SIAM J. NUMER. ANAL.
Vol. 46, No. 1, pp. 1–36
© 2007 Society for Industrial and Applied Mathematics

HOMOGENIZATION OF PARABOLIC EQUATIONS WITH A CONTINUUM OF SPACE AND TIME SCALES*

HOUMAN OWHADI† AND LEI ZHANG‡

Abstract. This paper addresses the issue of the homogenization of linear divergence form parabolic operators in situations where no ergodicity and no scale separation in time or space are available. Namely, we consider divergence form linear parabolic operators in $\Omega \subset \mathbb{R}^n$ with $L^\infty(\Omega \times (0,T))$-coefficients. It appears that the inverse operator maps the unit ball of $L^2(\Omega \times (0,T))$ into a space of functions which at small (time and space) scales are close in H^1 norm to a functional space of dimension n. It follows that once one has solved these equations at least n times it is possible to homogenize them both in space and in time, reducing the number of operation counts necessary to obtain further solutions. In practice we show under a Cordes-type condition that the first order time derivatives and second order space derivatives of the solution of these operators with respect to caloric coordinates are in L^2 (instead of H^{-1} with Euclidean coordinates). If the medium is time-independent, then it is sufficient to solve n times the associated elliptic equation in order to homogenize the parabolic equation.

Key words. multiscale problem, compensation, homogenization, upscaling, compression

AMS subject classifications. Primary, 34E13, 35B27; Secondary, 35B26, 68P30, 60F05, 35B05

DOI. 10.1137/060670420

1. Introduction and main results.

Let Ω be a bounded and convex domain of class C^2 of $\mathbb{R}^n$. Let $T > 0$. Consider the following parabolic PDE:

$$\begin{cases} \partial_t u = \operatorname{div}\big(a(x,t)\nabla u(x,t)\big) + g \quad \text{in} \quad \Omega \times (0,T), \\ u(x,t) = 0 \quad \text{for} \quad (x,t) \in \big(\partial\Omega \times (0,T)\big) \cup \big(\Omega \times \{t=0\}\big). \end{cases} \tag{1.1}$$

Write $\Omega_T := \Omega \times (0,T)$. g is a function in $L^2(\Omega_T)$. $(x,t) \to a(x,t)$ is a mapping from Ω_T into the space of symmetric positive definite matrices with entries in $L^\infty(\Omega_T)$. Assume a to be uniformly elliptic on the closure of Ω_T. This paper addresses the issue of the homogenization of (1.1) in space and time in situations where scale separation and ergodicity at small scales are not available (see [16], [43], [4] for an introduction to classical homogenization theory).

We will introduce in subsection 1.1 theorems establishing under Cordes-type conditions the increase of regularity of solutions of (1.1) when derivatives are taken with respect to caloric coordinates instead of Euclidean coordinates. In subsections 1.2 and 1.3 these results will be used to homogenize (1.1) in space and in time. More precisely, assume a to be written on a fine tessellation with N degrees of freedom. If a is time-independent, then by solving n times an elliptic boundary value problem associated to (1.1) (at a cost of $O(N(\ln N)^{n+3})$ operations, using the hierarchical matrix method [11] or in $O(N)$ operations using iterative methods (see [76], [77]) it

*Received by the editors September 22, 2006; accepted for publication (in revised form) August 10, 2007; published electronically December 12, 2007.
http://www.siam.org/journals/sinum/46-1/67042.html

†Department of Applied and Computational Mathematics, Control, and Dynamical Systems, California Institute of Technology MC 217-50, Pasadena, CA 91125 (owhadi@caltech.edu).

‡Department of Applied and Computational Mathematics, California Institute of Technology, MC 217-50, Pasadena, CA 91125 (zhanglei@acm.caltech.edu).

is possible to approximate the solutions of (1.1) by solving an homogenized operator with N^α degrees of freedom ($\alpha < 1$, $\alpha = 0.2$, for instance) or with a fixed number M of degrees of freedom (numerical experiments given at the end of this paper have been conducted with $N = 16641$ and $M = 9$). Therefore, in the case of time-independent a, it is possible to reduce the number of operations even if we solve the problem (1.1) only once by using large time steps. If a is also characterized by a continuum of time scales, then the method presented here does not reduce the number of operation counts necessary to solve (1.1) only one time. However, if one needs to solve (1.1) K ($K > n$) times (with different right-hand sides), then by solving (1.1) n times it is possible to obtain an approximation of the solutions of (1.1) by solving an homogenized (in space and time) parabolic equation written on a coarse tessellation with coarse time steps. From a broader point of view, this paper suggests that, when one is trying to solve the linear problem $Ax = y$ with many different y (where A is a matrix associated to a divergence form operator), one should never compute A^{-1}.

In [60] the authors have addressed the issue of the homogenization of divergence form elliptic equations with a continuum of space scales; however, this paper is not a technical extension of the previous one. Indeed the presence of temporal scales adds strong difficulties through the fact that for parabolic equations finite element methods are no longer optimal. The main difference when the coefficients are rough in both space and time lies in the fact that we have to introduce time-dependent elements and use dual methods for controlling the numerical homogenization errors instead of simply relying on the increased regularity. Furthermore the numerical scheme has to be implicit in time.

Numerical results given in section 3 show that the method can be implemented and work in practice. The idea to use a change of coordinates for a finite element method for divergence form elliptic equations with rough coefficients can be traced back to the work of Babuška et al. [9, 8] (in dimension two with coefficients depending only on one parameter). The idea to use a global change of coordinates to homogenize transport equations is currently being implemented in the industry and has been shown to give accurate results when local methods fail. For instance, we refer to [75], [74] for reservoir modeling in geophysics; in this situation the porosity of the medium is time-independent, and one has to solve an elliptic equation only at $t = 0$ to upscale the transport equations. Observe also that using an implicit method to solve divergence form parabolic equations with time-independent rough coefficients implies solving an elliptic problem at each time step. The method presented in this paper implies that these elliptic problems can be solved with an "homogenized" operator leading to a sharp improvement of the computational cost.

Although we use the term "homogenization" in this paper for lack of a better word, the method presented in this paper is more general than homogenization. Homogenization is typically associated with the limit points of a sequence of operators. Here we are speaking of a fixed operator.

1.1. Compensation phenomenon. Let F be the solution of the following parabolic equation:

$$\begin{cases} \partial_t F = \operatorname{div}\big(a(x,t)\nabla F(x,t)\big) \quad \text{in} \quad \Omega_T, \\ F(x,t) = x \quad \text{for} \quad (x,t) \in \big(\partial\Omega \times (0,T)\big), \\ \operatorname{div}\big(a(x,0)\nabla F(x,0)\big) = 0 \quad \text{in} \quad \Omega. \end{cases} \tag{1.2}$$

By (1.2) we mean that $F := (F_1, \dots, F_n)$ is a n-dimensional vector field such that each of its entries satisfies

$$\begin{cases} \partial_t F_i = \operatorname{div}\big(a(x,t)\nabla F_i(x,t)\big) & \text{in} \quad \Omega_T, \\ F_i(x,t) = x_i & \text{for} \quad (x,t) \in \big(\partial\Omega \times (0,T)\big), \\ \operatorname{div}\big(a(x,0)\nabla F_i(x,0)\big) = 0 & \text{in} \quad \Omega. \end{cases} \tag{1.3}$$

In the literature F is also called a caloric coordinate associated to (1.1). Observe that if a is time-independent, then F is the solution of an elliptic boundary value problem.

Definition 1.1. *Write*

$$\sigma := {}^t\nabla F a \nabla F. \tag{1.4}$$

Write β_σ as the Cordes parameter associated to σ defined by

$$\beta_\sigma := \operatorname{esssup}_{(x,t)\in\Omega_T} \left(n - \frac{\big(\operatorname{Trace}[\sigma]\big)^2}{\operatorname{Trace}[{}^t\sigma\sigma]} \right). \tag{1.5}$$

Observe that since

$$\beta_\sigma = \operatorname{esssup}_{(x,t)\in\Omega_T} \left(n - \frac{\big(\sum_{i=1}^n \lambda_{i,\sigma(x,t)}\big)^2}{\sum_{i=1}^n \lambda_{i,\sigma(x,t)}^2} \right), \tag{1.6}$$

where $(\lambda_{i,M})$ denotes the eigenvalues of M, β_σ is a measure of the anisotropy of σ. The greater β_σ is, the more anisotropic σ is, and $\beta_\sigma = 0$ corresponds to an isotropic σ.

1.1.1. Time-independent medium. In this subsection we assume that a does not depend on time t. Write for $p \geq 2$, $W_D^{2,p}$ (D for Dirichlet boundary condition) the Banach space $W^{2,p}(\Omega) \cap W_0^{1,p}(\Omega)$. Equip $W_D^{2,p}(\Omega)$ with the norm

$$\|v\|^p_{W_D^{2,p}(\Omega)} := \int_\Omega \left(\sum_{i,j} (\partial_i \partial_j v)^2 \right)^{\frac{p}{2}}. \tag{1.7}$$

Equip the space $L^p(0,T,W_D^{2,p}(\Omega))$ with the norm

$$\|v\|^p_{L^p(0,T,W_D^{2,p}(\Omega))} = \int_0^T \int_\Omega \left(\sum_{i,j} (\partial_i \partial_j v)^2 \right)^{\frac{p}{2}} dx\, dt. \tag{1.8}$$

Theorem 1.1. *Assume that $\partial_t a \equiv 0$, $g \in L^2(\Omega_T)$, Ω is convex, $\beta_\sigma < 1$, and $(\operatorname{Trace}[\sigma])^{\frac{n}{4}-1} \in L^\infty(\Omega)$; then $u \circ F^{-1} \in L^2(0,T,W_D^{2,2}(\Omega))$ and*

$$\|u \circ F^{-1}\|_{L^2(0,T,W_D^{2,2}(\Omega))} \leq \frac{C}{1-\beta_\sigma^{\frac{1}{2}}} \|g\|_{L^2(\Omega_T)}. \tag{1.9}$$

Remark 1.1. The constant C can be written

$$C = \frac{C_n}{(\lambda_{\min}(a))^{\frac{n}{4}}} \big\| (\operatorname{Trace}[\sigma])^{\frac{n}{4}-1} \big\|_{L^\infty(\Omega)}.$$

Throughout this paper, we write

$$\lambda_{\min}(a) := \inf_{(x,t)\in\Omega_T} \inf_{l\in\mathbb{R}^n, |l|=1} {}^t l.a(x,t).l. \tag{1.10}$$

All of the results are stated in a finite time interval [0, T]. T is just any given finite time, and it is easy to observe that if $\|g\|_{L^2(\Omega_\infty)} < \infty$, then one can take $T = \infty$.

Remark 1.2. According to Theorem 1.1 although the second order derivatives of u with respect to Euclidean coordinates are only in $L^2(0, T, H^{-1}(\Omega))$, they are in $L^2(\Omega_T)$ with respect to harmonic coordinates. The proof of that theorem (given in subsection 2.1.1) lies on the fact $u \circ F^{-1}$ satisfies nondivergence form parabolic equations, and the solutions of these equations are known to have a higher regularity under a Cordes-type condition on the coefficients and under a convexity assumption on the domain. One can obtain a similar result without the convexity assumption on the domain (Theorem 2.6), but the Cordes-type condition on the coefficients would have to be stronger.

Remark 1.3. Observe that if a is time-independent, then F and σ are time-independent, and F is the solution of the following elliptic problem:

$$\begin{cases} \operatorname{div} a\nabla F = 0 \quad \text{in} \quad \Omega, \\ F(x) = x \quad \text{on} \quad \partial\Omega. \end{cases} \tag{1.11}$$

In dimension one F is trivially a homeomorphism. In dimension two this property follows from the convexity of the domain [3] (even with $a_{i,j} \in L^\infty(\Omega)$), [6] (one can also deduce from [6] that for $n = 2$, if a is smooth, then the conditions $\beta_\sigma < 1$ and $(\text{Trace}[\sigma])^{-1} \in L^\infty(\Omega)$ are satisfied). In dimension three and higher F can be nonbijective even if a is smooth; we refer to [6], [22]. However in dimension three the assumption $(\text{Trace}[\sigma])^{\frac{n}{4}-1} \in L^\infty(\Omega_T)$ implies that F is a homeomorphism. If $n \geq 4$, we need to assume that F is a homeomorphism to prove the theorem.

Remark 1.4. In fact the condition $(\text{Trace}(\sigma))^{-1} \in L^p(\Omega_T)$, for $p < \infty$ depending on n, is sufficient to obtain Theorem 1.1 and the following compensation theorems. For the sake of clarity this paper has been restricted to $(\text{Trace}(\sigma))^{-1} \in L^\infty(\Omega_T)$.

Remark 1.5. Write

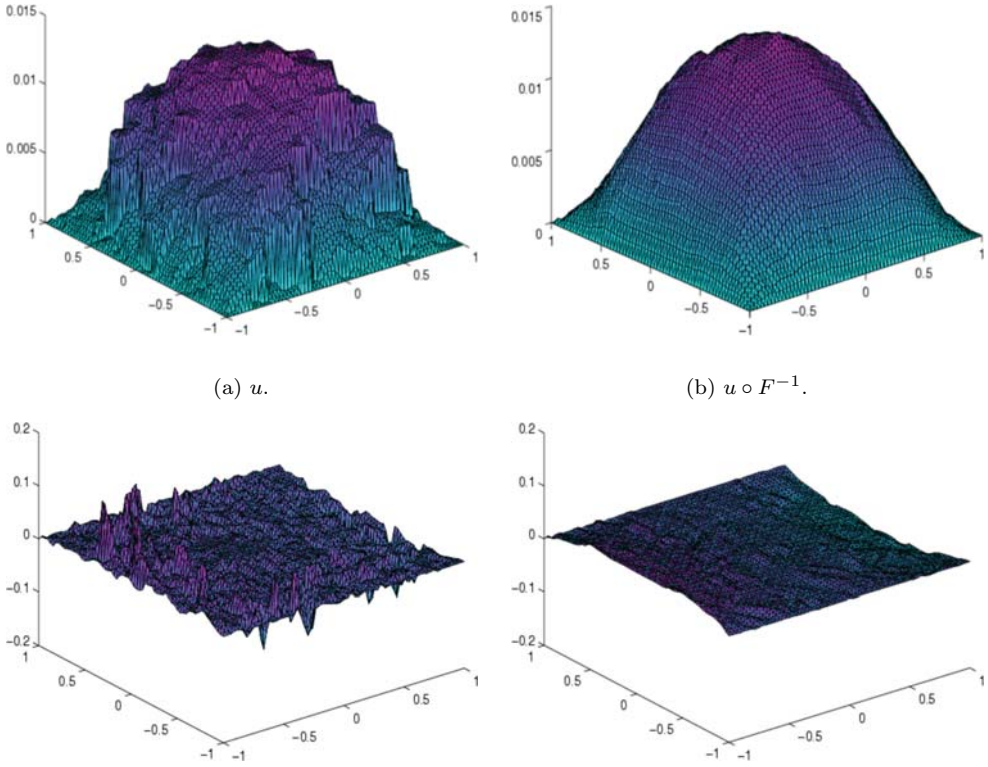
$$\mu_\sigma := \text{esssup}_{\Omega_T} \frac{\lambda_{\max}(\sigma)}{\lambda_{\min}(\sigma)}. \tag{1.12}$$

It is easy to check that μ_σ is bounded by an increasing function of $(1 - \beta_\sigma)^{-1}$, and in dimension two $\beta_\sigma < 1$ is equivalent to $\mu_\sigma < \infty$.

Remark 1.6. Theorem 1.1 has been called the compensation phenomenon because the composition by F^{-1} increases the regularity of $u \in L^2(0, T, H^1_0(\Omega))$. The choice of this name has been motivated by Murat and Tartar's work on H-convergence [54] which is also based on a regularization property called compensated compactness or the div-curl lemma introduced in the 1970s by Murat and Tartar [53], [70] (we also refer to [26] for refinements of the div-curl lemma).

The compensation phenomena presented in this subsection can be observed numerically. We consider the time-independent site percolation, and the value of a is set to be equal to 1 or 100 with probability 1/2 on each triangle of a fine mesh characterized by 16641 nodes and 32768 triangles. (1.1) has been solved numerically on that mesh with $g = 1$. u, $u \circ F^{-1}$, $\partial_x u$, and $\partial x(u \circ F^{-1})$ are plotted at time $t = 1$ in Figure 1.1.

It is easy to check that if $a = e(x)S(t)$, where e, is a time-independent symmetric uniformly elliptic matrix with $L^\infty(\Omega)$ entries and S is a smooth uniformly positive function, then the results given in this subsection and the homogenization schemes of subsection 1.2 remain valid with the time-independent harmonic coordinates associated to e, i.e., solution of $-\operatorname{div} e\nabla F = 0$.



(a) u. (b) $u \circ F^{-1}$.

(c) $\partial_x u$. (d) $\partial_x (u \circ F^{-1})$.

FIG. 1.1. *u, $u \circ F^{-1}$, $\partial_x u$, and $\partial_x (u \circ F^{-1})$ at time $t = 1$ for the time-independent site percolating medium.*

1.1.2. Medium with a continuum of time scales. In this subsection the entries of a are merely in $L^\infty(\Omega_T)$. We need to introduce the following Cordes-type condition.

CONDITION 1.1. *We say that Condition* 1.1 *is satisfied if and only if there exists $\delta \in (0, \infty)$ and $\epsilon > 0$ such that*

$$\operatorname{esssup}_{\Omega_T} \frac{\delta^2 \operatorname{Trace}[{}^t\sigma\sigma] + 1}{\left(\delta \operatorname{Trace}[\sigma] + 1\right)^2} \leq \frac{1}{n + \epsilon}. \tag{1.13}$$

Remark 1.7. This remark is due to Gyrya and Berlyand. Equation (1.13) can be written in terms of the Cordes parameter by

$$\operatorname{esssup}_{\Omega_T} \left\{ n - \frac{\left(\operatorname{Trace}(\sigma) + \delta^{-1}\right)^2}{\operatorname{Trace}(\sigma^t\sigma) + \delta^{-2}} \right\} \leq -\varepsilon. \tag{1.14}$$

An alternative way of interpreting the Cordes-type condition is by rewriting (1.14) as follows:

$$\beta_{\hat{\sigma}(\delta)} = \operatorname{esssup}_{\Omega_T} \left\{ (n+1) - \frac{\left(\operatorname{Trace}(\sigma) + \delta^{-1}\right)^2}{\operatorname{Trace}(\sigma^t\sigma) + \delta^{-2}} \right\} \leq 1 - \varepsilon, \tag{1.15}$$

where

$$\hat{\sigma}(\delta) = \begin{bmatrix} \delta^{-1} & 0 & \dots & 0 \\ 0 & & & \\ \vdots & & \sigma & \\ 0 & & & \end{bmatrix}.$$

Write

$$z_\sigma := \operatorname{esssup}_{\Omega_T} n \frac{\operatorname{Trace}[{}^t\sigma\sigma]}{(\operatorname{Trace}[\sigma])^2}. \tag{1.16}$$

Observe that z_σ is a measure of anisotropy of σ, in particular, $1 \leq z_\sigma \leq n$ and $z_\sigma = 1$ if σ is isotropic. Write

$$y_\sigma := \|\operatorname{Trace}[\sigma]\|_{L^\infty(\Omega_T)} \left\|(\operatorname{Trace}[\sigma])^{-1}\right\|_{L^\infty(\Omega_T)}. \tag{1.17}$$

PROPOSITION 1.1. *If* $\|\operatorname{Trace}[\sigma]\|_{L^\infty(\Omega_T)} < \infty$ *and* $\left\|(\operatorname{Trace}[\sigma])^{-1}\right\|_{L^\infty(\Omega_T)} < \infty$, *then Condition* 1.1 *is satisfied with*

$$\delta := n \left\|(\operatorname{Trace}[\sigma])^{-1}\right\|_{L^\infty(\Omega_T)} \tag{1.18}$$

and with $\epsilon := \frac{2ny_\sigma - n}{2ny_\sigma^2}$ *provided that* $z_\sigma \leq 1 + \frac{\epsilon}{n}$.

Remark 1.8. Observe that in dimension one $z_\sigma = 1$, and thus for $n = 1$ Condition 1.1 is satisfied if $\operatorname{Trace}[\sigma] \in L^\infty(\Omega_T)$ and $(\operatorname{Trace}[\sigma])^{-1} \in L^\infty(\Omega_T)$.

We have the following compensation theorem for a time-dependent medium.

THEOREM 1.2. *Assume that* Ω *is convex, and Condition* 1.1 *is satisfied, then* $u \circ F^{-1} \in L^2(0, T, W_D^{2,2}(\Omega))$, $\partial_t(u \circ F^{-1}) \in L^2(\Omega_T)$, *and*

$$\|u \circ F^{-1}\|_{L^2(0,T,W_D^{2,2}(\Omega))} + \|\partial_t(u \circ F^{-1})\|_{L^2(\Omega_T)} \leq C\|g\|_{L^2(\Omega_T)}, \tag{1.19}$$

where C *depends on* Ω, n, δ, *and* ϵ.

Remark 1.9. According to Theorem 1.2 although the second order space derivatives and first order time derivatives of u with respect to Euclidean coordinates are only in $L^2(0, T, H^{-1}(\Omega))$, they are in $L^2(\Omega_T)$ with respect to caloric coordinates. As in the time-independent, the proof (given in subsection 2.1.2) lies on the fact that $u \circ F^{-1}$ satisfies a nondivergence form parabolic equation.

The compensation phenomena can be observed numerically. We consider in dimension $n = 2$

$$a(x, y, t) = \frac{1}{6}\left(\sum_{i=1}^{5} \frac{1.1 + \sin(2\pi x'/\epsilon_i)}{1.1 + \sin(2\pi y'/\epsilon_i)} + \sin(4x'^2 y'^2) + 1\right), \tag{1.20}$$

with $x' = x + \sqrt{2}t$, $y' = y - \sqrt{2}t$, $\epsilon_1 = \frac{1}{5}$, $\epsilon_2 = \frac{1}{13}$, $\epsilon_3 = \frac{1}{17}$, $\epsilon_4 = \frac{1}{31}$, and $\epsilon_5 = \frac{1}{65}$. This medium has been plotted in Figure 1.2 at time 0 (observe that $\lambda_{\max}(a)/\lambda_{\min}(a) \sim 100$).

Equation (1.1) has been solved numerically on that mesh with $g \equiv 1$ on the fine mesh with 16641 nodes and 32768 triangles. Figure 1.3 shows $\partial_x u$ and $\partial_x(u \circ F^{-1})$ at time 0.3. In Figures 1.4 and 1.5, the value of x_0 is set to $x_0 := (0.75, -0.25)$, and the curves $t \to u(x_0, t), u \circ F^{-1}(x_0, t), \nabla u(x_0, t)$, and $\nabla u \circ F^{-1}(x_0, t)$ are plotted from $t = 0$ to $t = 0.3$.

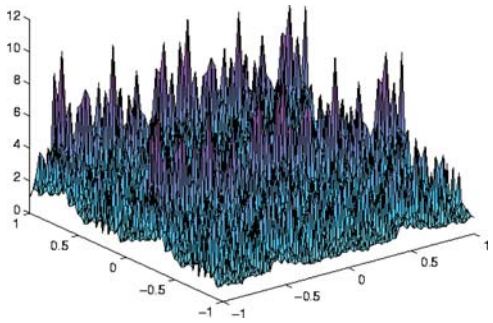
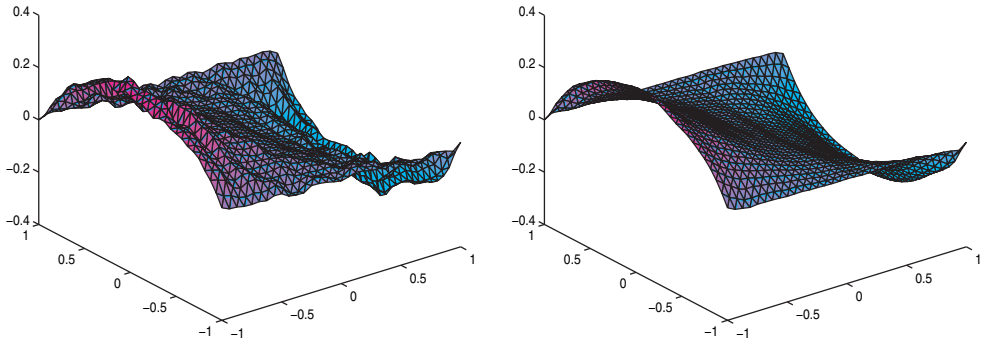
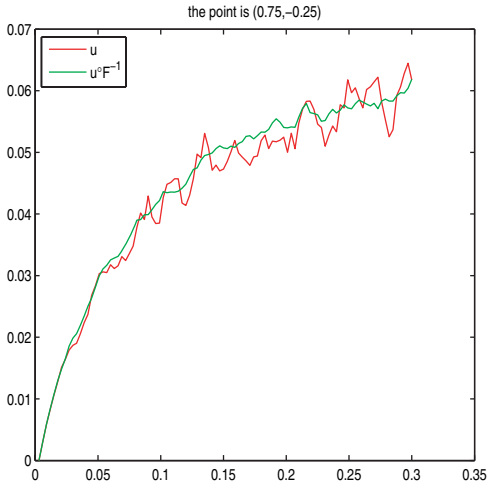


FIG. 1.2. *a at time* 0.



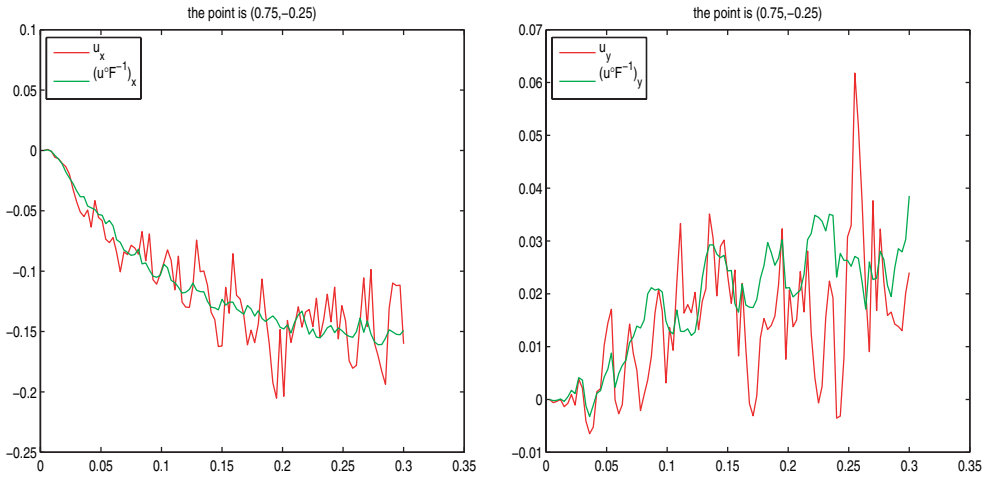
(a) $\partial_x u$. (b) $\partial_x (u \circ F^{-1})$.

FIG. 1.3. *$\partial_x u$ and $\partial_x (u \circ F^{-1})$ at time $t = 0.3$ for the multiscale trigonometric time-dependent medium.*



(a) u and $u \circ F^{-1}$.

FIG. 1.4. *$t \to u(x_0, t), u \circ F^{-1}(x_0, t)$ from $t = 0$ to $t = 0.3$, with $x_0 := (0.75, -0.25)$.*



(a) $\partial_x u$ and $\partial_x(u \circ F^{-1})$. (b) $\partial_y u$ and $\partial_y(u \circ F^{-1})$.

FIG. 1.5. $t \to \nabla u(x_0,t), \nabla u \circ F^{-1}(x_0,t)$ *from* $t=0$ *to* $t=0.3$, *with* $x_0 := (0.75,-0.25)$.

1.2. Numerical homogenization in space. Let X^h be a finite-dimensional subspace of $H_0^1(\Omega) \cap W^{1,\infty}(\Omega)$ ($W^{1,\infty}$ is the usual space of uniformly Lipschitz continuous functions) with the following approximation property: There exists a constant C_X such that for all $f \in W_D^{2,2}(\Omega)$

$$\inf_{v \in X^h} \|f - v\|_{H_0^1(\Omega)} \leq C_X h \|f\|_{W_D^{2,2}(\Omega)}. \tag{1.21}$$

It is known and easy to check that the set of piecewise linear functions on a triangulation of Ω satisfies condition (1.21) provided that the length of the edges of the triangles are bounded by h (C_X in (1.21) being given by the aspect ratio of the triangles).

For media characterized by a continuum of time scales we will consider twice differentiable elements satisfying the following usual inverse inequalities (see section 1.7 of [30]): For $v \in X^h$,

$$\|v\|_{W_D^{2,2}(\Omega)} \leq C_X h^{-1} \|v\|_{H_0^1(\Omega)}. \tag{1.22}$$

and

$$\|v\|_{H_0^1(\Omega)} \leq C_X h^{-1} \|v\|_{L^2(\Omega)}. \tag{1.23}$$

In this paper we will use splines to ensure that condition (1.22) is satisfied (observe that it requires the quasi uniformity of the (coarse) mesh, i.e., a bound on the aspect ratio of the (coarse) triangles).

For $t \in (0,T)$ let us define

$$V_h(t) := \{\varphi \circ F(x,t) \,:\, \varphi \in X^h\}. \tag{1.24}$$

Write $L^2(0,T;H_0^1(\Omega))$, the usual Sobolev space associated to the norm

$$\|v\|^2_{L^2(0,T;H_0^1(\Omega))} := \int_0^T \|v(.,t)\|^2_{H_0^1(\Omega)}\, dt. \tag{1.25}$$

Define Y_T^h, the subspace of $L^2(0,T;H_0^1(\Omega))$, as

$$Y_T^h := \{v \in L^2\big(0,T;H_0^1(\Omega)\big) : v(x,t) \in V_h(t)\ \forall t \in [0,T]\}. \tag{1.26}$$

Write u_h, the solution in Y_T^h of the following system of ordinary differential equations:

$$\begin{cases} (\psi, \partial_t u_h)_{L^2(\Omega)} + a[\psi, u_h] = (\psi, g)_{L^2(\Omega)} & \text{for all } t \in (0,T) \text{ and } \psi \in V_h(t), \\ u_h(x,0) = 0. \end{cases} \tag{1.27}$$

Write

$$a[v,w] := \int_\Omega {}^t\nabla v(x,t) a(x,t) \nabla w(x,t)\, dx. \tag{1.28}$$

1.2.1. Time-independent domain. We have the following theorem.

THEOREM 1.3. *Assume that $\partial_t a \equiv 0$, Ω is convex, $\beta_\sigma < 1$, and $(\mathrm{Trace}[\sigma])^{-1} \in L^\infty(\Omega_T)$; then*

$$\big\|(u-u_h)(.,T)\big\|_{L^2(\Omega)} + \big\|u-u_h\big\|_{L^2(0,T;H_0^1(\Omega))} \leq Ch\|g\|_{L^2(\Omega_T)}. \tag{1.29}$$

Remark 1.10. The constant C depends on C_X, n, Ω, $\lambda_{\min}(a)$, and $\|(\mathrm{Trace}[\sigma])^{-1}\|_{L^\infty(\Omega_T)}$. If $n \geq 5$, it also depends on $\|\,\mathrm{Trace}[\sigma]\|_{L^\infty(\Omega_T)}$, and if $n = 1$, it also depends on $\lambda_{\max}(a)$.

1.2.2. Medium with a continuum of time scales.

THEOREM 1.4. *Assume that Ω is convex, and Condition* 1.1 *is satisfied; then*

$$\big\|(u-u_h)(T)\big\|_{L^2(\Omega)} + \big\|u-u_h\big\|_{L^2(0,T,H_0^1(\Omega))} \leq Ch\|g\|_{L^2(\Omega_T)}. \tag{1.30}$$

Remark 1.11. The constant C depends on C_X, n, Ω, δ, ϵ, $\lambda_{\min}(a)$, and $\lambda_{\max}(a)$.

The system of ordinary differential equations (1.27) is still characterized by a continuum of time scales in situations where the entries of a merely belong to $L^\infty(\Omega_T)$. They need to be discretized (homogenized) in time in order to be solved numerically. This will be the object of the next subsection. Loosely speaking, although (1.1) is associated to a fine tessellation and fine time steps, it is possible to approximate its operator on a coarse tessellation with coarse time steps.

1.3. Numerical homogenization in space and time. Let $M \in \mathbb{N}$, and $\{t_n = n\frac{T}{M}\}_{0\leq n\leq M}$ is a discretization of $[0,T]$. Assume (φ_i) is a basis of X^h. Write Z_T^h, the subspace of Y_T^h, such that

$$Z_T^h = \Big\{ w \in Y_T^h : w(x,t) = \sum_i c_i(t)\varphi_i\big(F(x,t)\big),\ c_i(t) \text{ are constants on } (t_n, t_{n+1}] \Big\}. \tag{1.31}$$

Write U_T^h, the subspace of Y_T^h, such that

$$U_T^h = \left\{ \psi \in Y_T^h : \psi(x,t) = \sum_i d_i\varphi_i\big(F(x,t)\big),\ d_i \text{ are constants on } [0,T] \right\}. \tag{1.32}$$

For $w \in Y_T^h$, define $w_n \in U_T^h$ by

$$w_n(x,t) := \sum_i c_i(t_n)\varphi_i(F(x,t)). \tag{1.33}$$

Write v, the solution in Z_T^h of the following system of implicit weak formulation (such that $v(x,0)\equiv 0$): For $n\in\{0,\dots,M-1\}$ and $\psi\in V$,

$$\begin{aligned}\big(\psi(t_{n+1}),v_{n+1}(t_{n+1})\big)_{L^2(\Omega)} =&\big(\psi(t_n),v_n(t_n)\big)_{L^2(\Omega)}\\ &+\int_{t_n}^{t_{n+1}}\Big(\big(\partial_t\psi(t),v_{n+1}(t)\big)_{L^2(\Omega)}\\ &-a\big[\psi(t),v_{n+1}(t)\big]+\big(\psi(t),g(t)\big)_{L^2(\Omega)}\Big)\,dt.\end{aligned}\tag{1.34}$$

The following theorem shows the stability of the implicit scheme (1.34).

THEOREM 1.5. *Let $v\in Z_T$ be the solution of* (1.34)*. We have*

$$\big\|v(T)\big\|_{L^2(\Omega)}+\|v\|_{L^2(0,T,H_0^1(\Omega))}\leq C\|g\|_{L^2(\Omega_T)}.\tag{1.35}$$

Remark 1.12. The constant C depends on n, Ω, and $\lambda_{\min}(a)$.

The following theorem gives an error bound on the accuracy of time discretization scheme (1.34) when a does not depend on time.

THEOREM 1.6. *Let $v\in Z_T$ be the solution of* (1.34) *and u_h be the solution of* (1.27)*. Assume that $\partial_t a\equiv 0$. We have*

$$\begin{aligned}\big\|(u_h-v)(T)\big\|_{L^2(\Omega)}+&\|u_h-v\|_{L^2(0,T,H_0^1(\Omega))}\leq C|\Delta t|\\ &\Big(\|\partial_t g\|_{L^2(0,T,H^{-1}(\Omega))}+\big\|g(.,0)\big\|_{L^2(\Omega)}\Big).\end{aligned}\tag{1.36}$$

Remark 1.13. The constant C depends on n, Ω, and $\lambda_{\min}(a)$.

Remark 1.14. Observe that the combination of Theorem 1.6 with Theorem 1.3 leads to the following inequality:

$$\begin{aligned}\big\|(u-u_h)(.,T)\big\|_{L^2(\Omega)}+\big\|u-u_h\big\|_{L^2(0,T;H_0^1(\Omega))}&\leq C(h+\Delta t)\\ \Big(\|g\|_{L^2(\Omega_T)}+\|\partial_t g\|_{L^2(0,T,H^{-1}(\Omega))}+\big\|g(.,0)\big\|_{L^2(\Omega)}\Big).&\end{aligned}\tag{1.37}$$

Hence the accuracy of the method is controlled by $\max(h,\Delta t)$.

The following theorem gives an error bound on the accuracy of the time discretization scheme (1.34) when a has no bounded time derivatives.

THEOREM 1.7. *Assume that Ω is convex, and Condition* 1.1 *is satisfied. Let $v\in Z_T$ be the solution of* (1.34) *and u_h be the solution of* (1.27)*, and we have*

$$\big\|(u_h-v)(T)\big\|_{L^2(\Omega)}+\|u_h-v\|_{L^2(0,T,H_0^1(\Omega))}\leq C\frac{|\Delta t|}{h}\|g\|_{L^2(\Omega_T)},\tag{1.38}$$

where C depends on Ω, n, δ, ϵ, $\lambda_{\min}(a)$, and $\lambda_{\max}(a)$.

Remark 1.15. Observe that the accuracy of the time discretization scheme (1.34) requires that $|\Delta t|<<h$ when a has no bounded time derivatives. Indeed the combination of Theorem 1.7 with (1.30) leads to the following inequality:

$$\begin{aligned}\big\|(u-u_h)(.,T)\big\|_{L^2(\Omega)}+\big\|u-u_h\big\|_{L^2(0,T;H_0^1(\Omega))}&\leq C\left(h+\frac{\Delta t}{h}\right)\\ &\|g\|_{L^2(\Omega_T)}.\end{aligned}\tag{1.39}$$

Hence the accuracy of the method is controlled by h if one chooses $\Delta t\sim h^2$. The fact that one must chose $\Delta t\sim h^2$ for optimal accuracy is a standard phenomenon for parabolic equations, and we refer to [48] for a reminder.

We refer to section 3 for numerical experiments.

1.4. Remarks on caloric coordinates F. In numerical simulations a is given as mapping from a fine mesh triangulation Ω into the set of positive definite symmetric matrices. In other words (in the time-independent case, for instance), a is a constant matrix in each triangle of a given fine mesh. In the time-dependent case on each fine mesh triangle a is a step function from $[0,T]$ into the set of positive definite symmetric matrices.

The caloric coordinates F have been computed on the same fine triangulation of Ω associated to a through a standard finite element method.

The homogenization results given above assume an underlying continuous medium and that F is a function on that continuous medium. As has been mentioned in [60] the proofs given in this paper can be extended to the case where the underlying medium is discrete. Indeed the core of those proofs is based on Theorems 2.1 and 2.2, which are straightforward adaptations of Theorem 1.2.1 of [52], and those theorems can be extended to the discrete case.

A similar issue lies in the fact that a may be a mapping from the continuous medium Ω. In that case the equation solved in practice is not (1.1) but a discretization of (1.1) associated to a discretization of a over a fine mesh triangulation of Ω. The control of the error between the solution of (1.1) and the discrete counterpart of (1.1) is standard.

1.5. Literature and further remarks. For early works on homogenization with random mixing coefficients we refer to [62], [45], [61], [64], [78], [46], [47], [37]. Papanicolaou, Stroock, and Varadhan [63] have considered a two-component Markov process $(x(t), y(t))$, where $y(t)$ is rapidly varying (and is not assumed to be ergodic in dimension one) and enters in the coefficients of the stochastic process driving $x(t)$. They have studied the convergence properties of $x(t)$ as the fluctuations of $y(t)$ becomes more rapid using the martingale approach to diffusion, developed by Stroock and Varadhan [67], [66], [69], [68].

Methods dealing with problems involving multiple scales are numerous and not limited to the ones cited in this paper. As an example of a different approach, we refer to [2], where the concept of Young measure on micropatterns is introduced to study variational problems leading to multiple small scales involving a small parameter.

The numerical homogenization method implemented in this paper is a finite element method. The idea of using oscillating test functions can be tracked back to the work of Murat and Tartar on homogenization and H-convergence; we refer in particular to [71] and [54]. Those papers also contain convergence proofs for the finite element method in an abstract setting for a sequence of H-converging elliptic operators (recall that the framework of H-convergence is independent from ergodicity or scale separation assumptions and is based on the compactness of any sequence of solutions of $-\operatorname{div} a_\epsilon \nabla u_\epsilon = g$, with uniformly bounded and elliptic conductivities a_ϵ; we also refer to the initial work of Spagnolo [65] for G-convergence).

The numerical implementation and practical application of oscillating test functions in numerical finite element homogenization have been called multiscale finite element methods and have been studied by several authors [8], [29], [55], [41], [40], [35], [39], [5]. The work of Hou and Wu [41] has been a large source of inspiration in numerical applications (particularly for reservoir modeling in geophysics; we refer to [74], [49], [1], [75] for recent developments) since it was leading to a coarse scale operator while keeping the fine scale structures of the solutions. The construction of the base functions is decoupled from element to element, leading to a scheme adapted to parallel computers. A proof of the convergence of the method is given in periodic

settings when the size of the heterogeneities is smaller than the grid size, and an oversampling technique is proposed to remove the so-called cell resonance error [42] when the size of the heterogeneities is comparable to the grid size.

Allaire and Brizzi [5] have introduced the composition rule in the formulation of a multiscale finite element and observed that a multiscale finite element method with higher order Lagrange polynomials would have a better accuracy. In fact, Babuska et al. introduced a so-called change of variable technique in the general setting of the partition of unity method using a p-version of finite elements. In [8], a special class of second order elliptic problems with essentially one-dimensional rough coefficients $a(x,y)=a(x)$ was considered. In [56], [57], conformal mapping was used to map the rough solution to a smoother function for elliptic problems in two dimensions with corners or interfaces.

In [60], we proved that if u is the solution of the divergence form elliptic equation

$$\begin{cases} -\operatorname{div}\big(a(x)\nabla u(x)\big)=g \quad \text{in} \quad \Omega,\\ u=0 \quad \text{in} \quad \partial\Omega, \end{cases} \tag{1.40}$$

and F are caloric coordinates defined by

$$\begin{cases} \operatorname{div} a\nabla F=0 \quad \text{in} \quad \Omega,\\ F(x)=x \quad \text{on} \quad \partial\Omega, \end{cases} \tag{1.41}$$

then, under the Cordes-type condition $\beta_\sigma<1$ on σ given by (1.4), one has for some $p>2$

$$\|u\circ F^{-1}\|_{W^{2,p}(\Omega)}\leq C\|g\|_{L^p(\Omega)}. \tag{1.42}$$

It has been deduced from this compensation phenomenon that numerical homogenization methods based on oscillating finite elements can converge in the presence of a continuum of scales if one uses global caloric coordinates to obtain the test functions instead of solutions of a local cell problem [60]. In dimension three and higher it has been known since the work of Fenchenko and Khruslov [34], [44] that the homogenization of divergence form elliptic operators $-\operatorname{div} a_\epsilon\nabla u_\epsilon=g$ can lead to a nonlocal homogenized operator if the sequence of matrices a_ϵ is uniformly elliptic but with entries uniformly bounded only in $L^1(\Omega)$. From a numerical point of view these nonlocal effects imply that a nonlocal numerical homogenization method cannot be avoided to obtain accuracy. Hence in [60], it is shown that the accuracy of local methods depends on the aspect ratio of the triangles of the tessellation with respect to caloric coordinates (which is not the case if one uses nonlocal finite elements; we refer to [60] for further discussions on the apparition of nonlocal effects in numerical homogenization). Recently Briane has shown [21] that this nonlocal effect is absent in dimension two in the H-convergence setting.

The phenomenon is similar here; however, observe that, if one has solved the initial parabolic equation at least n times and those solutions are (locally) linearly independent, it is also possible to use them as new coordinates for numerical homogenization. Observe that in dimension higher than three the caloric coordinates are not always invertible; an idea to bypass this difficulty could be either to choose the change of coordinates locally and adaptively or to enrich the coordinates by writing down the initial equations as degenerate equations in a space of higher dimension [72]; these points have not been explored. For divergence form elliptic equations, recall that fast methods based on hierarchical matrices are available [14], [13], [12], [10], [11] for solving

(1.40) and (1.41) in $O\big(N(\ln N)^{n+3}\big)$ operations (or in $O(N)$ operations using iterative methods (see [76], [77]), N being the number of interior nodes of the fine mesh).

The issue of numerical homogenization partial differential equations with heterogeneous coefficients has received a great deal of attention, and many methods have been proposed. A few of them are cited below.

- multiscale finite volume methods [50];
- heterogeneous multiscale methods [28], [24];
- wavelet-based homogenization [38], [27], [25], [17], [7], [19];
- residual-free bubble methods [20];
- discontinuous enrichment methods [33], [32];
- partition of unity methods [36];
- energy-minimizing multigrid methods [73].

Following [60], it is possible to implement a finite volume method based on the compensation theorems given in this paper. The elements given in this paper contain the fine scale structure of F, and has been done in [60], and it is possible to approximate the initial parabolic operator by a homogenized parabolic operator associated to the coarse mesh (the test functions in this case would be piecewise linear on the coarse mesh, and the approximation error associated to the homogenized operator would depend on the aspect ratio of the triangles of the coarse mesh in the metric induced by F).

In this paper a has been assumed to be bounded and uniformly elliptic. Without these assumptions the diffusion associated to a homogenized operator can be anomalously slow [15], [58] or fast (superdiffusive) [59]. If a has an unbounded skew symmetric component, the homogenization of (1.1) can give rise to a degenerate operator [59].

2. Proofs.

2.1. Compensation.

2.1.1. Time-independent medium: Proof of Theorem 1.1. We will need the following lemmas. Let $\mathcal{A}_T$ be the bilinear form on $L^2\big(0,T;H_0^1(\Omega)\big)$ defined by

$$\mathcal{A}_T[v,w] := \int_0^T a[v,w](t)\,dt, \tag{2.1}$$

where

$$a[v,u](t) := \int_\Omega {}^t\nabla v(x,t)a(x,t)\nabla u(x,t)\,dx. \tag{2.2}$$

We write $\mathcal{A}_T[u] := \mathcal{A}_T[u,u]$. The following lemmas are standard energy estimates, and we refer to [31, section 7.1.2] for a reminder.

LEMMA 2.1. *We have*

$$\big\|u(.,T)\big\|^2_{L^2(\Omega)} + \mathcal{A}_T[u] \leq \frac{C_{n,\Omega}}{\lambda_{\min}(a)}\|g\|^2_{L^2(\Omega_T)}. \tag{2.3}$$

LEMMA 2.2. *Assume $\partial_t a \equiv 0$, and we have*

$$\big\|\partial_t u\big\|^2_{L^2(\Omega_T)} + a\big[u(.,T)\big] \leq \big\|g\big\|^2_{L^2(\Omega_T)}. \tag{2.4}$$

LEMMA 2.3. *Assume $\partial_t a \equiv 0$, and we have*

$$\big\|\partial_t u(.,T)\big\|^2_{L^2(\Omega)} + \mathcal{A}_T[\partial_t u] \leq \frac{C_{n,\Omega}}{\lambda_{\min}(a)}\|\partial_t g\|^2_{L^2(0,T,H^{-1}(\Omega))} + \big\|g(.,0)\big\|^2_{L^2(\Omega)}. \tag{2.5}$$

We now need a variation of Campanato's result [23] on nondivergence form elliptic operators. Let us write for a symmetric matrix M

$$\nu_M := \frac{\mathrm{Trace}(M)}{\mathrm{Trace}({}^tMM)}. \tag{2.6}$$

We consider the following Dirichlet problem:

$$L_M v = f, \tag{2.7}$$

with $L_M := \sum_{i,j=1}^n M_{ij}(x)\partial_i\partial_j$. The following Theorems 2.1 and 2.2 are straightforward adaptations of Theorem 1.2.1 of [52]. They are proven in [52] under the assumption that M is bounded and elliptic. It is easy to check that the conditions $\beta_M < 1$ and $\nu_M \in L^\infty(\Omega)$ are sufficient for the validity of those theorems. We refer to [60] for that adaptation.

THEOREM 2.1. *Assume that $\beta_M < 1$, $\nu_M \in L^\infty(\Omega)$, and Ω is convex. Then if $f \in L^2(\Omega)$, the Dirichlet problem* (2.7) *has a unique solution satisfying*

$$\|v\|_{W_D^{2,2}(\Omega)} \leq \frac{C}{1-\beta_M^{\frac{1}{2}}}\|\nu_M f\|_{L^2(\Omega)}. \tag{2.8}$$

Remark 2.1. β_M is the Cordes parameter associated to M.

THEOREM 2.2. *Assume that $\beta_M < 1$, $\nu_M \in L^\infty(\Omega)$, and Ω is convex. Then there exists a real number $p_0 > 2$ depending only on n, Ω, and β_M such that for each $f \in L^p(\Omega)$, $2 \leq p < p_0$, the Dirichlet problem* (2.7) *has a unique solution satisfying*

$$\|v\|_{W_D^{2,p}(\Omega)} \leq \frac{C_{n,\Omega,p}}{1-\beta_M^{\frac{1}{2}}}\|\nu_M f\|_{L^p(\Omega)}. \tag{2.9}$$

Let us now prove the compensation theorems. Choose

$$M := \frac{\sigma}{|\det(\nabla F)|^{\frac{1}{2}}} \circ F^{-1}. \tag{2.10}$$

It is easy to check that $\beta_\sigma < 1$ implies that F is a homeomorphism from Ω onto Ω, and thus (2.10) is well defined. Moreover observe that $\beta_M = \beta_\sigma$ and

$$\|\nu_M\|^2_{L^\infty(\Omega_T)} \leq \frac{C_n}{(\lambda_{\min}(a))^{\frac{n}{2}}}\big\|(\mathrm{Trace}[\sigma])^{\frac{n}{4}-1}\big\|^2_{L^\infty(\Omega_T)}. \tag{2.11}$$

Fix $t \in [0,T]$. Choose

$$f := \frac{(\partial_t u - g)}{|\det(\nabla F)|^{\frac{1}{2}}} \circ F^{-1}. \tag{2.12}$$

Observe that by the change of variable $y = F(x)$ one obtains that, if $\partial_t a \equiv 0$ (which implies that F is time-independent), $\partial_t u \in L^2(\Omega)$, and $g(.,t) \in L^2(\Omega)$, $f \in L^2(\Omega)$ and

$$\|f\|_{L^2(\Omega)} \leq \|\partial_t u\|_{L^2(\Omega)} + \|g\|_{L^2(\Omega)}. \tag{2.13}$$

It follows from Theorem 2.1 that there exists a unique $v \in W_D^{2,2}(\Omega)$ satisfying

$$\|v\|^2_{W_D^{2,2}(\Omega)} \leq \frac{C\|\nu_M\|^2_{L^\infty(\Omega_T)}}{(1-\beta_\sigma^{\frac{1}{2}})^2}\big(\|\partial_t u\|^2_{L^2(\Omega)} + \|g\|^2_{L^2(\Omega)}\big) \tag{2.14}$$

and the following equation:

$$\partial_t \hat{u}(y,t) = \sum_{i,j} \big(\sigma(F^{-1}(y,t),t)\big)_{i,j} \partial_i \partial_j v(y,t) + \hat{g}(y,t). \tag{2.15}$$

We use the notation $\hat{g} := g \circ F^{-1}$ and $\hat{u} := u \circ F^{-1}$. Using the change of variable $y = F(x)$ and using the property $\operatorname{div} a\nabla F = 0$ when $\partial_t a \equiv 0$, we obtain that (2.15) can be written

$$\partial_t u = \operatorname{div}\big(a\nabla(v \circ F)\big) + g. \tag{2.16}$$

If $\partial_t u \in L^2(\Omega)$ and $g(.,t) \in L^2(\Omega)$, we can use the uniqueness property of the solution of the divergence form elliptic Dirichlet problem

$$\operatorname{div}\big(a\nabla w\big) = \partial_t u - g \tag{2.17}$$

to obtain that $v \circ F = u$. Assume that $g \in L^2(\Omega_T)$ and $\partial_t u \in L^2(\Omega_T)$. It follows that, for $t \in [0,T] - B$, $g(.,t) \in L^2(\Omega)$ and $\partial_t u(.,t) \in L^2(\Omega)$, where B is a subset of $[0,T]$ of 0-Lebesgue measure. It follows from the previous arguments that, on $[0,T] - B$, $u \circ F^{-1}(.,t) \in W_D^{2,2}(\Omega)$ and satisfies

$$\|u \circ F^{-1}(.,t)\|^2_{W_D^{2,2}(\Omega)} \leq \frac{C\|\nu_M\|^2_{L^\infty(\Omega_T)}}{(1-\beta_\sigma^{\frac{1}{2}})^2}\big(\|\partial_t u(.,t)\|^2_{L^2(\Omega)} + \|g(.,t)\|^2_{L^2(\Omega)}\big). \tag{2.18}$$

Integrating (2.18) with respect to time we obtain that $u \circ F^{-1} \in L^2(0,T,W_D^{2,2}(\Omega))$ and that

$$\|u \circ F^{-1}\|^2_{L^2(0,T,W_D^{2,2}(\Omega))} \leq \frac{C\|\nu_M\|^2_{L^\infty(\Omega_T)}}{(1-\beta_\sigma^{\frac{1}{2}})^2}\big(\|\partial_t u\|^2_{L^2(\Omega_T)} + \|g\|^2_{L^2(\Omega_T)}\big). \tag{2.19}$$

Thus using Lemma 2.2 we have obtained Theorem 1.1.

In situations where $g \in L^\infty(0,T,L^2(\Omega))$, $\partial_t g \in L^2(0,T,H^{-1}(\Omega))$, or $g \in L^p(\Omega_T)$, with $p > 2$, one can obtain a higher regularity for $u \circ F^{-1}$. This is the object of the following theorems.

THEOREM 2.3. *Assume that Ω is convex, $g \in L^\infty(0,T,L^2(\Omega))$, $\partial_t g \in L^2(0,T,H^{-1}(\Omega))$, $\partial_t a \equiv 0$, $\beta_\sigma < 1$, and $(\operatorname{Trace}[\sigma])^{\frac{n}{4}-1} \in L^\infty(\Omega_T)$; then for all $t \in [0,T]$, $u \circ F^{-1}(.,t) \in W_D^{2,2}(\Omega)$ and*

$$\|u \circ F^{-1}(.,t)\|_{W_D^{2,2}(\Omega)} \leq \frac{C}{1-\beta_\sigma^{\frac{1}{2}}}\Big(\big\|g\big\|_{L^\infty(0,T,L^2(\Omega))} + \|\partial_t g\|_{L^2(0,T,H^{-1}(\Omega))}\Big). \tag{2.20}$$

Remark 2.2. The constant C can be written

$$C = \frac{C_{n,\Omega}}{(\lambda_{\min}(a))^{\frac{n}{4}}}\big\|(\operatorname{Trace}[\sigma])^{\frac{n}{4}-1}\big\|_{L^\infty(\Omega_T)}\left(1 + \frac{1}{\lambda_{\min}(a)}\right)^{\frac{1}{2}}.$$

THEOREM 2.4. *Assume that Ω is convex, $g(.,0) \in L^2(\Omega)$, $\partial_t g \in L^2(0,T,H^{-1}(\Omega))$, $g \in L^p(\Omega_T)$, $\partial_t a \equiv 0$, $\beta_\sigma < 1$, and $(\operatorname{Trace}[\sigma])^{\frac{n}{4}-1} \in L^\infty(\Omega_T)$; then there exists a real number $p_0 > 2$ depending only on n, Ω, and β_σ such that for each p such that $2 \leq p < p_0$ one has*

$$\begin{aligned}\|u \circ F^{-1}\|_{L^p(0,T,W_D^{2,p}(\Omega))} \leq & \frac{C}{1-\beta_\sigma^{\frac{1}{2}}}\big(\|g\|_{L^p(\Omega_T)} \\ & + \big\|g(.,0)\big\|_{L^2(\Omega)} + \|\partial_t g\|_{L^2(0,T,H^{-1}(\Omega))}\big).\end{aligned} \tag{2.21}$$

Remark 2.3. The constant C can be written

$$C = \frac{C_{n,\Omega,p}}{(\lambda_{\min}(a))^{\frac{n}{4}}} \left\| (\mathrm{Trace}[\sigma])^{\frac{n}{4}-1} \right\|_{L^\infty(\Omega_T)} \left(1 + \frac{1}{\lambda_{\min}(a)}\right)^{\frac{1}{2}}.$$

Write

$$\|v\|_{C^\gamma(\Omega)} := \sup_{x,y\in\Omega, x\neq y} \frac{|v(x)-v(y)|}{|x-y|^\gamma}. \tag{2.22}$$

THEOREM 2.5. *Assume that $n \leq 2$, Ω is convex, $g(.,0) \in L^2(\Omega)$, $\partial_t g \in L^2(0, T, H^{-1}(\Omega))$, $g \in L^p(\Omega_T)$, $\partial_t a \equiv 0$, $\beta_\sigma < 1$, $(\mathrm{Trace}[\sigma])^{-1} \in L^\infty(\Omega_T)$, and $g \in L^2[0,T; L^{p^*}(\Omega)]$, with $2 < p^*$. Then there exists $p \in (2, p^*]$ and $\gamma(p) > 0$ such that*

$$\left(\int_0^T \left\| \nabla(u \circ F^{-1})(.,t) \right\|^2_{C^\gamma(\Omega)} dt \right)^{\frac{1}{2}} \leq \frac{C}{1-\beta_\sigma^{\frac{1}{2}}} \big(\|g\|_{L^p(\Omega_T)} + \left\| g(.,0) \right\|_{L^2(\Omega)} + \|\partial_t g\|_{L^2(0,T,H^{-1}(\Omega))} \big). \tag{2.23}$$

Remark 2.4. The constant C in (2.23) depends on n, p, Ω, $\lambda_{\min}(a)$, and $\|(\mathrm{Trace}(\sigma))^{-1}\|_{L^\infty(\Omega_T)}$. It is easy to check that if $n = 1$, then the theorem is valid with $\gamma = 1/2$.

Using Lemma 2.3 we can prove Theorem 2.3.

Let us now prove Theorem 2.4. Assume that there exists $q_0 > 2$ such that, for $2 \leq p < q_0$, $\partial_t u \in L^p(\Omega_T)$ and $g \in L^p(\Omega_T)$. Let us now apply Theorem 2.2 with $p < \min(p_0, q_0)$, M given by (2.10), and f given by (2.12). It follows that, for $t \in [0,T] - B$ (where B is a subset of $[0,T]$ of 0-Lebesgue measure), $g(.,t) \in L^p(\Omega)$ and $\partial_t u(.,t) \in L^p(\Omega)$. We deduce from Theorem 2.2 and the argumentation related to (2.17) that, on $[0,T] - B$, $u \circ F^{-1}(.,t) \in W_D^{2,p}(\Omega)$ and

$$\|u \circ F^{-1}(.,t)\|^p_{W_D^{2,p}(\Omega)} \leq \frac{C_{n,p,\Omega} \|\nu_M\|^p_{L^\infty(\Omega_T)}}{(1-\beta_\sigma^{\frac{1}{2}})^p} \big(\|\partial_t u(.,t)\|^p_{L^p(\Omega)} + \|g(.,t)\|^p_{L^p(\Omega)} \big). \tag{2.24}$$

Integrating (2.24) with respect to time, we have $u \circ F^{-1} \in L^p(0,T,W_D^{2,p}(\Omega))$ and

$$\|u \circ F^{-1}\|_{L^p(0,T,W_D^{2,p}(\Omega))} \leq \frac{C_{n,p,\Omega} \|\nu_M\|_{L^\infty(\Omega_T)}}{1-\beta_\sigma^{\frac{1}{2}}} \big(\|\partial_t u\|_{L^p(\Omega_T)} + \|g\|_{L^p(\Omega_T)} \big). \tag{2.25}$$

It remains to show that, under the assumptions of Theorem 2.4, $\partial_t u \in L^p(\Omega_T)$.

In order to bound $\left\|\partial_t u(.,t)\right\|_{L^p(\Omega)}$ we use general Sobolev inequalities (chapter 5.6 of [31]).

- If $n \geq 3$, write $p^* = 2n/(n-2)$. We have for $2 < p \leq p^*$

$$\left(\int_\Omega (\partial_t u)^p \, dx \right)^{\frac{2}{p}} \leq C_{n,\Omega} \left(\int_\Omega (\partial_t u)^{p^*} \, dx \right)^{\frac{2}{p^*}}, \tag{2.26}$$

and thus, using the Gagliardo–Nirenberg–Sobolev inequality,

$$\left(\int_\Omega (\partial_t u)^p \, dx \right)^{\frac{2}{p}} \leq C_{n,p,\Omega} \frac{1}{\lambda_{\min}(a)} a[\partial_t u]. \tag{2.27}$$

- If $n = 2$, we write, for $(x_1, x_2, x_3) \in \Omega \times (0,1)$, $v(x_1, x_2, x_3) := \partial_t u(x_1, x_2)$. Using the Gagliardo–Nirenberg–Sobolev inequality in dimension three we obtain that for $2 < p \leq 6$

$$\left(\int_\Omega (\partial_t u)^p \, dx \right)^{\frac{2}{p}} \leq C_{n,p,\Omega} \int_\Omega (\nabla \partial_t u)^2 \, dx, \tag{2.28}$$

  which leads us to (2.27).
- If $n = 1$, then using Morrey's inequality we obtain that, with $\gamma := 1/2$,

$$\|\partial_t u\|^2_{C^{0,\gamma}(\Omega)} \leq C_\Omega \frac{1}{\lambda_{\min}(a)} a[\partial_t u]. \tag{2.29}$$

We conclude the proof of Theorem 2.4 by using Lemma 2.3.

We deduce Theorem 2.5 from Morrey's inequality and Theorem 2.4.

Hölder continuity for $n \geq 3$ or nonconvexity of Ω. In this paragraph we will not assume Ω to be convex. Let $N^{p,\lambda}(\Omega)$ $(1 < p < \infty,\ 0 < \lambda < n)$ be the weighted Morrey space formed by the functions $v : \Omega \to \mathbb{R}$ such that $\|v\|_{N^{p,\lambda}(\Omega)} < \infty$, with

$$\|v\|_{N^{p,\lambda}(\Omega)} := \sup_{x_0 \in \Omega} \left(\int_\Omega |x - x_0|^{-\lambda} |v(x)|^p \right)^{\frac{1}{p}}. \tag{2.30}$$

To obtain the Hölder continuity of $u \circ F^{-1}$ in dimension $n \geq 3$ we use Corollary 4.1 of [51]. We give the result of Leonardi below in a form adapted to our context. Consider the Dirichlet problem (2.7). We write $W^{2,p,\lambda}(\Omega)$, the functions in $W^{2,p}_D(\Omega)$, such that their second order derivatives belong to $N^{p,\lambda}(\Omega)$.

THEOREM 2.6. *There exist a constant $C^* = C^*(n, p, \lambda, \partial\Omega) > 0$ such that if $\beta_M < C^*$ and $f \in N^{p,\lambda}(\Omega)$, then the Dirichlet problem (2.7) has a unique solution in $W^{2,p,\lambda} \cap W^{1,p}_0(\Omega)$. Moreover, if $0 < \lambda < n < p$, then $\nabla v \in C^\alpha(\Omega)$, with $\alpha = 1 - n/p$ and*

$$\|\nabla v\|_{C^\alpha(\Omega)} \leq \frac{C}{\lambda_{\min}(M)} \|f\|_{N^{p,\lambda}(\Omega)}, \tag{2.31}$$

where $C = C(n, p, \lambda, \partial\Omega)$.

Now we have the compensation theorem in the nonconvex case.

THEOREM 2.7. *Assume $n \geq 2$ and $\partial_t a \equiv 0$. Let $p > 2$. There exists a constant $C^* = C^*(n, \partial\Omega) > 0$, a real number $\gamma > 0$, depending only on n, Ω, and p such that if $\beta_\sigma < C^*$, then*

$$\left(\int_0^T \left\| \nabla (u \circ F^{-1})(., t) \right\|^2_{C^\gamma(\Omega)} dt \right)^{\frac{1}{2}} \leq C \big(\|g\|_{L^p(\Omega_T)} + \|g(., 0)\|_{L^2(\Omega)} + \|\partial_t g\|_{L^2(0,T,H^{-1}(\Omega))} \big). \tag{2.32}$$

Remark 2.5. The constant C in (2.32) depends on n, γ, Ω, C^*, $\lambda_{\min}(a)$, and $\left\| (\mathrm{Trace}(\sigma))^{\frac{n}{2p}-1} \right\|_{L^\infty(\Omega_T)}$.

Remark 2.6. Compared with the convex case, we have to assume F is an automorphism. Theorem 2.7 shows that convexity is not needed, but without convexity the Cordes-type assumption $\beta_\sigma < C^*$ is stronger than $\beta_\sigma < 1$.

The proof of Theorem 2.7 is an application of Theorem 2.6. We just need to observe that from the Hölder inequality we have for $0 < \epsilon < 0.5$

$$\|f\|_{N^{p,\epsilon}(\Omega)} \leq C_{n,p,\Omega,\epsilon} \|f\|_{L^{p(1+\epsilon)}(\Omega)}. \tag{2.33}$$

From this point the proof is similar to the proof of Theorem 2.4.

2.1.2. Medium with a continuum of time scales: Proof of Theorem 1.2. We will need Theorems 1.6.2 and 1.6.3 of [52]. For the sake of completeness we will recall those theorems below in a version adapted to our framework. Consider the following parabolic problem:

$$\partial_t v = \sum_{i,j=1}^{n} M_{ij}(x)\partial_i\partial_j v + f. \tag{2.34}$$

We assume M to be symmetric bounded and elliptic and $v = 0$ at $t = 0$ and on the boundary $\partial\Omega$. Write

$$\eta_M := \sup_{x\in\Omega_T} \frac{\mathrm{Trace}[{}^tMM] + 1}{\big(\mathrm{Trace}[M] + 1\big)^2} \tag{2.35}$$

and

$$\alpha_M := \sup_{x\in\Omega_T} \frac{\mathrm{Trace}[M] + 1}{\mathrm{Trace}[{}^tMM] + 1}. \tag{2.36}$$

Write for $p \geq 2$

$$S_p(\Omega_T) := \Big\{ v \in L^p\big(0, T, W_D^{2,p}(\Omega)\big); \partial_t v \in L^p(\Omega_T); v(.,0) \equiv 0 \Big\} \tag{2.37}$$

and

$$\|v\|^p_{S_p(\Omega_T)} := \int_{\Omega_T} \left(\sum_{i,j} (\partial_i\partial_j v)^2 + (\partial_t v)^2 \right)^{\frac{p}{2}} dy\, dt. \tag{2.38}$$

THEOREM 2.8. *Assume Ω to be convex and that there exists $\epsilon \in (0,1)$ such that $\eta_M \leq 1/(n+\epsilon)$; then for each $f \in L^2(\Omega_T)$ the Cauchy–Dirichlet problem* (2.34) *admits a unique solution in $S_2(\Omega_T)$ which satisfies the bound*

$$\|v\|_{S_2(\Omega_T)} \leq \frac{\alpha_M}{1 - \sqrt{1-\epsilon}} \|f\|_{L^2(\Omega_T)}. \tag{2.39}$$

THEOREM 2.9. *Assume Ω to be convex and that there exists $\epsilon \in (0,1)$ such that $\eta_M \leq 1/(n+\epsilon)$; then there exists a number $p_0 > 2$ depending on Ω, n, ϵ such that for each $f \in L^p(\Omega_T)$ the Cauchy–Dirichlet problem* (2.34) *admits a unique solution in $S_p(\Omega_T)$ which satisfies the bound*

$$\|v\|_{S_p(\Omega_T)} \leq C_p \frac{\alpha_M}{1 - \sqrt{1-\epsilon}} \|f\|_{L^p(\Omega_T)}. \tag{2.40}$$

Remark 2.7. In fact Theorem 1.6.3 of [52] is written with $1 - C(p)\sqrt{1-\epsilon}$ in the denominator of (2.40), but it is easy to modify it to obtain (2.40) by lowering the value of p_0 and changing the value of C_p.

Let $\delta > 0$. Let us now prove Proposition 1.1. Write $x = \mathrm{Trace}[\sigma]$ and $z = n\frac{\mathrm{Trace}[{}^t\sigma\sigma]}{(\mathrm{Trace}[\sigma])^2}$ (observe that $1 \leq z \leq n$). It is easy to check that Condition 1.1 can be written

$$-\delta^2 x^2 \left(\frac{\epsilon + n}{n} z - 1 \right) + 2x\delta - (n + \epsilon - 1) \geq 0. \tag{2.41}$$

Choose $\delta = n\big\|(\mathrm{Trace}[\sigma])^{-1}\big\|_{L^\infty(\Omega_T)}$. Observing that $\delta x \geq n$ and $\delta x = n y_\sigma$ it is easy to conclude the proof of Proposition 1.1. Similarly one obtains the following lemma by straightforward computation from (2.41).

LEMMA 2.4. *Assume that Condition* 1.1 *is satisfied; then* $\mu_\sigma < C(n, \epsilon, \delta)$

$$\big\|(\mathrm{Trace}[\sigma])^{-1}\big\|_{L^\infty(\Omega_T)} \leq C(n, \epsilon, \delta) \tag{2.42}$$

and

$$\big\|\,\mathrm{Trace}[\sigma]\big\|_{L^\infty(\Omega_T)} \leq C(n, \epsilon, \delta). \tag{2.43}$$

Let us now apply Theorem 2.8 on $[0, T/\delta]$ with

$$M := \delta\sigma \circ F^{-1}(y, \delta t) \tag{2.44}$$

and

$$f := \delta(g \circ F^{-1})(y, \delta t). \tag{2.45}$$

Observe that if Condition 1.1 is satisfied, then F is a homeomorphism and M is well-defined, bounded, and elliptic. Moreover $\eta_M < \infty$ and $\alpha_M < \infty$ since

$$\mathrm{esssup}_{\Omega_{\frac{T}{\delta}}} \frac{\mathrm{Trace}[{}^t M M] + 1}{\big(\,\mathrm{Trace}[M] + 1\big)^2} = \mathrm{esssup}_{\Omega_T} \frac{\delta^2\,\mathrm{Trace}[{}^t\sigma\sigma] + 1}{\big(\delta\,\mathrm{Trace}[\sigma] + 1\big)^2}\,. \tag{2.46}$$

It follows that the following equation admits a unique solution in $S_2(\Omega_{\frac{T}{\delta}})$:

$$\partial_t w(y,t) = \sum_{i,j} M_{i,j}(y,t)\partial_i\partial_j w(y,t) + k(y,t), \tag{2.47}$$

with $k(y,t) = \delta\hat{g}(y, \delta t)$. We also have

$$\int_0^{\frac{T}{\delta}} \int_\Omega \left((\partial_t w)^2 + \sum_{i,j} (\partial_i\partial_j w)^2 \right) dy\, dt \leq \frac{C}{(1 - \sqrt{1-\epsilon})^2} \|f\|_{L^2(\Omega_{\frac{T}{\delta}})}. \tag{2.48}$$

Using the change of variables $t \to \delta t$ and writing

$$w(y,t) := v(y, \delta t), \tag{2.49}$$

we obtain that v satisfies the following equation on Ω_T:

$$\partial_t v(y,t) = \sum_{i,j} \big(\sigma(F^{-1}(y,t), t)\big)_{i,j} \partial_i\partial_j v(y,t) + \hat{g}(y,t). \tag{2.50}$$

Using the change of variable $y = F(x)$ and observing that $\partial_t F = \mathrm{div}\, a\nabla F$, we obtain that $v \circ F$ satisfies

$$\partial_t(v \circ F) = \mathrm{div}\,\big(a\nabla(v \circ F)\big) + g. \tag{2.51}$$

It follows from the uniqueness of the solution of (2.51) that $u = v \circ F$. In resume we have obtained Theorem 1.2 (we use Lemma 2.4 to control the constants).

Similarly we obtain the following theorems in situations where $g \in L^p(\Omega_T)$, with $p > 2$.

THEOREM 2.10. *Assume that Ω is convex and Condition 1.1 is satisfied; then there exists a number $p_0 > 2$ depending on n, Ω, ϵ such that, for $p \in (2, p_0)$, $u \circ F^{-1} \in L^p(0, T, W_D^{2,p}(\Omega))$, $\partial_t(u \circ F^{-1}) \in L^p(\Omega_T)$, and*

$$\|u \circ F^{-1}\|_{L^p(0,T,W_D^{2,p}(\Omega))} + \|\partial_t(u \circ F^{-1})\|_{L^p(\Omega_T)} \leq C\|g\|_{L^p(\Omega_T)}, \tag{2.52}$$

where C depends on Ω, n, δ, and ϵ.

THEOREM 2.11. *Assume that Ω is convex and Condition 1.1 is satisfied; then there exists a number $\alpha_0 > 2$ depending on n, Ω, ϵ, p such that, for $\alpha \in (0, \alpha_0)$, $\nabla(u \circ F^{-1}) \in L^2(0, T, C^\alpha(\Omega))$ and*

$$\|\nabla(u \circ F^{-1})\|_{L^2(0,T,C^\alpha(\Omega))} \leq C\|g\|_{L^p(\Omega_T)}, \tag{2.53}$$

where C depends on Ω, δ, n, and ϵ.

The proof of Theorem 2.10 is similar and based on Theorem 2.9. The proof of Theorem 2.11 follows from Theorem 2.10 and Morrey's inequality.

2.2. Convergence of the numerical homogenization. Write $\mathcal{R}_h$, the projection operator mapping $L^2\big(0, T; H_0^1(\Omega)\big)$ onto Y_T^h defined by: For all $v \in Y_T^h$

$$\mathcal{A}_T[v, u - \mathcal{R}_h u] = 0. \tag{2.54}$$

Write $\rho := u - \mathcal{R}_h u$ and $\theta := \mathcal{R}_h u - u_h$. By a standard energy estimate, we have the following lemmas without proof.

LEMMA 2.5.

$$\frac{1}{2}\big\|(u - u_h)(T)\big\|^2_{L^2(\Omega)} + \mathcal{A}_T[u - u_h] = \int_{\Omega_T} \rho\partial_t(u - u_h) + \mathcal{A}_T[\rho, u - u_h]. \tag{2.55}$$

Proof. Subtracting (1.1) (integrated against ψ) and (1.27) we obtain that

$$\big(\psi, \partial_t(u - u_h)\big) + a[\psi, u - u_h] = 0 \quad \text{for all } \psi \in V_h(t). \tag{2.56}$$

Integrating by parts with respect to time we deduce that

$$\big(\psi, (u - u_h)(., t)\big) + a[\psi, u - u_h] = \int_{\Omega_t} \partial_t\psi(u - u_h). \tag{2.57}$$

Taking $\psi = \theta$ in (2.57) we deduce that

$$\begin{aligned}\big\|(u - u_h)(., t)\big\|^2_{L^2(\Omega)} + \mathcal{A}_t[u - u_h] = \int_{\Omega_t} \partial_t\theta(u - u_h) + \big(\rho, (u - u_h)(., t)\big) \\ + \mathcal{A}_t[\rho, u - u_h].\end{aligned} \tag{2.58}$$

Observing that

$$\begin{aligned}\int_0^t \big(\partial_t\theta, u - u_h\big) + \big(\rho, (u - u_h)(., t)\big) =& \frac{1}{2}\big\|(u - u_h)(., t)\big\|^2_{L^2(\Omega)} \\ &+ \int_0^t (\rho, \partial_t(u - u_h))\end{aligned} \tag{2.59}$$

we deduce the lemma. ∎

2.2.1. Time-independent medium: Proof of Theorem 1.3.

LEMMA 2.6.

$$\begin{aligned}\big\|(u-u_h)(T)\big\|^2_{L^2(\Omega)} + \mathcal{A}_T[u-u_h] \le 2\Big(&\|\rho\|_{L^2(\Omega_T)}\|\partial_t u - \partial_t u_h\|_{L^2(\Omega_T)} \\ &+ \mathcal{A}_T[\rho]\Big).\end{aligned} \tag{2.60}$$

Proof. Lemma 2.6 is a straightforward consequence of Lemma 2.5 and Cauchy–Schwartz and Minkowski inequalities. □

Similar to Lemmas 2.1 and 2.2, we have the following energy estimate for u_h.

LEMMA 2.7. *Assume $\partial_t a \equiv 0$, and we have*

$$\big\|u_h(.,T)\big\|^2_{L^2(\Omega)} + \mathcal{A}_T[u_h] \le \frac{C_{n,\Omega}}{\lambda_{\min}(a)}\|g\|^2_{L^2(\Omega_T)}. \tag{2.61}$$

LEMMA 2.8. *Assume $\partial_t a \equiv 0$, and we have*

$$\big\|\partial_t u_h\big\|^2_{L^2(\Omega_T)} + a\big[u_h(.,T)\big] \le \big\|g\big\|^2_{L^2(\Omega_T)}. \tag{2.62}$$

Let $t \in [0,T]$ and $v \in H^1_0(\Omega)$, and we will write $\mathcal{R}_{h,t}v(.,t)$, the solution of

$$\int_\Omega {}^t\nabla\psi a(x,t)(\psi, v - \mathcal{R}_{h,t}v)\,dx = 0 \quad \text{for all } \psi \in V_h(t). \tag{2.63}$$

We will need the following lemma.

LEMMA 2.9. *Assume the mapping $x \to F(x,t)$ to be invertible, and then for $v \in H^1_0(\Omega)$ we have*

- *for $n = 1$,*

$$\big(a[v - \mathcal{R}_{h,t}v]\big)^{\frac{1}{2}} \le C_X h\|v \circ F^{-1}(.,t)\|_{W^{2,2}_D}\|a\nabla F\|^{\frac{1}{2}}_{L^\infty(\Omega_T)}; \tag{2.64}$$

- *for $n \ge 2$,*

$$\begin{aligned}\big(a[v - \mathcal{R}_{h,t}v]\big)^{\frac{1}{2}} \le &C_X h\|v \circ F^{-1}(.,t)\|_{W^{2,2}_D} \\ &\times C_n\mu_\sigma^{\frac{n-1}{4}}\big\|(\mathrm{Trace}[\sigma])^{-1}\big\|^{\frac{n-2}{4}}_{L^\infty(\Omega_T)}.\end{aligned} \tag{2.65}$$

Remark 2.8. Recall that μ_σ is given by (1.12), and it is easy to check that μ_σ is bounded by an increasing function of $(1-\beta_\sigma)^{-1}$.

Proof. Using the change of coordinates $y = F(x,t)$ we obtain that (we write $\hat{v} := v \circ F^{-1}$)

$$a[v] = Q[\hat{v}], \tag{2.66}$$

with

$$\mathcal{Q}[w] := \int_\Omega {}^t\nabla w(y,t)Q(y,t)\nabla w(y,t)\,dy \tag{2.67}$$

and

$$Q(y,t) := \frac{\sigma}{\det(\nabla F)} \circ F^{-1}. \tag{2.68}$$

Using the definition of $\mathcal{R}_{h,t}v$ we obtain that

$$\mathcal{Q}[\hat{v} - \widehat{\mathcal{R}_{h,t}v}] = \inf_{\varphi \in X^h} \mathcal{Q}[\hat{v} - \varphi]. \tag{2.69}$$

Using property (1.21) we obtain that

$$\mathcal{Q}[\hat{v} - \widehat{\mathcal{R}_{h,t}v}] \leq \lambda_{\max}(Q) C_X^2 h^2 \|\hat{v}\|^2_{W_D^{2,2}(T)}. \tag{2.70}$$

It is easy to obtain that

- for $n = 1$,

$$\lambda_{\max}(Q) \leq \|a \nabla F\|_{L^\infty(\Omega_T)}; \tag{2.71}$$

- for $n \geq 2$,

$$\lambda_{\max}(Q) \leq C_n \mu_\sigma^{\frac{n-1}{2}} \big\| (\mathrm{Trace}[\sigma])^{-1} \big\|^{\frac{n}{2}-1}_{L^\infty(\Omega_T)}. \tag{2.72}$$

□

LEMMA 2.10. *Assume that $\partial_t a \equiv 0$, Ω is convex, $\beta_\sigma < 1$, and $(\mathrm{Trace}[\sigma])^{-1} \in L^\infty(\Omega_T)$, and then*

$$\mathcal{A}_T[\rho] \leq C h^2 \|g\|^2_{L^2(\Omega_T)}. \tag{2.73}$$

Remark 2.9. The constant C depends on C_X, n, Ω, $\lambda_{\min}(a)$, and $\|(\mathrm{Trace}[\sigma])^{-1}\|_{L^\infty(\Omega_T)}$. If $n \geq 5$ it also depends on $\|\mathrm{Trace}[\sigma]\|_{L^\infty(\Omega_T)}$, and if $n = 1$ it also depends on $\lambda_{\max}(a)$.

Proof. The proof is a straightforward application of Lemma 2.9 and Theorem 1.1. Observe that in dimension one $a \nabla F = (\int_\Omega a^{-1})^{-1}$. □

LEMMA 2.11. *Assume that $\partial_t a \equiv 0$, Ω is convex, $\beta_\sigma < 1$, and $(\mathrm{Trace}[\sigma])^{-1} \in L^\infty(\Omega_T)$, and then*

$$\|\rho\|_{L^2(\Omega_T)} \leq C h^2 \|g\|_{L^2(\Omega_T)}. \tag{2.74}$$

Remark 2.10. The constant C depends on C_X, n, Ω, $\lambda_{\min}(a)$, and $\|(\mathrm{Trace}[\sigma])^{-1}\|_{L^\infty(\Omega_T)}$. If $n \geq 5$ it also depends on $\|\mathrm{Trace}[\sigma]\|_{L^\infty(\Omega_T)}$, and if $n = 1$ it also depends on $\lambda_{\max}(a)$.

Proof. The proof follows from standard duality techniques (see, for instance, Theorem 5.7.6 of [18]). We choose $v \in L^2(0, T, H^1_0(\Omega))$ to be the solution of the following linear problem: For all $w \in L^2(0, T, H^1_0(\Omega))$

$$A_T[w, v] = (w, \rho)_{L^2(\Omega_T)}. \tag{2.75}$$

Choosing $w = \rho$ in (2.75) we deduce that

$$\|\rho\|^2_{L^2(\Omega_T))} = \mathcal{A}_T[\rho, v - \mathcal{R}_{h,t}v]. \tag{2.76}$$

Using the Cauchy–Schwartz inequality we deduce that

$$\|\rho\|^2_{L^2(\Omega_T)} \leq \big(\mathcal{A}_T[\rho]\big)^{\frac{1}{2}} \big(\mathcal{A}_T[v - \mathcal{R}_{h,t}v]\big)^{\frac{1}{2}}. \tag{2.77}$$

Using Theorem 1.1 we obtain that

$$\|\hat{v}\|_{L^2(0,T,W_D^{2,2}(\Omega))} \leq C \|\rho\|_{L^2(\Omega_T)}. \tag{2.78}$$

Using Lemma 2.9 we obtain that

$$\big(\mathcal{A}_T[v-\mathcal{R}_{h,t}v]\big)^{\frac12}\le Ch\|\rho\|_{L^2(\Omega_T)}. \tag{2.79}$$

It follows that

$$\|\rho\|_{L^2(\Omega_T)}\le Ch\big(\mathcal{A}_T[\rho]\big)^{\frac12}. \tag{2.80}$$

We deduce the lemma by applying Lemma 2.10 to bound $A_T[\rho]$. ∎

The proof of Theorem 1.3 is a straightforward application of Lemmas 2.2, 2.6, 2.8, 2.10, and 2.11.

2.2.2. Medium with a continuum of time scales: Proof of Theorem 1.4. In this subsection we will assume that the finite elements are in $H^2(\Omega)\cap H^1_0(\Omega)$ and satisfy inverse inequality (1.22).

Lemma 2.12.

$$\begin{aligned}\frac12\big\|(u-u_h)(t)\big\|^2_{L^2(\Omega)}+\mathcal{A}_t[u-u_h]=\int_{\Omega_t}\frac{\hat\rho}{|\det\nabla F|\circ F^{-1}}\\ \left(\hat g+\sum_{i,j=1}^n\sigma_{i,j}\circ F^{-1}\partial_i\partial_j\hat u_h-\partial_t\hat u_h\right).\end{aligned} \tag{2.81}$$

Proof. Consider (2.55). We have

$$\int_{\Omega_t}\rho\partial_t(u-u_h)=\int_{\Omega_t}\frac{\hat\rho}{|\det\nabla F|\circ F^{-1}}\partial_t(\hat u-\hat u_h)+\int_{\Omega_t}\rho\partial_tF(\nabla F)^{-1}\nabla(u-u_h). \tag{2.82}$$

Using (1.2) we obtain that

$$\int_{\Omega_t}\rho\partial_tF(\nabla F)^{-1}\nabla(u-u_h)=-\mathcal{A}_t[\rho,u-u_h]-\sum_{i,j=1}^n\int_{\Omega_t}\hat\rho Q_{i,j}\partial_i\partial_j(\hat u-\hat u_h). \tag{2.83}$$ ∎

Lemma 2.13.

$$\left\|\frac{\partial_t\hat u_h}{|\det(\nabla F)|^{\frac12}\circ F^{-1}}\right\|_{L^2(\Omega_T)}\le 2\|g\|_{L^2(\Omega_T)}+C\|\hat u_h\|_{L^2(0,T,W^{2,2}_D(\Omega))}, \tag{2.84}$$

where the constant C depends on n, $\lambda_{\max}(a)$, $\|\operatorname{Trace}[\sigma]\|_{L^\infty(\Omega_T)}$, and μ_σ.

Proof. Using the change of variable $y=F(x,t)$ in (1.27) we obtain that for all $\varphi\in X^h$

$$\begin{cases}(\varphi,\frac{\partial_t\hat u_h}{|\det(\nabla F)|\circ F^{-1}})_{L^2(\Omega)}= & \sum_{i,j=1}^n\int_\Omega(\varphi,Q_{i,j}\partial_i\partial_j\hat u_h)_{L^2(\Omega)}\\ & +(\varphi,\frac{\hat g}{|\det(\nabla F)|\circ F^{-1}})_{L^2(\Omega)},\\ \hat u_h(x,0)=0.\end{cases} \tag{2.85}$$

Recall that Q is given by (2.68). We choose $\varphi=\partial_t\hat u$ and observe that

$$\frac{\sigma}{|\det\nabla F|^{\frac12}}=\frac{\sigma}{|\det\sigma|^{\frac14}}|\det a|^{\frac14}. \tag{2.86}$$

Thus

$$\left\|\frac{\sigma}{|\det\nabla F|^{\frac12}}\right\|\le C\big(n,\lambda_{\max}(a),\|\operatorname{Trace}[\sigma]\|_{L^\infty(\Omega_T)},\mu_\sigma\big). \tag{2.87}$$

We deduce the lemma by the Minkowski inequality. ∎

Combining Lemmas 2.12 and 2.13 we obtain the following lemma.

LEMMA 2.14.

$$\frac{1}{2}\big\|(u-u_h)(T)\big\|^2_{L^2(\Omega)} + \mathcal{A}_T[u-u_h] \leq \|\rho\|_{L^2(\Omega_T)}\Big(\|g\|_{L^2(\Omega_T)} + C\|\hat{u}_h\|_{L^2(0,T,W_D^{2,2}(\Omega))}\Big), \tag{2.88}$$

where the constant C depends on n, $\lambda_{\max}(a)$, $\|\operatorname{Trace}[\sigma]\|_{L^\infty(\Omega_T)}$, and μ_σ.

LEMMA 2.15. *Assume that Ω is convex and Condition 1.1 is satisfied; then*

$$\|\rho\|_{L^2(\Omega_T)} \leq Ch^2\|g\|_{L^2(\Omega_T)}. \tag{2.89}$$

Remark 2.11. The constant C depends on C_X, n, Ω, δ, ϵ, $\lambda_{\min}(a)$, and $\lambda_{\max}(a)$.

Proof. The proof is similar to the proof of Lemma 2.11. As in the proof of Lemma 2.11 we choose $v \in L^2(0,T,H_0^1(\Omega))$ to be the solution of the following linear problem: For all $w \in L^2(0,T,H_0^1(\Omega))$

$$A_T[w,v] = (w,\rho)_{L^2(\Omega_T)}. \tag{2.90}$$

Choosing $w = \rho$ in (2.90) we deduce that

$$\|\rho\|^2_{L^2(\Omega_T))} = \mathcal{A}_T[\rho, v - \mathcal{R}_{h,t}v]. \tag{2.91}$$

Using the Cauchy–Schwartz inequality we deduce that

$$\|\rho\|^2_{L^2(\Omega_T)} \leq \big(\mathcal{A}_T[\rho]\big)^{\frac{1}{2}}\big(\mathcal{A}_T[v-\mathcal{R}_{h,t}v]\big)^{\frac{1}{2}}. \tag{2.92}$$

Using Theorem 1.2 we obtain that

$$\|\hat{v}\|_{L^2(0,T,W_D^{2,2}(\Omega))} \leq C\|\rho\|_{L^2(\Omega_T)}. \tag{2.93}$$

Using Lemma 2.9 we obtain that

$$\big(\mathcal{A}_T[v-\mathcal{R}_{h,t}v]\big)^{\frac{1}{2}} \leq Ch\|\rho\|_{L^2(\Omega_T)}. \tag{2.94}$$

It follows that

$$\|\rho\|_{L^2(\Omega)} \leq Ch\big(\mathcal{A}_T[\rho]\big)^{\frac{1}{2}}. \tag{2.95}$$

We deduce the lemma by applying Lemma 2.9 and Theorem 1.2 to bound $A_T[\rho]$. ∎

LEMMA 2.16. *Assume that Ω is convex and that* $\operatorname{Trace}[\sigma] \in L^\infty(\Omega_T)$.

$$\|\hat{u}_h\|_{L^2(0,T,W_D^{2,2}(\Omega))} \leq \frac{C}{h}\|g\|_{L^2(\Omega_T)}. \tag{2.96}$$

Remark 2.12. The constant C depends on C_X, n, Ω, $\lambda_{\min}(a)$, $\lambda_{\max}(a)$, and $\|\operatorname{Trace}[\sigma]\|_{L^\infty(\Omega_T)}$.

Proof. Using the inverse inequality (1.22) of the finite elements we obtain that

$$\|\hat{u}_h\|_{L^2(0,T,W_D^{2,2}(\Omega))} \leq \frac{C_X}{h}\|\nabla\hat{u}_h\|_{L^2(0,T,W_D^{2,2}(\Omega))}. \tag{2.97}$$

Using the change of variables $y = F(x)$ we obtain that

$$\|\nabla\hat{u}_h\|^2_{L^2(0,T,W_D^{2,2}(\Omega))} \leq C\mathcal{A}_T[u_h], \tag{2.98}$$

where C depends on n, $\lambda_{\min}(a)$, and $\|\operatorname{Trace}[\sigma]\|_{L^\infty(\Omega_T)}$. We deduce the lemma by using Lemma 2.7. ∎

The proof of Theorem 1.4 is a straightforward application of Lemmas 2.14, 2.15, and 2.16.

2.2.3. Numerical homogenization in time and space: Proof of Theorems 1.5–1.7. We use the notation of subsection 1.3. First we will prove Theorem 1.5; let us observe that the numerical scheme associated to (1.34) is stable. Indeed choosing $\psi = v_{n+1}$ one gets

$$\begin{aligned}\big|v_{n+1}(t_{n+1})\big|^2_{L^2(\Omega)} &= \big(v_{n+1}(t_n), v_n(t_n)\big)_{L^2(\Omega)} \\ &\quad + \frac{1}{2}\big(\big|v_{n+1}(t_{n+1})\big|^2_{L^2(\Omega)} - \big|v_{n+1}(t_n)\big|^2_{L^2(\Omega)}\big) \\ &\quad - \int_{t_n}^{t_{n+1}} \Big(a\big[v_{n+1}(t)\big] + \big(v_{n+1}(t), g(t)\big)_{L^2(\Omega)}\Big)\, dt.\end{aligned} \tag{2.99}$$

It follows by the Cauchy–Schwartz inequality that

$$\begin{aligned}\frac{1}{2}\big|v_{n+1}(t_{n+1})\big|^2_{L^2(\Omega)} \le &\frac{1}{2}\big|v_n(t_n)\big|^2_{L^2(\Omega)} \\ &- \int_{t_n}^{t_{n+1}} \Big(a\big[v_{n+1}(t)\big] + \big(v_{n+1}(t), g(t)\big)_{L^2(\Omega)}\Big)\, dt.\end{aligned} \tag{2.100}$$

Hence using Poincaré and Minkowski inequalities one obtains that

$$\begin{aligned}\big|v_{n+1}(t_{n+1})\big|^2_{L^2(\Omega)} + \int_{t_n}^{t_{n+1}} a\big[v_{n+1}(t)\big]\, dt \le &\big|v_n(t_n)\big|^2_{L^2(\Omega)} \\ &+ \frac{C_{n,\Omega}}{\lambda_{\min}(a)} \int_{t_n}^{t_{n+1}} \big|g(t)\big|^2_{L^2(\Omega)}\, dt,\end{aligned} \tag{2.101}$$

which implies Theorem 1.5 and the stability of the scheme.

Integrating (1.27) with respect to time we obtain that, for $\psi \in U_T^h$,

$$\begin{aligned}\big(\psi(t_{n+1}), u_h(t_{n+1})\big)_{L^2(\Omega)} = &\big(\psi(t_n), u_h(t_n)\big)_{L^2(\Omega)} \\ &+ \int_{t_n}^{t_{n+1}} \Big(\big(\partial_t\psi(t), u_h(t)\big)_{L^2(\Omega)} \\ &- a\big[\psi(t), u_h(t)\big] + \big(\psi(t), g(t)\big)_{L^2(\Omega)}\Big)\, dt.\end{aligned} \tag{2.102}$$

Let us write (u^i) the coordinates of u_h associated to the basis $(\varphi_i \circ F)$, i.e.,

$$u_h(x,t) := \sum_i u^i(t)\varphi_i(F(x,t)), \tag{2.103}$$

and define

$$u_n(x,t) := \sum_i u^i(t_n)\varphi_i(F(x,t)). \tag{2.104}$$

Subtracting (2.102) and (1.34) we obtain that, for $\psi \in Z_T$,

$$\begin{aligned}\big(\psi(t_{n+1}), (u_{n+1} - v_{n+1})(t_{n+1})\big)_{L^2(\Omega)} = &\big(\psi(t_n), (u_n - v_n)(t_n)\big)_{L^2(\Omega)} \\ &+ \int_{t_n}^{t_{n+1}} \Big(\big(\partial_t\psi(t), (u_h - v_{n+1})(t)\big)_{L^2(\Omega)} \\ &- a\big[\psi(t), (u_h - v_{n+1})(t)\big]\Big)\, dt.\end{aligned} \tag{2.105}$$

Choosing $\psi = u_{n+1} - v_{n+1}$ we deduce using the Cauchy–Schwartz inequality that

$$\begin{aligned}
&\frac{1}{2}\big|(u_{n+1}-v_{n+1})(t_{n+1})\big|^2_{L^2(\Omega)} + \int_{t_n}^{t_{n+1}} a\big[(u_{n+1}-v_{n+1})(t)\big]\,dt \\
&\quad\le \frac{1}{2}\big|(u_n-v_n)(t_n)\big|^2_{L^2(\Omega)} + \int_{t_n}^{t_{n+1}} \Big(\big(\partial_t(u_{n+1}-v_{n+1})(t),(u_h-u_{n+1})(t)\big)_{L^2(\Omega)} \\
&\quad - a\big[(u_{n+1}-v_{n+1})(t),(u_h-u_{n+1})(t)\big]\Big)\,dt.
\end{aligned} \tag{2.106}$$

Time-independent medium. Observe that if the medium is time-independent, then (2.106) can be written

$$\begin{aligned}
&\frac{1}{2}\big|(u_{n+1}-v_{n+1})(t_{n+1})\big|^2_{L^2(\Omega)} + \int_{t_n}^{t_{n+1}} a\big[(u_{n+1}-v_{n+1})(t)\big]\,dt \\
&\quad\le \frac{1}{2}\big|(u_n-v_n)(t_n)\big|^2_{L^2(\Omega)} - \int_{t_n}^{t_{n+1}} a\big[(u_{n+1}-v_{n+1})(t),(u_h-u_{n+1})(t)\big]\,dt,
\end{aligned} \tag{2.107}$$

which leads us to

$$\begin{aligned}
&\frac{1}{2}\big|(u_{n+1}-v_{n+1})(t_{n+1})\big|^2_{L^2(\Omega)} + \int_{t_n}^{t_{n+1}} a\big[(u_{n+1}-v_{n+1})(t)\big]\,dt \\
&\quad\le \frac{1}{2}\big|(u_n-v_n)(t_n)\big|^2_{L^2(\Omega)} \\
&\quad + \int_{t_n}^{t_{n+1}}\int_{t_n}^{t_{n+1}} 1(t<s)\,a\big[(u_{n+1}-v_{n+1})(t),\partial_s u_h(s)\big]\,ds\,dt.
\end{aligned} \tag{2.108}$$

Write $\Delta t := t_{n+1} - t_n$. Using the Minkowski inequality we obtain that

$$\begin{aligned}
a\big[(u_{n+1}-v_{n+1})(t),\partial_s u_h(s)\big] \le \frac{1}{2\Delta t} a\big[(u_{n+1}-v_{n+1})(t)\big] \\
+ \frac{1}{2}\Delta t a\big[\partial_s u_h(s)\big].
\end{aligned} \tag{2.109}$$

It follows from (2.108) that

$$\begin{aligned}
&\big|(u_{n+1}-v_{n+1})(t_{n+1})\big|^2_{L^2(\Omega)} + \int_{t_n}^{t_{n+1}} a\big[(u_{n+1}-v_{n+1})(t)\big]\,dt \\
&\le \big|(u_n-v_n)(t_n)\big|^2_{L^2(\Omega)} + |\Delta t|^2 \int_{t_n}^{t_{n+1}} a\big[\partial_s u_h(s)\big]\,ds.
\end{aligned} \tag{2.110}$$

Observing that

$$\begin{aligned}
\int_{t_n}^{t_{n+1}} a\big[(u_{n+1}-v_{n+1})(t)\big]\,dt \ge 0.5\int_{t_n}^{t_{n+1}} a\big[(u_h-v_{n+1})(t)\big]\,dt \\
- \int_{t_n}^{t_{n+1}} a\big[(u_h-u_{n+1})(t)\big]\,dt
\end{aligned} \tag{2.111}$$

and

$$\int_{t_n}^{t_{n+1}} a\big[(u_h - u_{n+1})(t)\big]\,dt \le |\Delta t|^2 \int_{t_n}^{t_{n+1}} a\big[\partial_s u_h(s)\big]\,ds, \tag{2.112}$$

we obtain that

$$\begin{aligned} &\big|(u_{n+1} - v_{n+1})(t_{n+1})\big|^2_{L^2(\Omega)} + 0.5 \int_{t_n}^{t_{n+1}} a\big[(u_h - v_{n+1})(t)\big]\,dt \\ &\le \big|(u_n - v_n)(t_n)\big|^2_{L^2(\Omega)} + \frac{3}{2}|\Delta t|^2 \int_{t_n}^{t_{n+1}} a\big[\partial_s u_h(s)\big]\,ds. \end{aligned} \tag{2.113}$$

In conclusion we have obtained the following lemma.

LEMMA 2.17. *Let $v \in Z_T$ be the solution of* (1.34). *We have*

$$\big\|(u_h - v)(T)\big\|^2_{L^2(\Omega)} + \int_0^T a\big[(u_h - v)(t)\big]\,dt \le 3|\Delta t|^2 \int_0^T a\big[\partial_s u_h(s)\big]\,ds. \tag{2.114}$$

Combining Lemma 2.3 with Lemma 2.17 we obtain Theorem 1.6.

Time-dependent medium. Observe that

$$\partial_t (u_{n+1} - v_{n+1}) = \partial_t F (\nabla F)^{-1} \nabla (u_{n+1} - v_{n+1}).$$

It follows after writing $\partial_t F = \operatorname{div} a \nabla F$, integration by parts, and using the change of variables $y = F(x,t)$ in (2.106) that

$$\begin{aligned} &\frac{1}{2}\big|(u_{n+1} - v_{n+1})(t_{n+1})\big|^2_{L^2(\Omega)} + \int_{t_n}^{t_{n+1}} a\big[(u_{n+1} - v_{n+1})(t)\big]\,dt \\ &\le \frac{1}{2}\big|(u_n - v_n)(t_n)\big|^2_{L^2(\Omega)} - 2\int_{t_n}^{t_{n+1}} a\big[(u_{n+1} - v_{n+1})(t), (u_h - u_{n+1})(t)\big]\,dt \\ &- \sum_{i,j} \int_{t_n}^{t_{n+1}} \int_\Omega (\hat{u}_h - \hat{u}_{n+1}) Q_{i,j} \partial_i \partial_j (\hat{u}_{n+1} - \hat{v}_{n+1})\,dt\,dy. \end{aligned} \tag{2.115}$$

Hence using the Minkowski inequality we obtain that

$$\begin{aligned} &\big|(u_{n+1} - v_{n+1})(t_{n+1})\big|^2_{L^2(\Omega)} + \int_{t_n}^{t_{n+1}} a\big[(u_{n+1} - v_{n+1})(t)\big]\,dt \\ &\le \big|(u_n - v_n)(t_n)\big|^2_{L^2(\Omega)} + 4\int_{t_n}^{t_{n+1}} a\big[(u_h - u_{n+1})(t)\big]\,dt \\ &- 2\sum_{i,j} \int_{t_n}^{t_{n+1}} \int_\Omega (\hat{u}_h - \hat{u}_{n+1}) Q_{i,j} \partial_i \partial_j (\hat{u}_{n+1} - \hat{v}_{n+1})\,dt\,dy. \end{aligned} \tag{2.116}$$

Using the Minkowski inequality we obtain that

$$\begin{aligned}\Big|\sum_{i,j}\int_{t_n}^{t_{n+1}}\int_\Omega(\hat u_h-\hat u_{n+1})Q_{i,j}\partial_i\partial_j(\hat u_{n+1}-\hat v_{n+1})\,dt\,dy\Big|&\\ \le C_An^2\int_{t_n}^{t_{n+1}}\int_\Omega|\hat u_h-\hat u_{n+1}|^2\,dt\,dy&\\ +\frac{\lambda_{\max}(Q)}{C_A}\int_{t_n}^{t_{n+1}}\sum_{i,j}\int_\Omega|\partial_i\partial_j(\hat u_{n+1}-\hat v_{n+1})|^2\,dt\,dy.&\end{aligned}\tag{2.117}$$

Using the inverse inequality (1.22) and the change of variable $y=F(x)$ we obtain that

$$\begin{aligned}\int_{t_n}^{t_{n+1}}\sum_{i,j}\int_\Omega|\partial_i\partial_j(\hat u_{n+1}-\hat v_{n+1})|^2\,dt\,dy\le\frac{C_X}{h^2\lambda_{\min}(Q)}&\\ \int_{t_n}^{t_{n+1}}a\big[(u_{n+1}-v_{n+1})(t)\big]\,dt.&\end{aligned}\tag{2.118}$$

In resume, choosing $C_A=\frac{4C_X\lambda_{\max}(Q)}{h^2\lambda_{\min}(Q)}$ we have obtained that

$$\begin{aligned}\big|(u_{n+1}-v_{n+1})(t_{n+1})\big|^2_{L^2(\Omega)}+0.5\int_{t_n}^{t_{n+1}}a\big[(u_{n+1}-v_{n+1})(t)\big]\,dt&\\ \le\big|(u_n-v_n)(t_n)\big|^2_{L^2(\Omega)}+8\int_{t_n}^{t_{n+1}}a\big[(u_h-u_{n+1})(t)\big]\,dt&\\ +\frac{8C_X\lambda_{\max}(Q)}{h^2\lambda_{\min}(Q)}n^2\int_{t_n}^{t_{n+1}}\int_\Omega|\hat u_h-\hat u_{n+1}|^2\,dt\,dy.&\end{aligned}\tag{2.119}$$

A computation similar to the one leading to (2.109) gives us

$$\begin{aligned}\big|(u_{n+1}-v_{n+1})(t_{n+1})\big|^2_{L^2(\Omega)}+\frac14\int_{t_n}^{t_{n+1}}a\big[(u_h-v_{n+1})(t)\big]\,dt&\\ \le\big|(u_n-v_n)(t_n)\big|^2_{L^2(\Omega)}+9\int_{t_n}^{t_{n+1}}a\big[(u_h-u_{n+1})(t)\big]\,dt&\\ +\frac{8C_X\lambda_{\max}(Q)}{h^2\lambda_{\min}(Q)}n^2\int_{t_n}^{t_{n+1}}\int_\Omega|\hat u_h-\hat u_{n+1}|^2\,dt\,dy.&\end{aligned}\tag{2.120}$$

Moreover using the change of variables $F(x)=y$ and the inverse inequality (1.23) we obtain that

$$\int_{t_n}^{t_{n+1}}a\big[(u_h-u_{n+1})(t)\big]\,dt\le\frac{C_X\lambda_{\max}(Q)}{h^2}\int_{t_n}^{t_{n+1}}\int_\Omega|\hat u_h-\hat u_{n+1}|^2\,dt\,dy.\tag{2.121}$$

Let us also observe that

$$\int_{t_n}^{t_{n+1}}\int_\Omega|\hat u_h-\hat u_{n+1}|^2\,dt\,dy\le|\Delta t|^2\int_{t_n}^{t_{n+1}}\int_\Omega|\partial_t\hat u_h|^2\,dt\,dy.\tag{2.122}$$

It follows that

$$\begin{aligned}&\big|(u_{n+1}-v_{n+1})(t_{n+1})\big|^2_{L^2(\Omega)}+\frac{1}{4}\int_{t_n}^{t_{n+1}} a\big[(u_h-v_{n+1})(t)\big]\,dt\\&\le\big|(u_n-v_n)(t_n)\big|^2_{L^2(\Omega)}+C_B\frac{|\Delta t|^2}{h^2}\int_{t_n}^{t_{n+1}}\int_\Omega|\partial_t\hat{u}_h|^2\,dt\,dy,\end{aligned}\tag{2.123}$$

with

$$C_B=C_nC_X\lambda_{\max}(Q)\left(1+\frac{1}{\lambda_{\min}(Q)}\right).\tag{2.124}$$

We deduce that

$$\big\|(u_h-v)(T)\big\|^2_{L^2(\Omega)}+\frac{1}{4}\int_0^T a\big[(u_h-v)(t)\big]\,dt\le C_B\frac{|\Delta t|^2}{h^2}\int_{\Omega_T}|\partial_t\hat{u}_h|^2\,dt\,dy.\tag{2.125}$$

Using Lemma 2.4 to control C_B and combining (2.125) with Theorem 1.2 we obtain Theorem 1.7.

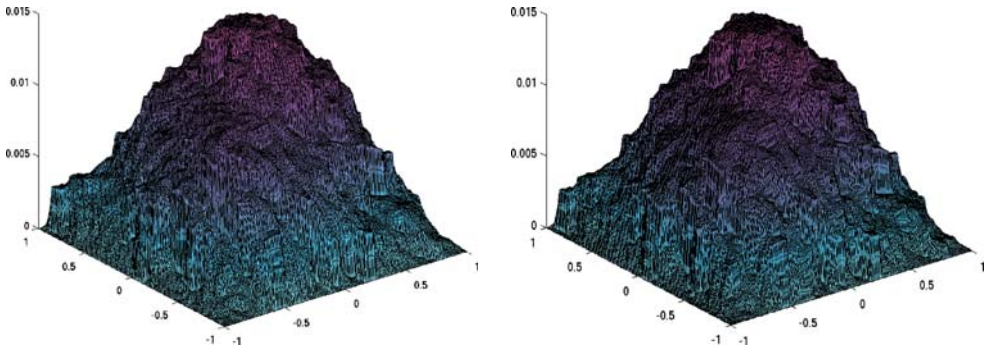
3. Numerical experiments. The purpose of this section is to give several illustrations of the implementation and performance of the numerical method. The computational domain is the unit square in dimension two. Equation (1.1) is solved on a fine tessellation characterized by 16129 interior nodes (degrees of freedom).

Three different coarse tessellations with 9, 49, and 225 degrees of freedoms are considered.

The parabolic operator associated to (1.1) has been homogenized onto these coarse meshes using the method presented in subsection 1.3. We have chosen splines to be the space X^h in subsection 1.2.

3.1. Time-independent examples.

Example 3.1 (time-independent site percolation). In this example we consider the site-percolating medium. a is set to be equal to 1 or 100 with probability 1/2 on each triangle of a fine mesh. (1.1) has been solved with $g=1$ and $g=\sin(2.4x-1.8y+2\pi t)$. Figure 3.1 shows u computed on 16641 interior nodes and u_h computed on 9 interior nodes in the case $g=1$ at time 1. The fine mesh and coarse mesh errors are given in Table 3.1.



(a) u. (b) u_h.

FIG. 3.1. *u computed on* 16641 *interior nodes and u_h computed on* 9 *interior nodes.*

TABLE 3.1

Experiment	dof	L^1	L^∞	L^2	H^1
Coarse mesh error.	9	0.0142	0.0389	0.0168	0.0366
Time-independent percolation	49	0.0077	0.0450	0.0101	0.0482
with $g = 1$.	225	0.0035	0.0228	0.0060	0.0293
Fine mesh error.	9	0.0196	0.0843	0.0251	0.1193
Time-independent percolation	49	0.0136	0.0698	0.0184	0.1028
with $g = 1$.	225	0.0040	0.0243	0.0070	0.0485
Coarse mesh error.	9	0.0236	0.0569	0.0262	0.0477
Time-independent percolation	49	0.0181	0.0571	0.0215	0.0558
with $g = \sin(2.4x - 1.8y + 2\pi t)$.	225	0.0119	0.0774	0.0167	0.0939
Fine mesh error.	9	0.0424	0.1099	0.0512	0.1712
Time-independent percolation	49	0.0277	0.0985	0.0348	0.1451
with $g = \sin(2.4x - 1.8y + 2\pi t)$.	225	0.0174	0.0886	0.0242	0.1192
Coarse mesh error. Multiscale	9	0.0018	0.0045	0.0019	0.0039
trigonometric time-dependent. $g = 1$.	49	0.0012	0.0054	0.0015	0.0060
Fine mesh error. Multiscale	9	0.0031	0.0096	0.0034	0.0242
trigonometric time-dependent. $g = 1$.	49	0.0014	0.0059	0.0016	0.0166
Coarse mesh error. Multiscale trigonometric	9	0.0043	0.0087	0.0044	0.0085
time-dependent. $g = \sin(2.4x - 1.8y + 2\pi t)$.	49	0.0033	0.0079	0.0035	0.0084
Fine mesh error. Multiscale trigonometric	9	0.0082	0.0199	0.0087	0.0379
time-dependent medium. $g = \sin(2.4x - 1.8y + 2\pi t)$.	49	0.0038	0.0104	0.0040	0.0244
Coarse mesh error.	9	0.0046	0.0074	0.0052	0.0065
Time-dependent random fractal.	49	0.0036	0.0046	0.0036	0.0059
Fine mesh error.	9	0.0039	0.0082	0.0043	0.0222
Time-dependent random fractal.	49	0.0033	0.0054	0.0034	0.0168

3.2. Time-dependent examples. In the following examples we consider media characterized by a continuum of time scales.

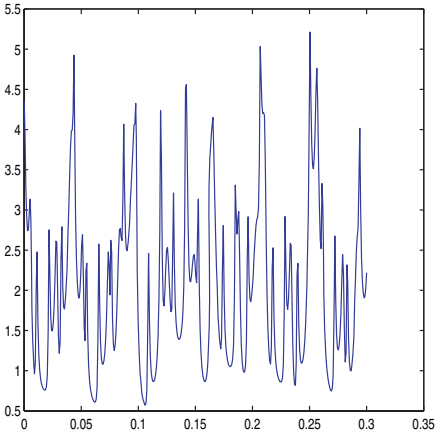
Example 3.2 (Time-dependent multiscale trigonometric). In this example a is given by (1.20). Although the number of fine time steps to solve (1.1) is 2663, only 134 coarse time steps have been used to solve the homogenized equation. Hence if one also takes into account homogenization in space, the compression factor is of the order of 35000 for the coarse mesh with 9 interior nodes.

Figure 3.2 shows the curves of $t \to a(x_0, t)$ and $t \to F(x_0, t)$ for a given $x_0 \in \Omega$ and the coarse and fine mesh relative to L^2 and H^1 errors with respect to time. The initial increase of the relative error is due to the initial value $u(x, 0) \equiv 0$.

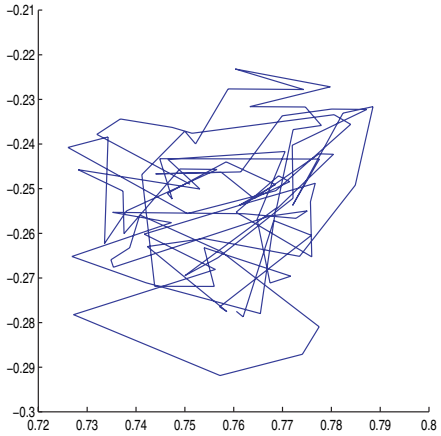
The coarse and fine meshes errors at $t = 0.1$ are given in Table 3.1 for $g = 1$. The errors at $t = 0.1$ are given for $g = \sin(2.4x - 1.8y + 2\pi t)$.

Example 3.3 (Time-dependent random fractal). In this case, a is given by a product of discontinuous functions oscillating randomly at different scales: $a(x, t) = a_1(x, t)a_2(x, t) \dots a_n(x, t)$ and $a_i(x, t) = c_{pq}$ for $x \in [\frac{p}{2^i}, \frac{p+1}{2^i}) \times [\frac{q}{2^i}, \frac{q+1}{2^i})$ in the time interval $0.1 \times [\frac{k}{4^i}, \frac{k+1}{4^i})$. c_{pq} is uniformly random in $[\frac{1}{\gamma}, \gamma]$, $n = 6$, and $\gamma = 0.7$. In this example, we have $\frac{\lambda_{\max}(a)}{\lambda_{\min}(a)} = 160.3295$. The number of fine time steps is 3482, and the number of coarse time steps is 175.

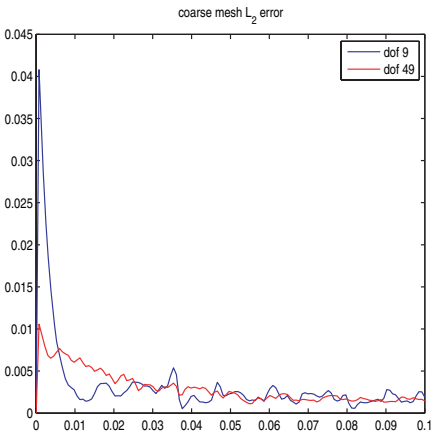
a and the map (F_1, F_2) are drawn in Figure 3.3. L^2 and H^1 errors are given in Figure 3.4. Coarse and fine mesh errors at $t = 0.1$ are given in Table 3.1. We have chosen $g = 1$ in this numerical experiment, and one obtains similar results by choosing $g = \sin(2.4x - 1.8y + 2\pi t)$.



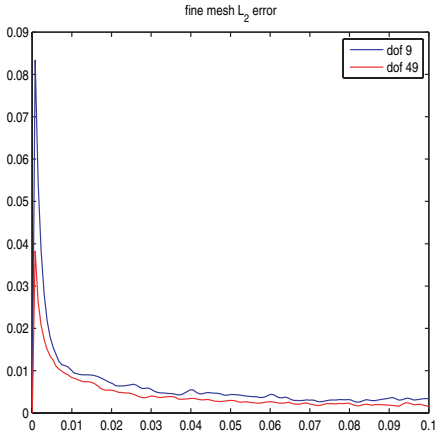
(a) $t \to a(x_0, t)$.



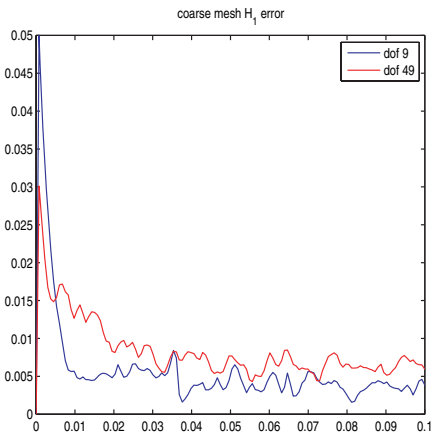
(b) Top view of $t \to F(x_0, t)$.



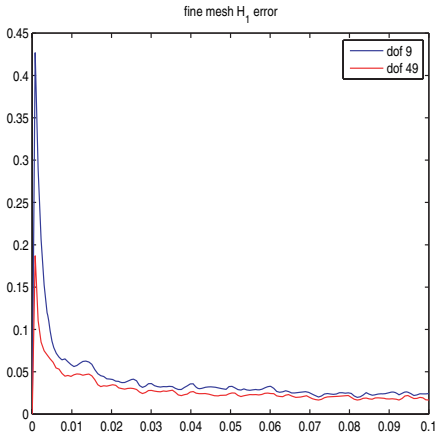
(c) Coarse mesh L^2 error.



(d) Fine Mesh L^2 error.

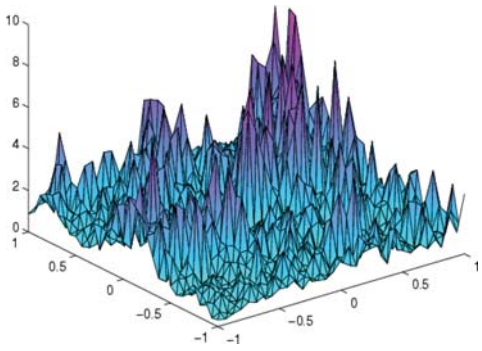


(e) Coarse Mesh H^1 error.

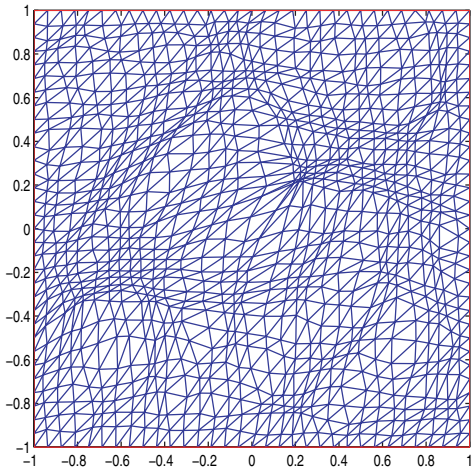


(f) Fine Mesh H^1 error.

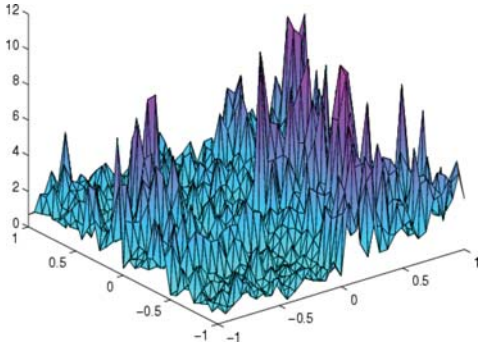
FIG. 3.2. *Multiscale time-dependent trigonometric medium.*



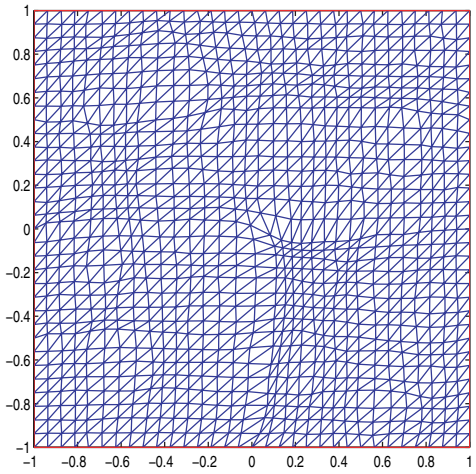
(a) a at $t = 0$.



(b) (F_1, F_2) at $t = 0$.

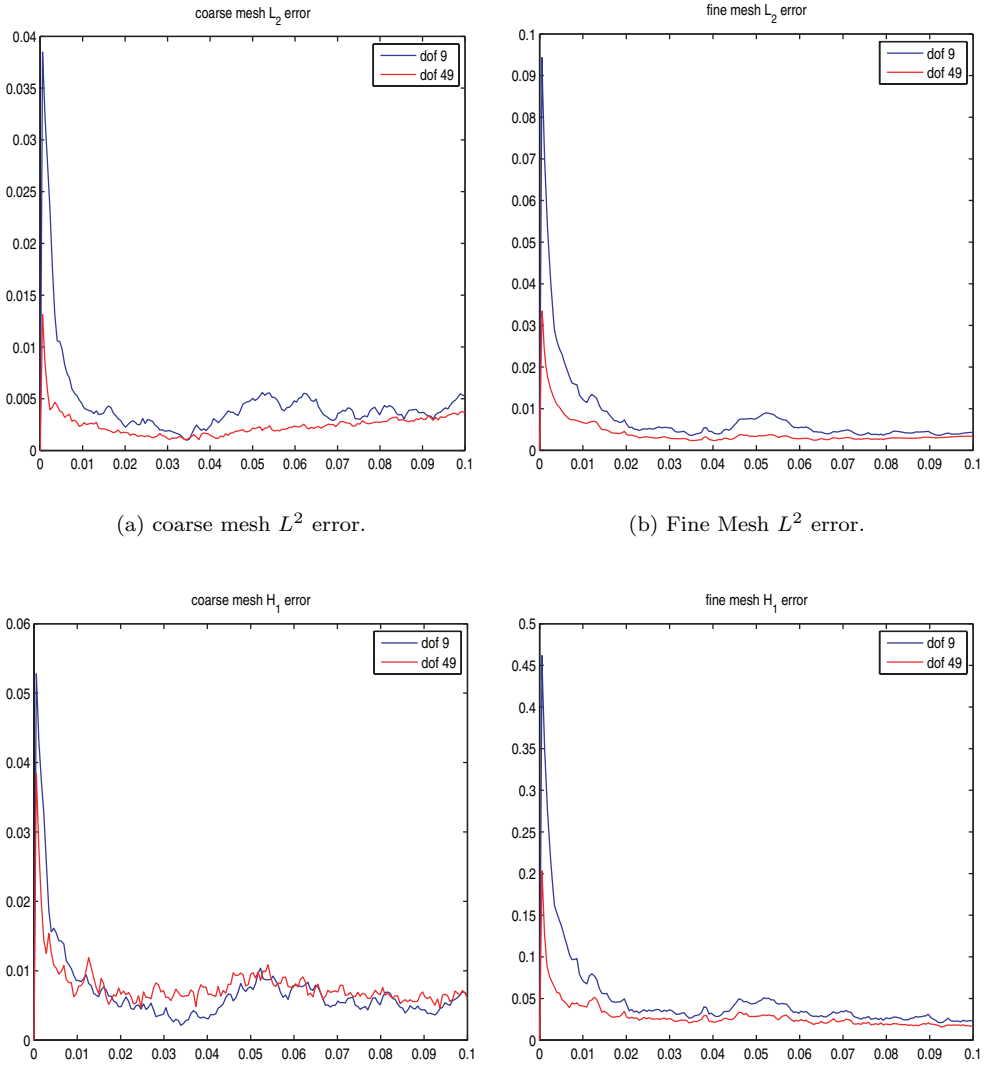


(c) a at $t = 0.1$.



(d) (F_1, F_2) at $t = 0.1$.

FIG. 3.3. *a and (F_1, F_2) at time $t = 0$, $t = 0.1$ for the time-dependent random fractal medium.*



(a) coarse mesh L^2 error.

(b) Fine Mesh L^2 error.

(c) coarse Mesh H^1 error.

(d) fine Mesh H^1 error.

FIG. 3.4. *Time-dependent random fractal medium at* $t = 0.1$.

Acknowledgment. We thank two anonymous referees for detail comments and suggestions.

REFERENCES

[1] J. E. Aarnes, *On the use of a mixed multiscale finite element method for greater flexibility and increased speed or improved accuracy in reservoir simulation*, Multiscale Model. Simul., 2 (2004), pp. 421–439.
[2] G. Alberti and S. Müller, *A new approach to variational problems with multiple scales*, Comm. Pure Appl. Math., 54 (2001), pp. 761–825.
[3] G. Alessandrini and V. Nesi, *Univalent σ-harmonic mappings: Connections with quasiconformal mappings*, J. Anal. Math., 90 (2003), pp. 197–215.
[4] G. Allaire, *Shape Optimization by the Homogenization Method*, Springer, New York, 2001.
[5] G. Allaire and R. Brizzi, *A multiscale finite element method for numerical homogenization*, Multiscale Model. Simul., 4 (2005), pp. 790–812.
[6] A. Ancona, *Some results and examples about the behavior of harmonic functions and Green's functions with respect to second order elliptic operators*, Nagoya Math. J., 165 (2002), pp. 123–158.
[7] A. Averbuch, G. Beylkin, R. Coifman, and M. Israeli, *Multiscale inversion of elliptic operators*, in Signal and Image Representation in Combined Spaces, Wavelet Anal. Appl. 7, Academic Press, San Diego, CA, 1998, pp. 341–359.
[8] I. Babuška, G. Caloz, and J. E. Osborn, *Special finite element methods for a class of second order elliptic problems with rough coefficients*, SIAM J. Numer. Anal., 31 (1994), pp. 945–981.
[9] I. Babuška and J. E. Osborn, *Generalized finite element methods: Their performance and their relation to mixed methods*, SIAM J. Numer. Anal., 20 (1983), pp. 510–536.
[10] M. Bebendorf, *Approximate inverse preconditioning of finite element discretizations of elliptic operators with nonsmooth coefficients*, SIAM J. Matrix. Anal., 27 (2006), pp. 909–929.
[11] M. Bebendorf, *Efficient inversion of Galerkin matrices of general second-order elliptic differential operators with nonsmooth coefficients*, Math. Comp., 74 (2005), pp. 1179–1199.
[12] M. Bebendorf, *Why Approximate Lu Decompositions of Finite Element Discretizations of Elliptic Operators can be Computed with Almost Linear Complexity*, preprint 8/2005, Max-Planck-Institute MiS, Leipzig, 2005.
[13] M. Bebendorf and Y. Chen, *Efficient Solution of Nonlinear Elliptic Problems Using Hierarchical Matrices with Broyden Updates*, preprint 51/2005, Max-Planck-Institut MiS, Leipzig, 2005.
[14] M. Bebendorf and W. Hackbusch, *Existence of $\mathcal{H}$-matrix approximants to the inverse FE-matrix of elliptic operators with L^∞-coefficients*, Numer. Math., 95 (2003), pp. 1–28.
[15] G. Ben Arous and H. Owhadi, *Multiscale homogenization with bounded ratios and anomalous slow diffusion*, Comm. Pure Appl. Math., 56 (2003), pp. 80–113.
[16] A. Bensoussan, J. L. Lions, and G. Papanicolaou, *Asymptotic Analysis for Periodic Structure*, North-Holland, Amsterdam, 1978.
[17] G. Beylkin and N. Coult, *A multiresolution strategy for reduction of elliptic PDEs and eigenvalue problems*, Appl. Comput. Harmon. Anal., 5 (1998), pp. 129–155.
[18] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, 2nd ed., Springer, New York, 2002.
[19] M. E. Brewster and G. Beylkin, *A multiresolution strategy for numerical homogenization*, Appl. Comput. Harmon. Anal., 2 (1995), pp. 327–349.
[20] F. Brezzi and D. Marini, *Subgrid phenomena and numerical schemes*, in Frontiers in Numerical Analysis (Durham, 2002), Universitext, Springer, Berlin, 2003, pp. 1–16.
[21] M. Briane, *Nonlocal effects in two-dimension conductivity*, Arch. Ration. Mech. Anal., 182 (2006), pp. 255–267.
[22] M. Briane, G. W. Milton, and V. Nesi, *Change of sign of the corrector's determinant for homogenization in three-dimensional conductivity*, Arch. Ration. Mech. Anal., 173 (2004), pp. 133–150.
[23] S. Campanato, *Un risultato relativo ad equazioni ellittiche del secondo ordine di tipo non variazionale*, Ann. Scuola Norm. Sup. Pisa (3), 21 (1967), pp. 701–707.
[24] S. Chen, W. E, and C.-W. Shu, *The heterogeneous multiscale method based on the discontinuous Galerkin method for hyperbolic and parabolic problems*, Multiscale Model. Simul., 3 (2005), pp. 871–894.
[25] A. Chertock and D. Levy, *On wavelet-based numerical homogenization*, Multiscale Model. Simul., 3 (2005), pp. 65–88.

[26] R. Coifman, P.-L. Lions, Y. Meyer, and S. Semmes, *Compensated compactness and Hardy spaces*, J. Math. Pures Appl. (9), 72 (1993), pp. 247–286.
[27] M. Dorobantu and B. Engquist, *Wavelet-based numerical homogenization*, SIAM J. Numer. Anal., 35 (1998), pp. 540–559.
[28] W. E, B. Engquist, X. Li, W. Ren, and E. Vanden-Eijnden, *The Heterogeneous Multiscale Method: A Review*, preprint, Princeton, http://www.math.princeton.edu/multiscale/review.pdf.
[29] Y. R. Efendiev, T. Y. Hou, and X.-H. Wu, *Convergence of a nonconforming multiscale finite element method*, SIAM J. Numer. Anal., 37 (2000), pp. 888–910.
[30] A. Ern and J.-L. Guermond, *Theory and Practice of Finite Element Methods*, Appl. Math. Sci. 159, Springer, New York, 2004. Two volumes bound as one.
[31] L. Evans, *Partial Differential Equations*, Grad. Stud. Math. 19, American Mathematical Society, Providence, RI, 1997.
[32] C. Farhat, I. Harari, and L. P. Franca, *The discontinuous enrichment method*, Comput. Methods Appl. Mech. Engrg., 190 (2001), pp. 6455–6479.
[33] C. Farhat, I. Harari, and U. Hetmaniuk, *The discontinuous enrichment method for multiscale analysis*, Comput. Methods Appl. Mech. Engrg., 192 (2003), pp. 3195–3209.
[34] V. N. Fenchenko and Ē. Ya. Khruslov, *Asymptotic behavior of solutions of differential equations with a strongly oscillating coefficient matrix that does not satisfy a uniform boundedness condition*, Dokl. Akad. Nauk Ukrain. SSR Ser. A, 95 (1981), pp. 24–27.
[35] J. Fish and A. Wagiman, *Multiscale finite element method for a locally nonperiodic heterogeneous medium*, Comput. Mech., 12 (1993), pp. 164–180.
[36] J. Fish and Y. Zheng, *Multi-scale enrichment based on partition of unity*, Internat. J. Numer. Methods Engrg., (2005).
[37] J. Garnier, *Homogenization in a periodic and time-dependent potential*, SIAM J. Appl. Math., 57 (1997), pp. 95–111.
[38] A. C. Gilbert, *A comparison of multiresolution and classical one-dimensional homogenization schemes*, Appl. Comput. Harmon. Anal., 5 (1998), pp. 1–35.
[39] O. Goubet, *Séparation des variables dans le problème de Stokes. Application à son approximation multiéchelles éléments finis*, C. R. Acad. Sci. Paris Sér. I Math., 315 (1992), pp. 1315–1318.
[40] V. H. Hoang and C. Schwab, *High-dimensional finite elements for elliptic problems with multiple scales*, Multiscale Model. Simul., 3 (2005), pp. 168–194.
[41] T. Y. Hou and X.-H. Wu, *A multiscale finite element method for elliptic problems in composite materials and porous media*, J. Comput. Phys., 134 (1997), pp. 169–189.
[42] T. Y. Hou, X.-H. Wu, and Y. Zhang, *Removing the cell resonance error in the multiscale finite element method via a Petrov-Galerkin formulation*, Commun. Math. Sci., 2 (2004), pp. 185–205.
[43] V. V. Jikov, S. M. Kozlov, and O. A. Oleinik, *Homogenization of Differential Operators and Integral Functionals*, Springer, Berlin, 1991.
[44] Ē. Ya. Khruslov, *Homogenized models of composite media*, in Composite Media and Homogenization Theory (Trieste, 1990), Progr. Nonlinear Differential Equations Appl. 5, Birkhäuser Boston, Boston, MA, 1991, pp. 159–182.
[45] C. Kipnis and S. R. S. Varadhan, *Central limit theorem for additive functional of reversible markov processes and application to simple exclusion*, Comm. Math. Phys., 104 (1986), pp. 1–19.
[46] S. M. Kozlov, *The method of averaging and walks in inhomogeneous environments*, Russian Math. Surveys, 2 (1985), pp. 73–145.
[47] S. M. Kozlov, *Averaging of difference schemes*, Math. USSR. Sbornik, 2 (1987), pp. 351–369.
[48] S. Larson and V. Thomee, *Partial Differential Equations with Numerical Methods*, Texts Appl. Math. 45, Springer, New York, 2003.
[49] S. H. Lee, P. Jenny, and H. A. Tchelepi, *A finite-volume method with hexahedral multiblock grids for modeling flow in porous media*, Comput. Geosci., 6 (2002), pp. 353–379.
[50] S. H. Lee, P. Jenny, and H. A. Tchelepi, *Multi-scale finite-volume method for elliptic problems in subsurface flow simulation*, J. Comput. Phys., 187 (2003), pp. 47–67.
[51] S. Leonardi, *Weighted Miranda-Talenti inequality and applications to equations with discontinuous coefficients*, Comment. Math. Univ. Carolin., 43 (2002), pp. 43–59.
[52] A. Maugeri, D. K. Palagachev, and L. G. Softova, *Elliptic and Parabolic Equations with Discontinuous Coefficients*, Mathematical Research 109, Wiley-VCH, New York, 2000.
[53] F. Murat, *Compacité par compensation*, Ann. Sc. Norm. Super. Pisa Cl. Sci. (4), 5 (1978), pp. 489–507.

[54] F. Murat and L. Tartar, *H-convergence*, in Topics in the Mathematical Modelling of Composite Materials, Progr. Nonlinear Differential Equations Appl. 31, Birkhäuser Boston, Boston, MA, 1997, pp. 21–43.
[55] A. A. Oberai and P. M. Pinsky, *A multiscale finite element method for the Helmholtz equation*, Comput. Methods Appl. Mech. Engrg., 154 (1998), pp. 281–297.
[56] H. Oh and I. Babuška, *The p-version of the finite element method for the elliptic boundary value problems with interfaces*, Comput. Methods Appl. Mech. Engrg., 97 (1992), pp. 211–231.
[57] H. Oh and I. Babuška, *The method of auxiliary mapping for the finite element solutions of elasticity problems containing singularities*, J. Comput. Phys., 121 (1995), pp. 193–212.
[58] H. Owhadi, *Anomalous slow diffusion from perpetual homogenization*, Ann. Probab., 31 (2003), pp. 1935–1969.
[59] H. Owhadi, *Averaging versus chaos in turbulent transport?*, Comm. Math. Phys., 247 (2004), pp. 553–599.
[60] H. Owhadi and L. Zhang, *Metric based upscaling*, Commun. Pure Appl. Math., 60 (2007), pp. 675–723.
[61] G. Papanicolaou and S. R. S. Varadhan, *Boundary value problems with rapidly oscillating random coefficients*, in Colloquia Mathematica Societatis János Bolay 27, 1979.
[62] G. Papanicolaou and S. R. S. Varadhan, *Diffusions with random coefficients*, in Statistics and Probability. Essays in Honour of C. R. Rao, N. Ikeda, ed., North-Holland, Amsterdam, 1982, pp. 547–552.
[63] G. C. Papanicolaou, D. Stroock, and S. R. S. Varadhan, *Martingale approach to some limit theorems*, in Papers from the Duke Turbulence Conference (Duke Univ., Durham, N.C., 1976), Paper No. 6, Duke Univ., Durham, NC, 1977, pp. ii+120 pp. Duke Univ. Math. Ser., Vol. III.
[64] G. C. Papanicolaou and S. R. S. Varadhan, *A limit theorem with strong mixing in Banach space and two applications to stochastic differential equations*, Comm. Pure Appl. Math., 26 (1973), pp. 497–524.
[65] S. Spagnolo, *Sulla convergenza di soluzioni di equazioni paraboliche ed ellittiche.*, Ann. Sc. Norm. Super. Pisa (3) 22 (1968), 571-597; errata, ibid. (3), 22 (1968), p. 673.
[66] D. W. Stroock and S. R. S. Varadhan, *Diffusion processes and martingales.* I, in Martingales (Report Meeting, Oberwolfach, 1970), Lecture Notes in Math. 190, Springer, Berlin, 1971, pp. 60–66.
[67] D. W. Stroock and S. R. S. Varadhan, *Diffusion processes and martingales.* II, in Martingales (Report Meeting, Oberwolfach, 1970, Lecture Notes in Math. 190, Springer, Berlin, 1971, pp. 67–75.
[68] D. W. Stroock and S. R. S. Varadhan, *Martingales.* I, II, III, in Topics in Probability Theory (Sem., Courant Inst. Math. Sci., New York, 1971–1972), Courant Inst. Math. Sci., New York, 1973, pp. 113–161.
[69] D. W. Stroock and S. R. Srinivasa Varadhan, *Multidimensional Diffusion Processes*, Grundlehren Math. Wis. 233, Springer, Berlin, 1979.
[70] L. Tartar, *Compensated compactness and applications to partial differential equations*, in Nonlinear Analysis and Mechanics: Heriot-Watt Symposium, Vol. IV, Res. Notes in Math. 39, Pitman, Boston, MA, 1979, pp. 136–212.
[71] L. Tartar, *Homogénéisation et compacité par compensation*, in Séminaire Goulaouic-Schwartz (1978/1979), École Polytech., Palaiseau, 1979, pp. Exp. No. 9, 9.
[72] S. R. S. Varadhan, *private communication*, 2005.
[73] W. L. Wan, T. F. Chan, and B. Smith, *An energy-minimizing interpolation for robust multigrid methods*, SIAM J. Sci. Comput., 21 (1999), pp. 1632–1649.
[74] X. H. Wu, Y. Efendiev, and T. Y. Hou, *Analysis of upscaling absolute permeability*, Discrete Contin. Dyn. Syst. Ser. B, 2 (2002), pp. 185–204.
[75] X. H. Wu, Y. Efendiev, and T. Y. Hou, *Accurate multiscale finite element methods for two-phase flow simulations*, J. Comput. Phys., 220 (2006), pp. 155–174.
[76] J. Xu, *An introduction to multigrid convergence theory*, in Iterative Methods in Scientific Computing (Hong Kong, 1995), Springer, Singapore, 1997, pp. 169–241.
[77] J. Xu and J. Qin, *Some remarks on a multigrid preconditioner*, SIAM J. Sci. Comput., 15 (1994), pp. 172–184.
[78] V. V. Zhikov, S. M. Kozlov, O. A. Oleinik, and K. T. Ngoan, *Averaging and g-convergence of differential operators*, Russian Math. Surveys, 34 (1979), pp. 69–147.

SIAM J. NUMER. ANAL.
Vol. 46, No. 1, pp. 37–60
© 2007 Society for Industrial and Applied Mathematics

CONVERGENCE ANALYSIS OF FINITE VOLUME SCHEME FOR NONLINEAR TENSOR ANISOTROPIC DIFFUSION IN IMAGE PROCESSING*

OLGA DRBLÍKOVÁ† AND KAROL MIKULA†

Abstract. In this article we design the semiimplicit finite volume scheme for coherence enhancing diffusion in image processing and prove its convergence to the weak solution of the problem. The finite volume methods are natural tools for image processing applications since they use piecewise constant representation of approximate solutions similarly to the structure of digital images. They have been successfully applied in image processing, e.g., for solving the Perona–Malik equation or curvature-driven level set equations, where the nonlinearities are represented by a scalar function dependent on a solution gradient. Design of suitable finite volume schemes for tensor diffusion is a nontrivial task here we present the first such scheme with a convergence proof for the practical nonlinear model used in coherence-enhancing image smoothing. We provide basic information about this type of nonlinear diffusion including a construction of its diffusion tensor, and we derive a semiimplicit finite volume scheme for this nonlinear model with the help of covolume mesh. This method is well known as the diamond-cell method owing to the choice of covolume as a diamond-shaped polygon. Further, we prove a convergence of a discrete solution given by our scheme to the weak solution of the problem. The proof is based on Kolmogorov's compactness theorem and a bounding of a gradient in the tangential direction by using a gradient in the normal direction. Finally computational results illustrated in figures are discussed.

Key words. image processing, nonlinear tensor diffusion, numerical solution, semiimplicit scheme, diamond-cell finite volume method, convergence

AMS subject classifications. 74S10, 65M12, 35K60, 68U10

DOI. 10.1137/070685038

1. Introduction.

Nonlinear diffusion models are widely used nowadays in many practical tasks of image processing. In this paper we deal with the numerical solution of the model of tensor nonlinear anisotropic diffusion introduced by Weickert (see [23, 24, 22]) in the following form:

$$\frac{\partial u}{\partial t} - \nabla \cdot (D\nabla u) = 0 \qquad \text{in } Q_T \equiv I \times \Omega, \tag{1.1}$$

$$u(x,0) = u_0(x) \quad \text{in } \Omega, \tag{1.2}$$

$$(D\nabla u) \cdot \mathbf{n} = 0 \qquad \text{on } I \times \partial\Omega, \tag{1.3}$$

where D is a matrix depending on the eigenvalues and eigenvectors of the so-called (regularized) structure tensor, $u_0 \in L^2(\Omega)$, and $\mathbf{n}$ is the outer normal unit vector to $\partial\Omega$. Such a model is useful in any situation, where strong smoothing is desirable in a preferred direction and a low smoothing is expected in the perpendicular direction, e.g., for images with interrupted coherence of structures. To that goal the matrix

$$J_0(\nabla u_{\tilde{t}}) = \nabla u_{\tilde{t}} \otimes \nabla u_{\tilde{t}} = \nabla u_{\tilde{t}} \nabla u_{\tilde{t}}^T, \tag{1.4}$$

*Received by the editors March 12, 2007; accepted for publication (in revised form) September 10, 2007; published electronically December 12, 2007. This work was supported by EC projects Embryomics and BioEmergences of the 6th Framework program and by the projects VEGA 1/3321/06 and APVV-RPEU-0004-06.

http://www.siam.org/journals/sinum/46-1/68503.html

†Department of Mathematics, Slovak University of Technology, Radlinského 11, 813 68 Bratislava, Slovakia (drblikov@math.sk, mikula@math.sk).

where

$$u_{\tilde{t}}(x,t) = (G_{\tilde{t}} * u(\cdot,t))(x) \qquad (\tilde{t} > 0) \tag{1.5}$$

is used. The matrix J_0 is symmetric and positive semidefinite, and its eigenvectors are parallel and orthogonal to $\nabla u_{\tilde{t}}$, respectively. We can average J_0 by applying another convolution with Gaussian G_ρ and define

$$J_\rho(\nabla u_{\tilde{t}}) = G_\rho * (\nabla u_{\tilde{t}} \otimes \nabla u_{\tilde{t}}) \qquad (\rho > 0). \tag{1.6}$$

In computer vision the matrix $J_\rho = \left(\begin{smallmatrix} a & b \\ b & c \end{smallmatrix}\right)$ is known as a structure tensor, interest operator, or second moment matrix (see [9]). It is again symmetric and positive semidefinite, and its eigenvalues are given by

$$\mu_{1,2} = \frac{1}{2}\left(a + c \pm \sqrt{(a-c)^2 + 4b^2}\right), \qquad \mu_1 \geq \mu_2. \tag{1.7}$$

Since the eigenvalues integrate the variation of the gray values within a neighborhood of size $O(\rho)$, they describe the average contrast in the eigendirections v and w.

With the help of the eigenvalues of J_ρ we can obtain useful information on the coherence. The expression $(\mu_1 - \mu_2)^2$ is large for anisotropic structures and tends to zero for isotropic structures; constant areas are characterized by $\mu_1 = \mu_2 = 0$, straight edges by $\mu_1 \gg \mu_2 = 0$, and corners by $\mu_1 \geq \mu_2 \gg 0$.

The corresponding orthogonal set of eigenvectors (v, w) to eigenvalues (μ_1, μ_2) is given by

$$\begin{aligned} &v = (v_1, v_2), && w = (w_1, w_2), \\ &v_1 = 2b, && v_2 = c - a + \sqrt{(a-c)^2 + 4b^2}, \\ &w \perp v, && w_1 = -v_2, \qquad w_2 = v_1. \end{aligned} \tag{1.8}$$

The orientation of the eigenvector w which corresponds to the smaller eigenvalue μ_2 is called coherence orientation. This orientation has the lowest fluctuations.

One can use the above-mentioned structure tensor information in a construction of specific nonlinear diffusion filter [23, 24, 22]. The idea of the tensor nonlinear diffusion filtering is as follows. We get a processed version $u(x,t)$ of an original image $u_0(x)$ with a scale parameter $t \geq 0$ as the solution of mathematical model (1.1)–(1.3), where matrix D depends on solution u, satisfies smoothness, symmetry, and uniform positive definiteness properties, and steers a filtering process such that diffusion is strong along the coherence direction w and increases with the coherence $(\mu_1 - \mu_2)^2$. To that goal D must possess the same eigenvectors v and w as the structure tensor $J_\rho(\nabla u_{\tilde{t}})$, and we choose the eigenvalues of D as

$$\begin{aligned} &\kappa_1 = \alpha, \quad \alpha \in (0,1),\ \alpha \ll 1, \\ &\kappa_2 = \begin{cases} \alpha & \text{if } \mu_1 = \mu_2, \\ \alpha + (1-\alpha)\exp\left(\frac{-C}{(\mu_1-\mu_2)^2}\right),\ C > 0 & \text{otherwise.} \end{cases} \end{aligned} \tag{1.9}$$

The matrix D then has the following form:

$$D = ABA^{-1}, \tag{1.10}$$

where $A = \left(\begin{smallmatrix} v_1 & -v_2 \\ v_2 & v_1 \end{smallmatrix}\right)$ and $B = \left(\begin{smallmatrix} \kappa_1 & 0 \\ 0 & \kappa_2 \end{smallmatrix}\right)$. The exponential function is used in (1.9) because it ensures that the smoothness of the structure tensor carries over to the

diffusion tensor and that κ_2 does not exceed 1. The positive parameter α guarantees that the process never stops. Even if $(\mu_1-\mu_2)^2$ tends to zero so the structure becomes isotropic, there still remains some small linear diffusion with diffusivity $\alpha > 0$. Such α is a regularization parameter, which keeps the diffusion tensor uniformly positive definite. C has the role of a threshold parameter. If $(\mu_1 - \mu_2)^2 \gg C$, then $\kappa_2 \approx 1$, and, in contrast, if $(\mu_1 - \mu_2)^2 \ll C$, then $\kappa_2 \approx \alpha$. Due to the convolutions in (1.5) and (1.6), the elements of matrix D are C^∞ functions. Such a model is a nontrivial extension of the regularized Perona–Malik equation [17, 1, 15], and, as well as further PDEs employing tensor diffusion, it is used in many practical image processing applications; see, e.g., [23, 24, 22, 6, 13, 19, 18]. In section 5 of this paper we also illustrate its usefulness by smoothing and segmenting the cell membrane images obtained by a confocal microscope. We show that after application of the nonlinear tensor anisotropic diffusion using our numerical scheme the coherent structures are attenuated. If such improved edge information is used in the so-called subjective surface segmentation method [20, 16, 2], the cell boundaries are correctly segmented.

There are only a few purely finite volume methods designed and studied from the numerical analysis point of view for solving tensor diffusion problems; see, e.g., [3, 4, 5, 26] devoted to discretization of the elliptic operators. On the other hand, finite volume schemes for nonlinear parabolic problems as arising in image analysis are natural since they use piecewise constant representation of approximate solutions similarly to the structure of digital images. Finite and complementary volume schemes have been used successively in image processing for solving the Perona–Malik equation and its generalizations (see, e.g., [15, 11, 12, 10, 7, 21]) and for solving the generalized curvature-driven level set equations (see, e.g., [8, 16, 2]) where the nonlinearities are represented by a scalar function dependent on a solution gradient. Here we present the first finite volume scheme with a convergence proof for the highly nonlinear anisotropic tensor diffusion model arising in coherence-enhancing image smoothing.

The next section is devoted to derivation of our numerical scheme, in section 3 we study the existence and uniqueness of discrete solutions, section 4 contains our convergence proof, and finally, in section 5, we discuss numerical experiments.

2. Finite volume scheme for nonlinear tensor anisotropic diffusion. The aim of this section is to derive our computational method. Let the image be represented by $n_1 \times n_2$ pixels (finite volumes) such that it looks like a mesh with n_1 rows and n_2 columns. Let $\Omega = (0, n_1 h)\times(0, n_2 h)$, h is a pixel size, and let the image $u(x)$ be given by a bounded mapping $u : \Omega \to R$. The filtering process is considered in a time interval $I = [0, T]$. Let $0 = t_0 \le t_1 \le \cdots \le t_{N_{max}} = T$ denote the time discretization, with $t_n = t_{n-1} + k$, where k is the length of a discrete time step. In our scheme we will look for u^n, an approximation of solution at time t_n, for every $n = 1, \ldots, N_{max}$. As usual in finite volume methods, we integrate (1.1) over finite volume K, then provide a semiimplicit time discretization, and use a divergence theorem to get

$$\frac{u_K^n - u_K^{n-1}}{k} m(K) - \sum_{\sigma \in \mathcal{E}_K \cap \mathcal{E}_{int}} \int_\sigma (D^{n-1} \nabla u^n) \cdot \mathbf{n}_{K,\sigma} ds = 0, \tag{2.1}$$

where u_K^n, $K \in \mathcal{T}_h$, represents the mean value of u^n on K. $\mathcal{T}_h$ is an admissible finite volume mesh (see [4]), and further quantities and notations are described as follows: $m(K)$ is the measure of the finite volume K with boundary ∂K, and $\sigma_{KL} = K \cap L = K|L$ is an edge of the finite volume K, where $L \in \mathcal{T}_h$ is an adjacent finite volume to K such that $m(K \cap L) \neq 0$. Due to simplifying notation, we use σ instead of

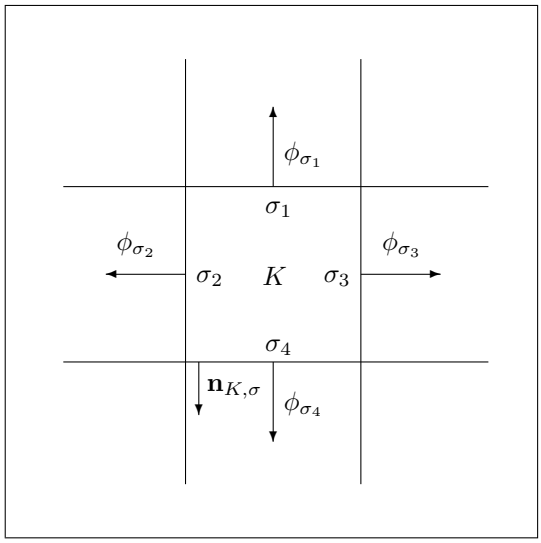


FIG. 2.1. *A detail of a finite volume mesh—a finite volume K, its boundaries σ_i, $i = 1, 2, 3, 4$, and the fluxes outward to a finite volume K.*

σ_{KL} at several places if no confusion can appear. $\mathcal{E}_K$ is a set of edges such that $\partial K = \bigcup_{\sigma\in\mathcal{E}_K} \sigma$ and $\mathcal{E} = \bigcup_{K\in\mathcal{T}_h} \mathcal{E}_K$. The set of boundary edges is denoted by $\mathcal{E}_{ext}$, that is, $\mathcal{E}_{ext} = \mathcal{E} \cap \partial\Omega$, and let $\mathcal{E}_{int} = \mathcal{E} \setminus \mathcal{E}_{ext}$. Υ is the set of pairs of adjacent finite volumes, defined by $\Upsilon = \{(K, L) \in \mathcal{T}_h^2,\ K \neq L,\ m(K|L) \neq 0\}$, and $\mathbf{n}_{K,\sigma}$ is the normal unit vector to σ outward to K. See Figure 2.1.

Let us define our discrete numerical solution by

$$u_{h,k}(x,t) = \sum_{n=0}^{N_{max}} \sum_{K\in\mathcal{T}_h} u_K^n \chi\{x \in K\}\chi\{t_{n-1} < t \le t_n\}, \tag{2.2}$$

where the function $\chi(A)$ is defined as

$$\chi_{\{A\}} = \begin{cases} 1 & \text{if A is true,} \\ 0 & \text{elsewhere.} \end{cases} \tag{2.3}$$

The extension of the function (2.2) outside Ω is given first by its periodic mirror reflection in $\Omega_{\tilde{t}}$, where $\tilde{t}$ is the width of the smoothing kernel:

$$\Omega_{\tilde{t}} = \Omega \cup B_{\tilde{t}}(x), \quad x \in \partial\Omega, \tag{2.4}$$

$B_{\tilde{t}}(x)$ is a ball centered at x with radius $\tilde{t}$, and then we extend this periodic mirror reflection by 0 outside $\Omega_{\tilde{t}}$ and denote it by $\tilde{u}_{h,k}$.

In our scheme we will start computation by defining initial values

$$u_K^0 = \frac{1}{m(K)} \int_K u_0(x)dx, \quad K \in \mathcal{T}_h \tag{2.5}$$

and letting $u_{h,k}^n(x) = \sum_{K\in\mathcal{T}_h} u_K^n \chi\{x \in K\}$ denote a finite volume approximation at the nth time step. In order to get the scheme we write (2.1) in the form $\frac{u_K^n - u_K^{n-1}}{k} - \frac{1}{m(K)} \sum_{\sigma\in\mathcal{E}_K\cap\mathcal{E}_{int}} \phi_\sigma^n(u_{h,k}^n) m(\sigma) = 0$, where $m(\sigma)$ is the measure of edge σ and $\phi_\sigma^n(u_{h,k}^n)$

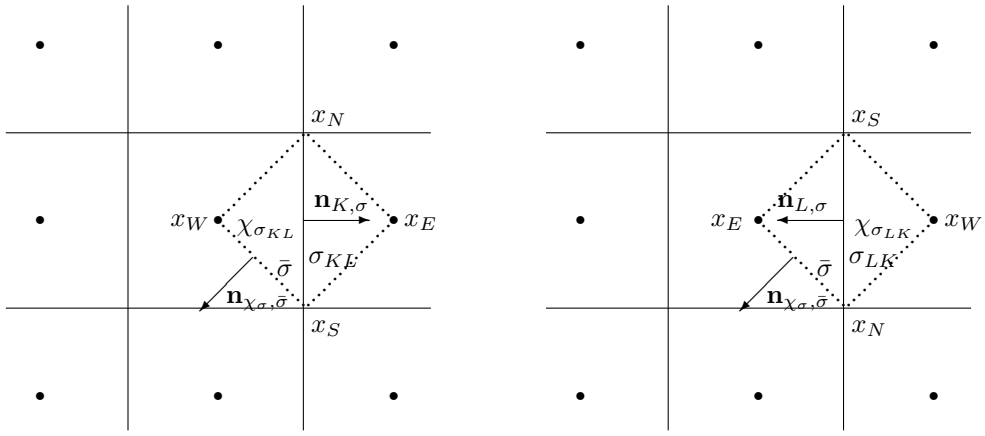


FIG. 2.2. *A detail of a mesh. The covolumes associated to edges $\sigma = \sigma_{KL}$ (left) and $\sigma = \sigma_{LK}$ (right).*

denotes an approximation of the exact averaged flux $\frac{1}{m(\sigma)}\int_\sigma (D^{n-1}\nabla u^n)\cdot\ \mathbf{n}_{K,\sigma}ds$ for any K and $\sigma \in \mathcal{E}_K$.

We construct $\phi^n_\sigma(u^n_{h,k})$ with the help of a covolume mesh (see, e.g., [3]). The covolume χ_σ associated to σ is constructed around each edge by joining end points of this edge and midpoints of finite volumes which are common to this edge; see Figure 2.2. We denote the end points of an edge $\bar{\sigma} \subset \partial\chi_\sigma$ by $N_1(\bar{\sigma})$ and $N_2(\bar{\sigma})$ and let $\mathbf{n}_{\chi_\sigma,\bar{\sigma}}$ be the normal unit vector to $\bar{\sigma}$ outward to χ_σ. In order to approximate diffusion flux, using the divergence theorem, we first derive an approximation of the averaged gradient on χ_σ, namely, $\frac{1}{m(\chi_\sigma)}\int_{\chi_\sigma}\nabla u^n dx = \frac{1}{m(\chi_\sigma)}\int_{\partial\chi_\sigma} u^n\mathbf{n}_{\chi_\sigma,\bar{\sigma}}ds$, and then we approximate it by $p^n_\sigma(u) = \frac{1}{m(\chi_\sigma)}\sum_{\bar{\sigma}\in\partial\chi_\sigma}\frac{1}{2}(u^n_{N_1(\bar{\sigma})} + u^n_{N_2(\bar{\sigma})})m(\bar{\sigma})\mathbf{n}_{\chi_\sigma,\bar{\sigma}}$. Let the values at x_E and x_W be taken as u_E and u_W, and let the values u_S and u_N at the vertices x_N and x_S be computed as the arithmetic mean of u_K, where K are finite volumes which are common to this vertex.

Since our mesh is uniform and squared, we can use the following relations: $m(\chi_\sigma) = \frac{h^2}{2}$ and $m(\bar{\sigma}) = \frac{\sqrt{2}}{2}h$, and after a short calculation we are ready to write

$$p^n_\sigma(u) = \frac{u^n_E - u^n_W}{h}\mathbf{n}_{K,\sigma} + \frac{u^n_N - u^n_S}{h}\mathbf{t}_{K,\sigma}, \tag{2.6}$$

where $\mathbf{t}_{K,\sigma}$ is a unit vector parallel to σ such that $(x_N - x_S)\cdot\mathbf{t}_{K,\sigma} > 0$. Although such u^n_N, u^n_W, u^n_E, and u^n_S correspond to a particular edge σ, and so we should denote them by $u^n_{N_\sigma}$, $u^n_{W_\sigma}$, $u^n_{E_\sigma}$, and $u^n_{S_\sigma}$, respectively, in (2.6), we will use the above simplified notations. By replacing the exact gradient ∇u^n by the numerical gradient $p^n_\sigma(u)$ in the approximation of $\phi^n_\sigma(u^n_{h,k})$, we get the numerical flux in the form

$$\phi^n_\sigma(u^n_{h,k}) = (D_\sigma p^n_\sigma(u))\cdot\mathbf{n}_{K,\sigma}, \tag{2.7}$$

where $D_\sigma = D^{n-1}_\sigma = \left(\begin{smallmatrix}\bar{\lambda}_\sigma & \bar{\beta}_\sigma\\ \bar{\beta}_\sigma & \bar{\nu}_\sigma\end{smallmatrix}\right)$ is an approximation of the mean value of matrix D along σ evaluated at the previous time step. To that goal we take $u^{n-1}_{h,k}$ for constructing the structure and diffusion tensor and evaluate them at x_{KL}, where x_{KL} is a point of $\sigma_{KL} = K|L$ intersecting the segment $x_K x_L$. From an implementation point of view, the structure and then diffusion tensor evaluation can be done in two ways.

Either we can replace gradients of u appearing in structure tensor by their numerical approximation $p_\sigma^n(u)$ and then smooth them by weighted average (convolution), or we can evaluate $\nabla G_{\tilde{t}} * u_{h,k}^{n-1}$ using weights given by $\nabla G_{\tilde{t}}$ applied to discrete piecewise constant values of $u_{h,k}^{n-1}$ as convolution realization. In the latter way we do not introduce additional approximation into the scheme, and in the part devoted to convergence analysis we use the latter approach, although both are realizable computationally.

It is important to note that in (2.7) we always consider the matrix D_σ written in the basis $(\mathbf{n}_{K,\sigma}, \mathbf{t}_{K,\sigma})$; cf. [3]. Although it may look artificial, it will simplify further considerations. In practice it means that (cf. Figure 2.1) if the matrix D is given in standard basis on edge σ by $\left(\begin{smallmatrix}\lambda_\sigma & \beta_\sigma\\ \beta_\sigma & \nu_\sigma\end{smallmatrix}\right)$, then $D_\sigma = \left(\begin{smallmatrix}\lambda_\sigma & \beta_\sigma\\ \beta_\sigma & \nu_\sigma\end{smallmatrix}\right)$, i.e., $\bar{\lambda}_\sigma = \lambda_\sigma$, $\bar{\beta}_\sigma = \beta_\sigma$, $\bar{\nu}_\sigma = \nu_\sigma$ for the two edges $\sigma = \sigma_2$ and σ_3. On the other hand, $D_\sigma = \left(\begin{smallmatrix}\nu_\sigma & -\beta_\sigma\\ -\beta_\sigma & \lambda_\sigma\end{smallmatrix}\right)$, i.e., $\bar{\lambda}_\sigma = \nu_\sigma$, $\bar{\beta}_\sigma = -\beta_\sigma$, $\bar{\nu}_\sigma = \lambda_\sigma$ for other two edges $\sigma = \sigma_1$ and σ_4. By using such a matrix representation, the definition (2.7) can be written in this compact form

$$\phi_\sigma^n(u_{h,k}^n) = \left[\begin{pmatrix}\bar{\lambda}_\sigma & \bar{\beta}_\sigma\\ \bar{\beta}_\sigma & \bar{\nu}_\sigma\end{pmatrix}\begin{pmatrix}\frac{u_E^n - u_W^n}{h}\\ \frac{u_N^n - u_S^n}{h}\end{pmatrix}\right]\cdot\begin{pmatrix}1\\0\end{pmatrix} = \bar{\lambda}_\sigma\frac{u_E^n - u_W^n}{h} + \bar{\beta}_\sigma\frac{u_N^n - u_S^n}{h}, \tag{2.8}$$

since in the basis $(\mathbf{n}_{K,\sigma}, \mathbf{t}_{K,\sigma})$ the formula (2.6) can be written for each edge as

$$p_\sigma^n(u) = \begin{pmatrix}\frac{u_E^n - u_W^n}{h}\\ \frac{u_N^n - u_S^n}{h}\end{pmatrix} \tag{2.9}$$

and $\mathbf{n}_{K,\sigma}$ is equal to $\binom{1}{0}$ in the basis $(\mathbf{n}_{K,\sigma}, \mathbf{t}_{K,\sigma})$ for each edge σ. Because of the convolutions in (1.5) and (1.6), the elements of matrix D_σ are C^∞ functions.

Finally, let us summarize our *semiimplicit finite volume scheme*:

$$\frac{u_K^n - u_K^{n-1}}{k} - \frac{1}{m(K)}\sum_{\sigma\in\mathcal{E}_K\cap\mathcal{E}_{int}}\phi_\sigma^n(u_{h,k}^n)m(\sigma) = 0, \tag{2.10}$$

where

$$\phi_\sigma^n(u_{h,k}^n) = \bar{\lambda}_\sigma\frac{u_E^n - u_W^n}{h} + \bar{\beta}_\sigma\frac{u_N^n - u_S^n}{h}. \tag{2.11}$$

3. Existence and uniqueness of the solution to the discrete scheme. In order to fulfill the main goal of this section, to prove the existence and uniqueness of u_K^n, $K\in\mathcal{T}_h$, we estimate the expressions $u_N^n - u_S^n$ by means of $u_E^n - u_W^n$ for all edges σ. To that goal we use mainly the results of [3] in our situation. Let us note that, due to simplification of notation, we do not use upper index n in what follows, and at some places we relate u_E and u_W to particular edge σ using u_{E_σ}, u_{W_σ}, etc. In the following we denote by C_i constants which may depend on the properties of the diffusion tensor.

DEFINITION 3.1. *Let P_σ be the set of all edges δ perpendicular to σ (see Figure* 3.1 *for two illustrative situations when $\sigma = \sigma_{WE}$ and $\sigma = \sigma_{EW}$), which have a common vertex with σ and fulfill the following conditions: $x_{E_\delta} - x_{W_\delta} > 0$ if $x_{N_\sigma} - x_{S_\sigma} > 0$ and $x_{E_\delta} - x_{W_\delta} < 0$ if $x_{N_\sigma} - x_{S_\sigma} < 0$. Let us note that $x_{W_\sigma} = x_{W_\delta}^1 = x_{E_\delta}^3$ for $\sigma = \sigma_{WE}$, $x_{E_\sigma} = x_{W_\delta}^2 = x_{E_\delta}^4$ for $\sigma = \sigma_{WE}$, $x_{W_\sigma} = x_{E_\delta}^2 = x_{W_\delta}^4$ for $\sigma = \sigma_{EW}$, and $x_{E_\sigma} = x_{E_\delta}^1 = x_{W_\delta}^3$ for $\sigma = \sigma_{EW}$.*

Using definitions given in the previous section we can write

$$u_N - u_S = \frac{1}{4}\left[(u_E^1 - u_W^1) + (u_E^3 - u_W^3) + (u_E^2 - u_W^2) + (u_E^4 - u_W^4)\right], \tag{3.1}$$

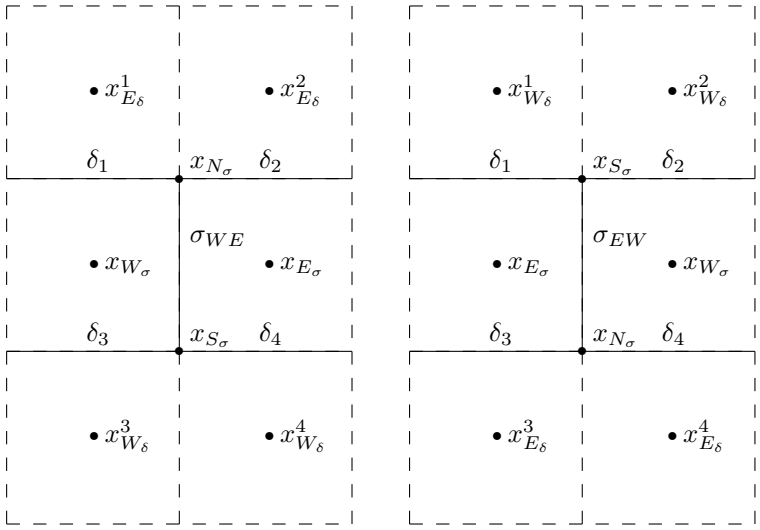


FIG. 3.1. *Left: An edge σ_{WE} and edges δ_1, δ_2, δ_3, $\delta_4 \in P_{\sigma_{WE}}$. Right: An edge σ_{EW} and edges δ_1, δ_2, δ_3, $\delta_4 \in P_{\sigma_{EW}}$.*

where $u^1_E = u_{E_{\delta_1}}$ and $u^1_W = u_{W_{\delta_1}}$ correspond to edge δ_1 and similarly u^2_E, u^2_W, u^3_E, u^3_W, u^4_E, and u^4_W correspond to edges δ_2, δ_3 and δ_4. Applying the inequality $(a+b)^2 \le 2a^2 + 2b^2$ we have

$$(u_{N_\sigma} - u_{S_\sigma})^2 \le \sum_{\delta \in P_\sigma \cap \mathcal{E}_{int}} \frac{1}{4}(u_{E_\delta} - u_{W_\delta})^2. \tag{3.2}$$

Multiplying (3.2) by $(\frac{\bar\beta_\sigma}{\bar\lambda_\sigma})^2 \frac{\bar\lambda_\sigma}{h^2}$ and summing for all $\sigma \in \mathcal{E}_{int}$ (by σ we mean σ_{WE}) we obtain

$$\sum_{\sigma \in \mathcal{E}_{int}} \left(\frac{\bar\beta_\sigma}{\bar\lambda_\sigma}\right)^2 \left(\frac{u_{N_\sigma} - u_{S_\sigma}}{h}\right)^2 \bar\lambda_\sigma \le \sum_{\sigma \in \mathcal{E}_{int}} \left(\frac{\bar\beta_\sigma}{\bar\lambda_\sigma}\right)^2 \sum_{\delta \in P_\sigma \cap \mathcal{E}_{int}} \frac{1}{4} \left(\frac{u_{E_\delta} - u_{W_\delta}}{h}\right)^2 \bar\lambda_\sigma. \tag{3.3}$$

Then we swap the two sums on the right-hand side of (3.3) to get

$$\sum_{\sigma \in \mathcal{E}_{int}} \left(\frac{\bar\beta_\sigma}{\bar\lambda_\sigma}\right)^2 \left(\frac{u_{N_\sigma} - u_{S_\sigma}}{h}\right)^2 \bar\lambda_\sigma \le \sum_{\delta \in \mathcal{E}_{int}} \gamma_\delta \left(\frac{u_{E_\delta} - u_{W_\delta}}{h}\right)^2 \bar\lambda_\delta, \tag{3.4}$$

where

$$\gamma_\delta = \sum_{\sigma \in P_\delta \cap \mathcal{E}_{int}} \frac{1}{4} \left(\frac{\bar\beta_\sigma}{\bar\lambda_\sigma}\right)^2 \frac{\bar\lambda_\sigma}{\bar\lambda_\delta}. \tag{3.5}$$

Let us consider the matrix $\left(\begin{smallmatrix} \bar\lambda_{\sigma^\perp} & \bar\beta_{\sigma^\perp} \\ \bar\beta_{\sigma^\perp} & \bar\nu_{\sigma^\perp} \end{smallmatrix}\right)$, which is the matrix D written in the basis $(\mathbf{t}_{K,\sigma}, -\mathbf{n}_{K,\sigma})$ on edge σ. Due to the smoothness of D we get

$$\bar\lambda_\sigma = \bar\nu_{\sigma^\perp} = \bar\nu_\delta(1 + O(h)) = \bar\lambda_{\delta^\perp}(1 + O(h)), \quad \delta \in P_\sigma, \tag{3.6}$$

$$\bar\beta_\sigma = -\bar\beta_{\sigma^\perp} = -\bar\beta_\delta(1 + O(h)) = \bar\beta_{\delta^\perp}(1 + O(h)), \quad \delta \in P_\sigma, \tag{3.7}$$

$$\bar\nu_\sigma = \bar\lambda_{\sigma^\perp} = \bar\lambda_\delta(1 + O(h)) = \bar\nu_{\delta^\perp}(1 + O(h)), \quad \delta \in P_\sigma. \tag{3.8}$$

By applying (3.6)–(3.8) in (3.5) and using $\frac{1}{1-Ch} \leq 1 + (C + 2C^2 h)h$ for h sufficiently small ($h \leq \frac{1}{2C}$), we have

$$\gamma_\delta \leq \sum_{\sigma \in P_\delta \cap \mathcal{E}_{int}} \frac{1}{4} \left(\frac{\bar{\beta}_{\delta^\perp}}{\bar{\lambda}_{\delta^\perp}} \right)^2 \frac{\bar{\lambda}_{\delta^\perp}}{\bar{\lambda}_\delta} (1 + O(h)) = \left(\frac{\bar{\beta}_{\delta^\perp}}{\bar{\lambda}_{\delta^\perp}} \right)^2 \frac{\bar{\lambda}_{\delta^\perp}}{\bar{\lambda}_\delta} (1 + O(h)).$$

By using the positive definiteness of the diffusion tensor written in a standard basis as $\begin{pmatrix} \lambda_\delta & \beta_\delta \\ \beta_\delta & \nu_\delta \end{pmatrix}$, we obtain for its determinant

$$\lambda_\delta \nu_\delta - \beta_\delta^2 > 0. \tag{3.9}$$

Now we have two possibilities for γ_δ. Let δ be an arbitrary edge in the mesh parallel to σ_3 (see Figure 2.1). Then $\gamma_\delta \leq (\frac{-\beta_\delta}{\nu_\delta})^2 \frac{\nu_\delta}{\lambda_\delta}(1 + O(h)) = \frac{(\beta_\delta)^2}{\lambda_\delta \nu_\delta}(1 + O(h)) < 1$ for h sufficiently small due to (3.9). Similarly, if δ is any edge oriented perpendicularly to σ_3, we have $\gamma_\delta \leq (\frac{\beta_\delta}{\lambda_\delta})^2 \frac{\lambda_\delta}{\nu_\delta}(1 + O(h)) = \frac{(\beta_\delta)^2}{\lambda_\delta \nu_\delta}(1 + O(h)) < 1$ for h sufficiently small. Thus, due to the fact that $\lambda_\sigma \geq C > 0$ and $\nu_\sigma \geq C > 0$, we obtain $0 \leq \gamma_\delta < 1$ for h sufficiently small. Since this condition is fulfilled for each edge δ, we can rewrite (3.4) as

$$\sum_{\sigma \in \mathcal{E}_{int}} \left(\frac{\bar{\beta}_\sigma}{\bar{\lambda}_\sigma} \right)^2 \left(\frac{u_N - u_S}{h} \right)^2 \bar{\lambda}_\sigma \leq \gamma \sum_{\sigma \in \mathcal{E}_{int}} \left(\frac{u_E - u_W}{h} \right)^2 \bar{\lambda}_\sigma, \tag{3.10}$$

where $0 \leq \gamma < 1, \quad \gamma = \max_{\sigma \in \mathcal{E}} \gamma_\sigma$.

Let us now introduce the space of piecewise constant functions associated to our mesh and discrete H^1 norm for this space. This discrete norm will be used to obtain some estimates on the approximate solution given by the finite volume scheme.

DEFINITION 3.2. *Let Ω be an open bounded polygonal subset of R^2. Let $\mathcal{T}_h$ be an admissible finite volume mesh (see [4]). We define $\mathcal{P}_0(\mathcal{T}_h)$ as the set of functions from Ω to R which are constant over each finite volume K of the mesh $\mathcal{T}_h$.*

DEFINITION 3.3. *Let Ω be an open bounded polygonal subset of R^2. For $u \in \mathcal{P}_0(\mathcal{T}_h)$ we define*

$$|u^n_{h,k}|_{1,\mathcal{T}_h} = \left(\sum_{(K,L) \in \Upsilon} \frac{(u_L - u_K)^2}{d_{K,L}} m(\sigma) \right)^{\frac{1}{2}}, \tag{3.11}$$

where $d_{K,L}$ is the Euclidean distance between x_K and x_L.

Remark that (3.11) can be rewritten for our uniform mesh into the following form:

$$|u^n_{h,k}|_{1,\mathcal{T}_h} = \left(2 \sum_{\sigma \in \mathcal{E}_{int}} \left(\frac{u_E - u_W}{h} \right)^2 m(\chi_\sigma) \right)^{\frac{1}{2}}, \tag{3.12}$$

where $\sigma = \sigma_{WE}$. Let us define a discrete operator $\mathcal{L}_h$ by

$$\mathcal{L}_h(u^n_{h,k}) = u^n_K m(K) - k \sum_{\sigma \in \mathcal{E}_K \cap \mathcal{E}_{int}} \phi^n_\sigma(u^n_{h,k}) m(\sigma).$$

Then solution $u^n_{h,k} \in \mathcal{P}_0(\mathcal{T}_h)$ of our scheme at time t_n is given by

$$\mathcal{L}_h(u^n_{h,k}) = f_{h,k}(u^{n-1}_{h,k}), \tag{3.13}$$

where $f_{h,k}(u_{h,k}^{n-1}) = u_K^{n-1} m(K)$, $K \in \mathcal{T}_h$, and u_K^{n-1} is the value of the piecewise constant function $u_{h,k}^{n-1}$ in K. This equality is a linear system of N equations with N unknowns u_K^n, $K \in \mathcal{T}_h$, $N = card(\mathcal{T}_h)$.

Multiplying $\mathcal{L}_h(u_{h,k})$ by u_K^n, summing over K, and splitting into parts A and B leads to

$$\sum_{K\in\mathcal{T}_h} \mathcal{L}_h(u_{h,k}^n) u_K^n = A + B, \tag{3.14}$$

with

$$A = \sum_{K\in\mathcal{T}_h} (u_K^n)^2 m(K) = ||u_{h,k}^n||^2_{L^2(\Omega)} \tag{3.15}$$

and

$$B = k \sum_{K\in\mathcal{T}_h} u_K^n \sum_{\sigma\in\mathcal{E}_K\cap\mathcal{E}_{int}} -\phi_\sigma^n(u_{h,k}^n) m(\sigma).$$

The above expression can be written in the following form:

$$\begin{aligned} B &= k \sum_{W\in\mathcal{T}_h} u_W^n \sum_{\sigma\in\mathcal{E}_W\cap\mathcal{E}_{int}} -\phi_\sigma^n(u_{h,k}^n) m(\sigma) \\ &= \frac{k}{2} \sum_{\sigma\in\mathcal{E}_{int}} \phi_\sigma^n(u_{h,k}^n) \frac{u_E - u_W}{h} 2m(\chi_\sigma) = Q(u_{h,k}^n) \end{aligned} \tag{3.16}$$

owing to property $\phi_\sigma^n(u_{h,k}^n) = \phi_{\sigma_{WE}}^n(u_{h,k}^n) = -\phi_{\sigma_{EW}}^n(u_{h,k}^n)$. Since $\phi_\sigma^n(u_{h,k}^n) = 0$ for $\sigma \in \mathcal{E}_{ext}$, we can extend the sum in (3.16) and write

$$Q(u_{h,k}^n) = \frac{k}{2} \sum_{\sigma\in\mathcal{E}} (D_\sigma p_\sigma^\star) \cdot p_\sigma 2m(\chi_\sigma) = k(D_h p_h^\star, p_h)_{L^2(\Omega)}$$

where $p_\sigma^\star = \frac{u_E - u_W}{h} \mathbf{n}_{W,\sigma}$ for $\sigma = \sigma_{WE}$ is the normal component of the gradient and $D_h, p_h, p_h^\star$ are piecewise constant functions with values extended from σ to χ_σ. Further, we use the following inequality:

$$(D_h p_h^\star, p_h)_{L^2(\Omega)} \geq (D_h p_h^\star, p_h^\star)_{L^2(\Omega)} - |(D_h p_h^\star, p_h - p_h^\star)_{L^2(\Omega)}|. \tag{3.17}$$

It is clear that $(D_h p_h^\star, p_h^\star)_{L^2(\Omega)} = \sum_{\sigma\in\mathcal{E}} \bar{\lambda}_\sigma \left(\frac{u_E - u_W}{h}\right)^2 m(\chi_\sigma)$, due to the fact that $u_E - u_W = 0$ for $\sigma \in \mathcal{E}_{ext}$ thanks to reflexion of $u_{h,k}$ in $\Omega_{\tilde{t}}$ (see the previous section). Applying Young's inequality in the second term on the right-hand side of (3.17) leads to

$$\begin{aligned} |(D_h p_h^\star, p_h - p_h^\star)_{L^2(\Omega)}| &= \left| \sum_{\sigma\in\mathcal{E}} \bar{\beta}_\sigma \frac{u_E - u_W}{h} \frac{u_N - u_S}{h} m(\chi_\sigma) \right| \\ &\leq \sum_{\sigma\in\mathcal{E}_{int}} \frac{1}{2} \left[\left(\frac{u_E - u_W}{h}\right)^2 + \left(\frac{\bar{\beta}_\sigma}{\bar{\lambda}_\sigma}\right)^2 \left(\frac{u_N - u_S}{h}\right)^2 \right] \bar{\lambda}_\sigma m(\chi_\sigma), \end{aligned} \tag{3.18}$$

since $\phi_\sigma^n(u_{h,k}^n) = 0$ for $\sigma \in \mathcal{E}_{ext}$. By using inequalities (3.10) we get

$$\begin{aligned} |(D_h p_h^\star, p_h - p_h^\star)_{L^2(\Omega)}| &\leq \frac{1+\gamma}{2} \sum_{\sigma\in\mathcal{E}_{int}} \bar{\lambda}_\sigma \left(\frac{u_E - u_W}{h}\right)^2 m(\chi_\sigma) \\ &= \frac{1+\gamma}{2} (D_h p_h^\star, p_h^\star)_{L^2(\Omega)}. \end{aligned} \tag{3.19}$$

Using (3.12), it in turn implies

$$Q(u_{h,k}^n) \geq \left(1 - \frac{1+\gamma}{2}\right) k (Dp_h^\star, p_h^\star)_{L^2(\Omega)} \geq \bar{\lambda}_{\min} \frac{1-\gamma}{2} \frac{k}{2} |u_{h,k}^n|_{1,\mathcal{T}_h}^2, \tag{3.20}$$

where $\bar{\lambda}_{\min} = \inf_{\sigma \in \mathcal{E}} \bar{\lambda}_\sigma \geq C > 0$. By applying (3.15), (3.16), and (3.20) in (3.14), we get for h sufficiently small and any $u_{h,k}^n \in \mathcal{P}_0(\mathcal{T}_h)$ that

$$\sum_{K \in \mathcal{T}_h} \mathcal{L}_h(u_{h,k}^n) u_K^n \geq \alpha \left(|u_{h,k}^n|_{1,\mathcal{T}_h}^2 + ||u_{h,k}^n||_{L^2(\Omega)}^2 \right),$$

with $\alpha = \min\left(\bar{\lambda}_{\min}(1-\gamma)\frac{k}{4}, 1\right)$.

THEOREM 3.4. *For h sufficiently small, there exists a unique solution $u_{h,k}^n$ given by the scheme* (2.10)–(2.11) *at any discrete time step t_n.*

Proof. Assume that u_K, $K \in \mathcal{T}_h$, satisfy the linear system (3.13), and let the right-hand side of (3.13) be equal to 0. Then

$$\alpha \left(|u_{h,k}^n|_{1,\mathcal{T}_h}^2 + ||u_{h,k}^n||_{L^2(\Omega)}^2 \right) \leq \sum_{K \in \mathcal{T}_h} \mathcal{L}_h(u_{h,k}^n) u_K^n = \sum_{K \in \mathcal{T}_h} f_{h,k}(u_{h,k}^{n-1}) u_K^n = 0. \tag{3.21}$$

Due to relation (3.21) and the strict positivity of α we know that $u_K^n = 0$ for all $K \in \mathcal{T}_h$. It means that the kernel of the linear transformation represented by the matrix of the system (3.13) contains only $\bar{0}$ vector, which implies that the matrix is regular. Thus it also implies that there exists a unique solution for any right-hand side. ∎

4. Convergence of the scheme to the weak solution.

DEFINITION 4.1. *The weak solution of the problem* (1.1)–(1.3) *is a function $u \in L^2(0,T;H^1(\Omega))$ which satisfies the identity*

$$\int_0^T \int_\Omega u \frac{\partial \varphi}{\partial t}(x,t) dx dt + \int_\Omega u_0(x)\varphi(x,0) dx - \int_0^T \int_\Omega (D\nabla u) \cdot \nabla \varphi dx dt = 0 \ \forall \varphi \in \Psi, \tag{4.1}$$

where $\Psi = \{\varphi \in C^{2,1}(\overline{\Omega} \times [0,T]), (D\nabla\varphi) \cdot \overrightarrow{n} = 0 \text{ on } \partial\Omega \times (0,T), \varphi(.,T) = 0\}$.

Remark 1. The existence and uniqueness of the weak solution and extremum principle for the model (1.1)–(1.3) are given in [24]. The proofs are based on theory built in [1].

In the proof of convergence we will use a strategy based on the application of Kolmogorov's compactness criterion in L^2 which gives the relative compactness of the approximate solutions given by the scheme refining the space and time discretization step. By using relative compactness we can choose a convergent subsequence which in the limit gives the weak solution. In order to use Kolmogorov's compactness criterion we shall prove the following four lemmas.

LEMMA 4.2 (uniform boundedness). *There exists a positive constant C such that*

$$\|u_{h,k}\|_{L^2(Q_T)} \leq C. \tag{4.2}$$

LEMMA 4.3 (time translate estimate). *For any $s \in (0,T)$ there exists a positive constant C such that*

$$\int_{\Omega \times (0,T-s)} \left(u_{h,k}(x,t+s) - u_{h,k}(x,t)\right)^2 dx dt \leq Cs. \tag{4.3}$$

LEMMA 4.4 (space translate estimate I). *There exists a positive constant C such that*

$$\int_{\Omega_\xi \times (0,T)} \left(u_{h,k}(x+\xi,t) - u_{h,k}(x,t)\right)^2 dx dt \leq C|\xi|(|\xi| + 2h) \tag{4.4}$$

for any vector $\xi \in R^d$, where $\Omega_\xi = \{x \in \Omega, [x, x+\xi] \in \Omega\}$.

LEMMA 4.5 (space translate estimate II). *There exists a positive constant C such that*

$$\int_{\Omega\times(0,T)} \left(u_{h,k}\left(x+\xi,t\right)-u_{h,k}\left(x,t\right)\right)^2 dxdt \le C\left|\xi\right| \tag{4.5}$$

for any vector $\xi \in R^d$.

To prove (4.2)–(4.5) we will use the following a priori estimates.

LEMMA 4.6. *The scheme* (2.10)–(2.11) *leads to the following estimates. For h sufficiently small, there exists a positive constant C which does not depend on h,k such that*

$$\max_{0\le n\le N_{\max}} \sum_{K\in\mathcal{T}_h} \left(u_K^n\right)^2 m(K) \le C, \tag{4.6}$$

$$\sum_{n=1}^{N_{\max}} k \sum_{(K,L)\in\Upsilon} \frac{\left(u_K^n-u_L^n\right)^2}{d_{K,L}} m\left(\sigma\right) \le C, \tag{4.7}$$

$$\sum_{n=1}^{N_{\max}} \sum_{K\in\mathcal{T}_h} \left(u_K^n-u_K^{n-1}\right)^2 m\left(K\right) \le C. \tag{4.8}$$

Proof. We multiply (2.10) by u_K^n, sum it over $K\in\mathcal{T}_h$ and over $n=1,\dots,m<N_{max}$, and use the property $(a-b)a=\frac{1}{2}a^2-\frac{1}{2}b^2+\frac{1}{2}(a-b)^2$ to obtain

$$\begin{aligned}&\frac{1}{2}\sum_{K\in\mathcal{T}_h}(u_K^m)^2m(K)+\frac{1}{2}\sum_{n=1}^{m}\sum_{K\in\mathcal{T}_h}(u_K^n-u_K^{n-1})^2m(K)\\ &-\sum_{n=1}^{m}k\sum_{K\in\mathcal{T}_h}u_K^n\sum_{\sigma\in\mathcal{E}_K\cap\mathcal{E}_{int}}\phi_\sigma^n(u_{h,k}^n)m(\sigma)=\frac{1}{2}\sum_{K\in\mathcal{T}_h}(u_K^0)^2m(K).\end{aligned} \tag{4.9}$$

Then by using (3.16) and (3.20) we have

$$\begin{aligned}&\frac{1}{2}\sum_{K\in\mathcal{T}_h}(u_K^m)^2m(K)+\frac{1}{2}\sum_{n=1}^{m}\sum_{K\in\mathcal{T}_h}(u_K^n-u_K^{n-1})^2m(K)\\ &+\bar{\alpha}\sum_{n=1}^{m}k|u_{h,k}^n|_{1,\mathcal{T}_h}^2\le\frac{1}{2}\sum_{K\in\mathcal{T}_h}(u_K^0)^2m(K),\end{aligned} \tag{4.10}$$

with positive constant $\bar{\alpha}=\bar{\lambda}_{\min}\frac{1-\gamma}{4}$. Since $u_0\in L^2(\Omega)$, the right-hand side is bounded by a positive constant C. By using the first term of (4.10) we get a priori estimate (4.6), and from the second term of (4.10) we get a priori estimate (4.8). From the strict positiveness of $\bar{\alpha}$ in the third term of (4.10) and from definition (3.11), we obtain a priori estimate (4.7). ∎

Proof of Lemma 4.2. It follows from the first $L^2(\Omega)$—a priori estimate (4.6). ∎

Proof of Lemma 4.3. First, for fixed $s\in(0,T)$, we define function

$$f(t)=\int_\Omega \left(u_{h,k}\left(x,t+s\right)-u_{h,k}\left(x,t\right)\right)^2 dx.$$

By using the fact that $u_{h,k}$ is a piecewise constant function, we get

$$f(t)=\sum_{K\in\mathcal{T}_h}\left(u_K^{n_{t+s}}-u_K^{n_t}\right)^2 m\left(K\right), \tag{4.11}$$

with $n_t = \lceil \frac{t}{k} \rceil$ and $n_{t+s} = \lceil \frac{t+s}{k} \rceil$, where $\lceil \cdot \rceil$ means the upper integer part of a positive real number. We rearrange (4.11) to obtain

$$f(t) = \sum_{K\in\mathcal{T}_h} \left(u_K^{n_{t+s}} - u_K^{n_t}\right) \sum_{t\le(n+1)k<t+s} \left(u_K^n - u_K^{n-1}\right) m\left(K\right). \tag{4.12}$$

By using the scheme (2.10)–(2.11) in (4.12) (replacing K by W) we get

$$f(t) = \sum_{t\le(n+1)k<t+s} k \sum_{W\in\mathcal{T}_h} \left(\left(u_W^{n_{t+s}} - u_W^{n_t}\right) \sum_{\sigma\in\mathcal{E}_W\cap\mathcal{E}_{int}} \bar{\lambda}_\sigma (u_E^n - u_W^n) + \bar{\beta}_\sigma (u_N^n - u_S^n) \right), \tag{4.13}$$

and due to conservativity of numerical fluxes (antisymmetry of term $\bar{\lambda}_\sigma(u_E^n - u_W^n) + \bar{\beta}_\sigma(u_N^n - u_S^n)$) we have

$$\begin{aligned} f(t) = \sum_{t\le(n+1)k<t+s} \frac{k}{2} \sum_{\sigma\in\mathcal{E}_{int}} & \left(u_W^{n_{t+s}} - u_W^{n_t} - u_E^{n_{t+s}} + u_E^{n_t}\right) \\ & \left(\bar{\lambda}_\sigma(u_E^n - u_W^n) + \bar{\beta}_\sigma(u_N^n - u_S^n)\right). \end{aligned} \tag{4.14}$$

Using Young's inequality leads to the relation

$$\begin{aligned} f(t) \le & \sum_{t\le(n+1)k<t+s} \frac{k}{4} \sum_{\sigma\in\mathcal{E}_{int}} \bar{\lambda}_\sigma (u_W^{n_{t+s}} - u_W^{n_t} - u_E^{n_{t+s}} + u_E^{n_t})^2 \\ & + \sum_{t\le(n+1)k<t+s} \frac{k}{4} \sum_{\sigma\in\mathcal{E}_{int}} \bar{\lambda}_\sigma \left((u_E^n - u_W^n) + \frac{\bar{\beta}_\sigma}{\bar{\lambda}_\sigma} (u_N^n - u_S^n) \right)^2, \end{aligned} \tag{4.15}$$

where the right-hand side can be further estimated and we get

$$f(t) \le f_1(t) + f_2(t) + f_3(t) + f_4(t), \tag{4.16}$$

$$f_1(t) = \sum_{t\le(n+1)k<t+s} \frac{k}{2} \sum_{\sigma\in\mathcal{E}_{int}} \bar{\lambda}_\sigma \left(u_E^{n_t} - u_W^{n_t}\right)^2, \tag{4.17}$$

$$f_2(t) = \sum_{t\le(n+1)k<t+s} \frac{k}{2} \sum_{\sigma\in\mathcal{E}_{int}} \bar{\lambda}_\sigma \left(u_E^{n_{t+s}} - u_W^{n_{t+s}}\right)^2, \tag{4.18}$$

$$f_3(t) = \sum_{t\le(n+1)k<t+s} \frac{k}{2} \sum_{\sigma\in\mathcal{E}_{int}} \bar{\lambda}_\sigma (u_E^n - u_W^n)^2, \tag{4.19}$$

$$f_4(t) = \sum_{t\le(n+1)k<t+s} \frac{k}{2} \sum_{\sigma\in\mathcal{E}_{int}} \bar{\lambda}_\sigma \left(\frac{\bar{\beta}_\sigma}{\bar{\lambda}_\sigma}\right)^2 (u_N^n - u_S^n)^2. \tag{4.20}$$

Next we integrate (4.16) in time interval $(0, T-s)$, and by replacing $\sum_{\sigma\in\mathcal{E}_{int}}$ by $\sum_{(K,L)\in\Upsilon}$ (the edge $\sigma \in \mathcal{E}_K$ is an intersection of K and its adjacent finite volume L) we get an estimate of the first integral term

$$\int_0^{T-s} f_1(t)dt = \int_0^{T-s} \frac{k}{2} \sum_{(K,L)\in\Upsilon} \bar{\lambda}_\sigma \left(u_L^{n_t} - u_K^{n_t}\right)^2 \sum_{n\in\mathbf{N}} \chi_{\{t\le(n+1)k<t+s\}} dt. \tag{4.21}$$

We substitute the integral over $(0, T-s)$ by the sum of time step intervals and use the property $\chi_{\{t\le(n+1)k<t+s\}} = \chi_{\{(n+1)k-s<t\le(n+1)k\}}$ to obtain

$$\int_0^{T-s} f_1(t)dt \le \sum_{n_t=0}^{N_{\max}-1} \frac{k}{2} \sum_{(K,L)\in\Upsilon} \bar{\lambda}_\sigma \left(u_L^{n_t} - u_K^{n_t}\right)^2 \int_{n_t k}^{(n_t+1)k} \sum_{n\in\mathbf{N^o}} \chi_{\{(n+1)k-s<t\le(n+1)k\}} dt. \tag{4.22}$$

Since $\int_{n_t k}^{(n_t+1)k} \sum_{n\in\mathbf{N^o}} \chi_{\{(n+1)k-s<t\le(n+1)k\}} dt = s$, and $m(\sigma) = d_{K,L}$ for our uniform mesh, the relation (4.22) leads to

$$\int_0^{T-s} f_1(t)dt \le s \sum_{n_t=0}^{N_{\max}} \frac{k}{2} \sum_{(K,L)\in\Upsilon} \frac{m(\sigma)}{d_{K,L}} \bar{\lambda}_\sigma \left(u_L^{n_t} - u_K^{n_t}\right)^2. \tag{4.23}$$

The next step is to prove the following relation:

$$0 < C_1 \le \bar{\lambda}_\sigma \le C_2 < \infty \ \text{ for all } \ \sigma \in \mathcal{E}. \tag{4.24}$$

Let K be any fixed finite volume. Since at any time step the matrix $D_\sigma = \left(\begin{smallmatrix} \bar{\lambda}_\sigma & \bar{\beta}_\sigma \\ \bar{\beta}_\sigma & \bar{\nu}_\sigma \end{smallmatrix}\right)$ is uniformly (strictly) positive definite, $\bar{\lambda}_\sigma \ge C_3 > 0$ and $\bar{\nu}_\sigma \ge C_4 > 0$ for all σ. The structure tensor evaluated numerically at point x_{KL} is given by

$$J_\rho\left(\nabla u_{h,k}^{n-1}{}_{\tilde{t}}\right)(x_{KL}) = G_\rho * \begin{pmatrix} A & B \\ B & C \end{pmatrix}, \tag{4.25}$$

where

$$A = \left(\left(\frac{\partial G_{\tilde{t}}}{\partial x} * \tilde{u}_{h,k}^{n-1}\right)(x_{KL})\right)^2, \quad C = \left(\left(\frac{\partial G_{\tilde{t}}}{\partial y} * \tilde{u}_{h,k}^{n-1}\right)(x_{KL})\right)^2, \tag{4.26}$$

$$B = \left(\frac{\partial G_{\tilde{t}}}{\partial x} * \tilde{u}_{h,k}^{n-1}\right)(x_{KL}) \left(\frac{\partial G_{\tilde{t}}}{\partial y} * \tilde{u}_{h,k}^{n-1}\right)(x_{KL}). \tag{4.27}$$

By using Young's inequality, a priori estimate (4.6), and the definition of extension $\tilde{u}_{h,k}^n$ (see (2.4)), we subsequently get for $i = 1, 2$ ($x_1 = x, x_2 = y$)

$$\begin{aligned} \left|\left(\frac{\partial}{\partial x_i} G_{\tilde{t}} * \tilde{u}_{h,k}^n\right)(x_{KL})\right| &\le \int_{R^d} \left|\frac{\partial}{\partial x_i} G_{\tilde{t}}(x_{KL} - \xi)\tilde{u}_{h,k}^n(\xi)\right| d\xi \\ &\le \frac{1}{2}\int_{R^d} \left|\frac{\partial}{\partial x_i} G_{\tilde{t}}(x_{KL} - \xi)\right|^2 d\xi + \frac{1}{2}\int_{R^d} |\tilde{u}_{h,k}^n(\xi)|^2 d\xi \le C_{\tilde{t}} \\ &\quad + C_5 \int_{\Omega_{\tilde{t}}} |\tilde{u}_{h,k}^n(\xi)|^2 d\xi \le C_{\tilde{t}} + C_5 \sum_{K\in\mathcal{T}_h} \left(u_K^n\right)^2 m(K) \le C_6. \end{aligned} \tag{4.28}$$

By inspecting relations (1.6)–(1.10) we may observe that if the elements of matrix J_ρ are finite, then also the elements of matrix D_σ are finite, which gives (4.24). By applying (4.24) and (4.7) in (4.23), we get $\int_0^{T-s} f_1(t)dt \le Cs$. By using similar approach as in [15] and relation (3.10), all further integrals can be estimated in the same way, which ends the proof. □

Proof of Lemma 4.4. Let us define $\xi_{K,L} = \frac{\xi}{|\xi|}.n_{K,\sigma}$ for all $(K, L) \in \Upsilon$ and let for all $x \in \Omega_\xi$

$$E(x, K, L) = \begin{cases} 1 & \text{if } [x, x+\xi] \text{ intersects } \sigma = \sigma_{KL},\ K \text{ and } L;\ \text{and } \xi_{K,L} > 0 \\ 0 & \text{otherwise.} \end{cases}$$

For any $t \in (0,T)$ there exists $n \in N$ which satisfies $(n-1)k < t \leq nk$. Then for almost all $x \in \Omega_\xi$ we can see that

$$u_{h,k}(x+\xi,t) - u_{h,k}(x,t) = u^n_{K(x+\xi)} - u^n_{K(x)} = \sum_{(K,L)\in\Upsilon} E(x,K,L)\,(u^n_L - u^n_K),$$

where $K(x)$ denotes the volume $K \in \mathcal{T}_h$, where $x \in K$. By using these notations we get the proof in similar lines as in [15]—proof of Lemma 3.2. □

Proof of Lemma 4.5. In this proof, for simplicity, let us consider that $\Omega_{\tilde{t}} = \Omega$; i.e., we extend $u_{h,k}$ outside Ω by 0. The results which are obtained below can be straightforwardly adjusted to the situation with reflexion to $\Omega_{\tilde{t}}$; the derivation is just technically more complicated, and for details we refer to [10]. Let us define the set

$$\mathcal{E}_{ext} = \{\varpi, \text{ such that there exists } K \in \mathcal{T}_h,\ \varpi \subset \partial K \cap \partial\Omega\}$$

and let $u_\varpi := u_K$, where $K \in \mathcal{T}_h,\ \varpi \subset \partial K \cap \partial\Omega$. Since now for $x \in \Omega - \Omega_\xi$ the point $x+\xi$ can be outside Ω, we see that

$$\begin{aligned} &u_{h,k}(x+\xi,t) - u_{h,k}(x,t) \\ &= \sum_{(K,L)\in\Upsilon} E(x,K,L)\,(u^n_L - u^n_K) - \sum_{\varpi\in\mathcal{E}_{ext}} \chi([x,x+\xi]\cap\varpi)\,u^n_\varpi. \end{aligned} \tag{4.29}$$

By using the Cauchy–Schwarz and Young inequalities we obtain

$$\begin{aligned} &(u_{h,k}(x+\xi,t) - u_{h,k}(x,t))^2 \\ &\leq 2\left(\sum_{(K,L)\in\Upsilon} E(x,p,q)\,\xi_{K,L}d_{K,L}\right)\left(\sum_{(K,L)\in\Upsilon} E(x,K,L)\,\frac{(u^n_L - u^n_K)^2}{\xi_{K,L}d_{K,L}}\right) \\ &\quad + 2\sum_{\varpi\in\mathcal{E}_{ext}} \chi([x,x+\xi]\cap\varpi)\,(u^n_\varpi)^2, \end{aligned} \tag{4.30}$$

$$\begin{aligned} &\int_{\Omega\times(0,T)} (u_{h,k}(x+\xi,t) - u_{h,k}(x,t))^2\,dxdt \\ &\leq (2h+|\xi|)\,|\xi|\,C + 2\sum_{n=0}^{N_{\max}} k\int_\Omega \sum_{\varpi\in\mathcal{E}_{ext}} \chi([x,x+\xi]\cap\varpi)\,(u^n_\varpi)^2\,dxdt, \end{aligned} \tag{4.31}$$

which can be written as

$$\begin{aligned} &\int_{\Omega\times(0,T)} (u_{h,k}(x+\xi,t) - u_{h,k}(x,t))^2\,dxdt \\ &\leq (2h+|\xi|)\,|\xi|\,C + 2\,|\xi| \sum_{n=0}^{N_{\max}} k \sum_{\varpi\in\mathcal{E}_{ext}} (u^n_\varpi)^2\,m(\varpi). \end{aligned} \tag{4.32}$$

For the last term in (4.32) we use the trace inequality given in [4].

LEMMA 4.7. *Let Ω be an open bounded polygonal connected subset of R^d. Let $\overline{\gamma}(u_{h,k})$ be defined by $\overline{\gamma}(u_{h,k}) = u_\varpi$ a.e. for the $(d-1)-$Lebesgue measure on $\varpi \in \mathcal{E}_{ext}$. Then there exists positive C depending only on Ω such that*

$$\left\|\overline{\gamma}\left(u^n_{h,k}\right)\right\|_{L^2(\partial\Omega)} \leq C\left(\left|u^n_{h,k}\right|_{1,\mathcal{T}_h} + \left\|u^n_{h,k}\right\|_{L^2(\Omega)}\right).$$

By using the trace operator $\overline{\gamma}(u_{h,k}) = u_{\varpi}$ we can write

$$\begin{aligned}&\int_{\Omega\times(0,T)} \left(u_{h,k}(x+\xi,t) - u_{h,k}(x,t)\right)^2 dxdt\\ &\leq (2h+|\xi|)\,|\xi|\,C + 2\,|\xi| \sum_{n=0}^{N_{\max}} k \left\|\overline{\gamma}\left(u_{h,k}^n\right)\right\|^2_{L^2(\partial\Omega)} \end{aligned} \tag{4.33}$$

and applying the trace inequality implies that

$$\begin{aligned}&\int_{\Omega\times(0,T)} \left(u_{h,k}(x+\xi,t) - u_{h,k}(x,t)\right)^2 dxdt\\ &\leq (2h+|\xi|)\,|\xi|\,C + 4C\,|\xi| \sum_{n=0}^{N_{\max}} k\Big(\left|u_{h,k}^n\right|^2_{1,\mathcal{T}_h} + \left\|u_{h,k}^n\right\|^2_{L^2(\Omega)}\Big). \end{aligned} \tag{4.34}$$

Then using a priori estimates (4.6) and (4.7) ends the proof. □

Lemmas 4.2, 4.5, and 4.3 guarantee that sequence $u_{h,k}$ is relatively compact in $L^2(Q_T)$, which implies following convergence result.

LEMMA 4.8. *There exists $u \in L^2(Q_T)$ such that $u_{h,k} \to u$ in $L^2(Q_T)$ as $h,k \to 0$ in the sense of subsequences.*

For the sake of simplicity, we denote the subsequence converging to u again by $u_{h,k}$, and we are going to prove that its limit u is the weak solution of (1.1)–(1.3) in the sense of Definition 4.1.

To that goal, let $\varphi \in \Psi$ be given, and multiply the scheme (2.10) by $\varphi(x_K, t_n)$. Then we sum it over all $K \in \mathcal{T}_h$ and for $n = 1, \ldots, N_{\max}$ to get

$$\begin{aligned}&\sum_{n=1}^{N_{\max}} k \sum_{K\in\mathcal{T}_h} \frac{\left(u_K^n - u_K^{n-1}\right)}{k} \varphi(x_K, t_{n-1})\, m(K)\\ &= \sum_{n=1}^{N_{\max}} k \sum_{K\in\mathcal{T}_h} \varphi(x_K, t_{n-1}) \sum_{\sigma\in\mathcal{E}_K\cap\mathcal{E}_{int}} \phi^n_{\sigma_{KL}}(u^n_{h,k}) m(\sigma). \end{aligned} \tag{4.35}$$

In order to have a structure similar to the weak solution identity (4.1), we rearrange (4.35) by using a discrete integration by parts and gathering the sums over $K \in \mathcal{T}_h$ and $\sigma \in \mathcal{E}_K \cap \mathcal{E}_{int}$, and we get

$$\begin{aligned}&\sum_{n=1}^{N_{\max}} k \sum_{K\in\mathcal{T}_h} u_K^n \frac{\varphi(x_K,t_n) - \varphi(x_K,t_{n-1})}{k} m(K) + \sum_{K\in\mathcal{T}_h} u_K^0 \varphi(x_K,0)\, m(K)\\ &-\frac{1}{2}\sum_{n=1}^{N_{\max}} k \sum_{\sigma\in\mathcal{E}_{int}} \phi^n_\sigma(u^n_{h,k}) m(\sigma)\left(\varphi(x_L,t_{n-1}) - \varphi(x_K,t_{n-1})\right) = 0. \end{aligned} \tag{4.36}$$

In the same way as in [15] we can prove that

$$\sum_{n=1}^{N_{\max}} k \sum_{K\in\mathcal{T}_h} u_K^n \frac{\varphi(x_K,t_n) - \varphi(x_K,t_{n-1})}{k} m(K) \to \int_0^T \int_\Omega u \frac{\partial\varphi}{\partial t}(x,t)\, dxdt, \tag{4.37}$$

$$\sum_{K\in\mathcal{T}_h} u_K^0 \varphi(x_K,0)\, m(K) \to \int_\Omega u_0(x)\,\varphi(x,0)\, dx \tag{4.38}$$

as $h, k \to 0$ for all $\varphi \in \Psi$. The main point in proving convergence of the scheme is to get that

$$\frac{1}{2}\sum_{n=1}^{N_{\max}} k \sum_{\sigma\in\mathcal{E}_{int}} \phi_\sigma^n(u_{h,k}^n) m(\sigma)\left(\varphi\left(x_L, t_{n-1}\right) - \varphi\left(x_K, t_{n-1}\right)\right)$$
$$\to \int_0^T \int_\Omega -\nabla\cdot\left(D\nabla\varphi\right) u dx dt \tag{4.39}$$

as $h, k \to 0$ for all $\varphi \in \Psi$. By using then the space translate estimate (4.4) we know (see, e.g., [4] or [14]) that the limit function u is in the space $L^2((0,T), H^1)$, so we can use Green's theorem, and by applying the boundary conditions we have

$$\int_0^T \int_\Omega -\nabla\cdot\left(D\nabla\varphi\right) u dx dt = \int_0^T \int_\Omega \left(D\nabla\varphi\right)\cdot\nabla u dx dt.$$

Proving (4.39) thus leads to overall convergence of the scheme to the weak solution in the sense of (4.1). To deal with (4.39) we rewrite it and then split into the sum of five terms

$$\frac{1}{2}\sum_{n=1}^{N_{\max}} k \sum_{(W,E)\in\Upsilon} \phi_\sigma^n(u_{h,k}^n) m(\sigma)\left(\varphi\left(x_E, t_{n-1}\right) - \varphi\left(x_W, t_{n-1}\right)\right)$$
$$+ \int_0^T \int_\Omega \nabla\cdot\left(D\nabla\varphi(x,t)\right) u(x,t) dx dt = \sum_{i=1}^{5} T_i, \tag{4.40}$$

where

$$T_1 = \frac{1}{2}\sum_{n=1}^{N_{\max}} k \sum_{(W,E)\in\Upsilon} \big[(D_\sigma p_\sigma(u))\cdot\mathbf{n}_{W,\sigma}(\varphi_E^{n-1} - \varphi_W^{n-1})$$
$$- (u_E^n - u_W^n)(D_\sigma p_\sigma(\varphi^{n-1}))\cdot\mathbf{n}_{W,\sigma}\big] m(\sigma),$$
$$T_2 = \frac{1}{2}\sum_{n=1}^{N_{\max}} k \sum_{(W,E)\in\Upsilon} (u_E^n - u_W^n) m(\sigma)\big[(D_\sigma p_\sigma(\varphi^{n-1}))$$
$$\cdot\mathbf{n}_{W,\sigma} - (D_\sigma\nabla\varphi(x_{WE}, t_{n-1}))\cdot\mathbf{n}_{W,\sigma}\big],$$
$$T_3 = \frac{1}{2}\sum_{n=1}^{N_{\max}} \sum_{(W,E)\in\Upsilon} (u_E^n - u_W^n)\left(m(\sigma) k (D_\sigma\nabla\varphi(x_{WE}, t_{n-1}))\right.$$
$$\left.\cdot\mathbf{n}_{W,\sigma} - \int_{t_{n-1}}^{t_n}\int_\sigma (D_\sigma\nabla\varphi(s,t))\cdot\mathbf{n}_{W,\sigma} ds dt\right),$$
$$T_4 = \frac{1}{2}\sum_{n=1}^{N_{\max}} \sum_{(W,E)\in\Upsilon} (u_E^n - u_W^n)\int_{t_{n-1}}^{t_n}\int_\sigma ((D_\sigma - D)\nabla\varphi(s,t))\cdot\mathbf{n}_{W,\sigma} ds dt,$$
$$T_5 = \int_0^T \int_\Omega \nabla\cdot\left(D\nabla\varphi(x,t)\right)\left(u(x,t) - u_{h,k}(x,t)\right) dx dt,$$

where $\varphi_W^{n-1} = \varphi(x_W, t_{n-1})$, $\varphi_E^{n-1} = \varphi(x_E, t_{n-1})$, and $\varphi^{n-1} = \varphi(x, t_{n-1})$. Since

$$\begin{aligned}
&\frac{1}{2}\sum_{n=1}^{N_{\max}}\sum_{(W,E)\in\Upsilon}(u_E^n - u_W^n)\int_{t_{n-1}}^{t_n}\int_\sigma (D\nabla\varphi(s,t))\cdot\mathbf{n}_{W,\sigma}dsdt\\
&= -\sum_{n=1}^{N_{\max}}\sum_{W\in\mathcal{T}_h} u_W^n\int_{t_{n-1}}^{t_n}\sum_{\sigma\in\mathcal{E}_W}\int_\sigma (D\nabla\varphi(s,t))\cdot\mathbf{n}_{W,\sigma}dsdt\\
&= -\sum_{n=1}^{N_{\max}}\sum_{W\in\mathcal{T}_h} u_W^n\int_{t_{n-1}}^{t_n}\int_W \nabla\cdot(D\nabla\varphi(x,t))dxdt\\
&= -\int_0^T\int_\Omega \nabla\cdot(D\nabla\varphi(x,t))u_{h,k}(x,t)dxdt,
\end{aligned}$$

one can see correspondence of terms in T_4 and T_5. First, let us deal with T_1 and rewrite it using (2.7)–(2.8) into the form

$$\begin{aligned}
T_1 = \frac{1}{2}\sum_{n=1}^{N_{\max}} k \sum_{(W,E)\in\Upsilon} \big([\bar\lambda_\sigma(u_E^n - u_W^n) + \bar\beta_\sigma(u_N^n - u_S^n)](\varphi_E^{n-1} - \varphi_W^{n-1})\\
- (u_E^n - u_W^n)[\bar\lambda_\sigma(\varphi_E^{n-1} - \varphi_W^{n-1}) + \bar\beta_\sigma(\varphi_N^{n-1} - \varphi_S^{n-1})]\big),
\end{aligned}$$

which can be easily simplified to

$$T_1 = \frac{1}{2}\sum_{n=1}^{N_{\max}} k \sum_{(W,E)\in\Upsilon}[\bar\beta_\sigma(u_N^n - u_S^n)(\varphi_E^{n-1} - \varphi_W^{n-1}) - \bar\beta_\sigma(u_E^n - u_W^n)(\varphi_N^{n-1} - \varphi_S^{n-1})]. \tag{4.41}$$

By applying (3.1) for u and similarly for φ we get

$$\begin{aligned}
T_1 = \frac{1}{2}\sum_{n=1}^{N_{\max}} k \sum_{(W,E)\in\Upsilon}\sum_{i=1}^{4}\Bigg[&\frac{\bar\beta_\sigma}{4}(\varphi_{E_\sigma} - \varphi_{W_\sigma})(u_{E_{\delta_i}} - u_{W_{\delta_i}})\\
&-\frac{\bar\beta_\sigma}{4}(u_{E_\sigma} - u_{W_\sigma})(\varphi_{E_{\delta_i}} - \varphi_{W_{\delta_i}})(1+O(h))\Bigg],
\end{aligned} \tag{4.42}$$

where we omit time indexes due to simplification; for graphical explanation of notations see Figures 3.1 and 4.1. For each term with positive sign in (4.42) one can find a corresponding term in the group of terms with negative signs ($\bar\beta_\sigma$ corresponds to some $\bar\beta_\delta$). We denote these couples by $T_{\sigma\delta}$. For example, for $\sigma = \sigma_{WE}$ and δ as plotted in Figure 4.1 (left) we can write the couple as follows:

$$\begin{aligned}
T_{\sigma_{WE}\delta} &= \frac{\bar\beta_{\sigma_{WE}}}{4}(\varphi^{n-1}_{E_{\sigma_{WE}}} - \varphi^{n-1}_{W_{\sigma_{WE}}})(u^n_{E_\delta} - u^n_{W_\delta})\\
&- \frac{\bar\beta_\delta}{4}(\varphi^{n-1}_{E_{\sigma_{EW}}} - \varphi^{n-1}_{W_{\sigma_{EW}}})(u^n_{E_\delta} - u^n_{W_\delta})(1+O(h))\\
= \frac{\bar\beta_{\sigma_{WE}}}{4}&(\varphi^{n-1}_{E_{\sigma_{WE}}} - \varphi^{n-1}_{W_{\sigma_{WE}}})(u^n_{E_\delta} v - u^n_{W_\delta}) - \frac{\bar\beta_\delta}{4}(\varphi^{n-1}_{W_{\sigma_{WE}}} - \varphi^{n-1}_{E_{\sigma_{WE}}})(u^n_{E_\delta} - u^n_{W_\delta})(1+O(h))
\end{aligned}$$

because $\bar\beta_{\sigma_{WE}} = \bar\beta_{\sigma_{EW}}$, $\varphi^{n-1}_{E_{\sigma_{WE}}} = \varphi^{n-1}_{W_{\sigma_{EW}}}$, and $\varphi^{n-1}_{W_{\sigma_{WE}}} = \varphi^{n-1}_{E_{\sigma_{EW}}}$; see Figure 4.1 (right). Using the previous expression for every $\sigma = \sigma_{WE}$ yields

$$\begin{aligned}
T_{\sigma\delta} &= \left[\frac{\bar\beta_\sigma}{4} + \left(-\frac{\bar\beta_\sigma}{4}(1+O(h))\right)\right](\varphi^{n-1}_{E_\sigma} - \varphi^{n-1}_{W_\sigma})(u^n_{E_\delta} - u^n_{W_\delta})\\
&= \frac{\bar\beta_\sigma}{4}O(h)(\varphi^{n-1}_{E_\sigma} - \varphi^{n-1}_{W_\sigma})(u^n_{E_\delta} - u^n_{W_\delta}).
\end{aligned} \tag{4.43}$$

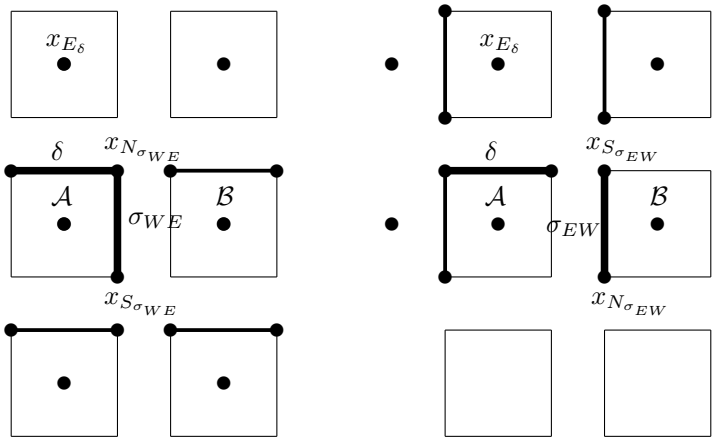


FIG. 4.1. *Left: The edge $\sigma = \sigma_{WE}$ and one $\delta \in P_\sigma$. Right: The same edge δ and one corresponding $\sigma = \sigma_{EW} \in P_\delta$. The thickest lines represent one particular couple $T_{\sigma\delta}$ in* (4.42)*, which vanishes up to the $O(h)$ term. $\mathcal{A} = x_{W_{\sigma_{WE}}} = x_{E_{\sigma_{EW}}} = x_{W_\delta}$, $\mathcal{B} = x_{E_{\sigma_{WE}}} = x_{W_{\sigma_{EW}}}$.*

Then T_1 can be estimated as follows:

$$|T_1| \le C_1 h \left| \sum_{n=1}^{N_{\max}} k \sum_{\sigma\in\mathcal{E}_{int}} \sum_{\delta\in P_\sigma\cap\mathcal{E}_{int}} (\varphi_{E_\sigma}^{n-1} - \varphi_{W_\sigma}^{n-1})(u_{E_\delta}^n - u_{W_\delta}^n) \right|,$$

with positive constant C_1 due to the fact that $\bar{\beta}_\sigma$ for each edge of mesh is finite (see (4.24)). By the Cauchy–Schwarz inequality we have

$$|T_1| \le C_2 h \left(\sum_{n=1}^{N_{\max}} k \sum_{\sigma\in\mathcal{E}_{int}} \sum_{\delta\in P_\sigma\cap\mathcal{E}_{int}} (\varphi_{E_\sigma}^{n-1} - \varphi_{W_\sigma}^{n-1})^2 \right)^{\frac{1}{2}} \left(\sum_{n=1}^{N_{\max}} k \sum_{\sigma\in\mathcal{E}_{int}} \sum_{\delta\in P_\sigma\cap\mathcal{E}_{int}} (u_{E_\delta}^n - u_{W_\delta}^n)^2 \right)^{\frac{1}{2}}. \tag{4.44}$$

It comes from the regularity of φ that there exists a positive constant C_3 such that $(\varphi_{E_\sigma}^{n-1} - \varphi_{W_\sigma}^{n-1})^2 \le C_3 h^2$. Thanks to geometrical arguments, we know that

$$\sum_{\sigma\in\mathcal{E}_{int}} d_{WE} m(\sigma) \le C_4 |\Omega|, \tag{4.45}$$

which straightforwardly gives for our uniform square mesh $\sum_{\sigma\in\mathcal{E}_{int}} h^2 \le C_4|\Omega|$. The above-mentioned facts lead to

$$\left(\sum_{n=1}^{N_{\max}} k \sum_{\sigma\in\mathcal{E}_{int}} \sum_{\delta\in P_\sigma\cap\mathcal{E}_{int}} (\varphi_{E_\sigma}^{n-1} - \varphi_{W_\sigma}^{n-1})^2 \right) \le C_5 T|\Omega|,$$

which together with a priori estimate (4.7) gives that $|T_1| \le C_6 h$, which implies

$$|T_1| \to 0 \text{ as } h, k \to 0. \tag{4.46}$$

Term T_2 can be written as $T_2 = \frac{1}{2}\sum_{n=1}^{N_{\max}} k \sum_{(W,E)\in\Upsilon}(u_E^n - u_W^n)m(\sigma)T_{2WE}$, where

$$T_{2WE} = \bar{\lambda}_\sigma \frac{\varphi_E^{n-1} - \varphi_W^{n-1}}{h} + \bar{\beta}_\sigma \frac{\varphi_N^{n-1} - \varphi_S^{n-1}}{h} - \bar{\lambda}_\sigma \left(\varphi_{WE}^{n-1}\right)_x - \bar{\beta}_\sigma \left(\varphi_{WE}^{n-1}\right)_y$$

and $\binom{(\varphi_{WE}^{n-1})_x}{(\varphi_{WE}^{n-1})_y} = \nabla\varphi(x_{WE}, t_{n-1})$ in the basis $(\mathbf{n}_{W,\sigma}, \mathbf{t}_{W,\sigma})$. Since $\varphi \in C^{2,1}\left(\overline{\Omega}\times[0,T]\right)$, there exist positive constants C_7 and C_8 such that

$$\left|\frac{\varphi_E^{n-1} - \varphi_W^{n-1}}{h} - \left(\varphi_{WE}^{n-1}\right)_x\right| \le C_7 h, \quad \left|\frac{\varphi_N^{n-1} - \varphi_S^{n-1}}{h} - \left(\varphi_{WE}^{n-1}\right)_y\right| \le C_8 h.$$

From there and the property that elements of D_σ are finite (see (4.28)) we have $|T_{2WE}| \le C_9 h$, with a positive constant C_9, which implies

$$|T_2| \le C_{10} h \left| \sum_{n=1}^{N_{\max}} k \sum_{(W,E)\in\Upsilon} (u_E^n - u_W^n) m(\sigma) \right|. \tag{4.47}$$

Then we multiply the right-hand side of (4.47) by $\frac{\sqrt{d_{WE}}}{\sqrt{d_{WE}}}$ and apply the Cauchy–Schwarz inequality to obtain

$$|T_2| \le C_{10} h \left(\sum_{n=1}^{N_{\max}} k \sum_{(W,E)\in\Upsilon} (u_E^n - u_W^n)^2 \frac{m(\sigma)}{d_{WE}} \right)^{\frac{1}{2}} \left(\sum_{n=1}^{N_{\max}} k \sum_{(W,E)\in\Upsilon} m(\sigma) d_{WE} \right)^{\frac{1}{2}}.$$

A priori estimate (4.7) together with (4.45) gives $|T_2| \le C_{10}(C_{11}C_4|\Omega|T)^{\frac{1}{2}} h$ and finally

$$|T_2| \to 0 \text{ as } h, k \to 0. \tag{4.48}$$

We consider the third term in the form $T_3 = \frac{1}{2}\sum_{n=1}^{N_{\max}} \sum_{(W,E)\in\Upsilon}(u_E^n - u_W^n)T_{3WE}$, where

$$T_{3WE} = m(\sigma)k(D_\sigma \nabla\varphi(x_{WE}, t_{n-1}))\cdot \mathbf{n}_{W,\sigma} - \int_{t_{n-1}}^{t_n}\int_\sigma (D_\sigma \nabla\varphi(s,t))\cdot \mathbf{n}_{W,\sigma} ds dt.$$

Due to the smoothness of φ, the mean value theorem, and the finiteness of D_σ, we have $|T_3| \le C_{12}(h+k)|\sum_{n=1}^{N_{\max}} k \sum_{(W,E)\in\Upsilon}(u_E^n - u_W^n)m(\sigma)|$, and similarly as above by using (4.7) and (4.45) we get

$$|T_3| \to 0 \text{ as } h, k \to 0. \tag{4.49}$$

Let us define the fourth term as $T_4 = \frac{1}{2}\sum_{n=1}^{N_{\max}} \sum_{(W,E)\in\Upsilon}(u_E^n - u_W^n)T_\sigma$, where T_σ is represented by the relation $T_\sigma = \int_{t_{n-1}}^{t_n}\int_\sigma (\bar{\lambda}_\sigma - \bar{\lambda})\varphi_x + (\bar{\beta}_\sigma - \bar{\beta})\varphi_y ds dt$ and $\bar{\lambda}(x,t)$ and $\bar{\beta}(x,t)$ are the elements of matrix $D(x,t)$ in the basis $(\mathbf{n}_{W,\sigma}, \mathbf{t}_{W,\sigma})$. Owing to the continuity of function $\nabla\varphi$, one can show that there exist positive constants C_{13} and C_{14} such that $|\varphi_x| \le C_{13}$ and $|\varphi_y| \le C_{14}$. By using these facts it will be sufficient to prove that

$$|\lambda_\sigma - \lambda| = |\lambda(x_{WE}, t_{n-1}) - \lambda(x,t)| \le C_{15} h, \quad x \in \sigma, \quad t \in [t_{n-1}, t_n], \tag{4.50}$$

$$|\beta_\sigma - \beta| = |\beta(x_{WE}, t_{n-1}) - \beta(x,t)| \le C_{16} h, \quad x \in \sigma, \quad t \in [t_{n-1}, t_n], \tag{4.51}$$

$$|\nu_\sigma - \nu| = |\nu(x_{WE}, t_{n-1}) - \nu(x,t)| \le C_{17} h, \quad x \in \sigma, \quad t \in [t_{n-1}, t_n] \tag{4.52}$$

for each edge $\sigma \in \mathcal{E}$. It comes from (1.10) that

$$|\lambda_\sigma - \lambda| = \left| \frac{\kappa_1 v_{1_\sigma}^2 + \kappa_{2_\sigma} v_{2_\sigma}^2}{v_{1_\sigma}^2 + v_{2_\sigma}^2} - \frac{\kappa_1 v_1^2 + \kappa_2 v_2^2}{v_1^2 + v_2^2} \right|,$$

where the terms with index σ correspond to λ_σ and the remaining terms to λ. Here κ_2 depends on μ_1 and μ_2 (see (1.9)), and μ_1 and μ_2 together with v_1 and v_2 (see (1.7) and (1.8)) depend on the elements of the structure tensor $J_\rho(\nabla u_{\tilde{t}}) = \left(\begin{smallmatrix} a & b \\ b & c \end{smallmatrix}\right)$ which is constructed by (1.5)–(1.6). Similarly, λ_σ depends on the elements of tensor $J_\rho(\nabla u_{h,k_{\tilde{t}}})(x_{KL}) = \left(\begin{smallmatrix} a_\sigma & b_\sigma \\ b_\sigma & c_\sigma \end{smallmatrix}\right)$, evaluated numerically at point x_{KL}, using values of $\tilde{u}_{h,k}$ and weights given by convolutions. Using (4.25)–(4.27) leads to

$$a_\sigma = G_\rho * \left(\frac{\partial G_{\tilde{t}}}{\partial x} * u_{h,k} \right)^2, \quad c_\sigma = G_\rho * \left(\frac{\partial G_{\tilde{t}}}{\partial y} * u_{h,k} \right)^2,$$
$$b_\sigma = G_\rho * \left[\left(\frac{\partial G_{\tilde{t}}}{\partial x} * u_{h,k} \right) \left(\frac{\partial G_{\tilde{t}}}{\partial y} * u_{h,k} \right) \right].$$

Elements of the matrix $J_\rho(\nabla u_{\tilde{t}})$ are given as follows:

$$a = G_\rho * \left(\frac{\partial G_{\tilde{t}}}{\partial x} * u \right)^2, \quad b = G_\rho * \left[\left(\frac{\partial G_{\tilde{t}}}{\partial x} * u \right) \left(\frac{\partial G_{\tilde{t}}}{\partial y} * u \right) \right], \quad c = G_\rho * \left(\frac{\partial G_{\tilde{t}}}{\partial y} * u \right)^2.$$

To prove (4.50), we use inequality $p^2 - q^2 \le |p+q||p-q|$, and then we get $a_\sigma - a \le G_\rho * (|\frac{\partial G_{\tilde{t}}}{\partial x} * (u_{h,k}+u)||\frac{\partial G_{\tilde{t}}}{\partial x} * (u_{h,k}-u)|)$. Applying (4.28), the boundedness of solution u, and Lemma 4.8 yields that $a_\sigma \to a$ as $h, k \to 0$. One can obtain convergences $c_\sigma \to c$ in the same way, just changing x to y. For b_σ and b we add and subtract $G_\rho * [(\frac{\partial G_{\tilde{t}}}{\partial x} * u)(\frac{\partial G_{\tilde{t}}}{\partial y} * u_{h,k})]$. Then we have

$$b_\sigma - b = G_\rho * \left[\left(\frac{\partial G_{\tilde{t}}}{\partial y} * u_{h,k} \right) \left(\frac{\partial G_{\tilde{t}}}{\partial x} * (u_{h,k} - u) \right) + \left(\frac{\partial G_{\tilde{t}}}{\partial x} * u \right) \left(\frac{\partial G_{\tilde{t}}}{\partial y} * (u_{h,k} - u) \right) \right],$$

and using (4.28) and Lemma 4.8 yields the result. Due to dependence of v_1, v_2, μ_1, μ_2, and κ_2 on a, b, and c, one can conclude that also $v_{1_\sigma} \to v_1$ as $h, k \to 0$ and v_{2_σ}, μ_{1_σ}, μ_{2_σ}, and κ_{2_σ} correspondingly. Then we use above-mentioned facts to get $\lambda_\sigma \to \lambda$ as $h, k \to 0$. Since $|\beta_\sigma - \beta| = |\frac{v_{1_\sigma} v_{2_\sigma} (\kappa_1 - \kappa_{2_\sigma})}{v_{1_\sigma}^2 + v_{2_\sigma}^2} - \frac{v_1 v_2 (\kappa_1 - \kappa_2)}{v_1^2 + v_2^2}|$ and $|\nu_\sigma - \nu| = |\frac{\kappa_{2_\sigma} v_{1_\sigma}^2 + \kappa_1 v_{2_\sigma}^2}{v_{1_\sigma}^2 + v_{2_\sigma}^2} - \frac{\kappa_2 v_1^2 + \kappa_1 v_2^2}{v_1^2 + v_2^2}|$, one can prove the inequalities (4.51) and (4.52) in the same way. Then by using the same technique as in estimates of T_2 and T_3 we get

$$|T_4| \to 0 \text{ as } h, k \to 0. \tag{4.53}$$

The last term is given by $T_5 = \int_0^T \int_\Omega \nabla \cdot (D \nabla \varphi(x,t)) \, (u(x,t) - u_{h,k}(x,t)) dx dt$. We use the property that $D \in C^\infty(R^{2\times 2}, R^{2\times 2})$ (see [24, p. 115]) to state that $\nabla \cdot (D \nabla \varphi(x,t)) \in L^2(Q_T)$. Then by using the strong convergence of $u_{h,k}$ to u one can see that

$$|T_5| \to 0 \text{ as } h, k \to 0. \tag{4.54}$$

Now we can state following convergence result.

THEOREM 4.9. *The sequence $u_{h,k}$ converges strongly in $L^2(Q_T)$ to the unique weak solution u of (1.1)–(1.3) as $h, k \to 0$.*

Proof. The result comes from (4.46), (4.48), (4.49), (4.53), (4.54), and the fact that the limit u of $u_{h,k}$ is in space $L^2(0,T; H^1(\Omega))$. Due to the uniqueness of the weak solution, which can be found in [24], not only the subsequence but also the sequence $u_{h,k}$ itself converges to u. ∎

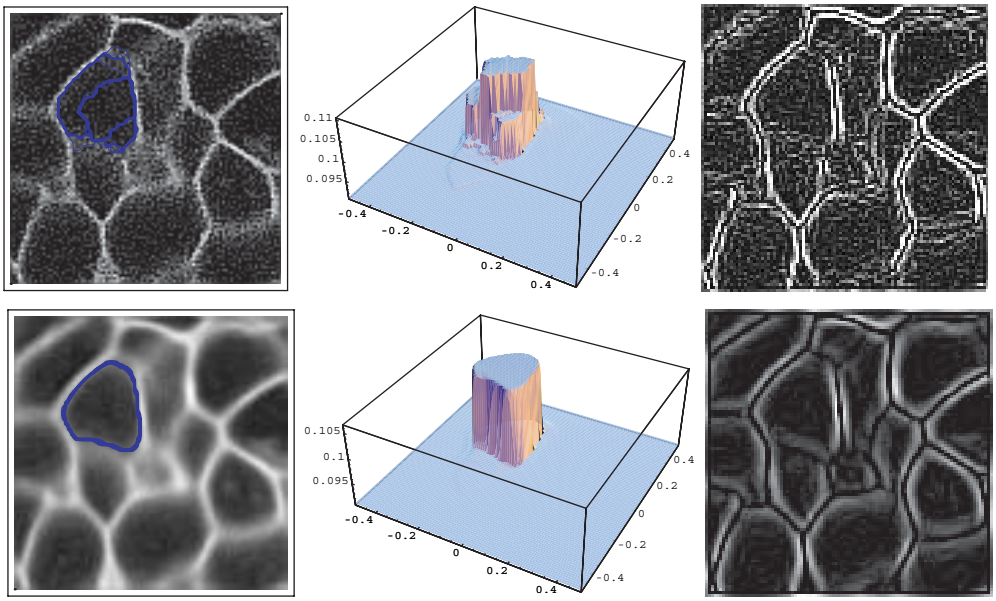


FIG. 5.1. *The cell segmentation and edge detection for the original image and the image filtered by* 4 *diffusion steps. Top row: The original image (*100×100 *pixels detail) with the isolines of the final state of the segmentation function (left), the graph of the final state of the segmentation function (middle), and the edge detection for the original image (right). Bottom row: The filtered image with the isolines of the final state of the segmentation function (left), the graph of the final state of the segmentation function (middle), and the edge detection for the filtered image (right).*

5. Numerical experiments. In this section we present results of several computational examples using real two-dimensional images coming from multiphoton laser scanning microscopy. They represent the membranes and nuclei of cells in the early stages of zebrafish embryogenesis. Especially the images of membranes are well suited for processing by this type of diffusion which is documented by comparing the edge detection and cell segmentation results before and after filtering in Figures 5.1–5.4. In the experiments we use the spatial step $h = 0.01$, time step $k = 0.0001$, $C = 1$, $\alpha = 0.001$, $\tilde{t} = 0.00001$, $\rho = 0.002$. The arising linear systems are solved using Gauss–Seidel iterations. Satisfactory results were obtained after a few filtering steps, so the denoising method is really fast. In the presented experiments we do not observe any stability problems which are a usual drawback of explicit schemes; cf. [25].

The nonlinear tensor anisotropic diffusion smoothes out the noise and improves significantly the connectivity of the coherent structures. Although the filtered image seems to be more blurred compared to the original one—cf. Figure 5.1 (left top and bottom)—the enhancement of the structure connectivity and improvement of the quality of edge detection—cf. Figure (5.1 right top and bottom)—enable us to get much more precise results of segmentation algorithms based on image intensity gradient information such as the subjective surface method [20, 16, 2]. In the subjective surface method, the initial segmentation function in the form of a peak centered in the approximate centroid of the segmented object is created. Then the initial function is evolved by nonlinear PDE, it forms a shock profile during the evolution, and the segmented object is detected as an isoline of the final (numerically steady) state of the segmentation function; for details we refer to [20, 16]. Since many spurious noisy structures can be seen in the original image, which is expressed in a highly noisy edge

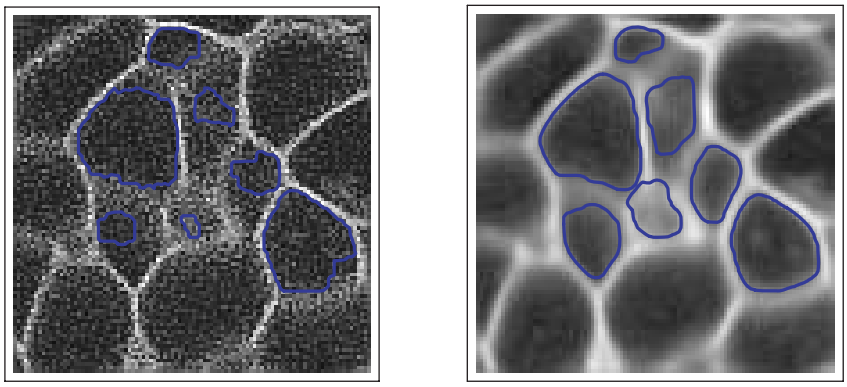


FIG. 5.2. *The results of subjective surface cell segmentation when using unfiltered and filtered images, respectively.*

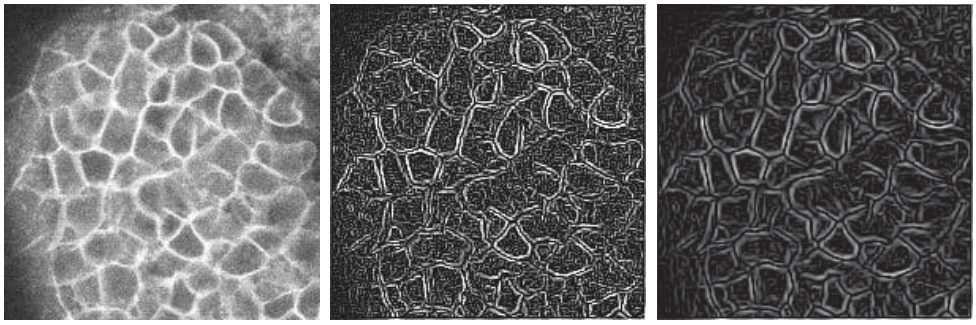


FIG. 5.3. *The image of the cell membranes (*200×200 *pixels, left), its edge detection (middle), and the edge detection for the image filtered by* 2 *diffusion steps where the strong improvement of structure coherence can be seen (right).*

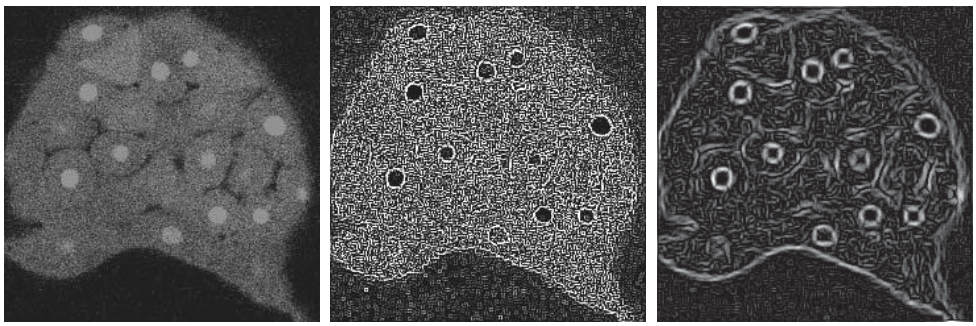


FIG. 5.4. *The image of the cell membranes and nuclei (*240×240 *pixels, left), its edge detection (middle), and the edge detection for the image filtered by* 5 *diffusion steps (right).*

detection result, the segmentation algorithm can hardly find the correct cell boundary using the noisy data. It is difficult to the chose the proper isoline when several shocks are formed in the irregular steady state which is created due to a noise in the image—cf. Figure 5.1 (left and middle top). On the other hand, using a few steps of the nonlinear tensor anisotropic diffusion, all level lines are accumulated along the

TABLE 5.1
Error in the $L_2(I, L_2(\Omega))$ norm and the EOC comparing numerical and exact solutions.

n	h	k	Error	EOC
10	0.2	0.04	$1.809572 \cdot 10^{-4}$	
20	0.1	0.01	$0.3835138 \cdot 10^{-4}$	2.2383
40	0.05	0.0025	$0.09159927 \cdot 10^{-4}$	2.06587
80	0.025	0.000625	$0.02238713 \cdot 10^{-4}$	2.03267
160	0.0125	0.00015625	$0.00495121 \cdot 10^{-4}$	2.17682

cell boundary (just one shock is created in the final state of segmentation function)—cf. Figure 5.1 (left and middle bottom), and the cell can be segmented precisely. Now it is easy to choose the isoline for the cell boundary representation, we take the average of minimal and maximal values of the final segmentation function, and in Figure 5.2 we show the segmentation results for several cells visualizing it for both unfiltered and filtered images. We use the same parameters of the subjective surface segmentation method in both cases, and one can see much more precise segmentation results after filtering.

In Figures 5.3–5.4 we show two other real images: originals (left) and edge detection results for originals (middle) and after a few steps of filtering (right). Again, one can clearly see coherence enhancement and edge detection improvement.

In the last experiment we test the experimental order of convergence (EOC) of our method. In the theoretical part we prove its convergence, and the rigorous error estimates will be an objective of a further study; cf. [7]. Here we consider function $u(x, y, t) = t\cos(\pi x)\cos(\pi y)$ which fulfills the boundary conditions in the domain $\Omega = [-1, 1]^2$ and in time interval $I = [0, 1]$. Putting this function into the model equation (1.1), without convolutions, because we do not need to smooth either the function or the structure tensor, we get the nonzero right-hand side, and we modify the scheme accordingly. We take $C = 1$ and $\alpha = 0.001$ so the diffusion matrix D has eigenvalues between α and 1 and the process is strongly anisotropic. Then we take subsequently refined grids with $M = n^2$ finite volumes, $n = 10, 20, 40, 80, 160$, and the time step $k = h^2$, and we measure errors in the $L_2(I, L_2(\Omega))$ norm, which is natural for testing the schemes for solving parabolic problems. In Table 5.1 we report the errors for different grid sizes, and we observe that the EOC of our numerical scheme is equal to 2.

6. Acknowledgments. The authors thank Dr. Nadine Peyrieras from CNRS-DEPSN, Paris, coordinator of the European project Embryomics, for providing data used in our numerical experiments. We thank also Prof. Robert Eymard for giving us a hint on how to treat the T_1 term in the convergence proof.

REFERENCES

[1] F. CATTÉ, P.-L. LIONS, J.-M. MOREL, AND T. COLL, *Image selective smoothing and edge detection by nonlinear diffusion*, SIAM J. Numer. Anal., 29 (1992), pp. 182–193.

[2] S. CORSARO, K. MIKULA, A. SARTI, AND F. SGALLARI, *Semi-implicit co-volume method in 3D image segmentation*, SIAM J. Sci. Comput., 28 (2006), pp. 2248–2265.

[3] Y. COUDIERE, J. P. VILA, AND P. VILLEDIEU, *Convergence rate of a finite volume scheme for a two-dimensional convection-diffusion problem*, M2AN Math. Model. Numer. Anal., 33 (1999), pp. 493–516.

[4] R. EYMARD, T. GALLOUËT, AND R. HERBIN, *Finite Volume Methods*, in Handbook for Numerical Analysis 7 Ph. Ciarlet and J. L. Lions, eds., Elsevier, New York, 2000.

[5] R. Eymard, T. Gallouët, and R. Herbin, *A cell-centred finite-volume approximation for anisotropic diffusion operators on unstructured meshes in any space dimension*, IMA J. Numer. Anal., 26 (2006), pp. 326–353.
[6] G. Gilboa, N. Sochen, and Y. Y. Zeevi, *Forward-and-backward diffusion processes for adaptive image enhancement and denoising*, IEEE Trans. Image Proc., 11 (2002), pp. 689–703.
[7] A. Handlovičová and Z. Krivá, *Error estimates for finite volume scheme for Perona-Malik equation*, Acta Math. Univ. Comenian. (N.S.), 74 (2005), pp. 79–94.
[8] A. Handlovičová, K. Mikula, and F. Sgallari, *Semi-implicit complementary volume scheme for solving level set like equations in image processing and curve evolution*, Numer. Math., 93 (2003), pp. 675–695.
[9] B. Jähne, *Spatio-temporal Image Processing*, Lecture Notes in Comput. Sci. 751, Springer, Berlin, 1993.
[10] Z. Krivá, *Explicit finite volume scheme for the Perona-Malik equation*, Comput. Methods Appl. Math., 5 (2005), pp. 170–200.
[11] Z. Krivá and K. Mikula, *An adaptive finite volume method in processing of color images*, in Proceedings of ALGORITMY 2000, Conference on Scientific Computing, Podbanské, 2000, pp. 174–187.
[12] Z. Krivá and K. Mikula, *An adaptive finite volume scheme for solving nonlinear diffusion equations arising in image processing*, J. Visual Comm. Image Repres., 13 (2002), pp. 22–35.
[13] C. Lange and K. Polthier, *Anisotropic smoothing of point sets*, Com. Geom. Design, 22 (2005), pp. 680–692.
[14] K. Mikula, *Image processing with partial differential equations*, Modern Methods in Scientific Computing and Applications, A. Bourlioux and M. J. Gander, eds., Springer, Berlin, 2002, pp. 283–321.
[15] K. Mikula and N. Ramarosy, *Semi-implicit finite volume scheme for solving nonlinear diffusion equations in image processing*, Numer. Math., 89 (2001), pp. 561–590.
[16] K. Mikula, A. Sarti, and F. Sgallari, *Semi-implicit co-volume level set method in medical image segmentation*, in Handbook of Biomedical Image Analysis: Segmentation and Registration Models, J. Suri et al., eds., Springer, New York, 2005, pp. 583–626.
[17] P. Perona and J. Malik, *Scale space and edge detection using anisotropic diffusion*, in Proceedings of the IEEE Computer Society Workshop on Computer Vision, 1987.
[18] T. Preusser and M. Rumpf, *An adaptive finite element method for large scale image processing*, J. Visual Comm. Image Repres., 11 (2002), pp. 183–195.
[19] M. Rumpf and T. Preusser, *A level set method for anisotropic geometric diffusion in* $3D$ *image processing*, SIAM J. Appl. Math., 62 (2002), pp. 1772–1793.
[20] A. Sarti, R. Malladi, and J. A. Sethian, *Subjective surfaces: A method for completing missing boundaries*, Proc. Nat. Acad. Sci. U.S.A., 12 (2000), pp. 6258–6263.
[21] A. Sarti, K. Mikula, and F. Sgallari, *Nonlinear multiscale analysis of* $3D$ *echocardiographic sequences*, IEEE Trans. Medical Imaging, 18 (1999), pp. 453–467.
[22] J. Weickert, *Anisotropic Diffusion in Computer Vision*, Teubner, Stuttgart, 1998.
[23] J. Weickert, *Coherence-enhancing diffusion of colour images*, Image and Vision Computing, 17 (1999), pp. 201–212.
[24] J. Weickert, *Coherence-enhancing diffusion filtering*, Internat. J. Comput. Vision, 31 (1999), pp. 111–127.
[25] J. Weickert and H. Scharr, *A scheme for coherence-enhancing diffusion filtering with optimized rotation invariance*, J. Visual Comm. Image Repres., 13 (2002), pp. 103–118.
[26] X. Ye, *A new discontinuous finite volume method for elliptic problems*, SIAM J. Numer. Anal., 42 (2004), pp. 1062–1072.

SIAM J. NUMER. ANAL.
Vol. 46, No. 1, pp. 61–87
© 2007 Society for Industrial and Applied Mathematics

FINITE ELEMENT APPROXIMATIONS OF WAVE MAPS INTO SPHERES*

SÖREN BARTELS†, XIAOBING FENG‡, AND ANDREAS PROHL§

Abstract. Three fully discrete finite element methods are developed for approximating wave maps into the sphere based on two different approaches. The first method is an explicit scheme and the numerical solution satisfies the sphere-constraint exactly at every node. The second and third methods are implicit schemes which are based on a penalization approach, their numerical solutions satisfy the sphere-constraint approximately, and the quality of approximations is controlled by a small penalization parameter. Discrete energy conservation laws which mimic the underlying differential conservation law are established, and convergence of all proposed methods is proved. Computational experiments are also provided to validate the proposed methods and to present numerical evidence for possible finite-time blow-ups of the wave maps.

Key words. wave maps, penalization, finite element method, fully discrete scheme, convergence analysis

AMS subject classifications. 65M12, 65M60, 83C35

DOI. 10.1137/060659971

1. Introduction. Let $\mathbf{R}\times\mathbf{R}^m$ for $m\geq 1$ denote the $(m+1)$-dimensional Minkowski space endowed with the Minkowski metric $\eta=(\eta_{\alpha\beta})=\eta^{-1}=(\eta^{\alpha\beta})=\mathrm{diag}(-1,1,1,\ldots,1)$, and let $(\mathbf{N},g)$ be an n-dimensional Riemannian manifold. A map $\mathbf{u}:(t,\mathbf{x})\in\mathbf{R}\times\mathbf{R}^m\to\mathbf{u}(t,\mathbf{x})\in\mathbf{N}$ is called a *wave map* if it satisfies the *wave map equation* (cf. [20, 25])

$$\Box\mathbf{u}=A(\mathbf{u})(\partial_\alpha\mathbf{u},\partial^\alpha\mathbf{u})\perp T_{\mathbf{u}}\mathbf{N}, \tag{1.1}$$

where $\partial^\alpha\mathbf{u}=\eta^{\alpha\beta}\partial_\beta\mathbf{u}$, and $\Box$ is the d'Alembert operator, $\Box=\partial_\alpha\partial^\alpha$. Moreover, $A(\mathbf{u})(\cdot,\cdot)$ denotes the second fundamental form of the manifold $\mathbf{N}$, and $T_{\mathbf{u}}\mathbf{N}$ denotes the tangent space of $\mathbf{N}$ at $\mathbf{u}$.

It is well known (cf. [20]) that (1.1) can also be interpreted as the Euler–Lagrange equation for the Lagrangian

$$\mathcal{L}(\mathbf{u}):=\frac{1}{2}\int_{\mathbf{R}\times\mathbf{R}^m}\langle\partial_\alpha\mathbf{u},\partial^\alpha\mathbf{u}\rangle_g\,\mathrm{d}x\mathrm{d}t, \tag{1.2}$$

where all the first order derivatives of $\mathbf{u}$ take values in the tangent space $T\mathbf{N}$ of $\mathbf{N}$, that is,

$$\partial_\alpha\mathbf{u}:\mathbf{R}\times\mathbf{R}^m\to T_{\mathbf{u}}\mathbf{N},\qquad \alpha=t,x_1,x_2,\ldots,x_m.$$

*Received by the editors May 15, 2006; accepted for publication (in revised form) May 11, 2007; published electronically December 19, 2007.
http://www.siam.org/journals/sinum/46-1/65997.html

†Department of Mathematics, Humboldt-Universität zu Berlin, Unter den Linden 6, D-10099 Berlin, Germany (sba@math.hu-berlin.de). This author was supported by the Deutsche Forschungsgemeinschaft through the DFG Research Center MATHEON "Mathematics for key technologies" in Berlin. Part of this work was done when this author visited Forschungsinstitut für Mathematik (ETH Zürich) in February, 2006.

‡Department of Mathematics, The University of Tennessee, Knoxville, TN 37996 (xfeng@math.utk.edu). The work of this author was partially supported by NSF grant DMS-0410266.

§Mathematisches Institut, Universität Tübingen, Auf der Morgenstelle 10, D-72076 Tübingen, Germany (prohl@na.uni-tuebingen.de).

The wave map equation is the simplest geometric nonlinear wave equation, thus providing a good setup for studying nonlinear wave interactions. Wave maps have many applications in physics; they arise as harmonic gauges in general relativity (cf. [10]) and as the nonlinear σ-models in particle physics (cf. [1]); under suitable hypotheses, the Einstein vacuum equations with cylindrical symmetry reduce to a wave map system on $(2+1)$-dimensional Minkowski space-time (cf. [7]). Moreover, wave maps also arise in the analysis of the more difficult hyperbolic Yang–Mills equations either as special cases or as equations for certain families of gauge transformations (cf. [11]).

In this paper we consider wave maps in the case when the target manifold $\mathbf{N}$ is an $(n-1)$-sphere, i.e., $\mathbf{N} = \mathbf{S}^{n-1} \subset \mathbf{R}^n$ $(n \geq 1)$. In this case, we have (cf. [20])

$$\begin{aligned} \Box &= \partial_t^2 - \Delta, \\ \langle \partial_\alpha \mathbf{u}, \partial^\alpha \mathbf{u} \rangle_g &= |\nabla \mathbf{u}|^2 - |\partial_t \mathbf{u}|^2, \\ A(\mathbf{u})(\partial_\alpha \mathbf{u}, \partial^\alpha \mathbf{u}) &= \big(|\,\nabla \mathbf{u}\,|^2 - |\,\partial_t \mathbf{u}\,|^2\big)\mathbf{u}, \end{aligned}$$

where $\nabla \mathbf{u} = (\partial_{x_1}, \ldots, \partial_{x_m})$, and Δ is the (spatial) Laplacian. Specifically, we shall study the following initial-boundary value problem.

Given $T > 0$ and $(\mathbf{u}_0(\mathbf{x}), \mathbf{u}_1(\mathbf{x})) \in \mathbf{S}^{n-1} \times T_{\mathbf{u}_0(\mathbf{x})}\mathbf{S}^{n-1}$ for all $\mathbf{x} \in \Omega$, find a wave map $\mathbf{u} : \mathbf{R} \times \mathbf{R}^m \to \mathbf{S}^{n-1}$ such that

$$\Box \mathbf{u} = \big(|\,\nabla \mathbf{u}\,|^2 - |\,\partial_t \mathbf{u}\,|^2\big)\mathbf{u} \qquad \text{in } \Omega_T := (0,T) \times \Omega\,, \tag{1.3}$$

$$\frac{\partial \mathbf{u}}{\partial \mathbf{n}} = 0 \qquad \text{on } \partial\Omega_T := (0,T) \times \partial\Omega\,, \tag{1.4}$$

$$\mathbf{u}(0,\cdot) = \mathbf{u}_0, \qquad \partial_t \mathbf{u}(0,\cdot) = \mathbf{u}_1 \qquad \text{in } \Omega\,, \tag{1.5}$$

where $\Omega \subset \mathbf{R}^m$ is a bounded domain. Here, $\mathbf{n}$ and $\frac{\partial \mathbf{u}}{\partial \mathbf{n}}$ denote the outward unit normal to, and the normal derivative of $\mathbf{u}$ on the boundary $\partial\Omega$, of Ω.

It is well known that the wave map equation (1.1) is *integrable* (so are many other wave equations), which implies that the solutions of (1.1) satisfy some conservation law, which is, for wave maps with the sphere target manifold (cf. [20]),

$$E\big(\partial_t \mathbf{u}(t,\cdot), \mathbf{u}(t,\cdot)\big) = E\big(\mathbf{u}_1(\cdot), \mathbf{u}_0(\cdot)\big) \qquad \forall\, t \geq 0, \tag{1.6}$$

where

$$E\big(\partial_t \mathbf{u}(t,\cdot), \mathbf{u}(t,\cdot)\big) := \frac{1}{2}\Big[\,\|\,\partial_t \mathbf{u}(t,\cdot)\,\|_{L^2}^2 + \|\,\nabla \mathbf{u}(t,\cdot)\,\|_{L^2}^2\Big]. \tag{1.7}$$

Wave map equation (1.1) has been extensively studied in the past twenty years from a theoretical point of view. It was shown that on a (1+1)-dimensional Minkowski space-time base the equation has global in time smooth solutions if the initial data are smooth. It has also been known that for $m \geq 3$ the wave maps on an $(m+1)$-dimensional Minkowski space-time base can blow up in finite time in the sense that $\lim_{t \to t^*} \|\nabla \mathbf{u}(\cdot,t)\|_{L^\infty} = \infty$ for some $t^* < \infty$. On the other hand, whether the wave map equation has global in time smooth solutions on a $(2+1)$-dimensional Minkowski space-time base remains an open problem; see Open Problems 5 and 6 in [25]. The dimension analysis identifies $(2+1)$ as the critical dimension for wave maps, and so there has been a considerable amount of interest in determining whether $(2+1)$-dimensional wave maps develop singularities in finite time for smooth initial data. Numerical experiments of [8, 17] suggest that finite-time collapses, which are identified as local concentrations of energy, may occur for the special class of spherically

equivariant wave maps in the case of large, smooth initial data with the spheres as target manifolds. For more discussions and up-to-date analytical results for wave maps, we refer the reader to the recent monograph of Shatah and Struwe [20] and the survey paper of Tataru [25].

Numerically, besides the works of [8, 17], no other results are known in the literature. However, despite the strong evidence of finite-time singularities in [8, 17], the numerical schemes proposed there are straightforward discretizations of reduced models and are not supported by rigorous numerical analysis, so unphysical numerical artifacts cannot be completely ruled out. The primary goal of this paper is to develop various finite element methods for approximating the initial-boundary value problem (1.3)–(1.5) based on two different approaches (see below and section 3). Rigorous convergence analysis will be provided, so our numerical solutions are guaranteed to converge to weak solutions of (1.3)–(1.5).

Like other well-known maps such as harmonic maps (cf. [4, 5, 22] and references therein), the main difficulty in analyzing and approximating wave maps is caused by the nonflatness of the target manifold $\mathbf{N}$, which is equivalent to imposing some (usually nonconvex) constraints on the solutions of the wave map equation, and is the reason for the nonlinearities. For example, when $\mathbf{N} = \mathbf{S}^{n-1}$, which is the focus of this paper, the constraint is $|\mathbf{u}| = 1$. Recall that the wave map equation (1.1) can be regarded as the Euler–Lagrange equation for the Lagrangian $\mathcal{L}$ defined in (1.2), interpreting this difficulty in the calculus of variation context, which means that the wave map problem is a variation problem with a holonomic constraint (cf. [16]).

The strong nonlinearity and nonconvex constraint make it very difficult to develop convergent numerical methods for (1.3)–(1.5). To the best of our knowledge, no such numerical method is known in the literature up to now. To overcome this difficulty, there are two approaches one can try. On the one hand, one may want to approximate (1.3)–(1.5) directly. To this end, one has to maintain the constraint $|\mathbf{u}| = 1$ at the discrete level, which is difficult to accomplish. Usually, a projection technique is employed for this purpose. Unfortunately, straightforward strategies do not give (discrete) energy conservation laws; consequently, the methods may not be convergent. On the other hand, one can approximate (1.3)–(1.5) indirectly. This approach often involves two steps: first, the wave map equation is penalized and the nonconvex constraint is relaxed by a singular perturbation technique; second, the penalized and nonconstrained problem is approximated by various numerical methods such as finite element, finite difference, and spectral methods. The trade-off is that one now needs to approximate singularly perturbed equations and to resolve the solutions on a very small scale introduced by the penalization parameter and to control numerical parameters according to the penalization parameter for both stability and convergence concerns.

In this paper, we shall develop numerical methods using both direct and indirect approaches. Our direct methods are inspired by the works of [2, 4, 6], where constraint-preserving finite element methods were developed respectively for the Landau–Lifshitz equation and the p-harmonic map heat flow. The direct methods are constructed based on the following equivalent reformulation of (1.3): find $\mathbf{u} : \Omega_T \to \mathbf{S}^{n-1}$ such that for all $\mathbf{w} \in C_0^\infty\big([0,T) \times \Omega, \mathbf{R}^n\big)$ with $\langle \mathbf{u}, \mathbf{w} \rangle = 0$ a.e. in Ω_T there holds

$$\int_0^T \big(\partial_t^2 \mathbf{u}, \mathbf{w}\big)\, \mathrm{d}t + \int_0^T \big(\nabla \mathbf{u}, \nabla \mathbf{w}\big)\, \mathrm{d}t = 0. \tag{1.8}$$

Our indirect methods are based on the following Ginzburg–Landau-type penalization

ansatz (cf. [20]):

$$\Box \mathbf{u}^\varepsilon + \frac{1}{\varepsilon}\big(|\,\mathbf{u}^\varepsilon\,|^2 - 1\big)\mathbf{u}^\varepsilon = 0 \qquad \text{in } \Omega_T\,, \tag{1.9}$$

$$\frac{\partial \mathbf{u}^\varepsilon}{\partial \mathbf{n}} = 0 \qquad \text{on } \partial\Omega_T\,, \tag{1.10}$$

$$\mathbf{u}^\varepsilon(0,\cdot) = \mathbf{u}_0\,, \qquad \partial_t \mathbf{u}^\varepsilon(0,\cdot) = \mathbf{u}_1 \qquad \text{in } \Omega \tag{1.11}$$

for $0 < \varepsilon << 1$. It was shown that for each $\varepsilon > 0$ the regularized problem (1.9)–(1.11) has a weak solution $\mathbf{u}^\varepsilon : \Omega_T \to \mathbf{R}^n$, which satisfies the following conservation law (hence, (1.9)–(1.11) is also integrable):

$$E\big(\partial_t \mathbf{u}^\varepsilon(t,\cdot), \mathbf{u}^\varepsilon(t,\cdot)\big) + \frac{1}{\varepsilon} F(\mathbf{u}^\varepsilon(t,\cdot)) = E\big(\mathbf{u}_1(\cdot), \mathbf{u}_0(\cdot)\big) \qquad \forall\, t \in [0,T], \tag{1.12}$$

where

$$F(\mathbf{w}) = \frac{1}{4}\int_\Omega \big(|\,\mathbf{w}\,|^2 - 1\big)^2\,\mathrm{d}\mathbf{x}. \tag{1.13}$$

Moreover, there exists a subsequence of $\{\mathbf{u}^\varepsilon\}$ which converges to a map $\mathbf{u}$, which turns out to be a weak solution of (1.3)–(1.5) as $\varepsilon \to 0^+$.

The remainder of this paper is organized as follows. In section 2 we introduce some notation and a weak formulation of (1.3)–(1.5). In section 3 we present our fully discrete finite element methods based on both direct and indirect approaches. The convergence analysis of the proposed numerical methods is given in sections 4 and 5. In section 6 we present several numerical experiment results and provide numerical evidence for finite-time (finite-energy) blow-ups of wave maps in the critical dimension $(2+1)$. Finally, we finish the paper with a few concluding remarks.

2. Preliminaries. Standard notations are adopted throughout this paper. $\langle\cdot,\cdot\rangle$ denotes the standard inner product of the Euclidean space $\mathbf{R}^m$, and $(\cdot,\cdot) := (\cdot,\cdot)_\Omega$ is the standard L^2-inner product over the domain Ω. $W^{m,p}(\Omega,\mathbf{R}^n)$ denotes the (m,p)-Sobolev space of vector-valued functions, and $\|\cdot\|_{W^{m,p}}$ denotes its norm. Note that $W^{0,p}(\Omega,\mathbf{R}^n) = L^p(\Omega,\mathbf{R}^n)$. Throughout this paper, $C > 0$ is used to denote a generic positive $(\,h,\tau,\varepsilon\,)$-independent constant which may take different values at different locations. We also introduce $\mathbf{u}_t := \partial_t \mathbf{u}, \nabla\mathbf{u} := (\partial_{x_1},\ldots,\partial_{x_m})$, and $D := (\partial_t, \nabla)$ and define the nonlinear Sobolev space

$$W^{1,2}(\Omega,\mathbf{S}^{n-1}) = \big\{\mathbf{v} \in W^{1,2}(\Omega,\mathbf{R}^n);\ \mathbf{v} \in \mathbf{S}^{n-1} \text{ a.e. in } \Omega\big\}.$$

We now give a definition of weak solutions to (1.3)–(1.5).

DEFINITION 2.1. *Given $T > 0$ and $(\,\mathbf{u}_0, \mathbf{u}_1\,) \in W^{1,2}(\Omega,\mathbf{S}^{n-1}) \times L^2(\Omega, T\mathbf{S}^{n-1})$, we call $\mathbf{u} : \Omega_T \to \mathbf{S}^{n-1}$ a weak solution of (1.3)–(1.5) if the following hold:*

1. *$D\mathbf{u} \in L^2(\Omega_T,\mathbf{R}^n)$.*
2. *$|\,\mathbf{u}\,| = 1$ a.e. in Ω_T.*
3. *Equation (1.3) holds (in wedged form) in the distributional sense; i.e., for all $\boldsymbol{\phi} \in C_0^\infty\big([0,T);\, W^{1,2}(\Omega,\mathbf{R}^n)\big)$ we have*

$$-\int_0^T \big(\mathbf{u}_t \wedge \mathbf{u}, \boldsymbol{\phi}_t\big)\,\mathrm{d}t + \int_0^T \big(\nabla\mathbf{u} \wedge \mathbf{u}, \nabla\boldsymbol{\phi}\big)\,\mathrm{d}t = \big(\mathbf{u}_1 \wedge \mathbf{u}_0, \boldsymbol{\phi}(0,\cdot)\big). \tag{2.1}$$

(4) *The initial conditions* $\mathbf{u}_0$ *and* $\mathbf{u}_1$ *are continuously attained by* $\mathbf{u}$ *and* $\mathbf{u}_t$ *in* $W^{1,2}(\Omega;\mathbf{R}^n)$ *and* $L^2(\Omega;\mathbf{R}^n)$*; i.e., there hold as* $t \to 0$

$$\mathbf{u}(t,\cdot) \to \mathbf{u}_0 \quad \textit{in } W^{1,2}(\Omega;\mathbf{R}^n), \qquad \mathbf{u}_t(t,\cdot) \to \mathbf{u}_1 \quad \textit{in } L^2(\Omega;\mathbf{R}^n).$$

(5) *The following energy inequality holds:*

$$E\big(\mathbf{u}_t(t,\cdot),\mathbf{u}(t,\cdot)\big) \le E\big(\mathbf{u}_1,\mathbf{u}_0\big) \qquad \forall\, t \in [0,T]\,. \tag{2.2}$$

Existence of weak solutions to (1.3)–(1.5) for $m \ge 1$ was proved in [19] by the Ginzburg–Landau penalization technique, and weak solutions were constructed as proper limits of $\{\mathbf{u}_\varepsilon\}$ which solve (1.9)–(1.11). Another method, which was proposed in [18], constructs weak solutions for general targets $\mathbf{N}$ as limits of convergent sequences of spatially discrete wave maps. The semidiscrete maps are direct approximations of the original wave maps and are constructed by (abstract) finite difference methods, so no penalization is involved. We remark that weak solutions to (1.3)–(1.5) are not unique in general (cf. [20]).

3. Fully discrete finite element methods.

3.1. Notation. For simplicity, let Ω be a bounded polygonal (when $m = 2$) or polyhedral (when $m = 3$) domain. Let $\mathcal{T}_h$ denote a quasiuniform triangulation of Ω into triangles or tetrahedrons with mesh-size $h > 0$ for $m = 2$ or $m = 3$, respectively. For a domain $A \subset \mathbf{R}^m$, let $\mathcal{P}(A,\mathbf{R}^n)$ stand for the set of all polynomials on A of degree ≤ 1. We define the Lagrange finite element space

$$\boldsymbol{\mathcal{V}}_h := \big\{\mathbf{w} \in C(\overline{\Omega},\mathbf{R}^n);\, \mathbf{w}|_K \in \mathcal{P}(K,\mathbf{R}^n)\ \forall\, K \in \mathcal{T}_h\big\}\,.$$

Let $\mathcal{N}_h$ denote the set of all nodes associated with the finite element space $\boldsymbol{\mathcal{V}}_h$, and let $\big\{\varphi_{\mathbf{q}_i};\, \mathbf{q}_i \in \mathcal{N}_h\big\}$ denote the nodal basis for $\boldsymbol{\mathcal{V}}_h$; we define the following nodal interpolation operator $I_h : C(\overline{\Omega},\mathbf{R}^n) \to \boldsymbol{\mathcal{V}}_h$ by

$$I_h\mathbf{w} := \sum_{\mathbf{q}_i \in \mathcal{N}_h} \mathbf{w}(\mathbf{q}_i)\varphi_{\mathbf{q}_i} \qquad \forall\, \mathbf{w} \in C(\overline{\Omega},\mathbf{R}^n).$$

For any two functions $\mathbf{v},\mathbf{w} \in C(\overline{\Omega},\mathbf{R}^n)$ we define a discrete L^2-inner product by

$$\big(\mathbf{v},\mathbf{w}\big)_h := \int_\Omega I_h\big(\langle \mathbf{v},\mathbf{w}\rangle\big)\,\mathrm{d}\mathbf{x} = \sum_{\mathbf{q}_i \in \mathcal{N}_h} \beta_{\mathbf{q}_i}\langle \mathbf{v}(\mathbf{q}_i),\mathbf{w}(\mathbf{q}_i)\rangle\,,$$

where $\beta_{\mathbf{q}_i} = \int_\Omega \varphi_{\mathbf{q}_i}\,\mathrm{d}\mathbf{x}$ for all $\mathbf{q}_i \in \mathcal{N}_h$. We also define $||\mathbf{w}||_h := (\mathbf{w},\mathbf{w})_h^{\frac12}$. It is easy to check that there hold for all $\mathbf{v}_h,\mathbf{w}_h \in \boldsymbol{\mathcal{V}}_h$

$$||\mathbf{w}_h||_{L^2} \le ||\mathbf{w}_h||_h \le (m+2)^{\frac12}||\mathbf{w}_h||_{L^2}, \tag{3.1}$$

$$\big|(\mathbf{v}_h,\mathbf{w}_h)_h - (\mathbf{v}_h,\mathbf{w}_h)\big| \le Ch\,||\mathbf{v}_h||_{L^2}\,||\nabla\mathbf{w}_h||_{L^2}. \tag{3.2}$$

Let τ_1 and τ be two (small) positive numbers, and

$$t_0 := 0,\ t_1 := \tau_1,\ t_j := t_1 + (j-1)\tau, \quad j = 1,2,\dots,J := \left[\frac{T-t_1}{\tau}\right].$$

Then $\{t_j\}$ forms a partition of the interval $[0,T]$. We introduce the following difference operators (cf. [13, 14]):

$$
\begin{aligned}
&d_t\mathbf{v}^1 := \frac{\mathbf{v}^1-\mathbf{v}^0}{\tau_1}, \qquad \delta_t\mathbf{v}^1 := \frac{\mathbf{v}^2-\mathbf{v}^0}{\tau_1+\tau},\\
&d_t\mathbf{v}^j := \frac{\mathbf{v}^j-\mathbf{v}^{j-1}}{\tau}, \quad \delta_t\mathbf{v}^j := \frac{\mathbf{v}^{j+1}-\mathbf{v}^{j-1}}{2\tau}, \quad j\geq 1,\\
&\mathbf{v}^{j\pm\frac12} := \frac{\mathbf{v}^{j\pm1}+\mathbf{v}^j}{2}, \quad \mathbf{v}^{j,\theta} := \theta\mathbf{v}^{j+1}+(1-2\theta)\mathbf{v}^j+\theta\mathbf{v}^{j-1} \qquad \forall\,\theta\in\left[0,\frac12\right].
\end{aligned}
$$

Given a sequence $\{\mathbf{w}^j\}_{1\leq j\leq J}\in \ell^\infty\big(0,J;\boldsymbol{\mathcal{V}}_h\big)$, we define the following constant and linear interpolations of the sequence in time: for every $t\in(t_{j-1},t_j]$ set

$$
\begin{aligned}
&\mathbf{w}^-(t,\cdot) := \mathbf{w}^{j-1}(\cdot)\,, && \mathbf{w}^+(t,\cdot) := \mathbf{w}^j(\cdot)\,,\\
&\overline{\mathbf{w}}(t,\cdot) := \frac12\big[\mathbf{w}^{j-1}(\cdot)+\mathbf{w}^j(\cdot)\big]\,, && \mathbf{w}(t,\cdot) := \frac{t-t_{j-1}}{\tau}\mathbf{w}^j(\cdot)+\frac{t_j-t}{\tau}\mathbf{w}^{j-1}(\cdot)\,.
\end{aligned}
$$

We note that τ needs to be replaced by τ_1 in the definition of $\mathbf{w}(t,\cdot)$ when $j=1$. It is easy to check that

$$
\|\,\mathbf{w}^\pm-\mathbf{w}\,\|+\|\,\overline{\mathbf{w}}-\mathbf{w}\,\|\leq 3\tau\,\|\,d_t\mathbf{w}\,\|. \tag{3.3}
$$

Hence, $\mathbf{w}^\pm,\overline{\mathbf{w}}$, and $\mathbf{w}$ converge simultaneously to the same limit provided that $\|\,d_t\mathbf{w}\,\|$ is bounded and one of the sequences converges. Finally, we introduce the following subsets of $\boldsymbol{\mathcal{V}}_h$:

$$
\begin{aligned}
\boldsymbol{\mathcal{M}}_h &:= \big\{\mathbf{w}\in\boldsymbol{\mathcal{V}}_h;\,|\,\mathbf{w}(\mathbf{q}_i)\,|=1\;\forall\,\mathbf{q}_i\in\mathcal{N}_h\big\}\,,\\
\boldsymbol{\mathcal{F}}_h(\boldsymbol{\chi}) &:= \big\{\mathbf{w}\in\boldsymbol{\mathcal{V}}_h;\,\langle\mathbf{w}(\mathbf{q}_i),\boldsymbol{\chi}(\mathbf{q}_i)\rangle=0\quad\forall\,\mathbf{q}_i\in\mathcal{N}_h\big\}\,,\quad \text{where } \boldsymbol{\chi}\in\boldsymbol{\mathcal{M}}_h\,.
\end{aligned}
$$

3.2. Fully discrete methods for (1.3)–(1.5). We introduce a family of fully discrete *explicit* finite element methods for (1.3)–(1.5). The starting point is the weak formulation (1.8). To discretize (1.8), the following fully discrete semi-implicit approximation of (1.8) is immediate: For $j=1\to(J-1)$, given $\hat{\mathbf{U}}^{j-1},\hat{\mathbf{U}}^j\in\boldsymbol{\mathcal{M}}_h$, find $\hat{\mathbf{U}}^{j+1}\in\boldsymbol{\mathcal{M}}_h$ such that

$$
(d_t^2\hat{\mathbf{U}}^{j+1},\mathbf{w})+(\nabla\hat{\mathbf{U}}^{j+1},\nabla\mathbf{w})=0 \qquad \forall\,\mathbf{w}\in\boldsymbol{\mathcal{F}}_h(\hat{\mathbf{U}}^j)\,, \tag{3.4}
$$

and $\hat{\mathbf{U}}^0,\hat{\mathbf{U}}^1\in\boldsymbol{\mathcal{M}}_h$, where $d_t^2\varphi^j=d_t(d_t\varphi^j)$. However, (3.4) is difficult to solve because of the nonconvex constraint on $\boldsymbol{\mathcal{M}}_h$. To overcome this difficulty, on noting $\langle\mathbf{u}_t,\mathbf{u}\rangle=0$ we may assume that $d_t\hat{\mathbf{U}}^{j+1}$ is almost an element of $\boldsymbol{\mathcal{F}}_h(\hat{\mathbf{U}}^j)$. This motivates the following explicit scheme for (1.8).

ALGORITHM 3.1.

(i) *Choose* $(\mathbf{U}^0,\mathbf{V}^0)\in\boldsymbol{\mathcal{M}}_h\times\boldsymbol{\mathcal{V}}_h$ *as a suitable approximation to* $(\mathbf{u}_0,\mathbf{u}_1)$.

(ii) *For* $j=0\to(J-1)$, *given* $(\mathbf{U}^j,\mathbf{V}^j)\in\boldsymbol{\mathcal{M}}_h\times\boldsymbol{\mathcal{V}}_h$, *find* $\mathbf{V}^{j+1}\in\boldsymbol{\mathcal{F}}_h(\mathbf{U}^j)$ *such that*

$$
(d_t\mathbf{V}^{j+1},\mathbf{w})_h=(\nabla\mathbf{U}^j,\nabla\mathbf{w}) \qquad \forall\,\mathbf{w}\in\boldsymbol{\mathcal{F}}_h(\mathbf{U}^j)\,,
$$

and define $\mathbf{U}^{j+1}\in\boldsymbol{\mathcal{M}}_h$ *by*

$$
\mathbf{U}^{j+1}(\mathbf{q}_i)=\frac{\mathbf{U}^j(\mathbf{q}_i)+\tau\mathbf{V}^{j+1}(\mathbf{q}_i)}{|\,\mathbf{U}^j(\mathbf{q}_i)+\tau\mathbf{V}^{j+1}(\mathbf{q}_i)\,|} \qquad \forall\,\mathbf{q}_i\in\mathcal{N}_h\,.
$$

Convergence analysis of Algorithm 3.1 will be given in section 4.

Remark 3.1. (a) The auxiliary function $\mathbf{V}^j$ is an approximation to $\mathbf{u}_t(t_j,\cdot)$, so the algorithm may be viewed as a mixed method which approximates $\mathbf{u}$ and $\mathbf{u}_t$ simultaneously.

(b) The last equation in the algorithm combines a discretization of $\mathbf{u}_t = \mathbf{v}$ with a projection to the sphere at the nodes.

(c) Note that owing to $\mathbf{U}^j(\mathbf{q}_i)\cdot\mathbf{V}^{j+1}(\mathbf{q}_i) = 0$ and $|\mathbf{U}^j(\mathbf{q}_i)| = 1$ for all $\mathbf{q}_i \in \mathcal{N}_h$ the projection step is well defined.

3.3. Fully discrete methods for (1.9)–(1.11). We construct two families of fully discrete *implicit* finite element methods for (1.9)–(1.11). Since the problem (1.9)–(1.11) converges to the problem (1.3)–(1.5) as $\varepsilon \to 0^+$ (cf. [20]), the methods constructed in this subsection are indirect methods for approximating the problem (1.3)–(1.5).

Because for each fixed $\varepsilon > 0$ (1.9) is a semilinear system of wave equations, a natural strategy is to adapt some of best known numerical schemes developed for the linear wave equation to the problem (1.9)–(1.11). Indeed, two families of such schemes will be considered in this subsection. The first family of schemes motivated by the work of Baker [3] are one-step implicit mixed methods in the sense of Remark 3.1 (a); the second family motivated by the work of Dupont [13] are two-step implicit methods.

Apart from the similarities of numerical methods to be given below and those developed in [3, 13], we emphasize that, in addition to the extra difficulty caused by the nonlinearity in (1.9), we also have to cope with a new scale which is due to the introduction of the penalization parameter ε. In particular, we need to make sure our numerical solutions satisfy a discrete conservation law which mimics the differential conservation law (1.12), and they are robust with respect to ε.

First, adapting the schemes of [3], we introduce the following fully discrete implicit discretization of (1.9)–(1.11).

ALGORITHM 3.2.

(i) *Choose* $(\mathbf{U}^0,\mathbf{V}^0) \in \big[\boldsymbol{\mathcal{V}}_h\big]^2$ *as a suitable approximation to* $(\,\mathbf{u}_0,\mathbf{u}_1\,)$.

(ii) *For* $j = 0 \to (J-1)$, *given* $(\mathbf{U}^j,\mathbf{V}^j) \in \big[\boldsymbol{\mathcal{V}}_h\big]^2$, *find* $(\mathbf{U}^{j+1},\mathbf{V}^{j+1}) \in \big[\boldsymbol{\mathcal{V}}_h\big]^2$ *such that for all* $(\mathbf{w},\boldsymbol{\chi}) \in [\boldsymbol{\mathcal{V}}_h]^2$

$$(d_t\mathbf{V}^{j+1},\mathbf{w})_h + (\nabla\mathbf{U}^{j+\frac12},\nabla\mathbf{w}) + \frac{1}{2\varepsilon}\big((|\,\mathbf{U}^{j+1}\,|^2 + |\,\mathbf{U}^{j}\,|^2 - 2)\mathbf{U}^{j+\frac12},\mathbf{w}\big)_h = 0\,,$$

$$(d_t\mathbf{U}^{j+1},\boldsymbol{\chi})_h = (\mathbf{V}^{j+\frac12},\boldsymbol{\chi})_h\,.$$

Next, adapting the schemes of [13], we introduce our second family of fully discrete implicit methods for (1.9)–(1.11).

ALGORITHM 3.3.

(i) *Choose* $\mathbf{U}^0,\mathbf{U}^1 \in \boldsymbol{\mathcal{V}}_h$ *as suitable approximations to* $\mathbf{u}_0$ *and* $\mathbf{u}(t_1,\cdot)$.

(ii) *For* $j = 1 \to (J-1)$, *given* $\{\mathbf{U}^i\}_{i=0}^j \subset \boldsymbol{\mathcal{V}}_h$, *find* $\mathbf{U}^{j+1} \in \boldsymbol{\mathcal{V}}_h$ *such that for all* $\mathbf{w} \in \boldsymbol{\mathcal{V}}_h$

$$(d_t^2\mathbf{U}^{j+1},\mathbf{w})_h + (\nabla\mathbf{U}^{j,\frac14},\nabla\mathbf{w}) + \frac{1}{2\varepsilon}\big((|\,\mathbf{U}^{j+\frac12}\,|^2 + |\,\mathbf{U}^{j-\frac12}\,|^2 - 2)\mathbf{U}^{j,\frac14},\mathbf{w}\big)_h = 0\,.$$

Convergence analysis of Algorithms 3.2 and 3.3 will be presented in section 5. We note that in the above equation $\mathbf{U}^{j,\frac14}$ could be replaced by $\mathbf{U}^{j,\theta}$ for any $\theta \in [0,\frac12]$. However, as pointed out in [13], although for $\theta \ge \frac14$ the analogues of the above

scheme are all absolutely stable, the time truncation error is minimized over this class at $\theta = \frac{1}{4}$, which is the main reason for the choice of $\theta = \frac{1}{4}$.

4. Convergence analysis for Algorithm 3.1. The main goal of this section is to prove that the solution sequence $\{\mathbf{U}^j\}$ of Algorithm 3.1 has a convergent subsequence whose limit is a weak solution of (1.3)–(1.5). To this end, we first need to establish some energy estimates for the solution sequence $\{\mathbf{U}^j\}$. Throughout this section we assume $\tau_1 = \tau$.

4.1. Stability. Recall that by definition every weak solution of (1.3)–(1.5) must satisfy the energy inequality (2.2). In what follows, we show that the solution sequence $\{(\mathbf{U}^j, \mathbf{V}^j)\}$ generated by Algorithm 3.1 satisfies a discrete energy inequality, which mimics (2.2), provided that the temporal mesh-size τ and the spatial mesh-size h satisfy the relation $\tau = o(h^{\frac{m+4}{3}})$. Specifically, we have the following lemma.

LEMMA 4.1. *Let $m \geq 1$ and $(\mathbf{U}^0, \mathbf{V}^0) \in \boldsymbol{\mathcal{M}}_h \times \boldsymbol{\mathcal{F}}_h(\mathbf{U}^0)$. Then, under the mesh condition $\tau = o(h^{\frac{4+m}{3}})$, the solution sequence $\{(\mathbf{U}^j, \mathbf{V}^j)\}_{j=0}^J$ of Algorithm 3.1 satisfies*

$$E_h\big(\mathbf{V}^{j+1}, \mathbf{U}^{j+1}\big) + \frac{\tau^2}{2}\sum_{\ell=0}^{j}\big[\,\|\, d_t\mathbf{V}^{\ell+1}\,\|_h^2 + \|\,\nabla d_t\mathbf{U}^{\ell+1}\,\|_{L^2}^2\big] = c_0\, E_h\big(\mathbf{V}^0, \mathbf{U}^0\big) \tag{4.1}$$

with a positive constant $c_0 = 1 + o(1)$ and where

$$E_h(\mathbf{V}^j, \mathbf{U}^j) := \frac{1}{2}\big[\,\|\,\mathbf{V}^j\,\|_h^2 + \|\,\nabla\mathbf{U}^j\,\|_{L^2}^2\,\big]. \tag{4.2}$$

Proof. Since the proof is long, we divide it into two steps.

Step 1 (auxiliary solution estimates). Setting $\mathbf{w} = \mathbf{V}^{j+1}$ in Step (ii) of Algorithm 3.1 and using the inverse inequality $\|\,\nabla\mathbf{v}_h\,\|_{L^2} \leq Ch^{-1}\|\,\mathbf{v}_h\,\|_{L^2}$ yield

$$\frac{1}{2}d_t\|\,\mathbf{V}^{j+1}\,\|_h^2 + \frac{\tau}{2}\|\,d_t\mathbf{V}^{j+1}\,\|_h^2 = -(\nabla\mathbf{U}^j, \nabla\mathbf{V}^{j+1}) \leq Ch^{-1}\,\|\,\nabla\mathbf{U}^j\,\|_{L^2}\|\,\mathbf{V}^{j+1}\,\|_{L^2}\,. \tag{4.3}$$

Next, we need to get an estimate for $d_t\|\,\nabla\mathbf{U}^{j+1}\,\|_{L^2}^2$, which in turn requires an estimate for $\|\,d_t\mathbf{U}^{j+1}\,\|_{L^2}^2$. We now derive these estimates by borrowing an argument from [2]. Let

$$\mathbf{R}^{j+1} := \mathbf{U}^{j+1} - \mathbf{U}^j - \tau\mathbf{V}^{j+1}.$$

Note that $\mathbf{R}^{j+1} \in \boldsymbol{\mathcal{V}}_h$ and $\tau^{-1}\mathbf{R}^{j+1}$ measures the discrepancy between $d_t\mathbf{U}^{j+1}$ and $\mathbf{V}^{j+1}$. By the definition of $\mathbf{U}^{j+1}$ in (ii) of Algorithm 3.1 we have

$$\begin{aligned}|\,\mathbf{R}^{j+1}(\mathbf{q}_i)\,| &= \left|\frac{\mathbf{U}^j(\mathbf{q}_i) + \tau\mathbf{V}^{j+1}(\mathbf{q}_i)}{|\,\mathbf{U}^j(\mathbf{q}_i) + \tau\mathbf{V}^{j+1}(\mathbf{q}_i)\,|} - \mathbf{U}^j(\mathbf{q}_i) - \tau\mathbf{V}^{j+1}(\mathbf{q}_i)\right| \\ &= \big|1 - \big|\mathbf{U}^j(\mathbf{q}_i) + \tau\mathbf{V}^{j+1}(\mathbf{q}_i)\big|\big| \qquad\qquad \forall\,\mathbf{q}_i \in \mathcal{N}_h.\end{aligned}$$

Since $|\,\mathbf{U}^j(\mathbf{q}_i) + \tau\mathbf{V}^{j+1}(\mathbf{q}_i)\,| = \sqrt{1 + \tau^2|\,\mathbf{V}^{j+1}(\mathbf{q}_i)\,|^2} \geq 1$, we conclude that

$$|\,\mathbf{R}^{j+1}(\mathbf{q}_i)\,| \leq \frac{\tau^2}{2}|\,\mathbf{V}^{j+1}(\mathbf{q}_i)\,|^2\,. \tag{4.4}$$

Hence,

$$\int_\Omega |\,\mathbf{R}^{j+1}\,|\,\mathrm{d}\mathbf{x} \le \int_\Omega I_h\big[|\,\mathbf{R}^{j+1}\,|\big]\,\mathrm{d}\mathbf{x} \le \frac{\tau^2}{2}\int_\Omega I_h\big[|\,\mathbf{V}^{j+1}\,|^2\big]\,\mathrm{d}\mathbf{x} \le \frac{\tau^2(m+2)}{2}\int_\Omega |\,\mathbf{V}^{j+1}\,|^2\,\mathrm{d}\mathbf{x}\,. \tag{4.5}$$

Similarly,

$$\|\,\mathbf{R}^{j+1}\,\|_{L^2}^2 \le \|\,\mathbf{R}^{j+1}\,\|_h^2 \le \frac{\tau^4}{4}\|\,\mathbf{V}^{j+1}\,\|_{L^\infty}^2\|\,\mathbf{V}^{j+1}\,\|_h^2 \le C\tau^4 h^{-m}\,\|\,\mathbf{V}^{j+1}\,\|_{L^2}^4\,;$$

thus

$$\begin{aligned}\|\,d_t\mathbf{U}^{j+1}\,\|_{L^2}^2 &\le \big[\|\,\mathbf{V}^{j+1}\,\|_{L^2} + \tau^{-1}\,\|\,\mathbf{R}^{j+1}\,\|_{L^2}\big]^2 \\ &\le \big[1 + C\tau h^{-\frac{m}{2}}\|\,\mathbf{V}^{j+1}\,\|_{L^2}\big]^2\|\,\mathbf{V}^{j+1}\,\|_{L^2}^2\,.\end{aligned} \tag{4.6}$$

Now, setting $\mathbf{w} = \mathbf{V}^{j+1} = d_t\mathbf{U}^{j+1} - \tau^{-1}\mathbf{R}^{j+1}$ in Step (ii) of Algorithm 3.1, we get

$$(d_t\mathbf{V}^{j+1}, \mathbf{V}^{j+1})_h + \big(\nabla\mathbf{U}^{j+1}, \nabla d_t\mathbf{U}^{j+1}\big) = \tau^{-1}(\nabla\mathbf{U}^j, \nabla\mathbf{R}^{j+1}) + \tau\|\,\nabla d_t\mathbf{U}^{j+1}\,\|_{L^2}^2;$$

this, an application of the inverse inequality, and (4.6) yield

$$\begin{aligned}&\frac{1}{2}d_t\|\,\mathbf{V}^{j+1}\,\|_h^2 + \frac{1}{2}d_t\|\,\nabla\mathbf{U}^{j+1}\,\|_{L^2}^2 + \frac{\tau}{2}\|\,d_t\mathbf{V}^{j+1}\,\|_h^2 + \frac{\tau}{2}\|\,\nabla d_t\mathbf{U}^{j+1}\,\|_{L^2}^2 \\ &\quad\le C\tau^{-1}h^{-2}\,\|\,\mathbf{U}^j\,\|_{L^\infty}\|\,\mathbf{R}^{j+1}\,\|_{L^1} + C\tau h^{-2}\,\|\,d_t\mathbf{U}^{j+1}\,\|_{L^2}^2 \\ &\quad\le C\tau^2 h^{-2}\,\|\,\mathbf{V}^{j+1}\,\|_h^2 + C\tau h^{-2}\,\|\,d_t\mathbf{U}^{j+1}\,\|_{L^2}^2 \\ &\quad\le C\tau h^{-2}\,\|\,\mathbf{V}^{j+1}\,\|_{L^2}^2 + C\tau h^{-2}\big[1 + \tau h^{-\frac{m}{2}}\|\,\mathbf{V}^{j+1}\,\|_{L^2}\big]^2\|\,\mathbf{V}^{j+1}\,\|_{L^2}^2\,.\end{aligned} \tag{4.7}$$

Step 2 (finishing up). We now conclude the proof by an induction argument. The assumptions on $\mathbf{u}_0$ and $\mathbf{U}^0$ imply that there exists a positive constant C_1 independent of τ and h such that $\|\,\nabla\mathbf{U}^0\,\|_{L^2} \le C_1$. Assume that $\|\,\nabla\mathbf{U}^i\,\|_{L^2} \le C_1$ for all $0 \le i \le j$; it follows from (4.3) and a discrete version of Gronwall's lemma that

$$\|\,\mathbf{V}^{j+1}\,\|_{L^2} \le \|\,\mathbf{V}^{j+1}\,\|_h \le Ch^{-1}\,. \tag{4.8}$$

If we use this result in (4.7), together with the mesh condition, we have existence of another $C_2 = C_2(T) > 0$, such that $\tau h^{-2}\tau^2 h^{-m}\,\|\,\mathbf{V}^{j+1}\,\|_{L^2}^2 \le C_2$. The discrete version of Gronwall's lemma then yields ($h \to 0$)

$$E_h\big(\mathbf{V}^{j+1}, \mathbf{U}^{j+1}\big) + \frac{\tau^2}{2}\sum_{\ell=0}^{j}\big[\|\,d_t\mathbf{V}^{\ell+1}\,\|_h^2 + \|\,\nabla d_t\mathbf{U}^{\ell+1}\,\|_{L^2}^2\big] \le \big(1 + o(1)\big)\,E_h\big(\mathbf{V}^0, \mathbf{U}^0\big).$$

Hence, by induction (4.1) holds for all $j = 0, 1, \ldots, J$. The proof is complete. ∎

The following corollary follows immediately from Lemma 4.1 and (4.6).

COROLLARY 4.2. *The solution sequence* $\{\mathbf{U}^j\}$ *of Algorithm* 3.1 *also satisfies*

$$\|\,d_t\mathbf{U}^{j+1}\,\|_{L^2} \le C\,, \qquad j = 0, 1, \ldots, J\,. \tag{4.9}$$

4.2. Convergence. To establish the convergence of Algorithm 3.1, we first need some preparations. Let $(\mathbf{U}^+_{\tau,h}, \mathbf{U}^-_{\tau,h}, \overline{\mathbf{U}}_{\tau,h}, \mathbf{U}_{\tau,h})$ and $(\mathbf{V}^+_{\tau,h}, \mathbf{V}^-_{\tau,h}, \overline{\mathbf{V}}_{\tau,h}, \mathbf{V}_{\tau,h})$ denote their respective constant and linear interpolations in time of the solution sequences $\{\mathbf{U}^j\}$ and $\{\mathbf{V}^j\}$ of Algorithm 3.1 as defined in section 3.1. However, for notational brevity, most of the time we omit subindices τ, h in this subsection and use $(\mathbf{U}^+, \mathbf{U}^-, \overline{\mathbf{U}}, \mathbf{U})$ and $(\mathbf{V}^+, \mathbf{V}^-, \overline{\mathbf{V}}, \mathbf{V})$ to stand for $(\mathbf{U}^+_{\tau,h}, \mathbf{U}^-_{\tau,h}, \overline{\mathbf{U}}_{\tau,h}, \mathbf{U}_{\tau,h})$ and $(\mathbf{V}^+_{\tau,h}, \mathbf{V}^-_{\tau,h}, \overline{\mathbf{V}}_{\tau,h}, \mathbf{V}_{\tau,h})$, respectively.

We first derive the following reformulation of Algorithm 3.1.

LEMMA 4.3. *Suppose that the assumptions of Lemma* 4.1 *are valid. There holds for all* $\mathbf{w} \in C_0^\infty\big([0,T); \boldsymbol{\mathcal{F}}_h(\mathbf{U}^-)\big)$

$$
\begin{aligned}
(4.10)\quad & \left| \int_0^T \big[-(\mathbf{U}_t, \mathbf{w}_t) + (\nabla \mathbf{U}, \nabla \mathbf{w}) \big] \, \mathrm{d}t + \big(\mathbf{V}^0, \mathbf{w}(0,\cdot) \big) \right| \\
& \qquad \le (C\tau h^{-\frac{m}{2}} + h) \left[\int_0^T \| \mathbf{w}_t \|_{L^2}^2 + \| \nabla \mathbf{w}_t \|_{L^2}^2 \, \mathrm{d}t + \| \nabla \mathbf{w}(0,\cdot) \|_{L^2}^2 \right]^{\frac{1}{2}} \\
& \qquad\qquad + \left| \int_0^T (\mathbf{V} - \mathbf{V}^+, \mathbf{w}_t) \, \mathrm{d}t \right| + \left| \int_0^T (\nabla [\mathbf{U} - \mathbf{U}^-], \nabla \mathbf{w}) \, \mathrm{d}t \right| .
\end{aligned}
$$

Proof. Using the new notation we rewrite Step (ii) in Algorithm 3.1 as

$$
\int_0^T \big[(\mathbf{V}_t, \mathbf{w})_h + (\nabla \mathbf{U}^-, \nabla \mathbf{w}) \big] \, \mathrm{d}t = 0 .
$$

Integrating by parts in t in the first term yields

$$
\int_0^T \big[-(\mathbf{V}, \mathbf{w}_t)_h + \big(\nabla \mathbf{U}^-, \nabla \mathbf{w} \big) \big] \, \mathrm{d}t + \big(\mathbf{V}^0, \mathbf{w}(0,\cdot) \big)_h = 0;
$$

this and the relation $\mathbf{V}^+ = \mathbf{U}_t - \tau^{-1} \mathbf{R}^+$ imply that

$$
\begin{aligned}
(4.11)\quad & - \int_0^T \big[-(\mathbf{U}_t, \mathbf{w}_t) + (\nabla \mathbf{U}, \nabla \mathbf{w}) \big] \, \mathrm{d}t - \big(\mathbf{V}^0, \mathbf{w}(0,\cdot) \big) \\
& \qquad = \int_0^T \big[(\mathbf{V}, \mathbf{w}_t) - (\mathbf{V}, \mathbf{w}_t)_h \big] \, \mathrm{d}t + \big(\mathbf{V}^0, \mathbf{w}(0,\cdot) \big)_h - \big(\mathbf{V}^0, \mathbf{w}(0,\cdot) \big) \\
& \qquad\qquad + \int_0^T \big[\tau^{-1} (\mathbf{R}^+, \mathbf{w}_t) + (\mathbf{V}^+ - \mathbf{V}, \mathbf{w}_t) + \big(\nabla [\mathbf{U}^- - \mathbf{U}], \nabla \mathbf{w} \big) \big] \, \mathrm{d}t .
\end{aligned}
$$

By (4.1) and the inequality after (4.5) we have

$$
\int_0^T \| \mathbf{R}^+ \|_{L^2}^2 \, \mathrm{d}t \le C \tau^4 h^{-m} .
$$

Finally, the assertion (4.10) follows from (4.11), (4.9), and (3.2). ∎

We are ready to state and prove the first main convergence theorem of this paper.

THEOREM 4.4. *Let* $T > 0$, *and* $\big(\mathbf{u}_0, \mathbf{u}_1\big) \in W^{1,2}(\Omega, \mathbf{S}^{n-1}) \times L^2(\Omega, T\mathbf{S}^{n-1})$. *Suppose that* $\big(\mathbf{U}^0, \mathbf{V}^0\big) \in \boldsymbol{\mathcal{M}}_h \times \boldsymbol{\mathcal{F}}_h(\mathbf{U}^0)$ *satisfies* $\mathbf{U}^0 \to \mathbf{u}_0$ *strongly in* $W^{1,2}(\Omega, \mathbf{R}^n)$, *and* $\mathbf{V}^0 \to \mathbf{u}_1$ *strongly in* $L^2(\Omega, \mathbf{R}^n)$ *as* $h \to 0$. *Let* $\tau = o(h^{\frac{4+m}{3}})$. *Then there exist a map*

$\mathbf{u} \in L^\infty\big(0,T;W^{1,2}(\Omega,\mathbf{R}^n)\big) \cap W^{1,\infty}\big(0,T;L^2(\Omega,\mathbf{R}^n)\big)$ *and a subsequence of solutions* $\{\mathbf{U}_{\tau,h}\}$ *(denoted by the same notation) of Algorithm* 3.1 *such that as* $h \to 0$

$$\begin{aligned} \mathbf{U}_{\tau,h} &\overset{*}{\rightharpoonup} \mathbf{u} \quad \text{in } L^\infty\big(0,T;W^{1,2}(\Omega,\mathbf{R}^n)\big)\,, \\ (\mathbf{U}_{\tau,h})_t &\rightharpoonup \mathbf{u}_t \quad \text{in } L^2(\Omega_T,\mathbf{R}^n)\,. \end{aligned}$$

Moreover, $\mathbf{u}$ *is a weak solution of* (1.3)–(1.5).

Proof. It follows from Lemma 4.1 that under the above mesh relation, there exist a map $\mathbf{u} \in W^{1,\infty}\big(0,T;L^2(\Omega,\mathbf{R}^n)\big) \cap L^\infty\big(0,T;W^{1,2}(\Omega,\mathbf{R}^n)\big)$ and a subsequence of $\{\mathbf{U} = \mathbf{U}_{\tau,h}\}$ (denoted by the same notation) such that as $h \to 0$

$$\begin{aligned} \mathbf{U}, \mathbf{U}^-, \mathbf{U}^+ &\overset{*}{\rightharpoonup} \mathbf{u} \quad \text{in } L^\infty\big(0,T;W^{1,2}(\Omega,\mathbf{R}^n)\big)\,, \\ &\to \mathbf{u} \quad \text{in } L^2(\Omega_T,\mathbf{R}^n), \end{aligned} \tag{4.12}$$

$$\mathbf{U}_t \overset{*}{\rightharpoonup} \mathbf{u}_t \quad \text{in } L^\infty\big(0,T;L^2(\Omega,\mathbf{R}^n)\big)\,. \tag{4.13}$$

Moreover, taking $h \to 0$ in (4.1) immediately shows that $\mathbf{u}$ satisfies the energy inequality in item (5) of Definition 2.1. Furthermore, from the inequality

$$\| \,|\,\mathbf{U}^+(t,\cdot)\,|^2 - 1\,\|_{L^2} \le Ch\,\|\,\nabla\mathbf{U}^+\,\|_{L^2} \qquad \forall\, t \in [0,T]$$

we conclude that $|\,\mathbf{u}\,| = 1$ a.e. in Ω_T.

Given $\boldsymbol{\phi} \in C_0^\infty\big([0,T);C^\infty(\Omega,\mathbf{R}^n)\big)$ let $\mathbf{w}_h = I_h\big(\mathbf{U}\wedge\boldsymbol{\phi}\big)$. By the properties of the interpolation operator I_h we get (cf. [9, 12])

$$\begin{aligned} &\| \left(I_h(\mathbf{U}\wedge\boldsymbol{\phi}) - \mathbf{U}\wedge\boldsymbol{\phi}\right)_t \|_{L^2}^2 \\ &\le Ch^4 \sum_{K\in\mathcal{T}_h} \Big(\| \nabla^2(\mathbf{U}_t\wedge\boldsymbol{\phi}) \|_{L^2(K)}^2 + \| \nabla^2(\mathbf{U}\wedge\boldsymbol{\phi}_t) \|_{L^2(K)}^2 \Big) \\ &\le Ch^4 \left[\| \,|\mathbf{U}_t|\,|\nabla^2\boldsymbol{\phi}|\, \|_{L^2}^2 + \| \,|\nabla\mathbf{U}_t|\,|\nabla\boldsymbol{\phi}|\, \|_{L^2}^2 + \| \nabla^2\boldsymbol{\phi}_t \|_{L^2}^2 + \| \,|\nabla\mathbf{U}|\,|\nabla\boldsymbol{\phi}_t|\, \|_{L^2}^2 \right] \\ &\le Ch^4 \left[\| \boldsymbol{\phi} \|_{W^{2,\infty}}^2 + h^{-2}\,\| \mathbf{U}_t \|_{L^2}^2 \| \nabla\boldsymbol{\phi} \|_{L^\infty}^2 + \| \boldsymbol{\phi}_t \|_{W^{2,\infty}}^2 \right]. \end{aligned}$$

This implies that $(\mathbf{w}_h - \mathbf{U}\wedge\boldsymbol{\phi})_t \to 0$ in $L^\infty\big(0,T;L^2(\Omega,\mathbf{R}^n)\big)$. Hence, as $h \to 0$, using $\mathbf{U}_t \cdot \big(\mathbf{U}_t\wedge\boldsymbol{\phi}\big) = 0$ and (4.12) and (4.13),

$$\begin{aligned} &\int_{\Omega_T} \langle \mathbf{U}_t, (\mathbf{w}_h)_t \rangle \,\mathrm{d}x\mathrm{d}t \\ &\quad = \int_{\Omega_T} \langle \mathbf{U}_t, \left(I_h(\mathbf{U}\wedge\boldsymbol{\phi}) - \mathbf{U}\wedge\boldsymbol{\phi}\right)_t \rangle \,\mathrm{d}x\mathrm{d}t + \int_{\Omega_T} \langle \mathbf{U}_t, \mathbf{U}\wedge\boldsymbol{\phi}_t \rangle \,\mathrm{d}x\mathrm{d}t \\ &\quad \to \int_{\Omega_T} \langle \mathbf{u}_t, \mathbf{u}\wedge\boldsymbol{\phi}_t \rangle \,\mathrm{d}x\mathrm{d}t\,. \end{aligned} \tag{4.14}$$

Similarly, we can show

$$\| \nabla\big[I_h(\mathbf{U}\wedge\boldsymbol{\phi}) - \mathbf{U}\wedge\boldsymbol{\phi}\big] \|_{L^2}^2 \le Ch^2 \left[\| \boldsymbol{\phi} \|_{W^{2,2}}^2 + \| \nabla\mathbf{U} \|_{L^2}^2 \| \nabla\boldsymbol{\phi} \|_{L^\infty}^2 \right].$$

On noting the vector identity $\langle \nabla\mathbf{z}, \nabla\mathbf{z}\wedge\boldsymbol{\phi} \rangle = \langle \nabla\mathbf{z}, \mathbf{z}\wedge\nabla\boldsymbol{\phi} \rangle$, from (4.12) and (4.13) we conclude that as $h \to 0$

$$\begin{aligned} &\int_{\Omega_T} \langle \nabla\mathbf{U}, \nabla\mathbf{w}_h \rangle \,\mathrm{d}\mathbf{x}\mathrm{d}t \\ &\quad = \int_{\Omega_T} \langle \nabla\mathbf{U}, \nabla\big[I_h(\mathbf{U}\wedge\boldsymbol{\phi}) - \mathbf{U}\wedge\boldsymbol{\phi}\big] \rangle \,\mathrm{d}\mathbf{x}\mathrm{d}t + \int_{\Omega_T} \langle \nabla\mathbf{U}, (\mathbf{U}\wedge\nabla\boldsymbol{\phi}) \rangle \,\mathrm{d}\mathbf{x}\mathrm{d}t \\ &\quad \to \int_{\Omega_T} \langle \nabla\mathbf{u}, \mathbf{u}\wedge\nabla\boldsymbol{\phi} \rangle \,\mathrm{d}\mathbf{x}\mathrm{d}t\,. \end{aligned} \tag{4.15}$$

Straightforward calculations show that

$$\left|\int_0^T (\mathbf{V}-\mathbf{V}^+,\mathbf{w}_t)\,\mathrm{d}t\right| \le C\tau^{\frac12}\left(\tau\int_0^T ||\mathbf{V}_t||_{L^2}^2\,\mathrm{d}t\right)^{\frac12}\left(\int_0^T ||\mathbf{w}_t||_{L^2}^2\,\mathrm{d}t\right)^{\frac12}.$$

Similarly, we can bound the last term in (4.10) and benefit from the estimates for $\tau\int_0^T ||\mathbf{V}_t||_{L^2}^2\,\mathrm{d}t$ and $\tau\int_0^T ||\nabla\mathbf{U}_t||_{L^2}^2\,\mathrm{d}t$ of Lemma 4.1. Letting $h\to 0$ in (4.10) and using (4.14) and (4.15) then give (2.1).

In order to verify item (4) of Definition 2.1, we adopt an argument from [20, p. 92]. $\mathbf{u}_0=\lim_{t\to0}\lim_{\tau,h\to0}\mathbf{U}(t,\cdot)$ in $L^2(\Omega,\mathbf{R}^n)$ follows immediately from (5.7). It remains to verify $\mathbf{U}_t(t,\cdot)\to\mathbf{u}_1$ in $L^2(\Omega,\mathbf{R}^n)$ as $t\to0$. To this end, multiply $\Box\mathbf{u}\wedge\mathbf{u}=0$ with $\mathbf{w}\in C_0^\infty\big([0,T);C^\infty(\Omega,\mathbf{R}^n)\big)$, integrate by parts on Ω_T, and then subtract the resulting equation from (4.10); together with Lemma 4.3 and (5.7), (5.8), as well as assumptions on $(\mathbf{U}^0,\mathbf{V}^0)$, we find for the limit $\tau,h\to0$,

$$\big((\mathbf{u}_t(0,\cdot)-\mathbf{u}_1)\wedge\mathbf{u}_0,\mathbf{w}\big)=0\qquad\forall\,\mathbf{w}\in C_0^\infty\big([0,T);C^\infty(\Omega,\mathbf{R}^n)\big)\,.$$

On noting that $\langle\mathbf{u}_1(\mathbf{x}),\mathbf{u}_0(\mathbf{x})\rangle=\langle\mathbf{u}_t(0,\mathbf{x}),\mathbf{u}_0(\mathbf{x})\rangle=0$ for a.e. $\mathbf{x}\in\Omega$, it follows from the vector identity $\mathbf{v}=\langle\mathbf{u},\mathbf{v}\rangle\mathbf{u}-\mathbf{u}\wedge(\mathbf{u}\wedge\mathbf{v})$ with $\mathbf{v}=\mathbf{u}_t(0,\cdot)-\mathbf{u}_1$ that $\mathbf{u}_t(t,\cdot)\rightharpoonup\mathbf{u}_1$ in $L^2(\Omega,\mathbf{R}^n)$ as $t\to0$.

We also need to show that $\mathbf{u}_t(t,\cdot)\to\mathbf{u}_1$ in $L^2(\Omega,\mathbf{R}^n)$ as $t\to0$. By weak lower semicontinuity of the L^2-norm and Fatou's lemma

$$\|\,D\mathbf{u}(t,\cdot)\,\|_{L^2}\le\liminf_{\tau,h\to0}\|\,D\mathbf{U}(t,\cdot)\,\|_{L^2},\qquad t\ge0\,.$$

Hence, for all $t\ge0$, by (3.2), Lemma 4.1, and the assumptions on initial data,

$$\begin{aligned}E\big(\mathbf{u}_t(t,\cdot),\mathbf{u}(t,\cdot)\big)&\le\liminf_{\tau,h\to0}E\big(\mathbf{U}_t(t,\cdot),\mathbf{U}(t,\cdot)\big)\\&=\liminf_{\tau,h\to0}E_h\big(\mathbf{U}_t(t,\cdot),\mathbf{U}(t,\cdot)\big)\le E\big(\mathbf{u}_1,\mathbf{u}_0\big)\,.\end{aligned}$$

Therefore,

$$\limsup_{t\to0}\|\,\mathbf{u}_t(t,\cdot)\,\|_{L^2}\le\|\,\mathbf{u}_1\,\|_{L^2}\,,\quad\limsup_{t\to0}\|\,\nabla\mathbf{u}(t,\cdot)\,\|_{L^2}\le\|\,\nabla\mathbf{u}_0\,\|_{L^2}\,,$$

and the weak convergence $\mathbf{u}_t(t,\cdot)\rightharpoonup\mathbf{u}_1(\cdot)$ in $L^2(\Omega,\mathbf{R}^n)$ and $\nabla\mathbf{u}(t,\cdot)\rightharpoonup\nabla\mathbf{u}_0(\cdot)$ in $L^2(\Omega)$ imply strong convergence $\mathbf{u}_t(t,\cdot)\to\mathbf{u}_1(\cdot)$ and $\nabla\mathbf{u}(t,\cdot)\to\nabla\mathbf{u}_0(\cdot)$ in $L^2(\Omega,\mathbf{R}^{mn})$ as $t\to0$. Consequently, $\mathbf{u}:\Omega_T\to\mathbf{R}^n$ attains prescribed initial data continuously in $W^{1,2}(\Omega,\mathbf{R}^n)\times L^2(\Omega,\mathbf{R}^n)$.

Since all requirements of Definition 2.1 are verified, the map $\mathbf{u}:\Omega_T\to\mathbf{R}^n$ is a weak solution to (1.3)–(1.5). The proof is complete. □

As a by-product, the above convergence result provides an alternative (to the one given in [20]) proof of existence of weak solutions to (1.3)–(1.5).

5. Convergence analysis for Algorithms 3.2 and 3.3. The goal of this section is to prove that the solution sequences $\{\mathbf{U}^j\}$ of Algorithms 3.2 and 3.3 converge uniformly in ε, and the limits of both sequences are weak solutions of (1.9)–(1.11).

5.1. Stability of Algorithms 3.2 and 3.3. Let $E_h(\cdot,\cdot)$ be the same as in (4.2), and define

$$F_h(\mathbf{U}^j) = \frac{1}{4}\| \,|\, \mathbf{U}^j \,|^2 - 1 \,\|_h^2.$$

The following lemma establishes a discrete conservation law for Algorithm 3.2, which mimics the differential conservation law (1.12).

LEMMA 5.1. *For $\varepsilon > 0$, let $\tau_1 = \tau$ and $\big(\mathbf{U}^0, \mathbf{V}^0\big) \in [\boldsymbol{\mathcal{V}}_h]^2$. Then the solution sequence $\{(\mathbf{U}^j, \mathbf{V}^j)\}$ generated by Algorithm 3.2 satisfies for $j = 0 \to (J-1)$*

$$E_h\big(\mathbf{V}^{j+1}, \mathbf{U}^{j+1}\big) + \frac{1}{\varepsilon} F_h(\mathbf{U}^{j+1}) = E_h\big(\mathbf{V}^0, \mathbf{U}^0\big) + \frac{1}{\varepsilon} F_h(\mathbf{U}^0)\,. \tag{5.1}$$

Moreover, if $\mathbf{U}^0 \in \boldsymbol{\mathcal{M}}_h$, then $F_h(\mathbf{U}^0) = 0$, and

$$\| \,d_t \mathbf{U}^{j+1} \,\|_h^2 \le 2E_h(\mathbf{V}^0, \mathbf{U}^0). \tag{5.2}$$

Proof. Taking test functions $(\mathbf{w}, \boldsymbol{\chi}) = (d_t\mathbf{U}^{j+1}, d_t\mathbf{V}^{j+1})$ in (ii) of Algorithm 3.2 and using binomial formulas, we obtain

$$\begin{aligned}
&(d_t\mathbf{V}^{j+1}, d_t\mathbf{U}^{j+1})_h + \frac{1}{2} d_t \| \,\nabla \mathbf{U}^{j+1} \,\|_{L^2}^2 + \frac{1}{\varepsilon} d_t F_h(\mathbf{U}^{j+1}) = 0\,,\\
&(d_t\mathbf{U}^{j+1}, d_t\mathbf{V}^{j+1})_h = \frac{1}{2} d_t \| \,\mathbf{V}^{j+1} \,\|_h^2\,.
\end{aligned}$$

Subtracting the second equation from the first one and summing up the resulting equation over the index from 1 to j gives (5.1).

Alternatively, taking $\boldsymbol{\chi} = d_t\mathbf{U}^{j+1}$ as a test function in the second equation of (ii) of Algorithm 3.2, it follows from (5.1) and the fact $F_h(\mathbf{U}^0) = 0$ if $\mathbf{U}^0 \in \boldsymbol{\mathcal{M}}_h$ that

$$\| \,d_t\mathbf{U}^{j+1} \,\|_h^2 \le \| \,\mathbf{V}^{j+\frac{1}{2}} \,\|_h^2 \le \frac{1}{2} \| \,\mathbf{V}^{j+1} \,\|_h^2 + \frac{1}{2} \| \,\mathbf{V}^{j} \,\|_h^2 \le 2E_h(\mathbf{V}^0, \mathbf{U}^0)\,.$$

The proof is complete. ∎

Likewise, we now prove that Algorithm 3.3 also satisfies a discrete conservation law which mimics the differential conservation law (1.12).

LEMMA 5.2. *Let $\varepsilon > 0, m \ge 1$, and $\big(\mathbf{U}^0, \mathbf{U}^1\big) \in [\boldsymbol{\mathcal{V}}_h]^2$. Then the solution sequence $\{\mathbf{U}^j\}$ generated by Algorithm 3.3 satisfies for $j = 1 \to (J-1)$*

$$E_h\big(d_t\mathbf{U}^{j+1}, \mathbf{U}^{j+\frac{1}{2}}\big) + \frac{1}{\varepsilon} F_h(\mathbf{U}^{j+\frac{1}{2}}) = E_h\big(d_t\mathbf{U}^{1}, \mathbf{U}^{\frac{1}{2}}\big) + \frac{1}{\varepsilon} F_h(\mathbf{U}^{\frac{1}{2}})\,. \tag{5.3}$$

Proof. Choose $\mathbf{w} = \delta_t\mathbf{U}^j$ in (ii) of Algorithm 3.3, and use the identities (cf. [14])

$$\delta_t\mathbf{U}^j = d_t\mathbf{U}^{j+\frac{1}{2}} = \frac{1}{2}\big(\,d_t\mathbf{U}^{j+1} + d_t\mathbf{U}^{j}\,\big), \qquad \mathbf{U}^{j,\frac{1}{4}} = \frac{1}{2}\big(\,\mathbf{U}^{j+\frac{1}{2}} + \mathbf{U}^{j-\frac{1}{2}}\,\big)$$

and binomial formulas to simplify the resulting equation. Assertion (5.3) follows easily from summing up the equation over the index from 1 to j. ∎

We remark that the existence of solutions to Algorithms 3.2 and 3.3 now follows from the discrete conservation laws (5.1) and (5.3), respectively, and an application of the Brouwer fixed point theorem.

5.2. Convergence of Algorithm 3.2. The same notation as that in section 4.2 will be used in this subsection. The following lemma prepares us to prove the convergence of Algorithm 3.2.

LEMMA 5.3. *In addition to the assumptions of Lemma* 5.1*, we assume* $\mathbf{U}^0 \in \boldsymbol{\mathcal{M}}_h$*; then for all* $\mathbf{w} \in C_0^\infty([0,T);\boldsymbol{\mathcal{V}}_h)$ *there holds*

$$
\begin{aligned}
(5.4)\quad &\left|\int_0^T \Big[-(\mathbf{U}_t,\mathbf{w}_t)+(\nabla\mathbf{U},\nabla\mathbf{w})+\frac{1}{\varepsilon}\big([|\,\mathbf{U}\,|^2-1]\mathbf{U},\mathbf{w}\big)\Big]\,\mathrm{d}t+\big(\mathbf{V}^0,\mathbf{w}(0,\cdot)\big)\right| \\
&\qquad\le C\,\left[h+\frac{\tau+h}{\varepsilon}\right]\left\{\int_0^T\big(\|\,\nabla\mathbf{w}\,\|_{L^2}^2+\|\,\nabla\mathbf{w}_t\,\|_{L^2}^2\big)\mathrm{d}t+\|\,\nabla\mathbf{w}(0,\cdot)\,\|_{L^2}^2\right\}^{\frac12} \\
&\qquad\qquad+\left|\int_0^T(\mathbf{V}-\overline{\mathbf{V}},\mathbf{w}_t)_h\,\mathrm{d}t\right|+\left|\int_0^T\big(\nabla[\mathbf{U}-\overline{\mathbf{U}}],\nabla\mathbf{w}\big)\,\mathrm{d}t\right|.
\end{aligned}
$$

Proof. Using the new notation and integration by parts, the equations in (ii) of Algorithm 3.2 can be rewritten as follows:

$$
\begin{aligned}
(5.5)\quad &\int_0^T\left\{-(\mathbf{V},\mathbf{w}_t)_h+(\nabla\overline{\mathbf{U}},\nabla\mathbf{w})+\frac{1}{2\varepsilon}\big([\,|\,\mathbf{U}^-\,|^2+|\,\mathbf{U}^+\,|^2-2]\overline{\mathbf{U}},\mathbf{w}\big)_h\right\}\mathrm{d}t \\
&\qquad\qquad=-\big(\mathbf{V}^0,\mathbf{w}(0,\cdot)\big)_h\,, \\
(5.6)\quad &\int_0^T(\mathbf{U}_t-\overline{\mathbf{V}},\mathbf{w})_h\,\mathrm{d}t=0\,.
\end{aligned}
$$

Using the second equation, we restate terms in the first equation as follows:

$$
\begin{aligned}
\int_0^T(\mathbf{V},\mathbf{w}_t)_h\,\mathrm{d}t&=\int_0^T(\mathbf{U}_t,\mathbf{w}_t)\,\mathrm{d}t-\int_0^T\big[(\mathbf{U}_t,\mathbf{w}_t)-(\mathbf{U}_t,\mathbf{w}_t)_h\big]\,\mathrm{d}t \\
&\qquad+\int_0^T\big(\mathbf{V}-\overline{\mathbf{V}},\mathbf{w}_t\big)_h\,\mathrm{d}t\,, \\
\int_0^T(\nabla\overline{\mathbf{U}},\nabla\mathbf{w})\,\mathrm{d}t&=\int_0^T(\nabla\mathbf{U},\nabla\mathbf{w})\,\mathrm{d}t-\int_0^T\big(\nabla[\mathbf{U}-\overline{\mathbf{U}}],\nabla\mathbf{w}\big)\,\mathrm{d}t\,, \\
\frac{1}{2\varepsilon}\int_0^T\big([\,|\,\mathbf{U}^-\,|^2+|\,\mathbf{U}^+\,|^2-2]\overline{\mathbf{U}},\mathbf{w}\big)_h\,\mathrm{d}t&-\frac{1}{\varepsilon}\int_0^T\big([|\,\mathbf{U}\,|^2-1]\mathbf{U},\mathbf{w}\big)\,\mathrm{d}t \\
&=-\frac{1}{\varepsilon}\int_0^T\Big\{\big([|\,\mathbf{U}\,|^2-1]\mathbf{U},\mathbf{w}\big)-\big([|\,\mathbf{U}\,|^2-1]\mathbf{U},\mathbf{w}\big)_h\Big\}\,\mathrm{d}t \\
&\qquad-\frac{1}{2\varepsilon}\int_0^T\big([2|\,\mathbf{U}\,|^2-|\,\mathbf{U}^-\,|^2-|\,\mathbf{U}^+\,|^2]\mathbf{U},\mathbf{w}\big)_h\,\mathrm{d}t \\
&\qquad-\frac{1}{2\varepsilon}\int_0^T\big([\,|\,\mathbf{U}^-\,|^2+|\,\mathbf{U}^+\,|^2-2](\mathbf{U}-\overline{\mathbf{U}}),\mathbf{w}\big)_h\,\mathrm{d}t\,.
\end{aligned}
$$

It follows from (3.2), the L^p-stability of I_h, and an inverse inequality that

$$\Big|\int_0^T \big[(\mathbf{V},\mathbf{w}_t)_h - (\mathbf{U}_t,\mathbf{w}_t)\big]\,\mathrm{d}t\Big| \le Ch\int_0^T \|\,\mathbf{U}_t\,\|_{L^2}\,\|\,\nabla\mathbf{w}_t\,\|_{L^2}\,\mathrm{d}t + \Big|\int_0^T (\mathbf{V}-\overline{\mathbf{V}},\mathbf{w}_t)_h\Big|,$$

$$\Big|\frac{1}{2\varepsilon}\int_0^T \Big\{\big([\,|\,\mathbf{U}^-\,|^2+|\,\mathbf{U}^+\,|^2-2]\overline{\mathbf{U}},\mathbf{w}\big)_h - 2\big([\,|\,\mathbf{U}\,|^2-1]\mathbf{U},\mathbf{w}\big)\Big\}\mathrm{d}t\Big|$$

$$\le \frac{C}{\varepsilon}\int_0^T \Big\{h\|\,I_h([\,|\,\mathbf{U}\,|^2-1]\mathbf{U})\,\|_{L^2}\|\,\nabla\mathbf{w}\,\|_{L^2} + h\|\,\nabla\big([\,|\,\mathbf{U}\,|^2-1]\mathbf{U}\big)\,\|_{L^1}\|\,\mathbf{w}\,\|_{L^\infty}$$

$$+\tau\,\|\,\mathbf{U}_t\,\|_{L^2}\|\,\mathbf{U}\,\|_{L^4}^2\|\,\mathbf{w}\,\|_{L^\infty} + \frac{\tau}{2}\,\|\,\mathbf{U}_t\,\|_{L^2}\|\,I_h(|\,\mathbf{U}^-\,|^2+|\,\mathbf{U}^+\,|^2-2)\,\|_{L^2}\|\,\mathbf{w}\,\|_{L^\infty}\Big\}\,\mathrm{d}t\,.$$

The assertion of the lemma now follows from combining the above estimates and appealing to Lemma 5.1. ∎

We are ready to state and prove the second main convergence theorem of this paper.

THEOREM 5.4. *Let $\varepsilon>0, m\ge 1, \tau_1=\tau$, and $\big(\mathbf{u}_0,\mathbf{u}_1\big)\in W^{1,2}(\Omega,\mathbf{S}^{n-1})\times L^2(\Omega,T\mathbf{S}^{n-1})$. Suppose that $\big(\mathbf{U}^0,\mathbf{V}^0\big)\in\boldsymbol{\mathcal{M}}_h\times\boldsymbol{\mathcal{V}}_h$ satisfies $\mathbf{U}^0\to\mathbf{u}_0$ strongly in $W^{1,2}(\Omega,\mathbf{R}^n)$ and $\mathbf{V}^0\to\mathbf{u}_1$ strongly in $L^2(\Omega,\mathbf{R}^n)$ as $h\to 0$. Then there exist a map $\mathbf{u}^\varepsilon\in L^\infty\big(0,T;W^{1,2}(\Omega,\mathbf{R}^n)\big)\cap W^{1,\infty}\big(0,T;L^2(\Omega,\mathbf{R}^n)\big)$ and a subsequence of the solution sequence $\{\mathbf{U}_{\varepsilon,\tau,h}\}$ (denoted by the same notation) of Algorithm 3.2 such that as $\tau,h\to 0$*

$$\begin{aligned}\mathbf{U}_{\varepsilon,\tau,h} &\overset{*}{\rightharpoonup} \mathbf{u}^\varepsilon \quad \text{in } L^\infty\big(0,T;W^{1,2}(\Omega,\mathbf{R}^n)\big)\,,\\ (\mathbf{U}_{\varepsilon,\tau,h})_t &\overset{*}{\rightharpoonup} \mathbf{u}_t^\varepsilon \quad \text{in } L^\infty\big(0,T;L^2(\Omega,\mathbf{R}^n)\big)\end{aligned}$$

uniformly in ε. Moreover, $\mathbf{u}^\varepsilon$ is a weak solution of (1.9)–(1.11). Hence, $\mathbf{u}:=\lim_{\varepsilon\to 0}\lim_{\tau,h\to 0}\mathbf{U}_{\varepsilon,\tau,h}$ exists and is a weak solution of (1.3)–(1.5), by [20].

Proof. It follows from Lemma 5.1 that there exist $\mathbf{u}^\varepsilon\in W^{1,\infty}\big(0,T;L^2(\Omega,\mathbf{R}^n)\big)\cap L^\infty\big(0,T;W^{1,2}(\Omega,\mathbf{R}^n)\big)$ and a subsequence of $\{\mathbf{U}=\mathbf{U}_{\varepsilon,\tau,h}\}$ (denoted by the same notation) such that as $\tau,h\to 0$

$$\begin{aligned}\mathbf{U},\mathbf{U}^-,\mathbf{U}^+,\overline{\mathbf{U}} &\overset{*}{\rightharpoonup} \mathbf{u}^\varepsilon \quad \text{in } L^\infty\big(0,T;W^{1,2}(\Omega,\mathbf{R}^n)\big)\,, \qquad (5.7)\\ &\to \mathbf{u}^\varepsilon \quad \text{in } L^2(\Omega_T,\mathbf{R}^n)\,,\\ \mathbf{U}_t &\overset{*}{\rightharpoonup} \mathbf{u}_t^\varepsilon \quad \text{in } L^\infty\big(0,T;L^2(\Omega,\mathbf{R}^n)\big)\,, \qquad (5.8)\end{aligned}$$

all uniformly in ε. For a.a. $\mathbf{x}\in\Omega$ and all $j=0,1,\dots,J-1$ there holds $\int_{t_j}^{t_{j+1}}(\mathbf{V}-\overline{\mathbf{V}})(\mathbf{x},t)\,\mathrm{d}t=0$; hence we can subtract the temporal average

$$\overline{\mathbf{w}}_t(x)=(1/\tau)\int_{t_j}^{t_{j+1}}\mathbf{w}_t(\mathbf{x},t)\,\mathrm{d}t$$

of w. Then by a Poincaré estimate we have

$$\begin{aligned}\left|\int_0^T(\mathbf{V}-\overline{\mathbf{V}},\mathbf{w}_t)\,\mathrm{d}t\right| &= \left|\int_0^T(\mathbf{V}-\overline{\mathbf{V}},\mathbf{w}_t-\overline{\mathbf{w}}_t)\,\mathrm{d}t\right|\\ &\le C\tau\left(\int_0^T||\mathbf{V}-\overline{\mathbf{V}}||_{L^2}^2\,\mathrm{d}t\right)^{1/2}\left(\int_0^T||\mathbf{w}_{tt}||_{L^2}^2\,\mathrm{d}t\right)^{1/2},\end{aligned}$$

and this tends to zero for $h,\tau \to 0$. Similarly, we can bound the last term in (5.4). Therefore, for any $\boldsymbol{\phi} \in C_0^\infty\big(\Omega_T, \mathbf{R}^n\big)$, let $\mathbf{w} = I_h\boldsymbol{\phi}$; using (5.1), (5.2), and the above inequality we can easily pass to the limit in (5.4) by setting $\tau, h \to 0$, and hence conclude that $\mathbf{u}^\varepsilon$ satisfies (1.9)–(1.11) in the distributional sense.

We now show that initial data are continuously taken for $t \to 0$ by the argument used at the end of the proof of Theorem 4.4. Since $\{D\mathbf{U}\}_{k,h}$ is bounded in $L^2(\Omega_T, \mathbf{R}^n)$, we have $\mathbf{U} \to \mathbf{u}^\varepsilon$ in $C^0\big(0,T; L^2(\Omega, \mathbf{R}^n)\big)$ as $\tau, h \to 0$; hence, $\mathbf{u}^\varepsilon : \Omega_T \to \mathbf{R}^n$ assumes data $\mathbf{u}_0 : \Omega \to \mathbf{R}^n$ continuously in $L^2(\Omega; \mathbf{R}^n)$. Multiply (1.9) with $\mathbf{w} \in C_0^\infty\big([0,T); C^\infty(\Omega, \mathbf{R}^n)\big)$, integrate by parts on Ω_T, and then subtract the resulting equation from (5.4); together with Lemma 5.3 and (5.7), (5.8), as well as the assumptions on the initial data $\mathbf{U}^0, \mathbf{V}^0$, we find that $\mathbf{u}_t^\varepsilon(t,\cdot) \rightharpoonup \mathbf{u}_1(\cdot)$ weakly in $L^2(\Omega, \mathbf{R}^n)$ as $t \to 0$. Moreover, by weak lower semicontinuity of the L^2-norm and Fatou's lemma we have

$$\| D\mathbf{u}^\varepsilon(t,\cdot) \|_{L^2} \le \liminf_{\tau,h\to 0} \| D\mathbf{U}(t,\cdot) \|_{L^2}\,, \quad F\big(\mathbf{u}^\varepsilon(t,\cdot)\big) \le \liminf_{\tau,h\to 0} F\big(\mathbf{U}(t,\cdot)\big).$$

Hence, for all $t \ge 0$, by (3.1), Lemma 5.1, and the assumptions on initial data,

$$\begin{aligned} E\big(\mathbf{u}_t^\varepsilon(t,\cdot), \mathbf{u}^\varepsilon(t,\cdot)\big) + \frac{1}{\varepsilon} F\big(\mathbf{u}^\varepsilon(t,\cdot)\big) &\le \liminf_{\tau,h\to 0} \left[E\big(\mathbf{U}_t(t,\cdot), \mathbf{U}(t,\cdot)\big) + \frac{1}{\varepsilon} F\big(\mathbf{U}(t,\cdot)\big) \right] \\ &\le \liminf_{\tau,h\to 0} \left[E_h\big(\mathbf{U}_t(t,\cdot), \mathbf{U}(t,\cdot)\big) + \frac{1}{\varepsilon} F_h\big(\mathbf{U}(t,\cdot)\big) \right] \\ &\le E\big(\mathbf{u}_1, \mathbf{u}_0\big)\,. \end{aligned}$$

Therefore,

$$\limsup_{t\to 0} \| \mathbf{u}_t^\varepsilon(t,\cdot) \|_{L^2} \le \| \mathbf{u}_1 \|_{L^2}\,, \quad \limsup_{t\to 0} \| \nabla\mathbf{u}^\varepsilon(t,\cdot) \|_{L^2} \le \| \nabla\mathbf{u}_0 \|_{L^2},$$

and weak convergence $\mathbf{u}_t^\varepsilon(t,\cdot) \rightharpoonup \mathbf{u}_1(\cdot)$ in $L^2(\Omega, \mathbf{R}^n)$ and $\nabla\mathbf{u}^\varepsilon(t,\cdot) \rightharpoonup \nabla\mathbf{u}_0(\cdot)$ in $L^2(\Omega)$ imply strong convergence $\mathbf{u}_t^\varepsilon(t,\cdot) \to \mathbf{u}_1(\cdot)$ and $\nabla\mathbf{u}^\varepsilon(t,\cdot) \to \nabla\mathbf{u}_0(\cdot)$ in $L^2(\Omega, \mathbf{R}^{mn})$ as $t \to 0$. Hence, $\mathbf{u}^\varepsilon : \Omega \to \mathbf{R}^n$ attains prescribed initial data continuously in $W^{1,2}(\Omega, \mathbf{R}^n) \times L^2(\Omega, \mathbf{R}^n)$ and is a weak solution of (1.9)–(1.11) in the sense of [20, Definition 5.1]. ∎

5.3. Convergence of Algorithm 3.3. Let $\{\mathbf{U}^j\}$ be the solution sequence generated by Algorithm 3.3, and define for $j = 1 \to J$

$$\mathbf{V}^j := d_t\mathbf{U}^j = \frac{1}{\tau}\big(\,\mathbf{U}^j - \mathbf{U}^{j-1}\,\big), \qquad \mathbf{Z}^j := \mathbf{U}^{j-\frac{1}{2}} = \frac{1}{2}\big(\,\mathbf{U}^{j-1} + \mathbf{U}^j\,\big).$$

We note that τ needs to be replaced by τ_1 in the definition of $\mathbf{V}^j$ when $j = 1$. Let $\mathbf{U}^+, \mathbf{U}^-, \overline{\mathbf{U}}, \mathbf{U}$ (respectively, $\mathbf{V}^+, \mathbf{V}^-, \overline{\mathbf{V}}, \mathbf{V}$ and $\mathbf{Z}^+, \mathbf{Z}^-, \overline{\mathbf{Z}}, \mathbf{Z}$) be the constant and linear interpolations of $\{\mathbf{U}^j\}$ (respectively, $\{\mathbf{V}^j\}$ and $\{\mathbf{Z}^j\}$) in time as defined in section 3.1.

Using the new notation the equation in (ii) of Algorithm 3.3 can be rewritten as for any $\mathbf{w} \in \boldsymbol{\mathcal{V}}_h$

$$\begin{aligned} &(d_t\mathbf{V}^{j+1}, \mathbf{w})_h + \big(\nabla\mathbf{Z}^{j+\frac{1}{2}}, \nabla\mathbf{w}\big) + \frac{1}{2\varepsilon}\big([\,|\mathbf{Z}^j|^2 + |\mathbf{Z}^{j+1}|^2 - 2]\mathbf{Z}^{j+\frac{1}{2}}, \mathbf{w}\big)_h = 0, \\ &(d_t\mathbf{U}^{j+1} - \mathbf{V}^{j+1}, \mathbf{w})_h = 0, \end{aligned}$$

or, equivalently, for any $\mathbf{w} \in C_0^\infty([0,T);\boldsymbol{\mathcal{V}}_h)$

$$\int_0^T \left\{ -(\mathbf{V},\mathbf{w}_t)_h + \big(\nabla\overline{\mathbf{Z}},\nabla\mathbf{w}\big) + \frac{1}{2\varepsilon}\big([\,|\mathbf{Z}^+|^2+|\mathbf{Z}^-|^2-2]\overline{\mathbf{Z}},\mathbf{w}\big)_h \right\} \mathrm{d}t = \big(\mathbf{V}^0,\mathbf{w}(0,\cdot)\big)_h, \tag{5.9}$$

$$\int_0^T (\mathbf{U}_t - \mathbf{V}^+,\mathbf{w})_h \,\mathrm{d}t = 0. \tag{5.10}$$

Comparing (5.9)–(5.10) with (5.5)–(5.6), we see that (5.9) can be obtained from (5.5) by replacing $\mathbf{Z}$ by $\mathbf{U}$, and (5.10) from (5.6) by replacing $\mathbf{V}^+$ by $\overline{\mathbf{V}}$. Moreover, the discrete conservation law (5.3) provides the same energy estimates for $(\mathbf{Z},\mathbf{V})$ here as those enjoyed by $(\mathbf{U},\mathbf{V})$ in subsection 5.2 as a result of the discrete conservation law (5.1). Hence, on noting $\mathbf{U}^j = \mathbf{Z}^j - \tau\mathbf{V}^j$ for Algorithm 3.3 and the inequality (3.3), by repeating the proof of Lemma 5.3 we easily get the following lemma.

LEMMA 5.5. *Let $\{\mathbf{U}^j\}$ be the solution sequence of Algorithm* 3.3 *and $\mathbf{V}^+,\mathbf{V},\mathbf{Z}^+$, $\mathbf{Z}^-$, and $\overline{\mathbf{Z}}$ be defined as above. In addition to the assumptions of Lemma* 5.2*, we assume $(\mathbf{U}^0,\mathbf{U}^1) \in [\boldsymbol{\mathcal{M}}_h]^2$ such that $E_h(d_t\mathbf{U}^1,\mathbf{U}^{\frac12}) + \frac{1}{\varepsilon}F_h(\mathbf{U}^{\frac12}) \le C$; then for all $\mathbf{w} \in C_0^\infty([0,T);\boldsymbol{\mathcal{V}}_h)$, there holds*

$$\begin{aligned}
&\left| \int_0^T \left[-(\mathbf{U}_t,\mathbf{w}_t) + (\nabla\mathbf{Z},\nabla\mathbf{w}) + \frac{1}{\varepsilon}\big([\,|\,\mathbf{Z}\,|^2-1]\mathbf{Z},\mathbf{w}\big)\right] \mathrm{d}t + \big(\mathbf{V}^0,\mathbf{w}(0,\cdot)\big)\right| \\
&\qquad \le C\left[h + \frac{\tau+h}{\varepsilon}\right]\left\{ \int_0^T \big(\|\,\nabla\mathbf{w}\,\|_{L^2}^2 + \|\,\nabla\mathbf{w}_t\,\|_{L^2}^2\big)\mathrm{d}t + \|\,\nabla\mathbf{w}(0,\cdot)\,\|_{L^2}^2 \right\}^{\frac12} \\
&\qquad\qquad + \left| \int_0^T (\mathbf{V}-\mathbf{V}^+,\mathbf{w}_t)_h \,\mathrm{d}t \right| + \left| \int_0^T \big(\nabla[\mathbf{Z}-\overline{\mathbf{Z}}],\nabla\mathbf{w}\big)\,\mathrm{d}t \right|.
\end{aligned} \tag{5.11}$$

We now state and show the third main convergence theorem of this paper.

THEOREM 5.6. *Let $\varepsilon > 0, m \ge 1$, and $\big(\mathbf{u}_0,\mathbf{u}_1\big) \in W^{1,2}(\Omega,\mathbf{S}^{n-1}) \times L^2(\Omega,T\mathbf{S}^{n-1})$. In addition to the assumptions of Lemma* 5.5*, suppose $\big(\mathbf{U}^0,\mathbf{U}^1\big) \in [\boldsymbol{\mathcal{M}}_h]^2$ satisfies $\mathbf{U}^0 \to \mathbf{u}_0$ strongly in $W^{1,2}(\Omega,\mathbf{R}^n)$ and $d_t\mathbf{U}^1 \to \mathbf{u}_1$ strongly in $L^2(\Omega,\mathbf{R}^n)$ as $\tau_1, h \to 0$. Then there exist a map $\mathbf{u}^\varepsilon \in L^\infty\big(0,T;W^{1,2}(\Omega,\,\mathbf{R}^n)\big) \cap W^{1,\infty}\big(0,T;L^2(\Omega,\mathbf{R}^n)\big)$ and a subsequence of the solution sequence $\{\mathbf{U}_{\varepsilon,\tau,h}\}$ (denoted by the same notation) of Algorithm* 3.3 *such that as $\tau, h \to 0$*

$$\begin{aligned}
\mathbf{U}_{\varepsilon,\tau,h} &\overset{*}{\rightharpoonup} \mathbf{u}^\varepsilon \quad \text{in } L^\infty\big(0,T;W^{1,2}(\Omega,\mathbf{R}^n)\big), \\
(\mathbf{U}_{\varepsilon,\tau,h})_t &\overset{*}{\rightharpoonup} \mathbf{u}_t^\varepsilon \quad \text{in } L^\infty\big(0,T;L^2(\Omega,\mathbf{R}^n)\big)
\end{aligned}$$

uniformly in ε. Moreover, $\mathbf{u}^\varepsilon$ is a weak solution of (1.9)–(1.11)*. Hence, $\mathbf{u} := \lim_{\varepsilon\to0}\lim_{\tau,h\to0}\mathbf{U}_{\varepsilon,\tau,h}$ exists and is a weak solution of* (1.3)–(1.5) *by* [20].

Proof. The proof is essentially the same as that of Theorem 5.4. A slight complication is that after extracting convergent subsequences of $\{\mathbf{U}\}$, $\{\mathbf{Z}\}$, and $\{\mathbf{V}\}$ (which is possible thanks to the discrete energy law (5.3)) and passing to the limit in (5.11), we now need to relate the three limiting functions $\mathbf{u}^\varepsilon, \mathbf{z}^\varepsilon$, and $\mathbf{v}^\varepsilon$ to each other.

To this end, first, on noting identity $\mathbf{U}^j = \mathbf{Z}^j - \tau\mathbf{V}^j$ (or equivalently, $\mathbf{U}^+ = \mathbf{Z}^+ - \tau\mathbf{V}^+$) and inequality (3.3), we conclude that $\mathbf{U}^+$, $\mathbf{Z}^+$, and $\mathbf{Z}$ must converge to the same limit. Hence, $\mathbf{u}^\varepsilon = \mathbf{z}^\varepsilon$. Next, it follows from identity $\mathbf{V}^j = d_t\mathbf{U}^j$ (or equivalently, $\mathbf{V}^+ = \mathbf{U}_t$) and inequality (3.3) that $\mathbf{v}^\varepsilon = \mathbf{u}_t^\varepsilon$. The proof is complete. ∎

Remark 5.1. (a) The use of Algorithm 3.3 requires choosing suitable starting values $\mathbf{U}^0$ and $\mathbf{U}^1$. It is easy to see that $\mathbf{U}^0$ can be constructed straightforwardly as the interpolation or the L^2-projection of $\mathbf{u}_0$. On the other hand, the choice of $\mathbf{U}^1$ is less obvious since U^1 must satisfy (i) $d_t\mathbf{U}^1 \to \mathbf{u}_1$ strongly in $L^2(\Omega,\mathbf{R}^n)$ as $\tau_1, h \to 0$, and (ii) $\frac{1}{\varepsilon}F_h(\mathbf{U}^{\frac{1}{2}}) \le C$. To this end, let $\mathbf{V}^0$ be the interpolation or the L^2-projection of $\mathbf{u}_1$, and define $\mathbf{U}^1 := \mathbf{U}^0 + \tau_1\mathbf{V}^0$. Clearly, requirement (i) is fulfilled. Since $\mathbf{U}^{\frac{1}{2}} = \mathbf{U}^0 + \frac{\tau_1}{2}\mathbf{V}^0$, then $|\mathbf{U}^{\frac{1}{2}}| = 1 + O(\tau_1)$ and $\frac{1}{\varepsilon}F_h(\mathbf{U}^{\frac{1}{2}}) = O(\frac{\tau_1^2}{\varepsilon})$. Hence, requirement (ii) is fulfilled provided that $\tau_1 = O(\varepsilon^{\frac{1}{2}})$. We also note that if $\mathbf{u}_0 \cdot \mathbf{u}_1 = 0$, $\mathbf{V}^0$ above can be chosen such that $\mathbf{U}^0 \cdot \mathbf{V}^0 = 0$; in this case we have $\frac{1}{\varepsilon}F_h(\mathbf{U}^{\frac{1}{2}}) \le C$ provided that $\tau_1 = O(\varepsilon^{\frac{1}{4}})$.

(b) It is interesting to note that Algorithm 3.3 is stable (and converges) uniformly in ε provided that $\tau_1 = O(\varepsilon^{\frac{1}{2}})$, which means that the *first time step* must be chosen very small. On the other hand, no restriction is imposed on h and the later time step τ for stability. Hence, as expected, the algorithm is absolutely stable.

6. Numerical experiments. Besides local existence results, the only global result is a unique smooth solution for all times, provided that the initial energy is sufficiently small [20]. In order to obtain global existence without the assumption of small energy, it would be sufficient to show that the energy cannot concentrate at the hypothetical singularity. Similar to [8], our results for the same initial data show a concentration effect in the case of large initial data at a finite time $T > 0$. All of our numerical results were obtained with MATLAB using the direct linear solver provided by MATLAB's backslash operator. Constraints were included through Lagrange multipliers, and the assembly of matrices was realized in C.

6.1. Consistent discretization versus penalized approximation. In our first numerical experiment we report the performance of Algorithms 3.1 and 3.2 on an example that leads to large gradients. In particular, we follow [6] and employ, for $\mathbf{x} = \big(x_1, x_2\big) \in \mathbf{R}^2$, $r := |\mathbf{x}|$, and $a(r) := (1-2r)^4$, the pair of initial data

$$\mathbf{u}_0(\mathbf{x}) := \begin{cases} \big(0,0,-1\big) & \text{for } r \ge 1/2, \\ \big(2x_1a, 2x_2a, a^2 - r^2\big)/\big(a^2+r^2\big) & \text{for } r \le 1/2, \end{cases} \qquad \mathbf{u}_1(\mathbf{x}) := \mathbf{0}.$$

The vector field $\mathbf{u}_0$ wraps about the unit sphere once and is therefore expected to lead to large gradients. We use a uniform triangulation of $\Omega = (-1/2, 1/2)^2$ into 2048 triangles which are squares halved along the direction $(1,1)$ and with maximal diameter $h = \sqrt{2}\,2^{-5}$. The time-step size that we use is $\tau = h^{2.1}$, where the additional power 0.1 is motivated by the condition $\tau = o(h^2)$. The upper plot in Figure 1 displays the $W^{1,\infty}$ seminorm

$$|\mathbf{U}(t)|_{1,\infty} = ||\nabla\mathbf{U}(t)||_{L^\infty(\Omega)}$$

as a function of $t \in [0,2]$ and with $\mathbf{U}$ obtained from Algorithm 3.1 (indicated as "consistent") and Algorithm 3.2 for various choices of ε. In addition, we display in the lower plot of Figure 1 the discrete energies

$$E_h\big(\mathbf{U}(t)\big) = E_h\big(\mathbf{V}^+(t), \mathbf{U}^+(t)\big)$$

and

$$E_h^\varepsilon\big(\mathbf{U}(t)\big) = E_h\big(\mathbf{V}^+(t), \mathbf{U}^+(t)\big) + \frac{1}{\varepsilon}F_h\big(\mathbf{U}^+(t)\big)$$

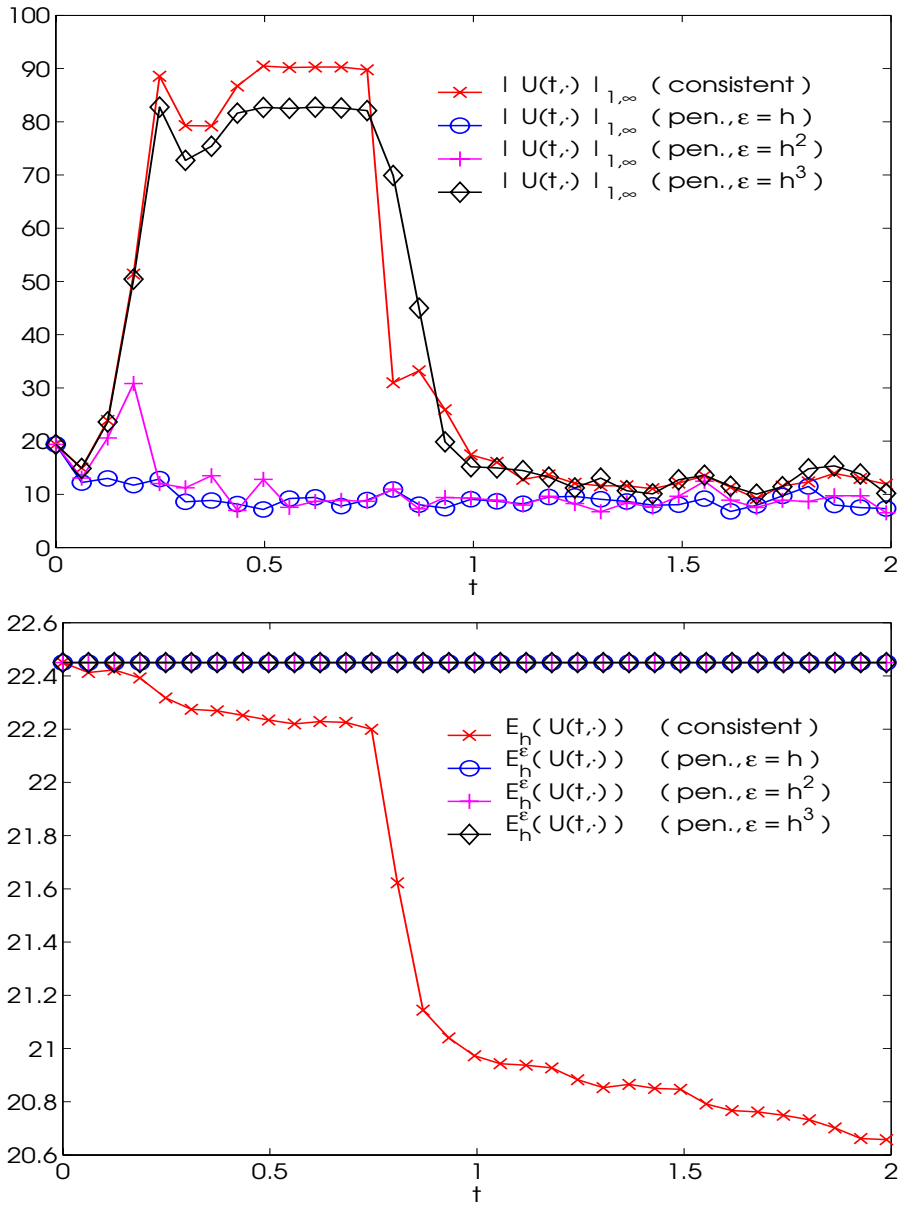


FIG. 1. *$W^{1,\infty}$ seminorm (upper plot) and discrete energy (lower plot) as functions of $t \in [0,2]$ for approximations obtained with Algorithms* 3.1 *and* 3.2 *for various choices of ε. The plots for the results obtained with the penalization method almost coincide in the lower plot.*

for the approximate solutions obtained with Algorithm 3.1 and Algorithm 3.2, respectively. We observe that the $W^{1,\infty}$ seminorm attains its maximum value

$$\max_{\mathbf{W}\in\boldsymbol{\mathcal{V}}_h;\,\forall \mathbf{q}_i\in\mathcal{N}_h,\,|\mathbf{W}(\mathbf{q}_i)|=1} ||\nabla \mathbf{W}||_{L^\infty(\Omega)} = 2^{6.5} \approx 90.51$$

among functions $\mathbf{W} \in \boldsymbol{\mathcal{V}}_h$ that satisfy $|\mathbf{W}(\mathbf{q}_i)| = 1$ for all nodes $\mathbf{q}_i \in \mathcal{N}_h$ for the solution obtained with Algorithm 3.1 at $t \approx 0.25$. Its energy $E_h\big(\mathbf{U}(t)\big)$ slightly de-

creases at the time of blow-up; however, the fact that it is not increasing indicates the robustness of our direct method. The numerical results differ significantly when the approximation is obtained with Algorithm 3.2 and when different values for ε are used. While the energies E_h^ε obtained with the unconstrained method appear to be constant, we observe that even for the extremely small choice $\varepsilon = h^3$ the $W^{1,\infty}$ seminorm of the numerical approximation remains bounded away from the maximum value. We note that such a choice of ε is critical since Lemma 5.3 suggests that convergence of the numerical approximations obtained by Algorithm 3.2 can be expected only if $\tau, h = o(\varepsilon)$. Nevertheless, it seems that the results obtained with the penalization approach those obtained with Algorithm 3.1 when ε becomes small.

6.2. Numerical evidence for finite-time blow-up and occurrence of singular solutions. Even though Algorithms 3.1 and 3.2 are mathematically equivalent in the sense that they both provide subsequences that converge weakly to a solution of the wave map equation, we believe that Algorithm 3.1 is better suited to studying the occurrence of singularities since the corresponding topological effects are not diffused as in the regularized (penalized) approximation scheme of Algorithm 3.2. The two plots in Figure 2 display the behavior of numerical solutions obtained with Algorithm 3.1 for decreasing mesh-sizes and with the initial data described above. As in the previous subsection we employed uniform triangulations of $\Omega = (-1/2, 1/2)^2$ into squares halved along the direction $(1, 1)$. The time-step size was defined through $\tau = h^{2.1}$ for $h = \sqrt{2}\,2^{-5}$, $\sqrt{2}\,2^{-6}$, $\sqrt{2}\,2^{-7}$ corresponding to a sequence of triangulations with 2048, 8192, and 32768 triangles, respectively. Figure 2 shows that for each triangulation the maximal $W^{1,\infty}$ seminorm is attained. This indicates that an existing exact solution for the choice of our initial data has unbounded $W^{1,\infty}$ norm, i.e., is a singular solution, but we stress that this behavior might change for sufficiently small mesh-sizes. This so-called finite-time blow-up behavior occurs in our example when the solution at the origin points in a direction different from all the surrounding vectors. We illustrate this behavior in Figures 3 and 4, where the first two components of the vector field $\mathbf{U}(t, \mathbf{x})$ for the nodes $\mathbf{q}_i \in (-1/4, 1/4)^2 \cap \mathcal{N}_h$ and the vectors in a neighborhood of the origin are shown, respectively, for the solution obtained with the triangulation that consists of 2048 triangles. After a small period of large (maximal) gradients, the vector at the origin changes its direction. This change of direction may be unstable and mesh-dependent. It is not clear to the authors whether such an effect also occurs in the continuous situation. We note that in this experiment it was not possible to recover the initial data by reversing time. The results in the following subsection indicate that the quantitative behavior does not change significantly when other triangulations with more symmetry are used.

6.3. Effect of symmetric/asymmetric triangulations. In order to obtain a better understanding of the mechanism that forces the vector at the origin to change its direction after the occurrence of a maximal $W^{1,\infty}$ norm, we applied Algorithm 3.1 to the initial data given above for different spatial triangulations of comparable mesh-size. The four different triangulations we used are displayed in Figure 5. The first one (upper left plot) is obtained from four uniform refinements of a coarse triangulation $\mathcal{T}_0$ of $\Omega = (-1/2, 1/2)$ into two triangles. The next triangulation (upper right plot) resulted from four uniform refinements of that coarse triangulation of Ω that is obtained from dividing Ω into four similar triangles by dividing it along its diagonals. The third triangulation shown (lower left plot) is obtained from the previous one by a random perturbation of those nodes that do not belong to the boundary $\partial\Omega$ by a value smaller than $h/10$ in each direction. Finally, we also employed an approximation of

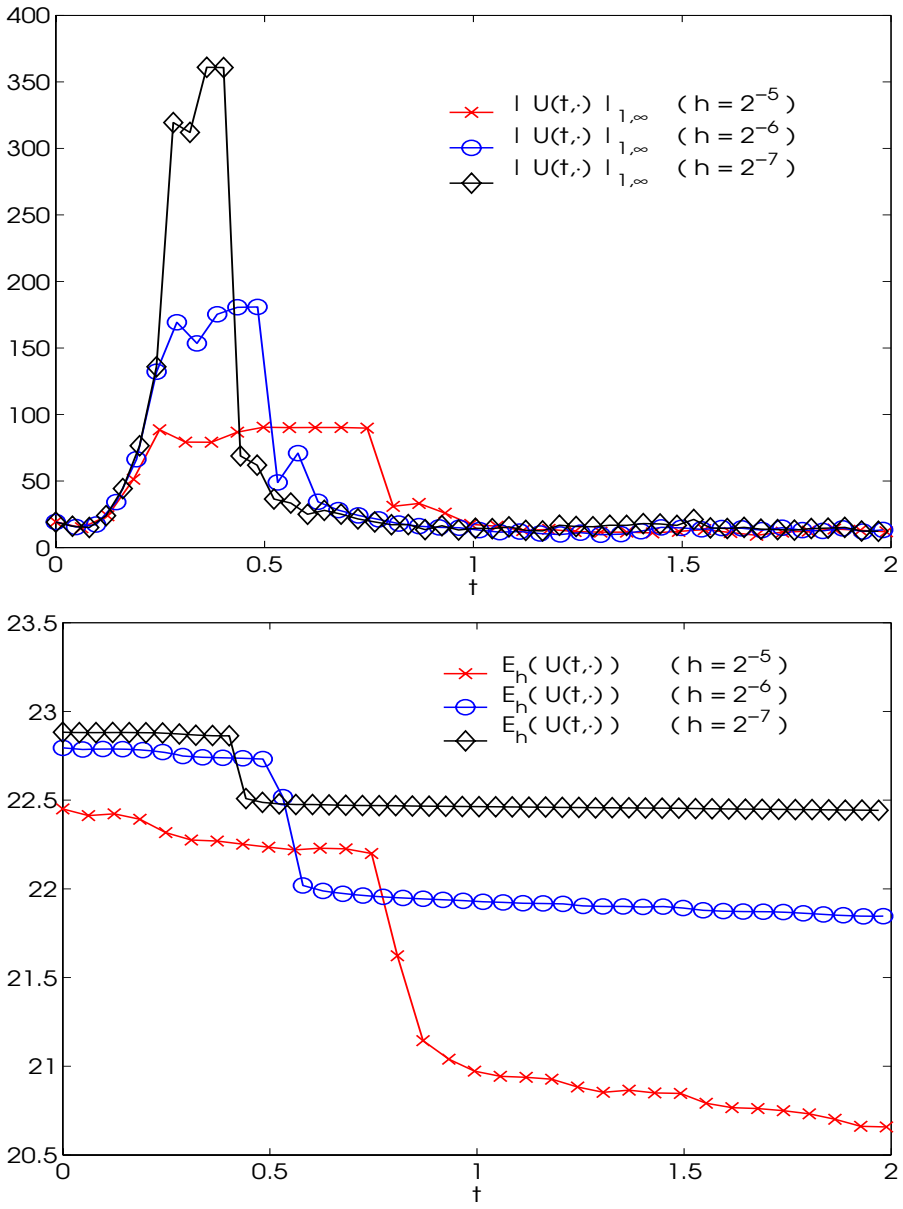


FIG. 2. *$W^{1,\infty}$ seminorm (upper plot) and discrete energy (lower plot) as functions of $t \in [0,2]$ for approximations obtained with Algorithm* 3.1 *on refining triangulations.*

the disk $\Omega = \{\mathbf{x} \in \mathbf{R}^2 : |\mathbf{x}| \leq 1/2\}$; cf. the last triangulation in Figure 5 (lower right plot). Figure 6 displays the $W^{1,\infty}$ seminorms of the numerical approximations using the four different triangulations. The results differ mostly after the occurrence of discrete blow-up, and we expect that this mesh-dependence reflects a loss of stability and a possibility of nonuniqueness of exact solutions. However, we observe that before the occurrence of maximal gradients, quantitatively the behavior of the numerical approximations does not change significantly when more symmetry is introduced in

the grid or when symmetry is entirely destroyed by perturbing the positions of the vertices. We thus believe that our Algorithm 3.1 provides an accurate approximation of the exact solution and is capable of detecting singularities in a reliable way.

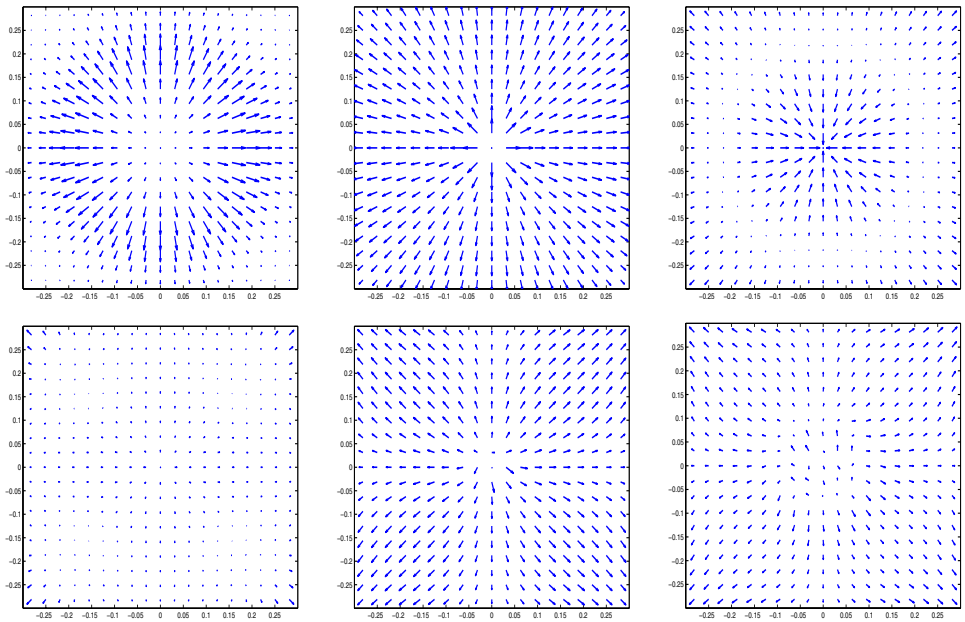


FIG. 3. *Scaled first two components of* $\mathbf{U}(t,\cdot)$ *for the solution provided by Algorithm* 3.1 *for* $t = 0$, 0.20025, 0.40051, 0.60076, 0.80102, 1.00130 *(from left to right, from top to bottom).*

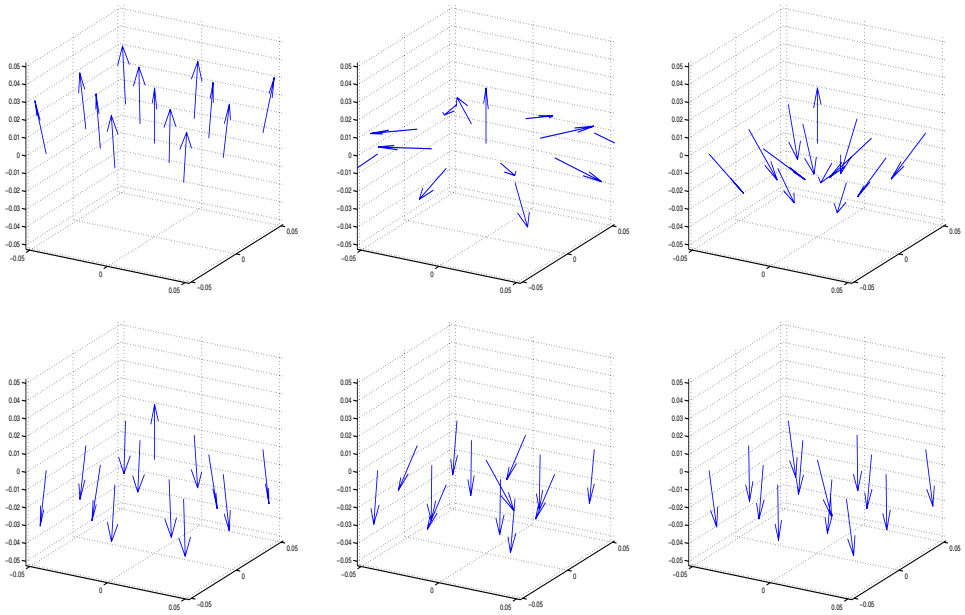


FIG. 4. *Scaled solution* $\mathbf{U}(t,\cdot)$ *obtained with Algorithm* 3.1 *in a neighborhood of the origin for* $t = 0$, 0.20025, 0.40051, 0.60076, 0.80102, 1.00130 *(from left to right, from top to bottom) on a criss-cross triangulation.*

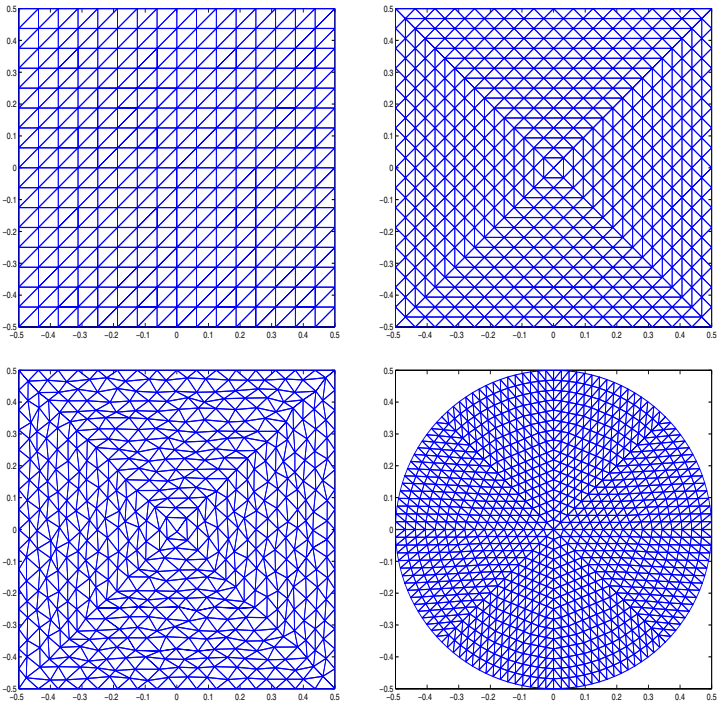


FIG. 5. *Triangulations with different symmetry properties.*

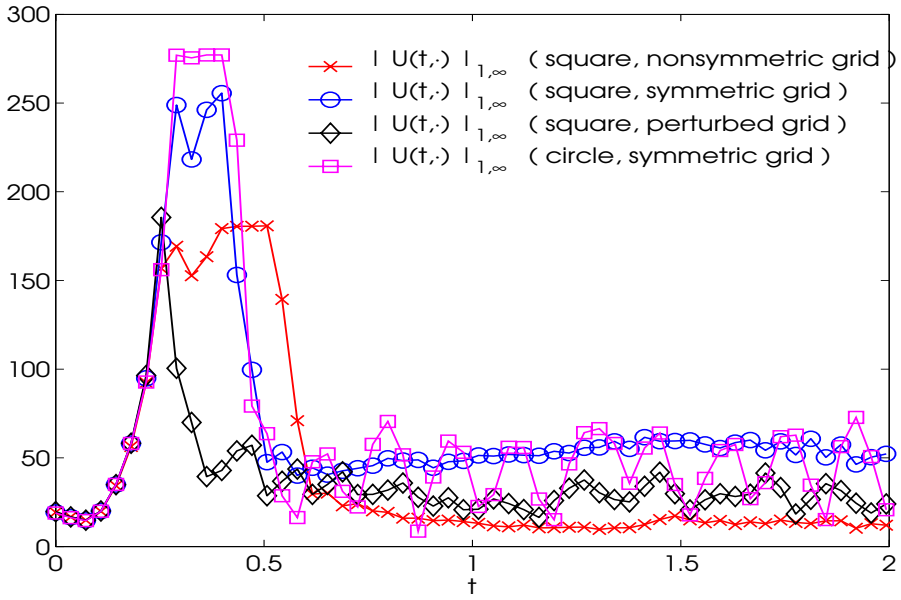


FIG. 6. *$W^{1,\infty}$ seminorm and discrete energy as functions of $t \in [0,2]$ for approximations obtained with Algorithm* 3.1 *for (refinements of) the triangulations shown in Figure* 5.

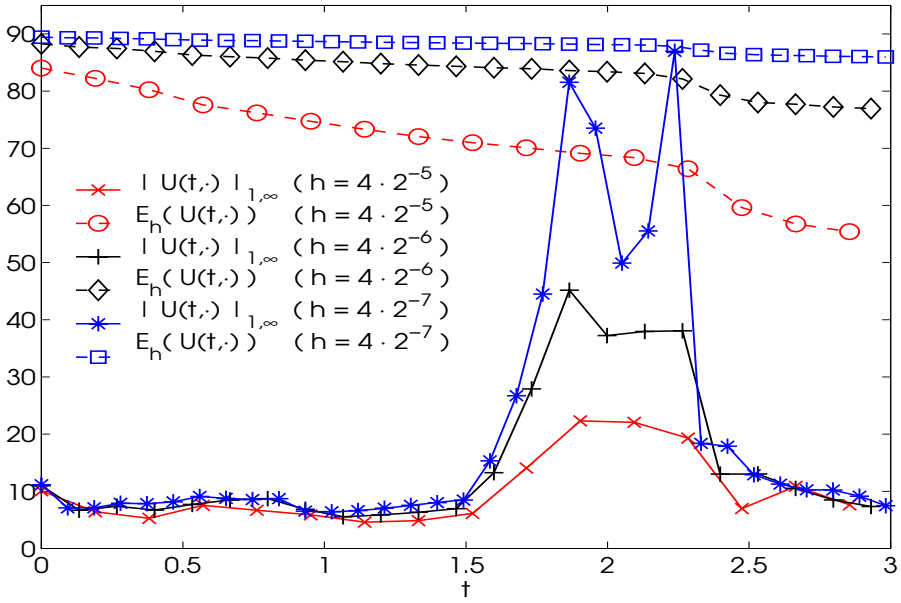


FIG. 7. *$W^{1,\infty}$ seminorm and discrete energy as functions of $t \in [0,3]$ for approximations obtained with Algorithms 3.1 for equivariant initial data on refining triangulations.*

6.4. Equivariant initial data. In our last series of experiments we follow the work of [8] and employ equivariant initial data defined by

$$\mathbf{u}_0(r,\theta) = \begin{pmatrix} \sin\chi(r,\theta)\,\sin\theta \\ \sin\chi(r,\theta)\,\cos\theta \\ \cos\chi(r,\theta) \end{pmatrix}, \qquad \mathbf{u}_1(r,\theta) \equiv 0,$$

where (r,θ) denotes polar coordinates in $\mathbf{R}^2$ and $\chi(r,\theta) = (r^3/4)\exp(-(4(r-2)/10)^4)$. We ran Algorithm 3.1 with these initial data on uniform triangulations of $\Omega = (-2,2)^2$ with 2048, 8192, and 32768 triangles. Figure 7 displays the $W^{1,\infty}$ seminorms and the energy as functions of $t \in [0,3]$ obtained with the three different triangulations. As in [8] we observe finite-time blow-up. We remark that in [8] the results were obtained by reducing the problem to a one-dimensional model that guarantees that the solution remains an equivariant map for all times. Here, this is not included as a constraint in the approximation scheme. Nevertheless, the sequence of snapshots of the numerical solution shown in Figure 8 for $t \approx 0, 2, 4, \ldots, 10$ indicates that this is true for our numerical solution. In particular, the vector at the origin does not change its direction on triangulations that obey certain symmetry properties. In Figure 9 we display the third component of the solution at the origin as a function of $t \in [0,10]$, i.e., the real number $\mathbf{U}^{(3)}(t,\mathbf{0})$, for the numerical solution obtained on the different triangulations described above (scaled by a factor of 4 for this experiment) and indicated in Figure 9 (for simplicity, we did not perturb the vertex located at the origin). For the unperturbed triangulations the solution vector at the origin does not change for $t \in (0,10)$, while for the perturbed triangulation the vector changes its direction for a small period of time when the blow-up occurs. These results are in good agreement with theoretical predictions stating that for equivariant data the vector at the origin remains constant [24]. Finally, we remark that in this experiment

(on the nonsymmetric, unperturbed triangulation) it was possible to reverse time and to reproduce the initial data.

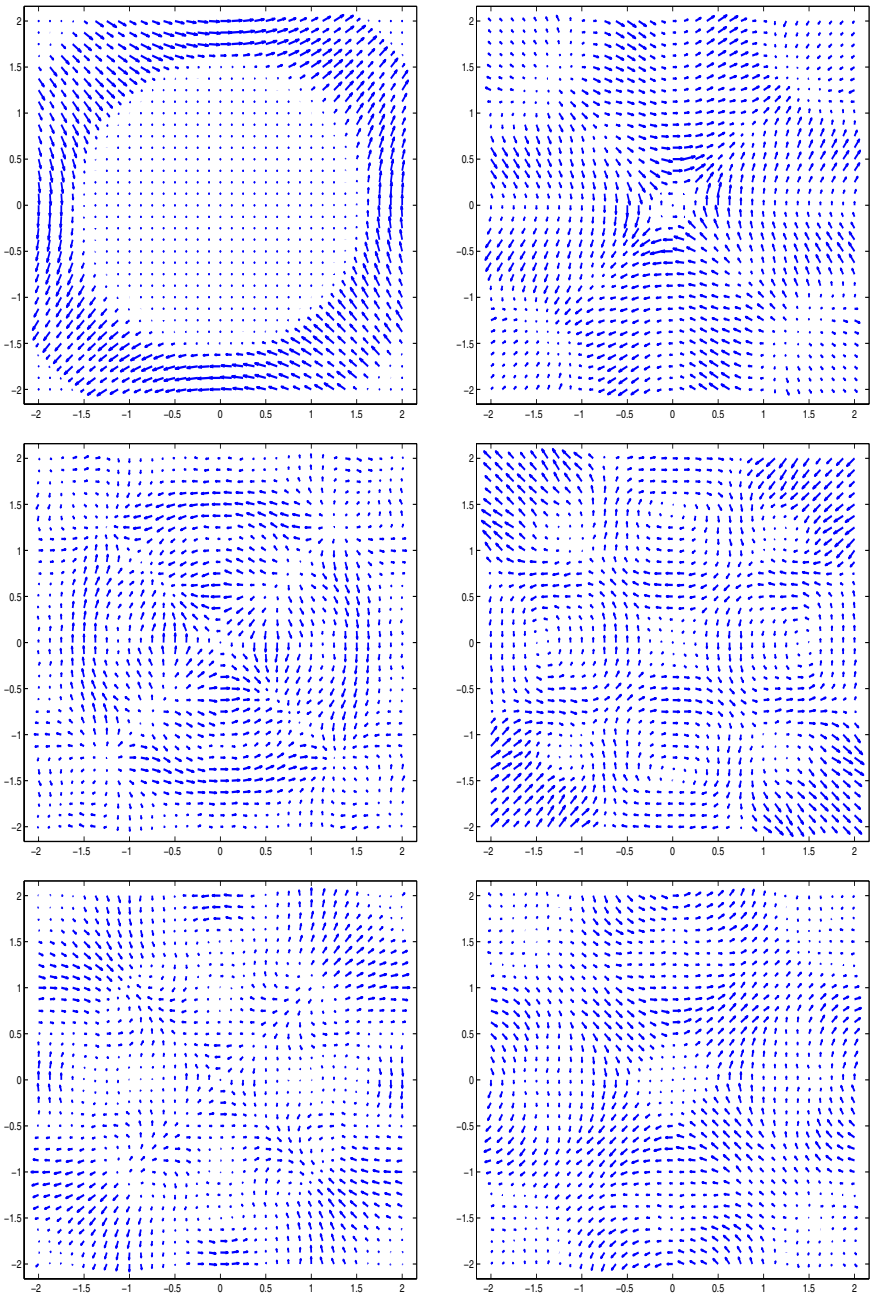


FIG. 8. *Scaled first two components of* $\mathbf{U}(t,\cdot)$ *for the solution provided by Algorithm* 3.1 *for equivariant initial data for* $t \approx 0$, 2.0, 4.0, 6.0, 8.0, 10.0 *(from left to right, from top to bottom).*

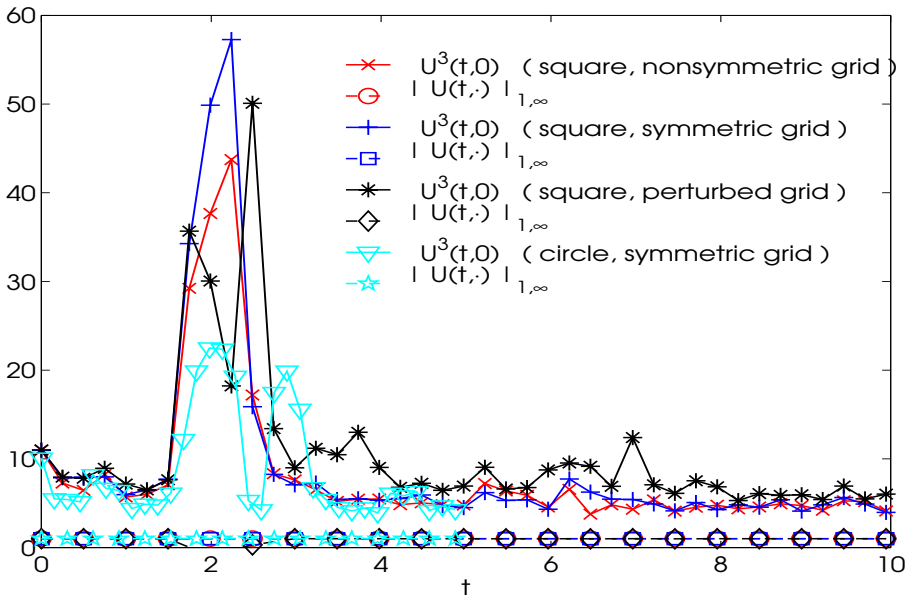


FIG. 9. $W^{1,\infty}$ *seminorm and third component of numerical solutions at origin of* $t \in [0,10]$ *for approximations obtained with Algorithm* 3.1 *for (refinements of) the triangulations shown in Figure* 5 *and equivariant initial data.*

REFERENCES

[1] I. J. R. AITCHISON AND A. J. G. HEY, *Gauge Theories in Particle Physics, Volume* I: *A Practical Introduction: From Relativistic Quantum Mechanics to QED*, Graduate Student Series in Physics, Taylor & Francis, London, 2002.

[2] F. ALOUGES AND P. JAISSON, *Convergence of a finite element discretization for the Landau Lifshitz equations*, Math. Models Methods Appl. Sci., 16 (2006), pp. 299–316.

[3] G. A. BAKER, *Error estimates for finite element methods for second order hyperbolic equations,* SIAM J. Numer. Anal., 13 (1976), pp. 564–576.

[4] J. W. BARRETT, S. BARTELS, X. FENG, AND A. PROHL, *A convergent and constraint-preserving finite element method for the p-harmonic flow into spheres*, SIAM J. Numer. Anal., 45 (2007), pp. 905–927.

[5] J. W. BARRETT, X. FENG, AND A. PROHL, *p-harmonic map heat flows for* $1 \le p < \infty$ *and their finite element approximations with applications to image processing*, SIAM J. Math. Anal., submitted.

[6] S. BARTELS, J. KO, AND A. PROHL, *Numerical approximation of the Landau–Lifshitz–Gilbert equation and finite time blow-up of weak solutions*, SIAM J. Numer. Anal., submitted.

[7] B. K. BERGER, P. T. CHRUSCIEL, AND V. MONCRIEF, *On "asymptotically flat" space-times with* G_2*-invariant Cauchy surfaces*, Ann. Physics, 237 (1995), pp. 322–354.

[8] P. BIZON, T. CHMAJ, AND Z. TABOR, *Formation of singularities for equivariant* $2+1$ *dimensional wave maps into two-sphere*, Nonlinearity, 14 (2001), pp. 1041–1053.

[9] S. C. BRENNER AND L. R. SCOTT, *The Mathematical Theory of Finite Element Methods*, Springer-Verlag, New York, 1994.

[10] S. M. CARROLL, *Spacetime and Geometry: An Introduction to General Relativity*, Addison–Wesley, Reading, MA, 2003.

[11] T.-P. CHENG AND L.-F. LI, *Gauge Theory of Elementary Particle Physics*, Oxford University Press, Oxford, UK, 1983.

[12] P. G. CIARLET, *Basic error estimates for elliptic problems*, in Handbook of Numerical Analysis, Vol. II, North–Holland, Amsterdam, 1991, pp. 17–351.

[13] T. DUPONT, L^2*-estimates for Galerkin methods for second order hyperbolic equations*, SIAM J. Numer. Anal., 10 (1973), pp. 880–889.

[14] X. Feng, *Analysis of finite element methods and domain decomposition algorithms for a fluid-solid interaction problem*, SIAM J. Numer. Anal., 38 (2000), pp. 1312–1336.
[15] A. Freire, *Global weak solutions of the wave map system to compact homogeneous spaces*, Manuscripta Math., 91 (1996), pp. 525–533.
[16] M. Giaquinta and S. Hildebrandt, *Calculus of Variations* I, Springer-Verlag, New York, 1996.
[17] J. Isenberg and S. L. Liebling, *Singularity formation in* $2+1$ *wave maps*, J. Math. Phys., 43 (2002), pp. 678–683.
[18] S. Müller and M. Struwe, *Spatially discrete wave maps on* $(1+2)$*-dimensional space-time*, Topol. Methods Nonlinear Anal., 11 (1998), pp. 295–320.
[19] J. Shatah, *Weak solutions and development of singularities in the* $SU(2)$ σ*-model*, Comm. Pure Appl. Math., 41 (1998), pp. 459–469.
[20] J. Shatah and M. Struwe, *Geometric Wave Equations*, New York University, Courant Institute of Mathematical Sciences, New York; AMS, Providence, RI, 1998.
[21] J. Shatah and M. Struwe, *The Cauchy problem for wave maps*, Int. Math. Res. Not., 11 (2002), pp. 555–571.
[22] M. Struwe, *Geometric evolution problems*, in Nonlinear Partial Differential Equations in Differential Geometry, IAS/Park City Math. Ser. 2, AMS, Providence, RI, 1996, pp. 257–339.
[23] M. Struwe, *Recent existence and regularity results for wave maps*, in Journées "Équations aux Dérivées Partielles" (Saint-Jean-de-Monts, 1997), Exp. No. XVII, 7 pp., École Polytech., Palaiseau, France, 1997.
[24] M. Struwe, *Equivariant wave maps in two space dimensions*, Comm. Pure Appl. Math., 56 (2003), pp. 815–823.
[25] D. Tataru, *The wave maps equation*, Bull. Amer. Math. Soc. (N.S.), 41 (2004), pp. 185–204.

SIAM J. NUMER. ANAL.
Vol. 46, No. 1, pp. 88–102
© 2007 Society for Industrial and Applied Mathematics

LOWER ORDER RECTANGULAR NONCONFORMING MIXED FINITE ELEMENTS FOR PLANE ELASTICITY*

JUN HU† AND ZHONG-CI SHI‡

Abstract. In this paper, we present two stable rectangular nonconforming mixed finite element methods for the equations of linear elasticity in two space dimensions which produce direct approximations for the stress and displacement. In the first method, the normal stress space of the matrix-valued stress space is taken as the second order rotated Brezzi–Douglas–Fortin–Marini element space [F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer-Verlag, New York, 1991], the enriched nonconforming rotated Q_1 element [Q. Lin, L. Tobiska, and A. H. Zhou, *IMA J. Numer. Anal.*, 25 (2005), pp. 160–181] is taken for the shear stress, and the lowest order Raviart–Thomas element space [P. A. Raviart and J. M. Thomas, in *Mathematical Aspects of the Finite Element Method*, Lecture Notes in Math. 606, Springer-Verlag, New York, 1977, pp. 292–315] is employed to approximate the vector displacement field. The second method is obtained from the first one through dropping the interior degrees of the normal stress on each element. A first order convergence rate is obtained for both the stress and the displacement for these methods based on the superconvergence of the enriched nonconforming rotated Q_1 element.

Key words. mixed method, nonconforming finite element, elasticity, rectangular

AMS subject classifications. 65N30, 65N15, 35J25

DOI. 10.1137/060669681

1. Introduction.

We consider the finite element approximation of a stress-displacement system derived from the Hellinger–Reissner variational principle for the plane linear elasticity, which reads: Given $f \in L^2(\Omega, \mathbb{R}^2)$ find $(\sigma, u) \in H(\mathrm{div}, \Omega, \mathbb{S}) \times L^2(\Omega, \mathbb{R}^2)$ with

$$(A\sigma, \tau)_{L^2(\Omega)} + (\mathrm{div}\,\tau, u)_{L^2(\Omega)} = 0,$$

$$(\mathrm{div}\,\sigma, v)_{L^2(\Omega)} = (f, v)_{L^2(\Omega)}$$

for any $(\tau, v) \in H(\mathrm{div}, \Omega, \mathbb{S}) \times L^2(\Omega, \mathbb{R}^2)$. Here and throughout this paper, the compliance tensor $A(x) : \mathbb{S} \to \mathbb{S}$ is bounded and symmetric positive definite uniformly for $x \in \Omega$ with $\mathbb{S} := \mathbb{R}^{2\times 2}_{\mathrm{sym}}$ the set of symmetric tensors. The space $H(\mathrm{div}, \Omega, \mathbb{S})$ is defined as

$$H(\mathrm{div}, \Omega, \mathbb{S}) := \{\tau \in L^2(\Omega, \mathbb{S}) \text{ and } \mathrm{div}\,\tau \in L^2(\Omega, \mathbb{R}^2)\}$$

equipped with the norm

$$\|\tau\|^2_{H(\mathrm{div},\Omega)} := \|\tau\|^2_{L^2(\Omega)} + \|\mathrm{div}\,\tau\|^2_{L^2(\Omega)}.$$

*Received by the editors September 12, 2006; accepted for publication (in revised form) June 12, 2007; published electronically December 19, 2007.
http://www.siam.org/journals/sinum/46-1/66968.html

†LMAM and School of Mathematical Sciences, Peking University, Beijing 100871, People's Republic of China (hujun@math.pku.edu.cn). This author's research was supported by the NSFC under grant 10601003 and sponsored by the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry, and the Foundation for the Author of National Excellent Doctoral Dissertation of the People's Republic of China 200718.

‡No 55, Zhong-Guan-Cun Dong Lu, Institute of Computational Mathematics, Chinese Academy of Sciences, Beijing 100080, People's Republic of China (shi@lsec.cc.ac.cn).

Because of the symmetry requirement imposed on the discrete matrix stress spaces and the usual stable conditions for the discrete stress spaces and vector displacement spaces from the theory of mixed finite element methods [4], it becomes very difficult to construct simple mixed elements. As D. N. Arnold pointed out, in his plenary address to the 2002 International Congress of Mathematicians, four decades of searching for mixed finite elements for elasticity did not yield any stable elements with polynomial shape functions. In recent years, there has been some progress in this direction. In [2], a family of stable conforming triangular elements are presented with the simplest and lowest order element taking 21 stress and 3 displacement degrees of freedom on each triangle. In order to keep the conformity, the vertex degrees of freedom are in particular employed in those methods. To avoid the complex of conforming mixed element and also vertex degrees of freedom, they proposed a first order nonconforming triangular mixed element with 12 degrees of freedom for the stress and 3 degrees of freedom for the displacement [3]. The conforming rectangular mixed elements are proposed and analyzed in [1] with the lowest order element needing 45 stress and 12 displacement degrees of freedom. In a recent paper [12], a nonconforming mixed finite element based on rectangular meshes is proposed with 19 stress and 6 displacement degrees of freedom on each element.

In this paper, we present two first order nonconforming rectangular mixed elements. In the first method, it takes 15 stress and 4 displacement degrees of freedom on each rectangle. The total degrees of freedom is $3|E|+7|K|$ with $|E|$ the number of edges and $|K|$ the number of elements of the partition $\mathcal{T}_h$. More precisely, we use the second order rotated Brezzi–Douglas–Fortin–Marini (hereafter BDFM) element space [4] to approximate the normal stress of the matrix stress field, the enriched nonconforming rotated Q_1 element [9] is taken for the shear stress, and the lowest order Raviart–Thomas (hereafter RT) element space [11] with 4 degrees on each element is taken for the displacement space. The second element is obtained from the first one by dropping the interior degrees of freedom of both diagonal components on each element and keeping the vector displacement space. Therefore, the total degrees of freedom becomes $3|E|+5|K|$. We stress that it is the simplest element so far. Based on the superconvergence, we analyze and obtain the optimal error estimate, i.e., the first order convergence rate. We also show that the discrete analogue of the following exact sequence:

(1.1)
$$0 \longrightarrow P_1(\Omega) \xrightarrow{\subset} C^\infty(\Omega) \xrightarrow{J} C^\infty(\Omega,\mathbb{S}) \xrightarrow{\mathrm{div}} C^\infty(\Omega,\mathbb{R}^2) \longrightarrow 0$$

holds for both elements with the enriched and original nonconforming Adini elements, respectively. Here, and throughout this paper, the Airy stress function corresponding to a scalar field q reads

$$Jq := \begin{pmatrix} \frac{\partial^2 q}{\partial y^2} & -\frac{\partial^2 q}{\partial x \partial y} \\ -\frac{\partial^2 q}{\partial x \partial y} & \frac{\partial^2 q}{\partial x^2} \end{pmatrix}.$$

This paper is organized as follows. In the following two sections, we shall present the new mixed elements and analyze their properties including the well-posedness. In section 4, we prove the discrete analogue of (1.1) for both elements. The error analysis will be investigated in section 5.

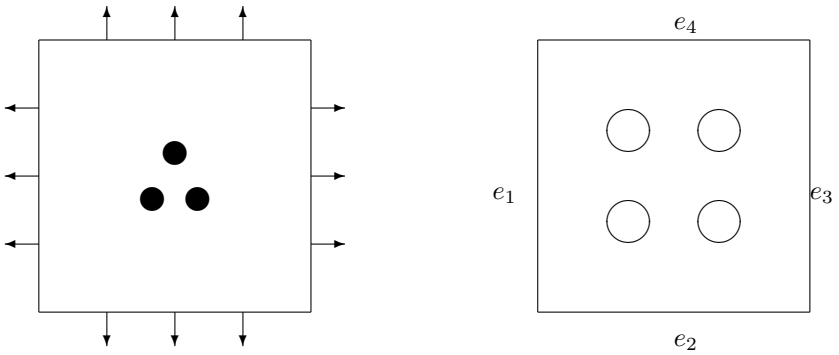


FIG. 1. *Element diagram for the stress and displacement.*

2. Nonconforming mixed finite element approximation. For approximating problem (1.1) by the finite element method, we introduce a rectangular triangulation $\mathcal{T}_h$ of the rectangular domain Ω; the regularity of the mesh $\mathcal{T}_h$ is assumed in the sense of Ciarlet [5] such that $\bigcup_{K\in\mathcal{T}_h} K = \bar{\Omega}$, two distinct elements K and K' in $\mathcal{T}_h$ are either disjoint or share the common edge e or a common vertex. Let $\mathcal{E}$ denote the set of all edges in $\mathcal{T}_h$ with $\mathcal{E}_{K,V}$ the two vertical edges of $K \in \mathcal{T}_h$ and $\mathcal{E}_{K,H}$ the two horizontal edges. Given any edge $e \in \mathcal{E}$ we assign one fixed unit normal n with (n_1, n_2) its components; we also let $t = (-n_2, n_1)$ denote the tangential vector.

For each $K \in \mathcal{T}_h$, we introduce the following affine invertible transformation:

$$F_K : \hat{K} \to K, \quad x = \frac{h_{x,K}}{2}\xi + x_{0,K}, \quad y = \frac{h_{y,K}}{2}\eta + y_{0,K}$$

with the center $(x_{0,K}, y_{0,K})$, the horizontal and vertical edge lengths $h_{x,K}$ and $h_{y,K}$, respectively, and the reference element $\hat{K} = [-1, 1]^2$. Given any integer k, let $P_k(\omega)$ denote the space of polynomials over ω of total degrees not greater than k.

The key idea in the designing of mixed finite elements for the elasticity is to construct the polynomial space pair (Σ_K, V_k) with the following conditions.

(A1) $\operatorname{div}\Sigma_K \subset V_K$.

(A2) There exists a bounded interpolation operator $\Pi_K : H^1(K, \mathbb{S}) \to \Sigma_K$ with the following commuting diagram property:

$$\operatorname{div}\Pi_K\sigma = P_K \operatorname{div}\sigma \quad \forall\sigma \in H^1(K, \mathbb{S})$$

with P_K the projection operator from $L^2(K, \mathbb{R}^2)$ to V_K.

These two conditions in turn ensure that there exists a unique critical point and provide a good approximation of the true solution.

On the element K, we choose the following RT element space [11] as the displacement space:

$$V_K := \begin{pmatrix} a + by \\ c + dx \end{pmatrix}.$$

As depicted in Figure 1, we define the space for the stress as

$$\Sigma_K^0 := \begin{pmatrix} P_2^1(K) & P_2^2(K) \\ P_2^2(K) & P_2^3(K) \end{pmatrix}_S, \tag{2.1}$$

the space of symmetric fields $\sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} \in \mathbb{S}$ such that $\sigma_{11} \in P_2^1(K)$, $\sigma_{12} = \sigma_{21} \in P_2^2(K)$, and $\sigma_{22} \in P_2^3(K)$. Here

$$P_2^1(K) := \operatorname{span}\{1, x, y, xy, y^2\},$$

$$P_2^2(K) := \operatorname{span}\{1, x, y, x^2, y^2\},$$

$$P_2^3(K) := \operatorname{span}\{1, x, y, xy, x^2\}.$$

A straightforward investigation gives that the dimension of Σ_K^0 is 15.

Remark 2.1. Note that the normal stress space $(P_2^1(K), P_2^3(K))$ is the second order rotated BDFM element space for second order elliptic problems [4], and the shear space $P_2^2(K)$ is the enriched nonconforming rotated Q_1 element space [9].

The degrees of freedom for Σ_K^0 follow

- the moments of degrees 0 and 1 on the four edges of K for $\sigma n \cdot n$,
- the moments of degree 0 on the four edges of K for $\sigma n \cdot t$,
- the moment of degree 0 on K for σ_{12},
- the values of $\int_K (y - y_{0,K})^2 \sigma_{11} dxdy$ and $\int_K (x - x_{0,K})^2 \sigma_{22} dxdy$.

LEMMA 2.2. *If the* 15 *degrees of freedom vanish for some* $\sigma \in \Sigma_K^0$, *then* $\sigma = 0$.

Proof. The existence and uniqueness for the component $\sigma_{12} = \sigma_{21}$ follow directly from [9]; we consider only those for σ_{11} and σ_{22}. Without loss of generality, we prove them on the reference element $\hat{K}$. In this case, σ_{11} can be written as

$$\sigma_{11} = a_0 + a_1\xi + a_2\eta + a_3\xi\eta + a_4\eta^2$$

with some interpolation constants a_i, $i = 0, \ldots, 4$. Since $\int_{e \in \mathcal{E}_{\hat{K},V}} \sigma_{11} q d\eta = 0$ for any $q \in P_1(e)$, we can obtain by a direct calculation that

$$a_0 - a_1 + \frac{1}{3}a_4 = 0, \quad a_2 - a_3 = 0,$$

$$a_0 + a_1 + \frac{1}{3}a_4 = 0, \quad a_2 + a_3 = 0,$$

$$\frac{1}{3}a_0 + \frac{1}{5}a_4 = 0,$$

which yield $\sigma_{11} = 0$. A similar argument gives $\sigma_{22} = 0$. ∎

Dropping the interior degrees of σ_{11} and σ_{22}, we can get one subspace of Σ_K^0 as follows:

$$\Sigma_K^1 := \begin{pmatrix} Q_1(K) & P_2^2(K) \\ P_2^2(K) & Q_1(K) \end{pmatrix}_S,$$

the space of symmetric fields $\sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} \in \mathbb{S}$ such that $\sigma_{11} \in Q_1(K)$, $\sigma_{12} = \sigma_{21} \in P_2^2(K)$, and $\sigma_{22} \in Q_1(K)$. See Figure 2. Therefore, the dimension of Σ_K^1 is 13. Here

$$Q_1(K) := \operatorname{span}\{1, x, y, xy\}.$$

The degrees of freedom for Σ_K^1 follow

- the moments of degrees 0 and 1 on the four edges of K for $\sigma n \cdot n$,
- the moments of degree 0 on the four edges of K for $\sigma n \cdot t$,

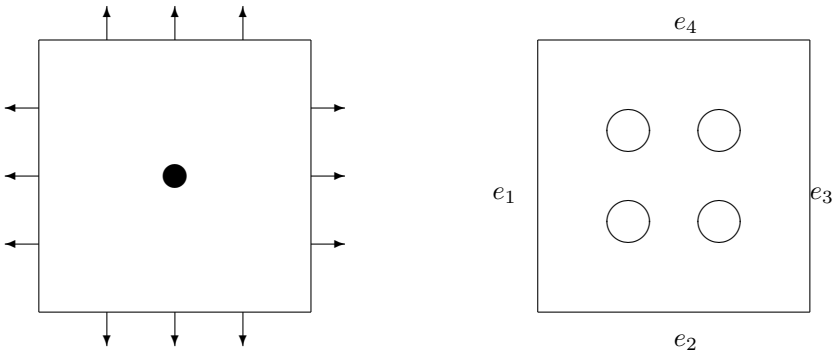


FIG. 2. *Element diagram for the stress and displacement.*

- the moment of degree 0 on K for σ_{12}.

A similar argument of Lemma 2.2 shows the following.

LEMMA 2.3. *If the* 13 *degrees of freedom vanish for some* $\sigma \in \Sigma_K^0$, *then* $\sigma = 0$.

For $i = 0, 1$, define the discrete spaces for the stress and displacement as, respectively,

$$\Sigma_h^i := \left\{ \sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} \in L^2(\Omega, \mathbb{S}), \sigma|_K \in \Sigma_K^i \text{ and the moments of degrees 0 and 1 on all the internal edges are continuous for } \sigma n \cdot n, \text{ and the moments of degree 0 on all the internal edges are continuous for } \sigma n \cdot t, \forall K \right\},$$

$$V_h := \{v \in L^2(\Omega, \mathbb{R}^2), v|_K \in V_K, \forall K\}.$$

LEMMA 2.4. *For* Σ_h^1, *we note from the definition that* σ_{11} *is continuous on interior vertical edges,* σ_{22} *continuous on interior horizontal edges, and* $\sigma_{12} = \sigma_{21}$ *continuous in the averaging sense on all interior edges.*

Proof. The proof immediately follows from the definitions. ∎

Remark 2.5. It follows from Lemma 2.4 that Σ_h^1 is conforming-nonconforming elements with $\Sigma_h^1 \not\subseteq H(\mathrm{div}, \Omega, \mathbb{S})$, which is a very good feature.

Since $\Sigma_h^i \not\subseteq H(\mathrm{div}, \Omega, \mathbb{S})$ with $i = 0, 1$, we denote by div_h the divergence operator defined elementwise with respect to $\mathcal{T}_h$, i.e.,

$$\mathrm{div}_h \tau|_K = \mathrm{div}(\tau|_K) \quad \text{for any } \tau \in \Sigma_h^i,\ i = 0, 1.$$

With these preparations, our nonconforming mixed element methods can be stated as follows: Find $(\sigma_h, u_h) \in \Sigma_h^i \times V_h$ such that

$$\begin{aligned} (A\sigma_h, \tau)_{L^2(\Omega)} + (\mathrm{div}_h \tau, u_h)_{L^2(\Omega)} &= 0, \\ (\mathrm{div}_h \sigma_h, v)_{L^2(\Omega)} &= (f, v)_{L^2(\Omega)} \end{aligned} \tag{2.2}$$

for any $(\tau, v) \in \Sigma_h^i \times V_h$ with $i = 0, 1$.

3. Well-posedness of the discrete problem. In this section, we consider the well-posedness of the discrete problem (2.2) for $i = 0, 1$, which in turn hangs on the following two assumptions.

1. K-ellipticity. There exists a constant $C > 0$ independent of the meshsize such that
$$(A\tau, \tau)_{L^2(\Omega)} \geq C\|\tau\|^2_{H(\mathrm{div}_h,\Omega)}$$
for any
$$\tau \in Z_h^i := \{\tau \in \Sigma_h^i | (\mathrm{div}_h\, \tau, v) = 0 \quad \forall v \in V_h\}.$$
2. Discrete B-B condition. There exists a positive constant C independent of the meshsize with
$$\sup_{\tau\in\Sigma_h^i} \frac{(\mathrm{div}_h\, \tau, v)}{\|\tau\|_{H(\mathrm{div}_h,\Omega)}} \geq C\|v\|_{L^2(\Omega)} \quad \text{for any } v \in V_h.$$

We first consider Σ_h^1, and to this end, we define for any $\sigma = \begin{pmatrix}\sigma_{11} & \sigma_{12}\\ \sigma_{21} & \sigma_{22}\end{pmatrix}$ the interpolation $\Pi_h^1\sigma \in \Sigma_h^1$ with

$$\int_e \Pi_h^1\sigma n \cdot nvds = \int_e \sigma n \cdot nvds \quad \text{for any edge } e \text{ and for any } v \in P_1(e),$$

$$\int_e \Pi_h^1\sigma n \cdot tds = \int_e \sigma n \cdot tds \quad \text{for any edge } e,$$

$$\int_K (\Pi_h^1\sigma)_{12}dxdy = \int_K \sigma_{12}dxdy \quad \text{for any } K \in \mathcal{T}_h.$$

It is not difficult to see that Π_h^1 is well defined. We infer from the definition of the interpolation operator Π_h^1 that

$$(\mathrm{div}_h(\sigma - \Pi_h^1\sigma), w_h)_{L^2(\Omega)} = 0 \quad \text{for any } w_h \in V_h.$$

In fact, an integration by parts yields that

$$(\mathrm{div}_h(\sigma - \Pi_h^1\sigma), w_h)_{L^2(\Omega)} = \sum_{K\in\mathcal{T}_h}\sum_{e\subset\partial K}\int_e (\sigma - \Pi_h^1\sigma)n_e \cdot w_h ds = 0. \tag{3.1}$$

Let B denote the Jacobian matrix for the affine mapping F_K introduced in section 2. Given $\hat{\tau} : \hat{K} \to \mathbb{S}$, define $\tau : K \to \mathbb{S}$ by the matrix Piola transformation,

$$\tau(x, y) = B\hat{\tau}(\xi, \eta)B^T \tag{3.2}$$

with $(x, y) = F_K(\xi, \eta)$. A direct calculation yields that

$$\mathrm{div}\,\tau = B\,\mathrm{div}\,\hat{\tau}. \tag{3.3}$$

It follows that $\tau \in H(\mathrm{div}, K, \mathbb{S})$ if and only if $\hat{\tau} \in H(\mathrm{div}, \hat{K}, \mathbb{S})$. Since B is a diagonal matrix, one claims that $\tau \in \Sigma_K^i$ with $i = 0, 1$ if and only if $\hat{\tau} \in \Sigma_{\hat{K}}^i$.

Let $\Pi_K^1 : H(\mathrm{div}, K, \mathbb{S}) \to \Sigma_K^1$ denote the restriction of Π_h^1 onto a single rectangle. We assert that

$$\Pi_K^1\tau = B\Pi_{\hat{K}}^1\hat{\tau}B^T. \tag{3.4}$$

Indeed, $n = \hat{n}$ and $B^T n = \|B^T n\|\hat{n}$; this leads to

$$\int_e B\Pi^1_{\hat{K}}\hat{\tau}B^T n \cdot n v ds = \frac{|e|}{2} B \int_{\hat{e}} (\Pi^1_{\hat{K}}\hat{\tau})B^T\hat{n} \cdot \hat{n}\hat{v}d\hat{s} = \frac{|e|}{2}\|B^T n\| B \int_{\hat{e}} \Pi^1_{\hat{K}}\hat{\tau}\hat{n} \cdot \hat{n}\hat{v}d\hat{s}$$

$$= \frac{|e|}{2}\|B^T n\| B \int_{\hat{e}} \hat{\tau}\hat{n} \cdot \hat{n}\hat{v}d\hat{s} = \int_e \tau n \cdot n v ds$$

for any edges and for any $v \in P_1(e)$. Therefore, we have the following properties for the finite element method in consideration.

LEMMA 3.1. *It holds that*

(1) $\operatorname{div}_h \Pi^1_h \sigma = P_h \operatorname{div}\sigma$, $\|\Pi^1_h\sigma\|_{L^2(\Omega)} \le C\|\sigma\|_{H^1(\Omega)}$ *for any* $\sigma \in H^1(\Omega, \mathbb{S})$, *where* $P_h : L^2(\Omega, \mathbb{R}^2) \to V_h$ *denotes the* L^2 *projection;*

(2) $\operatorname{div}_h \Sigma^1_h = V_h$.

Proof. It follows from (3.1) that $\operatorname{div}_h \Pi^1_h\sigma = P_h \operatorname{div}\sigma$. Since $\Pi^1_{\hat{K}}$ is bounded from $H^1(\hat{K}, \mathbb{S})$ to $L^2(\hat{K}, \mathbb{S})$, a scaling argument using (3.4) proves that

$$\|\Pi^2_K\tau\|_{L^2(K)} \le C(\|\tau\|_{L^2(K)} + h\|\tau\|_{H^1(K)}). \tag{3.5}$$

Summing over the elements shows (1).

It is immediate that $\operatorname{div}_h \Sigma^1_h \subset V_h$; we now prove the converse. Indeed, for any $v \in V_h$, there exists $\tau \in H^1(\Omega, \mathbb{S})$ such that

$$\operatorname{div}\tau = v, \|\tau\|_{H^1(\Omega)} \le C\|v\|_{L^2(\Omega)}.$$

With this choice we have $\Pi^1_h\tau \in \Sigma^1_h$, $\operatorname{div}\Pi^1_h\tau = P_h \operatorname{div}\tau = P_h v = v$, which proves that $V_h \subset \operatorname{div}_h \Sigma^1_h$. ∎

THEOREM 3.2. *K-ellipticity and the discrete B-B conditions hold uniformly in the meshsize; therefore problem* (2.2) *admits a unique solution* $(\sigma_h, u_h) \in \Sigma^1_h \times V_h$.

Proof. It follows from (2) of Lemma 3.1 that $\operatorname{div}_h \tau = 0$ for all $\tau \in Z^1_h$, which implies the K-ellipticity.

It remains to show the discrete B-B condition. For any $v \in V_h$, there exists $\tau \in H^1(\Omega, \mathbb{S})$ such that

$$\operatorname{div}\tau = v, \|\tau\|_{H^1(\Omega)} \le C_0\|v\|_{L^2(\Omega)}.$$

With this choice we have $\Pi^1_h\tau \in \Sigma^1_h$, $\operatorname{div}\Pi^1_h\tau = P_h \operatorname{div}\tau = P_h v = v$. This leads to

$$(\operatorname{div}_h \Pi^1_h\tau, v)_{L^2(\Omega)} = (P_h \operatorname{div}\tau, v)_{L^2(\Omega)} = (P_h v, v)_{L^2(\Omega)} \ge 1/C_0\|v\|_{L^2(\Omega)}\|\tau\|_{H^1(\Omega)}.$$

Thanks to (1) of Lemma 3.1, we conclude that

$$(\operatorname{div}_h \Pi^1_h\tau, v)_{L^2(\Omega)} \ge C\|\Pi^1_h\tau\|_{H(\operatorname{div}_h,\Omega)}\|v\|_{L^2(\Omega)},$$

which yields the desired result. ∎

Remark 3.3. One can define a similar interpolation operator Π^0_h for the space Σ^0_h. Moreover, Lemma 3.1 and Theorem 3.2 hold equally for the space pair (Σ^0_h, V_h).

4. The discrete exact sequence of differential complex. In this section, we investigate the relations between the elements in section 2 and some nonconforming elements for the fourth order elliptic problems. We show the discrete exact sequence of differential complex for Σ^0_h and Σ^1_h.

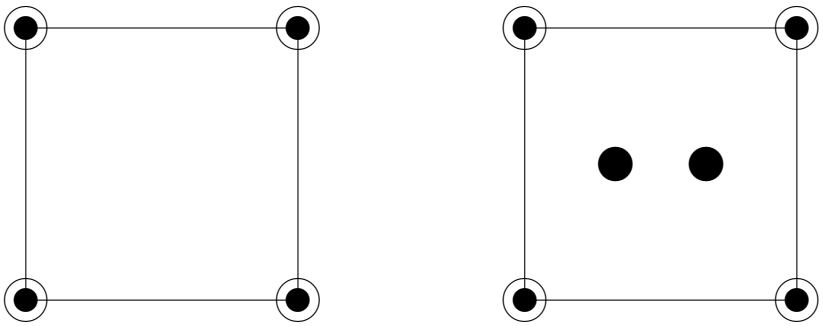


FIG. 3. *The Adini element and enriched Adini element.*

The second element Σ_h^1 is closely related to the nonconforming Adini element (see Figure 3) for the fourth order elliptic problems. In fact, the piecewise divergence-free elements in Σ_h^1 can be explicitly obtained from the Adini element on each element K. We show this observation in the following. For any $K \in \mathcal{T}_h$, the Adini element with dimension 12 reads as

$$AD(K) := P_3(K) \oplus \text{span}\{x^3y, y^3x\}.$$

Any element q in $AD(K)$ can be uniquely determined by the values of q and the two first order derivatives of q at each vertex (12 degrees of freedom). The associated finite element space AD_h is a subspace of $H^1(\Omega)$, but not a subspace of $H^2(\Omega)$ due to the discontinuity of the normal derivatives. However, the normal derivatives are weak continuous since they are continuous at nodes of the partition $\mathcal{T}_h$.

LEMMA 4.1. *It holds that*

$$J_h q \in \Sigma_h^1 \quad \textit{and} \quad \text{div}_h\, J_h q = 0 \quad \forall q \in AD_h.$$

Proof. Let J_K and q_K denote the restriction of J_h and q on K, respectively; then it is immediate that

$$J_K q_K \in \Sigma_K^1.$$

Therefore, we need only to prove that $J_h q$ satisfies the continuous conditions for the space Σ_h^1. For any vertical edge e with $n = (\pm 1, 0)$ and two endpoints V_1 and V_2, we have

$$\begin{aligned}\int_e \frac{\partial_h^2 q}{\partial y^2} v dy &= -\int_e \frac{\partial q}{\partial y}\frac{dv}{dy} dy + \left(\frac{\partial q}{\partial y} v\right)(V_2) - \left(\frac{\partial q}{\partial y} v\right)(V_1) \\ &= -\left(q\frac{dv}{dy}\right)(V_2) + \left(q\frac{dv}{dy}\right)(V_1) + \left(\frac{\partial q}{\partial y} v\right)(V_2) - \left(\frac{\partial q}{\partial y} v\right)(V_1)\end{aligned}$$

for any $v \in P_1(e)$. This, together with the definition of the Adini element space, shows the continuity of $\frac{\partial_h^2 q}{\partial y^2}$. The same argument and result hold equally for $\frac{\partial_h^2 q}{\partial x^2}$. It remains to prove that $\frac{\partial_h^2 q}{\partial y \partial x} = \frac{\partial_h^2 q}{\partial x \partial y}$ is continuous in the average sense on all interior

edges. In fact,

$$\int_e \frac{\partial_h^2 q}{\partial y \partial x} dy = \frac{\partial q}{\partial x}(V_2) - \frac{\partial q}{\partial x}(V_1) \tag{4.1}$$

for any vertical edge e. The corresponding continuity follows from that of the Adini element. The second assertion of $\mathrm{div}_h\, J_h q = 0$ is obvious. □

Define the interpolation $I_h^1 q \in AD_h$ for any $C^2(\Omega)$ by requiring that the values of $I_h^1 q$ and the first order derivatives of $I_h^1 q$ at vertexes all be equal to the corresponding values of q.

LEMMA 4.2. *There holds the commutativity property*

$$J_h I_h^1 q = \Pi_h^1 J q \quad \forall q \in C^2(\Omega).$$

Proof. Since $J_h I_h^1 q \in \Sigma_h^1$ and $\Pi_h^1 J q \in \Sigma_h^1$, we need only to establish the following identities:

$$\int_e (J_h I_h^1 q) n \cdot n v ds = \int_e J q n \cdot n v ds, \quad v \in P_1(e)\ \forall \text{ edges } e,$$

$$\int_e (J_h I_h^1 q) n \cdot t ds = \int_e J q n \cdot t ds \quad \forall \text{ edges } e,$$

$$\int_K \frac{\partial_h^2 I_h^1 q}{\partial x \partial y} dx dy = \int_K \frac{\partial^2 q}{\partial x \partial y} dx dy \quad \forall \text{ elements } K.$$

By the symmetry, we show only the case where $n = (1, 0)$ for the first and second identities in the following. As before, let V_1 and V_2 denote the bottom and top endpoints of e. Integrating by parts twice and taking into account that v is linear on e, the left-hand side of the first identity becomes

$$\int_e \frac{\partial_h^2 I_h^1 q}{\partial y^2} v dy = \left(\frac{\partial I_h^1 q}{\partial y} v\right)(V_2) - \left(\frac{\partial I_h^1 q}{\partial y} v\right)(V_1) - \left(I_h^1 q \frac{dv}{dy}\right)(V_2) + \left(I_h^1 q \frac{dv}{dy}\right)(V_1).$$

The right-hand side of the first identity becomes

$$\int_e \frac{\partial^2 q}{\partial y^2} v dy = \left(\frac{\partial q}{\partial y} v\right)(V_2) - \left(\frac{\partial q}{\partial y} v\right)(V_1) - \left(q \frac{dv}{dy}\right)(V_2) + \left(q \frac{dv}{dy}\right)(V_1).$$

These two identities and the definition of I_h^1 prove the first identity. The proof of the second identity is a direct consequence of the integration by parts with the definition of I_h^1. Now we establish the third identity. An integration by parts leads to

$$\begin{aligned}\int_K \frac{\partial_h^2 I_h^1 q}{\partial x \partial y} dx dy &= \int_{e_3} \frac{\partial I_h^1 q}{\partial y} dy - \int_{e_1} \frac{\partial I_h^1 q}{\partial y} dy \\ &= (I_h^1 q)(V_2^{e_3}) - (I_h^1 q)(V_1^{e_3}) - (I_h^1 q)(V_2^{e_1}) + (I_h^1 q)(V_1^{e_1}),\end{aligned}$$

$$\int_K \frac{\partial^2 q}{\partial x \partial y} dx dy = q(V_2^{e_3}) - q(V_1^{e_3}) - q(V_2^{e_1}) + q(V_1^{e_1}).$$

An application of the definition of I_h^1 completes the proof. □

With the aforementioned commutativity property, we have the following commuting diagram with exact rows:

$$\begin{array}{ccccccccccc}
0 & \longrightarrow & P_1(\Omega) & \xrightarrow{\subset} & C^\infty(\Omega) & \xrightarrow{J} & C^\infty(\Omega,\mathbb{S}) & \xrightarrow{\mathrm{div}} & C^\infty(\Omega,\mathbb{R}^2) & \longrightarrow & 0 \\
 & & \downarrow{\scriptstyle id} & & \downarrow{\scriptstyle I_h^1} & & \downarrow{\scriptstyle \Pi_h^1} & & \downarrow{\scriptstyle P_h} & & \\
0 & \longrightarrow & P_1(\Omega) & \xrightarrow{\subset} & AD_h & \xrightarrow{J_h} & \Sigma_h^1 & \xrightarrow{\mathrm{div}_h} & V_h & \longrightarrow & 0
\end{array}$$

To obtain the analogue for the first element space Σ_h^0, we define the following enriched Adini element:

$$EAD(K) := P_3(K) \oplus \mathrm{span}\{x^3y, y^3x, x^4, y^4\}.$$

Any element q in $EAD(K)$ can be uniquely determined by

- the values of q and the two first order derivatives of q at each node (12 degrees of freedom),
- the values of $\int_K (x - x_{0,K})^2 \frac{\partial_h^2 I_h^0 q}{\partial x^2} dxdy$ and $\int_K (y - y_{0,K})^2 \frac{\partial_h^2 I_h^0 q}{\partial y^2} dxdy$ (two degrees of freedom).

A variant of the argument for Lemma 4.1 shows the following.

LEMMA 4.3. *It holds that*

$$J_h q \in \Sigma_h^0 \quad \textit{and} \quad \mathrm{div}_h J_h q = 0 \quad \forall q \in EAD_h.$$

Define the interpolation $I_h^0 q \in EAD_h$ for any $C^2(\Omega)$ by requiring that the values of $I_h^0 q$, the first order derivatives of $I_h^0 q$ at vertexes, $\int_K (x - x_{0,K})^2 \frac{\partial_h^2 I_h^0 q}{\partial x^2} dxdy$, $\int_K (y - y_{0,K})^2 \frac{\partial_h^2 I_h^0 q}{\partial y^2} dxdy$, all be equal to the corresponding values of q.

LEMMA 4.4. *There holds the commutativity property*

$$J_h I_h^0 q = \Pi_h^0 J q \quad \forall q \in C^2(\Omega).$$

Proof. Thanks to the proof of Lemma 4.2, the proof of Lemma 4.4 is the direct consequence of the definitions of I_h^0 and Π_h^0. ∎

Therefore, we have the following commuting diagram with exact rows:

$$\begin{array}{ccccccccccc}
0 & \longrightarrow & P_1(\Omega) & \xrightarrow{\subset} & C^\infty(\Omega) & \xrightarrow{J} & C^\infty(\Omega,\mathbb{S}) & \xrightarrow{\mathrm{div}} & C^\infty(\Omega,\mathbb{R}^2) & \longrightarrow & 0 \\
 & & \downarrow{\scriptstyle id} & & \downarrow{\scriptstyle I_h^0} & & \downarrow{\scriptstyle \Pi_h^0} & & \downarrow{\scriptstyle P_h} & & \\
0 & \longrightarrow & P_1(\Omega) & \xrightarrow{\subset} & EAD_h & \xrightarrow{J_h} & \Sigma_h^0 & \xrightarrow{\mathrm{div}_h} & V_h & \longrightarrow & 0
\end{array}$$

5. Error analysis. This section presents the error analysis for the methods in consideration. We first collect the following error estimates for the interpolation operators Π_h^i, with $i = 0, 1$, and P_h, which can be shown using the standard scaling arguments and the properties of Π_h^i, $i = 0, 1$:

$$\|\Pi_h^i \tau - \tau\|_{L^2(\Omega)} \le Ch^m \|\tau\|_{H^m(\Omega)}, \qquad m = 1, 2,\ i = 0, 1, \tag{5.1}$$

$$\|\,\mathrm{div}_h \Pi_h^i \tau - \mathrm{div}\,\tau\|_{L^2(\Omega)} \le Ch \|\,\mathrm{div}\,\tau\|_{H^1(\Omega)}, \qquad i = 0, 1, \tag{5.2}$$

$$\|P_h v - v\|_{L^2(\Omega)} \le Ch \|v\|_{H^1(\Omega)}. \tag{5.3}$$

For the analysis, we need some further result concerning the superconvergence for the enriched nonconforming rotated Q_1 element which reads

$$NR_h := \{v \in L^2(\Omega), v|_K \in P_2^2(K) \text{ for any } K \in \mathcal{T}_h, \text{ and } v \text{ is continuous in the averaging sense on all the interior edges of } \mathcal{T}_h\}.$$

Given $\tau \in \Sigma_h^0$, it follows from the definition of Σ_h^0 that $\tau_{12} \in NR_h$.

LEMMA 5.1. *It holds for any $u \in H^2(\Omega, \mathbb{R}^2) \cap H_0^1(\Omega, \mathbb{R}^2)$ that*

$$\sum_{K\in\mathcal{T}_h} \int_{\partial K} u \cdot nv ds \le Ch^2 \|u\|_{H^2(\Omega)} |v|_{1,h} \quad \forall v \in NR_h,$$

where $n = (n_1, n_2)$ is the outer normal vector to ∂K. The broken seminorm $|\cdot|_{1,h}$ is defined as

$$|v|_{1,h}^2 := \sum_{K\in\mathcal{T}_h} \|\nabla v\|_{L^2(K)}^2.$$

Proof. For this rectangular case, a similar proof can be found in [10, 8, 6, 7, 9]. For the convenience of readers, we sketch the proof here. With $\mathcal{J}_e v = \frac{1}{|e|}\int_e v ds$ for any $v \in L^2(e)$, we have

$$\sum_{K\in\mathcal{T}_h} \int_{\partial K} u \cdot nv ds = \sum_{K\in\mathcal{T}_h} \sum_{e\subset\partial K} \int_e (u - \mathcal{J}_e u) \cdot n(v - \mathcal{J}_e v) ds$$

$$= \sum_{K\in\mathcal{T}_h} \sum_{i=1}^4 \int_{e_i} (u - \mathcal{J}_{e_i} u) \cdot n(v - \mathcal{J}_{e_i} v) ds = \sum_{K\in\mathcal{T}_h} \sum_{i=1}^4 I_K^i.$$

Let $(n_{1,1}, n_{2,1})$ (resp., $(n_{1,3}, n_{2,3})$) be the unit normal vector to e_1 (resp., e_3); we have

$$I_K^1 + I_K^3 = \int_{e_1} (u_1 - \mathcal{J}_{e_1} u_1) \times (v - \mathcal{J}_{e_1} v) n_{1,1} ds + \int_{e_3} (u_1 - \mathcal{J}_{e_3} u_1) \times (v - \mathcal{J}_{e_3} v) n_{1,3} ds$$

$$= -\frac{|e_1|}{2} \int_{-1}^1 (\widehat{u}_1(-1,\eta) - \widehat{\mathcal{J}_{e_1} u_1}) \times (\hat{v}(-1,\eta) - \widehat{\mathcal{J}_{e_1} v}) d\eta$$

$$+ \frac{|e_3|}{2} \int_{-1}^1 (\widehat{u}_1(1,\eta) - \widehat{\mathcal{J}_{e_3} u_1}) \times (\hat{v}(1,\eta) - \widehat{\mathcal{J}_{e_3} v}) d\eta.$$

Because $|\hat{e}_1| = |\hat{e}_3| = 2$,

$$\widehat{\mathcal{J}_{e_1} u_1} = \frac{1}{|\hat{e_1}|} \int_{-1}^1 \widehat{u}_1(-1,t) dt, \quad \widehat{\mathcal{J}_{e_3} u_1} = \frac{1}{|\hat{e_3}|} \int_{-1}^1 \widehat{u}_1(1,t) dt.$$

Note that

$$\hat{v}(-1,\eta) - \mathcal{J}_{\hat{e}_1} \hat{v} = \hat{v}(1,\eta) - \mathcal{J}_{\hat{e}_3} \hat{v}$$

and $\frac{\partial}{\partial\eta}\hat{v}(\xi,\eta) = c_0 + c_1\eta$ with c_0 and c_1 constant; then we derive

$$I_K^1 + I_K^3$$

$$= \frac{|e_1|}{8} \int_{-1}^1 \left\{ \int_{-1}^1 \int_t^\eta \int_{-1}^1 \frac{\partial^2}{\partial s \partial z} \widehat{u}_1(s,z) ds dz dt \right\} \times \left\{ \int_{-1}^1 \int_t^\eta \frac{\partial}{\partial z} \hat{v}(1,z) dz dt \right\} d\eta$$

$$\le Ch |\widehat{u}_1|_{H^2(\hat{K})} |\hat{v}|_{H^1(\hat{K})} \le Ch^2 |u_1|_{H^2(K)} |v|_{H^1(K)}.$$

Similarly,

$$I_K^2 + I_K^4 \le Ch^2 |u_2|_{H^2(K)} |v|_{H^1(K)}.$$

The summation over elements then completes the proof. ∎

THEOREM 5.2. *Let $(\sigma, u) \in H(\mathrm{div}, \Omega, \mathbb{S}) \times L^2(\Omega, \mathbb{R})$ be the unique solution to problem (1.1) and let $(\sigma_h, u_h) \in \Sigma_h^i \times V_h$ be the unique solution to problem (2.2) with the discrete space pairs $\Sigma_h^i \times V_h$, $i = 0, 1$, defined in section 2. Then*

$$\|\sigma - \sigma_h\|_{H(\mathrm{div}_h, \Omega)} + \|u - u_h\|_{L^2(\Omega)} \le Ch(\|u\|_{H^2(\Omega)} + \|\,\mathrm{div}\,\sigma\|_{H^1(\Omega)}). \tag{5.4}$$

Proof. Define

$$Z_f^i := \{\tau \in \Sigma_h^i | (\mathrm{div}_h\,\sigma, v) = (f, v) \quad \forall v \in V_h\}, \qquad i = 0, 1. \tag{5.5}$$

For any $\tau \in Z_f^i$, we have $\sigma_h - \tau \in Z_h^i$. It follows from the K-ellipticity (see Theorem 3.2 and Remark 3.3) and from $(\mathrm{div}_h(\sigma_h - \tau), v_h)_{L^2(\Omega)} = 0$ for any $v_h \in V_h$ that

$$\begin{aligned} C\|\sigma_h - \tau\|^2_{H(\mathrm{div}_h, \Omega)} &\le (A(\sigma_h - \tau), \sigma_h - \tau)_{L^2(\Omega)} \\ &= (A(\sigma_h - \sigma + \sigma - \tau), \sigma_h - \tau)_{L^2(\Omega)} \\ &= (A(\sigma_h - \sigma), \sigma_h - \tau)_{L^2(\Omega)} + (A(\sigma - \tau), \sigma_h - \tau)_{L^2(\Omega)} \\ &= (A(\sigma - \tau), \sigma_h - \tau)_{L^2(\Omega)} - (A\sigma, \sigma_h - \tau)_{L^2(\Omega)} \\ &\quad - (\mathrm{div}_h(\sigma_h - \tau), u)_{L^2(\Omega)} + (\mathrm{div}_h(\sigma_h - \tau), u - P_h u). \end{aligned}$$

Using $\epsilon(u) = A\sigma$ with $\epsilon(u) = 1/2(\nabla u + \nabla u^T)$ for the exact solution $u \in H_0^1(\Omega, \mathbb{R}^2)$ and Hölder's inequality, we obtain

$$\begin{aligned} \|\sigma_h - r\|_{H(\mathrm{div}_h, \Omega)} &\le C \inf_{\tau \in Z_f^i} \|\sigma - \tau\|_{L^2(\Omega)} \\ &\quad + C \sup_{\tau \in \Sigma_h^i} \frac{(\epsilon(u), \tau)_{L^2(\Omega)} + (\mathrm{div}_h\,\tau, u)_{L^2(\Omega)} + (\mathrm{div}_h\,\tau, P_h u - u)_{L^2(\Omega)}}{\|\tau\|_{H(\mathrm{div}_h, \Omega)}}. \end{aligned} \tag{5.6}$$

It follows from Lemma 3.1 (resp., Remark 3.3) that $\Pi_h^i \sigma \in Z_f^i$, which leads to

$$\inf_{\tau \in Z_f^i} \|\sigma - \tau\|_{L^2(\Omega)} \le Ch^m \|\sigma\|_{H^m(\Omega)}, \quad m = 1, 2.$$

Since $(\mathrm{div}_h\,\tau, P_h u - u)_{L^2(\Omega)} = 0$, we now focus on the second term in (5.6), which measures the nonconformity of Σ_h^i. An elementwise integration by parts yields

$$\begin{aligned} (\epsilon(u), \tau)_{L^2(\Omega)} + (\mathrm{div}_h\,\tau, u)_{L^2(\Omega)} &= \sum_{K \in \mathcal{T}_h} \sum_{e \subset \partial K} \int_e \tau n \cdot u ds \\ &= \sum_{K \in \mathcal{T}_h} \left(\sum_{e \in \mathcal{E}_{K,V}} \int_e (\tau_{11} u_1 + \tau_{21} u_2) n_1 ds + \sum_{e \in \mathcal{E}_{K,H}} \int_e (\tau_{12} u_1 + \tau_{22} u_2) n_2 ds \right). \end{aligned} \tag{5.7}$$

It follows from the definition of Σ_h^i that the moments of degree 0 and 1 of τ_{11} are continuous on vertical edges, the moments of degree 0 and 1 of τ_{22} are continuous on horizontal edges, and $\tau_{12}=\tau_{21}\in NR_h$. These with $u\in H_0^1(\Omega,\mathbb{R}^2)$ lead to

$$\begin{aligned}(\epsilon(u),\tau)_{L^2(\Omega)}+(\operatorname{div}_h\tau,u)_{L^2(\Omega)}=&\sum_{K\in\mathcal{T}_h}\sum_{e\in\mathcal{E}_{K,V}}\int_e(\tau_{21}-\Pi_0^e\tau_{21})(u_2-\Pi_0^e u_2)n_1ds\\&+\sum_{K\in\mathcal{T}_h}\sum_{e\in\mathcal{E}_{K,H}}\int_e(\tau_{12}-\Pi_0^e\tau_{12})(u_1-\Pi_0^e u_1)n_2ds\\&+\sum_{K\in\mathcal{T}_h}\sum_{e\in\mathcal{E}_{K,V}}\int_e(\tau_{11}-\Pi_1^e\tau_{11})(u_1-\Pi_1^e u_1)n_1ds\\&+\sum_{K\in\mathcal{T}_h}\sum_{e\in\mathcal{E}_{K,H}}\int_e(\tau_{22}-\Pi_1^e\tau_{22})(u_2-\Pi_1^e u_2)n_2ds.\end{aligned}$$

Here, Π_1^e denotes the projection operator from $L^2(e)$ to $P_1(e)$. It follows from Lemma 5.1 that

$$\begin{aligned}&\sum_{K\in\mathcal{T}_h}\sum_{e\in\mathcal{E}_{K,V}}\int_e(\tau_{21}-\Pi_0^e\tau_{21})(u_2-\Pi_0^e u_2)n_1ds\\&\quad+\sum_{K\in\mathcal{T}_h}\sum_{e\in\mathcal{E}_{K,H}}\int_e(\tau_{12}-\Pi_0^e\tau_{12})(u_1-\Pi_0^e u_1)n_2ds\\&\le Ch^2|\tau_{12}|_{1,h}|u|_{H^2(\Omega)}.\end{aligned}$$

In view of the approximation of Π_1^e and the trace theorem,

$$\begin{aligned}&\sum_{K\in\mathcal{T}_h}\sum_{e\in\mathcal{E}_{K,V}}\int_e(\tau_{11}-\Pi_1^e\tau_{11})(u_1-\Pi_1^e u_1)n_1ds\\&\qquad\le Ch^2|\tau_{11}|_{1,h}|u_1|_{H^2(\Omega)},\\&\sum_{K\in\mathcal{T}_h}\sum_{e\in\mathcal{E}_{K,H}}\int_e(\tau_{22}-\Pi_1^e\tau_{22})(u_2-\Pi_1^e u_2)n_2ds\\&\qquad\le Ch^2|\tau_{22}|_{1,h}|u_2|_{H^2(\Omega)}.\end{aligned}$$

A combination of the aforementioned three inequalities leads to

$$(\epsilon(u),\tau)_{L^2(\Omega)}+(\operatorname{div}_h\tau,u)_{L^2(\Omega)}\le Ch^2|\tau|_{1,h}|u|_{H^2(\Omega)}. \tag{5.8}$$

Taking into account the approximation property of P_h and using the inverse estimate, we get

$$\|\sigma_h-\sigma\|_{H(\operatorname{div}_h,\Omega)}\le Ch(\|u\|_{H^2}+\|\operatorname{div}\sigma\|_{H^1(\Omega)}). \tag{5.9}$$

We turn to the error estimate for u_h. In view of the discrete B-B condition (see Theorem 3.2 and Remark 3.3), we have for any $v \in V_h$ that

$$\begin{aligned} C\|u_h - v\|_{L^2(\Omega)} &\le \sup_{\tau\in\Sigma_h^i} \frac{(\operatorname{div}_h \tau, u_h - v)_{L^2(\Omega)}}{\|\tau\|_{H(\operatorname{div}_h,\Omega)}} \\ &\le \frac{(\operatorname{div}_h \tau, u_h - u)_{L^2(\Omega)}}{\|\tau\|_{H(\operatorname{div}_h,\Omega)}} + \inf_{v\in V_h} \|u - v\|_{L^2(\Omega)} \\ &\le C\big(\|\sigma_h - \sigma\|_{L^2(\Omega)} + \inf_{v\in V_h} \|u - v\|_{L^2(\Omega)}\big) \\ &\quad + \sup_{\tau\in\Sigma_h^i} \frac{(\epsilon(u), \tau)_{L^2(\Omega)} + (\operatorname{div}_h \tau, u)_{L^2(\Omega)}}{\|\tau\|_{H(\operatorname{div},\Omega)}} \\ &\le Ch(\|u\|_{H^2(\Omega)} + \|\operatorname{div}\sigma\|_{H^1(\Omega)}). \end{aligned} \tag{5.10}$$

An application of the triangle inequality with $v = P_h u$ gives

$$\|u - u_h\|_{L^2(\Omega)} \le Ch(\|u\|_{H^2(\Omega)} + \|\operatorname{div}\sigma\|_{H^1(\Omega)}). \qquad \square \tag{5.11}$$

Remark 5.3. With a minor modification of the above proof, one can show that

$$\|\sigma - \sigma_h\|_{L^2(\Omega)} \le Ch\|u\|_{H^2(\Omega)}, \tag{5.12}$$

$$\|\operatorname{div}\sigma - \operatorname{div}_h \sigma_h\|_{L^2(\Omega)} \le Ch\|\operatorname{div}\sigma\|_{H^1(\Omega)}, \tag{5.13}$$

$$\|u - u_h\|_{L^2(\Omega)} \le Ch\|u\|_{H^2(\Omega)}. \tag{5.14}$$

Remark 5.4. Note that for the quadrilateral mesh, the discrete space Σ_h cannot be defined from the space $\Sigma_{\hat{K}}$ on the reference element $\hat{K}$ through the usual Piola transformation. Therefore, it has to be directly defined on the present element $K \in \mathcal{T}_h$. The rectangular methods of this paper cannot be directly extended to the more general quadrilateral mesh. To get the quadrilateral versions of them, we need to enrich the shear stress space to $Q_2(K)\backslash\{x^2y^2\} \oplus \{x^3y - y^3x\}$ from [12]. In the first method, then the degrees of freedom become the moments of degrees 0 and 1 of τn on each edge and the average of τ on each element K. For the rectangular mesh, this results in a variant of the method proposed in [12]. The second method is obtained from the first one by dropping out the bubble functions in the normal stress, which goes to another variant of the method in [12] for a rectangular mesh. The displacement space V_h is the same as that in section 2. In addition, to get the stability, one can follow the idea of [12] to employ the modified energy norm $\|\tau\|_{L^2(\Omega)} + \|P_h \operatorname{div}_h \tau\|_{L^2(\Omega)}$ for $\tau \in \Sigma_h$ instead.

Remark 5.5. After the submission of this paper, we found that the second element (Σ_h^1, V_h) is independently proposed in [13]. The analysis therein is based on the superconvergence of the consistency term for $\tau \in Z_h^1$. However, the regularity $u \in (H^3(\Omega))^2$ is required in the proof of this superconvergence (see the proof of Lemmas 5.1 and 4.3 of [13]). In this paper, we get more practical results with the regularity $u \in (H^2(\Omega) \cap H_0^1(\Omega))^2$ (see Lemma 5.1 and (5.12)). We also stress that the analysis of [13] cannot be generalized to the more general quadrilateral mesh either.

REFERENCES

[1] D. N. Arnold and G. Awanou, *Rectangular mixed finite elements for elasticity*, Math. Models Methods Appl. Sci., 15 (2005), pp. 1417–1429.
[2] D. N. Arnold and R. Winther, *Mixed finite elements for elasticity*, Numer. Math., 42 (2002), pp. 401–419.
[3] D. N. Arnold and R. Winther, *Nonconforming mixed elements for elasticity*, Math. Models Methods Appl. Sci., 13 (2003), pp. 295–307.
[4] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer-Verlag, New York, 1991.
[5] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1978.
[6] J. Hu, *Quadrilateral Locking Free Elements in Elasticity*, Doctorate dissertation, Institute of Computaional Mathematics, Chinese Academy of Sciences, Beijing, 2004 (in Chinese).
[7] J. Hu and Z. C. Shi, *Constrained quadrilateral nonconforming rotated Q_1 element*, J. Comput. Math., 23 (2005), pp. 561–586.
[8] J. Hu and Z. C. Shi, *Analysis of Nonconforming Quadrilateral Rotated Q_1 Element for Reissner-Mindlin Plate*, Preprint 2003-10, Institute of Computational Mathematics, Chinese Academy of Science, 2003.
[9] Q. Lin, L. Tobiska, and A. H. Zhou, *On the superconvergence of nonconforming low order finite elements applied to the Poisson equation*, IMA J. Numer. Anal., 25 (2005), pp. 160–181.
[10] P. B. Ming, *Nonconforming Finite Element vs. Locking Problem*, Doctorate dissertation, Institute of Computational Mathematics, Chinese Academy of Sciences, Beijing, 1999 (in Chinese).
[11] P. A. Raviart and J. M. Thomas, *A mixed finite element method for second order elliptic problems*, in Mathematical Aspects of the Finite Element Method, Lecture Notes in Math. 606, I. Galligani and E. Magenes, eds., Springer-Verlag, New York, 1977, pp. 292–315.
[12] S. Y. Yi, *Nonconforming mixed finite element methods for linear elasticity using rectangular elements in two and three dimensions*, Calcolo, 42 (2005), pp. 115–133.
[13] S. Y. Yi, *A new nonconforming mixed finite element method for linear elasticity*, Math. Models Methods Appl. Sci., 7 (2006), pp. 979–999.

SIAM J. NUMER. ANAL.
Vol. 46, No. 1, pp. 103–123
© 2007 Society for Industrial and Applied Mathematics

STRANG SPLITTING FOR THE TIME-DEPENDENT SCHRÖDINGER EQUATION ON SPARSE GRIDS*

V. GRADINARU†

Abstract. The time-dependent Schrödinger equation is discretized in space by a sparse grid pseudospectral method. The Strang splitting for the resulting evolutionary problem features first or second order convergence in time, depending on the smoothness of the potential and of the initial data. In contrast to the full grid case, where the frequency domain is the working place, the proof of the sufficient conditions for the convergence is done in the space realm.

Key words. trigonometric interpolation, sparse grids, Strang splitting, Schrödinger equation

AMS subject classifications. 65M70, 65T40, 42A16, 81-08

DOI. 10.1137/050629823

1. Introduction.

Problems in high dimensional spaces represent one of the big challenges today. The standard method of solving the Schrödinger equation uses a representation of the Hamiltonian and of the wave-functions in appropriate product bases [8]. It is sobering that the required computational resources grow exponentially with the number of particles. Sparse grids are an attempt to lessen this curse of dimensionality. Sometimes called hyperbolic cross approximation, they gained increasing attention since their introduction not only for the efficient approximations of smooth functions [10, 4, 11, 36, 32, 34] but also for the solution of PDEs [45, 6, 7]. The recent survey paper on sparse grids [7] contains detailed references. The crucial idea of the sparse grid schemes is to drop certain insignificant contributions of the hierarchical representations of functions. Provided that the solutions are smooth, a tremendously reduced number of unknowns yields almost the same asymptotic rates of approximation as for standard grids.

Spectral methods are well suited for solving PDEs to high accuracy on simple domains and for smooth data. The extension of the sparse grid techniques to the spectral methods comes about by means of Fourier polynomials with harmonics from the hyperbolic cross [23]. Based on the Fourier transform on sparse grids [20, 4] many common pseudospectral methods (such as those based on Chebyshev and Legendre polynomials) arise on sparse grids in the same way as Fourier methods [29, 35]. Applied to various elliptic, parabolic, and hyperbolic problems, the sparse grid collocation method evidenced a trade of efficiency for robustness [23]. The order of convergence depends on the smoothness of the solution as well as on the initial and boundary conditions.

Problems from quantum mechanics were early on a target of the sparse grids methods [21]. The combination technique on sparse grids [18, 30, 28] is used for the six-dimensional eigenvalue problem resulting from the Born–Oppenheimer approximation of the helium atom in [14]. Limitations of this approach are the impossibility to compute eigenvalues from the higher end of the spectrum and the lack of adaptivity. Recently, cutting-edge work on the electronic Schrödinger equation showed surprising

*Received by the editors April 22, 2005; accepted for publication (in revised form) May 10, 2007; published electronically December 20, 2007.
http://www.siam.org/journals/sinum/46-1/62982.html
†Mathematisches Institut, Universität Tübingen, Auf der Morgenstelle 10, D-72076 Tübingen, Germany (gradinar@na.uni-tuebingen.de).

smoothness properties of their solutions [41, 42, 43], a fact that can help to break the complexity barriers using antisymmetrized sparse grid functions [17, 44].

The time-dependent setting of the quantum molecular problems is different, since we are faced with an initial value problem. In the Born–Oppenheimer approximation, the original problem is split into an electronic part and a nuclear part. The latter involves only the nuclei evolving in a smooth potential [37, 19]. Here, we tackle the time-dependent Schrödinger equation (TDSE) that appears in quantum molecular dynamics for nuclei [40]:

$$\begin{aligned} i\frac{\partial u}{\partial t} &= -\Delta u + Vu \quad \text{for } \mathbf{x} \in \Omega \text{ and } t > 0, \\ u(\mathbf{x}, 0) &= g(\mathbf{x}) \qquad\quad \text{for } \mathbf{x} \in \Omega, \end{aligned} \tag{1.1}$$

where the smooth potential $V : \mathbb{R}^d \to \mathbb{R}$ is 2π-periodic in each coordinate and the boundary conditions are periodic. In this setting, the pseudospectral method together with a Strang splitting for the time scheme are appropriate for the numerical solution [22]. However, the standard Fourier discretization of the space needs a full grid and hence confronts us with the curse of dimensionality. This limits the classical approach to dimensions $d \leq 3$; see [3]. Theoretically, the trigonometric polynomials on sparse grids permit a space discretization in higher dimensional cases. The context of the collocation method allows us to avoid the difficulties that the sparse grid schemes usually have in the presence of variable coefficients [1]. Let us note that we do not have at our disposal the symmetry conditions imposed by the Pauli principle, as we do not expect the nuclei of a molecule to be identical. Hence we are not able to really break the curse of dimensionality, as done for the electronic Schrödinger equation [17, pp. 41–44] but only to lessen it to some extent. Here, we study the sparse grid spectral collocation in space combined with the Strang splitting in time. First, we interpolate the initial data by Fourier polynomials on the sparse grid of a fixed resolution n. Then, we propagate the wave-function in time by means of a Strang splitting scheme. This is a widely used numerical approach to the linear initial value problem

$$u' = (A + B)u, \quad u(0) = u_0.$$

Also known as the symmetric operator splitting, the time iteration

$$u^{m+1} = e^{\frac{1}{2}\tau B} e^{\tau A} e^{\frac{1}{2}\tau B} u^m$$

determines recursively approximations u^m to $u(m\tau)$. The space discretization is reflected in the actual operators A and B. In [22], the analysis of the Strang splitting in the presence of unbounded operators was performed in an abstract setting. The resulting sufficient conditions are in the form of commutator bounds. The challenge in the sparse grid context is to prove such commutator bounds that hold uniformly with respect to the resolution. This is the central result of the paper. We show global error bounds ensuring order one, or, for smoother data, order two convergence in time. We impose high regularity assumptions for the initial data and the potential, and the simulations show that they are necessary. In the adiabatic case, one can prove that, away from the gap, the potential is analytic [19, 37]. Applications of such smooth potentials may be found in [26, 2] and [31, page 552].

We begin by resuming the trigonometric interpolation on sparse grids in the spirit of [20]. Then, we present several approximation estimates in Korobov spaces. Together with some direct and inverse inequalities, these estimates play a central role

in the convergence proofs. Next, we show that the interpolation error in the corresponding Korobov space transfers to the solution of the space-discretized problem, as is expected in collocation pseudospectral methods [9]. But for the convergence of the time scheme, commutator bounds for the involved unbounded operators are needed [22]. In the full grid case the proof may be done in the frequency domain by arguments relying on the aliasing formula. Lacking the corresponding tool for the Fourier coefficients of the sparse grid interpolator, we show that the commutator lemmas hold in the space domain. Here, the main ingredients are furnished by the decay of the space-residuals as established in the approximation estimates. In the last section, we resume the convergence results and the dependence on the smoothness of the potential and the initial data. We present numerical experiments in small dimensional cases. This makes possible a comparison to the full grid approach, because similar computations are then easy to perform. The focus is on the time convergence rates and on the influence of the smoothness of the initial data and of the potential. Algorithmic and convergence issues in the space discretization together with examples in higher dimensional cases are the matter of a separate paper [15].

2. Trigonometric interpolation on sparse grids. Spectral collocation methods work usually on tensor product grids [38, 13]. This is the prerequisite for a sparse grid approach as well, because it allows the reduction to the one-dimensional case. The mainstay of a sparse grid interpolation is a hierarchical multilevel representation of functions [7, 16]. We follow the approach of Hallatschek [20] for the design of the hierarchical trigonometrical basis.

We begin with the one-dimensional case discretizing the space domain $[-\pi, \pi]$ by

$$G_n := \{x_{nk} := 2\pi\, k2^{-n} \text{ for } k \in G'_n\}\,,$$

where G'_n is the discrete frequency domain

$$G'_n = \{1 - 2^{n-1}, \ldots, 2^{n-1}\}$$

for levels $n \geq 1$. For $n = 0$ we consider $G'_0 = \{0\}$.

The *nodal trigonometric representation* of a smooth function u is

$$\mathcal{I}_n u(x) := u_n(x) := \sum_{k \in G'_n} c_{nk} e^{i\,kx}\,,$$

with the *nodal Fourier coefficients* given by

$$c_{nk} := \frac{1}{2^n} \sum_{x \in G_n} u(x) e^{-i\,kx}\,.$$

For the manipulation of the indices it is useful to define the set

$$I_n := \begin{cases} \{0\} & \text{if } n = 0, \\ \{2^{n-1}, \ldots, 2^n - 1\} & \text{if } n \geq 1 \end{cases}$$

and the function

$$\sigma(q) := \begin{cases} -q/2 & \text{if } q \text{ is even}, \\ (q+1)/2 & \text{if } q \text{ is odd}. \end{cases}$$

It is then easy to see that $\sigma(I_{n+1}) = G'_{n+1} \setminus G'_n$ corresponds to the new nodes when gradually refining the grid from G_n to G_{n+1}. Let us change the order of summation according to σ:

$$u_n(x) = \sum_{q=0}^{2^n-1} c_{n\sigma(q)}\beta_{nq}(x) = \sum_{q=0}^{2^n-1} \tilde{c}_{nq}\beta_{nq}(x)\,,$$

where $\beta_{nq}$is the re-indexed exponential function:

$$\beta_{nq}(x) := \begin{cases} e^{-i\,\frac{q}{2}x} & \text{if } q \text{ is even,} \\ e^{i\,\frac{q+1}{2}x} & \text{if } q \text{ is odd.} \end{cases}$$

Note that the shape of β_{nq} does not depend explicitly on the level n but only on the index q. The meaning of n is to remind us that the range of the index q is allowed to vary: $\{0, \ldots, 2^n - 1\}$. The information gain when increasing the *resolution* n by one may then be written as

$$\begin{aligned} u_{n+1} - u_n = & \sum_{q \in I_{n+1}} \tilde{c}_{n+1\,q}(\beta_{n+1\,q} - \beta_{n+1\,2^{n+1}-q-1}) \\ & + \sum_{q=0}^{2^n-1} (\tilde{c}_{n+1\,2^{n+1}-q-1} + \tilde{c}_{n+1\,q} - \tilde{c}_{nq})\beta_{nq}\,. \end{aligned}$$

It interpolates at the new points $\sigma(I_{n+1})$, whereas it is zero at the old grid points G_n. Simple manipulations show that the coefficients of the second sum are zero:

$$\tilde{c}_{n+1\,2^{n+1}-q-1} + \tilde{c}_{n+1\,q} - \tilde{c}_{nq} = 0.$$

This yields to the definition of the *hierarchical trigonometrical basis* functions

$$\chi_{nq} := \beta_{nq} - \beta_{n\,2^n-q-1} \text{ for } q \in I_n, n \geq 1,$$

which have the property that they vanish at the grid points of level $n-1$:

$$\chi_{nq}(x) = 0 \text{ for all } x \in G_{n-1}.$$

In contrast to β_{nq}, the behavior of the hierarchical basis function is determined by q and n. We arrive at the *hierarchical representation* of the trigonometrical interpolator

$$\mathcal{I}_n u(x) = \tilde{c}_{00}\beta_{00}(x) + \sum_{l=1}^{n}\sum_{q\in I_l} \tilde{c}_{lq}\chi_{lq} = \sum_{l=0}^{n}\sum_{q\in I_l} d_{lq}\chi_{lq}\,,$$

where $d_{lq} := \tilde{c}_{lq} = c_{n\sigma(q)}$ and $\chi_{00} := \beta_{00}$. Recall the significance of c_{lk}: they are the discrete Fourier coefficients when *discretizing at the level* l. The aliasing formula for $c_{l\sigma(q)}$ gives a representation for the *hierarchical Fourier coefficients* d_{lq}:

$$d_{lq} = c_{l\sigma(q)} = \sum_{m\in\mathbb{Z}} \widehat{u}(\sigma(q) + m2^l),$$

where $\widehat{u}$ is the (continuous) Fourier transform of u: $\widehat{u}(k) = \int_{-\pi}^{\pi} e^{-ikx}u(x)dx$. One can switch from the nodal to the hierarchical coefficients and backward by simple recursive algorithms [20].

Now, we move to the higher dimensional case: $\Omega := [-\pi, \pi]^d$, $d \geq 2$. Writing $\mathbf{n} = (n_1, \ldots, n_d) \in \mathbb{N}^d$ for the *level*, and $\mathbf{k} = (k_1, \ldots, k_d) \in \mathbb{Z}^d$ for the index, the *full grid in space domain* is

$$G_{\mathbf{n}} := \{x_{\mathbf{nk}}, \text{ with } x_{n_s k_s} := 2\pi\, k_s 2^{-n_s}, \text{ for } k_s \in G'_{n_s}\}.$$

The *full grid in frequency domain* is the set of index vectors at level $\mathbf{n}$:

$$G'_{\mathbf{n}} = \{\boldsymbol{\omega}_{\mathbf{nk}} := \mathbf{k} \text{ for } k_s \in G'_{n_s}, s = 1, \ldots, d\}.$$

The *nodal trigonometric interpolation* of a function u on the full grid $G_{\mathbf{n}}$ reads

$$\mathcal{I}^{\text{full}}_{\mathbf{n}} u(\mathbf{x}) := u_{\mathbf{n}}(\mathbf{x}) := \sum_{\mathbf{k} \in G'_{\mathbf{n}}} c_{\mathbf{nk}} e^{i\, \mathbf{k}\cdot\mathbf{x}}, \tag{2.1}$$

with the *nodal Fourier coefficients*

$$c_{\mathbf{nk}} := \frac{1}{2^{|\mathbf{n}|_1}} \sum_{\mathbf{x} \in G_{\mathbf{n}}} u(\mathbf{x}) e^{-i\, \mathbf{k}\cdot\mathbf{x}},$$

where $|\mathbf{n}|_1 := n_1 + \cdots + n_d$. The tensor product allows us to perform the transformation to the hierarchical basis in each direction:

$$\mathcal{I}^{\text{full}}_{\mathbf{n}} u(\mathbf{x}) = u_{\mathbf{n}}(\mathbf{x}) = \sum_{|\mathbf{l}|_\infty \leq n} \sum_{\mathbf{q} \in I_{\mathbf{l}}} d_{\mathbf{lq}} \chi_{\mathbf{lq}}(\mathbf{x}), \tag{2.2}$$

where $|\mathbf{l}|_\infty = \max\{l_s, s = 1, \ldots, d\}$, the basis functions are constructed by tensor product $\chi_{\mathbf{lq}}(\mathbf{x}) := \prod_{s=1}^d \chi_{l_s q_s}(x_s)$, and $I_{\mathbf{l}} = I_{l_1} \times \cdots \times I_{l_d}$.

Denote $\mathbf{k} := \sigma(\mathbf{q}) := (\sigma(q_1), \ldots, \sigma(q_d))$. Due to the aliasing formula, we have for the *hierarchical Fourier coefficients*

$$d_{\mathbf{lq}} = c_{\mathbf{l}\,\sigma(\mathbf{q})} = c_{\mathbf{lk}} = \sum_{\mathbf{m} \in \mathbb{Z}^d} \widehat{u}(k_1 + m_1 2^{l_1}, \ldots, k_d + m_d 2^{l_d})\,. \tag{2.3}$$

The classical *sparse grid* in the space domain is

$$S_n := \{\mathbf{x}_{\mathbf{l}\,\sigma(\mathbf{q})}, \text{ where } |\mathbf{l}|_1 \leq n, \text{ and } \mathbf{q} \in I_{\mathbf{l}}\}$$

and in the frequency domain is

$$S'_n := \{\boldsymbol{\omega}_{\mathbf{l}\,\sigma(\mathbf{q})} := \sigma(\mathbf{q}), \text{ where } |\mathbf{l}|_1 \leq n, \text{ and } \mathbf{q} \in I_{\mathbf{l}}\}.$$

We display them in Figure 2.1 at the particular resolution $n = 4$ in two dimensions.

The thought behind these definitions is that the hierarchical coefficients $d_{\mathbf{lq}}$ furnish significant information only if $|\mathbf{l}|_1 \leq n$, as long as u is sufficiently smooth. In this case, it is enough to consider the trigonometric interpolation only on the sparse grid:

$$\mathcal{I}_n u(\mathbf{x}) := \sum_{|\mathbf{l}|_1 \leq n} \sum_{\mathbf{q} \in I_{\mathbf{l}}} d_{\mathbf{lq}} \chi_{\mathbf{lq}}(\mathbf{x})\,. \tag{2.4}$$

From this hierarchical representation one can switch to the nodal one,

$$\mathcal{I}_n u(\mathbf{x}) = u_n(\mathbf{x}) = \sum_{\boldsymbol{\omega} \in S'_n} a_{\boldsymbol{\omega}} e^{i\, \boldsymbol{\omega}\cdot\mathbf{x}}\,, \tag{2.5}$$

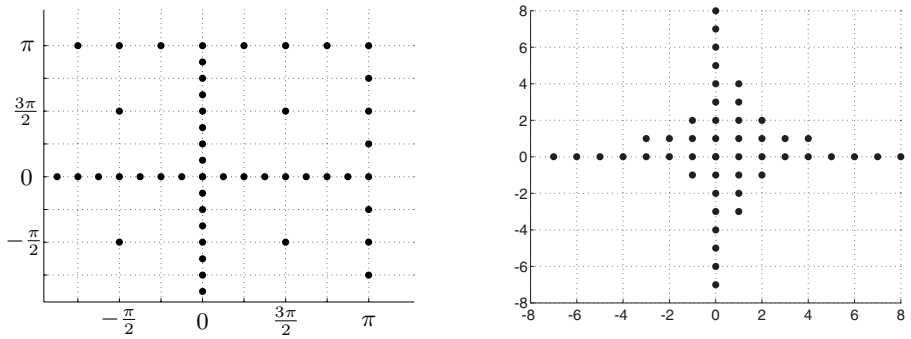


FIG. 2.1. *Sparse grids at resolution $n = 4$: in space (S_4) and in frequency (S_4').*

by recursively carrying out the inverse of the hierarchical transform [20]. Note that the obtained sparse grid Fourier coefficients $a_{\boldsymbol{\omega}}$ are not equal to the corresponding full grid Fourier coefficients $c_{\mathbf{nk}}$. For fixed d and large n, the number of function evaluations is

$$\#S_n = \frac{n^{d-1}2^n}{2^{d-1}(d-1)!} + O(n^{d-2}2^n)\,.$$

Accordingly, Hallatschek [20] showed that the complexity of the direct and reverse computations for fixed d and large n is

$$\frac{Cn^d 2^n}{2^{d-1}(d-1)!} + O(n^{d-1}2^n),$$

with some constant C; this is significantly cheaper than the work needed in the full grid approach, which requires $Cdn2^{nd}$ operations. However, it would be misleading to think that the curse of dimensionality is broken, because we still have an exponential dependence on d. This is typical for the sparse grid approach, where the curse of dimensionality can be circumvented only to some extent by restricting the class of functions under consideration [7].

3. Approximation estimates in Korobov spaces. Let us focus on the required smoothness of u and the corresponding interpolation error. Sobolev spaces of functions defined on a hypercube and having dominant mixed derivatives represent a common place of the sparse grid theory. Endowed with weights moderating the behavior of periodic functions with respect to successive variables, these so-called Korobov spaces make the approximation problem tractable [27, 39]. They reflect the blessing of dimensionality [12, 7]. Here, we stick to the *Korobov space* K^α consisting of complex-valued L^2-functions defined on Ω that are periodic in each variable with period 2π and that have finite $\|\cdot\|_\alpha$-norm:

$$\|u\|_\alpha^2 := \sum_{\mathbf{k}\in\mathbb{Z}^d} |\widetilde{k}_1 \cdot \dots \cdot \widetilde{k}_d|^{2\alpha} |\widehat{u}(\mathbf{k})|^2\,, \tag{3.1}$$

where $\widetilde{k} = 1$ if $k = 0$ and $\widetilde{k} = k$ otherwise. An equivalent norm results if we set $\widetilde{k} = \sqrt{1+k^2}$. This is useful in the proof of some inequalities involving Korobov spaces of different exponent α.

LEMMA 3.1 (direct inequalities). *Provided that u is in $K^{\alpha+2}$ or, respectively, in $K^{\alpha+2\zeta}$ with $\zeta > 0$, it holds that*

$$\|\Delta u\|_\alpha \le d\,\|u\|_{\alpha+2}\,,$$
$$\big\|(I-\Delta)^\zeta u\big\|_\alpha \le (2d)^\zeta\,\|u\|_{\alpha+2\zeta}\,.$$

Proof. In order to familiarize the reader with the Fourier techniques used in the paper, let us prove the direct inequalities explicitly. The involved norms are

$$\|\Delta u\|_\alpha^2 = \sum_{\mathbf{k}\in\mathbb{Z}^d} (\widetilde{k}_1^2+\cdots+\widetilde{k}_d^2)^2 |\widetilde{k}_1\cdot\dots\cdot\widetilde{k}_d|^{2\alpha} |\widehat{u}(\mathbf{k})|^2\,,$$
$$\big\|(I-\Delta)^\zeta u\big\|_\alpha = \sum_{\mathbf{k}\in\mathbb{Z}^d} (1+\widetilde{k}_1^2+\cdots+\widetilde{k}_d^2)^{2\zeta} |\widetilde{k}_1\cdot\dots\cdot\widetilde{k}_d|^{2\alpha} |\widehat{u}(\mathbf{k})|^2\,,$$
$$\|u\|_{\alpha+2\zeta} = \sum_{\mathbf{k}\in\mathbb{Z}^d} (\widetilde{k}_1^2\cdot\dots\cdot\widetilde{k}_d^2)^{2\zeta} |\widetilde{k}_1\cdot\dots\cdot\widetilde{k}_d|^{2\alpha} |\widehat{u}(\mathbf{k})|^2\,,$$

and hence the proof reduces to the trivial inequalities

$$\widetilde{k}_1^2+\cdots+\widetilde{k}_d^2 \le d\widetilde{k}_1^2\cdot\dots\cdot\widetilde{k}_d^2\,,$$
$$1+\widetilde{k}_1^2+\cdots+\widetilde{k}_d^2 \le 2d\widetilde{k}_1^2\cdot\dots\cdot\widetilde{k}_d^2\,. \qquad \square$$

LEMMA 3.2 (inverse inequality). *Take $r > 0$ and $\zeta > dr/2$. If u is sufficiently smooth, then*

$$\|u\|_{\alpha+r} \le C(d,r)\big\|(I-\Delta)^\zeta u\big\|_\alpha\,. \tag{3.2}$$

Proof. Writing the expressions of these norms with $\widetilde{k} = \sqrt{1+k^2}$, we see that we have to find some exponent ζ and a constant C such that $|\widetilde{k}_1\cdot\dots\cdot\widetilde{k}_d|^{2r} \le C(1+k_1^2+\dots k_d^2)^{2\zeta}$. Using the mean inequality

$$\sqrt[d]{\widetilde{k}_1^2\cdot\dots\cdot\widetilde{k}_d^2} \le \frac{\widetilde{k}_1^2+\cdots+\widetilde{k}_d^2}{d},$$

it is easy to see that each exponent $\zeta > dr/2$ gives the result. $\square$

In the following, we just omit $\widetilde{\ }$, the meaning being clear from the context. The scalar product

$$\langle u,v\rangle_\alpha := \sum_{\mathbf{k}\in\mathbb{Z}^d} |k_1\cdot\dots\cdot k_d|^{2\alpha}\widehat{u}(\mathbf{k})\overline{\widehat{v}(\mathbf{k})} \tag{3.3}$$

makes K^α a Hilbert space that is compactly embedded in $L^2(\Omega)$ for $\alpha > 0$. Moreover, for $\alpha > \frac{1}{2}$, it is a reproducing kernel Hilbert space and an algebra [27]:

$$\|uv\|_\alpha \le C(d,\alpha)\,\|u\|_\alpha\|v\|_\alpha\,. \tag{3.4}$$

Using the reasoning of the proof from [27] that K^α is an algebra, one can show that, for appropriately smooth u and v, it holds for $\alpha > 1/2$ that

$$\|\Delta(uv)\|_\alpha \le C(d,\alpha)\ \left(\|u\|_\alpha\|\Delta v\|_\alpha + \|v\|_\alpha\|\Delta u\|_\alpha\right)\,,$$
$$\big\|\Delta^2(uv)\big\|_\alpha \le C(d,\alpha)\ \left(\|u\|_\alpha\big\|\Delta^2 v\big\|_\alpha + \|v\|_\alpha\big\|\Delta^2 u\big\|_\alpha\right)\,,$$
$$\|(I-\Delta)^a(uv)\|_\alpha \le C(d,\alpha)\ \left(\|u\|_\alpha\|(I-\Delta)^a v\|_\alpha + \|v\|_\alpha\|(I-\Delta)^a u\|_\alpha\right)\,.$$

The direct inequalities (Lemma 3.1) give then upper bounds only in terms of different K^α-norms.

LEMMA 3.3. *If u and v are sufficiently smooth, we have*

$$\|\Delta(uv)\|_\alpha \le C(d,\alpha)\ \|u\|_{\alpha+2}\|v\|_{\alpha+2}\,, \tag{3.5}$$

$$\left\|\Delta^2(uv)\right\|_\alpha \le C(d,\alpha)\ \|u\|_{\alpha+4}\|v\|_{\alpha+4}\,, \tag{3.6}$$

$$\left\|(I-\Delta)^\zeta(uv)\right\|_\alpha \le C(d,\alpha)\ \|u\|_{\alpha+2\zeta}\|v\|_{\alpha+2\zeta}\,. \tag{3.7}$$

Note that in the definitions (3.1) and (3.3) the factor 2 in front of α is added for cosmetic reasons that will become clear in what follows. Keep in mind that the case $\alpha = 0$ embodies the L^2-space.

Let the sparse grid interpolation operator (2.4) act on K^α. The representation of the hierarchical Fourier coefficients in d dimensions (2.3) and the Cauchy–Schwarz inequality give for u in K^α

$$|d_{\mathbf{lq}}| \le \|u\|_\alpha \left(\sum_{\mathbf{m}\in\mathbb{Z}^d}\prod_{s=1}^{d}|k_s+m_s2^{l_s}|^{-2\alpha}\right)^{1/2}.$$

We focus on the last sum. Remember that the range of $\mathbf{k}=\sigma(\mathbf{q})$ depends on the level $\mathbf{l}$: for $l_s \ge 2$, we have $2^{l_s-2} \le |k_s| \le 2^{l_s-1}$. Thereby, the one-dimensional sum may be bounded for $\alpha > \frac{1}{2}$ by $(1+\frac{1}{\alpha-\frac{1}{2}})|k_s|^{-2\alpha}$, which is true even for $l_s = 1$ or $l_s = 0$. Hence

$$|d_{\mathbf{lq}}| \le \|u\|_\alpha \left(1+\frac{1}{\alpha-\frac{1}{2}}\right)^{d/2} |k_1\cdot\dots\cdot k_d|^{-\alpha}\,.$$

This brings us to the estimation of the hierarchical Fourier coefficients of a function u belonging to the Korobov space K^α with $\alpha > \frac{1}{2}$:

$$|d_{\mathbf{lq}}| \le \left(1+\frac{1}{\alpha-\frac{1}{2}}\right)^{d/2} 2^{(2d-|\mathbf{l}|_1)\alpha}\|u\|_\alpha.$$

The usual method for the derivation of the interpolation error on sparse grids passes across the hierarchical surpluses

$$\mathcal{H}_\mathbf{l}u(\mathbf{x}) := \sum_{\mathbf{q}\in I_\mathbf{l}} d_{\mathbf{lq}}\chi_{\mathbf{lq}}(\mathbf{x}),$$

because

$$\mathcal{I}_\mathbf{n} = \sum_{|\mathbf{l}|_\infty\le n}\mathcal{H}_\mathbf{l}\,,\quad \mathcal{I}_n = \sum_{|\mathbf{l}|_1\le n}\mathcal{H}_\mathbf{l}\,,\ \text{and}\quad \|\mathcal{I}_n u - u\| \le \sum_{|\mathbf{l}|_1>n}\|\mathcal{H}_\mathbf{l}u\|$$

for suitable norm $\|\cdot\|$. For instance, in the L^2-norm, the hierarchical contributions are estimated by means of the previously obtained bounds for the hierarchical Fourier coefficients

$$\|\mathcal{H}_\mathbf{l}u\|_{L^2(\Omega)} \le \sum_{\mathbf{q}\in I_\mathbf{l}}|d_{\mathbf{lq}}|\cdot\|\chi_{\mathbf{lq}}\|_{L^2(\Omega)} \le \sqrt{2}^d\sum_{\mathbf{q}\in I_\mathbf{l}}|d_{\mathbf{lq}}| \le C'(d,\alpha)\ \cdot 2^{|\mathbf{l}|_1(1-\alpha)}\|u\|_\alpha,$$

because $\|\chi_{\mathbf{lq}}\|_{L^2(\Omega)} = \sqrt{2}$, with the constant $C'(d,\alpha)$ not depending on n but exponentially in d:

$$C'(d,\alpha) = 2^{d\alpha} 2^{d(\alpha-\frac{1}{2})} \left(1 + \frac{1}{\alpha - \frac{1}{2}}\right)^{d/2}.$$

This yields the following bound for the L^2-norm of the interpolation error:

$$\|\mathcal{I}_n u - u\|_{L^2(\Omega)} \leq C(d,\alpha)\ (n+1)^{d-1} 2^{(1-\alpha)n} \|u\|_\alpha,$$

with the constant (with respect to n)

$$C(d,\alpha) = C'(d,\alpha) \cdot S_{\alpha,d},$$

with $S_{\alpha,d} = \sum_{r=1}^{\infty} r^{d-1} 2^{(1-\alpha)r}$. This constant suggests that even for small dimensions d, we need a high resolution n in order to get a reasonable approximation of a high frequency function. Note that $S(\alpha,d) \approx 2^{-\alpha}$ for large α and fixed d, while $S(\alpha,d) \approx e^d$ for fixed α. Hence, if u belongs to the Korobov space K^α with $\alpha > 1$, we achieve the usual convergence for the sparse grid schemes.

The method of finding a bound for the interpolation error may be also applied in the case of the K^β-norm:

$$\|\mathcal{H}_{\mathbf{l}} u\|_\beta \leq \sqrt{2}^d \cdot 2^{(|\mathbf{l}|_1 - d)\beta} \cdot C \cdot 2^{(2d - |\mathbf{l}|_1)\alpha} \|u\|_\alpha,$$

because $\|\chi_{\mathbf{lq}}\|_\beta \leq \sqrt{2} \cdot 2^{(|l|_1 - d)\beta}$. This shows convergence in the K^β-norm; if $\alpha - \beta > 1$,

$$\|\mathcal{I}_n u - u\|_\beta \leq C(d,\alpha,\beta)(n+1)^{d-1} 2^{(1-(\alpha-\beta))n} \|u\|_\alpha,$$

with $C(d,\alpha,\beta) = C'(d,\alpha-\beta) \cdot S(\alpha-\beta,d)$. As the involved constants C are all of the same kind, we often write just C and keep in mind their exponential dependence on d.

PROPOSITION 3.4. *If $u \in K^{\alpha+1+\gamma}$ with $\gamma > 0$, then there is a constant $C = C(d,\alpha,\alpha+1+\gamma)$ such that*

$$\|\mathcal{I}_n u - u\|_\alpha \leq C\ (n+1)^{d-1} 2^{-\gamma n} \|u\|_{\alpha+1+\gamma}.$$

The lack of an aliasing formula in the sparse grid case requires that we impose the condition of more smoothness of the function to interpolate u. The same reason inspires us to make use of the next three propositions regarding the decay of the error $\mathcal{I}_n u - u$ under the action of the operators Δ, Δ^2, and $(I-\Delta)^\zeta$. They are necessary in the derivation of the commutator bounds in section 5.

We begin with the case of Δ. A short calculation shows that $\|\chi''_{lq}\|_\beta \leq \sqrt{2} \cdot 2^{(l-1)(2+\beta)}$. This helps to compute the bound of the hierarchical contributions

$$|d_{\mathbf{lq}}| \cdot \|\Delta \chi_{\mathbf{lq}}\|_\beta \leq C(d,\alpha,\beta)\ 2^{(\beta-\alpha)|\mathbf{l}|_1} \sum_{s=1}^{d} 4^{l_s},$$

which add together giving as previously

$$\begin{aligned}\|\Delta(\mathcal{I}_n u - u)\|_\beta &\leq C(d,\alpha,\beta) \sum_{|\mathbf{l}|_1 > n} 2^{|\mathbf{l}|_1} 2^{(\beta-\alpha)|\mathbf{l}|_1} \sum_{s=1}^{d} 4^{l_s} \\ &\leq C \sum_{m=n+1}^{\infty} 2^{(1-(\alpha-\beta))m} \sum_{|\mathbf{l}|_1 = m} 4^{l_1}.\end{aligned}$$

The last sum is a particular case of the expression

$$S_m(r,d) = \sum_{|\mathbf{l}|_1=m} 4^{l_1+\cdots+l_r}.$$

By backward induction on the rank r with $1 \le r \le d$, it can be shown as in [16] that

$$S_m(r,d) \le \left(\frac{4}{3}\right)^{d-r} 4^m \binom{m+r-1}{r-1},$$

which finally yields for $r = 1$

$$\|\Delta(\mathcal{I}_n u - u)\|_\beta \le C(d,\alpha,\beta)\ 2^{(3-(\alpha-\beta))n}\|u\|_\alpha$$

if $\alpha - \beta > 3$. This results in the following proposition.

PROPOSITION 3.5. *If $u \in K^{\alpha+3+\gamma}$ with a $\gamma > 0$, then there is a constant $C = C(d,\alpha,\gamma)$ such that*

$$\|\Delta(\mathcal{I}_n u - u)\|_\alpha \le C\ 2^{-\gamma n}\|u\|_{\alpha+3+\gamma}.$$

Similar but long manipulations show that

$$\begin{aligned}\left\|\Delta^2(\mathcal{I}_n u - u)\right\|_\beta &\le C \sum_{m=n+1}^{\infty} 2^{(1-(\alpha-\beta))m} S_m(2,d)\\ &\le C\ (n+2)2^{(3-(\alpha-\beta))(n+1)}\|u\|_\alpha\end{aligned}$$

if $\alpha - \beta > 3$. This results in the following proposition.

PROPOSITION 3.6. *If $u \in K^{\alpha+3+\gamma}$ with a $\gamma > 0$, then there is a constant $C = C(d,\alpha,\gamma)$ such that*

$$\left\|\Delta^2(\mathcal{I}_n u - u)\right\|_\alpha \le C\ (n+2)2^{-\gamma n}\|u\|_{\alpha+3+\gamma}.$$

More care is needed in the treatment of $(I-\Delta)^\zeta$. As we cannot explicitly compute the local contributions $\left\|(I-\Delta)^\zeta \chi_{\mathbf{lq}}\right\|_\beta$, we have to use an upper bound

$$\left\|(I-\Delta)^\zeta \chi_{\mathbf{lq}}\right\|_\beta \le C\ 2^{\beta|\mathbf{l}|_1}(1 + 4^{l_1} + \ldots 4^{l_d})^\zeta$$

that furnishes

$$\left\|(I-\Delta)^\zeta(\mathcal{I}_n u - u)\right\|_\beta \le C\ (n+1)^{d-1}2^{(1+2\zeta-(\alpha-\beta))n}\|u\|_\alpha.$$

PROPOSITION 3.7. *For $\zeta \ge 0$ and $\gamma > 0$ given, there is a constant $C = C(d,\alpha,\gamma,\zeta)$ such that for all $u \in K^{\alpha+1+2\zeta+\gamma}$ it holds that*

$$\left\|(I-\Delta)^\zeta(\mathcal{I}_n u - u)\right\|_\alpha \le C\ (n+1)^{d-1}2^{-\gamma n}\|u\|_{\alpha+1+2\zeta+\gamma}.$$

Let us remark that the Korobov spaces and usual Sobolev spaces are connected by the trivial embeddings

$$H^s \subset K^s \subset H^{s/d} \text{ for } s \ge 0,$$

and the operator $(I-\Delta)^\zeta$ is a lift operator (bijection) between H^α and $H^{\alpha+\zeta}$ (see [33]). Hence, the approximation estimates of this section are not surprising. Putting them this way, the upper estimates for Δ^2 reduce to upper estimates for $(I-\Delta)$. Moreover, let us note that the error estimates for the interpolation operator $\mathcal{I}_n$ are closely related to the d-blended Boolean approximation techniques [10, 4, 11, 32, 34].

4. Discretization of the TDSE. We put to use the previously described interpolation operator for the discretization of the TDSE (1.1). Following the reasoning of [22], we first perform the space discretization by the pseudospectral method. Here, we look for a trigonometric polynomial on the sparse grid,

$$u_n(\mathbf{x}) = \sum_{|\mathbf{l}|_1 \leq n} \sum_{\mathbf{q} \in I_\mathbf{l}} d_{\mathbf{lq}}(t) \chi_{\mathbf{lq}}(\mathbf{x}) = \sum_{\boldsymbol{\omega} \in S'_n} a_{\boldsymbol{\omega}}(t) e^{i\,\boldsymbol{\omega}\cdot\mathbf{x}},$$

such that (1.1) is satisfied at the sparse grid points $\mathbf{x} \in S_n$:

$$\begin{aligned} i\frac{\partial u}{\partial t}(\mathbf{x},t) &= -\Delta u(\mathbf{x},t) + V(\mathbf{x})u(\mathbf{x},t) \quad && \text{for } \mathbf{x} \in S_n \text{ and } t > 0, \\ u(\mathbf{x},0) &= g(\mathbf{x}) && \text{for } \mathbf{x} \in S_n. \end{aligned}$$

Let $\mathbb{V}_n = \operatorname{span}\{e^{i\,\boldsymbol{\omega}\cdot\mathbf{x}}, \boldsymbol{\omega} \in S'_n\}$. We regard the operator $\mathcal{I}_n$ as the composition of the restriction $\mathcal{R}_n$ of the functions in K^α to the sparse grid points and the operator $\mathcal{Q}_n$ that constructs the sparse grid interpolation $u_n \in \mathbb{V}_n$:

$$K^\alpha \xrightarrow{\mathcal{R}_n} \mathbb{C}^{\#S_n} \xrightarrow{\mathcal{Q}_n} \mathbb{V}_n \text{ and } \mathcal{I}_n = \mathcal{Q}_n\mathcal{R}_n.$$

The space-discretized problem is then as follows: find $u_n \in \mathbb{V}_n$ such that

$$\begin{aligned} i\frac{\partial u_n}{\partial t} &= -\Delta u_n + \mathcal{Q}_n \operatorname{diag}(\mathcal{R}_n V)\mathcal{R}_n u_n \quad && \text{for } \mathbf{x} \in \Omega \text{ and } t > 0, \\ u(\mathbf{x},0) &= \mathcal{I}_n g(\mathbf{x}) && \text{for } \mathbf{x} \in \Omega\,. \end{aligned} \tag{4.1}$$

The potential V acts here via the operator B_n:

$$B_n u_n = \mathcal{Q}_n \operatorname{diag}(\mathcal{R}_n V)\mathcal{R}_n u_n = \mathcal{Q}_n\mathcal{R}_n(V u_n) = \mathcal{I}_n(V u_n)\,. \tag{4.2}$$

By the Fourier transform and the variation of constants formula, smooth initial condition g and potential V induce a smooth solution u. The restrictions that we actually set on V result from the sufficient conditions for the convergence of the time scheme below and will be revealed in the next section.

Denote $\mathcal{V}_n := \operatorname{diag}[(V(\mathbf{x}))_{\mathbf{x}\in S_n}]$ and $\mathcal{F}_n$ the discrete Fourier transform on the sparse grid S_n. Setting $\widehat{u}_n := (a_{\boldsymbol{\omega}})_{\boldsymbol{\omega}\in S'_n}$ as the vector of Fourier coefficients, (4.1) is equivalent to

$$\begin{aligned} i\widehat{u}'_n &= \mathcal{D}_n^2\widehat{u}_n + \mathcal{W}_n\widehat{u}_n \quad && \text{for } t > 0, \\ \widehat{u}_n(0) &= \widehat{g}, \end{aligned} \tag{4.3}$$

where $\mathcal{D}_n = \operatorname{diag}[(|\boldsymbol{\omega}|_2^2)_{\boldsymbol{\omega}\in S'_n}]$, with the l^2-norm $|\mathbf{k}|_2^2 = k_1^2 + \cdots + k_d^2$, the operator $\mathcal{W}_n = \mathcal{F}_n\mathcal{V}_n\mathcal{F}_n^{-1}$, and $\widehat{g} = \mathcal{F}_n g$.

We show now that the interpolation error for $u \in K^\alpha$ transfers to the error of the space-discretized problem, as is usual in collocation pseudospectral methods [9]. Denoting the error $e_n = u_n - \mathcal{I}_n u$ and the defect $d_n = \Delta(\mathcal{I}_n u) - \mathcal{I}_n(\Delta u)$, we get the equation in space domain,

$$i\frac{\partial e_n}{\partial t} = -\Delta e_n + \mathcal{I}_n(V e_n) + d_n \quad \text{for } \Omega \text{ and } t > 0,$$

and in frequency domain,

$$i\frac{d\widehat{e}_n}{dt} = \mathcal{D}_n^2\widehat{e}_n + \mathcal{W}_n\widehat{e}_n + \widehat{d}_n \quad \text{for } t > 0.$$

We focus on $\|e_n\|_\alpha$, the L^2-convergence arising for $\alpha = 0$. The counterpart in the frequency domain of the K^α-norm is the discrete norm $\|\widehat{u}\|_\alpha = \|u\|_\alpha$, arising from the scalar product $\langle \widehat{u}, \widehat{v}\rangle_\alpha = \langle u, v\rangle_\alpha$. In contrast to the full grid case, $\mathcal{F}_n$ is not orthogonal, so $\mathcal{W}_n$ is not symmetric, such that the standard arguments used in the full grid or in the Galerkin approach are not applicable. We have

$$\|\widehat{e}_n(t)\|_\alpha \frac{d}{dt}\|\widehat{e}_n(t)\|_\alpha \le \|\widehat{e}_n(t)\|_\alpha \cdot \left\|\mathcal{W}_n\widehat{e}_n + \widehat{d}_n(t)\right\|_\alpha .$$

A work-around involves the condition number of the Fourier transform on the sparse grid $c = \|\mathcal{F}_n\|_\alpha \left\|\mathcal{F}_n^{-1}\right\|_\alpha$, the upper bound B of the potential, and the Gronwall lemma:

$$\|e_n(t)\|_\alpha = \|\widehat{e}_n(t)\|_\alpha \le e^{Bct}\left(\|\widehat{e}_n(0)\|_\alpha + \int_0^t \left(1 + e^{-Bcs}\right)\|d_n(s)\|_\alpha ds\right).$$

The initial error for $g \in K^\alpha$ is given by Proposition 3.4:

$$\|\widehat{e}_n(0)\|_\alpha = \|g - \mathcal{I}_n g\|_\alpha \le C(d,\alpha)\ (n+1)^{d-1}2^{-\gamma n}\|g\|_{\alpha+1+\gamma}.$$

Concerning the defect, Proposition 3.5 intervenes as

$$\begin{aligned}\|d_n(t)\|_\alpha &\le \|\Delta(\mathcal{I}_n u(t) - u(t))\|_\alpha + \|\mathcal{I}_n(\Delta u(t)) - \Delta u(t)\|_\alpha \\ &\le C\ 2^{-\gamma n}\|u(t)\|_{\alpha+3+\gamma} + C\ (n+1)^{d-1}2^{-\gamma n}\|\Delta u(t)\|_{\alpha+1+\gamma}\end{aligned}$$

if u lies in $K^{\alpha+3+\gamma}$. We obtain K^α-convergence of u_n to u of order $O((n+1)^{d-1}2^{-\gamma n})$.

PROPOSITION 4.1. *Let the potential V and the initial condition g be so smooth that the solution u of the continuous Schrödinger equation* (1.1) *is in $K^{\alpha+3+\gamma}$. Then the solution u_n of the space-discretized problem* (4.1) *converges to u :*

$$\begin{aligned}\|u_n(t) - u(t)\|_\alpha \le e^{Bct}C\ (n+1)^{d-1}2^{-\gamma n}&\left(\|g\|_{\alpha+1+\gamma} + \int_0^t \left(1 + e^{-Bcs}\right)\|u(s)\|_{\alpha+2+\gamma}ds\right) \\ &+ C\ 2^{-\gamma n}\int_0^t \left(1 + e^{-Bcs}\right)\|u(s)\|_{\alpha+3+\gamma}ds\,.\end{aligned}$$

Note that the pessimistic factor e^{Bct} is not reflected in our numerical simulations. For the time discretization of the obtained system we carry out the Strang splitting. Over a time step τ, the differential system is solved approximately by computing recursively $\widehat{u}_n^m = (a_{\boldsymbol{\omega}})_{\boldsymbol{\omega}\in S_n'}^m$ via

$$\widehat{u}_n^{m+1} = e^{-\frac{i}{2}\tau\mathcal{W}_n}e^{-i\tau\mathcal{D}_n^2}e^{-\frac{i}{2}\tau\mathcal{W}_n}\widehat{u}_n^m\,,$$

where the action of $e^{-\frac{i}{2}\tau\mathcal{W}_n} = \mathcal{F}_n\,\mathrm{diag}[(e^{-\frac{i}{2}\tau\mathcal{V}(\mathbf{x})})_{\mathbf{x}\in S_n}]\mathcal{F}_n^{-1}$ is inexpensive to compute. This leads to the algorithm for the solution u_n^m that is discrete both in time and space. Note that, in contrast to the full grid case, $\mathcal{F}_n$ is not unitary; hence we do not expect the conservation of the L^2-norm. However, we have seen that the pseudospectral method on sparse grids furnishes a good approximation; hence the computed L^2-norm is nearly 1 and renorming is harmless.

For the convergence analysis of the time scheme, we need commutator lemmas [22]. While complicated, one can carry out the proof in the full grid case using the last formulation on the frequency domain (4.3). This method fails in the sparse grid case, since we do not have an aliasing formula for the Fourier coefficients $a_{\boldsymbol{\omega}}$. Moreover,

the hierarchical trigonometric basis functions $\chi_{\mathbf{lq}}$ are not orthogonal, and hence, the representation (2.3) of the hierarchical Fourier coefficients is not useful here. We are therefore constrained to study the convergence of the Strang splitting in the space domain:

$$u_n^{m+1} = e^{-\frac{i}{2}\tau B_n} e^{i\tau\Delta} e^{-\frac{i}{2}\tau B_n} u_n \,. \tag{4.4}$$

This is the subject of the next sections.

5. Commutator bounds. The local error of the Strang splitting is analyzed in an abstract setting in [22], where sufficient conditions in the presence of unbounded operators are derived, together with an application to the one-dimensional Schrödinger equation. Written in our context of the sparse grid discretization (4.1), the main ingredient for proving the first order convergence of the time scheme is the following commutator lemma. Remember from the equality (4.2) the action of the discrete potential $B_n u_n = \mathcal{Q}_n \operatorname{diag}(\mathcal{R}_n V)\mathcal{R}_n u_n = \mathcal{I}_n(V u_n)$. Denote $A = I - \Delta$ and the commutator $[A, B_n] = AB_n - B_n A = [-\Delta, B_n]$. Fix $\alpha \geq 0$ and $\gamma > 0$.

LEMMA 5.1. *For a smooth potential $V \in K^{\alpha+3+\gamma}$, the commutator bound*

$$\|[A, B_n]u_n\|_\alpha \leq C(d,\alpha) \left\|A^\zeta u_n\right\|_\alpha$$

holds with the constants $C(d,\alpha)$ and $\zeta > d(1 + \frac{1}{2} + \frac{\gamma}{2})$ independent of the level n and $u_n \in K^{\alpha+3+\gamma}$.

Proof. Obviously

$$[A, B_n]u_n = -\Delta\mathcal{I}_n(V u_n) + \mathcal{I}_n(V\Delta u_n)\,.$$

Denoting $v := V u_n \in K^\alpha$ and $w := V\Delta u_n \in K^\alpha$, the focus moves on $\|\Delta\mathcal{I}_n v\|_\alpha$ and $\|\mathcal{I}_n w\|_\alpha$. The bound of the interpolation error from Proposition 3.4 gives for $\gamma > 0$

$$\|\mathcal{I}_n w\|_\alpha \leq \|w\|_\alpha + C\,(n+1)^{d-1} 2^{-\gamma n} \|w\|_{\alpha+1+\gamma}.$$

For a sufficiently large $n \geq n_0(d,\alpha,\gamma)$ we obtain by the algebra property of the Korobov space (3.4)

$$\|\mathcal{I}_n w\|_\alpha \leq C\,\|V\|_\alpha \cdot \|\Delta u_n\|_\alpha + C\,\|V\|_{\alpha+1+\gamma} \cdot \|\Delta u_n\|_{\alpha+1+\gamma}\,.$$

Analogously, using Proposition 3.5 and inequality (3.5), we get

$$\|\Delta\mathcal{I}_n v\|_\alpha \leq C\,\|V\|_{\alpha+2} \cdot \|u_n\|_{\alpha+2} + C\,\|V\|_{\alpha+3+\gamma} \cdot \|u_n\|_{\alpha+3+\gamma}.$$

The first part of Lemma 3.1 and the monotony in α of the Korobov norm yield

$$\|[A, B_n]u_n\|_\alpha \leq C\,\|u_n\|_{\alpha+3+\gamma}.$$

This is the moment for the inverse inequality (3.2) to come into play, providing an exponent $\zeta > d(1 + \frac{1}{2} + \frac{\gamma}{2})$. ∎

Let us move now to the question of second order convergence of the time scheme. We split the corresponding sufficient condition into the following two lemmas.

LEMMA 5.2. *For a smooth potential $V \in K^{\alpha+5+\gamma}$, the commutator bound*

$$\|[A, [A, B_n]]u_n\|_\alpha \leq C(d,\alpha) \left\|A^\xi u_n\right\|_\alpha$$

holds with the constants $C(d,\alpha)$ and $\xi > d(2+\frac{1}{2}+\frac{\gamma}{2})$ independent of the level n and $u_n \in K^{\alpha+5+\gamma}$.

Proof. The double commutator is

$$[A,[A,B_n]]u_n = \Delta^2 B_n u_n - 2\Delta B_n \Delta u_n + B_n \Delta^2 u_n,$$

and therefore, by (4.2), the right-hand side is bounded by

$$\|[A,[A,B_n]]u_n\|_\alpha \le \left\|\Delta^2 \mathcal{I}_n(Vu_n)\right\|_\alpha + 2\|\Delta \mathcal{I}_n(V\Delta u_n)\|_\alpha + \left\|\mathcal{I}_n(V\cdot\Delta^2 u_n)\right\|_\alpha .$$

As in the proof of the previous lemma for $\mathcal{I}_n w$, we have

$$\left\|\mathcal{I}_n(V\cdot\Delta^2 u_n)\right\|_\alpha \le C\,\|V\|_\alpha \left\|\Delta^2 u_n\right\|_\alpha + C\,\|V\|_{\alpha+1+\gamma}\left\|\Delta^2 u_n\right\|_{\alpha+1+\gamma},$$

and hence, by the first part of Lemma 3.1,

$$\left\|\mathcal{I}_n(V\cdot\Delta^2 u_n)\right\|_\alpha \le C\,\|u_n\|_{\alpha+5+\gamma}.$$

Proposition 3.5 and inequality (3.5), together with the first part of Lemma 3.1, give an upper bound for the middle term

$$2\|\Delta \mathcal{I}_n(V\Delta u_n)\|_\alpha \le C\,\|V\|_{\alpha+2}\|u_n\|_{\alpha+4} + C\,\|V\|_{\alpha+3+\gamma}\|u_n\|_{\alpha+5+\gamma}.$$

Concerning the first term, we need Proposition 3.6, inequality (3.6), and the first part of Lemma 3.1 in order to obtain

$$\left\|\Delta^2 \mathcal{I}_n(Vu_n)\right\|_\alpha \le C\,\|V\|_{\alpha+4}\|u_n\|_{\alpha+4} + C\,\|V\|_{\alpha+3+\gamma}\|u_n\|_{\alpha+3+\gamma}.$$

All together, an upper bound for the left-hand side is given by $\|u_n\|_{\alpha+5+\gamma}$. As in the previous proof, the inverse inequality (3.2) shows that ξ should be chosen greater than $d(2+\frac{1}{2}+\frac{\gamma}{2})$. Note that we employed only $K^{\alpha+4+\gamma}$ smoothness of V, but more regularity of the potential (and the initial data) is required in order to have a solution $u_n \in K^{\alpha+5+\gamma}$. □

LEMMA 5.3. *Given $\zeta \ge 0$ and the potential $V \in K^{\alpha+1+2\zeta+\gamma}$, there is an exponent $\xi \ge \zeta$ such that the commutator bound*

$$\left\|[(-A)^\zeta, B_n]u_n\right\|_\alpha \le C\,\left\|(-A)^\xi u_n\right\|_\alpha$$

holds with the constant C independent of the level n and $u_n \in K^{\alpha+1+2\zeta+\gamma}$.

Proof. With $v = Vu_n$ and $w = VA^\zeta u_n$, we have

$$\left\|[(-A)^\zeta, B_n]u_n\right\|_\alpha \le \left\|A^\zeta \mathcal{I}_n v\right\|_\alpha + \|\mathcal{I}_n w\|_\alpha .$$

The convergence result from Proposition 3.4 for w produces a majoration of the last term

$$\|\mathcal{I}_n w\|_\alpha \le C\,\left(\|V\|_\alpha \left\|A^\zeta u_n\right\|_\alpha + \|V\|_{\alpha+1+\gamma}\left\|A^\zeta u_n\right\|_{\alpha+1+\gamma}\right);$$

that is, by the second part of Lemma 3.1,

$$\|\mathcal{I}_n w\|_\alpha \le C\,\|u\|_{\alpha+2\zeta+1+\gamma}.$$

We employ Proposition 3.7 and inequality (3.7) for the first term:

$$\left\|A^\zeta \mathcal{I}_n v\right\|_\alpha \le C\,\left(\|V\|_{\alpha+2\zeta}\|u_n\|_{\alpha+2\zeta} + \|V\|_{\alpha+1+2\zeta+\gamma}\|u_n\|_{\alpha+1+2\zeta+\gamma}\right).$$

Consequently, an upper bound of the left-hand side $\left\|[(-A)^\zeta, B_n]u_n\right\|_\alpha$ is $\|u_n\|_\beta$, with $\beta = \alpha+1+2\zeta+\gamma$. The inverse inequality (3.2) leads then to the choice $\xi \ge d\zeta + d(\frac{1}{2}+\frac{\gamma}{2})$. □

6. Convergence results. The previous lemmas gather into the convergence result, which is ensured by Theorem 2.1 and the immediately following remark from [22]. Let us quote it in full. Let A be the generator of a strongly continuous semigroup e^{tA} on a Banach space X and B a bounded linear operator on X.

THEOREM 6.1.

(a) *Under the condition*

$$\|[A,B]v\| \le c_1\|(-A)^\alpha v\| \text{ for all } v$$

with $\alpha > 0$, *we have*

$$\left\|e^{\frac{1}{2}\tau B}e^{\tau A}e^{\frac{1}{2}\tau B}v - e^{\tau(A+B)}v\right\| \le C_1\tau^2\|(-A)^\alpha v\| \text{ for all } v$$

with C_1 *depending only on* c_1 *and* $\|B\|$.

(b) *If we additionally have that*

$$\|[A,[A,B]]v\| \le c_2\left\|(-A)^\beta v\right\| \text{ for all } v,$$
$$\|[(-A)^\alpha,B]v\| \le c_1\left\|(-A)^\beta v\right\| \text{ for all } v$$

with $\beta \ge \alpha$, *then*

$$\left\|e^{\frac{1}{2}\tau B}e^{\tau A}e^{\frac{1}{2}\tau B}v - e^{\tau(A+B)}v\right\| \le C_2\tau^3\|(-A)^\alpha v\| \text{ for all } v$$

with C_2 *depending only on* c_1, c_2, *and* $\|B\|$.

The local bounds ensured by the previous lemmas yield the following global error bounds for the Strang splitting.

THEOREM 6.2. *Under the conditions of Lemma* 5.1 *and, respectively, of Lemmas* 5.2 *and* 5.3, *the following global error bounds of orders* 1 *and* 2 *hold for the Strang splitting* (4.4) *at time* $t = m\tau$:

$$\|u_n^m - u_n(t)\|_\alpha \le \tau C_1 t \max_{0\le s\le t}\left\|A^\zeta u_n(s)\right\|_\alpha,$$
$$\|u_n^m - u_n(t)\|_\alpha \le \tau^2 C_1 t \max_{0\le s\le t}\left\|A^\xi u_n(s)\right\|_\alpha.$$

It remains only to find a majoration of $\left\|A^\zeta u_n(t)\right\|_\alpha$ in terms of the initial data g and the potential V. In the continuous case, the formula of variation of constants together with the Gronwall lemma and inequality (3.7) yield

$$\left\|A^\zeta u(t)\right\|_\alpha \le \sqrt{2}\left\|A^\zeta g\right\|_\alpha e^{Ct^2\|V\|_\alpha^2} \tag{6.1}$$

and

$$\|u(t)\|_\alpha \le \sqrt{2}\|g\|_\alpha e^{Ct^2\|V\|_\alpha^2}. \tag{6.2}$$

The triangle inequality

$$\left\|A^\zeta u_n(t)\right\|_\alpha \le \left\|A^\zeta u_n(t) - A^\zeta u(t)\right\|_\alpha + \left\|A^\zeta u(t)\right\|_\alpha \tag{6.3}$$

reduces the question to the evaluation of the error $e_n = u_n - u$ in the norm induced by the operator A^ζ on K^α:

$$\|e_n\|_{\zeta,\alpha} := \left\|A^\zeta e_n\right\|_\alpha.$$

Analogously to the method in the proof of Proposition 4.1, we have

$$\|e_n(t)\|_{\zeta,\alpha} \leq e^{Bct}\left(\|e_n(0)\|_{\zeta,\alpha} + \int_0^t \left(1+e^{-Bc}\right)\|d_n(s)\|_{\zeta,\alpha} ds\right).$$

The initial error is provided by Proposition 3.7:

$$\|e_n(0)\|_{\zeta,\alpha} = \left\|A^{\zeta}(\mathcal{I}_n g - g)\right\|_{\alpha} \leq C\,(n+1)^{d-1}2^{-\gamma n}\|g\|_{\alpha+1+2\zeta+\gamma}.$$

We endeavor now to bound the defect d_n:

$$\begin{aligned}\|d_n(s)\|_{\zeta,\alpha} &= \left\|A^{\zeta}(\Delta\mathcal{I}_n u - \mathcal{I}_n\Delta u)\right\|_{\alpha}\\ &\leq \|\Delta(\mathcal{I}_n u - u)\|_{\zeta,\alpha} + \|\mathcal{I}_n(\Delta u) - \Delta u\|_{\zeta,\alpha}.\end{aligned}$$

The last term is known from Proposition 3.7:

$$\|\mathcal{I}_n(\Delta u) - \Delta u\|_{\zeta,\alpha} \leq C\,(n+1)^{d-1}2^{-\gamma n}\|\Delta u\|_{\alpha+1+2\zeta+\gamma}.$$

Remember that $A = I - \Delta$, and so the triangle inequality gives for the first term

$$\left\|A^{\zeta}(\Delta(\mathcal{I}_n u - u))\right\|_{\alpha} \leq \left\|A^{\zeta+1}(\mathcal{I}_n u - u)\right\|_{\alpha} + \left\|A^{\zeta}(\mathcal{I}_n u - u)\right\|_{\alpha},$$

and again Proposition 3.7 yields

$$\|\Delta(\mathcal{I}_n u - u)\|_{\zeta,\alpha} \leq C\,(n+1)^{d-1}2^{-\gamma n}\|u\|_{\alpha+3+2\zeta+\gamma}.$$

We obtain the upper bound for the defect

$$\|d_n(s)\|_{\zeta,\alpha} \leq C\,(n+1)^{d-1}2^{-\gamma n}\|u\|_{\alpha+3+2\zeta+\gamma},$$

with another constant C. Thereupon, the error is bounded by

$$\|e_n(t)\|_{\zeta,\alpha} \leq e^{Bct}C\,(n+1)^{d-1}2^{-\gamma n}\left(\|g\|_{\alpha+1+2\zeta+\gamma} + \int_0^t (1+e^{-Bcs})\|u(s)\|_{\alpha+3+2\zeta+\gamma} ds\right).$$

The last inequality, together with (6.3), (6.2) for $K^{\alpha+3+2\zeta+\gamma}$, and (6.1) produces a bound for $\left\|A^{\zeta}u_n(t)\right\|_{\alpha}$ only in terms of the initial data g and the potential V.

Note that the necessary smoothness increases with the dimension d because of the conditions set on ζ and ξ in the commutator bounds lemmas.

We now turn our attention to simulation results. The concrete implementation issues and the approximation properties are investigated numerically in [15]. Consider now the Schrödinger equation with the small parameter $\epsilon = 0.1$ or $\epsilon = 0.01$:

$$\begin{aligned} i\epsilon\frac{\partial u}{\partial t} &= -\tfrac{1}{2}\epsilon^2\Delta u + Vu \quad && \text{for } \mathbf{x}\in\Omega \text{ and } t>0,\\ u(\mathbf{x},0) &= g(\mathbf{x}) && \text{for } \mathbf{x}\in\Omega. \end{aligned} \tag{6.4}$$

We are concerned here by the time-convergence rate for different smoothness conditions. We consider a smooth potential, a harmonic potential, and a hat function potential:

$$V_1(\mathbf{x}) = \prod_{s=1}^{d}(1-\cos(x_s)), \quad V_2(\mathbf{x}) = \frac{1}{2}\sum_{s=1}^{d}(x_s-\pi)^2, \quad V_3(\mathbf{x}) = \prod_{s=1}^{d}\left(1-\frac{|x_s|}{\pi}\right).$$

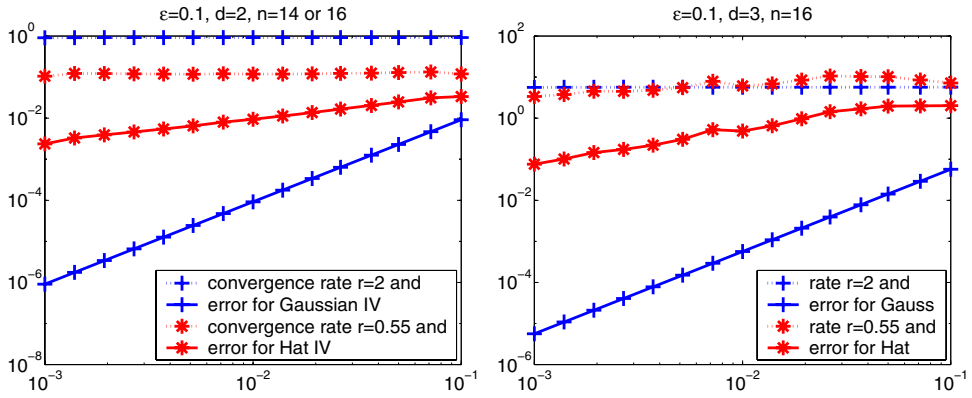


FIG. 6.1. *Error against time step for* $\epsilon = 0.1$*: Smooth potential.*

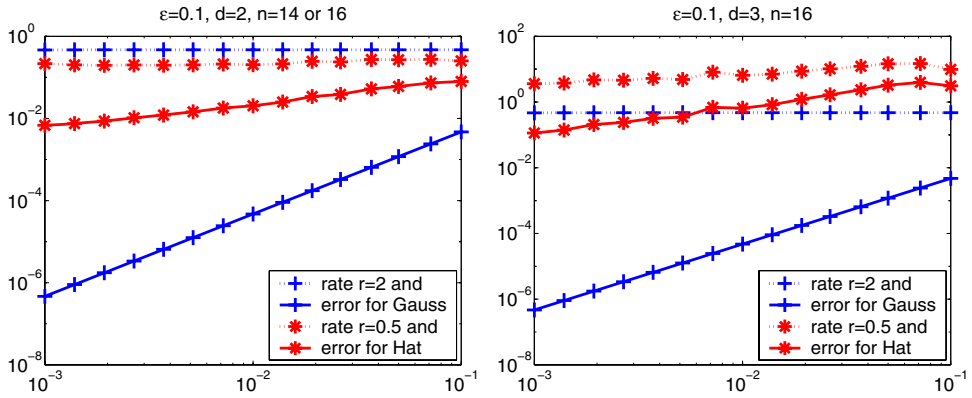


FIG. 6.2. *Error against time step for* $\epsilon = 0.1$*: Harmonic potential.*

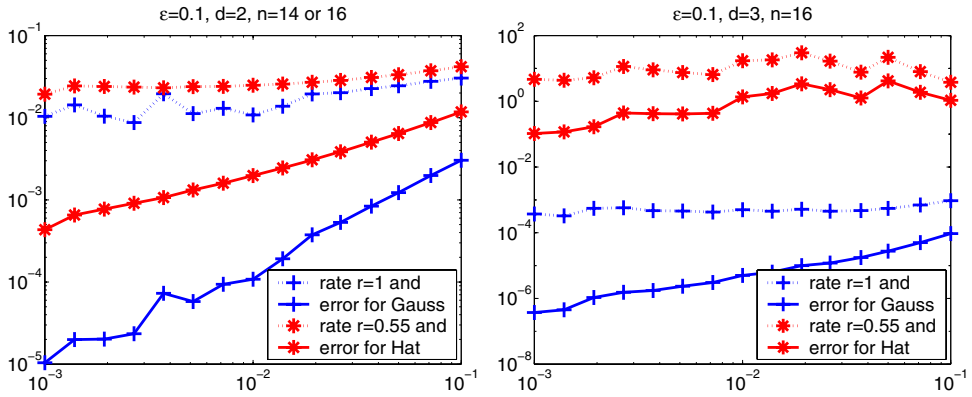


FIG. 6.3. *Error against time step for* $\epsilon = 0.1$*: Hat potential.*

The initial value is either the Gaussian

$$g(\mathbf{x}) = \left(\frac{2}{\pi\epsilon}\right)^{d/4} \exp\left(-\frac{\left(x_1 - \frac{3\pi}{2}\right)^2 + \sum_{s=2}^{d}(x_s - \pi)^2}{\epsilon}\right)$$

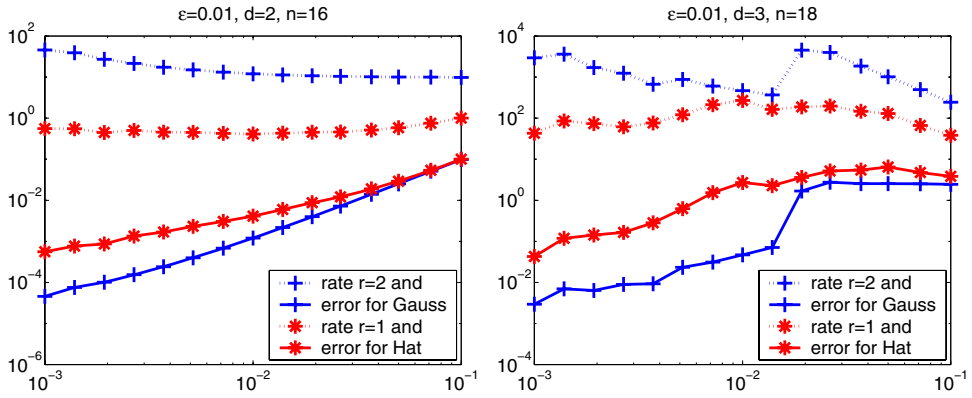


FIG. 6.4. *Error against time step for* $\epsilon = 0.01$*: Smooth potential.*

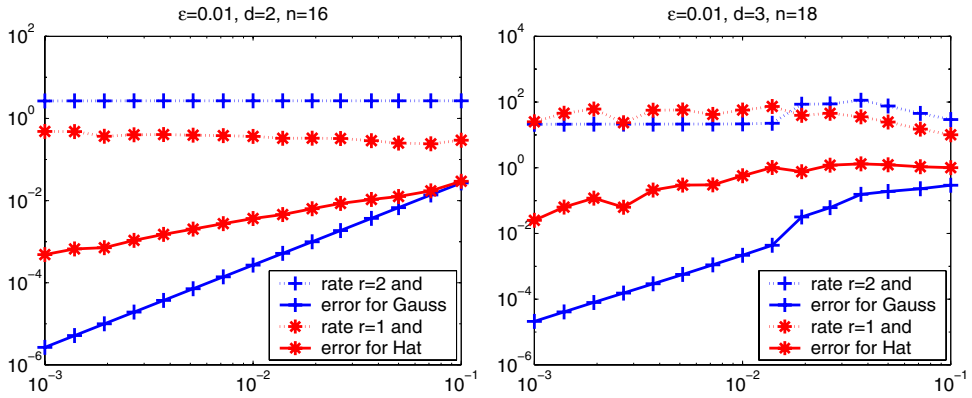


FIG. 6.5. *Error against time step for* $\epsilon = 0.01$*: Harmonic potential.*

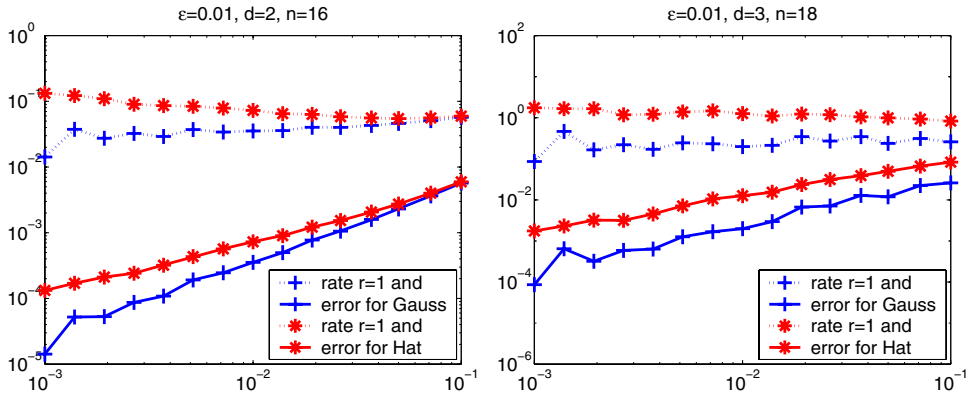


FIG. 6.6. *Error against time step for* $\epsilon = 0.01$*: Hat potential.*

or the hat function

$$h(\mathbf{x}) = \left(\frac{3}{\pi\sqrt{\epsilon}}\right)^{d/2} \cdot \left(1 - \left(\frac{2}{\pi\sqrt{\epsilon}}\right)\left|x_1 - \frac{3\pi}{2}\right|\right) \cdot \prod_{s=2}^{d}\left(1 - \left(\frac{2}{\pi\sqrt{\epsilon}}\right)\left|x_s - \frac{\pi}{2}\right|\right).$$

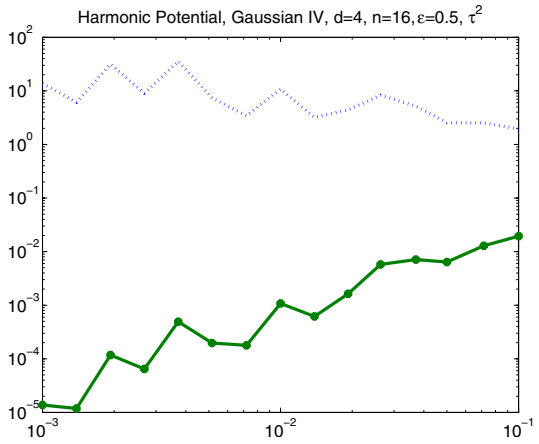


FIG. 6.7. *Error and convergence rate $r = 2$ against time step in four dimensions.*

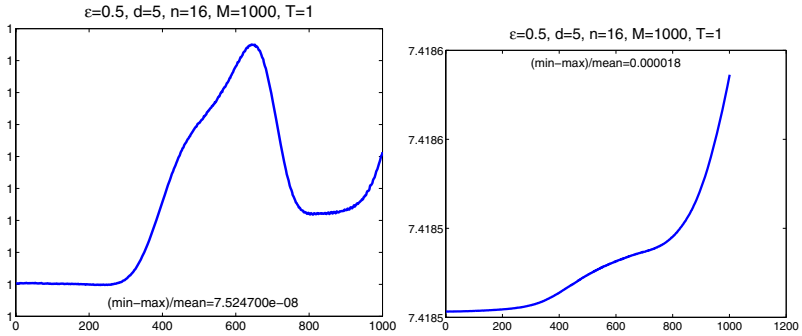


FIG. 6.8. *Time evolution of the norm and the energy for $d = 5$.*

Figures 6.1–6.7 show the norms $\left\|\widehat{u}_n^m - \widehat{u}_n^M\right\|_0$ at $t = m\tau = 1$ for $M = 10^4$ against different step sizes. The dashed lines indicate the errors divided by τ^r, with the convergence rate r. An aliasing error in the initial data may arise as an effect of multiple scales (that is, ϵ is small) or because of a too sparse grid in a high dimensional space. We try to avoid the aliasing in the interpolation of the initial values by adapting the resolution n to the dimension d and by reducing the computational domain Ω to a smaller expected domain. This results in the loss of smoothness of the potential V_1 at the ends of the interval and hence in the deterioration of the time convergence rate, as we can see in Figure 6.4. In the case of the harmonic oscillator on $[-\pi, \pi]^4$ with $\epsilon = 0.5$, the smoothness assumption does not hold on the boundary; this results in a slight deterioration of the convergence rate as shown in Figure 6.7. The improvement in the smoothness of the potential directly influences the numerical solution. The smoothness is inherent to this approach. At the same time, the accuracy is affected by the greater aliasing effect of the higher dimensions or of the smaller parameter ϵ. This is not surprising, since fixed domain truncation methods do not work well in a multiscaling context (see [5, Chapter 17]), and the sparse grids cannot cure this difficulty. For high resolutions of the space discretization we observed numerically the conservation of the L^2-norm for long times T (see Figure 6.8). This is expected if the convergence assumptions are fulfilled. What is remarkable is that the energy norm

$\langle u_n, (V - \frac{1}{2}\epsilon\Delta)u_n\rangle$ is almost conserved along the time iteration, too [15]. In the full grid case, the proof of the conservation properties resides on the orthogonality of the Fourier transform operator [22], a fact lacking here.

The exponential dependence on the dimension of the constants together with the need of a small parameter ϵ for relevant applications hamper us in using this technique on high dimensional problems. We may report simulations with norm conservation until dimension 5 and level 16.

7. Conclusion. Joined to the Strang splitting, the sparse grid discretization of the TDSE makes possible a good numerical solution in spaces of dimension up to 5. We proved first and second order convergence (in time) results under suitable smoothness assumptions. The method displays the typical efficiency of the approximations based on sparse grids. This paves the way for applications in computational chemistry such as conical crossings [24, 25]. A separate paper [15] addresses the concrete algorithmic details and more elaborate simulation results in higher dimensional spaces.

Acknowledgments. I am grateful to Christian Lubich for patient discussions on the Strang splitting and the commutator lemmas. I thank the referees for their very valuable comments and suggestions.

REFERENCES

[1] S. Achatz, *Adaptive finite Dünngitter-Elemente höherer Ordnung für elliptische partielle Differentialgleichungen mit variablen Koeffizienten*, Ph.D. thesis, TU München, Munich, Germany, 2003.
[2] N. Balakrishnan, C. Kalyanaraman, and N. Sathyamurthy, *Time-dependent quantum mechanical approach to reactive scattering and related processes*, Phys. Rep., 280 (1997), pp. 79–144.
[3] J. Bayfield, *Quantum evolution*, Wiley, New York, 1999.
[4] G. Bazenski and F. Delvos, *A Discrete Fourier Transform Scheme for Boolean Sums of Trigonometric Operators*, in Multivariate Approximation Theory IV, K. Z. C.K. Chui and W. Schempp, eds., Birkhäuser, Basel, 1989, pp. 15–24.
[5] J. Boyd, *Chebyshev and Fourier Spectral Methods*, 2nd ed., Dover, New York, 2000.
[6] H. Bungartz, *Finite Elements of Higher Order on Sparse Grids*, Shaker Verlag, Aachen, 1998.
[7] H.-J. Bungartz and M. Griebel, *Sparse grids*, Acta Numer., 13 (2004), pp. 147–269.
[8] E. Cancès, M. Defranceschi, W. Kutzelnigg, C. Le Bris, and Y. Maday, *Computational Quantum Chemistry: A Primer*, in Handbook of Numerical Analysis, P. Ciarlet and J. L. Lions, eds., Handb. Numer. Anal. 10, North–Holland, Amsterdam, 2003, pp. 3–270.
[9] C. Canuto, M. Hussaini, A. Quarteroni, and T. Zang, *Spectral Methods in Fluid Dynamics*, Springer-Verlag, New York, 1988.
[10] F. Delvos, *d–variate boolean interpolation*, J. Approx. Theory, 34 (1982), pp. 99–114.
[11] F. Delvos and W. Schempp, *Boolean methods in interpolation and approximation*, Pitman Research Notes in Mathematics 230, Longman Scientific and Technical, Harlow, UK, 1989.
[12] D. Donoho, *High-Dimensional Data Analysis: The Curses and Blessings of Dimensionality*, available online from http://www-stat. stanford. edu/donoho/Lectures/AMS2000/Curses. pdf, 2000.
[13] B. Fornberg, *A Practical Guide to Pseudospectral Methods*, Cambridge University Press, Cambridge, UK, 1996.
[14] J. Garcke and M. Griebel, *On the computation of the eigenproblems of hydrogen and helium in strong magnetic and electric fields with the sparse grid combination technique*, J. Comput. Phys., 165 (2000), pp. 694–716.
[15] V. Gradinaru, *Fourier transform on sparse grids: Code design and the time dependent Schrödinger equation*, Computing, 80 (2007), pp. 1–22.
[16] V. Gradinaru, *Whitney Elements on Sparse Grids*, Ph.D. thesis, Universität Tübingen, Tübingen, Germany, 2002. Also available online from http://w210.ub.uni-tuebingen.de/dbt/volltexte/2002/543/.
[17] M. Griebel and J. Hamaekers, *Sparse grids for the Schrödinger equation*, M2AN Math. Model. Numer. Anal., 41 (2007), pp. 215–247.

[18] M. Griebel, M. Schneider, and C. Zenger, *A combination technique for the solution of sparse grid problems*, in Iterative Methods in Linear Algebra, P. de Groen and R. Beauwens, eds., North–Holland, Amsterdam, 1992, pp. 263–281.
[19] G. Hagedorn and A. Joye, *A time-dependent Born-Oppenheimer approximation with exponentially small error estimates*, Comm. Math. Phys., 223 (2001), pp. 583–626.
[20] K. Hallatschek, *Fouriertransformation auf dünnen Gittern mit hierarchischen Basen*, Numer. Math., 63 (1992), pp. 83–97.
[21] S. Hilgenfeldt, R. Balder, and C. Zenger, *Sparse Grids: Applications to Multidimensional Schrödinger Problems*, Tech. rep., TU München, Munich, Germany, 1995.
[22] T. Jahnke and C. Lubich, *Error Bounds for Exponential Operator Splittings*, BIT, 40 (2000), pp. 735–744.
[23] F. Kupka, *Sparse Grid Spectral Methods for the Numerical Solution of Partial Differential Equations with Periodic Boundary Conditions*, Ph.D. thesis, Univerität Wien, Vienna, Austria, 1997.
[24] C. Lasser, T. Swart, and S. Teufel, *Construction and Validation of a Rigorous Surface Hopping Algorithm for Conical Crossings*, Tech. rep., Universität Tübingen, Tübingen, Germany, 2006.
[25] C. Lasser and S. Teufel, *Propagation through conical crossings: An asymptotic semigroup*, Comm. Pure Appl. Math., 58 (2005), pp. 1188–1230.
[26] A. Maynard, R. Wyatt, and C. Iung, *A quantum dynamical study of CH overtones in fluoroform.* I. *A nine-dimensional ab initio surface, vibrational spectra and dynamics*, J. Chem. Phys., 103 (1995), pp. 8372–8389.
[27] E. Novak, I. Sloan, and H. Wozniakowski, *Tractability of approximation for weighted Korobov spaces on classical and quantum computers*, Found. Comput. Math., (2004), pp. 121–156.
[28] C. Pflaum, *Diskretisierung elliptischer Differentialgleichungen mit dünnen Gittern*, Ph.D. thesis, TU München, Munich, Germany, 1996.
[29] G. Pöplau and F. Sprengel, *Some error estimates for periodic interpolation on full and sparse grids*, in Curves and Surfaces with Applications in CAGD, A. Le Méhauté, C. Rabut, and L. L. Schumaker, eds., Vanderbilt University Press, Nashville, TN, 1997, pp. 355–362.
[30] U. Rüde, H. Bungartz, and M. Griebel, *Extrapolation, combination, and sparse grid techniques for elliptic boundary value problems*, in Comput. Methods Appl. Mech. Engrg. 116 (1992), pp. 243–252.
[31] W. Schleich, *Quantum Optics in Phase Space*, Wiley, New York, 2001.
[32] H. Schmeisser and W. Sickel, *Spaces of functions of mixed smoothness and approximation from hyperbolic crosses*, J. Approx. Theory, (2004), pp. 115–150.
[33] H. Schmeisser and H. Triebel, *Topics in Fourier Analysis and Function Spaces*, Wiley, New York, 1987.
[34] W. Sickel, *Approximation from sparse grids and function spaces of dominating mixed smoothness*, in Approximation and Probability, Banach Center Publ. 672, Polish Academy of Sciences, Warsaw, Poland, 2006, pp. 271–283.
[35] F. Sprengel, *Interpolation and Wavelets on Sparse Gauß–Chebyshev Grids*, in Multivariate Approximation, Recent Trends and Results, W. Haußmann, K. Jetter, and M. Reimer, eds., Math. Res., 101, Akademie-Verlag, Berlin, 1997, pp. 269–286.
[36] V. Temlyakov, *Approximation of Periodic Functions*, Nova Science, New York, 1994.
[37] S. Teufel, *Adiabatic Perturbation Theory in Quantum Dynamics*, Lecture Notes in Math. 1821, Springer-Verlag Berlin, 2003.
[38] L. Trefethen, *Spectral Methods in MATLAB*, Software Environ. Tools 10, SIAM, Philadelphia, 2000.
[39] G. Wasilkowski and H. Wozniakowski, *Weighted tensor product algorithms for linear multivariate problems*, J. Complexity, 15 (1999), pp. 402–447.
[40] G. Worth and M. Robb, *Applying direct molecular dynamics to non-adiabatic systems*, Adv. Chem. Phys., 124 (2002), pp. 355–431.
[41] H. Yserentant, *The hyperbolic cross space approximation of electronic wavefunctions*, Numer. Math., 105 (2007), pp. 659–690.
[42] H. Yserentant, *On the regularity of the electronic Schrödinger equation in Hilbert spaces of mixed derivatives*, Numer. Math., 98 (2004), pp. 731–759.
[43] H. Yserentant, *Sparse grid spaces for the numerical solution of the electronic Schrödinger equation*, Numer. Math., 101 (2005), pp. 381–389.
[44] H. Yserentant, *Sparse grids, adaptivity, and symmetry*, Computing, 78 (2006), pp. 195–209.
[45] C. Zenger, *Sparse Grids*, Tech. rep. 18–90, SFB 342, TU München, Munich, Germany, 1990.

SIAM J. NUMER. ANAL.
Vol. 46, No. 1, pp. 124–150
© 2007 Society for Industrial and Applied Mathematics

L^∞- AND L^2-ERROR ESTIMATES FOR A FINITE VOLUME APPROXIMATION OF LINEAR ADVECTION*

BENOIT MERLET†

Abstract. We study the convergence of the upwind finite volume scheme applied to the linear advection equation with a Lipschitz divergence-free speed in $\mathbf{R}^d$. We prove an $h^{1/2-\varepsilon}$-error estimate in the $L^\infty(\mathbf{R}^d \times [0,T])$-norm for Lipschitz initial data. The expected optimal result is an $h^{1/2}$-error estimate. In a second part, we also prove an $h^{1/2}$-error estimate in the $L^\infty(0,T;L^2(\mathbf{R}^d))$-norm for initial data in $H^1(\mathbf{R}^d)$.

Key words. scalar conservation laws, advection equation, finite volume method, error estimate

AMS subject classifications. 35L65, 65M15

DOI. 10.1137/060664057

1. Introduction.

We consider here the numerical approximation by the explicit upwind finite volume (FV) scheme of the following general nonlinear hyperbolic scalar problem in d-dimensional space,

$$\left\{\begin{array}{rcccl} u_t + \operatorname{div}(Vf(u)) & = & 0 & \text{on} & \mathbf{R}_x^d \times \mathbf{R}_{+t}, \\ u(x,0) & = & u_0(x) & \text{on} & \mathbf{R}^d, \end{array}\right. \tag{1.1}$$

where the vector field V is a Lipschitz function of x and t with values in $\mathbf{R}^d$ and is divergence free.

FV schemes are widely used for the numerical approximation of such equations and more generally of conservative hyperbolic systems. They have several good behaviors: they can handle general meshes, they have a compact numerical stencil, they conserve the quantity $\int_{\mathbf{R}^d} u$, and under some hypothesis (CFL condition) they have stability properties. Moreover, in general, the solutions to (1.1) are not unique, and, again under some hypothesis concerning the time step and the regularity of the mesh, FV schemes do converge towards the physical (entropic) solution when the mesh size h tends to 0.

FV methods are also used to approximate convection-diffusion systems—see, e.g., [8, 9] for error estimates. In general the diffusion term provides energy estimates, the situation is more difficult if in some regions of the domain, the solution is derived by the convection terms. In this case, techniques developed for hyperbolic problems may be useful—see [10, 14] and the references therein.

However, the convergence theory for FV schemes applied to (1.1) is far from being complete. The first result is due to Kuznecov [11], who proved an $h^{1/2}$-error estimate in the $L^\infty(0,T;L^1_x)$-norm for BV data and a family of structured Cartesian meshes of mesh size h tending to 0. The method of Kuznecov has been extended to the case of nonregular Cartesian meshes or almost Cartesian meshes (see [3]). This method has also been extended to a family of unstructured uniformly regular meshes, but in this last case only an $h^{1/4}$-convergence rate is obtained (see Cockburn, Coquel,

*Received by the editors June 30, 2006; accepted for publication (in revised form) May 11, 2007; published electronically December 20, 2007.

http://www.siam.org/journals/sinum/46-1/66405.html

†Université Paris Nord - Institut Galilée, LAGA (Laboratoire d'Analyse, Géométrie, et Applications), Avenue J. B. Clément, 93430 Villetaneuse, France (merlet@math.univ-paris13.fr).

and LeFloch [2], Vila [15], and Eymard, Gallouët, and Herbin [7]). We emphasize that for BV initial data, numerical experiments show that the error behaves like $h^{1/2}$ regardless of whether the family of meshes is structured.

Remark 1.1. If the exact solution is more regular, the error may behave like h or even h^2. This supraconvergence has been linked to geometric properties of the meshes—see [1] and the references therein.

In this paper, we focus on the advection equation (i.e., the linear case, (1.1) with $f(u) := u$) with $W^{1,p}$ initial data. Even in this case, it is more difficult to establish convergence results than in the setting of finite difference (FD) methods. Indeed, in general, FV schemes are not consistent in the FD classical sense and Lax's theorem cannot be used. From numerical experiments and existing results on uniform Cartesian meshes, we expect an $h^{1/2}$ convergence rate in the $L^\infty(0,T;L^p_x)$-norm (provided that the CFL condition is satisfied and that the meshes are uniformly regular—see the condition (1.6) below).

In [12] the author and Vovelle proved this result for $p = 1$ (BV data) under a strict CFL condition in the general case and under a sharp CFL condition when the speed does not depend on the time variable.

In [5] Després showed the result in the case $p = 2$ under a strict CFL condition and for a constant speed. His proof is based on the study of the dissipation of the consistency error by the scheme and on a generalization of the Lax theorem. The lack of consistency of the scheme is compensated by the particular structure of the consistency error.

We also refer the reader to Vila and Villedieu [16], who proved (under a strict CFL condition) an $h^{1/2}$-error estimate in the L^2_{loc} space-time norm for the FV approximation of Friedrichs hyperbolic systems with H^1 data—the linear advection equation appears to be a particular case of these systems. Their proof is based on energy estimates and on the weak consistency of the scheme in the dual of H^1_{loc}.

The case $p > 2$ has been attacked by Després. In particular, in [6] an $h^{1/4-\varepsilon}$-estimate is obtained in the L^∞-norm for Lipschitz data.

Here, we study the case $2 \le p \le \infty$. For $p = 2$, we obtain the expected result under a strict CFL condition—Theorem 1.8. The ingredients of the proof are essentially contained in [12]: the conservation of the L^2-energy of the exact solution and energy estimates for the approximate solution coming from the dissipative properties of the scheme.

In the case $p = +\infty$, we prove an $h^{1/2-\varepsilon}$-error estimate in the L^∞-norm for Lipschitz data under a strict CFL condition—Theorem 1.6. In the proof, the fact that the scheme is consistent of order 1 for the dual norm of the BV seminorm is crucial—see section 2.3 for a precise description. This result is again the consequence of the structure of the consistency error: the size of the error is $\mathcal{O}(1)$, but the integral of this error on any finite union of cells ω is controlled by the measure of the boundary of ω times h—and not only by the measure of ω.

Remark 1.2. In general, in the nonlinear setting of (1.1), the flow is not continuous with respect to the $L^p(\mathbf{R}^d)$-norm when $p > 1$. Thus the extension of Theorems 1.6 and 1.8 to nonlinear problems does not seem relevant. Moreover, in this paper (and also in [12]) the linearity hypothesis is strongly used in several places. Nevertheless, in the case $p = 1$, one may hope proving error estimates for nonlinear problems via kinetic formulation of conservation laws. In this theory, (1.1) is replaced by a family of linear equations coupled by nonlinear constraints (see [13] for a result in this direction).

The proof of Theorem 1.6 may be sketched as follows. By duality, we are led to prove Theorem 1.7 concerning the behavior of the FV approximation when the initial data is a probability measure concentrated on one cell. Namely, if K_0 is this cell, the initial condition is $u_0 := 1/|K_0|\,\mathbf{1}_{K_0}$. Let u be the exact solution to the advection equation associated with this initial condition and let $\boldsymbol{u}$ be the corresponding FV approximation. Roughly speaking, Theorem 1.7 says that the following estimate holds:

$$\int_{\mathbf{R}^d\times\mathbf{R}^d} \|x-y\| u(x,T)\boldsymbol{u}(y,T)dxdy \leq Ch^{1/2-\varepsilon}. \tag{1.2}$$

For simplicity, assume that the speed is constant: $V(x,s)=\mathbf{V}$; in this case the exact solution is given by $u(\cdot,t) = 1/_{|K_0|}\,\mathbf{1}_{K_0+\mathbf{V}t}$. Letting $x_0\in K_0$ and $R>0$, we define $\Omega_R(t) := \{x : \|x-x_0-\mathbf{V}t\|\geq R\}$. The exact solution $u(\cdot,t)$ is concentrated in an h-neighborhood of $x_0+\mathbf{V}t$. So, in order to get (1.2), we have to prove that for $R\geq Ch^{1/2-\varepsilon}$, the quantity $\int_{\Omega_R(t)}\boldsymbol{u}(\cdot,t)$ remains small for $0<t\leq T$. For $k\geq 0$, we set $R_k := kh^{1/2}$ and $\omega_k(t) := \Omega_{R_k}(t)$. We will prove that when k tends to $+\infty$ the quantity $\int_{\omega_k(t)\setminus\omega_{k+1}(t)}\boldsymbol{u}(\cdot,t)$ decreases fast.

Via the Cauchy–Schwarz inequality, we are led to study the L^2-energy, $Q_k(t) := \int_{\omega_k(t)}\boldsymbol{u}^2(\cdot,t)$. At time 0, $Q_k(0)=0$. Next, the evolution of Q_k is driven by two phenomenons:

(1) the dissipation of the scheme decreases the energy and provides energy estimates,

(2) a source term located on the boundary of $\omega_k(t)$ may increase the energy.

We use the energy identity of section 2.2 and the weak consistency result of section 2.3 to prove (recursively on k) that $Q_k(t)$ is bounded by $t^k h^{-d}/k!$. In fact, in the proof, the characteristic functions of the sets $\omega_k(t)$ are replaced by smooth cutoff functions in the neighborhood of ∞.

The rest of the paper is organized as follows. First we supplement this introduction by describing the linear advection problem (section 1.1), the explicit upwind FV scheme (section 1.2), and the main results of the paper (section 1.3).

Section 2 is devoted to the proof of Theorems 1.6 and 1.7. In section 2.1, we prove by duality that Theorem 1.6 ensues from Theorem 1.7. In section 2.2, we establish an identity satisfied by the $L^2(\Psi^2 d\lambda(x))$-energy of the numerical solution—in the applications the weight Ψ^2 will be a cutoff function in the neighborhood of ∞. In section 2.3 we prove that the implicit in time downwind FV scheme is consistent of order 1 in a weak sense. Applying the results of sections 2.2 and 2.3, we prove in section 2.4 that if two solutions of the advection equation Ψ, $\overline{\Psi}$ satisfy $0\leq\Psi\leq 1$ and $\mathbf{1}_{|\operatorname{supp}\Psi}\leq\overline{\Psi}$, the growth of the $L^2(\Psi^2 d\lambda(x))$-energy of the numerical solution between times 0 and T is controlled by its $L^2(\overline{\Psi}^2 d\lambda(x)dt)$-energy. Finally, in section 2.5, we build a decreasing sequence of cutoff functions, we apply recursively the result of section 2.4 and we end the proof of Theorem 1.7.

Theorem 1.8 is proved in section 3. In section 3.1, we recall that the FV approximation satisfies the weak formulation of the advection equation up to an error term, we prove an energy inequality, and we give some technical results concerning some projections of the exact solution on the mesh. In section 3.2 we derive successive estimates yielding Theorem 1.8.

1.1. The linear advection equation. We consider the linear advection problem in $\mathbf{R}^d$,

$$\begin{cases} u_t + \operatorname{div}(Vu) = 0, & x \in \mathbf{R}^d, t \in \mathbf{R}_+, \\ u(x,0) = u_0(x), & x \in \mathbf{R}^d, \end{cases} \tag{1.3}$$

where we suppose that $V \in W^{1,\infty}(\mathbf{R}^d \times \mathbf{R}_+, \mathbf{R}^d)$ satisfies

$$\operatorname{div} V(\cdot,t) = 0 \qquad \forall t \in \mathbf{R}_+. \tag{1.4}$$

The problem (1.3) is well-posed in $L^1_{loc}(\mathbf{R}^d)$. Let us recall the following classical result (see Theorem 1 in [12] for a sketch of the proof).

THEOREM 1.1. *For every $u_0 \in L^1_{loc}(\mathbf{R}^d)$, the problem (1.3) has a unique weak solution u, in the sense that $u \in L^1_{loc}(\mathbf{R}_+ \times \mathbf{R}^d)$ is such that, for every compactly supported test function $\phi \in W^{1,\infty}(\mathbf{R}^d \times \mathbf{R}_+)$,*

$$\int_{\mathbf{R}_+} \int_{\mathbf{R}^d} u(\phi_t + V \cdot \nabla \phi) dx dt + \int_{\mathbf{R}^d} u_0(x)\, \phi(x,0) dx = 0. \tag{1.5}$$

This solution is given by the characteristic formula

$$u(x,t) = u_0(X(x,t)) \qquad \forall (x,t) \in \mathbf{R}^d \times \mathbf{R}_+$$

with X defined by $X(\cdot,t) := Y(\cdot,t)^{-1}$, where then Y solves the Cauchy problem $\partial_t Y(x,t) = V(Y(x,t),t)$; $Y(x,0) = x$.

Moreover, for every $t \geq 0$, $X(\cdot,t) : \mathbf{R}^d \to \mathbf{R}^d$ is one to one and onto, and the map X belongs to $W^{1,\infty}_{loc}(\mathbf{R}^d \times \mathbf{R}_+, \mathbf{R}^d)$. More precisely, for every $T \geq 0$, $X - Id_{\mathbf{R}^d_x}$ and $Y - Id_{\mathbf{R}^d_x}$ belong to $W^{1,\infty}(\mathbf{R}^d \times [0,T])$ and there exists $C_0 \geq 1$ depending only on T, $\|V\|_{W^{1,\infty}}$, and d such that $\forall x \in \mathbf{R}^d$, $\forall t \in [0,T]$,

$$\|X(x,t) - x\| \leq C_0\, t, \qquad \|\nabla X(x,t)\| \leq C_0, \qquad \|\partial_t X(x,t)\| \leq C_0,$$

$$\|Y(x,t) - x\| \leq C_0\, t, \qquad \|\nabla Y(x,t)\| \leq C_0, \qquad \|\partial_t Y(x,t)\| \leq C_0.$$

Finally, for every $t \geq 0$, $X(\cdot,t)$ preserves the Lebesgue measure λ on $\mathbf{R}^d$, i.e.,

$$\lambda(X(E,t)) = \lambda(E) \qquad \text{for every Borel subset } E \text{ of } \mathbf{R}^d.$$

For $t \geq 0$, the map $u_0 \mapsto u(\cdot,t)$ acts on $W^{1,\infty}(\mathbf{R}^d)$ as a continuous linear operator and if $u_0 \in W^{1,\infty}(\mathbf{R}^d)$, then $t \mapsto u(\cdot,t)$ is locally Lipschitz from $\mathbf{R}_+$ to $L^\infty(\mathbf{R}^d)$. More precisely, the characteristic formula yields the following.

COROLLARY 1.2. *Let $0 \leq t \leq T$, there exists a constant C_0 depending on $\|V\|_{W^{1,\infty}}$ and T such that if $u_0 \in W^{1,p}(\mathbf{R}^d)$, for some $1 \leq p \leq \infty$, then*

$$\|u(\cdot,s) - u(\cdot,t)\|_{L^p} \leq C_0 \|\nabla u_0\|_{L^p} |s-t|, \qquad 0 \leq s,t \leq T,$$

$$\|\nabla u(\cdot,t)\|_{L^p} \leq C_0 \|\nabla u_0\|_{L^p}, \qquad 0 \leq t \leq T.$$

1.2. FV scheme. The FV scheme which approximates (1.3) is defined on a mesh $\mathcal{T}$ which is a family of closed connected polygonal subsets with disjoint interiors covering $\mathbf{R}^d$; the time half-line is meshed by regular cells of size $\delta t > 0$. We also assume that the partition $\mathcal{T}$ satisfies the following properties: the common interface of two control volumes is included in a hyperplane of $\mathbf{R}^d$ and

$$\text{there exists } \alpha > 0 \text{ such that } \begin{cases} \alpha h^d \le |K|, \\ |\partial K| \le \alpha^{-1} h^{d-1} \quad \forall K \in \mathcal{T}, \end{cases} \tag{1.6}$$

where h is the size of the mesh: $h := \sup\{diam(K),\, K \in \mathcal{T}\}$, $|K|$ is the d-dimensional Lebesgue measure of K, and $|\partial K|$ is the $(d-1)$-dimensional Lebesgue measure of ∂K. If K and L are two control volumes having a common edge, we say that L is a neighbor of K and denote (quite abusively) $L \in \partial K$. We also denote by $K \vert L$ the common edge and by $\mathbf{n}_{KL}$ the unit normal to $K \vert L$ pointing outward from K. We set

$$V_{KL}^n := \frac{1}{\delta t} \int_{n\delta t}^{(n+1)\delta t} \int_{K \vert L} V \cdot \mathbf{n}_{KL}.$$

We denote by $K_n := K \times [n\delta t, (n+1)\delta t)$ a generic space-time cell and by $\mathcal{M} := \mathcal{T} \times \mathbf{N}$ the space-time mesh. The sets ∂K_n^-, ∂K_n^+ are defined by

$$\partial K_n^- := \{L \in \partial K, V_{KL}^n < 0\}, \qquad \partial K_n^+ := \{L \in \partial K, V_{KL}^n \ge 0\}.$$

In the same way we set $K \vert L_n := K \vert L \times [n\delta t, (n+1)\delta t)$.

We will assume that a strict CFL condition is satisfied:

$$\sum_{L \in \partial K_n^-} \delta t |V_{KL}^n| \le (1-\xi)|K| \qquad \forall K_n \in \mathcal{M} \text{ for some } 1 \ge \xi > 0. \tag{1.7}$$

Moreover, in order to avoid the occurrence of terms with factor $\delta t/h$ in our estimates, we add to (1.7) the following condition: there exists $c_0 \ge 0$ such that

$$\delta t \le c_0 h. \tag{1.8}$$

Remark 1.3. In general, (1.7) does not imply (1.8). Indeed, if for every nonnegative integer q, we have $\int_q^{q+1} V(\cdot, s)\, ds \equiv 0$, then the CFL condition (1.7) is satisfied with $\delta t = 1$.

The upwind FV method with explicit time discretization is defined by the following set of equations:

$$u_K^0 := \frac{1}{|K|} \int_K u_0(x)\, dx \qquad \forall K \in \mathcal{T}, \tag{1.9}$$

$$u_K^{n+1} := u_K^n + \frac{\delta t}{|K|} \sum_{L \in \partial K_n^-} -V_{KL}^n (u_L^n - u_K^n) \qquad \forall K_n \in \mathcal{M}. \tag{1.10}$$

We then denote by $\boldsymbol{u}$ the approximate solution of (1.3) defined by the FV scheme,

$$\boldsymbol{u}(\cdot, t) := u_K^n \quad \text{a.e. in K for every } t \in [n\delta t, (n+1)\delta t); \quad K_n \in \mathcal{M}. \tag{1.11}$$

Classical results. The divergence-free assumption (1.4) leads to the following identity.

LEMMA 1.3. *Under the hypothesis* (1.4) *on the divergence of* V, *one has*

$$\sum_{L\in\partial K} V_{KL}^n = \sum_{L\in\partial K_n^-} V_{KL}^n + \sum_{L\in\partial K_n^+} V_{KL}^n = 0 \qquad \forall\, K_n \in \mathcal{M}.$$

Under the CFL condition (1.7) the FV scheme is order preserving.

PROPOSITION 1.4. *Under condition* (1.7), *the linear application* $\mathcal{L} : (u_K^n) \mapsto (u_K^{n+1})$ *defined by* (1.10) *is order preserving and stable for the* L^∞*-norm.*

Proof. Equation (1.10) gives

$$u_K^{n+1} = \left(1 - \sum_{L\in\partial K_n^-} \frac{-V_{KL}^n \delta t}{|K|}\right) u_K^n + \sum_{L\in\partial K_n^-} \frac{-V_{KL}^n \delta t}{|K|} u_L^n;$$

i.e., under (1.7), u_K^{n+1} is a convex combination of $u_K^n, (u_L^n)_{L\in\partial K_n^-}$. The stability result follows from $\mathcal{L}(1) = 1$. ∎

From Lemma 1.3 it is not difficult to see that the quantity $\sum_{K\in\mathcal{T}} |K| u_K^n$ is conserved. By the Crandall–Tartar lemma [4], this fact and Proposition 1.4 imply the following.

PROPOSITION 1.5. *Under condition* (1.7), *the scheme* $\mathcal{L} : (u_K^n) \mapsto (u_K^{n+1})$ *is stable for the* L^1*-norm*,

$$\sum_{K\in\mathcal{T}} |K||u_K^{n+1}| \le \sum_{K\in\mathcal{T}} |K||u_K^n| \qquad \forall n \ge 0.$$

Interpolating between L^1*- and* L^∞*-stabilities, the scheme is stable for* L^p*-norms*, $1 < p < \infty$.

1.3. Main results. We now assume that $V \in W^{1,\infty}(\mathbf{R}^d \times \mathbf{R}_+, \mathbf{R}^d)$ satisfies (1.4), that $\mathcal{T}$ is a mesh of size $h > 0$ satisfying conditions (1.6), and that the time step $\delta t > 0$ satisfies the CFL conditions (1.7) and (1.8). We fix a time $T \ge \delta t$ and we assume that there exists an integer $N \ge 0$ such that $T = (N+1)\delta t$.

In what follows, given an initial condition $u_0 \in L^1_{loc}(\mathbf{R}^d)$, the function $u \in C(\mathbf{R}_+, L^1_{loc}(\mathbf{R}^d))$ denotes the exact solution to (1.3) and $\boldsymbol{u}$ its numerical approximation obtained by the upwind FV method (1.9)–(1.11). In the results and in the proofs, $C_0 \ge 1$ is the constant introduced in Theorem 1.1 and Corollary 1.2; this constant depends only on T, $\|V\|_{W^{1,\infty}}$, and d. The letter C denotes various constants which are nondecreasing functions of α^{-1}, c_0, $\|V\|_{W^{1,\infty}}$, and d but do not depend on h, δt, ξ, u_0, or T.

We have the following error estimate for Lipschitz data.

THEOREM 1.6. *Under a strict CFL condition* ($\xi > 0$), *there exists* $h_\star > 0$ *depending only on* C_0, ξ, T, *and* c_0 *such that if* $h \le h_\star$, *we have for every* $u_0 \in W^{1,\infty}(\mathbf{R}^d)$,

$$\|\boldsymbol{u}(\cdot,T) - u(\cdot,T)\|_{L^\infty} \le C C_0^2\, \xi^{-1/2} \|\nabla u_0\|_{L^\infty} T^{1/2} h^{1/2} \Gamma^{-1}(C_0^2 \xi^{-1} T h^{-1}),$$

where $\Gamma^{-1} : [1, +\infty) \to [2, +\infty)$ *is the inverse Euler* Γ *function—in particular* $\Gamma^{-1}(n!) = n + 1$.

This result will be the consequence of the following theorem.

THEOREM 1.7. *Let $K_0 \in \mathcal{T}$, $x_0 \in K_0$, and $u_0 := 1/|K_0|\,\mathbf{1}_{K_0}$; then under a strict CFL condition, for $h \le h_\star$, as above, we have*

$$\int_{\mathbf{R}^d} \|x - Y(x_0,T)\|\,\boldsymbol{u}(x,T)dx \le CC_0^2\xi^{-1/2}T^{1/2}h^{1/2}\Gamma^{-1}(C_0^2\xi^{-1}Th^{-1}).$$

Remark 1.4. The dependency on T in Theorems 1.6 and 1.7 is not interesting in the general case since C_0 depends on T. But in the case where the speed is independent of x: $V(x,t) := \mathbf{V}(t)$, we have $X(x,t) = x - \int_0^t \mathbf{V}(s)ds$ and we can choose $C_0 = \|\mathbf{V}\|_\infty$ in Theorem 1.1. This is also valid for Theorem 1.8 below.

Remark 1.5. The function Γ^{-1} is very slowly increasing (for example, $\Gamma^{-1}(R) = o(\ln R)$) but we do not obtain an $h^{1/2}$-estimate. Proving this result (or exhibiting a counterexample) remains an open problem.

In this L^2 setting, we need an additional assumption on the mesh: there is no small interface:

$$\alpha h^{d-1} \le |K|L| \qquad \forall K \in \mathcal{T},\ L \in \partial K. \tag{1.12}$$

THEOREM 1.8. *Assume that* (1.6), (1.12), (1.7), *and* (1.8) *hold. Then for every u_0 in $H^1(\mathbf{R}^d)$, we have*

$$\begin{aligned}
&\|u(\cdot,T) - \boldsymbol{u}(\cdot,T)\|_{L^2}^2 + \|\boldsymbol{u}(\cdot,0)\|_{L^2}^2 - \|\boldsymbol{u}(\cdot,T)\|_{L^2}^2 \\
&\qquad \le C(1+C_0^2)\xi^{-1}\,\|u_0\|_{H^1}^2\,Th + C\|u_0\|_{H^1}^2h^2, \\
&\mathcal{F}(\boldsymbol{u},T) \le C(1+C_0^2)\xi^{-2}\,\|u_0\|_{H^1}^2\,Th + C\|u_0\|_{H^1}^2h^2,
\end{aligned}$$

with

$$\mathcal{F}(\boldsymbol{u},T) := \sum_{n=0}^{N}\sum_{K\in\mathcal{T}} |K||u_K^{n+1} - u_K^n|^2 + \sum_{n=0}^{N}\sum_{K\in\mathcal{T}}\sum_{L\in\partial K_n^-} \delta t|V_{KL}^n|\,|u_L^n - u_K^n|^2.$$

Remark 1.6. Let us decompose the approximate solution $\boldsymbol{u}(\cdot,T)$ into collinear and orthogonal parts with respect to $u(\cdot,T)$: we write $\boldsymbol{u}(\cdot,T) = \lambda u(\cdot,T) + v(\cdot,T)$ with $\lambda \in \mathbf{R}$ and $v \in L^2(\mathbf{R}^d)$ such that $\int_{\mathbf{R}^d} u(\cdot,T)v = 0$.

We will see in the following that the L^2-norm of the discrete solution is nonincreasing. Since the L^2-energy of the exact solution is conserved, we deduce that $|\lambda| \le 1$. From the first estimate of Theorem 1.8 we have $0 \le \|u(\cdot,T)\|_{L^2}^2 - \|\boldsymbol{u}(\cdot,T)\|_{L^2}^2 + \|\boldsymbol{u}(\cdot,T) - u(\cdot,T)\|_{L^2}^2 \le \mathcal{O}(h)$, so

$$0 \ \le\ \left[(1-\lambda)^2 + (1-\lambda^2)\right]\|u(\cdot,T)\|_{L^2}^2 + \|v\|_{L^2}^2 \ \le\ \mathcal{O}(h).$$

Thus, $\lambda = 1 + \mathcal{O}(h)$ and $\|v\|_{L^2} = \mathcal{O}(h^{1/2})$. The main part of the error is orthogonal to $u(\cdot,T)$.

Remark 1.7. Using the technique of [12], we may prove that in the case of a constant speed $V(x,t) = \mathbf{V}$, the result also holds under a sharp CFL condition. If moreover the speed V is constant, then assumption (1.12) is not necessary—see Remark 3.2.

Remark 1.8. The conditions (1.6), (1.12) on the mesh are not stringent. They are satisfied in practice by every efficient mesh generator.

Conventions and notations. If (X,μ) is a measurable set with finite (positive) measure and $\phi \in L^1(X)$, we denote the mean value of ϕ over X by

$$\langle \phi \rangle_X := \fint_X \phi d\mu := \frac{1}{\mu(X)} \int_X \phi d\mu.$$

The characteristic function of a set X is denoted by $\mathbf{1}_X$.

For $R>0$, $B(R)$ denotes the open ball centered at 0 with radius R in $\mathbf{R}^d$.

In what follows, we will use several functions in $L^1_{loc}(\mathbf{R}^d \times \mathbf{R}_+)$ which are almost everywhere constant on each space-time cell K_n. We will use the following convention: when $(\phi_{K_n})_{K_n \in \mathcal{M}}$ is a family of real numbers, then the bold type letter $\boldsymbol{\phi}$ denotes the function $\boldsymbol{\phi} := \sum_{K \in \mathcal{M}} \phi_K^n \mathbf{1}_{K_n}$. When ϕ belongs to $C(\mathbf{R}^d \times \mathbf{R}_+, \mathbf{R})$, then, for $K_n \in \mathcal{M}$, $\phi_K^n := \langle \phi(\cdot, n\delta t) \rangle_K$ denotes the mean value of ϕ on K at time $n\delta t$ and in this case $\boldsymbol{\phi} = \sum_{K \in \mathcal{M}} \langle \phi(\cdot, n\delta t) \rangle_K \mathbf{1}_{K_n}$. This last convention holds for any letter φ but u (u is always the exact solution to (1.3) and $\boldsymbol{u}$ its FV approximation).

For every $K_n \in \mathcal{M}$, we set $V_K^n := -\sum_{L \in \partial K_n^-} V_{KL}^n = \sum_{L \in \partial K_n^+} V_{KL}^n$ and for every $t \geq 0$,

$$\mathcal{M}_t := \{K_n \in \mathcal{M} \ : \ 0 \leq n < t/\delta t\},$$

in particular, for $T = (N+1)\delta t$, $\mathcal{M}_T = \{K_n \ : \ K \in \mathcal{T}, 0 \leq n \leq N\}$.

Let $(\varphi_K^n)_{K_n \in \mathcal{M}}$ be a bounded family of real numbers. We define an energy associated with the weight $\boldsymbol{\varphi}$ and the FV solution $\boldsymbol{u}$,

$$\mathcal{E}(\boldsymbol{\varphi}, \boldsymbol{u}, t) := \sum_{K_n \in \mathcal{M}_t} \varphi_K^{n+1} e_K^n(\boldsymbol{u}),$$

where

$$e_K^n(\boldsymbol{u}) := \sum_{L \in \partial K_n^-} \left(1 - \frac{V_K^n \delta t}{|K|}\right) |V_{KL}^n| \delta t |u_L^n - u_K^n|^2 + \frac{1}{2} \sum_{L,M \in \partial K_n^-} \frac{V_{KL}^n V_{KM}^n \delta t^2}{|K|} |u_M^n - u_L^n|^2.$$

In section 3, we will write $\mathcal{E}(\boldsymbol{u}, t)$ for $\mathcal{E}(\mathbf{1}, \boldsymbol{u}, t)$.

We also define the local consistency error $Q_K^n(\boldsymbol{\Psi})$ of the implicit-in-time downwind FV scheme associated with (1.3) applied to a function $\Psi \in W^{1,\infty}(\mathbf{R}^d \times \mathbf{R}_+)$,

$$Q_K^n(\boldsymbol{\Psi}) := \frac{\Psi_K^{n+1} - \Psi_K^n}{\delta t} + \sum_{L \in \partial K_n^+} \frac{V_{KL}^n}{|K|} \left(\Psi_L^{n+1} - \Psi_K^{n+1}\right) \qquad \forall K_n \in \mathcal{M}.$$

2. Proof of Theorems 1.6 and 1.7.

2.1. Proof of Theorem 1.6. In this section, $u_0 \in W^{1,\infty}(\mathbf{R}^d)$. We have to estimate the quantity

$$\|\boldsymbol{u}(\cdot, T) - u(\cdot, T)\|_\infty = \sup_{K_0 \in \mathcal{T}} \sup_{x_0 \in K_0} \left| u_{K_0}^{N+1} - u(x_0, T) \right|.$$

Let $x_0 \in \mathbf{R}^d$ and let $K_0 \in \mathcal{T}$ such that $x_0 \in K_0$. Next, set $\boldsymbol{v}^0 := 1/|K_0| \, \mathbf{1}_{K_0}$ and define recursively for $K \in \mathcal{T}$ and $0 \leq n \leq N$,

$$v_K^{n+1} = v_K^n + \frac{\delta t}{|K|} \sum_{L \in \partial K_{N-n}^+} V_{KL}^{N-n} (v_L^n - v_K^n). \tag{2.1}$$

Since $u_{K_0}^{N+1} = \sum_{K\in\mathcal{T}} |K|\, u_K^{N+1} v_K^0$, (1.10) and repeated discrete integrations by parts yield

$$u_{K_0}^{N+1} = \sum_{K\in\mathcal{T}} \left(|K|\, u_K^N v_K^0 - \sum_{L\in\partial K_N^-} \delta t V_{KL}^N (u_L^N - u_K^N) v_K^0 \right) = \sum_{K\in\mathcal{T}} |K|\, u_K^N v_K^1$$

$$= \cdots = \sum_{K\in\mathcal{T}} |K|\, u_K^0 v_K^{N+1}.$$

Notice that the scheme (2.1) is the FV scheme with explicit time discretization associated with the dual continuous problem $v_t(x,t) + \operatorname{div}\{-V(x,T-t)v(x,t)\} = 0$ for $(x,t) \in \mathbf{R}^d \times \mathbf{R}_+$; $v(x,0) = \boldsymbol{v}^0(x)$ for $x \in \mathbf{R}^d$ whose solution at time T is $v(x,T) = \boldsymbol{v}^0(Y(x,T))$.

From the monotony and the L^1-stability of the scheme (2.1), we have $v_K^{N+1} \geq 0$ for every $K \in \mathcal{T}$ and $\sum_{K\in\mathcal{T}} |K| v_K^{N+1} = 1$; thus for every $x_0 \in K_0$, we have

$$\begin{aligned} \left| u_{K_0}^{N+1} - u(x_0,T) \right| &\leq \sum_{K\in\mathcal{T}} |K|\, |u_K^0 - u(x_0,T)| v_K^{N+1} \\ &= \sum_{K\in\mathcal{T}} |K|\, |u_K^0 - u_0(X(x_0,T))|\, v_K^{N+1} \\ &\leq \|\nabla u_0\|_\infty \left(\int_{\mathbf{R}^d} \|x - X(x_0,T)\| \boldsymbol{v}(x,T) dx + h \right), \end{aligned}$$

and therefore Theorem 1.6 is the consequence of Theorem 1.7 applied to the dual problem. □

Remark 2.1. In general, the dual problem associated with (1.3) is

$$\begin{cases} v_t + V(\cdot, T-\cdot)\cdot\nabla v = 0, & x \in \mathbf{R}^d, t \in [0,T], \\ v(x,0) = v_0(x), & x \in \mathbf{R}^d, \end{cases}$$

but since $\operatorname{div} V = 0$, we have $V\cdot\nabla v = \operatorname{div}(Vv)$ and the dual and original problems are similar.

2.2. Energy estimates. We now turn to the proof of Theorem 1.7. We begin by proving an energy identity.

LEMMA 2.1. *Let $u_0 \in L^2(\mathbf{R}^d)$, and let $(\varphi_K^n)_{K_n\in\mathcal{M}}$ be a family of real numbers. Then, for every $t \geq 0$,*

$$\int_{\mathbf{R}^d} \boldsymbol{u}^2(x,t)\, \boldsymbol{\varphi}(x,t) dx - \int_{\mathbf{R}^d} \boldsymbol{u}^2(x,0)\, \boldsymbol{\varphi}(x,0) dx + \mathcal{E}(\boldsymbol{\varphi}, \boldsymbol{u}, t)$$

$$= - \sum_{K_n\in\mathcal{M}_t} |K| \delta t (u_K^n)^2 Q_K^n(\boldsymbol{\varphi}).$$

Proof. We assume, without loss of generality, that $t = (n_0+1)\delta t$ for some $n_0 \in \mathbf{N}$. Let $n \geq 0$; we consider the quantity

$$e^n := \sum_{K\in\mathcal{T}} |K| \varphi_K^{n+1} (u_K^{n+1})^2 - \sum_{K\in\mathcal{T}} |K| \varphi_K^n (u_K^n)^2 \;=:\; A - B. \tag{2.2}$$

Using (1.10), we get

$$A = \sum_{K\in\mathcal{T}} |K|\varphi_K^{n+1} \left(\sum_{L\in\mathcal{V}_n^-(K)} a_{KL}^n u_L^n \right)^2, \tag{2.3}$$

where, for $K_n \in \mathcal{M}$, we have set $\mathcal{V}_n^-(K) := \{L \in \partial K_n^-\} \cup \{K\}$,

$$a_{KK}^n := 1 - V_K^n \delta t/|K|, \qquad \text{and for } L \in \partial K_n^-, \quad a_{KL}^n := -V_{KL}^n \delta t/|K|.$$

Next, using Lemma 1.3, we rewrite B as

$$B = \sum_{K\in\mathcal{T}} \left(|K| - V_K^n \delta t\right) \varphi_K^n (u_K^n)^2 - \sum_{K\in\mathcal{T}} \sum_{L\in\partial K_n^+} V_{LK}^n \delta t \varphi_K^n (u_K^n)^2.$$

In the last sum, we swap the indexes K and L to get

$$\begin{aligned} B &= \sum_{K\in\mathcal{T}} \left(|K| - V_K^n \delta t\right) \varphi_K^n (u_K^n)^2 + \sum_{K\in\mathcal{T}} \sum_{L\in\partial K_n^-} -V_{KL}^n \delta t \varphi_L^n (u_L^n)^2 \\ &= \sum_{K\in\mathcal{T}} |K| \sum_{L\in\mathcal{V}_n^-(K)} a_{KL}^n \varphi_L^n (u_L^n)^2 \\ &= \sum_{K\in\mathcal{T}} |K|\varphi_K^{n+1} \sum_{L\in\mathcal{V}_n^-(K)} a_{KL}^n (u_L^n)^2 + \sum_{K\in\mathcal{T}} \sum_{L\in\mathcal{V}_n^-(K)} |K| a_{KL}^n (\varphi_L^n - \varphi_K^{n+1})(u_L^n)^2. \end{aligned} \tag{2.4}$$

Thus, plugging (2.3) and (2.4) into (2.2), we get $e^n = \mathrm{I} + \mathrm{II}$, with

$$\mathrm{I} := \sum_{K\in\mathcal{T}} |K|\varphi_K^{n+1} \left\{ \left(\sum_{L\in\mathcal{V}_n^-(K)} a_{KL}^n u_L^n \right)^2 - \sum_{L\in\mathcal{V}_n^-(K)} a_{KL}^n (u_L^n)^2 \right\},$$

$$\mathrm{II} := \sum_{K\in\mathcal{T}} \sum_{L\in\mathcal{V}_n^-(K)} |K| a_{KL}^n (\varphi_K^{n+1} - \varphi_L^n)(u_L^n)^2.$$

For every $K_n \in \mathcal{M}$, using the identity $\sum_{M\in\mathcal{V}_n^-(K)} a_{KM}^n = 1$, an easy computation yields

$$\left(\sum_{L\in\mathcal{V}_n^-(K)} a_{KL}^n u_L^n \right)^2 - \sum_{L\in\mathcal{V}_n^-(K)} a_{KL}^n (u_L^n)^2 = -\frac{1}{2} \sum_{L,M\in\mathcal{V}_n^-(K)} a_{KL}^n a_{KM}^n (u_L^n - u_M^n)^2.$$

Thus, replacing the terms a_{KL}^n by their definitions and summing on $K \in \mathcal{T}$, we get

$$\mathrm{I} = -\sum_{K\in\mathcal{T}} \varphi_K^{n+1} e_K^n(\boldsymbol{u}).$$

Now, rearranging the terms in II, we compute

$$\mathrm{II} = \sum_{K\in\mathcal{T}} |K|\delta t (u_K^n)^2 \left\{ \frac{\varphi_K^{n+1} - \varphi_K^n}{\delta t} + \sum_{L\in\partial K_n^+} \frac{V_{KL}^n \delta t}{|K|} \left(\varphi_L^{n+1} - \varphi_K^{n+1}\right) \right\}.$$

Summing these two equations and then summing on $0 \le n \le n_0$, the lemma is proved. ∎

2.3. Weak consistency. In this section, the integer $0 \le n \le N$ is fixed. Let $(v_K)_{K\in\mathcal{T}}$ be a family of real numbers and set $\boldsymbol{v} := \sum_K v_K \mathbf{1}_K$; then let us define the weighted total variation of $\boldsymbol{v}$ with respect to the family $(V^n_{KL})_{K,L\in\mathcal{T}}$ by

$$\|\boldsymbol{v}\|_{V^n-TV} := \sum_{K\in\mathcal{T}} \sum_{L\in\partial K_n^-} |V^n_{KL}||v_K - v_L|.$$

Now for $\eta \in \mathbf{R}$, we define $\omega(\boldsymbol{v},\eta)$ to be the union of cells K such that $v_K < \eta < 0$ or $0 < \eta < v_K$ and we consider the level set decomposition $\boldsymbol{v}(x) = \int_{\mathbf{R}} \operatorname{sgn}(\eta)\, \mathbf{1}_{\omega(\boldsymbol{v},\eta)} d\eta$. When $\omega \subset \mathbf{R}^d$ is a finite union of elements of $\mathcal{T}$, we have $\|\mathbf{1}_\omega\|_{V^n-TV} = \frac{1}{2}\sum_{K|L\subset\partial\omega} |V^n_{KL}|$, and it is not difficult to see that

$$\|\boldsymbol{v}\|_{V^n-TV} = \int_{\mathbf{R}} \|\mathbf{1}_{\omega(\boldsymbol{v},\eta)}\|_{V^n-TV} d\eta.$$

This suggests the following definition. Let $(w_K)_{K\in\mathcal{T}} \subset \mathbf{R}$ and $\boldsymbol{w} := \sum_{K\in\mathcal{T}} w_K \mathbf{1}_K$. We set

$$\|\boldsymbol{w}\|'_{V^n-TV} := \sup_\omega \frac{\left|\int_\omega w(x)dx\right|}{\|\mathbf{1}_\omega\|_{V^n-TV}},$$

where the supremum is taken over all subsets $\omega \subset \mathbf{R}^d$ such that ω is a finite union of elements of $\mathcal{T}$. We do not prove that $\|\cdot\|'_{V^n-TV}$ is the dual norm associated with $\|\cdot\|_{V^n-TV}$. We will only need the following lemma, which is a direct consequence of the two last identities.

LEMMA 2.2. *Let $\boldsymbol{v}$, $\boldsymbol{w}$ be as above such that $\|\boldsymbol{v}\|_{V^n-TV} < \infty$ and $\|\boldsymbol{w}\|'_{V^n-TV} < \infty$. Then*

$$\int_{\mathbf{R}^d} \boldsymbol{v}(x)\boldsymbol{w}(x)dx = \sum_{K\in\mathcal{T}} |K| v_K w_K \le \|\boldsymbol{v}\|_{V^n-TV} \|\boldsymbol{w}\|'_{V^n-TV}.$$

We now prove the following consistency result for the implicit downwind FV scheme.

LEMMA 2.3. *Let $\Psi_0 \in W^{1,\infty}(\mathbf{R}^d)$. For $(x,t) \in \mathbf{R}^d \times [0,T]$, set $\Psi(x,t) := \Psi_0(X(x,t))$ and $\boldsymbol{Q}^n(\Psi) := \sum_{K\in\mathcal{T}} Q^n_K(\boldsymbol{\Psi})\mathbf{1}_{K_n}$. There exists $R^m_K(\Psi)$, $K_m \in \mathcal{M}$, such that for $n \le N$, $\boldsymbol{R}^n(\Psi) := \sum_{K\in\mathcal{T}} R^n_K(\Psi)\mathbf{1}_{K_n}$ and $\boldsymbol{S}^n(\Psi) := \boldsymbol{Q}^n(\Psi) - \boldsymbol{R}^n(\Psi)$ satisfy*

$$\|\boldsymbol{R}^n(\Psi)\|_\infty \le CC_0 \left(\|\nabla V\|_\infty + \|\partial_t V\|_\infty\right) \|\nabla\Psi_0\|_\infty h,$$

$$\|\boldsymbol{S}^n(\Psi)\|'_{V^n-TV} \le CC_0 \|V\|_\infty \|\nabla\Psi_0\|_\infty h.$$

Proof. For $K_m \in \mathcal{M}$, let us denote by Ψ^m_K the mean value of Ψ on K at time $m\delta t$: $\Psi^m_K = \langle \Psi(\cdot, m\delta t)\rangle_K$. Let $K \in \mathcal{T}$; since Ψ solves $\partial_t\Psi + \operatorname{div}(V\Psi) = 0$ on $\mathbf{R}^d \times \mathbf{R}_+$, we have

$$\frac{\Psi^{n+1}_K - \Psi^n_K}{\delta t} = \frac{1}{|K|\delta t}\int_{K_n} \partial_t\Psi(x,s)dxds$$

$$= -\frac{1}{|K|\delta t} \sum_{L\in\partial K} \int_{K|L_n} \Psi(x,s)V(x,s)\cdot\boldsymbol{n}_{KL}dxds.$$

For $K_m \in \mathcal{M}$, we set

$$R_K^m(\Psi) := \frac{1}{|K|\delta t} \sum_{L\in\partial K} \int_{K\vert L_m} \Psi(x,s)\Big\{\langle V\cdot \boldsymbol{n}_{KL}\rangle_{K\vert L_m} - V(x,s)\cdot \boldsymbol{n}_{KL}\Big\}dxds.$$

Since we have $\int_{K\vert L_m}\{\langle V\cdot \boldsymbol{n}_{KL}\rangle_{K\vert L_m} - V(x,s)\cdot \boldsymbol{n}_{KL}\}dxds = 0$ for every space-time edge $K\vert L_m$, we may replace $\Psi(x,s)$ by $\Psi(x,s) - \langle\Psi\rangle_{K\vert L_m}$ in the above equation. Then, for every $K\in\mathcal{T}$,

$$\begin{aligned}
|R_K^n(\Psi)| &\le \frac{1}{|K|\delta t}\sum_{L\in\partial K}\Bigg|\int_{K\vert L_n}\Big\{\Psi(x,s)-\langle\Psi\rangle_{K\vert L_n}\Big\}\\
&\qquad\qquad \times\Big\{V(x,s)\cdot\boldsymbol{n}_{KL} - \langle V\cdot\boldsymbol{n}_{KL}\rangle_{K\vert L_n}\Big\}dxds\Bigg|\\
&\le \frac{|\partial K|}{|K|}\big(\|\nabla\Psi\|_{L^\infty(\mathbf{R}^d\times[0,T])}h + \|\partial_t\Psi\|_{L^\infty(\mathbf{R}^d\times[0,T])}\delta t\big)\\
&\quad\times\big(\|\nabla V\|_\infty h + \|\partial_t V\|_\infty\delta t\big).
\end{aligned}$$

The regularity of the mesh (1.6) leads to $|\partial K|/|K| \le \alpha^{-2}h^{-1}$, and using the bounds on $\|\nabla X\|$ and $\|\partial_t X\|$ of Theorem 1.1 and the CFL condition (1.8), we get the first part of the lemma.

Next, since $V_{KL}^n = |K\vert L|\langle V\cdot\boldsymbol{n}_{KL}\rangle_{K\vert L_n}$, we have

$$\frac{\Psi_K^{n+1}-\Psi_K^n}{\delta t} - R_K^n = -\sum_{L\in\partial K}\frac{V_{KL}^n}{|K|}\langle\Psi\rangle_{K\vert L_n},$$

and using the identity of Lemma 1.3,

$$\begin{aligned}
&Q_K^n(\boldsymbol{\Psi}) - R_K^n(\Psi)\\
&\quad= -\sum_{L\in\partial K}\frac{V_{KL}^n}{|K|}\langle\Psi\rangle_{K\vert L_n} + \sum_{L\in\partial K_n^+}\frac{V_{KL}^n}{|K|}\big(\Psi_L^{n+1}-\Psi_K^{n+1}\big)\\
&\quad= \frac{1}{|K|}\left(\sum_{L\in\partial K_n^+} V_{KL}^n(\Psi_L^{n+1}-\langle\Psi\rangle_{K\vert L_n}) + \sum_{L\in\partial K_n^-} V_{KL}^n(\Psi_K^{n+1}-\langle\Psi\rangle_{K\vert L_n})\right).
\end{aligned}$$

Letting $\omega = \cup_{K\in I_\omega}K$ be a finite union of elements of $\mathcal{T}$, we have

$$\int_\omega \boldsymbol{S}^n(\Psi) = \sum_{K\in I_\omega}|K|(Q_K^n(\boldsymbol{\Psi}) - R_K^n(\Psi)).$$

If the cells K, L with $L\in\partial K_n^+$ are such that K and L belong to I_ω, then the edge $K\vert L$ has two contributions in the sum above: $V_{KL}^n(\Psi_L^{n+1}-\langle\Psi\rangle_{K\vert L_n})$ (due to $K\in I_\omega$, $L\in\partial K_n^+$) and $V_{LK}^n(\Psi_L^{n+1}-\langle\Psi\rangle_{L\vert K_n})$ (due to $L\in I_\omega$, $K\in\partial L_n^-$). These contributions cancel each other. Thus, if $\partial\omega^+$ (resp., $\partial\omega^-$) denotes the set $\{K\vert L,\, K\in I_\omega, L\in\partial K_n^+\setminus I_\omega\}$ (resp., $\{K\vert L,\, K\in I_\omega, L\in\partial K_n^-\setminus I_\omega\}$), we have

$$\int_\omega \boldsymbol{S}^n(\Psi) = \sum_{K\vert L\in\partial\omega^+} V_{KL}^n(\Psi_L^{n+1}-\langle\Psi\rangle_{K\vert L_n}) + \sum_{K\vert L\in\partial\omega^-} V_{KL}^n(\Psi_K^{n+1}-\langle\Psi\rangle_{K\vert L_n}),$$

and from the definition of $\|\cdot\|_{V^n-WTV}$, we get

$$\left|\int_\omega \boldsymbol{S}^n(\Psi)\right| \le \left(\frac{1}{2}\sum_{K\,|\,L\subset\partial\omega}|V_{KL}|\right)\left(\|\nabla\Psi\|_{L^\infty(\mathbf{R}^d\times[0,T])}h + \|\partial_t\Psi\|_{L^\infty(\mathbf{R}^d\times[0,T])}\delta t\right)$$

$$\le \|\mathbf{1}_\omega\|_{V^n-WTV}C_0(1+c_0)\|\nabla\Psi_0\|_\infty h,$$

which is the desired estimate. □

2.4. Key estimate. We use the results of sections 2.2 and 2.3 to estimate the growth of $\int_{\mathbf{R}}^d \boldsymbol{u}^2(x,t)\boldsymbol{\Psi}^2(x,t)dx$, when ψ is a cutoff function.

LEMMA 2.4. *Let Ψ_0, $\overline{\Psi}_0 \in W^{1,\infty}(\mathbf{R}^d)$ such that $0 \le \Psi_0 \le \overline{\Psi}_0 \le 1$. For $(x,t) \in \mathbf{R}^d \times \mathbf{R}_+$, let us define $\Psi(x,t) := \Psi_0(X(x,t))$ and $\overline{\Psi}(x,t) := \overline{\Psi}_0(X(x,t))$. Assume that $\overline{\Psi}_0 \equiv 1$ on $\operatorname{supp}\Psi_0 + B(0, C_0(2+c_0)h) = \{x \in \mathbf{R}^d \ : \ \Psi_{0|B(x,C_0(1+c_0)h)} \not\equiv 0\}$. Then, for every $0 \le t \le T$,*

$$\int_{\mathbf{R}^d} \boldsymbol{u}^2(\cdot,t)\,\boldsymbol{\Psi}^2(\cdot,t) \;-\; \int_{\mathbf{R}^d} u_0^2\,\boldsymbol{\Psi}^2(\cdot,0)$$

$$\le \left\{C_{1,\star}\|\nabla\Psi_0\|_\infty^2\, h + C_{2,\star}\|\nabla\Psi_0\|_\infty\, h\right\}\int_0^t\int_{R^d}\overline{\boldsymbol{\Psi}}^2\,\boldsymbol{u}^2$$

with $C_{1,\star} = CC_0^2\xi^{-1}$ and $C_{2,\star} = CC_0$.

Proof. Let $0 \le t \le T$; if $t < \delta t$, the left-hand side vanishes and there is nothing to prove. So we assume that there exists $n_0 \ge 0$ such that $(n_0+1)\delta t \le t < (n_0+2)\delta t$; since the left-hand side does not depend on t for $(n_0+1)\delta t \le t < (n_0+2)\delta t$ and the right-hand side is nondecreasing, we may assume that $t = (n_0+1)\delta t$. In this case $\mathcal{M}_t = \{K_n \in \mathcal{M} \ : \ 0 \le n \le n_0\}$.

Now, let $K \in \mathcal{M}$ and $0 \le n \le n_0+1$ such that $\Psi_K^n \ne 0$. There exists $x \in K$ such that $\Psi(x, n\delta t) \ne 0$; thus $\Psi_0(X(x,n\delta t)) \ne 0$. Let $z \in \mathbf{R}^d$ and $0 \le m \le N+1$ such that $|z-x| \le 2h$ and $|n-m| \le 1$. From the estimates on ∇X and $\partial_t X$ of Theorem 1.1, we have $\|X(x,n\delta t) - X(z,m\delta t)\| \le C_0(2h+\delta t) \le C_0(2+c_0)h$; thus $\overline{\Psi}(z,s) = \overline{\Psi}_0(X(z,s)) = 1$. We conclude that for every $K_n, L_m \in \mathcal{M}$ such that $0 \le n, m \le n_0+1$, we have

$$\left(L \in \partial K \cup \{K\}, \quad |m-n| \le 1, \quad \Psi_K^n > 0\right) \implies \overline{\Psi}_L^m = 1. \tag{2.5}$$

Let us write the identity of Lemma 2.1, with $\boldsymbol{\varphi} = \boldsymbol{\Psi}^2$:

$$\mathrm{I} := \int_{\mathbf{R}^d} \boldsymbol{u}^2(\cdot,t)\,\boldsymbol{\Psi}^2(\cdot,t) - \int_{\mathbf{R}^d} u_0^2\,\boldsymbol{\Psi}^2(\cdot,0) + \mathcal{E}(\boldsymbol{\Psi}^2,\boldsymbol{u},t) \tag{2.6}$$

$$= -\sum_{K_n\in\mathcal{M}_t} |K|\,\delta t (u_K^n)^2 Q_K^n(\boldsymbol{\Psi}^2).$$

Let $K_n \in \mathcal{M}$; a direct computation yields

$$Q_K^n(\boldsymbol{\Psi}^2) = 2\Psi_K^{n+1}Q_K^n(\boldsymbol{\Psi}) - \frac{(\Psi_K^{n+1}-\Psi_K^n)^2}{\delta t} + \sum_{L\in\partial K_n^+}\frac{V_{KL}^n}{|K|}(\Psi_L^{n+1}-\Psi_K^{n+1})^2.$$

We then have

$$\text{I} \le \sum_{K_n\in\mathcal{M}_t} |K|(u_K^n)^2(\Psi_K^{n+1}-\Psi_K^n)^2 - 2\sum_{K_n\in\mathcal{M}_t}\delta t|K|\Psi_K^{n+1}(u_K^n)^2 Q_K^n(\boldsymbol{\Psi}) \tag{2.7}$$

$$=: A - 2\sum_{n=0}^{n_0}\delta t B_n.$$

First, from the estimate on $\partial_t X$ of Theorem 1.1, we have $|\Psi_K^{n+1}-\Psi_K^n| \le C_0\|\nabla\Psi_0\|_\infty\delta t \le C_0c_0\|\nabla\Psi_0\|_\infty h$ for every $K_n\in\mathcal{M}_t$. Using (2.5), we deduce that

$$A \le CC_0^2\|\nabla\Psi_0\|_\infty^2\, h\sum_{K_n\in\mathcal{M}_t}|K|\,\delta t(u_K^n)^2(\overline{\Psi}_K^n)^2 \tag{2.8}$$

$$= CC_0^2\|\nabla\Psi_0\|_\infty^2\, h\int_0^t\int_{R^d}\overline{\boldsymbol{\Psi}}^2\,\boldsymbol{u}^2.$$

Next, let us fix $0\le n\le n_0$. With the notation of Lemma 2.3, we have

$$B_n = \sum_{K\in\mathcal{T}}|K|\Psi_K^{n+1}(u_K^n)^2R_K^n(\Psi) + \sum_{K\in\mathcal{T}}|K|\Psi_K^{n+1}(u_K^n)^2S_K^n(\Psi) \tag{2.9}$$

$$=: B_n^1 + B_n^2,$$

where $\|\boldsymbol{R}^n(\Psi)\|_\infty \le CC_0\|\nabla\Psi_0\|_\infty\, h$ and $\|\boldsymbol{S}^n(\Psi)\|'_{V^n-TV} \le CC_0\|\nabla\Psi_0\|_\infty h$. Thus,

$$\left|B_n^1\right| \le CC_0\|\nabla\Psi_0\|_\infty\, h\int_{\mathbf{R}^d}\overline{\boldsymbol{\Psi}}^2\,\boldsymbol{u}^2(\cdot,n\delta t). \tag{2.10}$$

Next, using Lemma 2.2 with $\boldsymbol{v}_K^n = \Psi_K^{n+1}(u_K^n)^2$, for $K\in\mathcal{T}$ and $\boldsymbol{w} = \boldsymbol{S}^n(\Psi)$, we get

(2.11)

$$\left|B_n^2\right| \le CC_0\|\nabla\Psi_0\|_\infty h\sum_{K\in\mathcal{T}}\sum_{L\in\partial K_n^-}|V_{KL}^n|\,\left|\Psi_K^{n+1}(u_K^n)^2 - \Psi_L^{n+1}(u_L^n)^2\right|$$

$$\le CC_0\|\nabla\Psi_0\|_\infty\, h\left(\sum_{K\in\mathcal{T}}\sum_{L\in\partial K_n^-}|V_{KL}^n|\left|\Psi_L^{n+1}-\Psi_K^{n+1}\right|(u_L^n)^2\right.$$

$$\left. + \sum_{K\in\mathcal{T}}\sum_{L\in\partial K_n^-}|V_{KL}^n|\,\Psi_K^{n+1}\left|(u_K^n)^2-(u_L^n)^2\right|\right)$$

$$=: CC_0\|\nabla\Psi_0\|_\infty\, h\,(C_n+D_n).$$

From (2.5), the estimate on $\|\nabla X\|$ of Theorem 1.1, and (1.6), we bound C_n by

$$C_n \le \sum_{K\in\mathcal{T}}\sum_{L\in\partial K_n^-}(\|V\|_\infty\,|K\shortmid L|)\,(C_0\|\nabla\Psi_0\|_\infty 2h)\,(u_L^n)^2(\overline{\Psi}_L^n)^2$$

$$= 2C_0\|\nabla\Psi_0\|_\infty\|V\|_\infty\sum_{K\in\mathcal{T}}\left(\sum_{L\in\partial K_n^+}|K\shortmid L|h\right)(u_K^n)^2(\overline{\Psi}_K^n)^2$$

$$\le 2C_0\|\nabla\Psi_0\|_\infty\|V\|_\infty\alpha^{-2}\sum_{K\in\mathcal{T}}|K|(u_K^n)^2\left(\overline{\Psi}_K^n\right)^2,$$

and finally,

$$C_n \le CC_0\|\nabla\Psi_0\|_\infty \int_{\mathbf{R}^d} \boldsymbol{u}^2\overline{\boldsymbol{\Psi}}^2(\cdot, n\delta t). \tag{2.12}$$

Writing

$$\begin{aligned}\Psi_K^{n+1}\left|(u_K^n)^2-(u_L^n)^2\right| &= \left|\Psi_K^{n+1}(u_K^n-u_L^n)(u_K^n+u_L^n)\right| \\ &\le \left|\Psi_K^{n+1}(u_K^n-u_L^n)\right| \times \left(\left|\overline{\Psi}_K^{n+1}u_K^n\right| + \left|\overline{\Psi}_L^{n+1}u_L^n\right|\right)\end{aligned}$$

and using the Cauchy–Schwarz inequality, we bound D_n by

$$\begin{aligned} D_n &\le \left(\sum_{K\in\mathcal{T}}\sum_{L\in\partial K_n^-} |V_{KL}^n|\,|u_K^n-u_L^n|^2\,(\Psi_K^{n+1})^2\right)^{1/2} \\ &\quad\times\left(2\sum_{K\in\mathcal{T}}\sum_{L\in\partial K} |V_{KL}^n|\,|u_K^n|^2\left(\overline{\Psi_K}^{n+1}\right)^2\right)^{1/2} \\ &\le \delta t^{-1/2}\left(\sum_{K\in\mathcal{T}}\sum_{L\in\partial K_n^-} |V_{KL}^n|\delta t\,|u_L^n-u_K^n|^2(\Psi_K^{n+1})^2\right)^{1/2} \\ &\quad\times\|V\|_\infty^{1/2}\alpha^{-1}h^{-1/2}\left(2\sum_{K\in\mathcal{T}}|K|(u_K^n)^2\left(\overline{\Psi}_K^n\right)^2\right)^{1/2} \\ &\le Ch^{-1/2}\delta t^{-1/2}\left(\sum_{K\in\mathcal{T}}\sum_{L\in\partial K_n^-} |V_{KL}^n|\delta t\,|u_L^n-u_K^n|^2(\Psi_K^{n+1})^2\right)^{1/2} \\ &\quad\times\left(\int_{\mathbf{R}^d}\boldsymbol{u}^2\overline{\boldsymbol{\Psi}}^2(\cdot,n\delta t)\right)^{1/2}. \end{aligned} \tag{2.13}$$

Collecting (2.9)–(2.13), we get

$$\begin{aligned}|B_n| &\le CC_0\|\nabla\Psi_0\|_\infty h^{1/2}(C_0\|\nabla\Psi_0\|_\infty h^{1/2}+Ch^{1/2})\int_{\mathbf{R}^d}\boldsymbol{u}^2\overline{\boldsymbol{\Psi}}^2(\cdot,n\delta t) \\ &\quad+CC_0\|\nabla\Psi_0\|_\infty\, h^{1/2}\delta t^{-1/2}\left(\sum_{K\in\mathcal{T}}\sum_{L\in\partial K_n^-} |V_{KL}^n|\delta t\,|u_L^n-u_K^n|^2(\Psi_K^{n+1})^2\right)^{1/2} \\ &\quad\times\left(\int_{\mathbf{R}^d}\boldsymbol{u}^2\overline{\boldsymbol{\Psi}}^2(\cdot,n\delta t)\right)^{1/2}.\end{aligned} \tag{2.14}$$

Plugging (2.14) and (2.8) into (2.7) and using again the Cauchy–Schwarz inequality yield

$$\mathrm{I} \le CC_0\|\nabla\Psi_0\|_\infty h^{1/2}(C_0\|\nabla\Psi_0\|_\infty h^{1/2} + Ch^{1/2}) \int_0^t \int_{R^d} \overline{\mathbf{\Psi}}^2\, \boldsymbol{u}^2 \tag{2.15}$$

$$+CC_0\|\nabla\Psi_0\|_\infty\, h^{1/2} \left(\sum_{K_n\in\mathcal{M}_t} \sum_{L\in\partial K_n^-} |V_{KL}^n|\delta t\, |u_L^n - u_K^n|^2 (\Psi_K^{n+1})^2 \right)^{1/2}$$

$$\times \left(\int_0^t \int_{R^d} \overline{\mathbf{\Psi}}^2\, \boldsymbol{u}^2 \right)^{1/2}.$$

To end the proof, we introduce the following lemma.

LEMMA 2.5. *For every $t \ge 0$, we have*

$$\sum_{K_n\in\mathcal{M}_t} \sum_{L\in\partial K_n^-} |V_{KL}^n|\delta t\, |u_L^n - u_K^n|^2 (\Psi_K^{n+1})^2 \le \xi^{-1}\mathcal{E}(\mathbf{\Psi}^2, \boldsymbol{u}, t).$$

Proof. It is sufficient to prove that, for every $K_n \in \mathcal{M}$, we have $1 \le \xi^{-1}(1 - V_K^n\delta t/|K|)$ which is a direct consequence of the CFL condition (1.7). ∎

Plugging (2.15) into (2.6) and using Lemma 2.5 and then Young's inequality ($ab \le \varepsilon a^2/2 + \varepsilon^{-1}b^2/2$) to absorb the term $\mathcal{E}(\mathbf{\Psi^2}, \boldsymbol{u}, t)$ in the left-hand side of (2.6), we obtain Lemma 2.4. ∎

2.5. Cutoff functions. Let $K_0 \in \mathcal{T}$, $x_0 \in K_0$ and define $u_0 := 1/|K_0|\,\mathbf{1}_{K_0}$. Without loss of generality, we assume that $x_0 = 0$.

With the notation of Lemma 2.4, we set

$$\sigma := 3C_{1,\star}^{1/2}T^{1/2}. \tag{2.16}$$

Let $\Psi_{0,0}$ be the constant function equal to 1 on $\mathbf{R}^d$, and for $k \ge 1$, let $\Psi_{k,0} \in W^{1,\infty}(\mathbf{R}^d, [0,1])$ be the unique cutoff function satisfying

$$\Psi_{k,0}\left(B(k+1/2)\sigma h^{1/2}\right) = \{0\}, \qquad \Psi_{k,0}\left(\mathbf{R}^d \setminus B((k+1)\sigma h^{1/2})\right) = \{1\},$$
$$\|\nabla\Psi_{k,0}\|_\infty \le \sigma^{-1}h^{-1/2}.$$

Then for $k \ge 0$ and $(x,t) \in \mathbf{R}^d \times \mathbf{R}_+$, set $\Psi_k(x,t) := \Psi_{k,0}(X(x,t))$.

We want to apply Lemma 2.4 with $(\Psi, \overline{\Psi}) = (\Psi_k, \Psi_{k-1})$, $k \ge 1$. The required hypothesis on the support of Ψ_k reads $C_0(2+c_0)h \le \sigma\, h^{1/2}/2$ and is true for $h \le h_{1,\star}$ small enough. Moreover, for $h \le h_{2,\star}$, we have $\int_{\mathbf{R}^d} u_0^2\, \mathbf{\Psi}_k^2(\cdot,0) = 0$, for $k \ge 1$, and from (2.16), for $h \le h_{3,\star}$, we have $4C_{1,\star}/\sigma^2 + 2C_{2,\star}h^{1/2}/\sigma \le T^{-1}$. We now assume $h \le h_\star := \min(h_{1,\star}, h_{2,\star}, h_{3,\star})$.

So, applying the lemma with $(\Psi, \overline{\Psi}) = (\Psi_k, \Psi_{k-1})$, we get

$$\int_{\mathbf{R}^d} \boldsymbol{u}^2(\cdot,t)\, \mathbf{\Psi}_k^2(\cdot,t) \le T^{-1} \int_0^t \int_{R^d} \boldsymbol{u}^2\, \mathbf{\Psi}_{k-1}^2, \qquad 0 \le t \le T, \quad k \ge 1.$$

Now, for $k = 0$, the L^2-stability of the scheme and (1.6) yield $\int_{\mathbf{R}^d} \boldsymbol{u}^2(\cdot,t)\, \mathbf{\Psi}_0^2(\cdot,t) \le \int_{\mathbf{R}^d} u_0^2 = 1/|K_0| \le \alpha^{-1}h^{-d}$ for every $t \ge 0$, and we obtain, recursively,

$$\int_{\mathbf{R}^d} \boldsymbol{u}^2(\cdot,t)\, \mathbf{\Psi}_k^2(\cdot,t) \le \alpha^{-1}\frac{(t/T)^k}{k!}h^{-d}, \qquad 0 \le t \le T,\ k \ge 0. \tag{2.17}$$

Let k_0 be a positive integer. Let $x \in \mathbf{R}^d$ such that $\|x\| \geq k_0\, \sigma h^{1/2}$, and k denotes the integer satisfying $k\sigma h^{1/2} \leq \|x\| < (k+1)\sigma h^{1/2}$. From the definition of $(\Psi_{j,0})_{j\geq 0}$, we have $(1 - \Psi_{k+1,0}(x))\, \Psi_{k-1,0}(x) = 1$ and since $\|x\| \leq (k+1)\sigma h^{1/2}$, we get

$$\|x\| \leq \sigma h^{1/2} \sum_{k\geq k_0} (k+1)\Psi_{k-1,0}(x)\,(1 - \Psi_{k+1,0}(x))\,.$$

So, for every $x \in \mathbf{R}^d$,

$$\|x\| \leq \sigma h^{1/2} \left(k_0 \mathbf{1}_{B(k_0\sigma h^{1/2})}(x) + \sum_{k\geq k_0} (k+1)\,(\Psi_{k-1,0}\,(1 - \Psi_{k+1,0}))\,(x) \right).$$

From the estimate on $\|\nabla Y\|$ of Theorem 1.1, we have $\|x - Y(0,T)\| \leq C_0 \|X(x,T)\|$. Thus,

$$\begin{aligned} &\|x - Y(0,T)\| \\ &\leq C_0 \sigma h^{1/2} \left(k_0 \mathbf{1}_{B(k_0\sigma h^{1/2})}(X(x,T)) + \sum_{k\geq k_0} (k+1)(\Psi_{k-1}\,(1 - \Psi_{k+1}))\,(x,T) \right). \end{aligned}$$

Multiplying by $\boldsymbol{u}(x,T)$ and integrating on $\mathbf{R}^d$, we get

$$\begin{aligned} \int_{\mathbf{R}^d} \|x - Y(0,T)\|\, \boldsymbol{u}(x,T)dx &\leq C_0\sigma h^{1/2} k_0 \int_{\mathbf{R}^d} \mathbf{1}_{B(\sigma h^{1/2})}(X(x,T))\, \boldsymbol{u}(x,T)dx \\ &+ C_0\sigma h^{1/2} \sum_{k\geq k_0} (k+1) \int_{R^d} (\boldsymbol{u}\, \Psi_{k-1}\,(1 - \Psi_{k+1}))\,(x,T)dx \\ &=:\ C_0\sigma h^{1/2} k_0\ A + C_0\sigma h^{1/2} \sum_{k\geq k_0} (k+1)\, B_k. \end{aligned} \tag{2.18}$$

From the monotony and conservation properties of (1.3), we have

$$A \leq 1. \tag{2.19}$$

For the terms B_k, we use the Cauchy–Schwarz inequality and (2.17). For every $k \geq 2$, we have

$$\begin{aligned} B_k &\leq \int_{R^d} (\boldsymbol{u}\Psi_{k-1}\,(1 - \Psi_{k+1}))\,(\cdot,T) \\ &\leq \left(\lambda \left\{ \mathrm{supp}\left(\Psi_{k-1}\,(1 - \Psi_{k+1}) \right)(\cdot,T) \right\} \right)^{1/2} \left(\int_{\mathbf{R}^d} \left(\boldsymbol{u}^2\, \Psi_{k-1}^2 \right)(\cdot,T) \right)^{1/2} \\ &\leq C\sigma^{d/2}(k+2)^{d/2} h^{d/4} \left(\int_{\mathbf{R}^d} \left(\boldsymbol{u}^2\, \boldsymbol{\Psi}_{k-2}^2 \right)(\cdot,T) \right)^{1/2} \\ &\leq C\sigma^{d/2}(k+2)^{d/2} h^{-d/4} ((k-2)!)^{-1/2}. \end{aligned} \tag{2.20}$$

Thus, (2.18)–(2.20) lead to

$$\int_{\mathbf{R}^d} \|x - Y(x_0,T)\| \boldsymbol{u}(x,T) dt$$

$$\leq CC_0 \sigma h^{1/2} \left(k_0 + \sigma^{d/2} h^{-d/4} \sum_{k \geq k_0} (k+2)^{(d-1)/2} ((k-2)!)^{-1/2} \right).$$

Clearly, $\sum_{k\geq k_0}(k+2)^{(d-1)/2}((k-2)!)^{-1/2} \overset{\infty}{\approx} ((k_0-d-1)!)^{-1/2} = (\Gamma(k_0-d))^{-1/2}$. Choosing k_0 to be the first integer such that $\sigma^d h^{-d/2}(\Gamma(k_0-d))^{-1} \leq 1$, we have

$$\int_{\mathbf{R}^d} \|x - Y(x_0,T)\| \boldsymbol{u}(x,T) dx \leq CC_0 \sigma h^{1/2}(k_0+1) \ \leq \ CC_0^2 \xi^{-1/2} T^{1/2} h^{1/2}(k_0+1).$$

And since $k_0 + 1 \leq \Gamma^{-1}(\sigma^d h^{-d/2}) + d + 2 \leq C\Gamma^{-1}(C_0^2 \xi^{-1} T h^{-1})$, Theorem 1.7 is proved. □

3. Proof of Theorem 1.8. In this section $u_0 \in L^2(\mathbf{R}^d)$. The function $u \in C(\mathbf{R}_+, L^2(\mathbf{R}^d))$ is the exact solution to (1.3) and $\boldsymbol{u}$ is the numerical approximation given by the scheme (1.9)–(1.11). We assume that (1.6) and the CFL conditions (1.7), (1.8) are satisfied with $\xi > 0$.

3.1. Weak formulation and technical results. The next lemma says that $\boldsymbol{u}$ satisfies the weak formulation (1.5) up to an error term, related to the consistency error of the scheme.

LEMMA 3.1 (Lemma 3.1 in [12]). *For every compactly supported test function $\phi \in W^{1,\infty}(\mathbf{R}^d \times \mathbf{R}_+)$, we have*

$$\int_{\mathbf{R}_+} \int_{\mathbf{R}^d} \boldsymbol{u}(\phi_t + V \cdot \nabla \phi) + \int_{\mathbf{R}^d} \boldsymbol{u}(\cdot,0)\phi(\cdot,0) = \mu(\phi) + \nu(\phi), \tag{3.1}$$

$$\mu(\phi) := \sum_{K_n \in \mathcal{M}} |K| (u_K^{n+1} - u_K^n)(\langle \phi \rangle_{K_n} - \langle \phi \rangle_K((n+1)\delta t)), \tag{3.2}$$

$$\nu(\phi) := \sum_{K_n \in \mathcal{M}} \sum_{L \in \partial K_n^-} \delta t (u_L^n - u_K^n)(V_{KL}^n \langle \phi \rangle_{K_n} - |K|L| \langle V \cdot \mathbf{n}\, \phi \rangle_{K|L_n}), \tag{3.3}$$

where $\langle \phi \rangle_{K|L_n} := \fint_{K|L_n} \phi V \cdot \mathbf{n}_{KL}$.

Remark 3.1. This equality is similar to the entropy inequality used in the Kuznecov method (see, e.g., [7]). Here, the initial problem (1.3) is linear and the notion of entropy is useless.

Proof. We develop the first term of the inequality and perform a discrete integration by parts,

$$\int_0^\infty \int_{\mathbf{R}^d} \boldsymbol{u}(\phi_t + V\cdot\nabla\phi)dxdt = \sum_{K_n\in\mathcal{M}} |K|\delta t\, u_K^n \fint_{K_n} (\phi_t + V\cdot\nabla\phi)dxdt \tag{3.4}$$

$$= \sum_{K_n\in\mathcal{M}} u_K^n \left(|K|\Big(\langle\phi\rangle_K((n+1)\delta t) - \langle\phi\rangle_K(n\delta t)\Big) + \sum_{L\in\partial K} \delta t|K\vert L|\langle V\cdot\mathbf{n}\,\phi\rangle_{KL}^n \right)$$

$$= \sum_{K_n\in\mathcal{M}} \left(|K|(u_K^n - u_K^{n+1})\langle\phi\rangle_K((n+1)\delta t) + \sum_{L\in\partial K} \delta t|K\vert L|\langle V\cdot\mathbf{n}\,\phi\rangle_{K\vert L_n} u_K^n \right)$$

$$- \sum_{K_n\in\mathcal{M}} |K|\delta t u_K^0\langle\phi\rangle_K(0).$$

The last sum is equal to $\int_{\mathbf{R}^d} \boldsymbol{u}(\cdot,0)\phi(\cdot,0)$. Rearranging the terms in the second sum and using the identity $\langle V\cdot\mathbf{n}\,\phi\rangle_{K\vert L_n} + \langle V\cdot\mathbf{n}\,\phi\rangle_{LK_n} = 0$, we compute, for every $n\geq 0$,

$$\sum_{K\in\mathcal{T}}\sum_{L\in\partial K} |K\vert L|\langle V\cdot\mathbf{n}\,\phi\rangle_{K\vert L_n} u_K^n = \sum_{K\in\mathcal{T}}\sum_{L\in\partial K_n^-} |K\vert L|\langle V\cdot\mathbf{n}\,\phi\rangle_{K\vert L_n}(u_K^n - u_L^n).$$

Plugging this identity and (1.10) into (3.4), we obtain (3.1)–(3.3). ∎

LEMMA 3.2. *Assume that the CFL condition* (1.7) *is satisfied with* $\xi\in(0,1)$; *then*

$$\mathcal{F}(\boldsymbol{u},T) \;\leq\; C\xi^{-1}\left(\|\boldsymbol{u}(\cdot,0)\|_{L^2(\mathbf{R}^d)}^2 - \|\boldsymbol{u}(\cdot,T)\|_{L^2(\mathbf{R}^d)}^2 \right).$$

Proof. From (1.10) and the Cauchy–Schwarz inequality, we have, for every $K_n\in\mathcal{M}$,

$$|K||u_K^{n+1} - u_K^n|^2 \leq \left(\sum_{L\in\partial K_n^-} \frac{|V_{KL}^n|\delta t}{|K|} \right) \sum_{L\in\partial K_n^-} |V_{KL}^n|\delta t|u_L^n - u_K^n|^2.$$

Due to the CFL condition (1.7), we have $\sum_{L\in\partial K_n^-} |V_{KL}^n|\delta t \leq |K|$; consequently, summing the preceding inequality on $K_n\in\mathcal{M}_T$, we obtain

$$\mathcal{F}(\boldsymbol{u},T) \leq 2\sum_{K_n\in\mathcal{M}_t}\sum_{L\in\partial K_n^-} |V_{KL}^n|\delta t|u_L^n - u_K^n|^2,$$

and we conclude by Lemma 2.1 applied with $\boldsymbol{\phi}=1$ and $t=T$. ∎

By continuity in H^1 of the L^2-projection on the space of functions which are constant with respect to the mesh $\mathcal{T}$, we have the following lemma.

LEMMA 3.3. *There exists* $c\geq 0$, *depending only on* d, *such that for every* $u_0\in H^1(\mathbf{R}^d)$,

$$\sum_{K\in\mathcal{T},\, L\in K_0^-} |K\vert L|\,|u_L^0 - u_K^0|^2 \leq c\alpha^{-4}\,\|\nabla u_0\|_{L^2}^2 h,$$

$$\|\boldsymbol{u}(\cdot,0) - u_0\|_{L^2} \leq c\alpha^{-1/2}\,\|\nabla u_0\|_{L^2} h.$$

Proof. Let $K \in \mathcal{T}$, $L \in \partial K$. We have

$$|u_K^0 - u_L^0| \leq \frac{1}{|K|\,|L|} \int_{K\times L} |u_0(x) - u_0(y)| dxdy.$$

Now from $|u_0(x) - u_0(y)| \leq \|x-y\| \int_0^1 |\nabla u_0((1-t)x + ty)| dt$, we have

$$|u_K^0 - u_L^0|^2 \leq \frac{1}{|K|\,|L|} \int_{K\times L} \|x-y\|^2 \int_0^1 |\nabla u_0((1-t)x+ty)|^2 dtdxdy$$

$$\leq \frac{4h^2}{|K|\,|L|} \int_{K\times L} \int_0^1 |\nabla u_0((1-t)x+ty)|^2 dtdxdy.$$

We perform the change of variables of Jacobian determinant $(-1)^d$, $(x,y,t) \mapsto (z = (1-t)x + ty, w = x - y, t = t)$, to obtain

$$|u_K^0 - u_L^0|^2 \leq \frac{4h^2}{|K|\,|L|} \int_{B(x_K,2h)} |\nabla u_0(z)|^2 \left(\int_0^1 \int_{\mathbf{R}^d} g(z,w,t) dwdt \right) dz,$$

where we have chosen $x_K \in K$ and where $g(z,w,t) = 1$ if $z + tw \in K$ and $z - (1-t)w \in L$, and $g(z,w,t) = 0$ otherwise. For $(z,t) \in B(x_K, 2h) \times [0,1]$, we have $\int_{\mathbf{R}^d} g(z,w,t) dw \leq 2^d |K|$ if $t \geq 1/2$ and $\int_{\mathbf{R}^d} g(z,w,t) dw \leq 2^d |L|$ if $t < 1/2$. Finally, we get

$$|u_K^0 - u_L^0|^2 \leq \frac{2^{d+2} h^2 \max(|K|,|L|)}{|K|\,|L|} \int_{B(x_K,2h)} |\nabla u_0(z)|^2 dz, \tag{3.5}$$

and the first inequality of (1.6) yields

$$|u_K^0 - u_L^0|^2 \leq 2^{d+2} h^{2-d} \alpha^{-1} \int_{B(x_K,2h)} |\nabla u_0(z)|^2 dz.$$

Multiplying by $|K|L|$, summing on $L \in \partial K$, and using the second part of (1.6), we get

$$\sum_{L\in\partial K} |K|L|\, |u_K^0 - u_L^0|^2 \leq M(K) 2^{d+1} h \alpha^{-2} \int_{B(x_K,2h)} |\nabla u_0(z)|^2 dz,$$

where $M(K)$ is the cardinal of $\{L : L \in \partial K\}$. From (1.6), we have

$$M(K)\alpha h^d \leq |\cup_{L\in\partial K} L| \leq |B(x_K, 3h)|, \tag{3.6}$$

and $M(K) \leq c\alpha^{-1}$. Summing on $K \in \mathcal{T}$, we obtain

$$\sum_{K\,,\,L\in\partial K} |K|L|\, |u_K^0 - u_L^0|^2 \leq ch\alpha^{-3} \int_{\mathbf{R}^d} |\nabla u_0(z)|^2 N(z) dz,$$

where $N(z)$ is the cardinal of $\{K : d(x_K, z) \leq 2h\}$. The integer $N(z)$ satisfies an inequality similar to (3.6) and so is bounded by $c\alpha^{-1}$. This leads to the first inequality of the lemma.

We now sketch the proof of the second inequality. Let $K \in \mathcal{T}$. Performing the computations and change of variables leading to (3.5) in the proof of the first part, we get successively

$$\begin{aligned}\int_K |\boldsymbol{u}(\cdot,0) - u_0|^2 &\le \frac{1}{|K|}\int_{K\times K} (u_0(x) - u_0(y))^2\,dx\,dy \\ &\le \frac{h^2}{|K|}\int_{K\times K}\int_0^1 |\nabla u_0((1-t)x+ty)|^2\,dt\,dx\,dy \\ &\le 2^d h^2 \int_{B(x_k,h)} |\nabla u_0(z)|^2.\end{aligned}$$

Summing on $K \in \mathcal{T}$ and using the fact that the cardinal of the set $\{K : d(K,z) \le h\}$ is bounded by $c\alpha^{-1}$, we get the second estimate. □

In what follows, we will need some estimates on the differences between u and its mean values on cells or on edges. They are contained in the next lemma. The three estimates are based on the trace theorem for H^1-functions. For (3.9), the conditions (1.6) are not sufficient, and we need hypothesis (1.12) to avoid the blowup of the constant in the trace theorem.

LEMMA 3.4. *Assume that* (1.6) *is satisfied. Then, for every* $u_0 \in H^1(\mathbf{R}^d)$ *we have*

$$\sum_{K_n\in\mathcal{M}_T}\sum_{L\in\partial K_n^-} \delta t|K|L|\,(\langle u\rangle_{K_n} - \langle u\rangle_{K|L_n})^2 \le CC_0^2\|\nabla u_0\|_{L^2}^2 Th, \tag{3.7}$$

$$\sum_{K_n\in\mathcal{M}_T}\sum_{L\in\partial K_n^-} \delta t|K|L| \left(\fint_{K|L_n}\fint_{n\delta t}^{(n+1)\delta t} |u(x,s)-u(x,t)|dsdtdx\right)^2 \le CC_0^2\|\nabla u_0\|_{L^2}^2 T\delta t. \tag{3.8}$$

If, moreover, assumption (1.12) *is satisfied, then*

$$\sum_{K_n\in\mathcal{M}_T}\sum_{L\in\partial K_n^-} \delta t|K|L| \left(\fint_{K|L_n}\fint_{K|L} |u(x,t)-u(y,t)|\,dxdydt\right)^2 \le CC_0^2\|\nabla u_0\|_{L^2}^2 Th. \tag{3.9}$$

Proof. We prove (3.7), (3.8), and (3.9) successively.

Let $K_n \in \mathcal{M}_T$, and let $L \in \partial K_n^-$. Let us fix $x_K \in K$ and $t \in [n\delta t, (n+1)\delta t)$. As in the proof of (3.5) in Lemma 3.3 (with L replaced by $B(x_K, 2h)$), we have

$$\left|\fint_K u(\cdot,t) - \fint_{B(x_K,2h)} u(\cdot,t)\right|^2 \le Ch^{2-d}\int_{B(x_K,2h)} |\nabla u(\cdot,t)|^2. \tag{3.10}$$

On the other hand, let P be the hyperplane containing $K|L$. We compute

$$\left(\fint_{K|L} u(\cdot,t) - \fint_{B(x_K,2h)} u(\cdot,t)\right)^2 \tag{3.11}$$

$$\leq \fint_{K|L} \left(u(x,t) - \fint_{B(x_K,2h)} u(y,t)\,dy\right)^2 dx$$

$$\leq \frac{1}{|K|L|}\int_{P\cap B(x_K,2h)} \left(u(x,t) - \fint_{B(x_K,2h)} u(y,t)\,dy\right)^2 dx$$

$$\leq\ C\frac{h}{|K|L|}\int_{B(x_K,2h)} |\nabla u(\cdot,t)|^2.$$

The last inequality is obtained by rescaling the following standard result. □

LEMMA 3.5. *Let $w \in H^1(B_{\mathbf{R}^d}(0,2))$ such that $\int_{B(0,2)} w = 0$, and let Q be a hyperplane such that $d(Q,0) \leq 1$; then the trace γ of w on $Q \cap B(0,2)$ satisfies $\|\gamma\|_{L^2} \leq C\|\nabla w\|_{L^2}$.*

Summing (3.10) and (3.11), integrating on $[n\delta t,(n+1)\delta t)$, and using (1.6), we get

$$\delta t|K|L|\left(\langle u\rangle_{K_n} - \langle u\rangle_{K|L_n}\right)^2 \leq C\int_{B(x_K,2h)\times[n\delta t,(n+1)\delta t)} |\nabla u|^2\ h.$$

Summing on $L \in \partial K_n^-$ and $K_n \in \mathcal{M}_T$ and using the fact that the cardinals of the sets $\{K\ :\ d(K,z) \leq 2h\}$ are uniformly bounded (see the proof of Lemma 3.3), we find that the left-hand side of (3.7) is bounded by $C\int_{\mathbf{R}^d\times[0,T]} |\nabla u|^2\, h$. We conclude by the second estimate of Corollary 1.2.

Before proving (3.8) and (3.9) we recall the definition of an $H^{1/2}$-seminorm. Let B be a ball in $\mathbf{R}^q$ centered at x_0, let P be an affine subspace of dimension $q-1$ containing x_0, and let $v \in H^1(B,\mathbf{R})$. Then γ, the trace of v on P, satisfies

$$|\gamma|^2_{H^{1/2}(B\cap P)} := \int_{(B\cap P)^2} \frac{|\gamma(x)-\gamma(y)|^2}{\|x-y\|^q} dxdy \leq C\int_B |\nabla v|^2. \tag{3.12}$$

Let $K_n \in \mathcal{M}_T$ and $L \in \partial K_n^-$. Let x_{KL} be any element of $K|L$ so that $K|L \subset B(x_{KL},h)$. We introduce new Cartesian coordinates: letting P be the affine subspace containing $K|L$, we define $(x'_1,\dots,x'_d)$ such that $x' \in P \Leftrightarrow x'_d = 0$ and we define $K|L'$ to be the set such that $x' \in K|L \Leftrightarrow (x'_d = 0$ and $\bar{x}' = (x'_1,\dots,x'_{d-1}) \in K|L')$.

From the Cauchy–Schwarz inequality, we have for every $\bar{x}' \in K|L'$

$$\left(\fint_{n\delta t}^{(n+1)\delta t}\fint_{n\delta t}^{(n+1)\delta t} |u((\bar{x}',0),s) - u((\bar{x}',0),t)|dsdt\right)^2$$

$$\leq \delta t^2 \fint_{n\delta t}^{(n+1)\delta t}\fint_{n\delta t}^{(n+1)\delta t} \frac{|u((\bar{x}',0),s) - u((\bar{x}',0),t)|^2}{|s-t|^2}\,dsdt$$

$$\leq C\int_0^{\delta t}\int_{n\delta t}^{(n+1)\delta t} |\partial_{x'_d} u((\bar{x}',x'_d),s)|^2 + |\partial_t u((\bar{x}',x'_d),s)|^2 ds\,dx'_d.$$

The last inequality comes from the trace theorem (3.12) applied with $q = 2$ and $P = \{x'_d = 0\}$ to the family of functions $(x'_d, s) \mapsto u((\bar{x}', x'_d), s)$. Integrating on $K|L'$ and using Jensen's inequality and the CFL condition $\delta t \le c_0 h$, we obtain

$$\left(\fint_{n\delta t}^{(n+1)\delta t} \fint_{K|L_n} |u(x,s) - u(u,t)| ds dt dx \right)^2 \le C \int_{n\delta t}^{(n+1)\delta t} \int_{B(x_{KL},(1+c_0)h)} |\nabla u|^2 + |\partial_t u|^2.$$

Multiplying by $\delta t |K|L|$ and summing on L, K_n, (3.8) follows from Corollary 1.2 and the fact that the cardinal of the neighbors of K is uniformly bounded.

We assume assumption (1.12) is satisfied. Let $K_n \in \mathcal{M}_T$, $L \in \partial K_n^-$, and $t \in [n\delta t, (n+1)\delta t]$. We denote again by P the affine subspace containing $K|L$ and $x_{KL} \in K|L$. From the Cauchy–Schwarz inequality, we have

$$\left(\fint_{(K|L)^2} |u(x,t) - u(y,t)| dx dy \right)^2 \le h^d \fint_{(K|L)^2} \frac{|u(x,t) - u(y,t)|^2}{\|x-y\|^d} dx dy,$$

and from the trace theorem (3.12),

$$\left(\fint_{(K|L)^2} |u(x,t) - u(y,t)| dx dy \right)^2 \le C \frac{h^d}{|K|L|^2} \int_{B(x_{KL},h)} |\nabla u(z,t)|^2 dz.$$

Integrating on $[n\delta t, (n+1)\delta t)$ and using Jensen's inequality and assumption (1.12) yield

$$\delta t |K|L| \left(\fint_{K|L_n} \fint_{K|L} |u(x,t) - u(y,t)| \, dx dy dt \right)^2 \le C \int_{B(x_{KL},h)} |\nabla u(z,t)|^2 dz \, \delta t \, h.$$

Summing on $L \in \partial K_n^-$ and $K_n \in \mathcal{M}_T$, the second inequality of Corollary 1.2 yields (3.9). □

3.2. End of the proof. We now assume that $u_0 \in H^1(\mathbf{R}^d)$ and end the proof of Theorem 1.8. Let us notice that from the L^2-stability of the scheme and of the evolution problem (1.3), it is sufficient to prove Theorem 1.8 for $u_0 \in C_c^\infty(\mathbf{R}^d)$ and conclude by density of such functions in $H^1(\mathbf{R}^d)$. In this case, the formula $u(x,t) = u_0(X(x,t))$ and the definition of X imply that $u \in W^{1,\infty}(\mathbf{R}^d \times [0, T+\delta t])$ and is compactly supported.

From now on, $u_0 \in C_c^\infty(\mathbf{R}^d)$. We want to apply the weak formulations (1.5) and (3.1) to a test function ϕ defined by $\phi(x,t) := \psi(t) u(x,t)$, for $x \in \mathbf{R}^d$, $s \ge 0$, with $\psi := \mathbf{1}_{[0,T]}$. Since the function $\mathbf{1}_{[0,T]}$ is not Lipschitz, we have to introduce a mollifying sequence to approximate ψ.

Let $(\psi_q)_{q \ge 1}$ be a sequence of $C^\infty(\mathbf{R}_+)$ functions satisfying $\psi_q \equiv 1$ on $[0,T]$, $\psi_q \equiv 0$ on $[T + \frac{\delta t}{q}, +\infty)$ and $0 \ge \psi_q' \ge \frac{-2q}{\delta t}$. We set $\phi_q(x,t) := u(x,t)\psi_q(t)$ for every $(x,t) \in \mathbf{R}^d \times \mathbf{R}_+$ and $q \ge 1$. Since u solves (1.3) and $\operatorname{div} V \equiv 0$, we have $\partial_t u + V \cdot \nabla u = \partial_t u + \operatorname{div}(Vu) = 0$ on $\mathbf{R}^d \times \mathbf{R}_+$. Thus, the weak formulation (3.1) reads

$$\int_T^{T+\delta t} \left(\int_{\mathbf{R}^d} \boldsymbol{u}(x,t) u(x,t) dx \right) \psi_q'(t) dt + \int_{\mathbf{R}^d} \boldsymbol{u}(\cdot,0) u(\cdot,0) = \mu(\phi^q) + \nu(\phi^q).$$

The sequence of Radon measures $(\psi_q'(t)dt)_q$ converges to $-\delta_T$, so the first term of the left-hand side converges to $-\int_{\mathbf{R}^d} \boldsymbol{u}(\cdot,T)u(\cdot,T)$. Clearly, for every $K_n \in \mathcal{M}$ and $L \in \partial K_n^-$, we have $(\langle\phi_q\rangle_K)_q \to \langle\phi\rangle_{K_n}$, $(\langle V\cdot\mathbf{n}\phi_q\rangle_{K|L_n})_q \to \langle V\cdot\mathbf{n}\phi\rangle_{K|L_n}$ and $(\langle\phi_q\rangle_K(n\delta t))_q \to \langle\phi\rangle_K(n\delta t)$. Thus, passing to the limit on q, we get

$$-\int_{\mathbf{R}^d} \boldsymbol{u}(\cdot,T)u(\cdot,T) + \int_{\mathbf{R}^d} \boldsymbol{u}(\cdot,0)u(\cdot,0) = \mu(\psi u) + \nu(\psi u).$$

Similarly, we obtain from (1.5) (or more directly from the conservation of λ by the flow) that

$$-\int_{\mathbf{R}^d} u(\cdot,T)^2 + \int_{\mathbf{R}^d} u(\cdot,0)^2 = 0.$$

Subtracting these equalities, we get

$$\int_{\mathbf{R}^d} ((u-\boldsymbol{u})u)(\cdot,T) - \int_{\mathbf{R}^d} ((u-\boldsymbol{u})u)(\cdot,0) = \mu(\psi u) + \nu(\psi u).$$

Using the identity $(a-b)a = [(a-b)^2]/2 + [a^2-b^2]/2$, we obtain

$$\begin{aligned}&\|(u-\boldsymbol{u})(\cdot,T)\|_{L^2}^2 + \|\boldsymbol{u}(\cdot,0)\|_{L^2}^2 - \|\boldsymbol{u}(\cdot,T)\|_{L^2}^2\\ &\quad= 2\mu(\psi u) + 2\nu(\psi u) + \|(u-\boldsymbol{u})(\cdot,0)\|_{L^2}^2 + \left(\|u(\cdot,0)\|_{L^2}^2 - \|u(\cdot,T)\|_{L^2}^2\right).\end{aligned}$$

The last term vanishes and Lemma 3.3 implies $\|(u-\boldsymbol{u})(\cdot,0)\|_{L^2}^2 \le C\|\nabla u_0\|_{L^2}^2 h^2$. Thus we have

$$\begin{aligned}&\|(u-\boldsymbol{u})(\cdot,T)\|_{L^2}^2 + \|\boldsymbol{u}(\cdot,0)\|_{L^2}^2 - \|\boldsymbol{u}(\cdot,T)\|_{L^2}^2\\ &\qquad\le 2\mu(\psi u) + 2\nu(\psi u) + C\|\nabla u_0\|_{L^2}^2 h^2.\end{aligned} \tag{3.13}$$

(a) We first estimate the term $\mu(\psi u)$ in (3.13). From the definition (3.2) of μ and the Cauchy–Schwarz inequality, we get

$$|\mu(\psi u)| \le \mathcal{F}(\boldsymbol{u},T)^{1/2}\left(\sum_{K_n\in\mathcal{M}_T} |K|\,|\langle u\rangle_{K_n} - \langle u\rangle_K((n+1)\delta t)|^2\right)^{1/2}.$$

By Jensen's inequality, we have

$$|\langle u\rangle_{K_n} - \langle u\rangle_K((n+1)\delta t)|^2 \le \langle |u-u((n+1)\delta t,\cdot)|^2\rangle_{K_n},$$

and by the first estimate of Corollary 1.2 we get

$$|\mu(\psi u)| \le C_0\mathcal{F}(\boldsymbol{u},T)^{1/2}\|\nabla u_0\|_{L^2}\,T^{1/2}\delta t^{1/2}. \tag{3.14}$$

(b) We now turn to the term $\nu(\psi u)$ in (3.13). Let us write

$$\begin{aligned}\nu(\psi u) &= \sum_{K_n\in\mathcal{M}_T}\sum_{L\in\partial K_n^-} \delta t(u_L^n - u_K^n)(V_{KL}^n\langle u\rangle_{K_n} - |K|L|\langle V\cdot\mathbf{n}\,u\rangle_{K|L_n})\\ &= \mathrm{I} + \mathrm{II},\end{aligned} \tag{3.15}$$

with

$$\mathrm{I} := \sum_{K_n\in\mathcal{M}_T}\sum_{L\in\partial K_n^-} \delta t(u_L^n - u_K^n)V_{KL}^n \left(\langle u\rangle_{K_n} - \fint_{K|L_n} u\right),$$

$$\mathrm{II} := \sum_{K_n\in\mathcal{M}_T}\sum_{L\in\partial K_n^-} \delta t(u_L^n - u_K^n)|K|L|\fint_{K|L_n}\left(-u + \fint_{K|L_n} u\right) V\cdot\mathbf{n}_{KL}.$$

From the Cauchy–Schwarz inequality, we deduce

$$|\mathrm{I}| \le \mathcal{F}(\boldsymbol{u},T)^{1/2}\left(\sum_{K_n\in\mathcal{M}_T}\sum_{L\in\partial K_n^-}\delta t|V_{KL}^n|\left(\langle u\rangle_{K_n} - \fint_{K|L_n} u\right)^2\right)^{1/2},$$

and since $|V_{KL}^n| \le \|V\|_\infty |K|L|$, (3.7) yields

$$|\mathrm{I}| \le CC_0\mathcal{F}(\boldsymbol{u},T)^{1/2}\|\nabla u_0\|_{L^2}T^{1/2}h^{1/2}. \tag{3.16}$$

We now estimate II. We split the integral in space and time differences,

$$\begin{aligned}\fint_{K|L_n}\left(-u + \fint_{K|L_n} u\right) V\cdot\mathbf{n}_{KL} &= \fint_{K|L_n}\fint_{K|L_n}(u(y,t)-u(x,s))V(x,s)\cdot\mathbf{n}_{KL}\,dxdrdyds\\ &=: A + B,\end{aligned} \tag{3.17}$$

with

$$A := \fint_{K|L_n}\fint_{n\delta t}^{(n+1)\delta t}(u(y,s)-u(x,s))V(x,s)\cdot\mathbf{n}_{KL},$$

$$B := \fint_{K|L_n}\fint_{K|L_n}(u(y,t)-u(y,s))V(x,s)\cdot\mathbf{n}_{KL}.$$

Since $\fint_{n\delta t}^{(n+1)\delta t}\fint_{n\delta t}^{(n+1)\delta t}(u(y,t)-u(y,s))drds = 0$, we have

$$\begin{aligned}|B| = \Bigg|\fint_{K|L_n}\fint_{K|L_n}&(u(y,s)-u(y,t))\\ &\times\left(V(x,s)\cdot\mathbf{n}_{KL} - \fint_{n\delta t}^{(n+1)\delta t} V(x,\cdot)\cdot\mathbf{n}_{KL}\right)dxdydrds\Bigg|\\ &\le C\delta t\fint_{K|L_n}\fint_{n\delta t}^{(n+1)\delta t}|u(y,s)-u(y,t)|dydrds.\end{aligned} \tag{3.18}$$

Similarly, $|A| \le Ch\fint_{K|L_n}\fint_{K|L}|u(x,s)-u(y,s)|$. Plugging this estimate and (3.18)

into (3.17) and using the Cauchy–Schwarz inequality, we get

$$
\begin{aligned}
|\mathrm{II}|^2 \le Ch & \left(\sum_{K_n \in \mathcal{M}_T,\, L \in \partial K_n^-} \delta t\, h^d |u_L^n - u_K^n|^2 \right) \\
& \times \left(\sum_{K_n \in \mathcal{M}_T,\, L \in \partial K_n^-} \delta t |K|L| \left(\fint_{K|L_n} \fint_{K|L} |u(x,t) - u(y,t)| \right)^2 \right) \\
& + C\delta t \left(\sum_{K_n \in \mathcal{M}_T,\, L \in \partial K_n^-} \delta t^2 |K|L| |u_L^n - u_K^n|^2 \right) \\
& \times \left(\sum_{K_n \in \mathcal{M}_T,\, L \in \partial K_n^-} \delta t |K|L| \left(\fint_{K|L_n} \fint_{n\delta t}^{(n+1)\delta t} |u(x,s) - u(x,t)| \right)^2 \right) \\
=: & \, ChA_1B_1 + C\delta t A_2 B_2.
\end{aligned}
$$

Using the L^2-stability of the scheme and assumption (1.6), we find that A_1 and A_2 are bounded by $CT\|u_0\|_{L^2}^2$. We use (3.9) and (3.8), respectively, to estimate B_1 and B_2 and obtain

$$
|\mathrm{II}| \le CC_0 \|\nabla u_0\|_{L^2} \|u_0\|_{L^2} Th. \tag{3.19}
$$

Remark 3.2. When V is constant, II vanishes, so (3.9) and then (1.12) are not needed.

Plugging (3.16) and (3.19) into (3.15) and then (3.15) and (3.14) into (3.13) (and using $\delta t \le c_0 h$), we get

$$
\begin{aligned}
& \|(u - \boldsymbol{u})(\cdot, T)\|_{L^2}^2 + \|\boldsymbol{u}(\cdot, 0)\|_{L^2}^2 - \|\boldsymbol{u}(\cdot, T)\|_{L^2}^2 \\
& \quad \le \; CC_0 \left(\|u_0\|_{H^1}^2 Th + \|u_0\|_{H^1} \mathcal{F}(\boldsymbol{u}, T)^{1/2} T^{1/2} h^{1/2} \right) + C\|u_0\|_{H^1}^2 h^2.
\end{aligned}
$$

And from Lemma 3.2,

$$
\begin{aligned}
& \|(u - \boldsymbol{u})(\cdot, T)\|_{L^2}^2 + \|\boldsymbol{u}(\cdot, 0)\|_{L^2}^2 - \|\boldsymbol{u}(\cdot, T)\|_{L^2}^2 \\
& \quad \le \; CC_0 \left(\|u_0\|_{H^1}^2 Th + \xi^{-1/2} \|u_0\|_{H^1} \left(\|\boldsymbol{u}(\cdot, 0)\|_{L^2}^2 - \|\boldsymbol{u}(\cdot, T)\|_{L^2}^2 \right)^{1/2} T^{1/2} h^{1/2} \right) \\
& \qquad + C\|u_0\|_{H^1}^2 h^2.
\end{aligned}
$$

We then use the Young inequality $ab \le \frac{1}{2}a^2 + \frac{1}{2}b^2$ with

$$
a = (CC_0)^{-1/2} \left(\|\boldsymbol{u}(\cdot, 0)\|_{L^2}^2 - \|\boldsymbol{u}(\cdot, T)\|_{L^2}^2 \right)^{1/2}, \qquad b = (CC_0)^{1/2} \xi^{-1/2} \|u_0\|_{H^1} T^{1/2} h^{1/2}
$$

to derive

$$
\begin{aligned}
& \|(u - \boldsymbol{u})(\cdot, T)\|_{L^2}^2 + \frac{1}{2} \left(\|\boldsymbol{u}(\cdot, 0)\|_{L^2}^2 - \|\boldsymbol{u}(\cdot, T)\|_{L^2}^2 \right) \\
& \qquad \le C(1 + C_0^2)\, \xi^{-1} \|u_0\|_{H^1}^2 Th + C\|u_0\|_{H^1}^2 h^2,
\end{aligned}
$$

which yields the first estimate of Theorem 1.8 and, by Lemma 3.2, the second estimate. ∎

REFERENCES

[1] D. Bouche, J.-M. Ghidaglia, and F. Pascal, *Error estimate and the geometric corrector for the upwind finite volume method applied to the linear advection equation*, SIAM J. Numer. Anal., 43 (2005), pp. 578–603.

[2] B. Cockburn, F. Coquel, and P. LeFloch, *An error estimate for finite volume methods for multidimensional conservation laws*, Math. Comp., 63 (1994), pp. 77–103.

[3] B. Cockburn, P.-A. Gremaud, and J. X. Yang, *A priori error estimates for numerical methods for scalar conservation laws. Part* III. *Multidimensional flux-splitting monotone schemes on non-Cartesian grids*, SIAM J. Numer. Anal., 35 (1998), pp. 1775–1803.

[4] M. G. Crandall and L. Tartar, *Some relations between nonexpansive and order preserving mappings*, Proc. Amer. Math. Soc., 78 (1980), pp. 385–390.

[5] B. Després, *Convergence of non-linear finite volume schemes for linear transport*, in Notes from the XIth Jacques-Louis Lions Hispano-French School on Numerical Simulation in Physics and Engineering (Spanish), Grupo Anal. Teor. Numer. Modelos Cienc. Exp. Univ. Cádiz, Cádiz, Spain, 2004, pp. 219–239.

[6] B. Després, *Lax theorem and finite volume schemes*, Math. Comp., 73 (2004), pp. 1203–1234.

[7] R. Eymard, T. Gallouët, and R. Herbin, *Finite volume methods*, in Handbook of Numerical Analysis, Vol. VII, Handb. Numer. Anal. VII, North-Holland, Amsterdam, 2000, pp. 713–1020.

[8] R. Eymard, T. Gallouët, and R. Herbin, *Error estimate for approximate solutions of a nonlinear convection-diffusion problem*, Adv. Differential Equations, 7 (2002), pp. 419–440.

[9] R. Herbin, *An error estimate for a finite volume scheme for a diffusion-convection problem on a triangular mesh*, Numer. Methods Partial Differential Equations, 11 (1995), pp. 165–173.

[10] D. Kröner and M. Rokyta, *Higher order finite volume method for convection dominated diffusion equation in* $2D$, in Numerical Modelling in Continuum Mechanics, M. Feistauer, R. Rannacher, K. Kozel, eds., Matfyzpress, Praha, Czech Republic, 1997, pp. 62–70.

[11] N. N. Kuznecov, *The accuracy of certain approximate methods for the computation of weak solutions of a first order quasilinear equation*, Ž. Vyčisl. Mat. i Mat. Fiz., 16 (1976), pp. 1489–1502, 1627.

[12] B. Merlet and J. Vovelle, *Error estimate for a finite volume scheme applied to the linear advection equation*, Numer. Math., to appear.

[13] S. Noelle and M. Westdickenberg, *A new convergence proof for finite volume schemes using the kinetic formulation of conservation laws*, SIAM J. Numer. Anal., 37 (2000), pp. 742–757.

[14] M. Ohlberger and C. Rohde, *Adaptive finite volume approximations for weakly coupled convection dominated parabolic systems*, IMA J. Numer. Anal., 22 (2002), pp. 253–280.

[15] J.-P. Vila, *Convergence and error estimates in finite volume schemes for general multidimensional scalar conservation laws.* I. *Explicit monotone schemes*, RAIRO Modél. Math. Anal. Numér., 28 (1994), pp. 267–295.

[16] J.-P. Vila and P. Villedieu, *Convergence of an explicit finite volume scheme for first order symmetric systems*, Numer. Math., 94 (2003), pp. 573–602.

SIAM J. NUMER. ANAL.
Vol. 46, No. 1, pp. 151–179

© 2007 Society for Industrial and Applied Mathematics

CONVERGENCE OF TIME-AVERAGED STATISTICS OF FINITE ELEMENT APPROXIMATIONS OF THE NAVIER–STOKES EQUATIONS*

V. JOHN†, W. LAYTON‡, AND C. C. MANICA‡

Abstract. When discussing numerical solutions of the Navier–Stokes equations, especially when turbulent flows are concerned, there are at least two questions that can be raised. *What is meaningful to compute? How to determine the fidelity of the computed solution with respect to the true solution?* This paper takes a step towards the answer of these questions for turbulent flows. We consider long-time averages of weak solutions of the Navier–Stokes equations, rather than strong solutions. We present error estimates for the time-averaged energy dissipation rate, drag and lift, most of them under the assumption of small Reynolds/generalized Grashof number. For shear flows, we address the question of fidelity of the computed solution with respect to the true solution, in view of Kolmogorov's energy cascade theory.

Key words. time-averaged statistics, Navier–Stokes equations, turbulence

AMS subject classifications. 74H15, 76D05, 76F65

DOI. 10.1137/060649550

1. Introduction.

The motion of an incompressible Newtonian fluid is governed by the incompressible Navier–Stokes equations. A fundamental problem of fluid motion is turbulence and a fundamental problem in the Navier–Stokes equations is uniqueness of weak solutions in the general case of no assumed extra regularity or small data. The Leray conjecture [20] states that these two problems are connected: the lack of uniqueness of weak solutions (which he called "turbulent solutions") is not an artifact of imperfect mathematical techniques, but it reflects fundamental physical mechanisms of turbulence.

The numerical analysis of turbulent flows is caught between the gaps in the physical understanding of turbulence and those in the mathematical foundations of the Navier–Stokes equations. For example, smooth strong solutions are not expected while, if the uniqueness of the weak solution is unknown, bounding the error in a numerical simulation is currently not possible without assuming extra regularity on the solution, or without assuming both the initial data u_0 and the body force $f(x,t)$ to be very small.

On the other hand, computational simulations are carried out and statistics of computed fluid velocities and pressures often reflect rather accurately statistics of physical flows even in the absence of mathematical justification for this accuracy. Further, statistics (by which we shall mean long-time averages) are often smooth, behaving deterministically (often in accordance with the Kolmogorov theory [18]) in both numerical simulations and physical experiments. From this situation, a challenge for the numerical analysis of fluid motion arises: develop a rigorous understanding

*Received by the editors January 10, 2006; accepted for publication (in revised form) June 6, 2007; published electronically December 20, 2007.
http://www.siam.org/journals/sinum/46-1/64955.html

†Fachbereich 6.1–Mathematik, Universität des Saarlandes, Postfach 15 11 50, 66041 Saarbrücken, Germany (john@math.uni-sb.de, http://www.math.uni-sb.de/ag/john/).

‡Department of Mathematics, University of Pittsburgh, Pittsburgh, PA 15260 (wjl@pitt.edu, http://www.math.pitt.edu/∼wjl; cac15@pitt.edu, http://www.pitt.edu/∼cac15). The research of these authors was partially supported by NSF grants 0207627 and 0508260.

of how statistics computed from numerical simulations reflect those for the unknown solution of the Navier–Stokes equations.

The incompressible Navier–Stokes equations (with homogeneous Dirichlet boundary conditions) are given by

$$\begin{aligned} u_t + u \cdot \nabla u - \nu \Delta u + \nabla p &= f \quad && \text{in } \Omega \times (0, \infty), \\ \nabla \cdot u &= 0 \quad && \text{in } \Omega \times [0, \infty), \\ u(x, 0) &= u_0 \quad && \text{in } \Omega, \\ u &= 0 \quad && \text{on } \Gamma \times [0, \infty). \end{aligned} \tag{1.1}$$

Here, Ω denotes a bounded and regular flow domain in $\mathbb{R}^d$, $d = 2, 3$, $\Gamma = \partial\Omega$ is the boundary of Ω, $u(x,t)$, $p(x,t)$ are the fluid velocity and pressure, ν is the kinematic viscosity, $f(x,t) \in L^\infty(0,\infty; L^2(\Omega)^d)$ are the body forces, and $u_0 \in L^2(\Omega)^d$ is a weakly divergence-free initial condition. The Reynolds number is defined by $Re = LU/\nu$, where the constants L and U are, respectively, a reference large-scale length and velocity. From the point of view of applications, the three-dimensional case is the important one. This case is mathematically much more challenging than the two-dimensional one, too.

We will study statistics related to the energy dissipation rate and the total kinetic energy of the flow. The energy dissipation rate per unit volume of the flow at time t is given by

$$\varepsilon(u) := \frac{\nu}{|\,\Omega\,|} \| \nabla u(\cdot, t) \|^2,$$

where $|\,\Omega\,|$ is the volume of Ω and $\|\cdot\|$ denotes the $L^2(\Omega)$-norm, and its total kinetic energy per unit volume is

$$k(u) := \frac{1}{2|\,\Omega\,|} \| u(\cdot, t) \|^2.$$

The time average $\langle q \rangle$ of a quantity q is defined by

$$\langle q \rangle = \limsup_{T \to \infty} \frac{1}{T} \int_0^T q(t)\, dt. \tag{1.2}$$

Consequently, the time average of the energy dissipation rate is

$$\langle \varepsilon(u) \rangle = \limsup_{T \to \infty} \frac{1}{T} \int_0^T \varepsilon(u)\, dt = \frac{\nu}{|\,\Omega\,|} \limsup_{T \to \infty} \frac{1}{T} \int_0^T \|\nabla u\|^2 = \frac{\nu}{|\,\Omega\,|} \langle \|\nabla u\|^2 \rangle,$$

and the time average of the kinetic energy is

$$\langle k(u) \rangle = \frac{1}{2|\,\Omega\,|} \limsup_{T \to \infty} \frac{1}{T} \int_0^T \| u \|^2\, dt = \frac{1}{2|\,\Omega\,|} \langle \| u \|^2 \rangle.$$

In practical simulations of turbulent flows, it is typical to compute time-averaged flow statistics, which are then matched against benchmark averages; see, e.g., [2, 17, 15, 16, 21, 5]. However, there is very little numerical analysis in support of this approach. Of course, if the error in certain norms of the velocity and the pressure is provably optimal over $0 \le t < \infty$, then time averages involving these norms are

convergent with optimal rates as well. But, the practical case is complementary: time averages seem to be predictable even when the dynamic flow behavior over bounded time intervals is irregular. This is the case we aim to study. However, a complete analysis seems to be still beyond the present mathematical tools.

In section 3.1, the case of large data is studied and estimates for the L^2-norm of the time-averaged error of the pressure (Theorem 3.1) and the time-averaged energy dissipation rate (Theorem 3.2) are given. Then, in section 3.2, we consider the case of arbitrary initial data u_0 and asymptotically small body forces (small generalized Grashof number) which converge to a stationary limit $f^*(x) = \lim_{t\to\infty} f(x,t)$. Let (u^h, p^h) be a finite element approximation of the velocity field and the pressure and assume that $f^*(x)$ satisfies a small data condition. Let u^* be the solution of the stationary Navier–Stokes equations with body force f^*. We show that

$$\langle \varepsilon(u - u^*)\rangle = 0, \quad \langle \varepsilon(u^h - u^{*\,h})\rangle = 0,$$

and we prove (Theorem 3.3) an error estimate which shows that the problem of estimating $\langle \varepsilon(u - u^h)\rangle$ reduces to the one of estimating $\|\nabla(u^* - u^{*\,h})\|^2$. Then the error goes to zero optimally as the mesh width $h \to 0$. This result is plausible because the possible irregularities caused by large initial data are washed out by the time averaging.

Section 4 studies the flow through a channel around a body. Under the assumption of bounded growth of the kinetic energy (in the case of inhomogeneous inflow boundary conditions), we give estimates for the time-averaged drag and lift coefficients on an immersed body (Theorem 4.1). It is shown that the assumption holds true if the inflow boundary conditions are sufficiently small.

In section 5, we consider the complementary situation of a flow driven by a large and persistent boundary condition. We are not (yet) able to perform a complete error analysis in this case. However, following the important work of Constantin and Doering [4] in the continuous case, we show that, provided the first mesh line in the finite element mesh is within $O(1/Re)$ of the moving wall which drives the flow, the computed time-averaged energy dissipation rate for the shear flow scales as predicted for the continuous flow by the Kolmogorov theory (Theorem 5.1):

$$\langle \varepsilon(u^h)\rangle \quad \leq \quad C\,\frac{U^3}{L}.$$

This restriction on the mesh size arises from mathematical analysis of constructible background flows in finite element spaces and their subsequent analysis. However, it is in accordance with an entirely different observation of the thickness of time-averaged turbulent boundary layers [24].

2. Mathematical preliminaries. Throughout the paper, we use the standard notations $L^p(\Omega)$, $W^{k,p}(\Omega)$, $H^k(\Omega) = W^{k,2}(\Omega)$ for the Lebesgue and Sobolev spaces, respectively. The inner product in the space $L^2(\Omega)$, $L^2(\Omega)^d$, and $L^2(\Omega)^{d\times d}$ will be denoted by $(\cdot,\cdot)$ and its norm by $\|\cdot\|$. Norms in Sobolev spaces $H^k, k > 0$, are denoted by $\|\cdot\|_k$. The symbol C stands for generic constants independent of the viscosity ν and the mesh size h.

The velocity at a given time t is sought in the space

$$\mathbb{X} = H_0^1(\Omega)^d = \{v \in L^2(\Omega)^d : \nabla v \in L^2(\Omega)^{d\times d} \quad \text{and} \quad v = 0 \ \text{ on } \ \Gamma\},$$

equipped with the norm $\|v\|_{\mathbb{X}} = \|\nabla v\|$. The norm of the dual space $\mathbb{X}^*$ of $\mathbb{X}$ is denoted by $\|\cdot\|_{-1}$. The pressure at time t is sought in the space

$$\mathbb{Q} = L_0^2(\Omega) = \left\{ q : \, q \in L^2(\Omega) \, , \int_\Omega q \, dx = 0 \right\}.$$

In addition, the space of weakly divergence-free functions is denoted by

$$\mathbb{V} = \{ v \in \mathbb{X} : (\nabla \cdot v, q) = 0 \; \forall \; q \in \mathbb{Q} \}.$$

For Y being a Banach space and for $0 < T < \infty$, the space $L^p(0,T;Y)$ with $1 \leq p < \infty$ (and the usual modification if $p = \infty$) consists of all functions $v : [0,\infty) \to Y$ for which

$$\int_0^T \| \, v(t) \, \|_Y^p \, dt < \infty.$$

Define the trilinear forms on $\mathbb{X} \times \mathbb{X} \times \mathbb{X}$:

$$b(u,v,w) = \int_\Omega u \cdot \nabla v \cdot w \, dx \qquad \text{and} \qquad b_s(u,v,w) = \frac{1}{2} b(u,v,w) - \frac{1}{2} b(u,w,v).$$

Note that the convective form and the skew-symmetric form are equal for $u \in \mathbb{V}$: $b(u,v,w) = b_s(u,v,w)$.

LEMMA 2.1. *Let $\Omega \subset \mathbb{R}^d, d = 2$ or 3. Then there exists a constant $C = C(\Omega) < \infty$ such that*

$$b(u,v,w) \leq C \| \, \nabla u \, \| \, \| \, \nabla v \, \| \, \| \, \nabla w \, \| \qquad \forall \; u,v,w \in \mathbb{X}. \tag{2.1}$$

If $d = 3$, this can be improved to

$$b(u,v,w) \leq C \sqrt{\| \, u \, \| \, \| \, \nabla u \, \|} \, \| \, \nabla v \, \| \, \| \, \nabla w \, \| \qquad \forall \; u,v,w \in \mathbb{X}, \tag{2.2}$$

or, respectively, to

$$b(u,v,w) \leq C \| \, \nabla u \, \| \, \| \, \nabla v \, \| \, \sqrt{\| \, w \, \| \, \| \, \nabla w \, \|} \qquad \forall \; u,v,w \in \mathbb{X}. \tag{2.3}$$

Proof. Hölder's inequality and Sobolev imbedding theorems give

$$|b(u,v,w)| \leq \|u\|_{L^3} \|\nabla v\|_{L^2} \|w\|_{L^6} \leq C \|u\|_{1/2} \|v\|_1 \|w\|_1.$$

Estimate (2.2) is obtained now by applying the interpolation inequality between $L^2(\Omega)$ and $H^1(\Omega)$ (see Adams [1]),

$$\|u\|_{1/2} \leq C \|u\|^{1/2} \|u\|_1^{1/2},$$

and Poincaré's inequality. Similarly, (2.3) follows from

$$b(u,v,w) \leq C \|u\|_1 \|v\|_1 \|w\|_{1/2}.$$

The proof of (2.1) uses, after the application of Hölder's inequality, the imbedding $H^1(\Omega) \subset L^3(\Omega)$. ∎

Estimates for the skew-symmetric form $b_s(\cdot,\cdot,\cdot)$ can be derived directly from (2.1)–(2.3). The norms of the trilinear forms $b_s(\cdot,\cdot,\cdot) : \mathbb{X} \to \mathbb{R}$ and of the restriction $b_s(\cdot,\cdot,\cdot) : \mathbb{V} \to \mathbb{R}$ are denoted by

(2.4)
$$M = \sup_{u,v,w\in\mathbb{X}} \frac{b_s(u,v,w)}{\|\nabla u\|\,\|\nabla v\|\,\|\nabla w\|} < \infty \text{ and } N = \sup_{u,v,w\in\mathbb{V}} \frac{b_s(u,v,w)}{\|\nabla u\|\,\|\nabla v\|\,\|\nabla w\|} < \infty.$$

We will consider in this paper finite element approximations of (1.1). Throughout the paper, we shall assume that the velocity-pressure finite element spaces $\mathbb{X}^h \subset \mathbb{X}$ and $\mathbb{Q}^h \subset \mathbb{Q}$ are conforming, have approximation properties typical of finite element spaces commonly in use, and satisfy the discrete inf-sup condition,

$$\inf_{q^h\in\mathbb{Q}^h} \sup_{v^h\in\mathbb{X}^h} \frac{(q^h,\, \nabla\cdot v^h)}{\|\nabla v^h\|\,\|q^h\|} \geq \beta^h > 0, \tag{2.5}$$

where β^h is bounded away from zero uniformly in h. The space of discretely divergence-free functions is defined by

$$\mathbb{V}^h = \{v^h \in \mathbb{X}^h : (q^h,\, \nabla\cdot v^h) = 0 \quad \forall\, q^h \in \mathbb{Q}^h\}.$$

For examples of such spaces see, e.g., Gunzburger [13], Brezzi and Fortin [3], and Girault and Raviart [11].

Norms of the trilinear form $b_s(\cdot,\cdot,\cdot)$ restricted to $\mathbb{X}^h$ and $\mathbb{V}^h$, respectively, are defined in the same way as in (2.4), leading to corresponding constants M^h and N^h. Note that $M \geq M^h, N^h, N$ and that $N^h \to N$ as $h \to 0$ (see [11]) and by the same argument $M^h \to M$.

2.1. The continuous-in-time finite element discretization. Consider the standard finite element discretization of the Navier–Stokes equations. The semidiscrete (continuous-in-time) finite element approximations $u^h = u^h(\cdot,t)$ and $p^h = p^h(\cdot,t)$ are the maps $u^h : [\,0,\infty\,) \to \mathbb{X}^h$, $p^h : (\,0,\infty\,) \to \mathbb{Q}^h$ satisfying

$$(u_t^h, v^h) + \nu(\nabla u^h, \nabla v^h) + b_s(u^h, u^h, v^h) - (p^h, \nabla\cdot v^h) = (f, v^h) \quad \forall\, v^h \in \mathbb{X}^h, \tag{2.6}$$

$$(\nabla\cdot u^h, q^h) = 0 \quad \forall\, q^h \in \mathbb{Q}^h, \tag{2.7}$$

$$(u^h(\cdot,0) - u_0, v^h) = 0 \quad \forall\, v^h \in \mathbb{X}^h. \tag{2.8}$$

Under the inf-sup condition (2.5), this is equivalent to: find $u^h : [\,0,\infty\,) \to \mathbb{V}^h$ satisfying

$$(u_t^h, v^h) + \nu(\nabla u^h, \nabla v^h) + b_s(u^h, u^h, v^h) = (f, v^h) \quad \forall\, v^h \in \mathbb{V}^h, \tag{2.9}$$

$$(u^h(\cdot,0) - u_0, v^h) = 0 \quad \forall\, v^h \in \mathbb{V}^h.$$

2.2. The associated equilibrium problem. In the case that $f(x,t) \to f^*(x)$ in $H^{-1}(\Omega)$ as $t \to \infty$, we can associate with (1.1) the following equilibrium problem: find $u^*(x)$, $p^*(x)$ satisfying

$$\begin{aligned} -\nu\Delta u^* + u^*\cdot\nabla u^* + \nabla p^* &= f^* \quad \text{in } \Omega, \\ \nabla\cdot u^* &= 0 \quad \text{in } \Omega, \\ u^* &= 0 \quad \text{on } \Gamma, \\ \int_\Omega p^*\, dx &= 0. \end{aligned} \tag{2.10}$$

The variational formulation of the equilibrium problem is as follows: Find $u^* \in \mathbb{X}$ and $p^* \in \mathbb{Q}$ such that

$$\nu(\nabla u^*, \nabla v) + b_s(u^*, u^*, v) - (p^*, \nabla \cdot v) = (f^*, v) \quad \forall\, v \in \mathbb{X}, \tag{2.11}$$
$$(\nabla \cdot u^*, q) = 0 \quad \forall\, q \in \mathbb{Q},$$

or, equivalently, find $u^* \in \mathbb{V}$ such that

$$\nu(\nabla u^*, \nabla v) + b_s(u^*, u^*, v) = (f^*, v) \qquad \forall\, v \in \mathbb{V}. \tag{2.12}$$

A finite element approximation $(u^{*\,h}, p^{*\,h})$ is given by the solution of

$$\nu(\nabla u^{*\,h}, \nabla v^h) + b_s(u^{*\,h}, u^{*\,h}, v^h) - (p^{*\,h}, \nabla \cdot v^h) = (f^*, v^h) \quad \forall\, v^h \in \mathbb{X}^h, \tag{2.13}$$
$$(\nabla \cdot u^{*\,h}, q^h) = 0 \quad \forall\, q^h \in \mathbb{Q}^h.$$

In $\mathbb{V}^h$ this becomes: find $u^{*\,h} \in \mathbb{V}^h$ such that

$$\nu(\nabla u^{*\,h}, \nabla v^h) + b_s(u^{*\,h}, u^{*\,h}, v^h) = (f^*, v^h) \qquad \forall\, v^h \in \mathbb{V}^h. \tag{2.14}$$

It is known that solutions of the equilibrium problem are nonsingular for small data, generically nonsingular for large data, and optimally approximated by $u^{*\,h}$ when nonsingular; see [12]. Setting $v = u^*$ in (2.11) and $v^h = u^{*h}$ in (2.13), it is easy to check the a priori bounds

$$\| \nabla u^* \| \le \nu^{-1} \| f^* \|_{-1}, \qquad \| \nabla u^{*\,h} \| \le \nu^{-1} \| f^* \|_{-1}. \tag{2.15}$$

Both bounds can be sharpened slightly by replacing $\| f^* \|_{-1}$ with the dual norms of $\mathbb{V}$ or $\mathbb{V}^h$, respectively.

It is known that if the problem data is small enough, concretely if

$$N\nu^{-2} \| f^* \|_{-1} < 1,$$

then u^* is unique. If additionally $f(x,t) \equiv f^*(x)$, then $u(x,t) \to u^*(x)$ in $L^2(\Omega)^d$ exponentially fast as $t \to \infty$ and (u^h, p^h) approximates (u, p) optimally; see [11, 19, 13].

2.3. Weak solutions of the Navier–Stokes equations. The problem of turbulence is perhaps intimately connected with questions about weak solutions versus strong solutions of the Navier–Stokes equations. It is well known that weak solutions exist but it is not known if they are unique. (Thus, different methods of proving existence might possibly lead to different solutions.) A strong solution is generally defined as a weak solution which has enough extra regularity to ensure global uniqueness, i.e., which fulfills Serrin's condition [25]. In $\mathbb{R}^3$, it is unknown if strong solutions exist for all time; see [10, 26]. But if a strong solution exists, it is unique. Strong solutions might conceivably describe all fluid motion. However, in at least one conjecture about turbulence the case of strong solutions is associated with laminar flow.

For clearness of notation, we will give the definition of a weak and a strong solution, following [10].

DEFINITION 2.1. *Let*

1. $\mathbf{D}_T = \left\{ v \in C^\infty(\Omega \times [0,T])^d \ : \ v(t) \in C_0^\infty(\Omega)^d \text{ for each } t \right\}$,
2. $\mathcal{D} = \left\{ \psi \in C_0^\infty(\Omega)^d \ : \nabla \cdot \psi = 0 \text{ in } \Omega \right\}$,
3. $H(\Omega) = \left\{ v \in L^2(\Omega)^d : \nabla \cdot v = 0 \text{ and } v \cdot \hat{n} = 0 \text{ on } \Gamma \right\}$, *where* $\hat{n}$ *is the unit outer normal on* Γ,

4. $\mathcal{D}_T = \{\ \phi(x,t) \in C^\infty(\Omega \times [0,T])^d : \phi(x,t) \in \mathcal{D}$ *for each* t, $0 \le t \le T\}$.

Let $u_0 \in H(\Omega)$, $f \in L^2(0,T;L^2(\Omega)^d)$. *A measurable function* $u(x,t) : \Omega \times [0,T] \to \mathbb{R}^d$ *is a weak solution of the Navier–Stokes equations if for all* $T > 0$

1. $u \in L^2(0,T;\mathbb{V}) \cap L^\infty(0,T;H(\Omega))$,
2. u *satisfies the integral relation*

$$(u(T),\phi(T)) - \int_0^T \left[\left(u, \frac{\partial\,\phi}{\partial\,t}\right) - \nu\,(\nabla\,u, \nabla\,\phi) - b(u,u,\phi)\right]\,dt = (u(0),\phi(0)) + \int_0^T (f,\phi)\,dt$$

for all $\phi \in \mathcal{D}_T$, *which is equivalent to*

$$\frac{d}{dt}(u,v) + \nu(\nabla u, \nabla v) + b(u,u,v) - (f,v) = 0$$

for all $v \in \mathbb{V}$,

3. u *is a strong solution if* u *is a weak solution and* $u \in L^\infty(0,T;\mathbb{V})$ *for any* $T > 0$.

We note that if Ω is a bounded domain with Γ satisfying a cone condition, then it is known that, given a weak solution u, there exists a pressure $p(x,t) \in L^\infty(0,T;L^2_0(\Omega))$ (see, e.g., [10, Remark 2.5]) satisfying

$$(u(T),\phi(T)) - \int_0^T \left[\left(u, \frac{\partial\phi}{\partial t}\right) - \nu\,(\nabla u, \nabla\phi) - b(u,u,\phi) + (p, \nabla\cdot\phi)\right]\,dt$$

$$= (u(0),\phi(0)) + \int_0^T (f,\phi)\,dt \qquad \forall\ \phi \in \mathbf{D}_T. \tag{2.16}$$

This is equivalent to

$$\frac{d}{dt}(u,v) + \nu\,(\nabla u, \nabla v) + b(u,u,v) - (p, \nabla\cdot v) = (f,v) \qquad \forall\ v \in \mathbb{X}. \tag{2.17}$$

It is well known (see [10]) that weak solutions satisfy the energy inequality: for any $t \in [0,T]$,

$$\frac{1}{2}\|\,u(T)\,\|^2 + \nu\int_0^T \|\,\nabla\,u(t)\,\|^2\,dt \le \frac{1}{2}\|\,u_0\,\|^2 + \int_0^T (u(t), f(t))\,dt. \tag{2.18}$$

Strong solutions satisfy even an energy equality, i.e., (2.18) with "$\le$" replaced by "=".

LEMMA 2.2. *Let* (u,p) *be a weak solution of the Navier–Stokes equations, let* (u^h,p^h) *be its finite element approximation defined by* (2.6)–(2.8), *and let* $e = u - u^h$. *Then, for any* C^1 *map* $v^h : [0,T] \to \mathbb{X}^h$, $q^h : (0,T] \to \mathbb{Q}^h$ *(for each* T, $0 < T < \infty$*),*

$$(e(T), v^h(T)) - \int_0^T \left[\left(e, \frac{\partial v^h}{\partial t}\right) - \nu(\nabla e, \nabla v^h) - b_s(u,u,v^h) + b_s(u^h,u^h,v^h)\right.$$

$$\left. + (p - p^h, \nabla\cdot v^h) + (\nabla\cdot e, q^h)\right]\,dt = (e(0), v^h(0)), \tag{2.19}$$

which is equivalent to

$$\frac{d}{dt}(e, v^h) - \left(e, \frac{\partial v^h}{\partial t}\right) + \nu\,(\nabla e, \nabla v^h) + b_s(u, u, v^h) \tag{2.20}$$

$$-b_s(u^h, u^h, v^h) - (p - p^h, \nabla \cdot v^h) = 0,$$

$$(\nabla \cdot e, q^h) = 0.$$

Proof. We shall prove (2.20). The connection between (2.20) and (2.19) is the same as that between (2.16) and (2.17). The second equation in (2.20) follows immediately from $u \in \mathbb{V}$ and (2.8).

First, note that both statements of the lemma follow by subtraction if the term with the time derivative of the test function is well defined, i.e., provided (2.16) can be shown to hold for $\phi \in C^1(0,T;\mathbb{X}^h)$ or (2.17) can be shown for $v \in C^1(0,T;\mathbb{X}^h)$ (since $\mathbb{X}^h \subset \mathbb{X}$). We show the latter.

Since $\mathbb{X}^h \subset \mathbb{X}$, (2.17) holds for all $v := \tilde{v}^h(x) \in \mathbb{X}^h$. Next, let $A(t)$ be a $C^1(0,T)$ function. Multiplication of (2.17) by $A(t)$ and using

$$A(t)\frac{d}{dt}(u, \tilde{v}^h) = \frac{d}{dt}(u, A(t)\tilde{v}^h) - (u, A'(t)\tilde{v}^h)$$

give that u and p satisfy

$$\frac{d}{dt}(u, \tilde{v}^h) - \left(u, \frac{\partial \tilde{v}^h}{\partial t}\right) + \nu\,(\nabla u, \nabla \tilde{v}^h) + b_s(u, u, \tilde{v}^h) - (p, \nabla \cdot \tilde{v}^h) = (f, \tilde{v}^h)$$

with $v^h = v^h(x,t) = A(t)\tilde{v}^h(x)$. The same equation can be derived from (2.6) for (u^h, p^h). Subtracting gives (2.20) for any v^h of the form $v^h = v^h(x,t) = A(t)\tilde{v}^h(x)$.

Since (2.20) is linear in v^h, it also follows for any v^h which is a finite linear combination of such functions,

$$v^h(x,t) = \sum_{i=1}^{N} A_i(t)\tilde{v}_i^h(x).$$

Picking $\tilde{v}_i^h(x)$ to be a basis for $\mathbb{X}^h$ completes the proof. ∎

2.4. Preliminaries on the time-averaging operator. Define the temporal mean value by

$$\langle q \rangle_T = \frac{1}{T}\int_0^T q(t)\,dt.$$

By properties of integrals, $|\langle q \rangle_T| \leq \langle |q| \rangle_T$, and similarly for any function $q(t,x)$, where $\|\cdot\|$ is a spacial norm of $q(t,x)$,

$$\|\langle q \rangle_T\| \leq \langle \|q\| \rangle_T. \tag{2.21}$$

Let $u \in L^p(\Omega)$, $v \in L^q(\Omega)$ with $p^{-1} + q^{-1} = 1$, $p, q \in [1, \infty]$. Using Hölder's

inequality for Lebesgue spaces, one can show a Hölder inequality of the form

$$|\langle (u,v)\rangle_T| \le \frac{1}{T}\int_0^T |(u,v)|\,dt \le \frac{1}{T}\int_0^T \|u\|_{L^p}\|v\|_{L^q}\,dt$$

$$\le \left(\frac{1}{T}\int_0^T \|u\|_{L^p}^p\,dt\right)^{1/p}\left(\frac{1}{T}\int_0^T \|v\|_{L^q}^q\,dt\right)^{1/q}$$

$$= \langle \|u\|_{L^p}^p\rangle_T^{1/p}\langle \|v\|_{L^q}^q\rangle_T^{1/q}. \tag{2.22}$$

With the same arguments, one obtains, for $u \in H_0^1(\Omega)$ and $v \in H^{-1}(\Omega)$,

$$|\langle (u,v)\rangle_T| \;\le\; \langle \|\nabla u\|_{L^2}^2\rangle_T^{1/2}\langle \|v\|_{H^{-1}}^2\rangle_T^{1/2}. \tag{2.23}$$

The time-averaging operator $\langle\cdot\rangle$ defined in (1.2) is just $\langle\cdot\rangle = \limsup_{T\to\infty}\langle\cdot\rangle_T$. One of the subtleties of this definition is that the limit superior is not additive. For example, let $\{a_n\}$ and $\{b_n\}$ be sequences in $\mathbb{R}$. Then

$$\limsup_{n\to\infty}(a_n + b_n) \le \limsup_{n\to\infty} a_n + \limsup_{n\to\infty} b_n, \tag{2.24}$$

$$|\limsup_{n\to\infty} a_n| \le \limsup_{n\to\infty} |a_n|, \tag{2.25}$$

$$\limsup_{n\to\infty}(a_n\, b_n) \le \limsup_{n\to\infty} a_n \limsup_{n\to\infty} b_n \quad \text{if } a_n,\, b_n \ge 0\,\forall\, n, \tag{2.26}$$

$$\limsup_{n\to\infty}(c\, a_n) = c \limsup_{n\to\infty} a_n \quad \text{for } c \ge 0, \tag{2.27}$$

$$\limsup_{n\to\infty} a_n \le \limsup_{n\to\infty} b_n \quad \text{if } a_n \le b_n\,\forall\, n, \tag{2.28}$$

$$\limsup_{n\to\infty}(a_n)^p = (\limsup_{n\to\infty} a_n)^p \quad \text{if } a_n \ge 0\,\forall\, n \text{ and } p > 0. \tag{2.29}$$

Note that equality holds in (2.24) if one of the limits on the right-hand side exists.

From these properties, it follows that

$$|\langle q\rangle| \le \langle |\,q\,|\rangle, \tag{2.30}$$

$$\|\langle q\rangle\| \le \langle \|\,q\,\|\rangle, \tag{2.31}$$

$$|\langle (u,v)\rangle| \le \langle \|u\|_{L^p}^p\rangle^{1/p}\langle \|v\|_{L^q}^q\rangle^{1/q}, \tag{2.32}$$

$$|\langle (u,v)\rangle| \le \langle \|\nabla u\|^2\rangle^{1/2}\langle \|v\|_{-1}^2\rangle^{1/2} \tag{2.33}$$

if the right-hand sides of these inequalities are well defined.

3. Analysis of the time-averaged velocity, pressure, and energy dissipation. First, in section 3.1, the case of arbitrary data fulfilling the conditions of Definition 2.1 is considered. In this case, estimates for the L^2-norm of the time-averaged error of the pressure, $\limsup_{T\to\infty}\|\langle p - p^h\rangle_T\|$ (Theorem 3.1), and the time-averaged energy dissipation rate, $\langle\varepsilon(u-u^h)\rangle$ (Theorem 3.2), are proven. Section 3.2 considers the case of a small body force (but large initial data) such that the solution of the Navier–Stokes equations tends to the solution of the steady state Navier–Stokes equations. It will be shown that in this case the time average of the energy dissipation rate of the error $\langle\varepsilon(u-u^h)\rangle$ will be bounded by the energy dissipation rate of the error of the steady state problem (Theorem 3.3).

3.1. The case of large data.

LEMMA 3.1. *Let u be a weak solution to the Navier–Stokes equations (obtained by the Leray–Hopf construction). If $f \in L^\infty(0,\infty;H^{-1}(\Omega)^d)$, then $\| u \|$ is uniformly bounded,*

$$\frac{1}{2}\| u(T) \|^2 \leq e^{-\nu C_{PF}^{-2} T}\| u(0) \|^2 + \frac{C_{PF}^2}{\nu^2}\| f \|^2_{L^\infty(0,\infty;H^{-1})}, \tag{3.1}$$

where C_{PF} is the Poincaré–Friedrich constant of Ω, and consequently

$$\lim_{T\to\infty} \frac{1}{T}\| u(T) \|^2 = 0.$$

Proof. Let V_N be the span of eigenfunctions of the Stokes operator. The Leray–Hopf construction of weak solutions gives a sequence $\{u_N\}$ in $\mathbb{V}_N$ satisfying

$$(u_{N,t}, v) + \nu(\nabla u_N, \nabla v) + b(u_N, u_N, v) = (f, v) \quad \forall\, v \in \mathbb{V}_N, \tag{3.2}$$

with a subsequence $\{u_{N_j}\}$ such that u_{N_j} converges to a weak solution u strongly in $L^2(0,T;H(\Omega)^d)$ and weakly in $L^2(0,T;\mathbb{V})$.

Setting $v = u_N$ in (3.2) and using the Cauchy–Schwarz and the Young inequalities, followed by the Poincaré–Friedrich inequality, we have

$$\frac{d}{dt}\| u_N(t) \|^2 + \nu C_{PF}^{-2}\| u_N(t) \|^2 \leq \frac{1}{\nu}\| f(t) \|^2_{-1}.$$

Using the integrating factor $e^{\nu C_{PF}^{-2} t}$, we obtain a differential inequality which can be integrated on $(0,T)$, yielding

$$\| u_N(T) \|^2 \leq e^{-\nu C_{PF}^{-2} T}\| u_N(0) \|^2 + \frac{C_{PF}^2}{\nu^2}\| f \|^2_{L^\infty(0,\infty;H^{-1})}.$$

This shows the uniform boundedness of $\| u_N(T) \|$. Taking the limit of both sides, using a weak convergence argument (which is standard for the Navier–Stokes equations and which we show in detail in the proof of Lemma 3.4) and letting $N \to \infty$, we recover (3.1). The second claim now follows from the first one. ∎

LEMMA 3.2. *Let $f \in L^\infty(0,\infty;H^{-1}(\Omega)^d)$ and let u^h satisfy* (2.9). *Then $\| u^h \|$ is uniformly bounded and consequently*

$$\lim_{T\to\infty} \frac{1}{T}\| u^h(T) \|^2 = 0 \qquad \textit{and} \qquad \lim_{T\to\infty} \frac{1}{T}\| (u - u^h)(T) \|^2 = 0.$$

Proof. Take $v^h = u^h$ in (2.9) (a step not possible in the continuous case of Lemma 3.1) and proceeding as in the proof of Lemma 3.1, we prove a similar uniform bound for $\| u^h \|$. This proves the first statement and the bound on $\| u - u^h \|$ follows by the triangle inequality and Lemma 3.1. ∎

We next consider time averages.

LEMMA 3.3. *Let u be a weak solution of the Navier–Stokes equations satisfying the energy inequality* (2.18). *Then*

$$\langle \varepsilon(u) \rangle \;\leq\; \limsup_{T\to\infty} \frac{1}{|\,\Omega\,|\,T} \int_0^T (\,f, u\,)\, dt = \frac{1}{|\,\Omega\,|}\langle (f,u) \rangle. \tag{3.3}$$

If u satisfies the energy equality, then the above inequality can be replaced by equality. Further, if $f \in L^\infty(0,T;H^{-1}(\Omega)^d) \cap L^2(0,T;L^2(\Omega)^d)$ for every $0 < T < \infty$, then

$$\langle \varepsilon(u) \rangle \;\leq\; \frac{1}{\nu |\,\Omega\,|} \langle \| f \|_{-1}^2 \rangle \;\leq\; \frac{1}{\nu |\,\Omega\,|} \| f \|_{L^\infty(0,\infty;H^{-1})}^2. \tag{3.4}$$

The semidiscrete finite element approximation u^h of u satisfies (3.3) *and* (3.4), *with u replaced by u^h, and with equality in* (3.3).

Proof. Since u satisfies the energy inequality (2.18), we have

$$\frac{1}{2T} \frac{1}{|\,\Omega\,|} \| u(T) \|^2 + \frac{1}{T} \int_0^T \frac{\nu}{|\,\Omega\,|} \| \nabla u(t) \|^2 \, dt \leq \frac{1}{2T} \frac{1}{|\,\Omega\,|} \| u_0 \|^2 + \frac{1}{T|\,\Omega\,|} \int_0^T (f,u) \, dt.$$

Since $\frac{1}{2T} \| u(T) \|^2 \to 0$ by Lemma 3.1 and $\frac{1}{2T} \| u_0 \|^2 \to 0$ as $T \to \infty$, we obtain (3.3). If we use as a starting point the energy equality, the equal sign will be preserved.

For proving (3.4), we apply inequality (2.33) and Young's inequality to (3.3)

$$\begin{aligned} \langle \varepsilon(u) \rangle &\leq \frac{\nu}{2|\,\Omega\,|} \langle \| \nabla u \|^2 \rangle + \frac{1}{2\nu|\,\Omega\,|} \langle \| f \|_{-1}^2 \rangle \\ &\leq \frac{1}{2} \langle \varepsilon(u) \rangle + \frac{1}{2\nu|\,\Omega\,|} \| f \|_{L^\infty(0,\infty;H^{-1})}^2. \end{aligned}$$

In the semidiscrete case, take u^h as a test function in (2.6). This gives

$$\frac{1}{2} \frac{d}{dt} \| u^h(t) \|^2 + \nu \| \nabla u^h(t) \|^2 = (u^h(t), f(t)).$$

Integration in $(0,T)$ shows that u^h fulfills an energy equality. Now, the arguments to derive the estimates of the form (3.3) and (3.4) for u^h are the same as in the continuous case. ∎

Remark 3.1.

1. Let $f \in L^\infty(0,T;H^{-1}(\Omega)^d) \cap L^2(0,T;L^2(\Omega)^d)$ for every $0 < T < \infty$; then one obtains from the triangle inequality that $\langle \varepsilon(u - u^h) \rangle$ is bounded by the data of the problem

$$|\langle \varepsilon(u - u^h) \rangle| \leq \frac{2}{\nu |\,\Omega\,|} \| f \|_{L^\infty(0,\infty;H^{-1})}^2.$$

2. From the Poincaré–Friedrichs inequality it follows that

$$\langle k(v) \rangle \leq \frac{C_{PF}^2}{2\nu} \langle \varepsilon(v) \rangle$$

for all $v \in \mathbb{X}$. Thus, all estimates of Lemma 3.3 as well as the first part of this remark carry over to the time average of the kinetic energy.

Next, we consider the time-averaged errors. It is important to note that there is a difference between $\| \langle \nabla(u - u^h) \rangle \|$ and $\langle \| \nabla(u - u^h) \| \rangle$; the second term is an upper bound for the first one, by (2.31). Experience with turbulent flows suggests that $\langle \nabla u \rangle$ might be smooth (and thus approximable). Thus, ideally we would like estimates for the first term, $\| \langle \nabla(u - u^h) \rangle \|$. In the case of the error in the pressure, we are able to prove such a bound.

THEOREM 3.1. *Let* $f \in L^\infty(0,\infty;H^{-1}(\Omega)^d) \cap L^2(0,T;L^2(\Omega)^d)$ *and let* $(\mathbb{X}^h,\mathbb{Q}^h)$ *satisfy the discrete inf-sup condition* (2.5); *then*

$$\limsup_{T\to\infty} \|\langle p-p^h\rangle_T\| \le \frac{\nu}{\beta^h}\left(1+\frac{2M}{\nu^2}\|f\|_{L^\infty(0,\infty;H^{-1})}\right)\langle\|\nabla(u-u^h)\|^2\rangle^{1/2} + \left(1+\frac{\sqrt{d}}{\beta^h}\right)\inf_{q^h\in\mathbb{Q}^h}\limsup_{T\to\infty}\|\langle p-q^h\rangle_T\|.$$

Proof. A straightforward calculation, using (2.7), shows that (2.19) is equivalent to

$$-\int_0^T (p^h-q^h,\nabla\cdot v^h)\,dt = (e(T),v^h(T)) - (e(0),v^h(0))$$

$$-\int_0^T \left[\left(e,\frac{\partial v^h}{\partial t}\right) - \nu(\nabla e,\nabla v^h) - b_s(u,e,v^h) - b_s(e,u^h,v^h) + (p-q^h,\nabla\cdot v^h)\right]dt$$

for all $(v^h,q^h)\in\mathbb{X}^h\times\mathbb{Q}^h$. Since the velocity finite element functions are continuous in $\overline{\Omega}$, all terms are well defined. Let $v^h = v^h(x)$. Division by T gives

$$\begin{aligned}(\langle q^h-p^h\rangle_T,\nabla\cdot v^h) = {}& \frac{1}{T}(e(T),v^h) - \frac{1}{T}(e(0),v^h) + \nu\,(\langle\nabla e\rangle_T,\nabla v^h)\\ &+\langle b_s(u,e,v^h)\rangle_T + \langle b_s(e,u^h,v^h)\rangle_T\\ &-(\langle p-q^h\rangle_T,\nabla\cdot v^h).\end{aligned} \tag{3.5}$$

For estimating (3.5), we use that v^h does not depend on time, (2.4), and $\|\nabla\cdot v^h\| \le \sqrt{d}\,\|\nabla v^h\|$ to obtain

$$\begin{aligned}\left|(\langle q^h-p^h\rangle_T,\nabla\cdot v^h)\right| \le {}& \frac{C_{PF}}{T}\|e(T)\|\,\|\nabla v^h\| + \frac{C_{PF}}{T}\|e(0)\|\,\|\nabla v^h\|\\ &+\nu\,\|\langle\nabla e\rangle_T\|\,\|\nabla v^h\| + M\langle\|\nabla u\|\,\|\nabla e\|\rangle_T\|\nabla v^h\|\\ &+M\langle\|\nabla e\|\,\|\nabla u^h\|\rangle_T\|\nabla v^h\|\\ &+\sqrt{d}\,\|\langle p-q^h\rangle_T\|\,\|\nabla v^h\|.\end{aligned}$$

Dividing by $\|\nabla v^h\|$ and applying the discrete inf-sup condition (2.5) and (2.21) to the left-hand side of this inequality leads to

$$\begin{aligned}\beta^h\|\langle q^h-p^h\rangle_T\| \le {}& \frac{C_{PF}}{T}\|e(T)\| + \frac{C_{PF}}{T}\|e(0)\| + \nu\,\|\langle\nabla e\rangle_T\|\\ &+M\langle\|\nabla u\|\,\|\nabla e\|\rangle_T + M\langle\|\nabla e\|\,\|\nabla u^h\|\rangle_T\\ &+\sqrt{d}\,\|\langle p-q^h\rangle_T\|.\end{aligned}$$

The terms on the right-hand side are estimated by (2.21) and an estimate similar to (2.22), resulting in

$$\begin{aligned}\beta^h\|\langle q^h-p^h\rangle_T\| \le {}& \frac{C_{PF}}{T}\|e(T)\| + \frac{C_{PF}}{T}\|e(0)\|\\ &+(\nu + M\langle\|\nabla u\|^2\rangle_T^{1/2} + M\langle\|\nabla u^h\|^2\rangle_T^{1/2})\langle\|\nabla e\|^2\rangle_T^{1/2}\\ &+\sqrt{d}\,\|\langle p-q^h\rangle_T\|.\end{aligned}$$

The triangle inequality implies that

$$\|\langle p - p^h\rangle_T\| \le \frac{C_{PF}}{\beta^h T}\|\, e(T)\,\| + \frac{C_{PF}}{\beta^h T}\|\, e(0)\,\|$$

$$+\frac{1}{\beta^h}\left(\nu + M\langle\|\,\nabla u\,\|^2\rangle_T^{1/2} + M\langle\|\,\nabla u^h\,\|^2\rangle_T^{1/2}\right)\langle\|\,\nabla e\,\|^2\rangle_T^{1/2}$$

$$+\left(1+\frac{\sqrt{d}}{\beta^h}\right)\,\|\langle p - q^h\rangle_T\|.$$

Taking lim sup as $T \to \infty$ on both sides of the inequality and using Lemmas 3.1 and 3.2, together with properties of lim sup, give

$$\limsup_{T\to\infty}\|\langle p - p^h\rangle_T\|$$

$$\le \frac{1}{\beta^h}(\nu + M\langle\|\,\nabla u\,\|^2\rangle^{1/2} + M\langle\|\,\nabla u^h\,\|^2\rangle^{1/2})\langle\|\,\nabla e\,\|^2\rangle^{1/2}$$

$$+\left(1+\frac{\sqrt{d}}{\beta^h}\right)\,\limsup_{T\to\infty}\|\langle p - q^h\rangle_T\|.$$

The norms of the weak and the discrete solution can be estimated with the results of Lemma 3.3. The proof concludes by taking the infimum over $q^h \in \mathbb{Q}^h$. ∎

COROLLARY 3.1. *If the assumptions of Theorem* 3.1 *hold true, then*

$$\|\langle p - p^h\rangle\| \le \frac{\nu}{\beta^h}\left(1+\frac{2M}{\nu^2}\|f\|_{L^\infty(0,\infty;H^{-1})}\right)\langle\,\|\,\nabla(u-u^h)\|^2\,\rangle^{1/2}$$

$$+\left(1+\frac{\sqrt{d}}{\beta^h}\right)\,\inf_{q^h\in\mathbb{Q}^h}\,\limsup_{T\to\infty}\|\langle p - q^h\rangle_T\|.$$

Proof. The lower bound on the left-hand side is a consequence of (2.25). ∎

The key idea in the above proofs was to restrict $v^h \in \mathbb{V}^h$ to be time independent. Then, time averaging can be applied and brought inside upon the pressure error directly. It is interesting that the equations of motion give a different realization of the time-averaged error for the velocity and pressure ($\langle\|\,\nabla(u-u^h)\,\|\rangle$ versus $\|\langle p-p^h\rangle\|$). This appears also in the time-averaged lift and drag error estimates in Theorem 4.1. At this point, we do not know if this distinction has other deeper causes or implications.

We next turn to the error inequalities for the time-averaged error of the energy dissipation rate $\langle\varepsilon(u-u^h)\rangle$.

THEOREM 3.2. *Let* $Y = L^2(0,\infty;H^1(\Omega)^d)\cap L^\infty(0,\infty;\mathbb{V}^h)$ *and assume* $u_t \in L^1(0,T;\mathbb{X}^*)$ *for every* $0 < T < \infty$. *Then the time-averaged error of the energy*

dissipation rate satisfies the following inequalities:

$$\begin{aligned}\langle\varepsilon(u-u^h)\rangle \le \inf_{\tilde{u}\in Y}\Big[& C\big(\langle\varepsilon(u-\tilde{u})\rangle + \nu^{-1}\langle\|\,(u-\tilde{u})_t\,\|_{-1}^2\rangle\big)\\ & + \frac{2}{|\Omega|}\Big(\langle\,|\,b_s(u,u-u^h,u-\tilde{u})\,|\rangle + \langle|\,b_s(u-u^h,u,u-\tilde{u})\,|\rangle\\ & + \langle|\,b_s(u-u^h,u-u^h,u-\tilde{u})\,|\,\rangle\Big)\Big]\\ & + C\inf_{q^h\in\mathbb{Q}^h}\Big[\nu^{-1}\langle\|\,p-q^h\,\|^2\rangle\Big] + \frac{2}{|\Omega|}\langle|\,b_s(u-u^h,u,u-u^h)\,|\rangle\end{aligned} \tag{3.6}$$

and

$$\begin{aligned}\langle\varepsilon(u-u^h)\rangle \le C\inf_{\tilde{u}\in Y}\Big[& \langle\varepsilon(u-\tilde{u})\rangle + \nu^{-1}\langle\|\,(u-\tilde{u})_t\,\|_{-1}^2\rangle\\ & + \nu^{-3}\langle\|\,u-u^h\,\|^{2/3}\|\,\nabla u\,\|^{4/3}\|\,\nabla(u-\tilde{u})\,\|^{4/3}\rangle\\ & + \nu^{-3}\langle\|\,u-u^h\,\|^2\|\,\nabla(u-\tilde{u})\,\|^4\rangle\\ & + \nu^{-1}\langle\|u\,\|\,\|\,\nabla u\,\|\,\|\,\nabla\,(u-\tilde{u})\,\|^2\rangle\Big]\\ & + C\inf_{q^h\in\mathbb{Q}^h}\Big[\nu^{-1}\langle\|\,p-q^h\,\|^2\rangle\Big] + C\nu^{-3}\langle\|\,\nabla u\,\|^4\|\,u-u^h\,\|^2\rangle.\end{aligned} \tag{3.7}$$

Remark 3.2. Estimate (3.7) is not closed in the sense that $\langle\varepsilon(u-u^h)\rangle$ is not estimated only by approximation errors $u-\tilde{u}$ but also by the last term. For this term, we cannot provide an estimate under the assumptions of Theorem 3.2. However, if $\|\,\nabla u\,\|$ is uniformly bounded in time, $\|\,\nabla u\,\| \in L^\infty(0,\infty;L^2(\Omega)^{d\times d})$, then these estimates can be closed provided $\langle\|\,e\,\|^2\rangle \le Ch^\alpha\langle\|\,\nabla e\,\|^2\rangle$ for some $\alpha>0$ and provided h is small enough,

$$\begin{aligned}\nu^{-3}\langle\|\,\nabla u\,\|^4\|\,u-u^h\,\|^2\rangle &\le \nu^{-3}\|\nabla u\|_{L^\infty(0,\infty;L^2)}\langle\|\,u-u^h\,\|^2\rangle \le Ch^\alpha\langle\|\nabla e\|^2\rangle\\ &\le (1-\beta)\langle\varepsilon(u-u^h)\rangle\end{aligned}$$

with $\beta\in(0,1)$ if h is sufficiently small. Now, the term can be absorbed by the left-hand side. However, the assumption on the regularity of u is again the case when pointwise accuracy in time is reasonable to expect rather than accuracy in time-averaged statistics. Thus, the problem of closing the circle in the velocity error equation for the time-averaged statistics seems to catch at the same point as in the standard error analysis.

We shall see, in section 3.2, that in at least one case the circle of analysis is closable. In more general cases, we believe the problem is due to the fact that we are estimating $\langle\|\,\nabla(u-u^h)\,\|^2\rangle$ rather than $\|\langle\nabla(u-u^h)\rangle\|^2$. Of course, an estimate for the latter term follows immediately from (2.31) and (3.7), but it is most probably not optimal.

Proof. The weak solution obtained by the Leray–Hopf construction satisfies

$$(u_t,v^h)+\nu(\nabla u,\nabla v^h)+b_s(u,u,v^h)-(p,\nabla\cdot v^h)=(f,v^h)\qquad \forall\, v^h\in L^\infty(0,T;\mathbb{X}^h)$$

as in the proof of Lemma 3.1. A similar equation holds for u^h. Subtraction and the fact that $(q^h, \nabla \cdot v^h) = 0$ for $q^h \in \mathbb{Q}^h$ give an equation for the error $e = u - u^h$:

$$(e_t, v^h) + \nu(\nabla e, \nabla v^h) + b_s(u, u, v^h) - b_s(u^h, u^h, v^h)$$

$$-(p - q^h, \nabla \cdot v^h) = 0 \quad \forall\, v^h \in L^\infty(0, T; \mathbb{V}^h),$$

where $q^h \in \mathbb{Q}^h$ is arbitrary.

Let $\tilde{u}$ be an interpolant of u with $\tilde{u} \in L^2(0, \infty; H^1(\Omega)^d) \cap L^\infty(0, \infty; \mathbb{V}^h)$ and write $e = \eta - \phi^h$, where $\eta = u - \tilde{u}$ and $\phi^h = u^h - \tilde{u}$. Adding

$$-b_s(e, e, e) + b_s(u, u^h, \phi^h) - b_s(u, u^h, \phi^h),$$

where the first term vanishes, we get with $v^h = \phi^h$

$$\frac{1}{2}\frac{d}{dt}\| \phi^h \|^2 + \nu \| \nabla \phi^h \|^2 = (\eta_t, \phi^h) + \nu(\nabla \eta, \nabla \phi^h) - (p - q^h, \nabla \cdot \phi^h)$$

$$+ b_s(e, u, \eta) - b_s(e, e, \eta) - b_s(e, u, e) + b_s(u, e, \eta).$$

Time-averaging this equation and observing that $\| \phi^h \|$ is uniformly bounded in time (since $\| u^h \|$ is bounded and $\tilde{u} \in L^\infty(0, \infty; L^2(\Omega)^d)$), we have

$$\langle \varepsilon(\phi^h) \rangle \le \frac{1}{|\Omega|}\Big[|\langle\, (\eta_t, \phi^h)\,\rangle| + |\langle \nu\,(\nabla \eta, \nabla \phi^h)\,\rangle| + |\langle\, (p - q^h, \nabla \cdot \phi^h)\,\rangle|$$

$$+ |\langle\, b_s(e, u, \eta)\,\rangle| + |\langle\, b_s(e, e, \eta)\,\rangle| + |\langle\, b_s(e, u, e)\,\rangle|$$

$$+ |\langle\, b_s(u, e, \eta)\,\rangle|\Big].$$

Inequalities (2.30), (2.32), (2.33), and Young's inequality give

$$\langle \varepsilon(\phi^h) \rangle \le C\Big[\nu^{-1}\langle \| \eta_t \|_{-1}^2 \rangle + \langle \varepsilon(\eta) \rangle + \nu^{-1}\langle \| p - q^h \|^2 \rangle\Big]$$

$$+ \frac{2}{|\Omega|}\Big[\langle |\, b_s(e, u, \eta)\,| \rangle + \langle |\, b_s(e, e, \eta)\,| \rangle + \langle |\, b_s(e, u, e)\,| \rangle$$

$$+ \langle |\, b_s(u, e, \eta)\,| \rangle\Big].$$

Now, (3.6) follows from the triangle inequality.

For proving (3.7), we use the following bounds on the trilinear forms, which can

be derived from Lemma 2.1, together with Young's inequality:

$$\begin{aligned}
\langle |\, b_s(e,u,e)\,| \rangle &\le C\,\langle \|\,\nabla u\,\|\,\|\,e\,\|^{1/2}\,\|\,\nabla e\,\|^{3/2}\rangle \\
&\le \frac{\nu}{16}\langle \|\,\nabla e\,\|^2\rangle + C\,\nu^{-3}\langle \|\,\nabla u\,\|^4\,\|\,e\,\|^2\rangle, \\
\langle |\, b_s(e,e,\eta)\,| \rangle &\le C\,\langle \|\,e\,\|^{1/2}\,\|\,\nabla e\,\|^{3/2}\,\|\,\nabla \eta\,\|\rangle \\
&\le \frac{\nu}{16}\langle \|\,\nabla e\,\|^2\rangle + C\,\nu^{-3}\langle \|\,e\,\|^2\,\|\,\nabla \eta\,\|^4\rangle, \\
\langle |\, b_s(e,u,\eta)\,| \rangle &\le C\,\langle \|\,e\,\|^{1/2}\,\|\,\nabla e\,\|^{1/2}\,\|\,\nabla u\,\|\|\,\nabla \eta\,\|\rangle \\
&\le \frac{\nu}{16}\langle \|\,\nabla e\,\|^2\rangle + C\,\nu^{-3}\langle \|\,e\,\|^{2/3}\,\|\,\nabla u\,\|^{4/3}\,\|\,\nabla \eta\,\|^{4/3}\rangle, \\
\langle |\, b_s(u,e,\eta)\,| \rangle &\le C\,\langle \|\,u\,\|^{1/2}\|\,\nabla\, u\,\|^{1/2}\|\,\nabla\, e\,\|\,\|\,\nabla \eta\,\|\rangle \\
&\le \frac{\nu}{16}\langle \|\,\nabla e\,\|^2\rangle + C\langle \nu^{-1}\|u\,\|\,\|\,\nabla u\,\|\,\|\,\nabla\,\eta\,\|^2\rangle.
\end{aligned}$$

Inserting these estimates into (3.6) proves (3.7). ∎

Remark 3.3. Alternative estimates of the trilinear terms in the proof are possible. Using (2.4), one obtains for $\langle |\, b_s(e,u,\eta)\,| \rangle$ the term $\nu^{-1}\langle \|\nabla u\|^2 \|\nabla \eta\|^2\rangle$ which stays on the right-hand side of the estimate (instead of the term

$$\nu^{-3}\langle \|\,u - u^h\,\|^{2/3}\|\,\nabla u\,\|^{4/3}\|\,\nabla \eta\,\|^{4/3}\rangle).$$

The dependency of the alternative term on ν is more favorable, however assuming $\|\nabla \eta\| = \mathcal{O}(h^k)$ and $\|\,u - u^h\,\| = \mathcal{O}(h^{k+\alpha})$, $\alpha > 0$, the order of convergence for the term in estimate (3.7) is higher.

3.2. The case of large u_0 and small $f^*(x)$. There is at least one interesting case in which the error equations for the time-averaged velocity error, $\langle \varepsilon(u - u^h)\rangle$, can be closed: the case of large initial condition u_0 and asymptotically small body force $f(x,t)$. In this subsection, we assume

$$f(x,t) \in L^\infty(0,\infty; H^{-1}(\Omega)^d), \qquad f(x,t) \to f^*(x) \text{ as } t \to \infty,$$

and

$$\nu^{-2} M\,\|\,f^*\,\|_{-1} =: \alpha < 1.$$

In this case, time averaging will eventually wash out the irregularities caused by the large initial condition. We show that this is indeed the case. To shorten the presentation, we shall simplify the condition on f to

$$f(x,t) \equiv f^*(x) \qquad \text{with} \qquad \nu^{-2} M\,\|\,f^*\,\|_{-1} = \alpha < 1. \tag{3.8}$$

This small data (or small Reynolds number) condition can also be formulated as a small generalized Grashof number condition. By a scaling argument, we find that there are constants M_0, M_1, and M_2 depending on the geometry of Ω, but not on $L = diam(\Omega)$, such that $M = M_0\, L^{2-d/2}$, and for $f^*(x) \in C^0(\Omega)$, $\|\,f^*\,\|_{-1} =$

$M_1 L^{1+d/2}$ and $\| f^* \| = M_2 L^{d/2}$. This implies that $\| f^* \|_{-1} = \frac{M_1}{M_2} L \| f^* \|$, so that (3.8) is equivalent to

$$C\, G_r < 1,$$

where G_r is the Grashof number, defined in [7] as

$$G_r = \frac{L^{3-d/2}}{\nu^2} \| f^* \|.$$

LEMMA 3.4. *Let u be a weak solution to the Navier–Stokes equations obtained by the Leray–Hopf construction and suppose* (3.8) *holds. Then*

$$\langle \varepsilon(u - u^*) \rangle = 0,$$

where u^* is the solution of the equilibrium Navier–Stokes equations (2.10).

Proof. Let V_N be the span of eigenfunctions of the Stokes operator. The Leray–Hopf construction gives a sequence $\{u_N\}$ in V_N satisfying

$$(u_{N,t}, v) + \nu(\nabla u_N, \nabla v) + b(u_N, u_N, v) = (f, v) \quad \forall\, v \in V_N, \tag{3.9}$$

with a subsequence $\{u_{N_j}\}$ in V_N converging to a weak solution u, as $N_j \to \infty$, strongly in $L^2(0,T;H(\Omega))$ and weakly in $L^2(0,T;\mathbb{V})$. Let $u_N^* \in V_N$ be the Galerkin projection of u^* in V_N. Then u_N^* satisfies

$$\nu(\nabla u_N^*, \nabla v) + b(u_N^*, u_N^*, v) = (f, v) \quad \forall\, v \in V_N, \tag{3.10}$$

and $u_N^* \to u^*$ in $\mathbb{X}$ and $\mathbb{V}$ as $N \to \infty$.

Set $\phi_N = u_N(x,t) - u_N^*(x)$ and subtract (3.10) from (3.9) to get

$$(\phi_{N,t}, v) + \nu(\nabla \phi_N, \nabla v) + b(u_N, u_N, v) - b(u_N^*, u_N^*, v) = 0 \qquad \forall\, v \in V_N.$$

Let $v = \phi_N$, add and subtract $b(u_N, u_N^*, v)$, and integrate from 0 to T to get

$$\frac{1}{2}\|\phi_N(T)\|^2 + \int_0^T \nu \|\nabla \phi_N\|^2\, dt = \frac{1}{2}\|\phi_N(0)\|^2 - \int_0^T b(\phi_N, u_N^*, \phi_N)\, dt.$$

Setting $v = u_N^*$ in (3.10) gives immediately the a priori bound $\|\nabla u_N^*\| \le \nu^{-1}\|f\|_{-1}$. Using this bound, the bound (2.4) on the trilinear form and the small data assumption (3.8) give

$$\frac{1}{2}\|\phi_N(T)\|^2 + (1-\alpha)\nu \int_0^T \|\nabla \phi_N\|^2\, dt \le \frac{1}{2}\|\phi_N(0)\|^2.$$

Thus, dropping the first term leads to

$$\int_0^T \nu \| \nabla \phi_N \|^2\, dt \le \frac{1}{2(1-\alpha)} \| \phi_N(0) \|^2. \tag{3.11}$$

Using classical properties of weak limits, we have

$$\liminf_{N\to\infty} \left(\int_0^T \| \nabla \phi_N \|^2\, dt \right) \ge \int_0^T \| \nabla \phi \|^2\, dt$$

with $\phi = u - u^*$. Therefore, taking the limit inferior on both sides of (3.11) gives

$$\int_0^T \nu \| \nabla \phi \|^2 \, dt \leq \frac{1}{2(1-\alpha)} \| \phi(0) \|^2.$$

Dividing by $T|\Omega|$ and taking the limit superior as $T \to \infty$ proves the result. ∎

The next lemma is needed to prove the desired error estimate on $\langle \varepsilon(u - u^h) \rangle$ given in Theorem 3.3.

LEMMA 3.5. *Assume that* (3.8) *holds. Then*

$$\langle \varepsilon(u^h - u^{*\,h}) \rangle = 0.$$

Proof. The proof works in the same way as that of Lemma 3.4. It is based on subtracting (2.9) and (2.14). ∎

Remark 3.4. By the Poincaré–Friedrichs inequality, the statements of Lemmas 3.4 and 3.5 hold for the kinetic energy as well.

THEOREM 3.3. *Suppose that the assumptions of Lemmas* 3.4 *and* 3.5 *hold. Then*

$$\langle \varepsilon(u - u^h) \rangle \;\; \leq \;\; 3 \; \varepsilon(u^* - u^{*\,h}). \tag{3.12}$$

Proof. The triangle inequality gives

$$\|\nabla(u - u^h)\|^2 \leq 3 \left(\|\nabla(u - u^*)\|^2 + \|\nabla(u^* - u^{*\,h})\|^2 + \|\nabla(u^{*\,h} - u^h)\|^2 \right).$$

Hence, we get

$$\langle \varepsilon(u - u^h) \rangle \leq 3 \left(\langle \varepsilon(u - u^*) \rangle + \varepsilon(u^* - u^{*\,h}) + \langle \varepsilon(u^{*\,h} - u^h) \rangle \right).$$

The first and the third terms on the right-hand side vanish by Lemmas 3.4 and Lemma 3.5, respectively. ∎

The statement of Theorem 3.3 says that the problem of estimating $\langle \varepsilon(u - u^h) \rangle$ reduces to the one of estimating $\|\nabla(u^* - u^{*\,h})\|^2$. Standard finite element error estimates thus immediately imply that $\langle \varepsilon(u - u^h) \rangle$ is optimal.

COROLLARY 3.2. *Suppose that the assumptions of Lemmas* 3.4 *and* 3.5 *hold and that* $(\mathbb{X}^h, \mathbb{Q}^h)$ *satisfies the inf-sup condition* (2.5). *Then*

$$\langle \varepsilon(u - u^h) \rangle \leq C \left[\inf_{v^h \in \mathbb{X}^h} \nu \| \nabla(u^* - v^h) \|^2 + \inf_{q^h \in \mathbb{Q}^h} \nu^{-1} \| p^* - q^h \|^2 \right].$$

Proof. This follows by inserting the error estimates for $\| \nabla(u^* - u^{*\,h}) \|$ from [11] into the right-hand side of (3.12). ∎

Concerning the time-averaged error in the pressure, we have the following corollary. It is a direct consequence of Corollary 3.1.

COROLLARY 3.3. *Let the assumptions of Theorem* 3.1 *and Lemmas* 3.4 *and* 3.5 *be fulfilled. Then*

$$\| \langle p - p^h \rangle \| \leq \frac{3\sqrt{3}\nu}{\beta^h} \| \nabla(u^* - u^{*\,h}) \| + \left(1 + \frac{\sqrt{d}}{\beta^h} \right) \inf_{q^h \in \mathbb{Q}^h} \limsup_{T \to \infty} \| \langle p - q^h \rangle_T \|.$$

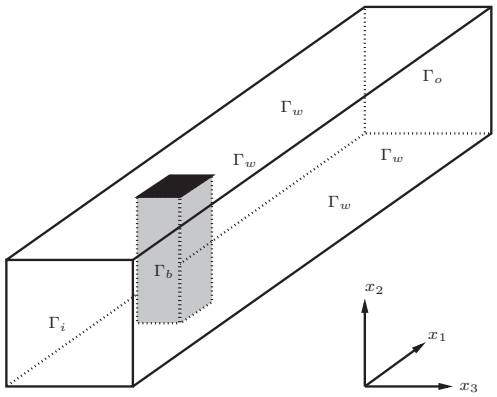


FIG. 1. *A channel with a body.*

4. Drag and lift. Consider the flow around a body in a channel with flow region Ω and boundary Γ, which consists of Γ_b (boundary of the body) and $\Gamma_c = \Gamma_i \cup \Gamma_o \cup \Gamma_w$ (where Γ_i, Γ_o correspond to the inflow and outflow and Γ_w correspond to the walls); see Figure 1 for an example.

Define

$$\sigma = -p\,\mathbb{I} + 2\,\mu\,\nabla^s u,$$

where ∇^s indicates the symmetric part of the operator ∇, and μ the dynamic viscosity. Then, we consider the Navier–Stokes equations written in the form

$$\begin{aligned} \rho(u_t + u\cdot\nabla u) &= \nabla\cdot\sigma + f && \text{in } \Omega\times(0,T],\\ \nabla\cdot u &= 0 && \text{in } \Omega\times[0,T],\\ u &= g && \text{on } \Gamma\times[0,T],\\ u(x,0) &= u_0(x) && \text{in } \Omega, \end{aligned} \tag{4.1}$$

satisfying the compatibility condition $\int_\Gamma g\cdot\hat{n}\,dS = 0$, where $\hat{n}$ is the outward pointing unit normal to Γ_b. We assume $g = 0$ on $\Gamma_w \cup \Gamma_b$.

We introduce the spaces

$$\mathbb{X}_g = \left\{ v \in H^1(\Omega)^d : v|_\Gamma = g \right\}, \quad \mathbb{X}_0 = H_0^1(\Omega)^d.$$

A weak formulation of (4.1) reads as follows: Find $u : [0,T] \to \mathbb{X}_g$ and $p : (0,T] \to \mathbb{Q}$ such that

$$\rho\,(u_t, v) + 2\,\mu(\nabla^s u, \nabla^s v) + \rho\,b(u,u,v) - (p, \nabla\cdot v) + (\nabla\cdot u, q) = (f, v) \qquad \forall (v,q) \in \mathbb{X}_0\times\mathbb{Q}. \tag{4.2}$$

We assume that the asymptotic growth of the kinetic energy of the solutions of (4.2) is bounded by

$$\limsup_{T\to\infty} \frac{1}{T} k(u(T)) = 0. \tag{4.3}$$

In particular, assumption (4.3) is valid if the solutions of (4.2) have bounded kinetic

energy.[1]

Drag and lift are defined by

$$D = -\int_{\Gamma_b} \hat{e}_1 \cdot \sigma \cdot \hat{n}\, d\gamma, \qquad L = -\int_{\Gamma_b} \hat{e}_2 \cdot \sigma \cdot \hat{n}\, d\gamma,$$

respectively, where $\hat{e}_i$ is the unit vector in the ith direction. We assume that $+\hat{e}_1$ is the direction of motion and $-\hat{e}_2$ is the direction of gravity. We note that if (u, p) are sufficiently regular and $v \in L^\infty(0, T; H^1(\Omega)^d)$ with $v = 0$ on Γ_c, then we have

$$\frac{1}{T}\int_0^T \int_{\Gamma_b} v \cdot \sigma \cdot \hat{n}\, d\gamma\, dt$$

$$= \frac{1}{T}\int_0^T \Big[\rho\,(u_t, v) + 2\,\nu(\nabla^s u, \nabla^s v) + \rho\, b(u, u, v) - (p, \nabla \cdot v) - (f, v)\Big]\, dt. \tag{4.4}$$

Whenever the right-hand side of (4.4) is well defined, it can be taken as the definition of the force on the immersed body, i.e., the left-hand side of (4.4). We will use this definition under the following regularity assumptions on (u, p):

$$u \in L^1(0, T; H^1(\Omega)^d) \cap \mathbb{X}_g, \quad u_t \in L^1(0, T; H^{-1}(\Omega)^d), \quad p \in L^1(0, T; L^2(\Omega)).$$

Choosing in (4.4) a function v satisfying $v = -\hat{e}_1$ on Γ_b gives a formula for the drag and $v = -\hat{e}_2$ gives a formula for the lift.

THEOREM 4.1. *Let the force on the immersed body be well defined by* (4.4), *let the kinetic energy of the solutions of* (4.2) *be bounded by* (4.3), *and let a similar bound be valid for the kinetic energy of the finite element approximation* u^h. *Then, the time-averaged drag and lift can be estimated as*

$$|\langle D - D^h \rangle|, |\langle L - L^h \rangle|$$

$$\leq C\Big[(\mu + M\rho\langle\|\nabla u\|^2 + \|\nabla u^h\|^2\rangle^{1/2})\langle\|\nabla (u - u^h)\|^2\rangle^{1/2}$$

$$+ \limsup_{T\to\infty} \|\langle p - p^h\rangle_T\|\Big]. \tag{4.5}$$

Proof. We present here a proof for the drag estimate. With the same arguments follows the estimate for the lift, according to an appropriate choice of v.

Let $v \in H^2(\Omega)^d$ be a time-independent vector field satisfying $v = -\hat{e}_1$ on Γ_b and $v = 0$ on Γ_c. From (4.4) it follows that

$$\int_0^T D\, dt = \int_0^T \Big[\rho\,(u_t, v) + 2\,\mu(\nabla^s u, \nabla^s v) + \rho\, b(u, u, v) - (p, \nabla v) - (f, v)\Big]\, dt.$$

Let $i_h v$ be a finite element interpolant to v such that $i_h v = -\hat{e}_1$ on Γ_b. In particular, we have $v, i_h v \in L^\infty(0, T; H^1(\Omega)^d)$. Then,

$$\int_0^T \Big[D - D^h\Big]\, dt = \int_0^T \Big[\rho\,(u_t, v) + 2\,\mu(\nabla^s u, \nabla^s v) + \rho\, b(u, u, v)$$

$$- (p, \nabla \cdot v) - (f, v) - \rho\,(u_t^h, i_h v) - 2\,\mu(\nabla^s u^h, \nabla^s i_h v)$$

$$-\rho\, b_s(u^h, u^h, i_h v) + (p^h, \nabla \cdot i_h v) + (f, i_h v)\Big]\, dt.$$

[1] This is known for geometries, such as simply connected $\partial\Omega$, which admit a Hopf extension of the boundary condition into Ω. It is unknown, for instance, if $\partial\Omega = \cup\Gamma_j$ is multiply connected and $\int_{\Gamma_j} g\, d\gamma = 0$, but $\int_{\Gamma_j} g\, d\gamma \neq 0$ for some connected component; see Theorem 4.1, p. 33, of Galdi [9].

Adding and subtracting terms and using $b(u,u,v) = b_s(u,u,v)$, this becomes

$$\begin{aligned}\int_0^T \Big[D - D^h\Big]\,dt = \int_0^T \Big[&\rho\,(u_t, v - i_h v) + \rho\,((u-u^h)_t, i_h v) + 2\,\mu(\nabla^s u, \nabla^s (v - i^h v))\\ &+ 2\,\mu(\nabla^s (u-u^h), \nabla^s\, i_h v) + \rho\, b_s(u,u,v-i_h v) + \rho\, b_s(u,u,i_h v)\\ &- \rho\, b_s(u^h,u^h,i_h v) - (p, \nabla\cdot(v - i_h v))\\ &- (p-p^h, \nabla\cdot i_h v) - (f, v - i_h v)\Big]\,dt. \qquad (4.6)\end{aligned}$$

Observe now that the term containing $(f, v - i_h v)$ in (4.6) can be rewritten by multiplying (4.1) by $v - i_h v$ and integrating:

$$(f, v-i_h v) = \rho\,(u_t, v-i_h v) + 2\,\mu(\nabla^s u, \nabla^s\,(v-i_h v)) + \rho\, b_s(u,u,v-i_h v) - (p, \nabla\cdot(v-i_h v)). \qquad (4.7)$$

Hence, combining (4.6) and (4.7) gives

$$\begin{aligned}\int_0^T \Big[D - D^h\Big]\,dt = \int_0^T \Big[&\rho\,((u-u^h)_t, i_h v) + 2\,\mu(\nabla^s\,(u-u^h), \nabla^s i^h v)\\ &+\rho\, b_s(u,u,i_h v) - \rho\, b_s(u^h,u^h,i_h v) + (p-p^h, \nabla\cdot i_h v)\Big]\,dt. \qquad (4.8)\end{aligned}$$

Let $e = u - u^h$ and divide each term in (4.8) by T. The first term on the right-hand side yields, with the time independence of $i_h v$,

$$\frac{1}{T}(e(T) - e(0), i_h v) \le \frac{C_{PF}}{T}\,(\|e(T)\| + \|e(0)\|)\,\|\nabla i_h v\|.$$

Note that the Poincaré–Friedrich inequality can be applied since $i_h v = 0$ on Γ_c. Now, by (2.22) and $\|\nabla^s v\| \le \|\nabla v\|$ for all $v \in H^1(\Omega)^d$, we obtain

$$\langle(\nabla^s e, \nabla^s\, i_h v)\rangle_T \;\le\; \langle\|\nabla e\|^2\rangle_T^{1/2}\,\|\nabla\, i_h v\|.$$

Next, the nonlinear terms are estimated by (2.4), (2.22), and Young's inequality:

$$\begin{aligned}\langle b_s(u,u,i_h v) - b_s(u^h,u^h,i_h v)\rangle_T &= \langle b_s(u,e,i_h v) + b_s(e,u^h,i_h v)\rangle_T\\ &\le M\langle(\|\nabla u\| + \|\nabla u^h\|)\|\nabla e\|\|\nabla i_h v\|\rangle_T\\ &\le M\|\nabla i_h v\|\langle(\|\nabla u\| + \|\nabla u^h\|)^2\rangle_T^{1/2}\langle\|\nabla e\|^2\rangle_T^{1/2}\\ &\le 2M\|\nabla i_h v\|\langle\|\nabla u\|^2 + \|\nabla u^h\|^2\rangle_T^{1/2}\langle\|\nabla e\|^2\rangle_T^{1/2}.\end{aligned}$$

For the pressure term, we obtain, with the time independence of $i_h v$ and the Cauchy–Schwarz inequality,

$$\langle(p-p^h, \nabla\cdot i_h v)\rangle_T = (\langle p-p^h\rangle_T, \nabla\cdot i_h v) \le \sqrt{d}\,\|\langle p-p^h\rangle_T\|\,\|\nabla\, i_h v\|.$$

Putting everything together, (4.8) becomes

$$\begin{aligned}|\langle D - D^h\rangle_T| \le \rho\frac{C_{PF}}{T}\,&(\|e(T)\| + \|e(0)\|)\,\|\nabla i_h v\|\\ &+(2\mu + 2M\rho\langle\|\nabla u\|^2 + \|\nabla u^h\|^2\rangle_T^{1/2})\,\langle\|\nabla\, e\|^2\rangle_T^{1/2}\,\|\nabla\, i_h v\|\\ &+\sqrt{d}\,\|\langle p-p^h\rangle_T\|\,\|\nabla\, i_h v\|.\end{aligned}$$

Taking the lim sup on both sides of the inequality, using (4.3) and the same property for u^h, and (2.26) give

$$\begin{aligned}
&\limsup_{T\to\infty} |\langle D - D^h \rangle_T| \\
&\qquad \le (2\mu + 2M\rho\langle \|\nabla u\|^2 + \|\nabla u^h\|^2\rangle^{1/2}) \langle \| \nabla e \|^2\rangle^{1/2} \| \nabla i_h v \| \\
&\qquad\quad + \sqrt{d} \limsup_{T\to\infty} \|\langle p - p^h\rangle_T\| \| \nabla i_h v \|.
\end{aligned}$$

Finally, applying (2.25) to the left-hand side of the estimate and $\| \nabla i_h v \| \le 2\| v \|_2 = C$ give the statement of theorem. ∎

The constant C in the estimate of Theorem 4.1 depends on the $H^2(\Omega)^d$-norm of the test function v and the size of Ω by the Poincaré–Friedrich constant.

In the last part of the proof of Theorem 4.1, the assumption on the boundedness of the growth of the kinetic energy is used. The boundedness of the kinetic energy has already been proven for the case $g = 0$ on Γ; see Lemmas 3.1 and 3.2. Thus, the statement of Theorem 4.1 applies for this case, and the respective bounds of Lemmas 3.1 and 3.2 can be inserted into (4.5).

The assumptions on g can be extended to sufficiently small Dirichlet data. We consider for simplicity of presentation the case when g does not depend on time. Let G be the Hopf extension of the boundary condition g to Ω, i.e., $G \in H^1(\Omega)^d$, $\nabla\cdot G = 0$, $G_t = 0$, and define $\tilde{u} = u - G$. Then $\tilde{u} \in \mathbb{V}$ and this function satisfies

$$\begin{aligned}
\rho(\tilde{u}_t + \tilde{u}\cdot\nabla\tilde{u}) &= \nabla\cdot\tilde{\sigma} + f + 2\mu\nabla\cdot\nabla^s G + \rho G\cdot\nabla G + \rho G \\
&\quad \cdot\nabla\tilde{u} + \rho\tilde{u}\cdot\nabla G \qquad \text{in } \Omega\times(0,T], \\
\nabla\cdot\tilde{u} &= 0 \qquad \text{in } \Omega\times[0,T], \\
\tilde{u} &= 0 \qquad \text{on } \Gamma\times[0,T], \\
\tilde{u}(x,0) &= (u_0 - G)(x) \qquad \text{in } \Omega,
\end{aligned} \tag{4.9}$$

where $\tilde{\sigma} = -\tilde{p}\,\mathbb{I} + 2\,\mu\,\nabla^s\tilde{u}$. Denote $\tilde{f} = f + 2\mu\nabla\cdot\nabla^s G + \rho G\cdot\nabla G$. Transforming (4.9) in the usual way to a variational formulation using $\tilde{u}$ as a test function, and applying the estimate for the dual pairing, Korn's inequality ($\|\nabla\tilde{u}\| \le C_{\text{Korn}}\|\nabla^s\tilde{u}\|$) and Young's inequality give

$$\frac{\rho}{2}\frac{d}{dt}\|\tilde{u}\|^2 + \frac{3\mu}{2}\|\nabla^s\tilde{u}\|^2 \le \frac{C}{\mu}\|\tilde{f}\|_{-1}^2 + \rho C_{\text{Korn}}^2\|\nabla^s\tilde{u}\|^2\|G\|_1. \tag{4.10}$$

If G is sufficiently small, for instance if

$$\rho C_{\text{Korn}}^2\|G\|_1 < \mu, \tag{4.11}$$

the last term on the right-hand side of (4.10) can be absorbed into the left-hand side and the boundedness of the kinetic energy of $\tilde{u}$ is obtained by integrating (4.10) on $(0,T)$. The kinetic energy of u is bounded by using the triangle inequality and the fact that $\|G\|$ is a constant. In the discrete case, by subtracting an interpolant of G^h from u^h and using a similar argument, one obtains this property for u^h. Thus for small G, the assumption of Theorem 4.1 of bounded kinetic energy holds.

COROLLARY 4.1. *Let the force on the immersed body be well defined by* (4.4), *let* $\tilde{f} = f + 2\mu\nabla\cdot\nabla^s G + \rho G\cdot\nabla G \in L^2(0,T;H^{-1}(\Omega))$, *and let the smallness condition*

(4.11) *on the inflow be fulfilled. Then, the time-averaged drag and lift can be estimated by* (4.5).

Remark 4.1. The estimate in Corollary 4.1 is not closed (the same problem as in Theorem 3.1) due to the term $\langle \| \nabla (u - u^h) \|^2 \rangle$ in (4.5). Closing this estimate requires the investigation (beyond the scope of this paper) of (4.9), which has additional terms in comparison with the Navier–Stokes equations. It is very likely that a small data assumption on $\tilde{f}$ (like in section 3.2) has to be used.

Note that the condition (4.11) can be reformulated with the Poincaré–Friedrichs inequality since G vanishes on $\Gamma_w \cup \Gamma_b$.

5. Persistent shear flows. We have seen in the previous sections that, provided that both the portion of the persistent body force driving the flow and the (nonhomogeneous) Dirichlet data are small, statistics, such as the time-averaged energy dissipation rate (or drag and lift), can be accurately predicted by a flow simulation. This accuracy holds quite generally without any further assumptions on u_0, ν, and *Re* typically needed to prove accuracy over bounded time intervals.

The case when the persistent forces driving the flow are not small is much more difficult. We shall prove in this section that the analogous estimate of the time-averaged energy dissipation rate is physically reasonable under a condition on the finite element mesh near the walls. Briefly, we consider the finite element approximation to the following shear flow problem: let $\Omega = [0, L]^3$, and find (u, p) satisfying

$$\begin{aligned} \frac{\partial u}{\partial t} + u \cdot \nabla u - \nu \Delta u + \nabla p &= 0 && \text{in } \Omega \times (0, T], \\ \nabla \cdot u &= 0 && \text{in } \Omega \times [0, T], \\ u &= u_0 && \text{at } t = 0, \\ u(x_1, x_2, x_3, t) &= \phi(x_3) && \text{for } x_3 \in \{0, L\}, \end{aligned}$$

where

$$\phi(x_3) = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \text{ if } x_3 = 0 \quad \text{and} \quad \phi(x_3) = \begin{pmatrix} U \\ 0 \\ 0 \end{pmatrix} \text{ if } x_3 = L,$$

and periodic boundary conditions in the x_1, x_2 directions; see Figure 2.

In this problem, the persistent force driving the flow is clearly the motion of the top wall and the time-averaged energy dissipation rate must balance the drag exerted by the walls on the fluid. For such problems, the Richardson–Kolmogorov energy

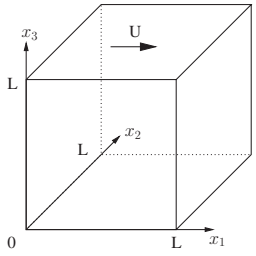


FIG. 2. *The shear flow problem.*

cascade predicts quite simply[2] that

$$\langle \varepsilon(u) \rangle \approx \frac{U^3}{L}$$

independently of ν and Re, respectively. Remarkably, the upper estimate has also been proven for weak solutions of the Navier–Stokes equations in full generality,

$$\langle \varepsilon(u) \rangle \leq C \frac{U^3}{L}$$

by Constantin and Doering [4] and Wang [27]; see also Doering and Foias [6]. Herein, we show in essence that provided the first mesh line of the finite element space is within $O(1/Re)$ of the top wall, then

$$\langle \varepsilon(u^h) \rangle \leq C \frac{U^3}{L};$$

i.e., the computed energy dissipation has the correct mathematical and physical scaling. This result depends on the precise construction of a discrete background flow. The mesh condition is necessary for the construction. Since the proof adapts the ideas of [27], we expect that a similar analysis would hold for other variational methods as well.

Let

$$\begin{aligned}
\mathbb{X} = \{ & v \in H^1(\Omega)^d \ : \ v(x_1, x_2, x_3, t) = v(x_1 + L, x_2, x_3, t) \text{ for } x_1 \in \Gamma, \\
& v(x_1, x_2, x_3, t) = v(x_1, x_2 + L, x_3, t) \text{ for } x_2 \in \Gamma, \\
& v(x_1, x_2, x_3, t) = \phi(x_3) \text{ for } x_3 \in \Gamma \}, \\
\mathbb{X}_0 = \{ & v \in H^1(\Omega)^d \ : \ v(x_1, x_2, x_3, t) = v(x_1 + L, x_2, x_3, t) \text{ for } x_1 \in \Gamma, \\
& v(x_1, x_2, x_3, t) = v(x_1, x_2 + L, x_3, t) \text{ for } x_2 \in \Gamma, \\
& v(x_1, x_2, x_3, t) = 0 \text{ for } x_3 \in \Gamma \}, \\
\mathbb{Q} = & L_0^2(\Omega)
\end{aligned}$$

and denote the corresponding conforming finite element spaces with a superscript h. We assume that the finite element space for the velocity contains linears, which is usually the case for conforming velocity finite element spaces. The finite element problem reads as follows: find $u^h : [0, T] \to \mathbb{X}^h$, $p^h : (0, T] \to \mathbb{Q}^h$ such that

$$\begin{aligned}
(u_t^h, v^h) + \nu(\nabla u^h, \nabla v^h) + b_s(u^h, u^h, v^h) - (p^h, \nabla \cdot v^h) = 0 \quad & \forall\, v^h \in \mathbb{X}_0^h, \\
(\nabla \cdot u^h, q^h) = 0 \quad & \forall\, q^h \in \mathbb{Q}^h, \\
(u^h(x, 0) - u_0(x), v^h) = 0 \quad & \forall\, v^h \in \mathbb{X}_0^h.
\end{aligned} \tag{5.1}$$

[2]Briefly, the largest coherent structures are associated with the motion of the upper wall. They thus have length scale L and characteristic velocity U. Their local Reynolds number is thus $(\frac{U\,L}{\nu} (= Re))$ and the viscous dissipation is negligible on them. These break up into smaller eddies (velocity u, length l, $Re\,(l) = \frac{u\,(l)\,l}{\nu}$) until $Re\,(l)$ is small enough for viscous dissipation to drive their kinetic energy to zero exponentially fast. Since viscous dissipation is negligible through this cascade, the energy dissipation rate is related then to the power input to the largest scales at the first step in the cascade. These largest eddies have energy $\frac{1}{2}U^2$ and time scale $\tau = \frac{L}{U}$ so the rate of energy transfer is $O(\frac{U^2}{\tau}) = O(\frac{U^3}{L})$.

Consider the background flow (an extension of the boundary condition ϕ to the interior of Ω) given by

$$\tilde{\phi}(x_3) = \begin{cases} 0 & \text{if } x_3 \in [0, L - \gamma L], \\ \frac{U}{\gamma L}(x_3 - (L - \gamma L)) & \text{if } x_3 \in [L - \gamma L, L], \end{cases}$$

and define

$$\Phi = \begin{bmatrix} \tilde{\phi}(x_3) \\ 0 \\ 0 \end{bmatrix},$$

where γ is a positive number, referred to as the "boundary layer thickness." For $\gamma \in (0, 1)$ this function is piecewise linear, continuous, and satisfies the boundary conditions. Straightforward calculations give

$$\| \Phi \|_{L^\infty} = U, \tag{5.2}$$

$$\| \nabla \Phi \|_{L^\infty} = \frac{U}{\gamma L}, \tag{5.3}$$

$$\begin{aligned} \| \Phi \|^2 &= L^2 \int_0^L | \tilde{\phi}(x_3) |^2 dx_3 \\ &= L^2 \int_{L-\gamma L}^L \frac{U^2}{(\gamma L)^2} (x_3 - (L - \gamma L))^2 dx_3 = \frac{U^2 \gamma L^3}{3}, \end{aligned} \tag{5.4}$$

$$\| \nabla \Phi \|^2 = \frac{U^2 L}{\gamma}. \tag{5.5}$$

For completeness, we include a short proof of the scaling of the constant in the Poincaré–Friedrichs inequality that will be helpful in the proof of the next theorem.

LEMMA 5.1. *Let $\mathcal{O}_{\gamma L} = \{(x_1, x_2, x_3) \in \Omega : L - \gamma L \le x_3 \le L\}$ be the region close to the upper boundary (where the background flow Φ does not vanish). Then*

$$\| u^h - \Phi \|_{L^2(\mathcal{O}_{\gamma L})} \le \gamma L \| \nabla (u^h - \Phi) \|_{L^2(\mathcal{O}_{\gamma L})}. \tag{5.6}$$

Proof. First, let v be a C^1 function on $\mathcal{O}_{\gamma L}$ that vanishes for $x_3 = L$. Then, componentwise ($i = 1, 2, 3$), we have

$$v_i(x_1, x_2, x_3) = v_i(x_1, x_2, L) - \int_{x_3}^L \frac{dv_i}{d\xi_3}(x_1, x_2, \xi_3) \, d\xi_3.$$

Observing that $v_i(x_1, x_2, L) = 0$, squaring both sides, and using the Cauchy–Schwarz inequality, we get

$$v_i^2(x_1, x_2, x_3) \le \gamma L \int_{L-\gamma L}^L \left(\frac{dv_i}{d\xi_3}(x_1, x_2, \xi_3) \right)^2 d\xi_3.$$

Integrating both sides with respect to x_3 gives

$$\int_{L-\gamma L}^L v_i^2(x_1, x_2, x_3) \, dx_3 \le (\gamma L)^2 \int_{L-\gamma L}^L \left(\frac{dv_i}{d\xi_3}(x_1, x_2, \xi_3) \right)^2 d\xi_3.$$

Then, integrating with respect to x_1 and x_2 and summing from $i = 1$ to 3, we obtain

$$\| v \|^2_{\mathcal{O}_{\gamma L}} \leq (\gamma L)^2 \| \nabla v \|^2_{\mathcal{O}_{\gamma L}},$$

which proves the lemma in the case of a C^1 function. The case $v \in \mathbb{X}^h \subset H^1(\Omega)$ follows by density. Finally, just take $v = u^h - \Phi$. ∎

THEOREM 5.1. *Let*

$$\gamma = \frac{1}{C_\gamma \, Re}, \quad \textit{where} \quad Re = \frac{U\,L}{\nu}, \quad C_\gamma \geq 5, \tag{5.7}$$

selected such that Φ belongs to the finite element space. Then, the time-averaged energy dissipation rate of the finite element velocity satisfies

$$\langle \varepsilon(u^h) \rangle \quad \leq C \, \frac{U^3}{L},$$

where $C = C(C_\gamma)$.

Proof. We take $v^h = u^h - \Phi$ (this is possible due to the choice of γ) in (5.1) to get

$$(u^h_t, u^h) - (u^h_t, \Phi) + \nu \| \nabla u^h \|^2 - \nu \, (\nabla \, u^h, \nabla \, \Phi) - b_s(u^h, u^h, \Phi) = 0 \tag{5.8}$$

since $b_s(\cdot, \cdot, \cdot)$ is skew-symmetric and Φ is divergence free.

Observing that $(u^h_t, \Phi) = \frac{d}{d\,t}(u^h, \Phi)$, since $\frac{d\,\Phi}{d\,t} = 0$, we rewrite (5.8) as

$$\frac{1}{2} \frac{d}{d\,t} \| u^h \|^2 + \nu \| \nabla u^h \|^2 = \frac{d}{d\,t}(u^h, \Phi) + b_s(u^h, u^h, \Phi) + \nu \, (\nabla \, u^h, \nabla \, \Phi)$$

and integrate in time to get

$$\begin{aligned} &\frac{1}{2} \| u^h(T) \|^2 - \frac{1}{2} \| u^h(0) \|^2 + \nu \int_0^T \| \nabla u^h \|^2 \, d\,t \\ &= (u^h(T), \Phi) - (u^h(0), \Phi) + \int_0^T b_s(u^h, u^h, \Phi) \, d\,t + \nu \int_0^T (\nabla \, u^h, \nabla \, \Phi) \, d\,t. \end{aligned} \tag{5.9}$$

We need to estimate each term on the right-hand side of (5.9). For that, we use (5.2)–(5.5) to derive upper bounds for the following terms:

$$(u^h(T), \Phi) \leq \frac{1}{2} \| u^h(T) \|^2 + \frac{1}{2} \| \Phi \|^2 = \frac{1}{2} \| u^h(T) \|^2 + \frac{1}{6} U^2 \gamma L^3,$$

$$(u^h(0), \Phi) \leq \| u^h(0) \| \, \| \Phi \| = \sqrt{\frac{\gamma}{3}} U L^{3/2} \| u^h(0) \|,$$

$$\begin{aligned} \nu \int_0^T (\nabla \, u^h, \nabla \, \Phi) \, d\,t &\leq \frac{\nu}{2} \int_0^T \| \nabla u^h \|^2 \, d\,t + \frac{\nu}{2} \int_0^T \| \nabla \, \Phi \|^2 \, d\,t \\ &= \frac{\nu}{2} \int_0^T \| \nabla u^h \|^2 \, d\,t + \frac{\nu}{2\,\gamma} L \, U^2 \, T. \end{aligned}$$

For the nonlinear term, we add and subtract terms, and use the fact that $b_s(\cdot,\cdot,\cdot)$ is skew-symmetric to write

$$\begin{aligned} b_s(u^h,u^h,\Phi) &= b_s(u^h-\Phi,u^h-\Phi,\Phi)+b_s(\Phi,u^h-\Phi,\Phi)\\ &= \frac{1}{2}b(u^h-\Phi,u^h-\Phi,\Phi)-\frac{1}{2}b(u^h-\Phi,\Phi,u^h-\Phi)\\ &\quad +\frac{1}{2}b(\Phi,u^h-\Phi,\Phi)-\frac{1}{2}b(\Phi,\Phi,u^h-\Phi). \end{aligned} \tag{5.10}$$

We use Lemma 5.1 together with (5.2)–(5.5) to analyze the terms in (5.10). In all cases, integration is restricted to $\mathcal{O}_{\gamma L}$ since $\mathrm{supp}(\Phi)=\overline{\mathcal{O}}_{\gamma L}$:

$$\begin{aligned} b(u^h-\Phi,u^h-\Phi,\Phi) &\le \|\Phi\|_{L^\infty(\mathcal{O}_{\gamma L})}\|\nabla(u^h-\Phi)\|_{L^2(\mathcal{O}_{\gamma L})}\|u^h-\Phi\|_{L^2(\mathcal{O}_{\gamma L})}\\ &\le U\gamma L\|\nabla(u^h-\Phi)\|^2_{L^2(\mathcal{O}_{\gamma L})}\\ &\le 2U\gamma L\|\nabla u^h\|^2+2U^3L^2,\\ b(u^h-\Phi,\Phi,u^h-\Phi) &\le \|\nabla\Phi\|_{L^\infty(\mathcal{O}_{\gamma L})}\|u^h-\Phi\|^2_{L^2(\mathcal{O}_{\gamma L})}\\ &\le U\gamma L\|\nabla(u^h-\Phi)\|^2\\ &\le 2U\gamma L\|\nabla u^h\|^2+2U^3L^2,\\ b(\Phi,u^h-\Phi,\Phi) &\le \|\Phi\|_{L^\infty(\mathcal{O}_{\gamma L})}\|\Phi\|_{L^2(\mathcal{O}_{\gamma L})}\|\nabla(u^h-\Phi)\|_{L^2(\mathcal{O}_{\gamma L})}\\ &\le U\left(\frac{1}{3}U^2\gamma L^3\right)^{1/2}\|\nabla u^h\|+U\left(\frac{1}{3}U^2\gamma L^3\right)^{1/2}\left(\frac{U^2L}{\gamma}\right)^{1/2}\\ &\le \frac{1}{2}U\gamma L\|\nabla u^h\|^2+\frac{5}{6}U^3L^2,\\ b(\Phi,\Phi,u^h-\Phi) &\le \|\Phi\|_{L^\infty(\mathcal{O}_{\gamma L})}\|\nabla\Phi\|_{L^2(\mathcal{O}_{\gamma L})}\|u^h-\Phi\|_{L^2(\mathcal{O}_{\gamma L})}\\ &\le U\gamma L\left(\frac{U^2L}{\gamma}\right)^{1/2}\|\nabla u^h\|+U\gamma L\left(\frac{U^2L}{\gamma}\right)^{1/2}\left(\frac{U^2L}{\gamma}\right)^{1/2}\\ &\le \frac{1}{2}U\gamma L\|\nabla u^h\|^2+\frac{3}{2}U^3L^2. \end{aligned}$$

Inserting these estimates into (5.10) gives

$$|b_s(u^h,u^h,\Phi)|\le\frac{5}{2}U\gamma L\|\nabla u^h\|^2+\frac{19}{6}U^3L^2.$$

Hence, estimate (5.9) becomes

$$\begin{aligned} \nu\int_0^T\|\nabla u^h\|^2\,dt &\le \frac{1}{6}U^2\gamma L^3+\frac{1}{2}\|u^h(0)\|^2+\sqrt{\frac{\gamma}{3}}UL^{3/2}\|u^h(0)\|+\frac{\nu}{2}\int_0^T\|\nabla u^h\|^2\,dt\\ &\quad +\frac{\nu}{2\gamma}LU^2T+\frac{5}{2}\gamma LU\int_0^T\|\nabla u^h\|^2\,dt+\frac{19}{6}U^3L^2T. \end{aligned}$$

Dividing by T, taking lim sup as $T \to \infty$, and using $|\Omega| = L^3$ lead to

$$L^3 \left(\frac{1}{2} - \frac{5}{2} \frac{L \gamma U}{\nu} \right) \langle \varepsilon(u^h) \rangle \leq \frac{\nu}{2\gamma} L U^2 + \frac{19}{6} U^3 L^2.$$

Inserting (5.7) on both sides of this estimate proves the theorem. ∎

6. Summary and outlook. There are many important and interesting questions concerning the computational approximation of statistics in different flow settings. We studied the simple yet interesting case of internal flow, for instance, for small data (Reynolds number, body force) but large initial condition. In this setting, we proved convergence of the time-averaged energy dissipation rate in three dimensions without assuming that the solution is a strong, smooth, and unique solution, or the initial data is smooth. We then analyzed the convergence of time-averaged drag and lift coefficients. The crucial assumption of the general theory, a certain control on the kinetic energy, was proven for sufficiently small inflow boundary conditions. Lastly, we considered a simple case of shear flow and showed that the statistics of the computed solutions scale as predicted by the Kolmogorov theory.

One next step could be to consider discretization in time and approximate infinite time averaging by finite time averaging [8]. Our long-term goal is to develop a theory paralleling and inspired by the theory of shadowing in approximation of dynamical systems [23, 22, 14]. Are the computed statistics the exact statistics of a perturbed flow? Answering this numerical analysis question is fundamental for turbulent flow simulation.

Acknowledgment. The authors would like to thank the referees for their careful analysis of this manuscript and their important suggestions.

REFERENCES

[1] R. A. Adams, *Sobolev Spaces*, Academic Press, New York, 1975.
[2] L. C. Berselli, T. Iliescu, and W. Layton, *Mathematics of Large Eddy Simulation of Turbulent Flows*, Sci. Comput., Springer-Verlag, 2006.
[3] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer-Verlag, New York, 1991.
[4] P. Constantin and C. Doering, *Energy dissipation in shear driven turbulence*, Phys. Rev. Lett., 69 (1992), pp. 1648–1651.
[5] J. W. Deardorff, *A numerical study of three-dimensional turbulent channel flow of large Reynolds numbers*, J. Fluid Mech., 41 (1970), pp. 453–480.
[6] C. R. Doering and C. Foias, *Energy dissipation in body-forced turbulence*, J. Fluid Mech., 467 (2002), pp. 289–306.
[7] C. R. Doering and J. D. Gibbon, *Applied Analysis of the Navier-Stokes Equations*, Cambridge University Press, Cambridge, UK, 1995.
[8] C. Foias, M. S. Jolly, O. P. Manley, R. Rosa, and R. Temam, *Kolmogorov theory via finite-time averages*, Phys. D, 212 (2005), pp. 245–270.
[9] G. P. Galdi, *An Introduction to the Mathematical Theory of the Navier-Stokes Equations. Volume* II. *Nonlinear Steady Problems*, Springer Tracts in Natural Philosophy 39, Springer-Verlag, New York, 1994.
[10] G. P. Galdi, *An introduction to the Navier-Stokes initial-boundary value problem*, in Fundamental Directions in Mathematical Fluid Dynamics, G. P. Galdi, J. G. Heywood, and R. Rannacher, eds., Birkhäuser, Basel, 2000, pp. 1–70.
[11] V. Girault and P.-A. Raviart, *Finite Element Approximation of the Navier-Stokes Equations*, Lecture Notes in Math. 749, Springer-Verlag, Berlin, 1979.
[12] V. Girault and P.-A. Raviart, *Finite Element Methods for Navier-Stokes Equations: Theory and Algorithms*, Springer-Verlag, Berlin, 1986.
[13] M. D. Gunzburger, *Finite Element Methods for Viscous Incompressible Flows: A Guide to Theory, Practice, and Algorithms*, Academic Press, New York, 1988.

[14] S. M. Hammel, J. A. Yorke, and C. Grebogi, *Numerical orbits of chaotic processes represent true orbits*, Bull. Amer. Math. Soc. (N.S.), 19 (1988), pp. 465–469.
[15] T. J. R. Hughes, A. A. Oberai, and L. Mazzei, *Large eddy simulation of turbulent channel flow by the variational multiscale method*, Phys. Fluids, 13 (2001), pp. 1784–1799.
[16] T. Iliescu and P. Fischer, *Backscatter in the rational LES model*, Comput. & Fluids, 33 (2004), pp. 783–790.
[17] V. John and S. Kaya, *A finite element variational multiscale method for the Navier-Stokes equations*, SIAM J. Sci. Comput., 26 (2005), pp. 1485–1503.
[18] A. V. Kolmogorov, *The local structure of turbulence in incompressible viscous fluids for very large Reynolds number*, Dokl. Akad. Nauk. SSR, 30 (1941), pp. 9–13.
[19] W. Layton, *Numerical Analysis of Finite Element Computational Fluid Dynamics*, lecture notes, 2002.
[20] J. Leray, *Sur le mouvemennt d'un fluide visquex remplissant l'espace*, Acta Math., 63 (1934), pp. 193–248.
[21] R. Moser, J. Kim, and N. Mansour, *Direct numerical simulation of turbulent channel flow up to* $Re_\tau = 590$, Phys. Fluids, 11 (1998), pp. 943–945.
[22] A. Ostermann and C. Palencia, *Shadowing for nonautonomous parabolic problems with applications to long-time error bounds*, SIAM J. Numer. Anal., 37 (2001), pp. 1399–1419.
[23] S. Yu Pilyugin, *Shadowing in Dynamical Systems*, Lecture Notes in Math. 1706, Springer-Verlag, Berlin, 1999.
[24] H. Schlichting, *Boundary-Layer Theory*, McGraw-Hill, New York, 1979.
[25] J. Serrin, *The initial value problem for the Navier–Stokes equations*, in Proceedings of the Symposium on Nonlinear Problems, Madison, WI, 1962, pp. 69–98.
[26] H. Sohr, *The Navier-Stokes Equations: An Elementary Functional Analytic Approach*, Birkhäuser Advanced Texts, Birkhäuser-Verlag, Basel, Boston, Berlin, 2001.
[27] X. Wang, *Time averaged energy dissipation rate for shear driven flows in* R^n, Phys. D, 99 (1997), pp. 555–563.

SIAM J. NUMER. ANAL.
Vol. 46, No. 1, pp. 180–200
© 2007 Society for Industrial and Applied Mathematics

RIGOROUS ERROR BOUNDS FOR THE OPTIMAL VALUE IN SEMIDEFINITE PROGRAMMING*

CHRISTIAN JANSSON†, DENIS CHAYKIN‡, AND CHRISTIAN KEIL†

Abstract. A wide variety of problems in global optimization, combinatorial optimization, as well as systems and control theory can be solved by using linear and semidefinite programming. Sometimes, due to the use of floating point arithmetic in combination with ill-conditioning and degeneracy, erroneous results may be produced. The purpose of this article is to show how rigorous error bounds for the optimal value can be computed by carefully postprocessing the output of a linear or semidefinite programming solver. It turns out that in many cases the computational costs for postprocessing are small compared to the effort required by the solver. Numerical results are presented including problems from the SDPLIB and the NETLIB LP library; these libraries contain many ill-conditioned and real-life problems.

Key words. semidefinite programming, linear programming, interval arithmetic, rigorous error bounds, sensitivity analysis, SDPLIB, NETLIB LP library

AMS subject classifications. 90C22, 65G30, 65N15

DOI. 10.1137/050622870

1. Introduction.

We consider the *(primal) semidefinite program* in block diagonal form

$$p^* := \min \sum_{j=1}^{n} \langle C_j, X_j \rangle \quad \text{s.t.} \quad \sum_{j=1}^{n} \langle A_{ij}, X_j \rangle = b_i \quad \text{for } i = 1, \ldots, m, \qquad X_j \succeq 0 \quad \text{for } j = 1, \ldots, n, \tag{1.1}$$

where C_j, A_{ij}, and X_j are symmetric $s_j \times s_j$ matrices, $b \in \mathbb{R}^m$, and

$$\langle C, X \rangle = \text{ trace } (C^T X) \tag{1.2}$$

denotes the *inner product* for the set of symmetric matrices. Moreover, $\succeq$ is the *Löwner partial order*; that is, $X \succeq Y$ iff $X - Y$ is positive semidefinite. In the case $n = 1$ we suppress the index j and write shortly C, X, A_i, and s for the dimension.

If $s_j = 1$ for $j = 1, \ldots, n$ (i.e., C_j, A_{ij}, and X_j are real numbers), then (1.1) defines the standard linear programming problem. Hence, semidefinite programming is an extension of linear programming.

The *Lagrangian dual* of (1.1) is

$$d^* := \max b^T y \quad \text{s.t.} \quad \sum_{i=1}^{m} y_i A_{ij} \preceq C_j \quad \text{for } j = 1, \ldots, n, \tag{1.3}$$

where $y \in \mathbb{R}^m$. The constraints $\sum_{i=1}^{m} y_i A_{ij} \preceq C_j$ are called *linear matrix inequalities (LMIs)*. We use the convention that $p^* = -\infty$ if (1.1) is unbounded and $p^* = \infty$ if (1.1) is infeasible. The analogous convention is used for (1.3).

*Received by the editors January 18, 2005; accepted for publication (in revised form) June 13, 2007; published electronically December 20, 2007.
http://www.siam.org/journals/sinum/46-1/62287.html

†Institute for Reliable Computing, Hamburg University of Technology, Schwarzenbergstraße 95, 21071 Hamburg, Germany (c.jansson@tu-harburg.de, c.keil@tu-harburg.de).

‡Institute of Chemical Reaction Engineering, Hamburg University of Technology, Eissendorfer Straße 38, 21071 Hamburg, Germany (denis.chaykin@tu-harburg.de).

The duality theory is similar to linear programming, but more subtle. The programs satisfy the *weak duality* condition

$$d^* \le p^*, \tag{1.4}$$

but strong duality requires, in contrast to linear programming, additional conditions (see, for example, Nemirovski [32], Ramana, Tunçel, and Wolkowicz [39], and Vandenberghe and Boyd [49]).

THEOREM 1.1 (strong duality theorem).

(a) *If* (1.1) *is strictly feasible (i.e., there exist feasible positive definite matrices* X_j *for* $j = 1, \dots, n$*) and* p^* *is finite, then* $p^* = d^*$ *and the dual supremum is attained.*

(b) *If* (1.3) *is strictly feasible (i.e., there exists some* $y \in \mathbb{R}^m$ *such that* $C_j - \sum_{i=1}^m y_i A_{ij}$ *are positive definite for* $j = 1, \dots, n$*) and* d^* *is finite, then* $p^* = d^*$ *and the primal infimum is attained.*

In general, one of the problems (1.1) and (1.3) may have optimal solutions while its dual is infeasible, or the duality gap may be positive at optimality. The strict feasibility assumptions in Theorem 1.1 are called *Slater's constraint qualifications.*

Semidefinite programming and LMI methods are documented by many applications and a number of survey papers (see, for example, Skelton and Iwasaki [45], Balakrishnan and Feron [3], and Vandenberghe and Boyd [49]). Applications include global optimization problems, optimal state space realizations, robust controller design, integer programming problems, as well as eigenvalue problems in the form of minimizing the largest, or minimizing the sum of the first few largest eigenvalues of a symmetric matrix X subject to linear constraints on X.

Semidefinite programs can be solved in polynomial time if an a priori bound for the size of their solution is known (see Grötschel, Lovász, and Schrijver [12]). This is a consequence of the ellipsoid method for convex programming. The ellipsoid method has not proven practical, and interior point methods turned out to be the methods of choice in semidefinite programming.

Conventionally, algorithms assume that the input data are given exactly, and they use floating point arithmetic for computing an approximate solution. Occasionally, wrong results may be produced, not solely but especially for ill-conditioned and ill-posed problems in the sense defined by Renegar [40]. He defines the condition number as the scale-invariant reciprocal of the smallest data perturbation that will render the perturbed data instance either primal or dual infeasible. It is set to ∞ if the distance to primal or dual infeasibility is 0, and in this case the problem is called ill-posed. Examples where commercial solvers fail to solve linear optimization problems can be found in Neumaier and Shcherbina [37] and in [17]. It cannot be answered how frequently such failures occur. Ill-conditioning is, however, frequently observed. In a paper by Ordóñez and Freund [38] it is stated that 71% of the LP instances in the NETLIB Linear Programming Library [33] are ill-posed; i.e., the problems have an infinite condition number. Recently, Freund, Ordóñez, and Toh [9] solved 85 out of the 92 problems of the SDPLIB [5] with SDPT3 [47] and investigated the interior-point iteration counts with respect to different measures for semidefinite programming problems. They omitted the four infeasible problems and three very large problems where SDPT3 ran out of memory. Of the remaining 85 problems they have shown 32 to be ill-posed.

As pointed out in Neumaier and Shcherbina [37], ill-conditioning is also likely to take place in combinatorial optimization when branch-and-cut procedures sequentially

generate linear or semidefinite programming relaxations. Therefore, the computation of rigorous error bounds, which take account of all rounding errors and of small errors in the input data, is valuable in practice.

The primary purpose of this paper is to show that by properly postprocessing the output of a semidefinite or linear solver, rigorous error bounds for the optimal value can be obtained. Moreover, existence of optimal solutions can be proved, or a certificate of infeasibility can be given. The input data are allowed to vary within small intervals. Our numerical experience with the NETLIB LP library and the SDPLIB demonstrates that, roughly speaking, rigorous lower and upper error bounds for the optimal value are computed even for ill-conditioned and degenerate problems. The quality of the error bounds depends on the quality of the computed approximations and the distances to dual and primal infeasibility. By comparing these bounds, one knows whether the computed results are good.

The presented results can be viewed as a further development of similar methods for linear programming (Neumaier and Shcherbina [37] and Jansson [17]) and convex programming [16].

The paper is organized as follows. Section 2 contains notation. In section 3 an algorithm for computing a rigorous lower bound of the global minimum value is considered, and in section 4 a rigorous upper bound of the optimal value, together with a certificate of existence of optimal solutions, is presented. In section 5 we show how these rigorous bounds can be used for obtaining certificates of infeasibility. Section 6 contains numerical results and some remarks on other software packages. Finally, in section 7 some conclusions are given.

2. Notation, interval arithmetic. Throughout this paper we use the following notation. We denote by $\mathbb{R}$, $\mathbb{R}^n$, $\mathbb{R}^n_+$, and $\mathbb{R}^{m\times n}$ the sets of real numbers, real vectors, real nonnegative vectors, and real $m \times n$ matrices, respectively. Comparisons $\leq$, absolute value $|\cdot|$, min, max, inf, and sup are used entrywise for vectors and matrices. The identity matrix is denoted by I_d.

For a symmetric matrix A the eigenvalues are sorted nonincreasingly: $\lambda_{\max}(A) = \lambda_1(A) \geq \lambda_2(A) \geq \cdots \geq \lambda_{\min}(A)$.

For $\mu \in \mathbb{R}$ the operator

$$\operatorname{svec}(A, \mu) := (A_{11}, \mu A_{21}, \ldots, \mu A_{n1}, A_{22}, \mu A_{32}, \ldots, \mu A_{nn-1}, A_{nn})^T \tag{2.1}$$

transforms symmetric $n \times n$ matrices into $((n+1)n/2)$-dimensional vectors with the property that the inner product of two symmetric matrices A, B is

$$\langle A, B\rangle = \operatorname{svec}(A, 2)^T \operatorname{svec}(B, 1) = \operatorname{svec}(A, \sqrt{2})^T \operatorname{svec}(B, \sqrt{2}), \tag{2.2}$$

and $\operatorname{svec}(A, \sqrt{2})$ is the customary svec operator. We prefer the first representation of the inner product, since this avoids conversion errors of the input data of semidefinite programs in its vector representation form. The inverse operator of svec is denoted by $\operatorname{smat}(a, \mu)$ where a is the vector representation (2.1).

For block matrices with blocks A_j for $j = 1, \ldots, n$, we define the concatenated vector

$$\operatorname{svec}((A_j), \mu) := (\operatorname{svec}(A_1, \mu); \ldots; \operatorname{svec}(A_n, \mu)). \tag{2.3}$$

We require only some elementary facts about interval arithmetic, which are described here. There are a number of textbooks on interval arithmetic and self-validating

methods that we highly recommend to readers. These include Alefeld and Herzberger [1], Moore [31], and Neumaier [34, 35].

If $\mathbb{V}$ is one of the spaces $\mathbb{R}$, $\mathbb{R}^n$, or $\mathbb{R}^{m\times n}$, and $\underline{v}, \overline{v} \in \mathbb{V}$, then the box

$$\mathbf{v} := [\underline{v}, \overline{v}] := \{v \in \mathbb{V} : \ \underline{v} \le v \le \overline{v}\} \tag{2.4}$$

is called an *interval quantity* in $\mathbb{IV}$ with *lower bound* $\underline{v}$ and *upper bound* $\overline{v}$. In particular, $\mathbb{IR}$, $\mathbb{IR}^n$, and $\mathbb{IR}^{m\times n}$ denote the set of real intervals $\mathbf{a} = [\underline{a}, \overline{a}]$, the set of real interval vectors $\mathbf{x} = [\underline{x}, \overline{x}]$, and the set of real interval matrices $\mathbf{A} = [\underline{A}, \overline{A}]$, respectively. The real operations $A \circ B$ with $\circ \in \{+, -, \cdot, /\}$ between real numbers, real vectors, and real matrices can be generalized to *interval operations*. The result $\mathbf{A} \circ \mathbf{B}$ of an interval operation is defined as the interval hull of all possible real results, that is,

$$\mathbf{A} \circ \mathbf{B} := \cap\{\mathbf{C} \in \mathbb{IV} : \ A \circ B \in \mathbf{C} \quad \text{for all} \quad A \in \mathbf{A}, B \in \mathbf{B}\}. \tag{2.5}$$

All interval operations can be easily executed by working appropriately with the lower and upper bounds of the interval quantities. For example, in the simple case of addition, we obtain

$$\mathbf{A} + \mathbf{B} = [\underline{A} + \underline{B}, \overline{A} + \overline{B}]. \tag{2.6}$$

Interval multiplications and divisions require a distinction of cases. Similarly all operations (2.5) between interval vectors and interval matrices can be executed. For example the i, j component of the product of two interval matrices $\mathbf{C}, \mathbf{X} \in \mathbb{IR}^{n\times n}$ is

$$(\mathbf{CX})_{ij} = \sum_{k=1}^{n} \mathbf{C}_{ik}\mathbf{X}_{kj}, \tag{2.7}$$

and the inner product is

$$\langle \mathbf{C}, \mathbf{X} \rangle = \ \text{trace}\ (\mathbf{C}^T\mathbf{X}) = \sum_{i,j=1}^{n} \mathbf{C}_{ij}\mathbf{X}_{ij}. \tag{2.8}$$

For interval quantities $\mathbf{A}, \mathbf{B} \in \mathbb{IV}$ we define

$$\text{mid}\mathbf{A} := (\underline{A} + \overline{A})/2 \quad \text{as the } \textit{midpoint}, \tag{2.9}$$

$$\text{rad}\mathbf{A} := (\overline{A} - \underline{A})/2 \quad \text{as the } \textit{radius}, \tag{2.10}$$

$$|\mathbf{A}| := \max\{|A| : \ A \in \mathbf{A}\} \quad \text{as the } \textit{absolute value}, \tag{2.11}$$

$$\mathbf{A}^+ := \max\{0, \overline{A}\}, \tag{2.12}$$

$$\mathbf{A}^- := \min\{0, \underline{A}\}. \tag{2.13}$$

Moreover, the comparison in $\mathbb{IV}$ is defined by

$$\mathbf{A} \le \mathbf{B} \quad \text{iff} \quad \overline{A} \le \underline{B},$$

and other relations are defined analogously. Real quantities v are embedded in the interval quantities by identifying $v = \mathbf{v} = [v, v]$.

We call $\mathbf{A} \in \mathbb{IR}^{n\times n}$ *symmetric* if $\mathbf{A}_{ij} = \mathbf{A}_{ji}$ for all i, j, and $\mathbf{A}$ is called positive semidefinite if all $A \in \mathbf{A}$ have this property.

For linear systems of equations with inexact input data, the aim frequently is to compute an interval vector $\mathbf{x} \in \mathbb{IR}^n$ containing the *solution set*

$$\Sigma(\mathbf{A}, \mathbf{b}) := \{x \in \mathbb{R}^n : Ax = b \text{ for some } (A, b) \in (\mathbf{A}, \mathbf{b})\}, \tag{2.14}$$

where $\mathbf{A} \in \mathbb{IR}^{n\times n}$ and $\mathbf{b} \in \mathbb{IR}^n$. This is an NP-hard problem, but there are several methods that compute enclosures $\mathbf{x}$. A precise description of such methods, required assumptions, and approximation properties can be found, for example, in Neumaier [34]. Roughly speaking, it turns out that for interval matrices with $\|I_d - R\mathbf{A}\| < 1$ (R is an approximate inverse of the midpoint mid$\mathbf{A}$), there are several methods which compute an enclosure $\mathbf{x}$ with $O(n^3)$ operations. The radius rad$\mathbf{x}$ decreases linearly with decreasing radii rad$\mathbf{A}$ and rad$\mathbf{b}$. For the computation of enclosures in the case of large-scale linear systems the reader is referred to Rump [42].

In interval arithmetic several methods for computing rigorous bounds for all or some eigenvalues of interval matrices were developed. Some important references are Floudas [7], Mayer [30], Neumaier [36], and Rump [42, 43].

3. Rigorous lower bound. In many applications some or all input data are uncertain. We model these uncertainties by intervals. In the case of semidefinite programming we assume that symmetric interval matrices $\mathbf{C}_j, \mathbf{A}_{ij} \in \mathbb{IR}^{s_j\times s_j}$, $i = 1, \ldots, m$, $j = 1, \ldots, n$, and an interval vector $\mathbf{b} \in \mathbb{IR}^m$ are given. This yields a family of semidefinite programs (1.1), where the input data $P = (A, b, C)$ are allowed to vary within interval bounds $\mathbf{P} := (\mathbf{A}, \mathbf{b}, \mathbf{C})$.

In order to indicate the dependency on the input data, we sometimes write $p^*(P)$, $d^*(P)$, $X^*(P)$, etc.

First, we state a lemma proving a lower bound for the inner product of two symmetric matrices.

LEMMA 3.1. *Let D, X be symmetric matrices of dimension s that satisfy*

$$\underline{d} \le \lambda_{\min}(D), \quad 0 \le \lambda_{\min}(X), \quad \textit{and} \quad \lambda_{\max}(X) \le \overline{x}. \tag{3.1}$$

Then

$$\langle D, X\rangle \ge s \cdot \underline{d}^- \cdot \overline{x}, \tag{3.2}$$

where $\underline{d}^- := \min\{0, \underline{d}\}$.

Proof. Let D have the eigenvalue decomposition

$$D = Q\Lambda(D)Q^T, \quad QQ^T = I_d,$$

where $\Lambda(D)$ is the diagonal matrix with eigenvalues of D on the diagonal. Then

$$\begin{aligned}\langle D, X\rangle &= \operatorname{trace}(Q\Lambda(D)Q^T X)\\ &= \operatorname{trace}(\Lambda(D)Q^T XQ)\\ &= \sum_{k=1}^{s} \lambda_k(D)Q(:,k)^T XQ(:,k).\end{aligned}$$

Because of (3.1), we have $0 \le Q(:,k)^T XQ(:,k) \le \overline{x}$ yielding

$$\langle D, X\rangle \ge \sum_{k=1}^{s} \lambda_k(D)^- \cdot \overline{x} \ge s \cdot \underline{d}^- \cdot \overline{x}. \qquad \square$$

We are now ready to prove a rigorous lower bound for the optimal value p^*.

THEOREM 3.2. *Let* $\mathbf{P}$ *define a family of semidefinite programs* (1.1) *with input data* $P \in \mathbf{P}$, *let* $\tilde{y} \in \mathbb{R}^m$, *set*

$$\mathbf{D}_j := \mathbf{C}_j - \sum_{i=1}^{m} \tilde{y}_i \mathbf{A}_{ij} \quad \text{for } j = 1, \ldots, n, \tag{3.3}$$

and suppose that

$$\underline{d}_j \le \lambda_{\min}(\mathbf{D}_j) \quad \text{for } j = 1, \ldots, n. \tag{3.4}$$

Assume further that upper bounds for the maximal eigenvalues of the primal feasible solutions of (1.1)

$$\lambda_{\max}(X_j) \le \overline{x}_j \quad \text{for } j = 1, \ldots, n \tag{3.5}$$

are known, where $\overline{x}_j$ *may be infinite. If*

$$\underline{d}_j \ge 0 \quad \text{for } \overline{x}_j = +\infty, \tag{3.6}$$

then for every $P \in \mathbf{P}$ *the inequality*

$$p^*(P) \ge \inf \left\{ \mathbf{b}^T \tilde{y} + \sum_{j=1}^{n} s_j \cdot \underline{d}_j^- \cdot \overline{x}_j \right\} \tag{3.7}$$

is satisfied, and the right-hand side of (3.7) *is finite.*[1] *Moreover, for every* $P \in \mathbf{P}$ *and every* j *with* $\underline{d}_j \ge 0$ *the LMI*

$$\sum_{i=1}^{m} y_i A_{ij} - C_j \preceq 0$$

is feasible with $y := \tilde{y}$.

Proof. Let $P = (A, b, C) \in \mathbf{P}$ be chosen fixed, and let $X_j = X_j(P)$ be primal feasible for P and $j = 1, \ldots, n$. Let

$$D_j = C_j - \sum_{i=1}^{n} \tilde{y}_i A_{ij} \quad \text{for } j = 1, \ldots, n;$$

then

$$\sum_{j=1}^{n} \langle C_j, X_j \rangle - b^T \tilde{y} = \sum_{j=1}^{n} \langle D_j, X_j \rangle.$$

Since $D_j \in \mathbf{D}_j$, Lemma 3.1 implies

$$\sum_{j=1}^{n} \langle D_j, X_j \rangle \ge \sum_{j=1}^{n} s_j \cdot \underline{d}_j^- \cdot \overline{x}_j,$$

[1] Notice that $\mathbf{b}^T y$ is an interval operation yielding an interval for the expression in the braces in (3.7). Hence, the infimum denotes the lower bound of this interval. This notation applies also for the supremum and subsequently.

which proves the inequality (3.7), and the assumption (3.6) yields a finite right-hand side. The last statement is an immediate consequence of $D_j \in \mathbf{D}_j$ and $\lambda_{\min}(D_j) \geq \underline{d}_j \geq 0$. ∎

Observe that $\tilde{y}$ is dual feasible provided $\underline{d}_j \geq 0$ for $j = 1, \ldots, n$. Hence in this case, (3.7) yields the lower bound $\inf\{\mathbf{b}^T \tilde{y}\}$ for the dual optimal value $d^*(P)$ for every $P \in \mathbf{P}$.

In order to judge the quality of the lower bound (3.7), we assume that

(i) exact input data $P = \mathbf{P}$ are given,
(ii) $D = \mathbf{D}$ is computed exactly, and
(iii) Slater's constraint qualifications (cf. section 1) are fulfilled.

Moreover, let $\tilde{y}$ be the optimal solution of the dual problem (1.2), and let $\underline{d}_j = \lambda_{\min}(D)$ for $j = 1, \ldots, n$. Then $\underline{d}_j \geq 0$ for $j = 1, \ldots, n$, and

$$p^*(P) = d^*(P) = b^T \tilde{y}.$$

Hence, no overestimation occurs, and it follows that the quality of this lower bound mainly depends on the quality of the $\underline{d}_j$ and on the computed approximation $\tilde{y}$.

An immediate consequence of Theorem 3.2 is the following error bound for linear programming problems:

$$p^* := \min c^T x \quad \text{s.t. } Ax = b, x \geq 0, \tag{3.8}$$

which is proved in [17], and in [37] for finite bounds $\overline{x}_j$. The input data are $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, $c \in \mathbb{R}^n$, and $P = (A, b, c) \in \mathbb{R}^{m \times n + m + n}$.

COROLLARY 3.1. *Let* $\mathbf{P} = (\mathbf{A}, \mathbf{b}, \mathbf{c}) \in \mathbb{IR}^{m \times n + m + n}$, $\tilde{y} \in \mathbb{R}^m$, *and let*

$$\mathbf{d} := \mathbf{c} - \mathbf{A}^T \tilde{y}. \tag{3.9}$$

Assume further that upper bounds of the primal feasible solutions

$$x_j \leq \overline{x}_j \quad \textit{for } j = 1, \ldots, n$$

are known for all $P \in \mathbf{P}$, *which may also be infinite. If*

$$\mathbf{d}_j \geq 0 \quad \textit{for } \overline{x}_j = +\infty, \tag{3.10}$$

then for every $P \in \mathbf{P}$ *the optimal value* $p^*(P)$ *satisfies the inequality*

$$p^*(P) \geq \inf \left\{ \mathbf{b}^T \tilde{y} + \sum_{j=1}^{n} \mathbf{d}_j^- \cdot \overline{x}_j \right\}. \tag{3.11}$$

Proof. Apply Theorem 3.2 to the semidefinite program where the symmetric matrices A_{ij}, C_j, and X_j are one dimensional. ∎

Next, we describe an algorithm for computing a lower bound of the optimal value, which is based on Theorem 3.2. We assume that an approximate dual optimal solution $\tilde{y} \in \mathbb{R}^m$ of the midpoint problem $\operatorname{mid} \mathbf{P}$ is known. If condition (3.6) is fulfilled, the only work is to compute the right-hand side of (3.7). Otherwise, the idea is to perturb all constraints which violate condition (3.6); that is, we solve a perturbed midpoint problem $P = (\operatorname{mid} \mathbf{A}, \operatorname{mid} \mathbf{b}, C(\varepsilon))$ with

$$C_j(\varepsilon) = \operatorname{mid} \mathbf{C}_j - \varepsilon_j I_d, \varepsilon_j = \begin{cases} > 0 & \text{if } \underline{d}_j < 0 \text{ and } \overline{x}_j = +\infty, \\ 0 & \text{otherwise.} \end{cases} \tag{3.12}$$

Then the dual optimal solution $y(\varepsilon)$ satisfies the constraints

$$\operatorname{mid} \mathbf{C}_j - \sum_{i=1}^{m} y_i(\varepsilon) \operatorname{mid} \mathbf{A}_{ij} \succeq \varepsilon_j I_d.$$

Hence, the minimal eigenvalues of the new defect

$$\mathbf{D}_j(\varepsilon) := \mathbf{C}_j - \sum_{i=1}^{m} y_i(\varepsilon) \mathbf{A}_{ij}$$

will increase. Choosing ε_j very large may imply dual infeasibility, choosing $\varepsilon_j > 0$ too small may not be sufficient for satisfying (3.6). Our current trade-off is to solve repeatedly perturbed programs until either condition (3.6) is satisfied or the dual is infeasible. The details are given in Algorithm 1. This algorithm requires interval arithmetic (or at least the monotonic rounding operations) for computing the defect matrices $\mathbf{D}$, an algorithm for computing rigorous lower bounds $\underline{d}_j$, and a semidefinite solver for computing approximate solutions of the perturbed problems.

The algorithm terminates during the first iteration in step 3 if all simple bounds $\overline{x}_j$ are finite or all $\underline{d}_j$ are nonnegative. In this case the computational costs are $O(m \cdot \sum_{j=1}^{n} s_j^2)$ for computing the $\mathbf{D}_j$'s, the lower bounds $\underline{d}_j$ require $O(\sum_{j=1}^{n} s_j^3)$ operations, and the bound $\underline{p}^*$ needs $O(m+n)$ operations. Hence the costs are negligible compared to the costs for approximately solving a semidefinite program.

In other cases, however, the computational costs may increase because perturbed semidefinite programs must be solved until either condition (3.6) is satisfied, the semidefinite programming solver indicates dual infeasibility of the perturbed problem, or the maximal number of iterations $l_{\max}$ is reached.

Several modifications of this algorithm are possible and may yield improvements. Here we have considered a simple choice of perturbations: In each step we add to ε_j the negative defects $-\underline{d}_j$ multiplied by a factor 2^{k_j}, where k_j counts the number of iterations that violated the inequality $\underline{d}_j \geq 0$.

In applications we recommend to use infinite bounds $\overline{x}_j$ instead of unreasonable large bounds, because otherwise the sum in (3.7) may yield an unnecessary overestimation.

If the upper bounds $\overline{x}_j = +\infty$ for $j = 1, \ldots, n$ and Algorithm 1 delivers a finite lower bound $\underline{p}^*$, then the lower eigenvalue bounds $\underline{d}_j$ must be nonnegative. This proves dual feasibility, and if $\underline{d}_j$ is positive for $j = 1, \ldots, n$, strict dual feasibility is verified.

4. Rigorous upper bound. In this section we investigate the computation of a rigorous upper bound for the optimal value of a semidefinite program together with a certificate of existence of primal feasible solutions. The basic idea is to compute interval matrices $\mathbf{X}_j$ for $j = 1, \ldots, n$ that contain a primal feasible solution for every semidefinite program $P \in \mathbf{P}$. The desirable characteristics of the matrices $\mathbf{X}_j$ are given in the next theorem.

THEOREM 4.1. *Let* $\mathbf{P}$ *define a family of semidefinite programs* (1.1), *and suppose that there exist interval matrices* $\mathbf{X}_j$ *for* $j = 1, \ldots, n$ *such that*

$$\begin{aligned} &\forall b \in \mathbf{b}, \forall A_{ij} \in \mathbf{A}_{ij}, i = 1, \ldots, m, j = 1, \ldots, n, \\ &\exists \text{ symmetric } X_j \in \mathbf{X}_j : \sum_{j=1}^{n} \langle A_{ij}, X_j \rangle = b_i, \end{aligned} \tag{4.1}$$

ALGORITHM 1. *Rigorous lower bound*

given: real or interval input data $\mathbf{P} = (\mathbf{A}, \mathbf{b}, \mathbf{c})$,
upper bounds $\overline{x}_j$ for $j = 1, \ldots, n$,
approximate dual optimal solution $\tilde{y}$ for $\operatorname{mid} \mathbf{P}$,
$\underline{p}^* := -\infty$,
ε, k are n-dimensional zero vectors,
maximal numbers of iterations $l_{\max}$,
$l := 0$.

while perturbed problem $P(\varepsilon)$ is dual feasible **and** $l \le l_{\max}$

1. Compute $\mathbf{D}_j = \mathbf{C}_j - \sum_{i=1}^{m} \tilde{y}_i \mathbf{A}_{ij}$, $j = 1, \ldots, n$.
2. Compute rigorous lower bounds $\underline{d}_j \le \lambda_{\min}(\mathbf{D}_j)$ for $j = 1, \ldots, n$.
3. **If** $\underline{d}_j \ge 0$ for every j with $\overline{x}_j = +\infty$ **then** compute

$$\underline{p}^* = \inf \left\{ \mathbf{b}^T \tilde{y} + \sum_{j=1}^{n} s_j \cdot \underline{d}_j^- \cdot \overline{x}_j \right\},$$

STOP.

4. Compute for $j = 1, \ldots, n$

$$k_j := \begin{cases} k_j + 1 & \text{if } \underline{d}_j < 0 \text{ and } \overline{x}_j = +\infty, \\ k_j & \text{otherwise,} \end{cases}$$

$$\varepsilon_j := \begin{cases} -2^{k_j} \underline{d}_j + \varepsilon_j & \text{if } \underline{d}_j < 0 \text{ and } \overline{x}_j = +\infty, \\ \varepsilon_j & \text{otherwise.} \end{cases}$$

5. Solve the perturbed midpoint problem $P(\varepsilon) = (\operatorname{mid} \mathbf{A}, \operatorname{mid} \mathbf{b}, C(\varepsilon))$, where $C_j(\varepsilon) = \operatorname{mid} \mathbf{C}_j - \varepsilon_j I_d$ for $j = 1, \ldots, n$, and set $\tilde{y} := \tilde{y}(\varepsilon)$ (approximate dual optimal solution).
6. $l := l + 1$.

end

and for $j = 1, \ldots, n$

$$X_j \succeq 0 \quad \textit{for all symmetric } X_j \in \mathbf{X}_j. \tag{4.2}$$

Then, the optimal value is bounded from above by

$$p^*(P) \le \sup \left\{ \sum_{j=1}^{n} \langle \mathbf{C}_j, \mathbf{X}_j \rangle \right\}. \tag{4.3}$$

Moreover, if all symmetric $X_j \in \mathbf{X}_j$ *are positive definite and* $p^*(P)$ *is bounded from below, then* $p^*(P) = d^*(P)$ *for every* $P \in \mathbf{P}$ *(no duality gap) and the dual supremum is attained.*

Proof. Let $P \in \mathbf{P}$ be a fixed chosen problem. Then the conditions (4.1) and (4.2) imply that there exists a primal feasible solution $X_j = X_j(P)$ for $j = 1, \ldots, n$. Hence, $\sum_{j=1}^{n} \langle C_j, X_j \rangle \ge p^*(P)$, and the inclusion property (2.5) yields (4.3). If all $X_j \in \mathbf{X}_j$

are positive definite, then (4.1) and (4.2) imply the existence of strictly primal feasible solutions, and hence Theorem 1.1 shows that the dual optimal solution is attained and strong duality holds. □

By weak duality the upper bound in (4.3) is also an upper bound of the dual optimal value. Moreover, if all $X_j \in \mathbf{X}_j$ are positive definite, then the strong duality theorem (Theorem 1.1) implies that the right-hand side of (3.7) is also a lower bound of the dual optimal value for all $P \in \mathbf{P}$. Hence, in this case it is not necessary to assume $\underline{d}_j \geq 0$ for $j = 1, \ldots, n$.

In the following, we describe an algorithm for computing this rigorous upper bound. This algorithm must find appropriate interval matrices $\mathbf{X}_j$ and verify the conditions (4.1) and (4.2). We discuss these items below.

To make sure that the upper bound (4.3) is close to the optimal value, the interval matrices $\mathbf{X}_j$ must be close to optimality. In general, the complementary slackness relations yield rank-deficient matrices, which are not positive definite. Therefore, we solve the slightly perturbed midpoint problem

$$\min \sum_{j=1}^{n} \langle C_j, X_j \rangle \quad \text{s.t.} \quad \sum_{j=1}^{n} \langle A_{ij}, X_j \rangle = b_i \quad \text{for } i = 1, \ldots, m, \tag{4.4}$$
$$X_j \succeq \varepsilon_j \cdot I_d \quad \text{for } j = 1, \ldots, n,$$

where ε_j is positive and the input data $(A, b, c) = \operatorname{mid} \mathbf{P}$. Then for small ε_j the optimal solution $(X_j(\varepsilon_j))$ is positive definite and close to the optimal solution of the midpoint problem.

In the following we show how we can construct an appropriate interval matrix $(\mathbf{X}_j)$ by using an approximate optimal solution $(X_j(\varepsilon_j))$ of (4.4).

The semidefinite program (1.1) can be written in the equivalent vector representation form,

$$\min c^T x \quad \text{s.t.} \quad A^{\text{mat}} x = b, X_j \succeq 0 \text{ for } j = 1, \ldots, n, \tag{4.5}$$

where

$$c := \operatorname{svec}((C_j), 2), \tag{4.6}$$

$$x := \operatorname{svec}((X_j), 1), \tag{4.7}$$

and the ith row of the $m \times \sum_{j=1}^{n} \frac{s_j(s_j+1)}{2}$ matrix A^{mat} is defined by

$$A^{\text{mat}}(i, :) = \operatorname{svec}((A_{ij})_{j=1}^{n}, 2). \tag{4.8}$$

If interval input data $\mathbf{P}$ are given, then we denote by $\mathbf{A}^{\text{mat}}$, $\mathbf{b}$, and $\mathbf{c}$ the corresponding interval quantities. Thus condition (4.1) is equivalent to

$$\forall b \in \mathbf{b}, \forall A^{\text{mat}} \in \mathbf{A}^{\text{mat}} \exists x \in \mathbf{x} \text{ such that } A^{\text{mat}} x = b, \tag{4.9}$$

which is an underdetermined system of linear equations with interval input data. Given an approximate optimal solution $(X_j(\varepsilon_j))_{j=1}^{n}$, it is straightforward to solve such a system.

We start by assuming that the $m \times m$ submatrix $\operatorname{mid} \mathbf{A}_I^{\text{mat}}$ with the m columns $\operatorname{mid} \mathbf{A}^{\text{mat}}(:, \beta_i)$ is nonsingular, where the index set $I := \{\beta_1, \ldots, \beta_m\}$. Let N denote

all indices of columns of $\operatorname{mid}\mathbf{A}^{\mathrm{mat}}$ which are not in I, let $\mathbf{A}_N^{\mathrm{mat}}$ be the matrix with columns corresponding to the indices of N, and let $\tilde{x} = \operatorname{svec}((X_j(\varepsilon_j)), 1)$. In our algorithm we choose the index set I by performing an LU decomposition on $(\operatorname{mid}\mathbf{A}^{\mathrm{mat}})^T$ and assembling the computed pivot columns to A_I^{mat}. Now we fix the variables $\tilde{x}_N$ and compute with some verification method for interval linear systems an enclosure $\mathbf{x}_I$ of the solution set

$$\Sigma_I := \left\{ x_I \in \mathbb{R}^m : A_I^{\mathrm{mat}} x_I = b - \sum_{\gamma \in N} A_N^{\mathrm{mat}} \tilde{x}_N, A \in \mathbf{A}^{\mathrm{mat}}, b \in \mathbf{b} \right\}. \tag{4.10}$$

Then $\mathbf{x} := (\mathbf{x}_I; \tilde{x}_N)$ fulfills (4.9), and therefore $(\mathbf{X}_j) := \operatorname{smat}(\mathbf{x}, 1)$ satisfies condition (4.1). Condition (4.2) must be verified by some method for computing a rigorous lower bound for the smallest eigenvalue of a symmetric interval matrix.

Algorithm 2 contains the details for computing a rigorous upper bound for the optimal value and for proving existence of primal feasible solutions. The algorithm needs verified solvers for interval linear systems and eigenvalue problems, and a semidefinite solver for computing approximations of the perturbed problems.

If Algorithm 2 delivers a finite upper bound $\overline{p}^*$, then the lower eigenvalue bounds $\underline{\lambda}_j$ must be nonnegative. If $\underline{\lambda}_j > 0$ for $j = 1, \ldots, n$, then strict primal feasibility is verified.

Krawczyk [27] was the first who solved nondegenerate interval linear programming problems by using the technique of fixing appropriate variables (the nonbasic variables) and solving a remaining quadratic interval linear system for the basic variables. In [15] this technique was used to compute enclosures of all optimal vertices in the degenerate case. Hansen [13] used this technique in order to prove existence of a feasible point for nonlinear equations within a bounded box. It was further modified and investigated numerically by Kearfott [20], and is also described in his book [19].

5. Certificate of infeasibility. In branch and bound algorithms a subproblem is discarded if the local nonlinear solver detects infeasibility. Occasionally local solvers do not find feasible solutions of a subproblem, although they exist (see, for example, the comments for use of SDPT3). A consequence is that the global minimum solutions may be cut off.

To avoid this disadvantage we can apply the algorithms for computing rigorous bounds described in the previous sections to a phase I problem in order to verify infeasibility for primal and dual semidefinite problems. In the literature there are several variations of the phase I method. It is common, however, that the auxiliary objective function describes the infeasibility in the sense that the problem has no feasible solutions, provided the optimal value is greater than zero. The latter property can be verified by the algorithms of the previous sections.

Another approach is based on certificates of infeasibility. For linear programs with bounded variables rigorous certificates of infeasibility are described in Neumaier and Shcherbina [37]. For infeasible semidefinite problems often (but not every time) certificates of infeasibility exposed by improving rays can be obtained (see also the discussion in Todd [46]).

The primal problem (1.1) has a primal improving ray if there exists a block-diagonal matrix (X_j) such that for all $i = 1, \ldots, m$ and $j = 1, \ldots, n$

$$X_j \succeq 0, \quad \sum_{j=1}^{n} \langle A_{ij}, X_j \rangle = 0, \quad \text{and} \quad \sum_{j=1}^{n} \langle C_j, X_j \rangle < 0. \tag{5.1}$$

ALGORITHM 2. *Rigorous upper bound, certificate of feasibility*

given: real or interval input data $\mathbf{P} = (\mathbf{A}, \mathbf{b}, \mathbf{c})$,
approximate primal optimal solution $(\tilde{X}_j)_{j=1}^n$ of the midpoint problem,
$\overline{p}^* := \infty$,
ε, k are n-dimensional zero vectors,
maximal number of iterations $l_{\max}$,
$l := 0$.

Choose an index set I such that the submatrix $\operatorname{mid} \mathbf{A}_I^{\mathrm{mat}}$ is (at least numerically) nonsingular (for example, by performing an LU decomposition on $(\operatorname{mid} \mathbf{A}^{\mathrm{mat}})^T$).

if there is no nonsingular submatrix **then STOP**.

while perturbed problem $P(\varepsilon)$ is primal feasible **and** $l \le l_{\max}$

1. Compute an enclosure $\mathbf{x}_I$ of the solution set Σ_I (4.10), and set $\mathbf{x} := (\mathbf{x}_I; \tilde{x}_N)$.
2. Set $(\mathbf{X}_j) = \operatorname{smat}(\mathbf{x}, 1)$, and compute rigorous bounds
$$\underline{\lambda}_j \le \lambda_{\min}(\mathbf{X}_j) \quad \text{for } j = 1, \ldots, n.$$
3. **if** $\underline{\lambda}_j \ge 0$ for $j = 1, \ldots, n$ **then** compute
$$\overline{p}^* = \sup\{\mathbf{c}^T \mathbf{x}\},$$
STOP.
4. Compute for $j = 1, \ldots, n$
$$k_j := \begin{cases} k_j + 1 & \text{if } \underline{\lambda}_j < 0, \\ k_j & \text{otherwise,} \end{cases}$$
$$\varepsilon_j := \begin{cases} -2^{k_j} \underline{\lambda}_j + \varepsilon_j & \text{if } \underline{\lambda}_j < 0, \\ \varepsilon_j & \text{otherwise.} \end{cases}$$
5. Solve the perturbed problem (4.4), set $\tilde{X}_j := \tilde{X}_j(\varepsilon)$ for $j = 1, \ldots, n$ (approximate primal optimal solution), and set $\tilde{x} := \operatorname{svec}((\tilde{X}_j), 1)$.
6. $l := l + 1$.

end

It is well known and straightforward to show that the existence of a primal improving ray implies dual infeasibility. If interval input data $\mathbf{P}$ are given, and for the midpoint problem of $\mathbf{P}$ an approximate primal improving ray is known, then we can try to verify dual infeasibility for all problems with $P \in \mathbf{P}$ by using a similar approach as in the previous section. We assume that the semidefinite solver has computed an approximate primal improving ray $(\tilde{X}_j)$ for the midpoint problem. Let $\beta \approx \sum_{j=1}^n \langle \operatorname{mid} \mathbf{C}_j, \tilde{X}_j \rangle$ be approximately calculated and assume that $\beta < 0$. Notice that for positive β the conditions (5.1) are in general not satisfied. Now, analogously to the previous section, we can compute enclosures $(\mathbf{X}_j)$ such that for every $A_{ij} \in \mathbf{A}_{ij}$ and for every $C_j \in \mathbf{C}_j$ there exist solutions $X_j \in \mathbf{X}_j$ of the underdetermined linear system

$$\sum_{j=1}^n \langle A_{ij}, X_j \rangle = 0 \quad \text{for } i = 1, \ldots, m, \qquad \sum_{j=1}^n \langle C_j, X_j \rangle = \beta. \tag{5.2}$$

If an enclosure $(\mathbf{X}_j)$ is computed and if all minimal eigenvalues $\lambda_{min}(\mathbf{X}_j)$ are nonnegative, then, because $\beta < 0$, it follows that for every $P \in \mathbf{P}$ there exists $X_j \in \mathbf{X}_j$ for $j = 1, \ldots, n$ such that (5.1) is satisfied. Therefore, all problems with $P \in \mathbf{P}$ are dual infeasible, and the block matrices $(\mathbf{X}_j)$ contain the primal improving rays.

The dual problem (1.3) has a dual improving ray if there is a vector $y \in \mathbb{R}^m$ such that

$$\sum_{i=1}^{m} y_i A_{ij} \preceq 0 \quad \text{for } j = 1, \ldots, n, \quad \text{and} \quad b^T y > 0. \tag{5.3}$$

The existence of a dual improving ray implies primal infeasibility. If interval input data $\mathbf{P}$ are given and for some problem $P \in \mathbf{P}$ an approximate dual improving ray $\tilde{y}$ is known, then we can try to verify primal infeasibility for all problems with $P \in \mathbf{P}$ as follows: First we compute by using interval arithmetic upper bounds of the maximal eigenvalues of $\sum_{i=1}^{m} \tilde{y}_i \mathbf{A}_{ij}$ for $j = 1, \ldots, n$. If these upper bounds are nonpositive and if $\mathbf{b}^T \tilde{y} > 0$, then $\tilde{y}$ is a dual improving ray for all $P \in \mathbf{P}$. Hence, $\tilde{y}$ is a rigorous certificate of primal infeasibility for all $P \in \mathbf{P}$.

6. Numerical results. In this section, we present some numerical experiments for semidefinite and linear programming problems. The results for the semidefinite programming problems were obtained by using VSDP [14], a MATLAB software package for verified semidefinite programming. This package uses the toolbox INTLAB [44] and allows interval input data. For all experiments we used the default values of the solver.

First we consider a semidefinite program of small size,

$$\begin{aligned}
\min \quad & \left\langle \begin{pmatrix} 0 & \frac{1}{2} & 0 \\ \frac{1}{2} & \delta & 0 \\ 0 & 0 & \delta \end{pmatrix}, X \right\rangle \\
\text{s.t.} \quad & \left\langle \begin{pmatrix} 0 & -\frac{1}{2} & 0 \\ -\frac{1}{2} & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, X \right\rangle = 1, \\
& \left\langle \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, X \right\rangle = \varepsilon, \\
& \left\langle \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}, X \right\rangle = 0, \\
& \left\langle \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, X \right\rangle = 0, \\
& X \succeq 0.
\end{aligned}$$

TABLE 6.1
Approximations $\tilde{p}^*$, $\tilde{d}^*$ and rigorous bounds $\overline{p}^*$, $\underline{p}^*$.

ε	δ	$\tilde{p}^*$	$\tilde{d}^*$	tc	$\overline{p}^*$	$\underline{p}^*$
0	0	−1.0004	−0.99355	0	∞	$-\infty$
10^{-8}	10^{-8}	−0.99184	−0.98372	0	∞	−0.98373
10^{-6}	10^{-10}	−1.0007	−1.0027	0	−0.99965	−1.0061
10^{-4}	10^{-3}	8.9004	8.9990	0	9.3353	8.9990
-10^{-4}	10^{-3}	28.228	142.86	0	∞	142.86
10^{-4}	-10^{-4}	−5.9323	−1.0361	−7	−5.9324	$-\infty$

The Lagrangian dual is

$$d^* = \max y_1 + \varepsilon y_2 \quad \text{s.t.} \quad Y := C - \sum_{i=1}^{4} A_{i1} y_i$$
$$= \begin{pmatrix} -y_2 & \frac{1+y_1}{2} & -y_3 \\ \frac{1+y_1}{2} & \delta & -y_4 \\ -y_3 & -y_4 & \delta \end{pmatrix} \succeq 0.$$

The linear constraints of the primal problem imply

$$X = \begin{pmatrix} \varepsilon & -1 & 0 \\ -1 & X_{22} & 0 \\ 0 & 0 & X_{33} \end{pmatrix},$$

and X is positive semidefinite iff $X_{22} \geq 0$, $X_{33} \geq 0$, and $\varepsilon \cdot X_{22} - (-1)^2 \geq 0$. Hence, for $\varepsilon \leq 0$, the problem is primal infeasible and $p^* = +\infty$. The dual problem is infeasible for $\delta < 0$ with $d^* = -\infty$.

For $\varepsilon = 0$ and $\delta = 0$ we obtain a duality gap with $p^* = +\infty$ and $d^* = -1$, and the problem is ill-posed. For $\varepsilon > 0$ and $\delta > 0$ Slater's constraint qualifications are satisfied and the optimal value of $p^* = d^* = -1 + \delta/\varepsilon$.

Numerical results for different values of ε and δ are summarized in Table 6.1. The termination code $tc = 0$ in SDPT3 means normal termination without warning, whereas $tc = -7$ indicates primal infeasibility. In the following, $\tilde{p}^*$ and $\tilde{d}^*$ denote the computed approximations of the primal and dual optimal value, respectively.

We see that SDPT3 is not backward stable, since in five cases $\tilde{p}^* < \tilde{d}^*$, violating the weak duality. This example demonstrates that the measures for termination and accepting an approximation are not appropriate for ill-conditioned or ill-posed problems. The rigorous bounds $\overline{p}^*$ and $\underline{p}^*$ recognize difficult problems much better and overestimate the optimal value only slightly, and this overestimation depends on the quality of the computed approximations. The bounds are infinite if the problem is infeasible or very ill-conditioned. For larger values $\varepsilon > 0$ and $\delta > 0$ the approximations and the rigorous bounds are almost identical, and are not displayed here.

In the following, we describe the numerical results on problems from the SDPLIB collection of Borchers [5]. In summary, VSDP could compute (by using SDPT3 as approximate solver), for all 85 problems discussed in [9], a rigorous lower bound of the optimal value and verify the existence of strictly dual feasible solutions which implies a zero duality gap. A finite rigorous upper bound could be computed for all well-posed problems with the exception of `hinf8`, which is ill-conditioned. For the 32 ill-posed problems VSDP has computed $\overline{p}^* = +\infty$, which reflects that the distance

to the next primal infeasible problem is zero as well as the infinite condition number. Detailed numerical results can be found in the tables contained in the appendix. We measure the accuracy by the quantity

$$\mu(a,b) := \frac{a-b}{\max\{1.0, (|a|+|b|)/2\}}.$$

Notice that we do not use the absolute value of $a-b$. Hence, a negative sign implies that $a < b$. Table A.1 contains the rigorous upper bound $\overline{p}^*$, the rigorous lower bound $\underline{p}^*$, the rigorous error $\mu(\overline{p}^*, \underline{p}^*)$, and the approximate duality gap $\mu(\tilde{p}^*, \tilde{d}^*)$. We have set $\mu(\overline{p}^*, \underline{p}^*) = NaN$ if the upper bound $\overline{p}^*$ is infinite. Table A.2 contains the termination code tc given by SDPT3 and the time t in seconds for computing the approximations. The times for computing the rigorous lower and upper bounds are $\underline{t}$ and $\overline{t}$, respectively.

Some major characteristics of our numerical results for the SDPLIB are as follows: The median of the time ratio for computing the rigorous lower bound and the approximations is $\mathrm{med}(\underline{t}/t) = 0.045$ and $\mathrm{med}(\overline{t}/t) = 2.4$. The median of the guaranteed accuracy $\mu(\overline{p}^*, \underline{p}^*)$ for the problems with finite condition number is $4.9 \cdot 10^{-7}$. We have used the median here because there are some outliers.

For the SDPLIB problems, SDPT3 (with default values) gave 7 warnings, and 2 warnings were given for well-posed problems. Hence, no warnings were given for 27 ill-posed problems with zero distance to primal infeasibility. In other words, there is no correlation between warnings and the difficulty of the problem. At least for this test set our rigorous bounds reflect the difficulty of the problems much better, and they provide safety, especially in the case where algorithms subsequently call other algorithms, as is done for example in branch-and-bound methods. In our opinion the usefulness of a warning is dubious if for ill-posed problems normal termination occurs. In contrast, VSDP shows a strong correlation between the rigorous bounds and the difficulty of the problem. Moreover, the negative signs of $\mu(\tilde{p}^*, \tilde{d}^*)$ show that the approximations do not satisfy weak duality. Therefore they are not backward stable; i.e., they are not exact solutions of slightly perturbed problems.

One of the largest problems which could be solved by VSDP is `thetaG51`, where the number of constraints is $m = 6910$, $n = 1$ and the dimension of the primal Solution X_1 is $s_1 = 1001$ (implying 501501 variables). For this problem SDPT3 gave the message out of memory, and we used SDPA [10] as approximate solver. The rigorous lower and upper bounds computed by VSDP are $\underline{p}^* = -3.4900 \cdot 10^2$, $\overline{p}^* = -3.4406 \cdot 10^2$, respectively. The guaranteed relative accuracy is only 0.014, which may be sufficient in several applications, but it is insufficient from a numerical point of view. The times in seconds for computing the approximations, the lower bounds, and the upper bounds of the optimal value are $t = 3687.95$, $\underline{t} = 45.17$, and $\overline{t} = 6592.52$, respectively. Existence of optimal solutions and strong duality is proved. For the 11 instances, `control1` to `control11`, the guaranteed relative accuracies are also low, ranging from $1.21 \cdot 10^{-4}$ to $5.93 \cdot 10^{-2}$.

The aim in designing VSDP was to write the programs in a convenient form, appropriate also for students. For example, the user can specify the block-diagonal structure in the more natural way of cell arrays. Operations such as extracting selected elements of a matrix or the operations `vsvec` (rounding-error-free vectorization operator) and `vsmat` (corresponding rounding-error-free inverse operator) are written in MATLAB [29]. This causes a loss of efficiency. At present a C++ implementation of these error bounds using the interval library PROFIL/BIAS [25] is in process [6].

We already have a C++ implementation of the rigorous bounds for the special case of linear programming called Lurupa [21]. Here we present a summary of our numerical results for the NETLIB suite of linear programming problems [33]. For details we refer the reader to [23]. The NETLIB collection contains problems with up to 15695 variables and 16675 constraints. They originate from various applications, for example forestry, flap settings on aircraft, and staff scheduling. In order to compare the results, we initially chose the set of problems that Ordóñez and Freund [38] have computed condition numbers for. Removing problems *DEGEN*3 and *PILOT* because the approximate solver failed to solve these within 24 hours, the final test set included 87 problems. We applied our algorithm to the original problems without preprocessing. As Fourer and Gay [8] observed, preprocessing can change the state of an LP from feasible to infeasible and vice versa.

Roughly speaking, a finite bound can be computed iff the corresponding distance to infeasibility is greater than 0. For 76 problems a finite lower bound could be computed with a median accuracy of $\operatorname{med}(\mu(\underline{p}^*, \tilde{p}^*)) = 2.2 \cdot 10^{-8}$ and a median time ratio of $\operatorname{med}(\underline{t}/t) = 0.5$. For 35 problems the algorithm yielded a finite upper bound with $\operatorname{med}(\mu(\overline{p}^*, \tilde{p}^*)) = 8.0 \cdot 10^{-9}$ and $\operatorname{med}(\overline{t}/t) = 5.3$. For 32 problems finite rigorous lower and upper bounds could be computed with $\operatorname{med}(\mu(\overline{p}^*, \underline{p}^*)) = 5.6 \cdot 10^{-8}$. Taking into account the approximate solver's default stopping tolerance of 10^{-9}, the guaranteed accuracy for the NETLIB LP suite is reasonable. The upper bound is more expensive, since linear systems have to be solved rigorously.

To our knowledge no other software packages compute rigorous results for semidefinite programs. There are several packages that compute verified results for optimization problems where the objective and the constraints are defined by smooth algebraic expressions. The most widely known solvers for global optimization that use interval arithmetic are probably GlobSol [18], Numerica [48], and COSY [4] (on inquiry we did not receive a copy of COSY). They do not permit the nonsmooth semidefinite constraints. But they can be used to solve linear programming problems (a special, smooth case of semidefinite programming) rigorously. Packages for verified constraint programming, like ICOS [28] and RealPaver [11], can also be used to compute rigorous results for an LP by adding bounds on the objective function to the set of constraints. Specifically dedicated to verified linear programming are solvers using a rational arithmetic. Examples are QSopt_ex [2], exlp [24], and, to a limited extent, perPlex [26]. The latter one only verifies the optimality of a given basis.

We want to see how these packages compare for the special case of linear programming. For this purpose we generate dense test problems with a method similar to the one described by Rosen and Suzuki [41]. Setting a timeout of one hour on each solver run, we then determine the largest problem size that the packages can solve. First we have a look at the packages that do not exploit the linear structure in the test problems. GlobSol, Numerica, and RealPaver fail to solve the test problems if the number of variables and constraints exceeds 10. ICOS solves the problems within one hour with up to 100 variables and constraints. Now looking at the packages taking account of the linear structure, we see that these succeed for larger problems. QSopt_ex, exlp, and perPlex solve problems with 200 variables and constraints. Lurupa does not fail until the dimension exceeds 1500. For larger dimensions it turns out that it already takes more than one hour to compute the initial approximate solution. The verification part needs only a fraction of this time. Our algorithm, however, produces rigorous bounds for the optimal value and for feasible solutions close to optimality; enclosures of optimal solutions are not computed. Elaborate numerical results can be

found in [22].

7. Conclusions. The computation of rigorous error bounds for linear or semidefinite optimization problems can be viewed as a carefully postprocessing tool that uses only approximate solutions computed by an LP or SDP solver. The numerical results show that such rigorous error bounds can be computed even for problems of large size.

Appendix. Results from the SDPLIB.

TABLE A.1

Rigorous bounds for the SDPLIB. Notice that all problems with $\overline{p}^* = \infty$ are ill-posed with the exception of `hinf8`.

Problem	$\overline{p}^*$	$\underline{p}^*$	$\mu(\overline{p}^*, \underline{p}^*)$	$\mu(\tilde{p}^*, \tilde{d}^*)$
arch0	$-5.6651e-01$	$-5.6652e-01$	$5.05e-06$	$-2.77e-07$
arch2	$-6.7151e-01$	$-6.7152e-01$	$5.16e-06$	$1.01e-07$
arch4	$-9.7263e-01$	$-9.7263e-01$	$3.66e-08$	$3.66e-08$
arch8	$-7.0570e+00$	$-7.0570e+00$	$1.84e-07$	$1.61e-07$
control1	$-1.7782e+01$	$-1.7785e+01$	$1.21e-04$	$-6.13e-08$
control2	$-8.2909e+00$	$-8.3000e+00$	$1.10e-03$	$3.74e-08$
control3	$-1.3615e+01$	$-1.3633e+01$	$1.36e-03$	$-1.89e-07$
control4	$-1.9671e+01$	$-1.9794e+01$	$6.24e-03$	$-7.71e-07$
control5	$-1.6795e+01$	$-1.6884e+01$	$5.25e-03$	$-1.72e-06$
control6	$-3.6736e+01$	$-3.7304e+01$	$1.53e-02$	$2.15e-06$
control7	$-2.0434e+01$	$-2.0625e+01$	$9.29e-03$	$1.36e-06$
control8	$-2.0019e+01$	$-2.0286e+01$	$1.33e-02$	$-5.36e-08$
control9	$-1.4036e+01$	$-1.4675e+01$	$4.45e-02$	$-2.57e-06$
control10	$-3.6890e+01$	$-3.8533e+01$	$4.36e-02$	$2.79e-06$
control11	$-3.0119e+01$	$-3.1959e+01$	$5.93e-02$	$3.40e-06$
equalG11	$-6.2915e+02$	$-6.2916e+02$	$1.08e-05$	$2.62e-05$
equalG51	$-4.0055e+03$	$-4.0056e+03$	$3.58e-05$	$4.17e-08$
gpp100	∞	$4.4943e+01$	NaN	$1.24e-04$
gpp124-1	∞	$7.3431e+00$	NaN	$6.94e-07$
gpp124-2	∞	$4.6862e+01$	NaN	$1.23e-07$
gpp124-3	∞	$1.5301e+02$	NaN	$1.28e-04$
gpp124-4	∞	$4.1899e+02$	NaN	$9.83e-05$
gpp250-1	∞	$1.5443e+01$	NaN	$6.85e-03$
gpp250-2	∞	$8.1869e+01$	NaN	$1.44e-06$
gpp250-3	∞	$3.0354e+02$	NaN	$4.30e-07$
gpp250-4	∞	$7.4733e+02$	NaN	$2.81e-07$
gpp500-1	∞	$2.5320e+01$	NaN	$2.04e-06$
gpp500-2	∞	$1.5606e+02$	NaN	$1.30e-03$
gpp500-3	∞	$5.1302e+02$	NaN	$1.21e-05$
gpp500-4	∞	$1.5670e+03$	NaN	$8.81e-08$
hinf1	∞	$-2.0328e+00$	NaN	$-1.12e-04$
hinf2	$-7.1598e+00$	$-1.0967e+01$	$4.20e-01$	$-3.38e-05$
hinf3	∞	$-5.6954e+01$	NaN	$-2.30e-04$
hinf4	∞	$-2.7477e+02$	NaN	$-1.43e-05$
hinf5	∞	$-3.6259e+02$	NaN	$2.20e-03$
hinf6	∞	$-4.4914e+02$	NaN	$-1.21e-03$
hinf7	∞	$-3.9082e+02$	NaN	$-8.98e-06$
hinf8	∞	$-1.1618e+02$	NaN	$-3.93e-04$
hinf9	∞	$-2.3709e+02$	NaN	$-3.43e-02$
hinf10	∞	$-1.0887e+02$	NaN	$-1.42e-03$
hinf11	∞	$-6.5944e+01$	NaN	$-1.24e-03$
hinf12	∞	$-7.8714e-01$	NaN	$-6.66e-01$
hinf13	∞	$-4.6009e+01$	NaN	$-3.62e-02$
hinf14	∞	$-1.3001e+01$	NaN	$1.70e-04$
hinf15	∞	$-2.6324e+01$	NaN	$-9.20e-02$
maxG11	$-6.2916e+02$	$-6.2916e+02$	$5.13e-08$	$3.01e-08$

continued...

Problem	$\overline{p}^*$	$\underline{p}^*$	$\mu(\overline{p}^*, \underline{p}^*)$	$\mu(\tilde{p}^*, \tilde{d}^*)$
maxG32	$-1.5676e+03$	$-1.5676e+03$	$3.16e-07$	$-2.99e-08$
maxG51	$-4.0063e+03$	$-4.0063e+03$	$5.49e-08$	$-2.23e-08$
mcp100	$-2.2616e+02$	$-2.2616e+02$	$1.85e-08$	$1.11e-08$
mcp124-1	$-1.4199e+02$	$-1.4199e+02$	$8.92e-09$	$8.79e-09$
mcp124-2	$-2.6988e+02$	$-2.6988e+02$	$3.31e-08$	$9.67e-10$
mcp124-3	$-4.6775e+02$	$-4.6775e+02$	$8.57e-09$	$-1.45e-09$
mcp124-4	$-8.6441e+02$	$-8.6441e+02$	$1.46e-08$	$-7.77e-09$
mcp250-1	$-3.1726e+02$	$-3.1726e+02$	$1.00e-08$	$3.43e-08$
mcp250-2	$-5.3193e+02$	$-5.3193e+02$	$6.85e-09$	$-1.14e-10$
mcp250-3	$-9.8117e+02$	$-9.8117e+02$	$2.04e-08$	$1.86e-08$
mcp250-4	$-1.6820e+03$	$-1.6820e+03$	$1.59e-08$	$3.43e-08$
mcp500-1	$-5.9815e+02$	$-5.9815e+02$	$4.91e-08$	$-1.16e-08$
mcp500-2	$-1.0701e+03$	$-1.0701e+03$	$1.28e-07$	$-5.93e-08$
mcp500-3	$-1.8480e+03$	$-1.8480e+03$	$1.31e-08$	$7.66e-09$
mcp500-4	$-3.5667e+03$	$-3.5667e+03$	$1.19e-08$	$-3.91e-09$
qap10	∞	$1.0925e+03$	NaN	$-1.10e-04$
qap5	∞	$4.3600e+02$	NaN	$7.22e-10$
qap6	∞	$3.8141e+02$	NaN	$-9.00e-05$
qap7	∞	$4.2479e+02$	NaN	$-7.27e-05$
qap8	∞	$7.5686e+02$	NaN	$-1.29e-04$
qap9	∞	$1.4099e+03$	NaN	$-5.35e-05$
qpG11	$-2.4487e+03$	$-2.4487e+03$	$6.42e-09$	$6.42e-09$
qpG51	$-1.1818e+04$	$-1.1818e+04$	$7.44e-09$	$7.44e-09$
ss30	$-2.0239e+01$	$-2.0240e+01$	$1.55e-06$	$1.55e-06$
theta1	$-2.3000e+01$	$-2.3000e+01$	$1.55e-08$	$9.89e-09$
theta2	$-3.2879e+01$	$-3.2879e+01$	$1.18e-06$	$1.17e-08$
theta3	$-4.2167e+01$	$-4.2167e+01$	$9.64e-07$	$-1.73e-08$
theta4	$-5.0321e+01$	$-5.0321e+01$	$1.09e-06$	$9.41e-09$
theta5	$-5.7232e+01$	$-5.7232e+01$	$4.92e-07$	$-6.68e-09$
thetaG11	$-4.0000e+02$	$-4.0000e+02$	$4.23e-07$	$2.02e-10$
truss1	$9.0000e+00$	$9.0000e+00$	$1.61e-07$	$3.08e-09$
truss2	$1.2338e+02$	$1.2338e+02$	$1.60e-06$	$-5.86e-09$
truss3	$9.1100e+00$	$9.1100e+00$	$1.11e-07$	$1.11e-07$
truss4	$9.0100e+00$	$9.0100e+00$	$1.19e-08$	$4.15e-09$
truss5	$1.3264e+02$	$1.3264e+02$	$9.09e-06$	$-2.28e-07$
truss6	$9.0119e+02$	$9.0100e+02$	$2.16e-04$	$-3.26e-07$
truss7	$9.0015e+02$	$9.0000e+02$	$1.70e-04$	$-4.67e-06$
truss8	$1.3312e+02$	$1.3311e+02$	$7.79e-05$	$-4.98e-06$

TABLE A.2
Computational effort for the SDPLIB.

Problem	tc	t	$\underline{t}$	$\overline{t}$
arch0	0	19.95	0.81	24.05
arch2	0	18.06	0.36	43.13
arch4	0	19.20	0.34	3.39
arch8	0	19.81	0.33	46.34
control1	0	1.28	0.05	1.66
control2	0	2.64	0.03	3.14
control3	0	6.20	0.09	6.56
control4	0	11.33	0.14	13.83
control5	0	23.11	0.28	34.75
control6	0	60.72	0.47	155.11
control7	0	87.67	0.72	138.31
control8	0	152.98	1.23	232.28
control9	0	238.88	635.77	390.06
control10	0	432.48	3.05	743.75
control11	0	661.59	3.95	1129.63

continued...

Problem	tc	t	$\underline{t}$	$\overline{t}$
equalG11	0	590.89	844.22	896.25
equalG51	0	1677.77	1369.69	1638.83
gpp100	0	2.34	0.30	35.88
gpp124-1	0	5.88	6.25	38.55
gpp124-2	0	6.11	0.56	41.75
gpp124-3	0	3.95	0.63	36.80
gpp124-4	0	4.16	0.59	43.89
gpp250-1	0	18.64	5.58	122.00
gpp250-2	0	31.14	5.53	112.63
gpp250-3	0	24.77	5.08	160.81
gpp250-4	0	24.66	5.56	130.88
gpp500-1	0	190.39	236.81	642.58
gpp500-2	0	145.53	46.69	809.75
gpp500-3	0	156.52	46.73	961.78
gpp500-4	0	205.06	46.42	815.34
hinf1	0	1.28	0.05	16.95
hinf2	0	1.03	0.06	12.23
hinf3	0	1.11	0.02	16.73
hinf4	0	1.09	0.06	5.98
hinf5	−4	1.23	0.05	6.81
hinf6	0	1.34	0.06	8.69
hinf7	−4	1.22	1.31	13.63
hinf8	−4	1.19	0.06	5.83
hinf9	2	1.38	0.06	7.98
hinf10	0	1.80	0.05	6.02
hinf11	0	1.94	0.05	11.25
hinf12	0	3.00	6.28	6.64
hinf13	0	1.94	0.08	7.39
hinf14	−4	1.44	0.06	6.91
hinf15	0	2.34	0.08	10.47
maxG11	0	449.28	13.14	678.13
maxG32	0	6910.72	176.53	9838.19
maxG51	0	990.22	26.47	1352.08
mcp100	0	2.44	0.09	5.53
mcp124-1	0	3.13	0.13	0.55
mcp124-2	0	3.69	0.13	9.16
mcp124-3	0	3.91	0.14	4.59
mcp124-4	0	4.09	0.17	4.91
mcp250-1	0	15.33	0.47	20.83
mcp250-2	0	16.41	0.52	22.16
mcp250-3	0	18.08	0.59	43.17
mcp250-4	0	18.80	0.70	23.08
mcp500-1	0	104.31	3.03	146.30
mcp500-2	0	125.19	3.58	167.39
mcp500-3	0	127.05	3.94	301.72
mcp500-4	0	134.67	4.58	176.47
qap10	0	12.03	2.50	42.36
qap5	0	0.84	0.06	6.08
qap6	0	1.42	0.13	4.27
qap7	0	2.28	0.25	6.19
qap8	0	3.84	0.58	11.22
qap9	0	7.08	1.06	21.95
qpG11	0	1081.17	58.72	998.61
qpG51	0	3520.34	113.42	1380.81
ss30	0	49.09	0.81	5.22
theta1	0	0.92	0.09	1.13
theta2	0	3.75	1.33	11.28
theta3	0	14.67	8.02	35.41

continued...

Problem	tc	t	$\underline{t}$	$\overline{t}$
theta4	0	45.06	22.16	217.53
theta5	0	141.27	59.50	446.95
thetaG11	−7	725.06	768.33	2142.47
truss1	0	1.09	0.06	3.42
truss2	0	6.95	0.28	29.55
truss3	0	1.36	0.05	0.09
truss4	0	1.05	0.06	1.19
truss5	0	14.44	0.66	45.80
truss6	0	41.75	2.20	183.61
truss7	0	33.86	1.39	72.20
truss8	0	18.69	1.63	23.88

Acknowledgments. We wish to thank Arnold Neumaier and Siegfried M. Rump for their stimulating input and the two anonymous referees for many suggestions on improving this paper.

REFERENCES

[1] G. Alefeld and J. Herzberger, *Introduction to Interval Computations*, Academic Press, New York, 1983.

[2] D. L. Applegate, W. Cook, S. Dash, and D. G. Espinoza, *QSopt_ex*, http://www.dii.uchile.cl/∼daespino/QSoptExact_doc/main.html.

[3] V. Balakrishnan and E. Feron, eds., *Linear matrix inequalities in control theory and applications*, Internat. J. Robust Nonlinear Control, 6 (1996), pp. 895–1100.

[4] M. Berz et al., *COSY Infinity*, http://www.bt.pa.msu.edu/index_files/cosy.htm.

[5] B. Borchers, *SDPLIB* 1.2, *a library of semidefinite programming test problems*, Optim. Methods Softw., 11 (1999), pp. 683–690.

[6] D. Chaykin, *A* C++ *implementation of rigorous error bounds in semidefinite programming*, in progress.

[7] C. A. Floudas, *Deterministic Global Optimization—Theory, Methods and Applications*, Nonconvex Optim. Appl. 37, Kluwer Academic Publishers, Dordrecht, The Netherlands, Boston, London, 2000.

[8] R. Fourer and D. M. Gay, *Experience with a primal presolve algorithm*, in Large Scale Optimization: State of the Art, W. W. Hager, D. W. Hearn, and P. M. Pardalos, eds., Kluwer Academic Publishers, Norwell, MA, Dordrecht, The Netherlands, 1994, pp. 135–154.

[9] R. M. Freund, F. Ordóñez, and K. C. Toh, *Behavioral measures and their correlation with IPM iteration on semi-definite programming problems*, Math. Program., 109 (2007), pp. 445–475.

[10] K. Fujisawa, Y. Futakata, M. Kojima, K. Nakata, and M. Yamashita, *SDPA-M (SemiDefinite Programming Algorithm in MATLAB) User's Manual—Version* 2.00, Research report B-359, Tokyo Institute of Technology, Tokyo, Japan, 2000.

[11] L. Granvilliers and F. Benhamou, *Algorithm* 852*: RealPaver: An interval solver using constraint satisfaction techniques*, ACM Trans. Math. Software, 32 (2006), pp. 138–156.

[12] M. Grötschel, L. Lovász, and A. Schrijver, *The ellipsoid method and its consequences in combinatorial optimization*, Combinatorica, 1 (1981), pp. 169–197.

[13] E. R. Hansen, *Global Optimization Using Interval Analysis*, Marcel Dekker, New York, 1992.

[14] C. Jansson, *VSDP: Verified SemiDefinite Programming, User's Guide*, Beta Version 0.1, to appear.

[15] C. Jansson, *A self-validating method for solving linear programming problems with interval input data*, in Scientific Computation with Automatic Result Verification, Comput. Suppl. 6, Springer-Verlag, Vienna, 1988, pp. 33–45.

[16] C. Jansson, *A rigorous lower bound for the optimal value of convex optimization problems*, J. Global Optim., 28 (2004), pp. 121–137.

[17] C. Jansson, *Rigorous lower and upper bounds in linear programming*, SIAM J. Optim., 14 (2004), pp. 914–935.

[18] R. B. Kearfott, *GlobSol*, http://interval.louisiana.edu.

[19] R. B. Kearfott, *Rigorous Global Search: Continuous Problems*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1996.

[20] R. B. Kearfott, *On proving existence of feasible points in equality constrained optimization problems*, Math. Program., 83 (1998), pp. 89–100.
[21] C. Keil, *Lurupa—Rigorous Error Bounds in Linear Programming*, in Algebraic and Numerical Algorithms and Computer-Assisted Proofs, B. Buchberger, S. Oishi, M. Plum, and S. Rump, eds., no. 05391 in Dagstuhl Seminar Proceedings, Internationales Begegnungs- und Forschungszentrum fuer Informatik (IBFI), Schloss Dagstuhl, Germany, 2006, http://drops.dagstuhl.de/opus/volltexte/2006/445.
[22] C. Keil, *Verified linear programming—a comparison*, Reliable Computing, submitted.
[23] C. Keil and C. Jansson, *Computational experience with rigorous error bounds for the Netlib linear programming library*, Reliab. Comput., 12 (2006), pp. 303–321.
[24] M. Kiyomi, *exlp*, http://members.jcom.home.ne.jp/masashi777/exlp.html.
[25] O. Knüppel, *PROFIL/BIAS and extensions, Version* 2.0, Technical report, Inst. f. Informatik III, Technische Universität Hamburg-Harburg, 1998.
[26] T. Koch, *perPlex*, urlhttp://www.zib.de/koch/perplex.
[27] R. Krawczyk, *Fehlerabschätzung bei linearer Optimierung*, in Interval Mathematics, K. Nickel, ed., Lecture Notes in Comput. Sci. 29, Springer-Verlag, Berlin, 1975, pp. 215–222.
[28] Y. Lebbah, *ICOS (Interval COnstraints Solver)*, http://www.essi.fr/∼lebbah/icos/index.html.
[29] *MATLAB User's Guide, Version* 6, The MathWorks Inc., Natick, MA, 2000.
[30] G. Mayer, *Result verification for eigenvectors and eigenvalues*, in Topics in Validated Computations, Proceedings of the IMACS-GAMM International Workshop, Oldenburg, Germany, Stud. Comput. Math. 5, J. Herzberger, ed., Elsevier, Amsterdam, 1994, pp. 209–276.
[31] R. E. Moore, *Methods and Applications of Interval Analysis*, SIAM, Philadelphia, 1979.
[32] A. Nemirovski, *Lectures on Modern Convex Optimization*, http://www2.isye.gatech.edu/∼nemirovs/Lect_ModConvOpt.pdf, 2003.
[33] *NETLIB linear programming library*, http://www.netlib.org/lp.
[34] A. Neumaier, *Interval methods for systems of equations*, Encyclopedia Math. Appl., Cambridge University Press, Cambridge, UK, 1990.
[35] A. Neumaier, *Introduction to Numerical Analysis*, Cambridge University Press, Cambridge, UK, 2001.
[36] A. Neumaier, *Complete search in continuous global optimization and constraint satisfaction*, Acta Numer., 13 (2004), pp. 271–369.
[37] A. Neumaier and O. Shcherbina, *Safe bounds in linear and mixed-integer programming*, Math. Program., 99 (2004), pp. 283–296.
[38] F. Ordóñez and R. M. Freund, *Computational experience and the explanatory value of condition measures for linear optimization*, SIAM J. Optim., 14 (2003), pp. 307–333.
[39] M. V. Ramana, L. Tunçel, and H. Wolkowicz, *Strong duality for semidefinite programming*, SIAM J. Optim., 7 (1997), pp. 641–662.
[40] J. Renegar, *Linear programming, complexity theory and elementary functional analysis*, Math. Program., 70 (1995), pp. 279–351.
[41] J. B. Rosen and S. Suzuki, *Construction of nonlinear programming test problems*, Communication of ACM, 8 (1965), p. 113.
[42] S. M. Rump, *Validated solution of large linear systems*, in Validation Numerics: Theory and Applications, Comput. Suppl. 9, R. Albrecht, G. Alefeld, and H. Stetter, eds., Springer-Verlag, Berlin, 1993, pp. 191–212.
[43] S. M. Rump, *Verification methods for dense and sparse systems of equations*, in Topics in Validated Computations—Studies in Computational Mathematics, J. Herzberger, ed., Elsevier, Amsterdam, 1994, pp. 63–136.
[44] S. M. Rump, *INTLAB—Interval Laboratory, a MATLAB Toolbox for Verified Computations, Version* 5.1, 2005.
[45] R. E. Skelton and T. Iwasaki, *Increased roles of linear algebra in control education*, IEEE Control Syst. Mag., 15 (1995), pp. 76–89.
[46] M. J. Todd, *Detecting infeasibility in infeasible-interior-point methods for optimization*, in Foundations of Computational Mathematics: Minneapolis, 2002, London Math. Soc. Lecture Note Ser. 312, F. Cucker, R. DeVore, P. Olver, and E. Süli, eds., Cambridge University Press, Cambridge, UK, 2004, pp. 157–192.
[47] R. H. Tütüncü, K. C. Toh, and M. J. Todd, *Solving semidefinite-quadratic-linear programs using SDPT*3, Math. Program., 95B (2003), pp. 189–217.
[48] P. Van Hentenryck, P. Michel, and Y. Deville, *Numerica: A Modelling Language for Global Optimization*, MIT Press, Cambridge, MA, 1997.
[49] L. Vandenberghe and S. Boyd, *Semidefinite programming*, SIAM Rev., 38 (1996), pp. 49–95.

SIAM J. NUMER. ANAL.
Vol. 46, No. 1, pp. 201–230
© 2007 Society for Industrial and Applied Mathematics

THE POSTPROCESSED MIXED FINITE-ELEMENT METHOD FOR THE NAVIER–STOKES EQUATIONS: REFINED ERROR BOUNDS*

JAVIER DE FRUTOS†, BOSCO GARCÍA-ARCHILLA‡, AND JULIA NOVO§

Abstract. A postprocessing technique for mixed finite-element methods for the incompressible Navier–Stokes equations is analyzed. The postprocess, which amounts to solving a (linear) Stokes problem, is shown to increase the order of convergence of the method to which it is applied by one unit (times the logarithm of the mesh diameter). In proving the error bounds, some superconvergence results are also obtained. Contrary to previous analysis of the postprocessing technique, in the present paper we take into account the loss of regularity suffered by the solutions of the Navier–Stokes equations at the initial time in the absence of nonlocal compatibility conditions of the data. As in [H. G. Heywood and R. Rannacher, *SIAM J. Numer. Anal.*, 25 (1988), pp. 489–512], where the same hypothesis is assumed, no better than fifth-order convergence is achieved.

Key words. Navier–Stokes equations, mixed finite-element methods, optimal regularity, error estimates

AMS subject classifications. 65M60, 65M20, 65M15, 65M12

DOI. 10.1137/06064458

1. Introduction.

It is well known that, no matter how regular the data are, solutions of the Navier–Stokes equations cannot be assumed to have more than second-order spatial derivatives bounded in L^2 up to initial time $t = 0$, since this requires the data to satisfy nonlocal compatibility conditions unlikely to be fulfilled in practical situations [36], [37]. Therefore, error analysis of numerical methods for the Navier–Stokes equations is more meaningful if this fact is taken into account. This is the case of the present paper, where we analyze a postprocessing technique that improves the errors of mixed finite-element (MFE) methods for the Navier–Stokes equations from $O(h^r)$ to $O(h^{r+1}\,|\log(h)|)$, where h is the mesh size.

This postprocessing technique was first developed for spectral methods in [26], [27], in connection with approximate inertial manifolds [14], [15]. Later it was extended to methods based on Chebyshev and Legendre polynomials [16], spectral-element methods [18], and finite-element methods [28], [19], together with being applied in the study of nonlinear shell vibrations [43]. Also, it has been effectively applied to reduce the order of practical engineering problems modeled by nonlinear differential systems [46], [45].

More recently, the postprocessing technique has also been developed for MFE methods for the Navier–Stokes equations in [4], [6], but for solutions more regular up to $t = 0$ than what can be realistically assumed in practice, and this allows the

*Received by the editors July 6, 2006; accepted for publication (in revised form) July 25, 2007; published electronically December 20, 2007.

http://www.siam.org/journals/sinum/46-1/66445.html

†Departamento de Matemática Aplicada, Universidad de Valladolid, 47011 Valladolid, Spain (frutos@mac.uva.es). This author's research was financed by DGI-MEC under project MTM2004-02847 (cofinanced by FEDER funds) and Junta de Castilla y León under grants VA079A06 and VA045A06.

‡Departamento de Matemática Aplicada II, Universidad de Sevilla, 41002 Sevilla, Spain (bosco@esi.us.es). This author's research was supported by DGI-MEC under project MTM2006-00847.

§Departamento de Matemáticas, Universidad Autónoma de Madrid, 28049 Madrid, Spain (julia.novo@uam.es). This author's research was financed by DGI-MEC under project MTM2004-02847 (cofinanced by FEDER funds) and Junta de Castilla y León under grants VA079A06 and VA045A06.

above-mentioned improvement of the error for $r \geq 2$. In the present paper, however, due to the loss of regularity at $t = 0$, no better than $O(h^5 |\log(h)|)$ error bounds are proved, a limitation similar to that already found in [37] for MFE methods. On the other hand, as in [6], although for simplicity we concentrate on Hood–Taylor [39] elements, the postprocessing technique can be easily adapted to other kinds of mixed LBB-stable elements.

We analyze the postprocessing technique for the incompressible Navier–Stokes equations

$$\begin{aligned} u_t - \Delta u + (u \cdot \nabla)u + \nabla p &= f, \\ \operatorname{div}(u) &= 0, \end{aligned} \tag{1.1}$$

in a bounded domain $\Omega \subset \mathbb{R}^d$ $(d = 2, 3)$ with a smooth boundary subject to homogeneous Dirichlet boundary conditions $u = 0$ on $\partial\Omega$. In (1.1), u is the velocity field, p the pressure, and f a given force field. We assume that the fluid density and viscosity have been normalized by an adequate change of scale in space and time.

The postprocessing technique to approximate up to time $T > 0$ the solution u and p corresponding to a given initial condition

$$u(\cdot, 0) = u_0 \tag{1.2}$$

is as follows. Compute first standard MFE approximations u_h and p_h to the velocity and pressure, respectively, by integrating in time the corresponding discretization of (1.1)–(1.2). Then compute MFE approximations $\tilde{u}_h$ and $\tilde{p}_h$ to the solution $\tilde{u}$ and $\tilde{p}$ of the following Stokes problem:

$$\left.\begin{aligned} -\Delta\tilde{u} + \nabla\tilde{p} = f - \tfrac{d}{dt}u_h - (u_h \cdot \nabla)u_h \\ \operatorname{div}(\tilde{u}) = 0 \end{aligned}\right\} \quad \text{in } \Omega, \qquad \tilde{u} = 0, \quad \text{on } \partial\Omega. \tag{1.3}$$

The MFE on this postprocessing step can be either the same-order Hood–Taylor element over a finer grid or a higher-order Hood–Taylor element over the same grid. Observe that, since the MFE approximations u_h and p_h do not depend on $\tilde{u}$ and $\tilde{p}$ or their approximations $\tilde{u}_h$ and $\tilde{p}_h$, these may be computed only when needed. If this happens only at the final time $t = T$, then an $O(h^{r+1} |\log(h)|)$ error is obtained at the cost of an $O(h^r)$ one, since only a single Stokes problem such as (1.3) is solved with the enhanced MFE method, and this is done when the time integration is completed.

The method studied in the present paper can be seen as a two-level method, where the postprocessed (or fine-mesh) approximations $\tilde{u}_h$ and $\tilde{p}_h$ are an improvement of the previously computed (coarse-mesh) approximations u_h and p_h. Recent research on two-level finite-element methods for the transient Navier–Stokes equations can be found in [30], [31], [32], [44] (see also [34], [33], [40], [42] for spectral discretizations), where the full nonlinear problem is dealt with on the coarse mesh, and a linear problem is solved on the fine mesh. Let us mention that, whereas in the present paper we are concerned with higher-order methods, in [30], [31], [32], [44] only low-order methods are dealt with. Also, in the present paper, computations on the finer grid are done only at the final time $t = T$ (or at those time levels where more accuracy is wanted), whereas in the previous works computations on the finer grids are carried out from $t = 0$ to $t = T$. It must be mentioned though, as opposed to [2], that the coarse-grid

component computation in [30], [31], [32], [44] is fully decoupled from that on the fine grid.

Although in the present paper we concentrate on the spatial discretization, practical computations are affected by the errors induced by the time discretization. Numerical experiments in [6], [16], [19], [17], [26], [28] have repeatedly shown, for the different discretizations considered, that the increase in accuracy and convergence rate predicted by the theory is also seen in practice (provided errors arising from the time discretization are kept sufficiently small). Nevertheless, in the present paper we give an explanation of this fact; that is, the gain in (spatial) accuracy in the postprocessing step takes place independently of errors arising from the temporal discretization being present or not. Similar results are obtained in [47] for finite-element methods of degree 3 or larger for reaction-diffusion problems when integrated by the implicit Euler method.

To prove our error estimates, we need to prove first some superconvergence results. These are not with respect to the Stokes projection (as in [17], [6], [4]) but with respect to the solution of a certain linear evolution problem. In this process, we also obtain error bounds for the pressure that improve those originally proved in [37, Theorem 3.1] by a factor of $t^{1/2}$.

Finally, we remark that in recent works [20], [21], [22] the postprocessing technique has shown itself useful in obtaining efficient a posteriori error estimators for partial differential equations of evolution, a field which is remarkably less developed than that of steady problems. The application of the postprocessing technique to a posteriori error estimates for the Navier–Stokes equations using the results obtained in the present paper will be the subject of future research.

The rest of the paper is as follows. In section 2 we introduce the notation and some standard material. In section 3 we comment on how to approximate the Stokes problem of the postprocessing step. Section 4 is devoted to the superconvergence analysis of the MFE approximation. In section 5 error bounds of the postprocessed approximation are obtained. Finally, we make some remarks on postprocessing when time discretization is taken into account.

2. Preliminaries and notations. We will assume that Ω is a bounded domain in $\mathbb{R}^d$, $d = 2, 3$, not necessarily convex, of class $\mathcal{C}^m$, for $m \geq 3$, and we consider the Hilbert spaces

$$H = \{u \in \left(L^2(\Omega)\right)^d \mid \operatorname{div}(u) = 0,\ u \cdot n_{|_{\partial\Omega}} = 0\},$$

$$V = \{u \in \left(H_0^1(\Omega)\right)^d \mid \operatorname{div}(u) = 0\},$$

endowed with the inner product of $L^2(\Omega)^d$ and $H_0^1(\Omega)^d$, respectively. For $l \geq 0$ integer and $1 \leq q \leq \infty$, we consider the standard Sobolev spaces $W^{l,q}(\Omega)^d$ of functions with derivatives up to order l in $L^q(\Omega)$, and $H^l(\Omega)^d = W^{l,2}(\Omega)^d$. We will denote by $\|\cdot\|_l$ the norm in $H^l(\Omega)^d$, and $\|\cdot\|_{-l}$ will represent the norm of its dual space. We consider also the quotient spaces $H^l(\Omega)/\mathbb{R}$ with norm $\|p\|_{H^l/\mathbb{R}} = \inf\{\|p + c\|_l \mid c \in \mathbb{R}\}$.

Let us recall the following Sobolev's imbeddings [1]: For $q \in [1, \infty)$, there exists a constant $C = C(\Omega, q)$ such that

$$\|v\|_{L^{q'}(\Omega)^d} \leq C\|v\|_{W^{s,q}(\Omega)^d}, \quad \frac{1}{q'} \geq \frac{1}{q} - \frac{s}{d} > 0, \quad q < \infty, \quad v \in W^{s,q}(\Omega)^d. \tag{2.1}$$

For $q' = \infty$, (2.1) holds with $\frac{1}{q} < \frac{s}{d}$.

Let $\Pi : L^2(\Omega)^d \longrightarrow H$ be the $L^2(\Omega)^d$ projection onto H. We denote by A the Stokes operator on Ω:

$$A : \mathcal{D}(A) \subset H \longrightarrow H, \quad A = -\Pi\Delta, \quad \mathcal{D}(A) = H^2(\Omega)^d \cap V.$$

Applying Leray's projector to (1.1), the equations can be written in the form

$$u_t + Au + B(u,u) = \Pi f \qquad \text{in } \Omega,$$

where $B(u,v) = \Pi(u\cdot\nabla)v$ for u, v in $H_0^1(\Omega)^d$.

We shall use the trilinear form $b(\cdot,\cdot,\cdot)$ defined by

$$b(u,v,w) = (F(u,v),w) \quad \forall u,v,w \in H_0^1(\Omega)^d,$$

where

$$F(u,v) = (u\cdot\nabla)v + \frac{1}{2}(\nabla\cdot u)v \quad \forall u,v \in H_0^1(\Omega)^d.$$

It is straightforward to verify that b enjoys the skew-symmetry property

$$b(u,v,w) = -b(u,w,v) \quad \forall u,v,w \in H_0^1(\Omega)^d. \tag{2.2}$$

Let us observe that $B(u,v) = \Pi F(u,v)$ for $u \in V$, $v \in H_0^1(\Omega)^d$.

We shall assume that

$$\|u(t)\|_1 \le M_1, \quad \|u(t)\|_2 \le M_2, \qquad 0 \le t \le T,$$

and, for $k \ge 2$ integer,

$$\sup_{0\le t\le T} \left\|\partial_t^{\lfloor k/2\rfloor} f\right\|_{k-1-2\lfloor k/2\rfloor} + \sum_{j=0}^{\lfloor (k-2)/2\rfloor} \sup_{0\le t\le T} \left\|\partial_t^j f\right\|_{k-2j-2} < +\infty,$$

so that, according to Theorems 2.4 and 2.5 in [36], there exist positive constants M_k and K_k such that for $k \ge 2$

$$\|u(t)\|_k + \|u_t(t)\|_{k-2} + \|p(t)\|_{H^{k-1}/\mathbb{R}} \le M_k \tau(t)^{1-k/2} \tag{2.3}$$

and for $k \ge 3$

$$\int_0^t \sigma_{k-3}(s)\big(\,\|u(s)\|_k^2 + \|u_s(s)\|_{k-2}^2 + \|p(s)\|_{H^{k-1}/\mathbb{R}}^2 + \|p_s(s)\|_{H^{k-3}/\mathbb{R}}^2\big)\,ds \le K_k^2, \tag{2.4}$$

where $\tau(t) = \min(t,1)$ and $\sigma_n = e^{-\alpha(t-s)}\tau^n(s)$ for some $\alpha > 0$. Observe that, for $t \le T < \infty$, we can take $\tau(t) = t$ and $\sigma_n(s) = s^n$. For simplicity, we will take these values of τ and σ_n. We note that no further than $k \le 6$ will be needed in the present paper.

Let $\mathcal{T}_h = (\tau_i^h, \phi_i^h)_{i\in I_h}$, $h > 0$, be a family of partitions of suitable domains Ω_h, where h is the maximum diameter of the elements $\tau_i^h \in \mathcal{T}_h$ and ϕ_i^h are the mappings of the reference simplex τ_0 onto τ_i^h. We restrict ourselves to quasi-uniform and regular meshes $\mathcal{T}_h$.

Let $r \ge 3$, and we consider the finite-element spaces

$$S_{h,r} = \big\{\chi_h \in \mathcal{C}(\overline{\Omega}_h)\,|\,\chi_h|_{\tau_i^h}\circ\phi_i^h \in P^{r-1}(\tau_0)\big\} \subset H^1(\Omega_h),$$

$$S_{h,r}^0 = \big\{\chi_h \in \mathcal{C}(\overline{\Omega}_h)\,|\,\chi_h|_{\tau_i^h}\circ\phi_i^h \in P^{r-1}(\tau_0),\ \chi_h(x) = 0\ \forall\, x \in \partial\Omega_h\big\} \subset H_0^1(\Omega_h),$$

where $P^{r-1}(\tau_0)$ denotes the space of polynomials of degree at most $r-1$ on τ_0. Since we are assuming that the meshes are quasi-uniform, the following inverse inequality holds for each $v_h \in (S^0_{h,r})^d$ (see, e.g., [9, Theorem 3.2.6])

$$\|v_h\|_{W^{m,q}(\tau)^d} \leq Ch^{l-m-d(\frac{1}{q'}-\frac{1}{q})}\|v_h\|_{W^{l,q'}(\tau)^d}, \tag{2.5}$$

where $0 \leq l \leq m \leq 1$, $1 \leq q' \leq q \leq \infty$, and τ is an element in the partition $\mathcal{T}_h$.

We shall denote by $(X_{h,r}, Q_{h,r-1})$ the so-called Hood–Taylor element, where

$$X_{h,r} = \left(S^0_{h,r}\right)^d, \quad Q_{h,r-1} = S_{h,r-1} \cap L^2(\Omega_h)/\mathbb{R}, \qquad r \geq 3.$$

For this mixed element a uniform inf-sup condition is satisfied (see [8]); that is, there exists a constant $\beta > 0$ independent of the mesh grid size h such that

$$\inf_{q_h \in Q_{h,r-1}} \sup_{v_h \in X_{h,r}} \frac{(q_h, \nabla \cdot v_h)}{\|v_h\|_1 \|q_h\|_{L^2/\mathbb{R}}} \geq \beta. \tag{2.6}$$

The approximate velocity belongs to the discretely divergence-free space

$$V_{h,r} = X_{h,r} \cap \left\{ \chi_h \in H^1_0(\Omega_h) \mid (q_h, \nabla \cdot \chi_h) = 0 \quad \forall q_h \in Q_{h,r-1} \right\}.$$

We observe that, for the Hood–Taylor element, $V_{h,r}$ is not a subspace of V.

Let $\Pi_h : L^2(\Omega)^d \longrightarrow V_{h,r}$ be the discrete Leray's projection defined by

$$(\Pi_h u, \chi_h) = (u, \chi_h) \quad \forall \chi_h \in V_{h,r}.$$

We will use the following well known bounds

$$\|(I - \Pi_h)u\|_j \leq Ch^{l-j} \|u\|_l, \qquad 1 \leq l \leq r, \quad j = 0, 1, \tag{2.7}$$

and, also, since we are assuming that Ω is at least $\mathcal{C}^2$.

$$\left\|A^{-m/2}\Pi(I - \Pi_h)u\right\|_0 \leq Ch^{l+\min(m,r-2)} \|u\|_l, \qquad 1 \leq l \leq r, \quad m = 1, 2, \tag{2.8}$$

We will denote by $A_h : V_{h,r} \to V_{h,r}$ the discrete Stokes operator defined by

$$(\nabla v_h, \nabla \phi_h) = (A_h v_h, \phi_h) = \left(A_h^{1/2} v_h, A_h^{1/2} \phi_h\right) \qquad \forall v_h, \phi_h \in V_{h,r}.$$

Let $\mathcal{A}$ denote either $\mathcal{A} = A$ or $\mathcal{A} = A_h$. Notice that both are positive self-adjoint operators with compact resolvent in H and V_h, respectively. Let us consider then for $\alpha \in \mathbb{R}$ and $t > 0$ the operators $\mathcal{A}^\alpha$ and $e^{-t\mathcal{A}}$, which are easily defined by means of the spectral properties of $\mathcal{A}$ (see, e.g., [10, p. 33], [24]). An easy calculation shows that

$$\|\mathcal{A}^\alpha e^{-t\mathcal{A}}\|_0 \leq (\alpha e^{-1})^\alpha t^{-\alpha}, \qquad \alpha \geq 0, \ t > 0, \tag{2.9}$$

where, here and in what follows, $\|\cdot\|_0$ when applied to an operator denotes the associated operator norm. Also, using the change of variables $\tau = s/t$, it is easy to show that

$$\int_0^t s^{-1/2} \left\|A_h^{1/2} e^{-(t-s)A_h}\right\|_0 ds \leq \frac{1}{\sqrt{2e}} B\left(\frac{1}{2}, \frac{1}{2}\right), \tag{2.10}$$

where B is the Beta function (see, e.g., [12]).

Let $(u,p) \in (H^2(\Omega)^d \cap V) \times (H^1(\Omega)/\mathbb{R})$ be the solution of a Stokes problem with right-hand side g, and we will denote by $S_h(u) \in V_{h,r}$ the so-called Stokes projection defined by (see [37])

$$(\nabla S_h(u), \nabla \chi_h) = (\nabla u, \nabla \chi_h) - (p, \nabla \cdot \chi_h) = (g, \chi_h) \quad \forall \chi_h \in V_{h,r}.$$

We will also consider the above definition of $S_h(u)$ written in mixed form: Find $(s_h, q_h) \in (X_h, Q_h)$ such that

$$(\nabla s_h, \nabla \phi_h) + (\nabla q_h, \phi_h) = (g, \phi_h) \quad \forall \phi_h \in X_h, \tag{2.11}$$

$$(\nabla \cdot s_h, \psi_h) = 0 \quad \forall \psi_h \in Q_h. \tag{2.12}$$

Obviously, $s_h = S_h(u)$. The following bound holds for $2 \le l \le r$:

$$\|u - s_h\|_0 + h\|u - s_h\|_1 \le Ch^l \left(\|u\|_l + \|p\|_{H^{l-1}/\mathbb{R}} \right). \tag{2.13}$$

The proof of (2.13) for $\Omega = \Omega_h$ can be found in [37]. For the general case superparametric approximation at the boundary is assumed [5]. Under the same conditions, the bound for the pressure is [29]

$$\|p - q_h\|_{L^2/\mathbb{R}} \le C_\beta h^{l-1} \left(\|u\|_l + \|p\|_{H^{l-1}/\mathbb{R}} \right), \tag{2.14}$$

where the constant C_β depends on the constant β in the inf-sup condition (2.6).

We will assume that the domain Ω is of class $\mathcal{C}^r$, so that standard bounds for the Stokes problem [5], [25] imply that

$$\left\| A^{-1} \Pi g \right\|_{2+j} \le \|g\|_j, \qquad -1 \le j \le r-2. \tag{2.15}$$

Then, using standard duality arguments and (2.13), it is easy to show that

$$\|u - s_h\|_{-m} \le Ch^{l+\min(m, r-2)} (\|u\|_l + \|p\|_{H^{l-1}/\mathbb{R}}), \quad m = 1, 2. \tag{2.16}$$

In what follows we will apply the above estimates to the particular case in which (u,p) is the solution of the Navier–Stokes problem (1.1)–(1.2). In that case s_h is the discrete velocity in problem (2.11)–(2.12) with $g = f - u_t - (u \cdot \nabla u)$. Note that the temporal variable t appears here merely as a parameter, and then, taking the time derivative, the error bounds (2.13), (2.16) can also be applied to the time derivative of s_h changing u, p by u_t, p_t.

Since we are assuming that Ω is of class $\mathcal{C}^r$ and $r \ge 3$, from (2.13) and standard bounds for the Stokes problem [5], [25], we deduce that

$$\|(A^{-1}\Pi - A_h^{-1}\Pi_h) f\|_j \le Ch^{2-j} \|f\|_0 \quad \forall f \in L^2(\Omega)^d, \quad j = 0, 1, \tag{2.17}$$

$$\|(A^{-1}\Pi - A_h^{-1}\Pi_h) f\|_1 \le Ch^2 \|f\|_1 \quad \forall f \in H^1(\Omega)^d. \tag{2.18}$$

In our analysis we shall frequently use the following relations for $f \in L^2(\Omega)^d$:

$$\|A_h^{-s/2} \Pi_h f\|_0 \le Ch^s \|f\|_0 + \|A^{-s/2} \Pi f\|_0, \quad s = 1, 2, \tag{2.19}$$

$$\|A^{-s/2} \Pi f\|_0 \le Ch^s \|f\|_0 + \|A_h^{-s/2} \Pi_h f\|_0, \quad s = 1, 2, \tag{2.20}$$

which are a consequence of (2.17) and the fact that for $v \in V$ and $v_h \in V_{h,r}$ we have $\|\nabla v\|_0 = \|A^{1/2}v\|_0$ and $\|\nabla v_h\|_0 = \|A_h^{1/2}v_h\|_0$. Notice also that, since $(A_h^{-1/2}\Pi_h f, v_h) = (f, A_h^{-1/2}v_h)$, for all $v_h \in V_{h,r}$, it follows that

$$\|A_h^{-1/2}\Pi_h f\|_0 \le C\|f\|_{-1}. \tag{2.21}$$

Finally, we will use the following inequality whose proof can be obtained by applying [36, Lemma 4.4]:

$$\|v_h\|_\infty \le C\,\|A_h v_h\|_0 \quad \forall v_h \in V_{h,r}. \tag{2.22}$$

Remark 2.1. We remark that our analysis applies also to any pair of LBB-stable mixed finite elements satisfying (2.7), (2.13), and especially (2.16). This is the case, for example, of the Crouzeix–Raviart element when $r \ge 3$ [11], [29]. In the case of low-order LBB-stable elements, the analysis is much simpler; see Remark 4.2 below. However, since to simplify the analysis we make use of several results from [6] which are stated and proved for Hood–Taylor elements, we will restrict ourselves to these elements in the present paper.

3. The postprocessed method. The postprocessing technique can be seen as a two-level method. In the first level, the mixed finite-element approximation to (1.1)–(1.2) for a given partition $\mathcal{T}_h$ of Ω_h is computed. That is, given $u_h(0) = \Pi_h(u_0)$, we compute $u_h(t) \in X_{h,r}$ and $p_h(t) \in Q_{h,r-1}$, $t \in (0,T]$, satisfying

$$(\dot u_h, \phi_h) + (\nabla u_h, \nabla \phi_h) + b(u_h, u_h, \phi_h) + (\nabla p_h, \phi_h) = (f, \phi_h) \quad \forall\, \phi_h \in X_{h,r}, \tag{3.1}$$

$$(\nabla \cdot u_h, \psi_h) = 0 \qquad \forall\, \psi_h \in Q_{h,r-1}. \tag{3.2}$$

In the second level, the discrete velocity and pressure $(u_h(t), p_h(t))$ are postprocessed by solving the following discrete Stokes problem: Find $(\tilde u_h(t), \tilde p_h(t)) \in (\widetilde X, \widetilde Q)$ satisfying

$$\big(\nabla \tilde u_h(t), \nabla \tilde\phi\big) + \big(\nabla \tilde p_h(t), \tilde\phi\big) = (f, \tilde\phi) - b(u_h(t), u_h(t), \tilde\phi) - (\dot u_h(t), \tilde\phi) \quad \forall\, \tilde\phi \in \widetilde X, \tag{3.3}$$

$$\big(\nabla \cdot \tilde u_h(t), \tilde\psi\big) = 0 \quad \forall\, \tilde\psi \in \widetilde Q, \tag{3.4}$$

where $(\widetilde X, \widetilde Q)$, is either

(a) the same-order Hood–Taylor element over a finer grid (that is, for $h' < h$, we choose $(\widetilde X, \widetilde Q) = (X_{h',r}, Q_{h',r-1})$), or

(b) a higher-order Hood–Taylor element over the same grid. In this case we choose $(\widetilde X, \widetilde Q) = (X_{h,r+1}, Q_{h,r})$.

In both cases, we will denote by $\widetilde V$ the corresponding discretely divergence-free space that can be either $\widetilde V = V_{h',r}$ or $\widetilde V = V_{h,r+1}$ depending on the selection of the postprocessing space. The discrete orthogonal projection into $\widetilde V$ will be denoted by $\widetilde\Pi_h$, and we will represent by $\widetilde A_h$ the discrete Stokes operator acting on functions in $\widetilde V$.

We remark that in (3.3)–(3.4) the time variable appears merely as a parameter. Thus, in practice, $(\tilde u_h, \tilde p_h)$ may be computed only for those $t \in (0,T]$ where improved accuracy is desired, which are usually a small set of selected times. Nevertheless, here we obtain error estimates for $t \in (0,T]$ since this adds no further difficulty.

Section 5 is devoted to the analysis of convergence of the postprocessed MFE method. In Theorems 5.2 and 5.3 the rate of convergence of the postprocessed method

is obtained. We notice that the order of approximation is optimal with respect to the pair $(\widetilde{X}, \widetilde{Q})$ in which the standard MFE approximation is postprocessed. Comparing with the (optimal) rate of convergence of the MFE method (see Corollaries 4.8, 4.16, and 4.19), we deduce that in case b an increase in the rate of convergence of one unit is achieved. Similar conclusions can be reached in case a by taking h' small enough.

The analysis of the postprocessed method will be done in two steps. We will first obtain superconvergence bounds for the MFE approximation in Theorems 4.7, 4.15, and 4.18 with respect to an auxiliary approximation to be introduced in section 4.1. In Theorems 4.7 and 4.15 we deal with the velocity, for the quadratic and cubic approximations, respectively, and with the pressure in Theorem 4.18. In the second step, the errors bounds of the postprocessed approximation are given in Theorem 5.2 for the velocity and in Theorem 5.3 for the pressure.

4. Analysis of the MFE approximation.

4.1. An auxiliary approximation. For a u and p solution of (1.1)–(1.2) let us consider the approximations $v_h : [0,T] \longrightarrow X_{h,r}$ and $g_h : [0,T] \longrightarrow Q_{h,r-1}$, respectively, solutions of

$$(\dot{v}_h, \phi_h) + (\nabla v_h, \nabla \phi_h) + (\nabla g_h, \phi_h) = (f, \phi_h) - b(u, u, \phi_h) \quad \forall \phi_h \in X_{h,r}, \tag{4.1}$$

$$(\nabla \cdot v_h, \psi_h) = 0 \qquad \forall\, \psi_h \in Q_{h,r-1}, \tag{4.2}$$

with initial condition $v_h(0) = \Pi_h u_0$.

Let us observe that the MFE approximation u_h and the recently defined v_h satisfy

$$\dot{u}_h + A_h u_h + B_h(u_h, u_h) = \Pi_h f, \qquad u_h(0) = \Pi_h u_0, \tag{4.3}$$

$$\dot{v}_h + A_h v_h + B_h(u, u) = \Pi_h f, \qquad v_h(0) = \Pi_h u_0, \tag{4.4}$$

respectively, where $B_h(u,v) = \Pi_h F(u,v)$. Then $e_h = v_h - u_h$ satisfies

$$\dot{e}_h + A_h e_h = B_h(u_h, u_h) - B_h(u, u), \qquad e_h(0). \tag{4.5}$$

We will also use the following notation:

$$z_h = \Pi_h u - v_h, \qquad \theta_h = \Pi_h u - s_h.$$

It is easy to show then that

$$\dot{z}_h + A_h z_h = A_h \theta_h, \quad z_h(0) = 0. \tag{4.6}$$

Some technical lemmas are stated in this section. For the convenience of the reader, we will reproduce here the following two lemmas from [6].

LEMMA 4.1. *Let $v \in (H^2(\Omega))^d \cap V$. Then there exists a constant $K = K(\|v\|_2)$ such that for all $w \in H_0^1(\Omega)^d$ we have that*

$$\left\|A^{-1}\Pi[F(v,v) - F(w,w)]\right\|_0 \leq K\big(\|v-w\|_{-1} + \|v-w\|_1 \|v-w\|_0\big). \tag{4.7}$$

LEMMA 4.2. *For any $f \in C([0,T]; L^2(\Omega)^d)$, the following estimate holds for all $t \in [0,T]$:*

$$\int_0^t \left\|A_h e^{-(t-s)A_h} \Pi_h f(s)\right\|_0 ds \leq C|\log(h)| \max_{0\leq t\leq T} \|f(t)\|_0.$$

LEMMA 4.3 (estimates for z_h). *Let (u,p) be the solution of (1.1)–(1.2). Then there exists a positive constant C such that the error $z_h = \Pi_h u - v_h$ of the discrete velocity v_h defined by (4.4) satisfies the following bounds:*

$$\|A_h^{-j/2} z_h\|_0 \le Ch^{2+j}, \qquad j = 0,1,2, \qquad r \ge 3, \tag{4.8}$$

$$\|A_h^{(-1+j)/2} z_h\|_0 \le \frac{C}{t^{1/2}} h^{4-j}, \qquad j = -1,0,1, \qquad r \ge 4, \tag{4.9}$$

$$\|A_h^{(-1+j)/2} z_h\|_0 \le \frac{C}{t^{1/2}} h^{4-j}, \qquad j = 0,1,2, \qquad r \ge 3, \tag{4.10}$$

$$\|A_h^{(-1+j)/2} z_h\|_0 \le \frac{C}{t} h^{5-j}, \qquad j = 0,1,2, \qquad r \ge 4, \tag{4.11}$$

$$\int_0^t \|z_h(s)\|_j^2 \, ds \le Ch^{6-2j}, \qquad j = 0,1, \qquad r \ge 3. \tag{4.12}$$

Proof. From (4.6) it follows that

$$z_h(t) = \int_0^t e^{-A_h(t-s)} A_h \theta_h(s) \, ds. \tag{4.13}$$

If we multiply both sides of this equation by A_h^{-1}, then using (2.10), (2.8), and (2.16) we have

$$\|A_h^{-1} z_h\|_0 \le CB\left(\frac{1}{2}, \frac{1}{2}\right) \max_{0\le s\le t} s^{1/2} \|A_h^{-1/2}\theta_h(s)\|_0 \le Ch^4 \max_{0\le s\le t} s^{1/2} \|u(s)\|_3,$$

which, applying estimates (2.3), proves (4.8) for the case $j = 2$. Since $\|A_h^{-1/2} z_h\|_0 \le Ch^{-1}\|A_h^{-1} z_h\|_0$ and $\|z_h\|_0 \le Ch^{-2}\|A_h^{-1} z_h\|_0$, the cases $j = 1$ and $j = 0$ are readily deduced, and then (4.8) is concluded.

We will now prove (4.9). Multiplying (4.6) by $t^{1/2} A_h^{-1/2}$, it is easy to see that $y_h = t^{1/2} A_h^{-1/2} z_h$ satisfies

$$\dot{y}_h + A_h y_h = t^{1/2} A_h^{1/2} \theta_h + (2t)^{-1/2} A_h^{-1/2} z_h,$$

so that we have

$$y_h(t) = \int_0^t e^{-A_h(t-s)} s^{1/2} A_h^{1/2} \theta_h(s) \, ds + \frac{1}{2} \int_0^t e^{-A_h(t-s)} s^{-1/2} A_h^{-1/2} z_h \, ds. \tag{4.14}$$

Then, using (2.10), (2.8), and (2.16), it follows that

$$\begin{aligned} t^{1/2}\|A_h^{-1} z_h\|_0 &\le CB\left(\frac{1}{2}, \frac{1}{2}\right)\left(\max_{0\le s\le t} s\|A_h^{-1/2}\theta_h\|_0 + \max_{0\le s\le t} \|A_h^{-3/2} z_h\|_0\right) \\ &\le C\left(M_4 h^5 + \max_{0\le s\le t} \|A_h^{-3/2} z_h\|_0\right). \end{aligned}$$

In order to estimate the last term on the right-hand side above, we multiple both sides of (4.13) by $A_h^{-3/2}$, so that using (2.10), (2.8), and (2.16) we obtain

$$\|A_h^{-3/2} z_h\|_0 \le CB\left(\frac{1}{2}, \frac{1}{2}\right) \max_{0\le s\le t} s^{1/2} \|A_h^{-1}\theta_h(s)\|_0 \le Ch^5 M_3,$$

and then (4.9) is proved in the case $j=-1$, from which, using the same arguments as with (4.8), the cases $j=0,1$ are inferred.

We will now prove (4.10) and (4.11) for the case $j=0$. As before, the cases $j=1$ and $j=2$ follow immediately. For $y_h(t)=tA_h^{-1/2}z_h(t)$, it is easy to see that

$$y_h(t)=\int_0^t e^{-A_h(t-s)}sA_h^{1/2}\theta_h(s)\,ds+\int_0^t e^{-A_h(t-s)}A_h^{-1/2}z_h(s)\,ds,$$

so that, integrating by parts, we have

$$\begin{aligned}y_h(t)&=tA_h^{-1/2}\theta_h(t)-\int_0^t e^{-A_h(t-s)}A_h^{-1/2}\big(\theta_h(s)+s\dot\theta_h(s)\big)\,ds\\&\quad+\int_0^t e^{-A_h(t-s)}A_h^{-1/2}z_h(s)\,ds.\end{aligned}\tag{4.15}$$

We deal first with (4.10). For the last term on the right-hand side above, by writing $e^{-A_h(t-s)}A_h^{-1/2}z_h(s)=A_h^{1/2}e^{-A_h(t-s)}A_h^{-1}z_h(s)$, then, thanks to (2.9) and (4.8), it follows that

$$\left\|\int_0^t e^{-A_h(t-s)}A_h^{-1/2}z_h(s)\,ds\right\|_0\le Ct^{1/2}\max_{0\le s\le t}\|A_h^{-1}z_h(s)\|_0\le Ct^{1/2}h^4.\tag{4.16}$$

For the second term on the right-hand side of (4.15), we start by writing $e^{-A_h(t-s)}=e^{-A_h(t-s)}s^{-1/2}s^{1/2}$, and then, since, trivially, $\|s^{-1/2}e^{-A_h(t-s)}\|_0\le s^{-1/2}$, we have

$$\begin{aligned}&\left\|\int_0^t e^{-A_h(t-s)}A_h^{-1/2}(\theta_h(s)+s\dot\theta_h(s))\,ds\right\|_0\\&\quad\le 2t^{1/2}\max_{0\le s\le t}s^{1/2}\big(\|A_h^{-1/2}(\theta_h(s)+s\dot\theta_h(s))\|_0\big)\\&\quad\le Ct^{1/2}h^4\max_{0\le s\le t}s^{1/2}\big(\|u(s)\|_3+s\,\|u_s(s)\|_3\big)\le Ct^{1/2}h^4(M_3+M_5),\end{aligned}\tag{4.17}$$

where in the last inequality we have applied (2.8) and (2.16). The same argument shows that the first term on the right-hand side of (4.15) can be bounded by $Ct^{1/2}h^4M_3$, which together with (4.16) and (4.17) finishes the proof of (4.10).

In order to prove (4.11), we go back to (4.15). Using (2.10) to estimate both integrals on the right-hand side of (4.15), we may write

$$\begin{aligned}\|y_h(t)\|_0&\le t\big\|A_h^{-1/2}\theta_h(t)\big\|_0\\&\quad+CB\left(\frac12,\frac12\right)\Big(\max_{0\le s\le t}s^{1/2}\big\|A_h^{-1}(\theta_h(s)+s\dot\theta_h(s))\big\|_0+\max_{0\le s\le t}s^{1/2}\|A_h^{-1}z_h(s)\|_0\Big),\end{aligned}$$

so that, thanks to (2.8), (2.16), and (4.9), it follows that (4.11) holds for $j=0$.

It only remains to prove (4.12) in the case $j=0$ (the case $j=1$ is deduced by applying inverse inequality (2.5)). From (4.6) we get

$$(A_h^{-1}\dot z_h,z_h)+(z_h,z_h)=(\theta_h,z_h).$$

Then

$$\frac{d}{dt}\|A_h^{-1/2}z_h\|_0^2 + \|z_h\|_0^2 \le \|\theta_h\|_0^2.$$

Finally, by integrating with respect to t and applying (2.7) and (2.13) we obtain

$$\int_0^t \|z_h(s)\|_0^2\, ds \le \int_0^t \|\theta_h(s)\|_0^2\, ds \le h^6 \int_0^t \|u(s)\|_3^2\, ds \le K_3^2 h^6,$$

which finishes the proof. ∎

4.2. Superconvergence for the velocity: $r = 3$. We need several auxiliary lemmas before proving Theorem 4.7. We start with the continuity of the nonlinear term in several different norms.

LEMMA 4.4. *For each $\alpha > 0$ there exists a constant $K > 0$ depending on α and M_2 such that, for every $w_h^1(\cdot), w_h^2(\cdot) \in V_{h,r}$ satisfying the threshold condition*

$$\|u(t) - w_h^1(t)\|_j \le \alpha h^{2-j}, \quad \|u(t) - w_h^2(t)\|_j \le \alpha h^{2-j}, \quad j = 0, 1,\ t \in [0, T], \tag{4.18}$$

the following bounds hold:

$$\|F(w_h^1, w_h^1) - F(w_h^2, w_h^2)\|_0 \le K \|w_h^1 - w_h^2\|_1, \tag{4.19}$$

$$\|F(w_h^1, w_h^1) - F(w_h^2, w_h^2))\|_{-1} \le K \|w_h^1 - w_h^2\|_0, \tag{4.20}$$

$$\|B_h(w_h^1, w_h^1) - B_h(w_h^2, w_h^2)\|_0 \le K \|w_h^1 - w_h^2\|_1, \tag{4.21}$$

$$\|A_h^{-1/2}(B_h(w_h^1, w_h^1) - B_h(w_h^2, w_h^2))\|_0 \le K \|w_h^1 - w_h^2\|_0, \tag{4.22}$$

$$\|A_h^{-1}(B_h(w_h^1, w_h^1) - B_h(w_h^2, w_h^2))\|_0 \le K\|A_h^{-1/2}(w_h^1 - w_h^2)\|_0. \tag{4.23}$$

Proof. The bounds (4.19) and (4.20) are obtained by arguing exactly as in the proof of [6, Lemma 3.1]. Using the stability of Π_h and (2.21), the proofs of (4.21) and (4.22) are easily deduced from (4.19) and (4.20).

Finally, let us prove (4.23). Let us write $w_h = w_h^1 - w_h^2$. Taking into account that

$$h^2\|w_h\|_1 \le C\|A_h^{-1/2}w_h\|_0 \tag{4.24}$$

and using (2.19) and (4.21), it follows that

$$\begin{aligned}\|A_h^{-1}(B_h(w_h^1, w_h^1) - B_h(w_h^2, w_h^2))\|_0 \le{}& \|A^{-1}\Pi(F(w_h^1, w_h^1) - F(w_h^2, w_h^2))\|_0 \\ &+ C\|A_h^{-1/2}w_h\|_0.\end{aligned} \tag{4.25}$$

We now estimate the first term on the right-hand side above by duality, since

$$\|A^{-1}\Pi(F(w_h^1, w_h^1) - F(w_h^2, w_h^2))\|_0 \le \max_{\phi\in H, \phi\neq 0} \frac{|(F(w_h^1, w_h^1) - F(w_h^2, w_h^2), A^{-1}\phi)|}{\|\phi\|_0}.$$

For this purpose, using the skew-symmetry property of b, we write

$$\begin{aligned}(F(w_h^1, w_h^1) - F(w_h^2, w_h^2), A^{-1}\phi) &= b(w_h, w_h^2, A^{-1}\phi) + b(w_h^1, w_h, A^{-1}\phi) \\ &= b(w_h, u(t), A^{-1}\phi) - b(u(t), A^{-1}\phi, w_h) \\ &\quad + b(w_h, A^{-1}\phi, u(t) - w_h^2) + b(u(t) - w_h^1, A^{-1}\phi, w_h).\end{aligned} \tag{4.26}$$

We now notice that $|b(w_h, A^{-1}\phi, u(t) - w_h^2)| \le C \|w_h\|_1 \|\phi\|_0 \|u(t) - w_h^2\|_0$, so that, thanks to (4.24) and (4.18), we have

$$\left|b(w_h, A^{-1}\phi, u(t) - w_h^2)\right| \le C \|w_h\|_1 h^2 \|\phi\|_0 \le C\left\|A_h^{-1/2} w_h\right\|_0 \|\phi\|_0 .$$

Similarly, $|b(u(t) - w_h^1, A^{-1}\phi, w_h)| \le C \|w_h\|_0 \|\phi\|_0 \|u(t) - w_h^1\|_1$. Since $h\|w_h\|_0 \le C\|A_h^{-1/2} w_h\|_0$, it follows that

$$\left|b(u(t) - w_h^1, A^{-1}\phi, w_h)\right| \le Ch \|w_h\|_0 \|\phi\|_0 \le C\left\|A_h^{-1/2} w_h\right\|_0 \|\phi\|_0 ,$$

so that from (4.26) we get

$$\begin{aligned}\left|(F(w_h^1, w_h^1) - F(w_h^2, w_h^2), A^{-1}\phi)\right| \le& \left|b(w_h, u(t), A^{-1}\phi)\right| + \left|b(u(t), A^{-1}\phi, w_h)\right| \\ &+ C\left\|A_h^{-1/2} w_h\right\|_0 \|\phi\|_0 .\end{aligned} \tag{4.27}$$

Finally, since by applying the divergence theorem we have

$$b(w_h, u(t), A^{-1}\phi) = (w_h, \nabla u(t) \cdot A^{-1}\phi) - \frac{1}{2}(w_h, \nabla(u(t) \cdot A^{-1}\phi)),$$

it follows that

$$\left|b(w_h, u(t), A^{-1}\phi)\right| \le C \|w_h\|_{-1} \left(\left\|\nabla u(t) \cdot A^{-1}\phi\right\|_1 + \left\|\nabla(A^{-1}\phi \cdot u(t))\right\|_1\right), \tag{4.28}$$

$$\left|b(u(t), A^{-1}\phi, w_h)\right| \le C \|w_h\|_{-1} \left(\left\|(u(t) \cdot \nabla)A^{-1}\phi\right\|_1 + \left\|(\nabla \cdot u(t))A^{-1}\phi\right\|_1\right). \tag{4.29}$$

Since

$$\begin{aligned}\left\|\nabla u \cdot A^{-1}\phi\right\|_1 &\le C\left(\|u\|_2 \left\|A^{-1}\phi\right\|_{L^\infty} + \|u\|_{W^{1,4}} \left\|A^{-1}\phi\right\|_{W^{1,4}}\right) \\ &\le C \|u\|_2 \left\|A^{-1}\phi\right\|_2 \le C\|u\|_2\|\phi\|_0,\end{aligned}$$

arguing similarly with the rest of the terms on the right-hand side of (4.28)–(4.29), from (4.25) and (4.27) the proof is easily finished. ∎

Remark 4.1. In what follows we will apply Lemma 4.4 to functions $w_h^1 = v_h$, satisfying (4.4), and $w_h^2 = u_h$, satisfying (4.3). Let us now check that the threshold condition (4.18) is satisfied. We first observe that $\|v_h - u\|_0 \le \|z_h\|_0 + \|u - \Pi_h u\|_0$ so that by applying (4.8) and (2.7) we get $\|v_h - u\|_0 \le Ch^2$, and then v_h satisfies (4.18). For the bound $\|u - u_h\|_0 \le Ch^2$, we refer the reader to [36, Theorem 3.1].

LEMMA 4.5. *Let (u, p) be the solution of* (1.1)–(1.2). *Then there exists a positive constant C such that the discrete velocity v_h defined by* (4.4) *satisfies*

$$\|A_h^{-1}(B_h(v_h(t), v_h(t)) - B_h(u(t), u(t)))\|_0 \le \frac{C}{t^{(r-2)/2}} h^{r+1}, \quad t \in (0, T], \quad r = 3, 4.$$

Proof. Let us write $\rho_h = B_h(v_h, v_h) - B_h(u, u)$. By applying (2.19) we have

$$\|A_h^{-1}\rho_h\|_0 \le Ch^2 \|F(v_h, v_h) - F(u, u)\|_0 + \|A^{-1}\Pi(F(v_h, v_h) - F(u, u))\|_0.$$

Using then (4.19) from Lemmas 4.4 and 4.1 we get

$$\|A_h^{-1}\rho_h\|_0 \le Ch^2 \|v_h - u\|_1 + C(\|v_h - u\|_1 \|v_h - u\|_0 + C\|v_h - u\|_{-1}).$$

Applying now the standard bounds for Π_h (see (2.7), (2.8)) together with the estimates (4.8), (4.10) for z_h in Lemma 4.3 when $r = 3$ and (4.11), (4.8) when $r = 4$, the conclusion is reached. ∎

LEMMA 4.6. *Let (u, p) be the solution of* (1.1)–(1.2). *Then there exists a positive constant C such that the discrete velocity v_h defined by* (4.4) *and the Hood–Taylor element approximation to u, u_h, satisfy the following bound:*

$$\|A_h^{-1/2}(v_h(t) - u_h(t))\|_0 \le Ch^4, \quad r \ge 3, \quad t \in (0, T].$$

Proof. Let us consider $y_h(t) = A_h^{-1/2} e_h(t)$ and $\rho_h = B_h(v_h, v_h) - B_h(u, u)$. From (4.5) it follows that

$$y_h(t) = \int_0^t e^{-(t-s)A_h} A_h^{-1/2} \big(B_h(u_h, u_h) - B_h(v_h, v_h)\big)\, ds + \int_0^t e^{-(t-s)A_h} A_h^{-1/2} \rho_h(s)\, ds.$$

Applying (2.9) and (4.23) we have that

$$\|y_h(t)\|_0 \le \int_0^t \frac{C}{\sqrt{t-s}} \|y_h(s)\|_0 \, ds + \left\| \int_0^t e^{-(t-s)A_h} A_h^{-1/2} \rho_h(s)\, ds \right\|_0,$$

so that a generalized Gronwall lemma [35, pp. 188–189] allows us to write

$$\max_{0 \le t \le T} \|y_h(t)\|_0 \le C \max_{0 \le t \le T} \left\| \int_0^t e^{-(t-s)A_h} A_h^{-1/2} \rho_h(s)\, ds \right\|_0.$$

Using (2.10) we obtain

$$\max_{0 \le t \le T} \|y_h(t)\|_0 \le CB\left(\frac{1}{2}, \frac{1}{2}\right) \max_{0 \le s \le T} s^{1/2} \big\| A_h^{-1} \big(B_h(v_h(s), v_h(s)) - B_h(u(s), u(s))\big) \big\|_0,$$

where, by applying Lemma 4.5 with $r = 3$, the proof is finished. ∎

THEOREM 4.7 (superconvergence for the velocity). *Let (u, p) be the solution of* (1.1)–(1.2). *There exists a positive constant C such that the discrete velocity v_h defined by* (4.4) *and the Hood–Taylor element approximation to u, u_h, satisfy the following bound:*

$$\|v_h(t) - u_h(t)\|_j \le \frac{C}{t^{1/2}} |\log(h)| h^{4-j}, \quad t \in (0, T], \quad j = 0, 1, \quad r \ge 3. \tag{4.30}$$

Proof. We prove the error bound for the L^2 norm, from which the bound in the H^1 norm is readily obtained by applying the inverse inequality (2.5). Let us consider $y_h(t) = t^{1/2} e_h(t)$. An easy calculation shows that

$$\dot{y}_h + A_h y_h + t^{1/2}(B_h(v_h, v_h) - B_h(u_h, u_h)) = t^{1/2} \rho_h + \frac{1}{2t^{1/2}} e_h,$$

where $\rho_h = B_h(v_h, v_h) - B_h(u, u)$, and, thus,

$$y_h(t) = \int_0^t e^{-A_h(t-s)} s^{1/2} (B_h(u_h, u_h) - B_h(v_h, v_h))\, ds + \int_0^t e^{-A_h(t-s)} \left(s^{1/2} \rho_h + \frac{1}{2s^{1/2}} e_h \right) ds.$$

For the first term on the right-hand side above, using (2.9) and (4.22) we have

$$\left\| \int_0^t e^{-A_h(t-s)} s^{1/2} (B_h(u_h,u_h) - B_h(v_h,v_h))\, ds \right\|_0 \leq C \int_0^t \frac{\|y_h\|_0}{\sqrt{t-s}}\, ds,$$

so that by applying a generalized Gronwall lemma [35, pp. 188–189], it follows that

$$\max_{0\leq t\leq T} \|y_h(t)\|_0$$
$$\leq C\left(\max_{0\leq t\leq T} \left\| \int_0^t e^{-A_h(t-s)} s^{1/2} \rho_h\, ds \right\|_0 + \max_{0\leq t\leq T} \left\| \int_0^t e^{-A_h(t-s)} \frac{e_h}{s^{1/2}}\, ds \right\|_0 \right).$$

Applying now Lemma 4.2 and (2.10) we have

$$\max_{0\leq t\leq T} \|y_h(t)\|_0$$
$$\leq C|\log(h)| \max_{0\leq s\leq T} \left\| s^{1/2} A_h^{-1} \rho_h(s) \right\|_0 + CB\left(\frac{1}{2},\frac{1}{2}\right) \max_{0\leq s\leq T} \left\| A_h^{-1/2} e_h(s) \right\|_0,$$

where Lemmas 4.5 and 4.6 finish the proof. ∎

COROLLARY 4.8. *Let (u,p) be the solution of* (1.1)–(1.2). *Then, for $r \geq 3$, there exists a positive constant C such that the Hood–Taylor element approximation to u, u_h, satisfies the following bound for $j = 0, 1$:*

$$\|u_h(t) - u(t)\|_j \leq \frac{C}{t^{1/2}} h^{3-j}, \quad t \in (0,T].$$

Proof. By rewriting $u - u_h = (u - \Pi_h u) + (\Pi_h u - v_h) + (v_h - u_h)$, rewriting $u - u_h = (u - \Pi_h u) + (\Pi_h u - v_h) + (v_h - u_h)$, and applying (2.7), Lemma 4.3, and Theorem 4.7, the corollary is proved. ∎

Remark 4.2. Notice that by applying Lemma 4.2 to (4.13), for the z_h solution of (4.6), we obtain

$$\|z_h(t)\|_0 \leq C M_2 h^2 |\log(h)|. \tag{4.31}$$

With this estimate and a much simpler analysis than the previous one, it is possible to prove superconvergence results in the H^1 norm for low-order LBB-stable pairs of mixed finite-element methods, such as the so-called mini element [3]. More precisely, for the e_h solution of (4.5), the following bound holds:

$$\max_{0\leq t\leq T} \|e_h(t)\|_1 \leq C |\log(h)|^2 h^2,$$

with C depending only on M_2.

4.3. Superconvergence for the velocity: $r = 4$. As before, several auxiliary lemmas are needed before Theorem 4.15. We start with a generalized Gronwall lemma.

LEMMA 4.9. *If u is a nonnegative function, continuous in $[0,T]$ and satisfying*

$$u(t) \leq u_0 + \beta \int_0^t \left(\frac{1}{(t-s)^{1/2}} + \frac{1}{s^{1/2}} \right) u(s)\, ds, \qquad 0 \leq t < T,$$

for some $u_0 > 0$ and $\beta > 0$, then

$$u(t) \leq E_{1/2}(2\beta(\pi t)^{1/2}) u_0, \qquad 0 \leq t < T,$$

where $E_\alpha(z)$ denotes the Mittag–Leffler function (see 4.36 *below).*

Proof. Let us consider the operator

$$Eu(t) = \beta \int_0^t \left(\frac{1}{(t-s)^{1/2}} + \frac{1}{s^{1/2}} \right) u(s)\, ds, \qquad 0 \le t < T.$$

We notice that

$$u(t) \le u_0 + Eu(t) \le u_0 + Eu_0 + E^2 u(t) \le (1 + E + \cdots + E^n)u_0 + E^{n+1}u(t). \tag{4.32}$$

We have that

$$Eu_0 = 4\beta t^{1/2} u_0 = (2 + B(1/2, 1))(\beta t^{1/2})u_0,$$

where $B(x, y)$ denotes the Beta function (see, e.g., [12]). We also notice that

$$\begin{aligned} E^2 u_0 &= 4\beta^2 u_0 \int_0^t \left(\frac{1}{(t-s)^{1/2}} + \frac{1}{s^{1/2}} \right) s^{1/2}\, ds \\ &= 4\beta^2 u_0 \left(t + \int_0^t \frac{s^{1/2}}{(t-s)^{1/2}}\, ds \right). \end{aligned}$$

Now taking into account that, by means of the change of variables, $s = tx$ we have

$$\int_0^t \frac{s^{1/2}}{(t-s)^{1/2}}\, ds = t \int_0^1 \frac{x^{1/2}}{(1-x)^{1/2}}\, dx = tB(1/2, 3/2),$$

and, thus,

$$E^2 u_0 = 4(1 + B(1/2, 3/2))\left(\beta t^{1/2}\right)^2 u_0 = \left(\prod_{l=1}^{2} \left(\frac{2}{l} + B\left(\frac{1}{2}, \frac{l+1}{2} \right) \right) \right) \left(\beta t^{1/2}\right)^2 u_0.$$

Assume that for $j = 2, \ldots, n$

$$E^j u_0 = \left(\prod_{l=1}^{j} \left(\frac{2}{l} + B\left(\frac{1}{2}, \frac{l+1}{2} \right) \right) \right) (\beta t^{1/2})^j, \tag{4.33}$$

Let us compute $E^{n+1}u_0$. Since $E^{n+1}u_0 = EE^n u_0$, we have

$$E^{n+1}u_0 = \beta^{n+1} u_0 \left(\prod_{l=1}^{n} \left(\frac{2}{l} + B\left(\frac{1}{2}, \frac{l+1}{2} \right) \right) \right) \int_0^t \left(\frac{1}{(t-s)^{1/2}} + \frac{1}{s^{1/2}} \right) s^{n/2}\, ds.$$

But

$$\int_0^t \left(\frac{s^{n/2}}{(t-s)^{1/2}} + s^{(n-1)/2} \right) ds = \frac{2}{n+1} t^{(n+1)/2} + \int_0^t \frac{s^{n/2}}{(t-s)^{1/2}}\, ds,$$

and, using again the change of variables $s = tx$,

$$\int_0^t \frac{s^{n/2}}{(t-s)^{1/2}}\, ds = t^{(n+1)/2} \int_0^1 \frac{x^{n/2}}{(1-x)^{1/2}}\, ds = t^{(n+1)/2} B\left(\frac{1}{2}, \frac{n}{2} + 1 \right),$$

and since $\frac{n}{2} + 1 = \frac{(n+1)+1}{2}$, it follows that (4.33) also holds for $j = n + 1$.

Since, on the one hand

$$B\left(\frac{1}{2},\frac{n+1}{2}\right)=\Gamma\left(\frac{1}{2}\right)\frac{\Gamma(\frac{n}{2}+\frac{1}{2})}{\Gamma(\frac{n}{2}+1)}, \tag{4.34}$$

where $\Gamma(z)$ denotes the Gamma function, and on the other hand (see, e.g., [12, 6.1.47])

$$1=\lim_{z\to\infty} z^{1/2}\frac{\Gamma(z)}{\Gamma(z+1/2)},$$

we have that

$$E^{n+1}u(t)\le E^{n+1}\left\|u\right\|_{L^\infty(0,T)}\to 0 \qquad \text{as } n\to\infty.$$

Thus, from (4.32) and (4.33) it follows that

$$u(t)\le u_0\sum_{n=0}^{\infty} c_n\big(\beta t^{1/2}\big)^n, \qquad 0\le t\le T, \tag{4.35}$$

where

$$c_0=1, \qquad \frac{c_n}{c_{n-1}}=\frac{2}{n}+B\left(\frac{1}{2},\frac{n+1}{2}\right), \quad n=1,3,\dots .$$

Furthermore, on the one hand, for $n=1$ we have $2/n=B(\frac{1}{2},\frac{n+1}{2})=2$, and, on the other hand, since for $x\in(0,1)$ we have $1/\sqrt{x(1-x)}\ge 2$, it follows that, for $n\ge 2$,

$$B\left(\frac{1}{2},\frac{n+1}{2}\right)=\int_0^1\frac{x^{(n-1)/2}}{\sqrt{1-x}}\,dx=\int_0^1\frac{x^{n/2}}{\sqrt{x(1-x)}}\,dx\ge 2\int_0^1 x^{n/2}\,dx=\frac{4}{n+2}\ge\frac{2}{n},$$

so that by $c_n/c_{n-1}\le 2B(\frac{1}{2},\frac{n+1}{2})$ for $n=1,2,\dots$. Thus, if we set $k_0=2\Gamma(1/2)=2\sqrt{\pi}$, then in view of (4.34) we have $c_n\le k_0\Gamma(\frac{n}{2}+\frac{1}{2})\Gamma(\frac{n}{2}+1)^{-1}c_{n-1}$, that is,

$$c_n\le k_0\frac{\Gamma(\frac{n}{2}+\frac{1}{2})}{\Gamma(\frac{n}{2}+1)}c_{n-1}\le k_0^2\frac{\Gamma(\frac{n}{2}+\frac{1}{2})}{\Gamma(\frac{n}{2}+1)}\frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n}{2}+\frac{1}{2})}c_{n-2}=k_0^2\frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n}{2}+1)}c_{n-2},$$

so that by iterating this process we have

$$c_n\le\frac{k_0^n}{\Gamma(\frac{n}{2}+1)}c_0, \qquad n=1,2,\dots .$$

Then from (4.35) it follows that

$$u(t)\le u_0\sum_{n=0}^{\infty}\frac{\big(\beta k_0 t^{1/2}\big)^n}{\Gamma(\frac{n}{2}+1)}=E_{1/2}(\beta k_0 t^{1/2})u_0,$$

where $E_\alpha(z)$ is the Mittag–Leffler function

$$E_\alpha(z)=\sum_{n=0}^{\infty}\frac{z^n}{\Gamma(n\alpha+1)}. \tag{4.36}$$ □

LEMMA 4.10. *Let $v,w\in H_0^1(\Omega)\cap H^3(\Omega)$, and let $g\in H_0^1(\Omega)$, satisfying*

$$(\nabla\cdot\ g,\chi_h)=0, \qquad \forall\chi_h\in Q_{h,r-1}, \quad r\ge 4. \tag{4.37}$$

Then there exists a positive constant C such that the following bounds hold:

$$|b(v,w,g)| \le C\big(\|A^{-1}\Pi g\|_0 + h^2\|g\|_0\big)\big(\|(v\cdot\nabla)w\|_2 + \|(\nabla\cdot v)w\|_2\big), \tag{4.38}$$

$$|b(g,w,v)| \le C\big(\|A^{-1}\Pi g\|_0 + h^2\|g\|_0\big)\|\nabla w\cdot v\|_2 + Ch^2\|g\|_0\|\nabla v\cdot w\|_2. \tag{4.39}$$

Proof. We will first prove the following result whose proof is similar to that of Lemma 3.11 in [6]. Let $v \in H_0^1(\Omega)^d$, satisfying $(\nabla\cdot v,\chi_h) = 0$ for all $\chi_h \in Q_{h,r-1}$. Then for $f \in H_0^1(\Omega)^d \cap H^s(\Omega)^d$ the following bound holds:

$$|(v,f)| \le \big(\|A^{-s/2}\Pi v\|_0 + Ch^s\|v\|_0\big)\|f\|_s, \qquad s = 1,2. \tag{4.40}$$

To prove (4.40) we decompose

$$f = \Pi f + (I-\Pi)f = \Pi f + \nabla q$$

for some $q \in H^{s+1}$. Since Leray's projector Π is continuous from $H_0^1(\Omega)\cap H^s(\Omega)$ onto H^s, we have

$$\|\Pi f\|_s \le C\|f\|_s, \qquad \|q\|_{H^{s+1}/\mathbb{R}} \le C\|f\|_s \tag{4.41}$$

for some positive constant C. Thus,

$$(v,f) = (v,\Pi f) + (v,\nabla q) = (A^{-s/2}\Pi v, A^{s/2}\Pi f) + (v,\nabla q). \tag{4.42}$$

Since, for any $\chi_h \in Q_{h,r-1}$, $|(v,\nabla q)| = |(v,\nabla(q-\chi_h))|$, by taking χ_h as the interpolant of q in $Q_{h,r-1}$, from (4.41) and (4.42) it follows that

$$|(v,f)| \le Ch^s\|v\|\|f\|_s + \big|(A^{-s/2}\Pi v, A^{s/2}\Pi f)\big|.$$

Finally, since $\|A^{s/2}\Pi f\|_0 \le C\|\Pi f\|_s \le C'\|f\|_s$, (4.40) follows.

Notice now that (4.38) is a direct consequence of (4.40). For the second bound, the same argument shows that $((g\cdot\nabla)w,v)) = (g,\nabla w\cdot v)$ can be bounded by the first term on the right-hand side of (4.39). Now by using the divergence theorem and (4.37) we have

$$\big((\nabla\cdot g)w,v\big) = -\big(g,\nabla(w\cdot v)\big) = -\big(g,\nabla(w\cdot v-\chi_h)\big) \qquad \forall\chi_h \in Q_{h,r-1}.$$

Taking χ_h as the interpolant of $(w\cdot v)$ in $Q_{h,r-1}$ and taking into account that $\nabla(w\cdot v) = \nabla w\cdot v + \nabla v\cdot w$, the proof is easily concluded. ∎

LEMMA 4.11. *There exists a positive constant C such that for any $\phi_h \in V_{h,r}$ the following bound holds for $\psi_h(t) = e^{-tA_h}A_h^{-1}\phi_h$ and $\Psi(t) = e^{-tA}A^{-1}\Pi\phi_h$:*

$$\|\Psi(t) - \psi_h(t)\|_1 \le \frac{Ch^2}{t^{1/2}}\|\phi_h\|_0, \quad t > 0. \tag{4.43}$$

Proof. Let us first observe that, since $\Psi_t = -A\Psi$, then, thanks to (2.9) and (2.15), for $t > 0$ we have

$$\|\Psi(t)\|_{2+j} \le C\|\Psi_t(t)\|_j \le Ct^{-j/2}\|\phi_h\|_0, \qquad j = 0,1, \tag{4.44}$$

$$\|\Psi_{tt}(t)\|_0 \le Ct^{-1}\|\phi_h\|_0. \tag{4.45}$$

We set $\varphi_h(t) = e^{-tA_h}\Pi_h A^{-1}\Pi\phi_h$. We decompose $\psi_h - \Psi = (\psi_h - \varphi_h) + (\varphi_h - \Psi)$, and we will bound the two terms on the right-hand side separately. For the first one, taking into account that $\psi_h(t)-\varphi_h(t) = e^{-tA_h}\Pi_h(A_h^{-1}\Pi_h - A^{-1}\Pi)\phi_h$, and using (2.9) and (2.17), we have

$$\|\psi_h(t) - \varphi_h(t)\|_1 \le \frac{C}{t^{1/2}}\left\|(A_h^{-1}\Pi_h - A^{-1}\Pi)\phi_h\right\|_0 \le \frac{Ch^2}{t^{1/2}}\|\phi_h\|_0.$$

In order to bound $\varphi_h - \Psi$, we write $\varphi_h - \Psi = \varphi_h - \Pi_h\Psi + (I-\Pi_h)\Psi$, so that, in view of (2.7) and (4.44) we only have to estimate $\varphi_h - \Pi_h\Psi$. For this purpose, on the one hand, we have $\dot\varphi_h + A_h\varphi_h = 0$. On other hand, since $\Psi_t = -A\Psi$, we have

$$\Pi_h\Psi_t + A_h\Pi_h\Psi = \Pi_h\Psi_t - A_h\Pi_h A^{-1}\Psi_t = A_h(A_h^{-1}\Pi_h - \Pi_h A^{-1})\Psi_t,$$

so that for $\epsilon_h = \varphi_h - \Pi_h\Psi$ and $\rho_h = (A_h^{-1}\Pi_h - \Pi_h A^{-1})\Psi_t$ we have

$$\dot\epsilon_h + A_h\epsilon_h = A_h\rho_h,$$

an equation which is similar to (4.6). Thus, arguing as in the proof of (4.8) we have

$$\|\epsilon_h(t)\|_0 \le C\max_{0\le s\le t} s^{1/2}\left\|A_h^{1/2}\rho_h(s)\right\|_0 \le Ch^2\left\|\phi_h\right\|_0, \tag{4.46}$$

where, in the last inequality, we have applied (2.7), (2.18), and the case $j=1$ in (4.44). Furthermore, since for $y_h(t) = A_h^{1/2}t^{1/2}\epsilon_h(t)$ it is easy to see that

$$y_h(t) = \int_0^t e^{-A_h(t-s)}s^{1/2}A_h^{3/2}\rho_h(s)\,ds + \frac{1}{2}\int_0^t e^{-A_h(t-s)}s^{-1/2}A_h^{1/2}\epsilon_h(s)\,ds,$$

so that by integrating by parts in the first integral above and using (2.10) we have

$$\|y_h(t)\|_0 \le t^{1/2}\|\rho_h(t)\|_1 + CB\left(\frac{1}{2},\frac{1}{2}\right)\left(\max_{0\le s\le t}\left\|\frac{1}{2}\rho_h(s) + s\dot\rho_h(s)\right\|_0 + \max_{0\le s\le t}\|\epsilon_h(s)\|_0\right).$$

The first and the last terms on the right-hand side above are already bounded in (4.46), whereas the second term, applying (2.7), (2.17), and (4.44)–(4.45), can also be bounded by $Ch^2\|\phi_h\|_0$, which completes the proof. ∎

LEMMA 4.12. *Let (u,p) be the solution of (1.1)–(1.2). There exists a constant $C>0$ such that for any $g\in H_0^1(\Omega)^d$ and $\phi_h\in V_{h,r}$ the following bounds hold for $\psi_h(t) = e^{-tA_h}A_h^{-1}\phi_h$, $\Psi(t) = e^{-tA}A^{-1}\Pi\phi_h$, and $0<s<t\le T$:*

$$|b(g,u(s),\psi_h(t-s))| \le |b(g,u(s),\Psi(t-s))| + C\frac{M_2}{\sqrt{t-s}}h^2\|g\|_0\|\phi_h\|_0, \tag{4.47}$$

$$|b(u(s),\psi_h(t-s),g)| \le |b(u(s),\Psi(t-s),g)| + C\frac{M_2}{\sqrt{t-s}}h^2\|g\|_0\|\phi_h\|_0. \tag{4.48}$$

Proof. We will prove only (4.47) since (4.48) can be proved by reasoning similarly. Also, we will drop the dependence on s and $t-s$. We first observe that

$$b(g,u,\psi_h) = b(g,u,\Psi) + b(g,u,\psi_h - \Psi).$$

On the one hand, we have

$$|((g\cdot\nabla)u,\psi_h - \Psi)| \le C\|g\|_0\|\nabla u\|_{L^4(\Omega)}\|\psi_h - \Psi\|_{L^4(\Omega)},$$

so that by using (2.1) and (4.43) it follows that

$$|((g\cdot\nabla)u,\psi_h-\Psi)|\le CM_2\,\|g\|_0\,\|\psi_h-\Psi\|_1\le C\frac{M_2}{\sqrt{t-s}}h^2\,\|g\|_0\,\|\phi_h\|_0\,. \tag{4.49}$$

On the other hand, since $((\nabla\cdot g)u,\psi_h-\Psi)=-(g,\nabla(u\cdot(\psi_h-\Psi)))$, we have

$$\|((\nabla\cdot g)u,(\psi_h-\Psi))\|_0\le C\,\|g\|_0\,\big(\|\nabla u\|_{L^4}\,\|\psi_h-\Psi\|_{L^4}+\|u\|_{L^\infty}\,\|\nabla(\psi_h-\Psi)\|_0\big).$$

Applying again (2.1) and (4.43) we get

$$\|((\nabla\cdot g)u,(\psi_h-\Psi))\|_0\le CM_2\,\|g\|_0\,\|\psi_h-\Psi\|_1\le C\frac{M_2}{\sqrt{t-s}}h^2\,\|g\|_0\,\|\phi_h\|_0\,,$$

which together with (4.49) allow us to conclude (4.47). ∎

LEMMA 4.13. *Let (u,p) be the solution of (1.1)–(1.2). There exists a positive constant C such that, for $r\ge 4$ and $0<s<t\le T$, the discrete velocity v_h defined by (4.4) and the Hood–Taylor element approximation to u, u_h, satisfy the following bound:*

$$\begin{aligned}&\big\|e^{-A_h(t-s)}A_h^{-1}\big(B_h(u_h(s),u_h(s))-B_h(v_h(s),v_h(s))\big)\big\|_0\\&\qquad\le C\left(\frac{1}{\sqrt{t-s}}+\frac{1}{\sqrt{s}}\right)\big\|A_h^{-1}\big(u_h(s)-v_h(s)\big)\big\|_0.\end{aligned}\tag{4.50}$$

Proof. We write $\rho_h=B_h(u_h(s),u_h(s))-B_h(v_h(s),v_h(s))$. We will bound the norm of $e^{-A_h(t-s)}A_h^{-1}\rho_h$ by duality. As usual, we will omit the dependence on s whenever this does not cause confusion. For $\phi_h\in V_{h,r}$ we set $\psi_h=e^{-A_h(t-s)}A_h^{-1}\phi_h$, so that

$$\big\|e^{-A_h(t-s)}A_h^{-1}\rho_h\big\|_0=\max_{\phi_h\in V_{h,r}\phi_h\ne 0}\frac{\big|(\rho_h,\psi_h)\big|}{\|\phi_h\|_0}.$$

We have

$$(\rho_h,\psi_h)=-b(e_h,\psi_h,u_h)-b(v_h,\psi_h,e_h),\tag{4.51}$$

where we recall that $e_h=v_h-u_h$. For the second term on the right-hand side of (4.51) we have

$$|b(v_h,\psi_h,e_h)|\le|b(u,\psi_h,e_h)|+|b(v_h-u,\psi_h,e_h)|$$

and

$$|b(v_h-u,\psi_h,e_h)|\le\|v_h-u\|_0\,\|\nabla\psi_h\|_\infty\,\|e_h\|_0+\|\nabla\cdot(v_h-u)\|_0\,\|\psi_h\|_\infty\,\|e_h\|_0\,.$$

Applying (2.5) and (2.22) we have $\|\nabla\psi_h\|_\infty\le Ch^{-1}\,\|\psi_h\|_\infty\le Ch^{-1}\,\|\phi_h\|_0$. Also, since $(v_h-u)=z_h+(\Pi_hu-u)$, by taking into account (2.7) and Lemma 4.3 we get

$$|b(v_h,\psi_h,e_h)|\le|b(u,\psi_h,e_h)|+\frac{CM_4}{s^{1/2}}h^2\,\|e_h\|_0\,\|\phi_h\|_0\,.\tag{4.52}$$

Similarly,

$$|b(e_h,\psi_h,u_h)|\le|b(e_h,\psi_h,u)|+|b(e_h,\psi_h,u_h-u)|\,,$$

and

$$|b(e_h,\psi_h,u_h-u)| \le \left(\|e_h\|_0 \|u_h-u\|_1 + \|e_h\|_1 \|u_h-u\|_0\right) \|\psi_h\|_\infty .$$

Since by virtue of (2.5) we have $\|e_h\|_1 \le Ch^{-1}\|e_h\|_0$, by applying Corollary 4.8 we get

$$|b(e_h,\psi_h,u_h)| \le |b(e_h,\psi_h,u)| + \frac{C}{s^{1/2}} h^2 \|e_h\|_0 \|\phi_h\|_0 . \tag{4.53}$$

Thus, from (4.51), (4.52), and (4.53) and, since

$$h^2\|e_h\|_0 \le C\|A_h^{-1}e_h\|_0, \tag{4.54}$$

it follows that

$$|(\rho_h,\psi_h)| \le |b(e_h,\psi_h,u)| + |b(u,\psi_h,e_h)| + \frac{C}{\sqrt{s}} \left\|A_h^{-1}e_h\right\|_0 \|\phi_h\|_0 . \tag{4.55}$$

Next, we apply Lemma 4.12 to get

$$\begin{aligned} |(\rho_h,\psi_h)| &\le |b(e_h,\Psi(t-s),u)| + |b(u,\Psi(t-s),e_h)| \\ &\quad + \left(\frac{C}{\sqrt{t-s}} + \frac{C}{\sqrt{s}}\right) \left\|A_h^{-1}e_h\right\|_0 \|\phi\|_0 . \end{aligned} \tag{4.56}$$

Now we apply Lemma 4.10 to the first two terms on the right-hand side of (4.56), with g replaced by e_h and v by u. Let us first observe that, using (2.20) and (4.54), we have $\|A^{-1}\Pi e_h\|_0 + h^2\|e_h\|_0 \le C\|A_h^{-1}e_h\|_0$, so that in order to finish the proof we only have to bound $\|(u\cdot\nabla)\Psi\|_2$, $\|\nabla\Psi\cdot u\|_2$, and $\|\nabla u\cdot\Psi\|_2$ (recall that $(\nabla\cdot u)=0$).

Since

$$\|(u\cdot\nabla)\Psi\|_2 \le \|u\|_2 \|\nabla\Psi\|_\infty + \|u\|_{W^{1,4}} \|\Psi\|_{W^{2,4}} + \|u\|_\infty \|\Psi\|_3 ,$$

taking into account that $\|\Psi\|_3 \le Ct^{-1/2}\|\phi_h\|_0$ and applying standard Sobolev inequalities (2.1) we obtain

$$\|(u(s)\cdot\nabla)\Psi(t-s)\|_2 \le \frac{CM_2}{\sqrt{t-s}} \|\phi_h\|_0 . \tag{4.57}$$

Arguing similarly we get

$$\|\nabla\Psi(t-s)\cdot u(s)\|_2 \le \frac{CM_2}{\sqrt{t-s}} \|\phi_h\|_0 . \tag{4.58}$$

Finally, since

$$\|\nabla u\cdot\Psi\|_2 \le \|\nabla u\|_{L^4} \|\Psi\|_{W^{2,4}} + \|\nabla u\|_1 \|\Psi\|_{W^{1,\infty}} + \|\nabla u\|_2 \|\Psi\|_\infty ,$$

taking into account that $\|\Psi\|_2 \le C\|\phi_h\|_0$ and using (2.1) again, we deduce

$$\begin{aligned} \|\nabla u(s)\cdot\Psi(t-s)\|_2 &\le C\left(\frac{M_2}{\sqrt{t-s}} + \|u(s)\|_3\right) \|\phi_h\|_0 \\ &\le C\left(\frac{M_2}{\sqrt{t-s}} + \frac{M_3}{\sqrt{s}}\right) \|\phi_h\|_0 . \end{aligned} \tag{4.59}$$

Thus, inequalities (4.57), (4.58), and (4.59), together with (4.56) and Lemma 4.10, allow us to conclude the proof. ∎

LEMMA 4.14. *Let (u,p) be the solution of* (1.1)–(1.2). *There exists a positive constant C such that the discrete velocity v_h defined by* (4.4) *and the Hood–Taylor element approximation to u, u_h, satisfy the following bound for $r \geq 4$:*

$$\|A_h^{-1}(v_h(t) - u_h(t))\|_0 \leq Ch^5, \quad t \in (0,T].$$

Proof. Let us consider $y_h(t) = A_h^{-1}e_h(t)$ and $\rho_h = B_h(v_h,v_h) - B_h(u,u)$. From (4.5) we get

$$y_h(t) = \int_0^t e^{-(t-s)A_h} A_h^{-1}\big(B_h(u_h,u_h) - B_h(v_h,v_h)\big)\,ds + \int_0^t e^{-(t-s)A_h}A_h^{-1}\rho_h(s)\,ds.$$

Thanks to Lemma 4.13 we obtain

$$\|y_h(t)\|_0 \leq C\int_0^t \left(\frac{1}{\sqrt{t-s}} + \frac{1}{\sqrt{s}}\right)\|y_h(t)\|_0 + \left\|\int_0^t e^{-(t-s)A_h}A_h^{-1}\rho_h(s)\,ds\right\|_0,$$

so that by applying Lemma 4.9, we only have to show that

$$\max_{0\leq t\leq T}\left\|\int_0^t e^{-A_h(t-s)}A_h^{-1}\rho_h(s)\,ds\right\|_0 \leq Ch^5. \tag{4.60}$$

For this purpose, we will first estimate $e^{-A_h(t-s)}A_h^{-1}\rho_h(s)$ by duality. Let us take $\phi_h \in V_{h,r}$ and set $\psi_h = e^{-A_h(t-s)}A_h^{-1}\phi_h$ and $z = u - v_h$, so that we have

$$(\rho_h,\psi_h) = b(z,u,\psi_h) + b(u,z,\psi_h) - b(z,z,\psi_h). \tag{4.61}$$

For the third term on the right-hand side above, by applying (2.22), we have

$$\big|b(z,z,\psi)\big| \leq C\,\|z\|_0\,\|z\|_1\,\|\psi_h\|_{L^\infty} \leq C\,\|z\|_0\,\|z\|_1\,\|\phi_h\|_0\,.$$

For the other two terms on the right-hand side of (4.61), Lemma 4.12 shows that

$$|b(z,u,\psi_h) + b(u,z,\psi_h)| \leq |b(z,\Psi(t-s),u)| + |b(u,\Psi(t-s),z)| + \frac{CM_2}{\sqrt{t-s}}h^2\,\|z\|_0\,\|\phi\|_0\,,$$

where, as in Lemma 4.12, $\Psi(t) = e^{-At}A^{-1}\Pi\phi_h$. Furthermore, by applying Lemma 4.10 to the first two terms on the right-hand side above, we have

$$|b(z,\Psi(t-s),u)| + |b(u,\Psi(t-s),z)| \leq Ch^2\,\|z\|_0\,\|\nabla u\cdot\Psi\|_2$$
$$+C\big(\|A^{-1}\Pi z\|_0 + h^2\,\|z\|_0\big)\big(\|(u\cdot\nabla)\Psi\|_2 + \|\nabla\Psi\cdot u\|_2\big).$$

Recalling the bounds of $\|(u\cdot\nabla)\Psi\|_2$, $\|\nabla\Psi\cdot u\|_2$ and $\|\nabla u\cdot\Psi\|_2$ in (4.57)–(4.59) we reach

$$\left\|e^{-(t-s)A_h}\rho_h(s)\right\|_0 \leq C\big(\|z\|_1 + h^2\,\|u\|_3\big)\,\|z\|_0 + \frac{CM_2}{\sqrt{t-s}}\big(h^2\,\|z\|_0 + \|A^{-1}\Pi z\|_0\big). \tag{4.62}$$

By applying Hölder's inequality, writing $z = (u - \Pi_h u) + z_h$, and using (2.7) and (4.12) we get

$$\int_0^t \left(\|z(s)\|_1 + h^2 \|u(s)\|_3\right) \|z(s)\|_0 \, ds \le Ch^5. \tag{4.63}$$

Using again (2.7) together with (4.9) we obtain $s^{1/2}\|z(s)\|_0 \le Ch^3$, and then

$$\int_0^t \frac{M_2}{\sqrt{t-s}} h^2 \|z(s)\|_0 \, ds \le CB\left(\frac{1}{2}, \frac{1}{2}\right) h^5. \tag{4.64}$$

Finally, as a consequence of (2.8) and (2.20) we have

$$\left\|A^{-1}\Pi z\right\|_0 \le \left\|A^{-1}\Pi z_h\right\|_0 + \left\|A^{-1}\Pi(I-\Pi_h)u\right\|_0 \le \left\|A_h^{-1} z_h\right\|_0 + C\left(h^2 \|z_h\|_0 + \|u\|_3 h^5\right),$$

so that using (4.9) again we get $s^{1/2}\|A^{-1}\Pi z(s)\|_0 \le Ch^5$, and then

$$\int_0^t \frac{M_2}{\sqrt{t-s}} \|A^{-1}\Pi z(s)\|_0 \, ds \le CB\left(\frac{1}{2}, \frac{1}{2}\right) h^5. \tag{4.65}$$

Now (4.60) follows from (4.62), (4.63), (4.64), and (4.65). ∎

THEOREM 4.15. *Let (u,p) be the solution of* (1.1)–(1.2). *There exists a positive constant C such that the discrete velocity v_h defined by* (4.4) *and the Hood–Taylor element approximation to u, u_h, satisfy the following bound:*

$$\|v_h(t) - u_h(t)\|_j \le \frac{C}{t} |\log(h)| h^{5-j}, \quad t \in (0,T], \quad j = 0, 1, \quad r \ge 4. \tag{4.66}$$

Proof. We prove the error bound for the L^2 norm from which the H^1 norm bound is obtained by applying inverse inequality (2.5). We consider $y_h(t) = te_h(t)$. By arguing similarly as in Theorem 4.7 we have

$$\max_{0\le t\le T} \|y_h(t)\|_0 \le C\left(\max_{0\le s\le T} \left\|sA_h^{-1}\rho_h(s)\right\|_0 |\log(h)| + \max_{0\le s\le T} \left\|A_h^{-1}e_h(s)\right\|_0 |\log(h)|\right).$$

To conclude the estimate we apply Lemmas 4.5 and 4.14 to the first and second terms on the right-hand side above, respectively. ∎

COROLLARY 4.16. *Let (u,p) be the solution of* (1.1)–(1.2), *and let u_h be the Hood–Taylor element approximation to u. Then, for $r \ge 4$, there exists a positive constant C such that the following bound holds for $j = 0, 1$:*

$$\|u_h(t) - u(t)\|_j \le \frac{C}{t} h^{4-j}, \quad t \in (0,T].$$

Proof. The proof is obtained by reasoning as in Corollary 4.8 using Theorem 4.15 instead of Theorem 4.7. ∎

4.4. Superconvergence for the pressure. We begin with some error estimates for the time derivative of $v_h - u_h$.

LEMMA 4.17. *Let (u,p) be the solution of* (1.1)–(1.2). *Then there exists a positive constant C such that the discrete velocity v_h defined by* (4.4) *and the Hood–Taylor*

element approximation to u, u_h, satisfy the following bounds:

$$\|\dot{v}_h(t) - \dot{u}_h(t)\|_0 \leq \frac{C}{t^{s/2}} |\log(h)| h^{1+s}, \quad t \in (0,T], \tag{4.67}$$

$$\|\dot{v}_h(t) - \dot{u}_h(t)\|_{-1} \leq \frac{C}{t^{s/2}} |\log(h)| h^{2+s}, \quad t \in (0,T], \tag{4.68}$$

$$\|A^{-1}\Pi(\dot{v}_h(t) - \dot{u}_h(t))\|_0 \leq \frac{C}{t^{s/2}} |\log(h)| h^{3+s}, \quad t \in (0,T], \tag{4.69}$$

where $s = 1$ in the case $r = 3$ and $s = 2$ in the case $r = 4$.

Proof. Let us first observe that

$$\dot{e}_h = -A_h e_h + B_h(u_h, u_h) - B_h(u,u),$$

where we recall that $e_h = v_h - u_h$ Then by using (4.21) we get

$$\|\dot{e}_h\|_0 \leq Ch^{-1}\|e_h\|_1 + \|u_h - u\|_1.$$

Applying now Theorems 4.7 and 4.15 and Corollaries 4.8 and 4.16 to bound the first and second terms on the right-hand side above, we conclude (4.67).

In order to bound $\|\dot{e}_h\|_{-1} \leq C\|A_h^{-1/2}\dot{e}_h\|_0$ we observe that

$$\|A_h^{-1/2}\dot{e}_h\|_0 \leq \|A_h^{1/2}e_h\|_0 + \|A_h^{-1/2}(B_h(u_h,u_h) - B_h(u,u))\|_0,$$

so that by applying (4.22) from Lemma 4.4 we get

$$\|A_h^{-1/2}\dot{e}_h\|_0 \leq C\|e_h\|_1 + C\|u - u_h\|_0,$$

and (4.68) is reached by applying again Theorems 4.7 and 4.15 and Corollaries 4.8 and (4.16).

We now prove (4.69). By using (2.20) and Lemma 4.1 we obtain

$$\begin{aligned}
\|A^{-1}\Pi\dot{e}_h\|_0 &\leq \|A^{-1}\Pi A_h e_h\|_0 + \|A^{-1}\Pi(B_h(u_h,u_h) - B_h(u,u))\|_0 \\
&\leq C(h^2\|A_h e_h\|_0 + \|A_h^{-1}A_h e_h\|_0) + K(\|u - u_h\|_{-1} + \|u_h - u\|_0\|u_h - u\|_1) \\
&\leq C\|e_h\|_0 + K(\|e_h\|_{-1} + \|v_h - u\|_{-1} + \|u_h - u\|_0\|u_h - u\|_1).
\end{aligned}$$

To finish the proof we notice that $\|e_h\|_{-1} \leq C\|e_h\|_0$, and then we apply Theorems 4.7 and 4.15 and Corollaries 4.8 and 4.16 together with Lemma 4.3. ∎

THEOREM 4.18 (superconvergence for the pressure). *Let (u,p) be the solution of (1.1)–(1.2). Then, for $r = 3,4$, there exists a positive constant C such that the discrete pressure g_h defined by (4.1) and the Hood–Taylor element approximation to p, p_h, satisfy the following bound:*

$$\|p_h(t) - g_h(t)\|_{L^2/\mathbb{R}} \leq \frac{C}{t^{(r-2)/2}} h^r \left|\log(h)\right|, \qquad t \in (0,T].$$

Proof. By subtracting (3.1) from (4.1), we obtain for the difference $g_h - p_h$

$$(g_h - p_h, \nabla\cdot\phi_h) = (\nabla e_h, \nabla\phi_h) + (F(u,u) - F(u_h,u_h), \phi_h) + (\dot{e}_h, \phi_h)$$

for all $\phi_h \in X_{h,r}$. Using the inf-sup condition (2.6),

$$\beta\|g_h - p_h\|_{L^2/\mathbb{R}} \leq \|e_h\|_1 + \|F(u,u) - F(u_h,u_h)\|_{-1} + \|\dot{e}_h\|_{-1}.$$

By applying Theorems 4.7 and 4.15, (4.20), and (4.68) we obtain

$$\beta\|g_h - p_h\|_{L^2/\mathbb{R}} \leq \frac{C}{t^{(r-2)/2}} h^r |\log(h)| + C\|u - u_h\|_0.$$

Finally, by using Corollaries 4.8 and 4.16, the conclusion is reached. ∎

COROLLARY 4.19. *Let (u,p) be the solution of (1.1)–(1.2). Then, for $r = 3,4$, there exists a positive constant C such that the Hood–Taylor element approximation to p, p_h, satisfies the following bound:*

$$\|p_h(t) - p(t)\|_{L^2/\mathbb{R}} \leq \frac{C}{t^{(r-2)/2}} h^{r-1}, \qquad t \in (0,T]. \tag{4.70}$$

Proof. Let g_h be the discrete pressure defined by (4.1). We decompose

$$\|p_h - p\|_{L^2/\mathbb{R}} \leq \|p_h - g_h\|_{L^2/\mathbb{R}} + \|g_h - p\|_{L^2/\mathbb{R}}$$

and apply Theorem 4.18 to bound the first term on the right-hand side above. To bound the second one we decompose $\|g_h - p\|_{L^2/\mathbb{R}} \leq \|g_h - q_h\|_{L^2/\mathbb{R}} + \|q_h - p\|_{L^2/\mathbb{R}}$, where q_h is that in (2.11). Using (2.14) it only remains to bound the first term above. By taking into account that

$$(u_t, \phi_h) + (\nabla s_h, \nabla\phi_h) + b(u,u,\phi_h) + (\nabla q_h, \phi_h) = (f,\phi_h) \quad \forall \phi_h \in X_{h,r},$$

$$(\dot{v}_h, \phi_h) + (\nabla v_h, \nabla\phi_h) + b(u,u,\phi_h) + (\nabla g_h, \phi_h) = (f,\phi_h) \quad \forall \phi_h \in X_{h,r},$$

and arguing as in the proof of Theorem 4.18, we deduce that

$$\|g_h - q_h\|_{L^2/\mathbb{R}} \leq C\|v_h - s_h\|_1 + \|\dot{v}_h - u_t\|_{-1}.$$

To bound the first term we decompose $\|v_h - s_h\|_1 \leq \|z_h\|_1 + \|\Pi_h u - u\|_1 + \|s_h - u\|_1$ and then apply Lemma 4.3, (2.7), and (2.13). To bound the second one, by using (2.8) we get

$$\|\dot{v}_h - u_t\|_{-1} \leq \|\dot{z}_h\|_{-1} + \|\Pi_h u_t - u_t\|_{-1} \leq \|\dot{z}_h\|_{-1} + Ch^{r-1}\|u_t\|_{r-2}.$$

Finally, since $\|\dot{z}_h\|_{-1} \leq C\|A_h^{-1/2}\dot{z}_h\|_0$, by using (4.6) we get

$$\|\dot{z}_h\|_{-1} \leq C\|z_h\|_1 + \|\theta_h\|_1,$$

and we conclude by applying Lemma 4.3, (2.7), and (2.13). ∎

Remark 4.3. We want to point out that the error bounds for the Galerkin approximation to the velocity in Corollaries 4.8 and 4.16 are analogous to those obtained in [37, Theorem 3.1]. If we compare the error bound for the pressure in Corollary 4.19 with that of [37, Theorem 3.1], we can observe that the singular behavior of pressure estimate (4.70) as t tends to zero behaves like $t^{-(r-2)/2}$ improving the estimate in [37, Theorem 3.1] that behaves like $t^{-(r-1)/2}$. Recently, in [7] improved convergence of order $O(t^{(-1/4)-\delta})$ for any $\delta > 0$ is shown for the approximation to the velocity using quadratic elements. Error estimates for the pressure are not considered in [7].

5. Analysis of the postprocessed method. We require some error estimates for the time derivative that we now prove.

LEMMA 5.1. *Let (u,p) be the solution of (1.1)–(1.2). There exists a positive constant C such that the Hood–Taylor element approximation to u, u_h, satisfies the following bounds:*

$$\|u_t - \dot{u}_h(t)\|_0 \le \frac{C}{t^{(l+1)/2}} h^{1+l} \left|\log(h)\right|, \quad t \in (0,T], \tag{5.1}$$

$$\|u_t - \dot{u}_h(t)\|_{-1} \le \frac{C}{t^{(l+1)/2}} h^{2+l} \left|\log(h)\right|, \quad t \in (0,T], \tag{5.2}$$

$$\|A^{-1}\Pi(u_t - \dot{u}_h(t))\|_0 \le \frac{C}{t^{(l+1)/2}} h^{3+l} \left|\log(h)\right|, \quad t \in (0,T], \tag{5.3}$$

where $l = 1$ in the case $r = 3$ and $l = 2$ in the case $r = 4$.

Proof. By writing

$$u_t - \dot{u}_h = (u_t - \Pi_h u_t) + (\Pi_h u_t - \dot{v}_h) + (\dot{v}_h - \dot{u}_h) = (u_t - \Pi_h u_t) + \dot{z}_h + \dot{e}_h$$

and applying (2.7) and (2.8) to bound the first term and Lemma 4.17 for the third one, we only need to bound the different norms of $\dot{z}_h$. Since the norms $\|A_h^{-1/2}\dot{z}_h\|_0$ and $\|\dot{z}_h\|_{-1}$ are equivalent and since in view of (2.20) $\|A^{-1}\Pi\dot{z}_h\|_0$ can be bounded in terms of $\|\dot{z}_h\|_0$ and $\|A_h^{-1}\dot{z}_h\|_0$, we only have to bound $\|A_h^{-k/2}\dot{z}_h\|_0$ for $k = 0,1,2$. For this purpose, we will argue as in the proof of (4.10). We notice that $\ddot{z}_h + A_h\dot{z}_h = A_h\dot{\theta}_h$, and if we let j be $j = (1+l)/2$ for $y_h = t^j A_h^{-1}\dot{z}_h$, we have $\dot{y}_h + A_h y_h = t^j\dot{\theta}_h + jt^{j-1}A_h^{-1}\dot{z}_h$, so that

$$A_h^{-1}t^j\dot{z}_h(t) = \int_0^t e^{-A_h(t-s)}s^j\dot{\theta}_h(s)\,ds + j\int_0^t e^{-A_h(t-s)}A_h^{-1}s^{j-1}\dot{z}_h(s)\,ds.$$

By applying (2.10) to both integrals we obtain

$$\left\|A_h^{-1}t^j\dot{z}_h(t)\right\|_0 \le CB\left(\frac{1}{2},\frac{1}{2}\right) \left(\max_{0\le s\le t} s^{j+1/2}\left\|A_h^{-1/2}\dot{\theta}_h(s)\right\|_0 + j\max_{0\le s\le t} s^{j-1/2}\left\|A_h^{-3/2}\dot{z}_h(s)\right\|_0\right).$$

Now, on the one hand, we have that (2.16) and (2.8) are valid if u, s_h, and p are replaced by their time derivatives. On the other hand, we have that $A_h^{-3/2}\dot{z}_h = A_h^{-1/2}\theta_h - A_h^{-1/2}z_h$, so that in view of (2.16), (2.8), (4.10), and (4.11) the proof of (5.3) is easily finished. Taking into account that $\left\|A_h^{k/2}\right\|_0 \le Ch^{-k}$ for $k = 1,2$, then (5.1) and (5.2) follow. ∎

THEOREM 5.2. *Let (u,p) be the solution of (1.1)–(1.2). There exists a positive constant C such that the postprocessed MFE approximation to u, $\tilde{u}_h$, satisfies the following bounds for $j = 0,1$:*

(i) *If the postprocessing element is $(\widetilde{X},\widetilde{Q}) = (X_{h',r}, Q_{h',r-1})$, then*

$$\|u(t) - \tilde{u}_h(t)\|_j \le \frac{C}{t^{1/2}}(h')^{3-j} + \frac{C}{t}h^{4-j}|\log(h)|, \quad t \in (0,T],\ r \ge 3, \tag{5.4}$$

$$\|u(t) - \tilde{u}_h(t)\|_j \le \frac{C}{t}(h')^{4-j} + \frac{C}{t^{3/2}}h^{5-j}|\log(h)|, \quad t \in (0,T],\ r \ge 4; \tag{5.5}$$

(ii) *if the postprocessing element is* $(\widetilde{X},\widetilde{Q}) = (X_{h,r+1}, Q_{h,r})$*, then*

$$\|u(t)-\tilde{u}_h(t)\|_j \le \frac{C}{t}h^{4-j}|\log(h)|, \quad t\in(0,T],\ r\ge 3, \tag{5.6}$$

$$\|u(t)-\tilde{u}_h(t)\|_j \le \frac{C}{t^{3/2}}h^{5-j}|\log(h)|, \quad t\in(0,T],\ r\ge 4. \tag{5.7}$$

Proof. The proof follows the same steps as [6, Theorem 3.14]. Let $\widetilde{S}_h(u(t))\in\widetilde{V}$ be the Stokes projection of the solution of (1.1)–(1.2) at time t. We decompose $\|u(t)-\tilde{u}_h(t)\|_l \le \|u(t)-\widetilde{S}_h(u(t))\|_l + \|\widetilde{S}_h(u(t))-\tilde{u}_h(t)\|_l$, $l=0,1$. We apply (2.13) to bound the first term, so that we will concentrate now on the second. It is easy to obtain

$$\|\tilde{u}_h(t)-\widetilde{S}_h(u(t))\|_1 \le C\|F(u(t),u(t))-F(u_h(t),u_h(t))\|_{-1} + C\|u_t(t)-\dot{u}_h(t)\|_{-1}.$$

For the first term above we apply (4.20) and Corollaries 4.8 and 4.16 to obtain

$$\|F(u(t),u(t))-F(u_h(t),u_h(t))\|_{-1} \le C\|u(t)-u_h(t)\|_0 \le \frac{C}{t^{(r-1)/2}}h^r.$$

The second term is already bounded in (5.2). Then the proof for the H^1 norm is complete. We next deal with the estimate in the L^2 norm. We first observe that

$$\widetilde{A}_h(\tilde{u}_h(t)-\widetilde{S}_h(u(t))) = \widetilde{\Pi}_h(F(u(t),u(t))-F(u_h(t),u_h(t))) + \widetilde{\Pi}_h(u_t(t)-\dot{u}_h(t)).$$

Then, by applying $\widetilde{A}_h^{-1}$ to both sides of the above equation, we obtain

$$\|\tilde{u}_h-\widetilde{S}_h(u)\|_0 \le \|\widetilde{A}_h^{-1}\widetilde{\Pi}_h(F(u(t),u(t))-F(u_h(t),u_h(t)))\|_0 + \|\widetilde{A}_h^{-1}\widetilde{\Pi}_h(u_t(t)-\dot{u}_h(t))\|_0.$$

As regards the nonlinear term, by taking into account (2.19) and applying Lemma 4.1 and (4.19) we get

$$\begin{aligned}&\|\widetilde{A}_h^{-1}\widetilde{\Pi}_h(F(u,u)-F(u_h,u_h))\|_0\\ &\qquad\le C\tilde{h}^2\|u-u_h\|_1 + C(\|u-u_h\|_{-1}+\|u-u_h\|_0\|u-u_h\|_1).\end{aligned}$$

The estimates for the L^2 and H^1 norms of the errors of the Galerkin approximation to the velocity are granted by Corollaries 4.8 and 4.16. As regards the estimate in the H^{-1} norm, we use the decomposition $\|u-u_h\|_{-1} \le \|u-\Pi_h u\|_{-1} + \|z_h\|_{-1} + \|v_h-u_h\|_0$ and apply (2.8) together with Lemma 4.3 and Theorems 4.7 and 4.15. Finally, to bound $\|\widetilde{A}_h^{-1}\widetilde{\Pi}_h(u_t(t)-\dot{u}_h(t))\|_0$ we apply (2.19) again and then use estimates (5.1) and (5.3) from Lemma 5.1. ∎

THEOREM 5.3. *Let* (u,p) *be the solution of* (1.1)–(1.2)*. There exists a positive constant* C *such that the postprocessed MFE approximation to* p*,* $\tilde{p}_h$*, satisfies the following bounds:*

(i) *If the postprocessing element is* $(\widetilde{X},\widetilde{Q}) = (X_{h',r}, Q_{h',r-1})$*, then*

$$\|p(t)-\tilde{p}_h(t)\|_{L^2/\mathbb{R}} \le \frac{C}{t^{(r-2)/2}}(h')^{r-1} + \frac{C}{t^{(r-1)/2}}h^r|\log(h)|, \quad t\in(0,T];$$

(ii) *if the postprocessing element is* $(\widetilde{X},\widetilde{Q}) = (X_{h,r+1}, Q_{h,r})$*, then*

$$\|p(t)-\tilde{p}_h(t)\|_{L^2/\mathbb{R}} \le \frac{C}{t^{(r-1)/2}}h^r|\log(h)|, \quad t\in(0,T]. \tag{5.8}$$

Proof. The proof is reached by reasoning as in [6, Theorem 3.15]. Let us denote by $\tilde{q}_h(t)$ the approximation to the pressure $p(t)$ such that $(\widetilde{S}_h(u(t)), \tilde{q}_h(t)) \in (\widetilde{X}, \widetilde{Q})$ solves (2.11)–(2.12) with $g = f - u_t - u \cdot \nabla u$. Then we use the decomposition

$$\|p(t) - \tilde{p}_h(t)\|_{L^2/\mathbb{R}} \leq \|p(t) - \tilde{q}_h(t)\|_{L^2/\mathbb{R}} + \|\tilde{q}_h(t) - \tilde{p}_h(t)\|_{L^2/\mathbb{R}}.$$

The first term is estimated by applying (2.14). Let us now bound the second term. Using the inf-sup condition (2.6) it is easy to deduce that

$$\beta\|\tilde{p}_h - \tilde{q}_h\|_{L^2/\mathbb{R}} \leq \|\tilde{u}_h - \widetilde{S}_h(u)\|_1 + \|F(u_h, u_h) - F(u, u)\|_{-1} + \|u_h - u_t\|_{-1}.$$

By applying Theorem 5.2 to bound the first term, (4.20) together with Corollaries 4.8 and 4.16 for the second, and (5.2) for the third, the conclusion is reached. ∎

6. Remarks on fully discrete postprocessing. In practice, the finite-element approximations u_h and p_h, being solutions of system (3.1)–(3.2), can rarely be computed analytically, and one has to compute approximations $u_h^{(n)} \approx u_h(t_n)$ and $p_h^{(n)} \approx p_h(t_n)$ on some time levels $0 = t_0 < t_1 \ldots < t_N = T$ by means of a time integrator. Consequently, instead of the postprocessed approximations $\tilde{u}_h(t_n)$ and $\tilde{p}_h(t_n)$, one obtains $\tilde{u}_h^{(n)}$ and $\tilde{p}_h^{(n)}$ as solutions of

$$\left(\nabla\tilde{u}_h^{(n)}, \nabla\tilde{\phi}\right) + \left(\nabla\tilde{p}_h^{(n)}, \tilde{\phi}\right) = (f, \tilde{\phi}) - b(u_h^{(n)}, u_h^{(n)}, \tilde{\phi}) - (d_t u_h^{(n)}, \tilde{\phi}) \quad \forall\, \tilde{\phi} \in \widetilde{X}, \tag{6.1}$$

$$\left(\nabla \cdot \tilde{u}_h^{(n)}, \tilde{\psi}\right) = 0 \quad \forall\, \tilde{\psi} \in \widetilde{Q}, \tag{6.2}$$

where $(\widetilde{X}, \widetilde{Q})$ is as in (3.3)–(3.4), and $d_t u_h^{(n)}$ is an approximation to $\dot{u}_h(t_n)$ computed in terms of as many values $u_h^{(n-j)}$, $j = 0, 1, \ldots, N - n$, as needed. Recent research in [22] suggests

$$d_t u^{(n)} = \Pi_h f - A_h u_h^{(n)} - B_h(u_h^{(n)}, u_h^{(n)}) \tag{6.3}$$

as a convenient and adequate approximation.

Notice then that the error $u(t_n) - \tilde{u}_h^{(n)}$ can be expressed as $u(t_n) - \tilde{u}_h^{(n)} = (u(t_n) - \tilde{u}_h(t_n)) + (\tilde{u}_h(t_n) - \tilde{u}_n^{(n)})$. The first term on the right-hand side is the spatial discretization error which we have analyzed in the previous sections. The second one is the error arising from temporal discretization that will be analyzed now.

Let us denote by $e_h^{(n)} = u_h(t_n) - u_h^{(n)}$ and $\tilde{e}_h^{(n)} = \tilde{u}_h(t_n) - \tilde{u}_h^{(n)}$ the errors induced by the temporal discretization in the Galerkin and postprocessed approximations, respectively. We estimate here $\tilde{e}_h^{(n)}$ in terms of $e_h^{(n)}$, which depends on the particular time integrator used to approximate the solution of (3.1)–(3.2). Estimates of $e_h^{(n)}$ (when $t_n = nk$, $n = 0, 1, \ldots, N = T/k$) for the Crank–Nicolson method can be found in [38], with $O(t_n^{-1}k^2)$ and $O(t_n^{-1}k)$ estimates in the L^2 and H^1 norms, respectively, (see also [31]). For the second-order backward differentiation formula (BDF), similar estimates are obtained in [13] (for the limit case $h \to 0$), with a $k^{1/2}$ improvement in the H^1 estimate with respect to that in [38]. Also, from results in [41], it is straightforward to obtain $O(k)$ estimates in both the L^2 and the H^1 norms for the implicit Euler method. Notice that all of these estimates together with the boundedness of $\|u_h(t)\|_1$, $0 \leq t \leq T$, and h sufficiently small [37] imply that $\|u_h^{(n)}\|_1$ is bounded for $0 \leq n \leq N$ and h sufficiently small. Further estimates for other integrators are outside the scope of the present paper.

From (6.3) it follows that

$$\dot{u}_h(t_n) - d_t u_h^{(n)} = -A_h e_h^{(n)} + \Pi_h\big(F(u_h^{(n)}, u_h^{(n)}) - F(u_h(t_n), u_h(t_n))\big), \tag{6.4}$$

so that by subtracting (6.1) from (3.3) we get

$$\widetilde{A}_h \tilde{e}_h^{(n)} = (\widetilde{\Pi}_h - \Pi_h)\big(F(u_h^{(n)}, u_h^{(n)}) - F(u_h(t_n), u_h(t_n))\big) - \widetilde{\Pi}_h A_h e_h^{(n)}.$$

On the one hand, we may write $\widetilde{A}_h^{-1}\widetilde{\Pi}_h A_h e_h^{(n)} = e_h^{(n)} + \big(\widetilde{A}_h^{-1}\widetilde{\Pi}_h - A_h^{-1}\big)A_h e_h^{(n)}$, so that by using (2.17) we get

$$\|\tilde{A}_h^{-1}\tilde{\Pi}_h A_h e_h^{(n)}\|_j \le \|e_h^{(n)}\|_j + Ch^{2-j}\|A_h e_h^{(n)}\|_0, \quad j = 0, 1.$$

On the other hand, a simple duality argument and (2.17) show that $\|\widetilde{A}_h^{-1}(\widetilde{\Pi}_h - \Pi_h)f\|_j \le Ch^{2-j}\,\|f\|_0$ for $f \in L^2(\Omega)^d$, $j = 0, 1$, so that first applying this inequality to $f = F(u_h^{(n)}, u_h^{(n)}) - F(u_h(t_n), u_h(t_n))$ and then well-known inequalities (e.g., equation [38, (3.7)]), we conclude that

$$\big\|\tilde{e}_h^{(n)}\big\|_j \le \big\|e_h^{(n)}\big\|_j + Ch^{2-j}\Big(\big\|e_h^{(n)}\big\|_1 + \big\|A_h e_h^{(n)}\big\|_0\Big), \tag{6.5}$$

where the constant C depends on $\|u_n^{(n)}\|_1$ (which, as mentioned before, is bounded for h sufficiently small) and on $\|A_h u_h(t_n)\|_0$, which can be easily seen to be bounded for $0 \le t \le T$ and h sufficiently small. Taking into account that $h^{2-j}\big\|A_h e_h^{(n)}\big\|_0 \le C\|e_h^{(n)}\|_j$, we conclude that the temporal error $\tilde{e}_h^{(n)}$ of the fully discrete postprocessed method is proportional to temporal error $e_h^{(n)}$ of the mixed finite-element approximation. Furthermore, for the second-order BDF formula, besides the estimates in [13], in [23] we obtain $O(t_n^{-3/2}k^2)$ and $O(t_n^{-2}k^2)$ estimates for $\|e_h^{(n)}\|_1$ and $\|A_h e_h^{(n)}\|_0$, respectively, so that, in view of (6.5), the two errors $\tilde{e}_h^{(n)}$ and $e_h^{(n)}$ are asymptotically equivalent as $h \to 0$.

Finally, for the pressure, arguing as in the proof of Theorem 5.3 and using the same arguments we have used for $\tilde{e}_h^{(n)}$, one can easily show that also

$$\big\|\tilde{p}_h(t_n) - \tilde{p}_h^{(n)}\big\|_{L^2/\mathbb{R}} \le C\big\|e_h^{(n)}\big\|_1.$$

REFERENCES

[1] R. Adams, *Sobolev Spaces*, Academic Press, New York, 1975.

[2] A. Ait Ou Ammi and M. Marion, *Nonlinear Galerkin methods and mixed finite elements: Two-grid algorithms for the Navier–Stokes equations*, Numer. Math., 68 (1994), pp. 189–213.

[3] D. N. Arnold, F. Brezzi, and M. Fortin, *A stable finite element for the Stokes equations*, Calcolo, 23 (1984), pp. 374–344.

[4] B. Ayuso, J. de Frutos, and J. Novo, *Improving the accuracy of the mini-element approximation to Navier–Stokes equations*, IMA J. Numer. Anal., 27 (2007), pp. 198–218.

[5] B. Ayuso and B. García-Archilla, *Regularity constants of the Stokes problem. Application to finite-element methods on curved domains*, Math. Models Methods Appl. Sci., 15 (2005), pp. 437–470.

[6] B. Ayuso, B. García-Archilla, and J. Novo, *The postprocessed mixed finite-element method for the Navier–Stokes equations*, SIAM J. Numer. Anal., 43 (2005), pp. 1091–1111.

[7] M. Bause, *On optimal convergence rates for higher-order Navier–Stokes approximations*. I. *Error estimates for the spatial discretization*, IMA J. Numer. Anal., 25 (2005), pp. 812–841.

[8] F. Brezzi and R. S. Falk, *Stability of higher-order Hood–Taylor methods*, SIAM J. Numer. Anal., 28 (1991), pp. 581–590.
[9] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1978.
[10] P. Constantin and C. Foias, *Navier–Stokes Equations*, Chicago Lectures in Math., The University of Chicago, Chicago, 1988.
[11] M. Crouzeix and P.-A. Raviart, *Conforming and nonconforming finite element methods for solving the stationary Stokes equations*, RAIRO, 7 (1973), pp. 33–75.
[12] P. J. Davis, *Gamma function and related functions*, in Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables, M. Abramowitz and I. A. Stegun, eds., Dover, New York, 1992 (reprint of the 1972 edition).
[13] E. Emmrich, *Error of the two-step BDF for the incompressible Navier–Stokes problem*, M2AN Math. Model. Numer. Anal., 38 (2004), pp. 757–764.
[14] C. Foias, O. Manley, and R. Temam, *Modelling of the interaction of small and large eddies in two dimensional turbulent flows*, RAIRO Modél. Math. Anal. Numér., 22 (1988), pp. 93–118.
[15] C. Foias, G. R. Sell, and E. S. Titi, *Exponential tracking and approximation of inertial manifolds for dissipative nonlinear equations*, J. Dynam. Differential Equations, 1 (1989), pp. 199–243.
[16] J. de Frutos, B. García-Archilla, and J. Novo, *A postprocessed Galerkin method with Chebyshev or Legendre polynomials*, Numer. Math., 86 (2000), pp. 419–442.
[17] J. de Frutos and J. Novo, *A spectral element method for the Navier–Stokes equations with improved accuracy*, SIAM J. Numer. Anal., 38 (2000), pp. 799–819.
[18] J. de Frutos and J. Novo, *A postprocess based improvement of the spectral element method*, Appl. Numer. Math., 33 (2000), pp. 217–223.
[19] J. de Frutos and J. Novo, *Postprocessing the linear finite element method*, SIAM J. Numer. Anal., 40 (2002), pp. 805–819.
[20] J. de Frutos and J. Novo, *A posteriori error estimation with the p-version of the finite element method for nonlinear parabolic differential equations*, Comput. Methods Appl. Mech. Engrg., 191 (2002), pp. 4893–4904.
[21] J. de Frutos and J. Novo, *Element-wise a posteriori estimates based on hierarchical bases for non-linear parabolic problems*, Internat. J. Numer. Methods Engrg., 63 (2005), pp. 1146–1173.
[22] J. de Frutos, B. García-Archilla, and J. Novo, *A posteriori error estimates for fully discrete nonlinear parabolic problems*, Comput. Methods Appl. Mech. Engrg., 196 (2007), pp. 3462–3474.
[23] J. de Frutos, B. García-Archilla, and J. Novo, *Postprocessing Finite-Element Methods for the Navier–Stokes Equations: The Fully Discrete Case*, manuscript.
[24] H. Fujita and T. Kato, *On the Navier–Stokes initial value problem.* I, Arch. Ration. Mech. Anal., 16 (1964), pp. 269–315.
[25] G. P. Galdi, *An Introduction to the Mathematical Theory of the Navier–Stokes Equations. Vol.* I. *Linearized Steady Problems*, Springer-Verlag, New York, 1994.
[26] B. García-Archilla, J. Novo, and E. S. Titi, *Postprocessing the Galerkin method: A novel approach to approximate inertial manifolds*, SIAM J. Numer. Anal., 35 (1998), pp. 941–972.
[27] B. García-Archilla, J. Novo, and E. S. Titi, *An approximate inertial manifold approach to postprocessing Galerkin methods for the Navier–Stokes equations*, Math. Comp., 68 (1999), pp. 893–911.
[28] B. García-Archilla and E. S. Titi, *Postprocessing the Galerkin method: The finite-element case*, SIAM J. Numer. Anal., 37 (1999), pp. 470–499.
[29] V. Girault and P. A. Raviart, *Finite Element Methods for Navier–Stokes Equations*, Springer-Verlag, Berlin, 1986.
[30] V. Girault and J. L. Lions, *Two-grid finite-element schemes for the transient Navier–Stokes problem*, M2AN Math. Model. Numer. Anal., 35 (2001), pp. 945–980.
[31] Y. He, *Two-level method based on finite element and Crank–Nicolson extrapolation for the time-dependent Navier–Stokes equations*, SIAM J. Numer. Anal., 41 (2003), pp. 1263–1285.
[32] Y. He and K. M. Liu, *A multilevel finite element method for the Navier–Stokes problem*, Numer. Methods Partial Differential Equations, 21 (2005), pp. 1052–1078.
[33] Y. He and K. M. Liu, *A multilevel spectral Galerkin method for the Navier–Stokes equations*, II: *Time discretization*, Adv. Comput. Math., 25 (2006), pp. 403–433.
[34] Y. He, K. M. Liu, and W. Sun, *Multi-level spectral Galerkin method for the Navier–Stokes problem* I: *Spatial discretization*, Numer. Math., 101 (2005), pp. 501–522.

[35] D. Henry, *Geometric Theory of Semilinear Parabolic Equations*, Springer-Verlag, Berlin, 1991.
[36] J. G. Heywood and R. Rannacher, *Finite element approximation of the nonstationary Navier–Stokes problem.* I. *Regularity of solutions and second-order error estimates for spatial discretization*, SIAM J. Numer. Anal., 19 (1982), pp. 275–311.
[37] J. G. Heywood and R. Rannacher, *Finite element approximation of the nonstationary Navier–Stokes problem.* III: *Smoothing property and higher order error estimates for spatial discretization*, SIAM J. Numer. Anal., 25 (1988), pp. 489–512.
[38] J. G. Heywood and R. Rannacher, *Finite-element approximation of the nonstationary Navier–Stokes problem.* IV: *Error analysis for second-order time discretization*, SIAM J. Numer. Anal., 27 (1990), pp. 353–384.
[39] P. Hood and C. Taylor, *A numerical solution of the Navier–Stokes equations using the finite element technique*, Comput. & Fluids, 1 (1973), pp. 73–100.
[40] Y. Hou and K. Lip, *A small eddy correction method for nonlinear dissipative evolutionary equations*, SIAM J. Numer. Anal., 41 (2003), pp. 1101–1130.
[41] C. Lubich and A. Ostermann, *Runge-Kutta time discretization of reaction-diffusion and Navier–Stokes equations: Nonsmooth-data error estimates and applications to long-time behaviour*, Appl. Numer. Math., 22 (1996), pp. 279–292.
[42] L. G. Margolin, E. S. Titi, and S. Wynne, *The postprocessing Galerkin and nonlinear Galerkin methods—A truncation analysis point of view*, SIAM J. Numer. Anal., 41 (2003), pp. 695–714.
[43] C. R. Laing, A. McRobie, and J. M. T. Thompson, *The post-processed Galerkin method applied to non-linear shell vibrations*, Dynam. Stability Systems, 14 (1999), pp. 163–181.
[44] M. A. Olshanskii, *Two-level method and some a priori estimates in unsteady Navier–Stokes calculations*, J. Comput. Appl. Math., 104 (1999), pp. 173–191.
[45] C. Sansour, P. Wriggers, and J, Sansour, *A finite element post-processed Galerkin method for dimensional reduction in the non-linear dynamics of solids. Applications to shells*, Comput. Mech., 32 (2003), pp. 104–114.
[46] S. C. Sinha, S. Redkar, V. Deshmukh, and E. A. Butcher, *Order reduction of parametrically excited nonlinear systems: Techniques and applications*, Nonlinear Dynam., 41 (2005), pp. 237–273.
[47] Y. Yan, *Postprocessing the finite element method for semilinear parabolic problems*, SIAM J. Numer. Anal., 44 (2006), pp. 1681–1702.

SIAM J. NUMER. ANAL.
Vol. 46, No. 1, pp. 231–259
© 2008 Society for Industrial and Applied Mathematics

UNIFORM l^1 BEHAVIOR FOR TIME DISCRETIZATION OF A VOLTERRA EQUATION WITH COMPLETELY MONOTONIC KERNEL II: CONVERGENCE*

XU DA†

Abstract. A previous work on the time discretization for the solution of a Volterra equation with completely monotonic convolution kernel is extended. The considered time discretization method comes from Part I [X. Da, *IMA J. Numer. Anal.*, 22 (2002), pp. 133–151], where the backward Euler is combined with order one convolution quadrature approximating the integral. In the present paper, the convergence properties of the discretization in time are given in the $l_t^1(0,\infty;H)\bigcap l_t^\infty(0,\infty;H)$ norm.

Key words. Volterra equation, completely monotonic kernel, time discretization, convolution quadrature, uniform ℓ^1 behavior, convergence

AMS subject classifications. 45K05, 65J60, 65D32

DOI. 10.1137/060668699

1. Introduction.

We continue our study, begun in Part I [3], for the discretization in time of the solution of the homogeneous linear equation

$$u_t + \int_0^t \beta(t-s)\, A\, u\,(s)\, ds = 0 \quad \text{for } t > 0, \tag{1.1}$$

$$u(0) = u_0 \in H,$$

where $u_t = \partial \mathrm{u}/\partial \mathrm{t}$ and A is a self-adjoint positive definite linear operator defined on a dense subspace $\mathcal{D}(A)$ of the real Hilbert space H, with a complete eigensystem $\{\lambda_m, \varphi_m\}_{m=1}^\infty$. The real-valued kernel $\beta(t)$ satisfies

$$\beta \in L^1(0,1),\ \beta \text{ is completely monotonic on } (0,\infty), \text{ and} \tag{1.2}$$

$$0 \le \beta(\infty) < \beta(0^+) \le \infty.$$

Following Part I, we discretize (1.1) by means of a backward Euler time difference approximation of the derivative, combined with order one convolution quadrature approximating the integral in the sense of Lubich [13]. Thus we choose a fixed time step $k > 0$, and let $t_n = n\,k$ for $n \ge 0$. This yields a numerical solution $U^n \approx u(t_n)$ satisfying

$$\overline{\partial}\, U^n + \sum_{j=1}^{n} \beta_{n-j}(k)\, A\, U^j = 0 \quad \text{for } n \ge 1, \tag{1.3}$$

$$U^0 = u_0,$$

*Received by the editors August 30, 2006; accepted for publication (in revised form) July 3, 2007; published electronically January 11, 2008. This work was supported in part by the National Natural Science Foundation of China under grant 10271046.

http://www.siam.org/journals/sinum/46-1/66869.html

†Department of Mathematics, Hunan Normal University, Changsha 410081, Hunan, People's Republic of China (daxu@hunnu.edu.cn).

where $\overline{\partial} U^n := (U^n - U^{n-1})/k$, and the convolution quadrature weights $\beta_j(k)$ are the coefficients of the power series

$$\widehat{\beta}\left(\frac{1-z}{k}\right) = \sum_{j=0}^{\infty} \beta_j(k)\, z^j.$$

Here $\widehat{\beta}(s)$ is the Laplace transform of the kernel function β, which is

$$\widehat{\beta}(s) = \int_0^{\infty} \beta(t)\, e^{-st}\, dt.$$

Part I established the stability in the norm $l_t^1(0,\infty;H) \bigcap l_t^{\infty}(0,\infty;H)$ for the discretized problem (1.3) by using the theory of Hardy spaces [4] and Laplace transform methods considered by Hannsgen and Wheeler [9]. Our present purpose is to study the convergence properties in the same norm for the problem (1.3). We shall prove the following theorem.

THEOREM 1. *Let $\beta(t)$ satisfy* (1.2). *Then we have, for the solutions of* (1.3) *and* (1.1) *and for k sufficiently small,*

$$k \sum_{n=1}^{\infty} \|U^n - u(t_n)\| \le C\, k\, |\log k|\, \|u_0\|_6, \tag{1.4}$$

where C is a positive constant independent of $u(t)$ and k, $\|.\|$ denotes the norm in H, and $\|v\|_\nu = \|A^{\nu/2} v\|$ denotes the norm in the space $\dot{H}^\nu = D(A^{\nu/2})$.

We recall some results concerning the asymptotic behavior of the solution of problem (1.1). Let $u(t,\lambda)$ denote the solution of the real equation

$$\frac{d\,u(t,\lambda)}{d\,t} + \lambda \int_0^t \beta(t-s)\, u(s,\lambda)\, ds = 0, \quad u(0,\lambda) = 1, \tag{1.5}$$

with parameter λ ($\lambda_1 \le \lambda < \infty,\ 0 \le t < \infty$).

Carr and Hannsgen give sufficient conditions in order that

$$\text{(i)} \quad \int_0^{\infty} \sup_{\lambda \ge \lambda_1} |u(t,\lambda)|\, dt < \infty, \text{ and} \tag{1.6}$$

$$\text{(ii)} \quad \int_0^{\infty} \sup_{\lambda \ge \lambda_1} |u'(t,\lambda)\, \lambda^{-1/2}| dt < \infty \qquad \left(' = \frac{d}{dt}\right)$$

hold (see [1, 2]). In particular, (1.6) holds when $\beta(t)$ satisfies

$$\beta(t) \in L^1_{loc}[0,\infty) \quad \text{is nonnegative, nonincreasing,} \tag{1.7}$$

$$\text{convex and} \quad 0 \le \beta(\infty) < \beta(0^+) \le \infty,$$

and

$$-\beta' \text{ is convex on } (0,\infty). \tag{1.8}$$

Also it is a result, from the proof of [1, Thm. 2.2] or [2, Thm. 2.3], with $\beta(t)$ as (1.7) and (1.8), that

$$\int_0^{\infty} \sup_{\lambda \ge \lambda_1} |u''(t,\lambda)\, \lambda^{-2}| dt < \infty. \tag{1.9}$$

We note that the completely monotonic kernel (1.2) satisfies (1.7) and (1.8).

The asymptotic analysis of problem (1.5) has been a very active field over the past several decades, and there are numerous very ingenious results; see, for example, the ones just stated and those of [7, 6, 9, 8, 5, 10, 11, 12, 18]. Hannsgen and Wheeler considered the completely monotonic kernel in [7, 6, 8]. Further, in [9] they studied the more general problem for (1.1) with certain families of completely monotonic kernels. Noren [18] extended their results and gave the sufficient conditions concerning the kernel $\beta(t)$ in order that

$$\int_0^\infty \sup_{\lambda\geq\lambda_1} |u''(t,\lambda)\ \lambda^{-1}|dt\ <\infty. \tag{1.10}$$

Now we state a condition from [18, Thm. 2.5].

Assume that (1.7) and (1.8) are satisfied, and in addition

$$\int_0^1 (-\log\ t)\,\beta(t)\,dt\ <\ \infty, \tag{1.11}$$

and there exist constants $c_1,\ c_2>0$ such that

$$c_1\,\tau\,A_1(\tau^{-1})\leq\ B(\tau^{-1})\leq\ c_2\,\tau\,A_1(\tau^{-1}),\qquad \rho/2\leq\ \tau, \tag{1.12}$$

where $A(\tau^{-1})$ and $A_1(\tau^{-1})$ are defined by

$$A(\tau^{-1})\ =\ \int_0^{\tau^{-1}}\beta(t)\,dt,\quad A_1(\tau^{-1})\ =\ \int_0^{\tau^{-1}} t\,\beta(t)\,dt\qquad(\tau>0),$$

and $\rho=\omega(\lambda_0)>0$ (see (2.4) below),

$$B(\tau^{-1})\ =\ \int_0^{\tau^{-1}} -t\,\beta'(t)\,dt\qquad(\tau>0).$$

Then (1.10) holds.

From [1] we know that the formula

$$u(t)\ =\ \sum_{j=1}^{\infty} u(t,\lambda_j)\ (u_0,\varphi_j)\ \varphi_j \tag{1.13}$$

gives a solution of (1.1) in the Hilbert space H. With

$$\|u(t)\|\ \leq\ \sup_{\lambda_1\leq\lambda<\infty}|u(t,\lambda)|\ \|u_0\|, \tag{1.14}$$

and using (1.6)(i), it follows that

$$\int_0^\infty \|u(t)\|dt\ \leq\ C\|u_0\|. \tag{1.15}$$

Similarly, the conditions (1.6)(ii), (1.9), and (1.10) imply that, respectively,

$$\int_0^\infty \|u'(t)\|dt\ \leq\ C\|u_0\|_1, \tag{1.16}$$

$$\int_0^\infty \|u''(t)\|dt\ \leq\ C\|u_0\|_4, \tag{1.17}$$

$$\int_0^\infty \|u''(t)\|dt\ \leq\ C\|u_0\|_2. \tag{1.18}$$

The numerical analogue of (1.5) follows such that, for $\lambda > 0$, assume that the real sequence $\{U^n(\lambda)\}$ is the solution

$$\frac{U^n(\lambda) \;-\; U^{n-1}(\lambda)}{k} \;+\; \lambda \sum_{j=1}^{n} \beta_{n-j}(k)\, U^j(\lambda) \;=\; 0, \qquad n \;\geq\; 1, \tag{1.19}$$

$$U^0(\lambda) \;=\; 1.$$

It is easy to see that the solution of (1.3) can be represented as

$$U^n \;=\; \sum_{j=1}^{\infty} U^n(\lambda_j)\, (u_0, \varphi_j)\, \varphi_j; \tag{1.20}$$

this is a discretized analogue of (1.13). We shall establish the following two additional l_t^∞ estimates.

THEOREM 2. *Let the conditions of Theorem* 1 *be in force. Then, for k sufficiently small,*

$$\|U^n - u(t_n)\| \;\leq\; C\, k\, |\log k|\, \|u_0\|_6 \;\; \textit{for } n \geq 0. \tag{1.21}$$

The proof will be given in section 3.

From Theorem 2 we get, with $0 \leq t_m \leq 1$,

$$k \sum_{n=1}^{m} \|U^n - u(t_n)\| \;\leq\; C\, k\, |\log k|\, \|u_0\|_6. \tag{1.22}$$

The error estimate (1.21) is a long time l_t^∞ error bound and contains a factor of the logarithm. This may be removed by imposing additional conditions on the kernel, for example, (1.11) (see Remark 3.2 below). Another is a long time l_t^∞ error estimate away from 0, given in the following theorem.

THEOREM 3. *If $\beta(t)$ satisfies* (1.2), *then for the solutions of* (1.3) *and* (1.1), *and for k sufficiently small, with $t_{n-2} > 0$,*

$$\|U^n - u(t_n)\| \;\leq\; C\, k\, t_{n-2}^{-2}\, \|u_0\|_6 \quad \textit{for } t_{n-2} > 0. \tag{1.23}$$

The proof will be given in section 4.

We show that Theorems 2 and 3 imply the conclusions of Theorem 1. In fact, with $t_{m-2} > 0$, we can deduce estimate

$$k \sum_{n=m}^{\infty} \|U^n - u(t_n)\| \;\leq\; C\, k\, \|u_0\|_6. \tag{1.24}$$

Formulas (1.22) and (1.24) gives us the conclusion of Theorem 1. We have shown that Theorem 1 is a consequence of Theorems 2 and 3, which we will prove in sections 3 and 4, respectively.

Formulas (1.21) and (1.23) are similar to those in [16, 17] for the classical weakly singular kernel $\beta(t) = t^{-\alpha}$ $(0 < \alpha < 1)$ [16] and the exponential decay kernel [17]; it is known that these types of kernels satisfy (1.2).

Comparing [15] and [14], Theorem 3 is similar to Theorem 2.1 of [15] for a scalar function, and (1.22) is similar to Theorem 3.3 of [14] (see also the references therein)

for a scalar function. Thus, Theorems 2 and 3 extend the results of [14] and [15] for a scalar function to a function in an infinite-dimensional Hilbert space H, where it is a solution of (1.1).

In sections 3 and 4 we shall use different techniques in [14, 15] to obtain the proofs of Theorems 2 and 3, respectively. [14, Theorem 3.3] is a shorter time error estimate, and [15, Theorem 2.1] contains a long time error bound. We do not know whether the results from those used in [14, 15] could be used to extend or simplify the proofs of Theorems 2 and 3, but they might be worth looking at as a next step.

A completely monotonic function on $(0,\infty)$ can be represented (see [20]) by the Stieltjes integral

$$\beta(t) \;=\; \int_0^\infty e^{-xt}\, d\,\alpha(x) \quad \text{for} \;\; t>0, \tag{1.25}$$

where $\alpha(0)=0$, α is nondecreasing, and $\alpha(x)=\alpha(x^-)$ for $0<x<\infty$. The conditions in (1.2) are equivalent, respectively, to

$$\int_0^\infty \frac{1-e^{-x}}{x}\, d\,\alpha(x) \;<\; \infty, \tag{1.26}$$

$$\alpha(0^+) \;<\; \alpha(\infty) \;=\; \beta(0^+). \tag{1.27}$$

Define

$$\phi(\sigma,\tau)=\int_0^\infty \frac{(x+\sigma)d\alpha(x)}{(x+\sigma)^2+\tau^2}, \;\; \theta(\sigma,\tau)=\int_0^\infty \frac{d\alpha(x)}{(x+\sigma)^2+\tau^2} \quad (\sigma+i\,\tau \;\notin\; (-\infty,0]),$$

$$D(s;\lambda)=D(\sigma+i\tau;\lambda)=\frac{s}{\lambda}+\widehat{\beta}(s)=\frac{\sigma+i\tau}{\lambda}+\widehat{\beta}(\sigma+i\tau),$$

$$\phi(\tau)=\phi(0,\tau), \qquad \theta(\tau)=\theta(0,\tau) \qquad (0 \;<\; \tau \;<\infty).$$

Note that

$$\begin{aligned}\phi(\sigma,\tau) \;-\; i\,\tau\,\theta(\sigma,\tau) \;&=\; \int_0^\infty \frac{d\alpha(x)}{x+\sigma+i\tau} \qquad (\sigma+i\,\tau \;\notin\; (-\infty,0])\\ &=\; \int_0^\infty e^{-t\,(\sigma+i\tau)}\,\beta(t)\,dt \qquad (\sigma>0).\end{aligned}$$

Then the Fourier transform of $\beta(t)$ is

$$\widetilde{\beta}(\tau)=\int_0^\infty e^{-i\,\tau\,t}\beta(t)\,dt=\int_0^\infty \frac{d\alpha(x)}{x+i\,\tau}=\widehat{\beta}(i\,\tau)=\phi(\tau)-i\,\tau\,\theta(\tau), \quad \tau\neq 0, \tag{1.28}$$

and (see [1, Thm. 2.2 and Cor. 2.1])

$$\limsup_{\tau\to\infty} \frac{\theta(\tau)}{\phi(\tau)} \;<\; \infty. \tag{1.29}$$

We note that (1.29) is a necessary condition in order that (1.6)(i) holds [1, Thm. 2.2 (iii)]. Also, using the definition of $\theta(\tau)$, we see that $\theta(\tau)$ is nonnegative, continuously differentiable, and strictly decreasing on R^+.

The result of Theorem 1 needs more restrictive conditions on the initial data u_0. $A^3 u_0 \in H$ is not desirable from a practical point of view. If we add the conditions on the kernel $\beta(t)$, for example,

$$\limsup_{\tau\to\infty} \frac{\tau\theta(\tau)}{\phi(\tau)} < \infty, \tag{1.30}$$

then we have the following improved estimate.

COROLLARY 1. *Let $\beta(t)$ satisfy* (1.2) *and* (1.30). *Then for the solutions of* (1.3) *and* (1.1) *and for k sufficiently small, we have that*

$$k \sum_{n=1}^{\infty} \|U^n - u(t_n)\| \le C\, k\, |\log k|\, \|u_0\|_4. \tag{1.31}$$

The proof will be given at the end of section 5.

On the reader's part we assume some familiarity with the introduction to Part I. Further material from Part I will be explicitly referenced where it is required.

From the conditions given by Noren [18] and by a comparison between (1.17) and (1.18), we can conjecture that the regularity requirements on the initial data may be reduced to that, a sharper case given by McLean and Thomée [17, Thm. 4.2]. They considered an exponential decay kernel and obtained the exponential decay error bounds in the norm $l_t^\infty(0,\infty;H)$ based on the same time discretization scheme. These error bounds imply the estimate (1.32) below. The completely monotonic kernel (1.2) does not have exponential decay. As examples, consider the following three kernels all satisfying (1.2) but neither (A1) nor (A2) of [17]: (i) $\beta(t) = (1+t)^{-1}$, (ii) $\beta(t) = t^{-1/2}$, (iii) $\beta(t) = t^{-1/2}[\log(t+3)]^{-1}$. Thus one cannot apply McLean and Thomée's [17] results in the present case. Note that the exponential decay kernels given by [17] are completely monotonic as well and satisfy (1.2); in particular, the exponential kernel $\beta(t) = e^{-t}$ satisfies (1.2), (1.11), and $\beta(0^+) = 1 < \infty$. Thus, for this type of kernel Theorem 1 holds and from [18, Thm. 2.5(a)] (1.10) holds, but Corollary 1 cannot follow because it does not satisfy (1.30) (see [3]). Lubich, Sloan, and Thomée [16] considered the classical weakly singular kernel $\beta(t) = t^{-\alpha}$ $(0 < \alpha < 1)$ and obtained an $l_t^\infty(0,\infty;H)$ estimate for nonsmooth data and positive time. An easy calculation shows that it satisfies (1.2), (1.11), and (1.12), as well as (1.30), but neither (A1) nor (A2) of [17]. With this class of kernels, (1.10) holds, and in a forthcoming paper we shall investigate in detail how to reduce the spatial regularity requirement on the initial data in the estimate (1.31), and on the contrary increase the regularity requirement in Theorem 3.2 of [16]] to derive a sharper l^1 estimate (1.32) below. Seeking sufficient and necessary conditions in order that (1.31) hold when the spatial regularity is reduced further, from $A^2 u_0 \in H$ to $Au_0 \in H$, requires further investigation. We guess that the conditions which yield that (1.10) holds guarantee that (1.32) below holds.

We close the introduction with the following.

Conjecture. If $\beta(t)$ satisfies (1.2), (1.11), and (1.12), then for the solutions of (1.3) and (1.1), and for k sufficiently small, we have that

$$k \sum_{n=1}^{\infty} \|U^n - u(t_n)\| \le C\, k\, \|u_0\|_2. \tag{1.32}$$

2. Preliminaries. We collect here some known results concerning $\phi, \theta, u(t,\lambda)$. Except where noted, only (1.7) is assumed to hold. In the papers cited, some of

the inequalities are proved for the case where $\beta(\infty) = 0$; the extension to our more general situation is straightforward.

Under the assumptions (1.7) and (1.8), by [1, (4.36)] we have the formula

$$u(t,\lambda) = \mathrm{Re}\,[\lambda^{-1}u_1(t) + \lambda^{-2}u_2(t) + \lambda^{-3}u_3(t) + u_4(t,\lambda) + u_5(t,\lambda) + u_6(t,\lambda)], \tag{2.1}$$

where

$$u_1(t) = \frac{1}{\pi\, t}\int_0^{\varepsilon} e^{i\theta\, t}\,\frac{\widehat{\beta}'(i\theta)\,d\theta}{[\widehat{\beta}(i\theta)]^2}, \tag{2.2}$$

$$u_2(t) = \frac{1}{\pi\, t}\int_0^{\varepsilon} e^{i\theta\, t}\,\frac{1}{[\widehat{\beta}(i\theta)]^2}\left(1 - \frac{2\,i\theta\widehat{\beta}'(i\theta)}{\widehat{\beta}(i\theta)}\right) d\theta,$$

$$u_3(t) = -\frac{1}{\pi\, t}\int_0^{\varepsilon} e^{i\theta\, t}\,\frac{2\,i\theta\,d\theta}{[\widehat{\beta}(i\theta)]^3},$$

$$u_4(t,\lambda) = \frac{1}{\pi\, t\,\lambda^3}\int_0^{\varepsilon} e^{i\theta\, t}\,\frac{(i\theta)^2\,D_s(i\theta;\lambda)}{[\widehat{\beta}(i\theta)]^2\,D(i\theta;\lambda)}\left(\frac{2}{\widehat{\beta}(i\theta)} + \frac{1}{D(i\theta;\lambda)}\right) d\theta,$$

$$u_5(t,\lambda) = \frac{1}{\pi\, t\,\lambda}\int_{\varepsilon}^{\frac{\pi}{k}} e^{i\theta\, t}\,\frac{D_s(i\theta;\lambda)}{[D(i\theta;\lambda)]^2}\,d\theta,$$

$$u_6(t,\lambda) = \frac{1}{\pi\, t\,\lambda}\int_{\frac{\pi}{k}}^{\infty} e^{i\theta\, t}\,\frac{D_s(i\theta;\lambda)}{[D(i\theta;\lambda)]^2}\,d\theta.$$

By [5], $\infty \geq \lim_{\tau\to 0^+}|\widetilde{\beta}(\tau)| = \int_0^{\infty}\beta(t)dt > 0$, so there exists an $\varepsilon > 0$ such that

$$|\widetilde{\beta}(\tau)| > 2\tau \qquad (0 < \tau \leq \varepsilon). \tag{2.3}$$

As in [1] (see also [3]), we can define an increasing function $\omega(\lambda) : [\lambda_0,\infty) \to [2\varepsilon,\infty)$ (here $\lambda_0 \geq 1$ and sufficiently large) with $\omega(\lambda) \to \infty\,(\lambda\to\infty)$ and

$$\theta(\omega(\lambda)) = \frac{1}{\lambda} \qquad (\lambda_0 \leq \lambda < \infty). \tag{2.4}$$

We extend ω to $[\lambda_1,\infty)$ if necessary by defining $\omega(\lambda) = 2\varepsilon$ $(\lambda_1 \leq \lambda \leq \lambda_0)$.

From [1, (4.24)] and [2, (2.10) and (2.11)], there exist positive constants c and C such that

$$c\,\lambda^{1/2} \leq \omega(\lambda) \leq C\,\lambda \quad (\lambda_1 \leq \lambda). \tag{2.5}$$

By [1, Lem. 4.1 and 5.2] and [2, Lem. 5.2] (see also [19]), $\phi(\tau)$ is continuously

differentiable on R^+, and there exists a positive constant C such that

$$\frac{1}{2\sqrt{2}}\,A(\tau^{-1}) \,\le\, |\widetilde{\beta}(\tau)| \,\le\, 4\,A(\tau^{-1}) \qquad (\tau \,>\, 0), \tag{2.6}$$

$$\frac{1}{5}\,A_1(\tau^{-1}) \,\le\, \theta(\tau) \,\le\, 12\,A_1(\tau^{-1}) \qquad (\tau \,>\, 0), \tag{2.7}$$

$$|\widetilde{\beta}'(\tau)| \,\le\, 40\,A_1(\tau^{-1}) \qquad (\tau \,>\, 0), \tag{2.8}$$

$$|D(i\tau;\lambda)| \,\ge\, C\,(\tau\theta(\tau) \,+\, |\widetilde{\beta}(\tau)|) \qquad \left(\varepsilon \le \tau \le \frac{\omega(\lambda)}{2}\right), \tag{2.9}$$

$$|D(i\tau;\lambda)|^2 \,\ge\, C\,\left(\phi^2(\tau) \,+\, \left(\frac{\tau-\omega}{\lambda}\right)^2\right) \qquad (\tau \ge \varepsilon). \tag{2.10}$$

By [1, Lem. 5.1], (1.8) implies that

$$|\widetilde{\beta}''(\tau)| \,\le\, 6000\int_0^{\frac{1}{\tau}} t^2\,\beta(t)\,dt \qquad (\tau \,>\, 0). \tag{2.11}$$

We recall an estimate of Shea and Wainger [19, (1.21)], namely,

$$\int_0^{\varepsilon} \frac{A_1(\tau^{-1})}{A^2(\tau^{-1})}\,d\tau \,<\, \infty. \tag{2.12}$$

Finally, we remark that

$$\phi(\tau) \,>\, 0, \qquad \tau \,>\, 0, \tag{2.13}$$

except for certain piecewise linear kernels β for which $\phi(\tau)$ vanishes on a sequence tending to infinity [5, Lem. 4]. This fact, together with (2.7), implies that (2.13) holds whenever $\limsup_{\tau\to\infty} \frac{\theta(\tau)}{\phi(\tau)} < \infty$. Note that (2.1) holds with (2.13) instead of (1.8).

3. Proof of Theorem 2. From now on, unless noted otherwise, we assume that (1.2) holds. Let ε denote an appropriately small positive constant, independent of k, n, and let C denote a positive constant which depends only on ε; λ_1, the value of C, may change from line to line. We introduce the notation

$$\sigma \,=\, \sigma(k,\theta) \,=\, \frac{1-\cos\,k\theta}{k}, \quad \tau \,=\, \tau(k,\theta) \,=\, \frac{\sin\,k\theta}{k},$$

$$s \,=\, s(k,\theta) \,=\, \sigma + \,i\,\tau \,=\, \frac{1-e^{-i\,k\theta}}{k}.$$

Multiplying (1.19) by z^n, and then summing from $n = 1$ to infinity, we find that

$$\widetilde{U}(z,\lambda) \,=\, U^1(\lambda)\,z \,+\, U^2(\lambda)\,z^2 \,+\cdots+\, U^n(\lambda)\,z^n \,+\cdots$$

satisfies

$$\widetilde{U}(z,\lambda) \,=\, \frac{z}{k}\,\frac{1}{\left(\frac{1-z}{k}\right) \,+\, \lambda\,\widehat{\beta}\left(\frac{1-z}{k}\right)} \,=\, \frac{z}{k}\,\widehat{u}\left(\frac{1-z}{k},\lambda\right), \tag{3.1}$$

where $\widehat{u}(s,\lambda)$ is the Laplace transform of the solution of problem (1.5).

Set $V^n(\lambda) = U^n(\lambda) - 1$. Then the generating function

$$\widetilde{V}(z,\lambda) \;=\; \sum_{n=0}^{\infty} V^n(\lambda)\, z^n$$

obeys the relation

$$\widetilde{V}(z,\lambda) \;=\; \frac{z}{k}\,\widehat{K}\left(\frac{1-z}{k},\lambda\right), \tag{3.2}$$

with

$$\widehat{K}(s,\lambda) \;=\; -\;\frac{\widehat{\beta}(s)}{s\,D(s;\lambda)}.$$

This function is analytic at $z=0$, so that $\widetilde{V}(z,\lambda)$ has a positive radius of convergence, and hence, for ρ sufficiently small,

$$\begin{aligned} V^n(\lambda) \;&=\; \frac{1}{2\pi\, i}\int_{|\zeta|=\rho} \zeta^{-n-1}\,\widetilde{V}(\zeta,\lambda)\,d\zeta \\ &=\; \frac{k^{-1}}{2\pi\, i}\int_{|\zeta|=\rho} \zeta^{-n}\,\widehat{K}\left(\frac{1-\zeta}{k},\lambda\right)\,d\zeta. \end{aligned} \tag{3.3}$$

We now change variables $\zeta = e^{-kz}$ and deform the resulting contour

$$\left\{z = k^{-1}\,\log(1/\rho) \;+\; i\,y; |y| \le\; \frac{\pi}{k}\right\}$$

to

$$\Gamma_k = \left\{z = i\tau, |\tau| \le\; \frac{\pi}{k}\right\}$$

(note that $\widehat{K}\big(\frac{1-e^{-z\,k}}{k};\lambda\big)$ is analytic in $\prod^+ = \{z:\ \operatorname{Re} z \;\ge\; 0,\ z \neq 0\}$ and the contributions from the line segments on $\operatorname{Im} z = \pm\frac{\pi}{k}$ cancel by periodicity), and we get

$$V^n(\lambda) \;=\; \operatorname{Re}\left\{\frac{1}{\pi}\int_0^{\frac{\pi}{k}}\; e^{i\theta\, t_{n-1}}\,\widehat{K}\left(\frac{1-e^{-i\,\theta\,k}}{k},\lambda\right)\,d\theta\right\}. \tag{3.4}$$

On the other hand, if we let $v(t,\lambda) = u(t,\lambda) - 1$, then its Laplace transform is

$$\widehat{v}(z,\lambda) \;=\; \widehat{u}(z,\lambda) \;-\; \frac{1}{z} \;=\; \widehat{K}(z,\lambda).$$

Thus, by applying the Laplace inversion formula, since $\widehat{K}(z,\lambda)$ is analytic in $\prod^+$,

$$v(t_n,\lambda) \;=\; \operatorname{Re}\left\{\frac{1}{\pi}\;\int_0^{\infty}\; e^{i\theta\, t_n}\;\widehat{K}(i\theta,\lambda)\,d\theta\right\}. \tag{3.5}$$

Subtracting (3.5) from (3.4), we obtain

$$\begin{aligned} U^n(\lambda) \;-\; u(t_n,\lambda) \;=\; \frac{1}{\pi}\operatorname{Re}\{&V_1^n(\lambda) \;-\; v_1(t_n,\lambda) \;+\; V_2^n(\lambda) \;-\; v_2(t_n,\lambda) \\ &+\; V_3^n(\lambda) \;-\; v_3(t_n,\lambda) \;-\; v_4(t_n,\lambda)\}, \end{aligned} \tag{3.6}$$

where

$$V_1^n(\lambda) - v_1(t_n,\lambda) = E_{1,1}^n(\lambda) + E_{1,2}^n(\lambda) \tag{3.7}$$
$$= \int_0^{\varepsilon} e^{i\theta\, t_{n-1}}\,(1 - e^{i\theta\, k})\,\widehat{K}(i\theta,\lambda)\,d\theta$$
$$+ \int_0^{\varepsilon} e^{i\theta\, t_{n-1}}\,(\widehat{K}(s,\lambda) - \widehat{K}(i\theta,\lambda))\,d\theta,$$

$$V_2^n(\lambda) - v_2(t_n,\lambda) = E_{2,1}^n(\lambda) + E_{2,2}^n(\lambda) \tag{3.8}$$
$$= \int_{\varepsilon}^{\frac{\varepsilon}{k}} e^{i\theta\, t_{n-1}}\,(1 - e^{i\theta\, k})\,\widehat{K}(i\theta,\lambda)\,d\theta$$
$$+ \int_{\varepsilon}^{\frac{\varepsilon}{k}} e^{i\theta\, t_{n-1}}\,(\widehat{K}(s,\lambda) - \widehat{K}(i\theta,\lambda))\,d\theta,$$

$$V_3^n(\lambda) - v_3(t_n,\lambda) = \int_{\frac{\varepsilon}{k}}^{\frac{\pi}{k}} e^{i\theta\, t_{n-1}}\,\widehat{K}(s,\lambda)\,d\theta \tag{3.9}$$
$$- \int_{\frac{\varepsilon}{k}}^{\frac{\pi}{k}} e^{i\theta\, t_n}\,\widehat{K}(i\theta,\lambda)\,d\theta,$$

$$v_4(t_n,\lambda) = \int_{\frac{\pi}{k}}^{\infty} e^{i\theta\, t_n}\,\widehat{K}(i\theta,\lambda)\,d\theta. \tag{3.10}$$

If $\beta(\infty) > 0$, then (2.6) shows that

$$|\widetilde{\beta}(\tau)| \le C\,\tau^{-1} \qquad (0 < \tau \le \varepsilon),$$

and (2.13) gives us

$$|D(i\tau;\lambda)| \ge \max\left\{\phi(\tau),\ \frac{\beta(\infty)-\tau^2}{\tau}\right\} \ge \frac{C}{\tau},\quad 0<\tau\le\varepsilon. \tag{3.11}$$

Thus,

$$\int_0^{\varepsilon} \frac{|\widetilde{\beta}(\tau)|}{|D(i\tau;\lambda)|}\,d\tau \le C. \tag{3.12}$$

On the other hand, if $\beta(\infty) = 0$, then (2.6) implies that $|\widetilde{\beta}(\tau)| \le 4\,A(\tau^{-1})$ and

$$|D(i\tau;\lambda)| \ge \max\left\{2^{-3/2}A(\tau^{-1}) - \tau,\ \phi(\tau)\right\}, \tag{3.13}$$

so (3.12) is again valid. Then a simple estimate and (3.12) lead to

$$|E_{1,1}^n(\lambda)| \le C\,k\int_0^{\varepsilon} \theta\,|\widehat{K}(i\theta,\lambda)|\,d\theta \tag{3.14}$$
$$= C\,k\int_0^{\varepsilon} \frac{|\widetilde{\beta}(\theta)|}{|D(i\theta;\lambda)|}\,d\theta \le C\,k\,.$$

We remark that (3.14) holds without the assumption (1.2), which needs only (1.7) and (2.13).

By the mean-value theorem we have that

$$|E^n_{1,2}(\lambda)| \;\le\; C\,k \int_0^\varepsilon \theta^2\,|\widehat{K}'(s_1,\lambda)|\,d\theta, \tag{3.15}$$

where $s_1 = s_1(k,\theta) = \sigma_1(k,\theta)+i\tau_1(k,\theta)$, with $0 \le \sigma_1 < \sigma(k,\theta),\ \tau(k,\theta) \le \tau_1(k,\theta) < \theta$, and

$$\widehat{K}'(s,\lambda) \;=\; -\frac{\widehat{\beta}'(s)}{s\,D(s;\lambda)} \;+\; \frac{\widehat{\beta}(s)}{s^2\,D(s;\lambda)} \;+\; \frac{\widehat{\beta}(s)\,D_s(s;\lambda)}{s\,D^2(s;\lambda)}.$$

In Part I [3, (3.8) and (3.10)] we have proved that when (1.2) holds and $0 \le \theta \le \varepsilon$,

$$c_0\,\theta \;\le\; \tau(k,\theta) \;\le\; \theta, \tag{3.16}$$

$$|D(\sigma_1+i\tau_1;\lambda)| \ge\; \frac{1}{10\sqrt{2}}\,|\widetilde{\beta}(\tau_1)| \ge\; \frac{1}{160}\,|\widetilde{\beta}(\varepsilon)|, \tag{3.17}$$

and for $\sigma_1 \ge 0,\ \tau_1 > 0$ (see [3, (3.13) and (3.14)]),

$$|\widehat{\beta}(\sigma_1+i\tau_1)| \le\; 4\,|\widetilde{\beta}(\tau_1)|, \tag{3.18}$$

$$|\widehat{\beta}'(\sigma_1+i\tau_1)| \;\le\; \theta(\tau_1), \tag{3.19}$$

with $c_0 < 1$ as close as we like to 1.

Combining these estimates, (2.6), and (2.7) yields, when $0 \le \theta \le \varepsilon$,

$$|D(\sigma_1+i\tau_1;\lambda)| \ge\; C\,|\widetilde{\beta}(\tau)|, \tag{3.20}$$

and for $\sigma_1 \ge 0,\ \tau_1 \ge \tau > 0$,

$$|\widehat{\beta}(\sigma_1+i\tau_1)| \le\; C\,|\widetilde{\beta}(\tau)|, \tag{3.21}$$

$$|\widehat{\beta}'(\sigma_1+i\tau_1)| \;\le\; C\,\theta(\tau), \tag{3.22}$$

so that, after trivial estimates, some simple computations, and using (3.20) through (3.22), (2.6), and (2.7),

$$\begin{aligned}
\int_0^\varepsilon \theta^2\,|\widehat{K}'(s_1,\lambda)|\,d\theta \;&\le\; \int_0^\varepsilon \theta^2\,\frac{|\widehat{\beta}'(s_1)|}{|s_1||D(s_1;\lambda)|}\,d\theta \\
&+\; \int_0^\varepsilon \theta^2\,\frac{|\widehat{\beta}(s_1)|}{|s_1|^2|D(s_1;\lambda)|}\,d\theta \;+\; \int_0^\varepsilon \theta^2\,\frac{|\widehat{\beta}(s_1)||D_s(s_1;\lambda)|}{|s_1||D(s_1;\lambda)|^2}\,d\theta \\
&\le\; C\left[\int_0^\varepsilon \frac{\tau\theta(\tau)}{|\widetilde{\beta}(\tau)|}\,d\tau \;+\; \int_0^\varepsilon \frac{|\widetilde{\beta}(\tau)|}{|\widetilde{\beta}(\tau)|}\,d\tau\right. \\
&+\; \left.\int_0^\varepsilon \frac{\tau\,|\widetilde{\beta}(\tau)||\lambda^{-1}+\theta(\tau)|}{|\widetilde{\beta}(\tau)|^2}\,d\tau\right] \;\le\; C.
\end{aligned} \tag{3.23}$$

Combining (3.15) and (3.23) shows

$$|E^n_{1,2}(\lambda)| \le C\,k. \tag{3.24}$$

From (3.7), (3.14), and (3.24) we obtain

$$|V^n_1(\lambda) - v_1(t_n,\lambda)| \le C\,k. \tag{3.25}$$

Next we consider the estimate of $E^n_{2,1}(\lambda)$. After trivial estimates we obtain that

$$|E^n_{2,1}(\lambda)| \le C\,k \int_\varepsilon^{\frac{\varepsilon}{k}} \frac{|\widetilde{\beta}(\theta)|}{D(i\theta;\lambda)|}\,d\theta. \tag{3.26}$$

In order to bound the integral of (3.26) we split it into five cases according to the position relation of ω and $\frac{\varepsilon}{k}$. We use the estimates of (2.9) and (2.10) for lower bounds on $|D(i\theta;\lambda)|$. Inequality (2.6) will give upper bounds on the numerators.

For the remainder of the proof, the estimates (1.27) and (2.4)–(2.8) will be exploited throughout without explicit mention.

Case 1. If $\omega+\varepsilon < \frac{\varepsilon}{k}$, then (2.9) and (2.10) imply

$$\begin{aligned}\int_\varepsilon^{\frac{\varepsilon}{k}} \frac{|\widetilde{\beta}(\theta)|}{D(i\theta;\lambda)|}\,d\theta &= \int_\varepsilon^{\frac{\omega}{2}} + \int_{\frac{\omega}{2}}^{\omega-\varepsilon} + \int_{\omega-\varepsilon}^{\omega+\varepsilon} + \int_{\omega+\varepsilon}^{\frac{\varepsilon}{k}}. \\ &\le C\left(\int_\varepsilon^{\frac{\omega}{2}} d\tau + \int_{\frac{\omega}{2}}^{\omega-\varepsilon} \frac{\lambda}{|\tau-\omega|}\,d\tau + \int_{\omega-\varepsilon}^{\omega+\varepsilon} \frac{1}{\phi(\tau)}\,d\tau\right. \\ &\quad \left. + \lambda\int_{\omega+\varepsilon}^{\frac{\varepsilon}{k}} \frac{|\widehat{\beta}(i\,\tau)|}{|\tau-\omega|}\,d\tau\right) \\ &\le C\,(\lambda + \lambda^2 + \lambda\,|\log k|).\end{aligned} \tag{3.27}$$

Case 2. If $\omega < \frac{\varepsilon}{k} \le \omega+\varepsilon$, then, again using (2.9) and (2.10), we obtain

$$\begin{aligned}\int_\varepsilon^{\frac{\varepsilon}{k}} \frac{|\widetilde{\beta}(\theta)|}{D(i\theta;\lambda)|}\,d\theta &= \int_\varepsilon^{\frac{\omega}{2}} + \int_{\frac{\omega}{2}}^{\omega-\varepsilon} + \int_{\omega-\varepsilon}^{\omega+\varepsilon}. \\ &\le C\left(\int_\varepsilon^{\frac{\omega}{2}} d\tau + \int_{\frac{\omega}{2}}^{\omega-\varepsilon} \frac{\lambda}{|\tau-\omega|}\,d\tau + \int_{\omega-\varepsilon}^{\omega+\varepsilon} \frac{1}{\phi(\tau)}\,d\tau\right) \\ &\le C\,(\lambda + \lambda^2).\end{aligned} \tag{3.28}$$

Case 3. If $\omega-\varepsilon < \frac{\varepsilon}{k} \le \omega$ or $\frac{\omega}{2} < \frac{\varepsilon}{k} \le \omega-\varepsilon$ or $\frac{\varepsilon}{k} \le \frac{\omega}{2}$, then we have

$$\int_\varepsilon^{\frac{\varepsilon}{k}} \frac{|\widetilde{\beta}(\theta)|}{D(i\theta;\lambda)|}\,d\theta \le C\,(\lambda+\lambda^2). \tag{3.29}$$

For example, when $\omega-\varepsilon < \frac{\varepsilon}{k} \le \omega$, by (2.9) and (2.10)

$$\begin{aligned}\int_\varepsilon^{\frac{\varepsilon}{k}} \frac{|\widetilde{\beta}(\theta)|}{D(i\theta;\lambda)|}\,d\theta &= \int_\varepsilon^{\frac{\omega}{2}} + \int_{\frac{\omega}{2}}^{\omega-\varepsilon} + \int_{\omega-\varepsilon}^{\omega}. \\ &\le C\left(\int_\varepsilon^{\frac{\omega}{2}} d\tau + \int_{\frac{\omega}{2}}^{\omega-\varepsilon} \frac{\lambda}{|\tau-\omega|}\,d\tau + \int_{\omega-\varepsilon}^{\omega} \frac{1}{\phi(\tau)}\,d\tau\right) \\ &\le C\,(\lambda + \lambda^2).\end{aligned} \tag{3.30}$$

Together the above five estimates yield

$$|E^n_{2,1}(\lambda)| \le C\,k\lambda^2|\log k|. \tag{3.31}$$

Now we turn to the estimate of $E^n_{2,2}(\lambda)$. First, applying the mean-value theorem we have

$$|E^n_{2,2}(\lambda)| \le C\,k\int_\varepsilon^{\frac{\varepsilon}{k}} \theta^2|\widehat{K}'(s_1,\lambda)|d\theta. \tag{3.32}$$

Then, comparing (3.32) with (3.15) and (3.23), we find that the integrand of (3.32) has the same form as these, so that

$$\begin{aligned}\int_\varepsilon^{\frac{\varepsilon}{k}} \theta^2|\widehat{K}'(s_1,\lambda)|d\theta \le C\,\Bigg(&\int_{c_0\varepsilon}^{\frac{\varepsilon}{k}} \tau^2\frac{|\widehat{\beta}'(s_1)|}{|s_1|\,|D(s_1;\lambda)|}\,d\tau\\ &+\int_{c_0\varepsilon}^{\frac{\varepsilon}{k}} \tau^2\frac{|\widehat{\beta}(s_1)|}{|s_1|^2\,|D(s_1;\lambda)|}\,d\tau+\int_{c_0\varepsilon}^{\frac{\varepsilon}{k}} \tau^2\frac{|\widehat{\beta}(s_1)|\,|D_s(s_1;\lambda)|}{|s_1|\,|D(s_1;\lambda)|^2}\,d\tau\Bigg),\end{aligned} \tag{3.33}$$

where c_0 is given by (3.34) below.

Here we have to bound the various terms on the right-hand side. For this purpose we recall from Part I [3, (3.22), (2.10), (3.23), and (3.24)] that, in the case $\varepsilon \le \theta \le \frac{\varepsilon}{k}$,

$$c_0\theta \le \tau(k,\theta) \le \theta, \tag{3.34}$$

$$|D(s_1;\lambda)| \ge \phi(\sigma_1,\tau_1) \ge C\,\phi(\tau_1), \tag{3.35}$$

$$|D(s_1;\lambda)| \ge C\left(\tau_1\,\theta(\tau_1) + A(\tau_1^{-1})\right), \quad \tau \in \left[c_0\varepsilon, \frac{c_0\omega}{2}\right] \cap \left[c_0\varepsilon, \frac{c_0\varepsilon}{k}\right], \tag{3.36}$$

$$|D(s_1;\lambda)| \ge C\frac{|\tau_1-\omega|}{\lambda}, \ \ \omega < \frac{\omega}{1-\widetilde{\varepsilon}} \le \tau, \ \ \widetilde{\varepsilon} \le \frac{1}{2}. \tag{3.37}$$

where $c_0 < 1$ can be chosen arbitrarily close to 1.

Also, using the definition of $\phi(\tau)$ in the introduction, we have that, when $\varepsilon \le \theta \le \frac{\varepsilon}{k}$,

$$\begin{aligned}|\phi(\tau)-\phi(\tau_1)| &= \left|\int_0^\infty \frac{x(\tau_1+\tau)(\tau_1-\tau)}{(x^2+\tau^2)(x^2+\tau_1^2)}\,d\alpha(x)\right|\\ &\le C\int_0^\infty \frac{x\theta(\theta-\tau)}{(x^2+\tau^2)\theta^2}\,d\alpha(x) \le C\,\varepsilon_1\,\phi(\tau),\end{aligned} \tag{3.38}$$

with $\left|\frac{\sin k\theta}{k\theta}-1\right| \le \varepsilon_1$ appropriately small.

Thus, these estimates, along with (2.6) and (2.7), give us, for $\varepsilon \le \theta \le \frac{\varepsilon}{k}$,

$$|D(s_1;\lambda)| \ge C\,\phi(\tau), \tag{3.39}$$

$$|D(s_1;\lambda)| \ge C\left(\tau\,\theta(\tau) + A(\tau^{-1})\right), \quad \tau \in \left[c_0\varepsilon, \frac{c_0\omega}{2}\right] \cap \left[c_0\varepsilon, \frac{c_0\varepsilon}{k}\right], \tag{3.40}$$

$$|D(s_1;\lambda)| \ge C\frac{|\tau-\omega|}{\lambda}, \ \ \omega < \frac{\omega}{1-\widetilde{\varepsilon}} \le \tau, \ \ \widetilde{\varepsilon} \le \frac{1}{2}. \tag{3.41}$$

In order to obtain a similar bound on (3.33), we split it into three cases. The estimates (3.21) and (3.22) will be used for estimates of the numerators. Inequalities (3.39)–(3.41) will give lower bounds on $|D(s_1;\lambda)|$, but in the proof below these estimates will be mentioned without going into detail.

When $2\,\omega < \frac{\varepsilon}{k}$,

$$\int_{\varepsilon}^{\frac{\varepsilon}{k}} \theta^2\,|\widehat{K}'(s_1,\lambda)|\,d\theta \;\le\; \left[\int_{c_0\varepsilon}^{\frac{c_0\omega}{2}} + \int_{\frac{c_0\omega}{2}}^{2\omega} + \int_{2\omega}^{\frac{\varepsilon}{k}}\ .\right] \tag{3.42}$$

$$\le\; C\left(\int_{c_0\,\varepsilon}^{\frac{c_0\omega}{2}} \left(1+\frac{\lambda^{-1}}{\theta(\tau)}\right)\,d\tau + \int_{\frac{c_0\omega}{2}}^{2\omega}\left(\frac{1}{\phi(\tau)}+\frac{\tau\,\lambda^{-1}}{\phi^2(\tau)}\right)\,d\tau\right.$$

$$\left. +\lambda\int_{2\omega}^{\frac{\varepsilon}{k}} \frac{|\widehat{\beta}(i\,\tau)|}{\tau-\omega}d\tau\right)$$

$$\le\; C\,(\lambda\;+\;\lambda^2\;+\;\lambda^3\;+\;\lambda\,|\log k|),$$

where we have used (see [1, (6.8)])

$$\theta(\tau)\;\le\; C\,\lambda^{-1},\qquad \tau\ge\;\frac{c_0\omega}{2}, \tag{3.43}$$

and an easily obtained inequality

$$|\widetilde{\beta}(\tau)|\le\;4\int_0^{\frac{1}{\tau}}\beta(t)\,dt \tag{3.44}$$

$$\le\;4\int_0^{\frac{1}{\varepsilon_0}}\beta(t)\,dt\;=\;C(\varepsilon_0)\,,\qquad \tau\ge\varepsilon_0.$$

When $\frac{\omega}{2}\le\;\frac{\varepsilon}{k}\le\;2\omega$,

$$\int_{\varepsilon}^{\frac{\varepsilon}{k}}\theta^2\,|\widehat{K}'(s_1,\lambda)|\,d\theta\le\left[\int_{c_0\varepsilon}^{\frac{c_0\omega}{2}}+\int_{\frac{c_0\omega}{2}}^{2\omega}\ .\right] \tag{3.45}$$

$$\le\;C\left(\int_{c_0\,\varepsilon}^{\frac{c_0\omega}{2}}\left(1\;+\;\frac{\lambda^{-1}}{\theta(\tau)}\right)\,d\tau\;+\;\int_{\frac{c_0\omega}{2}}^{2\omega}\left(\frac{1}{\phi(\tau)}\;+\;\frac{\tau\,\lambda^{-1}}{\phi^2(\tau)}\right)\,d\tau\;\right)$$

$$\le\;C\,(\lambda\;+\;\lambda^2\;+\;\lambda^3\;).$$

When $\frac{\varepsilon}{k}<\frac{\omega}{2}$,

$$\int_{\varepsilon}^{\frac{\varepsilon}{k}}\theta^2\,|\widehat{K}'(s_1,\lambda)|\,d\theta\le\left[\int_{c_0\varepsilon}^{\frac{c_0\omega}{2}}+\int_{\frac{c_0\omega}{2}}^{\frac{\omega}{2}}\ .\right] \tag{3.46}$$

$$\le\;C\left(\int_{c_0\,\varepsilon}^{\frac{c_0\omega}{2}}\left(1\;+\;\frac{\lambda^{-1}}{\theta(\tau)}\right)\,d\tau\;+\;\int_{\frac{c_0\omega}{2}}^{\frac{\omega}{2}}\left(\frac{1}{\phi(\tau)}\;+\;\frac{\tau\,\lambda^{-1}}{\phi^2(\tau)}\right)\,d\tau\;\right)$$

$$\le\;C\,(\lambda\;+\;\lambda^2\;+\;\lambda^3\;).$$

Together these three estimates and (3.32), (3.33) show that

$$|E^n_{2,2}(\lambda)| \;\le\; C\,k\,|\log\,k|\,\lambda^3. \tag{3.47}$$

Combining (3.31) and (3.47) we have obtained that

$$|V^n_2(\lambda) \;-\; v_2(t_n,\lambda)| \;\le\; C\,k\,|\log\,k|\;\lambda^3. \tag{3.48}$$

For the estimate of $V^n_3(\lambda)$ we recall from Part I [3, (3.18) and (3.19)] that when $\frac{\varepsilon}{k} \le \theta \le \frac{\pi}{k}$,

$$\sigma(k,\theta) \;\ge\; \frac{1-\cos\,\varepsilon}{k} = \frac{C(\varepsilon)}{k}, \;\; |D(\sigma+i\tau;\lambda)| \ge \;\frac{\sigma}{\lambda} \ge \;\frac{C(\varepsilon)}{k\lambda}, \tag{3.49}$$

and for $\sigma \ge 0$,

$$|\widehat{\beta}(\sigma+i\tau)| \le \int_0^\infty \;\frac{d\alpha(x)}{x+\sigma} = \widehat{\beta}(\sigma). \tag{3.50}$$

We use (3.49) and (3.50) to obtain

$$\begin{aligned} |V^n_3(\lambda)| \;&\le\; \int_{\frac{\varepsilon}{k}}^{\frac{\pi}{k}} \;|\widehat{K}(s,\lambda)|\;d\theta \\ &\le C\,k\int_{\frac{\varepsilon}{k}}^{\frac{\pi}{k}} \;\frac{|\widehat{\beta}(\sigma)|\lambda}{\sigma}\;d\theta \;\le\; C\,k\int_0^{\frac{k}{c(\varepsilon)}} \;\beta(t)\;dt\;\lambda. \end{aligned} \tag{3.51}$$

Finally, we will derive an estimate on

$$|v_3(t_n,\lambda)| \;+\; |v_4(t_n,\lambda)| \;\le\; \int_{\frac{\varepsilon}{k}}^\infty \;|\widehat{K}(i\theta,\lambda)|\;d\theta. \tag{3.52}$$

In the case $2\omega < \frac{\varepsilon}{k}$,

$$\begin{aligned} \int_{\frac{\varepsilon}{k}}^\infty \;|\widehat{K}(i\theta,\lambda)|\;d\theta \;&\le\; \int_{\frac{\varepsilon}{k}}^\infty \;\frac{\lambda|\widehat{\beta}(i\theta)|}{\theta|\theta-\omega|}\;d\theta \\ &\le\; C\,k\,\lambda\int_0^{\frac{k}{\varepsilon}} \;\beta(t)\;dt. \end{aligned} \tag{3.53}$$

In the case $\omega < \frac{\varepsilon}{k} \le\; 2\omega$ or $\frac{\omega}{2} \le\; \frac{\varepsilon}{k} \le\; \omega$,

$$\begin{aligned} \int_{\frac{\varepsilon}{k}}^\infty \;|\widehat{K}(i\theta,\lambda)|\;d\theta \;&\le\; \int_0^{\frac{k}{\varepsilon}} \;\beta(t)\;dt\;\left(\frac{k}{\varepsilon}\int_{\frac{\varepsilon}{k}}^{2\omega} \;\frac{d\theta}{\phi(\theta)} \;+\; \int_{2\omega}^\infty \frac{\lambda\,d\theta}{\theta(\theta-\omega)}\right) \\ &\le\; C\,k\,\lambda^2\int_0^{\frac{k}{\varepsilon}} \;\beta(t)\;dt. \end{aligned} \tag{3.54}$$

In the case $\frac{\varepsilon}{k} < \frac{\omega}{2}$,

$$\int_{\frac{\varepsilon}{k}}^\infty \;|\widehat{K}(i\theta,\lambda)|\;d\theta \;\le\; \int_{\frac{\varepsilon}{k}}^{\frac{\omega}{2}} \;.\; +\; \int_{\frac{\omega}{2}}^{2\omega} \;.\; +\; \int_{2\omega}^\infty \;.\; \le\; C\,k\int_0^{\frac{k}{\varepsilon}} \;\beta(t)\;dt\;(\lambda+\lambda^2). \tag{3.55}$$

Taken together with (3.53)–(3.55), this shows that

$$\int_{\frac{\varepsilon}{k}}^{\infty} |\widehat{K}(i\theta,\lambda)|\, d\theta \;\le\; C\,k\,\lambda^2 \int_0^{\frac{k}{\varepsilon}} \beta(t)\, dt. \tag{3.56}$$

Using (3.6)–(3.10), (3.25), (3.48), (3.51), (3.52), and (3.56), and combining (1.13) and (1.20), we have completed the proof of Theorem 2.

Remark 3.1. If adding the condition (1.30), then we can show that

$$|E^n_{2,2}(\lambda)| \;\le\; C\,k\,(\lambda+\lambda^2+\lambda\,|\log k|), \tag{3.47$'$}$$

and hence, we have the improved estimate

$$\|U^n - u(t_n)\| \;\le\; C\,k\,|\log k|\;\|u_0\|_4 \qquad \text{for } n \ge 0. \tag{3.57}$$

Remark 3.2. We guess that a thorough analysis may remove the factor $|\log k|$ from (1.21). On the other hand if we add the conditions (1.11) on the kernel $\beta(t)$, then we have that, in the case $\omega+\varepsilon < \frac{\varepsilon}{k}$,

$$\int_{\omega+\varepsilon}^{\frac{\varepsilon}{k}} \frac{|\widehat{\beta}(i\,\tau)|}{|\tau-\omega|}\, d\tau \;\le\; C, \tag{3.58}$$

and when $2\,\omega < \frac{\varepsilon}{k}$,

$$\int_{2\,\omega}^{\frac{\varepsilon}{k}} \frac{|\widehat{\beta}(i\,\tau)|}{|\tau-\omega|}\, d\tau \;\le\; C, \tag{3.59}$$

so that we obtain the desired estimate

$$\|U^n \;-\; u(t_n)\| \;\le\; C\;k\;\|u_0\|_6 \;\text{ for } n \ge 0. \tag{3.60}$$

4. Proof of Theorem 3. To prove Theorem 3 we need the expressions of (2.1) and (2.2) for the solution of (1.5). Similarly, we recall from [3, (4.7) and (4.8)] that for the solution of (1.19),

$$U^n(\lambda) \;=\; \mathrm{Re}\,[\lambda^{-1}U^n_1 \;+\; \lambda^{-2}U^n_2 \;+\; \lambda^{-3}U^n_3 \;+\; U^n_4(\lambda) \;+\; U^n_5(\lambda)\,], \tag{4.1}$$

where

$$\begin{aligned}
U^n_1 \;&=\; \frac{1}{\pi\,t_{n-1}} \int_0^{\varepsilon} e^{i\theta\,t_{n-2}}\, \frac{\widehat{\beta}'(s)\,d\theta}{[\widehat{\beta}(s)]^2}, \\
U^n_2 \;&=\; \frac{1}{\pi\,t_{n-1}} \int_0^{\varepsilon} e^{i\theta\,t_{n-2}}\, \frac{1}{[\widehat{\beta}(s)]^2} \left(1 \;-\; \frac{2\,s\widehat{\beta}'(s)}{\widehat{\beta}(s)}\right) d\theta, \\
U^n_3 \;&=\; -\;\frac{1}{\pi\,t_{n-1}} \int_0^{\varepsilon} e^{i\theta\,t_{n-2}}\, \frac{2\,s\,d\theta}{[\widehat{\beta}(s)]^3}, \\
U^n_4(\lambda) \;&=\; \frac{1}{\pi\,t_{n-1}\lambda^3} \int_0^{\varepsilon} e^{i\theta\,t_{n-2}}\, \frac{s^2\,D_s(s;\lambda)}{[\widehat{\beta}(s)]^2\,D(s;\lambda)} \left(\frac{2}{\widehat{\beta}(s)} \;+\; \frac{1}{D(s;\lambda)}\right) d\theta, \\
U^n_5(\lambda) \;&=\; \frac{1}{\pi\,t_{n-1}\,\lambda} \int_{\varepsilon}^{\frac{\pi}{k}} e^{i\theta\,t_{n-2}}\, \frac{D_s(s;\lambda)}{[D(s;\lambda)]^2}\, d\theta.
\end{aligned} \tag{4.2}$$

By subtracting (2.1) from (4.1) we can write

(4.3)
$$U^n(\lambda) - u(t_n,\lambda) \;=\; \mathrm{Re}\{\lambda^{-1}(U_1^n - u_1(t_n)) \;+\; \lambda^{-2}(U_2^n - u_2(t_n))$$
$$+\; \lambda^{-3}(U_3^n - u_3(t_n)) \;+\; (U_4^n(\lambda) - u_4(t_n,\lambda)) \;+\; (U_5^n(\lambda) - u_5(t_n,\lambda)) - u_6(t_n,\lambda)\}.$$

We shall estimate the above various terms. The estimates (1.29) and (2.4)–(2.12) will be used throughout without explicit mention.

Estimate 1.

$$|U_1^n \;-\; u_1(t_n)| \;\le\; C\,k\,t_{n-2}^{-2} \quad (t_{n-2} > 0). \tag{4.4}$$

Write $U_1^n - u_1(t_n) \;=\; e_{1,1}^n + e_{1,2}^n + e_{1,3}^n$, where

$$e_{1,1}^n \;=\; \frac{1}{\pi\, t_{n-1}} \int_0^{\varepsilon} e^{i\theta\, t_{n-2}} \left[\widehat{K}_1(s) \;-\; \widehat{K}_1(i\theta)\right] d\theta, \tag{4.5}$$

$$e_{1,2}^n \;=\; \frac{1}{\pi\, t_{n-1}} \int_0^{\varepsilon} e^{i\theta\, t_{n-2}} (1 \;-\; e^{2i\theta\, k})\, \widehat{K}_1(i\theta)\, d\theta, \tag{4.6}$$

$$e_{1,3}^n \;=\; \frac{1}{\pi} \left(\frac{1}{t_{n-1}} \;-\; \frac{1}{t_n}\right) \int_0^{\varepsilon} e^{i\theta\, t_n}\, \widehat{K}_1(i\theta)\, d\theta, \tag{4.7}$$

$$\widehat{K}_1(s) \;=\; \frac{\widehat{\beta}'(s)}{[\widehat{\beta}(s)]^2}.$$

Evidently, in view of (2.12),

$$|e_{1,3}^n| \;\le\; C\,k\,t_{n-2}^{-2} \quad (t_{n-2} > 0). \tag{4.8}$$

Integrate by parts in (4.6) to bring out another factor of t_{n-2}^{-1},

$$e_{1,2}^n = \frac{1}{i\,\pi\, t_{n-1}\, t_{n-2}}\, e^{i\theta\, t_{n-2}} \left(1 \;-\; e^{2i\theta\, k}\right) \widehat{K}_1(i\theta)\Big|_{\theta=0}^{\theta=\varepsilon} \tag{4.9}$$
$$+\; \frac{2\,k}{\pi\, t_{n-1}\, t_{n-2}} \int_0^{\varepsilon} e^{i\theta\, t_n}\, \widehat{K}_1(i\theta)\, d\theta$$
$$-\; \frac{1}{i\,\pi\, t_{n-1}\, t_{n-2}} \int_0^{\varepsilon} e^{i\theta\, t_{n-2}} \left(1 \;-\; e^{2i\theta\, k}\right) \widehat{K}_1'(i\theta)\, d\theta,$$

where

$$\widehat{K}_1'(s) \;=\; \frac{\widehat{\beta}''(s)}{\widehat{\beta}^2(s)} \;-\; \frac{2[\widehat{\beta}'(s)]^2}{\widehat{\beta}^3(s)}.$$

After some estimates we have

$$|e_{1,2}^n| \;\le\; C\,k\,t_{n-2}^{-2} \quad (t_{n-2} > 0), \tag{4.10}$$

for example, from (2.11), (2.6), (2.8), and (2.12) we see that

$$\int_0^{\varepsilon} \theta|\widehat{K}_1'(i\theta)|d\theta \le C\int_0^{\varepsilon}\left[\frac{\theta|\widetilde{\beta}''(\theta)|}{|\widetilde{\beta}(\theta)|^2} + \frac{\theta|\widetilde{\beta}'(\theta)|^2}{|\widetilde{\beta}(\theta)|^3}\right]d\theta \tag{4.11}$$

$$\le \int_0^{\varepsilon}\frac{A_1(\theta^{-1})}{A^2(\theta^{-1})}\,d\theta < \infty.$$

For $e_{1,1}^n$, again using integration by parts gives us

$$|e_{1,1}^n| \le \left|\frac{1}{i\,\pi\,t_{n-1}\,t_{n-2}}\,e^{i\theta\,t_{n-2}}\,\left(\widehat{K}_1(s) - \widehat{K}_1(i\theta)\right)\right|_{\theta=0}^{\theta=\varepsilon} \tag{4.12}$$

$$+\left|\frac{1}{\pi\,t_{n-1}\,t_{n-2}}\int_0^{\varepsilon} e^{i\theta\,t_{n-3}}\,\left(\widehat{K}_1'(s) - \widehat{K}_1'(i\theta)\right)\,d\theta\right|$$

$$+\left|\frac{1}{\pi\,t_{n-1}\,t_{n-2}}\int_0^{\varepsilon} e^{i\theta\,t_{n-3}}\,(1-e^{i\theta\,k})\,\widehat{K}_1'(i\theta)\,d\theta\right|$$

$$= e_{1,1,1}^n + e_{1,1,2}^n + e_{1,1,3}^n.$$

Clearly, (4.11) shows

$$e_{1,1,3}^n \le C\,k\,t_{n-2}^{-2} \quad (t_{n-2} > 0). \tag{4.13}$$

By the mean-value theorem we obtain

$$|\widehat{K}_1(s) - \widehat{K}_1(i\theta)| \le C\,k\,\theta^2\,|\widehat{K}_1'(s_1)|, \tag{4.14}$$

where s_1 is given as in the form of (3.15).

We recall from [3, (5.3) and (5.4)] that, for $0 < \theta \le \varepsilon,\ \tau \le \tau_1 \le \theta$,

$$|\widehat{\beta}(s_1)| \ge C\,|\widetilde{\beta}(\tau_1)| \ge C\,|\widetilde{\beta}(\tau)|, \tag{4.15}$$

and for $\sigma_1 \ge 0,\ \tau_1 > 0$,

$$|\widehat{\beta}''(s_1)| \le \frac{2}{\tau_1}\,\theta(\tau_1), \tag{4.16}$$

which implies that, for $\sigma_1 \ge 0,\ \tau_1 \ge \tau > 0$,

$$|\widehat{\beta}''(s_1)| \le C\,\tau^{-1}\,\theta(\tau). \tag{4.17}$$

Returning to (4.14), using the estimate of (4.15) for lower bounds on $|\widehat{\beta}(s_1)|$, and of (3.22) and (4.17) for upper bounds on the numerators, we obtain, for $0 \le \theta \le \varepsilon$,

$$\theta^2\,|\widehat{K}_1'(s_1)| \le C,$$

which is

$$|\widehat{K}_1(s) - \widehat{K}_1(i\theta)| \le C\,k. \tag{4.18}$$

Similarly, we have

$$|\widehat{K}_1'(s) \;-\; \widehat{K}_1'(i\theta)| \;\le\; C\,k\,\theta^2\;|\widehat{K}_1''(s_1)|, \tag{4.19}$$

where

$$\widehat{K}_1''(s) = \frac{\widehat{\beta}'''(s)}{\widehat{\beta}^2(s)} \;-\; \frac{6\widehat{\beta}'(s)\widehat{\beta}''(s)}{\widehat{\beta}^3(s)} \;+\; \frac{6[\widehat{\beta}'(s)]^3}{\widehat{\beta}^4(s)}.$$

However, for $\sigma_1 \ge 0,\ \tau_1 \ge \tau > 0$, we have

$$|s_1\widehat{\beta}'''(s_1)| \le 6\int_0^\infty \frac{d\alpha(x)}{|x+s_1|^3} \le \frac{6}{\tau_1}\theta(\tau_1) \le C\tau^{-1}\theta(\tau). \tag{4.20}$$

Using this estimate, along with (4.15), (4.17), and (3.22), we obtain

$$|\widehat{K}_1'(s) \;-\; \widehat{K}_1'(i\theta)| \;\le\; C\,k\frac{|\widetilde{\beta}'(\tau)|}{|\widetilde{\beta}(\tau)|^2}. \tag{4.21}$$

It then follows from (2.12) and (4.12) that

$$e^n_{1,1,2} \le C\,k\,t_{n-2}^{-2} \quad (t_{n-2} > 0). \tag{4.22}$$

Combining (4.12), (4.13), (4.18), and (4.22), we obtain

$$|e^n_{1,1}| \le C\,k\,t_{n-2}^{-2} \quad (t_{n-2} > 0). \tag{4.23}$$

Together the three estimates (4.8), (4.10), and (4.23) complete the proof of Estimate 1.

Estimate 2.

$$|U_2^n \;-\; u_2(t_n)| \le C\,k\,t_{n-2}^{-2} \quad (t_{n-2} > 0). \tag{4.24}$$

Also, integrating by parts on U_2^n and $u_2(t_n)$, we write $U_2^n - u_2(t_n) = e^n_{2,1} + e^n_{2,2}$, where

$$e^n_{2,1} \;=\; \frac{1}{i\,\pi\,t_{n-1}\,t_{n-2}}\;e^{i\theta\,t_{n-2}}\;\widehat{K}_{2,1}(s)\Big|_{\theta=0}^{\theta=\varepsilon} \;-\; \frac{1}{i\pi\,t_n^2}\;e^{i\theta\,t_n}\;\widehat{K}_{2,1}(i\theta)\Big|_{\theta=0}^{\theta=\varepsilon}$$

and

$$e^n_{2,2} \;=\; \frac{2}{\pi\,t_{n-1}t_{n-2}}\int_0^\varepsilon e^{i\theta\,t_{n-3}}\,\widehat{K}_{2,2}(s)\,d\theta \;-\; \frac{2}{\pi\,t_n^2}\int_0^\varepsilon e^{i\theta\,t_n}\widehat{K}_{2,2}(i\theta)\,d\theta,$$

$$\widehat{K}_{2,1}(s) \;=\; \frac{1}{[\widehat{\beta}(s)]^2}\left(1 - \frac{2\,s\widehat{\beta}'(s)}{\widehat{\beta}(s)}\right),$$

$$\widehat{K}_{2,2}(s) \;=\; \frac{\widehat{\beta}'(s)}{[\widehat{\beta}(s)]^3}\left(1 - \frac{2\,s\widehat{\beta}'(s)}{\widehat{\beta}(s)}\right) \;+\; \frac{\widehat{\beta}'(s) + s\widehat{\beta}''(s)}{[\widehat{\beta}(s)]^3} - \frac{s[\widehat{\beta}'(s)]^2}{[\widehat{\beta}(s)]^4}.$$

Furthermore, by simple computations we have

$$
\begin{aligned}
|e^n_{2,1}| &\le e^n_{2,1,1} \;+\; e^n_{2,1,2} \;+\; e^n_{2,1,3} \;+\; e^n_{2,1,4} \qquad (4.25)\\
&= \left| \frac{1}{i\,\pi\,t_{n-1}\,t_{n-2}}\; e^{i\theta\,t_{n-2}}\; \Big(\widehat{K}_{2,1}(s) \;-\; \widehat{K}_{2,1}(i\theta)\Big) \right|^{\theta=\varepsilon}_{\theta=0} \\
&\quad + \left| \frac{1}{i\,\pi\,t_{n-1}\,t_{n-2}}\; e^{\theta\,t_{n-2}}\; \big(1 \;-\; e^{2i\theta\,k}\big)\, \widehat{K}_{2,1}(i\theta) \right|^{\theta=\varepsilon}_{\theta=0} \\
&\quad + \left| \frac{1}{i\,\pi\,t_{n-1}} \left(\frac{1}{t_{n-2}} \;-\; \frac{1}{t_n}\right) e^{i\theta\,t_n}\; \widehat{K}_{2,1}(i\theta) \right|^{\theta=\varepsilon}_{\theta=0} \\
&\quad + \left| \frac{1}{i\,\pi\,t_n} \left(\frac{1}{t_{n-1}} \;-\; \frac{1}{t_n}\right) e^{i\theta\,t_n}\; \widehat{K}_{2,1}(i\theta) \right|^{\theta=\varepsilon}_{\theta=0}
\end{aligned}
$$

and

$$
\begin{aligned}
|e^n_{2,2}| &\le \left| \frac{2}{\pi\,t_{n-1}t_{n-2}} \int_0^{\varepsilon} e^{i\theta\,t_{n-3}}\; \Big(\widehat{K}_{2,2}(s) \;-\; \widehat{K}_{2,2}(i\theta)\Big)\; d\theta \;\right| \qquad (4.26)\\
&\quad + \left| \frac{2}{\pi\,t_{n-1}t_{n-2}} \int_0^{\varepsilon} e^{i\theta\,t_{n-3}}\; \big(1 \;-\; e^{3\,i\theta\,k}\big)\; \widehat{K}_{2,2}(i\theta)\; d\theta \;\right| \\
&\quad + \left| \frac{2}{\pi\,t_{n-1}} \left(\frac{1}{t_{n-2}} \;-\; \frac{1}{t_n}\right) \int_0^{\varepsilon} e^{i\theta\,t_n}\; \widehat{K}_{2,2}(i\theta)\; d\theta \;\right| \\
&\quad + \left| \frac{2}{\pi\,t_n} \left(\frac{1}{t_{n-1}} \;-\; \frac{1}{t_n}\right) \int_0^{\varepsilon} e^{i\theta\,t_n}\; \widehat{K}_{2,2}(i\theta)\; d\theta \;\right| \\
&= e^n_{2,2,1} \;+\; e^n_{2,2,2} \;+\; e^n_{2,2,3} \;+\; e^n_{2,2,4}.
\end{aligned}
$$

It is now a clear matter that

$$
e^n_{2,1,2} \;+\; e^n_{2,1,3} \;+\; e^n_{2,1,4} \;\le\; C\,k\,t^{-2}_{n-2}\,, \qquad (4.27)
$$

$$
\begin{aligned}
&e^n_{2,2,2} \;+\; e^n_{2,2,3} \;+\; e^n_{2,2,4} \qquad (4.28)\\
&\le\; C\,k\,t^{-2}_{n-2} \left(1 + \int_0^{\varepsilon} \frac{|\widetilde{\beta}'(\theta)|}{|\widetilde{\beta}(\theta)|^2}\, d\theta\right) \;\le\; C\,k\,t^{-2}_{n-2}.
\end{aligned}
$$

For $e^n_{2,1,1}$ and $e^n_{2,2,1}$ we first use the mean-value theorem to obtain

$$
|\widehat{K}_{2,1}(s) - \widehat{K}_{2,1}(i\theta)| \;\le\; C\,k\,\theta^2 |\widehat{K}'_{2,1}(s_1)|, \qquad (4.29)
$$

$$
|\widehat{K}_{2,2}(s) - \widehat{K}_{2,2}(i\theta)| \;\le\; C\,k\,\theta^2 |\widehat{K}'_{2,2}(s_1)|, \qquad (4.30)
$$

with

$$\widehat{K}'_{2,1}(s) = -\frac{4\widehat{\beta}'(s)}{\widehat{\beta}^3(s)} - \frac{2\,s\widehat{\beta}''(s)}{\widehat{\beta}^3(s)} + \frac{6\,s[\widehat{\beta}'(s)]^2}{\widehat{\beta}^4(s)},$$

$$\widehat{K}'_{2,2}(s) = \frac{3\widehat{\beta}''(s)}{\widehat{\beta}^3(s)} - \frac{9\,[\widehat{\beta}'(s)]^2}{\widehat{\beta}^4(s)} + \frac{s\,\widehat{\beta}'''(s)}{\widehat{\beta}^3(s)}$$

$$- \frac{9\,s\widehat{\beta}'(s)\widehat{\beta}''(s)}{\widehat{\beta}^4(s)} + \frac{12\,s[\widehat{\beta}'(s)]^3}{\widehat{\beta}^5(s)}.$$

From (3.22), (4.15), (4.17) we have that, for $0 \le \theta \le \varepsilon$,

$$\theta^2\,|\widehat{K}'_{2,1}(s_1)| \le C; \tag{4.31}$$

in other words

$$e^n_{2,1,1} \le C\,k\,t^{-2}_{n-2}. \tag{4.32}$$

Taken together with (4.25) and (4.27), this shows that

$$|e^n_{2,1}| \le C\,k\,t^{-2}_{n-2}. \tag{4.33}$$

Similarly, from (4.15), (4.17), (3.22), and (4.20) we see that

$$\theta^2\,|\widehat{K}'_{2,2}(s_1)| \le C, \tag{4.34}$$

and it follows that

$$e^n_{2,2,1} \le C\,k\,t^{-2}_{n-2}. \tag{4.35}$$

Combining (4.26), (4.28), and (4.35) we now have established the estimate

$$|e^n_{2,2}| \le C\,k\,t^{-2}_{n-2}. \tag{4.36}$$

This completes the proof of Estimate 2.

We now state and prove two more estimates needed to deduce the bounds for the third and fourth terms on the right-hand side of (4.3).

Estimate 3.

$$|U^n_3 - u_3(t_n)| \le C\,k\,t^{-2}_{n-2} \quad (t_{n-2} > 0). \tag{4.37}$$

We argue similarly to the proof of Estimate 2. Let $U^n_3 - u_3(t_n) = e^n_{3,1} + e^n_{3,2}$, where

$$e^n_{3,1} = -\left\{ \frac{1}{i\,\pi\,t_{n-1}\,t_{n-2}}\, e^{i\theta\,t_{n-2}}\, \widehat{K}_{3,1}(s)\,|_{\theta=\varepsilon} - \frac{1}{i\pi\,t_n^2}\, e^{i\theta\,t_n}\, \widehat{K}_{3,1}(i\theta)\,|_{\theta=\varepsilon} \right\}$$

and

$$e^n_{3,2} = \frac{2}{\pi\,t_{n-1}t_{n-2}} \int_0^\varepsilon e^{i\theta\,t_{n-3}}\, \widehat{K}_{3,2}(s)\,d\theta - \frac{2}{\pi\,t_n^2} \int_0^\varepsilon e^{i\theta\,t_n}\, \widehat{K}_{3,2}(i\theta)\,d\theta,$$

$$\widehat{K}_{3,1}(s) = \frac{2\,s}{[\widehat{\beta}(s)]^3},$$

$$\widehat{K}_{3,2}(s) = \frac{1}{[\widehat{\beta}(s)]^3} - \frac{3\,s\,\widehat{\beta}'(s)}{[\widehat{\beta}(s)]^4}.$$

Similarly, we have

$$|e^n_{3,1}| \le \left| \frac{1}{i\,\pi\,t_{n-1}\,t_{n-2}}\; e^{i\theta\,t_{n-2}}\; \left(\widehat{K}_{3,1}(s) \;-\; \widehat{K}_{3,1}(i\theta) \right) \Big|_{\theta=\varepsilon} \right| \tag{4.38}$$

$$+ \left| \frac{1}{i\,\pi\,t_{n-1}\,t_{n-2}}\; e^{i\theta\,t_{n-2}}\; \left(1 \;-\; e^{2i\theta\,k}\right) \widehat{K}_{3,1}(i\theta) \Big|_{\theta=\varepsilon} \right|$$

$$+ \left| \frac{1}{i\,\pi\,t_{n-1}} \left(\frac{1}{t_{n-2}} \;-\; \frac{1}{t_n} \right)\; e^{i\theta\,t_n}\; \widehat{K}_{3,1}(i\theta) \Big|_{\theta=\varepsilon} \right|$$

$$+ \left| \frac{1}{i\,\pi\,t_n} \left(\frac{1}{t_{n-1}} \;-\; \frac{1}{t_n} \right)\; e^{i\theta\,t_n}\; \widehat{K}_{3,1}(i\theta) \Big|_{\theta=\varepsilon} \right|$$

and

$$|e^n_{3,2}| \le \left| \frac{2}{\pi\,t_{n-1}t_{n-2}} \int_0^{\varepsilon} e^{i\theta\,t_{n-3}}\; \left(\widehat{K}_{3,2}(s) \;-\; \widehat{K}_{3,2}(i\theta) \right)\; d\theta \;\right| \tag{4.39}$$

$$+ \left| \frac{2}{\pi\,t_{n-1}t_{n-2}} \int_0^{\varepsilon} e^{i\theta\,t_{n-3}}\; \left(1 \;-\; e^{3\,i\theta\,k}\right)\; \widehat{K}_{3,2}(i\theta)\; d\theta \;\right|$$

$$+ \left| \frac{2}{\pi\,t_{n-1}} \left(\frac{1}{t_{n-2}} \;-\; \frac{1}{t_n} \right) \int_0^{\varepsilon} e^{i\theta\,t_n}\; \widehat{K}_{3,2}(i\theta)\; d\theta \;\right|$$

$$+ \left| \frac{2}{\pi\,t_n} \left(\frac{1}{t_{n-1}} \;-\; \frac{1}{t_n} \right) \int_0^{\varepsilon} e^{i\theta\,t_n}\; \widehat{K}_{3,2}(i\theta)\; d\theta \;\right|.$$

An analogous argument can be used to obtain the proof of Estimate 3.

Estimate 4.

$$|U^n_4(\lambda) - u_4(t_n,\lambda)| \;\le\; C\,k\,t^{-2}_{n-2}\,\lambda^{-3} \qquad (t_{n-2} > 0). \tag{4.40}$$

We first integrate by parts on $U^n_4(\lambda)$ and $u_4(t_n,\lambda)$ in order to bring out another factor of t^{-1}_{n-2} and t^{-1}_n, respectively. This gives us $U^n_4(\lambda) - u_4(t_n,\lambda) = e^n_{4,1}(\lambda) + e^n_{4,2}(\lambda)$, where

$$e^n_{4,1}(\lambda) \;=\; \frac{1}{i\,\pi\,t_{n-1}\,t_{n-2}\lambda^3}\; e^{i\theta\,t_{n-2}}\; \widehat{K}_{4,1}(s;\lambda) \Big|_{\theta=\varepsilon}$$

$$-\; \frac{1}{i\pi\,t^2_n\lambda^3}\; e^{i\theta\,t_n}\; \widehat{K}_{4,1}(i\theta;\lambda) \Big|_{\theta=\varepsilon}$$

and

$$\begin{aligned} e^n_{4,2}(\lambda) &= -\left\{ \frac{1}{\pi\, t_{n-1}t_{n-2}\lambda^3} \int_0^\varepsilon e^{i\theta\, t_{n-3}}\, \widehat{K}_{4,2}(s;\lambda)\, d\theta \right. \\ &\quad \left. - \frac{1}{\pi\, t_n^2\lambda^3} \int_0^\varepsilon e^{i\theta\, t_n} \widehat{K}_{4,2}(i\theta;\lambda)\, d\theta \right\}, \\ \widehat{K}_{4,1}(s;\lambda) &= \frac{s^2 D_s(s;\lambda)}{[\widehat{\beta}(s)]^2 D(s;\lambda)} \left(\frac{2}{\widehat{\beta}(s)} + \frac{1}{D(s;\lambda)} \right), \\ \widehat{K}_{4,2}(s;\lambda) &= \frac{2\, s\, D_s(s;\lambda) + s^2\widehat{\beta}''(s)}{[\widehat{\beta}(s)]^2 D(s;\lambda)} \left(\frac{2}{\widehat{\beta}(s)} + \frac{1}{D(s;\lambda)} \right) \\ &\quad - \frac{s^2\, D_s(s;\lambda)}{[\widehat{\beta}(s)]^2\, D(s;\lambda)} \left(\frac{6\widehat{\beta}'(s)}{[\widehat{\beta}(s)]^2} + \frac{4\,\widehat{\beta}'(s) + 2\,\lambda^{-1}}{\widehat{\beta}(s)\, D(s;\lambda)} + \frac{2D_s(s;\lambda)}{D^2(s;\lambda)} \right). \end{aligned}$$

Arguing similarly to (4.38) and (4.39), we have that

$$\begin{aligned} |e^n_{4,1}| \le & \left| \frac{1}{i\,\pi\, t_{n-1}\, t_{n-2}\lambda^3}\, e^{i\theta\, t_{n-2}} \left(\widehat{K}_{4,1}(s;\lambda) - \widehat{K}_{4,1}(i\theta;\lambda) \right) \Big|_{\theta=\varepsilon} \right| \\ & + \left| \frac{1}{i\,\pi\, t_{n-1}\, t_{n-2}\lambda^3}\, e^{i\theta\, t_{n-2}} \left(1 - e^{2i\theta\, k}\right) \widehat{K}_{4,1}(i\theta;\lambda) \Big|_{\theta=\varepsilon} \right| \\ & + \left| \frac{1}{i\,\pi\, t_{n-1}\,\lambda^3} \left(\frac{1}{t_{n-2}} - \frac{1}{t_n} \right) e^{i\theta\, t_n}\, \widehat{K}_{4,1}(i\theta;\lambda) \Big|_{\theta=\varepsilon} \right| \\ & + \left| \frac{1}{i\,\pi\, t_n\,\lambda^3} \left(\frac{1}{t_{n-1}} - \frac{1}{t_n} \right) e^{i\theta\, t_n}\, \widehat{K}_{4,1}(i\theta;\lambda) \Big|_{\theta=\varepsilon} \right| \end{aligned} \tag{4.41}$$

and

$$\begin{aligned} |e^n_{4,2}| \le & \left| \frac{1}{\pi\, t_{n-1}t_{n-2}\lambda^3} \int_0^\varepsilon e^{i\theta\, t_{n-3}} \left(\widehat{K}_{4,2}(s;\lambda) - \widehat{K}_{4,2}(i\theta;\lambda) \right) d\theta \right| \\ & + \left| \frac{1}{\pi\, t_{n-1}t_{n-2}\lambda^3} \int_0^\varepsilon e^{i\theta\, t_{n-3}} \left(1 - e^{3\,i\theta\, k}\right) \widehat{K}_{4,2}(i\theta;\lambda)\, d\theta \right| \\ & + \left| \frac{1}{\pi\, t_{n-1}\lambda^3} \left(\frac{1}{t_{n-2}} - \frac{1}{t_n} \right) \int_0^\varepsilon e^{i\theta\, t_n}\, \widehat{K}_{4,2}(i\theta;\lambda)\, d\theta \right| \\ & + \left| \frac{1}{\pi\, t_n\lambda^3} \left(\frac{1}{t_{n-1}} - \frac{1}{t_n} \right) \int_0^\varepsilon e^{i\theta\, t_n}\, \widehat{K}_{4,2}(i\theta;\lambda)\, d\theta \right|. \end{aligned} \tag{4.42}$$

It is easily seen that the second and fourth terms of (4.41) and (4.42) are bounded by $C\, k\, t_{n-2}^{-2}\lambda^{-3}$, respectively. For the first term of (4.41) and (4.42), we apply the mean-value theorem, then use the estimates of (3.22), (4.17), and (4.20) for upper bounds on the numerators, and of (3.20) and (4.15) for lower bounds on $|\widehat{\beta}(s_1)|$ and $|D(s_1;\lambda)|$. Therefore we deduce the proof of Estimate 4.

Now integrate by parts on $U_5^n(\lambda)$, $u_5(t_n,\lambda)$, and $u_6(t_n,\lambda)$ to obtain

$$U_5^n(\lambda) \;-\; u_5(t_n,\lambda) \;=\; e_{5,1}^n(\lambda) \;+\; e_{5,2}^n(\lambda) \;+\; e_{5,3}^n(\lambda), \tag{4.43}$$

where

$$\begin{aligned}
e_{5,1}^n(\lambda) &= \frac{1}{i\,\pi\,t_{n-1}\,t_{n-2}\lambda}\; e^{i\theta\,t_{n-2}}\;\widehat{K}_{5,1}(s;\lambda)\Big|_{\theta=\frac{\pi}{k}} \\
&\quad - \frac{1}{i\pi\,t_n^2\lambda}\; e^{i\theta\,t_n}\;\widehat{K}_{5,1}(i\theta;\lambda)\Big|_{\theta=\frac{\pi}{k}}, \\
e_{5,2}^n(\lambda) &= -\Big\{\frac{1}{i\,\pi\,t_{n-1}\,t_{n-2}\lambda}\; e^{i\theta\,t_{n-2}}\,\widehat{K}_{5,1}(s;\lambda)\Big|_{\theta=\varepsilon} \\
&\quad - \frac{1}{i\pi\,t_n^2\lambda}\; e^{i\theta\,t_n}\;\widehat{K}_{5,1}(i\theta;\lambda)\Big|_{\theta=\varepsilon}\Big\}, \\
e_{5,3}^n(\lambda) &= -\Big\{\frac{1}{\pi\,t_{n-1}t_{n-2}\lambda}\int_\varepsilon^{\frac{\pi}{k}}\; e^{i\theta\,t_{n-3}}\;\widehat{K}_{5,2}(s;\lambda)\;d\theta \\
&\quad - \frac{1}{\pi\,t_n^2\lambda}\int_\varepsilon^{\frac{\pi}{k}}\; e^{i\theta\,t_n}\,\widehat{K}_{5,2}(i\theta;\lambda)\,d\theta\Big\}, \\
\widehat{K}_{5,1}(s;\lambda) &= \frac{D_s(s;\lambda)}{[D(s;\lambda)]^2}, \\
\widehat{K}_{5,2}(s;\lambda) &= \frac{\widehat{\beta}''(s)}{D^2(s;\lambda)} \;-\; \frac{2\,D_s^2(s;\lambda)}{D^3(s;\lambda)},
\end{aligned}$$

and

$$\begin{aligned}
-u_6(t_n,\lambda) &= e_{6,1}^n(\lambda) \;+\; e_{6,2}^n(\lambda) \\
&= \frac{1}{i\pi\,t_n^2\lambda}\; e^{i\theta\,t_n}\;\widehat{K}_{5,1}(i\theta;\lambda)\Big|_{\theta=\frac{\pi}{k}} \\
&\quad + \frac{1}{\pi\,t_n^2\lambda}\int_{\frac{\pi}{k}}^{\infty}\; e^{i\theta\,t_n}\,\widehat{K}_{5,2}(i\theta;\lambda)\,d\theta.
\end{aligned} \tag{4.44}$$

A similar argument can be applied to show that for $t_{n-2} > 0$,

$$|e_{5,2}^n(\lambda)| \;\le\; C\,k\,\lambda^{-1}t_{n-2}^{-2}. \tag{4.45}$$

We recall as well that (see [3, (5.7)]),

$$\frac{|D_s(\frac{2}{k};\lambda)|}{|D(\frac{2}{k};\lambda)|^2} \;\le\; \frac{k^2\lambda}{4} \;+\; \frac{\frac{k}{2}\widehat{\beta}(\frac{2}{k})k\lambda}{2\phi(\frac{2}{k},0)} \;\le\; \frac{k^2\lambda}{2}, \tag{4.46}$$

which yields

$$|e_{5,1}^n(\lambda) \;+\; e_{6,1}^n(\lambda)| \;\le\; C\,k^2t_{n-2}^{-2}. \tag{4.47}$$

In section 5 we will show that

$$|e^n_{5,3}(\lambda)| \le C\,k\,\lambda^3 t_{n-2}^{-2} \quad (t_{n-2}>0), \tag{4.48}$$

$$|e^n_{6,2}(\lambda)| \le C\,k^2\,\lambda^2 t_n^{-2} \quad (t_n>0). \tag{4.49}$$

Referring to (4.3), (4.4), (4.24), (4.37), (4.40), and (4.43)–(4.45), along with the estimates (4.47)–(4.49), we obtain the proof of Theorem 3.

5. Proof of (4.48) and (4.49). In order to obtain a similar estimate on $e^n_{5,3}(\lambda)$ we put that $e^n_{5,3}(\lambda) = e^n_{5,3,1}(\lambda) + e^n_{5,3,2}(\lambda)$, where

$$e^n_{5,3,1}(\lambda) = -\left\{ \frac{1}{\pi\, t_{n-1}t_{n-2}\lambda} \int_{\varepsilon}^{\frac{\varepsilon}{k}} e^{i\theta\, t_{n-3}}\, \widehat{K}_{5,2}(s;\lambda)\, d\theta \right. \tag{5.1}$$

$$\left. - \frac{1}{\pi\, t_n^2\lambda} \int_{\varepsilon}^{\frac{\varepsilon}{k}} e^{i\theta\, t_n} \widehat{K}_{5,2}(i\theta;\lambda)\, d\theta \right\},$$

$$e^n_{5,3,2}(\lambda) = -\left\{ \frac{1}{\pi\, t_{n-1}t_{n-2}\lambda} \int_{\frac{\varepsilon}{k}}^{\frac{\pi}{k}} e^{i\theta\, t_{n-3}}\, \widehat{K}_{5,2}(s;\lambda)\, d\theta \right. \tag{5.2}$$

$$\left. - \frac{1}{\pi\, t_n^2\lambda} \int_{\frac{\varepsilon}{k}}^{\frac{\pi}{k}} e^{i\theta\, t_n} \widehat{K}_{5,2}(i\theta;\lambda)\, d\theta \right\}.$$

We argue as in (4.41) to obtain

$$\begin{aligned} |e^n_{5,3,1}(\lambda)| \le{}& \left| \frac{1}{\pi\, t_{n-1}t_{n-2}\lambda} \int_{\varepsilon}^{\frac{\varepsilon}{k}} e^{i\theta\, t_{n-3}} \left(\widehat{K}_{5,2}(s;\lambda) - \widehat{K}_{5,2}(i\theta;\lambda) \right) d\theta \right| \\ &+ \left| \frac{1}{\pi\, t_{n-1}t_{n-2}\lambda} \int_{\varepsilon}^{\frac{\varepsilon}{k}} e^{i\theta\, t_{n-3}} \left(1 - e^{3\,i\theta\, k}\right) \widehat{K}_{5,2}(i\theta;\lambda)\, d\theta \right| \\ &+ \left| \frac{1}{\pi\, t_{n-1}\lambda} \left(\frac{1}{t_{n-2}} - \frac{1}{t_n} \right) \int_{\varepsilon}^{\frac{\varepsilon}{k}} e^{i\theta\, t_n} \widehat{K}_{5,2}(i\theta;\lambda)\, d\theta \right| \\ &+ \left| \frac{1}{\pi\, t_n\lambda} \left(\frac{1}{t_{n-1}} - \frac{1}{t_n} \right) \int_{\varepsilon}^{\frac{\varepsilon}{k}} e^{i\theta\, t_n} \widehat{K}_{5,2}(i\theta;\lambda)\, d\theta \right| \\ ={}& e^n_{5,3,1,1}(\lambda) + e^n_{5,3,1,2}(\lambda) + e^n_{5,3,1,3}(\lambda) + e^n_{5,3,1,4}(\lambda). \end{aligned} \tag{5.3}$$

By the same computations we see that

$$
\begin{aligned}
(5.4)\quad e^n_{5,3,1,2}(\lambda) \;&\le\; C\,k\,t_{n-2}^{-2}\,\lambda^{-1}\int_{\varepsilon}^{\frac{\varepsilon}{k}}\theta\,|\widehat{K}_{5,2}(i\theta;\lambda)|\,d\theta\\
&\le\; C\,k\,t_{n-2}^{-2}\,\lambda^{-1}\int_{\varepsilon}^{\frac{\varepsilon}{k}}\tau\left(\left|\frac{\widehat{\beta}''(i\tau)}{D^2(i\tau;\lambda)}\right| \;+\; \left|\frac{D_S^2(i\tau;\lambda)}{D^3(i\tau;\lambda)}\right|\right)d\tau\\
&\le\; C\,k\,t_{n-2}^{-2}\,\lambda^{-1}\left\{\int_{\varepsilon}^{\frac{\omega}{2}}\frac{\lambda^{-2}+\theta^2(\tau)}{\tau^2\theta^3(\tau)}\,d\tau+\int_{\frac{\omega}{2}}^{\omega-\varepsilon}\left(\frac{\lambda}{|\tau-\omega|^2}+\frac{\tau\lambda}{|\tau-\omega|^3}\right)d\tau\right.\\
&+\left.\int_{\omega-\varepsilon}^{\omega+\varepsilon}\left(\frac{\lambda^{-1}}{\phi^2(\tau)}+\frac{\tau\lambda^{-2}}{\phi^3(\tau)}\right)d\tau+\int_{\omega+\varepsilon}^{\infty}\left(\frac{\lambda}{|\tau-\omega|^2}+\frac{\tau\lambda}{|\tau-\omega|^3}\right)d\tau\right\}\\
&\le\; C\,k\,t_{n-2}^{-2}\,\lambda.
\end{aligned}
$$

Similarly, we have that

$$
(5.5)\qquad e^n_{5,3,1,3}(\lambda) \;+\; e^n_{5,3,1,4}(\lambda) \;\le\; C\,k\,t_{n-2}^{-3}\,\lambda;
$$

for example, for $t_{n-2}>0$,

$$
\begin{aligned}
(5.6)\quad e^n_{5,3,1,3}(\lambda) \;&\le\; C\,k\,t_{n-2}^{-3}\,\lambda^{-1}\int_{\varepsilon}^{\frac{\varepsilon}{k}}|\widehat{K}_{5,2}(i\theta;\lambda)|\,d\theta\\
&\le\; C\,k\,t_{n-2}^{-3}\,\lambda^{-1}\int_{\varepsilon}^{\frac{\varepsilon}{k}}\left(\left|\frac{\widehat{\beta}''(i\tau)}{D^2(i\tau;\lambda)}\right| \;+\; \left|\frac{D_S^2(i\tau;\lambda)}{D^3(i\tau;\lambda)}\right|\right)d\tau\\
&\le\; C\,k\,t_{n-2}^{-3}\,\lambda^{-1}\left\{\int_{\varepsilon}^{\frac{\omega}{2}}\frac{\lambda^{-2}+\theta^2(\tau)}{\tau^3\theta^3(\tau)}\,d\tau+\int_{\frac{\omega}{2}}^{\omega-\varepsilon}\left(\frac{\lambda}{\tau\,|\tau-\omega|^2}+\frac{\lambda}{|\tau-\omega|^3}\right)d\tau\right.\\
&+\left.\int_{\omega-\varepsilon}^{\omega+\varepsilon}\left(\frac{\lambda^{-1}}{\tau\,\phi^2(\tau)}+\frac{\lambda^{-2}}{\phi^3(\tau)}\right)d\tau+\int_{\omega+\varepsilon}^{\infty}\left(\frac{\lambda}{\tau\,|\tau-\omega|^2}+\frac{\lambda}{|\tau-\omega|^3}\right)d\tau\right\}\\
&\le\; C\,k\,t_{n-2}^{-3}\,\lambda.
\end{aligned}
$$

Also, by the mean-value theorem we now have

$$
(5.7)\qquad e^n_{5,3,1,1}(\lambda)\;\le\; C\,k\,t_{n-2}^{-2}\lambda^{-1}\int_{\varepsilon}^{\frac{\varepsilon}{k}}\theta^2|\widehat{K}'_{5,2}(s_1;\lambda)|\,d\theta,
$$

where

$$
\widehat{K}'_{5,2}(s;\lambda) \;=\; \frac{\widehat{\beta}'''(s)}{D^2(s;\lambda)} \;-\; \frac{6\,\widehat{\beta}''(s)\,D_S(s;\lambda)}{D^3(s;\lambda)} \;+\; \frac{6\,D_S^3(s;\lambda)}{D^4(s;\lambda)}.
$$

From (3.22), (4.17), (4.20), and (3.39)–(3.41) we find that

(5.8)
$$\int_{\varepsilon}^{\frac{\varepsilon}{k}} \theta^2 |\widehat{K}'_{5,2}(s_1;\lambda)|\, d\theta \;\le\; C \int_{c_0\varepsilon}^{\frac{\varepsilon}{k}} \tau^2 |\widehat{K}'_{5,2}(s_1;\lambda)|\, d\tau$$
$$\le C \left(\int_{c_0\varepsilon}^{\frac{c_0\omega}{2}} \;+\; \int_{\frac{c_0\omega}{2}}^{2\omega} \;+\; \int_{2\omega}^{\infty} \tau^2 \;\left(\left| \frac{\widehat{\beta}'''(s_1)}{D^2(s_1;\lambda)} \right| \right.\right.$$
$$+ \;\left| \frac{\widehat{\beta}''(s_1)\, D_s(s_1;\lambda)}{D^3(s_1;\lambda)} \right| \;+\; \left.\left| \frac{D_s^3(s_1;\lambda)}{D^4(s_1;\lambda)} \right| \right) \, d\tau \Bigg)$$
$$\le\; C \Bigg\{ \int_{c_0\varepsilon}^{\frac{c_0\omega}{2}} \frac{\lambda^{-3} + \lambda^{-1}\theta^2(\tau) + \theta^3(\tau)}{\tau^2\theta^4(\tau)}\, d\tau + \int_{\frac{c_0\omega}{2}}^{2\omega} \left(\frac{\lambda^{-1}}{\phi^2(\tau)} + \frac{\tau\lambda^{-2}}{\phi^3(\tau)} + \frac{\tau^2\lambda^{-3}}{\phi^4(\tau)} \right) \, d\tau$$
$$+ \;\int_{2\omega}^{\infty} \left(\frac{\lambda}{|\tau-\omega|^2} + \frac{\tau\lambda}{|\tau-\omega|^3} + \frac{\tau^2\lambda}{|\tau-\omega|^4} \right) \, d\tau \Bigg\} \;\le\; C\,\lambda^4.$$

Combining (5.7) and (5.8) yields

$$e^n_{5,3,1,1}(\lambda) \;\le\; C\,k\,\lambda^3 t_{n-2}^{-2}. \tag{5.9}$$

Together the estimates (5.1) and (5.3)–(5.9) yield

$$|e^n_{5,3,1}(\lambda)| \le C\,k\,\lambda^3 t_{n-2}^{-2} \qquad (t_{n-2} > 0). \tag{5.10}$$

Next we turn to $e^n_{5,3,2}(\lambda)$. We know from [3, (5.12) and (5.13)] that

$$\lambda^{-1} \int_{\frac{\varepsilon}{k}}^{\frac{\pi-\varepsilon\,k}{k}} \left(\frac{|\widehat{\beta}''(s)|}{|D(s;\lambda)|^2} + \frac{|D_s^2(s;\lambda)|}{|D(s;\lambda)|^3} \right) \, d\theta \;\le\; C\,k^2 \tag{5.11}$$

and

$$\lambda^{-1} \int_{\frac{\pi-\varepsilon\,k}{k}}^{\frac{\pi}{k}} \left(\frac{|\widehat{\beta}''(s)|}{|D(s;\lambda)|^2} + \frac{|D_s^2(s;\lambda)|}{|D(s;\lambda)|^3} \right) \, d\theta \;\le\; C\,k^2. \tag{5.12}$$

We use (5.11) and (5.12) to obtain

$$\left| \frac{1}{\pi\, t_{n-1} t_{n-2} \lambda} \int_{\frac{\varepsilon}{k}}^{\frac{\pi}{k}} e^{i\theta\, t_{n-3}}\, \widehat{K}_{5,2}(s;\lambda)\, d\theta \right| \;\le\; C\,k^2\, t_{n-2}^{-2}\,. \tag{5.13}$$

It remains to establish an estimate on

$$\left| \frac{1}{\pi\, t_n^2 \lambda} \int_{\frac{\varepsilon}{k}}^{\frac{\pi}{k}} e^{i\theta\, t_n}\, \widehat{K}_{5,2}(i\theta;\lambda)\, d\theta \right| \tag{5.14}$$
$$\le\; \frac{1}{\pi\, t_n^2 \lambda} \int_{\frac{\varepsilon}{k}}^{\infty} \left(\left| \frac{\widehat{\beta}''(i\tau)}{D^2(i\tau;\lambda)} \right| \;+\; \left| \frac{D_s^2(i\tau;\lambda)}{D^3(i\tau;\lambda)} \right| \right) \, d\tau \;=\; \overline{e}^n_{5,3,2}(\lambda).$$

If $2\omega \le \frac{\varepsilon}{k}$, then $\frac{1}{\varepsilon - k\omega} \le \frac{2}{\varepsilon}$, and we have

$$\overline{e}^n_{5,3,2}(\lambda) \;\le\; C\, t_n^{-2} \left(\frac{k}{\varepsilon} \int_{\frac{\varepsilon}{k}}^{\infty} \frac{d\tau}{(\tau-\omega)^2} \;+\; \int_{\frac{\varepsilon}{k}}^{\infty} \frac{d\tau}{(\tau-\omega)^3} \right) \le\; C\, \frac{k^2 t_n^{-2}}{\varepsilon^2}. \tag{5.15}$$

If $\omega + \varepsilon \le \frac{\varepsilon}{k} < 2\omega$, then $\lambda^{-1} < \frac{2k}{\varepsilon}$, and arguing similarly to (5.15),

$$\overline{e}^n_{5,3,2}(\lambda) \;\le\; C\, t_n^{-2} \left(\int_{\frac{\varepsilon}{k}}^{\infty} \frac{d\tau}{\tau(\tau-\omega)^2} \;+\; \int_{\frac{\varepsilon}{k}}^{\infty} \frac{d\tau}{(\tau-\omega)^3} \right) \le\; C\, \frac{k^2 t_n^{-2} \lambda^2}{\varepsilon^4}. \tag{5.16}$$

If $\omega \le \frac{\varepsilon}{k} < \omega + \varepsilon < 2\omega$, then $\lambda^{-1} \le \frac{2k}{\varepsilon}$, and

$$\begin{aligned} \overline{e}^n_{5,3,2}(\lambda) \;\le\; C\, t_n^{-2} \lambda^{-1} \int_{\frac{\varepsilon}{k}}^{\omega+\varepsilon} \left(\frac{\lambda^{-1}}{\tau \phi^2(\tau)} \;+\; \frac{\lambda^{-2}}{\phi^3(\tau)} \right) d\tau \\ + C\, t_n^{-2} \int_{\omega+\varepsilon}^{\infty} \left(\frac{1}{\tau(\tau-\omega)^2} \;+\; \frac{1}{(\tau-\omega)^3} \right) d\tau \;\le\; C\, k^2 \lambda^2\, \frac{t_n^{-2}}{\varepsilon^4}. \end{aligned} \tag{5.17}$$

If $\omega - \varepsilon \le \frac{\varepsilon}{k} < \omega$, then $\lambda^{-1} < \frac{k}{\varepsilon}$, and

$$\overline{e}^n_{5,3,2}(\lambda) \;\le\; C\, t_n^{-2} \lambda^{-1} \left(\int_{\frac{\varepsilon}{k}}^{\omega+\varepsilon} . \;+\; \int_{\omega+\varepsilon}^{\infty} . \right) \;\le\; C\, k^2 \lambda^2\, \frac{t_n^{-2}}{\varepsilon^4}. \tag{5.18}$$

If $\frac{\omega}{2} \le \frac{\varepsilon}{k} < \omega - \varepsilon$, or $\frac{\varepsilon}{k} < \frac{\omega}{2}$, similar arguments apply to obtain

$$\overline{e}^n_{5,3,2}(\lambda) \;\le\; C t_n^{-2}\, k^2 \;\lambda^2. \tag{5.19}$$

Using (5.15)–(5.19) it follows that

$$\left| \frac{1}{\pi\, t_n^2 \lambda} \int_{\frac{\varepsilon}{k}}^{\frac{\pi}{k}} e^{i\theta\, t_n}\, \widehat{K}_{5,2}(i\theta;\lambda)\, d\theta \right| \;\le\; C\, k^2 \lambda^2\, t_n^{-2}, \tag{5.20}$$

and together with (5.13) show that

$$|e^n_{5,3,2}(\lambda)| \;\le\; C\, k^2 \lambda^2\, t_{n-2}^{-2} \quad (t_{n-2} > 0). \tag{5.21}$$

Together, (5.10), (5.21) give us

$$|e^n_{5,\,3}(\lambda)| \;\le\; C\, k\, \lambda^3\, t_{n-2}^{-2}, \tag{5.22}$$

which is (4.48).

Repeating the proof of (5.20) we now have

$$|e^n_{6,2}(\lambda)| \;\le\; C\, k^2 \lambda^2 t_n^{-2} \quad (t_n > 0), \tag{5.23}$$

which is (4.49).

Remark 5.1. Under the additional condition (1.30) on the kernel $\beta(t)$, we have that

$$e^n_{5,3,1,1}(\lambda) \;\le\; C\, k\, t_{n-2}^{-2}\; \lambda^2, \tag{5.9$'$}$$

and along with the proof of Theorem 3, this implies that

$$k \sum_{n=m}^{\infty} \|U^n \;-\; u(t_n)\| \;\le\; C\, k\, \|u_0\|_4 \quad (t_{m-2} > 0). \tag{5.24}$$

Combining Remarks 3.1 and 5.1 completes the proof of Corollary 1.

REFERENCES

[1] R. W. Carr and K. B. Hannsgen, *A nonhomogeneous integrodifferential equation in Hilbert space*, SIAM J. Math. Anal., 10 (1979), pp. 961–984.

[2] R. W. Carr and K. B. Hannsgen, *Resolvent formulas for a Volterra equation in Hilbert space*, SIAM J. Math. Anal., 13 (1982), pp. 459–483.

[3] X. Da, *Uniform l^1 behavior for time discretization of a Volterra equation with completely monotonic kernel* I: *Stability*, IMA J. Numer. Anal., 22 (2002), pp. 133–151.

[4] P. L. Duren, *The Theory of H^p Spaces*, Academic Press, New York, 1970.

[5] K. B. Hannsgen, *Indirect Abelian theorems and a linear Volterra equation*, Trans. Amer. Math. Soc., 142 (1969), pp. 539–555.

[6] K. B. Hannsgen, *A Volterra equation with completely monotonic convolution kernel*, J. Math. Anal. Appl., 31 (1970), pp. 459–471.

[7] K. B. Hannsgen, *The resolvent kernel of an integrodifferential equation in Hilbert space*, SIAM J. Math. Anal., 7 (1976), pp. 481–490.

[8] K. B. Hannsgen and R. L. Wheeler, *Complete monotonicity and resolvents of Volterra integrodifferential equations*, SIAM J. Math. Anal., 13 (1982), pp. 962–969.

[9] K. B. Hannsgen and R. L. Wheeler, *Uniform L^1 behavior in classes of integrodifferential equations with completely monotonic kernels*, SIAM J. Math. Anal., 15 (1984), pp. 579–594.

[10] J. J. Levin and D. F. Shea, *On the asymptotic behavior of the bounded solutions of some integral equations*, I, J. Math. Anal. Appl., 37 (1972), pp. 42–82.

[11] J. J. Levin and D. F. Shea, *On the asymptotic behavior of the bounded solutions of some integral equations*, II, J. Math. Anal. Appl., 37 (1972), pp. 288–326.

[12] J. J. Levin and D. F. Shea, *On the asymptotic behavior of the bounded solutions of some integral equations*, III, J. Math. Anal. Appl., 37 (1972), pp. 537–575.

[13] C. Lubich, *Convolution quadrature and discretized operational calculus* I, Numer. Math., 52 (1988), pp. 129–145.

[14] C. Lubich, *On convolution quadrature and Hille-Phillips operational calculus*, Appl. Numer. Math., 9 (1992), pp. 187–199.

[15] C. Lubich, *Convolution quadrature revisited*, BIT Numerical Mathematics, 44 (2004), pp. 503–514.

[16] C. Lubich, I. H. Sloan, and V. Thomée, *Nonsmooth data error estimates for approximations of an evolution equation with a positive-type memory term*, Math. Comp., 65 (1996), pp. 1–17.

[17] W. McLean and V. Thomée, *Asymptotic behaviour of numerical solutions of an evolution equation with memory*, Asymptot. Anal., 14 (1997), pp. 257–276.

[18] R. Noren, *Uniform L^1 behavior for the solution of a Volterra equation with a parameter*, SIAM J. Math. Anal., 19 (1988), pp. 270–286.

[19] D. F. Shea and S. Wainger, *Variants of the Wiener-Lévy theorem, with applications to stability problems for some Volterra integral equations*, Amer. J. Math., 97 (1975), pp. 312–343.

[20] D. V. Widder, *The Laplace Transform*, Princeton University Press, Princeton, NJ, 1946.

SIAM J. NUMER. ANAL.
Vol. 46, No. 1, pp. 260–279
© 2008 Society for Industrial and Applied Mathematics

ANALYSIS OF THE MULTISCALE FINITE ELEMENT METHOD FOR NONLINEAR AND RANDOM HOMOGENIZATION PROBLEMS*

ZHANGXIN CHEN† AND TATYANA Y. SAVCHUK‡

Abstract. In this paper we study the convergence of the multiscale finite element method for nonlinear and random homogenization problems. Error estimates similar to those for linear homogenization problems are obtained here.

Key words. multiscale problem, multiscale finite element method, finite element, nonlinear problems, random problems, oversampling technique, convergence, stability, error estimate

AMS subject classifications. 35K60, 35K65, 76S05, 76T05

DOI. 10.1137/060654207

1. Introduction. Hou and Wu [16] introduced the multiscale finite element method for numerical solution of multiscale problems that are described by partial differential equations with highly oscillatory coefficients. The main idea of this method is to incorporate the microscale information of a multiscale differential problem into finite element basis functions. It is through these modified bases and finite element formulations that the effect of microscales on macroscales can be correctly captured.

The convergence analysis of the method was given in [17] for a two-scale linear homogenization problem with periodic coefficients. It was proven that the multiscale finite element solution converges to the homogenized solution as $h,\ \epsilon \to 0$, where h is the mesh size and ϵ is the small scale in the solution. The analysis also indicated that a resonance error exists between the grid scale and the scales of the homogenization problem. This is a common feature in some numerical upscaling techniques. This error represents a mismatch between the local construction of the multiscale basis functions and the global nature of the continuous problem. An oversampling technique was analyzed in [13] to reduce the resonance error. The idea of this technique is to construct the local basis functions over a domain with size larger than h to reduce the boundary layer effect present in the first order corrector of the local solution.

A convergence analysis of the multiscale finite element method and its oversampled version for nonlinear problems was recently given by Efendiev [11] and Efendiev, Hou, and Ginting [12], and error estimates were obtained for monotone operators (see the next section). These monotone operators do not cover the nonlinear problem studied here, which is a natural extension of the usual linear homogenization problems with highly oscillatory coefficients [5, 16, 17]. The primary goal of this paper is to derive error estimates for the nonlinear problem under consideration. We show that error estimates similar to those for linear homogenization problems hold

*Received by the editors March 14, 2006; accepted for publication (in revised form) August 20, 2007; published electronically January 11, 2008.
http://www.siam.org/journals/sinum/46-1/65420.html

†Research Center for Science, Xi'an Jiaotong University, Xi'an 710049, China, Department of Chemical and Petroleum Engineering, Schulich School of Engineering, University of Calgary, 2500 University Drive, N.W. Calgary, AB T2N 1N4, Canada (zhachen@ucalgary.ca), Center for Scientific Computation, Southern Methodist University, Dallas, TX 75275-0156, and Center for Advanced Reservoir Modeling and Simulation, College of Engineering, Peking University, Beijing 100871, China.
‡Department of Mathematics, Southern Methodist University, Dallas, TX 75275-0156 (tsavchuk@mail.smu.edu).

for this problem as well. We also indicate how to extend this error analysis to random homogenization problems.

The present analysis is based on an equivalent formulation for the multiscale finite element method recently introduced by Chen [5], which utilizes standard basis functions of finite element spaces but modifies the bilinear (quadratic) form in the finite element formulation of the underlying multiscale problems. This new formulation captures the macroscale structure of the solution of a differential multiscale problem through the modification of this bilinear form. It is a general approach that can handle a large variety of differential problems, periodic or nonperiodic, linear or nonlinear, and stationary or dynamic, as shown here. A similar idea using the operator approach was employed by Arbogast [1].

The paper is organized as follows. In the next section we present a continuous two-scale nonlinear problem, the multiscale finite element method, and existing error estimates. Existence and uniqueness of a multiscale finite element solution is shown in the third section. An improved error analysis is given in the fourth section. An oversampling technique for the nonlinear problem is presented in the fifth section, and a simple reconstruction trick to retrieve the microscopic information is described in the sixth section. Finally, an extension to a random homogenization problem is given in the seventh section. As a general remark, the generic constant $C > 0$ (with or without a subscript) is assumed to be independent of the mesh size h and the microscale ϵ throughout this paper.

2. Existing error estimates for nonlinear problems. Let Ω be a bounded domain in $\mathbb{R}^d$, $1 \le d \le 3$, with Lipschitz boundary Γ. For a subdomain $D \subset \Omega$, each integer $m \ge 0$, and each real number $1 \le p \le \infty$, $W^{m,p}(D)$ indicates the usual Sobolev space of real functions that have all their weak derivatives of order up to m in the Lebesgue space $L^p(D)$. The norm and seminorm of $W^{m,p}(D)$ are denoted by $\|\cdot\|_{m,p,D}$ and $|\cdot|_{m,p,D}$, respectively. When $p = 2$, $W^{m,p}(D)$ is written as $H^m(D)$ with the norm $\|\cdot\|_{m,D}$ and the seminorm $|\cdot|_{m,D}$. We also use the space

$$W_0^{1,p}(D) = \{v \in W^{1,p}(D) : v|_{\partial D} = 0\}, \qquad p > 1.$$

Again, when $p = 2$, it is written as $H_0^1(D)$.

We consider the nonlinear problem

$$\begin{aligned} -\nabla \cdot \left(a_\epsilon \nabla u^\epsilon\right) &= f \qquad \text{in } \Omega, \\ u^\epsilon &= 0 \qquad \text{on } \Gamma, \end{aligned} \tag{2.1}$$

where $a_\epsilon = a(x, x/\epsilon, u^\epsilon)$ depends on the solution u^ϵ. In problem (2.1), the multiscale feature is reflected in the oscillatory nature of the coefficient a_ϵ for $\epsilon \ll 1$, which represents the microscale.

We assume that the coefficient $a(x, y, z)$ is equicontinuous in z uniformly with respect to x and y and periodic in y with period $I = [0, 1]^d$. Furthermore, it satisfies

$$a_*|\zeta|^2 \le \sum_{i,j=1}^{d} a_{ij}(x, y, q)\zeta_i\zeta_j \le a^*|\zeta|^2 \qquad \forall x \in \Omega,\ y,\ \zeta \in \mathbb{R}^d,\ q \in \mathbb{R}, \tag{2.2}$$

for some positive constants a_* and a^*. Under such assumptions, the solution u^ϵ of problem (2.1) converges weakly in $\mathcal{U} = W_0^{1,p}(\Omega)$ $(p > 1)$ to the solution of the homogenized equation [3]

$$\begin{aligned} -\nabla \cdot \left(\mathcal{A}(x, U_0)\nabla U_0\right) &= f \qquad \text{in } \Omega, \\ U_0 &= 0 \qquad \text{on } \Gamma, \end{aligned} \tag{2.3}$$

where the homogenized matrix $\mathcal{A} = (\mathcal{A}_{ij})$ is

$$\mathcal{A}_{ij}(x,q) = \frac{1}{|I|}\int_I \left(a_{ij}(x,y,q) + \sum_{k=1}^{d} \left(a_{ik}\frac{\partial \chi^j}{\partial y_k} \right)(x,y,q) \right) dy \qquad \forall q \in \mathbb{R},$$

and χ^j satisfies, with a periodic boundary condition in y,

$$\begin{aligned} -\nabla_y \cdot \left(a(x,y,q)\nabla_y \chi^j \right) &= \sum_{i=1}^{d} \frac{\partial}{\partial y_i} a_{ij}(x,y,q), \qquad && y \in I, \\ \int_I \chi^j(x,y,q)\, dy &= 0, && q \in \mathbb{R}. \end{aligned} \tag{2.4}$$

The variational form of (2.3) reads as follows: Find $U_0 \in \mathcal{U}$ such that

$$A(U_0, v) \equiv (\mathcal{A}(x, U_0)\nabla U_0, \nabla v) = (f, v) \qquad \forall v \in \mathcal{U}. \tag{2.5}$$

For $h > 0$, let T_h be a regular, quasi-uniform macroscale triangulation of Ω [4, 7] into simplices, where the mesh size h resolves the variations of Ω, f, and the slow variable of a_ϵ. Associated with T_h, let $\mathcal{U}_h \subset \mathcal{U}$ be the finite element space of piecewise linear functions over simplices so that for any $v \in \mathcal{U} \cap H^2(\Omega)$, there exists a $v_h \in \mathcal{U}_h$ satisfying the approximation property

$$\|v - v_h\|_{0,\Omega} + h\|v - v_h\|_{1,\Omega} \le Ch^2 |v|_{l,\Omega}. \tag{2.6}$$

For any $v \in \mathcal{U}_h$, we define its local solution $R_T(v) \in H^1(T)$, $T \in T_h$, by

$$\begin{aligned} (a(x, x/\epsilon, v)\nabla R_T(v), \nabla w)_T &= 0 \qquad && \forall w \in H_0^1(T), \\ R_T(v) &= v && \text{on } \partial T. \end{aligned} \tag{2.7}$$

The global operator R is then given by

$$R(v)|_T = R_T(v) \qquad \forall v \in \mathcal{U}_h,\ T \in T_h.$$

It is easy to see that $R(v) \in \mathcal{U}$, $v \in \mathcal{U}_h$. Note that the local problem (2.7) is linear. In the case without ambiguity in the context, the subscript T in R will be omitted.

Define

$$A_h(v,w) = \big(a(x, x/\epsilon, v)\nabla R(v), \nabla R(w)\big), \qquad v,\ w \in \mathcal{U}_h.$$

The multiscale finite element method (MsFEM) for (2.1) is as follows: Find $u_h \in \mathcal{U}_h$ such that

$$A_h(u_h, v) = (f, v) \qquad \forall v \in \mathcal{U}_h. \tag{2.8}$$

Note that the major difference between (2.8) and the standard Galerkin finite element method lies in the modification of the bilinear form, which needs the solution of local problems (2.7). It is through these local problems and the finite element formulation that the effect of microscales on macroscales can be correctly captured. Since these local problems are independent of each other, they can be solved in parallel.

For a linear counterpart of problem (2.1) where the coefficient $a_\epsilon = a(x, x/\epsilon)$ does not depend on the solution u^ϵ, the following error estimate was obtained for method (2.8) [5, 17]:

$$
\begin{aligned}
|u^\epsilon - R(u_h)|_{1,\Omega} &\le C\left\{(h+\epsilon)\|f\|_{0,\Omega} + \sqrt{\frac{\epsilon}{h}}\,|U_0|_{1,\infty,\Omega}\right\},\\
\|u^\epsilon - R(u_h)\|_{0,\Omega} &\le C\left\{(h^2+\epsilon)\|f\|_{0,\Omega} + \sqrt{\frac{\epsilon}{h}}\,|U_0|_{1,\infty,\Omega}\right\},
\end{aligned}
\tag{2.9}
$$

provided that $U_0 \in H^2(\Omega) \cap W^{1,\infty}(\Omega)$.

Recently, Efendiev, Hou, and Ginting [12] studied convergence of the MsFEM for a nonlinear problem analogous to (2.1):

$$
-\nabla \cdot a(x, u^\epsilon, \nabla u^\epsilon) = f. \tag{2.10}
$$

Under the assumptions that $a(x, u^\epsilon, \nabla u^\epsilon) = a(x/\epsilon, \nabla u^\epsilon)$ and

$$
\begin{aligned}
&|a(x,\xi)| \le C|\xi|^{p-1},\\
&(a(x,\xi_1) - a(x,\xi_2), \xi_1 - \xi_2) \ge C|\xi_1 - \xi_2|^p,\\
&(a(x,\xi),\xi) \ge C|\xi|^p,\\
&|a(x,\xi_1) - a(x,\xi_2)| \le CH(\xi_1,\xi_2,p-1-s)|\xi_1 - \xi_2|^s,
\end{aligned}
$$

where $s > 0$ and $p > 1$, the following error estimate was derived (see [12, Theorem 3.2]):

$$
\|u - u_h\|^p_{1,p,\Omega} \le C\left(\left(\frac{\epsilon}{h}\right)^{\frac{s}{(p-1)(p-s)}} + \left(\frac{\epsilon}{h}\right)^{\frac{p}{p-1}} + h^{\frac{p}{p-1}}\right). \tag{2.11}
$$

Obviously, the assumptions made on a_ϵ exclude the nonlinear problem (2.1). The convergence result in [12] was obtained for the general nonlinear case (2.10); however, there is no explicit convergence for the case considered in this paper because of very weak assumptions made in the way the coefficients depend on u^ϵ. The aim of this paper is to obtain error estimates for the MsFEM (2.8). In particular, we will derive error estimates similar to (2.9) under much weaker assumptions on the coefficient a_ϵ in (2.1). The error analysis is inspired by E [9] and E, Ming, and Zhang [10] for the heterogeneous multiscale method.

3. Existence and uniqueness of a solution. It is known that the error analysis of the MsFEM for the case $h < \epsilon$ is different from that for the case $h > \epsilon$. In the former case, the MsFEM has error estimates similar to those for the traditional finite element method [5, 17]. It is the latter case that is of interest and is being investigated in this paper. The argument below requires that $\epsilon > 0$ be sufficiently small.

Introduce the linearized differential operator at U_0

$$
L(U_0)v = -\nabla \cdot (\mathcal{A}(x, U_0)\nabla v + v\ \mathcal{A}_p(x, U_0)\nabla U_0)\,, \qquad v \in H^1(\Omega),
$$

and the corresponding bilinear form

$$
\hat{A}(U_0; v, w) = (\mathcal{A}(x, U_0)\nabla v, \nabla w) + (v\mathcal{A}_p(x, U_0)\nabla U_0, \nabla w) \qquad \forall v,\ w \in H^1(\Omega),
$$

where $\mathcal{A}_p(x,u) = \nabla_u \mathcal{A}(x,u)$. We assume that this linearized operator is an isomorphism from $H^1_0(\Omega)$ to $H^{-1}(\Omega)$, and so U_0 is an isolated solution of (2.5) and there is

$h_0 > 0$ such that for $0 < h < h_0$ [19],

$$\sup_{w \in \mathcal{U}_h} \frac{\hat{A}(U_0; v, w)}{\|w\|_{1,\Omega}} \geq C_0 \|v\|_{1,\Omega} \qquad \forall v \in \mathcal{U}_h, \tag{3.1}$$

where $C_0 > 0$ is independent of h.

For any $v, v_h, w \in \mathcal{U}$, we define

$$\mathcal{R}(v, v_h, w) = A(v_h, w) - A(v, w) - \hat{A}(v; v_h - v, w).$$

If $v, v_h \in \mathcal{U}$ satisfy $\|v\|_{1,\infty,\Omega} + \|v_h\|_{1,\infty,\Omega} \leq M$, then it follows from [20] that

$$|\mathcal{R}(v, v_h, w)| \leq C(M) \left(\|e_h\|_{0,2p}^2 + \|e_h \nabla e_h\|_{0,p} \right) \|\nabla w\|_{0,q}, \qquad e_h = v - v_h, \ \frac{1}{p} + \frac{1}{q} = 1. \tag{3.2}$$

It follows from the definition of $\mathcal{R}$ and (2.5) that $u_h \in \mathcal{U}_h$ is the solution of (2.8) if and only if

$$\hat{A}(U_0; U_0 - u_h, w) = \mathcal{R}(U_0, u_h, w) + A_h(u_h, w) - A(u_h, w) \qquad \forall w \in \mathcal{U}_h. \tag{3.3}$$

Define

$$E(v, w) = A_h(v, w) - A(v, w) \qquad \forall v, \ w \in \mathcal{U}_h,$$

and

$$\overline{E} = \max_{v \in \mathcal{U}_h \cap W^{1,\infty}(\Omega), w \in \mathcal{U}_h} \frac{|E(v, w)|}{\|v\|_{1,\Omega} \|w\|_{1,\Omega}}.$$

The following equivalence will be used [5]:

$$C_1' |v|_{1,T} \leq |R_T(v)|_{1,T} \leq C_2' |v|_{1,T} \qquad \forall T \in T_h, \ v \in \mathcal{U}_h. \tag{3.4}$$

To prove existence and uniqueness of a solution to (2.8), we introduce the projection of U_0 into $\mathcal{U}_h$ through the linearized bilinear form $\hat{A}$:

$$\hat{A}(U_0; P_h U_0, v) = \hat{A}(U_0; U_0, v) \qquad \forall v \in \mathcal{U}_h. \tag{3.5}$$

It follows from (3.1) that $P_h U_0$ exists and is unique for $0 < h < h_0$, and it satisfies [8]

$$\|U_0 - P_h U_0\|_{1,\infty,\Omega} \leq Ch, \quad \|U_0 - P_h U_0\|_{1,\Omega} \leq Ch \tag{3.6}$$

if $U_0 \in W^{2,\infty}(\Omega)$. When $U_0 \in W^{2,p}(\Omega)$ $(p > d)$, it holds that

$$\|U_0 - P_h U_0\|_{1,\infty,\Omega} \leq Ch^{1-d/p}. \tag{3.7}$$

Finally, for a given $x \in \Omega$, we define the discrete Green function $G_h^x \in \mathcal{U}_h$ by

$$\hat{A}(U_0; v, G_h^x) = \partial v(x) \qquad \forall v \in \mathcal{U}_h, \tag{3.8}$$

where ∂v indicates any of the partial derivatives $\partial v / \partial x_i$ $(i = 1, 2, \ldots, d)$. This function satisfies

$$\|G_h^x\|_{1,1,\Omega} \leq C |\ln h|. \tag{3.9}$$

THEOREM 3.1. *Assume that L is an isomorphism from $H_0^1(\Omega)$ to $H^{-1}(\Omega)$ and $U_0 \in \mathcal{U} \cap W^{2,p}(\Omega)$ with $p > d$. In addition, assume that $\overline{E}$ is bounded and there are constants C_1 and h_1 such that for $0 < h \le h_1$,*

$$\overline{E}^{1/2} |\ln h| \le C_1. \tag{3.10}$$

Then problem (2.8) *has a solution u_h satisfying*

$$\begin{aligned} &\|u_h - P_h U_0\|_{1,\infty,\Omega} \le \overline{E}^{1/2} + h^{1-d/p}, \\ &\|u_h - U_0\|_{1,\infty,\Omega} \le C\left(\overline{E}^{1/2} + h^{1-d/p}\right). \end{aligned} \tag{3.11}$$

Furthermore, if, for all $v_1, v_2 \in \mathcal{U}_h \cap W^{1,\infty}(\Omega)$ and $w \in \mathcal{U}_h$ with $\|v_1\|_{1,\infty,\Omega} + \|v_2\|_{1,\infty,\Omega} \le M$, there is a constant $\zeta_0(M)$, with $0 < \zeta_0 < 1$, such that

$$|E(v_1, w) - E(v_2, w)| \le \zeta_0(M) \|v_1 - v_2\|_{1,\Omega} \|w\|_{1,\Omega}, \tag{3.12}$$

then this solution u_h is locally unique.

Proof. We define the nonlinear mapping $\mathcal{L} : \mathcal{U}_h \to \mathcal{U}_h$ by

$$\hat{A}(U_0; \mathcal{L}(v), w) = \hat{A}(U_0; U_0, w) - \mathcal{R}(U_0, v, w) + A(v, w) - A_h(v, w) \qquad \forall w \in \mathcal{U}_h.$$

This mapping is continuous using (3.1) and (3.2). We also define the set

$$B = \left\{ v \in \mathcal{U}_h : \|v - P_h U_0\|_{1,\infty,\Omega} \le \overline{E}^{1/2} + h^{1-d/p} \right\}.$$

Note that, by (3.5),

$$\hat{A}(U_0; \mathcal{L}(v) - P_h U_0, w) = -\mathcal{R}(U_0, v, w) + A(v, w) - A_h(v, w) \qquad \forall w \in \mathcal{U}_h.$$

Choosing $w = G_h^x$ in this equation and applying (3.2), (3.7), (3.9), and (3.10), we see that, with $v \in B$ and $\|v\|_{1,\infty,\Omega} \le M$,

$$\begin{aligned} \|\mathcal{L}(v) - P_h U_0\|_{1,\infty,\Omega} &\le C(M) \left(\|U_0 - v\|_{1,\infty,\Omega}^2 + \overline{E} \right) |\ln h| \\ &\le C(M) \left(\|U_0 - P_h U_0\|_{1,\infty,\Omega}^2 + \|P_h U_0 - v\|_{1,\infty,\Omega}^2 + \overline{E} \right) |\ln h| \\ &\le C(M) \left(h^{2-2d/p} + \overline{E} \right) |\ln h|. \end{aligned}$$

Because $v \in B$ and $\overline{E}$ is bounded (e.g., $\overline{E} \le \mathcal{C}_1$), we see that

$$\begin{aligned} \|v\|_{1,\infty,\Omega} &\le \|v - P_h U_0\|_{1,\infty,\Omega} + \|P_h U_0\|_{1,\infty,\Omega} \\ &\le \overline{E}^{1/2} + C(U_0) \le \mathcal{C}_1^{1/2} + C(U_0) \equiv \mathcal{C}_0. \end{aligned} \tag{3.13}$$

Combining these two inequalities, we have

$$\|\mathcal{L}(v) - P_h U_0\|_{1,\infty,\Omega} \le C(\mathcal{C}_0) \left(h^{2-2d/p} + \overline{E} \right) |\ln h|.$$

Defining $C_1 = 1/C(\mathcal{C}_0)$, it follows from (3.10) that

$$\|\mathcal{L}(v) - P_h U_0\|_{1,\infty,\Omega} \le \overline{E}^{1/2} + C(\mathcal{C}_0) h^{2-2d/p} |\ln h|.$$

Thus there is a constant h_2 such that for $0 < h \le h_2$, we obtain

$$\|\mathcal{L}(v) - P_h U_0\|_{1,\infty,\Omega} \le \overline{E}^{1/2} + h^{1-d/p}.$$

Set $h_1 = \min(h_0, h_2)$. Then, for $0 < h \le h_1$, we see that $\mathcal{L}(B) \subset B$. The Brouwer fixed point theorem means that there is a $u_h \in B$ such that $\mathcal{L}(u_h) = u_h$.

To prove the uniqueness, let u_h^1 and u_h^2 be two solutions of (2.8). Then it follows from (3.3) that

$$\begin{aligned}\hat{A}(U_0; u_h^2 - u_h^1, w) &= \hat{A}(U_0; U_0 - u_h^1, w) - \hat{A}(U_0; U_0 - u_h^2, w)\\ &= \mathcal{R}(U_0, u_h^1, w) - \mathcal{R}(U_0, u_h^2, w) + E(u_h^1, w) - E(u_h^2, w).\end{aligned} \tag{3.14}$$

Because both u_h^1 and u_h^2 are in the set B, it follows from (3.13) that $\|u_h^1\|_{1,\infty,\Omega} + \|u_h^2\|_{1,\infty,\Omega} \le 2\mathcal{C}_0$, which, together with (3.1), (3.12), (3.14), and Poincaré's inequality, implies that

$$\|u_h^1 - u_h^2\|_{1,\Omega} \le \zeta_0(2\mathcal{C}_0)\|u_h^1 - u_h^2\|_{1,\Omega}.$$

Since $\zeta_0 < 1$ via assumption, then $u_h^1 = u_h^2$. Therefore, the solution u_h is locally unique. ∎

4. Improved error estimates for nonlinear problems. The main result in this section is stated in the next theorem. The multiscale finite element solution u_h used below refers to the one that satisfies the conditions in Theorem 3.1. To avoid unnecessary techniques, we perform an error analysis when system (2.7) is replaced by

$$\begin{aligned}&(a(x_T, x/\epsilon, v(x_T))\nabla R_T(v), \nabla w)_T = 0 \qquad && \forall w \in H_0^1(T),\\ &R_T(v) = v && \text{on } \partial T,\end{aligned} \tag{4.1}$$

where x_T is any point in $T \in T_h$ (e.g., the barycenter of T).

THEOREM 4.1. *Let U_0 and u_h be the solutions of* (2.5) *and* (2.8), *respectively, and $U_0 \in W^{2,\infty}(\Omega)$. Then there is $h_0 > 0$ such that for $0 < h < h_0$,*

$$\|U_0 - u_h\|_{1,\Omega} \le C\left(h + \sqrt{\frac{\epsilon}{h}}\right), \quad \|U_0 - u_h\|_{1,\infty,\Omega} \le C\left(h + \sqrt{\frac{\epsilon}{h}}\right)|\ln h|, \tag{4.2}$$

provided that ϵ/h is sufficiently small.

This theorem can be shown by combining the next two propositions.

PROPOSITION 4.2. *In addition to the assumptions of Theorem* 3.1, *let $U_0 \in W^{2,\infty}(\Omega)$. Then there are constants C_2 (see the proof below) and $h_0 > 0$ such that if $C_1 < C_2$ and $0 < h < h_0$, then*

$$\|U_0 - u_h\|_{1,\Omega} \le C\left(h + \overline{E}\right), \quad \|U_0 - u_h\|_{1,\infty,\Omega} \le C\left(h + \overline{E}\right)|\ln h|. \tag{4.3}$$

PROPOSITION 4.3. *Under the assumptions of Theorem* 4.1, *if $\sqrt{\epsilon/h}|\ln h|$ is sufficiently small, then conditions* (3.10) *and* (3.12) *hold. Moreover,*

$$\overline{E} \le C\sqrt{\frac{\epsilon}{h}}. \tag{4.4}$$

Proof of Proposition 4.2. Taking $w = P_h U_0 - u_h$ in (3.3) and using (3.1), (3.2), and (3.6), we see that

$$\|P_h U_0 - u_h\|_{1,\Omega} \le C\left(\|U_0 - u_h\|_{1,4,\Omega}^2 + \overline{E} + h\right). \tag{4.5}$$

Applying an interpolation inequality, we have

$$\|U_0 - u_h\|_{1,4,\Omega}^2 \le \|U_0 - u_h\|_{1,\Omega}\|U_0 - u_h\|_{1,\infty,\Omega}, \tag{4.6}$$

which, together with (3.6) and (3.13), yields

$$\|P_hU_0 - u_h\|_{1,\Omega} \le C_3\left(\overline{E}^{1/2} + h^{1-d/p}\right)\|P_hU_0 - u_h\|_{1,\Omega} + C\left(h + \overline{E}\right).$$

Choosing $w = G_h^x$ in (3.3) and using (3.2), (3.6), and (3.9), we have

$$\begin{aligned}\|P_hU_0 - u_h\|_{1,\infty,\Omega} \le{}& C_4(\overline{E} + h)|\ln h|\|P_hU_0 - u_h\|_{1,\infty,\Omega}\\ &+ C\left(\|U_0 - u_h\|_{1,\infty,\Omega}^2 + \overline{E} + h\right)|\ln h|.\end{aligned}$$

From (3.6), (3.7), and (3.10) it follows that

$$\begin{aligned}\|P_hU_0 - u_h\|_{1,\infty,\Omega} \le{}& C_4(\overline{E}^{1/2} + h^{1-d/p})|\ln h|\|P_hU_0 - u_h\|_{1,\infty,\Omega}\\ &+ C\left(\overline{E} + h\right)|\ln h|.\end{aligned}$$

Now, we set

$$C_2 = \min\left(\frac{|\ln h|}{2C_3}, \frac{1}{2C_4}\right)$$

and choose h_0 such that $\overline{E}^{1/2}|\ln h| \le C_2$. With these two choices, we obtain

$$\begin{aligned}&\|P_hU_0 - u_h\|_{1,\Omega} \le C\left(\overline{E} + h\right),\\ &\|P_hU_0 - u_h\|_{1,\infty,\Omega} \le C\left(\overline{E} + h\right)|\ln h|,\end{aligned}$$

which, together with (3.6), gives the desired result. □

Proposition 4.3 can be proven from the next three lemmas.

LEMMA 4.4. *For $v, w \in \mathcal{U}_h$, we have*

$$|R(v) - R(w)|_{1,T} \le C\left(\|v - w\|_{0,\infty,T}\left[|v|_{1,T} + |w|_{1,T}\right] + |v - w|_{1,T}\right), \quad T \in T_h. \tag{4.7}$$

Proof. It follows from (2.2) and (4.1) that

$$\begin{aligned}C|R(v) - R(w)|_{1,T}^2 \le{}& \left(a\left(x_T, \frac{x}{\epsilon}, v(x_T)\right)\nabla[R(v) - R(w)], \nabla[R(v) - R(w)]\right)_T\\ \equiv{}& J_1 + J_2,\end{aligned} \tag{4.8}$$

where

$$\begin{aligned}J_1 &= \left(a\left(x_T, \frac{x}{\epsilon}, v(x_T)\right)\nabla R(v) - a\left(x_T, \frac{x}{\epsilon}, w(x_T)\right)\nabla R(w), \nabla(v - w)\right)_T,\\ J_2 &= -\left(\left[a\left(x_T, \frac{x}{\epsilon}, v(x_T)\right) - a\left(x_T, \frac{x}{\epsilon}, w(x_T)\right)\right]\nabla R(w), \nabla[R(v) - R(w)]\right)_T.\end{aligned}$$

Note that

$$\begin{aligned}J_1 = {}& \left(\left[a\left(x_T, \frac{x}{\epsilon}, v(x_T)\right) - a\left(x_T, \frac{x}{\epsilon}, w(x_T)\right)\right]\nabla R(v), \nabla(v - w)\right)_T\\ &+ \left(a\left(x_T, \frac{x}{\epsilon}, w(x_T)\right)\nabla[R(v) - R(w)], \nabla(v - w)\right)_T.\end{aligned}$$

As a result, using (2.2), we bound J_1 and J_2 as follows:

$$|J_1| \le C\left(\|v-w\|_{0,\infty,T}|R(v)|_{1,T} + |R(v)-R(w)|_{1,T}\right)|v-w|_{1,T},$$
$$|J_2| \le C\|v-w\|_{0,\infty,T}|R(w)|_{1,T}|R(v)-R(w)|_{1,T}.$$

Substituting these two bounds for J_1 and J_2 into (4.8) and using (3.4) implies the desired result (4.7). □

For $v, w \in \mathcal{U}_h$, we define

$$Q(v) = v(x) + \epsilon \sum_{k=1}^{d} \chi^k\left(x_T, \frac{x}{\epsilon}, v(x_T)\right)\frac{\partial v}{\partial x_k}(x),$$
$$Q(w) = w(x) + \epsilon \sum_{k=1}^{d} \chi^k\left(x_T, \frac{x}{\epsilon}, w(x_T)\right)\frac{\partial w}{\partial x_k}(x).$$

Set $\theta_\epsilon(v) = R(v) - Q(v)$ and $\theta_\epsilon(w) = R(w) - Q(w)$. Note that

$$\begin{aligned} -\nabla\cdot\left(a\left(x_T, \frac{x}{\epsilon}, v(x_T)\right)\nabla\theta_\epsilon(v)\right) &= 0 \qquad \text{in } T,\\ \theta_\epsilon(v) &= -\epsilon\sum_{k=1}^{d}\chi^k\left(x_T, \frac{x}{\epsilon}, v(x_T)\right)\frac{\partial v}{\partial x_k} \qquad \text{on } \partial T, \end{aligned} \tag{4.9}$$

and an analogous definition can be given for $\theta_\epsilon(w)$. These two quantities satisfy [5, 17]

$$|\theta_\epsilon(v)|_{1,T} \le C\sqrt{\frac{\epsilon}{h}}\,|v|_{1,T}, \quad |\theta_\epsilon(w)|_{1,T} \le C\sqrt{\frac{\epsilon}{h}}\,|w|_{1,T}, \tag{4.10}$$

which hold because ∇v and ∇w are piecewise constant.

LEMMA 4.5. *For $v, w \in \mathcal{U}_h$, we have*

$$|\theta_\epsilon(v) - \theta_\epsilon(w)|_{1,T} \le C\sqrt{\frac{\epsilon}{h}}\left(\|v-w\|_{0,\infty,T}\left[|v|_{1,T}+|w|_{1,T}\right] + |v-w|_{1,T}\right), \quad T \in T_h. \tag{4.11}$$

Proof. Let $\xi_\epsilon \in C_0^\infty(T)$, $0 \le \xi_\epsilon \le 1$, be a cut-off function in T such that $\xi_\epsilon = 1$ outside a ϵ-neighborhood of the boundary ∂T and $|\nabla\xi_\epsilon| \le C\epsilon^{-1}$ with C independent of ϵ and T. Define, for $v \in \mathcal{U}_h$,

$$\varphi_\epsilon(v) = \theta_\epsilon(v) + \epsilon\sum_{k=1}^{d}\chi^k\left(x_T, \frac{x}{\epsilon}, v(x_T)\right)\frac{\partial v}{\partial x_k}(1-\xi_\epsilon),$$

and a similar meaning can be given for $\varphi_\epsilon(w)$, $w \in \mathcal{U}_h$. Calculations show that

$$\begin{aligned} &C|\varphi_\epsilon(v) - \varphi_\epsilon(w)|_{1,T}\\ &\quad\le \max_{x\in T}|a\big(x_T, x/\epsilon, v(x_T)\big) - a\big(x_T, x/\epsilon, w(x_T)\big)|\left(|\varphi_\epsilon(v)-\theta_\epsilon(v)|_{1,T} + |\varphi_\epsilon(w)|_{1,T}\right)\\ &\qquad + \max_{x\in T}|a\big(x_T, x/\epsilon, w(x_T)\big)|\;|\varphi_\epsilon(v)-\theta_\epsilon(v)-\varphi_\epsilon(w)+\theta_\epsilon(w)|_{1,T} \end{aligned}$$

and

$$|\varphi_\epsilon(v)-\theta_\epsilon(v)|_{1,T} \le C\sqrt{\frac{\epsilon}{h}}|v|_{1,T}, \quad |\varphi_\epsilon(w)-\theta_\epsilon(w)|_{1,T} \le C\sqrt{\frac{\epsilon}{h}}|w|_{1,T}.$$

Also, using (4.10), we see that

$$|\varphi_\epsilon(w)|_{1,T} \le |\varphi_\epsilon(w) - \theta_\epsilon(w)|_{1,T} + |\theta_\epsilon(w)|_{1,T} \le C\sqrt{\frac{\epsilon}{h}}|w|_{1,T}.$$

Observe that

$$\begin{aligned}
&\varphi_\epsilon(v) - \theta_\epsilon(v) - \varphi_\epsilon(w) + \theta_\epsilon(w) \\
&\quad = \epsilon \sum_{k=1}^{d} \left(\chi^k\left(x_T, \frac{x}{\epsilon}, v(x_T)\right) - \chi^k\left(x_T, \frac{x}{\epsilon}, w(x_T)\right)\right) \frac{\partial v}{\partial x_k}(1-\xi_\epsilon) \\
&\qquad + \epsilon \sum_{k=1}^{d} \chi^k\left(x_T, \frac{x}{\epsilon}, w(x_T)\right) \frac{\partial (v-w)}{\partial x_k}(1-\xi_\epsilon).
\end{aligned}$$

Consequently, by the continuity of $\{\chi^k\}_{k=1}^d$, we see that

$$|\varphi_\epsilon(v) - \theta_\epsilon(v) - \varphi_\epsilon(w) + \theta_\epsilon(w)|_{1,T} \le C\sqrt{\frac{\epsilon}{h}}\left(\|v-w\|_{0,\infty,T}|v|_{1,T} + |v-w|_{1,T}\right).$$

Combining these inequalities gives the desired result (4.11). ∎

The next lemma indicates that $E(\cdot,\cdot)$ has certain continuity with respect to its first argument.

LEMMA 4.6. *For $v_1, v_2, w \in \mathcal{U}_h$ satisfying $\|v_1\|_{1,\infty,\Omega} + \|v_2\|_{1,\infty,\Omega} \le M$, we have*

$$|E(v_1,w) - E(v_2,w)| \le C(M)\sqrt{\frac{\epsilon}{h}}\|v_1 - v_2\|_{1,\Omega}|w|_{1,\Omega}, \quad T \in T_h. \tag{4.12}$$

Proof. Define $l = [h/\epsilon]$, and let $I_{l\epsilon}$ be the cube of size $l\epsilon$ at x_T. By the definition of $R(v_1)$ and the relation that $R(v_1) = \theta_\epsilon(v_1) + Q(v_1)$, we see that

$$\begin{aligned}
\big(a\big(x_T, x/\epsilon, v_1(x_T)\big)\nabla R(v_1), \nabla R(w)\big)_T &= \big(a\big(x_T, x/\epsilon, v_1(x_T)\big)\nabla R(v_1), \nabla w\big)_T \\
&= \big(a\big(x_T, x/\epsilon, v_1(x_T)\big)\nabla [\theta_\epsilon(v_1) + Q(v_1)], \nabla w\big)_T.
\end{aligned}$$

Also, note that

$$\frac{1}{|I_{l\epsilon}|}\left(a\left(x_T, \frac{x}{\epsilon}, v_1(x_T)\right)\nabla Q(v_1), \nabla w\right)_{I_{l\epsilon}} = \nabla w \cdot \mathcal{A}\big(x_T, v_1(x_T)\big)\nabla v_1.$$

Then $E(v_1,w) - E(v_2,w)$ can be split as follows:

$$E(v_1,w) - E(v_2,w) = \sum_{T\in T_h} \left(E(v_1,w) - E(v_2,w)\right)|_T = \sum_{T\in T_h}(J_3 + J_4), \tag{4.13}$$

where

$$\begin{aligned}
J_3 &= \left(a\left(x_T, \frac{x}{\epsilon}, v_1(x_T)\right)\nabla\theta_\epsilon(v_1) - a\left(x_T, \frac{x}{\epsilon}, v_2(x_T)\right)\nabla\theta_\epsilon(v_2), \nabla w\right)_T, \\
J_4 &= \left(a\left(x_T, \frac{x}{\epsilon}, v_1(x_T)\right)\nabla Q(v_1), \nabla w\right)_T \\
&\quad - \frac{|T|}{|I_{l\epsilon}|}\left(a\left(x_T, \frac{x}{\epsilon}, v_1(x_T)\right)\nabla Q(v_1), \nabla w\right)_{I_{l\epsilon}} \\
&\quad - \left(a\left(x_T, \frac{x}{\epsilon}, v_2(x_T)\right)\nabla Q(v_2), \nabla w\right)_T \\
&\quad - \frac{|T|}{|I_{l\epsilon}|}\left(a\left(x_T, \frac{x}{\epsilon}, v_2(x_T)\right)\nabla Q(v_2), \nabla w\right)_{I_{l\epsilon}}.
\end{aligned}$$

It follows from Lemma 4.5 that

$$|J_3| \le C\sqrt{\frac{\epsilon}{h}}\big(\|v_1 - v_2\|_{0,\infty,T}\,[|v_1|_{1,T} + |v_2|_{1,T}] + |v_1 - v_2|_{1,T}\big)|w|_{1,T}. \tag{4.14}$$

The term J_4 can be rewritten as follows:

$$\begin{aligned} J_4 = &-\left(\frac{|T|}{|l\epsilon|^d} - 1\right)\left(a\left(x_T, \frac{x}{\epsilon}, v_1(x_T)\right)\nabla Q(v_1) - a\left(x_T, \frac{x}{\epsilon}, v_2(x_T)\right)\nabla Q(v_2), \nabla w\right)_T \\ &+ \frac{|T|}{|l\epsilon|^d}\left(a\left(x_T, \frac{x}{\epsilon}, v_1(x_T)\right)\nabla Q(v_1) - a\left(x_T, \frac{x}{\epsilon}, v_2(x_T)\right)\nabla Q(v_2), \nabla w\right)_{T\setminus I_{l\epsilon}}. \end{aligned}$$

Then it can be bounded by

$$|J_4| \le C\frac{\epsilon}{h}\left(\|v_1 - v_2\|_{0,\infty,T}\,[|v_1|_{1,T} + |v_2|_{1,T}] + |v_1 - v_2|_{1,T}\right)|w|_{1,T}. \tag{4.15}$$

Applying an inverse inequality, we see that

$$\begin{aligned} \|v_1 - v_2\|_{0,\infty,T}|v_1|_{1,T} &\le Ch_T^{-d/2}\|v_1 - v_2\|_{0,T}h_T^{d/2}\|v_1\|_{1,\infty,T} \\ &= C\|v_1 - v_2\|_{0,T}\|v_1\|_{1,\infty,T}, \end{aligned} \tag{4.16}$$

since $\nabla v_1|_T$ is a constant. A similar result holds for $\|v_1 - v_2\|_{0,\infty,T}|v_2|_{1,T}$. Substituting (4.14)–(4.16) into (4.13) implies the desired result. ∎

Proof of Proposition 4.3. Taking $v_2 = 0$ in (4.12) generates (4.4). Also, if $\sqrt{\epsilon/h}|\ln h|$ is sufficiently small, $\overline{E}^{1/2}|\ln h|$ can be made smaller than any given threshold, and so (3.10) holds. Finally, choose $\zeta_0(M) = C(M)\sqrt{\epsilon/h}$. Then if $\sqrt{\epsilon/h}|\ln h|$ is sufficiently small, $\zeta_0(M) < 1$, and thus (3.12) is verified. ∎

5. An oversampling technique. Note that estimates (4.2) deteriorate when ϵ is of the same order as the mesh size h. This phenomenon reveals a "resonance error" between the grid scale h and the scale ϵ of the continuous problem (2.1). This resonance is due to a mismatch between the local solution of (2.7) and the global solution of (2.1) on the boundary of each $T \in T_h$, which produces a boundary layer. Since this layer is thin, we can sample in a (local) domain with size larger than h and utilize only the interior sampled information. In this manner, the influence of the boundary layer in the larger domain can be greatly reduced. In this section, we extend an oversampling technique for linear problems [13, 16] to the nonlinear problem (2.1) in order to reduce the resonance error in (4.2).

For each $T \in T_h$, we indicate by $S(T)$ a macroelement which contains T and satisfies the following condition: There are positive constants C_3 and C_4, independent of h and ϵ, such that $h_S \le C_3 h_T$ and $\operatorname{dist}(\partial S, \partial T) \ge C_4 h_T$. For each $v \in \mathcal{U}_h(T)$, we extend it to $\mathcal{U}_h(S)$ as follows. Let $\{\phi_i^T\}_1^{d+1}$ and $\{\psi_i^S\}_1^{d+1}$ be the respective bases of $\mathcal{U}_h(T)$ and $\mathcal{U}_h(S)$. Set

$$v|_T = \sum_{i=1}^{d+1} c_i^T \phi_i^T, \quad \phi_i^T = \sum_{j=1}^{d+1} c_{ij}^T \psi_j^S|_T.$$

Then we define $\widehat{v} \in \mathcal{U}_h(S)$ by

$$\widehat{v} = \sum_{i,j=1}^{d+1} c_i^T c_{ij}^T \psi_j^S.$$

Now, for any $v \in \mathcal{U}_h$, we define $R_S(v) \in H^1(S)$, $T \subset S$, $T \in T_h$, by

$$\begin{aligned} &(a(x, x/\epsilon, \widehat{v})\nabla R_S(v), \nabla w)_S = 0 \qquad && \forall w \in H_0^1(S), \\ &R_S(v) = \widehat{v} && \text{on } \partial S. \end{aligned} \tag{5.1}$$

The global operator R is defined by

$$R(v)|_T = R_S(v)|_T \qquad \forall v \in \mathcal{U}_h,\ T \in T_h.$$

The oversampled MsFEM for (2.1) is to seek $u_h \in \mathcal{U}_h$ such that

$$\sum_{T \in T_h} \left(a(x, x/\epsilon, u_h)\nabla R_{S(T)}(u_h), \nabla R_{S(T)}(v)\right)_T = (f, v) \qquad \forall v \in \mathcal{U}_h. \tag{5.2}$$

The existence and uniqueness of a solution to (5.2) can be shown in Theorem 3.1. Furthermore, combining the error analysis in the previous section and that for the oversampled MsFEM for linear problems given in [5, 13], the following improved error estimates can be shown as in Theorem 4.1.

THEOREM 5.1. *Let U_0 and u_h be the solutions of* (2.5) *and* (5.2)*, respectively, and $U_0 \in W^{2,\infty}(\Omega)$. Then there is $h_0 > 0$ such that for $0 < h < h_0$,*

$$\|U_0 - u_h\|_{1,\Omega} \le C\left(h + \frac{\epsilon}{h}\right), \quad \|U_0 - u_h\|_{1,\infty,\Omega} \le C\left(h + \frac{\epsilon}{h}\right)|\ln h|, \tag{5.3}$$

provided that ϵ/h is sufficiently small.

Note that while these estimates improve those in (5.15), resonance persists.

6. Approximation to u^ϵ. Theorems 4.1 and 5.1 show that the multiscale finite element solution u_h of (2.8) is a good approximation of the macroscopic solution U_0. We now consider an approximation to the solution u^ϵ of (2.1).

Define $R(u_h)$ as the solution of (4.1) with v replaced by u_h:

$$\begin{aligned} &(a(x_T, x/\epsilon, u_h(x_T))\nabla R_T(u_h), \nabla w)_T = 0 \qquad && \forall w \in H_0^1(T), \\ &R_T(v) = u_h && \text{on } \partial T. \end{aligned} \tag{6.1}$$

Applying (2.4), $R_T(u_h)$ satisfies

$$\nabla R_T(u_h) = \nabla u_h + \sum_{k=1}^{d} \nabla_y \chi^k \left(x_T, \frac{x}{\epsilon}, u_h(x_T)\right) \frac{\partial u_h}{\partial x_k} + \nabla \theta_\epsilon(u_h), \qquad T \in T_h. \tag{6.2}$$

Also, define the first order approximation of u_1^ϵ by

$$u_1^\epsilon = U_0 + \epsilon \sum_{k=1}^{d} \chi^k \left(x, \frac{x}{\epsilon}, U_0(x)\right) \frac{\partial U_0}{\partial x_k}.$$

Clearly,

$$\nabla u_1^\epsilon = \nabla U_0 + \sum_{k=1}^{d} \left\{ \left(\epsilon \nabla \chi^k + \nabla_y \chi^k\right) \frac{\partial U_0}{\partial x_k} + \epsilon \frac{\partial \chi^k}{\partial U_0} \frac{\partial U_0}{\partial x_k} \nabla U_0 + \epsilon \chi^k \nabla \frac{\partial U_0}{\partial x_k} \right\}. \tag{6.3}$$

Combining (6.2) and (6.3) gives

(6.4)
$$|u_1^\epsilon - R(u_h)|_{1,T} \le C\left(|U_0 - u_h|_{1,T} + \epsilon\left[\|U_0\|_{2,T} + \|U_0\|_{1,4,T}^2\right]\right) + |\theta_\epsilon(u_h)|_{1,T}$$
$$+ \left\|\sum_{k=1}^d \left[\nabla_y \chi^k\left(x_T, \frac{x}{\epsilon}, u_h(x_T)\right) - \nabla_y \chi^k\left(x, \frac{x}{\epsilon}, U_0(x)\right)\right]\frac{\partial U_0}{\partial x_k}\right\|_{0,T}$$
$$\le C\Big\{|U_0 - u_h|_{1,T} + |U_0 - u_h|_{1,\infty,T}|U_0|_{1,T} + |\theta_\epsilon(u_h)|_{1,T}$$
$$+ (\epsilon + h)\left(\|U_0\|_{2,T} + \|U_0\|_{1,4,T}^2 + |U_0|_{1,\infty,T}|U_0|_{1,T}\right)\Big\}.$$

A classical estimate for $u^\epsilon - u_1^\epsilon$ [2, 3, 22] gives

$$\|u^\epsilon - u_1^\epsilon\|_{1,\Omega} \le C\sqrt{\epsilon}. \tag{6.5}$$

Finally, combining (4.2), (4.10), (6.4), and (6.5), we obtain the next theorem.

THEOREM 6.1. *Let u^ϵ be the solution of* (2.1), *$R(u_h)$ be defined by* (6.1), *and $U_0 \in W^{2,\infty}(\Omega)$. Then there is $h_0 > 0$ such that for $0 < h < h_0$,*

$$|u^\epsilon - R(u_h)|_{1,\Omega} \le C\left(h + \sqrt{\epsilon} + \sqrt{\frac{\epsilon}{h}}\right)|\ln h|, \tag{6.6}$$

provided that ϵ/h is sufficiently small.

Note that estimate (6.6) is similar to (2.9) obtained for a linear counterpart of (2.1). The oversampling technique discussed in the previous section can also be applied to (6.1).

7. A random homogenization problem. In the previous sections we have assumed that the coefficient a_ϵ in (2.1) is periodic. In many problems such as in porous media flows [6], this coefficient is often random. In this section we extend the multiscale finite element analysis performed for the nonlinear problem (2.1) to a multiscale problem with a random coefficient.

Let (D, F, P) be a probability space and $a(y,\omega) = \big(a_{ij}(y,\omega)\big)$ be a random field, $y \in \mathbb{R}^d$, $\omega \in D$, whose statistics is invariant under integer shifts. Furthermore, let a satisfy the uniform ellipticity condition (2.2); i.e.,

$$a_*|\zeta|^2 \le \sum_{i,j=1}^d a_{ij}(y,\omega)\zeta_i\zeta_j \le a^*|\zeta|^2 \qquad \forall \omega \in D,\ y,\ \zeta \in \mathbb{R}^d, \tag{7.1}$$

for some positive constants a_* and a^*. Problem (2.1) now takes the form

$$\begin{aligned} -\nabla\cdot\big(a(x/\epsilon,\omega)\nabla u^\epsilon\big) &= f \qquad \text{in } \Omega,\\ u^\epsilon &= 0 \qquad \text{on } \Gamma. \end{aligned} \tag{7.2}$$

7.1. Homogenization results. We collect some homogenization results for problem (7.2), following [21]. As in (2.4), let χ^j satisfy [18]

$$-\nabla_y\cdot\big(a(y,\omega)\nabla_y\chi^j\big) = \sum_{i=1}^d \frac{\partial}{\partial y_i}a_{ij}(y,\omega), \tag{7.3}$$

and $\nabla\chi^j$ is assumed to be stationary under integer shifts. χ^j is generally not stationary.

Define the average operator with respect to the measure P (mathematical expectation)

$$\langle v \rangle = \mathbb{E} \int_{[0,1]^d} v(y)\ dy.$$

Also, define

$$[v]_m \equiv [v; m] = \frac{1}{m^d} \int_{[0,m]^d} v(y)\ dy.$$

The homogenized coefficient $\mathcal{A}$ is given by

$$\mathcal{A} = \langle a(\mathcal{I} + \nabla\chi) \rangle\,, \tag{7.4}$$

where $\mathcal{I}$ is the identity matrix and $\chi = (\chi^1, \chi^2, \ldots, \chi^d)^T$.

For any $\rho > 0$, we consider the auxiliary problem

$$-\nabla_y \cdot (a(y,\omega)\nabla_y u) + \rho u = \sum_{i=1}^{d} \frac{\partial g_i}{\partial y_i}, \tag{7.5}$$

where

$$g_i \in \{v : \langle v^2 \rangle \le G^2\}, \qquad i = 1, 2, \ldots, d,$$

with v a random field whose statistics is stationary under integer shifts.

For each fixed realization of $\{a(y,\cdot)\}$, let η_x be the diffusion process generated by $-\nabla_y \cdot (a(y,\omega)\nabla_y)$ and starting from x at $t = 0$, and let M_x be the expectation with respect to η_x. Set

$$\Gamma(s) = \int_0^s e^{-\rho\tau} \sum_{i=1}^{d} \frac{\partial g_i}{\partial y_i}\big(\eta(\tau)\big)\ d\tau.$$

It is known [14] that the solution of problem (7.5) is

$$u_\rho(x) = M_x \Gamma(\infty).$$

Lemmas 7.1–7.3 and 7.5 below can be found in [21] and Lemma 7.4 in [10]. Note that the homogenization results in [21] may be overestimated because they are based on the Green function estimates that are not required for the computation of effective coefficients. Because of this, the convergence result here may be overestimated as well.

LEMMA 7.1. *For the solution u_ρ of (7.5), there are constants $C > 0$, independent of ρ, such that*

$$\begin{aligned} &\langle |\nabla u_\rho|^2 \rangle + \rho \langle u_\rho^2 \rangle \le C \langle |g|^2 \rangle\,, \\ &\left\langle \left(M_x \Gamma(\infty)\right)^2 \right\rangle^{1/2} \le \frac{CG^2}{\rho}, \\ &\left\langle M_x \left(\Gamma(\infty) - \Gamma(s)\right)^2 \right\rangle \le \frac{CG^2}{\rho} e^{-2s\rho}, \end{aligned} \tag{7.6}$$

where $g = (g_1, g_2, \ldots, g_d)^T$.

Due to the presence of the lower order term ρu, the Green function associated with the differential operator $-\nabla_y \cdot (a(y,\omega)\nabla_y) + \rho$ in the left-hand side of (7.5) decays exponentially with a rate of order $\mathcal{O}(\sqrt{\rho})$. To be specific, define

$$B_\rho = \left\{ x \in \mathbb{R}^d : \|x\|_\infty \le \rho^{-\frac{1}{2}|\ln \rho^{-1}|^{1/2}} \right\},$$

where $\|x\|_\infty = \max\{|x_i|, i = 1, 2, \ldots, d\}$.

LEMMA 7.2. *Let t be the first exit time of B_ρ starting at $x \in B_\rho$, and define $\hat{u}_\rho(x) = M_x \Gamma(t)$. Then, if ρ is small enough,*

$$\mathbb{E} \int_{\|x\|_\infty \le 10} |u_\rho(x) - \hat{u}_\rho(x)|^2 \, dx \le C G^2 e^{-C|\ln \rho^{-1}|^2},$$
$$\mathbb{E} \int_{\|x\|_\infty \le 1} |\nabla u_\rho(x) - \nabla \hat{u}_\rho(x)|^2 \, dx \le C G^2 e^{-C|\ln \rho^{-1}|^2}.$$

LEMMA 7.3. *Let $\{a^1, g^1\}$ and $\{a^2, g^2\}$ be two sets of data satisfying*

$$\{a^1(y), g^1(y)\} = \{a^2(y), g^2(y)\}, \qquad y \notin B,$$

where $B \subset \mathbb{R}^d$ is a domain, and u_ρ^1 and u_ρ^2 be the respective solutions of (7.5) *associated with $\{a^1, g^1\}$ and $\{a^2, g^2\}$. Then*

$$\int_{\mathbb{R}^d} |u_\rho^1(x) - u_\rho^2(x)|^2 \, dx \le \frac{C}{\rho} \int_{\mathbb{R}^d} \left(G^2 + |\nabla u_\rho^1(x)|^2 \right) I_B(x) \, dx,$$

where I_B is the indicator function of B.

For a subdomain $B \subset \mathbb{R}^d$, denote by $\Phi(B)$ the σ-algebra generated by the parameters $\{a(y,\omega) : y \in B\}$. Let ζ_1 and ζ_2 be two random variables that are measurable with respect to $\Phi(B_1)$ and $\Phi(B_2)$, respectively. We will use the exponential decay condition

$$|\mathbb{E}(\zeta_1 \zeta_2) - \mathbb{E}(\zeta_1)\mathbb{E}(\zeta_2)| \le e^{-C \operatorname{dist}(B_1, B_2)} \sqrt{\mathbb{E}\zeta_1^2} \sqrt{\mathbb{E}\zeta_2^2}. \tag{7.7}$$

This type of exponential decay condition is often used for geostatistical models [15].

LEMMA 7.4. *Under condition* (7.7), *we have*

$$\mathbb{E}\,[u_\rho; m]^2 \le C \left(\frac{G^2}{\rho} \left(\frac{|\ln \rho^{-1}|^2}{\rho^{1/2} m} \right) + e^{-C|\ln \rho^{-1}|^2} \right).$$

LEMMA 7.5. *Let $\chi_\rho = \left(\chi_\rho^1, \chi_\rho^2, \ldots, \chi_\rho^d\right)^T$, where χ_ρ^i is the solution of* (7.5) *with*

$$g = (a_{i1}, a_{i2}, \ldots, a_{id})^T.$$

Under condition (7.7), *for any $0 < \lambda < 1/2$, we have*

$$|\mathcal{A} - \langle a(\mathcal{I} + \nabla \chi_\rho) \rangle| \le C \rho^{(d-2-2\lambda)/(4+d)},$$

where $|\cdot|$ represents a matrix norm.

7.2. The MsFEM. Note that the definition of the MsFEM (2.8) does not utilize any periodicity or macroscopic model. Thus, in the random case it can be defined in the same manner as in the periodic case. That is, for any $v \in \mathcal{U}_h$, we define its local solution $R_T(v) \in H^1(T)$, $T \in T_h$, by

$$\begin{aligned} &(a_\epsilon \nabla R_T(v), \nabla w)_T = 0 \qquad && \forall w \in H_0^1(T), \\ &R_T(v) = v && \text{on } \partial T. \end{aligned} \tag{7.8}$$

Define

$$A_h(v,w) = \sum_{T \in T_h} \left(a_\epsilon \nabla R_T(v), \nabla R_T(w)\right)_T, \qquad v,\ w \in \mathcal{U}_h.$$

Then the MsFEM for (7.2) is as follows: Find $u_h \in \mathcal{U}_h$ such that

$$A_h(u_h, v) = (f, v) \qquad \forall v \in \mathcal{U}_h. \tag{7.9}$$

THEOREM 7.6. *Let U_0 and u_h be the respective solutions of* (2.5) *and* (7.9)*, where the homogenized coefficient $\mathcal{A}$ is now given by* (7.4)*, and $U_0 \in W^{2,\infty}(\Omega)$. Then, under condition* (7.7)*, we have*

$$\begin{aligned} &\mathbb{E}\ \|U_0 - u_h\|_{1,\Omega} \le C\left(h + \left(\frac{\epsilon}{h}\right)^\kappa\right), \\ &\mathbb{E}\ \|U_0 - u_h\|_{0,\Omega} \le C\left(h^2 + \left(\frac{\epsilon}{h}\right)^\kappa\right), \\ &\mathbb{E}\ \|U_0 - u_h\|_{1,\infty,\Omega} \le C\left(h + \left(\frac{\epsilon}{h}\right)^\kappa\right) |\ln h|, \end{aligned} \tag{7.10}$$

where

$$\kappa = \begin{cases} \dfrac{6 - 12\lambda}{25 - 8\lambda} & \text{if } d = 3, \\ \dfrac{1}{2} & \text{if } d = 1 \end{cases}$$

for any $0 < \lambda < 1/2$.

Note that the case $d = 2$ remains open. This theorem can be proven by combining the next two propositions. As in the nonlinear case in the third section, define

$$\overline{E} = \max_{v,w \in \mathcal{U}_h} \frac{|A_h(v,w) - A(v,w)|}{|v|_{1,\Omega}|w|_{1,\Omega}}.$$

PROPOSITION 7.7. *Let U_0 and u_h be the respective solutions of* (2.5) *and* (7.9)*, and let $U_0 \in W^{2,\infty}(\Omega)$. Then*

$$\begin{aligned} &\|U_0 - u_h\|_{1,\Omega} \le C\left(h + \overline{E}\right), \\ &\|U_0 - u_h\|_{0,\Omega} \le C\left(h^2 + \overline{E}\right). \end{aligned} \tag{7.11}$$

Furthermore, if there is a constant C_5 such that $\overline{E}|\ln h| < C_5$, then there is a constant h_0 such that for any $0 < h < h_0$,

$$\|U_0 - u_h\|_{1,\infty,\Omega} \le C\left(h + \overline{E}\right)|\ln h|. \tag{7.12}$$

PROPOSITION 7.8. *Under condition* (7.7)*, we have*

$$\mathbb{E}\overline{E} \le C\left(\frac{\epsilon}{h}\right)^{\kappa}, \tag{7.13}$$

where κ *is defined as in Theorem* 7.6.

As in (3.8), for a given $x \in \Omega$, we define the Green function $G^x \in H_0^1(\Omega)$ and its discrete counterpart $G_h^x \in \mathcal{U}_h$ by

$$\begin{aligned} A(G^x, v) &= \partial v(x) \qquad \forall v \in H_0^1(\Omega), \\ A(G_h^x, v) &= \partial v(x) \qquad \forall v \in \mathcal{U}_h. \end{aligned} \tag{7.14}$$

They satisfy

$$\|G^x - G_h^x\|_{1,1,\Omega} \le C, \quad \|G_h^x\|_{1,1,\Omega} \le C|\ln h|. \tag{7.15}$$

Below P_h indicates the standard Lagrange interpolation operator from $H_0^1(\Omega)$ into $\mathcal{U}_h$.

Proof of Proposition 7.7. It follows from the first Strang lemma [7] that

$$\|U_0 - u_h\|_{1,\Omega} \le C \inf_{v\in\mathcal{U}_h}\left(\|U_0 - v\|_{1,\Omega} + \sup_{w\in\mathcal{U}_h}\frac{|A(v,w) - A_h(v,w)|}{\|w\|_{1,\Omega}}\right).$$

Taking $v = P_h U_0$ and using the definition of $\overline{E}$ implies the first inequality in (7.11). The second inequality in (7.11) follows from a standard duality argument [4, 7].

To prove (7.12), using (7.14), we see that

$$\begin{aligned} \partial(U_0 - u_h)(x) &= A(G^x, U_0 - P_h U_0) + A(G^x, P_h U_0 - u_h) \\ &= A(G^x - G_h^x, U_0 - P_h U_0) + A(G_h^x, U_0 - u_h) \\ &= A(G^x - G_h^x, U_0 - P_h U_0) + A_h(u_h, G_h^x) - A(u_h, G_h^x) \\ &= A(G^x - G_h^x, U_0 - P_h U_0) + (A_h(P_h U_0, G_h^x) - A(P_h U_0, G_h^x)) \\ &\quad + (A_h(u_h - P_h U_0, G_h^x) - A(u_h - P_h U_0, G_h^x)), \end{aligned}$$

which, together with (7.15), gives

$$\begin{aligned} \|U_0 - u_h\|_{1,\infty,\Omega} \le C\|U_0 - P_h U_0\|_{1,\infty,\Omega} &+ |A_h(P_h U_0, G_h^x) - A(P_h U_0, G_h^x)| \\ &+ |A_h(u_h - P_h U_0, G_h^x) - A(u_h - P_h U_0, G_h^x)|. \end{aligned} \tag{7.16}$$

Applying (7.15) again and an inverse inequality, we have

$$|A_h(P_h U_0, G_h^x) - A(P_h U_0, G_h^x)| \le C\overline{E}|\ln h|\|U_0\|_{2,\infty,\Omega}. \tag{7.17}$$

A similar argument yields

$$\begin{aligned} &|A_h(u_h - P_h U_0, G_h^x) - A(u_h - P_h U_0, G_h^x)| \\ &\quad \le C\overline{E}|\ln h|\left(\|U_0 - u_h\|_{1,\infty,\Omega} + h\|U_0\|_{2,\infty,\Omega}\right). \end{aligned} \tag{7.18}$$

Substituting (7.17) and (7.18) into (7.16) implies that

$$\|U_0 - u_h\|_{1,\infty,\Omega} \le C\left(h + \overline{E}|\ln h| + \overline{E}|\ln h|\|U_0 - u_h\|_{1,\infty,\Omega}\right). \tag{7.19}$$

If $\overline{E}|\ln h| < C_5 = 1/(2C)$, the desired result (7.12) follows from (7.19). ∎

Proof of Proposition 7.8. Let $m = h/(2\epsilon)$, and let χ_m^j be the solution of (7.3) on $[0,m]^d$, with the boundary condition $\chi_m^j = 0$ on $\partial[0,m]^d$. Set $\chi_m = (\chi_m^1, \chi_m^2, \ldots, \chi_m^d)^T$. u_ρ^m can be similarly defined using (7.5). Then it follows from the definition of $\overline{E}$ that

$$\overline{E} = |\mathcal{A} - [a(\mathcal{I} + \nabla\chi_m)]_m| \le C(F_1 + F_2 + F_3), \tag{7.20}$$

where

$$\begin{aligned} F_1 &= |\mathcal{A} - [a(\mathcal{I} + \nabla u_\rho)]_m|, \\ F_2 &= |[a(\mathcal{I} + \nabla u_\rho) - a(\mathcal{I} + \nabla u_\rho^m)]_m|, \\ F_2 &= |[a\nabla(u_\rho^m - \chi_m)]_m|. \end{aligned}$$

Note that

$$F_1 = |\mathcal{A} - \langle a(\mathcal{I} + \nabla u_\rho)\rangle + [\varphi]_m|,$$

where $\varphi = \langle a(\mathcal{I} + \nabla u_\rho)\rangle - a(\mathcal{I} + \nabla u_\rho)$. It follows [21] that

$$\mathbb{E}|[\varphi]_m| \le \sqrt{\mathbb{E}[\varphi]_m^2} \le C\left(\frac{|\ln\rho^{-1}|^2}{\rho^{1/2}m}\right)^{d/2},$$

which, together with Lemma 7.5, gives

$$\mathbb{E}F_1 \le C\left(G\rho^{(d-2-2\lambda)/(4+d)} + \left(\frac{|\ln\rho^{-1}|^2}{\rho^{1/2}m}\right)^{d/2}\right). \tag{7.21}$$

Let t_m be the first exit time of the domain $[0,2m]^d$. Then $u_\rho^{2m} = M_x\Gamma(t_m)$, and for any $s > 0$,

$$\begin{aligned} |u_\rho - u_\rho^{2m}| &= \left|M_x\big(\Gamma(\infty) - \Gamma(t_m)\big)\right| \\ &\le M_x\left\{|\Gamma(\infty)| + |\Gamma(t_m)| : t_m \le s\right\} \\ &\quad + M_x\left\{e^{-s\rho}M_{\rho(s)}|\Gamma(\infty) - \Gamma(t_m)| : t_m > s\right\} \\ &\le C\left(M_x\left((\Gamma(\infty))^2 + (\Gamma(t_m))^2\right)\right)^{1/2}\left(P_x\{t_m \le s\}^{1/2} + e^{-s\rho}\right). \end{aligned}$$

Because $P_x\{t_m \le s\} \le e^{-Cm^2/s}$, we see that

$$\mathbb{E}\left[|u_\rho - u_\rho^{2m}|^2\right]_{2m} \le \frac{CG^2}{\rho}\left(e^{-Cm^2/s} + e^{-s\rho}\right)^2,$$

whose optimization in s gives

$$\mathbb{E}\left[|u_\rho - u_\rho^{2m}|^2\right]_{2m} \le \frac{CG^4}{\rho^2}e^{-Cm\rho^{1/2}}.$$

Now, it follows from standard interior estimates that

$$\begin{aligned} \mathbb{E}F_2 &\le C\left(\mathbb{E}\left[|\nabla(u_\rho - u_\rho^m)|^2\right]_m\right)^{1/2} \le \frac{C}{m}\left(\mathbb{E}\left[|u_\rho - u_\rho^{2m}|^2\right]_{2m}\right)^{1/2} \\ &\le \frac{CG^2}{m\rho}e^{-Cm\rho^{1/2}}. \end{aligned} \tag{7.22}$$

Also, in an argument similar to that for F_1, we have

$$\mathbb{E}F_3 \le C\left(G^2\rho^{(d-2-2\lambda)/(4+d)} + \left(\frac{|\ln\rho^{-1}|^2}{\rho^{1/2}m}\right)^{d/2}\right). \tag{7.23}$$

Finally, substituting (7.21)–(7.23) into (7.20) generates

$$\mathbb{E}\overline{E} \le C\left(G^2\rho^{(d-2-2\lambda)/(4+d)} + \left(\frac{|\ln\rho^{-1}|^2}{\rho^{1/2}m}\right)^{d/2} + \frac{G^2}{m\rho}e^{-Cm\rho^{1/2}}\right).$$

Optimizing in ρ with respect to the first two terms of this inequality, we see that

$$\rho_o = m^{-(2d)/(d+4\beta)},$$

where $\beta = (d-2-2\lambda)/(d+4)$. Consequently, we obtain

$$\mathbb{E}\overline{E} \le C\left(\frac{|\ln m|^d}{m^\kappa} + \frac{G^2}{m\rho_0}e^{-Cm\rho_0^{1/2}}\right) \le C\frac{|\ln m|^d}{m^\kappa}, \tag{7.24}$$

where

$$\kappa = \frac{d/2}{1+\frac{d(d+4)/4}{d-2-2\lambda}}, \quad 0<\lambda<\frac{1}{2}.$$

Absorbing the factor $|\ln m|^d$ into m^κ in (7.24) yields the desired result (7.13) for $d=3$.

For $d=1$, a direct evaluation gives

$$\mathbb{E}\overline{E}^2 \le \frac{C}{m},$$

where C is independent of m, which yields (7.13) with $\kappa = 1/2$. ∎

Acknowledgment. The authors would like to thank the referees for their comments that led to great improvements in content.

REFERENCES

[1] T. Arbogast, *Analysis of a two-scale, locally conservative subgrid upscaling for elliptic problems*, SIAM J. Numer. Anal., 42 (2004), pp. 576–598.
[2] A. Benssousan, J. L. Lions, and G. Papanicolaou, *Asymptotic Analysis of periodic Structures*, North–Holland, Amsterdam, 1978.
[3] L. Boccardo and T. Murat, *Homogénéisation de problémes equasi-linéaires*, Publ. IRMA, Lille, 3 (1981), pp. 1–17.
[4] Z. Chen, *Finite Element Methods and Their Applications*, Springer-Verlag, Heidelberg, New York, 2005.
[5] Z. Chen, *Multiscale methods for elliptic homogenization problems*, Numer. Methods Partial Differential Equations, 22 (2006), pp. 317–360.
[6] Z. Chen, G. Huan, and Y. Ma, *Computational Methods for Multiphase Flows in Porous Media*, Comput. Sci. Eng. 2, SIAM, Philadelphia, 2006.
[7] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North–Holland, Amsterdam, 1978.
[8] J. Douglas, Jr., and T. Dupont, *A Galerkin method for a nonlinear Dirichlet problem*, Math. Comp., 29 (1975), pp. 689–696.
[9] W. E and B. Engquist, *The heterogeneous multiscale methods*, Commun. Math. Sci., 1 (2003), pp. 87–132.

[10] W. E, P. Ming, and P. Zhang, *Analysis of the heterogeneous multiscale method for elliptic homogenization problems*, J. Amer. Math. Soc., 18 (2005), pp. 121–156.
[11] Y. R. Efendiev, *The Multiscale Finite Element Method (MsFEM) and Its Applications*, Ph.D. thesis, California Institute of Technology, Pasadena, CA, 1999.
[12] Y. R. Efendiev, T. Hou, and V. Ginting, *Multiscale finite element methods for nonlinear problems and their applications*, Commun. Math. Sci., 2 (2004), pp. 553–589.
[13] Y. R. Efendiev, T. Y. Hou, and X.-H. Wu, *Convergence of a nonconforming multiscale finite element method*, SIAM J. Numer. Anal., 37 (2000), pp. 888–910.
[14] M. I. Freidlin, *Functional Integration and Partial Differential Equations*, Princeton University Press, Princeton, NJ, 1985.
[15] M. I. Freidlin and A. D. Wentzell, *Random Perturbations of Dynamical Systems*, 2nd ed., Springer-Verlag, New York, 1998.
[16] T. Hou and X. Wu, *A multiscale finite element method for elliptic problems in composite materials and porous media*, J. Comput. Phys., 134 (1997), pp. 169–189.
[17] T. Hou, X. Wu, and Z. Cai, *Convergence of a multiscale finite element method for elliptic problems with rapidly oscillating coefficients*, Math. Comp., 68 (1999), pp. 913–943.
[18] S. M. Kozlov, *Homogenization of random operators*, Math. USSR-Sb., 37 (1980), pp. 167–180.
[19] A. Schatz, *An observation concerning Ritz-Galerkin methods with infinite bilinear forms*, Math. Comp., 28 (1974), pp. 959–962.
[20] J. Xu, *Two-grid discretization techniques for linear and nonlinear PDEs*, SIAM J. Numer. Anal., 33 (1996), pp. 1759–1777.
[21] V. V. Yurinskii, *Averaging of symmetric diffusion in random media*, Sibirsk. Mat. Zh., 23 (1982), pp. 176–188 (in Russian); Siberian Math. J., 27 (1986), pp. 603–613 (in English).
[22] V. V. Zhikov, S. M. Kozlov, and O. A. Oleinik, *Homogenization of Differential Operators and Integral Functionals*, Springer-Verlag, Heidelberg, 1994.

SIAM J. NUMER. ANAL.
Vol. 46, No. 1, pp. 280–294
© 2008 Society for Industrial and Applied Mathematics

SPECTRUM OF THE JACOBI TAU APPROXIMATION FOR THE SECOND DERIVATIVE OPERATOR*

MARIOS CHARALAMBIDES† AND FABIAN WALEFFE‡

Abstract. It is proved that the eigenvalues of the Jacobi tau method for the second derivative operator with Dirichlet boundary conditions are real, negative, and distinct for a range of the Jacobi parameters. Special emphasis is placed on the symmetric case of the Gegenbauer tau method where the range of parameters included in the theorems can be merged and characteristic polynomials given by successive order approximations interlace. This includes the common Chebyshev and Legendre tau and Galerkin methods.

Key words. Jacobi polynomials, Gegenbauer polynomials, stable polynomials, positive pairs, zeros of polynomials, spectral methods

AMS subject classifications. 65D30, 65L10, 65L15, 65M70, 65N35, 26C10

DOI. 10.1137/060665907

1. Introduction.

Constructing polynomial approximations to solutions of differential equations is the basic ingredient of most numerical methods. Approximations based on orthogonal polynomials have been widely used (e.g., [2], [3], [9]) because their rate of convergence is faster than algebraic for arbitrary boundary conditions when the solution is smooth. The purpose of this paper is to give a rigorous proof that the spectrum of the Jacobi tau approximation is real, negative, and distinct for the second order operator with Dirichlet boundary conditions. The Jacobi tau class of *spectral methods* includes the common Chebyshev and Legendre tau and Galerkin formulations, as demonstrated below. The general method of proof is similar to that used by Gottlieb and Lustman [7], [8] to prove such results for the Chebyshev collocation operator. However, we argue in Appendix B that Gottlieb and Lustman's proof for the collocation operator is not complete.

The spectrum of the Jacobi tau approximation for the first order operator has been considered elsewhere [5]. Here, we consider polynomial approximations to the eigenvalue problem

$$\frac{d^2u}{dx^2} = \lambda u, \qquad -1 < x < 1, \qquad \text{with} \qquad u(\pm 1) = 0. \tag{1.1}$$

The spectrum of Jacobi polynomial approximations to this eigenvalue problem is directly relevant to numerical simulations of the diffusion equation $u_t = u_{xx}$ which is itself a building block for numerical solution of various other problems including the Stokes and Navier–Stokes equations (e.g., [3, section 5.1], [13]).

The second order problem (1.1) is a self-adjoint, negative definite Sturm–Liouville differential eigenproblem, so its eigenvalues λ are real, negative, and distinct. The

*Received by the editors July 24, 2006; accepted for publication (in revised form) April 24, 2007; published electronically January 16, 2008. This work was supported in part by NSF grant DMS-0204636.

http://www.siam.org/journals/sinum/46-1/66590.html

†Department of Business Administration, Frederick Institute of Technology, 7 Yianni Frederickou Street, Pallouriotissa, P.O. Box 24729, 1303 Nicosia, Cyprus (bus.chm@fit.ac.cy).

‡Department of Mathematics, University of Wisconsin, Madison, WI 53706 (waleffe@math.wisc.edu).

eigenmodes separate into even modes $\cos\sqrt{|\lambda|}\,x$ with $\lambda = -(2k-1)^2\pi^2/4$ and odd modes $\sin\sqrt{|\lambda|}\,x$ with $\lambda = -k^2\pi^2$ for $k = 1, 2, 3, \ldots$.

If $u_n(x)$ is a polynomial approximation of degree n to (1.1), then $u_n(x)$ satisfies the differential equation

$$\lambda u_n(x) - D^2 u_n(x) = R_n(x), \tag{1.2}$$

where the *residual* $R_n(x)$ is a polynomial of degree n in x and $D = d/dx$. We can invert this relation to express the polynomial approximation $u_n(x)$ in terms of the residual $R_n(x)$ [8]:

$$u_n(x) = \mu \sum_{k=0}^{[n/2]} \mu^k D^{2k} R_n(x), \tag{1.3}$$

where $\mu = 1/\lambda$ and $[n/2]$ denotes the greatest integer less than or equal to $n/2$. We can assume that $\lambda \neq 0$ because $\lambda = 0$ with $u_n(\pm 1) = 0$ necessarily corresponds to the trivial solution $u_n(x) = 0$, for all x in $[-1, 1]$, as shown below. The inversion (1.3) follows from formal application of the geometric (Neumann) series for $(1 - \mu D^2)^{-1} = \sum_{k=0}^{\infty} \mu^k D^{2k}$ which terminates since $R_n(x)$ is a polynomial.

Spectral methods fit in the general framework of the *method of weighted residuals* [3]. In the *tau* method [3, section 10.4.2], the polynomial approximation $u_n(x)$ is determined from the boundary conditions $u_n(\pm 1) = 0$ and the requirement that $R_n(x)$ is orthogonal to all polynomials $p_{n-2}(x)$ of degree $n-2$ or less with respect to a weight function $W(x) \geq 0$ in the interval $(-1, 1)$:

$$\int_{-1}^{1} R_n(x) p_{n-2}(x) W(x) dx = 0. \tag{1.4}$$

For the Jacobi weight function $W_{\alpha,\beta}(x) = (1-x)^\alpha (1+x)^\beta$, the residual (1.2) can be written as

$$R_n(x) = \tau_0 \lambda\, P_n^{(\alpha,\beta)}(x) + \tau_1 \lambda\, P_{n-1}^{(\alpha,\beta)}(x) \tag{1.5}$$

for some x-independent coefficients τ_0 and τ_1, where $P_n^{(\alpha,\beta)}(x)$ is the Jacobi polynomial of degree n (see Appendix A). This follows from orthogonality of the Jacobi polynomials in $-1 < x < 1$ with respect to the Jacobi weight $W_{\alpha,\beta}(x)$ which implies orthogonality of the Jacobi polynomial of degree k to *any* polynomial of degree $k-1$ or less with respect to that weight function. Jacobi polynomials are the most general class of polynomial solutions of a Sturm–Liouville eigenproblem that is singular at ± 1 as required for faster than algebraic convergence [3, sections 9.2.2 and 9.6.1]. It is now easy to verify from (1.2) and (1.5) that if $\lambda = 0$, then $D^2 u_n(x) = R_n(x) = 0$ for all x in $(-1, 1)$, but the boundary conditions $u_n(\pm 1) = 0$ would then require that $u_n(x) = 0$ for all x in $[-1, 1]$. Therefore we can assume that $\lambda \neq 0$.

In the *Galerkin* approach, $u_n(x)$ is determined from the boundary conditions $u_n(\pm 1) = 0$ and orthogonality of the residual $R_n(x)$ to all polynomials of degree n that vanish at $x = \pm 1$, with respect to a weight function $W(x) \geq 0$. In other words, the *test functions* are in the same space (polynomials of degree n) as the trial functions, and they satisfy the same boundary conditions. Such polynomials can be written in the form $(1 - x^2) p_{n-2}(x)$, where $p_{n-2}(x)$ is an arbitrary polynomial of degree $n-2$, and the Galerkin equations can be written as

$$\int_{-1}^{1} R_n(x)(1 - x^2) p_{n-2}(x) W(x) dx = 0. \tag{1.6}$$

For the Jacobi weight $W(x) = W_{\alpha,\beta}(x) = (1-x)^\alpha(1+x)^\beta$, the Galerkin method is therefore equivalent to the tau method for the weight $W_{\alpha+1,\beta+1}(x)$, and the residual controlled by (1.6) has the form

$$R_n(x) = \tau_0 \lambda\, P_n^{(\alpha+1,\beta+1)}(x) + \tau_1 \lambda\, P_{n-1}^{(\alpha+1,\beta+1)}(x). \tag{1.7}$$

This residual can be written in terms of the derivatives of $P_{n+1}^{(\alpha,\beta)}(x)$ and $P_n^{(\alpha,\beta)}(x)$ using (A.4), but the key point is that since we consider a range of parameters α and β, the Jacobi tau method (1.4) also includes some Jacobi–Galerkin methods (1.6).

Characteristic polynomials for the eigenvalues $\mu = 1/\lambda$ are derived in section 2 from the explicit expression (1.3) for $u_n(x)$ in terms of the residual $R_n(x)$ specified by (1.5) and (1.7). The zeros of these characteristic polynomials are shown to be real, negative, and distinct in section 3, and the results are extended to Neumann and Dirichlet–Neumann boundary conditions. We conclude in section 4 with a brief discussion of further results involving recurrence relations for the characteristic polynomials in the Gegenbauer case and equivalence of these recurrences with an integration formulation. Those results are available in an extended online version of our article [4].

2. Characteristic polynomials.

2.1. Jacobi tau method. Substituting (1.5) into (1.3), the Jacobi tau approximation can be written explicitly in terms of the yet undetermined constants τ_0, τ_1 and the eigenvalue $\mu = 1/\lambda$ as

$$u_n(x) = \tau_0 \sum_{k=0}^{[\frac{n}{2}]} \mu^k D^{2k} P_n^{(\alpha,\beta)}(x) + \tau_1 \sum_{k=0}^{[\frac{n-1}{2}]} \mu^k D^{2k} P_{n-1}^{(\alpha,\beta)}(x). \tag{2.1}$$

The boundary conditions $u_n(\pm 1) = 0$ then yield the characteristic equations

$$\begin{cases} \tau_0 \displaystyle\sum_{k=0}^{[\frac{n}{2}]} \mu^k D^{2k} P_n^{(\alpha,\beta)}(1) + \tau_1 \sum_{k=0}^{[\frac{n-1}{2}]} \mu^k D^{2k} P_{n-1}^{(\alpha,\beta)}(1) = 0, \\ \tau_0 \displaystyle\sum_{k=0}^{[\frac{n}{2}]} \mu^k D^{2k} P_n^{(\alpha,\beta)}(-1) + \tau_1 \sum_{k=0}^{[\frac{n-1}{2}]} \mu^k D^{2k} P_{n-1}^{(\alpha,\beta)}(-1) = 0. \end{cases} \tag{2.2}$$

This system has a nontrivial solution $(\tau_0, \tau_1) \neq (0,0)$ if and only if

$$\left(\sum_{k=0}^{[\frac{n}{2}]} \mu^k D^{2k} P_n^{(\alpha,\beta)}(1) \right) \left(\sum_{k=0}^{[\frac{n-1}{2}]} \mu^k D^{2k} P_{n-1}^{(\beta,\alpha)}(1) \right) + \left(\sum_{k=0}^{[\frac{n-1}{2}]} \mu^k D^{2k} P_{n-1}^{(\alpha,\beta)}(1) \right) \left(\sum_{k=0}^{[\frac{n}{2}]} \mu^k D^{2k} P_n^{(\beta,\alpha)}(1) \right) = 0, \tag{2.3}$$

where we have used the symmetry $P_n^{(\alpha,\beta)}(-1) = (-1)^n P_n^{(\beta,\alpha)}(1)$ [1, eq. (22.4.1)], [11, eq. (1.5.2)]. This is the characteristic equation for the eigenvalue μ.

2.2. Gegenbauer tau method. Gegenbauer (a.k.a. ultraspherical) polynomials $G_n^{(\gamma)}(x)$ are the class of Jacobi polynomials $P_n^{(\alpha,\beta)}(x)$ with equal indices $\alpha = \beta = \gamma - 1/2$. The Gegenbauer polynomials are even in x for n even and odd for n odd [1, eq. (22.4.2)]. Chebyshev and Legendre polynomials are Gegenbauer polynomials with $\gamma = 0$ and $1/2$, respectively. We use a nonstandard normalization for Gegenbauer polynomials, denoted $G_n^{(\gamma)}(x)$ and defined in (A.6), since the standard normalization $C_n^{(\gamma)}(x)$ is singular in the Chebyshev case.

The symmetry of the differential equation (1.1) and of the Gegenbauer polynomials allows decoupling of the discrete problem into even and odd solutions. This parity reduction leads to simpler residuals and simpler forms for the corresponding characteristic polynomials. The residual (1.4) in the parity-separated Gegenbauer case contains only one term,

$$R_n(x) = \tau_0 \lambda\, G_n^{(\gamma)}(x), \tag{2.4}$$

where $G_n^{(\gamma)}(x)$ is the nth Gegenbauer polynomial and n is even for even solutions and odd for odd solutions. Substituting (2.4) into (1.3) provides the Gegenbauer tau approximation to (1.1) in terms of an undetermined constant τ_0 and the eigenvalue $\mu = 1/\lambda$:

$$u_n(x) = \tau_0 \sum_{k=0}^{[\frac{n}{2}]} \mu^k D^{2k} G_n^{(\gamma)}(x). \tag{2.5}$$

The boundary conditions $u_n(\pm 1) = 0$ lead to the characteristic polynomial equation

$$\sum_{k=0}^{[\frac{n}{2}]} \mu^k D^{2k} G_n^{(\gamma)}(1) = 0, \tag{2.6}$$

since by symmetry $G_n^{(\gamma)}(-1) = (-1)^n G_n^{(\gamma)}(1)$ and the two boundary conditions give the same equation.

3. Zeros of characteristic polynomials.

3.1. Stable polynomials and the Hermite–Biehler theorem. The general approach to proving that the eigenvalues are real, negative, and distinct is to construct a *stable* real polynomial $p(z)$ and then use the Hermite–Biehler theorem to deduce that the polynomials $\Omega_1(\mu)$ and $\Omega_2(\mu)$ such that $p(z) = \Omega_1(z^2) + z\Omega_2(z^2)$ have real, negative, and distinct zeros that interlace.

DEFINITION 3.1. *A real polynomial, $p(z)$, is stable (a.k.a.* Hurwitz*) if all its zeros lie in the open left half-plane, i.e., their real part is strictly less than zero, $\Re z < 0$.*

DEFINITION 3.2. *Two real polynomials $\Omega_1(\mu)$ and $\Omega_2(\mu)$ of degree n and $n-1$ (or n), respectively, form a positive pair if* (a) *the roots $\mu_1, \ldots, \mu_n$ of Ω_1 and $\mu'_1, \ldots, \mu'_{n-1}$ (or $\mu'_1, \ldots, \mu'_n$) of Ω_2 are real, negative, and distinct;* (b) *the roots strictly interlace (or alternate) as follows: $\mu_1 < \mu'_1 < \cdots < \mu'_{n-1} < \mu_n < 0$ (or $\mu'_1 < \mu_1 < \cdots < \mu'_n < \mu_n < 0$);* (c) *the highest coefficients of $\Omega_1(\mu)$ and $\Omega_2(\mu)$ are of like sign.*

THEOREM 3.3 (Hermite–Biehler). *A real polynomial $p(z) = \Omega_1(z^2) + z\Omega_2(z^2)$ is stable if and only if $\Omega_1(\mu)$ and $\Omega_2(\mu)$ form a positive pair.*

The Hermite–Biehler theorem (e.g., [11, p. 197], [10]) states that the even and odd parts of stable polynomials form positive pairs. This supplies us with a very

strong tool to prove reality and negativity of the roots of certain polynomials. We will also need the following lemmas about positive pairs [11, p. 198], [12, section 2].

LEMMA 3.4. *Any nontrivial real linear combination of two polynomials that form a positive pair has real roots.*

LEMMA 3.5. *Let $P(\mu)$ and $Q(\mu)$ be real standard polynomials (i.e., the leading coefficient is positive) with only nonpositive zeros. Then $P(\mu)$ interlaces (or alternates) $Q(\mu)$ (in the sense of Definition* 3.2 *but not strictly) if and only if for all $A > 0$ both $Q(\mu) + AP(\mu)$ and $Q(\mu) + A\mu P(\mu)$ have only nonpositive zeros.*

The polynomials P and Q of Theorem 3.5 do not necessarily form a *positive pair* since they are allowed to have common and/or multiple roots. We call this set of polynomials a *quasi-positive pair.*

LEMMA 3.6 (see [8, Lemma 3.4]). *If $\Omega_1(\mu), \Omega_2(\mu)$ and $\Theta_1(\mu), \Theta_2(\mu)$ are two positive pairs, then the zeros of $H(\mu) = \Omega_1(\mu)\Theta_2(\mu) + \Omega_2(\mu)\Theta_1(\mu)$ are real, negative, and distinct.*

Gottlieb and Lustman [8] used the Hermite–Biehler theorem to prove that the spectrum of the Chebyshev collocation operator for the heat equation is real, negative, and distinct for a variety of homogeneous boundary conditions. The basic strategy is to show that the characteristic polynomials for that method are the even or odd parts of a stable polynomial. Our results extend their strategy to a class of Jacobi and Gegenbauer tau methods that includes Chebyshev and Legendre tau and Galerkin formulations. Although the general approach is similar to that of Gottlieb [7] and Gottlieb and Lustman [8], the extension is technically nontrivial, and there are differences and some corrections. Their work is discussed in more detail in Appendix B, where we argue that their proof of stability for the collocation operator is not complete.

3.2. Important stable polynomials and positive pairs. Here we prove stability of certain real polynomials whose even and odd parts are directly related to the characteristic polynomials derived in section 2 for the Jacobi tau method.

THEOREM 3.7. *Let $P_n^{(\alpha,\beta)}(x)$ denote the Jacobi polynomial of degree n, where $n \geq 2$. If $-1 < \alpha \leq 1$ and $\beta > -1$, then the zeros of the polynomial*

$$\Phi_n(z) := \sum_{k=0}^{n} z^k \, D^k P_n^{(\alpha,\beta)}(1) \tag{3.1}$$

lie in the left half-plane; that is, $\Phi_n(z)$ is a stable polynomial.

The proof of this theorem is in [5] together with a discussion of its relation to zeros of Bessel polynomials. The next two theorems give generalizations of that result that are needed for this paper.

THEOREM 3.8. *Let $P_n^{(\alpha,\beta)}(x)$ denote the Jacobi polynomial of degree n with $n \geq 3$; then the polynomial*

$$\Phi_n^{(1)}(z) := \sum_{k=0}^{n} z^k \, D^k P_n^{(\alpha,\beta)}(1) + A \sum_{k=0}^{n-1} z^k \, D^k P_{n-1}^{(\alpha,\beta)}(1) \tag{3.2}$$

is stable for every $A \geq 0$ when $-1 < \alpha \leq 0$ and $\beta > -1$.

THEOREM 3.9. *Let $P_n^{(\alpha,\beta)}(x)$ denote the Jacobi polynomial of degree n with $n \geq 3$; then the polynomial*

$$\Phi_n^{(2)}(z) := \sum_{k=0}^{n} z^k \, D^k P_n^{(\alpha,\beta)}(1) + Az^2 \sum_{k=0}^{n-1} z^k \, D^k P_{n-1}^{(\alpha,\beta)}(1) \tag{3.3}$$

is stable for every $A \geq 0$ *when* $-1 < \alpha \leq 1$ *and* $\beta > -1$.

The proofs of these theorems are technical and are given in Appendix A. Our next theorem combines all the above theorems to get an important result.

THEOREM 3.10. *Let* $n \geq 3$*; then the polynomials*

$$\Omega_n^{(\alpha,\beta)}(\mu) := \sum_{k=0}^{[\frac{n}{2}]} \mu^k D^{2k} P_n^{(\alpha,\beta)}(1), \quad \Omega_{n-1}^{(\alpha,\beta)}(\mu) := \sum_{k=0}^{[\frac{n-1}{2}]} \mu^k D^{2k} P_{n-1}^{(\alpha,\beta)}(1) \tag{3.4}$$

form a positive pair if $-1 < \alpha, \beta \leq 0$ *or* $0 < \alpha, \beta \leq 1$.

Remark 1. It was shown in [5] that these polynomials have real, negative, and distinct roots for $-1 < \alpha \leq 1$ and $-1 < \beta$. The important addition of Theorem 3.10 is that the roots of these polynomials interlace, as was conjectured in [5].

Proof. Applying the Hermite–Biehler theorem to the stable polynomials of Theorem 3.7 for a given n and also for $n - 1 \geq 2$ proves that the polynomials $\Omega_n^{(\alpha,\beta)}(\mu)$ and $\Omega_{n-1}^{(\alpha,\beta)}(\mu)$ have real, negative, and distinct roots for $-1 < \alpha, \beta \leq 1$ (notice that we can interchange α and β). Applying the Hermite–Biehler theorem to the stable polynomials of Theorems 3.8 and 3.9 shows that the polynomials

$$\sum_{k=0}^{[\frac{n}{2}]} \mu^k D^{2k} P_n^{(\alpha,\beta)}(1) + A \sum_{k=0}^{[\frac{n-1}{2}]} \mu^k D^{2k} P_{n-1}^{(\alpha,\beta)}(1) = \Omega_n^{(\alpha,\beta)}(\mu) + A\Omega_{n-1}^{(\alpha,\beta)}(\mu), \tag{3.5}$$

$$\sum_{k=0}^{[\frac{n}{2}]} \mu^k D^{2k} P_n^{(\alpha,\beta)}(1) + A\mu \sum_{k=0}^{[\frac{n-1}{2}]} \mu^k D^{2k} P_{n-1}^{(\alpha,\beta)}(1) = \Omega_n^{(\alpha,\beta)}(\mu) + A\mu\Omega_{n-1}^{(\alpha,\beta)}(\mu) \tag{3.6}$$

have real, negative, and distinct roots for all $A > 0$ and $-1 < \alpha, \beta \leq 0$. These results provide sufficient information to apply Lemma 3.5 to deduce that the set of polynomials $(\Omega_n^{(\alpha,\beta)}(\mu), \Omega_{n-1}^{(\alpha,\beta)}(\mu))$ form quasi-positive pairs for $-1 < \alpha, \beta \leq 0$.

To show that these polynomials form a positive pair, recall that both $\Omega_n^{(\alpha,\beta)}(\mu)$ and $\Omega_{n-1}^{(\alpha,\beta)}(\mu)$ have real, negative, and distinct roots by Theorem 3.7. Thus it remains to show that they have no common roots. To do so, assume that μ is a common root. Then $P_1(\mu) = \Omega_n^{(\alpha,\beta)}(\mu) + A\Omega_{n-1}^{(\alpha,\beta)}(\mu) = 0$, and $P_2(\mu) = \Omega_n^{(\alpha,\beta)}(\mu) + A\mu\Omega_{n-1}^{(\alpha,\beta)}(\mu) = 0$. Since $D\Omega_n^{(\alpha,\beta)}(\mu) \neq 0$ and $D\Omega_{n-1}^{(\alpha,\beta)}(\mu) \neq 0$ (both Ω_n and Ω_{n-1} do not have a double root), set $A = -\frac{D\Omega_n^{(\alpha,\beta)}(\mu)}{D\Omega_{n-1}^{(\alpha,\beta)}(\mu)}$ or $A = -\frac{D\Omega_n^{(\alpha,\beta)}(\mu)}{\mu D\Omega_{n-1}^{(\alpha,\beta)}(\mu)}$, whichever one is positive (one of the two must be since μ is negative). But this will imply that $DP_1(\mu) = 0$ and $DP_2(\mu) = 0$, respectively, which is a contradiction since $P_1(\mu)$ and $P_2(\mu)$ have simple zeros. For the second range of parameters, replace n with $n + 1$ in Theorems 3.8 and 3.9 and apply the Hermite–Biehler theorem. This gives that the polynomials

$$\sum_{k=0}^{[\frac{n}{2}]} \mu^k D^{2k+1} P_{n+1}^{(\alpha,\beta)}(1) + A \sum_{k=0}^{[\frac{n-1}{2}]} \mu^k D^{2k+1} P_n^{(\alpha,\beta)}(1), \tag{3.7}$$

$$\sum_{k=0}^{[\frac{n}{2}]} \mu^k D^{2k+1} P_{n+1}^{(\alpha,\beta)}(1) + A\mu \sum_{k=0}^{[\frac{n-1}{2}]} \mu^k D^{2k+1} P_n^{(\alpha,\beta)}(1) \tag{3.8}$$

also have real, negative, and distinct roots. Using (A.4), these polynomials transform into

$$\sum_{k=0}^{[\frac{n}{2}]} \mu^k D^{2k} P_n^{(\alpha+1,\beta+1)}(1) + A' \sum_{k=0}^{[\frac{n-1}{2}]} \mu^k D^{2k} P_{n-1}^{(\alpha+1,\beta+1)}(1), \tag{3.9}$$

$$\sum_{k=0}^{[\frac{n}{2}]} \mu^k D^{2k} P_n^{(\alpha+1,\beta+1)}(1) + A'\mu \sum_{k=0}^{[\frac{n-1}{2}]} \mu^k D^{2k} P_{n-1}^{(\alpha+1,\beta+1)}(1). \tag{3.10}$$

The new polynomials have real, negative, and distinct roots for all $A' > 0$, where $A' = \frac{n+\alpha+\beta}{n+1+\alpha+\beta} A$, and thus the second range $0 < \alpha, \beta \leq 1$ follows from the first by a simple change of variables. ∎

3.3. Eigenvalues of the Gegenbauer and Jacobi tau methods. The previous subsection provides all necessary information needed for proving reality and negativity of the eigenvalues. First consider the Gegenbauer case.

THEOREM 3.11. *The eigenvalues of the Gegenbauer tau discretization of the second order operator with Dirichlet boundary conditions, problem* (1.1)*, are real, negative, and distinct for* $-1/2 < \gamma \leq 5/2$*. Also, eigenvalues given by successive order approximations (i.e.,* $n-1$ *and* n*) interlace.*

Proof. The eigenvalues are the roots of (2.6). The theorem follows directly from Theorem 3.10 since $\alpha = \beta = \gamma - 1/2$ in the Gegenbauer case and the two ranges for the indices (α, β) merge into the single range $-1/2 < \gamma \leq 5/2$. ∎

Remark 2. The results of Theorem 3.11 are sharp in the sense that well-conditioned numerical calculations give some complex conjugate pairs of eigenvalues for $\gamma > 5/2$ and Gegenbauer integration diverges in general for $\gamma \leq -1/2$.

Remark 3. As pointed out in the introduction, the Galerkin method with weight function $W_{\alpha,\beta}(x)$ for problem (1.1) is equivalent to the tau method with weight function $W_{\alpha+1,\beta+1}(x)$. For the Gegenbauer case, the Galerkin method with weight function $W_\gamma(x)$ is equivalent to the tau method with weight function $W_{\gamma+1}(x)$. Thus, a direct consequence of Theorem 3.11 is that the eigenvalues of the Gegenbauer–Galerkin method are real, negative, and distinct for $-1/2 < \gamma \leq 3/2$. Again, eigenvalues given by successive order approximations interlace.

Remark 4. Theorem 3.11 includes the Chebyshev and Legendre polynomials. For $\gamma = 0$, we have $T_n(x) = nG_n^{(0)}(x)$, where $T_n(x)$ denotes the nth Chebyshev polynomial of the first kind [11, p. 19]. Thus the theorem implies that the Chebyshev tau method has real, negative, and distinct eigenvalues that interlace for successive order approximations. For $\gamma = 1/2$ we have $G_n^{(1/2)}(x) = P_n(x)$, where $P_n(x)$ is the nth Legendre polynomial and for $\gamma = 1$ we have that $U_n(x) = 2G_n^{(1)}(x)$, where $U_n(x)$ is the Chebyshev polynomial of second kind. Therefore, the same result holds for both the Legendre tau and the Chebyshev tau of second kind method. Furthermore, in the Galerkin case (see Remark 3), the Galerkin–Chebyshev, the Galerkin–Legendre, and the Galerkin–Chebyshev of the second kind methods have real, negative, and distinct eigenvalues as well.

For the Jacobi case we have the following theorem.

THEOREM 3.12. *The eigenvalues of the Jacobi tau discretization of the second order operator with Dirichlet boundary conditions, problem* (1.1)*, are real, negative, and distinct if* $-1 < \alpha, \beta \leq 0$ *or* $0 < \alpha, \beta \leq 1$.

Proof. By Theorem 3.10 the polynomials $(\Omega_n^{(\alpha,\beta)}(\mu), \Omega_{n-1}^{(\alpha,\beta)}(\mu))$ form a positive pair for $-1 < \alpha, \beta \leq 0$ and $0 < \alpha, \beta \leq 1$. Interchanging the indices α, β, the same result holds for $(\Omega_n^{(\beta,\alpha)}(\mu), \Omega_{n-1}^{(\beta,\alpha)}(\mu))$. Application of Lemma 3.6 to these sets of positive pairs gives that the polynomial

$$B_n^{(\alpha,\beta)}(\mu) = \Omega_n^{(\alpha,\beta)}(\mu)\Omega_{n-1}^{(\beta,\alpha)}(\mu) + \Omega_n^{(\beta,\alpha)}(\mu)\Omega_{n-1}^{(\alpha,\beta)}(\mu) \tag{3.11}$$

has real, negative, and distinct roots. Equation (2.3) shows that $B_n^{(\alpha,\beta)}(\mu)$ is the characteristic polynomial for the Jacobi tau method. ∎

This paper focuses on the second order problem with Dirichlet boundary conditions. Naturally, questions arise about the equivalent results for different boundary conditions. The next remarks give some answers to those questions.

Remark 5. Consider problem (1.1) with Neumann boundary conditions, i.e., $\lambda u - D^2u = 0$ with $Du(\pm 1) = 0$. The Jacobi tau method gives $\lambda u - D^2u = \tau_0 P_n^{(\alpha,\beta)}(x) + \tau_1 P_{n-1}^{(\alpha,\beta)}(x)$. Notice that $\lambda = 0$ is an eigenvalue since $u = constant$ is a solution. For $\lambda \neq 0$, differentiate the last equation to get $\lambda Du - D^3u = \tau_0 DP_n^{(\alpha,\beta)}(x) + \tau_1 DP_{n-1}^{(\alpha,\beta)}(x)$ with $Du(\pm 1) = 0$. Now set $v = Du$; with the use of (A.4) the equation transforms to $\lambda v - D^2v = \tau_0' P_{n-1}^{(\alpha+1,\beta+1)}(x) + \tau_1' P_{n-2}^{(\alpha+1,\beta+1)}(x)$ with $v(\pm 1) = 0$. This is the Jacobi tau approximation of second kind, i.e., $(\alpha, \beta) \to (\alpha + 1, \beta + 1)$ for problem (1.1), and by Theorem 3.12 it has real, negative, and distinct eigenvalues for $-1 < \alpha, \beta \leq 0$.

Remark 6. Consider problem (1.1) with boundary conditions $u(-1) = 0$ and $Du(1) = 0$. Using equations (2.1) and (A.4) and following the ideas of subsection 2.1, the characteristic polynomial in this case is $B_n^{(\alpha,\beta)}(\mu) = k_{n-1}\Omega_n^{(\beta,\alpha)}(\mu)\Omega_{n-2}^{(\alpha+1,\beta+1)}(\mu) + k_n\Omega_{n-1}^{(\beta,\alpha)}(\mu)\Omega_{n-1}^{(\alpha+1,\beta+1)}(\mu)$, where $k_n = \frac{1}{2}(n + \alpha + \beta + 1)$. Since the polynomials $(\Omega_n^{(\beta,\alpha)}(\mu), \Omega_{n-1}^{(\beta,\alpha)}(\mu))$ and $(k_n\Omega_{n-1}^{(\alpha+1,\beta+1)}(\mu), k_{n-1}\Omega_{n-2}^{(\alpha+1,\beta+1)}(\mu))$ form positive pairs for $-1 < \alpha, \beta \leq 0$, by Theorem 3.6 the roots of $B_n^{(\alpha,\beta)}(\mu)$ are real, negative, and distinct for $-1 < \alpha, \beta \leq 0$.

An alternative implementation is to rescale the domain to $[-1, 0]$ with the Neuman boundary condition $Du(0) = 0$ and to use only the even Gegenbauer polynomials with a Gegenbauer tau approach on the entire domain $[-1, 1]$ and boundary conditions $u(\pm 1) = 0$ since such polynomials automatically satisfy $Du(0) = 0$.

4. Conclusions. We have proved that the eigenvalues of the Jacobi tau method for the second derivative operator with Dirichlet boundary conditions are real, negative, and distinct for ranges of the Jacobi indices α and β. These ranges include tau methods with Chebyshev and Legendre polynomials of the first and second kinds. Chebyshev and Legendre–Galerkin formulations are included as well, but collocation methods are not. Although our work owes much to earlier work by Gottlieb and Lustman [7, 8], we have raised doubts about the completeness of their proof for the Chebyshev collocation operator.

Special emphasis has been placed on the symmetric case of the Gegenbauer tau method where the range of parameters included in the theorems can be extended and characteristic polynomials given by successive order approximations interlace. The interlacing proof is between successive even and odd approximations [5, Conjecture 1], not between successive even or odd approximations, although we believe the latter is true as well [5, Conjecture 2].

In an extended version of this paper posted online [4], the characteristic polynomials (2.6) of both even and odd Gegenbauer tau approximations are shown to satisfy three term recurrences *plus* a constant term. That constant vanishes for the

Legendre tau and Galerkin methods ($\gamma = 1/2$ and $\gamma = 3/2$). Hence for those cases the characteristic polynomials are new classes of orthogonal polynomials, and their roots interlace; that is, the roots of successive even or odd Legendre tau or Galerkin approximations interlace. It is also shown in [4] that the integration formulation of spectral tau methods leads directly to the same tridiagonal plus one row matrices as those arising from the characteristic polynomial recurrences. These recurrences and integration formulation are perfectly conditioned, and some numerical results up to approximations of degree 2000 together with a MATLAB code are provided.

In a forthcoming paper we apply similar methods to the simplified Stokes eigenvalue problem $D^4u = \lambda D^2u$ with $u(\pm1) = Du(\pm1) = 0$ and rigorously identify classes of spectral methods that are free of spurious eigenvalues.

Appendix A. Proofs of Theorems 3.8 and 3.9.

We recall some properties of the Jacobi polynomials that are used in the proofs of Theorems 3.8 and 3.9.

A.1. Jacobi. The Jacobi polynomials $P_n^{(\alpha,\beta)}(x)$ [11, eq. (1.5.2)] are suitably standardized orthogonal polynomials with weight function $W_{\alpha,\beta}(x) = (1-x)^\alpha(1+x)^\beta$. They satisfy the orthogonality relation

$$\int_{-1}^{1}(1-x)^\alpha(1+x)^\beta\, P_m^{(\alpha,\beta)}P_n^{(\alpha,\beta)}\,dx = \begin{cases} 0, & m \neq n, \\ h_n^{\alpha,\beta}, & m = n, \end{cases} \tag{A.1}$$

where $h_n^{\alpha,\beta} > 0$ is given in [11, eq. (1.5.5)]. Orthogonal polynomials satisfy a three term recurrence relation; for the Jacobi polynomials this reads as

$$\begin{aligned} 2(n+1)(n+\alpha+\beta+1)(2n+\alpha+\beta)P_{n+1}^{(\alpha,\beta)}(x) \\ = \big((2n+\alpha+\beta+1)(\alpha^2-\beta^2) + (2n+\alpha+\beta)_3\, x\big)\, P_n^{(\alpha,\beta)}(x) \\ - 2(n+\alpha)(n+\beta)(2n+\alpha+\beta+2)P_{n-1}^{(\alpha,\beta)}(x), \end{aligned} \tag{A.2}$$

where $(2n+\alpha+\beta)_3 = (2n+\alpha+\beta)(2n+\alpha+\beta+1)(2n+\alpha+\beta+2)$. To ease the notation in our calculations we write the recurrence relation in the form

$$a_{1,n}P_{n+1}^{(\alpha,\beta)}(x) = (a_{2,n} + a_{3,n}x)P_n^{(\alpha,\beta)}(x) - a_{4,n}P_{n-1}^{(\alpha,\beta)}(x). \tag{A.3}$$

Two other useful relations involving derivatives of Jacobi polynomials [6] are

$$\frac{d}{dx}P_n^{(\alpha,\beta)}(x) = \frac{1}{2}(n+\alpha+\beta+1)P_{n-1}^{(\alpha+1,\beta+1)}(x) \tag{A.4}$$

and

$$\frac{d}{dx}P_{n+1}^{(\alpha,\beta)}(x) = B_nP_n^{(\alpha,\beta)}(x) + p_{n-1}(x) \tag{A.5}$$

with $B_n = \frac{(2n+\alpha+\beta+1)(2n+\alpha+\beta+2)}{2(n+\alpha+\beta+1)}$ and $p_{n-1}(x)$ a polynomial of degree $n-1$.

A.2. Gegenbauer. The Gegenbauer (a.k.a. ultraspherical) polynomials $C_n^{(\gamma)}(x)$, $\gamma > -1/2$, of degree n are the Jacobi polynomials with $\alpha = \beta = \gamma - 1/2$, up to normalization [1, eq. (22.5.20)]. Since the standard normalization [1, eq. (22.3.4)] is singular for the Chebyshev case $\gamma = 0$, we use a nonstandard normalization denoted by $G_n^{(\gamma)}(x)$ that includes the Chebyshev polynomials but preserves the simplicity of the Gegenbauer recurrences. Our normalization is defined as

$$G_0^{(\gamma)}(x) := 1, \qquad G_n^{(\gamma)}(x) := \frac{C_n^{(\gamma)}(x)}{2\gamma}, \quad n \geq 1. \tag{A.6}$$

A.3. Theorem 3.8.

Proof. For $A = 0$ the theorem reduces to Theorem 3.7. Now fix $A > 0$ but otherwise arbitrary. Let

$$f_n(x;z) = \sum_{k=0}^{n} z^k D^k P_n^{(\alpha,\beta)}(x) + A \sum_{k=0}^{n-1} z^k D^k P_{n-1}^{(\alpha,\beta)}(x) \tag{A.7}$$

with z such that $f_n(1;z) = 0$ and $D := d/dx$. Then $f_n(x;z)$ satisfies the following differential equation:

$$\left(f_n - P_n^{(\alpha,\beta)} - AP_{n-1}^{(\alpha,\beta)}\right) = z\frac{df_n}{dx}. \tag{A.8}$$

Multiplying by $\frac{df_n^*(x;z)}{dx}(1+x)$, integrating from -1 to 1 in the Jacobi norm, and adding the conjugate, we obtain

$$\int_{-1}^{1} \frac{d|f_n|^2}{dx}(1+x)W_{\alpha,\beta}dx - \int_{-1}^{1}\left(\frac{df_n}{dx} + \frac{df_n^*}{dx}\right)(1+x)P_n^{(\alpha,\beta)}(x)W_{\alpha,\beta}dx \tag{A.9}$$

$$-A\int_{-1}^{1}\left(\frac{df_n}{dx} + \frac{df_n^*}{dx}\right)(1+x)P_{n-1}^{(\alpha,\beta)}(x)W_{\alpha,\beta}dx = (z+z^*)\int_{-1}^{1}\left|\frac{df_n}{dx}\right|^2(1+x)W_{\alpha,\beta}dx.$$

For the first term of equation (A.9), integration by parts yields

$$\int_{-1}^{1} \frac{d|f_n|^2}{dx}(1+x)W_{\alpha,\beta}dx = -\int_{-1}^{1}|f_n|^2\frac{W_{\alpha,\beta}}{(1-x)}\left(\beta+1-\alpha-(\beta+1+\alpha)x\right)dx. \tag{A.10}$$

For $\beta > -1$ and $\alpha \le 0$ the factor $\beta + 1 - \alpha - (\beta + 1 + \alpha)x$ is nonnegative for all $x \in [-1, 1]$. The second term of (A.9) can be expanded as

$$\begin{aligned}\int_{-1}^{1}\left(\frac{df_n}{dx} + \frac{df_n^*}{dx}\right)(1+x)P_n^{(\alpha,\beta)}W_{\alpha,\beta}dx &= 2\int_{-1}^{1} xDP_n^{(\alpha,\beta)}P_n^{(\alpha,\beta)}W_{\alpha,\beta}dx \\ &= 2B_{n-1}\int_{-1}^{1} xP_{n-1}^{(\alpha,\beta)}P_n^{(\alpha,\beta)}W_{\alpha,\beta}dx \\ &= \frac{2B_{n-1}a_{1,n-1}}{a_{3,n-1}}\int_{-1}^{1}\left(P_n^{(\alpha,\beta)}\right)^2 W_{\alpha,\beta}dx \\ &= \frac{2B_{n-1}a_{1,n-1}}{a_{3,n-1}}h_n^{\alpha,\beta},\end{aligned} \tag{A.11}$$

where we have used expression (A.5), the Jacobi recurrence relation (A.3), and orthogonality of $P_n^{(\alpha,\beta)}(x)$ to all polynomials of degree less than n with respect to the

Jacobi weight $W_{\alpha,\beta}(x)$. Similarly, the third term of (A.9) can be calculated as

(A.12)
$$\begin{aligned}
&\int_{-1}^{1}\left(\frac{df_n}{dx}+\frac{df_n^*}{dx}\right)(1+x)P_{n-1}^{(\alpha,\beta)}W_{\alpha,\beta}dx\\
&=2\int_{-1}^{1}(1+x)DP_n^{(\alpha,\beta)}P_{n-1}^{(\alpha,\beta)}W_{\alpha,\beta}dx+(z+z^*)\int_{-1}^{1}xD^2P_n^{(\alpha,\beta)}P_{n-1}^{(\alpha,\beta)}W_{\alpha,\beta}dx\\
&\quad+2A\int_{-1}^{1}xDP_{n-1}^{(\alpha,\beta)}P_{n-1}^{(\alpha,\beta)}W_{\alpha,\beta}dx\\
&=2B_{n-1}\int_{-1}^{1}\left(P_{n-1}^{(\alpha,\beta)}\right)^2W_{\alpha,\beta}dx+2B_{n-1}\int_{-1}^{1}x\left(P_{n-1}^{(\alpha,\beta)}\right)^2W_{\alpha,\beta}dx\\
&\quad+(z+z^*)B_{n-1}B_{n-2}\int_{-1}^{1}xP_{N-2}^{(\alpha,\beta)}P_{n-1}^{(\alpha,\beta)}W_{\alpha,\beta}dx\\
&\quad+2AB_{n-2}\int_{-1}^{1}xP_{N-2}^{(\alpha,\beta)}P_{n-1}^{(\alpha,\beta)}W_{\alpha,\beta}dx\\
&=2B_{n-1}h_{n-1}^{\alpha,\beta}-2\frac{a_{2,n-1}}{a_{3,n-1}}B_{n-1}h_{n-1}^{\alpha,\beta}+(z+z^*)\frac{a_{1,n-2}}{a_{3,n-2}}B_{n-1}B_{n-2}h_{n-1}^{\alpha,\beta}\\
&\quad+2A\frac{a_{1,n-2}}{a_{3,n-2}}B_{n-2}h_{n-1}^{\alpha,\beta}.
\end{aligned}$$

Substituting these expressions back into (A.9) yields

$$\begin{aligned}
&-\left[\int_{-1}^{1}|f_n|^2\frac{W_{\alpha,\beta}}{(1-x)}(\beta+1-\alpha-(\beta+1+\alpha)x)dx+\frac{2B_{n-1}a_{1,n-1}}{a_{3,n-1}}h_n^{\alpha,\beta}\right.\\
&\qquad\left.+2A\left(\left(1-\frac{a_{2,n-1}}{a_{3,n-1}}\right)B_{n-1}+A\frac{a_{1,n-2}}{a_{3,n-2}}B_{n-2}\right)h_{n-1}^{\alpha,\beta}\right]\\
&\qquad=(z+z^*)\left[\int_{-1}^{1}\left|\frac{df_n}{dx}\right|^2(1+x)W_{\alpha,\beta}dx+A\frac{a_{1,n-2}}{a_{3,n-2}}B_{n-1}B_{n-2}h_{n-1}^{\alpha,\beta}\right].
\end{aligned}\tag{A.13}$$

Since $1>\frac{a_{2,n-1}}{a_{3,n-1}}$, it follows that the left-hand side is negative and $\Re(z)<0$, which ensures stability. ∎

A.4. Theorem 3.9.

Proof. For $A=0$ the theorem reduces again to Theorem 3.7. Now fix $A>0$. Let

$$f_n(x;z)=\sum_{k=0}^{n}z^kD^kP_n^{(\alpha,\beta)}(x)+Az^2\sum_{k=0}^{n-1}z^kD^kP_{n-1}^{(\alpha,\beta)}(x)\tag{A.14}$$

with $f_n(1;z)=0$. Then $f_n(x;z)$ satisfies the differential equation

$$\frac{1}{z}\left(f_n-P_n^{(\alpha,\beta)}-Az^2P_{n-1}^{(\alpha,\beta)}\right)=\frac{df_n}{dx}.\tag{A.15}$$

Multiplying by $f_n^*(x;z)\frac{(1+x)}{(1-x)}$, integrating from -1 to 1, and adding the conjugate yields

$$\left(\frac{1}{z}+\frac{1}{z^*}\right)\int_{-1}^{1}|f_n|^2\frac{(1+x)}{(1-x)}W_{\alpha,\beta}dx-\frac{1}{z}\int_{-1}^{1}f_n^*\frac{(1+x)}{(1-x)}P_n^{(\alpha,\beta)}(x)W_{\alpha,\beta}dx \tag{A.16}$$

$$-\frac{1}{z^*}\int_{-1}^{1}f_n\frac{(1+x)}{(1-x)}P_n^{(\alpha,\beta)}(x)W_{\alpha,\beta}dx-Az\int_{-1}^{1}f_n^*\frac{(1+x)}{(1-x)}P_{n-1}^{(\alpha,\beta)}(x)W_{\alpha,\beta}dx$$

$$-Az^*\int_{-1}^{1}f_n\frac{(1+x)}{(1-x)}P_{n-1}^{(\alpha,\beta)}(x)W_{\alpha,\beta}dx\;=\int_{-1}^{1}\frac{d|f_n|^2}{dx}\frac{(1+x)}{(1-x)}W_{\alpha,\beta}dx.$$

Integration by parts on the first term gives

$$\int_{-1}^{1}\frac{d|f_n|^2}{dx}(1+x)W_{\alpha,\beta}dx=-\int_{-1}^{1}|f_n|^2\frac{W_{\alpha,\beta}}{(1-x)^2}\left(\beta-\alpha+2-(\beta+\alpha)x\right)dx. \tag{A.17}$$

For $\beta>-1$ and $\alpha\le 1$ the factor $\beta-\alpha+2-(\beta+\alpha)x$ is nonnegative for all $x\in[-1,1]$. For the other terms on the left-hand side of (A.16), recall that $f_n(1;z)=0$ and write

$$f_n=(1-x)\sum_{k=0}^{n-1}c_kP_k^{(\alpha,\beta)}(x). \tag{A.18}$$

Then

$$\int_{-1}^{1}\frac{(1+x)}{(1-x)}f_n^*P_n^{(\alpha,\beta)}(x)W_{\alpha,\beta}dx=\int_{-1}^{1}c_{n-1}^*xP_{n-1}^{(\alpha,\beta)}(x)P_n^{(\alpha,\beta)}(x)W_{\alpha,\beta}dx \tag{A.19}$$

$$=c_{n-1}^*\frac{a_{1,n-1}}{a_{3,n-1}}\int_{-1}^{1}\left(P_n^{(\alpha,\beta)}(x)\right)^2W_{\alpha,\beta}dx=c_{n-1}^*\frac{a_{1,n-1}}{a_{3,n-1}}h_n^{\alpha,\beta}.$$

Also

$$\int_{-1}^{1}\frac{(1+x)}{(1-x)}f_n^*P_{n-1}^{(\alpha,\beta)}(x)W_{\alpha,\beta}dx \tag{A.20}$$

$$=\int_{-1}^{1}c_{n-1}^*(1+x)P_{n-1}^{(\alpha,\beta)}(x)P_{n-1}^{(\alpha,\beta)}(x)W_{\alpha,\beta}dx+\int_{-1}^{1}c_{n-2}^*xP_{n-2}^{(\alpha,\beta)}(x)P_{n-1}^{(\alpha,\beta)}(x)W_{\alpha,\beta}dx$$

$$=c_{n-1}^*\int_{-1}^{1}\left(P_{n-1}^{(\alpha,\beta)}(x)\right)^2W_{\alpha,\beta}dx-c_{n-1}^*\frac{a_{2,n-1}}{a_{3,n-1}}\int_{-1}^{1}\left(P_{n-1}^{(\alpha,\beta)}(x)\right)^2W_{\alpha,\beta}dx$$

$$+c_{n-2}^*\frac{a_{1,n-2}}{a_{3,n-2}}\int_{-1}^{1}\left(P_{n-1}^{(\alpha,\beta)}(x)\right)^2W_{\alpha,\beta}dx$$

$$=\left(c_{n-1}^*\left[1-\frac{a_{2,n-1}}{a_{3,n-1}}\right]+c_{n-2}^*\frac{a_{1,n-2}}{a_{3,n-2}}\right)h_{n-1}^{\alpha,\beta}.$$

Explicit values of c_{n-1} and c_{n-2} follow from (A.18):

$$
\begin{aligned}
f_n &= (1-x)\sum_{k=0}^{n-1} c_k P_k^{(\alpha,\beta)}(x) \\
&= c_{n-1}P_{n-1}^{(\alpha,\beta)}(x) - c_{n-1}xP_{n-1}^{(\alpha,\beta)}(x) - c_{n-2}xP_{n-2}^{(\alpha,\beta)}(x) + O(n-2) \\
&= -c_{n-1}\frac{a_{1,n-1}}{a_{3,n-1}}P_n^{(\alpha,\beta)}(x) \\
&\quad + \left(c_{n-1}\left[1+\frac{a_{2,n-1}}{a_{3,n-1}}\right] - c_{n-2}\frac{a_{1,n-2}}{a_{3,n-2}}\right)P_{n-1}^{(\alpha,\beta)}(x) + O(n-2).
\end{aligned} \tag{A.21}
$$

Now from (A.14),

$$
\begin{aligned}
f_n &= P_n^{(\alpha,\beta)}(x) + zDP_n^{(\alpha,\beta)}(x) + Az^2P_{n-1}^{(\alpha,\beta)}(x) + O(n-2) \\
&= P_n^{(\alpha,\beta)}(x) + B_{n-1}zP_{n-1}^{(\alpha,\beta)}(x) + Az^2P_{n-1}^{(\alpha,\beta)}(x) + O(n-2).
\end{aligned} \tag{A.22}
$$

Comparing these two expressions for f_n gives

$$
c_{n-1} = -\frac{a_{3,n-1}}{a_{1,n-1}}, \qquad c_{n-2} = \frac{a_{3,n-2}}{a_{1,n-2}}\left(-\frac{a_{2,n-1}}{a_{1,n-1}} - \frac{a_{3,n-1}}{a_{1,n-1}} - B_{n-1}z - Az^2\right). \tag{A.23}
$$

Substituting all these results back into (A.16) yields

$$
\begin{aligned}
&\left(\frac{1}{z}+\frac{1}{z^*}\right)\left[\int_{-1}^{1}|f_n|^2\frac{(1+x)}{(1-x)}W_{\alpha,\beta}dx + h_n^{\alpha,\beta}\right] + 2A(z+z^*)\frac{a_{3,n-1}}{a_{1,n-1}}h_{n-1}^{\alpha,\beta} \\
&+ 2AB_{n-1}|z|^2h_{n-1}^{\alpha,\beta} + (z+z^*)A^2|z|^2h_{n-1}^{\alpha,\beta} = -\int_{-1}^{1}|f_n|^2\frac{(\alpha-\beta+2-(\alpha+\beta)x)}{(1-x)^2}W_{\alpha,\beta}dx
\end{aligned} \tag{A.24}
$$

or, after rearranging some of the terms,

$$
\begin{aligned}
&(z+z^*)\left(\frac{1}{|z|^2}\left[\int_{-1}^{1}|f_n|^2\frac{(1+x)}{(1-x)}W_{\alpha,\beta}dx + h_n^{\alpha,\beta}\right] + \left(2A\frac{a_{3,n-1}}{a_{1,n-1}} + A^2|z|^2\right)h_{n-1}^{\alpha,\beta}\right) \\
&\qquad = -\int_{-1}^{1}|f_n|^2\frac{(\beta-\alpha+2-(\beta+\alpha)x)}{(1-x)^2}W_{\alpha,\beta}dx - 2AB_{n-1}|z|^2h_{n-1}^{\alpha,\beta}.
\end{aligned} \tag{A.25}
$$

The right-hand side is negative, so this implies that $\Re(z) < 0$. ∎

Appendix B. On stability for collocation.

Gottlieb and Lustman [8] focus on the Chebyshev collocation method where $u_n(x)$ is determined from the boundary conditions $u_n(\pm1) = 0$ and vanishing of the residual $R_n(x_j) = 0$ at $x_j = \cos j\pi/n$, $j = 1, \ldots, n-1$. The collocation residual has the form [8, eq. (4.5)] $R_n(x) = (A+Bx)DT_n(x)$, where $T_n(x)$ is the Chebyshev polynomial of degree n. Substituting into (1.3) leads to the characteristic polynomial as in section 2. Gottlieb and Lustman consider [8, eqs. (4.11), (4.12)]

$$
f_n(x;z) = \sum_{k=0}^{n-1} z^{k+1}D^{k+1}T_n(x) \quad \text{and} \quad g_n(x;z) = \sum_{k=0}^{n} z^{k+1}D^k\left(xDT_n(x)\right). \tag{B.1}
$$

To prove that $f_n(1;z)$ is stable, they first note that $f_n(x;z)$ satisfies $z^{-1}f_n - Df_n = DT_n$ and then consider $w_n(x,t;z) = e^{t/z}f_n(x;z)$ with the boundary condition $w_n(1,t;z) = e^{t/z}f_n(1;z) = 0$. That w_n satisfies the PDE

$$\partial_t w_n - \partial_x w_n = e^{t/z} DT_n(x). \tag{B.2}$$

Gottlieb and Lustman state *"In* [7] *it has been shown that w_n decays in time and therefore that $\Re(z) < 0$."* Presumably, this refers to the stability result [7, eq. (3.18)] for the polynomial approximation $v_n(x,t) = \sum_{k=0}^{n-1} d_k(t)T_k(x)$ to $\partial_t v - \partial_x v = 0$ with $v(1,t) = 0$ obtained by imposing $v_n(1,t) = 0$ and collocating at $x_j = \cos j\pi/n$, $j = 1,\dots,n-1$. The approximation $v_n(x,t)$ satisfies [7, eq. (3.11b)]

$$\partial_t v_n - \partial_x v_n = \tau_2(t)\, DT_n(x) \quad \text{with} \quad v_n(1,t) = 0. \tag{B.3}$$

Gottlieb shows stability of $v_n(x,t)$ by evaluating (B.3) at the interior point x_j where $DT_n(x_j) = 0$, multiplying by $(1+x_j)v_n(x_j,t)w_j$, and summing over $j = 1,\dots,n-1$, where $w_j > 0$ are the Chebyshev Gauss–Lobatto weights (Gottlieb's lemma [7, eq. (3.14)] is not necessary; one can directly use [7, eq. (3.15)] as done here). This yields

$$\begin{aligned} \frac{d}{dt}\sum_{j=1}^{n-1}(1+x_j)v_n^2(x_j,t)w_j &= \sum_{j=1}^{n-1}(1+x_j)v_n(x_j,t)\partial_x v_n(x_j,t)\, w_j \\ &= \int_{-1}^{1}(1+x)v_n\partial_x v_n \frac{dx}{\sqrt{1-x^2}} = \cdots \le 0. \end{aligned} \tag{B.4}$$

The last equality holds because v_n is of degree $n-1$ in x, so $(1+x)v_n\partial_x v_n$ is of degree $2n-2$ and vanishes at $x = \pm 1$; thus Gauss–Lobatto integration with $n-1$ interior points is exact. The integral can be calculated by parts to show that it is ≤ 0 and deduce stability since each term in the sum on the left-hand side is nonnegative.

That this result implies stability of w_n in (B.2) is not quite correct since v_n is assumed to be real, while $w_n(t,x;z)$ is complex since z is complex. However, this is a minor issue that is easily fixed. Our proof in [5] (Theorem 3.7 above) generalizes this result to a class of Jacobi polynomials without going back to a PDE such as (B.2) and without using Gauss integration.

For the proof of stability of $g_n(1;z)$, Gottlieb and Lustman simply state that *"a similar argument holds"* [8, eq. (4.12)], but we do not think it is that simple. Gottlieb [7] makes a similar claim that the stability of another approximation $u_n(x,t)$ which satisfies [7, eq. (3.11a)]

$$\partial_t u_n - \partial_x u_n = \tau_1(t)(1+x)DT_n(x) \quad \text{with} \quad u_n(1,t) = 0 \tag{B.5}$$

follows directly from the stability of (B.3) [7, eq. (3.11b)], but that does not seem correct since $\tau_1(t)$ and $\tau_2(t)$ are different functions of t. The Gauss integration procedure (B.4) does not work here since u_n is of degree n; thus $(1+x)u_n\partial_x u_n$ is of degree $2n$ and Gauss–Lobatto integration is not exact.

Gottlieb states another direct relation [7, eq. (3.8)] between $u_n(x,t)$ and $v_n(x,t)$, but that, too, does not seem correct. While it is true that one can write

$$u_n(x_j,t) = \sum_{k=0}^{n} a_k(t)T_k(x_j) = \sum_{k=0}^{n-1}\tilde{d}_k(t)T_k(x_j) \tag{B.6}$$

for $x_j = \cos j\pi/n$, $j = 0, \dots, n-1$, with $d_k = a_k + (-1)^{n-1} 2a_n (-1)^k / c_k$ [7, (3.8)], where $c_0 = c_n = 2$, $c_k = 1$ for $0 \neq k \neq n$, there is no reason why those $\tilde{d}_k(t)$ coefficients would be identical to the corresponding coefficients for $v_n(x_j, t)$ since $u_n(x_j, t)$ and $v_n(x_j, t)$ are distinct solutions determined by different sets of equations.[1] Hence, it appears that there is currently no proof of the stability of [7, eqs. (3.6), (3.11a)] and [8, eq. (4.12)], therefore invalidating Gottlieb and Lustman's proof that the eigenvalues of the Chebyshev collocation operator are real, negative, and distinct [8].

Acknowledgment. The authors thank Jue Wang for several helpful calculations in the early stages of this work.

REFERENCES

[1] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, Dover, New York, 1965.
[2] J. P. Boyd, *Chebyshev and Fourier Spectral Methods*, Dover, New York, 2001.
[3] C. Canuto, M. Y. Hussaini, A. Quarteroni, and T. A. Zang, *Spectral Methods in Fluid Dynamics*, Springer, New York, 1988.
[4] M. Charalambides and F. Waleffe, *Spectrum of the Jacobi Tau Approximation for the Second Derivative Operator*, 2006; available online at http://www.arxiv.org/abs/math.NA/0607640.
[5] G. Csordas, M. Charalambides, and F. Waleffe, *A new property of a class of Jacobi polynomials*, Proc. Amer. Math. Soc., 133 (2005), pp. 3551–3560.
[6] E. H. Doha, *On the coefficients of differentiated expansions and derivatives of Jacobi polynomials*, J. Phys. A., 35 (2002), pp. 3467–3478.
[7] D. Gottlieb, *The stability of pseudospectral-Chebyshev methods*, Math. Comp., 36 (1981), pp. 107–118.
[8] D. Gottlieb and L. Lustman, *The spectrum of the Chebyshev collocation operator for the heat equation*, SIAM J. Numer. Anal., 20 (1983), pp. 909–921.
[9] D. Gottlieb and S. A. Orszag, *Numerical Analysis of Spectral Methods: Theory and Applications*, SIAM, Philadelphia, 1977.
[10] O. Holtz, *Hermite-Biehler, Routh-Hurwitz, and total positivity*, Linear Algebra Appl., 372 (2003), pp. 105–110.
[11] Q. I. Rahman and G. Schmeisser, *Analytic Theory of Polynomials*, Oxford University Press, New York, 2002.
[12] D. Wagner, *Zeros of reliability polynomials and f-vectors of matroids*, Combin. Probab. Comput., 9 (2000), pp. 167–190.
[13] J. A. C. Weideman and L. N. Trefethen, *The eigenvalues of second-order spectral differentiation matrices*, SIAM J. Numer. Anal., 25 (1988), pp. 1279–1298.

[1] The reduction in (B.6) follows from the Chebyshev expansion of $(1-x)DT_n(x)$ which vanishes at all $x_j = \cos j\pi/n$, $j = 0, \dots, n-1$. One obtains $(1-x)DT_n(x) = \sum_{k=0}^{n} \frac{2}{nc_k}(-1)^k DT_n(-1)T_k(x)$ by Chebyshev transform (e.g., [7, eqs. (3.2), (3.3)]), from which one can express $T_n(x_j)$ as a combination of $T_k(x_j)$ for $j, k = 0, \dots, n-1$.

SIAM J. NUMER. ANAL.
Vol. 46, No. 1, pp. 295–324
© 2008 Society for Industrial and Applied Mathematics

THREE-SCALE FINITE ELEMENT DISCRETIZATIONS FOR QUANTUM EIGENVALUE PROBLEMS*

XIAOYING DAI† AND AIHUI ZHOU‡

Abstract. Based on globally and locally coupled discretizations, some three-scale finite element schemes are proposed in this paper for a class of quantum eigenvalue problems. It is shown that the solution of a quantum eigenvalue problem on a fine grid may be reduced to the solution of an eigenvalue problem on a relatively coarse grid, and the solutions of linear algebraic systems on a globally mesoscopic grid and the locally fine grid, and the resulting solution is still very satisfactory.

Key words. eigenvalue, finite element, ground state energy, local computation, three-scale, quantum chemistry

AMS subject classifications. 65N15, 65N25, 65N30, 65N50

DOI. 10.1137/06067780X

1. Introduction.

It is significant to obtain the ground state energies in the electronic structure study. In modern electronic structure computations, the ground state energies are usually obtained from the Kohn–Sham equations [7, 19, 25]. Since Kohn–Sham equations are nonlinear eigenvalue systems in three dimensions, an iteration procedure must be involved. More precisely, the central computation in solving Kohn–Sham equations is the repeated solution of the following type of singular eigenvalue problem, posed on a convex polyhedral domain $\Omega \subset \mathbb{R}^3$:

$$\begin{cases} -\Delta u + Vu = \lambda u & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega, \end{cases} \tag{1.1}$$

where $V = V_{ne} + V_0$ is some (density-dependent) effective potential that has finite singular points:

$$V_{ne}(x) = -\sum_{q=1}^{N_{atom}} \frac{Z_q}{|x - r_q|}$$

and V_0 is the remainder of the effective potential with $r_q \in \Omega$ and $Z_q \in (0, \infty)(q = 1, 2, \ldots, N_{atom})$ (see section 4). Because of the singularity of the potential V, the eigenvectors of (1.1) vary very rapidly near the singular points but diffuse further away (cf., e.g., [7]). Thus it is important to design some efficient approximation scheme for solving such a linear singular eigenvalue problem at each iteration step.

*Received by the editors December 15, 2006; accepted for publication (in revised form) July 9, 2007; published electronically January 16, 2008. This work was partially supported by the National Science Foundation of China under grant 10425105 and the National Basic Research Program under grant 2005CB321704.

http://www.siam.org/journals/sinum/46-1/67780.html

†LSEC, Institute of Computational Mathematics and Scientific/Engineering Computing, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, P.O. Box 2719, Beijing 100080, China, and Graduate University of Chinese Academy of Sciences, Beijing 100080, China (daixy@lsec.cc.ac.cn).

‡LSEC, Institute of Computational Mathematics and Scientific/Engineering Computing, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, P.O. Box 2719, Beijing 100080, China (azhou@lsec.cc.ac.cn).

As the finite element method is one of the most effective numerical methods, we shall consider using the finite element scheme to discretize such linear singular eigenvalue problems. Indeed, a significant strength of the finite element method lies in its ability to place adaptive/local refinements in regions where the desired functions vary rapidly while treating the distant zones with a coarser description. We refer the reader to [7, 26, 32] and the references cited therein for the application of the finite element method to electronic structure computations. In this paper, we will propose some new techniques for finite element eigenvalue approximations and then apply these new techniques to ground state energy computations. These techniques are based on our understanding of local behaviors of finite element solutions to some (linear) elliptic problems. By using these techniques, the computational complexity can be resolved through some globally and locally coupled discretizations that can be carried out in parallel. The main idea of our new algorithms is to use a relatively coarse finite element to approximate the low frequency of the solution and then to use some linear algebraic systems to correct the residual (which contains mostly high frequencies) in some globally and locally fine finite element spaces. With these new proposed schemes, solving an elliptic eigenvalue problem will not be much more difficult than the global and local solutions of some standard elliptic boundary value problems. Our schemes are some iterative approaches, which are related to those in [10, 20, 38, 39]. The schemes in the current work, however, are proposed for solving singular eigenvalue problems and based on globally and locally coupled discretizations on different finite element grids with different scales.

The main philosophy behind this paper is that we should treat different phenomena by different tools that are usually posed in different scale settings (cf. [37, 41]). In multigrid approaches, this kind of idea is used to devise iterative methods for solving a given eigenvalue discretization scheme (see, e.g., [7, 8, 12, 16, 21, 24, 35] and the references cited therein); while in our approach, we try to use this idea for designing discretization schemes for quantum eigenvalue problems, which are known to be singular.

Let us now give a little more detailed but informal description of the main ideas in this paper. Let $S_0^H(\Omega), S_0^k(\Omega)$, and $S_0^h(\Omega)$, satisfying $S_0^H(\Omega) \subset S_0^k(\Omega) \subset S_0^h(\Omega) \subset H_0^1(\Omega)$, be three finite element spaces which consist of piecewise polynomials of degree less than or equal to r associated with a coarse grid $T^H(\Omega)$ of mesh size H, a mesoscopic grid $T^k(\Omega)$ of mesh size k, and a fine grid $T^h(\Omega)$ of mesh size h, respectively. If the singular points are located in some subdomain, say $D \subset \Omega$, we may employ the following algorithm to discretize (1.1) to obtain eigenpair approximations (see section 3):

1. Solve an eigenvalue problem on a globally coarse grid (see Figure 1.1 for the two-dimensional version): find $(\lambda_H, u_H) \in S_0^H(\Omega) \times \mathbb{R}$ satisfying $\|u_H\|_{L^2(\Omega)} = 1$ and

$$\int_\Omega (\nabla u_H \cdot \nabla v + V u_H v) = \lambda_H \int_\Omega u_H v \quad \forall v \in S_0^H(\Omega).$$

2. Solve a linear boundary value problem on a globally mesoscopic grid (see Figure 1.2 for the two-dimensional version): find $u^k \in S_0^k(\Omega)$ satisfying

$$\int_\Omega \left(\nabla u^k \cdot \nabla v + V u^k v\right) = \lambda_H \int_\Omega u_H v \quad \forall v \in S_0^k(\Omega).$$

3. Solve a linear boundary value problem on a locally fine grid (see Figure 1.3

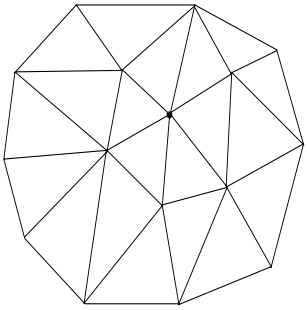


FIG. 1.1. *A coarse grid.*

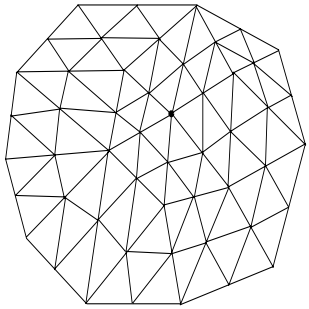


FIG. 1.2. *A mesoscopic grid.*

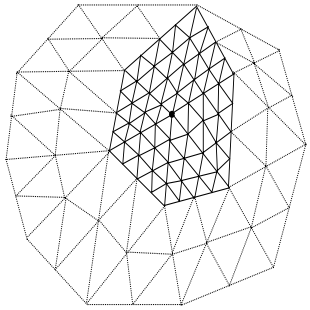


FIG. 1.3. *A locally fine grid.*

for the two-dimensional version): find $e_h \in S_0^h(\Omega_0)$ satisfying

$$\int_{\Omega_0} (\nabla e_h \cdot \nabla v + V e_h v)$$

$$= \lambda_H \int_{\Omega_0} u_H v - \int_{\Omega_0} \left(\nabla u^k \cdot \nabla v + V u^k v\right) \quad \forall v \in S_0^h(\Omega_0).$$

4. Set

$$u^{k,h} = \begin{cases} u^k + e_h & \text{on } \bar{\Omega}_0, \\ u^k & \text{in } \Omega \setminus \bar{\Omega}_0 \end{cases}$$

and

$$\lambda^{k,h} = \frac{\displaystyle\int_\Omega \left(|\nabla u^{k,h}|^2 + V|u^{k,h}|^2\right)}{\displaystyle\int_\Omega |u^{k,h}|^2},$$

where $D \subset\subset \Omega_0$, which means that $D \subset \Omega_0$ and dist $(\partial D \setminus \partial\Omega, \partial\Omega_0 \setminus \partial\Omega) > 0$.

In this algorithm, we use step 1 to handle the nonlinearity (note that the eigenvalue problem is nonlinear). To damp out the singularity (which contains mostly high frequencies), we set up step 3. And we apply step 2 to improve the accuracy of approximation. If, for example, λ_H is the first eigenvalue of the problem at the first step, then under some reasonable assumption we can establish the following results (see section 3):

$$\|u - u^{k,h}\|_{H^1(\Omega)} = \mathcal{O}(h^{r+s-1} + k^r + H^{r+s+\theta-1}) \tag{1.2}$$

and

$$\left|\lambda - \lambda^{k,h}\right| = \mathcal{O}(h^{2r+2s-2} + k^{2r} + H^{2r+2s+2\theta-2}), \tag{1.3}$$

where $s \in [0, 1)$ depends on the local regularity of the exact eigenvector associated with the first eigenvalue λ of (1.1) (more precisely, $u \in H_0^1(\Omega) \cap H^{r+s}(\Omega) \cap H^{r+1}(\Omega \setminus \bar{D})$), and $\theta \in (0, 1]$ depends on s and the regularity of Ω.

This is a very satisfactory result since we may choose $H \gg k \gg h$. As a consequence, for example, we can then design the following local and parallel scheme:

first, solve the eigenvalue problem on a globally coarse grid $T^H(\Omega)$, then solve a linear boundary value problem on a globally mesoscopic grid $T^k(\Omega)$ and some linear boundary value problems on a collection of nonoverlapped subdomains on locally fine grids $T^h(\Omega_j)$ in parallel, and, finally, compute some Rayleigh quotient to obtain a new eigenvalue approximation $\lambda^{k,h}$ which satisfies (1.3) (see section 3).

It is seen that, with this new scheme, the computational complexity is significantly reduced. More precisely, we can largely resolve the computational complexity and keep the same approximation accuracy as that of solving the eigenvalue problem in the globally fine grid. The theoretical tool for motivating this idea is the local error estimates for finite element approximations developed in [37, 39] (see also section 2.2). It is shown by numerical examples that this new scheme is a potentially efficient approach to solving quantum eigenvalue problems (see section 4). To apply the three-scale discretization schemes for ground state energies, which are of nonlinear eigenvalue problems, we need to employ a so-called self-consistent approach [7, 19, 32]. In our computations, a Broyden-type quasi-Newton method [34] will be used.

The rest of the paper is organized as follows. In the next section, we shall describe some basic notation and the finite element approximations to a class of linear second order elliptic boundary value and eigenvalue problems, including some local properties of finite element solutions discovered and studied in [37, 39]. Based on the local properties of finite element solutions and certain globally and locally coupled discretizations, some three-scale discretization schemes for linear eigenvalue approximations are devised and analyzed in section 3. Combining with the self-consistent approach, in section 4, the corresponding three-scale discretization schemes are constructed for Kohn–Sham equations. It is shown by the numerical experiments that our schemes are very efficient in electronic structure computations. Finally, some concluding remarks are given.

2. Preliminaries. In this section, we shall describe some basic notation and the standard finite element discretizations for a class of linear second order elliptic boundary value and eigenvalue problems.

Let Ω be a bounded domain in $\mathbb{R}^d (d \geq 1)$. We shall use the standard notation for Sobolev spaces $W^{s,p}(\Omega)$ and their associated norms and seminorms; see, e.g., [1, 11]. For $p = 2$, we denote $H^s(\Omega) = W^{s,2}(\Omega)$ and $H_0^1(\Omega) = \{v \in H^1(\Omega) : v\mid_{\partial\Omega} = 0\}$, where $v\mid_{\partial\Omega} = 0$ is in the sense of trace, $\|\cdot\|_{s,\Omega} = \|\cdot\|_{s,2,\Omega}$ and $\|\cdot\|_{\Omega} = \|\cdot\|_{0,2,\Omega}$. (In some places in this paper, $\|\cdot\|_{s,2,\Omega}$ should be viewed as piecewise defined if necessary.) The space $H^{-1}(\Omega)$, the dual of $H_0^1(\Omega)$, will also be used.

For $D \subset G \subset \Omega$, we use the notation $D \subset\subset G$ to mean that $\mathrm{dist}(\partial D \setminus \partial\Omega, \partial G \setminus \partial\Omega) > 0$ (see Figure 2.1 for two-dimensional cases). For simplicity, we assume that Ω and its subdomain (if any) involved in this paper are polytopic. Note that any $w \in H_0^1(\Omega_0)$ can be naturally extended to be a function in $H_0^1(\Omega)$ with zero outside of Ω_0 when $\Omega_0 \subset \Omega$; thus we shall state this fact by the slightly abused notation $H_0^1(\Omega_0) \subset H_0^1(\Omega)$.

Throughout this paper, we shall use the letter C (with or without subscripts) to denote a generic positive constant which may stand for different values at its different occurrences. For convenience, the symbol $\lesssim$ will be used in this paper. The notation that $x_1 \lesssim y_1$ means that $x_1 \leq Cy_1$ for some positive constant C that is independent of mesh parameters.

Assume that $T^h(\Omega) = \{\tau\}$ is a mesh of Ω with mesh-size function $h(x)$ whose value is the diameter h_τ of the element τ containing x. One basic assumption on the mesh is that it is not exceedingly overrefined locally; namely, we have the following.

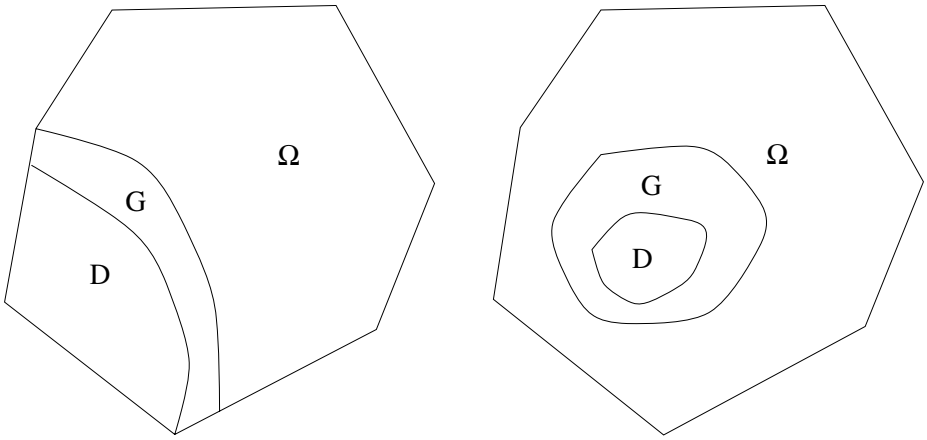


FIG. 2.1. *Subdomains.*

A.0. There exists $\nu \geq 1$ such that

$$h_{\Omega}^{\nu} \lesssim h(x), \quad x \in \Omega, \tag{2.1}$$

where $h_{\Omega} = \max_{x\in\Omega} h(x)$ is the (largest) mesh size of $T^h(\Omega)$.

This is apparently a very mild assumption from the theoretical point of view. Sometimes, we will drop the subscript in h_{Ω} to h for the mesh size on a domain that is clear from the context.

Let $T^h(\Omega)$ consist of shape-regular simplices and define $S^h(\Omega)(\equiv S^{h,r}(\Omega))$ to be a space of continuous functions on Ω such that for $v \in S^h(\Omega), v$ restricted to each τ is a polynomial of total degree $\leq r$, namely

$$S^h(\Omega) = \{v \in C(\bar{\Omega}) : v\,|_{\tau} \in P_{\tau}^r \quad \forall \tau \in T^h(\Omega)\}, \tag{2.2}$$

where P_{τ}^r is the space of polynomials of degree not greater than a positive integer r. Set $S_0^h(\Omega) = S^h(\Omega) \cap H_0^1(\Omega)$. These are Lagrange finite element spaces, and we refer the reader to [37] (see also [28, 29]) for their basic properties that will be used in our analysis.

Given $G \subset \Omega$, we define $S^h(G)$ and $T^h(G)$ to be the restriction of $S^h(\Omega)$ and $T^h(\Omega)$ to G, respectively, and

$$S_h^0(G) = \{v \in S^h(\Omega) : \text{supp } v \subset\subset G\}.$$

For any $G \subset \Omega$ mentioned in this paper, we assume that it aligns with $T^h(\Omega)$ when necessary.

2.1. A linear elliptic boundary value problem. In this subsection, we shall present some basic properties of a second order elliptic boundary value problem and its finite element approximations, which will be used in this paper.

Let $(\cdot,\cdot)$ be the standard inner product of $L^2(\Omega)$ and

$$a(u,v) = \int_{\Omega} \sum_{i,j=1}^{d} a_{ij} \frac{\partial u}{\partial x_i} \frac{\partial v}{\partial x_j} + cuv,$$

where (a_{ij}) is symmetric with $a_{ij} \in L^\infty(\Omega)(i,j = 1,2,\ldots,d)$ and c is some real function such that $a(\cdot,\cdot)$ is a bounded bilinear form in $H_0^1(\Omega)$ and

$$\|w\|_{1,\Omega}^2 - C^{-1}\|w\|_{0,\Omega}^2 \lesssim a(w,w) \quad \forall w \in H_0^1(\Omega)$$

holds for some positive constant C.

Consider the homogeneous boundary value problem

$$\begin{cases} Lu = f \ \text{ in } \ \Omega, \\ u = 0 \ \text{ on } \ \partial\Omega, \end{cases} \tag{2.3}$$

where L is a linear second order elliptic operator:

$$Lu = -\sum_{i,j=1}^{d} \frac{\partial}{\partial x_j}\left(a_{ij}\frac{\partial u}{\partial x_i}\right) + cu.$$

The weak form of (2.3) is as follows: Find $u \equiv L^{-1}f \in H_0^1(\Omega)$ such that

$$a(u,v) = (f,v) \quad \forall v \in H_0^1(\Omega). \tag{2.4}$$

Our basic assumption is that (2.4) is well posed; namely, (2.4) is uniquely solvable for any $f \in H^{-1}(\Omega)$. We have (cf. [15]) the following estimate for the regularity of the solution of (2.3) or (2.4): $u \in H_0^1(\Omega) \cap H^{1+\gamma}(\Omega)$ and

$$\|u\|_{1+\gamma,\Omega} \lesssim \|f\|_{-1+\gamma,\Omega} \tag{2.5}$$

for some $\gamma \in (0,1]$ depending on Ω and the coefficients of L.

For some $G \subset \Omega$, we need the following assumption.

R(G). For any $f \in L^2(G)$, there exists a $u \in H_0^1(G) \cap H^{1+\gamma}(G)$ satisfying

$$a(v,u) = (f,v) \quad \forall v \in H_0^1(G)$$

and

$$\|u\|_{1+\gamma,G} \lesssim \|f\|_{-1+\gamma,G}.$$

Define a Galerkin projection $P_h : H_0^1(\Omega) \mapsto S_0^h(\Omega)$ by

$$a(u - P_h u, v) = 0 \quad \forall v \in S_0^h(\Omega) \tag{2.6}$$

and apparently

$$\|P_h u\|_{1,\Omega} \lesssim \|u\|_{1,\Omega} \quad \forall u \in H_0^1(\Omega). \tag{2.7}$$

From (2.7), various global a priori error estimates for finite element approximation $P_h u$ to u can be obtained from the approximate properties of the finite element space $S^h(\Omega)$ (cf. [11, 37]).

Now we introduce the following quantity:

$$\rho_\Omega(h) = \sup_{f \in L^2(\Omega), \|f\|_{0,\Omega}=1} \inf_{v \in S_0^h(\Omega)} \|L^{-1}f - v\|_{1,\Omega}.$$

Then (see, e.g., [37]) $\rho_\Omega(h) \to 0$ as $h_\Omega \to 0$ and

$$\rho_\Omega(h) \lesssim h_\Omega^\gamma. \tag{2.8}$$

Similarly, if Assumption R(G) holds, we can define $\rho_G(h)$ well, too.

The following results can be found in [37] (cf. also [5, 30, 38]).

PROPOSITION 2.1.

$$\|(I - P_h)L^{-1}f\|_{1,\Omega} \lesssim \rho_\Omega(h)\|f\|_{0,\Omega} \quad \forall f \in L^2(\Omega) \tag{2.9}$$

and

$$\|u - P_h u\|_{0,\Omega} \lesssim \rho_\Omega(h)\|u - P_h u\|_{1,\Omega} \quad \forall u \in H_0^1(\Omega). \tag{2.10}$$

Local error estimates for finite element approximations can also be obtained (see, e.g., [11, 28, 29, 37]). For instance, we have the following a priori error estimate (see [37] for details), which will be employed in our discussions.

PROPOSITION 2.2. *Suppose that* $D \subset\subset \Omega_0$.

(i) *If* $f \in H^{-1}(\Omega)$ *and* $w \in S^h(\Omega_0)$ *satisfies*

$$a(w, v) = f(v) \quad \forall v \in S_h^0(\Omega_0), \tag{2.11}$$

then

$$\|w\|_{1,D} \lesssim \|w\|_{0,\Omega_0} + \|f\|_{-1,\Omega_0}. \tag{2.12}$$

(ii) *If* $u \in H_0^1(\Omega)$, *then*

$$\|u - P_h u\|_{1,D} \lesssim \inf_{v \in S_0^h(\Omega)} \|u - v\|_{1,\Omega_0} + \rho_\Omega(h)\|u - P_h u\|_{1,\Omega}. \tag{2.13}$$

2.2. A linear elliptic eigenvalue problem. A number λ is called an eigenvalue of the form $a(\cdot,\cdot)$ relative to the form $(\cdot,\cdot)$ if there is a nonzero vector $u \in H_0^1(\Omega)$, called an associated eigenvector, satisfying

$$a(u, v) = \lambda(u, v) \quad \forall v \in H_0^1(\Omega). \tag{2.14}$$

It is known that (2.14) has a countable sequence of real eigenvalues

$$\lambda_1 < \lambda_2 \le \lambda_3 \le \cdots$$

and the corresponding eigenvectors

$$u_1, u_2, u_3, \ldots,$$

which can be assumed to satisfy

$$(u_i, u_j) = \delta_{ij}, \quad i, j = 1, 2, \ldots.$$

In the sequence $\{\lambda_j\}$, the λ_j's are repeated according to geometric multiplicity.

The following straightforward property of eigenvalue and eigenvector is useful (see, e.g., [5, 6] or [38]).

PROPOSITION 2.3. *Let* (λ, u) *be an eigenpair of* (2.14). *For any* $w \in H_0^1(\Omega) \setminus \{0\}$,

$$\frac{a(w,w)}{(w,w)} - \lambda = \frac{a(w-u,w-u)}{(w,w)} - \lambda \frac{(w-u,w-u)}{(w,w)}. \tag{2.15}$$

A standard finite element scheme for (2.14) is the following: find a pair of (λ_h, u_h), where λ_h is a number and $0 \neq u_h \in S_0^h(\Omega)$, satisfying

$$a(u_h, v) = \lambda_h (u_h, v) \quad \forall v \in S_0^h(\Omega), \tag{2.16}$$

and use λ_h and u_h as approximations to λ and u (as $h \to 0$), respectively. One sees that (2.16) has a finite sequence of eigenvalues

$$\lambda_{1,h} < \lambda_{2,h} \leq \cdots \leq \lambda_{n_h,h}, \quad n_h = \dim\, S_0^h(\Omega),$$

and the corresponding eigenvectors

$$u_{1,h}, u_{2,h}, \ldots, u_{n_h,h},$$

which can be assumed to satisfy

$$(u_{i,h}, u_{j,h}) = \delta_{ij}, \quad i,j = 1,2,\ldots.$$

Set

$$M(\lambda_i) = \{w \in H_0^1(\Omega) : w \text{ is an eigenvector of (2.14) corresponding to } \lambda_i\}$$

and

$$\delta_h(\lambda_i) = \sup_{w \in M(\lambda_i), \|w\|_{0,\Omega}=1} \inf_{v \in S_0^h(\Omega)} \|w - v\|_{1,\Omega}.$$

The following results are standard and can be found in the literature (see, e.g., [5, 6, 10] or [38]).

PROPOSITION 2.4. (i) *For any eigenvector* $u_{i,h}$ *of* (2.16) $(i = 1,2,\ldots,n_h)$ *with* $\|u_{i,h}\|_{0,\Omega} = 1$, *there is an eigenvector* u^i *of* (2.14) *corresponding to* λ_i *satisfying* $\|u^i\|_{0,\Omega} = 1$ *and*

$$\|u^i - u_{i,h}\|_{1,\Omega} \leq C_i \delta_h(\lambda_i). \tag{2.17}$$

Moreover,

$$\|u^i - u_{i,h}\|_{0,\Omega} \leq C_i \rho_\Omega(h) \|u^i - u_{i,h}\|_{1,\Omega}. \tag{2.18}$$

(ii) *For eigenvalues,*

$$\lambda_i \leq \lambda_{i,h} \leq \lambda_i + C_i \delta_h^2(\lambda_i), \quad i = 1,2,\ldots,n_h. \tag{2.19}$$

Here C_i *is some positive constant depending on* i *but not on the mesh parameter* h.

For the eigenvectors, there is a local a priori estimate which was presented in [39].

THEOREM 2.5. *If* $D \subset\subset \Omega_0$, *then*

$$\|u^i - u_{i,h}\|_{1,D} \lesssim \inf_{v \in S_0^h(\Omega)} \|u^i - v\|_{1,\Omega_0} + \lambda_{i,h} - \lambda_i + \lambda_i \rho_\Omega(h) \|u^i - u_{i,h}\|_{1,\Omega}. \tag{2.20}$$

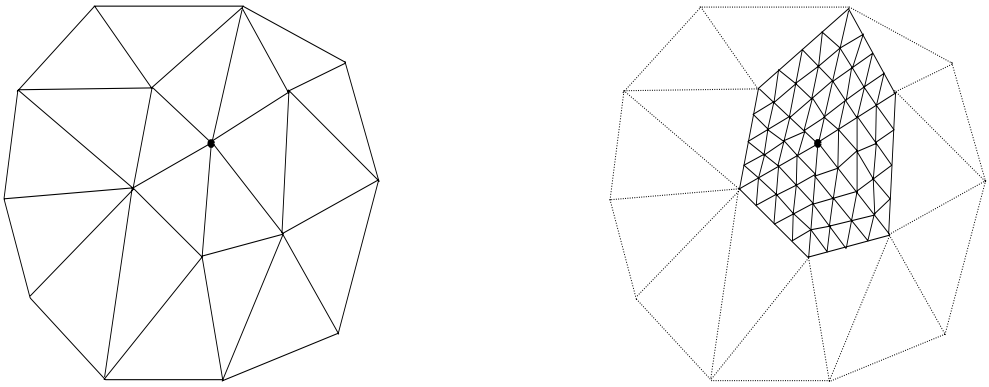


FIG. 3.1. *Localization.*

3. Discretization schemes for linear eigenvalue problems. In this section we shall present and analyze a three-scale discretization scheme for solving linear elliptic eigenvalue problems, based on some global and local corrections on different grids that are of different scales. The parallel version of the three-scale scheme is proposed, too. The schemes are motivated by the observation that for a solution of some elliptic problem, low frequency components can be approximated well on a relatively coarse grid and high frequency ones can be computed on some globally mesoscopic grid and locally fine grids. Although these schemes are theoretically interesting in their own right for linear eigenvalue problems, our main motivation is to use them to devise efficient and accurate numerical methods for quantum eigenvalue problems; see section 4.

Let $T^H(\Omega)$ be a shape-regular grid of size $H \in (0,1)$, $D \subset \Omega$ be a subdomain which contains a singular point, and Ω_0 be a slightly larger subdomain containing D (namely $D \subset\subset \Omega_0$). Let $h \ll H$ and a highly locally refined grid $T^h(\Omega_0)$ be obtained from $T^H(\Omega)$; see Figure 3.1. Let $S_0^H(\Omega)(\equiv S_0^{H,r}(\Omega))$ denote the finite element space which consists of piecewise polynomials of degree less than or equal to r associated with $T^H(\Omega)$. In our discussion, we shall use an auxiliary fine grid, say $T^h(\Omega)$, that is globally defined. Associated with $T^h(\Omega)$, the finite element space $S_0^h(\Omega)(\equiv S_0^{h,r}(\Omega))$ which consists of piecewise polynomials of degree less than or equal to r is then defined. One basic assumption for this auxiliary fine finite element space is that it should coincide with the locally fine finite element space in the subdomain of interest. To improve the efficiency, we also need to apply the third finite element space $S_0^k(\Omega)(\equiv S_0^{k,r}(\Omega))$ which consists of piecewise polynomials of degree less than or equal to r associated with a shape-regular grid $T^k(\Omega)$ of a mesoscopic size k. All these spaces satisfy $S_0^H(\Omega) \subset S_0^k(\Omega) \subset S_0^h(\Omega) \subset H_0^1(\Omega)$.

We consider the approximation of any eigenvalue λ of (2.14). Here and hereafter we let λ_H be the finite element eigenvalue of (2.16) corresponding to $S_0^H(\Omega)$ and satisfying

$$|\lambda_H - \lambda| \lesssim \delta_H(\lambda). \tag{3.1}$$

3.1. A two-scale discretization. Before we propose and analyze the three-scale discretization scheme, we shall discuss a simpler scheme—a so-called two-scale discretization scheme. The two-scale discretization scheme we shall now present can be used to obtain an approximate solution mostly by a local computation. The main

idea is that the more global component of a finite element eigenvector may be obtained in a relatively coarse finite element space, and the rest of the computation can then be localized using a locally fine finite element discretization. This scheme may be viewed as a postprocessing of the standard finite element discretization when the eigenvectors are known to be oscillating in some local domain.

A prototype of our two-scale discretization, which is mainly a local computational scheme, is stated as follows.

ALGORITHM A0.

1. *Solve an eigenvalue problem on a globally coarse grid: find* $(\lambda_H, u_H) \in S_0^H(\Omega) \times \mathbb{R}$ *satisfying* $\|u_H\|_{0,\Omega} = 1$ *and*

$$a(u_H, v) = \lambda_H(u_H, v) \quad \forall v \in S_0^H(\Omega).$$

2. *Solve a linear boundary value problem on a locally fine grid: find* $e_h \in S_0^h(\Omega_0)$ *satisfying*

$$a(e^h, v) = \lambda_H(u_H, v) - a(u_H, v) \quad \forall v \in S_0^h(\Omega_0).$$

3. *Set*

$$u^h = \begin{cases} u_H + e^h & \text{on } \bar{\Omega}_0, \\ u_H & \text{in } \Omega \setminus \bar{\Omega}_0 \end{cases}$$

and

$$\lambda^h = \frac{a(u^h, u^h)}{\| u^h \|_{0,\Omega}^2}.$$

It is seen from Proposition 2.4 that associated with the eigenvector u_H obtained by step 1 in Algorithm A0, there exists an exact eigenvector u of (2.14) satisfying $\|u\|_{0,\Omega} = 1$ and

$$\|u - u_H\|_{1,\Omega} \lesssim \delta_H(\lambda), \quad \|u - u_H\|_{0,\Omega} \lesssim \rho_\Omega(H)\delta_H(\lambda). \tag{3.2}$$

LEMMA 3.1. *Let* $\sigma \in \{H, k, h\}$ *and* $G \subset \Omega$. *If* $M(\lambda) \subset H^{r+s}(G)(0 \le s \le 1)$, *then*

$$\sup_{w \in M(\lambda), \|w\|_{0,\Omega}=1} \inf_{v \in S_0^\sigma(G)} \|w - v\|_{1,G} \lesssim \sigma^{r+s-1}. \tag{3.3}$$

In particular, if $M(\lambda) \subset H^{r+s}(\Omega)(0 \le s \le 1)$, *then*

$$|\lambda - \lambda_H| \lesssim H^{2r+2s-2}, \quad \|u - u_H\|_{1,\Omega} \lesssim H^{r+s-1}, \quad \|u - u_H\|_{0,\Omega} \lesssim H^{\gamma+r+s-1} \tag{3.4}$$

and

$$\|u - P_\sigma u\|_{1,\Omega} \lesssim \sigma^{r+s-1}, \quad \|u - P_\sigma u\|_{0,\Omega} \lesssim \sigma^{\gamma+r+s-1}. \tag{3.5}$$

Proof. Since Riesz–Schauder theory implies that $\dim M(\lambda) < \infty$, we immediately have

$$\sup_{w \in M(\lambda), \|w\|_{0,\Omega}=1} \|w\|_{r+s,G} < \infty,$$

from which we get (3.3). This completes the proof. ∎

In the following discussion, we assume that (λ, u) is the exact eigenpair of (2.14) satisfying (3.1), (3.2), and $M(\lambda) \subset H^{r+s}(\Omega)(1 < r+s$ and $0 \le s < 1)$. For simplicity, we set $\theta = \min(\gamma, r+s-1)$. Obviously, $\theta \in (0,1]$.

THEOREM 3.2. *Assume that u^h is obtained by Algorithm* A0 *and assumption $R(\Omega_0)$ holds. If $M(\lambda) \subset H_0^1(\Omega) \cap H^{r+s}(\Omega) \cap H^{r+1}(\Omega \setminus \bar{D})(1 < r+s$ and $0 \le s < 1)$, then*

$$\|u - u^h\|_{1,\Omega} \lesssim h^{r+s-1} + H^r + H^{r+s+\theta-1} \tag{3.6}$$

and

$$|\lambda - \lambda^h| \lesssim h^{2r+2s-2} + H^{2r} + H^{2r+2s+2\theta-2}. \tag{3.7}$$

Proof. Because of Lemma 3.1 and Proposition 2.3 it is sufficient to prove that

$$\|P_h u - u^h\|_{1,\Omega} \lesssim H^r + H^{r+s+\theta-1}. \tag{3.8}$$

We divide the proof of (3.8) into four steps. Let $G \subset \Omega$ satisfy $D \subset\subset G \subset\subset \Omega_0$. Due to the triangle inequality

$$\|P_h u - u^h\|_{1,\Omega} \lesssim \|P_h u - u^h\|_{1,D} + \|P_h u - u^h\|_{1,G\setminus\bar{D}} + \|P_h u - u^h\|_{1,\Omega\setminus\bar{G}},$$

we shall estimate $\|P_h u - u^h\|_{1,D}$, $\|P_h u - u^h\|_{1,G\setminus\bar{D}}$, and $\|P_h u - u^h\|_{1,\Omega\setminus\bar{G}}$, respectively.

First, we proceed to estimate $\|P_h u - u^h\|_{1,D}$. Note that the following equation can be derived from the construction of Algorithm A0:

$$a(u^h - P_h u, v) = \lambda_H(u_H, v) - \lambda(u, v) \quad \forall v \in S_0^h(\Omega_0). \tag{3.9}$$

Hence, from the identity

$$\lambda_H(u_H, v) - \lambda(u, v) = (\lambda_H - \lambda)(u, v) + \lambda_H(u_H - u, v) \ \ \forall v \in H_0^1(\Omega), \tag{3.10}$$

we obtain

$$a(u^h - P_h u, v) = (\lambda_H - \lambda)(u, v) + \lambda_H(u_H - u, v) \quad \forall v \in S_0^h(\Omega_0). \tag{3.11}$$

Since $S_h^0(\Omega_0) \subset S_0^h(\Omega)$, the estimation

$$\|P_h u - u^h\|_{1,D} \lesssim |\lambda - \lambda_H| + \|u - u_H\|_{0,\Omega_0} + \|u^h - P_h u\|_{0,\Omega_0}$$

or

$$\|P_h u - u^h\|_{1,D} \lesssim |\lambda - \lambda_H| + \|u - u_H\|_{0,\Omega_0} + \|u - P_h u\|_{0,\Omega_0} + \|e_h\|_{0,\Omega_0}$$

can be obtained from (3.11) and Proposition 2.2.

We now need to estimate $\|e^h\|_{0,\Omega_0}$ by using the Aubin–Nitsche duality argument. Given any $\phi \in L^2(\Omega_0)$, there exists $w \in H_0^1(\Omega_0)$ such that

$$a(v, w) = (\phi, v) \quad \forall v \in H_0^1(\Omega_0). \tag{3.12}$$

Let $w_h^0 \in S_0^h(\Omega_0)$ and $w_H^0 \in S_0^H(\Omega_0)$ satisfy

$$a(v, w_h^0) = a(v, w) \quad \forall v \in S_0^h(\Omega_0), \tag{3.13}$$

$$a(v, w_H^0) = a(v, w) \quad \forall v \in S_0^H(\Omega_0). \tag{3.14}$$

Then, Proposition 2.1 implies

$$\|w - w_h^0\|_{1,\Omega_0} \lesssim h^\gamma \|\phi\|_{0,\Omega_0}, \quad \|w - w_H^0\|_{1,\Omega_0} \lesssim H^\gamma \|\phi\|_{0,\Omega_0}.$$

It follows from the definitions of w, e_h, and w_h^0 that

$$\begin{aligned}(e^h, \phi) &= a(e^h, w) = a(e^h, w_h^0) = a(u^h - u_H, w_h^0)\\ &= a(P_h u - u_H, w_h^0) + a(u^h, w_h^0) - a(P_h u, w_h^0)\\ &= a(P_h u - u_H, w_h^0) + \lambda_H(u_H, w_h^0) - \lambda(u, w_h^0).\end{aligned}$$

Thus, we obtain

$$\begin{aligned}(e^h, \phi) &= a(P_h u - u_H, w_h^0 - w) + a(P_h u - u_H, w - w_H^0)\\ &\quad + a(P_h u - u_H, w_H^0) + \lambda_H(u_H, w_h^0) - \lambda(u, w_h^0)\\ &= a(P_h u - u_H, w_h^0 - w) + a(P_h u - u_H, w - w_H^0) + \lambda(u, w_H^0)\\ &\quad - \lambda_H(u_H, w_H^0) + \lambda_H(u_H, w_h^0) - \lambda(u, w_h^0).\end{aligned}$$

By some further simple operations, we then arrive at

$$\begin{aligned}(e^h, \phi) &= a(P_h u - u_H, w_h^0 - w) + a(P_h u - u_H, w - w_H^0)\\ &\quad + \lambda(u, w_H^0 - w_h^0) - \lambda_H(u_H, w_H^0 - w_h^0)\\ &= a(P_h u - u_H, w_h^0 - w) + a(P_h u - u_H, w - w_H^0)\\ &\quad + (\lambda - \lambda_H)(u, w_H^0 - w_h^0) + \lambda_H(u - u_H, w_H^0 - w_h^0).\end{aligned}$$

As a result, we get the estimation for any $\phi \in L^2(\Omega)$ that

$$|(e^h, \phi)| \lesssim (H^\gamma \|P_h u - u_H\|_{1,\Omega} + |\lambda - \lambda_H| + \|u - u_H\|_{0,\Omega_0}) \|\phi\|_{0,\Omega_0},$$

which implies

$$\|e^h\|_{0,\Omega_0} \lesssim |\lambda - \lambda_H| + \|u - u_H\|_{0,\Omega_0} + H^\gamma \|P_h u - u_H\|_{1,\Omega}. \tag{3.15}$$

Therefore

$$\|P_h u - u^h\|_{1,D} \lesssim |\lambda - \lambda_H| + \|u - u_H\|_{0,\Omega} + \|u - P_h u\|_{0,\Omega} + H^\gamma \|P_h u - u_H\|_{1,\Omega},$$

which together with Lemma 3.1 yields

$$\|P_h u - u^h\|_{1,D} \lesssim H^{r+s+\theta-1}. \tag{3.16}$$

Second, we turn to estimate $\|P_h u - u^h\|_{1,G\setminus\bar{D}}$. Note that $(G \setminus \bar{D}) \subset\subset \Omega_0$; hence the estimation

$$\|P_h u - u^h\|_{1,G\setminus\bar{D}} \lesssim |\lambda - \lambda_H| + \|u - u_H\|_{0,\Omega} + \|u - P_h u\|_{0,\Omega} + H^\gamma \|P_h u - u_H\|_{1,\Omega}$$

can be derived from (3.9) and Proposition 2.2, too. Consequently, we have

$$\|P_h u - u^h\|_{1,G\setminus\bar{D}} \lesssim H^{r+s+\theta-1}. \tag{3.17}$$

Third, let us investigate $\|P_h u - u^h\|_{1,\Omega\setminus\bar{G}}$. It is seen from the construction of Algorithm A0 that

$$\|P_h u - u^h\|_{1,\Omega\setminus\bar{\Omega}_0} = \|P_h u - u_H\|_{1,\Omega\setminus\bar{\Omega}_0},$$

and we may estimate $\|P_h u - u^h\|_{1,\Omega\setminus\bar{G}}$ as follows:

$$\begin{aligned}\|P_h u - u^h\|_{1,\Omega\setminus\bar{G}} &\le \|P_h u - u_H\|_{1,\Omega\setminus\bar{\Omega}_0} + \|P_h u - u_H\|_{1,\Omega_0\setminus\bar{G}} + \|e^h\|_{1,\Omega_0\setminus\bar{G}}\\ &\lesssim \|P_h u - u_H\|_{1,\Omega\setminus\bar{G}} + \|e^h\|_{1,\Omega_0\setminus\bar{G}}. \qquad (3.18)\end{aligned}$$

To analyze $\|e^h\|_{1,\Omega_0\setminus\bar{G}}$, we need the following equation:

$$a(e^h, v) = \lambda_H(u_H, v) - \lambda(u, v) - a(u_H - u, v) \quad \forall v \in S_0^h(\Omega_0), \qquad (3.19)$$

which is nothing but a variant of (3.9). Thus, from (3.10) and Proposition 2.2, we obtain

$$\|e^h\|_{1,\Omega_0\setminus\bar{G}} \lesssim \|e^h\|_{0,\Omega_0\setminus\bar{F}} + |\lambda - \lambda_H| + \|u - u_H\|_{1,\Omega_0\setminus\bar{F}}, \qquad (3.20)$$

where $F \subset \Omega$ satisfies $D \subset\subset F \subset\subset G$. Combining (3.18) and (3.20), we have

$$\|P_h u - u^h\|_{1,\Omega\setminus\bar{G}} \lesssim \|P_h u - u_H\|_{1,\Omega\setminus\bar{G}} + \|e^h\|_{0,\Omega_0\setminus\bar{F}} + |\lambda - \lambda_H| + \|u - u_H\|_{1,\Omega_0\setminus\bar{F}}$$

or

$$\|P_h u - u^h\|_{1,\Omega\setminus\bar{G}} \lesssim \|e^h\|_{0,\Omega_0} + |\lambda - \lambda_H| + \|u - u_H\|_{1,\Omega\setminus\bar{F}} + \|P_h u - u\|_{1,\Omega\setminus\bar{F}}. \qquad (3.21)$$

Note that $(\Omega \setminus \bar{F}) \subset\subset (\Omega \setminus \bar{D})$, and we may estimate by using Proposition 2.2 and Theorem 2.5 that

$$\begin{aligned}&\|u - P_h u\|_{1,\Omega\setminus\bar{F}} + \|u - u_H\|_{1,\Omega\setminus\bar{F}}\\ &\lesssim \inf_{v\in S_0^h(\Omega)} \|u - v\|_{1,\Omega\setminus\bar{D}} + \rho_\Omega(h)\|u - P_h u\|_{1,\Omega}\\ &\quad + \inf_{v\in S_0^H(\Omega)} \|u - v\|_{1,\Omega\setminus\bar{D}} + |\lambda - \lambda_H| + \rho_\Omega(H)\|u - u_H\|_{1,\Omega}\\ &\lesssim \inf_{v\in S_0^H(\Omega)} \|u - v\|_{1,\Omega\setminus\bar{D}} + |\lambda - \lambda_H| + \rho_\Omega(H)\|u - u_H\|_{1,\Omega} + \rho_\Omega(h)\|u - P_h u\|_{1,\Omega},\end{aligned}$$

which, together with (3.15), (3.21), and Lemma 3.1, leads to

$$\|P_h u - u^h\|_{1,\Omega\setminus\bar{G}} \lesssim H^r + H^{r+s+\theta-1}. \qquad (3.22)$$

Finally, the desired result then follows from (3.22), (3.16), and (3.17). ∎

3.2. A three-scale discretization. It is shown by the analysis above and the numerics in section 4 that the two-scale scheme may not be very efficient. It will be shown in this subsection that the efficiency can be improved if one more scale discretization is coupled. Let $T^k(\Omega)$ be a refined mesoscopic shape-regular grid (from $T^H(\Omega)$) and $T^h(\Omega_0)$ a locally refined grid (from $T^k(\Omega)$) that satisfy $h \ll k \ll H$; see Figure 3.2.

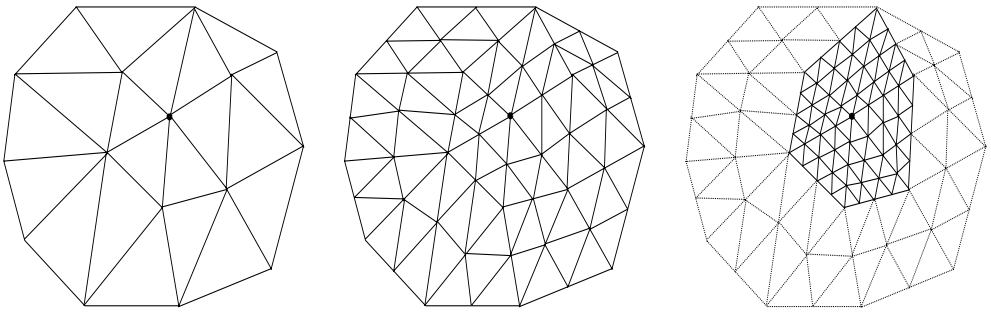


FIG. 3.2. *Globally mesoscopic grid + locally fine grid.*

ALGORITHM B0.

1. *Solve* (2.14) *on a globally coarse grid: find* $(\lambda_H, u_H) \in S_0^H(\Omega) \times \mathbb{R}$ *satisfying* $\|u_H\|_{0,\Omega} = 1$ *and*

$$a(u_H, v) = \lambda_H(u_H, v) \quad \forall v \in S_0^H(\Omega).$$

2. *Solve a linear boundary value problem on a globally mesoscopic grid: find* $u^k \in S_0^k(\Omega)$ *satisfying*

$$a(u^k, v) = \lambda_H(u_H, v) \quad \forall v \in S_0^k(\Omega).$$

3. *Solve a linear boundary value problem on a locally fine gird: find* $e^h \in S_0^h(\Omega_0)$ *satisfying*

$$a(e^h, v) = \lambda_H(u_H, v) - a(u^k, v) \quad \forall v \in S_0^h(\Omega_0).$$

4. *Set*

$$u^{k,h} = \begin{cases} u^k + e^h & \text{on } \bar{\Omega}_0, \\ u^k & \text{in } \Omega \setminus \bar{\Omega}_0 \end{cases}$$

and

$$\lambda^{k,h} = \frac{a(u^{k,h}, u^{k,h})}{\|u^{k,h}\|_{0,\Omega}^2}.$$

The above algorithm is in a way an improvement of the relevant scheme proposed in [20, 38] when the eigenvectors are oscillating in some local domain.

THEOREM 3.3. *Assume that* $u^{k,h}$ *is obtained by Algorithm* B0 *and assumption* $R(\Omega_0)$ *holds. If* $M(\lambda) \subset H_0^1(\Omega) \cap H^{r+s}(\Omega) \cap H^{r+1}(\Omega \setminus \bar{D})(1 < r + s$ *and* $0 \le s < 1)$*, then*

$$\|u - u^{k,h}\|_{1,\Omega} \lesssim h^{r+s-1} + k^r + H^{r+s+\theta-1}$$

and

$$|\lambda - \lambda^{k,h}| \lesssim h^{2r+2s-2} + k^{2r} + H^{2r+2s+2\theta-2}.$$

Proof. Following the proof of Theorem 3.2, we choose $G \subset \Omega$ satisfying $D \subset\subset G \subset\subset \Omega_0$. Since $\|P_h u - u^{k,h}\|_{1,\Omega}$ is bounded by

$$\|P_h u - u^{k,h}\|_{1,D} + \|P_h u - u^{k,h}\|_{1,G\setminus\bar{D}} + \|P_h u - u^{k,h}\|_{1,\Omega\setminus\bar{G}},$$

we need only estimate $\|P_h u - u^{k,h}\|_{1,D}$, $\|P_h u - u^{k,h}\|_{1,G\setminus\bar{D}}$, and $\|P_h u - u^{k,h}\|_{1,\Omega\setminus\bar{G}}$, respectively. For these goals, we take $F \subset \Omega$ such that $D \subset\subset F \subset\subset G \subset\subset \Omega_0$.

First of all, from the equation

$$a(P_k u - u^k, v) = \lambda(u, v) - \lambda_H(u_H, v) \quad \forall v \in S_0^k(\Omega),$$

we have

$$\|P_k u - u^k\|_{1,\Omega} \lesssim |\lambda - \lambda_H| + \|u - u_H\|_{0,\Omega}, \tag{3.23}$$

which will be used in the following analysis.

To estimate $\|P_h u - u^{k,h}\|_{1,D}$, we need the identity

$$a(u^{k,h} - P_h u, v) = \lambda_H(u_H, v) - \lambda(u, v) \quad \forall v \in S_0^h(\Omega_0), \tag{3.24}$$

which comes from the construction of Algorithm B0. It is seen that the estimation

$$\|P_h u - u^{k,h}\|_{1,D} \lesssim |\lambda - \lambda_H| + \|u - u_H\|_{0,\Omega_0} + \|P_h u - u^{k,h}\|_{0,\Omega_0} \tag{3.25}$$

can be derived from (3.10), (3.24), and Proposition 2.2. Due to the definition of $u^{k,h}$, we may estimate $\|P_h u - u^{k,h}\|_{0,\Omega_0}$ as follows:

$$\begin{aligned}\|P_h u - u^{k,h}\|_{0,\Omega_0} &\le \|P_h u - u^k\|_{0,\Omega_0} + \|e^h\|_{0,\Omega_0} \\ &\le \|P_h u - P_k u\|_{0,\Omega_0} + \|P_k u - u^k\|_{0,\Omega_0} + \|e^h\|_{0,\Omega_0}.\end{aligned}$$

Thus combining (3.23) and (3.25), we obtain

$$\|P_h u - u^{k,h}\|_{1,D} \lesssim |\lambda - \lambda_H| + \|u - u_H\|_{0,\Omega} + \|P_h u - P_k u\|_{0,\Omega} + \|e^h\|_{0,\Omega_0}. \tag{3.26}$$

Now we are going to estimate $\|e^h\|_{0,\Omega_0}$ by using the Aubin–Nitsche duality argument. Given any $\phi \in L^2(\Omega_0)$, there exists $w \in H_0^1(\Omega_0)$ satisfying (3.12). Let $w_h^0 \in S_0^h(\Omega_0)$ and $w_H^0 \in S_0^H(\Omega_0)$ be the solutions of (3.13) and (3.14), respectively. Similar to that in the proof of Theorem 3.2, it follows from the definitions of w, e_h, and w_h^0 that

$$\begin{aligned}(e^h, \phi) = {} & a(P_h u - u^k, w_h^0 - w) + a(P_h u - u^k, w - w_H^0) \\ & + (\lambda - \lambda_H)(u, w_H^0 - w_h^0) + \lambda_H(u - u_H, w_H^0 - w_h^0).\end{aligned}$$

As a result, we get the estimation for any $\phi \in L^2(\Omega)$ that

$$|(e^h, \phi)| \lesssim (H^\gamma \|P_h u - u^k\|_{1,\Omega} + |\lambda - \lambda_H| + \|u - u_H\|_{0,\Omega})\|\phi\|_{0,\Omega_0},$$

which leads to

$$\|e^h\|_{0,\Omega_0} \lesssim |\lambda - \lambda_H| + \|u - u_H\|_{0,\Omega} + H^\gamma \|P_h u - u^k\|_{1,\Omega}.$$

Hence taking (3.23) and the triangle inequality

$$\|P_h u - u^k\|_{1,\Omega} \le \|P_h u - P_k u\|_{1,\Omega} + \|P_k u - u^k\|_{1,\Omega}$$

into account, we then obtain

$$\|e^h\|_{0,\Omega_0} \lesssim |\lambda - \lambda_H| + \|u - u_H\|_{0,\Omega} + H^\gamma \|P_h u - P_k u\|_{1,\Omega}. \tag{3.27}$$

It is not difficult to see that the estimation

$$\|P_h u - P_k u\|_{0,\Omega} \lesssim H^{\gamma} \|P_h u - P_k u\|_{1,\Omega}$$

can also be derived from the Aubin–Nitsche duality argument. Therefore we have

$$\|P_h u - u^{k,h}\|_{1,D} \lesssim |\lambda - \lambda_H| + \|u - u_H\|_{0,\Omega} + H^{\gamma} \|P_h u - P_k u\|_{1,\Omega},$$

which together with Lemma 3.1 and (3.23) implies

$$\|P_h u - u^{k,h}\|_{1,D} \lesssim H^{r+s+\theta-1}. \tag{3.28}$$

Similarly, the estimation

$$\|P_h u - u^{k,h}\|_{1,G\setminus\bar{D}} \lesssim H^{r+s+\theta-1} \tag{3.29}$$

can be obtained from $(G \setminus \bar{D}) \subset\subset \Omega_0$ and Proposition 2.2, too.

The remainder is to analyze $\|P_h u - u^{k,h}\|_{1,\Omega\setminus\bar{G}}$. From the construction of Algorithm B0, we see that

$$\|P_h u - u^{k,h}\|_{1,\Omega\setminus\bar{\Omega}_0} = \|P_h u - u^{k}\|_{1,\Omega\setminus\bar{\Omega}_0},$$

which leads to

$$\begin{aligned}\|P_h u - u^{k,h}\|_{1,\Omega\setminus\bar{G}} &\le \|P_h u - u^{k}\|_{1,\Omega\setminus\bar{\Omega}_0} + \|P_h u - u^{k}\|_{1,\Omega_0\setminus\bar{G}} + \|e^h\|_{1,\Omega_0\setminus\bar{G}} \\ &\lesssim \|P_h u - u^{k}\|_{1,\Omega\setminus\bar{G}} + \|e^h\|_{1,\Omega_0\setminus\bar{G}}.\end{aligned}$$

Hence we have

$$\|P_h u - u^{k,h}\|_{1,\Omega\setminus\bar{G}} \lesssim \|P_h u - u\|_{1,\Omega\setminus\bar{G}} + \|u - u^{k}\|_{1,\Omega\setminus\bar{G}} + \|e^h\|_{1,\Omega_0\setminus\bar{G}}. \tag{3.30}$$

Note that the estimation

$$\|e^h\|_{1,\Omega_0\setminus\bar{G}} \lesssim \|e^h\|_{0,\Omega_0\setminus\bar{F}} + |\lambda - \lambda_H| + \|u - u_H\|_{0,\Omega_0} + \|u^k - u\|_{1,\Omega_0\setminus\bar{F}} \tag{3.31}$$

can be derived from Proposition 2.2, (3.10), and

$$a(e^h, v) = \lambda_H(u_H, v) - \lambda(u, v) - a(u^k - u, v) \quad \forall v \in S_0^h(\Omega_0); \tag{3.32}$$

we then arrive at

$$\begin{aligned}\|P_h u - u^{k,h}\|_{1,\Omega\setminus\bar{G}} \lesssim{}& \|e^h\|_{0,\Omega_0} + |\lambda - \lambda_H| + \|u - u_H\|_{0,\Omega} \\ &+ \|P_h u - u\|_{1,\Omega\setminus\bar{F}} + \|u^k - u\|_{1,\Omega\setminus\bar{F}}.\end{aligned} \tag{3.33}$$

Thanks to (3.23) and the triangle inequality

$$\|u^k - u\|_{1,\Omega\setminus\bar{F}} \le \|u - P_k u\|_{1,\Omega\setminus\bar{F}} + \|P_k u - u^k\|_{1,\Omega\setminus\bar{F}},$$

we get

$$\|u^k - u\|_{1,\Omega\setminus\bar{F}} \le \|u - P_k u\|_{1,\Omega\setminus\bar{F}} + |\lambda - \lambda_H| + \|u - u_H\|_{0,\Omega}.$$

Thus combining (3.27) and (3.33), we conclude that

$$\|P_h u - u^{k,h}\|_{1,\Omega\setminus\bar{G}} \lesssim |\lambda - \lambda_H| + \|u - u_H\|_{0,\Omega} + H^\gamma \|P_h u - P_k u\|_{1,\Omega} + \|u - P_h u\|_{1,\Omega\setminus\bar{F}} + \|u - P_k u\|_{1,\Omega\setminus\bar{F}}. \tag{3.34}$$

Since $(\Omega \setminus \bar{F}) \subset\subset (\Omega \setminus \bar{D})$, we obtain from Proposition 2.2 that

$$\begin{aligned}\|P_h u - u^{k,h}\|_{1,\Omega\setminus\bar{G}} \lesssim\ & |\lambda - \lambda_H| + \|u - u_H\|_{0,\Omega} + H^\gamma \|P_h u - P_k u\|_{1,\Omega} \\ & + \inf_{v\in S_0^h(\Omega)} \|u - v\|_{1,\Omega\setminus\bar{D}} + \rho_\Omega(h)\|u - P_h u\|_{1,\Omega} \\ & + \inf_{v\in S_0^k(\Omega)} \|u - v\|_{1,\Omega\setminus\bar{D}} + \rho_\Omega(k)\|u - P_k u\|_{1,\Omega},\end{aligned}$$

which together with Lemma 3.1 yields

$$\|P_h u - u^{k,h}\|_{1,\Omega\setminus\bar{G}} \lesssim k^r + H^{r+s+\theta-1}. \tag{3.35}$$

Combining (3.35), (3.28), (3.29), and Lemma 3.1, finally, we complete the proof. □

3.3. Parallel algorithms. For eigenvalue problems having at least two isolated singular points, such as the Kohn–Sham equation of a system that contains more than two atoms, we can use the parallel versions of the algorithms designed above. Let n represent the number of isolated singular points. For each singular point, we choose a subdomain D_j containing it and then enlarge each D_j to Ω_j, and all these Ω_j satisfy that $\Omega_i \bigcap \Omega_j = \emptyset$ if $i \neq j(\ i, j = 1, \ldots, n)$; see Figure 3.3. Based on Algorithms A0 and B0, we have the following two parallel algorithms.

Algorithm A.

1. *Solve* (2.14) *on a globally coarse grid: find* $(\lambda_H, u_H) \in S_0^H(\Omega) \times \mathbb{R}$ *satisfying* $\|u_H\|_{0,\Omega} = 1$ *and*

$$a(u_H, v) = \lambda_H(u_H, v) \quad \forall v \in S_0^H(\Omega).$$

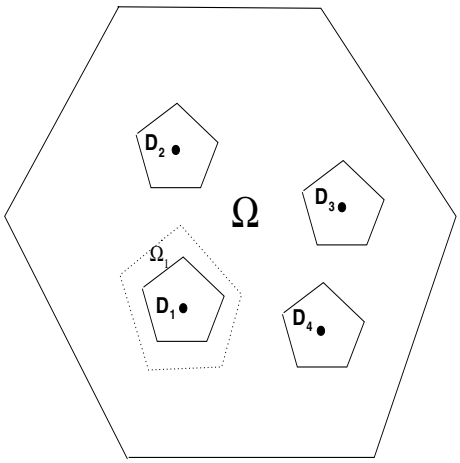


Fig. 3.3. *Localization.*

2. *Solve linear boundary value problems on locally fine grids in parallel: find* $e_h^j \in S_0^h(\Omega_j)(j = 1, 2, \ldots, n)$ *satisfying*

$$a(e_j^h, v) = \lambda_H(u_H, v) - a(u_H, v) \quad \forall v \in S_0^h(\Omega_j).$$

3. *Set*

$$u^h = \begin{cases} u_H + e_j^h & \text{on } \bar{\Omega}_j,\ j = 1, 2, \ldots, n, \\ u_H & \text{in } \Omega \backslash \left(\bigcup_{j=1}^n \bar{\Omega}_j \right) \end{cases}$$

and

$$\lambda^h = \frac{a(u^h, u^h)}{\|u^h\|_{0,\Omega}^2}.$$

Note that the assumption $\bigcup_{j=1}^n \bar{\Omega}_j = \bar{\Omega}$ is not required; hence Algorithm A is different from that in [39]. As a result, the local computational subdomains $\Omega_j (j = 1, 2, \ldots, n)$ may be much smaller. From the proof of Theorem 3.2, for this algorithm, we apparently have the following result.

THEOREM 3.4. *Assume that* u^h *is obtained by Algorithm* A *and assumption* $R(\Omega_j)$ *holds for* $j = 1, 2, \ldots, n$. *If* $u \in H_0^1(\Omega) \cap H^{r+s}(\Omega) \cap H^{r+1}(\Omega \setminus (\bigcup_{j=1}^n \bar{D}_j))(1 < r + s$ *and* $0 \le s < 1)$, *then*

$$\|u - u^h\|_{1,\Omega} \lesssim h^{r+s-1} + H^r + H^{r+s+\theta-1}$$

and

$$|\lambda - \lambda^h| \lesssim h^{2r+2s-2} + H^{2r} + H^{2r+2s+2\theta-2}.$$

ALGORITHM B.

1. *Solve* (2.14) *on a globally coarse grid: find* $(\lambda_H, u_H) \in S_0^H(\Omega) \times \mathbb{R}$ *satisfying* $\|u_H\|_{0,\Omega} = 1$ *and*

$$a(u_H, v) = \lambda_H(u_H, v) \quad \forall v \in S_0^H(\Omega).$$

2. *Solve a linear boundary value problem on a globally mesoscopic grid: find* $u^k \in S_0^k(\Omega)$ *satisfying*

$$a(u^k, v) = \lambda_H(u_H, v) \quad \forall v \in S_0^k(\Omega).$$

3. *Solve linear boundary value problems on locally fine grids in parallel: find* $e_j^h \in S_0^h(\Omega_j)(j = 1, 2, \ldots, n)$ *satisfying*

$$a(e_j^h, v) = \lambda_H(u_H, v) - a(u^k, v) \quad \forall v \in S_0^h(\Omega_j).$$

4. *Set*

$$u^{k,h} = \begin{cases} u^k + e_j^h & \text{on } \bar{\Omega}_j,\ j = 1, 2, \ldots, n, \\ u^k & \text{in } \Omega \backslash \left(\bigcup_{j=1}^n \bar{\Omega}_j \right) \end{cases}$$

and

$$\lambda^{k,h} = \frac{a(u^{k,h}, u^{k,h})}{\|u^{k,h}\|_{0,\Omega}^2}.$$

It should be mentioned that step 2 in Algorithm B can be localized and carried out in parallel, too (cf. [37]). For Algorithm B, we also have the following result.

THEOREM 3.5. *Assume that $u^{k,h}$ is obtained by Algorithm* B *and assumption $R(\Omega_j)$ holds for $j = 1, 2, \ldots, n$. If $u \in H_0^1(\Omega) \cap H^{r+s}(\Omega) \cap H^{r+1}(\Omega \backslash (\bigcup_{j=1}^n \bar{D}_j))(1 < r + s$ and $0 \le s < 1)$, then*

$$\|u - u^{k,h}\|_{1,\Omega} \lesssim h^{r+s-1} + k^r + H^{r+s+\theta-1}$$

and

$$|\lambda - \lambda^{k,h}| \lesssim h^{2r+2s-2} + k^{2r} + H^{2r+2s+2\theta-2}.$$

4. Discretization schemes for quantum eigenvalue problems. In this section, we shall construct some multiscale finite element discretization schemes for electronic structure computations and apply the schemes to obtain the ground state energies for several typical atomic and molecular systems. It is shown by our numerical experiments that the three-scale discretization schemes are very efficient and recommended to quantum eigenvalue computations.

4.1. Linearization of the Kohn–Sham equation. Modern electronic structure computations are usually based on the following Kohn–Sham equation (see, e.g., [7, 19, 25]):

$$\left(-\frac{1}{2}\Delta + V_{ne}(x) + V_H(\rho) + V_{xc}(\rho)\right) u_i = \lambda_i u_i \quad \text{in } \mathbb{R}^3 \tag{4.1}$$

with $\int_{\mathbb{R}^3} |u_i|^2 dx = 1 (i = 1, 2, \ldots, N)$, where the first term $-\frac{1}{2}\Delta u_i$ is the kinetic energy of the system, the second term $V_{ne}(x)$ is the potential of the electron-ion attraction,

$$V_{ne}(x) = -\sum_{q=1}^{N_{atom}} \frac{Z_q}{|x - r_q|},$$

the third term $V_H(\rho)$ is the Hartree approximation of the Coulomb potential of the system

$$V_H(\rho) = \int_{\mathbb{R}^3} \frac{\rho(y)}{|x - y|} dy,$$

the last term $V_{xc}(\rho)$ is the exchange-correlation potential, $\rho(x) = \sum_{i=1}^N |u_i(x)|^2$ is the electron density of the ground state of the system, N_{atom} is the total number of the atoms, Z_q is the electron number of the qth atom, r_q is the position of the qth atom, and N is the number of the electrons. Usually, $V_H(\rho)$ is obtained from solving the Poisson equation:

$$-\Delta V_H(\rho) = 4\pi\rho(x). \tag{4.2}$$

The most important thing for the molecular system, as we see, is the total energy of the ground state:

$$E_{total} = \sum_{i=1}^N \lambda_i - \frac{1}{2}\int_{\mathbb{R}^3}\int_{\mathbb{R}^3} \frac{\rho(x)\rho(y)dxdy}{|x - y|} + E_{xc}(\rho) - \int_{\mathbb{R}^3} \rho(x)V_{xc}(\rho)dx$$

$$+ \frac{1}{2} \sum_{p,q=1, p\neq q}^{N_{atom}} \frac{Z_p Z_q}{|r_p - r_q|},$$

where $E_{xc}(\rho)$ is the exchange-correlation energy, and $\lambda_i (i = 1, 2, \ldots, N)$ are the eigenvalues of (4.1). Since the Kohn–Sham equations (4.1) are nonlinear eigenvalue systems, we need to linearize and solve them iteratively, which is called the self-consistent approach [7, 19, 27, 32]. In our computations, a Broyden-type quasi-Newton method [34] will be used. The self-consistent iteration is described as follows.

SELF-CONSISTENT ITERATION.

1. Given an initial $\rho_{in}(x)$ by superposing atomic charge densities, obtain $V_{xc}(\rho_{in})$ and $V_H(\rho_{in})$.
2. Compute

$$V^{eff}(\rho_{in}) = V_{ne}(x) + V_H(\rho_{in}) + V_{xc}(\rho_{in})$$

and solve

$$\left(-\frac{1}{2}\Delta + V^{eff}(\rho_{in})\right) u_i = \lambda_i u_i \quad \text{in } \mathbb{R}^3 \tag{4.3}$$

with $\int_{\mathbb{R}^3} |u_i|^2 dx = 1 (i = 1, 2, \ldots, N)$.

3. Set $\rho_{out}(x) = \sum_{i=1}^{N} |u_i(x)|^2$.
4. Obtain $\rho_{in}^{new}(x)$ by "mixing" density with a Broyden-type quasi-Newton approach.
5. If self-consistent, stop; else repeat from step 2.

It is noted that the electron density ρ of the ground state of the system exponentially decays (see, e.g., [2, 14, 33]). Consequently, the central computation in solving the Kohn–Sham equations becomes the repeated solution of the linear eigenvalue problems (4.3) when $\mathbb{R}^3$ is replaced by some bounded domain $\Omega \subset \mathbb{R}^3$. We apply the multiscale finite element discretization schemes designed in section 3 to solve the linear eigenvalue problems (4.3).

4.2. Discretization schemes based on adaptive grids. Because of the singular potentials, we consider applying piecewise linear finite elements to solve (4.3). As we see, it is expensive to solve three-dimensional problems by using uniform grids, while it is much cheaper by using adaptive grids when accurate approximate solutions are required (cf. Tables 4.1–4.7). Thus, in our quantum eigenvalue computations, we will employ adaptive grids instead of uniform grids. The globally coarse grids we will use are some adaptive finite element grids, which are constructed from the bisection approaches [4] and the error indicators

$$\eta_\tau = \eta_{\tau,G} = \|A^{-1/2}(G_h u_h - A\nabla u_h)\|_{0,\tau}^2, \quad \tau \in T^h(\Omega),$$

where

$$A = \begin{pmatrix} 1/2 & 0 & 0 \\ 0 & 1/2 & 0 \\ 0 & 0 & 1/2 \end{pmatrix}$$

and the locally averaging operator $G_h : S_0^h(\Omega) \to S^h(\Omega) \times S^h(\Omega)$ is defined by (see, e.g., [40, 42])

$$G_h v = \sum_{z \in \partial^2 T^h} (A\nabla v)_z \phi_z, \qquad (A\nabla v)_z = \sum_{j=1}^{J_z} \alpha_z^j (A(z)\nabla v)_{\tau_z^j} \quad \forall v \in S_0^h(\Omega)$$

with $\bigcup_{j=1}^{J_z} \tau_z^j = \omega_z$, $\sum_{j=1}^{J_z} \alpha_z^j = 1$, $\omega_z = \bigcup_{z\in\bar{\tau},\tau\in T^h(\Omega)} \tau$, $\alpha_z^j \geq 0$ (for instance, $\alpha_z^j = \frac{1}{J_z}$, $\alpha_z^j = \frac{|\tau_z^j|}{|\omega_z|}$), and J_z is the number of elements containing z (see, e.g., [22]). While in the step of correction, the mesoscopic grid and the fine grid are obtained from the globally coarse grids directly by using some tetrahedral bisection strategy.

Remark 4.1. It should be pointed out that Assumption A.0 rules out most adaptively refined grids. However, the adaptive grids obtained from the bisection approach [4] satisfy Assumption A.0 if the initial grid, on which the bisection approach is posed, is fine enough. Indeed, let $T^{h_0}(\Omega)$ be the initial grid of size $h_0 \in (0,1)$ and $T^h(\Omega)$ be an adaptive grid of size h constructed from the bisection approach. Then there exists a positive integer m such that the finest size of elements in $T^h(\Omega)$ is $\mathcal{O}(h_0/2^m)$. Consequently, Assumption A.0 is satisfied when $\nu \geq 1 - \frac{m\ln 2}{\ln h_0}$.

We will use three types of algorithms to solve (4.3). The first one, which is denoted by Algorithm I, is the algorithm proposed in [39] for linear finite element eigenvalues: solve an eigenvalue problem in the linear finite element space associated with a relatively coarse grid and then find the correction by solving some Poisson boundary value problems over locally refined grids in parallel when there is more than one atom (the union of the subdomains is a partition of the whole domain); or it is the algorithm, which was studied in [38] in a general setting, presented in [20] for linear finite element eigenvalues: solve an eigenvalue problem in the linear finite element space associated with a relatively coarse grid and find the solution of the Poisson boundary value problem over a global refined grid when there is only one atom.

The second one, which is denoted by Algorithm II, is Algorithm A described in section 3: solve an eigenvalue problem in the globally coarse grid space and then find the correction by solving some Poisson boundary value problems over locally refined grids in parallel if there is more than one atom (the union of the subdomains may not be the exact whole domain); or it is Algorithm A0 described in section 3: solve an eigenvalue problem in the globally coarse grid space and then find the correction by solving some Poisson boundary value problem over a locally refined grid when there is only one atom.

The third one, which is denoted by Algorithm III, is Algorithm B described in section 3: solve an eigenvalue problem in the globally coarse grid space, then solve the Poisson boundary value problem in a globally mesoscopic grid, and, finally, find the correction by solving some Poisson boundary value problems in locally fine grids in parallel if there is more than one atom (the union of the subdomains may not be the exact whole domain); or it is Algorithm B0 described in section 3: solve an eigenvalue problem in the globally coarse grid space, then solve the Poisson boundary value problem in a globally mesoscopic grid, and, finally, find the correction by solving some Poisson boundary value problem in a locally fine grid when there is only one atom.

4.3. Numerical examples. In this subsection, we will report our numerical results of the ground state energies of the following typical atoms and molecules: hydrogen atom (H), helium atom (He), hydrogen molecule (H_2), and lithium hydride (LiH). The first two atoms are the simplest ones. While the diatomic hydrogen molecule H_2 is the most common molecule in space and the presence of H_2 also has important consequences regarding the origin of the universe and interpretation of the cosmological redshift [23], the molecule LiH is the simplest stable metallic hydride, and it is increasingly being studied as a potential storage agent for solid-state hydro-

gen storage with recent trials suggesting that it can exhibit the maximum reversible hydrogen capacity ever achieved in metal hydrides [3]. It is also believed that LiH molecules have played a crucial role in the early universe. The numerical examples were carried out on an SGI Origin 3800 in the State Key Laboratory of Scientific and Engineering Computing, Chinese Academy of Sciences.

Example 1. Consider the Kohn–Sham equation for hydrogen atoms:

$$\left(-\frac{1}{2}\Delta - \frac{1}{|x|}\right) u = \lambda u \quad \text{in } \mathbb{R}^3 \tag{4.4}$$

with $\int_{\mathbb{R}^3} |u|^2 dx = 1$. The first eigenvalue of (4.4), i.e., the total energy of the ground state for the hydrogen atom, is $\lambda = -0.5$.

In our computation for the total energy of the ground state, we solve the following eigenvalue problem by using one processor: find $(\lambda, u) \in H_0^1(\Omega) \times \mathbb{R}$ such that $\int_\Omega |u|^2 dx = 1$ and

$$\begin{cases} \left(-\frac{1}{2}\Delta - \frac{1}{|x|}\right) u = \lambda u & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega, \end{cases} \tag{4.5}$$

where $\Omega = (-6.4, 6.4)^3$. We choose $\Omega_0 = (-3.2, 3.2)^3$.

The numerical results obtained from the uniform grids are presented in Tables 4.1–4.3, while the numerical results obtained from the adaptive grids are presented in Tables 4.5–4.7. And the numerical results obtained from the standard discretization scheme on adaptive grids are reported in Table 4.4. Here and hereafter, we will use the following notation in our tables:

N_c : the number of degrees of freedom for solving the eigenvalue problem without any correction.

N_f: the total number of degrees of freedom for solving the boundary value problems for corrections.

λ_c: the numerical eigenvalue without any correction.

λ^h (or $\lambda^{k,h}$): the eigenvalue approximation obtained from the Rayleigh quotient.

TABLE 4.1
Example 1, *uniform grids, Algorithm* I: $H = 4h$.

N_c	λ_c	$\lvert(\lambda_c - \lambda)/\lambda\rvert$	N_f	λ^h	$\lvert(\lambda^h - \lambda)/\lambda\rvert$	CPU time(s)
27	−2.928560e-01	4.142880e-01	4913	−3.436870e-01	3.126261e-01	1.670403e+00
343	−3.690747e-01	2.618506e-01	35937	−4.228023e-01	1.543953e-01	1.600644e+01
3375	−4.317320e-01	1.365361e-01	274625	−4.733237e-01	5.335254e-02	1.529737e+02
6859	−4.474444e-01	1.051113e-01	531441	−4.826029e-01	3.479428e-02	3.207160e+02
19683	−4.662193e-01	6.756148e-02	1442897	−4.914295e-01	1.714091e-02	9.925273e+02
29791*	−4.720740e-01	5.585191e-02	-	-	-	6.561837e+01
59319*	−4.800278e-01	3.994440e-02	-	-	-	1.395323e+02

* *We are not able to carry out the computation on the fine grid correction on our SGI O*3800 *because the number of degrees of freedom is too large. The CPU time is that of getting the numerical solution on the coarse grid.*

It is shown by Tables 4.1–4.3 that, on uniform grids, Algorithms III and II are more efficient than Algorithm I, and the approximation accuracy of Algorithm III is better than that of Algorithm II, which coincide with the theory obtained in section 3. However, it is seen from comparing Tables 4.1–4.3 with Tables 4.5–4.7, respectively, that it is much cheaper to use adaptive grids than to use uniform grids to get numerical

TABLE 4.2
Example 1, *uniform grids, Algorithm* II: $H = 4h$.

N_c	λ_c	$\lvert(\lambda_c - \lambda)/\lambda\rvert$	N_f	λ^h	$\lvert(\lambda_c - \lambda)/\lambda\rvert$	CPU time(s)
27	−2.928560e-01	4.142880e-01	827	−3.297982e-01	3.404035e-01	3.030548e-01
343	−3.690747e-01	2.618506e-01	5517	−4.171119e-01	1.657762e-01	2.528012e+00
3375	−4.317320e-01	1.365361e-01	40121	−4.722106e-01	5.557884e-02	2.292948e+01
6859	−4.474444e-01	1.051113e-01	76851	−4.819518e-01	3.609643e-02	4.659957e+01
19683	−4.662193e-01	6.756148e-02	206207	−4.911308e-01	1.773831e-02	1.510508e+02
29791	−4.720740e-01	5.585195e-02	305649	−4.934068e-01	1.318640e-02	2.315993e+02
59319	−4.800278e-01	3.994432e-02	591101	−4.960204e-01	7.959192e-03	4.974919e+02

TABLE 4.3
Example 1, *uniform grids, Algorithm* III: $H = 2k = 4h$.

N_c	λ_c	$\lvert(\lambda_c - \lambda)/\lambda\rvert$	N_f	λ^h	$\lvert(\lambda_c - \lambda)/\lambda\rvert$	CPU time(s)
27	−2.928560e-01	4.142880e-01	1333	−3.491113e-01	3.017774e-01	4.802210e-01
343	−3.690747e-01	2.618506e-01	9097	−4.306885e-01	1.386230e-01	4.140220e+00
3375	−4.317320e-01	1.365361e-01	66961	−4.768579e-01	4.628429e-02	3.761074e+01
6859	−4.474444e-01	1.051113e-01	128581	−4.849741e-01	3.005177e-02	7.671814e+01
19683	−4.662193e-01	6.756148e-02	345997	−4.925891e-01	1.482173e-02	2.395772e+02
29791	−4.720740e-01	5.585195e-02	513313	−4.944662e-01	1.106756e-02	3.646883e+02
59319	−4.800278e-01	3.994432e-02	993961	−4.966182e-01	6.763663e-03	7.840637e+02

TABLE 4.4
Example 1, *adaptive grids, the standard discretization scheme (no correction).*

N_c	λ_c	$\lvert(\lambda_c - \lambda)/\lambda\rvert$	CPU time(s)
1313	−4.725548e-01	5.489033e-02	9.405628e+00
3213	−4.843072e-01	3.138552e-02	2.060037e+01
6241	−4.902593e-01	1.948136e-02	5.310144e+01
9257	−4.925712e-01	1.485750e-02	8.431472e+01
12343	−4.938080e-01	1.238395e-02	1.377305e+02
19193	−4.954112e-01	9.177673e-03	2.306809e+02
39489	−4.971723e-01	5.655475e-03	7.396562e+02
63313	−4.978759e-01	4.248137e-03	1.214609e+03

TABLE 4.5
Example 1, *adaptive grids, Algorithm* I: $H = 4h$.

N_c	λ_c	$\lvert(\lambda_c - \lambda)/\lambda\rvert$	N_f	λ^h	$\lvert(\lambda_c - \lambda)/\lambda\rvert$	CPU time(s)
241	−4.221418e-01	1.557164e-01	16809	−4.714494e-01	5.710129e-02	8.838912e+00
585	−4.451226e-01	1.097547e-01	37385	−4.849241e-01	3.015175e-02	2.570832e+01
963	−4.653817e-01	6.923651e-02	61457	−4.921725e-01	1.565495e-02	4.600226e+01
1313	−4.725548e-01	5.489033e-02	81577	−4.944553e-01	1.108933e-02	7.024775e+01
3213	−4.843072e-01	3.138552e-02	195257	−4.973894e-01	5.221128e-03	1.975070e+02
4167	−4.872962e-01	2.540769e-02	251393	−4.979776e-01	4.044742e-03	2.765680e+02
6241	−4.902594e-01	1.948117e-02	373929	−4.984659e-01	3.068125e-03	4.350539e+02
9257	−4.925713e-01	1.485747e-02	555433	−4.988557e-01	2.288552e-03	6.956501e+02
12343	−4.938081e-01	1.238386e-02	742017	−4.990644e-01	1.871178e-03	1.015238e+03

approximations. As a result, it is sufficient to show and discuss the numerical results for other examples on adaptive grids only.

Next we turn to discussing the numerical results obtained by our schemes on adaptive grids. First, it is seen from Tables 4.4–4.7 that our correction procedures are efficient. Second, it is observed from Tables 4.5 and 4.6 that when the scale of the coarse grids is the same, the time cost of Algorithm II is much less than that of Algorithm I, while the accuracy of Algorithm II is a little less than that of Algorithm I.

TABLE 4.6
Example 1, *adaptive grids, Algorithm* II: $H = 4h$.

N_c	λ_c	$\lvert(\lambda_c-\lambda)/\lambda\rvert$	N_f	λ^h	$\lvert(\lambda_c-\lambda)/\lambda\rvert$	CPU time(s)
241	−4.221418e-01	1.557164e-01	9027	−4.674482e-01	6.510358e-02	4.973507e+00
585	−4.451226e-01	1.097547e-01	24731	−4.816516e-01	3.669675e-02	1.651680e+01
963	−4.653817e-01	6.923651e-02	46457	−4.900794e-01	1.984110e-02	3.226771e+01
1313	−4.725548e-01	5.489033e-02	66577	−4.924359e-01	1.512812e-02	5.158988e+01
3213	−4.843072e-01	3.138552e-02	166661	−4.962501e-01	7.499826e-03	1.428840e+02
4167	−4.872962e-01	2.540769e-02	213725	−4.969481e-01	6.103848e-03	1.948279e+02
6241	−4.902593e-01	1.948136e-02	329043	−4.975815e-01	4.836956e-03	3.293773e+02
9257	−4.925712e-01	1.485750e-02	482275	−4.981307e-01	3.738594e-03	5.213350e+02
12343	−4.938080e-01	1.238395e-02	641919	−4.985082e-01	2.983695e-03	7.523030e+02

TABLE 4.7
Example 1, *adaptive grids, Algorithm* III: $H = 2k = 4h$.

N_c	λ_c	$\lvert(\lambda_c-\lambda)/\lambda\rvert$	N_f	$\lambda^{k,h}$	$\lvert(\lambda_c-\lambda)/\lambda\rvert$	CPU time(s)
241	−4.221418e-01	1.557164e-01	9917	−4.742024e-01	5.159524e-02	5.710342e+00
585	−4.451226e-01	1.097547e-01	26149	−4.863889e-01	2.722226e-02	1.826906e+01
963	−4.653817e-01	6.923651e-02	48121	−4.924374e-01	1.512518e-02	3.508268e+01
1313	−4.725548e-01	5.489033e-02	68241	−4.944089e-01	1.118230e-02	5.563193e+01
3213	−4.843072e-01	3.138552e-02	169753	−4.973025e-01	5.395056e-03	1.545872e+02
4167	−4.872962e-01	2.540769e-02	217785	−4.978502e-01	4.299644e-03	2.104204e+02
6241	−4.902593e-01	1.948136e-02	333877	−4.983160e-01	3.368087e-03	3.508001e+02
9257	−4.925712e-01	1.485750e-02	490301	−4.987160e-01	2.567987e-03	5.610495e+02
12343	−4.938080e-01	1.238395e-02	652885	−4.989565e-01	2.087074e-03	8.119487e+02

It should be noted that instead of a global boundary value problem, a local boundary value problem is solved in the step of correction; hence many more large-scale problems can be solved by Algorithm II than by Algorithm I. Third, thanks to Algorithm III, in which an additional correction is carried out over a global mesoscopic grid, the approximation accuracy can be improved. Different from Algorithm I, in Algorithm III, two small-scale boundary value problems, instead of a large-scale boundary value problem, are solved. As a result, for instance, it is seen from Tables 4.5 and 4.7 that the CPU times using Algorithm III are less than those of Algorithm I while keeping the same approximation accuracy. Consequently, Algorithm III is the first choice among the three algorithms taking into account both the computational complexity and the accuracy.

Example 2. Consider the Kohn–Sham equation for helium atoms:

$$\left(-\frac{1}{2}\Delta - \frac{2}{|x|} + \int \frac{\rho(y)}{|x-y|}dy + V_{xc}\right) u = \lambda u \quad \text{in } \mathbb{R}^3 \tag{4.6}$$

with $\int_{\mathbb{R}^3} |u|^2 = 1$, where $\rho = 2|u|$. In our computation for the total energy of the ground state, we solve the following nonlinear eigenvalue problem by using one processor: find $(\lambda, u) \in H_0^1(\Omega) \times \mathbb{R}$ such that $\int_\Omega |u|^2 dx = 1$ and

$$\begin{cases} \left(-\frac{1}{2}\Delta - \frac{2}{|x|} + \int \frac{\rho(y)}{|x-y|}dy + V_{xc}\right) u = \lambda u \text{ in } \Omega, \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad u = 0 \quad \text{on } \partial\Omega, \end{cases} \tag{4.7}$$

where $\Omega = (-10.0, 10.0)^3$. In computation, we choose $V_{xc}(\rho) = -\frac{3}{2}\alpha(\frac{3}{\pi}\rho)^{\frac{1}{3}}$, where $\alpha = 0.77298$. The local computation subdomain $\Omega_0 = (-5.0, 5.0)^3$.

The relevant numerical results in the literature are provided in Table 4.8. The exact total energy of the ground state may be viewed as -2.90 (cf. [9, 31]). Our numerical results are reported in Tables 4.9–4.11. It is shown that our results are good enough, and Algorithms II and III are better than Algorithm I taking into account both the computational complexity and the accuracy.

TABLE 4.8
He atoms.

Model	Method	Reference	Degree of freedom	Energy
Schrödinger	standard FEM	[31]	$6,720^*$	-2.90289
Schrödinger	standard FEM	[9]	unknown	-2.9036118
Hatree–Fock	standard FEM	[36]	$500,000^{**}$	-2.8522

* *Here special basis functions and special grids are used.*

** *Here the products of orthogonal or nonorthogonal one-dimensional finite element shape functions are used to form three-dimensional basis functions.*

TABLE 4.9
Example 2, *Algorithm* I: $H = 4h$.

N_c	N_f	λ	E_{total}	CPU time(s)
485	31225	−6.296393e-01	−2.917657e+00	4.530073e+01
931	58449	−6.536631e-01	−2.928375e+00	8.776221e+01
1483	92817	−6.307205e-01	−2.915374e+00	1.706035e+02
2495	152545	−6.300252e-01	−2.916329e+00	3.007054e+02
3673	221097	−6.213038e-01	−2.903806e+00	4.491209e+02
4969	300201	−6.169102e-01	−2.900137e+00	6.508143e+02
6255	370145	−6.120339e-01	−2.895762e+00	8.442268e+02
9719	579841	−6.063249e-01	−2.888356e+00	1.397660e+03
14547	859793	−6.023419e-01	−2.884202e+00	2.012624e+03
19501	1149849	−6.036066e-01	−2.884971e+00	2.927486e+03

TABLE 4.10
Example 2, *Algorithm* II: $H = 4h$.

N_c	N_f	λ^h	E_{total}	CPU time(s)
485	23443	−6.299719e-01	−2.918322e+00	3.662065e+01
931	45795	−6.535703e-01	−2.928190e+00	7.033673e+01
1483	80163	−6.305399e-01	−2.915013e+00	1.435414e+02
2495	139891	−6.299741e-01	−2.916227e+00	2.574028e+02
3673	208443	−6.212387e-01	−2.903676e+00	3.911174e+02
4969	287547	−6.168342e-01	−2.899985e+00	5.550541e+02
6255	357491	−6.119810e-01	−2.895656e+00	7.179750e+02
9719	567187	−6.062732e-01	−2.888252e+00	1.228756e+03
14547	847139	−6.022975e-01	−2.884114e+00	1.770803e+03
19501	1133001	−6.035775e-01	−2.884912e+00	2.645706e+03

Because there is only one singular point in Examples 1 and 2, we use only one processor to solve them, which results in the advantage of Algorithms II and III not being very apparent. In the following two examples, however, because there are two isolated singularity points, we solve them by using two processors in parallel: first, we solve an eigenvalue problem in a globally coarse grid (and a boundary value problem in a globally refined grid); then we solve two boundary value problems in locally refined grids in parallel. It is seen that the time cost of Algorithms II and III is much less than that of Algorithm I, while the corresponding accuracy is comparable with

TABLE 4.11
Example 2, *Algorithm* III: $H = 2k = 4h$.

N_c	N_f	$\lambda^{k,h}$	E_{total}	CPU time(s)
485	24333	−6.356068e-01	−2.929592e+00	3.896100e+01
931	47213	−6.574089e-01	−2.935867e+00	7.492060e+01
1483	81581	−6.318015e-01	−2.917536e+00	1.510373e+02
2495	141309	−6.304890e-01	−2.917257e+00	2.720404e+02
3673	209861	−6.215602e-01	−2.904319e+00	4.118891e+02
4969	288965	−6.170269e-01	−2.900370e+00	5.808058e+02
6255	358909	−6.121141e-01	−2.895923e+00	7.547003e+02
9719	568605	−6.063521e-01	−2.888410e+00	1.300196e+03
14547	848557	−6.023488e-01	−2.884216e+00	1.841649e+03
19501	1134857	−6.036085e-01	−2.884974e+00	2.826073e+03

that of Algorithm I. Taking into account both the computational complexity and the accuracy, Algorithms II and III are better than Algorithm I. And Algorithm III is recommended if more accurate approximations are required.

Example 3. Consider the Kohn–Sham equation of diatomic hydrogen molecules:

$$\left(-\frac{1}{2}\Delta - \frac{1}{|x-r_1|} - \frac{1}{|x-r_2|} + \int \frac{\rho(y)}{|x-y|}dy + V_{xc}\right)u = \lambda u \ \text{ in } \mathbb{R}^3 \tag{4.8}$$

with $\int_{\mathbb{R}^3} |u|^2 = 1$, where $\rho = 2|u|$, $r_j (j = 1, 2)$ are the positions of the two hydrogen atoms. Here we choose $r_1 = (-0.70, 0, 0), r_2 = (0.70, 0, 0)$. In our computation for the total energy of the ground state, we use two processors to solve the following nonlinear eigenvalue problem: find $(\lambda, u) \in H_0^1(\Omega) \times \mathbb{R}$ such that $\int_\Omega |u|^2 dx = 1$ and

$$\begin{cases} \left(-\frac{1}{2}\Delta - \frac{1}{|x-r_1|} - \frac{1}{|x-r_2|} + \int \frac{\rho(y)}{|x-y|}dy + V_{xc}\right)u = \lambda u \ \text{ in } \Omega, \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad u = 0 \ \text{ on } \partial\Omega, \end{cases} \tag{4.9}$$

where $\Omega = (-10.0, 10.0)^3$. Here we choose V_{xc} as follows (see [27]):

$$V_{xc}(\rho) = \begin{cases} \frac{\kappa}{1+\beta_1\sqrt{r_s}+\beta_2 r_s}\frac{1+\frac{7}{6}\beta_1\sqrt{r_s}+\frac{4}{3}\beta_2 r_s}{1+\beta_1\sqrt{r_s}+\beta_2 r_s} - (\frac{9}{4\pi^2})^{\frac{1}{3}}\frac{1}{r_s} & \text{if } r_s \geq 1, \\ \xi\ln(r_s) + \zeta - \frac{1}{3}\xi + \frac{2}{3}\eta r_s \ln(r_s) + \frac{1}{3}(2\mu - \eta)r_s - (\frac{9}{4\pi^2})^{\frac{1}{3}}\frac{1}{r_s} & \text{if } r_s < 1, \end{cases}$$

where $r_s = (\frac{3}{4\pi\rho})^{\frac{1}{3}}$, $\kappa = -0.1423$, $\beta_1 = 1.0529$, $\beta_2 = 0.3334$, $\xi = 0.0311$, $\zeta = -0.048$, $\eta = 0.0020$, and $\mu = -0.0116$. The local subdomains in Algorithms II and III are $\Omega_1 = (-5.0, 0) \times (-5.0, 5.0)^2$ and $\Omega_2 = (0, 5.0) \times (-5.0, 5.0)^2$.

The experimental value for the ground state energy and the equilibrium bond distance of H_2 are −1.1746 a.u. and 1.40 a.u., respectively [17]. Some numerical results in literature [13] are listed in Table 4.12. Our numerical results are reported in Tables 4.13–4.15. Compared with that of [13], our results are good enough, too.

Example 4. Consider the Kohn–Sham equations for LiH molecules:

$$\left(-\frac{1}{2}\Delta - \frac{3}{|x-r_1|} - \frac{1}{|x-r_2|} + \int \frac{\rho(y)}{|x-y|}dy + V_{xc}\right)u_i = \lambda_i u_i \ \text{ in } \mathbb{R}^3 \tag{4.10}$$

with $\int_{\mathbb{R}^3} |u_i|^2 = 1 (i = 1, 2)$, where $\rho = 2(|u_1|^2 + |u_2|^2)$, $r_j (j = 1, 2)$ are the positions of the lithium atom and the hydrogen atom, respectively. Here we choose $r_1 = (-1.0075, 0, 0)$ and $r_2 = (2.0075, 0, 0)$.

TABLE 4.12
Molecule H_2.

	MO^*	VB^{**}
E (a.u.)	−1.1282	−1.1389
% error	3.95	3.04

** Here MO means method considering molecular orbital wave function.*
*** Here VB means method considering valence bond wave function.*

TABLE 4.13
Example 3, *Algorithm* I: $H = 4h$.

N_c	λ^h	E_{total}	CPU time(s)
343	−3.132335e-01	−1.177330e+00	2.075182e+01
493	−3.440112e-01	−1.205555e+00	3.995146e+01
541	−3.785985e-01	−1.196054e+00	4.590348e+01
653	−4.006745e-01	−1.191450e+00	5.837562e+01
893	−3.910186e-01	−1.177869e+00	8.521689e+01
1171	−3.986831e-01	−1.178966e+00	1.116769e+02
1629	−3.963278e-01	−1.172159e+00	1.629191e+02
2285	−3.980473e-01	−1.169919e+00	2.359080e+02
2875	−3.986961e-01	−1.169829e+00	3.127561e+02
3817	−3.956656e-01	−1.165450e+00	4.304949e+02

TABLE 4.14
Example 3, *Algorithm* II: $H = 4h$.

N_c	λ^h	E_{total}	CPU time(s)
343	−2.551035e-01	−1.061154e+00	2.952539e+00
493	−3.278807e-01	−1.173303e+00	1.621324e+01
541	−3.628029e-01	−1.164471e+00	2.095883e+01
653	−3.863669e-01	−1.162842e+00	2.905838e+01
893	−3.810261e-01	−1.157890e+00	4.427377e+01
1171	−3.925215e-01	−1.166645e+00	5.985596e+01
1629	−3.918260e-01	−1.163158e+00	9.051947e+01
2285	−3.947462e-01	−1.163318e+00	1.274517e+02
2875	−3.959790e-01	−1.164395e+00	1.726272e+02
3817	−3.934625e-01	−1.161045e+00	2.383266e+02

TABLE 4.15
Example 3, *Algorithm* III: $H = 2k = 4h$.

N_c	$\lambda^{k,h}$	E_{total}	CPU time(s)
343	−2.987938e-01	−1.148463e+00	4.610338e+00
493	−3.400216e-01	−1.197577e+00	1.830246e+01
541	−3.751566e-01	−1.189171e+00	2.310248e+01
653	−3.976311e-01	−1.185365e+00	3.120803e+01
893	−3.891308e-01	−1.174095e+00	4.706458e+01
1171	−3.973738e-01	−1.176347e+00	6.370687e+01
1629	−3.952668e-01	−1.170038e+00	9.564380e+01
2285	−3.973787e-01	−1.168582e+00	1.345256e+02
2875	−3.981450e-01	−1.168726e+00	1.825406e+02
3817	−3.952144e-01	−1.164548e+00	2.496900e+02

In our computation for the total energy of the ground state, we use two processors to solve the following nonlinear eigenvalue problem: find $(\lambda_i, u_i) \in H_0^1(\Omega)\times\mathbb{R}(i = 1, 2)$

such that $\int_\Omega |u_i|^2 dx = 1$ and

$$\text{(4.11)}\quad \begin{cases} \left(-\frac{1}{2}\Delta - \frac{3}{|x-r_1|} - \frac{1}{|x-r_2|} + \int \frac{\rho(y)}{|x-y|}dy + V_{xc}\right) u_i = \lambda_i u_i \ \text{ in } \Omega, \\ u_i = 0 \ \text{ on } \partial\Omega, \end{cases}$$

where $\Omega = (-40.0, 40.0)^3$. In computation, we choose V_{xc} as that of Example 3 and the local subdomains in Algorithms II and III to be $\Omega_1 = (-20.0, 0) \times (-20.0, 20.0)^2$ and $\Omega_2 = (0, 20.0) \times (-20.0, 20.0)^2$.

For a fixed LiH bondlength 3.015 a.u., the experimental value of the ground state energy of LiH is -8.0705 a.u. and one numerical result in literature is -7.9891 a.u. [3]. Some other numerical results of LiH can also be found in [18], in which the energy of the ground state of LiH is -7.987352237 a.u. Our numerical results are reported in Tables 4.16–4.18. It is seen that our results are more agreeable with the experimental value than those of [3] and [18].

TABLE 4.16
Example 4, *Algorithm* I: $H = 4h$.

N_c	λ_1^h	λ_2^h	E_{total}	CPU time(s)
1431	−1.920090e+00	−1.337586e-01	−8.027624e+00	1.271456e+03
1779	−1.882749e+00	−1.293489e-01	−7.991300e+00	1.415080e+03
2179	−1.875230e+00	−1.318718e-01	−7.982886e+00	1.583497e+03
2199	−1.870919e+00	−1.314579e-01	−7.979606e+00	1.683251e+03
2662	−1.859604e+00	−1.329593e-01	−7.969458e+00	1.915602e+03
3827	−1.849146e+00	−1.420370e-01	−7.966170e+00	2.429021e+03
5394	−1.877563e+00	−1.804833e-01	−8.016103e+00	3.230563e+03
8022	−1.884737e+00	−1.965154e-01	−8.036745e+00	4.245198e+03
10134	−1.884991e+00	−2.031805e-01	−8.044572e+00	8.012743e+03

TABLE 4.17
Example 4, *Algorithm* II: $H = 4h$.

N_c	λ_1^h	λ_2^h	E_{total}	CPU time(s)
1431	−1.915160e+00	−1.133577e-01	−7.976961e+00	1.050816e+03
1779	−1.879557e+00	−1.143428e-01	−7.954904e+00	1.144955e+03
2179	−1.873086e+00	−1.184806e-01	−7.951815e+00	1.230697e+03
2199	−1.868878e+00	−1.184214e-01	−7.949451e+00	1.320936e+03
2662	−1.858165e+00	−1.210660e-01	−7.942792e+00	1.455666e+03
3827	−1.848310e+00	−1.314935e-01	−7.943412e+00	1.729450e+03
5394	−1.877026e+00	−1.734047e-01	−8.000871e+00	2.165690e+03
8022	−1.884389e+00	−1.903621e-01	−8.023741e+00	2.843285e+03
10134	−1.884752e+00	−1.972828e-01	−8.032297e+00	3.757715e+03

TABLE 4.18
Example 4, *Algorithm* III: $H = 2k = 4h$.

N_c	$\lambda_1^{k,h}$	$\lambda_2^{k,h}$	E_{total}	CPU time(s)
1431	−1.918868e+00	−1.278341e-01	−8.013331e+00	1.090022e+03
1779	−1.882012e+00	−1.252885e-01	−7.981708e+00	1.205499e+03
2179	−1.874744e+00	−1.283515e-01	−7.974875e+00	1.263880e+03
2199	−1.870463e+00	−1.280616e-01	−7.971897e+00	1.353980e+03
2662	−1.859275e+00	−1.299885e-01	−7.962852e+00	1.533450e+03
3827	−1.848957e+00	−1.394616e-01	−7.960640e+00	1.830007e+03
5394	−1.877442e+00	−1.794289e-01	−8.013752e+00	2.161640e+03
8022	−1.884659e+00	−1.956278e-01	−8.034816e+00	2.947668e+03
10134	−1.884939e+00	−2.023394e-01	−8.042786e+00	3.661563e+03

5. Concluding remarks. In this paper, we have proposed some three-scale finite element discretization schemes for a class of elliptic eigenvalue problems and especially applied these new schemes to several typical electronic structure computations successfully. We believe that these kinds of computational approaches are powerful techniques in obtaining accurate and efficient approximations for large-scale eigenvalue problems. Indeed, to apply these approaches to complex quantum chemistry computations is our ongoing project. For such applications to complex quantum chemistry computations, however, there are many practical issues, including the implementation details for local density approximations and pseudopotentials, that need to be addressed. We will present our progresses in our forthcoming papers.

Acknowledgments. The authors would like to thank the referees for their valuable comments and suggestions that have improved the presentation of this paper and Prof. X. Gong, Dr. L. Shen, and Mr. D. Zhang for stimulating discussions and fruitful cooperation on electronic structure computations that have motivated the topic in this paper.

REFERENCES

[1] R. A. Adams, *Sobolev Spaces*, Academic Press, New York, 1975.
[2] S. Agmon, *Lectures on the Exponential Decay of Solutions of Second-Order Elliptic Operators*, Princeton University Press, Princeton, NJ, 1981.
[3] B. K. Antony, K. N. Joshipura, N. J. Mason, and J. Tennyson, *R-matrix calculation of low-energy electron collisions with LiH*, J. Phys. B, 37 (2004), pp. 1689–1697.
[4] D. N. Arnold, A. Mukherjee, and L. Pouly, *Locally adapted tetrahedral meshes using bisection*, SIAM J. Sci. Comput., 22 (2000), pp. 431–448.
[5] I. Babuska and J. E. Osborn, *Finite element-Galerkin approximation of the eigenvalues and eigenvectors of selfadjoint problems*, Math. Comp., 52 (1989), pp. 275–297.
[6] I. Babuska and J. E. Osborn, *Eigenvalue problems*, in Handbook of Numerical Analysis, Vol. II, P. G. Ciarlet and J. L. Lions, eds., North–Holland, Amsterdam, 1991, pp. 641–792.
[7] T. L. Beck, *Real-space mesh techniques in density-function theory*, Rev. Modern Phys., 72 (2000), pp. 1041–1080.
[8] A. Brandt, S. McCormick, and J. Ruge, *Multigrid methods for differential eigenproblems*, SIAM J. Sci. Stat. Comput., 4 (1983), pp. 244–260.
[9] M. Braun, W. Schweizer, and H. Herold, *Finite-element calculations for the S states of helium*, Phys. Rev. A (3), 48 (1993), pp. 1916–1920.
[10] F. Chatelin, *Spectral Approximations of Linear Operators*, Academic Press, New York, 1983.
[11] P. G. Ciarlet and J. L. Lions, eds., *Handbook of Numerical Analysis*, Vol. II, North–Holland, Amsterdam, 1991.
[12] S. Costiner and S. Ta'asan, *Adaptive multigrid techniques for large-scale eigenvalue problems: Solutions of the Schrödinger problem in two and three dimensions*, Phys. Rev. E (3), 51 (1995), pp. 3704–3717.
[13] R. Frank, *The covalent bond in H_2*, Chem. Educator, 6 (2001), pp. 288–290.
[14] L. Gårding, *On the essential spectrum of Schrödinger operators*, J. Funct. Anal., 52 (1983), pp. 1–10.
[15] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Pitman, Boston, 1985.
[16] W. Hackbusch, *Multigrid Methods and Applications*, Springer-Verlag, Berlin, 1985.
[17] K. P. Huber and G. Herzberg, *Constants of Diatomic Molecules*, Vol. IV, Van Nostrand Reinhold, New York, 1979.
[18] J. Kobus, D. Moncrieff, and S. Wilson, *Comparison of polarizabilities and hyperpolarizabilities obtained from finite basis set and finite difference Hartree-Fock calculations for diatomic molecules*, J. Phys. B, 34 (2001), pp. 5127–5143.
[19] W. Kohn and L. J. Sham, *Self-consistent equations including exchange and correlation effects*, Phys. Rev. A, 140 (1965), pp. 1133–1138.
[20] Q. Lin and G. Xie, *Accelerating the finite element method in eigenvalue problems*, Kexue Tongbao, 26 (1981), pp. 449–452 (in Chinese).
[21] J. Mandel and S. McCormick, *A multilevel variational method for $Au = \lambda Bu$ on composite grids*, J. Comput. Phys., 80 (1989), pp. 442–452.

[22] D. Mao, L. Shen, and A. Zhou, *Adaptive finite algorithms for eigenvalue problems based on local averaging type a posteriori error estimates*, Adv. Comput. Math., 25 (2006), pp. 135–160.
[23] P. Marmet, *Discovery of H_2 in space explains dark matter and redshift*, 21th Century Science and Technology, 13 (2000), pp. 5–7.
[24] S. F. McCormick, *A mesh refinement method for $Ax = \lambda Bx$*, Math. Comp., 36 (1981), pp. 485–498.
[25] R. G. Parr and W. T. Yang, *Density-Functional Theory of Atoms and Molecules*, Oxford University Press, New York, Clarendon Press, Oxford, UK, 1994.
[26] J. E. Pask and P. A. Sterne, *Finite element methods in* ab initio *electronic structure calculations*, Model. Simul. Mater. Sci. Eng., 13 (2005), pp. 71–96.
[27] J. P. Perdew and A. Zunger, *Self-interaction correction to density-functional approximations for many-electron*, Phys. Rev. B (3), 23 (1981), pp. 5048–5079.
[28] A. H. Schatz and L. B. Wahlbin, *Interior maximum-norm estimates for finite element methods*, Math. Comp., 31 (1977), pp. 414–442.
[29] A. H. Schatz and L. B. Wahlbin, *Interior maximum-norm estimates for finite element methods, Part* II, Math. Comp., 64 (1995), pp. 907–928.
[30] A. H. Schatz and J. Wang, *Some new error estimates for Ritz-Galerkin methods with minimal regularity assumptions*, Math. Comp., 65 (1996), pp. 19–27.
[31] A. Scrinzi *A* 3*-dimensional finite elements procedure for quantum mechanical applications*, Comput. Phys. Comm., 86 (1995), pp. 67–80.
[32] L. Shen and A. Zhou, *A defect correction scheme for finite element eigenvalues with applications to quantum chemistry*, SIAM J. Sci. Comput., 28 (2006), pp. 321–338.
[33] B. Simon, *Schrödinger operators in the twentieth century*, J. Math. Phys., 41 (2000), pp. 3523–3555.
[34] G. P. Srivastava, *Broyden's method for self-consistent field convergence acceleration*, J. Phys. A, 17 (1984), pp. 317–321.
[35] L. Strakhovskaya, *An iterative method for evaluating the first eigenvalue of an elliptic operator*, USSR Comput. Math. Math. Phys., 17 (1977), pp. 88–101.
[36] S. R. White, J. W. Wilkins, and M. P. Teter, *Finite-element method for electronic structure*, Phys. Rev. B (3), 39 (1989), pp. 5819–5833.
[37] J. Xu and A. Zhou, *Local and parallel finite element algorithms based on two-grid discretizations*, Math. Comp., 69 (2000), pp. 881–909.
[38] J. Xu and A. Zhou, *A two-grid discretization scheme for eigenvalue problems*, Math. Comp., 70 (2001), pp. 17–25.
[39] J. Xu and A. Zhou, *Local and parallel element algorithms for eigenvalue problems*, Acta Math. Appl. Sin. Engl. Ser., 18 (2002), pp. 185–200.
[40] N. Yan and A. Zhou, *Gradient recovery type a posteriori error estimates for finite element approximations on irregular msshes*, Comput. Methods Appl. Mech. Engrg., 190 (2001), pp. 4289–4299.
[41] A. Zhou, C. L. Liem, T. M. Shih, and T. Lü, *Error analysis on bi-parameter finite elements*, Comput. Methods Appl. Mech. Engrg., 158 (1998), pp. 329–339.
[42] O. C. Zienkiewicz and J. Z. Zhu, *The superconvergence patch recovery and a posteriori error estimates*, Internat. J. Numer. Methods Engrg., 33 (1992), pp. 1331–1382.

SIAM J. NUMER. ANAL.
Vol. 46, No. 1, pp. 325–343
© 2008 Society for Industrial and Applied Mathematics

LIE SYMMETRY PRESERVATION AND SHOCK-CAPTURING METHODS*

ZHENG RAN†

Abstract. It is known that some traditional shock-capturing schemes, including Godunov- or Roe-type schemes, generate downstream oscillatory wave patterns that cannot be effectively dampened by the dissipation methods they provide. The theme of this paper is to understand the formation and behavior at these upstream/downstream patterns from the point view of the symmetry inherent in the PDE. This paper investigates symmetry breaking and spatial oscillations in finite difference solutions for the inviscid Burgers and one-dimensional Navier–Stokes shock equations using the Lax–Wendroff, TVD, and nonoscillatory, nonfree parameter dissipative schemes. Symmetry breaking is considered as a candidate for the inherent cause of the nonphysical oscillations in the vicinity of the shock.

Key words. difference scheme, symmetry, shock-capturing method

AMS subject classifications. 22E70, 39A11, 65M06

DOI. 10.1137/050648201

1. Introduction.

The past four decades have seen tremendous progress in the development of numerical methods for solving the equations governing the unsteady flow of an inviscid, adiabatic ideal gas, that is, the Euler equations [1, 2, 3, 4, 5]. The goal of shock-capturing methods is to capture discontinuities in solutions to these equations automatically, without explicitly tracking them. Discontinuities must then be smeared over one or more grid cells. Success requires that the method implicitly incorporates the correct jump conditions, reduces smearing to a minimum, and does not introduce nonphysical oscillations near the discontinuities. High-resolution finite difference/volume methods based on Riemann solutions often perform well and are simple to implement. There are several circumstances where even first order schemes (such as the Godunov method and the Roe scheme) may produce inferior results [6]. One known example is the computation of slow-moving shocks. Several references have reported the difficulty of computing slow-moving shocks, where first order Godunov- or Roe-type methods produce spurious long wave oscillations behind the shock [7, 8, 9]. These eventually ruin the downstream solution. A full understanding of the nature of this kind of numerical phenomenon is still lacking. Regardless of definition, spurious oscillations and overshoots are the most common symptoms of numerical instability. In one natural interpretation, these numerical phenomena are due to nonlinear stability. Linear stability theory has a strong history, dating back to the late 1940s, but the study of nonlinear stability is far newer. Starting with papers by Boris and Book [10] and van Leer [11], nonlinear stability theory has developed over roughly the past thirty years. Perhaps the best known nonlinear stability condition is the total variation diminishing (TVD) condition, suggested by Harten in 1983 [2]. A review of this can be found in [12]. This paper addresses the interesting problem of nonlinear instability from a new point of view. Before moving on, it should be noted that the Euler equations and most scalar conservation laws are highly nonlinear.

*Received by the editors December 22, 2005; accepted for publication (in revised form) August 8, 2007; published electronically January 23, 2008. This research was supported by the National Natural Science Foundation of China under grants 10272018 and 10572083.
http://www.siam.org/journals/sinum/46-1/64820.html
†Shanghai Institute of Applied Mathematics and Mechanics, Shanghai University, Shanghai 200072, People's Republic of China (zran@staff.shu.edu.cn).

Numerical approximations for these should contain at least as much nonlinearity as the governing equations. Most modern numerical methods add their own nonlinearity, making them highly nonlinear even for linear governing equations. Unfortunately, in general, there are still many unresolved issues in nonlinear stability theory. There are two primary drawbacks in the existing theoretical development: heuristics and one-dimensional analysis. Indeed, both of these reflect the difficulties presented by high nonlinearity. It is hoped that Lie symmetry analysis can be used to throw some light on this subject.

Symmetry has always played an important role in mechanics, from the fundamental formulation of basic principles to concrete application [13, 14, 15]. Applications of symmetry groups to numerical analysis are given extensive treatment by Shokin [16] and Dorodnistyn [17]. For gas dynamics, the noninvariance relative to Galileo transformation of a difference scheme which approximates the equations results in nonphysical fluctuations, which was remarked in the 1960s (Harlow 1964).[1] In 1970 Yanenko and Shokin [18] developed a method of differential approximations for the study of the group properties of difference schemes for hyperbolic systems of equations. They used the first differential approximation to perform a group analysis. The study of symmetry in difference equations is a much more recent phenomenon than for differential equations. Early work in this direction performed by Maeda [19] concentrates on transformations acting only on dependent variables. A more recent series of papers was devoted to the Lie point symmetries of differential difference equations on fixed regular lattices [20]. A different approach was developed mainly for linear or linearizable difference equations and involved transformations acting on more than one point of the lattice [21, 22]. The symmetries considered in this approach are really generalized; however, they reduce to point symmetries at the continuous limit. A more general class of generalized symmetries has also been investigated for difference equations, and differential-difference equations on fixed regular lattices [23, 24]. A further approach to investigating symmetries in discrete equations was originally suggested by Dorodnitsyn and collaborators [25, 26]. The main aim of this series of papers is to make differential equations discrete while preserving their Lie point symmetries. Symmetries of ordinary and partial difference schemes on lattices that are a prior given, but are allowed to transform under point transformations, were studied in [27, 28, 29]. In a series of lectures, presented at the CIMPA Winter School on Discrete Integrability Systems in February 2003, Winternitz [30] gave a useful review of the application of Lie point symmetries, and their generalizations, to the study of difference equations. The overall theme could be called "continuous symmetries of discrete equations."

Three important questions on numerical analysis, especially important for shock-capturing methods are discussed from the point of view of group theory below. They are

(1) the group classification of the Lax–Wendroff scheme,

(2) symmetry and Harten's entropy enforcement condition,

(3) the role of symmetry in various aspects of a proper manipulation of the coefficients of the third order derivative ν_3 in the nonoscillatory, nonfree parameter dissipative (NND) scheme.

The structure of this paper is as follows. First, the theory of symmetries of differential equations is very briefly discussed, by way of a theoretical introduction in section 2. The group classification of the nonlinear wave equations is discussed in section 3. The following three sections are devoted to a complete development of

[1] Thanks to the referee's information.

the Lie symmetry analysis method proposed here and its application to three special cases of interest. The final section contains concluding remarks.

2. Mathematical preliminaries. All of the problems addressed here can be described by a general system of nonlinear differential equations of the nth order

$$\Delta_\nu \left(x, u^{(n)}\right) = 0, \tag{2.1}$$

where $\nu = 1, \ldots, l$, and $x = \left(x^1, \ldots, x^p\right) \in X$ are independent variables, $u = \left(u^1, \ldots, u^q\right) \in U$ are dependent variables, and $\Delta_\nu(x, u^{(n)}) = (\Delta_1(x, u^{(n)}), \ldots, \Delta_l(x, u^{(n)}))$ is a smoothing function that depends on x, u, and derivatives of u up to order n with respect to $x^1, \ldots, x^p$. If we define a jet space $X \times U^{(n)}$ as a space whose coordinates are independent variables, dependent variables, and derivatives of dependent variables up to order n, then Δ is a smoothing mapping

$$\Delta : X \times U^{(n)} \to R^l. \tag{2.2}$$

Before studying the symmetries of difference schemes, let us very briefly review the theory of symmetries for differential equations. For all details, proofs, and further information we refer to the many excellent books on the subject, e.g., [13, 14, 15]. Here we follow the style of [13], but the Lie symmetry description is made concise by emphasizing the significant points and results. In order to provide the reader with a relatively quick and painless introduction to Lie symmetry theory, some important concepts must be introduced.

The main tool used in Lie group theory and working with transformation groups is "infinitesimal transformation." In order to present this, we need first to develop the concept of a vector field on a manifold. We begin with a discussion of tangent vectors. Suppose C is a smooth curve on a manifold M, parameterized by

$$\phi : I \to M, \tag{2.3}$$

where I is a subinterval of R. In local coordinates $x = \left(x^1, \ldots, x^p\right)$, C is given by p smoothing functions

$$\phi(\varepsilon) = \left(\phi^1(\varepsilon), \ldots, \phi^p(\varepsilon)\right) \tag{2.4}$$

of the real variable ε. At each point $x = \phi(\varepsilon)$ of C the curve has a tangent vector, namely, the derivative

$$\dot{\phi} = \frac{d\phi}{d\varepsilon} = \left(\frac{d\phi^1}{d\varepsilon}, \ldots, \frac{d\phi^p}{d\varepsilon}\right). \tag{2.5}$$

In order to distinguish between tangent vectors and local coordinate expressions for a point on the manifold, we adopt the notation

$$V = \frac{d\phi}{d\varepsilon} = \frac{d\phi^1}{d\varepsilon} \cdot \frac{\partial}{\partial x^1} + \cdots + \frac{d\phi^p}{d\varepsilon} \cdot \frac{\partial}{\partial x^p} \tag{2.6}$$

for the vector tangential to C at $x = \phi(\varepsilon)$. The collection of all tangent vectors to all possible curves passing through a given point x in M is called the tangent space to M at x and is denoted by TM. A vector field V on M assigns a tangent vector $V \in TM$ to each point $x \in M$, with V varying smoothly from point to point. In local coordinates, a vector field has the form

$$V = \xi^1(x) \cdot \frac{\partial}{\partial x^1} + \cdots + \xi^p(x) \cdot \frac{\partial}{\partial x^p}, \tag{2.7}$$

where each $\xi^i(x)$ is a smoothing function of x.

If V is a vector field, we denote the parameterized maximal integral curve passing through x in M by $\Psi(\varepsilon, x)$ and call Ψ the flow generated by V. Thus for each x in M, and ε in some interval I_x containing 0, $\Psi(\varepsilon, x)$ is a point on the integral curve passing through x in M. The flow of a vector field has the basic properties

$$\Psi(\delta, \Psi(\varepsilon, x)) = \Psi(\delta + \varepsilon, x) \tag{2.8}$$

for all $\delta, \varepsilon \in R$ such that both sides of equation are defined:

$$\Psi(0, x) = x, \tag{2.9}$$

and

$$\frac{d}{d\varepsilon}\Psi(\varepsilon, x) = V \tag{2.10}$$

for all ε where defined. We see that the flow generated by a vector field is the same as a local group action of the Lie group on the manifold M, often called a "one-parameter group of transformations." The vector field V is called the infinitesimal generator of the action since, by Taylor's theorem, in local coordinates

$$\Psi(\varepsilon, x) = x + \varepsilon\xi(x) + O\left(\varepsilon^2\right), \tag{2.11}$$

where $\xi = \left(\xi^1, \ldots, \xi^p\right)$ are the coefficients of V. The orbits of the one-parameter group action are the maximal integral curves of the vector field V.

Definition 1. *A symmetry group of* (2.1) *is a one-parameter group of transformations G, acting on $X \times U$, such that if $u = f(x)$ is an arbitrary solution of* (2.1) *and $g_\varepsilon \in G$, then $g_\varepsilon \cdot f(x)$ is also a solution of* (2.1).

The infinitesimal generator of a symmetry group is called an infinitesimal symmetry. Infinitesimal generators are used to formulate the conditions for a group G to make it a symmetry group. Working with infinitesimal generators is simple. First we define a prolongation of a vector field. The symmetry group of a system of differential equations is the largest local group of transformations acting on the independent and dependent variables of the system such that it can transform one system solution to another. The main goal of Lie symmetry theory is to determine a useful, systematic, computational method that explicitly determines the symmetry group of any given system of differential equations. The search for the symmetry algebra L of a system of differential equations is best formulated in terms of vector fields acting on the space $X \times U$ of independent and dependent variables. The vector field tells us how the variables x, u transform. We also need to know how the derivatives, that is, $u_x, u_{xx}, \ldots$, transform. This is given by the prolongation of the vector field V. Combining these, we have [13, p. 110, Theorem 2.36] the following.

Theorem 1. *Let*

$$V = \sum_{i=1}^{p} \xi^i(x, u)\,\partial_{x^i} + \sum_{a=1}^{q} \eta_a(x, u)\,\partial_{u^a}$$

be a vector defined on an open subset $M \subset X \times U$. The nth prolongation of the original vector field is the vector field:

$$pr^{(n)}V = V + \sum_{a=1}^{q} \sum_{J}^{\eta_a^J} \left(x, u^{(n)}\right) \frac{\partial}{\partial u_J^a}$$

defined on the corresponding jet space $M^{(n)} \subset X \times U^{(n)}$. *The second summation here is over all (unordered) multiindices* $J = (j_1, j_2, \ldots, j_k)$, *with* $1 \leq j_k \leq p$, $1 \leq k \leq n$. *The coefficient functions* ϕ_a^J *of* $pr^{(n)}V$ *are given by the following formula:*

$$\eta_a^J \left(x, u^{(n)}\right) = D_J \left(\eta_a - \sum_{i=1}^{p} \xi^i u_i^a\right) + \sum_{i=1}^{p} \xi^i u_{J,i}^a,$$

where $u_i^a = \frac{\partial u^a}{\partial x^i}$, $u_{J,i}^a = \frac{\partial u_J^a}{\partial x^i}$, *and* D_J *is the total derivative of* η *with respect to* x^j.

In the following analysis, we deal only with one-dimensional scalar differential equations that are assumed to be differentiable up to the necessary order.

Consider the special case where $p = 2, q = 1$ in the prolongation formula, so that we are looking at a partial differential equation involving the function $u = f(x,t)$. A general vector field on $X \times U \cong R^2 \times R$ then takes the form [13, p. 114]:

$$V = \xi(x,t,u)\frac{\partial}{\partial x} + \tau(x,t,u)\frac{\partial}{\partial t} + \eta(x,t,u)\frac{\partial}{\partial u}. \tag{2.12}$$

The first prolongation of V is the vector field:

$$pr^{(1)}V = V + [\eta_x]\frac{\partial}{\partial u_x} + [\eta_t]\frac{\partial}{\partial u_t}, \tag{2.13}$$

where

$$[\eta_x] = \eta_x + (\eta_u - \xi_x)u_x - \tau_x u_t - \xi_u u_x^2 - \tau_u u_x u_t$$

and

$$[\eta_t] = \eta_t + (\eta_u - \tau_t)u_t - \xi_t u_x - \tau_u u_t^2 - \xi_u u_x u_t.$$

The subscripts on η, ξ, τ denote partial derivatives. Similarly,

$$pr^{(2)}V = pr^{(1)}V + [\eta_{xx}]\frac{\partial}{\partial u_{xx}} + [\eta_{xt}]\frac{\partial}{\partial u_{xt}} + [\eta_{tt}]\frac{\partial}{\partial u_{tt}}, \tag{2.14}$$

where

$$\begin{aligned}
&[\eta_{xx}] = \eta_{xx} + (2\eta_{xu} - \xi_{xx})u_x - \tau_{xx}u_t + (\eta_{uu} - 2\xi_{xu})u_x^2 - 2\tau_{xu}u_x u_t - \xi_{uu}u_x^3 \\
&-\tau_{uu}u_x^2 u_t + (\eta_u - 2\xi_x)u_{xx} - 2\tau_x u_{xt} - 3\xi_u u_{xx}u_x - \tau_u u_{xx}u_t - 2\tau_u u_{xt}u_t, \\
&[\eta_{xt}] = \eta_{xt} + (\eta_{xu} - \tau_{tx})u_t + (\eta_{tu} - \xi_{tx})u_x - \tau_{xu}u_t^2 + (\eta_{uu} - \xi_{xu} - \tau_{ut})u_x u_t \\
&-\xi_{tu}u_x^2 - \tau_{uu}u_x u_t^2 - \xi_{uu}u_t u_x^2 - \tau_x u_{tt} + (\eta_u - \xi_x - \tau_t)u_{xt} - \xi_t u_{xx} - 2\tau_u u_t u_{xt} \\
&-2\xi_u u_x u_{xt} - \tau_u u_x u_{tt} - \xi_u u_t u_{xx}, \\
&[\eta_{tt}] = \eta_{tt} + (2\eta_{tu} - \tau_{tt})u_t - \xi_{tt}u_x + (\eta_{uu} - 2\tau_{tu})u_t^2 - 2\xi_{tu}u_x u_t - \tau_{uu}u_x^3 \\
&-\tau_{uu}u_t^2 u_x + (\eta_u - 2\tau_t)u_{tt} - 2\xi_t u_{xt} - 3\tau_u u_{tt}u_t - \xi_u u_{tt}u_x - 2\xi_u u_{xt}u_t.
\end{aligned}$$

From here on the analysis of difference equations concerns only modified equations from the third prolongation of the vector field. From work in computational fluid dynamics, we know that the right-hand side of the modified equation is written entirely in terms of x derivatives. So investigation can be limited to the terms of the spatial derivatives in the following analysis. The coefficients of the various monomials in the third order partial derivatives of u are given in the following:

$$pr^{(3)}V = pr^{(2)}V + [\eta_{xxx}]\frac{\partial}{\partial u_{xxx}} + [\eta_{xxt}]\frac{\partial}{\partial u_{xxt}} + [\eta_{xtt}]\frac{\partial}{\partial u_{xtt}} + [\eta_{ttt}]\frac{\partial}{\partial u_{ttt}}, \tag{2.15}$$

where

$$
\begin{aligned}
[\eta_{xxx}] &= \eta_{xxx} + (3\eta_{xxu} - \xi_{xxx})u_x - \tau_{xxx}u_t + 3(\eta_{xuu} - \xi_{xxu})u_x^2 - 3\tau_{xxu}u_xu_t \\
&+(\eta_{uuu} - 3\xi_{xuu})(u_x)^3 + 3(\eta_{xu} - \xi_{xx})u_{xx} - 3\tau_{xx}u_{xt} - 3\tau_{xuu}(u_x)^2u_t \\
&+3(\eta_{uu} - 3\xi_{xu})u_xu_{xx} - 3\tau_{xu}u_tu_{xx} - 6\tau_{xu}u_{xt}u_x - 3\tau_xu_{xxt} + (\eta_u - 3\xi_x)u_{xxx} \\
&-\xi_{xxx}(u_x)^4 - 6\xi_{uu}(u_x)^2u_{xx} - 3\tau_{uu}(u_x)^2u_{xt} - \tau_{uuu}(u_x)^3u_t - 3\xi_u(u_{xx})^2 \\
&-3\tau_uu_{xxt}u_x - 3\tau_uu_{xt}u_{xx} - 3\tau_{uu}u_{xx}u_xu_t - 4\xi_uu_{xxx}u_x - \tau_uu_{xxx}u_t.
\end{aligned}
$$

Suppose we are given an nth order system of differential equations or, equivalently, a subvariety of the jet space $M^{(n)} \subset X \times U^{(n)}$. A symmetry group of this system is a local transformation G acting on $M \subset X \times U$ which transforms solutions of the system to other solutions. We can reduce the important infinitesimals condition for a group G to be a symmetry group of a given system of differential equations. The following theorem [13, p. 104, Theorem 2.31] provides the infinitesimal conditions for a group G to be a symmetry group.

THEOREM 2. *Suppose*

$$\Delta_\nu\left(x, u^{(n)}\right) = 0, \quad \nu = 1, 2, \ldots, l,$$

is a system of differential equations of maximal rank defined over $M \subset X \times U$. If G is a local group of transformations acting on M, and

$$pr^{(n)}V \circ \left[\Delta_\nu\left(x, u^{(n)}\right)\right] = 0, \quad \nu = 1, 2, \ldots, l,$$

whenever

$$\Delta_\nu\left(x, u^{(n)}\right) = 0,$$

for every infinitesimal generator V of G, then G is a symmetry group of the system.

In the following sections, this theorem is used to deduce explicitly different infinitesimal conditions for specific problems. It must be remembered, however, that, in all cases, though only the scalar differential problem is being discussed, Δ_ν is still used to denote different differential equations.

3. The group classification of the nonlinear wave equations. The main prototype equation here is the inviscid Burgers equation. Its mathematical properties are very clear, and its group classification can be obtained from previous work. In 1981 Ames, Lohner, and Adams [34] (also see [14, p. 194]) considered the nonlinear wave equation:

$$u_{tt} = \left(c^2(u)\,u_x\right)_x, \tag{3.1}$$

where $c(u) \neq \text{const}$. Viewing this in terms of infinitesimal generators reveals that the group classification of (3.1) is

a. $c(u)$ arbitrary:

$$V_1 = \partial_x, \tag{3.2}$$

$$V_2 = \partial_t, \tag{3.3}$$

$$V_3 = x\partial_x + t\partial_t. \tag{3.4}$$

b. $c(u) = A(u+B)^C$, where A, B, C are arbitrary constants:

$$V_1, V_2, V_3, \tag{3.5}$$

$$V_4 = Cx\partial_x + (u+B)\partial_u. \tag{3.6}$$

c. $c(u) = A(u+B)^{-2}$, where A, B are arbitrary constants:

$$V_1, V_2, V_3, V_4(C=-2), \tag{3.7}$$

$$V_5 = t^2\partial_t + t(u+B)\partial_u. \tag{3.8}$$

d. $c(u) = A(u+B)^{-\frac{2}{3}}$, where A, B are arbitrary constants:

$$V_1, V_2, V_3, V_4\left(C=-\frac{2}{3}\right), \tag{3.9}$$

$$V_5 = x^2\partial_x - 3x(u+B)\partial_u. \tag{3.10}$$

For the inviscid Burgers equation:

$$u_t + uu_x = 0. \tag{3.11}$$

This equation can be rewritten as:

$$u_{tt} = \left(u^2 u_x\right)_x. \tag{3.12}$$

The corresponding wave velocity is then:

$$c(u) = u, \tag{3.13}$$

with the corresponding classification coefficients:

$$A = 1, \tag{3.14}$$

$$B = 0, \tag{3.15}$$

$$C = 1. \tag{3.16}$$

This indicates that there are four spans in the infinitesimal symmetry group of nonlinear wave equation vector fields:

$$V_1 = \partial_x, \tag{3.17}$$

$$V_2 = \partial_t, \tag{3.18}$$

$$V_3 = x\partial_x + t\partial_t, \tag{3.19}$$

$$V_4 = x\partial_x + u\partial_u. \tag{3.20}$$

The prolongation formula up to the third order can then be listed as follows:

$$pr^{(2)}V_1 = \partial_x, \tag{3.21}$$

$$pr^{(2)}V_2 = \partial_t, \tag{3.22}$$

$$pr^{(2)}V_3 = x\partial_x + t\partial_t - u_x\partial_{u_x} - u_t\partial_{u_t} - 2u_{xx}\partial_{u_{xx}} - 2u_{xt}\partial_{u_{xt}} - 2u_{tt}\partial_{u_{tt}}, \tag{3.23}$$

$$pr^{(2)}V_4 = x\partial_x + u\partial_u + u_t\partial_{u_t} - u_{xx}\partial_{u_{xx}} + u_{tt}\partial_{u_{tt}}, \tag{3.24}$$

$$pr^{(3)}V_3 = pr^{(2)}V_3 - 3u_{xxx}\partial_{u_{xxx}}, \tag{3.25}$$

$$pr^{(3)}V_4 = x\partial_x + u\partial_u + u_t\partial_{u_t} - u_{xx}\partial_{u_{xx}} + u_{tt}\partial_{u_{tt}}. \tag{3.26}$$

4. The group classification of the Lax–Wendroff scheme.

4.1. The Lax–Wendroff scheme. In this paper a rigorous systematic approach for the group classification of the Lax–Wendroff scheme is developed. To achieve this the difference approximations to the scalar conservation law

$$u_t + f(u)_x = 0 \tag{4.1}$$

require study. The equation in (4.1) is the simplest one-dimensional nonlinear model exhibiting behavior typical of the equations of inviscid fluid dynamics. The Lax–Wendroff difference approximation to (4.1) is defined by

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + \frac{f(u_{j+1}^n) - f(u_{j-1}^n)}{2\Delta x} = \frac{1}{2}\Delta t \frac{a_{j+\frac{1}{2}}^2(u_{j+1}^n - u_j^n) - a_{j-\frac{1}{2}}^2(u_j^n - u_{j-1}^n)}{\Delta x^2}, \tag{4.2}$$

where

$$\begin{aligned} u_j^n &= u\left(n\Delta t, j\Delta x\right), \\ a(u) &= \frac{\partial f\left(u\right)}{\partial u}, \\ a_{j+\frac{1}{2}} &= a\left(u_{j+\frac{1}{2}}^n\right), \\ u_{j+\frac{1}{2}}^n &= \frac{1}{2}(u_{j+1}^n + u_j^n). \end{aligned}$$

It is well known that the Lax–Wendroff scheme was designed to have the following desirable computational features [31, 32]: conservation of form; to have a three-point scheme; second order accuracy on smooth solutions. Numerical spikes and downstream oscillations are generated in the vicinity of the shock.

4.2. The modified differential equation of Lax–Wendroff scheme. Majda and Osher [32, eq. 1.5] have produced the modification equation:

$$\begin{aligned} u_t = -(f(u))_x + \frac{h^2}{6}\big[&-(f(u))_{xxx} + \lambda^2([a'(u)((f(u))_x)^2]_x \\ &+[a(u)[a(u)(f(u))_x]_x]_x)\big]. \end{aligned} \tag{4.3}$$

Some remarks are essential here:

1. The modified equation retains the conservation law form.

2. Even though the modified equation has been derived by assuming that the exact solution of (4.1) is smooth, numerical evidence indicates that, for shock solutions of (4.1), solution of the difference equation in (4.2) more closely mimics the solution of a modified equation in (4.3) than solutions of the original equation in (4.1).

Setting $f(u) = \frac{1}{2}u^2$, and performing some mechanical calculations, we arrive at the model differential equation for the Lax–Wendroff scheme [32, eq. 1.7]

$$\Delta_1 = u_t + uu_x + RHS = 0, \tag{4.4}$$

where

$$RHS = -h^2\left[\frac{u}{6}(\lambda^2u^2-1)u_{xxx} + \lambda^2uu_x^3 + \frac{1}{4}(3\lambda^2u^2-1)(2u_xu_{xx})\right],$$

$$h = \Delta x, \quad \lambda = \frac{\Delta t}{\Delta x}.$$

4.3. Group invariants of the Lax–Wendroff scheme. Before beginning the group analysis, some detailed but mechanical calculations must be performed:

$$d_1 = \partial_x \circ \Delta_1 = 0, \tag{4.5}$$

$$d_2 = \partial_t \circ \Delta_1 = 0, \tag{4.6}$$

$$d_3 = \partial_u \circ \Delta_1 = u_x - h^2\left[\frac{1}{6}\left(3\lambda^2u^2-1\right)u_{xxx} + \lambda^2u_x^3 + 3\lambda^2uu_xu_{xx}\right], \tag{4.7}$$

$$d_4 = \partial_{u_x} \circ \Delta_1 = u - h^2\left[3\lambda^2uu_x^2 + \frac{1}{2}\left(3\lambda^2u^2-1\right)u_{xx}\right], \tag{4.8}$$

$$d_5 = \partial_{u_t} \circ \Delta_1 = 1, \tag{4.9}$$

$$d_6 = \partial_{u_{xx}} \circ \Delta_1 = -h^2\left[\frac{1}{2}\left(3\lambda^2u^2-1\right)u_x\right], \tag{4.10}$$

$$d_7 = \partial_{u_{xt}} \circ \Delta_1 = 0, \tag{4.11}$$

$$d_8 = \partial_{u_{tt}} \circ \Delta_1 = 0, \tag{4.12}$$

$$d_9 = \partial_{u_{xxx}} \circ \Delta_1 = -h^2\left[\frac{1}{6}\left(\lambda^2u^3-u\right)\right]. \tag{4.13}$$

According to Theorem 2, the system of equations admits a space of operators with the basis V_3 (3.19) if and only if the following equations are satisfied:

$$\Delta_1 = 0, \tag{4.14}$$

$$pr^{(3)}V_3 \circ \Delta_1 = 0, \tag{4.15}$$

From (4.15), we have

$$-u_xd_4 - u_td_5 - 2u_{xx}d_6 - 3u_{xxx}d_9 = 0. \tag{4.16}$$

Expressing the functions $d_l, l = 1, \ldots, 9$, leads to a system of equations equivalent to:

$$\begin{aligned} &-u_x\left\{u - h^2\left[3\lambda^2 u u_x^2 + \frac{1}{2}\left(3\lambda^2 u^2 - 1\right)u_{xx}\right]\right\} - u_t \\ &- 2u_{xx}\left\{-h^2\left[\frac{1}{2}\left(3\lambda^2 u^2 - 1\right)u_x\right]\right\} \\ &- 3u_{xxx}\left\{-h^2\left[\frac{1}{6}\left(\lambda^2 u^3 - u\right)\right]\right\} = 0. \end{aligned} \tag{4.17}$$

Performing some mechanical calculations, we arrive at

$$u_{xxx} = \frac{3}{u}\frac{u_x}{\left(1 - \lambda^2 u^2\right)}\left[(3\lambda^2 u^2 - 1)u_{xx} + 2\lambda^2 u u_x^2\right]. \tag{4.18}$$

The right-hand side of the group invariant equation is written entirely in terms of x derivatives. Under the condition that λ is a small parameter, (4.18) can be recast into

$$u_{xxx} \approx -\frac{3}{u}u_x u_{xx}. \tag{4.19}$$

From the theory of gas dynamics we known that, for the Navier–Stokes shock, there are some inherent constraints among the spatial derivatives: $u_x, u_{xx}, u_{xxx}, \ldots$. In the following tables, the theoretical and group invariant conditions are compared, revealing a conflict.

Analysis at the upstream from the shock:

	u	u_x	u_{xx}	u_{xxx}
Theoretical solution	> 0	< 0	< 0	> 0
Lax–Wendroff scheme	> 0	< 0	< 0	< 0

Obviously, there is a conflict upstream from the shock. The signature of the third spatial derivative from group theory is inconsistent with the theoretical results, showing that this is the result of a lack of invariance in the similarity transformation upstream from the shock.

Analysis downstream from the shock:

	u	u_x	u_{xx}	u_{xxx}
Theoretical solution	> 0	< 0	> 0	> 0
Lax–Wendroff scheme	> 0	< 0	> 0	> 0

There is no conflict downstream from the shock, and the signature of the third spatial derivative from group theory is consistent with the theoretical results. The invariance of the similarity transformation is preserved upstream from the shock.

These group analysis results are consistent with previous numerical results produced by the Lax–Wendroff scheme [31, 33]. They exhibit some level of momentum spikes and postshock oscillations, while the group invariant conditions are broken in this region.

By using the same analysis, we finally obtain an exact group invariant equation for V_4

$$u_{xxx} = -\frac{3}{u}u_x u_{xx}. \tag{4.20}$$

Other results are the same as those for group analysis on the basis of V_3. Some remarks should be made here:

1. Based on the analysis above, we can see that the origin of the spurious wave oscillations behind the shock has some inherent connection with the lack of symmetry. Obviously, there is a conflict upstream from the shock as the signature of the third spatial derivative from group theory is not consistent with the theoretical results. This is a result of the lack of invariance of the similarity transformation upstream from the shock. The Lax–Wendroff scheme exhibits some level of momentum spikes and postshock oscillations, but the group invariant conditions are broken in this region.
2. In the above analysis (4.19), it is assumed that λ is a small parameter. This condition is met in many cases, but this is not generally applicable. It appears that, in general, modification equations do not admit similarity transformations for both cases.

5. Symmetry and Harten's entropy enforcement condition. The approach to development of the shock-capturing methods TVD and essentially nonoscillatory shock-capturing method (ENO) has been primarily mathematical. Most of these schemes do not directly consider the physical reasons for oscillation production in the numerical solutions. It is well known that Harten [2] has proved that some conditions must be met for the coefficient of numerical viscosity, in order for the TVD property to hold, but the general formula of these coefficients cannot be obtained from Harten's analysis. In his celebrated paper [2], he simply presented a heuristic analysis and provided the entropy enforcement formula. In this section, we will address the coefficient of numerical viscosity from the point of view of symmetry.

5.1. TVD scheme and its modified equation. This section studies the difference approximations to the scalar conservation law and discusses numerical approximations to weak solutions of (4.1), which are obtained by $(2k+1)$-point explicit schemes in conservation form:

$$v_j^{n+1} = v_j^n - \lambda\left(\bar{f}_{j+\frac{1}{2}}^n - \bar{f}_{j-\frac{1}{2}}^n\right), \tag{5.1}$$

where

$$\bar{f}_{j+\frac{1}{2}}^n = f\left(v_{j-k+1}^n, \ldots, v_{j+k}^n\right).$$

Here $v_j^n = v\left(j\Delta x, n\Delta t\right)$, and $\bar{f}$ is a numerical flux function. The numerical flux function must be consistent with the flux $f(u)$ in the following sense:

$$\bar{f}(u, \ldots, u) = f(u). \tag{5.2}$$

Let us consider a general three-point finite difference scheme in conservation form with a numerical flux $\bar{f}$ of the form [2, eq. 3.1a]

$$\bar{f}\left(v_j, v_{j+1}\right) = \frac{1}{2}\left[f\left(v_j\right) + f\left(v_{j+1}\right) - \frac{1}{\lambda}Q\left(\lambda\bar{a}_{j+\frac{1}{2}}\right)\Delta_{j+\frac{1}{2}}v\right], \tag{5.3}$$

where

$$\bar{a}_{j+\frac{1}{2}} = \frac{f\left(v_{j+1}\right) - f\left(v_j\right)}{\Delta_{j+\frac{1}{2}}v} \text{ when } \Delta_{j+\frac{1}{2}}v \neq 0,$$

TABLE 5.1
The coefficients of numerical viscosity for different conservation schemes.

Scheme	$Q(x)$
Lax–Friedrichs	$Q(x) = \frac{1}{\lambda}$
Lax–Wendroff	$Q(x) = x^2$
Roe	$Q(x) = \|x\|$
Harten	$Q(x) = \begin{cases} \|x\|, \|x\| \geq \epsilon \\ \frac{x^2+\epsilon^2}{2\epsilon}, \|x\| \leq \epsilon \end{cases}$
Boris–Book	$Q(x) = x^2 + \frac{1}{4}$

$$\bar{a}_{j+\frac{1}{2}} = a\left(v_j\right) \text{ when } \Delta_{j+\frac{1}{2}} v = 0.$$

Here $Q\left(x\right)$ is a function, which is often referred to as the coefficient of numerical viscosity. The modified equation of scheme (5.3), i.e., the equation that it approximates to second order accuracy, is

$$u_t + f\left(u\right)_x = \frac{\Delta t}{2\lambda^2}\left\{\left[Q\left(\nu\right) - \nu^2\right] u_x\right\}_x, \tag{5.4}$$

where

$$\lambda = \frac{\Delta t}{\Delta x},$$
$$\nu = \lambda a\left(u\right).$$

In Table 5.1, some typical coefficients of numerical viscosity are presented, including Harten's formula.

5.2. Group invariants of the TVD scheme. In this section, the modified equation is used instead of the first differential approximation, and the conditions under which a modified equation admits all groups of transformations allowed by the inviscid Burgers equation are formulated.

Setting $f(u) = \frac{1}{2}u^2$, we rewrite the modified equation as

$$u_t + uu_x = \left[D(u)u_x\right]_x, \tag{5.5}$$

where

$$D(u) = \frac{\Delta t}{2\lambda^2}\left[Q(\nu) - \nu^2\right].$$

Hence,

$$\Delta_2 = u_t + uu_x - D'(u)u_x^2 - D(u)u_{xx} = 0. \tag{5.6}$$

Using the same mathematical manipulation presented in section 4, we can obtain two different group invariant conditions as follows:

$$D'(u)u_x^2 + D(u)u_{xx} = 0, \tag{5.7}$$

$$\left[D'(u) - uD''(u)\right]u_x^2 + \left[2D(u) - uD'(u)\right]u_{xx} = 0. \tag{5.8}$$

Equations (5.7) and (5.8) consist of a homogeneous equation system about variables u_x^2, u_{xx}. The nondegenerate conditions for these variables are

$$\begin{vmatrix} D'(u) & D(u) \\ D'(u) - uD''(u) & 2D(u) - uD'(u) \end{vmatrix} = uD(u)D''(u) + D(u)D'(u) - uD'^2(u) = 0. \tag{5.9}$$

This is a nonlinear equation of variable u; the function is $D(u)$. The solution of this equation is

$$D(u) = D_0 u^a, \tag{5.10}$$

where D_0, a are constant parameters.

According to Harten's definition [2, eq. 3.15b], we have

$$Q(\nu) = \nu^2 + \frac{2\lambda^2}{\Delta t} D(u). \tag{5.11}$$

For the present problem,

$$\nu = \lambda u. \tag{5.12}$$

Using the group invariant condition (5.10), we have

$$Q(\nu) = \nu^2 + \frac{2\lambda^2}{\Delta t} D_0 u^a \tag{5.13}$$

or

$$Q(\nu) = \nu^2 + \frac{2\lambda^2}{\Delta t} D_0 \lambda^{-a} \nu^a. \tag{5.13'}$$

In order to obtain a universal formula the coefficients in it must be dimensionalless parameters. Δx and Δt denote, as usual, the dimensions of length and time, respectively. It is clear that D_0 can be rewritten as

$$D_0 = \hat{\varepsilon} (\Delta x)^b (\Delta t)^c, \tag{5.14}$$

where $\hat{\varepsilon}$ is a dimensionless parameter; meanwhile b, c are the exponents that must be chosen.

Finally, we have

$$Q(\nu) = \nu^2 + 2\hat{\varepsilon} \nu^a. \tag{5.15}$$

These dimensional exponents must be met:

$$\begin{aligned} Q_0 &= (\Delta x)^b (\Delta t)^c (\Delta t)^{-1} (\Delta t)^{2-a} (\Delta x)^{a-2} \\ &= (\Delta x)^{a+b-2} (\Delta t)^{c-a+1}. \end{aligned} \tag{5.16}$$

Hence:

$$a + b - 2 = 0, \tag{5.17}$$

$$c - a + 1 = 0. \tag{5.18}$$

That is,

$$b = 2 - a, \tag{5.19}$$

$$c = a - 1. \tag{5.20}$$

These exponents form the universal formula of the coefficients of numerical viscosity, inheriting the symmetries of the PDE. Different schemes could be deduced depending upon the choice of the parameter a.

Another dissipation condition reads as:

$$Q(0) > 0. \tag{5.21}$$

So, the choice of the parameter a is

$$a = 0. \tag{5.22}$$

Finally, we have a definition of the coefficients of numerical viscosity, deduced from the symmetries:

$$Q(x) = x^2 + 2\hat{\varepsilon}. \tag{5.23}$$

THEOREM 3. *If we let $Q(x)$ in (5.3) satisfy the equality*

$$Q(x) = x^2 + 2\hat{\varepsilon},$$

where $\hat{\varepsilon} > 0$ is a dimensionless constant, then the finite different scheme

$$\hat{f}_{j+\frac{1}{2}} = \frac{1}{2}\left[f(u_j^n) + f(u_{j+1}^n) - \frac{1}{\lambda}Q(\lambda \bar{a}_{j+\frac{1}{2}}^n)(u_{j+1}^n - u_j^n)\right]$$

satisfies all of the group invariant conditions of the nonlinear wave equation.

6. The reexploration of the spatial oscillations in finite difference solutions for Navier–Stokes shocks. In the calculation of complex flow fields containing shock waves, much attention has been paid to the shock capturing method [1, 2, 3, 4, 5, 6, 7, 8, 9]. In order to capture shock waves smoothly without spurious data near or in the shock region, the development of high-resolution nonoscillatory dissipative schemes containing no free parameters has been much emphasized. These include TVD schemes and ENO schemes. In [4, 36], Zhang found that, with a linearized analysis of one-dimensional Navier–Stokes equations and a proper manipulation of the coefficients of third order derivative ν_3, the spurious oscillation both upstream and downstream from shocks can be totally suppressed. The choice of ν_3 must be in accordance with the second law of thermodynamics. The numerical experiments with nonlinear Navier–Stokes equations confirm that conclusions reached through linearized analysis are valid in general. The importance of the role played by ν_3 has not been previously recognized [35]. While it has been usually thought that the third order dispersion terms in modified differential equations would not play an essential role in a linear problem, e.g., in stability analysis, etc., they do contribute through nonlinear interactions to the unwanted wave structure and even lead to chaos, if they are not properly treated. Based on the above physical concepts, Zhang suggested the celebrated NND schemes that have been applied to the numerical simulation of Navier–Stokes equations with some success [4, 35].

It is well known that, in application, Navier–Stokes equations are unsteady and nonlinear. The validity of the above linearized analysis is an open question. Difficulty in nonlinear analysis is primarily caused by the nonlinearity inherent in the Navier–Stokes equations, in particular the mechanism that allows the formation of shocks. The theme of this section is to emphasize the role of symmetry in various aspects of a proper manipulation of the coefficients of third order derivative ν_3.

6.1. The model equation and its symmetries. To avoid complicated mathematics, let us consider only the one-dimensional Navier–Stokes shock equation [36]:

$$\rho u_t + \frac{\gamma+1}{2\gamma}\rho_\infty u_\infty \left(1 - \frac{u_2 u_\infty}{u^2}\right) u_x = \frac{4}{3}(\mu u_x)_x, \tag{6.1}$$

where ρ is the mass per unit volume, u is the velocity, t is time, x is the spatial coordinate, γ is the specific heat ratio, and μ is dynamic viscosity.

Here

$$\rho = \frac{\rho_\infty u_\infty}{u},$$
$$\frac{\mu}{\mu_\infty} = \left(\frac{h}{h_\infty}\right)^n,$$
$$h = \frac{1}{2}u_\infty^2\left(1 + \frac{2}{\gamma-1}\frac{1}{M_\infty^2}\right) - \frac{1}{2}u^2.$$

For the propagating shock problem the Dirichlet boundary conditions for u are required up- and downstream from the shock. These are

$$u = u_\infty, \qquad x \to -\infty,$$
$$u_2 = \frac{\gamma-1}{\gamma+1}\left(1 + \frac{2}{\gamma-1}\frac{1}{M_\infty^2}\right) u_\infty. \tag{6.2}$$

The system of equations (6.1) admits a space of operators with the basis:

1. translation in time $V_1 = \partial_t$,
2. translation in a space coordinate $V_2 = \partial_x$,
3. Galileo transformation $V_3 = t\partial_x + \partial_u$,
4. similarity transformation $V_4 = x\partial_x + t\partial_t$,

together with the prolongation vectors as follows, only showing the spatial partial derivatives of u:

$$pr^{(3)}V_1 = \partial_t, \tag{6.3}$$

$$pr^{(3)}V_2 = \partial_x, \tag{6.4}$$

$$pr^{(3)}V_3 = t\partial_x + \partial_u - u_x\partial_{u_t}, \tag{6.5}$$

$$pr^{(3)}V_4 = x\partial_x + t\partial_t - u_x\partial_{u_x} - u_t\partial_{u_t} - 2u_{xx}\partial_{u_{xx}} - 3u_{xxx}\partial_{u_{xxx}}. \tag{6.6}$$

6.2. The modified equation and its symmetries. It can be stated that, generally when the leading term of the truncation error is the third order derivative, the modified equation of the above one-dimensional Navier–Stokes shock is

$$\rho u_t + \frac{\gamma+1}{2\gamma}\rho_\infty u_\infty \left(1 - \frac{u_2 u_\infty}{u^2}\right) u_x = \frac{4}{3}(\mu u_x)_x - (\nu_3 u_{xx})_x. \tag{6.7}$$

The physical viscid coefficient will be taken as constant wherever convenient, and the modified equation could be written as follows:

$$\Delta_3 = u_t + \frac{\gamma+1}{2\gamma}\left(u - \frac{u_2 u_\infty}{u}\right)u_x - \frac{\nu}{\rho_\infty u_\infty}uu_{xx} + \frac{\nu_3}{\rho_\infty u_\infty}uu_{xxx} = 0. \tag{6.8}$$

By using the same mathematical manipulation presented in section 4, we can obtain two different group invariance conditions listed as follows:

(1) For the invariants of the Galileo transformation, the group invariance condition reads

$$\nu_3 u_{xxx} = \nu u_{xx} + \rho_\infty u_\infty \left(\frac{\gamma-1}{2\gamma} - \frac{\gamma+1}{2\gamma}\frac{u_2 u_\infty}{u^2}\right)u_x. \tag{6.9}$$

(2) For the invariants of the similarity transformation, the group invariance condition reads:

$$\nu_3 u_{xxx} = \frac{1}{2}\nu u_{xx}. \tag{6.10}$$

6.3. The choice of ν_3 in accordance with the group invariants. First, considering the invariants of the similarity transformation, the corresponding signature of ν_3 can be calculated as follows.

Equation (6.10) can be recast into

$$\nu_3 = \frac{\nu}{2}\cdot\frac{u_{xx}}{u_{xxx}}. \tag{6.11}$$

Hence, we have

$$signature\,[\nu_3] = signature\,[u_{xx}] \times signature\,[u_{xxx}]\,. \tag{6.12}$$

From (6.12), we have

$$\nu_3 < 0 \tag{6.13}$$

in the upstream region, while

$$\nu_3 > 0 \tag{6.14}$$

in the downstream region.

ν_3 is chosen using the similarity invariance analysis developed above. Now it is obvious that, if we can keep $\nu_3 < 0$ upstream and $\nu_3 > 0$ downstream, the invariants of the similarity transformation condition can be met. This result is equivalent to the former linearized result.

6.4. The Galileo invariants. The invariants of the Galileo transformation are more complex than those of the similarity transformations. In general conditions, there is not a simple relation for the choice of ν_3. In this situation, it is thought that the invariants of the similarity transformations are met, and those for the Galileo transformation are not. Here we focus only on the region near the shock. In this region, we have

$$\frac{\gamma-1}{2\gamma} - \frac{\gamma+1}{2\gamma}\frac{u_2 u_\infty}{u^2} \approx \frac{\gamma-1}{2\gamma} - \frac{\gamma+1}{2\gamma}\frac{u_2}{u_\infty}. \tag{6.15}$$

The speed of shock propagation is the governing parameter for this problem. For a given speed of shock propagation, the boundary value of u_2 is obtained from the Rankine–Hugoniot relation:

$$\frac{u_2}{u_\infty} = \frac{\gamma-1}{\gamma+1}\left(1+\frac{2}{\gamma-1}\frac{1}{M_\infty^2}\right). \tag{6.16}$$

Hence,

$$\begin{aligned}\frac{\gamma-1}{2\gamma} - \frac{\gamma+1}{2\gamma}\frac{u_2 u_\infty}{u^2} &\approx \frac{\gamma-1}{2\gamma} - \frac{\gamma+1}{2\gamma}\frac{u_2}{u_\infty}\\ &= \frac{\gamma-1}{2\gamma} - \frac{\gamma+1}{2\gamma}\frac{\gamma-1}{\gamma+1}\left(1+\frac{2}{\gamma-1}\frac{1}{M_\infty^2}\right)\\ &= \frac{\gamma-1}{2\gamma} - \frac{\gamma-1}{2\gamma} - \frac{1}{\gamma M_\infty^2}\\ &= -\frac{1}{\gamma M_\infty^2}.\end{aligned} \tag{6.17}$$

The Galileo invariants can be written in compact form as:

$$\nu_3 u_{xxx} = \nu u_{xx} - \frac{1}{\gamma M_\infty^2}\rho_\infty u_\infty u_x. \tag{6.18}$$

If $M_\infty >> 1$, then the choice of ν_3 is the same as that of the similarity transformation. But, for the case $M_\infty << 1$,

$$\nu_3 u_{xxx} = -\frac{1}{\gamma M_\infty^2}\rho_\infty u_\infty u_x. \tag{6.19}$$

The Galileo invariants are not met up- or downstream.

From the above consideration, we known that for fast shock propagation the choice of ν_3 based on the linearized analysis is valid under unsteady, nonlinear conditions, but for slow shock propagation, the Galileo invariants are not met in some cases.

It should be remarked that:

> A NND scheme has been suggested based on linear analysis, which has good stability characteristics and convergence. Through the preliminary numerical calculations, confidence was gained in the use of the NND schemes. The principle by which the coefficients of the third order derivatives are chosen is still based on the linearized conclusion. Under unsteady, nonlinear conditions, the validity of this linear conclusion is an open question. The use of symmetries helps us to reach a new conclusion and demonstrate that the linear conclusion is met with the similarity invariants but not met with the Galileo invariants in the general case.

7. Conclusions. It is known that the numerical solutions calculated by finite difference schemes are always associated with numerical dissipation and dispersion. Such errors can lead to undesirable numerical effects, especially for shock capturing. This paper uses a Lie symmetry analysis method to investigate some noteworthy properties of several difference schemes for nonlinear equations. We have demonstrated that the loss of properties of invariance from a modification equation, which conserve the main properties of its difference scheme, can serve as the cause of nonphysical

oscillations that are capable of eventually completely destroying a numerical solution. Further, this points to symmetry breaking as the inherent cause of nonphysical oscillation in the vicinity of shock. Each of these results is vital to the further development of the application of Lie symmetry to the study of shock-capturing methods.

It should be pointed out that the conclusions presented in this paper have preliminary character and demand further study.

Acknowledgments. I appreciate the numerous corrections and very useful comments of the referees. Special thanks to the referee who provided valuable information about the history of this research.

REFERENCES

[1] S. K.Godunov, *A difference scheme for numerical computation of discontinuous solutions of hydrodynamics equations*, Mat. Sb., 47 (1959), pp. 271–306. English translation in U.S. Joint Publications Research Service, JPRS 7226, 1969.

[2] A. Harten, *High resolution schemes for hyperbolic conservation laws*, J. Comput. Phys., 49 (1983), pp. 357–393.

[3] A. Harten, B. Engquist, S. Osher, and S. R. Chakravarthy, *Uniformly high order accurate essentially non-oscillatory schemes*, J. Comput. Phys., 71 (1987), pp. 231–303.

[4] H. Zhang and Z. Fenggan, *NND schemes and their applications to numerical simulation of two- and three-dimensional flows*, Adv. Appl. Mech., 29 (1992), pp. 193–256.

[5] C. W. Shu and S. Osher, *Efficient implementation of essentially non-oscillatory shock capturing schemes*, J. Comput. Phys., 77 (1988), pp. 439–471.

[6] J. J.Quirk, *A contribution to the great Rieamnn solver debate*, Internat. J. Numer. Methods Fluids, 18 (1994), pp. 555–574.

[7] P. Woodward and P. Colella, *The numerical simulation of two-dimensional fluid flow with strong shock*, J. Comput. Phys., 54 (1984), pp. 115–173.

[8] T. W. Roberts, *The behavior of flux difference splitting schemes near slowly moving shock waves*, J. Comput. Phys., 90 (1990), pp. 141–160.

[9] S. Jin and J. G. Liu, *The effects of numerical viscosities*, J. Comput. Phys., 126 (1996), pp. 373–389.

[10] J. P. Boris and D. L. Book, *Flux-corrected transport* I. *SHASTA, a fluid transport algorithm that works*, J. Comput. Phys., 11 (1973), pp. 38–69.

[11] B. van Leer, *Towards the ultimate conservative difference scheme.* II. *Monotonicity and conservation combined in a second order scheme*, J. Comput. Phys., 14 (1974), pp. 361–370.

[12] C. B. Laney, *Computational Gasdynamics*, Cambridge University Press, London, 1998.

[13] P. J. Olver, *Applications of Lie Group to Differential Equations*, GTM 107, Springer-Verlag, New York, 1986.

[14] G. W. Bluman and S. Kumei, *Symmetries and Differential Equations*, Appl. Math. Sci., 81, Springer-Verlag, New York, 1991, pp. 194–195.

[15] J. E. Marsden and T. S. Ratiu, *Introduction to mechanics and symmetry: A basic exposition of classical mechanical systems*, Texta Appl. Math. 17, Springer-Verlag, New York, 1994.

[16] Y. I. Shokin, *The Method of Differential Approximation*, Springer-Verlag, New York, 1983.

[17] V. A. Dorodnitsyn, *Transformation groups in the net space*, J. Sov. Math., 55 (1991), pp. 1490–1517.

[18] N. N. Yanenko and Y. I. Shokin, *On the group classification of difference schemes for systems of equations in gas dynamics*, in Proceedings of the Second International Conference on Numerical Methods in Fluid Dynamics, Lecture Notes in Phys. 8, Springer-Verlag, Berlin, 1971.

[19] S. Maeda, *Canonical structure and symmetries for discrete systems*, Math. Japan, 25 (1980), pp. 405–420

[20] S. Lafortune, S. Tremblay, and P. Winternitz, *Symmetry classification of diatomic molecular chains*, J. Math. Phys., 42 (2001), pp. 5341–5357.

[21] R. Floreanini, J. Negro, L. M. Nieto, and L. Vinet, *Symmetries of the heat equation on a lattice*, Lett. Math. Phys., 36 (1996), pp. 351–355.

[22] R. Floreanini and L. Vinet, *Lie symmetries of finite difference equations*, J. Math. Phys., 36 (1995), pp. 7024–7042.

[23] R. H. Heredero, D. Levi, M. A. Rodriguez, and P. Winternitz, *Lie algebra contractions and symmetries of the Toda hierarch*, J. Phys. A, 33 (2000), pp. 5025–5040.

[24] D. Levi and L. Martina, *Integrable hierarchies of nonlinear difference-different equations and symmetries*, J. Phys. A, 34 (2001), pp. 10357–10368.
[25] V. A. Dorodnitsyn, *Transformation group in a space of difference variables*, J. Sov. Math., 55 (1991), pp. 1490–1517.
[26] V. A. Dorodnitsyn, *Finite difference models entirely inheriting continuous symmetries equations*, Internat. J. Modern Phys. C, 5 (1994), pp. 723–734.
[27] D. Levi, P. S. Tremblay, and P. Winternitz, *Lie point symmetries of difference equations and lattices*, J. Phys. A, 33 (2000), pp. 8507–8542.
[28] D. Levi, P. S. Tremblay, and P. Winternitz, *Lie symmetries of multidimensional difference equations*, J. Phys. A, 34 (2001), pp. 9507–9524.
[29] D. Levi and P. Winternitz, *Lie point symmetries and commuting flows for equations on lattice*, J. Phys. A, 35 (2002), pp. 2249–2262.
[30] P. Winternitz, *Symmetries of discrete system*, lecture presented at the CIMPA Winter School on Discrete Integrable System in February 2003.
[31] P. D. Lax and B. Wendroff, *Systems of conservation laws*, Comm. Pure. Appl. Math., 23 (1960), pp. 217–237.
[32] A. Majda and S. Osher, *A systematic approach for scalar conservation laws*, Numer. Math., 30 (1978), pp. 429–452.
[33] E. L Rubin and S. Z. Burstein, *Difference methods for inviscid and viscous equations of compressible gas*, J. Comput. Phys., l2 (1976), pp. 178–196.
[34] W. F. Ames, R. J. Lohner, and E. Adams, *Group properties of* $u_{tt} = [f(u)u_x]_x$, Internat. J. Non-Linear Mech., 16 (1981), pp. 439–447.
[35] F. Zhuang, *NND schemes and their application to numerical simulation of complex plume flow problems*, in Advances in Fluid Dynamics (A90-45726 20-34), Springer-Verlag, New York, 1989, pp. 45–56.
[36] H. Zhang, *The exploration of the spatial oscillations in finite difference solutions for Navier-Stokes shocks*, Acta Aerodynamica Sinica, 1 (1984), pp. 12–19 (in Chinese).

SIAM J. NUMER. ANAL.
Vol. 46, No. 1, pp. 344–364
© 2008 Society for Industrial and Applied Mathematics

ALTERNATING DIRECTION IMPLICIT ORTHOGONAL SPLINE COLLOCATION METHODS FOR AN EVOLUTION EQUATION WITH A POSITIVE-TYPE MEMORY TERM*

AMIYA KUMAR PANI†, GRAEME FAIRWEATHER‡, AND RYAN I. FERNANDES§

Abstract. New numerical techniques are presented for the solution of a class of linear partial integro-differential equations (PIDEs) with a positive-type memory term in the unit square. In these methods, orthogonal spline collocation (OSC) is used for the spatial discretization, and, for the time stepping, new alternating direction implicit (ADI) methods based on the backward Euler, the Crank–Nicolson, and the second order BDF methods combined with judiciously chosen quadrature rules are considered. The ADI OSC methods are proved to be of optimal accuracy in time and in the L^2 norm in space. Numerical results confirm the predicted convergence rates and also exhibit optimal accuracy in the L^∞ and H^1norms and superconvergence phenomena.

Key words. evolution equation with positive-type memory term, alternating direction implicit method, orthogonal spline collocation, backward Euler method, Crank–Nicolson method, second order backward differentiation formula method, quadrature rules, optimal order convergence.

AMS subject classifications. Primary, 65M12, 65M15, 65M60; Secondary, 45K05

DOI. 10.1137/050634967

1. Introduction.

We formulate and analyze alternating direction implicit (ADI) time-stepping schemes combined with an orthogonal spline collocation (OSC) spatial discretization for solving the evolution equation with a positive-type memory term

$$u_t = \int_0^t a(t-s)\Delta u(s)\,ds + f(x,y,t), \quad (x,y,t) \in \Omega_T \equiv \Omega \times (0,T], \tag{1.1}$$

subject to the initial and boundary conditions

$$u(x,y,0) = u_0(x,y), \quad (x,y) \in \Omega, \tag{1.2}$$

$$u(x,y,t) = 0, \quad (x,y,t) \in \partial\Omega \times (0,T], \tag{1.3}$$

respectively, where Δ is the two-dimensional Laplacian, $\Omega = (0,1)\times(0,1)$ with boundary $\partial\Omega$, and $u_0(x,y)$ and $f(x,y,t)$ are given functions. The kernel a is assumed to be smooth for $t \geq 0$ and to satisfy

$$(-1)^j a^{(j)}(t) \geq 0, \quad t > 0, \quad j = 0,1,2. \tag{1.4}$$

It follows from [18] that the kernel a is positive definite in the sense that, for all $T > 0$,

$$\int_0^T v(s)\left(\int_0^s a(s-\tau)v(\tau)\,d\tau\right) ds \geq 0 \quad \forall v \in C([0,T]).$$

*Received by the editors July 5, 2005; accepted for publication (in revised form) October 9, 2007; published electronically January 23, 2008.
http://www.siam.org/journals/sinum/46-1/63496.html

†Department of Mathematics, Industrial Mathematics Group, Indian Institute of Technology, Bombay, Powai, Mumbai 400076, India (akp@math.iitb.ac.in).

‡Department of Mathematical and Computer Sciences, Colorado School of Mines, Golden, CO 80401-1887 (gfairwea@mines.edu).

§Department of Mathematics, The Petroleum Institute, P.O. Box 2533, Abu Dhabi, UAE (rfernandes@pi.ac.ae).

Equations of the type (1.1) and nonlinear versions of it arise in viscoelasticity and heat conduction in materials with memory; see, for example, [12, 13, 21].

Various numerical methods have been developed for the solution of (1.1) for both smooth and nonsmooth kernels. For the one-dimensional form of (1.1) with a weakly singular kernel, Sanz-Serna [22] employed a spectral method, while a finite difference scheme is used in [11] to discretize the spatial variable. Methods based on an OSC spatial discretization combined with various time-stepping schemes for (1.1) are formulated in [7] and optimal error estimates derived. For a one-dimensional problem, a modified spline collocation method is also considered. In [24], the Laplace transform is used in time with OSC in space. For details of more recent work on the use of transforms for the time discretization, see [16] and the review in [23], which includes a preview of [14]. In [15, 17], finite element Galerkin methods combined with constant and variable time-stepping schemes for (1.1) with both smooth and nonsmooth kernels are discussed. In [19], an H^1-Galerkin mixed finite element method is formulated and analyzed. For a review of earlier work involving finite element methods for solving (1.1), see [4].

In this paper, we approximate the solution u of (1.1)–(1.3) using OSC with C^1 piecewise polynomials of arbitrary degree $r \geq 3$ in each space variable for the spatial discretization. Then the resulting systems of integro-differential equations in the time variable are discretized using ADI versions of the backward Euler method (section 3), the Crank–Nicolson method (section 4), and the second order backward differentiation formula (BDF) scheme (section 5), each combined with a judiciously chosen quadrature rule. For each method, we examine the existence and uniqueness of the OSC approximation, and derive optimal order error estimates in the L^2 norm at each time step, assuming that the solution u of (1.1)–(1.3) is sufficiently smooth. (For a discussion of regularity results under various assumptions on the kernel a and the data, see [15].) In section 6, we present numerical results obtained using the ADI Crank–Nicolson OSC method and the second order ADI BDF OSC scheme with tensor products of piecewise Hermite cubics. These results not only confirm the predicted convergence rates but also exhibit optimal accuracy in the L^∞ and H^1norms and superconvergence phenomena invariably observed in numerical experiments involving OSC methods but as yet unproved. In section 2, we introduce notation and basic lemmas used in the paper. A survey of orthogonal spline collocation methods for partial differential equations and partial integro-differential equations is given in [2]. Recent advances in the development and analysis of ADI OSC methods for parabolic and hyperbolic problems can be found in [3].

2. Preliminaries. Set $I = (0, 1)$, and let $\delta_x = \{x_i\}_{i=0}^{N_x}$ and $\delta_y = \{y_j\}_{j=0}^{N_y}$ be two partitions of $\bar{I}$ such that

$$0 = x_0 < x_1 < \cdots < x_{N_x - 1} < x_{N_x} = 1, \qquad 0 = y_0 < y_1 < \cdots < y_{N_y - 1} < y_{N_y} = 1.$$

Assume that the partition $\delta = \delta_x \otimes \delta_y$ of Ω is quasi-uniform. Let

$$I_i^x = (x_{i-1}, x_i), \quad I_j^y = (y_{j-1}, y_j), \quad I_{ij} = I_i^x \times I_j^y,$$
$$h_i^x = x_i - x_{i-1}, \quad h_j^y = y_j - y_{j-1}$$

for $1 \leq i \leq N_x$, $1 \leq j \leq N_y$, and set

$$h = \max\left(\max_i h_i^x, \max_j h_j^y\right).$$

Let $\mathcal{M}_r(\delta_x)$ be the spaces of piecewise polynomials of degree $\leq r$ defined by

$$\mathcal{M}_r(\delta_x) = \{v \in C^1([0,1]) : v|_{I_i^x} \in P_r,\ 1 \leq i \leq N_x, v(0) = v(1) = 0\},$$

where P_r denotes the set of polynomials of degree $\leq r$. Let $\mathcal{M}_r(\delta_y)$ be defined similarly, and set

$$\mathcal{M}_r(\delta) = \mathcal{M}_r(\delta_x) \otimes \mathcal{M}_r(\delta_y).$$

Then $\dim(\mathcal{M}_r(\delta)) = (r-1)^2 N_x N_y$.

Let $\{\lambda_\ell\}_{\ell=1}^{r-1}$ and $\{\gamma_\ell\}_{\ell=1}^{r-1}$ be the nodes and weights, respectively, of the $(r-1)$-point Gauss quadrature rule on I. Let

$$\mathcal{G}_x = \{\xi_p^x\}_{p=1}^{M_x} \quad \text{and} \quad \mathcal{G}_y = \{\xi_q^y\}_{q=1}^{M_y},$$

where $M_x = (r-1)N_x$ and $M_y = (r-1)N_y$, be the sets of Gauss points in the x- and y-directions, respectively, where

$$\xi_{i,\ell}^x = x_{i-1} + h_i^x \lambda_\ell, \qquad \xi_{j,\ell}^y = y_{j-1} + h_j^y \lambda_\ell$$

for $1 \leq i \leq N_x,\ 1 \leq j \leq N_y,\ 1 \leq \ell \leq r-1$. Let

$$\mathcal{G}_r = \{\xi = (\xi^x, \xi^y) : \xi^x \in \mathcal{G}_x, \xi^y \in \mathcal{G}_y\}$$

be the set of Gauss points in Ω.

For V and W defined on $\mathcal{G}_r$, let $\langle V, W\rangle_x$, $\langle V, W\rangle_y$, $\langle V, W\rangle$, and $|V|_h$ be defined by

$$\langle V, W\rangle_x(y) = \sum_{i=1}^{N_x} h_i^x \sum_{k=1}^{r-1} \gamma_k (VW)(\xi_{i,k}^x, y), \qquad y \in \bar{I}, \tag{2.1}$$

$$\langle V, W\rangle_y(x) = \sum_{j=1}^{N_y} h_j^y \sum_{l=1}^{r-1} \gamma_l (VW)(x, \xi_{j,l}^y), \qquad x \in \bar{I}, \tag{2.2}$$

$$\langle V, W\rangle = \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} h_i^x h_j^y \sum_{k=1}^{r-1} \sum_{l=1}^{r-1} \gamma_k \gamma_l (VW)(\xi_{i,k}^x, \xi_{j,l}^y), \tag{2.3}$$

and

$$|V|_h = \langle V, V\rangle^{\frac{1}{2}}. \tag{2.4}$$

Let $L^p(\Omega)$, $1 \leq p < \infty$, denote the Banach space with norm $\|\cdot\|_{L^p}$, where

$$\|v\|_{L^p} = \left(\int_\Omega |v|^p dxdy\right)^{1/p}, \qquad \|v\|_{L^\infty} = \sup_\Omega |v|.$$

With m a nonnegative integer, let $H^m(\Omega)$ denote the Sobolev space with norm

$$\|v\|_{H^m} = \left(\sum_{l=0}^{m} \sum_{i+j=l} \left\| \frac{\partial^{i+j} v}{\partial x^i \partial y^j} \right\|_{L^2}^2\right)^{1/2}.$$

If X is a normed space with norm $\|\cdot\|_X$, then we define $L^p(X)$ by

$$L^p(X) = \{v : v(\cdot\,,t) \in X,\ t \in [0,T];\ \|v\|_{L^p(X)} < \infty\},$$

where

$$\|v\|_{L^p(X)} = \left(\int_0^T \|v\|_X^p dt\right)^{1/p}, \qquad \|v\|_{L^\infty(X)} = \sup_{0\le t\le T} \|v\|_X.$$

We assume that $\{t_n\}_{n=0}^J$ is a uniform partition of $[0,T]$ such that $t_n = nk$, $k = T/J$, and we introduce the following notation:

$$V^n(\cdot\,) = V(\cdot\,,t_n), \qquad 0 \le n \le J,$$

$$\partial_t V^n = \frac{V^n - V^{n-1}}{k}, \quad V^{n-\frac{1}{2}} = \frac{1}{2}\left[V^n + V^{n-1}\right], \qquad 1 \le n \le J.$$

Also, we denote by C a generic positive constant that is independent of h and k.

In the error analysis, we use the differentiable map $W : [0,T] \to \mathcal{M}_r(\delta)$ defined by

$$\Delta(u - W) = 0 \quad \text{on } \mathcal{G} \times [0,T], \tag{2.5}$$

where u is the solution of (1.1)–(1.3). The following estimates of $u - W$ and its time derivatives are well known [10, 20].

LEMMA 2.1. *If $\partial^\ell u/\partial t^\ell \in L^p(H^{r+3-j})$, $\ell = 0,1$, $j = 0,1$, $p = 2,\infty$, then there exists a constant C, independent of h, such that*

$$\left\|\frac{\partial^\ell (u - W)}{\partial t^\ell}\right\|_{H^j} \le Ch^{r+1-j}\left\|\frac{\partial^\ell u}{\partial t^\ell}\right\|_{H^{r+3-j}}, \quad j = 0,1, \quad \ell = 0,1. \tag{2.6}$$

3. The backward Euler ADI OSC method. Using the composite right rectangle rule,

$$q_n(\phi) = k\sum_{j=1}^n a_{n-j}\phi^j \approx \int_0^{t_n} a(t_n - s)\phi(s)ds, \tag{3.1}$$

we define the backward Euler OSC method for the solution of (1.1) by

$$\partial_t U^n = k\sum_{j=1}^n a_{n-j}\Delta U^j + f^n \quad \text{on } \mathcal{G}, \quad n = 1,\dots,J, \tag{3.2}$$

with given U^0; see [7]. With $E^n = U^n - U^{n-1}$, we may rewrite (3.2) as

$$\begin{aligned} E^n - k^2 a_0 \Delta E^n &= k^2 \sum_{j=1}^{n-1} a_{n-j}\Delta U^j + k^2 a_0 \Delta U^{n-1} + kf^n \quad \text{on } \mathcal{G} \\ &\equiv F^n, \quad n = 1,\dots,J. \end{aligned} \tag{3.3}$$

If we add the term

$$a_0^2 k^4 \frac{\partial^4 E^n}{\partial x^2 \partial y^2}$$

to the left-hand side of (3.3), we obtain the backward Euler ADI OSC method

$$\left[1 - a_0k^2\Delta + a_0^2k^4\frac{\partial^4}{\partial x^2\partial y^2}\right]E^n = F^n \quad \text{on } \mathcal{G}, \quad n = 1, \ldots, J. \tag{3.4}$$

A unique feature of this method, and of the ADI methods of sections 4 and 5, is the use of a term from the quadrature rule in its formulation.

We now rewrite (3.4) in a more familiar ADI form. To this end, let $\{\phi_i\}_{i=1}^{M_x}$ and $\{\psi_j\}_{j=1}^{M_y}$ be bases for $\mathcal{M}_r(\delta_x)$ and $\mathcal{M}_r(\delta_y)$, respectively, set

$$U^n(x,y) = \sum_{i=1}^{M_x}\sum_{j=1}^{M_y}\alpha_{ij}^{(n)}\phi_i(x)\psi_j(y),$$

and let

$$\boldsymbol{\alpha}^{(n)} = [\alpha_{11}^{(n)}, \alpha_{12}^{(n)}, \ldots, \alpha_{1M_y}^{(n)}, \alpha_{21}^{(n)}, \ldots, \alpha_{M_xM_y}^{(n)}]^T.$$

Define

$$\begin{aligned} &A_x = \left(a_{ij}^x\right)_{i,j=1}^{M_x}, \quad a_{ij}^x = -\phi_j''(\xi_i^x), \quad B_x = \left(b_{ij}^x\right)_{i,j=1}^{M_x}, \quad b_{ij}^x = \phi_j(\xi_i^x), \\ &A_y = \left(a_{ij}^y\right)_{i,j=1}^{M_y}, \quad a_{ij}^y = -\psi_j''(\xi_i^y), \quad B_y = \left(b_{ij}^y\right)_{i,j=1}^{M_y}, \quad b_{ij}^y = \psi_j(\xi_i^y), \end{aligned} \tag{3.5}$$

and

$$\mathbf{F}^{(n)} = [F^n(\xi_1^x,\xi_1^y), F^n(\xi_1^x,\xi_2^y), \ldots, F^n(\xi_1^x,\xi_{M_y}^y), F^n(\xi_2^x,\xi_1^y), \ldots, F^n(\xi_{M_x}^x,\xi_{M_y}^y)].$$

Then (3.4) becomes

$$\left[B_x\otimes B_y + a_0k^2(A_x\otimes B_y + B_x\otimes A_y) + a_0^2k^4A_x\otimes A_y\right](\boldsymbol{\alpha}^{(n)} - \boldsymbol{\alpha}^{(n-1)}) = \mathbf{F}^{(n)},$$

or, with $\boldsymbol{\nu}^{(n)} = \boldsymbol{\alpha}^{(n)} - \boldsymbol{\alpha}^{(n-1)}$,

$$\left[(B_x + k^2a_0A_x)\otimes I_{M_y}\right]\left[I_{M_x}\otimes(B_y + k^2a_0A_y)\right]\boldsymbol{\nu}^{(n)} = \mathbf{F}^{(n)}, \tag{3.6}$$

where $\otimes$ denotes the matrix tensor product,[1] and I_{M_x} and I_{M_y} are unit matrices of orders M_x and M_y, respectively. Introducing the auxiliary vector $\hat{\boldsymbol{\nu}}^{(n)}$, (3.6) is equivalent to

$$\begin{aligned} &\left[(B_x + k^2a_0A_x)\otimes I_{M_y}\right]\hat{\boldsymbol{\nu}}^{(n)} = \mathbf{F}^{(n)}, \\ &\left[I_{M_x}\otimes(B_y + k^2a_0A_y)\right]\boldsymbol{\nu}^{(n)} = \hat{\boldsymbol{\nu}}^{(n)}. \end{aligned} \tag{3.7}$$

Thus we determine $\boldsymbol{\nu}^{(n)}$ by solving two sets of independent one-dimensional problems, first

$$(B_x + a_0k^2A_x)\hat{\boldsymbol{\nu}}_j^{(n)} = \mathbf{F}_j^{(n)}, \quad j = 1, \ldots, M_y, \tag{3.8}$$

[1]If the matrix $A = (a_{i,j})$ is $M_A \times N_A$ and B is $M_B \times N_B$, then the matrix $A \otimes B$ is the $M_AM_B \times N_AN_B$ block matrix whose (i,j) block is $a_{i,j}B$.

in the x-direction, where

$$\mathbf{v}_j = [v_{1j}, v_{2j}, \ldots, v_{M_x j}]^T,$$

followed by

$$(B_y + a_0 k^2 A_y)\boldsymbol{\nu}_i^{(n)} = \hat{\boldsymbol{\nu}}_i^{(n)}, \quad i = 1, \ldots, M_x, \tag{3.9}$$

in the y-direction, where

$$\mathbf{v}_i = [v_{i1}, v_{i2}, \ldots, v_{iM_y}]^T.$$

Then

$$\boldsymbol{\alpha}^{(n)} = \boldsymbol{\alpha}^{(n-1)} + \boldsymbol{\nu}^{(n)}.$$

Since $a_0 \geq 0$, the matrices $B_x + k^2 a_0 A_x$ and $B_y + k^2 a_0 A_y$ are nonsingular [5]. Hence, there exists a unique backward Euler ADI OSC approximation. With standard choices of bases for $\mathcal{M}_r(\delta_x)$ and $\mathcal{M}_r(\delta_y)$, the linear systems (3.8) and (3.9) are almost block diagonal and can be solved efficiently using Gauss elimination-based algorithms [8].

An optimal order L^2 error estimate for the backward Euler ADI OSC approximation is given in the following theorem.

THEOREM 3.1. *Let u and $\{U_n\}_{n=0}^J$ denote the solutions of* (1.1)–(1.3) *and* (3.4)*, respectively. Assume that $a \in C^1([0,T])$ and satisfies* (1.4)*. Then*

$$\|u(t_J) - U^J\|_{L^2} \leq C\Bigg([h^{r+1} + k^4 h^{r-1}]\|u\|_{L^\infty(H^{r+3})} + [h^{r+1} + k^4 h^{r-3}]\|u_t\|_{L^1(H^{r+3})}$$
$$+ k\left[\|\Delta u\|_{L^1(L^\infty)} + \|\Delta u_t\|_{L^1(L^\infty)} + \|u_{tt}\|_{L^1(L^\infty)}\right] + k^4 \left\| \frac{\partial^4 u_t}{\partial x^2 \partial y^2} \right\|_{L^1(L^\infty)} \Bigg),$$

provided U^0 is chosen so that

$$\|u_0 - U^0\|_{L^2} \leq C h^{r+1} \|u_0\|_{H^{r+1}}. \tag{3.10}$$

Proof. Let $\zeta^n = U^n - W^n$ and $\eta^n = u^n - W^n$, where W is defined by (2.5). Then

$$u^n - U^n = \eta^n - \zeta^n.$$

Since bounds on η^n are known from Lemma 2.1, it is sufficient to bound ζ^n and then use the triangle inequality to bound $u^n - U^n$.

First note that (3.4) can be written as

$$\partial_t U^n + a_0^2 k^4 \frac{\partial^4 [\partial_t U^n]}{\partial x^2 \partial y^2} = k \sum_{j=1}^n a_{n-j} \Delta U^j + f^n \quad \text{on } \mathcal{G}, \quad n = 1, \ldots, J, \tag{3.11}$$

which is equivalent to

$$\langle \partial_t U^n, v \rangle + \left\langle a_0^2 k^4 \frac{\partial^4 [\partial_t U^n]}{\partial x^2 \partial y^2}, v \right\rangle = k \sum_{j=1}^n a_{n-j} \langle \Delta U^j, v \rangle + \langle f^n, v \rangle, \quad v \in \mathcal{M}_r(\delta). \tag{3.12}$$

Also,

$$\langle \partial_t u^n, v\rangle + \left\langle a_0^2 k^4 \frac{\partial^4[\partial_t u^n]}{\partial x^2 \partial y^2}, v \right\rangle = \langle q_n(\Delta u), v\rangle + \langle f^n, v\rangle + \left\langle (\partial_t u^n - u_t^n) + a_0^2 k^4 \frac{\partial^4[\partial_t u^n]}{\partial x^2 \partial y^2} - \varepsilon_n(\Delta u), v \right\rangle, \quad v \in \mathcal{M}_r(\delta), \tag{3.13}$$

where ε_n is the quadrature error associated with the quadrature rule (3.1), that is,

$$\varepsilon_n(\phi) = q_n(\phi) - \int_0^{t_n} a(t_n - s)\phi(s)ds. \tag{3.14}$$

Then, using (3.12), (3.13), and (2.5) and choosing $v = \zeta^n$, we obtain

$$\langle \partial_t \zeta^n, \zeta^n\rangle + a_0^2 k^4 \left\langle \frac{\partial^4[\partial_t \zeta^n]}{\partial x^2 \partial y^2}, \zeta^n \right\rangle + k \sum_{j=1}^{n} a_{n-j} \left\langle -\Delta \zeta^j, \zeta^n \right\rangle = \left\langle \sigma^n, \zeta^n \right\rangle, \tag{3.15}$$

where $\sigma^n = \sigma_1^n + \sigma_2^n + \sigma_3^n + \sigma_4^n$ with

$$\sigma_1^n = u_t(t_n) - \partial_t u(t_n), \quad \sigma_2^n = -\varepsilon_n(\Delta u), \quad \sigma_3^n = \partial_t \eta^n,$$

and

$$\sigma_4^n = a_0^2 k^4 \frac{\partial^4}{\partial x^2 \partial y^2}[\partial_t \eta^n - \partial_t u(t_n)].$$

Then

$$\langle \partial_t \zeta^n, \zeta^n\rangle \geq \frac{1}{2}\partial_t |\zeta^n|_h^2, \tag{3.16}$$

and

$$\left\langle \frac{\partial^4[\partial_t \zeta^n]}{\partial x^2 \partial y^2}, \zeta^n \right\rangle \geq \frac{1}{2}\partial_t \left\langle \frac{\partial^4 \zeta^n}{\partial x^2 \partial y^2}, \zeta^n \right\rangle. \tag{3.17}$$

If we substitute (3.16) and (3.17) into (3.15), apply the Cauchy–Schwarz inequality to the right-hand-side term, multiply both sides by $2k$, and then sum the resulting expression from $n = 1$ to $n = m$, $1 \leq m \leq J$, we obtain

$$|\zeta^m|_h^2 + a_0^2 k^4 \left\langle \frac{\partial^4 \zeta^m}{\partial x^2 \partial y^2}, \zeta^m \right\rangle + 2k^2 \sum_{n=1}^{m} \sum_{j=1}^{n} a_{n-j} \left\langle -\Delta \zeta^j, \zeta^n \right\rangle \leq |\zeta^0|_h^2 + a_0^2 k^4 \left\langle \frac{\partial^4 \zeta^0}{\partial x^2 \partial y^2}, \zeta^0 \right\rangle + 2k \sum_{n=1}^{m} |\sigma^n|_h |\zeta^n|_h. \tag{3.18}$$

Now, from [9, Lemma 3.4],

$$a_0^2 k^4 \left\langle \frac{\partial^4 \zeta^m}{\partial x^2 \partial y^2}, \zeta^m \right\rangle \geq a_0^2 k^4 \left\| \frac{\partial^2 \zeta^m}{\partial x \partial y} \right\|_{L^2}^2 \geq 0, \quad 1 \leq m \leq J. \tag{3.19}$$

Then if $a \in C([0,T])$ and satisfies (1.4), it follows from [7, Lemma 3.5] that the quadrature rule (3.1) is positive, that is,

$$k \sum_{n=1}^{m} q_n(\phi)\phi^n = k^2 \sum_{n=1}^{m} \sum_{j=1}^{n} a_{n-j} \phi^j \phi^n \geq 0, \quad 1 \leq m \leq J. \tag{3.20}$$

Thus, from [7, Lemma 3.3] and (3.20), it follows that

$$2k^2 \sum_{n=1}^{m} \sum_{j=1}^{n} a_{n-j} \left\langle -\Delta \zeta^j, \zeta^n \right\rangle \geq 0, \quad 1 \leq m \leq J. \tag{3.21}$$

We bound the second term on the right-hand side of (3.18) using the inequality

$$\left\langle \frac{\partial^4 \zeta^0}{\partial x^2 \partial y^2}, \zeta^0 \right\rangle \leq C \|\zeta^0\|_{H^1}^2, \tag{3.22}$$

which is proved in the following way. If we choose a tensor product basis for $\mathcal{M}_r(\delta)$, it suffices to consider the case where $\zeta^0(x,y) = \zeta_1^0(x)\zeta_2^0(y)$. Then, using [5, Lemma 3.1],

$$\left\langle \frac{\partial^4 \zeta^0}{\partial x^2 \partial y^2}, \zeta^0 \right\rangle = \left\langle \frac{\partial^2 \zeta_1^0}{\partial x^2}, \zeta_1^0 \right\rangle_x \left\langle \frac{\partial^2 \zeta_2^0}{\partial y^2}, \zeta_2^0 \right\rangle_y \leq C \|\zeta^0\|_{H^1}^2,$$

as required. Using the inverse inequality [9, equation (2.4b)] and the equivalence of the norms $\|\cdot\|_{L^2}$ and $|\cdot|_h$ on $\mathcal{M}_r(\delta)$ proved in [20], we then obtain

$$a_0^2 k^4 \left\langle \frac{\partial^4 \zeta^0}{\partial x^2 \partial y^2}, \zeta^0 \right\rangle \leq C k^4 h^{-1} \|\zeta^0\|_{H^1} |\zeta^0|_h. \tag{3.23}$$

From (3.18)–(3.23), we obtain

$$\begin{aligned} |\zeta^m|_h^2 &\leq C \left(|\zeta^0|_h^2 + k^4 h^{-1} \|\zeta^0\|_{H^1} |\zeta^0|_h + k \sum_{n=1}^{m} |\sigma^n|_h |\zeta^n|_h \right) \\ &\leq C \left(|\zeta^0|_h + k^4 h^{-1} \|\zeta^0\|_{H^1} + k \sum_{n=1}^{m} |\sigma^n|_h \right) |\zeta^N|_h, \end{aligned}$$

where

$$|\zeta^N|_h = \max_{0 \leq m \leq J} |\zeta^m|_h.$$

Hence

$$|\zeta^N|_h^2 \leq C \left(|\zeta^0|_h + k^4 h^{-1} \|\zeta^0\|_{H^1} + k \sum_{n=1}^{J} |\sigma^n|_h \right) |\zeta^N|_h,$$

from which it follows that

$$|\zeta^J|_h \leq |\zeta^N|_h \leq C \left(|\zeta^0|_h + k^4 h^{-1} \|\zeta^0\|_{H^1} + k \sum_{n=1}^{J} |\sigma^n|_h \right),$$

and

$$\|\zeta^J\|_{L^2} \leq C \left(\|\zeta^0\|_{L^2} + k^4 h^{-1} \|\zeta^0\|_{H^1} + k \sum_{n=1}^{N} |\sigma^n|_h \right), \tag{3.24}$$

on using the equivalence of the norms $\|\cdot\|_{L^2}$ and $|\cdot|_h$ on $\mathcal{M}_r(\delta)$. Since, from (3.10) and (2.6),

$$\|\zeta^0\|_{H^j} \leq \|U^0 - u_0\|_{H^j} + \|u_0 - W^0\|_{H^j} \leq C h^{r+1-j} \|u_0\|_{H^{r+3}},$$

the first and second terms on the right-hand side of (3.24) are bounded by

$$\|\zeta^0\|_{L^2} + k^4 h^{-1} \|\zeta^0\|_{H^1} \le C \left(h^{r+1} + k^4 h^{r-1}\right) \|u_0\|_{H^{r+3}}. \tag{3.25}$$

It remains to estimate the last term on the right-hand side of (3.24). To estimate σ_1^n, we use Taylor's theorem with integral remainder to obtain

$$k \sum_{n=1}^{J} |\sigma_1^n|_h \le Ck \int_0^T \|u_{tt}\|_{L^\infty}\, ds. \tag{3.26}$$

If $a \in C^1([0,T])$, then

$$|\varepsilon_n(\phi)| \le Ck \int_0^{t_n} \left| \frac{\partial}{\partial s}[a(t_n - s)\phi(s)] \right| ds \le Ck \int_0^{t_n} \left(\|\phi\|_{L^\infty} + \|\phi_t\|_{L^\infty}\right) ds, \tag{3.27}$$

from which it follows that

$$|\sigma_2^n|_h = |\varepsilon_n(\Delta u)|_h \le Ck \int_0^{t_n} \left(\|\Delta u\|_{L^\infty} + \|\Delta u_t\|_{L^\infty}\right) ds,$$

and

$$k \sum_{n=1}^{J} |\sigma_2^n|_h \le Ck \int_0^T \left[\|\Delta u\|_{L^\infty} + \|\Delta u_t\|_{L^\infty}\right] ds. \tag{3.28}$$

Since

$$\partial_t \eta^n = \frac{1}{k} \int_{t_{n-1}}^{t_n} \eta_t(s)\, ds,$$

we obtain

$$\begin{aligned} k \sum_{n=1}^{J} |\sigma_3^n|_h \le \int_0^T |\eta_t|_h &\le Ch^2 \int_0^T \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} \|\eta_t\|_{L^\infty(I_{ij})}\, ds \\ &\le Ch^{r+1} \int_0^T \|u_t\|_{H^{r+3}}\, ds \end{aligned} \tag{3.29}$$

on using Lemmas 3.1 and 3.2 in [9]. Similarly,

$$k \sum_{n=1}^{J} \left| \frac{\partial^4 [\partial_t \eta^n]}{\partial x^2 \partial y^2} \right|_h \le Ch^{r-3} \int_0^T \|u_t\|_{H^{r+3}}\, ds,$$

and

$$k \sum_{n=1}^{J} \left| \frac{\partial^4 [\partial_t u^n]}{\partial x^2 \partial y^2} \right|_h \le \int_0^T \left| \frac{\partial^4 [u_t]}{\partial x^2 \partial y^2} \right|_h ds \le \int_0^T \left\| \frac{\partial^4 u_t}{\partial x^2 \partial y^2} \right\|_{L^\infty} ds.$$

Thus we obtain

$$k \sum_{n=1}^{J} |\sigma_4^n|_h \le Ck^4 \int_0^T \left(h^{r-3} \|u_t\|_{H^{r+3}} + \left\| \frac{\partial^4 u_t}{\partial x^2 \partial y^2} \right\|_{L^\infty} \right) ds. \tag{3.30}$$

On substituting (3.25)–(3.30) into (3.24), we obtain

$$\|\zeta^J\|_{L^2} \le C \left(\left[h^{r+1} + k^4 h^{r-1}\right] \|u_0\|_{H^{r+3}} + \left[h^{r+1} + k^4 h^{r-3}\right] \int_0^T \|u_t\|_{H^{r+3}}\, ds \right.$$
$$\left. + k \int_0^T \left[\|\Delta u\|_{L^\infty} + \|\Delta u_t\|_{L^\infty} + \|u_{tt}\|_{L^\infty}\right] ds + k^4 \int_0^T \left\| \frac{\partial^4 u_t}{\partial x^2 \partial y^2} \right\|_{L^\infty} ds \right). \tag{3.31}$$

Now we use the triangle inequality, (3.31), and Lemma 2.1 to complete the proof. ∎

4. The ADI Crank–Nicolson OSC method. In order to obtain higher order accuracy in time, we now consider a method in which the time discretization is done using a Crank–Nicolson-type approach. This method comprises

$$\partial_t U^n = q_{n-1/2}(-\Delta U) + f(t_{n-1/2}) \text{ on } \mathcal{G}, \quad n = 1, \ldots, J, \tag{4.1}$$

where $q_{n-1/2}(\phi)$, a variant of the composite midpoint quadrature rule, is defined by

$$q_{n-1/2}(\phi) = k \sum_{j=1}^{n-1} a(t_{n-1/2} - t_{j-1/2}) \phi^{j-1/2} + \frac{k}{2} a_0 \phi^{n-1/2}$$
$$= k \sum_{j=1}^{n-1} a_{n-j} \phi^{j-1/2} + \frac{k}{2} a_0 \phi^{n-1/2} \approx \int_0^{t_{n-1/2}} a(t_{n-1/2} - s) \phi(s) ds, \tag{4.2}$$

with

$$\phi^{j-1/2} = (\phi^j + \phi^{j-1})/2, \quad j = 1, \ldots, n.$$

Using [17, Lemma 2.1], it is easy to check that the quadrature rule (4.2) is positive; that is, for all ϕ, and $1 \le m \le J$,

$$k \sum_{n=1}^{m} q_{n-1/2}(\phi) \phi^{n-1/2} = k^2 \sum_{n=1}^{m} \left(\sum_{j=1}^{n-1} a_{n-j} \phi^{j-1/2} + \frac{1}{2} a_0 \phi^{n-1/2} \right) \phi^{n-1/2} \ge 0. \tag{4.3}$$

If $E^n = U^n - U^{n-1}$, then (4.1) can be written as

$$E^n - \frac{k^2}{4} a_0 \Delta E^n = F^n \quad \text{on } \mathcal{G}, \quad n = 1, \ldots, J, \tag{4.4}$$

where

$$F^n = k f_{n-1/2} + k^2 \sum_{j=1}^{n-1} a_{n-j} \Delta U^{j-1/2} + \frac{k^2}{2} a_0 \Delta U^{n-1}.$$

If we add the term

$$\frac{k^4}{16} a_0^2 \frac{\partial^4 E^n}{\partial x^2 \partial y^2}$$

to the left-hand side of (4.4), we obtain the ADI Crank–Nicolson OSC method

$$\left[1 - \frac{k^2}{4} a_0 \Delta + \frac{k^4}{16} a_0^2 \frac{\partial^4}{\partial x^2 \partial y^2}\right] E^n = F^n \quad \text{on } \mathcal{G}, \quad n = 1, \ldots, J. \tag{4.5}$$

Thus, using the notation of section 3, we may write (4.5) in the form (3.7) with k^2 replaced by $k^2/4$.

We denote the error associated with the quadrature rule (4.2) by $\varepsilon_{n-1/2}(\phi)$ with

$$\varepsilon_{n-1/2}(\phi) = q_{n-1/2}(\phi) - \int_0^{t_{n-1/2}} a(t_n - s)\phi(s)\, ds.$$

If $a \in C^2([0,T])$, then, for $0 \le t_n \le T$,

$$|\varepsilon_{n-1/2}(\phi)| \le C\left(k\int_{t_{n-1}}^{t_{n-1/2}} [|\phi| + |\phi'|]\, ds + k^2 \sum_{j=1}^{n-1} \int_{t_{j-1}}^{t_j} [|\phi| + |\phi'| + |\phi''|]\, ds\right). \tag{4.6}$$

THEOREM 4.1. *Let u and $\{U_n\}_{n=0}^J$ denote the solutions of* (1.1)–(1.3) *and* (4.5)*, respectively. Assume that $a \in C^2[0,T]$ and satisfies* (1.4)*. Then*

$$\begin{aligned}\|u(t_J) - U^J\|_{L^2} \le C\Big(&[h^{r+1} + k^4 h^{r-1}]\|u\|_{L^\infty(H^{r+3})} + [h^{r+1} + k^4 h^{r-3}]\|u_t\|_{L^1(H^{r+3})} \\ &+ k^2\left[\|\Delta u\|_{L^1(L^\infty)} + \|\Delta u_t\|_{L^1(L^\infty)} + \|\Delta u_{tt}\|_{L^1(L^\infty)} + \|u_{ttt}\|_{L^1(L^\infty)}\right] \\ &+ k^4 \left\|\frac{\partial^4 u_t}{\partial x^2 \partial y^2}\right\|_{L^1(L^\infty)}\Big),\end{aligned}$$

provided U^0 is chosen so that (3.10) *is satisfied.*

Proof. From (4.5), we obtain

$$\partial_t U^n + \frac{k^4}{16} a_0^2 \frac{\partial^4[\partial_t U^n]}{\partial x^2 \partial y^2} + q_{n-1/2}(-\Delta U) = f_{n-1/2} \quad \text{on } \mathcal{G}, \quad n = 1, \ldots, J,$$

which is equivalent to

$$\left\langle \partial_t U^n + \frac{k^4}{16} a_0^2 \frac{\partial^4[\partial_t U^n]}{\partial x^2 \partial y^2}, v \right\rangle + \left\langle q_{n-1/2}(-\Delta U), v\right\rangle = \left\langle f_{n-1/2}, v\right\rangle, \quad v \in \mathcal{M}_r(\delta).$$

With $\zeta^n = U^n - W^n$ and $\eta^n = u^n - W^n$ as before, we have

$$\langle \partial_t \zeta^n, v\rangle + \frac{k^4}{16} a_0^2 \left\langle \frac{\partial^4 \partial_t \zeta^n}{\partial x^2 \partial y^2}, v\right\rangle + \left\langle q_{n-1/2}(-\Delta\zeta), v\right\rangle = \left\langle \sigma^{n-1/2}, v\right\rangle, \tag{4.7}$$

where $\sigma^{n-1/2} = \sigma_1^{n-1/2} + \sigma_2^{n-1/2} + \sigma_3^{n-1/2} + \sigma_4^{n-1/2}$ with

$$\sigma_1^{n-1/2} = u_t(t_{n-1/2}) - \partial_t u^n, \quad \sigma_2^{n-1/2} = -\varepsilon_{n-1/2}(\Delta u), \quad \sigma_3^{n-1/2} = \partial_t \eta_n,$$

and

$$\sigma_4^{n-1/2} = \frac{k^4}{16} a_0^2 \frac{\partial^4[\partial_t \eta^n - \partial_t u^n]}{\partial x^2 \partial y^2}.$$

Choose $v = \zeta^{n-1/2}$ in (4.7). Then

$$\left\langle \partial_t \zeta^n, \zeta^{n-1/2}\right\rangle = \frac{1}{2}\partial_t \left\langle \zeta^n, \zeta^n\right\rangle,$$

and

$$\left\langle \frac{\partial^4 \partial_t \zeta^n}{\partial x^2 \partial y^2}, \zeta^{n-1/2} \right\rangle = \frac{1}{2} \partial_t \left\langle \frac{\partial^4 \zeta^n}{\partial x^2 \partial y^2}, \zeta^n \right\rangle.$$

On substituting these expressions into (4.7), we multiply the resulting equation by $2k$ and then sum from $n = 1, \ldots, m$, $1 \le m \le J$, to obtain

$$\begin{aligned} (4.8) \qquad & |\zeta^m|_h^2 + a_0^2 \frac{k^4}{16} \left\langle \frac{\partial^4 \zeta^m}{\partial x^2 \partial y^2}, \zeta^m \right\rangle + 2k \sum_{n=1}^{m} \left\langle q_{n-1/2}(-\Delta\zeta), \zeta^{n-1/2} \right\rangle \\ & \le |\zeta^0|_h^2 + a_0^2 \frac{k^4}{16} \left\langle \frac{\partial^4 \zeta^0}{\partial x^2 \partial y^2}, \zeta^0 \right\rangle + 2k \sum_{n=1}^{m} |\sigma^{n-1/2}|_h \left(|\zeta^n|_h + |\zeta^{n-1}|_h \right). \end{aligned}$$

Using [9, equation (3.7)] and the positivity property (4.3), it can be shown that

$$\sum_{n=1}^{m} \left\langle q_{n-1/2}(-\Delta\zeta), \zeta^{n-1/2} \right\rangle \ge 0.$$

The second term on the left-hand side of (4.8) is nonnegative from (3.19), and the second term on the right-hand side is bounded as in (3.23). Then we obtain

$$|\zeta^m|_h^2 \le C \left(|\zeta^0|_h^2 + k^4 h^{-1} \|\zeta^0\|_{H^1} |\zeta^0|_h + k \sum_{n=1}^{m} |\sigma^{n-1/2}|_h \left[|\zeta^n|_h + |\zeta^{n-1}|_h \right] \right).$$

With

$$|\zeta^N|_h = \max_{0 \le m \le J} |\zeta^m|_h,$$

we obtain

$$|\zeta^N|_h^2 \le C \left(|\zeta^0|_h + k^4 h^{-1} \|\zeta^0\|_{H^1} + k \sum_{n=1}^{J} |\sigma^{n-1/2}|_h \right) |\zeta^N|_h,$$

and hence

$$(4.9) \qquad |\zeta^J|_h \le |\zeta^N|_h \le C \left(|\zeta^0|_h + k^4 h^{-1} \|\zeta^0\|_{H^1} + k \sum_{n=1}^{J} |\sigma^{n-1/2}|_h \right).$$

As in (3.25), the first two terms on the right-hand side of (4.9) are bounded by

$$(4.10) \qquad C \left(h^{r+1} + k^4 h^{r-1} \right) \|u_0\|_{H^{r+3}}.$$

Thus, it remains to estimate the last term on the right-hand side of (4.9). Using Taylor's theorem with integral remainder, we obtain

$$(4.11) \qquad k \sum_{n=1}^{J} |\sigma_1^{n-1/2}|_h \le C k^2 \int_0^T \|u_{ttt}\|_{L^\infty} \, ds.$$

Then we use the quadrature error (4.6) to obtain

$$\begin{aligned} |\varepsilon_{n-1/2}(\Delta u)|_h \le C \Bigg(& k \int_{t_{n-1}}^{t_{n-1/2}} \left[\|\Delta u\|_{L^\infty} + \|\Delta u_t\|_{L^\infty} \right] ds \\ & + k^2 \sum_{j=1}^{n-1} \int_{t_{j-1}}^{t_j} \left[\|\Delta u\|_{L^\infty} + \|\Delta u_t\|_{L^\infty} + \|\Delta u_{tt}\|_{L^\infty} \right] ds \Bigg), \end{aligned}$$

and hence

$$k\sum_{n=1}^{J}|\sigma_2^{n-1/2}|_h \le Ck^2\int_0^T \left(\|\Delta u\|_{L^\infty}+\|\Delta u_t\|_{L^\infty}+\|\Delta u_{tt}\|_{L^\infty}\right)\,ds. \tag{4.12}$$

The terms involving $\sigma_3^{n-1/2}$ and $\sigma_4^{n-1/2}$ are bounded as in (3.29) and (3.30), respectively. If we substitute these bounds, (4.10)–(4.12), into (4.9) and use the equivalence of the norms $\|\cdot\|_{L^2}$ and $|\cdot|_h$ on $\mathcal{M}_r(\delta)$, we obtain

$$\begin{aligned}\|\zeta^J\|_{L^2} \le C\Bigg(&\left[h^{r+1}+k^4h^{r-1}\right]\|u_0\|_{H^{r+3}}+\left[h^{r+1}+k^4h^{r-3}\right]\int_0^T\|u_t\|_{H^{r+3}}\,ds\\ &+k^2\int_0^T\left[\|\Delta u\|_{L^\infty}+\|\Delta u_t\|_{L^\infty}+\|\Delta u_{tt}\|_{L^\infty}+\|u_{ttt}\|_{L^\infty}\right]ds\\ &+k^4\int_0^T\left\|\frac{\partial^4 u_t}{\partial x^2\partial y^2}\right\|_{L^\infty}ds\Bigg).\end{aligned} \tag{4.13}$$

The triangle inequality, Lemma 2.1, and (4.13) then yield the desired result. ∎

5. The second order ADI BDF OSC scheme. First we define the trapezoidal rule

$$q_n(\phi)=\sum_{j=0}^{n}\omega_{nj}\phi^j \approx \int_0^{t_n} a(t_n-s)\phi(s)\,ds,$$

where $w_{nn}=\frac{1}{2}ka_0$, $w_{n0}=\frac{1}{2}ka_n$, and $w_{nj}=ka_{n-j}, 1\le j\le n-1$. If $a\in C([0,T])$ and (1.4) holds, then the quadrature rule $q_n(\phi)$ is ω_0-positive [15]; that is,

$$k\sum_{n=1}^{m}q_n(\phi)\phi^n \ge -\omega_0(\phi^0)^2, \quad 1\le m\le J, \tag{5.1}$$

where $\omega_0=k^2a_0/4$. Further, if $a\in C^2([0,T])$, then

$$|\varepsilon_n(\phi)|\le Ck^2\int_0^{t_n}\left|\frac{\partial^2}{\partial s^2}[a(t_n-s)\phi(s)]\right|ds, \quad 0\le t_n\le T, \tag{5.2}$$

where

$$\varepsilon_n(\phi)=q_n(\phi)-\int_0^{t_n}a(t_n-s)\phi(s)ds.$$

The second order BDF OSC scheme for the solution of (1.1)–(1.3) consists of determining $\{U^n\}_{n=0}^J$ satisfying

$$\begin{aligned}&\partial_t U^1=q_1(\Delta U)+f^1 \quad \text{on } \mathcal{G},\\ &D_tU^n=q_n(\Delta U)+f^n \quad \text{on } \mathcal{G}, \qquad n=2,\dots,J,\end{aligned} \tag{5.3}$$

where

$$D_t\phi^n=\frac{3\phi^n-4\phi^{n-1}+\phi^{n-2}}{2k}=\frac{3}{2}\partial_t\phi^n-\frac{1}{2}\partial_t\phi^{n-1},$$

and U^0 satisfies (3.10). The second order ADI BDF OSC scheme then takes the form

$$\partial_t U^1 - \omega_{11}\Delta U^1 + k^2\omega_{11}^2 \frac{\partial^4[\partial_t U^1]}{\partial x^2 \partial y^2} = \omega_{10}\Delta U^0 + f^1, \tag{5.4}$$

$$\begin{aligned}\partial_t U^n - \frac{2}{3}\omega_{nn}\Delta U^n + \frac{4}{9}k^2\omega_{nn}^2 \frac{\partial^4[\partial_t U^n]}{\partial x^2 \partial y^2} \\ = \frac{1}{3}\partial_t U^{n-1} + \frac{2}{3}f^n + \frac{2}{3}\sum_{j=0}^{n-1}\omega_{nj}\Delta U^j, \quad n = 2, \dots, J,\end{aligned} \tag{5.5}$$

or

$$\begin{aligned}&\partial_t U^1 + k^2\omega_{11}^2 \frac{\partial^4[\partial_t U^1]}{\partial x^2 \partial y^2} = q_1(\Delta U) + f^1 \quad \text{on } \mathcal{G}, \\ &D_t U^n + \frac{2}{3}k^2\omega_{nn}^2 \frac{\partial^4[\partial_t U^n]}{\partial x^2 \partial y^2} = q_n(\Delta U) + f^n \quad \text{on } \mathcal{G}, \qquad n = 2, \dots, J.\end{aligned} \tag{5.6}$$

THEOREM 5.1. *Let u and $\{U_n\}_{n=0}^J$ denote the solutions of* (1.1)–(1.3) *and* (5.6), *respectively. Assume that $a \in C^2([0,T])$ and satisfies* (1.4). *Then*

$$\begin{aligned}\|u(t_J) - U^J\|_{L^2} \le C\Bigg(&\left[h^{r+1} + kh^r\right]\|u\|_{L^\infty(H^{r+3})} + \left[h^{r+1} + k^4h^{r-3}\right]\|u_t\|_{L^1(H^{r+3})} \\ &+ k^2\left[\|u_{tt}\|_{L^\infty(L^\infty)} + \|u_{ttt}\|_{L^1(L^\infty)} + \|\Delta u\|_{L^1(L^\infty)} + \|\Delta u_t\|_{L^1(L^\infty)} + \|\Delta u_{tt}\|_{L^1(L^\infty)}\right] \\ &+ k^4\left\|\frac{\partial^4 u_t}{\partial x^2 \partial y^2}\right\|_{L^1(L^\infty)}\Bigg),\end{aligned}$$

provided U^0 is chosen so that (3.10) *is satisfied.*

Proof. As before, we write

$$u(t_n) - U^n = [u(t_n) - W^n] - [U^n - W^n] = \eta^n - \zeta^n$$

and estimate ζ^n. Following the approach of earlier sections, it is easy to see that, for $n = 2, \dots, J$,

$$\langle D_t\zeta^n, \zeta^n\rangle + \frac{2}{3}k^2\omega_{nn}^2\left\langle \frac{\partial^4[\partial_t\zeta^n]}{\partial x^2\partial y^2}, \zeta^n\right\rangle + \langle q_n(-\Delta\zeta), \zeta^n\rangle = \langle \sigma^n, \zeta^n\rangle, \tag{5.7}$$

where $\sigma^n = \sigma_1^n + \sigma_2^n + \sigma_3^n + \sigma_4^n$ with

$$\sigma_1^n = u_t(t_n) - D_t u(t_n), \quad \sigma_2^n = -\varepsilon_n(\Delta u), \quad \sigma_3^n = \frac{2}{3}D_t\eta^n,$$

and

$$\sigma_4^n = \frac{2}{3}k^2\omega_{nn}^2\frac{\partial^4}{\partial x^2\partial y^2}[\partial_t\eta^n - \partial_t u(t_n)].$$

Also, for $n = 1$,

$$\left\langle \partial_t\zeta^1, \zeta^1\right\rangle + k^2\omega_{11}^2\left\langle \frac{\partial^4[\partial_t\zeta^1]}{\partial x^2\partial y^2}, \zeta^1\right\rangle + \left\langle q_1(-\Delta\zeta), \zeta^1\right\rangle = \left\langle \sigma^1, \zeta^1\right\rangle, \tag{5.8}$$

where $\sigma^1 = \sigma_1^1 + \sigma_2^1 + \sigma_3^1 + \sigma_4^1$ with

$$\sigma_1^1 = u_t(t_1) - \partial_t u(t_1), \quad \sigma_2^1 = -\varepsilon_1(\Delta u), \quad \sigma_3^1 = \partial_t \eta^1,$$

and

$$\sigma_4^1 = k^2\omega_{11}^2 \frac{\partial^4}{\partial x^2 \partial y^2}[\partial_t \eta^1 - \partial_t u(t_1)].$$

Then on summing (5.7) for $n = 2$ to $n = m$, $1 \le m \le J$, adding (5.8), and multiplying the resulting equation by k,

$$k\left\langle \partial_t \zeta^1, \zeta^1 \right\rangle + k\sum_{n=2}^{m} \langle D_t \zeta^n, \zeta^n \rangle + \frac{k^5}{6} a_0^2 \sum_{n=2}^{m} \left\langle \frac{\partial^4[\partial_t \zeta^n]}{\partial x^2 \partial y^2}, \zeta^n \right\rangle + \frac{k^5}{4} a_0^2 \left\langle \frac{\partial^4[\partial_t \zeta^1]}{\partial x^2 \partial y^2}, \zeta^1 \right\rangle + k\sum_{n=1}^{m} \langle q^n(-\Delta\zeta), \zeta^n \rangle = k\sum_{n=1}^{m} \langle \sigma^n, \zeta^n \rangle . \tag{5.9}$$

With $\delta_k U^n = U^n - U^{n-k},\ k = 1, 2$, we have

$$kD_t \zeta^n = \frac{3}{2}\zeta^n - 2\zeta^{n-1} + \frac{1}{2}\zeta^{n-2} = 2\delta_1 \zeta^n - \frac{1}{2}\delta_2 \zeta^n,$$

and

$$2\langle \delta_k \zeta^n, \zeta^n \rangle = \delta_k |\zeta^n|_h^2 + |\delta_k \zeta^n|_h^2.$$

Using these results, we have

$$k\sum_{n=2}^{m} \langle D_t \zeta^n, \zeta^n \rangle = \sum_{n=2}^{m} \left(\delta_1 |\zeta^n|_h^2 - \frac{1}{4}\delta_2 |\zeta^n|_h^2 \right) + \sum_{n=2}^{m} \left(|\delta_1 \zeta^n|_h^2 - \frac{1}{4}|\delta_2 \zeta^n|_h^2 \right). \tag{5.10}$$

The first term on the right-hand side of (5.10) becomes

$$\sum_{n=2}^{m} \left(\delta_1 |\zeta^n|_h^2 - \frac{1}{4}\delta_2 |\zeta^n|_h^2 \right) = \frac{3}{4}|\zeta^m|_h^2 - \frac{1}{4}|\zeta^{m-1}|_h^2 - \frac{3}{4}|\zeta^1|_h^2 + \frac{1}{4}|\zeta^0|_h^2. \tag{5.11}$$

Since $\delta_2 \zeta^n = \delta_1 \zeta^n + \delta_1 \zeta^{n-1}$ and $(a+b)^2 \le 2(a^2 + b^2), a, b \in \mathcal{R}$, we bound the second term on the right-hand side of (5.10) in the following way:

$$\begin{aligned} \sum_{n=2}^{m} \left(|\delta_1 \zeta^n|_h^2 - \frac{1}{4}|\delta_2 \zeta^n|_h^2 \right) &\ge \sum_{n=2}^{m} \left(|\delta_1 \zeta^n|_h^2 - \frac{1}{4}\left[|\delta_1 \zeta^n|_h + |\delta_1 \zeta^{n-1}|_h^2 \right]^2 \right) \\ &\ge \sum_{n=2}^{m} \left(|\delta_1 \zeta^n|_h^2 - \frac{1}{2}\left[|\delta_1 \zeta^n|_h^2 + |\delta_1 \zeta^{n-1}|_h \right]^2 \right) \\ &= \frac{1}{2}\sum_{n=2}^{m} \left(|\delta_1 \zeta^n|_h^2 - |\delta_1 \zeta^{n-1}|_h^2 \right) \\ &= \frac{1}{2}\left(|\delta_1 \zeta^m|_h^2 - |\delta_1 \zeta^1|_h^2 \right). \end{aligned} \tag{5.12}$$

Also,

$$k\left\langle \partial_t \zeta^1, \zeta^1 \right\rangle = \frac{1}{2}\left[|\zeta^1|_h^2 - |\zeta^0|_h^2 + |\delta_1 \zeta^1|_h^2 \right]. \tag{5.13}$$

Combining (5.10)–(5.13), we obtain

$$
\begin{aligned}
(5.14)\qquad & k\left\langle \partial_t\zeta^1,\zeta^1\right\rangle + k\sum_{n=2}^{m}\left\langle D_t\zeta^n,\zeta^n\right\rangle \geq \frac{1}{2}\left[|\zeta^1|_h^2-|\zeta^0|_h^2+|\delta_1\zeta^1|_h^2\right]\\
&+\left(\frac{3}{4}|\zeta^m|_h^2-\frac{1}{4}|\zeta^{m-1}|_h^2-\frac{3}{4}|\zeta^1|_h^2+\frac{1}{4}|\zeta^0|_h^2\right)+\frac{1}{2}\left(|\delta_1\zeta^m|_h^2-|\delta_1\zeta^1|_h^2\right)\\
&\geq \frac{3}{4}|\zeta^m|_h^2-\frac{1}{4}\left(|\zeta^{m-1}|_h^2+|\zeta^1|_h^2+|\zeta^0|_h^2\right).
\end{aligned}
$$

Now, on using [9, equation (6.37)] and (5.1),

$$
\begin{aligned}
(5.15)\qquad k\sum_{n=1}^{m}\left\langle q^n(-\Delta\zeta),\zeta^n\right\rangle &= k\sum_{n=1}^{m}\sum_{j=0}^{n}\omega_{nj}\left\langle -\Delta\zeta^j,\zeta^n\right\rangle\\
&\geq -\omega_0^2\left\langle -\Delta\zeta^0,\zeta^0\right\rangle \geq -Ck^4\left\|\zeta^0\right\|_{H^1}^2,
\end{aligned}
$$

since $\omega_0 = k^2a_0/4$ and, from [9, Lemma 3.2],

$$\left\langle -\Delta\zeta^0,\zeta^0\right\rangle \leq C\|\zeta^0\|_{H^1}^2.$$

Multiplying (5.9) by 4/3 and using (3.17), (5.14), and (5.15), we obtain

$$
\begin{aligned}
(5.16)\qquad & |\zeta^m|_h^2+\frac{k^4}{9}a_0^2\left\langle\frac{\partial^4\zeta^m}{\partial x^2\partial y^2},\zeta^m\right\rangle+\frac{k^4}{18}a_0^2\left\langle\frac{\partial^4\zeta^1}{\partial x^2\partial y^2},\zeta^1\right\rangle\\
&\leq \frac{1}{3}\left[|\zeta^{m-1}|_h^2+|\zeta^1|_h^2+|\zeta^0|_h^2\right]\\
&\quad+\frac{4}{3}k\sum_{n=1}^{m}|\sigma^n|_h\,|\zeta^n|_h+Ck^4\|\zeta^0\|_{H^1}^2+\frac{k^4}{6}a_0^2\left\langle\frac{\partial^4\zeta^0}{\partial x^2\partial y^2},\zeta^0\right\rangle.
\end{aligned}
$$

The second and third terms on the left-hand side are positive from (3.19). Setting

$$|\zeta^N|_h=\max_{0\leq m\leq J}|\zeta^m|_h,$$

we have

$$
\begin{aligned}
(5.17)\qquad |\zeta^N|_h^2 &\leq \frac{1}{3}|\zeta^N|_h^2+\frac{1}{3}\left[|\zeta^1|_h+|\zeta^0|_h+4k\sum_{n=1}^{J}|\sigma^n|_h\right]|\zeta^N|_h\\
&\quad+Ck^4\|\zeta^0\|_{H_1}^2+\frac{k^4}{6}a_0^2\left\langle\frac{\partial^4\zeta^0}{\partial x^2\partial y^2},\zeta^0\right\rangle.
\end{aligned}
$$

Thus, using (3.22), we obtain

$$
\begin{aligned}
(5.18)\quad |\zeta^N|_h^2 &\leq \left(\frac{1}{2}\left[|\zeta^1|_h+|\zeta^0|_h+4k\sum_{n=1}^{J}|\sigma^n|_h\right]\right)|\zeta^N|_h+Ck^4\|\zeta^0\|_{H^1}^2\\
&\leq \frac{1}{2}\left(\frac{1}{2}\left[|\zeta^1|_h+|\zeta^0|_h+4k\sum_{n=1}^{J}|\sigma^n|_h\right]\right)^2+Ck^4\|\zeta^0\|_{H^1}^2+\frac{1}{2}|\zeta^N|_h^2,
\end{aligned}
$$

so that

$$(5.19)\qquad |\zeta^J|_h\leq|\zeta^N|_h\leq\frac{1}{2}\left(|\zeta^1|_h+|\zeta^0|_h\right)+2k\sum_{n=1}^{J}|\sigma^n|_h+Ck^2\|\zeta^0\|_{H^1}.$$

From (5.8),

$$\left\langle \zeta^1-\zeta^0,\zeta^1\right\rangle+\frac{k^4}{4}a_0^2\left\langle \frac{\partial^4(\zeta^1-\zeta^0)}{\partial x^2\partial y^2},\zeta^1\right\rangle+k\left\langle q_1(-\Delta\zeta^1),\zeta^1\right\rangle=k\left\langle \sigma^1,\zeta^1\right\rangle.$$

Using (3.16) and (3.17) in the first and second terms, respectively, and bounding the third term as in (5.15), we obtain

$$|\zeta^1|_h^2\le C\left(|\zeta^0|_h^2+k^4\left\langle \frac{\partial^4\zeta^0}{\partial x^2\partial y^2},\zeta^0\right\rangle+k^4\|\zeta^0\|_{H^1}^2+k^2|\sigma^1|_h^2\right).$$

Thus, using (3.22),

$$|\zeta^1|_h\le C\left(|\zeta^0|_h+k^2\|\zeta^0\|_{H^1}+k|\sigma^1|_h\right).$$

On substitution into (5.19), it follows that

$$|\zeta^J|_h\le C\left(|\zeta^0|_h+k^2\|\zeta^0\|_{H^1}+Ck\sum_{n=1}^{J}|\sigma^n|_h\right). \tag{5.20}$$

To complete the proof, we estimate the term

$$k\sum_{n=1}^{J}|\sigma^n|_h.$$

First,

$$k\sum_{n=2}^{J}|\sigma_1^n|_h\le Ck^2\int_k^T\|u_{ttt}\|_{L^\infty}\,ds,$$

and

$$k|\sigma_1^1|_h\le Ck\int_0^k\|u_{tt}\|_{L^\infty}\,ds.$$

Since $a\in C^2([0,T])$, we obtain

$$k\sum_{n=1}^{J}|\sigma_2^n|_h\le Ck^2\int_0^T\left[\|\Delta u\|_{L^\infty}+\|\Delta u_t\|_{L^\infty}+\|\Delta u_{tt}\|_{L^\infty}\right]ds.$$

Following the argument in (3.29), we have

$$k\sum_{n=2}^{J}|\sigma_3^n|_h\le Ch^{r+1}\int_k^T\|u_t\|_{H^{r+3}}\,ds,$$

and

$$k|\sigma_3^1|_h\le Ch^{r+1}\int_0^k\|u_t\|_{H^{r+3}}\,ds.$$

The term

$$k\sum_{n=1}^{J}|\sigma_4^n|_h$$

is bounded as in (3.30). Thus, substituting into (5.20) and using the equivalence of the norms $\|\cdot\|_{L^2}$ and $|\cdot|_h$ on $\mathcal{M}_r(\delta)$, we obtain

$$\|\zeta^J\|_{L^2} \le C\left(\left[h^{r+1}+kh^r\right]\|u_0\|_{H^{r+3}} + \left[h^{r+1}+k^4h^{r-3}\right]\int_0^T \|u_t\|_{H^{r+3}}\,ds \right. \tag{5.21}$$
$$+ k\int_0^k \|u_{tt}\|_{L^\infty}\,ds + k^2\int_0^T \left[\|u_{ttt}\|_{L^\infty} + \|\Delta u\|_{L^\infty} + \|\Delta u_t\|_{L^\infty} + \|\Delta u_{tt}\|_{L^\infty}\right] ds$$
$$\left. + k^4\int_0^T \left\|\frac{\partial^4 u_t}{\partial x^2\partial y^2}\right\|_{L^\infty}\right).$$

An application of the triangle inequality with Lemma 2.1 and (5.21) completes the proof. ∎

6. Numerical results. We present numerical results obtained using the ADI Crank–Nicolson OSC scheme (4.5) and the second order ADI BDF OSC scheme (5.4)–(5.5). In our test problem, the kernel $a(t) = e^{-t}$ and the forcing function $f(x, y, t)$ is chosen so that

$$u(x, y, t) = e^{xy+t}xy(x-1)(y-1)$$

is the exact solution of the partial integro-differential equation (PIDE) problem.

In the implementation of the ADI methods, we use the space of tensor products of piecewise Hermite cubics, $\mathcal{M}_3(\delta)$, with standard value and scaled slope basis functions defined on uniform partitions $\{x_i\}_{i=0}^{N_x^{(l)}} = \{y_j\}_{j=0}^{N_y^{(l)}}$ of $[0, 1]$ with $N_x^{(l)} = N_y^{(l)} = N_l$; see [6]. We present spatial errors in various norms and the corresponding convergence rates determined by the formula

$$\text{Convergence rate} \approx \frac{\log(\varepsilon_l/\varepsilon_{l+1})}{\log(h_l/h_{l+1})},$$

where the stepsize $h_l = 1/N_l$ and ε_l is the norm of the error with $h = h_l$.

The initial condition can be approximated using either the Gauss or the Hermite interpolant of u_0; the results presented were obtained using the Gauss interpolant. We select $k_l = h_l^2$ ($J_l = N_l^2$) to verify the second order accuracy in t and fourth order accuracy in space for both methods.

The results for the ADI Crank–Nicolson OSC method and the second order ADI BDF OSC scheme are presented in Tables 6.1–6.3 and Tables 6.4–6.6, respectively, at $T = 1$. Specifically, the L^∞ and L^2 errors and the corresponding rates of convergence for the respective schemes are shown in Tables 6.1 and 6.4.

In Tables 6.2 and 6.5, the L^∞ errors in u_x and u_y, the H^1 errors, and the corresponding rates of convergence for the two schemes are given. The L^∞ errors are estimated by taking the maximum error at 10 by 10 points in each subrectangle $[x_{i-1}, x_i] \times [y_{j-1}, y_j]$. Finally, in Tables 6.3 and 6.6, the maximum absolute value errors in u, u_x, and u_y at the partition nodes together with the corresponding rates of convergence for the two schemes are presented.

The numerical results shown in Tables 6.1 and 6.3 indicate that the rate of convergence corresponding to the L^2 error is consistent with the theory for the two schemes (cf. Theorems 4.1 and 5.1), where we have taken $k = h^2$. One would also expect the rate of convergence using the H^1 error to be 3 for both schemes. This can be observed in the last columns of Tables 6.2 and 6.5.

TABLE 6.1
ADI Crank–Nicolson OSC scheme L^∞, L^2 errors in u.

N_l	L^∞ error	Rate	L^2 error	Rate
5	0.650–04		0.191–04	
10	0.449–05	3.854	0.119–05	3.997
15	0.881–06	4.016	0.235–06	4.008
20	0.279–06	3.989	0.745–07	3.995

TABLE 6.2
ADI Crank–Nicolson OSC scheme L^∞ errors in u_x, u_y; H^1 error in u.

N_l	L^∞ error in u_x	Rate	L^∞ error in u_y	Rate	H^1 error	Rate
5	0.694–03		0.694–03		0.233–03	
10	0.900–04	2.946	0.900–04	2.946	0.273–04	3.091
15	0.253–04	3.127	0.253–04	3.127	0.788–05	3.065
20	0.103–04	3.113	0.103–04	3.113	0.331–05	3.012

TABLE 6.3
ADI Crank–Nicolson OSC scheme nodal errors in u, u_x, u_y.

N_l	Nodal error in u	Rate	Nodal error in u_x	Rate	Nodal error in u_y	Rate
5	0.634–04		0.668–03		0.668–03	
10	0.449–05	3.820	0.396–04	4.077	0.396–04	4.077
15	0.881–06	4.016	0.861–05	3.762	0.861–05	3.762
20	0.280–16	3.989	0.254–05	4.244	0.254–05	4.244

TABLE 6.4
ADI BDF OSC scheme L^∞, L^2 errors in u.

N_l	L^∞ error	Rate	L^2 error	Rate
5	0.149–03		0.740–04	
10	0.883–05	4.074	0.436–05	4.086
15	0.174–05	4.007	0.851–06	4.028
20	0.550–06	4.002	0.268–06	4.013

TABLE 6.5
ADI BDF OSC scheme L^∞ errors in u_x, u_y; H^1 error in u.

N_l	L^∞ error in u_x	Rate	L^∞ error in u_y	Rate	H^1 error	Rate
5	0.836–03		0.836–03		0.393–03	
10	0.973–04	3.103	0.973–04	3.103	0.328–04	3.582
15	0.270–04	3.157	0.270–04	3.157	0.868–05	3.279
20	0.109–04	3.162	0.109–04	3.162	0.351–05	3.153

TABLE 6.6
ADI BDF OSC scheme nodal errors in u, u_x, u_y.

N_l	Nodal error in u	Rate	Nodal error in u_x	Rate	Nodal error in u_y	Rate
5	0.148–03		0.826–03		0.826–03	
10	0.883–05	4.072	0.501–04	4.042	0.501–04	4.042
15	0.174–05	4.006	0.102–04	3.927	0.102–04	3.927
20	0.550–06	4.001	0.324–05	3.990	0.324–05	3.990

For each scheme, the rates of convergence in u, u_x, and u_y when using the L^∞ error are $4, 3$, and 3, respectively (cf. Tables 6.1 and 6.4 and 6.2 and 6.5).

The results in Tables 6.3 and 6.6 demonstrate that, for each scheme, the phenomenon of superconvergence is observed; that is, the rates of convergence in the maximum norm error at the nodes in u_x and u_y are approximately 4 in each case.

It should be noted that, throughout, the errors in both methods are very similar.

7. Concluding remarks. We have formulated and analyzed ADI OSC methods for the solution of a PIDE with a positive-type memory term under smoothness conditions which yield optimal error estimates. The analysis in the case of nonsmooth problems is a topic for future study. Numerical results support the analysis and also exhibit optimal convergence and superconvergence phenomena not proved in the present paper. In fact, while superconvergence is frequently observed in numerical implementations of OSC methods, very little analysis has been produced to justify it; see, for example, [1].

In this paper, the problem is linear, and the spatial domain is assumed to be the unit square. Using the approach of [3], it may be possible to extend the present ADI methods to nonlinear equations in rectangular polygons.

REFERENCES

[1] B. Bialecki, *Superconvergence of orthogonal spline collocation in the solution of Poisson's equation*, Numer. Methods Partial Differential Equations, 15 (1999), pp. 285–303.

[2] B. Bialecki and G. Fairweather, *Orthogonal spline collocation methods for partial differential equations*, J. Comput. Appl. Math., 128 (2001), pp. 55–82.

[3] B. Bialecki and R. I. Fernandes, *An alternating-direction implicit orthogonal spline collocation scheme for nonlinear parabolic problems on rectangular polygons*, SIAM J. Sci. Comput., 28 (2006), pp. 1054–1077.

[4] C. Chen and T. Shih, *Finite Element Methods for Integrodifferential Equations*, World Scientific, Singapore, 1998.

[5] J. Douglas, Jr., and T. Dupont, *Collocation Methods for Parabolic Equations in a Single Space Variable*, Lecture Notes in Math. 385, Springer-Verlag, New York, 1974.

[6] G. Fairweather, *Finite Element Galerkin Methods for Differential Equations*, Lecture Notes in Pure and Appl. Math. 34, Marcel Dekker, New York, 1978.

[7] G. Fairweather, *Spline collocation methods for a class of hyperbolic partial integro-differential equations*, SIAM J. Numer. Anal., 31 (1994), pp. 444–460.

[8] G. Fairweather and I. Gladwell, *Algorithms for almost block diagonal linear systems,* SIAM Rev., 46 (2004), pp. 49–58.

[9] R. I. Fernandes and G. Fairweather, *Analysis of alternating direction collocation methods for parabolic and hyperbolic problems in two space variables*, Numer. Methods Partial Differential Equations, 9 (1993), pp. 191–211.

[10] C. E. Greenwell-Yanik and G. Fairweather, *Analysis of spline collocation methods for parabolic and hyperbolic problems in two space variables*, SIAM J. Numer. Anal., 23 (1986), pp. 282–296.

[11] J. C. López-Marcos, *A difference scheme for a nonlinear partial integrodifferential equation*, SIAM J. Numer. Anal., 27 (1990), pp. 20–31.

[12] R. C. MacCamy, *An integro-differential equation with application in heat flow*, Quart. Appl. Math., 35 (1977), pp. 1–19.

[13] R. C. MacCamy, *A model for one-dimensional nonlinear viscoelasticity*, Quart. Appl. Math., 35 (1977), pp. 21–33.

[14] W. McLean, I. H. Sloan, and V. Thomée, *Time discretization via Laplace transformation of an integro-differential equation of parabolic type*, Numer. Math., 102 (2006), pp. 497–522.

[15] W. McLean and V. Thomée, *Numerical solution of an evolution equation with a positive–type memory term*, J. Austral. Math. Soc. Ser. B., 35 (1993), pp. 23–70.

[16] W. McLean and V. Thomée, *Time discretization of an evolution equation via Laplace transforms*, IMA J. Numer. Anal., 24 (2004), pp. 439–463.

[17] W. McLean, V. Thomée, and L. B. Wahlbin, *Discretization with variable time steps of an evolution equation with a positive type memory term*, J. Comput. Appl. Math., 69 (1996), pp. 49–69.

[18] J. H. Nohel and D. F. Shea, *Frequency domain methods for Volterra equations*, Advances in Math., 22 (1976), pp. 278–304.

[19] A. K. Pani and G. Fairweather, *An H^1-Galerkin mixed finite element method for an evolution equation with a positive-type memory term*, SIAM J. Numer. Anal., 40 (2002), pp. 1475–1490.

[20] P. Percell and M. F. Wheeler, *A C^1 finite element collocation method for elliptic equations*, SIAM J. Numer. Anal., 17 (1980), pp. 605–622.

[21] M. Renardy, W. Hrusa, and J. Nohel, *Mathematical Problems in Viscoelasticity*, Pitman Monogr. Surveys Pure Appl. Math. 35, Wiley, New York, 1987.

[22] J. M. Sanz-Serna, *A numerical method for a partial integro-differential equation*, SIAM J. Numer. Anal., 25 (1988), pp. 319–327.

[23] V. Thomée, *A high order parallel method for time discretization of parabolic type equations based on Laplace transformation and quadrature*, Int. J. Numer. Anal. Model., 2 (2005), pp. 85–96.

[24] Y. Yan and G. Fairweather, *Orthogonal spline collocation methods for some partial integrodifferential equations*, SIAM J. Numer. Anal., 29 (1992), pp. 755–768.

SIAM J. NUMER. ANAL.
Vol. 46, No. 1, pp. 365–396
© 2008 Society for Industrial and Applied Mathematics

EXPONENTIALLY CONVERGENT DUHAMEL-LIKE ALGORITHMS FOR DIFFERENTIAL EQUATIONS WITH AN OPERATOR COEFFICIENT POSSESSING A VARIABLE DOMAIN IN A BANACH SPACE*

T. YU. BOHONOVA†, I. P. GAVRILYUK‡, V. L. MAKAROV§, AND V. VASYLYK§

Abstract. A suitable abstract setting of an initial value problem for the first order operator differential equation is introduced for the case when both the unbounded operator coefficient and its domain can depend on the differentiation variable t. A new exponentially convergent algorithm is proposed. The algorithm is based on a generalization of the Duhamel integral for vector-valued functions. The Duhamel-like technique makes it possible to translate the initial problem to an integral equation and then approximate it with exponential accuracy. The theoretical results are illustrated by examples associated with the heat transfer boundary value problems.

Key words. first order differential equations in Banach space, operator coefficient with a variable domain, Duhamel's integral, operator exponential, exponentially convergent algorithms

AMS subject classifications. 65J10, 65M12, 65M15, 46N20, 46N40, 47N20, 47N40

DOI. 10.1137/06065252X

1. Introduction.

This paper deals with a special class of parabolic partial differential equations with time-dependent boundary conditions in an abstract setting, which are associated with the first order differential equation in a Banach space X

$$\frac{du(t)}{dt} + A(t)u(t) = f(t), \quad u(0) = u_0. \tag{1.1}$$

Here t is a real variable, the unknown function $u(t)$ and the given function $f(t)$ take values in X, and $A(t)$ is a given function whose values are densely defined, closed linear operators in X with domains $D(A,t)$ depending on the parameter t.

In some special cases it is possible to translate problem (1.1) to an operator differential equation, in which operator coefficient $\tilde{A}$ possesses domain $D(\tilde{A})$ independent of t. For example, let us consider the problem

$$\begin{aligned} &\frac{\partial u(x,t)}{\partial t} = \frac{\partial^2 u}{\partial x^2} + f(x,t), \\ &u_x(0,t) - \alpha(t)u(0,t) = \phi_1(t), \quad u_x(1,t) + \beta(t)u(1,t) = \phi_2(t), \\ &u(x,0) = u_0(x), \end{aligned} \tag{1.2}$$

where f, α, β, ϕ_1, ϕ_2, and u_0 are given smooth functions. This problem belongs to

*Received by the editors February 21, 2006; accepted for publication (in revised form) July 19, 2007; published electronically January 25, 2008. This work was supported in part by the German Research Council (DFG).

http://www.siam.org/journals/sinum/46-1/65252.html

†National Aviation University of Ukraine, 1, Komarov Ave., 03058 Kyiv, Ukraine (bohonoff@astral.kiev.ua).

‡Staatliche Studienakademie Thüringen, Berufsakademie Eisenach, University of Cooperative Education, Am Wartenberg 2, D-99817 Eisenach, Germany (ipg@ba-eisenach.de).

§Institute of Mathematics of NAS of Ukraine, 3 Tereshchenkivs'ka Str., Kyiv-4, 01601, Ukraine (makarov@imath.kiev.ua, vasylyk@imath.kiev.ua).

class (1.1). The operator coefficient A is as follows:

(1.3)
$$D(A,t) = \{u(x) \in H^2(0,1) :\ u'(0) - \alpha(t)u(0) = \phi_1(t),\ u'(1) + \beta(t)u(1) = \phi_2(t)\},$$
$$Au = -u''(x) \quad \forall u \in D(A,t).$$

Substitution of (compare with [17])

$$u(x,t) = e^{r(x,t)}v(x,t) \tag{1.4}$$

with function $r(x,t)$ satisfying

$$r_x(0,t) + \alpha(t) = 0, \quad r_x(1,t) - \beta(t) = 0 \tag{1.5}$$

(one can, for example, choose $r(x,t) = -\alpha(t)\cdot x + \frac{1}{2}[\alpha(t) + \beta(t)]\cdot x^2$) transforms (1.2) into

(1.6)
$$\frac{\partial v(x,t)}{\partial t} = \frac{\partial^2 v}{\partial x^2} + 2r_x(x,t)\frac{\partial v}{\partial x} + (r_{xx} - r_t)v + e^{-r(x,t)}f(x,t),$$
$$v_x(0,t) = e^{-r(0,t)}\phi_1(t), \quad v_x(1,t) = e^{-r(0,t)}\phi_2(t),$$
$$v(x,0) = e^{-r(x,0)}u_0(x).$$

The new operator coefficient $\tilde{A}(t)$ is defined by

(1.7)
$$D(\tilde{A},t) = \{v(x) \in H^2(0,1) :\ v'(0) = e^{-r(0,t)}\phi_1(t),\ v'(1) = e^{-r(1,t)}\phi_2(t)\},$$
$$-\tilde{A}(t)u = v''(x) + 2r_x(x,t)v' + (r_{xx} - r_t)v \quad \forall v \in D(\tilde{A},t).$$

The operator coefficient $\tilde{A}$ becomes more complicated and is not more self-adjoint. However, in the homogeneous case $\phi_1(t) \equiv 0,\ \phi_2(t) \equiv 0$ due to substitution (1.4) we obtain a bounded operator $R(t)$ as the multiplicative operator $R(t) = e^{r(x,t)}$ such that the operator $\tilde{A}(t) = [R(t)]^{-1}AR(t)$ possesses the domain independent of t. Under this and some other assumptions it was shown in [17, 16] that the initial value problem (1.2) has a unique solution. Unfortunately there is not a constructive way to find such an operator $R(t)$ in the general case.

The variable domain of an operator can be described by a separate equation, and then we have an abstract problem of the kind

(1.8)
$$\frac{du(t)}{dt} = A(t)u(t), \quad 0 \le s \le t \le T,$$
$$L(t)u(t) = \Phi(t)u(t) + f(t), \quad 0 \le s \le t \le T,$$
$$u(s) = u_0$$

instead of (1.1). Here $L(t)$ and $\Phi(t)$ are some linear operators defined on the boundary of the spatial domain, and the second equation represents an abstract model of the time-dependent boundary condition. An existence and uniqueness result for this problem was proven in [5]. Incorporating the boundary condition into the definition of the operator coefficient of the first equation, one obtains problem (1.1).

The literature concerning discretizations of such problems in an abstract setting is rather limited (see, e.g., [25], where the Euler difference approximation of the first accuracy order for problem (1.1) with the time-dependent domain was considered, and the references therein). The classical Duhamel integral together with discretization of high accuracy order for the two-dimensional heat equation with the time-dependent inhomogeneous Dirichlet boundary condition was proposed in [13]. It is clear that the discretization (with respect to t) is more complicated than in the case of a t-independent domain $D(A)$, since the inclusion $y_k = y(t_k) \in D(A, t_k)$ of the approximate solution y_k at each discretization point t_k should be additionally checked and guaranteed.

In the present paper we consider the problem

$$\begin{aligned} &\frac{du(t)}{dt} + A(t)u(t) = f(t), \\ &\partial_1 u(t) + \partial_0(t)u(t) = g(t), \\ &u(0) = u_0, \end{aligned} \tag{1.9}$$

where $u(t)$ is the unknown function $u : (0,T) \to D(A) \subset X$ with values in a Banach space X, $f(t)$ is a given measurable function $f : (0,T) \to X$ from $L_q(0,T;X)$ with the norm $\|f\| = \{\int_0^T \|f\|_X^q dt\}^{1/q}$, $A(t) : D(A) \in X$ is a densely defined, closed linear operator in X with a time-independent domain $D(A)$, $g : (0,T) \to Y$ is a given function from $L_q(0,T;Y)$ with values in some other Banach space Y, and $\partial_1 : D(A) \to Y$ (independent of t!), $\partial_0(t) : D(A) \to Y$ (can depend on t!) are linear operators. In applications the second equation above is just the time-dependent boundary condition with appropriate operators ∂_1, ∂_0 acting on the boundary of the spatial domain (see section 6 for examples). For this reason we call this equation an abstract (time-dependent) boundary condition.

Incorporating the boundary condition into the definition of the operator coefficient in the first equation, we get a problem of the type (1.1) with a variable domain. The difficulties of the discretization of problems of this form were mentioned above and were well discussed in [25]. The separation of this condition in (1.9) will allow us below to use an abstract Duhamel-like technique in order to reduce the problem to another one including operators with t-independent domains.

The goal of the present paper is threefold. First, we want to generalize the classical Duhamel integral technique for abstract differential equation (1.9) pursuing a t-independent domain of the operators for the transformed statement. Second, we use this technique for reduction of problem (1.9) to a "boundary" integral equation in the case of t-independent operator coefficient A or to a system of boundary integral equations in the case $A = A(t)$. Finally, there is the task of constructing a temporal discretization (i.e., the discretization in t) of these equations with an exponential convergence order.

The boundary integral equations appearing in the framework of the second task deal with integrals over a simple t-interval, and the integrands are the values of some operators (boundary operators) defined on the boundary of the spatial domain. The studied integral equations differ dramatically from those in the boundary integral/boundary element method. The reason is that the last ones include integrals over a boundary of the spatial domain (see, e.g., [31] and the references therein).

The third aim has as a consequence algorithms of low complexity. Exponentially convergent algorithms were proposed recently for various problems. In [7, 8, 9, 10,

11, 14, 15], such algorithms based on the Dunford–Cauchy integral representation and on the Sinc-quadratures were proposed for parabolic and elliptic solution operators, the Silvester operator equation, the inverse of an elliptic operator, and other operator-valued functions. The paper [21] deals with exponentially convergent algorithms for parabolic partial differential equations based on the Runge–Kutta methods. In [23, 22] exponentially convergent algorithms for inverting Laplace transforms were proposed and justified. But to the authors' best knowledge there are no exponentially convergent temporal discretizations for problems of type (1.1) and (1.9) in the literature.

The paper is organized as follows. In section 2 we start from the case of an operator coefficient A independent of the parameter t. Using the Duhamel-like integral and the operator exponential we reduce problem (1.9) to an abstract boundary integral equation. Using the fixed point iteration we show in section 3 that this integral equation is uniquely solvable and that the iteration sequence converges to the exact solution with the factorial convergence rate. The general case of an operator coefficient $A(t)$ depending on the parameter t is considered in section 4. The Duhamel-like technique allows us to reduce problem (1.9) to a system of two boundary integral equations. In section 4 we introduce a prediscretization of this system. The prediscretization is based on the piecewise constant approximation of the operators involved. By a factorially convergent fixed point iteration we prove the existence and uniqueness result for the system of boundary integral equations. Then, using the piecewise constant prediscretization, the Chebyshev interpolation for unknown functions involved, and the collocation, we derive in section 5 a discretization of the boundary integral equations. The main theorem of this section shows an almost (i.e., within a polynomial factor) exponential convergence of the discretization for analytical input data. In sections 6 and 7 we give some examples of parabolic partial differential equations with time-dependent coefficients and boundary conditions which illustrate our theoretical assumptions and confirm the main results. Section 7 presents a numerical example illustrating the efficiency of our temporal discretization.

2. Duhamel-like technique for the first order differential equations in Banach space. In this section we consider a particular case of problem (1.9),

$$\begin{aligned} &\frac{du(t)}{dt} + Au(t) = f(t), \\ &\partial_1 u(t) + \partial_0(t)u(t) = g(t), \quad u(0) = u_0, \end{aligned} \tag{2.1}$$

where the operator A and its domain $D(A)$ in some Banach space X are independent of t, $f(t)$ is a given measurable function $f : (0,T) \to X$ from the space $L_q(0,T;X)$ with the norm $\|f\| = \{\int_0^T \|f\|_X^q dt\}^{1/q}$, $g : (0,T) \to Y$ is a given function from $L_q(0,T;Y)$ with values in some other Banach space Y, and $\partial_1 : D(A) \to Y$ (independent of t!), $\partial_0(t) : D(A) \to Y$ (can depend on t!) are linear operators.

We represent the solution in the form

$$u(t) = w(t) + v(t), \tag{2.2}$$

with v and w satisfying

$$\begin{aligned} &\frac{dv}{dt} + Av = f(t), \\ &\partial_1 v = 0, \quad v(0) = u_0 \end{aligned} \tag{2.3}$$

and

$$\begin{aligned} &\frac{dw(t)}{dt} + Aw = 0, \\ &\partial_1 w(t) = -\partial_0(t)u(t) + g(t), \\ &w(0) = 0, \end{aligned} \tag{2.4}$$

respectively. Introducing the operator $A^{(1)} : D(A^{(1)}) \to X$ independent of t and defined by

$$\begin{aligned} &D(A^{(1)}) = \{u \in D(A) : \partial_1 u = 0\}, \\ &A^{(1)}u = Au \quad \forall u \in D(A^{(1)}), \end{aligned} \tag{2.5}$$

we can rewrite problem (2.3) also in the form

$$\begin{aligned} &\frac{dv}{dt} + A^{(1)}v = f(t), \\ &v(0) = u_0. \end{aligned} \tag{2.6}$$

We suppose that the operator A is such that the operator $A^{(1)}$ is strongly positive [1, 12, 15] (m-sectorial in the sense of [6]); i.e., there exist a positive constant M_R and a fixed $\theta \in [0, \pi/2)$ such that on the rays from the origin building a sector $\Sigma_\theta = \{z \in \mathbb{C} : |\arg(z)| \le \theta\}$ and outside this sector, the following resolvent estimate holds:

$$\|(zI - A^{(1)})^{-1}\| \le \frac{M_R}{1 + |z|}. \tag{2.7}$$

Then under some assumptions with respect to $f(t)$ the solution of (2.6) can be represented by [26]

$$v(t) = e^{-A^{(1)}t}u_0 + \int_0^t e^{-A^{(1)}(t-\tau)}f(\tau)d\tau. \tag{2.8}$$

Note that problem (2.6) with a strongly positive operator $A^{(1)}$ can also be solved numerically by the recently proposed exponentially convergent algorithm from [15].

Returning to problem (2.4) we introduce the auxiliary function $W(\lambda, t)$ by

$$\begin{aligned} &\frac{\partial W(\lambda,t)}{\partial t} + AW(\lambda,t) = 0, \\ &\partial_1 W(\lambda,t) = -\partial_0(\lambda)u(\lambda) + g(\lambda), \\ &W(\lambda,0) = 0, \end{aligned} \tag{2.9}$$

where the abstract boundary condition is now independent of t. The next result gives a representation of the solution of problem (2.4) through the function W by the so-called Duhamel integral.

THEOREM 2.1. *The solution of problem* (2.4) *with a* t*-dependent boundary condition can be represented through the solution of problem* (2.9) *with a* t*-independent boundary condition by the following Duhamel integral:*

$$w(t) = \frac{d}{dt}\int_0^t W(\lambda, t-\lambda)d\lambda = \int_0^t \frac{\partial}{\partial t}W(\lambda, t-\lambda)d\lambda. \tag{2.10}$$

Proof. Let us show that the function (2.10) in fact satisfies (2.4) (compare with the classical representation by Duhamel's integral [13, 27]).

Actually, the initial condition $w(0) = 0$ obviously holds true. Due to the first representation in (2.10) and (2.9) we have

$$\begin{aligned}\partial_1 w(t) &= \frac{d}{dt}\int_0^t \partial_1 W(\lambda, t-\lambda)d\lambda = \frac{d}{dt}\int_0^t [-\partial_0(\lambda)u(\lambda) + g(\lambda)]d\lambda \\ &= -\partial_0(t)u(t) + g(t);\end{aligned} \tag{2.11}$$

i.e., the boundary condition $\partial_1 w(t) = -\partial_0(t)u(t) + g(t)$ is also fulfilled. Due to the second representation in (2.10) we obtain

$$\begin{aligned}&\frac{dw(t)}{dt} + Aw(t) = \frac{d}{dt}\int_0^t \frac{\partial}{\partial t}W(\lambda, t-\lambda)d\lambda + Aw(t) \\ &= \frac{\partial}{\partial t}W(\lambda, t-\lambda)\Big|_{\lambda=t} + \int_0^t \frac{\partial^2}{\partial t^2}W(\lambda, t-\lambda)d\lambda + \int_0^t \frac{\partial}{\partial t}AW(\lambda, t-\lambda)d\lambda \\ &= -AW(\lambda, t-\lambda)|_{\lambda=t} + \int_0^t \frac{\partial}{\partial t}\left[\frac{\partial}{\partial t}W(\lambda, t-\lambda) + AW(\lambda, t-\lambda)\right]d\lambda = 0;\end{aligned} \tag{2.12}$$

i.e., the first equation in (2.4) holds true, which completes the proof. □

In order to make the boundary condition in (2.9) homogeneous and independent of t, we introduce the operator $B : Y \to D(A)$ by

$$\begin{aligned}A(By) &= 0, \\ \partial_1 By &= y;\end{aligned} \tag{2.13}$$

i.e., $\partial_1 B$ is a projector on Y.

Remark 2.2. In order to show the existence of the operator B, we set $By = v$. Then the problem (2.13) is equivalent to the problem

$$Av = 0,$$

$$\partial_1 v = y.$$

We assume that there exists an operator $E : Y \to D(A)$ such that $\partial_1 Ey = y$ (compare with $E(t)$ from [25]). The existence of this operator for an elliptic operator A is proven in [30, section 5.4, Theorem 5.4.5]. Let us introduce $v_1 = v - Ey \in D(A)$, and then

$$Av_1 = -AEy,$$

$$\partial_1 v_1 = 0.$$

Taking into account (2.5) we have the problem

$$A^{(1)}v_1 = -AEy,$$

which has a solution since operator $A^{(1)}$ is strongly positive. Thus, we have

$$B = (I - [A^{(1)}]^{-1}A)E.$$

Given operator B we change the dependent variable by

$$W(\lambda,t)=B[-\partial_0(\lambda)u(\lambda)+g(\lambda)]+W_1(\lambda,t). \tag{2.14}$$

For the new dependent variable we have the problem

$$\begin{aligned}&\frac{\partial W_1(\lambda,t)}{\partial t}+AW_1(\lambda,t)=0,\\ &\partial_1 W_1(\lambda,t)=0,\\ &W_1(\lambda,0)=-B[-\partial_0(\lambda)u(\lambda)+g(\lambda)]\end{aligned} \tag{2.15}$$

or, equivalently,

$$\begin{aligned}&\frac{\partial W_1(\lambda,t)}{\partial t}+A^{(1)}W_1(\lambda,t)=0,\\ &W_1(\lambda,0)=-B[-\partial_0(\lambda)u(\lambda)+g(\lambda)].\end{aligned} \tag{2.16}$$

Using the operator exponential, the solution of this problem can be given in the form

$$W_1(\lambda,t)=-e^{-A^{(1)}t}B[-\partial_0(\lambda)u(\lambda)+g(\lambda)]. \tag{2.17}$$

Now taking into account the substitution (2.14) and the last representation, we obtain

$$W(\lambda,t)=-[e^{-A^{(1)}t}-I]B[-\partial_0(\lambda)u(\lambda)+g(\lambda)], \tag{2.18}$$

where $I:X\to X$ is the identity operator. Due to relations (2.2), (2.10), and (2.18) we arrive at the representation

$$\begin{aligned}u(t)&=v(t)+\frac{\partial}{\partial t}\int_0^t W(\lambda,t-\lambda)d\lambda\\ &=v(t)-\int_0^t\frac{\partial}{\partial t}\left\{e^{-A^{(1)}(t-\lambda)}B[-\partial_0(\lambda)u(\lambda)+g(\lambda)]\right\}d\lambda\end{aligned} \tag{2.19}$$

and further the following boundary integral equation:

$$\partial_0(t)u(t)=\partial_0(t)v(t)-\partial_0(t)\int_0^t\frac{\partial}{\partial t}\left\{e^{-A^{(1)}(t-\lambda)}B[-\partial_0(\lambda)u(\lambda)+g(\lambda)]\right\}d\lambda. \tag{2.20}$$

The last equation can also be written in the form

$$\partial_0(t)u(t)=\partial_0(t)v(t)+\partial_0(t)\int_0^t A^{(1)}e^{-A^{(1)}(t-\lambda)}B[-\partial_0(\lambda)u(\lambda)+g(\lambda)]d\lambda. \tag{2.21}$$

After determining $\partial_0(t)u(t)$ from (2.21), we can find the solution of problem (2.1) by (2.8), (2.19) using the exponentially convergent algorithms for the operator exponential from [15, 10].

Let us introduce the operator-valued kernel $K(t-\tau)=A^{(1)}e^{-A^{(1)}(t-\tau)}B$, the functions $\theta(t)=\partial_0(t)u(t)$ and $F(t)=\partial_0(t)v(t)+\partial_0(t)\int_0^t A^{(1)}e^{-A^{(1)}(t-\lambda)}Bg(\lambda)d\lambda$, and the operator $V:L_q(0,T;D(A))\to L_q(0,T;Y)$ defined by

$$V(t)y(\cdot)=\partial_0(t)\int_0^t K(t-\lambda)y(\lambda)d\lambda. \tag{2.22}$$

Then (2.20) can be written in the fixed point iteration form

$$\theta(t) = V(t)\theta(\cdot) + F(t). \tag{2.23}$$

3. Existence and uniqueness of the solution of the integral equation. In order to prove the existence and uniqueness result for the equivalent equations (2.20), (2.21), and (2.23), we make the following hypotheses (see section 6 for examples):

(A1) There exists a positive constant c such that

$$\|\partial_0(t)\|_{X\to Y} \le c \quad \forall t \in [0,T] \tag{3.1}$$

(compare with H1 from [25]).

(A2) For some $p \ge 1$ it holds that

$$\left[\int_0^t \|A^{(1)}e^{-A^{(1)}(t-\lambda)}B\|^p_{Y\to X}d\lambda\right]^{1/p} \le c \quad \forall t \in [0,T]. \tag{3.2}$$

(A3) The function

$$\begin{aligned} F(t) &= \partial_0(t)v(t) - \partial_0(t)\int_0^t \frac{\partial}{\partial t}e^{-A^{(1)}(t-\lambda)}Bg(\lambda)d\lambda \\ &= \partial_0(t)v(t) + \partial_0(t)\int_0^t A^{(1)}e^{-A^{(1)}(t-\lambda)}Bg(\lambda)d\lambda : (0,T)\to Y \end{aligned} \tag{3.3}$$

belongs to the Banach space $L_q(0,T;Y)$ and

$$\|F\|_{L_q(0,T;Y)} = \left\{\int_0^T \|F(\tau)\|^q_Y d\tau\right\}^{1/q} \le c, \tag{3.4}$$

where q is such that $\frac{1}{p}+\frac{1}{q}=1$ for p defined in (A2).

Let us define the sequences

$$\begin{aligned} \theta_j(t) &= V(t)\theta_{j-1}(\cdot) + F(t), \\ \delta_j(t) &= \theta_j(t) - \theta^*(t), \end{aligned} \tag{3.5}$$

where $\theta^*(t)$ is the exact solution of (2.23).

Now we are in a position to prove the following result.

THEOREM 3.1. *Let conditions* (A1)–(A3) *be fulfilled; then* (2.20) *(or, equivalently,* (2.21), (2.23)) *possesses the unique solution* $\theta^*(t) \in L_q(0,T;Y)$. *This solution is the limit of the sequence* $\{\theta_j(t)\}$ *from* (3.5) *with the factorial convergence, i.e., with the estimate*

$$\|\delta_j(t)\|_{L_q(0,T;Y)} = \|\theta_j(t)-\theta^*(t)\|_{L_q(0,T;Y)} \le \left(\frac{c^{2jq}T^n}{j!}\right)^{1/q}\|\theta_0(t)-\theta^*(t)\|_{L_q(0,T;Y)}. \tag{3.6}$$

For the solution $\theta^*(t) = \partial_0(t)u(t)$ *the following stability estimate holds:*

$$\int_0^t \|\partial_0(\tau)u(\tau)\|^q_Y d\tau \le 2^{q/p}\int_0^t \|F(\tau)\|^q_Y e^{2^{q/p}c^{2q}(t-\tau)}d\tau. \tag{3.7}$$

For the proof of the theorem see Appendix A.

Given the solution $\theta^*(t) = \partial_0(t)u(t)$ of integral equation (2.20), problem (2.1) takes the form

$$\begin{aligned} &\frac{du(t)}{dt} + Au(t) = f(t), \\ &\partial_1 u(t) = g_1(t), \quad u(0) = u_0, \end{aligned} \tag{3.8}$$

with a known function $g_1(t) = \theta^*(t) + g(t)$, and its solution is given by (see (2.19))

$$u(t) = v(t) + \int_0^t \frac{\partial}{\partial t} \left\{ e^{-A^{(1)}(t-\lambda)} B g_1(\lambda) \right\} d\lambda \tag{3.9}$$

and can also be computed with exponential accuracy by algorithms from [15, 10].

4. Generalization to a parameter-dependent operator. Existence and uniqueness result. Let $(X, \|\cdot\|), (\mathcal{W}, \|\cdot\|_{\mathcal{W}}), (Y, \|\cdot\|_Y)$ be three Banach spaces and $\mathcal{W} \subset X$. For each $t \in [0, T]$ we are given a densely defined closed linear unbounded operator $A(t) : \mathcal{W} \to X$ and linear bounded operator $\partial_0(t) : \mathcal{W} \to Y$. We suppose that the domain $D(A(t)) = D(A) \subset \mathcal{W}$ is independent of t.

Let us consider the problem

$$\begin{aligned} &\frac{du(t)}{dt} + A(t)u(t) = f(t), \\ &\partial_1 u(t) + \partial_0(t)u(t) = g(t), \quad t \in (0, T], \\ &u(0) = u_0, \end{aligned} \tag{4.1}$$

where u_0 is a given vector, $f(t)$ is a given vector-valued function, and the operator $\partial_0(t)$ is the product of two operators

$$\partial_0(t) = \mu(t)\partial_0, \tag{4.2}$$

with $\partial_0 : D(A) \to Y$, $\mu(t) : Y \to Y$. We suppose that problem (4.1) possesses a unique solution $u(t)$ for all $t \in (0, T)$ for input data f, g, u_0 from some set including the elements $f = 0, g = 0, u_0 = 0$.

We choose a mesh $\omega_n = \{t_k,\ k = 1, \ldots, n\}$ of n various points on $[0, T]$ and set

$$\begin{aligned} &\overline{A}(t) = A_k = A(t_k), \quad t \in (t_{k-1}, t_k], \\ &\overline{\mu}(t) = \mu_k = \mu(t_k), \quad t \in (t_{k-1}, t_k]. \end{aligned} \tag{4.3}$$

On each subinterval $(t_{k-1}, t_k]$ we define the operator $A_k^{(2)}$ with a t-independent domain by

$$\begin{aligned} &D(A_k^{(2)}) = \{u \in D(A) :\ \partial_1 u + \mu_k \partial_0 u = 0\}, \\ &A_k^{(2)} u = A_k u \quad \forall u \in D(A_k^{(2)}) \end{aligned} \tag{4.4}$$

and the operator $B_k^{(1)} : Y \to D(A)$ by

$$\begin{aligned} &A_k(B_k^{(1)} y) = 0, \\ &(\partial_1 + \mu_k \partial_0) B_k^{(1)} y = y. \end{aligned} \tag{4.5}$$

For all $t \in [0, T)$ we define the operators

$$\begin{aligned} &A^{(2)}(t) = A_k^{(2)}, \quad t \in (t_{k-1}, t_k], \\ &B^{(1)}(t) = B_k^{(1)}, \quad t \in (t_{k-1}, t_k], \quad \forall k = 1, \ldots, n \end{aligned} \tag{4.6}$$

(the existence of $B_k^{(1)}$ can be shown analogously to the existence of B from section 2).

Further, we accept the following hypotheses (see section 6 for examples):

(B1) We suppose the operator $A^{(2)}(t)$ to be strongly positive (see (2.7)).

This assumption implies that there exist a positive constant c and a fixed κ such that (see, e.g., [6, p. 103])

$$\|[A^{(2)}(t)]^{\kappa} e^{-sA^{(2)}(t)}\| \le cs^{-\kappa}, \quad s > 0, \quad \kappa > 0. \tag{4.7}$$

(B2) There exists a real positive ω such that

$$\|e^{-sA^{(2)}(t)}\| \le e^{-\omega s} \quad \forall s, t \in [0, T] \tag{4.8}$$

(see [26, Corollary 3.8, p. 12] for corresponding assumptions on $A(t)$).

We also assume that the following conditions hold:

(B3)

$$\|[A^{(2)}(t) - A^{(2)}(s)][A^{(2)}(t)]^{-\gamma}\| \le c|t - s| \quad \forall t, s, \ 0 \le \gamma \le 1; \tag{4.9}$$

(B4)

$$\|[A^{(2)}(t)]^{\beta}[A^{(2)}(s)]^{-\beta} - I\| \le c|t - s| \quad \forall t, s \in [0, T], \quad \beta \in (0, 1). \tag{4.10}$$

(B5) The operator $\mu(t)\partial_0$ satisfies the following conditions:

$$\begin{aligned} &\|\mu(t) - \mu(t')\|_{Y \to Y} \le M|t - t'|, \\ &\|\partial_0\|_{X \to Y} \le c. \end{aligned} \tag{4.11}$$

(B6) For some $p \ge 1$ and $\gamma \ge 0$ there holds

$$\left[\int_0^t \|[A^{(2)}(\eta)]^{1+\gamma} e^{-A^{(2)}(\eta)(t-\lambda)} B^{(1)}(\eta)\|_{Y \to X}^p d\lambda\right]^{1/p} \le c \quad \forall t, \ \eta \in [0, T]. \tag{4.12}$$

Let us rewrite problem (4.1) in the equivalent form (so-called prediscretization)

$$\begin{aligned} &\frac{du}{dt} + \overline{A}(t)u = [\overline{A}(t) - A(t)]u(t) + f(t), \\ &[\partial_1 + \overline{\mu}(t)\partial_0]u(t) = -[\mu(t) - \overline{\mu}(t)]\partial_0 u(t) + g(t), \quad t \in [0, T], \\ &u(0) = u_0. \end{aligned} \tag{4.13}$$

Note that now all operators on the left-hand side of these equations are constant on each subinterval and piecewise constant on the whole interval $[0, T)$.

From (4.13) analogously to (2.19), (2.21) we deduce

(4.14)

$$u(t) = e^{-A_k^{(2)}(t-t_{k-1})}u(t_{k-1}) + \int_{t_{k-1}}^{t} e^{-A_k^{(2)}(t-\tau)}\{-[A(\tau)-A_k]u(\tau)+f(\tau)\}\,d\tau$$

$$+ \int_{t_{k-1}}^{t} A_k^{(2)} e^{-A_k^{(2)}(t-\lambda)} B_k^{(1)}\{-[\mu(\lambda)-\mu_k]\partial_0 u(\lambda)+g(\lambda)\}d\lambda,$$

$$\partial_0 u(t) = \partial_0 e^{-A_k^{(2)}(t-t_{k-1})}u(t_{k-1})$$

$$+ \partial_0 \int_{t_{k-1}}^{t} e^{-A_k^{(2)}(t-\tau)}\{-[A(\tau)-A_k]u(\tau)+f(\tau)\}\,d\tau$$

$$+ \partial_0 \int_{t_{k-1}}^{t} A_k^{(2)} e^{-A_k^{(2)}(t-\lambda)} B_k^{(1)}\{-[\mu(\lambda)-\mu_k]\partial_0 u(\lambda)+g(\lambda)\}d\lambda,$$

$$t \in [t_{k-1}, t_k], \quad k = 1, \dots, n.$$

Thus, with the Duhamel-like technique we have obtained the system of two integral equations with respect to the unknown functions $u(t)$ and $\partial_0 u(t)$ which is equivalent to (4.1). This system is the starting point for our further investigations and for the numerical algorithm.

In order to prove the existence and uniqueness result it is sufficient to choose in the framework above $n = 1,\ t_0 = 0,\ t_1 = T,\ \overline{A}(t) = A(0) = A$, and $\overline{\mu}(t) = \mu(T)$. In addition, we introduce the vectors

$$\mathcal{U}(t) = \begin{pmatrix} u(t) \\ \theta(t) \end{pmatrix}, \quad \mathcal{U}_n(t) = \begin{pmatrix} u_n(t) \\ \theta_n(t) \end{pmatrix}, \quad \mathcal{F} = \begin{pmatrix} F_1(t) \\ F_2(t) \end{pmatrix} \tag{4.15}$$

and the matrices

(4.16)

$$\mathcal{K}(t,\tau) = \begin{pmatrix} K_{11}(t,\tau) & K_{12}(t,\tau) \\ K_{21}(t,\tau) & K_{22}(t,\tau) \end{pmatrix}, \quad \mathcal{D} = \begin{pmatrix} I & 0 \\ 0 & \partial_0 \end{pmatrix}, \quad \mathcal{E}_\gamma = \begin{pmatrix} [A^{(2)}]^\gamma & 0 \\ 0 & I \end{pmatrix},$$

with

(4.17)

$$K_{11}(t,\tau) = K_{21}(t,\tau) = -e^{-A^{(2)}(t-\tau)}[A(\tau)-A],$$

$$K_{12}(t,\tau) = K_{22}(t,\tau) = -A^{(2)}e^{-A^{(2)}(t-\tau)}B^{(1)}(t)[\mu_k-\mu(\tau)],$$

$$F_1(t) = e^{-A^{(2)}t}u_0 + \int_0^t e^{-A^{(2)}(t-\tau)}f(\tau)d\tau$$

$$+ \operatorname{int}_0^t A^{(2)}e^{-A^{(2)}(t-\lambda)}B^{(1)}g(\lambda)d\lambda, \quad F_2(t) = \partial_0 F_1(t).$$

We also introduce the space $\mathcal{Y}$ of vectors $\mathcal{U} = (u,v)^T$ with the norm

$$\|\mathcal{U}\|_{\mathcal{Y}} = \max\{\|u\|_{L_q(0,T;D([A^{(2)}]^\gamma))}, \|v\|_{L_q(0,T;Y)}\}, \tag{4.18}$$

and we equip the space of matrices $\mathcal{K}(t,\tau)$ from (4.16) with the matrix norm

$$\|\mathcal{K}\|_\infty = \max\{\|K_{1,1}\|_{X\to X} + \|K_{1,2}\|_{Y\to X}, \|K_{2,1}\|_{X\to X} + \|K_{2,2}\|_{Y\to X}\}, \tag{4.19}$$

which is consistent with the vector norm (4.18). Now the system (4.14) with $k = 1,\ t_0 = 0, t_1 = T$ can be written in the form

$$\mathcal{U}(t) = \mathcal{D}\int_0^t \mathcal{K}(t,\tau)\mathcal{U}(\tau)d\tau + \mathcal{F}(t). \tag{4.20}$$

The fixed point iteration for the system (4.20) is given by

$$\mathcal{U}_{j+1}(t) = \mathcal{D}\int_0^t \mathcal{K}(t,\tau)\mathcal{U}_j(\tau)d\tau + \mathcal{F}(t), \quad j = 0, 1, \ldots, \quad \mathcal{U}_0(t) = 0. \tag{4.21}$$

Since the operators $\mathcal{E}_\gamma$ and $\mathcal{D}(t)$ commute, we get from (4.21) that

$$\mathcal{U}_{\gamma,j+1}(t) = \mathcal{D}\int_0^t \mathcal{E}_\gamma\mathcal{K}(t,\tau)\mathcal{E}_\gamma^{-1}\mathcal{U}_{\gamma,j}(\tau)d\tau + \mathcal{F}_\gamma(t), \quad j = 0, 1, \ldots, \quad \mathcal{U}_0(t) = 0, \tag{4.22}$$

where $\mathcal{U}_{\gamma,j}(t) = \mathcal{E}_\gamma\mathcal{U}_j(t)$ and $\mathcal{F}_\gamma(t) = \mathcal{E}_\gamma\mathcal{F}(t)$.

Now we are in a position to formulate the following result (the proof is completely analogous to the one of Theorem 3.1 and is given in Appendix B).

THEOREM 4.1. *Let us assume that the conditions* (B1)–(B6) *are fulfilled; then the system of equations* (4.14) *possesses the unique solution* $\mathcal{U}^*_\gamma(t)$ *in* $\mathcal{Y}$. *This solution is the limit of the sequence* $\mathcal{U}_{\gamma,j}(t)$ *from* (4.22) *with the factorial convergence characterized by the estimate*

$$\|\mathcal{U}_{\gamma,j}(t) - \mathcal{U}^*_\gamma(t)\|_{\mathcal{Y}} \le \left(\frac{c^{2jq}T^j}{j!}\right)^{1/q} \|\mathcal{U}_{\gamma,0}(t) - \mathcal{U}^*_\gamma(t)\|_{\mathcal{Y}}. \tag{4.23}$$

For the solution $\mathcal{U}^*_\gamma(t)$ *the following stability estimate holds:*

$$\int_0^t \|\mathcal{U}^*_\gamma(\tau)\|^q_{\mathcal{Y}} d\tau \le 2^{q/p}\int_0^t \|\mathcal{F}(\tau)\|^q_{\mathcal{Y}} e^{2^{q/p}c^{2q}(t-\tau)}d\tau. \tag{4.24}$$

5. Numerical algorithm. In order to construct a discrete approximation of (4.14) we use the Chebyshev interpolation on the interval $[-1, 1]$ (if it is not the case one can reduce the problem (4.1) to this interval by the variable transform $t = 2t'/T - 1,\ t \in [-1, 1],\ t' \in [0, T]$). We choose a mesh $\omega_n = \{t_k = -\cos\frac{(2k-1)\pi}{2n}, k = 1, \ldots, n\}$ on $[-1, 1]$ of n zeros of Chebyshev orthogonal polynomial $T_n(t)$ and set $\tau_k = t_k - t_{k-1}$. It is well known that (see [28, Chapter 6, Theorem 6.11.12], [29, p. 123])

$$t_{\nu+1} - t_\nu < \frac{\pi}{n}, \quad \nu = 1, \ldots, n,$$

$$\tau = \tau_{max} = \max_{1\le k\le n}\tau_k < \frac{\pi}{n}. \tag{5.1}$$

Let

$$P_{n-1}(t; u) = P_{n-1}u = \sum_{j=1}^n u(t_j)L_{j,n-1}(t),$$

$$P_{n-1}(t; \partial_0 u) = P_{n-1}(\partial_0 u) = \sum_{j=1}^n \partial_0 u(t_j)L_{j,n-1}(t) \tag{5.2}$$

be the Lagrange interpolation polynomials for $u(t)$ and $\partial_0 u(t)$ on the mesh ω_n, where $L_{j,n-1} = \frac{T_n(t)}{T_n'(t_j)(t-t_j)}$, $j = 1, \dots, n$, are the Lagrange fundamental polynomials. For a given vector $v = (v_1, \dots, v_n)$ we introduce the interpolation polynomial

$$P_{n-1}(t; v) = P_{n-1} y = \sum_{j=1}^{n} v_j L_{j,n-1}(t), \tag{5.3}$$

so that $P_{n-1}(t_j; v) = v_j$, $j = 1, 2, \dots, n$. Let $x = (x_1, x_2, \dots, x_n)$ and $y = (y_1, y_2, \dots, y_n)$ be the approximating vectors for $U = (u(t_1), u(t_2), \dots, u(t_n))$ and $\partial U = (\partial_0 u(t_1), \partial_0 u(t_2), \dots, \partial_0 u(t_n))$, respectively; i.e., x_k approximates $u(t_k)$, and y_k approximates $\partial_0 u(t_k)$. Substituting $P_{n-1}(\eta; x)$ for $u(\eta)$ and $P_{n-1}(\eta; y)$ for $\partial_0 u(\lambda)$ and then setting $t = t_k$ in (4.14), we arrive at the following system of linear equations with respect to the unknown elements x_k, y_k:

$$\begin{aligned} x_k &= e^{-A_k^{(2)}\tau_k} x_{k-1} + \sum_{j=1}^{n} \alpha_{kj} x_j + \sum_{j=1}^{n} \beta_{kj} y_j + \phi_k, \\ y_k &= \partial_0 e^{-A_k^{(2)}\tau_k} x_{k-1} + \partial_0 \sum_{j=1}^{n} \alpha_{kj} x_j + \partial_0 \sum_{j=1}^{n} \beta_{kj} y_j + \partial_0 \phi_k, \\ k &= 1, \dots, n, \quad x_0 = u_0, \quad y_0 = \partial_0 u_0, \end{aligned} \tag{5.4}$$

which represents our algorithm. Here we use the notation

$$\begin{aligned} \alpha_{kj} &= \int_{t_{k-1}}^{t_k} e^{-A_k^{(2)}(t_k-\eta)} [A_k - A(\eta)] L_{j,n-1}(\eta) d\eta, \\ \beta_{kj} &= \int_{t_{k-1}}^{t_k} A_k^{(2)} e^{-A_k^{(2)}(t_k-\lambda)} B_k^{(1)} [\mu(\lambda) - \mu_k] L_{j,n-1}(\lambda) d\lambda, \\ \phi_k &= \int_{t_{k-1}}^{t_k} e^{-A_k^{(2)}(t_k-\eta)} f(\eta) d\eta + \int_{t_{k-1}}^{t_k} A_k^{(2)} e^{-A_k^{(2)}(t_k-\lambda)} B_k^{(1)} g(\lambda) d\lambda. \end{aligned} \tag{5.5}$$

Further, for the sake of simplicity we analyze this algorithm for the particular case of problem (4.1), where operator $A(t)$ is independent on t, i.e., $A(t) \equiv A$. In this case we have $\alpha_{kj} = 0$, and system (5.4) takes the form

$$\begin{aligned} x_k &= e^{-A_k^{(2)}\tau_k} x_{k-1} + \sum_{j=1}^{n} \beta_{kj} y_j + \phi_k, \\ y_k &= \partial_0 e^{-A_k^{(2)}\tau_k} x_{k-1} + \partial_0 \sum_{j=1}^{n} \beta_{kj} y_j + \partial_0 \phi_k, \\ k &= 1, \dots, n, \quad x_0 = u_0, \quad y_0 = \partial_0 u_0. \end{aligned} \tag{5.6}$$

Remark 5.1. Under the assumption that $f(t), \mu(t), g(t)$ are polynomials, the calculation of the operators β_{kj} and the elements ϕ_k can be reduced to the calculation of integrals of the kind $I_s = \int_{t_{k-1}}^{t_k} e^{-A_k^{(2)}(t_k-\lambda)} \lambda^s d\lambda$, which can be found by a simple recurrence algorithm: $I_l = -l[A_k^{(2)}]^{-1} I_{l-1} + [A_k^{(2)}]^{-1}(t_k^l I - t_{k-1}^l e^{-A_k^{(2)}\tau_k})$, $l = 1, 2, \dots, s$,

$I_0 = [A_k^{(2)}]^{-1}(I - e^{-A_k^{(2)}\tau_k})$, where the operator exponentials can be computed by the exponentially convergent algorithm from [15].

For the errors $z_x = (z_{x,1}, \dots, z_{x,n})$, $z_y = (z_{y,1}, \dots, z_{y,n})$, with $z_{x,k} = u(t_k) - x_k$ and $z_{y,k} = \partial_0 u(t_k) - y_k$, we have the relations

$$
\begin{aligned}
&z_{x,k} = e^{-A_k\tau_k} z_{x,k-1} + \sum_{j=1}^{n} \beta_{kj} z_{y,j} + \psi_{y,k},\\
&z_{y,k} = \partial_0 e^{-A_k\tau_k} z_{y,k-1} + \partial_0 \sum_{j=1}^{n} \beta_{kj} z_{y,j} + \partial_0 \psi_{y,k},\\
&k = 1, \dots, n, \quad z_{x,0} = 0, \quad z_{y,0} = 0,
\end{aligned}
\tag{5.7}
$$

where

$$
\psi_{y,k} = \int_{t_{k-1}}^{t_k} A_k^{(2)} e^{-A_k^{(2)}(t_k-\eta)} B_k^{(1)} [\mu_k - \mu(\eta)][\partial_0 u(\eta) - P_n(\eta; \partial_0 u)] d\eta. \tag{5.8}
$$

In order to represent algorithm (5.4) in a block-matrix form, we introduce the matrix

$$
S = \{s_{i,k}\}_{i,k=1}^{n} = \begin{pmatrix}
E_X & 0 & 0 & \cdot & \cdot & \cdot & 0 & 0 \\
-e^{-A_2^{(2)}\tau_2} & E_X & 0 & \cdot & \cdot & \cdot & 0 & 0 \\
0 & -e^{-A_3^{(2)}\tau_3} & E_X & \cdot & \cdot & \cdot & 0 & 0 \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
0 & 0 & 0 & \cdot & \cdot & \cdot & -e^{-A_n^{(2)}\tau_n} & E_X
\end{pmatrix}, \tag{5.9}
$$

with E_X being the identity operator in X, the matrix $D = \{\beta_{k,j}\}_{k,j=1}^{n}$, and the vectors

$$
x = \begin{pmatrix} x_1 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{pmatrix}, \quad
f_x = \begin{pmatrix} \phi_1 \\ \cdot \\ \cdot \\ \cdot \\ \phi_n \end{pmatrix}, \quad
\tilde{f}_x = \begin{pmatrix} e^{-A_1^{(2)}\tau_1} u_0 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{pmatrix}, \quad
\psi_y = \begin{pmatrix} \psi_{y,1} \\ \cdot \\ \cdot \\ \cdot \\ \psi_{y,n} \end{pmatrix},
$$

$$
f_y = \begin{pmatrix} \partial_0\phi_1 \\ \cdot \\ \cdot \\ \cdot \\ \partial_0\phi_n \end{pmatrix}, \quad
\tilde{f}_y = \begin{pmatrix} \partial_0 e^{-A_1^{(2)}\tau_1} u_0 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \end{pmatrix}, \quad
\psi_y^{(0)} = \begin{pmatrix} \partial_0\psi_{y,1} \\ \cdot \\ \cdot \\ \cdot \\ \partial_0\psi_{y,n} \end{pmatrix}. \tag{5.10}
$$

It is easy to see that for the (left) inverse

$$S^{-1} = \begin{pmatrix} E_X & 0 & \cdots & 0 & 0 \\ s_1 & E_X & \cdots & 0 & 0 \\ s_2 s_1 & s_2 & \cdots & 0 & 0 \\ \cdot & \cdot & \cdots & \cdot & \cdot \\ s_{n-1} \cdots s_1 & s_{n-1} \cdots s_2 & \cdots & s_{n-1} & E_X \end{pmatrix}, \tag{5.11}$$

where $s_k = -e^{-A^{(2)}_{k+1}\tau_{k+1}}$, it holds that

$$S^{-1}S = \begin{pmatrix} E_X & 0 & \cdots & 0 \\ 0 & E_X & \cdots & 0 \\ \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & \cdots & E_X \end{pmatrix}. \tag{5.12}$$

Remark 5.2. Using results of [15, 10, 11] one can get a parallel and sparse approximation with an exponential convergence rate of the operator exponentials contained in S^{-1} and as a consequence a parallel and sparse approximation of S^{-1}.

The first and second equations of (5.4) can be written in the matrix form as

$$\begin{aligned} Sx &= Dy + f_x + \tilde{f}_x, \\ y &= \Lambda[(I_X - S)x + Dy + f_x + \tilde{f}_x], \end{aligned} \tag{5.13}$$

where

$$\Lambda = \operatorname{diag}[\partial_0, \ldots, \partial_0], \quad I_X = \operatorname{diag}(E_X, \ldots, E_X). \tag{5.14}$$

The errors z_x, z_y satisfy the equations

$$\begin{aligned} Sz_x &= Dz_y + \psi_y, \\ z_y &= \Lambda\left[(I_X - S)z_x + Dz_y + \psi_y\right]. \end{aligned} \tag{5.15}$$

From the second equation in (5.13) and the second equation in (5.15) we obtain

$$\begin{aligned} [I_Y - \Lambda D]y &= \Lambda[(I_X - S)x + f_x + \tilde{f}_x], \\ [I_Y - \Lambda D]z_y &= \Lambda[(I_X - S)z_x + \psi_y], \end{aligned} \tag{5.16}$$

where $I_Y = \operatorname{diag}(E_Y, \ldots, E_Y)$ and E_Y is the identity operator in Y.

For a vector $v = (v_1, v_2, \ldots, v_n)^T$ and a block operator matrix $A = \{a_{ij}\}_{i,j=1}^n$, we introduce the vector norm

$$|||v||| \equiv |||v|||_\infty = \max_{1 \le k \le n} \|v_k\| \tag{5.17}$$

and the consistent matrix norm

$$|||A||| \equiv |||A|||_\infty = \max_{1 \le i \le n} \sum_{j=1}^{n} \|a_{i,j}\|. \tag{5.18}$$

For further analysis we need the following auxiliary result.

LEMMA 5.3. *Under assumptions* (B1)–(B6) *the following estimates hold true:*

$$
\begin{aligned}
&|||S^{-1}||| \le n,\\
&|||D||| \le cn^{-(1+1/q)} \ln n, \quad 1/p + 1/q = 1,\\
&|||\Lambda||| \le c,
\end{aligned}
\tag{5.19}
$$

with a positive constant c independent of n.

Proof. The assumption (B2) yields

$$
|||S^{-1}||| \le 1 + e^{-\omega\tau} + \cdots + e^{-\omega\tau(n-1)} \le n. \tag{5.20}
$$

For the matrix D we have from (4.12)

$$
\begin{aligned}
|||D||| &\le \max_{1\le k\le n} \sum_{j=1}^{n} \|\beta_{kj}\|\\
&= \max_{1\le k\le n} \sum_{j=1}^{n} \left\| \int_{t_{k-1}}^{t_k} A_k^{(2)} e^{-A_k^{(2)}(t_k-\tau)} B_k^{(1)} [\mu(t_k) - \mu(\tau)] L_{j,n-1}(\tau) d\tau \right\|\\
&\le \max_{1\le k\le n} \int_{t_{k-1}}^{t_k} \|A_k^{(2)} e^{-A_k^{(2)}(t_k-\tau)} B_k^{(1)}\| \|\mu(t_k) - \mu(\tau)\| \sum_{j=1}^{n} |L_{j,n-1}(\tau)| d\tau\\
&\le c\Lambda_n (t_k - t_{k-1}) \int_{t_{k-1}}^{t_k} \|A_k^{(2)} e^{-A_k^{(2)}(t_k-\tau)} B_k^{(1)}\| d\tau \le cn^{-(1+1/q)} \ln n,
\end{aligned}
\tag{5.21}
$$

where $\Lambda_n = \max_{-1\le\tau\le 1} \sum_{j=1}^{n} |L_{j,n-1}(\tau)|$ is the Lebesgue constant related to the Chebyshev interpolation nodes.

The last estimate is a simple consequence of assumption (B5). The lemma is proved. ∎

Due to (5.19) for n large enough there exists the inverse $(I_Y - \Lambda D)^{-1}$, which is bounded by a constant c independent of n; i.e.,

$$
|||(I_Y - \Lambda D)^{-1}||| \le c. \tag{5.22}
$$

Therefore, we obtain from (5.16)

$$
\begin{aligned}
y &= [I_Y - \Lambda D]^{-1} \Lambda \left[(I_X - S)x + f_x + \tilde{f}_x \right],\\
z_y &= [I_Y - \Lambda D]^{-1} \Lambda \left[(I_X - S) z_x + \psi_y \right].
\end{aligned}
\tag{5.23}
$$

Substituting these expressions into the first equation in (5.13) and (5.15), respectively, we obtain

$$
\begin{aligned}
Gx &= Q(f_x + \tilde{f}_x),\\
Gz_x &= Q\psi_y,
\end{aligned}
\tag{5.24}
$$

where

$$
\begin{aligned}
G &= S - D[I_Y - \Lambda D]^{-1} \Lambda (I_X - S),\\
Q &= D[I_Y - \Lambda D]^{-1} \Lambda + I_X.
\end{aligned}
\tag{5.25}
$$

The next lemma presents estimates for G^{-1} and Q.

LEMMA 5.4. *Under assumptions of Lemma* 5.3 *there exists* G^{-1}, *and there holds*

$$\begin{aligned} |||G^{-1}||| &\le cn, \\ |||Q||| &\le c \end{aligned} \tag{5.26}$$

with some constant independent on n.

Proof. We represent $G = S[I_X - G_1]$ with $G_1 = S^{-1}D[I_Y - \Lambda D]^{-1}\Lambda(I_X - S)$ and estimate $|||G_1|||$. We have

$$|||G_1||| \le |||S^{-1}||| \cdot |||D||| \cdot |||(I_Y - \Lambda D)^{-1}||| \cdot |||\Lambda||| \cdot |||I_X - S|||, \tag{5.27}$$

and now Lemma 5.3 implies that

$$|||G_1||| \le n\, n^{-(1+1/q)} c \ln n = c\frac{\ln n}{n^{1/q}}. \tag{5.28}$$

This estimate guarantees the existence of the bounded inverse operator $(I_X - G_1)^{-1}$, which together with the estimate $|||S^{-1}||| \le n$ proves the first assertion of the lemma. The second assertion is evident. The proof is complete. □

This lemma and (5.24) imply the following stability estimates:

$$\begin{aligned} |||x||| &\le cn|||f_x + \tilde{f}_x|||, \\ |||z_x||| &\le cn|||\tilde{\psi}_y|||. \end{aligned} \tag{5.29}$$

Substituting estimates (5.29) into (5.23) and taking into account the estimates $|||I_X - S||| \le c$, $|||\Lambda||| \le c$ as well as (5.22), we obtain

$$\begin{aligned} |||y||| &\le cn|||f_x + \tilde{f}_x|||, \\ |||z_y||| &\le cn|||\psi_y|||. \end{aligned} \tag{5.30}$$

Remark 5.5. We have reduced the interval length to $T = 2$, but in general the constants c involved depend on the interval length T.

Let Π_{n-1} be the set of all polynomials in t with vector coefficients of degree less than or equal to $n - 1$. In complete analogy with [2, 28, 29], the following Lebesgue inequality for vector-valued functions can be proved:

$$\|u(\eta) - P_{n-1}(\eta; u)\|_{C[-1,1]} \equiv \max_{\eta\in[-1,1]} \|u(\eta) - P_{n-1}(\eta; u)\| \le (1 + \Lambda_n)E_n(u), \tag{5.31}$$

with the error of the best approximation of u by polynomials of degree not greater than $n - 1$

$$E_n(u) = \inf_{p\in\Pi_{n-1}} \max_{\eta\in[-1,1]} \|u(\eta) - p(\eta)\|. \tag{5.32}$$

Now we can go to the main result of this section.

THEOREM 5.6. *Let the assumptions of Lemma* 5.3 *hold; then there exists a positive constant* c *such that*

1. *for n large enough it holds that*

$$|||z_x||| \le cn^{-1/q} \cdot \ln n \cdot E_n(\partial_0 u),$$
$$|||z_y||| \le cn^{-1/q} \cdot \ln n \cdot E_n(\partial_0 u), \tag{5.33}$$

where u is the solution of (4.1);

2. *the first equation in* (5.24) *can be written in the form*

$$x = G_1 x + S^{-1} Q(f_x + \tilde{f}_x) \tag{5.34}$$

and can be solved by the fixed point iteration

$$x^{(k+1)} = G_1 x^{(k)} + S^{-1} Q(f_x + \tilde{f}_x), \quad k = 0, 1, \ldots; \quad x^{(0)} \text{ is arbitrary} \tag{5.35}$$

with the convergence rate of a geometrical progression with the denominator $q \le c\frac{\ln n}{n^{1/q}} < 1$ for n large enough.

Proof. Using (B6) the norm of ψ_y can be estimated in the following way:

$$\begin{aligned} |||\psi_y||| &= \max_{1\le k\le n} \left| \int_{t_{k-1}}^{t_k} A_k^{(2)} e^{-A_k^{(2)}(t_k-\eta)} B_k^{(1)} [\mu_k - \mu(\eta)]\,[\theta(\eta) - P_n(\eta;\theta)] d\eta \right| \\ &\le c\|\theta(\cdot) - P_n(\cdot;\theta)\|_{C[-1,1]} \max_{1\le k\le n} (t_k - t_{k-1})^{1+1/q} \\ &\le cn^{-(1+1/q)}(1+\Lambda_n) E_n(\theta) \le cn^{-(1+1/q)} \cdot \ln n \cdot E_n(\theta). \end{aligned} \tag{5.36}$$

Now the first assertion of the theorem follows from (5.29) and (5.30) and the second one from (5.28). ∎

Remark 5.7. Assuming that the vector-valued functions $\partial_0 u(t)$ can be analytically extended from the interval $[-1, 1]$ into the domain $\mathcal{D}_\rho$ enveloped by the so-called Bernstein regularity ellipse $\mathcal{E}_\rho = \mathcal{E}_\rho(B)$ (with the foci at $z = \pm 1$ and the sum of semiaxes equal to $\rho > 1$),

$$\begin{aligned} \mathcal{E}_\rho &= \left\{ z \in \mathbb{C} : z = \frac{1}{2}\left(\rho e^{i\varphi} + \frac{1}{\rho} e^{-i\phi}\right) \right\} \\ &= \left\{ (x,y) : \frac{x^2}{a^2} + \frac{y^2}{b^2} = 1,\ a = \frac{1}{2}\left(\rho + \frac{1}{\rho}\right),\ b = \frac{1}{2}\left(\rho - \frac{1}{\rho}\right) \right\}, \end{aligned}$$

we obtain for the value of the best polynomial approximation [3, 15]

$$E_n(\partial_0 u) \le \frac{\rho^{-n}}{1-\rho} \sup_{z\in D_\rho} \|\partial_0 u(z)\|, \tag{5.37}$$

which together with (5.33) yields the exponential convergence rate with the estimates (5.33).

Remark 5.8. For an elliptic operator A and for its discrete counterpart A_h (say, A_h is the finite element method/finite difference stiffness matrix corresponding to A), it holds that (see, e.g., [6])

$$\|(zI - A)^{-1}\|_{X\to X} \le \frac{1}{|z| \sin(\theta_1 - \theta)} \quad \forall z \in \mathbb{C} :\ \theta_1 \le |\arg z| \le \pi \tag{5.38}$$

for any $\theta_1 \in (\theta, \pi)$, where $\cos\theta = \delta_0/C$; i.e., both are strongly positive in the sense of (2.7). Replacing A by A_h in (4.1) and then applying the temporal discretization described above, we arrive at the full discretization of problem (2.1). Since the bound (5.38) on the matrix resolvent of A_h is valid uniformly in the mesh size h, the full error is the sum of the temporal and the spatial errors which can be controlled independently from each other.

6. Examples. In this section we show that many applied parabolic problems with a time-dependent boundary condition can be fitted into our abstract framework. Our aim here is to illustrate the assumptions accepted in sections 3 and 4.

A special example of the problem from the class (1.1) is

$$\begin{aligned} &\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + f(x,t), \\ &u(0,t) = 0, \quad \frac{\partial u(1,t)}{\partial x} + b(t)u(1,t) = g(t), \\ &u(x,0) = u_0(x), \end{aligned} \tag{6.1}$$

where the operator $A : D(A) \in X \to X,\ X = L_q(0,1)$, is defined by

$$\begin{aligned} &D(A) = \{v \in W_q^2(0,1) : v(0) = 0\}, \\ &Av = -\frac{\partial^2 v}{\partial x^2}, \end{aligned} \tag{6.2}$$

the operators $\partial_1 : D(A) \to Y$ and $\partial_0(t) : D(A) \to Y$, $Y = \mathbb{R}$ are defined by

$$\begin{aligned} &\partial_1 u = \left.\frac{\partial u(x,t)}{\partial x}\right|_{x=1}, \\ &\partial_0(t)u = b(t) \cdot u(x,t)|_{x=1}, \end{aligned} \tag{6.3}$$

and $g(t) \in L_q(0,T;Y) = L_q(0,T)$.

We represent the solution of (6.1) in the form (compare with (2.3), (2.4))

$$u(x,t) = w(x,t) + v(x,t), \tag{6.4}$$

where the function $v(x,t)$ is the solution of the problem

$$\begin{aligned} &\frac{\partial v}{\partial t} = \frac{\partial^2 v}{\partial x^2} + f(x,t), \\ &v(0,t) = 0, \quad \frac{\partial v(1,t)}{\partial x} = 0, \\ &v(x,0) = u_0(x) \end{aligned} \tag{6.5}$$

and the function $w(x,t)$ satisfies

$$\begin{aligned} &\frac{\partial w}{\partial t} = \frac{\partial^2 w}{\partial x^2}, \\ &w(0,t) = 0, \quad \frac{\partial w(1,t)}{\partial x} = -b(t)u(1,t) + g(t), \\ &w(x,0) = 0. \end{aligned} \tag{6.6}$$

Introducing the operator $A^{(1)} : D(A^{(1)}) \to X$ defined by

$$\begin{aligned} &D(A^{(1)}) = \{u \in W_q^2(0,1) :\ u(0) = 0,\ u'(1) = 0\} = \{u \in D(A) : u'(1) = 0\}, \\ &A^{(1)}u = -\frac{d^2u}{dx^2} \quad \forall u \in D(A^{(1)}) \end{aligned} \tag{6.7}$$

(see (2.5) for an abstract setting), we can also write problem (6.5) in the form

$$\begin{aligned} &\frac{dv}{dt} + A^{(1)}v = f(t), \\ &v(0) = u_0, \end{aligned} \tag{6.8}$$

with the solution

$$v(t) = e^{-A^{(1)}t}u_0 + \int_0^t e^{-A^{(1)}(t-\tau)}f(\tau)d\tau. \tag{6.9}$$

In order to solve the problem (6.6) we use Duhamel's integral. We introduce the auxiliary function $W(x, \lambda, t)$ satisfying the following equations (compare with (2.10)):

$$\begin{aligned} &\frac{\partial W(x,\lambda,t)}{\partial t} = \frac{\partial^2 W(x,\lambda,t)}{\partial x^2}, \\ &W(0,\lambda,t) = 0, \quad \frac{\partial W(1,\lambda,t)}{\partial x} = -b(\lambda)u(1,\lambda) + g(\lambda), \\ &W(x,\lambda,0) = 0 \end{aligned} \tag{6.10}$$

with time-independent boundary conditions. Then the solution of problem (6.6) is given by (see Theorem 2.1)

$$w(x,t) = \int_0^t \frac{\partial}{\partial t} W(x,\lambda,t-\lambda)d\lambda. \tag{6.11}$$

Using this representation one can obtain an integral equation with respect to $u(1,t)$. Actually, we have

$$\begin{aligned} w(1,t) = u(1,t) - v(1,t) &= \frac{\partial}{\partial t}\int_0^t W(1,\lambda,t-\lambda)d\lambda \\ &= \int_0^t \frac{\partial}{\partial t} W(1,\lambda,t-\lambda)d\lambda. \end{aligned} \tag{6.12}$$

The substitution

$$W(x,\lambda,t) = B[-b(\lambda)u(1,\lambda) + g(\lambda)] + W_1(x,\lambda,t) \tag{6.13}$$

implies the following problem with homogeneous boundary conditions for $W_1(x, \lambda, t)$:

$$\begin{aligned} &\frac{\partial W_1(x,\lambda,t)}{\partial t} = \frac{\partial^2 W_1(x,\lambda,t)}{\partial x^2}, \\ &W_1(0,\lambda,t) = 0, \quad \frac{\partial W_1(1,\lambda,t)}{\partial x} = 0, \\ &W_1(x,\lambda,0) = -B[-b(\lambda)u(1,\lambda) + g(\lambda)], \end{aligned} \tag{6.14}$$

where the operator B is given by

$$B[-b(\lambda)u(1,\lambda)+g(\lambda)] = x[-b(\lambda)u(1,\lambda)+g(\lambda)]. \tag{6.15}$$

By separation of variables we get the solution of this problem explicitly in the form

$$\begin{aligned} W_1(x,\lambda,t) &= -2[-b(\lambda)u(1,\lambda)+g(\lambda)] \\ &\quad \cdot \sum_{n=1}^{\infty}(-1)^{n+1}\left[\frac{2}{(2n-1)\pi}\right]^2 e^{-[\pi(2n-1)/2]^2 t}\sin\left[\frac{\pi}{2}(2n-1)x\right]. \end{aligned} \tag{6.16}$$

Due to (6.15) the boundary integral equation (2.20) (or, equivalently, (2.23)) for the example problem (6.1) takes the form

$$\begin{aligned} b(t)u(1,t) &= b(t)v(1,t) + b(t)\int_0^t \frac{\partial}{\partial t}W_1(1,\lambda,t-\lambda)d\lambda \\ &= b(t)v(1,t) - b(t)\int_0^t K(t-\lambda)g(\lambda)d\lambda + b(t)\int_0^t K(t-\lambda)b(\lambda)u(1,\lambda)d\lambda, \end{aligned} \tag{6.17}$$

with

$$K(t) = 2\sum_{n=1}^{\infty} e^{-[\pi(2n-1)/2]^2 t}. \tag{6.18}$$

Remark 6.1. Note that in this particular case we can represent the integrand through the kernel $K(t-\lambda)$ analytically. In the general case one can use the exponentially convergent algorithm for the operator exponential in (2.21) like the one from [15].

Let us illustrate Theorem 3.1 for the example problem (6.1). It is easy to see that the condition (A1) is fulfilled provided that

$$|b(t)| \le c \quad \forall t \in [0,T]. \tag{6.19}$$

Let us show that there exists p for which the assumption (A2) holds. Actually, we have

$$\begin{aligned} \int_0^t \|A^{(1)}e^{-A^{(1)}(t-\lambda)}B\|_{Y\to X}^p d\lambda &= \int_0^t [K(t-\lambda)]^p\, d\lambda \\ &= \int_0^t \left[2\sum_{n=1}^{\infty} e^{-(\frac{\pi(2n-1)}{2})^2(t-\lambda)}\right]^p d\lambda. \end{aligned} \tag{6.20}$$

Using the Hölder inequality, the kernel $K(t-\lambda)$ can be estimated by

$$\begin{aligned} K(t-\lambda) &= 2\sum_{n=1}^{\infty} e^{-(\frac{\pi(2n-1)}{2})^2(t-\lambda)} \\ &= 2\sum_{n=1}^{\infty}(2n-1)^{\alpha}e^{-(\frac{\pi(2n-1)}{2})^2(t-\lambda)}\frac{1}{(2n-1)^{\alpha}} \\ &\le 2\left\{\sum_{n=1}^{\infty}(2n-1)^{p\alpha}e^{-p(\frac{\pi(2n-1)}{2})^2(t-\lambda)}\right\}^{1/p}\cdot\left\{\sum_{n=1}^{\infty}\frac{1}{(2n-1)^{q\alpha}}\right\}^{1/q}, \end{aligned} \tag{6.21}$$

where $\frac{1}{p}+\frac{1}{q}=1$.

Substituting this inequality into (6.20), we obtain

$$\int_0^t [K(t-\lambda)]^p d\lambda \le \frac{2^{p+1}}{p\pi^2}\left\{\sum_{n=1}^{\infty}\frac{1}{(2n-1)^{q\alpha}}\right\}^{p/q}$$

$$\cdot\sum_{n=1}^{\infty}\frac{1}{(2n-1)^{2-p\alpha}}\left[1-e^{-p(\frac{\pi(2n-1)}{2})^2 t}\right] \tag{6.22}$$

$$\le \frac{2^{p+1}}{p\pi^2}\left\{\sum_{n=1}^{\infty}\frac{1}{(2n-1)^{q\alpha}}\right\}^{p/q}\cdot\sum_{n=1}^{\infty}\frac{1}{(2n-1)^{2-p\alpha}}.$$

The series are bounded by a constant c if

$$\begin{cases} q\alpha>1, \\ 2-p\alpha>1, \end{cases} \tag{6.23}$$

or $\frac{1}{q}<\alpha<\frac{1}{p}$. Thus, we can choose an arbitrary $p\in[1,2)$, and then $\frac{1}{q}=1-\frac{1}{p}$ and the choice of an arbitrary α from the interval $(1-\frac{1}{p},\frac{1}{p})$ provide that the assumption (A2) holds.

The assumption (A3) for the example problem (6.1) reads as

$$\left\{\int_0^T\left|b(t)v(1,t)-b(t)\int_0^t K(t-\lambda)g(\lambda)d\lambda\right|^q dt\right\}^{1/q}\le c. \tag{6.24}$$

Since $K(t-\cdot)\in L_p(0,T)$, this inequality holds provided that

$$g(t)\in L_q(0,T),\quad v(1,t)\in L_q(0,T). \tag{6.25}$$

Due to the estimate

$$\left(\int_0^T |v(1,t)|^q dt\right)^{1/q}=\left(\int_0^T\left|\int_0^1\frac{\partial v(x,t)}{\partial x}dx\right|^q dt\right)^{1/q}$$

$$\le\left(\int_0^T\int_0^1\left|\frac{\partial v(x,t)}{\partial x}\right|^q dxdt\right)^{1/q}, \tag{6.26}$$

the condition $v(1,t)\in L_q(0,T)$ is fulfilled if the last integral exists. The corresponding sufficient conditions on the input data $f(x,t),\ u_0(x)$ of the problem (6.1) can be found in [26, 20].

Given the solution of (6.17) (which can be found numerically by algorithm (5.4) consisting in this case of the second equation only), the problem (6.1) takes the form

$$\frac{du}{dt}+Au=f(t),$$

$$u(0)=u_0,\quad \partial_1 u=g_1(t), \tag{6.27}$$

with a known function $g_1(t) = g(t) - \partial_0(t)u(t)$. Using the representation by Duhamel's integral (6.11) we have

$$u(t) = v(t) - \int_0^t A^{(1)} e^{-A^{(1)}(t-\lambda)} B g_1(\lambda) d\lambda. \tag{6.28}$$

Now let us show that the assumptions (B1)–(B6) hold for the following model problem from the class (1.1):

$$\begin{aligned} &\frac{\partial u(x,t)}{\partial t} = \frac{\partial^2 u}{\partial x^2} - q(x,t)u + f(x,t), \\ &u(0,t) = 0, \quad \frac{\partial u(1,t)}{\partial x} + b(t)u(1,t) = g(t), \\ &u(x,0) = u_0(x). \end{aligned} \tag{6.29}$$

The results of [20, section 4, paragraph 9, p. 388] yield that this problem possesses the unique solution $u \in W_q^{2,1}(Q_T)$ provided that $q(x,t),\ f(x,t) \in L_q(Q_T),\ Q_T = (0,1) \times (0,T),\ g(t) \in W_q^{1-1/(2q)}(0,T),\ b(t) \in W_q^{1/2}(0,T),\ u_0 \in W_q^{2-\frac{2}{q}}(0,1),\ q > 3/2$, and some compatibility conditions for initial and boundary conditions are fulfilled.

Here the operator $A(t) : D(A) \in X \to X,\ X = L_q(0,1)$, is defined by

$$\begin{aligned} &D(A(t)) = \{v \in W_q^2(0,1) : v(0) = 0\}, \\ &A(t)v = -\frac{d^2 v(x)}{dx^2} + q(x,t)v(x), \end{aligned} \tag{6.30}$$

the operators $\partial_1 : D(A) \to Y$ and $\partial_0(t) : D(A) \to Y$, $Y = \mathbb{R}$ are defined by

$$\begin{aligned} &\partial_1 u = \left.\frac{\partial u(x,t)}{\partial x}\right|_{x=1}, \\ &\partial_0(t)u = b(t) \cdot u(x,t)|_{x=1}, \end{aligned} \tag{6.31}$$

and $g(t) \in L_q(0,T;Y) = L_q(0,T)$. We suppose that the function $b(t)$ satisfies condition (6.19), so that our assumption (B5) is fulfilled.

The piecewise constant operator $A^{(2)}(t)$ from (4.6) is defined on each subinterval $[t_{k-1}, t_k]$ by

$$D(A^{(2)}(t)) = \{v \in W_q^2(0,1) : v(0) = 0, \quad v'(1) + \mu_k v(1) = 0\}, \quad \mu_k = b(t_k),$$

$$A_k v = -\frac{d^2 v(x)}{dx^2} + q(x,t_k)v(x)$$

and the operator $B_k^{(1)}$ by

$$B_k^{(1)} z = \frac{x}{\mu_k + 1} z, \quad z \in Y.$$

Assumptions (B1)–(B2) are fulfilled due to results from [6], and assumptions (B3)–(B4) are obviously fulfilled, too. It remains to check assumption (B6). Let

$(\lambda_j^{(k)}, e_j^{(k)})$, $j = 1, 2, \ldots$, be the system of eigenpairs of $A^{(2)}(t), t \in [t_{k-1}, t_k]$. Then we have

$$\int_0^t \|[A^{(2)}]^{1+\gamma} e^{-A^{(2)}(t-\lambda)} B^{(1)}\|_{Y\to X}^p d\lambda = \sum_{k=1}^n \int_{t_{k-1}}^{t_k} \left[K_\gamma^{(k)}(t-\lambda)\right]^p d\lambda, \tag{6.32}$$

where

$$K_\gamma^{(k)}(t-\lambda) = \sum_{j=1}^\infty [\lambda_j^{(k)}]^\gamma e^{-\lambda_j^{(k)}(t-\lambda)}. \tag{6.33}$$

Using the well-known asymptotic $c_L j^2 \le \lambda_j \le c_U j^2$ (see, e.g, [24]) and Hölder's inequality, the kernel $K_\gamma^{(k)}(t-\lambda)$ can be estimated in the following way:

(6.34)

$$\begin{aligned} K_\gamma^{(k)}(t-\lambda) &= \sum_{j=1}^\infty \left(\lambda_j^{(k)}\right)^\gamma e^{-\lambda_j^{(k)}(t-\lambda)} \le c_U^\gamma \sum_{j=1}^\infty j^{2\gamma} e^{-c_L j^2 (t-\lambda)} \\ &= c_U^\gamma \sum_{n=1}^\infty j^{2\gamma+\alpha} e^{-c_L j^2(t-\lambda)} \frac{1}{j^\alpha} \\ &\le c_U^\gamma \left\{\sum_{j=1}^\infty j^{p(\alpha+2\gamma)} e^{-pc_L j^2(t-\lambda)}\right\}^{1/p} \cdot \left\{\sum_{j=1}^\infty \frac{1}{j^{q\alpha}}\right\}^{1/q}, \quad \frac{1}{p}+\frac{1}{q} = 1. \end{aligned}$$

Substituting this inequality into (6.32) we obtain

(6.35)

$$\begin{aligned} &\sum_{k=1}^n \int_{t_{k-1}}^{t_k} [K_\gamma^{(k)}(t-\lambda)]^p d\lambda \le \sum_{k=1}^n c_U^\gamma \int_{t_{k-1}}^{t_k} \sum_{j=1}^\infty j^{p(\alpha+2\gamma)} e^{-pc_L j^2(t-\lambda)} d\lambda \cdot \left\{\sum_{j=1}^\infty \frac{1}{j^{q\alpha}}\right\}^{p/q} \\ &= c_U^\gamma \sum_{j=1}^\infty j^{p(\alpha+2\gamma)} (pc_L j^2)^{-1} \left[1 - e^{-pc_L j^2 t}\right] \cdot \left\{\sum_{j=1}^\infty \frac{1}{j^{q\alpha}}\right\}^{p/q} \\ &\le \frac{c_U^\gamma}{pc_L} \sum_{j=1}^\infty \frac{1}{j^{2-p(\alpha+2\gamma)}} \cdot \left\{\sum_{j=1}^\infty \frac{1}{j^{q\alpha}}\right\}^{p/q}. \end{aligned}$$

The series remains bounded if

$$\begin{cases} q\alpha > 1, \\ 2 - p(\alpha + 2\gamma) > 1, \end{cases} \tag{6.36}$$

or $\frac{1}{q} < \alpha < \frac{1}{p} - 2\gamma$. Thus, we can choose an arbitrary $p \in [1, 2)$; then $\frac{1}{q} = 1 - \frac{1}{p}$ and the assumption (B6) holds with an arbitrary γ from the interval $[0, \frac{1}{p} - \frac{1}{2})$.

7. Numerical example. In this section we show that the algorithm (5.4) possesses the exponential convergence with respect to the temporal discretization parameter n predicted by Theorem 5.6. In order to eliminate the influence of other errors (the spatial error, the error of approximation of the operator exponential and of the integrals in (5.5)), we calculate the coefficients of the algorithm (5.6) exactly using the computer algebra tool Maple.

We consider the model problem (6.1) with

$$u_0(x) = \frac{4\sqrt{2}}{\pi}\sin\frac{\pi}{4}x, \quad b(t) = \frac{\pi}{4}\exp\left(\frac{\pi^2}{16}t\right), \quad g(t) = 1 + \exp\left(-\frac{\pi^2}{16}t\right),$$

for which algorithm (5.4) consists of the second equation only, where $\alpha_{kj} = 0$.

The exact solution of this problem is

$$u(x,t) = \exp\left(-\frac{\pi^2}{16}t\right)\frac{4\sqrt{2}}{\pi}\sin\frac{\pi}{4}x.$$

For the function $v(x,t)$ from (6.5) we have the following problem:

$$\begin{aligned} &\frac{\partial v(x,t)}{\partial t} = \frac{\partial^2 v(x,t)}{\partial x^2}, \quad v(0,t) = 0, \quad \frac{\partial v(1,t)}{\partial x} = 0, \\ &v(x,0) = \frac{4\sqrt{2}}{\pi}\sin\frac{\pi}{4}x. \end{aligned} \tag{7.1}$$

It is easy to check that

$$v(x,t) = \sum_{k=1}^{\infty} \frac{(-1)^{k+1}32}{\pi^2(4k-1)(4k-3)} e^{-\left(\frac{\pi(2k-1)}{2}\right)^2 t}\sin\Big(\frac{\pi}{2}(2k-1)x\Big). \tag{7.2}$$

Equation (6.17) (compare also with (6.16), (6.22)) reads as follows:

$$\begin{aligned} u(1,t) &= v(1,t) + 2\int_0^t \left[1 + e^{\frac{-\pi^2}{16}\tau}\right]\sum_{k=1}^{\infty} e^{-\left(\frac{\pi(2k-1)}{2}\right)^2(t-\tau)}d\tau \\ &\quad -2\int_0^t \frac{\pi}{4}e^{\frac{\pi^2}{16}\tau}u(1,\tau)\sum_{k=1}^{\infty} e^{-\left(\frac{\pi(2k-1)}{2}\right)^2(t-\tau)}d\tau \\ &= f(t) - \frac{\pi}{2}\int_0^t e^{\frac{\pi^2}{16}\tau}u(1,\tau)\sum_{k=1}^{\infty} e^{-\left(\frac{\pi(2k-1)}{2}\right)^2(t-\tau)}d\tau, \end{aligned} \tag{7.3}$$

with

$$\begin{aligned} f(t) &= v(1,t) - 2\sum_{k=1}^{\infty} e^{-\left(\frac{\pi(2k-1)}{2}\right)^2 t}\left[\frac{1}{\left(\frac{\pi(2k-1)}{2}\right)^2} + \frac{1}{\left(\frac{\pi(2k-1)}{2}\right)^2 - \frac{\pi^2}{16}}\right] + 1 + \frac{4}{\pi}e^{-\frac{\pi^2}{16}t} \\ &= -\frac{8}{\pi^2}\sum_{k=1}^{\infty}\frac{e^{-\left(\frac{\pi(2k-1)}{2}\right)^2 t}}{(2k-1)^2} + 1 + \frac{4}{\pi}e^{-\frac{\pi^2}{16}t}. \end{aligned}$$

In order to reduce the problem to the interval $[-1,1]$ we change the variable t in (7.3) by

$$t=\frac{s+1}{2}T,\quad s\in[-1,1];\quad \tau=\frac{\xi+1}{2}T,\quad \xi\in[-1,s];\quad d\tau=\frac{T}{2}d\xi$$

and obtain

$$u\left(1,\frac{s+1}{2}T\right)=f\left(\frac{s+1}{2}T\right)-\frac{\pi T}{4}\sum_{k=1}^{\infty}\int_{-1}^{s}e^{\frac{\pi^2}{16}\frac{\xi+1}{2}T}u\left(1,\frac{\xi+1}{2}T\right)e^{-\left(\frac{\pi(2k-1)}{2}\right)^2\frac{T}{2}(s-\xi)}d\xi.$$

Note that this equation is, in fact, the explicit form of the operator equation (6.22) with an explicitly calculated integrand including the operator exponential.

Substituting the corresponding interpolation polynomials and collocating the obtained equation at the points

$$s_l=\cos\left[\frac{2l-1}{2n}\pi\right],\quad l=\overline{1,n},$$

analogously as in (5.6), we arrive at a system of linear algebraic equations with respect to $y_j\approx u(1,t_j)$, where the calculation of the values β_{kj} and $\partial_0\phi_k$ is reduced to the calculation of integrals of the type $\int_{-1}^{s}e^{-a(s-\tau)}\tau^n d\tau$ and can be performed exactly. The computations were provided in Maple 9. For the cases of $n=2,3,4,8$ we used Digits=20', whereas for the case $n=16$ we used Digits=40'. The results of the computation presented in Tables 7.1–7.4 confirm our theory above.

TABLE 7.1
The error in the case $n=2$, $T=1$.

Point t	ε
0.8535533905	0.64328761e-2
0.1464466094	0.1817083e-2

TABLE 7.2
The error in the case $n=4$, $T=1$.

Point t	ε
0.9619397662	0.11818295e-4
0.6913417161	0.98439031e-5
0.3086582838	0.71990794e-5
0.0380602337	0.18751078e-5

TABLE 7.3
The error in the case $n=8$, $T=1$.

Point t	ε
0.9903926402	0.23186560e-11
0.9157348061	0.26419501e-11
0.7777851165	0.2614635e-11
0.5975451610	0.2367290e-11
0.4024548389	0.2059591e-11
0.2222148834	0.1600175e-11
0.0842651938	0.1298227e-11
0.0096073597	0.3285404e-12

TABLE 7.4
The error in the case $n = 16$, $T = 1$.

Point t	ε
0.9975923633	0.25187858260140e-26
0.9784701678	0.17507671148106e-26
0.9409606321	0.26606504614832e-26
0.8865052266	0.19216415631333e-26
0.8171966420	0.14389447843978e-26
0.7356983684	0.8729247130003e-27
0.6451423386	0.12250410788313e-26
0.5490085701	0.20057933602212e-26
0.4509914298	0.27443715793522e-26
0.3548576613	0.2071356401636e-26
0.2643016315	0.1352407675441e-26
0.1828033579	0.4176006046401e-26
0.1134947733	0.5890397288154e-26
0.0590393678	0.49643071858013e-25
0.0215298321	0.4614912514700e-26
0.0024076366	0.12497053453746e-25

Appendix A. Proof of Theorem 3.1. Using the Hölder inequality and assumptions (A1)–(A3) we have

(A.1)

$$\left(\int_0^t \|V^j(\tau_1)s(\cdot)\|^q d\tau_1\right)^{1/q} = \left(\int_0^t \left\|\partial_0(\tau_1)\int_0^{\tau_1} K(\tau_1-\tau_2)V^{j-1}(\tau_2)s(\cdot)d\tau_2\right\|^q d\tau_1\right)^{1/q}$$

$$\le c\left(\int_0^t \left(\int_0^{\tau_1} \|K(\tau_1-\tau_2)\|\cdot\|V^{j-1}(\tau_2)s(\cdot)\|d\tau_2\right)^q d\tau_1\right)^{1/q}$$

$$\le c\left(\int_0^t \left(\int_0^{\tau_1} \|K(\tau_1-\tau_2)\|^p d\tau_2\right)^{q/p}\cdot\int_0^{\tau_1}\|V^{j-1}(\tau_2)s(\cdot)\|^q d\tau_2 d\tau_1\right)^{1/q}$$

$$\le c^2\left(\int_0^t\int_0^{\tau_1}\|V^{j-1}(\tau_2)s(\cdot)\|^q d\tau_2 d\tau_1\right)^{1/q} \le \cdots$$

$$\le c^{2j}\left(\int_0^t\int_0^{\tau_1}\cdots\int_0^{\tau_j}\|s(\tau_{j+1})\|^q d\tau_{j+1}\ldots d\tau_2 d\tau_1\right)^{1/q}$$

$$\le c^{2j}\left(\int_0^t\int_0^{\tau_1}\cdots\int_0^{\tau_{j-1}} d\tau_j\ldots d\tau_2 d\tau_1\right)^{1/q}\left(\int_0^t\|s(\tau_{j+1})\|^q d\tau_{j+1}\right)^{1/q}$$

$$\le c^{2j}\left(\frac{t^j}{j!}\right)^{1/q}\left(\int_0^t\|s(\tau_{j+1})\|^q d\tau_{j+1}\right)^{1/q} \qquad \forall s(t)\in L_q(0,T;Y),$$

which means that the spectral radius of the operator $V : L_q(0,T;Y) \to L_q(0,T;Y)$ is equal to zero. The general theory of the Volterra integral equations (see, for example, [18, 19] and section 2) yields the existence and uniqueness of the solution of (2.20).

Applying the Hölder inequality to the equation

$$\delta_j(t) = \partial_0(t)\int_0^t K(t-\lambda)\delta_{j-1}(\lambda)d\lambda \tag{A.2}$$

analogously as above, we deduce that

(A.3)
$$\left(\int_0^t \|\delta_j(\tau_1)\|^q d\tau_1\right)^{1/q}$$
$$\leq c^2\left(\int_0^t\int_0^{\tau_1}\|\delta_{j-1}(\tau_2)\|^q d\tau_2 d\tau_1\right)^{1/q} \leq \cdots \leq c^{2j}\left(\frac{t^j}{j!}\right)^{1/q}\left(\int_0^t \|\delta_0(\tau_{j+1})\|^q d\tau_{j+1}\right)^{1/q},$$

from which the factorial convergence (3.6) follows.

Further, let us prove the stability of the solution of (2.20) with respect to the right-hand side.

Using the Hölder inequality for integrals we obtain

(A.4)
$$\left(\int_0^t \|\partial_0(\tau)u(\tau)\|_Y^q d\tau\right)^{1/q} \leq \left(\int_0^t \|F(\tau)\|_Y^q d\tau\right)^{1/q}$$
$$+\left\{\int_0^t\left\|\partial_0(\tau)\int_0^\tau A^{(1)}e^{-A^{(1)}(\tau-\lambda)}B\partial_0(\lambda)u(\lambda)d\lambda\right\|_Y^q d\tau\right\}^{1/q} \leq \left(\int_0^t \|F(\tau)\|_Y^q d\tau\right)^{1/q}$$
$$+\left\{\int_0^t \|\partial_0(\tau)\|^q\left(\int_0^\tau \|A^{(1)}e^{-A^{(1)}(\tau-\lambda)}B\|_{Y\to X}\|\partial_0(\lambda)u(\lambda)\|_Y d\lambda\right)^q d\tau\right\}^{1/q}.$$

Applying condition (3.1) to the last summand and then again the Hölder inequality, we arrive at the estimate

(A.5)
$$\left(\int_0^t \|\partial_0(\tau)u(\tau)\|_Y^q d\tau\right)^{1/q} \leq \left(\int_0^t \|F(\tau)\|_Y^q d\tau\right)^{1/q}$$
$$+c\left\{\int_0^t\left(\int_0^\tau \|A^{(1)}e^{-A^{(1)}(\tau-\lambda)}B\|_{Y\to X}^p d\lambda\right)^{q/p}\int_0^\tau \|\partial_0(\lambda)u(\lambda)\|_Y^q d\lambda d\tau\right\}^{1/q}.$$

Due to (3.2) we further obtain

(A.6)
$$\left(\int_0^t \|\partial_0(\tau)u(\tau)\|_Y^q d\tau\right)^{1/q} \leq \left(\int_0^t \|F(\tau)\|_Y^q d\tau\right)^{1/q}$$
$$+c^2\left\{\int_0^t\int_0^\tau \|\partial_0(\lambda)u(\lambda)\|_Y^q d\lambda d\tau\right\}^{1/q}.$$

The well-known inequality $(a+b)^q \leq 2^{q/p}(a^q+b^q)$ and the last estimate imply

(A.7)
$$\int_0^t \|\partial_0(\tau)u(\tau)\|_Y^q d\tau \leq 2^{q/p}\left[\int_0^t \|F(\tau)\|_Y^q d\tau\right.$$
$$\left.+c^{2q}\int_0^t\int_0^\tau \|\partial_0(\lambda)u(\lambda)\|_Y^q d\lambda d\tau\right].$$

Now the Gronwall lemma [4, 3] yields (3.7).

The proof is complete. ∎

Appendix B. Proof of Theorem 4.1.

Let us introduce the linear operator $\mathcal{V}(t)$ by

$$\mathcal{V}(t)\mathcal{U}(\cdot) = \mathcal{D}(t)\int_0^t \mathcal{E}_\gamma \mathcal{K}(t,\tau)\mathcal{E}_\gamma^{-1}\mathcal{U}(\tau)d\tau. \tag{B.1}$$

Using assumptions (B1) and (B3) we get the following for the element $[\mathcal{E}_\gamma \mathcal{K}\mathcal{E}_\gamma^{-1}]_{1,1}$ of the matrix $\mathcal{E}_\gamma \mathcal{K}\mathcal{E}_\gamma^{-1}$ with the indexes $(1,1)$:

$$\left\| \left[\mathcal{E}_\gamma \mathcal{K}(\tau_1,\tau_2)\mathcal{E}_\gamma^{-1}\right]_{1,1} \right\| = \left\| \left[A^{(2)}\right]^\gamma e^{-A^{(2)}(\tau_1-\tau_2)} \left[A(\tau_2)-A\right] \left[A^{(2)}\right]^{-\gamma} \right\|$$

$$\le \left\| \left[A^{(2)}\right]^\gamma e^{-A^{(2)}(\tau_1-\tau_2)} \right\| \left\| [A(\tau_2)-A] \left[A^{(2)}\right]^{-\gamma} \right\| \le c^2(\tau_1-\tau_2)^{-\gamma}(T-\tau_2).$$

This estimate implies

$$\begin{aligned} &\int_0^{\tau_1} \| \left[\mathcal{E}_\gamma K(\tau_1,\tau_2)\mathcal{E}_\gamma^{-1}\right]_{1,1} \|^p d\tau_2 \le c^{2p}\int_0^{\tau_1} (\tau_1-\tau_2)^{-\gamma p}(T-\tau_2)^p d\tau_2 \\ &\le c^{2p}\tau_1^{1-\gamma p}T^p \int_0^1 (1-\eta)^{-\gamma p} d\eta = c^{2p}\tau_1^{1-\gamma p}\frac{T^p}{1-\gamma p} \le C^p, \end{aligned} \tag{B.2}$$

with some new constant C. This expression remains bounded for $\gamma p \in [0,1)$.

Assumption (B6) implies

$$\begin{aligned} &\left(\int_0^{\tau_1} \| \left[\mathcal{E}_\gamma K(\tau_1,\tau_2)\mathcal{E}_\gamma^{-1}\right]_{1,2} \|^p d\tau_2\right)^{1/p} \\ &= \left(\int_0^{\tau_1} \left\| \left[A^{(2)}\right]^{\gamma+1} e^{-A^{(2)}(\tau_1-\tau_2)} B^{(1)} \right\|^p d\tau_2\right)^{1/p} \le C, \\ &\left(\int_0^{\tau_1} \| \left[\mathcal{E}_\gamma K(\tau_1,\tau_2)\mathcal{E}_\gamma^{-1}\right]_{2,2} \|^p d\tau_2\right)^{1/p} \\ &= \left(\int_0^{\tau_1} \|A^{(2)} e^{-A^{(2)}(\tau_1-\tau_2)} B^{(1)}\|^p d\tau_2\right)^{1/p} \le C. \end{aligned} \tag{B.3}$$

Using (B2) and (B3) we have

$$\begin{aligned} &\left(\int_0^{\tau_1} \| \left[\mathcal{E}_\gamma K(\tau_1,\tau_2)\mathcal{E}_\gamma^{-1}\right]_{2,1} \|^p d\tau_2\right)^{1/p} \\ &= \left(\int_0^{\tau_1} \left\| e^{-A^{(2)}(\tau_1-\tau_2)} \left[A(\tau_2)-A\right] \left[A^{(2)}\right]^{-\gamma} \right\|^p d\tau_2\right)^{1/p} \\ &\le \left(\int_0^{\tau_1} \left\| [A(\tau_2)-A] \left[A^{(2)}\right]^{-\gamma} \right\|^p d\tau_2\right)^{1/p} \le c\left(\int_0^{\tau_1} (T-\tau_2)^p d\tau_2\right)^{1/p} \le C. \end{aligned} \tag{B.4}$$

Therefore

$$\|\mathcal{E}_\gamma K(\tau_1,\tau_2)\mathcal{E}_\gamma^{-1}\|_\infty \le 2\max_{i,j}\|\left[\mathcal{E}_\gamma K(\tau_1,\tau_2)\mathcal{E}_\gamma^{-1}\right]_{i,j}\| \le C. \tag{B.5}$$

Assumption (B5) yields

$$\|\mathcal{D}(\tau_1)\|_1 \le \max\{1, \|\partial_0(\tau)\|\} \le C. \tag{B.6}$$

Now using estimates (B.5) and (B.6) we get for the nth power of the operator $\mathcal{V}(t)$

(B.7)

$$\begin{aligned}
&\left(\int_0^t \|\mathcal{V}^j(\tau_1)s(\cdot)\|^q d\tau_1\right)^{1/q} \\
&= \left(\int_0^t \left\|\mathcal{D}(\tau_1)\int_0^{\tau_1}\mathcal{E}_\gamma K(\tau_1,\tau_2)\mathcal{E}_\gamma^{-1}\mathcal{V}^{j-1}(\tau_2)s(\cdot)d\tau_2\right\|^q d\tau_1\right)^{1/q} \\
&\le \left(\int_0^t \|\mathcal{D}(\tau_1)\|_\infty^q \left(\int_0^{\tau_1}\|\mathcal{E}_\gamma K(\tau_1,\tau_2)\mathcal{E}_\gamma^{-1}\mathcal{V}^{j-1}(\tau_2)s(\cdot)\|_\infty d\tau_2\right)^q d\tau_1\right)^{1/q} \\
&\le C\left(\int_0^t \left(\int_0^{\tau_1}\|\mathcal{E}_\gamma K(\tau_1,\tau_2)\mathcal{E}_\gamma^{-1}\|_\infty\|\mathcal{V}^{j-1}(\tau_2)s(\cdot)\|_\infty d\tau_2\right)^q d\tau_1\right)^{1/q} \\
&\le C\left(\int_0^t \left(\int_0^{\tau_1}\|\mathcal{E}_\gamma K(\tau_1,\tau_2)\mathcal{E}_\gamma^{-1}\|_\infty^p d\tau_2\right)^{q/p}\cdot\int_0^{\tau_1}\|\mathcal{V}^{j-1}(\tau_2)s(\cdot)\|_\infty^q d\tau_2 d\tau_1\right)^{1/q} \\
&\le C^2\left(\int_0^t\int_0^{\tau_1}\|\mathcal{V}^{j-1}(\tau_2)s(\cdot)\|_\infty^q d\tau_2 d\tau_1\right)^{1/q} \le \cdots \\
&\le C^{2j}\left(\int_0^t\int_0^{\tau_1}\cdots\int_0^{\tau_j}\|s(\tau_{j+1})\|_\infty^q d\tau_{j+1}\ldots d\tau_2 d\tau_1\right)^{1/q} \\
&\le C^{2j}\left(\int_0^t\int_0^{\tau_1}\cdots\int_0^{\tau_{j-1}} d\tau_j\ldots d\tau_2 d\tau_1\right)^{1/q}\left(\int_0^t\|s(\tau_{j+1})\|_\infty^q d\tau_{j+1}\right)^{1/q} \\
&\le C^{2j}\left(\frac{t^j}{j!}\right)^{1/q}\left(\int_0^t\|s(\tau_{j+1})\|_\infty^q d\tau_{j+1}\right)^{1/q} \qquad \forall s(t)\in\mathcal{Y}.
\end{aligned}$$

For the difference

$$\Delta_j(t) = \mathcal{U}_j(t) - \mathcal{U}^*(t)$$

between the jth iteration and the exact solution, we have the equation

$$\Delta_j(t) = \mathcal{D}(t)\int_0^t \mathcal{E}_\gamma K(t,\tau)\mathcal{E}_\gamma^{-1}\Delta_{j-1}(\tau)d\tau. \tag{B.8}$$

Applying the Hölder inequality analogously as above, we get

$$\begin{aligned}
&\left(\int_0^t \|\Delta_j(\tau_1)\|_\infty^q d\tau_1\right)^{1/q} \\
&\le c^2 \left(\int_0^t \int_0^{\tau_1} \|\Delta_{j-1}(\tau_2)\|_\infty^q d\tau_2 d\tau_1\right)^{1/q} \le \cdots \\
&\le c^{2j} \left(\frac{t^j}{j!}\right)^{1/q} \left(\int_0^t \|\Delta_0(\tau_{j+1})\|_\infty^q d\tau_{j+1}\right)^{1/q},
\end{aligned} \tag{B.9}$$

from which the factorial convergence (4.23) follows.

Finally, let us prove the stability of the solution of (4.19) with respect to the right-hand side.

Using the Hölder inequality for integrals we obtain

$$\begin{aligned}
\left(\int_0^t \|\mathcal{U}_\gamma(\tau)\|_\infty^q d\tau\right)^{1/q} &\le \left(\int_0^t \|\mathcal{F}(\tau)\|_\infty^q d\tau\right)^{1/q} \\
&+ \left\{\int_0^t \left\|\mathcal{D}(\tau_1) \int_0^{\tau_1} \mathcal{E}_\gamma K(\tau_1,\tau_2) \mathcal{E}_\gamma^{-1} \mathcal{U}_\gamma(\tau_2) d\tau_2\right\|_\infty^q d\tau_1\right\}^{1/q}.
\end{aligned}$$

Using estimates as above we obtain

$$\left(\int_0^t \|\mathcal{U}_\gamma(\tau)\|_\infty^q d\tau\right)^{1/q} \le c^2 \left(\int_0^t \int_0^{\tau_1} \|\mathcal{U}_\gamma(\tau_2)\|_\infty^q d\tau_2 d\tau_1\right)^{1/q} + \left(\int_0^t \|\mathcal{F}(\tau)\|_\infty^q d\tau\right)^{1/q}.$$

This estimate together with the inequality $(a+b)^q \le 2^{q/p}(a^q + b^q)$ implies

$$\int_0^t \|\mathcal{U}_\gamma(\tau)\|_\infty^q d\tau \le 2^{q/p} \left[c^{2q} \int_0^t \int_0^{\tau_1} \|\mathcal{U}_\gamma(\tau_2)\|_\infty^q d\tau_2 d\tau_1 + \int_0^t \|\mathcal{F}(\tau)\|_\infty^q d\tau\right]. \tag{B.10}$$

Now, the Gronwall lemma yields (4.24).

The proof is complete. □

Acknowledgment. The authors thank the anonymous referees for their suggestions which have led to an improvement of the presentation.

REFERENCES

[1] A. Ashyralyev and P. Sobolevskii, *Well-Posedness of Parabolic Difference Equations*, Birkhäuser-Verlag, Basel, 1994.

[2] K. I. Babenko, *Foundations of the Numerical Analysis*, Nauka, Moscow, 1986 (in Russian).

[3] C. Canuto, M. Y. Hussaini, A. Quarteroni, and T. A. Zang, *Spectral Methods in Fluid Dynamics*, Springer-Verlag, Berlin, 1988.

[4] Ju. L. Dalezkij and M. G. Krein, *Stability of Solutions of Differential Equations in Banach Space*, Nauka, Moscow, 1970 (in Russian).

[5] M. Filali and M. Moussi, *Non-autonomous inhomogeneous boundary Cauchy problem on retarded equations*, Southwest J. Pure Appl. Math., 2 (2003), pp. 26–37.

[6] H. Fujita, N. Saito, and T. Suzuki, *Operator Theory and Numerical Methods*, North–Holland, Amsterdam, 2001.

[7] I. P. Gavrilyuk, W. Hackbusch, and B. N. Khoromskij, *$\mathcal{H}$-matrix approximation for elliptic solution operators in cylinder domains*, East-West J. Numer. Math., 9 (2001), pp. 25–58.

[8] I. P. Gavrilyuk, W. Hackbusch, and B. N. Khoromskij, *$\mathcal{H}$-matrix approximation for the operator exponential with applications*, Numer. Math., 92 (2002), pp. 83–111.
[9] I. P. Gavrilyuk, W. Hackbusch, and B. N. Khoromskij, *Data-sparse approximation to the operator-valued functions of elliptic operator*, Math. Comp., 73 (2004), pp. 1297–1324.
[10] I. P. Gavrilyuk, W. Hackbusch, and B. N. Khoromskij, *Data-sparse approximation of a class of operator-valued functions*, Math. Comp., 74 (2005), pp. 681–708.
[11] I. P. Gavrilyuk, W. Hackbusch, and B. N. Khoromskij, *Tensor-product approximation to elliptic and parabolic solution operators in higher dimensions*, Computing, 74 (2005), pp. 131–157.
[12] I. P. Gavrilyuk and V. L. Makarov, *Representation and approximation of the solution of an initial value problem for a first order differential equation in Banach space*, Z. Anal. Anwendungen, 15 (1996), pp. 495–527.
[13] I. P. Gavrilyuk and V. L. Makarov, *An explicit boundary integral representation of the solution of the two-dimensional heat equation and its discretization*, J. Integral Equations Appl., 12 (2000), pp. 63–83.
[14] I. P. Gavrilyuk and V. L. Makarov, *Exponentially convergent parallel discretization methods for the first order evolution equations*, Comput. Methods Appl. Math., 1 (2001), pp. 333–355.
[15] I. P. Gavrilyuk and V. L. Makarov, *Exponentially convergent algorithms for the operator exponential with applications to inhomogeneous problems in Banach spaces*, SIAM J. Numer. Anal., 43 (2005), pp. 2144–2171.
[16] J. A. Goldstein, *Semigroups of Linear Operators and Applications*, The Clarendon Press, Oxford University Press, New York, 1985.
[17] T. Kato, *On linear differential equations in Banach spaces*, Comm. Pure Appl. Math., 9 (1956), pp. 479–486.
[18] M. A. Krasnosel'skij, G. M. Vainikko, P. P. Zabrejko, Ja. B. Rutizkij, and V. Ja. Stezenko, *Approximate Solution of Operator Equations*, Nauka, Moscow, 1969 (in Russian).
[19] M. A. Krasnosel'skii, G. M. Vainikko, P. P. Zabreiko, Ja. B. Rutitskii, and V. Ja. Stezenko, *Approximate Solution of Operator Equations*, Wolters-Noordhoff, Groningen, The Netherlands, 1972.
[20] O. A. Ladyzhenskaja, V. A. Solonnikov, and N. N. Uralzeva, *Linear and Quasilinear Parabolic Equations*, Nauka, Moscow, 1967 (in Russian).
[21] M. López-Fernández, C. Lubich, C. Palencia, and A. Schädle, *Fast Runge-Kutta approximation of inhomogeneous parabolic differential equations*, Numer. Math., 102 (2005), pp. 277–291.
[22] M. López-Fernández and C. Palencia, *On the numerical inversion of the Laplace transform of certain holomorphic mappings*, Appl. Numer. Math., 51 (2004), pp. 289–303.
[23] M. López-Fernández, C. Palencia, and A. Schädle, *A spectral order method for inverting sectorial Laplace transforms*, SIAM J. Numer. Anal., 44 (2006), pp. 1332–1350.
[24] V. A. Marchenko, *Sturm-Liouville Operators and Applications*, Oper. Theory Adv. Appl. 22, Birkhäuser-Verlag, Basel, 1986.
[25] C. Palencia, *Backward Euler method for abstract time-dependent parabolic equations with variable domains*, Numer. Math., 82 (1999), pp. 471–490.
[26] A. Pazy, *Semigroups of Linear Operator and Applications to Partial Differential Equations*, Springer-Verlag, New York, 1983.
[27] G. N. Polozhij, *Equations of Mathematical Physics*, Vysshaja shkola, Moscow, 1964 (in Russian).
[28] G. Szegö, *Orthogonal Polynomials*, AMS, Providence, RI, 1959.
[29] G. Szegö, *Orthogonal Polynomials*, State Publishing House of Physical and Mathematical Literature, Moscow, 1962 (in Russian).
[30] H. Triebel, *Interpolation Theory, Function Spaces, Differential Operators*, North-Holland, Amsterdam, 1978.
[31] W. Wendland, *Integral Equations Method for Boundary Value Problems*, Springer-Verlag, Berlin, 2002.

SIAM J. NUMER. ANAL.
Vol. 46, No. 1, pp. 397–418
© 2008 Society for Industrial and Applied Mathematics

MODIFIED NODAL CUBIC SPLINE COLLOCATION FOR POISSON'S EQUATION†

ABEER ALI ABUSHAMA‡ AND BERNARD BIALECKI†

Abstract. We present a new modified nodal cubic spline collocation scheme for solving the Dirichlet problem for Poisson's equation on the unit square. We prove existence and uniqueness of a solution of the scheme and show how the solution can be computed on an $(N+1)\times(N+1)$ uniform partition of the square with cost $O(N^2\log N)$ using a direct fast Fourier transform method. Using two comparison functions, we derive an optimal fourth order error bound in the continuous maximum norm. We compare our scheme with other modified nodal cubic spline collocation schemes; in particular, the one proposed by Houstis, Vavalis, and Rice in [*SIAM J. Numer. Anal.*, 25 (1988), pp. 54–74]. We believe that our paper gives the first correct convergence analysis of a modified nodal cubic spline collocation for solving partial differential equations.

Key words. nodal collocation, cubic splines, convergence analysis, interpolants

AMS subject classifications. 65N35, 65N12, 65N15, 65N22

DOI. 10.1137/050629033

1. Introduction.

De Boor [7] proved that classical nodal cubic spline collocation for solving two-point boundary value problems is only second order accurate and no better. For two-point boundary value problems, Archer [2] and, independently, Daniel and Swartz [6] developed a modified nodal cubic spline collocation (MNCSC) scheme which is fourth order accurate. The approximate solution in this scheme satisfies higher-order perturbations of the ordinary differential equation at the partition nodes. Based on the method of [2] and [6], Houstis, Vavalis, and Rice [8] derived a fourth order MNCSC scheme for solving elliptic boundary value problems on rectangles. For the Helmholtz equation, a direct fast Fourier transform (FFT) algorithm for solving this scheme was proposed recently in [3].

In this paper, we consider the Dirichlet boundary value problem for Poisson's equation

$$\Delta u = f \text{ in } \Omega, \qquad u = 0 \text{ on } \partial\Omega, \tag{1.1}$$

where Δ denotes the Laplacian, $\Omega = (0,1)\times(0,1)$, and $\partial\Omega$ is the boundary of Ω. Let $\rho_x = \{x_i\}_{i=0}^{N+1}$ be a uniform partition of $[0,1]$ in the x-direction such that $x_i = ih$, $i = 0,\dots,N+1$, where $h = 1/(N+1)$. For the sake of simplicity, we assume that a uniform partition $\rho_y = \{y_j\}_{i=0}^{N+1}$ of $[0,1]$ in the y-direction is such that $y_j = x_j$. Let S_3 be the space of cubic splines defined by

$$S_3 = \{v \in C^2[0,1] : v|_{[x_{i-1},x_i]} \in P_3, i = 1,\dots,N+1\},$$

where P_3 denotes the set of all polynomials of degree ≤ 3, and let

$$S^D = \{v \in S_3 : v(0) = v(1) = 0\}.$$

†Received by the editors April 12, 2005; accepted for publication (in revised form) August 17, 2007; published electronically January 25, 2008.
http://www.siam.org/journals/sinum/46-1/62903.html
‡Department of Mathematical and Computer Sciences, Colorado School of Mines, Golden, CO 80401-1887 (ashamaa@gmail.com, bbialeck@mines.edu).

Our MNCSC scheme for solving (1.1) is formulated as follows: Find $u_h \in S^D \otimes S^D$ such that

$$\Delta u_h(x_i,y_j) - \frac{h^2}{6} D_x^2 D_y^2 u_h(x_i,y_j) = f(x_i,y_j) - \frac{h^2}{12}\Delta f(x_i,y_j), \quad i,j = 0,\dots,N+1. \tag{1.2}$$

The scheme (1.2) is motivated by the fourth order finite difference method for (1.1); see, for example, equation (7) in section 4.5 of [9]. Using $u_h = u = 0$ on $\partial\Omega$ and (1.1), we see that (1.2) is equivalent to

$$2D_x^2 D_y^2 u_h(x_i,y_j) = \Delta f(x_i,y_j), \quad i,j = 0, N+1, \tag{1.3}$$

$$D_x^2 u_h(x_i,y_j) - \frac{h^2}{6} D_x^2 D_y^2 u_h(x_i,y_j) = f(x_i,y_j) - \frac{h^2}{12}\Delta f(x_i,y_j), \quad i = 0, N+1, \quad j = 1,\dots,N, \tag{1.4}$$

$$D_y^2 u_h(x_i,y_j) - \frac{h^2}{6} D_x^2 D_y^2 u_h(x_i,y_j) = f(x_i,y_j) - \frac{h^2}{12}\Delta f(x_i,y_j), \quad i = 1,\dots,N, \quad j = 0, N+1, \tag{1.5}$$

$$\Delta u_h(x_i,y_j) - \frac{h^2}{6} D_x^2 D_y^2 u_h(x_i,y_j) = f(x_i,y_j) - \frac{h^2}{12}\Delta f(x_i,y_j), \quad i,j = 1,\dots,N. \tag{1.6}$$

The scheme (4.2)–(4.4) of [8] for (1.1) is: Find $u_h \in S^D \otimes S^D$ satisfying (1.3) and

$$D_x^2 u_h(x_i,y_j) = f(x_i,y_j), \quad i = 0, N+1, \quad j = 1,\dots,N, \tag{1.7}$$

$$D_y^2 u_h(x_i,y_j) = f(x_i,y_j), \quad i = 1,\dots,N, \quad j = 0, N+1, \tag{1.8}$$

$$(L_x + L_y) u_h(x_i,y_j) = f(x_i,y_j), \quad i,j = 1,\dots,N, \tag{1.9}$$

where, for $i,j = 1,\dots,N$,

$$L_x v(x_i,y_j) = \frac{1}{12}\left[D_x^2 v(x_{i-1},y_j) + 10 D_x^2 v(x_i,y_j) + D_x^2 v(x_{i+1},y_j)\right], \tag{1.10}$$

$$L_y v(x_i,y_j) = \frac{1}{12}\left[D_y^2 v(x_i,y_{j-1}) + 10 D_y^2 v(x_i,y_j) + D_y^2 v(x_i,y_{j+1})\right].$$

Our scheme and that of [8] are identical at the corners of $\overline{\Omega}$. However, they are different at the remaining partition nodes. While (1.4)–(1.6) involve perturbations of both the left- and right-hand sides, (1.9) involves a perturbation of the left-hand side only. Numerical results show that our scheme exhibits superconvergence phenomena while that of [8] does not.

An outline of this paper is as follows. We give preliminaries in section 2. The matrix-vector form of our scheme, an existence and uniqueness proof of its solution, and a direct FFT algorithm for solving the scheme are presented in section 3. In section 4, using two comparison functions, we derive a fourth order error bound in the continuous maximum norm. In section 5, we give convergence analysis of the scheme in [4] that consists of (1.3)–(1.5) and (1.9). We also explain why convergence analysis of the scheme (1.3) and (1.7)–(1.9), given in [8], is incorrect. This is why, we believe, our paper gives the first correct convergence analysis of MNCSC for solving partial differential equations. Section 6 includes numerical results obtained using our scheme.

2. Preliminaries. We extend the uniform partition $\rho_x = \{x_i\}_{i=0}^{N+1}$ outside of $[0,1]$ using $x_i = ih, i = -3,-2,-1,N+2,N+3,N+4$, and introduce $I_i = [x_{i-1}, x_i]$, $i = -2,\dots,N+4$. Let $\{B_m\}_{m=-1}^{N+2}$ be the basis for S_3 defined by

$$B_m(x) = \begin{cases} g_1[(x - x_{m-2})/h], & x \in I_{m-1}, \\ g_2[(x - x_{m-1})/h], & x \in I_m, \\ g_2[(x_{m+1} - x)/h], & x \in I_{m+1}, \\ g_1[(x_{m+2} - x)/h], & x \in I_{m+2}, \\ 0 & \text{otherwise}, \end{cases} \tag{2.1}$$

where

$$g_1(x) = x^3, \qquad g_2(x) = 1 + 3x + 3x^2 - 3x^3. \tag{2.2}$$

The basis functions are such that, for $m = 0,\dots,N+1$,

$$\begin{array}{lll} B_{m-1}(x_m) = 1, & B_m(x_m) = 4, & B_{m+1}(x_m) = 1, \\ B''_{m-1}(x_m) = 6/h^2, & B''_m(x_m) = -12/h^2, & B''_{m+1}(x_m) = 6/h^2. \end{array} \tag{2.3}$$

Let $\{B_m^D\}_{m=0}^{N+1}$ be the basis for S^D defined by

$$\begin{array}{l} B_0^D = B_0 - 4B_{-1}, \qquad B_1^D = B_1 - B_{-1}, \\ B_m^D = B_m, \quad m = 2,\dots,N-1, \\ B_N^D = B_N - B_{N+2}, \qquad B_{N+1}^D = B_{N+1} - 4B_{N+2}. \end{array} \tag{2.4}$$

It follows from (2.3) that

$$\begin{array}{lll} B_0^D(x_1) = 1, & B_1^D(x_1) = 4, & B_1^D(x_2) = 1, \\ B_N^D(x_{N-1}) = 1, & B_N^D(x_N) = 4, & B_{N+1}^D(x_N) = 1, \end{array} \tag{2.5}$$

$$\begin{array}{lll} \left[B_0^D\right]''(x_0) = -36/h^2, & \left[B_1^D\right]''(x_0) = 0, & \\ \left[B_0^D\right]''(x_1) = 6/h^2, & \left[B_1^D\right]''(x_1) = -12/h^2, & \left[B_1^D\right]''(x_2) = 6/h^2, \\ \left[B_N^D\right]''(x_{N-1}) = 6/h^2, & \left[B_N^D\right]''(x_N) = -12/h^2, & \left[B_{N+1}^D\right]''(x_N) = 6/h^2, \\ \left[B_N^D\right]''(x_{N+1}) = 0, & \left[B_{N+1}^D\right]''(x_{N+1}) = -36/h^2. & \end{array} \tag{2.6}$$

It also follows from (2.5), (2.6), (2.4), and (2.3) that, for $i = 1,\dots,N$,

$$B_m^D(x_i) - \frac{h^2}{6}[B_m^D]''(x_i) = \begin{cases} 6, & m = i, \\ 0, & m \neq i, \end{cases} \qquad m = 0,\dots,N+1. \tag{2.7}$$

Throughout this paper, C denotes a generic positive constant that is independent of u and h.

LEMMA 2.1. *$\{B_m^D\}_{m=0}^{N+1}$ of (2.4) satisfy $\max_{x\in[0,1]}|B_m^D(x)| \le C, m = 0,\dots,N+1$.*

Proof. For each fixed $m = -1,\dots,N+2$, using $I_i = [x_{i-1}, x_i]$ and $x_i = ih$, we have

$$\begin{array}{llll} 0 \le (x - x_{m-2})/h \le 1, & x \in I_{m-1}, & 0 \le (x - x_{m-1})/h \le 1, & x \in I_m, \\ 0 \le (x_{m+1} - x)/h \le 1, & x \in I_{m+1}, & 0 \le (x_{m+2} - x)/h \le 1, & x \in I_{m+2}. \end{array} \tag{2.8}$$

Equations (2.2) and (2.8) give

$$\begin{array}{llll} |g_1[(x - x_{m-2})/h]| \le 1, & x \in I_{m-1}, & |g_2[(x - x_{m-1})/h]| \le 7, & x \in I_m, \\ |g_2[(x_{m+1} - x)/h]| \le 7, & x \in I_{m+1}, & |g_1[(x_{m+2} - x)/h]| \le 1, & x \in I_{m+2}. \end{array} \tag{2.9}$$

Using (2.1) and (2.9), we see that $\max_{x\in[0,1]}|B_m(x)| \le C, m = -1, \dots, N+2$. Hence the required inequality follows from (2.4), which implies that each B_m^D is a linear combination of at most two of the functions $\{B_n\}_{n=-1}^{N+2}$. ∎

For $\{B_m^D\}_{m=0}^{N+1}$ of (2.4), we introduce $N \times N$ matrices A and B defined by

$$A = (a_{i,m})_{i,m=1}^{N}, \quad a_{i,m} = [B_m^D]''(x_i), \quad B = (b_{j,n})_{j,n=1}^{N}, \quad b_{j,n} = B_n^D(y_j). \tag{2.10}$$

It follows from (2.4), (2.3), (2.5), and (2.6) that

$$A = 6h^{-2}T, \qquad B = T + 6I, \tag{2.11}$$

where I is the identity matrix and the $N \times N$ matrix T is given by

$$T = \begin{bmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ & & & 1 & -2 \end{bmatrix}. \tag{2.12}$$

LEMMA 2.2. *If $B[u_1, \dots, u_N]^T = [v_1, \dots, v_N]^T$, where B is defined in* (2.10), *then* $\max_{1\le i\le N}|u_i| \le C \max_{1\le i\le N}|v_i|$.

Proof. It follows from (2.10), (2.11), and (2.12) that $|b_{i,i}| - \sum_{i\ne j}|b_{i,j}| \ge 2$, $i = 1, \dots, N$. Hence the required result follows, for example, from the discussion on page 21 in [1]. ∎

In what follows, $[u_{1,1}, \dots, u_{N,N}]^T$ is the short notation for

$$[u_{1,1}, \dots, u_{1,N}, u_{2,1}, \dots, u_{2,N}, \dots, u_{N,1}, \dots, u_{N,N}]^T.$$

LEMMA 2.3. *If $\mathbf{u} = [u_{1,1}, \dots, u_{N,N}]^T$ and $\mathbf{v} = [v_{1,1}, \dots, v_{N,N}]^T$ are such that $(B\otimes B)\mathbf{u} = \mathbf{v}$, where B is defined in* (2.10), *then* $\max_{1\le i,j\le N}|u_{i,j}| \le C \max_{1\le i,j\le N}|v_{i,j}|$.

Proof. Since $B \otimes B = (B \otimes I)(I \otimes B)$, we have

$$\mathbf{v} = (B \otimes I)\mathbf{w}, \qquad \mathbf{w} = (I \otimes B)\mathbf{u}. \tag{2.13}$$

Using (2.13) and Lemma 2.2, we obtain

$$\max_{1\le i,j\le N}|u_{i,j}| \le C \max_{1\le i,j\le N}|w_{i,j}|, \quad \max_{1\le i,j\le N}|w_{i,j}| \le C \max_{1\le i,j\le N}|v_{i,j}|,$$

which imply the required inequality. ∎

It is well known (see Theorem 4.5.2 of [10]) that for T of (2.12), we have

$$QTQ = \Lambda, \qquad QQ = I, \tag{2.14}$$

where the $N \times N$ matrices Λ and Q are given by

$$\Lambda = \operatorname{diag}(\lambda_i)_{i=1}^{N}, \qquad \lambda_i = -4\sin^2\frac{i\pi}{2(N+1)}, \tag{2.15}$$

$$Q = (q_{i,j})_{i,j=1}^{N}, \quad q_{i,j} = \left(\frac{2}{N+1}\right)^{1/2} \sin\frac{ij\pi}{N+1}. \tag{2.16}$$

LEMMA 2.4. *If* $\mathbf{v} = [v_{1,1}, \ldots, v_{N,N}]^T$ *and* $\mathbf{w} = [w_{1,1}, \ldots, w_{N,N}]^T$ *are such that*

$$\left[\frac{T}{h^2} \otimes I + I \otimes \frac{T}{h^2} + \frac{h^2}{6}\left(\frac{T}{h^2} \otimes \frac{T}{h^2}\right)\right] \mathbf{v} = \mathbf{w}, \tag{2.17}$$

where T *is the matrix defined in* (2.12)*, then* $\max_{1\le i,j\le N} v_{i,j}^2 \le Ch^2 \sum_{i=1}^N \sum_{j=1}^N w_{i,j}^2$.

Proof. The matrix in (2.17) arises in the fourth order finite difference method for (1.1). Hence the desired result follows, for example, from the last unnumbered equation on page 296 in [9]. □

Finally, we observe that the matrix-vector form of

$$\phi_{i,j} = \sum_{m=1}^{N} c_{i,m}^{(1)} \sum_{n=1}^{N} c_{j,n}^{(2)} \psi_{m,n}, \qquad i,j = 1, \ldots, N, \tag{2.18}$$

is

$$\boldsymbol{\phi} = (C_1 \otimes C_2)\boldsymbol{\psi}, \tag{2.19}$$

where $C_1 = \left(c_{i,m}^{(1)}\right)_{i,m=1}^{N}$, $C_2 = \left(c_{j,n}^{(2)}\right)_{j,n=1}^{N}$, and

$$\boldsymbol{\phi} = [\phi_{1,1}, \ldots, \phi_{N,N}]^T, \qquad \boldsymbol{\psi} = [\psi_{1,1}, \ldots, , \psi_{N,N}]^T.$$

3. Matrix-vector form of scheme. Since $\dim(S^D \otimes S^D) = (N+2)^2$, the scheme (1.3)–(1.6) involves $(N+2)^2$ equations in $(N+2)^2$ unknowns. Using the basis $\{B_m^D\}_{m=0}^{N+1}$ of (2.4) for the space S^D, we have

$$u_h(x,y) = \sum_{m=0}^{N+1} \sum_{n=0}^{N+1} u_{m,n} B_m^D(x) B_n^D(y). \tag{3.1}$$

Substituting (3.1) into (1.3), we obtain

$$2 \sum_{m=0}^{N+1} \sum_{n=0}^{N+1} u_{m,n} [B_m^D]''(x_i)[B_n^D]''(y_j) = \Delta f(x_i, y_j), \qquad i,j = 0, N+1. \tag{3.2}$$

Using (2.6), we conclude that (3.2) gives

$$u_{i,j} = \frac{h^4}{2592} \Delta f(x_i, y_j), \qquad i,j = 0, N+1. \tag{3.3}$$

Substituting (3.1) into (1.4), we obtain

$$\begin{aligned} &\sum_{m=0}^{N+1} \sum_{n=0}^{N+1} u_{m,n} [B_m^D]''(x_i) \left(B_n^D(y_j) - \frac{h^2}{6}[B_n^D]''(y_j)\right) \\ &\quad = f(x_i, y_j) - \frac{h^2}{12} \Delta f(x_i, y_j), \qquad i = 0, N+1, \quad j = 1, \ldots, N. \end{aligned} \tag{3.4}$$

Using (2.6) and (2.7), we see that (3.4) gives

$$u_{i,j} = -\frac{h^2}{216} f(x_i, y_j) + \frac{h^4}{2592} \Delta f(x_i, y_j), \quad i = 0, N+1, \quad j = 1, \ldots, N. \tag{3.5}$$

Using (3.5) and symmetry with respect to x and y, we conclude that (1.5) gives

$$u_{i,j} = -\frac{h^2}{216} f(x_i, y_j) + \frac{h^4}{2592} \Delta f(x_i, y_j), \quad i = 1, \ldots, N, \quad j = 0, N+1. \tag{3.6}$$

Substituting (3.1) into (1.6), we obtain

$$\begin{aligned} \sum_{m=0}^{N+1} \sum_{n=0}^{N+1} u_{m,n} \left([B_m^D]''(x_i) B_n^D(y_j) + \left[B_m^D(x_i) - \frac{h^2}{6} [B_m^D]''(x_i) \right] [B_n^D]''(y_j) \right) \\ = f(x_i, y_j) - \frac{h^2}{12} \Delta f(x_i, y_j), \quad i, j = 1, \ldots, N. \end{aligned} \tag{3.7}$$

Moving the terms involving $\{u_{m,n}\}_{n=0}^{N+1}$, $m = 0, N+1$, $\{u_{m,n}\}_{m=1}^{N}$, $n = 0, N+1$, to the right-hand side of (3.7), we get

$$\begin{aligned} \sum_{m=1}^{N} \sum_{n=1}^{N} u_{m,n} \left([B_m^D]''(x_i) B_n^D(y_j) + \left[B_m^D(x_i) - \frac{h^2}{6} [B_m^D]''(x_i) \right] [B_n^D]''(y_j) \right) \\ = p_{i,j}, \quad i, j = 1, \ldots, N, \end{aligned} \tag{3.8}$$

where

$$\begin{aligned} p_{i,j} &= f(x_i, y_j) - \frac{h^2}{12} \Delta f(x_i, y_j) \\ &- \sum_{m=0,N+1} \sum_{n=0}^{N+1} u_{m,n} \left([B_m^D]''(x_i) B_n^D(y_j) + \left[B_m^D(x_i) - \frac{h^2}{6} [B_m^D]''(x_i) \right] [B_n^D]''(y_j) \right) \\ &- \sum_{m=1}^{N} \sum_{n=0,N+1} u_{m,n} \left([B_m^D]''(x_i) B_n^D(y_j) + \left[B_m^D(x_i) - \frac{h^2}{6} [B_m^D]''(x_i) \right] [B_n^D]''(y_j) \right). \end{aligned}$$

Using (2.18)–(2.19), we write (3.8) as

$$\left[A \otimes B + \left(B - \frac{h^2}{6} A \right) \otimes A \right] \mathbf{u} = \mathbf{p}, \tag{3.9}$$

where $\mathbf{u} = [u_{1,1}, \ldots, u_{N,N}]^T$, $\mathbf{p} = [p_{1,1}, \ldots, p_{N,N}]^T$, and A, B are defined in (2.10). Using (2.11), we see that

$$A \otimes B + \left(B - \frac{h^2}{6} A \right) \otimes A = \frac{6}{h^2} \left[6T \otimes I + (6I + T) \otimes T \right],$$

and hence the system (3.9) simplifies to

$$6h^{-2} \left[6T \otimes I + (6I + T) \otimes T \right] \mathbf{u} = \mathbf{p}. \tag{3.10}$$

We are now ready to prove the existence and uniqueness of u_h in $S^D \otimes S^D$ which satisfies (1.3)–(1.6).

THEOREM 3.1. *There exists unique u_h in $S^D \otimes S^D$ satisfying* (1.3)–(1.6).

Proof. Since the number of equations in (1.3)–(1.6) is equal to the number of unknowns, we assume that the right-hand side in (1.3)–(1.6) is zero, and show that $u_h = 0$ is the only solution of the resulting scheme. Using (3.1), (3.3), (3.5), and (3.6), we have

$$u_{m,n} = 0, \quad m = 0, N+1, n = 0, \ldots, N+1, \quad m = 1, \ldots, N, n = 0, N+1. \tag{3.11}$$

Clearly

$$6h^{-2}[6T \otimes I + (6I+T) \otimes T] = 36\left[\frac{T}{h^2} \otimes I + I \otimes \frac{T}{h^2} + \frac{h^2}{6}\left(\frac{T}{h^2} \otimes \frac{T}{h^2}\right)\right]. \tag{3.12}$$

Hence it follows from (3.10) with $\mathbf{p}$ replaced by $\mathbf{0}$, (3.12), and Lemma 2.4 that

$$u_{m,n} = 0, \qquad m,n = 1,\ldots,N. \tag{3.13}$$

Equations (3.1), (3.11), and (3.13) give $u_h = 0$. □

Using Q of (2.16), we see that (3.10) is equivalent to

$$6h^{-2}(Q \otimes I)\,[6T \otimes I + (6I+T) \otimes T]\,(Q \otimes I)(Q^{-1} \otimes I)\mathbf{u} = (Q \otimes I)\mathbf{p}. \tag{3.14}$$

Introducing $\mathbf{u}' = (Q^{-1} \otimes I)\mathbf{u}$ and $\mathbf{p}' = (Q \otimes I)\mathbf{p}$, and using (3.14) and (2.14), we obtain

$$6h^{-2}\,[6\Lambda \otimes I + (6I+\Lambda) \otimes T]\mathbf{u}' = \mathbf{p}', \tag{3.15}$$

where Λ is defined in (2.15). The system (3.15) reduces to the N independent systems

$$6h^{-2}\,[6\lambda_i I + (6+\lambda_i)T]\mathbf{u}_i' = \mathbf{p}_i', \qquad i = 1,\ldots,N, \tag{3.16}$$

where $\mathbf{u}_i' = [u_{i,1}',\ldots,u_{i,N}']^T$, $\mathbf{p}_i' = [p_{i,1}',\ldots,p_{i,N}']^T$, $i = 1,\ldots,N$.

We have the following algorithm for solving (3.10):

Step 1. Compute $\mathbf{p}' = (Q \otimes I)\mathbf{p}$.
Step 2. Solve the N systems in (3.16).
Step 3. Compute $\mathbf{u} = (Q \otimes I)\mathbf{u}'$.

It follows from (2.16) and section 4.4.5 in [10] that the cost of performing Steps 1 and 3 is $2.5N^2 \log_2 N$ each. In Step 2, the systems are tridiagonal, so this step is performed at a cost $O(N^2)$. Thus the total cost of the algorithm is $5N^2 \log_2 N + O(N^2)$.

4. Convergence analysis. In what follows, $C(u)$ denotes a generic positive constant that is independent of h, but depends on u.

Our goal is to show that if u in $C^6(\overline{\Omega})$ and u_h in $S^D \otimes S^D$ are the solutions of (1.1) and (1.3)–(1.6), respectively, then

$$\|u - u_h\|_{C(\overline{\Omega})} \le C(u)h^4, \tag{4.1}$$

where $\|g\|_{C(\overline{\Omega})} = \max_{x\in\overline{\Omega}} |g(x)|$ for g in $C(\overline{\Omega})$.

To prove (4.1), for u in $C^4(\overline{\Omega})$, we introduce two comparison functions, the spline interpolants S and Z in $S^D \otimes S^D$ of u defined, respectively, by

$$D_x^2 D_y^2 S(x_i,y_j) = D_x^2 D_y^2 u(x_i,y_j), \qquad i,j = 0,N+1, \tag{4.2}$$

$$\begin{aligned} D_x^2 S(x_i,y_j) - \frac{h^2}{6} D_x^2 D_y^2 S(x_i,y_j) &= D_x^2 u(x_i,y_j) - \frac{h^2}{12} D_x^4 u(x_i,y_j) \\ -\frac{h^2}{6} D_x^2 D_y^2 u(x_i,y_j), &\qquad i = 0,N+1, \quad j = 1,\ldots,N, \end{aligned} \tag{4.3}$$

$$\begin{aligned} D_y^2 S(x_i,y_j) - \frac{h^2}{6} D_x^2 D_y^2 S(x_i,y_j) &= D_y^2 u(x_i,y_j) - \frac{h^2}{12} D_y^4 u(x_i,y_j) \\ -\frac{h^2}{6} D_x^2 D_y^2 u(x_i,y_j), &\qquad i = 1,\ldots,N, \quad j = 0,N+1, \end{aligned} \tag{4.4}$$

$$S(x_i, y_j) = u(x_i, y_j), \quad i, j = 1, \ldots, N, \tag{4.5}$$

and

$$D_x^2 D_y^2 Z(x_i, y_j) = D_x^2 D_y^2 u(x_i, y_j), \quad i, j = 0, N+1, \tag{4.6}$$

$$D_x^2 Z(x_i, y_j) = D_x^2 u(x_i, y_j), \quad i = 0, N+1, \quad j = 1, \ldots, N, \tag{4.7}$$

$$D_y^2 Z(x_i, y_j) = D_y^2 u(x_i, y_j), \quad i = 1, \ldots, N, \quad j = 0, N+1, \tag{4.8}$$

$$Z(x_i, y_j) = u(x_i, y_j), \quad i, j = 1, \ldots, N. \tag{4.9}$$

It follows from (1.1) that

$$f = D_x^2 u + D_y^2 u, \quad \Delta f = D_x^4 u + D_y^4 u + 2 D_x^2 D_y^2 u. \tag{4.10}$$

Hence, using $u = 0$ on $\partial\Omega$, we see that (1.3)–(1.5) reduce, respectively, to

$$D_x^2 D_y^2 u_h(x_i, y_j) = D_x^2 D_y^2 u(x_i, y_j), \quad i, j = 0, N+1, \tag{4.11}$$

$$\begin{aligned} D_x^2 u_h(x_i, y_j) - \frac{h^2}{6} D_x^2 D_y^2 u_h(x_i, y_j) &= D_x^2 u(x_i, y_j) - \frac{h^2}{12} D_x^4 u(x_i, y_j) \\ -\frac{h^2}{6} D_x^2 D_y^2 u(x_i, y_j), &\quad i = 0, N+1, \quad j = 1, \ldots, N, \end{aligned} \tag{4.12}$$

$$\begin{aligned} D_y^2 u_h(x_i, y_j) - \frac{h^2}{6} D_x^2 D_y^2 u_h(x_i, y_j) &= D_y^2 u(x_i, y_j) - \frac{h^2}{12} D_y^4 u(x_i, y_j) \\ -\frac{h^2}{6} D_x^2 D_y^2 u(x_i, y_j), &\quad i = 1, \ldots, N, \quad j = 0, N+1. \end{aligned} \tag{4.13}$$

Comparing (4.11)–(4.13) and (4.2)–(4.4), we see that u_h and S are defined in the same way for $i = 0, N+1, j = 0, \ldots, N+1$, and $i = 1, \ldots, N$, $j = 0, N+1$. On the other hand, (4.6)–(4.8) are a simplified, tensor product version of (4.2)–(4.4).

The triangle inequality gives

$$\|u - u_h\|_{C(\overline{\Omega})} \le \|u - Z\|_{C(\overline{\Omega})} + \|Z - S\|_{C(\overline{\Omega})} + \|S - u_h\|_{C(\overline{\Omega})}. \tag{4.14}$$

In what follows, we bound the three terms on the right-hand side of (4.14).

4.1. Bounding $\|u - Z\|_{C(\overline{\Omega})}$. We need the following results.

LEMMA 4.1. *Let the interpolant $I_x v$ in S_3 of v in $C^2[0,1]$ be defined by*

$$(I_x v)''(x_i) = v''(x_i), \quad i = 0, N+1, \qquad I_x v(x_i) = v(x_i), \quad i = 0, \ldots, N+1. \tag{4.15}$$

Then

$$\max_{x \in [0,1]} |v(x) - I_x v(x)| \le C \max_{x \in [0,1]} |v''(x)| h^2. \tag{4.16}$$

If $v \in C^4[0,1]$, then

$$\max_{x \in [0,1]} |v(x) - I_x v(x)| \le C \max_{x \in [0,1]} |v^{(4)}(x)| h^4. \tag{4.17}$$

Proof. First, we prove (4.16). Using the discussion on page 404 in [5], we have

$$I_x v(x) = v(x_i) + B_i(x - x_i) + C_i(x - x_i)^2 + D_i(x - x_i)^3, \quad x \in [x_i, x_{i+1}], \tag{4.18}$$

for $i = 0, \ldots, N$, where

$$B_i = -\frac{h}{6} r_{i+1} - \frac{h}{3} r_i + \frac{1}{h}\left[v(x_{i+1}) - v(x_i)\right], \quad C_i = \frac{r_i}{2}, \quad D_i = \frac{1}{6h}\left(r_{i+1} - r_i\right), \tag{4.19}$$

and $r_i = (I_x v)''(x_i)$. Equations (4.18) and (4.19) give

$$\begin{aligned} I_x v(x) - v(x) = A_i(x) &- \frac{h}{6} r_{i+1}(x - x_i) - \frac{h}{3} r_i (x - x_i) + \frac{r_i}{2}(x - x_i)^2 \\ &+ \frac{1}{6h}(r_{i+1} - r_i)(x - x_i)^3, \quad x \in [x_i, x_{i+1}], \end{aligned} \tag{4.20}$$

where

$$A_i(x) = v(x_i) - v(x) + \frac{v(x_{i+1}) - v(x_i)}{h}(x - x_i), \quad x \in [x_i, x_{i+1}]. \tag{4.21}$$

Using (4.20) and the triangle inequality, we obtain, for $x \in [x_i, x_{i+1}]$,

$$|I_x v(x) - v(x)| \le |A_i(x)| + h^2 \left(|r_i| + \frac{1}{3}|r_{i+1}|\right) \le |A_i(x)| + \frac{4}{3} h^2 \max_{0 \le i \le N+1} |r_i|. \tag{4.22}$$

We introduce

$$E = (e_{i,j})_{i,j=0}^{N+1} = \begin{bmatrix} 1 & & & & & \\ 1 & 4 & 1 & & & \\ & \ddots & \ddots & \ddots & & \\ & & 1 & 4 & 1 & \\ & & & & 1 & \end{bmatrix}, \qquad \begin{aligned} \mathbf{r} &= [r_0, \ldots, r_{N+1}]^T, \\ \mathbf{p} &= [p_0, \ldots, p_{N+1}]^T, \end{aligned}$$

where

$$p_i = \begin{cases} v''(x_i), & i = 0, N+1, \\ h^{-2}[v(x_{i-1}) - 2v(x_i) + v(x_{i+1})], & i = 1, \ldots, N. \end{cases} \tag{4.23}$$

It follows from the discussion on pages 400 and 401 in [5] that $E\mathbf{r} = \mathbf{p}$. Since $|e_{i,i}| - \sum_{i \ne j} |e_{i,j}| \ge 1$, $i, j = 0, \ldots, N+1$, the discussion on page 21 in [1] implies that

$$\max_{0 \le i \le N+1} |r_i| \le C \max_{0 \le i \le N+1} |p_i|. \tag{4.24}$$

Using Taylor's theorem, we obtain

$$|v(x_{i-1}) - 2v(x_i) + v(x_{i+1})| \le C h^2 \max_{x \in [0,1]} |v''(x)|, \quad i = 1, \ldots, N. \tag{4.25}$$

It follows from (4.24), (4.23), and (4.25) that

$$\max_{0 \le i \le N+1} |r_i| \le C \max_{x \in [0,1]} |v''(x)|. \tag{4.26}$$

Using Taylor's theorem to expand $v(x)$, $x \in [x_i, x_{i+1}]$, around x_i, we have

$$v(x) = v(x_i) + (x - x_i)v'(x_i) + \frac{(x - x_i)^2}{2}v''(\xi_{i,x}), \qquad x_i \le \xi_{i,x} \le x. \tag{4.27}$$

Using (4.21), (4.27), and the triangle inequality, we obtain, for $x \in [x_i, x_{i+1}]$,

$$|A_i(x)| = \left| \frac{(x - x_i)^2}{2}v''(\xi_{i,x}) - \frac{h}{2}(x - x_i)v''(\xi_{i,x_{i+1}}) \right| \le h^2 \max_{x\in[0,1]} |v''(x)|. \tag{4.28}$$

Inequality (4.16) follows from (4.22), (4.26), and (4.28).

A proof of (4.17) is given in the proof of Theorem 2.3.4 in [1]. □

LEMMA 4.2. *If $u \in C^4(\overline{\Omega})$, Z in $S^D \otimes S^D$ is defined by* (4.6)–(4.9), *and $I_x u$ and $I_y u$ are defined in* (4.15), *then for (x, y) in $\overline{\Omega}$, we have*

$$Z(x, y) = I_x(I_y u)(x, y), \qquad D_x^2(I_y u)(x, y) = I_y(D_x^2 u)(x, y). \tag{4.29}$$

Proof. Let $\{C_i\}_{i=0}^{N+3}$ be the basis for S_3 such that

$$\begin{aligned}
&C_i(x_j) = \delta_{ij}, \qquad i, j = 0, \dots, N+1,\\
&C_i''(x_j) = 0, \qquad i = 0, \dots, N+1, \quad j = 0, N+1,\\
&C_{N+2}(x_j) = C_{N+3}(x_j) = 0, \qquad j = 0, \dots, N+1,\\
&C_{N+2}''(x_0) = C_{N+3}''(x_{N+1}) = 1, \qquad C_{N+2}''(x_{N+1}) = C_{N+3}''(x_0) = 0,
\end{aligned} \tag{4.30}$$

where δ_{ij} is the Kronecker delta. Using (4.15) and (4.30), we have for $(x, y) \in \overline{\Omega}$,

$$\begin{aligned}
&I_x(I_y u)(x, y)\\
&= I_x\left[\sum_{j=0}^{N+1} u(x, y_j)C_j(y) + D_y^2 u(x, y_0)C_{N+2}(y) + D_y^2 u(x, y_{N+1})C_{N+3}(y)\right]\\
&= \sum_{i=0}^{N+1}\left[\sum_{j=0}^{N+1} u(x_i, y_j)C_j(y) + D_y^2 u(x_i, y_0)C_{N+2}(y) + D_y^2 u(x_i, y_{N+1})C_{N+3}(y)\right] C_i(x)\\
&+\left[\sum_{j=0}^{N+1} D_x^2 u(x_0, y_j)C_j(y) + D_x^2 D_y^2 u(x_0, y_0)C_{N+2}(y)\right.\\
&\left.+D_x^2 D_y^2 u(x_0, y_{N+1})C_{N+3}(y)\right] C_{N+2}(x) + \left[\sum_{j=0}^{N+1} D_x^2 u(x_{N+1}, y_j)C_j(y)\right.\\
&\left.+D_x^2 D_y^2 u(x_{N+1}, y_0)C_{N+2}(y) + D_x^2 D_y^2 u(x_{N+1}, y_{N+1})C_{N+3}(y)\right] C_{N+3}(x).
\end{aligned}$$

Since $u = 0$ on $\partial\Omega$, all terms involving $C_0(x)$, $C_{N+1}(x)$, $C_0(y)$, $C_{N+1}(y)$ drop out which implies that $I_x(I_y u) \in S^D \otimes S^D$. Using (4.30), we verify that $I_x(I_y u)$ satisfies (4.6)–(4.9), that is, (4.6)–(4.9) hold with $I_x(I_y u)$ in place of Z. Hence, the uniqueness

of the interpolant Z implies the first equation in (4.29). To prove the second equation in (4.29), we use (4.15) and (4.30) to see that for $(x, y) \in \overline{\Omega}$,

$$
\begin{aligned}
&I_y(D_x^2 u)(x,y)\\
&= \sum_{j=0}^{N+1} D_x^2 u(x,y_j)C_j(y) + D_x^2 D_y^2 u(x,y_0)C_{N+2}(y) + D_x^2 D_y^2 u(x,y_{N+1})C_{N+3}(y)\\
&= D_x^2\left[\sum_{j=0}^{N+1} u(x,y_j)C_j(y) + D_y^2 u(x,y_0)C_{N+2}(y) + D_y^2 u(x,y_{N+1})C_{N+3}(y)\right]\\
&= D_x^2(I_y u)(x,y). \qquad \square
\end{aligned}
$$

THEOREM 4.1. *If $u \in C^4(\overline{\Omega})$ and Z in $S^D \otimes S^D$ is defined by* (4.6)–(4.9)*, then* $\|u - Z\|_{C(\overline{\Omega})} \le C(u)h^4$.

Proof. Using (4.29) and the triangle inequality, we have

$$
\begin{aligned}
\|u - Z\|_{C(\overline{\Omega})} \le\ & \|u - I_x u\|_{C(\overline{\Omega})} + \|I_x(u - I_y u) - (u - I_y u)\|_{C(\overline{\Omega})}\\
&+ \|u - I_y u\|_{C(\overline{\Omega})}.
\end{aligned} \tag{4.31}
$$

For any fixed y in $[0,1]$, $I_x u(\cdot, y)$ is the cubic spline interpolant of $u(\cdot, y)$. Using this, symmetry with respect to x and y, and (4.17), we have

$$
\|u - I_x u\|_{C(\overline{\Omega})} \le C(u)h^4, \qquad \|u - I_y u\|_{C(\overline{\Omega})} \le C(u)h^4. \tag{4.32}
$$

For any fixed y in $[0,1]$, $I_x(u - I_y u)(\cdot, y)$ is the cubic spline interpolant of $(u - I_y u)(\cdot, y)$. Hence it follows from (4.16) that

$$
\|I_x(u - I_y u) - (u - I_y u)\|_{C(\overline{\Omega})} \le C\|D_x^2(u - I_y u)\|_{C(\overline{\Omega})} h^2. \tag{4.33}
$$

Using (4.29) and(4.16), we obtain

$$
\|D_x^2(u - I_y u)\|_{C(\overline{\Omega})} = \|D_x^2 u - I_y(D_x^2 u)\|_{C(\overline{\Omega})} \le C\|D_x^2 D_y^2 u\|_{C(\overline{\Omega})} h^2. \tag{4.34}
$$

Combining (4.33) and (4.34), we have

$$
\|I_x(u - I_y u) - (u - I_y u)\|_{C(\overline{\Omega})} \le C(u)h^4. \tag{4.35}
$$

The desired inequality now follows from (4.31), (4.32), and (4.35). $\square$

4.2. Bounding $\|S - Z\|_{C(\overline{\Omega})}$. We prove the following result.

THEOREM 4.2. *If $u \in C^4(\overline{\Omega})$ and S, Z in $S^D \otimes S^D$ are defined by* (4.2)–(4.5) *and* (4.6)–(4.9)*, respectively, then* $\|S - Z\|_{C(\overline{\Omega})} \le C(u)h^4$.

Proof. Clearly,

$$
S(x,y) = \sum_{m=0}^{N+1}\sum_{n=0}^{N+1} s_{m,n} B_m^D(x) B_n^D(y), \quad Z(x,y) = \sum_{m=0}^{N+1}\sum_{n=0}^{N+1} z_{m,n} B_m^D(x) B_n^D(y). \tag{4.36}
$$

Using (4.2), (4.6), and following the derivation of (3.3) from (1.3), we obtain

$$
s_{m,n} = z_{m,n}, \qquad m, n = 0, N+1. \tag{4.37}
$$

Using (4.7), we have

$$D_x^2(S-Z)(x_0,y_j) = D_x^2 S(x_0,y_j) - D_x^2 u(x_0,y_j), \quad j=1,\dots,N. \tag{4.38}$$

It follows from (4.36), (4.37), and (2.6) that

$$D_x^2(S-Z)(x_0,y_j) = -36h^{-2}\sum_{n=1}^{N}(s_{0,n}-z_{0,n})B_n^D(y_j), \quad j=1,\dots,N. \tag{4.39}$$

Using (4.36), (2.6), (2.4), (2.3), and (2.5), we obtain, for $j=1,\dots,N$, that

$$D_x^2 S(x_0,y_j) = -36h^{-2}\sum_{n=0}^{N+1} s_{0,n}B_n^D(y_j) = -36h^{-2}(s_{0,j-1}+4s_{0,j}+s_{0,j+1}). \tag{4.40}$$

Substituting (4.39) and (4.40) into (4.38), and multiplying through by $-h^2/36$, we have

$$\sum_{n=1}^{N}(s_{0,n}-z_{0,n})B_n^D(y_j) = s_{0,j-1}+4s_{0,j}+s_{0,j+1}+\frac{h^2}{36}D_x^2u(x_0,y_j) \tag{4.41}$$

for $j=1,\dots,N$. Using (4.2), (4.3), and following the derivations of (3.3) from (1.3) and (3.5) from (1.4), we obtain

$$s_{0,j} = \frac{h^4}{1296}D_x^2D_y^2u(x_0,y_j), \quad j=0,N+1, \tag{4.42}$$

and

$$s_{0,j} = -\frac{h^2}{216}\left[D_x^2u(x_0,y_j) - \frac{h^2}{12}D_x^4u(x_0,y_j) - \frac{h^2}{6}D_x^2D_y^2u(x_0,y_j)\right] \tag{4.43}$$

for $j=1,\dots,N$. Since $u=0$ on $\partial\Omega$, (4.42) is the same as (4.43) with $j=0,N+1$. This observation and (4.43) imply that for $j=1,\dots,N$, we have

$$s_{0,j\pm1} = -\frac{h^2}{216}\left[D_x^2u(x_0,y_{j\pm1}) - \frac{h^2}{12}D_x^4u(x_0,y_{j\pm1}) - \frac{h^2}{6}D_x^2D_y^2u(x_0,y_{j\pm1})\right]. \tag{4.44}$$

Taylor's theorem gives

$$D_x^2u(x_0,y_{j\pm1}) = D_x^2u(x_0,y_j) \pm hD_x^2D_yu(x_0,y_j) + \frac{h^2}{2}D_x^2D_y^2u(x_0,\xi_j^{\pm}), \tag{4.45}$$

where $y_{j-1}\le\xi_j^-\le y_j$, $y_j\le\xi_j^+\le y_{j+1}$. Using (4.44), (4.43), and (4.45), we obtain

$$\left|s_{0,j-1}+4s_{0,j}+s_{0,j+1}+\frac{h^2}{36}D_x^2u(x_0,y_j)\right| \le C(u)h^4, \quad j=1,\dots,N. \tag{4.46}$$

It follows from (4.46) that (4.41) is a system in $\{s_{0,n}-z_{0,n}\}_{n=1}^N$ with the matrix B defined in(2.10) and with each entry on the right-hand side bounded in absolute value by $C(u)h^4$. Hence, Lemma 2.2 implies

$$\max_{1\le n\le N}|s_{0,n}-z_{0,n}| \le C(u)h^4. \tag{4.47}$$

Using (4.47) and symmetry with respect to x and y, we also have

$$\begin{aligned} &\max_{1\le n\le N}|s_{N+1,n}-z_{N+1,n}|\le C(u)h^4,\\ &\max_{1\le m\le N}|s_{m,n}-z_{m,n}|\le C(u)h^4, \qquad n=0,N+1. \end{aligned} \tag{4.48}$$

Using (4.5) and (4.9), we have $(S-Z)(x_i,y_j)=0$, $i,j=1,\dots,N$, which, by (4.36) and (4.37), can be written as

$$\sum_{m=1}^{N}\sum_{n=1}^{N}(s_{m,n}-z_{m,n})B_m^D(x_i)B_n^D(y_j)=d_{i,j}, \qquad i,j=1,\dots N, \tag{4.49}$$

where

$$d_{i,j}=\left(\sum_{m=0,N+1}\sum_{n=1}^{N}+\sum_{m=1}^{N}\sum_{n=0,N+1}\right)(z_{m,n}-s_{m,n})B_m^D(x_i)B_n^D(y_j).$$

Since for any fixed i,j, each of the above double sums reduces to at most three terms, using the triangle inequality, (4.47), (4.48), and Lemma 2.1, we obtain

$$|d_{i,j}|\le C(u)h^4, \qquad i,j=1,\dots,N. \tag{4.50}$$

It follows from (2.18)–(2.19) that (4.49) is a system in $\{s_{m,n}-z_{m,n}\}_{m,n=1}^{N}$ with the matrix $B\otimes B$, where B is defined in (2.10). Hence, (4.50), Lemma 2.3, (4.47), and (4.48) give

$$|s_{m,n}-z_{m,n}|\le C(u)h^4, \qquad m,n=0,\dots,N+1. \tag{4.51}$$

Since $S-Z$ is continuous on $\overline{\Omega}$, there is (x_*,y_*) in $\overline{\Omega}$ such that

$$\|S-Z\|_{C(\overline{\Omega})}=|(S-Z)(x_*,y_*)|.$$

Hence, (4.36) and the triangle inequality imply

$$\|S-Z\|_{C(\overline{\Omega})}\le\sum_{m=0}^{N+1}\sum_{n=0}^{N+1}|s_{m,n}-z_{m,n}||B_m^D(x_*)||B_n^D(y_*)|.$$

Since the above double sum reduces to at most nine terms, the required inequality follows from (4.51) and Lemma 2.1. ∎

4.3. Bounding $\|S-u_h\|_{C(\overline{\Omega})}$ and $\|u-u_h\|_{C(\overline{\Omega})}$. To bound $\|S-u_h\|_{C(\overline{\Omega})}$, we need two additional lemmas.

LEMMA 4.3. *If $u\in C^6(\overline{\Omega})$ and S in $S^D\otimes S^D$ is defined by (4.2)–(4.5), then for $i=0,N+1$, $j=1,\dots,N$,*

$$\left|D_x^2D_y^2S(x_i,y_j)-D_x^2D_y^2u(x_i,y_j)\right|\le C(u)h^2, \tag{4.52}$$

$$\left|D_x^2S(x_i,y_j)-D_x^2u(x_i,y_j)+\frac{h^2}{12}D_x^4u(x_i,y_j)\right|\le C(u)h^4. \tag{4.53}$$

Proof. We prove (4.52) for $i = 0$; for $i = N + 1$, (4.52) follows by symmetry with respect to x. Using (4.36), we obtain

$$D_x^2 D_y^2 S(x_0, y_j) = \sum_{m=0}^{N+1} \sum_{n=0}^{N+1} s_{m,n} \left[B_m^D\right]''(x_0) \left[B_n^D\right]''(y_j), \qquad j = 1, \ldots, N,$$

and hence (2.4), (2.3), and (2.6) imply

$$D_x^2 D_y^2 S(x_0, y_j) = -216h^{-4}(s_{0,j-1} - 2s_{0,j} + s_{0,j+1}), \quad j = 1, \ldots, N. \tag{4.54}$$

Equations (4.54), (4.43), and (4.44) give, for $j = 1, \ldots, N$,

$$\begin{aligned} &D_x^2 D_y^2 S(x_0, y_j) - D_x^2 D_y^2 u(x_0, y_j) = -D_x^2 D_y^2 u(x_0, y_j) \\ &+h^{-2}\left[D_x^2 u(x_0, y_{j-1}) - 2D_x^2 u(x_0, y_j) + D_x^2 u(x_0, y_{j+1})\right] \\ &-\frac{1}{12}\left[D_x^4 u(x_0, y_{j-1}) - 2D_x^4 u(x_0, y_j) + D_x^4 u(x_0, y_{j+1})\right] \\ &-\frac{1}{6}\left[D_x^2 D_y^2 u(x_0, y_{j-1}) - 2D_x^2 D_y^2 u(x_0, y_j) + D_x^2 D_y^2 u(x_0, y_{j+1})\right]. \end{aligned} \tag{4.55}$$

Using Taylor's theorem, we obtain

$$\begin{aligned} D_x^2 u(x_0, y_{j\pm1}) = D_x^2 u(x_0, y_j) \pm hD_x^2 D_y u(x_0, y_j) + \frac{h^2}{2} D_x^2 D_y^2 u(x_0, y_j) \\ \pm \frac{h^3}{3!} D_x^2 D_y^3 u(x_0, y_j) + \frac{h^4}{4!} D_x^2 D_y^4 u(x_0, \xi_j^{\pm}), \end{aligned} \tag{4.56}$$

$$D_x^4 u(x_0, y_{j\pm1}) = D_x^4 u(x_0, y_j) \pm hD_x^4 D_y u(x_0, y_j) + \frac{h^2}{2} D_x^4 D_y^2 u(x_0, \eta_j^{\pm}), \tag{4.57}$$

$$D_x^2 D_y^2 u(x_0, y_{j\pm1}) = D_x^2 D_y^2 u(x_0, y_j) \pm hD_x^2 D_y^3 u(x_0, y_j) + \frac{h^2}{2} D_x^2 D_y^4 u(x_0, \kappa_j^{\pm}), \tag{4.58}$$

where $y_{j-1} \le \xi_j^-, \eta_j^-, \kappa_j^- \le y_j$, $y_j \le \xi_j^+, \eta_j^+, \kappa_j^+ \le y_{j+1}$. Equations (4.56)–(4.58) give

$$\begin{aligned} \left|h^{-2}\left[D_x^2 u(x_0, y_{j-1}) - 2D_x^2 u(x_0, y_j) + D_x^2 u(x_0, y_{j+1})\right] - D_x^2 D_y^2 u(x_0, y_j)\right| \le C(u)h^2, \\ \left|D_x^4 u(x_0, y_{j-1}) - 2D_x^4 u(x_0, y_j) + D_x^4 u(x_0, y_{j+1})\right| \le C(u)h^2, \\ \left|D_x^2 D_y^2 u(x_0, y_{j-1}) - 2D_x^2 D_y^2 u(x_0, y_j) + D_x^2 D_y^2 u(x_0, y_{j+1})\right| \le C(u)h^2, \end{aligned}$$

and hence (4.52) for $i = 0$ follows from (4.55) and the triangle inequality. Using (4.3) and (4.52), we obtain (4.53). ∎

LEMMA 4.4. *If $u \in C^6(\overline{\Omega})$ and S in $S^D \otimes S^D$ is defined by* (4.2)–(4.5)*, then, for $i, j = 1, \ldots, N$, we have*

$$\left|D_x^2 S(x_i, y_j) - D_x^2 u(x_i, y_j) + \frac{h^2}{12} D_x^4 u(x_i, y_j)\right| \le C(u)h^4, \tag{4.59}$$

$$\left|D_y^2 S(x_i, y_j) - D_y^2 u(x_i, y_j) + \frac{h^2}{12} D_y^4 u(x_i, y_j)\right| \le C(u)h^4, \tag{4.60}$$

$$|D_x^2 D_y^2 S(x_i, y_j) - D_x^2 D_y^2 u(x_i, y_j)| \le C(u)h^2. \tag{4.61}$$

Proof. First, we prove (4.59). For $i = 0, \ldots, N+1$, $j = 1, \ldots, N$, we introduce

$$d_{i,j} = D_x^2 S(x_i, y_j) - \left[D_x^2 u(x_i, y_j) - \frac{h^2}{12} D_x^4 u(x_i, y_j) \right]. \tag{4.62}$$

Then

$$d_{i-1,j} + 4d_{i,j} + d_{i+1,j} = \phi_{i,j} - \psi_{i,j}, \qquad i, j = 1, \ldots, N, \tag{4.63}$$

where

$$\begin{aligned} \phi_{i,j} &= D_x^2 S(x_{i-1}, y_j) + 4D_x^2 S(x_i, y_j) + D_x^2 S(x_{i+1}, y_j) \\ &\quad - 6 \left[D_x^2 u(x_i, y_j) + \frac{h^2}{12} D_x^4 u(x_i, y_j) \right], \end{aligned} \tag{4.64}$$

$$\begin{aligned} \psi_{i,j} &= D_x^2 u(x_{i-1}, y_j) - \frac{h^2}{12} D_x^4 u(x_{i-1}, y_j) + 4 \left[D_x^2 u(x_i, y_j) - \frac{h^2}{12} D_x^4 u(x_i, y_j) \right] \\ &\quad + D_x^2 u(x_{i+1}, y_j) - \frac{h^2}{12} D_x^4 u(x_{i+1}, y_j) - 6 \left[D_x^2 u(x_i, y_j) + \frac{h^2}{12} D_x^4 u(x_i, y_j) \right] \\ &= D_x^2 u(x_{i-1}, y_j) - 2D_x^2 u(x_i, y_j) + D_x^2 u(x_{i+1}, y_j) \\ &\quad - \frac{h^2}{12} \left[D_x^4 u(x_{i-1}, y_j) + 10 D_x^4 u(x_i, y_j) + D_x^4 u(x_{i+1}, y_j) \right]. \end{aligned} \tag{4.65}$$

Since $S(\cdot, y_j) \in S_3$, (2.1.7) in [1], (4.5), and $S = u = 0$ on $\partial\Omega$, imply that

$$\begin{aligned} &D_x^2 S(x_{i-1}, y_j) + 4D_x^2 S(x_i, y_j) + D_x^2 S(x_{i+1}, y_j) \\ &= 6h^{-2} \left[u(x_{i-1}, y_j) - 2u(x_i, y_j) + u(x_{i+1}, y_j) \right], \quad i, j = 1, \ldots, N. \end{aligned} \tag{4.66}$$

Using Taylor's theorem, we obtain

$$\begin{aligned} u(x_{i\pm1}, y_j) &= u(x_i, y_j) \pm h D_x u(x_i, y_j) + \frac{h^2}{2} D_x^2 u(x_i, y_j) \pm \frac{h^3}{3!} D_x^3 u(x_i, y_j) \\ &\quad + \frac{h^4}{4!} D_x^4 u(x_i, y_j) \pm \frac{h^5}{5!} D_x^5 u(x_i, y_j) + \frac{h^6}{6!} D_x^6 u(\xi_i^{\pm}, y_j), \end{aligned}$$

where $x_{i-1} \le \xi_i^- \le x_i$, $x_i \le \xi_i^+ \le x_{i+1}$, and hence

$$\begin{aligned} &\Big| h^{-2} \left[u(x_{i-1}, y_j) - 2u(x_i, y_j) + u(x_{i+1}, y_j) \right] \\ &\quad - \left[D_x^2 u(x_i, y_j) + \frac{h^2}{12} D_x^4 u(x_i, y_j) \right] \Big| \le C(u)h^4, \quad i, j = 1, \ldots, N. \end{aligned} \tag{4.67}$$

Using (4.64), (4.66), and (4.67), we obtain

$$|\phi_{i,j}| \le C(u)h^4, \quad i, j = 1, \ldots, N. \tag{4.68}$$

Using Taylor's theorem, we obtain

$$\begin{aligned} D_x^2 u(x_{i\pm1}, y_j) &= D_x^2 u(x_i, y_j) \pm h D_x^3 u(x_i, y_j) + \frac{h^2}{2} D_x^4 u(x_i, y_j) \\ &\quad \pm \frac{h^3}{3!} D_x^5 u(x_i, y_j) + \frac{h^4}{4!} D_x^6 u(\xi_i^{\pm}, y_j), \\ D_x^4 u(x_{i\pm1}, y_j) &= D_x^4 u(x_i, y_j) \pm h D_x^5 u(x_i, y_j) + \frac{h^2}{2} D_x^6 u(\eta_i^{\pm}, y_j), \end{aligned}$$

where $x_{i-1} \le \xi_i^-, \eta_i^- \le x_i$, $x_i \le \xi_i^+, \eta_i^+ \le x_{i+1}$, and hence (4.65) gives

$$|\psi_{i,j}| \le C(u)h^4, \quad i,j = 1,\dots,N. \tag{4.69}$$

Using (4.62) and (4.53), we have

$$|d_{i,j}| \le C(u)h^4, \quad i = 0, N+1, \quad j = 1,\dots,N. \tag{4.70}$$

It follows from (4.68)–(4.70) that moving $d_{0,j}$ and $d_{N+1,j}$ to the right-hand side of (4.63), we obtain, for each $j = 1,\dots,N$, a system in $\{d_{i,j}\}_{i=1}^N$ with the matrix B of (2.10)–(2.12), and with each entry on the right-hand side bounded in absolute value by $C(u)h^4$. Hence (4.59) follows from (4.62) and Lemma 2.2, and (4.60) follows from (4.59) by symmetry with respect to x and y.

Next we prove (4.61). Since $S(x,\cdot) \in S_3$ for $x \in [0,1]$, (2.1.7) in [1] gives

$$\begin{aligned} &D_y^2 S(x,y_{j-1}) + 4D_y^2 S(x,y_j) + D_y^2 S(x,y_{j+1}) \\ &= 6h^{-2}\left[S(x,y_{j-1}) - 2S(x,y_j) + S(x,y_{j+1})\right], \quad j = 1,\dots,N, \quad x \in [0,1]. \end{aligned} \tag{4.71}$$

Differentiating (4.71) twice with respect to x, we obtain, for $j = 1,\dots,N$, $x \in [0,1]$,

$$\begin{aligned} &D_x^2 D_y^2 S(x,y_{j-1}) + 4D_x^2 D_y^2 S(x,y_j) + D_x^2 D_y^2 S(x,y_{j+1}) \\ &= 6h^{-2}\left[D_x^2 S(x,y_{j-1}) - 2D_x^2 S(x,y_j) + D_x^2 S(x,y_{j+1})\right]. \end{aligned} \tag{4.72}$$

Using (4.72) with $x = x_{i-1}, x_i, x_{i+1}$, we obtain, for $i,j = 1,\dots,N$,

$$\begin{aligned} &D_x^2 D_y^2 S(x_{i-1},y_{j-1}) + 4D_x^2 D_y^2 S(x_{i-1},y_j) + D_x^2 D_y^2 S(x_{i-1},y_{j+1}) \\ &= 6h^{-2}\left[D_x^2 S(x_{i-1},y_{j-1}) - 2D_x^2 S(x_{i-1},y_j) + D_x^2 S(x_{i-1},y_{j+1})\right], \end{aligned} \tag{4.73}$$

$$\begin{aligned} &D_x^2 D_y^2 S(x_i,y_{j-1}) + 4D_x^2 D_y^2 S(x_i,y_j) + D_x^2 D_y^2 S(x_i,y_{j+1}) \\ &= 6h^{-2}\left[D_x^2 S(x_i,y_{j-1}) - 2D_x^2 S(x_i,y_j) + D_x^2 S(x_i,y_{j+1})\right], \end{aligned} \tag{4.74}$$

$$\begin{aligned} &D_x^2 D_y^2 S(x_{i+1},y_{j-1}) + 4D_x^2 D_y^2 S(x_{i+1},y_j) + D_x^2 D_y^2 S(x_{i+1},y_{j+1}) \\ &= 6h^{-2}\left[D_x^2 S(x_{i+1},y_{j-1}) - 2D_x^2 S(x_{i+1},y_j) + D_x^2 S(x_{i+1},y_{j+1})\right]. \end{aligned} \tag{4.75}$$

Adding (4.73), (4.75), and (4.74) multiplied through by 4, and using (4.66) and $S = u = 0$ on $\partial\Omega$, we obtain

$$\begin{aligned} &D_x^2 D_y^2 S(x_{i-1},y_{j-1}) + 4D_x^2 D_y^2 S(x_{i-1},y_j) + D_x^2 D_y^2 S(x_{i-1},y_{j+1}) \\ &+4D_x^2 D_y^2 S(x_i,y_{j-1}) + 16D_x^2 D_y^2 S(x_i,y_j) + 4D_x^2 D_y^2 S(x_i,y_{j+1}) \\ &+D_x^2 D_y^2 S(x_{i+1},y_{j-1}) + 4D_x^2 D_y^2 S(x_{i+1},y_j) + D_x^2 D_y^2 S(x_{i+1},y_{j+1}) \\ &= 36h^{-4}\alpha_{i,j}, \quad i,j = 1,\dots,N, \end{aligned} \tag{4.76}$$

where

$$\begin{aligned} \alpha_{i,j} = {} & u(x_{i-1},y_{j-1}) - 2u(x_i,y_{j-1}) + u(x_{i+1},y_{j-1}) \\ & -2u(x_{i-1},y_j) + 4u(x_i,y_j) - 2u(x_{i+1},y_j) \\ & +u(x_{i-1},y_{j+1}) - 2u(x_i,y_{j+1}) + u(x_{i+1},y_{j+1}). \end{aligned} \tag{4.77}$$

Using (4.77) and the discussion on pages 290–292 in [9], we have

$$\left|h^{-4}\alpha_{i,j} - D_x^2 D_y^2 u(x_i,y_j)\right| \le C(u)h^2, \quad i,j = 1,\dots,N. \tag{4.78}$$

Equation (4.76) is equivalent to

$$\begin{aligned}
&D_x^2D_y^2(S-u)(x_{i-1},y_{j-1})+4D_x^2D_y^2(S-u)(x_{i-1},y_j)\\
&+D_x^2D_y^2(S-u)(x_{i-1},y_{j+1})+4D_x^2D_y^2(S-u)(x_i,y_{j-1})\\
&+16D_x^2D_y^2(S-u)(x_i,y_j)+4D_x^2D_y^2(S-u)(x_i,y_{j+1})\\
&+D_x^2D_y^2(S-u)(x_{i+1},y_{j-1})+4D_x^2D_y^2(S-u)(x_{i+1},y_j)\\
&+D_x^2D_y^2(S-u)(x_{i+1},y_{j+1})=36h^{-4}\alpha_{i,j}-\beta_{i,j},\qquad i,j=1,\dots,N,
\end{aligned}\tag{4.79}$$

where

$$\begin{aligned}
\beta_{i,j}=\ &D_x^2D_y^2u(x_{i-1},y_{j-1})+4D_x^2D_y^2u(x_{i-1},y_j)+D_x^2D_y^2u(x_{i-1},y_{j+1})\\
&+4D_x^2D_y^2u(x_i,y_{j-1})+16D_x^2D_y^2u(x_i,y_j)+4D_x^2D_y^2u(x_i,y_{j+1})\\
&+D_x^2D_y^2u(x_{i+1},y_{j-1})+4D_x^2D_y^2u(x_{i+1},y_j)+D_x^2D_y^2u(x_{i+1},y_{j+1}).
\end{aligned}\tag{4.80}$$

Using Taylor's theorem, we obtain

$$\begin{aligned}
D_x^2D_y^2u(x_{i-1},y_{j\pm1})&=D_x^2D_y^2u(x_i,y_j)-hD_x^3D_y^2u(x_i,y_j)\pm hD_x^2D_y^3D_yu(x_i,y_j)+\epsilon_{i,j}^{\pm},\\
D_x^2D_y^2u(x_{i+1},y_{j\pm1})&=D_x^2D_y^2u(x_i,y_j)+hD_x^3D_y^2u(x_i,y_j)\pm hD_x^2D_y^3D_yu(x_i,y_j)+\sigma_{i,j}^{\pm},\\
D_x^2D_y^2u(x_i,y_{j\pm1})&=D_x^2D_y^2u(x_i,y_j)\pm hD_x^2D_y^3u(x_i,y_j)+\mu_{i,j}^{\pm},\\
D_x^2D_y^2u(x_{i\pm1},y_j)&=D_x^2D_y^2u(x_i,y_j)\pm hD_x^3D_y^2u(x_i,y_j)+\nu_{i,j}^{\pm},
\end{aligned}$$

where $|\epsilon_{i,j}^{\pm}|$, $|\sigma_{i,j}^{\pm}|$, $|\mu_{i,j}^{\pm}|$, $|\nu_{i,j}^{\pm}|\le C(u)h^2$, $i,j=1,\dots,N$, and hence (4.80) gives

$$\left|\beta_{i,j}-36D_x^2D_y^2u(x_i,y_j)\right|\le C(u)h^2,\qquad i,j=1,\dots,N.\tag{4.81}$$

It follows from (4.78) and (4.81) that the right-hand side of (4.79) is bounded in absolute value by $C(u)h^2$. Using (4.2) and moving terms involving

$$D_x^2D_y^2(S-u)(x_i,y_j),\quad i=0,N+1,j=1,\dots,N,\quad i=1,\dots,N,j=0,N+1,$$

to the right-hand side of (4.79), we obtain a system in $\{D_x^2D_y^2(S-u)(x_i,y_j)\}_{i,j=1}^N$ with the matrix $B\otimes B$, where B is given in (2.10)–(2.12). By (4.52) and symmetry with respect to x and y, each entry on the right-hand side in this system is bounded in absolute value by $C(u)h^2$. Therefore, (4.61) follows from Lemma 2.3. ∎

THEOREM 4.3. *If $u\in C^6(\overline{\Omega})$ and u_h and S in $S^D\otimes S^D$ are defined by* (1.3)–(1.6) *and* (4.2)–(4.5)*, respectively, then $\|S-u_h\|_{C(\overline{\Omega})}\le C(u)h^4$.*

Proof. Clearly, u_h and S are given by the formulas in (3.1) and (4.36), respectively. Using (4.2)–(4.4), (4.11)–(4.13), and following the derivations of (3.3) from (1.3), (3.5) from (1.4), and (3.6) from (1.5), we conclude that

$$s_{m,n}=u_{m,n},\quad m=0,N+1,n=0,\dots,N+1,\quad m=1,\dots,N,n=0,N+1.\tag{4.82}$$

We define $\{w_{i,j}\}_{i,j=1}^N$ by

$$\Delta(S-u_h)(x_i,y_j)-\frac{h^2}{6}D_x^2D_y^2(S-u_h)(x_i,y_j)=w_{i,j},\quad i,j=1,\dots,N.\tag{4.83}$$

Using (4.83), (1.6), and (4.10), we obtain

$$\begin{aligned}
w_{i,j}=\ &D_x^2S(x_i,y_j)+D_y^2S(x_i,y_j)-\frac{h^2}{6}D_x^2D_y^2S(x_i,y_j)\\
&-D_x^2u(x_i,y_j)-D_y^2u(x_i,y_j)+\frac{h^2}{12}\left[D_x^4u(x_i,y_j)+D_y^4u(x_i,y_j)+2D_x^2D_y^2u(x_i,y_j)\right],
\end{aligned}$$

and hence (4.59)–(4.61) and the triangle inequality imply that

$$|w_{i,j}| \le C(u)h^4, \quad i,j = 1,\dots,N. \tag{4.84}$$

Introducing $\mathbf{v} = [s_{1,1} - u_{1,1},\dots,s_{N,N} - u_{N,N}]^T$, $\mathbf{w} = [w_{1,1},\dots,w_{N,N}]^T$, using (4.83), (3.1), (4.36), (4.82), and following the derivation of (3.10) from (1.6), we obtain

$$6h^{-2}\left[6T \otimes I + (6I + T) \otimes T\right]\mathbf{v} = \mathbf{w}. \tag{4.85}$$

Since $h = 1/(N+1)$, (4.84) gives $h^2\sum_{i,j=1}^{N} w_{i,j}^2 \le C^2(u)h^8$, and hence (4.85), (3.12), and Lemma 2.4 imply that

$$\max_{1\le m,n\le N} |s_{m,n} - u_{m,n}| \le C(u)h^4, \quad m,n = 1,\dots,N. \tag{4.86}$$

Since $S - u_h$ is continuous on $\overline{\Omega}$, there is (x_*, y_*) in $\overline{\Omega}$ such that

$$\|S - u_h\|_{C(\overline{\Omega})} = |(S - u_h)(x_*, y_*)|.$$

Hence, (3.1), (4.36), (4.82), and the triangle inequality give

$$\|S - u_h\|_{C(\overline{\Omega})} \le \sum_{m=1}^{N}\sum_{n=1}^{N} |s_{m,n} - u_{m,n}||B_m^D(x_*)||B_n^D(y_*)|.$$

Since the above double sum reduces to at most nine terms, the desired result follows from (4.86) and Lemma 2.1. □

THEOREM 4.4. *If u in $C^6(\overline{\Omega})$ and u_h in $S^D \otimes S^D$ are the solutions of* (1.1) *and* (1.3)–(1.6)*, respectively, then* $\|u - u_h\|_{C(\overline{\Omega})} \le C(u)h^4$.

Proof. The required inequality follows from (4.14) and Theorems 4.1, 4.2, and 4.3. □

5. Other schemes. Consider the scheme for solving (1.1) formulated as follows: Find $u_h \in S^D \otimes S^D$ satisfying (1.3)–(1.5) and (1.9). This scheme is essentially the same as the scheme (4.1)–(4.3) in [4], except that (1.5) is replaced in [4] with

$$\frac{1}{12}[13D_y^2u_h(x_i,y_j) - 2D_y^2u_h(x_i,y_{j+1}) + D_y^2u_h(x_i,y_{j+2})] = f(x_i,y_j), \quad j = 0,$$

$$\frac{1}{12}[D_y^2u_h(x_i,y_{j-2}) - 2D_y^2u_h(x_i,y_{j-1}) + 13D_y^2u_h(x_i,y_j)] = f(x_i,y_j), \quad j = N+1,$$

where $i = 1,\dots,N$. It follows from (3.1), the discussion in section 3, and (2.9) of [4] that the matrix-vector form of (1.3)–(1.5) and (1.9) is

$$(A \otimes B + B \otimes A)\mathbf{u} = \mathbf{p}, \tag{5.1}$$

where $\mathbf{u} = [u_{1,1},\dots,u_{N,N}]^T$, $\mathbf{p} = [p_{1,1},\dots,p_{N,N}]^T$,

$$p_{i,j} = f(x_i,y_j) - \sum_{m=0,N+1}\sum_{n=0}^{N+1} u_{m,n}\left[L_xB_m^D(x_i)B_n^D(y_j) + B_m^D(x_i)L_yB_m^D(y_j)\right]$$
$$-\sum_{m=1}^{N}\sum_{n=0,N+1} u_{m,n}\left[L_xB_m^D(x_i)B_n^D(y_j) + B_m^D(x_i)L_yB_n^D(y_j)\right],$$

$\{u_{i,j}\}_{j=0}^{N+1}$, $i = 0, N+1$, $\{u_{i,j}\}_{i=1}^{N}$, $j = 0, N+1$, are given in (3.3), (3.5), (3.6),

$$A = \frac{1}{2h^2}(T^2 + 12T), \quad B = T + 6I, \tag{5.2}$$

and T is defined in (2.12).

LEMMA 5.1. *Assume A, B are as in* (5.2) *and* $\mathbf{v} = [v_{1,1}, \ldots, v_{N,N}]^T$, $\mathbf{w} = [w_{1,1}, \ldots, w_{N,N}]^T$ *are such that* $(A \otimes B + B \otimes A)\mathbf{v} = \mathbf{w}$. *Then*

$$\max_{1 \le i,j \le N} v_{i,j}^2 \le Ch^2 \sum_{i=1}^{N} \sum_{j=1}^{N} w_{i,j}^2.$$

Proof. It follows from (5.2) that

$$\begin{aligned} A \otimes B + B \otimes A &= \frac{1}{2h^2}(T^2 \otimes T) + \frac{3}{h^2}(T^2 \otimes I) + \frac{6}{h^2}(T \otimes T) + \frac{36}{h^2}(T \otimes I) \\ &+ \frac{1}{2h^2}(T \otimes T^2) + \frac{3}{h^2}(I \otimes T^2) + \frac{6}{h^2}(T \otimes T) + \frac{36}{h^2}(I \otimes T) = 36[r(T) + s(T)], \end{aligned} \tag{5.3}$$

where for an $N \times N$ matrix P,

$$r(P) = h^{-2}(P \otimes I + I \otimes P), \tag{5.4}$$

$$s(P) = \frac{1}{3h^2}(P \otimes P) + \frac{1}{72h^2}\left(P^2 \otimes P + P \otimes P^2\right) + \frac{1}{12h^2}\left(P^2 \otimes I + I \otimes P^2\right). \tag{5.5}$$

First, we will show that

$$([r(T) + s(T)]\mathbf{z}, \mathbf{z}) \le \frac{2}{9}(r(T)\mathbf{z}, \mathbf{z}), \qquad \mathbf{z} \in R^{N^2}, \tag{5.6}$$

where $(\cdot, \cdot)$ is the standard inner product in R^{N^2}. It follows from (2.14) and $Q^T = Q$ for Q of (2.16) that

$$(r(T)\mathbf{z}, \mathbf{z}) = ([Q \otimes Q]r(\Lambda)[Q \otimes Q]\mathbf{z}, \mathbf{z}) = (r(\Lambda)[Q \otimes Q]\mathbf{z}, [Q \otimes Q]\mathbf{z}),$$

$$(s(T)\mathbf{z}, \mathbf{z}) = ([Q \otimes Q]s(\Lambda)[Q \otimes Q]\mathbf{z}, \mathbf{z}) = (s(\Lambda)[Q \otimes Q]\mathbf{z}, [Q \otimes Q]\mathbf{z}),$$

where Λ is given in (2.15). Hence (5.6) is equivalent to

$$([r(\Lambda) + s(\Lambda)]\mathbf{z}, \mathbf{z}) \le \frac{2}{9}(r(\Lambda)\mathbf{z}, \mathbf{z}), \qquad \mathbf{z} \in R^{N^2},$$

which, by (5.4), (5.5), and (2.15), is in turn equivalent to

$$g(\lambda_i, \lambda_j) \le 0, \qquad i, j = 1, \ldots, N, \tag{5.7}$$

where

$$g(x, y) = \frac{7}{9}(x + y) + \frac{1}{3}xy + \frac{1}{72}(x^2 y + xy^2) + \frac{1}{12}(x^2 + y^2).$$

It follows from (2.15) that $-4 \le \lambda_i \le 0$, $i = 1, \ldots, N$. Hence, (5.7) follows from

$$g(x, y) \le 0, \qquad x, y \in [-4, 0],$$

which is established using elementary calculus.

The matrices $r(T)$ and $s(T)$ are symmetric, $r(T)s(T) = s(T)r(T)$, and $-r(T)$ is positive definite. Hence, (5.6) and the considerations 6) on page 135 in [9] imply that

$$\|r(T)\mathbf{z}\|_2 \le \frac{9}{2}\|[r(T)+s(T)]\mathbf{z}\|_2, \qquad \mathbf{z} \in R^{N^2}, \tag{5.8}$$

where $\|\cdot\|_2$ is the two vector norm. It is known (see, for example, the embedding theorem on page 281 in [9]) that

$$\max_{1\le i,j\le N} z_{i,j}^2 \le \frac{1}{4}h^2\|r(T)\mathbf{z}\|_2^2, \qquad \mathbf{z} = [z_{1,1},\dots,z_{N,N}]^T \in R^{N^2}. \tag{5.9}$$

Hence the desired result follows from (5.9), (5.8), and (5.3). □

THEOREM 5.1. *If $u \in C^6(\overline{\Omega})$ and u_h and S are defined by* (1.3)–(1.5) *and* (1.9)*, and* (4.2)–(4.5)*, respectively, then $\|S-u_h\|_{C(\overline{\Omega})} \le C(u)h^4$.*

Proof. Following the proof of Theorem 4.3, we define $\{w_{i,j}\}_{i,j=1}^N$ by

$$(L_x+L_y)(S-u_h)(x_i,y_j) = w_{i,j}, \quad i,j = 1,\dots,N. \tag{5.10}$$

Using (5.10), (1.9), and (1.1), we obtain

$$w_{i,j} = L_xS(x_i,y_j) - D_x^2u(x_i,y_j) + L_yS(x_i,y_j) - D_y^2u(x_i,y_j).$$

Equations (1.10), (4.59), and (4.53) give, for $i,j = 1,\dots,N$,

$$\begin{aligned}
&L_xS(x_i,y_j) - D_x^2u(x_i,y_j) = D_x^2S(x_i,y_j) - D_x^2u(x_i,y_j)\\
&+\frac{1}{12}[D_x^2S(x_{i-1},y_j) - 2D_x^2S(x_i,y_j) + D_x^2S(x_{i+1},y_j)]\\
&= -\frac{h^2}{12}D_x^4u(x_i,y_j) + \frac{1}{12}[D_x^2u(x_{i-1},y_j) - 2D_x^2u(x_i,y_j) + D_x^2u(x_{i+1},y_j)]\\
&-\frac{h^2}{144}[D_x^4u(x_{i-1},y_j) - 2D_x^4u(x_i,y_j) + D_x^4u(x_{i+1},y_j)] + \epsilon_{i,j},
\end{aligned}$$

where $|\epsilon_{i,j}| \le C(u)h^4$, $i,j = 1,\dots,N$. Hence Taylor's theorem and similar considerations for $L_yS(x_i,y_j) - D_y^2u(x_i,y_j)$ show that (4.84) holds. It follows from (3.1), (4.36), and (4.82) that the matrix-vector form of (5.10) is

$$(A\otimes B + B\otimes A)\mathbf{v} = \mathbf{w},$$

where A, B are as in (5.2), $\mathbf{v} = [s_{1,1}-u_{1,1},\dots,s_{N,N}-u_{N,N}]^T$, $\mathbf{w} = [w_{1,1},\dots,w_{N,N}]^T$. Hence Lemma 5.1 implies (4.86) and the desired result follows from the proof of Theorem 4.3. □

THEOREM 5.2. *If u in $C^6(\overline{\Omega})$ and u_h in $S^D\otimes S^D$ are the solutions of* (1.1)*, and* (1.3)–(1.5) *and* (1.9)*, respectively, then*

$$\|u-u_h\|_{C(\overline{\Omega})} \le C(u)h^4. \tag{5.11}$$

Proof. The required inequality follows from (4.14) and Theorems 4.1, 4.2, and 5.1. □

It is claimed in Theorem 4.1 of [8] that for the scheme (1.3) and (1.7)–(1.9), one has (5.11) provided that $u \in C^6(\overline{\Omega})$. The proof of this claim in [8] is based on using Z

defined in (4.6)–(4.9) as a comparison function. It is claimed, for example, in Lemma 2.1 of [8] that Z has properties (4.59) and (4.60), that is, (4.59) and (4.60) hold with Z in place of S. Unfortunately, numerical examples indicate that such property does not hold even in the one-dimensional case. Specifically, for $u(x) = x(x-1)e^x$ and $Z \in S^D$ such that

$$Z(x_i) = u(x_i), \quad i = 1, \ldots, N, \qquad Z''(x_i) = u''(x_i), \quad i = 0, N+1,$$

we only have

$$\max_{1 \le i \le N} \left| Z''(x_i) - u''(x_i) + \frac{h^2}{12} u^{(4)}(x_i) \right| = Ch^2$$

and not better. It should be noted that the convergence analysis of [6] for two-point boundary value problems involves the comparison function $S \in S^D$ defined by

$$S(x_i) = u(x_i), \quad i = 1, \ldots, N, \qquad S''(x_i) = u''(x_i) - \frac{h^2}{12} u^{(4)}(x_i), \quad i = 0, N+1,$$

which, in part, was motivation for the definition (4.2)–(4.5). The convergence analysis of the scheme (4.2)–(4.4) in [8] remains an open problem. We believe that such analysis may require proving stability not only with respect to the right-hand side but also with respect to the boundary conditions.

6. Numerical results. We used scheme (1.3)–(1.6) and the algorithm in section 3 to solve a test problem (1.1). The computations were carried out in double precision. We determined the nodal and global errors using the formulas

$$\|w\|_h = \max_{0 \le i,j \le N+1} |w(x_i, y_j)|, \qquad \|w\|_{C(\overline{\Omega})} \approx \max_{0 \le i,j \le 501} |w(t_i, t_j)|,$$

where $t_i = i/501$, $i = 1, \ldots, 501$. Convergence rates were determined using the formula

$$\text{rate} = \frac{\log(e_{N/2}/e_N)}{\log[(N+1)/(N/2+1)]},$$

where e_N is the error corresponding to the partition $\rho_x \times \rho_y$.

We took f in (1.1) corresponding to the exact solution

$$u(x, y) = 3e^{xy}(x^2 - x)(y^2 - y).$$

We see from the results in Tables 1 and 2 that the scheme (1.3)–(1.6) produces fourth order accuracy for u in both the discrete and the continuous maximum norms. We also observe superconvergence phenomena since the derivative approximations at the partition nodes are of order four.

TABLE 1
Nodal errors and convergence rates for u, u_x, u_y, and u_{xy}.

	$\|u-u_h\|_h$		$\|(u-u_h)_x\|_h$		$\|(u-u_h)_y\|_h$		$\|(u-u_h)_{xy}\|_h$	
N	Error	Rate	Error	Rate	Error	Rate	Error	Rate
4	7.305–05		9.219–04		9.219–04		1.673–02	
8	6.574–06	4.097	8.789–05	3.998	8.789–05	3.998	2.134–03	3.504
16	5.036–07	4.040	6.715–06	4.044	6.715–06	4.044	2.158–04	3.603
32	3.585–08	3.983	4.712–07	4.005	4.712–07	4.005	1.881–05	3.678
64	2.380–09	4.001	3.129–08	4.001	3.129–08	4.001	1.497–06	3.734
128	1.534–10	4.000	2.017–09	4.000	2.017–09	4.000	1.127–07	3.774

TABLE 2
Global errors and convergence rates for u, u_x, u_y, and u_{xy}.

	$\|u-u_h\|_{C(\overline{\Omega})}$		$\|(u-u_h)_x\|_{C(\overline{\Omega})}$		$\|(u-u_h)_y\|_{C(\overline{\Omega})}$		$\|(u-u_h)_{xy}\|_{C(\overline{\Omega})}$	
N	Error	Rate	Error	Rate	Error	Rate	Error	Rate
4	8.630–05		1.125–03		1.125–03		1.654–02	
8	8.606–06	3.922	1.186–04	3.827	1.186–0	3.827	2.169–03	3.456
16	6.776–07	3.996	1.232–05	3.561	1.232–05	3.561	2.593–04	3.339
32	4.763–08	4.003	1.267–06	3.429	1.267–06	3.429	2.997–05	3.253
64	3.157–09	4.003	1.308–07	3.350	1.308–07	3.350	3.308–06	3.251
128	2.038–10	3.998	1.467–08	3.192	1.467–08	3.192	3.823–07	3.148

Acknowledgment. The authors would like to thank Prof. Thomas Lucas for his assistance during the review process.

REFERENCES

[1] J. H. AHLBERG, E. N. NILSON, AND J. L. WALSH, *The Theory of Splines and Their Applications*, Academic Press, New York, 1967.
[2] D. ARCHER, *An $O(h^4)$ cubic spline collocation method for quasilinear parabolic equations*, SIAM J. Number. Anal., 14 (1977), pp. 620–637.
[3] B. BIALECKI, G. FAIRWEATHER, AND A. KARAGEORGHIS, *Matrix decomposition algorithms for modified spline collocation for Helmholtz problems*, SIAM J. Sci. Comput., 24 (2003), pp. 1733–1753.
[4] B. BIALECKI, G. FAIRWEATHER, AND A. KARAGEORGHIS, *Optimal superconvergent one step nodal cubic spline collocation methods*, SIAM J. Sci. Comput., 27 (2005), pp. 575–598.
[5] W. CHENEY AND D. KINCAID, *Numerical Mathematics and Computing*, Brooks/Cole Publishing Co., Pacific Grove, CA, 1999.
[6] J. W. DANIEL AND B. K. SWARTZ, *Extrapolated collocation for two-point boundary-value problems using cubic splines*, J. Inst. Math Appl., 16 (1975), pp. 161–174.
[7] C. DE BOOR, *The Method of Projections as Applied to the Numerical Solution of Two Point Boundary Value Problems Using Cubic Splines*, Ph.D. thesis, University of Michigan, Ann Arbor, MI, 1966.
[8] E. N. HOUSTIS, E. A. VAVALIS, AND J. R. RICE, *Convergence of $O(h^4)$ cubic spline collocation methods for elliptic partial differential equations*, SIAM J. Numer. Anal., 25 (1988), pp. 54–74.
[9] A. A. SAMARSKI, *The Theory of Difference Schemes*, Marcel Dekker, Inc., New York, 2001.
[10] C. VAN LOAN, *Computational Frameworks for the Fast Fourier Transform*, Frontiers Appl. Math. 10, SIAM, Philadelphia, 1992.

SIAM J. NUMER. ANAL.
Vol. 46, No. 1, pp. 419–436
© 2008 Society for Industrial and Applied Mathematics

$\mathcal{Y}$-SCALE REGULARIZATION*

HERBERT EGGER†

Abstract. Many inverse problems are ill-posed in the sense that their solution is unstable with respect to data perturbations, and hence *regularization methods* have to be used for their stable solution. Two possible drawbacks of standard regularization methods are (a) saturation, i.e., only suboptimal approximations can be found for smooth solutions—this is the case, e.g., for Tikhonov regularization; (b) numerical effort, e.g., the large number of iterations of methods like Landweber iteration. A framework that allows us to overcome both effects at least for certain classes of inverse problems is regularization in Hilbert scales, where a solution is searched in a scale of spaces (Hilbert scale), but convergence is achieved in the original space. Regularization methods in Hilbert scales can be viewed as modified (preconditioned) versions of standard methods. In order to make the advantages of the Hilbert scale approach applicable to a new class of problems, we propose to use a scale of spaces over the image space instead. This results in a new family of methods which we call $\mathcal{Y}$*-scale regularization*, and whose (optimal) convergence properties are analyzed. One of the key steps in the analysis is the formulation of an adequate a posteriori stopping rule and the proof of optimal convergence rates. The theoretical results are illustrated in several numerical examples.

Key words. inverse problems, regularization, Hilbert scales, preconditioning

AMS subject classifications. 15A29, 65F10, 65F22, 65J22

DOI. 10.1137/060654426

1. Introduction.

We consider the solution of linear inverse problems of the form

$$Tx = y^\delta, \tag{1.1}$$

where $T \in L(\mathcal{X}, \mathcal{Y})$ is a linear bounded injective operator between Hilbert spaces $\mathcal{X}$ and $\mathcal{Y}$; without loss of generality we assume $\|T\| \le 1$. Such problems arise in a variety of applications; cf., e.g., [2, 3, 5, 8, 15] and the references therein. In practice, one tries to reconstruct the solution from perturbed measurements y^δ, and we assume that at least a bound on the data error

$$\|y - y^\delta\| \le \delta \tag{1.2}$$

is available, where $y = Tx^\dagger \in \mathcal{R}(T)$ denotes the unperturbed data corresponding to the *true* solution $x^\dagger$. It is well known (cf., e.g., [8, 11]) that in case $\mathcal{R}(T)$ is not closed in $\mathcal{Y}$, problem (1.1) is *ill-posed*, i.e., (1.1) might not admit a unique solution, and more severely, a solution is usually unstable with respect to data noise $y - y^\delta$. In order to construct a stable approximate solution of (1.1) we apply *regularization* methods of the form

$$x_\alpha^\delta := R_\alpha y^\delta := g_\alpha(T^*T)T^*y^\delta, \tag{1.3}$$

generated by a family of piecewise continuous (filter-) functions $\{g_\alpha\}_{\alpha \in \mathcal{R}_+}$ with the property $g_\alpha(\lambda) \to 1/\lambda$ as $\alpha \to 0$ pointwise for all $\lambda > 0$. For a convergence (rate)

*Received by the editors March 16, 2006; accepted for publication (in revised form) May 3, 2007; published electronically January 30, 2008. This work was partially supported by the Austrian Science Foundation (FWF) under grant SFB F013, "Numerical and Symbolic Scientific Computing," project F1308.

http://www.siam.org/journals/sinum/46-1/65442.html

†Center for Computational Engineering Science, RWTH Aachen University, Pauwelstraße 19, D-52074 Aachen, Germany (herbert.egger@rwth-aachen.de).

analysis in case of perturbed data, the following assumptions are usually made; cf. [8]:

$$\sup_{0\le\lambda\le 1} |g_\alpha(\lambda)| \le c_* \alpha^{-1}, \tag{1.4}$$

$$\sup_{0\le\lambda\le 1} \lambda^\mu |r_\alpha(\lambda)| \le c_\mu \alpha^\mu, \qquad 0 \le \mu \le \mu_0, \tag{1.5}$$

where $r_\alpha(\lambda) := 1 - \lambda g_\alpha(\lambda)$, and c_μ, c_* are positive constants independent of $\alpha \in \mathbb{R}_+$. A maximal constant μ_0 for which (1.5) holds is called *qualification* of the method. In case of iterative regularization methods, we use the notation $g_k = g_\alpha$ and replace α in (1.4), (1.5) by adequate functions of k. If g_k is a polynomial, then condition (1.4) already follows from (1.5) by Markov's inequality.

For illustration and later reference, we mention some important examples of regularization methods of the form (1.3).

1. *Tikhonov regularization* [9, 32]. The approximate solutions are defined by

   $$x_\alpha^\delta = (T^*T + \alpha)^{-1} T^* y^\delta. \tag{1.6}$$

   Here, g_α is defined by $g_\alpha(\lambda) := (\lambda + \alpha)^{-1}$, and (1.5) holds for all $\mu \le 1$; i.e., the qualification of Tikhonov regularization is $\mu_0 = 1$.
2. *Landweber iteration* [13, 18]. The approximations x_k^δ are defined by the iteration

   $$x_{k+1}^\delta = x_k^\delta + T^*(y^\delta - T x_k^\delta), \qquad x_0^\delta = 0.$$

   Landweber iteration has the form (1.3) with $g_k(\lambda) = \sum_{i=0}^{k-1}(1-\lambda)^i$ and $r_k(\lambda) = (1-\lambda)^k$, respectively, and satisfies (1.5) for all $\mu \ge 0$ with α replaced by $\frac{1}{k}$ and $g_\alpha = g_k$.
3. *The ν-methods by Brakhage* [4, 12]. Semi-iterative methods whose residual polynomials r_k form an orthogonal sequence with respect to some positive weight function satisfy a three-term recurrence, which also carries over to the iterates; i.e., there exist sequences μ_k and ω_k such that

   $$x_k^\delta = x_{k-1}^\delta + \mu_k(x_{k-1}^\delta - x_{k-2}^\delta) + \omega_k T^*(y^\delta - T x_{k-1}^\delta), \quad k \ge 1,$$

   with $x_0^\delta = x_{-1}^\delta = x_0$. The choice $\mu_1 = 0$, $\omega_1 = (4\nu+2)/(4\nu+1)$, and

   $$\mu_k = \frac{(k-1)(2k-3)(2k+2\nu-1)}{(k+2\nu-1)(2k+4\nu-1)(2k+2\nu-3)},$$
   $$\omega_k = 4\frac{(2k+2\nu-1)(k+\nu-1)}{(k+2\nu-1)(2k+2\nu-1)}, \qquad k > 1,$$

   yields the ν-methods by Brakhage. Each ν-method satisfies (1.5) with α replaced by $\frac{1}{k^2}$ for $0 \le \mu \le \nu$.

Under condition (1.4) the regularized solutions x_α^δ converge to $x^\dagger$ with $\delta \to 0$ if $\alpha \to 0$ such that $\delta^2/\alpha \to 0$. In general, this convergence will be arbitrarily slow [28], and convergence rates can be obtained if and only if the true solution satisfies certain smoothness requirements, so-called *source conditions*. For

$$x^\dagger \in \mathcal{R}((T^*T)^\mu), \tag{1.7}$$

for some $\mu \le \mu_0$, the estimate

$$\|x_\alpha^\delta - x^\dagger\| = O(\delta^{\frac{2\mu}{2\mu+1}}) \tag{1.8}$$

can be obtained for the parameter choice

$$\alpha = \alpha(\delta) \sim \delta^{\frac{2}{2\mu+1}}, \tag{1.9}$$

and one can show that the rate (1.8) is optimal under this source condition. Such a source condition is not only sufficient but also (almost) necessary in order to obtain convergence at a rate (1.8); cf., e.g., [8]. While in general, the meaning of (1.7) is difficult to understand, sufficient conditions for validity of (1.7) can be given for certain problems by a problem adapted analysis; cf. [10] for a nonlinear problem in parameter identification of a differential operator; see Assumption 3.1 below.

If the source condition (1.7) holds for some $\mu > \mu_0$, still only a rate of $\delta^{\frac{2\mu_0}{2\mu_0+1}}$ can be expected; this phenomenon is called *saturation*. Tikhonov regularization, for instance, does not provide a better rate than $\delta^{\frac{2}{3}}$, while for Landweber iteration the rate may be arbitrarily close to δ if μ is sufficiently large. For an iterative method like Landweber iteration the number of iterations yielding optimal convergence rates is $n_* = O(\delta^{-\frac{2}{2\mu+1}})$ by (1.9); i.e., for small μ (nonsmooth solution $x^\dagger$) one has $n_* \sim O(\delta^{-2})$. Hence, for $\delta = 10^{-3}$ one would need about 10^6 iterations to stay within the regime of optimal convergence.

One way to overcome saturation and/or reduce the number of iterations and to relate the source condition (1.7) to standard differentiability assumptions, is provided by *regularization in Hilbert scales* introduced by Natterer [23] for Tikhonov regularization, and later analyzed for more general regularization methods in [8, 16, 30] and also for nonlinear problems [26, 27, 31]. We will recall the definition of a Hilbert scale and the most important results on regularization in Hilbert scales in section 2. The general idea is the following: We assume that the smoothing properties of the operator T can be characterized in terms of a simple, e.g., differential, operator L. The standard analysis of regularization in Hilbert scales is then based on the condition

$$\underline{m}\|y\| \le \|L^a T^* y\| \le \overline{m}\|y\| \qquad \text{for all} \quad y \in \mathcal{Y} \tag{1.10}$$

for some a, $\underline{m}$, $\overline{m} > 0$. For illustration we give a short example.

Example 1.1. Let $T : L_2(0,1) \to L_2(0,1)$ denote the solution operator to

$$-q\,y'' = x \quad \text{in} \quad [0,1]; \quad u(0) = u(1) = 0, \tag{1.11}$$

i.e., $Tx = y$. We assume that $q \in L_2(0,1)$ satisfies $0 < q_0 \le q \le q_1$ for some positive constants q_0, $q_1 \in \mathbb{R}$. One can show that $\mathcal{R}(T) = H^2(0,1) \cap H_0^1(0,1)$, and by the closed range theorem T is a homeomorphism between the spaces $L_2(0,1)$ and $H^2(0,1) \cap H_0^1(0,1)$.

Now let $L : \mathcal{D}(L) = H^2(0,1) \cap H_0^1(0,1) \to L_2(0,1)$ be defined by $Lx = -x''$. Since L is self-adjoint and positive definite, one can define fractional powers of L; then $\|x\|_s := \|L^s x\|$ are the natural norms on a scale of (Sobolev) spaces (cf. Definition 2.1). If q is sufficiently smooth, e.g., $q \in H^2(0,1)$, then (1.10) holds for $a = 1$.

Instead of (1.3) the regularized solutions can now be defined by the modified method

$$x_\alpha^\delta := g_\alpha(L^{-2s}T^*T)L^{-2s}T^*y^\delta, \qquad s \ge -a/2. \tag{1.12}$$

As shown in [8, 23, 30], saturation effects of a method g_α due to finite qualification μ_0 can be overcome by choosing $s > 0$ in (1.12); in this case L^{-2s} is smoothing. Conversely, for $s < 0$ the operator L^{-2s} acts as a preconditioner, and in this way

iterative regularization methods can be accelerated; cf. [6, 7]. Note that for $s < 0$ the right inequality in (1.10) is already needed to ensure well-definedness of the method (1.12).

Condition (1.10) characterizes the smoothing properties of the operator T^* in terms of the operator L, respectively, the scale of spaces over the preimage space $\mathcal{X}$ generated by it. Regularization in Hilbert scales is applicable if the range of T^* consists of sufficiently smooth functions. Note that in Example 1.1 condition (1.10) cannot hold for any $a \geq 1/4$ if q has jumps. Nevertheless, the condition

$$\underline{m}\|x\| \leq \|L^a T x\| \leq \overline{m}\|x\| \qquad \text{for all} \quad x \in L_2(0,1) \tag{1.13}$$

holds with $a = 1$ also for nonsmooth parameters q. In view of (1.13) one might call T smoothing like two times integration, while (1.10) would only suggest smoothing properties like integration of power $1/2$ at most. Similar examples, where (1.13) characterizes the smoothing properties of T more precisely than (1.10), arise from integral operators $(Tx)(s) := \int k(s,t)x(t)dt$ with kernels $k(s,t)$ that are differentiable with respect to s but not with respect to t; see section 5 for a detailed example.

In this paper, we analyze the solution of the inverse problem (1.1) by modified regularization methods of the form

$$x_\alpha^\delta := g_\alpha(T^* L^{-2s} T) T^* L^{-2s} y^\delta, \qquad s \geq -a/2, \tag{1.14}$$

under condition (1.13) or relaxed versions of it. In this way, the advantages of Hilbert scale regularization—overcoming saturation by choosing $s > 0$, respectively, by preconditioning by setting $s < 0$—can be made accessible for a new class of problems. Note that for methods (1.14) the operator L acts on the image space $\mathcal{Y}$ and generates a scale of spaces over $\mathcal{Y}$. For this reason we call the methods of the form (1.14) *$\mathcal{Y}$-scale regularization*.

The outline of the paper is as follows. We start by recalling the basic results on regularization in Hilbert scales. In section 3, we discuss regularization methods of the form (1.14) and derive the main error estimates. Section 4 contains the analysis of $\mathcal{Y}$-scale regularization methods under relaxed assumptions with some emphasis on the preconditioning of iterative regularization methods and the design of an order optimal a posteriori stopping rule. We conclude with presenting several examples and numerical tests illustrating the theoretical results.

2. Regularization in Hilbert scales. Before we quote the most important results concerning regularization in Hilbert scales, we shortly recall the definition of a Hilbert scale; see [17] or [8, section 8.4] for details.

DEFINITION 2.1. *Let L be a densely defined, unbounded, self-adjoint, and strictly positive operator in a Hilbert space $\mathcal{Z}$, and let $\mathcal{M} := \bigcap_{k=0}^\infty D(L^k)$. By $\mathcal{Z}_s$ we denote the completion of $\mathcal{M}$ with respect to the norm $\|z\|_s := \|L^s z\|_{\mathcal{Z}}$; $(\mathcal{Z}_s)_{s\in\mathbb{R}}$ is called a Hilbert scale (induced by L).*

Note that obviously $\|z\|_0 = \|z\|_{\mathcal{Z}}$. Two implications of this construction are that $\mathcal{Z}_{-s} = \mathcal{Z}_s'$ and that the following interpolation inequality holds, cf., e.g., [8, section 8.4]:

$$\|z\|_r \leq \|z\|_q^{\frac{s-r}{s-q}} \|z\|_s^{\frac{r-q}{s-q}} \tag{2.1}$$

for $-\infty < q < r < s < \infty$ and $z \in \mathcal{Z}_s$.

Remark 2.2. Like in Example 1.1, L is chosen to be a (simple) differential operator in most situations and the spaces $\mathcal{Z}_s$ coincide with standard Sobolev spaces for a

certain range of values of s. Note that the Sobolev spaces H^s do not form a Hilbert scale in general; cf. [25]. Fractional powers of L can then be efficiently implemented by Fourier transform or multilevel techniques.

Another choice of L always yielding (1.10) with $a = 1$ is $L := (T^*T)^{-\frac{1}{2}}$. At least for preconditioning, e.g., for $s = -a/2$, such a choice is, however, impractical since applying $(T^*T)^{-\frac{1}{2}}$ in (1.12) or (1.14) is at least as difficult as solving the original problem (1.1). If T is self-adjoint, then $L^{-2s} = (T^*T)^{-\frac{1}{2}}$ can be applied without any effort by skipping the terms $L^{-2s}T^* = I$, and the method (1.3) simplifies to

$$x_\alpha^\delta = g_\alpha(T)\, y^\delta,$$

which would be the default formulation of a regularization method for symmetric problems.

The standard assumption for regularization in Hilbert scales is the following norm equivalence condition:

$$\underline{m}\,\|x\|_{-a} \le \|Tx\| \le \overline{m}\,\|x\|_{-a} \tag{2.2}$$

for some $\underline{m}$, $\overline{m}$, $a > 0$. This condition is equivalent to (1.10) (cf. [7]; a similar result is derived in Proposition 3.3). If $x^\dagger \in \mathcal{X}_u$ for some $0 < u \le a + 2s$ and the operator T satisfies (2.2), then the regularization method (1.12) yields order optimal convergence rates

$$\|x_{\alpha_*}^\delta - x^\dagger\| = O(\delta^{\frac{u}{a+u}}) \qquad \text{for} \qquad \alpha_* \sim \delta^{\frac{2(a+s)}{a+u}}; \tag{2.3}$$

cf. [8, 30]. By setting $u = 2a\mu$, one can see that the rate (2.3) coincides with the standard results (1.8), and since the rate holds for $u \le a + 2s$, saturation can be overcome by choosing s sufficiently large. Note that under assumption (2.2) the source condition $x^\dagger \in \mathcal{X}_u$ is equivalent to $x^\dagger \in \mathcal{R}((T^*T)^{\frac{u}{2a}})$ for $u \le a$.

A different view on regularization in Hilbert scale has been taken in [6, 7]: If an iterative regularization method is used for the solution of (1.1), then (2.3) yields that the number of iterations needed to obtain the optimal convergence rates increases with s. Thus, from a numerical point of view (in particular for nonsmooth solutions) it is advantageous to choose s as small as possible, eventually $s < 0$. In this case the operator L^{-2s} in (1.12) acts as a preconditioner, and by setting $s = -a/2$, the number of iterations needed to guarantee the optimal rates (2.3) can be reduced to the square root, e.g., $n(\delta) = O(\delta^{-\frac{a}{a+u}})$ for Landweber iteration. Additionally, the analysis can be carried out under the weaker condition

$$\|Tx\| \le \overline{m}\|x\|_{-a}, \quad \text{respectively,} \quad \|T^*y\|_a \le \overline{m}\|y\|_{\mathcal{Y}}, \tag{2.4}$$

which allows us to apply the approach for a significantly wider class of problems.

THEOREM 2.3 (cf. Theorem 1 in [7]). *Let condition* (2.4) *hold,* g_α *satisfy* (1.4), (1.5), *and* $-a/2 \le s \le 0$. *Assume that* $x^\dagger$ *satisfies the source condition*

$$x^\dagger = L^{-s}(B^*B)^{\frac{u}{2(a+s)}}\, w, \qquad B := TL^{-s}, \tag{2.5}$$

for some $w \in \mathcal{X}$, *and let* x_k^δ *be defined by* (1.12). *Then the rate* (2.3) *holds.*

If the stronger condition (1.10) is valid, then the source condition (2.5) is $u \le a+2s$ equivalent to $x^\dagger \in \mathcal{X}_u$ usually used in regularization in Hilbert scales; cf. [6, 7] for a detailed discussion. We turn now to the analysis of $\mathcal{Y}$-scale regularization methods and prove the convergence rates (2.3) for methods (1.14) under condition (1.13).

3. $\mathcal{Y}$-scale regularization. Let $\{\mathcal{Y}_s\}_{s\in\mathbb{R}}$ be a Hilbert scale induced by some densely defined self-adjoint strictly positive operator $L : \mathcal{D}(L) \subset \mathcal{Y} \to \mathcal{Y}$ (cf. Definition 2.1), and let $\|\cdot\|_r$ denote the norm of $\mathcal{Y}_r$, i.e., $\|y\|_r = \|L^r y\|$. In what follows we require that the operator T satisfies the following assumption.

Assumption 3.1. There exist positive real constants $\underline{m}$, $\overline{m}$, and a such that

$$\underline{m}\|x\| \le \|Tx\|_a \le \overline{m}\|x\| \tag{3.1}$$

holds uniformly for all $x \in X$.

Note that Assumption 3.1 already implies infectivity of the operator T. For the stable solution of the inverse problem (1.1) we consider modified regularization methods of the form

$$x_\alpha^\delta := g_\alpha(T^*L^{-2s}T)T^*L^{-2s}y^\delta = g_\alpha(B^*B)B^*L^{-s}y^\delta \qquad \text{with} \quad B := L^{-s}T. \tag{3.2}$$

Here and below we assume that the functions g_α satisfy the standard assumptions (1.4), (1.5); $s \in \mathbb{R}$ will be specified later. Throughout, A^* denotes the adjoint of a linear operator A with respect to the spaces $\mathcal{X}$ and $\mathcal{Y}$. Under Assumption 3.1, the operator B is bounded as operator from $\mathcal{X}$ to $\mathcal{Y}$ for $s \ge -a$. Without loss of generality, we will assume $\|B\| \le 1$, which can always be achieved by proper scaling.

Remark 3.2. Problem (1.1) is equivalent to

$$Bx = L^{-s}y^\delta, \tag{3.3}$$

where B has to be understood as an operator between $\mathcal{X}$ and $\mathcal{Y}$, respectively, T in (1.1) as an operator from $\mathcal{X}$ to $\mathcal{Y}_s$. Note that T^*L^{-2s} is the adjoint of T with respect to these spaces. The $\mathcal{Y}$-scale method (3.2) hence corresponds to applying the regularization method defined by g_α to the solution of (3.3). Note that in case $s < 0$, the right-hand side $L^{-s}y^\delta$ may not be an element of $\mathcal{Y}$. In any case the noise level estimate (1.2) is not given in the appropriate norm if $s \ne 0$.

For the subsequent error analysis, we need some preliminary results.

PROPOSITION 3.3. *Assumption* 3.1 *is equivalent to*

$$\underline{m}\|y\|_{-a} \le \|T^*y\| \le \overline{m}\|y\|_{-a}. \tag{3.4}$$

Proof. First assume that (3.1) holds, let $y \in \mathcal{Y}$, and observe that

$$\sup_{y\in\mathcal{Y}} \frac{\|T^*y\|}{\|y\|_{-a}} = \sup_{y\in\mathcal{Y}} \sup_{x\in\mathcal{X}} \frac{\langle T^*y, x\rangle}{\|x\|\|y\|_{-a}} = \sup_{x\in\mathcal{X}} \sup_{y\in\mathcal{Y}} \frac{\langle y, Tx\rangle}{\|x\|\|y\|_{-a}} \le \sup_{x\in\mathcal{X}} \frac{\|Tx\|}{\|x\|} \le \overline{m}.$$

Hence, T^* is extendable as a bounded linear operator to $\mathcal{Y}_{-a}$. The result then follows similar to the above by noting that

$$\sup_{y\in\mathcal{Y}_{-a}} \frac{\|T^*y\|}{\|y\|_{-a}} = \sup_{x\in\mathcal{X}} \frac{\|Tx\|}{\|x\|}.$$

The other implication follows similarly. ∎

PROPOSITION 3.4. *Let Assumption* 3.1 *hold and let* $s \ge -a$. *Then for* $|\nu| \le 1$

$$\underline{c}(\nu)\|y\|_{-\nu(a+s)} \le \|(BB^*)^{\frac{\nu}{2}}y\| \le \overline{c}(\nu)\|y\|_{-\nu(a+s)}$$

holds for $y \in \mathcal{D}((BB^*)^{\frac{\nu}{2}})$ *with* $\underline{c}(\nu) = \min(\underline{m}^\nu, \overline{m}^\nu)$ *and* $\overline{c}(\nu) = \max(\underline{m}^\nu, \overline{m}^\nu)$. *Moreover,*

$$\mathcal{D}((BB^*)^{-\frac{\nu}{2}}) = \mathcal{R}((BB^*)^{\frac{\nu}{2}}) = \mathcal{Y}_{\nu(a+s)},$$

where in the case $\nu > 0$, $(BB^*)^{\frac{\nu}{2}}$ is identified with its extension to $\mathcal{Y}$.

Proof. The result follows by the inequality of Heinz; see, e.g., [8, Corollary 8.22] for details. ∎

It is well known that in order to obtain convergence rates, the true solution $x^\dagger$ has to satisfy some smoothness assumptions; cf. [28]. As will become clear from our analysis the appropriate source condition for $\mathcal{Y}$-scale regularization has to be stated in terms of the operator B (see also Remark 3.2 and Theorem 2.3).

Assumption 3.5 (source condition). Let $s > -a$. There exist a real number $u > 0$ and an element $w \in \mathcal{X}$ such that

$$x^\dagger = (B^*B)^{\frac{u}{2(a+s)}} w. \tag{3.5}$$

The total error $\|x_\alpha^\delta - x^\dagger\|$ can be decomposed into the two main components, namely, the *approximation error* $\|x_\alpha - x^\dagger\|$ and the *propagated data error* $\|x_\alpha^\delta - x_\alpha\|$, where x_α denotes the approximate solution defined by (3.2) with y^δ replaced by the correct data $y = Tx^\dagger$. We will now derive estimates for the two error components in terms of α and δ.

THEOREM 3.6. *Let Assumption* 3.1 *hold, let y^δ satisfy* (1.2)*, and let x_α^δ, x_α be defined by the $\mathcal{Y}$-scale method* (3.2) *with some $s \geq -a/2$. Furthermore, let g_α satisfy* (1.4), (1.5) *for some $\mu_0 > 0$ and let $x^\dagger$ satisfy Assumption* 3.5 *for $u \leq 2(a+s)\mu_0$. Then the following a priori estimates hold:*

$$\|x_\alpha - x^\dagger\| \leq C_1\, \alpha^{\frac{u}{2(a+s)}} \|w\|, \tag{3.6}$$

$$\|x_\alpha^\delta - x_\alpha\| \leq C_2\, \alpha^{-\frac{a}{2(a+s)}}\, \delta, \tag{3.7}$$

where C_1, C_2 denote generic constants depending only on c_μ, c_*, $\underline{m}$, and $\overline{m}$. As a consequence

$$\|x^\dagger - x_{\alpha_*}^\delta\| = O(\delta^{\frac{u}{a+u}}) \qquad \text{for} \quad \alpha_* \sim \delta^{\frac{2(a+s)}{a+u}}. \tag{3.8}$$

Proof. We start by estimating the approximation error by using (1.5) and (3.5),

$$s\|x_\alpha^\delta - x^\dagger\| = \|g_\alpha(B^*B)B^*L^{-s}y - x^\dagger\| \tag{3.9}$$

$$= \|r_\alpha(B^*B)(B^*B)^{\frac{u}{2(a+s)}} w\| \tag{3.10}$$

$$\leq c_{\frac{u}{2(a+s)}}\, \alpha^{\frac{u}{2(a+s)}} \|w\|. \tag{3.11}$$

Now consider the propagated data error. By Proposition 3.4 we have $\mathcal{R}(L^{-s}) = \mathcal{Y}_s = \mathcal{R}((BB^*)^{\frac{s}{2(a+s)}})$, hence

$$x_\alpha^\delta - x_\alpha = g_\alpha(B^*B)B^*L^{-s}(y^\delta - y) = g_\alpha(B^*B)B^*(BB^*)^{\frac{s}{2(a+s)}} v$$

for some $v_\delta \in \mathcal{Y}$ with $\|v_\delta\| \leq c_1\|y - y^\delta\|$. Thus, we can further estimate

$$\|x_\alpha^\delta - x_\alpha\| = \|g_\alpha(B^*B)B^*(BB^*)^{\frac{s}{2(a+s)}} v_\delta\| \tag{3.12}$$

$$\leq c_2\, \|g_\alpha(B^*B)(B^*B)^{\frac{a+2s}{2(a+s)}}\| \|y - y^\delta\| \tag{3.13}$$

$$\leq c_3\, \alpha^{-\frac{a}{2(a+s)}}\, \delta, \tag{3.14}$$

where for the last estimate we utilized that

$$\sup_{\lambda \in [0,1]} |g_\alpha(\lambda)\lambda^\mu| \leq C_\mu\, \alpha^{\mu-1}, \qquad 0 \leq \mu \leq 1,$$

which follows readily from (1.4) and (1.5). Equation (3.8) is implied by (3.6) and (3.7). ∎

Remark 3.7. In contrast to standard regularization in Hilbert scales, the source condition (3.5) cannot be interpreted in terms of the spaces $\mathcal{Y}_r$ directly. However, the following interpretation is possible. Define z by $x = B^*z$. Then problem (1.1) is equivalent to

$$BB^*z = L^{-s}y^\delta, \qquad x = B^*z.$$

The source condition (3.5) can be restated for z by noting

$$B^*z^\delta = x^\dagger = (B^*B)^{\frac{u}{2(a+s)}}w = B^*(BB^*)^{\frac{u-a-s}{2(a+s)}}v,$$

hence $z^\delta \in \mathcal{R}((BB^*)^{\frac{u-a-s}{2(a+s)}}) = \mathcal{Y}_{u-a-s}$. The last identity is valid only for $u \le a+s$. A similar restriction also holds for standard regularization in Hilbert scales, where the the source condition $x^\dagger \in \mathcal{X}_u$ guarantees optimal convergence rates only for $u \le a+2s$; cf. [8, Remark 8.24]. Finally, let Assumption 3.1 hold and let $x^\dagger$ satisfy a standard source condition $x^\dagger = T^*w$ for some $w \in \mathcal{Y}$. Then we obtain by Proposition 3.4 that

$$x^\dagger = T^*w = B^*L^sw = B^*(BB^*)^{-\frac{s}{2(a+s)}}\tilde{w} = (B^*B)^{\frac{a}{2(a+s)}}\hat{w}$$

holds for some $\hat{w} \in \mathcal{X}$ with $\|\hat{w}\|_{\mathcal{X}} \sim \|w\|_{\mathcal{Y}}$. Hence by the inequality of Heinz we at least have $\mathcal{R}((T^*T)^\mu) = \mathcal{R}((B^*B)^{\frac{\mu a}{a+s}})$ for $0 \le \mu \le \frac{1}{2}$.

Since in practice the smoothness of the solution $x^\dagger$ is not known a priori, an optimal parameter choice (3.8) is not possible in general. Therefore, we next consider an a posteriori parameter selection criterion based on the discrepancy principle (cf. [8, 22]), i.e.,

$$\alpha_* := \sup\{\alpha : \|Tx_\alpha^\delta - y^\delta\| \le \tau\delta\}. \tag{3.15}$$

The advantage of such a criterion is that it requires no a priori information about the smoothness of $x^\dagger$. Moreover, the residual $r_k^\delta := Tx_k^\delta - y^\delta$ usually has to be calculated anyway, in particular for iterative regularization methods.

THEOREM 3.8. *Let the assumptions of Theorem* 3.6 *hold and let* $\alpha_* = \alpha_*(\delta)$ *be determined by the discrepancy principle* (3.15). *Then the rates* (3.8) *hold for* $0 < u \le 2(a+s)\mu_0 - a$.

Proof. We have to estimate the residual $\|Tx_\alpha^\delta - y^\delta\|$,

$$s\|Tx_\alpha^\delta - y^\delta\| = \|L^s(Bg_\alpha(B^*B)B^* - I)L^{-s}y^\delta\| \tag{3.16}$$

$$\le c_1\,\|r_\alpha(BB^*)B(B^*B)^{\frac{u-s}{2(a+s)}}\|\|w\| + c_2\,\delta \tag{3.17}$$

$$\le c_{\frac{u+a}{2(a+s)}}\,\alpha^{\frac{u+a}{2(a+s)}}\|w\| + c_2\,\delta, \tag{3.18}$$

where we used (1.5) and $u \le 2(a+s)\mu_0 - a$. Hence,

$$\alpha_* \ge C\,\delta^{\frac{2(a+s)}{a+u}} \tag{3.19}$$

as long as $\tau > c_2$, which yields a sufficient bound for the propagated data error by (3.7). It remains to consider the approximation error. First note that we have

$$\|T(x_{\alpha_*} - x^\dagger)\| \le \|Tx_{\alpha_*}^\delta - y^\delta\| + \|T(x_{\alpha_*} - x_{\alpha_*}^\delta)\| + \delta \le C\,\delta.$$

It then follows by the interpolation inequality and (3.5) that

$$\|x_{\alpha_*} - x^\dagger\| \le c_3 \|T(x_{\alpha_*} - x^\dagger)\|^{\frac{u}{a+u}} \|w\|^{\frac{a}{a+u}} \le c_4 \delta^{\frac{u}{a+u}} \|w\|^{\frac{a}{a+u}},$$

which completes the proof. ∎

Remark 3.9. For $s = 0$ the above results coincide with the standard regularization theory by noting that $u = 2a\mu$. The restriction $u \le 2(a+s)\mu_0 - a$ in Theorem 3.8 then amounts to $\mu \le \mu_0 - 1/2$, which is the well-known saturation for the discrepancy principle. As for regularization in Hilbert scales, saturation effects of the applied method can be overcome by choosing s large enough.

4. $\mathcal{Y}$-scale regularization under relaxed assumptions—preconditioning. Probably the most severe drawback of $\mathcal{Y}$-scale regularization as presented in the previous section is the restrictive assumptions (3.1) (corresponding to (1.10) for standard regularization in Hilbert scales). As we will show now, most of the results remain valid if only one of the estimates in (3.1) is available. We will treat in detail the case where the right inequality holds and $s \le 0$ has to be chosen; the corresponding results for the other inequality and the choice $s \ge 0$ are indicated in remarks. As can be seen from (3.8), the (optimal) regularization parameter α_* can be increased by choosing $s < 0$. For an iterative regularization method $g_\alpha = g_n$, this means that the corresponding stopping index decreases by setting $s < 0$. This motivates calling the operator L^{-2s} in (1.14) a preconditioner. In order to stay with the notation of the previous section, we continue to consider general regularization methods g_α and mention some implications for iterative regularization methods in separate remarks.

Preconditioning of iterative regularization methods in Hilbert scales has been investigated in [6, 7] under the standard assumption (1.10) and under weaker conditions. Here, we are using scales of spaces over $\mathcal{Y}$ and the assumption (4.1). The derivation of the convergence results of the following sections therefore requires different reasoning and a more sophisticated a posteriori stopping rule.

Let L and $\mathcal{Y}_s$ be defined as above. Throughout this section we require the following assumption.

Assumption 4.1. T is injective and there exists a positive real constant a such that

$$\|Tx\|_a \le \|x\| \tag{4.1}$$

holds uniformly for all $x \in X$.

Unlike (3.1), where the injectivity of the operator T was implied by the left estimate, it has to be assumed here additionally. Similarly as in the previous section we have the following equivalent characterization of Assumption 4.1.

PROPOSITION 4.2. *Assumption* 4.1 *is equivalent to* T *being injective and*

$$\|T^* y\| \le \|y\|_{-a} \qquad \text{for all } y \in \mathcal{Y}_{-a}. \tag{4.2}$$

Proof. Let x, y, $\hat{y} \ne 0$. Assumption 4.1 can be restated as

$$\begin{aligned} 1 &\ge \sup_{x \in \mathcal{X}} \frac{\|L^a T x\|}{\|x\|} = \sup_{x \in \mathcal{X}} \sup_{y \in \mathcal{Y}} \frac{\langle Tx, L^a y\rangle}{\|x\|\|y\|} \\ &= \sup_{\hat{y} \in \mathcal{Y}_{-a}} \sup_{x \in \mathcal{X}} \frac{\langle x, T^* \hat{y}\rangle}{\|x\| \|\hat{y}\|_{-a}} = \sup_{\hat{y} \in \mathcal{Y}_{-a}} \frac{\|T^* \hat{y}\|}{\|\hat{y}\|_{-a}}. \end{aligned}$$

The other implication follows in the same way. ∎

Due to the relaxed assumption (4.1), Proposition 3.4 is no longer valid in its general form. Instead, only the following weaker implications hold.

PROPOSITION 4.3. *Let Assumption* 4.1 *hold, and let* $B = L^{-s}T$ *for some* $s \geq -a$ *with* $\|B\| \leq 1$. *Then the estimates*

$$\|(BB^*)^{\frac{\nu}{2}} y\| \leq \|y\|_{-\nu(a+s)} \qquad \text{for all } y \in \mathcal{Y}_{-\nu(a+s)} \subset \mathcal{R}((BB^*)^{-\frac{\nu}{2}}), \tag{4.3}$$

$$\|(BB^*)^{-\frac{\nu}{2}} y\| \geq \|y\|_{\nu(a+s)} \qquad \text{for all } y \in \mathcal{R}((BB^*)^{\frac{\nu}{2}}) \subset \mathcal{Y}_{\nu(a+s)} \tag{4.4}$$

hold for $0 \leq \nu \leq 1$.

Proof. The results follow from the inequality of Heinz and the standard results on Hilbert scales; see [8, section 8.5] for details. ∎

Assumption 4.1 and Proposition 4.3 are sufficient to prove the following a priori error estimates corresponding to those of Theorem 3.6.

THEOREM 4.4. *Let* $-a/2 \leq s \leq 0$ *and let Assumption* 4.1 *hold. Additionally, let* g_α, $x^\dagger$, y^δ *satisfy the assumptions of Theorem* 3.6. *Then* (3.6), (3.7), *and* (3.8) *hold.*

Proof. The estimate for the approximation error $x_\alpha^\delta - x^\dagger$ follows as in the proof of Theorem 3.6. Now observe that $L^{-s}(y^\delta - y) \in \mathcal{R}(L^{-s}) = \mathcal{Y}_s \subset \mathcal{R}((BB^*)^{\frac{s}{2(a+s)}})$ is implied by (4.3) since $s \leq 0$. Hence, also the estimate for the propagated data error follows as in Theorem 3.6. ∎

Remark 4.5. In case $s \geq 0$, condition (4.1) has to be replaced by

$$\|x\| \leq \|Tx\|_{\tilde{a}}. \tag{4.5}$$

Proposition 4.3 then holds with opposite inclusions and the analogue of Theorem 4.4 follows with obvious modifications. Again, similarly as in the proof of Theorem 3.6, there exists a $v_\delta \in \mathcal{Y}$ such that $L^{-s}(y - y^\delta) = (BB^*)^{\frac{s}{2(\tilde{a}+s)}} v_\delta$, since for $s \geq 0$ and under condition (4.5),

$$\mathcal{R}(L^{-s}) = \mathcal{Y}_s \subset \mathcal{R}((BB^*)^{\frac{s}{2(\tilde{a}+s)}})$$

follows from the analogue of Proposition 4.3 for this case.

4.1. A posteriori parameter choice. The result of Theorem 4.4 is not of great use per se, since it would require exact knowledge of the smoothness of the solution $x^\dagger$. We will therefore consider a posteriori parameter choice rules in the following.

A discrepancy principle. In contrast to the results of the previous section, the discrepancy principle (3.15) does not provide parameters that guarantee optimal convergence rates if only Assumption 4.1 is used. To see this, let us estimate the residual

$$\|Tx_\alpha^\delta - y^\delta\| = \|L^s(Bg_\alpha(B^*B)B^* - I)L^{-s}y^\delta\| \tag{4.6}$$

$$\begin{aligned} &\leq \|L^s r_\alpha(BB^*)B(B^*B)^{\frac{u}{2(a+s)}} w\| \\ &\qquad + \|L^s g_\alpha(BB^*)BB^*L^{-s}(y^\delta - y)\| + \|y^\delta - y\| \end{aligned} \tag{4.7}$$

$$\leq c_1 \alpha^{\frac{u+a+s}{2(a+s)}} + c_2 \alpha^{\frac{s}{2(a+s)}} \delta \tag{4.8}$$

for some positive constants c_1 and c_2. A careful inspection of the proof of Theorem 3.8 reveals that the second term may disturb the convergence of the residual $Tx_\alpha^\delta - y^\delta$; i.e., one cannot guarantee that $\inf\{\|Tx_\alpha^\delta - y^\delta\| : \alpha > 0\} \leq \tau\delta$ holds for any τ (independent of δ and y^δ) and for all possible data y^δ. Hence, the discrepancy principle (3.15) might not yield a parameter α_* at all. A partial result can, however, be found for the following modified criterion: Let α_{max} be defined by

$$\alpha_{max} := \sup\{\alpha : \|Tx_\alpha^\delta - y^\delta\| \leq \tau\delta\alpha^{\frac{s}{2(a+s)}}\} \tag{4.9}$$

for some appropriate $\tau > 1$. We can estimate α_{max} from below by

$$\tau\delta\alpha_{max}^{\frac{s}{2(a+s)}} \le \|Tx_\alpha^\delta - y^\delta\| \le c_1\,\alpha^{\frac{u+a+s}{2(a+s)}} + c_2\,\alpha^{\frac{s}{2(a+s)}}\delta$$

for all $\alpha > \alpha_{max}$. Hence, assuming that $\tau > c_2$, we obtain with an appropriate constant $\overline{C} > 0$ that

$$\alpha_{max} \ge \overline{C}\delta^{\frac{2(a+s)}{a+u}}. \tag{4.10}$$

In view of Theorems 3.6 and 3.8, a parameter choice $\alpha_* = \alpha_{max}$ cannot provide optimal convergence rates in general, but at least the following suboptimal result holds.

PROPOSITION 4.6. *Let the iterative method* (1.14) *be stopped at* $\alpha_* = \alpha_{max}$. *Then the following (suboptimal) rate holds:*

$$\|x_\alpha^\delta - x^\dagger\| = O\left(\delta^{\frac{u+\frac{us}{2(a+s)(a+u)}}{a+u}}\right). \tag{4.11}$$

Proof. In the same manner as (4.6) one obtains that

$$\|Tx_\alpha - y\| \le \alpha^{\frac{s}{2(a+s)}}\delta.$$

The result then follows similarly as in the proof of Theorem 3.8. ∎

Remark 4.7. For $s = 0$ the rate (4.11) coincides with the optimal rate $O(\delta^{\frac{u}{a+u}})$. If $s < 0$, the convergence rate observed for this parameter choice is not optimal; e.g., for $s = -a/2$, (4.11) yields

$$\|x_\alpha^\delta - x^\dagger\| = O(\delta^{\frac{u}{2(a+u)}}),$$

which is only half of the optimal rate (3.8). We only mention that even if the standard discrepancy principle (3.15) provides a parameter, i.e., if there exists an α such that $\|Tx_\alpha^\delta - y^\delta\| \le \tau\delta$, such a choice will in general only yield suboptimal rates similar to (4.11).

We will show next how optimal rates can be restored.

A Lepskij principle. In view of Theorem 3.6, α_{max} is a reasonable upper bound for appropriate parameters yielding optimal convergence rates. On the other hand, according to Theorem 3.6 optimal rates are always obtained for the choice

$$\alpha_* = \alpha_{opt} := \overline{C}\delta^{\frac{2(a+s)}{a+u}} \ge \overline{C}\delta^{\frac{2(a+s)}{a}} =: \alpha_{min} \tag{4.12}$$

with $\overline{C}$ as in (4.10). Hence, α_{min} may serve as a reasonable lower bound for interesting regularization parameters. A specific parameter yielding optimal convergence rates can be found in the range $[\alpha_{min}, \alpha_{max}]$ by means of the Lepskij principle (cf. [19, 21]); we use here a formulation similar to the one presented in [1].

Let α_{min}, α_{max} be determined according to (4.9), (4.12), and let α_n be defined by $\alpha_n := q^n \cdot \alpha_{max}$ for some $0 < q < 1$. We define the set $\mathcal{M}$ (of admissible parameters) by

$$\mathcal{M} := \left\{ \alpha_n \in [\alpha_{min}, \alpha_{max}] : \begin{array}{l} \|x_{\alpha_n}^\delta - x_{\alpha_m}^\delta\| \le 4C\alpha_n^{-\frac{a}{2(a+s)}}\delta \\ \text{for all } m = n+1, \dots, N \end{array} \right\}, \tag{4.13}$$

where N is the largest integer such that $\alpha_N > \alpha_{min}$, and C is a generic constant such that

$$\max\{\|x_\alpha^\delta - x_\alpha\|, \|x_\alpha - x^\dagger\|\} \le C\alpha^{-\frac{a}{2(a+s)}}\delta \qquad \text{for all} \quad \alpha \le \alpha_{opt}. \tag{4.14}$$

Note that by the error estimates (3.6), (3.7), such a constant can be chosen depending on (a bound on) $\|w\|$, but without knowledge of u or α_{opt}. As appropriate parameter, we select

$$\alpha_* := \begin{cases} \max \mathcal{M} & \text{if } \mathcal{M} \neq \emptyset, \\ \alpha_N & \text{else.} \end{cases} \tag{4.15}$$

THEOREM 4.8. *Let the assumptions of Theorem* 4.4 *be satisfied and let* α_* *be defined by* (4.15). *Then the rate*

$$\|x_{\alpha_*} - x^\dagger\| = O(\delta^{\frac{u}{a+u}}) \tag{4.16}$$

holds.

Proof. By the definition of α_{opt} we have $\alpha_{opt} \in [\alpha_{min}, \alpha_{max}]$. Additionally, it follows from the error estimates (3.6), (3.7), and (4.14) that $\alpha_* \ge \alpha_{opt}$. Hence, we obtain together with (4.13) that

$$\|x_{\alpha_*}^\delta - x^\dagger\| \le \|x_{\alpha_{opt}}^\delta - x^\dagger\| + \|x_{\alpha_*}^\delta - x_{\alpha_{opt}}^\delta\| = O(\delta^{\frac{u}{a+u}}). \qquad \square$$

Remark 4.9. A disadvantage of the Lepskij principle is that its realization requires a higher computational effort than the standard discrepancy principle (3.15). Note that the construction of the set $\mathcal{M}$ requires us to generate solutions for a wide range of regularization parameters. Additionally, at least some solutions $x_{\alpha_n}^\delta$ have to be stored in order to check the conditions for the construction of $\mathcal{M}$. By using the upper bound α_{max} on admissible parameters, the number of solutions $x_{\alpha_n}^\delta$ which have to be stored can be reduced significantly; note that usually one chooses $\alpha_{max} \sim 1$. Additionally, the $\mathcal{Y}$-scale approach allows us to choose the smallest regularization parameter α_{min} relatively large, e.g., $\alpha_{min} \sim \delta$ for $s = -a/2$, whereas one would need $\alpha_{min} \sim \delta^2$ without preconditioning; cf. [1, 21].

Remark 4.10. The results derived for general regularization methods (1.3) hold with obvious modifications also for iterative regularization methods. If the strong condition (3.1) holds and $s = -a/2$, Theorem 3.8 yields optimal convergence rates when the method is stopped at iteration k_* defined by the discrepancy principle

$$\|Tx_{k_*}^\delta - y^\delta\| \le \tau\delta < \|Tx_k^\delta - y^\delta\| \qquad \text{for all} \quad k < k_*.$$

As can be seen from (3.8), respectively, (2.3), the number of iterations needed to get the optimal rates (2.3) can be reduced to about the square root by proper preconditioning ($s = -a/2$).

Under the weaker assumption (4.1), and with $s = -a/2$, one has to perform $k_{max} := 1/\alpha_{min} \sim \delta^{-1}$ iterations in order to apply the Lepskij principle (4.15). Note that if $x^\dagger$ is not very smooth, i.e., if μ in (1.9) is small, which is the case we are interested in, then one would expect $k_* = O(\delta^{\frac{2}{2\mu+1}}) \sim O(\delta^{-2})$ iterations for (the standard version of) Landweber iteration. Hence also in this case, the number of iterations can be reduced to about the square root by preconditioning.

Remark 4.11. If instead of Assumption 4.1 condition (4.5) holds and $s \ge 0$, then the results concerning the discrepancy principle are slightly different from

those presented here; e.g., the standard discrepancy principle (3.15) can be used to determine α_{max}. However, as in the case $s \le 0$, a choice $\alpha_* = \alpha_{max}$ does not yield optimal convergence rates in general. The result for the Lepskij principle follows as above with obvious modification. For $s \ge 0$ and if iterative regularization methods are applied, the number of iterations will be increased by application of L^{-2s}, which in case $s \ge 0$ is a smoothing operator. Such choice seems to be disadvantageous from a numerical point of view, at least for iterative methods.

We only mention that if a condition (4.5) holds for some $\tilde{a} \ge a$ additionally to (4.1) (cf. Example 5.2), then some of the previous results can be strengthened. In particular, for $\tilde{a} = a$, the results of section 3 apply.

5. Examples and numerical tests. In this section we verify Assumption 3.1, respectively, Assumption 4.1, for several examples and present the results of the numerical tests illustrating the theoretical results.

Example 5.1 (Radon transform). Let $f \in C_0^\infty(\Omega^n)$ for some compact domain $\Omega^n \subset \mathbb{R}^n$, $n \ge 2$. The Radon transform defined by

$$Rf(\omega, s) = \int_{x\cdot\omega=s} f(x)dx$$

is important in many applications, e.g., in computerized tomography, and has been studied extensively in the literature. The following very general stability estimate holds; cf. [24, 29]:

$$c(\gamma, n)\|f\|_{H_0^\gamma(\Omega^n)} \le \|Rf\|_{H^{\gamma+(n-1)/2}(Z)} \le C(\gamma, n)\|f\|_{H_0^\gamma(\Omega^n)}, \tag{5.1}$$

where Z denotes the unit cylinder in $\mathbb{R}^n$, and

$$\|g\|_{H^{\gamma+(n-1)/2}(Z)} := \int_{S^{n-1}} \int_{\mathbb{R}} (1+\sigma^2)^\gamma |\hat{g}(\theta, \sigma)|^2 d\sigma\, d\theta.$$

Here, the Fourier transform is only taken with respect to the second variable. Due to (5.1) with $\alpha = 0$, the results of section 3 are applicable with $Ly(\theta, s) = -\partial_{ss} y(\theta, s) + y(\theta, s)$; fractional powers of the operator L can be realized efficiently by Fourier transform.

Example 5.2 (an integral equation allowing a lower bound). Consider the Fredholm integral equation $Tx = y$ with operator $T : L_2(0,1) \to L_2(0,1)$ defined by

$$Tx(s) = \int_0^1 k(s,t)dt \quad \text{with} \quad k(s,t) = \sqrt{t} \begin{cases} s(1-t), & 0 \le s < t \le 1, \\ t(1-s), & 0 \le t \le s \le 1. \end{cases}$$

Since $k(s,t) \in L_2([0,1]^2)$, it follows that T is compact and $Tx = y$ is ill-posed. With

$$(Tx)(s) = (1-s)\int_0^s t^{3/2} x(t)dt + s\int_s^1 t^{1/2}(1-t)x(t)dt$$

we get $(Tx)(0) = (Tx)(1) = 0$. By twice differentiation we obtain that

$$(Tx)''(s) = -s^{1/2}x(s),$$

which yields

$$\mathcal{R}(T) = \{y \in H^2(0,1) \cap H_0^1(0,1) : s^{-1/2} y''(s) \in L_2(0,1)\}.$$

We utilize the Hilbert scale induced by the operator L defined by

$$L^s y := \sum_{n=1}^{\infty} (n\pi)^s \langle x, x_n \rangle y_n \qquad \text{with} \quad y_n := \sqrt{2}\sin(n\pi\cdot), \tag{5.2}$$

i.e., $L^2 x = -x''$ and $\mathcal{Y}_2 = H^2(0,1) \cap H_0^1(0,1)$. This choice yields $\mathcal{R}(T) \subsetneq \mathcal{Y}_2$ and $\mathcal{R}(T) \supset \mathcal{Y}_{2.5} = \{y \in H^{2.5}(0,1) \cap H_0^1(0,1) : \rho^{-1/2} y'' \in L_2(0,1)\}$ with $\rho(t) = t(1-t)$. By Theorem 11.7 in [20], we get

$$\underline{m}\,\|x\| \le \|Tx\|_{2.5} \quad \text{and} \quad \|Tx\|_2 \le \overline{m}\,\|x\|$$

for some positive constants $\underline{m}$, $\overline{m}$, and the results of section 4 can be applied.

Next we compare Landweber iteration and the ν-methods by Brakhage [4, 12] for the numerical solution of $Tx = y$. As parameter choice strategy we use the Lepskij principle (4.15). We shortly describe how this stopping rule is implemented for Landweber iteration: As soon as the modified discrepancy principle (4.9) is satisfied, we set $\mathcal{M}_x := \{k_{min}\}$, where k_{min} plays the role of α_{max} in (4.9). Note that we store the iterates x_k^δ rather than the iteration indices k. At each iteration $k_n := 2^n \cdot k_{min}$, $n \in \mathbb{N}$, we add the element $x_{k_n}^\delta$ to the set $\mathcal{M}_x$ and check if the condition

$$\|x_{k_n}^\delta - x_{k_m}^\delta\| \le 4C k_n^{\frac{a}{2(a+s)}}\,\delta \tag{5.3}$$

holds for all $x_{k_n}^\delta$, $x_{k_m}^\delta$ in the set $\mathcal{M}_x$ with $m > n$. If (5.3) does not hold for some n, we eliminate all elements $x_{k_m}^\delta$ with $m < n$. This procedure is continued until the final iteration k_N is reached, where N is the largest integer such that $k_N < k_{max} = C\delta^{-1}$. As stopping index k_* we then choose the smallest k_n with x_{k_n} in $\mathcal{M}_x$.

For a concrete numerical test, we set $x^\dagger(s) = \operatorname{sign}(1-2s)$. With $a = 2$, $s = -1$, and the above choice of a Hilbert scale, we have

$$x^\dagger \in \mathcal{R}(B^*B)^{\frac{u}{2(a+s)}} = \left\{ x \in \mathcal{Y}_2 : \frac{1}{\sqrt{s}} \left(\frac{x}{\sqrt{s}} \right)'' \in L_2(0,1) \right\}$$

for $0 \le u < 1$. The correct data $y = Tx^\dagger$ are perturbed by uniformly distributed random noise such that $\|y - y^\delta\| = \delta$ for several values of δ. In a first test, we compare the performance of the discrepancy principle and the Lepskij rule suggested in the previous section: we set $\delta = 0.002$, $\tau = 1.1$ and plot the two error components (approximation error and propagated data error) for Landweber iteration with preconditioning ($s = -a/2$) in Figure 5.1. The optimal stopping index, i.e., the one where the total error $\|x_k^\delta - x^\dagger)\|$ attains its minimum, is 230, whereas the standard discrepancy principle stops only after 469 iterations; cf. (4.6) and Remark 4.9. The Lepskij principle (4.15) with $C = 0.25$ yields $k_* = 96$. The corresponding errors $e_k = \|x_k^\delta - x^\dagger\|$ are $e_{230} = 0.26$, $e_{469} = 0.30$, and $e_{96} = 0.27$.

In a second test, we illustrate the effect of preconditioning and list the iteration numbers and errors for several values of δ obtained for Landweber iteration and the ν-methods by Brakhage [4, 12] with and without preconditioning in Table 5.1. Note that the numerical effort for an iteration step is about the same for all methods and consists of application of T and T^*. The preconditioned iteration requires further application of a differential operator, which is cheap in comparison with application of the integral operator T.

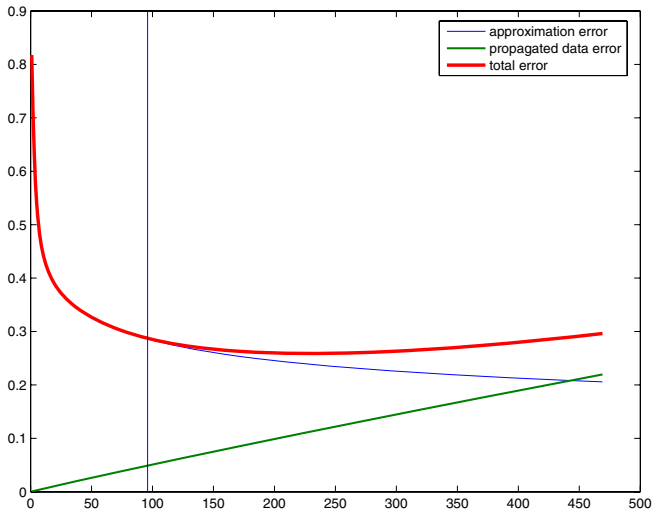


FIG. 5.1. *Approximation error, propagated data error, and total error, Example* 5.2*, with* $\delta = 0.002$*,* $\tau = 1.1$*, and iterates obtained by the preconditioned Landweber iteration. The vertical line denotes the actual stopping index* n_* *determined by the Lepskij rule.*

TABLE 5.1
Iteration numbers and relative errors for Landweber iteration (lw), the ν*-method with* $\nu = 2$*, and the preconditioned versions, Example* 5.2*.*

δ	$k_*^{(lw)}$	$e_{k_*}^{(lw)}$	$k_*^{(hslw)}$	$e_{k_*}^{(hslw)}$	$k_*^{(nu)}$	$e_{k_*}^{(nu)}$	$k_*^{(hsnu)}$	$e_{k_*}^{(hsnu)}$
0.02	46	0.4299	20	0.4244	13	0.4099	9	0.4078
0.01	177	0.3920	38	0.3655	21	0.3655	12	0.3634
0.005	819	0.3307	81	0.3312	43	0.3333	18	0.3300
0.002	3488	0.2690	166	0.2704	112	0.2587	29	0.2589

The corresponding convergence rates are $e_{k_*} = \|x_{k_*}^\delta - x^\dagger\| = O(\delta^{0.18})$ for all iterations. The iteration numbers behave like $k_*^{(lw)} \sim \delta^{-1.7}$, $k_*^{(nu)} \sim \delta^{-0.87}$ for the standard Landweber iteration and the ν-method. In both cases the stopping indices can be reduced to the square root by preconditioning, i.e., $k_*^{(hslw)} \sim \delta^{-0.90}$ and $k_*^{(hsnu)} \sim \delta^{-0.48}$.

The following two examples shall serve as models for inverse problems governed by partial differential equations. Note that in two dimensions or three dimensions, an assembly of the full operator is usually impossible for reasonable discretization levels, and even the application of the operator T which involves the solution of the governing equation may be rather expensive. Thus such problems will usually be solved by some iterative method, and the number of iterations should be kept as small as possible. As we will illustrate, this can effectively be achieved by preconditioning in $\mathcal{Y}$-scales.

Example 5.3 (source reconstruction). Let $\Omega \subset \mathbb{R}^n$, $n = 2, 3$, be a bounded convex domain and let $T : L_2(\Omega) \to L_2(\Omega)$ be defined by $Tf = u$ where u satisfies

$$-q\Delta u + cu = f, \qquad u\big|_{\partial\Omega} = 0,$$

and c, $q \in L_2(\Omega)$ are known functions bounded from below by some constants $\underline{c} \geq 0$,

TABLE 5.2
Iteration numbers and relative errors for Landweber iteration (lw), the ν-method with $\nu = 2$, and the preconditioned versions, Example 5.3.

δ	$k_*^{(lw)}$	$e_{k_*}^{(lw)}$	$k_*^{(hslw)}$	$e_{k_*}^{(hslw)}$	$k_*^{(nu)}$	$e_{k_*}^{(nu)}$	$k_*^{(hsnu)}$	$e_{k_*}^{(hsnu)}$
0.02	424	0.451	34	0.587	55	0.456	14	0.538
0.01	1639	0.374	52	0.445	116	0.375	18	0.425
0.005	4723	0.307	91	0.359	192	0.310	26	0.355
0.0025	12794	0.251	161	0.282	321	0.253	33	0.275

$\underline{q} > 0$. We consider the $\mathcal{Y}$-scale induced by

$$L^2 : \mathcal{D}(L^2) \subset L_2(\Omega) \to L_2(\Omega) \quad \text{with} \quad L^2 u = -\Delta u,$$

and domain $\mathcal{D}(L^2) = H^2(\Omega) \cap H_0^1(\Omega) =: \mathcal{Y}_2$. It follows from standard regularity results for elliptic equations that the solution u satisfies $u \in \mathcal{Y}_2$, i.e., $\mathcal{R}(T) \subset \mathcal{Y}_2$. Moreover, if q is additionally bounded from above, then also the opposite inclusion holds; i.e., Assumption 3.1 is valid with $a = 2$.

For a numerical test, we consider $\Omega = (0,1)^2$, $q = 2 + \text{sign}(1-2x)$, $c = 1$. Let the true solution be defined by $f^\dagger = \text{sign}(1-2y)$ and choose $s = -a/2 = -1$. In view of Theorem 3.8, we can apply the standard discrepancy principle for this example. The results of the numerical tests are listed in Table 5.2.

Example 5.4 (exponentially ill-posed problems). As a final example we consider the backward heat equation as a model problem for exponentially ill-posed problems: Let $T : L_2(0,1) \to L_2(0,1)$ defined by $(Tg)(x) = y(x) := u(x,\bar{t})$ with some $\bar{t} > 0$ and

$$-u_t + q u_{xx} = 0, \qquad u(0,t) = u(1,t) = 0, \qquad u(x,0) = g.$$

The solution of this equation has the Fourier expansion

$$u(x,\bar{t};g) = \sum_{n=1}^{\infty} \exp(-q\bar{t}\pi^2 n^2)\langle g, y_n\rangle y_n$$

with the basis functions y_n as in Example 5.2. Let L^s be defined by (5.2). Then we have

$$\|Tg\|_r \le c(r)\|g\| \qquad \text{for all} \quad r \in \mathbb{R};$$

thus Assumption 4.1 can be satisfied for arbitrary a (up to scaling by a constant).

For a numerical test, we let $q = 0.01$, set $s = -2$ for preconditioning, and try to reconstruct the initial condition

$$g^\dagger = 2x - \text{sign}(2x-1) - 1$$

from measurements of u at time $\bar{t} = 1$.

Since the problem under consideration is exponentially ill-posed, only logarithmic convergence rates can be expected; cf. [14]. The solution of all considered methods look almost identical. Note that for this problem, our theory does not actually imply smaller stopping indices of the discrepancy principle for the preconditioned methods. However, as can be seen from Table 5.3, the number of iterations is reduced significantly for finite noise levels $\delta > 0$; cf. also [7] for a discussion of this phenomenon.

TABLE 5.3

Iteration numbers and relative errors for Landweber iteration (lw), the ν-method with $\nu = 2$, and the preconditioned versions, Example 5.4.

δ	$k_*^{(lw)}$	$e_{k_*}^{(lw)}$	$k_*^{(hslw)}$	$e_{k_*}^{(hslw)}$	$k_*^{(nu)}$	$e_{k_*}^{(nu)}$	$k_*^{(hsnu)}$	$e_{k_*}^{(hsnu)}$
0.02	13	0.501	9	0.498	6	0.492	5	0.486
0.01	21	0.488	12	0.486	9	0.482	7	0.476
0.005	37	0.482	15	0.482	15	0.473	8	0.474
0.0025	402	0.435	58	0.433	76	0.434	21	0.432

REFERENCES

[1] F. BAUER AND T. HOHAGE, *A Lepskij-type stopping rule for regularized Newton methods*, Inverse Problems, 21 (2005), pp. 1975–1991.

[2] J. BECK, B. BLACKWELL, AND C. S. CLAIR, *Inverse Heat Conductions*, Wiley, Sussex, 1985.

[3] M. BERTERO AND P. BOCCACCI, *Introduction to Inverse Problems in Imaging*, Institute of Physics Publishing, London, 1998.

[4] H. BRAKHAGE, *On ill-posed problems and the method of conjugate gradients*, in Inverse and Ill-Posed Problems, H. W. Engl and C. W. Groetsch, eds., Academic Press, Boston, New York, London, 1987, pp. 165–175.

[5] D. COLTON AND R. KRESS, *Inverse Acoustic and Electromagnetic Scattering*, Springer, Berlin, 1992.

[6] H. EGGER, *Semi-iterative regularization in Hilbert scales*, SIAM J. Numer. Anal., 44 (2006), pp. 66–81.

[7] H. EGGER AND A. NEUBAUER, *Preconditioning Landweber iteration in Hilbert scales*, Numer. Math., 101 (2005), pp. 643–662.

[8] H. W. ENGL, M. HANKE, AND A. NEUBAUER, *Regularization of Inverse Problems*, Kluwer Academic Publishers, Dordrecht, 1996.

[9] H. W. ENGL, K. KUNISCH, AND A. NEUBAUER, *Convergence rates for Tikhonov regularization for nonlinear ill-posed problems*, Inverse Problems, 5 (1989), pp. 523–540.

[10] H. W. ENGL AND J. ZOU, *A new approach to convergence rate analysis of Tikhonov regularization for parameter identification in heat conduction*, Inverse Problems, 16 (2000), pp. 1907–1923.

[11] C. W. GROETSCH, *Generalized Inverses of Linear Operators*, Dekker, New York, Basel, 1977.

[12] M. HANKE, *Accelerated Landweber iterations for the solution of ill-posed equations*, Numer. Math., 60 (1991), pp. 341–373.

[13] M. HANKE, A. NEUBAUER, AND O. SCHERZER, *A convergence analysis of the Landweber iteration for nonlinear ill-posed problems*, Numer. Math., 72 (1995), pp. 21–37.

[14] T. HOHAGE, *Iterative Methods in Inverse Obstacle Scattering: Regularization Theory of Linear and Nonlinear Exponentially Ill-Posed Problems*, Ph.D. thesis, University of Linz, Austria, 1999.

[15] V. ISAKOV, *Inverse Problems in Partial Differential Equations*, Springer, Berlin, 1998.

[16] J. JANNO AND U. TAUTENHAHN, *On Lavrentiev regularization for ill-posed problems in Hilbert scales*, Numer. Funct. Anal. Optim., 24 (2003), pp. 531–555.

[17] S. G. KREIN AND J. I. PETUNIN, *Scales of Banach spaces*, Russian Math. Surveys, 21 (1966), pp. 85–160.

[18] L. LANDWEBER, *An iteration formula for Fredholm integral equations of the first kind*, Amer. J. Math., 73 (1951), pp. 615–624.

[19] O. V. LEPSKIJ, *A problem of adaptive estimation in Gaussian white noise*, Teor. Veroyatnost. i Primenen., 35 (1990), pp. 459–70.

[20] J. L. LIONS AND E. MAGENES, *Non-Homogeneous Boundary Value Problems and Applications: Volume* I, Springer, Berlin, Heidelberg, 1972.

[21] P. MATHÉ AND S. V. PEREVERZEV, *Geometry of linear ill-posed problems in variable Hilbert scales*, Inverse Problems, 19 (2003), pp. 789–803.

[22] V. A. MOROZOV, *On the solution of functional equations by the method of regularization*, Soviet Math. Dokl., 7 (1966), pp. 414–417.

[23] F. NATTERER, *Error bounds for Tikhonov regularization in Hilbert scales*, Appl. Anal., 18 (1984), pp. 29–37.

[24] F. NATTERER, *The Mathematics of Computerized Tomography*, Teubner, Stuttgart, 1986.

[25] A. NEUBAUER, *When do Sobolev spaces form a Hilbert scale?*, Proc. Amer. Math. Soc., 103 (1988), pp. 557–562.

[26] A. Neubauer, *Tikhonov regularization of nonlinear ill-posed problems in Hilbert scales*, Appl. Anal., 46 (1992), pp. 59–72.
[27] A. Neubauer, *On Landweber iteration for nonlinear ill-posed problems in Hilbert scales*, Numer. Math., 85 (2000), pp. 309–328.
[28] E. Schock, *Approximate solution of ill-posed equations: Arbitrarily slow convergence versus superconvergence*, in Constructive Methods for the Practical Treatment of Integral Equations, G. Hämmerlein and K. H. Hoffmann, eds., Birkhäuser, Basel, 1985, pp. 234–243.
[29] K. T. Smith, D. C. Solomon, and S. L. Wagner, *Practical and mathematical aspects of the problem of reconstructing a function from radiographs*, Bull. Amer. Math. Soc. (N.S.), 83 (1977), pp. 1227–1270.
[30] U. Tautenhahn, *Error estimates for regularization methods in Hilbert scales*, SIAM J. Numer. Anal., 33 (1996), pp. 2120–2130.
[31] U. Tautenhahn, *On a general regularization scheme for nonlinear ill-posed problems:* II. *Regularization in Hilbert scales*, Inverse Problems, 14 (1998), pp. 1607–1616.
[32] A. N. Tikhonov, *Regularization of incorrectly posed problems*, Soviet Math. Dokl., 4 (1963), pp. 1624–1627.

SIAM J. NUMER. ANAL.
Vol. 46, No. 1, pp. 437–471
© 2008 Society for Industrial and Applied Mathematics

FINITE ELEMENT APPROXIMATION OF STOCHASTIC PARTIAL DIFFERENTIAL EQUATIONS DRIVEN BY POISSON RANDOM MEASURES OF JUMP TYPE*

ERIKA HAUSENBLAS†

Abstract. The paper deals with stochastic partial differential equations driven by Poisson random measures of jump type and their numerical approximation. We investigate the accuracy of space and time approximation. As space approximation we consider finite elements and as time approximation the implicit Euler scheme.

Key words. stochastic evolution equations, stochastic partial differential equations, numerical approximation, time discretization, space discretization, Poisson random measure

AMS subject classifications. 60H15, 60H35, 65C30, 35R30

DOI. 10.1137/050654141

1. Introduction.

Our point of interest is the numerical approximation solutions of stochastic partial differential equations (SPDEs) of the following type:

$$\begin{cases} du(t,\xi) = \frac{\partial^2}{\partial \xi^2} u(t-,\xi)\, dt + f(u(t-,\xi),\xi)\, dt \\ \qquad\qquad + \int_{\zeta\in\mathbb{R}} g(u(t-,\xi),\xi;\zeta)\, \tilde{\eta}(d\zeta,\xi,dt),\ \xi\in\mathcal{O},\ 0<t<\infty, \\ u(0,\xi) = u_0(\xi), \end{cases} \tag{1.1}$$

where $f : \mathbb{R}\times\mathcal{O}\to\mathbb{R}$ and $g : \mathbb{R}\times\mathbb{R}\to\mathbb{R}$ are global Lipschitz continuous functions and η is a space time Poisson random measure of impulsive type (see, e.g., Saint Loubert Bié [SLB98] or Peszat and Zabczyk [PZ07]).

Non-Gaussian random processes play an increasing rôle in modeling stochastic dynamical systems. Typical examples of non-Gaussian stochastic processes are Lévy processes and processes arising by Poisson random measures. In neurophysiology the driving noise of the cable equation is basically impulsive, e.g., of a Poisson type (see Walsh [Wal86, Chapter 3])—or, on the other hand, Woyczyński describes in [Woy01] a number of phenomena from fluid mechanics, solid state physics, polymer chemistry, economic science, etc., for which non-Gaussian Lévy processes can be used as their mathematical model in describing the related stochastic behavior. Thus, from the point of view of applications one might feel that the restriction to Gaussian noise is unsatisfactory; to handle such cases one can replace the Gaussian noise by a Poisson random measure.

Concerning the numerical approximation of SPDEs driven by Wiener noise, several works already exist. Just to mention a few (see, e.g., Gyöngy [Gyö99, Gyö98]), Hausenblas [Hau03, Hau02], Kloeden and Shott [KS01], Printems [Pri01a, Pri01b], Shardlow [Sha99], and Yoo [Yoo00]. Finite element approximation of SPDEs driven by Wiener space time white noise were considered by Allen, Novosel, and Zhang

*Received by the editors March 13, 2006; accepted for publication (in revised form) June 14, 2007; published electronically January 30, 2008. This research was supported by Austrian Science Foundation grant P17273.

http://www.siam.org/journals/sinum/46-1/65414.html

†Department of Mathematics, University of Salzburg, Hellbrunnerstr. 34, 5020 Salzburg, Austria (erika.hausenblas@sbg.ac.at).

[ANZ98], Cardon-Weber and Millet [CWM04], Germani and Piccioni [GP88], Gyöngy [Gyö99, Gyö98], Gyöngy and Millet [GM05], Millet and Morien [MM05], Walsh [Wal05], and Yan [Yan05]. The numerical approximation of finite-dimensional SDEs driven by Poisson random measures is the topic of, e.g., the work of Gardoń [Gar04], Kuznetsov [Kuz01], Protter and Talay [PT97], or Rubenthaler [Rub03]. But the numerical approximation of SPDEs driven by Poisson random measures is barely investigated, and, in the cases known to the author, processes of finite variation are considered, e.g., by Kouritzin, Long, and Sun [KLS03] and Kouritzin and Long [KL02]. If the driving process is of infinite variation, there exists only some works about existence and uniqueness, e.g., Albeverio, Wu, and Zhang [AWZ98], Hausenblas [Hau05, Hau06], and Kallianpur and Xiong [KX87, KX96]. Applebaum and Wu [AW00] and Saint Loubert Bié [SLB98] considered SPDEs with space time Lévy noise.

First, we discretize the SPDE spatially by the finite element method, obtaining a system of high-dimensional SDEs that can be solved by the implicit Euler scheme. The noise will be approximated by using the projection onto finite elements.

2. Preliminaries. Let $\mathcal{O} \subset \mathbb{R}^d$ be a convex polygon, $1 < p \leq 2$, and A be a uniformly elliptic, dissipative, second order partial differential operator with either Dirichlet or Neumann boundary conditions. In particular, let $A : D(A) \subset L^p(\mathcal{O}) \to L^p(\mathcal{O})$ be the realization of

$$\begin{aligned} \mathcal{L} &= \mathcal{L}(x, D) \\ &= -\sum_{i,j=1}^{2} \frac{\partial}{\partial x_i} a_{i,j}(x) \frac{\partial}{\partial x_j} + \sum_{i=1}^{2} b_i(x) \frac{\partial}{\partial x_i} + c(x), \end{aligned} \tag{2.1}$$

with either

$$u = 0 \quad \text{on} \quad \partial\mathcal{O} \times (0, \infty)$$

or

$$\frac{\partial u}{\partial n_L} + \sigma u = 0,$$

where $\frac{\partial u}{\partial n_L}$ denotes the differentiation along the outer normal unit vector n on $\mathcal{O}$. We assume that $a_{ij} \in C^2(\mathcal{O})$, $1 \leq i, j \leq d$, strictly bounded away from zero, and $\sigma \in C^2(\mathcal{O})$.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a complete probability space. Let $\eta : \mathcal{B}(\mathbb{R}) \hat{\times} \mathcal{B}(\mathcal{O}) \hat{\times} \mathcal{B}(\mathbb{R}^+) \to \mathbb{R}^+$ be a Poisson random measure over $(\Omega, \mathcal{F}, \mathbb{P})$ with intensity[1] $\iota = \lambda \hat{\times} \nu$, $\nu \in \mathcal{L}(\mathbb{R}) := \{\sigma : \mathcal{B}(\mathbb{R}) \to \mathbb{R}^+, \sigma \text{ is a Lévy measure}\}$, γ be the compensator, and $\tilde{\eta} = \eta - \gamma$. Let $(\mathcal{F}_t)_{t \geq 0}$ be the natural filtration induced by η.

Our point of interest is solutions of SPDEs of the following form:

$$\begin{cases} du(t, \xi) = Au(t, \xi)\, dt + f(u(t-, \xi), \xi)\, dt \\ \qquad\qquad + \int_{\zeta \in \mathbb{R}} g(u(t-, \xi), \xi; \zeta)\, \tilde{\eta}(d\zeta, \xi, dt), \ \xi \in \mathcal{O}, \ t > 0, \\ u(0, \xi) = u_0(\xi), \quad \xi \in \mathcal{O}, \end{cases} \tag{2.2}$$

where $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \times \mathcal{O} \times \mathbb{R} \to \mathbb{R}$ are globally Lipschitz continuous functions. By a mild solution of (1.1) we understand a process u such that

[1] λ denotes the Lebesgue measure.

- u satisfies the integral equation

$$u(t,\xi) = \int_{\mathcal{O}} G_t(\zeta,\xi)u_0(\zeta)\,d\zeta + \int_{0+}^{t}\int_{\mathcal{O}} G_{t-s}(\zeta,\xi)f(u(s-,\zeta,\zeta))\,d\zeta\,ds + \int_{0+}^{t}\int_{\zeta\in\mathcal{O}}\int_{z\in\mathbb{R}} G_{t-s}(\zeta,\xi)g(u(s-,\zeta),\zeta;z)\,\eta(dz,d\zeta,ds),\ \text{a.s.},\quad t\geq 0,$$

  where $G = \{G_t(\cdot,\star), 0 < t < \infty\}$ is the Green kernel associated with $\frac{\partial}{\partial t} + \mathcal{L}$ and the boundary conditions specified above;
- for all $\phi \in C_b^\infty(\mathcal{O})$, the $\mathbb{R}$-valued process $t \mapsto \langle u(t), \phi\rangle$ is adapted and cádlág.

It is shown in Saint Loubert Bié [SLB98] or Hausenblas [Hau05, Example 2.3] that if $p < \frac{2}{d} + 1$, then there exists a unique solution of problem (1.1) such that $u \in L^p(\Omega; L^p(\mathcal{O}))$.

Our point of interest is the numerical approximation of the solution of (1.1). First, we introduce the space discretization by finite elements, and second we introduce the time discretization, i.e., the Euler scheme.

2.1. Space discretization. The space discretization is done by finite elements. For the definition for a finite element, we follow the book of Brenner and Scott [BS02].

DEFINITION 2.1. *Let*

- $K \subset \mathbb{R}^d$ *be a bounded closed set with nonempty interior and piecewise smooth boundary (the* element domain*);*
- $\mathcal{P}$ *be a finite-dimensional space of functions on* K *(the space of* shape functions*); and*
- $\mathcal{N} = \{N_1, \ldots, N_k\}$ *be a basis for* $\mathcal{P}'$ *(the set of* nodal variables*).*

Then $(K, \mathcal{P}, \mathcal{N})$ *is called a* finite element*. Moreover, we denote by* $diam(K) = \max_{x,y\in K} |x-y|_{\mathbb{R}^d}$ *the diameter of* K*.*

DEFINITION 2.2. *A set* K *is* star-shaped with respect to B*, if for all* $x \in K$ *the closed convex hull of* $\{x\} \cup B$ *is a subset of* K*.*

DEFINITION 2.3. *Suppose* K *has diameter* d *and is star-shaped with respect to a ball. Let* $\rho_{max} = \sup\{\rho : K \text{ is star-shaped with respect to a ball of radius } \rho\}$*. Then the* chunkiness parameter *of* K *is defined by*

$$\gamma := \frac{d}{\rho_{max}}.$$

Moreover, we denote by P_m the set of all polynomials in d variables of degree less than or equal to m.

DEFINITION 2.4 (compare the assumptions of Theorem 4.4.4 in [BS02]). *We assume that a finite element satisfies the* standard conditions for some l and m*, iff*

- K *is star-shaped with respect to some ball,*
- $P_{m-1} \subset \mathcal{P} \subset W_\infty^m(\mathcal{O})$*,*
- $\mathcal{N} \subset (C^l(\bar{K}))'$*.*

DEFINITION 2.5. *Let* $(K, \mathcal{P}, \mathcal{N})$ *be a finite element. The basis* $\{\phi_1, \ldots, \phi_k\}$ *of* $\mathcal{P}$ *dual to* $\mathcal{N}$ *is called the* nodal basis *of* $\mathcal{P}$*.*

DEFINITION 2.6. *Given a finite element* $(K, \mathcal{P}, \mathcal{N})$*, let the set* $\{\phi_1, \ldots, \phi_k\} \subset \mathcal{P}$ *be the basis dual to* $\mathcal{N}$*. If* v *is a function for which all* $N_i \in \mathcal{N}$*,* $i = 1, \ldots, k$*, are defined, then we define the* local interpolant *by*

$$\mathcal{I}_K v := \sum_{i=1}^{k} v(N_i)\phi_i.$$

DEFINITION 2.7. *A* subdivision *of a domain $\mathcal{O}$ is a finite collection of element domains $\{K_i, i = 1, \dots k\}$ such that*

- $\int K_i \cap \int K_j = \emptyset$ *if* $i \neq j$ *and*
- $\cup K_i = \mathcal{O}$.

Remark 2.1. The nodal basis of a subdivision $\{K_i, i = 1, \dots k\}$ induces a Voronoi decomposition $\{J_i; i = 1, \dots, k\}$ of $\mathcal{O}$ by

$$J_i := \left\{ \xi \in \mathcal{O},\ |\xi - N_i| \leq \min_{1 \leq i \leq k} \{|\xi - N_j|\} \right\}.$$

DEFINITION 2.8. *Suppose that $\mathcal{O}$ is a polyhedral domain with subdivision $\mathcal{T}$. Assume that each element domain K in the subdivision is equipped with some type of shape functions $\mathcal{P}$ and nodal variables $\mathcal{N}$ such that $(K, \mathcal{P}, \mathcal{N})$ forms a finite element. Let m be the order of the highest partial derivatives involved in the nodal variables. For $f \in C^m(\bar{\mathcal{O}})$, the* global interpolant *is defined by*

$$\mathcal{I}_{\mathcal{T}} f \Big|_{K_i} = \mathcal{I}_{K_i} f$$

for all $K_i \in \mathcal{T}$.

Example 2.1. Let $H = L^2([0,1])$. Let $\{x_i := \frac{i}{N}, i = 1, \dots, N-1\}$ be the set of grid points. For $0 \leq i \leq N$ let ϕ_i be a pyramid function, i.e., a function, which takes value 1 at the grid point $\frac{i}{N}$, vanishes at the other grid points, and is linear between the grid points. For $i = 1, \dots, N-1$, a finite element $(K_i, \mathcal{N}_i, \mathcal{P}_i)$ consists of the interval $[\frac{i-1}{N}, \frac{i}{N}]$, the nodal variables $N_i : \mathcal{C}^0([0,1]) \ni v \mapsto v(x_i) \in \mathbb{R}$, the shape functions $\phi_i|_{K_i}$ and $\phi_{i-1}|_{K_i}$, and $J_i = [\frac{i-1}{N} + \frac{1}{2N}, \frac{i}{N} + \frac{1}{2N}]$.

We can now define a family of subdivisions $\mathcal{F}_h$, $h = N_h^{-1}$. For each subdivision $\mathcal{T}_h$, let $\mathcal{V}_h$ the linear hull of the shape functions, i.e., the linear hull of $\{\phi_i\}_{i=1}^{N_h - 1}$. For any $v \in C([0,1])$, the global interpolant is given by

$$\mathcal{I}_h(v) = \sum_{i=1}^{N_h} N_i^h(v) \phi_i.$$

DEFINITION 2.9. *We call a family of subdivisions $\{\mathcal{T}_h, 0 < h \leq 1\}$* nondegenerate *if the chunkiness parameter is uniformly bounded for all $T \in \{\mathcal{T}_h\}$ and all $h \in (0,1]$.*

Let $\{\mathcal{T}_h, 0 < h \leq 1\}$ be a nondegenerate family of subdivisions of a polyhedral domain $\mathcal{O}$ in $\mathbb{R}^d$, consisting of k_h finite elements of reference type $(K, \mathcal{P}, \mathcal{N})$. We assume that the parameter h corresponds to the maximal diameter of the finite elements, i.e., $h = \max\{1 \leq i \leq k_h \mid \operatorname{diam}(K_i)\}$. Let $\{\phi_1^h, \dots, \phi_{k_h}^h\}$ be the set of shape functions and $\{N_1^h, \dots, N_{k_h}^h\}$ be the set of nodal variables. We denote by $\mathcal{V}_h$ the linear hull of the shape functions and $\mathcal{V}_h'$ its dual. The semidiscrete problem corresponding to (2.2) is to find the process $u_h = \{u_h(t) \in \mathbb{R}^{k_h}, 0 \leq t < \infty\}$

$$\text{(2.3)} \qquad \begin{cases} du_h(t) = \mathcal{A}_h u_h(t)\, dt f_h(u_h(t-))\, dt \\ \qquad\qquad + \int_{i \in \{1, \dots, k_h\}} \int_{\zeta \in \mathbb{R}} g_h(u_h(t-), i, \zeta)\, \tilde{\eta}_h(d\zeta, di, dt),\ t > 0, \\ u(0) = (\langle u_0, N_1 \rangle, \dots, \langle u_0, N_{k_h} \rangle) \in \mathbb{R}^{k_h} \end{cases}$$

where $\mathcal{A}_h : \mathcal{V}_h \to \mathcal{V}_h$ is the discrete analogue of A defined by

$$\text{(2.4)} \qquad \langle \mathcal{A}_h \phi, \chi \rangle = \langle A\phi, \chi \rangle \quad \forall \phi, \chi \in \mathcal{V}_h,$$

and $\tilde{\eta}_h : \mathcal{B}(\mathbb{R})\hat{\times}\mathcal{P}(\{1,\dots,k_h\})\hat{\times}\mathcal{B}(\mathbb{R}^+) \to \mathbb{R}^+$ is the compensated Poisson random measure with intensity ι_h given by

$$\mathcal{B}(\mathbb{R}) \times \mathcal{B}(\{1,\dots,k_h\}) \ni (A,B) \mapsto \iota_h((A,B)) := \cup_{i\in B}\lambda(J_i^h)\,\nu(A) \in \mathbb{R}^+.$$

Moreover, $f_h : \mathcal{V}_h \to \mathcal{V}_h$ and $g_h : \mathcal{V}_h \times \{1,\dots k_h\} \times \mathbb{R} \to \mathcal{V}_h$ are given by

$$\mathcal{V}_h \ni v \mapsto f_h(v) := \sum_{i=1}^{k_h} f(v(N_i), N_i)\phi_i^h \in \mathcal{V}_h$$

and

$$\mathcal{V}_h \times \{1,\dots k_h\} \times \mathbb{R} \ni (v,i,\zeta) \mapsto g_h(v,i,\zeta) = g(v(N_i), N_i;\zeta)\, c_i^h\ \phi_i^h \in \mathcal{V}_h,$$

where $c_i^h := \left[\int_{\mathcal{O}} \phi_i^h(\xi)\, d\xi\right]^{-1}$.

Remark 2.2. The Poisson random measure $\tilde{\eta}_h$ can also be written as a sum of k_h independent Poisson random measures (see, e.g., [Lin86, Proposition 5.3.2]). To be precise, let $\tilde{\eta}_h^i : \mathcal{B}(\mathbb{R})\hat{\times}\mathcal{B}(\mathbb{R}^+) \to \mathbb{R}^+$ be pairwise independent compensated Poisson random measures with intensity $\nu^i = \lambda(J_i^h)\,\nu$. Then we have for some $A \in \mathcal{B}(\{1,\dots,k_h\})$ and $B \in \mathcal{B}(\mathbb{R})$, $\nu(B) < \infty$,

$$\int_{i\in A}\int_{\xi\in B} \iota_h(di,d\xi) \overset{\mathrm{d}}{=} \sum_{i=1}^{k_h} 1_A(i)\lambda(J_i^h)\int_B \nu^i(d\xi).$$

In particular, we have for all $\mathcal{V}_h$-valued $\mathcal{F}_t$-adapted and cádlág processes $u = \{u(t), 0 \le t < \infty\}$

$$\int_0^t \int_{\{1,\dots,k_h\}} \int_{\zeta\in\mathbb{R}} g_h(u(s-),di,\zeta)\ \tilde{\eta}_h(di,d\zeta,ds) \overset{\mathrm{d}}{=} \sum_{i=1}^{k_h}\int_0^t\int_{\zeta\in\mathbb{R}} g_h(u(s-),i,\zeta)\tilde{\eta}_h^i(d\zeta,ds).$$

2.2. Time discretization. One usually prefers to discretize simultaneously in time and space. Thus, let τ_h be the time step size corresponding to the subdivision $\mathcal{T}_h$. One popular way for time discretization is to use the *implicit Euler scheme*, i.e.,

$$\frac{u_h(t+\tau_h)-u_h(t)}{\tau_h} + \mathcal{A}_h u_h(t+\tau_h) \approx f_h(u_h(t)) + \int_t^{t+\tau_h}\int_{z\in\mathbb{R}^{k_h}} g_h(u_h(t);z)\tilde{\eta}_h(dz,dt).$$

So, if v_h^m denotes the approximation of $u_h(m\tau_h)$, then

$$\begin{cases} v_h^{m+1} = (1-\tau_h A_h)^{-1}\left[v_h^m + \tau_h f_h(m\tau_h, v_h^m) + \xi_h^m(u_h(m\tau_h))\right], \\ v_h^0 = P_h u_0, \end{cases} \tag{2.5}$$

where we denote by $\xi_h^m(u_h(m\tau_h))$ the k_h dimensional random variable

$$\int_{m\tau_h}^{(m+1)\tau_h}\int_{\zeta\in\mathbb{R}^{k_h}} g_h(u_h(m\tau_h);\zeta)\ \tilde{\eta}_h(dz;ds). \tag{2.6}$$

The approximation at time $t = \tau_h m$, $m \in \mathbb{N}$, is given by

$$\hat{u}_h(m\tau) := \mathcal{I}_h v_h^m. \tag{2.7}$$

The solution is defined by linear interpolation; let $t \in (m\tau_h, (m+1)\tau_h)$; we define $v_h(t)$ by $v_h(t) = v_h^m$ and

$$\hat{u}_h(t) := \mathcal{I}_h\, v_h(t), \quad m\tau_h < t < (m+1)\tau_h.$$

Since the implicit Euler scheme is unconditionally stable, no stability conditions arise. Now we can state our main result.

THEOREM 2.1. *Let $\mathcal{O}$ be a polyhedral domain and A be a uniformly elliptic, second order differential operator given in* (2.1) *with Dirichlet or Neumann boundary conditions.*

For a fixed p, $1 \le p \le 2$, let $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\ge 0}, P)$ be the usual filtration and

$$\eta : \mathcal{O} \times \mathbb{R} \times \mathcal{B}(\mathbb{R}^+) \to \mathbb{R}^+$$

be a Poisson random measure of jump type with intensity $\iota := \lambda \times \nu$, where ν belongs to the set of p-integrable Lévy measures on $\mathbb{R}$.

We assume that the functions $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \times \bar{\mathcal{O}} \times \mathbb{R} \to \mathbb{R}$ are globally Lipschitz and g linear in the third variable. Moreover, we assume that the initial condition u_0 belongs to $L^p(\mathcal{O})$.

The space is approximated by a nondegenerate family of subdivisions $\{\mathcal{T}_h, 0 < h \le 1\}$ of finite elements, satisfying the conditions described in section 2.1. *Moreover, the reference element is supposed to be a standard element for $l \ge 0$ and $m = 2$. Let v_n^k, $k \in \mathbb{N}$, be the approximation of* (2.2) *given by* (2.7). *Then for all $N \in \mathbb{N}$, for all α such that*

$$\alpha < \frac{2}{p} - d + \frac{d}{p},$$

and for all δ_1, δ_2, and δ_3, with

$$\delta_1 = 2 + \frac{d}{2} - \frac{d}{p}, \quad \delta_2 < \frac{2}{p} - d + \frac{d}{p}, \text{ and } \delta_3 < \left(\frac{2}{p} - d + \frac{d}{p}\right)\left(2 + \frac{d}{2} - \frac{d}{p}\right),$$

there exist constants C_0, C_1, C_2, and C_3 depending on p, f, and g such that for all $0 < h \le 1$ the following estimate holds for $0 < k \le N$:

$$\left[\mathbb{E}\ \left\|u(k\tau_h) - v_h^k\right\|_{L^p(\mathcal{O})}^p\right]^{\frac{1}{p}} \le \frac{C_0}{\epsilon}\,\tau^\alpha + \frac{C_1}{(k\tau_h)^{\frac{1}{p}}}\,h^{\delta_1} + \frac{C_2}{\epsilon_2}\,h^{\delta_2} + \frac{C_3}{\epsilon_3}\,h^{\delta_3},$$

where $\epsilon_1 = \frac{2}{p} - d + \frac{d}{p} - \alpha$, $\epsilon_2 = \frac{2}{p} - d + \frac{d}{p} - \delta_2$ and $\epsilon_3 = \frac{2}{p} - d + \frac{d}{p} - \frac{\delta_3}{\delta_1}$,

In [Hau05, Example 2.3] we treated (2.2) by means of $L^p(\mathcal{O})$ spaces. But finite elements are designed for Hilbert spaces. In particular, the approximation of the operator A is given by the inner product (see problem 3.4), and usually the error estimates of the corresponding semigroups are given in $L^2(\mathcal{O})$ or $H^1(\mathcal{O})$ (see Theorem 3.3). Since the $L^p(\mathcal{O})$ spaces were essential in [Hau05]—on which this paper relies—we had to express the error estimates of the corresponding semigroups in $L^p(\mathcal{O})$ spaces. We have done this in Theorem 3.4. Therefore, the entity $2 + \frac{d}{2} - \frac{d}{p}$ appears.

Fix $p \in [1,2]$. Let E be a separable M-type p Banach space and Z be a Banach space. Let A be an operator generating a compact and analytic semigroup, $\nu \in \mathcal{L}(Z)$, $\tilde{\eta}$ be a compensated Poisson random measure on Z with intensity ν, and $G : E \to L(Z,E)$ be a mapping such that there exist some $\gamma < \frac{2}{p}$ and $C < \infty$, with

$$\int_Z \left| (-A)^{-\frac{\gamma}{2}} \left(G(u,z) - G(v,z) \right) \right|_E^p \nu(dz) \leq C \, |u-v|_E^p. \tag{2.8}$$

Then, by [Hau05], there exists a unique solution to the equation

$$\begin{cases} du(t) = (Au(t-)) \; dt + \int_Z g(u(t-);z) \tilde{\eta}(dz;dt), \\ u(0) = u_0. \end{cases} \tag{2.9}$$

In [Hau05, Example 2.3], the space time Poissonian noise was represented as a functional-valued Poisson random measure taking values in the Besov space $B_p^{\frac{d}{p}-d}(\mathcal{O})$. Let $\nu \in \mathcal{L}(E)$ and $1 \leq p \leq 2$. Let $g : \mathbb{R} \times \mathcal{O} \times \mathbb{R} \to \mathbb{R}$ be Lipschitz continuous in the following sense: There exists some constant $C < \infty$ such that

$$\int_{\mathcal{O}} \int_{\mathbb{R}} |g(x,\xi,\zeta) - g(y,\xi,\zeta)|^p \nu(d\zeta) d\xi \leq C |x-y|^p, \quad x,y \in \mathbb{R}. \tag{2.10}$$

Thus, if

$$\frac{d}{p} - d < \frac{2}{p},$$

then the Nemytskij operator G of g, defined in (3.20), satisfies by Proposition A.1 the Lipschitz condition (2.8). The remaining difference

$$\frac{2}{p} - d + \frac{d}{p}$$

is used for estimating the error of the approximation. Therefore, the entity $\frac{2}{p} - d + \frac{d}{p}$ appears.

Again, let $\nu \in \mathcal{L}(E)$, $1 \leq p_0 \leq 2$ and $g : \mathbb{R} \times \mathcal{O} \times \mathbb{R} \to \mathbb{R}$ be Lipschitz continuous satisfying (2.10), with $p = p_0$. Moreover, let $f = 0$. Then, if

$$\frac{2}{p_0} - d + \frac{d}{p_0} > 0, \tag{2.11}$$

problem (1.1) has a unique solution such that $u \in L^p(\Omega; L^p(\mathcal{O}))$ for all $p \in [p_0, p_1)$, where $p_1 = \frac{2}{d} + 1$ (compare [Hau05, Example 2.3]). Moreover, the assumptions of Theorem 2.1 are fulfilled for all $p \in [p_0, p_1)$, and we can choose any $p \in [p_0, p_1)$ such that $\min(\delta_1, \delta_2, \delta_3)$ is maximal. If

$$2 + \frac{d}{2} - \frac{d}{p_1} < 0,$$

then no convergence is given by Theorem 2.1.

3. Finite elements—a short account. In the existing literature several error bounds for the approximation by finite elements can be found. In this section, we will cite the estimates which are necessary to prove our main result.

In this chapter, let $\{\mathcal{T}_h, 0 < h \leq 1\}$ be a nondegenerate family of subdivisions of a polyhedral domain $\mathcal{O}$ in $\mathbb{R}^d$, consisting of k_h finite elements of reference type $(K, \mathcal{S}, \mathcal{N})$ satisfying the standard conditions for some m and l. We assume that the parameter h corresponds to the maximal diameter of the finite elements, i.e., $h = \max\{1 \leq i \leq N \mid \operatorname{diam}(K_i)\}$. Moreover, we assume throughout this section, that $(K, \mathcal{P}, \mathcal{N})$ satisfies the standard conditions for some l and m.

Let $\{\phi_1^h, \dots, \phi_{k_h}^h\}$ be the set of shape functions and $\{N_1^h, \dots, N_k^h\}$ be the set of nodal variables. By $\bar{\mathcal{N}}$ we denote the set of nodal variables which are inner points. We denote by $\mathcal{V}_h$ the linear hull of the shape functions and by $\mathcal{V}_h'$ its dual.

The quality of approximation is given by the following Bramble–Hilbert (Raviart–Ciarlet) inequality.

THEOREM 3.1 (see, e.g., Brenner and Scott [BS02, Corollary 4.4.24]). *Assume that the conditions above are satisfied. Suppose* $1 \leq p \leq \infty$ *and either* $m - l - \frac{d}{p} > 0$ *when* $p > 1$ *or* $m - l - d \geq 0$ *when* $p = 1$. *Then there exists a positive constant* C *depending on the reference element,* m, l, d, p, *and the chunkiness parameter such that for* $0 \leq s \leq m$

$$\left(\sum_{T \in \mathcal{T}_h} \|v - \mathcal{I}_h v\|_{W_p^s(T)}^p \right)^{\frac{1}{p}} \leq C\, h^{m-s}\, \|v\|_{W_p^m(\mathcal{O})}$$

for all $v \in W_p^m(\mathcal{O})$, *where the left-hand side should be interpreted, in the case where* $p = \infty$, *as* $\max_{T \in \mathcal{T}_h} \|v - \mathcal{I}_h v\|_{W_\infty^s(\mathcal{O})}$. *For* $0 \leq s \leq l$,

$$\max_{T \in \mathcal{T}_h} \|v - \mathcal{I}_h v\|_{W_\infty^s(\mathcal{O})} \leq C\, h^{m-s-\frac{d}{p}}\, \|v\|_{W_p^m(\mathcal{O})} \quad \forall v \in W_p^m(\mathcal{O}).$$

Since the space $\mathcal{V}_h$ is finite-dimensional, all norms which can be defined in $\mathcal{V}_h$ are equivalent. The purpose of the inverse inequalities is to specify how the equivalent constants depend on the parameter h.

THEOREM 3.2 (Brenner and Scott [BS02, Theorem 4.5.11]). *Assume that the conditions above are satisfied, and let* p, q, *and* l *be given constants such that* $1 \leq p \leq \infty$, $1 \leq q \leq \infty$, $0 \leq m \leq l$, *and* $\phi_i^h \in W_p^l(\mathcal{O}) \cap W_q^m(\mathcal{O})$ *for* $i = 1, \dots, k_h$. *Then there exists a positive constant* C *depending on the reference element,* d, m, p, *and* q *such that*

$$\left[\sum_{T \in \mathcal{T}_h} \|v\|_{W_p^l(T)}^p \right]^{\frac{1}{p}} \leq C\, h^{m-l+\min(0, \frac{d}{p} - \frac{d}{q})} \left[\sum_{T \in \mathcal{T}_h} \|v\|_{W_q^m(T)}^q \right]^{\frac{1}{q}}$$

for all $v \in \mathcal{V}_h$. *When* $p = \infty$, *then* $[\sum_{T \in \mathcal{T}_h} \|v\|_{W_p^s(T)}^p]^{\frac{1}{p}}$ *is interpreted as* $\max_{T \in \mathcal{T}_h} \|v\|_{W_\infty^l(T)}^q$.

COROLLARY 3.1. *Under the conditions above, there exists a positive constant* C *depending on the reference element such that we have for all* $1 > \delta \geq 0$

$$\left[\sum_{T \in \mathcal{T}_h} \|v\|_{L_p(T)}^p \right]^{\frac{1}{p}} \leq C\, h^{-\delta+\min(0, \frac{d}{p} - \frac{d}{q})} \left[\sum_{T \in \mathcal{T}_h} \|v\|_{W_q^{-\delta}(T)}^q \right]^{\frac{1}{q}}$$

for all $v \in \mathcal{V}_h$.

Proof. The proof is done similarly to the proof of Theorem 3.2 (see Brenner and Scott [BS02, Theorem 4.5.11]). Let $(\hat{K}, \hat{\mathcal{P}}, \hat{\mathcal{N}})$ be a finite element of reference type satisfying the standard conditions for m and l with $\operatorname{diam}(\hat{K}) = 1$ and $(K, \mathcal{P}, \mathcal{N})$ be a finite element of the family of subdivisions with $\operatorname{diam}(K) = h_K$. Since the subdivision is nondegenerate, a transformation exists, having the form

$$T_K : \hat{K} \ni \hat{x} \mapsto J_K \hat{x} + b_K \in K,$$

where $J_K \in \mathbb{R}^d \times \mathbb{R}^d$ being invertible and $b_K \in \mathbb{R}^d$ (see Proposition 4.4.11 in [BS02]). Let $\hat{v} : \hat{K} \to \mathbb{R}$ be an arbitrary function and $v : K \to \mathbb{R}$ be defined by

$$K \ni x \mapsto v(x) := \hat{v}(J_K^{-1} x) \in \mathbb{R}.$$

Then substitution gives

$$|v|_{L^p(K)} = \det(J_K)^{\frac{1}{p}} |\hat{v}|_{L^p(\hat{K})} = C\, h_K^{\frac{d}{p}} |\hat{v}|_{L^p(\hat{K})}, \quad 0 < h_K \le 1,$$

and

$$|v|_{L^\infty(K)} = |\hat{v}|_{L^\infty(\hat{K})}, \quad 0 < h_K \le 1.$$

Similarly there exists a constant $C > 0$ such that we have for all $0 < h_K \le 1$

$$|v|_{W_p^j(K)} = C\, \det(J_K)^{\frac{1}{p}} h_K^{-j} |\hat{v}|_{W_p^j(\hat{K})} = C\, h_K^{\frac{d}{p}-j} |\hat{v}|_{W_p^j(\hat{K})}$$

and

$$|v|_{W_\infty^j(K)} = C\, h_K^{-j} |\hat{v}|_{W_\infty^j(\hat{K})} = C\, h_K^{-j} |\hat{v}|_{W_\infty^j(\hat{K})}.$$

Duality arguments give for q conjugate to p and for all $0 < h_K \le 1$

$$\begin{aligned} |v|_{W_p^{-j}(K)} &= \sup_{w \in W_q^j(K)} \frac{\langle v, w \rangle}{|w|_{W_q^j(K)}} \le C \sup_{\hat{w} \in W_q^j(\hat{K})} \frac{\det(J_K) \langle \hat{v}, \hat{w} \rangle}{h_K^{\frac{d}{q}-j} |\hat{w}|_{W_q^j(\hat{K})}} \\ &= C\, h_K^{d(1-\frac{1}{q})+j} \sup_{\hat{w} \in W_q^j(K)} \frac{\langle \hat{v}, \hat{w} \rangle}{|\hat{w}|_{W_q^j(K)}} \le C\, h_K^{\frac{d}{p}+j} |\hat{v}|_{W_p^{-j}(\hat{K})}. \end{aligned} \tag{3.1}$$

Since $\hat{\mathcal{P}}$ is finite-dimensional, all norms are equivalent, and there exists a constant $\hat{C}$ such that

$$|\hat{v}|_{L^p(\hat{K})} \le \hat{C}\, |\hat{v}|_{W_q^{-\gamma}(\hat{K})} \quad \forall v \in \mathcal{P}.$$

Therefore (3.1) and interpolation arguments imply for $\gamma > 0$ that

$$h_K^{-\frac{d}{p}} |v|_{L^p(K)} \le C\, h_K^{-\frac{d}{q}-\gamma} |v|_{W_q^{-\gamma}(K)} \quad \forall v \in \mathcal{P},\ 0 < h_K \le 1.$$

Let $\{\mathcal{T}_h, 0 < h \leq 1\}$ be a nondegenerate family of subdivisions. Then, by taking the sum, one can show by the same arguments as in Theorem 4.5.11 in [BS02], that, if the family of subdivisions is nondegenerate, the assertion of Corollary 3.1 holds. ∎

3.1. Approximation of the operator. We can associate with the operator $(A, D(A))$ the Gelfand triple (V, H, V'), where $V \hookrightarrow H = H' \hookrightarrow V'$ densely. In fact, $V = D(-A)^{\frac{1}{2}}$ and V' its dual. In our case, where $(A, D(A))$ is a second order differential operator with boundary conditions, we have $V = H^1(\mathcal{O})$ if the boundary condition involves first order derivatives and $V = H_0^1(\mathcal{O})$ if the homogeneous Dirichlet condition is specified on the whole boundary. Now the operator $A : V \to V'$ defines a bilinear form on H by

$$a(u, v) := \langle Au, v\rangle, \quad u, v \in V.$$

The corresponding symmetric variational problem is posed as follows. Suppose that the following three conditions are valid:

$$\text{(3.2)} \qquad \begin{cases} (1) & (H, \langle\cdot,\cdot\rangle) \text{ is a Hilbert space.} \\ (2) & V \text{ is a (closed) subspace of } H. \\ (3) & a(\cdot,\cdot) \text{ is a } \textit{bounded, symmetric} \text{ bilinear form} \\ & \text{that is } \textit{coercive} \text{ on } V. \end{cases}$$

Then the *symmetric variational* problem is the following:

$$\text{(3.3)} \qquad \begin{cases} \text{For a given } F \in V', \text{ find } u \in V \text{ such that} \\ \quad a(u, v) = F(v) \quad \forall v \in V. \end{cases}$$

Given the family of nondegenerate subdivisions $\mathcal{T}_h$ of a polyhedral domain $\mathcal{O}$, the *Ritz–Galerkin approximation* of problem (3.3) is the following:

$$\text{(3.4)} \qquad \begin{cases} \text{For a given } F \in \mathcal{V}_h', \text{ find } u_h \in \mathcal{V}_h \text{ such that} \\ \quad a(u_h, v) = F(v) \quad \forall v \in \mathcal{V}_h. \end{cases}$$

Assume that u solves problem (3.3) and u_h solves problem (3.4). Then we can find the following estimates on $u - u_h$ depending on F. By means of the inequality above and duality arguments, one can verify bounds of the difference $u - u_h$ in $H^1(\mathcal{O})$ and $L^2(\mathcal{O})$. In particular, one can show that, if the family of subdivisions $\{\mathcal{T}_h, 0 < h \leq 1\}$ satisfies the assumptions of Theorem 3.1 with $m = 2$, $l \geq 0$ and $p = 2$, then

$$\text{(3.5)} \qquad \inf_{v \in \mathcal{V}_h} \|u - v\|_{H^1(\mathcal{O})} \leq C\, h\, \|u\|_{H^2(\mathcal{O})} \quad \forall u \in H^2(\mathcal{O}), \quad 0 < h \leq 1,$$

is valid, and therefore (for more details, see Theorem 5.4.8 in [BS02])

$$\text{(3.6)} \qquad \|u - u_h\|_{L^2(\mathcal{O})} \leq Ch^2 \|u\|_{H^2(\mathcal{O})} \leq Ch^2 \|F\|_{L^2(\mathcal{O})}, \quad 0 < h \leq 1.$$

Finally, by using duality arguments, one can show that there exists a $h_0 > 0$ such that (see Theorem 5.8.3 in [BS02])

$$\text{(3.7)} \qquad \|u - u_h\|_{H^{2,-\gamma}(\mathcal{O})} \leq Ch^{\gamma+1} \|u - u_h\|_{H^1(\mathcal{O})} \leq Ch^{2+\gamma} \|F\|_{L^2(\mathcal{O})}, \quad 0 < h \leq 1.$$

Since our problem (2.2) is well-posed in $L^p(\mathcal{O})$, the error bounds of the difference have to be transferred to $L^p(\mathcal{O})$ or $W^p_{-\gamma}(\mathcal{O})$, $\gamma > d - \frac{d}{p}$, $p \in (1, 2]$.

COROLLARY 3.2. *Assume that* (3.5) *holds. Then, under the conditions above, there exist positive constants* C *and* h_0 *such that for all* $h \le h_0$ *there is a unique solution to problem* (3.4) *satisfying*

$$\|u - u_h\|_{L^p(\mathcal{O})} \le Ch^{2-\frac{d}{p}+\frac{d}{2}} \|F\|_{L^p(\mathcal{O})}, \quad 0 < h \le 1,$$

and for $\gamma \in (0,1)$

$$\|u - u_h\|_{W^p_{-\gamma}(\mathcal{O})} \le Ch^{2-\frac{d}{p}+\frac{d}{2}} \|F\|_{W^p_{-\gamma}(\mathcal{O})}, \quad 0 < h \le 1.$$

Proof. Since $p \le 2$ and O is bounded we can infer that

$$\|u - u_h\|_{L^p(\mathcal{O})} \le C \ \|u - u_h\|_{L^2(\mathcal{O})}.$$

The estimate (3.6) gives

$$\|u - u_h\|_{L^p(\mathcal{O})} \le C\, h^2 \|F\|_{L^2(\mathcal{O})}.$$

Theorem 3.2 leads to

$$\|u - u_h\|_{L^p(\mathcal{O})} \le C\, h^{2+\frac{d}{2}-\frac{d}{p}} \|F\|_{L^p(\mathcal{O})}.$$

Similarly, we get

$$\|u - u_h\|_{W^p_{-\gamma}(\mathcal{O})} \le C \ \|u - u_h\|_{W^2_{-\gamma}(\mathcal{O})}.$$

Estimate (3.7) gives

$$\|u - u_h\|_{W^p_{-\gamma}(\mathcal{O})} \le C\, h^{2+\gamma} \|F\|_{L^2(\mathcal{O})}.$$

Next, we apply negative norm estimates. In particular, Theorem 3.2 leads to

$$\|u - u_h\|_{W^p_{-\gamma}(\mathcal{O})} \le C\, h^{2+\frac{d}{2}-\frac{d}{p}} \|F\|_{W^p_{-\gamma}(\mathcal{O})}. \qquad \square$$

3.2. Approximation of the semigroup. The approximation $\mathcal{A}_h : \mathcal{V}_h \to \mathcal{V}_h'$ of the operator A is given by the induced bilinear form on $\mathcal{V}_h$. In particular, for any $u \in \mathcal{V}_h$ we define $\mathcal{A}_h u$ by

$$\langle \mathcal{A}_h u, v\rangle := a(u,v), \quad v \in \mathcal{V}_h. \tag{3.8}$$

The semidiscrete approximation of the solution u to the abstract Cauchy problem

$$\begin{cases} \frac{d}{dt}u(t) + Au(t) = 0, & t > 0, \\ u(0) = u_0 \end{cases} \tag{3.9}$$

is the function $u_h : [0,\infty) \to \mathcal{V}_h$, determined by the following condition: For all $v \in \mathcal{V}_h$ we have

$$\begin{cases} \frac{d}{dt}\langle u_h(t), v\rangle + a(u_h(t), v) = 0, \ t > 0, \\ \langle u_h(0), v\rangle = \langle u_0, v\rangle. \end{cases} \tag{3.10}$$

In use of the operator $\mathcal{A}_h$, relation (3.10) is written as

$$\frac{du_h(t)}{dt} + \mathcal{A}_h u_h(t) = 0, \quad t > 0.$$

Let $\mathcal{H}_h$ be the space $\mathcal{V}_h$ equipped with the topology induced from H. Let $\mathcal{R}_h : H \to \mathcal{H}_h$ be the Ritz projection. Recall that the linear operator $\mathcal{A}_h : \mathcal{H}_h \to \mathcal{H}_h$ is defined by (3.8). Let $\mathcal{R}_h : V \to \mathcal{V}_h$ be the *Ritz operator* defined through

$$a(R_h u, v_h) = a(u, v) \quad \forall u \in \mathcal{V}_h \text{ and } u \in V.$$

By the Lax–Milgram theorem, it can be shown that the Ritz operator is well-defined, i.e., that the element $R_h u$ exists and is unique. Theorem 1.1 of [FSS01] shows that the operator $\mathcal{A}_h : \mathcal{H}_h \to \mathcal{H}_h$ generates an analytic semigroup on $\mathcal{H}_h$ independently of h. To be precise, there exist some $\theta \in (0, \pi)$ and some $M \geq 1$ such that $\mathbb{C} \setminus \Sigma_\theta \subset \rho(\mathcal{A}_h)$, where $\rho(\mathcal{A}_h)$ denotes the resolvent set of $\mathcal{A}_h$ and

$$\Sigma_\theta := \{z \in \mathcal{C}, 0 \leq |\arg(z)| \leq \theta\}.$$

The semidiscrete approximation admits the error estimate of optimal order for $t > 0$.

THEOREM 3.3 ([FSS01, Theorem 2.1]). *Given $u_0 \in H$, let u and u_h be the solutions of* (3.9) *and* (3.10)*, respectively. Then we have*

$$\|u(t) - u_h(t)\|_{L^2(\mathcal{O})} + h \|u(t) - u_h(t)\|_{H^1(\mathcal{O})} \leq C\, h^2 t^{-1} \|u_0\|_{L^2(\mathcal{O})}, \quad t > 0,\ 0 < h \leq 1.$$

Similarly to the Proof of Theorem 3.3 one can find error estimates in L^p spaces.

THEOREM 3.4. *Given $u_0 \in H$, let u and u_h be the solutions of* (3.9) *and* (3.10)*, respectively. Then we have for $1 \leq p \leq 2$ and $0 \leq \gamma < 1$*

$$\|u(t) - u_h(t)\|_{W^{p,-\gamma}(\mathcal{O})} \leq C\, h^{2-\frac{d}{p}+\frac{d}{2}} t^{-1} \|u_0\|_{W^{p,-\gamma}(\mathcal{O})}, \quad t > 0,\ 0 < h \leq 1.$$

Proof. The proof of Theorem 3.4 is similar to the proof of Theorem 3.3, i.e., Theorem 2.1 in [FSS01]. Assume that u solves problem (3.3) with $F = u_0$ and u_h solves problem (3.4) with $F = \mathcal{I}_h u_0$. In Corollary 3.2 we have shown that there exists estimates of the error in L^p spaces. In particular, we have

$$\|u - u_h\|_{L^p(\mathcal{O})} \leq C h^{2-\frac{d}{p}+\frac{d}{2}} \|F\|_{L^p(\mathcal{O})}.$$

Since $u = A^{-1}u_0$ and $\mathcal{A}_h^{-1}\mathcal{I}_h u_0$, we can write

$$\left\|A^{-1}u_0 - \mathcal{A}_h^{-1}\mathcal{I}_h u_0\right\|_{L^p(\mathcal{O})} \leq C h^{2-\frac{d}{p}+\frac{d}{2}} \|u_0\|_{L^p(\mathcal{O})}.$$

Next, since for all $\lambda_0, \lambda \in \rho(A) \cap \rho(\mathcal{A}_h)$ we have

$$\begin{aligned}(\lambda\mathcal{I}_h - \mathcal{A}_h)^{-1} - (\lambda - A)^{-1} &= \left(1 + (\lambda - \lambda_0)(\lambda\mathcal{I}_h - \mathcal{A}_h)^{-1}\mathcal{I}_h\right) \\ &\times \left((\lambda_0\mathcal{I}_h - \mathcal{A}_h)^{-1}\mathcal{I}_h - (\lambda_0 - A)^{-1}\right) \times (\lambda_0 - A)(\lambda_0 - A)^{-1}, \end{aligned} \tag{3.11}$$

we can conclude that the following estimates:

$$\left\|(\lambda I - A)^{-1} u_0 - (\lambda\mathcal{I}_h - \mathcal{A}_h)^{-1}\mathcal{I}_h u_0\right\|_{L^p(\mathcal{O})} \leq C h^{2-\frac{d}{p}+\frac{d}{2}} \|u_0\|_{L^p(\mathcal{O})} \tag{3.12}$$

and for $\gamma \in (0, 1)$

$$\left\|(\lambda I - A)^{-1} u_0 - (\lambda\mathcal{I}_h - \mathcal{A}_h)^{-1}\mathcal{I}_h u_0\right\|_{W^p_{-\gamma}(\mathcal{O})} \leq C h^{2-\frac{d}{p}+\frac{d}{2}} \|u_0\|_{W^p_{-\gamma}(\mathcal{O})} \tag{3.13}$$

are valid for all $\lambda \in \rho(A) \cap \rho(\mathcal{A}_h)$. The proof of Theorem 3.3 relies on the fact that we can represent the semigroup in terms of the resolvent. By comparing to [FSS01, page 58], or to [Paz83, Theorem 1.7.7], we have

$$\|u(t) - u_h(t)\mathcal{I}_h\| \leq \frac{1}{2\pi} \int_\Gamma \left|e^{-tz}\right| \cdot \left\| \left[(\lambda I - A)^{-1} - (\lambda \mathcal{I}_h - \mathcal{A}_h)^{-1}\mathcal{I}_h\right] u_0 \right\| \, dz, \tag{3.14}$$

where Γ is the boundary of $\Sigma_{\theta_1} = \{z, 0 \leq |\arg(z)| \leq \theta_1\}$, for $\theta_1 \in (\theta, \frac{\pi}{2})$ coming from $\infty e^{i\theta_1}$, passing through the origin and going out to $\infty e^{-i\theta_1}$. Using estimate (3.6) and taking into account the calculations in (3.11) gives

$$\begin{aligned} &\left\| \left[(\lambda I - A)^{-1} - (\lambda \mathcal{I}_h - \mathcal{A}_h)^{-1}\mathcal{I}_h\right] u_0 \right\|_{L^2(\mathcal{O})} \\ &\qquad \leq C\, h^m \left\|u_0\right\|_{L^2(\mathcal{O})}, \quad u_0 \in L^2(\mathcal{O}). \end{aligned} \tag{3.15}$$

By substituting estimate (3.15) into (3.14), Theorem 3.3 can be shown. Similarly, by substituting estimate (3.12) into (3.14) one gets

$$\begin{aligned} &\|u(t) - u_h(t)\mathcal{I}_h\|_{L^p(\mathcal{O})} \leq \frac{C}{2\pi}\, h^{2-\frac{d}{p}+\frac{d}{2}}\ \|u_0\|_{L^p(\mathcal{O})} \int_\Gamma \left|e^{-tz}\right| \, dz \\ &\leq C\, h^{2-\frac{d}{p}+\frac{d}{2}}\ \|u_0\|_{L^p(\mathcal{O})} \int_0^\infty e^{-tr\cos(\theta_1)}\, dr \leq C\, h^{2-\frac{d}{p}+\frac{d}{2}}\ t^{-1}\ \|u_0\|_{L^p(\mathcal{O})}. \end{aligned}$$

The case $\gamma = 0$ of Theorem 3.4 follows. To show Theorem 3.4 for $\gamma > 0$ one can use the same computations, but one has to substitute estimate (3.13) instead of (3.12) into (3.14). □

3.3. Mesh-dependent norms. For $x \in \mathcal{V}_h$, one can define the mesh-dependent inner product

$$\langle v, w\rangle_h := h^d \sum_{i=1}^{k_h} N_i^h(v) N_i^h(w)$$

and the norm

$$\|v\|^2_{L^2_h(\mathcal{O})} := \langle v, v\rangle = h^d \sum_{i=1}^{k_h} v(N_i^h)^2.$$

It can be shown that there exist some constants c and C such that we have for all $h \in (0, 1)$ (see Lemma 6.2.7 of [BS02])

$$c\|v\|_{L^2(\mathcal{O})} \leq \|v\|_{L^2_h(\mathcal{O})} \leq C\|v\|_{L^2(\mathcal{O})} \quad \forall v \in \mathcal{V}_h.$$

Since the natural spaces for SPDEs driven by Poisson random measure are L^p spaces, we will also define mesh-dependent L^p spaces; i.e., for $0 < h \leq 1$ and $v \in \mathcal{V}_h$ we define

$$\|v\|_{L^p_h(\mathcal{O})} := \left(h^d \sum_{i=1}^{k_h} \left|N_i^h(v)\right|^p \right)^{\frac{1}{p}}$$

and

$$L^p_h(\mathcal{O}) := \{v \in \mathcal{V}_h, \|v\|_{L^p_h(\mathcal{O})} < \infty\}.$$

Similarly to the case where $p = 2$, one can show that there exist some constants c and C such that we have for all h with $0 < h \leq 1$

$$c\|v\|_{L^p(\mathcal{O})} \leq \|v\|_{L^p_h(\mathcal{O})} \leq C\|v\|_{L^p(\mathcal{O})} \quad \forall v \in \mathcal{V}_h \cap L^p(\mathcal{O}). \tag{3.16}$$

Remark 3.1. It follows from the nondegeneracy assumptions of the subdivisions that there exist some constants c and C such that

$$c\, h^d \leq k_h \leq Ch^d, \quad 0 < h < 1.$$

Theorem 1.1 of [FSS01] shows that the operator $\mathcal{A}_h : \mathcal{H}_h \to \mathcal{H}_h$ generates an analytic semigroup on $\mathcal{H}_h$ with parameter M and ω independent of h. To be precise, there exist some $\theta \in (0, \pi)$ and some $M \geq 1$ such that $\mathcal{C} \setminus \Sigma_\theta \subset \rho(\mathcal{A}_h)$, where $\rho(\mathcal{A}_h)$ denotes the resolvent set of operator $\mathcal{A}_h$ and

$$\Sigma_\theta := \{z \in \mathcal{C}, 0 \leq |\arg(z)| \leq \theta\}.$$

Let X be a Banach space and $A : X \to X$ be a operator generating an analytic semigroup on X. Since the constants M, ω, and C in

$$\left\|e^{-tA}u\right\|_X \leq M\, e^{\omega t}\, \|u\|_X$$

and

$$\left\|Ae^{-tA}u\right\|_X \leq \frac{C}{t}\, e^{\omega t}\, \|u\|_X$$

depend on the resolvent set, in particular, on the angle θ which includes the numerical range, we can conclude that there exist some constants C_1 and C_2 such that

$$\left\|e^{-t\mathcal{A}_h}u\right\|_{L^p_h(\mathcal{O})} \leq C_1\, \|u\|_{L^p_h(\mathcal{O})} \quad \forall v \in L^p_h(\mathcal{O}), \quad 0 < h \leq 1, \tag{3.17}$$

and

$$\left\|Ae^{-t\mathcal{A}_h}u\right\|_{L^p_h(\mathcal{O})} \leq \frac{C_2}{t}\, \|u\|_{L^p_h(\mathcal{O})} \quad \forall v \in L^p_h(\mathcal{O}), \quad 0 < h \leq 1. \tag{3.18}$$

For more details, see [Paz83, Theorem 2.5.2] and [FSS01].

3.4. The mass and stiffness matrix. In this section we will investigate properties of the mass and stiffness matrix. Let $\mathcal{O}$ be a polyhedron in $\mathbb{R}^d$ and $\{\mathcal{T}_h, 0 < h \leq 1\}$ be a nondegenerate family of subdivisions, where the reference finite element $(\hat{K}, \hat{\mathcal{P}}, \hat{\mathcal{N}})$ satisfies the standard conditions for $m = l = 1$.

The mass matrix $\mathbf{M}_h$ of a subdivision is given by $\mathbf{M}_h = ((m^h_{i,j}))^{k_h}_{i,j=1}$, where

$$m^h_{i,j} := \langle \phi^h_i, \phi^h_j \rangle, \quad 1 \leq i, j \leq k_h.$$

Since the shape functions are locally defined, there exists a number m such that for all $h \in (0, 1)$ there exists a set $\mathcal{C}_h \subset \mathbb{N}^{k_h} \times \mathbb{N}^{k_h}$ such that

$$\operatorname{supp}(\phi^h_i) \cap \operatorname{supp}(\phi^h_j) = \emptyset \quad \forall (i, j) \not\subset \mathcal{C}_h, \quad 0 < h \leq 1,$$

and

$$|(i, \mathbb{N}^{k_h}) \cap \mathcal{C}_h| \leq m, \quad 0 < h \leq 1.$$

Remark 3.2. The mass matrix is given by Theorem 9.8, page 387 in Ern and Guermond [EG04]. The left-hand side is given by

$$\langle \mathcal{A}_h \phi_l^h, \phi_j^h \rangle = \langle \nabla \phi_i^h, \nabla \phi_j^h \rangle.$$

Since $\{\mathcal{T}_h, 0 < h \leq 1\}$ is a uniformly bounded family of subdivisions, we have

$$\langle \nabla \phi_l^h, \nabla \phi_j^h \rangle = \begin{cases} c\, h^{d-2}, & \text{if } (i,j) \subset \mathcal{C}_h, \\ 0, & \text{elsewhere.} \end{cases}$$

Fix some i, $1 \leq i \leq k_h$. Our point of interest is $\mathcal{A}_h^{-1}\phi_i$. Let $u_i = \mathcal{A}_h^{-1}\phi_i$. Then

$$\langle \mathcal{A}_h u_i, \phi_j^h \rangle = \langle \phi_i^h, \phi_j^h \rangle, \quad j = 1, \ldots, k_h.$$

Since $u_i \in \mathcal{V}_h$, there exists $u_{i1}, \ldots, u_{ik_h} \in \mathbb{R}$ such that

$$u_i = \sum_{l=1}^{k_h} u_{il} \phi_l^h.$$

Therefore,

$$\langle \mathcal{A}_h u_i, \phi_j^h \rangle = \sum_{l=1}^{k_h} u_{il} \langle \mathcal{A}_h \phi_l^h, \phi_j^h \rangle = \langle \phi_i^h, \phi_j^h \rangle$$

and by Remark 3.2

$$u_{il} = \begin{cases} c\, h^2, & \text{if } (i,j) \subset C, \\ 0, & \text{elsewhere.} \end{cases} \tag{3.19}$$

3.5. Approximation of the noise. The space time Poissonian noise of jump type is given by a Poisson random measure

$$\eta : \mathcal{B}(\mathcal{O}) \times \mathcal{B}(\mathbb{R}) \times \mathcal{B}([0,\infty)) \to \mathbb{R}^+$$

with intensity $\lambda \times \nu$, where ν is a Lévy measure over $\mathbb{R}$, such that

$$\int_{\mathbb{R}} |z|^p \, \nu(dz) < \infty.$$

The function

$$g : \mathbb{R} \times \mathcal{O} \times \mathbb{R} \to \mathbb{R}$$

is Lipschitz continuous with Nemytskij operator

$$L^p(\mathcal{O}) \times \mathbb{R} \ni (u,z) \mapsto G(u;z) = g(u(\cdot);\cdot;z) \in L^p(\mathcal{O}), \tag{3.20}$$

where $g(u(\cdot);\cdot;z)(\xi) := g(u(\xi);\xi;z)$. In [Hau05] we have seen that the space time Poissonian noise can be described by a Banach-valued Poisson random measure. In particular, let $Z = B_{p,\infty}^{\frac{d}{p}-d}(\mathcal{O})$ and U be the unit sphere in Z. Let χ be defined by

$$\mathcal{O} \ni \xi \mapsto \chi(\xi) := c\, \delta_\xi \in Z,$$

where $c = |\delta_\xi|_Z^{-1}$. The intensity ι of the Z-valued Poisson random measure is given by

$$\mathcal{B}(Z) \ni B \mapsto \iota(B) := \int_{\zeta\in\mathbb{R}} \int_{x\in U} 1_B(\zeta x)\sigma(dx)\nu(d\zeta) \in \mathbb{R}^+,$$

where $\sigma(A) := \lambda(\chi^{-1}(A \cap \chi(\mathcal{O})))$, $A \in \mathcal{B}(Z)$. By inserting the definition above into the definition of the Besov space, one gets by Proposition A.1

$$\begin{aligned}\int_{z\in Z} |G(u,z) - G(v,z)|_Z^p \iota(dz) &= \int_{\xi\in\mathcal{O}} \int_{\zeta\in\mathbb{R}} |g(u(\xi),\xi;\zeta) - g(v(\xi),\xi;\zeta)|^p \,\nu(d\zeta)\,d\xi \\ &\le C\,|u-v|^p_{L^p(\mathcal{O})}, \quad u,v \in C_b^0(\mathcal{O}).\end{aligned}$$

Extension to $L^p(\mathcal{O})$ leads to the Lipschitz property (2.10) of G. The approximation can be similarly interpreted. First, the Lebesgue measure is approximated by λ_h defined by

$$\mathcal{B}\left(\{1,2,\dots,k_h\}\right) \ni A \mapsto \lambda_h(A) := \sum_{i=1}^{k_h} \lambda(J_i^h) 1_A(i) \in \mathbb{R}^+.$$

Now the approximation $\iota_h : \mathcal{B}(\{1,\dots,k_h\}) \times \mathcal{B}(\mathbb{R}) \to \mathbb{R}^+$ of the intensity ι is defined by

$$\{1,\dots,k_h\} \times \mathcal{B}(\mathbb{R}) \ni (A,B) \mapsto \iota_h(A,B) := \sum_{i=1}^{k_h} 1_A(i)\lambda_h(\{i\}) \int_{\zeta\in\mathbb{R}} \nu(d\zeta) \in \mathbb{R}^+.$$

The approximated Nemytiskij operator G_h is defined by

$$\mathcal{V}_h \times \{1,\dots,k_h\} \times \mathcal{B}(\mathbb{R}) \ni (u,i,\zeta) \mapsto G_h(u,i,\zeta) := g(u(N_i^h), N_i^h;\zeta)\, c_i^h \phi_i^h.$$

Similarly to the above we can show the following proposition.

PROPOSITION 3.1. *There exists some $C > 0$ such that for any $h > 0$ and for any $\gamma > \frac{1}{2}(d - \frac{d}{p})$ and $\delta < 2\gamma - d - \frac{d}{p}$ the following inequality holds for all $u_h, v_h \in \mathcal{V}_h$:*

$$\int_{z\in Z_h} \left\|(-\mathcal{A}_h)^{-\gamma}\left[G_h(u_h,z) - G_h(v_h,z)\right]\right\|^p_{L^p(\mathcal{O})} \iota_h(dz) \le C\, h^{d\delta p} \left\|u_h - v_h\right\|^p_{L^p(\mathcal{O})}.$$

Proof. Let $u, v \in \mathcal{V}_h$. Then we have

$$\begin{aligned}&\int_{z\in Z} \left\|(-\mathcal{A}_h)^{-\gamma}\left[G_h(u_h,z) - G_h(v_h,z)\right]\right\|^p_{L^p(\mathcal{O})} \iota_h(dz) \\ &= \mathbb{E}\left\|\sum_{i=1}^{k_h} \int_{\mathbb{R}} (-\mathcal{A}_h)^{-\gamma}\left[G_h(u_h,\zeta) - G_h(v_h,\zeta)\right] \mu_i^h(d\zeta,ds)\right\|^p_{L^p(\mathcal{O})},\end{aligned}$$

where μ_i^h, $i = 1,\dots,k_h$, are independent Poisson random measures with intensity $\lambda_h(i)\nu$. The generalized Burkholder inequality gives

$$\begin{aligned}&\int_{z\in Z} \left\|(-\mathcal{A}_h)^{-\gamma}\left[G_h(u_h,z) - G_h(v_h,z)\right]\right\|^p_{L^p(\mathcal{O})} \iota_h(dz) \\ &\le \sum_{j=1}^{k_h} \int_{\mathbb{R}} \left\|(-\mathcal{A}_h)^{-\gamma}\left[g(u_h(N_j^h),N_j^h;\zeta) - g(v_h(N_j^h),N_j^h;\zeta)\right] c_j^h\phi_j^h\right\|^p_{L^p(\mathcal{O})} \lambda(J_j^h)\nu(d\zeta).\end{aligned}$$

Since $g(u_h(N_j^h), N_j^h; \zeta) - g(v_h(N_j^h), N_j^h; \zeta) \in \mathbb{R}$ and $\mathcal{A}_h$ is linear, we can continue

$$\int_{z\in Z_h} \left\|(-\mathcal{A}_h)^{-\gamma}\left[G_h(u_h,z) - G_h(v_h,z)\right]\right\|_{L^p(\mathcal{O})}^p \iota_h(dz) \le \sum_{j=1}^{k_h} \int_{\mathbb{R}}$$

$$\left\|\left[g(u_h(N_j^h), N_j^h; \zeta) - g(v_h(N_j^h), N_j^h; \zeta)\right](-\mathcal{A}_h)^{-\gamma} c_j^h \phi_j^h\right\|_{L^p(\mathcal{O})}^p \lambda(J_j^h)\nu(d\zeta).$$

Since the family of subdivisions is nondegenerate, we have $\left\|\phi_j^h\right\|_{L^p(\mathcal{O})}^p \sim h^{\frac{d}{p}}$, $c_j^h \sim h^{-d}$, and $\lambda(J_j^h) \sim h^d$. Therefore, by using also Remark 3.2, we get

$$\int_{z\in Z_h} \left\|(-\mathcal{A}_h)^{-\gamma}\left[G_h(u_h,z) - G_h(v_h,z)\right]\right\|_{L^p(\mathcal{O})}^p \iota_h(dz)$$

$$\le C \sum_{j=1}^{k_h} \int_{\mathbb{R}} \left|g(u_h(N_j^h), N_j^h; \zeta) - g(v_h(N_j^h), N_j^h; \zeta)\right|^p \left\|(-\mathcal{A}_h)^{-\gamma} c_j^h \phi_j^h\right\|_{L^p(\mathcal{O})}^p \lambda(J_j^h)\nu(d\zeta)$$

$$\le C\, h^{\gamma 2} \sum_{j=1}^{k_h} \int_{\mathbb{R}} \left|g(u_h(N_j^h), N_j^h; \zeta) - g(v_h(N_j^h), N_j^h; \zeta)\right|^p \left\|c_j^h \phi_j^h\right\|_{L^p(\mathcal{O})}^p \lambda(J_j^h)\nu(d\zeta)$$

$$\le C\, h^{\gamma 2 - dp + d} \sum_{j=1}^{k_h} \int_{\mathbb{R}} \left|g(u_h(N_j^h), N_j^h; \zeta) - g(v_h(N_j^h), N_j^h; \zeta)\right|^p \lambda(J_j^h)\nu(d\zeta).$$

According to the Lipschitz condition of g we have

$$\int_{z\in Z} \left\|(-\mathcal{A}_h)^{-\gamma}\left[G_h(u_h,z) - G_h(v_h,z)\right]\right\|_{L^p(\mathcal{O})}^p \iota_h(dz)$$

$$\le C\, h^{\gamma 2 - dp + d} \sum_{j=1}^{k_h} h^d \left|u_h(N_j^h) - v_h(N_j^h)\right|^p.$$

The equivalence between $L^p(\mathcal{O})$ and $L_h^p(\mathcal{O})$ gives the assertion. ∎

In order to calculate the error induced by the approximation of the Poisson random measure, we have to express the measure ι_h in terms of ι.

PROPOSITION 3.2. *There exists some $C > 0$ such that for any $h > 0$ and for any $\gamma > \frac{1}{2}(d - \frac{d}{p})$ and $\delta < 2\gamma - d - \frac{d}{p}$ the following inequality holds for all $u_h \in \mathcal{V}_h$:*

$$\int_Z \left\|(-\mathcal{A}_h)^{-\gamma}\mathcal{I}_h G(u,z)\right\|_{L^p(\mathcal{O})}^p (\iota - \iota_h)(dz) \le C\, h^{d\delta p}\left(1 + \|u_h\|_{L^p(\mathcal{O})}^p\right).$$

Remark 3.3. In our calculation we use the fact that for $\gamma > d - \frac{d}{p}$ in $W^{p,-\gamma}(\mathcal{O})$ the noise vanishes for $h \to 0$. But, by plugging the difference $\iota - \iota_h$ into the definition of Z, one can show directly that the approximated noise converges to the functional-valued representation of the space time Poissonian noise in Z.

Proof. Embedding Z_h into Z by $i \mapsto c_i^h \phi_i^h$ leads to

$$\begin{aligned}
&\int_Z \left\|(-\mathcal{A}_h)^{-\gamma}\mathcal{I}_h G(u,z)\right\|^p_{L^p(\mathcal{O})}\ (\iota-\iota_h)(dz) \le \int_{\xi\in\mathcal{O}}\int_{\zeta\in\mathbb{R}} \\
&\quad \left\|(-\mathcal{A}_h)^{-\gamma}\left[\mathcal{I}_h G(u,z)\delta_\xi - \sum_{j=1}^{k_h} 1_{J_j^h}(\xi) g(u(N_j^h),N_j^h,\zeta)c_j^h\phi_j^h\right]\right\|^p_{L^p(\mathcal{O})} d\zeta\ d\xi \\
&\le \sum_{i=1}^{k_h}\int_{\xi\in J_i^h}\int_{\zeta\in\mathbb{R}} \left\|(-\mathcal{A}_h)^{-\gamma}\left[\sum_{l=1}^{k_h} g(u(N_l^h),N_l^h,\zeta)\phi_l^h\delta_\xi\right.\right. \\
&\qquad\qquad \left.\left. - \sum_{j=1}^{k_h} 1_{J_j^h}(\xi) g(u(N_j^h),N_j^h,\zeta)c_j^h\phi_j^h\right]\right\|^p_{L^p(\mathcal{O})} d\zeta\ d\xi \\
&\le \sum_{j=1}^{k_h}\int_{\xi\in J_j^h}\int_{\zeta\in\mathbb{R}} \left\|(-\mathcal{A}_h)^{-\gamma}\left[\sum_{l=1}^{k_h} g(u(N_l^h),N_l^h,\zeta)\phi_l^h\delta_\xi\right]\right\|^p_{L^p(\mathcal{O})} d\zeta\ d\xi \\
&+ \sum_{j=1}^{k_h}\int_{\xi\in J_j^h}\int_{\zeta\in\mathbb{R}} \left\|(-\mathcal{A}_h)^{-\gamma}\left[g(u(N_j^h),N_j^h,\zeta)c_j^h\phi_j^h\right]\right\|^p_{L^p(\mathcal{O})}\ d\zeta\ d\xi \\
&=: S_1 + S_2.
\end{aligned}$$

By Corollary 3.1 we obtain

$$S_1 \le h^{-\delta p}\sum_{j=1}^{k_h}\int_{\xi\in J_j^h}\int_{\zeta\in\mathbb{R}} \left\|(-\mathcal{A}_h)^{-\gamma}\left[\sum_{l=1}^{k_h} g(u(N_l^h),N_l^h,\zeta)\phi_l^h\delta_\xi\right]\right\|^p_{Z} d\zeta\ d\xi.$$

Inequality (A.1) leads to

$$S_1 \le C\ h^{-\delta p}\sum_{j=1}^{k_h}\int_{\zeta\in\mathbb{R}} \left\|(-\mathcal{A}_h)^{-\gamma}\left[\sum_{l=1}^{k_h} g(u(N_l^h),N_l^h,\zeta)\phi_l^h\right]\right\|^p_{L^p(J_j^h)} d\zeta.$$

Since there exists a constant m such that $|\{i \mid \sup\phi_i^h \cap J_j^h \neq \emptyset\}| \le m$ for all $i,j = 1,\dots,k_h$ and $h>0$, and $\sum_{i=1}^{k_h}\phi_i^h(\xi) = 1$, we can infer by Remark 3.2 that

$$\begin{aligned}
S_1 &\le C\ h^{(2\gamma-\delta)p}\sum_{j=1}^{k_h}\int_{\zeta\in\mathbb{R}} \left\|\sum_{l,\sup\phi_l^h\cap J_j^h\neq\emptyset} g(u(N_l^h),N_l^h,\zeta)\phi_l^h\right\|^p_{L^p(J_j^h)} d\zeta \\
&\le C\ h^{(2\gamma-\delta)p}\ m\ \sum_{j=1}^{k_h}\int_{\zeta\in\mathbb{R}} \left\|g(u(N_j^h),N_j^h,\zeta)\phi_j^h\right\|^p_{L^p(\mathcal{O})}\ d\zeta.
\end{aligned}$$

By the linearity of g and the equivalence of $L_h^p(\mathcal{O})$ and $L^p(\mathcal{O})$ we get

$$S_1 \le C\ h^{(2\gamma-\delta)p} m\ \left(1 + \sum_{j=1}^{k_h} h^d\ \left|u(N_j^h)\right|^p\right) \le h^{(2\gamma-\delta)p}\ m\ \left(1 + |u|^p_{L^p(\mathcal{O})}\right).$$

By applying Remark 3.19 we get for S_2

$$\begin{aligned}
S_2 &\le \sum_{j=1}^{k_h} \lambda(J_j^h) \int_{\zeta\in\mathbb{R}} \left\|(-\mathcal{A}_h)^{-\gamma} g(u(N_j^h), N_j^h, \zeta) c_j^h \phi_j^h\right\|_{L^p(\mathcal{O})}^p \, d\zeta \\
&\le C\, h^{2\gamma p} \sum_{j=1}^{k_h} \lambda(J_j^h) \int_{\zeta\in\mathbb{R}} \left\|g(u(N_j^h), N_j^h, \zeta) c_j^h \phi_j^h\right\|_{L^p(\mathcal{O})}^p \, d\zeta \\
&\le C\, h^{2\gamma p - dp} \sum_{j=1}^{k_h} \lambda(J_j^h) h^d \int_{\zeta\in\mathbb{R}} \left|g(u(N_j^h), N_j^h, \zeta)\right| \left\|\phi_j^h\right\|_{L^p(\mathcal{O})}^p \, d\zeta \\
&\le C\, h^{2\gamma p - dp + d} \sum_{j=1}^{k_h} \lambda(J_j^h) h^d \int_{\zeta\in\mathbb{R}} \left|g(u(N_j^h), N_j^h, \zeta)\right| \, d\zeta \,.
\end{aligned}$$

The linearity of g and the equivalence of $L_h^p(\mathcal{O})$ and $L^p(\mathcal{O})$ give

$$S_2 \le C\, h^{2\gamma p - dp + d} \left(1 + \|u\|_{L^p(\mathcal{O})}^p\right). \qquad \square$$

4. Stability of the finite element approximation. In this section we shall investigate the stability of the approximation by finite elements. Since the semidiscrete problem (2.3) is finite-dimensional, for all $0 < h \le 1$ there exists a unique solution $u_h = \{u_h(t) \in \mathbb{R}^{k_h}, 0 \le t < \infty\}$ (see, e.g., Theorem 8 of [Pro04]). In this section we show that the family of unique solution $\{u_h, 0 < h \le 1\}$ is uniformly bounded.

THEOREM 4.1. *Let $u_h = \{u_h(t), 0 \le t < \infty\}$ be a solution to problem (2.3). Then, under the conditions of Theorem 2.1, for all $p \in (1, 2]$, with $p < \frac{2}{d} + 1$, and for all $T > 0$ there exists some constant such that*

$$\sup_{0\le t\le T} \mathbb{E}\, \|u_h(t)\|_{L^p(\mathcal{O})}^p \le C, \quad 0 < h \le 1.$$

Proof. First, we show that there exist some $q < \infty$ and some constants C_1 and C_2 such that

$$(\mathbb{E}\|u_h(t)\|^p)^q \le C_1 + C_2 \int_0^t (\mathbb{E}\, \|u_h(s)\|^p)^q \, ds, \quad t > 0 \text{ and } 0 < h \le 1.$$

The Gronwall lemma yields the assertion. The variation of constant formula gives

$$\begin{aligned}
u_h(t) = e^{-t\mathcal{A}_h} u_h(0) &+ \int_0^t e^{-(t-s)\mathcal{A}_h} f_h(u_h(s))\, ds \\
&+ \int_0^t \sum_{i=1}^{k_h} \int_{\zeta_i\in\mathbb{R}} e^{-(t-s)\mathcal{A}_h} g_h(u_h(s); i, \zeta_i)\, \tilde{\eta}_h^i(d\zeta_i; ds), \quad t > 0.
\end{aligned}$$

Taking the expectation leads to

$$\begin{aligned}
\mathbb{E}\, \|u_h(t)\|_{L^p(\mathcal{O}}^p &\le C_p \left\|e^{-t\mathcal{A}_h}\, u_h(0)\right\|_{L^p(\mathcal{O})}^p \\
&+ C_p \mathbb{E} \left\| \int_0^t e^{-(t-s)\mathcal{A}_h} f_h(u_h(s))\, ds \right\|_{L^p(\mathcal{O})}^p \\
&+ C_p \mathbb{E} \left\| \int_0^t \sum_{i=1}^{k_h} \int_{\zeta_i\in\mathbb{R}} e^{-(t-s)\mathcal{A}_h} g_h(u_h(s); i, \zeta_i)\, \tilde{\eta}_h^i(d\zeta_i; ds) \right\|_{L^p}^p \\
&=: S_1 + S_2 + S_3.
\end{aligned}$$

If $u_0 \in L^p(\mathcal{O})$, then the first entity S_1 is finite. The second entity can be calculated directly. The Minkowski inequality gives

$$\begin{aligned} S_2 &= \mathbb{E} \left\| \int_0^t e^{-(t-s)\mathcal{A}_h} f_h(u_h(s))\, ds \right\|^p_{L^p(\mathcal{O})} \\ &\leq \mathbb{E} \left(\int_0^t \left\| e^{-(t-s)\mathcal{A}_h} f_h(u_h(s)) \right\|_{L^p(\mathcal{O})} ds \right)^p. \end{aligned}$$

In section 3.2 we have seen that there exist $\omega \in \mathbb{R}$, $M \geq 1$, and $\theta \in (0, \pi)$ such that for all $h \in (0,1)$ the operator $\mathcal{A}_h$ generates an analytic semigroup of (ω, M, θ) on $L^p_h(\mathcal{O})$ and the inequality (3.17) holds. In section 3.3 we have seen that there exist some constants c and C such that for all $0 < h \leq 1$ we have (see (3.16))

$$c \|w\|_{L^p(\mathcal{O})} \leq \|w_h\|_{L^p_h(\mathcal{O})} \leq C \|w\|_{L^p(\mathcal{O})}, \quad w \in \mathcal{V}_h \tag{4.1}$$

where

$$\|w\|_{L^p_h(\mathcal{O})} := h^d \sum_{i=1}^{k_h} |w_h(N_i)|^p, \quad w \in \mathcal{V}_h.$$

Therefore, a constant $C < \infty$ exists such that for all $0 < h \leq 1$

$$\left\| e^{-(t-s)\mathcal{A}_h} \mathcal{A}_h v \right\|_{L^p(\mathcal{O})} \leq C \|v\|_{L^p(\mathcal{O})}, \quad v \in L^p(\mathcal{O}) \cap \mathcal{V}_h. \tag{4.2}$$

Thus,

$$S_2 \leq C\, \mathbb{E} \left(\int_0^t \|f_h(u_h(s))\|_{L^p(\mathcal{O})}\, ds \right)^p.$$

Again by (4.1) we can infer that

$$S_2 \ \leq C\, \mathbb{E} \left(\int_0^t \|f(u_h(s))\|_{L^p_h(\mathcal{O})}\, ds \right)^p.$$

By the globally Lipschitz property of f it follows that

$$S_2 \leq C\, \mathbb{E} \left(\int_0^t \left(1 + \|u_h(s)\|_{L^p_h(\mathcal{O})} \right) ds \right)^p.$$

Again by (4.1) we get

$$S_2 \leq C\, \mathbb{E} \left(\int_0^t \left(1 + \|u_h(s)\|_{L^p(\mathcal{O})} \right) ds \right)^p.$$

The Hölder inequality leads to

$$S_2 \leq C\, t^p \left(1 + \sup_{0 \leq s \leq t} \mathbb{E} \|u_h(s)\|^p_{L^p(\mathcal{O})} \right).$$

Next, we estimate the term

$$S_3 = \mathbb{E} \left\| \int_0^t \sum_{i=1}^{k_h} \int_{\zeta_i \in \mathbb{R}} e^{-(t-s)\mathcal{A}_h} g_h(u_h(s); i, \zeta_i)\, \tilde{\eta}^i_h(d\zeta_i; ds) \right\|^p_{L^p(\mathcal{O})}.$$

Since $L^p(\mathcal{O})$ is of M-type p, we can infer by Proposition 3.3 in [Hau05]

$$S_3 \leq \mathbb{E} \int_0^t \frac{1}{k_h} \sum_{i=1}^{k_h} \int_{\zeta_i \in \mathbb{R}} \mathbb{E} \left\| e^{-(t-s)\mathcal{A}_h} g_h(u_h(s); i; \zeta_i) \right\|_{L^p(\mathcal{O})}^p \nu(d\zeta_i)\, ds.$$

In section 3.2 we have seen that there exist $\omega \in \mathbb{R}$, $M \geq 1$, and $\theta \in (0, \pi)$ such that for all $h \in (0,1)$ the operator $\mathcal{A}_h$ generates an analytic semigroup of (ω, M, θ) on $L^p(\mathcal{O})$. In particular (3.18) holds; i.e., there exists a constant $C < \infty$ such that for all $h > 0$

$$\left\| e^{-(t-s)\mathcal{A}_h} \mathcal{A}_h v \right\|_{L_h^p(\mathcal{O})} \leq \frac{C}{t} \|v\|_{L_h^p(\mathcal{O})}, \quad v \in L_h^p(\mathcal{O}). \tag{4.3}$$

Because of inequality (4.3), and since the norms in $L_h^p(\mathcal{O})$ and $L^p(\mathcal{O})$ are equivalent, we have for all $\gamma < \frac{1}{p}$

$$S_3 \leq \int_0^t \frac{1}{k_h} \sum_{i=1}^{k_h} (t-s)^{-\gamma p} \int_{\zeta_i \in \mathbb{R}} \mathbb{E} \left\| \mathcal{A}_h^{-\gamma} g_h(u_h(s); i, \zeta_i) \right\|_{L^p(\mathcal{O})}^p \nu(d\zeta_i)\, ds.$$

Since $g_h(u_h(s); i, \zeta) = N_i(G(u_h(s); \zeta_i))\phi_i$ and since $\mathcal{A}_h$ is linear, we get

$$S_3 \leq \int_0^t \frac{1}{k_h} \sum_{i=1}^{k_h} (t-s)^{-\gamma p} \int_{\zeta_i \in \mathbb{R}} \mathbb{E} \left\| N_i(G(u_h(s); \zeta_i)) \mathcal{A}_h^{-\gamma} \phi_i \right\|_{L^p(\mathcal{O})}^p \nu(d\zeta_i)\, ds.$$

The Hölder inequality gives

$$\begin{aligned} S_3 \leq & \int_0^t \frac{1}{k_h} \\ & \sum_{i=1}^{k_h} (t-s)^{-\gamma p} \int_{\zeta_i \in \mathbb{R}} \mathbb{E} \left\| N_i(G(u_h(s); \zeta_i)) \right\|_{L^p(\mathcal{O})}^p \left\| \mathcal{A}_h^{-\gamma} \phi_i \right\|_{L^\infty(\mathcal{O})}^p \nu(d\zeta_i)\, ds. \end{aligned} \tag{4.4}$$

Moreover, by (3.19) we know that for any $i = 1, \ldots, k_h$ we have

$$\mathcal{A}_h^{-\gamma} \phi_i = \sum_{l=1}^{k_h} u_{il} \phi_l^h,$$

where

$$u_{il} = \begin{cases} c\, h^{2\gamma}, & \text{if } (i,l) \subset \mathcal{C}_h, \\ 0, & \text{elsewhere,} \end{cases}$$

and there exists some constant m_0 such that for any i and $0 < h < 1$

$$\left| \{(i, \mathbb{N}^{k_h}) \cap \mathcal{C}_h \right| \leq m_0. \tag{4.5}$$

Therefore, we have

$$\left\| \mathcal{A}_h^{-\gamma} \phi_i \right\|_{L^\infty(\mathcal{O})} \leq m_0 h^{2\gamma}. \tag{4.6}$$

By substituting (4.6) in (4.4) we obtain

$$S_3 \leq \int_0^t \sum_{i=1}^{k_h} \lambda(J_i^h)$$
$$(t-s)^{-\gamma p}\, m_0\, h^{2\gamma p} \int_{\zeta\in\mathbb{R}} \mathbb{E}\left\|g(u_h(s,N_i^h),N_i^h;\zeta)\right\|^p_{L^p(\mathcal{O})}\, \nu(d\zeta_i)\, ds. \tag{4.7}$$

The inverse inequality given in Corollary 3.1 gives for all $\delta > d - \frac{d}{p}$

$$\begin{aligned} S_3 &\leq \int_0^t C \sum_{i=1}^{k_h} \lambda(J_i^h) \\ &\quad (t-s)^{-\gamma p}\, h^{2\gamma p-\delta p} \int_{\zeta\in\mathbb{R}} \mathbb{E}\left\|g(u_h(s,N_i^h),N_i^h;\zeta_i)\right\|^p_{B^{\frac{d}{p}-d}_{p,\infty}(\mathcal{O})}\, \nu(d\zeta)\, ds \\ &\leq C \int_0^t \sum_{i=1}^{k_h} \lambda(J_i^h)(t-s)^{-\gamma p}\, h^{(2\gamma-\delta)p} \int_{\zeta_i\in\mathbb{R}} \mathbb{E}\left\|g(u_h(s,N_i^h),N_i^h;\zeta)\right\|^p_{B^{\frac{d}{p}-d}_{p,\infty}(\mathcal{O})}\, \nu(d\zeta)\, ds, \end{aligned}$$

which gives

$$S_3 \leq C \int_0^t \sum_{i=1}^{k_h} \lambda(J_i^h)(t-s)^{-\gamma p}\, h^{(2\gamma-\delta)p} \int_{\zeta_i\in\mathbb{R}} \mathbb{E}\left|g(u_h(s,N_i^h),N_i^h;\zeta)\right|^p\, \nu(d\zeta)\, ds.$$

The mapping g is linear in the last variable and globally Lipschitz in the first variable. Thus we can infer that

$$S_3 \leq C \int_0^t \sum_{i=1}^{k_h} \lambda(J_i^h)(t-s)^{-\gamma p}\, h^{(2\gamma-\delta)p} \left(1 + \mathbb{E}\left|u_h(s,N_i^h)\right|^p\right)\, ds.$$

By the equivalence of the $L^p_h(\mathcal{O})$ and $L^p(\mathcal{O})$ norms we infer that

$$S_3 \leq C \int_0^t (t-s)^{-\gamma p}\, h^{(2\gamma-\delta)p} \left(1 + \mathbb{E}\left|u_h(s)\right|^p_{L^p(\mathcal{O})}\right)\, ds.$$

Let q be so large that $\gamma p < 1 - \frac{1}{q}$. Since there exists a $\delta > d - \frac{d}{p}$ such that $2\gamma - \delta > 0$, the Hölder inequality gives

$$\left(\mathbb{E}\left|S_3\right|\right)^q \leq c(t) \int_0^t \left(1 + \mathbb{E}\left|u_h(s)\right|^p_{L^p(\mathcal{O})}\right)^q\, ds.$$

Summing up leads to the assertion. □

5. Proof of the main result. The error of the time discretization follows from Theorem 3.1 of [HM06a]. By the Lipschitz continuity of f and g we infer that

$$\int_{z\in Z} \|G(u;z)\|^p_{B^{\frac{d}{p}-d}_{p,\infty}(\mathcal{O})}\, \iota(dz) \leq C\|u\|^p_{L^p(\mathcal{O})}, \quad u \in L^p(\mathcal{O}),$$

and

$$\|F(u)\|_{L^p(\mathcal{O})} \leq C\|u\|_{L^p(\mathcal{O})}, \quad u \in L^p(\mathcal{O}).$$

Thus, Theorem 3.1 of [HM06a] gives the order of convergence in time for $\alpha < \frac{2}{p} - d + \frac{d}{p}$. Here we investigate the accuracy of the space discretization, i.e., the difference $\Delta(t) = u(t) - u_h(t)\mathcal{I}_h$ at time t. To show the assertion, we will show that for a certain $q < \infty$ there exist constants C_1, C_2, and C_3 such that for all $0 < h \le 1$

$$\begin{aligned} &\left(\mathbb{E}\,\|\Delta(t)\|^p_{L^p(\mathcal{O})}\right)^q \\ &\le \frac{C_1}{t^q}\, h^{pq\delta_1} + C_2\, h^{pq\delta_2} + C_3\, h^{pq\delta_3} + C_4 \int_0^t \left(\mathbb{E}\,\|\Delta(t)\|^p_{L^p(\mathcal{O})}\right)^q ds \end{aligned} \tag{5.1}$$

where δ_1, δ_2, and δ_3 are the constants specified in Theorem 2.1. Now the second part of the assertion of Theorem 2.1 follows by an extension of Grownwall's lemma which is a simple consequence of Exercise 3, page 190 in [Hen81] and the proof of which is omitted.

LEMMA 5.1 (Gronwall-type lemma). *Suppose that $\alpha > 0$ and $c_1, c_2, c_3 > 0$. Assume that $u : \mathbb{R}^+ \to \mathbb{R}^+$ is a measurable function such that the function $t \mapsto t^{1-\alpha}u(t)$ is locally integrable on $\mathbb{R}^+$ and a.s. on $\mathbb{R}^+$,*

$$u(t) \le c_1 t^{\alpha-1} + c_2 + c_3 \int_0^t u(s)\, ds.$$

Then, a.s. on $\mathbb{R}^+$,

$$u(t) \le \left(c_1 t^{\alpha-1}\Gamma(\alpha) + c_2\right) \sum_{m=0}^{\infty} \frac{[c_3 t]^m}{\Gamma(m+\alpha)}.$$

To show the second part of Theorem 2.1, we will show the following: Let q and γ be two real numbers such that

$$d - \frac{d}{p} < \gamma < 2p \quad \text{and} \quad \frac{\gamma p}{2} < 1 - \frac{1}{q} \tag{5.2}$$

are satisfied. Then under the conditions of Theorem 2.1 for all δ_1, δ_2, and δ_3, with

$$\delta_1 = 2 - \frac{d}{p} + \frac{d}{2}, \quad \delta_2 < \frac{2}{p} - d + \frac{d}{p}, \text{ and } \delta_3 < \left(\frac{2}{p} - d + \frac{d}{p}\right)\left(2 + \frac{d}{2} - \frac{d}{p}\right),$$

there exist constants C_1, C_2, and C_3 depending on p, f, g, δ_1, δ_2, and δ_3 such that for all $0 < h \le 1$ inequality (5.1) holds.

Proof. By using the variation of constants formula, i.e.,

$$\begin{aligned} u(t) = e^{-At}u_0 &+ \int_0^t e^{-A(t-s)} f(u(s-))\, ds \\ &+ \int_{0^+}^t \int_Z e^{-A(t-s)} g(u(s-); z)\eta(ds; dz) \end{aligned}$$

and

$$\begin{aligned} u_h(t) = e^{-\mathcal{A}_h t}u_0 &+ \int_0^t e^{-\mathcal{A}_h(t-s)} f(u(s-))\, ds \\ &+ \int_{0^+}^t \int_Z e^{-\mathcal{A}_h(t-s)} g u(s-); z)\eta(ds; dz), \end{aligned}$$

we obtain

$$u(t) - u_h(t) = \sum_{i=1}^{6} I_{ni}(t), \tag{5.3}$$

where

$$\begin{aligned}
I_0(t) &= \left(e^{-At} - e^{-\mathcal{A}_h t}\right) u_0,\\
I_1(t) &= \int_{0+}^{t} \left(e^{-A(t-s)} - e^{-\mathcal{A}_h(t-s)}\mathcal{I}_h\right) f(u(s))\, ds,\\
I_2(t) &= \int_{0+}^{t} e^{-\mathcal{A}_h(t-s)} \left(\mathcal{I}_h f(u(s)) - f_h(u_h(s))\right)\, ds,\\
I_3(t) &= \int_{0+}^{t}\int_{Z} e^{-A(t-s)} \left(g(u(s);z) - g(u_h(s);z)\right)\, \tilde{\eta}(ds;dz),\\
I_4(t) &= \int_{0+}^{t}\int_{Z} e^{-A(t-s)} \left(I - \mathcal{I}_h\right) g(u_h(s);z)\, \tilde{\eta}(ds;dz),\\
I_5(t) &= \int_{0+}^{t}\int_{Z} e^{-A(t-s)}\mathcal{I}_h g(u_h(s);z)\, \tilde{\eta}(ds;dz)\\
&\quad - \int_{0+}^{t}\sum_{i=1}^{k_h}\int_{\zeta\in\mathbb{R}} e^{-A(t-s)} g_h(u_h(s);i;\zeta)\, \tilde{\eta}_h^i(ds,d\zeta),\\
I_6(t) &= \int_{0+}^{t}\sum_{i=1}^{k_h}\int_{\zeta\in\mathbb{R}} \left(e^{-A(t-s)} - e^{-\mathcal{A}_h(t-s)}\right) g_h(u_h(s);i;\zeta)\, \tilde{\eta}_h^i(ds;d\zeta).
\end{aligned}$$

By substituting (5.3), we get

$$\mathbb{E}\left[\|\Delta(t)\|^p\right] \leq C \cdot \sum_{i=0}^{4} \mathbb{E}\left[\|I_i(t)\|^p\right].$$

To give an estimate of $\mathbb{E}\left[\|I_0(t)\|^p\right]$, we apply Theorem 3.1. Thus, we get

$$\mathbb{E}\left[\|I_0(t)\|^p\right] \leq \frac{C}{t}\, h^{p\left(2-\frac{d}{p}+\frac{d}{2}\right)}\, \|u_0\|^p_{L^p(\mathcal{O})}. \tag{5.4}$$

The next entity we have to estimate is

$$\mathbb{E}\,\|I_1(t)\|^p_{L^p(\mathcal{O})} \leq \mathbb{E}\left(\int_{0+}^{t} \left\|e^{-A(t-s)} - e^{-\mathcal{A}_h(t-s)}\mathcal{I}_h f(u(s))\right\|_{L^p(\mathcal{O})}\, ds\right)^p.$$

In Theorem 3.4, the factor t^{-1} appears on the right-hand side. Thus, fix $\gamma < 1$, take into account

$$\left\|e^{-A(t-s)} - e^{-\mathcal{A}_h(t-s)}\mathcal{I}_h\right\|_{L(L^p(\mathcal{O}),L^p(\mathcal{O}))} \leq C,$$

and apply Theorem 3.4. Thus, we get

$$\begin{aligned}
&\mathbb{E}\,\|I_1(t)\|^p_{L^p(\mathcal{O})}\\
&\le C\,\left(\int_{0+}^{t}\left\|e^{-A(t-s)}-e^{-\mathcal{A}_h(t-s)}\mathcal{I}_h\right\|_{L(L^p(\mathcal{O}),L^p(\mathcal{O}))}\,\mathbb{E}\,\|f(u(s))\|^p_{L^p(\mathcal{O})}\,ds\right)^p\\
&\le C\,\Big(\int_{0+}^{t}\left\|e^{-A(t-s)}-e^{-\mathcal{A}_h(t-s)}\mathcal{I}_h\right\|^{\gamma}_{L(L^p(\mathcal{O}),L^p(\mathcal{O}))}\\
&\quad\times\left\|e^{-A(t-s)}-e^{-\mathcal{A}_h(t-s)}\mathcal{I}_h\right\|^{1-\gamma}_{L(L^p(\mathcal{O}),L^p(\mathcal{O}))}\,\mathbb{E}\,\|f(u(s))\|^p_{L^p(\mathcal{O})}\,ds\Big)^p\\
&\le C\,h^{\gamma(2-\frac{d}{p}+\frac{d}{2})p}\,\left(\int_{0+}^{t}(t-s)^{-\gamma}\,\mathbb{E}\,\|f(u(s))\|^p_{L^p(\mathcal{O})}\,ds\right)^p\\
&\le C(\gamma)\,h^{\gamma(2-\frac{d}{p}+\frac{d}{2})p}\sup_{0\le t\le T}\mathbb{E}\,\|f(u(s))\|^p_{L^p(\mathcal{O})}\,,
\end{aligned}$$

where

$$C(\gamma)=\frac{C'}{1-\gamma}.$$

The Lipschitz property of f gives

$$\mathbb{E}\,\|I_1(t)\|^p_{L^p(\mathcal{O})}\le C\,h^{p(2-\frac{d}{p}+\frac{d}{2})\gamma}\,\left(1+\sup_{0\le s\le t}\mathbb{E}\,\|u(s)\|^p_{L^p(\mathcal{O})}\right).$$

Since $u\in L^p(\Omega;L^p(\mathcal{O}))$, there exists some C such that

$$\mathbb{E}\,\|I_1(t)\|^p_{L^p(\mathcal{O})}\le C\,h^{p(2-\frac{d}{p}+\frac{d}{2})\gamma},\quad 0<h<1. \tag{5.5}$$

In section 3.3 we have shown that $\mathcal{A}_h$ generates a semigroup in $L^p_h(\mathcal{O})$. Thus, by the Minkowski inequality (4.2) and by the equivalence of the norms in $L^p(\mathcal{O})$ and $L^p_h(\mathcal{O})$, i.e., (3.16), we can write

$$\begin{aligned}
\mathbb{E}\,\|I_2(t)\|^p_{L^p(\mathcal{O})}&\le\mathbb{E}\left(\int_{0+}^{t}\|\mathcal{I}_hf(u(s))-f_h(u_h(s))\|_{L^p(\mathcal{O})}\,ds\right)^p\\
&\le C\,\mathbb{E}\left(\int_{0+}^{t}\|\mathcal{I}_hf(u(s))-f_h(u_h(s))\|_{L^p_h(\mathcal{O})}\,ds\right)^p.
\end{aligned}$$

The mapping f is Lipschitz continuous, and therefore

$$\mathbb{E}\,\|I_2(t)\|^p_{L^p(\mathcal{O})}\le C\,\mathbb{E}\left(\int_{0+}^{t}\|\mathcal{I}_hu(s)-u_h(s)\|_{L^p_h(\mathcal{O})}\,ds\right)^p.$$

Again, by the equivalence of the norms in $L^p(\mathcal{O})$ and $L^p_h(\mathcal{O})$, we can infer that

$$\mathbb{E}\,\|I_2(t)\|^p_{L^p(\mathcal{O})}\le C\,\mathbb{E}\left(\int_{0+}^{t}\|u(s)-u_h(s)\|_{L^p(\mathcal{O})}\,ds\right)^p.$$

The Hölder inequality gives

$$\mathbb{E}\,\|I_2(t)\|^p_{L^p(\mathcal{O})}\le C\,\mathbb{E}\int_{0+}^{t}\|u(s)-u_h(s)\|^p_{L^p(\mathcal{O})}\,ds. \tag{5.6}$$

Next, we handle the third term. The generalized Burkholder inequality (compare Proposition 3.3 in [Hau05]) gives

$$\mathbb{E}\,\|I_3(t)\|^p_{L^p(\mathcal{O})} \leq \int_{0+}^t \int_Z \mathbb{E}\left\|e^{-A(t-s)}\left(g(u(s);z)-g(u_h(s);z)\right)\right\|^p_{L^p(\mathcal{O})}\,\iota(dz)\,ds.$$

The operator A generates an analytic semigroup in $L^p(\mathcal{O})$ (see, e.g., Theorem 7.3.5 in [Paz83]). Therefore, for each $\rho>0$ there exists some constant $C<\infty$ such that

$$\left\|e^{-(t-s)A}Av\right\|_{W_p^{-\rho}(\mathcal{O})} \leq \frac{C}{t^{\frac{\rho}{2}}}\,\|v\|_{L^p(\mathcal{O})}\,, \quad v\in L^p(\mathcal{O}). \tag{5.7}$$

Thus, we can infer for any $\gamma>0$ that

$$\mathbb{E}\,\|I_3(t)\|^p_{L^p(\mathcal{O})} \leq \int_{0+}^t \int_Z (t-s)^{-\frac{\gamma p}{2}}\,\mathbb{E}\,\|g(u(s);z)-g(u_h(s);z)\|^p_{W_p^{-\gamma}(\mathcal{O})}\,\iota(dz)\,ds.$$

By Sobolev inequalities for any $\gamma > d-\frac{d}{p}$ we have $Z = B^{\frac{d}{p}-d}_{p,\infty}(\mathcal{O}) \hookrightarrow W_p^{-\gamma}(\mathcal{O})$ continuously. Therefore,

$$\mathbb{E}\,\|I_3(t)\|^p_{L^p(\mathcal{O})} \leq \int_{0+}^t \int_Z (t-s)^{-\frac{\gamma p}{2}}\,\mathbb{E}\,\|(g(u(s);z)-g(u_h(s);z))\|^p_Z\,\iota(dz)\,ds.$$

The Lipschitz property of g yields

$$\mathbb{E}\,\|I_3(t)\|^p_{L^p(\mathcal{O})} \leq \int_{0+}^t (t-s)^{-\frac{\gamma p}{2}}\,\mathbb{E}\,\|u(s)-u_h(s)\|^p_{L^p(\mathcal{O})}\,ds.$$

If γ satisfies (5.2), then $\frac{\gamma p}{2} < 1-\frac{1}{q}$, and we can infer by the Hölder inequality that

$$\left(\mathbb{E}\,\|I_3(t)\|^p_{L^p(\mathcal{O})}\right)^q \leq c(t)\int_0^t \left(\mathbb{E}\,\|u(s)-u_h(s)\|^p_{L^p(\mathcal{O})}\right)^q ds. \tag{5.8}$$

Similarly by Proposition 3.2 in [Hau05] we get for the fourth term

$$\mathbb{E}\,\|I_4(t)\|^p_{L^p(\mathcal{O})} \leq \int_{0+}^t \int_Z \mathbb{E}\left\|e^{-A(t-s)}\left(I-\mathcal{I}_h\right)g(u_h(s);z)\right\|^p_{L^p(\mathcal{O})}\,\nu(dz)\,ds.$$

By the smoothing property of A, i.e., (5.7), and since $Z = B^{\frac{d}{p}-d}_{p,\infty}(\mathcal{O}) \hookrightarrow W_p^{-\gamma}(\mathcal{O})$ continuously for all $\gamma > d-\frac{d}{p}$, we obtain for any $\rho < \frac{2}{p}-d+\frac{d}{p}$

$$\begin{aligned}
&\mathbb{E}\,\|I_4(t)\|^p_{L^p}\\
&\leq \int_{0+}^t (t-s)^{-\frac{p}{2}\left(\rho+d-\frac{d}{p}\right)} \int_Z \mathbb{E}\left\|A^{-\rho}\left(I-\mathcal{I}_h\right)G(u_h(s);z)\right\|^p_Z\,\iota(dz)\,ds\\
&\leq \int_{0+}^t (t-s)^{-\frac{p}{2}\left(\rho+d-\frac{d}{p}\right)} \int_{\zeta\in\mathbb{R}} \mathbb{E}\left\|A^{-\rho}\left(I-\mathcal{I}_h\right)g(u_h(s),\cdot,\zeta)\right\|^p_{L^p(\mathcal{O})}\,\nu(d\zeta)\,ds,
\end{aligned}$$

where G is defined in (3.20). Theorem 3.1 leads to

$$\mathbb{E}\,\|I_4(t)\|^p_{L^p(\mathcal{O})} \leq C(t)\,h^{p\rho}\sup_{0\leq s\leq t}\int_{\zeta\in\mathbb{R}}\mathbb{E}\,\|g(u_h(s),\cdot;\zeta)\|^p_{L^p(\mathcal{O})}\,\nu(d\zeta)$$

and the Lipschitz properties of g to

$$\mathbb{E}\,\|I_4(t)\|_{L^p}^p \le C(t)\,h^{p\rho}\,\left(1+\sup_{0\le s\le t}\mathbb{E}\,\|u_h(s)\|_{L^p(\mathcal{O})}^p\right).$$

By Proposition 4.1 there exists some constant $C<\infty$ such that

$$\sup_{0\le s\le t}\mathbb{E}\,\|u_h(s)\|_{L^p(\mathcal{O})}^p \le C,\quad h>0.$$

Thus, we can infer that there exists some constant $C'<\infty$ such that

$$\mathbb{E}\,\|I_4(t)\|_{L^p(\mathcal{O})}^p \le C'\,h^{p\rho},\quad 0\le t\le T. \tag{5.9}$$

The Burkholder inequality leads for any $\frac{1}{2}(d-\frac{d}{p})<\gamma<\frac{1}{p}$ to

$$\begin{aligned}
&\mathbb{E}\,\|I_5(t)\|_{L^p(\mathcal{O})}^p\\
&\le \int_0^t\int_Z \left\|(-\mathcal{A}_h)^{\gamma}e^{-A(t-s)}(-\mathcal{A}_h)^{-\gamma}\mathcal{I}_hG(u(t),z)\right\|_{L^p(\mathcal{O})}^p\,(\iota-\iota_h)(dz)\\
&\le \int_0^t (t-s)^{-\gamma p}\,dt\times\sup_{0\le t\le T}\int_Z \left\|(-\mathcal{A}_h)^{-\gamma}\mathcal{I}_hG(u(t),z)\right\|_{L^p(\mathcal{O})}^p\,(\iota-\iota_h)(dz).
\end{aligned}$$

Proposition 3.2 gives for any $0\le\delta<\frac{1}{p}-\frac{1}{2}(d-\frac{d}{p})$

$$\mathbb{E}\,\|I_5(t)\|_{L^p(\mathcal{O})}^p \le C\,\frac{h^{\delta p}}{p\gamma-1}. \tag{5.10}$$

Next, we have to calculate the entity I_6. The Fubini theorem and the generalized Burkholder inequality give

$$\begin{aligned}
\mathbb{E}\,\|I_6(t)\|^p &= \mathbb{E}\left\|\int_{0+}^t\sum_{i=1}^{k_h}\int_{\zeta\in\mathbb{R}}\left(e^{-A(t-s)}-e^{-\mathcal{A}_h(t-s)}\right)g_h(u_h(s);i,\zeta)\,\tilde{\eta}_h^i(ds;dz)\right\|_{L^p(\mathcal{O})}^p\\
&\le \mathbb{E}\int_{0+}^t\sum_{i=1}^{k_h}\int_{\zeta\in\mathbb{R}}\left\|\left(e^{-A(t-s)}-e^{-\mathcal{A}_h(t-s)}\right)g_h(u_h(s);i,\zeta)\right\|_{L^p(\mathcal{O})}^p\,\nu^i(d\zeta)\,ds\\
&\le \int_{0+}^t\sum_{i=1}^{k_h}\int_{\zeta\in\mathbb{R}}\mathbb{E}\Big\|\Big(e^{-\frac{1}{2}A(t-s)}\\
&\qquad -e^{-\frac{1}{2}\mathcal{A}_h(t-s)}\mathcal{I}_h\Big)e^{-\frac{1}{2}A(t-s)}g_h(u_h(s);i,\zeta)\Big\|_{L^p(\mathcal{O})}^p\,\nu^i(d\zeta)\,ds\\
&+ \int_{0+}^t\sum_{i=1}^{k_h}\int_{\zeta\in\mathbb{R}}\mathbb{E}\Big\|e^{-\frac{1}{2}\mathcal{A}_h(t-s)}\mathcal{I}_h\Big(e^{-\frac{1}{2}A(t-s)}\\
&\qquad -e^{-\frac{1}{2}\mathcal{A}_h(t-s)}\Big)g_h(u_h(s);i,\zeta)\Big\|_{L^p(\mathcal{O})}^p\,\nu^i(d\zeta)\,ds\\
&:= I_{61}+I_{62}.
\end{aligned}$$

First, in Theorem 3.4 we have shown that

$$\left\|\left[e^{-\frac{1}{2}A(t-s)}-e^{-\frac{1}{2}\mathcal{A}_h(t-s)}\mathcal{I}_h\right]u\right\|_{L^p}\le\frac{C}{t}\,h^{2+\frac{d}{2}-\frac{d}{p}}\|u\|_{L^p(\mathcal{O})}\ \forall u\in L^p(\mathcal{O}). \tag{5.11}$$

Let us recall estimate (3.17). There exists a constant $C<\infty$ such that for all $h>0$ and $v\in L^p_h(\mathcal{O})\cap\mathcal{V}_h$

$$\left\|e^{-(t-s)\mathcal{A}_h}v\right\|_{L^p_h}\leq C\left\|v\right\|_{L^p_h},\quad v\in L^p_h.$$

Moreover, the norms in $L^p_h(\mathcal{O})$ and in $L^p(\mathcal{O})$ are equivalent (see (3.16)). Therefore

$$\left\|e^{-(t-s)\mathcal{A}_h}v\right\|_{L^p_h}\leq C\left\|v\right\|_{L^p(\mathcal{O})},\quad v\in L^p_h(\mathcal{O}).$$

Thus, we can infer for any $\gamma\in(0,1)$ that

$$\begin{aligned}I_{61}&\leq C\,h^{p\gamma(2+\frac{d}{2}-\frac{d}{p})}\\&\times\int_{0+}^t(t-s)^{-\frac{\gamma p}{2}}\sum_{i=1}^{k_h}\lambda(J_i^h)\int_{\zeta\in\mathbb{R}}\mathbb{E}\left\|e^{-\frac{1}{2}A(t-s)}g_h(u_h(s);i,\zeta)\right\|^p_{L^p(\mathcal{O})}\,\nu(d\zeta)\,ds\end{aligned}$$

holds. The operator A generates an analytic semigroup in $L^p(\mathcal{O})$. Therefore, for each $\rho>0$ there exists some constant $C<\infty$ such that

$$\left\|e^{-tA}Av\right\|_{L^p(\mathcal{O})}\leq\frac{C}{t^{\frac{\rho}{2}}}\left\|v\right\|_{W_p^{-\rho}(\mathcal{O})},\quad v\in W_p^{-\rho}(\mathcal{O}),\ t>0. \tag{5.12}$$

Inequality (5.12) leads for all $\rho\geq0$, with $p(\gamma+\rho)<2$, to

$$\begin{aligned}I_{61}&\leq C\,h^{p\gamma(2+\frac{d}{2}-\frac{d}{p})}\\&\quad\times\int_{0+}^t(t-s)^{-\frac{p}{2}(\gamma+\rho)}\sum_{i=1}^{k_h}\lambda(J_i^h)\int_{\zeta\in\mathbb{R}}\mathbb{E}\left\|g_h(u_h(s);i,\zeta)\right\|^p_{W_p^{-\rho}(\mathcal{O})}\,\nu(d\zeta)\,ds\\&\leq C(\gamma,\rho)\,h^{p\gamma(2+\frac{d}{2}-\frac{d}{p})}\sup_{0\leq s\leq t}\sum_{i=1}^{k_h}\lambda(J_i^h)\int_{\zeta\in\mathbb{R}}\mathbb{E}\left\|g_h(u_h(s);i,\zeta)\right\|^p_{W_p^{-\rho}(\mathcal{O})}\,\nu(d\zeta),\end{aligned}$$

where

$$C(\gamma,\rho)=\frac{C'}{2-p(\gamma+\rho)}.$$

By Sobolev inequalities, we know that $L^1(\mathcal{O})\hookrightarrow W_p^{-\rho}(\mathcal{O})$ continuously. Thus we obtain

$$\begin{aligned}I_{61}&\leq C\,h^{p\gamma(2+\frac{d}{2}-\frac{d}{p})}\sup_{0\leq s\leq t}\sum_{i=1}^{k_h}\lambda(J_i^h)\int_{\zeta\in\mathbb{R}}\mathbb{E}\left\|g_h(u_h(s);i,\zeta)\right\|^p_{L^1(\mathcal{O})}\,\nu(d\zeta)\\&\leq C\,h^{p\gamma(2+\frac{d}{2}-\frac{d}{p})}\sup_{0\leq s\leq t}\sum_{i=1}^{k_h}\lambda(J_i^h)\int_{\zeta\in\mathbb{R}}\mathbb{E}\left|g(u_h(s,N_i^h),N_i^h;\zeta)\right|^p\left\|c_i^h\phi_i^h\right\|^p_{L^1(\mathcal{O})}\,\nu(d\zeta)\\&\leq C\,h^{p\gamma(2+\frac{d}{2}-\frac{d}{p})}\sup_{0\leq s\leq t}\sum_{i=1}^{k_h}\lambda(J_i^h)\int_{\zeta\in\mathbb{R}}\mathbb{E}\left|g(u_h(s,N_i^h),N_i^h;\zeta)\right|^p\,\nu(d\zeta).\end{aligned}$$

Since g is Lipschitz continuous and linear in the third variable, we get

$$I_{61}\leq C\,h^{p\gamma(2+\frac{d}{2}-\frac{d}{p})}\sup_{0\leq s\leq t}\sum_{i=1}^{k_h}\lambda(J_i^h)\left(1+\mathbb{E}\left\|u_h(s,N_i^h)\right\|^p_{L^p_h}\right)\,ds.$$

In section 3.3 we have seen that there exist some constants c and C such that for all $0 < h < 1$ we have (see (3.16))

$$c \left\| w \right\|_{L^p(\mathcal{O})} \leq h^{p\gamma(2+\frac{d}{2}-\frac{d}{p})} \left\| w_h \right\|_{L_h^p} \leq C \left\| w \right\|_{L^p(\mathcal{O})}, \quad w \in \mathcal{V}_h.$$

Thus, by taking into account Remark 3.1 we can write

$$I_{61} \leq Ct^{1-p(\gamma+\rho)} h^{p\gamma(2+\frac{d}{2}-\frac{d}{p})} \sup_{0\leq s\leq t} \left(1 + \mathbb{E} \left\| u_h(s) \right\|_{L^p(\mathcal{O})}^p\right).$$

Since $\mathbb{E} \left\| u_h(s) \right\|_{L^p(\mathcal{O})}^p$ is bounded, there exists a constant $C < \infty$ such that

$$I_{61} \leq Ct^{1-p(\gamma+\rho)} h^{p\gamma(2+\frac{d}{2}-\frac{d}{p})}. \tag{5.13}$$

Next, we have to handle the term I_{62}. Because of (3.18) and (3.16), there exists a constant $C < \infty$ such that for all $0 < h \leq 1$

$$\left\| e^{-(t-s)\mathcal{A}_h} \mathcal{A}_h v \right\|_{L^p(\mathcal{O})} \leq \frac{C}{t} \left\| v \right\|_{L^p(\mathcal{O})}, \quad v \in L^p(\mathcal{O}). \tag{5.14}$$

Because of inequality (5.14), we have for all $0 < \delta < \frac{1}{p}$

$$\begin{aligned} I_{62} \leq & \int_0^t (t-s)^{-\delta p} \sum_{i=1}^{k_h} \lambda(J_i^h) \\ & \int_{\zeta\in\mathbb{R}} \mathbb{E} \left\| (-\mathcal{A}_h)^{-\delta} \left(\mathcal{I}_h e^{-\frac{1}{2}A(t-s)} - e^{-\frac{1}{2}\mathcal{A}_h(t-s)} \right) g_h(u_h(s); i, \zeta) \right\|_{L^p(\mathcal{O})}^p \nu(d\zeta)\, ds. \end{aligned} \tag{5.15}$$

In Remark 3.1 we have shown that there exist some constants c and C such that $h^d \leq ck_h \leq h^d$. By (3.19) we know that for any $i = 1, \ldots, k_h$ we have

$$(-\mathcal{A}_h)^{-\delta} \phi_i = \sum_{l=1}^{k_h} u_{il} \phi_l^h, \tag{5.16}$$

where

$$u_{il} = \begin{cases} c\, h^{2\delta}, & \text{if } (i,l) \subset \mathcal{C}_h, \\ 0, & \text{elsewhere.} \end{cases}$$

Moreover, there exists some constant m_0 such that for any i and $0 < h < 1$

$$\left| \{(i, \mathbb{N}^{k_h}) \cap \mathcal{C}_h\} \right| \leq m_0.$$

By substituting (5.16) in (5.15) we obtain

$$\begin{aligned} I_{62} \leq & h^{2\delta p} \int_0^t (t-s)^{-\delta p} m_0 \sum_{i=1}^{k_h} \lambda(J_i^h) \\ & \int_{\zeta\in\mathbb{R}} \mathbb{E} \left\| \left(\mathcal{I}_h e^{-\frac{1}{2}A(t-s)} - e^{-\frac{1}{2}\mathcal{A}_h(t-s)} \right) g_h(u_h(s); i, \zeta) \right\|_{L^p(\mathcal{O})}^p \nu^i(d\zeta)\, ds. \end{aligned}$$

By applying Theorem 3.4 we can infer that for any $\gamma \in (01)$, we have

$$I_{62} \leq C\, h^{p\delta+p\gamma(2+\frac{d}{2}-\frac{d}{p})} \int_0^t (t-s)^{-(\delta+\gamma)p} m_0 \sum_{i=1}^{k_h} \lambda(J_i^h) \int_{\zeta\in\mathbb{R}} \mathbb{E}\Big\| g_h(u_h(s);i,\zeta)\Big\|^p_{L^p(\mathcal{O})}\, \nu(d\zeta)\, ds,$$

where $\rho = \gamma/(2-\frac{d}{p}+\frac{d}{2})$. By Theorem 3.1 and since $Z = B^{\frac{d}{p}-d}_{p,\infty}(\mathcal{O}) \hookrightarrow W_p^{-\gamma}(\mathcal{O})$ continuously, we get for any $\rho > d - \frac{d}{p}$

$$I_{62} \leq C\, h^{p(\delta-\rho)+p\gamma(2+\frac{d}{2}-\frac{d}{p})} \int_0^t (t-s)^{-(\delta+\rho)p} m_0 \sum_{i=1}^{k_h} \lambda(J_i^h) \int_{\zeta\in\mathbb{R}} \mathbb{E}\Big\| g_h(u_h(s);i,\zeta)\Big\|^p_{Z}\, \nu(d\zeta)\, ds.$$

Provided that $\delta \geq \rho$ we can omit the first term in the exponent of h. The mapping g is linear in the last variable and globally Lipschitz in the first. Thus we can infer that

$$I_{62} \leq C\, h^{p\gamma(2+\frac{d}{2}-\frac{d}{p})} \int_0^t (t-s)^{-(\delta+\rho)p} m_0 \sum_{i=1}^{k_h} \lambda(J_i^h) \left(1+\mathbb{E}\left|u_h(s,N_i^h)\right|^p\right) ds.$$

By Remark 3.1 it follows that

$$I_{62} \leq C\, h^{p\gamma(2+\frac{d}{2}-\frac{d}{p})} \int_0^t (t-s)^{-(\delta+\rho)p} m_0 \sum_{i=1}^{k_h} h^d \left(1+\mathbb{E}\left|u_h(s,N_i^h)\right|^p\right) ds,$$

and by (3.16)

$$I_{62} \leq C\, h^{p\gamma(2+\frac{d}{2}-\frac{d}{p})} \int_0^t (t-s)^{-(\delta+\rho)p} \left(1+\mathbb{E}\,\|u_h(s)\|^p_{L^p(\mathcal{O})}\right) ds.$$

The Hölder inequality leads to

$$I_{62} \leq C\, h^{p\gamma(2+\frac{d}{2}-\frac{d}{p})}\, C(t) \left(1+\sup_{0\leq s\leq t} \mathbb{E}\,\|u_h(s)\|^p_{L^p(\mathcal{O})}\right)$$

and Theorem 4.1 that there exists a constant C such that

$$I_{62} \leq C\, h^{p\gamma(2+\frac{d}{2}-\frac{d}{p})}, \quad 0 < h \leq 1. \tag{5.17}$$

Collecting together (5.4), (5.5), (5.6), (5.8), (5.9), (5.10), (5.13), and (5.17) gives the assertion. ∎

6. Discussions and conclusions. The next step is to confirm the theory described in this paper by numerical experiments. One of the problems which arises here is that, if the intensity of the Poisson random measure is given, the distribution is known only in a few cases. Therefore, it is difficult to find a benchmark where the solution is known analytically and to compare the numerical result with the exact solution.

Moreover, because of this problem, the distribution of the random variables $\xi_h^m(\cdot)$, $m \in \mathbb{N}$, appearing in (2.6), is not given a priori. Different approaches exist to solve this problem; e.g., Rubenthaler [Rub03] cut off the jumps, which are smaller than a certain threshold, calculated the time of the first jump, and applied the Euler scheme until the time of the first jump. Then he simulated a random variable according to the distribution of size of the jump, with the condition that the jump is larger than the threshold. This step is repeated until the time T is reached.

In [HM06a, HM06b] we approximated the intensity by a simple function. In fact, first we cut off the small jumps by setting $\nu_h(B) = 0$ for a neighborhood B of the origin, and then we approximated ν_h by a simple function ν_h^n. To be precise, let $\tilde{\eta}_h^n$ be the real-valued compensated Poisson random measure given by the intensity ν_h^n. Now for any $x \in \mathbb{R}$ the random variable

$$\int_{m\tau_h}^{(m+1)\tau_h} \int_{\zeta \in \mathbb{R}^{k_h}} g(x;\zeta)\, \tilde{\eta}_h^n(dz;ds)$$

can be simulated directly by the sum of N independent random variables with distribution $\frac{1}{\nu_h^n(\mathbb{R})}\nu_h^n$, where N is exponentially distributed with parameter $\nu_h^n(\mathbb{R})$.

By taking into account that the small jumps are cut off, one may ask if the perturbation of the small jumps can be approximated by, e.g., a Wiener process. This has been done, e.g., by Kallianpur and Xiong [KX95]. Since we are in Banach spaces of M-type p, $1 \le p \le 2$, the convergence of a simple random walk is not necessarily given. One alternative is to choose another scaling, which we also consider in [HM06b].

Appendix A. Besov spaces and their properties. An introduction to Besov spaces is given by Runst and Sickel [RS96]. We are interested is the continuity of the mapping G described in (3.20). To be precise, we prove the following result.

PROPOSITION A.1. *If $f \in \mathcal{S}(\mathbb{R}^d)$, then for each $a \in \mathbb{R}^d$ the Schwartz distribution $f\delta_a$ belongs to the Besov space $B_{p,\infty}^{\frac{d}{p}-d}(\mathbb{R}^d)$, and moreover there exists a constant $C = C(d,p)$ independent of f such that*

$$\int_{\mathbb{R}^d} |f\delta_a|^p_{B_{p,\infty}^{\frac{d}{p}-d}(\mathbb{R}^d)}\, da = |f|^p_{L^p(\mathbb{R}^d)}\,.$$

Thus, there exists a unique bounded linear map $\Lambda : L^p(\mathbb{R}^d) \to L^p(\mathbb{R}^d, B_{p,\infty}^{\frac{n}{p}-n}(\mathbb{R}^d))$ such that $[\Lambda(f)](a) = f\delta_a$, $f \in \mathcal{S}(\mathbb{R}^d)$, $a \in \mathbb{R}^d$. Moreover, for any $s \ge 0$ there exists some constant C such that

$$\int_{\mathbb{R}^d} |f\delta_a|^p_{B_{p,\infty}^{\frac{d}{p}-d-s}(\mathbb{R}^d)}\, da \le C\ |f|^p_{W_p^{-s}(\mathbb{R}^d)}\,.$$

In the following we denote the value of $\Lambda(f)$ at a, where $f \in L^p(\mathbb{R}^d)$ by $f\delta_a$. Let us recall the definition of the Besov spaces as given in [RS96, Definition 2, pages 7–8]. First we choose a function $\psi \in \mathcal{S}(\mathbb{R}^d)$ such that $0 \le \psi(x) \le 1$, $x \in \mathbb{R}^d$, and

$$\psi(x) = \begin{cases} 1, & \text{if} \quad |x| \le 1, \\ 0 & \text{if} \quad |x| \ge \frac{3}{2}. \end{cases}$$

Then put

$$\begin{cases} \phi_0(x) = \psi(x),\ x \in \mathbb{R}^d, \\ \phi_1(x) = \psi(\frac{x}{2}) - \psi(x),\ x \in \mathbb{R}^d, \\ \phi_j(x) = \phi_1(2^{-j+1}x),\ x \in \mathbb{R}^d, \quad j = 2,3,\ldots. \end{cases}$$

With the choice of $\phi = \{\phi_j\}_{j=0}^{\infty}$ as above and $\mathcal{F}$ and $\mathcal{F}^{-1}$ being the Fourier and the inverse Fourier transformations, respectively, (acting in the space $\mathcal{S}'(\mathbb{R}^d)$ of Schwartz distributions), we have the following definition.

DEFINITION A.1. *Let $s \in \mathbb{R}$, $0 < p \leq \infty$, and $0 < q \leq \infty$. If $0 < q < \infty$ and $f \in \mathcal{S}'(\mathbb{R}^d)$, we put*

$$|f|_{B^s_{p,q}} = \left(\sum_{j=0}^{\infty} 2^{sjq} \left| \mathcal{F}^{-1} \left[\phi_j \mathcal{F} f \right] \right|_{L^p}^q \right)^{\frac{1}{q}},$$
$$|f|_{B^s_{p,\infty}} = \sup_{j \in \mathbb{N}} 2^{sj} \left| \mathcal{F}^{-1} \left[\phi_j \mathcal{F} f \right] \right|_{L^p}.$$

We denote by $B^s_{p,q}(\mathbb{R}^d)$ the space of all $f \in \mathcal{S}'(\mathbb{R}^d)$ for which $|f|_{B^s_{p,q}}$ is finite.

LEMMA A.2. *If $h \in \mathcal{S}(\mathbb{R}^d)$, $h \in \mathcal{S}(\mathbb{R}^d)$, and $a \in \mathbb{R}^d$, then*

$$|(h\delta_a) * g|_{L^p(\mathbb{R}^d)} = |h(a)||g|_{L^p(\mathbb{R}^d)}.$$

Proof. Let us recall that $h\delta_b \in \mathcal{S}'(\mathbb{R}^d)$ assigns to a function $g \in \mathcal{S}(\mathbb{R}^d)$ a value $\delta_a(hg) = h(a)g(a)$. Since by the definition of the convolution of a distribution with a test function

$$[(h\delta_a) * g](x) = (h\delta_a)(\tau_x \check{g}) = h(a)(\tau_x \check{g})(a) = h(a)\check{g}(a - x) = h(a)g(x - a),\ x \in \mathbb{R}^d,$$

we have

$$\begin{aligned} |(h\delta_a) * g|^p_{L^p(\mathbb{R}^d)} &= \int_{\mathbb{R}^d} |g(x - a)h(a)|^p \, dx \\ &= |h(a)|^p \int_{\mathbb{R}^d} |g(x)|^p \, dx = |h(a)|^p \ |g|^p_{L^p(\mathbb{R}^d)}. \end{aligned}$$

∎

LEMMA A.3. *If $\varphi \in \mathcal{S}(\mathbb{R}^d)$, $\lambda > 0$, and $g(x) := \varphi(\lambda x)$, $x \in \mathbb{R}^d$, then*

$$|\mathcal{F}^{-1} g|_{L^p(\mathbb{R}^d)} = \lambda^{d(\frac{1}{p} - 1)} |\mathcal{F}^{-1} \varphi|_{L^p(\mathbb{R}^d)}.$$

Proof. The proof follows with simple calculations. ∎

Proof of Proposition A.1. Obviously it is enough to prove the first part of the proposition. We will use the definition of the Fourier transform as in [RS96]. In particular, with $\langle \cdot, \cdot \rangle$ being the scalar product in $\mathbb{R}^d$, we put

$$(\mathcal{F} f)(\xi) := (2\pi)^{-d/2} \int_{\mathbb{R}^d} e^{-i\langle x, \xi \rangle} f(x) \, dx,\ \xi \in \mathbb{R}^d,\ f \in \mathcal{S}(\mathbb{R}^d).$$

Let us fix $f \in \mathcal{S}(\mathbb{R}^d)$. Since $\mathcal{F}^{-1}(\varphi u) = (2\pi)^{-d/2}(\mathcal{F}^{-1}\varphi) * (\mathcal{F}^{-1} u)$, $\varphi \in \mathcal{S}$, $u \in \mathcal{S}'$, we infer that, for $j \in \mathbb{N}^*$,

$$\begin{aligned} |\mathcal{F}^{-1}[\varphi_j \mathcal{F}(f\delta_a)]|_{L^p(\mathbb{R}^d)} &= (2\pi)^{-d/2} |(\mathcal{F}^{-1}\varphi_j) * (f\delta_a)|_{L^p(\mathbb{R}^d)} \\ &= (2\pi)^{-d/2} |f(a)| \mathcal{F}^{-1}\varphi_j|_{L^p(\mathbb{R}^d)} \\ &= (2\pi)^{-d/2} 2^{d(\frac{1}{p} - 1)} 2^{-jd(\frac{1}{p} - 1)} |f(a)| |\mathcal{F}^{-1}\varphi_1|_{L^p(\mathbb{R}^d)}. \end{aligned}$$

Hence, $f\delta_a$ belongs to the Besov space $B^{d(\frac{1}{p} - 1)}_{p,\infty}(\mathbb{R}^d)$ as requested, and the equality (A.1) follows immediately. In order to show (A.1) note that

$$\|f\|_{B^{s+\gamma}_{p,\infty}(\mathbb{R}^d)} = \|I(s) f\|_{B^{\gamma}_{p,\infty}(\mathbb{R}^d)},$$

where $I(s) = \mathcal{F}^{-1}\left(1+\xi^2\right)^{\frac{s}{2}}\mathcal{F}$. Next,

$$\begin{aligned}
&\left\|\mathcal{F}^{-1}\left(\varphi_j\left(1+\xi^2\right)^{\frac{s}{2}}\mathcal{F}\left(\int_{\mathbb{R}^d}\delta_a f\,da\right)\right)\right\|_{L^p(\mathbb{R}^d)}\\
&\le \left\|\mathcal{F}^{-1}\varphi_j \star \mathcal{F}^{-1}\left(1+\xi^2\right)^{\frac{s}{2}}\mathcal{F}\left(\int_{\mathbb{R}^d}\delta_a f\,da\right)\right\|_{L^p(\mathbb{R}^d)}\\
&\le \left\|\mathcal{F}^{-1}\varphi_j\right\|_{L^p(\mathbb{R}^d)}\left\|\mathcal{F}^{-1}\left(1+\xi^2\right)^{\frac{s}{2}}\mathcal{F}\left(\int_{\mathbb{R}^d}\delta_a f\,da\right)\right\|_{L^1(\mathbb{R}^d)}.
\end{aligned}$$

Moreover,

$$\begin{aligned}
&\left\|\mathcal{F}^{-1}\left(1+\xi^2\right)^{\frac{s}{2}}\mathcal{F}\left(\int_{\mathbb{R}^d}\delta_a f\,da\right)\right\|_{L^1(\mathbb{R}^d)}\\
&= \int_{\mathbb{R}^d}\left|\int_{\mathbb{R}^d} e^{i\langle y,\xi\rangle}\,\left(1+\xi^2\right)^{\frac{s}{2}}\int_{\mathbb{R}^d} e^{-i\langle a,\xi\rangle}f(a)\,d\xi\,da\right|\,dy\\
&\quad= \int_{\mathbb{R}^d}\left|\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} e^{i\langle y,\xi\rangle}\,\left(1+\xi^2\right)^{\frac{s}{2}} e^{-i\langle a,\xi\rangle}f(a)\,da\,d\xi\right|\,dy\\
&= \int_{\mathbb{R}^d}\left|\int_{\mathbb{R}^d} e^{i\langle y,\xi\rangle}\,\mathcal{F}\left[(I+\Delta)^{\frac{s}{2}}f\right](\xi)\,d\xi\right|dy = \int_{\mathbb{R}^d}\left(\mathcal{F}^{-1}\mathcal{F}\left[(I+\Delta)^{\frac{s}{2}}f\right]\right)(y)\,dy\\
&\le \left\|\mathcal{F}^{-1}\mathcal{F}\left[(I+\Delta)^{\frac{s}{2}}f\right]\right\|_{L^1(\mathbb{R}^d)} = \left\|(I+\Delta)^{\frac{s}{2}}f\right\|_{L^1(\mathbb{R}^d)} = \|f\|_{W_p^s(\mathbb{R}^d)}.
\end{aligned}$$

In summary, one can show that for any $s\in\mathbb{R}$ there exists some C such that

$$\left\|\int_{\mathbb{R}^d} f\delta_a\,da\right\|_{B_p^{s+\frac{d}{p}-d}(\mathbb{R}^d)} \le C\,\|f\|_{W_p^s(\mathbb{R}^d)}, \quad f\in C_b^0(\mathbb{R}^d).$$

Since the operator $I(s)$ is positive for $s\le 0$, it follows that

$$\int_{\mathbb{R}^d}\|f\delta_a\|_{B_p^{s+\frac{d}{p}-d}(\mathbb{R}^d)}\,da \le C\,\|f\|_{W_p^s(\mathbb{R}^d)}, \quad f\in C_b^0(\mathbb{R}^d),\ \ f\ge 0.$$

Theorem 5.1.2 in [BL76] gives inequality (A.1). □

REFERENCES

[ANZ98] E. J. Allen, S. J. Novosel, and Z. Zhang, *Finite element and difference approximation of some linear stochastic partial differential equations*, Stoch. Stoch. Rep., 64 (1998), pp. 117–142.

[AW00] D. Applebaum and J.-L. Wu, *Stochastic partial differential equations driven by Lévy space-time white noise*, Random Oper. Stochastic Equations, 8 (2000), pp. 245–259.

[AWZ98] S. Albeverio, J. Wu, and T. Zhang, *Parabolic SPDEs driven by Poisson white noise*, Stochastic Process. Appl., 74 (1998), pp. 21–36.

[BL76] J. Bergh and J. Löfström, *Interpolation Spaces: An Introduction*, Grundlehren Math. Wiss. 223, Springer-Verlag, New York, 1976.

[BS02] S. Brenner and R. Scott, *The Mathematical Theory of Finite Element Methods*, 2nd ed., Texts Appl. Math. 15, Springer-Verlag, New York, 2002.

[CWM04] C. Cardon-Weber and A. Millet, *On strongly Petrovskiĭ's parabolic SPDEs in arbitrary dimension and application to the stochastic Cahn–Hilliard equation*, J. Theoret. Probab., 17 (2004), pp. 1–49.

[EG04] Al. Ern and J. Guermond, *Theory and Practice of Finite Elements*, Appl. Math. Sci. 159, Springer-Verlag, New York, 2004.

[FSS01] H. Fujita, N. Saito, and T. Suzuki, *Operator Theory and Numerical Methods*, Stud. Math. Appl. 30, North–Holland, Amsterdam, 2001.

[Gar04] A. Gardoń, *The order of approximations for solutions of Itô-type stochastic differential equations with jumps*, Stochastic Anal. Appl., 22 (2004), pp. 679–699.

[Gyö98] I. Gyöngy, *Lattice approximations for stochastic quasi-linear parabolic partial differential equations driven by space-time white noise*. I, Potential Anal., 9 (1998), pp. 1–25.

[Gyö99] I. Gyöngy, *Lattice approximations for stochastic quasi-linear parabolic partial differential equations driven by space-time white noise*. II, Potential Anal., 11 (1999), pp. 1–37.

[GM05] I. Gyöngy and A. Millet, *On discretization schemes for stochastic evolution equations*, Potential Anal., 23 (2005), pp. 99–134.

[GP88] A. Germani and M. Piccioni, *Semidiscretization of stochastic partial differential equations on $\mathbf{R}^d$ by a finite-element technique*, Stochastics, 23 (1988), pp. 131–148.

[Hau02] E. Hausenblas, *Numerical analysis of semilinear stochastic evolution equations in Banach spaces,* J. Comput. Appl. Math., 147 (2002), pp. 485–516.

[Hau03] E. Hausenblas, *Approximation for semilinear stochastic evolution equations,* Potential Anal., 18 (2003), pp. 141–186.

[Hau05] E. Hausenblas, *Existence, uniqueness, and regularity of parabolic SPDEs driven by Poisson random measure*, Electron. J. Probab., 10 (2005), pp. 1496–1546.

[Hau06] E. Hausenblas, *SPDEs driven by Poisson random measure with non-Lipschitz coefficients*, Probab. Theory Related Fields, 137 (2007), pp. 161–200.

[HM06a] E. Hausenblas and I. Marchis, *The numerical approximation of parabolic stochastic partial differential equations driven by Poisson random measure*, BIT, 46 (2006), pp. 773–811.

[HM06b] E. Hausenblas and I. Marchis, *Numerical Simulations of Parabolic Stochastic Partial Differential Equations Driven by Poisson Random Measure*, manuscript, 2006.

[Hen81] D. Henry, *Geometric Theory of Semilinear Parabolic Equations*, Lecture Notes in Math. 840, Springer-Verlag, Berlin, 1981.

[KX95] G. Kallianpur and J. Xiong, *Diffusion approximation of nuclear space-valued stochastic-differential equations driven by Poisson random measures*, Ann. Appl. Probab., 5 (1995), pp. 493–517.

[KL02] M. Kouritzin and H. Long, *Convergence of Markov chain approximations to stochastic reaction-diffusion equations*, Ann. Appl. Probab., 12 (2002), pp. 1039–1070.

[KLS03] M. Kouritzin, H. Long, and W. Sun, *On Markov chain approximations to semilinear partial differential equations driven by Poisson measure noise*, Stochastic Anal. Appl., 21 (2003), pp. 419–441.

[KS01] P. E. Kloeden and S. Shott, *Linear-implicit strong schemes for Itô-Galerkin approximations of stochastic PDEs*, J. Appl. Math. Stoch. Anal., 14 (2001), pp. 47–53.

[Kuz01] D. F. Kuznetsov, *New representations of explicit one-step numerical methods for stochastic differential equations with a jump component*, Zh. Vychisl. Mat. Mat. Fiz., 41 (2001), pp. 922–937.

[KX87] G. Kallianpur and J. Xiong, *A nuclear space-valued stochastic differential equation driven by Poisson random measures*, in Stochastic Partial Differential Equations and Their Applications, Lecture Notes in Control and Inform. Sci. 176, Springer-Verlag, New York, 1987, pp. 135–143.

[KX96] G. Kallianpur and J. Xiong, *Stochastic Differential Equations in Infinite Dimensional Spaces*, Lecture Notes Monograph Series 26, Institute of Mathematical Statistics, Hayward, CA, 1996.

[Lin86] W. Linde, *Probability in Banach Spaces—Stable and Infinitely Divisible Distributions*, 2nd ed., Wiley-Interscience, New York, 1986.

[MM05] A. Millet and P. Morien, *On implicit and explicit discretization schemes for parabolic SPDEs in any dimension*, Stochastic Process. Appl., 115 (2005), pp. 1073–1106.

[Paz83] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Appl. Math. Sci. 44, Springer-Verlag, New York, 1983.

[PZ07] S. Peszat and J. Zabczyk, *Stochastic Partial Differential Equations with Lévy Noise*, Encyclopedia Math. Appl. 113, Cambridge University Press, Cambridge, 2007.

[Pri01a] J. Printems, *On the discretization in time of parabolic stochastic partial differential equations*, Monte Carlo Methods Appl., 7 (2001), pp. 359–368.

[Pri01b] J. Printems, *On the discretization in time of parabolic stochastic partial differential equations*, M2AN Math. Model. Numer. Anal., 35 (2001), pp. 1055–1078.

[Pro04] P. Protter, *Stochastic Integration and Differential Equations*, 2nd ed., Appl. Math. 21, Springer-Verlag, New York, 2004.

[PT97] P. Protter and D. Talay, *The Euler scheme for Lévy driven stochastic differential equations*, Ann. Probab., 25 (1997), pp. 393–423.

[Rub03] S. Rubenthaler, *Numerical simulation of the solution of a stochastic differential equation driven by a Lévy process*, Stochastic Process. Appl., 103 (2003), pp. 311–349.

[RS96] T. Runst and W. Sickel, *Sobolev Spaces of Fractional Order, Nemytskij Operators and Nonlinear Partial Differential Equations*, de Gruyter Ser. Nonlinear Anal. Appl. 3, de Gruyter, Berlin, 1996.

[SLB98] E. St. L. Bié, *Étude d'une EDPS conduite par un bruit Poissonnien. (Study of a stochastic partial differential equation driven by a Poisson noise)*, Probab. Theory Related Fields, 111 (1998), pp. 287–321.

[Sha99] T. Shardlow, *Numerical methods for stochastic parabolic PDEs*, Numer. Funct. Anal. Optim., 20 (1999), pp. 121–145.

[Wal86] J. Walsh, *An introduction to stochastic partial differential equations*, in Ecole d'Eté de Probabilités de Saint-Flour XIV - 1984, Lecture Notes in Math. 1180, Springer-Verlag, New York, 1986, pp. 265–437.

[Wal05] J. B. Walsh, *Finite element methods for parabolic stochastic PDE's*, Potential Anal., 23 (2005), pp. 1–43.

[Woy01] W. A. Woyczyński, *Lévy Processes in the Physical Sciences*, Lévy Processes, Birkhäuser Boston, Boston, MA, 2001, pp. 241–266.

[Yan05] Y. Yan, *Galerkin finite element methods for stochastic parabolic partial differential equations*, SIAM J. Numer. Anal., 43 (2005), pp. 1363–1384.

[Yoo00] H. Yoo, *Semi-discretization of stochastic partial differential equations on r^1 by a finite-difference method*, Math. Comp., 69 (2000), pp. 653–666.

SIAM J. NUMER. ANAL.
Vol. 46, No. 1, pp. 437–471
© 2008 Society for Industrial and Applied Mathematics

FINITE ELEMENT APPROXIMATION OF STOCHASTIC PARTIAL DIFFERENTIAL EQUATIONS DRIVEN BY POISSON RANDOM MEASURES OF JUMP TYPE*

ERIKA HAUSENBLAS†

Abstract. The paper deals with stochastic partial differential equations driven by Poisson random measures of jump type and their numerical approximation. We investigate the accuracy of space and time approximation. As space approximation we consider finite elements and as time approximation the implicit Euler scheme.

Key words. stochastic evolution equations, stochastic partial differential equations, numerical approximation, time discretization, space discretization, Poisson random measure

AMS subject classifications. 60H15, 60H35, 65C30, 35R30

DOI. 10.1137/050654141

1. Introduction.

Our point of interest is the numerical approximation solutions of stochastic partial differential equations (SPDEs) of the following type:

$$\begin{cases} du(t,\xi) = \frac{\partial^2}{\partial \xi^2} u(t-,\xi)\, dt + f(u(t-,\xi),\xi)\, dt \\ \qquad\qquad + \int_{\zeta\in\mathbb{R}} g(u(t-,\xi),\xi;\zeta)\, \tilde{\eta}(d\zeta,\xi,dt),\ \xi\in\mathcal{O},\ 0<t<\infty, \\ u(0,\xi) = u_0(\xi), \end{cases} \tag{1.1}$$

where $f : \mathbb{R}\times\mathcal{O}\to\mathbb{R}$ and $g : \mathbb{R}\times\mathbb{R}\to\mathbb{R}$ are global Lipschitz continuous functions and η is a space time Poisson random measure of impulsive type (see, e.g., Saint Loubert Bié [SLB98] or Peszat and Zabczyk [PZ07]).

Non-Gaussian random processes play an increasing rôle in modeling stochastic dynamical systems. Typical examples of non-Gaussian stochastic processes are Lévy processes and processes arising by Poisson random measures. In neurophysiology the driving noise of the cable equation is basically impulsive, e.g., of a Poisson type (see Walsh [Wal86, Chapter 3])—or, on the other hand, Woyczyński describes in [Woy01] a number of phenomena from fluid mechanics, solid state physics, polymer chemistry, economic science, etc., for which non-Gaussian Lévy processes can be used as their mathematical model in describing the related stochastic behavior. Thus, from the point of view of applications one might feel that the restriction to Gaussian noise is unsatisfactory; to handle such cases one can replace the Gaussian noise by a Poisson random measure.

Concerning the numerical approximation of SPDEs driven by Wiener noise, several works already exist. Just to mention a few (see, e.g., Gyöngy [Gyö99, Gyö98]), Hausenblas [Hau03, Hau02], Kloeden and Shott [KS01], Printems [Pri01a, Pri01b], Shardlow [Sha99], and Yoo [Yoo00]. Finite element approximation of SPDEs driven by Wiener space time white noise were considered by Allen, Novosel, and Zhang

*Received by the editors March 13, 2006; accepted for publication (in revised form) June 14, 2007; published electronically January 30, 2008. This research was supported by Austrian Science Foundation grant P17273.

http://www.siam.org/journals/sinum/46-1/65414.html

†Department of Mathematics, University of Salzburg, Hellbrunnerstr. 34, 5020 Salzburg, Austria (erika.hausenblas@sbg.ac.at).

[ANZ98], Cardon-Weber and Millet [CWM04], Germani and Piccioni [GP88], Gyöngy [Gyö99, Gyö98], Gyöngy and Millet [GM05], Millet and Morien [MM05], Walsh [Wal05], and Yan [Yan05]. The numerical approximation of finite-dimensional SDEs driven by Poisson random measures is the topic of, e.g., the work of Gardoń [Gar04], Kuznetsov [Kuz01], Protter and Talay [PT97], or Rubenthaler [Rub03]. But the numerical approximation of SPDEs driven by Poisson random measures is barely investigated, and, in the cases known to the author, processes of finite variation are considered, e.g., by Kouritzin, Long, and Sun [KLS03] and Kouritzin and Long [KL02]. If the driving process is of infinite variation, there exists only some works about existence and uniqueness, e.g., Albeverio, Wu, and Zhang [AWZ98], Hausenblas [Hau05, Hau06], and Kallianpur and Xiong [KX87, KX96]. Applebaum and Wu [AW00] and Saint Loubert Bié [SLB98] considered SPDEs with space time Lévy noise.

First, we discretize the SPDE spatially by the finite element method, obtaining a system of high-dimensional SDEs that can be solved by the implicit Euler scheme. The noise will be approximated by using the projection onto finite elements.

2. Preliminaries. Let $\mathcal{O} \subset \mathbb{R}^d$ be a convex polygon, $1 < p \leq 2$, and A be a uniformly elliptic, dissipative, second order partial differential operator with either Dirichlet or Neumann boundary conditions. In particular, let $A : D(A) \subset L^p(\mathcal{O}) \to L^p(\mathcal{O})$ be the realization of

$$\begin{aligned} \mathcal{L} &= \mathcal{L}(x, D) \\ &= -\sum_{i,j=1}^{2} \frac{\partial}{\partial x_i} a_{i,j}(x) \frac{\partial}{\partial x_j} + \sum_{i=1}^{2} b_i(x) \frac{\partial}{\partial x_i} + c(x), \end{aligned} \tag{2.1}$$

with either

$$u = 0 \quad \text{on} \quad \partial\mathcal{O} \times (0, \infty)$$

or

$$\frac{\partial u}{\partial n_L} + \sigma u = 0,$$

where $\frac{\partial u}{\partial n_L}$ denotes the differentiation along the outer normal unit vector n on $\mathcal{O}$. We assume that $a_{ij} \in C^2(\mathcal{O})$, $1 \leq i, j \leq d$, strictly bounded away from zero, and $\sigma \in C^2(\mathcal{O})$.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a complete probability space. Let $\eta : \mathcal{B}(\mathbb{R}) \hat{\times} \mathcal{B}(\mathcal{O}) \hat{\times} \mathcal{B}(\mathbb{R}^+) \to \mathbb{R}^+$ be a Poisson random measure over $(\Omega, \mathcal{F}, \mathbb{P})$ with intensity[1] $\iota = \lambda \hat{\times} \nu$, $\nu \in \mathcal{L}(\mathbb{R}) := \{\sigma : \mathcal{B}(\mathbb{R}) \to \mathbb{R}^+, \sigma \text{ is a Lévy measure}\}$, γ be the compensator, and $\tilde{\eta} = \eta - \gamma$. Let $(\mathcal{F}_t)_{t \geq 0}$ be the natural filtration induced by η.

Our point of interest is solutions of SPDEs of the following form:

$$\begin{cases} du(t, \xi) = Au(t, \xi)\, dt + f(u(t-, \xi), \xi)\, dt \\ \qquad\qquad + \int_{\zeta \in \mathbb{R}} g(u(t-, \xi), \xi; \zeta)\, \tilde{\eta}(d\zeta, \xi, dt), \ \xi \in \mathcal{O}, \ t > 0, \\ u(0, \xi) = u_0(\xi), \quad \xi \in \mathcal{O}, \end{cases} \tag{2.2}$$

where $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \times \mathcal{O} \times \mathbb{R} \to \mathbb{R}$ are globally Lipschitz continuous functions. By a mild solution of (1.1) we understand a process u such that

[1] λ denotes the Lebesgue measure.

- u satisfies the integral equation

$$u(t,\xi) = \int_{\mathcal{O}} G_t(\zeta,\xi)u_0(\zeta)\,d\zeta + \int_{0+}^{t}\int_{\mathcal{O}} G_{t-s}(\zeta,\xi)f(u(s-,\zeta,\zeta))\,d\zeta\,ds + \int_{0+}^{t}\int_{\zeta\in\mathcal{O}}\int_{z\in\mathbb{R}} G_{t-s}(\zeta,\xi)g(u(s-,\zeta),\zeta;z)\,\eta(dz,d\zeta,ds),\ \text{a.s.},\quad t\geq 0,$$

  where $G = \{G_t(\cdot,\star), 0 < t < \infty\}$ is the Green kernel associated with $\frac{\partial}{\partial t} + \mathcal{L}$ and the boundary conditions specified above;
- for all $\phi \in C_b^\infty(\mathcal{O})$, the $\mathbb{R}$-valued process $t \mapsto \langle u(t), \phi\rangle$ is adapted and cádlág.

It is shown in Saint Loubert Bié [SLB98] or Hausenblas [Hau05, Example 2.3] that if $p < \frac{2}{d} + 1$, then there exists a unique solution of problem (1.1) such that $u \in L^p(\Omega; L^p(\mathcal{O}))$.

Our point of interest is the numerical approximation of the solution of (1.1). First, we introduce the space discretization by finite elements, and second we introduce the time discretization, i.e., the Euler scheme.

2.1. Space discretization. The space discretization is done by finite elements. For the definition for a finite element, we follow the book of Brenner and Scott [BS02].

DEFINITION 2.1. *Let*

- $K \subset \mathbb{R}^d$ *be a bounded closed set with nonempty interior and piecewise smooth boundary (the* element domain*);*
- $\mathcal{P}$ *be a finite-dimensional space of functions on* K *(the space of* shape functions*); and*
- $\mathcal{N} = \{N_1, \ldots, N_k\}$ *be a basis for* $\mathcal{P}'$ *(the set of* nodal variables*).*

Then $(K, \mathcal{P}, \mathcal{N})$ *is called a* finite element*. Moreover, we denote by* $diam(K) = \max_{x,y\in K} |x-y|_{\mathbb{R}^d}$ *the diameter of* K*.*

DEFINITION 2.2. *A set* K *is* star-shaped with respect to B*, if for all* $x \in K$ *the closed convex hull of* $\{x\} \cup B$ *is a subset of* K*.*

DEFINITION 2.3. *Suppose* K *has diameter* d *and is star-shaped with respect to a ball. Let* $\rho_{max} = \sup\{\rho : K \text{ is star-shaped with respect to a ball of radius } \rho\}$*. Then the* chunkiness parameter *of* K *is defined by*

$$\gamma := \frac{d}{\rho_{max}}.$$

Moreover, we denote by P_m the set of all polynomials in d variables of degree less than or equal to m.

DEFINITION 2.4 (compare the assumptions of Theorem 4.4.4 in [BS02]). *We assume that a finite element satisfies the* standard conditions for some l and m*, iff*

- K *is star-shaped with respect to some ball,*
- $P_{m-1} \subset \mathcal{P} \subset W_\infty^m(\mathcal{O})$*,*
- $\mathcal{N} \subset (C^l(\bar{K}))'$*.*

DEFINITION 2.5. *Let* $(K, \mathcal{P}, \mathcal{N})$ *be a finite element. The basis* $\{\phi_1, \ldots, \phi_k\}$ *of* $\mathcal{P}$ *dual to* $\mathcal{N}$ *is called the* nodal basis *of* $\mathcal{P}$*.*

DEFINITION 2.6. *Given a finite element* $(K, \mathcal{P}, \mathcal{N})$*, let the set* $\{\phi_1, \ldots, \phi_k\} \subset \mathcal{P}$ *be the basis dual to* $\mathcal{N}$*. If* v *is a function for which all* $N_i \in \mathcal{N}$*,* $i = 1, \ldots, k$*, are defined, then we define the* local interpolant *by*

$$\mathcal{I}_K v := \sum_{i=1}^{k} v(N_i)\phi_i.$$

DEFINITION 2.7. *A* subdivision *of a domain $\mathcal{O}$ is a finite collection of element domains $\{K_i, i = 1, \dots k\}$ such that*

- $\int K_i \cap \int K_j = \emptyset$ *if* $i \neq j$ *and*
- $\cup K_i = \mathcal{O}$.

Remark 2.1. The nodal basis of a subdivision $\{K_i, i = 1, \dots k\}$ induces a Voronoi decomposition $\{J_i; i = 1, \dots, k\}$ of $\mathcal{O}$ by

$$J_i := \left\{ \xi \in \mathcal{O},\ |\xi - N_i| \leq \min_{1 \leq i \leq k} \{|\xi - N_j|\} \right\}.$$

DEFINITION 2.8. *Suppose that $\mathcal{O}$ is a polyhedral domain with subdivision $\mathcal{T}$. Assume that each element domain K in the subdivision is equipped with some type of shape functions $\mathcal{P}$ and nodal variables $\mathcal{N}$ such that $(K, \mathcal{P}, \mathcal{N})$ forms a finite element. Let m be the order of the highest partial derivatives involved in the nodal variables. For $f \in C^m(\bar{\mathcal{O}})$, the* global interpolant *is defined by*

$$\mathcal{I}_{\mathcal{T}} f \Big|_{K_i} = \mathcal{I}_{K_i} f$$

for all $K_i \in \mathcal{T}$.

Example 2.1. Let $H = L^2([0,1])$. Let $\{x_i := \frac{i}{N}, i = 1, \dots, N-1\}$ be the set of grid points. For $0 \leq i \leq N$ let ϕ_i be a pyramid function, i.e., a function, which takes value 1 at the grid point $\frac{i}{N}$, vanishes at the other grid points, and is linear between the grid points. For $i = 1, \dots, N-1$, a finite element $(K_i, \mathcal{N}_i, \mathcal{P}_i)$ consists of the interval $[\frac{i-1}{N}, \frac{i}{N}]$, the nodal variables $N_i : \mathcal{C}^0([0,1]) \ni v \mapsto v(x_i) \in \mathbb{R}$, the shape functions $\phi_i|_{K_i}$ and $\phi_{i-1}|_{K_i}$, and $J_i = [\frac{i-1}{N} + \frac{1}{2N}, \frac{i}{N} + \frac{1}{2N}]$.

We can now define a family of subdivisions $\mathcal{F}_h$, $h = N_h^{-1}$. For each subdivision $\mathcal{T}_h$, let $\mathcal{V}_h$ the linear hull of the shape functions, i.e., the linear hull of $\{\phi_i\}_{i=1}^{N_h - 1}$. For any $v \in C([0,1])$, the global interpolant is given by

$$\mathcal{I}_h(v) = \sum_{i=1}^{N_h} N_i^h(v) \phi_i.$$

DEFINITION 2.9. *We call a family of subdivisions $\{\mathcal{T}_h, 0 < h \leq 1\}$* nondegenerate *if the chunkiness parameter is uniformly bounded for all $T \in \{\mathcal{T}_h\}$ and all $h \in (0,1]$.*

Let $\{\mathcal{T}_h, 0 < h \leq 1\}$ be a nondegenerate family of subdivisions of a polyhedral domain $\mathcal{O}$ in $\mathbb{R}^d$, consisting of k_h finite elements of reference type $(K, \mathcal{P}, \mathcal{N})$. We assume that the parameter h corresponds to the maximal diameter of the finite elements, i.e., $h = \max\{1 \leq i \leq k_h \mid \operatorname{diam}(K_i)\}$. Let $\{\phi_1^h, \dots, \phi_{k_h}^h\}$ be the set of shape functions and $\{N_1^h, \dots, N_{k_h}^h\}$ be the set of nodal variables. We denote by $\mathcal{V}_h$ the linear hull of the shape functions and $\mathcal{V}_h'$ its dual. The semidiscrete problem corresponding to (2.2) is to find the process $u_h = \{u_h(t) \in \mathbb{R}^{k_h}, 0 \leq t < \infty\}$

$$\begin{cases} du_h(t) = \mathcal{A}_h u_h(t)\, dt f_h(u_h(t-))\, dt \\ \qquad\qquad + \int_{i \in \{1, \dots, k_h\}} \int_{\zeta \in \mathbb{R}} g_h(u_h(t-), i, \zeta)\, \tilde{\eta}_h(d\zeta, di, dt),\ t > 0, \\ \qquad u(0) = (\langle u_0, N_1 \rangle, \dots, \langle u_0, N_{k_h} \rangle) \in \mathbb{R}^{k_h} \end{cases} \tag{2.3}$$

where $\mathcal{A}_h : \mathcal{V}_h \to \mathcal{V}_h$ is the discrete analogue of A defined by

$$\langle \mathcal{A}_h \phi, \chi \rangle = \langle A\phi, \chi \rangle \quad \forall \phi, \chi \in \mathcal{V}_h, \tag{2.4}$$

and $\tilde{\eta}_h : \mathcal{B}(\mathbb{R})\hat{\times}\mathcal{P}(\{1,\dots,k_h\})\hat{\times}\mathcal{B}(\mathbb{R}^+) \to \mathbb{R}^+$ is the compensated Poisson random measure with intensity ι_h given by

$$\mathcal{B}(\mathbb{R}) \times \mathcal{B}(\{1,\dots,k_h\}) \ni (A,B) \mapsto \iota_h((A,B)) := \cup_{i\in B}\lambda(J_i^h)\,\nu(A) \in \mathbb{R}^+.$$

Moreover, $f_h : \mathcal{V}_h \to \mathcal{V}_h$ and $g_h : \mathcal{V}_h \times \{1,\dots k_h\} \times \mathbb{R} \to \mathcal{V}_h$ are given by

$$\mathcal{V}_h \ni v \mapsto f_h(v) := \sum_{i=1}^{k_h} f(v(N_i), N_i)\phi_i^h \in \mathcal{V}_h$$

and

$$\mathcal{V}_h \times \{1,\dots k_h\} \times \mathbb{R} \ni (v,i,\zeta) \mapsto g_h(v,i,\zeta) = g(v(N_i), N_i;\zeta)\, c_i^h\ \phi_i^h \in \mathcal{V}_h,$$

where $c_i^h := \left[\int_{\mathcal{O}} \phi_i^h(\xi)\, d\xi\right]^{-1}$.

Remark 2.2. The Poisson random measure $\tilde{\eta}_h$ can also be written as a sum of k_h independent Poisson random measures (see, e.g., [Lin86, Proposition 5.3.2]). To be precise, let $\tilde{\eta}_h^i : \mathcal{B}(\mathbb{R})\hat{\times}\mathcal{B}(\mathbb{R}^+) \to \mathbb{R}^+$ be pairwise independent compensated Poisson random measures with intensity $\nu^i = \lambda(J_i^h)\,\nu$. Then we have for some $A \in \mathcal{B}(\{1,\dots,k_h\})$ and $B \in \mathcal{B}(\mathbb{R})$, $\nu(B) < \infty$,

$$\int_{i\in A}\int_{\xi\in B} \iota_h(di,d\xi) \overset{\mathrm{d}}{=} \sum_{i=1}^{k_h} 1_A(i)\lambda(J_i^h)\int_B \nu^i(d\xi).$$

In particular, we have for all $\mathcal{V}_h$-valued $\mathcal{F}_t$-adapted and cádlág processes $u = \{u(t), 0 \le t < \infty\}$

$$\int_0^t \int_{\{1,\dots,k_h\}} \int_{\zeta\in\mathbb{R}} g_h(u(s-),di,\zeta)\ \tilde{\eta}_h(di,d\zeta,ds) \overset{\mathrm{d}}{=} \sum_{i=1}^{k_h}\int_0^t\int_{\zeta\in\mathbb{R}} g_h(u(s-),i,\zeta)\tilde{\eta}_h^i(d\zeta,ds).$$

2.2. Time discretization. One usually prefers to discretize simultaneously in time and space. Thus, let τ_h be the time step size corresponding to the subdivision $\mathcal{T}_h$. One popular way for time discretization is to use the *implicit Euler scheme*, i.e.,

$$\frac{u_h(t+\tau_h)-u_h(t)}{\tau_h} + \mathcal{A}_h u_h(t+\tau_h) \approx f_h(u_h(t)) + \int_t^{t+\tau_h}\int_{z\in\mathbb{R}^{k_h}} g_h(u_h(t);z)\tilde{\eta}_h(dz,dt).$$

So, if v_h^m denotes the approximation of $u_h(m\tau_h)$, then

$$\begin{cases} v_h^{m+1} = (1-\tau_h A_h)^{-1}\left[v_h^m + \tau_h f_h(m\tau_h, v_h^m) + \xi_h^m(u_h(m\tau_h))\right], \\ v_h^0 = P_h u_0, \end{cases} \tag{2.5}$$

where we denote by $\xi_h^m(u_h(m\tau_h))$ the k_h dimensional random variable

$$\int_{m\tau_h}^{(m+1)\tau_h}\int_{\zeta\in\mathbb{R}^{k_h}} g_h(u_h(m\tau_h);\zeta)\,\tilde{\eta}_h(dz;ds). \tag{2.6}$$

The approximation at time $t = \tau_h m$, $m \in \mathbb{N}$, is given by

$$\hat{u}_h(m\tau) := \mathcal{I}_h v_h^m. \tag{2.7}$$

The solution is defined by linear interpolation; let $t \in (m\tau_h, (m+1)\tau_h)$; we define $v_h(t)$ by $v_h(t) = v_h^m$ and

$$\hat{u}_h(t) := \mathcal{I}_h\, v_h(t), \quad m\tau_h < t < (m+1)\tau_h.$$

Since the implicit Euler scheme is unconditionally stable, no stability conditions arise. Now we can state our main result.

THEOREM 2.1. *Let $\mathcal{O}$ be a polyhedral domain and A be a uniformly elliptic, second order differential operator given in* (2.1) *with Dirichlet or Neumann boundary conditions.*

For a fixed p, $1 \le p \le 2$, let $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\ge 0}, P)$ be the usual filtration and

$$\eta : \mathcal{O} \times \mathbb{R} \times \mathcal{B}(\mathbb{R}^+) \to \mathbb{R}^+$$

be a Poisson random measure of jump type with intensity $\iota := \lambda \times \nu$, where ν belongs to the set of p-integrable Lévy measures on $\mathbb{R}$.

We assume that the functions $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \times \bar{\mathcal{O}} \times \mathbb{R} \to \mathbb{R}$ are globally Lipschitz and g linear in the third variable. Moreover, we assume that the initial condition u_0 belongs to $L^p(\mathcal{O})$.

The space is approximated by a nondegenerate family of subdivisions $\{\mathcal{T}_h, 0 < h \le 1\}$ of finite elements, satisfying the conditions described in section 2.1. *Moreover, the reference element is supposed to be a standard element for $l \ge 0$ and $m = 2$. Let v_n^k, $k \in \mathbb{N}$, be the approximation of* (2.2) *given by* (2.7). *Then for all $N \in \mathbb{N}$, for all α such that*

$$\alpha < \frac{2}{p} - d + \frac{d}{p},$$

and for all δ_1, δ_2, and δ_3, with

$$\delta_1 = 2 + \frac{d}{2} - \frac{d}{p}, \quad \delta_2 < \frac{2}{p} - d + \frac{d}{p}, \text{ and } \delta_3 < \left(\frac{2}{p} - d + \frac{d}{p}\right)\left(2 + \frac{d}{2} - \frac{d}{p}\right),$$

there exist constants C_0, C_1, C_2, and C_3 depending on p, f, and g such that for all $0 < h \le 1$ the following estimate holds for $0 < k \le N$:

$$\left[\mathbb{E}\ \left\|u(k\tau_h) - v_h^k\right\|_{L^p(\mathcal{O})}^p\right]^{\frac{1}{p}} \le \frac{C_0}{\epsilon}\,\tau^\alpha + \frac{C_1}{(k\tau_h)^{\frac{1}{p}}}\,h^{\delta_1} + \frac{C_2}{\epsilon_2}\,h^{\delta_2} + \frac{C_3}{\epsilon_3}\,h^{\delta_3},$$

where $\epsilon_1 = \frac{2}{p} - d + \frac{d}{p} - \alpha$, $\epsilon_2 = \frac{2}{p} - d + \frac{d}{p} - \delta_2$ and $\epsilon_3 = \frac{2}{p} - d + \frac{d}{p} - \frac{\delta_3}{\delta_1}$,

In [Hau05, Example 2.3] we treated (2.2) by means of $L^p(\mathcal{O})$ spaces. But finite elements are designed for Hilbert spaces. In particular, the approximation of the operator A is given by the inner product (see problem 3.4), and usually the error estimates of the corresponding semigroups are given in $L^2(\mathcal{O})$ or $H^1(\mathcal{O})$ (see Theorem 3.3). Since the $L^p(\mathcal{O})$ spaces were essential in [Hau05]—on which this paper relies—we had to express the error estimates of the corresponding semigroups in $L^p(\mathcal{O})$ spaces. We have done this in Theorem 3.4. Therefore, the entity $2 + \frac{d}{2} - \frac{d}{p}$ appears.

Fix $p \in [1,2]$. Let E be a separable M-type p Banach space and Z be a Banach space. Let A be an operator generating a compact and analytic semigroup, $\nu \in \mathcal{L}(Z)$, $\tilde{\eta}$ be a compensated Poisson random measure on Z with intensity ν, and $G : E \to L(Z,E)$ be a mapping such that there exist some $\gamma < \frac{2}{p}$ and $C < \infty$, with

$$\int_Z \left| (-A)^{-\frac{\gamma}{2}} \left(G(u,z) - G(v,z) \right) \right|_E^p \nu(dz) \leq C \, |u-v|_E^p. \tag{2.8}$$

Then, by [Hau05], there exists a unique solution to the equation

$$\begin{cases} du(t) = (Au(t-)) \; dt + \int_Z g(u(t-);z) \tilde{\eta}(dz;dt), \\ u(0) = u_0. \end{cases} \tag{2.9}$$

In [Hau05, Example 2.3], the space time Poissonian noise was represented as a functional-valued Poisson random measure taking values in the Besov space $B_p^{\frac{d}{p}-d}(\mathcal{O})$. Let $\nu \in \mathcal{L}(E)$ and $1 \leq p \leq 2$. Let $g : \mathbb{R} \times \mathcal{O} \times \mathbb{R} \to \mathbb{R}$ be Lipschitz continuous in the following sense: There exists some constant $C < \infty$ such that

$$\int_{\mathcal{O}} \int_{\mathbb{R}} |g(x,\xi,\zeta) - g(y,\xi,\zeta)|^p \nu(d\zeta) d\xi \leq C |x-y|^p, \quad x,y \in \mathbb{R}. \tag{2.10}$$

Thus, if

$$\frac{d}{p} - d < \frac{2}{p},$$

then the Nemytskij operator G of g, defined in (3.20), satisfies by Proposition A.1 the Lipschitz condition (2.8). The remaining difference

$$\frac{2}{p} - d + \frac{d}{p}$$

is used for estimating the error of the approximation. Therefore, the entity $\frac{2}{p} - d + \frac{d}{p}$ appears.

Again, let $\nu \in \mathcal{L}(E)$, $1 \leq p_0 \leq 2$ and $g : \mathbb{R} \times \mathcal{O} \times \mathbb{R} \to \mathbb{R}$ be Lipschitz continuous satisfying (2.10), with $p = p_0$. Moreover, let $f = 0$. Then, if

$$\frac{2}{p_0} - d + \frac{d}{p_0} > 0, \tag{2.11}$$

problem (1.1) has a unique solution such that $u \in L^p(\Omega; L^p(\mathcal{O}))$ for all $p \in [p_0, p_1)$, where $p_1 = \frac{2}{d} + 1$ (compare [Hau05, Example 2.3]). Moreover, the assumptions of Theorem 2.1 are fulfilled for all $p \in [p_0, p_1)$, and we can choose any $p \in [p_0, p_1)$ such that $\min(\delta_1, \delta_2, \delta_3)$ is maximal. If

$$2 + \frac{d}{2} - \frac{d}{p_1} < 0,$$

then no convergence is given by Theorem 2.1.

3. Finite elements—a short account. In the existing literature several error bounds for the approximation by finite elements can be found. In this section, we will cite the estimates which are necessary to prove our main result.

In this chapter, let $\{\mathcal{T}_h, 0 < h \leq 1\}$ be a nondegenerate family of subdivisions of a polyhedral domain $\mathcal{O}$ in $\mathbb{R}^d$, consisting of k_h finite elements of reference type $(K, \mathcal{S}, \mathcal{N})$ satisfying the standard conditions for some m and l. We assume that the parameter h corresponds to the maximal diameter of the finite elements, i.e., $h = \max\{1 \leq i \leq N \mid \operatorname{diam}(K_i)\}$. Moreover, we assume throughout this section, that $(K, \mathcal{P}, \mathcal{N})$ satisfies the standard conditions for some l and m.

Let $\{\phi_1^h, \dots, \phi_{k_h}^h\}$ be the set of shape functions and $\{N_1^h, \dots, N_k^h\}$ be the set of nodal variables. By $\bar{\mathcal{N}}$ we denote the set of nodal variables which are inner points. We denote by $\mathcal{V}_h$ the linear hull of the shape functions and by $\mathcal{V}_h'$ its dual.

The quality of approximation is given by the following Bramble–Hilbert (Raviart–Ciarlet) inequality.

THEOREM 3.1 (see, e.g., Brenner and Scott [BS02, Corollary 4.4.24]). *Assume that the conditions above are satisfied. Suppose* $1 \leq p \leq \infty$ *and either* $m - l - \frac{d}{p} > 0$ *when* $p > 1$ *or* $m - l - d \geq 0$ *when* $p = 1$. *Then there exists a positive constant* C *depending on the reference element,* m, l, d, p, *and the chunkiness parameter such that for* $0 \leq s \leq m$

$$\left(\sum_{T \in \mathcal{T}_h} \|v - \mathcal{I}_h v\|_{W_p^s(T)}^p \right)^{\frac{1}{p}} \leq C\, h^{m-s}\, \|v\|_{W_p^m(\mathcal{O})}$$

for all $v \in W_p^m(\mathcal{O})$, *where the left-hand side should be interpreted, in the case where* $p = \infty$, *as* $\max_{T \in \mathcal{T}_h} \|v - \mathcal{I}_h v\|_{W_\infty^s(\mathcal{O})}$. *For* $0 \leq s \leq l$,

$$\max_{T \in \mathcal{T}_h} \|v - \mathcal{I}_h v\|_{W_\infty^s(\mathcal{O})} \leq C\, h^{m-s-\frac{d}{p}}\, \|v\|_{W_p^m(\mathcal{O})} \quad \forall v \in W_p^m(\mathcal{O}).$$

Since the space $\mathcal{V}_h$ is finite-dimensional, all norms which can be defined in $\mathcal{V}_h$ are equivalent. The purpose of the inverse inequalities is to specify how the equivalent constants depend on the parameter h.

THEOREM 3.2 (Brenner and Scott [BS02, Theorem 4.5.11]). *Assume that the conditions above are satisfied, and let* p, q, *and* l *be given constants such that* $1 \leq p \leq \infty$, $1 \leq q \leq \infty$, $0 \leq m \leq l$, *and* $\phi_i^h \in W_p^l(\mathcal{O}) \cap W_q^m(\mathcal{O})$ *for* $i = 1, \dots, k_h$. *Then there exists a positive constant* C *depending on the reference element,* d, m, p, *and* q *such that*

$$\left[\sum_{T \in \mathcal{T}_h} \|v\|_{W_p^l(T)}^p \right]^{\frac{1}{p}} \leq C\, h^{m-l+\min(0, \frac{d}{p} - \frac{d}{q})} \left[\sum_{T \in \mathcal{T}_h} \|v\|_{W_q^m(T)}^q \right]^{\frac{1}{q}}$$

for all $v \in \mathcal{V}_h$. *When* $p = \infty$, *then* $[\sum_{T \in \mathcal{T}_h} \|v\|_{W_p^s(T)}^p]^{\frac{1}{p}}$ *is interpreted as* $\max_{T \in \mathcal{T}_h} \|v\|_{W_\infty^l(T)}^q$.

COROLLARY 3.1. *Under the conditions above, there exists a positive constant* C *depending on the reference element such that we have for all* $1 > \delta \geq 0$

$$\left[\sum_{T \in \mathcal{T}_h} \|v\|_{L_p(T)}^p \right]^{\frac{1}{p}} \leq C\, h^{-\delta+\min(0, \frac{d}{p} - \frac{d}{q})} \left[\sum_{T \in \mathcal{T}_h} \|v\|_{W_q^{-\delta}(T)}^q \right]^{\frac{1}{q}}$$

for all $v \in \mathcal{V}_h$.

Proof. The proof is done similarly to the proof of Theorem 3.2 (see Brenner and Scott [BS02, Theorem 4.5.11]). Let $(\hat{K}, \hat{\mathcal{P}}, \hat{\mathcal{N}})$ be a finite element of reference type satisfying the standard conditions for m and l with $\operatorname{diam}(\hat{K}) = 1$ and $(K, \mathcal{P}, \mathcal{N})$ be a finite element of the family of subdivisions with $\operatorname{diam}(K) = h_K$. Since the subdivision is nondegenerate, a transformation exists, having the form

$$T_K : \hat{K} \ni \hat{x} \mapsto J_K \hat{x} + b_K \in K,$$

where $J_K \in \mathbb{R}^d \times \mathbb{R}^d$ being invertible and $b_K \in \mathbb{R}^d$ (see Proposition 4.4.11 in [BS02]). Let $\hat{v} : \hat{K} \to \mathbb{R}$ be an arbitrary function and $v : K \to \mathbb{R}$ be defined by

$$K \ni x \mapsto v(x) := \hat{v}(J_K^{-1} x) \in \mathbb{R}.$$

Then substitution gives

$$|v|_{L^p(K)} = \det(J_K)^{\frac{1}{p}} |\hat{v}|_{L^p(\hat{K})} = C\, h_K^{\frac{d}{p}} |\hat{v}|_{L^p(\hat{K})}, \quad 0 < h_K \le 1,$$

and

$$|v|_{L^\infty(K)} = |\hat{v}|_{L^\infty(\hat{K})}, \quad 0 < h_K \le 1.$$

Similarly there exists a constant $C > 0$ such that we have for all $0 < h_K \le 1$

$$|v|_{W_p^j(K)} = C\, \det(J_K)^{\frac{1}{p}} h_K^{-j} |\hat{v}|_{W_p^j(\hat{K})} = C\, h_K^{\frac{d}{p}-j} |\hat{v}|_{W_p^j(\hat{K})}$$

and

$$|v|_{W_\infty^j(K)} = C\, h_K^{-j} |\hat{v}|_{W_\infty^j(\hat{K})} = C\, h_K^{-j} |\hat{v}|_{W_\infty^j(\hat{K})}.$$

Duality arguments give for q conjugate to p and for all $0 < h_K \le 1$

$$\begin{aligned} |v|_{W_p^{-j}(K)} &= \sup_{w \in W_q^j(K)} \frac{\langle v, w \rangle}{|w|_{W_q^j(K)}} \le C \sup_{\hat{w} \in W_q^j(\hat{K})} \frac{\det(J_K) \langle \hat{v}, \hat{w} \rangle}{h_K^{\frac{d}{q}-j} |\hat{w}|_{W_q^j(\hat{K})}} \\ &= C\, h_K^{d(1-\frac{1}{q})+j} \sup_{\hat{w} \in W_q^j(K)} \frac{\langle \hat{v}, \hat{w} \rangle}{|\hat{w}|_{W_q^j(K)}} \le C\, h_K^{\frac{d}{p}+j} |\hat{v}|_{W_p^{-j}(\hat{K})}. \end{aligned} \tag{3.1}$$

Since $\hat{\mathcal{P}}$ is finite-dimensional, all norms are equivalent, and there exists a constant $\hat{C}$ such that

$$|\hat{v}|_{L^p(\hat{K})} \le \hat{C}\, |\hat{v}|_{W_q^{-\gamma}(\hat{K})} \quad \forall v \in \mathcal{P}.$$

Therefore (3.1) and interpolation arguments imply for $\gamma > 0$ that

$$h_K^{-\frac{d}{p}} |v|_{L^p(K)} \le C\, h_K^{-\frac{d}{q}-\gamma} |v|_{W_q^{-\gamma}(K)} \quad \forall v \in \mathcal{P},\ 0 < h_K \le 1.$$

Let $\{\mathcal{T}_h, 0 < h \leq 1\}$ be a nondegenerate family of subdivisions. Then, by taking the sum, one can show by the same arguments as in Theorem 4.5.11 in [BS02], that, if the family of subdivisions is nondegenerate, the assertion of Corollary 3.1 holds. □

3.1. Approximation of the operator. We can associate with the operator $(A, D(A))$ the Gelfand triple (V, H, V'), where $V \hookrightarrow H = H' \hookrightarrow V'$ densely. In fact, $V = D(-A)^{\frac{1}{2}}$ and V' its dual. In our case, where $(A, D(A))$ is a second order differential operator with boundary conditions, we have $V = H^1(\mathcal{O})$ if the boundary condition involves first order derivatives and $V = H_0^1(\mathcal{O})$ if the homogeneous Dirichlet condition is specified on the whole boundary. Now the operator $A : V \to V'$ defines a bilinear form on H by

$$a(u, v) := \langle Au, v\rangle, \quad u, v \in V.$$

The corresponding symmetric variational problem is posed as follows. Suppose that the following three conditions are valid:

$$\text{(3.2)} \qquad \begin{cases} (1) & (H, \langle\cdot,\cdot\rangle) \text{ is a Hilbert space.} \\ (2) & V \text{ is a (closed) subspace of } H. \\ (3) & a(\cdot,\cdot) \text{ is a } \textit{bounded, symmetric} \text{ bilinear form} \\ & \text{that is } \textit{coercive} \text{ on } V. \end{cases}$$

Then the *symmetric variational* problem is the following:

$$\text{(3.3)} \qquad \begin{cases} \text{For a given } F \in V', \text{ find } u \in V \text{ such that} \\ \quad a(u, v) = F(v) \quad \forall v \in V. \end{cases}$$

Given the family of nondegenerate subdivisions $\mathcal{T}_h$ of a polyhedral domain $\mathcal{O}$, the *Ritz–Galerkin approximation* of problem (3.3) is the following:

$$\text{(3.4)} \qquad \begin{cases} \text{For a given } F \in \mathcal{V}_h', \text{ find } u_h \in \mathcal{V}_h \text{ such that} \\ \quad a(u_h, v) = F(v) \quad \forall v \in \mathcal{V}_h. \end{cases}$$

Assume that u solves problem (3.3) and u_h solves problem (3.4). Then we can find the following estimates on $u - u_h$ depending on F. By means of the inequality above and duality arguments, one can verify bounds of the difference $u - u_h$ in $H^1(\mathcal{O})$ and $L^2(\mathcal{O})$. In particular, one can show that, if the family of subdivisions $\{\mathcal{T}_h, 0 < h \leq 1\}$ satisfies the assumptions of Theorem 3.1 with $m = 2$, $l \geq 0$ and $p = 2$, then

$$\text{(3.5)} \qquad \inf_{v \in \mathcal{V}_h} \|u - v\|_{H^1(\mathcal{O})} \leq C\, h\, \|u\|_{H^2(\mathcal{O})} \quad \forall u \in H^2(\mathcal{O}), \quad 0 < h \leq 1,$$

is valid, and therefore (for more details, see Theorem 5.4.8 in [BS02])

$$\text{(3.6)} \qquad \|u - u_h\|_{L^2(\mathcal{O})} \leq Ch^2 \|u\|_{H^2(\mathcal{O})} \leq Ch^2 \|F\|_{L^2(\mathcal{O})}, \quad 0 < h \leq 1.$$

Finally, by using duality arguments, one can show that there exists a $h_0 > 0$ such that (see Theorem 5.8.3 in [BS02])

$$\text{(3.7)} \qquad \|u - u_h\|_{H^{2,-\gamma}(\mathcal{O})} \leq Ch^{\gamma+1} \|u - u_h\|_{H^1(\mathcal{O})} \leq Ch^{2+\gamma} \|F\|_{L^2(\mathcal{O})}, \quad 0 < h \leq 1.$$

Since our problem (2.2) is well-posed in $L^p(\mathcal{O})$, the error bounds of the difference have to be transferred to $L^p(\mathcal{O})$ or $W^p_{-\gamma}(\mathcal{O})$, $\gamma > d - \frac{d}{p}$, $p \in (1, 2]$.

COROLLARY 3.2. *Assume that* (3.5) *holds. Then, under the conditions above, there exist positive constants* C *and* h_0 *such that for all* $h \le h_0$ *there is a unique solution to problem* (3.4) *satisfying*

$$\|u - u_h\|_{L^p(\mathcal{O})} \le Ch^{2-\frac{d}{p}+\frac{d}{2}} \|F\|_{L^p(\mathcal{O})}, \quad 0 < h \le 1,$$

and for $\gamma \in (0,1)$

$$\|u - u_h\|_{W^p_{-\gamma}(\mathcal{O})} \le Ch^{2-\frac{d}{p}+\frac{d}{2}} \|F\|_{W^p_{-\gamma}(\mathcal{O})}, \quad 0 < h \le 1.$$

Proof. Since $p \le 2$ and O is bounded we can infer that

$$\|u - u_h\|_{L^p(\mathcal{O})} \le C \ \|u - u_h\|_{L^2(\mathcal{O})}.$$

The estimate (3.6) gives

$$\|u - u_h\|_{L^p(\mathcal{O})} \le C\, h^2 \|F\|_{L^2(\mathcal{O})}.$$

Theorem 3.2 leads to

$$\|u - u_h\|_{L^p(\mathcal{O})} \le C\, h^{2+\frac{d}{2}-\frac{d}{p}} \|F\|_{L^p(\mathcal{O})}.$$

Similarly, we get

$$\|u - u_h\|_{W^p_{-\gamma}(\mathcal{O})} \le C \ \|u - u_h\|_{W^2_{-\gamma}(\mathcal{O})}.$$

Estimate (3.7) gives

$$\|u - u_h\|_{W^p_{-\gamma}(\mathcal{O})} \le C\, h^{2+\gamma} \|F\|_{L^2(\mathcal{O})}.$$

Next, we apply negative norm estimates. In particular, Theorem 3.2 leads to

$$\|u - u_h\|_{W^p_{-\gamma}(\mathcal{O})} \le C\, h^{2+\frac{d}{2}-\frac{d}{p}} \|F\|_{W^p_{-\gamma}(\mathcal{O})}. \qquad \square$$

3.2. Approximation of the semigroup. The approximation $\mathcal{A}_h : \mathcal{V}_h \to \mathcal{V}_h'$ of the operator A is given by the induced bilinear form on $\mathcal{V}_h$. In particular, for any $u \in \mathcal{V}_h$ we define $\mathcal{A}_h u$ by

$$\langle \mathcal{A}_h u, v\rangle := a(u,v), \quad v \in \mathcal{V}_h. \tag{3.8}$$

The semidiscrete approximation of the solution u to the abstract Cauchy problem

$$\begin{cases} \frac{d}{dt}u(t) + Au(t) = 0, & t > 0, \\ u(0) = u_0 \end{cases} \tag{3.9}$$

is the function $u_h : [0,\infty) \to \mathcal{V}_h$, determined by the following condition: For all $v \in \mathcal{V}_h$ we have

$$\begin{cases} \frac{d}{dt}\langle u_h(t), v\rangle + a(u_h(t), v) = 0, \ t > 0, \\ \langle u_h(0), v\rangle = \langle u_0, v\rangle. \end{cases} \tag{3.10}$$

In use of the operator $\mathcal{A}_h$, relation (3.10) is written as

$$\frac{du_h(t)}{dt} + \mathcal{A}_h u_h(t) = 0, \quad t > 0.$$

Let $\mathcal{H}_h$ be the space $\mathcal{V}_h$ equipped with the topology induced from H. Let $\mathcal{R}_h : H \to \mathcal{H}_h$ be the Ritz projection. Recall that the linear operator $\mathcal{A}_h : \mathcal{H}_h \to \mathcal{H}_h$ is defined by (3.8). Let $\mathcal{R}_h : V \to \mathcal{V}_h$ be the *Ritz operator* defined through

$$a(R_h u, v_h) = a(u, v) \quad \forall u \in \mathcal{V}_h \text{ and } u \in V.$$

By the Lax–Milgram theorem, it can be shown that the Ritz operator is well-defined, i.e., that the element $R_h u$ exists and is unique. Theorem 1.1 of [FSS01] shows that the operator $\mathcal{A}_h : \mathcal{H}_h \to \mathcal{H}_h$ generates an analytic semigroup on $\mathcal{H}_h$ independently of h. To be precise, there exist some $\theta \in (0, \pi)$ and some $M \geq 1$ such that $\mathbb{C} \setminus \Sigma_\theta \subset \rho(\mathcal{A}_h)$, where $\rho(\mathcal{A}_h)$ denotes the resolvent set of $\mathcal{A}_h$ and

$$\Sigma_\theta := \{z \in \mathcal{C}, 0 \leq |\arg(z)| \leq \theta\}.$$

The semidiscrete approximation admits the error estimate of optimal order for $t > 0$.

THEOREM 3.3 ([FSS01, Theorem 2.1]). *Given $u_0 \in H$, let u and u_h be the solutions of* (3.9) *and* (3.10)*, respectively. Then we have*

$$\|u(t) - u_h(t)\|_{L^2(\mathcal{O})} + h \|u(t) - u_h(t)\|_{H^1(\mathcal{O})} \leq C\, h^2 t^{-1} \|u_0\|_{L^2(\mathcal{O})}, \quad t > 0,\ 0 < h \leq 1.$$

Similarly to the Proof of Theorem 3.3 one can find error estimates in L^p spaces.

THEOREM 3.4. *Given $u_0 \in H$, let u and u_h be the solutions of* (3.9) *and* (3.10)*, respectively. Then we have for $1 \leq p \leq 2$ and $0 \leq \gamma < 1$*

$$\|u(t) - u_h(t)\|_{W^{p,-\gamma}(\mathcal{O})} \leq C\, h^{2-\frac{d}{p}+\frac{d}{2}} t^{-1} \|u_0\|_{W^{p,-\gamma}(\mathcal{O})}, \quad t > 0,\ 0 < h \leq 1.$$

Proof. The proof of Theorem 3.4 is similar to the proof of Theorem 3.3, i.e., Theorem 2.1 in [FSS01]. Assume that u solves problem (3.3) with $F = u_0$ and u_h solves problem (3.4) with $F = \mathcal{I}_h u_0$. In Corollary 3.2 we have shown that there exists estimates of the error in L^p spaces. In particular, we have

$$\|u - u_h\|_{L^p(\mathcal{O})} \leq C h^{2-\frac{d}{p}+\frac{d}{2}} \|F\|_{L^p(\mathcal{O})}.$$

Since $u = A^{-1}u_0$ and $\mathcal{A}_h^{-1}\mathcal{I}_h u_0$, we can write

$$\left\|A^{-1}u_0 - \mathcal{A}_h^{-1}\mathcal{I}_h u_0\right\|_{L^p(\mathcal{O})} \leq C h^{2-\frac{d}{p}+\frac{d}{2}} \|u_0\|_{L^p(\mathcal{O})}.$$

Next, since for all $\lambda_0, \lambda \in \rho(A) \cap \rho(\mathcal{A}_h)$ we have

$$\begin{aligned}(\lambda\mathcal{I}_h - \mathcal{A}_h)^{-1} - (\lambda - A)^{-1} &= \left(1 + (\lambda - \lambda_0)(\lambda\mathcal{I}_h - \mathcal{A}_h)^{-1}\mathcal{I}_h\right) \\ &\times \left((\lambda_0\mathcal{I}_h - \mathcal{A}_h)^{-1}\mathcal{I}_h - (\lambda_0 - A)^{-1}\right) \times (\lambda_0 - A)(\lambda_0 - A)^{-1}, \end{aligned} \tag{3.11}$$

we can conclude that the following estimates:

$$\left\|(\lambda I - A)^{-1} u_0 - (\lambda\mathcal{I}_h - \mathcal{A}_h)^{-1}\mathcal{I}_h u_0\right\|_{L^p(\mathcal{O})} \leq C h^{2-\frac{d}{p}+\frac{d}{2}} \|u_0\|_{L^p(\mathcal{O})} \tag{3.12}$$

and for $\gamma \in (0, 1)$

$$\left\|(\lambda I - A)^{-1} u_0 - (\lambda\mathcal{I}_h - \mathcal{A}_h)^{-1}\mathcal{I}_h u_0\right\|_{W^p_{-\gamma}(\mathcal{O})} \leq C h^{2-\frac{d}{p}+\frac{d}{2}} \|u_0\|_{W^p_{-\gamma}(\mathcal{O})} \tag{3.13}$$

are valid for all $\lambda \in \rho(A) \cap \rho(\mathcal{A}_h)$. The proof of Theorem 3.3 relies on the fact that we can represent the semigroup in terms of the resolvent. By comparing to [FSS01, page 58], or to [Paz83, Theorem 1.7.7], we have

$$\|u(t) - u_h(t)\mathcal{I}_h\| \leq \frac{1}{2\pi} \int_\Gamma \left|e^{-tz}\right| \cdot \left\| \left[(\lambda I - A)^{-1} - (\lambda \mathcal{I}_h - \mathcal{A}_h)^{-1}\mathcal{I}_h\right] u_0 \right\| \, dz, \tag{3.14}$$

where Γ is the boundary of $\Sigma_{\theta_1} = \{z, 0 \leq |\arg(z)| \leq \theta_1\}$, for $\theta_1 \in (\theta, \frac{\pi}{2})$ coming from $\infty e^{i\theta_1}$, passing through the origin and going out to $\infty e^{-i\theta_1}$. Using estimate (3.6) and taking into account the calculations in (3.11) gives

$$\begin{aligned} &\left\| \left[(\lambda I - A)^{-1} - (\lambda \mathcal{I}_h - \mathcal{A}_h)^{-1}\mathcal{I}_h\right] u_0 \right\|_{L^2(\mathcal{O})} \\ &\qquad \leq C\, h^m \left\|u_0\right\|_{L^2(\mathcal{O})}, \quad u_0 \in L^2(\mathcal{O}). \end{aligned} \tag{3.15}$$

By substituting estimate (3.15) into (3.14), Theorem 3.3 can be shown. Similarly, by substituting estimate (3.12) into (3.14) one gets

$$\begin{aligned} &\|u(t) - u_h(t)\mathcal{I}_h\|_{L^p(\mathcal{O})} \leq \frac{C}{2\pi}\, h^{2-\frac{d}{p}+\frac{d}{2}}\ \|u_0\|_{L^p(\mathcal{O})} \int_\Gamma \left|e^{-tz}\right| \, dz \\ &\leq C\, h^{2-\frac{d}{p}+\frac{d}{2}}\ \|u_0\|_{L^p(\mathcal{O})} \int_0^\infty e^{-tr\cos(\theta_1)}\, dr \leq C\, h^{2-\frac{d}{p}+\frac{d}{2}}\ t^{-1}\ \|u_0\|_{L^p(\mathcal{O})}. \end{aligned}$$

The case $\gamma = 0$ of Theorem 3.4 follows. To show Theorem 3.4 for $\gamma > 0$ one can use the same computations, but one has to substitute estimate (3.13) instead of (3.12) into (3.14). □

3.3. Mesh-dependent norms. For $x \in \mathcal{V}_h$, one can define the mesh-dependent inner product

$$\langle v, w\rangle_h := h^d \sum_{i=1}^{k_h} N_i^h(v) N_i^h(w)$$

and the norm

$$\|v\|^2_{L^2_h(\mathcal{O})} := \langle v, v\rangle = h^d \sum_{i=1}^{k_h} v(N_i^h)^2.$$

It can be shown that there exist some constants c and C such that we have for all $h \in (0, 1)$ (see Lemma 6.2.7 of [BS02])

$$c\|v\|_{L^2(\mathcal{O})} \leq \|v\|_{L^2_h(\mathcal{O})} \leq C\|v\|_{L^2(\mathcal{O})} \quad \forall v \in \mathcal{V}_h.$$

Since the natural spaces for SPDEs driven by Poisson random measure are L^p spaces, we will also define mesh-dependent L^p spaces; i.e., for $0 < h \leq 1$ and $v \in \mathcal{V}_h$ we define

$$\|v\|_{L^p_h(\mathcal{O})} := \left(h^d \sum_{i=1}^{k_h} \left|N_i^h(v)\right|^p \right)^{\frac{1}{p}}$$

and

$$L^p_h(\mathcal{O}) := \{v \in \mathcal{V}_h, \|v\|_{L^p_h(\mathcal{O})} < \infty\}.$$

Similarly to the case where $p = 2$, one can show that there exist some constants c and C such that we have for all h with $0 < h \leq 1$

$$c\|v\|_{L^p(\mathcal{O})} \leq \|v\|_{L^p_h(\mathcal{O})} \leq C\|v\|_{L^p(\mathcal{O})} \quad \forall v \in \mathcal{V}_h \cap L^p(\mathcal{O}). \tag{3.16}$$

Remark 3.1. It follows from the nondegeneracy assumptions of the subdivisions that there exist some constants c and C such that

$$c\, h^d \leq k_h \leq Ch^d, \quad 0 < h < 1.$$

Theorem 1.1 of [FSS01] shows that the operator $\mathcal{A}_h : \mathcal{H}_h \to \mathcal{H}_h$ generates an analytic semigroup on $\mathcal{H}_h$ with parameter M and ω independent of h. To be precise, there exist some $\theta \in (0, \pi)$ and some $M \geq 1$ such that $\mathcal{C} \setminus \Sigma_\theta \subset \rho(\mathcal{A}_h)$, where $\rho(\mathcal{A}_h)$ denotes the resolvent set of operator $\mathcal{A}_h$ and

$$\Sigma_\theta := \{z \in \mathcal{C}, 0 \leq |\arg(z)| \leq \theta\}.$$

Let X be a Banach space and $A : X \to X$ be a operator generating an analytic semigroup on X. Since the constants M, ω, and C in

$$\left\|e^{-tA}u\right\|_X \leq M\, e^{\omega t}\, \|u\|_X$$

and

$$\left\|Ae^{-tA}u\right\|_X \leq \frac{C}{t}\, e^{\omega t}\, \|u\|_X$$

depend on the resolvent set, in particular, on the angle θ which includes the numerical range, we can conclude that there exist some constants C_1 and C_2 such that

$$\left\|e^{-t\mathcal{A}_h}u\right\|_{L^p_h(\mathcal{O})} \leq C_1\, \|u\|_{L^p_h(\mathcal{O})} \quad \forall v \in L^p_h(\mathcal{O}), \quad 0 < h \leq 1, \tag{3.17}$$

and

$$\left\|Ae^{-t\mathcal{A}_h}u\right\|_{L^p_h(\mathcal{O})} \leq \frac{C_2}{t}\, \|u\|_{L^p_h(\mathcal{O})} \quad \forall v \in L^p_h(\mathcal{O}), \quad 0 < h \leq 1. \tag{3.18}$$

For more details, see [Paz83, Theorem 2.5.2] and [FSS01].

3.4. The mass and stiffness matrix. In this section we will investigate properties of the mass and stiffness matrix. Let $\mathcal{O}$ be a polyhedron in $\mathbb{R}^d$ and $\{\mathcal{T}_h, 0 < h \leq 1\}$ be a nondegenerate family of subdivisions, where the reference finite element $(\hat{K}, \hat{\mathcal{P}}, \hat{\mathcal{N}})$ satisfies the standard conditions for $m = l = 1$.

The mass matrix $\mathbf{M}_h$ of a subdivision is given by $\mathbf{M}_h = ((m^h_{i,j}))^{k_h}_{i,j=1}$, where

$$m^h_{i,j} := \langle \phi^h_i, \phi^h_j \rangle, \quad 1 \leq i, j \leq k_h.$$

Since the shape functions are locally defined, there exists a number m such that for all $h \in (0, 1)$ there exists a set $\mathcal{C}_h \subset \mathbb{N}^{k_h} \times \mathbb{N}^{k_h}$ such that

$$\operatorname{supp}(\phi^h_i) \cap \operatorname{supp}(\phi^h_j) = \emptyset \quad \forall (i, j) \not\subset \mathcal{C}_h, \quad 0 < h \leq 1,$$

and

$$|(i, \mathbb{N}^{k_h}) \cap \mathcal{C}_h| \leq m, \quad 0 < h \leq 1.$$

Remark 3.2. The mass matrix is given by Theorem 9.8, page 387 in Ern and Guermond [EG04]. The left-hand side is given by

$$\langle \mathcal{A}_h \phi_l^h, \phi_j^h \rangle = \langle \nabla \phi_i^h, \nabla \phi_j^h \rangle.$$

Since $\{\mathcal{T}_h, 0 < h \leq 1\}$ is a uniformly bounded family of subdivisions, we have

$$\langle \nabla \phi_l^h, \nabla \phi_j^h \rangle = \begin{cases} c\, h^{d-2}, & \text{if } (i,j) \subset \mathcal{C}_h, \\ 0, & \text{elsewhere.} \end{cases}$$

Fix some i, $1 \leq i \leq k_h$. Our point of interest is $\mathcal{A}_h^{-1}\phi_i$. Let $u_i = \mathcal{A}_h^{-1}\phi_i$. Then

$$\langle \mathcal{A}_h u_i, \phi_j^h \rangle = \langle \phi_i^h, \phi_j^h \rangle, \quad j = 1, \ldots, k_h.$$

Since $u_i \in \mathcal{V}_h$, there exists $u_{i1}, \ldots, u_{ik_h} \in \mathbb{R}$ such that

$$u_i = \sum_{l=1}^{k_h} u_{il} \phi_l^h.$$

Therefore,

$$\langle \mathcal{A}_h u_i, \phi_j^h \rangle = \sum_{l=1}^{k_h} u_{il} \langle \mathcal{A}_h \phi_l^h, \phi_j^h \rangle = \langle \phi_i^h, \phi_j^h \rangle$$

and by Remark 3.2

$$u_{il} = \begin{cases} c\, h^2, & \text{if } (i,j) \subset C, \\ 0, & \text{elsewhere.} \end{cases} \tag{3.19}$$

3.5. Approximation of the noise. The space time Poissonian noise of jump type is given by a Poisson random measure

$$\eta : \mathcal{B}(\mathcal{O}) \times \mathcal{B}(\mathbb{R}) \times \mathcal{B}([0,\infty)) \to \mathbb{R}^+$$

with intensity $\lambda \times \nu$, where ν is a Lévy measure over $\mathbb{R}$, such that

$$\int_{\mathbb{R}} |z|^p \, \nu(dz) < \infty.$$

The function

$$g : \mathbb{R} \times \mathcal{O} \times \mathbb{R} \to \mathbb{R}$$

is Lipschitz continuous with Nemytskij operator

$$L^p(\mathcal{O}) \times \mathbb{R} \ni (u,z) \mapsto G(u;z) = g(u(\cdot);\cdot;z) \in L^p(\mathcal{O}), \tag{3.20}$$

where $g(u(\cdot);\cdot;z)(\xi) := g(u(\xi);\xi;z)$. In [Hau05] we have seen that the space time Poissonian noise can be described by a Banach-valued Poisson random measure. In particular, let $Z = B_{p,\infty}^{\frac{d}{p}-d}(\mathcal{O})$ and U be the unit sphere in Z. Let χ be defined by

$$\mathcal{O} \ni \xi \mapsto \chi(\xi) := c\, \delta_\xi \in Z,$$

where $c = |\delta_\xi|_Z^{-1}$. The intensity ι of the Z-valued Poisson random measure is given by

$$\mathcal{B}(Z) \ni B \mapsto \iota(B) := \int_{\zeta\in\mathbb{R}} \int_{x\in U} 1_B(\zeta x)\sigma(dx)\nu(d\zeta) \in \mathbb{R}^+,$$

where $\sigma(A) := \lambda(\chi^{-1}(A \cap \chi(\mathcal{O})))$, $A \in \mathcal{B}(Z)$. By inserting the definition above into the definition of the Besov space, one gets by Proposition A.1

$$\begin{aligned}\int_{z\in Z} |G(u,z) - G(v,z)|_Z^p \iota(dz) &= \int_{\xi\in\mathcal{O}} \int_{\zeta\in\mathbb{R}} |g(u(\xi),\xi;\zeta) - g(v(\xi),\xi;\zeta)|^p \,\nu(d\zeta)\,d\xi \\ &\le C\,|u-v|^p_{L^p(\mathcal{O})}, \quad u,v \in C_b^0(\mathcal{O}).\end{aligned}$$

Extension to $L^p(\mathcal{O})$ leads to the Lipschitz property (2.10) of G. The approximation can be similarly interpreted. First, the Lebesgue measure is approximated by λ_h defined by

$$\mathcal{B}\left(\{1,2,\dots,k_h\}\right) \ni A \mapsto \lambda_h(A) := \sum_{i=1}^{k_h} \lambda(J_i^h) 1_A(i) \in \mathbb{R}^+.$$

Now the approximation $\iota_h : \mathcal{B}(\{1,\dots,k_h\}) \times \mathcal{B}(\mathbb{R}) \to \mathbb{R}^+$ of the intensity ι is defined by

$$\{1,\dots,k_h\} \times \mathcal{B}(\mathbb{R}) \ni (A,B) \mapsto \iota_h(A,B) := \sum_{i=1}^{k_h} 1_A(i)\lambda_h(\{i\}) \int_{\zeta\in\mathbb{R}} \nu(d\zeta) \in \mathbb{R}^+.$$

The approximated Nemytiskij operator G_h is defined by

$$\mathcal{V}_h \times \{1,\dots,k_h\} \times \mathcal{B}(\mathbb{R}) \ni (u,i,\zeta) \mapsto G_h(u,i,\zeta) := g(u(N_i^h), N_i^h;\zeta)\, c_i^h \phi_i^h.$$

Similarly to the above we can show the following proposition.

PROPOSITION 3.1. *There exists some $C > 0$ such that for any $h > 0$ and for any $\gamma > \frac{1}{2}(d - \frac{d}{p})$ and $\delta < 2\gamma - d - \frac{d}{p}$ the following inequality holds for all $u_h, v_h \in \mathcal{V}_h$:*

$$\int_{z\in Z_h} \left\|(-\mathcal{A}_h)^{-\gamma}\left[G_h(u_h,z) - G_h(v_h,z)\right]\right\|^p_{L^p(\mathcal{O})} \iota_h(dz) \le C\, h^{d\delta p} \left\|u_h - v_h\right\|^p_{L^p(\mathcal{O})}.$$

Proof. Let $u, v \in \mathcal{V}_h$. Then we have

$$\begin{aligned}&\int_{z\in Z} \left\|(-\mathcal{A}_h)^{-\gamma}\left[G_h(u_h,z) - G_h(v_h,z)\right]\right\|^p_{L^p(\mathcal{O})} \iota_h(dz) \\ &= \mathbb{E}\left\|\sum_{i=1}^{k_h} \int_{\mathbb{R}} (-\mathcal{A}_h)^{-\gamma}\left[G_h(u_h,\zeta) - G_h(v_h,\zeta)\right] \mu_i^h(d\zeta,ds)\right\|^p_{L^p(\mathcal{O})},\end{aligned}$$

where μ_i^h, $i = 1,\dots,k_h$, are independent Poisson random measures with intensity $\lambda_h(i)\nu$. The generalized Burkholder inequality gives

$$\begin{aligned}&\int_{z\in Z} \left\|(-\mathcal{A}_h)^{-\gamma}\left[G_h(u_h,z) - G_h(v_h,z)\right]\right\|^p_{L^p(\mathcal{O})} \iota_h(dz) \\ &\le \sum_{j=1}^{k_h} \int_{\mathbb{R}} \left\|(-\mathcal{A}_h)^{-\gamma}\left[g(u_h(N_j^h),N_j^h;\zeta) - g(v_h(N_j^h),N_j^h;\zeta)\right] c_j^h\phi_j^h\right\|^p_{L^p(\mathcal{O})} \lambda(J_j^h)\nu(d\zeta).\end{aligned}$$

Since $g(u_h(N_j^h), N_j^h; \zeta) - g(v_h(N_j^h), N_j^h; \zeta) \in \mathbb{R}$ and $\mathcal{A}_h$ is linear, we can continue

$$\int_{z\in Z_h} \left\|(-\mathcal{A}_h)^{-\gamma}\left[G_h(u_h,z) - G_h(v_h,z)\right]\right\|_{L^p(\mathcal{O})}^p \iota_h(dz) \le \sum_{j=1}^{k_h} \int_{\mathbb{R}}$$

$$\left\|\left[g(u_h(N_j^h), N_j^h; \zeta) - g(v_h(N_j^h), N_j^h; \zeta)\right](-\mathcal{A}_h)^{-\gamma} c_j^h \phi_j^h\right\|_{L^p(\mathcal{O})}^p \lambda(J_j^h)\nu(d\zeta).$$

Since the family of subdivisions is nondegenerate, we have $\left\|\phi_j^h\right\|_{L^p(\mathcal{O})}^p \sim h^{\frac{d}{p}}$, $c_j^h \sim h^{-d}$, and $\lambda(J_j^h) \sim h^d$. Therefore, by using also Remark 3.2, we get

$$\begin{aligned}
&\int_{z\in Z_h} \left\|(-\mathcal{A}_h)^{-\gamma}\left[G_h(u_h,z) - G_h(v_h,z)\right]\right\|_{L^p(\mathcal{O})}^p \iota_h(dz)\\
&\le C \sum_{j=1}^{k_h} \int_{\mathbb{R}} \left|g(u_h(N_j^h), N_j^h; \zeta) - g(v_h(N_j^h), N_j^h; \zeta)\right|^p \left\|(-\mathcal{A}_h)^{-\gamma} c_j^h \phi_j^h\right\|_{L^p(\mathcal{O})}^p \lambda(J_j^h)\nu(d\zeta)\\
&\le C\, h^{\gamma 2} \sum_{j=1}^{k_h} \int_{\mathbb{R}} \left|g(u_h(N_j^h), N_j^h; \zeta) - g(v_h(N_j^h), N_j^h; \zeta)\right|^p \left\|c_j^h \phi_j^h\right\|_{L^p(\mathcal{O})}^p \lambda(J_j^h)\nu(d\zeta)\\
&\le C\, h^{\gamma 2 - dp + d} \sum_{j=1}^{k_h} \int_{\mathbb{R}} \left|g(u_h(N_j^h), N_j^h; \zeta) - g(v_h(N_j^h), N_j^h; \zeta)\right|^p \lambda(J_j^h)\nu(d\zeta).
\end{aligned}$$

According to the Lipschitz condition of g we have

$$\begin{aligned}
&\int_{z\in Z} \left\|(-\mathcal{A}_h)^{-\gamma}\left[G_h(u_h,z) - G_h(v_h,z)\right]\right\|_{L^p(\mathcal{O})}^p \iota_h(dz)\\
&\qquad \le C\, h^{\gamma 2 - dp + d} \sum_{j=1}^{k_h} h^d \left|u_h(N_j^h) - v_h(N_j^h)\right|^p.
\end{aligned}$$

The equivalence between $L^p(\mathcal{O})$ and $L_h^p(\mathcal{O})$ gives the assertion. ∎

In order to calculate the error induced by the approximation of the Poisson random measure, we have to express the measure ι_h in terms of ι.

PROPOSITION 3.2. *There exists some $C > 0$ such that for any $h > 0$ and for any $\gamma > \frac{1}{2}(d - \frac{d}{p})$ and $\delta < 2\gamma - d - \frac{d}{p}$ the following inequality holds for all $u_h \in \mathcal{V}_h$:*

$$\int_Z \left\|(-\mathcal{A}_h)^{-\gamma}\mathcal{I}_h G(u,z)\right\|_{L^p(\mathcal{O})}^p (\iota - \iota_h)(dz) \le C\, h^{d\delta p}\left(1 + \|u_h\|_{L^p(\mathcal{O})}^p\right).$$

Remark 3.3. In our calculation we use the fact that for $\gamma > d - \frac{d}{p}$ in $W^{p,-\gamma}(\mathcal{O})$ the noise vanishes for $h \to 0$. But, by plugging the difference $\iota - \iota_h$ into the definition of Z, one can show directly that the approximated noise converges to the functional-valued representation of the space time Poissonian noise in Z.

Proof. Embedding Z_h into Z by $i \mapsto c_i^h \phi_i^h$ leads to

$$\begin{aligned}
&\int_Z \left\|(-\mathcal{A}_h)^{-\gamma}\mathcal{I}_h G(u,z)\right\|_{L^p(\mathcal{O})}^p \ (\iota-\iota_h)(dz) \leq \int_{\xi\in\mathcal{O}}\int_{\zeta\in\mathbb{R}} \\
&\quad \left\|(-\mathcal{A}_h)^{-\gamma}\left[\mathcal{I}_h G(u,z)\delta_\xi - \sum_{j=1}^{k_h} 1_{J_j^h}(\xi) g(u(N_j^h),N_j^h,\zeta)c_j^h\phi_j^h\right]\right\|_{L^p(\mathcal{O})}^p d\zeta\ d\xi \\
&\leq \sum_{i=1}^{k_h}\int_{\xi\in J_i^h}\int_{\zeta\in\mathbb{R}} \left\|(-\mathcal{A}_h)^{-\gamma}\left[\sum_{l=1}^{k_h} g(u(N_l^h),N_l^h,\zeta)\phi_l^h\delta_\xi\right.\right. \\
&\qquad\qquad \left.\left. - \sum_{j=1}^{k_h} 1_{J_j^h}(\xi) g(u(N_j^h),N_j^h,\zeta)c_j^h\phi_j^h\right]\right\|_{L^p(\mathcal{O})}^p d\zeta\ d\xi \\
&\leq \sum_{j=1}^{k_h}\int_{\xi\in J_j^h}\int_{\zeta\in\mathbb{R}} \left\|(-\mathcal{A}_h)^{-\gamma}\left[\sum_{l=1}^{k_h} g(u(N_l^h),N_l^h,\zeta)\phi_l^h\delta_\xi\right]\right\|_{L^p(\mathcal{O})}^p d\zeta\ d\xi \\
&+ \sum_{j=1}^{k_h}\int_{\xi\in J_j^h}\int_{\zeta\in\mathbb{R}} \left\|(-\mathcal{A}_h)^{-\gamma}\left[g(u(N_j^h),N_j^h,\zeta)c_j^h\phi_j^h\right]\right\|_{L^p(\mathcal{O})}^p d\zeta\ d\xi \\
&=: S_1 + S_2.
\end{aligned}$$

By Corollary 3.1 we obtain

$$S_1 \leq h^{-\delta p}\sum_{j=1}^{k_h}\int_{\xi\in J_j^h}\int_{\zeta\in\mathbb{R}} \left\|(-\mathcal{A}_h)^{-\gamma}\left[\sum_{l=1}^{k_h} g(u(N_l^h),N_l^h,\zeta)\phi_l^h\delta_\xi\right]\right\|_Z^p d\zeta\ d\xi.$$

Inequality (A.1) leads to

$$S_1 \leq C\ h^{-\delta p}\sum_{j=1}^{k_h}\int_{\zeta\in\mathbb{R}} \left\|(-\mathcal{A}_h)^{-\gamma}\left[\sum_{l=1}^{k_h} g(u(N_l^h),N_l^h,\zeta)\phi_l^h\right]\right\|_{L^p(J_j^h)}^p d\zeta.$$

Since there exists a constant m such that $|\{i \mid \sup\phi_i^h \cap J_j^h \neq \emptyset\}| \leq m$ for all $i,j = 1,\dots,k_h$ and $h>0$, and $\sum_{i=1}^{k_h}\phi_i^h(\xi) = 1$, we can infer by Remark 3.2 that

$$\begin{aligned}
S_1 &\leq C\ h^{(2\gamma-\delta)p}\sum_{j=1}^{k_h}\int_{\zeta\in\mathbb{R}} \left\|\sum_{l,\sup\phi_l^h\cap J_j^h\neq\emptyset} g(u(N_l^h),N_l^h,\zeta)\phi_l^h\right\|_{L^p(J_j^h)}^p d\zeta \\
&\leq C\ h^{(2\gamma-\delta)p}\ m\ \sum_{j=1}^{k_h}\int_{\zeta\in\mathbb{R}} \left\|g(u(N_j^h),N_j^h,\zeta)\phi_j^h\right\|_{L^p(\mathcal{O})}^p d\zeta.
\end{aligned}$$

By the linearity of g and the equivalence of $L_h^p(\mathcal{O})$ and $L^p(\mathcal{O})$ we get

$$S_1 \leq C\ h^{(2\gamma-\delta)p} m\ \left(1 + \sum_{j=1}^{k_h} h^d\ \left|u(N_j^h)\right|^p\right) \leq h^{(2\gamma-\delta)p}\ m\ \left(1 + |u|_{L^p(\mathcal{O})}^p\right).$$

By applying Remark 3.19 we get for S_2

$$\begin{aligned}
S_2 &\le \sum_{j=1}^{k_h} \lambda(J_j^h) \int_{\zeta\in\mathbb{R}} \left\|(-\mathcal{A}_h)^{-\gamma} g(u(N_j^h), N_j^h, \zeta) c_j^h \phi_j^h\right\|_{L^p(\mathcal{O})}^p \, d\zeta \\
&\le C\, h^{2\gamma p} \sum_{j=1}^{k_h} \lambda(J_j^h) \int_{\zeta\in\mathbb{R}} \left\|g(u(N_j^h), N_j^h, \zeta) c_j^h \phi_j^h\right\|_{L^p(\mathcal{O})}^p \, d\zeta \\
&\le C\, h^{2\gamma p - dp} \sum_{j=1}^{k_h} \lambda(J_j^h) h^d \int_{\zeta\in\mathbb{R}} \left|g(u(N_j^h), N_j^h, \zeta)\right| \left\|\phi_j^h\right\|_{L^p(\mathcal{O})}^p \, d\zeta \\
&\le C\, h^{2\gamma p - dp + d} \sum_{j=1}^{k_h} \lambda(J_j^h) h^d \int_{\zeta\in\mathbb{R}} \left|g(u(N_j^h), N_j^h, \zeta)\right| \, d\zeta \,.
\end{aligned}$$

The linearity of g and the equivalence of $L_h^p(\mathcal{O})$ and $L^p(\mathcal{O})$ give

$$S_2 \le C\, h^{2\gamma p - dp + d} \left(1 + \|u\|_{L^p(\mathcal{O})}^p\right). \qquad \square$$

4. Stability of the finite element approximation. In this section we shall investigate the stability of the approximation by finite elements. Since the semidiscrete problem (2.3) is finite-dimensional, for all $0 < h \le 1$ there exists a unique solution $u_h = \{u_h(t) \in \mathbb{R}^{k_h}, 0 \le t < \infty\}$ (see, e.g., Theorem 8 of [Pro04]). In this section we show that the family of unique solution $\{u_h, 0 < h \le 1\}$ is uniformly bounded.

THEOREM 4.1. *Let $u_h = \{u_h(t), 0 \le t < \infty\}$ be a solution to problem (2.3). Then, under the conditions of Theorem 2.1, for all $p \in (1, 2]$, with $p < \frac{2}{d} + 1$, and for all $T > 0$ there exists some constant such that*

$$\sup_{0\le t\le T} \mathbb{E}\, \|u_h(t)\|_{L^p(\mathcal{O})}^p \le C, \quad 0 < h \le 1.$$

Proof. First, we show that there exist some $q < \infty$ and some constants C_1 and C_2 such that

$$(\mathbb{E}\|u_h(t)\|^p)^q \le C_1 + C_2 \int_0^t (\mathbb{E}\, \|u_h(s)\|^p)^q \, ds, \quad t > 0 \text{ and } 0 < h \le 1.$$

The Gronwall lemma yields the assertion. The variation of constant formula gives

$$\begin{aligned}
u_h(t) = e^{-t\mathcal{A}_h} u_h(0) &+ \int_0^t e^{-(t-s)\mathcal{A}_h} f_h(u_h(s))\, ds \\
&+ \int_0^t \sum_{i=1}^{k_h} \int_{\zeta_i\in\mathbb{R}} e^{-(t-s)\mathcal{A}_h} g_h(u_h(s); i, \zeta_i)\, \tilde{\eta}_h^i(d\zeta_i; ds), \quad t > 0.
\end{aligned}$$

Taking the expectation leads to

$$\begin{aligned}
\mathbb{E}\, \|u_h(t)\|_{L^p(\mathcal{O}}^p &\le C_p \left\|e^{-t\mathcal{A}_h}\, u_h(0)\right\|_{L^p(\mathcal{O})}^p \\
&+ C_p \mathbb{E} \left\| \int_0^t e^{-(t-s)\mathcal{A}_h} f_h(u_h(s))\, ds \right\|_{L^p(\mathcal{O})}^p \\
&+ C_p \mathbb{E} \left\| \int_0^t \sum_{i=1}^{k_h} \int_{\zeta_i\in\mathbb{R}} e^{-(t-s)\mathcal{A}_h} g_h(u_h(s); i, \zeta_i)\, \tilde{\eta}_h^i(d\zeta_i; ds) \right\|_{L^p}^p \\
&=: S_1 + S_2 + S_3.
\end{aligned}$$

If $u_0 \in L^p(\mathcal{O})$, then the first entity S_1 is finite. The second entity can be calculated directly. The Minkowski inequality gives

$$\begin{aligned} S_2 &= \mathbb{E} \left\| \int_0^t e^{-(t-s)\mathcal{A}_h} f_h(u_h(s))\, ds \right\|^p_{L^p(\mathcal{O})} \\ &\leq \mathbb{E} \left(\int_0^t \left\| e^{-(t-s)\mathcal{A}_h} f_h(u_h(s)) \right\|_{L^p(\mathcal{O})} ds \right)^p. \end{aligned}$$

In section 3.2 we have seen that there exist $\omega \in \mathbb{R}$, $M \geq 1$, and $\theta \in (0, \pi)$ such that for all $h \in (0,1)$ the operator $\mathcal{A}_h$ generates an analytic semigroup of (ω, M, θ) on $L^p_h(\mathcal{O})$ and the inequality (3.17) holds. In section 3.3 we have seen that there exist some constants c and C such that for all $0 < h \leq 1$ we have (see (3.16))

$$c \|w\|_{L^p(\mathcal{O})} \leq \|w_h\|_{L^p_h(\mathcal{O})} \leq C \|w\|_{L^p(\mathcal{O})}, \quad w \in \mathcal{V}_h \tag{4.1}$$

where

$$\|w\|_{L^p_h(\mathcal{O})} := h^d \sum_{i=1}^{k_h} |w_h(N_i)|^p, \quad w \in \mathcal{V}_h.$$

Therefore, a constant $C < \infty$ exists such that for all $0 < h \leq 1$

$$\left\| e^{-(t-s)\mathcal{A}_h} \mathcal{A}_h v \right\|_{L^p(\mathcal{O})} \leq C \|v\|_{L^p(\mathcal{O})}, \quad v \in L^p(\mathcal{O}) \cap \mathcal{V}_h. \tag{4.2}$$

Thus,

$$S_2 \leq C\, \mathbb{E} \left(\int_0^t \|f_h(u_h(s))\|_{L^p(\mathcal{O})}\, ds \right)^p.$$

Again by (4.1) we can infer that

$$S_2 \ \leq C\, \mathbb{E} \left(\int_0^t \|f(u_h(s))\|_{L^p_h(\mathcal{O})}\, ds \right)^p.$$

By the globally Lipschitz property of f it follows that

$$S_2 \leq C\, \mathbb{E} \left(\int_0^t \left(1 + \|u_h(s)\|_{L^p_h(\mathcal{O})} \right) ds \right)^p.$$

Again by (4.1) we get

$$S_2 \leq C\, \mathbb{E} \left(\int_0^t \left(1 + \|u_h(s)\|_{L^p(\mathcal{O})} \right) ds \right)^p.$$

The Hölder inequality leads to

$$S_2 \leq C\, t^p \left(1 + \sup_{0 \leq s \leq t} \mathbb{E} \|u_h(s)\|^p_{L^p(\mathcal{O})} \right).$$

Next, we estimate the term

$$S_3 = \mathbb{E} \left\| \int_0^t \sum_{i=1}^{k_h} \int_{\zeta_i \in \mathbb{R}} e^{-(t-s)\mathcal{A}_h} g_h(u_h(s); i, \zeta_i)\, \tilde{\eta}^i_h(d\zeta_i; ds) \right\|^p_{L^p(\mathcal{O})}.$$

Since $L^p(\mathcal{O})$ is of M-type p, we can infer by Proposition 3.3 in [Hau05]

$$S_3 \leq \mathbb{E} \int_0^t \frac{1}{k_h} \sum_{i=1}^{k_h} \int_{\zeta_i \in \mathbb{R}} \mathbb{E} \left\| e^{-(t-s)\mathcal{A}_h} g_h(u_h(s); i; \zeta_i) \right\|_{L^p(\mathcal{O})}^p \nu(d\zeta_i)\, ds.$$

In section 3.2 we have seen that there exist $\omega \in \mathbb{R}$, $M \geq 1$, and $\theta \in (0, \pi)$ such that for all $h \in (0,1)$ the operator $\mathcal{A}_h$ generates an analytic semigroup of (ω, M, θ) on $L^p(\mathcal{O})$. In particular (3.18) holds; i.e., there exists a constant $C < \infty$ such that for all $h > 0$

$$\left\| e^{-(t-s)\mathcal{A}_h} \mathcal{A}_h v \right\|_{L_h^p(\mathcal{O})} \leq \frac{C}{t} \|v\|_{L_h^p(\mathcal{O})}, \quad v \in L_h^p(\mathcal{O}). \tag{4.3}$$

Because of inequality (4.3), and since the norms in $L_h^p(\mathcal{O})$ and $L^p(\mathcal{O})$ are equivalent, we have for all $\gamma < \frac{1}{p}$

$$S_3 \leq \int_0^t \frac{1}{k_h} \sum_{i=1}^{k_h} (t-s)^{-\gamma p} \int_{\zeta_i \in \mathbb{R}} \mathbb{E} \left\| \mathcal{A}_h^{-\gamma} g_h(u_h(s); i, \zeta_i) \right\|_{L^p(\mathcal{O})}^p \nu(d\zeta_i)\, ds.$$

Since $g_h(u_h(s); i, \zeta) = N_i(G(u_h(s); \zeta_i))\phi_i$ and since $\mathcal{A}_h$ is linear, we get

$$S_3 \leq \int_0^t \frac{1}{k_h} \sum_{i=1}^{k_h} (t-s)^{-\gamma p} \int_{\zeta_i \in \mathbb{R}} \mathbb{E} \left\| N_i(G(u_h(s); \zeta_i)) \mathcal{A}_h^{-\gamma} \phi_i \right\|_{L^p(\mathcal{O})}^p \nu(d\zeta_i)\, ds.$$

The Hölder inequality gives

$$\begin{aligned} S_3 \leq & \int_0^t \frac{1}{k_h} \\ & \sum_{i=1}^{k_h} (t-s)^{-\gamma p} \int_{\zeta_i \in \mathbb{R}} \mathbb{E} \left\| N_i(G(u_h(s); \zeta_i)) \right\|_{L^p(\mathcal{O})}^p \left\| \mathcal{A}_h^{-\gamma} \phi_i \right\|_{L^\infty(\mathcal{O})}^p \nu(d\zeta_i)\, ds. \end{aligned} \tag{4.4}$$

Moreover, by (3.19) we know that for any $i = 1, \ldots, k_h$ we have

$$\mathcal{A}_h^{-\gamma} \phi_i = \sum_{l=1}^{k_h} u_{il} \phi_l^h,$$

where

$$u_{il} = \begin{cases} c\, h^{2\gamma}, & \text{if } (i,l) \subset \mathcal{C}_h, \\ 0, & \text{elsewhere,} \end{cases}$$

and there exists some constant m_0 such that for any i and $0 < h < 1$

$$\left| \{(i, \mathbb{N}^{k_h}) \cap \mathcal{C}_h \right| \leq m_0. \tag{4.5}$$

Therefore, we have

$$\left\| \mathcal{A}_h^{-\gamma} \phi_i \right\|_{L^\infty(\mathcal{O})} \leq m_0 h^{2\gamma}. \tag{4.6}$$

By substituting (4.6) in (4.4) we obtain

$$S_3 \leq \int_0^t \sum_{i=1}^{k_h} \lambda(J_i^h)$$
$$(t-s)^{-\gamma p}\, m_0\, h^{2\gamma p} \int_{\zeta\in\mathbb{R}} \mathbb{E}\left\|g(u_h(s,N_i^h),N_i^h;\zeta)\right\|^p_{L^p(\mathcal{O})}\, \nu(d\zeta_i)\, ds. \tag{4.7}$$

The inverse inequality given in Corollary 3.1 gives for all $\delta > d - \frac{d}{p}$

$$\begin{aligned} S_3 &\leq \int_0^t C \sum_{i=1}^{k_h} \lambda(J_i^h) \\ &\quad (t-s)^{-\gamma p}\, h^{2\gamma p-\delta p} \int_{\zeta\in\mathbb{R}} \mathbb{E}\left\|g(u_h(s,N_i^h),N_i^h;\zeta_i)\right\|^p_{B^{\frac{d}{p}-d}_{p,\infty}(\mathcal{O})}\, \nu(d\zeta)\, ds \\ &\leq C \int_0^t \sum_{i=1}^{k_h} \lambda(J_i^h)(t-s)^{-\gamma p}\, h^{(2\gamma-\delta)p} \int_{\zeta_i\in\mathbb{R}} \mathbb{E}\left\|g(u_h(s,N_i^h),N_i^h;\zeta)\right\|^p_{B^{\frac{d}{p}-d}_{p,\infty}(\mathcal{O})}\, \nu(d\zeta)\, ds, \end{aligned}$$

which gives

$$S_3 \leq C \int_0^t \sum_{i=1}^{k_h} \lambda(J_i^h)(t-s)^{-\gamma p}\, h^{(2\gamma-\delta)p} \int_{\zeta_i\in\mathbb{R}} \mathbb{E}\left|g(u_h(s,N_i^h),N_i^h;\zeta)\right|^p\, \nu(d\zeta)\, ds.$$

The mapping g is linear in the last variable and globally Lipschitz in the first variable. Thus we can infer that

$$S_3 \leq C \int_0^t \sum_{i=1}^{k_h} \lambda(J_i^h)(t-s)^{-\gamma p}\, h^{(2\gamma-\delta)p} \left(1 + \mathbb{E}\left|u_h(s,N_i^h)\right|^p\right)\, ds.$$

By the equivalence of the $L^p_h(\mathcal{O})$ and $L^p(\mathcal{O})$ norms we infer that

$$S_3 \leq C \int_0^t (t-s)^{-\gamma p}\, h^{(2\gamma-\delta)p} \left(1 + \mathbb{E}\left|u_h(s)\right|^p_{L^p(\mathcal{O})}\right)\, ds.$$

Let q be so large that $\gamma p < 1 - \frac{1}{q}$. Since there exists a $\delta > d - \frac{d}{p}$ such that $2\gamma - \delta > 0$, the Hölder inequality gives

$$\left(\mathbb{E}\left|S_3\right|\right)^q \leq c(t) \int_0^t \left(1 + \mathbb{E}\left|u_h(s)\right|^p_{L^p(\mathcal{O})}\right)^q\, ds.$$

Summing up leads to the assertion. □

5. Proof of the main result. The error of the time discretization follows from Theorem 3.1 of [HM06a]. By the Lipschitz continuity of f and g we infer that

$$\int_{z\in Z} \|G(u;z)\|^p_{B^{\frac{d}{p}-d}_{p,\infty}(\mathcal{O})}\, \iota(dz) \leq C\|u\|^p_{L^p(\mathcal{O})}, \quad u \in L^p(\mathcal{O}),$$

and

$$\|F(u)\|_{L^p(\mathcal{O})} \leq C\|u\|_{L^p(\mathcal{O})}, \quad u \in L^p(\mathcal{O}).$$

Thus, Theorem 3.1 of [HM06a] gives the order of convergence in time for $\alpha < \frac{2}{p} - d + \frac{d}{p}$. Here we investigate the accuracy of the space discretization, i.e., the difference $\Delta(t) = u(t) - u_h(t)\mathcal{I}_h$ at time t. To show the assertion, we will show that for a certain $q < \infty$ there exist constants C_1, C_2, and C_3 such that for all $0 < h \le 1$

$$\begin{aligned} &\left(\mathbb{E}\,\|\Delta(t)\|^p_{L^p(\mathcal{O})}\right)^q \\ &\le \frac{C_1}{t^q}\, h^{pq\delta_1} + C_2\, h^{pq\delta_2} + C_3\, h^{pq\delta_3} + C_4 \int_0^t \left(\mathbb{E}\,\|\Delta(t)\|^p_{L^p(\mathcal{O})}\right)^q ds \end{aligned} \tag{5.1}$$

where δ_1, δ_2, and δ_3 are the constants specified in Theorem 2.1. Now the second part of the assertion of Theorem 2.1 follows by an extension of Grownwall's lemma which is a simple consequence of Exercise 3, page 190 in [Hen81] and the proof of which is omitted.

LEMMA 5.1 (Gronwall-type lemma). *Suppose that $\alpha > 0$ and $c_1, c_2, c_3 > 0$. Assume that $u : \mathbb{R}^+ \to \mathbb{R}^+$ is a measurable function such that the function $t \mapsto t^{1-\alpha}u(t)$ is locally integrable on $\mathbb{R}^+$ and a.s. on $\mathbb{R}^+$,*

$$u(t) \le c_1 t^{\alpha-1} + c_2 + c_3 \int_0^t u(s)\, ds.$$

Then, a.s. on $\mathbb{R}^+$,

$$u(t) \le \left(c_1 t^{\alpha-1}\Gamma(\alpha) + c_2\right) \sum_{m=0}^{\infty} \frac{[c_3 t]^m}{\Gamma(m+\alpha)}.$$

To show the second part of Theorem 2.1, we will show the following: Let q and γ be two real numbers such that

$$d - \frac{d}{p} < \gamma < 2p \quad \text{and} \quad \frac{\gamma p}{2} < 1 - \frac{1}{q} \tag{5.2}$$

are satisfied. Then under the conditions of Theorem 2.1 for all δ_1, δ_2, and δ_3, with

$$\delta_1 = 2 - \frac{d}{p} + \frac{d}{2}, \quad \delta_2 < \frac{2}{p} - d + \frac{d}{p}, \text{ and } \delta_3 < \left(\frac{2}{p} - d + \frac{d}{p}\right)\left(2 + \frac{d}{2} - \frac{d}{p}\right),$$

there exist constants C_1, C_2, and C_3 depending on p, f, g, δ_1, δ_2, and δ_3 such that for all $0 < h \le 1$ inequality (5.1) holds.

Proof. By using the variation of constants formula, i.e.,

$$\begin{aligned} u(t) = e^{-At}u_0 &+ \int_0^t e^{-A(t-s)} f(u(s-))\, ds \\ &+ \int_{0^+}^t \int_Z e^{-A(t-s)} g(u(s-); z)\eta(ds; dz) \end{aligned}$$

and

$$\begin{aligned} u_h(t) = e^{-\mathcal{A}_h t}u_0 &+ \int_0^t e^{-\mathcal{A}_h(t-s)} f(u(s-))\, ds \\ &+ \int_{0^+}^t \int_Z e^{-\mathcal{A}_h(t-s)} g u(s-); z)\eta(ds; dz), \end{aligned}$$

we obtain

$$u(t) - u_h(t) = \sum_{i=1}^{6} I_{ni}(t), \tag{5.3}$$

where

$$\begin{aligned}
I_0(t) &= \left(e^{-At} - e^{-\mathcal{A}_h t}\right) u_0,\\
I_1(t) &= \int_{0+}^{t} \left(e^{-A(t-s)} - e^{-\mathcal{A}_h(t-s)}\mathcal{I}_h\right) f(u(s))\, ds,\\
I_2(t) &= \int_{0+}^{t} e^{-\mathcal{A}_h(t-s)} \left(\mathcal{I}_h f(u(s)) - f_h(u_h(s))\right)\, ds,\\
I_3(t) &= \int_{0+}^{t}\int_{Z} e^{-A(t-s)} \left(g(u(s);z) - g(u_h(s);z)\right)\, \tilde{\eta}(ds;dz),\\
I_4(t) &= \int_{0+}^{t}\int_{Z} e^{-A(t-s)} \left(I - \mathcal{I}_h\right) g(u_h(s);z)\, \tilde{\eta}(ds;dz),\\
I_5(t) &= \int_{0+}^{t}\int_{Z} e^{-A(t-s)}\mathcal{I}_h g(u_h(s);z)\, \tilde{\eta}(ds;dz)\\
&\quad - \int_{0+}^{t}\sum_{i=1}^{k_h}\int_{\zeta\in\mathbb{R}} e^{-A(t-s)} g_h(u_h(s);i;\zeta)\, \tilde{\eta}_h^i(ds,d\zeta),\\
I_6(t) &= \int_{0+}^{t}\sum_{i=1}^{k_h}\int_{\zeta\in\mathbb{R}} \left(e^{-A(t-s)} - e^{-\mathcal{A}_h(t-s)}\right) g_h(u_h(s);i;\zeta)\, \tilde{\eta}_h^i(ds;d\zeta).
\end{aligned}$$

By substituting (5.3), we get

$$\mathbb{E}\left[\|\Delta(t)\|^p\right] \leq C \cdot \sum_{i=0}^{4} \mathbb{E}\left[\|I_i(t)\|^p\right].$$

To give an estimate of $\mathbb{E}\left[\|I_0(t)\|^p\right]$, we apply Theorem 3.1. Thus, we get

$$\mathbb{E}\left[\|I_0(t)\|^p\right] \leq \frac{C}{t}\, h^{p\left(2-\frac{d}{p}+\frac{d}{2}\right)}\, \|u_0\|^p_{L^p(\mathcal{O})}. \tag{5.4}$$

The next entity we have to estimate is

$$\mathbb{E}\,\|I_1(t)\|^p_{L^p(\mathcal{O})} \leq \mathbb{E}\left(\int_{0+}^{t} \left\|e^{-A(t-s)} - e^{-\mathcal{A}_h(t-s)}\mathcal{I}_h f(u(s))\right\|_{L^p(\mathcal{O})}\, ds\right)^p.$$

In Theorem 3.4, the factor t^{-1} appears on the right-hand side. Thus, fix $\gamma < 1$, take into account

$$\left\|e^{-A(t-s)} - e^{-\mathcal{A}_h(t-s)}\mathcal{I}_h\right\|_{L(L^p(\mathcal{O}),L^p(\mathcal{O}))} \leq C,$$

and apply Theorem 3.4. Thus, we get

$$\begin{aligned}
&\mathbb{E}\,\|I_1(t)\|^p_{L^p(\mathcal{O})}\\
&\le C\,\left(\int_{0+}^t \left\|e^{-A(t-s)}-e^{-\mathcal{A}_h(t-s)}\mathcal{I}_h\right\|_{L(L^p(\mathcal{O}),L^p(\mathcal{O}))}\,\mathbb{E}\,\|f(u(s))\|^p_{L^p(\mathcal{O})}\,ds\right)^p\\
&\le C\,\Big(\int_{0+}^t \left\|e^{-A(t-s)}-e^{-\mathcal{A}_h(t-s)}\mathcal{I}_h\right\|^{\gamma}_{L(L^p(\mathcal{O}),L^p(\mathcal{O}))}\\
&\quad\times\left\|e^{-A(t-s)}-e^{-\mathcal{A}_h(t-s)}\mathcal{I}_h\right\|^{1-\gamma}_{L(L^p(\mathcal{O}),L^p(\mathcal{O}))}\,\mathbb{E}\,\|f(u(s))\|^p_{L^p(\mathcal{O})}\,ds\Big)^p\\
&\le C\,h^{\gamma(2-\frac{d}{p}+\frac{d}{2})p}\,\left(\int_{0+}^t (t-s)^{-\gamma}\,\mathbb{E}\,\|f(u(s))\|^p_{L^p(\mathcal{O})}\,ds\right)^p\\
&\le C(\gamma)\,h^{\gamma(2-\frac{d}{p}+\frac{d}{2})p}\sup_{0\le t\le T}\mathbb{E}\,\|f(u(s))\|^p_{L^p(\mathcal{O})}\,,
\end{aligned}$$

where

$$C(\gamma)=\frac{C'}{1-\gamma}.$$

The Lipschitz property of f gives

$$\mathbb{E}\,\|I_1(t)\|^p_{L^p(\mathcal{O})}\le C\,h^{p(2-\frac{d}{p}+\frac{d}{2})\gamma}\,\left(1+\sup_{0\le s\le t}\mathbb{E}\,\|u(s)\|^p_{L^p(\mathcal{O})}\right).$$

Since $u\in L^p(\Omega;L^p(\mathcal{O}))$, there exists some C such that

$$\mathbb{E}\,\|I_1(t)\|^p_{L^p(\mathcal{O})}\le C\,h^{p(2-\frac{d}{p}+\frac{d}{2})\gamma},\quad 0<h<1. \tag{5.5}$$

In section 3.3 we have shown that $\mathcal{A}_h$ generates a semigroup in $L^p_h(\mathcal{O})$. Thus, by the Minkowski inequality (4.2) and by the equivalence of the norms in $L^p(\mathcal{O})$ and $L^p_h(\mathcal{O})$, i.e., (3.16), we can write

$$\begin{aligned}
\mathbb{E}\,\|I_2(t)\|^p_{L^p(\mathcal{O})}&\le\mathbb{E}\left(\int_{0+}^t\|\mathcal{I}_hf(u(s))-f_h(u_h(s))\|_{L^p(\mathcal{O})}\,ds\right)^p\\
&\le C\,\mathbb{E}\left(\int_{0+}^t\|\mathcal{I}_hf(u(s))-f_h(u_h(s))\|_{L^p_h(\mathcal{O})}\,ds\right)^p.
\end{aligned}$$

The mapping f is Lipschitz continuous, and therefore

$$\mathbb{E}\,\|I_2(t)\|^p_{L^p(\mathcal{O})}\le C\,\mathbb{E}\left(\int_{0+}^t\|\mathcal{I}_hu(s)-u_h(s)\|_{L^p_h(\mathcal{O})}\,ds\right)^p.$$

Again, by the equivalence of the norms in $L^p(\mathcal{O})$ and $L^p_h(\mathcal{O})$, we can infer that

$$\mathbb{E}\,\|I_2(t)\|^p_{L^p(\mathcal{O})}\le C\,\mathbb{E}\left(\int_{0+}^t\|u(s)-u_h(s)\|_{L^p(\mathcal{O})}\,ds\right)^p.$$

The Hölder inequality gives

$$\mathbb{E}\,\|I_2(t)\|^p_{L^p(\mathcal{O})}\le C\,\mathbb{E}\int_{0+}^t\|u(s)-u_h(s)\|^p_{L^p(\mathcal{O})}\,ds. \tag{5.6}$$

Next, we handle the third term. The generalized Burkholder inequality (compare Proposition 3.3 in [Hau05]) gives

$$\mathbb{E}\,\|I_3(t)\|^p_{L^p(\mathcal{O})} \leq \int_{0+}^t \int_Z \mathbb{E}\left\|e^{-A(t-s)}\left(g(u(s);z)-g(u_h(s);z)\right)\right\|^p_{L^p(\mathcal{O})}\,\iota(dz)\,ds.$$

The operator A generates an analytic semigroup in $L^p(\mathcal{O})$ (see, e.g., Theorem 7.3.5 in [Paz83]). Therefore, for each $\rho>0$ there exists some constant $C<\infty$ such that

$$\left\|e^{-(t-s)A}Av\right\|_{W_p^{-\rho}(\mathcal{O})} \leq \frac{C}{t^{\frac{\rho}{2}}}\,\|v\|_{L^p(\mathcal{O})}\,,\quad v\in L^p(\mathcal{O}). \tag{5.7}$$

Thus, we can infer for any $\gamma>0$ that

$$\mathbb{E}\,\|I_3(t)\|^p_{L^p(\mathcal{O})} \leq \int_{0+}^t \int_Z (t-s)^{-\frac{\gamma p}{2}}\,\mathbb{E}\,\|g(u(s);z)-g(u_h(s);z)\|^p_{W_p^{-\gamma}(\mathcal{O})}\,\iota(dz)\,ds.$$

By Sobolev inequalities for any $\gamma > d-\frac{d}{p}$ we have $Z = B^{\frac{d}{p}-d}_{p,\infty}(\mathcal{O}) \hookrightarrow W_p^{-\gamma}(\mathcal{O})$ continuously. Therefore,

$$\mathbb{E}\,\|I_3(t)\|^p_{L^p(\mathcal{O})} \leq \int_{0+}^t \int_Z (t-s)^{-\frac{\gamma p}{2}}\,\mathbb{E}\,\|(g(u(s);z)-g(u_h(s);z))\|^p_Z\,\iota(dz)\,ds.$$

The Lipschitz property of g leads to

$$\mathbb{E}\,\|I_3(t)\|^p_{L^p(\mathcal{O})} \leq \int_{0+}^t (t-s)^{-\frac{\gamma p}{2}}\,\mathbb{E}\,\|u(s)-u_h(s)\|^p_{L^p(\mathcal{O})}\,ds.$$

If γ satisfies (5.2), then $\frac{\gamma p}{2} < 1-\frac{1}{q}$, and we can infer by the Hölder inequality that

$$\left(\mathbb{E}\,\|I_3(t)\|^p_{L^p(\mathcal{O})}\right)^q \leq c(t)\int_0^t \left(\mathbb{E}\,\|u(s)-u_h(s)\|^p_{L^p(\mathcal{O})}\right)^q ds. \tag{5.8}$$

By arguments similar to those used in Proposition 3.2 in [Hau05] we get for the fourth term

$$\mathbb{E}\,\|I_4(t)\|^p_{L^p(\mathcal{O})} \leq \int_{0+}^t \int_Z \mathbb{E}\left\|e^{-A(t-s)}\left(I-\mathcal{I}_h\right)g(u_h(s);z)\right\|^p_{L^p(\mathcal{O})}\,\nu(dz)\,ds.$$

By the smoothing property of A, i.e., (5.7), and since $Z = B^{\frac{d}{p}-d}_{p,\infty}(\mathcal{O}) \hookrightarrow W_p^{-\gamma}(\mathcal{O})$ continuously for all $\gamma > d-\frac{d}{p}$, we obtain for any $\rho < \frac{2}{p}-d+\frac{d}{p}$

$$\begin{aligned}
&\mathbb{E}\,\|I_4(t)\|^p_{L^p(\mathcal{O})}\\
&\leq \int_{0+}^t (t-s)^{-\frac{p}{2}\left(\rho+d-\frac{d}{p}\right)} \int_Z \mathbb{E}\left\|A^{-\rho}\left(I-\mathcal{I}_h\right)G(u_h(s);z)\right\|^p_Z\,\iota(dz)\,ds\\
&\leq \int_{0+}^t (t-s)^{-\frac{p}{2}\left(\rho+d-\frac{d}{p}\right)} \int_{\zeta\in\mathbb{R}} \mathbb{E}\left\|A^{-\rho}\left(I-\mathcal{I}_h\right)g(u_h(s),\cdot,\zeta)\right\|^p_{L^p(\mathcal{O})}\,\nu(d\zeta)\,ds,
\end{aligned}$$

where G is defined in (3.20). Theorem 3.1 leads to

$$\mathbb{E}\,\|I_4(t)\|^p_{L^p(\mathcal{O})} \leq C(t)\,h^{p\rho}\sup_{0\leq s\leq t}\int_{\zeta\in\mathbb{R}}\mathbb{E}\,\|g(u_h(s),\cdot;\zeta)\|^p_{L^p(\mathcal{O})}\,\nu(d\zeta)$$

and the Lipschitz property of g to

$$\mathbb{E}\,\|I_4(t)\|^p_{L^p(\mathcal{O})} \le C(t)\,h^{p\rho}\,\Big(1+\sup_{0\le s\le t}\mathbb{E}\,\|u_h(s)\|^p_{L^p(\mathcal{O})}\Big).$$

By Proposition 4.1 there exists some constant $C<\infty$ such that

$$\sup_{0\le s\le t}\mathbb{E}\,\|u_h(s)\|^p_{L^p(\mathcal{O})} \le C, \quad h>0.$$

Thus, we can infer that there exists some constant $C'<\infty$ such that

$$\mathbb{E}\,\|I_4(t)\|^p_{L^p(\mathcal{O})} \le C'\,h^{p\rho}, \quad 0\le t\le T. \tag{5.9}$$

The Burkholder inequality leads for any $\frac{1}{2}(d-\frac{d}{p})<\gamma<\frac{1}{p}$ to

$$\begin{aligned}
&\mathbb{E}\,\|I_5(t)\|^p_{L^p(\mathcal{O})}\\
&\le \int_0^t\int_Z \Big\|(-\mathcal{A}_h)^{\gamma}e^{-A(t-s)}(-\mathcal{A}_h)^{-\gamma}\mathcal{I}_h G(u(t),z)\Big\|^p_{L^p(\mathcal{O})}\ (\iota-\iota_h)(dz)\\
&\le \int_0^t (t-s)^{-\gamma p}\,dt \times \sup_{0\le t\le T}\int_Z \big\|(-\mathcal{A}_h)^{-\gamma}\mathcal{I}_h G(u(t),z)\big\|^p_{L^p(\mathcal{O})}\ (\iota-\iota_h)(dz).
\end{aligned}$$

Proposition 3.2 gives for any $0\le\delta<\frac{1}{p}-\frac{1}{2}(d-\frac{d}{p})$

$$\mathbb{E}\,\|I_5(t)\|^p_{L^p(\mathcal{O})} \le C\,\frac{h^{\delta p}}{p\gamma-1}. \tag{5.10}$$

Next, we have to calculate the entity I_6. The Fubini theorem and the generalized Burkholder inequality give

$$\begin{aligned}
\mathbb{E}\,\|I_6(t)\|^p_{L^p(\mathcal{O})} &= \mathbb{E}\left\|\int_{0+}^t\sum_{i=1}^{k_h}\int_{\zeta\in\mathbb{R}}\Big(e^{-A(t-s)}-e^{-\mathcal{A}_h(t-s)}\Big)\,g_h(u_h(s);i,\zeta)\,\tilde{\eta}^i_h(ds;dz)\right\|^p_{L^p(\mathcal{O})}\\
&\le \mathbb{E}\int_{0+}^t\sum_{i=1}^{k_h}\int_{\zeta\in\mathbb{R}}\Big\|\Big(e^{-A(t-s)}-e^{-\mathcal{A}_h(t-s)}\Big)\,g_h(u_h(s);i,\zeta)\Big\|^p_{L^p(\mathcal{O})}\,\nu^i(d\zeta)\,ds\\
&\le \int_{0+}^t\sum_{i=1}^{k_h}\int_{\zeta\in\mathbb{R}}\mathbb{E}\Big\|\Big(e^{-\frac{1}{2}A(t-s)}\\
&\qquad -e^{-\frac{1}{2}\mathcal{A}_h(t-s)}\mathcal{I}_h\Big)e^{-\frac{1}{2}A(t-s)}g_h(u_h(s);i,\zeta)\Big\|^p_{L^p(\mathcal{O})}\,\nu^i(d\zeta)\,ds\\
&+ \int_{0+}^t\sum_{i=1}^{k_h}\int_{\zeta\in\mathbb{R}}\mathbb{E}\Big\|e^{-\frac{1}{2}\mathcal{A}_h(t-s)}\mathcal{I}_h\Big(e^{-\frac{1}{2}A(t-s)}\\
&\qquad -e^{-\frac{1}{2}\mathcal{A}_h(t-s)}\Big)g_h(u_h(s);i,\zeta)\Big\|^p_{L^p(\mathcal{O})}\,\nu^i(d\zeta)\,ds\\
&:= I_{61}+I_{62}.
\end{aligned}$$

First, in Theorem 3.4 we have shown that

$$\Big\|\Big[e^{-\frac{1}{2}A(t-s)}-e^{-\frac{1}{2}\mathcal{A}_h(t-s)}\mathcal{I}_h\Big]\,u\Big\|_{L^p(\mathcal{O})} \le \frac{C}{t}\,h^{2+\frac{d}{2}-\frac{d}{p}}\|u\|_{L^p(\mathcal{O})}\ \forall u\in L^p(\mathcal{O}). \tag{5.11}$$

Let us recall estimate (3.17). There exists a constant $C<\infty$ such that for all $h>0$ and $v\in L^p_h(\mathcal{O})\cap\mathcal{V}_h$

$$\left\|e^{-(t-s)\mathcal{A}_h}v\right\|_{L^p_h}\leq C\left\|v\right\|_{L^p_h},\quad v\in L^p_h.$$

Moreover, the norms in $L^p_h(\mathcal{O})$ and in $L^p(\mathcal{O})$ are equivalent (see (3.16)). Therefore

$$\left\|e^{-(t-s)\mathcal{A}_h}v\right\|_{L^p_h}\leq C\left\|v\right\|_{L^p(\mathcal{O})},\quad v\in L^p_h(\mathcal{O}).$$

Thus, we can infer for any $\gamma\in(0,1)$ that

$$\begin{aligned}I_{61}&\leq C\,h^{p\gamma(2+\frac{d}{2}-\frac{d}{p})}\\&\times\int_{0+}^t(t-s)^{-\frac{\gamma p}{2}}\sum_{i=1}^{k_h}\lambda(J_i^h)\int_{\zeta\in\mathbb{R}}\mathbb{E}\left\|e^{-\frac{1}{2}A(t-s)}g_h(u_h(s);i,\zeta)\right\|^p_{L^p(\mathcal{O})}\,\nu(d\zeta)\,ds\end{aligned}$$

holds. The operator A generates an analytic semigroup in $L^p(\mathcal{O})$. Therefore, for each $\rho>0$ there exists some constant $C<\infty$ such that

$$\left\|e^{-tA}Av\right\|_{L^p(\mathcal{O})}\leq\frac{C}{t^{\frac{\rho}{2}}}\left\|v\right\|_{W_p^{-\rho}(\mathcal{O})},\quad v\in W_p^{-\rho}(\mathcal{O}),\ t>0.\tag{5.12}$$

Inequality (5.12) leads for all $\rho\geq0$, with $p(\gamma+\rho)<2$, to

$$\begin{aligned}I_{61}&\leq C\,h^{p\gamma(2+\frac{d}{2}-\frac{d}{p})}\\&\quad\times\int_{0+}^t(t-s)^{-\frac{p}{2}(\gamma+\rho)}\sum_{i=1}^{k_h}\lambda(J_i^h)\int_{\zeta\in\mathbb{R}}\mathbb{E}\left\|g_h(u_h(s);i,\zeta)\right\|^p_{W_p^{-\rho}(\mathcal{O})}\,\nu(d\zeta)\,ds\\&\leq C(\gamma,\rho)\,h^{p\gamma(2+\frac{d}{2}-\frac{d}{p})}\sup_{0\leq s\leq t}\sum_{i=1}^{k_h}\lambda(J_i^h)\int_{\zeta\in\mathbb{R}}\mathbb{E}\left\|g_h(u_h(s);i,\zeta)\right\|^p_{W_p^{-\rho}(\mathcal{O})}\,\nu(d\zeta),\end{aligned}$$

where

$$C(\gamma,\rho)=\frac{C'}{2-p(\gamma+\rho)}.$$

By Sobolev inequalities, we know that $L^1(\mathcal{O})\hookrightarrow W_p^{-\rho}(\mathcal{O})$ continuously. Thus we obtain

$$\begin{aligned}I_{61}&\leq C\,h^{p\gamma(2+\frac{d}{2}-\frac{d}{p})}\sup_{0\leq s\leq t}\sum_{i=1}^{k_h}\lambda(J_i^h)\int_{\zeta\in\mathbb{R}}\mathbb{E}\left\|g_h(u_h(s);i,\zeta)\right\|^p_{L^1(\mathcal{O})}\,\nu(d\zeta)\\&\leq C\,h^{p\gamma(2+\frac{d}{2}-\frac{d}{p})}\sup_{0\leq s\leq t}\sum_{i=1}^{k_h}\lambda(J_i^h)\int_{\zeta\in\mathbb{R}}\mathbb{E}\left|g(u_h(s,N_i^h),N_i^h;\zeta)\right|^p\left\|c_i^h\phi_i^h\right\|^p_{L^1(\mathcal{O})}\,\nu(d\zeta)\\&\leq C\,h^{p\gamma(2+\frac{d}{2}-\frac{d}{p})}\sup_{0\leq s\leq t}\sum_{i=1}^{k_h}\lambda(J_i^h)\int_{\zeta\in\mathbb{R}}\mathbb{E}\left|g(u_h(s,N_i^h),N_i^h;\zeta)\right|^p\,\nu(d\zeta).\end{aligned}$$

Since g is Lipschitz continuous and linear in the third variable, we get

$$I_{61}\leq C\,h^{p\gamma(2+\frac{d}{2}-\frac{d}{p})}\sup_{0\leq s\leq t}\sum_{i=1}^{k_h}\lambda(J_i^h)\left(1+\mathbb{E}\left\|u_h(s,N_i^h)\right\|^p_{L^p_h}\right)\,ds.$$

In section 3.3 we have seen that there exist some constants c and C such that for all $0 < h < 1$ we have (see (3.16))

$$c \|w\|_{L^p(\mathcal{O})} \leq h^{p\gamma(2+\frac{d}{2}-\frac{d}{p})} \|w_h\|_{L_h^p} \leq C \|w\|_{L^p(\mathcal{O})}, \quad w \in \mathcal{V}_h.$$

Thus, by taking into account Remark 3.1 we can write

$$I_{61} \leq Ct^{1-p(\gamma+\rho)} h^{p\gamma(2+\frac{d}{2}-\frac{d}{p})} \sup_{0\leq s\leq t} \left(1 + \mathbb{E}\|u_h(s)\|^p_{L^p(\mathcal{O})}\right).$$

Since $\mathbb{E}\|u_h(s)\|^p_{L^p(\mathcal{O})}$ is bounded, there exists a constant $C < \infty$ such that

$$I_{61} \leq Ct^{1-p(\gamma+\rho)} h^{p\gamma(2+\frac{d}{2}-\frac{d}{p})}. \tag{5.13}$$

Next, we have to handle the term I_{62}. Because of (3.18) and (3.16), there exists a constant $C < \infty$ such that for all $0 < h \leq 1$

$$\left\| e^{-(t-s)\mathcal{A}_h} \mathcal{A}_h v \right\|_{L^p(\mathcal{O})} \leq \frac{C}{t} \|v\|_{L^p(\mathcal{O})}, \quad v \in L^p(\mathcal{O}). \tag{5.14}$$

Because of inequality (5.14), we have for all $0 < \delta < \frac{1}{p}$

$$\begin{aligned} I_{62} \leq & \int_0^t (t-s)^{-\delta p} \sum_{i=1}^{k_h} \lambda(J_i^h) \\ & \int_{\zeta\in\mathbb{R}} \mathbb{E}\left\| (-\mathcal{A}_h)^{-\delta} \left(\mathcal{I}_h e^{-\frac{1}{2}A(t-s)} - e^{-\frac{1}{2}\mathcal{A}_h(t-s)} \right) g_h(u_h(s); i, \zeta) \right\|^p_{L^p(\mathcal{O})} \nu(d\zeta)\, ds. \end{aligned} \tag{5.15}$$

In Remark 3.1 we have shown that there exist some constants c and C such that $h^d \leq ck_h \leq h^d$. By (3.19) we know that for any $i = 1, \ldots, k_h$ we have

$$(-\mathcal{A}_h)^{-\delta} \phi_i = \sum_{l=1}^{k_h} u_{il} \phi_l^h, \tag{5.16}$$

where

$$u_{il} = \begin{cases} c\, h^{2\delta}, & \text{if } (i,l) \subset \mathcal{C}_h, \\ 0, & \text{elsewhere.} \end{cases}$$

Moreover, there exists some constant m_0 such that for any i and $0 < h < 1$

$$\left| \{(i, \mathbb{N}^{k_h}) \cap \mathcal{C}_h\} \right| \leq m_0.$$

By substituting (5.16) in (5.15) we obtain

$$\begin{aligned} I_{62} \leq & h^{2\delta p} \int_0^t (t-s)^{-\delta p} m_0 \sum_{i=1}^{k_h} \lambda(J_i^h) \\ & \int_{\zeta\in\mathbb{R}} \mathbb{E}\left\| \left(\mathcal{I}_h e^{-\frac{1}{2}A(t-s)} - e^{-\frac{1}{2}\mathcal{A}_h(t-s)} \right) g_h(u_h(s); i, \zeta) \right\|^p_{L^p(\mathcal{O})} \nu^i(d\zeta)\, ds. \end{aligned}$$

By applying Theorem 3.4 we can infer that for any $\gamma \in (01)$, we have

$$I_{62} \leq C\, h^{p\delta+p\gamma(2+\frac{d}{2}-\frac{d}{p})} \int_0^t (t-s)^{-(\delta+\gamma)p} m_0 \sum_{i=1}^{k_h} \lambda(J_i^h) \int_{\zeta\in\mathbb{R}} \mathbb{E}\Big\| g_h(u_h(s);i,\zeta)\Big\|_{L^p(\mathcal{O})}^p \nu(d\zeta)\, ds,$$

where $\rho = \gamma/(2-\frac{d}{p}+\frac{d}{2})$. By Theorem 3.1 and since $Z = B^{\frac{d}{p}-d}_{p,\infty}(\mathcal{O}) \hookrightarrow W_p^{-\gamma}(\mathcal{O})$ continuously, we get for any $\rho > d - \frac{d}{p}$

$$I_{62} \leq C\, h^{p(\delta-\rho)+p\gamma(2+\frac{d}{2}-\frac{d}{p})} \int_0^t (t-s)^{-(\delta+\rho)p} m_0 \sum_{i=1}^{k_h} \lambda(J_i^h) \int_{\zeta\in\mathbb{R}} \mathbb{E}\Big\| g_h(u_h(s);i,\zeta)\Big\|_{Z}^p \nu(d\zeta)\, ds.$$

Provided that $\delta \geq \rho$ we can omit the first term in the exponent of h. The mapping g is linear in the last variable and globally Lipschitz in the first. Thus we can infer that

$$I_{62} \leq C\, h^{p\gamma(2+\frac{d}{2}-\frac{d}{p})} \int_0^t (t-s)^{-(\delta+\rho)p} m_0 \sum_{i=1}^{k_h} \lambda(J_i^h) \left(1+\mathbb{E}\left|u_h(s,N_i^h)\right|^p\right) ds.$$

By Remark 3.1 it follows that

$$I_{62} \leq C\, h^{p\gamma(2+\frac{d}{2}-\frac{d}{p})} \int_0^t (t-s)^{-(\delta+\rho)p} m_0 \sum_{i=1}^{k_h} h^d \left(1+\mathbb{E}\left|u_h(s,N_i^h)\right|^p\right) ds,$$

and by (3.16)

$$I_{62} \leq C\, h^{p\gamma(2+\frac{d}{2}-\frac{d}{p})} \int_0^t (t-s)^{-(\delta+\rho)p} \left(1+\mathbb{E}\,\|u_h(s)\|_{L^p(\mathcal{O})}^p\right) ds.$$

The Hölder inequality leads to

$$I_{62} \leq C\, h^{p\gamma(2+\frac{d}{2}-\frac{d}{p})}\, C(t) \left(1+\sup_{0\leq s\leq t} \mathbb{E}\,\|u_h(s)\|_{L^p(\mathcal{O})}^p\right)$$

and Theorem 4.1 that there exists a constant C such that

$$I_{62} \leq C\, h^{p\gamma(2+\frac{d}{2}-\frac{d}{p})}, \quad 0 < h \leq 1. \tag{5.17}$$

Collecting together (5.4), (5.5), (5.6), (5.8), (5.9), (5.10), (5.13), and (5.17) gives the assertion. ∎

6. Discussions and conclusions. The next step is to confirm the theory described in this paper by numerical experiments. One of the problems which arises here is that, if the intensity of the Poisson random measure is given, the distribution is known only in a few cases. Therefore, it is difficult to find a benchmark where the solution is known analytically and to compare the numerical result with the exact solution.

Moreover, because of this problem, the distribution of the random variables $\xi_h^m(\cdot)$, $m \in \mathbb{N}$, appearing in (2.6), is not given a priori. Different approaches exist to solve this problem; e.g., Rubenthaler [Rub03] cut off the jumps, which are smaller than a certain threshold, calculated the time of the first jump, and applied the Euler scheme until the time of the first jump. Then he simulated a random variable according to the distribution of size of the jump, with the condition that the jump is larger than the threshold. This step is repeated until the time T is reached.

In [HM06a, HM06b] we approximated the intensity by a discrete distribution. In fact, first we cut off the small jumps by setting $\nu_h(B) = 0$ for a neighborhood B of the origin, and then we approximated ν_h by a discrete distribution ν_h^n. To be precise, let $\tilde{\eta}_h^n$ be the real-valued compensated Poisson random measure given by the intensity ν_h^n. Now for any $x \in \mathbb{R}$ the random variable

$$\int_{m\tau_h}^{(m+1)\tau_h} \int_{\zeta \in \mathbb{R}^{k_h}} g(x;\zeta)\, \tilde{\eta}_h^n(dz;ds)$$

can be simulated directly by the sum of N independent random variables with distribution $\frac{1}{\nu_h^n(\mathbb{R})}\nu_h^n$, where N is exponentially distributed with parameter $\nu_h^n(\mathbb{R})$.

By taking into account that the small jumps are cut off, one may ask if the perturbation of the small jumps can be approximated by, e.g., a Wiener process. This has been done, e.g., by Kallianpur and Xiong [KX95]. Since we are in Banach spaces of M-type p, $1 \le p \le 2$, the convergence of a simple random walk is not necessarily given. One alternative is to choose another scaling, which we also consider in [HM06b].

Appendix A. Besov spaces and their properties. An introduction to Besov spaces is given by Runst and Sickel [RS96]. We are interested is the continuity of the mapping G described in (3.20). To be precise, we prove the following result.

PROPOSITION A.1. *If $f \in \mathcal{S}(\mathbb{R}^d)$, then for each $a \in \mathbb{R}^d$ the Schwartz distribution $f\delta_a$ belongs to the Besov space $B_{p,\infty}^{\frac{d}{p}-d}(\mathbb{R}^d)$, and moreover there exists a constant $C = C(d,p)$ independent of f such that*

$$\int_{\mathbb{R}^d} |f\delta_a|^p_{B_{p,\infty}^{\frac{d}{p}-d}(\mathbb{R}^d)}\, da = |f|^p_{L^p(\mathbb{R}^d)}\,.$$

Thus, there exists a unique bounded linear map $\Lambda : L^p(\mathbb{R}^d) \to L^p(\mathbb{R}^d, B_{p,\infty}^{\frac{n}{p}-n}(\mathbb{R}^d))$ such that $[\Lambda(f)](a) = f\delta_a$, $f \in \mathcal{S}(\mathbb{R}^d)$, $a \in \mathbb{R}^d$. Moreover, for any $s \ge 0$ there exists some constant C such that

$$\int_{\mathbb{R}^d} |f\delta_a|^p_{B_{p,\infty}^{\frac{d}{p}-d-s}(\mathbb{R}^d)}\, da \le C\ |f|^p_{W_p^{-s}(\mathbb{R}^d)}\,.$$

In the following we denote the value of $\Lambda(f)$ at a, where $f \in L^p(\mathbb{R}^d)$ by $f\delta_a$. Let us recall the definition of the Besov spaces as given in [RS96, Definition 2, pages 7–8]. First we choose a function $\psi \in \mathcal{S}(\mathbb{R}^d)$ such that $0 \le \psi(x) \le 1$, $x \in \mathbb{R}^d$, and

$$\psi(x) = \begin{cases} 1, & \text{if} \quad |x| \le 1, \\ 0 & \text{if} \quad |x| \ge \frac{3}{2}. \end{cases}$$

Then put

$$\begin{cases} \phi_0(x) = \psi(x),\ x \in \mathbb{R}^d, \\ \phi_1(x) = \psi(\frac{x}{2}) - \psi(x),\ x \in \mathbb{R}^d, \\ \phi_j(x) = \phi_1(2^{-j+1}x),\ x \in \mathbb{R}^d, \quad j = 2,3,\ldots. \end{cases}$$

With the choice of $\phi = \{\phi_j\}_{j=0}^{\infty}$ as above and $\mathcal{F}$ and $\mathcal{F}^{-1}$ being the Fourier and the inverse Fourier transformations, respectively, (acting in the space $\mathcal{S}'(\mathbb{R}^d)$ of Schwartz distributions), we have the following definition.

DEFINITION A.1. *Let $s \in \mathbb{R}$, $0 < p \leq \infty$, and $0 < q \leq \infty$. If $0 < q < \infty$ and $f \in \mathcal{S}'(\mathbb{R}^d)$, we put*

$$|f|_{B^s_{p,q}} = \left(\sum_{j=0}^{\infty} 2^{sjq} \left| \mathcal{F}^{-1} \left[\phi_j \mathcal{F} f \right] \right|_{L^p}^q \right)^{\frac{1}{q}},$$
$$|f|_{B^s_{p,\infty}} = \sup_{j \in \mathbb{N}} 2^{sj} \left| \mathcal{F}^{-1} \left[\phi_j \mathcal{F} f \right] \right|_{L^p}.$$

We denote by $B^s_{p,q}(\mathbb{R}^d)$ the space of all $f \in \mathcal{S}'(\mathbb{R}^d)$ for which $|f|_{B^s_{p,q}}$ is finite.

LEMMA A.2. *If $h \in \mathcal{S}(\mathbb{R}^d)$, $h \in \mathcal{S}(\mathbb{R}^d)$, and $a \in \mathbb{R}^d$, then*

$$|(h\delta_a) * g|_{L^p(\mathbb{R}^d)} = |h(a)||g|_{L^p(\mathbb{R}^d)}.$$

Proof. Let us recall that $h\delta_b \in \mathcal{S}'(\mathbb{R}^d)$ assigns to a function $g \in \mathcal{S}(\mathbb{R}^d)$ a value $\delta_a(hg) = h(a)g(a)$. Since by the definition of the convolution of a distribution with a test function

$$[(h\delta_a) * g](x) = (h\delta_a)(\tau_x \check{g}) = h(a)(\tau_x \check{g})(a) = h(a)\check{g}(a - x) = h(a)g(x - a),\ x \in \mathbb{R}^d,$$

we have

$$\begin{aligned} |(h\delta_a) * g|^p_{L^p(\mathbb{R}^d)} &= \int_{\mathbb{R}^d} |g(x - a)h(a)|^p \, dx \\ &= |h(a)|^p \int_{\mathbb{R}^d} |g(x)|^p \, dx = |h(a)|^p \ |g|^p_{L^p(\mathbb{R}^d)}. \end{aligned}$$ ∎

LEMMA A.3. *If $\varphi \in \mathcal{S}(\mathbb{R}^d)$, $\lambda > 0$, and $g(x) := \varphi(\lambda x)$, $x \in \mathbb{R}^d$, then*

$$|\mathcal{F}^{-1} g|_{L^p(\mathbb{R}^d)} = \lambda^{d(\frac{1}{p} - 1)} |\mathcal{F}^{-1} \varphi|_{L^p(\mathbb{R}^d)}.$$

Proof. The proof follows with simple calculations. ∎

Proof of Proposition A.1. Obviously it is enough to prove the first part of the proposition. We will use the definition of the Fourier transform as in [RS96]. In particular, with $\langle \cdot, \cdot \rangle$ being the scalar product in $\mathbb{R}^d$, we put

$$(\mathcal{F} f)(\xi) := (2\pi)^{-d/2} \int_{\mathbb{R}^d} e^{-i\langle x, \xi \rangle} f(x) \, dx,\ \xi \in \mathbb{R}^d,\ f \in \mathcal{S}(\mathbb{R}^d).$$

Let us fix $f \in \mathcal{S}(\mathbb{R}^d)$. Since $\mathcal{F}^{-1}(\varphi u) = (2\pi)^{-d/2}(\mathcal{F}^{-1}\varphi) * (\mathcal{F}^{-1} u)$, $\varphi \in \mathcal{S}$, $u \in \mathcal{S}'$, we infer that, for $j \in \mathbb{N}^*$,

$$\begin{aligned} |\mathcal{F}^{-1}[\varphi_j \mathcal{F}(f\delta_a)]|_{L^p(\mathbb{R}^d)} &= (2\pi)^{-d/2} |(\mathcal{F}^{-1}\varphi_j) * (f\delta_a)|_{L^p(\mathbb{R}^d)} \\ &= (2\pi)^{-d/2} |f(a)| \mathcal{F}^{-1}\varphi_j|_{L^p(\mathbb{R}^d)} \\ &= (2\pi)^{-d/2} 2^{d(\frac{1}{p} - 1)} 2^{-jd(\frac{1}{p} - 1)} |f(a)| |\mathcal{F}^{-1}\varphi_1|_{L^p(\mathbb{R}^d)}. \end{aligned}$$

Hence, $f\delta_a$ belongs to the Besov space $B^{d(\frac{1}{p} - 1)}_{p,\infty}(\mathbb{R}^d)$ as requested, and the equality (A.1) follows immediately. In order to show (A.1) note that

$$\|f\|_{B^{s+\gamma}_{p,\infty}(\mathbb{R}^d)} = \|I(s) f\|_{B^{\gamma}_{p,\infty}(\mathbb{R}^d)},$$

where $I(s) = \mathcal{F}^{-1}\left(1+\xi^2\right)^{\frac{s}{2}}\mathcal{F}$. Next,

$$\begin{aligned}
&\left\|\mathcal{F}^{-1}\left(\varphi_j\left(1+\xi^2\right)^{\frac{s}{2}}\mathcal{F}\left(\int_{\mathbb{R}^d}\delta_a f\,da\right)\right)\right\|_{L^p(\mathbb{R}^d)}\\
&\le \left\|\mathcal{F}^{-1}\varphi_j \star \mathcal{F}^{-1}\left(1+\xi^2\right)^{\frac{s}{2}}\mathcal{F}\left(\int_{\mathbb{R}^d}\delta_a f\,da\right)\right\|_{L^p(\mathbb{R}^d)}\\
&\le \left\|\mathcal{F}^{-1}\varphi_j\right\|_{L^p(\mathbb{R}^d)}\left\|\mathcal{F}^{-1}\left(1+\xi^2\right)^{\frac{s}{2}}\mathcal{F}\left(\int_{\mathbb{R}^d}\delta_a f\,da\right)\right\|_{L^1(\mathbb{R}^d)}.
\end{aligned}$$

Moreover,

$$\begin{aligned}
&\left\|\mathcal{F}^{-1}\left(1+\xi^2\right)^{\frac{s}{2}}\mathcal{F}\left(\int_{\mathbb{R}^d}\delta_a f\,da\right)\right\|_{L^1(\mathbb{R}^d)}\\
&= \int_{\mathbb{R}^d}\left|\int_{\mathbb{R}^d} e^{i\langle y,\xi\rangle}\,\left(1+\xi^2\right)^{\frac{s}{2}}\int_{\mathbb{R}^d} e^{-i\langle a,\xi\rangle}f(a)\,d\xi\,da\right|\,dy\\
&\quad= \int_{\mathbb{R}^d}\left|\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} e^{i\langle y,\xi\rangle}\,\left(1+\xi^2\right)^{\frac{s}{2}} e^{-i\langle a,\xi\rangle}f(a)\,da\,d\xi\right|\,dy\\
&= \int_{\mathbb{R}^d}\left|\int_{\mathbb{R}^d} e^{i\langle y,\xi\rangle}\,\mathcal{F}\left[(I+\Delta)^{\frac{s}{2}}f\right](\xi)\,d\xi\right|dy = \int_{\mathbb{R}^d}\left(\mathcal{F}^{-1}\mathcal{F}\left[(I+\Delta)^{\frac{s}{2}}f\right]\right)(y)\,dy\\
&\le \left\|\mathcal{F}^{-1}\mathcal{F}\left[(I+\Delta)^{\frac{s}{2}}f\right]\right\|_{L^1(\mathbb{R}^d)} = \left\|(I+\Delta)^{\frac{s}{2}}f\right\|_{L^1(\mathbb{R}^d)} = \|f\|_{W_p^s(\mathbb{R}^d)}.
\end{aligned}$$

In summary, one can show that for any $s\in\mathbb{R}$ there exists some C such that

$$\left\|\int_{\mathbb{R}^d} f\delta_a\,da\right\|_{B_p^{s+\frac{d}{p}-d}(\mathbb{R}^d)} \le C\,\|f\|_{W_p^s(\mathbb{R}^d)}, \quad f\in C_b^0(\mathbb{R}^d).$$

Since the operator $I(s)$ is positive for $s\le 0$, it follows that

$$\int_{\mathbb{R}^d}\|f\delta_a\|_{B_p^{s+\frac{d}{p}-d}(\mathbb{R}^d)}\,da \le C\,\|f\|_{W_p^s(\mathbb{R}^d)}, \quad f\in C_b^0(\mathbb{R}^d),\ \ f\ge 0.$$

Theorem 5.1.2 in [BL76] gives inequality (A.1). □

REFERENCES

[ANZ98] E. J. Allen, S. J. Novosel, and Z. Zhang, *Finite element and difference approximation of some linear stochastic partial differential equations*, Stoch. Stoch. Rep., 64 (1998), pp. 117–142.

[AW00] D. Applebaum and J.-L. Wu, *Stochastic partial differential equations driven by Lévy space-time white noise*, Random Oper. Stochastic Equations, 8 (2000), pp. 245–259.

[AWZ98] S. Albeverio, J. Wu, and T. Zhang, *Parabolic SPDEs driven by Poisson white noise*, Stochastic Process. Appl., 74 (1998), pp. 21–36.

[BL76] J. Bergh and J. Löfström, *Interpolation Spaces: An Introduction*, Grundlehren Math. Wiss. 223, Springer-Verlag, New York, 1976.

[BS02] S. Brenner and R. Scott, *The Mathematical Theory of Finite Element Methods*, 2nd ed., Texts Appl. Math. 15, Springer-Verlag, New York, 2002.

[CWM04] C. Cardon-Weber and A. Millet, *On strongly Petrovskiĭ's parabolic SPDEs in arbitrary dimension and application to the stochastic Cahn–Hilliard equation*, J. Theoret. Probab., 17 (2004), pp. 1–49.

[EG04] Al. Ern and J. Guermond, *Theory and Practice of Finite Elements*, Appl. Math. Sci. 159, Springer-Verlag, New York, 2004.

[FSS01] H. Fujita, N. Saito, and T. Suzuki, *Operator Theory and Numerical Methods*, Stud. Math. Appl. 30, North–Holland, Amsterdam, 2001.

[Gar04] A. Gardoń, *The order of approximations for solutions of Itô-type stochastic differential equations with jumps*, Stochastic Anal. Appl., 22 (2004), pp. 679–699.

[Gyö98] I. Gyöngy, *Lattice approximations for stochastic quasi-linear parabolic partial differential equations driven by space-time white noise*. I, Potential Anal., 9 (1998), pp. 1–25.

[Gyö99] I. Gyöngy, *Lattice approximations for stochastic quasi-linear parabolic partial differential equations driven by space-time white noise*. II, Potential Anal., 11 (1999), pp. 1–37.

[GM05] I. Gyöngy and A. Millet, *On discretization schemes for stochastic evolution equations*, Potential Anal., 23 (2005), pp. 99–134.

[GP88] A. Germani and M. Piccioni, *Semidiscretization of stochastic partial differential equations on $\mathbf{R}^d$ by a finite-element technique*, Stochastics, 23 (1988), pp. 131–148.

[Hau02] E. Hausenblas, *Numerical analysis of semilinear stochastic evolution equations in Banach spaces,* J. Comput. Appl. Math., 147 (2002), pp. 485–516.

[Hau03] E. Hausenblas, *Approximation for semilinear stochastic evolution equations,* Potential Anal., 18 (2003), pp. 141–186.

[Hau05] E. Hausenblas, *Existence, uniqueness, and regularity of parabolic SPDEs driven by Poisson random measure*, Electron. J. Probab., 10 (2005), pp. 1496–1546.

[Hau06] E. Hausenblas, *SPDEs driven by Poisson random measure with non-Lipschitz coefficients*, Probab. Theory Related Fields, 137 (2007), pp. 161–200.

[HM06a] E. Hausenblas and I. Marchis, *The numerical approximation of parabolic stochastic partial differential equations driven by Poisson random measure*, BIT, 46 (2006), pp. 773–811.

[HM06b] E. Hausenblas and I. Marchis, *Numerical Simulations of Parabolic Stochastic Partial Differential Equations Driven by Poisson Random Measure*, manuscript, 2006.

[Hen81] D. Henry, *Geometric Theory of Semilinear Parabolic Equations*, Lecture Notes in Math. 840, Springer-Verlag, Berlin, 1981.

[KX95] G. Kallianpur and J. Xiong, *Diffusion approximation of nuclear space-valued stochastic-differential equations driven by Poisson random measures*, Ann. Appl. Probab., 5 (1995), pp. 493–517.

[KL02] M. Kouritzin and H. Long, *Convergence of Markov chain approximations to stochastic reaction-diffusion equations*, Ann. Appl. Probab., 12 (2002), pp. 1039–1070.

[KLS03] M. Kouritzin, H. Long, and W. Sun, *On Markov chain approximations to semilinear partial differential equations driven by Poisson measure noise*, Stochastic Anal. Appl., 21 (2003), pp. 419–441.

[KS01] P. E. Kloeden and S. Shott, *Linear-implicit strong schemes for Itô-Galerkin approximations of stochastic PDEs*, J. Appl. Math. Stoch. Anal., 14 (2001), pp. 47–53.

[Kuz01] D. F. Kuznetsov, *New representations of explicit one-step numerical methods for stochastic differential equations with a jump component*, Zh. Vychisl. Mat. Mat. Fiz., 41 (2001), pp. 922–937.

[KX87] G. Kallianpur and J. Xiong, *A nuclear space-valued stochastic differential equation driven by Poisson random measures*, in Stochastic Partial Differential Equations and Their Applications, Lecture Notes in Control and Inform. Sci. 176, Springer-Verlag, New York, 1987, pp. 135–143.

[KX96] G. Kallianpur and J. Xiong, *Stochastic Differential Equations in Infinite Dimensional Spaces*, Lecture Notes Monograph Series 26, Institute of Mathematical Statistics, Hayward, CA, 1996.

[Lin86] W. Linde, *Probability in Banach Spaces—Stable and Infinitely Divisible Distributions*, 2nd ed., Wiley-Interscience, New York, 1986.

[MM05] A. Millet and P. Morien, *On implicit and explicit discretization schemes for parabolic SPDEs in any dimension*, Stochastic Process. Appl., 115 (2005), pp. 1073–1106.

[Paz83] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Appl. Math. Sci. 44, Springer-Verlag, New York, 1983.

[PZ07] S. Peszat and J. Zabczyk, *Stochastic Partial Differential Equations with Lévy Noise*, Encyclopedia Math. Appl. 113, Cambridge University Press, Cambridge, 2007.

[Pri01a] J. Printems, *On the discretization in time of parabolic stochastic partial differential equations*, Monte Carlo Methods Appl., 7 (2001), pp. 359–368.

[Pri01b] J. Printems, *On the discretization in time of parabolic stochastic partial differential equations*, M2AN Math. Model. Numer. Anal., 35 (2001), pp. 1055–1078.

[Pro04] P. Protter, *Stochastic Integration and Differential Equations*, 2nd ed., Appl. Math. 21, Springer-Verlag, New York, 2004.

[PT97] P. Protter and D. Talay, *The Euler scheme for Lévy driven stochastic differential equations*, Ann. Probab., 25 (1997), pp. 393–423.

[Rub03] S. Rubenthaler, *Numerical simulation of the solution of a stochastic differential equation driven by a Lévy process*, Stochastic Process. Appl., 103 (2003), pp. 311–349.

[RS96] T. Runst and W. Sickel, *Sobolev Spaces of Fractional Order, Nemytskij Operators and Nonlinear Partial Differential Equations*, de Gruyter Ser. Nonlinear Anal. Appl. 3, de Gruyter, Berlin, 1996.

[SLB98] E. St. L. Bié, *Étude d'une EDPS conduite par un bruit Poissonnien. (Study of a stochastic partial differential equation driven by a Poisson noise)*, Probab. Theory Related Fields, 111 (1998), pp. 287–321.

[Sha99] T. Shardlow, *Numerical methods for stochastic parabolic PDEs*, Numer. Funct. Anal. Optim., 20 (1999), pp. 121–145.

[Wal86] J. Walsh, *An introduction to stochastic partial differential equations*, in Ecole d'Eté de Probabilités de Saint-Flour XIV - 1984, Lecture Notes in Math. 1180, Springer-Verlag, New York, 1986, pp. 265–437.

[Wal05] J. B. Walsh, *Finite element methods for parabolic stochastic PDE's*, Potential Anal., 23 (2005), pp. 1–43.

[Woy01] W. A. Woyczyński, *Lévy Processes in the Physical Sciences*, Lévy Processes, Birkhäuser Boston, Boston, MA, 2001, pp. 241–266.

[Yan05] Y. Yan, *Galerkin finite element methods for stochastic parabolic partial differential equations*, SIAM J. Numer. Anal., 43 (2005), pp. 1363–1384.

[Yoo00] H. Yoo, *Semi-discretization of stochastic partial differential equations on r^1 by a finite-difference method*, Math. Comp., 69 (2000), pp. 653–666.

SIAM J. NUMER. ANAL.
Vol. 46, No. 1, pp. 472–495
© 2008 Society for Industrial and Applied Mathematics

IMMERSED-INTERFACE FINITE-ELEMENT METHODS FOR ELLIPTIC INTERFACE PROBLEMS WITH NONHOMOGENEOUS JUMP CONDITIONS*

YAN GONG†, BO LI‡, AND ZHILIN LI§

Abstract. In this work, a class of new finite-element methods, called *immersed-interface finite-element methods,* is developed to solve elliptic interface problems with nonhomogeneous jump conditions. Simple non–body-fitted meshes are used. A single function that satisfies the same nonhomogeneous jump conditions is constructed using a level-set representation of the interface. With such a function, the discontinuities across the interface in the solution and flux are removed, and an equivalent elliptic interface problem with homogeneous jump conditions is formulated. Special finite-element basis functions are constructed for nodal points near the interface to satisfy the homogeneous jump conditions. Error analysis and numerical tests are presented to demonstrate that such methods have an optimal convergence rate. These methods are designed as an efficient component of the finite-element level-set methodology for fast simulation of interface dynamics that does not require remeshing.

Key words. elliptic interface problems, nonhomogeneous jump conditions, immersed-interface finite-element method, level-set functions, error estimates

AMS subject classification. 65N30

DOI. 10.1137/060666482

1. Introduction.

We consider numerical solution of the elliptic interface problem

$$-\nabla\cdot\beta^-\nabla u = f \qquad \text{in } \Omega^-, \tag{1.1}$$

$$-\nabla\cdot\beta^+\nabla u = f \qquad \text{in } \Omega^+, \tag{1.2}$$

$$[u]_\Gamma = w, \tag{1.3}$$

$$\left[\beta\frac{\partial u}{\partial n}\right]_\Gamma = Q, \tag{1.4}$$

$$u = g \qquad \text{on } \partial\Omega. \tag{1.5}$$

Here, $\Omega \subset \mathbb{R}^2$ is a bounded domain with its boundary $\partial\Omega$. Both Ω^- and Ω^+ are subdomains of Ω such that $\Omega^- \cap \Omega^+ = \emptyset$ and $\overline{\Omega^-} \cup \overline{\Omega^+} = \overline{\Omega}$, where an overline denotes the

*Received by the editors August 1, 2006; accepted for publication (in revised form) September 11, 2007; published electronically January 30, 2008.
http://www.siam.org/journals/sinum/46-1/66648.html

†Department of Mathematics, North Carolina State University, Raleigh, NC 27695-8205 (ygong@ncsu.edu).

‡Department of Mathematics, University of California, San Diego, 9500 Gilman Drive, Mail code: 0112, La Jolla, CA 92093-0112 (bli@math.ucsd.edu). This author's work was partially supported by the National Science Foundation through grant DMS-0451466 and the Department of Energy through grant DE-FG02-05ER25707.

§Center for Research in Scientific Computation and Department of Mathematics, North Carolina State University, Raleigh, NC 27695-8205 (zhilin@math.ncsu.edu). This author's work was partially supported by the National Science Foundation/National Institutes of Health through grant 0201094, ARO grant 49308-MA, AFSOR grant FA9550-06-1-024, National Science Foundation grant DMS-0412654, and the State Key Laboratory of Scientific and Engineering Computing of Chinese Academy of Sciences (summer, 2006).

closure; see Figure 1.1 for an illustration. For simplicity, we assume that $\overline{\Omega^-} \cap \partial\Omega = \emptyset$. We denote $\Gamma = \overline{\Omega^-} \cap \overline{\Omega^+}$ and call it the *interface* separating Ω^- and Ω^+. We shall assume that Γ is sufficiently smooth. We also denote by n the unit vector normal to Γ pointing from Ω^- to Ω^+, or the unit exterior normal to $\partial\Omega$, and by $\partial/\partial n$ or ∂_n the corresponding normal derivative.

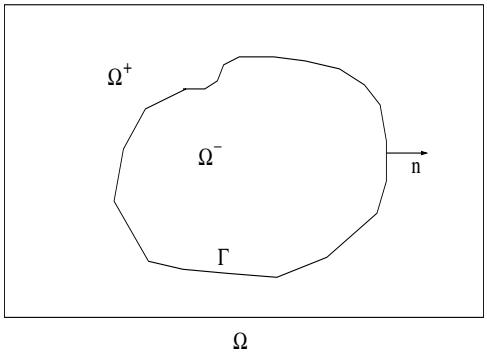


FIG. 1.1. *The geometry of an elliptic interface problem.*

All the functions β, $f : \Omega \to \mathbb{R}$, w, $Q : \Gamma \to \mathbb{R}$, and $g : \partial\Omega \to \mathbb{R}$ are given. As usual, for any function $\xi : \Omega \to \mathbb{R}$, we denote the restrictions of ξ on Ω^- and Ω^+ by

$$\xi^- = \xi|_{\Omega^-} \qquad \text{and} \qquad \xi^+ = \xi|_{\Omega^+},$$

respectively, and we denote by

$$[\xi]_\Gamma(x) = \lim_{y \to x, y \in \Omega^+} \xi^+(y) - \lim_{y \to x, y \in \Omega^-} \xi^-(y)$$

the jump of ξ across the interface Γ at $x \in \Gamma$, when the unique limiting values on Γ of ξ from both sides of Γ exist. We assume that β^- and β^+ are smooth and bounded on Ω^- and Ω^+, respectively,

$$\beta(x) \geq \beta_0 \qquad \forall x \in \Omega, \tag{1.6}$$

for some constant $\beta_0 > 0$, $f \in L^2(\Omega)$, all w, Q, and g are smooth and bounded.

Equations (1.3) and (1.4) are nonhomogeneous interface or jump conditions. Elliptic interface problems (1.1)–(1.5) with such jump conditions arise in many areas. For example, in a Burton–Cabrera–Frank-type model for epitaxial growth of thin films, the adatom (adsorbed atom) density that solves the diffusion equation on terraces, and the corresponding flux, can have jumps across interfaces that represent atomic steps [3, 5, 6]. Another example is that the reaction potential of electrostatics of a solvation energy satisfies a nonhomogeneous jump condition for the flux [13].

If the jump conditions are homogeneous, i.e., $w = 0$ and $Q = 0$ on Γ (cf. (1.3) and (1.4)), then the problem (1.1)–(1.5) is equivalent to that of finding $u \in H^1(\Omega)$ with $u = g$ on $\partial\Omega$ such that

$$\int_\Omega \beta \nabla u \cdot \nabla v \, dx = \int_\Omega f v \, dx \qquad \forall v \in H_0^1(\Omega).$$

In general, we can extend the nonhomogeneous jump $w : \Gamma \to \mathbb{R}$ to a piecewise smooth function $\hat{w} : \Omega \to \mathbb{R}$ such that

$$[\hat{w}]_\Gamma = w \qquad \text{and} \qquad \hat{w} = g \quad \text{on } \partial\Omega.$$

Then, the problem (1.1)–(1.5) is equivalent to the problem for $u = q + \hat{w}$ with $q \in H_0^1(\Omega)$ uniquely determined by

$$\int_\Omega \beta \nabla q \cdot \nabla v \, dx = \int_\Omega f v \, dx - \int_\Gamma Q v \, ds - \int_\Omega \beta \nabla \hat{w} \cdot \nabla v \, dx \qquad \forall v \in H_0^1(\Omega). \tag{1.7}$$

In a standard body-fitted finite-element method for solving the problem in the weak formulation (1.7), the interface is covered by a low dimensional finite-element mesh, and the extension of $w : \Gamma \to \mathbb{R}$ to $\hat{w} : \Omega \to \mathbb{R}$ can be made by finite-element approximations.

In this work, we develop a class of finite-element methods based on non–body-fitted finite-element meshes to solve the elliptic interface problem (1.1)–(1.5). Such a method has the following distinguished features.

(a) A fixed finite-element mesh that allows the interface to cut through edges of elements is used.
(b) Both the jumps $w : \Gamma \to \mathbb{R}$ and $Q : \Gamma \to \mathbb{R}$ are extended locally by a single function using a level-set representation of the interface Γ. The extensions are necessary to transfer the original problem to a new one with homogeneous jump conditions so that the immersed finite-element methods developed in [18] can be applied. The new methods make it possible to maintain second order accuracy on the elements near the interface; see sections 2 and 5.
(c) Special finite-element basis functions for nodal points near the interface are constructed to satisfy the homogeneous jump conditions; see section 4.
(d) The resulting linear system of discretization is symmetric positive definite.
(e) Optimal convergence rates, the same as those for a body-fitted finite-element method, are achieved; see section 3 for basic error estimates and section 6 for numerical results on such convergence properties.

Our methods incorporate the finite-element discretization into the framework of an immersed-interface method for interface problems, and we therefore call them *immersed-interface finite-element methods.* The basic idea of an immersed-interface method is to incorporate the jump conditions in constructing basis functions. In a finite-difference immersed-interface method, the jump conditions are enforced through finite-difference equations on grid points near the interface. In our immersed-interface finite-element methods, the jump conditions are enforced through the construction of special finite-element basis functions that satisfy the homogeneous interface conditions. Clearly, such basis functions depend on the interface location and the jump w. Some of the related work can be found in [7, 12].

The development of our methods is strongly motivated by the need for a fast and accurate solver of elliptic interface problems that is required in each time step in a long-time level-set simulation of interface dynamics without remeshing. Such simulation has been a powerful numerical approach in understanding material properties, biological processes, and many other important phenomena in science and engineering [20, 21, 22].

The finite-element spaces constructed in this paper are conforming ones that contain piecewise P_1 polynomials. The interpolation error is therefore of second order. If the solution to the elliptic interface problem is piecewise smooth, or, more precisely, $u^\pm \in C^2(\overline{\Omega^\pm})$, which is true for many applications, then the standard convergence results are true for our finite-element methods. This means that our finite-element solution is first order accurate in the H^1 norm and second order accurate in the L^2 norm. In this work, we construct only some two-dimensional linear and quadratic

triangular immersed-interface finite elements. However, our framework described here can be used for high order immersed-interface finite elements and for more general elliptic interface problems in two and three space dimensions.

We use the zero level set of a Lipschitz continuous function particularly, a good approximation of the signed distance function, to represent the interface Γ. For such a level-set function, the interface can cut the interface only once between two grid lines. The resolution of the interface Γ is determined by the level-set function. We note that there are plenty of discussions in the literature about how to construct the level-set function for an interface Γ; see, e.g., [20, 22]. If the interface is complicated in reference to a mesh, then it may not be resolved well by the level-set function. Consequently, the finite-element solution obtained from our method may not resolve all the fine details of the solution. A finer mesh is then needed to resolve the geometry and the solution.

Our work combines the development of a new formulation of underlying problems, construction of new elements, basic error estimates, and numerical experiments. The analysis in our work focuses on two parts: one is the algorithm of removing source singularities, and the other is the interpolation error estimate for some of the immersed-interface finite-element approximations. We have not tried to deal in this paper with the analysis of other errors, such as those of the approximation of the interface Γ by Γ_h, numerical quadrature, round-off errors, etc.

In section 2, we present a new weak formulation of the elliptic interface problem (1.1)–(1.5) with a special treatment of nonhomogeneous jump conditions. In section 3, we describe the framework of our immersed-interface finite-element methods and provide basic error estimates. In section 4, we construct several linear or quadratic immersed-interface finite-element spaces and obtain their corresponding interpolation errors. In section 5, we give some details of implementation of our methods. In section 6, we present numerical results. Finally, in section 7, we draw conclusions.

2. Weak formulation. We first review the level-set representation of the interface Γ which is assumed to be smooth. Let $\varphi : \Omega \to \mathbb{R}$ be a continuous function that satisfies

$$\varphi(x) \begin{cases} < 0 & \text{if } x \in \Omega^-, \\ = 0 & \text{if } x \in \Gamma, \\ > 0 & \text{if } x \in \Omega^+. \end{cases} \tag{2.1}$$

We call such a function $\varphi : \Omega \to \mathbb{R}$ a *level-set function* that represents the interface Γ. For $\rho > 0$, we denote the ρ-neighborhood of Γ in Ω by

$$N(\Gamma, \rho) = \{x \in \Omega : \text{dist}\,(x, \Gamma) < \rho\},$$

where $\text{dist}\,(x, \Gamma)$ is the distance from x to Γ. We assume that there exists $\rho_0 > 0$ such that $\overline{N(\Gamma, \rho_0)} \subset \Omega$, and

$$\varphi \text{ is smooth and } |\nabla\varphi| > 0 \text{ in } N(\Gamma, \rho_0). \tag{2.2}$$

We note that the unit normal n to Γ, pointing from Ω^- to Ω^+, is given by

$$n = \frac{\nabla\varphi}{|\nabla\varphi|}. \tag{2.3}$$

The signed distance function

$$\varphi(x) = \begin{cases} -\operatorname{dist}(x,\Gamma) & \text{if } x \in \Omega^-, \\ 0 & \text{if } x \in \Gamma, \\ +\operatorname{dist}(x,\Gamma) & \text{if } x \in \Omega^+ \end{cases}$$

is a typical level-set function that satisfies our assumptions.

We now turn to the description of our treatment of nonhomogeneous jump conditions. We first need the following lemma.

LEMMA 2.1. *Let $\rho > 0$ be small enough. Then, for any $x \in N(\Gamma,\rho)$, there exists a unique $\hat{x} \in \Gamma$ such that*

$$|x - \hat{x}| = dist(x,\Gamma). \tag{2.4}$$

Moreover,

$$\frac{x-\hat{x}}{|x-\hat{x}|} = \begin{cases} -n(\hat{x}) & \text{if } x \in \Omega^-, \\ +n(\hat{x}) & \text{if } x \in \Omega^+, \end{cases} \tag{2.5}$$

where $n(\hat{x})$ is the unit normal to Γ at $\hat{x}$, pointing from Ω^- to Ω^+.

Proof. Without loss of generality, let us assume that, in a local Cartesian coordinate system, $x = (0,0)$ is the origin, and the interface nearby is a graph of a smooth function $\eta = \eta(s)$ that is nonzero for any s in a certain range. The distance from the origin to any point $(s,\eta(s))$ on the interface is then given by $\sqrt{I(s)} < \rho$ with

$$I(s) = s^2 + [\eta(s)]^2.$$

Setting $I'(s) = 0$, we get $\eta'(s) = -s/\eta(s)$. This implies (2.5), since $(-1,\eta'(s))$ is parallel to the normal at $(s,\eta(s))$. Moreover,

$$I''(s) = 2 + 2[\eta'(s)]^2 + 2\eta(s)\eta''(s)$$

is positive, since $I(s)$, and hence $\eta(s)$, is small enough, whenever $\rho > 0$ is small enough. Thus, there exists a unique s that minimizes $I(s)$. This implies the existence and uniqueness of $\hat{x}$ that satisfies (2.4). ∎

In what follows, we fix $\rho_0 > 0$ as in (2.2). We let $\rho \in \mathbb{R}$ be given as in the above lemma and assume that $0 < \rho < \rho_0$. We define $w_\rho : \Omega \to \mathbb{R}$ and $Q_\rho : \Omega \to \mathbb{R}$ to be the extensions of $w : \Gamma \to \mathbb{R}$ and $Q : \Gamma \to \mathbb{R}$, respectively, that satisfy the following.

E1. Both w_ρ and Q_ρ are smooth on $\overline{\Omega}$.

E2. $w_\rho(x) = w(\hat{x})$ and $Q_\rho(x) = Q(\hat{x})$ for any $x \in N(\Gamma,\rho)$, where $\hat{x}$ is defined as in Lemma 2.1.

E3. $w_\rho(x) = g(x)$ for any $x \in \partial\Omega$, and $Q_\rho(x) = 0$ for any $x \in \Omega \setminus N(\Gamma,\rho)$.

To see the existence of w_ρ, we first apply Lemma 2.1 to define a function, say, w_1, by $w_1(x) = w(\hat{x})$ for any $x \in N(\Gamma,2\rho)$ for $\rho > 0$ small enough, and $w_1 = 0$ elsewhere in Ω. Using the lifting operator [2, 11], there exists a smooth function, say, w_2, on Ω such that the restriction of w_2 on $\partial\Omega$ is g. (The smoothness depends on that of $\partial\Omega$.) Now, assume that $N(\Gamma,2\rho)$ and $A := N(\partial\Omega,\varepsilon) = \{x \in \Omega : \operatorname{dist}(x,\partial\Omega) < \varepsilon\}$ are disjoint for some small $\varepsilon > 0$. Then, by applying mollifiers to $w_1 + \chi_A w_2$, we obtain the desired w_ρ. The existence of Q_ρ is seen similarly.

We define $u_\rho : \Omega \to \mathbb{R}$ by

$$u_\rho(x) = \chi_{\Omega^+}(x)\left(w_\rho(x) + \frac{Q_\rho(x)}{\beta^+(x)}\frac{\varphi(x)}{|\nabla\varphi(x)|}\right) \qquad \forall x \in \Omega, \tag{2.6}$$

where χ_{Ω^+} is the characteristic function of Ω^+. Note by E3 that $Q_\rho = 0$ outside $N(\Gamma, \rho_0)$ in which φ may not be smooth. Thus, implicitly, the second term in u_ρ is defined to be 0 outside $N(\Gamma, \rho_0)$.

The following statement summarizes some useful properties of $u_\rho : \Omega \to \mathbb{R}$.

LEMMA 2.2. *Both u_ρ^- and u_ρ^+ are smooth on Ω^- and Ω^+, respectively. Moreover, $u_\rho = g$ on $\partial\Omega$, and*

$$[u_\rho]_\Gamma = w, \qquad \left[\beta\frac{\partial u_\rho}{\partial n}\right]_\Gamma = Q.$$

Proof. Since β^+ is smooth and bounded on Ω^+, it follows from (1.6), E1, and (2.2) that u_ρ^+ is smooth on Ω^+. Obviously, $u_\rho^- = 0$ is smooth on Ω^-. By E3, $u_\rho = g$ on $\partial\Omega$. By (2.1) and E2,

$$[u_\rho]_\Gamma = u_\rho^+ - u_\rho^- = w.$$

Now, fix $x \in \Gamma$. It follows from Lemma 2.1 and E2 that $\partial_n w_\rho = 0$. Moreover, $\varphi(x) = 0$ by (2.1). Therefore, for any $x \in \Gamma$, we obtain by E2, (2.6), (2.3), and the fact that $\varphi(x) = 0$ that

$$\begin{aligned}
\left[\beta(x)\frac{\partial u_\rho(x)}{\partial n}\right]_\Gamma &= \beta^+(x)\frac{\partial u_\rho^+(x)}{\partial n} \\
&= \beta^+(x)\frac{\partial w_\rho(x)}{\partial n} + \beta^+(x)\frac{\partial}{\partial n}\left(\frac{Q_\rho(x)}{\beta^+(x)|\nabla\varphi(x)|}\right)\varphi(x) + \frac{Q_\rho(x)}{|\nabla\varphi(x)|}\frac{\partial\varphi(x)}{\partial n} \\
&= Q(x)\frac{n(x)\cdot\nabla\varphi(x)}{|\nabla\varphi|(x)} \\
&= Q(x).
\end{aligned}$$

The proof is completed. ∎

THEOREM 2.3. *There exists a unique $q \in H_0^1(\Omega)$ such that*

$$\int_\Omega \beta\nabla q\cdot\nabla v\,dx = \int_\Omega fv\,dx - \int_\Gamma Qv\,ds - \int_{\Omega^+}\beta^+\nabla u_\rho\cdot\nabla v\,dx \qquad \forall v \in H_0^1(\Omega). \tag{2.7}$$

This is equivalent to

$$\int_\Omega \beta\nabla q\cdot\nabla v\,dx = \int_\Omega fv\,dx + \int_{\Omega^+}(\nabla\cdot\beta^+\nabla u_\rho)\,v\,dx \qquad \forall v \in H_0^1(\Omega). \tag{2.8}$$

Moreover, $u = q + u_\rho$ solves the problem (1.1)–(1.5). In particular, $q \in H_0^1(\Omega)$ satisfies the homogeneous jump conditions

$$[q]_\Gamma = 0 \qquad \text{and} \qquad \left[\beta\frac{\partial q}{\partial n}\right]_\Gamma = 0. \tag{2.9}$$

Proof. Note that the line integral in (2.7) defines a continuous, linear functional on $H_0^1(\Omega)$. The existence and uniqueness of $q \in H_0^1(\Omega)$ satisfying (2.7) follow therefore

from the Lax–Milgram theorem. By integration by parts, (2.6), and Lemma 2.2, we imply that (2.7) and (2.8) are equivalent.

Choosing $v \in H_0^1(\Omega)$ with $\operatorname{supp} v \subset \Omega^-$ and $\operatorname{supp} v \subset \Omega^+$, respectively, applying the regularity theory of elliptic problems, we obtain from (2.7) that $u = q + u_\rho$ satisfies (1.1) and (1.2). The jump condition (1.3) follows from Lemma 2.2. Now, integrating by parts, we obtain by (1.1), (1.2), (2.6), and (2.7) that

$$\int_\Gamma \left(\left[\beta \frac{\partial u}{\partial n} \right]_\Gamma - Q \right) v \, ds = 0 \qquad \forall v \in H_0^1(\Omega).$$

This leads to (1.4). Clearly, the function $u = q + u_\rho$ satisfies (1.5).

Finally, since $q = u - u_\rho$, we have

$$[q]_\Gamma = [u]_\Gamma - [u_\rho]_\Gamma = w - w = 0$$

from (1.3) and Lemma 2.2. We also have

$$\left[\beta \frac{\partial q}{\partial n} \right]_\Gamma = \left[\beta \frac{\partial u}{\partial n} \right]_\Gamma - \left[\beta \frac{\partial u_\rho}{\partial n} \right]_\Gamma = Q - Q = 0$$

from (1.4) and Lemma 2.2. ∎

3. Immersed-interface finite-element approximations. In this section, we first state practically reasonable assumptions on how the interface can cut edges of elements in a finite-element mesh. We then define immersed-interface finite-element approximations of our underlying problem formulated in Theorem 2.3. Finally, we give rigorous error estimates for our method.

Let $\mathcal{T}_h$ be a finite-element mesh with mesh size h that covers $\overline{\Omega}$. We assume that the elements in $\mathcal{T}_h$ are all triangles. For simplicity, we shall assume that Ω is a convex polygonal domain and the mesh covers $\overline{\Omega}$ exactly. Standard finite-element techniques can be applied to treat a curved boundary without affecting our approximation properties. We remark that in practice the computational domain Ω can often be chosen as a rectangular domain with sides parallel to the coordinate axes; and the finite-element mesh can be uniform.

We call an element $T \in \mathcal{T}_h$ an *interface element* if $\Gamma \cap \operatorname{int} T \neq \emptyset$. Note that an element is a noninterface element if one of its edges is part of the interface. We assume, for any interface element $T \in \mathcal{T}_h$, that the set $\Gamma \cap \partial T$ consists of exactly two points that are on different edges of T.

We define an immersed-interface finite-element space V_h with respect to the mesh $\mathcal{T}_h$ to be a finitely dimensional subspace of $L^2(\Omega)$ that consists of all the linear combinations of the corresponding basis functions $\phi_1, \ldots, \phi_N$ for some integer $N \geq 1$:

$$V_h = \operatorname{span}\{\phi_1, \ldots, \phi_N\}. \tag{3.1}$$

The basis functions are the usual finite-element basis functions on a noninterface element and are piecewise polynomials on each interface element that is determined by the element and the interface Γ. All the basis functions satisfy the homogeneous jump conditions for both the function and flux. Moreover, there exists an interpolation operator from some functional space to V_h that enjoys the usual approximation properties.

To be more precise, we define a *conforming immersed-interface finite-element* space V_h for the approximation of a second order elliptic interface problem to be a subspace of $H^1(\Omega)$ that satisfies the following properties.

FE1. All the functions in V_h restricted onto the union of all noninterface elements form a usual finite-element space of piecewise polynomials of degree $\le k$ for some integer $k \ge 1$, such as a P_k-type Lagrange finite-element space when $\mathcal{T}_h$ is a triangular mesh, or a Q_k- or Q'_k-type Lagrange finite-element space when $\mathcal{T}_h$ is a quadrilateral mesh [4, 8]. On each interface element $T \in \mathcal{T}_h$, all functions in V_h are piecewise polynomials. The support of each basis function is a union of a few elements.

FE2. All the basis functions, and hence all the functions in V_h, satisfy the homogeneous jump conditions

$$[v_h]_{\Gamma_h} = 0 \quad \text{and} \quad \left[\beta(\hat{x})\frac{\partial v_h}{\partial n}\right]_{\Gamma_h} = 0 \qquad \forall v_h \in V_h, \tag{3.2}$$

where Γ_h denotes the union of line segments approximating the interface Γ, and the normal derivatives of $v_h \in V_h$ from each side of the interface are assumed to exist; $\hat{x}$ is the orthogonal projection on the interface of the midpoint of Γ_h within the triangle.[1] Note that $\left[\beta(\hat{x})\frac{\partial v_h}{\partial n}\right]_{\Gamma_h} = \beta^+(\hat{x})\frac{\partial v_h^+}{\partial n} - \beta^-(\hat{x})\frac{\partial v_h^-}{\partial n}$.

FE3. There exists an interpolation operator $I_h : D(\Omega) \to V_h$ such that

$$\|I_h v - v\|_{H^1(\Omega)} \le C h^{\sigma(k)} \|v\|_{H^{k+1}(\Omega\setminus\Gamma)} \qquad \forall v \in D(\Omega), \tag{3.3}$$

where $D(\Omega) \subset H^1(\Omega)$ is an infinite-dimensional space of functions that are smooth and bounded on both Ω^- and Ω^+, respectively, $\bar{k}$ is a constant such that $0 < \sigma(k) \le k$, and $C > 0$ denotes a generic constant independent of h and v.

The homogeneous jump condition for $v_h \in V_h$ in (3.2) is a consequence of the fact that $V_h \subset H^1(\Omega)$. Often the interpolation error in FE3 can be the same as usual, i.e., $\bar{k} = k$. But, it can be slightly worse due to the approximation of Γ by Γ_h.

We can similarly define a *nonconforming immersed-interface finite-element* space $V_h \subset L^\infty(\Omega)$ with $V_h \not\subset H^1(\Omega)$. For a nonconforming immersed-interface finite-element space V_h that is used to solve our underlying problem with the boundary condition (1.5), we also need to assume that a discrete Poincaré inequality,

$$\|v_h\|^2_{L^2(\Omega)} \le C \sum_{T\in T_h} \|\nabla v_h\|^2_{L^2(T\setminus\Gamma_h)} \qquad \forall v_h \in V_h,$$

is satisfied. Several examples of conforming and nonconforming, linear or quadratic, immersed-interface finite elements will be presented in the next section.

We consider now immersed-interface finite-element approximations of the solution $q \in H^1_0(\Omega)$ defined by (2.7). Let $u_\rho : \Omega \to \mathbb{R}$ be as given in (2.6) for some $\rho > 0$ small enough. Now let $V_h \subset H^1_0(\Omega)$ be a conforming immersed-interface finite-element space of piecewise polynomials of degree $\le k$ as in FE1–FE3 that accounts for the boundary condition (1.5). To be precise, we assume in our analysis below that the domain of the interpolation operator I_h is

$$D(\Omega) = \left\{ v \in H^1_0(\Omega) : v^- \in H^{k+1}(\Omega^-) \cap C(\overline{\Omega^-}) \text{ and } v^+ \in H^{k+1}(\Omega^+) \cap C(\overline{\Omega^+}) \right\}.$$

[1] Alternatively, we can approximate $\beta^\pm(\hat{x})$ using $\beta^\pm(\hat{X})$, where $\hat{X} \in \Gamma \cap \mathcal{T}_h$ is any point on the interface and in the triangle.

THEOREM 3.1. *There exists a unique $q_h \in V_h$ such that*

$$\int_\Omega \beta \nabla q_h \cdot \nabla v_h \, dx = \int_\Omega f v_h \, dx + \int_{\Omega^+} (\nabla \cdot \beta^+ \nabla u_\rho) \, v_h \, dx \qquad \forall v \in V_h. \tag{3.4}$$

Moreover, if the solution $q \in H_0^1(\Omega)$ of (2.7) satisfies that $q^- \in H^{k+1}(\Omega^-) \cap C(\overline{\Omega^-})$ and $q^+ \in H^{k+1}(\Omega^+) \cap C(\overline{\Omega^+})$, then

$$\|q - q_h\|_{H^1(\Omega)} \le C h^{\sigma(k)} \|q\|_{H^{k+1}(\Omega \backslash \Gamma)}, \tag{3.5}$$

$$\|q - q_h\|_{L^2(\Omega)} \le C h^{\sigma(k)+\sigma(1)} \|q\|_{H^{k+1}(\Omega \backslash \Gamma)}, \tag{3.6}$$

where $\sigma(k)$ is defined in FE3 *and $C > 0$ is a constant independent of h and q.*

Proof. Since $V_h \subset H_0^1(\Omega)$, the existence and uniqueness of $q_h \in V_h$ satisfying (3.4) follow from the Lax–Milgram theorem. Moreover, by (2.8), which is equivalent to (2.7) by Theorem 2.3, and (3.4), we obtain

$$\int_\Omega \beta \nabla (q - q_h) \cdot \nabla v_h \, dx = 0 \qquad \forall v_h \in V_h. \tag{3.7}$$

The error estimate (3.5) can be obtained by a standard argument using a Poincaré inequality, the ellipticity indicated by (1.6), the error equation (3.7), and the interpolation error estimate (3.3); see [4, 8].

To obtain the L^2 error estimate (3.6), we use the standard dual argument (cf., e.g., [4, 8]) with slight modification taking care of the lack of global regularity of a solution. Let $r \in H_0^1(\Omega)$ be the unique function satisfying

$$\int_\Omega \beta \nabla r \cdot \nabla v \, dx = \int_\Omega (q - q_h) v \, dx \qquad \forall v \in H_0^1(\Omega). \tag{3.8}$$

Since β^- and β^+ are smooth and bounded in Ω^- and Ω^+, respectively, we see that $r^- \in H^2(\Omega)$ and $r^+ \in H^2(\Omega)$. Moreover, it follows from (3.8) that $[\beta \partial_n r]_\Gamma = 0$. Therefore, we have (cf. Theorem 2.1 in [7]) that

$$\|r^\pm\|_{H^2(\Omega^\pm)} \le C \|q - q_h\|_{L^2(\Omega)}. \tag{3.9}$$

Setting $v = q - q_h \in H_0^1(\Omega)$ in (3.8), we have by (3.8), (3.7) with $v_n = I_h r$, the Cauchy–Schwarz inequality, (3.3) with $k = 1$, (3.5), and (3.9) that

$$\begin{aligned}
\|q - q_h\|_{L^2(\Omega)}^2 &= \int_\Omega \beta \nabla r \cdot \nabla (q - q_h) \, dx \\
&= \int_\Omega \beta \nabla (r - I_h r) \cdot \nabla (q - q_h) \, dx \\
&\le C \|r - I_h r\|_{H^1(\Omega)} \|q - q_h\|_{H^1(\Omega)} \\
&\le C h^{\sigma(k)+\sigma(1)} \|r\|_{H^2(\Omega \backslash \Gamma)} \|q\|_{H^{k+1}(\Omega \backslash \Gamma)} \\
&\le C h^{\sigma(k)+\sigma(1)} \|q\|_{H^{k+1}(\Omega \backslash \Gamma)} \|q - q_h\|_{L^2(\Omega)},
\end{aligned}$$

leading to (3.6). ∎

We remark that the removal of both jumps w and Q using a single function leads to finite-element equation (3.4) which does not involve line integrals as in (2.7). This much simplifies numerical implementation.

4. Construction of basis functions. We now describe how to construct some new classes of triangular immersed-interface finite-element basis functions that satisfy the homogeneous jump conditions.

Let $z_1, \ldots, z_N$ be all the vertices of triangular elements in the mesh $\mathcal{T}_h$. We define our linear or quadratic linear immersed-interface finite-element basis functions ϕ_i associated with z_i $(i = 1, \ldots, N)$ to be functions on Ω that satisfy the following properties.

B1. Each ϕ_i $(1 \le i \le N)$ is linear on a noninterface element and is piecewise linear or quadratic on an interface element.

B2. $\phi_i(z_j) = \delta_{ij}$ for $i, j = 1, \ldots, N$, where $\delta_{ij} = 1$ if $i = j$ and 0 if $i \ne j$.

B3. All ϕ_i $(1 \le i \le N)$ satisfy the approximate homogeneous jump conditions

$$[\phi_i]_{\Gamma_h} = 0, \qquad \left[\beta(\hat{x})\frac{\partial \phi_i}{\partial n}\right]_{\Gamma_h} = 0.$$

The corresponding finite-element space V_h is given by (3.1).

We call a vertex an *irregular vertex* or *irregular node* if it is a vertex of an interface element and a *regular vertex* or *regular node* otherwise. For a regular node z_i, we define ϕ_i to be the usual conforming linear finite-element basis function associated with z_i, i.e., $\phi_i \in C(\overline{\Omega}) \cap H_0^1(\Omega)$, ϕ_i is a linear polynomial on each element $T \in \mathcal{T}_h$, and $\phi_i(z_j) = \delta_{ij}$ for $j = 1, \ldots, N$.

For an irregular node z_i, we construct the corresponding basis function by modifying the usual linear finite-element basis function to satisfy the properties B1–B3. We need only to construct a local basis (or shape) function on an interface element for an irregular node.

4.1. A nonconforming linear element. This nonconforming immersed-interface finite element was first constructed in [18]; see also [12]. We briefly review it, since it will be used to construct our new linear and quadratic elements.

Fix an interface element $T \in \mathcal{T}_h$; cf. Figure 4.1. As in the common practice, we approximate the interface in T, $\Gamma \cap T$, by a line segment connecting the intersections of the interface and the edges of the triangle T. This line segment is $\overline{DE}$ in Figure 4.1. The line segment divides T into two parts T^+ and T^-, one triangular and the other quadrilateral. Note that there is a small region T_r in T,

$$T_r = T - \Omega^+ \cap T^+ - \Omega^- \cap T^-,$$

whose area is of order $O(h^3)$.

We now consider the element T in Figure 4.1 as a reference interface element T and define local basis functions for each vertex of T. The local basis function for a general interface element in the mesh $\mathcal{T}_h$ can be defined through the usual affine transformation. We denote

$$T = \{(x_1, x_2) : 0 \le x_1 \le h, \quad 0 \le x_2 \le hx_1 + x_2 \le h\}$$

and assume that the coordinates at A, B, C, D, and E are

$$(0, h), \quad (0, 0), \quad (h, 0), \quad (0, y_1), \quad (h - y_2, y_2),$$

respectively; cf. Figure 4.1.

Each of the three local basis functions corresponding to the nodes A, B, or C takes value 1 at one node and 0 at the other two. Once the values at nodes A, B, and

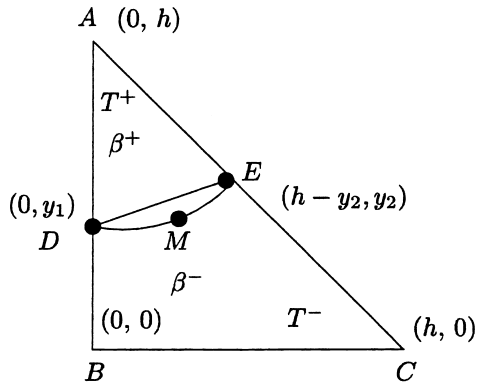


FIG. 4.1. *A typical interface element $T = \triangle ABC$. The arc DME is the part of the interface Γ in T. It is approximated by the line segment $\overline{DE}$. $T^+ = \triangle ADE$, $T^- = T - T^+$, and T_r is the region enclosed by the $\overline{DE}$ and the arc DME.*

C are specified, a local nonconforming finite-element basis function ϕ for this interface triangle is determined by

$$\phi(x) = \begin{cases} \phi^+(x) = a_0 + a_1 x_1 + a_2(x_2 - h) & \text{if } x = (x_1, x_2) \in T^+, \\ \phi^-(x) = b_0 + b_1 x_1 + b_2 x_2 & \text{if } x = (x_1, x_2) \in T^-. \end{cases} \tag{4.1}$$

The coefficients a_i and b_i $(i = 0, 1, 2)$ are determined by the conditions (cf. [18])

$$\phi^+(D) = \phi^-(D), \quad \phi^+(E) = \phi^-(E), \quad \beta^+ \frac{\partial \phi}{\partial n}^+ = \beta^- \frac{\partial \phi}{\partial n}^-, \tag{4.2}$$

where n is the unit normal direction of the line segment $\overline{DE}$. It is shown in [18] that this function ϕ is uniquely determined. Note that basis functions defined in this way can be discontinuous across edges of interface elements. So, this defines a nonconforming finite element.

4.2. A conforming linear element. This conforming element was proposed in [18]. Both error estimates and numerical experiments in [18] show that the corresponding interpolation errors that can depend on an angle condition are of second order, which is optimal [18].

Let $T = \Delta ABC$ be an interface element; cf. Figure 4.2. As before, we assume that the interface meets edges of this element at D and E. To construct a local basis function that is globally continuous, we extend the previously defined basis function at the same node (vertex) to one more triangle along the interface (cf. Figure 4.2 (b)). We require that the local basis functions in two adjacent interface elements, such as ΔABC and ΔAFB, take the same value at the interface point on their common edge, such as the point D. This will achieve the global continuity of a basis function associated with an irregular node.

We construct a local basis function ϕ by assigning its values at the vertices A, B, C, F, and I, respectively. This construction consists of the following five steps.

P1. Use the values at the nodes A, B, C, F, and I to construct the three nonconforming finite-element basis functions defined as in subsection 4.1 on the elements ΔABC, ΔAFB, and ΔACI, respectively.

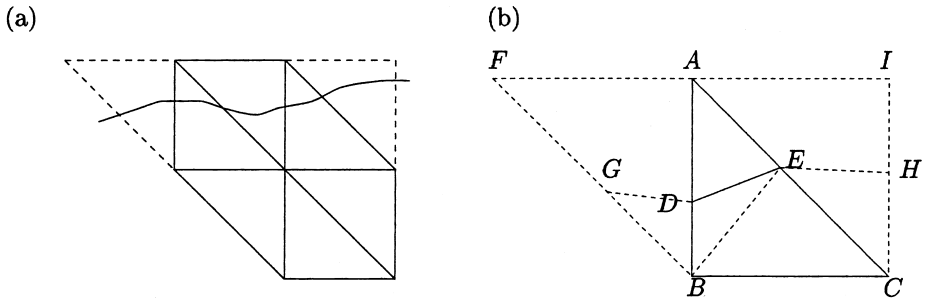


FIG. 4.2. (a) *The support of a local basis function.* (b) *A diagram for the construction of a local basis function on* $\triangle ABC$.

P2. Set the value at D as the average of the values at D of the nonconforming piecewise linear basis functions defined on ΔABC and ΔAFB constructed in P1.

P3. Similarly, set the value at E as the average of values at E of the nonconforming piecewise linear basis functions defined on the elements ΔABC and ΔACI constructed in P1.

P4. Partition the element ΔABC into three subtriangles by an auxiliary line, say, line segment $\overline{BE}$, or $\overline{DC}$. We choose the auxiliary line in such a way that at least one of angles (or complimentary angles if the angle is more than $\pi/4$) is bigger than or equal to $\pi/2$.

P5. Define the basis function to be the piecewise linear function in the three subtriangles determined by the values at the points A, B, C, D, and E.

For this type of conforming finite-element space and homogeneous jump conditions, i.e., $w = 0$ and $Q = 0$, we have the following error estimates of the interpolation function:

$$\|I_h u - u\|_\infty \le Ch^2 \|D^2 u\|_{\infty, \Omega \setminus (\cup T_r)}, \tag{4.3}$$

$$\|I_h u - u\|_{H^1(\Omega)} \le Ch \|D^2 u\|_{\infty, \Omega \setminus (\cup T_r)}, \tag{4.4}$$

assuming that $\beta(x)$ is a piecewise constant and the solution is piecewise smooth in Ω^- and Ω^+, where $C > 0$ is a generic constant independent of h and $\cup T_r$ is the region of mismatched regions between Γ and Γ_h; see [18] for the proof.

4.3. A conforming quadratic element. We now construct a conforming quadratic element. As before, the basis functions associated with regular nodes are the standard conforming linear finite-element basis functions. But, the basis functions associated with irregular nodes are piecewise quadratic. All the basis functions are globally continuous. The idea is to average the tangential derivatives in the construction in subsection 4.2. Referring the reader to Figure 4.2, we describe this procedure as follows.

P1. Use the values at the nodes A, B, C, F, and I to construct the three nonconforming linear finite-element basis functions defined in subsection 4.1 on the elements ΔABC, ΔAFB, and ΔACI, respectively.

P2. Set the value at D as the average of the values at D of the nonconforming linear finite-element basis functions defined ΔABC and ΔAFB constructed in P1. Assign the tangential derivative at D along AD to be the average of the values of the tangential derivatives (along AD) of the nonconforming

linear finite-element basis functions on ΔABC and ΔAFB constructed in P1. Similarly, assign the tangential derivative at D along DB as the average of the values of the tangential derivatives (along DB) of the nonconforming finite-element basis functions on ΔABC and ΔAFB constructed in P1.

P3. Repeat to set the value at E as the average of values at E of the nonconforming linear finite-element basis functions defined on the elements ΔABC and ΔACI in P1. Also assign the tangential derivatives along AE and EC as the average of those from the nonconforming finite-element basis functions along the same edge.

P4. Partition the element ΔABC into three subtriangles by an auxiliary line. We choose the auxiliary line in such a way that at least one of the angles (or complimentary angles if the angle is bigger than $\pi/4$) is bigger than or equal to $\pi/2$. In Figure 4.2, for example, the angles are $\angle EDB$, $\angle DEB$, and $\angle EBD$, where the complimentary angle of $\angle EDB$ is bigger than $\pi/4$.

P5. Assign the tangential derivatives along DE and BE exactly the same (no average) as those from the nonconforming finite-element basis functions.

P6. Set the values of the tangential derivatives along BE and BC exactly the same as those from the nonconforming finite-element basis function on ΔABC.

P7. Define the basis function ψ_i to be the piecewise quadratic function in the three subtriangles determined by the values at the points A, B, C, D, and E, respectively, and the tangential derivatives from each side of the triangles; see Figure 4.3 for an illustration.

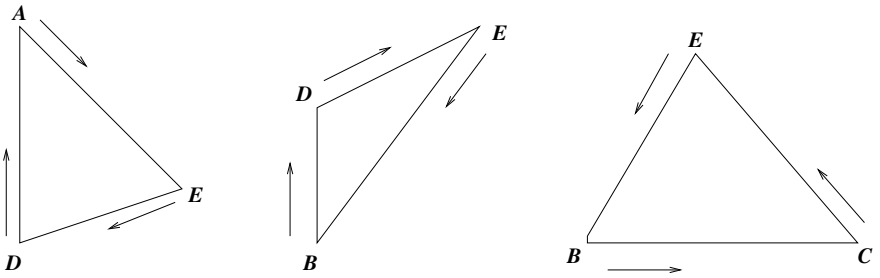


FIG. 4.3. *Three triangles on which piecewise continuous quadratic functions can be determined. The symbol "⟶" indicates a tangential derivative.*

We now give an error estimate for the interpolation error of this element for a given piecewise smooth function v. We first show that a quadratic function on a triangle is uniquely determined by its values at the three vertices and its tangential derivatives along the three sides at some particular points.

LEMMA 4.1. *Referring to the geometry in Figure* 4.4, *there is a unique quadratic polynomial*

$$I_h v(x,y) = v_A + a_{10}x + a_{01}y + a_{20}x^2 + a_{11}xy + a_{02}y^2 \tag{4.5}$$

that interpolates the values of v_A, v_D, and v_E, and three tangential derivatives v'_{AE}, v'_{AD}, and v'_{ED} defined at the three vertices A, E, and D, respectively.

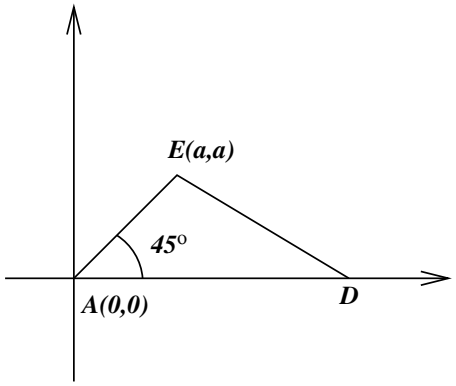


FIG. 4.4. *An illustration of determining the quadratic interpolation function.*

Proof. Using the undetermined coefficient method, we find the unique expression of the coefficients[2]

$$a_{10} = \frac{\tilde{h}v'_{AD} + 2v_D - 2v_A}{\tilde{h}},$$

$$a_{01} = -\frac{\tilde{h}v'_{AD} - \sqrt{2}v'_{AE}\tilde{h} + 2v_D - 2v_A}{\tilde{h}},$$

$$a_{20} = -\frac{\tilde{h}v'_{AD} + v_D - v_A}{\tilde{h}^2},$$

$$a_{11} = -\frac{-2\tilde{h}\alpha v'_{AD} - 2\alpha v_D + 2\alpha v_A + \tilde{h}^2 v'_{AD} + 2\tilde{h}v_D}{\tilde{h}^2\alpha}$$
$$- \frac{\sqrt{\tilde{h}^2 - 2\tilde{h}\alpha + 2\alpha^2}v'_{ED}\tilde{h} + \tilde{h}\alpha\sqrt{2}v'_{AE} - 2\tilde{h}v_E}{\tilde{h}^2\alpha},$$

$$a_{02} = \frac{\tilde{h}^2 v'_{AD}\alpha + 2\tilde{h}\alpha v_D - \tilde{h}^2\alpha\sqrt{2}v'_{AE} - \tilde{h}v'_{AD}\alpha^2 - \alpha^2 v_D + \alpha^2 v_A}{\tilde{h}^2\alpha^2}$$
$$+ \frac{v_E\tilde{h}^2 - v_A\tilde{h}^2 + \sqrt{\tilde{h}^2 - 2\tilde{h}\alpha + 2\alpha^2}v'_{ED}\tilde{h}\alpha + \tilde{h}\alpha^2\sqrt{2}v'_{AE} - 2v_E\tilde{h}\alpha}{\tilde{h}^2\alpha^2},$$

where $\tilde{h}$ is the distance from A to D, and the coordinates at E are (α, α). ∎

There are two steps in obtaining the interpolation error. First, the values of the quadratic function according to our algorithm at vertices either are directly copied from the conforming linear basis function (at the vertices of the right triangles) or are the average of values of the conforming linear basis function from the two adjacent triangles. Therefore, those values are $O(h^2)$ perturbations to v from the reference [18]. Similarly, the values of the tangential derivatives are $O(h)$ perturbations to that of v. Second, from the above reasoning, we can conclude that the values at the mid-points of each side of the triangle are $O(h^2)$ perturbations to v. Since the quadratic function can also be uniquely determined from its values at vertices and the three midpoints

[2]We use Maple to get the coefficients.

and it is insensitive to perturbations of those values (see [14]), we conclude that the interpolation function is an $O(h^2)$ perturbation to $v(x)$ in the entire triangle.

The second part of the argument is a direct consequence of the interpolation error estimate for the standard conforming P_2 finite element. The first part of the argument is a consequence of the following result.

LEMMA 4.2. *Let $a, b \in \mathbb{R}$ with $h := b - a > 0$. Let $v \in C^1[a, b]$. Then, there exists a unique quadratic polynomial p such that $p(a) = v(a)$, $p(b) = v(b)$, and $p'(a) = v'(a)$. Moreover, if v is smooth, then*

$$\|v - p\|_{L^\infty}(a, b) \le Ch^3,$$

$$\|v' - p'\|_{L^\infty}(a, b) \le Ch^2,$$

where $C > 0$ is a constant independent of h.

Proof. Note that any quadratic polynomial q can be written as

$$q(x) = q(a) + q'(a)(x - a) + \frac{q(b) - q(a) - q'(a)(b - a)}{(b - a)^2}(x - a)^2.$$

Therefore, the unique quadratic polynomial that interpolates $v(a)$, $v(b)$, and $v'(a)$ is given by

$$p(x) = v(a) + v'(a)(x - a) + \frac{v(b) - v(a) - v'(a)(b - a)}{(b - a)^2}(x - a)^2.$$

Now an application of the Bramble–Hilbert lemma [8] concludes the proof. ∎

The lemma is still true if we replace $p'(a)$ by $p'(b)$.

By the lemma, we see that if $p(a) = v(a) + O(h^2)$, $p(b) = v(b) + O(h^2)$, and $p'(a) = v'(a) + O(h)$, then $|v(x) - p(x)| = O(h^2)$ and $|v'(x) - p'(x)| = O(h)$ in the entire interval $[a, b]$ including the midpoint.

With the same local angle constraint given in [18] (e.g., at least one angle or complementary angle is larger than or equal to $\pi/4$), we have the following error estimates for the interpolation.

PROPOSITION 4.3. *Assume that $w = Q = 0$, $\beta(x)$ is a piecewise constant, and the solution u is piecewise C^2, i.e., $u^\pm \in C(\overline{\Omega^\pm})$. Then the following error estimates hold:*

$$\|I_h u - u\|_\infty \le Ch^2 \|D^2 u\|_{\infty, \Omega \setminus \sum T_r}, \tag{4.6}$$

$$\|I_h u - u\|_{H^1(\Omega)} \le Ch \|D^2 u\|_{\infty, \Omega \setminus \sum T_r}. \tag{4.7}$$

For the conforming linear finite-element space, the proof is given in [18]. So we just need to discuss the conforming quadratic finite-element space. The proof of the first inequality is straightforward, as has been discussed above. The proof of the second inequality is quite technical and tedious with different geometric configurations. This proof, though, is quite similar to that for the nonconforming immersed finite-element method (cf. [18]); we give only a sketch here. Taking the geometry in Figure 4.4 as an example, we have analytic expression of the interpolation function given in Lemma 4.1 in terms of its values and the tangential derivatives (averaging those of the nonconforming interpolation function). From the analytic expression of the

interpolation, it is easy to get $\|\nabla I_h u - \nabla I_h \tilde{u}\|_\infty \le Ch$, where $I_h\tilde{u}$ is any quadratic function when we perturb v_A, v_D, and v_E by $O(h^2)$, and v'_{AD}, v'_{AE}, and v'_{ED} by $O(h)$.

Our numerical tests reported in section 6 confirm such an estimate for the interpolation error. Note that, while the error bounds for the linear and quadratic basis functions are the same, numerical examples of our test problems show that the quadratic one behaves better numerically. Since the quadratic basis function is a conforming one, we can conclude that the finite-element solution is first order in the H^1 norm from the standard finite element theory.

Remark 1. Consider a uniform triangular mesh. For both conforming linear and conforming quadratic elements, the basis function ϕ_i associated with the ith node is nonzero only on the six surrounding triangular elements if the interface does not cut through any of these triangles. Otherwise, the support of a basis function includes two more triangular elements along the direction of the interface if the interface cuts through any of the surrounding triangle. A corresponding finite-difference scheme, however, generally has a nonstandard nine-point stencil.

Remark 2. If the coefficient β is continuous across the interface, i.e., $[\beta]_\Gamma = 0$, then both the linear and quadratic basis functions become the standard linear basis functions.

5. Implementation. In this section, we give details about the numerical extension of the jumps w and Q, the assembly of the stiffness matrix and load vector, and the evaluation of certain integrals on part of an interface element.

5.1. Numerical extension of the jumps w and Q. We use a Gaussian quadrature to compute the stiffness matrix and the load vector in each triangle $\triangle ADE$, $\triangle DEB$, and $\triangle BEC$ in Figure 4.2. Let x be such a Gaussian point which is close to the interface. We need to extend the jumps w and Q to this Gaussian point by the definition of w_ρ and Q_ρ given in E2 and E3 in section 2. This is done in two steps. The first step is to find an approximation of the orthogonal projection $\hat{x}$ of x on the interface Γ; see Lemma 2.1. The second step is to define the extensions of w and Q at x as $w(\hat{x})$ and $Q(\hat{x})$, respectively.

The orthogonal projection can be approximated by

$$\hat{x} = x + \alpha p,$$

where

$$p = \begin{bmatrix} \varphi_x(x) \\ \varphi_y(x) \end{bmatrix}.$$

Here, a subscript denotes a partial derivative. The scalar α is determined from the following quadratic equation:

$$\varphi(x) + (\nabla\varphi(x)\cdot p)\,\alpha + \frac{1}{2}\left(p^T \mathrm{He}(\varphi(x))\, p\right)\,\alpha^2 = 0,$$

where

$$p^T \mathrm{He}(\varphi)\, p = \varphi_x^2\,\varphi_{xx} + 2\,\varphi_x\varphi_y\varphi_{xy} + \varphi_y^2\,\varphi_{yy}.$$

The sign of α is chosen to be opposite that of $\varphi(x)$. If the underlying mesh is uniform, formulated from a Cartesian grid, then, the partial derivatives $\nabla\varphi(x)$ and the Hessian

matrix $\mathrm{He}(\varphi)$ can be computed at x using the standard centered five-point finite-difference formula. Using the method above, the computed projections have third order accuracy.

Note that in our implementation, we use the orthogonal projections $\hat{x} \in \Gamma$. It is also possible to use the orthogonal projections $\hat{x} \in \Gamma_h$. The difference in the finite-element solution using the two different implementations is small since the area of the mismatched region $\sum T_r$ is small. If the level set function is the signed distance function, or a good approximation of the signed distance function, then the error in approximating $\hat{x}$ would be $O(h^3)$; see [19]. We also refer the reader to [9,10] for some of the recent discussions on computing orthogonal projections.

Note that when one uses a finite number of piecewise linear basis functions $\phi_h' s \in V_h$ to seek the finite-element solution, we have introduced an error in substituting $V = H^1(\Omega)$ by V_h which is often $O(h^2)$. Other approximating errors, such as numerical integration, approximating orthogonal projections, have little effect on the finite-element solutions as long as they are higher order terms of h^3.

In practice, we need only to extend both jumps w and Q to $N(\Gamma, 2h)$, the $2h$ neighborhood of the interface Γ. This is described in (5.4), where the last two terms are identical but differ by a sign for noninterface triangles.

5.2. Assembly of stiffness matrix and load vector. On noninterface elements where the interface does not cut through, we can use the standard way in computing the contribution to the stiffness matrix and the load vector. On interface elements where the interface cuts through, we need to modify the load vector (the right-hand side); but the computation for the stiffness matrix remains the same as in the case of homogeneous jump conditions.

To fix this idea, let us focus on the quadratic conforming element (cf. subsection 4.3). We rewrite the finite-element equation (3.4) in terms of $u_h = q_h + u_\rho$:

(5.1)
$$\int_\Omega \beta \nabla u_h \cdot \nabla v_h \, dx = \int_\Omega f \, v_h \, dx + \int_\Omega \beta \nabla u_\rho \cdot \nabla v_h \, dx + \int_\Omega H(\varphi) \, \nabla \cdot (\beta \nabla \tilde{u}_\rho) \, v_h \, dx \qquad \forall v_h \in V_h,$$

where H is the Heaviside function and $\tilde{u}_\rho$ is given by

$$\tilde{u}_\rho(x) = w_\rho(x) + \frac{Q_\rho(x)}{\beta^+(x)} \frac{\varphi(x)}{|\nabla \varphi(x)|}, \qquad x \in \Omega. \tag{5.2}$$

Note that $\tilde{u}_\rho$ is smooth in a neighborhood of Γ and that $u_\rho = \chi_{\Omega^+} \tilde{u}_\rho$ and its corresponding flux have nonhomogeneous jumps across the interface.

Let $\{\phi_j(x)\}$ be the basis functions of the modified conforming finite-element space. Then, we define

(5.3)
$$u_h(x) = \begin{cases} \displaystyle\sum_j \alpha_j \phi_j(x) & \text{if } x \text{ belongs to a noninterface triangle,} \\ \displaystyle\sum_j \alpha_j \phi_j(x) + \sum_j u_\rho(x_j) \phi_j(x) & \text{otherwise,} \end{cases}$$

where $\phi_j(x)$ is the jth basis function centered at x_j. The left-hand side is exactly the same as in the case of the homogeneous jump condition. The entries of the stiffness matrix are

$$a_{ij} = \int_\Omega \beta \nabla \phi_i \cdot \nabla \phi_j \, dx.$$

Only the right-hand side of the system of equation needs to be modified for certain triangles near the interface.

At a noninterface triangle that is entirely in Ω^-, the last two terms of integration over the triangle are zero, since $H(\varphi(x)) = 0$ and $u_\rho = 0$. The situation is a little more complicated for triangles in Ω^+. If all the nonzero basis functions over a triangle in Ω^+ have no support from interface triangles, then the last two terms in (5.1) are canceled out. To see this, let T_k be such a triangle, and let ϕ_l be such a basis function with Ω_l being its support. We have

$$\begin{aligned}
&\int_{\Omega_l} \beta \nabla u_\rho \cdot \nabla \phi_l \, dx + \int_{\Omega_l} H(\varphi) \, \nabla \cdot (\beta \nabla \tilde{u}_\rho) \, \phi_l \, dx \\
&\qquad = \int_{\Omega_l} \beta \nabla u_\rho \cdot \nabla \phi_l \, dx + \int_{\Omega_l} \nabla \cdot (\beta \nabla \tilde{u}_\rho) \, \phi_l \, dx \\
&\qquad = \int_{\Omega_l} \beta \nabla u_\rho \cdot \nabla \phi_l \, dx + \int_{\partial \Omega_l} \frac{\partial \tilde{u}_\rho}{\partial n} \phi_l \, ds - \int_{\Omega_l} \beta \nabla \tilde{u}_\rho \cdot \nabla \phi_l \, dx = 0,
\end{aligned}$$

since $\tilde{u}_\rho = u_\rho$ and $\phi_l = 0$ along $\partial \Omega_l$. In other words, the total contribution of the line integral along the boundary of each triangle summed up to be zero. The right-hand side of the load vector can be summarized as

$$F_i = \begin{cases} \displaystyle\int_{\Omega_i} f \phi_i \, dx & \text{if } \Omega^s(\phi_i) \cap \Gamma = \emptyset, \\ \displaystyle\int_{\Omega_i} f \phi_i \, dx + \int_{\Omega^s} H(\varphi) \nabla \cdot (\beta \nabla \tilde{u}_\rho) \, \phi_i \, dx & \\ \qquad \displaystyle + \sum_j u_\rho(x_j) \int_{\Omega_i} \beta \nabla \phi_i \cdot \nabla \phi_j \, dx & \text{otherwise,} \end{cases} \tag{5.4}$$

where $\Omega^s(\phi_i)$ is the support of ϕ_i.

5.3. Evaluation of $\int_{\Omega^s} H(\varphi) \nabla \cdot (\beta \nabla \tilde{u_\rho}) \, \phi_i \, dx$. Take Figure 4.1 as an example. There are two ways to evaluate the integral. One way is to use

$$\int_{T_j} H(\varphi) \nabla \cdot (\beta \nabla \tilde{u}_\rho) \, \phi_i \, dx = \int_{\partial \triangle ADE} \beta \, \frac{\partial \tilde{u}_\rho}{\partial n} \, \phi \, ds - \int_{\triangle ADE} \nabla \tilde{u}_\rho \cdot \nabla \phi \, dx. \tag{5.5}$$

The line integral is evaluated using a Gaussian quadrature formula or some other numerical quadrature.

The second approach is to evaluate the double integral directly over the triangle using a quadrature formula, say, the four-point formula [14]. The coefficient β is approximated by a constant in the triangle. In order to evaluate the values of the integrand at a given point x, we first find a square $[x_i, x_{i+1}] \times [y_j, y_{j+1}]$ that contains the point x. The Laplacian at the vertices $(x_{i\pm1}, y_{j\pm1})$ is computed using the standard three-point central finite difference formula. Finally, the Laplacian at the point x is interpolated using the bilinear interpolation; see [17].

6. Numerical results. We report two examples of numerical calculations using our conforming quadratic immersed-interface finite-element method. In each of the calculations, we used ITPACK [23] to solve the resulting linear system of equations.

Example 1. *Homogeneous jump conditions.* This example is from [18]. We consider the problem (1.1)–(1.5) with $\Omega = (-1,1)\times(-1,1)$, Γ being the circle centered at point $(0,0)$ with radius $R = 0.5$, $\beta^- = 1$, and $\beta^+ = 100$. The source term $f(x,y)$ and the Dirichlet boundary data $u_0(x,y)$ are calculated from the exact solution $u(x,y)$:

$$u(x,y) = \begin{cases} \dfrac{r^3}{\beta^-} & \text{if } r \le R, \\ \dfrac{r^3}{\beta^+} + \left(\dfrac{1}{\beta^-} - \dfrac{1}{\beta^+}\right) R^3 & \text{otherwise,} \end{cases}$$

where $r = \sqrt{x^2+y^2}$. The exact solution satisfies the homogeneous jump conditions.

In Table 6.1 (a), we show a grid refinement analysis of our computations. The first column $E_N = \|u - u_h\|_\infty$ is the error of the finite-element solution measured in the L^∞ norm, and the second column is the estimate of the order of accuracy using the formula

$$\text{order} = \frac{\log\left(\|E_N\|_\infty / \|E_{2N}\|_\infty\right)}{\log 2}.$$

The sixth and eighth columns are defined as

$$e_{x,\infty} = \left\| \frac{\partial u}{\partial x} - \frac{\partial I_h u}{\partial x} \right\|_{\Omega',\infty}, \qquad e_{y,\infty} = \left\| \frac{\partial u}{\partial y} - \frac{\partial I_h u}{\partial y} \right\|_{\Omega',\infty},$$

respectively. We see clearly the second order accuracy. The fourth to the last columns are the interpolation errors. We see that the interpolation errors are of second order and its derivatives are of first order. These are optimal.

In Table 6.1 (b), we show the results and comparison of the finite-element solution in L^∞, L^2, and H^1 norms. We see second order convergence in L^∞ and L^2 norms and first order in the H^1 norm as expected.

Example 2. *A complicated interface and nonhomogeneous jump conditions.* We consider the problem (1.1)–(1.5) with $\Omega = (-1,1)\times(-1,1)$ and the interface Γ being the zero level set of the function

$$\varphi(x,y) = \sqrt{x^2+y^2} - 0.1\sin(5\theta - \pi/5) - 0.5,$$

where $\tan\theta = x/y$ and $0 \le \theta \le 2\pi$; see Figure 6.1 for the geometry. The function β is defined by

$$\beta(x,y) = \begin{cases} x^2+y^2+1 & \text{if } (x,y) \in \Omega^-, \\ b & \text{otherwise,} \end{cases}$$

where $b > 0$ is a parameter. The function f is defined by

$$f(x,y) = \begin{cases} -4\left(2x^2+2y^2+1\right) & \text{if } (x,y) \in \Omega^-, \\ 2\sin x\cos y & \text{otherwise.} \end{cases}$$

TABLE 6.1

A grid refinement analysis for Example 1, *where* p_i *are the approximated convergence order and the norms that involve the partial derivatives,* $e_{x,\infty} = \|\frac{\partial u}{\partial x} - \frac{\partial I_h u}{\partial x}\|_{\Omega',\infty}$ *and* $e_{y,\infty} = \|\frac{\partial u}{\partial y} - \frac{\partial I_h u}{\partial y}\|_{\Omega',\infty}$. (a) *The finite-element solution error and the interpolation error.* (b) *Results and comparison of the finite-element solution errors in* $L^\infty(\Omega')$, $L^2(\Omega')$, *and* $H^1(\Omega')$ *norms, where* $\Omega' = \Omega \setminus \sum T_r$.

(a)

N	$\|u-u_h\|_\infty$	p_1	$\|u-I_h u\|_\infty$	p_2	$e_{x,\infty}$	p_3	$e_{y,\infty}$	p_4
32	$1.314\ 10^{-3}$		$2.590\ 10^{-3}$		$1.056\ 10^{-1}$		$1.043\ 10^{-1}$	
64	$4.215\ 10^{-4}$	1.64	$6.799\ 10^{-4}$	1.93	$5.351\ 10^{-2}$	0.98	$5.351\ 10^{-2}$	0.96
128	$1.008\ 10^{-4}$	2.06	$1.780\ 10^{-4}$	1.93	$2.754\ 10^{-2}$	0.96	$2.754\ 10^{-2}$	0.96
256	$2.729\ 10^{-5}$	1.88	$4.501\ 10^{-5}$	1.99	$1.391\ 10^{-2}$	0.99	$1.393\ 10^{-2}$	0.99
512	$7.697\ 10^{-6}$	2.03	$1.138\ 10^{-5}$	1.99	$6.999\ 10^{-3}$	0.99	$6.999\ 10^{-3}$	0.99

(b)

N	$\|u-u_h\|_\infty$	p_5	$\|u-u_h\|_{L^2}$	p_6	$\|u-u_h\|_{H^1}$	p_7
32	1.314×10^{-3}		6.500×10^{-4}		5.777×10^{-2}	
64	4.215×10^{-4}	1.64	1.597×10^{-4}	2.03	2.661×10^{-2}	1.12
128	1.008×10^{-4}	2.06	4.001×10^{-5}	2.00	1.345×10^{-2}	0.99
256	2.729×10^{-5}	1.88	9.899×10^{-6}	2.01	6.593×10^{-3}	1.03
512	6.697×10^{-6}	2.03	2.489×10^{-6}	1.99	3.289×10^{-3}	1.00

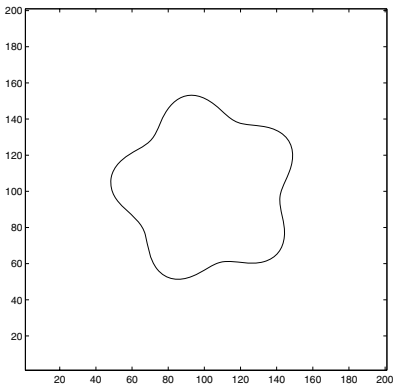


FIG. 6.1. *The domain and the interface for Example* 2.

Both β and f have nonzero jumps across the interface Γ. The exact solution is

$$u(x,y) = \begin{cases} x^2 + y^2 & \text{if}(x,y) \in \Omega^-, \\ \dfrac{1}{b}\left(\sin x \cos y + \log\sqrt{x^2+y^2}\right) & \text{otherwise.} \end{cases}$$

Notice that both the solution u and the normal flux $\beta\partial_n u$ have nonzero jumps across the interface Γ.

We remark that the solution behavior depends on the magnitude of the parameter b. If b is large, then the solution is close to a piecewise quadratic function. If b is small, then the jumps of the solution and its normal flux across the interface are very large. Numerically, this gives rise to difficulties in achieving optimal convergence

properties [1,16]. We test our method for various b and analyze the computed results.

In Table 6.2, we show a grid refinement analysis for $b = 100$. We see clearly the second order accuracy in L^∞ and L^2 norms and first order accuracy in the H^1 norm.

TABLE 6.2

A grid refinement analysis for Example 2 *with* $b = 100$*, where* p_i *are the approximated convergence order and the norms that involve the partial derivatives. Second order accuracy in* L^∞ *norm is observed.*

N	$\|u-u_h\|_\infty/\|u\|_\infty$	p_1	$\|u-u_h\|_{L^2}$	p_2	$\|u-u_h\|_{H^1}$	p_3
32	1.1995×10^{-1}		1.6705×10^{-2}		3.9175×10^{-1}	
64	2.4397×10^{-2}	2.30	1.8542×10^{-3}	3.17	1.9551×10^{-1}	1.00
128	5.3913×10^{-3}	2.18	3.2668×10^{-4}	2.51	9.8144×10^{-2}	0.99
256	1.1218×10^{-3}	2.27	5.1452×10^{-5}	2.66	4.9894×10^{-2}	0.94
512	2.7480×10^{-4}	2.03	9.4668×10^{-6}	2.44	2.5310×10^{-2}	0.98

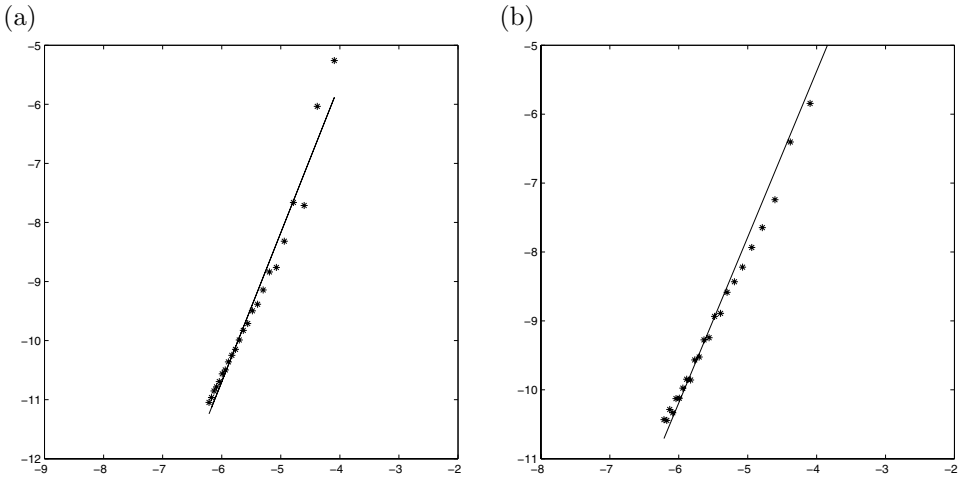


FIG. 6.2. *The linear regression analysis in the* L^∞ *norm in* log-log *scale with the mesh varying according to* $N = 40 + 20k$*,* $k = 0, 1, \ldots, 23$*.* (a) $b = 1$*; the slope (convergence order) is* 2.8122*.* (b) $b = 0.1$*; the slope is* 2.4061*.*

Where b gets smaller, the jumps in the solution and flux get larger. For interface problems, the errors obtained from non–body-fitted meshes usually do not decrease monotonically as we refine the mesh; see, for example, [15]. For small b, it is thus more realistic to find the asymptotic convergence rate as the slope of the line fitting of the experimental data $(\log(h_i), \log(E_i))$.

In Figure 6.2, we show the linear regression analysis for $b = 1$ and $b = 0.1$ for the computed finite-element solution. For these two cases, the convergence orders are 2.8122 and 2.4061. As the mesh gets finer, the linear regression analyses (by deleting the results from coarse meshes) get closer to number two, indicating a second order accuracy. In Figure 6.3, we also show the linear regression analysis in the L^2 and H^1 norms. The convergence orders are 1.9906 and 0.9135, respectively.

Figure 6.4 shows the result for $b = 0.01$ that is quite small, and hence there is a large ratio in the coefficient from both sides of the interface. The convergence order from the sample meshes ranging from 40 to 500 with 10 increments is 1.8875; see Figure 6.4(a). But as the mesh gets finer, the linear regression analyses done by cutting the results from coarse meshes get closer to number two, again indicating a second order accuracy. Figure 6.4(b) shows the convergence order to be 1.9811.

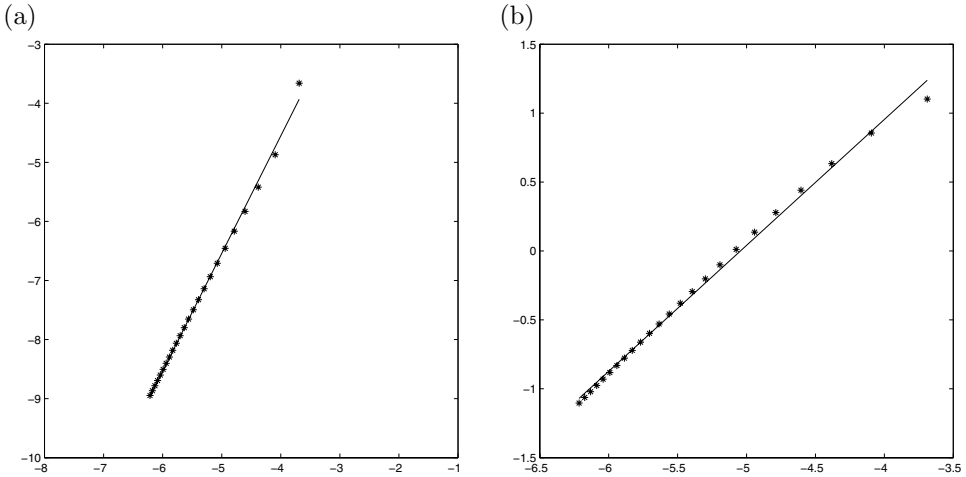


FIG. 6.3. *The linear regression analysis in the* L^2 *norm* (a)*, and in* H^1 *norm* (b)*, in* log-log *scale with the mesh varying according to* $N = 40 + 20k$*,* $k = 0, 1, \ldots, 23$*,* $b = 1$*. The slope (convergence order) is* 1.9906 *and* 0.9135*, respectively.*

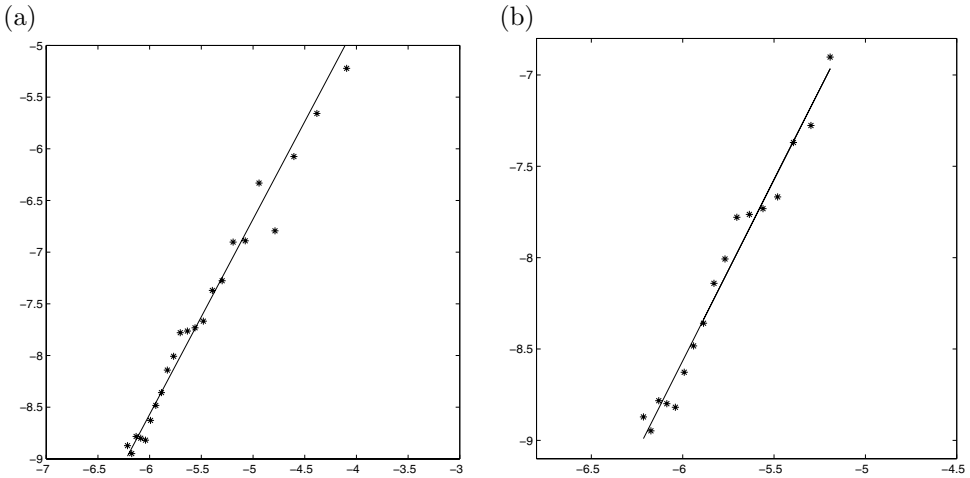


FIG. 6.4. *The linear regression analysis in the* L^∞ *norm in* log-log *scale for* $b = 0.01$. (a) $N = 40+20k$*,* $k = 0, 1, \ldots, 23$*; the slope (convergence order) is* 1.8875. (b) $N = 40+20k$*,* $k = 1, 15, \ldots, 23$*; the slope is* 1.9811.

7. Conclusions. We have developed a class of immersed-interface finite-element methods for solving elliptic interface problems with *nonhomogeneous jump conditions*. These methods consist of three parts:

(a) a weak formulation of the problem in which the nonhomogeneous jump conditions are removed by using the level-set representation of the interface;
(b) construction of immersed-interface finite-element basis functions for irregular nodes that satisfy the homogeneous jump conditions; and
(c) several techniques of numerical implementation for the resulting finite-element equations.

Our methods have several advantages. For instance, they result in symmetric positive definite systems of linear equations. Moreover, they can be used with the level-set method for fast simulations of interface dynamics. Our basic error analysis and numerical tests demonstrate that such methods have optimal convergence properties.

REFERENCES

[1] L. Adams and Z. Li, *The immersed interface/multigrid methods for interface problems*, SIAM J. Sci. Comput., 24 (2002), pp. 463–479.
[2] R. Adams, *Sobolev Spaces*, Academic Press, New York, 1975.
[3] E. Bänsch, F. Hausser, O. Lakkis, B. Li, and A. Voigt, *Finite element method for epitaxial growth with attachment-detachment kinetics*, J. Comput. Phys., 194 (2004), pp. 409–434.
[4] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, 2nd ed., Springer-Verlag, New York, 2002.
[5] W. K. Burton, N. Cabrera, and F. C. Frank, *The growth of crystals and the equilibrium of their surfaces*, Phil. Trans. Roy. Soc. London Ser. A, 243 (1951), pp. 299–358.
[6] R. E. Caflisch and B. Li, *Analysis of island dynamics in epitaxial growth of thin films*, Multiscale Model. Simul., 1 (2003), pp. 150–171.
[7] Z. Chen and J. Zou, *Finite element methods and their convergence for elliptic and parabolic interface problems*, Numer. Math., 79 (1998), pp. 175–202.
[8] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North–Holland, Amsterdam, 1978.
[9] A. Demlow and G. Dziuk, *An adaptive finite element method for the Laplace–Beltrami operator on implicity defined surfaces*, SIAM J. Numer. Anal., 45 (2007), pp. 421–442.
[10] P. Gremaud, C. Kuster, and Z. Li, *A study of numerical methods for the level set approach*, Appl. Numer. Math., 57 (2007), pp. 837–846.
[11] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Pitman Advanced Publishing Program, Boston, MA, 1985.
[12] S.-M. Hou and X.-D. Liu, *A numerical method for solving variable coefficient elliptic equation with interfaces*, J. Comput. Phys., 202 (2005), pp. 411–445.
[13] J. N. Israelachvili, *Intermolecular and Surface Forces with Applications to Colloidal and Biological Systems*, 2nd ed., Academic Press, New York, 1990.
[14] C. Johnson, *Numerical Solution of Partial Differential Equations by the Finite Element Method*, Cambridge University Press, Cambridge, UK, 1987.
[15] Z. Li, *A fast iterative algorithm for elliptic interface problems*, SIAM J. Numer. Anal., 35 (1998), pp. 230–254.
[16] Z. Li and K. Ito, *Maximum principle preserving schemes for interface problems with discontinuous coefficients*, SIAM J. Sci. Comput., 23 (2001), pp. 339–361.
[17] Z. Li and K. Ito, *The Immersed Interface Method: Numerical Solutions of PDEs Involving Interfaces and Irregular Domains*, Frontiers Appl. Math. 33, SIAM, Philadelphia, 2006.
[18] Z. Li, T. Lin, and X. Wu, *New Cartesian grid methods for interface problem using finite element formulation*, Numer. Math., 96 (2003), pp. 61–98.
[19] Z. Li, W.-C. Wang, I.-L. Chern, and M.-C. Lai, *New formulations for interface problems in polar coordinates*, SIAM J. Sci. Comput., 25 (2003), pp. 224–245.
[20] S. Osher and R. Fedkiw, *Level Set Methods and Dynamic Implicit Surfaces*, Springer-Verlag, New York, 2002.

[21] S. Osher and J. Sethian, *Front propagation with curvature dependent speed: Algorithms based on Hamilton-Jacobi formulations*, J. Comput. Phys., 79 (1988), pp. 12–49.
[22] J. A. Sethian, *Level Set Methods and Fast Marching Methods*, 2nd ed., Cambridge University Press, Cambridge, UK, 1999.
[23] D. Young and D. Kincaid, *The ITPACK package for large sparse linear systems*, in Elliptic Problem Solvers, M. Schultz et al., eds., Academic Press, New York, 1981, pp. 163–185.

SIAM J. NUMER. ANAL.
Vol. 46, No. 1, pp. 496–516
© 2008 Society for Industrial and Applied Mathematics

THE CROUZEIX–RAVIART FE ON NONMATCHING GRIDS WITH AN APPROXIMATE MORTAR CONDITION*

TALAL RAHMAN†, PETTER BJØRSTAD‡, AND XUEJUN XU§

Abstract. A new approximate mortar condition is proposed for the lowest order Crouzeix–Raviart finite element on nonmatching grids, which uses only the nodal values on the interface for the calculation of the mortar projection. This approach allows for improved and more flexible algorithms compared to those for the standard mortar condition where nodal values in the interior of a subdomain, those closest to a mortar side of the subdomain, are also required in the calculation.

Key words. Crouzeix–Raviart finite element, nonmatching grids, mortar condition, additive Schwarz

AMS subject classifications. 65F10, 65N30

DOI. 10.1137/060663593

1. Introduction. The general concept of the mortar technique was originally introduced by Bernardi, Maday, and Patera in [3], which provides a useful tool for coupling different discretization schemes. In recent years, the approach has largely been applied to nonmatching grids for the design of algorithms which are very well suited for parallel implementation and which can easily handle complicated geometries and heterogeneous materials. In order to ensure that the overall discretization makes sense, an optimal coupling between the meshes is required. In a standard mortar technique, this condition is realized by applying the condition of weak continuity on the solution, called the mortar condition, saying that the jump of the solution along the interface between two meshes is orthogonal to some suitable test space. The mortar technique has been extensively studied by many authors. A saddle point formulation for the mortar technique was studied in [2]. Later, an extension to three dimensions was introduced in [1]. Further analysis and extensions of the mortar technique can be found in [5, 9, 15, 21, 23, 24, 10] and the references therein.

In this paper, we are interested in the application of the mortar technique on nonmatching grids, where in each subgrid, the nonconforming P1 or the lowest order Crouzeix–Raviart (CR) finite element [13] is used for the discretization. The first analysis of the mortar technique for such an element was given by Marcinkowski in [15]. In the event of applying the mortar condition, it is necessary to know the function on the interface. For the conforming P1 finite element, it is enough to know the nodal values along the interface. However, for the nonconforming P1 finite element (the lowest order CR finite element), where the degrees of freedom are associated

*Received by the editors June 23, 2006; accepted for publication (in revised form) September 14, 2007; published electronically January 30, 2008.
http://www.siam.org/journals/sinum/46-1/66359.html

†Department of Mathematics, University of Bergen, c/o Center for Integrated Petroleum Research, Allégt. 41, 5007 Bergen, Norway (talal.rahman@math.uib.no).

‡Department of Mathematics, University of Bergen, Johannes Brunsgt. 12, 5008 Bergen, Norway (petter.bjorstad@math.uib.no).

§LSEC, Institute of Computational Mathematics, Academy of Mathematics and System Sciences, Chinese Academy of Sciences, P.O. Box 2719, 100080 Beijing, People's Republic of China (xxj@lsec.cc.ac.cn). The work of this author has been supported by the special funds for major state basic research projects (973) under 2005CB321701 and by the National Sciences Foundation (NSF) of China (10731060).

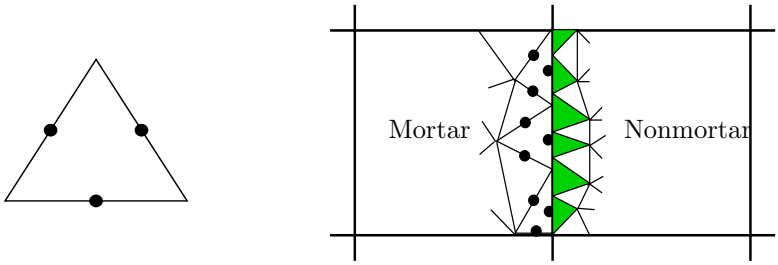


FIG. 1.1. *The lowest order CR finite element (left) and two nonmatching grids (right). CR basis functions associated with the nodes on the mortar side, denoted by dots (in the interior) and semidots (on the mortar side of the interface), have nonzero support on the nonmortar side, denoted by shaded triangles.*

with the edge midpoints, the function on the interface depends on the nodal values corresponding to interface nodes and some subdomain interior nodes lying closest to the interface; see [15] and Figure 1.1 for an illustration.

A variant of the mortar technique has recently been proposed in [17], where the standard mortar condition [15] has been modified so that the method will use only the nodal values on the interface. However, the error estimate as given in [17] requires the mesh sizes to be of the same order. In this paper, we improve this situation by modifying the mortar condition a step further, giving us an error estimate which is free from any such requirement on the mesh sizes.

The motivation to replace the standard mortar condition with an approximate mortar condition which uses nodal values only on the interface lies in the fact that the new approach will facilitate the design of improved and less intricate algorithms for the CR finite element on nonmatching grids. As we know, there already exists a number of algorithms for the conforming P1 mortar finite element, many of which, particularly the domain decomposition algorithms, can be extended to the CR mortar finite element; see, for instance, [15, 19]. In order to be able to readily extend an algorithm for the conforming P1 mortar finite element to the CR mortar finite element, it is desirable that the latter uses similar geometrical information as the former, for instance, using nodal values only on the interface for the mortar condition. An important application of the approximate mortar condition is the solution of elliptic problems with discontinuous coefficients using, for instance, parallel domain decomposition algorithms of the nonoverlapping type. Since, in the CR standard mortar case, basis functions corresponding to some interior nodes of a subdomain may have nonzero support on the neighboring subdomain, special care with respect to those nodes is required in defining the nonoverlapping subspaces in the algorithm; cf. [16, 19]. This results in a more intricate code as compared to the code in the approximate mortar case. In fact, the algorithm is affected even further; the stiffness matrices corresponding to subspaces living on the interfaces lose their block diagonal structure, which is natural in the approximate mortar approach; cf. [19]. It is also important to note that, in the application of the mortar condition for the CR finite element, the matrix on the nonmortar side, called the slave matrix, is a diagonal matrix. An application of the standard mortar condition simply destroys the advantage of having a diagonal slave matrix in the design of any effective algorithms. Another important application of the approximate mortar condition, which seems now obvious, is the construction of the intergrid transfer operator which is a vital and already an intricate part of a multilevel or multigrid algorithm; cf. [25, 18]. In the analysis of all such algorithms (those mentioned

here are for two-dimensional problems), one hardly notices the difference between using the standard and the approximate mortar conditions. However, the difference is significant in the implementation, especially for three-dimensional problems where the interfaces are no longer straight lines but planes. An analysis of a three-dimensional extension of the CR approximate mortar finite element method is already in progress.

The concept of using an approximate mortar condition has started to receive more and more attention. Recently, an additive Schwarz preconditioner for an approximate CR mortar finite element has been used to precondition the standard CR mortar finite element; see [16]. Approximate constraint has also been used in other contexts; see [4], for instance, where approximate constraint was indeed necessary for coupling the wavelet and the finite element.

The paper is organized as follows. In section 2, a detailed description of the new approximate mortar condition is given, and the corresponding discrete problem is then formulated. The analysis of the new approach is found in section 3, and a brief discussion of the implementation is given in section 5. For the purpose of experimenting, in section 4, we also propose an additive Schwarz preconditioner for the new approximate CR mortar finite element. Numerical results supporting our theory are presented in the final section.

2. The discrete problem. Let $\Omega \subset R^2$ be a bounded, simply connected polygonal domain, partitioned into a collection of nonoverlapping polygonal subdomains, $\Omega_i, i = 1, N$. We consider the following problem: Find $u^* \in H_0^1(\Omega)$ such that

$$a(u^*, v) = f(v), \quad v \in H_0^1(\Omega), \tag{2.1}$$

where

$$a(u,v) = \sum_{i=1}^{N} \int_{\Omega_i} \nabla u \cdot \nabla v \; dx \quad \text{and} \quad f(v) = \sum_{i=1}^{N} \int_{\Omega_i} fv \; dx.$$

We consider the subdomains to be geometrically conforming. With each subdomain Ω_i, we associate a quasi-uniform triangulation $\mathcal{T}_h(\Omega_i)$ of mesh size h_i. The resulting triangulation can be nonmatching across subdomain interfaces.

Let $X_{h_i}(\Omega_i)$ (or simply $X_h(\Omega_i)$) be the nonconforming P1 (CR) finite element space defined on a shape regular triangulation $\mathcal{T}_{h_i}(\Omega_i)$ (or simply $\mathcal{T}_h(\Omega_i)$) of Ω_i, consisting of functions which are piecewise linear in each triangle $\tau \in \mathcal{T}_h(\Omega_i)$, continuous at the interior edge midpoints of Ω_{ih}^{CR}, and vanishing at the edge midpoints of $\partial\Omega_{ih}^{CR} \cap \partial\Omega$ lying on the boundary $\partial\Omega$. We use a subscript for h only when we need to differentiate the discretization of one domain from the other. Let $X_h^0(\Omega_i) \subset X_h(\Omega_i)$ be the space of functions whose values at $\partial\Omega_{ih}^{CR}$ are equal to zero. Here, Ω_{ih}^{CR} and $\partial\Omega_{ih}^{CR}$ represent the sets of edge midpoints, i.e., the nonconforming P1 (CR) nodal points, of Ω_i and $\partial\Omega_i$, respectively. Note that $\overline{\Omega}_{ih}^{CR} = \Omega_{ih}^{CR} \cup \partial\Omega_{ih}^{CR}$ and $\Omega_{ih}^{CR} \cap \partial\Omega_{ih}^{CR}$ is an empty set. Without the superscript CR, the sets represent the corresponding sets of triangle vertices, i.e., the P1 conforming nodal points. Using $X_h(\Omega_i)$, we define the product space X_h on the whole domain as

$$X_h(\Omega) = X_h(\Omega_1) \times X_h(\Omega_2) \ldots \times X_h(\Omega_N).$$

Let Γ_{ij} be an open edge common to Ω_i and Ω_j, i.e., $\overline{\Gamma}_{ij} = \overline{\Omega}_i \cap \overline{\Omega}_j$. Note that each interface Γ_{ij} inherits two different discretizations from its two sides, $\mathcal{T}_{h_i}(\gamma_{ij})$ and $\mathcal{T}_{h_j}(\gamma_{ij})$. We select one side of Γ_{ij} as the master side, called the mortar, and the other

side as the slave side, called the nonmortar. We define the skeleton $\mathcal{S} = (\cup\partial\Omega_i) \setminus \partial\Omega$ of the decomposition as follows:

$$\overline{\mathcal{S}} = \bigcup_m \overline{\gamma}_m, \quad \text{with } \gamma_m \cap \gamma_n = \emptyset \ \text{ if } \ m \neq n,$$

where each γ_m denotes an open mortar edge. We write γ_m as $\gamma_{m(i)}$, if it is an edge of Ω_i, i.e., $\gamma_{m(i)} \subset \partial\Omega_i$. Let $\delta_m = \delta_{m(j)} \subset \partial\Omega_j$ be the corresponding open nonmortar edge of Ω_j that occupies the same geometrical space as $\gamma_{m(i)}$, i.e., $\gamma_{m(i)} = \Gamma_{ij} = \delta_{m(j)}$ (see Figure 2.1 for an illustration).

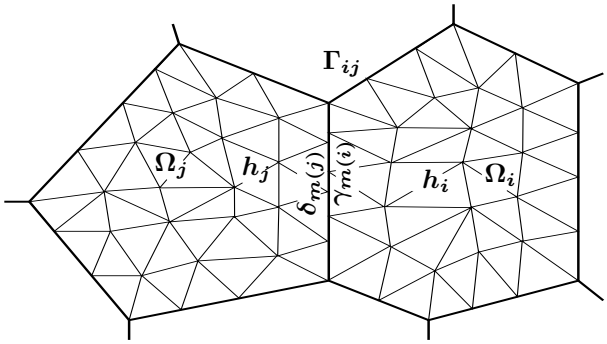


FIG. 2.1. *Illustrating a mortar side ($\gamma_{m(i)}$) and the corresponding nonmortar side ($\delta_{m(j)}$) of a subdomain interface (Γ_{ij}) with nonmatching meshes on both sides.*

Since the triangulations on Ω_i and Ω_j do not match on their common interface Γ_{ij}, the functions in $X_h(\Omega)$ are discontinuous on the set $\gamma^{CR}_{m(i)h}$ or $\delta^{CR}_{m(j)h}$ of edge midpoints on Γ_{ij}. In other words, $X_{h_i}(\gamma_{m(i)}) = X_{h_i}(\Omega_i)|_{\gamma_{m(i)}}$ differ from $X_{h_j}(\delta_{m(j)}) = X_{h_j}(\Omega_j)|_{\delta_{m(j)}}$ at those points. A weak continuity condition, called the mortar condition, is therefore imposed. Let $u_h \in X_h$, where $u_h = \{u_i\}_{i=1}^N$. A function $u_h \in X_h$ satisfies the mortar condition on $\delta_{m(j)} = \Gamma_{ij} = \gamma_{m(i)}$ if

$$Q_m J_m u_i = Q_m u_j, \tag{2.2}$$

where J_m is an interpolation operator to be defined below, and Q_m is the L^2-projection operator $Q_m : L^2(\Gamma_{ij}) \to M^{h_j}(\delta_{m(j)})$ defined as

$$(Q_m u, \psi)_{L^2(\delta_{m(j)})} = (u, \psi)_{L^2(\delta_{m(j)})} \qquad \forall \psi \in M^{h_j}(\delta_{m(j)}), \tag{2.3}$$

where $M^{h_j}(\delta_{m(j)}) \subset L^2(\Gamma_{ij})$ is the test space of functions which are piecewise constant on the triangulation of $\delta_{m(j)}$, and $(\cdot,\cdot)_{L^2(\delta_{m(j)})}$ denotes the L^2 inner product on $L^2(\delta_{m(j)})$. We remark that in the original mortar method (cf. [15]) the operator J_m is simply the identity operator.

Let $P^1(\tau)$ be the space of linear functions on a triangle $\tau \in \mathcal{T}_h(\Omega_i)$, uniquely determined by their values at the vertices. Denote by $Z_h(\Omega_i) = \Pi_{\tau\in\mathcal{T}_h(\Omega_i)} P^1(\tau)$ the space of piecewise linear functions defined on the triangulation $\mathcal{T}_h(\Omega_i)$. Let $\mathcal{T}_{\frac{h}{2}}(\Omega_i)$ be the triangulation associated with the Ω_i, which is obtained as a result of dividing the edges of $\mathcal{T}_h(\Omega_i)$. Denote by $Y_{\frac{h}{2}}(\Omega_i)$ the conforming space of piecewise linear and continuous functions on the triangulation $\mathcal{T}_{\frac{h}{2}}(\Omega_i)$. The functions of this space are

defined by their values at the set $\overline{\Omega}_{i\frac{h}{2}}$ of all edge endpoints of $\mathcal{T}_{\frac{h}{2}}(\Omega_i)$. Let $Y^0_{\frac{h}{2}}(\Omega_i) \subset Y_{\frac{h}{2}}(\Omega_i)$ be the space of functions whose values at $x \in \partial\Omega_{i\frac{h}{2}}$ are equal to zero. It is easy to see that $\overline{\Omega}_{i\frac{h}{2}} = \overline{\Omega}^{CR}_{ih} \cup \overline{\Omega}_{ih}$.

Let $\mathcal{E}$ denote an edge (a triangle edge) or an edge segment. The midpoint and the left and right endpoints of each edge (or edge segment) $\mathcal{E}$ are denoted by $x^{\mathcal{E}}_m$, $x^{\mathcal{E}}_l$, and $x^{\mathcal{E}}_r$, respectively. The length of $\mathcal{E}$ is denoted by $h_{\mathcal{E}}$.

We now define the operator $J_m : X_h(\gamma_m) \to Z_h(\gamma_m)$. It is based on the definition of another interpolation operator $I_m : X_h(\gamma_m) \to Y_{\frac{h}{2}}(\gamma_m)$; see Definition 2.1.

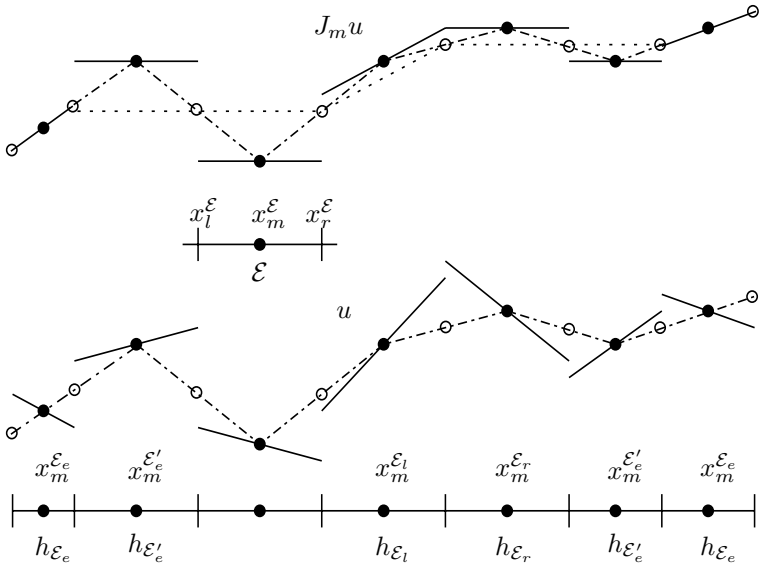


FIG. 2.2. *Illustrating the interpolation operator J_m. Here $u \in X_h(\gamma_m)$ is a (discontinuous) piecewise linear function shown (using solid lines) in the lower figure. The dashed-dotted lines (in both the upper and lower figures) correspond to $I_m u$; cf. Definition* 2.1. *The solid lines in the upper figure correspond to $J_m u$; cf. Definition* 2.2.

DEFINITION 2.1. *For $u \in X_h(\gamma_m)$, $I_m u \in Y_{\frac{h}{2}}(\gamma_m)$ is defined by the nodal values as*

$$I_m u(x) = \begin{cases} u(x), & x \in \gamma^{CR}_{mh}, \\ \frac{h_{\mathcal{E}_r}}{h_{\mathcal{E}_l}+h_{\mathcal{E}_r}} u(x^{\mathcal{E}_l}_m) + \frac{h_{\mathcal{E}_l}}{h_{\mathcal{E}_l}+h_{\mathcal{E}_r}} u(x^{\mathcal{E}_r}_m), & x \in \gamma_{mh}, \\ u(x^{\mathcal{E}_e}_m) + \frac{h_{\mathcal{E}_e}}{h_{\mathcal{E}_e}+h_{\mathcal{E}'_e}} \left(u(x^{\mathcal{E}_e}_m) - u(x^{\mathcal{E}'_e}_m) \right), & x \in \partial\gamma_{mh}. \end{cases}$$

Here, $\mathcal{E}_l$ and $\mathcal{E}_r$ are the left and the right neighboring edges of $x \in \gamma_{mh}$, respectively. $\mathcal{E}_e$ represents a triangle edge of $\mathcal{T}_h(\gamma_m)$, touching $\partial\gamma_m$, and $\mathcal{E}'_e$ is the corresponding neighboring edge.

The interpolation is done basically by first joining the neighboring edge midpoints using straight lines, and then simply extending the two end straight lines toward the end of the mortar γ_m; see Figure 2.2. The operator J_m can now be defined using I_m.

DEFINITION 2.2. *For $u \in X_h(\gamma_m)$, $J_m u \in Z_h(\gamma_m)$ is a piecewise linear function on the edges, $\{\mathcal{E}\}$, of γ_m, defined by its values at the two endpoints, $x^{\mathcal{E}}_l \in \overline{\gamma}_{mh}$ and*

$x_r^{\mathcal{E}} \in \overline{\gamma}_{mh}$, *of each such edge* $\mathcal{E}$. *If* $\mathcal{E}$ *is an interior edge of* γ_m, *then*

$$J_m u(x) = \begin{cases} u(x_m^{\mathcal{E}}) + \frac{1}{2}\left\{I_m u(x_l^{\mathcal{E}}) - I_m u(x_r^{\mathcal{E}})\right\}, & x = x_l^{\mathcal{E}}, \\ u(x_m^{\mathcal{E}}) + \frac{1}{2}\left\{I_m u(x_r^{\mathcal{E}}) - I_m u(x_l^{\mathcal{E}})\right\}, & x = x_r^{\mathcal{E}}. \end{cases}$$

It is easy to see that, if $\mathcal{E}$ *is a boundary edge of* γ_m, *then* $J_m u(x) = I_m u(x)$ *for* $x = x_l^{\mathcal{E}}, x_r^{\mathcal{E}}$.

As seen from Figure 2.2; for each edge $\mathcal{E}$ with edge midpoint $x_m^{\mathcal{E}}$, the straight line $J_m u$ passes through $u(x_m^{\mathcal{E}})$; in other words, $J_m u(x_m^{\mathcal{E}}) = I_m u(x_m^{\mathcal{E}}) = u(x_m^{\mathcal{E}})$. It is not difficult to see that the operator J_m preserves all linear functions on the mortar.

Let $V_h \subset X_h$ be the subspace of functions which satisfy the mortar condition for all $\delta_m \subset \mathcal{S}$. Since functions of V_h are not continuous, for $u, v \in X_h$, we use the broken bilinear form $a_h(\cdot,\cdot)$ defined by

$$a_h(u,v) = \sum_{i=1}^{N} a_i(u,v) = \sum_{\tau \in \mathcal{T}_h(\Omega_i)} \int_\tau \nabla u \cdot \nabla v \, dx$$

and the corresponding broken H^1 seminorm $|u|^2_{H^1_h(\Omega_i)} = \sum_{\tau \in \mathcal{T}_h(\Omega_i)} |u|^2_{H^1(\tau)}$. The discrete problem takes the following form: Find $u_h^* = \{u_i\}_{i=1}^N \in V_h$ such that

$$a_h(u_h^*, v_h) = f(v_h) \quad \forall v_h \in V_h. \tag{2.4}$$

In our analysis, we will be using the subspace $H^{\frac{1}{2}}_{oo}(\delta_m)$ which is defined as consisting of functions $v \in H^{\frac{1}{2}}(\delta_m)$ such that their trivial extensions by zero to all of $\partial\Omega_j$ belong to the space $H^{\frac{1}{2}}(\partial\Omega_j)$; cf. [14]. The corresponding norm $\|\cdot\|_{H^{\frac{1}{2}}_{oo}(\delta_m)}$ is defined as

$$\| v \|_{H^{\frac{1}{2}}_{oo}(\delta_m)} = \inf_{w \in H^1_{0,\partial\Omega_j \setminus \delta_m}(\Omega_j), w|_{\delta_m} = v} \| v \|_{H^1(\Omega_j)},$$

where $H^1_{0,\delta}$ is the closure in H^1 of all C^∞ functions vanishing on δ.

3. Analysis. Since, by construction, both $u \in X_h(\gamma_m)$ and $J_m u \in Z_h(\gamma_m)$ are linear on any edge $\mathcal{E}$ belonging to an interface γ_m, and they have the same value at the edge midpoint $x_m^{\mathcal{E}}$, their integral averages over the edge are the same; in other words, $\frac{1}{|\mathcal{E}|}\int_{\mathcal{E}} J_m u \, dx = \frac{1}{|\mathcal{E}|}\int_{\mathcal{E}} u \, dx$. This we will see is essential for our analysis.

DEFINITION 3.1. *For any interior edge* $\mathcal{E} \subset \gamma_m$, *where* γ_m *is a subdomain interface, such that* $\mathcal{E}$ *is not touching the boundary* $\partial\gamma_m$, *we call* $B_{\mathcal{E}}$ *the set of all triangles* τ *such that* $\partial\tau \cap \mathcal{E} \neq \emptyset$, *that is, all triangles touching* $\mathcal{E}$; *see Figure* 3.1 *(middle). If* $\mathcal{E}$ *is a boundary edge of* γ_m *(see Figure* 3.1 *(left)), then* $B_{\mathcal{E}}$ *is the set of all triangles touching* $\mathcal{E}$ *excluding the ones touching only at the endpoint* $\partial\mathcal{E} \cap \partial\gamma_m$ *of* $\mathcal{E}$. *We call* $B_{\mathcal{E}_i\mathcal{E}_j}$ *the set of triangles* τ *connecting the two triangle edges* $\mathcal{E}_i$ *and* $\mathcal{E}_j$ *(see Figure* 3.1 *(right)) in such a way that* $\partial\tau \cap \mathcal{E} \neq \emptyset$, *where* $\mathcal{E} = (x_m^{\mathcal{E}_i}, x_m^{\mathcal{E}_j})$ *is the straight line joining the two edge midpoints* $x_m^{\mathcal{E}_i}$ *and* $x_m^{\mathcal{E}_j}$.

Using the continuity of the function u at edge midpoints and the discrete equivalence of the H^1-seminorm, it is easy to see that, for any two neighboring edges $\mathcal{E}_i$ and $\mathcal{E}_j$ with $\partial\mathcal{E}_i \cap \partial\mathcal{E}_j \neq \emptyset$,

$$\left(u(x_m^{\mathcal{E}_i}) - u(x_m^{\mathcal{E}_j})\right)^2 \leq c \sum_{\tau \in B_{\mathcal{E}_i\mathcal{E}_j}} |u|^2_{H^1(\tau)}, \tag{3.1}$$

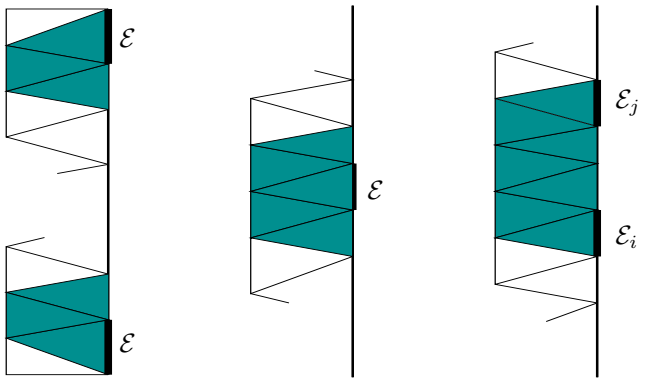


FIG. 3.1. *Shaded areas representing $B_{\mathcal{E}}$ (left and middle) and $B_{\mathcal{E}_i\mathcal{E}_j}$ (right).*

where the sum is taken over the elements of the set $B_{\mathcal{E}_i\mathcal{E}_j}$.

Now, we follow the definition of I_m; cf. Definition 2.1. If $\mathcal{E}$ is an interior edge of γ_m, then a simple calculation yields

$$I_m u(x_r^{\mathcal{E}}) - I_m u(x_l^{\mathcal{E}}) = \left[\frac{h_{\mathcal{E}}}{h_{\mathcal{E}} + h_{\mathcal{E}_r}} \left\{ u(x_m^{\mathcal{E}_r}) - u(x_m^{\mathcal{E}}) \right\} + \frac{h_{\mathcal{E}}}{h_{\mathcal{E}} + h_{\mathcal{E}_l}} \left\{ u(x_m^{\mathcal{E}}) - u(x_m^{\mathcal{E}_l}) \right\} \right],$$

where $\mathcal{E}_l$ and $\mathcal{E}_r$ are, respectively, the left and right neighboring edges of $\mathcal{E}$. If $\mathcal{E}$ is the right boundary edge of γ_m, then we have the special case

$$I_m u(x_r^{\mathcal{E}}) - I_m u(x_l^{\mathcal{E}}) = \frac{2h_{\mathcal{E}}}{h_{\mathcal{E}} + h_{\mathcal{E}_l}} \left\{ u(x_m^{\mathcal{E}}) - u(x_m^{\mathcal{E}_l}) \right\},$$

and, similarly, if $\mathcal{E}$ is the left boundary edge of γ_m, then

$$I_m u(x_r^{\mathcal{E}}) - I_m u(x_l^{\mathcal{E}}) = \frac{2h_{\mathcal{E}}}{h_{\mathcal{E}} + h_{\mathcal{E}_r}} \left\{ u(x_m^{\mathcal{E}_r}) - u(x_m^{\mathcal{E}}) \right\}.$$

Hence, by applying (3.1), we get

$$(I_m u(x_r^{\mathcal{E}}) - I_m u(x_l^{\mathcal{E}}))^2 \;\le\; c \sum_{\tau \in B_{\mathcal{E}}} |u|^2_{H^1(\tau)} \tag{3.2}$$

for each edge $\mathcal{E} \subset \gamma_m$, where the sum is taken over the elements of the set $B_{\mathcal{E}}$. Note that $B_{\mathcal{E}}$ equals either $B_{\mathcal{E}\mathcal{E}_l}$ or $B_{\mathcal{E}\mathcal{E}_r}$ if $\mathcal{E}$ is a boundary edge and equals $B_{\mathcal{E}\mathcal{E}_l} \cup B_{\mathcal{E}\mathcal{E}_r}$ if $\mathcal{E}$ is an interior edge.

Now we follow the definition of J_m; cf. Definition 2.2. For any edge $\mathcal{E} \subset \gamma_m$, the following is true:

$$J_m u(x_r^{\mathcal{E}}) - u(x_m^{\mathcal{E}}) = -\left(J_m u(x_l^{\mathcal{E}}) - u(x_m^{\mathcal{E}}) \right) = \frac{1}{2} \left[I_m u(x_r^{\mathcal{E}}) - I_m u(x_l^{\mathcal{E}}) \right].$$

If $x \in \mathcal{E}$ is a point lying on the right of $x_m^{\mathcal{E}}$, then

$$J_m u(x) - u(x_m^{\mathcal{E}}) = \frac{\operatorname{dist}(x, x_m^{\mathcal{E}})}{h_{\mathcal{E}}} \left[I_m u(x_r^{\mathcal{E}}) - I_m u(x_l^{\mathcal{E}}) \right],$$

or, if it is lying on the left of $x_m^{\mathcal{E}}$, then

$$J_m u(x) - u(x_m^{\mathcal{E}}) = -\frac{\operatorname{dist}(x, x_m^{\mathcal{E}})}{h_{\mathcal{E}}} \left[I_m u(x_r^{\mathcal{E}}) - I_m u(x_l^{\mathcal{E}}) \right],$$

where $\operatorname{dist}(x, x_m^{\mathcal{E}})$ is the distance between the points x and $x_m^{\mathcal{E}}$. If $x^{\mathcal{E}}$ is a point anywhere on the edge $\mathcal{E}$, then by applying (3.2), we get

$$(J_m u(x^{\mathcal{E}}) - u(x_m^{\mathcal{E}}))^2 \ \le \ c \sum_{\tau \in B_{\mathcal{E}}} |u|^2_{H^1(\tau)} \tag{3.3}$$

for edge $\mathcal{E} \subset \gamma_m$, where the sum is taken over the elements of the set $B_{\mathcal{E}}$.

The properties of J_m can be stated in the following lemmas.

LEMMA 3.2. *For $u_i \in X_h(\Omega_i)$, and each $\mathcal{E} \subset \gamma_m$, we have*

$$\| u_i - J_m u_i \|^2_{L^2(\mathcal{E})} \le c h_i \sum_{\tau \in B_{\mathcal{E}}} |u_i|^2_{H^1(\tau)}$$

and

$$\| u_i - J_m u_i \|^2_{L^2(\gamma_{m(i)})} \le c h_i |u_i|^2_{H^1_h(\Omega_i)}.$$

Proof. Let $\overline{(u_i)}_{\mathcal{E}}$ denote the average of u_i over the edge $\mathcal{E}$; that is, $\overline{(u_i)}_{\mathcal{E}} = \frac{1}{|\mathcal{E}|} \int_{\mathcal{E}} u_i \, dx$. From the trace and the Poincaré inequalities and (3.3), we get

$$\begin{aligned}
&\| u_i - J_m u_i \|^2_{L^2(\mathcal{E})} \\
&\quad \le 2 \| u_i - \overline{(u_i)}_{\mathcal{E}} \|^2_{L^2(\mathcal{E})} + 2 \| J_m u_i - \overline{(u_i)}_{\mathcal{E}} \|^2_{L^2(\mathcal{E})} \\
&\quad \le c h_i |u_i|^2_{H^1(\tau_{\mathcal{E}})} + c h_i \sum_{x = x_r^{\mathcal{E}}, x_m^{\mathcal{E}}, x_l^{\mathcal{E}}} (J_m u_i(x) - u_i(x_m^{\mathcal{E}}))^2 \\
&\quad \le c h_i |u_i|^2_{H^1(\tau_{\mathcal{E}})} + c h_i \sum_{\tau \in B_{\mathcal{E}}} |u_i|^2_{H^1(\tau)} \\
&\quad \le c h_i \sum_{\tau \in B_{\mathcal{E}}} |u_i|^2_{H^1(\tau)}.
\end{aligned}$$

The second inequality follows immediately from the first one by taking the sum over the triangle edges $\mathcal{E}$ of $\gamma_{m(i)}$. ∎

The next lemma is crucial for the analysis. The factor $h_j^{1/2}$ in its estimate is essential for our estimate of the consistency error. In the first part of its proof, we will need the operator $\mathcal{M}_i : X_h(\Omega_i) \to Y_{\frac{h}{2}}(\Omega_i)$ (cf. [20]): for $u \in X_h(\Omega_i)$,

$$\mathcal{M}_i u(x) = \begin{cases} u(x), & x \in \overline{\Omega}_{ih}^{CR}, \\ \frac{1}{n_x} \sum_{\tau \in B_x} u|_\tau(x), & x \in \Omega_{ih}, \\ \frac{h_{\mathcal{E}_l}}{h_{\mathcal{E}_l} + h_{\mathcal{E}_r}} u(x_m^{\mathcal{E}_l}) + \frac{h_{\mathcal{E}_r}}{h_{\mathcal{E}_l} + h_{\mathcal{E}_r}} u(x_m^{\mathcal{E}_r}), & x \in \partial\Omega_{ih}. \end{cases}$$

Here, B_x is the set of all triangles of $\mathcal{T}_h(\Omega_i)$ sharing the vertex x, and n_x is the number of such elements. The following are the properties of $\mathcal{M}_i$ which we will use:

$$\| u - \mathcal{M}_i u \|^2_{L^2(\mathcal{E})} \le c h_i \sum_{\tau \in \hat{B}_{\mathcal{E}}} |u_i|^2_{H^1(\tau)} \quad \text{and} \quad |\mathcal{M}_i u_i|^2_{H^1(\Omega_i)} \le c |u_i|^2_{H^1(\Omega_i)}. \tag{3.4}$$

In the first inequality above, $\mathcal{E}$ is an edge of $\partial\Omega_i$, where $\hat{B}_{\mathcal{E}}$ is the set of all triangles τ touching $\mathcal{E}$. We remark here that the number of triangles touching any vertex in a subdomain is bounded independently of the mesh size. The first estimate of (3.4) can be proved in a similar way as the proof of the first estimate in Lemma 3.2, and so we omit the proof here. For the proof of the second estimate we refer the reader to [20].

LEMMA 3.3. *For $u_i \in X_h(\Omega_i)$, we have*

$$\| J_m u_i - Q_m J_m u_i \|_{L^2(\gamma_{m(i)})} \leq c h_j^{1/2} |u_i|_{H_h^1(\Omega_i)}. \tag{3.5}$$

Proof. We partition γ_m into a collection of nonoverlapping subintervals or edge segments, $\mathcal{E}_o$, which are intersections between the edges from the mortar side and the edges from the nonmortar side. For simplicity we assume that each $\mathcal{E}_o$, shown as a thick line segment in Figures 3.2–3.3 (left pictures), corresponds to a complete edge. For the case where $\mathcal{E}_o$ may not be a complete edge but rather an edge segment shown as a thick line segment in Figures 3.2–3.3 (right pictures), the analyses are similar and will contribute only to the constant c. For the proof of the present lemma, we consider separately the two different cases of the mesh sizes: the mesh on the mortar side is finer than the mesh on the nonmortar side (Case A), and vice versa (Case B). The general case where the mesh on the mortar side may be finer than the mesh on the nonmortar side at some region along the interface and coarser in the other region can be formed as a combination of the two cases.

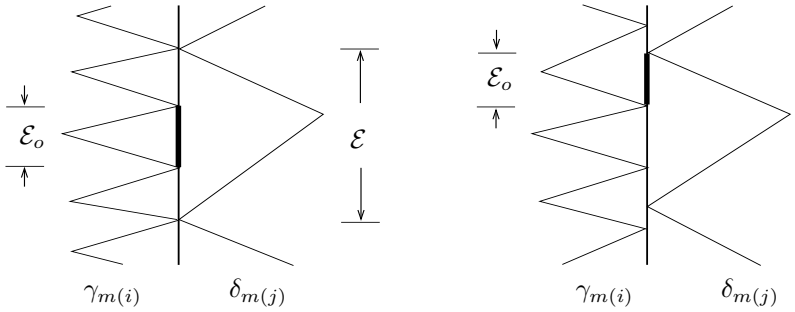


FIG. 3.2. *Illustrating $h_i < h_j$.*

Case A. Let $\mathcal{E} \subset \delta_{m(j)}$ and $\mathcal{E}_o \subset \gamma_{m(i)}$, where $\mathcal{E}_o \subset \mathcal{E}$ such that $h_{\mathcal{E}_o} \leq h_{\mathcal{E}}$; cf. Figure 3.2. Now, by the triangle inequality and the L^2 stability of the projection operator Q_m, we can write

$$\begin{aligned} &\| J_m u_i - Q_m J_m u_i \|^2_{L^2(\gamma_{m(i)})} \\ &\quad \leq c \Big\{ \| J_m u_i - u_i \|^2_{L^2(\gamma_{m(i)})} + \| u_i - Q_m u_i \|^2_{L^2(\gamma_{m(i)})} \\ &\qquad + \| Q_m(u_i - J_m u_i) \|^2_{L^2(\gamma_{m(i)})} \Big\} \\ &\quad \leq c \Big\{ \sum_{\mathcal{E} \subset \delta_{m(j)}} \| u_i - J_m u_i \|^2_{L^2(\mathcal{E})} + \| u_i - Q_m u_i \|^2_{L^2(\gamma_{m(i)})} \Big\}. \end{aligned} \tag{3.6}$$

Owing to the same L^2 stability, the second term above can further be estimated as the following:

$$\| u_i - Q_m u_i \|^2_{L^2(\gamma_{m(i)})}$$

$$\leq c\Big\{ \sum_{\mathcal{E}\subset\delta_{m(j)}} \| u_i - \mathcal{M}_i u_i \|^2_{L^2(\mathcal{E})} + \| \mathcal{M}_i u_i - Q_m \mathcal{M}_i u_i \|^2_{L^2(\gamma_{m(i)})} \Big\}.$$

Using Lemma 3.2 and the first estimate of (3.4) for each edge $\mathcal{E}_o \subset \mathcal{E}$, adding their contributions, and finally noting that $h_{\mathcal{E}_o} \leq h_{\mathcal{E}}$, the terms $\| u_i - J_m u_i \|^2_{L^2(\mathcal{E})}$ and $\| u_i - \mathcal{M}_i u_i \|^2_{L^2(\mathcal{E})}$ can be estimated as $ch_j \sum_{\tau\in B_{\mathcal{E}}} |u_i|^2_{H^1(\tau)}$ and $ch_j \sum_{\tau\in \hat{B}_{\mathcal{E}}} |u_i|^2_{H^1(\tau)}$, respectively. Now summing over $\mathcal{E} \subset \delta_{m(j)}$, in each case, we get the bound equal to $ch_j|u_i|^2_{H^1_h(\Omega_i)}$. What remains now is to estimate the term $\| \mathcal{M}_i u_i - Q_m \mathcal{M}_i u_i \|_{L^2(\gamma_{m(i)})}$. To do so, we first note that (cf. [1])

$$\| \mathcal{M}_i u_i - Q_m \mathcal{M}_i u_i \|^2_{L^2(\mathcal{E})} \leq ch_j |\mathcal{M}_i u_i|^2_{H^{\frac{1}{2}}(\mathcal{E})}.$$

Now, by adding together the estimates over $\mathcal{E} \subset \delta_{m(j)}$, using the trace inequality $|\mathcal{M}_i u_i|_{H^{\frac{1}{2}}(\gamma_{m(i)})} \leq c|\mathcal{M}_i u_i|_{H^1(\Omega_i)}$, and finally applying the second estimate of (3.4) we again get the bound $ch_j|u_i|^2_{H^1_h(\Omega_i)}$. Replacing the three estimates into (3.6) proves the lemma for the present case.

Case B. Let $\mathcal{E} \subset \gamma_{m(i)}$ and $\mathcal{E}_o \subset \delta_{m(j)}$, and $\mathcal{E}_o \subset \mathcal{E}$ such that $h_{\mathcal{E}_o} \leq h_{\mathcal{E}}$; cf. Figure 3.3. We note that the function $J_m u_i - Q_m J_m u_i$ is linear over $\mathcal{E}_o$. Now,

$$\begin{aligned}
&\| J_m u_i - Q_m J_m u_i \|^2_{L^2(\delta_{m(j)})} \\
&\quad = \sum_{\mathcal{E}_o\subset\delta_{m(j)}} \| J_m u_i - Q_m J_m u_i \|^2_{L^2(\mathcal{E}_o)} \\
&\quad \leq c \sum_{\mathcal{E}_o\subset\delta_{m(j)}} h_{\mathcal{E}_o} \sum_{x=x_l^{\mathcal{E}_o}, x_m^{\mathcal{E}_o}, x_r^{\mathcal{E}_o}} (J_m u_i(x) - Q_m J_m u_i(x))^2.
\end{aligned}$$

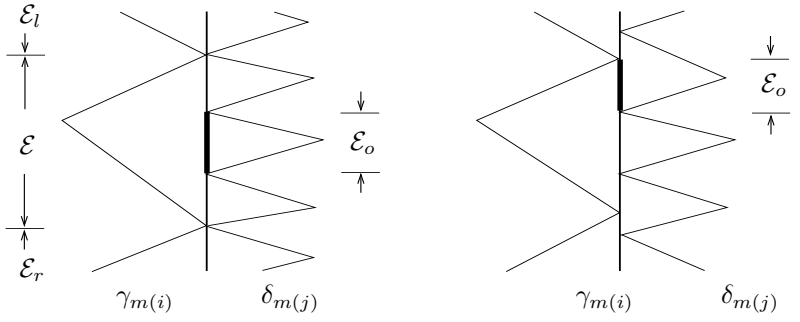


FIG. 3.3. *Illustrating* $h_j < h_i$.

On the edge $\mathcal{E}_o$, $Q_m J_m u_i$ is a constant equal to $J_m u_i(x_m^{\mathcal{E}_o})$. The difference $(J_m u_i(x) - Q_m J_m u_i(x))$ is zero at $x = x_m^{\mathcal{E}_o}$, and at $x = x_l^{\mathcal{E}_o}$ and $x_r^{\mathcal{E}_o}$ it can be written as

$$\begin{aligned}
J_m u_i(x_l^{\mathcal{E}_o}) - Q_m J_m u_i(x_l^{\mathcal{E}_o}) &= -\left(J_m u_i(x_r^{\mathcal{E}_o}) - Q_m J_m u_i(x_r^{\mathcal{E}_o})\right) \\
&= -\left\{ \frac{1}{h_{\mathcal{E}}} \left(I_m u_i(x_r^{\mathcal{E}}) - I_m u_i(x_l^{\mathcal{E}})\right) \right\} \frac{h_{\mathcal{E}_o}}{2},
\end{aligned}$$

where the term inside the curly brackets represents the slope of $J_m u_i$ or, equivalently, of $I_m u_i$, along the mortar. Hence, for $x = x_l^{\mathcal{E}_o}$ and $x_r^{\mathcal{E}_o}$, it follows from (3.2) that

$$(J_m u_i(x) - Q_m J_m u_i(x))^2 \le c \left(\frac{h_{\mathcal{E}_o}}{h_{\mathcal{E}}}\right)^2 \left(I_m u_i(x_r^{\mathcal{E}}) - I_m u_i(x_l^{\mathcal{E}})\right)^2$$

$$\le c \left(\frac{h_j}{h_i}\right)^2 \sum_{\tau \subset B_{\mathcal{E}}} |u_i|^2_{H^1(\tau)},$$

where the sum is taken over the set $B_{\mathcal{E}}$ of triangles along the mortar side, those touching the edge $\mathcal{E}$. The number of occurrences of the above sum in the final calculation is $\approx h_i/h_j$. This results in an h_j in the final calculation, since $h_{\mathcal{E}_o}(\frac{h_j}{h_i})^2 \frac{h_i}{h_j} \le h_j$ due to $\frac{h_j}{h_i} \le 1$. Hence, we arrive at

$$\| J_m u_i - Q_m J_m u_i \|^2_{L^2(\gamma_{m(i)})} \le c h_j |u_i|^2_{H^1_h(\Omega_i)}.$$

The lemma thus holds for both cases. For the general case where both Cases A and B exist along the interface $\gamma_{m(i)} = \delta_{m(j)}$, we simply need to take an appropriate combination of the estimates, that is, the estimates on $\mathcal{E}$ from Case A and the estimates on $\mathcal{E}_o$ from Case B, to show that the lemma holds. ∎

In the following, we briefly introduce some special operators from [15, 20] which we will need for our analysis. Let $\Pi_m : L^2(\delta_m(j)) \to Y_h^0(\delta_{m(j)})$, where $Y_h^0(\delta_{m(j)})$ is the set of continuous and piecewise linear functions on $\delta_{m(j)}$, defined uniquely by their values at $x \in \delta^{CR}_{m(j)h}$, and taking zero values at the boundary. $\Pi_m u$ on $\delta_{m(j)}$ is defined as follows (cf. [15]):

$$\Pi_m u(x) = Q_m u(x) \quad \forall x \in \delta^{CR}_{m(j)h}.$$

From the definition of Q_m (cf. (2.3)), it follows that $Q_m(Q_m u)(x) = (Q_m u)(x)$ for $x \in \delta^{CR}_{m(j)h}$, giving $(\Pi_m(Q_m u))(x) = (Q_m(Q_m u))(x) = Q_m u(x) = \Pi_m u(x)$ for all $x \in \delta^{CR}_{m(j)h}$. Hence, on $\overline{\delta}_m(j)$,

$$\Pi_m u = \Pi_m(Q_m u). \tag{3.7}$$

The stability of Π_m is stated in the following lemma; see [15] for proof.

LEMMA 3.4. *For* $u \in H^{\frac{1}{2}}_{oo}(\delta_{m(j)})$,

$$\| \Pi_m u \|_{H^{\frac{1}{2}}_{oo}(\delta_{m(j)})} \le c \| u \|_{H^{\frac{1}{2}}_{oo}(\delta_{m(j)})},$$

and for $u \in L^2(\delta_{m(j)})$,

$$\| \Pi_m u \|_{L^2(\delta_{m(j)})} \le c \| u \|_{L^2(\delta_{m(j)})}.$$

In the following, we define the two discrete harmonic extension operators, $\mathcal{H}_h^{CR} : X_h(\partial\Omega_j) \to X_h(\Omega_j)$ and $\mathcal{H}_{\frac{h}{2}} : Y_{\frac{h}{2}}(\partial\Omega_j) \to Y_{\frac{h}{2}}(\Omega_j)$, corresponding to the nonconforming P1 functions and the conforming P1 functions, respectively. For $u \in X_h(\partial\Omega_j)$, $\mathcal{H}_h^{CR} u \in X_h(\Omega_j)$ is the solution of

$$a_h(\mathcal{H}_h^{CR} u, v) = 0 \quad \forall v \in X_h^0(\Omega_j), \tag{3.8}$$

where $\mathcal{H}_h^{CR} u(x) = u(x)$ for $x \in \partial\Omega^{CR}_{jh}$. Similarly, for $u \in Y_{\frac{h}{2}}(\partial\Omega_j)$, $\mathcal{H}^{CR}_{\frac{h}{2}} u \in X_h(\Omega_j)$ is the solution of

$$a_j(\mathcal{H}_{\frac{h}{2}} u, v) = 0 \quad \forall v \in Y^0_{\frac{h}{2}}(\Omega_j), \tag{3.9}$$

where $\mathcal{H}_{\frac{h}{2}} u(x) = u(x)$ for $x \in \partial\Omega_{jh}$.

We state another useful operator, $\mathcal{M}_j^+ : Y_{\frac{h}{2}}(\Omega_j) \to X_h(\Omega_j)$ (cf. [20]), which is defined as follows: for $u \in Y_{\frac{h}{2}}(\Omega_j)$,

$$\mathcal{M}_j^+ u(x) = u(x) \quad \forall x \in \Omega_{jh}^{CR}.$$

For any $u \in Y_{\frac{h}{2}}(\Omega_j)$, the following holds (cf. [20]):

$$|\mathcal{M}_j^+ u|_{H_h^1(\Omega_j)} \le c|u|_{H_h^1(\Omega_j)}. \tag{3.10}$$

In the next lemma we estimate the H_h^1-seminorm over a subdomain by the $H_{oo}^{\frac{1}{2}}$-norm on a nonmortar side of the subdomain.

LEMMA 3.5. *Let $u \in X_h(\Omega_j)$ be a discrete harmonic function in Ω_j, in the sense of* (3.8)*, with $u = 0$ at $x \in \partial\Omega_{jh}^{CR} \setminus \delta_{m(j)h}^{CR}$. Then*

$$|u|_{H_h^1(\Omega_j)} \le c \parallel \Pi_m u \parallel_{H_{oo}^{\frac{1}{2}}(\delta_{m(j)})}. \tag{3.11}$$

Proof.

$$\begin{aligned} |u|_{H_h^1(\Omega_j)} &\le c|\mathcal{M}_j^+ \mathcal{H}_{\frac{h}{2}} E_m \Pi_m u|_{H_h^1(\Omega_j)} \\ &\le c|\mathcal{H}_{\frac{h}{2}} E_m \Pi_m u|_{H_h^1(\Omega_j)} \\ &\le c \parallel \Pi_m u \parallel_{H_{oo}^{\frac{1}{2}}(\delta_{m(j)})}. \end{aligned}$$

The first inequality follows from the fact that u is itself discrete harmonic and therefore has the minimal energy over all functions whose boundary values at $x \in \partial\Omega_{jh}^{CR}$ are the same as those of u. The second inequality follows from (3.10), and the third inequality follows from the property of a conforming discrete harmonic function; cf. [7]. E_m is the zero extension operator from $L^2(\delta_{m(j)})$ to $L^2(\partial\Omega_j)$. ∎

THEOREM 3.6 (error). *Let $u \in H^2(\Omega)$ and $u_h \in V_h$ be the exact solution of* (2.1) *and* (2.4)*, respectively. Then*

$$|u - u_h|_{H_h^1(\Omega)} \le c \left(\sum_{k=1}^{N} h_k^2 |u|_{H^2(\Omega_k)}^2 \right)^{\frac{1}{2}}. \tag{3.12}$$

The proof follows from the second Strang lemma (cf. [12]),

$$\begin{aligned} |u - u_h|_{H_h^1(\Omega)} \le c \inf_{v \in V_h} |u - v|_{H_h^1(\Omega)} \\ + \sup_{w \in V_h} \sum_{k=1}^{N} \sum_{\tau \in \mathcal{T}_h(\Omega_k)} \int_{\partial\tau} \frac{\partial u}{\partial \eta} \frac{w}{|w|}_{H_h^1(\Omega)} ds, \end{aligned} \tag{3.13}$$

and Lemmas 3.7 and 3.8.

LEMMA 3.7 (consistency error). *Let $u \in H^2(\Omega)$ be the exact solution of* (2.1)*. Then for any $w = \{w_k\} \in V_h$,*

$$\sum_{k=1}^{N} \sum_{\tau \in \mathcal{T}_h(\Omega_k)} \int_{\partial\tau} \frac{\partial u}{\partial \eta} \frac{w}{|w|}_{H_h^1(\Omega)} ds \le c \left(\sum_{k=1}^{N} h_k^2 |u|_{H^2(\Omega_k)}^2 \right)^{1/2}. \tag{3.14}$$

Proof. Note that

$$\sum_{k=1}^{N} \sum_{\tau \in \mathcal{T}_h(\Omega_k)} \int_{\partial\tau} \frac{\partial u}{\partial \eta} w_k \, ds = \sum_{\tau \in \mathcal{T}_h(\Omega_k),\ \mathcal{E} \subset (\partial\tau \setminus \mathcal{S})} \int_{\mathcal{E}} \frac{\partial u}{\partial \eta} [w] \, ds + \sum_{\Gamma_{ij} \subset \mathcal{S}} \int_{\Gamma_{ij}} \frac{\partial u}{\partial \eta} [w] \, ds.$$

Using Lemmas 8.3.7 and 8.3.9 of [12], for the interior and boundary edges, respectively, one can estimate the first term as follows:

$$\sum_{\tau \in \mathcal{T}_h(\Omega_k),\ \mathcal{E} \subset (\partial\tau \setminus \mathcal{S})} \int_{\mathcal{E}} \frac{\partial u}{\partial \eta} [w] \, ds \leq c \left(\sum_{k=1}^{N} h_k^2 |u|^2_{H^2(\Omega_k)} \right)^{1/2} |w|_{H_h^1(\Omega)}.$$

In the following, we estimate the second term.

Let $[w] = w_i - w_j$ be the jump across Γ_{ij}, where $\gamma_{m(i)} = \Gamma_{ij} = \delta_{m(j)}$. Let $\alpha = \{\alpha_{\mathcal{E}}\}_{\mathcal{E} \subset \gamma_{m(i)}}$ be the piecewise constant function on $\gamma_{m(i)}$, where $\alpha_{\mathcal{E}} = \frac{1}{|\mathcal{E}|} \int_{\mathcal{E}} \frac{\partial u}{\partial \eta_{\mathcal{E}}} \, d\sigma$ for each triangle edge $\mathcal{E} \subset \gamma_{m(i)}$. Similarly, we define the piecewise constant function on $\delta_{m(j)}$ as $\beta = \{\beta_{\mathcal{E}}\}_{\mathcal{E} \subset \delta_{m(j)}}$, where $\beta_{\mathcal{E}} = \frac{1}{|\mathcal{E}|} \int_{\mathcal{E}} \frac{\partial u}{\partial \eta_{\mathcal{E}}} \, d\sigma$ for each triangle edge $\mathcal{E} \subset \delta_{m(j)}$. We get immediately the two relations

$$\int_{\gamma_{m(i)}} \alpha (w_i - J_m w_i) \, d\sigma = 0$$

and

$$\int_{\delta_{m(j)}} \beta (J_m w_i - w_j) \, d\sigma = 0.$$

The first one follows from the fact that $\frac{1}{|\mathcal{E}|} \int_{\mathcal{E}} J_m w_i \, dx = \frac{1}{|\mathcal{E}|} \int_{\mathcal{E}} w_i \, dx$ by construction, for each $\mathcal{E} \subset \gamma_{m(i)}$, and the second one is due to the mortar condition, giving, for each $\mathcal{E} \subset \delta_{m(j)}$, that $w_j(x_m^{\mathcal{E}}) = \frac{1}{|\mathcal{E}|} \int_{\mathcal{E}} J_m w_i \, dx$. Now, using these relations, we get

$$\begin{aligned}
\int_{\Gamma_{ij}} \frac{\partial u}{\partial \eta} [w] \, d\sigma &\leq \left| \int_{\gamma_{m(i)}} \left(\frac{\partial u}{\partial \eta} - \alpha \right) (w_i - J_m w_i) \, d\sigma \right| \\
&\quad + \left| \int_{\delta_{m(j)}} \left(\frac{\partial u}{\partial \eta} - \beta \right) (J_m w_i - w_j) \, d\sigma \right| \\
&\leq \left\| \frac{\partial u}{\partial \eta} - \alpha \right\|_{L^2(\gamma_{m(i)})} \| w_i - J_m w_i \|_{L^2(\gamma_{m(i)})} \\
&\quad + \left\| \frac{\partial u}{\partial \eta} - \beta \right\|_{L^2(\delta_{m(j)})} \| J_m w_i - w_j \|_{L^2(\delta_{m(j)})}.
\end{aligned}$$

We have

$$\left\| \frac{\partial u}{\partial \eta} - \alpha \right\|^2_{L^2(\gamma_{m(i)})} = \sum_{\mathcal{E} \subset \gamma_{m(i)}} \left\| \frac{\partial u}{\partial \eta} - \alpha_{\mathcal{E}} \right\|^2_{L^2(\mathcal{E})}$$

$$\leq c \sum_{\mathcal{E} \subset \gamma_{m(i)}} h_{\mathcal{E}} \left| \frac{\partial u}{\partial \eta_{\mathcal{E}}} \right|^2_{H^1(\tau_{\mathcal{E}})}$$

$$\leq c h_i |u|^2_{H^2(\Omega_i)},$$

where the last sum is taken over the triangles $\tau_{\mathcal{E}} \in \mathcal{T}_h(\Omega_i)$—those having $\mathcal{E} \subset \gamma_{m(i)}$ as one of their edges. Similarly,

$$\left\| \frac{\partial u}{\partial \eta} - \beta \right\|^2_{L^2(\delta_{m(j)})} \leq c h_j |u|^2_{H^2(\Omega_j)}.$$

Also, it follows from Lemma 3.2 that

$$\|w_i - J_m w_i\|^2_{L^2(\gamma_{m(i)})} \leq c h_i |w_i|^2_{H^1_h(\Omega_i)}.$$

Now, using the mortar condition $Q_m J_m w_i = Q_m w_j$ and Lemma 3.3, we get

$$\begin{aligned} \| J_m w_i - w_j \|^2_{L^2(\delta_{m(j)})} &\leq 2 \| J_m w_i - Q_m J_m w_i \|^2_{L^2(\delta_{m(j)})} \\ &\quad + 2 \| Q_m w_j - w_j \|^2_{L^2(\delta_{m(j)})} \\ &\leq c \left(h_j |w_i|^2_{H^1_h(\Omega_i)} + h_j |w_j|^2_{H^1_h(\Omega_j)} \right). \end{aligned}$$

Finally,

$$\begin{aligned} \int_{\Gamma_{ij}} \frac{\partial u}{\partial \eta} [w] \, d\sigma &\leq c \left[h_i |u|_{H^2(\Omega_i)} |w_i|_{H^1(\Omega_i)} \right. \\ &\quad \left. + h_j |u|_{H^2(\Omega_j)} (|w_i|_{H^1_h(\Omega_i)} + |w_j|_{H^1_h(\Omega_j)}) \right]. \end{aligned}$$

Now, summing over the interfaces $\Gamma_{ij} \subset \mathcal{S}$, and using the Cauchy–Schwarz inequality, $|\sum_i (a_i b_i)| \leq (\sum_i a_i^2)^{\frac{1}{2}} (\sum_i b_i^2)^{\frac{1}{2}}$, we get the proof. ∎

LEMMA 3.8 (approximation error). *For any function $u \in H^1_0(\Omega)$, with $u_{|\Omega_k} \in H^2(\Omega_k)$, we have*

$$\inf_{v \in V_h} |u - v|_{H^1_h(\Omega)} \leq c \left(\sum_{k=1}^{N} h_k^2 |u|^2_{H^2(\Omega_k)} \right)^{1/2}. \tag{3.15}$$

Proof. Let I_h be the conforming P1-interpolant of u on $\mathcal{T}_h(\Omega_k)$ for $k = 1, \ldots N$, and define $\tilde{v} = I_h u = \{\tilde{v}_k\}_{k=1,\ldots,N} \in X_h$. We note that $\tilde{v}$ does not satisfy the mortar condition, and so $\tilde{v} \notin V_h$. Let $v = \tilde{v} + w$, where $w = \{w_k\}_{k=1,\ldots,N} \in X_h$, and $w_k = \sum_{\delta_{m(k)} \subset \partial\Omega_k} w_{m(k)} \in X_h(\Omega_k)$. The function $w_{m(k)} \in X_h(\Omega_k)$ is defined as follows:

$$w_{m(k)}(x) = \begin{cases} Q_m(J_m \tilde{v}_l - \tilde{v}_k)(x), & x \in \delta^{CR}_{m(k)h}, \\ 0, & x \in \partial\Omega^{CR}_{kh} \setminus \delta^{CR}_{m(k)h}, \\ \text{discrete harmonic (cf. 3.8)}, & x \in \Omega^{CR}_{kh}. \end{cases}$$

It is easy to see that $v \in V_h$:

$$\begin{aligned}|u-v|^2_{H^1_h(\Omega_k)} &\le |u-\tilde{v}_k|^2_{H^1(\Omega_k)} + |w_k|^2_{H^1_h(\Omega_k)} \\ &\le ch_k^2|u|^2_{H^2(\Omega_k)} + \sum_{\delta_{m(k)}\subset\partial\Omega_k} |w_{m(k)}|^2_{H^1_h(\Omega_k)}.\end{aligned}$$

Here we have used $|u - I_h u|_{H^1(\Omega_k)} \le ch_k|u|_{H^2(\Omega_k)}$ from the Bramble–Hilbert lemma. Now using Lemma 3.5 and (3.7), we have

$$\begin{aligned}&|w_{m(k)}|^2_{H^1_h(\Omega_k)} \\ &\le \| \Pi_m(J_m\tilde{v}_l - \tilde{v}_k) \|^2_{H^{\frac{1}{2}}_{oo}(\delta_{m(k)})} \\ &\le \| \Pi_m(J_m\tilde{v}_l - \tilde{v}_l) \|^2_{H^{\frac{1}{2}}_{oo}(\delta_{m(k)})} + \| \Pi_m(\tilde{v}_l - \tilde{v}_k) \|^2_{H^{\frac{1}{2}}_{oo}(\delta_{m(k)})}.\end{aligned}$$

From [15], we can bound the second term as

$$\| \Pi_m(\tilde{v}_l - \tilde{v}_k) \|^2_{H^{\frac{1}{2}}_{oo}(\delta_{m(k)})} \le c\left(h_l^2|u|^2_{H^2(\Omega_l)} + h_k^2|u|^2_{H^2(\Omega_k)}\right).$$

For the second term we use a trace inequality to have

$$\begin{aligned}\| \Pi_m(J_m\tilde{v}_l - \tilde{v}_l) \|^2_{H^{\frac{1}{2}}_{oo}(\delta_{m(k)})} &\le ch_l^{-1} \| \Pi_m(J_m\tilde{v}_l - \tilde{v}_l) \|^2_{L^2(\delta_{m(k)})} \\ &\le ch_l^{-1} \| J_m\tilde{v}_l - \tilde{v}_l \|^2_{L^2(\gamma_{m(l)})} \\ &= ch_l^{-1} \sum_{\mathcal{E}\subset\gamma_{m(l)}} \| J_m\tilde{v}_l - \tilde{v}_l \|^2_{L^2(\mathcal{E})} \\ &\le ch_l^{-1} \sum_{\mathcal{E}\subset\gamma_{m(l)}} h_l \sum_{\tau\in B_\mathcal{E}} |\tilde{v}_l|^2_{H^1(\tau)}.\end{aligned}$$

For each $\mathcal{E} \subset \gamma_{m(l)}$, define Q^2u inside $B_\mathcal{E}$, where Q^2u is the averaged Taylor polynomial of order 2 of u as defined in Chapter 4 of [12]. We note that $(J_m I_h u - I_h u) = J_m I_h(u - Q^2u) - I_h(u - Q^2u)$ on $\mathcal{E}$, and hence using the Bramble–Hilbert lemma, we have (cf. [12])

$$\begin{aligned}\| J_m\tilde{v}_l - \tilde{v}_l \|^2_{L^2(\mathcal{E})} &\le ch_l \sum_{\tau\in B_\mathcal{E}} |I_h(u - Q^2u)|^2_{H^1(\tau)} \\ &\le ch_l \sum_{\tau\in B_\mathcal{E}} |u - Q^2u|^2_{H^1(\tau)} \\ &\le ch_l^3 \sum_{\tau\in B_\mathcal{E}} |u|^2_{H^2(\tau)}.\end{aligned}$$

Hence,

$$\| \Pi_m(J_m\tilde{v}_l - \tilde{v}_l) \|^2_{H^{\frac{1}{2}}_{oo}(\delta_{m(k)})} \le ch_l^2|u|^2_{H^2(\Omega_l)}. \qquad \square$$

Theorem 3.6 gives an estimate for the error in the finite element solution in the broken H^1-seminorm. In the following we remark on the L^2-norm estimate of the error, which can be obtained using the same arguments as is in [11] for the conforming P1 finite element. Using (3.14), (3.15), and a duality technique in Lemma III.1.4 of [8], we can get the estimate

$$\|u - u_h\|_{L^2(\Omega)} \leq ch^2 |u|_{H^2(\Omega)},$$

where $h = \max_k h_k$. A careful manipulation of the estimates will yield the following estimate:

$$\|u - u_h\|_{L^2(\Omega)} \leq Ch \left(\sum_{k=1}^{N} h_k^2 |u|^2_{H^2(\Omega_k)^2} \right)^{\frac{1}{2}},$$

which explicitly includes the mesh sizes of the subdomains. However, the factor h appearing in the estimate is due to the use of the global regularity assumption: $|u|_{H^2(\Omega)} \leq c\|f\|_{L^2(\Omega)}$.

4. An additive Schwarz method. Recently, an additive Schwarz method for the CR mortar finite element, which uses the standard mortar condition, has been proposed in [19]. A P1 mortar finite element version of the method can be found in [6]. In this section, we propose a similar method for our discrete problem (2.4), which is formed as a natural extension of the original method from the standard mortar case to the new approximate mortar case. This is done by a simple adjustment in the definition of the subspaces involved in the decomposition of the discrete space, so that the subspaces adapt naturally to the new situation. In the standard mortar case, we recall from the definition of the subspaces that part of the subdomain interior nodes, that is, the set of edge midpoints lying closest to a mortar side, were treated as if they were on the mortar side; in other words, the mortar side became thicker. Note that due to the new mortar condition we do not require this arrangement.

So, following the general framework for additive Schwarz methods (cf. [22]), we can decompose V_h as $V_h = V^{\mathcal{S}} + V^0 + \sum_{i=1}^{N} V^i$. For $i = 1, \ldots, N$, V^i is the restriction of V_h to Ω_i, with functions vanishing at subdomain boundary edge midpoints $\partial\Omega_{ih}^{CR}$ as well as on the remaining subdomains.

$V^{\mathcal{S}}$ is the space of functions given by their values on the skeleton edge midpoints $\mathcal{S}_h^{CR} = \bigcup_{\gamma_m} \gamma_{mh}^{CR}$, $V^{\mathcal{S}} = \{v \in V_h : v(x) = 0,\ x \in \overline{\Omega}_h^{CR} \setminus \mathcal{S}_h^{CR}\}$. We assume that there are no corner triangles, that is, triangles having more than one edge on a subdomain boundary. It is then easy to see that the corresponding stiffness matrix is a block diagonal matrix with each block being associated to one mortar side only.

The coarse space V^0 is a special space having a dimension equal to the number of subdomains. It is defined using the function $\chi_i \in X_h(\Omega_i)$ associated with the subdomain Ω_i. χ_i is defined by its nodal values as $\chi_i(x) = 1/\sum_j \rho_j(x)$ at $x \in \overline{\Omega}_{ih}^{CR}$, where the sum is taken over the subdomains Ω_j to which x belongs, and $\rho_j = 1$ for all j. Note that the ρ_j's may represent physical parameters with jumps across interfaces; see [19]. V^0 is given as the span of its basis functions, $\Phi_i, i = 1, \ldots, N$,

i.e., $V^0 = span\{\Phi_i : i = 1, \ldots, N\}$, where Φ_i associated with Ω_i is defined as

$$\Phi_i(x) = \begin{cases} 1, & x \in \Omega_{ih}^{CR}, \\ \rho_i \chi_i(x), & x \in \gamma_{m(i)h}^{CR}, \\ \rho_i Q_m(J_m \chi_j)(x), & x \in \delta_{m(i)h}^{CR},\ \delta_{m(i)} = \gamma_{m(j)}, \\ \rho_i Q_m(J_m \chi_i)(x), & x \in \delta_{m(j)h}^{CR},\ \delta_{m(j)} = \gamma_{m(i)}, \\ \rho_i \chi_j(x), & x \in \gamma_{m(j)h}^{CR},\ \gamma_{m(j)} = \delta_{m(i)}, \\ 0, & x \in \partial\Omega_{ih}^{CR} \cap \partial\Omega, \end{cases} \tag{4.1}$$

and $\Phi_i(x) = 0$ at all other x in $\overline{\Omega}_h^{CR}$. We use exact bilinear forms for all our subproblems. The projection-like operators $T^i : V_h \to V^i$ are defined in the standard way; i.e., for $i \in \{\mathcal{S}, 0, \ldots, N\}$ and $u \in V_h$, $T^i u \in V^i$ is the solution of $a_h(T^i u, v) = a_h(u, v)$, $v \in V^i$. Let $T = T^{\mathcal{S}} + T^0 + T^1 + \cdots + T^N$. The problem (2.4) is now replaced by an equivalent system (cf. [22]),

$$Tu_h^* = g, \tag{4.2}$$

where $g = T^{\mathcal{S}} u_h^* + \sum_{i=0}^N T^i u_h^*$. Let c and C represent constants independent of the mesh sizes $h = \inf_i h_i$ and $H = \max_i H_i$. Then the following theorem holds.

THEOREM 4.1. *For all $u \in V_h$,*

$$c\frac{h}{H} a_h(u, u) \le a_h(Tu, u) \le C a_h(u, u). \tag{4.3}$$

The theorem can be shown in the same way as the proof in [19], which uses the general theory for Schwarz methods; cf. [22]. It follows from the theorem that the condition number of the operator T grows as $\frac{H}{h}$.

5. Implementation issues. The best way to understand how to implement the method is to look into the matrix representation of the method. When it comes to implementation, the mortar method of this paper differs from the one in [17] mainly in the mortar condition. We consider therefore only the mortar condition and present its matrix representation here. For the rest we refer to the matrix formulation section of [17].

For each mortar γ_m, let $\{\mathcal{E}_k : \mathcal{E}_k \in \mathcal{T}_h(\gamma_m)\}_{k=1,\ldots,n}$ and $\{\mathcal{E}_o : \mathcal{E}_o \in \mathcal{T}_h(\delta_m)\}_{o=1,\ldots,p}$ be the sets of n and p triangle edges along the mortar and the corresponding nonmortar sides, respectively. Let $\mathbf{J}_m$ be the matrix representation of the interpolation operator $J_m : X_h(\gamma_m) \to Z_h(\gamma_m)$ in the sense that if $\mathbf{u}_{\gamma_m}$ is the vector containing the values of the function $u \in X_h$ at the nodes of γ_{mh}^{CR}, that is, the set of edge midpoints of $\mathcal{T}_h(\gamma_m)$, then $\mathbf{J}_m \mathbf{u}_{\gamma_m}$ will return the values of the function $J_m u$ at the edge left and right endpoints $\overline{\gamma}_{mh}^{LR} = \{x : x = x_l^{\mathcal{E}_k}, x_r^{\mathcal{E}_k}, \mathcal{E}_k \in \mathcal{T}_h(\gamma_m)\}$. We note that $\overline{\gamma}_{mh}^{LR}$ contains edge endpoints which geometrically occupy the same space as those of $\overline{\gamma}_{mh}$. In the set $\overline{\gamma}_{mh}^{LR}$, each edge endpoint in the interior of γ_m occurs twice, once as the left endpoint of an edge and once as the right endpoint of the neighboring edge, or vice versa (the superscript "LR" stands for "left and right"). Let $\mathbf{I}_m$ be the matrix representation of the interpolation operator $I_m : X_h(\gamma_m) \to Y_{\frac{h}{2}}(\gamma_m)$ in the following sense. Let $\mathbf{u}_{\gamma_m}$ be the same vector as above; then $\mathbf{I}_m \mathbf{u}_{\gamma_m}$ will be the vector containing the values of the function $I_m u$ at the nodes of $\overline{\gamma}_{m\frac{h}{2}}$, that is, the set of edge endpoints of $\mathcal{T}_{\frac{h}{2}}(\gamma_m)$. Note that $\overline{\gamma}_{m\frac{h}{2}}$ is also the set of edge endpoints and edge midpoints of $\mathcal{T}_h(\gamma_m)$. Let h_k be

the length of the edge $\mathcal{E}_k$; then

$$\mathbf{I}_m = \begin{bmatrix} \frac{2h_1+h_2}{h_1+h_2} & -\frac{h_1}{h_1+h_2} & & & & & \\ 1 & & & & & & \\ \frac{h_2}{h_1+h_2} & \frac{h_1}{h_1+h_2} & & & & & \\ & 1 & & & & & \\ & \frac{h_3}{h_2+h_3} & * & & & & \\ & & * & & & & \\ & & * & \frac{h_{n-2}}{h_{n-2}+h_{n-1}} & & & \\ & & & 1 & & & \\ & & & \frac{h_n}{h_{n-1}+h_n} & \frac{h_{n-1}}{h_{n-1}+h_n} & & \\ & & & & 1 & & \\ & & & -\frac{h_n}{h_{n-1}+h_n} & \frac{h_{n-1}+2h_n}{h_{n-1}+h_n} & & \end{bmatrix}.$$

We follow Definition 2.2 and introduce an intermediate matrix $\mathbf{K}_m$ so that we have

$$\mathbf{J}_m = \mathbf{K}_m \mathbf{I}_m.$$

$\mathbf{K}_m$ is a block matrix consisting of n rectangular blocks, each corresponding to an edge $\mathcal{E}_k \in \mathcal{T}_h(\gamma_m)$ and being equal to the following 2×3 rectangular matrix:

$$\begin{bmatrix} \frac{1}{2} & 1 & -\frac{1}{2} \\ -\frac{1}{2} & 1 & \frac{1}{2} \end{bmatrix}.$$

The columns of this rectangular matrix correspond to the left endpoint $x_l^{\mathcal{E}_k}$, the midpoint $x_m^{\mathcal{E}_k}$, and the right endpoint $x_r^{\mathcal{E}_k}$ of the edge $\mathcal{E}_k$, respectively, and the rows correspond to the endpoints $x_l^{\mathcal{E}_k}$ and $x_r^{\mathcal{E}_k}$ of the edge $\mathcal{E}_k$, respectively. Consequently, the columns of $\mathbf{K}_m$ correspond to the set $\overline{\gamma}_{m\frac{h}{2}}$, while the rows correspond to the set $\overline{\gamma}_{mh}^{LR}$. We remark that the extra work involved in our algorithm, compared to the one presented in [17], is associated with the application of $\mathbf{K_m}$.

Let φ_k and φ_l be the standard CR basis functions associated with the edge midpoints $x_m^{\mathcal{E}_k} \in \gamma_{mh}^{CR}$ ($\mathcal{E}_k \in \mathcal{T}_h(\gamma_m)$) and $x_m^{\mathcal{E}_l} \in \delta_{mh}^{CR}$ ($\mathcal{E}_l \in \mathcal{T}_h(\delta_m)$), respectively. Let ψ_o be the basis function of the $M^h(\delta_m)$ associated with the edge midpoint $x_m^{\mathcal{E}_o}$ ($\mathcal{E}_o \in \mathcal{T}_h(\delta_m)$), defined as $\psi_o = 1$ in $\mathcal{E}_o$ and zero otherwise. Finally, let ξ_q be the basis function of $P^1(\mathcal{E}_q)$, associated with one of the edge endpoints $x_l^{\mathcal{E}_q}$ (left) and $x_r^{\mathcal{E}_q}$ (right).

Define the master matrix as $\mathbf{M}_{\gamma_m} = \left\{(J_m\varphi_k, \psi_o)_{L^2(\delta_m)}\right\}$ and the corresponding slave matrix as $\mathbf{S}_{\delta_m} = \left\{(\varphi_l, \psi_o)_{L^2(\delta_m)}\right\}$, where $x_m^{\mathcal{E}_k} \in \gamma_{mh}^{CR}$ and $x_m^{\mathcal{E}_l}, x_m^{\mathcal{E}_o} \in \delta_{mh}^{CR}$. Now, let $\mathbf{u}_{\gamma_m}$ be the vector defined as above, and let $\mathbf{u}_{\delta_m}$ be the corresponding vector containing the values of the function $u \in X_h$ at the edge midpoints δ_{mh}^{CR}; then

$$\mathbf{S}_{\delta_m}\mathbf{u}_{\delta_m} = \mathbf{M}_{\gamma_m}\mathbf{u}_{\gamma_m} = \mathbf{N}_{\gamma_m}\mathbf{K}_m\mathbf{I}_m\mathbf{u}_{\gamma_m}$$

is the matrix (discrete) representation of the mortar condition, where the supporting master matrix $\mathbf{N}_{\gamma_m} = \left\{(\xi_q, \psi_o)_{L^2(\delta_m)}\right\}$ with $x_l^{\mathcal{E}_q}, x_r^{\mathcal{E}_q} \in \overline{\gamma}_{mh}^{LR}$, and $x_m^{\mathcal{E}_o} \in \delta_{mh}^{CR}$. We note that $\mathbf{S}_{\delta_m}$ is a diagonal matrix containing the lengths of the triangle edges $\mathcal{E}_o \in \mathcal{T}_h(\delta_m)$ along the nonmortar side as entries.

6. Numerical results. The numerical results are presented here. We consider our model problem on a unit square domain with the function $f = 2\pi^2 \sin(\pi x)\sin(\pi y)$ and homogeneous Dirichlet boundary condition. The exact solution u equals the function $\sin(\pi x)\sin(\pi y)$. For all our experiments, as a general rule, we decompose the domain into a $d \times d = d^2$ number of square subdomains (subregions) and then uniformly triangulate each subdomain. In order to get nonmatching grids across all interfaces, each pair of subdomains sharing an interface are triangulated using a fixed $2m^2$ and $2n^2$ number of right angle triangles, where m is different from n. Note that the number d is inversely proportional to the subdomain size H, whereas the numbers m and n are proportional to $\frac{H}{h}$, the ratio between the subdomain size and the mesh size h. All sides of a subdomain are chosen to be either mortar or nonmortar. For comparison, we consider both the standard mortar condition of [15] and the new approximate mortar condition of the present paper for the lowest order CR finite element for solving the boundary value problem. The two resulting algebraic systems, which differ from each other due to different mortar conditions, are then solved using the conjugate gradient (CG) method.

TABLE 6.1

L^2-norm (error_{L^2}) and H^1_h-seminorm (error_{H^1}) of the error, using the standard and the approximate mortar conditions for the lower order CR finite element on 3×3 nonmatching grids.

		Standard CR Mortar		Approximate CR Mortar	
$\{m, n\}$	# dofs	error_{L^2}	$\text{error}_{H^1_h}$	error_{L^2}	error_{H1}
$\{07, 05\}$	1009	2.023e-3	5.815e-2	1.982e-3	5.858e-2
$\{14, 10\}$	4088	5.019e-4	2.905e-2	4.970e-4	2.916e-2
$\{28, 20\}$	16456	1.249e-4	1.454e-2	1.243e-4	1.456e-2
$\{56, 40\}$	66032	3.116e-5	7.272e-3	3.109e-5	7.278e-3

In our first experiment, we look at the accuracy of the computed solution u_h for the approximate mortar condition and compare it with that for the standard mortar condition. The L^2-norm and the H^1_h-seminorm of the error $u - u_h$ for different mesh sizes are calculated and shown in Table 6.1. As we can see from the table, the errors in both mortar cases are very close, and they vary as h^2 in the L^2-norm and as h in the H^1_h-seminorm, which is in accordance with the theory.

In our next experiment, we study the convergence behavior of their corresponding additive Schwarz methods; those are the method proposed for the approximate mortar condition in section 4 and the method proposed for the standard mortar condition in [19]. The additive Schwarz methods are used as preconditioners for the CG method for their respective algebraic systems. The condition number estimates and the number of iterations required to reduce the discrete L^2-norm of the residual by a factor of 10^{-6} are shown in the Table 6.2. As seen from the table, the condition number estimates as well as the iteration counts remain bounded for fixed $\frac{H}{h}$ ratio, which is as predicted by the theory. Again, the condition number estimates in both mortar cases are very close; however, the approximate mortar case seems to require slightly fewer iterations for the same tolerance.

We conclude by saying that both approaches, the standard mortar and the proposed approximate mortar, exhibit very similar numerical behavior. The approximate mortar approach, however, has the advantage that the mortar condition uses only the

TABLE 6.2
Condition number estimates and iteration counts using the standard and the approximate mortar conditions for the lower order CR finite element. $\frac{H}{h}$ *is fixed with* $m = 07$ *and* $n = 05$.

		Standard CR Mortar		Approximate CR Mortar	
$d \times d$	# dofs	κ_2	# iterations	κ_2	# iterations
03×03	1009	31.31	27	33.70	24
06×06	3984	35.47	41	35.85	39
12×12	16104	36.35	46	36.07	43
24×24	64752	36.58	48	36.14	43

nodal values on the interface, thereby making all its algorithms comparatively simpler and less intricate.

Acknowledgment. The authors would like to thank the anonymous referees for their valuable comments and suggestions to improve the paper.

REFERENCES

[1] F. B. BELGACEM AND Y. MADAY, *The mortar element method for three dimensional finite elements*, RAIRO Modél. Math. Nal. Numér., 31 (1997), pp. 289–302.
[2] F. BEN BELGACEM, *The Mortar Element Method with Lagrange Multipliers*, Université Paul Sabatier, Toulouse, France, 1994.
[3] C. BERNARDI, Y. MADAY, AND A. T. PATERA, *A new non conforming approach to domain decomposition: The mortar element method*, in Nonlinear Partial Differential Equations and Their Applications, Collége de France Seminar, H. Brezis and J.-L. Lions, eds., Longman Scientific and Technical, Harlow, UK, 1994, pp. 13–51.
[4] S. BERTOLUZZA AND S. FALLETTA, *The mortar method with approximate constraint*, in Proceedings of the Fourteenth International Conference on Domain Decomposition Methods, I. Herrera, D. E. Keyes, O. B. Widlund, and R. Yates, eds., National Autonomous University of Mexico, Mexico City, Mexico, 2003, pp. 357–364.
[5] S. BERTOLUZZA AND V. PERRIER, *The mortar method in the wavelet context*, M2AN Math. Model Numer. Anal., 35 (2001), pp. 647–674.
[6] P. BJØRSTAD, M. DRYJA, AND T. RAHMAN, *Additive Schwarz methods for elliptic mortar finite element problems*, Numer. Math., 95 (2003), pp. 427–457.
[7] P. BJØRSTAD AND O. B. WIDLUND, *Iterative methods for the solution of elliptic problems on regions partitioned into substructures*, SIAM J. Numer. Anal., 23 (1986), pp. 1097–1120.
[8] D. BRAESS, *Finite Elements: Theory, Fast Solvers, and Applications in Solid Mechanics*, 2nd ed., Cambridge University Press, Cambridge, UK, 2001.
[9] D. BRAESS AND W. DAHMEN, *Stability estimates of the mortar finite element method for* 3*-dimensional problems*, East-West J. Numer. Math., 6 (1998), pp. 249–264.
[10] D. BRAESS AND W. DAHMEN, *The mortar element method revisited—what are the right norms?*, in Proceedings of the Thirteenth International Conference on Domain Decomposition, N. Debit, M. Garbey, R. Hoppe, J. Périaux, D. Keyes, and Y. Kuznetsov, eds., Springer-Verlag, Berlin, 2001, pp. 245–252.
[11] D. BRAESS, W. DAHMEN, AND C. WIENERS, *A multigrid algorithm for the mortar finite element method*, SIAM J. Numer. Anal., 37 (1999), pp. 48–69.
[12] S. C. BRENNER AND L. R. SCOTT, *The Mathematical Theory of Finite Element Methods*, 2nd ed., Texts in Applied Mathematics 15, Springer-Verlag, New York, 2002.
[13] M. CROUZEIX AND P. A. RAVIART, *Conforming and nonconforming finite element methods for solving the stationary Stokes equations*, I, RAIRO Anal. Numer., 7 (1973), pp. 33–76.
[14] J.-L. LIONS AND E. MAGENES, *Non-Homogeneous Boundary Value Problems and Applications* I, Springer-Verlag, New York, 1972.
[15] L. MARCINKOWSKI, *The mortar element method with locally nonconforming elements*, BIT, 39 (1999), pp. 716–739.
[16] L. MARCINKOWSKI, *Additive Schwarz method for mortar discretization of elliptic problems with* P_1 *nonconforming finite element*, BIT, 45 (2005), pp. 375–394.

[17] T. Rahman and X. Xu, *A new variant of the mortar technique for the Crouzeix-Raviart finite element*, in Proceedings of the Sixteenth International Conference on Domain Decomposition Methods, D. Keyes and O. B. Widlund, eds., Springer-Verlag, New York, 2007, pp. 463–470.

[18] T. Rahman and X. Xu, *A Multilevel Preconditioner for the Mortar Method for Nonconforming $P1$ Finite Element*, BCCS technical report 16, Bergen Center for Computational Science, Bergen, Norway, 2006 (submitted).

[19] T. Rahman, X. Xu, and R. Hoppe, *Additive Schwarz methods for the Crouzeix-Raviart mortar finite element for elliptic problems with discontinuous coefficients*, Numer. Math., 101 (2005), pp. 551–572.

[20] M. Sarkis, *Nonstandard coarse spaces and Schwarz methods for elliptic problems with discontinuous coefficients using non-conforming element*, Numer. Math., 77 (1997), pp. 383–406.

[21] P. Seshaiyer and M. Suri, *Uniform hp convergence results for the mortar finite element method*, Math. Comput., 69 (1999), pp. 521–546.

[22] B. Smith, P. E. Bjørstad, and W. Gropp, *Domain Decomposition: Parallel Multilevel Methods for Elliptic Partial Differential Equations*, Cambridge University Press, Cambridge, UK, 1996.

[23] B. Wohlmuth, *A residual based error estimator for mortar finite element discretizations*, Numer. Math., 84 (1999), pp. 143–171.

[24] B. I. Wohlmuth, *A mortar finite element method using dual spaces for the Lagrange multiplier*, SIAM J. Numer. Anal., 38 (2000), pp. 989–1012.

[25] X. Xu and J. Chen, *Multigrid for the mortar element method for P_1 nonconforming element*, Numer. Math., 88 (2001), pp. 381–398.

SIAM J. NUMER. ANAL.
Vol. 46, No. 1, pp. 517–543
© 2008 Society for Industrial and Applied Mathematics

OPTIMAL PANEL-CLUSTERING IN THE PRESENCE OF ANISOTROPIC MESH REFINEMENT*

I. G. GRAHAM†, L. GRASEDYCK‡, W. HACKBUSCH‡, AND S. A. SAUTER§

Abstract. In this paper we consider the numerical solution of discrete boundary integral equations on polyhedral surfaces in three dimensions. When the solution contains typical edge singularities, highly stretched meshes are preferred to uniform meshes, since they reduce the number of degrees of freedom needed to obtain a fixed accuracy. The classical panel-clustering method can still be applied in the presence of such highly stretched meshes. However, we will show that the savings in computation time and storage become suboptimal because the nearfield matrix arising in the panel-clustering algorithm is no longer as sparse as it is in the case of uniform meshes. Hence, a natural question arises as to whether a new enhanced panel-clustering algorithm can be designed which performs efficiently even in the presence of highly stretched meshes. The main result of this paper is to formulate such an enhanced version of the panel-clustering algorithm. The key features of the algorithm are (i) the employment of partial analytic integration in the direction of stretching, yielding a new kernel function on a one-dimensional manifold where the influence of high aspect ratios in the stretched elements is removed, and (ii) the introduction of a generalized admissibility condition with respect to the partially integrated kernel, which ensures that certain stretched clusters which are inadmissible in the classical sense now become admissible. In the context of a model problem, we prove that our algorithm yields an accurate (up to the discretization error) matrix-vector multiplication which requires $\mathcal{O}(Nlog^k N)$ operations, where N is the number of degrees of freedom and k is small and independent of the aspect ratio. The generalized admissibility condition can be viewed as an addition to the classical method which may be useful in general when stretched meshes are present. We also have performed a numerical experiment which shows that the sparsity of the nearfield matrix for the enhanced panel-clustering method is not negatively affected by stretched elements, and the method will perform optimally.

Key words. boundary elements, panel-clustering, anisotropic meshes, matrix compression

AMS subject classifications. 65N12, 65N22, 65N38, 65R20

DOI. 10.1137/060677987

1. Introduction.

The boundary integral method is a well-known method for solving several important classes of boundary-value problems. Since boundary integral operators are nonlocal, their straightforward discretization via the boundary element method (BEM) leads to dense linear systems, which, for large applications, have to be solved iteratively. The storage requirement and the cost of a matrix-vector multiplication scale quadratically with the number of unknowns N, and this is a major bottleneck in the solution process.

The panel-clustering algorithm [19, 18, 20, 27, 28, 3, 4] employs separable polynomial expansions of the kernel function in the far field (i.e., for matrix entries corresponding to well separated pairs of freedoms) to obtain an approximate matrix for which the complexity of storage and matrix-vector multiplication is $O(N \log^\kappa N)$, for

*Received by the editors December 19, 2006; accepted for publication (in revised form) September 24, 2007; published electronically January 30, 2008. The first and fourth authors would like to thank the Max-Planck-Institut für Mathematik in den Naturwissenschaften, Leipzig and Universität Zürich for financial support.

http://www.siam.org/journals/sinum/46-1/67798.html

†Department of Mathematical Sciences, University of Bath, Bath BA2 7AY, United Kingdom (i.g.graham@bath.ac.uk).

‡Department of Scientific Computing, Max Planck Institute for Mathematics in the Sciences, Inselstr. 22-26, D-04103 Leipzig, Germany (lgr@mis.mpg.de, wh@mis.mpg.de).

§Institut für Mathematik, Universität Zürich, Winterthurerstr. 190, CH-8057 Zürich, Switzerland, (stas@amath.unizh.ch).

moderate κ, but for which the asymptotic accuracy of the resulting numerical solution (as the mesh is refined) remains unchanged.

To perform a thorough analysis of the panel-clustering algorithm we are required to address both the stability/consistency of the approximation and the complexity of the resulting matrix-vector multiplication. For quasi-uniform meshes this analysis can be found, e.g., in [19, 20, 29]. For more general classes of meshes, including certain degenerate meshes, a stability/consistency analysis was recently given in [11]. However, the complexity analysis has so far not been extended beyond the quasi-uniform case.

In this paper we focus on a class of degenerate meshes which are typically used in connection with the h-version of the Galerkin boundary element method on domains with nonsmooth boundaries. Because the solution being sought features anisotropic edge singularities, it can be shown (cf. [23, 8]) that meshes which are *algebraically graded* towards edges and corners allow a much better approximation of the solution than quasi-uniform meshes. These graded meshes are neither quasi-uniform nor shape-regular: Elements which are near edges but away from corners have an aspect ratio which tends to infinity as $N \to \infty$. Unfortunately, for such meshes, the standard panel-clustering algorithm will have suboptimal complexity, due to the fact that the near field becomes too large in the vicinity of the stretched elements. (This is explained in sections 5 and 7 below and was also the subject of previous unpublished work [10].) An alternative way to treat edge and corner singularities is the use of singular functions in the test and trial spaces of a Galerkin discretization which reflect the characteristic behavior of the solution [30, 22]. However, for complicated three-dimensional domains, the proper design of these functions, their implementation, and their numerical integration is far from trivial, and we will not investigate this method here.

In this paper we present a new version of the panel-clustering method which does not suffer from this defect. In order for the fundamental idea to be made clear, we will present the algorithm for general operators of order $> 1/2$ on polyhedra, discretised with piecewise constant elements. To keep the paper at reasonable length, the proofs will be limited to the case of the single layer potential for the Laplace operator. However, we also indicate throughout how the method and, in principle, the proofs can be extended to more general situations which are encountered in practical applications. The results of this paper complement those in [12], where the conditioning of Galerkin matrices on anisotropic meshes is analyzed, along with simple preconditioners which remove ill-conditioning due to the anisotropy. The fast multipole method (FMM) [26, 16] is closely related to the classical panel-clustering method. However, the generalization of the FMM to a version where the complexity is independent of an anisotropic mesh grading seems not to be straightforward and—to our best knowledge—not available in the literature. We remark that in addition to panel-clustering and FMM, other fast methods for realising discretizations of boundary integral operators can be based on Fourier methods; see, e.g., [1, 25].

The paper is organized as follows. In section 2, we briefly review the boundary integral equations, the graded meshes, and corresponding boundary element spaces having optimal approximation property. In section 3, we introduce the new (generalized) version of the panel-clustering method for stretched meshes and the algorithms for its efficient realization. The key idea is the introduction of a generalized admissibility condition which is based on the partial analytic integration of the kernel function on stretched elements in the direction parallel to the nearest edge. Here we make the (natural) assumption that the degenerate elements are parallel to the edge in

question. Section 4 is devoted to the error analysis of the generalized panel-clustering approximation. Specifically we show that the approximate system yields a solution which satisfies the same asymptotic error estimate as the true Galerkin solution. In section 5, we will show that the complexity of the new panel-clustering algorithm is of order $N \log^{\kappa} N$ for moderate κ. We will also show that the *standard* panel-clustering method for the stretched meshes considered here has complexity $N^{1+\delta}$, where $\delta \in (0, 1/2)$ and depends on the grading of the mesh. (In an extreme case, δ approaches $1/2$.) In section 6, explicit representations are presented for the partially integrated kernel function which is employed in section 3, along with the properties of the partially integrated kernel which are required for the proofs in sections 4 and 5. Numerical experiments are reported in section 7.

2. Boundary elements on anisotropically refined meshes. Let Γ be the surface of a closed bounded Lipschitz polyhedron in $\mathbb{R}^3$. Linear elliptic boundary-value problems with constant coefficients in the domain interior or exterior to Γ may be reformulated as boundary integral equations

$$(\lambda I + \mathcal{K})u(\mathbf{x}) := \lambda u(\mathbf{x}) + \int_{\Gamma} k(\mathbf{x}, \mathbf{y}) u(\mathbf{y}) ds_{\mathbf{y}} = f(\mathbf{x}), \quad \mathbf{x} \in \Gamma, \tag{2.1}$$

where $\lambda \in \mathbb{R}$, k is a known kernel function, and f is a known right-hand side. We will assume (as is true in very many cases) that for some $\mu \in \mathbb{R}$ the corresponding weak form,

$$\text{Find } u \in H^{\mu}(\Gamma) \quad \text{s.t. } a(u, v) := ((\lambda I + \mathcal{K})u, v) = (f, v) \quad \text{for all } v \in H^{\mu}(\Gamma), \tag{2.2}$$

constitutes a well-posed variational principle in the Sobolev space $H^{\mu}(\Gamma)$. (In some cases the energy space must be chosen as a suitable closed subspace of $H^{\mu}(\Gamma)$.) The continuous extension of the $L^2(\Gamma)$-scalar product $(\cdot, \cdot)$ to the duality pairing $H^{-\mu}(\Gamma) \times H^{\mu}(\Gamma)$ in (2.2) is again denoted by $(\cdot, \cdot)$. Typical examples include all the standard boundary integral formulations of Laplace's equation and the Helmholtz equation (see, e.g., [21, 29]). While the algorithm which we discuss in this paper is applicable to general boundary integral equations, we will present proofs for the following particular case.

Example 2.1. The Laplace single layer integral equation is of the form (2.1) with $\lambda = 0$ and $k(\mathbf{x}, \mathbf{y}) = 1/(4\pi \|\mathbf{x} - \mathbf{y}\|)$. The formulation (2.2) holds with $\mu = -1/2$, and the corresponding bilinear form is symmetric, bounded, and coercive in $H^{-1/2}(\Gamma)$.

To approximate (2.2) we assume that we have a mesh $\mathcal{T}$ of elements (here called *panels*) $\tau \in \mathcal{T}$. Each panel may be a triangle or quadrilateral and is assumed to be the image of a unit triangle or quadrilateral under an affine map. On this mesh we can introduce a space of piecewise polynomial functions $\mathcal{S} \subset H^{\mu}(\Gamma)$. The Galerkin method then seeks an approximate solution $U \in \mathcal{S}$ by solving

$$a(U, V) = (f, V) \qquad \text{for all } V \in \mathcal{S}. \tag{2.3}$$

Suppose $\mathbf{u}$ is the vector of coefficients of U with respect to a basis $\{\phi_i : i \in \mathcal{I}\}$, for some index set $\mathcal{I}$ with cardinality N. Then we have to solve the $N \times N$ linear system

$$(\lambda M + K)\mathbf{u} = \mathbf{f}, \tag{2.4}$$

with $M_{i,j} = (\phi_i, \phi_j)$, $K_{i,j} = (\mathcal{K}\phi_j, \phi_i)$, and $f_i = (f, \phi_i)$. The panel-clustering algorithm is used to approximate the dense matrix K. To be more precise, the panel-clustering is best thought of, *not* as an approximation of the *matrix entries* of K but as a data-sparse representation of the associated linear operator $K : \mathbb{R}^{\mathcal{I}} \to \mathbb{R}^{\mathcal{I}}$.

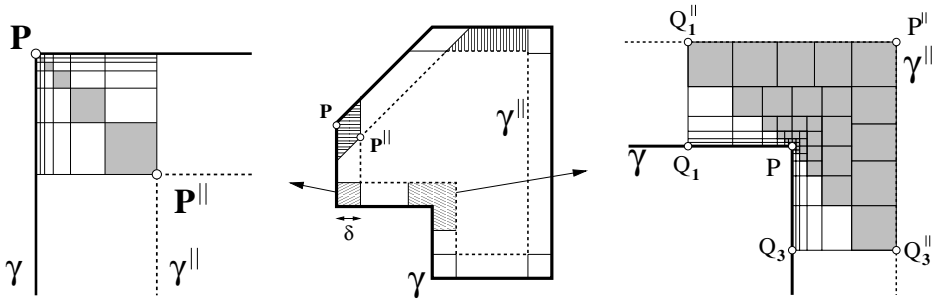


FIG. 2.1. *The face* Γ *(middle) is partitioned into segments with convex corners (left), concave corners (right), and the segments parallel to* γ *in between. Grey panels are shape-regular.*

It is well known (cf. [23, 29]) that the exact solution u exhibits a characteristic singular behavior in the *normal direction* near edges of Γ and in the *radial direction* near corners of Γ. Moreover, u is smooth in the interior of the plane faces of Γ and also in the *tangential* direction relative to the edges. This suggests that it may be advantageous to employ graded meshes whose elements have the following properties (as the meshes are refined): (i) they become increasingly small in the normal direction near the edges but are larger in the tangential direction, (ii) they become (uniformly) small when lying close to corners, and (iii) they are quasi-uniform when bounded away from the corners and edges.

We briefly describe the construction of suitable meshes. In general we are considering polyhedral surfaces, but without loss of generality we can explain the meshing algorithm by restricting to the case when Γ is a single planar polygon, with polygonal boundary $\gamma = \partial\Gamma$. For sufficiently small $\delta > 0$ we insert a polygon $\gamma^{\|}$ into Γ which is parallel to γ and has distance δ from γ (cf. Figure 2.1(middle)).[1] The portion of Γ between γ and $\gamma^{\|}$ we call "the δ-strip." We assume that δ is sufficiently small, so that for any vertex P in γ there exists a corresponding vertex $P^{\|}$ in $\gamma^{\|}$.

If P is a convex vertex of Γ, we extend both segments of $\gamma^{\|}$ which meet at $P^{\|}$ so that they hit γ.

If P is a concave vertex of Γ, we insert two lines into the δ-strip which are orthogonal to γ and a distance δ along γ from P.

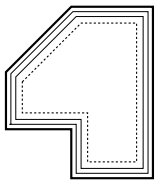
If δ is chosen sufficiently small, this procedure subdivides Γ into (i) rhombus-shaped regions of the δ-strip near convex corners, (ii) L-shaped regions of the δ-strip near concave corners, (iii) trapezia-shaped regions near edges but away from corners, and (iv) a polygonal inner region bounded by $\gamma^{\|}$.

The mesh on Γ depends on two parameters: $N \in \mathbb{N}$ (the number of elements in Γ) and $g \geq 1$ (the grading parameter which determines the concentration of panels towards edges). We shall study the convergence and complexity of our method as $N \to \infty$, with g fixed (although, in general, g may depend on the integral equation being solved, the polynomial degree of the elements, and the angles at the corners of Γ). Since some part of the mesh on Γ will be defined using tensor product constructions, we also introduce a coordinatewise discretization parameter $n \in \mathbb{N}$ such that $N \sim n^2$. The maximum mesh diameter will be proportional to n^{-1}.

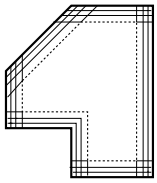
We will use the symbol "$\lesssim$" to compare two quantities $A \lesssim B$ if there exists a constant $C > 0$ which is independent of N such that $A \leq CB$.

[1] To be more precise, for each segment γ_i in γ there is a corresponding segment $\gamma_i^{\|}$ in $\gamma^{\|}$ so that the infinite lines through γ_i and $\gamma_i^{\|}$ are parallel and are a distance δ apart.

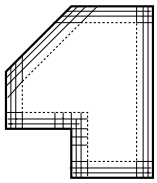
ALGORITHM 2.2 (construction of the mesh). *The meshing is applied to the four different types of region separately after the introduction of the δ-strip.*



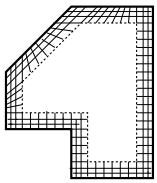
Step 1: *Edge-parallel mesh lines. For $n \in \mathbb{N}$, insert further polygons $\gamma_\ell^{\parallel}$, $1 \le \ell \le n-1$, into Γ which are parallel to γ and have distance $\delta\,(\ell/n)^g$ from γ. For any vertex P_ℓ of γ, there exists a corresponding vertex $P_\ell^{\parallel}$ in $\gamma_\ell^{\parallel}$.*



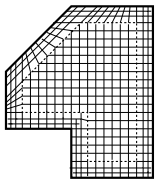
Step 2: *Rectangular meshes on rhombi near convex vertices. The straight line segments which meet at each $P_\ell^{\parallel}$ are prolonged until they hit γ. This produces a locally rectangular (non–shape-regular) mesh on each near-vertex rhombus which is graded towards the outer edges of the rhombus (cf. Figure* 2.1*(left)). The panels on the diagonal (towards the corner) are congruent to the near-vertex rhombus; i.e., they are shape-regular.*



Step 3: *Rectangular meshes on L-shaped regions at concave corners. We consider an L-shaped region as in Figure* 2.1*(right). In the two triangles $Q_1PQ_1^{\parallel}$ and $PQ_3Q_3^{\parallel}$ we use the same mesh as in a convex corner: graded towards the corner. Elements on the two lines $Q_1^{\parallel}P$ and $Q_3^{\parallel}P$ are then shape-regular. In the remaining part of the L-shaped region we use a shape-regular mesh that can be either conforming or nonconforming. We show the nonconforming case in Figure* 2.1*(right).*



Step 4: *Quadrilateral meshes on near-edge trapezia. Consider a near-edge trapezium with vertices $ABCD$ so that $\overline{AB} \subset \gamma$ and $\overline{CD} \subset \gamma^{\parallel}$. Subdivide $\overline{AB}$ into n equidistant segments and $\overline{CD}$ into n equidistant segments, and connect the opposite grid points in $\overline{AB}$ with those in $\overline{CD}$ by lines.*



Step 5: *Refinement of the inner region. The inner region is finally partitioned with a quasi-uniform mesh consisting of triangles and quadrilaterals of diameter $O\left(n^{-1}\right)$ such that the mesh points on $\gamma^{\parallel}$ already defined above coincide with the mesh points of the inner mesh at $\gamma^{\parallel}$. The quadrilaterals can be subdivided into triangles if required.*

Remark 2.3. The insertion of the lines as described in Algorithm 2.2, Step 4, subdivides the trapezium with vertices A, B, C, D into n smaller trapezia which we call *containers*. Each container Q consists of n parallel panels and has one and only one edge which lies in γ (namely, the edge $\overline{AB}$). The extension of this edge to an infinite line is called base line and denoted by γ_Q.

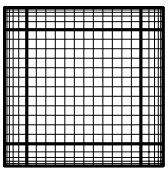
Since we will need it later, we now give the following explicit example.

Example 2.4. Let $\Gamma = [0,1] \times [0,1]$, and define the rectangular mesh:

$$\mathcal{T} := \{[p_{i-1}, p_i] \times [p_{j-1}, p_j] \ : \ i,j = 1, \dots, 3n\}, \text{ where}$$

$$p_i := \begin{cases} \delta(\frac{i}{n})^g, & 0 \le i \le n, \\ \delta + (1-2\delta)(\frac{i-n}{n}), & n \le i \le 2n, \\ 1 - \delta(\frac{3n-i}{n})^g, & 2n \le i \le 3n, \end{cases} \quad \text{and}$$

$$\delta := 1/(g+2).$$



(The inner square $\gamma^{\|}$ is denoted by the thick lines in the figure.)

Recalling that $N \sim n^2$, we see that near edges but away from corners we have degenerate elements with length $\sim N^{-1/2}$ and width $\sim \rho_\tau \sim N^{-g/2}$, i.e., aspect ratio growing with $N^{(g-1)/2}$ as $N \to \infty$. The reason for the specific choice of δ is that it ensures that adjacent elements on each side of any one of the thick lines have approximately the same sizes (as $N \to \infty$).

Now let $\mathcal{S}$ denote the space of piecewise constant functions with respect to the mesh on Γ. The analysis of the use of such graded meshes for approximating boundary integral equations on polyhedra can be found, for example, in [23, 24, 8, 9, 13]. In particular (see, e.g., [9, Theorem 1.4]) it is known that the Galerkin solution $U \in \mathcal{S}$ (see (2.3)) to the model problem in Example 2.1 on a Lipschitz polyhedral surface (for sufficiently smooth data f) satisfies

$$\|u - U\|_{H^{-1/2}(\Gamma)} \lesssim N^{-3/4} \qquad \text{when } g > 3. \tag{2.5}$$

Note that (2.5) constitutes an optimal estimate in terms of number of degrees of freedom N, since in the case of quasi-uniform meshes on smooth surfaces we expect $O(h^{3/2})$ convergence in the $H^{-1/2}$ norm, and the mesh diameter $h \sim N^{-1/2}$. Suboptimal convergence is attained for $1 \le g \le 3$. (This result also extends to some higher order elements [23].)

The central problem which we are concerned with in this paper is the construction of a data-sparse representation of the dense stiffness matrix K appearing in (2.4) which has optimal complexity but for which the convergence rates (2.5) are preserved.

We note in passing that as well as having a fast matrix-vector multiplication, the number of iterations required by an iterative method is of key importance for the speed of the method. It is known that anisotropic mesh refinement can produce considerable ill-conditioning. This can be substantially reduced by scaling the linear equations appropriately [12]. Additional preconditioning techniques involving approximate inverses such as the $\mathcal{H}$-matrix techniques [15, 14] or SPAI techniques [5] can be applied in addition.

3. Panel-clustering for graded meshes. The basic idea of the panel-clustering algorithm is to approximate the kernel k in (2.1) by a (short) separable expansion when $\mathbf{x}$ and $\mathbf{y}$ belong to sufficiently well separated sets. We briefly recall the essential ingredients of the panel-clustering method before we explain the modifications for stretched meshes.

From now on, the mesh $\mathcal{T}$ on Γ will be written $\mathcal{T} = \{\tau_i : i \in \mathcal{I}\}$, where $\mathcal{I}$ is a suitable index set. For each i, let ξ_i denote the incenter (center of the largest inscribed circle), let h_i denote the diameter, and let ρ_i denote the diameter of the largest inscribed circle in the panel τ_i.

3.1. Classical panel-clustering (isotropic admissibility).

DEFINITION 3.1 (cluster). *A cluster t is a subset of $\mathcal{I}$. If t is a cluster, the corresponding subdomain of Γ is $\Gamma_t := \bigcup_{i \in t} \tau_i$.*

The clusters are collected in a hierarchical cluster tree $T_{\mathcal{I}}$.

DEFINITION 3.2 (cluster tree). *A tree $T_{\mathcal{I}}$ is a cluster tree if the following conditions are satisfied:*

1. *The nodes in $T_{\mathcal{I}}$ are clusters.*
2. *The root of $T_{\mathcal{I}}$ is $\mathcal{I}$.*
3. *The leaves of $T_{\mathcal{I}}$ are denoted by $\mathcal{L}(T_{\mathcal{I}})$, and the tree hierarchy is given by a father/son relation: For each interior node $t \in T_{\mathcal{I}} \setminus \mathcal{L}(T_{\mathcal{I}})$, the set of sons of*

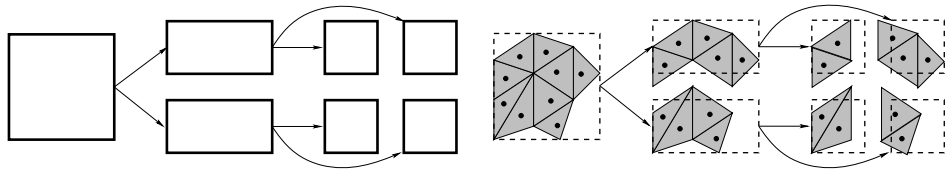


FIG. 3.1. *Left: Tree of regular boxes $\tilde{\mathcal{B}}$ obtained by bisection. Right: Tree of corresponding clusters $t = t(\tilde{\mathcal{B}})$. The black dots mark the corresponding incenters ξ_i of the panels $\tau_i \in \mathcal{T}$.*

t, $\mathrm{sons}(t)$, is the minimal subset in $\mathcal{T}_{\mathcal{I}} \setminus \{t\}$ such that

$$t = \dot{\bigcup_{s \in \mathrm{sons}(t)}} s$$

holds. Conversely, the father of any $s \in \mathrm{sons}(t)$ is t.

The standard (geometrically regular) construction of the cluster tree $T_{\mathcal{I}}$ is as follows. Choose a bounding box

$$\tilde{\mathcal{B}} := [a_1, b_1) \times [a_2, b_2) \times [a_3, b_3) \supset \Gamma.$$

Then subdivide $\tilde{\mathcal{B}}$ by bisecting a longest side, thus obtaining two sons $\tilde{\mathcal{B}}_1, \tilde{\mathcal{B}}_2$, e.g.,

$$\tilde{\mathcal{B}}_1 = [a_1, b_1') \times [a_2, b_2) \times [a_3, b_3), \quad \tilde{\mathcal{B}}_2 = [a_1', b_1) \times [a_2, b_2) \times [a_3, b_3),$$

where $a_1' = b_1' := (a_1 + b_1)/2$. Finally, applying the same bisection process recursively, starting with $\tilde{\mathcal{B}}_1$ and $\tilde{\mathcal{B}}_2$, yields an infinite tree T_{box} with root $\tilde{\mathcal{B}}$. Letting $\tilde{\mathcal{B}}_j$ denote a typical box in this tree, we can define a corresponding cluster by (cf. Figure 3.1)

$$t(\tilde{\mathcal{B}}_j) := \{i \in \mathcal{I} \mid \xi_i \in \tilde{\mathcal{B}}_j\}.$$

This yields an infinite cluster tree with root $t(\tilde{\mathcal{B}})$. We construct a finite cluster tree by deleting (not constructing) sons of clusters below a minimal cardinality $n_{\min}$, e.g., $n_{\min} := 8$. The cardinality of the cluster tree for the meshes introduced in section 2, as well as the complexity of its construction, will be estimated in section 5. (Note that for the special case that Γ is a flat polygon we may choose $\tilde{\mathcal{B}}$ as a two-dimensional bounding box.)

For a cluster, $t \subset \mathcal{I}$, let $\mathcal{B}_t$ denote the minimal bounding box (of the type from above) of Γ_t.

Conventionally, in order to obtain the approximation of the matrix K in (2.4), the kernel function k is approximated on $\Gamma_t \times \Gamma_s$, where (t, s) is a pair of clusters which satisfy the following condition.

DEFINITION 3.3 (isotropic admissibility condition). *Choose an isotropic admissibility parameter $\eta^{\mathrm{iso}} > 0$. A pair of clusters $(t, s) \in T_{\mathcal{I}} \times T_{\mathcal{I}}$ is isotropically (η^{iso}-) admissible if*

$$\max\{\mathrm{diam}\mathcal{B}_t,\ \mathrm{diam}\mathcal{B}_s\} \leq \eta^{\mathrm{iso}}\ \mathrm{dist}(\mathcal{B}_t,\ \mathcal{B}_s). \tag{3.1}$$

The standard choice for the parameter in (3.1) is $\eta^{\mathrm{iso}} := 3$.

The admissibility condition (3.1) allows us to define a nonoverlapping partition of $\mathcal{I} \times \mathcal{I}$ into admissible pairs of clusters (called block clusters) and nonadmissible pairs of (small) clusters. This partition is formed by the leaves of the block cluster tree $T^{\mathrm{iso}}_{\mathcal{I} \times \mathcal{I}}$, which we now define.

DEFINITION 3.4 (block cluster tree $T^{\mathrm{iso}}_{\mathcal{I}\times\mathcal{I}}$). *Let $T_{\mathcal{I}}$ be a cluster tree. Then the block cluster tree $T^{\mathrm{iso}}_{\mathcal{I}\times\mathcal{I}}$ based on the isotropic admissibility condition* (3.1) *is defined by the root $(\mathcal{I},\mathcal{I})$ and the father-son relation for $\mathbf{b}=(t,s)\in T^{\mathrm{iso}}_{\mathcal{I}\times\mathcal{I}}$:*

$$\operatorname{sons}(\mathbf{b}) := \begin{cases} \emptyset & \text{if } (t,s) \text{ is isotropically } \eta^{\mathrm{iso}}\text{-admissible,} \\ \operatorname{sons}(t)\times\operatorname{sons}(s) & \text{otherwise.} \end{cases}$$

The practical construction of the block cluster tree $T^{\mathrm{iso}}_{\mathcal{I}\times\mathcal{I}}$ from the root to the leaves is straightforward (cf. [19, 20]). The leaves $\mathcal{L}(T^{\mathrm{iso}}_{\mathcal{I}\times\mathcal{I}})$ consist of two different types of block clusters,

$$\mathcal{L}^{\mathrm{iso}}_{\mathrm{far}} := \{\mathbf{b}\in\mathcal{L}(T^{\mathrm{iso}}_{\mathcal{I}\times\mathcal{I}}) \,:\, \mathbf{b} \text{ is isotropically } \eta^{\mathrm{iso}}\text{-admissible }\}, \quad \mathcal{L}^{\mathrm{iso}}_{\mathrm{near}} := \mathcal{L}(T^{\mathrm{iso}}_{\mathcal{I}\times\mathcal{I}})\backslash\mathcal{L}^{\mathrm{iso}}_{\mathrm{far}}.$$

These are known as the nearfield block clusters ($\mathcal{L}^{\mathrm{iso}}_{\mathrm{near}}$) and the farfield block clusters ($\mathcal{L}^{\mathrm{iso}}_{\mathrm{far}}$). For each block cluster $\mathbf{b}=(t,s)$ we denote the corresponding domain by $\Gamma_{\mathbf{b}} := \Gamma_t\times\Gamma_s$. The set of all subdomains $P^{\mathrm{iso}} = \{\Gamma_{\mathbf{b}} \mid \mathbf{b}\in\mathcal{L}(T^{\mathrm{iso}}_{\mathcal{I}\times\mathcal{I}})\}$ is a partition of the tensor surface $\Gamma\times\Gamma$. The partition P^{iso} consists of two different types of subdomains, the farfield $P^{\mathrm{iso}}_{\mathrm{far}}$ and the nearfield $P^{\mathrm{iso}}_{\mathrm{near}}$:

$$P^{\mathrm{iso}}_{\mathrm{far}} := \left\{\Gamma_{\mathbf{b}} \,:\, \mathbf{b}\in\mathcal{L}^{\mathrm{iso}}_{\mathrm{far}}\right\}, \quad P^{\mathrm{iso}}_{\mathrm{near}} := P^{\mathrm{iso}}\setminus P^{\mathrm{iso}}_{\mathrm{far}}.$$

On each admissible block $\Gamma_{\mathbf{b}}\in P^{\mathrm{iso}}_{\mathrm{far}}$ the kernel function k is approximated by a degenerate expansion of the form

$$\tilde{k}_{\mathbf{b}}(\mathbf{x},\mathbf{y}) = \sum_{\nu=1}^{M} \Phi_{\mathbf{b},\nu}(\mathbf{x})\Psi_{\mathbf{b},\nu}(\mathbf{y}), \qquad (\mathbf{x},\mathbf{y})\in\Gamma_t\times\Gamma_s. \tag{3.2}$$

Remark 3.5. For the kernel function of the single layer potential for the Laplace problem, allowable expansion functions in (3.2) include, e.g., polynomial interpolation, multipole expansions, and Taylor expansion. One can show (for the proofs we refer the reader to, e.g., [29, 2]) that, for sufficiently small η^{iso}, there exist constants $0<\sigma<1$ (depending on η^{iso}) and $0<C<\infty$ with the following property: For all isotropically admissible blocks $\mathbf{b}=(t,s)\in\mathcal{L}^{\mathrm{iso}}_{\mathrm{far}}$ the Taylor expansion, Chebyshev interpolation, or multipole expansion of order m satisfies

$$\left|k(\mathbf{x},\mathbf{y}) - \tilde{k}_{\mathbf{b}}(\mathbf{x},\mathbf{y})\right| \;\le\; C\sigma^m \left(\operatorname{dist}(\mathcal{B}_t,\mathcal{B}_s)\right)^{-1} \quad \text{for all } (\mathbf{x},\mathbf{y})\in\Gamma_t\times\Gamma_s. \tag{3.3}$$

The number of terms M in (3.2) depends on m through $M(m) = m^q$ for some $q\in\{2,3\}$. For example, if the kernel $k(\mathbf{x},\mathbf{y})$ is interpolated (with respect to $\mathbf{y}$) using polynomials of degree m in each of the three coordinate directions, then there are $M=(m+1)^3$ terms in (3.2).

For interpolation and multipole expansions, (3.2) can be generalized so that the expansion functions depend only on the clusters and not on the blocks.

In the following, we briefly will sketch the algorithmic realization in the case of the Laplace single layer potential in Example 2.1 discretized by the Galerkin BEM with piecewise constant boundary elements.

For $\mathbf{b}=(t,s)\in\mathcal{L}^{\mathrm{iso}}_{\mathrm{far}}$ we define the matrices known as the "farfield coefficients" by

$$L^{\mathbf{b}}_{i,\nu} := \int_{\tau_i} \Phi_{\mathbf{b},\nu}(\mathbf{x})ds_{\mathbf{x}}, \quad R^{\mathbf{b}}_{\nu,j} := \int_{\tau_j} \Psi_{\mathbf{b},\nu}(\mathbf{y})ds_{\mathbf{y}}, \quad i\in t,\ j\in s,\ \nu\in\{1,\dots,M\},$$

and the nearfield matrix $\mathbf{V}^{\mathrm{near}}$ with entries

$$\mathbf{V}^{\mathrm{near}}_{i,j} := \begin{cases} \int_{\tau_i\times\tau_j} k(\mathbf{x},\mathbf{y})ds_{\mathbf{x}}ds_{\mathbf{y}} & \text{if } i\in t, j\in s, (t,s)\in\mathcal{L}^{\mathrm{iso}}_{\mathrm{near}},\\ 0 & \text{otherwise.}\end{cases}$$

The number of columns of $L^{\mathbf{b}}$ and the number of rows of $R^{\mathbf{b}}$ is equal to M. We have in mind that M will be small compared to N (typically $M\sim\log^q N$ for some $q\in\{2,3\}$). For the Laplace single layer potential on the screen (Example 2.1) $M\sim\log^2 N$ (see [2]).

The bilinear form

$$a(U,V)=\int_{\Gamma\times\Gamma} k(\mathbf{x},\mathbf{y})U(\mathbf{x})V(\mathbf{y})ds_{\mathbf{x}}ds_{\mathbf{y}} \qquad \text{for all } U,V\in S$$

is split into the blockwise bilinear forms

$$a(U,V)=\sum_{\mathbf{b}\in\mathcal{L}^{\mathrm{iso}}_{\mathrm{near}}\cup\mathcal{L}^{\mathrm{iso}}_{\mathrm{far}}} \underbrace{\int_{\Gamma_{\mathbf{b}}} k(\mathbf{x},\mathbf{y})U(\mathbf{x})V(\mathbf{y})ds_{\mathbf{x}}ds_{\mathbf{y}}}_{=:a_{\mathbf{b}}(U,V)}.$$

The panel-clustering approximation $\tilde{a}(U,V)$ of the bilinear form $a(U,V)$ is then obtained by replacing the blockwise bilinear form $a_{\mathbf{b}}(U,V)$ in the farfield blocks $\mathbf{b}=(t,s)\in\mathcal{L}^{\mathrm{iso}}_{\mathrm{far}}$ by

$$\begin{aligned}\tilde{a}_{\mathbf{b}}(U,V) &:= \int_{\Gamma_{\mathbf{b}}} \tilde{k}_{\mathbf{b}}(\mathbf{x},\mathbf{y})V(\mathbf{x})U(\mathbf{y})ds_{\mathbf{x}}ds_{\mathbf{y}}\\ &= \sum_{\nu=1}^{M}\int_{\Gamma_{\mathbf{b}}} \Phi_{\mathbf{b},\nu}(\mathbf{x})\Psi_{\mathbf{b},\nu}(\mathbf{y})V(\mathbf{x})U(\mathbf{y})ds_{\mathbf{x}}ds_{\mathbf{y}}\\ &= \sum_{i\in t}\sum_{j\in s}\left(L^{\mathbf{b}}R^{\mathbf{b}}\right)_{i,j} V(\xi_i)U(\xi_j). \qquad (3.4)\end{aligned}$$

The final equality follows because here we have restricted our analysis to piecewise constant elements.

The nearfield matrix $\mathbf{V}^{\mathrm{near}}$ contains all contributions from the nearfield blocks $\mathbf{b}\in\mathcal{L}^{\mathrm{iso}}_{\mathrm{near}}$. It is sparse since $P^{\mathrm{iso}}_{\mathrm{near}}$ covers only a small part of the tensor surface $\Gamma\times\Gamma$. The farfield coefficients $L^{\mathbf{b}}$ and $R^{\mathbf{b}}$ for blocks $\mathbf{b}\in\mathcal{L}^{\mathrm{iso}}_{\mathrm{far}}$ can be computed by standard quadrature techniques, since the degenerate expansion functions are smooth in the far field. They contain only $M(\#t+\#s)$ entries instead of the $\#t\#s$ entries of the full block $t\times s$.

In [11], the stability and convergence of the standard panel-clustering method is analyzed for general classes of meshes, including the stretched meshes discussed in section 2. It is shown there that the mesh degeneracy does not affect the accuracy of the approximation. However, it turns out that the standard algorithm is not optimal in terms of complexity, as the following example shows.

Example 3.6. Consider the Laplace single layer equation from Example 2.1, and consider piecewise constant elements on the meshes on the unit square in Example 2.4. In order to obtain optimal convergence rates we have to choose $g>3$ (see (2.5)).

We fix an arbitrary index $n+1\le i\le 2n$. All panels $\tau_{i,j}:=[p_{i-1},p_i]\times[p_{j-1},p_j]$ satisfy

$$p_i-p_{i-1}=(1-2\delta)n^{-1}=\left(\frac{g}{g+2}\right)n^{-1}>\frac{3}{5}n^{-1}.$$

On the other hand, for all indices $j = 1, \ldots, n^{1-1/g} + 1$ we have

$$p_{j-1} = \delta(j-1)^g n^{-g} \leq \delta n^{g-1} n^{-g} \leq \frac{1}{5} n^{-1} < \frac{3}{5} n^{-1},$$

so that for all $j, \ell \in J := \{1, \ldots, n^{1-1/g} + 1\}$

$$\operatorname{dist}(\tau_{i,j}, \tau_{i,\ell}) < \frac{3}{5} n^{-1}.$$

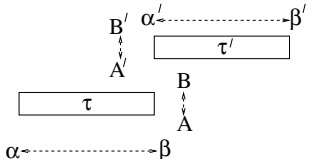
Thus any pair of clusters (t, s), where both t and s contain a panel $\tau_{i,j}$ with $j \in J$, cannot be isotropically admissible when the parameter η^{iso} is chosen such that $\eta^{iso} \leq 1$. Therefore, in the rows of the matrix K corresponding to the indices $\{(i,j) \mid i = n+1, \ldots, 2n,\ j \in J\}$ there are at least $\#J = n^{1-1/g}$ nearfield entries that must be computed and stored. There are $n^{2-1/g}$ such rows, and so we have to store at least $n^{3-2/g} = \mathcal{O}(N^{3/2-1/g})$ entries; i.e., the (storage) complexity of conventional panel-clustering is at least this. As g increases, the complexity approaches $\mathcal{O}(N^{3/2})$.

3.2. Anisotropic admissibility. The reason for the suboptimal performance of the panel-clustering method for stretched meshes is the isotropic admissibility condition. The near field of the long thin elements which lie close to edges of Γ becomes too large.

This problem is overcome by a modified panel-clustering algorithm which will be presented next.

3.2.1. The case of the screen. To illustrate the underlying idea we first consider the screen domain Γ from Example 2.4. Let two panels parallel to the x-axis be of the form

(3.5)



$$\begin{aligned} \tau &= (\alpha, \beta) \times (A, B), \\ \tau' &= (\alpha', \beta') \times (A', B'), \end{aligned}$$

with $\alpha < \beta$, $\alpha' < \beta'$, $A < B < A' < B'$.

Suppose also that

(3.6)

$$\max\{\operatorname{diam}(\tau), \operatorname{diam}(\tau')\} \geq \max\{\beta - \alpha, \beta' - \alpha'\} > \eta^{\text{iso}}(A' - B) = \eta^{\text{iso}} \operatorname{dist}(\tau, \tau'),$$

but at the same time

$$\max\{B - A, B' - A'\} \leq \eta^{\text{iso}} (A' - B). \tag{3.7}$$

Clearly (3.6) indicates that this pair of panels is inadmissible in the sense of condition (3.1). The problem is that the long sides (with length $\beta - \alpha$ and $\beta' - \alpha'$) are much longer than the distance between the elements. On the other hand, if somehow the lengths of the long sides could be ignored, then these elements would become admissible by virtue of (3.7). The essence of our new method is the observation that, provided the elements are parallel, analytic partial integration may be performed, leading to a new admissibility criterion which admits pairs satisfying (3.7) but not (3.1).

Continuing with the example (3.5), assuming U and V are piecewise constant functions and using the notation $\mathbf{x} = (x_1, x_2), \mathbf{y} = (y_1, y_2)$, we obtain

$$\begin{aligned} \int_\tau \int_{\tau'} k(\mathbf{x}, \mathbf{y}) V(\mathbf{x}) U(\mathbf{y}) \mathrm{d}\mathbf{x} \mathrm{d}\mathbf{y} &= V(\xi) U(\xi') \int_A^B \int_{A'}^{B'} \int_\alpha^\beta \int_{\alpha'}^{\beta'} k((x_1, x_2), (y_1, y_2)) \mathrm{d}\mathbf{x} \mathrm{d}\mathbf{y} \\ &= V(\xi) U(\xi') \int_A^B \int_{A'}^{B'} k^{\text{ad}}(x_2, y_2) \mathrm{d}y_2 \mathrm{d}x_2, \end{aligned}$$

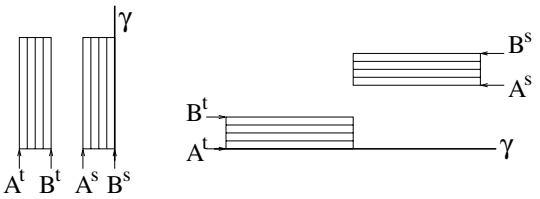


FIG. 3.2. *Examples of pairs (t,s) of anisotropically admissible clusters.*

where ξ and ξ' are the incenters of τ, τ' and $k^{\mathrm{ad}}(x_2,y_2) := \int_\alpha^\beta \int_{\alpha'}^{\beta'} k((x_1,x_2),(y_1,y_2))\mathrm{d}y_1\mathrm{d}x_1$ is the antiderivative of k. We will show that the twofold integration over the *long* directions (α,β), (α',β') leads to a kernel function $k^{\mathrm{ad}} : (A,B)\times(A',B') \to \mathbb{R}$ which is admissible for the one-dimensional intervals (A,B), (A',B') via the one-dimensional admissibility condition (3.7).

The idea can be carried over to pairs of clusters (instead of pairs of panels), provided they fulfil the following anisotropic admissibility condition.

DEFINITION 3.7 (anisotropic admissibility condition for the screen). *Choose $\eta^{\mathrm{aniso}} > 0$. A pair of clusters $(t,s) \in T_{\mathcal{I}} \times T_{\mathcal{I}}$ is anisotropically (η^{aniso}-) admissible if all panels τ_i, $i \in t$, and τ_j, $j \in s$, are of the form (in some suitable coordinate system; cf. Figure* 3.2*)*

$$\tau_i = (A_i^t, B_i^t)\times(\alpha^t,\beta^t), \quad \tau_j = (A_j^s, B_j^s)\times(\alpha^s,\beta^s), \tag{3.8}$$

or

$$\tau_i = (\alpha^t,\beta^t)\times(A_i^t, B_i^t), \quad \tau_j = (\alpha^s,\beta^s)\times(A_j^s, B_j^s) \tag{3.9}$$

and

$$\max\{|B^t - A^t|, |B^s - A^s|\} \;\le\; \eta^{\mathrm{aniso}}\,\mathrm{dist}(\mathcal{B}_t,\mathcal{B}_s), \tag{3.10}$$

where $A^t := \min\{A_i^t : i \in t\}$, $B^t := \max\{B_i^t : i \in t\}$, and A^s, B^s are defined analogously.

Let t and s, $\mathbf{b} := (t,s)$, be two clusters with corresponding panels of the form (3.8) or (3.9). The evaluation of the blockwise bilinear form $a_{\mathbf{b}}(U,V)$ for two piecewise constant finite element functions U and V can be written in the form

$$\begin{aligned} a_{\mathbf{b}}(U,V) &= \int_{\Gamma_{\mathbf{b}}} k(\mathbf{x},\mathbf{y})V(\mathbf{x})U(\mathbf{y})\mathrm{d}x\mathrm{d}y \\ &= \int_{\alpha^t}^{\beta^t}\int_{\alpha^s}^{\beta^s} \sum_{i\in t}\sum_{j\in s}\int_{A_i^t}^{B_i^t}\int_{A_j^s}^{B_j^s} V(x_1,x_2)U(y_1,y_2)k((x_1,x_2),(y_1,y_2))\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y} \\ &= \sum_{i\in t}\sum_{j\in s} V(\xi_i)U(\xi_j)\int_{A_i^t}^{B_i^t}\int_{A_j^s}^{B_j^s} k_{\mathbf{b}}^{\mathrm{ad}}(x_2,y_2)\mathrm{d}y_2\mathrm{d}x_2, \end{aligned}$$

where $k_{\mathbf{b}}^{\mathrm{ad}}$, as before, is the antiderivative of k,

$$k_{\mathbf{b}}^{\mathrm{ad}}(x_2,y_2) = \int_{\alpha^t}^{\beta^t}\int_{\alpha^s}^{\beta^s} k((x_1,x_2),(y_1,y_2))\mathrm{d}y_1\mathrm{d}x_1. \tag{3.11}$$

In section 4 we will prove that for an anisotropically admissible block $\mathbf{b} = (t, s)$ the integrated kernel $k_{\mathbf{b}}^{\mathrm{ad}}$ can be approximated by a degenerate kernel $\tilde{k}_{\mathbf{b}}^{\mathrm{ad}}$ of the form

$$\tilde{k}_{\mathbf{b}}^{\mathrm{ad}}(x_2, y_2) = \sum_{\nu=1}^{m} \Phi_{\mathbf{b},\nu}(x_2)\Psi_{\mathbf{b},\nu}(y_2). \tag{3.12}$$

The bilinear form $a_{\mathbf{b}}(U, V)$ is then approximated by

$$\begin{aligned}
\tilde{a}_{\mathbf{b}}(U,V) &:= \sum_{i\in t}\sum_{j\in s} V(\xi_i)\, U(\xi_j) \int_{A_i^t}^{B_i^t} \int_{A_j^s}^{B_j^s} \sum_{\nu=1}^{m} \Phi_{\mathbf{b},\nu}(x_2)\Psi_{\mathbf{b},\nu}(y_2)\mathrm{d}y_2\mathrm{d}x_2 \\
&= \sum_{i\in t}\sum_{j\in s} V(\xi_i)\, U(\xi_j) \sum_{\nu=1}^{m} \underbrace{\int_{A_i^t}^{B_i^t} \Phi_{\mathbf{b},\nu}(x_2)\mathrm{d}x_2}_{=:L_{i,\nu}^{\mathbf{b}}} \underbrace{\int_{A_j^s}^{B_j^s} \Psi_{\mathbf{b},\nu}(y_2)\mathrm{d}y_2}_{=:R_{\nu,j}^{\mathbf{b}}} \\
&= \sum_{i\in t}\sum_{j\in s} (L^{\mathbf{b}}R^{\mathbf{b}})_{i,j} V(\xi_i)U(\xi_j), \qquad (3.13)
\end{aligned}$$

where $L^{\mathbf{b}} \in \mathbb{R}^{\#t\times m}$ and $R^{\mathbf{b}} \in \mathbb{R}^{m\times \#s}$. Since it will turn out that m can be chosen proportional to $\log(N)$, the storage and matrix by vector complexity is $\mathcal{O}(\log(N)(\#t+\#s))$ instead of the quadratic complexity for the standard representation.

Combining the isotropic and anisotropic admissibilities, we arrive at a final block cluster tree $T_{\mathcal{I}\times\mathcal{I}}$ described as follows.

DEFINITION 3.8 (block cluster tree $T_{\mathcal{I}\times\mathcal{I}}$). *Let $T_{\mathcal{I}}$ be a cluster tree. Then the block cluster tree $T_{\mathcal{I}\times\mathcal{I}}$ based on the admissibility conditions* (3.1), (3.10) *is defined by the root $(\mathcal{I},\mathcal{I})$ and the father-son relation for $\mathbf{b} = (t, s) \in T_{\mathcal{I}\times\mathcal{I}}$:*

$$\mathrm{sons}\,(\mathbf{b}) := \begin{cases} \emptyset & \text{if } (t,s) \text{ is isotropically or anisotropically admissible,} \\ \mathrm{sons}(t)\times\mathrm{sons}(s) & \text{otherwise.} \end{cases}$$

The leaves $\mathbf{b} \in \mathcal{L}(T_{\mathcal{I}\times\mathcal{I}})$ of the block cluster tree form a partition of $\mathcal{I}\times\mathcal{I}$, and the corresponding domains $\Gamma_{\mathbf{b}}$ yield a partition of the tensor surface $\Gamma\times\Gamma$. In contrast to the isotropic case discussed in Definition 3.4, three different types of blocks will appear in this case:

- $\mathcal{L}_{\mathrm{far}}^{II} := \{\mathbf{b} \in \mathcal{L}(T_{\mathcal{I}\times\mathcal{I}}) : \mathbf{b}$ is anisotropically, but not isotropically, admissible$\}$. These blocks allow an approximation by $\tilde{a}_{\mathbf{b}}$ given in (3.13) and involve the integrated kernel $\tilde{k}_{\mathbf{b}}^{\mathrm{ad}}$.
- $\mathcal{L}_{\mathrm{far}}^{I} := \{\mathbf{b} \in \mathcal{L}(T_{\mathcal{I}\times\mathcal{I}}) : \mathbf{b}$ is isotropically admissible$\}$. These are treated by the standard techniques (3.4).
- $\mathcal{L}_{\mathrm{near}} := \mathcal{L}(T_{\mathcal{I}\times\mathcal{I}})\backslash\mathcal{L}_{\mathrm{far}}$, $\quad \mathcal{L}_{\mathrm{far}} := \mathcal{L}_{\mathrm{far}}^{II} \cup \mathcal{L}_{\mathrm{far}}^{I}$. For these blocks we use the original bilinear form $a_{\mathbf{b}}$.

For the block cluster tree $T_{\mathcal{I}\times\mathcal{I}}$ we will prove in section 5 that the nearfield matrix $\mathbf{V}^{\mathrm{near}}$ corresponding to entries from nearfield blocks $\mathbf{b} \in \mathcal{L}_{\mathrm{near}}$ is sparse (i.e., $\mathcal{O}(N)$ entries). This means that the generalized panel-clustering approximation

$$\tilde{a}(U,V) := \sum_{\mathbf{b}\in\mathcal{L}_{\mathrm{near}}} a_{\mathbf{b}}(U,V) + \sum_{\mathbf{b}\in\mathcal{L}_{\mathrm{far}}} \tilde{a}_{\mathbf{b}}(U,V) \tag{3.14}$$

can be evaluated in $\mathcal{O}(N\log(N)M)$ operations, where $M \approx \log^2 N$. In section 4 we prove the consistency of the panel-clustering approximation, i.e., that the complexity of $\mathcal{O}(N\log^3 N)$ is attained without damaging the accuracy of the underlying boundary element method.

3.2.2. The case of a polyhedral surface. The anisotropic admissibility condition (Definition 3.7) can be extended to the case of polyhedral surfaces Γ which are composed of faces with polygonal boundaries. Recall the definition of the containers Q and their base line γ_Q as in Remark 2.3.

DEFINITION 3.9 (anisotropic admissibility condition for polyhedral surfaces). *A pair of clusters (t,s) is anisotropically admissible with respect to $\gamma \in \mathcal{E}$ if there exist containers Q_t, Q_s with the same base line $\gamma_{Q_t} = \gamma_{Q_s}$ such that*

$$\Gamma_t \subset Q_t \quad \textit{and} \quad \Gamma_s \subset Q_s$$

and

$$\max\{H_t, H_s\} \;\leq\; \eta^{\mathrm{aniso}} \;\operatorname{dist}(\mathcal{B}_t, \mathcal{B}_s),$$

where for $i = s, t$

$$H_i := \sup_{\mathbf{x},\mathbf{y}\in\Gamma_i} \{\|\mathbf{x}-\mathbf{y}\| : (\mathbf{x}-\mathbf{y}) \perp (\gamma_{Q_i})\}.$$

Analogously to the screen case, the integration parallel to $\gamma_{Q_s} = \gamma_{Q_i}$ is carried out analytically, leading again to an antiderivative of the kernel function which can be approximated efficiently by panel-clustering with respect to the remaining (orthogonal) variables.

To be more precise, let (t,s) be a pair of clusters which is anisotropically admissible. Let Q_t, Q_s denote the containers for t, s which have the common base line $\gamma_Q = \gamma_{Q_t} = \gamma_{Q_s}$. Note that Q_t and Q_s may be subsets on the same face of Γ or subsets of two adjoining faces of Γ. We choose a local coordinate system such that the x_1-axis points in the direction of γ_Q and t lies in the (x_1,x_2)-plane. Because of the meshing algorithm, we have

$$t = \left\{(x_1,x_2,0)^\intercal \in \mathbb{R}^3 \mid A_t \leq x_2 \leq B_t \text{ and } \alpha_t(x_2) \leq x_1 \leq \beta_t(x_2)\right\}, \tag{3.15}$$

where $A_t < B_t$ are constants and α_t, β_t are affine functions which satisfy $\alpha_t(x_2) < \beta_t(x_2)$ for all $A_t \leq x_2 \leq B_t$. Analogously s is described by

$$s = \{(w_1, w_2\cos\varphi, w_2\sin\varphi)^\intercal \mid A_s \leq w_2 \leq B_s \text{ and } \alpha_s(w_2) \leq w_1 \leq \beta_s(w_2)\}.$$

Again, $A_s < B_s$ are constants, and α_s, β_s are affine functions which satisfy $\alpha_s(w_2) < \beta_s(w_2)$. The angle $\varphi \in [0, 2\pi[$ is fixed and denotes the angle between Γ_t and Γ_s. The antiderivative of the kernel function is then defined (analogously to (3.11)) by

$$k^{\mathrm{ad}}_{\mathbf{b}}(x_2,y_2) := \int_{\alpha_s(y_2)}^{\beta_s(y_2)} \int_{\alpha_t(x_2)}^{\beta_t(x_2)} k(\chi_t(x_1,x_2), \chi_s(w_1,y_2))\, dx_1 dw_1, \tag{3.16}$$

where χ_t, χ_s denote a unitary transformation of the two-dimensional coordinate systems to the surface panels.

An illustration of a pair of admissible clusters in this case is given in Figure 3.3.

The panel-clustering method for anisotropic elements uses an approximation of $k^{\mathrm{ad}}_{\mathbf{b}}$ by a degenerate expansion $\tilde{k}^{\mathrm{ad}}_{\mathbf{b}}$ of the form

$$\tilde{k}^{\mathrm{ad}}_{\mathbf{b}}(x_2,y_2) = \sum_{\nu=1}^{m} \Phi_{\mathbf{b},\nu}(x_2)\, \Psi_{\mathbf{b},\nu}(y_2). \tag{3.17}$$

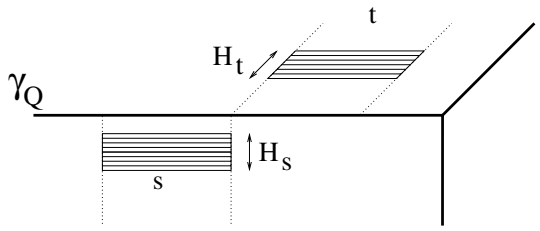


FIG. 3.3. *A pair of admissible clusters* (t, s).

The following assumption states that the approximation $\tilde{k}_{\mathbf{b}}^{\mathrm{ad}}$ converges exponentially towards $k_{\mathbf{b}}^{\mathrm{ad}}$.

ASSUMPTION 3.10. *Let k denote the kernel of the single layer operator, and let $k_{\mathbf{b}}^{\mathrm{ad}}$ be as in* (3.16). *For sufficiently small η^{aniso}, there exist constants $0 < \delta < 1$ (depending on η^{aniso}) and $0 < C < \infty$ with the following property: For all anisotropically admissible blocks $\mathbf{b} = (t, s) \in \mathcal{L}_{\mathrm{far}}^{\mathrm{aniso}}$ there exists a family of degenerate expansions $\tilde{k}_{\mathbf{b}}^{\mathrm{ad}}(\cdot, \cdot)$ depending on a parameter $m \in \mathbb{N}$ of the form* (3.17) *with*

$$\left| k_{\mathbf{b}}^{\mathrm{ad}}(\mathbf{x}, \mathbf{y}) - \tilde{k}_{\mathbf{b}}^{\mathrm{ad}}(\mathbf{x}, \mathbf{y}) \right| \leq C m^2 \delta^m \qquad \text{for all } (\mathbf{x}, \mathbf{y}) \in \Gamma_t \times \Gamma_s. \tag{3.18}$$

Remark 3.11. The proof of (3.18) in the case of the screen problem is given in Theorem 6.4, while the general case can be analyzed along the same lines.

Remark 3.12. Higher order elements can also be treated by anisotropic panel-clustering, provided we write the basis functions as a sum of products of the form $\psi(x_1)\phi(x_2)$, with x_1 being the coordinate in the direction orthogonal to the base line γ_Q.

Along the edges of the surface Γ it usually happens that stretched elements from two different containers (meeting at the edge) are contained in the same cluster t. In this situation, we will consider a pair (t, s) admissible even if this is only true for the two subsets of t coming from each of the two containers.

DEFINITION 3.13 (generalized anisotropic admissibility condition). *Let C_{cl} be a given constant (cf. section 5, typically $C_{\mathrm{cl}} = 2$). A pair of clusters (t, s) is (generalized) anisotropically admissible if there exists a partition*

$$t = \dot{\bigcup_{i=1,\ldots,C_{\mathrm{cl}}}} t_i, \qquad s = \dot{\bigcup_{j=1,\ldots,C_{\mathrm{cl}}}} s_j$$

such that each pair (t_i, s_j) is anisotropically admissible as defined in Definition 3.9.

The generalized anisotropic admissibility is in accordance with Assumption 3.10, because after the subdivision into $C_{\mathrm{cl}} \times C_{\mathrm{cl}}$ subblocks one can apply the one-dimensional expansion as in (3.17) for the arising subblocks.

4. Error analysis. In this section, we will prove the consistency and stability of the generalized panel-clustering algorithm for the single layer potential equation from Example 2.1 discretized by piecewise constant elements on the stretched meshes defined in Algorithm 2.2. In [11] the consistency and stability of the standard panel-clustering method for such meshes was proved. Thus the results here are a generalization of those, with particular emphasis on the approximations on the anisotropic blocks. The error analysis will be based on the second Strang lemma, which we

briefly recall. Let $\tilde{a} : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$ denote the bilinear form corresponding to the panel-clustering approximation (3.14). The corresponding approximate Galerkin method is to seek $\tilde{U} \in \mathcal{S}$ such that

$$\tilde{a}(\tilde{U}, V) = (g, V). \tag{4.1}$$

The following theorem follows from [7] (see in particular [11, Lemma 4.2, equation (4.24)]).

THEOREM 4.1. *Assume that there exists a continuous function $\mu : \mathbb{R}_+ \to \mathbb{R}_+$ with $\mu(N) \to 0$ as $N \to \infty$ such that, for all $V, W \in \mathcal{S}$,*

(i) $|a(V,W) - \tilde{a}(V,W)| \lesssim \mu(N)\|V\|_{H^{-1/2}(\Gamma)}\|W\|_{H^{-1/2}(\Gamma)}$,

(ii) $|a(V,W) - \tilde{a}(V,W)| \lesssim N^{-3/4}\|V\|_{L^2(\Gamma)}\ \|W\|_{H^{-1/2}(\Gamma)}$.

Then, for N sufficiently large, a unique solution $\tilde{U} \in \mathcal{S}$ of (4.1) exists and satisfies

$$\left\|u - \tilde{U}\right\|_{H^{-1/2}(\Gamma)} \lesssim N^{-1/4} \inf_{Z \in \mathcal{S}} \|u - Z\|_{L^2(\Gamma)} + N^{-3/4}\|u\|_{L^2(\Gamma)}.$$

The fact that $u \in L^2(\Gamma)$ follows from [6, Theorem 3]. Now it is shown in [23, Satz 3.7] that a suitable choice of g can be made which ensures that $\inf_{Z\in\mathcal{S}} \|u - Z\|_{L^2(\Gamma)} = N^{-1/2}$, and so in this case we obtain the estimate

$$\left\|u - \tilde{U}\right\|_{H^{-1/2}(\Gamma)} \lesssim N^{-3/4}. \tag{4.2}$$

In the remaining part of this section, we will prove that the bilinear form $\tilde{a}(\cdot,\cdot)$ satisfies the assumptions (i) and (ii) of Theorem 4.1. To reduce technicalities we assume for the following that the boundary element mesh $\mathcal{T}$ is conforming; i.e., the intersection $\overline{\tau} \cap \overline{\sigma}$ of any two nonidentical panels $\tau, \sigma \in \mathcal{T}$ is either empty or a common vertex or a common edge. This can be easily achieved by performing an appropriate closure of the mesh for the L-shaped regions in Algorithm 2.2, Step 3. Furthermore, we assume that all panels are convex and the interior angles are bounded above by a constant $\alpha_{\max} < \pi$. We introduce a positive-valued continuous function ρ as follows. For each panel τ, let ρ_τ denote the diameter of the largest inscribed disc. Let $\widehat{S}$ denote the unit triangle with vertices $(0,0)$, $(1,0)$, $(0,1)$, and let $\widehat{Q} := (0,1)^2$ denote the unit square. For each $\tau \in \mathcal{T}$, if τ is a triangle, let $\chi_\tau : \widehat{S} \to \tau$ denote an affine bijection, and if τ is a quadrilateral, let $\chi_\tau : \widehat{Q} \to \tau$ denote a bijection which is affine in each variable. Then ρ is defined by requiring the following:

- For all panels $\tau : \rho \circ \chi_\tau$ is $\begin{cases} \text{affine} & \text{if } \tau \text{ is a triangle,} \\ \text{affine in each variable} & \text{if } \tau \text{ is a quadrilateral.} \end{cases}$
- For all panel vertices $\mathbf{x}_p : \rho(\mathbf{x}_p) = \max\{\rho_\tau : \mathbf{x}_p \in \overline{\tau}\}$.

We use the function ρ in the proof of the following result.

THEOREM 4.2. *Assume that we are solving the single layer potential equation in Example 2.1 on a Lipschitz polyhedral surface Γ, meshed as in Algorithm 2.2, and choose $g = 3 + \epsilon$, where $\epsilon > 0$ is any small number. Let σ be as in Remark 3.5, and let Assumption 3.10 be satisfied for some $0 < \delta < 1$. Choose the orders m of the panel-clustering according to*

$$m = \left\lceil \frac{(2+\epsilon)\ \log N}{|\log \sigma|} \right\rceil \quad \textit{for the isotropic panel-clustering,} \tag{4.3}$$

$$m = \left\lceil \frac{(5/2+\epsilon)\ \log N}{|\log \delta|} \right\rceil \quad \textit{for the anisotropic panel-clustering,} \tag{4.4}$$

where $\lceil x \rceil$ *denotes the smallest integer* z *with* $z \geq x$. *Then the assumptions* (i) *and* (ii) *of Theorem* 4.1 *hold.*

Proof. For a block $\mathbf{b} \in \mathcal{L}^I_{\text{far}}$, let $\tilde{k}_{\mathbf{b}}$ denote the approximation as in (3.4), and let $\varepsilon_{\mathbf{b}} := k - \tilde{k}_{\mathbf{b}}$ denote the corresponding error. For a block $\mathbf{b} \in \mathcal{L}^{II}_{\text{far}}$, the partially integrated kernel function is denoted by $k^{\text{ad}}_{\mathbf{b}}$ (cf. 3.16) and the approximation by $\tilde{k}^{\text{ad}}_{\mathbf{b}}$ (cf. (3.12)). The corresponding error is denoted by $\varepsilon^{\text{ad}}_{\mathbf{b}} = k^{\text{ad}}_{\mathbf{b}} - \tilde{k}^{\text{ad}}_{\mathbf{b}}$. Thus

$$\begin{aligned} |a(U,V) - \tilde{a}(U,V)| &\leq \sum_{\mathbf{b}\in\mathcal{L}^I_{\text{far}}} \int_{\mathbf{b}} |\varepsilon_{\mathbf{b}}(\mathbf{x},\mathbf{y})|\,|U(\mathbf{x})|\,|V(\mathbf{y})|\,ds_{\mathbf{x}}ds_{\mathbf{y}} \\ &\quad + \sum_{\mathbf{b}\in\mathcal{L}^{II}_{\text{far}}} \int_{A^t}^{B^t} \int_{A^s}^{B^s} \left|\varepsilon^{\text{ad}}_{\mathbf{b}}(x_2,y_2)\right|\,|U(x_2)|\,|V(y_2)|\,dx_2dy_2 \\ &=: E^I(U,V) + E^{II}(U,V), \end{aligned} \tag{4.5}$$

where for all $\mathbf{b} = (s,t) \in \mathcal{L}^{II}_{\text{far}}$ and $(i,j) \in (s,t)$ we have set

$$\begin{aligned} U(x_2) &:= U(\xi_i) \qquad \text{for all } x_2 : (x_1,x_2) \in \tau_i, \\ V(y_2) &:= V(\xi_j) \qquad \text{for all } y_2 : (y_1,y_2) \in \tau_j. \end{aligned}$$

In the definition of E^{II} we have assumed the scenario and notation as in (3.9), i.e., $\mathbf{b} = (t,s)$, with

$$\Gamma_t = \bigcup\nolimits_{i=1}^{\#t} \tau_i, \quad \text{with } \tau_i = (\alpha^t,\beta^t)\times(A^t_i,B^t_i),\ h_i = \beta^t - \alpha^t,\ \rho_i = B^t_i - A^t_i, \tag{4.6}$$

and the analogous definition of Γ_s.

We first turn to the estimate of E^I and proceed as in [11]. From Assumption 3.5 and the fact that $\theta = 1$, we obtain

$$\begin{aligned} E^I(U,V) &\lesssim \sigma^m \sum_{(t,s)\in\mathcal{L}^I_{\text{far}}} \int_{\Gamma_t\times\Gamma_s} \frac{1}{\operatorname{dist}(\mathcal{B}_t,\mathcal{B}_s)} |U(\mathbf{x})||V(\mathbf{y})| ds_{\mathbf{y}}ds_{\mathbf{x}} \\ &= \sigma^m \sum_{(t,s)\in\mathcal{L}^I_{\text{far}}} \frac{1}{\operatorname{dist}(\mathcal{B}_t,\mathcal{B}_s)} \int_{\Gamma_t} |U(\mathbf{x})| ds_{\mathbf{x}} \int_{\Gamma_s} |V(\mathbf{y})| ds_{\mathbf{y}}. \end{aligned}$$

Since (t,s) is an isotropically admissible pair, from (3.3), we have

$$\operatorname{dist}(\mathcal{B}_t,\mathcal{B}_s)^{-1} \ \leq\ \eta^{iso} \max(\operatorname{diam}\mathcal{B}_t, \operatorname{diam}\mathcal{B}_s)^{-1} \leq \eta^{iso} h_i^{-1} \quad \text{for all } i \in t.$$

So, using the fact that the sets $\{\Gamma_t \times \Gamma_s : (t,s) \in \mathcal{L}^I_{\text{far}}\}$ are pairwise disjoint (and also $|\tau_i| \leq h_i^2$), we have

$$\begin{aligned} E^I(U,V) &\lesssim \eta^{iso}\sigma^m \sum_{i\in\mathcal{I}} \sum_{j\in\mathcal{I}} h_i^{-1} \|U\|_{L^1(\tau_i)} \|V\|_{L^1(\tau_j)} \\ &\leq \eta^{iso}\sigma^m \sum_{i\in\mathcal{I}} \sum_{j\in\mathcal{I}} h_j\ \|U\|_{L^2(\tau_i)} \|V\|_{L^2(\tau_j)}. \end{aligned}$$

Now, for any $\theta \in [0,1/2]$, using $h_j \lesssim N^{-1/2}$ and $\rho_j \gtrsim N^{-g/2}$, we obtain

$$\begin{aligned} E^I(U,V) &\lesssim \eta^{iso}\sigma^m N^{-1/2} \sum_{i\in\mathcal{I}} \sum_{j\in\mathcal{I}} \rho_i^{-\theta} \|\rho^\theta U\|_{L^2(\tau_i)} \rho_j^{-1/2} \|\rho^{1/2} V\|_{L^2(\tau_j)} \\ &\lesssim \eta^{iso}\sigma^m N^{-1/2} N^{g\theta/2} N^{g/4} \sum_{i\in\mathcal{I}} \|\rho^\theta U\|_{L^2(\tau_i)} \sum_{j\in\mathcal{I}} \|\rho^{1/2} V\|_{L^2(\tau_j)}. \end{aligned}$$

Hence, using the Cauchy–Schwarz inequality and then the inverse estimates in [11, Theorem 3.6], we have

$$\begin{aligned} E^I(U,V) &\leq \eta^{iso}\sigma^m N^{-1/2} N^{g\theta/2} N^{g/4} N \|\rho^\theta U\|_{L^2(\Gamma)} \|\rho^{1/2} V\|_{L^2(\Gamma)} \\ &\leq \eta^{iso}\sigma^m N^{1/2+g\theta/2+g/4} \|U\|_{H^{-\theta}(\Gamma)} \|V\|_{H^{-1/2}(\Gamma)}. \end{aligned}$$

The required estimate for $E^I(U,V)$ follows by observing that by taking m as given in (4.3) and with $g = 3+\epsilon$,

$$\sigma^m N^{1/2+g\theta/2+g/4} \lesssim \begin{cases} N^{-3(1+\epsilon)/4} & \text{if } \theta = 0, \\ N^{-\epsilon/2} & \text{if } \theta = 1/2. \end{cases}$$

In order to estimate $E^{II}(U,V)$, take a typical block $\mathbf{b} = (t,s)$ in $\mathcal{L}^{II}_{\text{far}}$ and assume the notation as in (4.6). Assumption 3.10 implies

$$\left|\varepsilon^{\text{ad}}_{\mathbf{b}}(x_2,y_2)\right| \lesssim m^2\delta^m.$$

Hence

$$E^{II}(U,V) \lesssim m^2\delta^m \sum_{\mathbf{b}=(t,s)\in\mathcal{L}^{II}_{\text{far}}} \int_{A^t} |U(x_2)|dx_2 \int_{A^s} |V(x_2)|dx_2. \tag{4.7}$$

Moreover, for any $\theta \in [0,1/2]$

$$\begin{aligned} \int_{A^t} |U(x_2)|dx_2 &= \sum_{i\in t} \int_{A^t_i} |U(x_2)|dx_2 \lesssim \sum_{i\in t} \rho_i^{1/2} \left\{ \int_{A^t_i} |U(x_2)|^2 dx_2 \right\}^{1/2} \\ &\lesssim \sum_{i\in t} \rho_i^{1/2-\theta} h_i^{-1/2} \|\rho^\theta U\|_{L^2(\tau_i)}. \end{aligned}$$

Since $\rho_i \leq h_i$ and $1/2-\theta \geq 0$, we have $\rho_i^{1/2-\theta} h_i^{-1/2} \lesssim h_i^{-\theta} \lesssim N^{g\theta/2}$, and so

$$\int_{A^t} |U(x_2)|dx_2 \lesssim N^{g\theta/2} \sum_{i\in t} \|\rho^\theta U\|_{L^2(\tau_i)}. \tag{4.8}$$

Hence combining (4.7) and (4.8) and using the Cauchy–Schwarz inequality, we get

$$\begin{aligned} E^{II}(U,V) &\lesssim m^2\delta^m N^{(g\theta/2+g/4)} \sum_{i\in\mathcal{I}} \|\rho^\theta U\|_{L^2(\tau_i)} \sum_{j\in\mathcal{I}} \|\rho^{1/2} V\|_{L^2(\tau_j)} \\ &\lesssim m^2\delta^m N^{(g\theta/2+g/4+1)} \|\rho^\theta U\|_{L^2(\Gamma)} \|\rho^{1/2} V\|_{L^2(\Gamma)}. \end{aligned}$$

Then, making use of [11, Theorem 3.6], we have

$$E^{II}(U,V) \lesssim m^2\delta^m N^{g\theta/2+g/4+1} \|U\|_{H^{-\theta}(\Gamma)} \|V\|_{H^{-1/2}(\Gamma)}. \tag{4.9}$$

Now if we choose m as in (4.4), it follows that

$$\delta^m N^{g\theta/2+g/4+1} \lesssim \begin{cases} N^{-3/4(1+\epsilon)}, & \theta = 0, \\ N^{-\epsilon/2}, & \theta = 1/2, \end{cases}$$

and so the required estimates for E^{II} follow. ∎

5. Complexity analysis. In this section we analyze the complexity of the storage requirements and of the matrix-vector multiplication for our new version of the panel-clustering algorithm. First we introduce a measure of the sparsity of the block cluster tree $T_{\mathcal{I}\times\mathcal{I}}$.

DEFINITION 5.1 (sparsity). *The sparsity of a block cluster tree $T_{\mathcal{I}\times\mathcal{I}}$ based on a cluster tree $T_{\mathcal{I}}$ is characterized by the quantity*

$$C_{\mathrm{sp}} := \max\left\{\max_{s\in T_{\mathcal{I}}} \#\{t\in T_{\mathcal{I}} \mid s\times t\in T_{\mathcal{I}\times\mathcal{I}}\},\ \max_{t\in T_{\mathcal{I}}} \#\{s\in T_{\mathcal{I}} \mid s\times t\in T_{\mathcal{I}\times\mathcal{I}}\}\right\}.$$

The importance of C_{sp} was discussed in [15]. In the case of shape-regular meshes and with cluster trees created using standard bisection algorithms, using the isotropic admissibility condition (3.1), it was shown in [15, Lemma 4.5] that C_{sp} is bounded independently of N (the number of degrees of freedom in the mesh). Here we shall extend this result to the anisotropic case. Before we do this, we will state the complexity bounds for the storage and matrix-vector product of the panel-clustering approximation $\tilde{a}(U,V)$ of the bilinear form $a(U,V)$ based on the measure C_{sp} and the depth p of the cluster tree. Both quantities will be bounded in Theorem 5.11 under quite general conditions, which are fulfilled for the screen mesh from Example 2.4.

THEOREM 5.2 (storage and matrix by vector complexity). *Let $T := T_{\mathcal{I}\times\mathcal{I}}$ be a block cluster tree with sparsity C_{sp} and depth $p\geq 1$. Let m be the (maximal) interpolation order used for the interpolation of the kernel function in isotropically or anisotropically admissible blocks. Then the storage requirements $N_{\mathrm{St}}(T,m)$ and the matrix by vector complexity $N_{\tilde{a}(U,V)}(T,m)$ for the panel-clustering approximation* (3.13) *are bounded by*

$$N_{\mathrm{St}}(T,m) = \mathcal{O}(C_{\mathrm{sp}}m^2p\#\mathcal{I}), \qquad N_{\tilde{a}(U,V)}(T,m)\sim N_{\mathrm{St}}(T,m).$$

In particular for the screen mesh from Example 2.4*, we have $pm^2\#\mathcal{I} = \mathcal{O}(gN\log^3 N)$.*

Proof. (a) For each block $(t,s)\in\mathcal{L}_{\mathrm{near}}$ we have to store the entries V_{ij}^{near} for all $(i,j)\in t\times s$. The construction of the block cluster tree ensures that either t or s is a leaf of $T_{\mathcal{I}}$, i.e., $\#t\leq n_{\min}$ or $\#s\leq n_{\min}$, so that at most $n_{\min}(\#t+\#s)$ entries have to be stored.

(b) For each block $\mathbf{b}=(t,s)\in\mathcal{L}_{\mathrm{far}}$ we have to store the entries $L_{i,\nu}^{\mathbf{b}}$ and $R_{\nu,j}^{\mathbf{b}}$ for all $i\in t, j\in s, \nu\in M$, where $M=\{1,\ldots,m\}$ for anisotropically admissible blocks and $M=\{1,\ldots,m^q\}$ (q from Remark 3.5) for isotropically admissible blocks. In total we can bound the number of entries to be stored for the block $\mathbf{b}$ by $m^q(\#t+\#s)$.

(c) Due to (a) and (b) the storage complexity for each block $\mathbf{b}\in\mathcal{L}(T_{\mathcal{I}\times\mathcal{I}})$ is bounded by $\max\{n_{\min},m^q\}(\#t+\#s)$. We conclude

$$\begin{aligned}
N_{\mathrm{St}}(T,m) &\leq \sum_{(t,s)\in\mathcal{L}(T_{\mathcal{I}\times\mathcal{I}})} \max\{n_{\min},m^q\}(\#t+\#s)\\
&\leq \sum_{t\in T_{\mathcal{I}}}\sum_{(t,s)\in T_{\mathcal{I}\times\mathcal{I}}} \max\{n_{\min},m^q\}\#t + \sum_{s\in T_{\mathcal{I}}}\sum_{(t,s)\in T_{\mathcal{I}\times\mathcal{I}}} \max\{n_{\min},m^q\}\#s\\
&\leq 2C_{\mathrm{sp}}\max\{n_{\min},m^q\}\sum_{t\in T_{\mathcal{I}}}\#t\\
&\overset{[15,\ \mathrm{R.1.10}]}{\leq} 2C_{\mathrm{sp}}\max\{n_{\min},m^q\}(p+1)\#\mathcal{I} = \mathcal{O}(C_{\mathrm{sp}}m^qp\#\mathcal{I}).
\end{aligned}$$

(d) The matrix by vector complexity is proportional to the storage complexity because every stored entry of V^{near}, $L^{\mathbf{b}}$, and $R^{\mathbf{b}}$ is accessed exactly once when computing a matrix-vector product. ∎

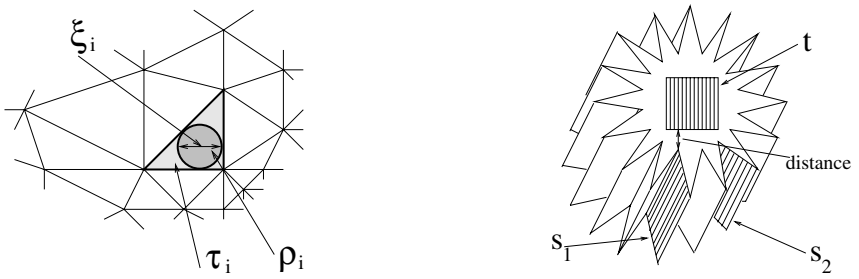


FIG. 5.1. *Left: In the inscribed circle around the incenter ξ_i there lies no other incenter $\xi_j, j \neq i$, i.e., $C_{\rm ov} = 1$. Right: The cluster t is surrounded by many clusters of the type of s_1, s_2 with very small distance to t.*

COROLLARY 5.3 (complexity for the cluster tree construction). *A cluster tree $T_{\mathcal{I}}$ of depth $p \geq 1$ has at most $\#T_{\mathcal{I}} = \mathcal{O}(p\#\mathcal{I})$ nodes. The complexity for the standard construction (so-called geometrically regular clustering) as introduced in section* 3.1 *is $\mathcal{O}(p\#\mathcal{I})$.*

Proof. Using [15, Remark 1.10], we obtain $\#\mathcal{L}(T_{\mathcal{I}}) \leq \#\mathcal{I}$, so that the property $t = \dot{\bigcup}_{s \in \mathrm{sons}(t)} s$ for interior nodes yields $\#T_{\mathcal{I}} = \mathcal{O}(p\#\mathcal{I})$. In the standard construction of the cluster tree $T_{\mathcal{I}}$ we compute the boxes $\tilde{\mathcal{B}}_1, \tilde{\mathcal{B}}_2$ for each node t in $\mathcal{O}(1)$ and determine for each point $(\xi_i)_{i \in t}$ the box to which it belongs, in total $\mathcal{O}(\#t)$ operations. Due to [15, Remark 1.10] the sum over all $t \in T_{\mathcal{I}}$ is $\mathcal{O}(p\#\mathcal{I})$. ∎

For the following bound on the sparsity measure $C_{\rm sp}$ we will now introduce some mesh-dependent quantities. The first of them is the quantity C_{ov} defined below. This quantifies the maximal number of nearby panels.

DEFINITION 5.4. *We define (cf. Figure* 5.1*)*

$$C_{\rm ov} := \max_{i \in \mathcal{I}} \#\left\{ j \in \mathcal{I} \mid \mathrm{dist}(\xi_i, \xi_j) \leq \rho_i/2 \right\},$$

where ρ_i is the diameter of the largest inscribed circle in τ_i and ξ_i, ξ_j are the incenters of τ_i, τ_j, respectively.

In many cases, and in particular for the screen mesh from Example 2.4, we have $C_{\rm ov} = 1$. In general (e.g., for highly folded surfaces) $C_{\rm ov} > 1$ might hold.

Next we want to characterize the alignment (in the sense of (3.8), (3.9), or (3.15)) of stretched panels. A situation where the constants in the complexity estimates might become large is depicted in Figure 5.1: The cluster t containing stretched panels is surrounded by a large number C of clusters $s_1, \ldots, s_C$ that also contain stretched panels. The clusters s_i are pairwise unaligned (the union of any two of those clusters would not fulfil the requirements of the anisotropic admissibility condition). In the following we want to consider the neighborhood of a panel.

DEFINITION 5.5. *Let $C_{\rm str} > 2 + 3\eta^{\rm iso}$. We define the neighborhood of a panel τ_i, $i \in \mathcal{I}$, by*

$$N_{str}(\tau_i) := \{ j \in \mathcal{I} \mid \mathrm{dist}(\tau_j, \tau_i) < C_{\rm str} (\eta^{\rm iso})^{-1} \max\{\mathrm{diam}(\tau_i), \mathrm{diam}(\tau_j)\}.$$

A partition of the form $N_{str}(\tau_i) = \dot{\cup}_{\nu=1}^{C} N_{str}^{\nu}(\tau_i)$ is an alignment-classification of $N_{str}(\tau_i)$, if for all $\nu = 1, \ldots, C$ the set $N_{str}^{\nu}(\tau_i)$ is aligned, i.e., of the form (3.8), (3.9), *or* (3.15).

ASSUMPTION 5.6. *We assume that for any panel τ_i, $i \in \mathcal{I}$, with $h_i > C_{\rm aniso}\, \rho_i$ there exists an alignment-classification of $N_{str}(\tau_i)$ with not more than $C_{\rm cl}$ classes,*

where $C_{\rm cl}$ depends on the given mesh but is bounded independently of i, the grading parameter g, and the number of unknowns N.

Remark 5.7. For the screen mesh the number of classes is $C_{\rm cl} = \mathcal{O}(C_{\rm str}/\eta^{\rm iso})$ independent of $C_{\rm aniso}$.

The following three auxiliary lemmata are needed to prove the final theorem of this section, Theorem 5.11.

LEMMA 5.8. *Let $t \in T_{\mathcal{I}}$ be a cluster of cardinality $\#t > C_{\rm ov}$. If the diameter of the minimal bounding box $\mathcal{B}_t$ is larger than the diameter of the regular box $\tilde{\mathcal{B}}_t$ (cf. section* 3.1*, Figure* 3.1*) of the cluster tree, i.e.,*

$$\operatorname{diam}(\mathcal{B}_t) = C\operatorname{diam}(\tilde{\mathcal{B}}_t), \quad C > 1,$$

then there exists an index $i \in t$ such that

$$h_i \geq (C-1)/C\operatorname{diam}(\mathcal{B}_t) \qquad \text{and} \qquad h_i > \frac{1}{2}(C-1)\rho_i.$$

Proof. Let $\operatorname{diam}(\mathcal{B}_t) = C\operatorname{diam}(\tilde{\mathcal{B}}_t)$. Then there exists at least one panel τ_i with centroid $\xi_i \in \tilde{\mathcal{B}}_t$ and a point $\zeta \in \tau_i \cap \partial\mathcal{B}_t$ with $\operatorname{dist}(\zeta, \tilde{\mathcal{B}}_t) \geq (C-1)/2\operatorname{diam}(\tilde{\mathcal{B}}_t)$. We conclude $h_i \geq 2|\zeta - \xi_i| \geq (C-1)\operatorname{diam}(\tilde{\mathcal{B}}_t) = (C-1)/C\operatorname{diam}(\mathcal{B}_t)$. Assume that $h_i \leq (C-1)\rho_i/2$. Then $\rho_i \geq 2h_i/(C-1) \geq 2\operatorname{diam}(\mathcal{B}_t)/C = 2\operatorname{diam}(\tilde{\mathcal{B}}_t)$, so that $|\xi_i - \xi_j| \leq \rho_i/2$ holds for all $j \in t$. Due to the assumption $\#t > C_{\rm ov}$, this is not possible. ∎

LEMMA 5.9. *Let $t, s \in T_{\mathcal{I}}$ be two clusters of cardinality $\#t > C_{\rm ov}, \#s > C_{\rm ov}$ that are not isotropically admissible: $\eta^{\rm iso}\operatorname{dist}(\mathcal{B}_t, \mathcal{B}_s) < \max\{\operatorname{diam}(\mathcal{B}_t), \operatorname{diam}(\mathcal{B}_s)\}$. Let $\operatorname{diam}(\mathcal{B}_t) = C\operatorname{diam}(\tilde{\mathcal{B}}_t)$ or $\operatorname{diam}(\mathcal{B}_s) = C\operatorname{diam}(\tilde{\mathcal{B}}_s)$ for some $C \geq C_{\rm str}/(C_{\rm str} - 1 - 2\eta^{\rm iso})$ and $\operatorname{diam}(\tilde{\mathcal{B}}_t) = \operatorname{diam}(\tilde{\mathcal{B}}_s)$. Then one of the following two cases is true:*

1. *$\exists i \in t : s \subset N_{str}(\tau_i)$ and $h_i \geq \frac{1}{2}(C-1)\rho_i$, or*
2. *$\exists j \in s : t \subset N_{str}(\tau_j)$ and $h_j \geq \frac{1}{2}(C-1)\rho_j$.*

Proof. Case 1. Let $\operatorname{diam}(\mathcal{B}_t) = C\operatorname{diam}(\tilde{\mathcal{B}}_t)$ and $\operatorname{diam}(\mathcal{B}_t) \geq \operatorname{diam}(\mathcal{B}_s)$. According to Lemma 5.8 there exists $i \in t$ such that $h_i \geq \frac{1}{2}(C-1)\rho_i$ and $h_i \geq \frac{C-1}{C}\operatorname{diam}(\mathcal{B}_t)$. We can now bound the distance of τ_i to any $\tau_j, j \in s$, by

$$\begin{aligned}\operatorname{dist}(\tau_i, \tau_j) &\leq \operatorname{diam}(\mathcal{B}_t) + \operatorname{diam}(\mathcal{B}_s) + \operatorname{dist}(\mathcal{B}_t, \mathcal{B}_s) \;\leq\; (2 + (\eta^{\rm iso})^{-1})\operatorname{diam}(\mathcal{B}_t) \\ &\leq \frac{C}{C-1}(2 + (\eta^{\rm iso})^{-1})h_i \;=\; \frac{C}{C-1}(1 + 2\eta^{\rm iso})(\eta^{\rm iso})^{-1}h_i \;\leq\; C_{\rm str}(\eta^{\rm iso})^{-1}h_i.\end{aligned}$$

According to Definition 5.5, $j \in N_{str}(\tau_i)$; i.e., $s \subset N_{str}(\tau_i)$.

Case 2. Let $\operatorname{diam}(\mathcal{B}_s) = C\operatorname{diam}(\tilde{\mathcal{B}}_s)$ and $\operatorname{diam}(\mathcal{B}_s) \geq \operatorname{diam}(\mathcal{B}_t)$. Analogously to case 1 we have $t \subset N_{str}(\tau_j)$ for some $j \in s$ and $h_j \geq \frac{1}{2}(C-1)\rho_j$. ∎

LEMMA 5.10. *Let $t, s_1, \ldots, s_\ell \in T_{\mathcal{I}}$ be clusters of cardinality $\#t > C_{\rm ov}, \#s_i > C_{\rm ov}$ that are not isotropically admissible and $\max\{\operatorname{diam}(\mathcal{B}_{s_i}), \operatorname{diam}(\mathcal{B}_t)\} \geq C\operatorname{diam}(\tilde{\mathcal{B}}_{s_i}) = \operatorname{diam}(\tilde{\mathcal{B}}_t)$, where $C := max\{2C_{\rm aniso} + 1, C_{\rm str}/(C_{\rm str} - 2 - 3\eta^{\rm iso})\}$. Then there exists $i \in t \cup_{i=1}^{\ell} s_i$ such that $t \cup_{i=1}^{\ell} s_i \subset N_{str}(\tau_i)$ and the panel τ_i is stretched in the sense $h_i > C_{\rm aniso}\rho_i$ required in Assumption* 5.6.

Proof. Let s_j denote the cluster with maximal diameter $\operatorname{diam}(\mathcal{B}_{s_j})$ among the $s_1, \ldots, s_\ell, t$. Due to Lemma 5.9 there is an element $i \in s_j$ with $\operatorname{diam}(\mathcal{B}_{s_j}) \leq \frac{C}{C-1}\operatorname{diam}(\tau_i)$. For any element $\nu \in s \in \{t, s_1, \ldots, s_\ell\}$ we can bound

$$\begin{aligned}\operatorname{dist}(\tau_i, \tau_\nu) &\leq \operatorname{diam}(\mathcal{B}_{s_j}) + \operatorname{dist}(\mathcal{B}_{s_j}, \mathcal{B}_t)\operatorname{diam}(\mathcal{B}_t) + \operatorname{dist}(\mathcal{B}_t, \mathcal{B}_s) + \operatorname{diam}(\mathcal{B}_s) \\ &\leq (3 + 2(\eta^{\rm iso})^{-1})\operatorname{diam}(\mathcal{B}_{s_j}) \;\leq\; (3 + 2(\eta^{\rm iso})^{-1})\frac{C}{C-1}\operatorname{diam}(\tau_i) \\ &\leq (\eta^{\rm iso})^{-1}C_{\rm str}\operatorname{diam}(\tau_i),\end{aligned}$$

i.e., $\nu \in N_{str}(\tau_i)$. Lemma 5.9 gives the bound for h_i. ∎

For the following theorem we use the notation level(t) for the distance of a cluster $t \in T_{\mathcal{I}}$ to the root. The construction of the block cluster tree $T_{\mathcal{I}\times\mathcal{I}}$ ensures that for any block cluster (t, s) there holds level(t) = level(s), and the corresponding boxes $\tilde{\mathcal{B}}_t, \tilde{\mathcal{B}}_s$ used for the clustering are identical up to translation.

THEOREM 5.11. *Let $T := T_{\mathcal{I}\times\mathcal{I}}$ be the block cluster tree constructed in section* 3.1 *from the cluster tree $T_{\mathcal{I}}$, where $n_{\min} \geq C_{\rm ov}$; i.e., a cluster with not more than $C_{\rm ov}$ elements is not further subdivided. We assume that $N^{-g} \lesssim |\xi_i - \xi_j|$. Under Assumption* 5.6 *the following statements hold:*

(a) *The depth of the tree is bounded by*

$$\operatorname{depth}(T) = \mathcal{O}(g \log N).$$

(b) *The sparsity constant is bounded (independently of the mesh parameters g and N) by*

$$C_{\rm sp} = \mathcal{O}(1).$$

Proof. (a) Let $t \in T_{\mathcal{I}}$ be a nonleaf node and ℓ the level of t. We denote the box corresponding to t by $\tilde{\mathcal{B}}_t$, where due to the construction, $\operatorname{diam}(\tilde{\mathcal{B}}_t) \sim 2^{-\ell/d}$. Since t is a nonleaf node, the size of t is at least $\#t > 1$. Let $i \neq j \in t$ be two different indices. Due to the assumed bound $N^{-g} \lesssim |\xi_i - \xi_j|$ we have $N^{-g} \lesssim \operatorname{diam}(\tilde{\mathcal{B}}_t) \lesssim \sqrt{d}2^{-\ell/d}$ so that $\ell \lesssim g \log N$ follows.

(b) We exploit the structure of the regular subdivision of the box $\tilde{\mathcal{B}} = [a_1, b_1) \times [a_2, b_2) \times [a_3, b_3)$ used for the clustering in section 3.1.

1. Let $t \in T_{\mathcal{I}}$ be a node with level$(t) = \ell$ and $\#t > C_{\rm ov}$. We count the number of clusters s with level$(t) = \ell$ and $\#s > C_{\rm ov}$ that are *not admissible* to t. Let

$$C := \max\{2C_{\rm aniso} + 1, C_{\rm str}/(C_{\rm str} - 2 - 3\eta^{\rm iso})\}.$$

Case (1). $\operatorname{diam}(\mathcal{B}_t) \leq C\operatorname{diam}(\tilde{\mathcal{B}}_t)$ and $\operatorname{diam}(\mathcal{B}_s) \leq C\operatorname{diam}(\tilde{\mathcal{B}}_s)$. Since (t, s) is not admissible, the relation

$$\begin{aligned}\operatorname{dist}(\mathcal{B}_t, \mathcal{B}_s) &\leq (\eta^{\rm iso})^{-1} \max\{\operatorname{diam}(\mathcal{B}_t), \operatorname{diam}(\mathcal{B}_s)\} \\ &\leq (\eta^{\rm iso})^{-1} C \max\{\operatorname{diam}(\tilde{\mathcal{B}}_t), \operatorname{diam}(\tilde{\mathcal{B}}_s)\} \;=\; C(\eta^{\rm iso})^{-1}\operatorname{diam}(\tilde{\mathcal{B}}_t)\end{aligned}$$

is valid. On the other hand,

$$\begin{aligned}\operatorname{dist}(\mathcal{B}_t, \mathcal{B}_s) &\geq \operatorname{dist}(\tilde{\mathcal{B}}_t, \tilde{\mathcal{B}}_s) - (\operatorname{diam}(\mathcal{B}_t) + \operatorname{diam}(\mathcal{B}_s))/2 \\ &\geq \operatorname{dist}(\tilde{\mathcal{B}}_t, \tilde{\mathcal{B}}_s) - C\operatorname{diam}(\tilde{\mathcal{B}}_t),\end{aligned}$$

i.e., $\operatorname{dist}(\tilde{\mathcal{B}}_t, \tilde{\mathcal{B}}_s) \leq C(1 + (\eta^{\rm iso})^{-1})\operatorname{diam}(\tilde{\mathcal{B}}_t)$.
Now the distance of the regular boxes $\tilde{\mathcal{B}}_t, \tilde{\mathcal{B}}_s$ can be measured in layers around t (cf. Figure 5.2): There are 3^d boxes touching $\tilde{\mathcal{B}}_t$. By induction it follows that all boxes $\tilde{\mathcal{B}}_r$ with $\operatorname{dist}(\tilde{\mathcal{B}}_t, \tilde{\mathcal{B}}_r) \leq p\operatorname{diam}(\tilde{\mathcal{B}}_t)/(2\sqrt{d})$ are contained in the first p layers, which are at most $(2p + 1)^d$ boxes.
For $p \geq 2\sqrt{d}C(1 + (\eta^{\rm iso})^{-1})$ the box $\tilde{\mathcal{B}}_s$ is included in the first p layers; i.e., s is one of the $(2p + 1)^d = \mathcal{O}(1)$ clusters of the first p layers.

Case (2). Either $\operatorname{diam}(\mathcal{B}_t) > C\operatorname{diam}(\tilde{\mathcal{B}}_t)$ or $\operatorname{diam}(\mathcal{B}_s) > C\operatorname{diam}(\tilde{\mathcal{B}}_s)$. Lemma 5.10 shows that there exists an index i such that t and all clusters s (with the above property) (isotropically) inadmissible to t are contained

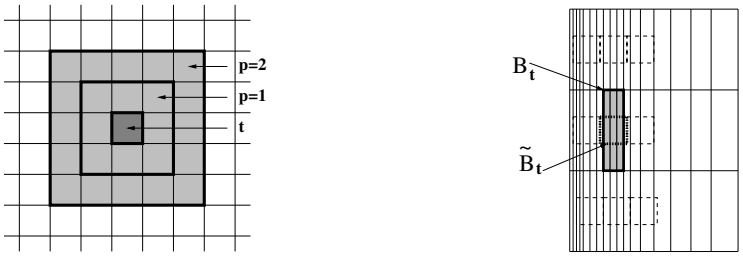


FIG. 5.2. *Left: The first two layers $p = 1$ and $p = 2$ around the box $\tilde{\mathcal{B}}_t$ of t. The first layer consists of 9 boxes, the second one of 25 boxes (it includes the first). Right: A cluster t with box $\tilde{\mathcal{B}}_t$ that is much smaller than the bounding box $\mathcal{B}_t$ of Γ_t. The dashed boxes $\tilde{\mathcal{B}}_s$ correspond to nonempty clusters s.*

in $N_{str}(\tau_i)$ and that τ_i is stretched in the sense $h_i > C_{\mathrm{aniso}}\rho_i$. Assumption 5.6 says that we can find an alignment-classification of $N_{str}(\tau_i)$ with $C_{\mathrm{cl}} = \mathcal{O}(1)$ classes. In each class, the one-dimensional anisotropic admissibility (3.15) applies. In this situation at most $\mathcal{O}((\eta^{\mathrm{ansio}})^{-1})$ clusters s form an inadmissible pair (t, s).

2. As a consequence of 1, the number of nodes $s \in T_{\mathcal{I}}$ (with $\mathrm{level}(s) = \mathrm{level}(t)$, $\#s > C_{\mathrm{ov}}$ and $\#t > C_{\mathrm{ov}}$) not admissible to t is bounded by $\mathcal{O}(1)$.
3. Let $t' \in T_{\mathcal{I}}$ be arbitrary. If t' is the root of $T_{\mathcal{I}}$, then there is exactly one cluster on the same level, namely, t'. Therefore a sparsity constant $C_{\mathrm{sp}} \geq 1$ would be sufficient. If t' is not the root, then the father cluster t of t' fulfils $\#t > n_{\min} \geq C_{\mathrm{ov}}$. Due to 2 we conclude that there are at most $\mathcal{O}(1)$ clusters $s \in T_{\mathcal{I}}$ with $s \times t \in T_{\mathcal{I}\times\mathcal{I}}$, so that there are at most $\mathcal{O}(1)$ clusters $s' \in T_{\mathcal{I}}$ with $s' \times t' \in T_{\mathcal{I}\times\mathcal{I}}$. This is the desired bound for C_{sp}. ∎

The previous theorem gives a rigorous proof that the sparsity constant is independent of N or the geometry and behaves like $\mathcal{O}(1)$ for fixed $\eta^{\mathrm{iso}}, \eta^{\mathrm{aniso}}$ from the admissibility condition and fixed constants $C_{\mathrm{ov}}, C_{\mathrm{str}}, C_{\mathrm{aniso}}, C_{\mathrm{cl}}$ describing the geometry and mesh. For the screen mesh with $\eta^{\mathrm{iso}} = \eta^{\mathrm{aniso}} = 1$ the constants $C_{\mathrm{ov}} := 1$, $C_{\mathrm{str}} := 6$, $C_{\mathrm{aniso}} := 2$, and $C_{\mathrm{cl}} := 12$ fulfil all the requirements.

6. Exact integration.

6.1. The partially integrated kernel: Simple case. In this section we investigate the analytical properties of the partially integrated kernel k^{ad} defined in (3.11). We show that the Assumption 3.10 holds true in the case of the Laplace single layer potential kernel in Example 2.1. Thus we are concerned with the function

$$k^{\mathrm{ad}}_{\mathbf{b}}(x, y) = \frac{1}{4\pi} \int_{\alpha^t}^{\beta^t} \int_{\alpha^s}^{\beta^s} ((y_1 - x_1)^2 + (y - x)^2)^{-1/2} \mathrm{d}y_1 \mathrm{d}x_1. \tag{6.1}$$

A suitable analytic formula is given in the following lemma.

LEMMA 6.1. *For any $x, y \in \mathbb{R}$,*

$$\begin{aligned} k^{\mathrm{ad}}_{\mathbf{b}}(x, y) = {} & \kappa(\alpha^s - \beta^t, x - y) + \kappa(\beta^s - \alpha^t, x - y) \\ & - \kappa(\alpha^s - \alpha^t, x - y) - \kappa(\beta^s - \beta^t, x - y), \end{aligned} \tag{6.2}$$

where

$$\kappa(\xi, z) := \frac{1}{4\pi} \left[\sqrt{\xi^2 + z^2} - \xi \ln \left(\xi + \sqrt{\xi^2 + z^2} \right) \right]. \tag{6.3}$$

Proof. The assertion follows by using

$$\frac{\partial \kappa\,(\xi,z)}{\partial \xi} = -\frac{\ln\left(\xi+\sqrt{z^2+\xi^2}\right)}{4\pi} \quad \text{and} \quad \frac{\partial^2 \kappa\,(\xi,z)}{\partial \xi^2} = -\frac{1}{4\pi\sqrt{z^2+\xi^2}}$$

and some obvious substitutions. □

(Note that this lemma could be generalized to the case of higher order basis functions, in which case an additional polynomial factor appears in (6.1); cf. [17]. The treatment of more general kernel functions such as the kernel for the double layer potential can be treated in the same way because analytic formulae for the antiderivatives with respect to the (x_1, y_1)-coordinates are available.)

Our next result obtains bounds for the derivatives with respect to x of a typical term from the right-hand side of (6.2).

LEMMA 6.2. *There exists an absolute constant C such that*

$$\left|\left\{\frac{\partial^j \kappa}{\partial z^j}\right\}(\xi,z)\right| \;\le\; C\; j!\;|\xi|\;\left(\frac{2}{\sqrt{\xi^2+z^2}}\right)^j \quad \textit{for all} \quad j \ge 1,\ldots$$

for all $\xi, z \in \mathbb{R}$ with $(\xi, z) \neq (0,0)$. (The factor 2 on the right-hand side can be replaced by 1 by a more refined analysis.)

Proof. First consider fixed $\xi \neq 0$, and write $\kappa(\xi,z) = (f(\xi,z) - g(\xi,z))\,/\,(4\pi)$, where

$$f\,(\xi,z) := \sqrt{\xi^2+z^2} \quad \text{and} \quad g(\xi,z) := \xi \ln\left(\xi + f\,(\xi,z)\right).$$

To bound the derivatives of f, note first that $\{\partial^j f/\partial z^j\}\,(\xi,z) = \xi^{1-j}\,\hat{f}^{(j)}\,(z/\xi)$, where $\hat{f}(z) := \sqrt{1+z^2}$. Then for any $z \in \mathbb{R}$, let $\mathcal{D}_z$ denote the disc centered on z with radius $\frac{\sqrt{1+z^2}}{2}$, and let $\mathcal{C}_z$ denote its boundary. Since $\hat{f}$ is analytic in $\mathcal{D}_z$, Cauchy's integral formula yields

$$\frac{1}{j!}\hat{f}^{(j)}\,(z) = \frac{1}{2\pi i}\oint_{\mathcal{C}_z} \frac{\hat{f}\,(t)}{(t-z)^{j+1}}\,dt. \tag{6.4}$$

Elementary but somewhat tedious arguments lead to the estimates

$$\frac{3}{8}\sqrt{1+z^2} \;\le\; \left|\hat{f}\,(t)\right| \;\le\; \frac{\sqrt{13}}{2}\sqrt{1+z^2} \quad \text{for } t \in \mathcal{C}_z, \tag{6.5}$$

and using this in (6.4) yields

$$\left|f^{(j)}\,(z)\right| \;\le\; \frac{\sqrt{13}}{4}\; j!\; 2^{j+1}\; \left\{\sqrt{\xi^2+z^2}\right\}^{1-j}. \tag{6.6}$$

To estimate the derivatives of g, let $\hat{g}\,(z) := \ln(1+\hat{f}\,(z))$, and observe that, for $j \ge 1$, $\left\{\partial^j g/\partial z^j\right\}(\xi,z) = \xi^{1-j}\hat{g}^{(j)}\,(z/\xi)$. Now, to estimate $\hat{g}^{(j)}$ for $j \ge 1$, we shall apply (6.4) with $\hat{f}$ replaced by $\hat{g}'$. To do this note that, for $t \in \mathcal{C}_z$,

$$|\hat{g}'\,(t)| = \left|\left(\frac{t}{\hat{f}\,(t)}\right)\left(\frac{1}{1+\hat{f}\,(t)}\right)\right| \;\le\; \frac{|t-z|+|z|}{|\hat{f}(t)|}\frac{1}{|1+\hat{f}\,(t)\,|} \;\le\; \frac{4}{|1+\hat{f}\,(t)\,|}$$

(where the last inequality follows from (6.5)). Since $\operatorname{Re}\hat{f}(t) > 0$ for all $t \in \mathcal{C}_z$, we have $|1+\hat{f}(t)| \geq |\hat{f}(t)|$ and hence (using also the left-hand side of (6.5))

$$|\hat{g}'(t)| \;\leq\; \frac{4}{|f(t)|} \;\leq\; \frac{32}{3\sqrt{1+z^2}}\,.$$

Using this in (6.4), with $\hat{f}$ replaced by $\hat{g}'$, we obtain, for $j \geq 1$,

$$\left|\hat{g}^{(j)}(z)\right| \;\leq\; \frac{32}{3}\;(j-1)!\;2^{j-1}\;\left\{\sqrt{1+z^2}\right\}^{-j}.$$

For the original function g this implies

$$\left|\left\{\frac{\partial^j g}{\partial z^j}\right\}(\xi,z)\right| \;\leq\; \frac{32}{3}\;(j-1)!\;2^{j-1}\;|\xi|\;\left\{\sqrt{\xi^2+z^2}\right\}^{-j} \quad \text{for} \quad j \geq 1. \tag{6.7}$$

The required result for $\xi \neq 0$ now follows on collection of (6.6) and (6.7). The result for $\xi = 0$ is trivial since then $\kappa(0,z) = |z|$. ∎

The bounds proved in Lemma 6.2 can now be used to bound the error in polynomial approximation of κ.

LEMMA 6.3. *Let $-\infty < a < b < \infty$, and suppose $y \notin [a,b]$. Then, for any $\xi \in \mathbb{R}$ and any $m \geq 1$, the function $\kappa(\xi, \cdot - y)$ can be interpolated on $[a,b]$ by an mth order Chebyshev polynomial $\tilde{\kappa}(\xi, \cdot - y)$ with error bounded by*

$$\|\kappa(\xi,\cdot-y) - \tilde{\kappa}(\xi,\cdot-y)\|_{L^\infty([a,b])} \;\leq\; C|\xi|\,(m+1)^2\,(1+\eta)\left(1+\frac{1}{\eta}\right)^{-m-1},$$

where C is an absolute constant and $\eta = \frac{b-a}{\operatorname{dist}(y,I)}$.

Proof. We apply [2, Theorem 3.2], where the constants C_u, γ_u appearing there are estimated using Lemma 6.2. ∎

THEOREM 6.4. *Let $([\alpha^t,\beta^t]\times[A,B],\ [\alpha^s,\beta^s]\times[A',B'])$ be the domain corresponding to an η^{aniso}-admissible block cluster with respect to* (3.10)*. The integrated kernel $k^{\text{ad}}_{\mathbf{b}} : [A,B]\times[A',B'] \to \mathbb{R}$ (see* (3.11)*) can be approximated by an mth order Chebyshev interpolation $\tilde{k}^{\text{ad}}_{\mathbf{b}}$ with error bounded by*

$$|k^{\text{ad}}_{\mathbf{b}} - \tilde{k}^{\text{ad}}_{\mathbf{b}}|_{[A,B]\times[A',B'],\infty} \;\lesssim (m+1)^2(1+\eta^{\text{aniso}})\left(1+\frac{1}{\eta^{\text{aniso}}}\right)^{-m-1}.$$

Proof. From Lemma 6.3 we get a bound on the interpolation error for each of the four functions on the right-hand side of (6.2). Each of the functions is of the form considered in in Lemma 6.3, where the parameter $\xi \in \{\beta^t-\alpha^s, \beta^t-\beta^s, \alpha^t-\alpha^s, \alpha^t-\beta^s\}$. We can trivially bound $|\xi| \leq \operatorname{diam}\Gamma$. Hence, the assertion is a direct consequence of Lemma 6.3. ∎

6.2. General polyhedral domains. The key idea of the anisotropic panel-clustering relies on the analytic integration of pairs of stretched panels. In the previous section, the relevant antiderivatives of the kernel function have been developed and analyzed for the special case of the screen mesh and piecewise constant shape functions.

In this section, we will present the antiderivatives of the kernel function for general polyhedral domains. We restrict our analysis to piecewise constant elements in order not to overload this paper with technicalities. The generalization to higher order

elements is straightforward because the antiderivatives of the kernel function can be computed analytically also for this case and the resulting integrated kernel has analyticity properties analogous to those for piecewise constant elements.

We consider pairs of panels which are trapezias of the following form. By a suitable translation and rotation of the coordinate system we may assume that one panel has the form

$$t = \left\{(x_1, x_2, 0) \in \mathbb{R}^3 \mid A_t \leq x_2 \leq B_t \text{ and } \alpha_t(x_2) \leq x_1 \leq \beta_t(x_2)\right\},$$

where $0 \leq A_t < B_t$ are constants and α_t, β_t are affine functions which satisfy $\alpha_t(x_2) < \beta_t(x_2)$ for all $0 \leq x_2 \leq B_t$ and $\alpha_t(0) = 0$. The second trapezia is described by

$$s = \left\{(w_1, w_2\cos\varphi, w_2\sin\varphi)^\intercal \mid A_s \leq w_2 \leq B_s \wedge \alpha_s(w_2) \leq w_1 \leq \beta_s(w_2)\right\}.$$

Again, $0 \leq A_s < B_s$ are constants, and α_s, β_s are affine functions which satisfy $\alpha_s(w_2) < \beta_2(w_2)$ for all $0 \leq w_2 \leq B_S$. The compatibility conditions are $\alpha_s(0) = \alpha_t(0) = 0$ and $\beta_s(0) = \beta_t(0)$. The angle $\varphi \in [0, 2\pi[$ is fixed and denotes the angle between t and s with respect to the x_1-axes. The antiderivative of the single layer kernel is given by

$$k_{\mathbf{b}}^{\mathrm{ad}}(x_2, y_2) = \frac{1}{4\pi}\int_{\alpha_s(y_2)}^{\beta_s(y_2)}\int_{\alpha_t(x_2)}^{\beta_t(x_2)} \frac{1}{\sqrt{(x_1 - w_1)^2 + z^2}}\mathrm{d}x_1\mathrm{d}w_1,$$

where $z = z(x_2, y_2) = \sqrt{(x_2 - y_2\cos\varphi)^2 + (y_2\sin\varphi)^2}$. Explicit calculations yield (cf. Lemma 6.1)

$$k_{\mathbf{b}}^{\mathrm{ad}}(x_2, y_2) = \kappa_{\mathbf{b}}(z, d)\big|_{d=\alpha_s(y_2)-\alpha_t(x_2)}^{\alpha_s(y_2)-\beta_t(x_2)} - \kappa_{\mathbf{b}}(z, d)\big|_{d=\beta_s(y_2)-\alpha_t(x_2)}^{\beta_s(y_2)-\beta_t(x_2)}, \tag{6.8}$$

where $\kappa_{\mathbf{b}}$ is as in Lemma 6.1.

Remark 6.5. The qualitative analysis of $k_{\mathbf{b}}^{\mathrm{ad}}$ can be developed along the same lines as for the screen problem. However, the analysis is more subtle since we can assume neither $B_s < A_t$ nor $B_t < A_s$ and the integration bounds in (6.8) have to be taken into account.

7. Numerical experiments. In this section we consider the model problem from Example 2.4 for three different grading parameters $g \in \{1, 3, 5\}$ and an increasing number N of panels, where $N = 9n^2$. We construct the cluster tree $T_{\mathcal{I}}$ by geometrically regular clustering as was introduced in section 3.1, with minimal size $n_{\min} = 4$. The block cluster tree $T_{\mathcal{I}\times\mathcal{I}}$ and corresponding matrix partition is based either on the isotropic admissibility alone or, additionally, on the anisotropic admissibility condition, both with parameters $\eta^{\mathrm{iso}} = \eta^{\mathrm{aniso}} = 3$. For the isotropic admissibility condition we expect the number of nearfield entries to be at least $N^{3/2-1/g}$; cf. Example 3.6.

In the first numerical test we use the isotropic admissibility condition and a blockwise rank of $k = 1$ for all farfield blocks (for a larger rank $k > 1$ one has to multiply the respective numbers in the column "Far" in Table 7.1 by k). We measure the storage requirements in kilobytes per degree of freedom, separately for the nearfield part (isotropically inadmissible leaves of the block cluster tree) and the farfield part (isotropically admissible leaves of the block cluster tree). In Table 7.1 the results are reported along with the sparsity constant C_{sp} (cf. Definition 5.1) from the

(theoretical) complexity estimates. The column "Near" contains the nearfield storage requirements in kilobytes per degree of freedom, and the column "Far" the respective farfield storage requirements. We observe an increase of the nearfield part and the unboundedness of the sparsity constant C_{sp} for $g > 1$ as $n \to \infty$. Since the columns "Far" grow with $O(\log N)$, this shows that the farfield part of the representation is growing with $O(N \log N)$.

TABLE 7.1
Matrix storage requirements using isotropic admissibility.

	$g=1$			$g=3$			$g=5$		
n	Near	Far	C_{sp}	Near	Far	C_{sp}	Near	Far	C_{sp}
100	0.6	1.7	24	0.7	1.6	30	1.0	1.5	30
200	0.6	2.0	24	0.8	2.0	30	1.3	1.9	30
400	0.6	2.3	24	0.8	2.3	36	1.7	2.2	42
800	0.6	2.6	24	1.0	2.6	48	2.4	2.4	66
1600	0.6	2.9	24	1.1	2.8	72	3.3	2.7	120

For the second numerical test we employ the combined isotropic and anisotropic admissibility conditions and measure the storage requirements in kilobyte per degree of freedom, separately for the nearfield part (isotropically and anisotropically inadmissible leaves of the block cluster tree) and the farfield part (isotropically or anisotropically admissible leaves of the block cluster tree). The results in Table 7.2 confirm the theoretical estimates that the storage requirements are $\mathcal{O}(N \log N)$ and the sparsity constant C_{sp} remains bounded, independently of g and N.

TABLE 7.2
Matrix storage requirements using isotropic and anisotropic admissibility.

	$g=1$			$g=3$			$g=5$		
n	Near	Far	C_{sp}	Near	Far	C_{sp}	Near	Far	C_{sp}
100	0.6	1.7	24	0.6	1.7	30	0.8	1.7	30
200	0.6	2.0	24	0.6	2.0	30	0.7	2.1	30
400	0.6	2.3	24	0.6	2.4	30	0.6	2.5	30
800	0.6	2.6	24	0.6	2.7	30	0.6	2.8	30
1600	0.6	2.9	24	0.5	3.0	30	0.6	3.1	30

REFERENCES

[1] O. P. BRUNO AND L. A. KUNYANSKY, *Surface scattering in three dimensions: An accelerated high-order solver*, Proc. Roy. Soc. Lond. A, 457 (2001), pp. 2921–2934.

[2] S. BÖRM AND L. GRASEDYCK, *Low-rank approximation of integral operators by interpolation*, Computing, 72 (2004), pp. 325–332.

[3] S. BÖRM AND L. GRASEDYCK, *Hybrid cross approximation of integral operators*, Numer. Math., 101 (2005), pp. 221–249.

[4] S. BÖRM AND S. SAUTER, *Classical BEM with linear complexity*, Math. Comp., 74 (2005), pp. 1139–1177.

[5] B. CARPENTIERI, I. S. DUFF, L. GIRAUD, AND M. MAGOLU MONGA MADE, *Sparse symmetric preconditioners for dense linear systems in electromagnetism*, Numer. Linear Algebra Appl., 11 (2004), pp. 753–771.

[6] M. COSTABEL, *Boundary integral operators on Lipschitz domains: Elementary results*, SIAM J. Math. Anal., 19 (1988), pp. 613–626.

[7] P. G. CIARLET, *The Finite Element Method for Elliptic Problems*, North–Holland, Amsterdam, 1978.

[8] J. ELSCHNER, *The double layer potential operator over polyhedral domains* II: *Spline Galerkin methods*, Math. Methods Appl. Sci., 15 (1992), pp. 23–37.

[9] V. J. Ervin, E. P. Stephan, and A. Abou El-Seoud, *An improved boundary element method for the charge density of a thin electrified plate in* $\mathbb{R}^3$, Math. Methods Appl. Sci., 13 (1990), pp. 291–303.
[10] K. Giebermann and S. A. Sauter, *manuscript*, 2000.
[11] I. G. Graham, S. A. Sauter, and W. Hackbusch, *Finite elements on degenerate meshes: Inverse-type inequalities and applications*, IMA J. Numer. Anal., 25 (2005), pp. 379–407.
[12] I. G. Graham and W. Mclean, *Anisotropic mesh refinement: The conditioning of Galerkin boundary element matrices and simple preconditioners*, SIAM J. Numer. Anal., 44 (2006), pp. 1487–1513.
[13] I. G. Graham, S. A. Sauter, and W. Hackbusch, *Hybrid Galerkin boundary elements on degenerate meshes*, in Mathematical Aspects of Boundary Element Methods, M. Bonnet, A.-M. Sändig, and W. L. Wendland, eds., CRC Press, London, 1999, pp. 140–164.
[14] L. Grasedyck, *Adaptive recompression of* $\mathcal{H}$*-matrices for BEM*, Computing, 74 (2005), pp. 205–223.
[15] L. Grasedyck and W. Hackbusch, *Construction and arithmetics of* $\mathcal{H}$*-matrices*, Computing, 70 (2003), pp. 295–334.
[16] L. Greengard and V. Rokhlin, *A new version of the fast multipole method for the Laplace equation in three dimensions*, Acta Numer., 6 (1997), pp. 229–269.
[17] W. Hackbusch, *Direct integration of the Newton potential over cubes*, Computing, 68 (2002), pp. 193–216.
[18] W. Hackbusch and Z. Nowak, *On the complexity of the panel method*, in Proceedings of the Conference: Modern Problems in Numerical Analysis, Moscow, 1986 (in Russian).
[19] W. Hackbusch and Z. Nowak, *On the fast matrix multiplication in the boundary element method by panel-clustering*, Numer. Math., 54 (1989), pp. 463–491.
[20] W. Hackbusch and S. Sauter, *On the efficient use of the Galerkin method to solve Fredholm integral equations*, Appl. Math., 38 (1993), pp. 301–322.
[21] W. Hackbusch, *Integral Equations*, Birkhäuser, Basel, 1995.
[22] U. Lamp, T. Schleicher, E. P. Stephan, and W. L. Wendland, *Galerkin collocation for an improved boundary element method for a plane mixed boundary value problem*, Computing, 33 (1984), pp. 269–296.
[23] T. v. Petersdorff, *Elasticity problems in Polyhedra—Singularities and Approximation with Boundary Elements*, Ph.D. Dissertation, Institut Für Angewandte Mathematik, Technische Hochschule Darmstadt, Darmstadt, Germany, 1989.
[24] T. Petersdorff and E. Stephan, *Improved Galerkin Boundary Element Methods for the Mixed Dirichlet-Neumann Problem in Polyhedra*, Technical Report 1202, Technische Hochschule Darmstadt, Darmstadt, Germany, 1989.
[25] J. R. Phillips and J. K. White, *A precorrected-FFT method for electrostatic analysis of complicated* 3*-d structures*, IEEE Trans. Computer-Aided Design Integr. Circuits Syst., 16 (1997), pp. 1059–1072.
[26] V. Rokhlin, *Rapid solutions of integral equations of classical potential theory*, J. Comput. Phys., 60 (1985), pp. 187–207.
[27] S. A. Sauter, *The panel clustering method in* 3*-d BEM*, in Wave Propagation in Complex Media, G. Papanicolaou, ed., IMA Vol. Math. Appl. 96, Springer, Heidelberg, 1997, pp. 199–224.
[28] S. A. Sauter, *Variable order panel clustering*, Computing, 64 (2000), pp. 223–261.
[29] S. A. Sauter and C. Schwab, *Randelementmethoden*, Teubner, Leipzig, 2004.
[30] E. P. Stephan and W. L. Wendland, *An augmented Galerkin procedure for the boundary integral method applied to two-dimensional screen and crack problems*, Appl. Anal., 18 (1984), pp. 183–219.

SIAM J. NUMER. ANAL.
Vol. 46, No. 2, pp. 545–576
© 2008 Society for Industrial and Applied Mathematics

CONVERGENCE ANALYSIS FOR THE NUMERICAL BOUNDARY CORRECTOR FOR ELLIPTIC EQUATIONS WITH RAPIDLY OSCILLATING COEFFICIENTS*

MARCUS SARKIS‡ AND HENRIQUE VERSIEUX†

Abstract. We develop the convergence analysis of a numerical scheme for approximating the solution of the elliptic problem $L_\epsilon u_\epsilon = -\frac{\partial}{\partial x_i} a_{ij}(x/\epsilon)\frac{\partial}{\partial x_j} u_\epsilon = f$ in $\Omega, u_\epsilon = 0$ on $\partial\Omega$, where $a(y) = (a_{ij}(y))$ is a periodic symmetric positive definite matrix and $\Omega = (0,1)^2$. The major goal of the numerical scheme is to capture the ϵ-scale of the oscillations of the solution u_ϵ on a mesh size $h > \epsilon$ (or $h >> \epsilon$). The numerical scheme is based on asymptotic expansions, constructive boundary corrector, and finite element approximations. New a priori error estimates are established for the asymptotic expansions and for the constructive boundary correctors under weak assumptions on the regularity of the problem. These estimates permit to establish sharp finite element error estimates and to consider composite materials applications. Depending on the regularity of the problem, we establish for the numerical scheme a priori error estimates of $O(h^2 + \epsilon^{3/2} + \epsilon h)$ on the L^2-norm, and $O(h + \epsilon^{1+\hat{\delta}})$ for the broken H^1-norm where $\hat{\delta} \in (-\frac{1}{4}, 0]$.

Key words. finite elements, homogenization, elliptic equations, multiscaling, boundary layer, mixed finite elements, discontinuous coefficients, composite materials

AMS subject classifications. 65N30, 35B27

DOI. 10.1137/060654773

1. Introduction.

This paper develops the convergence analysis of the numerical scheme proposed in [46] to approximate the solution of the following problem:

$$L_\epsilon u_\epsilon = -\frac{\partial}{\partial x_i}\left(a_{ij}(x/\epsilon)\frac{\partial}{\partial x_j} u_\epsilon\right) = f \ \text{ in } \ \Omega, \quad u_\epsilon = 0 \ \text{ on } \ \partial\Omega. \tag{1.1}$$

Here, ϵ is a small scale and $a(y) = (a_{ij}(y))$ is a periodic symmetric positive definite matrix with period $Y = (0,1)^2$, and $\Omega = (0,1)^2$. We assume that $a_{ij} \in L^\infty_{\text{per}}(Y)$, i.e., a_{ij} is Y-periodic and $a_{ij} \in L^\infty(\mathbb{R}^2)$, and that there exists a positive constant γ_a such that $\gamma_a \|\xi\|^2 \le a_{ij}(y)\xi_i\xi_j$ for all $\xi \in \mathbb{R}^2$ and $y \in Y$. Throughout the text we consider the Einstein summation convention, i.e., repeated indices indicate summation, and domains are always considered open.

We note that standard finite element methods do not yield good numerical approximations when the mesh size $h > \epsilon$; see [27]. To overcome this, new numerical methods have been recently proposed for solving the problem (1.1) such as the multiscale finite element methods [22, 26, 4, 15, 23], the residual-free bubble function methods [14, 6, 5, 39, 13], and the generalized FEM for homogenization problems [40]. There are also related methods for the case in which the homogenized equation is not known; see the heterogeneous multiscale method [18, 19, 1, 20, 21]. The numerical method considered here, as opposed to the methods in [6, 26, 39, 4, 14], is

*Received by the editors March 21, 2006; accepted for publication (in revised form) June 14, 2007; published electronically February 1, 2008.
http://www.siam.org/journals/sinum/46-2/65477.html
†Instituto Nacional de Matemática Pura e Aplicada, 22460-320, Rio de Janeiro, Brazil (msarkis@impa.br, versieux@impa.br). The first author's research was supported in part by CNPQ (Brazil) under grant 305539/2003-8. The second author's research was supported in part by ANP-PRH32.
‡Department of Mathematical Sciences, Worcester Polytechnic Institute, Worcester, MA 01609 (msarkis@wpi.edu).

based strongly on the asymptotic expansion of u_ϵ. We also explore the periodicity of the matrix a to obtain a very efficient numerical method for approximating u_ϵ.

A theoretical tool commonly used to treat this problem is homogenization theory [8, 9, 10, 11, 16, 28, 37] where, for instance, a first order term expansion for u_ϵ plus a boundary corrector term are considered. Based on such theory, [45, 46] propose numerical schemes to approximate such terms. These numerical methods were designed to work with a mesh size $h > \epsilon$ (or $h >> \epsilon$); however, they also work for the case $h < \epsilon$. In [45] is presented a numerical algorithm for the case in which the domain Ω is a rectangular region, while [46] generalizes the method to the case where the domain Ω is a convex polygon with rational boundary normals.

The convergence analysis for the numerical method is performed in two parts. In the first part we estimate the error between u_ϵ and $u_0 + \epsilon u_1 + \epsilon\phi_\epsilon$ in L^2- and H^1-norms, where ϕ_ϵ is a constructive approximation for the boundary corrector term θ_ϵ; see [3, 35, 36]. To estimate this error, we first estimate the error between u_ϵ and $u_0+\epsilon u_1+\epsilon\theta_\epsilon$ on the H^1- and L^2-norms (see Propositions 3.1 and 3.4) and then estimate the error between ϕ_ϵ and θ_ϵ; see Propositions 3.2 and 3.3. Propositions 3.1 and 3.4 extend some results of [3, 35] using fewer regularity requirements. More specifically, Proposition 3.1 gives the same error estimate of Theorem 2.2 in [3]; however, here we assume $u_0 \in W^{2,p}(\Omega)$ and $\chi^j \in W^{1,q}_{per}(\Omega)$ for $1/p + 1/q \leq 1/2$ and for Lipschitz domains $\Omega \subset \mathbb{R}^{2,3}$, while in Theorem 2.2 in [3] it is assumed that $u_0 \in W^{2,\infty}(\Omega)$ and $\chi^j \in H^1_{per}(Y)$ for Lipschitz domains $\Omega \subset \mathbb{R}^d$, d a natural number. We also note that Propositions 3.1 and 3.4 generalize, respectively, Propositions 2.1 and 2.3 of [35]. In Proposition 3.1, we assume that $a_{ij} \in L^\infty_{per}(Y)$, $u_0 \in W^{2,p}(\Omega)$, and $\chi^j \in W^{1,q}_{per}(\Omega)$ for $1/p + 1/q \leq 1/2$, and $\Omega \subset \mathbb{R}^{2,3}$ is a Lipschitz domain, while in Proposition 2.1 in [35] it is assumed that $a_{ij} \in C^{1,\beta}_{per}(Y)$, $u_0 \in H^2(\Omega)$, and $\Omega \subset \mathbb{R}^2$ is a smooth domain. Proposition 3.4 assumes that $a_{ij} \in L^\infty_{per}(Y)$, $u_0 \in W^{3,p}(\Omega)$, χ^j and $\chi^{ij} \in W^{1,q}_{per}(\Omega)$ for $1/p + 1/q \leq 1/2$, $\Omega \subset \mathbb{R}^{2,3}$ is a Lipschitz domain, while Proposition 2.3 in [35] assumes that $a_{ij} \in C^{1,\beta}_{per}(Y)$, $u_0 \in H^3(\Omega)$, and $\Omega \subset \mathbb{R}^2$ is smooth. The importance of considering a theory that handles the case $a_{ij} \in L^\infty_{per}(Y)$ comes from applications to composite materials where the coefficients a_{ij} are often piecewise constants; see also Theorem 1.1 in [31] which gives conditions on the discontinuities of the functions a_{ij} so that χ^j and $\chi^{ij} \in W^{1,\infty}_{per}(Y)$. We also observe that Proposition 2.1 in [35] is used in the convergence analysis of the numerical methods presented in [22, 27, 39], and therefore, the analysis presented here can be used to extend the analysis of these numerical methods with fewer regularity requirements on a, u_0, or Ω. To the best of our knowledge, Propositions 3.2 and 3.3 have not been considered in the literature; however, a technique developed in [35] is used in part of the proof of Proposition 3.2.

In the second part of this paper we develop the finite element analysis to estimate the a priori error estimates due to the finite element approximations. One difficulty lies in the fact that we need a discrete approximation for $\partial_\eta u_0$ to define the Dirichlet boundary condition for the discrete boundary corrector problem; If u_0^h is a finite element approximation for u_0, then $\partial_\eta u_0^h$ does not necessarily belong to $H^{1/2}(\partial\Omega)$, i.e., the space required to define the boundary corrector terms. To overcome this, we introduce a Lagrange multiplier space to approximate $\partial_\eta u_0$ and to develop error estimates between $\partial_\eta u_0$ and its discrete approximation μ_h using fractional-order Sobolev norms. The Lagrange multiplier plays an important role in obtaining convergence rates under weak assumptions on the regularity of u_0 and χ^j.

We now introduce some norms and semi-norms. Let $B \subset \mathbb{R}^{2,3}$ be a Lipschitz

domain and define

$$\|v\|_{m,\infty,B} = \max_{|\alpha|\le m}\left\{\text{ess. sup}_{x\in B}|\partial^\alpha v(x)|\right\} \quad \text{and} \quad |v|_{m,\infty,B} = \max_{|\alpha|=m}\left\{\text{ess. sup}_{x\in B}|\partial^\alpha v(x)|\right\},$$

and for $1 \le q < \infty$

$$\|v\|_{m,q,B} = \left(\int_B \sum_{|\alpha|\le m} |D^\alpha v|^q dx\right)^{1/q} \quad \text{and} \quad |v|_{m,q,B} = \left(\int_B \sum_{|\alpha|=m} |D^\alpha v|^q dx\right)^{1/q}.$$

We also define the nonconforming norms related to a partition $\mathcal{T}_h = K_1, K_2, \ldots, K_N$ of B by

$$\|v\|_{m,h} = \sqrt{\sum_{K_j\in\mathcal{T}_h} \|v\|^2_{H^m(K_j)}}.$$

Throughout this paper when we do not make reference to the domain B or to the norm q, it is because $B = \Omega$ or $q = 2$, respectively. In what follows c denotes a generic constant independent of ϵ and mesh parameters.

This paper is organized as follows: we devote section 2 to introducing the asymptotic expansion of u_ϵ and to describing the continuous approximation for the boundary corrector terms; in section 3 we develop the main results and proofs associated to error estimates due to the asymptotic expansion approximation. In section 4 we consider the numerical algorithm introduced in [46] for the case $\Omega = (0,1)^2$, while in section 5 we analyze the discretization errors due to the finite element approximation. In section 6 we test numerically the numerical schemes, and in section 7 we make some conclusions.

2. Periodic homogenization and asymptotic expansions.

2.1. Asymptotic expansions.

Consider the following ansatz:

$$u_\epsilon(x) = u_0(x, x/\epsilon) + \epsilon u_1(x, x/\epsilon) + \epsilon^2 u_2(x, x/\epsilon) + \cdots, \tag{2.1}$$

where the functions $u_j(x, y)$ are Y periodic in y. Using (2.1) in (1.1) and matching the terms with the same order in ϵ, one may define functions u_j such that $u_0(x, x/\epsilon) + \epsilon u_1(x, x/\epsilon) + \epsilon^2 u_2(x, x/\epsilon)$ approximates u_ϵ as

$$\|u_\epsilon(x) - u_0(x, x/\epsilon) - \epsilon u_1(x, x/\epsilon)\|_1 \le c\epsilon^{1/2}\|u_0\|_{2,\infty},$$

where $u_0 \in C^2(\Omega)$ and $\chi^j \in W^{1,\infty}(Y)$, and the constant c depends only on a, χ^j, and $\Omega \subset \mathbb{R}^{2,3}$ a Lipschitz domain. These terms are defined in detail below; for more details, including the proof of the above inequality, see [11, 28].

Let $\chi^j \in H^1_{\text{per}}(Y)$, i.e., $\chi^j \in H^1_{loc}(\mathbb{R}^{2,3})$ and Y-periodic, be the weak solution with zero average over Y of

$$\nabla_y \cdot a(y)\nabla_y \chi^j = \nabla_y \cdot a(y)\nabla_y y_j = \frac{\partial}{\partial y_i} a_{ij}(y), \tag{2.2}$$

and define the matrix

$$A_{ij} = \frac{1}{|Y|}\int_Y a_{lm}(y)\frac{\partial}{\partial y_l}(y_i - \chi^i)\frac{\partial}{\partial y_m}(y_j - \chi^j)dy. \tag{2.3}$$

It is easy to check that the matrix A is symmetric positive definite. Define $u_0 \in H_0^1(\Omega)$ as the weak solution of

$$-\nabla.A\nabla u_0 = f \ \text{ in } \ \Omega, \qquad u_0 = 0 \ \text{ on } \ \partial\Omega, \tag{2.4}$$

and let

$$u_1\left(x, \frac{x}{\epsilon}\right) = -\chi^j\left(\frac{x}{\epsilon}\right)\frac{\partial u_0}{\partial x_j}(x). \tag{2.5}$$

Note that $u_0 + \epsilon u_1$ does not satisfy the zero Dirichlet boundary condition on $\partial\Omega$. To overcome this, the boundary corrector term $\theta_\epsilon \in H^1(\Omega)$ is introduced and defined as the solution of

$$-\nabla \cdot a(x/\epsilon)\nabla\theta_\epsilon = 0 \ \text{ in } \ \Omega, \qquad \theta_\epsilon = -u_1\left(x, \frac{x}{\epsilon}\right) \ \text{ on } \ \partial\Omega. \tag{2.6}$$

Hence, we have that $u_0 + \epsilon u_1 + \epsilon\theta_\epsilon \in H_0^1(\Omega)$. Propositions 3.1 and 3.4 provide error estimates between u_ϵ and $u_0 + \epsilon u_1 + \epsilon\theta_\epsilon$ in the norms $\|\cdot\|_1$ and $\|\cdot\|_0$, respectively.

We also define the term u_2 (see (2.8)), which is used in the proof of Proposition 3.4. Set

$$b_{ij} = -a_{ij} + a_{im}\frac{\partial\chi^j}{\partial y_m} + \frac{\partial}{\partial y_m}(a_{mi}\chi^j)$$

and observe that $\overline{b}_{ij} = A_{ij}$, where $\overline{b}_{ij} = \int_Y b_{ij}dy$. Define $\chi^{ij} \in H^1_{per}(Y)$ as the weak solution with zero average over Y of

$$\nabla_y \cdot a\nabla_y\chi^{ij} = b_{ij} - \overline{b}_{ij}, \tag{2.7}$$

and let

$$u_2\left(x, \frac{x}{\epsilon}\right) = -\chi^{ij}\left(\frac{x}{\epsilon}\right)\frac{\partial^2 u_0}{\partial x_i \partial x_j}(x). \tag{2.8}$$

Note that $u_0 + \epsilon u_1 + \epsilon\theta_\epsilon + \epsilon^2 u_2$ does not satisfy the zero Dirichlet boundary condition on $\partial\Omega$ and to overcome this the boundary corrector term $\varphi_\epsilon \in H^1(\Omega)$ is introduced as the weak solution of

$$-\nabla \cdot a(x/\epsilon)\nabla\varphi_\epsilon = 0 \ \text{ in } \Omega, \ \text{ and } \ \varphi_\epsilon(x) = -u_2(x, x/\epsilon) \ \text{ on } \ \partial\Omega. \tag{2.9}$$

2.2. The boundary corrector terms. When dealing with constructive boundary corrector terms, we assume that $\Omega = (0,1)^2$, although the same theory holds for the case $\Omega = \prod_{i=1}^2 (a_i, b_i)$, $a_i < b_i \in \mathbb{R}$, or convex polygon or polyhedron domains with rational boundary normals; see [46, 44].

The coefficients $a_{ij}(x/\epsilon)$ and the Dirichlet boundary data $-u_1(x, \frac{x}{\epsilon})$ in (2.6) are highly oscillatory; thus it is not a trivial problem to obtain a good discrete approximation for θ_ϵ. We consider an analytical approximation for θ_ϵ, denoted by ϕ_ϵ, which satisfies the oscillating boundary condition and is suitable for numerical approximation; see [3, 35]. Analytical and numerical approximations for φ_ϵ will not be considered in this paper.

Denote η as the unity outward normal vector to $\partial\Omega$ and let $\partial_\eta u_0$ be the unity outward normal derivative of u_0 on $\partial\Omega$. In order to define the approximation ϕ_ϵ for θ_ϵ, we first introduce a decomposition for $\theta_\epsilon = \tilde{\theta}_\epsilon + \bar{\theta}_\epsilon$ as

$$-\nabla \cdot a(x/\epsilon)\nabla\tilde{\theta}_\epsilon = 0 \ \text{ in } \ \Omega, \quad \tilde{\theta}_\epsilon = \left(\chi^j\left(\frac{x}{\epsilon}\right)\eta_j - \chi^*\right)\partial_\eta u_0 \ \text{ on } \ \partial\Omega \tag{2.10}$$

and

$$-\nabla \cdot a(x/\epsilon)\nabla \bar{\theta}_\epsilon = 0 \ \text{ in } \ \Omega, \qquad \bar{\theta}_\epsilon = \chi^* \partial_\eta u_0 \ \text{ on } \ \partial\Omega, \tag{2.11}$$

where $\chi^*|_{\Gamma_k} = \chi_k^*$, $k \in \{e, w, n, s\}$, are properly chosen constants defined in subsection 2.2.1, and $\Gamma_e = \{1\} \times [0,1]$, $\Gamma_w = \{0\} \times [0,1]$, $\Gamma_n = [0,1] \times \{1\}$, and $\Gamma_s = [0,1] \times \{0\}$ are the edges of the domain $\Omega = (0,1)^2$. In Remark 2.1 we show that $\chi^* \partial_\eta u_0$ and $\chi^j(\frac{x}{\epsilon})\eta_j \partial_\eta u_0$ belong to $H^{1/2}(\partial\Omega)$; therefore, problems (2.10) and (2.11) are well posed. Later in this section we define the functions $\tilde{\phi}_\epsilon$ and $\bar{\phi}_\epsilon$, which are the approximations for $\tilde{\theta}_\epsilon$ and $\bar{\theta}_\epsilon$, respectively, and define $\phi_\epsilon = \tilde{\phi}_\epsilon + \bar{\phi}_\epsilon$.

Remark 2.1. Let $\Omega \subset \mathbb{R}^2$ be a convex polygon and assume $u_0 \in H^2(\Omega) \cap H_0^1(\Omega)$. We have by Theorem A.2 in [38] that $\partial_\eta u_0|_{\Gamma_k} \in H_{00}^{1/2}(\Gamma_k)$ and $\|\partial_\eta u_0\|_{H_{00}^{1/2}(\Gamma_k)} \leq c\|u_0\|_2$; therefore,

$$\|\chi^* \partial_\eta u_0\|_{H^{1/2}(\partial\Omega)} \leq c \max_k |\chi_k^*| \|u_0\|_2,$$

where $|\chi_k^*|$ can be bounded by $c \max_j \|\chi^j\|_{1,2,Y}$; see the proof of Lemma 4.4 in [3]. Note also that $u_1(x, \frac{x}{\epsilon}) = -\chi^j\left(\frac{x}{\epsilon}\right)\frac{\partial u_0}{\partial x_j}(x)$ and $\frac{\partial u_1}{\partial x_l} = -(\frac{\partial \chi^j}{\partial x_l})\frac{\partial u_0}{\partial x_j} - \chi^j(\frac{\partial^2 u_0}{\partial x_l \partial x_j})$. If we assume $u_0 \in W^{2,p}(\Omega)$ and $\chi^j \in W_{per}^{1,q}(Y)$ for $\frac{1}{p} + \frac{1}{q} \leq \frac{1}{2}$, $2 \leq p, q \leq \infty$, then $u_1 \in H^1(\Omega)$. In addition, using $u_0 \in H_0^1(\Omega)$ and density arguments we also have $u_1|_{\Gamma_k} \in H_{00}^{1/2}(\Gamma_k)$.

2.2.1. Calculating the constants χ_k^*. We now define properly the constants χ_k^* in order to make the function $\tilde{\phi}_\epsilon$ decay exponentially to zero away from the boundary $\partial\Omega$ and to satisfy the Dirichlet boundary condition $\tilde{\phi}_\epsilon(x) = -u_1(x, \frac{x}{\epsilon}) - \chi^* \partial_\eta u_0(x)$ for $x \in \partial\Omega$.

For $k \in \{e, w, n, s\}$ let $G_k = I_k^1 \times I_k^2$, where $I_e^1 = I_n^2 = (-\infty, 0)$, $I_e^2 = I_w^2 = (0,1)$, $I_w^1 = I_s^2 = (0, \infty)$, and $I_n^1 = I_s^1 = (0,1)$. Associated to each side Γ_k of Ω define the function v_k as the solution of

$$\begin{aligned} &-\nabla_y \cdot a(y_1, y_2)\nabla_y v_k = 0 \ \text{ in } \ G_k, \\ &\text{and } \ \partial_{y_i} v_n \exp(\gamma(-1)^{m_k} y_{n_k}) \in L^2(G_k), \quad i = 1, 2, \end{aligned}$$

where $n_k = 1$ if $k \in \{e, w\}$, $n_k = 2$ if $k \in \{s, n\}$, $m_k = 1$ if $k \in \{e, n\}$, and $m_k = 2$ if $k \in \{w, s\}$. For each $k \in \{e, w, s, n\}$ we impose the following boundary conditions:

$$\begin{aligned} &v_e(0, y_2) = \chi^1(1/\epsilon, y_2) \ \text{ for } 0 < y_2 < 1, && v_w(0, y_2) = -\chi^1(0, y_2) \ \text{ for } 0 < y_2 < 1, \\ &v_e(y_1, \cdot) \ \ [0,1]\text{-periodic for } -\infty < y_1 < 0, && v_w(y_1, \cdot) \ \ [0,1]\text{-periodic for } 0 < y_1 < \infty, \\ &v_n(y_1, 0) = \chi^2(y_1, 1/\epsilon) \ \text{ for } \ 0 < y_1 < 1, && v_s(y_1, 0) = -\chi^2(y_1, 0) \ \text{ for } \ 0 < y_1 < 1, \\ &v_n(\cdot, y_2) \ \ [0,1]\text{-periodic for } -\infty < y_2 < 0, && v_s(\cdot, y_2) \ \ [0,1]\text{-periodic for } 0 < y_2 < \infty. \end{aligned}$$

These boundary layer problems have been studied by several authors; see [30, 32, 28, 35]. Theorem 10.1 in section 10.4 in [32] guarantees existence and uniqueness of v_k and χ_k^*.

2.2.2. Approximating $\tilde{\theta}_\epsilon$. We note by Remark 2.1 that $(u_1(x, \frac{x}{\epsilon}) - \chi^* \partial_\eta u_0)|_{\Gamma_k} \in H_{00}^{1/2}(\Gamma_k)$. Thus, we can split $\tilde{\theta}_\epsilon = \sum_{k \in \{e,w,n,s\}} \tilde{\theta}_\epsilon^k$, where

$$L_\epsilon \tilde{\theta}_\epsilon^k = 0 \ \text{ in } \ \Omega, \quad \text{and} \quad \tilde{\theta}_\epsilon^k = \begin{cases} -u_1(x, \frac{x}{\epsilon}) - \chi_k^* \partial_\eta u_0 & \text{on } \ \Gamma_k, \\ 0 & \text{on } \ \partial\Omega \setminus \Gamma_k. \end{cases} \tag{2.12}$$

We propose to approximate $\tilde{\theta}_\epsilon^k$ by

$$\tilde{\phi}_\epsilon^k(x_1,x_2) = \varphi_k(x_1,x_2)\left(v_k\left(\frac{x-\delta_k}{\epsilon}\right) - \chi_k^*\right)\frac{\partial u_0}{\partial x_{i_k}}(x_1,x_2), \tag{2.13}$$

where $\delta_e = (1,0)$, $\delta_n = (0,1)$, $\delta_w = (0,0)$, and $\delta_s = (0,0)$. The cutting-off functions φ_k are nonnegative smooth functions satisfying $\varphi_k(x_1,x_2) = \varphi_k(x_1)$ for $k \in \{e,w\}$, $\varphi_k(x_1,x_2) = \varphi_k(x_2)$ for $k \in \{n,s\}$, and

$$\varphi_e(s) = \varphi_n(s) = \begin{cases} 1 & \text{if } s \in [2/3,1], \\ 0 & \text{if } s \in [0,1/3], \end{cases} \qquad \varphi_w(s) = \varphi_s(s) = \begin{cases} 0 & \text{if } s \in [2/3,1], \\ -1 & \text{if } s \in [0,1/3]. \end{cases}$$

Therefore,

$$\tilde{\phi}_\epsilon = \sum_{k\in\{e,w,n,s\}} \tilde{\phi}_\epsilon^k \tag{2.14}$$

approximates $\tilde{\theta}_\epsilon$ with $\tilde{\phi}_\epsilon = \tilde{\theta}_\epsilon$ on the boundary of Ω.

2.2.3. Approximating $\bar{\theta}_\epsilon$. The boundary condition imposed on (2.11) does not depend on ϵ. An effective approximation for $\bar{\theta}_\epsilon$ is given by $\bar{\phi} \in H^1(\Omega)$, the weak solution of

$$-\nabla \cdot A\nabla\bar{\phi} = 0 \ \text{ in } \ \Omega, \qquad \bar{\phi} = \chi^* \partial_\eta u_0 \ \text{ on } \ \partial\Omega, \tag{2.15}$$

where Remark 2.1 guarantees the well posedness of this problem.

2.2.4. Approximating u_ϵ. We finally define the analytical approximation for u_ϵ as $u_0 + \epsilon u_1 + \epsilon\phi_\epsilon$, where

$$\phi_\epsilon = \tilde{\phi}_\epsilon + \bar{\phi}. \tag{2.16}$$

Note that $\phi_\epsilon|_{\partial\Omega} = \theta_\epsilon|_{\partial\Omega}$; therefore, $u_0 + \epsilon u_1 + \epsilon\phi_\epsilon = 0$ on $\partial\Omega$.

3. Convergence results of asymptotic expansions. We now prove the propositions that are used in the proofs of Theorems 3.1 and 3.2.

Proposition 3.1 generalizes Theorem 2.2 in [3]. Here we assume that $a_{ij} \in L^\infty_{per}(Y)$, $u_0 \in W^{2,p}(\Omega)$, and $\chi^j \in W^{1,q}_{per}(\Omega)$ for $1/p + 1/q \le 1/2$, while Theorem 2.2 in [3] assumes that $a_{ij} \in L^\infty_{per}(Y)$, $u_0 \in W^{2,\infty}(\Omega)$, and $\chi^j \in H^1_{per}(\Omega)$. Proposition 3.1 also generalizes Proposition 2.1 in [35], where it is assumed that $a_{ij} \in C^{1,\beta}_{per}(Y)$, $u_0 \in H^2(\Omega)$, and $\Omega \subset \mathbb{R}^2$ is a smooth domain. We also refer the reader to Theorem 1.1 in [31] for results associated to sufficient conditions on a_{ij} in order to have $\chi^j \in W^{1,\infty}_{per}(Y)$.

PROPOSITION 3.1. *Let $\Omega \subset \mathbb{R}^d$, $d = 2,3$, be a bounded domain with a Lipschitz boundary. Let u_ϵ be the solution of the problem* (1.1) *and u_0, u_1, and θ_ϵ be defined by* (2.4), (2.5), *and* (2.6), *respectively. Assume $a_{ij} \in L^\infty_{per}(Y)$, $u_0 \in W^{2,p}(\Omega)$, and $\chi^j \in W^{1,q}_{per}(Y)$ for $1/p + 1/q \le 1/2$, $2 \le p,q \le \infty$. Then there exists a constant c, independent of u_0, χ^j, and ϵ, such that*

$$\|u_\epsilon(\cdot) - u_0(\cdot) - \epsilon u_1(\cdot,\cdot/\epsilon) - \epsilon\theta_\epsilon(\cdot)\|_1 \le c\epsilon \max_j \|\chi^j\|_{1,q,Y}\|u_0\|_{2,p}. \tag{3.1}$$

In addition, if $p,q > d$, then

$$\|u_\epsilon - u_0\|_0 \le c\epsilon \max_j \|\chi^j\|_{1,q,Y}\|u_0\|_{2,p}. \tag{3.2}$$

Proof. Define

$$v_0(x,y) = a(y)\nabla_x u_0(x) + a(y)\nabla_y u_1(x,y) = a(y)(\nabla_y y_j - \nabla_y \chi^j(y))\frac{\partial u_0}{\partial x_j}(x). \tag{3.3}$$

From the definition of χ^j (see (2.2)) we have

$$\int_Y \left(a(y)(e_j - \nabla_y \chi^j(y)) - Ae_j\right)\nabla_y \phi(y)dy = 0 \ \ \forall\, \phi \in H^1_{per}(Y).$$

Since the vector $a(y)(e_j - \nabla_y \chi^j(y)) - Ae_j$ is Y periodic and has zero average entries over Y, then the Lemma 3.1 guarantees the existence of a $\phi_j(y) \in H^1_{per}(Y)$ with zero average over Y and such that

$$a(y)(\nabla_y y_j - \nabla_y \chi^j(y)) - Ae_j = -curl_y \phi_j(y). \tag{3.4}$$

Let

$$\phi(x,y) = \phi_j(y)\frac{\partial u_0}{\partial x_j}(x). \tag{3.5}$$

For the case $d = 2$ define

$$\begin{aligned} v_1(x,y) &= -curl_x \phi(x,y) \\ &= \begin{pmatrix} -\phi_j(y)\frac{\partial^2 u_0}{\partial x_2 \partial x_j}(x) \\ \phi_j(y)\frac{\partial^2 u_0}{\partial x_1 \partial x_j}(x) \end{pmatrix}, \end{aligned}$$

and note that $|curl_y \phi_j|_{0,q} = |\phi_j|_{1,q}$. Since $\chi^j \in W^{1,q}_{per}(Y)$ and ϕ_j have zero average over Y, we apply a Poincaré inequality to obtain

$$\|\phi_j\|_{1,q,Y} \le c|curl_y \phi_j|_{0,q,Y} \le c(\|\chi^1\|_{1,q,Y} + \|\chi^2\|_{1,q,Y}).$$

For the case $d = 3$, the Remark 3.11 in [24] guarantees the existence of $\phi_j \in W^{1,q}_{per}(Y)^3$. Since $\chi^j \in W^{1,q}_{per}(Y)$ and $u_0 \in W^{2,p}(\Omega)$ for $1/p + 1/q \le 1/2$, we obtain $v_1(x, x/\epsilon) \in L^2(\Omega)$ and is estimated by $\|v_1\|_0 \le c\sum_{j=1}^d \|\chi^j\|_{1,q,Y}\|u_0\|_{2,p}, d = 2, 3$. Moreover, by Lemma 3.1, we have

$$\nabla_x \cdot v_1(x,y) = 0, \tag{3.6}$$

and using elementary calculations we obtain

$$\begin{aligned} \nabla_y \cdot v_1(x,y) &= \nabla_y \cdot curl_x\left(\phi_j(y)\partial_{x_j} u_0(x)\right) \\ &= -\nabla_x \cdot curl_y\left(\phi_j(y)\partial_{x_j} u_0(x)\right) \\ &= -\nabla_x \cdot v_0(x,y) - f. \end{aligned} \tag{3.7}$$

Let

$$z_\epsilon(x) = u_\epsilon(x) - u_0(x) - \epsilon u_1(x, x/\epsilon)$$

and

$$\eta_\epsilon(x) = a(x/\epsilon)\nabla u_\epsilon(x) - v_0(x, x/\epsilon) - \epsilon v_1(x, x/\epsilon). \tag{3.8}$$

We have

$$\begin{aligned}
&a(x/\epsilon)\nabla z_\epsilon(x) - \eta_\epsilon(x)\\
&\quad = a(x/\epsilon)\nabla u_\epsilon(x) - a(x/\epsilon)\nabla_x u_0(x) - \epsilon a(x/\epsilon)\nabla_x u_1(x, x/\epsilon)\\
&\qquad - a(x/\epsilon)\nabla_y u_1(x, x/\epsilon) - a(x/\epsilon)\nabla u_\epsilon(x) + v_0(x, x/\epsilon) + \epsilon v_1(x, x/\epsilon)\\
&\quad = \epsilon(v_1(x, x/\epsilon) - a(x/\epsilon)\nabla_x u_1(x, x/\epsilon)),
\end{aligned}$$

and using the notation $a^\epsilon(x) = a(x/\epsilon)$ we obtain

$$\|a^\epsilon \nabla z_\epsilon - \eta_\epsilon\|_0 \leq \epsilon \|v_1(\cdot, \cdot/\epsilon) - a^\epsilon(\cdot)\nabla_x u_1(\cdot, \cdot/\epsilon)\|_0. \tag{3.9}$$

Given $g \in L^2(\Omega)$, let $w_\epsilon \in H_0^1(\Omega)$ be the solution of

$$\int_\Omega a(x/\epsilon)\nabla w_\epsilon(x)\nabla\psi(x)dx = \int_\Omega g(x)\psi(x)dx \quad \forall \psi \in H_0^1(\Omega). \tag{3.10}$$

Hence, using (2.6), we obtain

$$\begin{aligned}
\int_\Omega g(z_\epsilon - \epsilon\theta_\epsilon)dx &= \int_\Omega a^\epsilon \nabla w_\epsilon \cdot \nabla(z_\epsilon - \epsilon\theta_\epsilon)dx,\\
\int_\Omega a^\epsilon \nabla w_\epsilon \cdot \nabla z_\epsilon dx - \epsilon \int_\Omega a^\epsilon \nabla w_\epsilon \cdot \nabla \theta_\epsilon dx &= \int_\Omega a^\epsilon \nabla w_\epsilon \cdot \nabla z_\epsilon dx. \qquad (3.11)
\end{aligned}$$

Now observe that

$$\int_\Omega a^\epsilon \nabla w_\epsilon \cdot \nabla z_\epsilon dx = \int_\Omega \nabla w_\epsilon \cdot (a^\epsilon \nabla z_\epsilon - \eta_\epsilon)dx + \int_\Omega \eta_\epsilon \cdot \nabla w_\epsilon dx. \tag{3.12}$$

In order to estimate the second term on the right-hand side of (3.12) we apply the definition of η_ϵ (see (3.8)) to obtain

$$\begin{aligned}
\int_\Omega \eta_\epsilon \cdot \nabla w_\epsilon dx &= \int_\Omega (a(x/\epsilon)\nabla u_\epsilon(x) - v_0(x, x/\epsilon) - \epsilon v_1(x, x/\epsilon)) \cdot \nabla w_\epsilon(x)dx\\
&= \int_\Omega f w_\epsilon dx - \int_\Omega (v_0(x, x/\epsilon) - \epsilon v_1(x, x/\epsilon)) \cdot \nabla w_\epsilon(x)dx. \qquad (3.13)
\end{aligned}$$

We note that

$$\begin{aligned}
\int_\Omega v_1(x, x/\epsilon) \cdot \nabla w_\epsilon(x)dx &= \int_\Omega \nabla \cdot v_1(x, x/\epsilon) w_\epsilon(x)dx\\
= \int_\Omega (\nabla_x + 1/\epsilon \nabla_y) \cdot v_1(x, y)|_{(y=x/\epsilon)} w_\epsilon(x)dx &= -\frac{1}{\epsilon}\int_\Omega (\nabla_x \cdot v_0 + f) w_\epsilon dx, \qquad (3.14)
\end{aligned}$$

where we have used (3.6) and (3.7) to obtain (3.14). Using the definition of v_0 (see (3.3)) we have

$$\int_\Omega v_0(x, x/\epsilon) \cdot \nabla w_\epsilon(x)dx = \int_\Omega a(x/\epsilon)(e_j - \nabla_y \chi^j(x/\epsilon))\frac{\partial u_0}{\partial x_j}(x) \cdot \nabla w_\epsilon(x)dx,$$

and by the chain rule we obtain

$$\begin{aligned}
\int_\Omega v_0(x, x/\epsilon) \cdot \nabla w_\epsilon dx &= \int_\Omega a(x/\epsilon)(e_j - \nabla_y \chi^j(x/\epsilon)) \cdot \nabla \left(\frac{\partial u_0}{\partial x_j} w_\epsilon(x)\right) dx\\
&\quad - \int_\Omega a(x/\epsilon)(e_j - \nabla_y \chi^j(x/\epsilon)) \cdot \left(w_\epsilon \nabla \frac{\partial u_0}{\partial x_j}(x)\right) dx. \qquad (3.15)
\end{aligned}$$

In this paragraph we evaluate the first term on the right-hand side of (3.15). Let $(\frac{\epsilon}{3}Y_i)_{i=1,\dots,i_m}$ be a finite set of translated cells of $\frac{\epsilon}{3}Y$, covering $\overline{\Omega}$, and consider a partition of unity ρ_i, such that $\text{supp}\rho_i \subset \frac{2\epsilon}{3}Y_i$, where $\frac{2\epsilon}{3}Y_i$ denotes the cell $\frac{2\epsilon}{3}Y$ centered at $\frac{\epsilon}{3}Y_i$. We note that

$$\text{supp}(\rho_i w_\epsilon) \subset \frac{2\epsilon}{3}Y_i \cap \overline{\Omega} \subset \epsilon Y_i. \tag{3.16}$$

Thus,

$$\begin{aligned}&\int_\Omega a(x/\epsilon)(e_j - \nabla_y \chi^j(x/\epsilon)) \cdot \nabla \left(\frac{\partial u_0}{\partial x_j} w_\epsilon(x)\right) dx \\ &= \sum_{i=1:i_m} \int_{\epsilon Y_i} a(x/\epsilon)(e_j - \nabla_y \chi^j(x/\epsilon)) \cdot \nabla \left(\rho_i \frac{\partial u_0}{\partial x_j} w_\epsilon(x)\right) dx = 0.\end{aligned} \tag{3.17}$$

Here we have used that for Lipschitz domains, the function u_0 defined in (2.4) has a stable extension to $W^{2,p}(\mathbb{R}^d)$ (see [43]) which we have also denoted by u_0. We also note that from (3.16) it follows that the function $\rho_i \partial_{x_j} u_0 w_\epsilon$ is well defined and vanishes outside of Ω. In addition, since $1/p + 1/q \le 1/2$, we have $\rho_i \partial_{x_j} u_0 w_\epsilon \in W^{1,q'}(\mathbb{R}^d)$ for $1/q' = 1 - 1/q$. In addition, $\chi^j \in W^{1,q}_{per}(Y)$ and (2.2) imply that

$$\int_Y a_{lm}(y) \partial_{y_l}(\chi^j - y_j) \partial_{y_m} \psi = 0 \quad \forall\, \psi \in W^{1,q'}_{per}(Y).$$

Therefore, since $\rho_i \partial_{x_j} u_0 w_\epsilon$ has a compact support contained in the interior of ϵY_i (see (3.16)) then $\rho_i \partial_{x_j} u_0 w_\epsilon \in W^{1,q'}_{per}(\epsilon Y_i)$ and (3.17) follows.

For the second term on the right-hand side of (3.15), we use the definition of v_0 (see (3.3)) and $1/p + 1/q \le 1/2$ to obtain

$$-\int_\Omega a(x/\epsilon)(e_j - \nabla_y \chi^j(x/\epsilon)) \cdot \left(w_\epsilon \nabla \frac{\partial u_0}{\partial x_j}(x)\right) dx = -\int_\Omega \nabla_x \cdot v_0(x, x/\epsilon) w_\epsilon(x) dx.$$

Hence,

$$\int_\Omega v_0(x, x/\epsilon) \cdot \nabla w_\epsilon(x) dx = -\int_\Omega \nabla_x \cdot v_0(x, x/\epsilon) w_\epsilon(x) dx. \tag{3.18}$$

From (3.13), (3.14), and (3.18) we obtain

$$\int_\Omega \eta_\epsilon \cdot \nabla w_\epsilon dx = 0,$$

and from (3.12)

$$\int_\Omega a^\epsilon \nabla w_\epsilon \cdot \nabla z_\epsilon dx = \int_\Omega (a^\epsilon \nabla z_\epsilon - \eta_\epsilon) \cdot \nabla w_\epsilon dx. \tag{3.19}$$

Also from (3.11) and (3.19) we have

$$\begin{aligned}\left|\int_\Omega g(z_\epsilon - \epsilon\theta_\epsilon) dx\right| &\le c \|a^\epsilon \nabla z_\epsilon - \eta_\epsilon\|_0 \|w_\epsilon\|_1 \\ &\le \epsilon \|v_1(\cdot, \cdot/\epsilon) - a^\epsilon \nabla_x u_1(\cdot, \cdot/\epsilon)\|_0 \|g\|_{-1} \quad \text{by (3.9).}\end{aligned}$$

Dividing by $\|g\|_{-1}$ and taking the supremum for $g \neq 0$, the estimate (3.1) follows since

$$\|z_\epsilon(x) - \epsilon\theta_\epsilon\|_1 \leq c\epsilon\|v_1(\cdot,\cdot/\epsilon) - a^\epsilon \nabla_x u_1(\cdot,\cdot/\epsilon)\|_0 \leq c\epsilon \max_j \|\chi^j\|_{1,q,Y}\|u_0\|_{2,p}.$$

We now prove the estimate (3.2). The proof is based on the maximum principle; see also [7] for a proof based on L^p boundary data. Since $u_0 \in W^{2,p}(\Omega)$ and $p > d$, we have $\partial_{x_i} u_0 \in C(\Omega)$, and since $\chi^j \in W^{1,q}_{per}(Y)$ and $q > d$, we obtain $\chi^j \in C(Y)$; see (2.5) and (2.6). Using a Sobolev embedding theorem we obtain

$$\|u_1\|_{0,\infty,\partial\Omega} \leq c \max_j \|\chi^j\|_{0,\infty,Y}\|u_0\|_{1,\infty} \leq c \max_j \|\chi^j\|_{1,q,Y}\|u_0\|_{2,p},$$

and using the maximum principle in [42], we have

$$\|\theta_\epsilon\|_0 \leq c\|\theta_\epsilon\|_{0,\infty,\partial\Omega} = c\|u_1\|_{0,\infty,\partial\Omega} \leq c \max_j \|\chi^j\|_{1,q,Y}\|u_0\|_{2,p}. \tag{3.20}$$

The estimate (3.2) follows from (3.20) and (3.1). ∎

The following remark is used in the proof of Proposition 3.5.

Remark 3.1. *Assume that the solution u_0 of*

$$-\nabla.A\nabla u_0 = f \quad in \quad \Omega, \qquad u_0 = g \quad on \quad \partial\Omega$$

belongs to $W^{2,p}(\Omega)$. Also let $u_\epsilon \in H^1(\Omega)$ be the weak solution of

$$L_\epsilon u_\epsilon = f \quad in \quad \Omega, \quad u_\epsilon = g \quad on \quad \partial\Omega.$$

Then it is easy to see that Proposition 3.1 *extends immediately to the nonhomogeneous Dirichlet boundary condition case.*

Proposition 3.2 estimates the H^1-norm of $\tilde{\theta}_\epsilon - \tilde{\phi}_\epsilon$ in terms of a parameter r; see Remark 3.2 for the discussion on r. Proposition 3.2 is also used in the proof of Theorems 3.1 and 3.2.

PROPOSITION 3.2. *Let $\Omega = (0,1)^2$ be the unit square. Let u_0, $\tilde{\theta}_\epsilon$, and $\tilde{\phi}_\epsilon$ be defined by* (2.4), (2.10), *and* (2.14), *respectively. Assume that $a_{ij} \in L^\infty_{per}(Y)$ and $u_0 \in W^{2,p}(\Omega)$. Let the functions v_k be defined as in subsection* 2.2.1. *Assume that $\exp(-\gamma y_1)\nabla v_e \in L^{\tilde{q}}(G_e)$ and $\exp(-\gamma y_1)(v_e - \chi^*_e) \in L^{\tilde{q}}(G_e)$ and that similar conditions also hold for the other functions v_k, $k \in \{w,n,s\}$. Let l be defined by $\frac{1}{p} + \frac{1}{\tilde{q}} + \frac{1}{l} = 1$, $1 \leq p, \tilde{q} \leq \infty$. Let r be any number such that $r > 2$. Then, there is positive constant $c = c(\gamma, l, r)$ and $\delta = \delta(p,l,r) = -\frac{1}{p} + \frac{1}{lr}$ such that*

$$\begin{aligned}\|\tilde{\theta}_\epsilon - \tilde{\phi}_\epsilon\|_1 \leq c\epsilon^\delta \max_k \big(&\|\exp(-\gamma y\cdot\eta^k)\nabla v_k\|_{0,\tilde{q},G_k} \\ &+ \|\exp(-\gamma y\cdot\eta^k)v_k - \chi^*_k\|_{1,\tilde{q},G_k}\big)\,\|u_0\|_{2,p}.\end{aligned}$$

In addition, when $p, q = \infty$, then $\delta = \frac{1}{2}$. Also, when $\frac{3}{p} + \frac{1}{\tilde{q}} < 1$, then there exists an $r > 2$ such that $\delta > 0$. Lastly, when $\frac{1}{p} + \frac{1}{\tilde{q}} < \frac{1}{2}$, then there exists an $r > 2$ such that $\delta > -\frac{1}{p} + \frac{1}{4}$.

Proof. By definition

$$\|\tilde{\theta}_\epsilon - \tilde{\phi}_\epsilon\|_1 \leq \sum_{k\in\{e,w,n,s\}} \|\tilde{\theta}^k_\epsilon - \tilde{\phi}^k_\epsilon\|_1.$$

Consider the case $k = e$; the other cases are treated similarly. Denote $a^\epsilon(x) = a(x/\epsilon)$, let $v_e^\epsilon(x) = v_e(\frac{x_1-1}{\epsilon}, \frac{x_2}{\epsilon})$, and take $g \in H_0^1(\Omega)$. Then applying the definition of $\tilde{\phi}_\epsilon^e$ and $\tilde{\theta}^e$ (see (2.13) and (2.12), respectively) we obtain

$$\int_\Omega a^\epsilon \nabla(\tilde{\theta}_\epsilon^e - \tilde{\phi}_\epsilon^e)\cdot\nabla g dx = \int_\Omega -a^\epsilon\nabla\left((v_e^\epsilon - \chi_e^*)\varphi_e\frac{\partial u_0}{\partial x_1}\right)\cdot\nabla g dx$$

$$= -\int_\Omega\left(\varphi_e\frac{\partial u_0}{\partial x_1}a^\epsilon\nabla v_e^\epsilon\right)\cdot\nabla g dx - \int_\Omega\left((v_e^\epsilon-\chi_e^*)a^\epsilon\nabla\left(\varphi_e\frac{\partial u_0}{\partial x_1}\right)\right)\cdot\nabla g dx. \tag{3.21}$$

We note that due to the Sobolev embedding Theorem 5.4 in [2], the integrals above are well defined. For estimating the first term on the right-hand side of (3.21) we let

$$\int_\Omega\left(\varphi_e\frac{\partial u_0}{\partial x_1}a^\epsilon\nabla v_e^\epsilon\right)\cdot\nabla g dx$$

$$= \int_\Omega a^\epsilon\nabla v_e^\epsilon\cdot\nabla\left(\varphi_e\frac{\partial u_0}{\partial x_1}g\right)dx - \int_\Omega a^\epsilon\nabla v_e^\epsilon\cdot g\nabla\left(\varphi_e\frac{\partial u_0}{\partial x_1}\right)dx. \tag{3.22}$$

We now estimate the first term of the right-hand side of (3.22). Let $I_i = \{(i-1)\epsilon/6 - \epsilon/6 < x_2 < i\epsilon/6+\epsilon/6,\}$, $i_{max} = 1+\sup_{i\in\mathbb{N}}(i3/\epsilon < 1)$, and consider a partition of unity ρ_i of Ω, subject to $(0,1)\times I_i$. Let I_i^ϵ be the interval centered in I_i with $|I_i^\epsilon| = \epsilon$. Since $\mathrm{supp}(\rho_i g)\subset[0,1]\times I_i^\epsilon$ we have

$$\int_\Omega a^\epsilon\nabla v_e^\epsilon\cdot\nabla\left(\varphi_e\frac{\partial u_0}{\partial x_1}g\right)dx$$

$$= \sum_{i=0:i_{max}}\int_0^1\int_{I_i^\epsilon} a^\epsilon\nabla v_e^\epsilon\cdot\nabla\left(\rho_i\varphi_e\frac{\partial u_0}{\partial x_1}g\right)dx_2dx_1 = 0, \tag{3.23}$$

where to arrive at (3.23) we have used the definition of v_e and arguments similar to the ones given to obtain (3.17).

For estimating the second term on the right-hand side of (3.22) we apply Hölder's inequality to obtain

$$\left|\int_\Omega a^\epsilon\nabla v_e^\epsilon\cdot\nabla\left(\varphi_e\frac{\partial u_0}{\partial x_1}\right)g dx\right|$$

$$\le \|a\|_\infty|\varphi_e\nabla u_0|_{1,p}\left\|\exp\left(-\gamma\frac{x_1-1}{\epsilon}\right)\nabla v_e^\epsilon\right\|_{0,\tilde{q}}\left(\frac{\epsilon}{\gamma}\right)^{1/l}\left\|(\gamma/\epsilon)^{1/l}\exp\left(\gamma\frac{x_1-1}{\epsilon}\right)g\right\|_{0,l}, \tag{3.24}$$

where $1/l = 1-1/p-1/\tilde{q}$. Taking $y_1 = (x_1-1)/\epsilon$ and $y_2 = x_2/\epsilon$, and exploring the $[0,1]$-periodicity of $v_e(y_1,\cdot)$, we have

$$\left\|\nabla(v_e^\epsilon)\exp\left(-\gamma\frac{x_1-1}{\epsilon}\right)\right\|_{0,\tilde{q}} \le \left(\left(\frac{1}{\epsilon}+1\right)\int_{-1/\epsilon}^0\int_0^1|\exp(-\gamma y_1)\nabla_y v_e|^s\epsilon^{2-\tilde{q}}dy_2dy_1\right)^{\frac{1}{\tilde{q}}}$$

$$\le c\epsilon^{(\frac{1}{\tilde{q}}-1)}\|\exp(-\gamma y_1)\nabla_y v_e\|_{0,\tilde{q},G_e}. \tag{3.25}$$

Now let $g_n \in C_0^\infty(\Omega)$ such that $g_n \to g$ in H^1, and let $I_n = (0,1)\cap|g_n| > 0$.

Using integration by parts in x_1 we obtain

$$\left\|(\gamma/\epsilon)^{1/l}\exp\left(\gamma\frac{x_1-1}{\epsilon}\right)g_n\right\|_{0,l} = \left(\int_0^1\int_{I_n}\frac{\gamma}{\epsilon}\exp\left(l\gamma\frac{x_1-1}{\epsilon}\right)|g_n|^l dx_1dx_2\right)^{1/l}$$

$$= \left(-\int_0^1\int_{I_n}\frac{1}{l}\exp\left(l\gamma\frac{x_1-1}{\epsilon}\right)\frac{\partial|g_n|^l}{\partial x_1}dx_1dx_2\right)^{1/l} \tag{3.26}$$

$$\leq c\left(\left\|\exp\left(l\gamma\frac{x_1-1}{\epsilon}\right)\right\|_{0,r}\|g_n\|^{l-1}_{0,s(l-1)}\left\|\frac{\partial g_n}{\partial x_1}\right\|_0\right)^{1/l} \tag{3.27}$$

$$\leq c(s(l-1))^{(l-1)/l}\left(\frac{\epsilon}{rl\gamma}\right)^{1/(rl)}|g_n|_1, \tag{3.28}$$

where we have used Hölder's inequality with $1/s+1/r=1/2, s<\infty$ to obtain (3.27). To obtain (3.28), we use the following inequality used in the proof of Lemma 5.10 in [2], i.e., there exists $1\leq t<2$ such that

$$\begin{aligned}\|g_n\|_{0,r(l-1)} &\leq 2^{(t-1)/t}\left(\frac{2t-t}{2-t}\right)\|g_n\|_{1,t} \quad \text{for} \quad 2t/(2-t)=s(l-1)\\ &\leq 2^{(t-1)/t}\left(\frac{2t-t}{2-t}\right)\mathrm{vol}(\Omega)^{(1/t-1/2)}\|g_n\|_1 \text{ by Theorem 2.8 in [2]}\\ &\leq c\left(\frac{2t-t}{2-t}\right)|g_n|_1 \text{ by a Poincaré inequality.}\end{aligned}$$

Hence, (3.28) follows from (3.27). Now taking the limit $n\to\infty$ in (3.28), we obtain the inequality (3.28) for g. Thus, using (3.22), (3.23), (3.24), (3.25), and (3.28), it follows that

$$\begin{aligned}\int_\Omega \varphi_e\frac{\partial u_0}{\partial x_1}a^\epsilon\nabla v_e^\epsilon\cdot\nabla g dx &\leq c(\gamma)(s(l-1))^{(l-1)/l}\epsilon^\delta\|a\|_\infty|\varphi_e\nabla u_0|_{1,p}\\ &\quad \|\exp(-\gamma y_1)\nabla v_e\|_{0,\tilde q,G_e}|g|_1,\end{aligned} \tag{3.29}$$

where $\delta=-\frac{1}{p}+\frac{1}{lr}$. When $p,\tilde q\to\infty$ and then $r\to 2$, then $(s(l-1))^{(l-1)/l}\,(\epsilon/(rl\gamma))^{1/(rl)}$ $\to \epsilon^{1/2}/(2\gamma)$. Also it is easy to see that $\frac{3}{p}+\frac{1}{\tilde q}<1$ implies that there exists $r<2$ such that $\delta=-\frac{1}{p}+\frac{1}{rl}>0$, and $\frac{1}{p}+\frac{1}{\tilde q}<\frac{1}{2}$ implies that there exists $r<2$ such that $\delta=-\frac{1}{p}+\frac{1}{rl}>-\frac{1}{4}$.

For estimating the second term on the right-hand side of (3.21), we apply Hölder's inequality to obtain

$$\begin{aligned}&\left|\int_\Omega(v_e^\epsilon-\chi_e^*)a^\epsilon\nabla\left(\varphi_e\frac{\partial u_0}{\partial x_1}\right)\cdot\nabla g dx\right|\\ &\leq\|a\|_{0,\infty}\left|\varphi_e\frac{\partial u_0}{\partial x_1}\right|_{1,p}\left\|(v_e^\epsilon-\chi_e^*)\exp\left(-\gamma\frac{x_1-1}{\epsilon}\right)\right\|_{0,\tilde q}\left\|\exp\left(\gamma\frac{x_1-1}{\epsilon}\right)\right\|_{0,l}\\ &\leq c(\gamma)\epsilon^{\frac{1}{2}-\frac{1}{p}}\|a\|_{0,\infty}\left|\varphi_e\frac{\partial u_0}{\partial x_1}\right|_{1,p}\|(v_e-\chi_e^*)\exp(-\gamma y_1)\|_{0,\tilde q,G_e}|g|_1,\end{aligned} \tag{3.30}$$

and we note that $\delta<-\frac{1}{p}+\frac{1}{2}$ since $0\leq\frac{1}{l}\leq 1$ and $2<r$.

Taking $g = \tilde{\theta}_\epsilon^e - \tilde{\phi}_\epsilon^e$ and using the ellipticity of a, we obtain

$$\begin{aligned}|\tilde{\theta}_\epsilon^e - \tilde{\phi}_\epsilon^e|^2_{H_0^1(\Omega)} &\le \gamma_a^{-1} \int_\Omega (a^\epsilon \nabla(\tilde{\theta}_\epsilon^e - \tilde{\phi}_\epsilon^e)) \cdot \nabla(\tilde{\theta}_\epsilon^e - \tilde{\phi}_\epsilon^e) dx \\ &\le c\epsilon^\delta \|a\|_{0,\infty} |\varphi_e \nabla u_0|_{1,p} \left(\|\nabla(v_e - \chi_e^*) \exp(-\gamma y_1)\|_{0,\tilde{q},G_e} \right. \\ &\quad \left. + \|(v_e - \chi_e^*)\exp(-\gamma y_1)\|_{0,\tilde{q},G_e} \right). \qquad \square\end{aligned}$$

Remark 3.2. When $r \to 2$, then $\delta(p,l,r) = \frac{1}{p} - \frac{1}{rl}$ increases to $\frac{1}{p} - \frac{1}{2l}$. We note, however, that $r \to 2$ implies that $s \to \infty$ and, therefore, the a priori constant $c(\gamma, l, r) \to \infty$ if $l \neq 1$; see the term $(s(l-1))^{(l-1)/l}(\frac{\epsilon}{r\gamma l})^{\frac{1}{rl}}$ in (3.28). The optimal choice of r (denoted by r^*) to minimize this term depends on ϵ, l, and weakly on γ. When $l \to 1$, then $r^*(l) \to 2$. Therefore, $\delta \to -\frac{1}{p} + \frac{1}{2}$. Also, it is easy to see that when $\epsilon \to 0$, then $r^* \to 2$. Hence, for the worst case, i.e., $p = 2$ and $\tilde{q} = \infty$, when ϵ is very small, then δ is very close to $-\frac{1}{4}$.

We next prove the last proposition used in the proof of Theorem 3.1. Proposition 3.3 estimates the H^1-norm of $\bar{\phi} - \bar{\theta}_\epsilon$.

PROPOSITION 3.3. *Let $\Omega = (0,1)^2$ be the unit square. Let the functions u_0, $\bar{\theta}_\epsilon$, and $\bar{\phi}$ be defined by* (2.4), (2.11), *and* (2.15), *respectively. Assume that $a_{ij} \in L^\infty_{per}(Y)$ and $u_0 \in H^2(\Omega)$. Then there exists a positive constant c, independent of ϵ, u_0, and χ^j, such that*

$$\|\bar{\phi} - \bar{\theta}_\epsilon\|_1 \le c \max_j \|\chi^j\|_{1,2,Y} \|u_0\|_2.$$

Proof. Using that $(\bar{\phi} - \bar{\theta}_\epsilon) = 0$ on $\partial\Omega$ we have

$$\begin{aligned}\int_\Omega a_{ij}^\epsilon \frac{\partial(\bar{\phi} - \bar{\theta}_\epsilon)}{\partial x_i} \frac{\partial(\bar{\phi} - \bar{\theta}_\epsilon)}{\partial x_j} dx &= \int_\Omega a_{ij}^\epsilon \frac{\partial \bar{\phi}}{\partial x_i} \frac{\partial(\bar{\phi} - \bar{\theta}_\epsilon)}{\partial x_j} dx \\ &\le \|a\|_{0,\infty,Y} \left(\int_\Omega |\nabla \bar{\phi}|^2 dx \right)^{1/2} \left(\int_\Omega |\nabla(\bar{\phi} - \bar{\theta}_\epsilon)|^2 dx \right)^{1/2},\end{aligned}$$

and from the ellipticity of a and using Remark 2.1 we obtain

$$|\bar{\phi} - \bar{\theta}_\epsilon|_1 \le \frac{\|a\|_{0,\infty,Y}}{\gamma_a} |\bar{\phi}|_1 \le \tilde{c} \frac{\|a\|_{0,\infty,Y}}{\gamma_a} \max_j \|\chi^j\|_{1,2,Y} \|u_0\|_2.$$

The proposition follows from a Poincaré inequality. $\square$

The following proposition generalizes Proposition 2.3 in [35], where it is assumed that $a_{ij} \in C^{1,\beta}_{per}(Y)$, $u_0 \in H^3(\Omega)$, and $\Omega \subset \mathbb{R}^2$ is a smooth domain.

PROPOSITION 3.4. *Let $\Omega \subset \mathbb{R}^d$, $d = 2, 3$, be a bounded domain with a Lipschitz boundary. Let u_ϵ be the solution of the problem* (1.1), *and let χ^j, u_0, u_1, θ_ϵ, χ^{ij}, u_2, and φ_ϵ be defined by* (2.2), (2.4), (2.5), (2.6), (2.7), (2.8), *and* (2.9), *respectively. Assume that $a_{ij} \in L^\infty_{per}(Y)$, $u_0 \in W^{3,p}(\Omega)$, and χ^j and $\chi^{ij} \in W^{1,q}_{per}(Y)$ for $1/p + 1/q \le 1/2$. Then there exists a constant c independent of u_0, ϵ, χ^j, and χ^{ij} such that*

$$\begin{aligned}\|u_\epsilon(\cdot) - u_0(\cdot) - \epsilon u_1(\cdot, \cdot/\epsilon) - \epsilon\theta_\epsilon(\cdot) - \epsilon^2 u_2(\cdot, \cdot/\epsilon) - \epsilon^2 \varphi_\epsilon(\cdot)\|_1 \\ \le C\epsilon^2 \left(\max_j \|\chi^j\|_{0,q,Y} + \max_{kj} \|\chi^{kj}\|_{1,q,Y} \right) \|u_0\|_{3,p}.\end{aligned}$$

In addition, if $p, q > d$, then

$$\|u_\epsilon(\cdot) - u_0(\cdot) - \epsilon u_1(\cdot, \cdot/\epsilon) - \epsilon\theta_\epsilon(\cdot)\|_0 \le c\epsilon^2 \left(\max_j \|\chi^j\|_{0,q} + \max_{kj} \|\chi^{kj}\|_{1,q} \right) \|u_0\|_{3,p}.$$

Proof. Define the field v_1 by

$$(v_1(x,y))_k = -a_{ki}(y)\chi^j \frac{\partial^2 u_0}{\partial x_j \partial x_i}(x) + a_{kl}(y)\frac{\partial \chi^{ij}}{\partial y_l}\frac{\partial^2 u_0}{\partial x_j \partial x_i}(x); \tag{3.31}$$

then

$$a(y)\nabla_x u_1(x,y) + a(y)\nabla_y u_2(x,y) = v_1(x,y). \tag{3.32}$$

Let $q(y) = \phi(y)$, where ϕ is defined by (3.5), and let $\psi_{ij} \in W^{1,q}_{\text{per}}(Y)$ such that

$$curl_y \psi_{1j} = \tilde{\psi}_{1j} = \begin{pmatrix} -a_{11}\chi^j + a_{1l}\partial_l \chi^{1,j} - c^1_{1j} \\ -a_{21}\chi^j + a_{2l}\partial_l \chi^{1,j} - \phi^{(3)}_j - c^2_{1j} \\ -a_{31}\chi^j + a_{3l}\partial_l \chi^{1,j} + \phi^{(2)}_j - c^3_{1j} \end{pmatrix},$$

$$curl_y \psi_{2j} = \tilde{\psi}_{2j} = \begin{pmatrix} -a_{12}\chi^j + a_{1l}\partial_l \chi^{2,j} + \phi^{(3)}_j - c^1_{2j} \\ -a_{22}\chi^j + a_{2l}\partial_l \chi^{2,j} - c^2_{2j} \\ -a_{32}\chi^j + a_{3l}\partial_l \chi^{2,j} - \phi^{(1)}_j - c^3_{2j} \end{pmatrix},$$

and

$$curl_y \psi_{3j} = \tilde{\psi}_{3j} = \begin{pmatrix} -a_{13}\chi^j + a_{1l}\partial_l \chi^{3,j} - \phi^{(2)}_j - c^1_{3j} \\ -a_{23}\chi^j + a_{2l}\partial_l \chi^{3,j} + \phi^{(1)}_j - c^2_{3j} \\ -a_{33}\chi^j + a_{3l}\partial_l \chi^{3,j} - c^3_{3j} \end{pmatrix}.$$

Here the constants c^l_{ij} are chosen so that each entry of the vectors $\tilde{\psi}_{ij}$ has zero value integral over Y; for instance, $c^1_{1j} = \int_Y -a_{11}\chi^j + a_{1l}\partial_l \chi^{1,j} dy$. It is easy to check that $\nabla_y \cdot \tilde{\psi}_{kj} = 0$; hence, the existence of such functions ψ_{kj} is guaranteed by Lemma 3.1 and by Remark 3.11 in [24]. We then obtain

$$\|\psi_{kj}\|_{1,q} \leq c(\|\chi^j\|_{0,q} + \|\chi^{kj}\|_{1,q}). \tag{3.33}$$

Define

$$p(x,y) = \psi_{kj}(y)\frac{\partial^2 u_0}{\partial x_k \partial x_j}(x), \tag{3.34}$$

and let

$$v_2(x,y) = -curl_x p(x,y).$$

An elementary calculation gives

$$\nabla_y \cdot v_2 = -\nabla_x \cdot v_1, \quad \nabla_x \cdot v_2 = 0 \tag{3.35}$$

and

$$\begin{aligned} \|v_2(\cdot,\cdot/\epsilon)\|_0 &\leq c\|u_0\|_{3,p} \max_{kj} \|\psi_{kj}\|_{1,q,Y} \\ &\leq c\|u_0\|_{3,p}(\|\chi^j\|_{0,q} + \|\chi^{kj}\|_{1,q}) \quad \text{by (3.33).} \end{aligned} \tag{3.36}$$

Define

$$\psi_\epsilon(x) = u_\epsilon(x) - u_0(x) - \epsilon u_1(x, x/\epsilon) - \epsilon^2 u_2(x, x/\epsilon)$$

and

$$\xi_\epsilon(x) = a(x/\epsilon)\nabla u_\epsilon(x) - v_0(x, x/\epsilon) - \epsilon v_1(x, x/\epsilon) - \epsilon^2 v_2(x, x/\epsilon),$$

where v_0 is defined by (3.3). Then

$$\begin{aligned} a(x/\epsilon)\nabla\psi_\epsilon - \xi_\epsilon(x) &= a(x/\epsilon)\nabla u_\epsilon(x) - a(x/\epsilon)\nabla u_0(x) - \epsilon a(x/\epsilon)\nabla u_1(x, x/\epsilon) \\ &\quad -\epsilon^2 a(x/\epsilon)\nabla u_2(x, x/\epsilon) \\ &\quad -a(x/\epsilon)\nabla u_\epsilon(x) + v_0(x, x/\epsilon) + \epsilon v_1(x, x/\epsilon) + \epsilon^2 v_2(x, x/\epsilon) \\ &= -a(x/\epsilon)\nabla_x u_0(x) - \epsilon a(x/\epsilon)\nabla_x u_1(x, x/\epsilon) - a(x/\epsilon)\nabla_y u_1(x, x/\epsilon) \\ &\quad -\epsilon^2 a(x/\epsilon)\nabla_x u_2(x, x/\epsilon) - \epsilon a(x/\epsilon)\nabla_y u_2(x, x/\epsilon) \\ &\quad +v_0(x, x/\epsilon) + \epsilon v_1(x, x/\epsilon) + \epsilon^2 v_2(x, x/\epsilon) \\ &= \epsilon^2(v_2(x, x/\epsilon) - a(x/\epsilon)\nabla_x u_2(x, x/\epsilon)) \quad \text{by (3.3) and (3.32).} \end{aligned}$$

Using (3.36) and the definition of u_2 (see (2.8)) we obtain

$$\|a(x/\epsilon)\nabla\psi_\epsilon - \xi_\epsilon\|_0 \le c\epsilon^2 \|u_0\|_{3,p} \max_{kj}(\|\chi^j\|_{0,q} + \|\chi^{kj}\|_{1,q}). \tag{3.37}$$

Given $g \in L^2(\Omega)$, let $w_\epsilon \in H^1(\Omega)$ denote the solution of

$$\int_\Omega a(x/\epsilon)\nabla w_\epsilon(x)\nabla\psi(x)dx = \int_\Omega g(x)\psi(x)dx \quad \forall\psi \in H_0^1(\Omega). \tag{3.38}$$

Since $\psi_\epsilon - \epsilon\theta_\epsilon - \epsilon^2\varphi_\epsilon \in H_0^1(\Omega)$ we obtain

$$\begin{aligned} \int_\Omega g(\psi_\epsilon - \epsilon\theta_\epsilon - \epsilon^2\varphi_\epsilon)dx &= \int_\Omega a(x/\epsilon)(\nabla\psi_\epsilon - \epsilon\nabla\theta_\epsilon - \epsilon^2\nabla\varphi_\epsilon)\nabla w_\epsilon(x)dx \\ &= \int_\Omega a(x/\epsilon)\nabla\psi_\epsilon\nabla w_\epsilon(x)dx, \end{aligned} \tag{3.39}$$

where we have used the definition of θ_ϵ and φ_ϵ to obtain (3.39). We observe that

$$\int_\Omega a^\epsilon\nabla\psi_\epsilon\nabla w_\epsilon dx = \int_\Omega (a^\epsilon\nabla\psi_\epsilon - \xi_\epsilon)\cdot\nabla w_\epsilon dx + \int_\Omega \xi_\epsilon\cdot\nabla w_\epsilon dx, \tag{3.40}$$

and we estimate the second term on the right-hand side of (3.40) as follows:

$$\begin{aligned} \int_\Omega \xi_\epsilon\cdot\nabla w_\epsilon dx &= \int_\Omega (a(x/\epsilon)\nabla u_\epsilon(x) - v_0(x, x/\epsilon) - \epsilon v_1(x, x/\epsilon) \\ &\quad -\epsilon^2 v_2(x, x/\epsilon))\cdot\nabla w_\epsilon(x)dx \\ &= \int_\Omega f(x)w_\epsilon(x) + \nabla_x\cdot v_0(x, x/\epsilon)w_\epsilon(x) \\ &\quad -\epsilon v_1(x, x/\epsilon)\cdot\nabla w_\epsilon(x) + \epsilon\nabla_x v_1(x, x/\epsilon)w_\epsilon(x)dx, \end{aligned} \tag{3.41}$$

where we have used the definition of u_ϵ, (3.18), integration by parts, and (3.35) to obtain (3.41). Using (3.31) we have

$$\begin{aligned} \int_\Omega v_1(x, x/\epsilon)\cdot\nabla w_\epsilon(x) = \int_\Omega \Bigg(&-a^\epsilon_{ki}\chi^j_\epsilon \frac{\partial^2 u_0}{\partial x_j \partial x_i}(x) \\ &+ a^\epsilon_{kl}\frac{\partial\chi^{ij}_\epsilon}{\partial y_l}\frac{\partial^2 u_0}{\partial x_j\partial x_i}(x)\Bigg)\frac{\partial w_\epsilon}{\partial x_k}(x)dx. \end{aligned} \tag{3.42}$$

Considering the partition of unit ρ_i defined in the proof of Proposition 3.1, we obtain

$$\begin{aligned}
&\int_\Omega a_{kl}^\epsilon \frac{\partial \chi_\epsilon^{ij}}{\partial y_l}\frac{\partial^2 u_0}{\partial x_j \partial x_i}\frac{\partial w_\epsilon}{\partial x_k}(x)dx = \sum_1^{i_m}\int_{\epsilon Y_i} a_{kl}^\epsilon \frac{\partial \chi_\epsilon^{ij}}{\partial y_l}\rho_i \frac{\partial^2 u_0}{\partial x_j \partial x_i}\frac{\partial w_\epsilon}{\partial x_k}dx \\
&= \sum_1^{i_m}\int_{\epsilon Y_i} a_{kl}^\epsilon \frac{\partial \chi_\epsilon^{ij}}{\partial y_l}\frac{\partial}{\partial x_k}\left(\rho_i \frac{\partial^2 u_0}{\partial x_j \partial x_i}w_\epsilon(x)\right) - a_{kl}^\epsilon \frac{\partial \chi_\epsilon^{ij}}{\partial y_l}w_\epsilon(x)\frac{\partial}{\partial x_k}\left(\rho_i \frac{\partial^2 u_0}{\partial x_j \partial x_i}\right)dx \\
&= \sum_1^{i_m}\int_{\epsilon Y_i} \epsilon^{-1}\left(a_{ij}^\epsilon - a_{ik}^\epsilon \frac{\partial \chi_\epsilon^j}{\partial_{y_k}} + A_{ij}\right)\rho_i \frac{\partial^2 u_0}{\partial x_j \partial x_i}w_\epsilon \\
&\quad + a_{ki}^\epsilon \chi_\epsilon^j \left(\frac{\partial}{\partial_{x_k}}\left(\rho_i \frac{\partial^2 u_0}{\partial x_j \partial x_i}(x)\right)w_\epsilon(x) + \rho_i \frac{\partial^2 u_0}{\partial x_j \partial x_i}(x)\frac{\partial w_\epsilon}{\partial_{x_k}}(x)\right)dx \\
&\quad - \int_\Omega a_{kl}\frac{\partial \chi_\epsilon^{ij}}{\partial y_l}w_\epsilon(x)\frac{\partial}{\partial x_k}\left(\frac{\partial^2 u_0}{\partial x_j \partial x_i}\right)dx \qquad (3.43) \\
&= \int_\Omega \epsilon^{-1}\left(\nabla_x v_0 \frac{\partial^2 u_0}{\partial x_j \partial x_i}(x) - f\right)w_\epsilon(x)dx \\
&\quad - \int_\Omega a_{ki}^\epsilon \chi_\epsilon^j \frac{\partial^2 u_0}{\partial x_j \partial x_i}\frac{\partial w_\epsilon}{\partial_{x_k}}(x)dx - \int_\Omega \nabla_x \cdot v_1 dx. \qquad (3.44)
\end{aligned}$$

Here, we have used the definition of χ^{ij} to arrive at (3.43). From (3.41), (3.42), and (3.44) we obtain

$$\int_\Omega \xi_\epsilon \cdot \nabla w_\epsilon(x)dx = 0,$$

and, therefore, from (3.37) and (3.40) we obtain

$$\begin{aligned}
\left|\int_\Omega g(\psi_\epsilon - \epsilon\theta_\epsilon - \epsilon^2\varphi_\epsilon)dx\right| &\le \|a^\epsilon \nabla\psi_\epsilon - \xi_\epsilon\|_0 \|w_\epsilon\|_1 \\
&\le c\epsilon^2 \|u_0\|_{3,p}(\|\chi^j\|_{0,q,Y} + \|\chi^{kj}\|_{1,q,Y})\|g\|_{-1}.
\end{aligned}$$

Dividing by g and taking the supremum over g, (3.31) follows since

$$\|u_\epsilon - u_0 - \epsilon u_1 - \epsilon\theta_\epsilon - \epsilon^2 u_2 - \epsilon^2\varphi_\epsilon\|_1 \le c\epsilon^2\|u_0\|_{3,p}\max_{kj}(\|\chi^j\|_{0,q} + \|\chi^{kj}\|_{1,q}).$$

The remainder of the proof uses the same arguments given to prove (3.2). ∎

The following proposition estimates the L^2-norm of $\bar\phi - \bar\theta_\epsilon$, and it is used in the proof of Theorem 3.2.

PROPOSITION 3.5. *Let $\Omega = (0,1)^2$ be the unit square. Let u_0, χ^j, $\bar\theta_\epsilon$, and $\bar\phi$ be defined by* (2.4), (2.2), (2.11), *and* (2.15), *respectively. Assume that $a_{ij} \in L^\infty_{per}(Y)$, $u_0 \in W^{3,p}(\Omega)$, and $\chi^j \in W^{1,q}_{per}(Y)$ for $1/p + 1/q \le 1/2$, $2 \le p, q \le \infty$. Then we have*

$$\|\bar\theta_\epsilon - \bar\phi\|_0 \le c\epsilon \max_j \|\chi^j\|_{1,2,Y}\|u_0\|_{3,p}.$$

Proof. Using the arguments given in Remark 2.1 we have

$$\left\|\sum_k \varphi_k \chi_k^* \nabla u_0 \cdot \eta_k\right\|_{2,p} \le c\max_k |\chi_k^*|\|u_0\|_{3,p} \le c\max_j \|\chi^j\|_{1,2,Y}\|u_0\|_{3,p}, \qquad (3.45)$$

where we have used that $|\chi_k^*| \leq c \max_j \|\chi^j\|_{1,2,Y}$; see the proof of Lemma 4.4 in [3]. By assumption, $u_0 \in H_0^1(\Omega) \cap W^{3,p}(\Omega)$. Using a trace theorem and Remarks 2.1 and 5.1, we obtain that

$$\bar{\phi}_{|\partial\Omega} = \sum_k \varphi_k \chi_k^* \nabla u_0 \cdot \eta_{k|\partial\Omega}$$

belongs to $W^{2-\frac{1}{p},p}(\Gamma_k) \cap W_{00}^{1-\frac{1}{p},p}(\Gamma_k)$, $k = \{e, w, n, s\}$. Following Chapter 5 in [25], when Ω is a is a polygonal convex domain, then the problem is $W^{2,p}(\Omega)$ regular. Therefore, the solution of (2.16) belongs to $W^{2,p}(\Omega)$ and satisfies $\|\bar{\phi}\|_{2,p} \leq c \max_j \|\chi^j\|_{1,2,Y} \|u_0\|_{3,p}$. From Remark 3.1 and (3.2) we obtain

$$\|\bar{\theta}_\epsilon - \bar{\phi}\|_0 \leq c\epsilon \max_j \|\chi^j\|_{1,2,Y} \|\bar{\phi}\|_{2,p},$$

and the proposition follows. ∎

The following proposition is used in the proof of Propositions 3.1 and 3.4.

LEMMA 3.1. *A function* $\mathbf{v} \in L^2_{per}(Y)^2$, $(\mathbf{v} \in L^2_{per}(Y)^3)$ *satisfies*

$$\nabla \cdot \mathbf{v} = \mathbf{0}, \tag{3.46}$$

and $\int_Y v_i dy = 0$ *iff there exists a function* $\phi \in H^1_{per}(Y)$ $(\phi \in H^1_{per}(Y)^3)$ *such that*

$$\mathbf{v} = curl\phi. \tag{3.47}$$

Proof. The proof is similar to the proof of Theorem 3.4 in [24], where here we use discrete Fourier transforms rather than continuous Fourier transforms; see [44]. ∎

The following theorems provide error estimates between u_ϵ and $u_0 - \epsilon u_1 - \epsilon\phi_\epsilon$ on the H^1- and L^2-norms.

THEOREM 3.1. *Let* $\Omega = (0,1)^2$ *be the unit square, and let* u_ϵ *be the solution of the problem* (1.1). *Let* u_0, u_1, ϕ_ϵ *be defined by* (2.4), (2.5), *and* (2.16), *respectively. Assume* $a_{ij} \in L^\infty_{per}(Y)$, $u_0 \in W^{2,p}(\Omega)$, *and* $\chi^j \in W^{1,q}_{per}(Y)$ *for* $1/p + 1/q \leq 1/2$. *Let the functions* v_k *be defined as in subsection* 2.2.1. *Assume that* $\exp(-\gamma y_1)\nabla v_e \in L^{\tilde{q}}(G_e)$ *and* $\exp(-\gamma y_1)(v_e - \chi_e^*) \in L^{\tilde{q}}(G_e)$ *for* $\frac{1}{p} + \frac{1}{\tilde{q}} \leq \frac{1}{2}$ *and that similar conditions also hold for the other functions* v_k, $k \in \{w, n, s\}$. *Define* $\hat{\delta} = \min\{0, \delta\}$, δ *defined by Proposition* 3.2 *and Remark* 3.2. *Then there exists a constant* c *independent of* ϵ *such that*

$$\|u_\epsilon(\cdot) - u_0(\cdot) - \epsilon u_1(\cdot, \cdot/\epsilon) - \epsilon\phi_\epsilon(\cdot)\|_1 \leq c\epsilon^{1+\hat{\delta}} \|u_0\|_{2,p}.$$

In addition, if $\frac{1}{p} + \frac{1}{\tilde{q}} < \frac{1}{2}$, *then* $\hat{\delta} > -\frac{1}{p} + \frac{1}{4}$. *Also, if* $\frac{3}{p} + \frac{1}{\tilde{q}} < 1$, *then* $\hat{\delta} = 0$.

Proof. By the triangular inequality we have

$$\begin{aligned} |u_\epsilon - u_0 - \epsilon u_1 - \epsilon\phi_\epsilon|_1 \leq\ & |u_\epsilon - u_0 - \epsilon u_1 - \epsilon\theta_\epsilon|_1 \\ & + \epsilon|\tilde{\theta}_\epsilon - \tilde{\phi}_\epsilon|_1 + \epsilon|\bar{\theta}_\epsilon - \bar{\phi}|_1, \end{aligned}$$

and applying Propositions 3.1, 3.2, and 3.3 and Remark 3.2, the theorem then follows. ∎

Remark 3.3. Theorem 1 in [34] guarantees that $\chi^j \in W^{1,q}_{per}(Y)$ for all $q < \infty$. See also Theorem 1.1 in [31] for conditions on a_{ij} in order to have $\chi^j \in W^{1,\infty}_{per}(Y)$.

The following theorem gives higher error estimates when more regularity is assumed.

THEOREM 3.2. *Let $\Omega = (0,1)^2$ be the unit square, and let u_ϵ be the solution of problem (1.1). Let u_0, u_1, ϕ_ϵ be defined by (2.4), (2.5), and (2.16), respectively. Assume $u_0 \in W^{3,p}(\Omega)$, $\bar{\phi} \in W^{2,p}(\Omega)$, and χ^j and $\chi^{ij} \in W^{1,q}_{per}(Y)$ for $1/p + 1/q \leq 1/2$, $p, q > 2$. Let the functions v_k be defined as in subsection 2.2.1. Assume that $\exp(-\gamma y_1)\nabla v_e \in L^\infty(G_e)$ and $\exp(-\gamma y_1)(v_e - \chi^*_e) \in L^\infty(G_e)$ and that similar conditions also hold for the other functions v_k, $k \in \{w, n, s\}$. Then there exists a constant c independent of ϵ such that*

$$\|u_\epsilon(\cdot) - u_0(\cdot) - \epsilon u_1(\cdot, \cdot/\epsilon) - \epsilon\phi_\epsilon(\cdot)\|_0 \leq c\epsilon^{3/2}\|u_0\|_{3,p}.$$

Proof. Use a triangular inequality similar to the one used in the proof of Theorem 3.1, use a Sobolev embedding theorem to have $u_0 \in W^{2,\infty}(\Omega)$, and then apply the Propositions 3.4, 3.2, and 3.5. We obtain the factor $O(\epsilon^{3/2})$ rather than the (ϵ^2) as in Proposition 3.4 since $\|\tilde{\theta}_\epsilon - \tilde{\phi}_\epsilon\|_0$ is $O(\epsilon^{1/2})$; see Remark 3.2 and Proposition 3.2. ∎

Remark 3.4. Note that if $a_{ij} \in C^{1,\beta}_{per}(Y)$, $\beta > 0$, then from regularity theory we have $\chi^j \in C^{1,\beta}_{per}$, $v_e \in C^{1,\beta}$, and $\nabla(v_e - \chi^*_e)\exp(-\gamma y_1) \in L^\infty(G_e)$; see Theorem 15.1 in [29] and Remark 6.4 in [35]; see also Theorem 3 in [30] for related issues.

4. Finite element approximation. For sections 4 and 5 we assume that A, χ^j, v_k, and χ^* are computed exactly. We now describe the finite element approximations for u_0, u_1, $\tilde{\phi}_\epsilon$, and $\bar{\phi}$.

- Let $V^h(\Omega)$ be the P_1 or the Q_1 conforming finite element space on the mesh $\mathcal{T}_h(\Omega)$, and let $V^h_0(\Omega) = V^h(\Omega) \cap H^1_0(\Omega)$. Define $u^h_0 \in V^h_0(\Omega)$ as the solution of

$$\int_\Omega A\nabla u^h_0 \cdot \nabla v^h dx = \int_\Omega f v^h dx \quad \forall v^h \in V^h_0(\Omega). \tag{4.1}$$

- We now introduce the discrete approximation for $\partial_\eta u_0$. Note that $\partial_\eta u_0$ is used on the boundary condition imposed for problem (2.15). In order to approximate $\partial_\eta u_0$, we define $Y^h = V^h(\Omega)|_{\partial\Omega}$, $Y^h_k = Y^h|_{\Gamma_k}$, and $Y^h_{0,k} = \{\lambda^h \in Y^h_k; \lambda^h = 0 \text{ at } \partial\Gamma_k\}$. For all $\phi^h \in V^h(\Omega)$ and $\phi^h|_{\partial\Omega\setminus\Gamma_k} = 0$, let $\lambda^h_k \in Y^h_{0,k}$ be defined as the solution of

$$\int_{\Gamma_k} \lambda^h_k \phi^h d\sigma = \int_\Omega A\nabla u^h_0 \cdot \nabla\phi^h dx - \int_\Omega f\phi^h dx; \tag{4.2}$$

  see [48]. Later in Lemma 5.3 we show that λ^h_k is a good approximation for $A\nabla u_0 \cdot \eta_k$ on Γ_k; hence, we approximate $\partial_\eta u_0$ by μ^h, where

$$\mu^h|_{\Gamma_k} = \lambda^h_k / A_{l_k l_k},\ l_k = \begin{cases} 1 & \text{if} \quad k = e, w, \\ 2 & \text{if} \quad k = n, s. \end{cases} \tag{4.3}$$

- We note that we use μ^h as the approximation for $\partial_\eta u_0$ in (4.7). Therefore, in order to guarantee that the final numerical approximation for u_ϵ satisfies the zero Dirichlet boundary condition, we define the approximation for ∇u_0 as

$$\Psi^h = \nabla u^h_0 + \sum_{k\in\{e,w,n,s\}} E^h_k(\mu^h - \nabla u^h_0 \cdot \eta^k)\eta^k. \tag{4.4}$$

  Here $E^h_k(\cdot)$ denotes the trivial zero nonconforming discrete extension of $\mu^h - \nabla u^h_0 \cdot \eta^k$ in Ω. More specifically, $E^h_k(\mu^h - \nabla u^h_0 \cdot \eta^k)(z) = 0$ if z is a nodal point

of $\mathcal{T}_h(\overline{\Omega})\setminus\Gamma_k$, $E_k^h(\mu^h-\nabla u_0^h\cdot\eta^k)=\mu^h-\nabla u_0^h\cdot\eta^k$ on Γ_k, and $E_k^h(\mu^h-\nabla u_0^h\cdot\eta^k)$ inside each element $K_i\in\mathcal{T}_h(\Omega)$ belongs to $V_h(K_i):=V^h(\Omega)|_{K_i}$.

- Define

$$u_1^h(x,x/\epsilon)=-\Psi_j^h(x)\chi^j(x/\epsilon). \tag{4.5}$$

Note that this leads to a nonconforming approximation for u_1 in $\mathcal{T}_h(\Omega)$.

- Define

$$\tilde{\phi}_\epsilon^h=\sum_{k\in\{e,w,n,s\}}\tilde{\phi}_\epsilon^{k,h}, \tag{4.6}$$

where

$$\tilde{\phi}_\epsilon^{e,h}(x_1,x_2)=\varphi_e(x_1,x_2)\left(v_e\left(\frac{x_1-1}{\epsilon},\frac{x_2}{\epsilon}\right)-\chi_e^*\right)\Psi^h$$

(see (2.13) and (4.4)) and the others terms $\tilde{\phi}_\epsilon^{k,h}$ are defined similarly.

- Let $\bar{\phi}^h\in V_h(\Omega)$ be the solution of

$$\int_\Omega A\nabla\bar{\phi}^h\cdot\nabla v^h dx=0\quad\forall v^h\in V_0^h(\Omega)\quad\text{and}\quad\bar{\phi}^h=\chi^*\mu^h\ \text{ on }\ \partial\Omega. \tag{4.7}$$

The well posedness of (4.7) follows from Remark 4.1.

- Approximate ϕ_ϵ by $\phi_\epsilon^h=\tilde{\phi}_\epsilon^h+\bar{\phi}^h$ and finally define the numerical solution for (1.1) as

$$u_\epsilon^h=u_0^h+\epsilon u_1^h+\epsilon\phi_\epsilon^h. \tag{4.8}$$

Remark 4.1. By construction μ^h vanishes at the corners of Ω; therefore, $\chi^*\mu^h\in H^{1/2}(\partial\Omega)$. This implies that (4.7) is well posed. In addition, $\chi^*\mu^h\in V^h|_{\partial\Omega}$; hence, we can look for a numerical solution of (4.7) in $V^h(\Omega)$.

5. Finite element error analysis. The main results of this section are Theorems 5.1 and 5.2, which provide the a priori error estimates between the exact solution u_ϵ and its numerical approximation u_ϵ^h. Theorem 5.1 provides estimates on the a priori error estimates on the broken H^1-norm, while Theorem 5.2 provides the L_2-norm.

In this section we recall a few results on discrete Sobolev norms and spaces, prove the propositions and lemmas required in the proofs of Theorems 5.1 and 5.2, and then close the section proving Theorems 5.1 and 5.2.

5.1. Preliminaries. We first review general finite element results that are used later in this section. Let u_0 and u_0^h be the solutions of problems (2.4) and (4.1), respectively. We consider $\mathcal{P}_1$ or $\mathcal{Q}_1$ finite elements spaces for approximating u_0. We assume that the domain Ω is a convex polygon; therefore, problem (2.4) is $W^{2,p}$ regular for some $2<p$; see [25]. Standard finite element analysis provides

$$\|u_0-u_0^h\|_{1,p}\le ch\|u_0\|_{2,p}\text{ for }2\le p\le\infty, \tag{5.1}$$

$$\|u_0-u_0^h\|_{0,p}\le ch^2\|u_0\|_{2,p}\text{ for }2\le p<\infty, \tag{5.2}$$

and

$$\|u_0-u_0^h\|_{2,p,h}\le c\|u_0\|_{2,p}\text{ for }2\le p\le\infty; \tag{5.3}$$

see Corollary 7.1.12, inequality (7.5.4), and Theorem 4.4.20 in [12] to obtain (5.1), and standard duality arguments to obtain (5.2). To obtain the stability result (5.3), we first introduce the interpolation operator $\mathcal{I}^h$, i.e., the usual local pointwise interpolation $\mathcal{P}_1$ or $\mathcal{Q}_1$ in $\mathcal{T}_h(\Omega)$. Then, for each element $K \in \mathcal{T}_h(\Omega)$, we have

$$|u_0 - u_0^h|_{2,p,K} \leq |u_0 - \mathcal{I}^h u_0|_{2,p,K} + |\mathcal{I}^h u_0 - u_0^h|_{2,p,K}, \tag{5.4}$$

and using an interpolation error estimate (see Theorem 4.4.20 in [12]) we obtain

$$|u_0 - \mathcal{I}^h u_0|_{s,p,h} \leq c h^{2-s} |u_0|_{2,p,h} \text{ for } 1 \leq p \leq \infty \text{ for } 0 \leq s \leq 2. \tag{5.5}$$

Now apply an inverse inequality (see Lemma 4.5.3 in [12]) to obtain

$$|\mathcal{I}^h u_0 - u_0^h|_{2,p,K} \leq c h^{-1} \|\mathcal{I}^h u_0 - u_0^h\|_{1,p,K} \text{ for } 1 \leq p \leq \infty, \tag{5.6}$$

and hence, (5.3) follows from (5.4), (5.5), (5.6), and (5.1).

In order to estimate the L^2- and the broken H^1-norms of $u_1 - u_1^h$ (see Proposition 5.1) we apply Hölder's inequality on $u_1 - u_1^h = (\partial_{x_j} u_0 - \Psi_j^h)\chi^j$ to obtain $\|u_1 - u_1^h\|_{1,h} \leq c\|\partial_{x_j} u_0 - \Psi_j^h\|_{1,p,h}\|\chi^j\|_{1,q}$ and $\|u_1 - u_1^h\|_0 \leq c\|\partial_{x_j} u_0 - \Psi_j^h\|_{0,p}\|\chi^j\|_{0,q}$ for $\frac{1}{p} + \frac{1}{q} \leq \frac{1}{2}$. Hence, it requires estimating the error between Ψ^h and ∇u_0 on the L^p-norm and the broken $W^{1,p}$-norm; see Lemma 5.4. Lemma 5.4 requires estimating the errors between $A\nabla u_0 \cdot \eta$ and λ^h (see (4.2)) over Γ_k using fractional-order Sobolev norms.

Remark 5.1. Here, we review some facts about fractional-order Sobolev norms and spaces that are used throughout the text; see [25, 33].

Case $2 < p < \infty$. Since $W^{1-\frac{1}{p},p}(\Gamma_k) \hookrightarrow C^0(\Gamma_k)$, we define the spaces $W_{00}^{1-\frac{1}{p},p}(\Gamma_k) = \{\varphi \in W^{1-\frac{1}{p},p}(\Gamma_k);\ \varphi = 0 \text{ on } \partial\Gamma_k\}$ equipped with the norm $\|\cdot\|_{W_{00}^{1-\frac{1}{p},p}(\Gamma_k)} = \|\cdot\|_{W^{1-\frac{1}{p},p}(\Gamma_k)}$.

Case $p = 2$. We set $W_{00}^{1-\frac{1}{p},p}(\Gamma_k) = H_{00}^{1/2}(\Gamma_k)$ equipped with the norm $\|\cdot\|_{W_{00}^{1-\frac{1}{p},p}(\Gamma_k)} = \|\cdot\|_{H_{00}^{1/2}(\Gamma_k)}$; see [33] for the definition of $H_{00}^{1/2}(\Gamma_k)$.

Case $1 < p < 2$. We define $W_{00}^{1-\frac{1}{p},p}(\Gamma_k) = W^{1-\frac{1}{p},p}(\Gamma_k)$ equipped with the norm $\|\cdot\|_{W_{00}^{1-\frac{1}{p},p}(\Gamma_k)} = \|\cdot\|_{W^{1-\frac{1}{p},p}(\Gamma_k)}$.

We note that for Lipschitz domains, we have that $W_{00}^{1-\frac{1}{p},p}(\Gamma_k)$ are also equivalent to $[L^p(\Gamma_k), W_0^{1,p}(\Gamma_k)]_{1-1/p,p}$, i.e., via real interpolation theory on Banach spaces; see [25]. This property is used throughout the text to establish stabilities and a priori error estimates results.

The spaces $W_{00}^{1-\frac{1}{p},p}(\Gamma_k)$ also have the following important feature. Denote by $\tilde{\varphi}$ the extension by zero to $\partial\Omega \setminus \Gamma_k$ of a given function $\varphi \in W_{00}^{1-1/p,p}(\Gamma_k)$. Then, by the trace theorem and the lift theorem, Theorem 1.5.2.3 in [25], there exists a function $\psi_\varphi \in W^{1,p}(\Omega)$ such that $\psi_\varphi|_{\partial\Omega} = \tilde{\varphi}$ and

$$c_1\|\varphi\|_{W_{00}^{1-\frac{1}{p},p}(\Gamma_k)} \leq \|\psi_\varphi\|_{1,p} \leq c_2\|\tilde{\varphi}\|_{W^{1-\frac{1}{p},p}(\partial\Omega)} \leq c_3\|\varphi\|_{W_{00}^{1-\frac{1}{p},p}(\Gamma_k)}. \tag{5.7}$$

The dual space of $W_{00}^{1-\frac{1}{p},p}(\Gamma_k)$ is denoted by $W^{-1+\frac{1}{p},p'}(\Gamma_k)$, where $\frac{1}{p} + \frac{1}{p'} = 1$.

5.2. Finite element convergence analysis. We now prove the lemmas and propositions required to complete the proof of Theorems 5.1 and 5.2.

The following inverse inequality is required in the proofs of Lemma 5.3.

LEMMA 5.1. *Let $1 < p < \infty$ and $v^h \in Y^h_{0,k}$. Then*

$$\|v^h\|_{W_{00}^{1-\frac{1}{p},p}(\Gamma_k)} \leq ch^{-1}\|v^h\|_{W^{-1+\frac{1}{p'},p}(\Gamma_k)}. \tag{5.8}$$

Proof. Considering the inverse inequality given by Theorem 4.5.11 in [12] followed by the real interpolation method (see Proposition 12.1.5 in [12]), we have

$$\|v^h\|_{s,q,\partial\Omega} \leq ch^{-s}\|v^h\|_{0,q,\partial\Omega} \quad \forall\, v^h \in Y^h,\ 1 \leq p \leq \infty, \text{ and } 0 \leq s \leq 1. \tag{5.9}$$

Given $v^h \in Y^h_{0,k}$, let $\tilde{v}^h \in Y^h$ be the extension of v^h to $\partial\Omega \setminus \Gamma_k$ by zero. Using (5.7) and (5.9) we obtain

$$\begin{aligned}\|v^h\|_{W_{00}^{1-\frac{1}{p},p}(\Gamma_k)} &\leq c\|\tilde{v}^h\|_{W^{1-\frac{1}{p},p}(\partial\Omega)} \\ &\leq ch^{-1+\frac{1}{p}}\|\tilde{v}^h\|_{L^p(\partial\Omega)} = ch^{-1+\frac{1}{p}}\|v^h\|_{L^p(\Gamma_k)}.\end{aligned} \tag{5.10}$$

Let $\mathcal{P}_{0,k}$ denote the L^2 projection on $Y^h_{0,k}$, and assume that $v^h \in Y^h_{0,k}$. Then

$$\|v^h\|_{L^p(\Gamma_k)} = \sup_{\phi \in L^{p'}(\Gamma_k)} \frac{\langle v^h, \phi\rangle}{\|\phi\|_{L^{p'}(\Gamma_k)}} = \sup_{\phi \in L^{p'}(\Gamma_k)} \frac{\langle v^h, \mathcal{P}_{0,k}\phi\rangle}{\|\phi\|_{L^{p'}(\Gamma_k)}}.$$

We also have by Theorem 1 in [17] that

$$\|\mathcal{P}_{0,k}\phi\|_{L^{p'}(\Gamma_k)} \leq c\|\phi\|_{L^{p'}(\Gamma_k)}, \quad 1 \leq p' \leq \infty. \tag{5.11}$$

Thus,

$$\begin{aligned}\|v^h\|_{L^p(\Gamma_k)} &\leq c \sup_{\phi \in L^{p'}(\Gamma_k)} \frac{\|v^h\|_{W^{-1+\frac{1}{p'},p}(\Gamma_k)}\|\mathcal{P}_{0,k}\phi\|_{W_{00}^{1-\frac{1}{p'},p'}(\Gamma_k)}}{\|\mathcal{P}_{0,k}\phi\|_{L^{p'}(\Gamma_k)}} \\ &\leq ch^{-1+\frac{1}{p'}}\|v^h\|_{W^{-1+\frac{1}{p'},p}(\Gamma_k)},\end{aligned} \tag{5.12}$$

where in the last inequality we have used (5.10). Combining inequalities (5.10) and (5.12), we obtain (5.8). ∎

The following lemma provides stability and approximation results associated to $\mathcal{P}_{0,k}$, i.e., the L^2 projection on $Y^h_{0,k}$. These results are required in the proof of Lemma 5.3.

LEMMA 5.2. *Let $1 < p < \infty$ and $\mathcal{P}_{0,k} : W^{-1+\frac{1}{p},p'}(\Gamma_k) \to Y^h_{0,k}$ be the L^2 projection on $Y^h_{0,k}$. Then we have*

$$\|\mathcal{P}_{0,k}\phi\|_{W_{00}^{1-\frac{1}{p},p}(\Gamma_k)} \leq c\|\phi\|_{W_{00}^{1-\frac{1}{p},p}(\Gamma_k)} \quad \forall\, \phi \in W_{00}^{1-\frac{1}{p},p}(\Gamma_k), \tag{5.13}$$

$$\|\phi - \mathcal{P}_{0,k}\phi\|_{L^p(\Gamma_k)} \leq ch^{1-\frac{1}{p}}\|\phi\|_{W_{00}^{1-\frac{1}{p},p}(\Gamma_k)} \quad \forall\, \phi \in W_{00}^{1-\frac{1}{p},p}(\Gamma_k), \tag{5.14}$$

$$\|\mathcal{P}_{0,k}\phi\|_{W^{-1+\frac{1}{p},p'}(\Gamma_k)} \leq c\|\phi\|_{W^{-1+\frac{1}{p},p'}(\Gamma_k)} \quad \forall\, \phi \in W^{-1+\frac{1}{p},p'}(\Gamma_k), \tag{5.15}$$

and

$$\|\phi - \mathcal{P}_{0,k}\phi\|_{W^{-1+\frac{1}{p},p'}(\Gamma_k)} \leq ch^{1-\frac{1}{p}}\|\phi\|_{L^{p'}(\Gamma_k)} \quad \forall\ \phi \in L^{p'}(\Gamma_k). \tag{5.16}$$

Proof of (5.13). Observe that $\mathcal{P}_{0,k} : L^p(\Gamma_k) \to Y_{0,k}^h$ is stable in L^p and $W_0^{1,p}$, $1 \leq p \leq \infty$, i.e., $\|\mathcal{P}_{0,k}\phi\|_{L^p(\Gamma_k)} \leq c\|\phi\|_{L^p(\Gamma_k)}$ for all $\phi \in L^p(\Gamma_k)$, and $\|\mathcal{P}_{0,k}\phi\|_{W^{1,p}(\Gamma_k)} \leq c\|\phi\|_{W^{1,p}(\Gamma_k)}$ for all $\phi \in W_0^{1,p}(\Gamma_k)$, respectively; see Theorems 1 and 2 in [17]. Since $W_{00}^{1-1/p,p}(\Gamma_k)$ is equivalent to $[L^p(\Gamma_k), W_0^{1,p}(\Gamma_k)]_{1-1/p,p}$, the stability of $\mathcal{P}_{0,k}$ in the norm $W_{00}^{1-1/p,p}(\Gamma_k)$ follows from the real interpolation method; see Proposition 12.1.5 in [12]. The inequality (5.13) then follows.

Proof of (5.14). Let $\mathcal{Q}^h : L^p(\Gamma_k) \to V^h(\Omega)|_{\Gamma_k}$ denote the $\mathcal{P}_1$ or $\mathcal{Q}_1$ Clement interpolation operator defined by (2.13) in [41]. Then we have

$$\begin{aligned}
\|\phi - \mathcal{P}_{0,k}\phi\|_{L^p(\Gamma_k)} &\leq \|\phi - \mathcal{Q}^h\phi\|_{L^p(\Gamma_k)} + \|\mathcal{P}_{0,k}(\phi - \mathcal{Q}^h\phi)\|_{L^p(\Gamma_k)} \\
&\leq c\|\phi - \mathcal{Q}^h\phi\|_{L^p(\Gamma_k)} \text{ by (5.11)} \\
&\leq ch^{1-\frac{1}{p}}\|\phi\|_{W_{00}^{1-\frac{1}{p},p}(\Gamma_k)}, \qquad (5.17)
\end{aligned}$$

where to obtain (5.17) we have used Theorem 4.1 and Lemma 4.1 [41] followed by the real interpolation method.

Proofs of (5.15) *and* (5.16). These proofs follow from standard duality arguments, the properties of the L_2 projection $\mathcal{P}_{0,k}$, and estimates (5.13) and (5.14), respectively. ∎

The following lemma estimates the error between $\lambda = A\nabla u_0 \cdot \eta$ and its numerical approximation λ^h defined by (4.2). This lemma is used in the proof of Lemma 5.4.

LEMMA 5.3. *Assume that* $u_0 \in W^{2,p}(\Omega)$. *Then we have*

$$\|\lambda^h\|_{W_{00}^{1-\frac{1}{p},p}(\Gamma_k)}, \|\lambda - \lambda^h\|_{W_{00}^{1-\frac{1}{p},p}(\Gamma_k)} \leq c\|u_0\|_{2,p} \quad \textit{for} \quad 2 \leq p < \infty, \tag{5.18}$$

$$\|\lambda - \lambda^h\|_{W^{-1+\frac{1}{p'},p}(\Gamma_k)} \leq ch\|u_0\|_{2,p} \quad \textit{for} \quad 2 \leq p < \infty, \tag{5.19}$$

and

$$\|\lambda - \lambda^h\|_{L^p(\Gamma_k)} \leq ch^{1-\frac{1}{p}}\|u_0\|_{2,p} \quad \textit{for} \quad 2 \leq p \leq \infty. \tag{5.20}$$

Proof of (5.18). From Remark 2.1 and a trace theorem we have

$$\|\lambda\|_{W_{00}^{1-\frac{1}{p},p}(\Gamma_k)} \leq c\|u_0\|_{2,p}. \tag{5.21}$$

In order to prove inequality (5.18) observe that

$$\|\lambda - \lambda^h\|_{W_{00}^{1-\frac{1}{p},p}(\Gamma_k)} \leq \|\lambda\|_{W_{00}^{1-\frac{1}{p},p}(\Gamma_k)} + \|\lambda^h\|_{W_{00}^{1-\frac{1}{p},p}(\Gamma_k)}, \tag{5.22}$$

and

$$\|\lambda^h\|_{W_{00}^{1-\frac{1}{p},p}(\Gamma_k)} = \sup_{\phi \in W^{-1+\frac{1}{p},p'}(\Gamma_k)} \frac{\langle \lambda^h, \phi\rangle}{\|\phi\|_{W^{-1+\frac{1}{p},p'}(\Gamma_k)}}.$$

Since $\lambda^h \in Y^h_{0,k}$, then $\langle \lambda^h, \phi\rangle = \langle \lambda^h, \mathcal{P}_{0,k}\phi\rangle$, and using (5.15) we obtain

$$\|\lambda^h\|_{W_{00}^{1-\frac{1}{p},p}(\Gamma_k)} \le c \sup_{\phi\in W^{-1+\frac{1}{p},p'}(\Gamma_k)} \frac{\langle \lambda^h, \mathcal{P}_{0,k}\phi\rangle}{\|\mathcal{P}_{0,k}\phi\|_{W^{-1+\frac{1}{p},p'}(\Gamma_k)}}. \tag{5.23}$$

Let $E : W^{-1+\frac{1}{p},p'}(\partial\Omega) \to W^{1,p'}(\Omega)$ be a stable lift extension and let $\mathcal{Q}_h$ (see [41]) be the modified Clement L_2-quasi-projection satisfying the discrete boundary condition. Then $E_h = \mathcal{Q}_h E : Y^h \to V^h(\Omega)$ (see (5.5) in [41]) satisfies

$$\|E_h g\|_{1,p'} \le c\|g\|_{W^{1-\frac{1}{p'},p'}(\partial\Omega)}$$

for $g \in Y^h$. Hence, if $g^h \in Y^h_{0,k}$ and $\tilde{g}^h$ denotes the extension of g^h by zero to $\partial\Omega \setminus \Gamma_k$, it follows that

$$\|E_h\tilde{g}^h\|_{1,p'} \le c\|g^h\|_{W_{00}^{1-\frac{1}{p'},p'}(\Gamma_k)}. \tag{5.24}$$

Let $\tilde{\mathcal{P}}_{0,k}\phi$ denote the discrete extension of $\mathcal{P}_{0,k}\phi$ to $\partial\Omega \setminus \Gamma_k$ by zero. From the definition of λ^h (see (4.2)) inequalities (5.1), and (5.24), and the inverse inequality (5.8), we obtain

$$\begin{aligned}\langle \lambda^h, \mathcal{P}_{0,k}\phi\rangle &= \langle \lambda, \mathcal{P}_{0,k}\phi\rangle + \int_\Omega A\nabla(u_0^h - u_0)\cdot\nabla(E_h\tilde{\mathcal{P}}_{0,k}\phi)dx \\ &\le \|\lambda\|_{W_{00}^{1-\frac{1}{p},p}(\Gamma_k)}\|\mathcal{P}_{0,k}\phi\|_{W^{-1+\frac{1}{p},p'}(\Gamma_k)} + ch\|u_0\|_{2,p}\|\mathcal{P}_{0,k}\phi\|_{W_{00}^{1-\frac{1}{p'},p'}(\Gamma_k)} \\ &\le \left(\|\lambda\|_{W_{00}^{1-\frac{1}{p},p}(\Gamma_k)} + c\|u_0\|_{2,p}\right)\|\mathcal{P}_{0,k}\phi\|_{W^{-1+\frac{1}{p},p'}(\Gamma_k)}. \end{aligned} \tag{5.25}$$

The inequality (5.18) follows from (5.23), (5.25), (5.22), and (5.21).

Proof of (5.19). We observe that

$$\begin{aligned}\|\lambda - \lambda^h\|_{W^{-1+\frac{1}{p'},p}(\Gamma_k)} &= \sup_{\phi\in W_{00}^{1-\frac{1}{p'},p'}(\Gamma_k)} \frac{\langle \lambda - \lambda^h, \phi\rangle}{\|\phi\|_{W_{00}^{1-\frac{1}{p'},p'}(\Gamma_k)}} \\ &\le \sup_{\phi\in W_{00}^{1-\frac{1}{p'},p'}(\Gamma_k)} c\frac{\langle \lambda - \lambda^h, \mathcal{P}_{0,k}\phi\rangle}{\|\mathcal{P}_{0,k}\phi\|_{W_{00}^{1-\frac{1}{p'},p'}(\Gamma_k)}} + \sup_{\phi\in W_{00}^{1-\frac{1}{p'},p'}(\Gamma_k)} c\frac{\langle \lambda - \lambda^h, \phi - \mathcal{P}_{0,k}\phi\rangle}{\|\phi\|_{W_{00}^{1-\frac{1}{p'},p'}(\Gamma_k)}}. \end{aligned} \tag{5.26}$$

In order to estimate the first term on the right-hand side of (5.26), we use the definition of λ and λ^h and inequality (5.24) to obtain

$$\begin{aligned}\langle \lambda - \lambda^h, \mathcal{P}_{0,k}\phi\rangle &= \int_\Omega A\nabla(u_0^h - u_0)\cdot\nabla(E_h\tilde{\mathcal{P}}_{0,k}\phi)dx \\ &\le ch\|u_0\|_{2,p}\|\mathcal{P}_{0,k}\phi\|_{W_{00}^{1-\frac{1}{p'},p'}(\Gamma_k)}. \end{aligned} \tag{5.27}$$

For estimating the second term on the right-hand side of (5.26), we use (5.18) and (5.16) to obtain

$$\begin{aligned}\langle \lambda - \lambda^h, \phi - \mathcal{P}_{0,k}\phi\rangle &\le \|\lambda - \lambda^h\|_{W_{00}^{1-\frac{1}{p},p}(\Gamma_k)}\|\phi - \mathcal{P}_{0,k}\phi\|_{W^{-1+\frac{1}{p},p'}(\Gamma_k)} \\ &\le ch\|u_0\|_{2,p}\|\phi\|_{W_{00}^{1-\frac{1}{p'},p'}(\Gamma_k)}, \end{aligned} \tag{5.28}$$

and inequality (5.19) follows from (5.26), (5.27), and (5.28).

Proof of (5.20).

Case $2 \leq p < \infty$. We have

$$\|\lambda - \lambda^h\|_{L^p(\Gamma_k)} \leq \sup_{\phi \in L^{p'}(\Gamma_k)} \frac{\langle \lambda - \lambda^h, \phi - \mathcal{P}_{0,k}\phi \rangle}{\|\phi\|_{L^{p'}(\Gamma_k)}} + \sup_{\phi \in L^{p'}(\Gamma_k)} \frac{\langle \lambda - \lambda^h, \mathcal{P}_{0,k}\phi \rangle}{\|\phi\|_{L^{p'}(\Gamma_k)}}. \tag{5.29}$$

The first term on the right-hand side of (5.29) is bounded as follows:

$$\begin{aligned} \sup_{\phi \in L^{p'}(\Gamma_k)} \frac{\langle \lambda - \lambda^h, \phi - \mathcal{P}_{0,k}\phi \rangle}{\|\phi\|_{L^{p'}(\Gamma_k)}} &\leq \sup_{\phi \in L^{p'}(\Gamma_k)} \frac{\|\lambda - \lambda^h\|_{W_{00}^{1-\frac{1}{p},p}(\Gamma_k)} \|\phi - \mathcal{P}_{0,k}\phi\|_{W^{-1+\frac{1}{p},p'}(\Gamma_k)}}{\|\phi\|_{L^{p'}(\Gamma_k)}} \\ &\leq c h^{1-\frac{1}{p}} \|u_0\|_{2,p}. \end{aligned} \tag{5.30}$$

Here we have used (5.16) and (5.18) to arrive at (5.30). In order to estimate the second term on the right-hand side of (5.29) we use the definition of λ and λ^h to obtain

$$\begin{aligned} \sup_{\phi \in L^{p'}(\Gamma_k)} \frac{\langle \lambda - \lambda^h, \mathcal{P}_{0,k}\phi \rangle}{\|\phi\|_{L^{p'}(\Gamma_k)}} &\leq c \sup_{\phi \in L^{p'}(\Gamma_k)} \frac{\int_\Omega A\nabla(u_0 - u_0^h) \cdot \nabla(E_h \tilde{\mathcal{P}}_{0,k}\phi) dx}{\|\mathcal{P}_{0,k}\phi\|_{L^{p'}(\Gamma_k)}} \\ &\leq ch \sup_{\phi \in L^{p'}(\Gamma_k)} \frac{\|u_0\|_{2,p} \|\mathcal{P}_{0,k}\phi\|_{W_{00}^{1-\frac{1}{p'},p'}}}{\|\mathcal{P}_{0,k}\phi\|_{L^{p'}(\Gamma_k)}} \\ &\leq ch^{1-\frac{1}{p}} \|u_0\|_{2,p}, \text{ by (5.10)}. \end{aligned}$$

Case $p = \infty$. Let $z \in \Gamma_k$; then

$$|\lambda(z) - \lambda^h(z)| \leq |\lambda(z) - \mathcal{P}_{0,k}\lambda(z)| + |\lambda^h(z) - \mathcal{P}_{0,k}\lambda(z)|. \tag{5.31}$$

To estimate the first term on the right-hand side of (5.31), we use that the L^2 projection is stable in any L^p-norm, $1 \leq p \leq \infty$; see Lemma 3.5 in [47].

To estimate the second term on the right-hand side of (5.31) we use the techniques developed in [47]. Let $E_z \subset \Gamma_k$ denote an edge of an element $K_z \in \mathcal{T}^h(\Omega)$ such that $z \in E_z$, and define δ_z as the polynomial of degree 1 on E_z such that

$$\int_{E_z} \delta_z(s) v(s) ds = v(z) \text{ for any } v \text{ polynomial of degree 1.}$$

Regard δ_z as extended by zero to $\Gamma_k \setminus E_z$, and denote by $\tilde{\delta}_z^h \in V^h(\Omega)$ the extension by zero of $\mathcal{P}_{0,k}\delta_z$ to Ω. Then we have

$$\begin{aligned} \lambda^h(z) - \mathcal{P}_{0,k}\lambda(z) &= \int_{\Gamma_k} \mathcal{P}_{0,k}(\lambda^h - \lambda)\delta_z ds = \int_{\Gamma_k} (\lambda^h - \lambda)\mathcal{P}_{0,k}\delta_z ds \\ &= \int_\Omega A\nabla(u_0 - u_0^h) \cdot \nabla \tilde{\delta}_z^h dx, \end{aligned} \tag{5.32}$$

where we have used the definition of λ^h to obtain (5.32). From (5.1) and (5.32) it follows that

$$|\lambda^h(z) - \mathcal{P}_{0,k}\lambda(z)| \leq ch \|u_0\|_{2,\infty} \|\tilde{\delta}_z^h\|_{1,1}.$$

Using an inverse inequality followed by a discrete Sobolev inequality for trivial extensions, we have

$$\|\tilde{\delta}_z^h\|_{1,1} \le ch^{-1}\|\tilde{\delta}_z^h\|_{0,1} \le c\|\mathcal{P}_{0,k}\delta_z\|_{0,1,\Gamma_k}.$$

Finally, we use the fact that $\|\mathcal{P}_{0,k}\delta_z\|_{0,1,\Gamma_k} \le c$ (see Lemma 3.5 in [47]) to obtain (5.20). ∎

The next lemma estimates the error between ∇u_0 and its numerical approximation Ψ^h. This proposition is required in the proof of Proposition 5.1.

LEMMA 5.4. *Let u_0 and Ψ^h be defined by (2.4) and (4.4), respectively. Assume that $u_0 \in W^{2,p}(\Omega)$ for $2 \le p \le \infty$. Then*

$$\|(\nabla u_0 - \Psi^h)\cdot\nu\|_{0,p} \le ch\|u_0\|_{2,p} \quad \forall\, \nu \in \mathbb{R}^2 \text{ with } |\nu| = 1 \tag{5.33}$$

and

$$\|(\nabla u_0 - \Psi^h)\cdot\nu\|_{1,p,h} \le c\|u_0\|_{2,p} \quad \forall\, \nu \in \mathbb{R}^2 \text{ with } |\nu| = 1. \tag{5.34}$$

Proof of (5.33). From the triangular inequality we have

$$\|(\nabla u_0 - \Psi^h)\cdot\nu\|_{0,p} \le \|(\nabla u_0 - \nabla u_0^h)\cdot\nu\|_{0,p} + \|(\nabla u_0^h - \Psi^h)\cdot\nu\|_{0,p}. \tag{5.35}$$

The first term on the right-hand side of (5.35) is estimated by (5.1), while for the second term we use (4.4) to obtain

$$\|(\nabla u_0^h - \Psi^h)\cdot\nu\|_{0,p} \le c \sum_{k\in\{e,w,n,s\}} \|E_k^h(\mu^h - \nabla u_0^h\cdot\eta^k)\|_{0,p}.$$

Consider the case $k = e$ and $\mathcal{Q}_1$ conforming finite elements for approximating u_0; the other cases, i.e., $k \in \{w,n,s\}$ or $\mathcal{P}_1$ conforming finite elements, can be treated similarly. The function $E_e^h(\mu^h - \frac{\partial u_0^h}{\partial x_1})$ is piecewise linear in the x_1 direction and equal to zero for $x_1 \le 1-h$; see the definition of E_e^h in (4.4). Hence, using a discrete Sobolev inequality for trivial extensions we obtain

$$\|E_e^h(\mu^h - \nabla u_0^h\cdot\eta^k)\|_{0,p} \le h^{\frac{1}{p}} \left\|\partial_{x_1}u_0^h - \mu^h\right\|_{0,p,\Gamma_e} \quad \text{if } \ 2 \le p < \infty$$

or

$$\|E_e^h(\mu^h - \nabla u_0^h\cdot\eta^k)\|_{0,\infty} \le \left\|\partial_{x_1}u_0^h - \mu^h\right\|_{0,\infty,\Gamma_e} \quad \text{if } \ p = \infty.$$

Case $2 \le p < \infty$. The triangular inequality gives

$$\left\|\partial_{x_1}u_0^h - \mu^h\right\|_{0,p,\Gamma_e} \le \left\|\partial_{x_1}u_0^h - \partial_{x_1}u_0\right\|_{0,p,\Gamma_e} + \left\|\partial_{x_1}u_0 - \mu^h\right\|_{0,p,\Gamma_e}. \tag{5.36}$$

In order to estimate the first term on the right-hand side of (5.36), let K be an element of $\mathcal{T}_h(\Omega)$ containing an edge $E \subset \Gamma_k$. Applying a trace theorem we have

$$\begin{aligned}&\left\|\partial_{x_1}u_0^h - \partial_{x_1}u_0\right\|_{0,p,E}\\ &\qquad \le c\left(h^{-1}\left\|\partial_{x_1}u_0^h - \partial_{x_1}u_0\right\|_{0,p,K}^p + h^{p-1}\|\partial_{x_1}u_0^h - \partial_{x_1}u_0\|_{1,p,K}^p\right)^{\frac{1}{p}}.\end{aligned} \tag{5.37}$$

From (5.37), (5.1), and (5.3) we obtain

$$\left\|\partial_{x_1}u_0^h - \partial_{x_1}u_0\right\|_{0,p,\Gamma_e} \le ch^{1-\frac{1}{p}}\|u_0\|_{2,p}. \tag{5.38}$$

We now estimate the second term on the right-hand side of (5.36). Let $\lambda = A\nabla u_0 \cdot \eta$, and let λ_h and μ_h be defined by (4.2) and (4.3), respectively. Using that the tangential derivative of u_0 vanishes on $\partial\Omega$, we obtain

$$\partial_{x_1} u_0 - \mu^h = A_{11}(\lambda - \lambda^h). \tag{5.39}$$

Hence, from (5.20) we have

$$\|\partial_{x_1} u_0 - \mu^h\|_{0,p,\Gamma_e} \leq A_{11}\|\lambda - \lambda^h\|_{0,p,\Gamma_e} \leq ch^{1-\frac{1}{p}}\|u_0\|_{2,p}. \tag{5.40}$$

From (5.36), (5.38), and (5.40) we obtain

$$\|E_e(\mu^h - \nabla u_0^h \cdot \eta^e)\|_{0,p} \leq ch\|u_0\|_{2,p},$$

and thus the estimate (5.33) holds for $2 \leq p \leq \infty$.

Case $p = \infty$. We have

$$\left\|\partial_{x_1} u_0^h - \mu^h\right\|_{0,\infty,\Gamma_e} \leq \|\partial_{x_1} u_0^h - \partial_{x_1} u_0\|_{0,\infty,\Gamma_e} + \|\partial_{x_1} u_0 - \mu^h\|_{0,\infty,\Gamma_e},$$

and applying (5.1), (5.39), and (5.20) we have

$$\left\|\partial_{x_1} u_0 - \mu^h\right\|_{0,\infty,\Gamma_e} \leq ch\|u_0\|_{2,\infty}.$$

Hence, estimate (5.33) follows for $p = \infty$.

Proof of (5.34). We have

$$\begin{aligned}\|(\nabla u_0^h - \Psi^h)\cdot\nu\|_{0,p} &\leq c\|(\nabla u_0 - \Psi^h)\cdot\nu\|_{0,p} + \|(\nabla u_0 - \nabla u_0^h)\cdot\nu\|_{0,p} \\ &\leq ch\|u_0\|_{2,p} \quad \text{by (5.1) and (5.33)},\end{aligned} \tag{5.41}$$

and from an inverse inequality (see Lemma 4.5.3 in [12]) it follows that

$$\|(\nabla u_0^h - \Psi^h)\cdot\nu\|_{1,p,h} \leq c\|u_0\|_{2,p}. \tag{5.42}$$

Since

$$\|(\nabla u_0 - \Psi^h)\cdot\nu\|_{1,p,h} \leq c\left(\|(\nabla u_0^h - \nabla u_0)\cdot\nu\|_{1,p,h} + \|(\nabla u_0^h - \Psi^h)\cdot\nu\|_{1,p,h}\right),$$

we obtain (5.34) from (5.3) and (5.42). ∎

The following proposition is required in the proofs of Theorems 5.1 and 5.2.

PROPOSITION 5.1. *Let u_1 and u_1^h be defined by* (2.5) *and* (4.5)*, respectively. Assume that $u_0 \in W^{2,p}(\Omega)$ and $\chi^j \in W^{1,q}_{per}(Y)$ for $\frac{1}{p} + \frac{1}{q} \leq \frac{1}{2}$ and $2 \leq p, q \leq \infty$. Then there exists a constant c independent of ϵ and h such that*

$$|u_1 - u_1^h|_{1,h} \leq c\left(\frac{h^2}{\epsilon^2} + 1\right)^{1/2} \max_j \|\chi^j\|_{1,q,Y}\|u_0\|_{2,p} \tag{5.43}$$

and

$$\|u_1 - u_1^h\|_0 \leq ch \max_j \|\chi^j\|_{1,q,Y}\|u_0\|_{2,p}. \tag{5.44}$$

Proof of (5.43). We have

(5.45)

$$\begin{aligned}&|u_1 - u_1^h|_{1,h}^2 \\ &\leq 2\sum_{K\in\mathcal{T}_h(\Omega)}\int_K \sum_{i\in 1,2}((\partial_{x_j} u_0 - \Psi_j^h)\partial_{x_i}\chi^j(\cdot/\epsilon))^2 + (\chi^j(\cdot/\epsilon)\cdot\partial_{x_i}(\partial_{x_j} u_0 - \Psi_j^h))^2 dx.\end{aligned}$$

For the first term on the right-hand side of (5.45) we have

$$\sum_{K\in\mathcal{T}_h(\Omega)}\int_K((\partial_{x_j}u_0-\Psi_j^h)\partial_{x_i}\chi^j(\cdot/\epsilon))^2dx \le |\partial_{x_j}u_0-\Psi_j^h|_{0,p}^2\|\partial_{x_i}\chi^j(\cdot/\epsilon)\|_{0,q}^2$$

$$\le \epsilon^{-2}|\partial_{x_j}u_0-\Psi_j^h|_{0,p}^2\|\chi^j\|_{1,q,Y}^2 \le c\epsilon^{-2}h^2\|u_0\|_{2,p}^2\|\chi^j\|_{1,q,Y}^2, \tag{5.46}$$

where we have used (5.33) to obtain (5.46).

For the second term on the right-hand side of (5.45) we use Hölder's inequality to obtain $\|\chi^j\partial_{x_i}(\partial_{x_j}u_0-\Psi_j^h)\|_0^2 \le \|\chi^j\|_{0,q}^2|\partial_{x_j}u_0-\Psi_j^h|_{1,p,h}^2$.

Proof of (5.44). This proof follows from a direct application of Hölder's inequality and the approximation error estimate (5.33). ∎

The next proposition is required in the proofs of Theorems 5.1 and 5.2.

PROPOSITION 5.2. *Let $\tilde{\phi}_\epsilon$ and $\tilde{\phi}_\epsilon^h$ be defined by* (2.14) *and* (4.6)*, respectively. Let the functions v_k be defined as in subsection* 2.2.1*. Assume that $u_0 \in W^{2,p}(\Omega)$ and $v_k \in W^{1,\tilde{q}}(G_k)$ for $\frac{1}{p}+\frac{1}{\tilde{q}} \le \frac{1}{2}$ and $2 \le p,\tilde{q} \le \infty$. Then*

$$|\tilde{\phi}_\epsilon-\tilde{\phi}_\epsilon^h|_{1,h} \le c\left(\frac{h^2}{\epsilon^2}+1\right)^{\frac{1}{2}}\max_k\|v_k\|_{1,\tilde{q},G_k}\|u_0\|_{2,p} \tag{5.47}$$

and

$$\|\tilde{\phi}_\epsilon-\tilde{\phi}_\epsilon^h\|_0 \le ch\max_k\|v_k-\chi_k^*\|_{0,\tilde{q},G_k}\|u_0\|_{2,p}. \tag{5.48}$$

Proof. From the definition of $\tilde{\phi}_\epsilon$ and $\tilde{\phi}_\epsilon^h$ we have

$$|\tilde{\phi}_\epsilon-\tilde{\phi}_\epsilon^h|_{1,h} \le \sum_{k\in\{e,w,n,s\}}|\tilde{\phi}_\epsilon^k-\tilde{\phi}_\epsilon^{k,h}|_{1,h},$$

and using the arguments similar to the ones given in the proof of Proposition 5.1, the proof follows. ∎

We now prove the last proposition used in the proof of Theorems 5.1 and 5.2.

PROPOSITION 5.3. *Let $\bar{\phi}$ and $\bar{\phi}^h$ be defined by* (2.15) *and* (4.7)*, respectively, and assume that $u_0 \in H^2(\Omega)$. Then we have*

$$\|\bar{\phi}-\bar{\phi}^h\|_1 \le c\|u_0\|_2 \tag{5.49}$$

and

$$\|\bar{\phi}-\bar{\phi}^h\|_0 \le ch\|u_0\|_2. \tag{5.50}$$

Proof of (5.49). Consider the triangular inequality

$$\|\bar{\phi}-\bar{\phi}^h\|_1 \le \|\bar{\phi}^h-\psi\|_1+\|\bar{\phi}-\psi\|_1,$$

where $\psi \in H^1(\Omega)$ is defined as the weak solution of

$$-\nabla\cdot A\nabla\psi = 0 \ \text{ in } \ \Omega \quad \text{and} \quad \psi = \chi^*\mu^h \ \text{ on } \ \partial\Omega. \tag{5.51}$$

We note that problem (5.51) is well defined since $\chi^*\mu^h \in H^{1/2}(\partial\Omega)$; see Remark 4.1. From regularity theory, (5.18), (5.39), and Remark 2.1 we have the following estimate:

$$\|\psi\|_1 \le \sum_k c\|\chi^*\mu^h\|_{H_{00}^{1/2}(\Gamma_k)} \le c\|u_0\|_2. \tag{5.52}$$

Note also that $\chi^* \mu^h$ belongs to Y^h. Hence, by using (5.52) and standard finite element analysis for conforming finite element discretizations we obtain

$$|\bar{\phi}^h - \psi|_1 \leq c \|u_0\|_2.$$

Finally, from regularity theory, (5.39), and (5.20) we obtain

$$\begin{aligned} |\bar{\phi} - \psi|_1 &\leq \|\chi^* \mu^h - \chi^* \partial_\eta u_0\|_{H^{1/2}(\partial\Omega)} \\ &\leq \sum_k \|\chi^* \mu^h - \chi^* \partial_\eta u_0\|_{H_{00}^{1/2}(\Gamma_k)} \leq c \|u_0\|_2. \end{aligned}$$

Proof of (5.50). From the triangular inequality

$$\|\bar{\phi} - \bar{\phi}^h\|_0 \leq c \|\bar{\phi} - \psi\|_0 + \|\bar{\phi}^h - \psi\|_0,$$

and from standard finite element analysis and (5.52) we obtain

$$\|\bar{\phi}^h - \psi\|_0 \leq ch \|\psi\|_1 \leq ch \|u_0\|_2.$$

Finally, from Theorem 6.1 in [38], (5.39), and (5.19) we obtain

$$\begin{aligned} \|\bar{\phi} - \psi\|_0 &\leq c \left(\sum_k \|\chi^* \partial_\eta u_0 - \chi^* \mu^h\|^2_{H^{-1/2}(\Gamma_k)} \right)^{1/2} \\ &\leq ch \|u_0\|_2 \quad \text{by (5.19).} \end{aligned}$$

∎

Finally, we prove Theorems 5.1 and 5.2.

THEOREM 5.1. *Assume the same conditions of Theorem* 3.1*, and let* u_ϵ *be defined by* (4.8)*. Then there exists a constant* c *independent of* ϵ*,* u_0*, and* h *such that*

$$|u_\epsilon - u_\epsilon^h|_{1,h} \leq c(h + \epsilon^{1+\hat{\delta}}) \|u_0\|_{2,p} \tag{5.53}$$

and

$$\|u_\epsilon - u_\epsilon^h\|_0 \leq c(h^2 + \epsilon^{1+\hat{\delta}}) \|u_0\|_{2,p}. \tag{5.54}$$

Proof. From the triangular inequality we have

$$\begin{aligned} |u_\epsilon - u_\epsilon^h|_{1,h} \leq |u_\epsilon - u_0 - u_1 - \phi_\epsilon|_1 + |u_0 - u_0^h|_1 + \epsilon |u_1 - u_1^h|_{1,h} \\ + \epsilon |\tilde{\phi}_\epsilon - \tilde{\phi}_\epsilon^h|_{1,h} + \epsilon |\bar{\phi} - \bar{\phi}^h|_1, \end{aligned}$$

and (5.53) follows from Theorem 3.1, the approximation error (5.1), and Propositions 5.1, 5.2, and 5.3. Using similar arguments and (5.2) the estimate (5.54) follows. ∎

THEOREM 5.2. *Assume the same conditions of Theorem* 3.2*, and let* u_ϵ *be defined by* (4.8)*. Then there exists a constant* c *independent of* ϵ*,* u_0*, and* h *such that*

$$\|u_\epsilon - u_\epsilon^h\|_0 \leq c(h^2 + \epsilon^{\frac{3}{2}} + \epsilon h) \|u_0\|_{3,p}.$$

Proof. Repeat the proof of Theorem 5.1 and use Theorem 3.2. ∎

6. Numerical results. The objective of this section is to compare the difference between u_ϵ and u_ϵ^h on the L^2-norm and on the broken H^1-semi-norm. We note, however, that an explicit formula for u_ϵ is not known; hence, we replace u_ϵ by a regular $\mathcal{P}_1$ conforming finite element solution of the problem (1.1) on a very fine uniform mesh of size h_f; denote this numerical solution by u_ϵ^*.

We note that we have assumed in sections 4 and 5 that the cell and boundary layer problems were solved exactly in order to define χ, A, v_k, and χ^*. In [44, 46] we have developed an accurate numerical algorithm for approximating these values and, as a result, the errors between u_ϵ^h and u_ϵ^* do not depend on whether the approximated or the exact values were used or not. We now describe how to approximate χ, A, v_k, χ^*, and $\tilde{\phi}_\epsilon$ and display only the changes that should be made in the algorithm in section 4 in order to obtain the numerical results of this section.

- Replace χ^j by the finite element solution of the problem (2.2) using a $\mathcal{P}_1$ or $\mathcal{Q}_1$ conforming finite element method on a partition $\mathcal{T}_{\hat{h}}(Y)$.
- Replace A_{ij} by $\frac{1}{|Y|}\int_Y a_{lm}(y)\frac{\partial}{\partial y_l}(y_i-\chi^i)\frac{\partial}{\partial y_m}(y_j-\chi^j)dy$; see (2.3).
- Let τ be a positive integer and $G_e^\tau = \{y \in \mathbb{R}^2;\ -\tau \le y_1 \le 0 \text{ and } 0 \le y_2 \le 1\}$. Using a $\mathcal{P}_1$ or $\mathcal{Q}_1$ conforming finite element, replace v_e by the finite element solution associated to the following problem:

$$\begin{aligned}
&-\nabla_y \cdot a(y)\nabla_y v_e = 0 \ \text{ in } \ G_e^\tau,\\
&v_e(y) = \chi^1(1/\epsilon, y_2) \ \text{ on } \ \{y \in G_e^\tau, y_1 = 0\},\\
&\partial_\eta v_e = 0 \ \text{ on } \{y \in G_e^\tau;\ y_1 = -\tau\},\\
&\text{and } v_e(y_1, 0) = v_k(y_1, 1) \ \text{ for } \ -\tau \le y_1 \le 0,
\end{aligned}$$

and replace χ_e^* by $\int_0^1 v_e(-\tau, y_2)dy_2$. We replace $\tilde{\phi}_\epsilon^e$ by

$$\tilde{\phi}_\epsilon^e(x_1,x_2) = \begin{cases} \left(v_e\left(\frac{x_1-1}{\epsilon}, \frac{x_2}{\epsilon}\right) - \chi_e^*\right)\Psi^h & \text{if } \ x_1 > 1-\epsilon\tau,\\ 0 & \text{otherwise.}\end{cases}$$

The other cases $k \in \{w, n, s\}$ are treated similarly.

In the numerical experiments we use a $\mathcal{P}_1$ conforming finite element with a uniform mesh $\hat{h} = \frac{1}{128}$ and $\tau = 2$.

For the first test of experiments we consider a smooth coefficient $a(y_1, y_2)$ case (see [26]) where

$$a(y) = \left(\frac{2+1.8\sin(2\pi y_1)}{2+1.8\cos(2\pi y_2)} + \frac{2+\sin(2\pi y_2)}{2+1.8\sin(2\pi y_1)}\right) I_{2\times 2} \ \text{ and } \ f(x) = -1.$$

Table 6.1 shows the errors between u_ϵ^* and u_ϵ^h. We see that for $\epsilon << h$ the errors are of order h^2 and h on the L^2-norm and the on broken H^1-semi-norm, respectively. We can see that when we fix h and decrease ϵ the errors then change very little. This is evidence that the dominant error is related to h, not to ϵ. Also, looking at the diagonal values of Table 6.1 we see that the numerical errors are in agreement with the theoretical convergent rates of Theorems 5.1 and 5.2.

Table 6.2 shows the level of approximation when each term of $u_\epsilon = u_0^h + \epsilon u_1^h + \epsilon\bar{\phi}^h + \epsilon\tilde{\phi}_\epsilon^h$ is taken into account separately. We can see that the numerical boundary layer treatment $\bar{\phi}^h$ is very effective on the L_2-norm rather than on the broken H^1-semi-norm. The improvement on the L^2-norm is evidence that we were able to obtain, through the proper calculation of χ^*, the L^2 asymptotic behavior of the boundary corrector θ_ϵ in the interior of the domain Ω. We also see from Table 6.2 that the

TABLE 6.1

$\|u_\epsilon^* - u_\epsilon^h\|_0$ error, $h_f = 1/2048$

$\epsilon \downarrow \quad h \rightarrow$	1/8	1/16	1/32	1/64
1/16	2.7085e-04	7.7993e-05		
1/32	2.6300e-04	6.6246e-05	1.7773e-05	
1/64	2.5388e-04	5.8069e-05	1.6020e-05	1.2137e-05

$|u_\epsilon^* - u_\epsilon^h|_{1,h}$ error,$h_f = 1/2048$

1/16	0.0097	0.0067		
1/32	0.0086	0.0051	0.0036	
1/64	0.0086	0.0044	0.0025	0.0018

TABLE 6.2

$\epsilon = 1/64$, $h = 1/32$, $h_f = 1/1024$

	$\|\cdot\|_0$	$\|\cdot\|_{1,h}$
$u_\epsilon^* - u_0^h$	0.0287	0.0215
$u_\epsilon^* - u_0^h - \epsilon u_1^h$	0.0213	0.0026
$u_\epsilon^* - u_0^h - \epsilon u_1^h - \epsilon\bar{\phi}^h$	5.0450e-05	0.0026
$u_\epsilon^* - u_0^h - \epsilon u_1^h - \epsilon(\bar{\phi}^h + \tilde{\phi}_\epsilon^h)$	5.1865e-05	0.0025

term $\tilde{\phi}_\epsilon$ ($\bar{\phi}^h$ and $\tilde{\phi}_\epsilon$) forces the final approximation u_ϵ^h to satisfy the zero Dirichlet boundary condition without deteriorating the L^2-norm (broken H^1-semi-norm).

The second test of experiments, reported in Table 6.3, consider the discontinuous coefficient $a(y_1, y_2)$ case,

$$a(y) = \begin{cases} 2 & \text{if } 2/5 < y_1 < 3/5 \text{ or } 2/5 < y_2 < 3/5, \\ 1 & \text{otherwise,} \end{cases} \qquad \text{and } f = -1,$$

and we obtain the same behavior as above.

TABLE 6.3

$\|u_\epsilon^* - u_\epsilon^h\|_0$ error, $h_f = 1/2000$

$\epsilon \downarrow \quad h \rightarrow$	1/10	1/20	1/40
1/20	4.8318e-04	1.3043e-04	
1/40	4.7578e-04	1.1954e-04	3.0805e-05
1/64	2.5388e-04	5.9446e-05	1.4414e-05

$|u_\epsilon^* - u_\epsilon^h|_{1,h}$ error, $h_f = 1/2000$

$\epsilon \downarrow \quad h \rightarrow$	1/10	1/20	1/40
1/20	0.0180	0.0092	
1/40	0.0179	0.0090	0.0046
1/64	0.0086	0.0045	0.0026

Although the convergence analysis presented here is not intended for the quasi-periodic case $a_{ij}(x, x/\epsilon)$, the numerical approximation presented here can be generalized for this case. This would be done by approximating the matrix $a(x, x/\epsilon)$ by $\sum_{K_j \in \mathcal{T}_h(\Omega)} a_j(x/\epsilon) I_{K_j}(x)$, where $a_j(x/\epsilon) = \frac{1}{|K_j|} \int_{z \in K_j} a_j(z, x/\epsilon) dz$, and I_{K_j} is the characteristic function of K_j. In such case, it is required to solve a cell problem for each element K_j.

7. Conclusions. We perform the convergence analysis for the proposed numerical method for approximating the solution of (1.1). The error estimates obtained in the numerical experiments agree with the theoretical error estimates from Theorems 5.1 and 5.2. The numerical method presented here is strongly based on the periodicity of the coefficients a_{ij} and on the construction of computable boundary correctors.

For this reason the method has relative low computational cost with optimal error convergence rates.

We generalize results in the literature for estimating the error between u_ϵ and its first and second order asymptotic expansion under weak assumptions on the regularity of the solutions of the cell problems and the homogenized equation. We also develop and analyze constructive boundary correctors. Such analysis permits us to develop finite element error estimates with very weak assumptions on the regularity of $a(y)$, including composite materials applications.

REFERENCES

[1] A. Abdulle and C. Schwab, *Heterogeneous multiscale FEM for diffusion problems on rough surfaces*, Multiscale Model. Simul., 3 (2005), pp. 195–220.

[2] R. A. Adams, *Sobolev Spaces*, Pure and Applied Mathematics 65, Academic Press, New York, London, 1975.

[3] G. Allaire and M. Amar, *Boundary layer tails in periodic homogenization*, ESAIM Control Optim. Calc. Var., 4 (1999), pp. 209–243.

[4] G. Allaire and R. Brizzi, *A multiscale finite element method for numerical homogenization*, Multiscale Model. Simul., 4 (2005), pp. 790–812.

[5] T. Arbogast, *An overview of subgrid upscaling for elliptic problems in mixed form*, in Current Trends in Scientific Computing (Xi'an, 2002), Contemp. Math. 329, Amer. Math. Soc., Providence, RI, 2003, pp. 21–32.

[6] T. Arbogast, *Analysis of a two-scale, locally conservative subgrid upscaling for elliptic problems*, SIAM J. Numer. Anal., 42 (2004), pp. 576–598.

[7] M. Avellaneda and F.-H. Lin, *Homogenization of elliptic problems with L^p boundary data*, Appl. Math. Optim., 15 (1987), pp. 93–107.

[8] I. Babuška, *Solution of interface problems by homogenization.* I, SIAM J. Math. Anal., 7 (1976), pp. 603–634.

[9] I. Babuška, *Solution of interface problems by homogenization.* II, SIAM J. Math. Anal., 7 (1976), pp. 635–645.

[10] I. Babuška, *Solution of interface problems by homogenization.* III, SIAM J. Math. Anal., 8 (1977), pp. 923–937.

[11] A. Bensoussan, J.-L. Lions, and G. Papanicolaou, *Asymptotic Analysis for Periodic Structures*, Studies in Mathematics and Its Applications 5, North–Holland, Amsterdam, 1978.

[12] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, Texts in Applied Mathematics 15, Springer-Verlag, New York, 1994.

[13] F. Brezzi, *Interacting with the subgrid world*, in Numerical Analysis 1999 (Dundee), Chapman & Hall/CRC Res. Notes Math. 420, Chapman & Hall/CRC, Boca Raton, FL, 2000, pp. 69–82.

[14] F. Brezzi, M. O. Bristeau, L. P. Franca, M. Mallet, and G. Rogé, *A relationship between stabilized finite element methods and the Galerkin method with bubble functions*, Comput. Methods Appl. Mech. Engrg., 96 (1992), pp. 117–129.

[15] Z. Chen and T. Y. Hou, *A mixed multiscale finite element method for elliptic problems with oscillating coefficients*, Math. Comp., 72 (2003), pp. 541–576.

[16] D. Cioranescu and P. Donato, *An Introduction to Homogenization*, Oxford Lecture Series in Mathematics and Its Applications 17, The Clarendon Press, Oxford University Press, New York, 1999.

[17] M. Crouzeix and V. Thomée, *The stability in L_p and W_p^1 of the L_2-projection onto finite element function spaces*, Math. Comp., 48 (1987), pp. 521–532.

[18] W. E and B. Engquist, *The heterogeneous multiscale methods*, Commun. Math. Sci., 1 (2003), pp. 87–132.

[19] W. E and B. Engquist, *The heterogeneous multi-scale method for homogenization problems*, in Multiscale Methods in Science and Engineering, Lect. Notes Comput. Sci. Eng. 44, Springer, Berlin, 2005, pp. 89–110.

[20] Y. Efendiev, T. Hou, and V. Ginting, *Multiscale finite element methods for nonlinear problems and their applications*, Commun. Math. Sci., 2 (2004), pp. 553–589.

[21] Y. Efendiev and A. Pankov, *Numerical homogenization and correctors for nonlinear elliptic equations*, SIAM J. Appl. Math., 65 (2004), pp. 43–68.

[22] Y. R. Efendiev, T. Y. Hou, and X.-H. Wu, *Convergence of a nonconforming multiscale finite*

element method, SIAM J. Numer. Anal., 37 (2000), pp. 888–910.
[23] Y. R. Efendiev and X.-H. Wu, *Multiscale finite element for problems with highly oscillatory coefficients*, Numer. Math., 90 (2002), pp. 459–486.
[24] V. Girault and P.-A. Raviart, *Finite Element Methods for Navier-Stokes Equations*, Springer Series in Computational Mathematics 5, Springer-Verlag, Berlin, 1986.
[25] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Monographs and Studies in Mathematics 24, Pitman (Advanced Publishing Program), Boston, 1985.
[26] T. Y. Hou and X.-H. Wu, *A multiscale finite element method for elliptic problems in composite materials and porous media*, J. Comput. Phys., 134 (1997), pp. 169–189.
[27] T. Y. Hou, X.-H. Wu, and Z. Cai, *Convergence of a multiscale finite element method for elliptic problems with rapidly oscillating coefficients*, Math. Comp., 68 (1999), pp. 913–943.
[28] V. V. Jikov, S. M. Kozlov, and O. A. Oleĭnik, *Homogenization of Differential Operators and Integral Functionals*, Springer-Verlag, Berlin, 1994.
[29] O. A. Ladyzhenskaya and N. N. Ural′tseva, *Linear and Quasilinear Elliptic Equations*, Academic Press, New York, 1968.
[30] E. M. Landis and G. P. Panasenko, *A theorem on the asymptotic behavior of the solutions of elliptic equations with coefficients that are periodic in all variables, except one*, Dokl. Akad. Nauk SSSR, 235 (1977), pp. 1253–1255.
[31] Y. Y. Li and M. Vogelius, *Gradient estimates for solutions to divergence form elliptic equations with discontinuous coefficients*, Arch. Ration. Mech. Anal., 153 (2000), pp. 91–151.
[32] J.-L. Lions, *Some methods in the mathematical analysis of systems and their control*, Kexue Chubanshe (Science Press), Beijing, 1981.
[33] J.-L. Lions and E. Magenes, *Non-homogeneous Boundary Value Problems and Applications. Vol. III*, Springer-Verlag, New York, 1973.
[34] N. G. Meyers, *An L^p e-estimate for the gradient of solutions of second order elliptic divergence equations*, Ann. Scuola Norm. Sup. Pisa (3), 17 (1963), pp. 189–206.
[35] S. Moskow and M. Vogelius, *First-order corrections to the homogenised eigenvalues of a periodic composite medium. A convergence proof*, Proc. Roy. Soc. Edinburgh Sect. A, 127 (1997), pp. 1263–1299.
[36] S. Moskow and M. Vogelius, *First order corrections to the eigenvalues of a periodic composite medium. the case of neumann boundary conditions*, Indiana Univ. Math. J., (2007), accepted.
[37] O. A. Oleĭnik, A. S. Shamaev, and G. A. Yosifian, *Mathematical Problems in Elasticity and Homogenization*, Studies in Mathematics and Its Applications 26, North–Holland, Amsterdam, 1992.
[38] P. Peisker, *On the numerical solution of the first biharmonic equation*, RAIRO Modél. Math. Anal. Numér., 22 (1988), pp. 655–676.
[39] G. Sangalli, *Capturing small scales in elliptic problems using a residual-free bubbles finite element method*, Multiscale Model. Simul., 1 (2003), pp. 485–503.
[40] C. Schwab and A.-M. Matache, *Generalized FEM for homogenization problems*, in Multiscale and Multiresolution Methods, Lect. Notes Comput. Sci. Eng. 20, Springer, Berlin, 2002, pp. 197–237.
[41] L. R. Scott and S. Zhang, *Finite element interpolation of nonsmooth functions satisfying boundary conditions*, Math. Comp., 54 (1990), pp. 483–493.
[42] G. Stampacchia, *Le problème de Dirichlet pour les équations elliptiques du second ordre à coefficients discontinus*, Ann. Inst. Fourier (Grenoble), 15 (1965), pp. 189–258.
[43] E. M. Stein, *Singular Integrals and Differentiability Properties of Functions*, Princeton Mathematical Series 30, Princeton University Press, Princeton, NJ, 1970.
[44] H. M. Versieux, *Numerical Boundary Corrector Methods and Analysis for a Second Order Elliptic PDE with Highly Oscillatory Periodic Coefficients*, Ph.D. thesis. IMPA-TR Serie C 47, Instituto de Matematica Pura e Aplicada, 2006.
[45] H. M. Versieux and M. Sarkis, *Numerical boundary corrector for elliptic equations with rapidly oscillating periodic coefficients*, Comm. Numer. Methods Engrg., 22 (2006), pp. 577–589.
[46] H. M. Versieux and M. Sarkis, *A three-scale finite element method for elliptic equations with rapidly oscillating periodic coefficients*, in Domain Decomposition Methods in Science and Engineering XVI, O. Widlund and D. Keyes, eds., Lect. Notes Comput. Sci. Eng. 55, Springer, Berlin, 2006, pp. 763–770.
[47] L. B. Wahlbin, *Local behavior in finite element methods*, in Handbook of Numerical Analysis, Vol. II, Handb. Numer. Anal., II, North–Holland, Amsterdam, 1991, pp. 353–522.
[48] M. F. Wheeler, *A Galerkin procedure for estimating the flux for two-point boundary value problems*, SIAM J. Numer. Anal., 11 (1974), pp. 764–768.

SIAM J. NUMER. ANAL.
Vol. 46, No. 2, pp. 577–613
© 2008 Society for Industrial and Applied Mathematics

RECOVERY ALGORITHMS FOR VECTOR-VALUED DATA WITH JOINT SPARSITY CONSTRAINTS*

MASSIMO FORNASIER† AND HOLGER RAUHUT‡

Abstract. Vector-valued data appearing in concrete applications often possess sparse expansions with respect to a preassigned frame for each vector component individually. Additionally, different components may also exhibit common sparsity patterns. Recently, there were introduced sparsity measures that take into account such joint sparsity patterns, promoting coupling of nonvanishing components. These measures are typically constructed as weighted ℓ_1 norms of componentwise ℓ_q norms of frame coefficients. We show how to compute solutions of linear inverse problems with such joint sparsity regularization constraints by fast thresholded Landweber algorithms. Next we discuss the adaptive choice of suitable weights appearing in the definition of sparsity measures. The weights are interpreted as indicators of the sparsity pattern and are iteratively updated after each new application of the thresholded Landweber algorithm. The resulting two-step algorithm is interpreted as a double-minimization scheme for a suitable target functional. We show its ℓ_2-norm convergence. An implementable version of the algorithm is also formulated, and its norm convergence is proven. Numerical experiments in color image restoration are presented.

Key words. linear inverse problems, joint sparsity, thresholded Landweber iterations, curvelets, subdifferential inclusion, color image reconstruction

AMS subject classifications. 65J22, 65K10, 65T60, 90C25, 52A41, 49M30, 68U10

DOI. 10.1137/0606668909

1. Introduction. Inverse problems.

We address the problem of recovering an element u of a Hilbert space $\mathcal{K}$ from the observed datum $g = Tu$ in the Hilbert space $\mathcal{H}$, where $T : \mathcal{K} \to \mathcal{H}$ is a bounded linear operator, possibly noninvertible or with an unbounded inverse. A simple approach to this problem is to minimize the discrepancy

$$\mathcal{T}(u) := \|Tu - g|\mathcal{H}\|^2.$$

If $\ker(T) = \{0\}$, then there exists a unique solution given by $u^* = (T^*T)^{-1}T^*g$. However, if T has an unbounded inverse, i.e., $(T^*T)^{-1}$ is unbounded, then this approach is very unstable.

Thus, if T is noninvertible or has an unbounded inverse (or an inverse with a high norm), one has to take into account further features of the expected solution. Indeed, a well-known way out is to consider the regularized problem [26]

$$u_\alpha^* := \operatorname{argmin}_{u\in\mathcal{K}} \mathcal{T}(u) + \alpha\|u|\mathcal{K}\|^2,$$

for which the corresponding solution operator $T_\alpha^\dagger : g \mapsto u_\alpha^*$ is bounded. Unfortunately, the minimal norm constraint is often not appropriate. A recent approach is

*Received by the editors August 4, 2006; accepted for publication (in revised form) July 18, 2007; published electronically February 1, 2008.
http://www.siam.org/journals/sinum/46-2/66680.html

†Johann Radon Institute for Computational and Applied Mathematics, Austrian Academy of Sciences, Altenbergerstrasse 69, A-4040 Linz, Austria (massimo.fornasier@oeaw.ac.at). This author's research was supported by the European Union's Human Potential Programme under contract MEIF-CT-2004-501018.

‡Numerical Harmonic Analysis Group, Faculty of Mathematics, University of Vienna, Nordbergstrasse 15, A-1090 Vienna, Austria (holger.rauhut@univie.ac.at). This author's research was supported by the European Union's Human Potential Programme under contracts HPRN–CT–2002–00285 (HASSIP) and MEIF-CT-2006-022811.

to substitute this particular constraint with a more general one

$$u_\Phi^* := \operatorname{argmin}_{u \in \mathcal{K}} \mathcal{T}(u) + \Phi(u),$$

where Φ is a suitable *sparsity measure*.

Sparse frame expansions. A *sparse representation* of an element of a Hilbert space is a series expansion with respect to orthonormal bases or a frame that has only a small number of large coefficients. Several types of signals appearing in nature admit sparse frame expansions, and thus, sparsity is a realistic assumption for a very large class of problems. For instance, images are well-represented by sparse expansions with respect to wavelets or curvelets, while for audio signals a Gabor frame is a good choice.

Sparsity has already had a long history of successes. The design of frames for sparse representations of digital signals has led to extremely efficient compression methods, such as JPEG2000 and MP3 [34]. Successively a new generation of optimal numerical schemes has been developed for the computation of sparse solutions of differential and integral equations, exploiting adaptive and greedy strategies [12, 39, 14, 15]. The use of sparsity in inverse problems for data recovery has been the most recent step of this long career of "simplifying and understanding complexity," with an enormous potential in applications [2, 17, 18, 20, 22, 21, 36, 40, 10, 13, 16]. Another field, which has gained much attention recently, is the observation that it is possible to reconstruct sparse signals from vastly incomplete information [6, 9, 23, 33, 37]. This line of research is called sparse recovery or compressed sensing.

From sparsity to joint sparsity. Most of the contributions appearing in the literature are addressed to the recovery of sparse scalar functions. Multichannel signals (i.e., vector-valued functions) appearing in concrete applications may not only possess sparse frame expansions for each channel individually, but additionally the different channels can also exhibit common sparsity patterns. Recently, new sparsity measures have been introduced that promote such coupling of the nonvanishing components through different channels [3, 30, 41]. These measures are typically constructed as *weighted* ℓ_1 norms of channel ℓ_q norms, with $q > 1$. We will use this concept for the solution of vector-valued inverse problems and combine it with another approach further promoting the coupling of sparsity patterns along channels. We note that sparse recovery of vector-valued signals was already investigated in [2], although joint sparsity was not considered in that work.

Our main results. We show how to compute solutions of linear inverse problems with joint sparsity regularization constraints by fast thresholded Landweber algorithms, similar to those presented in [17, 36, 40]. We discuss the adaptive choice of suitable *weights* appearing in the definition of the sparsity measures. The *weights* are interpreted as indicators of the sparsity pattern and are iteratively updated after each new application of the thresholded Landweber algorithm. The resulting two-step algorithm is interpreted as a double-minimization scheme for a suitable target functional. We prove that our algorithm converges to its minimizer. Since the functional is not smooth, this is done by subdifferential inclusions [38]. We prove that the thresholded Landweber algorithm, which constitutes the inner iteration of the double-minimization algorithm, converges linearly. The second step of the double-minimization has actually a simple explicit solution. Finally, we show that the full exact double-minimization scheme converges linearly, and we provide an implementable version which is also ensured to converge.

Morphological analysis of signals and sparsity patterns. The use of sparseness measures not only allows one to reconstruct a signal. At the same time it gives information about (joint) sparsity patterns which may encode morphological features of the signal. Well-known examples are the microlocal analysis properties of wavelets [31, 32] for singularity and regularity detection and the characterization of edges and curves by curvelets for natural images [8]. For instance, the weight sequences appearing in the sparsity measures we define, and interpret as indicators of the sparsity pattern, play a similar role as the discontinuity set is playing in the Mumford–Shah functional [35]. In fact, it is well known that wavelet or curvelet coefficients have high absolute values at high scales as soon as we are in the neighborhood of discontinuities. Even more illuminating and suggestive is the parallel between the sparsity measure and its indicator weights with the Ambrosio–Tortorelli [1] approximation of the Mumford–Shah functional. Here the discontinuity set is indicated by an auxiliary function, which is 1 where the image is smooth and 0 where edges and discontinuities are detected.

Joint sparsity patterns of vector-valued (i.e., multichannel) signals encode even finer properties of the morphology which do not belong only to one channel but are a common feature of all of the channels. Here the parallel is with generalizations of the Mumford–Shah functional as appearing, for example, in [5], where polyconvex functions of gradients couple discontinuity sets through different color channels of images.

Applications. We expect that our scheme can be applied in several different contexts. In this paper we limit ourselves to an application in color image reconstruction, modeling a real-world problem in art restoration. Indeed, color images have the advantage of being nontrivial multivariate and multichannel signals, exhibiting a very rich morphology and structure. In particular, discontinuities (jump sets) may appear in all of the channels at the same locations, which will be reflected in their curvelet representation (for instance). For these reasons, color images are a good model to test the effectiveness of our scheme promoting joint sparsity, also because the solution can be easily checked just by a visual analysis. Of course, the range of applicability of our approach is not limited to color image restoration. Neuroimaging (functional magnetic resonance imaging, magnetoencephalography), distributed compressed sensing [3], and several other problems with coupled vector-valued solutions are fields where we expect that our scheme can be fruitfully used. The numerical solution of differential and integral operator equations can also be addressed within this framework, and we refer, for example, to [14, 39, 15] for implementations by adaptive strategies.

Content of the paper. The paper is organized as follows. In section 2 we introduce the mathematical setting. We formulate our model of joint sparsity for multichannel signals and the corresponding functional to be minimized in order to solve a given linear inverse problem. The functional depends on two variables. The first belongs to the space of signals to be reconstructed; the second belongs to the space of sparsity indicator weights. Convexity properties of the functional are discussed. Section 3 is dedicated to the formulation of the double-minimization algorithm and to its weak convergence. The scheme is based on alternating minimizations in the first and in the second variable individually. In section 4 we discuss an efficient thresholded Landweber algorithm for the minimization with respect to the first variable. Its strong convergence is shown following the analysis in [17]. The minimization with respect to the second variable has an explicit solution, and no elaboration is needed. We provide an implementable version of the full scheme in section 5. To prove its convergence we

develop an error analysis. As a byproduct of the results in this section we show that the double-minimization scheme converges strongly. In section 6 we present an application in color image reconstruction. Numerical experiments are shown and discussed.

2. The functional.

2.1. Notation. Before starting our discussion let us briefly introduce some of the spaces we will use in the following. For some countable index set Λ we denote by $\ell_p = \ell_p(\Lambda)$, $1 \le p \le \infty$, the space of real sequences $u = (u_\lambda)_{\lambda\in\Lambda}$ with norm

$$\|u\|_p \;=\; \|u|\ell_p\| \;:=\; \left(\sum_{\lambda\in\Lambda} |u_\lambda|^p\right)^{1/p}, \quad 1 \le p < \infty,$$

and $\|u\|_\infty := \sup_{\lambda\in\Lambda} |u_\lambda|$ as usual. If (v_λ) is a sequence of positive weights, then we define the weighted spaces $\ell_{p,v} = \ell_{p,v}(\Lambda) = \{u, (u_\lambda v_\lambda) \in \ell_p(\Lambda)\}$ with norm

$$\|u\|_{p,v} \;=\; \|u|\ell_{p,v}\| \;=\; \|(u_\lambda v_\lambda)\|_p \;=\; \left(\sum_{\lambda\in\Lambda} v_\lambda^p |u_\lambda|^p)\right)^{1/p}$$

(with obvious modification for $p = \infty$). If the entries u_λ are actually vectors in a Banach space X with norm $\|\cdot\|_X$, then we denote

$$\ell_{p,v}(\Lambda, X) \;:=\; \{(u_\lambda)_{\lambda\in\Lambda}, u_\lambda \in X, (\|u_\lambda\|_X)_{\lambda\in\Lambda} \in \ell_{p,v}(\Lambda)\},$$

with norm $\|u|\ell_{p,v}(\Lambda, X)\| = \|(\|u_\lambda\|_X)_{\lambda\in\Lambda}|\ell_{p,v}(\Lambda)\|$. Usually X will be $\mathbb{R}^M$ endowed with the Euclidean norm or the M-dimensional space ℓ_q^M, i.e., $\mathbb{R}^M$ endowed with the ℓ_q norm. By $\mathbb{R}_+$ we denote the nonnegative real numbers.

2.2. Inverse problems with joint sparsity constraints. Let $\mathcal{K}$ and $\mathcal{H}_j$, $j = 1, \dots, N$, be (separable) Hilbert spaces and $A_{\ell,j} : \mathcal{K} \to \mathcal{H}_j$, $j = 1, \dots, M$, $\ell = 1, \dots, N$, some bounded linear operators. Assume we are given data $g_j \in \mathcal{H}_j$,

$$g_j \;=\; \sum_{\ell=1}^{M} A_{\ell,j} f_\ell, \quad j = 1, \dots, N.$$

Then our basic task consists in reconstructing the (unknown) elements $f_\ell \in \mathcal{K}$, $\ell = 1, \dots, M$.

In practice, it happens that the corresponding mapping from the vector (f_ℓ) to the vector (g_j) is not invertible or ill-conditioned. Moreover, the data g_j, $j = 1, \dots, N$, are often corrupted by noise. Thus, in order to deal with our reconstruction problem we have to regularize it.

Our basic assumption throughout this paper will be that the "channels" f_ℓ, $\ell = 1, \dots, M$, are correlated by means of joint sparsity patterns. Our aim is to model the joint sparsity within a regularization term. In the following we develop this idea.

For the sake of short notation we resume the data vector into

$$g \;=\; (g_j)_{j=1,\dots,N} \in \mathcal{H} := \bigoplus_{j=1}^{N} \mathcal{H}_j,$$

where the Hilbert space $\mathcal{H}$ is equipped with the usual inner product $\langle \sum_j g_j, \sum_j h_j \rangle := \sum_j \langle g_j, h_j \rangle$, with $g_j, h_j \in \mathcal{H}_j$. We also combine the operators $A_{\ell,j}$ into one operator

$$A : \bigoplus_{\ell=1}^{M} \mathcal{K} \to \mathcal{H}, \quad A(f_\ell)_{\ell=1}^M \;=\; \left(\sum_{\ell=1}^{M} A_{\ell,j} f_\ell\right)_{j=1}^{N}.$$

In order to exploit sparsity ideas we assume that we have given a suitable frame $\{\psi_\lambda : \lambda \in \Lambda\} \subset \mathcal{K}$ indexed by a countable set Λ. This means that there exist constants $C_1, C_2 > 0$ such that

$$C_1\|f\|_{\mathcal{K}}^2 \le \sum_{\lambda\in\Lambda} |\langle f, \psi_\lambda\rangle|^2 \le C_2\|f\|_{\mathcal{K}}^2 \qquad \text{for all } f \in \mathcal{K}. \tag{2.1}$$

Orthonormal bases are particular examples of frames. Frames allow for a (stable) series expansion of any $f \in \mathcal{K}$ of the form

$$f \;=\; Fu \;:=\; \sum_{\lambda\in\Lambda} u_\lambda \psi_\lambda, \tag{2.2}$$

where $u = (u_\lambda)_{\lambda\in\Lambda} \in \ell_2(\Lambda)$. The linear operator $F : \ell_2(\Lambda) \to \mathcal{K}$ is called the *synthesis map* in frame theory. It is bounded due to the frame inequality (2.1). In contrast to orthonormal bases, the coefficients u_λ need not be unique, in general. For more information on frames we refer to [11].

A main assumption here is that the f_ℓ to be reconstructed are sparse with respect to the frame $\{\psi_\lambda\}$. This means that f_ℓ can be well-approximated by a series of the form (2.2) with only a small number of nonvanishing coefficients u_λ. This can be modeled by assuming that the sequence u is contained in a (weighted) $\ell_1(\Lambda)$ space. Indeed, the minimization of the $\ell_1(\Lambda)$ norm promotes that only a few entries are nonzero. Taking, for instance, a wavelet frame and a suitable weight, the ℓ_1 constraint implies that the element to be reconstructed lies in a certain Besov space $B^s_{1,1}$; see [17].

Analogously as in [17] such considerations lead to the regularized functional

$$\begin{aligned}\mathcal{J}(u) &= \|g - Tu|\mathcal{H}\|^2 + \|u|\ell_{1,v}(\Lambda^M)\| \\ &= \sum_{j=1}^{N} \left\| g_j - \sum_{\ell=1}^{M} A_{\ell,j} F u^\ell \right\|_{\mathcal{H}_j}^2 + \sum_{\ell=1}^{M}\sum_{\lambda\in\Lambda} v_\lambda |u_\lambda^\ell|,\end{aligned} \tag{2.3}$$

which has to be minimized with respect to the vector of coefficients $u = (u_\lambda^\ell)_{\lambda\in\Lambda}^{\ell=1,\dots,M}$. The $\ell_{1,v}$ norm in this functional clearly represents the regularization term. The numbers v_λ, $\lambda \in \Lambda$, are some suitable positive weights. Once the minimizer $u = (u_\lambda^\ell)$ is determined, we obtain a reconstruction of the vectors of interest by means of $f_\ell = Fu^\ell = \sum_\lambda u_\lambda^\ell \psi_\lambda$. The algorithm in [17] can be taken to perform the minimization with respect to u.

The functional $\mathcal{J}(u)$ in the form stated, however, does not necessarily model any correlation between the vectors (channels) f_ℓ, $\ell = 1, \dots, M$. A way to incorporate such correlation is the assumption of joint sparsity; see also [30, 41]. By this we mean that the pattern of nonzero coefficients representing f_ℓ is (approximately) the same for all of the channels. In other words, for some *finite* set of indexes $\Lambda_0 \subset \Lambda$ and for all $\ell = 1, \dots, N$ there is an expansion

$$f_\ell \;\approx\; \sum_{\lambda\in\Lambda_0} u_\lambda^\ell \psi_\lambda.$$

In particular, the *same* Λ_0 can be chosen for all f_ℓ's.

We propose two approaches (that can be combined) to model joint sparsity. The first one assumes that the mixed norm

$$\|u|\ell_{1,v}(\Lambda,\ell_q^M)\| \,=\, \sum_{\lambda\in\Lambda} v_\lambda \|u_\lambda\|_q$$

of $u=(u_\lambda^\ell)$ is small. Here u_λ denotes the vector $(u_\lambda^{(\ell)})_{\ell=1}^M$ in $\mathbb{R}^M$. (Recall also that ℓ_q^M denotes $\mathbb{R}^M$ endowed with the ℓ_q norm). Here $q>1$ and, in particular, $q=2$ or $q=\infty$ represent the interesting cases, since for $q=1$ the above norm reduces to the usual weighted $\ell_{1,v}$ norm. In fact if q is large and some $|u_\lambda^\ell|$ is large, then the channel entries $|u_\lambda^{\ell'}|$ are also allowed to be large for $\ell'\neq\ell$, without increasing significantly the norm $\|u_\lambda|\ell_q^M\|$. The minimization of the above norm promotes that all entries of the "interchannel" vector u_λ may become significant, once at least one of the components $|u_\lambda^\ell|$ is large.

Introduce the operators $T_{\ell,j}=A_{\ell,j}F:\ell_2(\Lambda)\to\mathcal{H}_\ell$ and

$$T:\ell_2(\Lambda,\mathbb{R}^M)\to\mathcal{H},\quad Tu \,=\, \left(\sum_{\ell=1}^M T_{\ell,j}u^\ell\right)_{j=1}^N \,=\, \left(\sum_{\ell=1}^M A_{\ell,j}Fu^\ell\right)_{j=1}^N.$$

The above reasoning leads to the functional

$$\begin{aligned} K(u) \,=\, K_v^{(q)}(u) \,&:=\, \|Tu-g|\mathcal{H}\|^2+\|u|\ell_{1,v}(\Lambda,\ell_q^M)\| \\ &=\, \sum_{j=1}^N \left\|\sum_{\ell=1}^M T_{\ell,j}u^\ell-g_j\right\|_{\mathcal{H}_j}^2+\sum_{\lambda\in\Lambda} v_\lambda\|u_\lambda\|_q \end{aligned} \tag{2.4}$$

to be minimized with respect to u. In section 4 we will develop an iterative thresholding algorithm similar as in [17] to perform this minimization.

The second approach to support joint sparsity is to encode the joint sparsity information in some sort of indicator function. This can in fact be done by using the weight (v_λ) as a second minimization variable. To this end we add an additional term to the original functional (2.3), punishing small values of v_λ. We obtain the functional

$$J_0(u,v) \,:=\, J_{\theta,\rho,0}^{(1)}(u,v) \,:=\, \|Tu-g|\mathcal{H}\|^2+\sum_{\lambda\in\Lambda} v_\lambda\|u_\lambda\|_1+\sum_\lambda \theta_\lambda(\rho_\lambda-v_\lambda)^2$$

restricted to $v_\lambda\geq 0$. Here $(\theta_\lambda)_\lambda$ and $(\rho_\lambda)_\lambda$ are some suitable positive sequences.

Now the task is to minimize $J_0(u,v)$ jointly with respect to both u and v. (Again, once this minimizer is determined we obtain $f_\ell=Fu^\ell$). Analyzing $J_0(u,v)$ we realize that for the minimizer (u,v) we will have $v_\lambda=0$ (or close to 0) if $\|u_\lambda\|_1=\sum_{\ell=1}^M|u_\lambda^\ell|$ is large so that $v_\lambda\|u_\lambda\|_1$ becomes small. On the other hand, if $\|u_\lambda\|_1$ is small, then the term $\theta_\lambda(\rho_\lambda-v_\lambda)$ dominates and forces v_λ to be close to ρ_λ. Thus, v_λ serves indeed as an indicator of large values of $\|u_\lambda\|_1$. It has the effect that, if v_λ is chosen small due to one large u_λ^ℓ, then also the other coefficients $u_\lambda^{\ell'}$, $\ell'\neq\ell$, can be chosen large without making the functional considerably bigger.

Unfortunately, in contrast to the previous functionals, $J(u,v)$ as stated above is no longer jointly convex in (u,v) in general (although it is convex as a functional of u and of v alone). Thus, it cannot be ensured that a local minimum of the functional will be a global one, a property that is very crucial for an efficient minimization method.

To overcome this problem we may add an additional suitable quadratic term. Moreover, we can, of course, combine the second approach with the first one and use an ℓ_q norm instead of an ℓ_1 norm for the interchannel vectors u_λ. This leads to the most general form of the regularized functional considered in this paper:

$$J(u,v) = J^{(q)}_{\theta,\rho,\omega}(u,v) \tag{2.5}$$

$$:= \|Tu - g|\mathcal{H}\|^2 + \sum_{\lambda\in\Lambda} v_\lambda \|u_\lambda\|_q + \sum_{\lambda\in\Lambda} \omega_\lambda \|u_\lambda\|_2^2 + \sum_{\lambda\in\Lambda} \theta_\lambda(\rho_\lambda - v_\lambda)^2.$$

Here ω_λ is a suitably chosen sequence of positive numbers, and $1 \le q \le \infty$.

We will provide a sufficient condition depending on θ_λ and ρ_λ in the next subsection ensuring the strict joint convexity of $J(u,v)$ in (u,v). Although there is an extra term, $J(u,v)$ has similar properties as $J_0(u,v)$. In particular, v can still be seen as a sort of indicator function.

Observe that in the minimum we will always have $0 \le v_\lambda \le \rho_\lambda$. Therefore, we can assume the domain of J to be $\ell_2(\Lambda,\mathbb{R}^M) \times \ell_{\infty,\rho^{-1}}(\Lambda)_+$, where $\ell_{\infty,\rho^{-1}}(\Lambda)_+$ denotes the (convex) cone of all nonnegative sequences $(v_\lambda) \in \ell_{\infty,\rho^{-1}}(\Lambda)$.

Our main contribution consists in providing an algorithm for the minimization of $J(u,v)$. It consists in alternately minimizing with respect to u and with respect v. The minimization with respect to v can be done explicitly. For the minimization with respect to u we propose an efficient iterative algorithm.

We will mainly study the problem in the real-valued case. The complex-valued case can be treated with the same methods (in principle) by observing that $\mathbb{C}^M$ is isomorphic to $\mathbb{R}^{2M}$, so passing from M complex-valued channels to $2M$ real-valued channels. We note, however, that slight complications may arise from the fact that an ℓ_q norm on $\mathbb{C}^M$ is not isometric to an ℓ_q norm on $\mathbb{R}^{2M}$ if $q \ne 2$. (In particular, the thresholding operator on $\mathbb{C}^M$ for $q = \infty$ will have a different form than the one provided in the next section for the real-valued case.)

Remark. In [2] Anthoine considered a similar functional as in (2.3) for multichannel signals,

$$\mathcal{J}'_v(u) \;=\; \|Tu - g|\mathcal{H}\|^2 + \sum_{\ell=1}^{M} \sum_{\lambda\in\Lambda} v_\lambda^\ell |u_\lambda|^{p_\ell},$$

with $1 \le p_\ell \le 2$ for $\ell = 1,\dots,M$. Clearly, the generalization consists in the latter exponents p_ℓ. An algorithm for the minimization of $\mathcal{J}'$ was developed in [2]. However, Anthoine's functional does not promote joint sparsity like our functionals in (2.4) or (2.5), since the penalty term decouples with respect to the channels u^ℓ, $\ell = 1,\dots,M$.

2.3. Convexity of the functional J. At several places in the following it will be convenient to write

$$J(u,v) \;=\; \mathcal{T}(u) + \Phi^{(q)}(u,v), \tag{2.6}$$

where

$$\mathcal{T}(u) \;=\; \|Tu-g\|_{\mathcal{H}}^2 \;=\; \sum_{j=1}^{N}\left\|\sum_{\ell=1}^{M} T_{\ell,j}u^{\ell}-g_j\right\|_{\mathcal{H}_j}^2,$$

$$\begin{aligned}\Phi^{(q)}(u,v) \;&=\; \sum_{\lambda\in\Lambda} v_\lambda\|u_\lambda\|_q+\sum_{\lambda\in\Lambda}\omega_\lambda\|u_\lambda\|_2^2+\sum_{\lambda\in\Lambda}\theta_\lambda(\rho_\lambda-v_\lambda)^2\\ &=\; \|(v_\lambda\|u_\lambda\|_q)_{\lambda\in\Lambda}\|_1+\left\|u|\ell_{2,\omega^{1/2}}(\Lambda,\ell_2^M)\right\|^2+\|\rho-v|\ell_{2,\theta^{1/2}}(\Lambda)\|^2\end{aligned}$$

are the discrepancy with respect to the data and the joint sparsity measure, respectively.

Also it is useful to observe that $\Phi^{(q)}$ decouples with respect to λ, i.e.,

$$\Phi^{(q)}(u,v) \;=\; \sum_{\lambda\in\Lambda}\Phi_\lambda^{(q)}(u_\lambda,v_\lambda), \tag{2.7}$$

where

$$\begin{aligned}\Phi_\lambda^{(q)}(x,y) \;&=\; y\|x\|_q+\omega_\lambda\|x\|_2^2+\theta_\lambda(\rho_\lambda-y)^2\\ &=\; y\left(\sum_{\ell=1}^{M}|x_\ell|^q\right)^{1/q}+\omega_\lambda\sum_{\ell=1}^{M}x_\ell^2+\theta_\lambda(\rho_\lambda-y)^2,\qquad x\in\mathbb{R}^M,y\geq 0\end{aligned} \tag{2.8}$$

(with the usual modification for $q=\infty$).

In the following we give necessary and sufficient conditions for the (strict) convexity of the functional $\Phi^{(q)}$ for the most interesting cases $q=1,2,\infty$. These imply sufficient conditions for the (strict) convexity of $J=J_{\theta,\rho,\omega}^{(q)}$.

PROPOSITION 2.1. *Let $q\in\{1,2,\infty\}$. The sparsity measure $\Phi^{(q)}$ is convex if and only if $\omega_\lambda\theta_\lambda\geq\frac{\kappa_q}{4}$ for all $\lambda\in\Lambda$, where $\kappa_1=M$ and $\kappa_2=\kappa_\infty=1$. In particular, if $\omega_\lambda\theta_\lambda\geq\frac{\kappa_q}{4}$ for all $\lambda\in\Lambda$, then J is convex. In the case of a strict inequality $\omega_\lambda\theta_\lambda>\frac{\kappa_q}{4}$ we can replace "convexity" by "strict convexity" in all of these statements.*

Proof. It is easy to see that $\Phi^{(q)}$ is (strictly) convex if and only if all of the $\Phi_\lambda^{(q)}$, $\lambda\in\Lambda$, are (strictly) convex.

Let us first consider $q=1$. Observe that we can write $\Phi_\lambda^{(1)}(x,y)=\sum_{\ell=1}^{M}F_\lambda^{(1)}(x_\ell,y)$, with

$$\begin{aligned}F_\lambda^{(1)}(z,y) \;&=\; y|z|+\omega_\lambda|z|^2+M^{-1}\theta_\lambda(\rho_\lambda-y)^2\\ &=\; \left(y|z|+\omega_\lambda|z|^2+M^{-1}\theta_\lambda y^2\right)+M^{-1}\theta_\lambda\left(\rho_\lambda^2-2\rho_\lambda y\right),\quad z\in\mathbb{R},y\geq 0.\end{aligned} \tag{2.9}$$

The function in the second bracket is obviously linear and hence convex. The function in the first bracket can be written as the composition $G_\lambda\circ L$, with $L(z,y)=(|z|,y)$ and

$$G_\lambda(z,y) \;=\; yz+\omega_\lambda z^2+M^{-1}\theta_\lambda y^2 \;=\; \frac{1}{2}(z,y)\,H^{(1)}\,(z,y)^T,\qquad z,y\in\mathbb{R},$$

where

$$H^{(1)} \;=\; \begin{pmatrix}2\omega_\lambda & 1\\ 1 & 2\theta_\lambda M^{-1}\end{pmatrix}.$$

Since L is convex and has range $\mathbb{R}^2_+$, and G_λ is monotonically increasing in each coordinate on $\mathbb{R}^2_+$, it suffices to show that $G_\lambda(z,y)$ is convex if and only if $\theta_\lambda\omega_\lambda \geq M/4$; see, e.g., [4, p. 86]. The convexity of G_λ is equivalent to $H^{(1)}$ being positive semidefinite. The latter is clearly equivalent to $\theta_\lambda\omega_\lambda \geq M/4$. Strict convexity is equivalent to a strict inequality $\theta_\lambda\omega_\lambda > M/4$.

Now let $q = 2$. Observe that $\Phi^{(2)}_\lambda = F^{(2)}_\lambda \circ L^{(2)}$, where $L^{(2)} : \mathbb{R}^M \times \mathbb{R}_+ \to \mathbb{R}^2_+$, $L^{(2)}(x,y) = (\|x\|_2, y)$, and

$$F^{(2)}_\lambda(z,y) = yz + \omega_\lambda z^2 + \theta_\lambda(\rho_\lambda - y)^2 \tag{2.10}$$
$$= \frac{1}{2}(z,y)\, H^{(2)}\, (z,y)^T + \theta_\lambda(\rho_\lambda^2 - 2\rho_\lambda y), \quad z,y \in \mathbb{R},$$

with

$$H^{(2)} = \begin{pmatrix} 2\omega_\lambda & 1 \\ 1 & 2\theta_\lambda \end{pmatrix}.$$

By a similar argument as above, $\Phi^{(2)}_\lambda$ is convex if and only if $H^{(2)}$ is positive semidefinite. The latter is the case if and only if $\omega_\lambda\theta_\lambda \geq 1/4$, and a strict convexity is equivalent to a strict inequality.

Finally, let $q = \infty$. Observe that

$$\Phi^{(\infty)}_\lambda(x,y) = \max_{\ell=1,\dots,M} \left\{ y|x_\ell| + \omega_\lambda \sum_{m=1}^{M} |x_m|^2 + \theta_\lambda(\rho_\lambda - y)^2 \right\}.$$

Since $\Phi^{(\infty)}_\lambda$ is the pointwise maximum of M functions, it is sufficient (see [4, p. 80]) to investigate the (strict) convexity of each of the functions

$$f_{\lambda,\ell}(x,y) = yx_\ell + \omega_\lambda \sum_{m=1}^{M} (x_m)^2 + \theta_\lambda(\rho_\lambda - y)^2$$
$$= \frac{1}{2}(y,x)\, H^{(\infty)}_\ell\, (y,x)^T + \theta_\lambda(\rho_\lambda^2 - 2y\rho_\lambda), \quad x \in \mathbb{R}^M, y \geq 0,$$

with

$$H^{(\infty)}_\ell = \begin{pmatrix} 2\omega_\lambda & 0 & \cdots & 0 & \delta_{1,\ell} \\ 0 & 2\omega_\lambda & \cdots & 0 & \delta_{2,\ell} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 2\omega_\lambda & \delta_{M,\ell} \\ \delta_{1,\ell} & \delta_{2,\ell} & \cdots & \delta_{M,\ell} & 2\theta_\lambda \end{pmatrix}.$$

One can show by induction that

$$\det(H^{(\infty)}_\ell) = 2^{M-1}\omega_\lambda^{M-1}(4\theta_\lambda\omega_\lambda - 1).$$

Thus, $H^{(\infty)}_\ell$ is positive semidefinite if and only if $\theta_\lambda\omega_\lambda \geq 1/4$, and the convexity of $\Phi^{(q)}_\lambda$ is equivalent to the latter condition. Once again the strict convexity is equivalent to the strict inequality. □

We do not pursue the task to obtain conditions for the convexity of $\Phi^{(q)}$ and J for general $q \neq 1, 2, \infty$ but rather assume that $\Phi^{(q)}$ and hence J are always convex also in this case.

3. The minimizing algorithm and its convergence. In this section we propose and analyze an algorithm for the computation of the minimizer (u^*, v^*) of the functional $J(u,v) = J^{(q)}_{\theta,\rho,\omega}(u,v)$ defined in (2.5). The algorithm consists in alternating a minimization with respect to u and a minimization with respect to v. More formally, for an arbitrary choice $v^{(0)}$, for example, $v^{(0)} = (\rho_\lambda)_{\lambda\in\Lambda}$, we define

$$\begin{aligned} u^{(n)} &:= \arg\min_{u\in\ell_2(\Lambda,\mathbb{R}^M)} J(u, v^{(n-1)}), \\ v^{(n)} &:= \arg\min_{v\in\ell_{\infty,\rho^{-1}}(\Lambda)_+} J(u^{(n)}, v). \end{aligned} \tag{3.1}$$

The minimization of $J(u, v^{(n-1)})$ with respect to u can be done by means of the iterative thresholding algorithm that we will study in the next section. The minimizer $v^{(n)}$ of $J(u^{(n)}, v)$ for fixed $u^{(n)}$ can be computed explicitly. Indeed, it follows from elementary calculus that

$$v^{(n)}_\lambda = \begin{cases} \rho_\lambda - \frac{1}{2\theta_\lambda}\|u^{(n)}{}_\lambda\|_q & \text{if } \|u^{(n)}{}_\lambda\|_q < 2\theta_\lambda\rho_\lambda, \\ 0 & \text{otherwise.} \end{cases} \tag{3.2}$$

We have the following result about the convergence of the above algorithm.

THEOREM 3.1. *Let $1 \le q \le \infty$, and assume that $\Phi^{(q)}$ and hence J are strictly convex (see also Proposition 2.1). Moreover, we assume that $\ell_{2,\omega^{1/2}}(\Lambda,\mathbb{R}^M)$ is embedded into $\ell_2(\Lambda,\mathbb{R}^M)$, i.e., $\omega_\lambda \ge \gamma > 0$ for all $\lambda\in\Lambda$. Then the sequence $(u^{(n)}, v^{(n)})_{n\in\mathbb{N}}$ converges to the unique minimizer $(u^*, v^*) \in \ell_2(\Lambda,\mathbb{R}^M) \times \ell_{\infty,\rho^{-1}}(\Lambda)_+$ of J. The convergence of $u^{(n)}$ is weak in $\ell_2(\Lambda,\mathbb{R}^M)$ and that of $v^{(n)}$ holds componentwise.*

For the most interesting cases $q \in \{1, 2, \infty\}$, if in addition $\theta_\lambda\omega_\lambda \ge \sigma > \kappa_q/4$ for all $\lambda\in\Lambda$, where $\kappa_1 = M$ and $\kappa_2 = \kappa_\infty = 1$, then the convergence of $u^{(n)}$ to u^* is also strong in $\ell_2(\Lambda,\mathbb{R}^M)$ and $v^{(n)} - v^*$ converges to 0 strongly in $\ell_{2,\theta}(\Lambda)$.

The rest of the section will be spent with the proof of the weak convergence of the algorithm. The strong convergence and the full proof of the Theorem 3.1 will be established only in section 5.3 later.

3.1. Subdifferential calculus. A main tool in the analysis of nonsmooth functionals and their minima is the concept of subdifferential. Recall that, for a convex functional F on some Banach space V, its subdifferential $\partial F(x)$ at a point $x \in V$, with $F(x) < \infty$, is defined as the set

$$\partial F(x) = \{x^* \in V^*, x^*(z-x) + F(x) \le F(z) \text{ for all } z \in V\},$$

where V^* denotes the dual space of V. It is obvious from this definition that $0 \in \partial F(x)$ if and only if x is a minimizer of F. In the following, we investigate the subdifferential of J. In order to have J defined on the whole Banach space $\ell_2(\Lambda,\mathbb{R}^M) \times \ell_{\infty,\rho^{-1}}(\Lambda)$ rather than just for positive v_λ's (which is needed to use subdifferentials), we simply extend $J(u,v)$ by

$$J(u,v) = \infty \qquad \text{if } v_\lambda < 0 \text{ for some } \lambda\in\Lambda$$

as usual. This extension preserves convexity and does not change the minimizer.

Recall that J can be written as $J(u,v) = \mathcal{T}(u) + \Phi^{(q)}(u,v)$; see (2.6). Since both $\mathcal{T}$ and $\Phi^{(q)}$ are convex we have (see, e.g., [24, Proposition 5.6])

$$\partial J(u,v) = \partial\mathcal{T}(u) \times \{0\} + \partial\Phi^{(q)}(u,v). \tag{3.3}$$

Concerning the subdifferential of $\mathcal{T}$ we have the following result.

LEMMA 3.2. *The subdifferential of* $\mathcal{T}$ *at* $u \in \ell_2(\Lambda, \mathbb{R}^M)$ *consists of one element:*

$$\partial\mathcal{T}(u) = \{2T^*(Tu - g)\}.$$

Proof. Since $\mathcal{T}$ is convex and Gateaux-differentiable, by Proposition 5.3 [24] we have $\partial\mathcal{T}(u) = \{\mathcal{T}'(u)\}$, where its Gateaux derivative is characterized by $\langle \mathcal{T}'(u), z\rangle = \lim_{h\to 0^+} \frac{\mathcal{T}(u+hz)-\mathcal{T}(u)}{h}$ for all $z \in \ell_2(\Lambda, \mathbb{R}^M)$. It is straightforward to check that the Gateaux derivative of a functional of the type $u \to \|Tu - g\|^2$ (with linear T) at u applied on z is given by $2\langle Tu-g, Tz\rangle = 2\langle T^*(Tu-g), z\rangle$. This proves the claim. ∎

Let us now consider the subdifferential of $\partial\Phi^{(q)}(u, v)$. Recall its domain $\ell_2(\Lambda, \mathbb{R}^M) \times \ell_{\infty,\rho^{-1}}(\Lambda)$. Since the dual of $\ell_{\infty,\rho^{-1}}$ is a bit inconvenient to handle, we restrict the subdifferential to the predual $\ell_{1,\rho}$. This will be enough for our purposes. Moreover, recall that $\Phi^{(q)}$ decouples into a sum of functionals $\Phi_\lambda^{(q)}$ depending on only (u_λ, v_λ); see (2.7). It is straightforward to show the following lemma.

LEMMA 3.3. *The subdifferential of* $\Phi^{(q)}$ *at a point* $(u, v) \in \ell_2(\Lambda, \mathbb{R}^M) \times \ell_{\infty,\rho^{-1}}(\Lambda)$, *with* $\Phi^{(q)}(u, v) < \infty$, *satisfies*

$$\begin{aligned} D\Phi^{(q)}(u, v) &:= \partial\Phi^{(q)}(u, v) \cap (\ell_2(\Lambda, \mathbb{R}^M) \times \ell_{1,\rho}(\Lambda)) \\ &= \{(\xi, \eta) \in \ell_2(\Lambda, \mathbb{R}^M) \times \ell_{1,\rho}(\Lambda) : \ (\xi_\lambda, \eta_\lambda) \in \partial\Phi_\lambda^{(q)}(u_\lambda, v_\lambda) \text{ for all } \lambda \in \Lambda\}. \end{aligned}$$

We are left with investigating the subdifferential of the functional $\Phi_\lambda^{(q)}$ defined in (2.8). Similarly as J we extend it to $\mathbb{R}^M \times \mathbb{R}$ by $\Phi_\lambda^{(q)}(x, y) = \infty$ for $y < 0$.

LEMMA 3.4. *Let* $1 \le q \le \infty$. *Assume that* $\Phi_\lambda^{(q)}$ *is convex (see also Proposition* 2.1*). Then for* $(x, y) \in \mathbb{R}^M \times \mathbb{R}_+$ *we have*

$$\begin{aligned} \partial\Phi_\lambda^{(q)}(x, y) = \{(\xi, \eta) \in \mathbb{R}^M \times \mathbb{R} : \ &\xi \in y\partial\|\cdot\|_q(x) + 2\omega_\lambda x, \\ &\eta \in \|x\|_q \partial s^+(y) + 2\theta_\lambda(y - \rho_\lambda)\}, \end{aligned} \tag{3.4}$$

where $s^+(y) := y$ *for* $y \ge 0$ *and* $s^+(y) = \infty$ *for* $y < 0$. *In particular,* $\partial s^+(y) = \{1\}$ *for* $y > 0$ *and* $\partial s^+(0) = (-\infty, 1]$.

Remark. We recall that the subdifferential of the q norm on $\mathbb{R}^M$ is given as follows. If $1 < q < \infty$, then

$$\partial\|\cdot\|_q(x) = \begin{cases} B^{q'}(1) & \text{if } x = 0, \\ \left\{ \left(\frac{|x_\ell|^{q-1}\operatorname{sign}(x_\ell)}{\|x\|_q^{q-1}} \right)_{\ell=1}^M \right\} & \text{otherwise,} \end{cases}$$

where $B^{q'}(1)$ denotes the ball of radius 1 in the dual norm, i.e., in $\ell_{q'}^M$ with $1/q + 1/q' = 1$.

If $q = 1$, then

$$\partial\|\cdot\|_1(x) = \{\xi \in \mathbb{R}^M : \ \xi_\ell \in \partial|\cdot|(x_\ell), \ell = 1, \dots, M\}, \tag{3.5}$$

where $\partial|\cdot|(z) = \{\operatorname{sign}(z)\}$ if $z \neq 0$ and $\partial|\cdot|(0) = [-1, 1]$.

If $q = \infty$, then

$$\partial\|\cdot\|_\infty(x) = \begin{cases} B^1(1) & \text{if } x = 0, \\ \operatorname{conv}\{(\operatorname{sign}(x_\ell))e_\ell : |x_\ell| = \|x\|_\infty\} & \text{otherwise,} \end{cases} \tag{3.6}$$

where conv A denotes the convex hull of a set A and e_ℓ the ℓth canonical unit vector in $\mathbb{R}^M$.

Proof. Recall that

$$\Phi_\lambda^{(q)}(x,y) \;=\; s^+(y)\|x\|_q + \omega_\lambda \|x\|_2^2 + \theta_\lambda(\rho_\lambda - y)^2.$$

Let $y \geq 0$ so that $\Phi_\lambda^{(q)}(x,y)$ is finite. The subdifferential $\partial(\Phi_\lambda^{(q)})_x(x,y)$ of $\Phi^{(q)}(x,y)$ considered as a function of x alone (i.e., for fixed y) is clearly given by

$$\partial(\Phi_\lambda^{(q)})_x(x,y) \;=\; y\partial\|\cdot\|_q(x) + 2\omega_\lambda x, \tag{3.7}$$

while keeping x fixed gives

$$\partial(\Phi_\lambda^{(q)})_y(x,y) \;=\; \partial s^+(y)\|x\|_q + 2\theta_\lambda(y-\rho_\lambda).$$

This shows the inclusion "$\subset$" in (3.4). Moreover, for all of the points $(x,y) \in \mathbb{R}^M \times \mathbb{R}_+$ where $\Phi_\lambda^{(q)}$ is differentiable we even have equality in (3.4) since $\Phi_\lambda^{(q)}$ is convex, and, thus, all of the subdifferentials appearing consist of precisely one point, i.e., the usual gradient.

Let $1 < q < \infty$. Then for $x \neq 0$, $y > 0$ the differentiability assumption is clearly satisfied. For the other cases $x = 0$ or $y = 0$ we note that by convexity of $\Phi_\lambda^{(q)}$ we have (see [38, Corollary 10.11])

$$\partial(\Phi_\lambda^{(q)})_x(x,y) \;=\; \{\xi : \exists \eta \text{ such that } (\xi,\eta) \in \partial\Phi_\lambda^{(q)}(x,y)\} \tag{3.8}$$

and the corresponding relation for $\partial(\Phi_\lambda^{(q)})_y(x,y)$. Now, if $y > 0$, then $\Phi_\lambda^{(q)}(x,y)$ is differentiable with respect to y, and thus, η in the right-hand side of (3.8) is unique; indeed $\eta = \eta_0 := \frac{\partial}{\partial y}\Phi_\lambda^{(q)}(x,y)$. We conclude that for $y > 0$

$$\partial(\Phi_\lambda^{(q)})(x,y) = \{(\xi,\eta_0), \xi \in \partial(\Phi_\lambda^{(q)})_x(x,y)\}.$$

In particular this holds for $x = 0$, even for general $1 \leq q \leq \infty$. The same argument applies for the case $y = 0$ and $x \neq 0$ (and $1 < q < \infty$), which shows (3.4) in these cases. Now let $x = 0$ and $y = 0$. Then the right-hand side of (3.4) contains precisely one point, i.e., $(\xi,\eta) = (0, -2\theta_\lambda\rho_\lambda)$. Since the subdifferential $\Phi_\lambda^{(q)}(0,0)$ contains at least one point by convexity, it must coincide with (ξ,η) by the trivial inclusion $\subset$. (It is easy to check also directly that $(0,-2\theta_\lambda\rho_\lambda) \in \Phi_\lambda^{(q)}(0,0)$.) Note that this argument applies also for $q = 1, \infty$.

It remains to treat the cases $q = 1, \infty$ with $x \neq 0$ and arbitrary $y \geq 0$. Let us start with $q = 1$. In the proof of Proposition 2.1 it was noted that

$$\Phi_\lambda^{(1)}(x,y) \;=\; \sum_{\ell=1}^M F_\lambda^{(1)}(x_\ell, y),$$

with $F_\lambda^{(1)} : \mathbb{R}^2 \to \mathbb{R}$ defined in (2.9). The subdifferential of F_λ can be obtained in the same way as above (expressing, e.g., formally the modulus as a 2-norm on $\mathbb{R}^1$). For $(z,y) \in \mathbb{R} \times \mathbb{R}_+$ this yields

$$\partial F_\lambda^{(1)}(z,y) \;=\; \{(\tau,\eta) :\; \tau \in y\partial|\cdot|(z) + 2\omega_\lambda z, \eta \in |z|\partial s^+(y) + 2M^{-1}\theta_\lambda(y-\rho_\lambda)\}.$$

By convexity we have

$$\partial\Phi_\lambda^{(1)}(x,y) \;=\; \sum_{\ell=1}^{M} \Big\{ (e_\ell z_\ell, \eta) :\; (z_\ell,\eta) \in \partial F_\lambda^{(1)}(x_\ell, y) \Big\},$$

where e_ℓ denotes the ℓth unit vector in $\mathbb{R}^M$. By the explicit form of the subdifferential of the ℓ_1 norm (3.5), this gives (3.4) for $q = 1$.

Finally, let $q = \infty$. Similarly as in the proof of Proposition 2.1 we write

$$\Phi_\lambda^{(\infty)}(x,y) \;=\; \max_{\ell=1,\dots,M} F_\ell(x,y),$$

with

$$F_\ell(x,y) \;=\; y|x_\ell| + \omega_\lambda \|x\|_2^2 + \theta_\lambda(\rho_\lambda - y)^2.$$

If $x_\ell \neq 0$, then $F_\ell(x,y)$ is differentiable with respect to x and

$$\partial F_\ell(x,y) \;=\; \{(\xi,\eta) : \xi \;=\; y\,\mathrm{sign}(x_\ell) e_\ell + 2\omega_\lambda x,\; \eta \in \partial s^+(y)|x_\ell| + 2\theta_\lambda(y-\rho_\lambda)\},$$

where e_ℓ denotes the ℓth canonical unit vector in $\mathbb{R}^M$. This even holds for $y = 0$ by an analogous argument as above; see (3.8). The subdifferential of $\Phi_\lambda^{(\infty)}(x,y)$ for $x \neq 0$ is then given by (see, e.g., [38, Exercise 8.31])

$$\partial\Phi_\lambda^{(\infty)}(x,y) \;=\; \mathrm{conv}\{\partial F_\ell(x,y) : F_\ell(x,y) \;=\; \max_{m=1,\dots,M} F_m(x,y)\}.$$

Since $x \neq 0$, we have $x_\ell \neq 0$ if $|x_\ell| = \|x\|_\infty$, and the latter is the case if and only if $F_\ell(x,y) = \max_m F_m(x,y)$. Thus, we obtain

$$\begin{aligned}
&\partial\Phi_\lambda^{(\infty)}(x,y) \\
&\quad= \mathrm{conv} \bigcup_{\ell:|x_\ell|=\|x\|_\infty} \{(\xi,\eta) : \xi = y\,\mathrm{sign}(x_\ell)e_\ell + 2\omega_\lambda x, \eta \in \partial s^+(y)|x_\ell| + 2\theta_\lambda(y-\rho_\lambda)\} \\
&\quad= \{(\xi,\eta) : \xi \in \mathrm{conv}\{y\,\mathrm{sign}(x_\ell)e_\ell, |x_\ell| = \|x\|_\infty\} + 2\omega_\lambda x, \\
&\qquad \eta \in \|x\|_\infty \partial s^+(y) + 2\theta_\lambda(y-\rho_\lambda)\big\}.
\end{aligned}$$

By the characterization of the subdifferential of the ∞ norm in (3.6), we obtain the claimed equality in (3.4) for $q = \infty$ and $x \neq 0$. This finishes the proof. ∎

Combining the previous lemmas we obtain the following result.

PROPOSITION 3.5. *Let $1 \le q \le \infty$. Assume that $\Phi^{(q)}$ is convex, and let $(u,v) \in \ell_2(\Lambda,\mathbb{R}^M) \times \ell_{\infty,\rho^{-1}}(\Lambda)$ such that $\Phi^{(q)}(u,v) < \infty$. Then we have*

$$\begin{aligned}
D\Phi^{(q)}(u,v) = \{&(\xi,\eta) \in \ell_2(\Lambda,\mathbb{R}^M) \times \ell_{1,\rho}(\Lambda), \\
&\xi_\lambda \in v_\lambda \partial \|\cdot\|_q(u_\lambda) + 2\omega_\lambda u_\lambda, \\
&\eta_\lambda \in \|u_\lambda\|_q \partial s^+(v_\lambda) + 2\theta_\lambda(v_\lambda - \rho_\lambda), \lambda \in \Lambda\} \\
&\subset \partial\Phi^{(q)}(u,v)
\end{aligned} \tag{3.9}$$

and

$$\begin{aligned}
DJ(u,v) &= \partial J(u,v) \cap \big(\ell_2(\Lambda,\mathbb{R}^M) \times \ell_{1,\rho}(\Lambda)\big) \\
&= (2T^*(Tu-g),0) + D\Phi_\lambda^{(q)}(u,v) \;\subset\; \partial J(u,v).
\end{aligned}$$

3.2. Weak convergence of the double minimization. Before we actually start proving the weak convergence of the algorithm in (3.1) we recall the following definition [38].

DEFINITION 3.6. *Let V be a topological space and $\mathcal{A} = (A_n)_{n\in\mathbb{N}}$ a sequence of subsets of V. The subset $A \subseteq V$ is called the* limit of the sequence $\mathcal{A}$*, and we write $A = \lim_n A_n$, if*

$$A = \{a \in V : \exists a_n \in A_n, a = \lim_n a_n\}.$$

The following observation will be useful for us; see, e.g., [38, Proposition 8.7].

LEMMA 3.7. *Assume that Γ is a convex function on $\mathbb{R}^M$ and $(x_n) \subset \mathbb{R}^M$ a convergent sequence with limit x such that $\Gamma(x_n), \Gamma(x) < \infty$. Then the subdifferentials satisfy*

$$\lim_n \partial\Gamma(x_n) \subseteq \partial\Gamma(x).$$

In other words, the subdifferential $\partial\Gamma$ of a convex function is an outer semicontinuous set-valued function.

In the following we agree on the convention that the upper index n at $u^{(n)} \in \ell_2(\Lambda, \mathbb{R}^M)$ always denotes the nth iterate, and $u_\lambda^{(n)} \in \mathbb{R}^M$ denotes the (vector-valued) entry at λ of the nth iterate. In the following proof we will never refer to the ℓth component of the M-dimensional vector $u_\lambda^{(n)}$, so hopefully no confusion can arise. Also, we denote by $(DJ(u,v))_\lambda$ the restriction of $DJ(u,v) \subset \ell_2(\Lambda, \mathbb{R}^M) \times \ell_{1,\rho}(\Lambda)$ to the index λ. By the previous section it holds that

$$(DJ(u,v))_\lambda = (2T^*T(u-g))_\lambda, 0) + \partial\Phi_\lambda^{(q)}(u_\lambda, v_\lambda). \tag{3.10}$$

Now the proof is developed as follows. First, we recall that $(u^*, v^*) = \arg\min J(u, v)$ if and only if $0 \in \partial J(u^*, v^*)$. Next, we show that there exist weakly convergent subsequences of $(u^{(n)}, v^{(n)})$ (again denoted by $(u^{(n)}, v^{(n)})$) which converge to $(u^{(\infty)}, v^{(\infty)})$ and that

$$0 \in \lim_n DJ(u^{(n)}, v^{(n)}) \subseteq \partial J(u^{(\infty)}, v^{(\infty)}). \tag{3.11}$$

Due to the strict convexity of J we conclude that $(u^{(\infty)}, v^{(\infty)}) = (u^*, v^*)$. Now let us detail the argument.

By definition of $u^{(n)}$ and $v^{(n)}$ we have

$$\begin{aligned} &J(u^{(n)}, v^{(n)}) - J(u^{(n+1)}, v^{(n+1)}) \\ &= J(u^{(n)}, v^{(n)}) - J(u^{(n+1)}, v^{(n)}) + J(u^{(n+1)}, v^{(n)}) - J(u^{(n+1)}, v^{(n+1)}) \geq 0. \end{aligned}$$

Thus, $(J(u^{(n)}, v^{(n)}))_n$ is a nonincreasing sequence, and since $J \geq 0$ this implies that $(J(u^{(n)}, v^{(n)}))_n$ converges. Furthermore, we have

$$J(u^{(n)}, v^{(n)}) - J(u^{(n+1)}, v^{(n+1)}) \geq J(u^{(n+1)}, v^{(n)}) - J(u^{(n+1)}, v^{(n+1)}) \geq 0.$$

Observe that

$$\begin{aligned} &J(u^{(n+1)}, v^{(n)}) - J(u^{(n+1)}, v^{(n+1)}) \\ &= \sum_{\lambda\in\Lambda} \left(v_\lambda^{(n)} \|u_\lambda^{(n+1)}\|_q + \theta_\lambda (v_\lambda^{(n)} - \rho_\lambda)^2 - v_\lambda^{(n+1)} \|u_\lambda^{(n+1)}\|_q - \theta_\lambda (\rho_\lambda - v_\lambda^{(n+1)})^2 \right). \end{aligned}$$

A straightforward computation shows that the terms in the latter sum equal

$$\left(\|u_\lambda^{(n+1)}\|_q - 2\theta_\lambda(\rho_\lambda - v_\lambda^{(n+1)})\right)(v_\lambda^{(n)} - v_\lambda^{(n+1)}) + \theta_\lambda(v_\lambda^{(n)} - v_\lambda^{(n+1)})^2.$$

If $\|u_\lambda^{(n+1)}\|_q < 2\theta_\lambda\rho_\lambda$, then by (3.2) and a simple computation the first term vanishes. If $\|u_\lambda^{(n+1)}\|_q \geq 2\theta_\lambda\rho_\lambda$, then $v_\lambda^{(n+1)} = 0$ and the first term is nonnegative. Altogether, this shows that

$$J(u^{(n)}, v^{(n)}) - J(u^{(n+1)}, v^{(n+1)}) \geq \sum_{\lambda\in\Lambda} \theta_\lambda(v_\lambda^{(n)} - v_\lambda^{(n+1)})^2 \geq \theta_\lambda(v_\lambda^{(n)} - v_\lambda^{(n+1)})^2.$$

As noted above, the term on the left-hand side converges to 0, and since $\theta_\lambda > 0$ we deduce that

$$\lim_{n\to\infty} |v_\lambda^{(n)} - v_\lambda^{(n+1)}| = 0 \quad \text{for all } \lambda \in \Lambda. \tag{3.12}$$

Moreover, we have

$$J(u^{(0)}, v^{(0)}) \geq J(u^{(n)}, v^{(n)}) \geq \sum_{\lambda\in\Lambda} \omega_\lambda \|u_\lambda^{(n)}\|_2^2.$$

Therefore, $(u^{(n)})_n$ is uniformly bounded in $\ell_{2,\omega^{1/2}}(\Lambda, \mathbb{R}^M)$, and thus, there exists a subsequence $(u^{(n_k)})_k$ that converges to $u^{(\infty)} \in \ell_{2,\omega^{1/2}}(\Lambda, \mathbb{R}^M)$ weakly in both $\ell_{2,\omega^{1/2}}(\Lambda, \mathbb{R}^M)$ and $\ell_2(\Lambda, \mathbb{R}^M)$, due to our assumption $\omega_\lambda \geq \gamma > 0$ for all $\lambda \in \Lambda$. For simplicity, let us denote again $u^{(n_k)} = u^{(n)}$.

First of all, observe that weak convergence implies componentwise convergence, so that $u_\lambda^{(n)} \to u_\lambda^{(\infty)}$ and $[T^*Tu^{(n)}]_\lambda \to [T^*Tu^{(\infty)}]_\lambda$ for all $\lambda \in \Lambda$. By the explicit formula (3.2) for $v_\lambda^{(n)}$, this implies that $v^{(n)}$ converges pointwise to the limit

$$v_\lambda^{(\infty)} := \lim_n v_\lambda^{(n)} = \begin{cases} \rho_\lambda - \frac{1}{2\theta_\lambda}\|u^{(\infty)}{}_\lambda\|_q & \text{if } \|u^{(\infty)}{}_\lambda\|_q < 2\theta_\lambda\rho_\lambda, \\ 0 & \text{otherwise.} \end{cases} \tag{3.13}$$

By the definition of $u^{(n)}$ in (3.1), we have $0 \in \partial J_u(u, v^{(n)})$ (where $\partial J_u(u, v)$ denotes the subdifferential of J considered as a functional of u only). This means that

$$0 \in \left[2T^*(Tu^{(n)} - g)\right]_\lambda + v_\lambda^{(n-1)}\partial\|\cdot\|_q(u_\lambda^{(n)}) + 2\omega_\lambda u_\lambda^{(n)} \quad \text{for all } \lambda \in \Lambda$$

(see also Lemma 3.2 and (3.7)); in other words

$$0 = \left[2T^*(Tu^{(n)} - g)\right]_\lambda + v_\lambda^{(n-1)}\zeta_\lambda^{(n)} + 2\omega_\lambda u_\lambda^{(n)} \tag{3.14}$$

for a suitable $\zeta_\lambda^{(n)} \in \partial\|\cdot\|_q(u_\lambda^{(n)})$. Now let $(\xi^{(n)}, \eta^{(n)}) \in DJ(u^{(n)}, v^{(n)})$. Our goal is to show that we can choose $(\xi^{(n)}, \eta^{(n)})$ such that it converges to $(0, 0)$ as $n \to \infty$. By the definition of DJ and by (3.14), we have

$$\begin{aligned} \xi_\lambda^{(n)} &\in [2T^*(Tu^{(n)} - g)]_\lambda + v_\lambda^{(n)}\partial\|\cdot\|_q(u_\lambda^{(n)}) + 2\omega_\lambda u_\lambda^{(n)} \\ &= v_\lambda^{(n)}\partial\|\cdot\|_q(u_\lambda^{(n)}) - v_\lambda^{(n-1)}\zeta_\lambda^{(n)} \end{aligned}$$

for a suitable $\zeta_\lambda^{(n)} \in \partial\|\cdot\|_q(u_\lambda^{(n)})$. By (3.12) it is possible to choose the sequence $\xi^{(n)}$ such that $\lim_{n\to\infty} \xi_\lambda^{(n)} = 0$ for all $\lambda \in \Lambda$. From (3.13) it is straightforward to check that

$$0 \in \partial s^+(v_\lambda^{(\infty)})\|u_\lambda^{(\infty)}\|_q + 2\theta_\lambda(v_\lambda^{(\infty)} - \rho_\lambda) \quad \text{for all } \lambda \in \Lambda, \tag{3.15}$$

and similarly

$$0 \in \partial s^+(v_\lambda^{(n)})\|u_\lambda^{(n)}\|_q + 2\theta_\lambda(v_\lambda^{(n)} - \rho_\lambda).$$

Thus by the definition of $DJ(u^{(n)}, v^{(n)})$, we can choose $\eta^{(n)} = 0$ so that $\lim_n \eta_\lambda^{(n)} = 0$ for all $\lambda \in \Lambda$. Altogether we conclude that $0 \in \lim_n (DJ(u^{(n)}, v^{(n)}))_\lambda$ for all $\lambda \in \Lambda$. By the continuity of T and Lemma 3.7, we conclude that

$$\begin{aligned} 0 &\in \lim_n \left[(2(T^*T(u^{(n)} - g))_\lambda, 0) + \partial\Phi_\lambda^{(q)}(u_\lambda^{(n)}, v_\lambda^{(n)})\right] \\ &\subset (2T^*T(u^{(\infty)} - g)_\lambda, 0) + \partial\Phi_\lambda^{(q)}(u_\lambda^{(\infty)}, v_\lambda^{(\infty)}) = DJ(u^{(\infty)}, v^{(\infty)})_\lambda \end{aligned}$$

for all $\lambda \in \Lambda$. It follows that $0 \in DJ(u^{(\infty)}, v^{(\infty)}) \subset \partial J(u^{(\infty)}, v^{(\infty)})$, the latter inclusion by Proposition 3.5. Hence, by strict convexity $(u^*, v^*) = (u^{(\infty)}, v^{(\infty)})$. With this we have shown the weak convergence of the sequence $u^{(n)}$ to u^*.

To establish the strong convergence we need to develop a more detailed analysis of the minimization of J with respect to u. The next section is devoted to this end, and it will allow us to use some further tools for the full proof of Theorem 3.1 in section 5.3.

4. An iterative thresholding algorithm for the minimization with respect to u. One step of the minimization algorithm in the previous section consists in minimizing $J(u, v) = J_{\theta,\rho,\omega}^{(q)}(u, v)$ for some fixed v. Moreover, keeping v fixed is also interesting on its own—in particular, if one is interested in minimizing the functional $K = K_v^{(q)}$ defined in (4.2). Indeed, for $\omega = 0$ and $\rho = v$ we have $J_{\theta,v,0}^{(q)}(u, v) = K_v^{(q)}(u)$. As we will describe in the following this minimization task can be performed by a thresholded Landweber algorithm similar to the one analyzed by Daubechies, Defrise, and DeMol in [17].

With v fixed our task is equivalent to minimizing

$$K(u) = K_{v,\omega}^{(q)} := \|Tu - g\|_{\mathcal{H}}^2 + \Psi(u) \tag{4.1}$$

with respect to $u \in \ell_2(\Lambda, \mathbb{R}^M)$, where

$$\Psi(u) := \Psi_{v,\omega}^{(q)}(u) := \sum_{\lambda\in\Lambda} v_\lambda\|u_\lambda\|_q + \sum_{\lambda\in\Lambda} \omega_\lambda\|u_\lambda\|_2^2. \tag{4.2}$$

We assume that T is nonexpansive, i.e., $\|T\| < 1$, which can always be achieved by rescaling. Also we suppose that K is strictly convex. This is ensured if, e.g., the kernel of T is trivial or $\omega_\lambda > 0$ for all $\lambda \in \Lambda$.

We define a surrogate functional by

$$K^s(u, a) := K(u) - \|Tu - Ta\|_{\mathcal{H}}^2 + \|u - a\|_2^2 = \|Tu - g\|_{\mathcal{H}}^2 + \Psi(u) - \|Tu - Ta\|_{\mathcal{H}}^2 + \|u - a\|_2^2.$$

Since $\|T\| < 1$, K^s is also convex; see [17] for a rigorous argument. Now starting with some $u^{(0)} \in \ell_2(\Lambda, \mathbb{R}^M)$ we define a sequence $u^{(m)}$ by

$$u^{(m+1)} = \arg\min_{u\in\ell^2(\Lambda,\mathbb{R}^M)} K^s(u, u^{(m)}).$$

The minimizer of $K^s(u,a)$ (for fixed a) can be determined explicitly as follows. First, we claim that

$$\arg\min_u K^s(u,a) \;=\; U_\Psi(a + T^*(g - Ta)),$$

where the "thresholding" operator U_Ψ is defined as

$$U_\Psi(u) \;:=\; \arg\min_{z\in\ell_2(\Lambda,\mathbb{R}^M)} \|u-z\|_2^2 + \Psi(z). \tag{4.3}$$

Indeed, a direct calculation shows that

$$\begin{aligned} K^s(u,a) &= \|Tu-g\|_{\mathcal{H}}^2 - \|Tu-Ta\|_{\mathcal{H}}^2 + \|u-a\|_2^2 + \Psi(u) \\ &= \|(a+T^*(g-Ta))-u\|_2^2 + \Psi(u) \\ &\quad - \|a+T^*(g-Ta)\|_2^2 + \|g\|_{\mathcal{H}}^2 - \|Ta\|_{\mathcal{H}}^2 + \|a\|_2^2. \end{aligned}$$

Since the last terms (after $\Psi(u)$) do not depend on u, they can be discarded when minimizing with respect to u, and the above claim follows. (The same argument works also for general "sparseness measures" Ψ.) Thus, the iterative algorithm reads

$$u^{(m+1)} \;=\; U_\Psi(u^{(m)} + T^*(g - Tu^{(m)})). \tag{4.4}$$

In the following we give more details about U_Ψ and analyze the convergence of this algorithm.

4.1. The thresholding operator. Let us derive more information about U_Ψ for our specific $\Psi = \Psi_{v,\omega}^{(q)}$ in (4.2). We have the following lemma.

LEMMA 4.1. *Let $1 \le q \le \infty$. It holds that*

$$(U_{v,\omega}^{(q)}(u))_\lambda \;:=\; (U_{\Psi_{v,\omega}^{(q)}}(u))_\lambda \;=\; (1+\omega_\lambda)^{-1} S_{v_\lambda}^{(q)}(u_\lambda),$$

where

$$S_v^{(q)}(x) \;=\; \arg\min_{z\in\mathbb{R}^M} \|z-x\|_2^2 + v\|z\|_q, \quad x\in\mathbb{R}^M. \tag{4.5}$$

Furthermore, $S_v^{(q)}$ is given by

$$S_v^{(q)}(x) \;=\; x - P_{v/2}^{q'}(x), \tag{4.6}$$

where $P_{v/2}^{q'}$ denotes the orthogonal projection onto the norm ball of radius $v/2$ with respect to the dual norm of $\|\cdot\|_q$, i.e., the $\|\cdot\|_{q'}$ norm with q' denoting the dual index, $1/q + 1/q' = 1$. (The analogous result holds also if the norm $\|\cdot\|_q$ is replaced by an arbitrary norm on $\mathbb{R}^M$.)

Proof. For $\Psi_{v,\omega}$ the minimizing problem defining $U_\Psi = U_{v,\omega}^{(q)}$ decouples with respect to $\lambda\in\Lambda$. Thus, we have

$$(U_{v,\omega}^{(q)}(x))_\lambda = \arg\min_{z\in\mathbb{R}^M} \|x_\lambda - z\|_2^2 + \omega_\lambda\|z\|_2^2 + v_\lambda\|z\|_q.$$

If z minimizes the latter term, then necessarily $0 \in 2(1+\omega_\lambda)z - 2x + v_\lambda\partial\|\cdot\|_q(z)$, where $\partial\|\cdot\|_q$ denotes the subdifferential of the q norm. In other words,

$$(1+\omega_\lambda)z - x \in -\frac{v_\lambda}{2}\partial\|\cdot\|_q(z).$$

Since $\|\cdot\|_q$ is 1-homogeneous, we have $\partial\|\cdot\|_q(z) = \partial\|\cdot\|_q((1+\omega_\lambda)z)$. Setting $y = (1+\omega_\lambda)z$ gives $y - x \in -\frac{v_\lambda}{2}\partial\|\cdot\|_q(y)$, which is the above relation for $\omega_\lambda = 0$. From this we deduce the first claim.

Let us show the second claim, i.e., the explicit form of the operator $S_v^{(q)}$. We already know that if z minimizes the left-hand side of (4.5), then $x - z \in \partial\frac{v}{2}\|z\|_q$. Let $\psi(z) = \frac{v}{2}\|z\|_q$ and ψ^* be its Fenchel conjugate function defined by $\psi^*(y) = \sup_x(\langle x, y\rangle - \psi(x))$. It is well known [4, p. 93] that

$$\psi^*(y) \;=\; \chi_{B^{q'}(v/2)}(y) \;:=\; \begin{cases} 0 & \text{if } \|y\|_{q'} \le v/2, \\ \infty & \text{otherwise.} \end{cases}$$

Here $B^{q'}(v/2)$ denotes the norm ball of radius $v/2$ with respect to the dual norm of $\|\cdot\|_q$. It is a standard result (see, e.g., [38, Proposition 11.3], [24, Corollary 5.2]) that $w \in \partial\psi(y)$ if and only if $y \in \partial\psi^*(w)$ yielding $z \in \partial\psi^*(x-z)$ in our case, and hence,

$$x \in x - z + \partial\psi^*(x-z) \;=\; x - z + \partial\chi_{B^{q'}(v/2)}(x-z).$$

Now if $y \in w + \partial\chi_{B^{q'}(v/2)}(w)$, then it is straightforward to see that w must be the orthogonal projection of y onto $B^{q'}(v/2)$, i.e., $w = \arg\min_{w' \in B^{q'}(v/2)} \|w' - y\|_2$; see also [38, Example 10.2 and p. 20]. For our situation this means that $x - z = P_{v/2}^{q'}(x)$, i.e., $z = x - P_{v/2}^{q'}(x)$. This shows the second claim.

Clearly, all arguments work also for a general norm rather than the q norm. ∎

Let us give $S_v^{(q)}$ explicitly for $q = 1, 2, \infty$.

LEMMA 4.2. *Let $x \in \mathbb{R}^M$ and $v \ge 0$.*

(a) *For $q = 1$ we have $S_v^{(1)}(x) = (s_v^{(1)}(x_\ell))_{\ell=1}^M$, where for $y \in \mathbb{R}$*

$$s_v^{(1)}(y) \;=\; \begin{cases} 0 & \text{if } |y| \le \frac{v}{2}, \\ \operatorname{sign}(y)(|y| - \frac{v}{2}) & \text{otherwise.} \end{cases}$$

(b) *For $q = 2$ it holds that*

$$S_v^{(2)}(x) := \begin{cases} 0 & \text{if } \|x\|_2 \le \frac{v}{2}, \\ \frac{(\|x\|_2 - v/2)}{\|x\|_2} x & \text{otherwise.} \end{cases}$$

(c) *Let $q = \infty$. Order the entries of x by magnitude such that $|x_{i_1}| \ge |x_{i_2}| \ge \cdots \ge |x_{i_M}|$.*

1. *If $\|x\|_1 < v/2$, then $S_v^{(\infty)}(x) = 0$.*
2. *Suppose $\|x\|_1 > v/2$. If $|x_{i_2}| < |x_{i_1}| - v/2$, then choose $n = 1$. Otherwise, let $n \in \{2, \dots, M\}$ be the largest index satisfying*

$$|x_{i_n}| \;\ge\; \frac{1}{n-1}\left(\sum_{k=1}^{n-1} |x_{i_k}| - \frac{v}{2}\right). \tag{4.7}$$

Then

$$(S_v^{(\infty)}(x))_{i_j} \;=\; \frac{\operatorname{sign}(x_{i_j})}{n}\left(\sum_{k=1}^{n} |x_{i_k}| - \frac{v}{2}\right), \quad j = 1, \dots, n,$$

$$(S_v^{(\infty)}(x))_{i_j} \;=\; x_{i_j}, \quad j = n+1, \dots, M.$$

Proof. (b) The projection $P^2_{v/2}(x)$ of x onto an ℓ_2 ball of radius $v/2$ is clearly given by

$$P^2_{v/2}(x) \;=\; \begin{cases} x & \text{if } \|x\|_2 \le v/2, \\ \frac{v/2}{\|x\|_2} x & \text{otherwise.} \end{cases}$$

Since by the previous lemma $S^{(2)}_v(x) = x - P^2_{v/2}(x)$, this gives the assertion.

(a) Although this is well known, we give a simple argument. For $q = 1$ the functional in (4.5) defining $S^{(1)}$ decouples, i.e.,

$$S^{(1)}_v(x) \;=\; \arg\min_{z\in\mathbb{R}^M} \sum_{\ell=1}^{M} \left(|z_\ell - x_\ell|^2 + v|z_\ell|\right).$$

Thus, $S^{(1)}_v(x)_\ell = \arg\min_{z_\ell\in\mathbb{R}} |z_\ell - x_\ell|^2 + v|x_\ell|$ for all $\ell = 1,\dots,M$. The latter can be interpreted as the problem for $q = 2$ on $\mathbb{R}^1$, and hence, the assertion follows from (b).

(c) If $\|x\|_1 \le v/2$, then $P^1_{v/2}(x) = x$, and by the previous lemma $S^{(\infty)}_v(x) = x - P^1_{v/2}(x) = 0$. Now assume $\|x\|_1 > v/2$. Let $z = S^{(\infty)}_v(x)$. This is equivalent to 0 being contained in the subdifferential of the functional in (4.5) defining $S^{(\infty)}_v$. This means

$$2(z - x) \in -v\partial\|\cdot\|_\infty(z). \tag{4.8}$$

We recall that the subdifferential of the maximum norm is given by (3.6).

Now assume for the moment that the maximum norm of z is attained in $z_{i_1},\dots,z_{i_n}$. We will later check whether this was really the case. Further, we assume for simplicity that all of the entries $x_{i_1},\dots,x_{i_n}$ are positive. (The other cases can be carried through in the same way.) Then certainly also the numbers $z_{i_1},\dots,z_{i_n}$ are positive, because choosing them with the opposite sign would certainly increase the functional defining $S^{(\infty)}_v$. Then by (3.6) we obtain $2(z_{i_j} - x_{i_j}) = 0$ for the entries z_{i_j} not giving the maximum, i.e., $z_{i_j} = x_{i_j}, j = n+1,\dots,M$.

Moreover, if $n = 1$ (i.e., the maximum norm of z is attained at only one entry), then $2(z_{i_1} - x_{i_1}) = -v$; in other words, $z_{i_1} = x_{i_1} - v/2$. Thus, the initial hypothesis that the maximum norm of z is attained only at z_{i_1} is true if and only if the second largest entry x_{i_2} satisfies $|x_{i_2}| < |x_{i_1}| - v/2$.

So if the latter inequality is not satisfied, then the maximum norm of z is at least attained at two entries, i.e., $n \ge 2$. In this case by (3.6) the entries $z_{i_1} = z_{i_2} = \cdots = z_{i_n} = t$ satisfy

$$2t - 2x_{i_j} \;=\; -va_j, \quad j = 1,\dots,n-1,$$

$$2t - 2x_{i_n} \;=\; -v\left(1 - \sum_{k=1}^{n-1} a_k\right)$$

for some numbers $a_1,\dots,a_{n-1} \in [0,1]$ satisfying $\sum_j a_j \le 1$. This is a system of n linear equations in t and $a_1,\dots,a_{n-1}$. Writing it in matrix form, we get

$$\begin{pmatrix} 1 & v/2 & 0 & 0 & \cdots & 0 \\ 1 & 0 & v/2 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & -v/2 & -v/2 & -v/2 & \cdots & -v/2 \end{pmatrix} \begin{pmatrix} t \\ a_1 \\ \vdots \\ a_{n-1} \end{pmatrix} = \begin{pmatrix} x_{i_1} \\ \vdots \\ x_{i_{n-1}} \\ x_{i_n} - v/2 \end{pmatrix}.$$

Denoting the matrix on the left-hand side by B, a simple computation verifies that

$$B^{-1} = \frac{1}{n}\begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ \frac{2(n-1)}{v} & -\frac{2}{v} & -\frac{2}{v} & \cdots & -\frac{2}{v} \\ -\frac{2}{v} & \frac{2(n-1)}{v} & -\frac{2}{v} & \cdots & -\frac{2}{v} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ -\frac{2}{v} & \cdots & -\frac{2}{v} & \frac{2(n-1)}{v} & -\frac{2}{v} \end{pmatrix}.$$

This gives

$$z_{i_1} = \cdots = z_{i_n} = t = \frac{1}{n}\left(\sum_{j=1}^{n} x_{i_j} - v/2\right)$$

and $a_j = \frac{2}{nv}(v/2 + (n-1)x_{i_j} - \sum_{k\in\{1,\ldots,n\}\setminus\{j\}} x_{i_k})$. Thus, all a_j are nonnegative if for all $j \in \{1,\ldots,n-1\}$

$$x_{i_j} \geq \frac{1}{n-1}\left(\sum_{k\in\{1,\ldots,n\}\setminus\{j\}} x_{i_k} - v/2\right).$$

Moreover, a simple calculation gives $\sum_{j=1}^{n-1} a_j = \frac{n-1}{n} + \frac{2}{nv}(\sum_{j=1}^{n-1} x_{i_j} - (n-1)x_{i_n})$. Thus, it holds that $1 - \sum_{j=1}^{n-1} a_j \geq 0$ if and only if

$$x_{i_n} \geq \frac{1}{n-1}\left(\sum_{j=1}^{n-1} x_{i_j} - v/2\right).$$

Therefore, the initial assumption that the maximum norm of u is attained precisely at $z_{i_1},\ldots,z_{i_n}$ can be true only if $x_{i_1},\ldots,x_{i_n}$ are the largest entries of the vector x and

$$z_{i_{n+1}} = x_{i_{n+1}} < t = \frac{1}{n}\left(\sum_{j=1}^{n} x_{i_j} - v/2\right),$$

i.e., $|x_{i_{n+1}}| < n^{-1}(\sum_{j=1}^{n} |x_{i_j}| - v/2)$. Pasting all of the pieces together shows the assertion of the lemma. □

4.2. Weak convergence. In the following we will prove that $u^{(m)}$ converges weakly and strongly to the unique minimizer of K. We first establish the weak convergence. Following the proof of Proposition 3.11 in [17], one may extract essentially three conditions on a general sparsity measure Ψ such that weak convergence is ensured. Let us collect them in the following proposition.

PROPOSITION 4.3. *Assume K is given by* (4.1) *with a general sparsity measure Ψ, and suppose K is strictly convex. Let U_Ψ be the associated "thresholding operator" given by* (4.3). *Assume that the following conditions hold:*

(1) *U_Ψ is nonexpansive; i.e., $\|U_\Psi(x) - U_\Psi(y)\|_2 \leq \|x-y\|_2$ for all $x, y \in \ell_2(\Lambda, \mathbb{R}^M)$.*

(2) *It holds that* $\|f\|_2 \le H(\Psi(f))$ *for all* $f \in \ell_2(\Lambda, \mathbb{R}^M)$ *and some monotonically increasing function* H *on* $\mathbb{R}_+$. *(This ensures that a sequence* f_n *satisfying* $\Psi(f_n) \le C$ *is bounded in* $\ell_2(\Lambda, \mathbb{R}^M)$.*)*

(3) *For all* $x, h \in \ell_2(\Lambda, \mathbb{R}^M)$ *it holds that*

$$\Psi(U_\Psi(x) + h) - \Psi(U_\Psi(x)) + 2\langle h, U_\Psi(x) - x\rangle \ge 0.$$

Then the sequence $u^{(m)}$ *defined by* (4.4) *converges weakly to the minimizer of* K *independently of the choice of* $u^{(0)}$.

Proof. First we claim that the condition in (3) implies that the surrogate functional K^s satisfies

$$K^s(u + h, a) - K^s(u, a) \ \ge \ \|h\|_2^2 \tag{4.9}$$

for $u = \arg\min_{u'} K^s(u', a) = U_\Psi(a - T^*(g - Ta))$. Indeed, set $x := a - T^*(g - Ta)$, i.e., $u = U_\Psi(x)$. Then an elementary calculation yields

$$\begin{aligned} K^s(u + h, a) - K^s(u) &= \|T(u + h) - g\|_2^2 + \Psi(u + h) - \|T(u + h) - Ta\|_2^2 + \|u \\ &\quad + h - a\|_2^2 - \|Tu - g\|_2^2 - \Psi(u) + \|Tu - Ta\|_2^2 - \|u - a\|_2^2 \\ &= 2\langle h, u - a - T^*(g - Ta)\rangle + \Psi(u + h) - \Psi(u) + \|h\|_2^2 \\ &= 2\langle h, U_\Psi(x) - x\rangle + \Psi(U_\Psi(x) + h) - \Psi(U_\Psi(x)) + \|h\|_2^2 \ge \|h\|_2^2. \end{aligned}$$

The relation in (3) was used in the last inequality.

Now with (4.9) and properties (1) and (2) one can justify that the proofs of the analogues of Theorem 3.2 until Proposition 3.11 in [17] go through completely in the same way, which finally leads to the statement of this proposition. ∎

Let us now show that for our specific choice of $\Psi = \Psi_{v,\omega}^{(q)}$ properties (1)–(3) in the previous proposition hold, and thus, $u^{(n)}$ converges weakly to a minimizer of K.

LEMMA 4.4. $U_{v,\omega}^{(q)} = U_{\Psi_{v,\omega}^{(q)}}$ *is nonexpansive.*

Proof. Clearly, the map $x \mapsto (1 + \omega_\lambda)^{-1} x$ is nonexpansive. By (4.6) we have $S_v^{(q)} = I - P_{v/2}^{q'}$. Since $P_{v/2}^{q'}$ is an orthogonal projection onto a convex set, $S_v^{(q)}$ is also nonexpansive; see, e.g., [40]. Hence, $U_{v,\omega}^{(q)}$ is nonexpansive, since on each component x_λ, $\lambda \in \Lambda$, it is a composition of nonexpansive operators. ∎

LEMMA 4.5. *If* (v_λ) *or* (ω_λ) *are bounded away from* 0, *then condition* (2) *in Proposition* 4.3 *holds.*

Proof. This follows by a standard argument. ∎

If we consider the problem of minimizing $J(u, v)$ jointly over u and v, then we certainly cannot assume that v is bounded away from 0, but in this case we require that ω_λ is bounded away from 0. (By Proposition 2.1 this is needed anyway to ensure that $J(u, v)$ is jointly convex in u and v.) In the case where we only minimize $J(u, v)$ with respect to u (i.e., when minimizing $\Psi(u)$ defined in (4.2)), we may take ω_λ arbitrary (and even $\omega_\lambda = 0$), but then we have to require a lower bound on v_λ.

Now consider the third condition in the proposition. The next lemma reduces it to the respective condition for $S_v^{(q)}$.

LEMMA 4.6. *Assume that for all* $x, h \in \mathbb{R}^M$ *it holds that*

$$v(\|S_v^{(q)}(x) + h\|_q - \|S_v^{(q)}(x)\|_q) + 2\langle h, S_v^{(q)}(x) - x\rangle \ge 0.$$

Then condition (3) *in Proposition* 4.3 *is satisfied.*

Proof. By the definition of $\Psi_{v,\omega}^{(q)}$ we need to show that for $\omega, v \geq 0$ and all $x, h \in \mathbb{R}^M$

$$\begin{aligned} & v(\|(1+\omega)^{-1}S_v^{(q)}(x)+h\|_q - \|(1+\omega)^{-1}S_v^{(q)}(x)\|_q) \\ & \quad + \omega(\|(1+\omega)^{-1}S_v^{(q)}(x)+h\|_2^2 - \|(1+\omega)^{-1}S_v^{(q)}(x)\|_2^2) \\ & \quad + 2\langle h, (1+\omega)^{-1}S_v^{(q)}(x) - x\rangle \geq 0. \end{aligned}$$

Setting $h' = (1+\omega)h$, a straightforward computation yields for the left-hand side of this inequality

$$(1+\omega)^{-1}\left[v(\|S_v^{(q)}(x)+h'\|_q - \|S_v^{(q)}(x)\|_q) + 2\langle h', S_v^{(q)}(x) - x\rangle\right]$$

$$+(1+\omega)^{-2}\omega\left[\|S_v^{(q)}(x)+h'\|_2^2 - \|S_v^{(q)}(x)\|_2^2 - 2\langle h', S_v^{(q)}(x)\rangle\right] \geq (1+\omega)^{-2}\omega\|h'\|_2^2 \geq 0.$$

This completes the proof. ∎

LEMMA 4.7. *The condition in the previous lemma holds for* $S_v^{(q)}$, $1 \leq q \leq \infty$ *(and even if the* ℓ_q *norm is replaced by a general norm on* $\mathbb{R}^M$*).*

Proof. First note that by definition (4.5) and duality we have

$$\|S_v^{(q)}(x)\|_q \;=\; (v/2)^{-1} \sup_{k \in B^{q'}(v/2)} \langle k, x - P_{v/2}^{q'}x\rangle.$$

A characterization of the orthogonal projection tells us that $\langle k - P_{v/2}^{q'}x, x - P_{v/2}^{q'}x\rangle \leq 0$ for all $k \in B^{q'}(v/2)$; see, e.g., [40, Lemma 8]. This gives

$$\begin{aligned} \|S_v^{(q)}(x)\|_q \;&=\; (v/2)^{-1} \sup_{k \in B^{q'}(v/2)} \left(\langle k - P_{v/2}^{q'}x, x - P_{v/2}^{q'}(x)\rangle + \langle P_{v/2}^{q'}(x), S_v^{(q)}(x)\rangle\right) \\ &\leq\; (v/2)^{-1}\langle P_{v/2}^{q'}(x), S_v^{(q)}(x)\rangle. \end{aligned}$$

Using once more that $S_v^{(q)}(x) - x = -P_{v/2}^{q}(x)$, we further obtain

$$\begin{aligned} & v(\|S_v^{(q)}(x)+h\|_q - \|S_v(q)(x)\|_q) + 2\langle h, S_v^{(q)}(x) - x\rangle \\ & \quad \geq v\|S_v^{(q)}(x)+h\|_q - 2\langle P_{v/2}^{q'}(x), S_v^{(q)}(x)\rangle - 2\langle P_{v/2}^{q'}(x), h\rangle \\ & \quad = v\|S_v^{(q)}(x)+h\|_q - 2\langle P_{v/2}^{q'}(x), S_v^{(q)}(x)+h\rangle \\ & \quad \geq v\|S_v^{(q)}(x)+h\|_q - 2\|P_{v/2}^{q'}x\|_{q'}\|S_v^{(q)}(x)+h\|_q \;\geq\; 0. \end{aligned}$$

Here we used that $P_{v/2}^{q'}$ is a projection onto $B^{q'}(v/2)$, so $\|P_{v/2}^{q'}x\|_{q'} \leq v/2$. This finishes the proof. ∎

To summarize we have the following result about weak convergence.

COROLLARY 4.8. *Let* $1 \leq q \leq \infty$*, and assume that* (v_λ) *or* (ω_λ) *is bounded from below. Then the sequence* $u^{(m)}$ *defined in* (4.4) *converges weakly to a minimizer of* K*, where* $\Psi = \Psi_{v,\omega}^{(q)}$ *is the sparsity measure defined in* (4.2)*. (The* q *norm in* (4.2) *can be replaced by any other norm on* $\mathbb{R}^M$*.)*

4.3. Strong convergence. The next result establishes the strong convergence.

PROPOSITION 4.9. *Let $1 \le q \le \infty$, and assume that (v_λ) or (ω_λ) are bounded away from 0. In case (v_λ) is not bounded away from 0, assume further that there is a constant $c > 0$ such that $v_\lambda < c$ for only finitely many λ. Then $u^{(m)}$ converges strongly to a minimizer of K.*

Proof. The analogues of Lemmas 3.15 and 3.17 in [17] are proven in completely the same way. It remains to justify the analogue of [17, Lemma 3.18]: If for some $a \in \ell_2(\Lambda, \mathbb{R}^M)$ and some sequence $(h^{(m)}) \subset \ell_2(\Lambda, \mathbb{R}^M)$ converging weakly to 0 it holds that $\lim_{m\to\infty} \|U^{(q)}_{v,\omega}(a+h^{(m)}) - U^{(q)}_{v,\omega}(a) - h^{(m)}\|_2 = 0$, then $\|h^{(m)}\|_2 \to 0$ for $m \to \infty$. To this end we mainly follow the argument in [17].

Let c be the constant such that $v_\lambda < c$ for $\lambda \in \Lambda_{00}$ for Λ_{00} finite. Then let Λ_{01} be a finite set such that $\sum_{\lambda\in\Lambda\setminus\Lambda_{01}} \|a_\lambda\|^2_{q'} \le \sigma^2$ for some $\sigma < c/2$. (Such a set Λ_{01} exists since $\|\cdot\|_{q'}$ and $\|\cdot\|_2$ are equivalent norms on $\mathbb{R}^M$ and by assumption $a \in \ell_2(\Lambda, \mathbb{R}^M)$.) Since $\Lambda_0 = \Lambda_{00} \cup \Lambda_{01}$ is also finite, we have $\sum_{\lambda\in\Lambda_0} \|h^{(m)}_\lambda\|^2_2 \to 0$ for $m \to \infty$ by the weak convergence of $h^{(m)}$ to 0. Thus, we are left with proving that $\sum_{\lambda\in\Lambda\setminus\Lambda_0} \|h^{(m)}_\lambda\|^2_2 \to 0$ for $m \to \infty$.

For each m we split $\Lambda_1 := \Lambda\setminus\Lambda_0$ into the subsets $\Lambda_{1,m} := \{\lambda \in \Lambda_1 : \|h^{(m)}_\lambda + a_\lambda\|_{q'} < v_\lambda/2\}$ and $\widetilde{\Lambda}_{1,m} = \Lambda_1 \setminus \Lambda_{1,m}$. If $\lambda \in \Lambda_{1,m}$, then $U^{(q)}_{v,\omega}(a+h^{(m)})_\lambda = U^{(q)}_{v,\omega}(a)_\lambda = 0$ since $\|a_\lambda + h^{(m)}_\lambda\|_{q'}, \|a_\lambda\|_{q'} \le v_\lambda/2$. Thus, $\|h^{(m)}_\lambda - U^{(q)}_{v,\omega}(a+h^{(m)})_\lambda + U^{(q)}_{v,\omega}(a)_\lambda\|^2_2 = \|h^{(m)}_\lambda\|^2_2$ and by assumption

$$\sum_{\lambda\in\Lambda_1} \|h^{(m)}_\lambda\|^2_2 \le \|h^{(m)} - U^{(q)}_{v,\omega}(a+h^{(m)}) + U^{(q)}_{v,\omega}(a)\|^2_2 \to 0 \quad \text{as } m \to \infty.$$

Now let $\lambda \in \widetilde{\Lambda}_{1,m}$. We first consider the case that $\omega_\lambda = 0$, i.e., $U^{(q)}_{v,\omega}(x)_\lambda = S^{(q)}_v(x)_\lambda$. Since $\|a_\lambda\|_{q'} \le \sigma < v_\lambda/2$, we have $S^{(q)}_v(a) = 0$, and thus,

$$\begin{aligned}
&\|h^{(m)}_\lambda - S^{(q)}_{v_\lambda}(h^{(m)}_\lambda + a_\lambda)\|_{q'} \\
&\quad = \|h^{(m)}_\lambda - (h^{(m)}_\lambda + a_\lambda) + P^{q'}_{v_\lambda/2}(h^{(m)}_\lambda + a_\lambda)\|_{q'} \\
&\quad = \|P^{q'}_{v_\lambda/2}(h^{(m)}_\lambda + a_\lambda) - a_\lambda\|_{q'} \ge \|P^{q'}_{v_\lambda/2}(h^{(m)}_\lambda + a_\lambda)\|_{q'} - \|a_\lambda\|_{q'} \\
&\quad \ge v_\lambda/2 - \sigma \ge c/2 - \sigma.
\end{aligned}$$

Here we used that $\|P^{q'}_{v_\lambda/2}(h^{(m)}_\lambda + a_\lambda)\|_{q'} = v_\lambda/2$ (because $\|h^{(m)}_\lambda + a_\lambda\|_{q'} \ge v_\lambda/2$). Since every norm on a finite-dimensional space is equivalent, there is a constant C such that

$$\|h^{(m)}_\lambda - S^{(q)}_{v_\lambda}(h^{(m)}_\lambda + a_\lambda) + S^{(q)}_{v_\lambda}(a_\lambda)\|^2_2 \ge C^2(c/2 - \sigma)^2 > 0.$$

However, since by assumption $\sum_{\lambda\in\Lambda} \|h^{(m)}_\lambda - S^{(q)}_{v_\lambda}(h^{(m)}_\lambda + a_\lambda) + S^{(q)}_{v_\lambda}(a)\|^2_2 \to 0$ as $m \to \infty$, there must exist an m_0 such that $\widetilde{\Lambda}_{1,m}$ is empty for all $m \ge m_0$.

In the case that ω_λ does not vanish, we have

$$\begin{aligned}
&\|h^{(m)}_\lambda - U^{(q)}_{v,\omega}(h^{(m)} + a)_\lambda + U^{(q)}_{v,\omega}(a)_\lambda\|_2 \\
&\quad = \|h^{(m)}_\lambda - (1+\omega_\lambda)^{-1} S^{(q)}_{v_\lambda}(h^{(m)}_\lambda + a_\lambda)\|_2 \\
&\quad = (1+\omega_\lambda)^{-1}\|(1+\omega_\lambda)h^{(m)}_\lambda - S^{(q)}_{v_\lambda}(h^{(m)}_\lambda + a_\lambda)\|_2.
\end{aligned} \tag{4.10}$$

We claim that

$$\|(1+\omega_\lambda)h_\lambda^{(m)} - S_{v_\lambda}^{(q)}(h_\lambda^{(m)}+a_\lambda)\|_2 \;\geq\; \|h_\lambda^{(m)} - S_{v_\lambda}^{(q)}(h_\lambda^{(m)}+a_\lambda)\|_2 \tag{4.11}$$

so that we can apply the argument for $\omega_\lambda = 0$ to conclude that $\widetilde{\Lambda}_{1,m}$ is empty for m sufficiently large. Let us omit for the moment all indexes λ and m for the sake of simpler notation. We have

$$\|(1+\omega)h - S_v^{(q)}(h+a)\|_2^2 - \|h - S_v^{(q)}(h+a)\|_2^2 \;=\; 2\omega\langle h, h - S_v^{(q)}(h+a)\rangle + \omega^2\|h\|_2^2 \tag{4.12}$$

and, furthermore,

$$\begin{aligned}\langle h, h - S_v^{(q)}(h+a)\rangle &= \langle h, P_{v/2}^{q'}(h+a) - a\rangle \\ &= -\langle h + a - P_{v/2}^{q'}(h+a), a - P_{v/2}^{q'}(h+a)\rangle \\ &\quad + \|a - P_{v/2}^{q'}(h+a)\|_2^2 \geq 0. \end{aligned} \tag{4.13}$$

Here we used that $a \in B^{q'}(\sigma) \subset B^{q'}(v/2)$ and the fact that $\langle k - P_{v/2}^{q'}(x), x - P_{v/2}^{q'}(x)\rangle \leq 0$ for all $k \in B^{q'}(v/2)$ and $x \in \mathbb{R}^M$. Thus, the term in (4.12) is nonnegative, and therefore our claim (4.11) holds. □

Let us shortly comment on the condition that, if v_λ is not bounded from below, there is at least some $c > 0$ such that $v_\lambda > c$ except for a finite set of indexes λ. This condition is mainly relevant when considering also a minimization over (v_λ). Then the term $\sum \theta_\lambda(\rho_\lambda - v_\lambda)^2$ in the functional $J(u,v)$ ensures that the sequence $(\rho_\lambda - v_\lambda)$ is contained in $\ell_{2,\theta^{1/2}}$. If θ_λ and ρ_λ are bounded from below, this implies that v_λ can be less than $1/2 \min_\lambda \rho_\lambda$, say, only for finitely many λ.

5. Numerical implementation and convergence analysis. The scope of this section is twofold: We want to formulate an implementable version of the double-minimization algorithm and show its strong convergence. To this end we develop an error analysis.

5.1. Numerical implementation. Let us compose the two iterative algorithms described in (3.1) and (4.4), respectively, into a unique scheme.

ALGORITHM 1. **JOINTSPARSE**

Input: Data vector $(g_j)_{j=1}^N$, initial points $u^{(0)} \in \ell_2(\Lambda, \mathbb{R}^M)$, $v^{(0)}$, with $0 \leq v_\lambda^{(0)} \leq \rho_\lambda$, number $n_{\max}$ of outer iterations, number of inner iterations L_n, $n = 1, \ldots, n_{\max}$.

Parameters: $q \in [1, \infty]$, positive weights (θ_λ), (ρ_λ), (ω_λ), with $\omega_\lambda \geq c > 0$, such that $\Phi^{(q)}$ and hence J are convex; see Proposition 2.1.

Output: Approximation (u^*, v^*) of the minimizer of $J_{\theta,\rho,\omega}^{(q)}$

$u^{(0,0)} := u^{(0)}$;

for $n := 0$ *to* $n_{\max}$ *do*

  for $m := 0$ *to* L_n *do*

    $u^{(n,m+1)} := U_{v^{(n)},\omega}^{(q)}\left(u^{(n,m)} + T^*(g - Tu^{(n,m)})\right)$;

  endfor

  $u^{(n+1,0)} := u^{(n,L_n)}$;

  $v^{(n+1)} := \left(\begin{cases} \rho_\lambda - \frac{1}{2\theta_\lambda}\|u^{(n+1,0)}{}_\lambda\|_q, & \|u^{(n+1,0)}{}_\lambda\|_q < 2\theta_\lambda\rho_\lambda, \\ 0, & \text{otherwise.} \end{cases} \right)_{\lambda\in\Lambda}$;

endfor
$u^* := u^{(n_{\max}, L_{n_{\max}})}$;
$v^* := v^{(n_{\max})}$.

We will discuss below how to choose the number of inner iterations L_n and outer iterations $n_{\max}$. Our previous analysis ensures the (weak) convergence of this scheme only if the inner loop computes exactly the minimizer of $J(u, v^{(n)})$ for fixed $v^{(n)}$, i.e., if $L_n = \infty$. Of course, this cannot be numerically realized, so we need to analyze what happens if the inner loop makes a small error in computing this minimizer. In other words, we will clarify how large we have to choose $n_{\max}$ and $L_n, n = 1, \dots, n_{\max}$, in order to ensure that we have approximately computed the minimizer (u^*, v^*) within a given error tolerance.

The role of the other parameters $\rho_\lambda, \theta_\lambda, \omega_\lambda$ are investigated in the subsequent contribution [27]. Observe further that each (inner) iteration of the above algorithm involves an application of T^*T and of the thresholding operator $U^{(q)}_{v,\omega}$. The latter can be applied rapidly. So if there is also a fast algorithm for the computation of T^*T, then each iteration can be done rapidly.

5.2. Strong convergence of JOINTSPARSE. First of all we want to establish the convergence rate of the inner loop, i.e., the iterative thresholding algorithm of the previous section.

PROPOSITION 5.1. *Assume that $\omega_\lambda \geq \gamma > 0$ for all $\lambda \in \Lambda$ (implying that $K(u) = K^{(q)}_{v,\omega}(u)$ is strictly convex) and $\|T\| < 1$. Set $\alpha := (1+\gamma)^{-1}\|I - T^*T\| < 1$. Then the iterative thresholding algorithm*

$$u^{(n,m+1)} := U^{(q)}_{v^{(n)},\omega}\left(u^{(n,m)} + T^*(g - Tu^{(n,m)})\right)$$

converges linearly:

$$\|u^{(n,\infty)} - u^{(n,m+1)}\|_2 \leq \alpha \|u^{(n,\infty)} - u^{(n,m)}\|_2. \tag{5.1}$$

Proof. Note that

$$u^{(n,\infty)} := U^{(q)}_{v^{(n)},\omega}\left(u^{(n,\infty)} + T^*(g - Tu^{(n,\infty)})\right).$$

By nonexpansiveness of $S^{(q)}_v$ (see Lemma 4.4 and its proof) we obtain

$$\begin{aligned}
&\|u^{(n,\infty)} - u^{(n,m+1)}\|_2 \\
&\quad = \left\| U^{(q)}_{v^{(n)},\omega}\left(u^{(n,\infty)}_\lambda + T^*(g - Tu^{(n,\infty)}_\lambda)\right) - U^{(q)}_{v^{(n)},\omega}\left(u^{(n,m)} + T^*(g - Tu^{(n,m)})\right)\right\|_2 \\
&\quad = \Bigg(\sum_{\lambda\in\Lambda}(1+\omega_\lambda)^{-2}\|S^{(q)}_{v_\lambda}(u^{(n,\infty)}_\lambda + T^*(g - Tu^{(n,\infty)}_\lambda)_\lambda) \\
&\qquad\quad - S^{(q)}_{v_\lambda}(u^{(n,m)} + T^*(g - Tu^{(n,m)})_\lambda)\|_2^2\Bigg)^{1/2} \\
&\quad \leq \sup_{\lambda\in\Lambda}(1+\omega_\lambda)^{-1}\|(I - T^*T)(u^{(n,\infty)} - u^{(n,m)})\|_2 \\
&\quad \leq (1+\gamma)^{-1}\|I - T^*T\|\,\|u^{(n,\infty)} - u^{(n,m)}\|_2 \\
&\quad = \alpha\|u^{(n,\infty)} - u^{(n,m)}\|_2.
\end{aligned}$$

This establishes the claim. ∎

Remark. Clearly, the error estimation in (5.1) holds also if one is interested only in analyzing the iterative thresholding algorithm from the last section (i.e., without doing the outer iteration). Then it might also be interesting to consider the case that $\omega = 0$. According to what we have proven in the previous section, the algorithm still converges provided the weight v is bounded away from zero. However, then the error estimation (5.1) has a useful meaning only if $\alpha = \|I - T^*T\| < 1$. So this applies if T^*T is boundedly invertible. For a usual inverse problem, however, we will have a noninvertible T or at least one with an unbounded inverse resulting in $\|I - T^*T\| = 1$. So in this case we know only that the algorithm converges, but an error estimate does not seem to be available.

For simplicity we restrict the following analysis to the most interesting cases $q \in \{1, 2, \infty\}$. We first need the following technical result.

LEMMA 5.2. *For $q \in \{1, 2, \infty\}$ the projection P_v^q onto the ball $B^q(v) \subset \mathbb{R}^M$ is a Lipschitz function with respect to $v \in \mathbb{R}_+$. In particular, we have*

$$\|P_v^q(x) - P_w^q(x)|\ell_2^M\| \le L|v - w| \quad \text{for all } x \in \mathbb{R}^M, \tag{5.2}$$

where $L = 1$ for $q \in \{1, 2\}$ and $L = M^{1/2}$ for $q = \infty$.

Proof. Let us start with $q = 2$. By distinguishing cases it is not difficult to show that

$$\|P_v^2(x) - P_w^2(x)|\ell_2^M\| \le |v - w|.$$

For $q = \infty$ we have $P_v^\infty(x) = (p_v^\infty(x_\ell))_{\ell=1}^M$, where for $y \in \mathbb{R}$

$$p_v^\infty(y) = \begin{cases} y & \text{if } |y| \le v, \\ y - \operatorname{sign}(y)(|y| - v) & \text{otherwise.} \end{cases}$$

Since p_v^∞ can be interpreted as a projection onto the ℓ_2 ball in dimension 1, we obtain that

$$|p_v^\infty(y) - p_w^\infty(y)| \le |v - w|,$$

and

$$\|P_v^\infty(x) - P_w^\infty(x)|\ell_2^M\| = \left(\sum_{\ell=1}^M |p_v^\infty(x_\ell) - p_w^\infty(x_\ell)|^2 \right)^{1/2} \le M^{1/2}|v - w|.$$

The case $q = 1$ requires a bit more effort. By Lemma 4.2(c) we have the following. Let x_{i_k} denote the reordering of the entries of x by magnitude as in Lemma 4.2. Let $n \in \{1, \dots, M\}$ be the largest index satisfying

$$|x_{i_n}| \ge \frac{1}{n-1} \left(\sum_{k=1}^{n-1} |x_{i_k}| - v \right).$$

Then

$$(P_v^1(x))_{i_j} = x_{i_j} - \frac{\operatorname{sign}(x_{i_j})}{n} \left(\sum_{k=1}^{n} |x_{i_k}| - v \right), \quad j = 1, \dots, n,$$

$$(P_v^1(x))_{i_j} = 0, \quad j = n+1, \dots, M.$$

Observe first that for all $x \in \mathbb{R}^M$ there exists $\varepsilon_0 > 0$ such that for all $0 < \varepsilon < \varepsilon_0$ the same $n \in \{1, \dots, M\}$ is the largest index satisfying

$$|x_{i_n}| \geq \frac{1}{n-1} \left(\sum_{k=1}^{n-1} |x_{i_k}| - (v + \varepsilon) \right).$$

For $0 < \varepsilon < \varepsilon_0$, a simple computation yields

$$\frac{(P^1_{v+\varepsilon}(x) - P^1_v(x))_{i_j}}{\varepsilon} = \begin{cases} \frac{\operatorname{sign}(x_{i_j})}{n} & \text{for } j = 1, \dots, n, \\ 0 & j = n+1, \dots, M. \end{cases}$$

This means that the map $v \to P^1_v(x)$ is right-differentiable; i.e., the limit

$$(P^1_v(x))'_+ = \lim_{\varepsilon \to 0_+} \frac{P^1_{v+\varepsilon}(x) - P^1_v(x)}{\varepsilon}$$

exists in $\mathbb{R}^M$. Moreover, it also follows that

$$\|(P^1_v(x))'_+\|_2 = \frac{1}{\sqrt{n}} \leq 1. \tag{5.3}$$

To conclude the proof we use the following standard result.

LEMMA 5.3. *Let $f : \mathbb{R} \to \mathbb{R}^M$ and $\varphi : \mathbb{R} \to \mathbb{R}$ be two continuous and right-differentiable functions such that*

$$\|f'_+(v)\| \leq \varphi'_+(v)$$

for all $v \in \mathbb{R}$. Then

$$\|f(v) - f(w)\| \leq \varphi(v) - \varphi(w) \quad \text{for all } v \geq w.$$

According to the notation of this latter lemma, let us set $f(v) = P^1_v(x)$ and $\varphi(v) = v$. Since $P^1_v(x)$ is a continuous function with respect to v (in fact this is true for any projection onto convex sets; see [38]), by (5.3) and an application of the lemma we conclude that $\|P^1_v(x) - P^1_w(x)\| \leq |v - w|$. ∎

Observe that the strict convexity of $\Phi^{(q)}(u, v)$ is equivalent to $\theta_\lambda \omega_\lambda > \kappa_q/4$; see Proposition 2.1. In the following proposition we require the slightly stronger condition that $\theta_\lambda \omega_\lambda$ is bounded strictly away from $\kappa_q/4$.

PROPOSITION 5.4. *Let $q \in \{1, 2, \infty\}$. Assume that $\theta_\lambda \omega_\lambda \geq \sigma > \kappa_q/4$ for all $\lambda \in \Lambda$, where $\kappa_1 = M$ and $\kappa_2 = \kappa_\infty = 1$, implying that $\Phi^{(q)}(u, v)$ and $J(u, v)$ are strictly convex; see Proposition* 2.1. *Moreover, let us assume that $\omega_\lambda \geq \gamma > 0$ for all $\lambda \in \Lambda$. Suppose $\|T\| < 1$ resulting in $\|I - T^*T\| \leq 1$. Set*

$$\beta := \sup_{\lambda \in \Lambda} \frac{\kappa_q}{4\theta_\lambda \omega_\lambda + 4\theta_\lambda (1 - \|I - T^*T\|)} \leq \frac{\kappa_q}{4\sigma} < 1. \tag{5.4}$$

Then for each $n \in \mathbb{N}$ one has the following error estimate:

$$\|u^{(n,\infty)} - u^* | \ell_2(\Lambda, \mathbb{R}^M)\| \leq \beta \|u^{(n,0)} - u^* | \ell_2(\Lambda, \mathbb{R}^M)\|.$$

Proof. Let us consider the nth iteration of the outer loop. We have

$$u^{(n,\infty)} = U^{(q)}_{v^{(n)}, \omega}(\underbrace{u^{(n,\infty)} + T^*(g - Tu^{(n,\infty)})}_{:= y^{(n)}}).$$

By the weak convergence of the double-minimization algorithm, the minimum solution u^* also satisfies a similar relation:

$$u^* = U^{(q)}_{v^*,\omega}(\underbrace{u^* + T^*(g - Tu^*)}_{:=y^*}).$$

Recall that $U^{(q)}_{v,\omega}(y)_\lambda := (1+\omega_\lambda)^{-1} S^{(q)}_{v_\lambda}(y_\lambda)$. By nonexpansiveness of $S^{(q)}_v$ (Lemma 4.4) we have

$$\begin{aligned}
\|u^{(n,\infty)}_\lambda - u^*_\lambda\|_2 &\le \|U^{(q)}_{v^{(n)},\omega}(y^{(n)})_\lambda - U^{(q)}_{v^{(n)},\omega}(y^*)_\lambda\|_2 + \|U^{(q)}_{v^{(n)},\omega}(y^*)_\lambda - U^{(q)}_{v^*,\omega}(y^*)_\lambda\|_2 \\
&= (1+\omega_\lambda)^{-1}\|y^{(n)}_\lambda - y^*_\lambda\|_2 + \|U^{(q)}_{v^{(n)},\omega}(y^*)_\lambda - U^{(q)}_{v^*,\omega}(y^*)_\lambda\|_2 \\
&\le (1+\omega_\lambda)^{-1}\|I - T^*T\|\,\|u^{(n,\infty)}_\lambda - u^*_\lambda\|_2 \\
&\quad + \|U^{(q)}_{v^{(n)},\omega}(y^*)_\lambda - U^{(q)}_{v^*,\omega}(y^*)_\lambda\|_2.
\end{aligned}$$

This implies

$$\begin{aligned}
\|u^{(n,\infty)}_\lambda - u^*_\lambda\|_2 &\le \left(1 - (1+\omega_\lambda)^{-1}\|I - T^*T\|\right)^{-1} \|U^{(q)}_{v^{(n)},\omega}(y^*)_\lambda - U^{(q)}_{v^*,\omega}(y^*)_\lambda\|_2 \\
&= (1 + \omega_\lambda - \|I - T^*T\|)^{-1}\,\|S_{v^{(n)}_\lambda}(y^*_\lambda) - S_{v^*_\lambda}(y^*_\lambda)\|_2.
\end{aligned}$$

Recall from Lemma 4.1 that $S^{(q)}_v(x) = x - P^{q'}_{v/2}(x)$, where $P^{q'}_{v/2}$ denotes the orthogonal projection of x onto the $\ell_{q'}$ ball of radius $v/2$. By Lemma 5.2 we have that for any $z \in \mathbb{R}^M$

$$\|S^{(q)}_{v^{(n)}_\lambda}(z) - S^{(q)}_{v^*_\lambda}(z)\|_2 = \|P^{q'}_{v^*_\lambda/2}(z) - P^{q'}_{v^{(n)}_\lambda/2}(z)\|_2 \le \frac{L}{2}|v^{(n)}_\lambda - v^*_\lambda|.$$

So $S^{(q)}_v(z)$ is also Lipschitz in v. Let us recall that

$$v^{(n)}_\lambda := \begin{cases} \rho_\lambda - \frac{1}{2\theta_\lambda}\|u^{(n,0)}_\lambda\|_q & \|u^{(n,0)}_\lambda\|_q < 2\theta_\lambda\rho_\lambda, \\ 0 & \text{otherwise} \end{cases}$$

and

$$v^*_\lambda := \begin{cases} \rho_\lambda - \frac{1}{2\theta_\lambda}\|u^*_\lambda\|_q & \|u^*_\lambda\|_q < 2\theta_\lambda\rho_\lambda, \\ 0 & \text{otherwise.} \end{cases}$$

By distinguishing cases we can show that

$$|v^{(n)}_\lambda - v^*_\lambda| \le \frac{1}{2\theta_\lambda}\left|\|u^{(n,0)}_\lambda\|_q - \|u^*_\lambda\|_q\right| \le \frac{1}{2\theta_\lambda}\|u^{(n,0)}_\lambda - u^*_\lambda\|_q \le \frac{R}{2\theta_\lambda}\|u^{(n,0)}_\lambda - u^*_\lambda\|_2,$$

where $R = 1$ for $q \in \{2,\infty\}$ and $R = M^{1/2}$ for $q = 1$. Pasting the pieces together and using $RL = \kappa_q$ yields

$$\|u^{(n,\infty)}_\lambda - u^*_\lambda\|_2 \le \frac{\kappa_q}{4\theta_\lambda\,(\omega_\lambda + 1 - \|I - T^*T\|)}\|u^{(n,0)}_\lambda - u^*_\lambda\|_2 \le \beta\|u^{(n,0)}_\lambda - u^*_\lambda\|_2.$$

Summation over $\lambda \in \Lambda$ completes the proof. □

Let us combine the previous two results to obtain the error estimation for the finite algorithm, i.e., for $L_n < \infty$.

THEOREM 5.5. *Make the same assumptions as in Propositions* 5.1 *and* 5.4. *Choose* L_n *such that*

$$\delta_n := \left(\alpha^{L_n}(1+\beta)+\beta\right) \le \delta < 1 \quad \text{for all } n \in \mathbb{N}.$$

(This is possible since $\alpha, \beta < 1$.*) Then we have linear convergence of our algorithm, i.e.,*

$$\|u^{(n,0)} - u^*\|_2 \le \delta_n \|u^{(n-1,0)} - u^*\|_2.$$

Proof. Using Propositions 5.1 and 5.4 we get

$$\begin{aligned}
\|u^{(n,0)} - u^*\|_2 &\le \|u^{(n,0)} - u^{(n-1,\infty)}\|_2 + \|u^{(n-1,\infty)} - u^*\|_2 \\
&\le \alpha^{L_{n-1}} \|u^{(n-1,0)} - u^{(n-1,\infty)}\|_2 + \beta \|u^{(n-1,0)} - u^*\|_2 \\
&\le \alpha^{L_{n-1}} \left(\|u^{(n-1,0)} - u^*\|_2 + \|u^* - u^{(n-1,\infty)}\|_2 \right) + \beta \|u^{(n-1,0)} - u^*\|_2 \\
&\le \alpha^{L_{n-1}} \left(\|u^{(n-1,0)} - u^*\|_2 + \beta \|u^{(n-1,0)} - u^*\|_2 \right) + \beta \|u^{(n-1,0)} - u^*\|_2 \\
&\le \left(\alpha^{L_{n-1}}(1+\beta)+\beta\right) \|u^{(n-1,0)} - u^*\|_2.
\end{aligned}$$

This concludes the proof. ∎

Remark. The last theorem shows that it is possible to choose the number L_n of inner iterations constant with respect to n.

5.3. Strong convergence of the double-minimization algorithm. Finally, we can establish the strong convergence of the double-minimization algorithm and conclude the full proof of Theorem 3.1.

COROLLARY 5.6. *Under the assumptions of Proposition* 5.4, *if the minimizer of* $J(u, v^{(n)})$ *for fixed* $v^{(n)}$ *could be computed exactly, i.e.,* $L_n = \infty$ *for all* $n \in \mathbb{N}$, *then the outer loop converges with exponential rate, and we have*

$$\|u^{(n)} - u^*\|_2 \le \beta^n \|u^{(0)} - u^*\|_2,$$

where we have denoted $u^{(n)} := u^{(n,\infty)}$. *Moreover, the sequence* $v^{(n)}$ *converges componentwise, and* $v^{(n)} - v^*$ *converges to* 0 *strongly in* $\ell_{2,\theta}(\Lambda)$.

Proof. The first part of the statement is a direct application of Proposition 5.4. It remains to show that $v^{(n)} - v^*$ converges to 0 strongly in $\ell_{2,\theta}(\Lambda)$. Using that all norms on $\mathbb{R}^M$ are equivalent it follows that

$$\begin{aligned}
4 \sum_{\lambda \in \Lambda} \theta_\lambda^2 |v^{(n)} - v^*|^2 &= \sum_{\lambda \in \Lambda} |\|u_\lambda^*\|_q - \|u_\lambda^{(n)}\|_q|^2 \le \sum_{\lambda \in \Lambda} \|u_\lambda^* - u_\lambda^{(n)}\|_q^2 \\
&\le C \sum_{\lambda \in \Lambda} \|u_\lambda^* - u_\lambda^{(n)}\|_2^2 = C \|u^* - u^{(n)} | \ell_2(\Lambda, \mathbb{R}^M)\|^2.
\end{aligned}$$

Thus, $v^{(n)} - v^{(\infty)}$ converges also strongly in $\ell_{2,\theta}(\Lambda)$. ∎

6. Color image reconstruction. With this section we illustrate the application of the algorithms for color image recovery. The scope is to furnish a qualitative description of the behavior of the scheme. In a subsequent work we plan to provide a finer quantitative analysis in the context of *distributed compressed sensing* [3].

We begin by illustrating an interesting real-world problem occurring in art restoration. On March 11, 1944, a group of bombs launched from an Allied airplane hit the famous Italian Eremitani's Church in Padua, destroying it together with the inestimable frescoes by Andrea Mantegna et al. contained in the Ovetari Chapel. Details on "the state of the art" can be found in [29, 28]. In 1920 a collection of high quality gray level pictures of these frescoes was made by Alinari. The only color images of the frescoes are dated to 1940, but unfortunately their quality (i.e., the intrinsic resolution of the printouts) is much lower; see Figure 6.6. Inspired by the fresco application, we model the problem of the recovery of a high-resolution color image from a low-resolution color datum and a high-resolution gray datum. We will implement the solution to the model problem as a nontrivial application of the algorithms we have presented in this paper.

6.1. Color images, curvelets, and joint sparsity. Let us assume that the color images are encoded into YIQ channels. The Y component represents the luminance information (gray level), while I and Q give the chrominance information. Of course, one may also choose a different encoding system, e.g., RGB or CMYK. Clearly, the color image $f = (f_1, f_2, f_3)$ can be represented as a 3-channel signal. In order to apply our algorithm, we need to fix a frame for which we can assume color images being jointly sparse.

It is well known that *curvelets* [8] are well suited for sparse approximations of curved singularities. A natural image can in fact be modeled as a function which is piecewise smooth except on a discontinuity set, the latter being described as the union of rectifiable curves. Moreover, there are fast algorithms available for the computation of curvelet coefficients of digital images [7].

In the following, let us assume that a color image f is encoded into a vector of curvelet coefficients $(u_\lambda^\ell)_{\lambda\in\Lambda}^{\ell=1,2,3}$. The image can be reconstructed by the synthesis formula

$$f \;=\; (Fu^\ell)_{\ell=1,2,3} := \left(\sum_{\lambda\in\Lambda} u_\lambda^\ell \psi_\lambda\right)_{\ell=1,2,3},$$

where $\{\psi_\lambda : \lambda \in \Lambda\}$ is the collection of curvelets. The index λ consists of 3 different parameters: $\lambda = (j, p, k)$, where j corresponds to scale, p to a rotation, and k to the spatial location of the curvelet ψ_λ. We do not enter into further details, especially of the discrete and numerical implementation, which one can find in [8, 7].

Let us instead observe that significant curvelet coefficients $u_\lambda = (u_\lambda^1, u_\lambda^2, u_\lambda^3)$ will appear simultaneously at the same $\lambda \in \Lambda$ for all of the channels, as soon as the corresponding curvelet overlaps with a (curved) singularity (appearing simultaneously in all of the channels) and is approximately tangent to it. This justifies the joint sparsity assumption for color images with respect to curvelets.

6.2. The model of the problem. The datum of our problem is a 3-channel signal $g = (g_1, g_2, g_3) \in \ell_2(\mathbb{Z}_{N_0}^2, \mathbb{R}^3)$, where g_i, $i = 2, 3$, are the low-resolution chrominance channels I and Q, and g_1 is the high-resolution gray channel Y. We assume that g was produced by $g = Tu$, where $u = (u^1, u^2, u^3)$ are the curvelet coefficients of the three channels of the high-resolution color image that we want to reconstruct. The

operator $T = (T_{\ell,j})_{\ell,j=1,2,3}$ can be expressed by the matrix

$$T = \begin{pmatrix} F & 0 & 0 \\ 0 & AF & 0 \\ 0 & 0 & AF \end{pmatrix}. \tag{6.1}$$

Here A is the linear operator that transforms the high-resolution image into the low-resolution image. In particular, A can be taken as a convolution operator (with a Gaussian, for instance) followed by downsampling. Eventually, we may assume a suitable scaling in order to make $\|T\| < 1$ and a different weighting of the gray channel and the I,Q channels in the discrepancy term. Since A is not invertible, the operator T is also not invertible, and the minimization of $\mathcal{T}(u)$ requires a regularization. Clearly, for this task we use the functional K defined in (2.4) or J defined in (2.5).

6.3. On the choice of the parameters. What remains to clarify is the choice of the parameters ω_λ, θ_λ, and ρ_λ. The parameter $\omega_\lambda \geq \gamma > 0$ has been introduced for the sole purpose of making J strictly convex. A large value of this parameter actually produces an image u which is significantly blurred, and no information about edges is recovered. Thus, we rather put $\omega_\lambda = \gamma = \varepsilon > 0$ small. Due to the convexity requirements (see section 2.3), we select $\theta_\lambda \sim \frac{M}{\varepsilon}$. The choice of ρ_λ requires a deeper understanding of the information encoded by the curvelet coefficients.

Indeed, in [8] it was observed that those curvelets that overlap with a discontinuity decay like $\|u_\lambda\|_q \lesssim 2^{-3/4j}$, while the others satisfy $\|u_\lambda\|_q \lesssim 2^{-3/2j}$ (where j denotes the scale). Since we want to recover joint discontinuities, we may choose $\rho_\lambda := \rho_{j,p,k} \sim 2^{-js}$, with $s \in [3/4, 3/2]$. By this choice and by (3.2) the locations λ for which $v_\lambda = 0$ will indicate a potential joint discontinuity.

Of course, this is just one possible choice of the parameters, and further information might be extracted from the joint sparsity pattern indicated by v, by the use of different parameters. We believe that a deeper study of the characterization of the morphological properties of signals encoded by frames (e.g., curvelets and wavelets) is fundamental for the right choice of these parameters. We refer to [31, 32] for deeper insights in this direction, concerning fine properties of functions encoded by the distribution of wavelet coefficients.

6.4. Numerical experiments. According to the previous subsections, we illustrate here the application of **JOINTSPARSE** for the recovery of a high-resolution color image from a low-resolution color datum and a high-resolution gray datum. In Figure 6.1 we illustrate the data of the problem. In this case the resolution of the color image has been reduced by a factor of 4 in each direction by using a Gaussian filter and a downsampling. We have conducted several experiments for different choices of $q \in \{1, 2, \infty\}$, with fixed parameters as indicated in section 6.3. We have chosen $L_n = 105$ and $n_{\max} = 1$, as well as $L_n = 7$ and $n_{\max} = 15$ (the numbers of inner and outer iterations, respectively). In the first case, only the minimization of $J(u, v^{(0)})$ with respect to u has been performed; i.e., no iterative adaptation of the joint sparsity pattern indicated by v occurred. In order to estimate the different behavior depending on the parameters above, we have evaluated at each iteration the ℓ_2 error between the reconstructed I and Q color channels and the *original* I and Q color channels. Figures 6.2 and 6.3 indicate that the error decreases for increasing values of q. This means that the increased coupling due to the q parameter is significant in order to improve the recovery. Recall that the choice $q = 1$ does not induce any coupling between channels.

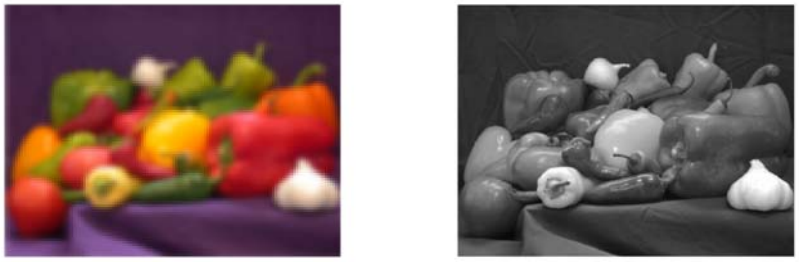


FIG. 6.1. Left: *The low-resolution color image is here presented after Gaussian filtering and downsampling. In the numerical experiments the I and Q channels are used.* Right: *The high-resolution gray level image encodes several morphological information useful to recover the high-resolution color image.*

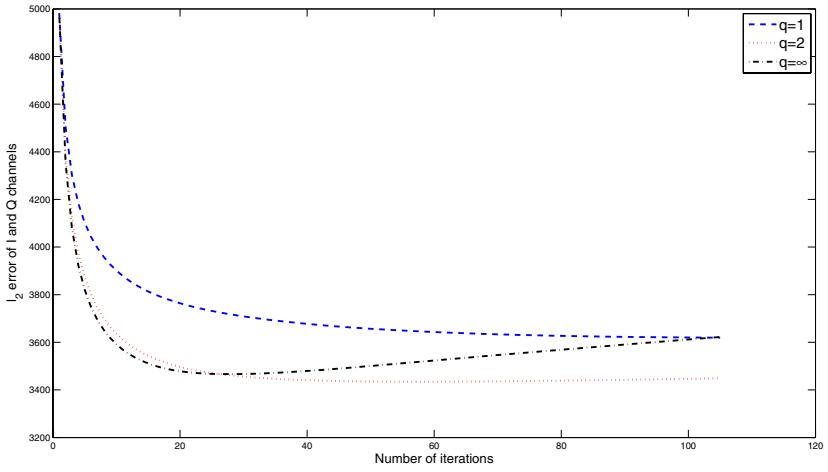


FIG. 6.2. *The ℓ_2 error between the* original color image *and the iterations of the algorithm is shown for different values of $q = 1, 2, \infty$. We have considered $L_n = 105$ and $n_{\max} = 1$ the numbers of inner and outer iterations, respectively. We have fixed here $\omega_\lambda = 0$, $\theta_\lambda = 10$, and $\rho_\lambda = 20 \times 2^{-j}$.*

This coupling effect due to $q > 1$ is even more evident in Figure 6.3, where the adaptation of the weight v occurs. The left and the central pictures in the second row of Figure 6.4 show a reduced color distortion at edges, passing from the case $q = 1$ (without coupling) to the case $q = \infty$, respectively, and consequently a better edge resolution. Nevertheless, the differences are not so remarkable. This is due to the fact that, although the functional J promotes coupling at edges, it does not necessarily enforce a significant edge enhancement. Thus, we may modify the functional by adding an additional total variation (TV) constraint on the I and Q channels:

$$J_{\mathrm{TV}}(u, v) := J(u, v) + \left(|Fu^2|_{\mathrm{TV}} + |Fu^3|_{\mathrm{TV}}\right).$$

The effect of this modification is to promote edge enhancing together with their simultaneous coupling through different channels. For the minimization of J_{TV} we use a heuristic scheme as in [25], by alternating iterations for the minimization of J and for

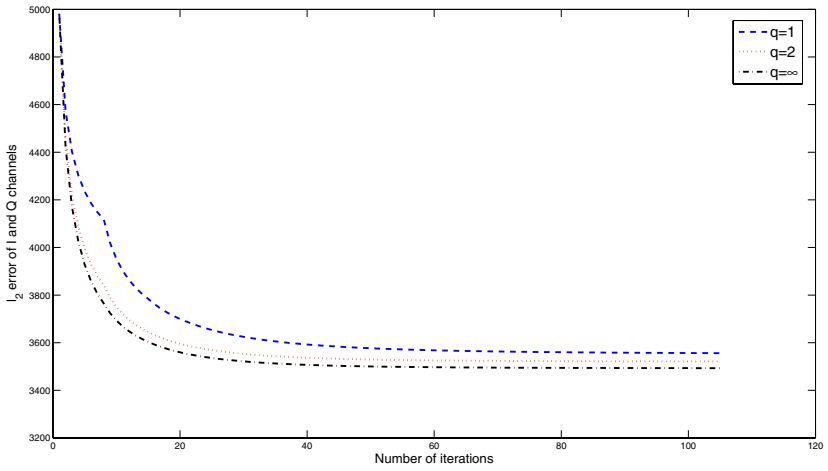


FIG. 6.3. *The ℓ_2 error between the* original color image *and the iterations of the algorithm is shown for different values of $q = 1, 2, \infty$. We have considered $L_n = 7$ and $n_{\max} = 15$ the numbers of inner and outer iterations, respectively. We have fixed here $\omega_\lambda = 1/20$, $\theta_\lambda = 10$, and $\rho_\lambda = 20 \times 2^{-j}$.*

FIG. 6.4. First row. Left: *Portion of the low-resolution color image.* Center: *Portion of the reconstructed color image by Gaussian interpolation and substitution of the Y channel with the gray level datum. Evident color artifacts appear at edges.* Right: *Portion of the original color image.* Second row. Left: *Portion of the reconstructed color image for $q = 1$, $L_n = 105$, and $n_{\max} = 1$.* Center: *Portion of the reconstructed color image for $q = \infty$, $L_n = 7$, and $n_{\max} = 15$.* Right: *Portion of the reconstructed color image for $q = \infty$, $L_n = 7$, $n_{\max} = 15$, and TV minimization.*

the minimization of $\left(|Fu^2|_{\mathrm{TV}} + |Fu^3|_{\mathrm{TV}}\right)$; compare also [19]. The corresponding results are shown in Figure 6.5, where the effect of the coupling (for the cases $q = 2, \infty$) is further enhanced. The right picture in the second row of Figure 6.4 shows the result of the reconstruction in this latter case. The edges are perfectly recovered.

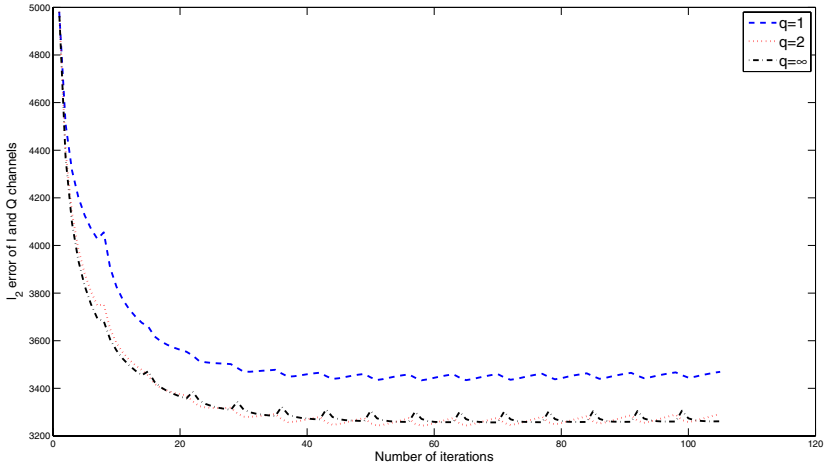


FIG. 6.5. *The ℓ_2 error between the* original color image *and the iterations of the algorithm is shown for different values of $q = 1, 2, \infty$. We have considered $L_n = 7$ and $n_{\max} = 15$ the numbers of inner and outer iterations, respectively. We have fixed here $\omega_\lambda = 1/20$, $\theta_\lambda = 10$, and $\rho_\lambda = 20 \times 2^{-j}$. Further iterations of TV minimization are added in the outer loop to enforce edge enhancing.*

FIG. 6.6. Left: *Low quality color image of the fresco dated to* 1940. Center: *High quality gray image of the fresco dated to* 1920. *Some details are not visible in the color version.* Right: *The reconstructed image after* 6 *outer iterations with* 7 *inner iterations each, for $q = \infty$. The final Y channel is substituted with the high-resolution gray level datum. The discontinuities are enhanced, and no artifact colors appear.*

These numerical experiments confirm that the use of the joint sparsity measure $\Phi^{(q)}$ associated to the curvelet representation can improve significantly the quality of the reconstructed color image. Better results are achieved by choosing $q = \infty$ and by the adaptive choice of weights as indicators of the sparsity pattern. Further improvements can be achieved by channelwise edge enhancing, e.g., via total variation minimization. An application to the real case of the art frescoes is illustrated in Figure 6.6.

7. Final remarks. 1. If the index set Λ is infinite, then T^*T is represented as a biinfinite matrix, and thus its evaluation might not be exactly numerically implementable. In a subsequent work we will consider the case $\#\Lambda = \infty$ and the treatment

of sparse (approximate) evaluations of bi-infinite matrices in order to realize fast and convergent schemes also in this situation; compare also [39, 14, 15].

2. In the subsequent paper [27] we investigated the role of the various parameters. To exploit the optimal performance of the scheme, an extensive campaign of numerical experiments should be conducted. It is also crucial to investigate the deeper relations among the parameter ρ_λ, the *multifractal analysis* as, e.g., in [31], and morphological image analysis. In particular, the parallel between the functional J and the Γ-approximation of the Mumford–Shah functional by Ambrosio and Tortorelli [1, 5] is suggestive:

$$F_\varepsilon(u,v) := \underbrace{\int_\Omega (u-g)^2 dx}_{\sim \mathcal{T}(u)} + \underbrace{\int_\Omega v^2 f(\nabla u) dx}_{\sim \sum_\lambda v_\lambda \|u\|_q} + \underbrace{\int_\Omega \varepsilon |\nabla u|^2 dx}_{\sim \sum_\lambda \omega_\lambda \|u\|_2^2}$$
$$+ \underbrace{\int_\Omega \left(\varepsilon |\nabla v|^2 + \frac{1}{4\varepsilon}(1-v)^2 \right) dx}_{\sim \sum_\lambda \theta_\lambda (\rho_\lambda - v_\lambda)^2},$$

where f is a suitable polyconvex function, e.g., $f(\nabla u) = (|\nabla u|^2 + |u_x \times u_y|) = (|\nabla u|^2 + |\mathrm{adj}_2(\nabla u)|)$, where $\mathrm{adj}_2(A)$ is the matrix of all 2×2 minors of A. The minimization of this term enforces the fact that derivatives of different channels are large only in the same directions.

3. While we were finishing this paper, we were informed by Teschke of the results in [19]. In that paper the authors consider linear inverse problems where the solution is assumed to fulfill some general 1-homogeneous convex constraint. They develop an algorithm that amounts to a projected Landweber iteration and that provides an iterative approach to the solution of this inverse problem. In particular, for the case $\omega = (\omega_\lambda)_\lambda = 0$, some of our results stated in section 4 can be reformulated in this more general setting and therefore derived from [19]. However, for $\omega \neq 0$ the sparsity measure $\Psi^{(q)}_{v,\omega}$ as in (4.2) is not 1-homogeneous, and the elaborations in section 4 are needed. Moreover, for the relevant cases $q = 1, 2, \infty$, we express explicitly the projection $P^{q'}_{v/2}$. Due to their generality, the results in [19] do not provide concrete recipes to compute such projections.

8. Conclusion. We have investigated joint sparsity measures with respect to frame expansions of vector-valued functions. These sparsity measures generalize approaches valid for scalar functions and take into account common sparsity patterns through different channels. We have analyzed linear inverse problems with joint sparsity regularization as well as their efficient numerical solution by means of a novel algorithm based on thresholded Landweber iterations. We have provided the convergence analysis for a wide range of parameters. The role of the joint sparsity measure is twofold: to tighten the characterization of solutions of interest and to extract significant morphological properties which are a common feature of all of the channels. By numerical applications in color image restoration, we have shown that joint sparsity significantly outperforms uncoupled constraints. We have presented the results of an application to a relevant real-world problem in art restoration. The wide range of applicability of our approach includes several other problems with coupled vector-valued solutions, e.g., neuroimaging and distributed compressed sensing.

Acknowledgment. Massimo Fornasier thanks NuHAG for its warm hospitality.

REFERENCES

[1] L. Ambrosio and V. M. Tortorelli, *Approximation of functionals depending on jumps by elliptic functionals via* Γ*-convergence*, Comm. Pure Appl. Math., 43 (1990), pp. 999–1036.

[2] S. Anthoine, *Different Wavelet-based Approaches for the Separation of Noisy and Blurred Mixtures of Components. Application to Astrophysical Data*, Ph.D. thesis, Princeton University, Princeton, NJ, 2005.

[3] D. Baron, M. B. Wakin, M. F. Duarte, S. Sarvotham, and R. G. Baraniuk, *Distributed Compressed Sensing*, preprint, 2005.

[4] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, Cambridge, 2004.

[5] A. Brook, R. Kimmel, and N. A. Sochen, *Variational restoration and edge detection for color images*, J. Math. Imaging Vision, 18 (2003), pp. 247–268.

[6] E. Candès and T. Tao, *Near optimal signal recovery from random projections: Universal encoding strategies?*, IEEE Trans. Inf. Theory, 52 (2006), pp. 5406–5425.

[7] E. J. Candès, L. Demanet, D. L. Donoho, and L. Ying, *Fast Discrete Curvelet Transforms*, 2005.

[8] E. J. Candès and D. L. Donoho, *New tight frames of curvelets and optimal representations of objects with piecewise* C^2 *singularities*, Comm. Pure Appl. Math., 57 (2004), pp. 219–266.

[9] E. J. Candès, J. Romberg, and T. Tao, *Exact signal reconstruction from highly incomplete frequency information*, IEEE Trans. Inform. Theory, 52 (2006), pp. 489–509.

[10] C. Canuto and K. Urban, *Adaptive optimization of convex functionals in Banach spaces*, SIAM J. Numer. Anal., 42 (2005), pp. 2043–2075.

[11] O. Christensen, *An Introduction to Frames and Riesz Bases*, Birkhäuser, Boston, 2003.

[12] A. Cohen, *Numerical Analysis of Wavelet Methods*, Stud. Math. Appl., 32, North-Holland, Amsterdam, 2003.

[13] A. Cohen, M. Hoffmann, and M. Reiss, *Adaptive wavelet Galerkin methods for linear inverse problems*, SIAM J. Numer. Anal., 42 (2004), pp. 1479–1501.

[14] S. Dahlke, M. Fornasier, and T. Raasch, *Adaptive frame methods for elliptic operator equations*, Adv. Comput. Math., 27 (2007), pp. 27–63.

[15] S. Dahlke, M. Fornasier, T. Raasch, R. Stevenson, and M. Werner, *Adaptive frame methods for elliptic operator equations: The steepest descent approach*, IMA J. Numer. Anal., 27 (2007), pp. 717–740.

[16] S. Dahlke and P. Maass, *An outline of adaptive wavelet Galerkin methods for Tikhonov regularization of inverse parabolic problems*, Recent Development in Theories and Numerics. Proceedings of the International Conference on Inverse Problems, Y.-C. Hon et al., eds., Hong Kong, China, 2002, World Scientific, River Edge, NJ, pp. 56–66, 2003.

[17] I. Daubechies, M. Defrise, and C. DeMol, *An iterative thresholding algorithm for linear inverse problems*, Comm. Pure Appl. Math., 57 (2004), pp. 1413–1457.

[18] I. Daubechies and G. Teschke, *Variational image restoration by means of wavelets: Simultaneous decomposition, deblurring, and denoising*, Appl. Comput. Harmon. Anal., 19 (2005), pp. 1–16.

[19] I. Daubechies, G. Teschke, and L. Vese, *Iteratively solving linear inverse problems under general convex constraints*, Inverse Probl. Imaging, 1 (2007), pp. 29–46.

[20] D. L. Donoho, *Superresolution via sparsity constraints*, SIAM J. Math. Anal., 23 (1992), pp. 1309–1331.

[21] D. L. Donoho, *De-noising by soft-thresholding*, IEEE Trans. Inform. Theory, 41 (1995), pp. 613–627.

[22] D. L. Donoho, *Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition*, Appl. Comput. Harmon. Anal., 2 (1995), pp. 101–126.

[23] D. L. Donoho, *Compressed sensing*, IEEE Trans. Inform. Theory, 52 (2006), pp. 1289–1306.

[24] I. Ekeland and R. Témam, *Convex Analysis and Variational Problems*, SIAM, Philadelphia, 1999.

[25] M. Elad, J.-L. Starck, P. Querre, and D. L. Donoho, *Simultaneous cartoon and texture image inpainting using morphological component analysis (MCA)*, Appl. Comput. Harmon. Anal., 19 (2005), pp. 340–358.

[26] H. W. Engl, M. Hanke, and A. Neubauer, *Regularization of Inverse Problems*, Math. Appl. 375, Kluwer Academic Publishers, Dordrecht, 1996.

[27] M. Fornasier and H. Rauhut, *Iterative thresholding algorithms*, Appl. Comput. Harmon. Anal., to appear.

[28] M. Fornasier and D. Toniolo, *Computer-based recomposition of the frescoes in the Ovetari Chapel in the Church of the Eremitani in Padua. Methodology and initial results*, in Mantegna nella chiesa degli Eremitani a Padova. Il recupero possibile, Skira, ed., 2003 (English/Italian).

[29] M. Fornasier and D. Toniolo, *Fast, robust, and efficient* 2*D pattern recognition for reassembling fragmented digital images*, Pattern Recognition, 38 (2005), pp. 2074–2087.

[30] A. C. Gilbert, M. J. Strauss, and J. Tropp, *Algorithms for simultaneous sparse approximation. Part* I: *Greedy pursuit*, Signal Processing, 86 (2006), pp. 572–588.

[31] S. Jaffard, *Beyond Besov spaces.* I: *Distributions of wavelet coefficients*, J. Fourier Anal. Appl., 10 (2004), pp. 221–246.

[32] S. Jaffard, *Beyond Besov spaces.* II: *Oscillation spaces*, Constr. Approx., 21 (2005), pp. 29–61.

[33] S. Kunis and H. Rauhut, *Random sampling of sparse trigonometric polynomials II—orthogonal matching pursuit versus basis pursuit*, Found. Comput. Math., to appear.

[34] S. Mallat, *A Wavelet Tour of Signal Processing*, 2nd ed., Academic Press, San Diego, CA, 1999.

[35] D. Mumford and J. Shah, *Optimal approximations by piecewise smooth functions and associated variational problems*, Comm. Pure Appl. Math., 42 (1989), pp. 577–685.

[36] R. Ramlau and G. Teschke, *Tikhonov replacement functionals for iteratively solving nonlinear operator equations*, Inverse Problems, 21 (2005), pp. 1571–1592.

[37] H. Rauhut, *Random sampling of sparse trigonometric polynomials*, Appl. Comput. Harmon. Anal., 22 (2007), pp. 16–42.

[38] R. T. Rockafellar and R. J. B. Wets, *Variational Analysis*, Grundlehren Math. Wiss., 317, Springer-Verlag, Berlin, 1998.

[39] R. Stevenson, *Adaptive solution of operator equations using wavelet frames*, SIAM J. Numer. Anal., 41 (2003), pp. 1074–1100.

[40] G. Teschke, *Multi-frame representations in linear inverse problems with mixed multiconstraints*, Appl. Comput. Harmon. Anal., 22 (2007), pp. 43–60.

[41] J. Tropp, *Algorithms for simultaneous sparse approximation. Part* II: *Convex relaxation*, Signal Processing, 86 (2006), pp. 589–602.

SIAM J. NUMER. ANAL.
Vol. 46, No. 2, pp. 614–638
© 2008 Society for Industrial and Applied Mathematics

LINEAR CONVERGENCE OF AN ADAPTIVE FINITE ELEMENT METHOD FOR THE p-LAPLACIAN EQUATION*

LARS DIENING† AND CHRISTIAN KREUZER‡

Abstract. We study an adaptive finite element method for the p-Laplacian like PDEs using piecewise linear, continuous functions. The error is measured by means of the quasi norm of Barrett and Liu. We provide residual based error estimators without a gap between the upper and lower bound. We show linear convergence of the algorithm which is similar to the one of Morin, Nochetto, and Siebert. All results are obtained without extra marking for the oscillation.

Key words. convergence, a posteriori error estimator, p-Laplacian, quasi norms, error reduction, oscillation

AMS subject classifications. 65N15, 65N30, 35J60, 35J70

DOI. 10.1137/070681508

1. Introduction.

Let Ω be a polyhedral, bounded domain in $\mathbb{R}^d$, $d \in \mathbb{N}$. We consider the following system of nonlinear structure:

$$\begin{aligned} -\operatorname{div}(\mathbf{A}(\nabla u)) &= f \qquad \text{in } \Omega, \\ u &= 0 \qquad \text{on } \partial\Omega. \end{aligned} \tag{1.1}$$

Our considerations include in particular the case of the p-Laplacian, where

$$\mathbf{A}(\nabla u) = (\kappa + |\nabla u|)^{p-2}\, \nabla u, \tag{1.2}$$

with $1 < p < \infty$, $\kappa \geq 0$, $f \in L^{p'}(\Omega)$, and $\frac{1}{p} + \frac{1}{p'} = 1$.

The purpose of this paper is to present a linear convergence result for an adaptive finite element method (AFEM) applied to the nonlinear Laplace equation (1.1). As is common practice the AFEM consists of a loop

$$\text{SOLVE} \to \text{ESTIMATE} \to \text{MARK} \to \text{REFINE} \tag{AFEM}$$

starting from a initial triangulation of Ω. To be more specific, the finite element problem on the current mesh is solved, then the a posteriori error estimator is computed, and finally with its help elements are marked for refinement. The marking strategy resorts to *Dörfler's marking* introduced in [Dör96]. The algorithm uses piecewise linear, continuous finite elements, whereas the refinement is realized by newest vertex bisection. This produces a sequence of weak finite element solutions u_k of (1.1) in nested finite element spaces V_k.

*Received by the editors February 20, 2007; accepted for publication (in revised form) August 6, 2007; published electronically February 1, 2008.
http://www.siam.org/journals/sinum/46-2/68150.html

†Abteilung für angewandte Mathematik, Universität Freiburg, Eckerstr. 1, 79104 Freiburg i. Br., Germany (diening@mathematik.uni-freiburg.de, http://www.mathematik.uni-freiburg.de/IAM/homepages/diening). This author's research was supported in part by the Landesstiftung Baden-Württemberg.

‡Institut für Mathematik, Universität Augsburg, Universitätsraße 14, 86159 Augsburg, Germany (kreuzer@math.uni-augsburg.de, http://www.math.uni-augsburg.de/de/prof/lam/mitarbeiter/christian_kreuzer). This author's research was supported by Project C.1 of the DFG-Research-Unit "Nonlinear Partial Differential Equations": Generalized Newtonian Fluids and Electrorheological Fluids.

The main result states linear convergence of u_k to the weak solution u of (1.1). In particular, we show that there exists $\alpha \in (0,1)$, $C > 0$ with

$$\|\mathbf{F}(\nabla u_k) - \mathbf{F}(\nabla u)\|_2^2 + \operatorname{osc}_k^2(f) \leq \alpha^{2k} C,$$

where the vector field $\mathbf{F}$ arises from the vector field $\mathbf{A}$ by $\mathbf{F}(\mathbf{a}) := |\mathbf{A}(\mathbf{a})|^{\frac{1}{2}} |\mathbf{a}|^{-\frac{1}{2}} \mathbf{a}$. The L^2 norm of the error $\|\mathbf{F}(\nabla u_k) - \mathbf{F}(\nabla u)\|_2^2$ measured in terms of $\mathbf{F}$ is equivalent to the so-called *quasi norm* $\|\nabla u_k - \nabla u\|_{(p)}^2$ introduced by Barrett and Liu; cf. [BL94a, BL94b] and Remark 5. The *quasi norm* was a breakthrough in the numerical investigation of (1.1). In particular, Barrett and Liu obtained the best approximation property of the conforming, finite element solution $u_h \in V_h$ in terms of quasi norms, i.e.,

$$\|\nabla v - \nabla v_h\|_{(p)}^2 \leq c \min_{\psi_h \in V_h} \|\nabla v - \nabla \psi_h\|_{(p)}^2.$$

In [EL05] it has been proved by Ebmeyer and Liu that for piecewise linear, continuous finite elements and $p > \frac{2d}{d+2}$ the best approximation error can be estimated as

$$\min_{\psi_h \in V_h} \|\nabla v - \nabla \psi_h\|_{(p)}^2 \leq c\, h^2 \int_\Omega (\kappa + |\nabla u|)^{p-2} \left|\nabla^2 u\right|^2 dx. \tag{1.3}$$

Recently, Diening and Růžička improved these results in [DR07] to the case $p > 1$, admitting also more general finite element spaces. In particular, they showed that

$$\|\mathbf{F}(\nabla v) - \mathbf{F}(\nabla v_h)\|_2^2 \leq c \min_{\psi_h \in V_h} \|\mathbf{F}(\nabla v) - \mathbf{F}(\nabla \psi_h)\|_2^2$$

and

$$\|\mathbf{F}(\nabla v) - \mathbf{F}(\nabla \Pi_h v)\|_2^2 \leq c\, h^2 \|\nabla \mathbf{F}(\nabla u)\|_2^2, \tag{1.4}$$

where Π_h is a suitable interpolation operator, e.g., the Scott–Zhang operator. We want to mention that the right-hand sides of (1.3) and (1.4) are proportional. They express the natural regularity of a *strong* solution of (1.1) (cf. [Giu03], [BL93a], [ELS05], [Ebm05]).

The technique of quasi norms found its way into a posteriori analysis in the work of Liu and Yan [LY01, LY02]. They show that

$$c\, \eta_h^2 - C \operatorname{osc}_h^2(f) \leq \|\nabla u - \nabla u_h\|_{(p)}^2 \leq C \left(\eta_h^2 + \widetilde{\eta}^2\right)$$

for residual based estimators η_h. Numerical experiments [LY01, CK06] indicate that these new estimators are indeed sharper and lead to more efficient meshes than existing ones. However, for convergence analysis the additional term $\widetilde{\eta}^2$ causes problems, since it forms a gap between the left- and the right-hand side. In this work we are able to overcome this drawback and prove estimates avoiding $\widetilde{\eta}^2$ (see Lemma 8 and Corollary 11):

$$c\, \eta_h^2 - C \operatorname{osc}_h^2(f) \leq \|\mathbf{F}(\nabla u) - \mathbf{F}(\nabla u_h)\|_2^2 \leq C\, \eta_h^2. \tag{1.5}$$

Recently a posteriori error estimators by gradient recovery have been studied in [CLY06].

Dörfler [Dör96] was the first to prove linear error reduction of (AFEM) for the linear Laplacian if the data oscillation is small enough. Later, this additional assumption has been removed by Morin, Nochetto, and Siebert in [MNS00] by additional marking for oscillation. The first convergence result for the nonlinear Laplacian is stated by Veeser [Ve02]. There residual based estimators for the $W^{1,p}$ norm are used. Since a gap appears in the power between the upper and lower estimates this prevents proving linear convergence.

The convergence results in the linear case are heavily based on Galerkin orthogonality and the Pythagorean theorem, which yield

$$|||u_h - u|||^2 = |||u_H - u|||^2 - |||u_H - u_h|||^2 \tag{1.6}$$

in the energy norm. To overcome the lack of orthogonality in the nonlinear case we proceed as follows: We prove that the energy difference of weak solutions in nested spaces $V_1 \subset V_2$ is proportional to the quasi-norm distance, i.e.,

$$\mathcal{J}(u_1) - \mathcal{J}(u_2) \sim \|\nabla u_1 - \nabla u_2\|_{(p)}^2 \sim \|\mathbf{F}(\nabla u_1) - \mathbf{F}(\nabla u_2)\|_2^2,$$

where $\mathcal{J}(u)$ is the energy functional of (1.1), $u_1 \in V_1$, and $u_2 \in V_2$; see (2.10) and Lemma 16. This property and the trivial equality

$$\mathcal{J}(u_h) - \mathcal{J}(u) = \big(\mathcal{J}(u_H) - \mathcal{J}(u)\big) - \big(\mathcal{J}(u_H) - \mathcal{J}(u_h)\big)$$

are our substitute for the orthogonality of the error (1.6).

In the linear, symmetric case it is possible to consider the reduction of the error and the oscillation independently, since the oscillation is solely dependent on the data f. Mekchay and Nochetto showed linear reduction of the sum of error and oscillation for nonsymmetric second order linear elliptic PDEs in [MN05]. In this case oscillation and error are coupled. A similar effect appears in our nonlinear setting. We introduce a new proof for error reduction, which enables us to manage without extra marking for oscillation; see Remark 14. Our proof permits us to use the fact that oscillation is dominated by the error indicator. Moreover, we prove a strict reduction of the difference of energies plus the oscillation in each step.

An essential tool in our calculations is the use of *shifted* N-functions, namely φ_a. They are closely related to the quasi norms, which are best expressed by the relation

$$\big(\mathbf{A}(\mathbf{a}) - \mathbf{A}(\mathbf{b})\big) \cdot (\mathbf{a} - \mathbf{b}) \sim \big|\mathbf{F}(\mathbf{a}) - \mathbf{F}(\mathbf{b})\big|^2 \sim \varphi_{|\mathbf{a}|}(|\mathbf{a} - \mathbf{b}|)$$

for $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$. See Lemma 3 for more details. The shifted N-functions enable us to handle more general nonlinear equations than the p-Laplacian, namely the φ-Laplacian from (2.1). But most important, the shifted N-functions simplify and clarify the calculations significantly also in the case of the p-Laplacian.

2. Preliminaries. We first introduce the nonlinear Dirichlet problem. Thereby the nonlinear partial differential operator called φ-Laplacian is defined via a certain function $\varphi : \mathbb{R}^{\geq 0} \to \mathbb{R}^{\geq 0}$. The most popular case of such operators is the p-Laplacian which corresponds to the function $\varphi(t) := \int_0^t (\kappa + s)^{p-2} s \, ds$. As mentioned before, the treatment of the nonlinear Laplacian via *N-functions* simplifies and clarifies calculations. Assumptions on φ and related properties are discussed subsequently. Afterwards the weak formulation of the problem is stated along with the corresponding minimizing problem.

Let Ω be a polyhedral, bounded domain in $\mathbb{R}^d$, $d \in \mathbb{N}$. In the center of our considerations are solutions of the φ-Laplacian problem, i.e.,

$$\begin{aligned} -\operatorname{div}(\mathbf{A}(\nabla u)) &= f \qquad \text{in } \Omega, \\ u &= 0 \qquad \text{on } \partial\Omega, \end{aligned} \tag{2.1}$$

with

$$\mathbf{A}(\nabla u) = \varphi'(|\nabla u|)\frac{\nabla u}{|\nabla u|}. \tag{2.2}$$

2.1. Assumptions on φ and resulting properties. Now we shed light on the considered function φ. To not go beyond the scope of this work we give only a short sketch of the underlying theory. The following definitions and results are standard in the theory of Orlicz functions and can be found, for example, in [RR91]. A few assertions are also proved in the appendix.

We use $c, C > 0$ (no index) as generic constants; i.e., their value may change from line to line but does not depend on the important variables. Furthermore, we write $f \sim g$ iff $c\,f \le g \le C\,f$.

A continuous function $\varphi : \mathbb{R}^{\ge 0} \to \mathbb{R}^{\ge 0}$ is said to be an *N-function* iff

- φ is continuous and convex with $\varphi(0) = 0$;
- there exists a derivative φ' of φ which is right continuous, nondecreasing, and satisfies $\varphi'(0) = 0$, $\varphi'(t) > 0$ for $t > 0$, and $\lim_{t\to\infty} \varphi'(t) = \infty$.

The N in N-function stands for *nice* function. Furthermore, φ satisfies the Δ_2-condition iff it holds that

- $\varphi(2t) \le C\,\varphi(t)$ uniformly in $t \ge 0$.

We denote the smallest such constant by $\Delta_2(\varphi)$. Since $\varphi(t) \le \varphi(2t)$ the Δ_2-condition means that $\varphi(t)$ and $\varphi(2t)$ are proportional. Note that if $\Delta_2(\varphi) < \infty$, then $\varphi(t) \sim \varphi(a\,t)$ uniformly in $t \ge 0$ for any fixed $a > 1$.

Define $(\varphi')^{-1} : \mathbb{R}^{\ge 0} \to \mathbb{R}^{\ge 0}$ as

$$(\varphi')^{-1}(t) := \sup\{u \in \mathbb{R}^{\ge 0} : \varphi'(u) \le t\}.$$

If φ' is strictly increasing, then $(\varphi')^{-1}$ is the inverse function of φ'. By the definition

$$\varphi^*(t) := \int_0^t (\varphi')^{-1}(s)\,ds$$

the so-called complementary function φ^* of φ is again an N-function and $(\varphi^*)'(t) = (\varphi')^{-1}(t)$ for $t > 0$. Note that $(\varphi^*)^* = \varphi$.

Assume that $\Delta_2(\varphi), \Delta_2(\varphi^*) < \infty$. Then for all $\delta > 0$ there exists $C_\delta > 0$ (depending only on $\Delta_2(\varphi)$ and $\Delta_2(\varphi^*)$) such that for all $s, t \ge 0$ it holds that

$$\begin{aligned} s\,t &\le C_\delta\,\varphi(s) + \delta\,\varphi^*(t), \\ s\,t &\le \delta\,\varphi(s) + C_\delta\,\varphi^*(t). \end{aligned} \tag{2.3}$$

This inequality is called Young's inequality. Moreover, for all $t > 0$ there exists $s > 0$ such that

$$s\,t = \varphi(s) + \varphi^*(t). \tag{2.4}$$

Further basic inequalities are for all $t \geq 0$

$$t \leq \varphi^{-1}(t)\,(\varphi^*)^{-1}(t) \leq 2t,$$

$$\frac{t}{2}\varphi'\Big(\frac{t}{2}\Big) \leq \varphi(t) \leq t\,\varphi'(t), \tag{2.5}$$

$$\varphi\bigg(\frac{\varphi^*(t)}{t}\bigg) \leq \varphi^*(t) \leq \varphi\bigg(\frac{2\,\varphi^*(t)}{t}\bigg).$$

Therefore, uniformly in $t \geq 0$

$$\varphi^{-1}(t)\,(\varphi^*)^{-1}(t) \sim t, \qquad \varphi(t) \sim \varphi'(t)\,t, \qquad \varphi^*\big(\varphi'(t)\big) \sim \varphi(t), \tag{2.6}$$

where the constants depend only on $\Delta_2(\varphi)$ and $\Delta_2(\varphi^*)$.

As in [DE08, DR07] we require the following properties about our function φ.

Assumption 1. Let φ be an N-function with $\Delta_2(\varphi) < \infty$, $\Delta_2(\varphi^*) < \infty$, and $\varphi \in C^2(0,\infty)$ such that

$$\varphi'(t) \sim t\,\varphi''(t) \tag{2.7}$$

uniformly in $t \geq 0$.

It is shown in [DE08] that φ satisfies Assumption 1 iff φ^* satisfies Assumption 1. Moreover, it is shown that we have for all $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$

$$\begin{aligned} \big(\mathbf{A}(\mathbf{a}) - \mathbf{A}(\mathbf{b})\big) \cdot (\mathbf{a} - \mathbf{b}) &\geq c\,\varphi''(|\mathbf{a}| + |\mathbf{b}|)\,|\mathbf{a} - \mathbf{b}|^2, \\ |\mathbf{A}(\mathbf{a}) - \mathbf{A}(\mathbf{b})| &\leq C\,\varphi''(|\mathbf{a}| + |\mathbf{b}|)\,|\mathbf{a} - \mathbf{b}|, \end{aligned} \tag{2.8}$$

where c, C depend only on $\Delta_2(\varphi)$, $\Delta_2(\varphi)$, and the constant in (2.7).

Remark 2. The most important example of such functions is certainly the p-Laplacian. Thereby $\varphi(t) := \int_0^t (\kappa + s)^{p-2}\,s\,ds$ with $1 < p < \infty$ and $\kappa \geq 0$. This function satisfies Assumption 1. If $\kappa = 0$, then Young's inequality (2.3) coincides with the well-known classical Young inequality

$$s\,t \leq \delta \frac{1}{p} t^p + \delta^{\frac{-1}{p-1}} \frac{1}{q} s^q,$$

where $q \in (1,\infty)$ with $\frac{1}{p} + \frac{1}{q} = 1$. Moreover, with $\mathbf{A}(\mathbf{a}) = \varphi'(|\mathbf{a}|)\frac{\mathbf{a}}{|\mathbf{a}|} = (\kappa + |\mathbf{a}|)^{p-2}\mathbf{a}$ (2.8) corresponds to the well-known monotonicity and coercivity inequalities

$$\begin{aligned} \big((\kappa + |\mathbf{a}|)^{p-2}\mathbf{a} - (\kappa + |\mathbf{b}|)^{p-2}\mathbf{b}\big)(\mathbf{a} - \mathbf{b}) &\geq c\,(\kappa + |\mathbf{a}| + |\mathbf{b}|)^{p-2}\,|\mathbf{a} - \mathbf{b}|^2, \\ |(\kappa + |\mathbf{a}|)^{p-2}\mathbf{a} - (\kappa + |\mathbf{b}|)^{p-2}\mathbf{b}| &\leq C\,(\kappa + |\mathbf{a}| + |\mathbf{b}|)^{p-2}\,|\mathbf{a} - \mathbf{b}| \end{aligned}$$

for all $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$ (see, e.g., [GM75, BL93a]).

2.2. Weak formulation of the φ-Laplacian problem and corresponding minimizing problem. We introduce the weak form of the nonlinear equation (2.1). To proceed we first have to introduce analytical background. For details we refer the reader to [RR91].

In the following the function φ will be a fixed N-function as stated in Assumption 1. By L^φ and $W^{1,\varphi}$ we denote the classical Orlicz and Sobolev–Orlicz spaces; i.e., $g \in L^\varphi$ iff $\int \varphi(|g|)\,dx < \infty$ and $g \in W^{1,\varphi}$ iff $g, \nabla g \in L^\varphi$. The norm on L^φ is

given by $\|f\|_\varphi = \inf\{\lambda > 0 : \int \varphi(f/\lambda)\,dx \le 1\}$. By $W_0^{1,\varphi}(\Omega)$ we denote the closure of $C_0^\infty(\Omega)$ in $W^{1,\varphi}(\Omega)$.

The weak formulation reads as follows: For $f \in L^{\varphi^*}(\Omega) \subset (W_0^{1,\varphi}(\Omega))^*$ find $u \in W_0^{1,\varphi}(\Omega)$ with

$$\langle \mathbf{A}(\nabla u), \nabla v\rangle = \langle f, v\rangle \qquad \text{for all } v \in W_0^{1,\varphi}(\Omega). \tag{2.9}$$

The theory of monotone operators ensures the unique existence of u. Moreover, u is the unique minimizer of the *energy functional*

$$\mathcal{J}(u) := \int_\Omega \varphi(|\nabla u|)\,dx - \int_\Omega u\,f\,dx \to \min! \tag{2.10}$$

3. A posteriori analysis. In this section we first discuss our concept of distance. In particular we introduce shifted N-functions and construct a measure of distance related to the nonlinear problem (2.1). In case of the p-Laplacian this concept is equivalent to the quasi norm introduced by Barrett and Liu [BL93a]. In the following we introduce the finite element spaces and note an interpolation inequality. In the second part of this section we construct a posteriori upper and lower bounds for the error of a finite element solution to the continuous solution. Finally, we discuss a discrete local lower bound, i.e., a lower bound for the distance between two different finite element solutions.

3.1. Concept of distance. Let φ again be a fixed N-function. We define another N-function ψ by

$$\psi'(t) := \sqrt{\varphi'(t)\,t}$$

and set for $\mathbf{a} \in \mathbb{R}^d$

$$\mathbf{F}(\mathbf{a}) := \psi'(|\mathbf{a}|)\,\frac{\mathbf{a}}{|\mathbf{a}|} = \sqrt{\varphi'(|\mathbf{a}|)\,|\mathbf{a}|}\;\frac{\mathbf{a}}{|\mathbf{a}|} = \sqrt{|\mathbf{A}(\mathbf{a})|\,|\mathbf{a}|}\,\frac{\mathbf{a}}{|\mathbf{a}|}. \tag{3.1}$$

It is shown in [DE08] that since φ satisfies Assumption 1 also ψ, ψ^*, and φ^* satisfy Assumption 1. It is also shown that as a consequence (2.8) holds with $\mathbf{A}, \varphi$ replaced by $\mathbf{F}, \psi$. In addition we introduce a family of N-function $\{\varphi_a\}_{a\ge 0}$ by

$$\frac{\varphi_a'(t)}{t} := \frac{\varphi'(a+t)}{a+t} \tag{3.2}$$

which owing to (2.7) implies that $\varphi_a''(t) \sim \varphi''(a+t)$ uniformly in $a, t \ge 0$. The functions φ_a are called *shifted* N-functions. The basic properties of φ_a are summarized in the appendix. The connection between $\mathbf{A}$, $\mathbf{F}$, and $\{\varphi_a\}_{a\ge 0}$ is best reflected in the following lemma from [DE08].

LEMMA 3. *Let φ satisfy Assumption* 1 *and let* $\mathbf{A}$ *and* $\mathbf{F}$ *be defined by* (2.2) *and* (3.1). *Then*

$$\big(\mathbf{A}(\mathbf{a}) - \mathbf{A}(\mathbf{b})\big)\cdot\big(\mathbf{a}-\mathbf{b}\big) \sim \big|\mathbf{F}(\mathbf{a}) - \mathbf{F}(\mathbf{b})\big|^2 \tag{3.3a}$$

$$\sim \varphi_{|\mathbf{a}|}(|\mathbf{a}-\mathbf{b}|), \tag{3.3b}$$

$$\sim |\mathbf{a}-\mathbf{b}|^2\,\varphi''\big(|\mathbf{a}| + |\mathbf{b}|\big) \tag{3.3c}$$

uniformly in $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$. *If* $\varphi''(0)$ *does not exists, the expression in* (3.3c) *is continuously extended by zero for* $|\mathbf{a}| = |\mathbf{b}| = \mathbf{0}$. *Moreover,*

$$|\mathbf{A}(\mathbf{a}) - \mathbf{A}(\mathbf{b})| \sim \varphi'_{|\mathbf{a}|}(|\mathbf{a} - \mathbf{b}|), \tag{3.3d}$$

$$\mathbf{A}(\mathbf{b}) \cdot \mathbf{b} \sim |\mathbf{F}(\mathbf{b})|^2 \sim \varphi(|\mathbf{b}|) \tag{3.3e}$$

uniformly in $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$.

We use the equivalences (3.3) extensively in this paper without referring.

An immediately consequence of Lemma 3 is the following corollary.

COROLLARY 4. *We have for all* $u, v \in W^{1,\varphi}(\Omega)$

$$\int_\Omega \big(\mathbf{A}(\nabla u) - \mathbf{A}(\nabla v)\big) \cdot (\nabla u - \nabla v)\, dx \sim \|\mathbf{F}(\nabla u) - \mathbf{F}(\nabla v)\|_2^2 \sim \int_\Omega \varphi_{|\nabla u|}(|\nabla u - \nabla v|)\, dx.$$

Remark 5. In the case of the p-Laplacian, i.e., $\mathbf{A}(\nabla v) = (\kappa + |\nabla v|)^{p-2}\nabla v$ and $\varphi'(t) = (\kappa + t)^{p-2}\, t$ with $1 < p < \infty$ and $\kappa \geq 0$, we have for all $\mathbf{a} \in \mathbb{R}^d$, $t \geq 0$

$$\mathbf{F}(\mathbf{a}) := (\kappa + |\mathbf{a}|)^{\frac{p-2}{2}}\mathbf{a}, \qquad \psi'(t) := (\kappa + t)^{\frac{p-2}{2}}\, t.$$

Moreover, for the p-Laplacian all expressions in Corollary 4 are proportional to the quasi norm introduced by Barrett and Liu in [BL93a]. This follows from the relation

$$\varphi'_{|\mathbf{a}|}(t) = (\kappa + |\mathbf{a}| + t)^{p-2}t$$

and

$$\|\nabla u - \nabla v\|_{(p)}^2 = \int_\Omega (\kappa + |\nabla u| + |\nabla u - \nabla v|)^{p-2}|\nabla u - \nabla v|^2\, dx = \int_\Omega \varphi'_{|\nabla u|}(|\nabla u - \nabla v|)|\nabla u - \nabla v|\, dx \sim \int_\Omega \varphi_{|\nabla u|}(|\nabla u - \nabla v|)\, dx.$$

This ensures in the case of the p-Laplacian that all the results below can also be expressed in terms of the quasi norm.

Additionally, we need the following direct consequence of Lemma 3.

COROLLARY 6. *Let* $\mathbf{A}, \varphi, \mathbf{F}$ *be as in Lemma* 3. *Then for all* $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$

$$(\varphi_{|\mathbf{a}|})^*\big(\big|\mathbf{A}(\mathbf{a}) - \mathbf{A}(\mathbf{b})\big|\big) \sim \varphi_{|\mathbf{a}|}\big(\big|\mathbf{a} - \mathbf{b}\big|\big) \sim \big|\mathbf{F}(\mathbf{a}) - \mathbf{F}(\mathbf{b})\big|^2. \tag{3.4}$$

Proof. The second relation is contained in Lemma 3. The first relation follows from (3.3d), (2.6), and $\Delta_2(\varphi) < \infty$. ∎

3.2. Finite element spaces. Let $\mathcal{T}_H$ be a conforming triangulation of Ω consisting of closed simplices $T \in \mathcal{T}_H$. Let h_T denote the diameter of the (closed) simplex $T \in \mathcal{T}_H$ and ρ_T the maximal radius of a ball that is contained in T. The maximal quotient h_T/ρ_T with $T \in \mathcal{T}_h$ is called the *shape regularity* (chunkiness) of $\mathcal{T}_H$.

Let $V_H := V(\mathcal{T}_H)$ be the space of continuous, piecewise linear finite elements over $\mathcal{T}_H$ with boundary values zero; then $V_H \subset V$. By $u_H \in V_H$ we denote the finite element solution of (2.9) with respect to V_H, i.e.,

$$\langle \mathbf{A}(\nabla u_H), \nabla v_H\rangle = \langle f, v_H\rangle \qquad \text{for all } v_H \in V_H, \tag{3.5}$$

where $\langle f, v_H\rangle = \int_\Omega f\, v_H\, dx$. The theory of monotone operators ensures the existence of a unique solution.

We denote by Γ_H the set of interior faces (sides) of the triangulation $\mathcal{T}_H$. For $\gamma \in \Gamma_H$ we define N_γ as the set of elements sharing γ and S_γ as the union of these elements, i.e.,

$$N_\gamma := \{T_1, T_2 \in \mathcal{T}_H \,:\, T_1 \cap T_2 = \gamma\}, \qquad S_\gamma := \bigcup_{T\in N_\gamma} T.$$

For $T \in \mathcal{T}_H$ define the set of neighbors N_T and the neighborhood S_T by

$$N_T := \{T' \in \mathcal{T}_H \,:\, T' \cap T \in \Gamma_H\}, \qquad S_T := \bigcup_{T'\in N_T} T'.$$

For interpolation estimates (see (3.6)) we additionally need to define the patch Ω_T around T and the set of its elements ω_T by

$$\omega_T := \{T' \in \mathcal{T}_H \,:\, T' \cap T \neq \emptyset\}, \qquad \Omega_T := \bigcup_{T'\in\omega_T} T'.$$

For $\gamma \in \Gamma_H$ let $h_\gamma := \operatorname{diam}(\gamma)$. For $T \in \mathcal{T}_H$ it holds that $h_T \sim h_\gamma$ for each face (side) $\gamma \subset \partial T$ depending only on the shape regularity of $\mathcal{T}_H$.

Let $\Pi_H : V \to V_H$ be the Scott–Zhang interpolation operator which respects zero boundary values; see [SZ90]. Additionally Π_H satisfies $\Pi_H v_H = v_H$ for all $v_H \in V_H$. It was shown in [DR07] that for all $v \in W^{1,\varphi}(\Omega)$, $a \geq 0$, and $T \in \mathcal{T}_h$

$$\int_T \varphi_a(|v - \Pi_H v|)\, dx + \int_T \varphi_a(h_T\,|\nabla v - \nabla \Pi_H v|)\, dx \leq C \int_{\Omega_T} \varphi_a(h_T\,|\nabla v|)\, dx, \tag{3.6}$$

where C depends only on $\Delta_2(\varphi)$ and the shape regularity of $\mathcal{T}_H$.

We introduce residual based error estimators for our system (2.1). For $\gamma \in \Gamma_H$ and $T \in \mathcal{T}_H$ define the (local) *interior and the jump estimators* by

$$\eta_E^2(u_H, T) := \int_T (\varphi_{|\nabla u_H|})^*\big(h_T\,|f|\big)\, dx,$$

$$\eta_J^2(u_H, \gamma) := \int_\gamma h_\gamma \left|[\![\mathbf{F}(\nabla u_H)]\!]_\gamma\right|^2 dx,$$

where $[\![\mathbf{F}(\nabla u_H)]\!]_\gamma$ denotes the jump of $\mathbf{F}(\nabla u_H)$ over the face γ. Furthermore, we define for $T \in \mathcal{T}_H$ the (local) element based error indicators and the oscillation as

$$\begin{aligned} \eta^2(u_H, T) &:= \eta_E^2(u_H, T) + \sum_{\gamma\in\Gamma_H, \gamma\subset\partial T} \eta_J^2(u_H, \gamma), \\ \operatorname{osc}^2(u_H, T) &:= \inf_{f_T\in\mathbb{R}} \int_T (\varphi_{|\nabla u_H|})^*\big(h_T\,|f - f_T|\big)\, dx. \end{aligned} \tag{3.7}$$

For a subset $\hat{\mathcal{T}}_H \subset \mathcal{T}_H$ we define the *total error estimator* over $\hat{\mathcal{T}}_H$ by

$$\eta^2(u_k, \hat{\mathcal{T}}_H) := \sum_{T\in\hat{\mathcal{T}}_H} \eta^2(u_k, T).$$

Note that $\eta^2(u_H, \{T\}) = \eta^2(u_H, T)$, and so there is no confusion between this definition of η^2 and (3.7). Similarly, we define the *total oscillation* on subsets of $\mathcal{T}_H$.

Furthermore, we use $e_H := u - u_H$ for the difference of the solutions.

Remark 7. In the case of the p-Laplacian we can translate the above definitions of the estimators and the oscillation. For the jump estimator we obtain by Corollary 4, Lemma 22, and Remark 5

$$\begin{aligned}\eta_J^2(u_H,\gamma) &= \int_\gamma h_\gamma \left|[\![\mathbf{F}(\nabla u_H)]\!]_\gamma\right|^2 dx \sim \int_{S_\gamma} \varphi_{|\nabla u_H|}\big([\![\nabla u_H]\!]_\gamma\big)\\ &\sim \int_{S_\gamma} \big(\kappa + |\nabla u_H| + |[\![\nabla u_H]\!]_\gamma|\big)^{p-2} |[\![\nabla u_H]\!]_\gamma|\, dx.\end{aligned}$$

The element residual and the oscillation can be treated in the same way. We use Lemma 22, (2.6), and Remark 5 to obtain

$$\begin{aligned}\eta_E^2(u_H,T) &= \int_T (\varphi_{|\nabla u_H|})^*\big(h_T\,|f|\big)\,dx \sim \int_T (\varphi^*)_{\varphi'(|\nabla u_H|)}\big(h_T\,|f|\big)\,dx\\ &\sim \int_T \big(\kappa + |\nabla u_H|^{p-1} + h_T\,|f|\big)^{q-2} h_T^2\,|f|^2\,dx.\end{aligned}$$

Let f_T denote the mean value of f over T. Then the same calculations yield for the oscillation

$$\operatorname{osc}^2(u_H,T) \sim \int_T \big(\kappa + |\nabla u_H|^{p-1} + h_T\,|f - f_T|\big)^{q-2} h_T^2\,|f - f_T|^2\,dx,$$

where we have used that for any N-function ϱ with $\Delta_2(\varrho) < \infty$ it holds that

$$\inf_{f_T} \int_T \varrho(|f - f_T|)\,dx \le \int_T \varrho(|f - f_T|)\,dx \le c \inf_{f_T} \int_T \varrho(|f - f_T|)\,dx$$

with constants c depending only on $\Delta_2(\varphi)$. Thus our a posteriori estimators improve the one in [LY02].

3.3. Upper bound. To obtain the upper bound we use Lemma 3, the Galerkin orthogonality, and $\Pi_H : V \to V_H$:

$$\begin{aligned}\big\|\mathbf{F}(\nabla u_H) - \mathbf{F}(\nabla u)\big\|_2^2 &\sim \big\langle \mathbf{A}(\nabla u) - \mathbf{A}(\nabla u_H), \nabla(u - u_H)\big\rangle\\ &= \big\langle \mathbf{A}(\nabla u) - \mathbf{A}(\nabla u_H), \nabla(e_H - \Pi_H e_H)\big\rangle.\end{aligned}$$

By integration by parts on each $T \in \mathcal{T}_h$ we get

$$\begin{aligned}&\big\|\mathbf{F}(\nabla u_H) - \mathbf{F}(\nabla u)\big\|_2^2\\ &\quad\sim -\sum_{T\in\mathcal{T}_H} \int_{\partial T} \big(\mathbf{A}(\nabla u_H)\cdot\mathbf{n}\big)\big(e_H - \Pi_H e_H\big)\,dx + \big\langle f, e_H - \Pi_H e_H\big\rangle\\ &\quad= -\sum_{\gamma\in\Gamma_H} \int_\gamma [\![\mathbf{A}(\nabla u_H)\cdot\mathbf{n}]\!]_\gamma \big(e_H - \Pi_H e_H\big)\,dx + \big\langle f, e_H - \Pi_H e_H\big\rangle\\ &\quad= (\mathtt{Upper}_1) + (\mathtt{Upper}_2),\end{aligned} \tag{3.8}$$

where $\mathbf{n}$ is the outer normal of ∂T. We handle the two terms jump residual ($\mathtt{Upper}_1$) and element residual ($\mathtt{Upper}_2$) separately. First we estimate

$$(\mathtt{Upper}_1) \le \sum_{\gamma\in\Gamma_H} \big|[\![\mathbf{A}(\nabla u_H)]\!]_\gamma\big| \int_\gamma \big|e_H - \Pi_H e_H\big|\,dx, \tag{3.9}$$

where we have used that ∇u_H is constant on each $T \in \mathcal{T}_H$. The trace theorem $W^{1,1}(S_\gamma) \hookrightarrow L^1(\gamma)$ and the $W^{1,1}$-approximability of Π_H (see [SZ90] or use (3.6) with $\varphi(t) = t$ and $a = 0$) imply that

$$\begin{aligned}\int_\gamma \big|e_H - \Pi_H e_H\big|\,dx &\le C\,h_\gamma \int_{S_\gamma} |e_H - \Pi_H e_H|\,dx + C \int_{S_\gamma} |\nabla e_H - \nabla\Pi_H e_H|\,dx\\ &\le c \sum_{T\in N_\gamma} \int_{\Omega_T} |\nabla e_h|\,dx.\end{aligned}$$

If we combine this with (3.9), then we get

$$(\mathtt{Upper}_1) \le C \sum_{\gamma\in\Gamma_H} \left(\big|[\![\mathbf{A}(\nabla u_H)]\!]_\gamma\big| \sum_{T\in N_\gamma} \int_{\Omega_T} \big|\nabla e_H\big|\,dx \right).$$

Now Young's inequality (2.3) for $\varphi_{|\nabla u_H|}$ on each element yields

$$(\mathtt{Upper}_1) \le \sum_{\gamma\in\Gamma_H} \sum_{T\in N_\gamma} \int_{\Omega_T} C_\delta\,(\varphi_{|\nabla u_H|})^*\big(\big|[\![\mathbf{A}(\nabla u_H)]\!]_\gamma\big|\big) + \delta\,\varphi_{|\nabla u_H|}\big(\big|\nabla e_H\big|\big)\,dx.$$

Due to (3.4) we have

$$(\varphi_{|\nabla u_H|})^*\big(\big|[\![\mathbf{A}(\nabla u_H)]\!]_\gamma\big|\big) \sim \varphi_{|\nabla u_H|}\big(\big|[\![\nabla u_H]\!]_\gamma\big|\big) \sim \big|[\![\mathbf{F}(\nabla u_H)]\!]_\gamma\big|^2. \tag{3.10}$$

Note that due to shape regularity it holds that $h_{T'} \sim h_T \sim h_\gamma$ for $T' \in \omega_T$, $T \in N_\gamma$. With the help of (3.10), Lemma 3, the finite overlapping of the Ω_T, and $\#N_T = 2$ we get

$$\begin{aligned}(\mathtt{Upper}_1) &\le C_\delta\,C \sum_{\gamma\in\Gamma_H} \sum_{T\in N_\gamma} \int_{\Omega_T} \big|[\![\mathbf{F}(\nabla u_H)]\!]_\gamma\big|^2\\ &\quad + \delta\,C \sum_{\gamma\in\Gamma_H} \sum_{T\in N_\gamma} \int_{\Omega_T} \big|\mathbf{F}(\nabla u) - \mathbf{F}(\nabla u_H)\big|^2\,dx\\ &\le C_\delta\,C \sum_{\gamma\in\Gamma_H} \int_\gamma h_\gamma \big|[\![\mathbf{F}(\nabla u_H)]\!]_\gamma\big|^2 \qquad (3.11)\\ &\quad + \delta\,C \sum_{\gamma\in\Gamma_H} \sum_{T\in N_\gamma} \int_{\Omega_T} \big|\mathbf{F}(\nabla u) - \mathbf{F}(\nabla u_H)\big|^2\,dx\\ &\le C_\delta\,C \sum_{\gamma\in\Gamma_H} \eta_J^2(u_H,\gamma) + \delta\,C\,\big\|\mathbf{F}(\nabla u_H) - \mathbf{F}(\nabla u)\big\|_2^2.\end{aligned}$$

We treat the element residual as follows:

$$\begin{aligned}(\mathtt{Upper}_2) &\le \sum_{T\in\mathcal{T}_H}\int_T |f|\,|e_H-\Pi_H e_H|\,dx\\ &\le \sum_{T\in\mathcal{T}_H}\int_T C_\delta\,(\varphi_{|\nabla u_H|})^*\big(h_T\,|f|\big)+\delta\,(\varphi_{|\nabla u_H|})\bigg(\frac{|e_H-\Pi_H e_H|}{h_T}\bigg)\,dx,\end{aligned}$$

where we have used Young's inequality (2.3). Since ∇u_H is constant on each $T\in\mathcal{T}_H$ we can use (3.6) to obtain

$$(\mathtt{Upper}_2)\le C_\delta\,C\sum_{T\in\mathcal{T}_H}\eta_E^2(u_H,T)+\delta\,C\sum_{T\in\mathcal{T}_H}\int_{\Omega_T}\varphi_{|\nabla u_H(T)|}(|\nabla e_H|)\,dx,$$

where we write $\nabla u_H(T)$ to indicate that the shift on the whole Ω_T depends on the value of ∇u_H on the triangle T. In order to get $\varphi_{|\nabla u_H|}(|\nabla e_H|)$ instead of $\varphi_{|\nabla u_H(T)|}(|\nabla e_H|)$ we need a *change of shift*. We apply Corollary 26 on each $T'\in\omega_T$ and get

$$\begin{aligned}(\mathtt{Upper}_2)&\le C_\delta\,C\sum_{T\in\mathcal{T}_H}\eta_E^2(u_H,T)+\delta\,C\sum_{T\in\mathcal{T}_H}\int_{\Omega_T}\varphi_{|\nabla u_H|}(|\nabla e_H|)\,dx\\ &\quad+\delta\,C\sum_{T\in\mathcal{T}_H}\sum_{T'\in\omega_T}\int_{T'}\big|\mathbf{F}(\nabla u_H(T))-\mathbf{F}(\nabla u_H(T'))\big|^2\,dx.\end{aligned}$$

Now we transform the last term. Since one can reach T' from T by passing through a finite number of faces (depending on the shape regularity of $\mathcal{T}_H$), we can estimate each $\big|\mathbf{F}(\nabla u_H(T))-\mathbf{F}(\nabla u_H(T'))\big|$ for $T'\in\omega_T$ by a sum of jumps $\big|[\![\mathbf{F}(\nabla u_H)]\!]_\gamma\big|$ over these faces. In particular,

$$\begin{aligned}\sum_{T\in\mathcal{T}_H}\sum_{T'\in\omega_T}\int_{T'}\big|\mathbf{F}(\nabla u_H(T))-\mathbf{F}(\nabla u_H(T'))\big|^2\,dx&\sim\sum_{\gamma\in\Gamma_H}\int_{S_\gamma}\big|[\![\mathbf{F}(\nabla u_H)]\!]_\gamma\big|^2\,dx\\ &\sim\sum_{\gamma\in\Gamma_H}h_\gamma\int_\gamma\big|[\![\mathbf{F}(\nabla u_H)]\!]_\gamma\big|^2\,dx\;=\sum_{\gamma\in\Gamma_H}\eta_J^2(u_H,\gamma)\end{aligned}$$

using that ∇u_H is piecewise constant and $|S_\gamma|\sim h_\gamma\,|\gamma|$. Hence, with Lemma 3

$$\begin{aligned}&(\mathtt{Upper}_2)\\ &\le C_\delta\,C\sum_{T\in\mathcal{T}_H}\eta_E^2(u_H,T)+\delta\,C\sum_{T\in\mathcal{T}_H}\int_{\Omega_T}\varphi_{|\nabla u_H|}(|\nabla e_H|)\,dx+\delta\,C\sum_{\gamma\in\Gamma_H}\eta_J^2(u_H,\gamma)\\ &\le C_\delta\,C\sum_{T\in\mathcal{T}_H}\eta_E^2(u_H,T)+\delta\,C\,\big\|\mathbf{F}(\nabla u_H)-\mathbf{F}(\nabla u)\big\|_2^2+\delta\,C\sum_{\gamma\in\Gamma_H}\eta_J^2(u_H,\gamma).\end{aligned}$$

Now taking $\delta>0$ small enough we obtain from (3.8), (3.11), the last inequality, (3.7), and the fact that each side $\gamma\in\Gamma_H$ is shared by at most two elements the following lemma.

LEMMA 8 (upper bound). *For finite element solutions u_H of (3.5) it holds that*

$$\big\|\mathbf{F}(\nabla u_H)-\mathbf{F}(\nabla u)\big\|_2^2\le C\,\eta^2(u_H,\mathcal{T}_H),\tag{3.12}$$

where the constant C depends only on $\Delta_2(\varphi)$, $\Delta_2(\varphi^*)$, and the shape regularity of $\mathcal{T}_H$.

3.4. Lower bound. In this section we show that the error can be locally estimated from below by the error estimators. We begin with the element estimator. As is well known, for each $T \in \mathcal{T}_H$ there exists a *bubble function* $w_T \in W_0^{1,\varphi}(T)$ with $w_T \geq 0$ and

$$\int_T w_T \, dx = |T|, \qquad \|w_T\|_\infty \leq C, \qquad \|\nabla w_T\|_\infty \leq \frac{C}{h_T}, \tag{3.13}$$

where $C > 0$ depends only on the shape regularity of $\mathcal{T}_H$. Then for $s \in \mathbb{R}$

$$\big\langle \mathbf{A}(\nabla u) - \mathbf{A}(\nabla u_H), \nabla(s\, w_T) \big\rangle = \langle f, s\, w_T \rangle, \tag{3.14}$$

where we have used that ∇u_H is constant on T. For $f_T \in \mathbb{R}$ by (2.4) applied to $\varphi_{|\nabla u_H|}$ there exists $s_T \in \mathbb{R}$ such that

$$s_T\,(h_T f_T) = (\varphi_{|\nabla u_H(T)|})^*(h_T|f_T|) + \varphi_{|\nabla u_H|}(|s_T|); \tag{3.15}$$

i.e., Young's inequality is sharp. We obtain with (3.15) and (3.14) taking $s = h_T s_T$

$$\begin{aligned} |T|\,(\varphi_{|\nabla u_H(T)|})^*(h_T|f_T|) + |T|\,\varphi_{|\nabla u_H(T)|}(|s_T|) &= |T|\, f_T\, h_T\, s_T \\ &= \big\langle \mathbf{A}(\nabla u) - \mathbf{A}(\nabla u_H(T)), \nabla(h_T\, s_T\, w_T) \big\rangle + \langle f_T - f, h_T\, s_T\, w_T \rangle \\ &= (\mathtt{Lower}_1) + (\mathtt{Lower}_2). \end{aligned} \tag{3.16}$$

With the help of (3.13), Young's inequality (2.3), (3.4), and the fact that the integrand is constant $\varphi_{|\nabla u_H|}(|s_T|) = \varphi_{|\nabla u_H(T)|}(|s_T|)$ on T, we have

$$\begin{aligned} (\mathtt{Lower}_1) &\leq C \int_T |\mathbf{A}(\nabla u) - \mathbf{A}(\nabla u_H)|\, |s_T|\, dx \\ &\leq C_\delta\, C \int_T |\mathbf{F}(\nabla u) - \mathbf{F}(\nabla u_H)|^2\, dx + \delta\, C\, |T|\, \varphi_{|\nabla u_H(T)|}(|s_T|). \end{aligned} \tag{3.17}$$

Similarly, with (3.13) and Young's inequality (2.3) we get

$$\begin{aligned} (\mathtt{Lower}_2) &\leq C \int_T h_T\, s_T\, |f - f_T|\, dx \\ &\leq C_\delta\, C \int_T (\varphi_{|\nabla u_H|})^*(h_T|f - f_T|)\, dx + \delta\, C|T|\, \varphi_{|\nabla u_H(T)|}(|s_T|). \end{aligned} \tag{3.18}$$

Now taking $\delta > 0$ small enough we obtain from (3.16), (3.17), and (3.18) that

$$\begin{aligned} |T|\,(\varphi_{|\nabla u_H(T)|})^*(h_T|f_T|) \leq C \int_T |\mathbf{F}(\nabla u) - \mathbf{F}(\nabla u_H)|^2\, dx \\ + C \int_T (\varphi_{|\nabla u_H|})^*(h_T|f - f_T|)\, dx. \end{aligned} \tag{3.19}$$

Observe that by convexity of $(\varphi_{|\nabla u_H(T)|})^*$, $\Delta_2(\varphi^*), \Delta_2(\varphi) < \infty$, and Lemma 22

$$\begin{aligned} \eta_E^2(u_H, T) &= \int_T (\varphi_{|\nabla u_H|})^*(h_T|f|) \\ &\leq \int_T (\varphi_{|\nabla u_H|})^*(2h_T|f - f_T|) + (\varphi_{|\nabla u_H|})^*(2h_T|f_T|) \\ &\leq C \int_T (\varphi_{|\nabla u_H|})^*(h_T|f - f_T|)\, dx + C\,|T|\,(\varphi_{|\nabla u_H(T)|})^*(h_T|f_T|) \end{aligned}$$

with $C > 0$ depending only on $\Delta_2(\varphi)$ and $\Delta_2(\varphi^*)$. This and (3.19) give

$$\eta_E^2(u_H,T) \le C \int_T |\mathbf{F}(\nabla u) - \mathbf{F}(\nabla u_H)|^2 \,dx + C \int_T (\varphi_{|\nabla u_H|})^*(h_T|f - f_T|)\,dx. \tag{3.20}$$

Taking the infimum over all $f_T \in \mathbb{R}$ proves the following assertion.

LEMMA 9. *For finite element solutions u_H of (3.5) and $T \in \mathcal{T}_h$ it holds that*

$$\eta_E^2(u_H,T) \le C\,\|\mathbf{F}(\nabla u_H) - \mathbf{F}(\nabla u)\|_{L^2(T)}^2 + C\,\mathrm{osc}^2(u_H,T), \tag{3.21}$$

where the constant C depends only on $\Delta_2(\varphi)$, $\Delta_2(\varphi^*)$, and the shape regularity of $\mathcal{T}_H$.

Now we estimate the jump estimator. As is well known, for each $\gamma \in \Gamma_H$ there exists a *bubble function* $w_\gamma \in W_0^{1,\varphi}(S_\gamma)$ with $w_\gamma \ge 0$ and

$$\int_\gamma w_\gamma \,dx = |\gamma|, \qquad \|w_\gamma\|_\infty \le C, \qquad \|\nabla w_\gamma\|_\infty \le \frac{C}{h_\gamma}, \tag{3.22}$$

where $C > 0$ depends only on the shape regularity of $\mathcal{T}_H$. Then for $s \in \mathbb{R}$

$$\begin{aligned} \big\langle \mathbf{A}(\nabla u) - \mathbf{A}(\nabla u_H), \nabla(s\,w_\gamma)\big\rangle &= \langle f, s\,w_\gamma\rangle - \int_\gamma [\![\mathbf{A}(\nabla u_H)\cdot\mathbf{n}]\!]_\gamma\, s\, w_\gamma \,dx \\ &= \langle f, s\,w_\gamma\rangle - |\gamma|\,[\![\mathbf{A}(\nabla u_H)\cdot\mathbf{n}]\!]_\gamma\, s, \end{aligned} \tag{3.23}$$

where we have used partial integration and that ∇u_H is piecewise constant. Let T_0, T_1 be the two triangles sharing γ. Then by (2.4) applied to $\varphi_{|\nabla u_H(T_0)|}$ there exists $s_\gamma \in \mathbb{R}$ such that

$$[\![\mathbf{A}(\nabla u_H)\cdot\mathbf{n}]\!]_\gamma\, s_\gamma = (\varphi_{|\nabla u_H(T_0)|})^*\big(\big|[\![\mathbf{A}(\nabla u_H)\cdot\mathbf{n}]\!]_\gamma\big|\big) + \varphi_{|\nabla u_H(T_0)|}(|s_\gamma|); \tag{3.24}$$

i.e., Young's inequality is sharp. We have chosen $|\nabla u_H(T_0)|$ as the shift, which puts T_0 into a special position, but we will see later that it is not important which of the two triangles is chosen. Let $s = \frac{|S_\gamma|}{|\gamma|} s_\gamma$ in (3.23); then we obtain with (3.24)

$$\begin{aligned} &|S_\gamma|\,(\varphi_{|\nabla u_H(T_0)|})^*\big([\![\mathbf{A}(\nabla u_H)\cdot\mathbf{n}]\!]_\gamma\big) + |S_\gamma|\,\varphi_{|\nabla u_H(T_0)|}(|s_\gamma|) \\ &\quad = |S\gamma|\,[\![\mathbf{A}(\nabla u_H)\cdot\mathbf{n}]\!]_\gamma\, s_\gamma \\ &\quad = -\frac{|S_\gamma|}{|\gamma|}\big\langle \mathbf{A}(\nabla u) - \mathbf{A}(\nabla u_H), \nabla(s_\gamma\, w_\gamma)\big\rangle + \frac{|S_\gamma|}{|\gamma|}\,\langle f, s_\gamma\, w_\gamma\rangle \\ &\quad = (\mathtt{Lower}_3) + (\mathtt{Lower}_4). \end{aligned} \tag{3.25}$$

Before we proceed with the estimates for $(\mathtt{Lower}_3)$ and $(\mathtt{Lower}_4)$ we simplify the term $(\varphi_{|\nabla u_H(T_0)|})^*\big(\big|[\![\mathbf{A}(\nabla u_H)\cdot\mathbf{n}]\!]_\gamma\big|\big)$. First we show that

$$\big|[\![\mathbf{A}(\nabla u_H)\cdot\mathbf{n}]\!]_\gamma\big| \sim \varphi'_{|\nabla u_H|}\big(\big|[\![\nabla u_H]\!]_\gamma\big|\big) \sim \big|[\![\mathbf{A}(\nabla u_H)]\!]_\gamma\big|. \tag{3.26}$$

The last part of (3.26) is an immediate consequence of Lemma 3. If $[\![\nabla u_H]\!]_\gamma = 0$, then also $[\![\mathbf{A}(\nabla u_H)\cdot\mathbf{n}]\!]_\gamma = 0$ and all terms in (3.26) are zero. So we can assume that $[\![\nabla u_H]\!]_\gamma \ne 0$. Since $u_H \in C(S_\gamma)$, the tangential derivatives of u_H are continuous on γ and do not jump. Hence, $|[\![\nabla u_H]\!]_\gamma| = |[\![\nabla u_H]\!]_\gamma \cdot \mathbf{n}|$ and

$$\mathbf{n} = \pm \frac{[\![\nabla u_H]\!]_\gamma}{|[\![\nabla u_H]\!]_\gamma|}.$$

This and (3.3) imply that

$$|[\![\mathbf{A}(\nabla u_H)\cdot\mathbf{n}]\!]_\gamma|\cdot|[\![\nabla u_H]\!]_\gamma| = \big|[\![\mathbf{A}(\nabla u_H)]\!]_\gamma\cdot[\![\nabla u_H]\!]_\gamma\big| \sim \varphi_{|\nabla u_H|}(|[\![\nabla u_H]\!]_\gamma|).$$

Now (2.6) proves (3.26). Lemma 22 states that $(\varphi_{|u_H|})^*$ is an N-function with $\Delta_2((\varphi_{|u_H|})^*) < \infty$ depending solely on $\Delta_2(\varphi)$ and $\Delta_2(\varphi^*)$. Therefore, we get with (3.4) and (3.26)

$$\begin{aligned}(\varphi_{|\nabla u_H(T_0)|})^*\big(\big|[\![\mathbf{A}(\nabla u_H)\cdot\mathbf{n}]\!]_\gamma\big|\big) &\sim (\varphi_{|\nabla u_H(T_0)|})^*\big(\big|[\![\mathbf{A}(\nabla u_H)]\!]_\gamma\big|\big)\\ &\sim \big|[\![\mathbf{F}(\nabla u_H)]\!]_\gamma\big|^2.\end{aligned} \tag{3.27}$$

Now it becomes clear why the preference of T_0 is not important: The expression $|[\![\mathbf{F}(\nabla u_H)]\!]_\gamma|^2$ in (3.27) is symmetrical in T_0 and T_1 and therefore independent of the choice T_0.

We proceed with the estimate for ($\mathtt{Lower}_3$). With $|S_\gamma| \sim h_\gamma\,|\gamma|$, (3.22), Young's inequality (2.3), and (3.4) we get

$$\begin{aligned}(\mathtt{Lower}_3) &\le C\int_{S_\gamma}|\mathbf{A}(\nabla u)-\mathbf{A}(\nabla u_H)|\,|s_\gamma|\,dx\\ &\le C_\delta\,C\int_{S_\gamma}|\mathbf{F}(\nabla u)-\mathbf{F}(\nabla u_H)|^2\,dx+\delta\,C\sum_{T\in N_\gamma}|T|\,\varphi_{|\nabla u_H(T)|}(|s_\gamma|).\end{aligned} \tag{3.28}$$

With $|S_\gamma| \sim h_\gamma\,|\gamma|$, (3.22), and Young's inequality (2.3) we deduce that

($\mathtt{Lower}_4$)

$$\begin{aligned}&\le C\int_{S_\gamma}|f|\,h_\gamma\,|s_\gamma|\,dx\\ &\le C\sum_{T\in N_\gamma}\inf_{f_T\in\mathbb{R}}\int_T|f-f_T|\,h_\gamma\,|s_\gamma|+|f_T|\,h_\gamma\,|s_\gamma|\,dx\\ &\le C_\delta\,C\sum_{T\in N_\gamma}\inf_{f_T\in\mathbb{R}}\int_T(\varphi_{|\nabla u_H|})^*(h_\gamma\,|f-f_T|)\,dx+(\varphi_{|\nabla u_H|})^*(h_\gamma|f_T|)\,dx\\ &\quad+\delta\,C\sum_{T\in N_\gamma}\int_T\varphi_{|\nabla u_H|}(|s_\gamma|)\,dx\\ &\le C_\delta\,C\sum_{T\in N_\gamma}\operatorname{osc}^2(u_H,T)+C_\delta\,C\sum_{T\in N_\gamma}\eta_E^2(u_H,T)+\delta\,C\sum_{T\in N_\gamma}|T|\varphi_{|\nabla u_H(T)|}(|s_\gamma|).\end{aligned}$$

This, (3.25), (3.28), and $|T_0|, |T_1| \le |S_\gamma|$ imply that

$$\begin{aligned}&|S_\gamma|\,(\varphi_{|\nabla u_H(T_0)|})^*\big(\big|[\![\mathbf{A}(\nabla u_H)\cdot\mathbf{n}]\!]_\gamma\big|\big)+|S_\gamma|\,\varphi_{|\nabla u_H(T_0)|}(|s_\gamma|)\\ &\quad\le C_\delta\,C\int_{S_\gamma}|\mathbf{F}(\nabla u)-\mathbf{F}(\nabla u_H)|^2\,dx+C_\delta\,C\sum_{T\in N_\gamma}\operatorname{osc}^2(u_H,T)\\ &\qquad+C_\delta\,C\sum_{T\in N_\gamma}\eta_E^2(u_H,T)+\delta\,C\sum_{T\in N_\gamma}|S_\gamma|\varphi_{|\nabla u_H(T)|}(|s_\gamma|).\end{aligned}$$

For small $\delta > 0$ the summand of the last term with $T = T_0$ could be absorbed on the left-hand side, but the term with $T = T_1$ bothers us, since it has the wrong shift $|\nabla u_H(T_1)|$. With Corollary 26 and (3.27) we get rid of this term:

$$\begin{aligned}\varphi_{|\nabla u_H(T_1)|}(|s_\gamma|) &\le C\,\varphi_{|\nabla u_H(T_0)|}(|s_\gamma|) + C\,\big|[\![\mathbf{F}(\nabla u_H)]\!]_\gamma\big|^2 \\ &\le C\,\varphi_{|\nabla u_H(T_0)|}(|s_\gamma|) + C\,(\varphi_{|\nabla u_H(T_0)|})^*\big(\big|[\![\mathbf{A}(\nabla u_H)\cdot\mathbf{n}]\!]_\gamma\big|\big).\end{aligned}$$

This and (3.28) give for $\delta > 0$ small

$$\begin{aligned}&|S_\gamma|\,(\varphi_{|\nabla u_H(T_0)|})^*\big(\big|[\![\mathbf{A}(\nabla u_H)\cdot\mathbf{n}]\!]_\gamma\big|\big) \\ &\qquad\le C\,\|\mathbf{F}(\nabla u) - \mathbf{F}(\nabla u_H)\|^2_{L^2(S_\gamma)} + C\sum_{T\in N_\gamma}\operatorname{osc}^2(u_H,T) + C\sum_{T\in N_\gamma}\eta_E^2(u_H,T).\end{aligned}$$

Now an application of (3.27) and $|S_\gamma| \sim h_\gamma\,|\gamma|$ prove the following assertion.

LEMMA 10. *For finite element solutions u_H of (3.5) and $\gamma \in \Gamma_H$ it holds that*

$$\eta_J^2(u_H,\gamma) \le C\,\|\mathbf{F}(\nabla u_H) - \mathbf{F}(\nabla u)\|^2_{L^2(S_\gamma)} + C\operatorname{osc}^2(u_H,N_\gamma) + C\,\eta_E^2(u_H,N_\gamma),$$

where the constant C depends only on $\Delta_2(\varphi)$, $\Delta_2(\varphi^*)$, and the shape regularity of $\mathcal{T}_H$.

Lemmas 9 and 10 can be combined.

COROLLARY 11. *For finite element solutions u_H of (3.5) and $T \in \mathcal{T}_H$ it holds that*

$$\eta^2(u_H,T) \le C\,\|\mathbf{F}(\nabla u_H) - \mathbf{F}(\nabla u)\|^2_{L^2(S_T)} + C\operatorname{osc}^2(u_H,N_T), \tag{3.29}$$

where the constant C depends only on $\Delta_2(\varphi)$, $\Delta_2(\varphi^*)$, and the shape regularity of $\mathcal{T}_H$.

3.5. Discrete lower estimates. In the following let $\mathcal{T}_h$ be a refinement of $\mathcal{T}_H$, which is generated from $\mathcal{T}_H$ by finitely many bisections. Then $V_h := V(\mathcal{T}_h)$ and $V_H := V(\mathcal{T}_H)$ are nested, i.e., $V_H \subset V_h \subset V$. Let u_h, u_H denote the unique solution of (2.1) with respect to V_h (resp., V_H).

Our aim is to generalize Corollary 11 from $V_H \subset V$ to $V_H \subset V_h$. Therefore, we have to ensure that V_h is a sufficient refinement of V_H. In particular, we have to ensure the existence of bump functions as required in (3.13) and (3.22). We say that $T \in \mathcal{T}_H$ is *fully refined in* $\mathcal{T}_h$ if T and each of its faces contain a node in their interior. This yields the existence of a bump function $w_T \in V_h$ on T which satisfies (3.13) and bump functions $w_\gamma \in V_h$ on S_γ for all $\gamma \in \Gamma_H \cup T$ which satisfy (3.22). Thus, to obtain the local lower bound (3.29) on S_T for a certain $T \in \mathcal{T}_H$ it suffices to assume that each $T' \in N_T$ is fully refined in $\mathcal{T}_h$. With these additional assumptions we can now transfer the estimates from section 3.4 to $V_H \subset V_h$.

LEMMA 12. *If for $T \in \mathcal{T}_H$ each $T' \in N_T$ is fully refined in $\mathcal{T}_h$, then*

$$\eta^2(u_H,T) \le C\,\|\mathbf{F}(\nabla u_H) - \mathbf{F}(\nabla u_h)\|^2_{L^2(S_T)} + C\operatorname{osc}^2(u_H,N_T), \tag{3.30}$$

where the constant C depends only on $\Delta_2(\varphi)$, $\Delta_2(\varphi^*)$, and the shape regularity of $\mathcal{T}_H$.

4. Algorithm and convergence. Let us first state the adaptive algorithm which produces a sequence u_k of solutions in nested spaces $V_k := V(\mathcal{T}_k)$ over triangulations $\mathcal{T}_k$. We substitute the index H (resp., h) of the previous sections by the index k (resp., $k+1$) to indicate the underlying mesh. Then we introduce the concept of energy reduction and thereafter prove convergence of the algorithm.

4.1. Adaptive algorithm. We propose the following adaptive algorithm to solve (1.1).

ALGORITHM 13 (AFEM). Choose $\theta \in (0,1)$. Construct an initial triangulation $\mathcal{T}_0$ of Ω and set $k := 0$.

1. (*"Solve"*) Compute the solution $u_k \in V_k$ of problem (3.5).
2. (*"Estimate"*) Compute $\eta^2(u_k, T)$ for all $T \in \mathcal{T}_k$.
3. If $\eta^2(u_k, \mathcal{T}_k) = 0$, then STOP.
4. (*"Mark"*) Choose a subset $\mathcal{T}_k^m \subset \mathcal{T}_k$ such that
$$\eta^2(u_k, \mathcal{T}_k^m) \geq \theta^2 \eta^2(u_k, \mathcal{T}_k). \tag{4.1}$$
5. (*"Refine"*) Perform a (minimal) conforming refinement of $\mathcal{T}_k$ using newest vertex bisections to obtain $\mathcal{T}_{k+1}$ such that each element of the neighborhood $N(\mathcal{T}_k^m) := \bigcup_{T \in \mathcal{T}_k^m} N_T \subset \mathcal{T}_k$ of $\mathcal{T}_k^m$ is fully refined in $\mathcal{T}_{k+1}$. This ensures that each $T \in N(\mathcal{T}_k^m)$ as well as each of its faces contain a node of $\mathcal{T}_{k+1}$ in their interior. Increment k and go to step 1.

Remark 14. Note that our marking strategy differs from the one proposed by Morin, Nochetto, and Siebert in [MNS00]. They used separate marking steps for the error estimator and the data oscillation. In our setting this would correspond to the following strategy: Construct $\mathcal{T}_k^m$ as in step *"Mark."* Then enlarge $\mathcal{T}_k^m$ such that for $\hat{\theta} > 0$ also
$$\operatorname{osc}^2(u_k, \mathcal{T}_k^m) \geq \hat{\theta}^2 \operatorname{osc}^2(u_k, \mathcal{T}_k). \tag{4.2}$$
This requires the calculation of the oscillation in step *"Estimate."* We want to point out that by the marking strategy (*"Mark"*) our new proof of convergence overcomes the drawback of additional marking for oscillation. This reflects the practical experience that the effect of oscillation plays a minor role (see, e.g., [MNS00]). We prove the success of most adaptive strategies which disregard the issue of oscillation altogether. Since $\eta^2(u_k, T) \geq \operatorname{osc}^2(u_k, T)$, this implies that (4.1) is equivalent to
$$\eta^2(u_k, \mathcal{T}_k^m) + \operatorname{osc}^2(u_k, \mathcal{T}_k^m) \geq \overline{\theta}^2 \big(\eta^2(u_k, \mathcal{T}_k) + \operatorname{osc}^2(u_k, \mathcal{T}_k)\big) \tag{4.3}$$
with $\overline{\theta} \in (0,1)$. Based on this cognition we give a new proof to show that the combination of energy difference and oscillation is reduced in each step (see Theorem 20).

Remark 15. Note that the condition in *"Refine"* of fully refined $T \in \mathcal{T}_k^m$ can be obtained by bisecting each $T \in \mathcal{T}_k^m$ three times in two dimensions (resp., six times in three dimensions); see [MNS00]. With this property we have a reduction factor $\lambda < 1$ of element size; i.e., if $T' \in \mathcal{T}_{k+1}$ is obtained by refining $T \in \mathcal{T}_k$, it holds that $h_{T'} \leq \lambda\, h_T$. By using the method of *newest vertex bisection* the shape regularity of $(\mathcal{T}_k)$ is uniformly bounded with respect to k depending on the shape regularity of $\mathcal{T}_0$.

4.2. Energy reduction in nested spaces. Assume as before that $V_H \subset V_h \subset V$. One main ingredient of proving linear convergence in [MNS00] for the linear case is the error reduction property for the energy norm
$$|||u_h - u|||^2 = |||u_H - u|||^2 - |||u_H - u_h|||^2.$$
This is a consequence of the Galerkin orthogonality and the Pythagorean theorem, which is related to Hilbert spaces. We do not have this property in the general case of the φ-Laplacian. But there is another way to interpret this property. In the linear case it is equivalent to
$$\mathcal{J}(u_h) - \mathcal{J}(u) = \mathcal{J}(u_H) - \mathcal{J}(u) - \big(\mathcal{J}(u_H) - \mathcal{J}(u_h)\big). \tag{4.4}$$

Obviously, this equality holds also in our case. Since $V_H \subset V_h \subset V$ and the minimizing property of u, u_h, and u_H we have

$$\mathcal{J}(u) \le \mathcal{J}(u_h) \le \mathcal{J}(u_H).$$

Thus we have a reduction of energy difference. Now it remains to find a link between the energy differences and the error. This is the content of the following lemma. For $v, w \in V$ we define the energy difference by

$$\varepsilon(v,w) := \mathcal{J}(v) - \mathcal{J}(w).$$

LEMMA 16. *Let u_1, u_2 be minimizers of the energy functional $\mathcal{J}$ with respect to the $V_1 \subset V_2 \subset V$. Then*

$$\mathcal{J}(u_1) - \mathcal{J}(u_2) \sim \|\mathbf{F}(\nabla u_1) - \mathbf{F}(\nabla u_2)\|_2^2,$$

where the constants depend only on $\Delta_2(\varphi)$, $\Delta_2(\varphi^*)$, and (2.7).

Proof. Define $\Phi : \mathbb{R}^d \to \mathbb{R}$ by $\Phi(\mathbf{a}) := \varphi(|\mathbf{a}|)$; then $\mathcal{J}(u) = \int_\Omega \Phi(\nabla u)\,dx - \int_\Omega u\,f\,dx$. Let $g(t) := \mathcal{J}([u_2,u_1]_t)$ for $t \in \mathbb{R}$, where $[u_2,u_1]_t := (1-t)\,u_2 + t\,u_1$. Since u_2 is the minimizer of $\mathcal{J}$ on $V_2 \supset V_1$, we have $g'(0) = 0$. We estimate by Taylor's formula

(4.5)

$$\mathcal{J}(u_1) - \mathcal{J}(u_2) = g(1) - g(0) = \frac{1}{2}\int_0^1 g''(t)\,(1-t)\,dt$$

$$= \frac{1}{2}\sum_{k,m}\int_0^1\int_\Omega (\partial_k\partial_m\Phi)([\nabla u_2, \nabla u_1]_t)\,(\partial_k u_1 - \partial_k u_2)(\partial_m u_1 - \partial_m u_2)\,dx\,(1-t)\,dt.$$

Note that for $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$

$$\sum_{k,m}(\partial_k\partial_m\Phi)(\mathbf{a})\,b_k\,b_m = \frac{\varphi'(|\mathbf{a}|)}{|\mathbf{a}|}\left(|\mathbf{b}|^2 - \frac{|\mathbf{a}\cdot\mathbf{b}|}{|\mathbf{a}|^2}\right) + \varphi''(|\mathbf{a}|)\frac{|\mathbf{a}\cdot\mathbf{b}|^2}{|\mathbf{a}|^2}.$$

By Assumption 1 we have $c\,\varphi'(t) \le t\,\varphi''(t) \le C\,\varphi'(t)$ uniformly in $t \ge 0$. Therefore,

$$\sum_{k,m}(\partial_k\partial_m\Phi)(\mathbf{a})\,b_k\,b_m \le (1+C)\,\frac{\varphi'(|\mathbf{a}|)}{|\mathbf{a}|}|\mathbf{b}|^2$$

and

$$\sum_{k,m}(\partial_k\partial_m\Phi)(\mathbf{a})\,b_k\,b_m \ge \frac{\varphi'(|\mathbf{a}|)}{|\mathbf{a}|}|\mathbf{b}|^2 + (c-1)\,\frac{\varphi'(|\mathbf{a}|)}{|\mathbf{a}|}\frac{|\mathbf{a}\cdot\mathbf{b}|^2}{|\mathbf{a}|^2} \ge c\,\frac{\varphi'(|\mathbf{a}|)}{|\mathbf{a}|}|\mathbf{b}|^2.$$

In other words $\sum_{k,m}(\partial_k\partial_m\Phi)(\mathbf{a})\,b_k\,b_m \sim \frac{\varphi'(|\mathbf{a}|)}{|\mathbf{a}|}|\mathbf{b}|^2$ uniformly in $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$. Combining the last estimates with (4.5) yields

$$\mathcal{J}(u_1) - \mathcal{J}(u_2) \sim \int_0^1\int_\Omega \frac{\varphi'(|[\nabla u_2, \nabla u_1]_t|)}{|[\nabla u_2, \nabla u_1]_t|}|\nabla u_1 - \nabla u_2|^2\,dx\,(1-t)\,dt. \tag{4.6}$$

Now we cite Lemma 19 from [DE08], which states that for any $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$

$$\int_0^1 \frac{\varphi'(|[\mathbf{a},\mathbf{b}]_t|)}{|[\mathbf{a},\mathbf{b}]_t|}\,dt \sim \frac{\varphi'(|\mathbf{a}|+|\mathbf{b}|)}{|\mathbf{a}|+|\mathbf{b}|}$$

with constants depending only on $\Delta_2(\varphi)$ and $\Delta_2(\varphi^*)$. In particular, this, (4.6), Assumption 1, and Lemma 3 give

$$\begin{aligned}\mathcal{J}(u_1)-\mathcal{J}(u_2) &\le C\int_\Omega \frac{\varphi'(|\nabla u_2|+|\nabla u_1|)}{|\nabla u_2|+|\nabla u_1|}\,|\nabla u_1-\nabla u_2|^2\,dx\\ &\le C\int_\Omega \varphi''(|\nabla u_2|+|\nabla u_1|)|\nabla u_1-\nabla u_2|^2\,dx\\ &\le C\int_\Omega |\mathbf{F}(\nabla u_1)-\mathbf{F}(\nabla u_2)|^2\,dx.\end{aligned}$$

On the other hand, (4.6), $\varphi'(t)\,t\sim\varphi(t)$ by (2.6), and Jensen's inequality give

$$\begin{aligned}\mathcal{J}(u_1)-\mathcal{J}(u_2) &\ge c\int_\Omega\int_0^1 \frac{\varphi(|[\nabla u_2,\nabla u_1]_t|)}{(|\nabla u_2|+|\nabla u_1|)^2}\,(1-t)\,dt\,|\nabla u_1-\nabla u_2|^2\,dx\\ &\ge c\int_\Omega \frac{\varphi(\int_0^1 |[\nabla u_2,\nabla u_1]_t|\,2\,(1-t)\,dt)}{(|\nabla u_2|+|\nabla u_1|)^2}\,|\nabla u_1-\nabla u_2|^2\,dx.\end{aligned}$$

Uniformly in $\mathbf{a},\mathbf{b}\in\mathbb{R}^d$ it holds that $\int_0^1 |[\mathbf{a},\mathbf{b}]_t|\,2(1-t)\,dt\sim|\mathbf{a}|+|\mathbf{b}|$, because both sides are a norm for the couple $(\mathbf{a},\mathbf{b})$. This and $\varphi''(t)\,t^2\sim\varphi(t)$ imply that

$$\mathcal{J}(u_1)-\mathcal{J}(u_2)\ge c\int_\Omega \varphi''(|\nabla u_2|+|\nabla u_1|)\,|\nabla u_1-\nabla u_2|^2\,dx.$$

Now Lemma 3 proves $\mathcal{J}(u_1)-\mathcal{J}(u_2)\ge c\int_\Omega |\mathbf{F}(\nabla u_1)-\mathbf{F}(\nabla u_2)|^2\,dx$. ∎

We recall that by Algorithm 13 for each element $T\in\mathcal{T}_k^m$ it holds that $N_T\subset N(\mathcal{T}_k^m)$. Thus each of the elements in N_T is marked for full refinement. Thus by Lemma 12 we have a discrete lower bound for each $T\in\mathcal{T}_k^m$. Summing over all sides in $\mathcal{T}_k^m$ yields together with Lemma 16 the following corollary.

COROLLARY 17. *For the sequence of finite element solutions produced by Algorithm* 13 *it holds that*

$$\eta^2(u_k,\mathcal{T}_k^m)\le C\,\varepsilon(u_k,u_{k+1})+C\,\operatorname{osc}^2(u_k,N(\mathcal{T}_k^m)),\tag{4.7}$$

where the constant C depends only on $\Delta_2(\varphi)$, $\Delta_2(\varphi^*)$, (2.7), and the shape regularity of $\mathcal{T}_0$.

4.3. Convergence. To prove that Algorithm 13 produces a sequence (u_k) which converges to the weak solution u of (1.1) we need an auxiliary lemma which deals with oscillation.

LEMMA 18. *Then for the sequence of finite element solutions produced by Algorithm* 13 *there exists $\rho\in(0,1)$ such that*

$$\operatorname{osc}^2(u,\mathcal{T}_{k+1})\le\operatorname{osc}^2(u,\mathcal{T}_k)-\rho\,\operatorname{osc}^2(u,N(\mathcal{T}_k^m)),\tag{4.8}$$

with $\rho=1-\lambda$ and λ from Remark 15.

Proof. Recall that for any $T \in \mathcal{T}_k$ and any $T' \in \mathcal{T}_{k+1}$ with $T' \subset T$ we have $h_{T'} \leq h_T$. Moreover, if T is refined in $\mathcal{T}_{k+1}$, then we even have $h_{T'} \leq \lambda\, h_T$. In particular, this holds if $T \in \mathcal{T}_k^m$ since these T are fully refined in $\mathcal{T}_{k+1}$. First, for every $T \in \mathcal{T}_k$ we estimate

$$\begin{aligned}
\sum_{T' \in \mathcal{T}_{k+1}, T' \subset T} \operatorname{osc}^2(u, T') &= \sum_{T' \in \mathcal{T}_{k+1}, T' \subset T} \inf_{f_{T'} \in \mathbb{R}} \int_{T'} (\varphi_{|\nabla u|})^*(h_{T'}\, |f - f_{T'}|)\, dx \\
&\leq \sum_{T' \in \mathcal{T}_{k+1}, T' \subset T} \inf_{f_{T'} \in \mathbb{R}} \int_{T'} (\varphi_{|\nabla u|})^*(h_T\, |f - f_{T'}|)\, dx \\
&\leq \inf_{f_T \in \mathbb{R}} \int_T (\varphi_{|\nabla u|})^*(h_T\, |f - f_T|)\, dx \\
&= \operatorname{osc}^2(u, T).
\end{aligned} \tag{4.9}$$

Second, for $T \in N(\mathcal{T}_k^m)$ we have a better estimate, since all elements in $N(\mathcal{T}_k^m)$ are fully refined in $\mathcal{T}_{k+1}$. Therefore, it holds that with the convexity of N-functions

$$\begin{aligned}
\sum_{T' \in \mathcal{T}_{k+1}, T' \subset T} \operatorname{osc}^2(u, T') &= \sum_{T' \in \mathcal{T}_{k+1}, T' \subset T} \inf_{f_{T'} \in \mathbb{R}} \int_{T'} (\varphi_{|\nabla u|})^*(h_{T'}\, |f - f_{T'}|)\, dx \\
&\leq \sum_{T' \in \mathcal{T}_{k+1}, T' \subset T} \inf_{f_{T'} \in \mathbb{R}} \int_{T'} (\varphi_{|\nabla u|})^*(\lambda\, h_T\, |f - f_{T'}|)\, dx \\
&\leq \inf_{f_T \in \mathbb{R}} \int_T (\varphi_{|\nabla u|})^*(\lambda\, h_T\, |f - f_T|)\, dx \\
&\leq \lambda \operatorname{osc}^2(u, T).
\end{aligned} \tag{4.10}$$

Now (4.9) and (4.10) imply that

$$\begin{aligned}
\operatorname{osc}^2(u, \mathcal{T}_{k+1}) &\leq \operatorname{osc}^2(u, \mathcal{T}_k \setminus \mathcal{T}_k^m) + \lambda \operatorname{osc}^2(u, \mathcal{T}_k^m) \\
&= \operatorname{osc}^2(u, \mathcal{T}_k) - (1 - \lambda) \operatorname{osc}^2(u, \mathcal{T}_k^m).
\end{aligned}$$

This proves the lemma. ∎

LEMMA 19. *For the sequence of finite element solutions produced by Algorithm* 13 *it holds that*

$$\eta^2(u_k, \mathcal{T}_k) + \operatorname{osc}^2(u_k, \mathcal{T}_k) \sim \varepsilon(u_k, u) + \operatorname{osc}^2(u_k, \mathcal{T}_k) \tag{4.11a}$$

$$\sim \varepsilon(u_k, u) + \operatorname{osc}^2(u, \mathcal{T}_k). \tag{4.11b}$$

Proof. First we prove (4.11a). From Corollary 11, Lemma 16, and Lemma 8 it follows that

$$\begin{aligned}
\eta^2(u_k, \mathcal{T}_k) &\leq C\, \varepsilon(u_k, u) + C \operatorname{osc}^2(u_k, \mathcal{T}_k), \\
\varepsilon(u_k, u) &\leq C\, \eta^2(u_k, \mathcal{T}_k) + C \operatorname{osc}^2(u_k, \mathcal{T}_k),
\end{aligned}$$

which immediately implies (4.11a). With the help of Corollary 28 we can change the shift in the last term, i.e.,

$$\begin{aligned}
&\operatorname{osc}^2(u_k,\mathcal{T}_k)\\
&\quad=\sum_{T\in\mathcal{T}_k}\inf_{f_T\in\mathbb{R}}\int_T(\varphi_{|\nabla u_k|})^*(h_K\,|f-f_T|)\,dx\\
&\quad\le\sum_{T\in\mathcal{T}_k}\left(\inf_{f_T\in\mathbb{R}}C\int_T(\varphi_{|\nabla u|})^*(h_K\,|f-f_T|)\,dx+C\,|\mathbf{F}(\nabla u_k)-\mathbf{F}(\nabla u)|^2\right)\\
&\quad=C\operatorname{osc}^2(u,\mathcal{T}_k)+C\,\|\mathbf{F}(\nabla u_k)-\mathbf{F}(\nabla u)\|_2^2\le C\operatorname{osc}^2(u,\mathcal{T}_k)+C\,\varepsilon(u_k,u).
\end{aligned}\tag{4.12}$$

The same calculation with u_k and u exchanged proves

$$\operatorname{osc}^2(u,\mathcal{T}_k)\le C\operatorname{osc}^2(u_k,\mathcal{T}_k)+C\,\varepsilon_k(u_k,u).\tag{4.13}$$

Now (4.12) and (4.13) prove (4.11b). ∎

Now we are able to prove our main result.

THEOREM 20 (energy/oscillation reduction). *There exists a constant $\alpha\in(0,1)$ such that for the sequence (u_k) of finite element solutions produced by Algorithm* AFEM

$$\varepsilon(u_{k+1},u)+\operatorname{osc}^2(u,\mathcal{T}_{k+1})\le\alpha^2\big(\varepsilon(u_k,u)+\operatorname{osc}^2(u,\mathcal{T}_k)\big).\tag{4.14}$$

Here α depends only on the shape regularity of the sequence $(\mathcal{T}_k)$, $\Delta_2(\varphi)$, $\Delta_2(\varphi^*)$, (2.7), and θ from the marking strategy (4.1). *In particular, for $k\in\mathbb{N}$*

$$\varepsilon(u_k,u)+\operatorname{osc}^2(u,\mathcal{T}_k)\le\alpha^{2k}\big(\varepsilon(u_0,u)+\operatorname{osc}^2(u,\mathcal{T}_0)\big).\tag{4.15}$$

Proof. We start with the energy reduction (4.4)

$$\varepsilon(u_{k+1},u)=\varepsilon(u_k,u)-\varepsilon(u_k,u_{k+1})\tag{4.16}$$

and note that by Lemma 18 we have a similar relation for oscillation; i.e.,

$$\operatorname{osc}^2(u,\mathcal{T}_{k+1})\le\operatorname{osc}^2(u,\mathcal{T}_k)-\rho\operatorname{osc}^2(u,N(\mathcal{T}_k^m))$$

for some $\rho\in(0,1)$. Summing these two terms yields

$$\begin{aligned}
&\varepsilon(u_{k+1},u)+\operatorname{osc}^2(u,\mathcal{T}_{k+1})\\
&\qquad\le\varepsilon(u_k,u)+\operatorname{osc}^2(u,\mathcal{T}_k)-\big(\varepsilon(u_k,u_{k+1})+\rho\operatorname{osc}^2\big(u,N(\mathcal{T}_k^m)\big)\big).
\end{aligned}\tag{4.17}$$

The main idea is to apply the discrete lower bound to $\varepsilon(u_k,u_{k+1})$, whereas the oscillation on the right-hand side of (3.30) shall be absorbed at the oscillation term in the above inequality. Since in (3.30) the oscillation appears with shift u_k, we use Corollary 28 (change of shift) as in (4.12) to obtain,

$$\operatorname{osc}^2\big(u_k,N(\mathcal{T}_k^m)\big)\le(1+C_\delta)C\operatorname{osc}^2\big(u,N(\mathcal{T}_k^m)\big)+\delta\,C\,\varepsilon_k(u_k,u)$$

for $\delta>0$ and a corresponding $C_\delta\ge0$. We set $D=(1+C_\delta)>1$ and obtain, equivalently,

$$\frac{1}{C\,D}\operatorname{osc}^2\big(u_k,N(\mathcal{T}_k^m)\big)-\frac{\delta}{D}\,\varepsilon(u_k,u)\le\operatorname{osc}^2\big(u,N(\mathcal{T}_k^m)\big).$$

This, (4.17), and (4.7) imply that

$$\begin{aligned}
&\varepsilon(u_{k+1},u)+\operatorname{osc}^2\big(u,\mathcal{T}_{k+1}\big)\\
&\quad\le \varepsilon(u_k,u)+\operatorname{osc}^2(u,\mathcal{T}_k)-\bigg(\varepsilon(u_k,u_{k+1})+\frac{\rho}{CD}\operatorname{osc}^2\big(u_k,N(\mathcal{T}_k^m)\big)-\frac{\rho\,\delta}{D}\,\varepsilon(u_k,u)\bigg)\\
&\quad\le \bigg(1+\frac{\rho\,\delta}{D}\bigg)\,\varepsilon(u_k,u)+\operatorname{osc}^2(u,\mathcal{T}_k)-\bigg(\varepsilon(u_k,u_{k+1})+\frac{\rho}{CD}\operatorname{osc}^2\big(u_k,N(\mathcal{T}_k^m)\big)\bigg).
\end{aligned}$$

Since $D\ge 1$ and $\rho\in(0,1)$ we have $\frac{\rho}{2\,C\,D}\le 1$. Thus

$$\begin{aligned}
&\varepsilon(u_{k+1},u)+\operatorname{osc}^2(u,\mathcal{T}_{k+1})\\
&\quad\le \bigg(1+\frac{\rho\,\delta}{D}\bigg)\,\varepsilon(u_k,u)+\operatorname{osc}^2(u,\mathcal{T}_k)-\frac{\rho}{2\,CD}\bigg(\varepsilon(u_k,u_{k+1})+2\operatorname{osc}^2\big(u_k,N(\mathcal{T}_k^m)\big)\bigg),
\end{aligned}$$

the discrete lower estimate (4.7), and $\mathcal{T}_k^m\subset N(\mathcal{T}_k^m)$ yield

$$\begin{aligned}
&\le \bigg(1+\frac{\rho\,\delta}{D}\bigg)\,\varepsilon(u_k,u)+\operatorname{osc}^2(u,\mathcal{T}_k)-\frac{\rho}{2\,CD}\bigg(\frac{1}{C}\,\eta^2(u_k,\mathcal{T}_k^m)+\operatorname{osc}^2\big(u_k,N(\mathcal{T}_k^m)\big)\bigg)\\
&\le \bigg(1+\frac{\rho\,\delta}{D}\bigg)\varepsilon(u_k,u)+\operatorname{osc}^2(u,\mathcal{T}_k)-\frac{\rho}{2\,CD}\bigg(\eta^2(u_k,\mathcal{T}_k^m)+\operatorname{osc}^2(u_k,\mathcal{T}_k^m)\bigg).
\end{aligned}$$

Now it remains to utilize the marking strategy. By the equivalent form (4.3) of our marking strategy *"Mark"* and (4.11) we get

$$C\,\big(\varepsilon(u_k,u)+\operatorname{osc}_k^2(u,\mathcal{T}_k)\big)\le\overline{\theta}^2\,\big(\eta^2(u_k,\mathcal{T}_k^m)+\operatorname{osc}_k^2(u_k,\mathcal{T}_k^m)\big).\tag{4.18}$$

This observation and Lemma 8 finally imply that

$$\begin{aligned}
&\varepsilon(u_{k+1},u)+\operatorname{osc}^2(u,\mathcal{T}_{k+1})\\
&\quad\le \bigg(1+\frac{\rho\,\delta}{D}\bigg)\,\varepsilon(u_k,u)+\operatorname{osc}^2(u,\mathcal{T}_k)-\frac{\rho\,\overline{\theta}^2}{2\,CD}\bigg(\eta^2(u_k,\mathcal{T}_k)+\operatorname{osc}^2(u,\mathcal{T}_k)\bigg)\\
&\quad\le \bigg(1+\frac{\rho\,\delta}{D}\bigg)\,\varepsilon(u_k,u)+\operatorname{osc}^2(u,\mathcal{T}_k)-\frac{\rho\,\overline{\theta}^2}{2\,CD}\bigg(\frac{1}{C}\,\varepsilon(u_k,u)+\operatorname{osc}^2(u,\mathcal{T}_k)\bigg)\\
&\quad\le \bigg(1+\frac{\rho\,\delta}{D}-\frac{\rho\,\overline{\theta}^2}{2\,CD}\bigg)\,\varepsilon(u_k,u)+\bigg(1-\frac{\rho\,\overline{\theta}^2}{2\,CD}\bigg)\operatorname{osc}^2(u,\mathcal{T}_k).
\end{aligned}$$

Substituting D by its definition, we set

$$\alpha^2:=\min\left\{1+\frac{\rho\,\delta}{(1+C_\delta)}-\frac{\rho\,\overline{\theta}^2}{2\,C\,(1+C_\delta)},1-\frac{\rho\,\overline{\theta}^2}{2\,C\,(1+C_\delta)}\right\}$$

and obtain the proposition for $\delta>0$ small enough. ∎

From the energy/oscillation reduction of Theorem 20 the next corollary follows with the help of Lemma 16.

COROLLARY 21 (energy/oscillation reduction). *For the sequence (u_k) of finite element solutions produced by Algorithm* AFEM *there exists constant $\alpha \in (0,1)$ such that for all $k \in \mathbb{N}$*

$$\begin{aligned} &\|\mathbf{F}(\nabla u_k) - \mathbf{F}(\nabla u)\|^2_{L^2(\Omega)} + \operatorname{osc}^2(u, \mathcal{T}_k) \\ &\qquad \le C\,\alpha^{2k}\left(\|\mathbf{F}(\nabla u_0) - \mathbf{F}(\nabla u)\|^2_{L^2(\Omega)} + \operatorname{osc}^2(u, \mathcal{T}_0)\right). \end{aligned} \tag{4.19}$$

Here α and the constants depend only on the shape regularity of $\mathcal{T}_0$, $\Delta_2(\varphi)$, $\Delta_2(\varphi^*)$, (2.7), and θ from the marking strategy (4.1).

5. Appendix. In this section we summarize the properties of the *shifted* N-functions φ_a. Recall that for given N-function φ with $\Delta_2(\varphi), \Delta_2(\varphi^*) < \infty$ we define φ_a as in (3.2) and $\mathbf{F}$ as in (3.1). The following results are from [DE08], and we present them here without proof. We use $C \ge 1$ as a generic constant that does solely depend on the Δ_2-constants of φ (resp., φ^*) and may change from line to line.

LEMMA 22. *Let φ be an N-function with $\Delta_2(\varphi), \Delta_2(\varphi^*) < \infty$. Then for all $a \ge 0$ the functions φ_a and $(\varphi_a)^*$ are N-functions. Moreover, the families φ_a and $(\varphi_a)^*$ satisfy the Δ_2 condition uniformly in $\alpha \ge 0$, i.e., $c_0 := \sup_{a\ge 0}(\Delta_2(\varphi_a), \Delta_2((\varphi_a)^*)) < \infty$. The constant c_0 depends on φ only by $\Delta_2(\varphi)$ and $\Delta_2(\varphi^*)$. Moreover,*

$$(\varphi_a)^*(t) \sim (\varphi^*)_{\varphi'(a)}(t) \tag{5.1}$$

uniformly in $a, t \ge 0$, where the constants depend only on $\Delta_2(\varphi)$ and $\Delta_2(\varphi^*)$.

LEMMA 23. *Let φ be an N-function with $\Delta_2(\varphi), \Delta_2(\varphi^*) < \infty$. Then uniformly in $\mathbf{a} \in \mathbb{R}^d$*

$$\begin{aligned} \varphi''(|\mathbf{a}| + |\mathbf{b}|)\,|\mathbf{a}-\mathbf{b}| &\sim \varphi'_{|\mathbf{a}|}(|\mathbf{a}-\mathbf{b}|) \sim \varphi'_{|\mathbf{b}|}(|\mathbf{a}-\mathbf{b}|), \\ \varphi''(|\mathbf{a}| + |\mathbf{b}|)\,|\mathbf{a}-\mathbf{b}|^2 &\sim \varphi_{|\mathbf{a}|}(|\mathbf{a}-\mathbf{b}|) \sim \varphi_{|\mathbf{b}|}(|\mathbf{a}-\mathbf{b}|), \end{aligned} \tag{5.2}$$

with constants depending only on $\Delta_2(\varphi)$ and $\Delta_2(\varphi^*)$.

LEMMA 24. *Let φ be as in Assumption* 1. *Then φ^* satisfies Assumption* 1, *too. If we define the N-function ψ by*

$$\frac{\psi'(t)}{t} := \left(\frac{\varphi'(t)}{t}\right)^{\frac{1}{2}},$$

then ψ and ψ^* satisfy Assumption 1. *Moreover, $\psi''(t) \sim \sqrt{\varphi''(t)}$ uniformly in $t > 0$. The Δ_2-constants of ψ and ψ^* and the constants of $\psi''(t) \sim \sqrt{\varphi''(t)}$ depend only on $\Delta_2(\varphi)$.*

The following lemma is proven already in [DR07], but we give a shorter proof.

LEMMA 25. *Let φ be an N-function with $\Delta_2(\varphi), \Delta_2(\varphi^*) < \infty$. Then there exists $C \ge 1$, which depends only on $\Delta_2(\varphi)$ such that for all $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$ and $t \ge 0$*

$$\varphi'_{|\mathbf{a}|}(t) \le C\,\varphi'_{|\mathbf{b}|}(t) + C\,\varphi'_{|\mathbf{a}|}(|\mathbf{a}-\mathbf{b}|). \tag{5.3}$$

Proof. Since $\varphi'_{|\mathbf{a}|}(t) \sim \varphi_{|\mathbf{a}|}(t)/t$ and $\varphi_{|\mathbf{a}|}(2t) \sim \varphi_{|\mathbf{a}|}(t)$, we have $\varphi'_{|\mathbf{a}|}(2t) \sim \varphi'_{|\mathbf{a}|}(t)$. In particular, $\varphi'_{|\mathbf{a}|}(2t) \le C\,\varphi'_{|\mathbf{a}|}(t)$ for some $C > 0$ uniformly in $t \ge 0$. All constants depend only on $\Delta_2(\varphi_{|\mathbf{a}|})$, and so by Lemma 22 the constants depend only on $\Delta_2(\varphi)$.

Case $|\mathbf{a}-\mathbf{b}| \le \frac{1}{2}\,t$. From $|\mathbf{a}-\mathbf{b}| \le \frac{1}{2}\,t$ it follows that $0 \le \frac{1}{2}(|\mathbf{b}|+t) \le |\mathbf{a}|+t \le 2(|\mathbf{b}|+t)$. Hence,

$$\varphi'_{|\mathbf{a}|}(t) = \frac{\varphi'(|\mathbf{a}|+t)}{|\mathbf{a}|+t}\,t \le \frac{\varphi'(2\,(|\mathbf{b}|+t))}{\frac{1}{2}(|\mathbf{b}|+t)}\,t \le 2\,C\,\frac{\varphi'(|\mathbf{b}|+t)}{|\mathbf{b}|+t}\,t = 2\,C\,\varphi'_{|\mathbf{b}|}(t).$$

Case $|\mathbf{a}-\mathbf{b}| \ge \frac{1}{2}\,t$. We estimate

$$\varphi'_{|\mathbf{a}|}(t) \le \varphi'_{|\mathbf{a}|}(2\,|\mathbf{a}-\mathbf{b}|) \le C\,\varphi'_{|\mathbf{a}|}(|\mathbf{a}-\mathbf{b}|).$$

Combining the two cases proves the lemma. ∎

COROLLARY 26 (change of shift). *Let φ be an N-function with $\Delta_2(\varphi), \Delta_2(\varphi^*) < \infty$. Then for any $\delta > 0$ there exists $C_\delta > 0$, which depends only on δ and $\Delta_2(\varphi)$ such that for all $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$ and $t \ge 0$*

$$\varphi_{|\mathbf{a}|}(t) \le (1+C_\delta)C\,\varphi_{|\mathbf{b}|}(t) + \delta\,\varphi_{|\mathbf{a}|}(|\mathbf{a}-\mathbf{b}|). \tag{5.4}$$

If φ satisfies Assumption 1, *then for any $\delta > 0$ there exists $C_\delta > 0$, which depends only on δ, $\Delta_2(\varphi)$, $\Delta_2(\varphi^*)$, and* (2.7) *such that for all $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$ and $t \ge 0$*

$$\varphi_{|\mathbf{a}|}(t) \le (1+C_\delta)C\,\varphi_{|\mathbf{b}|}(t) + \delta\,C\,\big|\mathbf{F}(\mathbf{a}) - \mathbf{F}(\mathbf{b})\big|^2. \tag{5.5}$$

Proof. Due to (2.7) it holds that $\varphi_{|\mathbf{a}|}(t) \sim \varphi'_{|\mathbf{a}|}(t)\,t$. Now inequality (5.4) follows by (5.3), (5.2), an application of Young's inequality (2.3), (5.2), and (2.5):

$$\begin{aligned}
\varphi_{|\mathbf{a}|}(t) &\le C\,\varphi'_{|\mathbf{a}|}(t)\,t \\
&\le C\,\varphi'_{|\mathbf{b}|}(t)\,t + C\,\varphi'_{|\mathbf{a}|}(|\mathbf{a}-\mathbf{b}|)\,t \\
&\le C\,\varphi'_{|\mathbf{b}|}(t)\,t + C\,\varphi'_{|\mathbf{b}|}(|\mathbf{a}-\mathbf{b}|)\,t \\
&\le C\,\varphi_{|\mathbf{b}|}(t) + C\,\delta\,(\varphi_{|\mathbf{b}|})^*\big(\varphi'_{|\mathbf{b}|}(|\mathbf{a}-\mathbf{b}|)\big) + C\,C_\delta\,\varphi_{|\mathbf{b}|}(t) \\
&\le C\,C_\delta\,\varphi_{|\mathbf{b}|}(t) + C\,\delta\,\varphi_{|\mathbf{b}|}(|\mathbf{a}-\mathbf{b}|) + C\,C_\delta\,\varphi_{|\mathbf{b}|}(t).
\end{aligned}$$

On the other hand, (5.5) follows from (5.4) with the help of Lemma 3. ∎

The following lemma is new. It generalizes the *change of shift* to complementary functions.

LEMMA 27. *If φ satisfies Assumption* 1, *then for any $\delta > 0$ there exists $C_\delta > 0$, which depends only on δ, $\Delta_2(\varphi)$, and $\Delta_2(\varphi^*)$ such that for all $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$ and $t \ge 0$*

$$\big((\varphi_{|\mathbf{a}|})^*\big)'(t) \le C\,\big((\varphi_{|\mathbf{b}|})^*\big)'(t) + C\,|\mathbf{a}-\mathbf{b}|. \tag{5.6}$$

Proof. With Lemma 22, $|\mathbf{A}(\mathbf{a})| = \varphi'(|\mathbf{a}|)$, Lemma 25, Lemma 3, and $\Delta_2(\varphi) < \infty$ we estimate that

$$\begin{aligned}
\big((\varphi_{|\mathbf{a}|})^*\big)'(t) &\le C\,\big((\varphi^*)_{|\mathbf{A}(\mathbf{a})|}\big)'(t) \\
&\le C\,\big((\varphi^*)_{|\mathbf{A}(\mathbf{a})|}\big)'(t) + C\,\big((\varphi^*)_{|\mathbf{A}(\mathbf{a})|}\big)'(|\mathbf{A}(\mathbf{a})-\mathbf{A}(\mathbf{b})|) \\
&\le C\,\big((\varphi_{|\mathbf{a}|})^*\big)'(t) + C\,\big((\varphi_{|\mathbf{a}|})^*\big)'(C\,\varphi'_{|\mathbf{a}|}(|\mathbf{a}-\mathbf{b}|)) \\
&\le C\,\big((\varphi_{|\mathbf{a}|})^*\big)'(t) + C\,\big((\varphi_{|\mathbf{a}|})^*\big)'(\varphi'_{|\mathbf{a}|}(|\mathbf{a}-\mathbf{b}|)) \\
&= C\,\big((\varphi_{|\mathbf{a}|})^*\big)'(t) + C\,|\mathbf{a}-\mathbf{b}|.
\end{aligned}$$

This proves the lemma. ∎

COROLLARY 28 (change of shift $*$). *If φ satisfies Assumption* 1, *then for any $\delta > 0$ there exists $C_\delta > 0$, which depends only on δ, $\Delta_2(\varphi)$, and $\Delta_2(\varphi^*)$ such that for all $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$ and $t \geq 0$*

$$(\varphi_{|\mathbf{a}|})^*(t) \leq (1 + C_\delta) C\, (\varphi_{|\mathbf{b}|})^*(t) + \delta\, C\, \varphi_{|\mathbf{a}|}(|\mathbf{a} - \mathbf{b}|). \tag{5.7}$$

If φ satisfies Assumption 1, *then for any $\delta > 0$ there exists $C_\delta > 0$, which depends only on δ, $\Delta_2(\varphi)$, $\Delta_2(\varphi^*)$, and* (2.7) *such that for all $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$ and $t \geq 0$*

$$(\varphi_{|\mathbf{a}|})^*(t) \leq (1 + C_\delta) C\, (\varphi_{|\mathbf{b}|})^*(t) + \delta\, C\, |\mathbf{F}(\mathbf{a}) - \mathbf{F}(\mathbf{b})|^2. \tag{5.8}$$

Proof. Due to (2.7) it holds that $(\varphi_{|a|})^*(t) \sim ((\varphi_{|a|})^*)'(t)\, t$. Now inequality (5.7) follows by (5.6), an application of Young's inequality, (2.3), (5.2), and (2.5), similarly to the proof of Corollary 26. On the other hand, (5.8) follows from (5.7) with the help of Lemma 3. ∎

REFERENCES

[BL93a] J. W. BARRETT AND W. B. LIU, *Finite element approximation of the p-Laplacian*, Math. Comp., 61 (1993), pp. 523–537.

[BL94a] J. W. BARRETT AND W. B. LIU, *Finite element approximation of degenerate quasilinear elliptic and parabolic problems*, in Numerical Analysis 1993 (Dundee, 1993), Pitman Res. Notes Math. Ser. 303, Longman Scientific and Technical, Harlow, UK, 1994, pp. 1–16.

[BL94b] J. W. BARRETT AND W. B. LIU, *Finite element approximation of the parabolic p-Laplacian*, SIAM J. Numer. Anal., 31 (1994), pp. 413–428.

[CK06] C. CARSTENSEN AND R. KLOSE, *A posteriori finite element error control for the p-Laplace problem*, SIAM J. Sci. Comput., 25 (2003), pp. 792–814.

[CLY06] C. CARSTENSEN, W. B. LIU, AND N. YAN, *A posteriori FE error control for p-Laplacian by gradient recovery in quasi-norm*, Math. Comp., 75 (1994), pp. 1599–1616.

[DE08] L. DIENING AND F. ETTWEIN, *Fractional estimates for non-differentiable elliptic systems with general growth*, Forum Math., 20 (2008), pp. 523–556.

[Dör96] W. DÖRFLER, *A convergent adaptive algorithm for Poisson's equation*, SIAM J. Numer. Anal., 33 (1996), pp. 1106–1124.

[DR07] L. DIENING AND M. RŮŽIČKA, *Interpolation operators in Orlicz-Sobolev spaces*, Numer. Math., 107 (2007), pp. 107–129.

[Ebm05] C. EBMEYER, *Gobal regularity in Sobolev spaces for elliptic problems with p-structure on bounded domains*, in Trends in Partial Differential Equations of Mathematical Physics, Progr. Nonlinear Differential Equations Appl. 61, Birkhäuser, Basel, 2005, pp. 81–89.

[EL05] C. EBMEYER AND W. B. LIU, *Quasi-norm interpolation error estimates for the piecewise linear finite element approximation of p-Laplacian problems*, Numer. Math., 100 (2005), pp. 233–258.

[ELS05] C. EBMEYER, W. B. LIU, AND M. STEINHAUER, *Global regularity in fractional order Sobolev spaces for the p-Laplace equation on polyhedral domains*, Z. Anal. Anwendungen, 24 (2005), pp. 353–374.

[Giu03] E. GIUSTI, *Direct Methods in the Calculus of Variations*, World Scientific, River Edge, NJ, 2003.

[GM75] R. GLOWINSKI AND A. MAROCCO, *Sur l'approximation par éléments finis d'ordre un, et la résolution, par pénalisation-dualité, d'une classe de problèmes de Dirichlet non lineaires*, Rev. Française Automat. Informat. Recherche Opérationelle Ser. Rouge Anal. Numér., 9 (1975), pp. 41–76.

[LY01] W. LIU AND N. YAN, *Quasi-norm local error estimators for p-Laplacian*, SIAM J. Numer. Anal., 39 (2001), pp. 100–127.

[LY02] W. LIU AND N. YAN, *On quasi-norm interpolation error estimation and a posteriori error estimates for p-Laplacian*, SIAM J. Numer. Anal., 40 (2002), pp. 1870–1895.

[MN05] K. MEKCHAY AND R. H. NOCHETTO, *Convergence of adaptive finite element methods for general second order linear elliptic PDEs*, SIAM J. Numer. Anal., 43 (2005), pp. 1803–1827.

[MNS00] P. Morin, R. H. Nochetto, and K. G. Siebert, *Data oscillation and convergence of adaptive FEM*, SIAM J. Numer. Anal., 38 (2000), pp. 466–488.

[RR91] M. M. Rao and Z. D. Ren, *Theory of Orlicz Spaces*, Monogr. Textbooks Pure Appl. Math. 146, Marcel Dekker, New York, 1991.

[SZ90] L. R. Scott and S. Zhang, *Finite element interpolation of nonsmooth functions satisfying boundary conditions*, Math. Comp., 54 (1990), pp. 483–493.

[Ve02] A. Veeser, *Convergent adaptive finite elements for the nonlinear Laplacian*, Numer. Math., 92 (2002), pp. 743–770.

SIAM J. NUMER. ANAL.
Vol. 46, No. 2, pp. 639–670
© 2008 Society for Industrial and Applied Mathematics

CONVERGENCE OF A HIGH-ORDER SEMI-LAGRANGIAN SCHEME WITH PROPAGATION OF GRADIENTS FOR THE ONE-DIMENSIONAL VLASOV–POISSON SYSTEM*

NICOLAS BESSE†

Abstract. In this paper we give a proof of convergence of a new numerical method introduced in [N. Besse and E. Sonnendrücker, *J. Comput. Phys.*, 191 (2003), pp. 341–376] for the Vlasov equation. The numerical method is based on the semi-Lagrangian principle and the transport of the gradient of the statistical distribution function in order to get a high-order and stable reconstruction. These kinds of new schemes have been successfully implemented on unstructured meshes of four-dimensional phase space (cf. [N. Besse, *Etude mathématique et numérique de l'equation de Vlasov sur des maillages non structurés de l'espace des phases*, thèse de l'Université Louis Pasteur, Strasbourg, France, 2003; N. Besse, J. Segré, and E. Sonnendrücker, *Transport Theory Statist. Phys.*, 34 (2005), pp. 311–332]). In order to make a rigorous proof of convergence of this method and simplify the convergence analysis, we have considered the periodic one-dimensional Vlasov–Poisson system in phase space on a grid. The distribution $f(t,x,v)$ and the electric field are shown to converge to the exact solution values in H^1 norm. The rate of convergence is of $\mathcal{O}(\Delta t^2 + \frac{\Delta x^{4-|\alpha|}}{\Delta t} + \frac{\Delta v^{4-|\alpha|}}{\Delta t})$, $\alpha \in \mathbb{N}^2$, $|\alpha| \leq 1$.

Key words. Vlasov–Poisson system, semi-Lagrangian methods, high-order schemes, convergence analysis

AMS subject classifications. 65M12, 65M15, 82D10

DOI. 10.1137/050635171

1. Introduction.

The numerical resolution of the Vlasov equation is most of the time performed by Lagrangian methods such as particle-in-cell methods which consist of approximating the plasma by a finite number of macroparticles (see Birdsall and Langdon [10] for more details). Although this method allows us to obtain satisfying results with a small number of particles, it is well known that the numerical noise inherent to the particle method becomes too significant to allow a precise description of the tail of the distribution function, which plays an important role in charged particle beams. To remedy this problem, Eulerian methods which consist in discretizing the Vlasov equation on a mesh of phase space have been proposed. One of these methods is semi-Lagrangian methods, which have efficiently been implemented using parallel computers [15] and give considerable promise in displaying the detailed structure of distribution functions in weak densities regions.

The author extends semi-Lagrangian schemes on unstructured meshes on two-, three-, and four-dimensional phase spaces with different kinds of high-order local interpolation operators and with the possibility of having a positive and conservative method by introducing a linear combination of low-order solution and high-order solution tempered by a limiter coefficient (cf. [5]). The interpolation operator considered involves the knowledge of the gradient of the distribution function which is obtained by solving a transport problem on the gradient or by differentiating the convected solution. For more details on how we get these schemes we refer the reader to [5, 8, 9].

*Received by the editors July 5, 2005; accepted for publication (in revised form) August 22, 2007; published electronically February 1, 2008.
http://www.siam.org/journals/sinum/46-2/63517.html
†Institut de Mathématiques Elie Cartan, UMR CNRS 7502, Université Henri Poincaré Nancy-I, B.P. 239, 54506 Vandoeuvre-lès-Nancy Cedex, France (besse@iecn.u-nancy.fr), and Laboratoire de Physique des Milieux Ionisés, UMR CNRS 7040, Université Henri Poincaré Nancy-I, B.P. 239, 54506 Vandoeuvre-lès-Nancy Cedex, France (Nicolas.Besse@lpmi.uhp-nancy.fr).

To simplify the convergence analysis we have considered this scheme on a uniform grid, without limiter coefficient, for the one-dimensional Vlasov–Poisson system.

Let us note that a first work on the convergence of one-dimensional particle methods is stated in [23], where Neunzert and Wick consider nonuniform initial loadings of particles asymptotically distributed with respect to initial data. Cottet and Raviart [14] present a mathematical analysis of the particle method for solving the one-dimensional Vlasov–Poisson system where uniform initial loadings of particles are considered. Then a lot of authors have studied the convergence of particle methods for the multidimensional Vlasov–Poisson system [18, 27, 28, 29]. They have also proved convergence results on random and deterministic particle methods for the Vlasov–Poisson–Fokker–Planck kinetic equations [21, 22]. Moreover Glassey and Schaeffer have done the convergence analysis of a particle method for the relativistic Vlasov–Maxwell system [20]. Schaeffer [24] has proved the convergence of a finite difference scheme for the one-dimensional Vlasov–Poisson–Fokker–Planck system, and Filbet [17] has shown the convergence of a finite volume scheme for the one-dimensional Vlasov–Poisson system. Finally the author [7] has proved the convergence of a semi-Lagrangian scheme on unstructured meshes (triangulation) of phase space in which a linear reconstruction is used, for the one-dimensional Vlasov–Poisson system. In the latter the author obtains a convergence rate in $\mathcal{O}(\Delta t^2 + h^2 + h^2/\Delta t)$ when $f \in \mathscr{C}_c^2([0,T] \times \mathbb{R}_x \times \mathbb{R}_v)$ (best convergence rate is in $\mathcal{O}(h^{4/3})$ when $\Delta t = h^{2/3}$) and a rate in $\mathcal{O}(\Delta t + h + h/\Delta t)$ when $f \in W_c^{1,\infty}([0,T] \times \mathbb{R}_x \times \mathbb{R}_v)$ (best convergence rate is in $\mathcal{O}(h^{1/2})$ when $\Delta t = h^{1/2}$).

Although a lot of papers present satisfactory numerical results using semi-Lagrangian methods [1, 2, 5, 13, 16, 25, 26], few rigorous mathematical results on convergence analysis of semi-Lagrangian methods have been stated. Despite interesting a priori estimates that have been pointed out (cf. [3, 4, 16]), a lot of work still remains to give complete and rigorous results in more general situations. The more difficult step in the convergence analysis of semi-Lagrangian methods is to obtain a stability result for the interpolation operators. If stability results in the $W^{1,\infty}$ norm seem inaccessible for high-order interpolation operators because of the Runge phenomena (artificial oscillations, whose amplitude increases with the degree of the polynomial in the case of Lagrange interpolation, appear at the edges of finite elements), a more appropriate mathematical framework is H^1 stability. If Fourier analysis tools such as Fourier series are useful for proving H^1 stability in case of grids, for unstructured meshes such as triangulation, convenient mathematical tools are lacking and have to be developed in the future. Nevertheless new results on convergence analysis of classes of high-order schemes can be found in [6].

This paper is organized as follows. In the first part we present the continuous problem. In the second part we expose the discrete problem and the numerical scheme to solve it. Then we study the convergence of our numerical scheme.

2. The continuous problem. Denoting by $f(t,x,v) \geq 0$ the distribution function of electrons in phase space (with mass normalized to one and the charge to plus one) and by $E(t,x)$ the self-consistent electric field, the adimensional Vlasov–Poisson system reads as

$$\frac{\partial f}{\partial t} + v\frac{\partial f}{\partial x} + E(t,x)\frac{\partial f}{\partial v} = 0, \tag{2.1}$$

$$\frac{dE}{dx}(t,x) = \rho(t,x) = \int_{-\infty}^{+\infty} f(t,x,v)dv - 1, \tag{2.2}$$

where x and v are independent variables. We consider a periodic plasma of period L. Hence in (2.1) and (2.2) we have $x \in [0, L]$, $v \in \mathbb{R}$, $t \geq 0$, and the functions f and E satisfy the periodic boundary conditions

$$f(t, 0, v) = f(t, L, v), \quad v \in \mathbb{R}, \quad t \geq 0, \tag{2.3}$$

and

$$E(t, 0) = E(t, L) \iff \frac{1}{L} \int_0^L \int_{-\infty}^{+\infty} f(t, x, v) dv dx = 1, \quad t \geq 0, \tag{2.4}$$

which means that the plasma is globally neutral. In order to have a well-posed problem we add to (2.1)–(2.4) a zero-mean electrostatic condition

$$\int_0^L E(t, x) dx = 0, \quad t \geq 0, \tag{2.5}$$

and an initial condition

$$f(0, x, v) = f_0(x, v), \quad x \in [0, L], \quad v \in \mathbb{R}. \tag{2.6}$$

Besides, by assuming that the electric field E is smooth enough, we can solve (2.1), (2.3), and (2.6) in the classical sense as follows. For the existence, the uniqueness, and the regularity of the solutions of the following differential system, we refer the reader to [11].

We consider the first-order differential system

$$\begin{aligned} \frac{dX}{dt}(t; s, x, v) &= V(t; s, x, v), \\ \frac{dV}{dt}(t; s, x, v) &= E(t, X(t; s, x, v)) \end{aligned} \tag{2.7}$$

and denote by $t \to (X(t; s, x, v), V(t; s, x, v))$ the characteristic curves, which are the solution of (2.7) with the initial condition

$$X(s; s, x, v) = x, \qquad V(s; s, x, v) = v. \tag{2.8}$$

Then the solution of problem (2.1), (2.6) is given by

$$f(t, x, v) = f_0(X(0; t, x, v), V(0; t, x, v)), \quad x,\ v \in \mathbb{R}, \quad t \geq 0. \tag{2.9}$$

We note that the periodicity in x of $f_0(x, v)$ and $E(t, x)$ implies the periodicity in x of $f(t, x, v)$. Moreover as

$$\left| \frac{\partial(X, V)}{\partial(x, v)} \right| = 1,$$

we get

$$\frac{1}{L} \int_0^L \int_{-\infty}^{+\infty} f(t, x, v) dv dx = \frac{1}{L} \int_0^L \int_{-\infty}^{+\infty} f_0(x, v) dv dx = 1.$$

Therefore, according to the previous considerations, an equivalent form of the Vlasov–Poisson periodic problem is to find a couple (f, E), smooth enough, periodic with respect to x, with period L, and solving (2.2), (2.7), (2.8), and (2.9).

If we introduce the electrostatic potential $\phi = \phi(t,x)$ such that $E(t,x) = -\partial_x \phi(t,x)$, and if we denote by $G = G(x,y)$ the fundamental solution of the Laplacian operator in one dimension with a periodic boundary condition, therefore we obtain

$$E(t,x) = \int_0^L K(x,y) \left(\int_{-\infty}^{+\infty} f(t,y,v) dv - 1 \right) dy,$$

where

$$K(x,y) = -\partial_x G(x,y) = \begin{cases} \left(\dfrac{y}{L} - 1 \right), & 0 \le x < y, \\ \dfrac{y}{L}, & y < x \le L. \end{cases}$$

2.1. Existence, uniqueness, and regularity of the solution of the continuous problem. In this section we recall a theorem of the existence of a classical solution for the Vlasov–Poisson system. The following theorem gives the existence, the uniqueness, and the regularity of the classical solutions, global in time, of the Vlasov–Poisson periodic system in one dimension. Let us note that a theory of weak solutions (in BV spaces) for the Vlasov equation has been recently developed in [12].

THEOREM 2.1. *Assuming $f_0 \in \mathscr{C}^1_{c,per_x}(\mathbb{R}_x \times \mathbb{R}_v)$ (continuously differentiable functions which are periodic with respect to x and compactly supported with respect to v), positive, periodic with respect to the variable x with period L, and $Q(0) \le R$, with $R > 0$ and $Q(t)$ defined as*

$$Q(t) = 1 + \sup \left\{ |v| : \exists x \in [0,L], \quad \tau \in [0,t] \mid f(\tau,x,v) \neq 0 \right\},$$

and

$$\frac{1}{L} \int_0^L \int_{-\infty}^{+\infty} f_0(x,v) dv dx = 1,$$

then the periodic Vlasov–Poisson system has a unique classical solution (f,E) periodic in x, with period L, for all time t in $[0,T]$, such that

$$f \in \mathscr{C}^1_b \left(0, T; \mathscr{C}^1_{c,per_x}(\mathbb{R}_x \times \mathbb{R}_v) \right),$$

$$E \in \mathscr{C}^1_b \left(0, T; \mathscr{C}^1_{b,per_x}(\mathbb{R}) \right),$$

and there exists a constant $C = C(R, f_0)$ dependent on R and f_0 such that

$$Q(T) \le CT.$$

Moreover if we assume $f_0 \in \mathscr{C}^m_{c,per_x}(\mathbb{R}_x \times \mathbb{R}_v)$, then $(f,E) \in \mathscr{C}^m_b(0,T;\mathscr{C}^m_{c,per_x}(\mathbb{R}_x \times \mathbb{R}_v)) \times \mathscr{C}^m_b(0,T;\mathscr{C}^m_{b,per_x}(\mathbb{R}))$ for all finite time T.

Proof. For a proof we refer the reader to the treatise by Glassey [19]. □

3. The discrete problem. This section presents the description of the numerical scheme. Sections 3.1–3.3 are devoted to some notation and definitions (definitions of the interpolation operators and the transport operators in order to discretize in time and space the nonlinear transport equations, on a fixed phase-space grid). Section 3.4 resumes the algorithm in a simple way and refers the reader to the previous subsections for more details.

3.1. Interpolation operator. Let $\Omega = [0, L] \times [-R, R]$, with $R > Q(T)$, and let $\mathcal{M}_h$ be a Cartesian mesh of the phase space Ω. The grid $\mathcal{M}_h$ is given by a first increasing sequence $(x_i)_{i\in\{0,\dots,N_x\}}$ of the interval $[0, L]$ and a second increasing sequence $(v_i)_{i\in\{0,\dots,N_v\}}$ of the interval $[-R, R]$.

Let $\Delta x_i = x_{i+1} - x_i$ be the physical space set and $\Delta v_i = v_{i+1} - v_i$ be the velocity space set. In order to simplify the convergence analysis we suppose that $\Delta x_i = \Delta x$ and $\Delta v_i = \Delta v$. We call h a generic discretization parameter which stands for Δx or Δv. We define the one-dimensional Hermite interpolation operator $\mathcal{I}_h^{\mathcal{H}}$ as

$$\mathcal{I}_h^{\mathcal{H}} f(z)_{|_{[z_i, z_{i+1}]}} = f(z_i)\phi_i(z) + f(z_{i+1})\phi_{i+1}(z) + \dot{f}(z_i)\psi_i(z) + \dot{f}(z_{i+1})\psi_{i+1}(z),$$

where

$$\phi_i(z) = \frac{(z - z_{i+1})^2[(z_i - z_{i+1}) - 2(z - z_i)]}{(z_i - z_{i+1})^3}, \qquad \psi_i(z) = \frac{(z - z_i)(z - z_{i+1})^2}{(z_i - z_{i+1})^2},$$
$$\phi_{i+1}(z) = \frac{(z - z_i)^2[(z_i - z_{i+1}) + 2(z - z_{i+1})]}{(z_i - z_{i+1})^3}, \qquad \psi_{i+1}(z) = \frac{(z - z_i)^2(z - z_{i+1})}{(z_i - z_{i+1})^2},$$

and the one-dimensional Lagrange interpolation operator of degree three $\mathcal{I}_h^{\mathcal{L}}$ as

$$\mathcal{I}_h^{\mathcal{L}} f(z)_{|_{[z_i, z_{i+1}]}} = \sum_{k=i-1}^{i+2} f(z_k)\ell_k^3(z), \quad \text{where } \ell_k^3(z) = \prod_{\substack{i=k-1 \\ i\neq k}}^{i=k+2} \frac{(z - z_i)}{(z_k - z_i)}.$$

Let $\delta(y, y_i) = 1$ if $y = y_i$ and zero otherwise; then we consider $\mathcal{H}_h$ the interpolation operator defined by $\mathcal{H}_h f(x, v) = \sum_{j=0}^{N_v} \mathcal{I}_h^{\mathcal{H}} f(x, v_j)\delta(v, v_j)$ for advection in the x-direction and $\mathcal{H}_h f(x, v) = \sum_{i=0}^{N_x} \mathcal{I}_h^{\mathcal{H}} f(x_i, v)\delta(x, x_i)$ for advection in the v-direction. In the same way we define $\mathcal{L}_h$ thanks to $\mathcal{I}_h^{\mathcal{L}}$.

3.2. Transport operators. In this section we introduce some transport operators. Let $\mathbb{S}_{h,\xi}$ be the translation operator defined by $\mathbb{S}_{h,\xi} f(t, x, v) = f(t, x - h, v - \xi)$. We suppose that $t \in [t^n, t^{n+1}]$, and then we define $\mathcal{T}_1^{0,0}$, $\mathcal{T}_2^{0,0}$, $\mathcal{T}_1^{1,0}$, $\mathcal{T}_1^{0,1}$, $\mathcal{T}_2^{1,0}$, and $\mathcal{T}_2^{0,1}$ as follows:

$$\begin{aligned}
\mathcal{T}_1^{0,0} f(t, x, v) &= f(t, x - v\Delta t/2, v), \\
\mathcal{T}_2^{0,0} f(t, x, v) &= f\left(t, x, v - \Delta t E(t^{n+1/2}, x)\right), \\
\mathcal{T}_1^{1,0} f(t, x, v) &= \partial_x \mathcal{T}_1^{0,0} f(t, x, v) \\
&= \partial_x \left(f(t, x - v\Delta t/2, v)\right) = \partial_x f(t, x - v\Delta t/2, v), \\
\mathcal{T}_1^{0,1} f(t, x, v) &= \partial_v \mathcal{T}_1^{0,0} f(t, x, v) \\
&= \partial_v \left(f(t, x - v\Delta t/2, v)\right) \\
&= \partial_v f(t, x - v\Delta t/2, v) - \frac{\Delta t}{2} \partial_x f(t, x - v\Delta t/2, v), \\
\mathcal{T}_2^{1,0} f(t, x, v) &= \partial_x \mathcal{T}_2^{0,0} f(t, x, v) \\
&= \partial_x \left(f\left(t, x, v - \Delta t E(t^{n+1/2}, x)\right)\right) \\
&= \partial_x f\left(t, x, v - \Delta t E(t^{n+1/2}, x)\right) \\
&\quad - \Delta t \partial_x E\left(t^{n+1/2}, x\right) \partial_v f\left(t, x, v - \Delta t E(t^{n+1/2}, x)\right),
\end{aligned}$$

$$\begin{aligned}\mathcal{T}_2^{0,1} f(t,x,v) &= \partial_v \mathcal{T}_2^{0,0} f(t,x,v) \\ &= \partial_v \left(f\left(t,x,v-\Delta t E(t^{n+1/2},x)\right)\right) = \partial_v f\left(t,x,v-\Delta t E(t^{n+1/2},x)\right),\end{aligned}$$

where $E(t,x)$ is the solution of the Poisson problem (2.2) and (2.5). For a given couple (x_k, v_l) there exists a real $\theta_k \in]0,1[$ and an integer m with $(k+m,l) \in \mathcal{M}_h$ such that $x_k - v_l \Delta t/2 = x_{k+m} + \theta_k \Delta x$. We set

$$\begin{aligned}\phi_0(z) &= \frac{1}{h^3}(h+2z)(h-z)^2, \qquad & \phi_1(z) &= \frac{z^2}{h^3}(3h-2z), \\ \psi_0(z) &= \frac{z}{h^2}(h-z)^2, & \psi_1(z) &= \frac{z^2}{h^2}(z-h),\end{aligned}$$

where h denotes Δx or Δv. Therefore, using the notation $f_{k,l}(t) = f(t,x_k,v_l)$ and $\partial_x f_{k,l}(t) = (\partial_x f)(t,x_k,v_l)$, we define $\widetilde{\mathcal{T}}_1^{0,0}$, $\widetilde{\mathcal{T}}_1^{1,0}$, and $\widetilde{\mathcal{T}}_1^{0,1}$ as follows:

$$\begin{aligned}\left(\widetilde{\mathcal{T}}_1^{0,0} f(t)\right)_{k,l} &= \left(\mathbb{S}_{v\Delta t/2,0} \mathcal{H}_h f\right)(t,x_k,v_l) \\ &= f_{m+k,l}(t)\phi_0(\theta_k\Delta x) + f_{m+k+1,l}(t)\phi_1(\theta_k\Delta x) \\ &\quad + \partial_x f_{m+k,l}(t)\psi_0(\theta_k\Delta x) + \partial_x f_{m+k+1,l}(t)\psi_1(\theta_k\Delta x), \\ \left(\widetilde{\mathcal{T}}_1^{1,0} f(t)\right)_{k,l} &= \left(\mathbb{S}_{v\Delta t/2,0} \partial_x \mathcal{H}_h f\right)(t,x_k,v_l), \\ &= f_{m+k,l}(t)\dot{\phi}_0(\theta_k\Delta x) + f_{m+k+1,l}(t)\dot{\phi}_1(\theta_k\Delta x) \\ &\quad + \partial_x f_{m+k,l}(t)\dot{\psi}_0(\theta_k\Delta x) + \partial_x f_{m+k+1,l}(t)\dot{\psi}_1(\theta_k\Delta x), \\ \left(\widetilde{\mathcal{T}}_1^{0,1} f(t)\right)_{k,l} &= \left(\mathbb{S}_{v\Delta t/2,0} \mathcal{L}_h \partial_v f\right)(t,x_k,v_l) - \frac{\Delta t}{2}\left(\mathbb{S}_{v\Delta t/2,0} \partial_x \mathcal{H}_h f\right)(t,x_k,v_l) \\ &= \sum_{i=k-1}^{k+2} \partial_v f_{m+i,l}(t)\ell_i^3((m+i+\theta_k)\Delta x)) \\ &\quad - \frac{\Delta t}{2}\{f_{m+k,l}(t)\dot{\phi}_0(\theta_k\Delta x) + f_{m+k+1,l}(t)\dot{\phi}_1(\theta_k\Delta x) \\ &\quad + \partial_x f_{m+k,l}(t)\dot{\psi}_0(\theta_k\Delta x) + \partial_x f_{m+k+1,l}(t)\dot{\psi}_1(\theta_k\Delta x)\}.\end{aligned}$$

Here we have used the Lagrange interpolation of degree three on the gradient in order to keep high-order accuracy.

Then we introduce the following two Poisson problems:

$$\text{(P)} \qquad \left\{ \left(\frac{d}{dx} E_h^{n+1/2}\right)(x) = \int_v \widetilde{\mathcal{T}}_1^{0,0} f_h(t^n) dv - 1, \quad \int_0^L E_h^{n+1/2}(x) dx = 0 \right.$$

and

$$\text{(P}') \left\{ \left(\frac{d}{dx}\left(\frac{d}{dx} E_h^{n+1/2}\right)\right)(x) = \int_v \widetilde{\mathcal{T}}_1^{1,0} f_h(t^n) dv, \quad \int_0^L \left(\frac{d}{dx} E_h^{n+1/2}\right)(x) dx = 0. \right.$$

In order to discretize (P) and (P′) we use a finite difference formula of order four for the x-derivative and a quadrature formula of order four for the integration in v. As a

consequence we have the following two schemes:

$$(\mathrm{P}_h)\quad \begin{cases} \dfrac{1}{12\Delta x}\{8(E_h^{n+1/2}(x_{k+1})-E_h^{n+1/2}(x_{k-1}))-(E_h^{n+1/2}(x_{k+2})-E_h^{n+1/2}(x_{k-2}))\} \\ = -1+\dfrac{\Delta v}{3}\Bigg\{ \left(\widetilde{\mathcal{T}}_1^{0,0} f_h(t^n)\right)_{k,0} + \left(\widetilde{\mathcal{T}}_1^{0,0} f_h(t^n)\right)_{k,N_v} \\ +4\displaystyle\sum_{l,\mathrm{odd}} \left(\widetilde{\mathcal{T}}_1^{0,0} f_h(t^n)\right)_{k,l} + 2\sum_{l,\mathrm{even}} \left(\widetilde{\mathcal{T}}_1^{0,0} f_h(t^n)\right)_{k,l} \Bigg\}, \\ E_h^{n+1/2}(x_0)+E_h^{n+1/2}(x_{N_x})+4\displaystyle\sum_{l,\mathrm{odd}} E_h^{n+1/2}(x_l)+2\sum_{l,\mathrm{even}} E_h^{n+1/2}(x_l)=0, \end{cases}$$

$$(\mathrm{P}'_h)\quad \begin{cases} \dfrac{1}{12\Delta x}\{8(\partial_x E_h^{n+1/2}(x_{k+1})-\partial_x E_h^{n+1/2}(x_{k-1}))-(\partial_x E_h^{n+1/2}(x_{k+2}) \\ -\partial_x E_h^{n+1/2}(x_{k-2}))\} = \dfrac{\Delta v}{3}\Bigg\{ \left(\widetilde{\mathcal{T}}_1^{1,0} f_h(t^n)\right)_{k,0} + \left(\widetilde{\mathcal{T}}_1^{1,0} f_h(t^n)\right)_{k,N_v} \\ +4\displaystyle\sum_{l,\mathrm{odd}} \left(\widetilde{\mathcal{T}}_1^{1,0} f_h(t^n)\right)_{k,l} + 2\sum_{l,\mathrm{even}} \left(\widetilde{\mathcal{T}}_1^{1,0} f_h(t^n)\right)_{k,l} \Bigg\}, \\ \partial_x E_h^{n+1/2}(x_0)+\partial_x E_h^{n+1/2}(x_{N_x})+4\displaystyle\sum_{l,\mathrm{odd}} \partial_x E_h^{n+1/2}(x_l)+2\sum_{l,\mathrm{even}} \partial_x E_h^{n+1/2}(x_l)=0. \end{cases}$$

As we choose periodic boundary conditions in the x-direction we have

$$E_h^{n+1/2}(x_{l+N_x+1}) = E_h^{n+1/2}(x_l), \quad \partial_x E_h^{n+1/2}(x_{l+N_x+1}) = \partial_x E_h^{n+1/2}(x_l), \quad l\in\mathbb{Z}.$$

Finally we build $E_h^{n+1/2}(x)$ by using the Hermite interpolation operator $\mathcal{I}_h^{\mathcal{H}}$, and, with $x = x_i+\theta\Delta x$, $0\le\theta\le 1$, we get

$$\begin{aligned} E_h^{n+1/2}(x)_{|_{[x_i,x_{i+1}]}} &= E_h^{n+1/2}(x_i)\phi_0(\theta\Delta x)+E_h^{n+1/2}(x_{i+1})\phi_1(\theta\Delta x) \\ &\quad +\partial_x E_h^{n+1/2}(x_i)\psi_0(\theta\Delta x)+\partial_x E_h^{n+1/2}(x_{i+1})\psi_1(\theta\Delta x). \end{aligned}$$

For a given couple (x_k, v_l) there exists a real $\theta_l\in\,]0,1[$ and an integer m with $(k, l+m)\in\mathcal{M}_h$ such that $v_l-\Delta t E_h^{n+1/2}(x_k) = v_{l+m}+\theta_l\Delta v$. Therefore, using the notation $\partial_v f_{k,l}(t) = (\partial_v f)(t,x_k,v_l)$, we consider the transport operators $\widehat{\mathcal{T}}_2^{0,0}$, $\widehat{\mathcal{T}}_2^{1,0}$, and $\widehat{\mathcal{T}}_2^{0,1}$, defined by

$$\begin{aligned} \widehat{\mathcal{T}}_2^{0,0} f(t,x,v) &= f\left(t,x,v-\Delta t E_h^{n+1/2}(x)\right), \\ \widehat{\mathcal{T}}_2^{1,0} f(t,x,v) &= \partial_x \widehat{\mathcal{T}}_2^{0,0} f(t,x,v) \\ &= \partial_x\left(f\left(t,x,v-\Delta t E_h^{n+1/2}(x)\right)\right) \\ &= \partial_x f\left(t,x,v-\Delta t E_h^{n+1/2}(x)\right) \\ &\quad -\Delta t\partial_x E_h^{n+1/2}(x)\partial_v f\left(t,x,v-\Delta t E_h^{n+1/2}(x)\right), \\ \widehat{\mathcal{T}}_2^{0,1} f(t,x,v) &= \partial_v \widehat{\mathcal{T}}_2^{0,0} f(t,x,v) \\ &= \partial_v\left(f\left(t,x,v-\Delta t E_h^{n+1/2}(x)\right)\right) = \partial_v f\left(t,x,v-\Delta t E_h^{n+1/2}(x)\right), \end{aligned}$$

and $\widetilde{\mathcal{T}}_2^{0,0}$, $\widetilde{\mathcal{T}}_2^{1,0}$, and $\widetilde{\mathcal{T}}_2^{0,1}$, defined by

$$\begin{aligned}
\left(\widetilde{\mathcal{T}}_2^{0,0} f(t)\right)_{k,l} &= \left(\mathbb{S}_{0,\Delta t E_h} \mathcal{H}_h f\right)(t, x_k, v_l) \\
&= f_{k,m+l}(t)\phi_0(\theta_l \Delta v) + f_{k,m+l+1}(t)\phi_1(\theta_l \Delta v) \\
&\quad + \partial_v f_{k,m+l}(t)\psi_0(\theta_l \Delta v) + \partial_v f_{k,m+l+1}(t)\psi_1(\theta_l \Delta v), \\
\left(\widetilde{\mathcal{T}}_2^{1,0} f(t)\right)_{k,l} &= \left(\mathbb{S}_{0,\Delta t E_h} \mathcal{L}_h \partial_x f\right)(t, x_k, v_l) - \Delta t \partial_x E_h^{n+1/2}(x_k)\left(\mathbb{S}_{0,\Delta t E_h} \partial_v \mathcal{H}_h f\right)(t, x_k, v_l) \\
&= \sum_{i=l-1}^{l+2} \partial_x f_{k,m+i}(t) \ell_i^3((m + i + \theta_l)\Delta v) \\
&\quad - \Delta t \partial_x E_h^{n+1/2}(x_k)\{\partial_v f_{k,m+l}(t)\dot{\phi}_0(\theta_l \Delta v) + f_{k,m+l+1}(t)\dot{\phi}_1(\theta_l \Delta v) \\
&\quad + \partial_v f_{k,m+l}(t)\dot{\psi}_0(\theta_l \Delta v) + \partial_v f_{k,m+l+1}(t)\dot{\psi}_1(\theta_l \Delta v)\}, \\
\left(\widetilde{\mathcal{T}}_2^{0,1} f(t)\right)_{k,l} &= \left(\mathbb{S}_{0,\Delta t E_h} \partial_v \mathcal{H}_h f\right)(t, x_k, v_l) \\
&= f_{k,m+l}(t)\dot{\phi}_0(\theta_l \Delta v) + f_{k,m+l+1}(t)\dot{\phi}_1(\theta_l \Delta v) \\
&\quad + \partial_v f_{k,m+l}(t)\dot{\psi}_0(\theta_l \Delta v) + \partial_v f_{k,m+l+1}(t)\dot{\psi}_1(\theta_l \Delta v).
\end{aligned}$$

Let us notice that the transport operators $\widetilde{\mathcal{T}}_1^{\alpha}$ and $\widetilde{\mathcal{T}}_2^{\alpha}$ act only on a Cartesian grid function and that the operators $\mathcal{T}_1^{\alpha}$, $\mathcal{T}_2^{\alpha}$, and $\widehat{\mathcal{T}}_2^{\alpha}$ act only on functions defined in a Cartesian coordinate system. Therefore when we make the composition of these operators we must take into account this constraint. In order to be more precise we give an abstract example of the composition of such an operator. Let φ and ψ be two applications from $\mathbb{R}^2$ into $\mathbb{R}^2$:

$$\begin{array}{rcl}
\varphi : & \mathbb{R}^2 \longrightarrow \mathbb{R}^2 \\
& (x, v) \longrightarrow \varphi(x, v) = (\varphi_1(x, v), \varphi_2(x, v))
\end{array}
\quad \Big| \quad
\begin{array}{rcl}
\psi : & \mathbb{R}^2 \longrightarrow \mathbb{R}^2 \\
& (x, v) \longrightarrow \psi(x, v) \\
& \qquad = (\psi_1(x, v), \psi_2(x, v)).
\end{array}$$

Let Γ_1 and Γ_1 be two transport operators which act only on functions defined in a Cartesian coordinate system. The operators Γ_1 and Γ_2 are defined as follows:

$$\Gamma_1 f(x, v) = f(\varphi_1(x, v), \varphi_2(x, v)) \quad \text{and} \quad \Gamma_2 f(x, v) = f(\psi_1(x, v), \psi_2(x, v)).$$

Let us now compute $\Gamma_1 \circ \Gamma_2 \circ \Gamma_1 f(x, v)$. First by applying Γ_1 to f we get

$$\Gamma_1 \circ \Gamma_2 \circ \Gamma_1 f(x, v) = \Gamma_1 \circ \Gamma_2 f(\varphi_1(x, v), \varphi_2(x, v)).$$

Now the function $f(\cdot, \cdot)$ is no longer defined in a Cartesian coordinate system, but it is defined in a curvilign coordinate system given by the mapping φ. Nevertheless there exists a function g defined in the original Cartesian coordinate system (x, v) such that $g(x, v) = f(\varphi_1(x, v), \varphi_2(x, v))$. We can now apply the operator Γ_2 to g, and we obtain

$$\begin{aligned}
\Gamma_1 \circ \Gamma_2 \circ \Gamma_1 f(x, v) &= \Gamma_1 \circ \Gamma_2 g(x, v) \\
&= \Gamma_1 g(\psi_1(x, v), \psi_2(x, v)) \\
&= \Gamma_1 f(\varphi_1(\psi_1(x, v), \psi_2(x, v)), \varphi_2(\psi_1(x, v), \psi_2(x, v))).
\end{aligned}$$

Following the same argument as before, there exists a function h which is defined in the original Cartesian coordinate system (x, v) such that $h(x, v) = g(\psi_1(x, v), \psi_2(x, v))$.

Therefore we get

$$\begin{aligned}\Gamma_1\circ\Gamma_2\circ\Gamma_1 f(x,v) &= \Gamma_1\circ\Gamma_2 g(x,v) \;=\; \Gamma_1 g(\psi_1(x,v),\psi_2(x,v)) \;=\; \Gamma_1 h(x,v)\\ &= h(\varphi_1(x,v),\varphi_2(x,v))\\ &= g(\psi_1(\varphi_1(x,v),\varphi_2(x,v)),\psi_2(\varphi_1(x,v),\varphi_2(x,v)))\\ &= f\left(\varphi_1(\psi_1(\varphi_1(x,v),\varphi_2(x,v)),\psi_2(\varphi_1(x,v),\varphi_2(x,v))),\right.\\ &\qquad \left.\varphi_2(\psi_1(\varphi_1(x,v),\varphi_2(x,v)),\psi_2(\varphi_1(x,v),\varphi_2(x,v)))\right).\end{aligned}$$

3.3. Notation and definitions. In this section we introduce some notation and definitions. The grid function $f=\{f_{i,j}\}_{(i,j)\in\mathcal{M}_h}$ belongs to functional space $L^2_h(\Omega)$ if the norm $\|\cdot\|_{L^2_h(\Omega)}$ defined as

$$\|f\|_{L^2_h(\Omega)}=\left(\Delta x\Delta v\sum_{(i,j)\in\mathcal{M}_h}|f_{i,j}|^2\right)^{1/2}$$

is uniformly bounded.

Let

$$\begin{aligned}G:\quad \mathcal{M}_h &\longrightarrow \mathbb{R}^2,\\ (i,j) &\longrightarrow (G_{1_{i,j}},G_{2_{i,j}}),\end{aligned}$$

with $G_1,\ G_2\in L^2_h(\Omega)$. Then we define the norm $\|\cdot\|_{\mathbb{L}^2_h(\Omega)}$ as

$$\|G\|^2_{\mathbb{L}^2_h(\Omega)}=\|G_1\|^2_{L^2_h(\Omega)}+\|G_2\|^2_{L^2_h(\Omega)}.$$

Let $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ be the vectors $(\mu_0,\dots,\mu_j,\dots,\mu_{N_v})$ and $(\nu_0,\dots,\nu_i,\dots,\nu_{N_x})$, with $0\le\mu_j,\nu_i\le 1$ for all $(i,j)\in[0,N_x]\times[0,N_v]$. Then we define the norm $\|\cdot\|_{L^2_h,\Delta^{\boldsymbol{\mu},\boldsymbol{\nu}}_h}$ by

$$\|f\|_{L^2_h,\Delta^{\boldsymbol{\mu},\boldsymbol{\nu}}_h}=\left(\Delta x\Delta v\sum_{(i,j)\in\mathcal{M}_h}|f_{i+\mu_j,j+\nu_i}|^2\right)^{1/2}.$$

We also define the translation operator $\tau_{\boldsymbol{\mu},\boldsymbol{\nu}}$ as follows:

$$(\tau_{\boldsymbol{\mu},\boldsymbol{\nu}}f)_{i,j}=f_{i+\mu_j,j+\nu_i}=f(x_i+\mu_j\Delta x,v_j+\nu_i\Delta v).$$

We continue by introducing the space $\ell^\infty(0,T;X)$ defined by

$$\begin{aligned}\ell^\infty(0,T;X):=\Big\{&f:\{t^0,\dots,t^M\}\to X|M\Delta t=T,\\ &||f||_{\ell^\infty(0,T;X)}=\max_{1\le n\le M}||f(t^n)||_X<\infty\Big\},\end{aligned}$$

where X denotes a functional space with norm $||.||_X$, and the norm $\|\cdot\|_{L^\infty_h}$ defined by

$$\|f\|_{L^\infty_h}=\max_{(i,j)\in\mathcal{M}_h}|f_{i,j}|.$$

We introduce now the notation

$$\mathcal{T}^\alpha,\quad \mathcal{T}\in\left\{\mathcal{T}_1,\mathcal{T}_2,\widetilde{\mathcal{T}}_1,\widetilde{\mathcal{T}}_2,\widehat{\mathcal{T}}_2\right\},$$

where $\alpha=(\alpha_1,\alpha_2)\in\{(0,0),(1,0),(0,1)\}$, with $|\alpha|=\alpha_1+\alpha_2$. In order to simplify the notation the following conventions are used:

$$\nabla\mathcal{T}=\begin{pmatrix}\mathcal{T}^{1,0}\\ \mathcal{T}^{0,1}\end{pmatrix},\qquad \partial^\alpha f\in\{\partial^{0,0}f,\partial^{1,0}f,\partial^{0,1}f\}=\{f,\partial_x f,\partial_v f\}.$$

3.4. The numerical scheme. We suppose that we know $f_h(t^n) = \left\{f_{i,j}^n, \partial_x f_{i,j}^n, \partial_v f_{i,j}^n\right\}_{(i,j)\in\mathcal{M}_h}$. Therefore the numerical scheme which allows us to go from time t^n to t^{n+1} and compute $f_h(t^{n+1}) = \left\{f_{i,j}^{n+1}, \partial_x f_{i,j}^{n+1}, \partial_v f_{i,j}^{n+1}\right\}_{(i,j)\in\mathcal{M}_h}$ can be described in four steps:

(A1) We evaluate the distribution function and its partial derivatives at time t^n at the foot of the field-free characteristics starting at $\{(x_k, v_l)\}_{(k,l)\in\mathcal{M}_h}$ at time $t^{n+1/2}$ using $\mathcal{L}_h$ and $\mathcal{H}_h$, the Lagrange and Hermite interpolation operators, respectively. Then we obtain three new grid functions: One approximates the particle distribution, and the other two approximate its gradients. These actions are described by the transport operators $\widetilde{\mathcal{T}}_1^{\alpha}$. The superscripts denote the position or velocity gradient under consideration.

(A2) The output from (A1) is integrated with respect to velocity (by a quadrature formula of order four) to provide an approximation for the density and its gradient at time $t^{n+1/2}$, which are then substituted into the Poisson equations (P_h) and (P_h') to compute the approximation of the electric field and its gradient at time $t^{n+1/2}$, respectively.

(A3) Therefore the result obtained from (A1) is evaluated at the foot of the velocity characteristic starting at $\{(x_k, v_l)\}_{(k,l)\in\mathcal{M}_h}$ at time t^{n+1} with the acceleration field found in (A2) using $\mathcal{L}_h$ and $\mathcal{H}_h$, the Lagrange and Hermite interpolation operators, respectively. Then we obtain three new grid functions: One approximates the particle distribution, and the other two approximate its gradients. These actions are described by the transport operators $\widetilde{\mathcal{T}}_2^{\alpha}$, with the superscripts denoting the appropriate gradient under consideration.

(A4) Between time $t^{n+1/2}$ and t^{n+1}, we apply step (A1) to the output from (A3). This action is described by the transport operators $\widetilde{\mathcal{T}}_1^{\alpha}$. Then we obtain $f_h(t^{n+1})$, the distribution function and its gradients at time t^{n+1}, which are the new initial data for the algorithm (A1)–(A4).

By using transport operators defined in section 3.2 the numerical scheme can be written as

$$\left(\partial^{\alpha} f_h(t^{n+1})\right)_{k,l} = \left(\widetilde{\mathcal{T}}_1^{\alpha} \circ \widetilde{\mathcal{T}}_2^{\alpha} \circ \widetilde{\mathcal{T}}_1^{\alpha} f_h(t^n)\right)_{k,l}, \quad |\alpha| \leq 1,\ k \in [0, N_x],\ l \in [0, N_v],$$

with $(\partial^{\alpha} f_h^0) = (\partial^{\alpha} f_0)_{k,l}$, $|\alpha| \leq 1$, as a discretization of the initial data f_0, $\partial^{\alpha} f_h^n(x_0 + L, v_l) = \partial^{\alpha} f_h^n(x_0, v_l)$, $|\alpha| \leq 1$, for all $l \in [0, N_v]$, the boundary condition in the x-direction, and $\partial^{\alpha} f_h^n(x, v_l) = 0$, for all $|v_l| > R$, for all $x \in [0, L]$, $|\alpha| \leq 1$, the boundary condition in the v-direction.

Let us note that the method presented in this paper belongs to the class of semi-Lagrangian methods but has the distinctive feature of using a splitting à la Strang for the operators associated to the transport of the distribution function and its gradient.

4. Convergence analysis. Let us state the convergence theorem and give the a priori error estimates.

THEOREM 4.1. *Assume that* $f \in \mathscr{C}^2(0, T; \mathscr{C}^4_{c,per_x}(\mathbb{R}_x \times \mathbb{R}_v))$, *is positive, and is periodic with respect to the variable* x *with period* L; *then the numerical solution of the Vlasov–Poisson system* (f_h, E_h), *computed by the numerical scheme exposed in section* 3.4, *converges towards the solution* (f, E) *of the periodic Vlasov–Poisson system, and there exists a constant* $C = C(||f||_{\mathscr{C}^2(0,T;\mathscr{C}^4_b(\Omega))})$ *independent of* Δt, Δx,

and Δv *such that for* $|\alpha| \leq 1$ *we have*

$$||\partial^\alpha f - \partial^\alpha f_h||_{l^\infty\left(0,T;L^2_h(\Omega)\right)} \leq C\Bigg(\Delta t^2 + \frac{\Delta x^{4-|\alpha|} + \Delta v^{4-|\alpha|}}{\Delta t}. \\ + \frac{\left(\Delta x^4 + \Delta v^4\right)}{\Delta x^{|\alpha|}}\left(1 + \frac{\Delta x}{\Delta t} + \frac{1}{\Delta t}\right)\Bigg)$$

and

$$||\partial^\alpha E - \partial^\alpha E_h||_{l^\infty\left(0,T;L^\infty_h([0,L])\right)} \leq C\Bigg(\Delta t^2 + \Delta x^4 + \Delta v^4 + \Delta x^{4-|\alpha|} \\ + \frac{\left(\Delta x^4 + \Delta v^4\right)}{\Delta x^{|\alpha|}}\left(1 + \frac{\Delta x}{\Delta t} + \frac{1}{\Delta t}\right) \\ + \frac{\Delta x^{4-|\alpha|} + \Delta v^{4-|\alpha|}}{\Delta t}\Bigg).$$

Remark 4.2. Theorem 4.1 tells us that the convergence occurred with a rate in $\mathcal{O}(\Delta t^2 + \Delta t^{-1}(\Delta x^{4-|\alpha|} + \Delta v^{4-|\alpha|}))$, $\alpha \in \mathbb{N}^2$, $|\alpha| \leq 1$. If $f \in \mathscr{C}^1(0,T;\mathscr{C}^2_{c,per_x}(\mathbb{R}_x \times \mathbb{R}_v))$, the convergence rate will be in $\mathcal{O}(\Delta t + \Delta t^{-1}(\Delta x^{2-|\alpha|} + \Delta v^{2-|\alpha|}))$, $\alpha \in \mathbb{N}^2$, $|\alpha| \leq 1$.

Idea of the proof. We want to evaluate the global error at the time t^{n+1} in the L^2_h norm

$$\|\partial^\alpha e^{n+1}\|_{L^2_h(\Omega)} = ||\partial^\alpha f(t^{n+1}) - \partial^\alpha f_h(t^{n+1})||_{L^2_h(\Omega)}.$$

Therefore we decompose $\partial^\alpha f(t^{n+1}, x_k, v_l) - \partial^\alpha f_h(t^{n+1}, x_k, v_l)$ as follows:

$$\begin{aligned}
&\partial^\alpha f(t^{n+1}, x_k, v_l) - \partial^\alpha f_h(t^{n+1}, x_k, v_l) \\
&\qquad = \partial^\alpha f(t^{n+1}, x_k, v_l) - \left(\mathcal{T}_1^\alpha \circ \mathcal{T}_2^\alpha \circ \mathcal{T}_1^\alpha f(t^n)\right)_{k,l} && \text{(I1)} \\
&\qquad + \left(\mathcal{T}_1^\alpha \circ \mathcal{T}_2^\alpha \circ \mathcal{T}_1^\alpha f(t^n)\right)_{k,l} - \left(\mathcal{T}_1^\alpha \circ \widehat{\mathcal{T}}_2^\alpha \circ \mathcal{T}_1^\alpha f(t^n)\right)_{k,l} && \text{(I2)} \\
&\qquad + \left(\mathcal{T}_1^\alpha \circ \widehat{\mathcal{T}}_2^\alpha \circ \mathcal{T}_1^\alpha f(t^n)\right)_{k,l} - \left(\widetilde{\mathcal{T}}_1^\alpha \circ \widetilde{\mathcal{T}}_2^\alpha \circ \widetilde{\mathcal{T}}_1^\alpha f(t^n)\right)_{k,l} && \text{(I3)} \\
&\qquad + \left(\widetilde{\mathcal{T}}_1^\alpha \circ \widetilde{\mathcal{T}}_2^\alpha \circ \widetilde{\mathcal{T}}_1^\alpha (f(t^n) - f_h(t^n))\right)_{k,l}. && \text{(I4)}
\end{aligned}$$

In order to estimate $\|\partial^\alpha e^{n+1}\|_{L^2_h(\Omega)}$ we will estimate the four right-hand terms of the previous equation. These estimates are described in the following section.

A priori estimates. The following lemma gives an estimate for the term (I1), which can be viewed as a discretization error in time.

LEMMA 4.3. *Assuming that* $f \in \mathscr{C}^2_b(0,T;\mathscr{C}^2_{c,per_x}(\mathbb{R}_x \times \mathbb{R}_v))$, *then there exists a constant* C *such that for* $|\alpha| \leq 1$ *we get*

$$\left\|\partial^\alpha f(t^{n+1}) - \mathcal{T}_1^\alpha \circ \mathcal{T}_2^\alpha \circ \mathcal{T}_1^\alpha f(t^n)\right\|_{L^2_h(\Omega)} \leq C\Delta t^3.$$

Proof. On one hand we can write

$$\begin{aligned}
f(t^{n+1}, x, v) &= f(t^{n+1}, X(t^{n+1}; t^{n+1}, x, v), V(t^{n+1}; t^{n+1}, x, v)) \\
&= f(t^n, X(t^n; t^{n+1}, x, v), V(t^n; t^{n+1}, x, v)) \\
&= f(t^n, X(t^n), V(t^n))
\end{aligned}$$

and

$$\begin{aligned}\nabla f(t^{n+1},x,v) &= \begin{pmatrix} \dfrac{\partial f}{\partial x}(t^{n+1},x,v) \\ \dfrac{\partial f}{\partial v}(t^{n+1},x,v) \end{pmatrix} \\ &= \begin{pmatrix} \dfrac{\partial X}{\partial x}(t^n) & \dfrac{\partial V}{\partial x}(t^n) \\ \dfrac{\partial X}{\partial v}(t^n) & \dfrac{\partial V}{\partial v}(t^n) \end{pmatrix} \begin{pmatrix} \dfrac{\partial f}{\partial X}(t^n,X(t^n),V(t^n)) \\ \dfrac{\partial f}{\partial V}(t^n,X(t^n),V(t^n)) \end{pmatrix} \\ &= \nabla(X(t^n),V(t^n))\nabla f(t^n,X(t^n),V(t^n)).\end{aligned}$$

Now we introduce $\mathbb{T}_1$ and $\mathbb{T}_2$ defined by

$$\mathbb{T}_1 = \begin{pmatrix} \mathcal{T}_1^{0,0} \\ \mathcal{T}_1^{1,0} \\ \mathcal{T}_1^{0,1} \end{pmatrix}, \qquad \mathbb{T}_2 = \begin{pmatrix} \mathcal{T}_2^{0,0} \\ \mathcal{T}_2^{1,0} \\ \mathcal{T}_2^{0,1} \end{pmatrix}.$$

On the other hand we get

$$\begin{aligned}\mathbb{T}_1 \circ \mathbb{T}_2 \circ \mathbb{T}_1 f(t^n) &= \mathbb{T}_1 \circ \mathbb{T}_2 \circ \mathbb{T}_1 f(t^n,x,v) \\ &= \mathbb{T}_1 \circ \mathbb{T}_2 \begin{pmatrix} f(t^n,x-v\Delta t/2,v) \\ \partial_x f(t^n,x-v\Delta t/2,v) \\ \partial_v f(t^n,x-v\Delta t/2,v) - \frac{\Delta t}{2}\partial_x f(t^n,x-v\Delta t/2,v) \end{pmatrix} \\ &= \mathbb{T}_1 \begin{pmatrix} f(t^n,x-v\frac{\Delta t}{2}+\frac{\Delta t^2}{2}E(t^{n+1/2},x),v-\Delta t E(t^{n+1/2},x)) \\ \\ \partial_x f(t^n,x-v\frac{\Delta t}{2}+\frac{\Delta t^2}{2}E(t^{n+1/2},x),v-\Delta t E(t^{n+1/2},x)) \\ -\Delta t \partial_x E(t^{n+1/2},x) \\ \times\{\partial_v f(t^n,x-v\frac{\Delta t}{2}+\frac{\Delta t^2}{2}E(t^{n+1/2},x),v-\Delta t E(t^{n+1/2},x)) \\ -\frac{\Delta t}{2}\partial_x f(t^n,x-v\frac{\Delta t}{2}+\frac{\Delta t^2}{2}E(t^{n+1/2},x),v-\Delta t E(t^{n+1/2},x))\} \\ \\ \partial_v f(t^n,x-v\frac{\Delta t}{2}+\frac{\Delta t^2}{2}E(t^{n+1/2},x),v-\Delta t E(t^{n+1/2},x)) \\ -\frac{\Delta t}{2}\partial_x f(t^n,x-v\frac{\Delta t}{2}+\frac{\Delta t^2}{2}E(t^{n+1/2},x),v-\Delta t E(t^{n+1/2},x)) \end{pmatrix} \\ &= \begin{pmatrix} f(t^n,\widetilde{X}(t^n),\widetilde{V}(t^n)) \\ \dfrac{\partial \widetilde{X}}{\partial x}(t^n)\partial_x f(t^n,\widetilde{X}(t^n),\widetilde{V}(t^n)) + \dfrac{\partial \widetilde{V}}{\partial x}(t^n)\partial_v f(t^n,\widetilde{X}(t^n),\widetilde{V}(t^n)) \\ \dfrac{\partial \widetilde{X}}{\partial v}(t^n)\partial_x f(t^n,\widetilde{X}(t^n),\widetilde{V}(t^n)) + \dfrac{\partial \widetilde{V}}{\partial v}(t^n)\partial_v f(t^n,\widetilde{X}(t^n),\widetilde{V}(t^n)) \end{pmatrix},\end{aligned}$$

where

$$\widetilde{X}(t^n) = x - v\Delta t + \frac{\Delta t^2}{2}E(t^{n+1/2},x-v\Delta t/2)$$

and

$$\widetilde{V}(t^n) = v - \Delta t E(t^{n+1/2}, x - v\Delta t/2).$$

A Taylor expansion gives

$$\begin{aligned} X(t^{n+1/2}) - (x - v\Delta t/2) &= X(t^{n+1/2}) - \left(X(t^{n+1}) - V(t^{n+1})\frac{\Delta t}{2}\right) \\ &= X(t^{n+1/2}) - \left(X(t^{n+1}) - \frac{\Delta t}{2}\dot{X}(t^{n+1})\right) \\ &= \mathcal{O}(\Delta t^2). \end{aligned} \tag{4.1}$$

By using (4.1) we get

$$\begin{aligned} V(t^n) - \widetilde{V}(t^n) &= V(t^n) - \left(V(t^{n+1}) - \Delta t E\left(t^{n+1/2}, X(t^{n+1}) - V(t^{n+1})\frac{\Delta t}{2}\right)\right) \\ &= V(t^n) - \left(V(t^{n+1}) - \Delta t E\left(t^{n+1/2}, X(t^{n+1/2}) + \mathcal{O}(\Delta t^2)\right)\right) \\ &= V(t^n) - (V(t^{n+1}) - \Delta t E(t^{n+1/2}, X(t^{n+1/2})) + \mathcal{O}(\Delta t^3) \\ &= V(t^n) - V(t^{n+1}) + \Delta t \dot{V}(t^{n+1/2}) + \mathcal{O}(\Delta t^3). \end{aligned}$$

We deduce that

$$\begin{aligned} &\sup\left\{\left|\partial^\alpha V(t^n; t^{n+1}, x, v) - \partial^\alpha \widetilde{V}(t^n; t^{n+1}, x, v)\right| \mid \forall (x,v) \in [0,L]\times[-R,R],\ |\alpha| \le 1\right\} \\ &\quad \le C\Delta t^3. \end{aligned}$$

In the same way we have

$$\begin{aligned} &X(t^n) - \widetilde{X}(t^n) \\ &\quad = X(t^n) - \left(X(t^{n+1}) - \Delta t V(t^{n+1}) + \frac{\Delta t^2}{2} E\left(t^{n+1/2}, X(t^{n+1}) - V(t^{n+1})\frac{\Delta t}{2}\right)\right) \\ &\quad = X(t^n) - \left(X(t^{n+1}) - \Delta t V(t^{n+1}) + \frac{\Delta t^2}{2} E(t^{n+1/2}, X(t^{n+1/2}) + \mathcal{O}(\Delta t^2))\right) \\ &\quad = X(t^n) - \left(X(t^{n+1}) - \Delta t V(t^{n+1}) + \frac{\Delta t^2}{2} E(t^{n+1/2}, X(t^{n+1/2}))\right) + \mathcal{O}(\Delta t^4) \\ &\quad = X(t^n) - \left(X(t^{n+1}) - \Delta t \dot{X}(t^{n+1}) + \frac{\Delta t^2}{2}\ddot{X}(t^{n+1/2})\right) + \mathcal{O}(\Delta t^4) \\ &\quad = X(t^n) - \left(X(t^{n+1}) - \Delta t \dot{X}(t^{n+1}) + \frac{\Delta t^2}{2}\ddot{X}(t^{n+1})\right) + \mathcal{O}(\Delta t^3). \end{aligned}$$

It follows that

$$\sup\left\{\left|\partial^\alpha X(t^n) - \partial^\alpha \widetilde{X}(t^n)\right| \mid \forall (x,v) \in [0,L]\times[-R,R],\ |\alpha| \le 1\right\} \le C\Delta t^3.$$

We note that

$$X_{k,l}(t^n) = X(t^n; t^{n+1}, x_k, v_l), \qquad V_{k,l}(t^n) = V(t^n; t^{n+1}, x_k, v_l).$$

Finally we get

$$\begin{aligned}(\mathcal{T}_1 \circ \mathcal{T}_2 \circ \mathcal{T}_1 f(t^n))_{k,l} &= f(t^n, X_{k,l}(t^n) + \mathcal{O}(\Delta t^3), V_{k,l}(t^n) + \mathcal{O}(\Delta t^3)) \\ &= f(t^n, X_{k,l}(t^n), V_{k,l}(t^n)) + \nabla f(t^n, X_{k,l}(t^n), V_{k,l}(t^n)).\mathcal{O}(\Delta t^3) \\ &= f(t^{n+1}, X_{k,l}(t^{n+1}), V_{k,l}(t^{n+1})) + \nabla f(t^n, X_{k,l}(t^n), V_{k,l}(t^n)).\mathcal{O}(\Delta t^3) \\ &= f(t^{n+1}, x_k, v_l) + \nabla f(t^n, X_{k,l}(t^n), V_{k,l}(t^n)).\mathcal{O}(\Delta t^3)\end{aligned}$$

and

$$\|f(t^{n+1}) - \mathcal{T}_1 \circ \mathcal{T}_2 \circ \mathcal{T}_1 f(t^n)\|_{L^2_h(\Omega)} \leq C\Delta t^3 \|\nabla f\|_{l^\infty(0,T;L^\infty(\Omega))}.$$

Moreover we have

$$\begin{aligned}&\left|\nabla f(t^{n+1}, x_k, v_l) - (\nabla\mathcal{T}_1 \circ \nabla\mathcal{T}_2 \circ \nabla\mathcal{T}_1 f(t^n))_{k,l}\right| \\ &\qquad \leq \left|\left(\nabla\widetilde{X}_{k,l}(t^n), \nabla\widetilde{V}_{k,l}(t^n)\right) - (\nabla X_{k,l}(t^n), \nabla V_{k,l}(t^n))\right| \\ &\qquad\quad \cdot \left|\nabla f\left(t^n, \widetilde{X}_{k,l}(t^n), \widetilde{V}_{k,l}(t^n)\right)\right| + |(\nabla X_{k,l}(t^n), \nabla V_{k,l}(t^n))| \\ &\qquad\quad \cdot \left|\nabla f\left(t^n, \widetilde{X}_{k,l}(t^n), \widetilde{V}_{k,l}(t^n)\right) - \nabla f\left(t^n, X_{k,l}(t^n), V_{k,l}(t^n)\right)\right|,\end{aligned}$$

so that we get

$$\begin{aligned}&\left\|\nabla f(t^{n+1}) - \nabla\mathcal{T}_1 \circ \nabla\mathcal{T}_2 \circ \nabla\mathcal{T}_1 f(t^n)\right\|_{\mathbb{L}^2_h(\Omega)} \\ &\qquad \leq C\Delta t^3 \left(\|\nabla f\|_{l^\infty(0,T;L^\infty(\Omega))} + \left\|\nabla^2 f\right\|_{l^\infty(0,T;L^\infty(\Omega))}\right),\end{aligned}$$

which ends the proof. □

We go on with a proposition which gives information on the L^2_h-stability of the interpolation operators.

PROPOSITION 4.4. *Let $\mathcal{H}_h$ and $\mathcal{L}_h$ be the interpolation operators and $\tau_{\boldsymbol{\mu},\boldsymbol{\nu}}$ the translation operator defined in section* 3.1. *Then for all functions $f \in \mathscr{C}\left(0,T;\mathscr{C}^1_b(\Omega)\right)$, we get the estimates*

$$\|\tau_{\boldsymbol{\mu},0}\mathcal{L}_h f\|_{L^2_h(\Omega)} \leq \|f\|_{L^2_h(\Omega)}, \tag{4.2}$$

$$\|\tau_{\boldsymbol{\mu},0}\mathcal{H}_h f\|_{L^2_h(\Omega)} \leq \|f\|_{L^2_h(\Omega)}, \tag{4.3}$$

$$\|\tau_{\boldsymbol{\mu},0}\partial_x\mathcal{H}_h f\|_{L^2_h(\Omega)} \leq \|\partial_x f\|_{L^2_h(\Omega)}, \tag{4.4}$$

$$\|\tau_{0,\boldsymbol{\nu}}\mathcal{L}_h f\|_{L^2_h(\Omega)} \leq \|f\|_{L^2_h(\Omega)}, \tag{4.5}$$

$$\|\tau_{0,\boldsymbol{\nu}}\mathcal{H}_h f\|_{L^2_h(\Omega)} \leq \|f\|_{L^2_h(\Omega)}, \tag{4.6}$$

$$\|\tau_{0,\boldsymbol{\nu}}\partial_v\mathcal{H}_h f\|_{L^2_h(\Omega)} \leq \|\partial_v f\|_{L^2_h(\Omega)}. \tag{4.7}$$

Proof. We give only the proof for the estimates (4.2)–(4.4), because the proof of inequalities (4.5)–(4.7) is the same.

Let us start with the estimate (4.2). We have

$$\|\tau_{\boldsymbol{\mu},0}\mathcal{L}_h f\|_{L^2_h(\Omega)} = \|\mathcal{L}_h f\|_{L^2_h,\Delta_h^{\boldsymbol{\mu},0}}.$$

Let $\boldsymbol{\omega} = (\omega_x, \omega_v)$ be a two-component vector belonging to $\mathbb{Z}^2$. $\mathbf{z}$ is an abbreviation for (x, v), and $\mathbf{k}$ is an abbreviation for (k_x, k_v). k_x (resp., k_v) takes the values $2\pi\omega_x/L$

(resp., $2\pi\omega_v/(2R)$), where ω_x and ω_v are integers. If we note that $\mathbf{N} = (N_x, N_v)$ and $\mathbf{z}_{k,l} = (x_k, v_l)$, the Fourier series decomposition of f gives

$$f_{i,j} = \frac{1}{|\Omega|^{1/2}} \sum_{|\boldsymbol{\omega}| \le \mathbf{N}/2} \widehat{f}(\boldsymbol{\omega}) e^{i\langle \mathbf{k}(\boldsymbol{\omega}), \mathbf{z}_{i,j}\rangle}.$$

Therefore we get

$$\begin{aligned}(\mathcal{L}_h f)_{i+\mu_j,j} &= \frac{1}{|\Omega|^{1/2}} \sum_{|\boldsymbol{\omega}| \le \mathbf{N}/2} \sum_{k=i-1}^{i+2} \widehat{f}(\boldsymbol{\omega}) \ell_k^3(x_i + \mu_j \Delta x) e^{i\langle \mathbf{k}(\boldsymbol{\omega}), \mathbf{z}_{k,j}\rangle} \\ &= \frac{1}{|\Omega|^{1/2}} \sum_{|\boldsymbol{\omega}| \le \mathbf{N}/2} \widehat{f}(\boldsymbol{\omega}) \varrho(\mu_j, \omega_x) e^{i\langle \mathbf{k}(\boldsymbol{\omega}), \mathbf{z}_{i-1,j}\rangle},\end{aligned}$$

where

$$\varrho(\mu_j, \omega_x) = \sum_{k=0}^{3} \ell_k^3((1+\mu_j)\Delta x) e^{ik_x(\omega_x)x_k}.$$

As a consequence we obtain

$$\begin{aligned}\|\mathcal{L}_h f\|^2_{L_h^2, \Delta_h^{\boldsymbol{\mu},0}} &= \Delta x \Delta v \sum_{i=0}^{N_x} \sum_{j=0}^{N_v} (\mathcal{L}_h f)_{i+\mu_j,j} \overline{(\mathcal{L}_h f)}_{i+\mu_j,j} \\ &= \frac{\Delta x \Delta v}{|\Omega|} \sum_{i=0}^{N_x} \sum_{j=0}^{N_v} \sum_{|\boldsymbol{\omega}| \le \mathbf{N}/2} \sum_{|\boldsymbol{\omega}'| \le \mathbf{N}/2} \widehat{f}(\boldsymbol{\omega}) \overline{\widehat{f}(\boldsymbol{\omega}')} \varrho(\mu_j, \omega_x) \overline{\varrho(\mu_j, \omega_x')} \\ &\qquad e^{-i\langle \mathbf{k}(\boldsymbol{\omega}) - \mathbf{k}(\boldsymbol{\omega}'), (1,0)\rangle} e^{i\langle \mathbf{k}(\boldsymbol{\omega}) - \mathbf{k}(\boldsymbol{\omega}'), \mathbf{z}_{i,j}\rangle}.\end{aligned}$$

Since

$$\frac{\Delta x}{|L|} \sum_{i=0}^{N_x} e^{i(k_x(\omega_x) - k_x(\omega_x'))x_i} = \delta_{\omega_x, \omega_x'},$$

we find

$$\|\mathcal{L}_h f\|^2_{L_h^2, \Delta_h^{\boldsymbol{\mu},0}} = \frac{\Delta v}{|2R|} \sum_{j=0}^{N_v} \sum_{|\boldsymbol{\omega}| \le \mathbf{N}/2} \sum_{|\omega_v'| \le N_v/2} \widehat{f}(\boldsymbol{\omega}) \overline{\widehat{f}(\omega_x, \omega_v')} |\varrho(\mu_j, \omega_x)|^2 e^{i(k_v(\omega_v) - k_v(\omega_v'))v_j}.$$

Finally as

$$\frac{\Delta v}{|2R|} \sum_{j=0}^{N_v} e^{i(k_v(\omega_v) - k_v(\omega_v'))v_j} = \delta_{\omega_v, \omega_v'},$$

we get

$$\begin{aligned}\|\mathcal{L}_h f\|^2_{L_h^2, \Delta_h^{\boldsymbol{\mu},0}} &\le \sup\{|\varrho(\mu, \omega_x)|^2,\ |\omega_x \Delta x| \le L/2,\ 0 \le \mu \le 1\} \\ &\quad \times \frac{\Delta v}{|2R|} \sum_{j=0}^{N_v} \sum_{|\boldsymbol{\omega}| \le \mathbf{N}/2} \sum_{|\omega_v'| \le N_v/2} \widehat{f}(\boldsymbol{\omega}) \overline{\widehat{f}(\omega_x, \omega_v')} e^{i(k_v(\omega_v) - k_v(\omega_v'))v_j} \\ &\le \sup\{|\varrho(\mu, \omega_x)|^2,\ |\omega_x \Delta x| \le L/2,\ 0 \le \mu \le 1\} \sum_{|\boldsymbol{\omega}| \le \mathbf{N}/2} |\widehat{f}(\boldsymbol{\omega})|^2 \\ &\le \sup\{|\varrho(\mu, \omega_x)|^2,\ |\omega_x \Delta x| \le L/2,\ 0 \le \mu \le 1\} \|f\|^2_{L_h^2}.\end{aligned}$$

Without loss of generality we can suppose that $L = 2\pi$, $\Delta x = h$, $\xi = \omega_x h$, and $\zeta = \mu - 1/2$. Then we have to prove that

$$\sup\{|\varrho(\zeta,\xi)|,\ |\xi| \leq \pi,\ |\zeta| \leq 1/2\} \leq 1,$$

with

$$\varrho(\zeta,\xi) = \sum_{k=-3/2}^{3/2} \ell_k^3(\zeta) e^{ik\xi},$$

and

$$\ell_k^3(\zeta) = \prod_{\substack{i=-3/2 \\ i\neq k}}^{3/2} \frac{(\zeta - i)}{(k-i)},\ k \in \{-3/2, -1/2, 1/2, 3/2\}.$$

If we set $\theta = 1 - \cos(\xi)$, after some algebra we find

$$|\varrho(\zeta,\theta)|^2 = 1 - \left(\left(\frac{1}{2}\right)^2 - \zeta^2\right)\left(\left(\frac{3}{2}\right)^2 - \zeta^2\right)\theta^2\left[3 + 2\theta\left(\left(\frac{1}{2}\right)^2 - \zeta^2\right)\right]\Big/9,$$

which ends the proof of the estimate (4.2).

Proceeding in the same way, we get

$$\begin{aligned}
\|\tau_{\boldsymbol{\mu},0}\mathcal{H}_h f\|_{L_h^2(\Omega)} &= \|\mathcal{H}_h f\|_{L_h^2,\Delta_h^{\boldsymbol{\mu},0}} \\
&\leq \sup\{|\mathcal{Q}(\mu,\xi,h)|,\ 0 \leq \mu \leq 1,\ |\xi| \leq \pi\}\left(\sum_{|\boldsymbol{\omega}|\leq \mathbf{N}/2} |\widehat{f}(\boldsymbol{\omega})|^2\right)^{1/2} \\
&\leq \sup\{|\mathcal{Q}(\mu,\xi,h)|,\ 0 \leq \mu \leq 1,\ |\xi| \leq \pi\}\|f\|_{L_h^2},
\end{aligned}$$

where

$$\mathcal{Q}(\mu,\xi,h) = \phi_0(\mu h) + \phi_1(\mu h)e^{i\xi} + i\frac{\xi}{h}\psi_0(\mu h) + i\frac{\xi}{h}\psi_1(\mu h)e^{i\xi}$$

and

$$\begin{aligned}
\|\tau_{\boldsymbol{\mu},0}\partial_x\mathcal{H}_h f\|_{L_h^2(\Omega)} &= \|\partial_x\mathcal{H}_h f\|_{L_h^2,\Delta_h^{\boldsymbol{\mu},0}} \\
&\leq \sup\{|\widehat{\mathcal{Q}}(\mu,\xi,h)|,\ 0 \leq \mu \leq 1,\ |\xi| \leq \pi\}\left(\sum_{|\boldsymbol{\omega}|\leq \mathbf{N}/2} |\omega_x|^2|\widehat{f}(\boldsymbol{\omega})|^2\right)^{1/2} \\
&\leq \sup\{|\widehat{\mathcal{Q}}(\mu,\xi,h)|,\ 0 \leq \mu \leq 1,\ |\xi| \leq \pi\}\|\partial_x f\|_{L_h^2},
\end{aligned}$$

where

$$|\widehat{\mathcal{Q}}(\mu,\xi,h)| \cdot |\xi| = |\dot{\mathcal{Q}}(\mu,\xi,h)|,$$

with

$$\dot{\mathcal{Q}}(\mu,\xi,h) = \dot{\phi_0}(\mu h) + \dot{\phi_1}(\mu h)e^{i\xi} + i\xi\dot{\psi_0}(\mu h) + i\xi\dot{\psi_1}(\mu h)e^{i\xi}.$$

Therefore we have to prove that

$$\mathcal{Q}_{\sup} = \sup\{|\mathcal{Q}(\mu,\xi,h)|,\ 0 \le \mu \le 1,\ |\xi| \le \pi\} \le 1$$

and

$$\widehat{\mathcal{Q}}_{\sup} = \sup\{|\widehat{\mathcal{Q}}(\mu,\xi,h)|,\ 0 \le \mu \le 1,\ |\xi| \le \pi\} \le 1.$$

Let us start by $\mathcal{Q}_{\sup}$. First we note that $\mathcal{Q}$ can be rewritten as

$$\mathcal{Q}(\mu,\xi,h) = \mathcal{Q}(\mu,\xi) = (1+2\mu)(1-\mu)^2 + (3-2\mu)\mu^2 e^{i\xi} + i\xi\mu(1-\mu)^2 + i\xi\mu^2(\mu-1)e^{i\xi}.$$

If we take the modulus of $\mathcal{Q}$, after some algebra we get

$$\begin{aligned}|\mathcal{Q}(\mu,\xi)|^2 = 1 - \mu^2(1-\mu)^2 \big\{&2(1-\cos\xi)(3-2\mu)(1+2\mu) - \xi^2(4\mu^2-4\mu+1)\\ &- 2\,\xi^2(1-\cos\xi)\mu(1-\mu) - 2\xi\sin\xi(-4\mu^2+4\mu+1)\big\}.\end{aligned}$$

If we make the change of variable $\nu = \mu - 1/2$, after some algebra we get

$$|\mathcal{Q}(\nu,\xi)|^2 = 1 - \left(\left(\frac{1}{2}\right)^2 - \nu^2\right)^2 \sigma(\nu,\xi),$$

where

$$\sigma(\nu,\xi) = -\nu^2(4\sin(\xi/2) - 2\xi\sin(\xi/2))^2 + 8(1-\cos\xi) - \frac{1}{2}\xi^2(1-\cos\xi) - 4\xi\sin^2\xi.$$

In order to show that $\mathcal{Q}_{\sup} \le 1$ it is enough to show that $\sigma \ge 0$ for all $|\xi| \le \pi,\ |\nu| \le 1/2$. Since σ is a polynomial of degree two with respect to the variable ν whose term in ν^2 is negative, we have to prove that for all ξ such that $|\xi| \le \pi$ the roots of σ are real and of modulus greater than or equal to $1/2$. The roots of σ are

$$\nu_r(\xi) = \pm\sqrt{\frac{-8(1-\cos\xi) + \frac{1}{2}\xi^2(1-\cos\xi) + 4\xi\sin^2\xi}{-\nu^2(4\sin(\xi/2) - 2\xi\sin(\xi/2))^2}}.$$

We must prove that $|\nu_r(\xi)|^2 > \frac{1}{4}$, which is equivalent to showing that $g(\xi) \le 0$, where $g(\xi) = -6(1-\cos\xi) + \xi^2 + 2\xi\sin\xi$. In order to prove that g is negative on $[-\pi,\pi]$, we study the function g. For that we need to derive g three times.

Now, let us show that $\widehat{Q}_{sup} \le 1$. In order to prove that $\widehat{Q}_{sup} \le 1$ it is enough to prove that

$$|\dot{\mathcal{Q}}(\mu,\xi,h)|^2 - |\xi|^2 \le 0 \quad \forall |\xi| \le \pi,\ 0 \le \mu \le 1.$$

In fact $\dot{\mathcal{Q}}(\mu,\xi,h)$ can be recast as

$$\begin{aligned}\dot{\mathcal{Q}}(\mu,\xi,h) &= \dot{\mathcal{Q}}(\mu,\xi)\\ &= 6\mu(\mu-1) + 6\mu(1-\mu)e^{i\xi} + i\xi(1-\mu)(1-3\mu) + i\xi\mu(3\mu-2)e^{i\xi}.\end{aligned}$$

Taking the modulus of $\dot{\mathcal{Q}}$, we get

$$\begin{aligned}|\dot{\mathcal{Q}}(\mu,\xi)|^2 = {}&72\mu^2(1-\mu)^2(1-\cos\xi) + 12\xi\sin\xi\mu(1-\mu)(6\mu^2-6\mu+1)\\ &+ \xi^2(1-6\mu+6\mu^2)^2 + 2(\cos\xi-1)\xi^2\mu(1-\mu)(1-3\mu)(3\mu-2).\end{aligned}$$

If we make the change of variable $\nu = \mu - 1/2$, after some algebra we find

$$|\dot{\mathcal{Q}}(\nu,\xi)|^2 - |\xi|^2 = a(\xi)\nu^4 + b(\xi)\nu^2 + c(\xi),$$

where

$$a(\xi) = \left[12\sin\frac{\xi}{2} - 6\xi\cos\frac{\xi}{2}\right]^2, \quad c(\xi) = \left[\left(3\sin\frac{\xi}{2} - \frac{\xi}{2}\cos\frac{\xi}{2}\right)^2 - \xi^2\right],$$

$$b(\xi) = \left[-2\left(6\sin\frac{\xi}{2} - 2\xi\cos\frac{\xi}{2}\right)^2 + \xi^2\left(1 + \sin^2\frac{\xi}{2}\right)\right].$$

First we note that $|\dot{\mathcal{Q}}(\pm\frac{1}{2},\xi)|^2 - |\xi|^2 = 0$. Consequently $|\dot{\mathcal{Q}}|^2 - |\xi|^2$ can be rewritten as follows:

$$|\dot{\mathcal{Q}}(\nu,\xi)|^2 - |\xi|^2 = \left(\nu^2 - \left(\frac{1}{2}\right)^2\right)(\nu^2 + \beta(\xi)).$$

Let us show that β is positive. By following a classical result on the equation of second degree we get $\beta(\xi) = -4a(\xi)c(\xi)$. It remains to prove that $c(\xi) \leq 0$ for all $|\xi| \leq \pi$. We have $c(\xi) = l_1(\xi)l_2(\xi)$, where

$$l_1(\xi) = 3\sin\frac{\xi}{2} - \frac{\xi}{2}\cos\frac{\xi}{2} - \xi, \quad l_2(\xi) = 3\sin\frac{\xi}{2} - \frac{\xi}{2}\cos\frac{\xi}{2} + \xi.$$

In order to prove that $l_1 l_2$ is negative on $[-\pi,\pi]$, we study the sign of l_1 and l_2. For that we need to compute the derivatives of l_1 and l_2 until the order three. Finally we get

$$|\dot{\mathcal{Q}}(\nu,\xi)|^2 - |\xi|^2 \leq 0, \quad |\nu| \leq 1/2, \quad |\xi| \leq \pi,$$

which ends the proof of the proposition. □

We continue with the following lemma, which gives an estimate of the term (I2).

LEMMA 4.5. *Let $f \in \mathscr{C}_b\left(0,T;\mathscr{C}_b^2(\Omega)\right)$, and then*

$$\begin{aligned}
&\left\|\mathcal{T}_1^{0,0}\circ\mathcal{T}_2^{0,0}\circ\mathcal{T}_1^{0,0}f(t^n) - \mathcal{T}_1^{0,0}\circ\widehat{\mathcal{T}}_2^{0,0}\circ\mathcal{T}_1^{0,0}f(t^n)\right\|_{L_h^2(\Omega)} \\
&\qquad \leq C\Delta t\left(\Delta t^2 + \Delta x^4 + \Delta v^4 + \Delta x^5 + \Delta x\Delta v^4 + \left\|\left(\mathcal{T}_1^{0,0} - \widetilde{\mathcal{T}}_1^{0,0}\right)f(t^n)\right\|_{L_h^2(\Omega)}\right. \\
&\qquad + \left\|\widetilde{\mathcal{T}}_1^{0,0}\left(f(t^n) - f_h(t^n)\right)\right\|_{L_h^2(\Omega)} + \Delta x\left\|\left(\mathcal{T}_1^{1,0} - \widetilde{\mathcal{T}}_1^{1,0}\right)f(t^n)\right\|_{L_h^2(\Omega)} \\
&\qquad \left. + \Delta x\left\|\widetilde{\mathcal{T}}_1^{1,0}\left(f(t^n) - f_h(t^n)\right)\right\|_{L_h^2(\Omega)}\right)
\end{aligned}$$

and

$$\begin{aligned}
&\left\|\nabla\mathcal{T}_1\circ\nabla\mathcal{T}_2\circ\nabla\mathcal{T}_1 f(t^n) - \nabla\mathcal{T}_1\circ\nabla\widehat{\mathcal{T}}_2\circ\nabla\mathcal{T}_1 f(t^n)\right\|_{\mathbb{L}_h^2(\Omega)} \\
&\leq C\Delta t\left(\Delta t^2 + \Delta x^3 + \Delta x^{-1}\Delta v^4 + \Delta v^4 + \Delta x^4 + \Delta x^{-1}\left\|\left(\mathcal{T}_1^{0,0} - \widetilde{\mathcal{T}}_1^{0,0}\right)f(t^n)\right\|_{L_h^2(\Omega)}\right. \\
&+\Delta x^{-1}\left\|\widetilde{\mathcal{T}}_1^{0,0}\left(f(t^n) - f_h(t^n)\right)\right\|_{L_h^2(\Omega)} + \left\|\left(\mathcal{T}_1^{1,0} - \widetilde{\mathcal{T}}_1^{1,0}\right)f(t^n)\right\|_{L_h^2(\Omega)} \\
&\left. + \left\|\widetilde{\mathcal{T}}_1^{1,0}\left(f(t^n) - f_h(t^n)\right)\right\|_{L_h^2(\Omega)}\right).
\end{aligned}$$

Proof. We introduce the characteristic curves $(\widehat{X}(t^n), \widehat{V}(t^n))$ defined by

$$\widehat{X}(t^n, x, v) = x - v\Delta t - \frac{1}{2}\Delta t^2 E_h^{n+1/2}(x - v\Delta t/2),$$
$$\widehat{V}(t^n, x, v) = v - \Delta t E_h^{n+1/2}(x - v\Delta t/2).$$

As has been done in Lemma 4.3 we have

$$\left(\mathcal{T}_1^{0,0} \circ \widehat{\mathcal{T}}_2^{0,0} \circ \mathcal{T}_1^{0,0} f(t^n)\right)_{k,l} = f\left(t^n, \widehat{X}_{k,l}(t^n), \widehat{V}_{k,l}(t^n)\right),$$
$$\left(\nabla\mathcal{T}_1 \circ \nabla\widehat{\mathcal{T}}_2 \circ \nabla\mathcal{T}_1 f(t^n)\right)_{k,l} = \nabla\left(\widehat{X}_{k,l}(t^n), \widehat{V}_{k,l}(t^n)\right) \nabla f\left(t^n, \widehat{X}_{k,l}(t^n), \widehat{V}_{k,l}(t^n)\right).$$

As a consequence we can write

(4.8)
$$\left|\left(\mathcal{T}_1^{0,0} \circ \mathcal{T}_2^{0,0} \circ \mathcal{T}_1^{0,0} f(t^n)\right)_{k,l} - \left(\mathcal{T}_1^{0,0} \circ \widehat{\mathcal{T}}_2^{0,0} \circ \mathcal{T}_1^{0,0} f(t^n)\right)_{k,l}\right|$$
$$\leq \left|\nabla f\left(\widetilde{X}_{k,l}(t^n), \widetilde{V}_{k,l}(t^n)\right)\right| \cdot \left(\left|\widehat{X}_{k,l}(t^n) - \widetilde{X}_{k,l}(t^n)\right| + \left|\widehat{V}_{k,l}(t^n) - \widetilde{V}_{k,l}(t^n)\right|\right)$$

and

(4.9)
$$\left|(\nabla\mathcal{T}_1 \circ \nabla\mathcal{T}_2 \circ \nabla\mathcal{T}_1 f(t^n))_{k,l} - \left(\nabla\mathcal{T}_1 \circ \nabla\widehat{\mathcal{T}}_2 \circ \nabla\mathcal{T}_1 f(t^n)\right)_{k,l}\right|$$
$$\leq \left|\nabla f\left(\widetilde{X}_{k,l}(t^n), \widetilde{V}_{k,l}(t^n)\right)\right| \cdot \left(\left|\nabla\widehat{X}_{k,l}(t^n) - \nabla\widetilde{X}_{k,l}(t^n)\right| + \left|\nabla\widehat{V}_{k,l}(t^n) - \nabla\widetilde{V}_{k,l}(t^n)\right|\right)$$
$$+ \left(\left|\widehat{X}_{k,l}(t^n) - \widetilde{X}_{k,l}(t^n)\right| + \left|\widehat{V}_{k,l}(t^n) - \widetilde{V}_{k,l}(t^n)\right|\right)$$
$$\cdot \left|\nabla^2 f\left(\widetilde{X}_{k,l}(t^n), \widetilde{V}_{k,l}(t^n)\right)\right| \cdot \left|\nabla\left(\widetilde{X}_{k,l}(t^n), \widetilde{V}_{k,l}(t^n)\right)\right|.$$

Let us prove a bound for $|\partial^\alpha \widehat{X}_{k,l}(t^n) - \partial^\alpha \widetilde{X}_{k,l}(t^n)|$ and $|\partial^\alpha \widehat{V}_{k,l}(t^n) - \partial^\alpha \widetilde{V}_{k,l}(t^n)|$. By using the definition of $(\widetilde{X}(t^n), \widetilde{V}(t^n))$ and of $(\widehat{X}(t^n), \widehat{V}(t^n))$ we get

(4.10)
$$\left(\left|\widehat{X}_{k,l}(t^n) - \widetilde{X}_{k,l}(t^n)\right| + \left|\widehat{V}_{k,l}(t^n) - \widetilde{V}_{k,l}(t^n)\right|\right)$$
$$\leq C\Delta t \left\|E_h^{n+1/2}(x) - E(t^{n+1/2}, x)\right\|_{L^\infty}$$
$$\leq C\Delta t \left(\left\|E(t^{n+1/2}, x) - \widetilde{E}(t^{n+1/2}, x)\right\|_{L^\infty} + \left\|\widetilde{E}(t^{n+1/2}, x) - E_h^{n+1/2}(x)\right\|_{L^\infty}\right),$$

$$\left(\left|\partial_x\widehat{X}_{k,l}(t^n) - \partial_x\widetilde{X}_{k,l}(t^n)\right| + \left|\partial_x\widehat{V}_{k,l}(t^n) - \partial_x\widetilde{V}_{k,l}(t^n)\right|\right)$$
$$\leq C\Delta t \left(\left\|\partial_x E(t^{n+1/2}, x) - \partial_x\widetilde{E}(t^{n+1/2}, x)\right\|_{L^\infty}\right.$$
$$\left. + \left\|\partial_x\widetilde{E}(t^{n+1/2}, x) - \partial_x E_h^{n+1/2}(x)\right\|_{L^\infty}\right) \tag{4.11}$$

and

$$\left(\left|\partial_v\widehat{X}_{k,l}(t^n) - \partial_v\widetilde{X}_{k,l}(t^n)\right| + \left|\partial_v\widehat{V}_{k,l}(t^n) - \partial_v\widetilde{V}_{k,l}(t^n)\right|\right)$$
$$\leq C\Delta t^2 \left(\left\|\partial_x E(t^{n+1/2}, x) - \partial_x\widetilde{E}(t^{n+1/2}, x)\right\|_{L^\infty}\right.$$
$$\left. + \left\|\partial_x\widetilde{E}(t^{n+1/2}, x) - \partial_x E_h^{n+1/2}(x)\right\|_{L^\infty}\right), \tag{4.12}$$

where $\widetilde{E}(t^{n+1/2},x)$ and $\partial_x\widetilde{E}(t^{n+1/2},x)$ are solutions of the Poisson problems

$$(\widetilde{\mathrm{P}})\quad\begin{cases}\left(\dfrac{d}{dx}\widetilde{E}\right)(t^{n+1/2},x)=\displaystyle\int_v \mathcal{T}_1 f(t^n,x,v)dv-1=\int_v f(t^n,x-v\Delta t/2,v)dv-1,\\ \displaystyle\int_0^L \widetilde{E}(t^{n+1/2},x)dx=0\end{cases}$$

and

$$(\widetilde{\mathrm{P}}')\quad\begin{cases}\left(\dfrac{d}{dx}\left(\dfrac{d}{dx}\widetilde{E}\right)\right)(t^{n+1/2},x)=\displaystyle\int_v \mathcal{T}_1^{1,0} f(t^n,x,v)dv=\int_v \partial_x f(t^n,x-v\Delta t/2,v)dv,\\ \displaystyle\int_0^L \left(\frac{d}{dx}\widetilde{E}\right)(t^{n+1/2},x)dx=0.\end{cases}$$

First we give an estimate for $\|E(t^{n+1/2},x)-\widetilde{E}(t^{n+1/2},x)\|_{L^\infty}$ and for $\|\partial_x E(t^{n+1/2},x)-\partial_x\widetilde{E}(t^{n+1/2},x)\|_{L^\infty}$. By using the Vlasov equation we get that

$$\begin{aligned}\frac{f(t^{n+1/2},x,v)-f(t^n,x-v\Delta t/2,v)}{\Delta t/2}&=\partial_t f(t^{n+1/2},x,v)+v\partial_x f(t^{n+1/2},x,v)+\mathcal{O}(\Delta t)\\&=-E(t^{n+1/2},x)\partial_v f(t^{n+1/2},x,v)+\mathcal{O}(\Delta t),\end{aligned}$$

and it follows that

(4.13)
$$\begin{aligned}&\left\|E(t^{n+1/2})-\widetilde{E}(t^{n+1/2})\right\|_{L^\infty}\\&\quad=\left\|\int_0^L K(x,y)\left(\int_{-\infty}^{+\infty}[f(t^{n+1/2},y,v)-f(t^n,y-v\Delta t/2,v)]dv\right)dy\right\|_{L^\infty}\\&\quad\le C\Delta t^2.\end{aligned}$$

In the same manner, if we derive the Vlasov equation with respect to the variable x, we get

$$\begin{aligned}\frac{\partial_x f(t^{n+1/2},x,v)-\partial_x f(t^n,x-v\Delta t/2,v)}{\Delta t/2}&=-\partial_x E(t^{n+1/2},x)\partial_v f(t^{n+1/2},x,v)\\&\quad-E(t^{n+1/2},x)\partial_{vx} f(t^{n+1/2},x,v)+\mathcal{O}(\Delta t),\end{aligned}$$

and we find

(4.14)
$$\begin{aligned}&\left\|\partial_x E(t^{n+1/2})-\partial_x\widetilde{E}(t^{n+1/2})\right\|_{L^\infty}\\&\quad=\left\|\int_0^L K(x,y)\left(\int_{-\infty}^{+\infty}[\partial_x f(t^{n+1/2},y,v)-\partial_x f(t^n,y-v\Delta t/2,v)]dv\right)dy\right\|_{L^\infty}\\&\quad\le C\Delta t^2.\end{aligned}$$

Now we estimate the terms $\|E_h^{n+1/2}(x)-\widetilde{E}(t^{n+1/2},x)\|_{L^\infty}$ and $\|\partial_x E_h^{n+1/2}(x)-\partial_x\widetilde{E}(t^{n+1/2},x)\|_{L^\infty}$. Until the end of the proof the following notation is used:

$$\left\|\widetilde{E}(t^{n+1/2})-E_h^{n+1/2}\right\|_\infty=\max_k\left|\widetilde{E}(t^{n+1/2},x_k)-E_h^{n+1/2}(x_k)\right|,$$

$$\left|\widetilde{E}(t^{n+1/2}) - E_h^{n+1/2}\right|_{1,\infty} = \max_k \left|\partial_x \widetilde{E}(t^{n+1/2}, x_k) - \partial_x E_h^{n+1/2}(x_k)\right|,$$

and

$$\|\cdot\|_{1,\infty} = \|\cdot\|_\infty + |\cdot|_{1,\infty}.$$

Suppose that $x = x_k + \theta\Delta x$, with $0 \le \theta \le 1$, and then we get

$$\begin{aligned}
&\left|\widetilde{E}(t^{n+1/2}, x) - E_h^{n+1/2}(x)\right| \le \left|\widetilde{E}(t^{n+1/2}, x) - \phi_0(\theta\Delta x)\widetilde{E}(t^{n+1/2}, x_k)\right.\\
&\left. - \phi_1(\theta\Delta x)\widetilde{E}(t^{n+1/2}, x_{k+1}) - \psi_0(\theta\Delta x)\partial_x\widetilde{E}(t^{n+1/2}, x_k) - \psi_1(\theta\Delta x)\partial_x\widetilde{E}(t^{n+1/2}, x_{k+1})\right|\\
&+ \left|\phi_0(\theta\Delta x)\left(\widetilde{E}(t^{n+1/2}, x_k) - E_h^{n+1/2}(x_k)\right)\right.\\
&+ \phi_1(\theta\Delta x)\left(\widetilde{E}(t^{n+1/2}, x_{k+1}) - E_h^{n+1/2}(x_{k+1})\right)\\
&+\psi_0(\theta\Delta x)\left(\partial_x\widetilde{E}(t^{n+1/2}, x_k) - \partial_x E_h^{n+1/2}(x_k)\right)\\
&\left.+ \psi_1(\theta\Delta x)\left(\partial_x\widetilde{E}(t^{n+1/2}, x_{k+1}) - \partial_x E_h^{n+1/2}(x_{k+1})\right)\right|.
\end{aligned}$$

We deduce that

$$\begin{aligned}
\left\|\widetilde{E}(t^{n+1/2}) - E_h^{n+1/2}\right\|_{L^\infty} \le C\left(\Delta x^4 \left\|\widetilde{E}\right\|_{\mathscr{C}(0,T;\mathscr{C}^4([0,L]))} + \left\|\widetilde{E}(t^{n+1/2}) - E_h^{n+1/2}\right\|_\infty\right.\\
\left. + \Delta x\left|\widetilde{E}(t^{n+1/2}) - E_h^{n+1/2}\right|_{1,\infty}\right).
\end{aligned} \tag{4.15}$$

In the same way we show that

$$\begin{aligned}
\left\|\partial_x\widetilde{E}(t^{n+1/2}) - \partial_x E_h^{n+1/2}\right\|_{L^\infty} \le C\left(\Delta x^4 \left\|\widetilde{E}\right\|_{\mathscr{C}(0,T;\mathscr{C}^4([0,L]))}\right.\\
\left. + \Delta x^{-1}\left\|\widetilde{E}(t^{n+1/2}) - E_h^{n+1/2}\right\|_\infty + \left|\widetilde{E}(t^{n+1/2}) - E_h^{n+1/2}\right|_{1,\infty}\right).
\end{aligned} \tag{4.16}$$

Now we estimate the terms $\|\widetilde{E}(t^{n+1/2}) - E_h^{n+1/2}\|_\infty$ and $|\widetilde{E}(t^{n+1/2}) - E_h^{n+1/2}|_{1,\infty}$. We proceed as follows:

$$\begin{aligned}
&\frac{1}{12\Delta x}\left\{8\left(\widetilde{E}(t^{n+1/2}, x_{k+1}) - \widetilde{E}(t^{n+1/2}, x_{k-1})\right) - \left(\widetilde{E}(t^{n+1/2}, x_{k+2}) - \widetilde{E}(t^{n+1/2}, x_{k-2})\right)\right\}\\
&\qquad = \partial_x\widetilde{E}(t^{n+1/2}, x_k) + \mathcal{O}\left(\Delta x^4\right)\\
&\qquad = -1 + \int_v \mathcal{T}_1^{0,0} f(t^n, x_k, v)dv + \mathcal{O}\left(\Delta x^4\right)\\
&\qquad = -1 + \mathcal{O}\left(\Delta x^4\right) + \mathcal{O}\left(\Delta v^4\right)\\
&\qquad + \frac{\Delta v}{3}\left\{\left(\mathcal{T}_1^{0,0} f(t^n)\right)_{k,0} + \left(\mathcal{T}_1^{0,0} f(t^n)\right)_{k,N_v} + 4\sum_{l,\text{odd}}\left(\mathcal{T}_1^{0,0} f(t^n)\right)_{k,l}\right.\\
&\qquad\qquad \left. + 2\sum_{l,\text{even}}\left(\mathcal{T}_1^{0,0} f(t^n)\right)_{k,l}\right\}.
\end{aligned}$$

If we set $\delta E_k = \widetilde{E}(t^{n+1/2}, x_k) - E_h^{n+1/2}(x_k)$, we get

(4.17)
$$\begin{aligned}
&\frac{1}{12\Delta x}\left\{8\left(\delta E_{k+1} - \delta E_{k-1}\right) - \left(\delta E_{k+2} - \delta E_{k-2}\right)\right\} = \mathcal{O}\left(\Delta x^4\right) + \mathcal{O}\left(\Delta v^4\right) \\
&\quad + \frac{\Delta v}{3}\Bigg\{\left(\mathcal{T}_1^{0,0} f(t^n)\right)_{k,0} - \left(\widetilde{\mathcal{T}}_1^{0,0} f_h(t^n)\right)_{k,0} + \left(\mathcal{T}_1^{0,0} f(t^n)\right)_{k,N_v} - \left(\widetilde{\mathcal{T}}_1^{0,0} f_h(t^n)\right)_{k,N_v} \\
&\qquad + 4\sum_{l,\text{odd}}\left[\left(\mathcal{T}_1^{0,0} f(t^n)\right)_{k,l} - \left(\widetilde{\mathcal{T}}_1^{0,0} f_h(t^n)\right)_{k,l}\right] \\
&\qquad + 2\sum_{l,\text{even}}\left[\left(\mathcal{T}_1^{0,0} f(t^n)\right)_{k,l} - \left(\widetilde{\mathcal{T}}_1^{0,0} f_h(t^n)\right)_{k,l}\right]\Bigg\}.
\end{aligned}$$

By using the periodicity of δE_k we obtain

(4.18)
$$\begin{aligned}
&\frac{1}{12\Delta x}\sum_{k=0}^{N_x}\left|8(\delta E_{k+1} - \delta E_{k-1}) - (\delta E_{k+2} - \delta E_{k-2})\right| \\
&\quad \geq \frac{1}{12\Delta x}\sum_{k=0}^{N_x}\left|8\left|\delta E_{k+1} - \delta E_{k-1}\right| - \left|\delta E_{k+2} - \delta E_{k-2}\right|\right| \\
&\quad \geq \frac{1}{12\Delta x}\left|8\sum_{k=0}^{N_x}\left|\delta E_{k+1} - \delta E_{k-1}\right| - \sum_{k=0}^{N_x}\left|\delta E_{k+2} - \delta E_{k-2}\right|\right| \\
&\quad \geq \frac{1}{12\Delta x}\left\{8\sum_{k=0}^{N_x}\left|\delta E_{k+1} - \delta E_{k-1}\right| - \sum_{k=0}^{N_x}\left|\delta E_{k+2} - \delta E_k\right| - \sum_{k=0}^{N_x}\left|\delta E_k - \delta E_{k-2}\right|\right\} \\
&\quad \geq \frac{1}{2\Delta x}\sum_{k=0}^{N_x}\left|\delta E_{k+1} - \delta E_{k-1}\right|.
\end{aligned}$$

By using (4.17), (4.18), and the Cauchy–Schwarz inequality we obtain

(4.19)
$$\begin{aligned}
&\sum_{k=0}^{N_x}\left|\delta E_{k+1} - \delta E_{k-1}\right| \leq \mathcal{O}\left(\Delta x^4\right) + \mathcal{O}\left(\Delta v^4\right) \\
&\qquad + C\Delta x\Delta v \sum_{(k,l)\in\mathcal{M}_h}\left|\left(\mathcal{T}_1^{0,0} f(t^n)\right)_{k,l} - \left(\widetilde{\mathcal{T}}_1^{0,0} f_h(t^n)\right)_{k,l}\right| \\
&\qquad\quad \leq \mathcal{O}\left(\Delta x^4\right) + \mathcal{O}\left(\Delta v^4\right) + C\left(\left\|\left(\mathcal{T}_1^{0,0} - \widetilde{\mathcal{T}}_1^{0,0}\right) f(t^n)\right\|_{L_h^1(\Omega)}\right. \\
&\qquad\qquad \left. + \left\|\widetilde{\mathcal{T}}_1^{0,0}\left(f(t^n) - f_h(t^n)\right)\right\|_{L_h^1(\Omega)}\right) \\
&\qquad\quad \leq \mathcal{O}\left(\Delta x^4\right) + \mathcal{O}\left(\Delta v^4\right) + C\sqrt{|\Omega|}\left(\left\|\left(\mathcal{T}_1^{0,0} - \widetilde{\mathcal{T}}_1^{0,0}\right) f(t^n)\right\|_{L_h^2(\Omega)}\right. \\
&\qquad\qquad \left. + \left\|\widetilde{\mathcal{T}}_1^{0,0}\left(f(t^n) - f_h(t^n)\right)\right\|_{L_h^2(\Omega)}\right).
\end{aligned}$$

We set $\mathcal{M}$ as a mean value of δE_k

$$\mathcal{M} = \frac{\Delta v}{3L}\left\{\delta E_0 + \delta E_{N_x} + 4\sum_{l,\text{odd}} \delta E_l + 2\sum_{l,\text{even}} \delta E_l\right\},$$

and then

$$\min_k \ \delta E_k \le |\mathcal{M}|.$$

By using a discrete version of the Taylor formula with the integral remainder we get

$$\begin{aligned}\max_k \ \delta E_k &\le \min_k \ \delta E_k + \sum_{k=0}^{N_x} |\delta E_{k+1} - \delta E_{k-1}| \\ &\le |\mathcal{M}| + \sum_{k=0}^{N_x} |\delta E_{k+1} - \delta E_{k-1}|\,. \end{aligned} \tag{4.20}$$

Thanks to the discrete zero-mean electrostatic condition on E_h (see Problem (P_h)) we get

(4.21)

$$\begin{aligned}&\delta E_0 + \delta E_{N_x} + 4\sum_{l,\text{odd}} \delta E_l + 2\sum_{l,\text{even}} \delta E_l \\ &\quad= \widetilde{E}(t^{n+1/2}, x_0) + \widetilde{E}(t^{n+1/2}, x_{N_x}) + 4\sum_{l,\text{odd}} \widetilde{E}(t^{n+1/2}, x_l) + 2\sum_{l,\text{even}} \widetilde{E}(t^{n+1/2}, x_l) \\ &\quad\quad - \left\{E_h^{n+1/2}(x_0) + E_h^{n+1/2}(x_{N_x}) + 4\sum_{l,\text{odd}} E_h^{n+1/2}(x_l) + 2\sum_{l,\text{even}} E_h^{n+1/2}(x_l)\right\} \\ &\quad= \int_0^L \widetilde{E}(t^{n+1/2}, x)dx + \mathcal{O}(\Delta x^4) = \mathcal{O}(\Delta x^4).\end{aligned}$$

Finally, using (4.21) and (4.20) gives

$$|\delta E_k| \le \mathcal{O}(\Delta x^4) + \sum_{k=0}^{N_x} |\delta E_{k+1} - \delta E_{k-1}|\,. \tag{4.22}$$

Thanks to (4.22) and (4.19) we obtain

$$\begin{aligned}\left\|\widetilde{E}(t^{n+1/2}) - E_h^{n+1/2}\right\|_\infty \le C\Bigg(&\Delta x^4 + \Delta v^4 + \left\|\left(\mathcal{T}_1^{0,0} - \widetilde{\mathcal{T}}_1^{0,0}\right) f(t^n)\right\|_{L_h^2(\Omega)} \\ &+ \left\|\widetilde{\mathcal{T}}_1^{0,0}\left(f(t^n) - f_h(t^n)\right)\right\|_{L_h^2(\Omega)}\Bigg).\end{aligned} \tag{4.23}$$

In the same manner we get

$$\begin{aligned}\left|\widetilde{E}(t^{n+1/2}) - E_h^{n+1/2}\right|_{1,\infty} \le C\Bigg(&\Delta x^4 + \Delta v^4 + \left\|\left(\mathcal{T}_1^{1,0} - \widetilde{\mathcal{T}}_1^{1,0}\right) f(t^n)\right\|_{L_h^2(\Omega)} \\ &+ \left\|\widetilde{\mathcal{T}}_1^{1,0}\left(f(t^n) - f_h(t^n)\right)\right\|_{L_h^2(\Omega)}\Bigg).\end{aligned} \tag{4.24}$$

By putting together the estimates (4.8), (4.10), (4.13), (4.15), and (4.23)–(4.24) we obtain the first estimate of Lemma 4.5. Therefore by assembling the estimates (4.9)–(4.16) and (4.23)–(4.24) we get the second estimate of Lemma 4.5. ∎

We continue the proof, with the following proposition which states the H^1-stability of the transport operators.

PROPOSITION 4.6. *Let $f \in \mathscr{C}_b(0,T,\mathscr{C}_b^1(\Omega))$, and then there exists a constant C independent of Δx, Δv, and Δt such that*

$$\left\|\widetilde{\mathcal{T}}_1^{0,0} f(t)\right\|_{L_h^2(\Omega)} \leq \|f(t)\|_{L_h^2(\Omega)}, \tag{4.25}$$

$$\left\|\widetilde{\mathcal{T}}_1^{1,0} f(t)\right\|_{L_h^2(\Omega)} \leq \|\partial_x f(t)\|_{L_h^2(\Omega)}, \tag{4.26}$$

$$\left\|\widetilde{\mathcal{T}}_1^{0,1} f(t)\right\|_{L_h^2(\Omega)} \leq \|\partial_v f(t)\|_{L_h^2(\Omega)} + C\Delta t \|\partial_x f(t)\|_{L_h^2(\Omega)}, \tag{4.27}$$

$$\left\|\widetilde{\mathcal{T}}_2^{0,0} f(t)\right\|_{L_h^2(\Omega)} \leq \|f(t)\|_{L_h^2(\Omega)}, \tag{4.28}$$

$$\left\|\widetilde{\mathcal{T}}_2^{1,0} f(t)\right\|_{L_h^2(\Omega)} \leq \|\partial_x f(t)\|_{L_h^2(\Omega)} + C\Delta t \|\partial_v f(t)\|_{L_h^2(\Omega)}, \tag{4.29}$$

$$\left\|\widetilde{\mathcal{T}}_2^{0,1} f(t)\right\|_{L_h^2(\Omega)} \leq \|\partial_v f(t)\|_{L_h^2(\Omega)}. \tag{4.30}$$

Proof. By using (4.3) we get

$$\left\|\widetilde{\mathcal{T}}_1^{0,0} f\right\|_{L_h^2(\Omega)} = \left\|\mathbb{S}_{v\Delta t/2,0}\mathcal{H}_h f\right\|_{L_h^2(\Omega)} \leq \|f\|_{L_h^2(\Omega)}.$$

From (4.4) it follows that

$$\left\|\widetilde{\mathcal{T}}_1^{1,0} f\right\|_{L_h^2(\Omega)} = \left\|\mathbb{S}_{v\Delta t/2,0}\partial_x\mathcal{H}_h f\right\|_{L_h^2(\Omega)} \leq \|\partial_x f\|_{L_h^2(\Omega)}.$$

Thanks to (4.2) we find that

$$\begin{aligned}\left\|\widetilde{\mathcal{T}}_1^{0,1} f\right\|_{L_h^2(\Omega)} &\leq \left\|\mathbb{S}_{v\Delta t/2,0}\mathcal{L}_h\partial_v f\right\|_{L_h^2(\Omega)} + \frac{\Delta t}{2}\left\|\widetilde{\mathcal{T}}_1^{1,0} f\right\|_{L_h^2(\Omega)} \\ &\leq \|\partial_v f\|_{L_h^2(\Omega)} + C\Delta t \|\partial_x f\|_{L_h^2(\Omega)}.\end{aligned}$$

By using (4.3) we get

$$\left\|\widetilde{\mathcal{T}}_2^{0,0} f\right\|_{L_h^2(\Omega)} = \|\mathbb{S}_{0,\Delta t E_h}\mathcal{H}_h f\|_{L_h^2(\Omega)} \leq \|f\|_{L_h^2(\Omega)},$$

and from (4.6) it follows that

$$\left\|\widetilde{\mathcal{T}}_2^{0,1} f\right\|_{L_h^2(\Omega)} = \|\mathbb{S}_{0,\Delta t E_h}\partial_v\mathcal{H}_h f\|_{L_h^2(\Omega)} \leq \|\partial_v f\|_{L_h^2(\Omega)}.$$

Now let us show that $|E_h^{n+1/2}|_{1,\infty}$ is bounded. To do this, we suppose that $\|\partial_x f_h(t^n)\|_{L_h^2(\Omega)}$

is bounded. Then from Problem (P'_h) we get that

$$\begin{aligned}\sum_{k=0}^{N_x}\left|\partial_x E_h^{n+1/2}(x_{k+1})-\partial_x E_h^{n+1/2}(x_{k-1})\right| &\le 2L + C\Delta x\Delta v\sum_{k,l}\left|\left(\widetilde{\mathcal{T}}_1^{1,0}f_h(t^n)\right)_{k,l}\right| \\ &\le L + C\left\|\widetilde{\mathcal{T}}_1^{1,0}f_h(t^n)\right\|_{L_h^1(\Omega)} \\ &\le L + C\sqrt{|\Omega|}\left\|\widetilde{\mathcal{T}}_1^{1,0}f_h(t^n)\right\|_{L_h^2(\Omega)} \\ &\le C,\end{aligned}$$

and it follows that

$$\begin{aligned}\left|E_h^{n+1/2}\right|_{1,\infty} \le \frac{\Delta v}{3L}\Big|&\partial_x E_h^{n+1/2}(x_0)+\partial_x E_h^{n+1/2}(x_{N_x}) \\ &+4\sum_{l,\mathrm{odd}}\partial_x E_h^{n+1/2}(x_l)+2\sum_{l,\mathrm{even}}\partial_x E_h^{n+1/2}(x_l)\Big| \\ &+\sum_{k=0}^{N_x}\left|\partial_x E_h^{n+1/2}(x_{k+1})-\partial_x E_h^{n+1/2}(x_{k-1})\right| \\ &\le C.\end{aligned}$$

Now let us prove inequality (4.29). By using (4.5) it follows that

$$\begin{aligned}\left\|\widetilde{\mathcal{T}}_2^{1,0}f\right\|_{L_h^2(\Omega)} &\le \left\|\mathbb{S}_{0,\Delta t E_h}\mathcal{L}_h\partial_x f\right\|_{L_h^2(\Omega)} + \frac{\Delta t}{2}\left\|\partial_x E_h^{n+1/2}\right\|_\infty\left\|\widetilde{\mathcal{T}}_2^{0,1}f\right\|_{L_h^2(\Omega)} \\ &\le \|\partial_x f\|_{L_h^2(\Omega)} + C\Delta t\,\|\partial_v f\|_{L_h^2(\Omega)},\end{aligned}$$

which ends the proof of the lemma. ∎

We continue with the following lemma, which gives an estimate of the precision of the approximation of the transport operators.

LEMMA 4.7. *Let $f\in\mathscr{C}_b(0,T;\mathscr{C}_b^4(\Omega))$, and then there exists a constant C independent of Δx, Δv, and Δt such that for $|\alpha|\le 1$ we have*

$$\begin{aligned}\left\|\left(\mathcal{T}_1^\alpha-\widetilde{\mathcal{T}}_1^\alpha\right)f(t)\right\|_{L_h^2(\Omega)} &\le C\Delta x^{4-|\alpha|}\,\|f\|_{\mathscr{C}_b\left(0,T;\mathscr{C}_b^4(\Omega)\right)}, \\ \left\|\left(\widehat{\mathcal{T}}_2^\alpha-\widetilde{\mathcal{T}}_2^\alpha\right)f(t)\right\|_{L_h^2(\Omega)} &\le C\Delta v^{4-|\alpha|}\,\|f\|_{\mathscr{C}_b\left(0,T;\mathscr{C}_b^4(\Omega)\right)}.\end{aligned}$$

Proof. If $f\in\mathscr{C}^{2m}([a,b])$ and $\pi_h f$ is the Hermite polynomial of degree $2m-1$ which interpolates f and its $m-1$ first derivatives at the points a and b, then

$$\left\|f^{(j)}-\pi_h f^{(j)}\right\|_{L^\infty([a,b])} \le \frac{\left\|f^{(2m)}\right\|_{L^\infty([a,b])}}{(2m-j)!2^{2m-[j]}}h^{2m-j},\ 0\le j\le 2m-1,$$

where $[j]=j$ if j is even and $[j]=j+1$ if j is odd. Then we have

$$\begin{aligned}\left\|\left(\mathcal{T}_1^{0,0}-\widetilde{\mathcal{T}}_1^{0,0}\right)f(t)\right\|_{L_h^2(\Omega)} &\le C\left\|\mathbb{S}_{v\Delta t/2,0}\left(f(t)-\mathcal{H}_h f(t)\right)\right\|_{L_h^\infty(\Omega)} \\ &\le C\left\|\mathbb{S}_{v\Delta t/2,0}\left(f(t)-\mathcal{H}_h f(t)\right)\right\|_{L_x^\infty(L_{h,v}^\infty)} \\ &\le C\left\|f(t)-\mathcal{H}_h f(t)\right\|_{L_x^\infty(L_{h,v}^\infty)} \le C\Delta x^4\,\|f\|_{\mathscr{C}_b\left(0,T;\mathscr{C}_b^4(\Omega)\right)},\end{aligned}$$

$$\begin{aligned}
\left\|\left(\mathcal{T}_1^{1,0}-\widetilde{\mathcal{T}}_1^{1,0}\right)f(t)\right\|_{L_h^2(\Omega)} &\leq C\left\|\mathbb{S}_{v\Delta t/2,0}\left(\partial_x f(t)-\partial_x\mathcal{H}_h f(t)\right)\right\|_{L_h^\infty(\Omega)}\\
&\leq C\left\|\mathbb{S}_{v\Delta t/2,0}\left(\partial_x f(t)-\partial_x\mathcal{H}_h f(t)\right)\right\|_{L_x^\infty(L_{h,v}^\infty)}\\
&\leq C\left\|\partial_x f(t)-\partial_x\mathcal{H}_h f(t)\right\|_{L_x^\infty(L_{h,v}^\infty)}\leq C\Delta x^3\left\|f\right\|_{\mathscr{C}_b\left(0,T;\mathscr{C}_b^4(\Omega)\right)},
\end{aligned}$$

$$\begin{aligned}
\left\|\left(\mathcal{T}_1^{0,1}-\widetilde{\mathcal{T}}_1^{0,1}\right)f(t)\right\|_{L_h^2(\Omega)} &\leq C\left\|\mathbb{S}_{v\Delta t/2,0}\left(\partial_v f(t)-\mathcal{L}_h\partial_v f(t)\right)\right\|_{L_h^\infty(\Omega)}\\
&\quad+C\Delta t\left\|\mathbb{S}_{v\Delta t/2,0}\left(\partial_x f(t)-\partial_x\mathcal{H}_h f(t)\right)\right\|_{L_h^\infty(\Omega)}\\
&\leq C\left\|\mathbb{S}_{v\Delta t/2,0}\left(\partial_v f(t)-\mathcal{L}_h\partial_v f(t)\right)\right\|_{L_x^\infty(L_{h,v}^\infty)}\\
&\quad+C\Delta t\left\|\mathbb{S}_{v\Delta t/2,0}\left(\partial_x f(t)-\partial_x\mathcal{H}_h f(t)\right)\right\|_{L_x^\infty(L_{h,v}^\infty)}\\
&\leq C\left\|\partial_v f(t)-\mathcal{L}_h\partial_v f(t)\right\|_{L_x^\infty(L_{h,v}^\infty)}\\
&\quad+C\Delta t\left\|\partial_x f(t)-\partial_x\mathcal{H}_h f(t)\right\|_{L_x^\infty(L_{h,v}^\infty)}\\
&\leq C(\Delta t\Delta x^3+\Delta x^4)\left\|f\right\|_{\mathscr{C}_b\left(0,T;\mathscr{C}_b^4(\Omega)\right)}\\
&\leq C\Delta x^3\left\|f\right\|_{\mathscr{C}_b\left(0,T;\mathscr{C}_b^4(\Omega)\right)},
\end{aligned}$$

$$\begin{aligned}
\left\|\left(\widehat{\mathcal{T}}_2^{0,0}-\widetilde{\mathcal{T}}_2^{0,0}\right)f(t)\right\|_{L_h^2(\Omega)} &\leq C\left\|\mathbb{S}_{0,\Delta tE_h}\left(f(t)-\mathcal{H}_h f(t)\right)\right\|_{L_h^\infty(\Omega)}\\
&\leq C\left\|\mathbb{S}_{0,\Delta tE_h}\left(f(t)-\mathcal{H}_h f(t)\right)\right\|_{L_{h,x}^\infty(L_v^\infty)}\\
&\leq C\left\|f(t)-\mathcal{H}_h f(t)\right\|_{L_{h,x}^\infty(L_v^\infty)}\leq C\Delta v^4\left\|f\right\|_{\mathscr{C}_b\left(0,T;\mathscr{C}_b^4(\Omega)\right)},
\end{aligned}$$

$$\begin{aligned}
\left\|\left(\widehat{\mathcal{T}}_2^{0,1}-\widetilde{\mathcal{T}}_2^{0,1}\right)f(t)\right\|_{L_h^2(\Omega)} &\leq C\left\|\mathbb{S}_{0,\Delta tE_h}\left(\partial_x f(t)-\partial_x\mathcal{H}_h f(t)\right)\right\|_{L_h^\infty(\Omega)}\\
&\leq C\left\|\mathbb{S}_{0,\Delta tE_h}\left(\partial_x f(t)-\partial_x\mathcal{H}_h f(t)\right)\right\|_{L_{h,x}^\infty(L_v^\infty)}\\
&\leq C\left\|\partial_x f(t)-\partial_x\mathcal{H}_h f(t)\right\|_{L_{h,x}^\infty(L_v^\infty)}\leq C\Delta v^3\left\|f\right\|_{\mathscr{C}_b\left(0,T;\mathscr{C}_b^4(\Omega)\right)},
\end{aligned}$$

$$\begin{aligned}
\left\|\left(\widehat{\mathcal{T}}_2^{1,0}-\widetilde{\mathcal{T}}_2^{1,0}\right)f(t)\right\|_{L_h^2(\Omega)} &\leq C\left\|\mathbb{S}_{0,\Delta tE_h}\left(\partial_x f(t)-\mathcal{L}_h\partial_x f(t)\right)\right\|_{L_h^\infty(\Omega)}\\
&\quad+C\Delta t\left\|\partial_x E_h^{n+1/2}\right\|_\infty\left\|\mathbb{S}_{0,\Delta tE_h}\left(\partial_v f(t)-\partial_v\mathcal{H}_h f(t)\right)\right\|_{L_h^\infty(\Omega)}\\
&\leq C\left\|\mathbb{S}_{0,\Delta tE_h}\left(\partial_x f(t)-\mathcal{L}_h\partial_x f(t)\right)\right\|_{L_{h,x}^\infty(L_v^\infty)}\\
&\quad+C\Delta t\left\|\mathbb{S}_{0,\Delta tE_h}\left(\partial_v f(t)-\partial_v\mathcal{H}_h f(t)\right)\right\|_{L_{h,x}^\infty(L_v^\infty)}\\
&\leq C\left\|\partial_x f(t)-\mathcal{L}_h\partial_x f(t)\right\|_{L_{h,x}^\infty(L_v^\infty)}\\
&\quad+C\Delta t\left\|\partial_v f(t)-\partial_v\mathcal{H}_h f(t)\right\|_{L_{h,x}^\infty(L_v^\infty)}\\
&\leq C(\Delta t\Delta v^3+\Delta v^4)\left\|f\right\|_{\mathscr{C}_b\left(0,T;\mathscr{C}_b^4(\Omega)\right)}\\
&\leq C\Delta v^3\left\|f\right\|_{\mathscr{C}_b\left(0,T;\mathscr{C}_b^4(\Omega)\right)}.\qquad\square
\end{aligned}$$

Now let us evaluate the term (I3):

$$\left\|\mathcal{T}_1^\alpha\circ\widehat{\mathcal{T}}_2^\alpha\circ\mathcal{T}_1^\alpha f(t^n)-\widetilde{\mathcal{T}}_1^\alpha\circ\widetilde{\mathcal{T}}_2^\alpha\circ\widetilde{\mathcal{T}}_1^\alpha f(t^n)\right\|_{L_h^2(\Omega)}.$$

LEMMA 4.8. *Let $f \in \mathscr{C}_b(0,T;\mathscr{C}_b^4(\Omega))$, and then*

$$\begin{aligned}
&\left\|\mathcal{T}_1^{0,0}\circ\widehat{\mathcal{T}}_2^{0,0}\circ\mathcal{T}_1^{0,0}f(t^n)-\widetilde{\mathcal{T}}_1^{0,0}\circ\widetilde{\mathcal{T}}_2^{0,0}\circ\widetilde{\mathcal{T}}_1^{0,0}f(t^n)\right\|_{L_h^2(\Omega)}\\
&\quad\le C\Bigg(\Delta x^4+\Delta v^4+\Delta t\Bigg(\Delta t^2+\Delta x^4+\Delta v^4+\Delta x^5+\Delta x\Delta v^4\\
&\quad+\left\|\left(\mathcal{T}_1^{0,0}-\widetilde{\mathcal{T}}_1^{0,0}\right)f(t^n)\right\|_{L_h^2(\Omega)}+\left\|\widetilde{\mathcal{T}}_1^{0,0}\left(f(t^n)-f_h(t^n)\right)\right\|_{L_h^2(\Omega)}\\
&\quad+\Delta x\left\|\left(\mathcal{T}_1^{1,0}-\widetilde{\mathcal{T}}_1^{1,0}\right)f(t^n)\right\|_{L_h^2(\Omega)}+\Delta x\left\|\widetilde{\mathcal{T}}_1^{1,0}\left(f(t^n)-f_h(t^n)\right)\right\|_{L_h^2(\Omega)}\Bigg)\Bigg)
\end{aligned}$$

and

$$\begin{aligned}
&\left\|\nabla\mathcal{T}_1\circ\nabla\widehat{\mathcal{T}}_2\circ\nabla\mathcal{T}_1f(t^n)-\nabla\widetilde{\mathcal{T}}_1\circ\nabla\widetilde{\mathcal{T}}_2\circ\nabla\widetilde{\mathcal{T}}_1f(t^n)\right\|_{\mathbb{L}_h^2(\Omega)}\\
&\quad\le C\Bigg(\Delta x^3+\Delta v^3+\Delta t\Bigg(\Delta t^2+\Delta x^3+\Delta x^{-1}\Delta v^4+\Delta v^4+\Delta x^4\\
&\quad+\left(1+\Delta x^{-1}\right)\left\|\left(\mathcal{T}_1^{0,0}-\widetilde{\mathcal{T}}_1^{0,0}\right)f(t^n)\right\|_{L_h^2(\Omega)}\\
&\quad+\left(1+\Delta x^{-1}\right)\left\|\widetilde{\mathcal{T}}_1^{0,0}\left(f(t^n)-f_h(t^n)\right)\right\|_{L_h^2(\Omega)}\\
&\quad+\left\|\left(\mathcal{T}_1^{1,0}-\widetilde{\mathcal{T}}_1^{1,0}\right)f(t^n)\right\|_{L_h^2(\Omega)}+\left\|\widetilde{\mathcal{T}}_1^{1,0}\left(f(t^n)-f_h(t^n)\right)\right\|_{L_h^2(\Omega)}\Bigg)\Bigg).
\end{aligned}$$

Proof. First we consider the following decomposition:

$$\begin{aligned}
\left\|\mathcal{T}_1^{\alpha}\circ\widehat{\mathcal{T}}_2^{\alpha}\circ\mathcal{T}_1^{\alpha}f(t^n)-\widetilde{\mathcal{T}}_1^{\alpha}\circ\widetilde{\mathcal{T}}_2^{\alpha}\circ\widetilde{\mathcal{T}}_1^{\alpha}f(t^n)\right\|_{L_h^2(\Omega)} &\le \left\|\left(\mathcal{T}_1^{\alpha}-\widetilde{\mathcal{T}}_1^{\alpha}\right)\circ\widehat{\mathcal{T}}_2^{\alpha}\circ\mathcal{T}_1^{\alpha}f(t^n)\right\|_{L_h^2(\Omega)} &(4.31)\\
&+\left\|\widetilde{\mathcal{T}}_1^{\alpha}\circ\left(\widehat{\mathcal{T}}_2^{\alpha}-\widetilde{\mathcal{T}}_2^{\alpha}\right)\circ\mathcal{T}_1^{\alpha}f(t^n)\right\|_{L_h^2(\Omega)} &(4.32)\\
&+\left\|\widetilde{\mathcal{T}}_1^{\alpha}\circ\widetilde{\mathcal{T}}_2^{\alpha}\circ\left(\mathcal{T}_1^{\alpha}-\widetilde{\mathcal{T}}_1^{\alpha}\right)f(t^n)\right\|_{L_h^2(\Omega)}. &(4.33)
\end{aligned}$$

Let us start with the term (4.31). The term (4.31) can be decomposed in the following manner:

$$\begin{aligned}
\left\|\left(\mathcal{T}_1^{\alpha}-\widetilde{\mathcal{T}}_1^{\alpha}\right)\circ\widehat{\mathcal{T}}_2^{\alpha}\circ\mathcal{T}_1^{\alpha}f(t^n)\right\|_{L_h^2(\Omega)} &\le \left\|\left(\mathcal{T}_1^{\alpha}-\widetilde{\mathcal{T}}_1^{\alpha}\right)\circ\mathcal{T}_2^{\alpha}\circ\mathcal{T}_1^{\alpha}f(t^n)\right\|_{L_h^2(\Omega)} &(4.34)\\
&+\left\|\mathcal{T}_1^{\alpha}\circ\left(\widehat{\mathcal{T}}_2^{\alpha}-\mathcal{T}_2^{\alpha}\right)\circ\mathcal{T}_1^{\alpha}f(t^n)\right\|_{L_h^2(\Omega)} &(4.35)\\
&+\left\|\widetilde{\mathcal{T}}_1^{\alpha}\circ\left(\mathcal{T}_2^{\alpha}-\widehat{\mathcal{T}}_2^{\alpha}\right)\circ\mathcal{T}_1^{\alpha}f(t^n)\right\|_{L_h^2(\Omega)}. &(4.36)
\end{aligned}$$

Let us proceed with (4.34):

$$\begin{aligned}\left\|\left(\mathcal{T}_1^\alpha-\widetilde{\mathcal{T}}_1^\alpha\right)\circ\mathcal{T}_2^\alpha\circ\mathcal{T}_1^\alpha f(t^n)\right\|_{L^2_h(\Omega)}&\leq C\Delta x^{4-|\alpha|}\left\|\mathcal{T}_2^{0,0}\circ\mathcal{T}_1^{0,0}f(t^n)\right\|_{\mathscr{C}^4_b(\Omega)}\\&\leq C\left(\|f\|_{\mathscr{C}_b\left(0,T;\mathscr{C}^4_b(\Omega)\right)},\|E\|_{\mathscr{C}^4_b(\Omega)}\right)\Delta x^{4-|\alpha|}\\&\leq C\Delta x^{4-|\alpha|}.\end{aligned}$$

The term (4.35) has already been estimated, and the result is given by Lemma 4.5. Let us evaluate the term (4.36). If we introduce the characteristic curves

$$\begin{aligned}\widetilde{X}^\star(t^n)&=x-v\Delta t+\frac{\Delta t^2}{2}E(t^{n+1/2},x),\\\widetilde{V}^\star(t^n)&=v-\Delta tE(t^{n+1/2},x),\\\widehat{X}^\star(t^n)&=x-v\Delta t+\frac{\Delta t^2}{2}E_h^{n+1/2}(x),\\\widehat{V}^\star(t^n)&=v-\Delta tE_h^{n+1/2}(x),\end{aligned}$$

we get

$$\begin{aligned}\left|\left(\left(\widehat{\mathcal{T}}_2^{0,0}-\mathcal{T}_2^{0,0}\right)\circ\mathcal{T}_1^{0,0}f(t^n)\right)_{k,l}\right|&\leq\left|f\left(t^n,\widehat{X}^\star_{k,l}(t^n),\widehat{V}^\star_{k,l}(t^n)\right)-f\left(t^n,\widetilde{X}^\star_{k,l}(t^n),\widetilde{V}^\star_{k,l}(t^n)\right)\right|\\&\leq C\Delta t\left\|E_h^{n+1/2}-E(t^{n+1/2})\right\|_\infty\|\nabla f\|_{L^\infty(\Omega)}.\end{aligned}$$

We deduce that

$$\left\|\left(\widehat{\mathcal{T}}_2^{0,0}-\mathcal{T}_2^{0,0}\right)\circ\mathcal{T}_1^{0,0}f(t^n)\right\|_{L^2_h(\Omega)}\leq C|\Omega|\Delta t\left\|E_h^{n+1/2}-E(t^{n+1/2})\right\|_\infty,$$

and by using (4.25) we finally get

$$\begin{aligned}\left\|\widetilde{\mathcal{T}}_1^{0,0}\circ\left(\widehat{\mathcal{T}}_2^{0,0}-\mathcal{T}_2^{0,0}\right)\circ\mathcal{T}_1^{0,0}f(t^n)\right\|_{L^2_h(\Omega)}&\leq C|\Omega|\Delta t\left\|E_h^{n+1/2}-E(t^{n+1/2})\right\|_\infty\\&\leq C\Delta t\Big(\Delta t^2+\Delta x^4+\Delta v^4\\&+\left\|\left(\mathcal{T}_1^{0,0}-\widetilde{\mathcal{T}}_1^{0,0}\right)f(t^n)\right\|_{L^2_h(\Omega)}\\&+\left\|\widetilde{\mathcal{T}}_1^{0,0}\left(f(t^n)-f_h(t^n)\right)\right\|_{L^2_h(\Omega)}\Big).\end{aligned}$$

On the other hand we have

$$\begin{aligned}&\left|\left(\left(\nabla\widehat{\mathcal{T}}_2-\nabla\mathcal{T}_2\right)\circ\nabla\mathcal{T}_1f(t^n)\right)_{k,l}\right|\\&\leq\left|\nabla f\left(\widehat{X}^\star_{k,l}(t^n),\widehat{V}^\star_{k,l}(t^n)\right)\right|\cdot\left(\left|\nabla\widehat{X}^\star_{k,l}(t^n)-\nabla\widetilde{X}^\star_{k,l}(t^n)\right|+\left|\nabla\widehat{V}^\star_{k,l}(t^n)-\nabla\widetilde{V}^\star_{k,l}(t^n)\right|\right)\\&+\left|\nabla f\left(\widehat{X}^\star_{k,l}(t^n),\widehat{V}^\star_{k,l}(t^n)\right)-\nabla f\left(\widetilde{X}^\star_{k,l}(t^n),\widetilde{V}^\star_{k,l}(t^n)\right)\right|\cdot\left|\nabla\left(\widetilde{X}^\star_{k,l}(t^n),\widetilde{V}^\star_{k,l}(t^n)\right)\right|\\&\leq\left|\nabla f\left(\widehat{X}^\star_{k,l}(t^n),\widehat{V}^\star_{k,l}(t^n)\right)\right|\cdot\left(\left|\nabla\widehat{X}^\star_{k,l}(t^n)-\nabla\widetilde{X}^\star_{k,l}(t^n)\right|+\left|\nabla\widehat{V}^\star_{k,l}(t^n)-\nabla\widetilde{V}^\star_{k,l}(t^n)\right|\right)\\&+\left(\left|\widehat{X}^\star_{k,l}(t^n)-\widetilde{X}^\star_{k,l}(t^n)\right|+\left|\widehat{V}^\star_{k,l}(t^n)-\widetilde{V}^\star_{k,l}(t^n)\right|\right)\\&\quad\cdot\left|\nabla^2f\left(\widetilde{X}^\star_{k,l}(t^n),\widetilde{V}^\star_{k,l}(t^n)\right)\right|\cdot\left|\nabla\left(\widetilde{X}^\star_{k,l}(t^n),\widetilde{V}^\star_{k,l}(t^n)\right)\right|\end{aligned}$$

$$
\begin{aligned}
&\leq C\left(\|\nabla f\|_{\mathscr{C}_b(0,T;\mathscr{C}_b(\Omega))}, \left\|\nabla^2 f\right\|_{\mathscr{C}_b(0,T;\mathscr{C}_b(\Omega))}, \|\partial_x E\|_{\mathscr{C}_b(0,T;\mathscr{C}_b(\Omega))}\right)\\
&\times\left(\left|\widehat{X}^\star_{k,l}(t^n)-\widetilde{X}^\star_{k,l}(t^n)\right|+\left|\widehat{V}^\star_{k,l}(t^n)-\widetilde{V}^\star_{k,l}(t^n)\right|\right.\\
&\left.+\left|\nabla\widehat{X}^\star_{k,l}(t^n)-\nabla\widetilde{X}^\star_{k,l}(t^n)\right|+\left|\nabla\widehat{V}^\star_{k,l}(t^n)-\nabla\widetilde{V}^\star_{k,l}(t^n)\right|\right)\\
&\leq C\Delta t\left\|E_h^{n+1/2}-E(t^{n+1/2})\right\|_{1,\infty}.
\end{aligned}
$$

From inequalities (4.26), (4.27), (4.29), and (4.30) we find that

$$
\begin{aligned}
\left\|\nabla\widetilde{\mathcal{T}}_1\left(\nabla\widehat{\mathcal{T}}_2-\nabla\mathcal{T}_2\right)\circ\nabla\mathcal{T}_1 f(t^n)\right\|_{\mathbb{L}^2_h(\Omega)} &\leq (1+C\Delta t)\left\|\left(\nabla\widehat{\mathcal{T}}_2-\nabla\mathcal{T}_2\right)\circ\nabla\mathcal{T}_1 f(t^n)\right\|_{\mathbb{L}^2_h(\Omega)}\\
&\leq C|\Omega|\Delta t\left\|E_h^{n+1/2}-E(t^{n+1/2})\right\|_{1,\infty}\\
&\leq C\Delta t\left(\Delta t^2+\Delta x^4+\Delta v^4+\left\|\left(\mathcal{T}_1^{0,0}-\widetilde{\mathcal{T}}_1^{0,0}\right)f(t^n)\right\|_{L^2_h(\Omega)}\right.\\
&+\left\|\widetilde{\mathcal{T}}_1^{0,0}\left(f(t^n)-f_h(t^n)\right)\right\|_{L^2_h(\Omega)}+\left\|\left(\mathcal{T}_1^{1,0}-\widetilde{\mathcal{T}}_1^{1,0}\right)f(t^n)\right\|_{L^2_h(\Omega)}\\
&\left.+\left\|\widetilde{\mathcal{T}}_1^{1,0}\left(f(t^n)-f_h(t^n)\right)\right\|_{L^2_h(\Omega)}\right).
\end{aligned}
$$

Let us continue with the term (4.32). By using (4.25) we get that

$$
\begin{aligned}
\left\|\widetilde{\mathcal{T}}_1^{0,0}\circ\left(\widehat{\mathcal{T}}_2^{0,0}-\widetilde{\mathcal{T}}_2^{0,0}\right)\circ\mathcal{T}_1^{0,0}f(t^n)\right\|_{L^2_h(\Omega)} &\leq \left\|\left(\widehat{\mathcal{T}}_2^{0,0}-\widetilde{\mathcal{T}}_2^{0,0}\right)\circ\mathcal{T}_1^{0,0}f(t^n)\right\|_{L^2_h(\Omega)}\\
&\leq C\Delta v^4\left\|\mathcal{T}_1^{0,0}f(t^n)\right\|_{\mathscr{C}^4_b(\Omega)}\\
&\leq C\Delta v^4\|f\|_{\mathscr{C}\left(0,T;\mathscr{C}^4_b(\Omega)\right)}.
\end{aligned}
$$

By using (4.26) and (4.27) we have

$$
\begin{aligned}
\left\|\nabla\widetilde{\mathcal{T}}_1\circ\left(\nabla\widehat{\mathcal{T}}_2-\nabla\widetilde{\mathcal{T}}_2\right)\circ\nabla\mathcal{T}_1 f(t^n)\right\|_{\mathbb{L}^2_h(\Omega)} &\leq (1+C\Delta t)\left\|\left(\nabla\widehat{\mathcal{T}}_2-\nabla\widetilde{\mathcal{T}}_2\right)\circ\nabla\mathcal{T}_1 f(t^n)\right\|_{\mathbb{L}^2_h(\Omega)}\\
&\leq C\Delta v^3\left\|\mathcal{T}_1^{0,0}f(t^n)\right\|_{\mathscr{C}^4_b(\Omega)}\\
&\leq C\Delta v^3\|f\|_{\mathscr{C}\left(0,T;\mathscr{C}^4_b(\Omega)\right)}.
\end{aligned}
$$

Let us finish with the term (4.33). By using (4.25) and (4.28) we get that

$$
\begin{aligned}
\left\|\widetilde{\mathcal{T}}_1^{0,0}\circ\widetilde{\mathcal{T}}_2^{0,0}\circ\left(\mathcal{T}_1^{0,0}-\widetilde{\mathcal{T}}_1^{0,0}\right)f(t^n)\right\|_{L^2_h(\Omega)} &\leq \left\|\widetilde{\mathcal{T}}_2^{0,0}\circ\left(\mathcal{T}_1^{0,0}-\widetilde{\mathcal{T}}_1^{0,0}\right)f(t^n)\right\|_{L^2_h(\Omega)}\\
&\leq \left\|\left(\mathcal{T}_1^{0,0}-\widetilde{\mathcal{T}}_1^{0,0}\right)f(t^n)\right\|_{L^2_h(\Omega)}\\
&\leq C\Delta x^4\left\|\mathcal{T}_1^{0,0}f(t^n)\right\|_{\mathscr{C}^4_b(\Omega)}\\
&\leq C\Delta x^4\|f\|_{\mathscr{C}\left(0,T;\mathscr{C}^4_b(\Omega)\right)}.
\end{aligned}
$$

By using (4.26), (4.27), (4.29), and (4.30) we get that

$$\begin{aligned}\left\|\nabla\widetilde{\mathcal{T}}_1\circ\nabla\widetilde{\mathcal{T}}_2\circ\left(\nabla\mathcal{T}_1-\nabla\widetilde{\mathcal{T}}_1\right)f(t^n)\right\|_{\mathbb{L}^2_h(\Omega)}&\leq(1+C\Delta t)\left\|\nabla\widetilde{\mathcal{T}}_2\circ\left(\nabla\mathcal{T}_1-\nabla\widetilde{\mathcal{T}}_1\right)f(t^n)\right\|_{\mathbb{L}^2_h(\Omega)}\\&\leq(1+C\Delta t)^2\left\|\left(\nabla\mathcal{T}_1-\nabla\widetilde{\mathcal{T}}_1\right)f(t^n)\right\|_{\mathbb{L}^2_h(\Omega)}\\&\leq C\Delta x^3\left\|f(t^n)\right\|_{\mathscr{C}^4_b(\Omega)}\\&\leq C\Delta x^3\left\|f\right\|_{\mathscr{C}\left(0,T;\mathscr{C}^4_b(\Omega)\right)}.\end{aligned}$$

□

Now let us evaluate the term (I4).

LEMMA 4.9. *Let $f\in\mathscr{C}_b(0,T;\mathscr{C}^1_b(\Omega))$, and then there exists a constant C independent of Δx, Δv, and Δt such that*

$$\left\|\widetilde{\mathcal{T}}_1^{0,0}\circ\widetilde{\mathcal{T}}_2^{0,0}\circ\widetilde{\mathcal{T}}_1^{0,0}(f(t^n)-f_h(t^n))\right\|_{L^2_h(\Omega)}\leq\left\|f(t^n)-f_h(t^n)\right\|_{L^2_h(\Omega)}$$

and

$$\left\|\nabla\widetilde{\mathcal{T}}_1\circ\nabla\widetilde{\mathcal{T}}_2\circ\nabla\widetilde{\mathcal{T}}_1(f(t^n)-f_h(t^n))\right\|_{\mathbb{L}^2_h(\Omega)}\leq(1+C\Delta t)\left\|\nabla f(t^n)-\nabla f_h(t^n)\right\|_{\mathbb{L}^2_h(\Omega)}.$$

Proof. By using (4.25) and (4.28) we get that

$$\begin{aligned}\left\|\widetilde{\mathcal{T}}_1^{0,0}\circ\widetilde{\mathcal{T}}_2^{0,0}\circ\widetilde{\mathcal{T}}_1^{0,0}(f(t^n)-f_h(t^n))\right\|_{L^2_h(\Omega)}&\leq\left\|\widetilde{\mathcal{T}}_2^{0,0}\circ\widetilde{\mathcal{T}}_1^{0,0}(f(t^n)-f_h(t^n))\right\|_{L^2_h(\Omega)}\\&\leq\left\|\widetilde{\mathcal{T}}_1^{0,0}(f(t^n)-f_h(t^n))\right\|_{L^2_h(\Omega)}\\&\leq\left\|f(t^n)-f_h(t^n)\right\|_{L^2_h(\Omega)}.\end{aligned}$$

By using (4.26), (4.27), (4.29), and (4.30) it follows that

$$\begin{aligned}&\left\|\nabla\widetilde{\mathcal{T}}_1\circ\nabla\widetilde{\mathcal{T}}_2\circ\nabla\widetilde{\mathcal{T}}_1(f(t^n)-f_h(t^n))\right\|_{\mathbb{L}^2_h(\Omega)}\\&\qquad\leq(1+C\Delta t)\left\|\nabla\widetilde{\mathcal{T}}_2\circ\nabla\widetilde{\mathcal{T}}_1(f(t^n)-f_h(t^n))\right\|_{\mathbb{L}^2_h(\Omega)}\\&\qquad\leq(1+C\Delta t)^2\left\|\nabla\widetilde{\mathcal{T}}_1(f(t^n)-f_h(t^n))\right\|_{\mathbb{L}^2_h(\Omega)}\\&\qquad\leq(1+C\Delta t)^3\left\|\nabla f(t^n)-\nabla f_h(t^n)\right\|_{\mathbb{L}^2_h(\Omega)}\\&\qquad\leq(1+C\Delta t)\left\|\nabla f(t^n)-\nabla f_h(t^n)\right\|_{\mathbb{L}^2_h(\Omega)}.\end{aligned}$$

□

Proof of Theorem 4.1. By putting together Lemmas 4.3, 4.5, 4.7, 4.8, and 4.9 we obtain

$$\begin{pmatrix}\left\|e_{n+1}\right\|_{L^2_h(\Omega)}\\\left\|\nabla e_{n+1}\right\|_{\mathbb{L}^2_h(\Omega)}\end{pmatrix}=\begin{pmatrix}1+C\Delta t&\Delta t\Delta x\\\Delta t\left(1+\Delta x^{-1}\right)&1+C\Delta t\end{pmatrix}\begin{pmatrix}\left\|e_n\right\|_{L^2_h(\Omega)}\\\left\|\nabla e_n\right\|_{\mathbb{L}^2_h(\Omega)}\end{pmatrix}+\begin{pmatrix}\gamma\\\partial\gamma\end{pmatrix},$$

where

$$\gamma=C\left(\Delta t^3+(1+\Delta t+\Delta x)(\Delta x^4+\Delta v^4)+\Delta x^4+\Delta x^4\right)$$

and

$$\partial\gamma=C\left(\Delta t^3+\Delta x^{-1}(1+\Delta t+\Delta x)(\Delta x^4+\Delta v^4)+\Delta x^3+\Delta x^3\right).$$

The eigenvalues of the matrix

$$\begin{pmatrix} 1+C\Delta t & \Delta t\Delta x \\ \Delta t\left(1+\Delta x^{-1}\right) & 1+C\Delta t \end{pmatrix}$$

are equivalent to $1+\mathcal{O}(\Delta t)$. The formula of recurrence gives

$$\begin{pmatrix} \|e_{n+1}\|_{L_h^2(\Omega)} \\ \|\nabla e_{n+1}\|_{\mathbb{L}_h^2(\Omega)} \end{pmatrix} = \frac{1}{\Delta t}\begin{pmatrix} \gamma \\ \partial\gamma \end{pmatrix} + e^{CT}\begin{pmatrix} \|e_0\|_{L_h^2(\Omega)} \\ \|\nabla e_0\|_{\mathbb{L}_h^2(\Omega)} \end{pmatrix},$$

which is the error estimate of $\partial^\alpha f - \partial^\alpha f_h$ stated in Theorem 4.1, assuming $\|e_0\|_{L_h^2(\Omega)} = \|\nabla e_0\|_{\mathbb{L}_h^2(\Omega)} = 0$. By using the estimates (4.23) and (4.24) and thanks to the error estimate of $\partial^\alpha f - \partial^\alpha f_h$, we find the estimate $\partial^\alpha E - \partial^\alpha E_h$ stated in Theorem 4.1. ∎

REFERENCES

[1] M. L. Begue, A. Ghizzo, and P. Bertrand, *Two-dimensional Vlasov simulation of Raman scattering and plasma beatwave acceleration on parallel computers*, J. Comput. Phys., 151 (1999), pp. 458–478.

[2] M. L. Begue, A. Ghizzo, P. Bertrand, E. Sonnendrücker, and O. Coulaud, *Two-dimensional semi-Lagrangian Vlasov simulations of laser-plasma interaction in the relativistic regime*, J. Comput. Phys., 62 (1999), pp. 367–388.

[3] R. Bermejo, *Analysis of an algorithm for the Galerkin-characteristic method*, Numer. Math., 60 (1991), pp. 163–194.

[4] R. Bermejo, *Analysis of a class of quasi-monotone and conservative semi-Lagrangian advection schemes*, Numer. Math., 87 (2001), pp. 597–623.

[5] N. Besse and E. Sonnendrücker, *Semi-Lagrangian schemes for the Vlasov equation on an unstructured mesh of phase space*, J. Comput. Phys., 191 (2003), pp. 341–376.

[6] N. Besse and M. Mehrenberger, *Convergence of classes of high-order semi-Lagrangian schemes for the Vlasov-Poisson system*, Math. Comp., 77 (2008), pp. 93–123.

[7] N. Besse, *Convergence of a semi-Lagrangian scheme for the one-dimensional Vlasov–Poisson system*, SIAM J. Numer. Anal., 42 (2004), pp. 350–382.

[8] N. Besse, *Etude mathématique et numérique de l'equation de Vlasov sur des maillages non structurés de l'espace des phases*, thèse de l'Université Louis Pasteur, Strasbourg, France, 2003.

[9] N. Besse, J. Segré, and E. Sonnendrücker, *Semi-Lagrangian schemes for the two-dimensional Vlasov-Poisson system on unstructured meshes*, Transport Theory Statist. Phys., 34 (2005), pp. 311–332.

[10] C. K. Birdsall and A. B. Langdon, *Plasma Physics via Computer Simulation*, McGraw-Hill, New York, 1985.

[11] F. Bouchut, F. Golse, and M. Pulvirenti, *Kinetic Equations and Asymptotic Theory*, Ser. Appl. Math., P. G. Ciarlet and P.-L. Lions, eds., Gauthier-Villars, Paris, 2000.

[12] F. Bouchut, *Renormalized solutions to the Vlasov equation with coefficients of bounded variation*, Arch. Ration. Mech. Anal., 157 (2001), pp. 75–90.

[13] C. Z. Cheng and G. Knorr, *The integration of the Vlasov equation in configuration space*, J. Comput. Phys., 22 (1976), pp. 330–351.

[14] G.-H. Cottet and P.-A. Raviart, *Particle methods for the one-dimensional Vlasov–Poisson equations*, SIAM J. Numer. Anal., 21 (1984), pp. 52–76.

[15] O. Coulaud, E. Sonnendrücker, E. Dillon, P. Bertrand, and A. Ghizzo, *Parallelization of semi-Lagrangian Vlasov codes*, J. Comput. Phys., 61 (1999), pp. 435–448.

[16] M. Falcone and R. Ferretti, *Convergence analysis for a class of high-order semi-Lagrangian advection schemes*, SIAM J. Numer. Anal., 35 (1998), pp. 909–940.

[17] F. Filbet, *Convergence of a finite volume scheme for the Vlasov–Poisson system*, SIAM J. Numer. Anal., 39 (2001), pp. 1146–1169.

[18] K. Ganguly and H. D. Victory, Jr., *On the convergence of particle methods for multidimensional Vlasov–Poisson systems*, SIAM J. Numer. Anal., 26 (1989), pp. 249–288.

[19] R. T. Glassey, *The Cauchy Problem in Kinetic Theory*, SIAM, Philadelphia, 1996.

[20] R. Glassey and J. Schaeffer, *Convergence of a particle method for the relativistic Vlasov–Maxwell system*, SIAM J. Numer. Anal., 28 (1991), pp. 1–25.

[21] K. J. Havlak and H. D. Victory, Jr., *The numerical analysis of random particle methods applied to Vlasov–Poisson–Fokker–Planck kinetic equations*, SIAM J. Numer. Anal., 33 (1996), pp. 291–317.

[22] K. J. Havlak and H. D. Victory, Jr., *On deterministic particle methods for solving Vlasov–Poisson–Fokker–Planck systems*, SIAM J. Numer. Anal., 35 (1998), pp. 1473–1519.

[23] H. Neunzert and J. Wick, *Die Theorie der asymptotischen Verteiling und die numerische Lösung von Integrodifferentialgleichungen*, Numer. Math., 21 (1974), pp. 234–243.

[24] J. Schaeffer, *Convergence of a difference scheme for the Vlasov–Poisson–Fokker–Planck system in one dimension*, SIAM J. Numer. Anal., 35 (1998), pp. 1149–1175.

[25] E. Sonnendrücker, J. Roche, P. Bertrand, and A. Ghizzo, *The semi-Lagrangian method for the numerical resolution of Vlasov equations*, J. Comput. Phys., 149 (1996), pp. 841–872.

[26] A. Staniforth and J. Cote, *Semi-Lagrangian integration schemes for atmospheric models—a review*, Monthly Weather Review, 119 (1991), pp. 2206–2223.

[27] H. D. Victory, Jr., and E. J. Allen, *The convergence theory of particle-in-cell methods for multidimensional Vlasov–Poisson systems*, SIAM J. Numer. Anal., 28 (1991), pp. 1207–1241.

[28] H. D. Victory, Jr., G. Tucker, and K. Ganguly, *The convergence analysis of fully discretized particle methods for solving Vlasov–Poisson systems*, SIAM J. Numer. Anal., 28 (1991), pp. 955–989.

[29] S. Wollman, *On the approximation of the Vlasov–Poisson system by particle methods*, SIAM J. Numer. Anal., 37 (2000), pp. 1369–1398.

SIAM J. NUMER. ANAL.
Vol. 46, No. 2, pp. 671–690
© 2008 Society for Industrial and Applied Mathematics

ON SINC-BASED METHOD IN COMPUTING EIGENVALUES OF BOUNDARY-VALUE PROBLEMS*

M. H. ANNABY† AND R. M. ASHARABI‡

Abstract. We aim in this paper to give a complete treatment of the error analysis of the use of the sinc method in computing eigenvalues of second order Birkhoff-regular eigenvalue problems. For this task we study truncation and amplitude errors on $\mathbb{C}$ associated with cardinal series representation of band-limited functions as well as those of their derivatives. These results, which seem to be considered for the first time, give us the possibility to establish the error analysis when the eigenvalues are not real or not algebraically simple.

Key words. sinc methods, Birkhoff-regular problems

AMS subject classifications. 34L16, 65L15, 94A20

DOI. 10.1137/060664653

1. Introduction.

Sinc methods, which are based on the classical sampling theorem of Whittaker, Kotelnikov, and Shannon (WKS), enable solutions to many computational problems. Illustrations are given in [19, 20, 24, 25]. The WKS sampling theorem states that if $f(\mu)$ is entire in μ of exponential type σ, $\sigma > 0$, which belongs to $L^2(\mathbb{R})$ when restricted to $\mathbb{R}$, then $f(\mu)$ can be reconstructed via the cardinal series representation

$$f(\mu) = \sum_{n=-\infty}^{\infty} f\left(\frac{n\pi}{\sigma}\right) \operatorname{sinc}(\sigma\mu - n\pi), \quad \mu \in \mathbb{C}, \tag{1.1}$$

where $\operatorname{sinc} z$ is the function

$$\operatorname{sinc} z = \begin{cases} \dfrac{\sin z}{z}, & z \neq 0, \\ 1, & z = 0, \end{cases} \tag{1.2}$$

and the points $\{n\pi/\sigma\}_{n\in\mathbb{Z}}$ are called the sampling nodes (points). The series (1.1) is absolutely convergent for all $\mu \in \mathbb{C}$, and, moreover, it is uniformly convergent for $\mu \in \mathbb{R}$ and also for all μ on compact subsets of $\mathbb{C}$. The space of all functions f that satisfy the above conditions is the Paley–Wiener space B_σ^2 of band-limited functions with bandwidth σ, i.e., $L^2(\mathbb{R})$-functions whose Fourier transforms vanish on $\{x \in \mathbb{R} : \ |x| > \sigma\}$. Methods based on using the expansion (1.1) are called sinc methods. This paper is devoted to an application of sinc methods, namely, that of computing eigenvalues of second order boundary-value problems as is established in [4, 5, 6, 10, 11]. In such applications, two types of errors appear, namely, truncation and amplitude errors. By the truncation error of $f \in B_\sigma^2$ we mean

$$R_N(\mu) := f(\mu) - f_N(\mu), \quad N \in \mathbb{Z}^+, \quad \mu \in \mathbb{C}, \tag{1.3}$$

*Received by the editors July 10, 2006; accepted for publication (in revised form) October 3, 2007; published electronically February 1, 2008.
http://www.siam.org/journals/sinum/46-2/66465.html

†On leave from Department of Mathematics, Faculty of Science, Cairo University, Giza, Egypt (mhannaby@yahoo.com). Current address: Department of Mathematics and Physics, Qatar University, P.O. Box 2713, Doha, Qatar.

‡Department of Mathematics, Faculty of Science, Sana'a University, Yemen (rashad1974@hotmail.com).

where $f_N(\mu)$ is the truncated series

$$f_N(\mu) := \sum_{n=-N}^{N} f\left(\frac{n\pi}{\sigma}\right) \operatorname{sinc}(\sigma\mu - n\pi), \quad \mu \in \mathbb{C}.$$

The amplitude error arises when the exact values $f\left(\frac{n\pi}{\sigma}\right)$, $n \in \mathbb{Z}$, of (1.1) are replaced by approximate ones $\widetilde{f}\left(\frac{n\pi}{\sigma}\right)$, $n \in \mathbb{Z}$. So if the differences $\varepsilon_n := f\left(\frac{n\pi}{\sigma}\right) - \widetilde{f}\left(\frac{n\pi}{\sigma}\right)$ are bounded by a positive number ε, i.e., $|\varepsilon_n| \le \varepsilon$, $\varepsilon > 0$, the amplitude error is defined to be

$$(A_\varepsilon)f(\mu) = \sum_{n=-\infty}^{\infty} \left\{ f\left(\frac{n\pi}{\sigma}\right) - \widetilde{f}\left(\frac{n\pi}{\sigma}\right) \right\} \operatorname{sinc}(\sigma\mu - n\pi), \quad \mu \in \mathbb{C}. \tag{1.4}$$

In [18] it is proved that if $\mu \in \mathbb{R}$, $f(\mu) \in B_\sigma^2$, and $\mu f(\mu) \in L^2(\mathbb{R})$, then for $N \in \mathbb{Z}^+$, $|\mu| < N\pi/\sigma$, one gets

$$|R_N(\mu)| \le \frac{2E}{\pi^2\sqrt{3}} \, \frac{\sigma\sqrt{\sigma}|\sin\sigma\mu|}{N+1} \left(\frac{1}{\sqrt{N\pi - \sigma\mu}} + \frac{1}{\sqrt{N\pi + \sigma\mu}} \right), \tag{1.5}$$

where

$$E := \sqrt{\int_{-\infty}^{\infty} |\mu f(\mu)|^2 dt}. \tag{1.6}$$

Butzer, Splettstösser, and Stens [8] have studied the amplitude error and proved that if $f(\mu) \in B_\sigma^2$ and there is $M_f > 0$ such that

$$|\varepsilon_n| \le |f(n\pi/\sigma)|, \quad n \in \mathbb{Z}, \tag{1.7}$$

$$|f(\mu)| \le M_f \, |\mu|^{-1} \; (|\mu| \ge 1), \quad \mu \in \mathbb{R}, \tag{1.8}$$

hold. We should mention here that condition (1.7) is concluded from [8, p. 20]. Then for $\sigma \ge \pi$, $0 < \varepsilon \le \min\{\pi/\sigma, 1/\sqrt{e}\}$, we have

$$|(A_\varepsilon)f(\mu)| \le \mathcal{A}(\varepsilon) := 4\left(\sqrt{3}e + \sqrt{2}M_f \exp(1/4)\right) \varepsilon \log(1/\varepsilon), \quad \mu \in \mathbb{R}. \tag{1.9}$$

In [4, 5, 6, 7, 10, 11, 12] eigenvalues of particular cases of the second order eigenvalue problem

$$\ell(y) := -y''(x,\mu) + q(x)y(x,\mu) = \mu^2 y(x,\mu), \qquad 0 \le x \le b, \tag{1.10}$$

$$U_1(y) := \alpha_{11}y(0) + \alpha_{12}y'(0) + \beta_{11}y(b) + \beta_{12}y'(b) = 0, \tag{1.11}$$

$$U_2(y) := \alpha_{21}y(0) + \alpha_{22}y'(0) + \beta_{21}y(b) + \beta_{22}y'(b) = 0 \tag{1.12}$$

are computed approximately by the sinc method. Here $q(\cdot) \in L^1(0,b)$, $\mu \in \mathbb{C}$, and $(\alpha_{i1}, \alpha_{i2}, \beta_{i1}, \beta_{i2}) \in \mathbb{C}^4$, $i = 1, 2$, are linearly independent. Some additional details:

1. The self-adjoint eigenvalue problem with a separate type of condition, i.e., when $q(\cdot)$ is an $L^1(0,b)$-real-valued function and

   $$\beta_{11}, \beta_{12}, \alpha_{21}, \alpha_{22} \in \mathbb{R}, \quad \beta_{11} = \beta_{12} = \alpha_{21} = \alpha_{22} = 0, \tag{1.13}$$

   is considered in [4, 10]. The error analysis of the authors of [4, 10] considers only the truncation error, although they should also have considered the amplitude error.

2. The case when $q(\cdot)$ is a real-valued, $L^1(0,b)$ function and

$$\alpha_{11}=\alpha_{22}=1, \beta_{11}=\beta_{22}=-1, \quad \alpha_{ij}=\beta_{ij}=0,\ 1\le i\ne j\ge 2, \tag{1.14}$$

is considered in [6]. The error analysis when eigenvalues are double is not rigorously studied, although almost all eigenvalues of this problem are double; cf., e.g., [14, 17, 21]. When the eigenvalues are simple, only the truncation error is considered, while the amplitude error also arises.

3. The case of a non-self-adjoint eigenvalue problem with a separate type of condition is considered in [5], i.e., when $q \in L^1(0,b)$ is a complex-valued function and

$$\beta_{11}=\beta_{12}=\alpha_{21}=\alpha_{22}=0, \quad \beta_{11},\beta_{12},\alpha_{21},\alpha_{22}\in\mathbb{C}, \tag{1.15}$$

although no error analysis is given in [5].

The aim of this paper is twofold. The first is to derive the method of using the sinc technique in computing eigenvalues of a general Birkhoff-regular second order eigenvalue problem; cf. [2, 21, 22] for definitions. Our second aim is to derive a complete error analysis for the general method, which will also cover the previous omissions. The definition of Birkhoff-regularity is technical [21, section 4.8]. Fortunately Naimark has categorized all regular second order problems as follows.

Case I. $\alpha_{12}\beta_{22}-\alpha_{22}\beta_{12}\ne 0$.

Case II. $\alpha_{12}\beta_{22}-\alpha_{22}\beta_{12}=0, \quad |\alpha_{12}|+|\beta_{12}|>0, \beta_{12}\alpha_{21}+\beta_{21}\alpha_{12}\ne 0$.

Case III. $\alpha_{12}=\beta_{22}=\alpha_{22}=\beta_{12}=0,\ \alpha_{11}\beta_{21}-\beta_{11}\alpha_{21}\ne 0$.

Obviously conditions (1.11)–(1.12) are regular if any of (1.13), (1.14), and (1.15) is fulfilled. Therefore the application of the sinc method to compute the eigenvalues of general regular second order problems will contain all of these cases as special cases. On the other hand, the choice of a regular boundary condition guarantees the existence of a countable set of eigenvalues without finite limit points; cf. [21, p. 64]. The eigenvalues are not necessarily real, and each eigenvalue has algebraic multiplicity which is less than or equals 2. Also the asymptotic behavior of these eigenvalues is given in [21, section 4.8]. Since no error analysis can be established in the general case without error estimates for truncation and amplitude errors on $\mathbb{C}$ for (1.1) and its derivative, we will study truncation and amplitude errors on $\mathbb{C}$ for (1.1) and its derivative. This is the second aim of this paper, and it will be done in the next section. Section 3 is devoted to the numerical computation of the eigenvalues of (1.10)–(1.12) and the associated error analysis. The last section is devoted to giving some new examples exhibiting the technique. In these new examples we study various situations and indicate the effect of the amplitude error, which has been neglected in previous studies. The following complex mean value theorem [15] will be used in what follows. Assume that $f(\mu)$ is holomorphic on an open convex set $\Omega\in\mathbb{C}$. Let a and b be distinct points in Ω and $\Gamma_{a,b}$ be the line joining a and b in Ω. Then there exist $\xi_1,\xi_2\in\Gamma_{a,b}$ such that

$$\Re\left(\frac{f(b)-f(a)}{b-a}\right)=\Re(f'(\xi_1)), \qquad \Im\left(\frac{f(b)-f(a)}{b-a}\right)=\Im(f'(\xi_2)). \tag{1.16}$$

2. Error analysis on $\mathbb{C}$. In this section we establish the error analysis associated with (1.1) on $\mathbb{C}$ which will be needed in what follows. We consider both truncation and amplitude errors. We derive three results. Our first result extends (1.5) to the case of the complex plane. For our second result, we extend a previous

derivation of the derivative sampling theorem to the complex plane $\mathbb{C}$. The last result gave an estimate of the amplitude error of (1.1) and the derivative sampling series on $\mathbb{C}$. Now let $N \in \mathbb{Z}^+$ and Ω_N be the disk

$$\Omega_N := \left\{ \mu \in \mathbb{C} : |\mu| < \frac{N\pi}{\sigma}, \quad N \in \mathbb{Z}^+ \right\}. \tag{2.1}$$

THEOREM 2.1. *Let $N \in \mathbb{Z}^+$, $\mu \in \mathbb{C}$, and $f(\mu) \in B_\sigma^2$. Assume that $\mu f(\mu) \in L^2(\mathbb{R})$ when restricted to $\mathbb{R}$. Then we have*

$$|R_N(\mu)| \le \frac{2E}{\pi^2\sqrt{3}} \frac{\sigma\sqrt{\sigma}|\sin\sigma\mu|}{N+1} \left(\frac{1}{\sqrt{N\pi - \sigma\Re\,\mu}} + \frac{1}{\sqrt{N\pi + \sigma\Re\,\mu}} \right), \quad \mu \in \Omega_N, \tag{2.2}$$

where E is given in (1.6) *above.*

Proof. By applying triangle and Cauchy–Schwarz inequalities to $R_N(\mu)$ (cf. (1.3)), we obtain

$$\begin{aligned} |R_N(\mu)| \le{}& \left(\sum_{n>N} \left| f\left(\frac{n\pi}{\sigma}\right) \right|^2 \right)^{1/2} \left(\sum_{n>N} |\mathrm{sinc}\,(\sigma\mu - n\pi)|^2 \right)^{1/2} \\ &+ \left(\sum_{n<-N} \left| f\left(\frac{n\pi}{\sigma}\right) \right|^2 \right)^{1/2} \left(\sum_{n<-N} |\mathrm{sinc}\,(\sigma\mu - n\pi)|^2 \right)^{1/2}. \end{aligned} \tag{2.3}$$

Since $f(\mu) \in B_\sigma^2$ and $\mu f(\mu) \in L^2(\mathbb{R})$ when restricted to $\mathbb{R}$, then [18, p. 716]

$$\left(\sum_{n>N} \left| f\left(\frac{n\pi}{\sigma}\right) \right|^2 \right)^{1/2} \le \frac{2\sigma^{3/2}E}{\pi^{3/2}\sqrt{3}(N+1)}, \quad \left(\sum_{n<-N} \left| f\left(\frac{n\pi}{\sigma}\right) \right|^2 \right)^{1/2} \le \frac{2\sigma^{3/2}E}{\pi^{3/2}\sqrt{3}(N+1)}. \tag{2.4}$$

Let $\mu := x + iy \in \Omega_N$, $x, y \in \mathbb{R}$. Then

$$\begin{aligned} \sum_{n>N} |\mathrm{sinc}\,(\sigma\mu - n\pi)|^2 &= \sum_{n>N} \frac{|\sin\sigma\mu|^2}{|\sigma\mu - n\pi|^2} \le |\sin\sigma\mu|^2 \sum_{n>N} \frac{1}{(\sigma x - n\pi)^2} \\ &< |\sin\sigma\mu|^2 \int_N^\infty \frac{du}{(\sigma x - u\pi)^2} \le \frac{|\sin\sigma\mu|^2}{\pi(N\pi - \sigma x)}. \end{aligned} \tag{2.5}$$

Similarly

$$\sum_{n<-N} |\mathrm{sinc}\,(\sigma\mu - n\pi)|^2 < \frac{|\sin\sigma\mu|^2}{\pi(N\pi + \sigma x)}, \quad \mu \in \Omega_N. \tag{2.6}$$

By combining (2.3)–(2.6) we obtain (2.2). □

Remark 2.2. Letting $\mu \in \mathbb{R}$ in (2.2) leads to Jagerman estimation (1.5).

In [5] sinc methods are applied to compute eigenvalues for non-self-adjoint eigenvalue problems. However, no error analysis is given. The previous result gives us a chance to compute the error analysis when the eigenvalues are complex. If the eigenvalues are not algebraically simple, estimates (1.5) and (2.2) are not sufficient to compute the error bounds. We need to derive a truncation error estimate of the derivative of sampling series

$$f'(\mu) := \sum_{n=-\infty}^{\infty} f\left(\frac{n\pi}{\sigma}\right) \mathrm{sinc}\,'(\sigma\mu - n\pi), \quad \mu \in \mathbb{C}. \tag{2.7}$$

Series (2.7) converges absolutely on $\mathbb{C}$ and uniformly on $\mathbb{R}$ and on compact subsets of $\mathbb{C}$ [16, p. 52]. Let $N \in \mathbb{Z}^+$ and $R_{1,N}(\mu)$ be

$$R_{1,N}(\mu) := f'(\mu) - f'_N(\mu), \quad f'_N(\mu) := \sum_{n=-N}^{N} f\left(\frac{n\pi}{\sigma}\right) \operatorname{sinc}'(\sigma\mu - n\pi), \quad \mu \in \mathbb{C}. \tag{2.8}$$

THEOREM 2.3. *Let $N \in \mathbb{Z}^+$, $\mu \in \mathbb{C}$, and $f(\mu) \in B^2_\sigma$. Assume that $\mu f(\mu) \in L^2(\mathbb{R})$ when restricted to $\mathbb{R}$. Then we have*

$$\begin{aligned} |R_{1,N}(\mu)| \le{}& \frac{2E}{\pi^2\sqrt{3}} \frac{\sigma^{5/2}|\cos\sigma\mu|}{N+1} \left(\frac{1}{\sqrt{N\pi - \sigma\Re\mu}} + \frac{1}{\sqrt{N\pi + \sigma\Re\mu}}\right) \\ &+ \frac{2E}{3\pi^2} \frac{\sigma^{5/2}\,|\sin\sigma\mu|}{N+1} \left(\frac{1}{(N\pi - \sigma\Re\mu)^{3/2}} + \frac{1}{(N\pi + \sigma\Re\mu)^{3/2}}\right), \quad \mu \in \Omega_N. \end{aligned} \tag{2.9}$$

Proof. From triangle and Cauchy–Schwarz inequalities we obtain, for $\mu \in \Omega_N$,

$$\begin{aligned} |R_{1,N}(\mu)| \le{}& \sigma \left(\sum_{n>N} \left|f\left(\frac{n\pi}{\sigma}\right)\right|^2\right)^{1/2} \left\{\left(\sum_{n>N} \frac{|\cos\sigma\mu|^2}{|\sigma\mu - n\pi|^2}\right)^{1/2} + \left(\sum_{n>N} \frac{|\sin\sigma\mu|^2}{|\sigma\mu - n\pi|^4}\right)^{1/2}\right\} \\ &+ \sigma \left(\sum_{n<-N} \left|f\left(\frac{n\pi}{\sigma}\right)\right|^2\right)^{1/2} \left\{\left(\sum_{n<-N} \frac{|\cos\sigma\mu|^2}{|\sigma\mu - n\pi|^2}\right)^{1/2} \right. \\ &\left. + \left(\sum_{n<-N} \frac{|\sin\sigma\mu|^2}{|\sigma\mu - n\pi|^4}\right)^{1/2}\right\}. \end{aligned} \tag{2.10}$$

Let $\Omega_N \ni \mu := x + iy,\ x, y \in \mathbb{R}$. Then

$$\sum_{n>N} \frac{1}{|\sigma\mu - n\pi|^2} \le \sum_{n>N} \frac{1}{(\sigma x - n\pi)^2} < \int_N^\infty \frac{du}{(\sigma x - u\pi)^2} = \frac{1}{\pi} \frac{1}{N\pi - \sigma x}, \tag{2.11}$$

$$\sum_{n>N} \frac{1}{|\sigma\mu - n\pi|^4} \le \sum_{n>N} \frac{1}{(\sigma x - n\pi)^4} < \int_N^\infty \frac{du}{(\sigma x - u\pi)^4} = \frac{1}{3\pi} \frac{1}{(N\pi - \sigma x)^3}. \tag{2.12}$$

Similarly,

$$\sum_{n<-N} \frac{1}{|\sigma\mu - n\pi|^2} < \frac{1}{\pi} \frac{1}{N\pi + \sigma x}, \quad \sum_{n<-N} \frac{1}{|\sigma\mu - n\pi|^4} < \frac{1}{3\pi} \frac{1}{(N\pi + \sigma x)^3}. \tag{2.13}$$

Substituting from (2.11)–(2.13) and (2.4) in (2.10) completes the proof. ∎

Now we derive estimates for the amplitude error of (1.1) and (2.7) on $\mathbb{C}$. We start with the following two lemmas.

LEMMA 2.4. *Given $p, q > 1$ such that $\frac{1}{p} + \frac{1}{q} = 1$. We have for $r \in \mathbb{N}$*

$$\sum_{n=-\infty}^{\infty} \left(\frac{1}{\pi|n - 1/2|}\right)^{q(r+1)} < p. \tag{2.14}$$

Proof. Obviously

$$\sum_{n=-\infty}^{\infty} \left(\frac{1}{\pi|n - 1/2|}\right)^{q(r+1)} = 2\left(\frac{2}{\pi}\right)^{q(r+1)} + \frac{2}{\pi^{q(r+1)}} \sum_{n=1}^{\infty} \frac{1}{(n + 1/2)^{q(r+1)}}. \tag{2.15}$$

By using the inequalities

$$(2/\pi)^{q(r+1)} < 1, \quad (\pi(n-1/2))^{q(r+1)} \geq (\pi(n-1/2))^q,$$
$$(\pi(n+1/2))^{q(r+1)} \geq (\pi(n+1/2))^q,$$

and Higgins estimate (cf. [16, pp. 114–115])

$$1 + \frac{1}{\pi^q} \sum_{n=1}^{\infty} \left\{ \frac{1}{(n-1/2)^q} + \frac{1}{(n+1/2)^q} \right\} < p \tag{2.16}$$

in (2.15), we obtain (2.14). □

LEMMA 2.5. *Let p, $q > 1$ such that $\frac{1}{p} + \frac{1}{q} = 1$. For $r = 0, 1$, we have*

$$\left(\sum_{n=-\infty}^{\infty} \left| \operatorname{sinc}^{(r)}(\sigma\mu - n\pi) \right|^q \right)^{1/q} < (r+1)\, p\, \sigma^r \exp(\sigma|\Im\mu|), \quad \mu \in \mathbb{C}. \tag{2.17}$$

Proof. The sum in (2.17) is periodic with period π/σ. Indeed,

$$\sum_{n=-\infty}^{\infty} \left| \operatorname{sinc}^{(r)} \left(\sigma \left(\mu + \frac{\pi}{\sigma} \right) - n\pi \right) \right|^q = \sum_{n=-\infty}^{\infty} \left| \operatorname{sinc}^{(r)} (\sigma\mu - (n-1)\pi) \right|^q,$$

$r = 1, 2, \mu \in \mathbb{C}$. By replacing $n-1$ by n, we proved the π/σ-periodicity of the sum of (2.17). Define the strip $\mathcal{S}_\sigma$ to be $\mathcal{S}_\sigma := \{\mu \in \mathbb{C} : |\Re\mu| \leq \pi/2\sigma\}$. From the periodicity of the sum in (2.17), it is sufficient to prove the lemma only on the strip $\mathcal{S}_\sigma$. We distinguish between two cases. Let $r = 0$, $\mathcal{S}_\sigma \ni \mu := x + iy$, $x, y \in \mathbb{R}$.

$$|\operatorname{sinc}(\sigma\mu - n\pi)| \leq \frac{|\sin \sigma\mu|}{|\sigma\mu - n\pi|} \leq \frac{\exp(\sigma|y|)}{|\sigma x - n\pi|}. \tag{2.18}$$

Noting that

$$\operatorname{sinc} \sigma\mu := \frac{1}{2\sigma} \int_{-\sigma}^{\sigma} e^{i\mu\omega} d\omega,$$

then

$$|\operatorname{sinc} \sigma\mu| \leq \exp(\sigma|y|). \tag{2.19}$$

By using (2.18), (2.19), and the fact that $|\sin \sigma\mu| \leq \exp(\sigma|y|)$, we obtain

$$\begin{aligned} &\left(\sum_{n=-\infty}^{\infty} |\operatorname{sinc}(\sigma\mu - n\pi)|^q \right)^{1/q} \\ &\quad \leq \left(1 + \sum_{n=1}^{\infty} \left\{ \frac{1}{(n\pi - \sigma x)^q} + \frac{1}{(n\pi + \sigma x)^q} \right\} \right)^{1/q} \exp(\sigma|y|) \\ &\quad \leq \left(1 + \frac{1}{\pi^q} \sum_{n=1}^{\infty} \left\{ \frac{1}{(n-1/2)^q} + \frac{1}{(n+1/2)^q} \right\} \right)^{1/q} \exp(\sigma|y|), \end{aligned} \tag{2.20}$$

where $\mu \in \mathcal{S}_\sigma$ and we have used the fact that the function $1/2(n\pi - \sigma x)^q + 1/(n\pi + \sigma x)^q$ has a maximum on $[-\pi/2\sigma, \pi/2\sigma]$ at $x := \pm\pi/2\sigma$. Substituting from (2.16) in (2.20) yields (2.17) when $r = 0$. Now we study the case when $r = 1$. For $\mu \in \mathcal{S}_\sigma$ and $n \neq 0$,

we obtain

$$\left|\operatorname{sinc}'(\sigma\mu - n\pi)\right| \le \frac{\sigma|\cos\sigma\mu|}{|\sigma\mu - n\pi|} + \frac{\sigma|\sin\sigma\mu|}{|\sigma\mu - n\pi|^2} \le \frac{\sigma\exp(\sigma|y|)}{|\sigma x - n\pi|} + \frac{\sigma\exp(\sigma|y|)}{|\sigma x - n\pi|^2}. \tag{2.21}$$

Also

$$\operatorname{sinc}'\sigma\mu = \frac{i}{2\sigma}\int_{-\sigma}^{\sigma} \omega\, e^{i\mu\omega} d\omega,$$

and therefore

$$\left|\operatorname{sinc}'\sigma\mu\right| \le \sigma\exp(\sigma|y|). \tag{2.22}$$

By using (2.21) and (2.22), we obtain for $\mu \in \mathcal{S}_\sigma$

$$\begin{aligned}
&\frac{1}{\sigma}\left(\sum_{n=-\infty}^{\infty}\left|\operatorname{sinc}'(\sigma\mu - n\pi)\right|^q\right)^{1/q} \\
&\quad\le \left(\left|\operatorname{sinc}'\sigma\mu\right|^q + \sum_{n\in\mathbb{Z}\setminus\{0\}}\left|\operatorname{sinc}'(\sigma\mu - n\pi)\right|^q\right)^{1/q} \\
&\quad\le \left(1 + \sum_{n\in\mathbb{Z}\setminus\{0\}}\left(\frac{1}{|n\pi - \sigma x|} + \frac{1}{|n\pi - \sigma x|^2}\right)^q\right)^{1/q} \exp(\sigma|y|). \qquad (2.23)
\end{aligned}$$

For $n \in \mathbb{Z}^+$ the functions

$$g_n(x) := \left(\frac{1}{(n\pi - \sigma x)} + \frac{1}{(n\pi - \sigma x)^2}\right)^q + \left(\frac{1}{(n\pi + \sigma x)} + \frac{1}{(n\pi + \sigma x)^2}\right)^q, \quad |x| \le \pi/2\sigma, \tag{2.24}$$

are positive even functions that attain their maximum at $x := \pm\pi/2\sigma$. Therefore by noting that $2/\pi + (2/\pi)^2 > 1$, we obtain

$$\begin{aligned}
&\left(1 + \sum_{n\in\mathbb{Z}\setminus\{0\}}\left(\frac{1}{|n\pi - \sigma x|} + \frac{1}{|n\pi - \sigma x|^2}\right)^q\right)^{1/q} \\
&\quad\le \left(1 + \sum_{n\in\mathbb{Z}\setminus\{0\}}\left(\frac{1}{\pi|n - 1/2|} + \frac{1}{\pi^2|n - 1/2|^2}\right)^q\right)^{1/q} \\
&\quad\le \left(\sum_{n=-\infty}^{\infty}\left(\frac{1}{\pi|n - 1/2|} + \frac{1}{\pi^2|n - 1/2|^2}\right)^q\right)^{1/q}. \qquad (2.25)
\end{aligned}$$

By applying the Minkowsky inequality to (2.25) and blogging in (2.23), we obtain

$$\begin{aligned}
&\left(\sum_{n=-\infty}^{\infty}\left|\operatorname{sinc}'(\sigma\mu - n\pi)\right|^q\right)^{1/q} \\
&\quad\le \sigma\left(\left(\sum_{n=-\infty}^{\infty}\frac{1}{\pi^q|n - 1/2|^q}\right)^{1/q} + \left(\sum_{n=-\infty}^{\infty}\frac{1}{\pi^{2q}|n - 1/2|^{2q}}\right)^{1/q}\right)\exp(\sigma|y|). \qquad (2.26)
\end{aligned}$$

Estimation (2.17) could be obtained by combining (2.14) and (2.26), $r = 1$. ∎

Now we derive an amplitude error estimate which will be used in our investigations.

THEOREM 2.6. *Let $\mu \in \mathbb{C}$. Assume that $f \in B_\sigma^2$ such that* (1.7) *and* (1.8) *hold. Then for $0 < \varepsilon \le \min\{\sigma/\pi, \pi/\sigma, 1/\sqrt{e}\}$ we have*

$$|(A_\varepsilon) f^{(r)}(\mu)| \le (r+1)\sigma^r A(\varepsilon) \exp(\sigma|\Im\mu|), \quad r = 0, 1,\ \mu \in \mathbb{C}. \tag{2.27}$$

Proof. Let $\frac{1}{q} + \frac{1}{p} = 1$, $p > 1$, and $r \in \{0, 1\}$. From the Hölder inequality and (2.17), we obtain

$$\begin{aligned} \left|(A_\varepsilon) f^{(r)}(\mu)\right| &\le \left(\sum_{n=-\infty}^{\infty} \left| \operatorname{sinc}^{(r)}(\sigma\mu - n\pi) \right|^q \right)^{1/q} \left(\sum_{n=-\infty}^{\infty} |\varepsilon_n|^p \right)^{1/p} \\ &\le (r+1)\sigma^r \left(\sum_{n=-\infty}^{\infty} |\varepsilon_n|^p \right)^{1/p} \exp(\sigma|\Im\mu|), \quad \mu \in \mathbb{C}. \end{aligned} \tag{2.28}$$

Let $N_0 := \left[\frac{\sigma}{\pi}\right]$ and $N \in \mathbb{Z}^+$ such that $N \ge N_0$. By applying the Minkowsky inequality, we obtain

$$\left(\sum_{n=-\infty}^{\infty} |\varepsilon_n|^p \right)^{1/p} \le \left(\sum_{|n|\le N} |\varepsilon_n|^p \right)^{1/p} + \left(\sum_{|n|>N} |\varepsilon_n|^p \right)^{1/p}. \tag{2.29}$$

Since $|\varepsilon_n| < \varepsilon$, then

$$\left(\sum_{n=-N}^{N} |\varepsilon_n|^p \right)^{1/p} \le \varepsilon(2N+1)^{1/p}. \tag{2.30}$$

By using (1.7) and (1.8), then for $p \ge 2$, we get

$$\left(\sum_{|n|>N} |\varepsilon_n|^p \right)^{1/p} \le \left(\sum_{|n|>N} \left| f\left(\frac{n\pi}{\sigma}\right) \right|^p \right)^{1/p} \le \sqrt{2} M_f \left(\frac{\sigma}{\pi}\right) N^{(1-p)/p}. \tag{2.31}$$

By combining (2.31), (2.30), and (2.28) we obtain, for $\mu \in \mathbb{C}$,

$$\left|(A_\varepsilon) f^{(r)}(\mu)\right| \le (r+1)\sigma^r p \left(\varepsilon(2N+1)^{1/p} + \sqrt{2} M_f \left(\frac{\sigma}{\pi}\right) N^{(1-p)/p} \right) \exp(\sigma|\Im\mu|). \tag{2.32}$$

The parameters N and p will be chosen suitably to obtain the desired estimate. We distinguish between two cases. When $\sigma \ge \pi$ we take

$$N := \left[\varepsilon^{-1} \left(\frac{\sigma}{\pi}\right)^{\frac{p}{p-1}} \right], \quad p = 4\log(1/\varepsilon). \tag{2.33}$$

Since $\varepsilon \le \{\pi/\sigma, 1/\sqrt{e}\}$, then $(\frac{\sigma}{\pi}) \le \frac{1}{\varepsilon}$ and $N \ge N_0$. Therefore

$$(2N+1)^{1/p} \le \sqrt{3}\, \varepsilon^{-1/p} \left(\frac{\pi}{\sigma}\right)^{\frac{1}{p-1}} \le \sqrt{3} \left(\frac{1}{\varepsilon}\right)^{\frac{2p-1}{p(p-1)}} \le \sqrt{3} \left(\frac{1}{\varepsilon}\right)^{\frac{4}{p}} \le \sqrt{3}\, e, \tag{2.34}$$

where the last inequality comes by taking the logarithm of both sides and using (2.33). Also

$$\left(\frac{\sigma}{\pi}\right) N^{(1-p)/p}) \leq \varepsilon^{\frac{p-1}{p}} = e^{1/4}\varepsilon. \tag{2.35}$$

By substituting from (2.35) and (2.34) in (2.32) and noting that $p = 4\log(1/\varepsilon)$, we obtain (2.27). If $\sigma < \pi$, we choose

$$N := \left[\varepsilon^{-1}\left(\frac{\pi}{\sigma}\right)^{\frac{p}{p-1}}\right]. \tag{2.36}$$

In this case $N_0 = 0$, and we proceed as in the previous case. This ends the proof. □

Remark 2.7. Letting $\mu \in \mathbb{R}$, estimation (2.17) turns out to be

$$\left(\sum_{n=-\infty}^{\infty} \left|\operatorname{sinc}^{(r)}(\sigma\mu - n\pi)\right|^q\right)^{1/q} < (r+1)\, p\, \sigma^r, \quad r = 0, 1, \tag{2.37}$$

leading to the special case obtained by Splettstösser, Stens, and Wilmes [23] when $r = 0$. Also if $\mu \in \mathbb{R}$, then (2.27) becomes

$$|(A_\varepsilon) f^{(r)}(\mu)| \leq (r+1)\sigma^r A(\varepsilon), \quad r = 0, 1. \tag{2.38}$$

In the special case $r = 0$ and $\sigma \geq \pi$ we obtain the result of Butzer, Splettstösser, and Stens (1.9).

3. The scheme and error estimates. In this section we discuss the method and study the error analysis. Let $y_1(\cdot, \mu)$ and $y_2(\cdot, \mu)$ denote the solutions of (1.10) satisfying the following initial conditions:

$$y_j^{(i-1)}(0, \mu) = \delta_{ij}, \qquad i, j = 1, 2, \quad \mu \in \mathbb{C}. \tag{3.1}$$

It is known [13, 21] that $y_i(x, \mu)$ is entire in μ for $x \in [0, b]$, and the eigenvalues of the problem (1.10)–(1.12) are the zeros of the characteristic determinant

$$\Delta(\mu) := \begin{vmatrix} U_1(y_1) & U_1(y_2) \\ U_2(y_1) & U_2(y_2) \end{vmatrix}. \tag{3.2}$$

The characteristic determinant $\Delta(\mu)$ is also an entire function of μ. By using the method of variation of constants we can write $y_1(x, \mu)$ and $y_2(x, \mu)$, respectively, as the solutions of the following Volterra integral equations:

$$y_1(x, \mu) = \cos(\mu x) + (\mathcal{T} y_1)(x, \mu), \qquad y_2(x, \mu) = \frac{\sin(\mu x)}{\mu} + (\mathcal{T} y_2)(x, \mu), \tag{3.3}$$

where $\mathcal{T}$ is the Volterra operator defined by

$$(\mathcal{T} y)(x, \mu) := \int_0^x \frac{\sin \mu(x - t)}{\mu}\, q(t) y(t, \mu) dt; \tag{3.4}$$

cf., e.g., [13, p. 92]. By differentiating (3.3) we obtain

$$y_1'(x, \mu) = -\mu \sin(\mu x) + (\widetilde{\mathcal{T}} y_1)(x, \mu), \qquad y_2'(x, \mu) = \cos(\mu x) + (\widetilde{\mathcal{T}} y_2)(x, \mu), \tag{3.5}$$

where $\widetilde{\mathcal{T}}$ is the Volterra operator defined by

$$(\widetilde{\mathcal{T}}y)(x,\mu) := \int_0^x \cos(\mu(x-t))\, q(t) y(t,\mu) dt; \tag{3.6}$$

see [13, p. 92]. The method of successive approximation yields

$$y_1(x,\mu) = \sum_{n=0}^{\infty} \mathcal{T}^n \cos(\mu x), \quad y_2(x,\mu) = \sum_{n=0}^{\infty} \mathcal{T}^n \frac{\sin(\mu x)}{\mu}, \tag{3.7}$$

where $\mathcal{T}^0$ is the identity operator and the convergence in (3.7) is uniform on $[0,b]$ for any $\mu \in \mathbb{C}$. By differentiating (3.7) we have [3], with uniform convergence on $[0,b]$,

$$y_1'(x,\mu) = -\mu \sin(\mu x) + \sum_{n=0}^{\infty} \widetilde{\mathcal{T}}\mathcal{T}^n \cos(\mu x), \quad y_2'(x,\mu) = \cos(\mu x) + \sum_{n=0}^{\infty} \widetilde{\mathcal{T}}\mathcal{T}^n \frac{\sin(\mu x)}{\mu}. \tag{3.8}$$

For convenience we let

$$\begin{aligned} &C_1 := \alpha_{11}\alpha_{22} - \alpha_{12}\alpha_{21}, \quad && C_2 := \alpha_{11}\beta_{22} - \alpha_{21}\beta_{12}, \quad && C_3 := \alpha_{12}\beta_{22} - \alpha_{22}\beta_{12}, \\ &C_4 := \alpha_{22}\beta_{11} - \alpha_{12}\beta_{21}, \quad && C_5 := \beta_{22}\beta_{11} - \beta_{12}\beta_{21}, \quad && C_6 := \alpha_{11}\beta_{21} - \alpha_{21}\beta_{11}. \end{aligned} \tag{3.9}$$

By substituting from (3.7) and (3.8) in (3.2) we obtain

$$\begin{aligned} \Delta(\mu) = {} & C_1 + C_2 \cos[b\mu] + C_3 \mu \sin[b\mu] + (C_4 + C_5 \cos[b\mu]) \sum_{n=0}^{\infty} \mathcal{T}^n[\cos(\mu x)](b) \\ & + (C_6 + C_5 \mu \sin[b\mu]) \sum_{n=0}^{\infty} \mathcal{T}^n \left[\frac{\sin(\mu x)}{\mu}\right](b) + C_3 \sum_{n=0}^{\infty} \widetilde{\mathcal{T}}\mathcal{T}^n[\cos(\mu x)](b) \\ & + C_2 \sum_{n=0}^{\infty} \widetilde{\mathcal{T}}\mathcal{T}^n \left[\frac{\sin(\mu x)}{\mu}\right](b) + C_5 \sum_{n=0}^{\infty} \mathcal{T}^n \left[\frac{\sin(\mu x)}{\mu}\right](b) \sum_{n=0}^{\infty} \widetilde{\mathcal{T}}\mathcal{T}^n[\cos(\mu x)](b) \\ & + C_5 \sum_{n=0}^{\infty} \mathcal{T}^n[\cos(\mu x)](b) \sum_{n=0}^{\infty} \widetilde{\mathcal{T}}\mathcal{T}^n \left[\frac{\sin(\mu x)}{\mu}\right](b). \end{aligned} \tag{3.10}$$

Let us split $\Delta(\mu)$ into two parts via

$$\Delta(\mu) := \mathcal{K}(\mu) + \mathcal{U}(\mu), \tag{3.11}$$

where $\mathcal{K}(\mu)$ is the known part

$$\begin{aligned} \mathcal{K}(\mu) := {} & C_1 + C_2 \cos[b\mu] + C_3 \mu \sin[b\mu] + (C_4 + C_5 \cos[b\mu]) \sum_{n=0}^{1} \mathcal{T}^n[\cos(\mu x)](b) \\ & + (C_6 + C_5 \mu \sin[b\mu]) \sum_{n=0}^{1} \mathcal{T}^n \left[\frac{\sin(\mu x)}{\mu}\right](b) \\ & + C_3 \sum_{n=0}^{1} \widetilde{\mathcal{T}}\mathcal{T}^n[\cos(\mu x)](b) + C_2 \widetilde{\mathcal{T}} \left[\frac{\sin(\mu x)}{\mu}\right](b) \\ & + C_5 \frac{\sin[b\mu]}{\mu} \sum_{n=0}^{1} \widetilde{\mathcal{T}}\mathcal{T}^n[\cos(\mu x)](b) + C_5 \widetilde{\mathcal{T}} \left[\frac{\sin[x\mu]}{\mu}\right](b) \sum_{n=0}^{1} \mathcal{T}^n[\cos(\mu x)](b) \end{aligned} \tag{3.12}$$

and $\mathcal{U}(\mu)$ is the infinite sum

(3.13)
$$\begin{aligned}\mathcal{U}(\mu) := &\left(C_4 + C_5\cos[b\mu] + C_5\widetilde{\mathcal{T}}\left[\frac{\sin(\mu x)}{\mu}\right](b)\right)\sum_{n=2}^{\infty}\mathcal{T}^n[\cos(\mu x)](b)\\ &+ C_2\sum_{n=1}^{\infty}\widetilde{\mathcal{T}}\mathcal{T}^n\left[\frac{\sin(\mu x)}{\mu}\right](b) + \left(C_3 + C_5\frac{\sin(\mu b)}{\mu}\right)\sum_{n=2}^{\infty}\widetilde{\mathcal{T}}\mathcal{T}^n[\cos(\mu x)](b)\\ &+ \ (C_6 + C_5\mu\sin[b\mu])\sum_{n=2}^{\infty}\mathcal{T}^n\left[\frac{\sin(\mu x)}{\mu}\right](b)\\ &+ C_5\sum_{n=0}^{1}\widetilde{\mathcal{T}}\mathcal{T}^n[\cos(\mu x)](b)\sum_{n=1}^{\infty}\mathcal{T}^n\left[\frac{\sin(\mu x)}{\mu}\right](b)\\ &+ C_5\sum_{n=1}^{\infty}\mathcal{T}^n\left[\frac{\sin(\mu x)}{\mu}\right](b)\sum_{n=2}^{\infty}\widetilde{\mathcal{T}}\mathcal{T}^n[\cos(\mu x)](b)\\ &+ C_5\sum_{n=0}^{1}\mathcal{T}^n[\cos(\mu x)](b)\sum_{n=1}^{\infty}\widetilde{\mathcal{T}}\mathcal{T}^n\left[\frac{\sin(\mu x)}{\mu}\right](b)\\ &+ C_5\sum_{n=2}^{\infty}\mathcal{T}^n[\cos(\mu x)](b)\sum_{n=1}^{\infty}\widetilde{\mathcal{T}}\mathcal{T}^n\left[\frac{\sin(\mu x)}{\mu}\right](b).\end{aligned}$$

The next step is to show that $\mathcal{U}(\mu)$ lies in a Paly–Wiener space B_σ^2 for some $\sigma > 0$. We start with the following introductory estimates.

LEMMA 3.1. *For $n \in \mathbb{N}$ and $\mu \in \mathbb{R}$, we have*

$$\left|\mathcal{T}^n\left[\frac{\sin(\mu x)}{\mu}\right](b)\right| \le \frac{\tau^n}{|\mu|^{n+1}}, \qquad \left|\mathcal{T}^n\big[\cos(\mu x)\big](b)\right| \le \frac{\tau^n}{|\mu|^n}, \tag{3.14}$$

$$\left|\widetilde{\mathcal{T}}\mathcal{T}^n\left[\frac{\sin(\mu x)}{\mu}\right](b)\right| \le \frac{\tau^{n+1}}{|\mu|^{n+1}}, \qquad \left|\widetilde{\mathcal{T}}\mathcal{T}^n\big[\cos(\mu x)\big](b)\right| \le \frac{\tau^{n+1}}{|\mu|^n}, \tag{3.15}$$

where $\tau := \int_0^b |q(t)|dt$.

Proof. We prove the first inequality of (3.14) by induction on n. Assume it holds when $n = m \in \mathbb{N}$, $m \ge 1$. Then

$$\left|\mathcal{T}^{m+1}\left[\frac{\sin(\mu x)}{\mu}\right](b)\right| \le \frac{1}{|\mu|}\int_0^b |q(t)|\left|\mathcal{T}^m\left[\frac{\sin(\mu x)}{\mu}\right](t)\right| dt \le \frac{\tau^{m+1}}{|\mu|^{m+2}}. \tag{3.16}$$

This proves the first inequality of (3.14) for all $n \in N$ since it is obvious when $n = 1$. Similarly, the second inequality of (3.14) holds. By using (3.14) and the definition of $\widetilde{\mathcal{T}}$, we can easily prove (3.15). ∎

THEOREM 3.2. *Assume that $q(\cdot) \in L^1(0,b)$, and then $\mathcal{U}(\mu) \in B_{2b}^2$.*

Proof. Since $\Delta(\mu)$ and $\mathcal{K}(\mu)$ are entire functions in μ, then so is $\mathcal{U}(\mu)$. From Lemma 3.1 above, $\mathcal{U}(\mu) = \mathcal{O}(\frac{1}{|\mu|^2})$ when $|\mu| \to \pm\infty$. Then $S(\mu) \in L^2(\mathbb{R})$. It remains to show that there is a fixed constant $M > 0$ such that

$$|\mathcal{U}(\mu)| \le M\exp(2b|\Im\mu|), \quad \mu \in \mathbb{C}. \tag{3.17}$$

Indeed, for $\mu \in \mathbb{C}$ (see [9]) we have

$$\begin{aligned} &\left|\sum_{n=k}^{\infty} \mathcal{T}^n[\cos(t\mu)](b)\right| \le \left(\sum_{n=k}^{\infty} \frac{\gamma(\mu)^n}{n!}\right) \exp(b|\Im\mu|), \\ &\left|\sum_{n=k}^{\infty} \widetilde{\mathcal{T}}\mathcal{T}^n[\cos(t\mu)](b)\right| \le \tau \left(\sum_{n=k}^{\infty} \frac{\gamma(\mu)^n}{n!}\right) \exp(b|\Im\mu|), \end{aligned} \tag{3.18}$$

where τ is defined above, and

$$\gamma(\mu) := \frac{cb\tau}{1+|\mu|b}, \qquad c := 1.709.$$

Similarly (cf. [3, p. 7])

$$\begin{aligned} &\left|\sum_{n=k}^{\infty} \mathcal{T}^n\left[\frac{\sin(t\mu)}{\mu}\right](b)\right| \le \frac{cb}{1+|\mu|b}\left(\sum_{n=k}^{\infty} \frac{\delta(\mu)^n}{n!}\right) \exp(b|\Im\mu|), \\ &\left|\sum_{n=k}^{\infty} \widetilde{\mathcal{T}}\mathcal{T}^n\left[\frac{\sin(t\mu)}{\mu}\right](b)\right| \le \frac{cb\tau}{1+|\mu|b}\left(\sum_{n=k}^{\infty} \frac{\delta(\mu)^n}{n!}\right) \exp(b|\Im\mu|), \end{aligned} \tag{3.19}$$

where

$$\delta(\mu) := \int_0^b \frac{ct}{1+|\mu|t}|q(t)|dt.$$

By using the inequalities (cf. [9])

$$|\cos b\mu| \le \exp(b|\Im\mu|), \quad |\sin b\mu| \le \frac{c\,b|\mu|}{1+b|\mu|}\ \exp(b|\Im\mu|), \quad \mu \in \mathbb{C}, \tag{3.20}$$

and the fact that $\gamma(\mu)$, $\delta(\mu)$ are bounded functions, we can find $M > 0$ such that (3.17) is satisfied, completing the proof. □

Remark 3.3. In the special case when $C_5 = 0$, $\mathcal{U}(\mu) \in B_b^2$.

The following lemma is necessary for the applicability of Theorem 2.6.

LEMMA 3.4. *Assume that $q(\cdot) \in L^1(0,b)$. Then there exists a positive constant M such that*

$$|\mathcal{U}(\mu)| \le M|\mu|^{-1}, \quad |\mu| \ge 1, \quad \mu \in \mathbb{R}. \tag{3.21}$$

Proof. From the definitions of $\gamma(\mu)$ and $\delta(\mu)$, we obtain

$$\gamma(\mu), \delta(\mu) \le \frac{cb\tau}{1+b|\mu|}, \quad \mu \in \mathbb{C}, \tag{3.22}$$

$$\sum_{n=2}^{\infty} \frac{\gamma(\mu)^n}{n!} \le \gamma(\mu)^2 e^{\gamma(\mu)} \le \gamma(\mu)^2 e^{c\,b\tau}, \quad \mu \in \mathbb{C}. \tag{3.23}$$

Similarly

$$\sum_{n=2}^{\infty} \frac{\delta(\mu)^n}{n!} \le \delta(\mu)^2 e^{c\,b\tau}, \quad \mu \in \mathbb{C}. \tag{3.24}$$

For $\mu \in \mathbb{R}$ and by combining (3.24), (3.23), (3.22), (3.19), and (3.18), we can find a constant M such that

$$|\mathcal{U}(\mu)| \leq \frac{b^2 M}{(1+b|\mu|)^2}, \quad \mu \in \mathbb{R}, \tag{3.25}$$

which is reducible to (3.21) when $|\mu| \geq 1$. □

Applying the WKS sampling theorem to $\mathcal{U}(\mu)$, we obtain, for $\sigma = 2b$,

$$\mathcal{U}(\mu) := \sum_{n=-\infty}^{\infty} \mathcal{U}\left(\frac{n\pi}{2b}\right) \operatorname{sinc}(2b\mu - n\pi), \qquad \mu \in \mathbb{C}. \tag{3.26}$$

Let $N \in \mathbb{Z}^+$ and $\mathcal{U}_N(\mu)$ be the truncated sum of (3.26)

$$\mathcal{U}_N(\mu) := \sum_{n=-N}^{N} \mathcal{U}\left(\frac{n\pi}{2b}\right) \operatorname{sinc}(2b\mu - n\pi), \qquad \mu \in \mathbb{C}. \tag{3.27}$$

The use of $\mathcal{U}_N(\mu)$ produces a truncation error, i.e., $|\mathcal{U}(\mu) - \mathcal{U}_N(\mu)| \leq T_N(\mu)$; cf. Theorem 2.1 above, where $\mu \in \Omega_N$ and

$$T_N(\mu) := \frac{4\sqrt{2}E}{\pi^2\sqrt{3}} \, \frac{b\sqrt{b}\,|\sin 2b\mu|}{N+1} \left(\frac{1}{\sqrt{N\pi - 2b\Im\mu}} + \frac{1}{\sqrt{N\pi + 2b\Im\mu}} \right). \tag{3.28}$$

The samples $\mathcal{U}(n\pi/2b) = \Delta(n\pi/2b) - \mathcal{K}(n\pi/2b)$, $|n| \leq N$, cannot be computed explicitly in the general case. We approximate these samples by solving $4N+2$ initial-value problems defined by (1.10) and (3.1) to obtain the approximate values $\widetilde{\mathcal{U}}(n\pi/2b)$, $|n| \leq N$, i.e., $\widetilde{\mathcal{U}}_N\left(\frac{n\pi}{2b}\right) = \widetilde{\Delta}\left(\frac{n\pi}{2b}\right) - \mathcal{K}\left(\frac{n\pi}{2b}\right)$, and therefore we define $\widetilde{\mathcal{U}}(\mu)$ to be

$$\widetilde{\mathcal{U}}_N(\mu) := \sum_{n=-N}^{N} \widetilde{\mathcal{U}}\left(\frac{n\pi}{2b}\right) \operatorname{sinc}(2b\mu - n\pi). \tag{3.29}$$

For convenience let $A(\varepsilon, \mu) := A(\varepsilon)\exp(\sigma|\Im\mu|)$, where $A(\varepsilon)$ is defined above. Since $\mathcal{U}(\mu)$ satisfies the condition (3.21), then

$$\left|\mathcal{U}_N(\mu) - \widetilde{\mathcal{U}}_N(\mu)\right| \leq A(\varepsilon, \mu), \quad \mu \in \mathbb{C}. \tag{3.30}$$

Now let $\widetilde{\Delta}_N(\mu) := \mathcal{K}(\mu) + \widetilde{\mathcal{U}}_N(\mu)$. From (3.11) and (3.30) we obtain

$$|\Delta(\mu) - \widetilde{\Delta}_N(\mu)| \leq T_N(\mu) + A(\varepsilon, \mu), \quad \mu \in \Omega_N. \tag{3.31}$$

By similar treatment we can get for the derivative of $\Delta(\mu)$

$$|\Delta'(\mu) - \widetilde{\Delta'}_N(\mu)| \leq T_{1,N}(\mu) + 4bA(\varepsilon, \mu) \quad \mu \in \Omega_N. \tag{3.32}$$

Let $(\mu^*)^2$ be an eigenvalue and $(\mu_N)^2$ be its desired approximation, i.e., $\Delta(\mu^*) = 0$ and $\widehat{\Delta}_N(\mu_N) = 0$. From (3.31) we have $|\widetilde{\Delta}_N(\mu^*)| \leq T_N(\mu^*) + A(\varepsilon, \mu^*)$. Let Π_N be the region

$$\mu^* \in \Pi_N := \left\{\mu \in \Omega_N \text{ such that } \left|\widetilde{\Delta}_N(\mu)\right| \leq T_N(\mu) + A(\varepsilon, \mu)\right\}. \tag{3.33}$$

From Π_N we determine an enclosure annulus where μ_N lies in it. We solve the equation $\widetilde{\Delta}_N(\mu) = \pm(T_N(\mu) + A(\varepsilon, \mu))$. Let A_N be the set of all of these solutions. Then $\mu^* \in A_N$. Let $\mu_0, \mu_1 \in \Omega_N$ be

$$\mu_0 := \inf\{|\mu| : \ \mu \in \mathcal{A}_N\}, \quad \mu_1 := \sup\{|\mu| : \ \mu \in \mathcal{A}_N\}. \tag{3.34}$$

The enclosure annulus where μ_N lies is $\mathcal{A}_N := \{\mu \in \Omega_N : \ \mu_0 \le \mu \le \mu_1\}$. If $\mu^* \in \mathbb{R}$, the enclosure annulus becomes an interval I_N. The computation of the error bounds of $|\mu^* - \mu_N|$ depends on whether μ^* is simple or double.

THEOREM 3.5. *Let $(\mu^*)^2$ be a simple eigenvalue of* (1.10)–(1.12). *Then for all $\epsilon > 0$, there exists $N_0 \in \mathbb{N}$ such that for any positive integer N, $N \ge N_0$, there exists μ_N such that $|\mu^* - \mu_N| < \epsilon$. Furthermore, we have the following estimate:*

$$|\mu^* - \mu_N| < \frac{T_N(\mu_N) + A(\varepsilon, \mu_N)}{\inf_{\xi_1, \xi_2 \in \Gamma_{\mu^*, \mu_N}} \sqrt{(\Re\Delta'(\xi_1))^2 + (\Im\Delta'(\xi_2))^2}}. \tag{3.35}$$

Proof. The first part arises from the fact that $\widetilde{\Delta}_N(\mu)$ converges uniformly to $\Delta(\mu)$ on compact subsets of $\mathbb{C}$. By replacing μ by μ_N in (3.31) we obtain

$$|\Delta(\mu_N) - \Delta(\mu^*)| < T_N(\mu_N) + A(\varepsilon, \mu_N), \tag{3.36}$$

where we have used $\widetilde{\Delta}_N(\mu_N) = \Delta(\mu^*) = 0$. From the complex mean value theorem (cf. (1.16)), there exist $\xi_1, \xi_2 \in \Gamma_{\mu^*, \mu_N}$ such that

$$\begin{aligned} |(\overline{\mu_N} - \overline{\mu^*})(\Delta(\mu_N) - \Delta(\mu^*))| &= \big[(\Re\{(\overline{\mu_N} - \overline{\mu^*})(\Delta(\mu_N) - \Delta(\mu^*))\})^2 \\ &\quad + (\Im\{(\overline{\mu_N} - \overline{\mu^*})(\Delta(\mu_N) - \Delta(\mu^*))\})^2\big]^{1/2} \\ &= |\mu_N - \mu^*|^2 \sqrt{(\Re\Delta'(\xi_1))^2 + (\Im\Delta'(\xi_2))^2}. \end{aligned} \tag{3.37}$$

By combining (3.36) and (3.37) we obtain

$$|\mu_N - \mu^*| \sqrt{(\Re\Delta'(\xi_1))^2 + (\Im\Delta'(\xi_2))^2} < T_N(\mu_N) + A(\varepsilon, \mu_N). \tag{3.38}$$

Since μ^* is a simple zero of $\Delta(\mu)$ and N is sufficiently large, then

$$\inf_{\xi_1, \xi_2 \in \Gamma_{\mu^*, \mu_N}} \sqrt{(\Re\Delta'(\xi_1))^2 + (\Im\Delta'(\xi_2))^2} > 0,$$

and hence

$$|\mu^* - \mu_N| < \frac{T_N(\mu_N) + A(\varepsilon, \mu_N)}{\inf_{\xi_1, \xi_2 \in \Gamma_{\mu^*, \mu_N}} \sqrt{(\Re\Delta'(\xi_1))^2 + (\Im\Delta'(\xi_2))^2}},$$

which is (3.35). ∎

Now we estimate the error when $(\mu^*)^2$ is a double zero of $\Delta(\mu)$, i.e., $\Delta(\mu^*) = 0$, $\Delta'(\mu^*) = 0$ and $\Delta''(\mu^*) \ne 0$. In this case estimate (3.35) has no meaning.

THEOREM 3.6. *Let $(\mu^*)^2$ be a double eigenvalue of* (1.10)–(1.12). *Then for all $\epsilon > 0$, there exists $N_0 \in \mathbb{N}$ such that for any positive integer N, $N \ge N_0$, there exists μ_N such that $|\mu^* - \mu_N| < \epsilon$. Furthermore, we have the following estimate:*

$$|\mu^* - \mu_N| < \frac{T_{1,N}(\mu_N) + 4bA(\varepsilon, \mu_N)}{\inf_{\xi_1, \xi_2 \in \Gamma_{\mu^*, \mu_N}} \sqrt{(\Re\Delta''(\xi_1))^2 + (\Im\Delta''(\xi_2))^2}}. \tag{3.39}$$

Proof. The first part arises from the uniform convergence of $\widetilde{\Delta'}_N(\mu)$ to $\Delta'(\mu)$. Since $\widetilde{\Delta'}_N(\mu_N) = \Delta'(\mu^*) = 0$, then

$$|\Delta'(\mu_N) - \Delta'(\mu^*)| \le T_{1,N}(\mu_N) + 4bA(\varepsilon, \mu_N).$$

By similar treatments of (3.37) and using the complex mean value theorem, we obtain

(3.40)
$$|(\mu^* - \mu_N)| \sqrt{(\Re\Delta''(\xi_1))^2 + (\Im\Delta''(\xi_2))^2} < T_{1,N}(\mu_N) + 4bA(\varepsilon, \mu_N), \quad \xi_1, \xi_2 \in \Gamma_{\mu^*,\mu_N}.$$

Since μ^* is a double zero of $\Delta(\mu)$ and N is sufficiently large, then

$$\inf_{\xi_1,\xi_2 \in \Gamma_{\mu^*,\mu_N}} \sqrt{(\Re\Delta''(\xi_1))^2 + (\Im\Delta''(\xi_2))^2} > 0,$$

and hence

$$|\mu^* - \mu_N| < \frac{T_{1,N}(\mu_N) + 4bA(\varepsilon, \mu_N)}{\inf_{\xi_1,\xi_2 \in \Gamma_{\mu^*,\mu_N}} \sqrt{(\Re\Delta''(\xi_1))^2 + (\Im\Delta''(\xi_2))^2}},$$

which is (3.39). ∎

Remark 3.7. Let μ^* be a real and simple eigenvalue. Then the estimate in (3.35) becomes

$$|\mu^* - \mu_N| < \frac{T_N(\mu_N) + A(\varepsilon)}{\inf_{\xi \in I_{\mu^*,\mu_N}} |\Delta'(\xi)|}, \tag{3.41}$$

where $I_{\mu^*,\mu_N} \subset \mathbb{R}$. Also, when μ^* is real and double, the estimate (3.39) turns out to be

$$|\mu^* - \mu_N| < \frac{T_{1,N}(\mu_N) + 4bA(\varepsilon)}{\inf_{\xi \in \Gamma_{\mu^*,\mu_N}} |\Delta''(\xi)|}. \tag{3.42}$$

The authors of [7, 11] established the so-called higher order approximations for the eigenvalues of special cases of (1.10)–(1.12). In these approximations $\mathcal{K}(\mu)$ of (3.12) has more terms. However, the omissions mentioned in the introduction exist here, too. We can do the same in the present situation since if we take more terms in $\mathcal{K}(\mu)$, the remainder will also lie in a Paley–Wiener space. The error analysis will be as established above with better estimates. In [1] the authors established the technique for second order pencils and considered both amplitude and truncation errors. Finally we compute an estimate for the constants E and M which appear in the quantities $T_N(\mu)$, $T_{1,N}(\mu)$, and $\mathcal{A}(\varepsilon, \mu)$. These constants play a major role in the numerical computations, especially to find enclosure annuli and intervals.

LEMMA 3.8. *The constants E and M may be estimated by*

$$\begin{aligned} E \le & \sqrt{2}\rho_1 \sqrt{\int_0^{cb\tau} \left(\frac{f_1(\eta)}{\eta}\right)^2 d\eta} + \sqrt{2}\rho_2 \sqrt{\int_0^{cb\tau} \left(\frac{f_2(\eta)}{\eta}\right)^2 d\eta} \\ & + 2\sqrt{2}c\,\tau |C_5| \sqrt{\int_0^{cb\tau} \left(\frac{f_1(\eta)^2}{\eta}\right)^2 d\eta}, \end{aligned} \tag{3.43}$$

$$M \le \rho_1 f_1(cb\tau) + \rho_2 f_2(cb\tau) + 2\sqrt{2}c\,\tau |C_5| (f_1(cb\tau))^2, \tag{3.44}$$

where

$$f_k(\eta) := \sum_{n\geq k} \frac{\eta^n}{n!}, \quad k = 1, 2,$$

(3.45)
$$\rho_1 := c\,\tau|C_4| + c\,\tau^2|C_3| + (3c\,\tau + 2b\,c^2\,\tau^2)|C_5|, \quad \rho_1 := c\,\tau|C_2| + c\,\tau|C_5| + c\,|C_6|.$$

Proof. By substituting from (3.18)–(3.20) in (3.13), for $\mu \in \mathbb{R}$, we obtain

$$\begin{aligned}|\mathcal{U}(\mu)| \leq{}& (|C_4| + |C_5|) \sum_{n\geq 2} \frac{\gamma(\mu)^n}{n!} + \left(\frac{cb|C_6|}{1+|\mu|b} + c\,|C_5| + |C_2|\gamma(\mu)\right) \sum_{n\geq 2} \frac{\delta(\mu)^n}{n!} \\ &+ \tau|C_3| \sum_{n\geq 2} \frac{\gamma(\mu)^n}{n!} + |C_5|\gamma(\mu) \sum_{n\geq 2} \frac{\delta(\mu)^n}{n!} + 2|C_5|\gamma(\mu) \sum_{n\geq 2} \frac{\gamma(\mu)^n}{n!} \sum_{m\geq 1} \frac{\delta(\mu)^m}{m!} \\ &+ |C_5|\gamma(\mu) \sum_{n\geq 2} \frac{\gamma(\mu)^n}{n!} + 2|C_5|\gamma(\mu)\{1+\gamma(\mu)\} \sum_{n\geq 1} \frac{\delta(\mu)^n}{n!}. \end{aligned} \tag{3.46}$$

Since $0 < \delta(\mu) < \gamma(\mu)$, $\gamma(\mu) < c\,b\tau$, and $|\mu|\,\gamma(\mu) < c\tau$, we have

$$|\mu\,\mathcal{U}(\mu)| \leq \rho_1 \sum_{n\geq 1} \frac{\gamma(\mu)^n}{n!} + \rho_2 \sum_{n\geq 2} \frac{\gamma(\mu)^n}{n!} + 2c\,\tau|C_5| \left(\sum_{n\geq 1} \frac{\gamma(\mu)^n}{n!}\right)^2, \tag{3.47}$$

implying (3.44). From the definitions of E and f_k, inequality (3.47), and the triangle inequality, we obtain

$$E \leq \sqrt{2}\rho_1 \|f_1(\gamma(\mu))\|_2 + \sqrt{2}\rho_2 \|f_2(\gamma(\mu))\|_2 + 2\sqrt{2}c\,\tau|C_5|\;\|f_1(\gamma(\mu))^2\|_2,$$

where

$$\|f_k\|_2 := \sqrt{\int_0^\infty |f_k(\gamma(\mu))|^2 d\mu}.$$

Now we estimate the norm $\|f_k(\gamma(\mu))\|_2$. It is easy to see, from the definition of $\gamma(\mu)$, that

$$\gamma(\mu)^2 = -c\,\tau\,\gamma'(\mu), \quad \text{where} \quad 0 < \mu < \infty.$$

Thus

$$\int_0^\infty |f_k(\gamma(\mu))|^2 d\mu \leq c\tau \int_\infty^0 \frac{f_k(\gamma(\mu))^2}{d(\mu)^2} d[\gamma(\mu)] \leq c\tau \int_0^{cb\tau} \frac{f_k(\eta)^2}{\eta^2} d\eta. \tag{3.48}$$

By substituting from (3.48) in (3.48) one gets (3.43). □

The estimates (3.43) and (3.44) can be simplified with greater bounds. Such simplifications will affect the accuracy of the numerical results.

4. Examples and comparisons. This section includes some new examples illustrating the above method. We derive five examples exhibiting the previous scheme in different situations. The first example is a self-adjoint boundary-value problem

with simple eigenvalues. The second is also self-adjoint but all eigenvalues are double. Then we give two non-self-adjoint examples with double and complex eigenvalues, respectively. The last example is an eigenvalue problem with $q(x) = x^2$. In the following, μ_k denotes the exact eigenvalue, μ_k^* denotes the approximate eigenvalue computed by the method of section 3, and $I_{k,N}$ denotes the enclosure interval. Every example is illustrated with a table where the effect of the amplitude error, which has not been considered in similar studies, is indicated through different choices of the bound ε. We have used *Mathematica* to approximate results to the nearest suitable decimal representations.

Example 4.1. Consider the self-adjoint boundary-value problem

$$-y''(x) + q(x)y(x) = \mu^2 y(x), \quad 0 \le x \le 1, \tag{4.1}$$

$$U_1(y) := y(0) + y'(0) + 2y(1) = 0, \tag{4.2}$$

$$U_2(y) := 2y'(0) + y(1) + y'(1) = 0, \tag{4.3}$$

where

$$q(x) = \begin{cases} 1, & 0 \le x < 1/2, \\ -1, & 1/2 \le x \le 1. \end{cases} \tag{4.4}$$

This problem is a special case of problem (1.10)–(1.12) when $\alpha_{11} = \alpha_{12} = \beta_{21} = 1$, $\alpha_{22} = \beta_{11} = 2$, and $\alpha_{21} = \beta_{12} = 0$. It is obviously self-adjoint. The characteristic determinant is

$$\Delta(\mu) := 4 + 4\cos\sqrt{1+\mu^2} + \sqrt{1+\mu^2}\sin\sqrt{1+\mu^2} + \frac{\sin\sqrt{1+\mu^2}}{\sqrt{1+\mu^2}}. \tag{4.5}$$

The boundary conditions are strongly regular (cf. [21]) and the eigenvalues are simple.

$$N = 20, \quad E = 85.82, \quad M = 338.92$$

μ	μ_k^*	μ_k	$\lvert\mu_k - \mu_k^*\rvert$	$I_{k,N}$, $\varepsilon = 10^{-5}$	$I_{k,N}$, $\varepsilon = 10^{-10}$
μ_1	2.9781814	2.9781881	6.68×10^{-6}	[2.83659, 3.08105]	[2.93715, 3.00631]
μ_2	4.8639142	4.8639061	8.19×10^{-6}	[4.78021, 4.96777]	[4.84235, 4.89277]
μ_3	9.3715736	9.3715761	2.60×10^{-6}	[9.35701, 9.39491]	[9.36730, 9.37526]
μ_4	11.8812419	11.8812150	2.691×10^{-5}	[11.82362, 11.93596]	[11.84949, 11.91175]

Example 4.2. The antiperiodic boundary-value problem

$$y''(x) + y(x) = -\mu^2 y(x), \quad 0 \le x \le 1, \tag{4.6}$$

$$U_1(y) := y(0) + y(1) = 0, \tag{4.7}$$

$$U_2(y) := y'(0) + y'(1) = 0 \tag{4.8}$$

is self-adjoint. Here $q(x) = -1$, $\alpha_{11} = \alpha_{22} = \beta_{22} = \beta_{11} = 1$, and $\beta_{12} = \beta_{21} = \alpha_{21} = \alpha_{12} = 0$. The characteristic determinant is

$$\Delta(\mu) := 4\cos^2\frac{\sqrt{1+\mu^2}}{2}. \tag{4.9}$$

Thus the eigenvalues are $\mu_k^2 = ((2k-1)\pi)^2 - 1,\ k \in \mathbb{Z}$. The eigenvalues are all double from geometric and algebraic points of view.

$$N = 20, \quad E = 19.9980, \quad M = 123.443$$

μ	μ_k^*	μ_k	$\lvert\mu_k^* - \mu_k\rvert$	$I_{k,N}$, $\varepsilon = 10^{-5}$	$I_{k,N}$, $\varepsilon = 10^{-10}$
μ_1	2.9781853	2.9781881	2.85×10^{-6}	[2.52216, 3.28217]	[2.7186, 3.0814]
μ_2	9.3714923	9.3715761	9.51×10^{-6}	[8.9513, 9.6402]	[9.1697, 9.4140]
μ_3	15.6756866	15.6761000	1.79×10^{-5}	[15.2540, 15.9349]	[15.4769, 15.7037]
μ_4	21.9683690	21.9684004	3.14×10^{-5}	[21.5334, 22.2151]	[21.7500, 21.9891]

Example 4.3. The non-self-adjoint problem

$$-y''(x) + q(x)y(x) = \mu^2 y(x), \quad 0 \le x \le 1, \tag{4.10}$$

$$U_1(y) := 2y(0) + y(1) = 0, \tag{4.11}$$

$$U_1(y) := y'(0) + y'(1) = 0 \tag{4.12}$$

is a special case of problem (1.10)–(1.12) when $\alpha_{12} = \alpha_{21} = \beta_{12} = \beta_{21} = 0$, $\beta_{11} = \beta_{22} = \alpha_{22} = 1$, $\alpha_{11} = 2$, and

$$q(x) := \begin{cases} -1, & 0 \le x < 1/2, \\ 0, & 1/2 \le x \le 1. \end{cases} \tag{4.13}$$

Its characteristic determinant is

$$\Delta(\mu) := 3 + 3\cos\sqrt{\mu^2 + 1}. \tag{4.14}$$

Thus the exact eigenvalues are $\mu_k^2 = ((2k-1)\pi)^2 - 1,\ k \in \mathbb{Z}$. Obviously all eigenvalues are algebraically double.

$$N = 20, \quad E = 11.2777, \quad M = 7.8348$$

μ	μ_k^*	μ_k	$\lvert\mu_k - \mu_k^*\rvert$	$I_{k,N}$, $\varepsilon = 10^{-5}$	$I_{k,N}$, $\varepsilon = 10^{-10}$
μ_1	2.9767465	2.9781881	1.44×10^{-3}	[2.8139, 3.0806]	[2.8403, 3.0542]
μ_2	9.3710727	9.3715762	5.03×10^{-4}	[9.2407, 9.44124]	[9.2759, 9.4059]
μ_3	15.6759211	15.676100	1.79×10^{-4}	[15.5485, 15.7380]	[15.5867, 15.6996]
μ_4	21.9684223	21.9684004	2.19×10^{-5}	[21.8358, 22.0251]	[21.8739, 21.9868]

Example 4.4. Consider the non-self-adjoint boundary-value problem

$$y''(x) + (1+i)y(x) = -\mu^2 y(x), \quad 0 \le x \le 1, \tag{4.15}$$

$$U_1(y) := y(0) + y(1) = 0, \tag{4.16}$$

$$U_2(y) := 2y(0) + y(1) = 0. \tag{4.17}$$

In this example we have $\alpha_{12} = \alpha_{22} = \beta_{22} = 0$, $\alpha_{11} = \beta_{12} = \beta_{11} = \beta_{21} = 1$, $\alpha_{21} = 2$, and $g(x) = 1 + i$. The characteristic determinant is

$$\Delta(\mu) := \frac{\sin\sqrt{1 + i + \mu^2}}{\sqrt{1 + i + \mu^2}}, \tag{4.18}$$

The exact eigenvalues are $\mu_k^2 = (k\pi)^2 - 1 - i,\ k \in \mathbb{Z}\backslash\{0\}$. Here also all eigenvalues are simple.

$$N = 20$$

μ	μ_k^*	μ_k	$\lvert\mu_k - \mu_k^*\rvert$
μ_1	2.98407-0.16864i	2.98290-0.16762i	1.55×10^{-3}
μ_2	6.20305-0.08001i	6.20362-0.08060i	8.20×10^{-4}
μ_3	9.37212-0.05375i	9.37173-0.05335i	5.63×10^{-4}
μ_4	12.52658-0.03991i	12.52629-0.03960i	4.30×10^{-4}

Example 4.5. The boundary-value problem

$$-y''(x) + x^2 y(x) = \mu^2 y(x), \quad 0 \le x \le 1, \tag{4.19}$$

$$U_1(y) := y(0) + y'(0) = 0, \tag{4.20}$$

$$U_2(y) := y(1) + y'(1) = 0 \tag{4.21}$$

with the above separate boundary conditions is a special case of problem (1.10)–(1.12) when $\alpha_{11} = \alpha_{12} = \beta_{21} = \beta_{22} = 1$, $\alpha_{21} = \alpha_{22} = \beta_{11} = \beta_{12} = 0$, and $q(x) = x^2$. All eigenvalues are simple.

$$E = 5.832, \quad M = 1.463$$

μ	μ_k^*	μ_k	$\lvert\mu_k - \mu_k^*\rvert$	$I_{k,N}$, $\varepsilon = 10^{-5}$	$I_{k,N}$, $\varepsilon = 10^{-10}$
μ_1	3.2154614	3.2154536	7.82×10^{-6}	[3.21358, 3.21737]	[3.21451, 3.21642]
μ_2	6.3126426	6.3126433	7.93×10^{-7}	[6.31191, 6.31339]	[6.31243, 6.31286]
μ_3	9.4433444	9.4433442	1.51×10^{-7}	[9.44308, 9.44360]	[9.44325, 9.44344]
μ_4	12.5800077	12.5800078	8.36×10^{-8}	[12.57979, 12.58022]	[12.57995, 12.58006]

Acknowledgment. The authors thank the referee who suggested the simplification of equality (2.15) and adding reference [19].

REFERENCES

[1] M. H. Annaby and M. M. Tharwat, *On computing eigenvalues of second-order linear pencils*, IMA J. Numer. Anal., 27 (2007), pp. 366–380.
[2] G. D. Birkhoff, *Boundary value and expansion problems of ordinary linear differential operators*, Trans. Amer. Math. Soc., 9 (1908), pp. 373–395.
[3] A. Boumenir, *The sampling method for Sturm-Liouville problems with the eigenvalue parameter in the boundary condition*, Numer. Funct. Anal. Optim., 21 (2000), pp. 67–75.
[4] A. Boumenir and B. Chanane, *Eigenvalues of Sturm-Liouville systems using sampling theory*, Appl. Anal., 62 (1996), pp. 323–334.
[5] A. Boumenir, *Sampling and eigenvalues of non-self-adjoint Sturm–Liouville problems*, SIAM J. Sci. Comput., 23 (2001), pp. 219–229.
[6] A. Boumenir, *Computing eigenvalues of a periodic Sturm-Liouville problem by the Shannon Whittaker sampling theorem*, Math. Comp., 68 (1999), pp. 1057–1066.
[7] A. Boumenir, *Higher approximation of eigenvalues by sampling*, BIT, 40 (2000), pp. 215–225.
[8] P. L. Butzer, W. Splettstösser, and R. L. Stens, *The sampling theorem and linear prediction in signal analysis*, Jahresber. Deutsch. Math.-Verein., 90 (1988), pp. 1–70.
[9] K. Chadan and P. C. Sabatier, *Inverse Problems in Quantum Scattering Theory*, Springer-Verlag, Berlin, 1989.
[10] B. Chanane, *Computing eigenvalues of regular Sturm-Liouville problems*, Appl. Math. Lett., 12 (1999), pp. 119–125.

[11] B. Chanane, *High order approximations of the eigenvalues of regular Sturm-Liouville problems*, J. Math. Anal. Appl., 226 (1998), pp. 121–129.
[12] B. Chanane, *High order approximations of the eigenvalues of Sturm-Liouville problems with coupled self-adjoint boundary conditions*, Appl. Anal., 80 (2001), pp. 317–330.
[13] M. S. P. Eastham, *Theory of Ordinary Differential Operators*, Van Nostrand Reinhold, London, 1970.
[14] M. S. P. Eastham, *The Spectral Theory of Periodic Differential Equations*, Scottish Academic Press, London, 1973.
[15] J.-Cl. Evard and F. Jafari, *A complex Rolle's theorem*, Amer. Math. Monthly, 99 (1992), pp. 858–861.
[16] J. R. Higgins, *Sampling Theory in Fourier and Signal Analysis*, Clarendon Press, Oxford, 1996.
[17] H. Hochstadt, *The Functions of Mathematical Physics*, Dover, New York, 1971.
[18] D. Jagerman, *Bounds for truncation error of the sampling expansion*, SIAM J. Appl. Math., 14 (1966), pp. 714–723.
[19] M. Kowalski, K. Sikorski, and F. Stenger, *Selected Topics in Approximation and Computation*, Oxford University Press, Oxford, 1995.
[20] J. Lund and K. Bowers, *Sinc Methods for Quadrature and Differential Equations*, SIAM, Philadelphia, PA, 1992.
[21] M. A. Naimark, *Linear Differential Operators*, George Harrap, London, 1967.
[22] S. Salaff, *Regular boundary conditions for ordinary differential operators*, Trans. Amer. Math. Soc., 134 (1968), pp. 355–373.
[23] W. Splettstösser, R. L. Stens, and G. Wilmes, *On approximation by the interpolation series of G. Valiron*, Funct. Approx. Comment. Math., 11 (1981), pp. 39–56.
[24] F. Stenger, *Numerical methods based on Whittaker cardinal, or sinc functions*, SIAM Rev., 23 (1981), pp. 165–224.
[25] F. Stenger, *Numerical Methods Based on Sinc and Analytic Functions*, Springer-Verlag, New York, 1993.

SIAM J. NUMER. ANAL.
Vol. 46, No. 2, pp. 691–721
© 2008 Society for Industrial and Applied Mathematics

CONVERGENCE OF THE MASS-TRANSPORT STEEPEST DESCENT SCHEME FOR THE SUBCRITICAL PATLAK–KELLER–SEGEL MODEL*

ADRIEN BLANCHET†, VINCENT CALVEZ‡, AND JOSÉ A. CARRILLO§

Abstract. Variational steepest descent approximation schemes for the modified Patlak–Keller–Segel equation with a logarithmic interaction kernel in any dimension are considered. We prove the convergence of the suitably interpolated in time implicit Euler scheme, defined in terms of the Euclidean Wasserstein distance, associated with this equation for subcritical masses. As a consequence, we recover the recent result about the global in time existence of weak solutions to the modified Patlak–Keller–Segel equation for the logarithmic interaction kernel in any dimension in the subcritical case. Moreover, we show how this method performs numerically in dimension one. In this particular case, this numerical scheme corresponds to a standard implicit Euler method for the pseudoinverse of the cumulative distribution function. We demonstrate its capabilities to reproduce the blow-up of solutions for supercritical masses easily without the need of mesh-refinement.

Key words. Patlak–Keller–Segel model, steepest descent, Wasserstein distance

AMS subject classifications. 65M99, 35K55, 35Q99

DOI. 10.1137/070683337

1. Introduction.

The Patlak–Keller–Segel (PKS) equation is widely used in mathematical biology to model the collective motion of cells which are attracted by a self-emitted chemical substance, the slime mold amoebae *Dictyostelium discoideum* being a prototype organism for this behavior. Moreover, the PKS equation has become a paradigmatic mathematical problem since it shows a concentration-collapse dichotomy: for masses larger than a critical value, solutions aggregate their mass, as delta Diracs, in finite time, while below this critical mass threshold solutions exist globally and disperse, collapsing down to zero. These coexistence of phenomena in this simple-looking mathematical model makes appealing and difficult to develop numerical schemes capable of dealing with both situations.

Historically, the first mathematical models in chemotaxis were introduced in 1953 by Patlak in [40] and in 1970 by Keller and Segel in [33]. Here, we focus on the modified PKS system for the log interaction kernel introduced by Perthame, the second author, and Sharifi tabar in [14]:

*Received by the editors February 21, 2007; accepted for publication (in revised form) October 5, 2007; published electronically February 1, 2008.
http://www.siam.org/journals/sinum/46-2/68333.html

†Centre de Recerca Matemàtica, Universitat Autònoma de Barcelona, E-08193 Bellaterra, Spain (blanchet@ceremade.dauphine.fr, http://www.ceremade.dauphine.fr/∼blanchet/). This author's research was supported by ESF Network "Global" and bourse Lavoisier, and partially supported by the Acc. Integ./Picasso program HF2006-0198.

‡Département de Mathématiques et Applications, Ecole Normale Supérieure, 45 rue d'Ulm, 75005 Paris, France(vincent.calvez@ens.fr, http://www.dma.ens.fr/∼vcalvez/).

§ICREA and Departament de Matemàtiques, Universitat Autònoma de Barcelona, E-08193 Bellaterra, Spain (carrillo@mat.uab.es, http://kinetic.mat.uab.es/∼carrillo/). This author's research was supported by DGI-MEC (Spain) project MTM2005-08024, and partially supported by the Acc. Integ./Picasso program HF2006-0198.

$$
\text{(1.1)}\qquad \begin{cases} \dfrac{\partial n}{\partial t}(t,x) = \Delta n(t,x) - \chi \nabla \cdot [n(t,x) \nabla c(t,x)]\,, & t>0\,,\ x \in \mathbb{R}^d, \\ c(t,x) = -\dfrac{1}{d\pi} \displaystyle\int_{\mathbb{R}^d} \log|x-y|\, n(t,y)\, \mathrm{d}y, & t>0\,,\ x \in \mathbb{R}^d, \\ n(0,x) = n_0 \geq 0, & x \in \mathbb{R}^d\,. \end{cases}
$$

Here $(t,x) \mapsto n(t,x)$ represents the cell density, and $(t,x) \mapsto c(t,x)$ is the concentration of chemo-attractant. The constant $\chi > 0$ is the *sensitivity* of the bacteria to the chemo-attractant. Mathematically, it measures the interaction force between cells and hence the strength of the nonlinear coupling. Note that, only in dimension two, the logarithmic kernel is the Poisson kernel, and the system (1.1) corresponds to the variant of the PKS model in the simplified version introduced by Jäger and Luckhaus in [31]. In other dimensions the log kernel has no physical interpretation. However, as we will see below, this is a good "toy model" for studying the competition blow-up phenomenon versus global existence. Initial data are assumed to verify

$$
\text{(1.2)}\qquad (1+|x|^2)\, n_0 \in L^1_+(\mathbb{R}^d) \quad \text{and} \quad n_0 \log n_0 \in L^1(\mathbb{R}^d)\,.
$$

The solutions satisfy the formal conservation of the total mass of the system

$$
\int_{\mathbb{R}^d} n_0(x)\, \mathrm{d}x = \int_{\mathbb{R}^d} n(t,x)\, \mathrm{d}x.
$$

Without loss of generality we assume that the total mass is 1 such that all the parameters of the system are contained in the reduced parameter χ. The center of mass is also conserved as time evolves, and thus we fix it to be zero for the sake of simplicity,

$$
\int_{\mathbb{R}^d} x\, n(t,x)\, \mathrm{d}x = \int_{\mathbb{R}^d} x\, n_0(x)\, \mathrm{d}x = 0.
$$

We first recall that a notion of weak solution n in the space $C^0([0,T); L^1_{\text{weak}}(\mathbb{R}^d))$, with fixed $T > 0$, using the symmetry in x, y for the concentration gradient, was introduced in [42] able to handle measure solutions; see also [41] for an alternative theory. We shall say that n is a weak solution to the system (1.1) if for all test functions $\zeta \in \mathcal{D}(\mathbb{R}^d)$,

$$
\text{(1.3)}\qquad \begin{aligned} \frac{\mathrm{d}}{\mathrm{dt}} \int_{\mathbb{R}^d} \zeta(x)\, n(t,x)\, \mathrm{d}x &= \int_{\mathbb{R}^d} \Delta\zeta(x)\, n(t,x)\, \mathrm{d}x \\ &\quad - \frac{\chi}{2\,d\,\pi} \iint_{\mathbb{R}^d \times \mathbb{R}^d} [\nabla\zeta(x) - \nabla\zeta(y)] \cdot \frac{x-y}{|x-y|^2}\, n(t,s)\, n(t,y)\, \mathrm{d}x\, \mathrm{d}y \end{aligned}
$$

together with $n(t=0) = n_0$ in the distributional sense in $(0,T)$.

As proved in [12, 8, 14], this problem presents the following dichotomy: either solutions blow-up in finite time for the supercritical case $\chi > 2d^2\pi$ or rather solutions exist globally in time and spread in space decaying towards a stationary solution in rescaled variables as $t \to \infty$ in the subcritical case $\chi < 2d^2\pi$.

Global improved weak solutions have been constructed for the system (1.1) in the subcritical case, $\chi < 2d^2\pi$, for $d = 2$ [24, 8] and $d \neq 2$ in [14]; and in the critical case for $d = 2$ in [7]. Very recently, in [25], Dolbeault and Schmeiser investigated the supercritical case in dimension two. Accordingly, the regularization of the logarithmic

kernel produces a defect measure when passing to the limit, accounting for a blow-up phenomenon. In the subcritical case, the proof of global existence of the improved weak solutions relies on the decreasing character of a free energy functional for the PKS equation given by

$$t \mapsto \mathcal{F}[n](t) := \mathcal{S}[n](t) + \mathcal{W}[n](t), \tag{1.4}$$

where $\mathcal{S}[n]$ is the standard Boltzmann's entropy and $\mathcal{W}[n]$ is the interaction energy defined by

$$\mathcal{S}[n](t) := \int_{\mathbb{R}^d} n(t,x) \log n(t,x)\, \mathrm{d}x$$

$$\text{and} \quad \mathcal{W}[n](t) := \frac{\chi}{2d\pi} \iint_{\mathbb{R}^d} n(t,x)\, n(t,y)\, \log|x-y|\, \mathrm{d}x\, \mathrm{d}y.$$

The free energy $\mathcal{F}[n]$ is related to its time derivative, the corresponding Fisher information, in the following way: consider a nonnegative solution $n \in \mathcal{C}^0([0,T), L^1(\mathbb{R}^d))$ of the PKS system (1.1) such that $n(1+|x|^2)$, $n \log n$ are bounded in $L^\infty((0,T), L^1(\mathbb{R}^d))$, $\nabla\sqrt{n} \in L^1((0,T), L^2(\mathbb{R}^d))$, and $\nabla c \in L^\infty((0,T)\times\mathbb{R}^d)$; then

$$\frac{\mathrm{d}}{\mathrm{dt}}\mathcal{F}[n](t) = -\int_{\mathbb{R}^d} n(t,x)\, |\nabla \log n(t,x) - \chi \nabla c(t,x)|^2\, \mathrm{d}x. \tag{1.5}$$

The functional $\mathcal{F}$ structurally belongs to the general class of free energies for interacting particles introduced in [36, 18, 19] and further analyzed in [22, 2]. The functionals treated in those references are of the general form

$$\mathcal{E}[n] := \int_{\mathbb{R}^d} U[n(x)]\, \mathrm{d}x + \int_{\mathbb{R}^d} n(x)\, V(x)\, \mathrm{d}x + \frac{1}{2} \iint_{\mathbb{R}^d\times\mathbb{R}^d} W(x-y)\, n(x)\, n(y)\, \mathrm{d}x\, \mathrm{d}y \tag{1.6}$$

under the basic assumptions that $U : \mathbb{R}^+ \to \mathbb{R}$ is a density of internal energy, $V : \mathbb{R}^d \to \mathbb{R}$ is a convex confinement potential, and $W : \mathbb{R}^d \to \mathbb{R}$ is a symmetric convex interaction potential. The internal energy U should satisfy the following dilation condition, introduced in McCann [36]:

$$\lambda \longmapsto \lambda^d U(\lambda^{-d}) \qquad \text{is convex nonincreasing on } \mathbb{R}^+. \tag{1.7}$$

The most important case of application, as it is for our case, is $U(s) = s \log s$, which identifies the internal energy with Boltzmann's entropy.

Continuity equations where the velocity field is formally derived from the variational derivative of free energy functionals of the type (1.6), given by

$$\frac{\partial \rho}{\partial t} = \operatorname{div}\left(\rho \nabla \frac{\delta \mathcal{E}}{\delta \rho}\right) \qquad \text{in } (0,+\infty)\times\mathbb{R}^d, \tag{1.8}$$

appear in various contexts: the interest for a convex interaction potential energy arose from its use in the modeling of granular flows; see the works of Benedetto and coworkers [5, 6] and of Toscani [44] and Villani [49] and the references therein for the physical background and related mathematical analysis. Nice mathematical and physical reviews are provided in [48, Chapter 5] and [49].

A very powerful theory has been developed in the past decade starting from the seminal paper by McCann [36] where the notion of displacement convexity for

a functional acting on probability measures was introduced. This notion provides functionals of the form (1.6) with a natural convexity structure. However, the interacting kernel W is itself required to be convex. Later, Otto [39] introduced a formal Riemannian structure giving sense to this family of equations (1.8) as the gradient flow of the convex free-energy functional (1.8) with respect to a metric that induces the Euclidean Wasserstein distance for measures. Geodesics in Otto's interpretation correspond to optimal transportation pathways (or *displacement interpolation*),

$$\rho^t = \Big((1-t)\mathrm{Id} + t\nabla\varphi\Big)\#\rho^0,$$

where $T = \nabla\varphi$ is the optimal static transport map between the endpoints ρ^0 and ρ^1.

On the other hand, a steepest descent scheme based on optimal transport of measures was introduced in [32] for the linear Fokker–Planck equation, exhibiting very nice properties. This scheme is now well understood and has been formalized for a large class of degenerate parabolic equations in [1], and in a more abstract setting, by Ambrosio, Gigli, and Savaré [3] with the name of "minimising movement scheme." The idea corresponds to a discrete version of the gradient flow or steepest descent of the free energy under the Wasserstein metric structure; see section 2 below for precise definitions.

In our case, the free-energy functional shows a nonconvex interaction potential, characteristic also of other models in mathematical biology [9, 12] and swarming [43]. To weaken the convexity assumption on the interaction kernel and to find under which conditions stationary states continue to be global attractors of the dynamics are issues of great interest for applications in mathematical biology.

The main results of this work, Theorem 3.4 and Proposition 4.3, show the convergence of the Jordan–Kinderlehrer–Otto steepest descent discrete method [32] using Otto's interpretation of the PKS equation (1.1) as the gradient flow of the free-energy functional for the subcritical case and the exponential convergence towards a unique stationary profile in scaled variables for the subcritical one-dimensional case. The first result recovers the available global existence results in the subcritical cases for the PKS equation in [12, 8, 14]. Moreover, we numerically solve this scheme in the one-dimensional case showing its abilities to capture the blow-up for supercritical cases without the need of mesh-refinement.

The plan of this paper is the following: we first recall in section 2.1 some recent results on free energies and rescaled variables, which allow us to obtain a priori estimates. We recall in section 2.2 notions on optimal transport and on the Wasserstein distance that we will use in section 3 to prove the convergence of the scheme (3.1). The exponential convergence towards a unique equilibrium is shown in the scaled one-dimensional setting in section 4.1. Finally, one-dimensional numerical simulations are given in section 4.2.

2. Preliminaries.

2.1. A priori estimates in the subcritical case. Here, we review some aspects of the PKS model that were already used in [12, 8, 14, 7] as the main tools for the proof of global existence of weak solutions in the subcritical and critical cases.

We will make a fundamental use of the logarithmic Hardy–Littlewood–Sobolev inequality [4, 15]: let f be a nonnegative function in $L^1(\mathbb{R}^d)$ such that $f\log f$ and $f\log(1+|x|^2)$ belong to $L^1(\mathbb{R}^d)$. If

$$\int_{\mathbb{R}^d} f(x)\,\mathrm{d}x = 1,$$

then

$$\int_{\mathbb{R}^d} f(x)\log f(x)\,\mathrm{d}x + d\iint_{\mathbb{R}^d\times\mathbb{R}^d} f(x)f(y)\log|x-y|\,\mathrm{d}x\,\mathrm{d}y \geq -C(d) \tag{2.1}$$

with $C(d) := (1/2)\log\pi + (1/d)\log[\Gamma(d/2)/\Gamma(d)] + (1/2)[\psi(d) - \psi(d/2)]$, where ψ is the logarithmic derivative of the Γ-function. The logarithmic Hardy–Littlewood–Sobolev inequality (2.1) implies that the functional energy (1.4) is bounded from below if $\chi = \chi_c := 2\,d^2\pi$.

Since we will work mainly in the subcritical case $\chi < \chi_c$, it is clearer, although not necessary, to solve the equation in rescaled variables. Let us define the rescaled functions ρ and v by

$$n(t,x) = \frac{1}{R^d(t)}\,\rho\left(\tau(t), \frac{x}{R(t)}\right) \quad \text{and} \quad c(x,t) = v\left(\tau(t), \frac{x}{R(t)}\right) \tag{2.2}$$

with

$$R(t) = \sqrt{1+2t} \quad \text{and} \quad \tau(t) = \log R(t).$$

The rescaled system is

$$\begin{cases} \dfrac{\partial\rho}{\partial t}(t,x) = \Delta\rho(t,x) + \nabla\cdot\{\rho(t,x)\,[x - \chi\nabla v(t,x)]\}\,, & t>0\,,\ x\in\mathbb{R}^d\,, \\ v(t,x) = -\dfrac{1}{d\,\pi}\,\log|\cdot| * \rho(t,x) - \dfrac{1}{d\,\pi}\,\tau(t), & t>0\,,\ x\in\mathbb{R}^d\,, \\ \rho(0,x) = \rho^0 = n_0 \geq 0, & x\in\mathbb{R}^d\,. \end{cases} \tag{2.3}$$

In the rescaled variables, the confinement potential $V(x) = \frac{1}{2}|x|^2$ is added, and the free energy becomes

$$\begin{aligned} \mathcal{G}[\rho] = &\int_{\mathbb{R}^d} \rho(x)\log\rho(x)\,\mathrm{d}x + \frac{1}{2}\int_{\mathbb{R}^d} |x|^2\,\rho(x)\,\mathrm{d}x \\ &+ \frac{\chi}{2\,d\,\pi}\iint_{\mathbb{R}^d\times\mathbb{R}^d} \log|x-y|\,\rho(x)\,\rho(y)\,\mathrm{d}x\,\mathrm{d}y. \end{aligned} \tag{2.4}$$

With the definition (1.3) we shall say that ρ is a weak solution to the system (2.3) if for all test functions $\zeta\in\mathcal{D}(\mathbb{R}^d)$,

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathbb{R}^d}\zeta(x)\,\rho(t,x)\,\mathrm{d}x = &\int_{\mathbb{R}^d}\Delta\zeta(x)\,\rho(t,x)\,\mathrm{d}x - \int_{\mathbb{R}^d}\nabla\zeta(x)\cdot x\,\rho(t,x)\,\mathrm{d}x \\ &- \frac{\chi}{2\,d\,\pi}\iint_{\mathbb{R}^d\times\mathbb{R}^d}[\nabla\zeta(x) - \nabla\zeta(y)]\cdot\frac{x-y}{|x-y|^2}\,\rho(t,x)\,\rho(t,y)\,\mathrm{d}x\,\mathrm{d}y, \end{aligned} \tag{2.5}$$

together with $\rho(t=0) = \rho^0$ in the distributional sense in $(0,T)$. The following lemma extracts enough information from this decreasing free energy to proceed.

LEMMA 2.1 (a priori estimates). *The functional $\mathcal{G}$ is bounded from below on the set*

$$\mathcal{K} := \left\{\rho\in L^1_+(\mathbb{R}^d)\ :\ \int_{\mathbb{R}^d}\rho(t,x) = 1,\ |x|^2\,\rho\in L^1(\mathbb{R}^d),\ \int_{\mathbb{R}^d}\rho(t,x)\,|\log\rho(t,x)|\,\mathrm{d}x < \infty\right\}$$

if and only if $\chi\leq\chi_c$. In addition, if $\chi<\chi_c$, we have on every subset $\{\mathcal{G}\leq C\}$ the following:

(i) *no concentration:* $\int_{\mathbb{R}^d} \rho|\log\rho| \le C$,
(ii) *mass confinement:* $\int_{\mathbb{R}^d} |x|^2\rho \le C$.
As a consequence, every level subset $\{\mathcal{G} \le C\}$ *is equi-integrable.*

Proof. The first use of the logarithmic Hardy–Littlewood–Sobolev inequality (2.1) to bound the free energy $\mathcal{G}$ from below is due to [24]. Rewrite

(2.6)
$$\mathcal{G}[\rho](t) = (1-\theta)\int_{\mathbb{R}^d} \rho(t,x)\,\log\rho(t,x)\,\mathrm{d}x + \frac{1}{2}\int_{\mathbb{R}^d} |x|^2\,\rho(t,x)\,\mathrm{d}x$$
$$+\,\theta d\left[\frac{1}{d}\int_{\mathbb{R}^d} \rho(t,x)\,\log\rho(t,x)\,\mathrm{d}x + \frac{\chi}{2\,d^2\,\pi\,\theta}\iint_{\mathbb{R}^d\times\mathbb{R}^d} \rho(t,x)\,\rho(t,y)\,\log|x-y|\,\mathrm{d}x\,\mathrm{d}y\right].$$

The logarithmic Hardy–Littlewood–Sobolev inequality (2.1) controls the third term if we choose $\theta = \chi/\chi_c$. Because the function $\rho\log\rho$ is negative for small ρ, we need to somehow control the density for large x. We use in fact the confinement potential, i.e., the second momentum of ρ.

LEMMA 2.2 (Carleman's estimates). *For any probability density* $u \in L^1_+(\mathbb{R}^d)$*, if the second moment* $\int_{\mathbb{R}^d} |x|^2\,u(x)\,\mathrm{d}x$ *and the entropy* $\int_{\mathbb{R}^d} u\log u\,\mathrm{d}x$ *are bounded from above, then* $u\log u$ *is uniformly bounded in* $L^1(\mathbb{R}^d)$*, and we have*

$$\int_{\mathbb{R}^d} u(x)\,|\log u(x)|\,\mathrm{d}x \le \int_{\mathbb{R}^d} u(x)\,\left(\log u(x) + \frac{1}{2}|x|^2\right)\,\mathrm{d}x + d\log(4\pi) + \frac{2}{e}\,.$$

Proof. The proof goes as follows. Let $\bar u := u\,\mathbb{1}_{\{u\le 1\}}$ and $m = \int_{\mathbb{R}^d} \bar u(x)\,\mathrm{d}x \le \int_{\mathbb{R}^d} u(x)\,\mathrm{d}x = 1$. Then

$$\int_{\mathbb{R}^d} \bar u(x)\,\left(\log\bar u(x) + \frac{1}{4}|x|^2\right)\,\mathrm{d}x = \int_{\mathbb{R}^d} U(x)\log U(x)\,\mathrm{d}\mu - m\frac{d}{2}\log(4\pi)\,,$$

where $U := \bar u/\mu$, $d\mu(x) = (4\pi)^{-d/2}e^{-|x|^2/4}\,\mathrm{d}x$. The Jensen inequality yields

$$\int_{\mathbb{R}^d} U(x)\,\log U(x)\,\mathrm{d}\mu \ge \left(\int_{\mathbb{R}^d} U(x)\,\mathrm{d}\mu\right)\,\log\left(\int_{\mathbb{R}^d} U(x)\,\mathrm{d}\mu\right) = m\,\log m$$

and

$$\int_{\mathbb{R}^d} \bar u(x)\,\log\bar u(x)\,\mathrm{d}x \ge m\log m - m\frac{d}{2}\log 4\pi - \frac{1}{4}\int_{\mathbb{R}^d} |x|^2\,\bar u(x)\,\mathrm{d}x$$
$$\ge -\frac{1}{e} - \frac{d}{2}\log(4\pi) - \frac{1}{4}\int_{\mathbb{R}^d} |x|^2\,u(x)\,\mathrm{d}x.$$

Using

$$\int_{\mathbb{R}^d} u(x)\,|\log u(x)|\,\mathrm{d}x = \int_{\mathbb{R}^d} u(x)\,\log u(x)\,\mathrm{d}x - 2\int_{\mathbb{R}^d} \bar u(x)\,\log\bar u(x)\,\mathrm{d}x,$$

this completes the proof. ∎

We apply this lemma to obtain the first part of the result from (2.6):

$$\mathcal{G}[\rho](t) \ge (1-\theta)\int_{\mathbb{R}^d} \rho(t,x)\,|\log\rho(t,x)|\,\mathrm{d}x + \frac{\theta}{2}\int_{\mathbb{R}^d} |x|^2\,\rho(t,x)\,\mathrm{d}x + C.$$

On the other hand, the functional $\mathcal{G}[\rho]$ has an interesting scaling property. For a given ρ, let $\rho_\lambda(x) = \lambda^d \rho(\lambda x)$. It is straightforward to check that $\|\rho_\lambda\|_{L^1(\mathbb{R}^d)} = 1$ and

$$\mathcal{G}[\rho_\lambda] = \mathcal{G}[\rho] + d\left(1 - \frac{\chi}{\chi_c}\right)\log\lambda + \frac{\lambda^{-2}-1}{2}\int_{\mathbb{R}^d} |x|^2\,\rho\,\mathrm{d}x.$$

Since $\lambda \mapsto \mathcal{G}[\rho_\lambda]$ is clearly not bounded from below if $\chi > \chi_c$, the proof concludes here. ∎

We shall also state another technical lemma, which will play a major role when passing to the limit in the quadratic interaction contribution.

LEMMA 2.3 (doubling of variables). *Assume $f_i \rightharpoonup f$ in L^1 and that the family $\{f_i\}$ is equi-integrable in the sense of Lemma 2.1; then $f_i \otimes f_i \rightharpoonup f \otimes f$ in $L^1(\mathbb{R}^d \times \mathbb{R}^d)$.*

Proof. Let $\psi(x,y)$ denote any test function in $L^\infty(\mathbb{R}^d \times \mathbb{R}^d)$. For almost every $x \in \mathbb{R}^d$ define

$$\lim \Psi^i(x) := \lim \int_{\mathbb{R}^d} f_i(y)\,\psi(x,y)\,\mathrm{d}y = \int_{\mathbb{R}^d} f(y)\,\psi(x,y)\,\mathrm{d}y =: \Psi(x).$$

Note that for any x, $|\Psi^i(x)|$ and $|\Psi(x)|$ are bounded by $\|\psi\|_{L^\infty}$.

By Egorov's theorem for any $R > 0$ and $\delta > 0$, there exists a measurable set X_δ such that $|X_\delta| < \delta$ and Ψ^i uniformly converges to Ψ in $B_R \setminus X_\delta$. We have

$$\begin{aligned}
&\left|\int_{\mathbb{R}^d} \left[\Psi^i(x)\,f_i(x) - \Psi(x)\,f(x)\right]\,\mathrm{d}x\right| \\
&\quad\le \int_{B_R\setminus X_\delta} \left|\Psi^i(x)\,f_i(x) - \Psi(x)\,f(x)\right|\,\mathrm{d}x + \int_{X_\delta} \left|\Psi^i(x)\,f_i(x) - \Psi(x)\,f(x)\right|\,\mathrm{d}x \\
&\qquad\quad + \int_{B_R^c} \left|\Psi^i(x)\,f_i(x) - \Psi(x)\,f(x)\right|\,\mathrm{d}x \\
&\quad\le \int_{B_R\setminus X_\delta} \left|\Psi^i(x)\,f_i(x) - \Psi(x)\,f(x)\right|\,\mathrm{d}x + \|\psi\| \int_{X_\delta} f_i(x) + f(x)\,\mathrm{d}x \\
&\qquad\quad + \|\psi\|\,\frac{1}{R^2}\int_{B_R^c} |x|^2\,[f_i(x) + f(x)]\,\mathrm{d}x.
\end{aligned}$$

Egorov's theorem and the weak-L^1 convergence of f_i towards f ensures that the first term is as small as desired by choosing i large enough. By the a priori estimates in Lemma 2.1, $\int_{X_\delta} f_i(x)\,\mathrm{d}x$ and $\int_{X_\delta} f(x)\,\mathrm{d}x$ can be made as small as desired by choosing δ small enough; further, the third term can be made as small as desired by choosing R large enough. ∎

2.2. Optimal transport and the Wasserstein distance. We recall some standard results related to optimal transportation and Wasserstein distance that we will use in the remainder of this paper. The interested reader can refer to the books of Villani [48, 50] and the book of Ambrosio, Gigli, and Savaré [3]. A short summary of properties of the Wasserstein distance can be seen in [21].

Let μ and ν be in $\mathcal{P}(\mathbb{R}^d)$ the space of probability measure in $\mathbb{R}^d$, $\mathcal{P}_2(\mathbb{R}^d)$ the subset of probability measures with finite second-momentum, $\mathcal{P}_2^{\mathrm{ac}}(\mathbb{R}^d)$ its subset formed by the absolutely continuous measures with respect to Lebesgue, and T be a measurable map $\mathbb{R}^d \to \mathbb{R}^d$. We say that T *transports* μ onto ν, and we note $\nu = T\#\mu$ if for any measurable set $B \subset \mathbb{R}^d$, $\nu(B) = \mu \circ T^{-1}(B)$. We also say ν is the *push-forward* or *the*

image measure of μ by T; i.e.,

$$\int_{\mathbb{R}^d} \zeta \circ T(x)\, \mathrm{d}\mu(x) = \int_{\mathbb{R}^d} \zeta(y)\, \mathrm{d}\nu(y) \quad \forall \zeta \in \mathcal{C}_b^0(\mathbb{R}^d). \tag{2.7}$$

The Wasserstein distance between μ and ν, d_W can be defined by

$$d_W^2(\mu,\nu) := \inf_{T:\ \nu = T\#\mu} \int_{\mathbb{R}^d} |x - T(x)|^2\, \mathrm{d}\mu(x).$$

By Brenier's theorem [10, 35, 37] (see [48, Theorem 2.32, p. 85] for a review), if μ is absolutely continuous with respect to Lebesgue measure, then there is one measurable plan T such that $\nu = T\#\mu$ and $T = \nabla\varphi$ for some convex function φ. As a consequence,

$$d_W^2(\mu,\nu) = \int_{\mathbb{R}^d} |x - \nabla\varphi(x)|^2\, \mathrm{d}\mu(x). \tag{2.8}$$

The variational problem leading to the definition of the Wasserstein distance can be relaxed to the linear program

$$d_W^2(\mu,\nu) = \inf_{\Pi\in\Gamma} \left\{ \int_{\mathbb{R}^d\times\mathbb{R}^d} |x-y|^2\, \mathrm{d}\Pi(x,y) \right\},$$

where Π runs over the set of transference plans Γ, that is, the set of joint probability measures on $\mathbb{R}^d\times\mathbb{R}^d$ with marginals μ and ν. In fact, the infimum above is a minimum by Kantorovich duality theorems [48, Chapter 1]. The optimal transference plan, in case Brenier's theorem applies, is given by $\Pi_o = (\mathrm{id}_{\mathbb{R}^d} \otimes \nabla\varphi)\#\mu$.

Let us recall a simple consequence of the definition of the Wasserstein distance for controlling averages [21, Corollary 2.4].

LEMMA 2.4 (convergence of averages with d_W). *Given ζ a Lipschitz function with Lipschitz constant L and $\mu,\nu \in \mathcal{P}_2(\mathbb{R}^d)$, then we have*

$$\left| \int_{\mathbb{R}^d} \zeta(x)\, \mathrm{d}\mu - \int_{\mathbb{R}^d} \zeta(x)\, \mathrm{d}\nu \right| \le L\, d_W(\mu,\nu).$$

Proof. Let Π_o be the optimal plan between μ and $\nu \in \mathcal{P}_2(\mathbb{R}^d)$ for d_W. Then

$$\int_{\mathbb{R}^d\times\mathbb{R}^d} |x-y|^2\, \mathrm{d}\Pi_o(x,y) = d_W^2(\mu,\nu),$$

and we can write

$$\int_{\mathbb{R}^d} \zeta(x)\, \mathrm{d}\mu - \int_{\mathbb{R}^d} \zeta(x)\, \mathrm{d}\nu = \int_{\mathbb{R}^d\times\mathbb{R}^d} (\zeta(x) - \zeta(y))\ \mathrm{d}\Pi_o(x,y).$$

Using that ζ is Lipschitz with constant L and estimating by Hölder's inequality, we get

$$\begin{aligned} \left| \int_{\mathbb{R}^d} \zeta(x)\, \mathrm{d}\mu - \int_{\mathbb{R}^d} \zeta(x)\, \mathrm{d}\nu \right| &\le \int_{\mathbb{R}^d\times\mathbb{R}^d} |\zeta(x) - \zeta(y)|\ \mathrm{d}\Pi_o(x,y) \\ &\le L \int_{\mathbb{R}^d\times\mathbb{R}^d} |x-y|\ \mathrm{d}\Pi_o(x,y) \le L\, d_W(\mu,\nu), \end{aligned}$$

giving the assertion. ∎

3. Time discretization. We consider a time-step $\tau > 0$ and an initial datum $\rho^0 \in \mathcal{P}_2^{\mathrm{ac}}(\mathbb{R}^d)$. We introduce the sequence $(\rho_\tau^n)_{n\in\mathbb{N}}$ recursively defined by $\rho_\tau^0 = \rho^0$ and

$$\rho_\tau^{n+1} \in \arg\inf_{\rho\in\mathcal{K}} \left\{ \mathcal{G}[\rho] + \frac{1}{2\,\tau} d_W^2(\rho_\tau^n, \rho) \right\} . \tag{3.1}$$

The Jordan–Kinderlehrer–Otto (JKO) steepest descent scheme can be viewed formally as a time discretization of the abstract gradient flow equation

$$\frac{\partial \rho}{\partial t} = -\text{“}\nabla_W\text{”}\mathcal{G}[\rho] ,$$

where the space $\mathcal{K} \subset \mathcal{P}_2^{\mathrm{ac}}(\mathbb{R}^d)$ is endowed with a formal Riemannian structure compatible with the Wasserstein d_W distance [39]. We refer to [39, 48, 3, 19] for a deeper discussion and rigorous sense of the “∇_W” definition. The next lemma ensures that this discrete scheme is well defined.

LEMMA 3.1 (existence of minimizers). *Let ρ_0 satisfy* (1.2) *and $\chi < \chi_c$; then there recursively exists a minimizer to* (3.1).

Proof. Introduce the function

$$\mathcal{K} \ni \xi \mapsto \mathcal{G}[\xi] + \frac{1}{2\,\tau} d_W^2(\rho_\tau^n, \xi) . \tag{3.2}$$

By the a priori estimates in Lemma 2.1, this function is bounded from below. Consider $(\xi_k)_{k\in\mathbb{N}}$ a minimizing sequence; without loss of generality, we can assume that it satisfies $\mathcal{G}[\xi_k] \le \mathcal{G}[\rho_\tau^n]$ for all $k \in \mathbb{N}$. Proceeding as in Lemma 2.1, we get

$$(1-\theta) \int_{\mathbb{R}^d} \xi_k(x)\, |\log \xi_k(x)|\, \mathrm{d}x + \frac{\theta}{2} \int |x|^2 \xi_k(x) \le \mathcal{G}[\rho_\tau^n] + \theta\, C(d).$$

If $\chi < \chi_c$, then $\theta = \frac{\chi}{\chi_c} < 1$, and this shows that $\xi_k \log \xi_k$ is bounded in $L^1(\mathbb{R}^d)$.

The bound on the second-momentum avoids vanishing, while the $L^1(\mathbb{R}^d)$-bound on $\xi_k \log \xi_k$ avoids concentration: indeed,

$$\int_{\{\xi_k \ge Q\}} \xi_k \, \mathrm{d}x \le \frac{1}{\log Q} \int_{\{\xi_k \ge Q\}} \xi_k \log \xi_k \, \mathrm{d}x \le \frac{1}{\log Q} \int_{\mathbb{R}^d} \xi_k \, |\log \xi_k| \, \mathrm{d}x$$

can be made as small as desired for $Q > 1$ large enough. Hence the family $\{\xi_k\}_{k\in\mathbb{N}}$ verifies the hypotheses in the Dunford–Pettis theorem, and thus there exists a subsequence still denoted $(\xi_k)_{k\in\mathbb{N}}$ which converges weakly L^1 to a density ξ_*.

It remains to prove that this candidate ξ_* realizes in fact a minimum of (3.2). The weak-L^1 lower semicontinuity of the entropy $\mathcal{S}$, the second-momentum, and the Wasserstein distance are well known; see [32, 3] and the references therein. We will prove that the quadratic interaction term is continuous for the weak-L^1 convergence in our situation. We split it into

$$\iint_{\mathbb{R}^d\times\mathbb{R}^d} \xi_k(x)\, \xi_k(y)\, \log|x-y|\, \mathrm{d}x\, \mathrm{d}y = \mathrm{A} + \mathrm{B} + \mathrm{C},$$

with

$$\mathrm{A} := \iint_{|x-y|<\varepsilon} \xi_k(x)\,\xi_k(y)\,\log|x-y|\,\mathrm{d}x\,\mathrm{d}y,$$

$$\mathrm{B} := \iint_{\varepsilon\le|x-y|\le R} \xi_k(x)\,\xi_k(y)\,\log|x-y|\,\mathrm{d}x\,\mathrm{d}y$$

$$\text{and}\quad \mathrm{C} := \iint_{|x-y|>R} \xi_k(x)\,\xi_k(y)\,\log|x-y|\,\mathrm{d}x\,\mathrm{d}y,$$

where $\varepsilon < 1$ and $R > \sqrt{e}$.

Control of $\mathrm{A}(t)$. We use the duality inequality, $ab \le e^a + b\log b - b$:

$$\begin{aligned}
|\mathrm{A}| &\le \iint_{|x-y|<\varepsilon} \xi_k(x)\xi_k(y)\log\frac{1}{|x-y|}\,\mathrm{d}x\,\mathrm{d}y\\
&\le \int\left\{\xi_k(x)\int_{|x-y|<\varepsilon}\alpha^{-1}\xi_k(y)\log(\alpha^{-1}\xi_k(y)) - \alpha^{-1}\xi_k(y) + \exp\left(\alpha\log\frac{1}{|x-y|}\right)\right\}\\
&\le \alpha^{-1}\int \xi_k(y)\log(\alpha^{-1}\xi_k(y))\int_{|x-y|<\varepsilon}\xi_k(x)\,\mathrm{d}x\,\mathrm{d}y + \int\xi_k(x)\,\mathrm{d}x\int_{|z|<\varepsilon}\frac{1}{|z|^\alpha}\,\mathrm{d}z.
\end{aligned}$$

By the $L^1(\mathbb{R}^d)$-bound on $\xi_k\log\xi_k$, $\int_{X_\varepsilon}\xi_k$ is uniformly small on small sets X_ε, and therefore the last term can be made as small as desired uniformly in k by choosing ε small enough and $\alpha < d$.

Control of B. We shall use Lemma 2.3 because $\log|x-y|$ is bounded on the set $\{\varepsilon < |x-y| < R\}$.

Control of C. For $R > \sqrt{e}$, $R \mapsto R^2/\log R$ is an increasing function, so that

$$0 \le \mathrm{C} \le \frac{2\log R}{R^2}\int_{\mathbb{R}^2}|x|^2\,\xi_k(t,x)\,\mathrm{d}x.$$

Because we uniformly bound the second-momentum, this can be made as small as desired while choosing R large enough.

Finally, collecting terms, we get that the difference

$$\left|\iint_{\mathbb{R}^d\times\mathbb{R}^d}\xi_k(x)\,\xi_k(y)\log|x-y|\,\mathrm{d}x\,\mathrm{d}y - \iint_{\mathbb{R}^d\times\mathbb{R}^d}\xi_*(x)\,\xi_*(y)\log|x-y|\,\mathrm{d}x\,\mathrm{d}y\right|$$

can be made as small as desired by choosing ε and δ small enough and r, R, and k large enough. ∎

Remark 3.2 (uniqueness of minimizers). Since the functional $\mathcal{G}[n]$ is not convex, we cannot conclude the uniqueness of minimizers for the discrete scheme, and thus the scheme (3.1) is defined by choosing any element realizing the infimum as ρ_τ^{n+1}. Each choice might in principle give rise to a solution in the limit $\tau \to 0$. It is an open problem to deal with the uniqueness of solutions in the subcritical case.

Now, we define the time interpolation of the discrete scheme as a family of Lipschitz curves $(\rho_\tau)_{\tau>0}$ connecting every pair $\{\rho_\tau^n, \rho_\tau^{n+1}\}$ with a constant speed geodesic in the interval $[n\,\tau, (n+1)\,\tau)$. Accordingly for any $t \in [n\,\tau, (n+1)\,\tau)$ we have

$$d_W\left(\rho_\tau^n, \rho_\tau(t)\right) = \frac{t-n\tau}{\tau}\,d_W\left(\rho_\tau^n, \rho_\tau^{n+1}\right).$$

Obviously $\rho_\tau(n\tau) = \rho_\tau^n$. This is possible due to the Brenier theorem by defining the displacement interpolant

$$\rho_\tau(t) = \left(\frac{(n+1)\tau - t}{\tau} \mathrm{Id} + \frac{t - n\tau}{\tau} \nabla\varphi^n \right) \# \rho_\tau^n$$

with $\nabla\varphi^n$ being the optimal map transporting ρ_τ^n onto ρ_τ^{n+1}.

Remark 3.3 (comparison to literature). Let us point out that once the free energy $\mathcal{G}[n]$ is bounded from below (Lemma 2.1) and the approximation scheme is well defined (Lemma 3.1), then [3, Theorem 11.1.6, pp. 288–289] applies. For the convenience of the reader we give a shorter proof adapted to our problem. Our proof is based on the founding idea of [32] proving the convergence of the ad hoc scheme for the linear Fokker–Planck equation. A nice sketch of the proof of [32] can be found in [48, section 8.4, pp. 256–262].

THEOREM 3.4 (convergence of the scheme as $\tau \to 0$). *Under assumption* (1.2)*, if* $\chi < \chi_c$*, then the family* $(\rho_\tau)_{\tau>0}$ *admits a subsequence converging in* $C^0([0,T]; L^1_{\mathrm{weak}}(\mathbb{R}^d))$ *to a weak solution of* (2.5).

Proof. We proceed in three steps.

Step 1. A priori estimates in space and time: Since ρ_τ^{n+1} minimizes (3.1), we obviously have

$$\mathcal{G}[\rho_\tau^{n+1}] + \frac{1}{2\,\tau} d_W^2(\rho_\tau^n, \rho_\tau^{n+1}) \le \mathcal{G}[\rho_\tau^n] \,.$$

As a consequence we obtain an *energy estimate*

$$\sup_{n\in\mathbb{N}} \mathcal{G}[\rho_\tau^n] \le \mathcal{G}[\rho_\tau^0], \tag{3.3}$$

and a *total square estimate*

$$\frac{1}{2\,\tau} \sum_{n\in\mathbb{N}} d_W^2(\rho_\tau^n, \rho_\tau^{n+1}) \le \mathcal{G}[\rho_\tau^0] - \inf_{n\in\mathbb{N}} \mathcal{G}[\rho_\tau^n] \,. \tag{3.4}$$

The right-hand side is bounded thanks to Lemma 2.1.

From the total square estimate (3.4) we deduce a $\frac{1}{2}$-*Hölder-estimate* in time of $(\rho_\tau)_{\tau\ge0}$: indeed, for any $0 \le m \le n$,

$$d_W(\rho_\tau^m, \rho_\tau^n) \le \sqrt{2\,\tau} \sum_{k=m}^{n-1} \frac{1}{\sqrt{2\,\tau}} d_W(\rho_\tau^k, \rho_\tau^{k+1}) \le \sqrt{2\,\tau}\,\sqrt{(n-m)} \sqrt{\mathcal{G}[\rho_\tau^0] - \inf_{n\in\mathbb{N}} \mathcal{G}[\rho_\tau^n]} \,.$$

As a consequence for any $0 \le s \le t$

$$\begin{aligned}
d_W(\rho_\tau(s), \rho_\tau(t)) &\le d_W\Big(\rho_\tau(s), \rho_\tau^{\left[\frac{s}{\tau}+1\right]\tau}\Big) + d_W\Big(\rho_\tau^{\left[\frac{s}{\tau}+1\right]\tau}, \rho_\tau^{\left[\frac{t}{\tau}\right]\tau}\Big) + d_W\Big(\rho_\tau^{\left[\frac{t}{\tau}\right]\tau}, \rho_\tau(t)\Big) \\
&\le \left(\left[\frac{s}{\tau}+1\right] - \frac{s}{\tau} \right) d_W\Big(\rho_\tau^{\left[\frac{s}{\tau}\right]\tau}, \rho_\tau^{\left[\frac{s}{\tau}+1\right]\tau}\Big) \\
&\quad + \sqrt{2\,\tau \left(\left[\frac{t}{\tau}\right] - \left[\frac{s}{\tau}+1\right] \right) \Big(\mathcal{G}(\rho_\tau^0) - \inf_{n\in\mathbb{N}} \mathcal{G}(\rho_\tau^n) \Big)} \\
&\quad + \left(\frac{t}{\tau} - \left[\frac{t}{\tau}\right] \right) d_W\Big(\rho_\tau^{\left[\frac{t}{\tau}\right]\tau}, \rho_\tau^{\left[\frac{t}{\tau}+1\right]\tau}\Big) \\
&\le \sqrt{6 \left(\mathcal{G}(\rho^0) - \inf_{n\in\mathbb{N}} \mathcal{G}(\rho^n) \right)} \, (t-s)^{\frac{1}{2}} \,.
\end{aligned} \tag{3.5}$$

Step 2. Compactness: By the $\frac{1}{2}$-Hölder-estimate (3.5), ρ_τ is bounded in $\mathcal{P}_2(\mathbb{R}^d)$, so the family $\{\rho_\tau\}_{\tau>0}$ is tight. By the a priori estimates (Lemma 2.1), the family $\{\rho_\tau(t)\}_{\tau>0}$ can neither concentrate nor vanish and the family $\{\rho_\tau(t)\}_{\tau>0}$ is equi-integrable. In the other hand, by the estimate (3.5) the curves $\rho_\tau(t,\cdot)$ are $\frac{1}{2}$-Hölder continuous in time. The Ascoli–Arzelà theorems yield the relative compactness of the family $(\rho_\tau(t,\cdot))_{\tau>0}$.

Finally $\{\rho_\tau\}_{\tau>0}$ is relatively compact in $\mathcal{C}^0([0,T],L^1_{\text{weak}}(\mathbb{R}^d))$ for any $T>0$. As a consequence, for any $T>0$, there exists a subsequence still denoted $(\rho_\tau)_{\tau>0}$ such that $(\rho_\tau)_{\tau>0}$ converges in $\mathcal{C}^0([0,T],L^1_{\text{weak}}(\mathbb{R}^d))$ to a function ρ when τ goes to 0.

Step 3. Approximate Euler–Lagrange equation in weak formulation: *Weak space derivative.* Let ζ be a test function and $\nabla\zeta$ be a smooth vector field with compact support. Let us define $\nabla\varphi_\varepsilon$ with $\varphi_\varepsilon(x) := \frac{|x|^2}{2} + \varepsilon\,\zeta$. For ε small enough $\nabla\varphi_\varepsilon$ is a $\mathcal{C}^\infty$-diffeomorphism and $\operatorname{Det} D^2\varphi_\varepsilon = \operatorname{Det}(\operatorname{Id} + \varepsilon\, D^2\zeta) > 0$. We define $\overline{\rho_\varepsilon}$ to be the push-forward perturbation of ρ_τ^{n+1} by $\nabla\varphi_\varepsilon$:

$$\overline{\rho_\varepsilon} = \nabla\varphi_\varepsilon \# \rho_\tau^{n+1}\ .$$

Changing variables and using (2.7), we have

$$\mathcal{G}[\overline{\rho_\varepsilon}](t) = \int_{\mathbb{R}^d} \log\frac{\rho_\tau^{n+1}(x)}{\operatorname{Det} D^2\varphi_\varepsilon}\rho_\tau^{n+1}(x)\,\mathrm{d}x + \int_{\mathbb{R}^d}\frac{1}{2}|x-\varepsilon\,\nabla\zeta(x)|^2\rho_\tau^{n+1}(x)\,\mathrm{d}x + \iint_{\mathbb{R}^d\times\mathbb{R}^d}\frac{\chi}{2\,d\,\pi}\log|\nabla\varphi_\varepsilon(x)-\nabla\varphi_\varepsilon(y)|\,\rho_\tau^{n+1}(x)\,\rho_\tau^{n+1}(y)\,\mathrm{d}x\,\mathrm{d}y.$$

Alternatively we introduce the optimal map $\nabla\varphi^n$ which transports ρ_τ^n onto ρ_τ^{n+1}. By (2.8)

$$d_W^2(\rho_\tau^n,\rho_\tau^{n+1}) = \int_{\mathbb{R}^d}|x-\nabla\varphi^n(x)|^2\rho_\tau^n(x)\,\mathrm{d}x. \tag{3.6}$$

The map $\nabla\varphi_\varepsilon\circ\nabla\varphi^n$ transports ρ_τ^n on $\overline{\rho_\varepsilon}$. We do not know if this is the optimal map, however, by definition of the Wasserstein distance

$$d_W^2(\rho_\tau^n,\overline{\rho_\varepsilon}) \le \int_{\mathbb{R}^d}\left|x-\nabla\varphi_\varepsilon\circ\nabla\varphi^n(x)\right|^2\rho_\tau^n(x)\,\mathrm{d}x. \tag{3.7}$$

Using the minimizing property of the scheme (3.1) and combining (3.6) and (3.7), we obtain

$$\begin{aligned}0 &\le \frac{1}{2\,\tau}d_W^2(\rho_\tau^n,\overline{\rho_\varepsilon}) + \mathcal{G}[\overline{\rho_\varepsilon}] - \frac{1}{2\,\tau}d_W^2(\rho_\tau^{n+1},\rho_\tau^n) - \mathcal{G}[\rho_\tau^{n+1}]\\ &\le \frac{1}{2\,\tau}\int_{\mathbb{R}^d}\left(|x-\nabla\varphi^n(x)-\varepsilon\,\nabla\zeta\circ\nabla\varphi^n(x)|^2 - |x-\nabla\varphi^n(x)|^2\right)\rho_\tau^n(x)\,\mathrm{d}x\\ &\quad+\frac{1}{2}\int_{\mathbb{R}^d}\left(\left(|x-\varepsilon\nabla\zeta(x)|^2-|x|^2\right) - \log\left[\operatorname{Det}(\operatorname{Id}+\varepsilon\,D^2\zeta)\right]\right)\rho_\tau^{n+1}(x)\,\mathrm{d}x\\ &\quad+\frac{\chi}{2\,d\,\pi}\iint\left(\log\left|x-y+\varepsilon(\nabla\zeta(x)-\nabla\zeta(y))\right| - \log|x-y|\right)\rho_\tau^{n+1}(x)\,\rho_\tau^{n+1}(y)\,\mathrm{d}x\,\mathrm{d}y.\end{aligned}$$

Dividing by ε and letting $\varepsilon>0$ go to zero, we find

$$0 \le \frac{1}{\tau}\int_{\mathbb{R}^d}\left\langle\nabla\varphi^n(x)-x,\nabla\zeta\circ\nabla\varphi^n(x)\right\rangle\rho_\tau^n(x)\,\mathrm{d}x + \int_{\mathbb{R}^d}\left[-\Delta\zeta(x)-x\cdot\nabla\zeta(x)\right]\rho_\tau^{n+1}(x)\,\mathrm{d}x + \frac{\chi}{2\,d\,\pi}\iint_{\mathbb{R}^d\times\mathbb{R}^d}\frac{[\nabla\zeta(x)-\nabla\zeta(y)]\cdot(x-y)}{|x-y|^2}\rho_\tau^{n+1}(x)\,\rho_\tau^{n+1}(y)\,\mathrm{d}x\,\mathrm{d}y.$$

Since we can change ε into $-\varepsilon$ we have in fact

$$\frac{1}{\tau}\int_{\mathbb{R}^d}\left\langle \nabla\varphi^n(x)-x, \nabla\zeta\circ\nabla\varphi^n(x)\right\rangle \rho_\tau^n(x)\,\mathrm{d}x = \int_{\mathbb{R}^d}\left[\Delta\zeta(x)+x\cdot\nabla\zeta(x)\right]\rho_\tau^{n+1}(x)\,\mathrm{d}x$$
$$-\frac{\chi}{2\,d\,\pi}\iint_{\mathbb{R}^d\times\mathbb{R}^d}\frac{[\nabla\zeta(x)-\nabla\zeta(y)]\cdot(x-y)}{|x-y|^2}\rho_\tau^{n+1}(x)\,\rho_\tau^{n+1}(y)\,\mathrm{d}x\,\mathrm{d}y. \tag{3.8}$$

Weak time derivative. Using the Taylor's expansion

$$\zeta(\nabla\varphi^n(x))-\zeta(x)=\left\langle \nabla\varphi^n(x)-x, \nabla\zeta\circ\nabla\varphi^n(x)\right\rangle + O\Big(\left|x-\nabla\varphi^n(x)\right|^2\Big),$$

we can recast the left-hand side of (3.8) as

$$\frac{1}{\tau}\int_{\mathbb{R}^d}\left[\zeta\circ\nabla\varphi^n(x)-\zeta(x)\right]\rho_\tau^n(x)\,\mathrm{d}x + O\left(\frac{1}{\tau}\int_{\mathbb{R}^d}\left|x-\nabla\varphi^n(x)\right|^2\rho_\tau^n(t)\,\mathrm{d}x\right)$$
$$=\frac{1}{\tau}\left(\int_{\mathbb{R}^d}\zeta(x)\rho_\tau^{n+1}(x)\,\mathrm{d}x-\int_{\mathbb{R}^d}\zeta(x)\rho_\tau^n(x)\,\mathrm{d}x\right)+O\left(\frac{1}{\tau}d_W^2\left(\rho_\tau^n,\rho_\tau^{n+1}\right)\right).$$

We multiply (3.8) by τ and eventually obtain

$$\int_{\mathbb{R}^d}\zeta(x)\left[\rho_\tau^{n+1}(x)-\rho_\tau^n(x)\right]\,\mathrm{d}x + O\big(d_W^2\left(\rho_\tau^n,\rho_\tau^{n+1}\right)\big)$$
$$=\tau\int_{\mathbb{R}^d}\left[\Delta\zeta(x)+x\cdot\nabla\zeta(x)-\frac{\chi}{2\,d\,\pi}\int_{\mathbb{R}^d}\frac{[\nabla\zeta(x)-\nabla\zeta(y)]\cdot(x-y)}{|x-y|^2}\rho_\tau^{n+1}(y)\,\mathrm{d}y\right]\rho_\tau^{n+1}(x)\,\mathrm{d}x. \tag{3.9}$$

Let $0\le t_1<t_2$ be fixed times, $m=[t_1/\tau]+1$, and $n=[t_2/\tau]$. By summing (3.9) we have, thanks to the total square estimate (3.4),

$$\int_{\mathbb{R}^d}\zeta(x)\left[\rho_\tau^n(x)-\rho_\tau^m(x)\right]\,\mathrm{d}x+O(\tau)=\sum_{k=m}^{n-1}\tau\int_{\mathbb{R}^d}\left[\Delta\zeta(x)+x\cdot\nabla\zeta(x)\right]\rho_\tau^{k+1}(x)\,\mathrm{d}x$$
$$-\frac{\chi}{2\,d\,\pi}\sum_{k=m}^{n-1}\tau\iint_{\mathbb{R}^d\times\mathbb{R}^d}\frac{[\nabla\zeta(x)-\nabla\zeta(y)]\cdot(x-y)}{|x-y|^2}\rho_\tau^{k+1}(x)\rho_\tau^{k+1}(y)\,\mathrm{d}x\,\mathrm{d}y. \tag{3.10}$$

On the other hand, we can split

$$\int_{\mathbb{R}^d}\zeta(x)\left[\rho_\tau(t_2,x)-\rho_\tau(t_1,x)\right]\,\mathrm{d}x=\int_{\mathbb{R}^d}\zeta(x)\left[\rho_\tau(t_2,x)-\rho_\tau^n(x)\right]\,\mathrm{d}x$$
$$+\int_{\mathbb{R}^d}\zeta(x)\left[\rho_\tau^n(x)-\rho_\tau^m(x)\right]\,\mathrm{d}x+\int_{\mathbb{R}^d}\zeta(x)\left[\rho_\tau^m(x)-\rho_\tau(t_1,x)\right]\,\mathrm{d}x.$$

By Lemma 2.4, we control the bordering averages,

$$\int_{\mathbb{R}^d}\zeta(x)\left|\rho_\tau^m(x)-\rho_\tau(t_1,x)\right|\,\mathrm{d}x\le C\,d_W(\rho_\tau^m,\rho_\tau(t_1,x))\le C\,d_W(\rho_\tau^m,\rho_\tau^{m-1})\le O(\tau^{1/2}).$$

In addition, integrating Lemma 2.4 in time implies for all integers k, $m\le k\le n-1$,

$$\tau\int_{\mathbb{R}^d}\psi(x)\rho_\tau^{k+1}(x)\,\mathrm{d}x=\int_{k\tau}^{(k+1)\tau}\int_{\mathbb{R}^d}\psi(x)\rho_\tau(s,x)\,\mathrm{d}x\,\mathrm{d}s+O(\tau\;d_W(\rho_\tau^k,\rho_\tau^{k+1})),$$

where ψ denotes any bounded test function. Hence we can transform the discrete in time sum (3.10) into a continuous time integration. Finally, the test contributions are bounded in $L^\infty((0,T)\times\mathbb{R}^d)$ so that the bordering time integrands are negligible,

$$\int_{t_1}^{m\tau}\int_{\mathbb{R}^d}[\Delta\zeta(x)+x\cdot\nabla\zeta(x)]\,\rho_\tau(s,x)\,\mathrm{d}x\,\mathrm{d}s$$
$$-\frac{\chi}{2\,d\,\pi}\int_{t_1}^{m\tau}\iint_{\mathbb{R}^d\times\mathbb{R}^d}\frac{[\nabla\zeta(x)-\nabla\zeta(y)]\cdot(x-y)}{|x-y|^2}\rho_\tau(s,y)\rho_\tau(s,x)\,\mathrm{d}y\,\mathrm{d}x\,\mathrm{d}s=O(\tau).$$

Collecting all the terms, we end up with

(3.11)
$$\int_{\mathbb{R}^d}\zeta(x)\,[\rho_\tau(t_2,x)-\rho_\tau(t_1,x)]\,\mathrm{d}x=\int_{t_1}^{t_2}\int_{\mathbb{R}^d}[\Delta\zeta(x)+x\cdot\nabla\zeta(x)]\,\rho_\tau(s,x)\,\mathrm{d}x\,\mathrm{d}s$$
$$-\frac{\chi}{2\,d\,\pi}\int_{t_1}^{t_2}\iint_{\mathbb{R}^d\times\mathbb{R}^d}\frac{[\nabla\zeta(x)-\nabla\zeta(y)]\cdot(x-y)}{|x-y|^2}\rho_\tau(s,y)\,\rho_\tau(s,x)\,\mathrm{d}y\,\mathrm{d}x\,\mathrm{d}s+O(\tau^{1/2}).$$

Step 4. Passing to the limit: The relative compactness of the family of curves $\{\rho_\tau\}_{\tau>0}$ in $\mathcal{C}^0([0,T],L^1_{\mathrm{weak}}(\mathbb{R}^d))$ for any $T>0$ allows us to pass to the limit in the linear parts of (3.11) because $\nabla\zeta$ and $\Delta\zeta$ are bounded in $L^\infty(\mathbb{R}^d)$. The quadratic last term of (3.11), coming from the concave interaction energy, is more difficult to handle. Actually, we shall make another use of the doubling of variables trick: $\rho_\tau\otimes\rho_\tau\rightharpoonup\rho\otimes\rho$, thanks to the equi-integrability, obtained in Lemma 2.3. In addition, recall that the convergence is uniform in time thanks to the Ascoli–Arzelà theorems, and that equi-integrability bounds are also uniform with respect to τ. We can thus pass to the limit when τ goes to zero in (3.11) to obtain for any t_1, t_2

$$\int_{\mathbb{R}^d}\zeta(x)\,[\rho(t_2,x)-\rho(t_1,x)]\,\mathrm{d}x=\int_{t_1}^{t_2}\int_{\mathbb{R}^d}[\Delta\zeta(x)+x\cdot\nabla\zeta(x)]\,\rho(s,x)\,\mathrm{d}x\,\mathrm{d}s$$
$$-\frac{\chi}{2\,d\,\pi}\int_{t_1}^{t_2}\int_{\mathbb{R}^d\times\mathbb{R}^d}\frac{[\nabla\zeta(x)-\nabla\zeta(y)]\cdot(x-y)}{|x-y|^2}\rho(s,y)\rho(s,x)\,\mathrm{d}y\,\mathrm{d}x\,\mathrm{d}s,$$

which is a formulation of the weak solution as defined in (2.5). ∎

Remark 3.5 (original variables). This theorem is true in original variables (1.1) locally in time, when we consider the free energy (1.4) with minor changes; see also [32]. However, the long-time asymptotics are better understood in scaled variables (see the numerical results, section 4.2.1).

Remark 3.6 (comparison to literature). Rigorous convergence of numerical schemes was first obtained by Filbet in [28] in the nonradial case. His approach relies on the entropy and the Gagliardo–Nirenberg–Sobolev inequality. However, in our case as we crucially use the free-energy functional, we are able to prove in Theorem 3.4 the convergence of the JKO scheme for all values $\chi<\chi_c$, where χ_c is known to be optimal. It is worth noticing that the minimizing movement scheme is very expensive numerically in dimension greater than one, at least as a direct implicit discretization detailed in section 4.2. In contrast, the two-dimensional PKS model is much more manageable following [28].

Beyond the optimal threshold, there is another result arguing on this alternative scheme's behalf. Namely, for a given τ the solution $\{\rho_\tau^n\}_n$ is shown to converge towards the unique stationary state, with explicit exponential rate in the one-dimensional case; see Proposition 4.3 in the next section.

Remark 3.7 (nonlinear diffusions). The ongoing work extends to nonlinear diffusions, under suitable assumptions, $\Delta f(n)$ without deeper difficulty. The main point is that the free energy should be bounded from below—this results from a balance between diffusion at high density level (diffusion must be superlinear at infinity above the optimal threshold) and at low density level (basically $f(u) \gtrsim u^\alpha$, where $\alpha > \max(1/2, 1 - 2/d)$ preventing mass from escaping too fast towards infinity); see [13] and [17].

Remark 3.8 (radial case). In the two-dimensional radial case, one can write a closed equation for the mass inside a ball; see [38]. The equation for this radially cumulative function is a local one-dimensional parabolic equation of Burger's type which can be solved in a standard way. As we shall see, we also deal, in dimension one, with the cumulative distribution functions, but this leads to a nonlocal equation. The radial case has already been treated by several authors; see, for instance, [11]. Let us also note that in radial coordinates we lose the gradient flow structure of the equation, which is the basis of the JKO scheme.

4. One-dimensional case. In the case of the real line, consider μ and ν two absolutely continuous measures with respect to the Lebesgue measure, of respective densities f and g. Let $\nabla\varphi$ be the optimal map which transports μ onto ν. In dimension bigger than one, the optimal map is the solution of the following so-called Monge–Ampére equation:

$$f = g \circ \nabla\varphi \operatorname{Det} D^2\varphi,$$

which is fully nonlinear and cannot be solved in general. However, in dimension one it is a increasing rearrangement, as we will discuss below. Let F and G be the cumulative functions of f and g. As the cumulative distribution function is nondecreasing, we can define the pseudoinverse function by

$$V(z) = F^{-1}(z) := \inf\{x \,:\, F(x) \geq z\}.$$

By Brenier's theorem and the definition of the image measure (2.7), we have

$$F(x) = \int_{-\infty}^{x} f(y)\,\mathrm{d}y = \int_{-\infty}^{\varphi'(x)} g(y)\,\mathrm{d}y = G \circ \varphi'(x).$$

Hence, the transport map can be stated explicitly as $\varphi' = G^{-1} \circ F$, and the Wasserstein distance can be expressed in the following more tractable way:

$$d_W^2(\mu, \nu) = \int_0^1 |F^{-1}(w) - G^{-1}(w)|^2\,\mathrm{d}w. \tag{4.1}$$

In fact, $G^{-1} \circ F$ is the optimal map for all convex costs in dimension one [48, Theorem 2.18].

This expression of the one-dimensional Wasserstein distance has been used for nonlinear diffusions and nonlinear nonlocal friction equations in granular media [20, 34, 16] to analyze the long-time asymptotics and the contraction properties with respect to Wasserstein distances of those equations. Moreover, these ideas have first been used in [29, 30] for numerical purposes. Explicit in time numerical schemes for the equations of the inverse distribution function are proposed, keeping the contraction of the Wasserstein distance at the discrete level. Here, we prefer to solve it by an implicit in time Euler scheme since it coincides with the JKO scheme through the

representation (4.1) and since, moreover, the contraction property of the Wasserstein distance is not true due to the lack of convexity of the functional $\mathcal{G}$.

More precisely, let F_n and F_{n+1} be the cumulative distribution functions associated respectively to ρ_τ^n and ρ_τ^{n+1}. By the expression (4.1) of the Wasserstein distance on the real line, the scheme (3.1) can be rewritten in terms of $V_n = F_n^{-1}$ and $V_{n+1} = F_{n+1}^{-1}$ as the gradient flow of the inverse distribution function subject to L^2-metric structure:

$$V_{n+1} \in \arg\inf_{\{W\,:\,(W^{-1})'\in\mathcal{K}\}} \left[\tilde{\mathcal{G}}[W] + \frac{1}{2\,\tau}\|W - V_n\|^2_{L^2(0,1)}\right],$$

where

$$\tilde{\mathcal{G}}[W] := -\int_0^1 \log W'(w)\,\mathrm{d}w + \int_0^1 |W(w)|^2\,\mathrm{d}w + \frac{\chi}{\pi}\int_0^1\int_0^1 \log|W(w) - W(z)|\,\mathrm{d}w\,\mathrm{d}z.$$

Here the metric is Euclidean, and hence the Euler–Lagrange equation associated to this minimization problem is

$$\frac{V_{n+1} - V_n}{\tau} = -\nabla\tilde{\mathcal{G}}[V_{n+1}]\,,$$

where ∇ is the usual gradient operator in $L^2(\mathbb{R})$. This Euler–Lagrange equation can be rewritten as

$$-\frac{V_{n+1}(w) - V_n(w)}{\tau} = \frac{\partial}{\partial w}\left[\left(\frac{\partial V_{n+1}(w)}{\partial w}\right)^{-1}\right] + V_{n+1}(w) + \chi H[V_{n+1}], \tag{4.2}$$

where $\tilde{H}$ is related to the Hilbert transform and is defined by

$$\tilde{H}[V](w) := \frac{1}{\pi}\lim_{\varepsilon\to 0}\int_{|V(w)-V(z)|\geq\varepsilon}\frac{1}{V(w) - V(z)}\,\mathrm{d}z\,.$$

For the sake of simplicity we assume that the space step is constant, equal to h. If we set $V_n^i := V_n(ih)$, for any $i = 0\cdots N$, and $Nh = 1$, the finite difference discretization in space of (4.2) is the following implicit Euler scheme in rescaled variables:

$$\begin{aligned}-\frac{V_{n+1}^i - V_n^i}{\tau} = \frac{1}{V_{n+1}^{i+1} - V_{n+1}^i} - \frac{1}{V_{n+1}^i - V_{n+1}^{i-1}} + V_{n+1}^i \\ + \frac{\chi h}{\pi}\lim_{\varepsilon\to 0}\sum_{j\,:\,|V_{n+1}^i - V_{n+1}^j|\geq\varepsilon}\frac{1}{V_{n+1}^i - V_{n+1}^j}\end{aligned} \tag{4.3}$$

with initial condition $V_0 = \rho_0$. We impose Neumann boundary conditions in the ρ-problem (2.3), i.e., for any n, $\frac{1}{V_n^N - V_n^{N-1}} = 0$ and $\frac{1}{V_n^1 - V_n^0} = 0$, so that the "center of mass" is conserved:

$$\forall n \quad \sum_{i=0}^N V_n^i = 0.$$

The solution at each time-step of the nonlinear system of equations is obtained by an iterative Newton–Raphson procedure.

Remark 4.1 (higher dimensions). A way to extend the previous numerical scheme to higher dimensions is to solve an equivalent L^2 gradient flow for diffeomorphisms proposed in [27]. The price to pay is that one needs to solve a fully nonlinear coupled system of d partial differential equations.

4.1. Exponential convergence in one dimension. Here comes the second main improvement and motivation for this numerical scheme. In addition to convergence as the time step goes to zero, we are able to show that for a fixed $\tau > 0$ the discrete solution converges to a unique steady state as time goes to infinity. First we prove existence of this steady state equilibrium; see Proposition 4.2. Uniqueness and convergence follow from Proposition 4.3.

PROPOSITION 4.2. *Let $(V^i)_i$ be an increasing sequence. The discrete free energy functional defined as*

$$\tilde{\mathcal{G}}_h[V] = -\sum_{i=1}^{N-1} h \log\left(\frac{V^{i+1}-V^i}{h}\right) + \frac{1}{2}\sum_{i=1}^{N} h|V^i|^2 + \frac{\chi}{2\pi}\sum_{i=1}^{N}\sum_{j=1,\ j\neq i}^{N} h^2 \log|V^j - V^i|$$

is coercive and bounded from below if $\chi(1-h) < \chi_c = 2\,\pi$.

Proof. We shall adapt to the discrete setting the so-called logarithmic Hardy–Littlewood–Sobolev (2.1) and Carleman inequalities; see Lemma 2.2.

Step 1. A discrete logarithmic Hardy–Littlewood–Sobolev inequality: we will prove the following discrete logarithmic Hardy–Littlewood–Sobolev inequality:

$$\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N} h^2 \log|V^j - V^i| \geq (1-h)\sum_{i=1}^{N-1} h \log\left(\frac{V^{i+1}-V^i}{h}\right) + C_h. \tag{4.4}$$

Observe that the space step h modifies the apparent space dimension: the factor $d_h := 1 - h$ replaces $d = 1$ in front of the entropy term. Observe also that the correcting factor d_h is the only admissible one for homogeneity reasons. In fact, when V is multiplied by $\lambda > 0$ (think of $\lambda \to 0$ or $\lambda \to \infty$), then inequality (4.4) is violated if $d_h \neq 1 - h$: the additional term $h(N-1)\big(hN - (1-h)\big)\log\lambda$ comes out, and $hN = d_h$.

The left-hand side in (4.4) can be reorganized as following, where the increasingness of $(V^i)_i$ is crucially used:

$$\begin{aligned}
\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N} h^2 \log|V^j - V^i| &= \sum_{i=1}^{N}\sum_{j=i+1}^{N} h^2 \log(V^j - V^i) + \sum_{i=1}^{N}\sum_{j=1}^{i-1} h^2 \log(V^i - V^j)\\
&\geq \sum_{i=1}^{N-1}\sum_{j=i+1}^{N} h^2 \log(V^{i+1} - V^i) + \sum_{k=1}^{N-1}\sum_{j=1}^{k} h^2 \log(V^{k+1} - V^j)\\
&\geq \sum_{i=1}^{N-1} h^2 (N-i)\log(V^{i+1} - V^i) + \sum_{k=1}^{N-1} h^2 k \log(V^{k+1} - V^k)\\
&= \sum_{i=1}^{N-1} h(hN)\log(V^{i+1} - V^i)\\
&= d_h \sum_{i=1}^{N-1} h \log\left(\frac{V^{i+1}-V^i}{h}\right) + C_h,
\end{aligned}$$

which proves the discrete logarithmic Hardy–Littlewood–Sobolev inequality (4.4).

Step 2. A discrete Carleman inequality: we shall prove next the following inequality:

$$-\sum_{i=1}^{N-1} h \log\left(\frac{V^{i+1}-V^i}{h}\right) + \frac{1}{2}\sum_{i=1}^{N} h|V_i|^2 \geq C_h. \tag{4.5}$$

Introduce the median of the sequence $(V^i)_i$; that is, select k^0 such that $|V^{k_0}| = \min |V^k|$. Within this choice we have in particular $V^k \le 0$ for $k < k_0$ and $V^k \ge 0$ for $k > k_0$. Thus we obtain

$$
\begin{aligned}
&- \sum_{i=1}^{N-1} h \log\left(\frac{V^{i+1}-V^i}{h}\right) + \frac{1}{2}\sum_{i=1}^{N} h|V^i|^2 \\
&\ge - \sum_{i=1}^{N-1} h \log\left(\frac{V^{i+1}-V^i}{h}\right) + \frac{1}{2}\sum_{i=1, i\neq k_0}^{N} h|V^i|^2 \\
&= -h(N-1)\sum_{i=1}^{k_0-1} \frac{1}{N-1}\log\left((V^{i+1}-V^i)\mathrm{e}^{-\frac{1}{2}|V^i|^2}\right) \\
&\qquad - h(N-1)\sum_{i=k_0}^{N-1} \frac{1}{N-1}\log\left((V^{i+1}-V^i)\mathrm{e}^{-\frac{1}{2}|V^{i+1}|^2}\right) + C_h \\
&\ge -h(N-1)\log\left[\frac{1}{N-1}\left(\sum_{i=1}^{k_0-1}(V^{i+1}-V^i)\mathrm{e}^{-\frac{1}{2}|V^i|^2} + \sum_{i=k_0}^{N-1}(V^{i+1}-V^i)\mathrm{e}^{-\frac{1}{2}|V^{i+1}|^2}\right)\right] + C_h \\
&\ge C_h,
\end{aligned}
$$

thanks to the Jensen inequality and the peculiar choice of k_0. The last minoration step comes from

$$
\begin{aligned}
&\sum_{i=1}^{k_0-1} \mathrm{e}^{-\frac{1}{2}|V^i|^2}(V^{i+1}-V^i) + \sum_{i=k_0}^{N-1} \mathrm{e}^{-\frac{1}{2}|V^{i+1}|^2}(V^{i+1}-V^i) \\
&\le \sum_{i=1}^{k_0-1} \mathrm{e}^{-\frac{1}{2}|V^i|^2}|V^i| + \mathrm{e}^{-\frac{1}{2}|V^{k_0-1}|^2}|V^{k_0}| + \mathrm{e}^{-\frac{1}{2}|V^{k_0+1}|^2}|V^{k_0}| + \sum_{i=k_0}^{N-1} \mathrm{e}^{-\frac{1}{2}|V^{i+1}|^2}|V^{i+1}| \\
&\le C_h,
\end{aligned}
$$

since the function $\mathrm{e}^{-\frac{1}{2}|X|^2}|X|$ is uniformly bounded and $|V^{k_0}| \le \min\{|V^{k_0-1}|, |V^{k_0+1}|\}$.

Step 3. Coercivity of the discrete free energy: as a by-product of discrete logarithmic Hardy–Littlewood–Sobolev and Carleman inequalities we obtain the following coercivity a priori estimate ensuring the existence of a minimizer (which is the stationary state we are looking for). Define $\theta = \frac{\chi(1-h)}{2\pi} < 1$ and deduce from the discrete logarithmic Hardy–Littlewood–Sobolev inequality (4.4) that

$$
\tilde{\mathcal{G}}_h[V] \ge (1-\theta)\left(- \sum_{i=1}^{N-1} h \log\left(\frac{V^{i+1}-V^i}{h}\right)\right) + \frac{1}{2}\sum_{i=1}^{N} h|V^i|^2 + C_h. \tag{4.6}
$$

Then apply the discrete Carleman inequality (4.5) to eventually obtain

$$
\tilde{\mathcal{G}}_h[V] \ge \frac{\theta}{2}\sum_{i=1}^{N} h|V^i|^2 + C_h. \qquad \square
$$

As a by-product we obtain existence of a stationary state realized as a minimizer of the free energy functional. Indeed, it remains to show that a minimizing sequence cannot blow-up in the sense that $V_{i+1} - V_i$ is prevented from vanishing. If so, this would contradict the upper bound obtained above in (4.6).

PROPOSITION 4.3 (convergence in the subcritical case). *Assume that* $\chi(1-h) < \chi_c$. *Then the solution of the numerical scheme* (4.3) *converges to the unique steady state of the problem with exponential rate.*

Proof. First we need the following two characterizations of the unique equilibrium state. The uniqueness will in fact follow from the convergence proof, as we shall see later (Remark 4.4). The discrete function $(V_*^i)_i$ is an equilibrium if and only if

$$\forall i, \quad 0 = \frac{1}{V_*^{i+1} - V_*^i} - \frac{1}{V_*^i - V_*^{i-1}} + V_*^i + \frac{\chi}{\pi} \sum_{j \neq i} h \frac{1}{V_*^i - V_*^j}, \tag{4.7}$$

or equivalently

$$\forall k, \quad (V_*^{k+1} - V_*^k) \left\{ \frac{\chi}{\pi} \sum_{j=0}^{k} \sum_{i=k+1}^{N} h \frac{1}{V_*^i - V_*^j} - \sum_{i=0}^{k} V_*^i \right\} = 1. \tag{4.8}$$

To see that (4.7) and (4.8) are equivalent, rewrite the latter as

$$\forall k, \quad \frac{1}{V_*^{k+1} - V_*^k} = \frac{\chi}{\pi} \sum_{j=0}^{k} \sum_{i=k+1}^{N} h \frac{1}{V_*^i - V_*^j} - \sum_{i=0}^{k} V_*^i$$

and then "derive" the ongoing expression in a discrete way,

$$\begin{aligned} \frac{1}{V_*^{k+1} - V_*^k} - \frac{1}{V_*^k - V_*^{k-1}} &= \frac{\chi}{\pi} \sum_{j=0}^{k} \sum_{i=k+1}^{N} h \frac{1}{V_*^i - V_*^j} - \frac{\chi}{\pi} \sum_{j=0}^{k-1} \sum_{i=k}^{N} h \frac{1}{V_*^i - V_*^j} - V_*^k \\ &= \frac{\chi}{\pi} \sum_{i=k+1}^{N} h \frac{1}{V_*^i - V_*^k} - \frac{\chi}{\pi} \sum_{j=0}^{k-1} h \frac{1}{V_*^k - V_*^j} - V_*^k. \end{aligned}$$

We proceed as computing the time evolution of the L^2-distance between V_n and the stationary state V_*:

$$\begin{aligned} \frac{1}{2\tau} \Big(\|V_{n+1} - V_*\|^2 - \|V_n - V_*\|^2 \Big) &= \frac{1}{2\tau} \sum_i h (V_{n+1}^i - V_n^i)(V_{n+1}^i + V_n^i - 2V_*^i) \\ &= \sum_i h \frac{V_{n+1}^i - V_n^i}{\tau} (V_{n+1}^i - V_*^i) - \frac{1}{2\tau} \sum_i h (V_{n+1}^i - V_n^i)^2. \end{aligned}$$

We then input the evolution equation for $V_{n+1} - V_n$, and obtain, thanks to (4.7),

$$\begin{aligned} \frac{1}{2\tau} \Big(\|V_{n+1} - V_*\|^2 - \|V_n - V_*\|^2 \Big) \leq - \sum_i h \Bigg(& \frac{1}{V^{i+1} - V^i} - \frac{1}{V^i - V^{i-1}} - \frac{1}{V_*^{i+1} - V_*^i} \\ & + \frac{1}{V_*^i - V_*^{i-1}} + V^i - V_*^i + \frac{\chi}{\pi} \sum_{j \neq i} h \frac{1}{V^i - V^j} \\ & - \frac{\chi}{\pi} \sum_{j \neq i} h \frac{1}{V_*^i - V_*^j} \Bigg) (V^i - V_*^i) \\ = A_n + B_n + C_n, \end{aligned}$$

where V stands for V_{n+1} without any ambiguity. We integrate by parts the first (diffusion) contribution, to obtain

$$\begin{aligned} A &= -\sum_i h\left(\frac{1}{V^{i+1}-V^i}-\frac{1}{V^i-V^{i-1}}-\frac{1}{V_*^{i+1}-V_*^i}+\frac{1}{V_*^i-V_*^{i-1}}\right)(V^i-V_*^i)\\ &= \sum_i h\left(\frac{1}{V^{i+1}-V^i}-\frac{1}{V_*^{i+1}-V_*^i}\right)(V^{i+1}-V_*^{i+1}-V^i+V_*^i). \end{aligned}$$

We have carefully used the boundary conditions $\frac{1}{V^N-V^{N-1}}=0$ and $\frac{1}{V^1-V^0}=0$. We can rewrite A using zero-homogeneity of the last expression, namely,

$$A=\sum_i h\gamma\left(\frac{V^{i+1}-V^i}{V_*^{i+1}-V_*^i}\right),$$

where $\gamma(\lambda)=2-\lambda-\lambda^{-1}$ is concave and nonpositive. The second contribution coming from variable rescaling is obvious but crucial, namely,

$$B=-\sum_i h(V^i-V_*^i)^2=-\|V_{n+1}-V_*\|^2.$$

The last (interaction) contribution is more tricky to handle and involves variables doubling, as is known from granular media; see [20]. We have

$$\begin{aligned} C &= -\frac{\chi}{\pi}\sum_i h\left(\sum_{j\neq i} h\frac{1}{V^i-V^j}-\sum_{j\neq i} h\frac{1}{V_*^i-V_*^j}\right)(V^i-V_*^i)\\ &= -\frac{\chi}{2\pi}\sum_{i,j,i\neq j} h^2\left(\frac{1}{V^i-V^j}-\frac{1}{V_*^i-V_*^j}\right)(V^i-V^j-V_*^i+V_*^j)\\ &= -\frac{\chi}{2\pi}\sum_{i,j,i\neq j} h^2\gamma\left(\frac{V^i-V^j}{V_*^i-V_*^j}\right). \end{aligned}$$

Notice that the above expression is symmetric between the two possible choices $i<j$ and $j<i$. We shall also make use of the concavity property of γ as follows:

$$\begin{aligned} C &= -2\frac{\chi}{2\pi}\sum_{j<i} h^2\gamma\left(\sum_{j\le k<i}\frac{V^{k+1}-V^k}{V_*^{k+1}-V_*^k}\cdot\frac{V_*^{k+1}-V_*^k}{V_*^i-V_*^j}\right)\\ &\le -\frac{\chi}{\pi}\sum_{i<j} h^2\sum_{j\le k<i}\gamma\left(\frac{V^{k+1}-V^k}{V_*^{k+1}-V_*^k}\right)\frac{V_*^{k+1}-V_*^k}{V_*^i-V_*^j}\\ &= -\frac{\chi}{\pi}\sum_k h\gamma\left(\frac{V^{k+1}-V^k}{V_*^{k+1}-V_*^k}\right)(V_*^{k+1}-V_*^k)\sum_{j=0}^{k}\sum_{i=k+1}^{N} h\frac{1}{V_*^i-V_*^j}\\ &\le -\sum_k h\gamma\left(\frac{V^{k+1}-V^k}{V_*^{k+1}-V_*^k}\right)(V_*^{k+1}-V_*^k)\left\{\frac{\chi}{\pi}\sum_{j=0}^{k}\sum_{i=k+1}^{N} h\frac{1}{V_*^i-V_*^j}-\sum_{i=0}^{k}V_*^i\right\}, \end{aligned}$$

where we have used the fact that $D_k:=\sum_{i=0}^k V_*^i$ is a nonpositive quantity for all k.

Let us prove this last claim. Indeed, $D_{k+1} - D_{k-1} - 2\,D_k = V_*^{k+1} - V_*^k \geq 0$. Hence, $D_{k+1} - D_k \geq D_{k-1} - D_k$. Since $D_N = 0$ and $D_0 = V_*^0 \leq 0$, there exists k_0 such that $(D_k)_{1\leq k\leq k_0}$ is nondecreasing and $(D_k)_{k_0\leq k\leq N}$ is nonincreasing. As a consequence, $D_k \leq D_0 \leq 0$ for any integer k, $1 \leq k \leq k_0$, and $D_k \leq D_N = 0$ for any integer k, $k_0 + 1 \leq k \leq N$, which proves the assertion.

At this stage we bring in the alternative representation of the stationary solution (4.8), so that $A + C \leq 0$. As a consequence we obtain

$$\frac{1}{2\tau}\Big(\|V_{n+1} - V_*\|^2 - \|V_n - V_*\|^2\Big) \leq -\|V_{n+1} - V_*\|^2. \tag{4.9}$$

We finally get the exponential convergence rate,

$$\|V_n - V_*\|^2 \leq \left(\frac{1}{1+2\tau}\right)^n \|V^0 - V_*\|^2.$$

If τ is small, we can thus approximate $\log(1+2\tau) \approx 2\tau$ and $(\frac{1}{1+2\tau})^n \approx \exp(-2n\tau) \approx \exp(-2t)$. Thus, the bound on the rate, we find, does not depend on the parameter $\chi < \chi_c$. □

Remark 4.4 (uniqueness of the stationary solution). We can deduce a posteriori the uniqueness of the equilibrium. As a matter of fact, let us consider another equilibrium state $\tilde{V}_*$ and set $V_{n+1} = V_n = \tilde{V}_*$ in the above computations. We eventually obtain $\|\tilde{V}_* - V_*\| \leq 0$ from (4.9), which proves the uniqueness.

To complete the picture, we shall prove here that the discrete approximation (4.3) (without the drift term) blows-up in finite time in the supercritical case $(1-h)\chi > \chi_c$, in the sense that the iterations cannot be recursively handled forever. In fact, at some point some increment $V_n^{i+1} - V_n^i$ may vanish, breaking the scheme. We proceed as for the continuous virial method, by computing the evolution of the second-momentum, that is, the L^2-norm of (V_n) in this context. It is shown to vanish eventually in finite time, which is an obstruction.

PROPOSITION 4.5 (blow-up of the discrete solution). *Assume we are in the supercritical case $\chi(1-h) > \chi_c$; then the Keller–Segel steepest descent scheme (without a drift term) blows-up in finite time.*

Proof. We follow exactly the proof of Proposition 4.3, where integration by parts and symmetrization of the "kernel" were already performed.

$$\begin{aligned}
\frac{1}{2\tau}\left(\|V_{n+1}\|^2 - \|V_n\|^2\right) &= \frac{1}{2\tau}\sum_i h(V_{n+1}^i - V_n^i)\left(V_{n+1}^i + V_n^i\right)\\
&= \sum_i h\frac{V_{n+1}^i - V_n^i}{\tau}V_{n+1}^i - \frac{1}{2\tau}\sum_i h\left(V_{n+1}^i - V_n^i\right)^2\\
&\leq \sum_i h\frac{V_{n+1}^i - V_n^i}{\tau}V_{n+1}^i\\
&\leq \sum_i h\left(\frac{1}{V^{i+1} - V^i} - \frac{1}{V^i - V^{i-1}} + \frac{\chi}{\pi}\sum_{j\neq i} h\frac{1}{V^i - V^j}\right)V^i\\
&= A + B,
\end{aligned}$$

where V stands for V_{n+1} without any ambiguity. In fact,

$$A = \sum_i h\,\frac{V^{i+1} - V^i}{V^{i+1} - V^i} = 1,$$

and, doubling the sum,

$$C = -\frac{\chi}{2\pi} \sum_{i,j,i\neq j} h^2 \frac{V^i - V^j}{V^i - V^j}$$
$$= -\frac{\chi h^2}{2\pi} N(N+1) = -\frac{\chi(1-h)}{2\pi} .$$

Thus, in case of $(1-h)\chi > \chi_c$ the second momentum decreases at least linearly in time, ensuring blow-up after a finite time. □

Remark 4.6 (original variables). When coming back in the continuous setting to the original variables $n(t,x)$ by (2.2), we are not able to show even that

$$n(t,x) = n^\infty(t,x) + O_W(1),$$

where n^∞ is the dilatation of the stationary state,

$$n^\infty(t,x) = \frac{1}{1+2t} V_* \left(\frac{x}{\sqrt{1+2t}} \right),$$

and O_W means infinitesimal in the d_W sense. The reason is that the found estimate on the speed of convergence does not depend on the reduced parameter χ, but only on the variable rescaling, and the change of variables restores exactly the factor e^t due to the scaling properties of d_W [48, 21]. This result should be improved as seen from the numerical experiments below, and it is an open problem how to get a faster speed of convergence in the scaled equation leading to a polynomial decay in original variables. In fact, we conjecture that if we fix the center of mass, then the rate of convergence of the solution to the stationary solution in rescaled variables is of order e^{-2t}. This fact coincides with other situations as in nonlinear diffusions in which fixing certain invariants of the equation improves the rate of convergence [23]. Certainly the situation is close to the heat equation for small mass solutions [26].

In the next sections, we will show some numerical experiments for the PKS equation using the scheme (4.3). We begin with the subcritical case $\chi = \pi$ (recall that in dimension one the critical parameter is $\chi_c = 2\pi$) for the not rescaled (section 4.2.1) as opposed to the rescaled (section 4.2.2) system. We next approach the critical parameter, plugging $\chi = 1.8\pi$, and initialize the scheme with a two-peaks density n_0 (resp., sections 4.2.3 and 4.2.4). We then investigate the upper-critical case starting with respectively a single peak (section 4.3.1), two symmetric peaks attracting each other (section 4.3.2), and two asymmetric peaks (section 4.4).

In the following, we assume a uniform in space discretization $w_i = ih$, $i = 0, \ldots, N$, and $Nh = 1$.

4.2. Subcritical case.

4.2.1. Not rescaled case. Starting with the centered initial data,

$$V_0^i = 2 \frac{w_i - 0.5}{[(w_i + 0.01)(1.01 - w_i)]^{1/4}} ,$$

corresponding to a compact supported density n_0, we numerically solve the PKS equation on the time interval $[0, 400]$ with $\chi = \pi$. Figures 4.1 and 4.2 show the evolution of the solution both for the density (Figure 4.2) and its inverse distribution function (Figure 4.1).

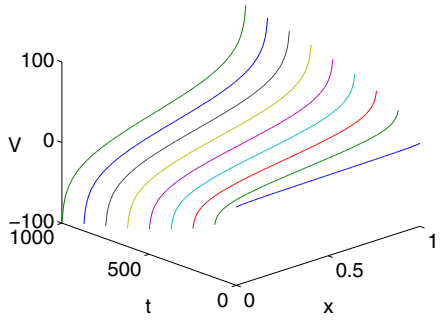


FIG. 4.1. *Inverse cumulative distribution function for $\chi = \pi$. Note that the initial data seems to be flat relative to the very large scale on the V-axis, as opposed to Figure* 4.5.

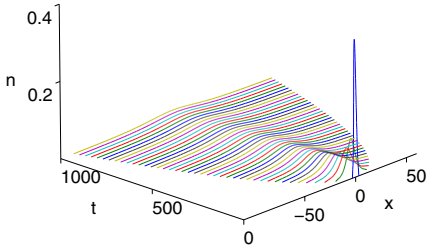


FIG. 4.2. *Cell density n as time evolves, obtained from its inverse cumulative distribution function. According to Figure* 4.1*, the space scale is also very large, and therefore the density seems highly concentrated at $t = 0$.*

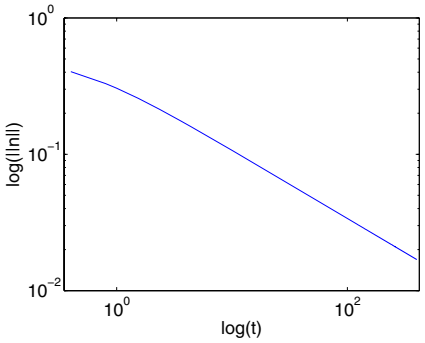


FIG. 4.3. *L^2-norm's evolution for the cell density n, in a* log-log *scale. The decay appears to be polynomial.*

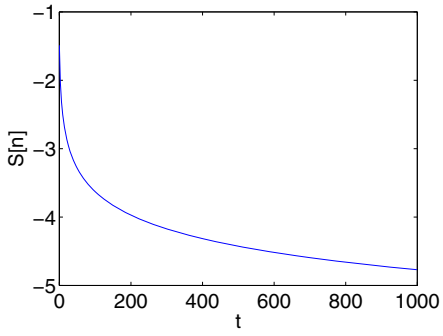


FIG. 4.4. *Evolution of the entropy $\mathcal{S}[n]$ showing slow decay.*

Observe in Figure 4.1 that the branches of the inverse cumulative function V go eventually to $\pm\infty$. This is expected because the support of the cell density spreads as time goes on, and therefore the distribution tails are wider. Recall that in the subcritical regime, the diffusion process dominates. The scheme captures well the collapse down to zero of the cell density and the spreading of the solution. Interestingly, this scheme handles easily with moving density's support (note that finite speed of propagation is a numerical artifact), whereas the reference domain $[0, 1]$ is fixed because we deal with probability densities (mass is conserved).

Moreover, the spreading towards zero seems to be polynomial from Figure 4.3 showing the evolution of the L^2-norm of the cell density in log-log scale. The entropy decay is plotted in Figure 4.4.

4.2.2. Rescaled variables: $\chi = \pi$. Given the compactly supported initial data,

$$V_0^i = 2\,\frac{w_i - 0.5}{\left[(w_i + 0.01)\,(1.01 - w_i)\right]^{1/4}}\,,$$

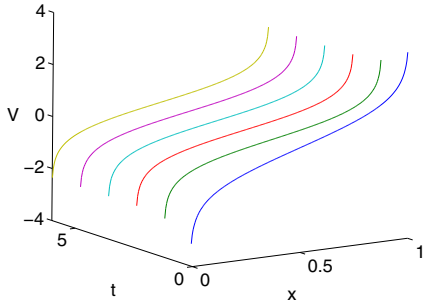


FIG. 4.5. *Fast convergence towards the stationary solution for $\chi = \pi$ and rescaled variables.*

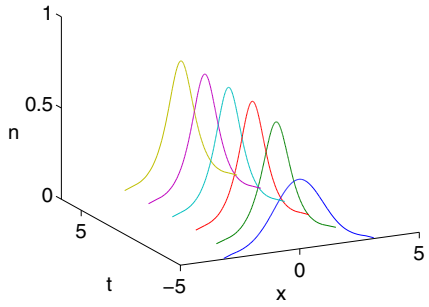


FIG. 4.6. *Evolution of the corresponding cell density n.*

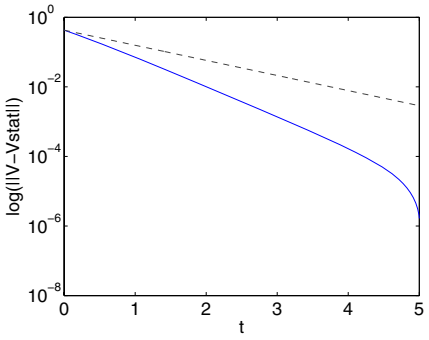


FIG. 4.7. *Wasserstein distance between the density at time t and the final computed density, assumed to be almost the stationary solution. Dashed line: decay rate proved in Proposition* 4.3.

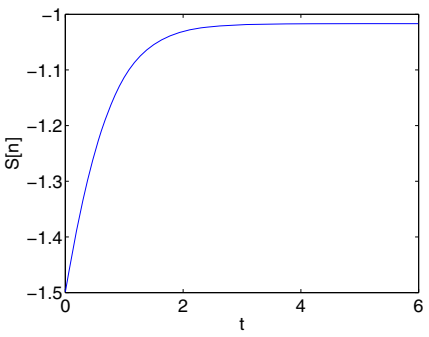


FIG. 4.8. *Evolution of the entropy.*

we solve the PKS equation numerically in rescaled variables on the time interval $[0, 5]$ with $\chi = \pi$ (corresponding results are shown in Figures 4.5 and 4.6, 4.7 and 4.8).

In contrast to previous section 4.2.1, we observe an exponential convergence towards the stationary solution (see Figures 4.5 and 4.7). When computing the Wasserstein distance between the density at time t and the expected stationary solution (last computed time), we find out that the convergence is faster than e^{-t} (Figure 4.7) obtained in Proposition 4.3. This confirms the open problem of trying to find a better decay rate in scaled variables that will eventually lead to a polynomial decay rate towards self-similarity in original variables for subcritical masses.

4.2.3. Rescaled variables: $\chi = 1.8\pi$. Given the compactly supported initial data,

$$V_0^i = 2\,\frac{w_i - 0.5}{\left[(w_i + 0.01)\,(1.01 - w_i)\right]^{1/4}}\,,$$

we numerically solve the PKS equation in rescaled variables on the time interval $[0, 5.5]$ with $\chi = 1.8\pi$. Figures 4.9 and 4.10 show the evolution of the solution.

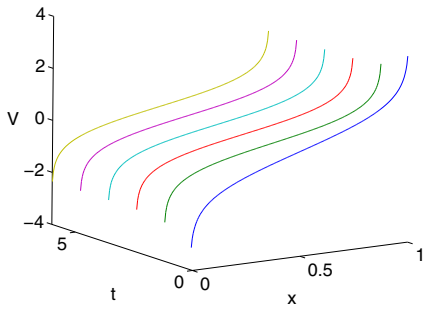


FIG. 4.9. *Cumulative distribution function V for the subcritical case $\chi = 1.8\pi < \chi_c$.*

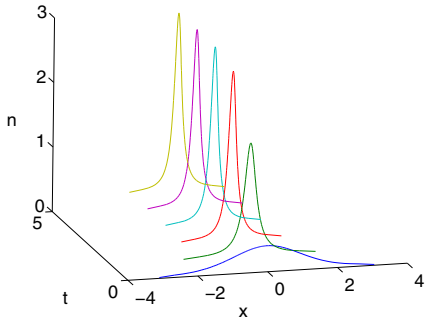


FIG. 4.10. *Evolution of the cell density n.*

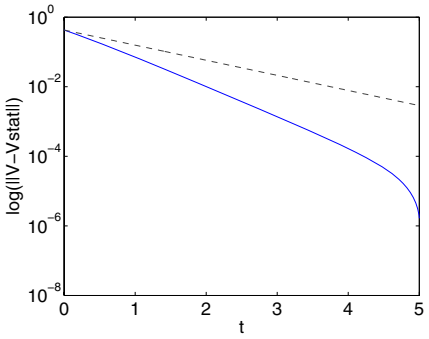


FIG. 4.11. *Wasserstein distance between the density at time t and the final computed density, assumed to be almost the stationary solution. Dashed line: decay rate proved in Proposition* 4.3.

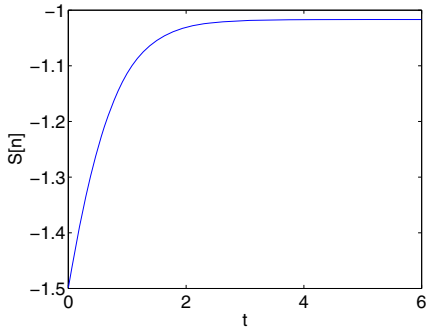


FIG. 4.12. *Evolution of the entropy.*

The initial data is the same as in section 4.2.2, but χ is closer to the critical parameter χ_c. The solution again converges exponentially to the stationary solution (see Figures 4.9 and 4.10). According to Proposition 4.3, the rate of convergence is at least the same as in section 4.2.2; compare Figure 4.11 to Figure 4.7 and the slopes are the same (of order e^{-2t}). On the other hand, the equilibrium state is more concentrated (Figure 4.10), corresponding to a flat plateau in Figure 4.9, as we expect it converges to a Dirac mass. The entropy seems to converge; see Figure 4.12.

4.2.4. Two-peaks initial data. Initialize with the centered cumulative distribution function,

$$V_0^i = \frac{\exp\left[10\ (w_i - 0.5)\right] - 1}{\left[(w_i + 0.01)\,(1.01 - w_i)\right]^{1/4}}\ ,$$

corresponding to a two-peaks-like density with compact support. We numerically solve the PKS equation in rescaled variables on the time interval $[0, 5]$ with $\chi = \pi$.

Whereas the parameter χ is the same as in section 4.2.2, the initial data is qualitatively different. The two peaks diffuse, eventually merging and finally converging to the stationary solution with exponential speed (see Figures 4.13 and 4.14). Let

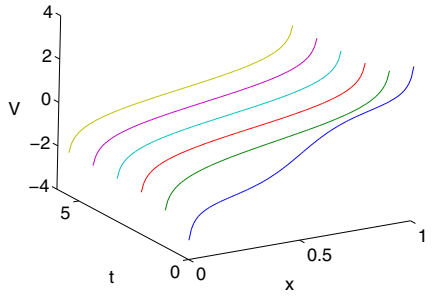


FIG. 4.13. *Cumulative distribution function V for $\chi = \pi$ and a two-peaks initial condition.*

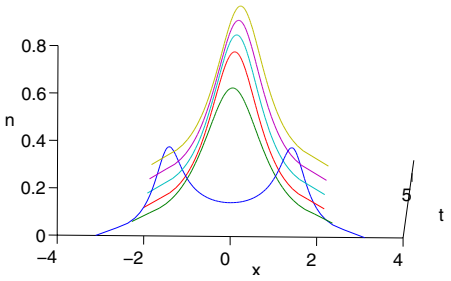


FIG. 4.14. *Cell density n.*

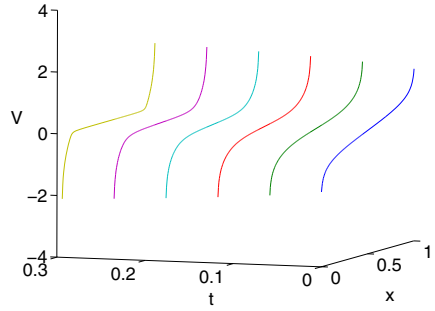


FIG. 4.15. *Cumulative distribution function V for $\chi > \chi_c$. The solution blows-up exhibiting a plateau in finite time.*

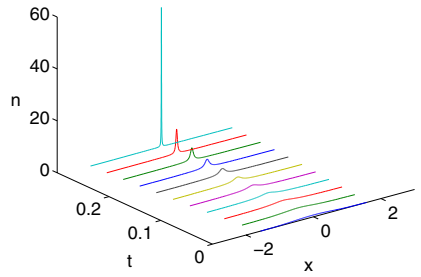


FIG. 4.16. *Cell density n. We observe blow-up in finite time.*

us mention finally that the numerical scheme does not preserve the exact value of the critical mass. The scheme does not preserve the law of evolution of the second moment. However, with the same initial data as in the first two subsections, the numerical critical mass is situated between 1.973π and 1.974π.

4.3. Supercritical case.

4.3.1. Single-peak initial data. Given the compactly supported initial data,

$$V_0^i = 2\,\frac{w_i - 0.5}{\left[(w_i + 0.01)\,(1.01 - w_i)\right]^{1/4}}\,,$$

we numerically solve the PKS equation in original variables on the time interval $[0, 0.32]$ with $\chi = (5/2)\,\pi$. Note that in the upper-critical case, the variables' rescaling seems to play no role.

The solution blows-up in finite time (either a flat portion or a highly concentrated region appears; see, resp., Figures 4.15 and 4.16). Visualized in Wasserstein distance (namely, the square root of the second-momentum), the convergence to the Dirac mass located at zero seems to be linear in time (see Figure 4.17), as it should be from the theoretical viewpoint. However, the computed distance does not reach zero in finite

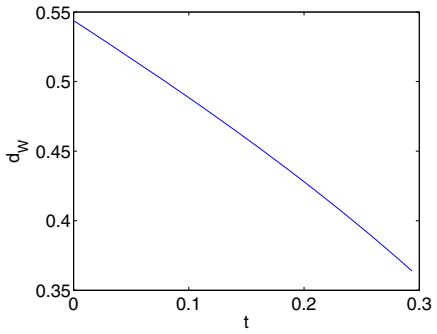


FIG. 4.17. *Wasserstein distance to the Dirac mass at zero δ_0. Blow-up occurs previously, and part of the mass has not yet reached zero at this time.*

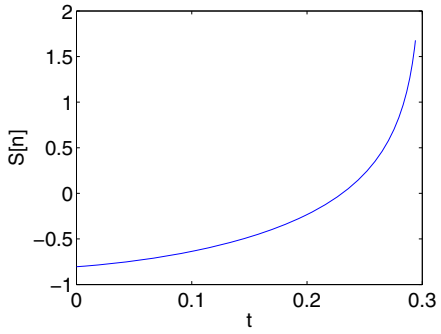


FIG. 4.18. *The variation of the entropy $\mathcal{S}[n]$ seems to blow-up.*

time. This is not surprising, because when blow-up occurs, part of the mass is still away from the blow-up point (here, zero). In order to see some vanishing distance, one can normalize the process in the following away: localize the Wasserstein distance in the transport variable (L^2-distance for the cumulative distribution function), to capture only the final plateau. This plateau is known a priori because it is entirely determined by the ratio χ_c/χ. However, this does not provide any new insight into what happens after blow-up, and it is known from theoretical works that the behavior depends highly upon the regularization procedure [46, 47, 25].

Interestingly, numerics are able to track the blow-up phenomenon quite precisely, without mesh refinement. Indeed, if the space step is even uniform, the number of space points at the density level adapts to the highly concentrated (blow-up) regions, corresponding to plateaus (compare Figure 4.15 and Figure 4.16). This is the counterpart of the "moving support" observed in section 4.2.1. The evolution of the entropy is plotted in Figure 4.18.

4.3.2. Two symmetric peaks: Case $\chi = 3\,\pi$. Given the compactly supported initial data,

$$V_0^i = \frac{\exp\left[10\ (w_i - 0.5)\right] - 1}{\left[(w_i + 0.01)\,(1.01 - w_i)\right]^{1/4}}\ ,$$

we numerically solve the PKS equation in original variables on the time interval $[0, 1.3]$ with $\chi = 3\,\pi$.

The factor χ is supercritical but is less than $2\,\chi_c$. Then, according to the conjectures in [45, 46, 47], there should be only one blow-up point. The density first diffuses (see Figure 4.20) and then concentrates in a delta Dirac (see Figures 4.19 and 4.21).

4.3.3. Two symmetric peaks: Case $\chi = 5\,\pi$. Starting with the centered cumulative distribution function,

$$V_0^i = \frac{\exp\left[10\ (w_i - 0.5)\right] - 1}{\left[(w_i + 0.01)\,(1.01 - w_i)\right]^{1/4}}\ ,$$

corresponding to a two-peaks-like initial density, we numerically solve the PKS equation in original variables on the time interval $[0, 0.45]$ with $\chi = 5\,\pi$.

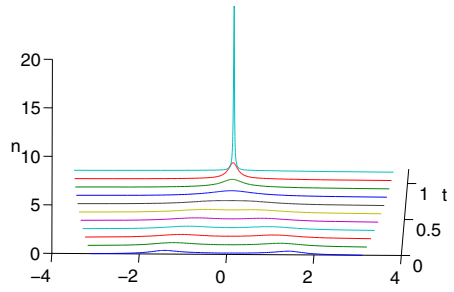


FIG. 4.19. *Cell density n for $\chi = 3\pi$ and two initial peaks.*

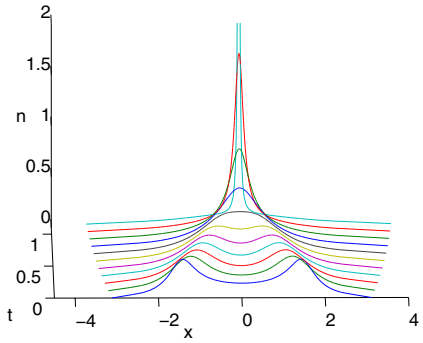


FIG. 4.20. *Zoom of Figure 4.19. Because both peaks do not contain enough mass to blow-up far from each other, they first merge, and then the solution blows-up.*

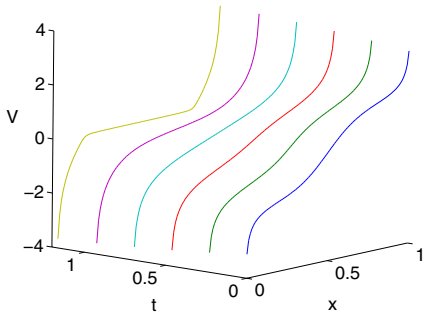


FIG. 4.21. *Cumulative function distribution function V for $\chi = 3\pi$ and two initial plateaus (that is, density peaks). The solution flattens into a single plateau.*

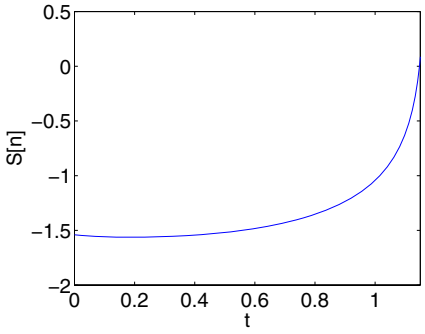


FIG. 4.22. *Evolution of the entropy.*

The initial condition is the same as in section 4.3.2, but χ is now bigger than $2\,\chi_c$. The blow-up occurs in two different points (see Figures 4.23 and 4.24). The entropy seems to blow-up; see Figure 4.22.

4.4. Two asymmetric peaks. Given the compactly supported initial data,

$$V_0^i = \frac{\exp\left[10\ (w_i - 0.45)\right] - 1}{\left[(w_i + 0.01)\,(1.01 - w_i)\right]^{1/4}}\ ,$$

we numerically solve the PKS equation in original variables on the time interval $[0, 1.1]$ with $\chi = 3\,\pi$. Note that the initial density is not centered, but it has no effect because Proposition 4.3 does not hold in this case.

When the parameter is between the critical parameter χ_c and twice the critical parameter $2\,\chi_c$, if the peaks are asymmetric, the blow-up occurs at the center of mass which is closer to the highest peak. The peaks diffuses, and then the density blows-up at the center of mass (see Figures 4.25 and 4.26).

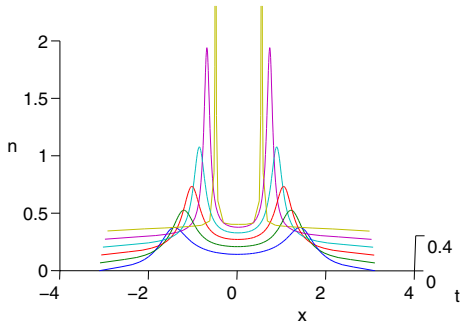


FIG. 4.23. *Cell density n for two initial peaks and $\chi = 5\pi$. In contrast to the previous section 4.3.2, each peak contains enough mass to blow-up itself.*

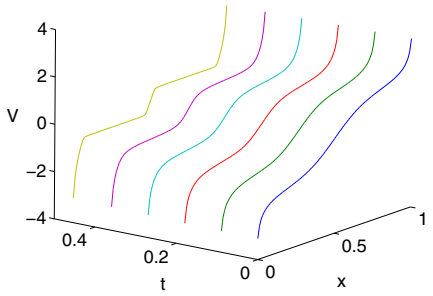


FIG. 4.24. *Cumulative distribution function V. Two distinct plateaus may appear when χ is above twice the critical parameters χ_c.*

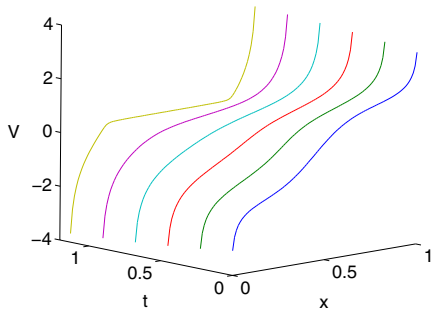


FIG. 4.25. *Cumulative distribution function V when $\chi \in (\chi_c, 2\,\chi_c)$ and initial data is a two-peaks-like density.*

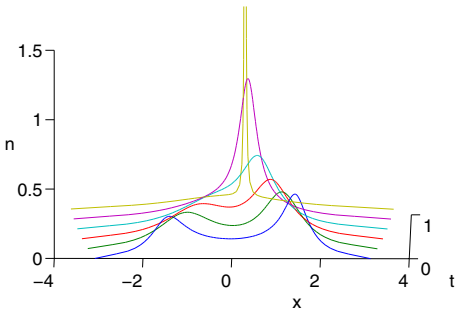


FIG. 4.26. *Evolution of the cell density n.*

Acknowledgment. We thank the Centre de Recerca Matemàtica (Barcelona) for providing an excellent atmosphere for research.

REFERENCES

[1] M. AGUEH, *Existence of solutions to degenerate parabolic equations via the Monge-Kantorovich theory*, Adv. Differential Equations., 10 (2005), pp. 309–360.

[2] M. AGUEH, N. GHOUSSOUB, AND X. KANG, *Geometric inequalities via a general comparison principle for interacting gases*, Geom. Funct. Anal., 14 (2004), pp. 215–244.

[3] L.A. AMBROSIO, N. GIGLI, AND G. SAVARÉ, *Gradient Flows in Metric Spaces and in the Space of Probability Measures*, Lectures in Math., Birkhäuser, Cambridge, MA, 2005.

[4] W. BECKNER, *Sharp Sobolev inequalities on the sphere and the Moser-Trudinger inequality*, Ann. of Math. (2), 138 (1993), pp. 213–242.

[5] D. BENEDETTO, E. CAGLIOTI, AND M. PULVIRENTI, *A kinetic equation for granular media*, RAIRO Modél. Math. Anal. Numér., 31 (1997), pp. 615–641.

[6] D. BENEDETTO, E. CAGLIOTI, J.A. CARRILLO, AND M. PULVIRENTI, *A non-Maxwellian steady distribution for one-dimensional granular media*, J. Statist. Phys., 91 (1998), pp. 979–990.

[7] A. BLANCHET, J.A. CARRILLO, AND N. MASMOUDI, *Infinite time aggregation for the critical Patlak–Keller–Segel model in $\mathbb{R}^2$*, Comm. Pure Appl. Math., to appear.

[8] A. BLANCHET, J. DOLBEAULT, AND B. PERTHAME, *Two-dimensional Keller-Segel model: Optimal critical mass and qualitative properties of the solutions*, Electron. J. Differential Equations, 44 (2006).

[9] M. Bodnar and J.J.L. Velazquez, *An integro-differential equation arising as a limit of individual cell-based models*, J. Differential Equations, 222 (2006), pp. 341–380.
[10] Y. Brenier, *Polar factorization and monotone rearrangement of vector-valued functions*, Comm. Pure Appl. Math., 44 (1991), pp. 375–417.
[11] M. Brenner, P. Constantin, L. Kadanoff, A. Schenkel, and S. Venkataramani, *Diffusion, attraction and collapse*, Nonlinearity, 12 (1999), pp. 1071–1098.
[12] M. Burger and M. Di Francesco, *Large Time Behavior of Nonlocal Aggregation Models with Nonlinear Diffusion*, preprint.
[13] V. Calvez and J.A. Carrillo, *Volume effects in the Keller-Segel model: Energy estimates preventing blow-up*, J. Math. Pures Appl., 86 (2006), pp. 155–175.
[14] V. Calvez, B. Perthame, and M. Sharifi tabar, *Modified Keller-Segel system and critical mass for the log interaction kernel*, in Nonlinear Partial Differential Equations and Related Analysis, Contemp. Math. 429, AMS, Providence, RI, 2007.
[15] E. Carlen and M. Loss, *Competing symmetries, the logarithmic HLS inequality and Onofri's inequality on S^n*, Geom. Funct. Anal., 2 (1992), pp. 90–104.
[16] J.A. Carrillo, M.P. Gualdani, and G. Toscani, *Finite speed of propagation for the porous medium equation by mass transportation methods*, C. R. Math. Acad. Sci. Paris Ser. I, 338 (2004), pp. 815–818.
[17] J.A. Carrillo, A. Jüngel, P.A. Markowich, G. Toscani, and A. Unterreiter, *Entropy dissipation methods for degenerate parabolic problems and generalized Sobolev inequalities*, Monatsh. Math., 133 (2001), pp. 1–82.
[18] J.A. Carrillo, R.J. McCann, and C. Villani, *Kinetic equilibration rates for granular media and related equations: Entropy dissipation and mass transportation estimates*, Rev. Mat. Iberoamericana, 19 (2003), pp. 1–48.
[19] J.A. Carrillo, R.J. McCann, and C. Villani, *Contractions in the 2-Wasserstein length space and thermalization of granular media*, Arch. Ration. Mech. Anal., 179 (2006), pp. 217–263.
[20] J.A. Carrillo and G. Toscani, *Wasserstein metric and large-time asymptotics of nonlinear diffusion equations*, in New Trends in Mathematical Physics (In Honour of the Salvatore Rionero 70th Birthday), World Scientific, Hackensack, NJ, 2004, pp. 234–244.
[21] J.A. Carrillo and G. Toscani, *Contractive probability metrics and asymptotic behavior of dissipative kinetic equations*, Riv. Mat. Univ. Parma, 6 (2007), pp. 75–198.
[22] D. Cordero-Erausquin, W. Gangbo, and C. Houdre, *Inequalities for generalized entropy and optimal transportation*, in Recent Advances in the Theory and Applications of Mass Transport, Contemp. Math. 353, AMS, Providence, RI, 2004, pp. 73–94.
[23] J. Denzler and R. McCann, *Fast diffusion to self-similarity: Complete spectrum, long-time asymptotics and numerology*, Arch. Ration. Mech. Anal., 175 (2004), pp. 301–342.
[24] J. Dolbeault and B. Perthame, *Optimal critical mass in the two-dimensional Keller-Segel model in $\mathbb{R}^2$*, C. R. Math. Acad. Sci. Paris, 339 (2004), pp. 611–616.
[25] J. Dolbeault and C. Schmeiser, *The Two-Dimensional Keller-Segel Model after Blow-up*, preprint.
[26] J. Duoandikoetxea and E. Zuazua, *Moments, masses de Dirac et décomposition de fonctions*, C. R. Acad. Sci. Paris Sér. I Math., 315 (1992), pp. 693-698.
[27] L.C. Evans, O. Savin, and W. Gangbo, *Diffeomorphisms and nonlinear heat flows*, SIAM J. Math. Anal., 37 (2005), pp. 737–751.
[28] F. Filbet, *A finite volume scheme for the Patlak-Keller-Segel chemotaxis model*, Numer. Math., 104 (2006), pp. 457–488.
[29] L. Gosse and G. Toscani, *Identification of asymptotic decay to self-similarity for one-dimensional filtration equations*, SIAM J. Numer. Anal., 43 (2006), pp. 2590–2606.
[30] L. Gosse and G. Toscani, *Lagrangian numerical approximations to one-dimensional convolution-diffusion equations*, SIAM J. Sci. Comput., 28 (2006), pp. 1203–1227.
[31] W. Jäger and S. Luckhaus, *On explosions of solutions to a system of partial differential equations modelling chemotaxis*, Trans. Amer. Math. Soc., 329 (1992), pp. 819–824.
[32] R. Jordan, D. Kinderlehrer, and F. Otto, *The variational formulation of the Fokker–Planck equation*, SIAM J. Math. Anal., 29 (1998), pp. 1–17.
[33] E.F. Keller and L.A. Segel, *Initiation of slide mold aggregation viewed as an instability*, J. Theoret. Biol., 26 (1970), pp. 399–415.
[34] H. Li and G. Toscani, *Long-time asymptotics of kinetic models of granular flows*, Arch. Ration. Mech. Anal., 172 (2004), pp. 407–428.
[35] R.J. McCann, *Existence and uniqueness of monotone measure-preserving maps*, Duke Math. J., 80 (1995), pp. 309–323.
[36] R.J. McCann, *A convexity principle for interacting gases*, Adv. Math., 128 (1997), pp. 153–179.

[37] R.J. McCann, *Polar factorization of maps on Riemannian manifolds*, Geom. Funct. Anal., 11 (2001), pp. 589–608.
[38] T. Nagai, *Blow-up of radially symmetric solutions to a chemotaxis system*, Adv. Math. Sci., 5 (1995), pp. 581–601.
[39] F. Otto, *The geometry of dissipative evolution equations: The porous medium equation*, Comm. Partial Differential Equations, 26 (2001), pp. 101–174.
[40] C.S. Patlak, *Random walk with persistence and external bias*, Bull. Math. Biophys., 15 (1953), pp. 311–338.
[41] F. Poupaud, *Diagonal defect measures, adhesion dynamics and Euler equation*, Methods Appl. Anal., 9 (2002), pp. 533–561.
[42] T. Senba and T. Suzuki, *Weak solutions to a parabolic-elliptic system of chemotaxis*, J. Funct. Anal., 191 (2002), pp. 17–51.
[43] C.M. Topaz, A.L. Bertozzi, and M.A. Lewis, *A nonlocal continuum model for biological aggregation*, Bull. Math. Biol., 68 (2006), pp. 1601–1623.
[44] G. Toscani, *One-dimensional kinetic models of granular flows*, RAIRO Modél. Math. Anal. Numér., 34 (2000), pp. 1277–1291.
[45] J.J.L. Velázquez, *Stability of some mechanisms of chemotactic aggregation*, SIAM J. Appl. Math., 62 (2002), pp. 1581–1633.
[46] J.J.L. Velázquez, *Point dynamics in a singular limit of the Keller–Segel model* 1: *Motion of the concentration regions*, SIAM J. Appl. Math., 64 (2004), pp. 1198–1223.
[47] J.J.L. Velázquez, *Point dynamics in a singular limit of the Keller–Segel model* 2: *Formation of the concentration regions*, SIAM J. Appl. Math., 64 (2004), pp. 1224–1248.
[48] C. Villani, *Topics in Optimal Transportation*, Grad. Stud. Math. 58, AMS, Providence, RI, 2003.
[49] C. Villani, *Mathematics of granular materials*, J. Statist. Phys., 124 (2006), pp. 781–822.
[50] C. Villani, *Optimal Transport, Old and New*, Lecture Notes for the 2005 Saint-Flour summer school, Springer, 2007, to appear.

SIAM J. NUMER. ANAL.
Vol. 46, No. 2, pp. 722–746
© 2008 Society for Industrial and Applied Mathematics

ADAPTIVE MULTIPRECISION PATH TRACKING*

DANIEL J. BATES†, JONATHAN D. HAUENSTEIN‡, ANDREW J. SOMMESE‡, AND CHARLES W. WAMPLER II§

Abstract. This article treats numerical methods for tracking an implicitly defined path. The numerical precision required to successfully track such a path is difficult to predict a priori, and indeed it may change dramatically through the course of the path. In current practice, one must either choose a conservatively large numerical precision at the outset or rerun paths multiple times in successively higher precision until success is achieved. To avoid unnecessary computational cost, it would be preferable to adaptively adjust the precision as the tracking proceeds in response to the local conditioning of the path. We present an algorithm that can be set to either reactively adjust precision in response to step failure or proactively set the precision using error estimates. We then test the relative merits of reactive and proactive adaptation on several examples arising as homotopies for solving systems of polynomial equations.

Key words. homotopy continuation, numerical algebraic geometry, polynomial systems, Bertini

AMS subject classifications. Primary, 65H10; Secondary, 65H20, 65G50, 14Q99

DOI. 10.1137/060658862

1. Introduction.

Path tracking is the task of tracing out a one-real-dimensional solution curve described implicitly by a system of equations, typically n equations in $n+1$ variables, given an initial point on, or close to, the path. This can arise in many ways, but our motivation is the solution of systems of polynomials via homotopy continuation (see [1, 9, 10, 13, 14, 21]). In this method, to find the isolated solutions of the system $f(z) = 0$ for given polynomials $f : \mathbb{C}^n \to \mathbb{C}^n$, one constructs a homotopy, $H(z,t)$, $H : \mathbb{C}^n \times \mathbb{C} \to \mathbb{C}^n$ such that $H(z,0) = f(z)$ is the target system to be solved, while $H(z,1)$ is a starting system whose isolated solutions are known. There is a well-developed theory on how to construct such homotopies to guarantee, with probability one, that every isolated solution of $f(z) = 0$ is the endpoint in the limit as $t \to 0$ of at least one smooth path $z_i(t)$, where $H(z_i(t),t) = 0$ on $t \in (0,1]$ and where $z_i(1)$, $i = 1, 2, \ldots$, are the known isolated solutions of $H(z,1) = 0$. Similar constructions arise in other contexts, where the existence of a path leading to the desired solutions may or may not be guaranteed. Even when there is no guarantee, experience shows that in some application domains, continuation techniques yield solutions more reliably than Newton's method, especially when good initial guesses are not available. While in our applications the path is just a means of arriving at the endpoint, in other applications one may desire to accurately trace out the path itself, such as when plotting the response of a mathematical model as one of its parameters

*Received by the editors May 3, 2006; accepted for publication (in revised form) July 9, 2007; published electronically February 8, 2008. This work was supported by NSF grant DMS-0410047 and the Institute for Mathematics and Its Applications in Minneapolis (IMA).
http://www.siam.org/journals/sinum/46-2/65886.html

†Institute for Mathematics and its Applications, University of Minnesota, Minneapolis, MN 55122 (dbates1@nd.edu, www.nd.edu/∼dbates1). This author was supported by the Duncan Chair of the University of Notre Dame and the Arthur J. Schmitt Foundation.

‡Department of Mathematics, University of Notre Dame, Notre Dame, IN 46556 (jhauenst@nd.edu, www.nd.edu/∼jhauenst sommese@nd.edu, www.nd.edu/∼sommese). The second and third authors were supported by the Duncan Chair of the University of Notre Dame.

§General Motors Research and Development, Mail Code 480-106-359, 30500 Mound Road, Warren, MI 48090 (Charles.W.Wampler@gm.com, www.nd.edu/∼cwample1).

is varied.

The most common path tracking algorithms are predictor-corrector methods: from an approximate solution point on the path, a predictor gives a new approximate point a given step size along the path, then a corrector brings this new point closer to the path. For example, one may use an Euler predictor, which steps ahead along the tangent to the path, or a higher order predictor that uses one or more recent points and possibly derivatives of the homotopy function at them to extrapolate to the predicted point. Typically, the prediction is then used as the initial point for correction by Newton's method. Since the solution set is one-dimensional, an extra constraint is introduced to isolate the target of the correction. For general homotopies, a useful constraint is to find where the solution path intersects the hyperplane normal to the last computed tangent direction. In the more restrictive setting of polynomial systems, the homotopy can be designed such that the paths advance monotonically with t, that is, there are no turning points, in which case it is acceptable (and simpler) to perform corrections by holding t fixed. The adaptive precision algorithm we discuss here is compatible with any of these prediction and correction schemes.

For good results, the predictor step size must be chosen appropriately. Too large a step size may result in a prediction outside the zone of convergence of the corrector, while too small a step size means progress is slow and computationally costly. Consequently, it has long been recognized that adaptive control of the step size is crucial for obtaining good reliability without undue computational cost.

While step size control is well established, less attention has been paid to efficient handling of precision. With wider availability of software packages for higher precision arithmetic, along with faster computers to execute the software, it becomes interesting to consider how adjustable precision might be deployed to improve the performance of path tracking algorithms. The issue at stake is analogous to step size control: without enough precision, path tracking might fail, but the use of excessive precision is inefficient. To address this trade-off, this paper proposes an algorithm that dynamically adjusts the number of digits used in computations according to the evolution of the numerical conditioning of the homotopy function.

In our primary application of interest, the solution of polynomial systems, there are several factors driving the need for higher precision. It is well known that high degree polynomials often lead to ill-conditioned problems. When treating polynomial systems in several variables, the total degree of the system, being the product of the degrees of the individual equations, quickly becomes large even for low degree polynomials, which can also lead to ill-conditioning. Thus, one driving force is the desire to solve larger systems of higher total degree.

A second motivation is that our systems often have some, or possibly many, singular solutions, and thus, the solution paths leading to these solutions are necessarily ill-conditioned near the end. While endgame methods exist for enhancing the accuracy with which such endpoints can be estimated, for singularities of high enough multiplicity, more precision is required.

A third motivation is that many polynomial systems from applications come with natural parameters. These systems with the parameters included typically have exact coefficients, which are relatively small integers. For "general values" of the parameters the structure of the solution set of the systems are the same [15], i.e., the number of multiplicity k points for $k = 1, 2, 3, \ldots$ remains constant. If the coefficients of the polynomials are computed to sufficient accuracy using the given choice of parameter values, the solutions of the system will typically have a clean structure for multiple

points, i.e., they will not have a cluster of solutions near each other unless the points computed more precisely converge to the same point. Here it is important that the coefficients are computed accurately as functions of the parameters. For example, in section 5.5, the polynomial system has coefficients computed accurately to 308 digits. The coefficients are polynomial functions of certain lengths, e.g., offsets and link lengths, and of cosines and sines of what are called twist angles. If only eight digits are used for the coefficients, as printed in the table in [14], then the algebraic relationships that should exist between the coefficients fail to hold to more than eight digits. For example, some coefficients depend on $\cos\alpha$ and $\sin\alpha$, where α is a twist angle, and when these are rounded to eight digits, the algebraic relation $\cos^2(\alpha)+\sin^2(\alpha) = 1$ no longer holds at higher precision. The effect of this is that the multiple points, e.g., the ones of multiplicity 24, form clusters that do not come together as precision is increased.

Finally, although the homotopy constructions guarantee, with probability one, that no path passes exactly through a singularity before reaching its endpoint, there is always a chance that a near-singular condition can be encountered. To obtain the highest reliability possible, we need to detect this and allocate sufficient digits to successfully track past such obstructions.

The paper is organized as follows. In section 2, we review the behavior of Newton's method in floating point, revealing how its accuracy and convergence properties depend on precision. In section 3, we discuss path tracking with adaptive step size control and identify how it fails when precision is insufficient. This leads, in section 4, to a novel technique for path tracking using adaptive precision. This new adaptive precision path tracking algorithm has been implemented in a software package, Bertini [2], currently under development by the authors. Several examples are presented in section 5 to illustrate the usefulness of adaptive precision. Section 6 contains a brief discussion of the value of adaptive precision in the presence of endgames. Finally, in section 7, a few related ideas that would make interesting studies are discussed.

2. Background: Newton's method. The core numerical process in the path tracker is the corrector, which in our case is Newton's method. A good predictor speeds up the path tracker by allowing a large step while still supplying an initial guess within the convergence region of the corrector. However, it is the loss of convergence that causes path tracking to fail. In exact arithmetic, as long as the path remains nonsingular, there must be a region surrounding the path within which Newton's method converges quadratically. With a small enough step Δt in t, we can be assured of advancing along the path, although possibly very slowly. This holds even if we use only a zeroth order predictor, i.e., if the point from the last value t_k is used to initialize the corrector for the new value $t_{k+1} = t_k + \Delta t$. In contrast, in inexact floating point arithmetic, the convergence region can disappear, thus halting the path tracker. Short of this, an unacceptably slow linear rate of convergence might dominate, causing the step size to plummet. It can also happen that the corrector appears to converge, judging by a diminishing corrector step, but since the limit point is a zero of the inexactly evaluated function, it may be further than the desired tolerance away from the exact root. Consequently, success requires enough precision to produce a convergence region and to ensure that the limit point is within the prescribed tolerance of the exact root.

Due to these considerations, an analysis of Newton's method in floating point is of interest and will help us derive rules for setting the precision used in our path tracker. The following analysis resembles that of [22]. Let $F(z) : \mathbb{C}^n \to \mathbb{C}^n$ be continuously differentiable, and denote its Jacobian matrix of partial derivatives as $J(z)$. To solve

$F(z) = 0$ by Newton's method given an initial guess z_0, one iteratively updates z_i, $i = 1, 2, \ldots$, as

$$\begin{aligned} &\text{Solve } J(z_i)\Delta z_i = -F(z_i) \text{ for } \Delta z_i, \\ &z_{i+1} = z_i + \Delta z_i. \end{aligned} \tag{1}$$

In the following analysis, we will assume a submultiplicative matrix norm, i.e., one that satisfies $\|AB\| \leq \|A\|\|B\|$. In particular, the 2-norm suffices. When we apply the formulas, we may save computation by using other norms with an adjustment factor. For example, for an $n \times n$ matrix A, $\|A\|_2 \leq \sqrt{n}\|A\|_p$ for $p = 1$ or $p = \infty$.

Suppose that we work in floating point with unit roundoff u. In other words, if we compute with a mantissa of b bits in binary or with P digits in decimal, $u = 2^{-b} = 10^{-P}$. We may consider evaluating the residuals $F(z_i)$ in higher precision, say $\bar{u} \leq u$, before rounding back to working precision u to compute the Newton correction. Let $\hat{F}(z)$ be the floating point output of the procedure that evaluates $F(z)$. We assume that there exists a function ψ depending on z, u, and $\bar{u}$ such that the error $e(z) = \hat{F}(z) - F(z)$ obeys

$$\|e(z)\| \leq u\|F(z)\| + \psi(z, u, \bar{u}). \tag{2}$$

By definition, at a solution point z_*, we have $F(z_*) = 0$, so it is clear that the function ψ drives the final error. To determine ψ, one must examine the function F and the program that implements it. We will give approximations to ψ later for the systems we treat.

In solving (1) for the correction Δz_i, there is error in evaluating $J(z_i)$ and in solving the linear system. Both errors can be absorbed into an error matrix E_i such that the computed correction is

$$\Delta z_i = (J(z_i) + E_i)^{-1}(F(z_i) + e(z_i)). \tag{3}$$

We assume this error is bounded by

$$\|E_i\| \leq \mathcal{E}\,(u\|J(z_i)\| + \phi(z_i, u)) \tag{4}$$

for some constant $\mathcal{E} > 1$ and positive function ϕ. We expect the first term because of roundoff of the Jacobian, whereas ϕ accounts for errors in evaluating J that do not vanish even when J does. The constant $\mathcal{E}$ accounts for the subsequent growth in the error during the linear solve.

For simplicity of notation, let $v = z_i$ be the current guess, $\bar{v} = z_{i+1}$ the new guess after a single iteration, and v_* the solution point near v. Also, let's use the shorthand notation $F = F(v)$, $J = J(v)$, $J_* = J(v_*)$, $\Delta = \|v - v_*\|$, and $\bar{\Delta} = \|\bar{v} - v_*\|$. In the next paragraph, we will establish a bound on $\bar{\Delta}$ in terms of Δ. Whenever $\bar{\Delta} < \Delta$, the Newton step successfully reduces the error in the estimate of the root v_*.

Since $F(v_*) = 0$, the Taylor series of $F(z)$ at v_* gives

$$F(z) = J_* \cdot (z - v_*) + H.O.T.,$$

where the higher order terms, $H.O.T.$, are quadratic or higher in $z - v_*$. Similarly,

$$J(z) = J_* + H.O.T.,$$

where the higher order terms are linear in $z - v_*$. Consequently, in a ball $B = \{z : \|z - v_*\| \leq R\}$ centered at v_* with $v \in B$, there exist positive constants α and β such that

$$\|F(z)\| \leq \|J_*\|\|z - v_*\| + \alpha\|z - v_*\|^2, \tag{5}$$

$$\|F(z) - J_*(z - v_*)\| \leq \alpha\|z - v_*\|^2, \tag{6}$$

$$\|J_*\| \leq \|J\| + \beta\|z - v_*\|, \qquad \|J - J_*\| \leq \beta\|z - v_*\|. \tag{7}$$

From these, we may conclude that

$$\|F(z)\| \leq \|J\|\Delta + (\alpha + \beta)\Delta^2. \tag{8}$$

In Newton's method, we solve

$$(J + E)d = -(F(v) + e) \tag{9}$$

for d and take the step

$$\bar{v} = v + d + \varepsilon, \tag{10}$$

where ε is the error in forming the sum. The standard model of round-off error in floating point addition [26] gives

$$\|\varepsilon\| \leq u(\|v\| + \|d\|) \leq u(\Delta + \|v_*\| + \|d\|), \tag{11}$$

so subtracting v_* from both sides of (10), we have

$$\bar{\Delta} \leq \|v - v_* + d\| + u(\Delta + \|v_*\| + \|d\|). \tag{12}$$

If J is nonsingular and $\|J^{-1}\|\|E\| < 1$, then $(J + E)$ is nonsingular, the Newton step is well defined, and

$$\|(J + E)^{-1}\| \leq K\|J^{-1}\|, \qquad K = \frac{1}{1 - \|J^{-1}\|\|E\|}. \tag{13}$$

Accordingly, from (9) we have

$$\|d\| \leq K\|J^{-1}\|(\|F\| + \|e\|). \tag{14}$$

Also, after adding $(J + E)(v - v_*)$ to both sides of (9) and simplifying using (5)–(8), (13), we have

$$\|v - v_* + d\| \leq K\|J^{-1}\|(\|E\|\Delta + (\alpha + \beta)\Delta^2 + \|e\|). \tag{15}$$

Substituting from (14), (15) into (12) and using (2), (4), (8), we obtain the bound

$$\begin{aligned}\bar{\Delta} &\leq K\|J^{-1}\|(1 + u)^2(\alpha + \beta)\Delta^2 \\ &\quad + (K\|J^{-1}\|\,[(2 + \mathcal{E} + u)u\|J\| + \mathcal{E}\phi] + u)\Delta + K\|J^{-1}\|(1 + u)\psi + u\|v_*\|.\end{aligned} \tag{16}$$

This relation holds as long as $\|J^{-1}\|\|E\| < 1$, so that the linear solve in the Newton step is well defined, and v is in the ball B, so that (5)–(7) hold.

If $\bar{\Delta} < \Delta$, the Newton step reduces the error in the approximation of the root. In exact arithmetic, we have $u = \phi = \psi = 0$ and $K = 1$, so $\bar{\Delta} \leq \|J^{-1}\|(\alpha + \beta)\Delta^2$. The error contracts if the initial guess is accurate enough so that $\Delta < 1/(\|J^{-1}\|(\alpha + \beta))$. If we also have $\Delta < 1/(\|J_*^{-1}\|(\alpha + \beta))$, it is clear that all subsequent iterates are nonsingular and contractive, from which one has the well-known result that Newton's method converges quadratically to a nonsingular solution for a sufficiently accurate initial guess. One sees that the more singular the Jacobian is at the root, the slower the convergence and the smaller the convergence zone.

In floating point arithmetic, we cannot expect the error to converge to zero. From (16), one may expect the error to contract until

$$\Delta \approx (1 + u)K\|J^{-1}\|\psi + u\|v_*\| \approx K\|J_*^{-1}\|\psi + u\|v_*\|. \tag{17}$$

The second term is the error inherent in representing v_* in floating point. The first term depends on the accuracy, ψ, with which the function is evaluated. This can be reduced by using higher precision, $\bar{u}$, in the function evaluation per (2). The precision of the Jacobian and the linear solve do not affect the final error.

On the other hand, the precision of the Jacobian and the linear solve do affect convergence. Without enough precision, $\|J^{-1}\|\|E\|$ may approach or surpass 1, which means that the linear solve may fail due to singularity or may yield such an inaccurate step that the error diverges. Notice that $\|J^{-1}\|\|E\| = u\|J^{-1}\|\|J\| + \|J^{-1}\|\phi = u\kappa + \|J^{-1}\|\phi$, where $\kappa = \text{cond}(J)$. The first term, $u\kappa$, reflects the well-known result that in solving a linear system, floating point roundoff is magnified by the condition number of the matrix.

3. Step length control and failure. To produce an improved path tracking algorithm, it is useful to first examine a standard predictor-corrector algorithm to see why adaptive step size control generally succeeds when conditioning is mild and why it may fail when conditioning is adverse.

A simple and effective approach for step size control is to adjust the step size up or down according to the success or failure of a complete prediction-correction cycle. Suppose the homotopy function $H(z, t) = 0$ defines a one-dimensional nonsingular path $z(t)$. We are given a start point approximately on the path, $z_0 \approx z(t_0)$, an ending value of t, and a tolerance to which points on the path are to be found. Then, in brief, a predictor-corrector path tracker with adaptive step size control may be constructed as follows.

Initialize Select: an initial step size, s; the number of corrector iterations allowed per step, $N \geq 1$; the step adjustment factor, $a \in (0, 1)$; the step expansion integer, $M \geq 1$; and a minimum step size $s_{\min}$.

Predict Estimate a new point near the path whose distance from the current point is the step size s.

Correct Iteratively improve the new path point, constraining its distance from the prior path point. Allow at most N iterations to reach the specified tolerance.

On success If the tolerance is achieved:

- Update the current path point to be the newly found path point.
- If we have reached the final value of t, exit with *success*.
- If there have been M successes in a row, expand the step size to $s = s/a$.

On failure If the tolerance is not achieved:

- Cut the step size to $s = as$.
- If $s < s_{\min}$, exit with *failure*.

Loop Go back to **Predict**.

The key is to allow only a small number of iterations in the corrector, typically only $N = 2$ or $N = 3$. This forces the prediction to stay within a good convergence region surrounding the path. If a large number of iterations is allowed, a bad prediction might ultimately converge, but it may wander first and become attracted to a completely different path in the homotopy. Keeping N small, the step size adaptation slows down to negotiate sharp turns in the path and accelerates whenever the path is relatively straight. Properly implemented, this results in a robust and efficient path tracking algorithm.

We can be a bit more precise. Let us do so by specifically considering an Euler predictor with a Newton corrector. Both of these derive from the linearized local model of the path. The Taylor series at (z_1, t_1) is

$$H(z_1 + \Delta z, t_1 + \Delta t) = H(z_1, t_1) + \frac{\partial H}{\partial z}(z_1, t_1)\Delta z + \frac{\partial H}{\partial t}(z_1, t_1)\Delta t + H.O.T., \tag{18}$$

where the higher order terms, $H.O.T.$, are quadratic or higher in $(\Delta z, \Delta t)$. Ignoring the higher order terms and setting $H(z_1 + \Delta z, t_1 + \Delta t) = 0$, one has the basic Euler predictor and Newton corrector relations. These are a system of n equations in $n + 1$ unknowns; as long as the combined matrix $[\partial H/\partial z \ \partial H/\partial t]$ is rank n, there is a well-defined tangent direction and tracking may proceed. The predictor adds a constraint on the length of the step along the tangent, whereas corrector steps are constrained to move transverse to the tangent. The extra constraints are particularly simple in the case where $\partial H/\partial z$ is rank n, for then the path progresses monotonically in t, and the step can be controlled via the advance of t. Accordingly, one has a linear system to be solved for Δz:

$$\left[\frac{\partial H}{\partial z}(z_1, t_1)\right] \Delta z = -\left(H(z_1, t_1) + \frac{\partial H}{\partial t}(z_1, t_1)\Delta t\right). \tag{19}$$

For prediction, we set $\Delta t = s$, the current step size, and for correction, we set $\Delta t = 0$.

Since the neglected terms are quadratic, the prediction error is order $\mathcal{O}(s^2)$. Thus, in the case of a failed step, cutting the step size from s to as reduces the prediction error by a factor of a^2. In this way, cuts in the step size quickly reduce the prediction error until it is within the convergence region of the corrector. With a kth order predictor, the prediction error scales as a^{k+1}, potentially allowing larger step sizes. In any case, the adaptive approach quickly settles to a step size s just small enough so that the corrector converges, while the next larger step of s/a fails. With $a = 1/2$ and $M = 5$, the step size adapts to within a factor of 2 of its optimum, with an approximate overhead of 20% spent checking if a larger step size is feasible.

Failure of path tracking with an adaptive step size can be understood from the discussion of Newton's method in section 2. For small enough initial error and exact arithmetic, the Newton corrector gives quadratic convergence to a nonsingular root. Near a singularity, $\|J_*^{-1}\|$ is large, which can lead to a small quadratic convergence zone and a slower rate of quadratic convergence. Inexact arithmetic can further shrink the convergence zone, degrade the convergence rate from quadratic to linear, and introduce error into the final answer. From these considerations, we see that there are two ways for the adaptive step size path tracker to halt prematurely near a singularity.

1. The predictor is limited to a tiny step size to keep the initial guess within the convergence zone of the corrector. If this is too small, we may exceed the total computational cost allotted for the path.
2. The path may approach a point where the final error of the corrector is larger than the requested path tracking tolerance.

The first mode of failure can occur even with infinite precision, but degradation of the convergence properties with too low a precision increases the occurrence of this failure. The second mode of failure is entirely a consequence of lack of precision. By allocating enough precision, we can eliminate the second mode of failure and reduce the occurrence of the first mode. It is important to note that in some applications there is flexibility in the definition of the homotopy, which can be used to enlarge convergence zones and thereby speed up path tracking. For example, rescaling of the equations and variables can sometimes help. However, such maneuvers are beyond the scope of this paper, which concentrates only on tracking the path of a given homotopy.

4. Adaptive precision. The use of high precision can largely eliminate both types of path tracking failure identified above. However, high precision arithmetic is expensive, so it must be employed judiciously. One might be tempted to ratchet precision up or down in response to step failures as in the adaptive step size algorithm. This presents the difficulty that there is just one stimulus, step failure, and two possible responses, cut the step size or increase precision. In the following paragraphs, we outline two possible approaches for adapting both step size and precision.

Both our "reactive" adaptation algorithm and our "proactive" one are encompassed in the same flowchart, Figure 1. One may see that the heart of the method is still the adaptive step size tracker described in section 3. In the diagram, P is the number of digits of precision and $P_{\max}$ is the largest precision we allow. The step size of the tracker is Δt, and $\varepsilon(P)$ is a lower limit for it, depending on P, as discussed further below. The magnitude of the most recent Newton step in the corrector is $\|d\|$ and $10^{-\tau}$ is the accuracy to which we wish to track the path. The final value of t to which we wish to track is t_{f}. Limits on P and on the total number of steps guarantee that the algorithm always terminates, declaring failure if too much computation has occurred before reaching t_{f}.

The algorithm as presented in Figure 1 assumes that the final value t_{f} can be reached without encountering an exact singularity. The homotopies used for solving polynomial systems are constructed such that, with probability one, no singularities will occur until possibly at the end, say $t = t^*$, where the homotopy function $H(z, t^*) = F(z)$ is the target system. In this situation, one should not use the corrector at $t = t^*$. Instead, one may predict the endpoint from values of t near t^* and accept the predictions as accurate when two successive ones agree to the desired precision. Both the prediction to t^* and the prediction to the next point on the path near t^* benefit from the use of a more sophisticated predictor than Euler's method, accounting for the possible existence of a singularity at the endpoint. For example, endgames that estimate the winding number of the root and use it to compute a fractional power series can be very effective [16, 18]. Such endgames still use a path tracker to approach t^*. In this paper, the endgame itself is out of scope: we discuss only the tracking algorithm and consider the final value $t_{\mathrm{f}} \neq t^*$ to be the last point on the path required by the endgame.

4.1. Reactive adaptation. Reactive adaptation changes precision in response to the observed behavior of the tracking process. In particular, if path tracking fails at a given precision, one may try again at a higher precision. The simplest approach is to run the entire path in a fixed precision with adaptive reruns. That is, if the path tracking fails, one reruns it in successively higher precision until the whole path is tracked successfully or until limits in computing resources force termination. The advantage of this approach is that adaptation is completely external to the core path tracking routine. Thus, this strategy can be applied to any path tracker that enables requests

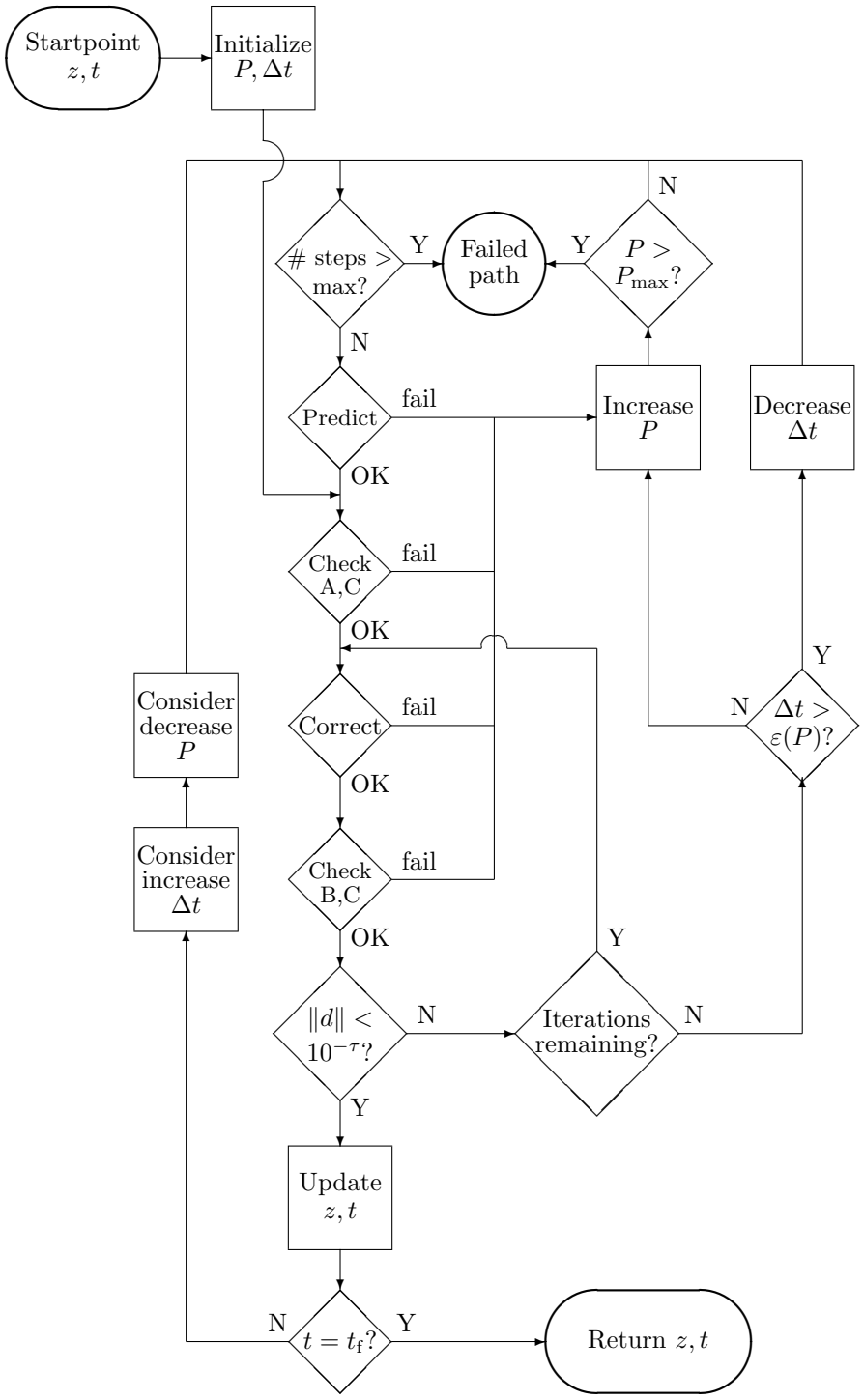


FIG. 1. *Adaptive precision path tracker with proactive precision checks. The purely reactive algorithm skips all checks of conditions A, B, and C. The purely proactive approach sets* $\varepsilon(P) \equiv 0$.

for higher precision. For example, in the polynomial domain, the package PHC [23] offers multiple precision, although the precision must be set when calling the program.

The path rerun strategy has two main disadvantages. First, when too low a precision is specified, the tracker wastes computation near the point of failure before giving up and initiating a rerun in higher precision. Second, and more important, the whole path is computed in high precision when it may be needed only in a small section of the path, often near the end in the approach to a singular solution.

A slightly more sophisticated treatment avoids recomputing the segment of the path leading up to the failure point. Upon failure at one precision, higher precision is initiated from the last successfully computed point on the path. In this approach, the adaptive step size module may still waste time cutting the step size before finally triggering a needed increase in precision. Moreover, the minimum step size needs to be set appropriately so that the precision is increased correctly. Once the precision has been increased, it may take many steps before the step size is increased again to a reasonable size. There is also a theoretical possibility that even though more precision is needed to accurately track the path, the tracking module may wander instead of failing, jumping to another solution path in the homotopy.

Figure 1 presents a flowchart that encompasses both the reactive adaptation method just described and the proactive one of the following section. For the reactive method, the checks labeled A, B, and C are skipped. There are three branches where the reactive method increases the precision, P: the prediction or correction steps may fail or the step size, Δt, may become too small. Failure in the predictor or corrector steps means that the linear solve of (19) has aborted early due to singularity. Using the magnitude of the largest entry in J as $\|J\|$, Gaussian elimination with row pivoting may declare such a failure when the magnitude of the largest available pivot is smaller than $u\mathcal{E}\|J\|$, for then the answer is meaningless.

The check for a step size that is too small is to compare Δt to $\varepsilon(P)$, a predetermined minimum step size for each level of precision. It is important to note that making $\varepsilon(P)$ too large may cause a runaway condition where precision is rapidly increased to its maximum, thus causing the path to declare failure. Setting it too small will result in temporary stalling of the path, wasting computation until the precision increase is finally triggered and wasting more computation afterwards as the algorithm has by then adopted an unnecessarily small step size. Our settings for $\varepsilon(P)$ are given in section 5.1. We determined by experimentation that values of $\varepsilon(P) = 10^{a-P}$ for $2 \le a \le 4$ seem to work well.

A path may have a difficult spot somewhere other than near the end of the path. Approaching this spot, the step size will decrease and the precision may increase. After passing the difficulty, the algorithm needs a means of readjusting back to a larger step size and smaller precision. For the box "Consider increase Δt," we use the criterion of the classic step size control method of section 3, that is, we increase Δt if several successive steps have succeeded. It is not clear at this time how to implement decreases in precision in a purely reactive manner. In contrast, the proactive approach of the following section gives a clear signal for when precision can be decreased.

4.2. Proactive adaptation. Instead of waiting for the adaptive step size method to fail before initiating higher precision, we propose to proactively monitor the conditioning of the homotopy to judge the level of precision needed at each step. In this way, the computational burden of higher precision is incurred only as needed, adjusting up and down as the tracker proceeds. If done properly, this avoids the computational cost incurred in the reactive approach of unnecessarily cutting the step

size when lack of precision is the root cause. Of course, monitoring the numerical properties of the homotopy introduces a new computational burden of its own. We will explore whether the trade-off is advantageous with respect to total computation, but even if the trade-off is in favor of the reactive approach, the proactive approach could still be preferable to avoid wandering due to lack of precision.

To decide how much precision is needed, we turn to the analysis of Newton's method from section 2. We wish to ensure that the achievable accuracy is within the specified tolerance and that convergence is fast enough.

In what follows, we need to evaluate $\|J\|$ and $\|J^{-1}\|$. These do not need to be very accurate, as we will always include safety margins in the formulas that use them. $\|J\|$ is readily available in the max norm, i.e., the matrix norm induced by the infinity norm for vectors. $\|J^{-1}\|$ is more difficult, as we do not wish to compute the full inverse of the matrix. This issue has been widely studied in terms of estimating the condition number $\kappa = \|J\|\|J^{-1}\|$. A relatively inexpensive method, suggested in [25] and elsewhere, is to choose a unit vector b at random and solve $Jy = b$ for y. Then we use the estimate $\|J^{-1}\| \approx \|y\|$. Although this underestimates $\|J^{-1}\|$, tests of matrices up to size 10×10 show the approximation to be reliably within a factor of 10 of the true value, which is easily absorbed into our safety margins.

One requirement is that $\|J^{-1}\|\|E\|$ should be small enough to ensure that the error-perturbed Jacobian is nonsingular. Minimally, we require $\|J^{-1}\|\|E\| < 1$, but by requiring it to be a bit smaller, say $\|J^{-1}\|\|E\| < 10^{-\sigma_1}$ for some $\sigma_1 \geq 1$, we force $K \approx 1$, as defined in (13). This removes the growth of K as one possible source of failure. Suppose that the error function ϕ in (4) is of the form $\phi = \Phi u$. Then our first rule is to require

$$\|J^{-1}\|\mathcal{E}\left(\|J\| + \Phi\right) u < 10^{-\sigma_1}. \tag{20}$$

Using P decimal digits of arithmetic results in precision $u = 10^{-P}$, so we may restate this rule as

$$P > \sigma_1 + \log_{10}[\|J^{-1}\|\mathcal{E}\left(\|J\| + \Phi\right)]. \tag{A}$$

A second requirement is that the corrector must converge within N iterations, where we keep N small as in the usual adaptive step size algorithm, typically 2 or 3. Let us say that the tolerance for convergence is $\Delta = \|v - v_*\| < 10^{-\tau}$. Recall that in each step of Newton's method, we compute d and take the step $\bar{v} = v + d$. The best estimate available of the accuracy is $\Delta \approx \|d\|$, so we declare success when $\|d\| < 10^{-\tau}$. Suppose that after $i < N$ iterations this is not yet satisfied. We still have $N - i$ iterations to meet the tolerance, and we would like to be sure that a lack of precision does not prevent success. Pessimistically, we assume that the linear factor in Δ in (16) dominates the quadratic one and that the rate of convergence does not improve with subsequent iterations. We force $K \approx 1$, and we have $u \ll 1$. Including the same safety margin as before, 10^{σ_1}, the requirement becomes

$$\left[10^{\sigma_1}\|J^{-1}\|\left((2+\mathcal{E})u\|J\| + \mathcal{E}\phi\right) + u\right]^{N-i}\|d\| < 10^{-\tau}. \tag{21}$$

As before, let's assume $\phi = \Phi u$. Taking logarithms, the number of decimal digits of precision must satisfy

$$P > \sigma_1 + \log_{10}\left[\|J^{-1}\|\left((2+\mathcal{E})\|J\| + \mathcal{E}\Phi\right) + 1\right] + (\tau + \log_{10}\|d\|)/(N-i). \tag{B}$$

Since we apply this formula only when the tolerance is not yet satisfied, we have $\|d\| > 10^{-\tau}$, or equivalently, $\tau + \log_{10}\|d\| > 0$. This implies that between corrector

iterations, requirement B is always more stringent than A. However, we still use A outside the corrector, because $\|d\|$ is not then available.

Our third requirement is that the precision must be high enough to ensure that the final accuracy of the corrector is within the tolerance at full convergence. For this, (17) is binding, so including a safety margin of $10^{-\sigma_2}$ and using the norm of the current approximate solution, $\|v\|$, as the best available estimate of $\|v_*\|$, we require

$$\|J^{-1}\|\psi + u\|v\| < 10^{-\tau-\sigma_2}. \tag{22}$$

Suppose the error in evaluating the homotopy function is given by $\psi = \Psi\bar{u}$. If the function is evaluated in the same precision as the rest of the calculations, i.e., $\bar{u} = u$, we have the requirement

$$P > \sigma_2 + \tau + \log_{10}(\|J^{-1}\|\Psi + \|v\|). \tag{C}$$

If instead we evaluate the function to higher precision, say $\bar{u} = 10^{-P'} < u = 10^{-P}$, we have the dual criteria

$$P > \sigma_2 + \tau + \log_{10}\|v\|, \qquad P' > \sigma_2 + \tau + \log_{10}\|J^{-1}\| + \log_{10}\Psi. \tag{C$'$}$$

The effect of adding the two errors is absorbed into the safety factor σ_2.

Conditions A, B, and C (or C$'$) allow one to adjust the precision as necessary without waiting for the adaptive step size to fail. If necessary, the precision can even be increased between corrector iterations. An algorithm using these criteria is described by the flowchart in Figure 1. With the proactive checks in place, it is no longer useful to trigger reactive increases in precision due a to small step size. Accordingly, we may remove the step size check, which is equivalent to setting $\varepsilon(P) \equiv 0$.

Unlike in the reactive method, the proactive method has a clear signal for when it is safe to decrease precision after passing a locally ill-conditioned point on the path. In the "Consider decrease P" box of Figure 1, one may safely do so if the checks of criteria A, B, and C in the last tracking step indicate that a lower precision would have sufficed.

4.3. Error estimates. To use the foregoing procedures, we need the function evaluation error, ψ, and the errors contributing to E, namely, $\mathcal{E}$ and ϕ. There is a trade-off between using rigorously safe bounds for highest reliability or using less stringent figures reflecting typical behavior to avoid the overuse of high precision. Rough figures are acceptable as this is just a means of setting the precision.

While ψ and ϕ concern the errors in evaluating the function and its Jacobian, the factor $\mathcal{E}$ concerns the stability of the linear solve. Roundoff errors can accumulate through each stage of elimination. When Gaussian elimination with partial pivoting is used, the worst-case error bound grows as $\mathcal{E} = 2^n$ for solving an $n \times n$ system [3]. However, as indicated in [3], $\mathcal{E}$ rarely exceeds n with the average case around $n^{\frac{2}{3}}$ or $n^{\frac{1}{2}}$. Setting $\mathcal{E} = n^2$ should therefore be sufficient for almost all cases.

It is important to note that the error in the function depends on the error in its parameters. For example, consider the simple function $f(x) = x^2 - 1/3$. If this is rounded off to $g(x) = x^2 - 0.333$ before we use high precision to solve $g(x) = 0$, we will obtain an accurate value of $\sqrt{0.333}$ but we will never get an accurate value of $1/\sqrt{3}$. Although this is an obvious observation, it can easily be forgotten in passing a homotopy function from some application to the adaptive precision path tracking algorithm. If coefficients in the function are frozen at fixed precision, the algorithm

tracks the solutions of the frozen function, not the exact problem that was intended. Whether the difference is significant depends on the nature of the application and the sensitivity of the function.

In the remainder of this section, we discuss how to obtain error estimates for several classes of functions.

4.3.1. General functions. A user of the path tracker will not usually wish to expend a lot of effort in developing error bounds, so an automated error analysis is desirable. A rigorous and automated way of establishing error bounds for almost any continuous function is to use interval arithmetic. Software packages such as IntBis [5], IntLab [20], Alias [12], and MPFI [19] provide interval evaluation for any real-valued function composed from a wide range of continuous elementary functions and arithmetic operations. Among these packages IntLab and MPFI currently support multiple precision, although IntLab's multiple precision is slow while MPFI does not support complex arithmetic. While a relatively efficient complex multiprecision interval arithmetic could be built on top of MPFI, it is beyond our scope to pursue this.

One could go further and employ interval techniques to obtain a path tracker with fully rigorous step size control, as in [6], guaranteeing that path crossing never occurs. However, this can be expensive, due partially to the cost of interval arithmetic but more significantly due to the cost of overconservative error bounds, which slow the algorithm's progress by driving the step size smaller than necessary. Still, when rigorous results are desired, it may be worth the computational cost. The method of [6] does not explicitly include adaptive precision, so something along the lines discussed here could be useful in modifying that approach.

Since our main area of interest is in systems of polynomial equations, we devote the rest of this article to that important subclass.

4.3.2. Straight-line polynomials. Suppose $f : \mathbb{C}^n \to \mathbb{C}$ is a polynomial whose terms are indexed by the set $\mathcal{I}$, that is,

$$f(z) = f(z_1, \ldots, z_n) = \sum_{i \in \mathcal{I}} c_i z_1^{d_{1i}} \cdots z_n^{d_{ni}}. \tag{23}$$

We wish to estimate the error in evaluating $f(z)$ using floating point arithmetic with unit roundoff $\bar{u}$. Evaluating the fully expanded form of a polynomial, that is, evaluating each term of (23) and then summing, is not the most efficient approach in terms of operation counts. For example, to evaluate a dense polynomial in one variable, e.g., $f(x) = c_0x^3 + c_1x^2 + c_2x + c_3$, Horner's rule, $f(x) = (((c_0x) + c_1)x + c_2)x + c_3$, uses the fewest multiplications. See [6, section 4.6] for more discussion. Different evaluation procedures can result in different floating point roundoff errors as well, for example, in floating point arithmetic distributivity, $a(b + c) = ab + ac$, and associativity, $(a + b) + c = a + (b + c)$, do not hold exactly [6, section 4.2].

Suppose the program to evaluate $f(z)$ has been parsed into a straight-line program, that is, a sequence of unary and binary operations free of branches or loops. Then we may approximate errors by accumulating their effects across successive operations. Interval analysis by circular arithmetic [4] is a convenient method for this. Let $\langle a, r\rangle$, $a \in \mathbb{C}$, $r \in \mathbb{R}^+$, denote the closed disk in the complex plane with center a and radius r. For two disks A_1, A_2 and an operation "$\circ$," let

$$A_1 \circ A_2 = \{x_1 \circ x_2 | x_1 \in A_1, x_2 \in A_2\}.$$

We are interested in the cases when "$\circ$" is addition or multiplication. Then one may confirm the inclusions [4]

$$\begin{aligned} \langle a_1, r_1\rangle + \langle a_2, r_2\rangle &= \langle a_1 + a_2, r_1 + r_2\rangle, \\ \langle a_1, r_1\rangle \times \langle a_2, r_2\rangle &\subset \langle a_1 a_2, |a_1| r_2 + |a_2| r_1 + r_1 r_2\rangle. \end{aligned} \tag{24}$$

This shows that if a_1 and a_2 are the computed values of the operands whose exact values lie in the respective disks, then the true sum or product must lie within the disks on the right-hand side of these relations. However, when we evaluate $a_1 \circ a_2$ in floating point, roundoff error must be taken into consideration. By appropriately increasing the radius of the output disk, we can guarantee that it still contains the exact answer. Since the roundoff for any point in the disk is bounded by $|x|u$ where $|x|$ is the largest magnitude of any point in the disk, we have

$$\begin{aligned} \langle a_1, r_1\rangle + \langle a_2, r_2\rangle &\subset \langle a_1 + a_2, r_1 + r_2 + (|a_1 + a_2| + r_1 + r_2)u\rangle, \\ \langle a_1, r_1\rangle \times \langle a_2, r_2\rangle &\subset \langle a_1 a_2, \rho + (\rho + |a_1 a_2|)u\rangle, \\ \text{where } \rho &= |a_1| r_2 + |a_2| r_1 + r_1 r_2. \end{aligned} \tag{25}$$

These relations can be used to bound the accumulated floating point error in the straight-line evaluation of a polynomial. In computing the radius in either of these formulas in floating point, one should consistently use upward rounding to be sure of obtaining inclusion.

To apply proactive adaptation based on criteria A, B, and C (or C′), we would like to develop a linearized model of how the error depends on precision. That is, we would like to express the evaluation error $\psi(z, u)$ as $\psi(z, u) \le \Psi(z)u + H.O.T.$, where the higher order terms depend on quadratic and higher powers of u. To this end, suppose that the error radii of all the initial data are expressed in terms of the unit roundoff. For example, the exact number $2{+}3i$ is $\langle 2{+}3i, 0u\rangle$, whereas the floating point representation $(1/3)_{\text{fp}}$ of $1/3$ is inexact: $(1/3) \in \langle (1/3)_{\text{fp}}, (1/3)_{\text{fp}} u\rangle$. Applying (25) to each operation of the straight-line program and keeping u as an indeterminant, the result after each operation is a disk of the form $\langle a, r(u)\rangle$, where $a \in \mathbb{C}$ and $r(u)$ is a polynomial in u with nonnegative real coefficients. At the end of the computation, one obtains a disk $f(z) \in \langle \hat{f}(z), r_1(z)u + r_2(z)u^2 + H.O.T.\rangle$, where $\hat{f}(z)$ is the floating point evaluation of $f(z)$, and r_1 and r_2 are numeric coefficients. For a given unit roundoff u_0, if $r_2 u_0 \ll r_1$, then the linear term dominates the quadratic one and presumably all higher powers of u as well. Then we may safely use $\psi(z, u) \approx r_1(z)u$, i.e., $\Psi = r_1(z)$, in the adaptive precision routines.

The foregoing methodology computes the function value and the error bound simultaneously. Unfortunately, the error bound is at least as expensive to compute as the function value, which may be an unacceptable computational burden. A less burdensome approach is to compute an error bound over a region and update this only when the path tracking algorithm leaves that region. To do so, we again just apply (25), but we begin with a disk for each coordinate z_i of z, say $z_i \in \langle a_i, \rho_i\rangle$, $\rho_i \in \mathbb{R}^+$, $i = 1, \ldots, n$. The method proceeds as above, with the only difference being that the radii of the results of each operation have a constant coefficient. At the end, we have a disk $\langle \hat{f}(z), r_0(z, \rho) + r_1(z, \rho)u + r_2(z, \rho)u^2 + H.O.T.\rangle$ that contains all possible values that the polynomial may take on the region defined by the initial disks.

Throughout this section, we have spoken only of how to bound the roundoff error in function evaluation. It should be clear that the same approach applies to bound the

errors in the entries of the Jacobian matrix, since each of these is also a polynomial. After estimating the error in each entry, one may estimate the error $\phi(z,u)$ in (4) using the relations $\|A\|_2 \leq \sqrt{n}\,\|A\|_p$ for $p=1$ or $p=\infty$.

4.3.3. Approximations for expanded polynomials. To avoid program complexity and save computation time, it is preferable not to perform a complete interval analysis of errors. In many cases a rough analysis is sufficient and easily derived. This is indeed possible for the case of expanded polynomials.

Referring to (23), we denote the degree of the ith term as $d_i = \sum_{j=1}^{n} d_{ji}$. If the coefficient c_i is represented in floating point with unit roundoff $\bar{u}$, then the error in the floating point evaluation of the ith term is $\bar{u}(d_i+1)|c_i||z_1|^{d_{1i}}\cdots|z_n|^{d_{ni}}$, ignoring quadratic and higher terms in $\bar{u}$. (If c_i has an exact floating point representation, such as if it is an integer of magnitude smaller than $1/\bar{u}$, then a slightly smaller bound applies: $\bar{u}d_i|c_i||z_1|^{d_{1i}}\cdots|z_n|^{d_{ni}}$.) Summing the terms and rounding off to working precision u gives the error between the computed value $\hat{f}(z)$ and the true $f(z)$ of

$$e(z) = |\hat{f}(z) - f(z)| \leq u|f(z)| + \bar{u}\sum_{i\in\mathcal{I}}(d_i+1)|c_i||z_1|^{d_{1i}}\cdots|z_n|^{d_{ni}}. \tag{26}$$

Thus, comparing to (2), we have

$$\psi(z,\bar{u}) = \bar{u}\sum_{i\in\mathcal{I}}(d_i+1)|c_i||z_1|^{d_{1i}}\cdots|z_n|^{d_{ni}}. \tag{27}$$

Equation (27) is as computationally expensive to evaluate as $f(z)$. However, this cost is usually small compared to evaluating the Jacobian matrix and solving for the Newton correction. Even so, one might prefer to replace $\psi(z,\bar{u})$ with an upper bound that is cheaper to compute. Suppose $D = \max_{i\in\mathcal{I}} d_i$ and $|z|_\infty = \max_{j\in[1,n]}|z_j|$. Then clearly

$$\psi(z,\bar{u}) \leq \bar{u}|z|_\infty^D \sum_{i\in\mathcal{I}}(d_i+1)|c_i| = \bar{u}|z|_\infty^D S, \tag{28}$$

where S is the sum indicated in the intermediate expression. Since S does not depend on z, it can be computed once at the beginning of the path tracking algorithm.

For a system of polynomials, $F:\mathbb{C}^n\to\mathbb{C}^m$, we obtain an error estimate for each polynomial in the system, i.e., $\psi_j(z,\bar{u})$, $j=1,\ldots,m$. Then the 2-norm error of the system is simply $\psi(z,\bar{u}) = \|[\psi_j(z,\bar{u})]\|_2$, which if we use (28) becomes

$$\psi(z,\bar{u}) = \bar{u}|z|_\infty^D(S_1^2+\cdots+S_m^2)^{1/2},$$

where S_j is the sum appearing in (28) applied to the jth polynomial.

At first glance, it may seem that errors could be reduced by simply scaling the functions, and thereby scaling their coefficients, by some small factor. But $\|J_*^{-1}\|$ will scale oppositely, so the error predicted by (17) is unchanged.

4.3.4. Homogeneous polynomials. Homogeneous polynomials, in which every term has the same degree $D = d_i$, $i\in\mathcal{I}$, are a special case. Some problems are naturally homogeneous, but for those that are not, it is often convenient and numerically advantageous to homogenize the equations before solving the system. Any inhomogeneous polynomial $f(z_1,\ldots,z_n)$ can be easily homogenized to obtain a related function $F(x_0,x_1,\ldots,x_n)$, with

$$F(1,x_1,\ldots,x_n) = f(x_1,\ldots,x_n).$$

If f is as in (23), then

$$F(x) = \sum_{i \in \mathcal{I}} c_i x_0^{D-d_i} x_1^{d_{1i}} \cdots x_n^{d_{ni}}, \tag{29}$$

where D and d_i are as above. Hence, for any solution z_* of $f(z) = 0$ there is a corresponding solution $x_* = (1, z_*)$ of $F(x) = 0$.

One advantage of homogenization is that we can rescale any solution x_* of $F(x) = 0$ to make $\|x_*\| = 1$, which often helps numerical conditioning. More precisely, since $F(\lambda x) = \lambda^D F(x)$, we have that if $F(x) = 0$, then also $F(\lambda x) = 0$. Consequently, the solution set of a system of homogeneous polynomials can be said to lie in projective space $\mathbb{P}^n$, the set of lines through the origin in $\mathbb{C}^{n+1}$. (Similarly, the solutions of multihomogeneous polynomials lie in a product of projective spaces; see [21].) To numerically solve for a point $x_* \in \mathbb{P}^n$, we intersect the corresponding line in $\mathbb{C}^{n+1}$ with an inhomogeneous hyperplane $a^T x - 1 = 0$, where $a \in \mathbb{C}^{n+1}$ can be chosen arbitrarily excepting a such that $a^T x_* = 0$. In other words, $a^T x - 1 = 0$ defines a $\mathbb{C}^n$ patch on $\mathbb{P}^n$, and $a^T x = 0$ is the hyperplane at infinity for that patch. In practice, we may choose a at random, with a zero probability of picking it such that the one-real-dimensional homotopy path intersects the hyperplane at infinity. However, during path tracking, if $\|x\|$ should grow, one may reassign $a = x'/\|x\|_2$, where x' denotes the complex conjugate of x, so that the new representative of the same projective point is $x/\|x\|_2$. In this way, the magnitudes of both a and x can be kept near 1 throughout the homotopy without changing the underlying path in projective space.

Suppose we use homogenization to force each homogenized coordinate x_i, $i = 0, 1, \ldots, n$, to lie in a unit disk at the origin: $x_i \in \langle 0, 1 \rangle$. Then we may obtain in one initial computation an error bound that will hold throughout the entire path tracking procedure by applying the method of section 4.3.2. Similarly, since $\|x\|_\infty \leq 1$, (28) becomes

$$\psi(x, u) \leq u(D+1) \sum_{i \in \mathcal{I}} |c_i|, \tag{30}$$

or just $uD \sum_{i \in \mathcal{I}} |c_i|$ if the coefficients are exact. Similarly, the derivatives have an approximate error bound of

$$\phi(x, u) \approx uD(D+1) \sum_{i \in \mathcal{I}} |c_i|. \tag{31}$$

The downside is that this bound may be overly conservative.

The effect of homogenization on the interpretation of the desired tolerance should be noted. The homogeneous coordinates $x = [x_0, x_1, \ldots, x_n] \in \mathbb{P}^n$ correspond to dehomogenized coordinates $z = (z_1, \ldots, z_n) = (x_1/x_0, \ldots, x_n/x_0)$. Suppose each coordinate of x is computed to accuracy $\epsilon \ll 1$ and consider what this implies about the accuracy of z. Dehomogenization of the result approximates z_i as $\hat{z}_i = (x_i + \delta_i)/(x_0 + \delta_0)$, where $|\delta_i| \leq \epsilon$, $i = 0, \ldots, n$. If $\|z\|_\infty < 1$, our homogenization will compute it with $|x_0| \approx 1$ and so $|\hat{z}_i - z_i| \approx \delta_i + \delta_0 \leq 2\epsilon$. On the other hand, if $\|z\|_\infty > 1$, the largest coordinate of x, say x_k will be computed with magnitude 1, and dehomogenization gives $|\hat{z}_k - z_k| \approx |z_k|\delta_i + |z_k|^2\delta_0 \leq (|z_k| + |z_k|^2)\epsilon$. So small solutions are found to absolute error 2ϵ, whereas large ones are found to absolute error $\epsilon\|z\|_\infty^2$. This implies a total loss of precision for $\|z\|_\infty \geq 1/\epsilon$, as one must expect since $[\epsilon, 1] \in \mathbb{P}$ and $[0, 1] \in \mathbb{P}$ become indistinguishable to accuracy ϵ, yet they dehomogenize to $1/\epsilon$ and ∞, respectively.

TABLE 1
Number of digits for mantissas at selected levels of precision.

	IEEE double	MPFR						
Bits	52	64	96	128	192	256	512	1024
Decimal digits	16	19	28	38	57	77	154	308

TABLE 2
Values of $\varepsilon(P)$ for reactive adaptive precision.

P	52	64	96	128	160	192	224
$\varepsilon(P)$	10^{-14}	10^{-16}	10^{-25}	10^{-34}	10^{-43}	10^{-52}	10^{-61}

5. Computational results. This section contains a brief discussion of the implementation details for multiprecision arithmetic and for evaluating the rules for adapting precision. We then discuss the application of the adaptive precision methods to several polynomial systems.

5.1. Implementation details. Bertini [2] is a software package for computation in numerical algebraic geometry currently under development by the authors. Bertini is written in the C programming language and makes use of straight-line programs for the representation, evaluation, and differentiation of polynomials. All the examples discussed here were run on an Opteron 250 processor running Linux, using a version of Bertini similar to the beta release.

The adaptation rules, A, B, and C (or C$'$), leave some implementation choices open. For the runs reported here, we chose to evaluate function residuals to the same precision as the computation of Newton corrections, so rule C applied, not rule C$'$. Also, in rules A and B, we chose to use $\mathcal{E} = n^2$, where n is the number of variables, which is somewhat conservative for typical cases but underestimates the worst pathological cases. (See section 4.3 for more on this issue.) The rules require formulas for evaluating the error bounds $\psi(x,u) = \Psi u$ and $\phi(x,u) = \Phi u$. These are problem dependent, so we report our choices for each of the example problems below. Also, since the size of the first step is problem and path dependent, Bertini automatically adjusts this based on the results of the Euler prediction of the first step rather than relying on the user to provide an appropriate size.

To adaptively change precision, Bertini relies on the open source MPFR library for multiprecision support. Bertini has data types and functions for double precision (based on the IEEE double standard) and higher precision (using MPFR). Although the program would be simpler if MPFR data types and functions were used exclusively, the standard double precision types and functions in C are more efficient, so Bertini uses these whenever the adaptation rules indicate that double precision is sufficient. Additional details regarding the use of multiple precision may be found on the Bertini website. Since the use of adaptive precision variables is highly implementation-specific, no other details are described here.

MPFR requires the addition of precision to the mantissa in packets of 32 bits. Since the discussion of the examples below involves both binary and decimal digits, Table 1 provides the conversion between the two for selected levels of precision.

The reactive adaptation method requires a minimum step size, $\varepsilon(P)$, for each level of precision, P. The values we used in our experiments are shown in Table 2.

5.2. Numerical error for a Wilkinson polynomial. Before testing adaptive precision in a path tracking application, let's consider Newton's method applied to a Wilkinson polynomial and compare its results to the predictions of our error analysis. The kth Wilkinson polynomial, $w_k(z)$, is defined as $w_k(z) = \prod_{i=1}^{k}(z-k)$, and we report here on the specific case of $k = 11$. To solve $w_{11}(z) = 0$ numerically, we use straight-line programs to evaluate $w_{11}(z)$ and $J(z) = w'_{11}(z)$ and apply Newton's method. Let us consider two different straight-line programs for $w_{11}(z)$:

- the product form $f_1(z) = (z-1)(z-2)\cdots(z-11)$ evaluated left to right as written;
- the expanded form $f_2(z) = z^{11} - 66z^{10} + 1925z^9 - \cdots - 150917976z^2 + 120543840z - 39916800$, where powers of z are formed by multiplication, then the terms are summed from left to right.

In the expanded form, the coefficient of largest magnitude is that of the quadratic term, which is large, but still small enough to be represented exactly in double precision. (For large enough k, an error analysis in double precision would have to include roundoff error in the coefficients.)

Consider Newton's method in the vicinity of the root $z = 7$. For $z \in \langle 7, .1\rangle$ with exact coefficients, error analysis of the product form gives

$$f_1(\langle 7, .1\rangle) \in \langle 0, 2682 + 52492u + 488015u^2\rangle,$$

while the expanded form gives

$$f_2(\langle 7, .1\rangle) \in \langle 0, 1.2 \times 10^{11} + 9.4 \times 10^{12}u + 2.6 \times 10^{13}u^2\rangle.$$

In both cases, the quadratic term in u is negligible for all u under consideration. Now, the Jacobian at $z = 7$ is $J(7) = 17280$, so (17) predicts that Newton's method will converge to $\Delta \approx (7 + \Psi/17000)u$, where $\Psi = 53000$ for the product form and $\Psi = 9.4 \times 10^{12}$ for the expanded form. For double precision, $u \approx 2.2 \times 10^{-16}$, so we have $\Delta \approx 2.2 \times 10^{-15}$ for the product form and $\Delta \approx 1.2 \times 10^{-7}$ for the expanded form. As Newton's method converges towards the root, the bound Ψ may improve. For example, consider the tighter disk $z \in \langle 7, 1 \times 10^{-7}\rangle$ for which the error analysis gives

$$f_1(\langle 7, 1 \times 10^{-7}\rangle) \in \langle 0, 0.00173 + 0.035u + 0.33u^2\rangle,$$

while the expanded form gives

$$f_2(\langle 7, 1 \times 10^{-7}\rangle) \in \langle 0, 1.2 \times 10^{5} + 8.7 \times 10^{12}u + 2.3 \times 10^{13}u^2\rangle.$$

Due to the fact that $z = 7$ has an exact representation in floating point (this will not be the case for the roots of most polynomials one would solve numerically!), a better estimate of the final error for Newton's method is in this case just $\Delta \approx u\Psi/17000$. One finds that as the disk tightens around $z = 7$, $\Psi \to 0$ for the product form, and so we expect Newton's method to converge to the exact root. In contrast, for the expanded form of $w_{11}(z)$, Ψ remains at 8.7×10^{12}, implying that one cannot expect a final answer better than $\Delta \approx 1.1 \times 10^{-7}$ using double precision.

This is a good estimate of the behavior one observes by running Newton's method with initial guesses in the vicinity of $z = 7$. Using the product form to evaluate the polynomial and starting with an initial guess of $z = 7.03 + 0.07i$, one obtains errors $|z_i - 7|$ in the first five iterates of 1.96×10^{-3}, 1.42×10^{-6}, 7.36×10^{-13}, 1.54×10^{-25}, 0, at

TABLE 3
Estimates of Ψ and Φ for the Chebyshev problems of varying degrees.

n	Ψ	Φ
10	54	230
50	520,000	$1.4{\cdot}10^{7}$
100	$1.3{\cdot}10^{10}$	$7.2{\cdot}10^{11}$
150	$2.3{\cdot}10^{14}$	$2.0{\cdot}10^{16}$

which point the exact root has been found. For the expanded form and the same initial guess, the errors are 1.96×10^{-3}, 1.41×10^{-6}, 4.22×10^{-10}, 2.40×10^{-9}, 2.43×10^{-10}. The iterates dither around the exact root, staying within $|z_i - 7| < 3.6 \times 10^{-9}$ for $3 \leq i \leq 50$. This is somewhat better than the predicted residual of 1×10^{-7}, which is not surprising because the estimate of Ψ is conservative.

5.3. High degree Chebyshev polynomials. To illustrate our method's applicability to high degree polynomials, let's consider the Chebyshev polynomials, which have been studied extensively and are known to have many interesting properties [3]. There is one Chebyshev polynomial in each degree, and scaled so that the leading coefficient is 1; they may be defined recursively by

$$\begin{aligned} T_0(x) &:= 2, \\ T_1(x) &:= x, \text{ and} \\ T_i(x) &:= xT_{i-1}(x) - T_{i-2}(x)/4 \quad \text{for } i \geq 2. \end{aligned}$$

The roots of $T_{n+1}(x)$ are all nonsingular and given by

$$\cos\left(\frac{(2n+1-2k)\,\pi}{2n+2}\right)$$

for $k = 0, 1, \ldots, n$.

Several of these polynomials, with degrees ranging from 10 to 150, were solved using Bertini. The systems were run with a linear homotopy with a total-degree start system and without homogenizing. Bounds on Φ and Ψ were set using the method of section 4.3.2 and are displayed in Table 3. Since the solutions of the Chebyshev polynomials are known to be nonsingular, the paths were tracked to $t_{\mathrm{f}} = 0$ using basic homotopy continuation with a tolerance of 10^{-8}, i.e., $\tau = 8$, and both safety digits set to 1.

With this setup, double precision is sufficient for path tracking with $n = 10$. This allows for the demonstration of the current cost of higher precision relative to double precision as shown in Table 4. Notably, increasing from double precision (52 bits) to 64-bit multiple precision increases the computational time by a factor of 13.3, while increasing from 64 bits to 1024 bits causes only a factor of 3.8 increase. This clearly shows the high computational cost to make the first move from double precision to higher precision.

Table 5 compares the computational times of fixed precision with those of proactive and reactive adaptive precision. For fixed precision, we report the time for the minimum precision that successfully tracked all paths. For $n = 10$, which runs successfully in double precision, one can see in the proactive run time the overhead cost of checking conditions A, B, and C to see if an increase in precision is needed. Since higher precision is not needed for $n = 10$, the proactive method is the worst performer.

TABLE 4
Average time, in seconds, of 10 *runs of the Chebyshev polynomial of degree* 10 *with* $\tau = 8$, *for different levels of fixed precision.*

Double (52 bits)	64 bits	96 bits	128 bits	256 bits	512 bits	1024 bits
2.447	32.616	35.456	35.829	50.330	73.009	124.401

TABLE 5
Average time of 10 *runs of Chebyshev polynomials of degree* n *with* $\tau = 8$, *in seconds.*

n	Minimum fixed precision		Proactive	Reactive
10	52 bits	2.447	2.937	2.447
50	96 bits	1698.350	797.064	319.251
100	128 bits	9763.147	9739.978	3730.438
150	192 bits	32709.488	28944.029	21121.190

For larger n, however, increases in precision are needed, so both adaptive methods outperform fixed precision. As n increases, so does the fraction of the path which must be tracked in higher precision, so relative differences in the times decrease. However, since one does not know a priori how much fixed precision is needed, the real cost of a fixed precision rerun strategy would be much higher.

In these runs, the reactive method uses less time than the proactive approach. This is because the proactive error bounds are conservative, causing an increase in precision somewhat earlier than the reactive method. The high cost of leaving double precision outweighs any savings in avoiding failed steps, thus giving the reactive method the advantage.

5.4. A problem from chemistry. To illustrate the effect of tightening the tracking tolerance (i.e., increasing τ) on adaptive precision path tracking, we will consider a polynomial system coming from chemistry. To determine chemical equilibria, one may pass from a set of reaction and conservation equations to polynomials, as described in [13, 21]. One such polynomial system, discussed in [17], is the following:

$$f = \begin{bmatrix} 14z_1^2 + 6z_1z_2 + 5z_1 - 72z_2^2 - 18z_2 - 850z_3 + 0.000000002 \\ 0.5z_1z_2^2 + 0.01z_1z_2 + 0.13z_2^2 + 0.04z_2 - 40000 \\ 0.03z_1z_3 + 0.04z_3 - 850 \end{bmatrix}.$$

Bertini homogenized this system using one variable group and found the solutions by using a linear homotopy with a compatible total-degree start system, that is, one whose polynomials have degrees 2, 3, and 2, respectively. The solution set of this system consists of eight nonsingular finite solutions and a multiplicity-4 solution at infinity.

Using the method of section 4.3.2, the error bounds were set to $\Psi = 120{,}030$ and $\Phi = 240{,}008$. Both safety digits were set to 1. To exhibit the effect of adaptive precision tracking for tighter tolerances, the paths were tracked from $t = 1$ to $t_{\mathrm{f}} = 10^{-30}$ using $\tau = 8$ and $\tau = 14$. Each setting was run with 10 different choices of random numbers, and the average run time for the entire system is provided in Table 6.

With $\tau = 8$, the paths leading to the infinite solution started to fail around $t = 10^{-7}$ when using double precision due to the inadequacy of double precision to complete the linear algebra calculations. With $\tau = 14$, in each of the ten tests performed, one of the paths leading to the infinite solution failed before $t = 0.9$ when using double precision due to Newton iterations not converging to the required

TABLE 6
Average time of 10 *runs of the chemical system, in seconds.*

τ	Minimum fixed precision		Proactive	Reactive
8	96 bits	6.018	1.565	0.368
14	128 bits	203.67	198.67	86.68

tolerance. By rerunning the system with increased (fixed) precision, the path tracking completed successfully.

Both the proactive and reactive methods tracked all of the paths successfully in both tolerances. With $\tau = 8$, both methods started tracking all of the paths in double precision. For the paths in which precision needed to be increased, this first occurred around $t = 10^{-2}$ and $t = 10^{-7}$ for the proactive and reactive method, respectively. With $\tau = 14$, the proactive method increased precision immediately and started tracking all paths in 96-bit precision, while the reactive method was able to start tracking in double precision. These differences of when the precision was increased directly caused the difference in average tracking time between the proactive and reactive method.

5.5. A multivariate system from robotics. When using homotopy continuation methods to solve polynomial systems, one should always make use of an endgame. The purpose of the following example is to compare how the adaptive precision tracking methods of this paper perform when solving a typical, nonpathological example by simple tracking and with the added power of an endgame. In particular, the fractional power series endgame of [18] was employed in solving the following polynomial system [14], an inverse kinematics problem for a general six-revolute serial-link robot.

$$f = \begin{bmatrix} x_1^2 + x_2^2 - 1 \\ x_3^2 + x_4^2 - 1 \\ x_5^2 + x_6^2 - 1 \\ x_7^2 + x_8^2 - 1 \\ -0.24915068x_1x_3 + 1.6091354x_1x_4 + 0.27942343x_2x_3 + 1.4348016x_2x_4 \\ + 0.40026384x_5x_8 - 0.80052768x_6x_7 + 0.074052388x_1 - 0.083050031x_2 - 0.38615961x_3 \\ - 0.75526603x_4 + 0.50420168x_5 - 1.0916287x_6 + 0.40026384x_8 + 0.04920729 \\ 0.12501635x_1x_3 - 0.68660736x_1x_4 - 0.11922812x_2x_3 - 0.71994047x_2x_4 \\ - 0.43241927x_5x_7 - 0.86483855x_6x_8 - 0.03715727x_1 + 0.035436896x_2 + 0.085383482x_3 \\ - 0.039251967x_5 - 0.43241927x_7 + 0.013873010 \\ -0.63555007x_1x_3 - 0.11571992x_1x_4 - 0.66640448x_2x_3 + 0.11036211x_2x_4 \\ + 0.29070203x_5x_7 + 1.2587767x_5x_8 - 0.62938836x_6x_7 + 0.58140406x_6x_8 \\ + 0.19594662x_1 - 1.2280342x_2 - 0.079034221x_4 + 0.026387877x_5 - 0.057131430x_6 \\ - 1.1628081x_7 + 1.2587767x_8 + 2.162575 \\ 1.4894773x_1x_3 + 0.23062341x_1x_4 + 1.3281073x_2x_3 - 0.25864503x_2x_4 + 1.1651720x_5x_7 \\ - 0.26908494x_5x_8 + 0.53816987x_6x_7 + 0.58258598x_6x_8 - 0.20816985x_1 + 2.6868320x_2 \\ - 0.69910317x_3 + 0.35744413x_4 + 1.2499117x_5 + 1.4677360x_6 + 1.1651720x_7 \\ + 1.1076340x_8 - 0.69686809 \end{bmatrix}.$$

The exact formulation of this problem includes nonalgebraic functions as coefficients, which Bertini does not currently accept. As a result, the coefficients were calculated in 1024-bit precision, available at [2], with the first 8 digits matching the coefficients given above. Using a homotopy similar to that of section 5.4, there are

TABLE 7
Average time of 10 *runs of the IPP system for the two tests, in seconds.*

	Minimum fixed precision		Proactive	Reactive
Test 1	192 bits	300.169	85.404	176.591
Test 2	96 bits	196.143	47.393	37.630

256 paths, 32 leading to finite regular solutions, 64 leading to thirty-two multiplicity-2 infinite solutions, 64 leading to four multiplicity-16 infinite solutions, and 96 leading to four multiplicity-24 infinite solutions. Even though this problem is 2-homogeneous with the variable groups $\{x_1, x_2, x_5, x_6\}$ and $\{x_3, x_4, x_7, x_8\}$, this was not used. Not taking advantage of the special structure of the system makes it a bit more representative of systems that come up in practice, where, even taking account of all special structure, the number of singular solutions greatly outnumbers the actual physical solutions, as in [24].

We ran this system using two different endgames: the first tracking until the path stabilizes and the second using a power series endgame. In both cases, a geometric series of sample values for t was set: $t = 10^{-k}$ for $k = 1, 2, \ldots$. The paths were tracked between the sample points with a tolerance of 10^{-6}, and sample points for the endgame were refined to full precision. The endgame stops when successive predictions of the endpoint at $t = 0$ agree to within 10^{-12}. In the first test, the prediction to $t = 0$ is taken as just the current path point, whereas in the second test, a fractional power series was used following the method in [18]. In both tests, using the method of section 4.3.2, we set Φ to 14.5 and Ψ to 42.2, and both safety digits to 1. The average times are provided in Table 7. One observes a marked reduction in run time using the more advanced endgame.

In both tests, the proactive adaptive precision tracking used higher precision and larger step size than the reactive method to track between the sample points. In the first test, the proactive method was faster since taking fewer steps in higher precision was advantageous over taking more steps in lower precision utilized by the reactive method. In the second test, fewer sample points were needed to generate an accurate prediction, which kept the tracking sufficiently far away from the endpoints that have high multiplicity. In this case, the reactive method was faster since taking more steps in lower precision was advantageous over using higher precision with fewer steps utilized by the proactive method.

6. Discussion regarding singular endgames. Multiple precision becomes necessary when numerical stability is poor, such as at a point having a near-singular Jacobian matrix. In the case of an exact singularity at the end of a continuation path, an endgame based on fractional power (Puiseux) series or Cauchy integration [18, 21] can increase the accuracy that can be obtained at a given precision, thereby reducing the computational burden of higher precision, especially in cases where a double precision endgame suffices. For the endgame to succeed, the homotopy path must be sampled in an "endgame operating zone" inside the branchpoint closest to the endpoint but before the path becomes too ill-conditioned at the end [18, 21]. Without enough precision, this operating zone is empty. Adaptive precision tracking will ensure that enough precision is allocated to succeed. Likewise, endgames based on deflating the system to derive a related nonsingular one [8] may need higher than double precision to make a correct decision regarding the rank of the Jacobian at each stage of deflation. The method of [11] may lower the precision needed to make a correct rank decision, but there is no guarantee that double precision, or any

prespecified precision, will suffice. So while either type of endgame can help reduce computation and sometimes can avoid higher precision, ultimately an adaptive precision method is needed to ensure success in difficult cases. Among the causes leading to difficult cases, the most common are high degree polynomials, badly scaled functions, endpoints having high multiplicity, and a tight tolerance on the final accuracy.

7. Conclusion. We have presented an adaptive precision predictor-corrector path tracking algorithm that automatically allocates enough precision to allow convergence of Newton's method in the corrector phase. The adaptation can be driven reactively in response to failures in convergence or proactively using rules based on a numerical analysis of Newton's coupled to bounds established on the floating point error incurred in evaluating the homotopy function. The reactive approach is simpler in that no error bound is required and no computation time is spent evaluating the adaptation rules. It does, however, require the careful setting of the minimum step size, $\varepsilon(P)$, to ensure that the need for higher precision is triggered appropriately. We determined good settings for $\varepsilon(P)$ by experimentation.

The proactive approach requires an error bound. Floating point errors depend not only on the mathematical function being evaluated but also on the sequence of operations used, for example, $a(b+c)$ and $ab+ac$ may have different errors. We provide rough formulas for polynomials evaluated in expanded form, based on just the degree and the size of the coefficients, and we give a more precise method using circular arithmetic, a form of interval analysis, to bound the error in any straight-line polynomial function. To apply the proactive approach to nonpolynomial systems, one must extend the interval analysis to handle whatever new elementary functions arise, such as trigonometric functions and the like. Although we have not carried this out in our experiments, there appears to be no fundamental barrier to such extensions of the approach.

In focusing on Newton's method in the corrector phase, we have neglected the prediction phase. While the corrector dominates the convergence to the path, we recognize that a careful study of predictor methods is certainly warranted. The use of different predictor schemes, e.g., Adams–Bashforth rather than Euler, is well worth considering. Moreover, step size control in the current algorithm is entirely reactive, i.e., based on the history of success and failure of the predictor-corrector steps. A careful analysis of the predictor might be combined with the convergence criteria of the corrector to proactively determine a safe step size, possibly making decisions balancing the computational cost of a small step size against that of higher precision. Along this line, the method presented in [6] presents a rigorous step size control algorithm based on interval arithmetic, but it does not consider adaptive precision. Instead, if not enough precision is used, the safe step size goes to zero. An approach like ours may remove this mode of failure. The main intent of [6] is to prevent path crossing when two or more paths in the homotopy pass close to each other. We have not addressed this important issue here.

In conclusion, the adaptive precision approach presented here appears to be highly effective for tracking numerically difficult homotopy paths. Open questions remain about prediction methods and step size control, and additional work is needed to extend the error analysis to nonpolynomial functions. While our error analysis goes far in explaining the conditions that drive an increase in precision, at present the conservative nature of the bounds causes the proactive method to increase precision earlier than necessary. Due to the high cost of leaving double precision, the reactive method, which uses double precision as long as possible, takes less computational time than

the proactive method in all our standard tracking experiments. When endgames are utilized and the tracking between the sample points needs to be completed in precision higher than double, the reactive method may become slower than the proactive method due to the increased number of steps needed to track between the sample points.

Acknowledgment. We would like to thank the referees for their insightful comments, which have resulted in this being a much better article.

REFERENCES

[1] E. L. Allgower and K. Georg, *Numerical Continuation Methods. An Introduction*, Springer Ser. Comput. Math. 13, Springer-Verlag, Berlin, Heidelberg, New York, 1990; reprinted as *Introduction to Numerical Continuation Methods*, Classics Appl. Math. 45, SIAM, Philadelphia, 2003.

[2] D. J. Bates, J. D. Hauenstein, A. J. Sommese, and C. W. Wampler. *Bertini: Software for Numerical Algebraic Geometry*, available online from www.nd.edu/∼sommese/bertini.

[3] J. W. Demmel, *Applied Numerical Linear Algebra*, SIAM, Philadelphia, 1997.

[4] I. Gargantini and P. Henrici, *Circular arithmetic and the determination of polynomial zeros*, Numer. Math., 18 (1971/1972), pp. 305–320.

[5] R. B. Kearfott and M. Novoa, *Algorithm* 681: *INTBIS, a portable interval Newton/bisection package*, ACM Trans. Math. Softw., 16 (1990), pp. 152–157.

[6] R. B. Kearfott and Z. Xing, *An interval step control for continuation methods*, SIAM J. Numer. Anal., 31 (1994), pp. 892–914.

[7] D. E. Knuth, *The Art of Computer Programming, Vol.* 2: *Seminumerical Algorithms*, Addison-Wesley, Reading, MA, 1969.

[8] A. Leykin, J. Verschelde, and A. Zhao, *Evaluation of Jacobian matrices for Newton's method with deflation for isolated singularities of polynomial systems*, in Proceedings of the International Workshop on Symbolic-Numeric Computation (SNC 2005), D. Wang and L. Zhi, eds., Xi'an, China, 2005, pp. 19–28.

[9] T. Y. Li, *Numerical solution of multivariate polynomial systems by homotopy continuation methods*, Acta Numer., 6 (1997), pp. 399–436.

[10] T. Y. Li, *Numerical solution of polynomial systems by homotopy continuation methods*, in Handbook of Numerical Analysis. Volume XI. Special Volume: Foundations of Computational Mathematics, F. Cucker, ed., North-Holland, Amsterdam, 2003, pp. 209–304.

[11] T. Y. Li and Z. Zeng, *A rank-revealing method with updating, downdating, and applications*, SIAM J. Matrix Anal. Appl., 26 (2005), pp. 918–946.

[12] J.-P. Merlet. *A parser for the interval evaluation of analytical functions and its applications to engineering problems*, J. Symbolic Comput., 31 (2001), pp. 475–486.

[13] A. P. Morgan, *Solving Polynomial Systems Using Continuation for Engineering and Scientific Problems*, Prentice-Hall, Englewood Cliffs, NJ, 1987.

[14] A. P. Morgan and A. J. Sommese, *Computing all solutions to polynomial systems using homotopy continuation*, Appl. Math. Comput., 24 (1987), pp. 115–138; errata: Appl. Math. Comput., 51 (1992), p. 209.

[15] A. P. Morgan and A. J. Sommese, *Coefficient-parameter polynomial continuation*, Appl. Math. Comput., 29 (1989), pp. 123–160; errata: Appl. Math. Comput., 51 (1992), p. 207.

[16] A. P. Morgan, A. J. Sommese, and C. W. Wampler, *Computing singular solutions to nonlinear analytic systems*, Numer. Math., 58 (1991), pp. 669–684.

[17] A. P. Morgan, A. J. Sommese, and C. W. Wampler, *Computing singular solutions to polynomial systems*, Adv. Appl. Math., 13 (1992), pp. 305–327.

[18] A. P. Morgan, A. J. Sommese, and C. W. Wampler, *A power series method for computing singular solutions to nonlinear analytic systems*, Numer. Math., 63 (1992), pp. 391–409.

[19] N. Revol and F. Rouillier, *Motivations for an arbitrary precision interval arithmetic and the MPFI library*, Reliable Comput., 11 (2005), pp. 275–290.

[20] S. M. Rump, *INTLAB—INTerval LABoratory*, in Developments in Reliable Computing: Proceedings of the GAMM-IMACS International Symposium on Scientific Computing, Computer Arithmetic and Validated Numerics (SCAN-98), T. Csendes, ed., Budapest, 1998, Kluwer Academic, Dordrecht, The Netherlands, 1999, pp. 77–104.

[21] A. J. Sommese and C. W. Wampler, *Numerical Solution of Systems of Polynomials Arising in Engineering and Science*, World Scientific, Singapore, 2005.

[22] F. Tisseur, *Newton's method in floating point arithmetic and iterative refinement of generalized eigenvalue problems*, SIAM J. Matrix Anal. Appl., 22 (2001), pp. 1038–1057.
[23] J. Verschelde, *Algorithm* 795: *PHCpack: A general-purpose solver for polynomial systems by homotopy continuation*, ACM T. Math. Software, 25 (1999), pp. 251–276; software available from www.math.uic.edu/∼jan.
[24] C. W. Wampler, A. Morgan, and A. J. Sommese, *Complete solution of the nine-point path synthesis problem for four-bar linkages*, ASME J. Mech. Design, 114 (1992), pp. 153–159.
[25] D. Watkins, *Fundamentals of Matrix Computations*, John Wiley, New York, 1991.
[26] J. H. Wilkinson, *Rounding Errors in Algebraic Processes*, Dover, New York, 1994; reprint of the 1963 Prentice-Hall original.

SIAM J. NUMER. ANAL.
Vol. 46, No. 2, pp. 747–775
© 2008 Society for Industrial and Applied Mathematics

AN OPTIMAL ADAPTIVE FINITE ELEMENT METHOD FOR THE STOKES PROBLEM*

YAROSLAV KONDRATYUK† AND ROB STEVENSON‡

Abstract. A new adaptive finite element method for solving the Stokes equations is developed, which is shown to converge with the best possible rate. The method consists of 3 nested loops. The outermost loop consists of an adaptive finite element method for solving the pressure from the (elliptic) Schur complement system that arises by eliminating the velocity. Each of the arising finite element problems is a Stokes-type problem, with the pressure being sought in the current trial space and the divergence-free constraint being reduced to orthogonality of the divergence to this trial space. Such a problem is still continuous in the velocity field. In the middle loop, its solution is approximated using the Uzawa scheme. In the innermost loop, the solution of the elliptic system for the velocity field that has to be solved in each Uzawa iteration is approximated by an adaptive finite element method.

Key words. adaptive finite element method, convergence rates, computational complexity, a posteriori error estimators, Uzawa iteration, Stokes equations

AMS subject classifications. 65N30, 65N50, 65N15, 65Y20, 41A25

DOI. 10.1137/06066566X

1. Introduction.

Often the solution of a boundary value problem exhibits singularities, e.g., due to a nonsmooth boundary. Then, because of the lacking (Sobolev) smoothness of the solution, finite element methods based on quasi-uniform partitions converge with a rate smaller than allowed by the polynomial degrees used in the finite element spaces. This can be repaired when suitable refinements are made near those singularities. The optimal distribution of the size of elements as a function of the distance to a singularity depends on the strength of the singularity, which is generally unknown.

With adaptive finite element methods (AFEMs), a sequence of nested partitions is created, where, when creating the next partition, the decision of where to refine is made on the basis of an a posteriori estimator of the error in the current finite element approximation. Although they had been used successfully for more than 25 years, in more than one space dimension, even for second order *elliptic* equations, their convergence was not shown before the works of Dörfler [Dör96] and that of Morin, Nochetto, and Siebert [MNS00]. Convergence alone, however, does not show that the use of an adaptive method for a solution that has singularities improves upon, or even competes with, that of a nonadaptive one. Recently, after the derivation of such a result by Binev, Dahmen, and DeVore [BDD04] for an AFEM extended with a so-called coarsening or derefinement routine, in [Ste07] we could prove that standard AFEMs converge with the best possible rate in linear complexity.

*Received by the editors July 21, 2006; accepted for publication (in revised form) July 10, 2007; published electronically February 8, 2008. This work was supported by the Netherlands Organization for Scientific Research and by the European Community's Human Potential Programme under contract HPRN-CT-2002-00286.

http://www.siam.org/journals/sinum/46-2/66566.html

†Department of Mathematics, Humboldt-Universität zu Berlin, Unter den Linden 6, D-10099 Berlin, Germany (kondraty@mathematik.hu-berlin.de).

‡Korteweg–de Vries Institute for Mathematics, Plantage Muidergracht 24, 1018 TV Amsterdam, The Netherlands (rstevens@science.uva.nl).

In this paper, as a model *saddle point problem*, we consider the Stokes equations

$$\begin{cases} -\triangle \mathbf{u} + \nabla p = \mathbf{f} & \text{on } \Omega \subset \mathbf{R}^d, \\ \qquad \text{div}\mathbf{u} = 0 & \text{on } \Omega, \\ \qquad\quad \mathbf{u} = 0 & \text{on } \partial\Omega. \end{cases}$$

(Although, in this introduction, we write equations in strong form, actually we always mean the corresponding variational formulations.) In [DDU02], Dahlke, Dahmen, and Urban analyzed an adaptive wavelet method for solving these equations. The starting point was the application of the Uzawa iteration on the continuous level, i.e., given some p_0, compute for $j = 0, 1, \ldots$

$$\left\{ \begin{array}{ll} -\triangle \mathbf{u}_{j+1} = \mathbf{f} - \nabla p_j & \text{on } \Omega \\ \qquad\quad \mathbf{u} = 0 & \text{on } \partial\Omega \end{array} \right\}, \quad p_{j+1} = p_j - \text{div}\mathbf{u}_j.$$

Of course, this iteration cannot be performed exactly, and in each iteration the solution of the elliptic system was approximated using an adaptive wavelet method within decreasing tolerances as the iteration proceeds. Convergence was shown, and by the inclusion of coarsening steps, even optimal rates and linear complexity were demonstrated. Since no Galerkin discretizations were formed from the mixed problem, the so-called LBB stability for the discrete problem was not required.

In [BMN02], Bänsch, Morin, and Nochetto studied the above solution method with the adaptive wavelet method replaced by an AFEM. They proved convergence, and despite the fact that they did not include coarsening, in numerical experiments they observed optimal rates, at least when the elliptic problems were solved not too accurately. When prescribing an a priori tolerance of the form γ^j in the jth iteration, it was necessary to take γ in the range $[\approx .95, 1)$. By the addition of coarsening to this method, in [Kon06] optimal computational complexity was demonstrated.

When starting this work, our aim was to prove optimal computational complexity of basically the method from [BMN02], i.e., without coarsening. For a reason that will be indicated later (in Remark 6.6), we did not succeed in doing this. Instead, we prove this result for a somewhat more complicated algorithm involving an additional outer loop, without coarsening though. Below we briefly describe the loops of our algorithm starting from the outermost one.

The pressure p can be found as the solution of the Schur complement equation that one obtains by eliminating the velocity $\mathbf{u}$ from the Stokes equations. This equation is *elliptic*, with corresponding energy norm equivalent to the $L_2(\Omega)$-norm. Given a finite element space $\mathbb{P}_{\sigma_i}$, where σ_i denotes the underlying finite element partition, the best approximation from this space to p with respect to this energy norm is the Galerkin solution $p_i \in \mathbb{P}_{\sigma_i}$. With Q_{σ_i} denoting the $L_2(\Omega)$-orthogonal projection onto $\mathbb{P}_{\sigma_i}$, this p_i can be shown to be the unique solution of

$$\begin{cases} -\triangle \mathbf{u}^{(i)} + \nabla p_i = \mathbf{f} & \text{on } \Omega, \\ Q_{\sigma_i}\text{div}\mathbf{u}^{(i)} = 0 & \text{on } \Omega, \\ \qquad\quad \mathbf{u}^{(i)} = 0 & \text{on } \partial\Omega, \end{cases}$$

i.e., the Stokes equations in which the divergence-free condition has been relaxed. We refer to this system as the reduced Stokes equations.

Concerning the velocity, this is still a problem posed over an infinite-dimensional space. Inside the inner loops, a sufficiently accurate approximate solution will be

determined, which here again is denoted as $(\mathbf{u}^{(i)}, p_i)$. We will have that $p_i \in \mathbb{P}_{\sigma_i}$, while $\mathbf{u}^{(i)}$ is from a finite element space $\mathbb{V}_{\tau_i}$ with respect to a partition τ_i that is a refinement of σ_i. It will be shown that the energy norm of $p - p_i$ is equivalent to the *a posteriori error estimator* $\|\mathrm{div}\mathbf{u}^{(i)}\|_{L_2(\Omega)}$, and that for any refinement σ_{i+1} of σ_i, the energy norm of $p_{i+1} - p_i$ is equivalent to $\|Q_{\sigma_{i+1}}\mathrm{div}\mathbf{u}^{(i)}\|_{L_2(\Omega)}$. Now following along the lines of [Dör96, MNS00] for Poisson-type problems, by selecting σ_{i+1} such that for some $\theta \in (0, 1]$, $\|Q_{\sigma_{i+1}}\mathrm{div}\mathbf{u}^{(i)}\|_{L_2(\Omega)} \geq \theta\|\mathrm{div}\mathbf{u}^{(i)}\|_{L_2(\Omega)}$ (the "bulk criterion"), the so-called saturation property is guaranteed, and a linearly convergent sequence $(p_i)_i$ towards p is obtained. Moreover, by taking θ to be small enough, and σ_{i+1} with quasi-minimal cardinality, following along the lines of [Ste07] convergence of the $p_i \in \mathbb{P}_{\sigma_i}$ towards p with the optimal rate is shown.

Compared to the adaptive methods for Poisson-type problems, a complication is that to find such a σ_{i+1}, it is generally not sufficient to search it within the set of partitions that can be created by refining each element of σ_i some fixed number of times. On the other hand, by considering finite element spaces such that $\mathrm{div}\mathbb{V}_{\tau_i} \subseteq \mathbb{P}_{\tau_i}$, we know that we never have to subdivide elements that are in τ_i. To determine σ_{i+1} with the required properties we apply an adaptive tree algorithm developed by Binev and DeVore in [BD04].

For solving the reduced Stokes problem for given i, i.e., fixed σ_i, we follow the approach from [BMN02] for the full Stokes problem. That is, we apply Uzawa, where the pressure update then reads as $p_{j+1}^{(i)} = p_j^{(i)} - Q_{\sigma_i}\mathrm{div}\mathbf{u}_j^{(i)}$. Each of the inner elliptic systems $-\triangle\mathbf{u}_{j+1}^{(i)} = \mathbf{f} - \nabla p_j^{(i)}$ on Ω, $\mathbf{u} = 0$ on $\partial\Omega$ is solved with a standard AFEM producing increasingly more refined partitions, with σ_i being the initial partition for the first call. Already having optimal control over $\#\sigma_i$, we are now able to prove also optimal rates of the velocity approximations towards $\mathbf{u}$. Note that other than in [BMN02], we have two different partitions underlying pressure and velocity approximations, with the latter being always a refinement of the first. Throughout the algorithm, both partitions become increasingly more refined, i.e., no derefinements are made.

This paper is organized as follows: In section 2, we recall some properties of the Stokes problem. In section 3, we define the finite element spaces that we will use. We give some properties of a procedure for refining partitions, which is a generalization to arbitrary space dimensions of the newest vertex bisection method in two space dimensions. An overview of the solution method will be given in section 4. In section 5, a posteriori error estimators are derived for the various problems that occur in our solution method. In section 6, the adaptive refinement routines for pressure and velocity partitions are given. In section 7, we give the detailed description of the adaptive method in the simplified situation that the right-hand side $\mathbf{f}$ is piecewise polynomial with respect to the initial partition. We prove convergence with the optimal rates. In this section, we assume that the arising finite-dimensional linear systems are solved exactly, ignoring the question of computational complexity. In section 8, we give the method for general right-hand sides, and replace the direct solvers by iterative solution methods, with which we end up with a method of optimal computational complexity. Finally, in section 9, we present numerical results, and compare them with those obtained with the method from [BMN02]. As we will see, in this example both methods give similar results.

In this paper, by $C \lesssim D$ we will mean that C can be bounded by a multiple of D, independently of parameters which C and D may depend on. Obviously, $C \gtrsim D$ is defined as $D \lesssim C$, and $C \eqsim D$ as $C \lesssim D$ and $C \gtrsim D$.

2. Stokes problem. Let Ω be a polygonal domain in $\mathbf{R}^d$. We consider the Stokes problem in variational form: With

$$\mathbb{V} := H^1_0(\Omega)^d, \quad \mathbb{P} := L_{2,0}(\Omega) := \left\{ q \in L_2(\Omega) : \int_\Omega q = 0 \right\},$$

and given an $\mathbf{f} \in \mathbb{V}'$, throughout this paper $\mathbf{u} \in \mathbb{V}$ (the velocity) and $p \in \mathbb{P}$ (the pressure) will denote the solutions of

$$\boxed{a(\mathbf{u},\mathbf{v}) + b(\mathbf{v},p) + b(\mathbf{u},q) = \mathbf{f}(\mathbf{v}) \quad (\mathbf{v} \in \mathbb{V},\, q \in \mathbb{P}),} \tag{2.1}$$

where $a : \mathbb{V} \times \mathbb{V} \to \mathbf{R}$, $b : \mathbb{V} \times \mathbb{P} \to \mathbf{R}$ are defined by

$$a(\mathbf{w},\mathbf{v}) := \int_\Omega \nabla\mathbf{w} : \nabla\mathbf{v}, \quad b(\mathbf{v},q) := -\int_\Omega q\, \mathrm{div}\mathbf{v}.$$

It is well known that

$$\|\mathbf{v}\|_{\mathbb{V}} := a(\mathbf{v},\mathbf{v})^{\frac{1}{2}} \eqsim \|\mathbf{v}\|_{H^1(\Omega)^d} \quad (\mathbf{v} \in \mathbb{V}),$$

b is bounded, and

$$\beta := \inf_{0\neq q\in\mathbb{P}} \sup_{0\neq \mathbf{v}\in\mathbb{V}} \frac{b(\mathbf{v},q)}{\|\mathbf{v}\|_{\mathbb{V}}\|q\|_{L_2(\Omega)}} > 0. \tag{2.2}$$

As a consequence, the Stokes problem is well-posed, meaning that

$$\|\mathbf{w}\|_{\mathbb{V}} + \|r\|_{L_2(\Omega)} \eqsim \sup_{0\neq(\mathbf{v},q)\in\mathbb{V}\times\mathbb{P}} \frac{a(\mathbf{w},\mathbf{v}) + b(\mathbf{v},r) + b(\mathbf{w},q)}{\|\mathbf{v}\|_{\mathbb{V}} + \|q\|_{L_2(\Omega)}} \quad (\mathbf{w} \in \mathbb{V},\, r \in \mathbb{P}). \tag{2.3}$$

Remark 2.1. Since $a(\cdot,\cdot)$ is elliptic on the whole of $\mathbb{V}$ instead of only on the space of the divergence-free velocities, clearly the Stokes problem (2.1) with $\mathbb{P}$ replaced by any subspace is also well-posed, uniformly in the choice of such a subspace.

Defining $A : \mathbb{V} \to \mathbb{V}'$, $B : \mathbb{V} \to \mathbb{P}'$, and $B' : \mathbb{P} \to \mathbb{V}'$ by $(A\mathbf{v})(\mathbf{w}) = a(\mathbf{v},\mathbf{w})$, $(B\mathbf{v})(q) = b(\mathbf{v},q) = (B'q)(\mathbf{v})$, problem (2.1) can be equivalently written as

$$\begin{bmatrix} A & B' \\ B & 0 \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ p \end{bmatrix} = \begin{bmatrix} \mathbf{f} \\ 0 \end{bmatrix},$$

and, with the Schur complement $S := BA^{-1}B'$, p is also uniquely determined by

$$Sp = BA^{-1}f.$$

LEMMA 2.2. *$(Sq)(q) = \sup_{0\neq\mathbf{v}\in\mathbb{V}} \frac{b(\mathbf{v},q)^2}{a(\mathbf{v},\mathbf{v})}$ and $\|q\|_{\mathbb{P}} := (Sq)(q)^{\frac{1}{2}} \eqsim \|q\|_{L_2(\Omega)}$ $(q \in \mathbb{P})$.*

Proof. With $\langle\cdot,\cdot\rangle = \langle\cdot,\cdot\rangle_{H^1(\Omega)^d}$ (or $\langle\cdot,\cdot\rangle = a(\cdot,\cdot)$), let $R : \mathbb{V} \to \mathbb{V}'$ be the mapping such that $\mathbf{g}(\mathbf{v}) = \langle\mathbf{v}, R\mathbf{g}\rangle$ $(\mathbf{v} \in \mathbb{V},\, \mathbf{g} \in \mathbb{V}')$. Writing $\tilde{B}' = RB'$, $\tilde{A} = RA$, we have

$$\sup_{0\neq\mathbf{v}\in\mathbb{V}} \frac{b(\mathbf{v},q)^2}{a(\mathbf{v},\mathbf{v})} = \sup_{0\neq\mathbf{v}\in\mathbb{V}} \frac{(B'q)(\mathbf{v})^2}{(A\mathbf{v})(\mathbf{v})} = \sup_{0\neq\mathbf{v}\in\mathbb{V}} \frac{\langle\mathbf{v},\tilde{B}'q\rangle^2}{\langle\mathbf{v},\tilde{A}\mathbf{v}\rangle} = \sup_{0\neq\mathbf{w}\in\mathbb{V}} \frac{\langle\mathbf{w},\tilde{A}^{-\frac{1}{2}}\tilde{B}'q\rangle^2}{\langle\mathbf{w},\mathbf{w}\rangle}$$
$$= \langle\tilde{A}^{-\frac{1}{2}}\tilde{B}'q, \tilde{A}^{-\frac{1}{2}}\tilde{B}'q\rangle = \langle A^{-1}B'q, RB'q\rangle = (Sq)(q) \quad (q \in \mathbb{P}).$$

The second statement follows from the ellipticity of a, the boundedness of b, and (2.2). ∎

For $\mathbf{g} \in \mathbb{V}'$, we set $\|\mathbf{g}\|_{\mathbb{V}'} = \sup_{0\neq\mathbf{v}\in\mathbb{V}} \frac{|\mathbf{g}(\mathbf{v})|}{\|\mathbf{v}\|_{\mathbb{V}}}$. Equipped with norms $\|\cdot\|_{\mathbb{V}}$ and $\|\cdot\|_{\mathbb{V}'}$, respectively, $A : \mathbb{V} \to \mathbb{V}'$ is an isomorphism.

Functions $\mathbf{g} \in L_2(\Omega)^d$ will be interpreted as functionals by means of $\mathbf{g}(\mathbf{v}) := \int_\Omega \mathbf{g}\cdot\mathbf{v}$.

3. Finite element approximation. Given some fixed $m \in \mathbf{N}_{>0}$, and partitions τ and σ of $\bar{\Omega}$ into essentially disjoint (closed) d-simplices, we will search approximations for $\mathbf{u}$ and p from the finite element spaces

$$\mathbb{V}_\tau := \mathbb{V} \cap \prod_{T\in\tau} P_m(T)^d \quad \text{and} \quad \mathbb{P}_\sigma := \mathbb{P} \cap \prod_{T\in\sigma} P_{m-1}(T),$$

respectively. For doing so, furthermore we will approximate the right-hand side $\mathbf{f}$ by functions from

$$\mathbb{V}_\tau^* := \prod_{T\in\tau} P_{m-1}(T)^d.$$

At any moment in our algorithm we will have that $\tau \supseteq \sigma$, meaning that $\tau = \sigma$ or that it is a proper refinement of σ. Note that $(\mathbb{V}_\tau, \mathbb{P}_\tau)$ is *not* an LBB stable pair.

Sometimes we will view $\mathbb{V}$ and $\mathbb{P}$ formally as finite element spaces corresponding to the infinitely fine partition ∞, and denote them as $\mathbb{V}_\infty$ and $\mathbb{P}_\infty$, respectively.

Remark 3.1. The fact that the approximate pressure is a piecewise polynomial of degree not larger than $m-1$ will be used only in the forthcoming Proposition 5.2. It is most likely that also there higher degree polynomials can be permitted at the expense of having a more complicated refinement rule for the velocity partitions (it will be needed to create more interior vertices; cf. Figure 5.1). On the other hand, at least for our analysis, it will be essential that $\mathbb{P}_\tau \supseteq \mathrm{div}\mathbb{V}_\tau$ (cf. Remark 6.1).

We will exclusively consider partitions that can be created by a certain recursive *bisection* procedure starting from some fixed conforming initial partition τ_0. The procedure we consider is a generalization to any space dimension of the well-known newest vertex bisection rule in two space dimensions (cf. [Mit89]). Here we recall a few properties of this generalized newest vertex bisection method that are relevant for the current exposition, and we refer to [Ste06] for a complete description of the method.

The way of bisecting any simplex in any partition that can be created by the method is uniquely determined. It depends only on a local numbering of the vertices of the simplices in the initial partition τ_0. As a consequence, any partition can be represented by a subtree of the infinite binary master tree $\mathcal{T}_*$ having as roots the simplices of τ_0, and in which the parent–child relation corresponds to the unique bisection of the parent. The partitions that can be created are *uniformly shape regular*, dependent only on τ_0.

For applying a posteriori error estimators, we will need the partitions τ underlying the velocity approximation spaces $\mathbb{V}_\tau$ to be conforming. So in the following,

$$\tau, \tau', \tilde{\tau}, \textit{ etc., will always denote } \text{conforming } \textit{partitions}.$$

Bisecting one or more simplices in a conforming partition τ generally results in a nonconforming partition ϱ. Conformity has to be restored by (recursively) bisecting any simplex $T \in \varrho$ that contains a vertex v of a $T' \in \varrho$ that does not coincide with any vertex of T (such a v is called a hanging vertex). This process, called completion, results into the smallest conforming refinement of ϱ.

Our adaptive method will be of the following form:

`for` $j := 1$ `to` M
`do` *create some, possibly nonconforming refinement* ϱ_j *of* τ_{j-1}
complete ϱ_j *to its smallest conforming refinement* τ_j
`endfor`

As we will see, we will be able to bound $\sum_{j=1}^{M} \#\varrho_j - \#\tau_{j-1}$. Because of the additional bisections made in the completion steps, however, generally $\#\tau_M - \#\tau_0$ will be larger. The following crucial result shows that these additional bisections inflate the total number of simplices by at most an absolute constant factor. It is valid with a proper local numbering of the vertices of the simplices in the initial partition τ_0, which we will assume in the following. In two dimensions such a numbering exists for any conforming partition. In more than two dimensions, it always exists after some initial refinement that inflates the number of simplices by not more than a constant factor.

THEOREM 3.2 (see [Ste06], generalizes upon [BDD04, Theorem 2.4] for $d = 2$).

$$\#\tau_M - \#\tau_0 \lesssim \sum_{j=1}^{M} \#\varrho_j - \#\tau_{j-1},$$

dependent only on τ_0, and in particular thus independent of M.

Remark 3.3. Note that this result in particular implies that any descendant ϱ of τ_0 has a conforming refinement τ with $\#\tau \lesssim \#\varrho$, dependent only on τ_0 and d.

We end this section by introducing some more notation. The smallest partition that is a refinement of partitions ϱ_1 and ϱ_2, i.e., their smallest common refinement, will be denoted as $\varrho_1 \cup \varrho_2$. For a partition τ (thus a conforming one), E_τ will denote the set of all internal $(d-1)$-dimensional hyperfaces in τ. For $T \in \tau$, $\mathcal{F}_\tau(T)$ denotes the union of T and the set of its neighbors in τ, i.e., those simplices that share a $(d-1)$-dimensional hyperface with T.

4. Overview of the solution method. For a partition σ_i, we consider the Galerkin problem of finding $p^{(i)} \in \mathbb{P}_{\sigma_i}$ such that

$$(Sp^{(i)})(q) = (BA^{-1}\mathbf{f})(q) \quad (q \in \mathbb{P}_{\sigma_i}).$$

With $\mathbf{u}^{(i)} := A^{-1}(\mathbf{f} - B'p^{(i)})$, this problem is equivalent to the semidiscrete problem of finding $(\mathbf{u}^{(i)}, p^{(i)}) \in \mathbb{V} \times \mathbb{P}_{\sigma_i}$ such that

$$a(\mathbf{u}^{(i)}, \mathbf{v}) + b(\mathbf{v}, p^{(i)}) + b(\mathbf{u}^{(i)}, q) = \mathbf{f}(\mathbf{v}) \quad (\mathbf{v} \in \mathbb{V},\ q \in \mathbb{P}_{\sigma_i}). \tag{4.1}$$

Since this is just the Stokes problem in which the divergence-free constraint has been relaxed, we will refer to this problem as the *reduced Stokes problem*. The solution $p^{(i)}$ *is the best approximation to p from $\mathbb{P}_{\sigma_i}$ with respect to $\|\cdot\|_{\mathbb{P}}$*, and by creating a suitable *adaptively refined* sequence of partitions $\tau_0 =: \sigma_0 \subset \sigma_1 \subset \cdots$, a convergent sequence $(p^{(i)})_i$ towards p is obtained.

The reduced Stokes problem, however, cannot be solved exactly. With the Riesz operator $R_{\sigma_i} : \mathbb{P}' \to \mathbb{P}_i$ being defined by $g(q) = \langle q, R_{\sigma_i} g\rangle_{L_2(\Omega)}$ $(q \in \mathbb{P}_{\sigma_i})$, it can be written as $R_{\sigma_i} Sp^{(i)} = R_{\sigma_i} BA^{-1}\mathbf{f}$. Equipping $\mathbb{P}_{\sigma_i}$ with $\langle\cdot,\cdot\rangle_{L_2(\Omega)}$, the operator $R_{\sigma_i} S : \mathbb{P}_{\sigma_i} \to \mathbb{P}_{\sigma_i}$ is symmetric, bounded, and positive definite, with a spectrum in $[\beta^2, 1]$ (cf. [NP04]). So to solve the reduced Stokes problem approximately, we may apply Richardson's iteration:

$$p_{j+1}^{(i)} = p_j^{(i)} - (R_{\sigma_i} Sp_j^{(i)} - R_{\sigma_i} BA^{-1}\mathbf{f}) \quad (j = 0, 1, \dots).$$

With $Q_{\sigma_i} : \mathbb{P} \to \mathbb{P}_{\sigma_i}$ denoting the $L_2(\Omega)$-orthogonal projector onto $\mathbb{P}_{\sigma_i}$, we have $R_{\sigma_i} B = -Q_{\sigma_i}\mathrm{div}$. Writing $\mathbf{u}_j^{(i)} := A^{-1}(\mathbf{f} - B'p_j^{(i)})$, we arrive at the equivalent formulation

$$\begin{cases} a(\mathbf{u}_j^{(i)}, \mathbf{v}) \;=\; \mathbf{f}(\mathbf{v}) - b(\mathbf{v}, p_j^{(i)}) \quad (\mathbf{v} \in \mathbb{V}), \\ \qquad p_{j+1}^{(i)} \;=\; p_j^{(i)} - Q_{\sigma_i}\mathrm{div}\mathbf{u}_j^{(i)}, \end{cases} \tag{4.2}$$

known as the Uzawa iteration. The properties of $R_{\sigma_i}S$ show that

$$\|p^{(i)} - p^{(i)}_{j+1}\|_{L_2(\Omega)} \leq [1-\beta^2]\|p^{(i)} - p^{(i)}_j\|_{L_2(\Omega)}. \tag{4.3}$$

Also the Uzawa iteration for solving the reduced Stokes problem cannot be performed exactly since it involves solving an elliptic problem posed over $\mathbb{V}$. To solve this problem approximately, we will again consider Galerkin approximations: Given a partition $\tau^{(i)}_{j,k}$, let $\mathbf{u}^{(i)}_{j,k} \in \mathbb{V}_{\tau^{(i)}_{j,k}}$ be the solution of

$$a(\mathbf{u}^{(i)}_{j,k}, \mathbf{v}) = \mathbf{f}(\mathbf{v}) - b(\mathbf{v}, p^{(i)}_j) \quad (\mathbf{v} \in \mathbb{V}_{\tau^{(i)}_{j,k}}). \tag{4.4}$$

It is the best approximation to $\mathbf{u}^{(i)}_j$ from $\mathbb{V}_{\tau^{(i)}_{j,k}}$ with respect to $\|\cdot\|_{\mathbb{V}}$, and by creating a suitable *adaptively refined* sequence of partitions $\sigma_i \subseteq \tau^{(i)}_{j,0} \subset \tau^{(i)}_{j,1} \subset \cdots$, a convergent sequence $(\mathbf{u}^{(i)}_{j,k})_k$ towards $\mathbf{u}^{(i)}_j$ is obtained. To guarantee that $\mathbf{u}^{(i)}_{j,k+1}$ is indeed a better approximation than $\mathbf{u}^{(i)}_{j,k}$, we will require that $\mathbf{f} \in \mathbb{V}'$ be sufficiently well approximated by a vector field from $\mathbb{V}^*_{\tau^{(i)}_{j,k}}$. To implement the latter requirement, instead of working with $\mathbf{f}$, we will replace it by suitable piecewise polynomial vector fields of degree $m-1$, which should become increasingly more accurate when the iterations proceed.

Finally, instead of solving the finite-dimensional linear systems (4.4) exactly, in order to obtain a method of (quasi-)optimal computational complexity, we will use approximate solutions obtained by employing optimal iterative solvers.

In order to stop each of the nested loops in time, as well as for both Galerkin problems, to create adaptively refined partitions such that the corresponding approximations converge towards the solution, and such that these partitions have quasi-optimal cardinalities, we need *a posteriori error estimators*, which will be discussed in the next section. On the one hand, stopping a loop in time means that it should not stop too early in order to guarantee convergence of the overall process. On the other hand, the iterations should not proceed too long in order to control both the cardinalities of the partitions, which grow by the refinements, and the computational complexity.

In view of our solution method, we fix some notation. Throughout this paper, given some $r \in \mathbb{P}$, where we have in mind an approximation to p, and a partition τ, $\mathbf{u}^r_\tau \in \mathbb{V}_\tau$ will denote the solution of the (discretized) elliptic problem

$$\boxed{a(\mathbf{u}^r_\tau, \mathbf{v}_\tau) = \mathbf{f}(\mathbf{v}_\tau) - b(\mathbf{v}_\tau, r) \quad (\mathbf{v}_\tau \in \mathbb{V}_\tau).} \tag{4.5}$$

As a special case of (4.5), $\mathbf{u}^r = \mathbf{u}^r_\infty \in \mathbb{V}$ thus denotes the solution of

$$a(\mathbf{u}^r, \mathbf{v}) = \mathbf{f}(\mathbf{v}) - b(\mathbf{v}, r) \quad (\mathbf{v} \in \mathbb{V}). \tag{4.6}$$

So the subscript τ in the notation $\mathbf{u}^r_\tau$ refers to its membership of the finite element space $\mathbb{V}_\tau$, whereas the superscript r in the notation $\mathbf{u}^r_\tau$ and $\mathbf{u}^r$ refers to their dependence on the approximate pressure r on the right-hand side. Note that $\mathbf{u} = \mathbf{u}^p$. Often we will consider (4.5) and (4.6) for $r = r_\sigma \in \mathbb{P}_\sigma$ for some $\sigma \subseteq \tau$ giving rise to the notation $\mathbf{u}^{r_\sigma}_\tau$ and $\mathbf{u}^{r_\sigma}$.

Given a partition σ, $(\mathbf{u}^{p_\sigma}, p_\sigma) \in \mathbb{V} \times \mathbb{P}_\sigma$ will denote the solution of the reduced Stokes problem

$$\boxed{a(\mathbf{u}^{p_\sigma}, \mathbf{v}) + b(\mathbf{v}, p_\sigma) + b(\mathbf{u}^{p_\sigma}, q_\sigma) = \mathbf{f}(\mathbf{v}) \quad (\mathbf{v} \in \mathbb{V},\, q_\sigma \in \mathbb{P}_\sigma).}$$

Note that $a(\mathbf{u}^{p_\sigma}, \mathbf{v}) = \mathbf{f}(\mathbf{v}) - b(\mathbf{v}, p_\sigma)$ $(\mathbf{v} \in \mathbb{V})$ so that the notation $\mathbf{u}^{p_\sigma}$ is consistent with (4.6).

5. A posteriori error estimators.

5.1. A posteriori error estimator for the inner elliptic problem. For a given right-hand side $\mathbf{f}$ and partitions $\tau \supseteq \sigma$ and $r_\sigma \in \mathbb{P}_\sigma$, we study an a posteriori error estimator for $\|\mathbf{u}^{r_\sigma} - \mathbf{u}_\tau^{r_\sigma}\|_{\mathbb{V}}$. Since even the Galerkin problem (4.5) will be solved only inexactly, whenever possible we even estimate $\|\mathbf{u}^{r_\sigma} - \mathbf{w}_\tau\|_{\mathbb{V}}$ for some general $\mathbf{w}_\tau \in \mathbb{V}_\tau$.

For $T \in \tau$, we set the local error indicator

$$\begin{aligned}&\eta_T(\mathbf{f}, r_\sigma, \mathbf{w}_\tau)\\ &\quad := \operatorname{diam}(T)^2 \|\mathbf{f} - \nabla r_\sigma + \triangle \mathbf{w}_\tau\|^2_{L_2(T)^d} + \operatorname{diam}(T) \|[\![r_\sigma \mathbf{n} - \nabla \mathbf{w}_\tau \cdot \mathbf{n}]\!]_{\partial T}\|^2_{L_2(\partial T)^d},\end{aligned}$$

where $[\![\cdot]\!]_{\partial T}$ denotes the jump of its argument over ∂T in the direction of $\mathbf{n}$, being a unit vector normal to ∂T. This jump is defined to be zero over $\partial\Omega$. We set the elliptic error estimator

$$\mathcal{E}^{\mathrm{E}}(\tau, \mathbf{f}, r_\sigma, \mathbf{w}_\tau) := \left[\sum_{T\in\tau} \eta_T(\mathbf{f}, r_\sigma, \mathbf{w}_\tau)\right]^{\frac{1}{2}}.$$

Note that the definition of the error estimator requires $\mathbf{f} \in L_2(\Omega)^d$, which we therefore assume here.

In Proposition 5.1 below, for $\tau' \supseteq \tau$ it is shown that the sum of the local error indicators corresponding to the simplices that were refined when creating τ' from τ, or those that have nonempty intersection with such simplices, is an *upper bound* for $\|\mathbf{u}_{\tau'}^{r_\sigma} - \mathbf{u}_\tau^{r_\sigma}\|^2_{\mathbb{V}}$. Substituting $\tau' = \infty$, it generalizes the known upper bound for $\|\mathbf{u}^{r_\sigma} - \mathbf{u}_\tau^{r_\sigma}\|_{\mathbb{V}}$ [BMN02, Lemma 5.1, eq. (5.4)], [Ver96]. Since our generalization is not so difficult to derive, we refer to the extended preprint version [KS07] of this work for a proof.

PROPOSITION 5.1. *Let $\tau' \supseteq \tau \supseteq \sigma$ be partitions, and let $r_\sigma \in \mathbb{P}_\sigma$, $\mathbf{f} \in L_2(\Omega)^d$, and*

$$\overline{F} = \overline{F}(\tau, \tau') := \{T \in \tau : T \cap \tilde{T} \neq \emptyset \text{ for some } \tilde{T} \in \tau \text{ that has been refined in } \tau'\}.$$

Then we have

$$\|\mathbf{u}_{\tau'}^{r_\sigma} - \mathbf{u}_\tau^{r_\sigma}\|_{\mathbb{V}} \leq C_1 \left[\sum_{T\in\overline{F}} \eta_T(\mathbf{f}, r_\sigma, \mathbf{u}_\tau^{r_\sigma})\right]^{\frac{1}{2}}$$

for some absolute constant $C_1 > 0$. Note that $\#\overline{F} \lesssim \#\tau' - \#\tau$.

In particular, by taking $\tau' = \infty$, we have $\|\mathbf{u}^{r_\sigma} - \mathbf{u}_\tau^{r_\sigma}\|_{\mathbb{V}} \leq C_1 \mathcal{E}^{\mathrm{E}}(\tau, \mathbf{f}, r_\sigma, \mathbf{u}_\tau^{r_\sigma})$.

Next we study whether the error estimator also provides a *lower bound* for the error $\|\mathbf{u}^{r_\sigma} - \mathbf{u}_\tau^{r_\sigma}\|_{\mathbb{V}}$ and, when τ' is a sufficient refinement of τ, for $\|\mathbf{u}_{\tau'}^{r_\sigma} - \mathbf{u}_\tau^{r_\sigma}\|_{\mathbb{V}}$. In order to derive such estimates, for the time being we further restrict the type of right-hand sides to piecewise polynomials of degree $m-1$ with respect to τ. We will call $\tau' \supset \tau$ *a full refinement with respect to $T \in \tau$*, when

> *all $\tilde{T} \in \mathcal{F}_\tau(T)$, as well as all $(d-1)$-dimensional hyperfaces of T, contain a vertex of τ' in their interiors.*

See Figure 5.1 for an illustration of the case $d = 2$. The following proposition was shown in [BMN02, Lemma 5.3]. (Actually, a somewhat stronger condition on the

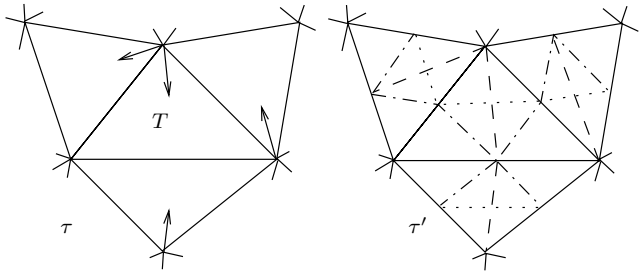


FIG. 5.1. *A refinement τ' of τ, which is a full refinement with respect to a triangle $T \in \tau$. The arrows in τ emanate from the newest vertices.*

refinement was imposed there, but not used; a more general $\mathbf{f}$ was considered at the expense of an additional "oscillation" term in the expression; and, finally, there our general $\mathbf{w}_\tau \in \mathbb{V}_\tau$ reads as $\mathbf{u}_\tau^{r_\sigma}$, whose additional property of being the solution of (4.5) was not used though.)

PROPOSITION 5.2. *Let $\tau \supseteq \sigma$ be partitions, let $r_\sigma \in \mathbb{P}_\sigma$, and let us assume that $\mathbf{f} \in \mathbb{V}_\tau^*$. Let $\tau' \supset \tau$ be a full refinement of τ with respect to $T \in \tau$. Then for any $\mathbf{w}_\tau \in \mathbb{V}_\tau$, we have*

$$\eta_T(\mathbf{f}, r_\sigma, \mathbf{w}_\tau) \lesssim \sum_{\tilde{T} \in \mathcal{F}_\tau(T)} |\mathbf{u}_{\tau'}^{r_\sigma} - \mathbf{w}_\tau|^2_{H^1(\tilde{T})^d}.$$

As a straightforward consequence we have the following.

COROLLARY 5.3. *In the situation of Proposition* 5.2, *let $\tau' \supseteq \tau$ be a full refinement of τ with respect to all T from some $\underline{F} \subset \tau$. Then*

$$c_2 \left[\sum_{T \in \underline{F}} \eta_T(\mathbf{f}, r_\sigma, \mathbf{w}_\tau) \right]^{\frac{1}{2}} \leq \|\mathbf{u}_{\tau'}^{r_\sigma} - \mathbf{w}_\tau\|_{\mathbb{V}}$$

for some absolute constant $c_2 > 0$. In particular, we have

$$c_2 \mathcal{E}^{\mathrm{E}}(\tau, \mathbf{f}, r_\sigma, \mathbf{w}_\tau) \leq \|\mathbf{u}^{r_\sigma} - \mathbf{w}_\tau\|_{\mathbb{V}}. \tag{5.1}$$

Finally in this subsection, we investigate the stability of the elliptic error estimator.

PROPOSITION 5.4. *Let $\tau \supseteq \sigma$ be partitions, and let $r_\sigma \in \mathbb{P}_\sigma$, $\mathbf{f} \in L_2(\Omega)^d$, and $\mathbf{v}_\tau, \mathbf{w}_\tau \in \mathbb{V}_\tau$. Then*

$$c_2 |\mathcal{E}^{\mathrm{E}}(\tau, \mathbf{f}, r_\sigma, \mathbf{v}_\tau) - \mathcal{E}^{\mathrm{E}}(\tau, \mathbf{f}, r_\sigma, \mathbf{w}_\tau)| \leq \|\mathbf{v}_\tau - \mathbf{w}_\tau\|_{\mathbb{V}}.$$

Proof. For $\mathbf{g} \in L_2(\Omega)^d$, $q_\sigma \in \mathbb{P}_\sigma$, by two applications of the triangle inequality in the form $\big|\|\cdot\| - \|\cdot\|\big|^2 \leq \|\cdot - \cdot\|^2$, first for vectors and then for functions, we have

$$|\mathcal{E}^{\mathrm{E}}(\tau, \mathbf{f}, r_\sigma, \mathbf{v}_\tau) - \mathcal{E}^{\mathrm{E}}(\tau, \mathbf{g}, q_\sigma, \mathbf{w}_\tau)| \leq \mathcal{E}^{\mathrm{E}}(\tau, \mathbf{f} - \mathbf{g}, r_\sigma - q_\sigma, \mathbf{v}_\tau - \mathbf{w}_\tau).$$

By substituting $\mathbf{g} = \mathbf{f}$ and $q_\sigma = r_\sigma$, and by applying (5.1), the proof is completed. ∎

5.2. A posteriori error estimator for the (reduced) Stokes problem. For given $\mathbf{f} \in L_2(\Omega)^d$ and partitions $\tau \supseteq \sigma$ and ϱ, where we think of either $\varrho = \infty$ (full Stokes) or $\varrho = \sigma$ (reduced Stokes), $r_\sigma \in \mathbb{P}_\sigma$ being an approximation for p_ϱ, we study an a posteriori error estimator for $\|\mathbf{u}^{p_\varrho} - \mathbf{u}_\tau^{r_\sigma}\|_{\mathbb{V}} + \|p_\varrho - r_\sigma\|_{\mathbb{P}}$, or whenever possible, for this quantity with $\mathbf{u}_\tau^{r_\sigma}$ replaced by a general $\mathbf{w}_\tau \in \mathbb{V}_\tau$. We set the estimator

$$\mathcal{E}^{\mathrm{S}}(\varrho, \tau, \mathbf{f}, r_\sigma, \mathbf{w}_\tau) := \mathcal{E}^{\mathrm{E}}(\tau, \mathbf{f}, r_\sigma, \mathbf{w}_\tau) + \|Q_\varrho \mathrm{div} \mathbf{w}_\tau\|_{L_2(\Omega)}.$$

PROPOSITION 5.5. *For partitions $\tau \supseteq \sigma$ and ϱ, $r_\sigma \in \mathbb{P}_\sigma$, and $\mathbf{f} \in L_2(\Omega)^d$, we have*

$$\|\mathbf{u}^{p_\varrho} - \mathbf{u}_\tau^{r_\sigma}\|_{\mathbb{V}} + \|p_\varrho - r_\sigma\|_{\mathbb{P}} \leq C_3 \mathcal{E}^{\mathrm{S}}(\varrho, \tau, \mathbf{f}, r_\sigma, \mathbf{u}_\tau^{r_\sigma})$$

for some absolute constant $C_3 > 0$.

The proof given in [Ver96] (cf. [BMN02, Lemma 4.1]) for $\varrho = \infty$ easily generalizes to $\varrho \subsetneq \infty$. More details can be found in [KS07].

The first statement of the following proposition was shown in [Ver96] (cf. [BMN02, Lemma 4.3]) for the case $\varrho = \infty$, but the proof generalizes immediately to general partitions ϱ. The second statement follows easily by using $Q_\varrho \mathrm{div} \mathbf{u}^{p_\varrho} = 0$.

PROPOSITION 5.6. *Let $\tau \supseteq \sigma$ and ϱ be partitions, let $r_\sigma \in \mathbb{P}_\sigma$ and $\mathbf{w}_\tau \in \mathbb{V}_\tau$, and let us assume that $\mathbf{f} \in \mathbb{V}_\tau^*$. Then*

$$\eta_T(\mathbf{f}, r_\sigma, \mathbf{w}_\tau) \lesssim \sum_{\tilde{T} \in \mathcal{F}_\tau(T)} \left[|\mathbf{u}^{p_\varrho} - \mathbf{w}_\tau|^2_{H^1(\tilde{T})^d} + \|p_\varrho - r_\sigma\|^2_{L_2(\tilde{T})^d} \right] \quad (T \in \tau)$$

and

$$c_4 \mathcal{E}^{\mathrm{S}}(\varrho, \tau, \mathbf{f}, r_\sigma, \mathbf{w}_\tau) \leq \|\mathbf{u}^{p_\varrho} - \mathbf{w}_\tau\|_{\mathbb{V}} + \|p_\varrho - r_\sigma\|_{\mathbb{P}}$$

for some absolute constant $c_4 > 0$.

The last result in this subsection provides an a posteriori error estimator for the outer elliptic problem.

PROPOSITION 5.7. *For a partition ϱ, and for $r \in \mathbb{P}$, we have*

$$c_6 \|Q_\varrho \mathrm{div} \mathbf{u}^r\|_{L_2(\Omega)} \leq \|p_\varrho - r\|_{\mathbb{P}} \leq C_5 \|Q_\varrho \mathrm{div} \mathbf{u}^r\|_{L_2(\Omega)}$$

for some absolute constants $C_5, c_6 > 0$.

Proof. Use $\sup_{0 \neq \mathbf{v} \in \mathbb{V}} \frac{b(\mathbf{v}, p_\varrho - r)}{\|\mathbf{v}\|_{\mathbb{V}}} = \sup_{0 \neq \mathbf{v} \in \mathbb{V}} \frac{a(\mathbf{u}^r - \mathbf{u}^{p_\varrho}, \mathbf{v})}{\|\mathbf{v}\|_{\mathbb{V}}} = \|\mathbf{u}^{p_\varrho} - \mathbf{u}^r\|_{\mathbb{V}}$, and thus

$$\beta \leq \frac{\|\mathbf{u}^{p_\varrho} - \mathbf{u}^r\|_{\mathbb{V}}}{\|p_\varrho - r\|_{\mathbb{P}}} \leq 1, \tag{5.2}$$

and

$$\begin{aligned} &\|p_\varrho - r\|_{\mathbb{P}} + \|\mathbf{u}^{p_\varrho} - \mathbf{u}^r\|_{\mathbb{V}} \\ &\eqsim \sup_{0 \neq (\mathbf{v}, q_\varrho) \in \mathbb{V} \times \mathbb{P}_\varrho} \frac{a(\mathbf{u}^{p_\varrho} - \mathbf{u}^r, \mathbf{v}) + b(\mathbf{v}, p_\varrho - r) + b(\mathbf{u}^{p_\varrho} - \mathbf{u}^r, q_\varrho)}{\|\mathbf{v}\|_{\mathbb{V}} + \|q_\varrho\|_{\mathbb{P}}} \\ &= \|Q_\varrho \mathrm{div} \mathbf{u}^r\|_{L_2(\Omega)}. \qquad \square \end{aligned}$$

Clearly, the evaluation of this estimator $\|Q_\varrho \mathrm{div} \mathbf{u}^r\|_{L_2(\Omega)}$ is not feasible, and so later $\mathbf{u}^r$ will be replaced by an approximation.

6. Adaptive refinements resulting in error reduction. For both elliptic problems $Sp = BA^{-1}f$ and $a(\mathbf{u}^r, \mathbf{v}) = \mathbf{f}(\mathbf{v}) - b(\mathbf{v}, r)$ $(\mathbf{v} \in \mathbb{V})$, the latter for some given $r \in \mathbb{P}$, we construct adaptive refinement routines based on the a posteriori error estimators. Given (approximate) Galerkin solutions from $\mathbb{P}_\sigma$ or $\mathbb{V}_\tau$, respectively, they produce refinements $\tilde{\sigma} \supset \sigma$ or $\tilde{\tau} \supset \tau$ such that the Galerkin solutions with respect to these partitions have strictly smaller errors. Moreover, we will give bounds on the number of refined simplices which eventually will lead to the conclusion that our adaptive Stokes solver generates quasi-optimal partitions.

6.1. Adaptive pressure refinements. With C_5, c_6 being the constants from Proposition 5.7, for some absolute constants

$$d \in \left(1 - \frac{c_6^2}{C_5^2}, 1\right], \quad \theta \in \left(0, \left[1 - \frac{1 - c_6^2/C_5^2}{d}\right]^{\frac{1}{2}}\right), \quad D \geq 1, \tag{6.1}$$

we assume that we have the following routine available. We think of its argument $\mathbf{w}$ as being an approximation to $\mathbf{u}^{r_\sigma}$, where r_σ is some approximation to p from $\mathbb{P}_\sigma$.

REFpres$[\sigma, \mathbf{w}] \to \tilde{\sigma}$
% σ is a partition and $\mathbf{w} \in \mathbb{V}$.
Select a partition $\tilde{\sigma} \supseteq \sigma$ with

$$\|Q_{\tilde{\sigma}} \mathrm{div}\mathbf{w}\|_{L_2(\Omega)} \geq \theta \|\mathrm{div}\mathbf{w}\|_{L_2(\Omega)}, \tag{6.2}$$

such that

$$\#\tilde{\sigma} - \#\sigma \leq D(\#\breve{\sigma} - \#\sigma)$$

for any $\breve{\sigma} \supseteq \sigma$ with

$$\|Q_{\breve{\sigma}} \mathrm{div}\mathbf{w}\|_{L_2(\Omega)} \geq \sqrt{1 - d(1 - \theta^2)} \|\mathrm{div}\mathbf{w}\|_{L_2(\Omega)}. \tag{6.3}$$

Remark 6.1. Eventually, we will make calls $\tilde{\sigma} := \mathbf{REFpres}[\sigma, \mathbf{w}]$ only for the argument $\mathbf{w}$ from $\mathbb{V}_\tau$ for some $\tau \supseteq \sigma$. Since then $\mathrm{div}\mathbf{w} \in \mathbb{P}_\tau$, in those cases we may always assume that $\sigma \subseteq \tilde{\sigma} \subseteq \tau$. The fact that the partition underlying the pressure approximation is always contained in or equal to that underlying the velocity approximation will be essential for the forthcoming adaptive refinement routine for reducing the error in the inner elliptic problem.

The benefit of **REFpres** can be seen from the following two lemmas.

LEMMA 6.2. *Let σ be a partition, and let $r_\sigma \in \mathbb{P}_\sigma$. Then for $\tilde{\sigma} = \mathbf{REFpres}[\sigma, \mathbf{u}^{r_\sigma}]$, we have*

$$\|p - p_{\tilde{\sigma}}\|_{\mathbb{P}} \leq \left[1 - \tfrac{c_6^2 \theta^2}{C_5^2}\right]^{\frac{1}{2}} \|p - r_\sigma\|_{\mathbb{P}}.$$

Moreover,

$$\#\tilde{\sigma} - \#\sigma \leq D[\#\bar{\sigma} - \#\tau_0]$$

for any *partition $\bar{\sigma}$ for which*

$$\inf_{q_{\bar{\sigma}} \in \mathbb{P}_{\bar{\sigma}}} \|p - q_{\bar{\sigma}}\|_{\mathbb{P}} \leq \left[1 - \tfrac{C_5^2}{c_6^2}(1 - d(1 - \theta^2))\right]^{\frac{1}{2}} \|p - r_\sigma\|_{\mathbb{P}}. \tag{6.4}$$

(Note that (6.1) implies that $\frac{C_5^2}{c_6^2}(1 - d(1 - \theta^2)) < 1$.)

Proof. Recall that $p_{\tilde\sigma}$ denotes the solution of a reduced Stokes problem, i.e., it is the best approximation to p from $\mathbb{P}_{\tilde\sigma}$ with respect to $\|\cdot\|_{\mathbb{P}}$. Therefore, the first statement follows from

$$\|p-r_\sigma\|_{\mathbb{P}}^2=\|p-p_{\tilde\sigma}\|_{\mathbb{P}}^2+\|p_{\tilde\sigma}-r_\sigma\|_{\mathbb{P}}^2$$

and $\|p_{\tilde\sigma}-r_\sigma\|_{\mathbb{P}}\ge c_6\|Q_{\tilde\sigma}\mathrm{div}\mathbf{u}^{r_\sigma}\|_{L_2(\Omega)}\ge c_6\theta\|\mathrm{div}\mathbf{u}^{r_\sigma}\|_{L_2(\Omega)}\ge\frac{c_6\theta}{C_5}\|p-r_\sigma\|_{\mathbb{P}}$ by Proposition 5.7.

For a $\bar\sigma$ satisfying (6.4), let $\breve\sigma=\sigma\cup\bar\sigma$. Then from $\|p-p_{\breve\sigma}\|_{\mathbb{P}}\le\inf_{q_{\bar\sigma}\in\mathbb{P}_{\bar\sigma}}\|p-q_{\bar\sigma}\|_{\mathbb{P}}$ and Proposition 5.7, with $\lambda:=\frac{C_5^2}{c_6^2}(1-d(1-\theta^2))$ we have

$$\begin{aligned}C_5^2\|Q_{\breve\sigma}\mathrm{div}\mathbf{u}^{r_\sigma}\|_{L_2(\Omega)}^2&\ge\|p_{\breve\sigma}-r_\sigma\|_{\mathbb{P}}^2=\|p-r_\sigma\|_{\mathbb{P}}^2-\|p-p_{\breve\sigma}\|_{\mathbb{P}}^2\\&\ge\lambda\|p-r_\sigma\|_{\mathbb{P}}^2\ge\lambda c_6^2\|\mathrm{div}\mathbf{u}^{r_\sigma}\|_{L_2(\Omega)}^2.\end{aligned}$$

Noting that $\frac{\lambda c_6^2}{C_5^2}=1-d(1-\theta^2)$, by construction of $\tilde\sigma$, we conclude that

$$\#\tilde\sigma-\#\sigma\le D[\#\breve\sigma-\#\sigma]\le D[\#\bar\sigma-\#\tau_0].\qquad\square$$

Now we generalize Lemma 6.2 to the practical relevant situation in which only an approximation to $\mathbf{u}^{r_\sigma}$ is available.

LEMMA 6.3. *Let $\omega\in(0,\theta)$ be a constant, let σ be a partition, let $r_\sigma\in\mathbb{P}_\sigma$, and let $\mathbf{w}\in\mathbb{V}$ with*

$$\|\mathrm{div}\mathbf{u}^{r_\sigma}-\mathrm{div}\mathbf{w}\|_{L_2(\Omega)}\le\omega\|\mathrm{div}\mathbf{w}\|_{L_2(\Omega)}.$$

Then for $\tilde\sigma=\mathbf{REFpres}[\sigma,\mathbf{w}]$, we have

$$\|p-p_{\tilde\sigma}\|_{\mathbb{P}}\le\big[1-\tfrac{c_6^2(\theta-\omega)^2}{C_5^2(1+\omega)^2}\big]^{\frac12}\|p-r_\sigma\|_{\mathbb{P}}.$$

Moreover, if ω is sufficiently small such that $\frac{\omega+\sqrt{1-d(1-\theta^2)}}{1-\omega}<\frac{c_6}{C_5}$, then

$$\#\tilde\sigma-\#\sigma\le D[\#\bar\sigma-\#\tau_0]$$

for any *partition $\bar\sigma$ for which, with $\xi:=\big[1-\big(\frac{\omega+\sqrt{1-d(1-\theta^2)}}{1-\omega}\frac{C_5}{c_6}\big)^2\big]^{\frac12}$,*

$$\inf_{q_{\bar\sigma}\in\mathbb{P}_{\bar\sigma}}\|p-q_{\bar\sigma}\|_{\mathbb{P}}\le\xi\|p-r_\sigma\|_{\mathbb{P}}.\tag{6.5}$$

Proof. The proof is similar to that of Lemma 6.2. For the first part use that

$$\|Q_{\tilde\sigma}\mathrm{div}\mathbf{u}^{r_\sigma}\|_{L_2(\Omega)}\ge\|Q_{\tilde\sigma}\mathrm{div}\mathbf{w}\|_{L_2(\Omega)}-\omega\|\mathrm{div}\mathbf{w}\|_{L_2(\Omega)}\ge\tfrac{\theta-\omega}{1+\omega}\|\mathrm{div}\mathbf{u}^{r_\sigma}\|_{L_2(\Omega)},$$

and, for the second part, with any $\bar\sigma$ satisfying (6.5) and $\breve\sigma=\sigma\cup\bar\sigma$, that

$$\begin{aligned}&C_5[\|Q_{\breve\sigma}\mathrm{div}\mathbf{w}\|_{L_2(\Omega)}+\omega\|\mathrm{div}\mathbf{w}\|_{L_2(\Omega)}]\ge C_5\|Q_{\breve\sigma}\mathrm{div}\mathbf{u}^{r_\sigma}\|_{L_2(\Omega)}\ge\|p_{\breve\sigma}-r_\sigma\|_{\mathbb{P}}\\&=\big[\|p-r_\sigma\|_{\mathbb{P}}^2-\|p-p_{\breve\sigma}\|_{\mathbb{P}}^2\big]^{\frac12}\ge\sqrt{1-\xi^2}\|p-r_\sigma\|_{\mathbb{P}}\ge c_6\sqrt{1-\xi^2}\|\mathrm{div}\mathbf{u}^{r_\sigma}\|_{L_2(\Omega)}\\&\ge(1-\omega)c_6\sqrt{1-\xi^2}\|\mathrm{div}\mathbf{w}\|_{L_2(\Omega)},\end{aligned}$$

or, equivalently, $\|Q_{\breve\sigma}\mathrm{div}\mathbf{w}\|_{L_2(\Omega)}\ge\sqrt{1-d(1-\theta^2)}\|\mathrm{div}\mathbf{w}\|_{L_2(\Omega)}$. $\square$

As mentioned previously, we will make calls $\tilde{\sigma} := \mathbf{REFpres}[\sigma, \mathbf{w}]$ only for the argument $\mathbf{w}$ from $\mathbb{V}_\tau$ for some $\tau \supseteq \sigma$, so that $\mathrm{div}\mathbf{w} \in \mathbb{P}_\tau$. Obviously, if **REFpres** is implemented as the selection for some $\theta \in (0, c_6/C_5)$ of the *smallest* partition $\tilde{\sigma} \supseteq \sigma$ such that (6.2) is valid, it satisfies its requirements with $d = 1 = D$. Yet, at least a naive implementation of this algorithm would require computing $\|Q_{\breve{\sigma}}\mathrm{div}\mathbf{w}\|_{L_2(\Omega)}$ for all partitions $\sigma \subsetneq \breve{\sigma} \subsetneq \tau$, which is prohibitively expensive.

Recalling that any partition corresponds to a subtree of the infinite binary master tree $\mathcal{T}_*$ that is determined by the initial partition τ_0, alternatively one may apply an adaptive tree approximation algorithm from [BD04]. Note that (6.2) and (6.3) are equivalent to $\|(I-Q_{\tilde{\sigma}})\mathrm{div}\mathbf{w}\|_{L_2(\Omega)} \leq \sqrt{1-\theta^2}\,\|\mathrm{div}\mathbf{w}\|_{L_2(\Omega)}$ and $\|(I-Q_{\breve{\sigma}})\mathrm{div}\mathbf{w}\|_{L_2(\Omega)} \leq \sqrt{d}\sqrt{1-\theta^2}\,\|\mathrm{div}\mathbf{w}\|_{L_2(\Omega)}$, respectively. Prescribing a $\theta \in (0, c_6/C_5)$, a call of the *thresholding second algorithm* starting with initial partition σ, "error functional" $e(T) = \inf_{q\in P_{m-1}(T)} \|(\mathrm{div}\mathbf{w})|_T - q\|^2_{L_2(T)}$ $(T \in \mathcal{T}_*)$, and tolerance $(1-\theta^2)\|\mathrm{div}\mathbf{w}\|^2_{L_2(\Omega)}$ fulfills the requirements on **REFpres** for any given $d < 1$ and $D := D(d) < \infty$, independent of σ, and thus for d, θ, D as in (6.1). This was demonstrated to the authors by Binev [Bin07], using the properties of $\mathcal{T}_*$ being a binary tree and $e(T) \geq e(T_1) + e(T_2)$ whenever $T_1 \cup T_2 = T$. (The proof of [BD04, Corollary 5.4] shows the properties mentioned in **REFpres** for some specific constants $d < 1 < D$ that might not fulfill the conditions from (6.1).) After precomputing the values $e(T)$ for any simplex T from any partition $\sigma \subsetneq \breve{\sigma} \subsetneq \tau$, which can be done in $\mathcal{O}(\#\tau)$ operations, this algorithm produces $\tilde{\sigma}$ as in (6.2) in $\mathcal{O}(\#\tilde{\sigma})$ additional operations. With this implementation of **REFpres** we thus have that for $\sigma \subseteq \tau$ and $\mathbf{w} \in \mathbb{V}_\tau$, the call $\mathbf{REFpres}[\sigma, \mathbf{w}]$ takes $\mathcal{O}(\#\tau)$ operations.

6.2. Adaptive velocity refinements. For some *fixed*

$$\zeta \in \left(0, \frac{c_2}{C_1}\right),$$

we will make use of the following routine to determine a suitable adaptive refinement for an update of the velocity:

$\mathbf{REFvel}[\tau, \mathbf{g}, r_\sigma, \mathbf{w}_\tau] \to \tilde{\tau}$
% τ is a partition, $\mathbf{g} \in L_2(\Omega)^d$, $r_\sigma \in \mathbb{P}_\sigma$ for some $\sigma \subseteq \tau$, and $\mathbf{w}_\tau \in \mathbb{V}_\tau$.
Select a set $\underline{F} \subset \tau$ with, up to some absolute factor, minimal cardinality *such that*

$$\sum_{T\in\underline{F}} \eta_T(\mathbf{g}, r_\sigma, \mathbf{w}_\tau) \geq \zeta^2\, \mathcal{E}^{\mathrm{E}}(\tau, \mathbf{g}, r_\sigma, \mathbf{w}_\tau)^2. \tag{6.6}$$

Construct the smallest $\tilde{\tau} \supset \tau$ which is a full refinement with respect to all $T \in \underline{F}$.

The next lemma will show the benefit of **REFvel**. It applies under the (unrealistic) assumptions that $\mathbf{f} \in \mathbb{V}^*_\tau$ and that the Galerkin problems are solved exactly. In Lemma 8.2, inexact Galerkin solutions will be allowed, and the given right-hand side $\mathbf{f} \in \mathbb{V}'$ will be replaced by approximations from $\mathbb{V}^*_\tau$ and $\mathbb{V}^*_{\tilde{\tau}}$, respectively.

Note that when $\mathbf{f} \in \mathbb{V}^*_\tau$, the computation of all $\eta_T(\mathbf{f}, r_\tau, \mathbf{w}_\tau)$ $(T \in \tau)$ can be done in $\mathcal{O}(\#\tau)$ operations. By doing an *approximate* sorting of the $\eta_T(\mathbf{f}_\tau, r_\tau, \mathbf{w}_\tau)$ by their values, $\mathbf{REFvel}[\tau, \mathbf{f}, \cdot, \cdot]$ can be implemented in $\mathcal{O}(\#\tau)$ operations (cf. [Ste07]).

LEMMA 6.4. *Let $\tau \supseteq \sigma$ be partitions, let $\mathbf{f} \in \mathbb{V}^*_\tau$, and let $r_\sigma \in \mathbb{P}_\sigma$. Then for $\tilde{\tau} = \mathbf{REFvel}[\tau, \mathbf{f}, r_\sigma, \mathbf{u}^{r_\sigma}_\tau]$, we have*

$$\|\mathbf{u}^{r_\sigma} - \mathbf{u}^{r_\sigma}_{\tilde{\tau}}\|_{\mathbb{V}} \leq \left[1 - \tfrac{c_2^2\zeta^2}{C_1^2}\right]^{\frac{1}{2}} \|\mathbf{u}^{r_\sigma} - \mathbf{u}^{r_\sigma}_{\tau}\|_{\mathbb{V}}.$$

Moreover, if $\zeta < \frac{c_2}{C_1}$*, and, for some absolute constant* $\vartheta > 0$*,*

$$\|\mathbf{u}^{r_\sigma} - \mathbf{u}_\tau^{r_\sigma}\|_{\mathbb{V}} \geq \vartheta \|\mathbf{u} - \mathbf{u}^{r_\sigma}\|_{\mathbb{V}}, \tag{6.7}$$

then for the set of marked simplices $\underline{F}$ *inside* **REFvel***, we have*

$$\#\underline{F} \lesssim \#\bar{\tau} + \#\bar{\sigma} + \#\sigma \tag{6.8}$$

for any partitions $\bar{\tau}$ *and* $\bar{\sigma}$ *for which*

$$\inf_{\mathbf{v}_{\bar{\tau}} \in \mathbb{V}_{\bar{\tau}}} \|\mathbf{u} - \mathbf{v}_{\bar{\tau}}\|_{\mathbb{V}} \leq \tfrac{1}{3}\big[1 - \tfrac{C_1^2\zeta^2}{c_2^2}\big]^{\frac{1}{2}} \|\mathbf{u}^{r_\sigma} - \mathbf{u}_\tau^{r_\sigma}\|_{\mathbb{V}}, \tag{6.9}$$

$$\inf_{q_{\bar{\sigma}} \in \mathbb{P}_{\bar{\sigma}}} \|p - q_{\bar{\sigma}}\|_{\mathbb{V}} \leq \tfrac{1}{3}\big[1 - \tfrac{C_1^2\zeta^2}{c_2^2}\big]^{\frac{1}{2}} \|\mathbf{u}^{r_\sigma} - \mathbf{u}_\tau^{r_\sigma}\|_{\mathbb{V}}. \tag{6.10}$$

Remark 6.5. Note that the bound on $\#\underline{F}$ in terms of $\mathbf{u}$ and p (via $\bar{\tau}$ and $\bar{\sigma}$) and σ can be shown only when (the variational formulation of) $-\triangle \mathbf{u}^{r_\sigma} = \mathbf{f} - \nabla r^\sigma$ is *solved not too accurately*, a restriction imposed by (6.7). This will be enforced by a timely stopping of the AFEM for solving this elliptic problem guided by a posteriori error estimators (cf. (7.20)). By assuming that $\mathbf{u}$ and p are in certain approximation classes, i.e., that these functions *can* be approximated with certain rates by finite element functions with respect to the *best partitions*, later we will derive quasi-optimal bounds for $\#\bar{\tau}$ and $\#\bar{\sigma}$, as well as for $\#\sigma$ via Lemma 6.3, and so in the end on $\#\underline{F}$. Without imposing (6.7), we would arrive at a similar bound on $\#\underline{F}$ only when we would assume that all approximations $\mathbf{u}^{r_\sigma}$ to $\mathbf{u}$ corresponding to all approximate pressures r_σ that are created in the adaptive method are similarly easy to approximate as $\mathbf{u}$ itself, which is an unverifiable assumption.

Proof. The first statement follows from

$$\|\mathbf{u}^{r_\sigma} - \mathbf{u}_\tau^{r_\sigma}\|_{\mathbb{V}}^2 = \|\mathbf{u}^{r_\sigma} - \mathbf{u}_{\tilde{\tau}}^{r_\sigma}\|_{\mathbb{V}}^2 + \|\mathbf{u}_{\tilde{\tau}}^{r_\sigma} - \mathbf{u}_\tau^{r_\sigma}\|_{\mathbb{V}}^2,$$

and $\|\mathbf{u}_{\tilde{\tau}}^{r_\sigma} - \mathbf{u}_\tau^{r_\sigma}\|_{\mathbb{V}} \geq c_2\zeta\mathcal{E}^{\mathrm{E}}(\tau, \mathbf{f}, r_\sigma, \mathbf{u}_\tau^{r_\sigma}) \geq \frac{c_2\zeta}{C_1}\|\mathbf{u}^{r_\sigma} - \mathbf{u}_\tau^{r_\sigma}\|_{\mathbb{V}}$ by Corollary 5.3 and Proposition 5.1.

Let $\hat{\tau}$ be a partition for which

$$\inf_{\mathbf{v}_{\hat{\tau}} \in \mathbb{V}_{\hat{\tau}}} \|\mathbf{u}^{r_\sigma} - \mathbf{v}_{\hat{\tau}}\|_{\mathbb{V}} \leq \big[1 - \tfrac{C_1^2\zeta^2}{c_2^2}\big]^{\frac{1}{2}} \|\mathbf{u}^{r_\sigma} - \mathbf{u}_\tau^{r_\sigma}\|_{\mathbb{V}}, \tag{6.11}$$

and let $\breve{\tau} = \tau \cup \hat{\tau}$. Then with $\overline{F} = \overline{F}(\tau, \breve{\tau})$ from Proposition 5.1, we have

$$\begin{aligned} C_1^2 \sum_{T \in \overline{F}} \eta_T(\mathbf{f}, r_\sigma, \mathbf{u}_\tau^{r_\sigma}) &\geq \|\mathbf{u}_{\breve{\tau}}^{r_\sigma} - \mathbf{u}_\tau^{r_\sigma}\|_{\mathbb{V}}^2 = \|\mathbf{u}^{r_\sigma} - \mathbf{u}_\tau^{r_\sigma}\|_{\mathbb{V}}^2 - \|\mathbf{u}^{r_\sigma} - \mathbf{u}_{\breve{\tau}}^{r_\sigma}\|_{\mathbb{V}}^2 \\ &\geq \tfrac{C_1^2\zeta^2}{c_2^2}\|\mathbf{u}^{r_\sigma} - \mathbf{u}_\tau^{r_\sigma}\|_{\mathbb{V}}^2 \geq C_1^2\zeta^2\mathcal{E}^{\mathrm{E}}(\tau, \mathbf{f}, r_\sigma, \mathbf{u}_\tau^{r_\sigma})^2. \end{aligned}$$

By construction of $\underline{F}$, we infer that

$$\#\underline{F} \lesssim \overline{F} \lesssim \#\breve{\tau} - \#\tau \leq \#\hat{\tau} - \#\tau_0. \tag{6.12}$$

It remains to bound $\#\hat{\tau} - \#\tau_0$ for a $\hat{\tau}$ as in (6.11). With $\bar{\sigma}$ as in (6.10), we write $\mathbf{u}^{r_\sigma} = (\mathbf{u}^{r_\sigma} - \mathbf{u}^{p_{\bar{\sigma}}}) + (\mathbf{u}^{p_{\bar{\sigma}}} - \mathbf{u}) + \mathbf{u}$ and approximate each of the three terms within

tolerance $\frac{1}{3}\big[1-\frac{C_1^2\zeta^2}{c_2^2}\big]^{\frac{1}{2}}$ with finite element functions (the second one with zero). From (5.2), we have

$$\|\mathbf{u}-\mathbf{u}^{p_{\bar\sigma}}\|_{\mathbb{V}} \le \|p-p_{\bar\sigma}\|_{\mathbb{P}} \le \tfrac{1}{3}\big[1-\tfrac{C_1^2\zeta^2}{c_2^2}\big]^{\frac{1}{2}}\|\mathbf{u}^{r_\sigma}-\mathbf{u}_\tau^{r_\sigma}\|_{\mathbb{V}}.$$

The vector field $\mathbf{w}=\mathbf{u}^{r_\sigma}-\mathbf{u}^{p_{\bar\sigma}}$ solves

$$a(\mathbf{w},\mathbf{v}) = -b(\mathbf{v}, r_\sigma - p_{\bar\sigma}) \quad (\mathbf{v}\in\mathbb{V}).$$

With $\hat\sigma=\sigma\cup\bar\sigma$, the error in its best approximation $\mathbf{w}_{\hat\sigma}$ from $\mathbb{V}_{\hat\sigma}$ can be bounded by

$$\begin{aligned}\|\mathbf{w}-\mathbf{w}_{\hat\sigma}\|_{\mathbb{V}} \le \|\mathbf{w}\|_{\mathbb{V}} &\le \|\mathbf{u}^{r_\sigma}-\mathbf{u}\|_{\mathbb{V}} + \|\mathbf{u}-\mathbf{u}^{p_{\bar\sigma}}\|_{\mathbb{V}}\\ &\le \left(\vartheta^{-1}+\frac{1}{3}\left[1-\frac{C_1^2\zeta^2}{c_2^2}\right]^{\frac{1}{2}}\right)\|\mathbf{u}^{r_\sigma}-\mathbf{u}_\tau^{r_\sigma}\|_{\mathbb{V}}.\end{aligned}$$

With $\hat\sigma_0=\hat\sigma$, for $k=1,2,\dots$ let $\hat\sigma_k\supset\hat\sigma_{k-1}$ be the smallest partition that is a full refinement with respect to all $T\in\hat\sigma_{k-1}$. Then using the fact that $r_\sigma-p_{\bar\sigma}\in\mathbb{P}_{\hat\sigma_0}$, as in the first part of this lemma, now with $\zeta=1$, we have $\|\mathbf{w}-\mathbf{w}_{\hat\sigma_k}\|_{\mathbb{V}} \le \big[1-\frac{C_1^2}{c_2^2}\big]^{\frac{k}{2}}\|\mathbf{w}-\mathbf{w}_{\hat\sigma}\|_{\mathbb{V}}$, where $\mathbf{w}_{\hat\sigma_k}$ denotes the best approximation to $\mathbf{w}$ from $\mathbb{V}_{\sigma_k}$. With k being the smallest integer with $\big[1-\frac{C_1^2}{c_2^2}\big]^{\frac{k}{2}}\big(\vartheta^{-1}+\frac{1}{3}\big[1-\frac{C_1^2\zeta^2}{c_2^2}\big]^{\frac{1}{2}}\big) \le \frac{1}{3}\big[1-\frac{C_1^2\zeta^2}{c_2^2}\big]^{\frac{1}{2}}$, we conclude that

$$\begin{aligned}\inf_{\mathbf{v}_{\bar\tau}\in\mathbb{V}_{\bar\tau}}\|\mathbf{u}^{r_\sigma}-(\mathbf{v}_{\bar\tau}+\mathbf{w}_{\hat\sigma_k})\|_{\mathbb{V}} &\le \|\mathbf{u}^{r_\sigma}-\mathbf{u}^{p_{\bar\sigma}}-\mathbf{w}_{\hat\sigma_k}\|_{\mathbb{V}} + \|\mathbf{u}^{p_{\bar\sigma}}-\mathbf{u}\|_{\mathbb{V}} + \inf_{\mathbf{v}_{\bar\tau}\in\mathbb{V}_{\bar\tau}}\|\mathbf{u}-\mathbf{v}_{\bar\tau}\|_{\mathbb{V}}\\ &\le \big[1-\tfrac{C_1^2\zeta^2}{c_2^2}\big]^{\frac{1}{2}}\|\mathbf{u}^{r_\sigma}-\mathbf{u}_\tau^{r_\sigma}\|_{\mathbb{V}}.\end{aligned}$$

Since $\mathbf{v}_{\bar\tau}+\mathbf{w}_{\hat\sigma_k}\in\mathbb{V}_{\bar\tau\cup\hat\sigma_k}$, and $\#(\bar\tau\cup\hat\sigma_k)-\#\tau_0 \lesssim \#\bar\tau+\#\bar\sigma+\#\sigma$ (dependent on k and thus on ϑ), in view of (6.11) and (6.12) the proof is complete. ∎

Remark 6.6. If in (6.7) $\vartheta > \big[1-\frac{C_1^2\zeta^2}{c_2^2}\big]^{-\frac{1}{2}}$, then by a simplification of the above proof, instead of (6.8), one obtains that

$$\#\underline{F} \lesssim \#\bar\tau - \#\tau_0$$

for any $\bar\tau$ with

$$\inf_{\mathbf{v}_{\bar\tau}\in\mathbb{V}_{\bar\tau}}\|\mathbf{u}-\mathbf{v}_{\bar\tau}\|_{\mathbb{V}} \le \big(\big[1-\tfrac{C_1^2\zeta^2}{c_2^2}\big]^{\frac{1}{2}}-\vartheta^{-1}\big)\|\mathbf{u}^{r_\sigma}-\mathbf{u}_\tau^{r_\sigma}\|_{\mathbb{V}},$$

which bound on $\#\underline{F}$ is in particular independent of the pressure p. It is, however, not clear whether under the condition $\|\mathbf{u}^{r_\sigma}-\mathbf{u}_\tau^{r_\sigma}\|_{\mathbb{V}} \ge \vartheta\|\mathbf{u}-\mathbf{u}^{r_\sigma}\|_{\mathbb{V}}$ for such ϑ the inner elliptic problem is solved with sufficient accuracy to obtain a convergent inexact Uzawa algorithm for solving the reduced Stokes problem of finding $(\mathbf{u}^{p_\sigma}, p_\sigma)$. Knowing that $\#\sigma$ is controlled in our outermost loop that produces Galerkin approximations to p, Lemma 6.4 provides a way to avoid the condition that $\vartheta > \big[1-\frac{C_1^2\zeta^2}{c_2^2}\big]^{-\frac{1}{2}}$. This point is the exact reason why we did not succeed in proving quasi-optimality of the Uzawa iteration for solving the full Stokes problem, i.e., without our outermost loop. Indeed, with that method there is no separate control over the partitions that underlie the pressure approximations.

7. An adaptive method for the Stokes problem in an idealized setting. For $s > 0$, we define the approximation class

$$\mathcal{A}^s_{\mathbb{V}} = \{\mathbf{v} \in \mathbb{V} : |\mathbf{v}|_{\mathcal{A}^s_{\mathbb{V}}} := \sup_{\varepsilon>0} \varepsilon \inf_{\{\tau:\inf_{\mathbf{v}_\tau\in\mathbb{V}_\tau}\|\mathbf{v}-\mathbf{v}_\tau\|_{\mathbb{V}}\le\varepsilon\}} [\#\tau - \#\tau_0]^s < \infty\}$$

and equip it with norm $\|\mathbf{v}\|_{\mathcal{A}^s_{\mathbb{V}}} := \|\mathbf{v}\|_{\mathbb{V}} + |\mathbf{v}|_{\mathcal{A}^s_{\mathbb{V}}}$. So $\mathcal{A}^s_{\mathbb{V}}$ is the class of vector fields that *can* be approximated within any given tolerance $\varepsilon > 0$ by a $\mathbf{v}_\tau \in \mathbb{V}_\tau$ for *some* partition τ with $\#\tau - \#\tau_0 \lesssim \varepsilon^{-1/s}|\mathbf{u}|^{1/s}_{\mathcal{A}^s_{\mathbb{V}}}$. Similarly, we define

$$\mathcal{A}^s_{\mathbb{P}} = \{q \in \mathbb{P} : |q|_{\mathcal{A}^s_{\mathbb{P}}} := \sup_{\varepsilon>0} \varepsilon \inf_{\{\sigma:\inf_{q_\sigma\in\mathbb{P}_\sigma}\|q-q_\sigma\|_{\mathbb{P}}\le\varepsilon\}} [\#\sigma - \#\tau_0]^s < \infty\}$$

and equip it with norm $\|q\|_{\mathcal{A}^s_{\mathbb{P}}} := \|q\|_{\mathbb{P}} + |q|_{\mathcal{A}^s_{\mathbb{P}}}$.

Because of the polynomial degrees of our approximations, only for $s \le m/d$ can membership of $\mathbf{u} \in \mathcal{A}^s_{\mathbb{V}}$ or $p \in \mathcal{A}^s_{\mathbb{P}}$ be enforced by imposing suitable smoothness conditions on $\mathbf{u}$ or p, respectively. These smoothness conditions, however, are much milder than requiring that $\mathbf{u} \in H^{1+sd}(\Omega)^d$ or $p \in H^{sd}(\Omega)$, which would be needed when only uniformly refined partitions were considered. The approximation classes can be (nearly) characterized as certain Besov spaces (see [BDDP02] for details). In any case for $d = 2$, polygonal domains, and sufficiently smooth $\mathbf{f}$, it is known (see [Dah99]) that $\mathbf{u}$ and p have sufficient Besov smoothness so that they are in $\mathcal{A}^s_{\mathbb{V}}$ or $\mathcal{A}^s_{\mathbb{P}}$ for any $s < m/d$, and thus also in the presence of reentrant corners.

The results derived in section 6.2 concerning adaptive velocity refinements were valid under the assumption that $\mathbf{f}$ was piecewise polynomial of degree $m-1$ with respect to the current partition. In order to make our exposition not too complicated, *in this section we will assume that* $\mathbf{f}$ *is piecewise polynomial of degree* $m-1$ *with respect to any partition that we encounter, i.e., that is in* $\mathbb{V}^*_{\tau_0}$. In the next section, we will remove this restriction. Furthermore, *in this section we assume that the arising finite-dimensional linear systems are solved exactly*, i.e., we do not care about the computational cost. In the next section, by applying iterative solvers, we will show quasi-optimal computational complexity.

The following algorithm is an implementation of the solution method that was announced in section 4 with the simplifications mentioned above. Note that the `do`-loop in the algorithm actually consists of 3 nested loops over i, j, and k. The loop over i concerns an adaptive method for solving p from the Schur complement equation. The loop over j concerns the Uzawa method for solving p^{σ_i} from the reduced Stokes problem. Finally, the loop over k concerns an adaptive finite element method for approximating the solution $\mathbf{u}^{p_j^{(i)}} \in \mathbb{V}$ of $a(\mathbf{u}^{p_j^{(i)}}, \mathbf{v}) = \mathbf{f}(\mathbf{v}) - b(\mathbf{v}, p_j^{(i)})$ $(\mathbf{v} \in \mathbb{V})$. We have formulated these loops as one loop to deal efficiently with the complicated stopping criteria. For example, the innermost one stops when either $\mathcal{E}^{\mathrm{E}}(\cdots) \le \alpha\mathcal{E}^{\mathrm{S}}(\sigma_i, \dots)$ or $\mathcal{E}^{\mathrm{S}}(\sigma_i, \dots) \le \kappa\mathcal{E}^{\mathrm{S}}(\infty, \dots)$ or $\mathcal{E}^{\mathrm{S}}(\infty, \dots) \le \varepsilon$, i.e., when either

$$\|\mathbf{u}^{p_j^{(i)}} - \mathbf{u}^{(i)}_{j,k}\|_{\mathbb{V}} \le C_1\alpha c_4^{-1}[\|p_{\sigma_i} - p_j^{(i)}\|_{\mathbb{P}} + \|\mathbf{u}^{p_{\sigma_i}} - \mathbf{u}^{(i)}_{j,k}\|_{\mathbb{V}}] \quad \text{or}$$

$$\|p^{\sigma_i} - p_j^{(i)}\|_{\mathbb{P}} + \|\mathbf{u}^{p_{\sigma_i}} - \mathbf{u}^{(i)}_{j,k}\|_{\mathbb{V}} \le C_3\kappa c_4^{-1}[\|p - p_j^{(i)}\|_{\mathbb{P}} + \|\mathbf{u} - \mathbf{u}^{(i)}_{j,k}\|_{\mathbb{V}}] \quad \text{or}$$

$$\|p - p_j^{(i)}\|_{\mathbb{P}} + \|\mathbf{u} - \mathbf{u}^{(i)}_{j,k}\|_{\mathbb{V}} \le \varepsilon.$$

STOKESSOLVE$_0[f,\varepsilon] \to [\sigma_j^{(i)}, p_j^{(i)}, \tau_{j,k}^{(i)}, \mathbf{u}_{j,k}^{(i)}]$
% For this preliminary version of the adaptive solver it is assumed
% that $f \in \mathbb{V}_{\tau_0}^*$.
% Let the parameter ζ *from* **REFvel** *satisfy* $\zeta \in (0, \frac{c_2}{C_1})$, *and* θ *from* **REFpres**
% satisfy $\theta \in (0, [1 - \frac{1-c_6^2/C_5^2}{d}]^{\frac{1}{2}})$. *For some* $\omega \in (0,\theta)$ *small enough such that*
% $\frac{\omega+\sqrt{1-d(1-\theta^2)}}{1-\omega} < \frac{c_6}{C_5}$, *fix some sufficiently small constants* $\kappa, \alpha > 0$ *such that*
% $\kappa < 1$, $C_1\frac{\kappa}{1-\kappa} \le \omega$, $\kappa C_1 < c_4$, $[1 - \frac{2\kappa C_3}{c_4 - \kappa C_1}]^{-1}[1 - \frac{c_6^2(\theta-\omega)^2}{c_5^2(1+\omega)^2}]^{\frac{1}{2}} < 1$, $\alpha C_1 < c_4$,
% and $1 - \beta^2 + \frac{2\alpha C_1}{c_4 - \alpha C_1} < 1$.
$p_0^{(0)} := 0$, $\sigma_0 := \tau_{0,0}^{(0)} := \tau_0$
$i := j := k := 0$
`do` $\mathbf{u}_{j,k}^{(i)} := \mathbf{u}_{\tau_{j,k}^{(i)}}^{p_j^{(i)}}$ *% i.e.,* $\mathbf{u}_{j,k}^{(i)} \in \mathbb{V}_{\tau_{j,k}^{(i)}}$, $a(\mathbf{u}_{j,k}^{(i)}, \mathbf{v}) = \mathbf{f}(\mathbf{v}) - b(\mathbf{v}, p_j^{(i)})$, $(\mathbf{v} \in \mathbb{V}_{\tau_{j,k}^{(i)}})$
  `if` $C_3\mathcal{E}^{\mathrm{S}}(\infty, \tau_{j,k}^{(i)}, \mathbf{f}, p_j^{(i)}, \mathbf{u}_{j,k}^{(i)}) \le \varepsilon$ `then stop`
  `elseif` $\mathcal{E}^{\mathrm{S}}(\sigma_i, \tau_{j,k}^{(i)}, \mathbf{f}, p_j^{(i)}, \mathbf{u}_{j,k}^{(i)}) \le \kappa\mathcal{E}^{\mathrm{S}}(\infty, \tau_{j,k}^{(i)}, \mathbf{f}, p_j^{(i)}, \mathbf{u}_{j,k}^{(i)})$ `then`
    $\sigma_{i+1} := \mathbf{REFpres}[\sigma_i, \mathbf{u}_{j,k}^{(i)}]$
    $p_0^{(i+1)} := p_j^{(i)}$, $\tau_{0,0}^{(i+1)} := \tau_{j,k}^{(i)}$
    $i{+}{+}$, $j := k := 0$
  `elseif` $\mathcal{E}^{\mathrm{E}}(\tau_{j,k}^{(i)}, \mathbf{f}, p_j^{(i)}, \mathbf{u}_{j,k}^{(i)}) \le \alpha\mathcal{E}^{\mathrm{S}}(\sigma_i, \tau_{j,k}^{(i)}, \mathbf{f}, p_j^{(i)}, \mathbf{u}_{j,k}^{(i)})$ `then`
    $p_{j+1}^{(i)} := p_j^{(i)} - Q_{\sigma_i}\mathrm{div}\mathbf{u}_{j,k}^{(i)}$
    $\tau_{j+1,0}^{(i)} := \tau_{j,k}^{(i)}$
    $j{+}{+}$, $k := 0$
  `else` $\tau_{j,k+1}^{(i)} := \mathbf{REFvel}[\tau_{j,k}^{(i)}, \mathbf{f}, p_j^{(i)}, \mathbf{u}_{j,k}^{(i)}]$
    $k{+}{+}$
  `endif`
`enddo`

THEOREM 7.1. (I) *Let* $\mathbf{f} \in \mathbb{V}_{\tau_0}^*$. *Then* $[\sigma_j^{(i)}, p_j^{(i)}, \tau_{j,k}^{(i)}, \mathbf{u}_{j,k}^{(i)}] := \mathbf{STOKESSOLVE}_0[f,\varepsilon]$ *terminates, and* $\|\mathbf{u} - \mathbf{u}_{j,k}^{(i)}\|_{\mathbb{V}} + \|p - p_j^{(i)}\|_{\mathbb{P}} \le \varepsilon$. (II) *If, for some* $s > 0$, $p \in \mathcal{A}_{\mathbb{P}}^{\mathbf{s}}$, *then* $\#\sigma_j^{(i)} - \#\tau_0 \lesssim \varepsilon^{-1/s}|p|_{\mathcal{A}_{\mathbb{P}}^{\mathbf{s}}}^{1/s}$, *dependent only on* τ_0, *and on* s *when it tends to* 0 *or infinity. If, in addition, for some* $\tilde{s} > 0$, $\mathbf{u} \in \mathcal{A}_{\mathbb{V}}^{\tilde{\mathbf{s}}}$, *then with* $\bar{s} = \min(s, \tilde{s})$, $\#\tau_{j,k}^{(i)} - \#\tau_0 \lesssim \varepsilon^{-1/\bar{s}}(\|p\|_{\mathcal{A}_{\mathbb{P}}^{\bar{\mathbf{s}}}}^{1/\bar{s}} + \|\mathbf{u}\|_{\mathcal{A}_{\mathbb{V}}^{\bar{\mathbf{s}}}}^{1/\bar{s}})$, *dependent only on* τ_0, *and on* $\bar{s}$ *when it tends to* 0 *or infinity.*

Remark 7.2. In view of the assumptions, note that generally $\#\sigma_j^{(i)} - \#\tau_0$ is at most a constant multiple larger than this expression for the *best partition* $\sigma_j^{(i)}$ giving rise to such an error in the pressure. Similarly, generally $\#\tau_{j,k}^{(i)} - \#\tau_0$ is at most a constant multiple larger than this expression for the *best partition* $\tau_{j,k}^{(i)}$ on which p and $\mathbf{u}$ can be approximated by piecewise polynomials of degree $m-1$, or continuous piecewise polynomials of degree m with errors less than or equal to ε in $\|\cdot\|_{\mathbb{P}}$ or $\|\cdot\|_{\mathbb{V}}$, respectively.

Proof. (I) Given i and j, $\underline{k} = \underline{k}(i,j)$ will denote the maximum value attained by k for those i and j. Given an i, $\underline{j} = \underline{j}(i)$ will denote the maximum value attained by j for that i, and $\underline{\underline{k}} = \underline{\underline{k}}(i) := \underline{k}(i, \underline{j}(i))$ is the the maximum value attained by k for that i. Finally, $\underline{\underline{i}}$ will denote the maximum value attained by i.

At the moment $C_3\mathcal{E}^{\mathrm{S}}(\infty,\tau^{(i)}_{j,k},\mathbf{f},p^{(i)}_j,\mathbf{u}^{(i)}_{j,k})\le\varepsilon$ is passed, i.e., the loop over i, and thus the algorithm terminates, then, by definition $(i,j,k)=(\underline{i},\underline{j}(\underline{i}),\underline{\underline{k}}(\underline{i}))$, and we have $\|\mathbf{u}-\mathbf{u}^{(\underline{i})}_{\underline{j},\underline{\underline{k}}}\|_{\mathbb{V}}+\|p-p^{(\underline{i})}_{\underline{j}}\|_{\mathbb{P}}\le\varepsilon$ by Proposition 5.5. Next, we show that indeed the loop over i terminates. In doing so, we first assume that each inner loop over j terminates, which will be shown afterwards.

The inequality $\mathcal{E}^{\mathrm{S}}(\sigma_i,\tau^{(i)}_{j,k},\mathbf{f},p^{(i)}_j,\mathbf{u}^{(i)}_{j,k})\le\kappa\mathcal{E}^{\mathrm{S}}(\infty,\tau^{(i)}_{j,k},\mathbf{f},p^{(i)}_j,\mathbf{u}^{(i)}_{j,k})$ is equivalent to $\mathcal{E}^{\mathrm{E}}(\tau^{(i)}_{j,k},\mathbf{f},p^{(i)}_j,\mathbf{u}^{(i)}_{j,k})\le(1-\kappa)^{-1}(\kappa\|\mathrm{div}\mathbf{u}^{(i)}_{j,k}\|_{L_2(\Omega)}-\|Q_{\sigma_i}\mathrm{div}\mathbf{u}^{(i)}_{j,k}\|_{L_2(\Omega)})$. So at the moment that this test is passed, i.e., a loop over j terminates, then by definition $(j,k)=(\underline{j}(i),\underline{\underline{k}}(i))$, and by Propositions 5.1, 5.6, and 5.5,

$$C_1^{-1}\|\mathbf{u}^{p^{(i)}_{\underline{j}}}-\mathbf{u}^{(i)}_{\underline{j},\underline{\underline{k}}}\|_{\mathbb{V}}\le\mathcal{E}^{\mathrm{E}}(\tau^{(i)}_{\underline{j},\underline{\underline{k}}},\mathbf{f},p^{(i)}_{\underline{j}},\mathbf{u}^{(i)}_{\underline{j},\underline{\underline{k}}})\le\tfrac{\kappa}{1-\kappa}\|\mathrm{div}\mathbf{u}^{(i)}_{\underline{j},\underline{\underline{k}}}\|_{L_2(\Omega)},\tag{7.1}$$

$$C_1^{-1}\|\mathbf{u}^{p^{(i)}_{\underline{j}}}-\mathbf{u}^{(i)}_{\underline{j},\underline{\underline{k}}}\|_{\mathbb{V}}\le\kappa c_4^{-1}(\|\mathbf{u}-\mathbf{u}^{(i)}_{\underline{j},\underline{\underline{k}}}\|_{\mathbb{V}}+\|p-p^{(i)}_{\underline{j}}\|_{\mathbb{P}}),\tag{7.2}$$

$$C_3^{-1}\|p_{\sigma_i}-p^{(i)}_{\underline{j}}\|_{\mathbb{P}}\le\kappa c_4^{-1}(\|\mathbf{u}-\mathbf{u}^{(i)}_{\underline{j},\underline{\underline{k}}}\|_{\mathbb{V}}+\|p-p^{(i)}_{\underline{j}}\|_{\mathbb{P}}).\tag{7.3}$$

Furthermore, by (7.2), $\kappa\le c_4C_1^{-1}$, and (5.2), we have

$$\begin{aligned}\boxed{\|\mathbf{u}-\mathbf{u}^{(i)}_{\underline{j},\underline{\underline{k}}}\|_{\mathbb{V}}}&\le\|\mathbf{u}-\mathbf{u}^{p^{(i)}_{\underline{j}}}\|_{\mathbb{V}}+\tfrac{\kappa C_1}{c_4-\kappa C_1}(\|\mathbf{u}-\mathbf{u}^{p^{(i)}_{\underline{j}}}\|_{\mathbb{V}}+\|p-p^{(i)}_{\underline{j}}\|_{\mathbb{P}})\\&\le\boxed{\tfrac{c_4+\kappa C_1}{c_4-\kappa C_1}\|p-p^{(i)}_{\underline{j}}\|_{\mathbb{P}}},\end{aligned}\tag{7.4}$$

and so by (7.3),

$$\|p_{\sigma_i}-p^{(i)}_{\underline{j}}\|_{\mathbb{P}}\le\tfrac{2\kappa C_3}{c_4-\kappa C_1}\|p-p^{(i)}_{\underline{j}}\|_{\mathbb{P}}.\tag{7.5}$$

By (7.1), by $\|\mathrm{div}\cdot\|_{L_2(\Omega)}\le\|\cdot\|_{\mathbb{V}}$ on $\mathbb{V}$, and by $C_1\frac{\kappa}{1-\kappa}\le\omega$, Lemma 6.3 shows that for the output σ_{i+1} of the routine **REFpres** called subsequently, it holds that

$$\|p-p_{\sigma_{i+1}}\|_{\mathbb{P}}\le\big[1-\tfrac{c_6^2(\theta-\omega)^2}{C_5^2(1+\omega)^2}\big]^{\frac12}\|p-p^{(i)}_{\underline{j}}\|_{\mathbb{P}}.\tag{7.6}$$

Furthermore, since $\frac{\omega+\sqrt{1-d(1-\theta^2)}}{1-\omega}<\frac{c_6}{C_5}$ and in view of the definition of $\mathcal{A}^s_{\mathbb{P}}$, we have that

$$\#\sigma_{i+1}-\#\sigma_i\lesssim\|p-p^{(i)}_{\underline{j}}\|_{\mathbb{P}}^{-1/s}|p|^{1/s}_{\mathcal{A}^s_{\mathbb{P}}},\tag{7.7}$$

which will be used in the second part of the proof. Concerning (7.6), it is not $p_{\sigma_{i+1}}$ that is computed in the next loop over j, but instead its approximation $p^{(i+1)}_{\underline{j}(i+1)}$. Combining (7.5), with i reading as $i+1$, together with (7.6) shows that with $\rho_1:=[1-\frac{2\kappa C_3}{c_4-\kappa C_1}]^{-1}[1-\frac{c_6^2(\theta-\omega)^2}{c_5^2(1+\omega)^2}]^{\frac12}<1$,

$$\boxed{\|p-p^{(i+1)}_{\underline{j}(i+1)}\|_{\mathbb{P}}\le\rho_1\|p-p^{(i)}_{\underline{j}(i)}\|_{\mathbb{P}}.}\tag{7.8}$$

Since $c_4\mathcal{E}^{\mathrm{S}}(\infty,\tau^{(i)}_{\underline{j},\underline{\underline{k}}},\mathbf{f},p^{(i)}_{\underline{j}},\mathbf{u}^{(i)}_{\underline{j},\underline{\underline{k}}})\le\|\mathbf{u}-\mathbf{u}^{(i)}_{\underline{j},\underline{\underline{k}}}\|_{\mathbb{V}}+\|p-p^{(i)}_{\underline{j}}\|_{\mathbb{P}}$, together (7.8) and (7.4) show that indeed the loop over i, and thus the algorithm, terminates once we have

shown that each loop over j does, which is shown next. In doing so, we first assume that each inner loop over k terminates, which will be shown afterwards.

At the moment $\mathcal{E}^{\mathrm{E}}(\tau_{j,k}^{(i)},\mathbf{f},p_j^{(i)},\mathbf{u}_{j,k}^{(i)}) \leq \alpha\mathcal{E}^{\mathrm{S}}(\sigma_i,\tau_{j,k}^{(i)},\mathbf{f},p_j^{(i)},\mathbf{u}_{j,k}^{(i)})$ is passed, i.e., a loop over k terminates, then by definition $k=\underline{k}(i,j)$, and

$$\|\mathbf{u}^{p_j^{(i)}}-\mathbf{u}_{j,\underline{k}}^{(i)}\|_{\mathbb{V}} \leq \alpha C_1 c_4^{-1}(\|p_{\sigma_i}-p_j^{(i)}\|_{\mathbb{P}}+\|\mathbf{u}^{\sigma_i}-\mathbf{u}_{j,\underline{k}}^{(i)}\|_{\mathbb{V}}). \tag{7.9}$$

Estimating $\|\mathbf{u}^{\sigma_i}-\mathbf{u}_{j,\underline{k}}^{(i)}\|_{\mathbb{V}} \leq \|\mathbf{u}^{\sigma_i}-\mathbf{u}^{p_j^{(i)}}\|_{\mathbb{V}}+\|\mathbf{u}^{p_j^{(i)}}-\mathbf{u}_{j,\underline{k}}^{(i)}\|_{\mathbb{V}}$, and applying (7.9) and $\|\mathbf{u}^{\sigma_i}-\mathbf{u}^{p_j^{(i)}}\|_{\mathbb{V}} \leq \|p_{\sigma_i}-p_j^{(i)}\|_{\mathbb{P}}$ ((5.2)), we obtain that

$$\|\mathbf{u}^{\sigma_i}-\mathbf{u}_{j,\underline{k}}^{(i)}\|_{\mathbb{V}} \leq \tfrac{c_4+\alpha C_1}{c_4-\alpha C_1}\|p_{\sigma_i}-p_j^{(i)}\|_{\mathbb{P}} \tag{7.10}$$

and, by substituting this in (7.9), that

$$\boxed{\|\mathbf{u}^{p_j^{(i)}}-\mathbf{u}_{j,\underline{k}}^{(i)}\|_{\mathbb{V}} \leq \tfrac{2\alpha C_1}{c_4-\alpha C_1}\|p_{\sigma_i}-p_j^{(i)}\|_{\mathbb{P}}.} \tag{7.11}$$

For $p_{j+1}^{(i)} := p_j^{(i)}-Q_{\sigma_i}\mathrm{div}\mathbf{u}_{j,\underline{k}}^{(i)}$ that is evaluated subsequently, from $\|\mathrm{div}\cdot\|_{L_2(\Omega)} \leq \|\cdot\|_{\mathbb{V}}$, (7.11), and the statement (4.3) concerning convergence of the exact Uzawa iteration, with $\rho_2 := 1-\beta^2+\frac{2\alpha C_1}{c_4-\alpha C_1} < 1$, we have

$$\boxed{\|p_{\sigma_i}-p_{j+1}^{(i)}\|_{\mathbb{P}} \leq \rho_2\|p_{\sigma_i}-p_j^{(i)}\|_{\mathbb{P}}.} \tag{7.12}$$

Since $c_4\mathcal{E}^{\mathrm{S}}(\sigma_i,\tau_{j,\underline{k}}^{(i)},\mathbf{f},p_j^{(i)},\mathbf{u}_{j,\underline{k}}^{(i)}) \leq \|\mathbf{u}^{\sigma_i}-\mathbf{u}_{j,\underline{k}}^{(i)}\|_{\mathbb{V}}+\|p_{\sigma_i}-p_j^{(i)}\|_{\mathbb{P}}$, together (7.11) and (7.12) show that indeed each loop over j terminates by either $\mathcal{E}^{\mathrm{S}}(\sigma_i,\tau_{j,k}^{(i)},\mathbf{f},p_j^{(i)},\mathbf{u}_{j,k}^{(i)}) \leq \kappa\mathcal{E}^{\mathrm{S}}(\infty,\tau_{j,k}^{(i)},\mathbf{f},p_j^{(i)},\mathbf{u}_{j,k}^{(i)})$ or $C_3\mathcal{E}^{\mathrm{S}}(\infty,\tau_{j,k}^{(i)},\mathbf{f},p_j^{(i)},\mathbf{u}_{j,k}^{(i)}) \leq \varepsilon$ once we have shown that each loop over k terminates, which is shown next.

With $\rho_3 := [1-\frac{c_2^2\zeta^2}{C_1^2}]^{\frac{1}{2}} < 1$, Lemma 6.4 shows that

$$\boxed{\|\mathbf{u}^{p_j^{(i)}}-\mathbf{u}_{j,k+1}^{(i)}\|_{\mathbb{V}} \leq \rho_3\|\mathbf{u}^{p_j^{(i)}}-\mathbf{u}_{j,k}^{(i)}\|_{\mathbb{V}}.} \tag{7.13}$$

Since $c_2\mathcal{E}^{\mathrm{E}}(\tau_{j,k}^{(i)},\mathbf{f},p_j^{(i)},\mathbf{u}_{j,k}^{(i)}) \leq \|\mathbf{u}^{p_j^{(i)}}-\mathbf{u}_{j,k}^{(i)}\|_{\mathbb{V}}$, from (7.13) we conclude that indeed each loop over k terminates by either $\mathcal{E}^{\mathrm{E}}(\tau_{j,k}^{(i)},\mathbf{f},p_j^{(i)},\mathbf{u}_{j,k}^{(i)}) \leq \alpha\mathcal{E}^{\mathrm{S}}(\sigma_i,\tau_{j,k}^{(i)},\mathbf{f},p_j^{(i)},\mathbf{u}_{j,k}^{(i)})$ or $\mathcal{E}^{\mathrm{S}}(\sigma_i,\tau_{j,k}^{(i)},\mathbf{f},p_j^{(i)},\mathbf{u}_{j,k}^{(i)}) \leq \kappa\mathcal{E}^{\mathrm{S}}(\infty,\tau_{j,k}^{(i)},\mathbf{f},p_j^{(i)},\mathbf{u}_{j,k}^{(i)})$ or $C_3\mathcal{E}^{\mathrm{S}}(\infty,\tau_{j,k}^{(i)},\mathbf{f},p_j^{(i)},\mathbf{u}_{j,k}^{(i)}) \leq \varepsilon$. With this, part (I) of the theorem is proved.

Before starting with the second part, we collect some estimates for $\|p-p_{\underline{j}(0)}^{(0)}\|_{\mathbb{P}}$, $\|p_{\sigma_i}-p_0^{(i)}\|_{\mathbb{P}}$, and $\|\mathbf{u}^{p_j^{(i)}}-\mathbf{u}_{j,0}^{(i)}\|_{\mathbb{V}}$, i.e., the initial values for the recursions (7.8), (7.12), and (7.13) over i, j, and k, respectively.

From (7.5) we infer that

$$\|p-p_{\underline{j}(0)}^{(0)}\|_{\mathbb{P}} \leq [1-\tfrac{2\kappa C_3}{c_4-\kappa C_1}]^{-1}\|p-p_{\tau_0}\|_{\mathbb{P}} \leq [1-\tfrac{2\kappa C_3}{c_4-\kappa C_1}]^{-1}\|p\|_{\mathbb{P}}. \tag{7.14}$$

For $j=0$ and $i=0$, we have that

$$\|p_{\sigma_0}-p_0^{(0)}\|_{\mathbb{P}} \leq \|p_{\sigma_0}-p\|_{\mathbb{P}}+\|p\|_{\mathbb{P}} \leq 2\|p\|_{\mathbb{P}} \tag{7.15}$$

and, for $j=0$ and $i>0$, that

$$\|p_{\sigma_i}-p_0^{(i)}\|_{\mathbb{P}}=\|p_{\sigma_i}-p_{\underline{j}(i-1)}^{(i-1)}\|_{\mathbb{P}}\le\|p_{\sigma_i}-p\|_{\mathbb{P}}+\|p-p_{\underline{j}(i-1)}^{(i-1)}\|_{\mathbb{P}}$$

$$\le\big(\big[1-\tfrac{c_6^2(\theta-\omega)^2}{C_5^2(1+\omega)^2}\big]^{\frac12}+1\big)\|p-p_{\underline{j}(i-1)}^{(i-1)}\|_{\mathbb{P}} \tag{7.16}$$

by (7.6).

For $k=j=i=0$, we have

$$\|\mathbf{u}^{p_0^{(0)}}-\mathbf{u}_{0,0}^{(0)}\|_{\mathbb{V}}\le\|\mathbf{u}^{p_0^{(0)}}\|_{\mathbb{V}}=\|\mathbf{f}\|_{\mathbb{V}'}. \tag{7.17}$$

For $k=j=0$ and $i>0$, we have $p_0^{(i)}=p_{\underline{j}(i-1)}^{(i-1)}$ and $\tau_{0,0}^{(i)}=\tau_{\underline{j}(i-1),\underline{\underline{k}}(i-1)}^{(i-1)}$, and so

$$\|\mathbf{u}^{p_0^{(i)}}-\mathbf{u}_{0,0}^{(i)}\|_{\mathbb{V}}=\|\mathbf{u}^{p_{\underline{j}(i-1)}^{(i-1)}}-\mathbf{u}_{\underline{j}(i-1),\underline{\underline{k}}(i-1)}^{(i-1)}\|_{\mathbb{V}}\le\tfrac{2C_1\kappa}{c_4-\kappa C_1}\|p-p_{\underline{j}(i-1)}^{(i-1)}\|_{\mathbb{P}} \tag{7.18}$$

by (7.2) and (7.4). For $k=0$ and $j>0$, we have

$$\|\mathbf{u}^{p_j^{(i)}}-\mathbf{u}_{j,0}^{(i)}\|_{\mathbb{V}}\le\|\mathbf{u}^{p_j^{(i)}}-\mathbf{u}_{j-1,\underline{k}(j-1)}^{(i)}\|_{\mathbb{V}}$$

$$\le\|\mathbf{u}^{p_j^{(i)}}-\mathbf{u}^{p_{j-1}^{(i)}}\|_{\mathbb{V}}+\|\mathbf{u}^{p_{j-1}^{(i)}}-\mathbf{u}_{j-1,\underline{k}(j-1)}^{(i)}\|_{\mathbb{V}}.$$

Now using that

$$\|\mathbf{u}^{p_j^{(i)}}-\mathbf{u}^{p_{j-1}^{(i)}}\|_{\mathbb{V}}\le\|p_j^{(i)}-p_{j-1}^{(i)}\|_{\mathbb{P}}\le\|\mathbf{u}^{\sigma_i}-\mathbf{u}_{j-1,\underline{k}(j-1)}^{(i)}\|_{\mathbb{V}}$$

$$\le\|\mathbf{u}^{\sigma_i}-\mathbf{u}^{p_{j-1}^{(i)}}\|_{\mathbb{V}}+\|\mathbf{u}^{p_{j-1}^{(i)}}-\mathbf{u}_{j-1,\underline{k}(j-1)}^{(i)}\|_{\mathbb{V}}$$

$$\le\|p_{\sigma_i}-p_{j-1}^{(i)}\|_{\mathbb{P}}+\|\mathbf{u}^{p_{j-1}^{(i)}}-\mathbf{u}_{j-1,\underline{k}(j-1)}^{(i)}\|_{\mathbb{V}}$$

by (5.2) and

$$\|\mathbf{u}^{p_{j-1}^{(i)}}-\mathbf{u}_{j-1,\underline{k}(j-1)}^{(i)}\|_{\mathbb{V}}\le\tfrac{2\alpha C_1}{c_4-\alpha C_1}\|p_{\sigma_i}-p_{j-1}^{(i)}\|_{\mathbb{P}}$$

by (7.11), we find that

$$\|\mathbf{u}^{p_j^{(i)}}-\mathbf{u}_{j,0}^{(i)}\|_{\mathbb{V}}\le\big(1+\tfrac{4\alpha C_1}{c_4-\alpha C_1}\big)\|p_{\sigma_i}-p_{j-1}^{(i)}\|_{\mathbb{P}}. \tag{7.19}$$

(II) At the moment of a call $\sigma_{i+1}:=\mathbf{REFpres}[\sigma_i,\mathbf{u}_{\underline{j},\underline{\underline{k}}}^{(i)}]$, we have

$$\|p-p_{\underline{j}}^{(i)}\|_{\mathbb{P}}\ge c_4\mathcal{E}^{\mathrm{S}}(\infty,\tau_{\underline{j},\underline{\underline{k}}}^{(i)},\mathbf{f},p_{\underline{j}}^{(i)},\mathbf{u}_{\underline{j},\underline{\underline{k}}}^{(i)})>c_4C_3^{-1}\varepsilon,$$

so that in view of (7.7) and (7.8), we have

$$\#\sigma_{\underline{i}}-\#\tau_0\le\sum_{i=0}^{\underline{i}}\#\sigma_i-\#\sigma_{i-1}\lesssim\varepsilon^{-1/s}|p|_{\mathcal{A}_{\mathbb{P}}^s}^{1/s}.$$

At the moment of a call $\tau_{j,k+1}^{(i)}:=\mathbf{REFvel}[\tau_{j,k}^{(i)},\mathbf{f},p_j^{(i)},\mathbf{u}_{j,k}^{(i)}]$, we have

$$\|\mathbf{u}^{p_j^{(i)}}-\mathbf{u}_{j,k}^{(i)}\|_{\mathbb{V}}>c_2\mathcal{E}^{\mathrm{E}}(\tau_{j,k}^{(i)},\mathbf{f},p_j^{(i)},\mathbf{u}_{j,k}^{(i)})>c_2\alpha\mathcal{E}^{\mathrm{S}}(\sigma_i,\tau_{j,k}^{(i)},\mathbf{f},p_j^{(i)},\mathbf{u}_{j,k}^{(i)})$$

$$>c_2\alpha\kappa\mathcal{E}^{\mathrm{S}}(\infty,\tau_{j,k}^{(i)},\mathbf{f},p_j^{(i)},\mathbf{u}_{j,k}^{(i)})\ge c_2\alpha\kappa C_3^{-1}\|\mathbf{u}-\mathbf{u}_{j,k}^{(i)}\|_{\mathbb{V}}, \tag{7.20}$$

where for the first time in this proof we use the fact that the innermost iteration is stopped in time. In view of the definitions of $\mathcal{A}_{\mathbb{P}}^{\bar{s}}$ and $\mathcal{A}_{\mathbb{V}}^{\bar{s}}$, Lemma 6.4 now shows that the set of marked simplices $\underline{F}_{i,j,k} = \underline{F}$ inside $\mathbf{REFvel}[\tau_{j,k}^{(i)}, \mathbf{f}, p_j^{(i)}, \mathbf{u}_{j,k}^{(i)}]$ satisfies

$$\#\underline{F}_{i,j,k} \lesssim \|\mathbf{u}^{p_j^{(i)}} - \mathbf{u}_{j,k}^{(i)}\|_{\mathbb{V}}^{-1/\bar{s}}(|\mathbf{u}|_{\mathcal{A}_{\mathbb{V}}^{\bar{s}}}^{1/\bar{s}} + |p|_{\mathcal{A}_{\mathbb{P}}^{\bar{s}}}^{1/\bar{s}}) + \#\sigma_i - \#\tau_0 + \#\tau_0. \tag{7.21}$$

From $\|\mathbf{u}^{p_j^{(i)}} - \mathbf{u}_{j,k}^{(i)}\|_{\mathbb{V}} \le \|\mathbf{u}^{p_j^{(i)}}\|_{\mathbb{V}} \lesssim \|\mathbf{u}\|_{\mathbb{V}} + \|p\|_{\mathbb{P}}$, we have

$$\#\tau_0 \lesssim 1 \lesssim \|\mathbf{u}^{p_j^{(i)}} - \mathbf{u}_{j,k}^{(i)}\|_{\mathbb{V}}^{-1/\bar{s}}(\|\mathbf{u}\|_{\mathbb{V}}^{1/\bar{s}} + \|p\|_{\mathbb{P}}^{1/\bar{s}}).$$

For $i > 0$, we have $\#\sigma_i - \#\tau_0 \lesssim \|p - p_{\underline{j}(i-1)}^{(i-1)}\|_{\mathbb{P}}^{-1/\bar{s}}|p|_{\mathcal{A}_{\mathbb{P}}^{\bar{s}}}^{1/\bar{s}}$ by (7.7). By (7.18), (7.16), and (7.19), and the decrease of $\|p_{\sigma_i} - p_j^{(i)}\|_{\mathbb{P}}$ and $\|\mathbf{u}^{p_j^{(i)}} - \mathbf{u}_{j,k}^{(i)}\|_{\mathbb{V}}$ as a function of j or k, respectively, we have

$$\begin{aligned}\|p - p_{\underline{j}(i-1)}^{(i-1)}\|_{\mathbb{P}} &\gtrsim \begin{cases} \|\mathbf{u}^{p_0^{(i)}} - \mathbf{u}_{0,0}^{(i)}\|_{\mathbb{V}} \\ \|p_{\sigma_i} - p_0^{(i)}\|_{\mathbb{P}} \ge \|p_{\sigma_i} - p_{j-1}^{(i)}\|_{\mathbb{P}} \gtrsim \|\mathbf{u}^{p_j^{(i)}} - \mathbf{u}_{j,0}^{(i)}\|_{\mathbb{V}} \quad (j > 0) \end{cases} \\ &\gtrsim \|\mathbf{u}^{p_j^{(i)}} - \mathbf{u}_{j,k}^{(i)}\|_{\mathbb{V}}\end{aligned}$$

uniformly in i, j, k. We conclude that

$$\#\underline{F}_{i,j,k} \lesssim \|\mathbf{u}^{p_j^{(i)}} - \mathbf{u}_{j,k}^{(i)}\|_{\mathbb{V}}^{-1/\bar{s}}(\|\mathbf{u}\|_{\mathcal{A}_{\mathbb{V}}^{\bar{s}}}^{1/\bar{s}} + \|p\|_{\mathcal{A}_{\mathbb{P}}^{\bar{s}}}^{1/\bar{s}}),$$

and so, by Theorem 3.2 and (7.13),

$$\begin{aligned}\#\tau_{\underline{j}(\underline{i}),\underline{k}(\underline{i})}^{(\underline{i})} - \tau_0 &\lesssim (\|\mathbf{u}\|_{\mathcal{A}_{\mathbb{V}}^{\bar{s}}}^{1/\bar{s}} + \|p\|_{\mathcal{A}_{\mathbb{P}}^{\bar{s}}}^{1/\bar{s}}) \sum_{i=0}^{\underline{i}} \sum_{j=0}^{\underline{j}(i)} \sum_{k=0}^{\underline{k}(j)-1} \|\mathbf{u}^{p_j^{(i)}} - \mathbf{u}_{j,k}^{(i)}\|_{\mathbb{V}}^{-1/\bar{s}} \\ &\lesssim (\|\mathbf{u}\|_{\mathcal{A}_{\mathbb{V}}^{\bar{s}}}^{1/\bar{s}} + |p|_{\mathcal{A}_{\mathbb{P}}^{\bar{s}}}^{1/\bar{s}}) \sum_{i=0}^{\underline{i}} \sum_{j=0}^{\underline{j}(i)} \|\mathbf{u}^{p_j^{(i)}} - \mathbf{u}_{j,\underline{k}(j)-1}^{(i)}\|_{\mathbb{V}}^{-1/\bar{s}}.\end{aligned}$$

We will bound this expression by using the fact that each of the three nested iterations is stopped in time.

For any i and j, by the definition of $\underline{k}$, we have

$$\begin{aligned}\mathcal{E}^{\mathrm{E}}(\tau_{j,\underline{k}-1}^{(i)}, \mathbf{f}, p_j^{(i)}, \mathbf{u}_{j,\underline{k}-1}^{(i)}) &> \alpha\mathcal{E}^{\mathrm{S}}(\sigma_i, \tau_{j,\underline{k}-1}^{(i)}, \mathbf{f}, p_j^{(i)}, \mathbf{u}_{j,\underline{k}-1}^{(i)}) \\ &> \alpha\kappa\mathcal{E}^{\mathrm{S}}(\infty, \tau_{j,\underline{k}-1}^{(i)}, \mathbf{f}, p_j^{(i)}, \mathbf{u}_{j,\underline{k}-1}^{(i)}) > \alpha\kappa C_3^{-1}\varepsilon,\end{aligned}$$

or

$$\begin{aligned}\|\mathbf{u}^{p_j^{(i)}} - \mathbf{u}_{j,\underline{k}-1}^{(i)}\|_{\mathbb{V}} &\gtrsim \|\mathbf{u}^{\sigma_i} - \mathbf{u}_{j,\underline{k}-1}^{(i)}\|_{\mathbb{V}} + \|p_{\sigma_i} - p_j^{(i)}\|_{\mathbb{P}} \\ &\gtrsim \|\mathbf{u} - \mathbf{u}_{j,\underline{k}-1}^{(i)}\|_{\mathbb{V}} + \|p - p_j^{(i)}\|_{\mathbb{P}} \gtrsim \varepsilon.\end{aligned}$$

Similarly, for any i, by the definition of $\underline{j}$, we have

$$\mathcal{E}^{\mathrm{S}}(\sigma_i, \tau_{\underline{j}-1,\underline{k}}^{(i)}, \mathbf{f}, p_{\underline{j}-1}^{(i)}, \mathbf{u}_{\underline{j}-1,\underline{k}}^{(i)}) > \kappa\mathcal{E}^{\mathrm{S}}(\infty, \tau_{\underline{j}-1,\underline{k}}^{(i)}, \mathbf{f}, p_{\underline{j}-1}^{(i)}, \mathbf{u}_{\underline{j}-1,\underline{k}}^{(i)}) > \kappa C_3^{-1}\varepsilon,$$

or

$$\|p_{\sigma_i}-p^{(i)}_{\underline{j}-1}\|_{\mathbb{P}}\eqsim\|\mathbf{u}^{\sigma_i}-\mathbf{u}^{(i)}_{\underline{j}-1,\underline{k}}\|_{\mathbb{V}}+\|p_{\sigma_i}-p^{(i)}_{\underline{j}-1}\|_{\mathbb{P}}\gtrsim\|\mathbf{u}-\mathbf{u}^{(i)}_{\underline{j}-1,\underline{k}}\|_{\mathbb{V}}+\|p-p^{(i)}_{\underline{j}-1}\|_{\mathbb{P}}\gtrsim\varepsilon,$$

where for "$\eqsim$" we used (7.10). Finally, by the definition of $\underline{i}$, we have

$$\mathcal{E}^{\mathrm{S}}(\infty,\tau^{(\underline{i}-1)}_{\underline{j},\underline{\underline{k}}},\mathbf{f},p^{(\underline{i}-1)}_{\underline{j}},\mathbf{u}^{(\underline{i}-1)}_{\underline{j},\underline{\underline{k}}})>C_3^{-1}\varepsilon,$$

or

$$\|p-p^{(\underline{i}-1)}_{\underline{j}}\|_{\mathbb{P}}\eqsim\|\mathbf{u}-\mathbf{u}^{(\underline{i}-1)}_{\underline{j},\underline{\underline{k}}}\|_{\mathbb{V}}+\|p-p^{(\underline{i}-1)}_{\underline{j}}\|_{\mathbb{P}}\gtrsim\varepsilon,$$

where for "$\eqsim$" we used (7.4). By using in addition (7.12) and (7.8), we find that

$$\begin{aligned}
&\sum_{i=0}^{\underline{i}}\sum_{j=0}^{\underline{j}(i)}\|\mathbf{u}^{p^{(i)}_j}-\mathbf{u}^{(i)}_{j,\underline{k}(j)-1}\|_{\mathbb{V}}^{-1/\bar{s}}\\
&\lesssim\sum_{i=0}^{\underline{i}}\sum_{j=0}^{\underline{j}(i)-1}\|p_{\sigma_i}-p^{(i)}_j\|_{\mathbb{P}}^{-1/\bar{s}}+\sum_{i=0}^{\underline{i}-1}\|p-p^{(i)}_{\underline{j}(i)}\|_{\mathbb{P}}^{-1/\bar{s}}+\varepsilon^{-1/\bar{s}}\\
&\lesssim\sum_{i=0}^{\underline{i}}\|p_{\sigma_i}-p^{(i)}_{\underline{j}(i)-1}\|_{\mathbb{P}}^{-1/\bar{s}}+\|p-p^{(\underline{i}-1)}_{\underline{j}(\underline{i}-1)}\|_{\mathbb{P}}^{-1/\bar{s}}+\varepsilon^{-1/\bar{s}}\\
&\lesssim\sum_{i=0}^{\underline{i}-1}\|p-p^{(i)}_{\underline{j}(i)-1}\|_{\mathbb{P}}^{-1/\bar{s}}+\|p_{\sigma_{\underline{i}}}-p^{(\underline{i})}_{\underline{j}(\underline{i})-1}\|_{\mathbb{P}}^{-1/\bar{s}}+\varepsilon^{-1/\bar{s}}\\
&\lesssim\sum_{i=0}^{\underline{i}-1}\|p-p^{(i)}_{\underline{j}(i)}\|_{\mathbb{P}}^{-1/\bar{s}}+\varepsilon^{-1/\bar{s}}\lesssim\varepsilon^{-1/\bar{s}},
\end{aligned}$$

which completes the proof of the theorem. ∎

8. A practical adaptive method for the Stokes problem. In the previous section, we assumed that with respect to any current partition τ, the right-hand side $\mathbf{f}$ was in $\mathbb{V}^*_\tau$. This led to the severe limitation of $\mathbf{f}$ being in $\mathbb{V}^*_{\tau_0}$. In the practical method, given a general right-hand side $\mathbf{f}$ and any current partition τ, $\mathbf{f}$ is going to be *replaced* by a sufficiently accurate approximation from $\mathbb{V}^*_\tau$ or, when such an approximation does not exist, by an approximation from $\mathbb{V}^*_{\tau'}$ for some $\tau'\supset\tau$. We assume the availability of the following routine, **RHS**.

RHS$[\tau,\eta]\to[\tau',\mathbf{f}_{\tau'}]$
% $\eta>0$. The output consists of an $\mathbf{f}_{\tau'}\in\mathbb{V}^*_{\tau'}$, where $\tau'=\tau$, or, if necessary,
% $\tau'\supset\tau$, such that $\|\mathbf{f}-\mathbf{f}_{\tau'}\|_{\mathbb{V}'}\le\eta$.

Assuming that $p\in\mathcal{A}^s_{\mathbb{P}}$ and $\mathbf{u}\in\mathcal{A}^{\tilde{s}}_{\mathbb{V}}$ for some $s,\tilde{s}>0$, the cost of approximating the right-hand side $\mathbf{f}$ using **RHS** will generally not dominate the other cost of our practical adaptive method only then when there is some constant $c_{\mathbf{f}}$ such that, with $\bar{s}=\max(s,\tilde{s})$, for any $\eta>0$ and any partition τ, for $[\tau',\cdot]:=\mathbf{RHS}[\tau,\eta]$, it holds that

$$\#\tau'-\#\tau\le c_{\mathbf{f}}^{1/\bar{s}}\eta^{-1/\bar{s}},$$

and the number of arithmetic operations required by the call is $\mathcal{O}(\#\tau')$. We will call such an **RHS** $\bar{s}$-*optimal* with constant $c_{\mathbf{f}}$. Obviously, given $\bar{s}$, such a routine can exist only when $\mathbf{f} \in \mathcal{A}^{\bar{s}}_{\mathbb{V}'}$, defined by

$$\mathcal{A}^{\bar{s}}_{\mathbb{V}'} = \{\mathbf{g} \in \mathbb{V}' : \sup_{\varepsilon>0} \varepsilon \inf_{\{\tau:\inf_{\mathbf{g}_\tau \in \mathbb{V}^*_\tau} \|\mathbf{g}-\mathbf{g}_\tau\|_{\mathbb{V}'} \le \varepsilon\}} [\#\tau - \#\tau_0]^{\bar{s}} < \infty\}.$$

Knowing that $\mathbf{f} \in \mathcal{A}^{\bar{s}}_{\mathbb{V}'}$ is a different thing than knowing how to *construct* suitable approximations. If $\bar{s} \in [1/d, (m+1)/d]$ and $\mathbf{f} \in H^{\bar{s}d-1}(\Omega)^d$, then $\mathbf{f} \in \mathcal{A}^{\bar{s}}_{\mathbb{V}'}$, and $\mathbf{f}_{\tau'}$ constructed as the best approximation from $\mathbb{V}^*_{\tau'}$ to $\mathbf{f}$ with respect to $L_2(\Omega)^d$ using (the smallest common refinement of the input partition and) uniform refined partitions τ' are known to converge at the required rate. For general $\mathbf{f} \in \bar{\mathcal{A}}^{\bar{s}}_{\mathbb{V}'}$, however, a realization of a suitable routine **RHS** has to depend on $\mathbf{f}$ at hand.

The maximal values of s, $\tilde{s}$, and $\bar{s}$ for which membership $p \in \mathcal{A}^s_{\mathbb{P}}$, $\mathbf{u} \in \mathcal{A}^{\tilde{s}}_{\mathbb{V}}$, and $\mathbf{f} \in \mathcal{A}^{\bar{s}}_{\mathbb{V}'}$ can be expected are m/d, m/d, and $(m+1)/d$, respectively. Therefore, usually one can expect that asymptotically calls of **RHS** will not give rise to refinements.

Remark 8.1. Instead of replacing $\mathbf{f}$ by a piecewise polynomial approximation both for setting up the Galerkin system for the inner elliptic problem and for computing the a posteriori error estimators, one may work with the original $\mathbf{f}$. In that case, generally one commits quadrature errors, which one may view as a consequence of an implicit replacement of $\mathbf{f}$ by a piecewise polynomial approximation. This approach, however, is restricted to $\mathbf{f} \in L_2(\Omega)^d$, since otherwise the error estimators are not defined.

Since we are going to consider modified right-hand sides, we introduce the following notation: Given $r \in \mathbb{P}$, $\mathbf{g} \in \mathbb{V}'$, and a partition τ, $\mathbf{u}^{r,\mathbf{g}}_\tau \in \mathbb{V}_\tau$ will denote the solution of the Galerkin problem

$$\boxed{a(\mathbf{u}^{r,\mathbf{g}}_\tau, \mathbf{v}_\tau) = \mathbf{g}(\mathbf{v}_\tau) - b(\mathbf{v}_\tau, r) \quad (\mathbf{v}_\tau \in \mathbb{V}_\tau).} \tag{8.1}$$

So in view of the notation introduced in (4.5), we have $\mathbf{u}^{r,\mathbf{f}}_\tau = \mathbf{u}^r_\tau$. Note that $\|\mathbf{u}^r_\tau - \mathbf{u}^{r,\mathbf{g}}_\tau\|_{\mathbb{V}} \le \|\mathbf{f}-\mathbf{g}\|_{\mathbb{V}}$.

In the previous section, we solved the arising Galerkin systems exactly. When aiming at a method of optimal computational complexity we cannot afford this. Therefore, we assume that we have an iterative solver of optimal type available:

GALSOLVE$[\tau, \mathbf{f}_\tau, r_\tau, \mathbf{w}^{(0)}_\tau, \eta] \to \bar{\mathbf{w}}_\tau$
% τ is a partition, $\mathbf{f}_\tau \in \mathbb{V}^*_\tau$, $r_\sigma \in \mathbb{P}_\sigma$ for some $\sigma \supseteq \tau$, $\mathbf{w}^{(0)}_\tau \in \mathbb{V}_\tau$, the latter being an
% initial approximation for an iterative solver, and $\eta > 0$.
% The output $\bar{\mathbf{w}}_\tau \in \mathbb{V}_\tau$ satisfies

$$\|\mathbf{u}^{r_\sigma,\mathbf{f}_\tau}_\tau - \bar{\mathbf{w}}_\tau\|_{\mathbb{V}} \le \eta.$$

% The call requires $\lesssim \max\{1, \log(\eta^{-1}\|\mathbf{u}^{r_\sigma,\mathbf{f}_\tau}_\tau - \mathbf{w}^{(0)}_\tau\|_{\mathbb{V}})\}\#\tau$ arithmetic operations.

Additive or multiplicative multigrid methods with local smoothing are known to be of this type [WC06].

In Lemma 6.4, the adaptive refinement routine **REFvel** was analyzed. For a call $\tilde{\tau} = \mathbf{REFvel}[\tau, \mathbf{f}, r_\sigma, \mathbf{u}^{r_\sigma}_\tau]$, it was shown that $\|\mathbf{u}^{r_\sigma} - \mathbf{u}^{r_\sigma}_{\tilde{\tau}}\|_{\mathbb{V}}$ was strictly smaller than $\|\mathbf{u}^{r_\sigma} - \mathbf{u}^{r_\sigma}_\tau\|_{\mathbb{V}}$, and for $\mathbf{u}^{r_\sigma}_\tau$ not too close to $\mathbf{u}^{r_\sigma}$, a bound for the number of marked simplices inside the call was given. In the following lemma, these results are extended to the situation in which we work with approximations to $\mathbf{f}$, and where we have only inexact Galerkin solutions available.

LEMMA 8.2. *There exist constants* $\chi_1 = \chi_1(\zeta, C_1, c_2) > 0$ *and* $\lambda = \lambda(\chi_1, C_1, c_2) \in (0, \frac{1}{3}\big[1 - \frac{C_1^2\zeta^2}{c_2^2}\big]^{\frac{1}{2}}]$ *such that if for* $\mathbf{f} \in \mathbb{V}'$, *partitions* $\tau \supseteq \sigma$, $r_\sigma \in \mathbb{P}_\sigma$, $\mathbf{f}_\tau \in \mathbb{V}_\tau^*$, $\mathbf{w}_\tau \in \mathbb{V}_\tau$,

$$\|\mathbf{f} - \mathbf{f}_\tau\|_{\mathbb{V}'} + \|\mathbf{u}_\tau^{r_\sigma, \mathbf{f}_\tau} - \mathbf{w}_\tau\|_{\mathbb{V}} \leq \chi_1 \mathcal{E}(\tau, \mathbf{f}_\tau, r_\sigma, \mathbf{w}_\tau) \tag{8.2}$$

and, for some absolute constant $\vartheta > 0$,

$$\|\mathbf{u}^{r_\sigma} - \mathbf{w}_\tau\|_{\mathbb{V}} \geq \vartheta \|\mathbf{u} - \mathbf{u}^{r_\sigma}\|_{\mathbb{V}},$$

then the set of marked simplices $\underline{F}$ *inside the call* $\tilde{\tau} := \mathbf{REFvel}[\tau, \mathbf{f}_\tau, r_\sigma, \mathbf{w}_\tau]$ *satisfies*

$$\#\underline{F} \lesssim \#\bar{\tau} + \#\bar{\sigma} + \#\sigma$$

for any partitions $\bar{\tau}$ *and* $\bar{\sigma}$ *for which*

$$\inf_{\mathbf{v}_{\bar{\tau}} \in \mathbb{V}_{\bar{\tau}}} \|\mathbf{u} - \mathbf{v}_{\bar{\tau}}\|_{\mathbb{V}} \leq \lambda \|\mathbf{u}^{r_\sigma} - \mathbf{w}_\tau\|_{\mathbb{V}}, \quad \inf_{q_{\bar{\sigma}} \in \mathbb{P}_{\bar{\sigma}}} \|p - q_{\bar{\sigma}}\|_{\mathbb{V}} \leq \lambda \|\mathbf{u}^{r_\sigma} - \mathbf{w}_\tau\|_{\mathbb{V}}. \tag{8.3}$$

Furthermore, given a

$$\mu \in \big(\big[1 - \tfrac{c_2^2\zeta^2}{C_1^2}\big]^{\frac{1}{2}}, 1\big),$$

there exists an $\chi_2 = \chi_2(\mu, \zeta, C_1, c_2) > 0$, *such that if* (5.2) *is valid with* χ_1 *reading as* χ_2, *and for* $\tau' \supseteq \tilde{\tau}$, $\mathbf{f}_{\tau'} \in \mathbb{V}'$, *and* $\mathbf{w}_{\tau'} \in \mathbb{V}_{\tau'}$,

$$\|\mathbf{f} - \mathbf{f}_{\tau'}\|_{\mathbb{V}'} + \|\mathbf{u}_{\tau'}^{r_\sigma, \mathbf{f}_{\tau'}} - \mathbf{w}_{\tau'}\|_{\mathbb{V}} \leq \chi_2 \mathcal{E}^{\mathrm{E}}(\tau, \mathbf{f}_\tau, r_\sigma, \mathbf{w}_\tau),$$

then

$$\|\mathbf{u}^{r_\sigma} - \mathbf{w}_{\tau'}\|_{\mathbb{V}} \leq \mu \|\mathbf{u}^{r_\sigma} - \mathbf{w}_\tau\|_{\mathbb{V}}.$$

Proof. Following along the lines of [Ste07, proof of Lemma 6.1], for suitable constants χ_1 and λ, one can show that $\#\underline{F} \lesssim \#\hat{\tau} - \#\tau_0$ for any partition $\hat{\tau}$ with $\inf_{\mathbf{v}_{\hat{\tau}} \in \mathbb{V}_{\hat{\tau}}} \|\mathbf{u}^{r_\sigma} - \mathbf{v}_{\hat{\tau}}\|_{\mathbb{V}} \leq \lambda \|\mathbf{u}^{r_\sigma} - \mathbf{w}_\tau\|_{\mathbb{V}}$. Then, following the proof of Lemma 6.4, we infer that $\#\hat{\tau} - \#\tau_0 \leq \#\bar{\tau} + \#\bar{\sigma} + \#\sigma$, with $\bar{\tau}$ and $\#\bar{\sigma}$ from (8.3).

The second statement can be proven as in [Ste07, Lemma 6.2]. ∎

Now we are ready to formulate our practical adaptive Stokes solver, **STOKESSOLVE**.

STOKESSOLVE$[f, \varepsilon] \to [\sigma_j^{(i)}, p_j^{(i)}, \tau_{j,k}^{(i)}, \mathbf{w}_{j,k}^{(i)}]$

% Let the parameter ζ *from* **REFvel** *satisfy* $\zeta \in (0, \frac{c_2}{C_1})$, *and let* θ *from* **REFpres**

% satisfy $\theta \in (0, [1 - \frac{1 - c_6^2/C_5^2}{d}]^{\frac{1}{2}})$. *Let* $\kappa, \beta, \alpha, \chi > 0$ *be sufficiently small constants.*

$\sigma_0 := \tau_{0,0}^{(0)} := \tau_0$, $p_0^{(0)} := 0$, $\mathbf{w}_{0,0}^{(0)} := 0$, $\delta \eqsim \|\mathbf{f}\|_{\mathbb{V}'}$, $i := j := k := 0$

`do` $[\mathbf{f}_{j,k}^{(i)}, \tau_{j,k}^{(i)}] := \mathbf{RHS}[\delta/2, \tau_{j,k}^{(i)}]$

  $\mathbf{w}_{j,k}^{(i)} := \mathbf{GALSOLVE}[\tau_{j,k}^{(i)}, \mathbf{f}_{j,k}^{(i)}, p_j^{(i)}, \mathbf{w}_{j,k}^{(i)}, \delta/2]$

  `if` $C_3\mathcal{E}^{\mathrm{S}}(\infty, \tau_{j,k}^{(i)}, \mathbf{f}_{j,k}^{(i)}, p_j^{(i)}, \mathbf{w}_{j,k}^{(i)}) + (1 + C_3(2c_2^{-1} + 1))\delta/2 \leq \varepsilon$ `then stop`

  `elseif` $\mathcal{E}^{\mathrm{S}}(\sigma_i, \tau_{j,k}^{(i)}, \mathbf{f}_{j,k}^{(i)}, p_j^{(i)}, \mathbf{w}_{j,k}^{(i)}) + \delta \leq \kappa\mathcal{E}^{\mathrm{S}}(\infty, \tau_{j,k}^{(i)}, \mathbf{f}_{j,k}^{(i)}, p_j^{(i)}, \mathbf{w}_{j,k}^{(i)})$ `then`

    $\sigma_{i+1} := \mathbf{REFpres}[\sigma_i, \mathbf{w}_{j,k}^{(i)}]$

    $p_0^{(i+1)} := p_j^{(i)}$, $\tau_{0,0}^{(i+1)} := \tau_{j,k}^{(i)}$, $\mathbf{w}_{0,0}^{(i+1)} := \mathbf{w}_{j,k}^{(i)}$

$\delta := \beta(\mathcal{E}^{\mathrm{S}}(\infty, \tau_{j,k}^{(i)}, \mathbf{f}_{j,k}^{(i)}, p_j^{(i)}, \mathbf{w}_{j,k}^{(i)}) + \delta)$
$i{+}{+},\ j := k := 0$
elseif $\mathcal{E}^{\mathrm{E}}(\tau_{j,k}^{(i)}, \mathbf{f}_{j,k}^{(i)}, p_j^{(i)}, \mathbf{w}_{j,k}^{(i)}) + \delta \leq \alpha \mathcal{E}^{\mathrm{S}}(\sigma_i, \tau_{j,k}^{(i)}, \mathbf{f}_{j,k}^{(i)}, p_j^{(i)}, \mathbf{w}_{j,k}^{(i)})$ **then**
$p_{j+1}^{(i)} := p_j^{(i)} - Q_{\sigma_i} \mathrm{div} \mathbf{w}_{j,k}^{(i)}$
$\tau_{j+1,0}^{(i)} := \tau_{j,k}^{(i)},\ \mathbf{w}_{j+1,0}^{(i)} := \mathbf{w}_{j,k}^{(i)}$
$\delta := \beta(\mathcal{E}^{\mathrm{S}}(\sigma_i, \tau_{j,k}^{(i)}, \mathbf{f}_{j,k}^{(i)}, p_j^{(i)}, \mathbf{w}_{j,k}^{(i)}) + \delta)$
$j{+}{+},\ k := 0$
elseif $\delta \leq \chi \mathcal{E}^{\mathrm{E}}(\tau_{j,k}^{(i)}, \mathbf{f}_{j,k}^{(i)}, , p_j^{(i)}, \mathbf{w}_{j,k}^{(i)})$
$\tau_{j,k+1}^{(i)} := \mathbf{REFvel}[\tau_{j,k}^{(i)}, \mathbf{f}_{j,k}^{(i)}, p_j^{(i)}, \mathbf{w}_{j,k}^{(i)}],\ \mathbf{w}_{j,k+1}^{(i)} := \mathbf{w}_{j,k}^{(i)}$
$\delta := \beta(\mathcal{E}^{\mathrm{E}}(\tau_{j,k}^{(i)}, \mathbf{f}_{j,k}^{(i)}, p_j^{(i)}, \mathbf{w}_{j,k}^{(i)}) + \delta)$
$k{+}{+}$
else $\delta := \delta/2$
endif
enddo

Compared to the preliminary version $\mathbf{STOKESSOLVE}_0$, in **STOKESSOLVE** the exact solution $\mathbf{u}_{j,k}^{(i)}$ of the inner elliptic problem is replaced by an approximate one, $\mathbf{w}_{j,k}^{(i)}$, which is moreover computed using an approximation $\mathbf{f}_{j,k}^{(i)} \in \mathbb{V}_{\tau_{j,k}^{(i)}}^{*}$ of the right-hand side $\mathbf{f} \in \mathbb{V}'$. As a consequence, the results on the a posteriori error estimators from section 5 cannot be applied directly. Using the stability of the problems defining $\mathbf{u}$ and p with respect to perturbations in $\mathbf{f}$, and the stability of the error estimator $\mathcal{E}^{\mathrm{E}}$ demonstrated in Proposition 5.4, for sufficiently small α and κ, stopping by either $\mathcal{E}^{\mathrm{E}}(\cdots)+\delta \leq \alpha \mathcal{E}^{\mathrm{S}}(\sigma_i, \dots)$, $\mathcal{E}^{\mathrm{S}}(\sigma_i, \dots)+\delta \leq \kappa \mathcal{E}^{\mathrm{S}}(\infty, \dots)$ or $\mathcal{E}^{\mathrm{S}}(\infty, \dots)+(1+C_3(2c_2^{-1}+1))\delta/2 \leq \varepsilon$ means that

$$\|\mathbf{u}^{p_j^{(i)}} - \mathbf{w}_{j,k}^{(i)}\|_{\mathbb{V}} \leq D_1(\alpha)[\|p_{\sigma_i} - p_j^{(i)}\|_{\mathbb{P}} + \|\mathbf{u}^{p_{\sigma_i}} - \mathbf{w}_{j,k}^{(i)}\|_{\mathbb{V}}],$$
$$\|p^{\sigma_i} - p_j^{(i)}\|_{\mathbb{P}} + \|\mathbf{u}^{p_{\sigma_i}} - \mathbf{w}_{j,k}^{(i)}\|_{\mathbb{V}} \leq D_2(\kappa)[\|p - p_j^{(i)}\|_{\mathbb{P}} + \|\mathbf{u} - \mathbf{w}_{j,k}^{(i)}\|_{\mathbb{V}}], \quad \text{or}$$
$$\|p - p_j^{(i)}\|_{\mathbb{P}} + \|\mathbf{u} - \mathbf{w}_{j,k}^{(i)}\|_{\mathbb{V}} \leq \varepsilon,$$

respectively, with $D_1(\alpha), D_2(\kappa) > 0$ being some constants that tend to zero when α or κ tend to zero.

Since the tolerance $\delta/2$ for the approximations for the right-hand side and for the approximate solution of the inner Galerkin problem decreases until $\delta \leq \chi \mathcal{E}^{\mathrm{E}}(\cdots)$, where, for β small enough, the next δ respects the same bound, the second part of Lemma 8.2 shows that for each i and j, $\mathbf{w}_{j,k}^{(i)}$ converges linearly towards $\mathbf{u}^{p_j^{(i)}}$. Here, we silently assumed that by halving δ, at some point $\delta \leq \chi \mathcal{E}^{\mathrm{E}}(\cdots)$ is valid. This, however, is not necessarily true since $\mathcal{E}^{\mathrm{E}}(\cdots)$ changes as well. For example, think of the (unlikely) situation in which we have reached a partition on which $\mathbf{u}^{p_j^{(i)}}$ can be represented exactly. Yet, in that case we have also linear convergence of $\mathbf{w}_{j,k}^{(i)}$ towards $\mathbf{u}^{p_j^{(i)}}$. Indeed, if during the process of halving δ it remains larger than $\chi \mathcal{E}^{\mathrm{E}}(\cdots)$, then $\chi \mathcal{E}^{\mathrm{E}}(\cdots) + \delta$, being up to some constant factor an upper bound for $\|\mathbf{u}^{p_j^{(i)}} - \mathbf{w}_{j,k}^{(i)}\|_{\mathbb{V}}$, decreases linearly.

Based on these observations, similarly as in the proof of part (I) of Theorem 8.3, for sufficiently small κ and α, one shows convergence of **STOKESSOLVE**, with

$\|p-p_j^{(i)}\|_{\mathbb{P}}+\|\mathbf{u}-\mathbf{w}_{j,k}^{(i)}\|_{\mathbb{V}}\le\varepsilon$ at termination.

When **REFvel** is called, from $\delta\le\chi\mathcal{E}^{\mathrm{E}}(\cdots)$ for sufficiently small χ, $\mathcal{E}^{\mathrm{E}}(\cdots)+\delta>\alpha\mathcal{E}^{\mathrm{S}}(\sigma_i,\cdots)$ and $\mathcal{E}^{\mathrm{S}}(\sigma_i,\cdots)+\delta>\kappa\mathcal{E}^{\mathrm{S}}(\infty,\cdots)$, we have

$$\|\mathbf{u}^{p_j^{(i)}}-\mathbf{w}_{j,k}^{(i)}\|_{\mathbb{V}}\gtrsim\mathcal{E}^{\mathrm{E}}(\cdots)-\delta\gtrsim\mathcal{E}^{\mathrm{E}}(\cdots)+\delta\gtrsim\mathcal{E}^{\mathrm{S}}(\sigma_i,\dots)+\delta\gtrsim\mathcal{E}^{\mathrm{S}}(\infty,\dots)+\delta\gtrsim\|\mathbf{u}-\mathbf{w}_{j,k}^{(i)}\|_{\mathbb{V}},$$

meaning that we may apply the first part of Lemma 8.2 to bound the cardinality of the set $\underline{F}_{i,j,k}$ of marked simplices. Assuming that $p\in\mathcal{A}_{\mathbb{P}}^{\bar{s}}$ and $\mathbf{u}\in\mathcal{A}_{\mathbb{V}}^{\bar{s}}$, as in (7.21) we find

$$\#\underline{F}_{i,j,k}\lesssim\|\mathbf{u}^{p_j^{(i)}}-\mathbf{w}_{j,k}^{(i)}\|_{\mathbb{V}}^{-1/\bar{s}}(|\mathbf{u}|_{\mathcal{A}_{\mathbb{V}}^{\bar{s}}}^{1/\bar{s}}+|p|_{\mathcal{A}_{\mathbb{P}}^{\bar{s}}}^{1/\bar{s}})+\#\sigma_i-\#\tau_0+\#\tau_0.$$

Other than in **STOKESSOLVE**$_0$, in **STOKESSOLVE** refinements can also be made by calls $[\mathbf{f}_{j,k}^{(i)},\tau_{j,k}^{(i)}]:=\mathbf{RHS}[\delta/2,\tilde{\tau}_{j,k}^{(i)}]$. Assuming that **RHS** $\bar{s}$-optimal with constant $c_{\mathbf{f}}$, then $\#\tau_{j,k}^{(i)}-\#\tilde{\tau}_{j,k}^{(i)}\le c_{\mathbf{f}}^{-1/\bar{s}}(\delta/2)^{-1/\bar{s}}$. Using the fact that such a call can be made only when

$$\delta\gtrsim\mathcal{E}^{\mathrm{E}}(\cdots)+\delta\gtrsim\mathcal{E}^{\mathrm{S}}(\sigma_i,\dots)+\delta\gtrsim\mathcal{E}^{\mathrm{S}}(\infty,\dots)+\delta\gtrsim\varepsilon,$$

by applying similar techniques to those in the proof of Theorem 7.1 one can prove that **STOKESSOLVE** outputs quasi-optimal partitions.

Finally, one can prove that **STOKESSOLVE** is of quasi-optimal computational complexity. The main point is that with a call $\mathbf{GALSOLVE}[\tau_{j,k}^{(i)},p_j^{(i)},\mathbf{f}_{j,k}^{(i)},\mathbf{w}_{j,k}^{(i)},\delta/2]$, it holds that $\|\mathbf{u}^{p_j^{(i)},\mathbf{f}_{j,k}^{(i)}}-\mathbf{w}_{j,k}^{(i)}\|_{\mathbb{V}}\lesssim\delta/2$, so that the error has to be reduced by a constant factor only.

Along the lines indicated above, we end up with the following theorem. We have chosen not to include a full proof, since this would require another level of technicalities on top of those from the proof of Theorem 7.1.

THEOREM 8.3. (I) $[\sigma_j^{(i)},p_j^{(i)},\tau_{j,k}^{(i)},\mathbf{w}_{j,k}^{(i)}]:=\mathbf{STOKESSOLVE}[f,\varepsilon]$ *terminates, and* $\|\mathbf{u}-\mathbf{w}_{j,k}^{(i)}\|_{\mathbb{V}}+\|p-p_j^{(i)}\|_{\mathbb{P}}\le\varepsilon$. (II) *If, for some* $s>0$, $p\in\mathcal{A}_{\mathbb{P}}^{s}$, *then* $\#\sigma_j^{(i)}-\#\tau_0\lesssim\varepsilon^{-1/s}|p|_{\mathcal{A}_{\mathbb{P}}^{s}}^{1/s}$, *dependent only on* τ_0, *and on* s *when it tends to* 0 *or infinity. If, in addition, for some* $\tilde{s}>0$, $\mathbf{u}\in\mathcal{A}_{\mathbb{V}}^{\tilde{s}}$ *and, with* $\bar{s}=\min(s,\tilde{s})$, **RHS** *is* $\bar{s}$*-optimal with constant* $c_{\mathbf{f}}$, *then* $\#\tau_{j,k}^{(i)}-\#\tau_0\lesssim\varepsilon^{-1/\bar{s}}(\|p\|_{\mathcal{A}_{\mathbb{P}}^{\bar{s}}}^{1/\bar{s}}+\|\mathbf{u}\|_{\mathcal{A}_{\mathbb{V}}^{\bar{s}}}^{1/\bar{s}}+c_{vcf}^{1/\bar{s}})$, *dependent only on* τ_0, *and on* $\bar{s}$ *when it tends to* 0 *or infinity. Moreover, when* $\varepsilon\lesssim\|\mathbf{f}\|_{\mathbb{V}'}$, *the number of arithmetic operations and storage locations required by the call is also bounded by a multiple of* $\varepsilon^{-1/\bar{s}}(\|p\|_{\mathcal{A}_{\mathbb{P}}^{\bar{s}}}^{1/\bar{s}}+\|\mathbf{u}\|_{\mathcal{A}_{\mathbb{V}}^{\bar{s}}}^{1/\bar{s}}+c_{vcf}^{1/\bar{s}})$.

9. Numerical experiment. We consider $d=2$, the L-shaped domain $\Omega=(0,1)^2\backslash(\frac{1}{2},1)^2$, $\mathbf{f}:\mathbf{x}\mapsto 25(4\mathbf{x}_2-1,1-4\mathbf{x}_1)$, and $m=1$, i.e., continuous piecewise linear approximation for the velocity, piecewise constant approximation for the pressure, and an initial partition into 24 triangles.

The "bulk chasing" parameters θ and ζ inside **REFpres** or **REFvel** were chosen to be 0.7 and $\sqrt{0.3}$, respectively. Since $\mathbf{f}$ is smooth, we followed the approach discussed in Remark 8.1 and skipped the calls of **RHS**. Furthermore, instead of solving each arising finite-dimensional linear system within tolerance $\delta/2$, for the first value of $\delta/2$ obtained by successively halving that is less than $\chi\mathcal{E}^{\mathrm{E}}(\tau_{j,k}^{(i)},\mathbf{f}_{j,k,}^{(i)},p_j^{(i)},\mathbf{w}_{j,k}^{(i)})$, we always approximately solved it by 3 multigrid iterations with local smoothing (cf. [WC06])

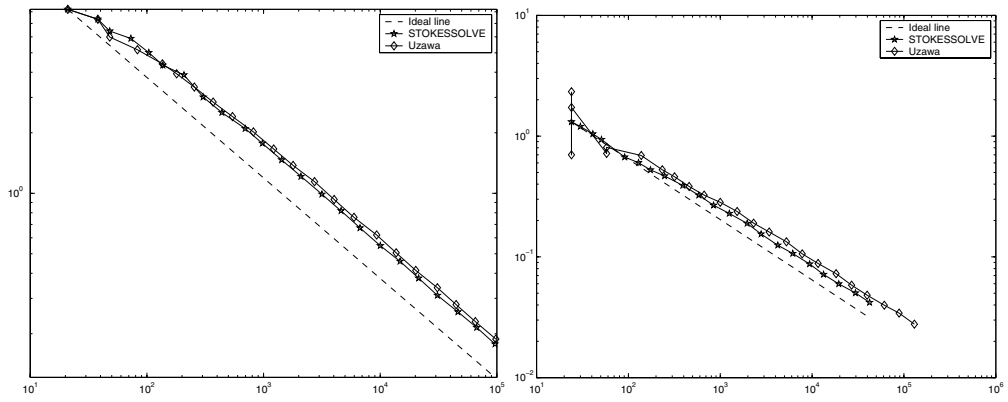


FIG. 9.1. *Estimated Stokes error* (left) *and estimated pressure error* (right) *vs. cardinality underlying partition.*

starting with the previously computed approximate velocity. The parameters κ and α are chosen to be 0.88 and 0.9, respectively.

For comparison, we also implemented the adaptive Uzawa method for solving the full Stokes problem, i.e., **STOKESSOLVE** without the outermost loop over i, i.e., without the part starting from the first `elseif` statement until the second one, and all remaining occurrences of σ_i replaced by ∞. In particular the pressure update $p_{j+1}^{(i)} := p_j^{(i)} + Q_{\sigma_i}\mathrm{div}\mathbf{w}_{j,k}^{(i)}$ then reads $p_{j+1} := p_j + \mathrm{div}\mathbf{w}_{j,k}$. This is the algorithm studied in [BMN02], apart from the replacement of a priori prescribed tolerances by a posteriori ones.

In Figure 9.1 (left), we plotted the full Stokes error estimator $\mathcal{E}^{\mathrm{S}}(\infty, \tau_{\underline{\underline{j}},\underline{\underline{k}}}^{(i)}, \mathbf{f}, p_{\underline{\underline{j}}}^{(i)}, \mathbf{w}_{\underline{\underline{j}},\underline{\underline{k}}}^{(i)})$ vs. $\#\tau_{\underline{\underline{j}},\underline{\underline{k}}}^{(i)}$ (or $\mathcal{E}^{\mathrm{S}}(\infty, \tau_{j,\underline{k}}, \mathbf{f}, p_j, \mathbf{w}_{j,\underline{k}})$ vs. $\#\tau_{j,\underline{k}}$). Ignoring the fact that generally $\mathbf{w}_{\underline{\underline{j}},\underline{\underline{k}}}^{(i)} \neq \mathbf{u}_{\tau_{\underline{\underline{j}},\underline{\underline{k}}}^{(i)}}^{p_{\underline{\underline{j}}}^{(i)}}$ (or $\mathbf{w}_{j,\underline{k}} \neq \mathbf{u}_{\tau_{j,\underline{k}}}^{p_j}$), modulo a constant factor this estimator is an upper bound for $\|\mathbf{u}-\mathbf{w}_{\underline{\underline{j}},\underline{\underline{k}}}^{(i)}\|_{\mathbb{V}} + \|p-p_{\underline{\underline{j}}}^{(i)}\|_{\mathbb{P}}$ (or $\|\mathbf{u}-\mathbf{w}_{j,\underline{k}}\|_{\mathbb{V}} + \|p-p_j\|_{\mathbb{P}}$). (For $\mathbf{f} \in \mathbb{V}^*_{\tau_{\underline{\underline{j}},\underline{\underline{k}}}^{(i)}}$ (or $\mathbf{f} \in \mathbb{V}^*_{\tau_{j,\underline{k}}}$), it would also be a lower bound.) In Figure 9.1 (right), we plotted the pressure error estimator $\|\mathrm{div}\mathbf{w}_{\underline{\underline{j}},\underline{\underline{k}}}^{(i)}\|_{L_2(\Omega)}$ vs. $\#\sigma_i$ (or $\|\mathrm{div}\mathbf{w}_{j,\underline{k}}\|_{L_2(\Omega)}$ vs. $\#\tau_{j,0}$). Ignoring the fact that generally $\mathbf{w}_{\underline{\underline{j}},\underline{\underline{k}}}^{(i)} \neq \mathbf{u}^{p_{\underline{\underline{j}}}^{(i)}}$ (or $\mathbf{w}_{j,\underline{k}} \neq \mathbf{u}^{p_j}$), this estimator is equivalent to $\|p-p_{\underline{\underline{j}}}^{(i)}\|_{\mathbb{P}}$ (or $\|p-p_j\|_{\mathbb{P}}$). As predicted by the theory, the approximations produced by **STOKESSOLVE** converge with the best possible rates. In this example, we observe that the same is true for the adaptive Uzawa method.

Our implementation is partly written in C and, for our convenience, partly in MATLAB, making use of the PDE toolbox. This toolbox was modified such that it performs newest vertex bisection. Due to the data structures used in the MATLAB part, our code is not of optimal computational complexity; the time needed for a call of **REFvel** is not proportional to the cardinality of its output partition. Subtracting the times spent for this routine, we observed computing times that are proportional to the cardinality of the final partition.

In Figure 9.2 we show two partitions σ_i that were produced by **STOKESSOLVE**. Note that these partitions are nonconforming. In Figure 9.3 we give two partitions $\tau_{j,k}^{(i)}$. Finally, in Figure 9.4 plots of p and $\mathbf{u}$ are given.

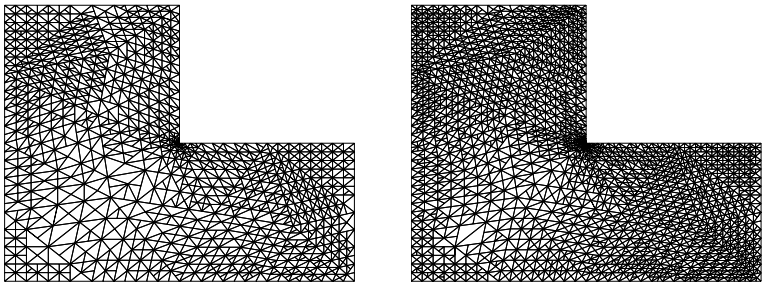


FIG. 9.2. *Pressure partitions with* 1965 *or* 4235 *triangles.*

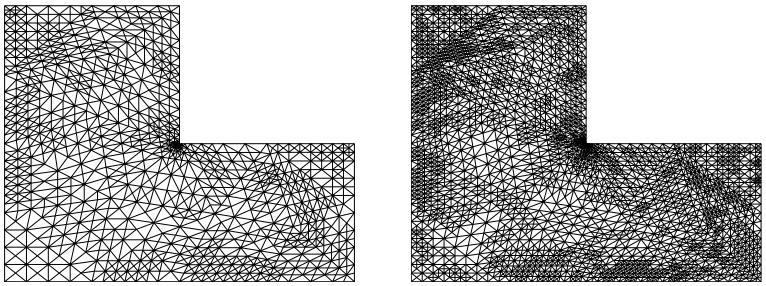


FIG. 9.3. *Velocity partitions with* 1842 *or* 6126 *triangles.*

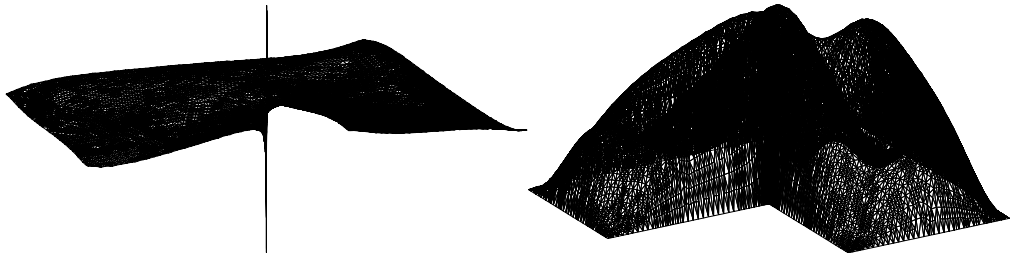


FIG. 9.4. *Pressure* p *and the norm* $|\mathbf{u}|$ *of the velocity.*

Acknowledgments. The authors are indebted to Dr. Haijun Wu (Nanjing University, China) for making available to us his code implementing multigrid on locally refined partitions. We thank Dr. Peter Binev (University of South Carolina, USA) for proving a new estimate for an adaptive tree algorithm that we needed for the **REFpres** routine.

REFERENCES

[BMN02] E. BÄNSCH, P. MORIN, AND R. H. NOCHETTO, *An adaptive Uzawa FEM for the Stokes problem: Convergence without the inf-sup condition*, SIAM J. Numer. Anal., 40 (2002), pp. 1207–1229.

[Bin07] P. BINEV, *Private communication*, 2007.

[BDD04] P. BINEV, W. DAHMEN, AND R. DEVORE, *Adaptive finite element methods with convergence rates*, Numer. Math., 97 (2004), pp. 219–268.

[BDDP02] P. BINEV, W. DAHMEN, R. DEVORE, AND P. PETRUCHEV, *Approximation classes for adaptive methods*, Serdica Math. J., 28 (2002), pp. 391–416.

[BD04] P. Binev and R. DeVore, *Fast computation in adaptive tree approximation*, Numer. Math., 97 (2004), pp. 193–217.

[Dah99] S. Dahlke, *Besov regularity for the Stokes problem*, in Advances in Multivariate Approximation, W. Haussmann, K. Jetter, and M. Reimer, eds., Math. Res. 107, Wiley-VCH, Berlin, 1999, pp. 129–138.

[DDU02] S. Dahlke, W. Dahmen, and K. Urban, *Adaptive wavelet methods for saddle point problems—Optimal convergence rates*, SIAM J. Numer. Anal., 40 (2002), pp. 1230–1262.

[Dör96] W. Dörfler, *A convergent adaptive algorithm for Poisson's equation*, SIAM J. Numer. Anal., 33 (1996), pp. 1106–1124.

[Kon06] Y. Kondratyuk, *Adaptive Finite Element Algorithms for the Stokes Problem: Convergence Rates and Optimal Computational Complexity*, Preprint 1346, Department of Mathematics, Utrecht University, The Netherlands, 2006.

[KS07] Y. Kondratyuk and R. P. Stevenson, *An Optimal Adaptive Finite Element Method for the Stokes Problem—Extended Version*, Technical report, Department of Mathematics, Utrecht University, The Netherlands, 2007; available at http://staff.science.uva.nl/~rstevens/publ.html.

[Mit89] W. F. Mitchell, *A comparison of adaptive refinement techniques for elliptic problems*, ACM Trans. Math. Software, 15 (1989), pp. 326–347.

[MNS00] P. Morin, R. H. Nochetto, and K. G. Siebert, *Data oscillation and convergence of adaptive FEM*, SIAM J. Numer. Anal., 38 (2000), pp. 466–488.

[NP04] R. Nochetto and J. H. Pyo, *Optimal relaxation parameter for the Uzawa method*, Numer. Math., 98 (2004), pp. 695–702.

[Ste06] R. P. Stevenson, *The completion of locally refined simplicial partitions created by bisection*, Math. Comp., 77 (2008), pp. 227–241.

[Ste07] R. P. Stevenson, *Optimality of a standard adaptive finite element method*, Found. Comput. Math., 7 (2007), pp. 245–269.

[Ver96] R. Verfürth, *A Review of A Posteriori Error Estimation and Adaptive Mesh-Refinement Techniques*, Wiley-Teubner, Chichester, UK, 1996.

[WC06] H. Wu and Z. Chen, *Uniform convergence of multigrid V-cycle on adaptively refined finite element meshes for second order elliptic problems*, Sci. China Ser. A, 49 (2006), pp. 1405–1429.

SIAM J. NUMER. ANAL.
Vol. 46, No. 2, pp. 776–804
© 2008 Society for Industrial and Applied Mathematics

DISCONTINUOUS GALERKIN METHODS FOR FRIEDRICHS' SYSTEMS. PART III. MULTIFIELD THEORIES WITH PARTIAL COERCIVITY*

ALEXANDRE ERN† AND JEAN-LUC GUERMOND‡

Abstract. This paper is the third and last part of a work attempting to give a unified analysis of discontinuous Galerkin methods. The purpose of this paper is to extend the framework that has been developed in Part II for two-field Friedrichs' systems associated with second-order PDEs. We now consider two-field Friedrichs' systems with partial L^2-coercivity and three-field Friedrichs' systems with an even weaker L^2-coercivity hypothesis. In particular, this work generalizes the discontinuous Galerkin methods of Part II to compressible and incompressible linear continuum mechanics. We also show how the stabilizing parameters of the method must be set when the two-field Friedrichs' system is composed of terms that may be of different magnitude, thus accounting, for instance, for advection–diffusion equations at high Péclet numbers.

Key words. Friedrichs' systems, finite elements, partial differential equations, discontinuous Galerkin method

AMS subject classifications. 65N30, 65M60, 35F15

DOI. 10.1137/060664045

1. Introduction.

The framework of Friedrichs' systems [21] is well adapted to the analysis of first-order PDEs for it accommodates many types of PDEs in a unified setting and systematically handles various types of boundary conditions. Friedrichs' systems were initially developed to deal with change of type in transonic flows, but the theory developed by Friedrichs turned out to give a "unified treatment" of PDEs that goes beyond the elliptic/hyperbolic/parabolic divide. The key ingredient of the theory is that the sum of the operator associated with the PDE and its formal adjoint is L^2-coercive (implying that the coefficients of the PDE satisfy symmetry properties).

As shown in Part I of this work [15], approximating Friedrichs' systems using discontinuous Galerkin (DG) methods is somewhat natural, and the setting of Friedrichs' systems lends itself nicely to analysis. Actually this observation is not new, and the analysis of DG approximations of Friedrichs' systems had been initiated in the 1970s by Lesaint and Raviart [25, 26] and further pursued by Johnson, Nävert, and Pitkäranta [24], but for some reason research remained focused for many years on hyperbolic equations. Meanwhile the development of DG methods for solving elliptic equations followed a different route and has been essentially based on the early work of Nitsche on boundary-penalty methods [27] and the use of interior penalties (IP) to weakly enforce continuity conditions imposed on the solution or its derivatives across the interfaces between adjoining elements; see Douglas and Dupont [12], Baker [4], Wheeler [29], Arnold [1], Bassi and Rebay [5], and Cockburn and Shu [11]. All these approaches have been recast into a single framework amenable to a unified error analysis in Arnold et al. [2] for the Poisson problem.

*Received by the editors June 30, 2006; accepted for publication (in revised form) July 12, 2007; published electronically February 8, 2008.
http://www.siam.org/journals/sinum/46-2/66404.html
†CERMICS, Ecole des Ponts, Université Paris-Est, 77455 Marne la Vallée Cedex 2, France (ern@cermics.enpc.fr).
‡Department of Mathematics, Texas A&M, College Station, TX 77843-3368 (guermond@math.tamu.edu); on leave from LIMSI (CNRS-UPR 3251), BP 133, 91403, Orsay, France.

In Part I [15] we reunited the two above diverging viewpoints by showing that the theory of Friedrichs' systems is a proper setting to analyze DG methods irrespective of the type of the PDE: The DG analysis of a transport equation and that of the Laplace equation are identical when all the components of the unknown are given equal attention. Sometimes, however, the unknown, say z, possesses a two-field structure, $z = (z^\sigma, z^u)$, and the σ-component can be eliminated to yield a system of second-order PDEs for the u-component. (We have in mind here the Maxwell equations, the mixed form of the Laplace equation, or the mixed form of the linear elasticity equations.) This situation, which gives more weight to one component of the unknown, say the u-component, has been analyzed in Part II [16]. Therein we developed a theory of two-field Friedrichs' systems and analyzed two-field DG methods for which z^σ can be locally eliminated on each mesh cell.

The goal of the third and last part of this work is to extend the analysis of Part II in three directions by weakening the assumption of L^2-coercivity on which the theory of the two-field Friedrichs' systems is based. First, the L^2-coercivity is assumed to hold only on the σ-component of the field $z = (z^\sigma, z^u)$. Examples include advection–diffusion equations and compressible linear continuum mechanics problems. Second, further weakening of the partial coercivity framework is done by introducing a three-field theory of Friedrichs' systems. This framework encompasses incompressible linear continuum mechanics, e.g., Stokes and Oseen flows. Third, the two-field DG method is revisited by performing a singular perturbation analysis. The goal of this third extension is to determine how the stabilizing parameters of the method must be set when the elliptic-like PDE associated with the two-field Friedrichs' system under scrutiny is composed of a second-order term and a first-order term that may be of different magnitude. The situation covered by this theory is that of advection–diffusion equations at high Péclet numbers. The recurrent theme of the present paper is to construct robust multifield DG methods that can efficiently handle type changes (elliptic/hyperbolic, compressible/incompressible, etc.) and are such that one subfield of the unknown, say z^σ, can be eliminated on each mesh element.

This paper is organized as follows. Section 2 sets the notation and briefly reviews the main results obtained in Parts I and II. This section can be skipped by readers who are familiar with the material. Section 3 treats two-field Friedrichs' systems for which L^2-coercivity holds only on the σ-component. The key difference with Part II is that a Poincaré-like inequality must be invoked to transfer the L^2-stability from z^σ to z^u. Section 4 deals with three-field Friedrichs' systems where the partial coercivity framework is further weakened. In both cases, the well-posedness of the Friedrichs' systems is established and the convergence of their DG approximation is analyzed under general design conditions. Finally, section 5 presents a singular perturbation analysis relevant to second-order PDEs where first- and second-order terms are not of the same magnitude. Sections 3, 4, and 5 are independent and can be read separately. The theme of section 4 is robustness with respect to the compressible/incompressible-type change and that of section 5 is robustness with respect to the elliptic/hyperbolic-type change.

2. DG approximation of Friedrichs' systems. The objective of this section is to set the notation and briefly restate the main results of Parts I and II. The reader familiar with this material can jump to section 3.

2.1. One-field Friedrichs' systems. Let Ω be a bounded, open, and connected Lipschitz domain in $\mathbb{R}^d$. Let m be a positive integer and set $L = [L^2(\Omega)]^m$ with inner product $(\cdot,\cdot)_L$. The two ingredients to build a Friedrichs' system are an operator

$K \in \mathcal{L}(L;L)$ and a family $\{\mathcal{A}^k\}_{1\le k\le d}$ of d functions on Ω with values in $\mathbb{R}^{m,m}$ such that

(A1) $\quad \forall k \in \{1,\dots,d\},\ \mathcal{A}^k \in [L^\infty(\Omega)]^{m,m} \quad \text{and} \quad \sum_{k=1}^d \partial_k \mathcal{A}^k \in [L^\infty(\Omega)]^{m,m},$

(A2) $\quad \forall k \in \{1,\dots,d\},\ \mathcal{A}^k = (\mathcal{A}^k)^t$ a.e. in Ω,

(A3) $\quad \exists \mu_0 > 0, \quad \forall z \in L, \quad ((K + K^* - \nabla\cdot A)z, z)_L \ge 2\mu_0 \|z\|_L^2,$

where K^* is the adjoint of K in $\mathcal{L}(L;L)$ and $\nabla\cdot A \in \mathcal{L}(L;L)$ is defined such that $\nabla\cdot A(z) = (\sum_{k=1}^d \partial_k \mathcal{A}^k)z$ for all $z \in L$.

Let $\mathfrak{D}(\Omega)$ denote the space of $\mathfrak{C}^\infty$ functions that are compactly supported in Ω. A function z in L is said to have an A-weak derivative in L if the linear form $[\mathfrak{D}(\Omega)]^m \ni \phi \longmapsto -\int_\Omega \sum_{k=1}^d z^t \partial_k(\mathcal{A}^k\phi) \in \mathbb{R}$ is bounded on L. In this case, the function in L that can be associated with the above linear form by means of the Riesz representation theorem is denoted by Az. The so-called graph space $W = \{z \in L;\ Az \in L\}$ is endowed with a Hilbert structure when equipped with the scalar product $(z,y)_L + (Az, Ay)_L$. Define the operators $A \in \mathcal{L}(W;L)$ and $\tilde{A} \in \mathcal{L}(W;L)$ by

$$Az = \sum_{k=1}^d \mathcal{A}^k \partial_k z, \qquad \tilde{A}z = -\sum_{k=1}^d \partial_k(\mathcal{A}^k z), \tag{2.1}$$

and set $T = K + A$, $\tilde{T} = K^* + \tilde{A}$. $\tilde{A}$ and $\tilde{T}$ are the formal adjoints of A and T, respectively. Assumption (A3), which implies that $T + \tilde{T}$ is L-coercive on L, is the full L^2-coercivity property alluded to in section 1.

Let $f \in L$ and consider the PDE system $Tz = f$. An important question we are facing now is to equip this problem with proper boundary conditions. The key idea underlying the theory of Friedrichs' systems is that boundary conditions can be enforced by making use of a boundary operator $M \in \mathcal{L}(W;W')$ such that

(M1) $\quad \forall z \in W,\ \langle Mz, z\rangle_{W',W} \ge 0,$

(M2) $\quad W = \mathrm{Ker}(D - M) + \mathrm{Ker}(D + M),$

where $D \in \mathcal{L}(W;W')$ is defined by

$$\forall (z,y) \in W \times W, \quad \langle Dz, y\rangle_{W',W} = (Az, y)_L - (z, \tilde{A}y)_L. \tag{2.2}$$

Observe that (2.2) is just an integration-by-parts formula and that D is self-adjoint by construction. It is shown in [17] that by setting $V = \mathrm{Ker}(D - M)$ and $V^* = \mathrm{Ker}(D + M^*)$, where M^* is the adjoint operator of M, the following problems are well-posed:

$$\text{Seek } z \in V \text{ such that } Tz = f. \qquad \text{Seek } z^* \in V^* \text{ such that } \tilde{T}z^* = f. \tag{2.3}$$

The key idea sustaining the entire DG theory developed in Parts I, II, and hereafter is that it is possible to enforce boundary conditions weakly by introducing the following bilinear forms on $W \times W$,

$$a(z,y) = (Tz, y)_L + \tfrac{1}{2}\langle (M - D)z, y\rangle_{W',W}, \tag{2.4}$$

$$a^*(z,y) = (\tilde{T}z, y)_L + \tfrac{1}{2}\langle (M^* + D)z, y\rangle_{W',W}, \tag{2.5}$$

and by reformulating (2.3) as follows:

$$\text{Seek } z \in W \text{ such that } a(z,y) = (f,y)_L \quad \forall y \in W. \tag{2.6}$$

$$\text{Seek } z^* \in W \text{ such that } a^*(z^*,y) = (f,y)_L \quad \forall y \in W. \tag{2.7}$$

The key well-posedness result established in Part I is the following.

THEOREM 2.1. *Assume* (A1)–(A3) *and* (M1)–(M2). *Then there are unique solutions to* (2.6) *and* (2.7) *and these solutions solve* (2.3).

We finish this section by giving local representations of the operators D and M. Let $n = (n_1, \ldots, n_d)^t$ be the unit outward normal to $\partial\Omega$. Whenever the fields $\{\mathcal{A}^k\}_{1\le k\le d}$ are sufficiently smooth for the field $\mathcal{D} = \sum_{k=1}^d n_k \mathcal{A}^k : \partial\Omega \longrightarrow \mathbb{R}^{m,m}$ to be meaningful at the boundary, the following representation of D holds:

$$\langle Dz, y\rangle_{W',W} = \int_{\partial\Omega} y^t \mathcal{D} z \tag{2.8}$$

for every smooth function z and y. Likewise, we henceforth assume that there is a field $\mathcal{M} : \partial\Omega \longrightarrow \mathbb{R}^{m,m}$ such that following representation of M holds for every smooth function z and y:

$$\langle Mz, y\rangle_{W',W} = \int_{\partial\Omega} y^t \mathcal{M} z. \tag{2.9}$$

2.2. Two-field Friedrichs' systems. We now briefly recall the two-field theory developed in Part II. Elliptic-like PDEs in mixed form lead to Friedrichs' systems with the following 2×2 structure: There are two positive integers m_σ and m_u such that $m = m_\sigma + m_u$ and $L = L_\sigma \times L_u$, where $L_\sigma = [L^2(\Omega)]^{m_\sigma}$ and $L_u = [L^2(\Omega)]^{m_u}$, yielding the decomposition $v = (v^\sigma, v^u)$ for all $v \in L$. With obvious notation this leads to the following block decompositions

$$K = \left[\begin{array}{c:c} K^{\sigma\sigma} & K^{\sigma u} \\ \hdashline K^{u\sigma} & K^{uu} \end{array}\right], \qquad \mathcal{A}^k = \left[\begin{array}{c:c} \mathcal{A}^{\sigma\sigma,k} & \mathcal{B}^k \\ \hdashline (\mathcal{B}^k)^t & \mathcal{C}^k \end{array}\right], \tag{2.10}$$

where for all $k \in \{1, \ldots, d\}$, $\mathcal{B}^k$ is an $m_\sigma \times m_u$ matrix field and $\mathcal{C}^k$ an $m_u \times m_u$ matrix field. Assume now that the block $K^{\sigma\sigma}$ has a local representation, i.e., there is $\mathcal{K}^{\sigma\sigma} \in [L^\infty(\Omega)]^{m_\sigma, m_\sigma}$ such that $K^{\sigma\sigma} y^\sigma = \mathcal{K}^{\sigma\sigma} y^\sigma$ for all $y^\sigma \in L_\sigma$ (this localization hypothesis is needed to locally eliminate the σ-component in the two-field DG method described below). The two key hypotheses on which the two-field theory is based are

$$\forall k \in \{1, \ldots, d\}, \quad \mathcal{A}^{\sigma\sigma,k} = 0, \tag{A4}$$

$$\exists k_0 > 0, \quad \mathcal{K}^{\sigma\sigma} \ge k_0 \mathcal{I}_{m_\sigma}, \tag{A5}$$

where $\mathcal{I}_{m_\sigma}$ is the identity matrix in $\mathbb{R}^{m_\sigma, m_\sigma}$. Assumptions (A4)–(A5) allow us to eliminate the σ-component of z in the PDE system $Tz = f$ leading to an elliptic-like PDE for the u-component. Furthermore, assumption (A4) yields

$$\mathcal{D} = \left[\begin{array}{c:c} 0 & \mathcal{D}^{\sigma u} \\ \hdashline \mathcal{D}^{u\sigma} & \mathcal{D}^{uu} \end{array}\right], \tag{2.11}$$

with $\mathcal{D}^{\sigma u} = \sum_{k=1}^d n_k \mathcal{B}^k$, $\mathcal{D}^{u\sigma} = (\mathcal{D}^{\sigma u})^t$, and $\mathcal{D}^{uu} = \sum_{k=1}^d n_k \mathcal{C}^k$.

Henceforth, boundary conditions are enforced by taking

$$\mathcal{M} = \left[\begin{array}{c|c} 0 & -\alpha\mathcal{D}^{\sigma u} \\ \hline \alpha\mathcal{D}^{u\sigma} & \mathcal{M}^{uu} \end{array}\right], \tag{2.12}$$

where $\mathcal{M}^{uu} \in \mathbb{R}^{m_u,m_u}$ is positive and $\alpha \in \{-1,+1\}$. The choice $\alpha = +1$ leads to the Dirichlet boundary condition $z^u \in \mathrm{Ker}(\mathcal{D}^{\sigma u}) \cap \mathrm{Ker}(\mathcal{M}^{uu} - \mathcal{D}^{uu})$. The choice $\alpha = -1$ yields the Robin-type boundary condition $2\mathcal{D}^{u\sigma}z^\sigma + (\mathcal{D}^{uu} - \mathcal{M}^{uu})z^u = 0$; the boundary condition is of Neumann-type if $\mathcal{M}^{uu} = \mathcal{D}^{uu}$, provided $\mathcal{D}^{uu}$ is positive. In practice (see the examples in sections 3.3 and 3.4), $\mathrm{Ker}(\mathcal{D}^{\sigma u}) = \{0\}$, so that the Dirichlet boundary condition amounts to $z^u = 0$, while the Robin-type boundary condition is enforced by taking $\mathcal{M}^{uu} = |\mathcal{D}^{uu}|$.

It will prove convenient in what follows to define the operators $B = \sum_{k=1}^d \mathcal{B}^k\partial_k$, $B^\dagger = \sum_{k=1}^d [\mathcal{B}^k]^t\partial_k$, and $C = \sum_{k=1}^d \mathcal{C}^k\partial_k$.

2.3. The discrete setting. Let $\{\mathcal{T}_h\}_{h>0}$ be a family of meshes of Ω. To simplify, we assume that the meshes are affine and that Ω is a polyhedron. For all $K \in \mathcal{T}_h$, $n_K = (n_{K,1},\ldots,n_{K,d})^t$ denotes the unit outward normal to K and h_K is the diameter of K. We set $h = \max_{K\in\mathcal{T}_h} h_K$ and we denote by $\mathfrak{h}$ the piecewise constant function such that for all $K \in \mathcal{T}_h$, $\mathfrak{h}|_K = h_K$. Henceforth, the notation $\xi \lesssim \zeta$ means that there is a positive c, *independent of* h, such that $\xi \le c\zeta$.

We denote by $\mathcal{F}_h^{\mathrm{i}}$ the set of mesh interfaces, i.e., $F \in \mathcal{F}_h^{\mathrm{i}}$ if F is a $(d-1)$-manifold and there are $K_1(F)$ and $K_2(F) \in \mathcal{T}_h$ such that $F = K_1(F) \cap K_2(F)$. For $F \in \mathcal{F}_h^{\mathrm{i}}$, we set $\mathcal{T}(F) = K_1(F) \cup K_2(F)$. We denote by $\mathcal{F}_h^\partial$ the set of the faces that separate the mesh from the exterior of Ω, i.e., $F \in \mathcal{F}_h^\partial$ if F is a $(d-1)$-manifold and there is $K(F) \in \mathcal{T}_h$ such that $F = K(F) \cap \partial\Omega$. For $F \in \mathcal{F}_h^\partial$, we set $\mathcal{T}(F) = K(F)$. For all $F \in \mathcal{F}_h^{\mathrm{i}}$, n_F is the unit normal vector on F pointing from $K_1(F)$ to $K_2(F)$, and for all $F \in \mathcal{F}_h^\partial$, n_F is the unit normal vector on F pointing outside Ω. Finally, we set $\mathcal{F}_h = \mathcal{F}_h^{\mathrm{i}} \cup \mathcal{F}_h^\partial$ and for all $F \in \mathcal{F}_h$, h_F denotes the diameter of F. The sole assumption we make on the matching of element faces is that for all $F \in \mathcal{F}_h$, $\max_{K\in\mathcal{T}(F)} h_K \lesssim h_F$. This assumption implies, in particular, that the mesh family $\{\mathcal{T}_h\}_{h>0}$ is shape-regular.

For any measurable subset E of Ω, $(\cdot,\cdot)_{L,E}$ denotes the usual scalar product in $[L^2(E)]^m$. For ease of notation, we define the operators B_h, $B_h^\dagger$, and C_h as the elementwise versions of B, $B^\dagger$, and C, respectively; for instance, for v smooth enough, $(B_hv)|_K = \sum_{k=1}^d \mathcal{B}^k\partial_k(v|_K)$ for all $K \in \mathcal{T}_h$.

Let p be a *nonnegative* integer and consider the DG finite element space

$$W_h = [P_{h,p}]^m, \qquad P_{h,p} = \{v_h \in L^2(\Omega);\, \forall K \in \mathcal{T}_h,\, v_h|_K \in \mathbb{P}_p\}, \tag{2.13}$$

$\mathbb{P}_p$ denoting the vector space of polynomials with real coefficients and with total degree less than or equal to p. Observe that the usual inverse and trace inverse inequalities hold in W_h. For a function v that admits a possibly two-valued trace on mesh interfaces, we define the jump and mean-value of v on $F \in \mathcal{F}_h^{\mathrm{i}}$ as

$$[\![v]\!] = v^1 - v^2, \qquad \{v\} = \tfrac{1}{2}(v^1 + v^2), \tag{2.14}$$

where $v^\gamma(x) = \lim_{y\to x} v(y)|_{K_\gamma(F)}$, $\gamma \in \{1,2\}$. The field $\mathcal{D}$ is extended to $\mathcal{F}_h$ by setting for all $K \in \mathcal{T}_h$, $\mathcal{D} = \sum_{k=1}^d n_{K,k}\mathcal{A}^k$ a.e. on ∂K. Observe that $\mathcal{D}$ is two-valued on $\mathcal{F}_h^{\mathrm{i}}$ with $\{\mathcal{D}\} = 0$ on $\mathcal{F}_h^{\mathrm{i}}$ and that $|\mathcal{D}|$ is well-defined and single-valued since $\mathcal{D}$ is symmetric. We also define $\mathcal{D}_F = \sum_{k=1}^d n_{F,k}\mathcal{A}^k$.

To write a DG method starting from (2.6), we introduce three families of boundary and interface operators: $\{M_F\}_{\mathcal{F}_h^\partial}$, $\{S_F\}_{\mathcal{F}_h^{\mathrm{i}}}$, and $\{R_F\}_{\mathcal{F}_h}$. For all $F \in \mathcal{F}_h^\partial$, the role

of the operator $M_F \in \mathcal{L}([L^2(F)]^m;[L^2(F)]^m)$ is to weakly enforce the boundary conditions on F. For all $F \in \mathcal{F}_h^{\mathrm{i}}$, the role of the operator $S_F \in \mathcal{L}([L^2(F)]^m;[L^2(F)]^m)$ is to penalize the jump of the discrete unknowns across F. For all $F \in \mathcal{F}_h$, the operator $R_F \in \mathcal{L}([L^2(\mathcal{F}_h)]^m;[L^2(F)]^m)$ is user-defined so as to facilitate the implementation of the method. The default option is to take $R_F \equiv 0$ in general for the one-field approach, but a nonzero choice must be made if, when using the multifield approach, one insists on obtaining an IP-like method; cf. Part II. The reader can take $R_F \equiv 0$ in a first reading. The design of the above operators depends on whether the one-field, two-field, or three-field approach is used.

Let $W(h) = W_h + [H^1(\Omega)]^m$ and define on $W(h) \times W(h)$ the DG bilinear form

(2.15)
$$a_h(z,y) = \sum_{K\in\mathcal{T}_h} (Tz,y)_{L,K} + \sum_{F\in\mathcal{F}_h^\partial} \tfrac{1}{2}(M_F(z) - \mathcal{D}z, y)_{L,F} - \sum_{F\in\mathcal{F}_h^{\mathrm{i}}} 2(\{\mathcal{D}z\},\{y\})_{L,F} + \sum_{F\in\mathcal{F}_h^{\mathrm{i}}} (S_F([\![z]\!]),[\![y]\!])_{L,F} + \sum_{F\in\mathcal{F}_h} (R_F([\![z]\!]),[\![y]\!])_{L,F}.$$

The first and second terms on the right-hand side are the discrete counterparts of (2.4). The third term is a consistency term; it is zero whenever z is smooth and it is meant to guarantee the L-coercivity of a_h (recall that (A3) and (M1) imply that a is L-coercive). The discrete counterpart of (2.6) is formulated as follows:

$$\text{Seek } z_h \in W_h \text{ such that } a_h(z_h,y_h) = (f,y_h)_L \quad \forall y_h \in W_h. \tag{2.16}$$

Problem (2.16) can be equivalently reformulated in local form by introducing the notion of flux: Seek $z_h \in W_h$ such that for all $K \in \mathcal{T}_h$ and for all $y \in [\mathbb{P}_p(K)]^m$,

$$(Kz_h,y)_{L,K} + (z_h,\tilde{A}y)_{L,K} + (\phi_{\partial K}(z_h),y)_{L,\partial K} = (f,y)_{L,K}, \tag{2.17}$$

where the element fluxes are defined on a face $F \subset \partial K$ by

$$\phi_{\partial K}(z)|_F = \begin{cases} \frac{1}{2}\mathcal{D}_F z + \frac{1}{2}M_F(z) + R_F(z) & \text{if } F \in \mathcal{F}_h^\partial, \\ n_F{\cdot}n_K(\mathcal{D}_F\{z\} + S_F([\![z]\!]) + R_F([\![z]\!])) & \text{if } F \in \mathcal{F}_h^{\mathrm{i}}. \end{cases} \tag{2.18}$$

2.4. One-field DG approximation. For the one-field DG method, the operators $\{R_F\}_{\mathcal{F}_h}$ are generally set to zero and the operators $\{M_F\}_{\mathcal{F}_h^\partial}$ and $\{S_F\}_{\mathcal{F}_h^{\mathrm{i}}}$ are designed as follows: For all $v,w \in [L^2(F)]^m$,

(DG1A) $$\mathrm{Ker}(\mathcal{M}-\mathcal{D}) \subset \mathrm{Ker}(M_F - \mathcal{D}),$$

(DG1B) $$(M_F(v),v)_{L,F} \geq 0,$$

(DG1C) $$|(M_F(v) - \mathcal{D}v, w)_{L,F}| \lesssim |v|_{M,F}\|w\|_{L,F},$$

(DG1D) $$|(M_F(v) + \mathcal{D}v, w)_{L,F}| \lesssim \|v\|_{L,F}|w|_{M,F},$$

(DG1E) $$S_F = (S_F)^* \text{ and } |\mathcal{D}| \lesssim S_F \lesssim \mathcal{I}_m,$$

where $|v|_{M,F}^2 = (M_F(v),v)_{L,F}$, $\mathcal{I}_m$ is the identity matrix in $\mathbb{R}^{m,m}$, and $(S_F)^*$ is the adjoint operator of S_F. Assumption (DG1A) is a consistency assumption meaning that for all $F \in \mathcal{F}_h^\partial$ and for all $v \in [L^2(F)]^m$, $\mathcal{M}v = \mathcal{D}v$ implies $M_F(v) = \mathcal{D}v$. Design conditions for $\{S_F\}_{\mathcal{F}_h^{\mathrm{i}}}$ slightly more general than (DG1E) are stated in Part I.

To formulate the convergence result, we equip $W(h)$ with the following norms:

$$\|y\|_{h,A}^2 = \|y\|_L^2 + |y|_J^2 + |y|_M^2 + \sum_{K\in\mathcal{T}_h} h_K \|Ay\|_{L,K}^2, \tag{2.19}$$

$$\|y\|_{h,\frac{1}{2}}^2 = \|y\|_{h,A}^2 + \sum_{K\in\mathcal{T}_h} [h_K^{-1}\|y\|_{L,K}^2 + \|y\|_{L,\partial K}^2], \tag{2.20}$$

with $|y|_J^2 = \sum_{F\in\mathcal{F}_h^{\mathrm{i}}} |[\![y]\!]|_{J,F}^2$, $|y|_{J,F}^2 = (S_F([\![y]\!]),[\![y]\!])_{L,F}$, and $|y|_M^2 = \sum_{F\in\mathcal{F}_h^{\partial}} |y|_{M,F}^2$.

The main convergence result derived in Part I is the following.

THEOREM 2.2. *Assume* (DG1A)–(DG1E) *and* $\mathcal{A}^k \in [\mathfrak{C}^{0,\frac{1}{2}}(K)]^{m,m}$ *for all* $K\in\mathcal{T}_h$ *and all* $1\le k\le d$. *Let* $z\in[H^1(\Omega)]^m\cap V$ *solve* (2.6) *and let* z_h *solve* (2.16). *Then*

$$\|z-z_h\|_{h,A} \lesssim \inf_{y_h\in W_h} \|z-y_h\|_{h,\frac{1}{2}}. \tag{2.21}$$

Theorem 2.2 yields $(p+\frac{1}{2})$-order convergence in the L-norm and p-order convergence in the broken graph norm if the mesh family is quasi-uniform and the exact solution is in the broken Sobolev space $[H^{p+1}(\mathcal{T}_h)]^m$.

2.5. Two-field DG approximation. Let $p_u>0$ be a *positive* integer and take $p_\sigma\in\mathbb{N}$ such that $p_u-1\le p_\sigma$. Define the finite element spaces

$$\Sigma_h = [P_{h,p_\sigma}]^{m_\sigma}, \qquad U_h = [P_{h,p_u}]^{m_u}, \qquad W_h = \Sigma_h\times U_h. \tag{2.22}$$

The bilinear form a_h is still defined by (2.15) and the discrete problem is still (2.16).

The design of the operators $\{M_F\}_{F\in\mathcal{F}_h^{\partial}}$, $\{S_F\}_{F\in\mathcal{F}_h^{\mathrm{i}}}$, and $\{R_F\}_{F\in\mathcal{F}_h^{\partial}}$ for the two-field DG approximation hinges on the requirement that we be able to locally eliminate the discrete component z_h^σ. To this purpose, these operators are designed such that

(DG2A) $$M_F = \left[\begin{array}{c|c} 0 & -\alpha\mathcal{D}^{\sigma u} \\ \hline \alpha\mathcal{D}^{u\sigma} & M_F^{uu}\end{array}\right],\ \alpha\in\{-1,+1\}, \quad S_F = \left[\begin{array}{c|c} 0 & 0 \\ \hline 0 & S_F^{uu}\end{array}\right], \quad R_F = \left[\begin{array}{c|c} 0 & 0 \\ \hline 0 & R_F^{uu}\end{array}\right];$$

(DG2B) if $\alpha=+1$, $$\begin{cases} M_F^{uu} = (M_F^{uu})^* \quad\text{and}\quad \mathrm{Ker}(\mathcal{D}^{\sigma u})\subset\mathrm{Ker}(M_F^{uu}-\mathcal{D}^{uu}),\\ h_F^{-1}(\mathcal{D}^{u\sigma}\mathcal{D}^{\sigma u})^{\frac{1}{2}} + h_F|\mathcal{D}^{uu}| \lesssim M_F^{uu} \lesssim h_F^{-1}\mathcal{I}_{m_u};\end{cases}$$

(DG2C) if $\alpha=-1$, $M_F^{uu}(v) = \mathcal{M}^{uu}v$ and $|\mathcal{D}^{uu}|\lesssim\mathcal{M}^{uu}\lesssim\mathcal{I}_{m_u}$;

(DG2D) $S_F^{uu} = (S_F^{uu})^*$ and $h_F^{-1}(\mathcal{D}^{u\sigma}\mathcal{D}^{\sigma u})^{\frac{1}{2}} + h_F|\mathcal{D}^{uu}| \lesssim S_F^{uu} \lesssim h_F^{-1}\mathcal{I}_{m_u}$.

Furthermore, letting $|y^u|_J^2 = \sum_{F\in\mathcal{F}_h^{\mathrm{i}}} |y^u|_{J,F}^2$ and $|y^u|_M^2 = \sum_{F\in\mathcal{F}_h^{\partial}} |y^u|_{M,F}^2$, where

$$|y^u|_{J,F}^2 = (S_F^{uu}([\![y^u]\!]),[\![y^u]\!])_{L_u,F}, \qquad |y^u|_{M,F}^2 = (M_F^{uu}(y^u),y^u)_{L_u,F}, \tag{2.23}$$

the user-dependent operator R_F^{uu} must be designed such that for all $z_h^u\in U_h$ and all $(z^u,y_h^u)\in U(h)\times U_h$,

(DG2E) $$\sum_{F\in\mathcal{F}_h} (R_F^{uu}([\![z_h^u]\!]),[\![z_h^u]\!])_{L_u,F} \ge -\frac{1}{4}(|z_h^u|_J^2 + |z_h^u|_M^2),$$

(DG2F) $$\sum_{F\in\mathcal{F}_h} (R_F^{uu}([\![z^u]\!]),[\![y_h^u]\!])_{L_u,F} \lesssim (|z^u|_J + |z^u|_M)(|y_h^u|_J + |y_h^u|_M).$$

Owing to (DG2A), the σ-component of the element fluxes defined by (2.18) does not depend on z_h^σ, thus allowing for the local elimination of z_h^σ. Design conditions slightly more general than (DG2A)–(DG2D) are stated in Part II. Furthermore, assumptions (DG2A)–(DG2C) imply the consistency condition $\mathrm{Ker}(\mathcal{M}-\mathcal{D}) \subset \mathrm{Ker}(M_F-\mathcal{D})$ and the adjoint-consistency condition $\mathrm{Ker}(\mathcal{M}^t+\mathcal{D}) \subset \mathrm{Ker}(M_F^*+\mathcal{D})$. Indeed, both conditions are evident if $\alpha=-1$ since (DG2A) and (DG2C) imply $M_F=\mathcal{M}$. If $\alpha=+1$, the consistency condition directly results from (DG2B), while the adjoint-consistency condition results from the fact that if $z \in \mathrm{Ker}(\mathcal{M}^t+\mathcal{D})$, $\mathcal{D}^{\sigma u} z^u = 0$ and using (DG2B) yields $(\mathcal{M}^{uu})^t z^u = -\mathcal{D}^{uu} z^u = -M_F^{uu}(z^u) = -(M_F^{uu})^*(z^u)$. Finally, we observe that the second part of assumption (DG2C) imposes a condition on the way the Robin–Neumann boundary condition is enforced rather than on the DG setting; in practice, $\mathcal{M}^{uu} = |\mathcal{D}^{uu}|$ (see section 3.3.2) so that (DG2C) holds.

Equip $W(h)$ with the following norms:

$$\|y\|_{h,B}^2 = \|y\|_L^2 + |y^u|_J^2 + |y^u|_M^2 + \|B_h y^u\|_{L_\sigma}^2, \tag{2.24}$$

$$\|y\|_{h,1}^2 = \|y\|_{h,B}^2 + \sum_{K\in\mathcal{T}_h} [h_K^{-2}\|z^u\|_{L_u,K}^2 + h_K^{-1}\|z^u\|_{L_u,\partial K}^2 + h_K\|z^\sigma\|_{L_\sigma,\partial K}^2], \tag{2.25}$$

$$\|y\|_{h,1^+}^2 = \|y\|_{h,1}^2 + \sum_{K\in\mathcal{T}_h} [h_K^2\|y^\sigma\|_{[H^1(K)]^{m_\sigma}}^2 + h_K\|y^\sigma\|_{L_\sigma,\partial K}^2]. \tag{2.26}$$

The main convergence result proved in Part II is the following.

THEOREM 2.3. *Assume* (DG2A)–(DG2F) *and* $\mathcal{B}^k \in [\mathfrak{C}^{0,1}(K)]^{m_\sigma,m_u}$ *for all* $K\in\mathcal{T}_h$ *and* $1\le k\le d$. *Let* $z\in[H^1(\Omega)]^m\cap V$ *solve* (2.6) *and let* z_h *solve* (2.16). *Then*

$$\|z-z_h\|_{h,B} \lesssim \inf_{y_h\in W_h} \|z-y_h\|_{h,1}. \tag{2.27}$$

Moreover, if for every $y^u\in L_u$, *the solution* $\psi\in V^*$ *to the dual problem* $\tilde{T}\psi=(0,y^u)$ *is such that* $\|\psi^u\|_{[H^2(\Omega)]^{m_u}} + \|\psi^\sigma\|_{[H^1(\Omega)]^{m_\sigma}} \lesssim \|y^u\|_{L_u}$, *then*

$$\|z^u-z_h^u\|_{L_u} \lesssim h \inf_{y_h\in W_h} \|z-y_h\|_{h,1^+}. \tag{2.28}$$

If the exact solution is in $[H^{p_u}(\mathcal{T}_h)]^{m_\sigma}\times[H^{p_u+1}(\mathcal{T}_h)]^{m_u}$, Theorem 2.3 yields p_u-order convergence in the L_σ-norm for the σ-component, (p_u+1)-order convergence in the L_u-norm for the u-component, and p_u-order convergence in the broken graph norm for the u-component.

3. Two-field theory with L_σ-coercivity only. The goal of this section is to weaken assumption (A3), so as to be able to account for two-field Friedrichs' systems with no L_u-coercivity on the u-component. The model problems we have in mind are advection–diffusion equations with no zero-order term, i.e., no reaction (see section 3.3), and compressible linear continuum mechanics (see section 3.4).

3.1. Two-field Friedrichs' systems with L_σ-coercivity only. We assume

$$\text{(A3A)} \qquad \forall z\in L,\quad ((K+K^*-\nabla\cdot A)z,z)_L \gtrsim \|z^\sigma\|_{L_\sigma}^2,$$

$$\text{(A3B}') \qquad K^{\sigma u} = (K^{u\sigma})^* = 0 \text{ and the fields } \mathcal{B}^k \text{ are constant over } \Omega,$$

$$\text{(A3B}'') \qquad \forall z\in V\cup V^*,\quad \|z^u\|_{L_u} \lesssim a(z,z)^{\frac12} + \|Bz^u\|_{L_\sigma}.$$

Observe that (A3B″) is meaningful since a is L_σ-coercive on W. Indeed, the definition of D and assumptions (M1) and (A3A) yield, for all $z \in W$,

$$a(z,z) = \tfrac{1}{2}[(Tz,z)_L + (z,\tilde{T}z)_L] + \tfrac{1}{2}\langle Mz,z\rangle_{W',W} \gtrsim \|z^\sigma\|_{L_\sigma}^2. \tag{3.1}$$

Moreover, T (resp., T^*) is L_σ-coercive on V (resp., V^*).

THEOREM 3.1. *The conclusions of Theorem* 2.1 *still hold if assumption* (A3) *is replaced by assumptions* (A3A)–(A3B′)–(A3B″).

Proof. (1) Let us first prove that $T : V \to L$ is an isomorphism by using the so-called Banach–Nečas–Babuška (BNB) theorem, which states that the bijectivity of $T \in \mathcal{L}(V;L)$ is equivalent to the following conditions [14, p. 85]:

$$\forall z \in V, \quad \sup_{y\in L\setminus\{0\}} \frac{(Tz,y)_L}{\|y\|_L} = \|Tz\|_L \gtrsim \|z\|_W, \tag{3.2}$$

$$\forall y \in L, \quad ((Tz,y)_L = 0 \ \forall z \in V) \implies (y = 0). \tag{3.3}$$

Recall that the graph norm is $\|z\|_W = \|z\|_L + \|Az\|_L$ with $\|z\|_L = \|z^\sigma\|_{L_\sigma} + \|z^u\|_{L_u}$ and $\|Az\|_L = \|Bz^u\|_{L_\sigma} + \|B^\dagger z^\sigma + Cz^u\|_{L_u}$.

(1a) Proof of (3.2). Let $z \in V$. Owing to (A3B′), $Bz^u = (Tz)^\sigma - \mathcal{K}^{\sigma\sigma} z^\sigma$. Hence, $\|Bz^u\|_{L_\sigma} \lesssim \|z^\sigma\|_{L_\sigma} + \|Tz\|_L$. Then, since $a(z,z) = (Tz,z)_L$, (A3B″) implies

$$\|z^u\|_{L_u} \lesssim (Tz,z)_L^{\frac{1}{2}} + c(\|z^\sigma\|_{L_\sigma} + \|Tz\|_L) \lesssim \gamma\|z^u\|_{L_u} + (\|z^\sigma\|_{L_\sigma} + \|Tz\|_L),$$

where $\gamma > 0$ can be chosen as small as needed. Hence, $\|z^u\|_{L_u} \lesssim \|z^\sigma\|_{L_\sigma} + \|Tz\|_L$. Combining this result with the L_σ-coercivity of T on V yields

$$\|z^\sigma\|_{L_\sigma}^2 \lesssim \frac{(Tz,z)_L}{\|z\|_L}(\|z^\sigma\|_{L_\sigma} + \|z^u\|_{L_u}) \lesssim \|Tz\|_L(\|z^\sigma\|_{L_\sigma} + \|Tz\|_L),$$

whence $\|z^\sigma\|_{L_\sigma} \lesssim \|Tz\|_L$. Collecting the above bounds leads to $\|z\|_L \lesssim \|Tz\|_L$, and hence $\|z\|_W = \|z\|_L + \|Az\|_L \lesssim \|z\|_L + \|Tz\|_L \lesssim \|Tz\|_L$.

(1b) Proof of (3.3). Assume that $y \in L$ is such that $(Tz,y)_L = 0$ for all $z \in V$. Following the same arguments as in the proof of [15, Theorem 2.5] in Part I or Corollary 5.8 in [14], we infer that $y \in V^*$ and $\tilde{T}y = 0$. The L_σ-coercivity of $\tilde{T}$ on V^* yields $y^\sigma = 0$. Proceeding as above using (A3B′) leads to $\|z^u\|_{L_u} \lesssim \|z^\sigma\|_{L_\sigma} + \|\tilde{T}z\|_L$ for all $z \in V^*$. Applying this estimate to y yields $y^u = 0$. Hence, $y = 0$.

(2) Since $T : V \to L$ is an isomorphism and $V = \mathrm{Ker}(D - M)$, a solution to (2.6) is readily constructed by setting $z = T^{-1}f$. To prove uniqueness, let us prove that the only solution to (2.6) with $f = 0$ is $z = 0$. Since a is L_σ-coercive on W, $z^\sigma = 0$. In addition, taking $y \in [\mathfrak{D}(\Omega)]^m$ in (2.6) yields $Tz = 0$ in L and $z \in V$. Hence, the bound $\|Bz^u\|_{L_\sigma} \lesssim \|z^\sigma\|_{L_\sigma} + \|Tz\|_L$ implies $Bz^u = 0$ and owing to (A3B″), $z^u = 0$.

(3) Proceed similarly to prove that problem (2.7) is well-posed. ∎

3.2. Two-field DG approximation with L_σ-coercivity only. Consider the two-field DG method introduced in section 2.5 and assume that conditions (DG2A)–(DG2F) are fulfilled. The objective of this section is to analyze the convergence of the two-field DG approximation in the framework of the partial coercivity assumptions (A3A), (A3B′), and (A3B″). The discrete counterpart of assumption (A3B″) is

$$\forall z_h \in W_h, \quad \|z_h^u\|_{L_u} \lesssim a_h(z_h,z_h)^{\frac{1}{2}} + \|B_h z_h^u\|_{L_\sigma}. \tag{3.4}$$

Recall that the norm $\|\cdot\|_{h,B}$ is defined by (2.24).

LEMMA 3.2. *Assume* (A3A)–(A3B′), (3.4), *and* (DG2A)–(DG2F). *Then*

$$\forall z_h \in W_h, \qquad \|z_h\|_{h,B} \lesssim \sup_{y_h \in W_h\setminus\{0\}} \frac{a_h(z_h,y_h)}{\|y_h\|_{h,B}}. \tag{3.5}$$

Proof. Let $z_h \in W_h$. Owing to the definition of a_h, (DG2A), (DG2E), and (A3A),

$$\|z_h^\sigma\|_{L_\sigma}^2 + |z_h^u|_J^2 + |z_h^u|_M^2 \lesssim a_h(z_h,z_h). \tag{3.6}$$

Set $\varpi_h = (B_h z_h^u, 0)$ and observe that $\varpi_h \in W_h$ since the fields $\mathcal{B}^k$ are constant over Ω and $p_u - 1 \le p_\sigma$. Moreover,

$$\begin{aligned}\|B_h z_h^u\|_{L_\sigma}^2 &= a_h(z_h,\varpi_h) - (\mathcal{K}^{\sigma\sigma} z_h^\sigma, B_h z_h^u)_{L_\sigma} - \sum_{F\in\mathcal{F}_h^\partial} \tfrac{\alpha+1}{2} (\mathcal{D}^{\sigma u} z_h^u, B_h z_h^u)_{L_{\sigma},F} \\ &\quad + \sum_{F\in\mathcal{F}_h^{\mathrm{i}}} 2(\{\mathcal{D}^{\sigma u} z_h^u\}, \{B_h z_h^u\})_{L_\sigma,F} := a_h(z_h,\varpi_h) + R_1 + R_2 + R_3.\end{aligned}$$

Clearly, $|R_1| \lesssim \|z_h^\sigma\|_{L_\sigma}^2 + \gamma \|B_h z_h^u\|_{L_\sigma}^2$, where $\gamma > 0$ can be chosen as small as needed. If $\alpha = +1$, use (DG2B) and a trace inverse inequality to infer

$$|R_2| \lesssim \sum_{F\in\mathcal{F}_h^\partial} h_F^{\frac12} |z_h^u|_{M,F} h_F^{-\frac12} \|B_h z_h^u\|_{L_\sigma,\mathcal{T}(F)} \lesssim |z_h^u|_M^2 + \gamma \|B_h z_h^u\|_{L_\sigma}^2,$$

while if $\alpha = -1$, $R_2 = 0$. Finally, using $\{\mathcal{D}^{\sigma u}\} = 0$ and (DG2D) leads to $|R_3| \lesssim |z_h^u|_J^2 + \gamma \|B_h z_h^u\|_{L_\sigma}^2$. Collecting the above bounds yields

$$\|B_h z_h^u\|_{L_\sigma}^2 \lesssim a_h(z_h,\varpi_h) + a_h(z_h,z_h),$$

and owing to (3.4) and (3.6), it is inferred that $\|z_h\|_{h,B}^2 \lesssim a_h(z_h,\varpi_h) + a_h(z_h,z_h)$. Conclude using the fact that $\|\varpi_h\|_{h,B} = \|B_h z_h^u\|_{L_\sigma} \lesssim \|z_h\|_{h,B}$. ∎

It is now straightforward to verify the following convergence result.

THEOREM 3.3. *The statement of Theorem* 2.3 *remains valid under the assumptions of Lemma* 3.2.

3.3. Example 1: Advection–diffusion. Let $f \in L^2(\Omega)$ and let $\beta \in [L^\infty(\Omega)]^d$ with $\nabla\cdot\beta \in L^\infty(\Omega)$. Let $\mu \in L^\infty(\Omega)$ and let $\kappa = (\kappa_{kl})_{1\le k,l\le d}$ be a symmetric positive definite tensor-valued field defined on Ω whose lowest eigenvalue is uniformly bounded away from zero. Consider the PDE $-\nabla\cdot(\kappa\nabla u) + \beta\cdot\nabla u + \mu u = f$ in mixed form

$$\begin{cases} \kappa^{-1}\sigma + \nabla u = 0, \\ \mu u + \nabla\cdot\sigma + \beta\cdot\nabla u = f. \end{cases} \tag{3.7}$$

Letting $m = d+1$, $m_\sigma = d$, and $m_u = 1$, the mixed formulation (3.7) fits the two-field framework by setting, for all $z \in L$ and for all $k \in \{1,\dots,d\}$,

$$K(z) = \left[\begin{array}{c:c} \kappa^{-1} & 0 \\ \hdashline 0 & \mu \end{array}\right] z, \qquad \mathcal{A}^k = \left[\begin{array}{c:c} 0 & e^k \\ \hdashline (e^k)^t & \beta^k \end{array}\right], \tag{3.8}$$

where e^k is the kth vector in the canonical basis of $\mathbb{R}^d$, and β^k is the kth component of β. Clearly, hypotheses (A1), (A2), (A4), and (A5) hold. We further assume that

$$\inf_\Omega \operatorname{ess}(\mu - \tfrac12 \nabla\cdot\beta) \ge 0 \tag{3.9}$$

so that (A3) does not hold, but (A3A) holds instead with μ_0 equal to the reciprocal of the largest eigenvalue of κ. This situation covers, in particular, the Laplace/Poisson equation where $\mu = 0$ and $\beta = 0$.

The graph space is $W = H(\mathrm{div};\Omega) \times H^1(\Omega)$ and the boundary operator D is such that for all $z, y \in W$,

$$\langle Dz, y\rangle_{W',W} = \langle z^\sigma \cdot n, y^u\rangle_{-\frac{1}{2},\frac{1}{2}} + \langle y^\sigma \cdot n, z^u\rangle_{-\frac{1}{2},\frac{1}{2}} + \int_{\partial\Omega} (\beta\cdot n) z^u y^u, \tag{3.10}$$

where $\langle , \rangle_{-\frac{1}{2},\frac{1}{2}}$ denotes the duality pairing between $H^{-\frac{1}{2}}(\partial\Omega)$ and $H^{\frac{1}{2}}(\partial\Omega)$. Dirichlet boundary conditions can be enforced by setting

$$\langle Mz, y\rangle_{W',W} = \langle z^\sigma \cdot n, y^u\rangle_{-\frac{1}{2},\frac{1}{2}} - \langle y^\sigma \cdot n, z^u\rangle_{-\frac{1}{2},\frac{1}{2}}, \tag{3.11}$$

yielding $V = H(\mathrm{div};\Omega)\times H^1_0(\Omega)$. Furthermore, mixed Robin–Neumann boundary conditions can be enforced by setting

$$\langle Mz, y\rangle_{W',W} = -\langle z^\sigma \cdot n, y^u\rangle_{-\frac{1}{2},\frac{1}{2}} + \langle y^\sigma \cdot n, z^u\rangle_{-\frac{1}{2},\frac{1}{2}} + \int_{\partial\Omega} (2\varrho + \beta\cdot n) z^u y^u, \tag{3.12}$$

where $\varrho \in L^\infty(\partial\Omega)$ is such that $2\varrho + \beta\cdot n \geq 0$ a.e. on $\partial\Omega$. Then $V = \{z \in W;\, z^\sigma\cdot n - \varrho z^u|_{\partial\Omega} = 0\}$ and $V^* = \{z \in W;\, z^\sigma\cdot n + (\varrho + \beta\cdot n) z^u|_{\partial\Omega} = 0\}$. In terms of boundary fields, (3.10), (3.11), (3.12), respectively, yield

$$\mathcal{D} = \left[\begin{array}{c:c} 0 & n \\ \hdashline n^t & \beta\cdot n \end{array}\right], \qquad \mathcal{M} = \left[\begin{array}{c:c} 0 & -n \\ \hdashline n^t & 0 \end{array}\right], \qquad \mathcal{M} = \left[\begin{array}{c:c} 0 & n \\ \hdashline -n^t & 2\varrho + \beta\cdot n \end{array}\right]. \tag{3.13}$$

Observe that $\mathrm{Ker}(\mathcal{D}^{\sigma u}) = \{0\}$. Furthermore, a possible choice for the mixed Robin–Neumann boundary condition is $\varrho = -\min(\beta\cdot n, 0)$ (or, equivalently, $\mathcal{M}^{uu} = |\beta\cdot n| = |\mathcal{D}^{uu}|$), yielding the usual Robin (inflow) condition $(z^\sigma + \beta z^u)\cdot n = 0$ on $\partial\Omega^- = \{x \in \partial\Omega;\, \beta(x)\cdot n(x) < 0\}$ and the usual Neumann (outflow) condition $z^\sigma\cdot n = 0$ on $\partial\Omega\backslash\partial\Omega^-$.

3.3.1. Well-posedness. Let us verify that the advection–diffusion equation equipped with the above boundary conditions fits the theoretical framework analyzed in section 3.1.

PROPOSITION 3.4. *Assumptions* (A3A), (A3B′), *and* (A3B″) *hold for Dirichlet boundary conditions and for mixed Robin–Neumann boundary conditions, provided either* $\mu - \frac{1}{2}\nabla\cdot\beta \neq 0$ *or* $\varrho + \frac{1}{2}\beta\cdot n \neq 0$ *(this means that either* $\mu - \frac{1}{2}\nabla\cdot\beta$ *or* $\varrho + \frac{1}{2}\beta\cdot n$ *is uniformly bounded away from zero on a measurable subset of* Ω *or of* $\partial\Omega$*, respectively, of nonzero measure).*

Proof. Assumptions (A3A) and (A3B′) are evident. Observe also that $Bz^u = \nabla z^u$.

(1) For Dirichlet boundary conditions, (A3B″) directly results from the Poincaré inequality since $z \in V = V^*$ implies $z^u \in H^1_0(\Omega)$ so that $\|z^u\|_{L_u} \lesssim \|\nabla z^u\|_{L_\sigma}$.

(2) For mixed Robin–Neumann boundary conditions, observe that for all $z \in W$,

$$a(z,z) \geq \int_\Omega (\mu - \tfrac{1}{2}\nabla\cdot\beta)(z^u)^2 + \int_{\partial\Omega} (\varrho + \tfrac{1}{2}\beta\cdot n)(z^u)^2. \tag{3.14}$$

We apply Lemma 3.5 below; it is a simple variant of the Petree–Tartar lemma (the proof is omitted for brevity). Take $X = H^1(\Omega)$, $Y = [L^2(\Omega)]^d$, $Z = L^2(\Omega)$, $Fx = \nabla x$, $Gx = x$, and $\Phi(x) = \|(\mu - \frac{1}{2}\nabla\cdot\beta)^{\frac{1}{2}} x\|_{L^2(\Omega)} + \|(\varrho + \frac{1}{2}\beta\cdot n)^{\frac{1}{2}} x\|_{L^2(\partial\Omega)}$. Properties (i) and (iii) are evident. Property (ii) results from the fact that if $\|Fx\|_Y + \Phi(x) = 0$,

then x is constant (since Ω is connected) and $\Phi(x) = 0$ implies that $x = 0$ since $\mu - \frac{1}{2}\nabla\cdot\beta \neq 0$ or $\varrho + \frac{1}{2}\beta\cdot n \neq 0$. Moreover, Φ is continuous since for all $x, y \in X$, $|\Phi(x) - \Phi(y)| \leq \Phi(x - y) \lesssim \|x - y\|_X$. Hence, Lemma 3.5 yields $\|z^u\|_{L_u} \lesssim \Phi(z^u) + \|\nabla z^u\|_{L_\sigma}$, whence (A3B$''$) follows owing to (3.14). ∎

LEMMA 3.5. *Let X, Y, Z be Banach spaces, let $F \in \mathcal{L}(X;Y)$, and let $G \in \mathcal{L}(X;Z)$. Let $\Phi : X \to \mathbb{R}_+$ be a continuous functional. Assume that*

(i) *G is compact;*

(ii) *for all $x \in X$, $(\|Fx\|_Y + \Phi(x) = 0) \Rightarrow (x = 0)$;*

(iii) *there is $\gamma_1 > 0$ such that for all $x \in X$, $\gamma_1\|x\|_X \leq \|Fx\|_Y + \|Gx\|_Z$.*

Then there is $\gamma_2 > 0$ such that for all $x \in X$, $\gamma_2\|x\|_X \leq \|Fx\|_Y + \Phi(x)$.

Remark 3.1. When mixed Robin–Neumann boundary conditions are enforced and $\mu - \frac{1}{2}\nabla\cdot\beta = 0$ and $2\varrho + \beta\cdot n = 0$, (A3B$''$) no longer holds and the analysis proceeds as follows. If $\mu \neq 0$ or $\varrho \neq 0$, it is easily verified that T is injective on V and that $\tilde{T}$ is injective on V^*. The injectivity of T on V combined with Lemma 3.5 easily yields (3.2), while the injectivity of $\tilde{T}$ on V^* yields (3.3); thus well-posedness holds. If $\mu = \nabla\cdot\beta = 0$ and $\varrho = \beta\cdot n = 0$, then T is no longer injective on V, the compatibility condition $\langle f\rangle_\Omega = 0$ must be imposed on the data, and the solution u is subjected to the constraint $\langle u\rangle_\Omega = 0$ (here, for a function $\phi \in L^2(\Omega)$, $\langle\phi\rangle_\Omega := \frac{1}{|\Omega|}\int_\Omega \phi$, where $|\Omega|$ denotes the measure of Ω). Hence, we modify (3.7) as follows:

$$\begin{cases} \kappa^{-1}\sigma + \nabla u = 0, \\ \nabla\cdot\sigma + \beta\cdot\nabla u + \langle u\rangle_\Omega = f, \end{cases} \tag{3.15}$$

equipped with the boundary condition $\sigma\cdot n|_{\partial\Omega} = 0$. Since $\langle\nabla\cdot\sigma + \beta\cdot\nabla u\rangle_\Omega = \langle f\rangle_\Omega = 0$, the second PDE implies $\langle u\rangle_\Omega = 0$, i.e., (3.15) is equivalent to (3.7) together with $\langle u\rangle_\Omega = 0$. Moreover, for all $z \in W$, $a(z,z) \geq |\Omega|\langle z^u\rangle_\Omega^2$, so that (A3B$''$) results from the fact that for all $\phi \in H^1(\Omega)$, $\|\phi\|_{L_u} \lesssim \langle\phi\rangle_\Omega + \|\nabla\phi\|_{L_\sigma}$.

3.3.2. Two-field DG approximation. When Dirichlet boundary conditions are enforced, let $\eta_1 > 0$, $\eta_2 > 0$ (these parameters can vary from face to face), and set

$$M_F^{uu}(v) = \eta_1 h_F^{-1} v, \qquad S_F^{uu}(v) = \eta_2 h_F^{-1} v, \qquad R_F^{uu} \equiv 0. \tag{3.16}$$

Since $\mathcal{D}^{u\sigma}\mathcal{D}^{\sigma u} = 1$ and $\mathcal{D}^{uu} = \beta\cdot n$, properties (DG2A)–(DG2F) hold. Many other choices can be considered for M_F^{uu}, S_F^{uu}, and R_F^{uu}; see Part II [16] for details.

PROPOSITION 3.6. *Property* (3.4) *holds.*

Proof. The proof is a direct consequence of the fact that for all $v_h \in U_h$,

$$\|v_h\|_{L_u}^2 \lesssim \sum_{K\in\mathcal{T}_h} \|\nabla v_h\|_{L_\sigma,K}^2 + \sum_{F\in\mathcal{F}_h^{\mathrm{i}}} h_F^{-1}\|[\![v_h]\!]\|_{L_u,F}^2 + \sum_{F\in\mathcal{F}_h^{\partial}} h_F^{-1}\|v_h\|_{L_u,F}^2.$$

See [1, 6] or [14, p. 134] for the proof. ∎

When mixed Robin–Neumann boundary conditions are enforced, (DG2C) holds for $M_F^{uu}(v) = \mathcal{M}^{uu}v = (2\varrho + \beta\cdot n)v$, provided $\varrho \geq -\min(\beta\cdot n, 0)$, while S_F^{uu} can be chosen as in (3.16).

PROPOSITION 3.7. *Assume that either $\mu - \frac{1}{2}\nabla\cdot\beta \neq 0$ or $\varrho + \frac{1}{2}\beta\cdot n \neq 0$ and that Ω is such that $H^{\frac{3}{2}+\epsilon}$-elliptic regularity holds, $\epsilon > 0$. Then property* (3.4) *holds.*

Proof. Let v_h be an arbitrary function in U_h. Let $\psi \in H^1(\Omega)$ solve

$$(\mu - \tfrac{1}{2}\nabla\cdot\beta)\psi - \Delta\psi = v_h, \qquad \partial_n\psi|_{\partial\Omega} = -(\rho + \tfrac{1}{2}\beta\cdot n)\psi.$$

This problem is well-posed (proceed as in the proof of Proposition 3.4), and the $H^{\frac{3}{2}+\epsilon}$-elliptic regularity hypothesis means that $\|\psi\|_{H^{\frac{3}{2}+\epsilon}} \lesssim \|v_h\|_{L_u}$. Testing with v_h yields

$$\begin{aligned}\|v_h\|_{L_u}^2 &= ((\mu - \tfrac{1}{2}\nabla\cdot\beta)\psi, v_h)_{L_u} + \sum_{K\in\mathcal{T}_h} (\nabla\psi, \nabla v_h)_{L_\sigma,K} \\ &\quad - \sum_{F\in\mathcal{F}_h^{\mathrm{i}}} \int_F 2\nabla\psi\cdot\{nv_h\} + \sum_{F\in\mathcal{F}_h^{\partial}} \int_F (\varrho + \tfrac{1}{2}\beta\cdot n)\psi v_h \\ &\lesssim \|\psi\|_{H^{\frac{3}{2}+\epsilon}} \Big(\|(\mu - \tfrac{1}{2}\nabla\cdot\beta)^{\frac{1}{2}} v_h\|_{L_u} + \Big(\sum_{K\in\mathcal{T}_h} \|\nabla v_h\|_{L_\sigma,K}^2 \Big)^{\frac{1}{2}} + |v_h|_J \\ &\qquad + \|(\varrho + \tfrac{1}{2}\beta\cdot n)^{\frac{1}{2}} v_h\|_{L_u,\partial\Omega} \Big).\end{aligned}$$

The conclusion follows readily. ∎

3.4. Example 2: Compressible linear continuum mechanics. Let $\beta \in [L^\infty(\Omega)]^d$ with $\nabla\cdot\beta \in L^\infty(\Omega)$, let λ, $\delta \in L^\infty(\Omega)$, and let $f \in [L^2(\Omega)]^d$. Consider the set of PDEs

$$\begin{cases} \sigma - \frac{1}{d+\delta}\mathrm{tr}(\sigma)\mathcal{I}_d - \frac{1}{2}(\nabla u + (\nabla u)^t) = 0, \\ -\frac{1}{2}\nabla\cdot(\sigma + \sigma^t) + \beta\cdot\nabla u + \lambda u = f, \end{cases} \tag{3.17}$$

where σ is $\mathbb{R}^{d,d}$-valued and u is $\mathbb{R}^d$-valued. Assuming $\delta \neq 0$, the first equation in (3.17) implies $\sigma = \frac{1}{2}(\nabla u + (\nabla u)^t) + \delta^{-1}(\nabla\cdot u)\mathcal{I}_d$. Equations (3.17) are a linearized version of the compressible Navier–Stokes equations and model also linear continuum mechanics (when $\lambda = 0$ and $\beta = 0$). In the context of linear continuum mechanics these equations are usually referred to as the displacement-stress formulation. The tensor field σ with values in $\mathbb{R}^{d,d}$ can be identified with the vector field $\overline{\sigma}$ with values in $\mathbb{R}^{d^2}$ by setting $\overline{\sigma}_{[ij]} = \sigma_{ij}$ with $1 \le i,j \le d$ and $[ij] := d(j-1)+i$. To alleviate the notation, we henceforth use the same symbol for both fields.

Set $m = d^2 + d$, $m_\sigma = d^2$, and $m_u = d$. The mixed formulation (3.17) fits in the framework of two-field Friedrichs' systems by setting, for all $z \in L$ and for all $k \in \{1,\dots,d\}$,

$$K(z) = \left[\begin{array}{c:c} \mathcal{I}_{d^2} - \frac{1}{d+\delta}\mathcal{Z} & 0 \\ \hdashline 0 & \lambda\mathcal{I}_d \end{array}\right] z, \qquad \mathcal{A}^k = \left[\begin{array}{c:c} 0 & \mathcal{E}^k \\ \hdashline (\mathcal{E}^k)^t & \mathcal{C}^k \end{array}\right], \tag{3.18}$$

where $\mathcal{Z} \in \mathbb{R}^{d^2,d^2}$ is such that $\mathcal{Z}_{[ij][kl]} = \delta_{ij}\delta_{kl}$ with $1 \le i,j,k,l \le d$, and for all $k \in \{1,\dots,d\}$, $\mathcal{C}^k = \beta^k\mathcal{I}_d \in \mathbb{R}^{d,d}$ and $\mathcal{E}^k \in \mathbb{R}^{d^2,d}$ is such that $\mathcal{E}^k_{[ij],l} = -\frac{1}{2}(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk})$ with $1 \le i,j,l \le d$; here, the δ's are Kronecker symbols. Clearly, hypotheses (A1), (A2), and (A4) hold. We further assume that

$$\lambda_0 := \inf_\Omega \mathrm{ess}(\lambda - \tfrac{1}{2}\nabla\cdot\beta) \ge 0, \qquad \delta_0 := \inf_\Omega \mathrm{ess}\,\delta > 0. \tag{3.19}$$

Then observing that

$$\begin{aligned}\mathcal{K}^{\sigma\sigma}\sigma\cdot\sigma = (\sigma - \tfrac{1}{d+\delta}\mathrm{tr}(\sigma)\mathcal{I}_d) : \sigma &= \tfrac{\delta}{d+\delta}\|\sigma\|^2 + \tfrac{d}{d+\delta}\|\sigma - \tfrac{1}{d}\mathrm{tr}(\sigma)\mathcal{I}_d\|^2 \\ &= \tfrac{\delta}{(d+\delta)^2}|\mathrm{tr}(\sigma)|^2 + \|\sigma - \tfrac{1}{d+\delta}\mathrm{tr}(\sigma)\mathcal{I}_d\|^2,\end{aligned} \tag{3.20}$$

where $\|\cdot\|$ is the Frobenius norm, we deduce that (A5) also holds. The incompressible limit $\delta_0 = 0$ is treated in sections 3.5 and 4.3. Note that (A3) does not hold ((A3) would hold if $\lambda_0 > 0$. The case $\lambda_0 \geq 0$ covers the usual compressible solid mechanics problems for which $\lambda = 0$ and $\beta = 0$.

The graph space is $W = H_\sigma \times [H^1(\Omega)]^d$ with $H_\sigma = \{\sigma \in [L^2(\Omega)]^{d,d};\ \nabla\cdot(\sigma+\sigma^t) \in [L^2(\Omega)]^d\}$ and the boundary operator D is such that for all $z, y \in W$ with $z = (\sigma, u)$ and $y = (\tau, v)$,

$$\langle Dz, y\rangle_{W',W} = -\langle \tfrac{1}{2}(\tau+\tau^t)\cdot n, u\rangle_{-\frac{1}{2},\frac{1}{2}} - \langle \tfrac{1}{2}(\sigma+\sigma^t)\cdot n, v\rangle_{-\frac{1}{2},\frac{1}{2}} + \int_{\partial\Omega} (\beta\cdot n)uv, \tag{3.21}$$

where $\langle , \rangle_{-\frac{1}{2},\frac{1}{2}}$ denotes the duality pairing between $[H^{-\frac{1}{2}}(\partial\Omega)]^d$ and $[H^{\frac{1}{2}}(\partial\Omega)]^d$. An homogeneous Dirichlet boundary condition on the u-component is obtained by setting

$$\langle Mz, y\rangle_{W',W} = \langle \tfrac{1}{2}(\tau+\tau^t)\cdot n, u\rangle_{-\frac{1}{2},\frac{1}{2}} - \langle \tfrac{1}{2}(\sigma+\sigma^t)\cdot n, v\rangle_{-\frac{1}{2},\frac{1}{2}}, \tag{3.22}$$

yielding $V = V^* = H_\sigma \times [H^1_0(\Omega)]^d$. Similarly, a mixed Robin–Neumann boundary condition is obtained by setting

$$\langle Mz, y\rangle_{W',W} = -\langle \tfrac{1}{2}(\tau+\tau^t)\cdot n, u\rangle_{-\frac{1}{2},\frac{1}{2}} + \langle \tfrac{1}{2}(\sigma+\sigma^t)\cdot n, v\rangle_{-\frac{1}{2},\frac{1}{2}} + \int_{\partial\Omega} (2\varrho+\beta\cdot n)uv, \tag{3.23}$$

where $\varrho \in L^\infty(\partial\Omega)$ is such that $2\varrho + \beta\cdot n \geq 0$ a.e. on $\partial\Omega$. Then $V = \{(\sigma, u) \in W;\ \frac{1}{2}(\sigma+\sigma^t)\cdot n - \varrho u|_{\partial\Omega} = 0\}$ and $V^* = \{(\sigma, u) \in W;\ \frac{1}{2}(\sigma+\sigma^t)\cdot n + (\varrho+\beta\cdot n)u|_{\partial\Omega} = 0\}$. A standard choice is $\varrho = -\min(\beta\cdot n, 0)$.

In terms of boundary fields, letting $\mathcal{N} = \sum_{k=1}^d n_k \mathcal{E}^k \in \mathbb{R}^{d^2,d}$, (3.21), (3.22), and (3.23), respectively, yields

$$\mathcal{D} = \left[\begin{array}{c:c} 0 & \mathcal{N} \\ \hdashline \mathcal{N}^t & (\beta\cdot n)\mathcal{I}_d \end{array}\right], \quad \mathcal{M} = \left[\begin{array}{c:c} 0 & -\mathcal{N} \\ \hdashline \mathcal{N}^t & 0 \end{array}\right], \quad \mathcal{M} = \left[\begin{array}{c:c} 0 & \mathcal{N} \\ \hdashline -\mathcal{N}^t & (2\varrho+\beta\cdot n)\mathcal{I}_d \end{array}\right]. \tag{3.24}$$

Observe that $\mathrm{Ker}(\mathcal{D}^{\sigma u}) = \{0\}$ and that $\mathcal{M}^{uu} = |\mathcal{D}^{uu}|$ in the Robin–Neumann case, provided $\varrho = -\min(\beta\cdot n, 0)$.

3.4.1. Well-posedness. For the sake of brevity, we restrict ourselves to homogeneous Dirichlet boundary conditions. The case of mixed Robin–Neumann boundary conditions can be treated by proceeding as in section 3.3.1.

PROPOSITION 3.8. *Assume $\lambda_0 = 0$ and $\delta_0 > 0$ and that Dirichlet boundary conditions are enforced. Then assumptions* (A3A)*,* (A3B′)*, and* (A3B″) *hold.*

Proof. Assumptions (A3A)–(A3B′) are evident. Moreover, since Dirichlet boundary conditions are enforced on the u-component, $z \in V = V^*$ implies $z^u \in [H^1_0(\Omega)]^d$; hence, (A3B″) results from Korn's first inequality. ∎

3.4.2. Two-field DG approximation. We assume again for simplicity that homogeneous Dirichlet boundary conditions are enforced. Let $\eta_1 > 0$, $\eta_2 > 0$ (these parameters can vary from face to face), and

$$M_F^{uu}(v) = \eta_1 h_F^{-1} v, \qquad S_F^{uu}(v) = \eta_2 h_F^{-1} v, \qquad R_F^{uu} \equiv 0. \tag{3.25}$$

Since $\mathcal{D}^{u\sigma}\mathcal{D}^{\sigma u} = \frac{1}{2}(\mathcal{I}_d + n\otimes n)$ and $\mathcal{D}^{uu} = (\beta\cdot n)\mathcal{I}_d$, properties (DG2A)–(DG2D) hold. Many other choices can be considered for M_F^{uu}, S_F^{uu}, and R_F^{uu}.

PROPOSITION 3.9. *Property* (3.4) *holds.*

Proof. The proof is a direct consequence of the fact that for all $v_h \in U_h$,

$$\|v_h\|_{L_u}^2 \lesssim \sum_{K\in\mathcal{T}_h} \|\nabla v_h + (\nabla v_h)^t\|_{L_\sigma,K}^2 + \sum_{F\in\mathcal{F}_h^{\mathrm{i}}} h_F^{-1}\|[\![v_h]\!]\|_{L_u,F}^2 + \sum_{F\in\mathcal{F}_h^{\partial}} h_F^{-1}\|v_h\|_{L_u,F}^2.$$

See [7, 13] for a proof. ∎

3.5. Remarks on the incompressible limit. The above analysis is not uniform with respect to the compressibility factor δ, i.e., the analysis fails as $\delta \to 0$; this is the well-known locking phenomenon. The origin of the problem is that the operator $K^{\sigma\sigma}$ is not invertible at the limit $\delta = 0$; indeed, since $K^{\sigma\sigma}\sigma = \sigma - \frac{1}{d+\delta}\mathrm{tr}(\sigma)\mathcal{I}_d$, it is clear that $K^{\sigma\sigma}\sigma = 0$ when $\delta = 0$ and σ is proportional to the identity. In other words, it is our insisting on eliminating σ that is responsible for locking (i.e., the primal formulation locks).

Actually the above two-field DG approximation is robust with respect to δ if one *abandons* the idea of eliminating σ; in other words, the mixed formulation is robust. This fact has already been recognized in Franca and Stenberg [20] using the GaLS technique, and this is also true for conforming mixed approximations; see Brezzi and Fortin [8] and Schwab and Suri [28]. (A DG method (IP) where σ is eliminated and which does not lock for small δ is analyzed in Hansbo and Larson [23]. This method, though, yields a linear system with a condition number that grows unboundedly as $\delta \to 0$.)

Let us now substantiate the above claim. Since it is not more difficult to analyze the general case, let us forget problem (3.17) and let us deal with the general DG approximation (2.16). We now weaken the L_σ-coercivity requirement used in section 3.1 as follows. We replace (A3A) by the following: Assume that there exists an operator $\pi \in \mathcal{L}(L_\sigma, L_\sigma)$ such that

$$(\text{A3A}') \qquad \forall z \in L, \qquad ((K + K^* - \nabla\cdot A)z, z)_L \gtrsim \|\mathcal{K}^{\sigma\sigma}z^\sigma\|_{L_\sigma}^2 + \|\pi z^\sigma\|_{L_\sigma}^2.$$

We also modify (DG2A) as follows:

$$(\text{DG2A}') \qquad M_F = \left[\begin{array}{c|c} 0 & -\alpha\mathcal{D}^{\sigma u} \\ \hline \alpha\mathcal{D}^{u\sigma} & M_F^{uu} \end{array}\right],\ \alpha\in\{-1,+1\}, \quad S_F = \left[\begin{array}{c|c} S_F^{\sigma\sigma} & 0 \\ \hline 0 & S_F^{uu} \end{array}\right], \quad R_F \equiv 0,$$

where $S_F^{\sigma\sigma}$ is defined by

$$(\text{DG2A}'') \qquad (S_F^{\sigma\sigma}(z^\sigma), y^\sigma)_{L_\sigma,F} = \eta h_F (z^\sigma - \pi z^\sigma, y^\sigma - \pi y^\sigma)_{L_\sigma,F},$$

with user-dependent parameter $\eta > 0$. Note that owing to the presence of $S_F^{\sigma\sigma}$, z_h^σ can no longer be locally eliminated. We introduce the notation $|z^\sigma|_{J'}^2 := \sum_{F\in\mathcal{F}_h^{\mathrm{i}}} |z^\sigma|_{J',F}^2$, $|z^\sigma|_{J',F}^2 = (S_F^{\sigma\sigma}([\![z^\sigma]\!]), [\![z^\sigma]\!])_{L_\sigma,F}$, and we redefine the $\|\cdot\|_{h,B}$ discrete norm as

$$\|y\|_{h,B}^2 := \|y\|_L^2 + |y^u|_J^2 + |y^\sigma|_{J'}^2 + |y^u|_M^2 + \|B_h y^u\|_{L_\sigma}^2. \tag{3.26}$$

To ascertain control over $z_h^\sigma - \pi z_h^\sigma$ we make the two following assumptions:

$$\forall z_h^\sigma \in \Sigma_h, \quad \|z_h^\sigma - \pi z_h^\sigma\|_{L_\sigma}^2 \lesssim \|\pi z_h^\sigma\|_{L_\sigma}^2 + |z_h^\sigma|_{J'}^2 + \left(\sup_{0\neq v_h^u\in U_h} \frac{\mathbb{B}_h(z_h^\sigma, v_h^u)}{\|(0, v_h^u)\|_{h,B}}\right)^2, \tag{3.27}$$

$$\forall z_h^u \in U_h, \quad \|C_h z_h^u\|_{L_u} \lesssim \|B_h z_h^u\|_{L_\sigma} + \|z_h^u\|_{L_u}, \tag{3.28}$$

where $\mathbb{B}_h(z_h^\sigma, v_h^u) := a_h((z_h^\sigma, 0), (0, v_h^u))$, i.e.,

$$\mathbb{B}_h(z_h^\sigma, v_h^u) := (B_h^\dagger z_h^\sigma, v_h^u)_{L_u} - 2\sum_{F\in\mathcal{F}_h^{\rm i}} (\{\mathcal{D}^{u\sigma} z_h^\sigma\}, \{v_h^u\})_{L_u,F} + \tfrac{\alpha-1}{2}\sum_{F\in\mathcal{F}_h^{\partial}} (\mathcal{D}^{u\sigma} z_h^\sigma, v_h^u)_{L_u,F}. \tag{3.29}$$

THEOREM 3.10. *The statement of Theorem* 2.3 *remains true under the assumptions* (A3A′)–(A3B′), (DG2A′)–(DG2A″), (DG2B)–(DG2F), *and* (3.4), (3.27), *and* (3.28).

Proof. We just sketch the proof since most of the arguments will be repeated with more details in section 4.2. Let us prove the inf-sup inequality

$$\forall z_h \in W_h, \qquad \|z_h\|_{h,B} \lesssim \sup_{y_h\in W_h\setminus\{0\}} \frac{a_h(z_h, y_h)}{\|y_h\|_{h,B}}. \tag{3.30}$$

Let $z_h \in W_h$ and define $\mathbb{S} := \sup_{y_h\in W_h\setminus\{0\}} \frac{a_h(z_h,y_h)}{\|y_h\|_{h,B}}$. Owing to the definition of a_h, (DG2A′)–(DG2A″), and (A3A′),

$$\|\mathcal{K}^{\sigma\sigma} z_h^\sigma\|_{L_\sigma}^2 + \|\pi z_h^\sigma\|_{L_\sigma}^2 + |z_h^u|_J^2 + |z_h^\sigma|_{J'}^2 + |z_h^u|_M^2 \lesssim a_h(z_h, z_h) \lesssim \mathbb{S}\,\|z_h\|_{h,B}.$$

Set $\varpi_h = (B_h z_h^u, 0)$ and observe that $\varpi_h \in W_h$ since the fields $\mathcal{B}^k$ are constant over Ω (see (A3B′)) and $p_u - 1 \le p_\sigma$. Then

$$\|B_h z_h^u\|_{L_\sigma}^2 = a_h(z_h, \varpi_h) - (\mathcal{K}^{\sigma\sigma} z_h^\sigma, B_h z_h^u)_{L_\sigma} - \sum_{F\in\mathcal{F}_h^{\partial}} \tfrac{\alpha+1}{2} (\mathcal{D}^{\sigma u} z_h^u, B_h z_h^u)_{L_\sigma,F} + \sum_{F\in\mathcal{F}_h^{\rm i}} 2(\{\mathcal{D}^{\sigma u} z_h^u\}, \{B_h z_h^u\})_{L_\sigma,F} - (S_F^{\sigma\sigma}(z_h^\sigma), B_h z_h^u)_{L_\sigma,F}.$$

Proceeding as in the proof of Lemma 3.2 and observing that $\|\varpi_h\|_{h,B} \lesssim \|B_h z_h^u\|_{L_\sigma} + |B_h z_h^u|_{J'} \lesssim \|z_h\|_{h,B}$, we infer

$$\|B_h z_h^u\|_{L_\sigma}^2 \lesssim a_h(z_h, \varpi_h) + a_h(z_h, z_h) \lesssim \mathbb{S}\,\|z_h\|_{h,B}.$$

Using (3.4) then yields $\|z_h^u\|_{L_u}^2 \lesssim \mathbb{S}\,\|z_h\|_{h,B}$. Finally, to control $\|z_h^\sigma - \pi z_h^\sigma\|_{L_\sigma}$, observe that $\mathbb{B}_h(z_h^\sigma, v_h^u) = a_h(z_h, (0, v_h^u)) - a_h(0, z_h^u), (0, v_h^u))$ and using (3.27)–(3.28) proceed as in the proof of Lemma 4.2, step (4c). This yields $\|z_h^\sigma - \pi z_h^\sigma\|_{L_\sigma}^2 \lesssim \mathbb{S}\,\|z_h\|_{h,B}$. The rest of the proof is standard. ∎

Coming back to our compressible linear elasticity problem (3.17), observe that the above analysis applies by setting $\pi\sigma = \sigma - \frac{1}{d}\mathrm{tr}(\sigma)\mathcal{I}_d$. Inequality (A3A′) holds *uniformly* with respect to δ, as can be seen by taking the mean of the two equations in (3.20). Assumption (3.28) results from Korn's second inequality. A proof for (3.27) is given in the proof of Proposition 4.5. In conclusion, the present two-field DG formulation is robust with respect to δ. However, owing to the presence of the stabilization operator $S_F^{\sigma\sigma}$ which penalizes the jumps of $\mathrm{tr}(\sigma_h)$, σ_h cannot be eliminated locally. This motivates the introduction of a new scalar variable (the pressure) whose jumps will be penalized instead of those of $\mathrm{tr}(\sigma_h)$.

4. Three-field theory. The goal of this section is to weaken even further the set of hypotheses (A3A), (A3B′), and (A3B″). We have in mind the linear elasticity equations, and we want to come up with a method which is robust in the incompressible limit and for which the σ-component can be eliminated locally. To this purpose, we introduce a three-field theory of Friedrichs' systems and adapt the DG approximation to this setting. Thus, we assume that z can be decomposed into three fields, z^σ, z^p, and z^u, and we have in mind to locally eliminate z^σ.

4.1. Three-field Friedrichs' systems. We now assume that $L = L_\sigma \times L_p \times L_u$, where $L_\sigma = [L^2(\Omega)]^{m_\sigma}$, $L_p = [L^2(\Omega)]^{m_p}$, and $L_u = [L^2(\Omega)]^{m_u}$, with $m = m_\sigma + m_p + m_u$. We also assume that the operator K and the matrices $\mathcal{A}^k$, $1 \le k \le d$, admit the following 3×3 structure:

$$K = \left[\begin{array}{c:c:c} K^{\sigma\sigma} & K^{\sigma p} & 0 \\ \hdashline K^{p\sigma} & K^{pp} & 0 \\ \hdashline 0 & 0 & K^{uu} \end{array}\right] \qquad \text{and} \qquad \mathcal{A}^k = \left[\begin{array}{c:c:c} 0 & 0 & \mathcal{B}^k \\ \hdashline 0 & 0 & 0 \\ \hdashline [\mathcal{B}^k]^t & 0 & \mathcal{C}^k \end{array}\right]. \tag{4.1}$$

We assume that the block $K^{\sigma\sigma}$ has a local representation, i.e., there is $\mathcal{K}^{\sigma\sigma} \in [L^\infty(\Omega)]^{m_\sigma, m_\sigma}$ such that $K^{\sigma\sigma}(y^\sigma) = \mathcal{K}^{\sigma\sigma} y^\sigma$ for all $y^\sigma \in L_\sigma$ (this local representation is needed in the three-field DG method described below to locally eliminate the discrete σ-component). We denote by K^σ (resp., K^p) the canonical projection of K onto L_σ (resp., L_p). Finally, we assume that there exists $\pi \in \mathcal{L}(L_\sigma; L_\sigma)$ such that

(A3C) $$\forall z \in L, \quad ((K + K^* - \nabla\cdot A)z, z)_L \gtrsim \|\pi z^\sigma\|_{L_\sigma}^2 + \|K^\sigma z\|_{L_\sigma}^2 + \|K^p z\|_{L_p}^2,$$

(A3D) $$\forall z \in V \cup V^*, \quad \|z^u\|_{L_u} \lesssim a(z,z)^{\frac12} + \|Bz^u\|_{L_\sigma},$$

(A3E) $$\forall z \in W, \quad \|Cz^u\|_{L_u} \lesssim \|z^u\|_{L_u} + \|Bz^u\|_{L_\sigma},$$

(A3F) $$\forall z \in V \cup V^*, \quad \|z^\sigma - \pi z^\sigma\|_{L_\sigma} \lesssim \|\pi z^\sigma\|_{L_\sigma} + \|B^\dagger z^\sigma\|_{L_u},$$

(A3G) $$\forall z \in L, \quad \|z^p\|_{L_p}^2 \lesssim ((K + K^* - \nabla\cdot A)z, z)_L + \|K^{pp} z^p\|_{L_p}^2,$$

(A3H) $$\forall k \in \{1, \dots, k\}, \quad \mathcal{B}^k \text{ is constant over } \Omega,$$

and that (A3C) and (A3G) also hold when K^σ, K^p, and K^{pp} are substituted by the corresponding terms in K^*, say $K^{*\sigma}$, K^{*p}, and K^{*pp}.

THEOREM 4.1. *The conclusions of Theorem* 2.1 *still hold if assumption* (A3) *is replaced by assumptions* (A3C)–(A3H).

Proof. The proof, which is similar to that of Theorem 3.1, is only sketched. We only prove that $T : V \longrightarrow L$ is an isomorphism, since the rest of the proof is unchanged or goes along the same lines.

(1) Proof of (3.2). Let $z \in V$. Recall that $(Tz, z)_L = a(z,z) \ge ((K + K^* - \nabla\cdot A)z, z)_L$. Hence, property (A3C) implies

$$\|K^p z\|_{L_p} + \|K^\sigma z\|_{L_\sigma} + \|\pi z^\sigma\|_{L_\sigma} \lesssim (Tz, z)_L^{\frac12}.$$

Since $Bz^u + K^\sigma z = (Tz)^\sigma$, we infer $\|Bz^u\|_{L_\sigma} \lesssim \|Tz\|_L + (Tz, z)_L^{\frac12}$. Then, owing to (A3D)–(A3E),

$$\|z^u\|_{L_u} + \|Cz^u\|_{L_u} \lesssim \|Tz\|_L + (Tz, z)_L^{\frac12}.$$

Now we use (A3F), the above bounds, and the fact that $B^\dagger z^\sigma = (Tz)^u - Cz^u - K^{uu} z^u$ to deduce

$$\|z^\sigma - \pi z^\sigma\|_{L_\sigma} \lesssim \|B^\dagger z^\sigma\|_{L_u} + (Tz, z)_L^{\frac12} \lesssim \|Tz\|_L + (Tz, z)_L^{\frac12},$$

and the same bound holds for $\|z^\sigma\|_{L_\sigma}$ by the triangle inequality. To derive a bound on $\|z^p\|_{L_p}$ we use the fact that $K^p z = (Tz)^p$ to infer

$$\|K^{pp} z^p\|_{L_p} \le \|K^p z\|_{L_p} + \|K^{p\sigma} z^\sigma\|_{L_p} \lesssim \|Tz\|_L + \|z^\sigma\|_{L_\sigma}.$$

Hence, (A3G) and the above bounds yield $\|z^p\|_{L_p} \lesssim \|Tz\|_L + (Tz,z)_L^{\frac12}$. Combining the previous estimates yields $\|z\|_L \lesssim \|Tz\|_L + (Tz,z)_L^{\frac12}$ and we conclude as usual.

(2) Proof of (3.3). Assume that $y \in L$ is such that $(Tz,y)_L = 0$ for all $z \in V$. Then $y \in V^*$ and $\tilde T y = 0$. Hence (A3C) implies $\pi y^\sigma = 0$ and $K^{*\sigma} y = 0$. Then, observing that $0 = (\tilde T y)^\sigma = K^{*\sigma} y - B y^u$ since the fields $\mathcal{B}^k$ are constant, we infer $By^u = 0$. Then (A3D) and (A3E) yield $y^u = 0$ and $Cy^u = 0$. Using $0 = (\tilde T y)^u$ and the fact that the fields $\mathcal{B}^k$ are constant, we then infer $B^\dagger y^\sigma = 0$ so that (A3F) implies $y^\sigma - \pi y^\sigma = 0$, and hence $y^\sigma = 0$. Finally, since $0 = (\tilde T y)^p = K^{*pp} y^p$, using (A3G) we infer $y^p = 0$, thus completing the proof. ∎

4.2. Three-field DG approximation. We analyze in this section a DG method to approximate the three-field Friedrichs' systems introduced in section 4.1. We assume that hypotheses (A3C)–(A3H) hold so that the continuous problem is well-posed.

Let $p_u > 0$ be a *positive* integer and let p_σ and p_p be such that

$$p_u - 1 \le p_\sigma \le p_u + 1, \qquad p_\sigma \le p_p. \tag{4.2}$$

Consider the finite elements spaces

$$\Sigma_h = [P_{h,p_\sigma}]^{m_\sigma}, \quad P_h = [P_{h,p_p}]^{m_p}, \quad U_h = [P_{h,p_u}]^{m_u}, \quad W_h = \Sigma_h \times P_h \times U_h. \tag{4.3}$$

Consider the discrete problem (2.16) with the bilinear form still defined by (2.15).

The key property of the three-field DG approximation developed hereafter is that the operators M_F and S_F are designed in such a way that the discrete σ-component can be locally eliminated. We consider either Dirichlet boundary conditions or mixed Robin–Neumann boundary conditions enforced by setting $\mathcal{M}^{uu} = |\mathcal{D}^{uu}|$; see sections 3.3 and 3.4. The design conditions of the three-field DG method are the following:

(DG3A) $$M_F = \left[\begin{array}{c|c|c} 0 & 0 & -\alpha\mathcal{D}^{\sigma u} \\ \hline 0 & 0 & 0 \\ \hline \alpha\mathcal{D}^{u\sigma} & 0 & M_F^{uu} \end{array}\right], \quad S_F = \left[\begin{array}{c|c|c} 0 & 0 & 0 \\ \hline 0 & S_F^{pp} & 0 \\ \hline 0 & 0 & S_F^{uu} \end{array}\right], \quad R_F = \left[\begin{array}{c|c|c} 0 & 0 & 0 \\ \hline 0 & 0 & 0 \\ \hline 0 & 0 & R_F^{uu} \end{array}\right];$$

(DG3B) $$\text{if } \alpha = +1, \begin{cases} M_F^{uu} = (M_F^{uu})^* \quad \text{and} \quad \operatorname{Ker}(\mathcal{D}^{\sigma u}) \subset \operatorname{Ker}(M_F^{uu} - \mathcal{D}^{uu}), \\ h_F^{-1}(\mathcal{D}^{u\sigma}\mathcal{D}^{\sigma u})^{\frac12} + h_F^{-1}|\mathcal{D}^{uu}| \lesssim M_F^{uu} \lesssim h_F^{-1}\mathcal{I}_{m_u}; \end{cases}$$

(DG3C) $$\text{if } \alpha = -1,\ M_F^{uu}(v) = \mathcal{M}^{uu} v;$$

(DG3D) $$S_F^{uu} = (S_F^{uu})^*, \text{ and } h_F^{-1}(\mathcal{D}^{u\sigma}\mathcal{D}^{\sigma u})^{\frac12} + h_F^{-1}|\mathcal{D}^{uu}| \lesssim S_F^{uu} \lesssim h_F^{-1}\mathcal{I}_{m_u};$$

(DG3E) $$S_F^{pp} = (S_F^{pp})^*, \quad \text{and} \quad h_F \mathcal{I}_{m_p} \lesssim S_F^{pp} \lesssim h_F \mathcal{I}_{m_p},$$

where $\alpha \in \{-1,+1\}$ in the definition of M_F in (DG3A).

Our aim is to control the approximation error in the norm $\|\cdot\|_{h,B}$ defined by

$$\|y\|_{h,B}^2 = \|y\|_L^2 + |y^u|_M^2 + |y^u|_{J^u}^2 + |y^p|_{J^p}^2 + \|B_h y^u\|_{L_\sigma}^2, \tag{4.4}$$

with $|y^u|_{J^u}^2 = \sum_{F\in\mathcal{F}_h^{\mathrm{i}}} |y^u|_{J^u,F}^2$, $|y^p|_{J^p}^2 = \sum_{F\in\mathcal{F}_h^{\mathrm{i}}} |y^p|_{J^p,F}^2$, $|y^u|_M^2 = \sum_{F\in\mathcal{F}_h^{\partial}} |y^u|_{M,F}^2$,

$$|y^u|_{J^u,F}^2 = (S_F^{uu}(\llbracket y^u \rrbracket), \llbracket y^u \rrbracket)_{L_u,F}, \quad |y^p|_{J^p,F}^2 = (S_F^{pp}(\llbracket y^p \rrbracket), \llbracket y^p \rrbracket)_{L_p,F}, \tag{4.5}$$

$$|y^u|_{M,F}^2 = (M_F^{uu}(y^u), y^u)_{L_u,F}. \tag{4.6}$$

The user-dependent operator R_F^{uu} must be designed so that for all $z_h^u \in U_h$ and all $(z^u, y_h^u) \in U(h)\times U_h$,

$$\text{(DG3F)} \qquad \sum_{F\in\mathcal{F}_h} (R_F^{uu}([\![z_h^u]\!]), [\![z_h^u]\!])_{L_u,F} \geq -\frac{1}{4}(|z_h^u|_{J^u}^2 + |z_h^u|_M^2),$$

$$\text{(DG3G)} \qquad \sum_{F\in\mathcal{F}_h} (R_F^{uu}([\![z^u]\!]), [\![y_h^u]\!])_{L_u,F} \lesssim (|z^u|_{J^u} + |z^u|_M)(|y_h^u|_{J^u} + |y_h^u|_M).$$

The discrete counterpart of assumption (A3D) is still (3.4), while the discrete counterpart of assumption (A3F) is the following: For all $z_h^\sigma \in \Sigma_h$,

$$\|z_h^\sigma - \pi z_h^\sigma\|_{L_\sigma}^2 \lesssim \|\pi z_h^\sigma\|_{L_\sigma}^2 + \sum_{F\in\mathcal{F}_h^{\mathrm{i}}} h_F \|[\![z_h^\sigma - \pi z_h^\sigma]\!]\|_{L_\sigma,F}^2 + \left(\sup_{0\neq v_h^u\in U_h} \frac{\mathbb{B}_h(z_h^\sigma, v_h^u)}{\|(0,0,v_h^u)\|_{h,B}}\right)^2, \tag{4.7}$$

where $\mathbb{B}_h(z_h^\sigma, v_h^u)$ is defined by (3.29). Finally, the discrete counterpart of assumption (A3E) is

$$\forall z_h^u \in U_h, \qquad \|C_h z_h^u\|_{L_u} \lesssim \|B_h z_h^u\|_{L_\sigma} + \|z_h^u\|_{L_u}. \tag{4.8}$$

Since the jumps of the σ-component are not controlled in the three-field DG method (so as to eliminate this component locally), stability must come from the control on the jumps of the p-component. The link between the jumps of the σ- and p-components is provided by the equation for the p-component. This motivates the following additional assumptions:

$$\forall z_h \in W_h, \quad K^p z_h \in P_h, \tag{4.9}$$

$$\forall z_h \in W_h, \quad \forall F \in \mathcal{F}_h^{\mathrm{i}}, \quad \|[\![z_h^\sigma - \pi z_h^\sigma]\!]\|_{L_\sigma,F} \lesssim \|[\![K^p z_h]\!]\|_{L_p,F} + \|[\![z_h^p]\!]\|_{L_p,F}. \tag{4.10}$$

LEMMA 4.2. *Assume that* (A3C)–(A3H) *and the discrete assumptions* (DG3A)–(DG3G), (3.4), (4.2), (4.7), (4.8), (4.9), *and* (4.10) *hold. Then the following holds:*

$$\forall z_h \in W_h, \qquad \|z_h\|_{h,B} \lesssim \sup_{y_h\in W_h\setminus\{0\}} \frac{a_h(z_h, y_h)}{\|y_h\|_{h,B}}. \tag{4.11}$$

Proof. Let $z_h \in W_h$ and set $\mathbb{S} = \sup_{y_h\in W_h\setminus\{0\}} \frac{a_h(z_h,y_h)}{\|y_h\|_{h,B}}$.

(1) Owing to the definition of a_h, (DG3A), (DG3F), and (A3C),

$$\|K^\sigma z_h\|_{L_\sigma}^2 + \|K^p z_h\|_{L_p}^2 + |\pi z_h^\sigma|_{L_\sigma}^2 + |z_h^u|_M^2 + |z_h^u|_{J^u}^2 + |z_h^p|_{J^p}^2 \lesssim a_h(z_h, z_h) \leq \mathbb{S}\,\|z_h\|_{h,B}. \tag{4.12}$$

(2) Control on z_h^u, $C_h z_h^u$, and $B_h z_h^u$. Set $y_h = (B_h z_h^u, 0, 0)$ and observe that $y_h \in W_h$ owing to (A3H) and the fact that $p_u - 1 \leq p_\sigma$. Moreover, $\|y_h\|_{h,B} \lesssim \|z_h\|_{h,B}$. Furthermore,

$$\|B_h z_h^u\|_{L_\sigma}^2 = a_h(z_h, y_h) - (K^\sigma z_h, B_h z_h^u)_{L_\sigma} - \tfrac{\alpha+1}{2} \sum_{F\in\mathcal{F}_h^\partial} (\mathcal{D}^{\sigma u} z_h^u, B_h z_h^u)_{L_\sigma,F} + \sum_{F\in\mathcal{F}_h^{\mathrm{i}}} 2(\{\mathcal{D}^{\sigma u} z_h^u\}, \{B_h z_h^u\})_{L_\sigma,F} := T_1 + T_2 + T_3 + T_4.$$

Clearly, $|T_1| \lesssim \mathbb{S}\,\|z_h\|_{h,B}$ and owing to (4.12),

$$|T_2| \lesssim \|K^\sigma z_h\|_{L_\sigma}\|B_h z_h^u\|_{L_\sigma} \lesssim \mathbb{S}\,\|z_h\|_{h,B} + \gamma\|B_h z_h^u\|_{L_\sigma}^2,$$

where $\gamma > 0$ can be chosen as small as needed. Similarly, using (DG3B), $\{\mathcal{D}^{\sigma u}\} = 0$, (DG3D), and a trace inverse inequality leads to

$$|T_3| + |T_4| \lesssim (|z_h^u|_M + |z_h^u|_{J^u})\|B_h z_h^u\|_{L_\sigma} \lesssim \mathbb{S}\,\|z_h\|_{h,B} + \gamma\|B_h z_h^u\|_{L_\sigma}^2.$$

Collecting the above bounds and choosing the γ's small enough, it is inferred that $\|B_h z_h^u\|_{L_\sigma}^2 \lesssim \mathbb{S}\,\|z_h\|_{h,B}$. Then, owing to (3.4), (4.8), and (4.12), this in turn implies

$$\|z_h^u\|_{L_u}^2 + \|C_h z_h^u\|_{L_u}^2 + \|B_h z_h^u\|_{L_\sigma}^2 \lesssim \mathbb{S}\,\|z_h\|_{h,B}. \tag{4.13}$$

(3) Control on z_h^σ. It remains to control $z_h^\sigma - \pi z_h^\sigma$. The idea is to use (4.7) by estimating the three terms on the right-hand side, say R_1–R_3.

(3a) Clearly, $R_1 \lesssim \mathbb{S}\,\|z_h\|_{h,B}$.

(3b) To control R_2, use (4.10) and the fact that $\|[\![K^p z_h]\!]\|_{L_p,F} \lesssim h_F^{-1/2}\|K^p z_h\|_{L_p,\mathcal{T}(F)}$ owing to a trace inverse inequality (which can be used by (4.9)) to infer

$$\|[\![z_h^\sigma - \pi z_h^\sigma]\!]\|_{L\sigma,F}^2 \lesssim h_F^{-1}\|K^p z_h\|_{L_p,\mathcal{T}(F)}^2 + h_F^{-1}|z_h^p|_{J^p,F}^2.$$

Hence, owing to the estimates for $\|K^p z_h\|_{L_p}$ and $|z_h^p|_{J^p,F}$ resulting from step (1), $R_2 \lesssim \mathbb{S}\,\|z_h\|_{h,B}$.

(3c) For R_3, observe that $\mathbb{B}_h(z_h^\sigma, v_h^u) = a_h(z_h,(0,0,v_h^u)) - a_h((0,0,z_h^u),(0,0,v_h^u))$ and that for all $v_h^u \in U_h$, the following holds:

$$a_h((0,0,z_h^u),(0,0,v_h^u)) \lesssim \mathbb{A}_h := (\mathbb{S}\,\|z_h\|_{h,B})^{\frac{1}{2}}\|(0,0,v_h^u)\|_{h,B}. \tag{4.14}$$

Indeed,

$$\begin{aligned} a_h((0,0,z_h^u),(0,0,v_h^u)) &= (K^{uu}z_h^u, v_h^u)_{L_u} + (C_h z_h^u, v_h^u)_{L_u} \\ &+ \sum_{F\in\mathcal{F}_h^\partial} \tfrac{1}{2}(M_F^{uu}(z_h^u) - \mathcal{D}^{uu}z_h^u, v_h^u)_{L_u,F} - \sum_{F\in\mathcal{F}_h^{\mathrm{i}}} 2(\{\mathcal{D}^{uu}z_h^u\},\{v_h^u\})_{L_u,F} \\ &+ \sum_{F\in\mathcal{F}_h^{\mathrm{i}}} ((S_F^{uu}+R_F^{uu})([\![z_h^u]\!]),[\![v_h^u]\!])_{L_u,F} := T_1+T_2+T_3+T_4+T_5. \end{aligned}$$

Using (4.13) yields $|T_1| + |T_2| \lesssim \mathbb{A}_h$, while (4.12) together with (DG3G) readily yields $|T_5| \lesssim \mathbb{A}_h$. Since $\{\mathcal{D}^{uu}\} = 0$, using (DG3D) and a trace inverse inequality leads to

$$|T_4| \lesssim \sum_{F\in\mathcal{F}_h^{\mathrm{i}}} h_F^{\frac{1}{2}}|z_h^u|_{J^u,F}h_F^{-\frac{1}{2}}\|v_h^u\|_{L_u,\mathcal{T}(F)} \lesssim \mathbb{A}_h.$$

Finally, to control T_3, we proceed similarly using (DG3B) if $\alpha = +1$ to infer $|T_3| \lesssim |z_h^u|_M(|v_h^u|_M + \|v_h^u\|_{L_u}) \lesssim \mathbb{A}_h$, while if $\alpha = -1$, we use (DG3C) and the assumption $\mathcal{M}^{uu} = |\mathcal{D}^{uu}|$ to infer $|T_3| \lesssim |z_h^u|_M|v_h^u|_M \lesssim \mathbb{A}_h$. Collecting the bounds for T_1–T_5 yields (4.14), whence the bound $R_3 \lesssim \mathbb{S}^2 + \mathbb{S}\,\|z_h\|_{h,B}$ is readily inferred.

(3d) Collecting the bounds for R_1–R_3 yields

$$\|z_h^\sigma - \pi z_h^\sigma\|_{L_\sigma}^2 \le R_1 + R_2 + R_3 \lesssim \mathbb{S}^2 + \mathbb{S}\,\|z_h\|_{h,B}. \tag{4.15}$$

The same bound is inferred for $\|z_h^\sigma\|_{L_\sigma}^2$ by the triangle inequality.

(4) Control on z_h^p. Using (A3G), the fact that $K^{pp}z_h^p = K^p z_h - K^{p\sigma}z_h^\sigma$, and the derived bounds on $\|K^p z_h\|_{L_p}$ and $\|z_h^\sigma\|_{L_\sigma}$ leads to

$$\|z_h^p\|_{L_p}^2 \lesssim a_h(z_h,z_h) + \|K^p z_h\|_{L_p}^2 + \|K^{p\sigma}z_h^\sigma\|_{L_p}^2 \lesssim \mathbb{S}^2 + \mathbb{S}\,\|z_h\|_{h,B}. \tag{4.16}$$

(5) Conclusion. Collecting the above bounds yields $\|z_h\|_{h,B}^2 \lesssim \mathbb{S}^2 + \mathbb{S}\,\|z_h\|_{h,B}$, whence (4.11) readily follows. ∎

Remark 4.1. Assumption (DG3D) (resp., (DG3B)) requires a stronger control on $|\mathcal{D}^{uu}|$ with respect to (DG2D) (resp., (DG2B)). This stronger control is needed to prove (4.14) in step (3c) of the above proof. This is not really a restriction, since in practice $|\mathcal{D}^{uu}| \lesssim (\mathcal{D}^{u\sigma}\mathcal{D}^{\sigma u})^{\frac{1}{2}}$ so that $h_F^{-1}(\mathcal{D}^{u\sigma}\mathcal{D}^{\sigma u})^{\frac{1}{2}} \lesssim S_F^{uu}$ already yields $h_F^{-1}|\mathcal{D}^{uu}| \lesssim S_F^{uu}$ (see the examples in sections 3.3 and 3.4).

It is now straightforward to verify the following convergence result.

THEOREM 4.3. *The statement of Theorem* 2.3 *remains valid under the assumptions of Lemma* 4.2.

4.3. Example: Linear continuum mechanics. Let us consider again problem (3.17) in section 3.4 and let us introduce a third field $p := -\frac{1}{d+\delta}\mathrm{tr}(\sigma)$ so that (3.17) can be recast into

$$\begin{cases} \sigma + p\mathcal{I}_d - \frac{1}{2}(\nabla u + (\nabla u)^t) = 0, \\ \mathrm{tr}(\sigma) + (d+\delta)p = 0, \\ -\frac{1}{2}\nabla\cdot(\sigma+\sigma^t) + \beta\cdot\nabla u + \lambda u = f; \end{cases} \tag{4.17}$$

p is often referred to as the pressure. This type of problem arises in the modeling of non-Newtonian fluids. (4.17) is the limit of the so-called upper convected Maxwell model for small relaxation times; see Fortin and Pierre [19], Fortin, Guénette, and Pierre [18], and Schwab and Suri [28]. Although three-field models may not be easy to implement or may lead to increased computational costs, they appear to be more robust for high Weissenberg numbers; see Baaijens for a review on this question [3].

Instead of (3.19), we now assume that

$$\lambda_0 \geq 0, \qquad \delta_0 = 0. \tag{4.18}$$

This setting covers incompressible materials (i.e., $\lambda = 0$, $\beta = 0$, and $\delta = 0$) and incompressible fluid flows. For the sake of simplicity, we henceforth restrict ourselves to homogeneous Dirichlet boundary conditions, i.e., $\alpha = 1$. Then the pressure is defined up to a constant, and to avoid this arbitrariness, we choose the representative of the pressure which is of zero mean, i.e., $\langle p\rangle_\Omega = 0$. Accordingly, we modify slightly the equations as follows:

$$\begin{cases} \sigma + p\mathcal{I}_d - \frac{1}{2}(\nabla u + (\nabla u)^t) = 0, \\ \mathrm{tr}(\sigma) + (d+\delta)p + \langle p\rangle_\Omega = 0, \\ -\frac{1}{2}\nabla\cdot(\sigma+\sigma^t) + \beta\cdot\nabla u + \lambda u = f. \end{cases} \tag{4.19}$$

Note that by taking the trace of the first equation and integrating over Ω, we obtain $\langle \mathrm{tr}(\sigma) + dp\rangle_\Omega = \langle \nabla\cdot u\rangle_\Omega = 0$ for homogeneous Dirichlet boundary conditions on u. Accounting for the second equation in (4.19) yields $\langle \mathrm{tr}(\sigma)\rangle_\Omega = \langle p\rangle_\Omega = 0$. Hence, we are really solving (4.17) with $\langle p\rangle_\Omega = 0$. Note that K is self-adjoint and that K^{pp} is the only nonlocal block in the operator K.

4.3.1. Well-posedness. Define $\pi \in \mathcal{L}(L_\sigma; L_\sigma)$ such that for all $\sigma \in L_\sigma$,

$$\pi\sigma = \sigma - \tfrac{1}{d}(\mathrm{tr}(\sigma) - \langle \mathrm{tr}(\sigma)\rangle_\Omega)\mathcal{I}_d. \tag{4.20}$$

PROPOSITION 4.4. *Assumptions* (A3C)–(A3H) *hold.*

Proof. Let us first prove (A3C). A straightforward calculation yields, for all $z = (\sigma, p, 0) \in W$,

$$\begin{aligned}(Kz, z)_L &= \|\sigma + p\mathcal{I}_d\|_{L_\sigma}^2 + \delta\|p\|_{L_p}^2 + \langle p\rangle_\Omega^2 \\ &= \tfrac{d}{d+\delta}\|\sigma - \tfrac{1}{d}\mathrm{tr}(\sigma)\mathcal{I}_d\|_{L_\sigma}^2 + \tfrac{1}{d+\delta}\|\mathrm{tr}(\sigma) + (d+\delta)p\|_{L_p}^2 + \tfrac{\delta}{d+\delta}\|\sigma\|_{L_\sigma}^2 + \langle p\rangle_\Omega^2.\end{aligned}$$

Hence, taking the mean of the two above equations and using triangle inequalities,

$$\gamma(Kz, z)_L \geq \|\pi\sigma\|_{L_\sigma}^2 + \|K^\sigma z\|_{L_\sigma}^2 + \|K^p z\|_{L_p}^2 + \langle p\rangle_\Omega^2,$$

with γ independent of δ_0. Since K is self-adjoint, $\lambda_0 \geq 0$, and the fields $\mathcal{B}^k$ are constant over Ω, it is inferred that for all $z = (\sigma, p, u) \in W$,

$$\tfrac{1}{2}((K + K^* - \nabla\cdot A)z, z)_L \geq (K(\sigma, p, 0), (\sigma, p, 0))_L.$$

This proves (A3C). Assumption (A3D) is a simple consequence of Korn's first inequality since $V = V^* = H_\sigma \times L^2(\Omega) \times [H_0^1(\Omega)]^d$. Assumption (A3E) results from Korn's second inequality. Let us prove (A3F). Let $z \in V$ be such that $\sigma \neq \pi\sigma$; then $\mathrm{tr}(\sigma - \pi\sigma) \neq 0$. Since $\mathrm{tr}(\sigma - \pi\sigma) = \mathrm{tr}(\sigma) - \langle \mathrm{tr}(\sigma)\rangle_\Omega \in L_0^2(\Omega)$, there is $0 \neq v \in [H_0^1(\Omega)]^d$ such that $\nabla\cdot v = \mathrm{tr}(\sigma - \pi\sigma)$ and $\|v\|_{[H^1(\Omega)]^d} \lesssim \|\sigma - \pi\sigma\|_{L_\sigma}$. Since $\langle \nabla\cdot v\rangle_\Omega = 0$, we have $Bv - \pi Bv = \sigma - \pi\sigma$. Hence,

$$\begin{aligned}\|\sigma - \pi\sigma\|_{L_\sigma} &\lesssim \tfrac{1}{\|v\|_{[H^1(\Omega)]^d}}(Bv - \pi Bv, \sigma - \pi\sigma)_{L_\sigma} \lesssim \tfrac{1}{\|v\|_{[H^1(\Omega)]^d}}(Bv, \sigma - \pi\sigma)_{L_\sigma} \\ &\lesssim \|\pi\sigma\|_{L_\sigma} + \tfrac{1}{\|v\|_{[H^1(\Omega)]^d}}(Bv, \sigma)_{L_\sigma} \lesssim \|\pi\sigma\|_{L_\sigma} + \|B^\dagger\sigma\|_{L_u},\end{aligned}$$

whence (A3F). To prove (A3G), observe that $\|z^p\|_{L_p} \lesssim \|K^{pp}z^p\|_{L_p} + \langle p\rangle_\Omega$ and that $\langle p\rangle_\Omega \lesssim (Kz, z)_L^{1/2}$. Finally, assumption (A3H) is evident. ∎

4.3.2. Three-field DG approximation. Let $\eta_1 > 0$, $\eta_2 > 0$, $\eta_3 > 0$ (these parameters can vary from face to face) and

$$M_F^{uu}(v) = \eta_1 h_F^{-1} v, \qquad S_F^{uu}(v) = \eta_2 h_F^{-1} v, \qquad S_F^{pp}(q) = \eta_3 h_F q, \qquad R_F^{uu} \equiv 0. \tag{4.21}$$

Clearly, assumptions (DG3B)–(DG3G) hold. Other choices can be considered for M_F^{uu}, S_F^{uu}, and S_F^{pp}; see, e.g., [10, 9] for a similar DG method to approximate the Stokes and the Oseen equations.

PROPOSITION 4.5. *The discrete assumptions* (3.4)*,* (4.7)*,* (4.8)*,* (4.9)*, and* (4.10) *hold.*

Proof. The discrete Poincaré inequality (3.4) has already been shown to hold in section 3.4. Furthermore, assumptions (4.8), (4.9), and (4.10) are evident. It remains to prove (4.7). Let $z_h \in W_h$ be such that $z_h^\sigma \neq \pi z_h^\sigma$. Proceeding as in the proof of (A3E) in Proposition 4.4, there is $0 \neq v \in [H_0^1(\Omega)]^d$ such that $\nabla\cdot v = \mathrm{tr}(z_h^\sigma - \pi z_h^\sigma)$, $\|v\|_{[H^1(\Omega)]^d} \lesssim \|z_h^\sigma - \pi z_h^\sigma\|_{L_\sigma}$, and

$$\|z_h^\sigma - \pi z_h^\sigma\|_{L_\sigma} \lesssim \|\pi z_h^\sigma\|_{L_\sigma} + \tfrac{1}{\|v\|_{[H^1(\Omega)]^d}}(Bv, z_h^\sigma)_{L_\sigma}.$$

Since $v \in [H_0^1(\Omega)]^d$, integration by parts yields

$$(Bv, z_h^\sigma \pi z_h^\sigma)_{L_\sigma} = -(v, B_h^\dagger z_h^\sigma)_{L_u} + 2 \sum_{F \in \mathcal{F}_h^{\mathrm{i}}} (\{\mathcal{D}^{u\sigma} z_h^\sigma\}, v)_{L_u,F}.$$

Let v_h be the L_u-orthogonal projection of v onto U_h. Then

$$\begin{aligned}(Bv, z_h^\sigma)_{L_\sigma} = & -(v - v_h, B_h^\dagger z_h^\sigma)_{L_u} + 2 \sum_{F \in \mathcal{F}_h^{\mathrm{i}}} (\{\mathcal{D}^{u\sigma} z_h^\sigma\}, \{v - v_h\})_{L_u,F} \\ & -(v_h, B_h^\dagger z_h^\sigma)_{L_u} + 2 \sum_{F \in \mathcal{F}_h^{\mathrm{i}}} (\{\mathcal{D}^{u\sigma} z_h^\sigma\}, \{v_h\})_{L_u,F} := T_1 + T_2 + T_3 + T_4.\end{aligned}$$

By construction, $T_1 = 0$ and by definition $T_3 + T_4 = -\mathbb{B}_h(z_h^\sigma, v_h)$. Moreover, adding and subtracting πz_h^σ in T_2, using trace and inverse inequalities and the facts that $p_\sigma \le p_u + 1$ and $\|v - v_h\|_{L_u,\partial K} \lesssim h_K^{\frac{1}{2}} \|v\|_{[H^1(K)]^d}$ for all $K \in \mathcal{T}_h$, leads to

$$|T_2| \lesssim \left(\|\pi z_h^\sigma\|_{L_\sigma}^2 + \sum_{F \in \mathcal{F}_h^{\mathrm{i}}} h_F \|[\![z_h^\sigma - \pi z_h^\sigma]\!]\|_{L_\sigma,F}^2 \right)^{\frac{1}{2}} \|v\|_{[H^1(\Omega)]^d},$$

whence (4.7). ∎

5. Singular perturbation analysis. The situation we want to analyze in this section is that of two-field Friedrichs' systems where the off-diagonal term $\mathcal{B}^k$ coupling the σ- and u-components and the diagonal term $\mathcal{C}^k$ may have different magnitude. The model situation we have in mind is that of an elliptic/hyperbolic-type change. We show that this situation is properly accounted for by appropriately defining (tuning) the stabilizing parameters controlling the interface jumps of z_h^u.

To avoid irrelevant technicalities, we henceforth assume that (A1)–(A5) hold, i.e., full L-coercivity holds. Hypothesis (A3) can be replaced by the weaker hypotheses introduced in section 3, but these developments are omitted for brevity. The singular perturbation analysis for the three-field DG approximation will be reported elsewhere.

5.1. The setting. Let $1 \ge \epsilon > 0$ be a positive real number. The setting of section 2.2 is modified by considering the following two-field structure:

$$K = \left[\begin{array}{c|c} K^{\sigma\sigma} & K^{\sigma u} \\ \hline K^{u\sigma} & K^{uu} \end{array}\right], \qquad \mathcal{A}^k = \left[\begin{array}{c|c} 0 & \epsilon^{\frac{1}{2}} \mathcal{B}^k \\ \hline \epsilon^{\frac{1}{2}} [\mathcal{B}^k]^t & \mathcal{C}^k \end{array}\right], \tag{5.1}$$

where it is assumed that all the blocks of the operator K as well as the fields $\mathcal{B}^k$ and $\mathcal{C}^k$ are independent of the parameter ϵ.

Owing to (5.1), the definitions (2.11) and (2.12) are now replaced by

$$\mathcal{D} = \left[\begin{array}{c|c} 0 & \epsilon^{\frac{1}{2}} \mathcal{D}^{\sigma u} \\ \hline \epsilon^{\frac{1}{2}} \mathcal{D}^{u\sigma} & \mathcal{D}^{uu} \end{array}\right], \qquad \mathcal{M} = \left[\begin{array}{c|c} 0 & -\alpha\epsilon^{\frac{1}{2}} \mathcal{D}^{\sigma u} \\ \hline \alpha\epsilon^{\frac{1}{2}} \mathcal{D}^{u\sigma} & \mathcal{M}^{uu} \end{array}\right]. \tag{5.2}$$

The discrete problem we consider is (2.16) with the bilinear form a_h still defined by (2.15). As in section 2.5, we assume that p_u is a *positive* integer and that $p_u - 1 \le p_\sigma \le p_u$.

Henceforth the notation $\xi \lesssim \zeta$ now means that there is a positive c, *independent of h and ϵ*, such that $\xi \le c\zeta$.

5.2. Design of the boundary and jump operators. To avoid unnecessary technicalities we assume that the user-dependent operator R_F^{uu} is zero. Everything that is said hereafter extends to IP-like methods provided the assumptions (DG2E)–(DG2F) are localized. The details are left to the reader. To account for the presence of ϵ, we modify the design conditions (DG2A)–(DG2D) for the operators M_F and S_F as follows:

(DG2$_\epsilon$A) $$M_F = \left[\begin{array}{c:c} 0 & -\alpha\epsilon^{\frac12}\mathcal{D}^{\sigma u} \\ \hdashline \alpha\epsilon^{\frac12}\mathcal{D}^{u\sigma} & M_F^{uu} \end{array}\right],\ \alpha\in\{-1,+1\},\qquad S_F = \left[\begin{array}{c:c} 0 & 0 \\ \hdashline 0 & S_F^{uu} \end{array}\right];$$

(DG2$_\epsilon$B) $$\text{if } \alpha=+1,\ \begin{cases} M_F^{uu} = (M_F^{uu})^* \quad\text{and}\quad \mathrm{Ker}(\mathcal{D}^{\sigma u})\subset\mathrm{Ker}(M_F^{uu}-\mathcal{D}^{uu}),\\ \epsilon(\mathcal{D}^{u\sigma}\mathcal{D}^{\sigma u})^{\frac12} + h_F|\mathcal{D}^{uu}| \lesssim h_F M_F^{uu} \lesssim \theta_F\mathcal{I}_{m_u}; \end{cases}$$

(DG2$_\epsilon$C) $$\text{if } \alpha=-1,\ M_F^{uu}(v) = \mathcal{M}^{uu}v \text{ and } |\mathcal{D}^{uu}| \lesssim \mathcal{M}^{uu} \lesssim \mathcal{I}_{m_u};$$

(DG2$_\epsilon$D) $$S_F^{uu} = (S_F^{uu})^* \text{ and } \epsilon(\mathcal{D}^{u\sigma}\mathcal{D}^{\sigma u})^{\frac12} + h_F|\mathcal{D}^{uu}| \lesssim h_F S_F^{uu} \lesssim \theta_F\mathcal{I}_{m_u},$$

where we have set

$$\theta_S = \max(\epsilon, h_S)\qquad \forall S\in\mathcal{F}_h\cup\mathcal{T}_h. \tag{5.3}$$

If $\epsilon \ll h$, (DG2$_\epsilon$D) amounts to $|\mathcal{D}^{uu}| \lesssim S_F^{uu} \lesssim \mathcal{I}_{m_u}$, that is, assumption (DG1E) for one-field Friedrichs' systems is recovered for the (uu)-blocks. If $\epsilon\sim 1$, (DG2$_\epsilon$D) leads to assumption (DG2D) for two-field Friedrichs' systems concerning the control on $(\mathcal{D}^{u\sigma}\mathcal{D}^{\sigma u})^{\frac12}$ and leads to a slightly stronger control on $\mathcal{D}^{uu}$, namely, $|\mathcal{D}^{uu}| \lesssim M_F^{uu}$. The reason for this difference is that in the present analysis, we aim at obtaining a sharper convergence result for the u-component.

5.3. Convergence analysis. For all $z\in W(h)$, we introduce the following norms:

$$\|z\|_{h\epsilon,A}^2 = \|z\|_L^2 + |z^u|_J^2 + |z^u|_M^2 + \|\epsilon^{\frac12}B_h z^u\|_{L_\sigma}^2 + \|\mathfrak{h}^{\frac12}C_h z^u\|_{L_u}^2, \tag{5.4}$$

$$\|z\|_{h\epsilon,1}^2 = \|z\|_{h\epsilon,A}^2 + \sum_{K\in\mathcal{T}_h}[\theta_K h_K^{-1}(h_K^{-1}\|z^u\|_{L_u,K}^2 + \|z^u\|_{L_u,\partial K}^2) + h_K\|z^\sigma\|_{L_\sigma,\partial K}^2]. \tag{5.5}$$

We denote by $\mathcal{T}_h^+$ the set of mesh cells K such that $h_K\ge\epsilon$. We also denote by $\mathcal{F}_h^{i+}$ the set of faces F such that $\max_{K\in\mathcal{T}(F)}h_K\ge\epsilon$; observe that $h_F\gtrsim\epsilon$ whenever $F\in\mathcal{F}_h^{i+}$. The same definition applies for $\mathcal{F}_h^{\partial+}$.

LEMMA 5.1. *Assume $\mathcal{B}^k\in[\mathfrak{C}^{0,1}(K)]^{m_\sigma,m_u}$ and $\mathcal{C}^k\in[\mathfrak{C}^{0,\frac12}(K)]^{m_u,m_u}$ for all $K\in\mathcal{T}_h$ and all $1\le k\le d$, and that*

$$\forall z_h\in W_h,\quad \|C_h z_h^u\|_{L_u} \lesssim \|B_h z_h^u\|_{L_\sigma} + \|z_h^u\|_{L_u}. \tag{5.6}$$

Then

$$\forall z_h\in W_h,\qquad \|z_h\|_{h\epsilon,A} \lesssim \sup_{y_h\in W_h\setminus\{0\}}\frac{a_h(z_h,y_h)}{\|y_h\|_{h\epsilon,A}}. \tag{5.7}$$

Proof. Let $z_h\in W_h$ and set $\mathbb{S} = \sup_{y_h\in W_h\setminus\{0\}}\frac{a_h(z_h,y_h)}{\|y_h\|_{h\epsilon,A}}$.

(1) Using the definition of a_h together with (DG2$_\epsilon$A) and (A3) yields

$$\|z_h\|_L^2 + |z_h^u|_J^2 + |z_h^u|_M^2 \lesssim a_h(z_h,z_h) \le \mathbb{S}\,\|z_h\|_{h\epsilon,A}. \tag{5.8}$$

(2) Define the field π_h^σ such that for all $K \in \mathcal{T}_h$, $\pi_h^\sigma|_K = \sum_{k=1}^d \epsilon^{\frac{1}{2}} \overline{\mathcal{B}_K^k} \partial_k z_h^u$, where $\overline{\mathcal{B}_K^k}$ denotes the mean-value of $\mathcal{B}^k$ over K. Owing to the regularity of the $\mathcal{B}^k$'s, a standard inverse inequality, and the fact that $\epsilon \le 1$,

$$\|(\pi_h^\sigma, 0)\|_{h\epsilon,A} = \|\pi_h^\sigma\|_{L_\sigma} \lesssim \|\epsilon^{\frac{1}{2}} B_h z_h^u\|_{L_\sigma} + \epsilon^{\frac{1}{2}} \|z_h^u\|_{L_u} \lesssim \|z_h\|_{h\epsilon,A}. \tag{5.9}$$

From the definition of a_h and (DG2_ϵA), it follows that

$$\begin{aligned}\|\epsilon^{\frac{1}{2}} B_h z_h^u\|_{L_\sigma}^2 = a_h(z_h, (\pi_h^\sigma, 0)) + (\epsilon^{\frac{1}{2}} B_h z_h^u, \epsilon^{\frac{1}{2}} B_h z_h^u - \pi_h^\sigma)_{L_\sigma} \\ - (K^{\sigma\sigma} z_h^\sigma + K^{\sigma u} z_h^u, \pi_h^\sigma)_{L_\sigma} + \sum_{F\in\mathcal{F}_h^\partial} \tfrac{\alpha+1}{2} \epsilon^{\frac{1}{2}} (\mathcal{D}^{\sigma u} z_h^u, \pi_h^\sigma)_{L_\sigma,F} \\ + \sum_{F\in\mathcal{F}_h^{\mathrm{i}}} 2\epsilon^{\frac{1}{2}} (\{\mathcal{D}^{\sigma u} z_h^u\}, \{\pi_h^\sigma\})_{L_\sigma,F}.\end{aligned}$$

Then proceeding as in the proof of [16, lemma 5.5] in Part II yields

$$\|\epsilon^{\frac{1}{2}} B_h z_h^u\|_{L_\sigma}^2 \lesssim \mathbb{S} \|z_h\|_{h\epsilon,A}. \tag{5.10}$$

(3) Let $\overline{\mathcal{C}_K^k}$ denote the mean-value of $\mathcal{C}^k$ over K and define the field π_h^u such that $\pi_h^u|_K = \sum_{k=1}^d h_K \overline{\mathcal{C}_K^k} \partial_k z_h^u$ if $h_K \ge \epsilon$, and $\pi_h^u|_K = 0$ otherwise. Owing to the regularity of the $\mathcal{C}^k$'s and a standard inverse inequality,

$$h_K^{-\frac{1}{2}} \|\pi_h^u\|_{L_u,K} \lesssim h_K^{\frac{1}{2}} \|C_h z_h^u\|_{L_u,K} + \|z_h^u\|_{L_u,K}, \tag{5.11}$$

whence it is inferred, using inverse inequalities, the fact that $h_K \ge \epsilon$ in the support of π_h^u, and the upper bounds in (DG2_ϵB)–(DG2_ϵD), that $\|(0, \pi_h^u)\|_{h\epsilon,A} \lesssim \|z_h\|_{h\epsilon,A}$. Set $\mathtt{C} = \sum_{K\in\mathcal{T}_h^+} h_K \|C_h z_h^u\|_{L_u,K}^2$. From the definition of a_h, it follows that

$$\begin{aligned}\mathtt{C} = a_h(z_h, (0, \pi_h^u)) - (K^{u\sigma} z_h^\sigma + K^{uu} z_h^u, \pi_h^u)_{L_u} \\ - \sum_{K\in\mathcal{T}_h^+} (\epsilon^{\frac{1}{2}} B_h^\dagger z_h^\sigma, \pi_h^u)_{L_u,K} + \sum_{K\in\mathcal{T}_h^+} (C_h z_h^u, h_K C_h z_h^u - \pi_h^u)_{L_u,K} \\ - \sum_{F\in\mathcal{F}_h^{\partial+}} \tfrac{\alpha-1}{2} \epsilon^{\frac{1}{2}} (\mathcal{D}^{u\sigma} z_h^\sigma, \pi_h^u)_{L_u,F} - \sum_{F\in\mathcal{F}_h^{\partial+}} \tfrac{1}{2} (M_F^{uu}(z_h^u) - \mathcal{D}^{uu} z_h^u, \pi_h^u)_{L_u,F} \\ + \sum_{F\in\mathcal{F}_h^{i+}} 2\epsilon^{\frac{1}{2}} (\{\mathcal{D}^{u\sigma} z_h^\sigma\}, \{\pi_h^u\})_{L_u,F} + \sum_{F\in\mathcal{F}_h^{i+}} 2(\{\mathcal{D}^{uu} z_h^u\}, \{\pi_h^u\})_{L_u,F} \\ - \sum_{F\in\mathcal{F}_h^{i+}} (S_F^{uu}([\![z_h^u]\!]), [\![\pi_h^u]\!])_{L_u,F} = a_h(z_h, (0, \pi_h^u)) + R_1 + \cdots + R_8.\end{aligned}$$

Let us estimate the remainder terms R_i, $1 \le i \le 8$, on the right-hand side. Clearly, $|R_1| \lesssim \|z_h\|_L \|\pi_h^u\|_{L_u}$. Furthermore, using an inverse inequality and the fact that $\epsilon \le h_K$ for $K \in \mathcal{T}_h^+$,

$$|R_2| \lesssim \sum_{K\in\mathcal{T}_h^+} \epsilon^{\frac{1}{2}} h_K^{-1} \|z_h^\sigma\|_{L_\sigma,K} \|\pi_h^u\|_{L_u,K} \lesssim \sum_{K\in\mathcal{T}_h^+} \|z_h^\sigma\|_{L_\sigma,K} h_K^{-\frac{1}{2}} \|\pi_h^u\|_{L_u,K},$$

and $|R_3| \lesssim \sum_{K\in\mathcal{T}_h^+} h_K^{\frac12}\|C_h z_h^u\|_{L_u,K}\|z_h^u\|_{L_u,K}$. If $\alpha = +1$, then $R_4 = 0$, while if $\alpha = -1$, then

$$\begin{aligned}|R_4| &\lesssim \sum_{F\in\mathcal{F}_h^{\partial+}} \epsilon^{\frac12} h_F^{-\frac12}\|z_h^\sigma\|_{L_\sigma,\mathcal{T}(F)} h_F^{-\frac12}\|\pi_h^u\|_{L_u,\mathcal{T}(F)}\\ &\lesssim \sum_{F\in\mathcal{F}_h^{\partial+}} \|z_h^\sigma\|_{L_\sigma,\mathcal{T}(F)} h_F^{-\frac12}\|\pi_h^u\|_{L_u,\mathcal{T}(F)}.\end{aligned}$$

Moreover, since $|\mathcal{D}^{uu}| \lesssim M_F^{uu} \lesssim \mathcal{I}_{m_u}$ for all $F \in \mathcal{F}_h^{\partial+}$ in both cases for α,

$$|R_5| \lesssim \sum_{F\in\mathcal{F}_h^{\partial+}} |z_h^u|_{M,F}(|\pi_h^u|_{M,F} + \|\pi_h^u\|_{L_u,F}) \lesssim \sum_{F\in\mathcal{F}_h^{\partial+}} |z_h^u|_{M,F} h_F^{-\frac12}\|\pi_h^u\|_{L_u,\mathcal{T}(F)}.$$

Similarly,

$$|R_6| + |R_7| \lesssim \sum_{F\in\mathcal{F}_h^{i+}} (\|z_h^\sigma\|_{L_\sigma,\mathcal{T}(F)} + |z_h^u|_{J,F}) h_F^{-\frac12}\|\pi_h^u\|_{L_u,\mathcal{T}(F)},$$

and $|R_8| \le \sum_{F\in\mathcal{F}_h^{i+}} |z_h^u|_{J,F}|\pi_h^u|_{J,F} \lesssim \sum_{F\in\mathcal{F}_h^{i+}} |z_h^u|_{J,F} h_F^{-\frac12}\|\pi_h^u\|_{L_u,\mathcal{T}(F)}$. Collecting the above bounds and using (5.8) and (5.11), we deduce $|R_1|+\cdots+|R_8| \lesssim \gamma\mathfrak{C} + a_h(z_h,z_h)$, where γ can be chosen as small as needed. Hence, $\mathfrak{C} \lesssim \mathbb{S}\,\|z_h\|_{h\epsilon,A}$, and using (5.6) and (5.10) leads to the same bound for $\|\mathfrak{h}^{\frac12} C_h z^u\|_{L_u}^2$.

(4) Collecting the above bounds yields $\|z_h\|_{h\epsilon,A}^2 \lesssim \mathbb{S}\,\|z_h\|_{h\epsilon,A}$ and hence (5.7). ∎

LEMMA 5.2. *The following holds:*

$$\forall (z,y_h) \in W(h)\times W_h, \quad a_h(z,y_h) \lesssim \|z\|_{h\epsilon,1}\|y_h\|_{h\epsilon,A}. \tag{5.12}$$

Proof. Use integration by parts to infer

$$\begin{aligned}a_h(z,y_h) = &\sum_{K\in\mathcal{T}_h} (z,\tilde{T}y_h)_{L,K} + \sum_{F\in\mathcal{F}_h^\partial} \tfrac12 (M_F(z) + \mathcal{D}z, y_h)_{L,F}\\ &+ \sum_{F\in\mathcal{F}_h^{\mathrm{i}}} \tfrac12 (\llbracket \mathcal{D}z\rrbracket, \llbracket y_h\rrbracket)_{L,F} + \sum_{F\in\mathcal{F}_h^{\mathrm{i}}} (S_F^{uu}(\llbracket z^u\rrbracket), \llbracket y_h^u\rrbracket)_{L_u,F}.\end{aligned}$$

Let R_1 to R_4 be the four terms on the right-hand side. Observe that

$$\begin{aligned}(z,\tilde{T}y_h)_{L,K} &\lesssim \|z\|_{L,K}\|y_h\|_{L,K} + \|z^\sigma\|_{L_\sigma,K}\epsilon^{\frac12}\|B_h y_h^u\|_{L_\sigma,K}\\ &\quad + \|z^u\|_{L_u,K}\epsilon^{\frac12}\|B_h^\dagger y_h^\sigma\|_{L_u,K} + \|z^u\|_{L_u,K}\|C_h y_h^u\|_{L_u,K}\\ &\lesssim \|z\|_{L,K}\|y_h\|_{L,K} + \|z^\sigma\|_{L_\sigma,K}\epsilon^{\frac12}\|B_h y_h^u\|_{L_\sigma,K}\\ &\quad + \epsilon^{\frac12} h_K^{-1}\|z^u\|_{L_u,K}\|y_h^\sigma\|_{L_\sigma,K} + h_K^{-\frac12}\|z^u\|_{L_u,K} h_K^{\frac12}\|C_h y_h^u\|_{L_u,K}.\end{aligned}$$

Hence, $|R_1| \lesssim \|z\|_{h\epsilon,1}\|y_h\|_{h\epsilon,A}$. Furthermore, if $\alpha = +1$,

$$\begin{aligned}|R_2| &\le \sum_{F\in\mathcal{F}_h^\partial} [|\epsilon^{\frac12}(\mathcal{D}^{u\sigma}z^\sigma, y_h^u)_{L_u,F}| + \tfrac12 |(M_F^{uu}(z^u) + \mathcal{D}^{uu}z^u, y_h^u)_{L_u,F}|]\\ &\lesssim \sum_{F\in\mathcal{F}_h^\partial} [\|z^\sigma\|_{L_\sigma,F} h_F^{\frac12}|y_h^u|_{M,F} + \theta_F^{\frac12} h_F^{-\frac12}\|z^u\|_{L_u,F}|y_h^u|_{M,F}],\end{aligned}$$

while if $\alpha = -1$, $|R_2| \lesssim \sum_{F\in\mathcal{F}_h^\partial}[\epsilon^{\frac12}\|z^u\|_{L_u,F}h_F^{-\frac12}\|y_h^\sigma\|_{L_\sigma,\mathcal{T}(F)} + \|z^u\|_{L_u,F}|y_h^u|_{M,F}]$. Hence, in both cases, $|R_2| \lesssim \|z\|_{h\epsilon,1}\|y_h\|_{h\epsilon,A}$. Similarly,

$$\epsilon^{\frac12}([\![\mathcal{D}^{\sigma u}z^u]\!],[\![y_h^\sigma]\!])_{L_\sigma,F} \lesssim \epsilon^{\frac12}\|\{z^u\}\|_{L_u,F}h_F^{-\frac12}\|y_h^\sigma\|_{L_\sigma,\mathcal{T}(F)},$$
$$\epsilon^{\frac12}([\![\mathcal{D}^{u\sigma}z^\sigma]\!],[\![y_h^u]\!])_{L_u,F} \lesssim \|\{z^\sigma\}\|_{L_\sigma,F}h_F^{\frac12}|y_h^u|_{J,F},$$
$$([\![\mathcal{D}^{uu}z^u]\!],[\![y_h^u]\!])_{L_u,F} \lesssim \|\{z^u\}\|_{L_u,F}|y_h^u|_{J,F}.$$

Hence, $|R_3| \lesssim \|z\|_{h\epsilon,1}\|y_h\|_{h\epsilon,A}$. Finally, it is clear that $\sum_{F\in\mathcal{F}_h^{\mathrm{i}}}(S_F^{uu}([\![z^u]\!]),[\![y_h^u]\!])_{L_u,F} \le \|z\|_{h\epsilon,1}\|y_h\|_{h\epsilon,A}$, thereby completing the proof. ∎

It is now straightforward to derive the following result.

THEOREM 5.3. *Keep the hypotheses of Lemma* 5.1. *Assume that* $z \in [H^1(\Omega)]^m$. *Then*

$$\|z - z_h\|_{h\epsilon,A} \lesssim \inf_{y_h\in W_h}\|z - y_h\|_{h\epsilon,1}. \tag{5.13}$$

In particular, if $z \in [H^{p_u+1}(\Omega)]^m$,

$$\|z - z_h\|_{h\epsilon,A} \lesssim \theta^{\frac12}h^{p_u}\|z\|_{[H^{p_u+1}(\Omega)]^m}. \tag{5.14}$$

The convergence estimate in Theorem 5.3 is consistent with that from the two-field DG theory when $1 \sim \epsilon \ge h$, and it degenerates into that from the one-field theory for the u-component when $h \ge \epsilon$. Indeed, if $1 \sim \epsilon \ge h$,

$$\|z^u - z_h^u\|_{L_u} + \|B_h(z^u - z_h^u)\|_{L_\sigma} \lesssim h^{p_u}\|z\|_{[H^{p_u+1}(\Omega)]^m}, \tag{5.15}$$

and the L_u-norm error estimate can be improved if elliptic regularity holds, while if $h \ge \epsilon$,

$$\|z^u - z_h^u\|_{L_u} + \|\mathfrak{h}^{\frac12}C_h(z^u - z_h^u)\|_{L_u} \lesssim h^{p_u+\frac12}\|z\|_{[H^{p_u+1}(\Omega)]^m}. \tag{5.16}$$

Remark 5.1. A similar result to that of Theorem 5.3 has been proved in Gopalakrishnan and Kanschat [22, Thm. 5.1] for the advection–diffusion equation (5.17).

5.4. Example: Advection dominated advection–diffusion. Consider the advection–diffusion equation introduced in section 3.3 with a diffusion coefficient $\epsilon > 0$. The PDE $-\epsilon\Delta u + \beta\cdot\nabla u + \mu u = f$ in mixed form becomes

$$\begin{cases} \sigma + \epsilon^{\frac12}\nabla u = 0, \\ \mu u + \epsilon^{\frac12}\nabla\cdot\sigma + \beta\cdot\nabla u = f, \end{cases} \tag{5.17}$$

so that the off-diagonal blocks of $\mathcal{A}^k$ in (3.8) are rescaled by $\epsilon^{\frac12}$, while the operator K is unchanged.

In the case of a Dirichlet boundary condition, the boundary and interface operators can be redesigned to fit the above analysis by modifying (3.16) as follows:

$$M_F^{uu}(v) = \eta_1(|\beta\cdot n_F| + \epsilon h_F^{-1})v, \qquad S_F^{uu}(v) = \eta_2(|\beta\cdot n_F| + \epsilon h_F^{-1})v, \tag{5.18}$$

where n_F is a unit normal vector to F and $\eta_1 > 0$, $\eta_2 > 0$ (these two parameters can vary from face to face). It is easily verified that properties (DG2_ϵB) and (DG2_ϵD) hold.

Assuming $\varrho \ge -\min(\beta\cdot n, 0)$, mixed Robin–Neumann boundary conditions can be enforced by redesigning the boundary and interface operators as follows:

$$M_F^{uu}(v) = (2\varrho + \beta\cdot n)v, \qquad S_F^{uu}(v) = \eta_2(|\beta\cdot n_F| + \epsilon h_F^{-1})v. \tag{5.19}$$

Again, it is easily verified that properties (DG2_ϵC) and (DG2_ϵD) hold.

6. Conclusion. We have analyzed various DG methods in Parts I, II, and III. We have attempted to give a unified analysis for all of these methods. Following the seminal ideas of Lesaint and Raviart [25, 26], we have shown that the framework of symmetric positive Friedrichs' systems is the natural setting for this theory insofar as boundary conditions can be enforced weakly for all of these systems. The first building block of the theory is the bilinear form (2.4) along with the weak formulation (2.6). All the DG methods that we have analyzed can be put into the unified bilinear form (2.15) and the unified formulation (2.16). The differences between all of these methods reside solely in the design of the boundary and interface operators.

The method described in Part I is the most robust in the sense that it is essentially independent of the type of the PDE (which is the main argument for our working with Friedrichs' systems). The price for robustness is the slightly suboptimal convergence rate $\mathcal{O}(h^{p+/2})$ in the L^2-norm. In Part II, two components of the unknown are identified, say (z^σ, z^u), and the generic DG method is set so that the z^σ unknown can be locally eliminated on each mesh cell. The underlying model is that of elliptic equations. Compared to the method of Part I, the convergence rate in the L^2-norm for the z^u unknown is now optimal, i.e., $\mathcal{O}(h^{p+1})$, the downside being a slight loss of robustness with respect to type change, which is addressed in Part III, section 5. Part III revisits the results of Part II in two aspects: (i) weakening of the L^2-coercivity assumption on which Part II is based; (ii) questions regarding robustness with respect to type change.

REFERENCES

[1] D. N. Arnold, *An interior penalty finite element method with discontinuous elements*, SIAM J. Numer. Anal., 19 (1982), pp. 742–760.

[2] D. N. Arnold, F. Brezzi, B. Cockburn, and L. D. Marini, *Unified analysis of discontinuous Galerkin methods for elliptic problems*, SIAM J. Numer. Anal., 39 (2002), pp. 1749–1779.

[3] F. P. T. Baaijens, *Mixed finite element method for viscoelastic flow analysis: A review*, J. Non-Newtonian Fluid Mech., 79 (1998), pp. 361–386.

[4] G. A. Baker, *Finite element methods for elliptic equations using nonconforming elements*, Math. Comp., 31 (1977), pp. 45–59.

[5] F. Bassi and S. Rebay, *A high-order accurate discontinuous finite element method for the numerical solution of the compressible Navier-Stokes equations*, J. Comput. Phys., 131 (1997), pp. 267–279.

[6] S. C. Brenner, *Poincaré–Friedrichs inequalities for piecewise H^1 functions*, SIAM J. Numer. Anal., 41 (2003), pp. 306–324.

[7] S. C. Brenner, *Korn's inequalities for piecewise H^1 vector fields*, Math. Comp., 73 (2004), pp. 1067–1087.

[8] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer-Verlag, New York, 1991.

[9] B. Cockburn, G. Kanschat, and D. Schötzau, *The local discontinuous Galerkin method for the Oseen equations*, Math. Comp., 73 (2003), pp. 569–593.

[10] B. Cockburn, G. Kanschat, D. Schötzau, and C. Schwab, *Local discontinuous Galerkin methods for the Stokes system*, SIAM J. Numer. Anal., 40 (2002), pp. 319–343.

[11] B. Cockburn and C.-W. Shu, *The local discontinuous Galerkin method for time-dependent convection-diffusion systems*, SIAM J. Numer. Anal., 35 (1998), pp. 2440–2463.

[12] J. Douglas, Jr., and T. Dupont, *Interior penalty procedures for elliptic and parabolic Galerkin methods*, in Computing Methods in Applied Sciences Lecture Notes in Phys. 58, Springer-Verlag, Berlin, 1976, pp. 207–216.

[13] A. V. C. Duarte, E. G. D. do Carmo, and F. A. Rochinha, *Consistent discontinuous finite elements in elastodynamics*, Comput. Methods Appl. Mech. Engrg., 190 (2000), pp. 193–223.

[14] A. Ern and J.-L. Guermond, *Theory and Practice of Finite Elements*, Appl. Math. Sci. 159, Springer-Verlag, New York, 2004.

[15] A. Ern and J.-L. Guermond, *Discontinuous Galerkin methods for Friedrichs' systems.* I. *General theory*, SIAM J. Numer. Anal., 44 (2006), pp. 753–778.

[16] A. Ern and J.-L. Guermond, *Discontinuous Galerkin methods for Friedrichs' systems.* Part II. *Second-order elliptic PDEs*, SIAM J. Numer. Anal., 44 (2006), pp. 2363–2388.
[17] A. Ern, J.-L. Guermond, and G. Caplain, *An intrinsic criterion for the bijectivity of Hilbert operators related to Friedrichs' systems*, Comm. Partial Differential Equations, 32 (2007), pp. 317–341.
[18] M. Fortin, R. Guénette, and R. Pierre, *Numerical analysis of the modified EVSS method*, Comput. Methods Appl. Mech. Engrg., 143 (1997), pp. 79–95.
[19] M. Fortin and R. Pierre, *On the convergence of the mixed method of Crochet and Marchal for viscoelastic flows*, Comput. Methods Appl. Mech. Engrg., 73 (1989), pp. 341–350.
[20] L. P. Franca and R. Stenberg, *Error analysis of some Galerkin least squares methods for the elasticity equations*, SIAM J. Numer. Anal., 28 (1991), pp. 1680–1697.
[21] K. O. Friedrichs, *Symmetric positive linear differential equations*, Comm. Pure Appl. Math., 11 (1958), pp. 333–418.
[22] J. Gopalakrishnan and G. Kanschat, *A multilevel discontinuous Galerkin method*, Numer. Math., 95 (2003), pp. 527–550.
[23] P. Hansbo and M. G. Larson, *Discontinuous Galerkin methods for incompressible and nearly incompressible elasticity by Nitsche's method*, Comput. Methods Appl. Mech. Engrg., 191 (2002), pp. 1895–1908.
[24] C. Johnson, U. Nävert, and J. Pitkäranta, *Finite element methods for linear hyperbolic equations*, Comput. Methods Appl. Mech. Engrg., 45 (1984), pp. 285–312.
[25] P. Lesaint, *Sur la résolution des systèmes hyperboliques du premier ordre par des méthodes d'éléments finis*, Ph.D. thesis, University of Paris VI, 1975.
[26] P. Lesaint and P.-A. Raviart, *On a finite element method for solving the neutron transport equation*, in Mathematical Aspects of Finite Elements in Partial Differential Equations, Publication No. 33. Math. Res. Center, University of Wisconsin–Madison, Academic Press, New York, 1974, pp. 89–123.
[27] J. Nitsche, *Über ein Variationsprinzip zur Lösung von Dirichlet-Problemen bei Verwendung von Teilräumen, die keinen Randbedingungen unterworfen sind*, Abh. Math. Sem. Univ. Hamburg, 36 (1971), pp. 9–15.
[28] Ch. Schwab and M. Suri, *Mixed hp finite element methods for Stokes and non-Newtonian flow*, Comput. Methods Appl. Mech. Engrg., 175 (1999), pp. 217–241.
[29] M. F. Wheeler, *An elliptic collocation-finite element method with interior penalties*, SIAM J. Numer. Anal., 15 (1978), pp. 152–161.

SIAM J. NUMER. ANAL.
Vol. 46, No. 2, pp. 805–831
© 2008 Society for Industrial and Applied Mathematics

DISCONTINUOUS GALERKIN METHODS FOR ANISOTROPIC SEMIDEFINITE DIFFUSION WITH ADVECTION*

DANIELE A. DI PIETRO†‡, ALEXANDRE ERN†, AND JEAN-LUC GUERMOND§

Abstract. We construct and analyze a discontinuous Galerkin method to solve advection-diffusion-reaction PDEs with anisotropic and semidefinite diffusion. The method is designed to automatically detect the so-called elliptic/hyperbolic interface on fitted meshes. The key idea is to use consistent weighted average and jump operators. Optimal estimates in the broken graph norm are proven. These are consistent with well-known results when the problem is either hyperbolic or uniformly elliptic. The theoretical results are supported by numerical evidence.

Key words. discontinuous Galerkin, advection-diffusion-reaction, discontinuous coefficients, anisotropic diffusion, coupled elliptic-hyperbolic, weighted averages

AMS subject classifications. 65N30, 65M60, 35G15

DOI. 10.1137/060676106

1. Introduction.

Discontinuous Galerkin (DG) methods were originally introduced to solve transport equations in [23, 24, 25] and later extended to problems involving second-order elliptic operators in [1, 4, 26]. For many years, the development and analysis of DG methods have followed two somewhat parallel routes according to the hyperbolic or elliptic nature of the problem at hand. A unifying viewpoint has recently been proposed in a series of papers [10, 11, 13], where the authors rely on the Friedrichs framework originally proposed in [17] to perform an abstract analysis valid for a variety of (linear) PDE systems.

The goal of the present work is to merge even further the hyperbolic and elliptic points of view by considering advection-diffusion-reaction problems with discontinuous, anisotropic, and semidefinite diffusivity (or, using the terminology from [22], scalar-valued second-order PDEs with nonnegative characteristic form). One major difficulty associated with this class of problems is that only the conservative flux is continuous and, at variance with uniformly elliptic problems, the scalar field can be discontinuous across hyperbolic/elliptic interfaces (see (2.4) for the definition of these interfaces). The literature in numerical analysis dedicated to this topic is scarce. This issue has been investigated in one space dimension by Gastaldi and Quarteroni in [18], where interface conditions are derived using asymptotic analysis. In [22], Houston, Schwab, and Süli propose and analyze a DG method for PDEs with nonnegative characteristic form in higher space dimensions. They treat hyperbolic/elliptic interfaces in a numerical example at the end of the paper (see [22, section 6.4]). Another example can be found in [14]. Both proposed techniques are ad hoc in the sense that they require removing a priori some penalty terms at the hyperbolic/elliptic interface. Our

*Received by the editors November 28, 2006; accepted for publication (in revised form) October 8, 2007; published electronically February 8, 2008. This work was partially supported by the GdR MoMaS (CNRS-2439, ANDRA, BRGM, CEA, EDF, IRSN).
http://www.siam.org/journals/sinum/46-2/67610.html

†CERMICS, École des Ponts, Université Paris-Est, 77455 Marne la Vallée Cedex 2, France (ern@cermics.enpc.fr).

‡Current address: Direction Technologie, Informatique, et Mathématiques Appliquées, Institut Français du Pétrole, 92852 Rueil Malmaison Cedex, France (daniele-antonio.di-pietro@ifp.fr).

§Department of Mathematics, Texas A&M University, College Station, TX 77843-3368, and LIMSI (CNRS-UPR 3251), BP 133, 91403, Orsay, France (guermond@math.tamu.edu).

objective is to go beyond [14, 18, 22] by proposing a DG method that automatically detects the so-called hyperbolic/elliptic interface under the mild assumption that the mesh fits the discontinuities of the diffusion tensor.

The material is organized as follows. We introduce the model problem in section 2. It is a multidimensional scalar-valued second-order PDE with nonnegative characteristic from. We derive the multidimensional counterpart of the one-dimensional interface condition introduced in [18]. We reformulate the problem as a symmetric Friedrichs system (in the spirit of [10, 13, 17]) by identifying appropriate interface and boundary operators. A well-posedness result is proven under some assumptions. In section 3 we focus our attention on the discrete problem. We propose a DG approximation based on the weak formulation of the continuous problem with boundary and interface conditions enforced weakly. The discrete bilinear form is designed so that the correct interface conditions are automatically enforced on fitted meshes without identifying a priori the hyperbolic/elliptic interfaces. The basic requirements on the bilinear form are that it be strongly consistent, continuous, and (in the spirit of Friedrichs) L^2-coercive. The convergence of the method is proven in section 4, and the main results are stated in Theorems 4.5 and 4.7. The error estimates are optimal in the broken graph norm and compatible with those presented in [10, 11] when the problem is either hyperbolic or uniformly elliptic. Implementation issues are addressed in section 5, and variants of the method are discussed. Section 6 is devoted to numerical experiments illustrating the performance of the proposed method. Concluding remarks are contained in section 7.

2. The continuous problem. In this section we introduce the model problem, and we recast it as a first-order PDE system endowed with a Friedrichs-like structure. We then derive a weak formulation where boundary and interface conditions are enforced naturally. This formulation will be the starting point for the design of the DG method constructed in section 3.

2.1. The PDE setting. Let $\Omega \subset \mathbb{R}^d$ be a bounded, open, and connected Lipschitz domain with boundary $\partial\Omega$ and outward normal n. The problem investigated in the present work consists of the following scalar-valued PDE:

$$\nabla\cdot(-\nu\nabla u + \beta u) + \mu u = f, \tag{2.1}$$

with data $f \in L^2(\Omega)$. Boundary conditions are specified later in this section. The following assumptions are made on the coefficients:

(i) $\nu \in [L^\infty(\Omega)]^{d,d}$ is a symmetric positive *semi*definite tensor field, meaning that for all $r \in \mathbb{R}^d$ and for a.e. (almost every) $x \in \Omega$, $r^t\nu(x)r \geq 0$;

(ii) $\beta \in [\mathfrak{C}^1(\overline{\Omega})]^d$ and $\mu \in L^\infty(\Omega)$;

(iii) there is some $\mu_0 > 0$ so that $\mu + \frac{1}{2}\nabla\cdot\beta \geq \mu_0$ for a.e. $x \in \Omega$.

Throughout the rest of this work, the symbols $\lesssim$ and $\gtrsim$ are used for inequalities that hold up to a real positive multiplicative constant that is independent of ν (and discretization parameters like the meshsize) but may depend on β and μ (and regularity parameters of the mesh family considered later on).

We finally assume that the field ν is discontinuous and that we know the location of the discontinuities. More precisely, we assume that we are given a partition of Ω into Lipschitz connected subdomains $P_\Omega \stackrel{\text{def}}{=} \{\Omega_i\}_{i=1}^N$ so that the following holds:

(iv) ν is piecewise constant on the partition P_Ω, and the problem is normalized so that $\|\nu\|_{[L^\infty(\Omega)]^{d,d}} \leq 1$.

We introduce the symbol Γ to denote the union of the inner boundaries of the subdomains Ω_i, i.e.,

$$\Gamma \stackrel{\text{def}}{=} \{x \in \Omega;\ \exists i_1, i_2 \in \{1, \dots, N\},\ i_1 \neq i_2,\ x \in \partial\Omega_{i_1} \cap \partial\Omega_{i_2}\}. \tag{2.2}$$

The unit outward normals to Ω_{i_1} and Ω_{i_2} are denoted by n_{i_1} and n_{i_2}. For a.e. $x \in \Gamma$, the two indices i_1, i_2 are chosen such that $(n^t\nu n)(x)|_{\Omega_{i_1}} \geq (n^t\nu n)(x)|_{\Omega_{i_2}}$. We also denote by n the two-valued field on Γ such that, for $x \in \partial\Omega_{i_1} \cap \partial\Omega_{i_2}$, $n(x) = \{n_{i_1}, n_{i_2}\}$.

For every two-valued function φ on Γ, we denote by φ_1 the value of φ which is defined on the side of Ω_{i_1} and by φ_2 the value of φ which is defined on the side of Ω_{i_2}. For example, when applying this convention to the normal vector n, we have $n_1 = n_{i_1}$ and $n_2 = n_{i_2}$. Mean values and jumps across Γ are defined as follows:

$$\{\varphi\} \stackrel{\text{def}}{=} \tfrac{1}{2}(\varphi_1 + \varphi_2), \qquad [\![\varphi]\!] \stackrel{\text{def}}{=} \varphi_1 - \varphi_2. \tag{2.3}$$

2.2. The interface/transmission conditions. Since we are not requiring that ν be uniformly positive definite, the mathematical nature of the PDE can change over the domain. To account for this, we define the subset $I \subset \Gamma$ that we subsequently refer to as the elliptic/hyperbolic interface:

$$I \stackrel{\text{def}}{=} \{x \in \Gamma;\ (n^t\nu n)(x)|_{\Omega_{i_1}} > 0 \ \text{ and } \ (n^t\nu n)(x)|_{\Omega_{i_2}} = 0\}. \tag{2.4}$$

For any x in I, we refer to Ω_{i_1} as the elliptic side of I at x, and we refer to Ω_{i_2} as the hyperbolic side of I at x. Note that the terms elliptic and hyperbolic are not volumic properties but interface properties. Ω_{i_1} being the elliptic side of I at x does not mean that the diffusivity is positive definite in Ω_{i_1}. The diffusivity may not be positive definite in Ω_{i_1}, but still have a nonzero component in the normal direction at x and vice versa for the hyperbolic side. We now define

$$\Omega^\dagger \stackrel{\text{def}}{=} \Omega \setminus I, \qquad I^+ \stackrel{\text{def}}{=} \{x \in I;\ \beta{\cdot}n_1 > 0\}, \qquad I^- \stackrel{\text{def}}{=} \{x \in I;\ \beta{\cdot}n_1 < 0\}. \tag{2.5}$$

I^+ (resp., I^-) is the subset of I where the advection field flows from the elliptic side to the hyperbolic side (resp., hyperbolic side to the elliptic side). The reader is referred to Figures 2(a)–2(b) for examples illustrating these definitions.

Let $\kappa \stackrel{\text{def}}{=} \nu^{1/2}$. The assumptions on ν imply that κ is bounded and positive semidefinite. We now rewrite (2.1) in mixed form by introducing the auxiliary unknown σ so that

$$\begin{cases} \sigma + \kappa\nabla u = 0 & \text{in } \Omega^\dagger, \\ \nabla{\cdot}(\kappa\sigma + \beta u) + \mu u = f & \text{in } \Omega, \end{cases} \tag{2.6}$$

and we require the following continuity property to hold:

$$[\![u]\!] = 0 \quad \text{on } I^+. \tag{INT1}$$

Observe that σ is defined in $\Omega^\dagger$ only. Indeed, u may be discontinuous across I, in which case $\kappa\nabla u$ cannot be defined in the distributional sense, since the product of a discontinuous function and a distribution is not legitimate within the standard distribution theory. Note also that (INT1) demands that u be continuous only on the portion of I where the advection field flows from the elliptic to the hyperbolic side.

The above formulation implies the following formal property:

$$\{(\kappa\sigma + \beta u)\cdot n\} = 0 \quad \text{on } \Gamma, \tag{INT2}$$

since the second equation in (2.6) is expected to hold in the whole domain Ω. Similarly, the first equation in (2.6) formally implies that

$$[\![u]\!] = 0 \quad \text{across } \Gamma \setminus I. \tag{2.7}$$

Finally by combining (INT1)–(INT2) on I^+ and using Lemma 2.1 together with the continuity of β, we observe that (INT1) amounts to enforcing

$$n_1^t \kappa_1 \nabla u_1 = 0 \quad \text{on } I^+. \tag{2.8}$$

LEMMA 2.1. *Let ν be a $d\times d$ positive semidefinite matrix; then*

$$\forall r \in \mathbb{R}^d, \quad (\nu r = 0) \Leftrightarrow (r^t \nu r = 0).$$

This simple lemma will be frequently invoked in the paper.

2.3. Asymptotic justification. In one space dimension, (INT1)–(INT2) are the interface conditions derived by Gastaldi and Quarteroni in [18], and these are the transmission conditions used in [8, 14]. These conditions are deduced by considering the following regularized problem supplemented with suitable boundary conditions:

$$(-\nu u_\epsilon' + \beta u_\epsilon)' + \mu u_\epsilon - \epsilon u_\epsilon'' = f.$$

Under the hypothesis that β is a nonzero constant, it can be proved that, as $\epsilon \to 0$, u_ϵ converges in $L^2(\Omega)$ to the so-called viscosity solution of (2.6) which satisfies (INT1)–(INT2).

As an example, consider $\Omega = (0,1)$ partitioned into $\Omega_1 \stackrel{\text{def}}{=} (0, \frac{1}{3})$, $\Omega_2 \stackrel{\text{def}}{=} (\frac{1}{3}, \frac{2}{3})$, $\Omega_3 \stackrel{\text{def}}{=} (\frac{2}{3}, 1)$. Take $f = 0$, $\mu = 0$, $\beta = 1$, and set $\nu|_{\Omega_1 \cup \Omega_3} = 1$ and $\nu|_{\Omega_2} = 0$. The viscosity solution of (2.6) corresponding to the Dirichlet boundary conditions $u(0) = 1$, $u(1) = 0$ is

$$u|_{\Omega_1} = 1, \quad u|_{\Omega_2} = 1, \quad u|_{\Omega_3} = 1 - e^{(x-1)}. \tag{2.9}$$

It can be verified that this solution satisfies (INT1)–(INT2), so that u is continuous at $x = \frac{1}{3}$ and discontinuous at $x = \frac{2}{3}$.

Let us mention at this point that there is a theoretical difficulty in the above regularization process if the advection field is zero and $\mu = 0$. In this case, the limit solution can be shown to be

$$u|_{\Omega_1} = 1, \quad u|_{\Omega_2} = 2 - 3x, \quad u|_{\Omega_3} = 0. \tag{2.10}$$

Comparing (2.10) with (2.9), we conclude that the limit process $\lim_{\epsilon \to 0,\, \beta \to 0}$ is not uniform.

We hereafter assume that in higher space dimensions (INT1)–(INT2) can be obtained by means of a regularization process and that there is no ambiguity on the limit, provided $\mu + \frac{1}{2}\nabla\cdot\beta \geq \mu_0 > 0$. The goal of the present paper is not to justify (INT1)–(INT2) but to show that these conditions yield a well-posed problem which we propose to approximate using a DG method.

2.4. The functional setting. We now reformulate the above problem in an appropriate functional setting. To this end, we set

$$L_u \stackrel{\text{def}}{=} L^2(\Omega), \quad L_\sigma \stackrel{\text{def}}{=} [L^2(\Omega^\dagger)]^d, \quad L \stackrel{\text{def}}{=} L_\sigma \times L_u,$$

so that $L = \{(z^\sigma, z^u);\ z^\sigma \in L_\sigma,\ z^u \in L_u\}$. L is a Hilbert space when equipped with the product norm. We define the so-called graph space:

$$W \stackrel{\text{def}}{=} \{z \in L;\ \kappa\nabla z^u \in L_\sigma,\ \nabla\cdot(\kappa z^\sigma + \beta z^u) \in L_u\},$$

where all the derivatives are understood in the weak sense. Then we consider the following operators:

$$\begin{aligned} &K : L \ni z \mapsto (z^\sigma, \mu z^u) \in L,\\ &A : W \ni z \mapsto (\kappa\nabla z^u, \nabla\cdot\Phi(z)) \in L,\\ &\tilde{A} : W \ni z \mapsto (-\kappa\nabla z^u, (\nabla\cdot\beta)z^u - \nabla\cdot\Phi(z)) \in L, \end{aligned}$$

where, for all $y \in W$, $\Phi(y) \stackrel{\text{def}}{=} \kappa y^\sigma + \beta y^u$. A and $\tilde{A} \in \mathcal{L}(W; L)$ are formal adjoints of each other. W is clearly a Hilbert space when equipped with the following norm:

$$\|y\|_W^2 \stackrel{\text{def}}{=} \|y\|_L^2 + \|Ay\|_L^2.$$

Moreover, K and A are bounded operators; i.e., $K \in \mathcal{L}(L; L)$ and $A \in \mathcal{L}(W; L)$. We refer to W as the graph space of A, and the norm of W is called the graph norm. Note that functions in W satisfy (INT2) but not necessarily (INT1).

2.5. Boundary operators. Following [10, 13], we consider the operator $D : W \longrightarrow W'$ defined by

$$\langle Dz, y\rangle_{W',W} \stackrel{\text{def}}{=} (Az, y)_L - (z, \tilde{A}y)_L. \tag{2.11}$$

Clearly $D \in \mathcal{L}(W; W')$, and D is a boundary operator in the following sense.

LEMMA 2.2. *The following holds:*

$$\langle Dz, y\rangle_{W',W} = \int_{\partial\Omega} [\Phi(z)\cdot n y^u + \Phi(y)\cdot n z^u - (\beta\cdot n) z^u y^u] - \int_I (\beta\cdot n_1)[\![z^u]\!][\![y^u]\!] \tag{2.12}$$

for all $(z, y) \in W \times W$ smooth enough for the integrals to make sense.

Proof. Integrating by parts over $\Omega^\dagger$ in (2.11) yields

$$\begin{aligned} \langle Dz, y\rangle_{W',W} = \int_I 2\left\{z^u n^t \kappa y^\sigma + y^u n^t \kappa z^\sigma + (\beta\cdot n) z^u y^u\right\} \\ + \int_{\partial\Omega} \left[z^u n^t \kappa y^\sigma + y^u n^t \kappa z^\sigma + (\beta\cdot n) z^u y^u\right]. \end{aligned}$$

We conclude using the fact that on I, $n_1^t \kappa_1 z_1^\sigma = -\beta\cdot n_1 [\![z^u]\!]$ and $n_2^t \kappa_2 = 0$, so that $2\{z^u n^t \kappa y^\sigma + y^u n^t \kappa z^\sigma + \beta\cdot n z^u y^u\} = -(\beta\cdot n_1)[\![y^u]\!] z_1^u - (\beta\cdot n_1)[\![z^u]\!] y_1^u + (\beta\cdot n_1) y_1^u z_1^u + (\beta\cdot n_2) y_2^u z_2^u = -(\beta\cdot n_1)[\![z^u]\!][\![y^u]\!]$. ∎

In other words, if z and y are smooth enough, D admits the following integral representation:

$$\langle Dz, y\rangle_{W',W} = \int_{\partial\Omega} y^t \mathcal{D} z - \int_I (\beta\cdot n_1)[\![z^u]\!][\![y^u]\!], \quad \text{where} \quad \mathcal{D} \stackrel{\text{def}}{=} \left[\begin{array}{c:c} 0 & \kappa n \\ \hdashline (\kappa n)^t & \beta\cdot n \end{array}\right]. \tag{2.13}$$

When the traces of z and y are not in $L^2(\partial\Omega \cup I)$, the above integrals have to be understood in some duality sense that we do not try to identify here.

Let $\alpha \in \{-1, +1\}$. Still following [10, 13], we assume that there is a second boundary operator $M \in \mathcal{L}(W; W')$ such that we have the following:

(i) M is monotone, meaning that $\langle My, y\rangle_{W',W} \geq 0$ for all $y \in W$.

(ii) Letting

$$V \stackrel{\text{def}}{=} \operatorname{Ker}(M - D), \qquad V^* \stackrel{\text{def}}{=} \operatorname{Ker}(M^* + D), \tag{2.14}$$

where M^* is the adjoint of M, the following holds:

$$D(V)^\perp = V^*, \qquad D(V^*)^\perp = V, \tag{2.15}$$

where, for all $E \subset W'$, $E^\perp$ denotes the polar set of E; i.e., $E^\perp$ is composed of those linear forms on W that vanish on E.

(iii) If $(z, y) \in W \times W$ are smooth enough for the integrals to make sense, M admits the following integral representation:

$$\langle Mz, y\rangle_{W',W} = \int_{\partial\Omega} y^t \mathcal{M} z + \int_I |\beta{\cdot}n| [\![z^u]\!][\![y^u]\!], \quad \text{where} \quad \mathcal{M} \stackrel{\text{def}}{=} \left[\begin{array}{c:c} 0 & -\alpha\kappa n \\ \hdashline \alpha(\kappa n)^t & |\beta{\cdot}n| \end{array}\right]. \tag{2.16}$$

The above hypotheses essentially assert the existence of surjective trace operators on I and $\partial\Omega$. The purpose of the operator M is to impose boundary conditions and to enforce (INT1). The choice $\alpha = +1$ (resp., $\alpha = -1$) is used to enforce Dirichlet (resp., Neumann) boundary conditions. Indeed, one can verify that if $z \in W$ is smooth enough, then

(i) if $\alpha = +1$,

$$\{z \in V\} \iff \{[\![z^u]\!]|_{I^+} = 0 \text{ and } z^u|_{\{x\in\partial\Omega;\ \kappa n\neq 0 \text{ or } \beta\cdot n<0\}} = 0\}, \tag{2.17}$$

$$\{z \in V^*\} \iff \{[\![z^u]\!]|_{I^-} = 0 \text{ and } z^u|_{\{x\in\partial\Omega;\ \kappa n\neq 0 \text{ or } \beta\cdot n>0\}} = 0\}; \tag{2.18}$$

(ii) if $\alpha = -1$,

$$\{z \in V\} \iff \{[\![z^u]\!]|_{I^+} = 0 \text{ and } \Phi(z){\cdot}n = \tfrac{1}{2}(\beta{\cdot}n + |\beta{\cdot}n|)z^u\}, \tag{2.19}$$

$$\{z \in V^*\} \iff \{[\![z^u]\!]|_{I^-} = 0 \text{ and } \Phi(z){\cdot}n = \tfrac{1}{2}(\beta{\cdot}n - |\beta{\cdot}n|)z^u\}. \tag{2.20}$$

For instance, taking $\alpha = +1$, if $z \in V$ is smooth enough, then for all $y \in W$, $\langle (M - D)z, y\rangle_{W',W} = 0$, so that if y is smooth enough,

$$-\int_{\partial\Omega} [(n^t \kappa y^\sigma) z^u + (\beta{\cdot}n)^- y^u z^u] + \int_{I^+} (\beta{\cdot}n_1)[\![y^u]\!][\![z^u]\!] = 0, \tag{2.21}$$

whence (2.17) is inferred.

2.6. Well-posedness. Let $a_0 \in \mathcal{L}(W \times L; \mathbb{R})$, $a_0^* \in \mathcal{L}(W \times L; \mathbb{R})$ be defined by

$$a_0(z, y) \stackrel{\text{def}}{=} (Kz, y)_L + (Az, y)_L \qquad \forall (z, y) \in W \times L, \tag{2.22}$$

$$a_0^*(z, y) \stackrel{\text{def}}{=} (Kz, y)_L + (\tilde{A}z, y)_L \qquad \forall (z, y) \in W \times L. \tag{2.23}$$

LEMMA 2.3 (L-coercivity). *a_0 and a_0^* are L-coercive on V and V^*, respectively, in the following sense:*

(2.24)
$$\forall y \in V, \quad a_0(y,y) \geq \|y^\sigma\|_{L_\sigma}^2 + \mu_0\|y^u\|_{L_u}^2 + \tfrac{1}{2}\|[\![y^u]\!]\|_{L^2(|\beta\cdot n|;I)}^2 + \tfrac{1}{2}\|y^u\|_{L^2(|\beta\cdot n|;\partial\Omega)}^2,$$

(2.25)
$$\forall y \in V^*, \quad a_0^*(y,y) \geq \|y^\sigma\|_{L_\sigma}^2 + \mu_0\|y^u\|_{L_u}^2 + \tfrac{1}{2}\|[\![y^u]\!]\|_{L^2(|\beta\cdot n|;I)}^2 + \tfrac{1}{2}\|y^u\|_{L^2(|\beta\cdot n|;\partial\Omega)}^2.$$

Proof. Using the definition of D and V, we infer, for all $y \in V$,

$$\begin{aligned}((K+A)y,y)_L &= ((K+\tfrac{1}{2}(A+\tilde{A}))y,y)_L + (\tfrac{1}{2}(A-\tilde{A})y,y)_L \\ &= \|y^\sigma\|_{L_\sigma}^2 + ((\mu + \tfrac{1}{2}\nabla\cdot\beta)y^u, y^u)_{L_u} + \tfrac{1}{2}\langle My, y\rangle_{W',W}.\end{aligned}$$

The desired result then follows from the construction of M. Proceed similarly to prove (2.25). ∎

Consider the following problem: For $f \in L_u$,

$$\text{(2.26)} \qquad \begin{cases}\text{find } z \in V \text{ such that, } \forall y \in L,\\ a_0(z,y) = (f, y^u)_{L_u}.\end{cases}$$

PROPOSITION 2.4. *Assume that there is an $M \in \mathcal{L}(W;W')$ satisfying* (2.15). *Then, problem* (2.26) *is well-posed.*

Proof. See the appendix. ∎

Remark 2.1. Note that the boundary and interface conditions in problem (2.26) are strongly enforced by requiring the solution to be a member of V.

Having in mind that boundary conditions are weakly enforced in DG methods, we introduce the following bilinear form:

$$\text{(2.27)} \qquad a(z,y) \stackrel{\text{def}}{=} a_0(z,y) + \tfrac{1}{2}\langle (M-D)z, y\rangle_{W',W} \qquad \forall (z,y) \in W \times W.$$

Clearly, all the terms above are well defined and $a \in \mathcal{L}(W \times W; \mathbb{R})$.

LEMMA 2.5 (L-coercivity). *a is L-coercive on W.*

Proof. Clearly for all $y \in W$,

$$\begin{aligned}a(y,y) &= ((K+A)y,y)_L + \tfrac{1}{2}\langle (M-D)y,y\rangle_{W',W} \\ &= ((K+\tfrac{1}{2}(A+\tilde{A}))y,y)_L + \tfrac{1}{2}((A-\tilde{A})y,y)_L + \tfrac{1}{2}\langle (M-D)y,y\rangle_{W',W} \\ &\geq \|y^\sigma\|_{L_\sigma}^2 + \mu_0\|y^u\|_{L_u}^2 + \tfrac{1}{2}\|[\![y^u]\!]\|_{L^2(|\beta\cdot n|;I)}^2 + \tfrac{1}{2}\|y^u\|_{L^2(|\beta\cdot n|;\partial\Omega)}^2;\end{aligned}$$

that is to say, a is L-coercive. ∎

We henceforth consider the following reformulation of problem (2.26): For $f \in L_u$,

$$\text{(2.28)} \qquad \begin{cases}\text{find } z \in W \text{ such that, } \forall y \in W,\\ a(z,y) = (f, y^u)_{L_u}.\end{cases}$$

PROPOSITION 2.6 (well-posedness). *Assume that there is an $M \in \mathcal{L}(W;W')$ satisfying* (2.15). *Then, problem* (2.28) *is well-posed, and the solutions to* (2.26) *and* (2.28) *coincide.*

Proof. The unique solution to (2.26) solves (2.28). Moreover, coercivity (see Lemma 2.5) immediately implies that the solution to (2.28) is unique. ∎

3. The discrete problem. The goal of this section is to construct a DG approximation of the model problem (2.28). After introducing the DG setting in section 3.1, we explain in detail in section 3.2 how the DG bilinear form a_h (see (3.17)) is constructed.

3.1. The discrete setting. Let $\{\mathcal{T}_h\}_{h>0}$ be a family of affine meshes of Ω compatible with the partition P_Ω, which, for simplicity, is supposed to be composed of polyhedra. The elements are not necessarily simplices, and the interfaces are not required to match. We denote by $\mathcal{F}_h^i$ the set of element interfaces, i.e., $F \in \mathcal{F}_h^i$ if F is a $(d-1)$-manifold and there are $T_1, T_2 \in \mathcal{T}_h$ such that $F = \partial T_1 \cap \partial T_2$. The set of the faces that separate the mesh from the exterior of Ω is denoted by $\mathcal{F}_h^\partial$; i.e., $F \in \mathcal{F}_h^\partial$ if F is a $(d-1)$-manifold and there is a $T \in \mathcal{T}_h$ such that $F = \partial T \cap \partial\Omega$. The set of all the faces is denoted by $\mathcal{F}_h$; i.e., $\mathcal{F}_h \stackrel{\text{def}}{=} \mathcal{F}_h^i \cup \mathcal{F}_h^\partial$. Moreover, for every face $F \in \mathcal{F}_h$ we introduce the set $\mathcal{T}_h(F) \stackrel{\text{def}}{=} \{T \in \mathcal{T}_h;\ F \subset \partial T\}$. The diameters of $T \in \mathcal{T}_h$ and $F \in \mathcal{F}_h$ are denoted by h_T and h_F, respectively. Without loss of generality, we assume that $h \le 1$.

For every nonnegative integer p, we define

$$P_{h,p} \stackrel{\text{def}}{=} \{v_h \in L^2(\Omega);\ \forall T \in \mathcal{T}_h,\ v_h|_T \in \mathbb{P}_p(T)\}, \tag{3.1}$$

where $\mathbb{P}_p(T)$ denotes the set of d-variate polynomials of total degree at most p on T. The mesh family $\{\mathcal{T}_h\}_{h>0}$ is assumed regular in the sense that

$$h_T \lesssim h_F, \quad F \subset \partial T, \tag{3.2}$$

$$\|\nabla v_h\|_{[L^2(T)]^d} \lesssim h_T^{-1} \|v_h\|_{L^2(T)} \quad \forall T \in \mathcal{T}_h,\ \forall v_h \in P_{h,p}, \tag{3.3}$$

$$\|v_h\|_{L^2(F)} \lesssim h_F^{-1/2} \|v_h\|_{L^2(\mathcal{T}_h(F))} \quad \forall F \in \mathcal{F}_h,\ \forall v_h \in P_{h,p}. \tag{3.4}$$

Let p_u and p_σ be two nonnegative integers such that $p_u - 1 \le p_\sigma \le p_u$, and define the following spaces:

$$\Sigma_h = [P_{h,p_\sigma}]^d, \qquad U_h = P_{h,p_u}, \qquad W_h = \Sigma_h \times U_h.$$

Observe that the inverse and trace inequalities (3.3) and (3.4) are local and that they can be applied componentwise to the functions in Σ_h. Let $H^s(\mathcal{T}_h) \stackrel{\text{def}}{=} \{v \in L^2(\Omega);\ v \in H^s(T)\ \forall T \in \mathcal{T}_h\}$ be equipped with the usual broken Sobolev norm denoted by $\|\cdot\|_{H^s(\mathcal{T}_h)}$, and define

$$W(h) \stackrel{\text{def}}{=} W \cap [H^1(\mathcal{T}_h)]^{d+1} + W_h.$$

According to the assumptions listed in section 2.1 and since $\{\mathcal{T}_h\}_{h>0}$ is compatible with the partition P_Ω, we have that

$$\nu \in [P_{h,0}]^{d,d} \quad \text{and} \quad \kappa \in [P_{h,0}]^{d,d}. \tag{3.5}$$

Remark 3.1. (i) The hypotheses on ν could be slightly weakened by assuming that ν is Lipschitz in each subdomain, but to avoid unnecessary technicalities we restrict ourselves to piecewise constant diffusivity. (ii) The mesh being compatible with the partition P_Ω means that the meshes fit the discontinuities of the diffusivity.

3.2. Design of the DG bilinear form. We now construct a discrete DG counterpart of the bilinear form a defined in (2.27) under the following constraints: (i) The discrete bilinear form should satisfy a discrete version of Lemma 2.5 (L-coercivity) and be strongly consistent. (ii) It should not require that the elliptic-hyperbolic interface I be identified a priori. (Since computers work in finite precision arithmetic, it may happen in practice that $n^t\nu n$ takes small values instead of being exactly zero, so that I is possibly difficult to identify.) (iii) It should include stabilizing terms suitable to weakly enforcing boundary and interface conditions.

3.2.1. Step 0: Discrete analogue of a. We start by localizing the volume and boundary/interface integrals in (2.27) and by deriving a discrete counterpart of the operator M defined by (2.16).

The localization of a_0 and D does not pose any problem and is done by splitting the volume and boundary/interface integrals over the mesh elements and faces. Now let us now focus our attention on the discretization of M.

We first construct a discrete counterpart of the boundary integral over $\partial\Omega$ in (2.16). To this end, we take inspiration from [11, 12] and we proceed as follows. For every $F \in \mathcal{F}_h^\partial$ we define the operator $M_F \in \mathcal{L}([L^2(F)]^{d+1}; [L^2(F)]^{d+1})$ such that, for all $(z, y) \in W(h) \times W(h)$,

$$(M_F(z), y)_{L,F} \stackrel{\text{def}}{=} -\alpha(z^u, \kappa y^\sigma \cdot n)_{L_u,F} + \alpha(\kappa z^\sigma \cdot n, y^u)_{L_u,F} + (M_F^{uu}(z^u), y^u)_{L_u,F},$$

where the boundary operator $M_F^{uu} \in \mathcal{L}(L^2(F); L^2(F))$ is defined by

$$M_F^{uu}(y^u) \stackrel{\text{def}}{=} \left(|\beta\cdot n| + \tfrac{1}{2}(\alpha+1)\lambda^2 h_F^{-1}\right) y^u, \tag{3.6}$$

and we have set $\lambda \stackrel{\text{def}}{=} \sqrt{n^t\nu n}$. Observe that M_F satisfies the following consistency conditions:

$$\mathrm{Ker}(\mathcal{M} - \mathcal{D}) \subset \mathrm{Ker}(M_F - \mathcal{D}), \qquad \mathrm{Ker}(\mathcal{M} + \mathcal{D}) \subset \mathrm{Ker}(M_F + \mathcal{D}). \tag{3.7}$$

Since M_F^{uu} is nonnegative, we define

$$|y^u|_{M,F}^2 \stackrel{\text{def}}{=} (M_F^{uu}(y^u), y^u)_{L_u,F}, \qquad |y^u|_M^2 \stackrel{\text{def}}{=} \textstyle\sum_{F\in\mathcal{F}_h^\partial} |y^u|_{M,F}^2.$$

The $|\cdot|_M$-seminorm will be used to measure the error due to the weak enforcement of boundary conditions.

Second, we construct a discrete counterpart of the boundary integral over the interface I in (2.16). To this end we extend the definition of λ to interior faces. For every $F \in \mathcal{F}_h^i$, the two-valued field $\lambda|_F = \{\lambda_1, \lambda_2\}$ is defined by

$$\lambda_i \stackrel{\text{def}}{=} \sqrt{n_i^t \nu n_i}, \quad i \in \{1, 2\}. \tag{3.8}$$

Recall that the indexing is defined such that $\lambda_1 \geq \lambda_2$. A discrete counterpart of the manifold I is defined by

$$\mathcal{I}_h \stackrel{\text{def}}{=} \{F \in \mathcal{F}_h^i;\ \lambda_1 > 0 \text{ and } \lambda_2 = 0\}.$$

To avoid unnecessary extra technicalities we suppose for the time being that the sign of $\beta\cdot n$ is constant on every interface F in $\mathcal{F}_h^i$. Then we identify two subsets of $\mathcal{I}_h$, say $\mathcal{I}_h^\pm$, representing the discrete versions of $I^\pm$. The boundary integral over the interface

I in (2.16) is then discretized by splitting it over $\mathcal{I}_h$. The sets $\mathcal{I}_h, \mathcal{I}_h^{\pm}$ and the above assumption on the sign of $\beta{\cdot}n$ are introduced solely to explicate the design of the DG bilinear form. These sets are not required in the final definition (3.17).

Combining all the steps above, we obtain the discrete bilinear form $a_h^{(0)}$ such that, for all $(z,y) \in W(h) \times W(h)$,

$$\begin{aligned} a_h^{(0)}(z,y) \stackrel{\text{def}}{=} & \sum_{T\in\mathcal{T}_h} [(Kz,y)_{L,T} + (Az,y)_{L,T}] + \tfrac{1}{2} \sum_{F\in\mathcal{F}_h^\partial} ((M_F - \mathcal{D})z, y)_{L,F} \\ & + \sum_{F\in\mathcal{I}_h^+} ((\beta{\cdot}n_1)[\![z^u]\!], [\![y^u]\!])_{L_u,F}. \end{aligned} \tag{3.9}$$

3.2.2. Step 1: *L*-coercivity and consistency. The bilinear form $a_h^{(0)}$ is not appropriate since it does not satisfy the design criteria (i) (L-coercivity) and (ii) (the interface $\mathcal{I}_h^+$ is explicitly identified), as stated in the introduction of section 3.2.

To remedy the above two problems, we introduce weighted averages. Let $\Gamma_h \stackrel{\text{def}}{=} \bigcup_{F\in\mathcal{F}_h^i} F$, and let ω be a two-valued weight function such that $[L^2(\Gamma_h)]^2 \ni \omega = (\omega_1, \omega_2)$ with $\omega_1 + \omega_2 = 1$ for a.e. $x \in \Gamma_h$. For all $y \in W(h)$ and every $F \in \mathcal{F}_h^i$, we define the weighted average and weighted jump of y^u across F as follows:

$$\{y^u\}_\omega \stackrel{\text{def}}{=} \omega_1 y_1^u + \omega_2 y_2^u, \qquad [\![y^u]\!]_\omega \stackrel{\text{def}}{=} 2(\omega_2 y_1^u - \omega_1 y_2^u), \tag{3.10}$$

where, for a.e. $x \in F$, $y_i^u(x) = \lim_{y\to x} y^u(y)|_{T_i}$, $i \in \{1,2\}$. We adopt similar definitions for $\{\Phi(y){\cdot}n\}_\omega$ and $[\![\Phi(y){\cdot}n]\!]_\omega$. To alleviate notation we set $\{\cdot\} \stackrel{\text{def}}{=} \{\cdot\}_{(\frac12,\frac12)}$ and $[\![\cdot]\!] \stackrel{\text{def}}{=} [\![\cdot]\!]_{(\frac12,\frac12)}$. The following algebraic formula will be used repeatedly in the rest of the paper:

$$\{ab\} = \{a\}\{b\}_\omega + \tfrac14 [\![a]\!]_\omega [\![b]\!]. \tag{3.11}$$

LEMMA 3.1. *The following integration by parts formula holds:*

$$\sum_{T\in\mathcal{T}_h} [(Az,y)_{L,T} - (z,\tilde{A}y)_{L,T}] = \sum_{F\in\mathcal{F}_h^\partial} (\mathcal{D}z,y)_{L,F} + \sum_{F\in\mathcal{F}_h^i} 2\left[\chi_{F,\omega}(z,y) + \chi_{F,\omega}(y,z)\right], \tag{3.12}$$

where

$$\chi_{F,\omega}(z,y) \stackrel{\text{def}}{=} (\{\Phi(z){\cdot}n\}, \{y^u\}_\omega)_{L_u,F} + \left([\![z^u]\!], \frac14 [\![\Phi(y){\cdot}n]\!]_\omega - \frac{\beta{\cdot}n_1}{2}\{y^u\}\right)_{L_u,F}. \tag{3.13}$$

Proof. Indeed, let *LHS* be the left-hand side of (3.12), and observe that

$$LHS = \sum_{F\in\mathcal{F}_h^\partial} (\mathcal{D}z,y)_{L,F} + 2\sum_{F\in\mathcal{F}_h^i} \int_F \left[\{\Phi(z){\cdot}n y^u\} + \{\Phi(y){\cdot}n z^u\} - \{(\beta{\cdot}n) z^u y^u\}\right].$$

Apply (3.11) to the averages involving $\Phi(z)$ and $\Phi(y)$, and observe that $\{(\beta{\cdot}n)z^u y^u\} = \frac{\beta{\cdot}n_1}{2}[\![z^u]\!]\{y^u\} + \frac{\beta{\cdot}n_1}{2}[\![y^u]\!]\{z^u\}$. ∎

Applying (3.12) in (3.9), we infer that

$$\begin{aligned} a_h^{(0)}(y,y) = & \sum_{T\in\mathcal{T}_h} (K + \tfrac12(A+\tilde{A})y, y)_{L,T} + \tfrac12 \sum_{F\in\mathcal{F}_h^\partial} (M_F(y), y)_{L,F} \\ & + \sum_{F\in\mathcal{I}_h^+} (|\beta{\cdot}n|[\![y^u]\!], [\![y^u]\!])_{L_u,F} + 2\sum_{F\in\mathcal{F}_h^i} \chi_{F,\omega}(y,y). \end{aligned}$$

This expression suggests that the following bilinear form can be considered in order to restore L-coercivity:

$$\hat{a}_h^{(1)}(z,y) \stackrel{\text{def}}{=} a_h^{(0)}(z,y) - 2\sum_{F\in\mathcal{F}_h^i} \chi_{F,\omega}(z,y). \tag{3.14}$$

Let us now examine the consistency of the new bilinear form $\hat{a}_h^{(1)}$. Let $z \in V \cap W(h)$ and let $y_h \in W_h$. Since $a_h^{(0)}$ is consistent, we just have to verify the consistency of the sum involving $\chi_{F,\omega}(z,y)$. For all $F \in \mathcal{F}_h^i \setminus \mathcal{I}_h$, it is clear that $\chi_{F,\omega}(z,y_h) = 0$ owing to (INT1)–(INT2). Let us now consider a face F in $\mathcal{I}_h$. From this point on, we design ω so that

$$\forall F \in \mathcal{I}_h,\ \forall x \in F, \quad \omega(x) = (1,0). \tag{3.15}$$

This implies that $\{y_h^u\}_\omega = y_h^u|_{T_1}$ and, since $z \in V$ and $F \in \mathcal{I}_h$,

$$\frac{1}{4}[\![\Phi(y_h)\cdot n]\!]_\omega - \frac{\beta\cdot n_1}{2}\{y_h^u\} = \frac{\beta\cdot n_1}{2} y_h^u|_{T_2(F)} - \frac{\beta\cdot n_1}{2}\{y_h^u\} = -\frac{\beta\cdot n_1}{4}[\![y_h^u]\!].$$

Thus, the following simplification occurs:

$$\begin{aligned} -2\chi_{F,\omega}(z,y_h) &= -2(\{\Phi(z)\cdot n\}, y_h^u|_{T_1(F)})_{L_u,F} + \left(\frac{\beta\cdot n_1}{2}[\![z^u]\!], [\![y_h^u]\!]\right)_{L_u,F} \\ &= \left(\frac{\beta\cdot n_1}{2}[\![z^u]\!], [\![y_h^u]\!]\right)_{L_u,F}. \end{aligned}$$

In this form it is now clear that $\chi_{F,\omega}(z,y_h) \neq 0$ over $\mathcal{I}_h$ (actually over $\mathcal{I}_h^-$ only); i.e., $\hat{a}_h^{(1)}$ is not consistent. To make the bilinear form consistent, we remove the trouble-making term by defining the new bilinear form

$$\begin{aligned} a_h^{(1)}(z,y) &\stackrel{\text{def}}{=} \hat{a}_h^{(1)}(z,y) - \sum_{F\in\mathcal{I}_h}\left(\frac{\beta\cdot n_1}{2}[\![z^u]\!], [\![y^u]\!]\right)_{L_u,F} \\ &= \sum_{T\in\mathcal{T}_h}[(Kz,y)_{L,T} + (Az,y)_{L,T}] + \frac{1}{2}\sum_{F\in\mathcal{F}_h^\partial}((M_F - \mathcal{D})z,y)_{L,F} \\ &\quad - 2\sum_{F\in\mathcal{F}_h^i}\chi_{F,\omega}(z,y) + \sum_{F\in\mathcal{I}_h}\left(\frac{|\beta\cdot n|}{2}[\![z^u]\!], [\![y^u]\!]\right)_{L_u,F}. \end{aligned} \tag{3.16}$$

Then, the above arguments imply that $a_h^{(1)}$ is strongly consistent and satisfies the following coercivity property: For all $y \in W(h)$,

$$a_h^{(1)}(y,y) \geq \|y^\sigma\|_{L_\sigma}^2 + \mu_0\|y^u\|_{L_u}^2 + \tfrac{1}{2}|y^u|_M^2 + \tfrac{1}{2}\|[\![y^u]\!]\|_{L^2(|\beta\cdot n|;\mathcal{I}_h)}^2.$$

3.2.3. Step 2: Elimination of $\mathcal{I}_h$ and stability. Note that $a_h^{(1)}$ is not yet satisfactory since it still requires $\mathcal{I}_h$ be explicitly identified by the user, thus violating our design constraint (ii).

The key idea to eliminate any reference to $\mathcal{I}_h$ in (3.16) is to observe that the bilinear form $(|\beta\cdot n|[\![z^u]\!], [\![y^u]\!])_{L_u,F}$ is positive and consistent on every face which is not in $\mathcal{I}_h$. In other words, L-coercivity and consistency are preserved by extending the sum over $\mathcal{I}_h$ to $\mathcal{F}_h^i$. We thus replace $a_h^{(1)}$ by

(3.17)
$$
\begin{aligned}
a_h(z,y) \stackrel{\text{def}}{=} & \sum_{T\in\mathcal{T}_h} \left[(Kz,y)_{L,T} + (Az,y)_{L,T}\right] + \frac{1}{2}\sum_{F\in\mathcal{F}_h^\partial} (M_F(z)-\mathcal{D}z,y)_{L,F} \\
& - 2\sum_{F\in\mathcal{F}_h^i} \left[\left(\{\Phi(z)\cdot n\}, \{y^u\}_\omega\right)_{L_u,F} + \left([\![z^u]\!], \frac{1}{4}[\![\Phi(y)\cdot n]\!]_\omega - \frac{\beta\cdot n_1}{2}\{y^u\}\right)_{L_u,F}\right] \\
& + \sum_{F\in\mathcal{F}_h^i} (S_F([\![z^u]\!]), [\![y^u]\!])_{L,F},
\end{aligned}
$$

where the interface operator $S_F \in \mathcal{L}(L^2(F);L^2(F))$ is defined for all $F\in\mathcal{F}_h^i$ by

$$
S_F(v) \stackrel{\text{def}}{=} \left(\frac{|\beta\cdot n|}{2} + \frac{\lambda_2^2}{h_F}\right) v. \tag{3.18}
$$

Observe that λ_2 is by definition the minimum of λ_1 and λ_2. Hence, automatically

$$
S_F(v) = \frac{|\beta\cdot n|}{2} v \qquad \forall F\in\mathcal{I}_h. \tag{3.19}
$$

Finally, to avoid any reference to $\mathcal{I}_h$ in the definition of the weighting function ω while satisfying condition (3.15), we define ω as follows:

$$
\omega \stackrel{\text{def}}{=} \begin{cases} (\frac{\lambda_1}{2\{\lambda\}}, \frac{\lambda_2}{2\{\lambda\}}) & \text{if } \lambda_1 > 0, \\ (\frac{1}{2},\frac{1}{2}) & \text{otherwise.} \end{cases} \tag{3.20}
$$

Although other expressions for ω are possible, this one is simple and yields robust error estimates. A similar choice is made in [6, 15].

The discrete problem is now formulated as follows:

$$
\begin{cases} \text{Seek } z_h\in W_h \text{ such that} \\ a_h(z_h,y_h) = (f,y_h^u)_{L_u} \quad \forall y_h\in W_h. \end{cases} \tag{3.21}
$$

Observe that the σ-component of the unknown can be eliminated locally since the jumps of this quantity across element interfaces are not penalized; see, e.g., [11, section 4.4].

Remark 3.2. The use of weights in DG methods has been highlighted in several articles (see, e.g., [16, 20, 19, 21]). Although some of the above references point out that using weights may yield higher accuracy, they do not connect the weights with the coefficients of the problem. The weight/diffusivity dependence has recently been investigated in [6, 15], where the authors show that the use of a particular weighted average improves the stability of the numerical scheme in problems with high diffusivity contrasts, the diffusivity being still positive definite. In the present case, resorting to weighted average and jump operators is required for the method to select the proper interface conditions automatically, i.e., the design constraint (ii).

4. Convergence analysis. In this section we carry out the convergence analysis of the discrete problem (3.21). The main results are Theorems 4.5 and 4.7.

4.1. Basic convergence estimates. For all $y\in W(h)$ we introduce the following seminorm:

$$
|y^u|_{J,F}^2 \stackrel{\text{def}}{=} (S_F([\![y^u]\!]), [\![y^u]\!])_{L_u,F}, \qquad |y^u|_J^2 \stackrel{\text{def}}{=} \sum_{F\in\mathcal{F}_h^i} |y^u|_{J,F}^2. \tag{4.1}
$$

The space $W(h)$ is equipped with the following discrete norm:

$$\|y\|_{h,\kappa}^2 \stackrel{\text{def}}{=} \|y\|_L^2 + |y^u|_J^2 + |y^u|_M^2 + \sum_{T\in\mathcal{T}_h} \|\kappa\nabla y^u\|_{L_\sigma,T}^2. \tag{4.2}$$

The following two lemmata follow from the design procedure outlined in section 3.2.

LEMMA 4.1 (consistency). *Let z solve* (2.6) *and z_h solve* (3.21). *Assume, moreover, that $z \in [H^1(\mathcal{T}_h)]^{d+1}$. Then,*

$$\forall y_h \in W_h, \quad a_h(z - z_h, y_h) = 0.$$

LEMMA 4.2 (L-coercivity). *For all h and for all y in $W(h)$,*

$$a_h(y,y) \gtrsim \|y\|_L^2 + |y^u|_J^2 + |y^u|_M^2.$$

In order to estimate the L^2-norm of the diffusive derivative $\kappa\nabla z^u$ we need the following result.

LEMMA 4.3 (stability). *The following bound holds:*

$$\forall z_h \in W_h, \quad \|z_h\|_{h,\kappa} \lesssim \sup_{y_h\in W_h\setminus\{0\}} \frac{a_h(z_h,y_h)}{\|y_h\|_{h,\kappa}}.$$

Proof. Let $z_h \in W_h$ and set $\mathbb{S} \stackrel{\text{def}}{=} \sup_{y_h\in W_h\setminus\{0\}} \frac{a_h(z_h,y_h)}{\|y_h\|_{h,\kappa}}$.

(1) Owing to Lemma 4.2,

$$\|z_h\|_L^2 + |z_h^u|_J^2 + |z_h^u|_M^2 \lesssim a_h(z_h,z_h) \lesssim \mathbb{S}\|z_h\|_{h,\kappa}. \tag{4.3}$$

(2) Control of $\mathtt{B} \stackrel{\text{def}}{=} \sum_{T\in\mathcal{T}_h} \|\kappa\nabla z_h^u\|_{L_\sigma,T}^2$. Let $\pi_h^\sigma \in \Sigma_h$ be the field such that, for all $T \in \mathcal{T}_h$, $\pi_h^\sigma|_T \stackrel{\text{def}}{=} \kappa\nabla z_h^u|_T$. From the definition of a_h it follows that

$$\begin{aligned}\mathtt{B} = a_h(z_h,(\pi_h^\sigma,0)) - (z_h^\sigma,\pi_h^\sigma)_{L_\sigma} + \sum_{F\in\mathcal{F}_h^\partial} \frac{1+\alpha}{2}(\kappa n z_h^u, \pi_h^\sigma)_{L_\sigma,F} \\ + \frac{1}{2}\sum_{F\in\mathcal{F}_h^i} (\llbracket z_h^u\rrbracket, \llbracket n^t\kappa\pi_h^\sigma\rrbracket_\omega)_{L_u,F}.\end{aligned}$$

Let R_i, $i \in \{1,2,3\}$ denote the last three terms in the right-hand side. The first term is bounded from above as follows:

$$|R_1| \le \|z_h^\sigma\|_{L_\sigma}\|\pi_h^\sigma\|_{L_\sigma} \lesssim \|z_h^\sigma\|_{L_\sigma}^2 + \gamma\mathtt{B},$$

where γ can be chosen as small as needed.

The second term vanishes if $\alpha = -1$. If $\alpha = +1$, use trace inequality (3.4) together with (3.6) to get

$$|R_2| \lesssim \sum_{F\in\mathcal{F}_h^\partial} h_F^{-\frac12} |(n^t\nu n z_h^u, z_h^u)_{L_u,F}|^{\frac12} \|\pi_h^\sigma\|_{L_\sigma,\mathcal{T}_h(F)} \lesssim \sum_{F\in\mathcal{F}_h^\partial} |z_h^u|_{M,F}\|\pi_h^\sigma\|_{L_\sigma,\mathcal{T}_h(F)}.$$

Consequently, $|R_2| \lesssim |z_h^u|_M^2 + \gamma\mathtt{B}$. According to (3.20), for all $\mathcal{F}_h^i \ni F = \partial T_1 \cap \partial T_2$,

$$\|\llbracket n^t\kappa\pi_h^\sigma\rrbracket_\omega\|_{L_u,F} = \frac{\lambda_1\lambda_2}{\{\lambda\}}\|\llbracket (n^t\kappa/\lambda)\pi_h^\sigma\rrbracket\|_{L_u,F} \lesssim h_F^{-\frac12}\frac{\lambda_1\lambda_2}{\{\lambda\}}\|\pi_h^\sigma\|_{L_\sigma,\mathcal{T}_h(F)}.$$

Using the above relation together with (3.18) yields

$$|R_3| \lesssim \sum_{F\in\mathcal{F}_h^i} \frac{\lambda_1}{\{\lambda\}} \left(\frac{\lambda_2^2}{h_F}\|[\![z_h^u]\!]\|_{L_u,F}^2\right)^{\frac12} \|\pi_h^\sigma\|_{L_\sigma,\mathcal{T}_h(F)} \lesssim |z_h^u|_J^2 + \gamma\mathsf{B}.$$

The above bounds with $\gamma = \frac16$ together with Lemma 4.2 give

$$\tfrac12 \sum_{T\in\mathcal{T}_h} \|\kappa\nabla z_h^\sigma\|_{L_\sigma,T}^2 \lesssim a_h(z_h,(\pi_h^\sigma,0)) + a_h(z_h,z_h) \lesssim \mathbb{S}\|z_h\|_{h,\kappa}, \tag{4.4}$$

where we used the fact that, by definition, $\|(\pi_h^\sigma,0)\|_{h,\kappa} = \|\pi_h^\sigma\|_{L_\sigma} \le \|z_h\|_{h,\kappa}$. Observe that, owing to the choice of the weight function ω, the above estimate is robust with respect to the possible discontinuity and anisotropy of ν.

(3) Equations (4.3)–(4.4) yield $\|z_h\|_{h,\kappa}^2 \lesssim \mathbb{S}\|z_h\|_{h,\kappa}$, i.e., the desired result. ∎

Let us now introduce

$$W_h^\perp \overset{\text{def}}{=} \{y\in W(h);\ \forall w_h\in W_h,\ (y,w_h)_L = 0\}. \tag{4.5}$$

Moreover, we define the following norm on $W(h)$:

$$[\![y]\!]^2 \overset{\text{def}}{=} \|y\|_{h,\kappa}^2 + \sum_{T\in\mathcal{T}_h}\left[\frac{\mathfrak{h}_T}{h_T^2}\|y^u\|_{L_u,T}^2 + h_T\|y^\sigma\|_{L_\sigma,\partial T}^2 + \sum_{F\subset\partial T}\frac{\mathfrak{h}_F}{h_F}\|y^u\|_{L_u,F}^2\right], \tag{4.6}$$

where, for all $T\in\mathcal{T}_h$ and for all $F\in\mathcal{F}_h$, we have defined

$$\mathfrak{h}_T \overset{\text{def}}{=} \max(\|\nu\|_{[L^\infty(T)]^{d,d}}, h_T), \quad \mathfrak{h}_F \overset{\text{def}}{=} \begin{cases}\max(\lambda_1^2,h_F) & \text{if } F\in\mathcal{F}_h^i,\\ \max(\lambda^2,h_F) & \text{if } F\in\mathcal{F}_h^\partial.\end{cases} \tag{4.7}$$

The last property needed to prove convergence is stated in the following.

LEMMA 4.4 (continuity). *The following holds:*

$$\forall (z,y_h)\in W_h^\perp\times W_h,\quad a_h(z,y_h) \lesssim [\![z]\!]\|y_h\|_{h,\kappa}.$$

Proof. Let $(z,y_h)\in W_h^\perp\times W_h$. Using the integration by parts formula (3.12), we obtain

$$\begin{aligned} a_h(z,y_h) = {}& \sum_{T\in\mathcal{T}_h}(z,(K+\tilde A)y_h)_{L,T} + 2\sum_{F\in\mathcal{F}_h^i}\chi_{F,\omega}(y_h,z)\\ &+ \tfrac12\sum_{F\in\mathcal{F}_h^\partial}(M_F(z)+\mathcal{D}z,y_h)_{L,F} + \sum_{F\in\mathcal{F}_h^i}(S_F([\![z^u]\!]),[\![y_h^u]\!])_{L_u,F}. \end{aligned} \tag{4.8}$$

We now derive bounds for the four terms in the right-hand side, say $R_1,\dots,R_4$. For the first one we have

$$\begin{aligned} &(z,(K+\tilde A)y_h)_{L,T}\\ &\quad = (z^\sigma, y_h^\sigma - \kappa\nabla y_h^u)_{L_\sigma,T} + (z^u, \mu y_h^u - \nabla\cdot(\kappa y_h^\sigma) - \overline{\beta}\cdot\nabla y_h^u - (\beta-\overline{\beta})\cdot\nabla y_h^u)_{L_u,T}, \end{aligned}$$

where, for all $T\in\mathcal{T}_h$, $\overline{\beta}|_T$ is the mean value of the field β over T. Observe that, since $\kappa\nabla y_h^u\in\Sigma_h$, $\overline{\beta}\cdot\nabla y_h^u\in U_h$, and $z\in W_h^\perp$, $(z^\sigma,\kappa\nabla y_h^u)_{L_\sigma}=0$ and $(z^u,\overline{\beta}\cdot\nabla y_h^u)_{L_u}=0$. As a result,

$$|R_1| \lesssim \|z\|_L \|y_h\|_L + \sum_{T\in\mathcal{T}_h} \left[\|z^u\|_{L_u,T} h_T^{-1} \|\kappa y_h^\sigma\|_{L_\sigma,T} + (z^u, (\beta-\overline{\beta})\cdot\nabla y_h^u)_{L_u,T}\right]$$
$$\lesssim \|z\|_L \|y_h\|_L + \sum_{T\in\mathcal{T}_h} \left[\frac{\mathfrak{h}_T^{\frac{1}{2}}}{h_T} \|z^u\|_{L_u,T} \|y_h^\sigma\|_{L_\sigma,T} + \|\beta\|_{[\mathfrak{C}^1(\Omega)]^d} \|y_h^u\|_{L_u,T} \|z^u\|_{L_u,T}\right],$$

and, therefore, $|R_1| \lesssim [\![z]\!] \|y_h\|_L$. The second term R_2 can be simplified as follows:

$$R_2 = 2 \sum_{F\in\mathcal{F}_h^i} \left[(\{z^u\}_\omega, \{n^t \kappa y_h^\sigma\})_{L_u,F} + \frac{1}{4} ([\![n^t \kappa z^\sigma]\!]_\omega, [\![y_h^u]\!])_{L_u,F} + \left(\frac{\beta\cdot n_1}{2} \{z^u\}, [\![y_h^u]\!]\right)_{L_u,F}\right].$$

Let $R_{2,i}$, $i = 1, 2, 3$, be the addends of R_2. Using the definition of the weight function ω, (3.20), together with the inverse trace inequality (3.4) and definition (4.7), we infer

$$|R_{2,1}| \lesssim \sum_{F\in\mathcal{F}_h^i} \mathfrak{h}_F^{\frac{1}{2}} h_F^{-\frac{1}{2}} \left(\|z_1^u\|_{L_u,F} + \|z_2^u\|_{L_u,F}\right) \|y_h^\sigma\|_{L_\sigma,\mathcal{T}_h(F)},$$
$$|R_{2,2}| \lesssim \sum_{F\in\mathcal{F}_h^i} \frac{\lambda_1}{\{\lambda\}} h_F^{\frac{1}{2}} \left(\|z_1^\sigma\|_{L_\sigma,F} + \|z_2^\sigma\|_{L_\sigma,F}\right) \lambda_2 h_F^{-\frac{1}{2}} \|[\![y_h^u]\!]\|_{L_u,F}$$
$$\lesssim \sum_{F\in\mathcal{F}_h^i} h_F^{\frac{1}{2}} \left(\|z_1^\sigma\|_{L_\sigma,F} + \|z_2^\sigma\|_{L_\sigma,F}\right) |y_h^u|_{J,F},$$
$$|R_{2,3}| \lesssim \left(\|z_1^u\|_{L_u,F} + \|z_2^u\|_{L_u,F}\right) |y_h^u|_{J,F} \lesssim \mathfrak{h}_F^{\frac{1}{2}} h_F^{-\frac{1}{2}} \left(\|z_1^u\|_{L_u,F} + \|z_2^u\|_{L_u,F}\right) |y_h^u|_{J,F}.$$

Therefore, $|R_2| \lesssim [\![z]\!] \|y_h\|_L$. The third term is expanded as follows:

$$|R_3| = \sum_{F\in\mathcal{F}_h^\partial} \left[\frac{1-\alpha}{2} (z^u, n^t \kappa y_h^\sigma)_{L_u,F} + \frac{1+\alpha}{2} (n^t \kappa z^\sigma, y_h^u)_{L_u,F} + \frac{1}{2} ((M_F^{uu} + \beta\cdot n) z^u, y_h^u)_{L_u,F}\right].$$

Let $R_{3,i}$, $i = 1, \ldots, 3$, be the addends of R_3. If $\alpha = -1$, then $R_{3,2} = 0$, and using (4.7) and (3.4), we infer that

$$|R_{3,1}| \lesssim \sum_{F\in\mathcal{F}_h^\partial} \lambda h_F^{-\frac{1}{2}} \|z^u\|_{L_u,F} \|y_h^\sigma\|_{L_\sigma,\mathcal{T}_h(F)} \lesssim \sum_{F\in\mathcal{F}_h^\partial} \mathfrak{h}_F^{\frac{1}{2}} h_F^{-\frac{1}{2}} \|z^u\|_{L_u,F} \|y_h^\sigma\|_{L_\sigma,\mathcal{T}_h(F)},$$

whereas, if $\alpha = +1$, then $R_{3,1} = 0$, and (3.18) implies that

$$|R_{3,2}| \lesssim \sum_{F\in\mathcal{F}_h^\partial} h_F^{\frac{1}{2}} \|z^\sigma\|_{L_\sigma,F} \lambda h_F^{-\frac{1}{2}} \|y_h^u\|_{L_u,F} \lesssim \sum_{F\in\mathcal{F}_h^\partial} h_F^{\frac{1}{2}} \|z^\sigma\|_{L_\sigma,F} |y_h^u|_{M,F}.$$

Finally, (3.18) yields $|R_{3,3}| \lesssim |z^u|_M |y_h^u|_M$. Therefore, $|R_3| \lesssim [\![z]\!] \|y_h\|_L$. For the fourth term we immediately have $|R_4| \le |z^u|_J |y_h^u|_J$. The desired result is obtained by collecting the above bounds. ∎

Let π_h be the L^2-projection onto W_h. Upon collecting the above results (consistency, stability, and continuity) and observing that $z - \pi_h z \in W_h^\perp$, the second Strang lemma immediately yields the following convergence result.

THEOREM 4.5 (convergence). *Let z solve* (2.28) *and z_h solve* (3.21). *Assume that $z \in [H^1(\mathcal{T}_h)]^{d+1}$. Then,*

$$\|z - z_h\|_{h,\kappa} \lesssim [\![z - \pi_h z]\!].$$

Owing to the regularity of the mesh family $\{\mathcal{T}_h\}_{h>0}$, the following interpolation property holds: For all $z \in [H^{r_\sigma}(\mathcal{T}_h)]^d \times H^{r_u}(\mathcal{T}_h)$,

$$\begin{aligned}|\!|\!| z - \pi_h z |\!|\!| \lesssim & \left(\sum_{T\in\mathcal{T}_h} h_T^{2s_\sigma} \|z^\sigma\|^2_{[H^{s_\sigma}(T)]^d} + \mathfrak{h}_T h_T^{2s_u-2} \|z^u\|^2_{H^{s_u}(T)} \right)^{\frac12} \\ & + \left(\sum_{F\in\mathcal{F}_h} \mathfrak{h}_F h_F^{2s_u-2} \|z^u\|^2_{H^{s_u}(\mathcal{T}_h(F))} \right)^{\frac12}, \end{aligned} \tag{4.9}$$

where $s_\sigma \overset{\text{def}}{=} \min(r_\sigma, p_\sigma+1)$ and $s_u \overset{\text{def}}{=} \min(r_u, p_u+1)$. Since $p_u - 1 \le p_\sigma$ and provided $(r_\sigma, r_u) \ge (p_\sigma+1, p_u+1)$, the above interpolation error is of order h^{p_u}, i.e.,

$$|\!|\!| z - \pi_h z |\!|\!| \lesssim h^{p_u} \|z\|_{[H^{p_\sigma+1}(\mathcal{T}_h)]^d \times H^{p_u+1}(\mathcal{T}_h)}. \tag{4.10}$$

Remark 4.1. The above estimate is optimal for the $\|\cdot\|_{h,\kappa}$-norm but yields suboptimal convergence in the L^2-norm. If $p_\sigma = p_u - 1$, the error estimate is optimal in the L^2-norm for z_h^σ but is still suboptimal for the L^2-norm of z_h^u. From a theoretical viewpoint, it is optimal to set $\Sigma_h = [P_{h,p_u-1}]^d$, i.e., to work with different approximation orders for the σ- and u-components. However, working with equal-order approximation may be more convenient for implementation purposes.

Remark 4.2 (positive definite diffusivity). If the diffusivity is such that $\nu \ge \nu_0 \mathcal{I}_d$ with $\nu_0 = \mathcal{O}(1)$, the estimate (4.10) can be improved using a duality argument. More precisely, consider the mapping $L_u \ni y^u \longmapsto \psi \in V^*$ defined by

$$(K + \tilde{A})\psi = (0, y^u),$$

and assume that the following bound holds:

$$\|\psi^u\|_{H^2(\mathcal{T}_h)} + \|\psi^\sigma\|_{[H^1(\mathcal{T}_h)]^d} \lesssim \|y^u\|_{L_u}. \tag{4.11}$$

Adapting the reasoning in [11, section 5.3], if $r_u \ge p_u + 1$ and $p_u \ge 1$, it can be proved that

$$\|z - z_h\|_{L_u} \lesssim h^{p_u+1} \|z\|_{[H^{p_\sigma+1}(\mathcal{T}_h)]^d \times H^{p_u+1}(\mathcal{T}_h)}.$$

4.2. Improved convergence estimates. Owing to the definition of the $\|\cdot\|_{h,\kappa}$-norm, the convergence result of Theorem 4.5 does not contain an estimate involving the advective derivative. Such an estimate can be obtained, assuming that

$$\kappa \text{ is scalar-valued}, \tag{4.12}$$

and we still admit that κ may vanish over a portion of the domain. Define the following new discrete norm on $W(h)$:

$$\|y\|^2_{h,\kappa,\beta} \overset{\text{def}}{=} \|y\|^2_{h,\kappa} + \|y^u\|^2_{h,\beta} \quad \text{with} \quad \|y^u\|^2_{h,\beta} \overset{\text{def}}{=} \sum_{T\in\mathcal{T}_h} h_T \|\beta\cdot\nabla y^u\|^2_{L_u,T}. \tag{4.13}$$

LEMMA 4.6. *Assume that κ satisfies* (4.12). *Then the following bound holds:*

$$\forall z_h \in W_h, \quad \|z_h\|_{h,\kappa,\beta} \lesssim \sup_{y_h \in W_h\setminus\{0\}} \frac{a_h(z_h, y_h)}{\|y_h\|_{h,\kappa,\beta}}.$$

Proof. Let $z_h \in W_h$ and set $\mathbb{S} \overset{\text{def}}{=} \sup_{y_h \in W_h \setminus \{0\}} \frac{a_h(z_h, y_h)}{\|y_h\|_{h,\kappa,\beta}}$.

(1) Proceeding as in Lemma 4.3 and observing that $\|(\pi_h^\sigma, 0)\|_{h,\kappa,\beta} = \|(\pi_h^\sigma, 0)\|_{h,\kappa}$, we conclude that

$$\|z_h\|_{h,\kappa}^2 \lesssim \mathbb{S}\|z_h\|_{h,\kappa,\beta}. \tag{4.14}$$

(2) We define the field $W_h \ni \pi_h \overset{\text{def}}{=} (0, \pi_h^u)$ in such a way that, for all $T \in \mathcal{T}_h$, $\pi_h^u|_T = h_T \overline{\beta}\cdot\nabla z_h^u$, where $\overline{\beta}$ is the mean of β over T. Using (3.3) together with the regularity of β and the fact that $h_T \le 1$, for all $T \in \mathcal{T}_h$, we have

$$\begin{aligned} h_T^{-\frac{1}{2}}\|\pi_h^u\|_{L_u,T} &\le h_T^{\frac{1}{2}}\|(\overline{\beta} - \beta)\cdot\nabla z_h^u\|_{L_u,T} + h_T^{\frac{1}{2}}\|\beta\cdot\nabla z_h^u\|_{L_u,T} \\ &\le h_T^{\frac{1}{2}}\|\beta\|_{[\mathfrak{C}^1(\Omega)]^d}\|z_h^u\|_{L_u,T} + h_T^{\frac{1}{2}}\|\beta\cdot\nabla z_h^u\|_{L_u,T}. \end{aligned} \tag{4.15}$$

(i) We first show that $\|\pi_h\|_{h,\kappa,\beta} \lesssim \|z_h\|_{h,\kappa,\beta}$. According to the above bound, it is clear that $\|\pi_h^u\|_{L_u} \lesssim \|z_h\|_{h,\kappa,\beta}$. Commuting the operators $\kappa\nabla$ and $\overline{\beta}\cdot\nabla$ and applying the inverse inequality (3.3), we infer that

$$\sum_{T\in\mathcal{T}_h}\|\kappa\nabla\pi_h^u\|_{L_u,T}^2 = \sum_{T\in\mathcal{T}_h} h_T^2\|\overline{\beta}\cdot\nabla(\kappa\nabla z_h^u)\|_{L_u,T}^2 \lesssim \sum_{T\in\mathcal{T}_h}\|\kappa\nabla z_h^u\|_{L_\sigma,T}^2.$$

Moreover, the regularity of β and again (3.3) yield

$$\begin{aligned} \|\pi_h^u\|_{h,\beta}^2 &\lesssim \sum_{T\in\mathcal{T}_h} h_T^3\left[\|\partial_{\overline{\beta}}(\beta\cdot\nabla z_h^u)\|_{L_u,T} + \|\partial_{\overline{\beta}}\beta\|_{[L^\infty(T)]^d}h_T^{-1}\|z_h^u\|_{L_u,T}\right]^2 \\ &\lesssim \sum_{T\in\mathcal{T}_h} h_T\left[\|\beta\cdot\nabla z_h^u\|_{L_u,T} + \|z_h^u\|_{L_u,T}\right]^2. \end{aligned}$$

The term $|\pi_h^u|_J$ is treated as follows:

$$|\pi_h^u|_J^2 \lesssim \sum_{F\in\mathcal{F}_h^i}\||\beta\cdot n|^{\frac{1}{2}}[\![\pi_h^u]\!]\|_{L_u,F}^2 + \sum_{F\in\mathcal{F}_h^i}\|\lambda_2 h_F^{-\frac{1}{2}}[\![\pi_h^u]\!]\|_{L_u,F}^2 \overset{\text{def}}{=} R_1 + R_2.$$

Using (3.4) together with (4.15), we immediately conclude that

$$|R_1| \lesssim \sum_{F\in\mathcal{F}_h^i} h_F^{-1}\|\pi_h^u\|_{L_u,\mathcal{T}_h(F)}^2 \lesssim \|z_h\|_{h,\kappa,\beta}^2.$$

The second term is zero if $\lambda_2 = 0$. On the other hand, by definition, if $\lambda_2 > 0$, then $\lambda_1 > 0$; i.e., κ is nonzero on both sides of the considered element interface. We proceed using the trace inequality (3.4) together with assumption (4.12) (i.e., $\kappa_1 = \lambda_1 I$, $\kappa_2 = \lambda_2 I$, and $\lambda_1 \ge \lambda_2$), to get

$$|R_2| \lesssim \sum_{F\in\mathcal{F}_h^i}\frac{\lambda_2^2}{h_F}h_F\|\overline{\beta}\cdot\nabla z_h^u\|_{L_u,\mathcal{T}_h(F)}^2 \lesssim \sum_{F\in\mathcal{F}_h^i}\lambda_2^2\|\nabla z_h^u\|_{L_\sigma,\mathcal{T}_h(F)}^2 \le \sum_{F\in\mathcal{F}_h^i}\|\kappa\nabla z_h^u\|_{L_\sigma,\mathcal{T}_h(F)}^2,$$

whence $|\pi_h^u|_J \lesssim \|z_h\|_{h,\kappa,\beta}$. In a similar way we can prove that $|\pi_h^u|_M \lesssim \|z_h\|_{h,\kappa,\beta}$.

(ii) Estimate for $\|z_h^u\|_{h,\beta}$. Integrating by parts only the diffusive terms and setting $\tilde{\mu} \stackrel{\text{def}}{=} \mu + \nabla\cdot\beta$, we obtain

$$\begin{aligned}\|z_h^u\|_{h,\beta}^2 = a_h(z_h,\pi_h)&\\ &+ \sum_{T\in\mathcal{T}_h}\left[h_T(\beta\cdot\nabla z_h^u, (\beta-\overline{\beta})\cdot\nabla z_h^u)_{L_u,T} + (z_h^\sigma, \kappa\nabla\pi_h^u)_{L_\sigma,T} - (\tilde{\mu} z_h^u, \pi_h^u)_{L_u,T}\right]\\ &+ 2\sum_{F\in\mathcal{F}_h^i}\left[-\frac{1}{4}([\![n^t\kappa z_h^\sigma]\!]_\omega, [\![\pi_h^u]\!])_{L_u,F} + \left(\frac{\beta\cdot n_1}{2}[\![z_h^u]\!], \{\pi_h^u\}\right)_{L_u,F}\right]\\ &- \frac{1}{2}\sum_{F\in\mathcal{F}_h^\partial}\left[(1+\alpha)(n^t\kappa z_h^\sigma, \pi_h^u)_{L_u,F} + ((M_F^{uu}-\beta\cdot n)z_h^u, \pi_h^u)_{L_u,F}\right]\\ &- \sum_{F\in\mathcal{F}_h^i}(S_F([\![z_h^u]\!]), [\![\pi_h^u]\!])_{L_u,F}.\end{aligned}$$

Let R_i, $i=1,\dots,9$, be the nine terms in the right-hand side, and observe that

$$|R_1| \lesssim \mathbb{S}\|\pi_h\|_{h,\kappa,\beta} \lesssim \mathbb{S}\|z_h\|_{h,\kappa,\beta}.$$

Furthermore,

$$|R_2| \lesssim \sum_{T\in\mathcal{T}_h} h_T\|\beta\cdot\nabla z_h^u\|_{L_u,T}\|\overline{\beta}-\beta\|_{[L^\infty(T)]^d} h_T^{-1}\|z_h^u\|_{L_u,T} \lesssim \gamma\|z_h^u\|_{h,\beta}^2 + \|z_h\|_{h,\kappa}^2.$$

Moreover,

$$|R_3| + |R_4| + |R_5| + |R_7| + |R_9| \lesssim \|z_h\|_{h,\kappa}\|\pi_h\|_{h,\kappa,\beta} \lesssim \mathbb{S}^{\frac{1}{2}}\|z_h\|_{h,\kappa,\beta}^{\frac{3}{2}},$$
$$|R_6| \lesssim \|z_h\|_{h,\kappa}\|z_h\|_{h,\kappa,\beta} \lesssim \mathbb{S}^{\frac{1}{2}}\|z_h\|_{h,\kappa,\beta}^{\frac{3}{2}},$$

and $|R_8| \lesssim \|z_h\|_{h,\kappa}(\|\pi_h\|_{h,\kappa,\beta} + \|z_h\|_{h,\kappa,\beta}) \lesssim \mathbb{S}^{\frac{1}{2}}\|z_h\|_{h,\kappa,\beta}^{\frac{3}{2}}$. Hence,

$$\|z_h^u\|_{h,\beta}^2 \lesssim \mathbb{S}\|z_h\|_{h,\kappa,\beta} + \mathbb{S}^{\frac{1}{2}}\|z_h\|_{h,\kappa,\beta}^{\frac{3}{2}},$$

whence it follows, using (4.14), that $\|z_h^u\|_{h,\beta}^2 \lesssim \mathbb{S}^2$. ∎

By using Lemma 4.6 and proceeding as usual, we infer the next result.

THEOREM 4.7 (convergence). *Let z solve (2.28) and z_h solve (3.21). Assume that $z \in [H^1(\mathcal{T}_h)]^{d+1}$ and that κ satisfies (4.12). Then,*

$$\|z - z_h\|_{h,\kappa,\beta} \lesssim [\![z - \pi_h z]\!].$$

Remark 4.3 (purely hyperbolic case). A special situation is obtained when the diffusivity is identically zero over the entire domain, since, for all $T\in\mathcal{T}_h$ and for all $F\in\mathcal{F}_h$, $\mathfrak{h}_T = h_T$ and $\mathfrak{h}_F = h_F$. In such a case it is readily seen that

$$\|z^u - z_h^u\|_{L_u} + \|z^u - z_h^u\|_{h,\beta} \lesssim h^{p_u+\frac{1}{2}}\|z^u\|_{H^{p_u+1}(\mathcal{T}_h)}, \tag{4.16}$$

which is exactly the estimate for the problem investigated in [10, section 3.1].

5. Implementation issues. In this section we discuss important implementation aspects of the method. We show how it can be interpreted in terms of so-called numerical fluxes so as to compare it with other known approximation techniques that are defined in these terms in the literature. We also present two variants of the method that yield substantial computational savings.

5.1. Flux formulation. The notion of (numerical) fluxes is widely used by engineers. This concept originally introduced in the context of finite volume methods, naturally extends to DG methods. The link between DG methods and the concept of flux has been explored in [3] for the Laplace equation and in [11] for more general cases. A number of methods have originally been presented in terms of fluxes, and it is therefore interesting to recast our formulation in this framework so as to facilitate comparisons. To this purpose, let us define

$$\phi^u_{\partial T}(z^\sigma, z^u)|_F \stackrel{\text{def}}{=} \begin{cases} \frac{1+\alpha}{2} n^t \kappa z^\sigma + \frac{1}{2}(\beta\cdot n) z^u + \frac{1}{2} M^{uu}_F z^u & \text{if } F \in \mathcal{F}_h^\partial, \\ n_T^t \{\kappa z^\sigma\}_{\overline{\omega}} + (\beta\cdot n_T)\{z^u\} + (n_T\cdot n_F) S_F([\![z^u]\!]) & \text{if } F \in \mathcal{F}_h^i, \end{cases} \tag{5.1}$$

$$\phi^\sigma_{\partial T}(z^u)|_F \stackrel{\text{def}}{=} \begin{cases} \frac{1-\alpha}{2}(\kappa n)^t z^u & \text{if } F \in \mathcal{F}_h^\partial, \\ (\kappa n)^t|_T \{z^u\}_\omega & \text{if } F \in \mathcal{F}_h^i, \end{cases} \tag{5.2}$$

where $\overline{\omega} \stackrel{\text{def}}{=} (1,1) - \omega$ and n_T is the outward normal to the element T. It is possible to prove (see [11, section 4.3] for the details) that the discrete problem (3.21) can be equivalently reformulated in terms of the following local problems:

$$\begin{cases} \text{Seek } z_h \in W_h \text{ such that, } \forall T \in \mathcal{T}_h \text{ and } \forall q \in [\mathbb{P}_{p_\sigma}(T)]^d \times \mathbb{P}_{p_u}(T), \\ (z_h, (K + \tilde{A})q)_{L,T} + (\phi_{\partial T}(z_h), q)_{L,\partial T} = (f, q^u)_{L_u,T}. \end{cases}$$

The above form is known as the flux formulation of (3.21). Observe that the above flux definitions lead to the use of harmonic averages of the normal component of the diffusion tensor at mesh interfaces; see (3.20). Harmonic averaging of the diffusion matrix has been considered for dual-mixed formulations, e.g., in [2].

5.2. Interior penalty (IP) variant. In this section we discuss a variant of the method designed in section 3.2 which reduces the size of the local problems to be solved to eliminate the σ-component of the unknown. The advantages of such a variant are that it is easier to implement and that the associated matrix pattern is sparser. To this purpose we introduce the lifting operator defined as follows: For all $F \in \mathcal{F}_h$ and for all $\varphi \in L^2(F)$, $r_{F,\kappa}(\varphi) \in \Sigma_h$ is defined by

$$\forall \tau_h \in \Sigma_h, \quad (r_{F,\kappa}(\varphi), \tau_h)_{L_\sigma} \stackrel{\text{def}}{=} \begin{cases} \frac{\alpha+1}{2}(\varphi n, \kappa\tau_h)_{L_\sigma,F} & \text{if } F \in \mathcal{F}_h^\partial, \\ (\varphi n_1, \{\kappa\tau_h\}_{\overline{\omega}})_{L_\sigma,F} & \text{if } F \in \mathcal{F}_h^i. \end{cases} \tag{5.3}$$

Moreover, we let $R_\kappa(\varphi) \stackrel{\text{def}}{=} \sum_{F\in\mathcal{F}_h} r_{F,\kappa}(\varphi)$. Observe that, unlike in [3], the lifting operator depends on the diffusivity. Moreover, for a given face $F \in \mathcal{F}_h$, it is clear that $\text{supp}(r_{F,\kappa}(\varphi)) = \mathcal{T}_h(F)$. In what follows we shall extend the definition of the jump operator to boundary faces by setting

$$[\![y^u]\!] \stackrel{\text{def}}{=} y^u \quad \forall F \in \mathcal{F}_h^\partial, \ \forall y \in W(h).$$

The following result holds.

LEMMA 5.1. *For all $F \in \mathcal{F}_h$ and for all $v_h \in U_h$,*

$$\|r_{F,\kappa}([\![v_h]\!])\|_{L_\sigma} \lesssim \begin{cases} \lambda h_F^{-\frac{1}{2}} \|v_h\|_{L_u,F} & \text{if } F \in \mathcal{F}_h^\partial, \\ \lambda_2 h_F^{-\frac{1}{2}} \|[\![v_h]\!]\|_{L_u,F} & \text{if } F \in \mathcal{F}_h^i. \end{cases}$$

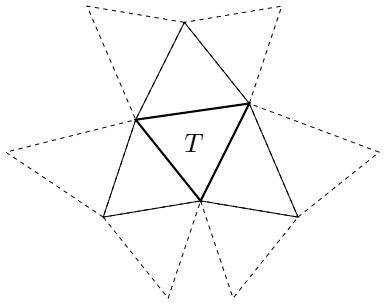


FIG. 1. *Elimination of the σ-component on element T. Stencil for the IP variant of the method (solid lines) and for the LDG variant (solid and dashed lines).*

Proof. Let $F \in \mathcal{F}_h^i$. Then, using (5.3), (3.20), and (3.4), we have that

$$\|r_{F,\kappa}(\llbracket v_h \rrbracket)\|_{L_\sigma}^2 = (\llbracket v_h \rrbracket n_1, \{\kappa r_{F,\kappa}(\llbracket v_h \rrbracket)\}_{\overline{\omega}})_{L_\sigma,F} \lesssim \|\llbracket v_h \rrbracket\|_{L_u,F} \frac{\lambda_1\lambda_2}{2\{\lambda\}} h_F^{-\frac{1}{2}} \|r_{F,\kappa}(\llbracket v_h \rrbracket)\|_{L_\sigma},$$

from which the assertion follows readily. The proof is carried out similarly for $F \in \mathcal{F}_h^\partial$. ∎

Proceeding in a way similar to that in [3, section 3.2] and using the fact that, owing to assumption (3.5), $\kappa\tau_h$ is in Σ_h for all $\tau_h \in \Sigma_h$, it is possible to prove that, for all $(\sigma, u) \in W(h)$ such that $\sigma = \kappa\nabla u - R_\kappa(\llbracket u \rrbracket)$ and for all $(0, v) \in W(h)$,

$$\begin{aligned} a_h((\sigma,u),(0,v)) = &\sum_{T\in\mathcal{T}_h} [(\kappa\nabla u - R_\kappa(\llbracket u\rrbracket), \kappa\nabla v - R_\kappa(\llbracket v\rrbracket))_{L_\sigma,T} + (\mu u, v)_{L_u,T}] \\ &- \sum_{T\in\mathcal{T}_h} (u, \beta\cdot\nabla v)_{L_u,T} + \sum_{F\in\mathcal{F}_h^\partial} \tfrac{1}{2}(M_F^{uu}(u) + (\beta\cdot n)u, v)_{L_u,F} \\ &+ \sum_{F\in\mathcal{F}_h^i} ((\beta\cdot n_1)\{u\}, \llbracket v\rrbracket)_{L_u,F} + \sum_{F\in\mathcal{F}_h^i} (S_F(\llbracket u\rrbracket), \llbracket v\rrbracket)_{L_u,F}. \end{aligned} \tag{5.4}$$

Notice that σ does not appear in the expression on the right-hand side; i.e., we have found a decoupled problem for the sole primal unknown u. The expression (5.4) will henceforth be referred to as the *LDG (local discontinuous Galerkin) variant* of the discrete bilinear form because of the similarity with the method for convection-diffusion systems proposed in [7].

One can verify that, when the basis functions are defined in such a way that their support is restricted to one element of the triangulation, the stencil resulting from (5.4) is composed of all the elements shown in Figure 1 (solid and dashed lines). But by having a closer look at (5.4), one realizes that the only term involving the dashed elements in Figure 1 is the following:

$$\sum_{T\in\mathcal{T}_h} (R_\kappa(\llbracket u\rrbracket), R_\kappa(\llbracket v\rrbracket))_{L_\sigma,T} \stackrel{\text{def}}{=} -\rho_h(u,v).$$

Hence, in order to reduce the stencil, it seems reasonable to consider the following perturbation of a_h:

$$a_h^{\text{IP}}(z,y) \stackrel{\text{def}}{=} a_h(z,y) + \rho_h(z^u, y^u). \tag{5.5}$$

Let us define the following seminorm:

$$|y^u|^2_{\mathrm{LDG}} \stackrel{\mathrm{def}}{=} \frac{\alpha+1}{2}\sum_{F\in\mathcal{F}_h^\partial}\|\lambda h_F^{-\frac{1}{2}} y^u\|^2_{L_u,F} + \sum_{F\in\mathcal{F}_h^i}\|\lambda_2 h_F^{-\frac{1}{2}}[\![y^u]\!]\|^2_{L_u,F} \quad \forall y\in W(h). \tag{5.6}$$

The following lemma is crucial to accommodating the proofs of Lemmata 4.2–4.4 to the new bilinear form a_h^{IP}.

LEMMA 5.2. *The following properties, uniform in h, hold:*

(i) *For all $(z,y_h)\in(V\cap W(h))\times W_h$ we have*

$$\forall y_h\in W_h,\quad \rho_h(z^u,y_h^u)=0.$$

(ii) *For all y in $W(h)$,*

$$\rho_h(y^u,y^u)\le CN_F|y^u|^2_{\mathrm{LDG}},$$

N_F being the maximum number of faces of one mesh element and C a positive parameter depending only on the mesh geometry and on the polynomial order of approximation.

(iii) *For all $(z,y_h)\in W_h^\perp\times W_h$,*

$$\rho_h(z,y_h)\lesssim |z^u|_{\mathrm{LDG}}|y_h^u|_{\mathrm{LDG}}. \tag{5.7}$$

Proof. (i) We know that $[\![z^u]\!]=0$, and consequently $r_{F,\kappa}([\![z^u]\!])=0$, on all $F\in\mathcal{F}_h\setminus\mathcal{I}_h^-$. On the other hand, let $\mathcal{I}_h^-\ni F=\partial T_1\cap\partial T_2$ and $\tau_h\in\Sigma_h$. Then, since $n^t\kappa|_{T_2}=0$ entails $\lambda_2=0$,

$$n_1^t\{\kappa\tau_h\}_{\overline{\omega}} = \frac{\lambda_2}{2\{\lambda\}}n_1^t\cdot\kappa\tau_h|_{T_1}+\frac{\lambda_1}{2\{\lambda\}}n_1^t\cdot\kappa\tau_h|_{T_2}=0,$$

i.e., $r_{F,\kappa}([\![z^u]\!])=0$, which gives the desired result.

(ii) The second point can be proved as follows. Observe that

$$\|R_\kappa([\![y^u]\!])\|^2_{L_\sigma}\le\sum_{F\in\mathcal{F}_h}\sum_{F'\in\mathcal{F}_h}\|r_{F,\kappa}([\![y^u]\!])\|_{L_\sigma}\|r_{F',\kappa}([\![y^u]\!])\|_{L_\sigma}.$$

Let $\mathcal{F}_h^i\ni F=\partial T_1\cap\partial T_2$. Since $\mathrm{supp}(r_{F,\kappa}([\![y^u]\!]))=T_1\cup T_2$, only a few products on the right-hand side are nonzero. In particular, the nonzero products are those for which $F'\in\Delta_F$, where $\Delta_F\stackrel{\mathrm{def}}{=}\{F'\in\mathcal{F}_h;\ F'\subset\partial T_1 \text{ or } F'\subset\partial T_2\}$. Therefore, the only terms involving F are

$$\sum_{F'\in\Delta_F}\|r_{F,\kappa}([\![y^u]\!])\|_{L_\sigma}\|r_{F',\kappa}([\![y^u]\!])\|_{L_\sigma}\le\tfrac{1}{2}\sum_{F'\in\Delta_F}\left(\|r_{F,\kappa}([\![y^u]\!])\|^2_{L_\sigma}+\|r_{F',\kappa}([\![y^u]\!])\|^2_{L_\sigma}\right).$$

We realize that $\|r_{F,\kappa}([\![y^u]\!])\|^2_{L_\sigma}$ is added at most N_F times. The desired result follows by repeating this argument for the other faces and using Lemma 5.1.

(iii) Deriving (5.7) is a simple application of Lemma 5.1. ∎

Modifying Lemmata 4.2–4.4 so as to hold for a_h^{IP} instead of a_h is now simple in view of the above result. However, observe that, according to the second point of Lemma 5.2, in order to preserve the L-coercivity, (3.6) and (3.18) should be modified as follows:

$$M_F^{uu}(v)\stackrel{\mathrm{def}}{=}\left(\frac{|\beta\cdot n|}{2}+N_F\eta\frac{\alpha+1}{2}\frac{\lambda^2}{h_F}\right)v,\qquad S_F(v)\stackrel{\mathrm{def}}{=}\left(\frac{|\beta\cdot n|}{2}+N_F\eta\frac{\lambda_2^2}{h_F}\right)v, \tag{5.8}$$

where the multiplicative factor η must be strictly greater than the constant C appearing in (5.7). The term ρ_h that has been added to simplify the elimination of the σ-component is thus counterbalanced by adding "more stabilization." The resulting method is termed the *IP variant* because of the similarity with the IP method proposed in [1, 4]. The method recently proposed in [15] also belongs to this class, although some modifications are introduced in the definition of the penalty parameter.

5.3. BRMPS variant. The parameter C in (5.7), and, consequently, η in (5.8), is possibly difficult to estimate in practical applications. To avoid this difficulty, we consider the following alternative expression for the boundary and interface operators:

$$M_F^{uu}(v) \stackrel{\text{def}}{=} \frac{|\beta\cdot n|}{2}v + N_F\eta r_{F,\kappa}(v), \qquad S_F(v) \stackrel{\text{def}}{=} \frac{|\beta\cdot n|}{2}v + N_F\eta\{r_{F,\kappa}(v)\}_{\overline{\omega}}. \tag{5.9}$$

A closer look at the proof of the second point of Lemma 5.2 shows that it is sufficient to take $\eta > 1$ to preserve L-coercivity. Owing to the similarities with the approach first presented in [5], the resulting numerical method is termed the *BRMPS variant* (after the initials of the authors of [5]).

6. Numerical results. In this section we evaluate the performance of the proposed method on two test cases. The simulations are done using the BRMPS variant discussed in section 5.3. In both test cases, we assume that the exact solution is smooth enough elementwise.

6.1. Convergence. In order to assess the theoretical convergence estimates, we consider the problem described in Figure 2(a). Here (r,θ) denote the standard cylindrical coordinates with the angle θ measured in the anticlockwise sense starting from the positive x-axis. The domain is taken to be $(-1,1)^2 \setminus [-0.5,0.5]^2$, while the coefficients are set to

$$\kappa = \begin{cases} \pi & \text{if } 0 < \theta < \pi, \\ 0 & \text{if } \pi < \theta < 2\pi, \end{cases} \qquad \beta = \frac{e_\theta}{r}, \qquad \mu = 10^{-3},$$

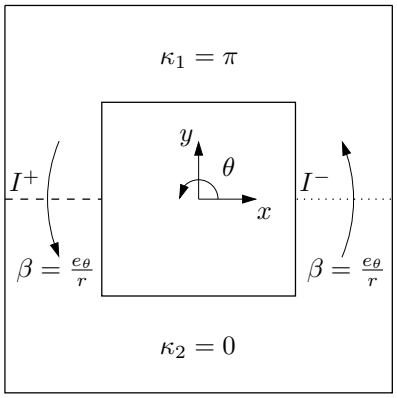
where e_θ is the unit azimuthal vector. The exact solution for a suitable right-hand side f is

$$u = \begin{cases} (\theta-\pi)^2 & \text{if } 0 \le \theta \le \pi, \\ 3\pi(\theta-\pi) & \text{if } \pi < \theta < 2\pi. \end{cases}$$

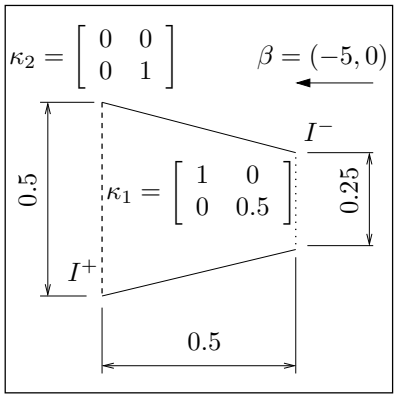
Observe that, although piecewise polynomial in θ, the above solution does *not* belong to the discrete space U_h since we are solving the problem in Cartesian coordinates. Moreover, according to the interface condition (INT1), the solution is continuous across I^+, while only (INT2) is verified on I^-. We introduce the following norm:

$$\|u\|_{h,\text{BRMPS}}^2 \stackrel{\text{def}}{=} \|u\|_{L_u}^2 + |u|_J^2 + |u|_M^2 + \sum_{T\in\mathcal{T}_h} \|\kappa\nabla u\|_{L_\sigma,T}^2.$$

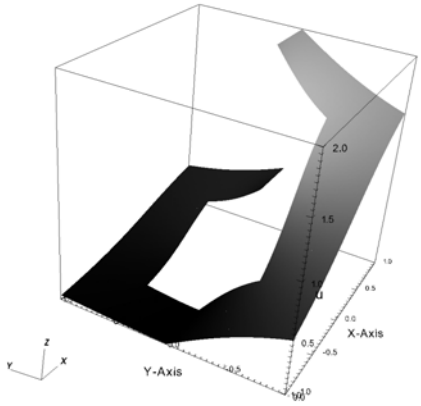
Let (σ_h, u_h) solve the discrete problem associated with the BRMPS variant. Then, observing that $\sigma_h = \kappa\nabla u_h + R([\![u_h]\!])$, it can be proved that $\|u-u_h\|_{h,\text{BRMPS}}$ is equivalent to $\|(\sigma,u)-(\sigma_h,u_h)\|_{h,\kappa}$. Coherently with the desire to avoid the additional cost coming from the computation of σ_h, $\|u-u_h\|_{h,\text{BRMPS}}$ was reported in Table 1. The convergence results confirm the sharpness of the estimates derived in sections 4.1 and 4.2. The L^2-norm is also reported for completeness, showing that convergence at order p_u+1 can be expected.



(a) Description of the test case of section 6.1.



(b) Description of the test case of section 6.2.



(c) Exact solution of the test case of section 6.1.

FIG. 2. *Problem setting for the numerical test cases. I^+ and I^- are plotted as dashed and dotted lines, respectively.*

6.2. Strongly anisotropic diffusivity. To demonstrate the behavior of the method in the presence of strongly anisotropic diffusivity we consider the test of Figure 2(b). The domain $\Omega = (0,1)^2$ is partitioned into two subdomains where the diffusivity takes different values; it is positive definite in one region and positive semidefinite in the other region. The advection field is $\beta = (-5,0)^t$, and the reaction coefficient is $\mu = 1$. The solution is discontinuous across the interface $I^- = \{x = 0.75; 0.375 \leq y \leq 0.625\}$. The solutions obtained for different polynomial degrees are displayed in Figure 3, showing that the predicted behavior is captured accurately.

7. Conclusion. In this work we developed and analyzed a DG method to approximate advection-diffusion-reaction equations with discontinuous, anisotropic, and semidefinite diffusivity. The proposed method is capable of treating in a robust fashion the semidefinite diffusivity case owing to our design of the boundary and penalty terms and provided the mesh fits the discontinuities of the diffusivity field. This is

TABLE 1
Convergence results.

h	$P_{h,1}$		$P_{h,2}$		$P_{h,3}$		$P_{h,4}$	
	Err	Ord	Err	Ord	Err	Ord	Err	Ord
$\|u-u_h\|_{h,\mathrm{BRMPS}}$								
1/2	3.15e+0		7.27e−1		1.74e−1		3.99e−2	
1/4	1.63e+0	0.95	2.05e−1	1.83	2.69e−2	2.70	3.51e−3	3.51
1/8	8.19e−1	0.99	5.32e−2	1.94	3.59e−3	2.91	2.51e−4	3.81
1/16	4.08e−1	1.00	1.34e−2	1.99	4.54e−4	2.98	1.63e−5	3.95
1/32	2.04e−1	1.00	3.36e−3	2.00				
$\|u-u_h\|_{h,\beta}$								
1/2	1.97e−0		4.50e−1		1.13e−1		2.65e−2	
1/4	7.46e−1	1.40	9.87e−2	2.18	1.40e−2	3.01	1.92e−3	3.79
1/8	2.73e−1	1.45	1.90e−2	2.38	1.44e−3	3.29	1.06e−4	4.18
1/16	9.82e−2	1.48	3.44e−3	2.46	1.34e−4	3.43	5.03e−6	4.40
1/32	3.50e−2	1.49	6.08e−4	2.50				
$\|u-u_h\|_{L_u}$								
1/2	2.92e−1		3.30e−2		5.79e−3		1.17e−3	
1/4	7.49e−2	1.96	4.75e−3	2.80	4.62e−4	3.65	5.50e−5	4.41
1/8	1.91e−2	1.97	6.09e−4	2.96	3.26e−5	3.83	2.01e−6	4.77
1/16	4.86e−3	1.97	7.76e−5	2.97	2.10e−6	3.96	6.32e−8	4.99
1/32	1.23e−3	1.98	9.82e−6	2.98				

achieved by resorting to weighted average and jump operators. The convergence analysis yields estimates that are uniform with respect to the diffusivity. The theoretical results are supported by numerical evidence.

Appendix. Proof of Proposition 2.4.

Proof. According to the so-called Banach–Nečas–Babuška (BNB) theorem (see, e.g., [9, section 2.1.3]), the statement amounts to proving that the following conditions hold:

$$\text{(BNB1)} \qquad \forall z \in V, \quad \sup_{y \in L\setminus\{0\}} \frac{a_0(z,y)}{\|y\|_L} \gtrsim \|z\|_V,$$

$$\text{(BNB2)} \qquad \forall z \in V, \quad (\forall y \in L, \quad a_0(z,y) = 0) \Longrightarrow (y = 0).$$

(i) Let us prove (BNB1). Let $z \in V$, and set $\mathbb{S} \overset{\text{def}}{=} \sup_{y\in L\setminus\{0\}} \frac{a_0(z,y)}{\|y\|_L}$. Using the definition of the L^2-norm, we deduce

$$\mathbb{S} \gtrsim \sup_{y\in L\setminus\{0\}} \frac{(Az,y)_L}{\|y\|_L} - \|z\|_L \gtrsim \|Az\|_L - \|z\|_L.$$

Then Lemma 2.3 gives

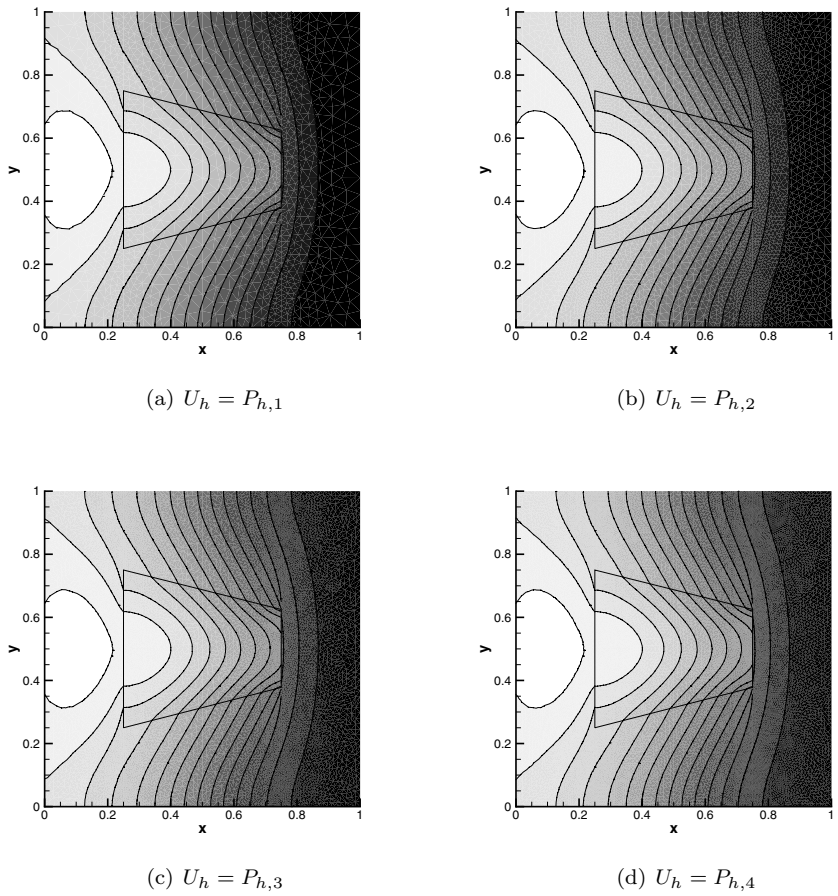
$$\|z\|_L \lesssim \frac{a_0(z,z)}{\|z\|_L} \lesssim \mathbb{S} \Longrightarrow \|z\|_L + \|Az\|_L \lesssim \mathbb{S},$$

i.e., $\|z\|_V \lesssim \mathbb{S}$, which proves (BNB1).

(ii) Let us prove (BNB2). Let $y \in L$ be such that

$$\text{(A.1)} \qquad a_0(z,y) = 0 \quad \forall z \in V.$$

Let us prove that y is necessarily zero.



(a) $U_h = P_{h,1}$ (b) $U_h = P_{h,2}$

(c) $U_h = P_{h,3}$ (d) $U_h = P_{h,4}$

FIG. 3. *Numerical results for the test case of section* 6.2.

(1) Let $\Omega_i \in P_\Omega$, take $z = (z^\sigma, 0)$ with $z^\sigma \in [\mathcal{D}(\Omega_i)]^d$, and observe that z is a member of V. Then using z to test (A.1), we obtain

$$a_0((z^\sigma, 0), y) = \langle y^\sigma - \kappa \nabla y^u, z^\sigma \rangle_{[\mathcal{D}(\Omega_i)]^d} = 0 \quad \forall z^\sigma \in [\mathcal{D}(\Omega_i)]^d,$$

meaning that $y^\sigma - \kappa \nabla y^u = 0$ in Ω_i, i.e., $\kappa \nabla y^u|_{\Omega_i} \in L^2(\Omega_i)$. Now let us prove that y^u is continuous across $\Gamma \setminus I$. Let x be a point in $\Gamma \setminus I$. Let Ω_i and Ω_j be the two subdomains that are on each side of the interface $\Gamma \setminus I$ at x. We assume that x is an interior point of $\overline{\Omega_i} \cap \overline{\Omega_j}$; i.e., all the points in a small neighborhood of x, say $\mathcal{V}$, belong to either $\overline{\Omega_i}$ or $\overline{\Omega_j}$. Up to a Lipschitz map, we can assume that the restriction of Γ to $\mathcal{V}$ is a hyperplane. We choose a local Cartesian coordinate system $(x_1, \ldots, x_d)$ so that the two normals n_1, n_2 are aligned with the x_1-axis. The neighborhood $\mathcal{V}$ can be chosen small enough so that $\mathcal{V} \cap I = \emptyset$. κ being piecewise constant, this means that $\kappa \cdot n_1$ is uniformly bounded away from zero in $\mathcal{V}$; in other words, the component κ_{11} is uniformly bounded away from zero in $\mathcal{V}$.

Let $\psi \in \mathcal{D}(\mathcal{V})$ and define the d-vector field $\tilde{\sigma}$ by $\tilde{\sigma}_1 = \frac{1}{\kappa_{11}}\psi$, and $\tilde{\sigma}_l = 0$ for $l > 1$. Then $\kappa\tilde{\sigma} = (1, \frac{\kappa_{21}}{\kappa_{11}}, \ldots, \frac{\kappa_{d1}}{\kappa_{11}})\psi$. Observe that $\tilde{\sigma} \in L_\sigma$ and $\nabla \cdot (\kappa\tilde{\sigma}) \in L^2(\Omega)$; in other

words, the pair $(\tilde{\sigma}, 0)$ is in V. By testing (A.1) with $(\tilde{\sigma}, 0)$, we obtain

$$\begin{aligned} 0 &= \int_{\mathcal{V}} [\tilde{\sigma}\cdot y^{\sigma} + \nabla\cdot(\kappa\tilde{\sigma}) y^{u}] \\ &= -\langle \psi, \partial_{x_1} y^{u}\rangle_{\mathfrak{D}(\Omega)} + \int_{\mathcal{V}} \psi \frac{1}{\kappa_{11}} (y_1^{\sigma} - \kappa_{12}\partial_{x_2} y^{u} - \cdots - \kappa_{1d}\partial_{x_d} y^{u}), \end{aligned}$$

where we used that $\kappa_{12}\partial_{x_2} y^{u}, \ldots, \kappa_{1d}\partial_{x_d} y^{u}$ are in $L^2(\mathcal{V})$ from the first part of the argument. A standard distribution argument implies

$$\partial_{x_1} y^{u} = \frac{1}{\kappa_{11}} (y_1^{\sigma} - \kappa_{12}\partial_{x_2} y^{u} - \cdots - \kappa_{1d}\partial_{x_d} y^{u}) \in L^2(\mathcal{V}).$$

Then we conclude that $\partial_{x_1} y^{u} \in L^2(\mathcal{V})$; i.e., y^{u} is continuous across Γ in a neighborhood of x. Using a standard argument together with x being an interior point of $\overline{\Omega_i} \cap \overline{\Omega_j}$, we infer that $\kappa\nabla y^{u}$ is in L_{σ}.

(2) Use $z = (0, z^{u})$ with $z^{u} \in \mathfrak{D}(\Omega)$ as a test function in (2.22), and observe that again z is a member of V. A distributional argument gives

$$\langle (\mu + \nabla\cdot\beta) y^{u} - \nabla\cdot(\kappa y^{\sigma} + \beta y^{u}), z^{u}\rangle_{\mathfrak{D}(\Omega)} = 0.$$

Owing to the regularity assumptions on μ and β listed in section 2.1, we conclude that $\nabla\cdot(\kappa y^{\sigma} + \beta y^{u}) \in L^2(\Omega)$, i.e., y is a member of W and

$$(K + \tilde{A}) y = 0.$$

(3) We then deduce that, for all $z \in V$,

$$\langle Dz, y\rangle_{W',W} = ((K + A) z, y)_L - ((K + \tilde{A}) y, z)_L = 0;$$

i.e., y is a member of $D(V)^{\perp} = V^*$. In conclusion, $a_0^*(y, w) = 0$ for all $w \in L$ and $y \in V^*$. Finally, the L-coercivity of a_0^* (see Lemma 2.3) implies that $y = 0$. ∎

REFERENCES

[1] D. N. Arnold, *An interior penalty finite element method with discontinuous elements*, SIAM J. Numer. Anal., 19 (1982), pp. 742–760.

[2] D. N. Arnold and F. Brezzi, *Mixed and nonconforming finite element methods: Implementation, postprocessing and error estimates*, RAIRO Modél. Math. Anal. Numér., 19 (1985), pp. 7–32.

[3] D. N. Arnold, F. Brezzi, B. Cockburn, and L. D. Marini, *Unified analysis of discontinuous Galerkin methods for elliptic problems*, SIAM J. Numer. Anal., 39 (2002), pp. 1749–1779.

[4] G. A. Baker, *Finite element methods for elliptic equations using nonconforming elements*, Math. Comp., 31 (1977), pp. 45–49.

[5] F. Bassi, S. Rebay, G. Mariotti, S. Pedinotti, and M. Savini, *A high-order accurate discontinuous finite element method for inviscid and viscous turbomachinery flows*, in Proceedings of the 2nd European Conference on Turbomachinery Fluid Dynamics and Thermodynamics, R. Decuypere and G. Dibelius, eds., 1997, pp. 99–109.

[6] E. Burman and P. Zunino, *A domain decomposition method based on weighted interior penalties for advection-diffusion-reaction problems*, SIAM J. Numer. Anal., 44 (2006), pp. 1612–1638.

[7] B. Cockburn and C.-W. Shu, *The local discontinuous Galerkin method for time-dependent convection-diffusion systems*, SIAM J. Numer. Anal., 35 (1998), pp. 2440–2463.

[8] J.-P. Croisille, A. Ern, T. Lelièvre, and J. Proft, *Analysis and simulation of a coupled hyperbolic/parabolic model problem*, J. Numer. Math., 13 (2005), pp. 81–103.

[9] A. Ern and J.-L. Guermond, *Theory and Practice of Finite Elements*, Appl. Math. Sci. 159, Springer-Verlag, New York, 2004.

[10] A. Ern and J.-L. Guermond, *Discontinuous Galerkin methods for Friedrichs' systems.* I. *General theory*, SIAM J. Numer. Anal., 44 (2006), pp. 753–778.

[11] A. Ern and J.-L. Guermond, *Discontinuous Galerkin methods for Friedrichs' systems. Part* II. *Second-order elliptic PDEs*, SIAM J. Numer. Anal., 44 (2006), pp. 2363–2388.

[12] A. Ern and J.-L. Guermond, *Discontinuous Galerkin methods for Friedrichs' systems. Part* III. *Multi-field theories with partial coercivity*, SIAM J. Numer. Anal., 46 (2008), pp. 776–804.

[13] A. Ern, J.-L. Guermond, and G. Caplain, *An intrinsic criterion for the bijectivity of Hilbert operators related to Friedrichs' systems*, Comm. Partial Differential Equations, 32 (2007), pp. 317–341.

[14] A. Ern and J. Proft, *Multi-algorithmic methods for coupled hyperbolic-parabolic problems*, Int. J. Numer. Anal. Model., 1 (2006), pp. 94–114.

[15] A. Ern, A. F. Stephansen, and P. Zunino, *A discontinuous Galerkin method with weighted averages for advection-diffusion equations with locally vanishing and anisotropic diffusivity*, IMA J. Numer. Anal., to appear.

[16] J. Freund and R. Stenberg, *On weakly imposed boundary conditions for second order problems*, in Proceedings of the International Conference on Finite Elements in Fluids—New Trends and Applications, Venice, Italy, 1995, M. Morandi Cecchi et al., eds., pp. 327–336.

[17] K. O. Friedrichs, *Symmetric positive linear differential equations*, Comm. Pure Appl. Math., 11 (1958), pp. 333–418.

[18] F. Gastaldi and A. Quarteroni, *On the coupling of hyperbolic and parabolic systems: Analytical and numerical approach*, Appl. Numer. Math., 6 (1989), pp. 3–31.

[19] B. Heinrich and S. Nicaise, *The Nitsche mortar finite-element method for transmission problems with singularities*, IMA J. Numer. Anal., 23 (2003), pp. 331–358.

[20] B. Heinrich and K. Pietsch, *Nitsche type mortaring for some elliptic problem with corner singularities*, Computing, 68 (2002), pp. 217–238.

[21] B. Heinrich and K. Pönitz, *Nitsche type mortaring for singularly perturbed reaction-diffusion problems*, Computing, 75 (2005), pp. 257–279.

[22] P. Houston, C. Schwab, and E. Süli, *Discontinuous hp-finite element methods for advection-diffusion-reaction problems*, SIAM J. Numer. Anal., 39 (2002), pp. 2133–2163.

[23] P. Lesaint, *Finite element methods for symmetric hyperbolic equations*, Numer. Math., 21 (1973/74), pp. 244–255.

[24] P. Lesaint and P.-A. Raviart, *On a finite element method for solving the neutron transport equation*, in Mathematical Aspects of Finite Elements in Partial Differential Equations, Publication 33, Math. Res. Center, University of Wisconsin-Madison, Academic Press, New York, 1974, pp. 89–123.

[25] W. H. Reed and T. R. Hill, *Triangular Mesh Methods for the Neutron Transport Equation*, Technical Report LA-UR-73-479, Los Alamos Scientific Laboratory, Los Alamos, NM, 1973.

[26] M. F. Wheeler, *An elliptic collocation-finite element method with interior penalties*, SIAM J. Numer. Anal., 15 (1978), pp. 152–161.

SIAM J. NUMER. ANAL.
Vol. 46, No. 2, pp. 832–868
© 2008 Society for Industrial and Applied Mathematics

LOCAL REGULARIZATION FOR THE NONLINEAR INVERSE AUTOCONVOLUTION PROBLEM*

ZHEWEI DAI† AND PATRICIA K. LAMM‡

Abstract. We develop a local regularization theory for the nonlinear inverse autoconvolution problem. Unlike classical regularization techniques such as Tikhonov regularization, this theory provides regularization methods that preserve the causal nature of the autoconvolution problem, allowing for fast sequential numerical solution ($\mathcal{O}(rN^2 - r^2N)$ flops, where $r \ll N$ for the method discussed in this paper as applied to the nonlinear problem; in comparison, the cost for Tikhonov regularization applied to a general *linear* problem is $\mathcal{O}(N^3)$ flops). We prove the convergence of the regularized solutions to the true solution as the noise level in the data shrinks to zero and supply convergence rates for the cases of both L_2 and continuous data. We propose several regularization methods and provide a theoretical basis for their convergence; of note is that this class of methods does not require an initial guess of the unknown solution. Our numerical results confirm the effectiveness of the methods, with results comparing favorably to numerical examples found in the literature for the autoconvolution problem (e.g., [G. Fleischer, R. Gorenflo, and B. Hofmann, *ZAMM Z. Angew. Math. Mech.*, 79 (1999), pp. 149–159] for examples using Tikhonov regularization with total variation constraints and [J. Janno, *Inverse Problems*, 16 (2000), pp. 333–348] for examples using the method of Lavrent'ev); this especially seems to be true when it comes to the recovery of sharp features in the unknown solution. We also show the effectiveness of our method in cases not covered by the theory.

Key words. local regularization of inverse problems, nonlinear Volterra problem, numerical methods, inverse autoconvolution problem

AMS subject classifications. 65R20, 65R30, 65R32, 45Q05

DOI. 10.1137/070679247

1. Introduction.

In this paper we develop a local regularization theory for the inverse autoconvolution problem of finding $\bar{x} \in L_2(0,T)$ solving

$$G(x) = f, \tag{1.1}$$

where G is the nonlinear Volterra operator given by

$$G(x)(t) = \int_0^t x(t-s)\,x(s)\,ds, \quad \text{a.e. } t \in (0,T), \tag{1.2}$$

and where $f \in \mathit{Range}(G) \subseteq L_2(0,T)$. Without loss of generality we will henceforth let $T = 1$.

Autoconvolution has been of interest to scientists for decades because of its applications in various fields. It arises in stochastics, where the density function of a continuous random variable V is reconstructed after observing the density function of the random variable $S = V_1 + V_2$, where V_1 and V_2 are identically and independently distributed random variables of V [14]. Another application of autoconvolution occurs in spectroscopy. Baumeister presents in [1] a reference list of physically motivated papers concerning this class of problems. He also discusses in detail the mathematical

*Received by the editors January 5, 2007; accepted for publication (in revised form) July 18, 2007; published electronically February 20, 2008. This work was supported in part by the National Science Foundation under contracts NSF DMS-0104003 and DMS-0405978.
http://www.siam.org/journals/sinum/46-2/67924.html
†Department of Mathematics, Alma College, Alma, MI 48801 (dai@alma.edu).
‡Department of Mathematics, Michigan State University, East Lansing, MI 48824-1027 (lamm@math.msu.edu).

model of deconvolution of "appearance potential"(AP) spectra to investigate electronic properties of solids in their surface region. In this context, the density of unoccupied states in the surface region of a solid is recovered from the measured AP-spectrum data.

Properties of the autoconvolution problem have been explored extensively in [12, 13, 11, 14] (to name just a few of the references in this area) and include analysis of the ill-posedness of the problem due to the lack of continuous dependence of solutions $\bar{x}$ on data f. Indeed, for $D(G) \equiv \{x \in L_2(0,1),\ x(t) \geq 0,\ \text{a.e. } t \in (0,1)\}$, the operator $G : D(G) \subset L_2(0,1) \mapsto L_2(0,1)$ is such that the inverse autoconvolution operator G^{-1} is discontinuous at every point f in the range of G; i.e., the autoconvolution equation is locally ill-posed at every point x in $D(G)$. Degrees of ill-posedness for the autoconvolution equation are discussed in [13, 14]. In short, we expect a correlation between the degree of ill-posedness and both the smoothness of the solution $\bar{x}$ and the behavior of $\bar{x}$ at 0.

Various regularization methods have been studied for the autoconvolution equation. One can utilize the Tikhonov regularization theory for nonlinear inverse problems, since the autoconvolution operator G is continuous and weakly closed on $D(G)$ and G has a compact Fréchet derivative at all $x \in L_2(0,1)$ satisfying assumptions needed to guarantee stability in the nonlinear Tikhonov theory [9, 14, 30]. However, as is well known in the case of the linear Volterra problem, classical regularization methods such as Tikhonov regularization have inherent disadvantages in solving Volterra problems; indeed, the use of Tikhonov regularization transforms problems that are structurally causal in nature into noncausal or full-domain problems (see, e.g., [21] for a discussion of these difficulties). Causal-based methods have advantages because they are sequential in nature, and we will see in section 5.1 that the cost of the local regularization algorithm discussed in that section is $\mathcal{O}(rN^2 - r^2N)$, where $r \ll N$, while the cost for Tikhonov regularization on a general *linear* problem is $\mathcal{O}(N^3)$.

In addition, source conditions needed for convergence rates for Tikhonov regularization as applied to the nonlinear problem (1.1) require that $\bar{x} - x^\star = G'(\bar{x})^\star w$ for some $w \in L_2(0,1)$ [10], where $x^\star$ is an initial guess for $\bar{x}$ and $\bar{x}$ is the true solution of (1.1). It is not hard to show that the source condition requires that $x^\star(1) = \bar{x}(1)$; since $x^\star$ is part of the Tikhonov algorithm, the method then requires knowledge of the value of the unknown solution $\bar{x}$ at $t = 1$.

In the late 1960s, Beck developed a regularization scheme for the discretized inverse heat conduction problem (IHCP), a linear Volterra inverse problem, which retains the causal nature of the original problem [2]. Because of its sequential formulation, the numerical implementation of the Beck method is also more efficient than classical regularization methods. In the mid-1990s, the second author established the theoretical basis for the convergence of the sequential local regularization method, and Beck's approach was generalized to a wide class of linear first-kind Volterra problems [17, 18, 19]. While Beck's method was an approach developed to handle a finite-dimensional approximation of the IHCP, the current theory of local regularization methods can be placed in both finite- and infinite-dimensional settings. In addition, the theory has been extended more recently to nonlinear Volterra problems [29] of the Hammerstein type,

$$\int_0^t k(t-s)g(x(s),s)\,ds = f(t), \quad \text{a.e. } t \in (0,T). \tag{1.3}$$

In [24], the special structure of this equation was coupled with the linear convergence

theory so that local regularization was used to solve

$$\int_0^t k(t-s)v(s)\,ds = f(t), \quad \text{a.e. } t \in (0,T),$$

for v, and then u, was recovered as a solution of the nonlinear problem $v(t) = g(u(t), t)$ for all $t \in [0,1]$. More recently in [26] this work has been modified so that the nonlinear equation (1.3) need only be solved on an initial (small) interval $[0,\varepsilon]$; then for $t > \varepsilon$, the solution x is stably recovered explicitly and sequentially from a local regularization equation which is an extension of the method found in [24].

In [15, 16], Janno applied the Lavrent'ev regularization method to the autoconvolution equation and provided both theoretical convergence estimates and numerical examples. The advantage of the Lavrent'ev method is that it preserves the causal nature of Volterra problems and therefore leads to a fast sequential method. The drawback for the method is that it requires and depends on an initial guess $x^\star$ of the true solution $\bar{x}$, and it must be that $x^\star(0) = \bar{x}(0)$, which means that $\bar{x}(0)$ must be known in advance (or at least computed using regularized differentiation of noisy data; we note that in section 5 we propose an algorithm which approximates $\bar{x}(0)$ in exactly the same way). The performance of both the Lavrent'ev and Tikhonov methods seems fairly dependent on the choice of the initial guess $x^\star$ and of its closeness to $\bar{x}$.

In this paper, we develop a local regularization theory for the nonlinear autoconvolution problem of finding $x \in D(G) \subset L_2(0,1)$ solving (1.1) with G given by (1.2). The result is an effective regularization method which essentially preserves the causal nature of the problem and does not require an initial guess of the true solution (such as is needed for the Lavrent'ev and Tikhonov methods).

2. The regularized autoconvolution equation. We will first motivate the equation which will become the basis for our regularized approximation of the true solution $\bar{x}$ of (1.1). We assume that our data and solution are available on an extended interval $[0, 1+\bar{R}]$ for some small $\bar{R} \in (0,1]$ and note that this can always be accomplished by simply decreasing the size of our original interval slightly. Thus, for any $0 < R < \bar{R}$, $\bar{x}$ solves

$$\int_0^{t+\rho} x(t+\rho-s)\,x(s)\,ds = f(t+\rho), \quad \text{a.e. } t \in (0,1),\ \rho \in (0,R).$$

We split the integral at ρ and t and then change the variable of integration to obtain

$$2\int_0^\rho x(t+\rho-s)\,x(s)\,ds + \int_\rho^t x(t+\rho-s)\,x(s)\,ds = f(t+\rho) \tag{2.1}$$

for a.e. $t \in (0,1), \rho \in (0,R)$. In order to consolidate the local future information introduced by the variable ρ, we follow the idea of local regularization in the case of the linear Volterra problem (see, e.g., [17, 18]) and integrate both sides of (2.1) with respect to a suitable measure $\eta = \eta(\rho) > 0$ (which we will clarify later) on $[0,R]$, leading to the equation

$$\begin{aligned} 2\int_0^R \int_0^\rho x(t+\rho-s)\,x(s)\,ds\,d\eta(\rho) + \int_0^R \int_\rho^t x(t+\rho-s)\,x(s)\,ds\,d\eta(\rho) \\ = \int_0^R f(t+\rho)\,d\eta(\rho), \quad \text{a.e. } t \in (0,1). \end{aligned} \tag{2.2}$$

Note that $\bar{x}$ still satisfies (2.2) exactly.

In reality, instead of having the exact data f, we have access only to some approximation f^δ to f. Because of the local ill-posedness of the original problem (1.1), some form of regularization is needed when the perturbed data f^δ are used in place of the "ideal" data f. The results of regularization differ depending on the smoothness of the perturbed data, so we will investigate the properties of our method using data f^δ in two different spaces $\mathcal{F}$ of functions defined on $[0, 1+\bar{R}]$. Our assumption regarding data will be as follows.

DEFINITION 2.1. *We will say the data f^δ satisfy the* $\mathcal{F}$-data assumption *if $f^\delta \in \mathcal{F}$ and*

$$\|f^\delta - f\|_{\mathcal{F}} \leq \delta \tag{2.3}$$

for some $\delta > 0$, where either $\mathcal{F} = C[0, 1+\bar{R}]$ or $\mathcal{F} = L_2(0, 1+\bar{R})$.

Returning then to local regularization in the presence of perturbed data f^δ, the method is suggested by the idea of momentarily holding x constant on a small local interval $[t, t+R]$; i.e., we replace $x(t+\rho-s)$ by $x(t)$ in the first term of (2.2) for values of ρ, s such that $\rho - s \in [0, R]$. Here the length of this local interval R serves as the regularization parameter. We then obtain the regularization equation

$$\alpha_R(x)x + F_R(x) = f_R^\delta, \tag{2.4}$$

where, for a.e. $t \in (0,1)$,

$$\alpha_R(x) \equiv 2\int_0^R \int_0^\rho x(s)\, ds\, d\eta(\rho), \tag{2.5}$$

$$F_R(x)(t) \equiv \int_0^R \int_\rho^t x(t+\rho-s)\, x(s)\, ds\, d\eta(\rho), \tag{2.6}$$

$$f_R^\delta(t) \equiv \int_0^R f^\delta(t+\rho)\, d\eta(\rho). \tag{2.7}$$

The R-dependent measure η is defined via

$$\int_0^R g(\rho)\, d\eta(\rho) \equiv \int_0^R g(\rho)\, \omega(\rho, R)\, d\rho, \quad g \in L_2(0,R), \tag{2.8}$$

where the family $\{\omega(\cdot, R) \in L_\infty(0,R)\}_{R\in(0,\bar{R}]}$ is assumed to be such that there exist $\underline{\omega}$, $\bar{\omega} > 0$ independent of R so that

$$0 < \underline{\omega} \leq \omega(\rho, R) \leq \bar{\omega}, \quad \text{a.e. } \rho \in (0, R], \tag{2.9}$$

for all R sufficiently small. It follows then that for any real $m \geq 0$ there exist constants $K(m) > 0$ independent of R so that

$$\frac{\int_0^R \rho^m\, d\eta(\rho)}{\int_0^R \rho\, d\eta(\rho)} \leq K(m) R^{m-1} \tag{2.10}$$

for all $R > 0$ sufficiently small. We note that this class of measures is a special case of the measures discussed in previous papers on local regularization for linear Volterra problems (see, e.g., [5, 6, 7, 17, 18, 19, 20, 22, 23, 25, 27, 28, 31]) and that the discrete

measures discussed in these references are not suitable for use in (2.7) in the case of $f^\delta \in L_2(0, 1+R)$.

Our main results involving the well-posedness of (2.4) and the convergence of the solution x of (2.4) to the true solution $\bar{x}$ of the original autoconvolution equation will be given in the next section. First we will introduce some additional spaces and norms, which will facilitate further analysis below of the operator F_R, as well as the theory given in section 3.

We will define a new R-dependent topology on $L_2(0,1)$ via

$$\langle x, y\rangle_{\sigma,R} \equiv \frac{\mathcal{C}^2 - 1}{R}\int_0^R e^{-2\sigma t}x(t)\,y(t)\,dt + \int_0^1 e^{-2\sigma t}x(t)\,y(t)\,dt, \tag{2.11}$$

where $\sigma > 0$ and $\mathcal{C} > 1$, with further conditions on these constants to be specified later. We will designate by $\|\cdot\|_{\sigma,R}$ the usual norm associated with this inner product and let $L_2^{\sigma,R}(0,1)$ denote the space $L_2(0,1)$ coupled with this R-dependent topology. The closed ball in $L_2^{\sigma,R}(0,1)$ centered at x_0 with radius r will be defined by

$$\mathcal{B}(x_0, r) = \{z \in L_2^{\sigma,R}(0,1),\ \|z - x_0\|_{\sigma,R} \le r\}. \tag{2.12}$$

The following lemma provides a technical estimate useful in the theory which follows.

LEMMA 2.2. *Let $g \in L_\infty(0,1)$. Then for σ, $\mathcal{C}$ defined above and $R \in (0, \bar{R}]$,*

$$\|g\|_{\sigma,R} \le \mathcal{C}\|g\|_\infty. \tag{2.13}$$

Proof. For any $T \in (0,1]$, $\int_0^T e^{-2\sigma t}|g(t)|^2\,dt \le T\|g\|_\infty^2$ so that

$$\|g\|_{\sigma,R}^2 \le \left(\frac{\mathcal{C}^2-1}{R}R + 1\right)\|g\|_\infty^2. \qquad \square$$

LEMMA 2.3. *The operator $F_R : L_2^{\sigma,R}(0,1) \to L_2^{\sigma,R}(0,1)$ as defined in (2.6) is Fréchet differentiable with*

$$F_R'(x)(h)(t) = 2\int_0^R \int_\rho^t x(t+\rho-s)\,h(s)\,ds\,d\eta(\rho), \quad \text{a.e. } t \in (0,1), \tag{2.14}$$

for $h \in L_2^{\sigma,R}(0,1)$. Further, F_R' is uniformly Lipschitz in $L_2^{\sigma,R}(0,1)$ for all $R > 0$ sufficiently small; i.e., there exists some constant $k > 0$ such that

$$\|F_R'(x_1) - F_R'(x_2)\| \le k\|x_1 - x_2\|_{\sigma,R}$$

for $x_1, x_2 \in L_2^{\sigma,R}(0,1)$, where $\|\cdot\|$ denotes the usual $\mathcal{L}(L_2^{\sigma,R}(0,1))$ operator norm.

Proof. Standard arguments (e.g., see [8]) give (2.14). To show the Lipschitz condition, we let $x_1, x_2, h \in L_2^{\sigma,R}(0,1)$ and define

$$g(t;R) \equiv 2\int_0^R \int_\rho^t (x_1(t+\rho-s) - x_2(t+\rho-s))h(s)\,ds\,d\eta(\rho), \quad \text{a.e. } t \in (0,1).$$

Then

$$\begin{aligned}
&|g(t;R)| \\
&\le 2\int_0^R \left(\int_\rho^t e^{2\sigma\tau}e^{-2\sigma\tau}(x_1(\tau) - x_2(\tau))^2\,d\tau\right)^{1/2}\left(\int_\rho^t e^{2\sigma s}e^{-2\sigma s}h^2(s)\,ds\right)^{1/2} d\eta(\rho) \\
&\le 2e^{2\sigma}\,\|x_1 - x_2\|_{\sigma,R}\,\|h\|_{\sigma,R}\int_0^R d\eta(\rho)
\end{aligned}$$

for a.e. $t \in (0,1)$. It follows then from Lemma 2.2 that

$$\|F_R'(x_1) - F_R'(x_2)\| \le \left(2\,\mathcal{C}e^{2\sigma}\int_0^R d\eta(\rho)\right)\|x_1 - x_2\|_{\sigma,R}. \qquad \square$$

LEMMA 2.4. *Let $v, v_1, v_2 \in \mathcal{B}(0,r) \subseteq L_2^{\sigma,R}(0,1)$ and $x \in L_2^{\sigma,R}(0,1)$. Then the remainder*

$$\mathcal{R}_R(x,v) \equiv F_R(x+v) - F_R(x) - F_R'(x)v \tag{2.15}$$

of the Fréchet derivative $F_R'(x)$ satisfies

$$\|\mathcal{R}_R(x,v)\|_{\sigma,R} \le e^{2\sigma}\int_0^R d\eta(\rho)\,\|v\|_{\sigma,R}^2 \tag{2.16}$$

and

$$\begin{aligned}\|\mathcal{R}_R(x,v_1) - \mathcal{R}_R(x,v_2)\|_{\sigma,R} \\ \le 2\,e^{2\sigma}\int_0^R d\eta(\rho)\,\max\{\|v_1\|_{\sigma,R},\,\|v_2\|_{\sigma,R}\}\,\|v_1 - v_2\|_{\sigma,R}\end{aligned} \tag{2.17}$$

for all $R > 0$ sufficiently small.

Proof. As in Lemma 1 of [16] and Lemma 2.2 of [8], we may write

$$\mathcal{R}_R(x,v) = \int_0^1 (F_R'(x+tv) - F_R'(x))\,v\,dt,$$

$$\mathcal{R}_R(x,v_1) - \mathcal{R}_R(x,v_2) = \int_0^1 (F_R'(x+tv_1+(1-t)v_2) - F_R'(x))(v_1 - v_2)\,dt$$

and then apply the results of Lemma 2.3. $\square$

3. Convergence and well-posedness results. Our main convergence results are stated below. They are immediate consequences of Theorem 3.7 and Corollary 3.8, respectively, which are proved at the end of this section.

THEOREM 3.1. *Assume f^δ satisfies the $\mathcal{F}$-data assumption, and let $\tau_{\text{data}} = 1/2$ if $\mathcal{F} = C[0, 1+\bar{R}]$ and $\tau_{\text{data}} = 2/5$ if $\mathcal{F} = L_2(0, 1+\bar{R})$. Let the measure η be given satisfying* (2.8)–(2.9).

Then there exist $\bar{C} > 0$ and $\kappa_1 > 0$ independent of R such that, if the true solution $\bar{x} \in W^{2,\infty}[0, 1+\bar{R}]$ of the autoconvolution equation is positive and satisfies

$$\bar{x}(0) > \bar{C}\,\|\bar{x}'\|_\infty, \tag{3.1}$$

then, for $R = R(\delta) > 0$ selected satisfying

$$\kappa_1 \delta^{\tau_{\text{data}}} \le R(\delta) \le c\,\delta^p \tag{3.2}$$

for any $c > 0$ and $p \in (0, \tau_{\text{data}}]$ ($c \ge \kappa_1$ if $p = \tau_{\text{data}}$) as $\delta \to 0$, it follows that there is a unique solution $x^\delta_{R(\delta)}$ of the regularization equation (2.4) *associated with data f^δ which depends continuously on $f^\delta \in \mathcal{F}$ and which satisfies*

$$\|x^\delta_{R(\delta)} - \bar{x}\|_{L_2(0,1)} = \mathcal{O}(\delta^p) \tag{3.3}$$

as $\delta \to 0$, with an optimal rate obtained for $p = \tau_{\text{data}}$. Thus the optimal convergence rate for continuous data is $\mathcal{O}(\delta^{1/2})$, and the optimal convergence rate for L_2 data is $\mathcal{O}(\delta^{2/5})$.

Proof. Let $\kappa_1 = k_1^{-\tau_{\text{data}}}$, where k_1 is defined in the proof of Theorem 3.7 below. Then assumption (3.2) becomes $\delta \leq k_1 R^{1/(\tau_{\text{data}})}$, and with this assumption Theorem 3.7 gives

$$\|x_R^\delta - \bar{x}\|_{\sigma,R} \leq \hat{C} R$$

for all $R > 0$ sufficiently small. Since

$$\|x_R^\delta - \bar{x}\|_{L_2(0,1)}^2 \leq e^{2\sigma} \int_0^1 e^{-2\sigma t} |x_\delta^R - \bar{x}|^2 \, dx \leq e^{2\sigma} \|x_R^\delta - \bar{x}\|_{\sigma,R}^2,$$

the result follows. ∎

Condition (3.1) is an a priori condition or source condition on the true solution $\bar{x}$, a sufficient condition needed to obtain the convergence rate (3.3) in theory. We note that Example 2 in section 6 below illustrates that this a priori condition is evidently *not* always needed in practice. The particular nature of this a priori condition reflects the fact that the ill-posedness of the original problem increases as $\bar{x}(0)$ approaches zero; thus if $\bar{x}$ is such that $\bar{x}(0)$ is close to zero, the a priori condition says that convergence is assured provided $\bar{x}$ is relatively "flat." To further explore this restriction, note that, for any $y \in W^{1,\infty}(0, 1 + \bar{R})$,

$$\frac{\int_0^R y(t+\rho)\, d\eta(\rho)}{\int_0^R d\eta(\rho)} = y(t)\left(1 + \mathcal{O}(R\,\|y'\|_\infty)\right), \quad \text{a.e. } t \in (0,1),$$

so that the degree to which, for a given value of $R > 0$, the measure is able to recover information about y depends to a large extent on the size of $\|y'\|_\infty$; i.e., flat functions y are handled better under a local averaging process than are functions y with a lot of local variation. Thus, for very ill-posed autoconvolution problems, the source condition says that a method based on local averaging will perform better if the true solution $\bar{x}$ is relatively flat. It is worth noting that a condition like this arises in other areas associated with local regularization; for example, in [4], where a discrepancy principle is developed for the determination of the appropriate value of $R \in (0, \bar{R}]$ in the case of linear Volterra problems, one needs to have the true data f satisfy a similar condition on $\|f'\|_\infty$, a condition which is less restrictive if the value of $\bar{R}$ is small and more restrictive if $\bar{R}$ is relatively large.

As the next corollary shows, we may dispense with the source condition (3.1) entirely if we have knowledge of the actual value of $\bar{x}(0) > 0$ (or an approximation $\hat{x}(0; R)$ to it, as is discussed in Remark 3.3 below) and make use of this information in the construction of our measure η. The proof of this result follows immediately from Corollary 3.8.

COROLLARY 3.2. *Let $\bar{x} \in W^{2,\infty}[0, 1 + \bar{R}]$ be positive. Assume that the perturbed data f^δ satisfy the $\mathcal{F}$-data assumption, and let $\tau_{\text{data}} = 1/2$ if $\mathcal{F} = C[0, 1 + \bar{R}]$ and $\tau_{\text{data}} = 2/5$ if $\mathcal{F} = L_2(0, 1 + \bar{R})$. Assume that the measure η satisfies* (2.8)–(2.9) *with $\bar{\omega}$, $\underline{\omega}$ in* (2.9) *satisfying*

$$\frac{\bar{\omega}}{\underline{\omega}} \leq \tilde{K}\, \bar{x}(0) \tag{3.4}$$

for all $R > 0$ sufficiently small, where $\tilde{K}$ is independent of R and $\bar{x}$.

Then there exists $\kappa_1 > 0$ independent of R such that, if $R = R(\delta) > 0$ is selected satisfying

$$\kappa_1 \delta^{\tau_{\text{data}}} \le R(\delta) \le c\,\delta^p$$

for any $c > 0$ and $p \in (0, \tau_{\text{data}}]$ ($c \ge \kappa_1$ if $p = \tau_{\text{data}}$) as $\delta \to 0$, it follows that there is a unique solution $x^\delta_{R(\delta)}$ of the regularization equation (2.4) associated with data f^δ which depends continuously on $f^\delta \in \mathcal{F}$ and which satisfies

$$\|x^\delta_{R(\delta)} - \bar{x}\|_{L_2(0,1)} = \mathcal{O}(\delta^p)$$

as $\delta \to 0$, with the optimal rate obtained for $p = \tau_{\text{data}}$. Thus the optimal convergence rate for continuous data is $\mathcal{O}(\delta^{1/2})$, and the optimal convergence rate for L_2 data is $\mathcal{O}(\delta^{2/5})$.

Remark 3.3. Despite the presence of $\bar{x}(0)$ in condition (3.4) of Corollary 3.2, it is not necessary to know $\bar{x}(0)$ exactly in order to have such a condition hold. It is enough to have some basic information about the size of $\bar{x}$ and to have an approximation $\hat{x}(0;R)$ of $\bar{x}(0)$ satisfying

$$|\bar{x}(0) - \hat{x}(0;R)| = \mathcal{O}(R) \tag{3.5}$$

as $R \to 0$. (In the next section we will show that an $\mathcal{O}(R)$ approximation to $\bar{x}(0)$ is easily found.) For example, in the case where it is known that $\bar{x}(0) > 1$, then for $\hat{R}$ sufficiently small we have $1 < \hat{x}(0,\hat{R}) < (\hat{x}(0,\hat{R}))^2$, and we may define

$$\underline{\omega} = \hat{x}(0,\hat{R}), \quad \bar{\omega} = (\hat{x}(0;\hat{R}))^2.$$

Then

$$\frac{\bar{\omega}}{\underline{\omega}} = \hat{x}(0,\hat{R}) \le 2\bar{x}(0)$$

for $\hat{R}$ sufficiently small, so that (3.4) holds with $\tilde{K} = 2$. Once $\bar{\omega}$ and $\underline{\omega}$ have been specified, one can always find a suitable R-dependent family of measures η satisfying (2.8)–(2.9) using these choices of constants; for example, let η be given by (2.8) where $\omega(\cdot;R)$ is the line on $[0,R]$ joining $(0,\underline{\omega})$ to $(R,\bar{\omega})$.

To begin the arguments needed to prove the main convergence results given above, we first rewrite (2.2) using similar notation to that used in (2.4). That is, the solution $\bar{x}$ of the original autoconvolution equation satisfies

$$\alpha_R(\bar{x})\bar{x} + F_R(\bar{x}) = f_R + \epsilon_R, \tag{3.6}$$

where

$$\epsilon_R(t) \equiv 2\int_0^R \int_0^\rho (\bar{x}(t) - \bar{x}(t+\rho-s))\,\bar{x}(s)\,ds\,d\eta(\rho), \quad \text{a.e. } t \in (0,1), \tag{3.7}$$

$\alpha_R(\bar{x})$ and $F_R(\bar{x})$ are defined as in (2.5)–(2.6), here using $\bar{x}$ instead of x, and f_R is defined as in (2.7) using f instead of f^δ.

For $h \in L_2^{\sigma,R}(0,1)$ we define bounded linear operators in $L_2^{\sigma,R}(0,1)$:

$$B_R(\bar{x})(h)(t) \equiv 2\int_0^R \int_0^t \bar{x}(t+\rho-s)\,h(s)\,ds\,d\eta(\rho), \tag{3.8}$$

$$D_R(\bar{x})(h)(t) \equiv 2\int_0^R \int_0^\rho \bar{x}(t+\rho-s)\,h(s)\,ds\,d\eta(\rho) \tag{3.9}$$

for a.e. $t \in (0,1)$ and note that

$$F_R'(\bar{x})(h)(t) = B_R(\bar{x})(h)(t) - D_R(\bar{x})(h)(t), \quad \text{a.e. } t \in (0,1). \tag{3.10}$$

We may then use the expansion of $F_R(x)$ in (2.4) to write

$$\alpha_R(x)x + F_R(\bar{x}) + F_R'(\bar{x})(x-\bar{x}) + \mathcal{R}_R(\bar{x}, x-\bar{x}) = f_R^\delta. \tag{3.11}$$

Combining (3.6) and (3.10) with (3.11), we obtain

$$\begin{aligned}(\alpha_R(\bar{x})I + B_R(\bar{x}))(x-\bar{x}) = f_R^\delta - f_R - \epsilon_R - \mathcal{R}_R(\bar{x}, x-\bar{x}) \\ + D_R(\bar{x})(x-\bar{x}) + (\alpha_R(\bar{x}) - \alpha_R(x))x,\end{aligned} \tag{3.12}$$

where I is the identity operator on $L_2^{\sigma,R}(0,1)$. Let us further denote for $v \in L_2^{\sigma,R}(0,1)$ that

$$E_R(\bar{x}, v) = D_R(\bar{x})(v) - \alpha_R(v)\,\bar{x}, \tag{3.13}$$

where we note that $v \to E_R(\bar{x}, v)$ is linear. Then (3.12) becomes

$$\begin{aligned}(\alpha_R(\bar{x})I + B_R(\bar{x}))(x-\bar{x}) = f_R^\delta - f_R - \epsilon_R - \mathcal{R}_R(\bar{x}, x-\bar{x}) \\ + E_R(\bar{x}, x-\bar{x}) + (\alpha_R(x) - \alpha_R(\bar{x}))(\bar{x}-x).\end{aligned} \tag{3.14}$$

The following two lemmas will be used to establish the invertibility of the operator $(\alpha_R(\bar{x})I + B_R(\bar{x})) \in \mathcal{L}(L_2^{\sigma,R}(0,1))$.

LEMMA 3.4. *If $\bar{x} \in W^{2,\infty}[0, 1+\bar{R}]$ and $\bar{x}(t) > 0$ for $t \in [0, 1+\bar{R}]$, then there exists $\sigma_0 > 0$ independent of $R > 0$ such that the operator $B_R(\bar{x})$ is accretive in $L_2^{\sigma,R}(0,1)$ for $\sigma \geq \sigma_0$; i.e.,*

$$\langle B_R(\bar{x})v, v\rangle_{\sigma,R} \geq 0 \qquad \textit{for any} \quad v \in L_2^{\sigma,R}(0,1),$$

where B_R is defined using a measure η satisfying (2.8)–(2.9).

Proof. Following the ideas of [16], only now with R-dependent quantities, we make the definition for $t \in [0,1]$, $a_R(t) \equiv 2\int_0^R \bar{x}(t+\rho)\,d\eta(\rho)$, so that we may write, for $h \in L_2^{\sigma,R}(0,1)$,

$$B_R(\bar{x})(h)(t) = \int_0^t a_R(t-s)h(s)\,ds, \quad \text{a.e. } t \in (0,1). \tag{3.15}$$

For all σ sufficiently large, we will first show that the operator $B_R^\sigma(\bar{x})$ is accretive in $L_2(0,1)$, where $B_R^\sigma(\bar{x})$ is defined as in (3.15) except that the kernel $a_R(t)$ for B_R is replaced by $a_R[\sigma](t) = e^{-\sigma t}a_R(t)$. Defining

$$A_0 = \min_{t\in[0,1+\bar{R}]} \bar{x}(t) > 0, \qquad A_1 = \|\bar{x}'\|_{L^\infty[0,1+\bar{R}]}, \qquad A_2 = \|\bar{x}''\|_{L^\infty[0,1+\bar{R}]},$$

then

$$\min_{t\in[0,1]} a_R(t) \geq 2\,A_0 \int_0^R d\eta(\rho),$$

$$\|a_R'\|_{L^\infty[0,1]} \leq 2\,A_1 \int_0^R d\eta(\rho),$$

$$\|a_R''\|_{L^\infty[0,1]} \leq 2\,A_2 \int_0^R d\eta(\rho),$$

and, for a.e. $t \in (0,1)$,

$$a_R[\sigma](t) \ge e^{-\sigma t}\left(2\,A_0\int_0^R d\eta(\rho)\right),$$

$$a_R'[\sigma](t) \le e^{-\sigma t}\,2\,(-\sigma A_0 + A_1)\int_0^R d\eta(\rho),$$

$$a_R''[\sigma](t) \ge e^{-\sigma t}\,2\,(\sigma^2 A_0 - 2\sigma A_1 - A_2)\int_0^R d\eta(\rho).$$

For $\sigma \ge \sigma_0$, where

$$\sigma_0 \ge \frac{A_1 + \sqrt{A_1^2 + A_0 A_2}}{A_0} \qquad \text{(independent of } R\text{)},$$

we then have from the positivity of η that

$$a_R[\sigma](t) \ge 0, \qquad a_R[\sigma]'(t) \le 0, \qquad a_R[\sigma]''(t) \ge 0$$

for a.e. $t \in (0,1)$.

Therefore, the kernel $a_R[\sigma](\cdot)$ is nonnegative, nonincreasing, and convex. Using Lemma 2 of [16] for $\sigma \ge \sigma_0$, we have $B_R^\sigma(\bar{x})$ accretive on $L_2(0,1)$, from which it follows that $B_R(\bar{x})$ is accretive in $L_2(0,1)$ with a weighted $e^{-2\sigma t}$ norm. A similar statement may be made for $B_R(\bar{x})$ using a weighted $L_2(0,R)$ norm, from which it follows that $B_R(\bar{x})$ is accretive in $L_2^{\sigma,R}(0,1)$ for all $\sigma \ge \sigma_0$. ∎

LEMMA 3.5. *Let $\bar{x} \in C^1[0, 1+\bar{R}]$ satisfy $\bar{x}(0) > 0$, and assume that the measure η satisfies* (2.8)–(2.9). *Then for $R > 0$ sufficiently small we have $\alpha_R(\bar{x}) > 0$ and*

$$\frac{1}{\alpha_R(\bar{x})} \le \frac{1}{\bar{x}(0)\int_0^R \rho\, d\eta(\rho)}. \tag{3.16}$$

Proof. We can write

$$\alpha_R(\bar{x}) = 2\,\bar{x}(0)\int_0^R \rho\, d\eta(\rho)\;[1 + h(R)]\,,$$

where

$$h(R) = \frac{1}{\bar{x}(0)\int_0^R \rho\, d\eta(\rho)}\int_0^R\int_0^\rho s\,\bar{x}'(\xi(s))\,ds\,d\eta(\rho)$$

for some $\xi(s) \in [0,R]$. But

$$|h(R)| \le \frac{1}{\bar{x}(0)\int_0^R \rho\, d\eta(\rho)}\,\|\bar{x}'\|_\infty \int_0^R \frac{\rho^2}{2}\,d\eta(\rho) \le \frac{\|\bar{x}'\|_\infty K(2)}{2\,\bar{x}(0)}R,$$

where $K(2)$ is defined in (2.10), so it follows that

$$\frac{1}{\alpha_R(\bar{x})} = \frac{1}{2\,\bar{x}(0)\int_0^R \rho\, d\eta(\rho)}\,(1 + \mathcal{O}(R)) \quad \text{as } R \to 0. \qquad ∎$$

As a consequence of Lemmas 3.4 and 3.5, we have that if $\bar{x} \in W^{2,\infty}[0, 1+\bar{R}]$ with $\bar{x}(0) > 0$, then $\alpha_R(\bar{x}) > 0$ and $(\alpha_R(\bar{x})I + B_R(\bar{x}))^{-1} \in \mathcal{L}(L_2^{\sigma,R}(0,1))$ for all $R > 0$ sufficiently small; in addition, we have the following estimate (see [33]):

$$\|(\alpha_R(\bar{x})I + B_R(\bar{x}))^{-1}\| \leq \frac{1}{\alpha_R(\bar{x})} \tag{3.17}$$

for $\|\cdot\|$ the $\mathcal{L}(L_2^{\sigma,R}(0,1))$ operator norm and for $\sigma \geq \sigma_0$, where $\sigma_0 > 0$ is independent of R.

We are now ready to rewrite our regularized equation (2.4) as

$$x = H_R\, x, \tag{3.18}$$

where, from (3.14), $H_R : L_2^{\sigma,R}(0,1) \mapsto L_2^{\sigma,R}(0,1)$ is given by

$$\begin{aligned} H_R\, x = (\alpha_R(\bar{x})I + B_R(\bar{x}))^{-1}[f_R^\delta - f_R - \epsilon_R - \mathcal{R}_R(\bar{x}, x-\bar{x}) \\ + E_R(\bar{x}, x-\bar{x}) + (\alpha_R(x) - \alpha_R(\bar{x}))(\bar{x}-x)] + \bar{x}. \end{aligned} \tag{3.19}$$

In the next lemma we bound relevant quantities on the right-hand side of (3.19).

LEMMA 3.6. *Let $\bar{x} \in C^1[0, 1+\bar{R}]$ satisfy $\bar{x}(0) > 0$, let f^δ satisfy the $\mathcal{F}$-data assumption, assume the measure η satisfies* (2.8)–(2.9), *and let $\sigma > 0$. Let $x_1, x_2 \in L_2^{\sigma,R}(0,1)$. Then*

$$\|f_R^\delta - f_R\|_{\sigma,R} \leq \begin{cases} \delta\, \mathcal{C}\, \bar{\omega}\, R & \text{if } \mathcal{F} = C[0, 1+\bar{R}], \\ \delta\, \mathcal{C}\, \bar{\omega}\, R^{1/2} & \text{if } \mathcal{F} = L_2(0, 1+\bar{R}), \end{cases} \tag{3.20}$$

$$\|E_R(\bar{x}, x_1 - x_2)\|_{\sigma,R} \leq \mathcal{C}\, \frac{2e^{\sigma,R}}{\sqrt{3}} \|\bar{x}'\|_\infty \|x_1 - x_2\|_{\sigma,R} \int_0^R \rho^{3/2}\, d\eta(\rho), \tag{3.21}$$

$$\|\epsilon_R\|_{\sigma,R} \leq 2\, \mathcal{C} \sqrt{\frac{2}{3}} \|\bar{x}'\|_\infty \left(\bar{x}(0) \int_0^R \rho^2\, d\eta(\rho) + \frac{\|\bar{x}'\|_\infty}{\sqrt{3}} \int_0^R \rho^3\, d\eta(\rho) \right), \tag{3.22}$$

and

$$|\alpha_R(x_1) - \alpha_R(x_2)| \leq 2\, e^{\sigma R} \sqrt{\frac{R}{\mathcal{C}^2 - 1}}\, \|x_1 - x_2\|_{\sigma,R} \int_0^R \rho^{1/2}\, d\eta(\rho). \tag{3.23}$$

Proof. The bounds in (3.20)–(3.22) each follow from applications of Lemma 2.2 using obvious choices of the function $g = g(\cdot; R)$. Inequality (3.20) is straightforward in the case of $f^\delta \in C[0, 1+\bar{R}]$ since

$$\begin{aligned} |f_R^\delta(t) - f_R(t)| &= \left| \int_0^R (f^\delta(t+\rho) - f(t+\rho))\, \omega(\rho; R)\, d\rho \right| \\ &\leq \delta\, \bar{\omega}\, R. \end{aligned}$$

For the case of $f^\delta \in L_2(0, 1+\bar{R}]$ we observe that $f_R^\delta - f_R \in L_\infty(0,1)$ since

$$\begin{aligned} |f_R^\delta(t) - f_R(t)| &\leq \left(\int_0^R |f^\delta(t+\rho) - f(t+\rho)|^2 d\rho \right)^{1/2} \left(\int_0^R |\omega(\rho; R)|^2 d\rho \right)^{1/2} \\ &\leq \|f^\delta - f\|_{L_2(0,1+\bar{R})}\, \bar{\omega}\, R^{1/2} \end{aligned}$$

for a.e. $t \in (0, 1+\bar{R})$. The result then follows from Lemma 2.2.

For (3.21) we use the fact that, for a.e. $t \in (0,1)$,

$$\begin{aligned}
&|E_R(\bar{x}, x_1 - x_2)(t)| \\
&\le 2\int_0^R \int_0^\rho |\bar{x}(t+\rho-s) - \bar{x}(t)|\, |x_1(s) - x_2(s)|\, ds\, d\eta(\rho) \\
&\le 2\|\bar{x}'\|_\infty \int_0^R \left(\int_0^\rho (\rho - s)^2\, ds\right)^{1/2} \left(\int_0^\rho e^{2\sigma s} e^{-2\sigma s} |x_1(s) - x_2(s)|^2\, ds\right)^{1/2} d\eta(\rho).
\end{aligned}$$

For (3.22) we observe that, for some $\xi \in (0, R)$,

$$\begin{aligned}
|\epsilon_R(t)| &\le 2\|\bar{x}'\|_\infty \int_0^R \left(\int_0^\rho (\rho - s)^2\, ds\right)^{1/2} \left(\int_0^\rho \left(\bar{x}(0)\, ds + \bar{x}'(\xi)\, s\right)^2\, ds\right)^{1/2} d\eta(\rho) \\
&\le 2\|\bar{x}'\|_\infty \int_0^R \left[\frac{\rho^3}{3} 2\left(\bar{x}(0)^2\, \rho + \|\bar{x}'\|_\infty^2\, \frac{\rho^3}{3}\right)\right]^{1/2} d\eta(\rho) \\
&\le 2\|\bar{x}'\|_\infty \int_0^R \sqrt{\frac{2}{3}}\, \rho^{3/2} \left(\bar{x}(0)\rho^{1/2} + \|\bar{x}'\|_\infty \frac{\rho^{3/2}}{\sqrt{3}}\right) d\eta(\rho)
\end{aligned}$$

for a.e. $t \in (0,1)$, from which the bound in (3.22) follows.

Finally, by using the fact that

$$\begin{aligned}
|\alpha_R(x_1) - \alpha_R(x_2)| &\le 2\int_0^R \left|\int_0^\rho (x_1(s) - x_2(s))\, ds\right| d\eta(\rho) \\
&\le 2\, e^{\sigma R} \left(\int_0^R e^{-2\sigma t} |x_1(t) - x_2(t)|^2\, dt\right)^{1/2} \int_0^R \rho^{1/2} d\eta(\rho),
\end{aligned}$$

the estimate in (3.23) is obtained. ∎

We are now ready to state our main convergence results which are given below in Theorem 3.7 and Corollary 3.8.

THEOREM 3.7. *Assume that f^δ satisfies the $\mathcal{F}$-data condition, and let $\tau_{\text{data}} = 1/2$ in the case of $\mathcal{F} = C[0, 1+\bar{R}]$ and $\tau_{\text{data}} = 2/5$ in the case of $\mathcal{F} = L_2(0, 1+\bar{R})$. Assume that the measure $\eta(\rho) > 0$ satisfies (2.8)–(2.9) and that the autoconvolution problem (1.1) has a positive solution $\bar{x} \in W^{2,\infty}[0, 1+\bar{R}]$ satisfying*

$$\bar{x}(0) > 9\, b^2\, e^{2\sigma}\, \|\bar{x}'\|_\infty \tag{3.24}$$

for $\sigma \ge \sigma_0$ and $b \ge 2\bar{\omega}/\underline{\omega}$, where $\bar{\omega}, \underline{\omega}$ are given in (2.8). Then for any $\mathcal{C} \in (1, 9/8]$ there exist constants $k_1 > 0$ and $\hat{C} = \hat{C}(\mathcal{C}) > 0$ independent of R such that, if $\delta = \delta(R) > 0$ satisfies

$$\delta \le k_1 R^{1/\tau_{\text{data}}}, \tag{3.25}$$

then for all $R > 0$ sufficiently small the regularized equation (3.18) has a unique solution $x_R^\delta \in L_2^{\sigma,R}(0,1)$ satisfying

$$\|x_R^\delta - \bar{x}\|_{\sigma,R}^2 \le \hat{C}^2 R^2,$$

where $\|\cdot\|_{\sigma,R}$ *is defined from* (2.11) *using the value of* $\mathcal{C}$ *given above. Further,* $x_R^\delta \in L_2^{\sigma,R}(0,1)$ *depends continuously on* $f^\delta \in \mathcal{F}$ *for all* $R > 0$ *sufficiently small.*

Proof. We will apply the contraction mapping principle to the regularized equation (3.18) in the ball $\mathcal{B}(\bar{x}, \hat{C}R)$. From Lemmas 2.4, 3.5, and 3.6 and inequality (3.17) we have

$$\begin{aligned}
\|H_R x - \bar{x}\|_{\sigma,R} &\le \frac{1}{\alpha(\bar{x})}\|f_R^\delta - f_R\|_{\sigma,R} + \frac{1}{\alpha(\bar{x})}\|\mathcal{R}_R(\bar{x}, x-\bar{x})\|_{\sigma,R} + \frac{1}{\alpha(\bar{x})}\|\epsilon_R\|_{\sigma,R} \\
&\quad + \frac{1}{\alpha(\bar{x})}\|E_R(\bar{x}, x-\bar{x})\|_{\sigma,R} + \frac{|\alpha_R(x) - \alpha_R(\bar{x})|}{\alpha_R(\bar{x})}\|x-\bar{x}\|_{\sigma,R} \\
&\le \frac{\delta\,\mathcal{C}}{\bar{x}(0)}\,\frac{2\bar{\omega}}{\underline{\omega}}\,R^p + \frac{e^{2\sigma}}{\bar{x}(0)}\,\frac{2\bar{\omega}}{\underline{\omega}}\,R^{-1}\,\|x-\bar{x}\|_{\sigma,R}^2 \\
&\quad + 2\,\mathcal{C}\sqrt{\frac{2}{3}}\,\|\bar{x}'\|_\infty\left(\frac{2\bar{\omega}}{3\underline{\omega}}\,R + \frac{\|\bar{x}'\|_\infty}{\sqrt{3}\,\bar{x}(0)}\,K(3)\,R^2\right) \\
&\quad + \frac{2\,\mathcal{C}\,\|\bar{x}'\|_\infty\,e^{\sigma R}}{\sqrt{3}\,\bar{x}(0)}\,K(\tfrac{3}{2})\,R^{1/2}\,\|x-\bar{x}\|_{\sigma,R} \\
&\quad + \frac{2\,e^{\sigma R}}{\bar{x}(0)}\,K(\tfrac{1}{2})\,R^{-1/2}\,\|x-\bar{x}\|_{\sigma,R}^2\,\sqrt{\frac{R}{\mathcal{C}^2-1}}
\end{aligned}$$

for $\sigma \ge \sigma_0$, where $p = -1$ for $\mathcal{F} = C[0, 1+\bar{R}]$ and $p = -3/2$ for $\mathcal{F} = L_2(0, 1+\bar{R})$. Using assumption (3.25) and the fact that $\|x-\bar{x}\|_{\sigma,R} \le \hat{C}R$, we have

$$\begin{aligned}
\|H_R x - \bar{x}\|_{\sigma,R} &\le \frac{k_1\mathcal{C}}{\bar{x}(0)}\,\frac{2\bar{\omega}}{\underline{\omega}}R + \frac{e^{2\sigma}}{\bar{x}(0)}\,\frac{2\bar{\omega}}{\underline{\omega}}\,\hat{C}^2R + 2\mathcal{C}\sqrt{\frac{2}{3}}\,\|\bar{x}'\|_\infty\,\frac{2\bar{\omega}}{3\underline{\omega}}\,R \\
&\quad + \frac{2\sqrt{2}\,\mathcal{C}\|\bar{x}'\|_\infty^2}{3\bar{x}(0)}\,K(3)R^2 + \frac{2\,\mathcal{C}\,\|\bar{x}'\|_\infty e^{\sigma R}}{\sqrt{3}\,\bar{x}(0)}\,K(\tfrac{3}{2})\,\hat{C}R^{\frac{3}{2}} + \frac{2\,e^{\sigma R}\,K(\frac{1}{2})}{\bar{x}(0)}\,\frac{\hat{C}^2R^2}{\sqrt{\mathcal{C}^2-1}}.
\end{aligned}$$

Therefore to have $\|H_R x - \bar{x}\|_{\sigma,R} \le \hat{C}R$ for some $\hat{C} > 0$ and all $R > 0$ sufficiently small, a sufficient condition is that

$$\frac{1}{\bar{x}(0)}\,b\,k_1\,\mathcal{C} + \|\bar{x}'\|_\infty\,b\,\mathcal{C} + \frac{e^{2\sigma}}{\bar{x}(0)}\,b\,\hat{C}^2 < \hat{C}. \tag{3.26}$$

If we let

$$k_1 = \|\bar{x}'\|_\infty\,\bar{x}(0), \tag{3.27}$$

then (3.26) becomes

$$L(\hat{C}) \equiv \frac{e^{2\sigma}\,b}{\bar{x}(0)}\,\hat{C}^2 - \hat{C} + 2\,b\,\mathcal{C}\,\|\bar{x}'\|_\infty < 0. \tag{3.28}$$

It is not hard to see that $L(\hat{C}) = 0$ has two distinct positive solutions $0 < \hat{C}_1 < \hat{C}_2$ by assumption (3.24) when $\mathcal{C} \in (1, 9/8]$. Then for $\hat{C}$ satisfying $\hat{C}_1 < \hat{C} < \hat{C}_2$ we have $L(\hat{C}) < 0$, and thus $\|H_R x - \bar{x}\|_{\sigma,R} \le \hat{C}R$ for all $R > 0$ sufficiently small.

To further demonstrate that H_R is a contraction on $\mathcal{B}(\bar{x}, \hat{C}R)$, we let $x_1, x_2 \in \mathcal{B}(\bar{x}, \hat{C}R)$ and note that

$$
\begin{aligned}
(3.29)\quad & \|H_R\, x_1 - H_R\, x_2\|_{\sigma,R} \\
&= \| \left(\alpha_R(\bar{x})I + B_R(\bar{x})\right)^{-1} \{\mathcal{R}_R(\bar{x}, x_2 - \bar{x}) - \mathcal{R}_R(\bar{x}, x_1 - \bar{x}) + E_R(\bar{x}, x_1 - x_2) \\
&\qquad - [(\alpha_R(x_1) - \alpha_R(\bar{x}))(x_1 - \bar{x}) - (\alpha_R(x_2) - \alpha_R(\bar{x}))(x_2 - \bar{x})]\,\} \|_{\sigma,R} \\
&\le T_R(x_1, x_2),
\end{aligned}
$$

where

$$
\begin{aligned}
T_R(x_1, x_2) \equiv & \frac{1}{\alpha_R(\bar{x})} \|\mathcal{R}_R(\bar{x}, x_2 - \bar{x}) - \mathcal{R}_R(\bar{x}, x_1 - \bar{x})\|_{\sigma,R} \\
& + \frac{1}{\alpha_R(\bar{x})} \|E_R(\bar{x}, x_1 - x_2)\|_{\sigma,R} \\
& + \frac{1}{\alpha_R(\bar{x})} \|(\alpha_R(x_1) - \alpha_R(\bar{x}))(x_1 - \bar{x}) - (\alpha_R(x_2) - \alpha_R(\bar{x}))(x_2 - \bar{x})\|_{\sigma,R}.
\end{aligned}
$$

Since

$$
\begin{aligned}
& \frac{1}{\alpha_R(\bar{x})} \|(\alpha_R(x_1) - \alpha_R(\bar{x}))(x_1 - \bar{x}) - (\alpha_R(x_2) - \alpha_R(\bar{x}))(x_2 - \bar{x})\|_{\sigma,R} \\
&= \frac{1}{\alpha_R(\bar{x})} \|(\alpha_R(x_1) - \alpha_R(x_2))(x_1 - \bar{x}) + (\alpha_R(x_2) - \alpha_R(\bar{x}))(x_1 - x_2)\|_{\sigma,R} \\
&\le \frac{|\alpha_R(x_1) - \alpha_R(x_2)|}{\alpha(\bar{x})} \|x_1 - \bar{x}\|_{\sigma,R} + \frac{|\alpha_R(x_2) - \alpha_R(\bar{x})|}{\alpha(\bar{x})} \|x_1 - x_2\|_{\sigma,R} \\
&\le \frac{2e^{\sigma R} \int_0^R \rho^{\frac{1}{2}}\, d\eta(\rho)}{\bar{x}(0) \int_0^R \rho\, d\eta(\rho)} \sqrt{\frac{R}{\mathcal{C}^2 - 1}}\ \{\|x_1 - x_2\|_{\sigma,R} \|x_1 - \bar{x}\|_{\sigma,R} + \|x_2 - \bar{x}\|_{\sigma,R}\, \|x_1 - x_2\|_{\sigma,R}\} \\
&\le \frac{4e^{\sigma R} K(\frac{1}{2}) \hat{C} R}{\bar{x}(0)\sqrt{\mathcal{C}^2 - 1}} \|x_1 - x_2\|_{\sigma,R},
\end{aligned}
$$

we have

$$
\begin{aligned}
T_R(x_1, x_2) \le & \left[\frac{2e^{2\sigma}}{\bar{x}(0)}\, \frac{2\bar{\omega}}{\underline{\omega}}\, \hat{C}\right] \|x_1 - x_2\|_{\sigma,R} \\
& + \left[\frac{2\,\mathcal{C}\, \|\bar{x}'\|_\infty e^{\sigma R}}{\sqrt{3}\bar{x}(0)}\, K\left(\frac{3}{2}\right) R^{\frac{1}{2}} + \frac{4e^{\sigma R} K(\frac{1}{2}) \hat{C} R}{\bar{x}(0)\sqrt{\mathcal{C}^2 - 1}}\right] \|x_1 - x_2\|_{\sigma,R}.
\end{aligned}
$$

Thus if we require that

$$
(3.30)\qquad \hat{C} < \frac{\bar{x}(0)}{2\, e^{2\sigma}\, b},
$$

it follows that

$$
\hat{C} < \frac{\bar{x}(0)}{2e^{2\sigma}}\, \frac{\underline{\omega}}{2\bar{\omega}}
$$

and, for $R > 0$ sufficiently small,

$$\|H_R\, x_1 - H_R\, x_2\|_{\sigma,R} \le T_R(x_1, x_2) \le q\, \|x_1 - x_2\|_{\sigma,R}$$

for any $x_1, x_2 \in \mathcal{B}(\bar{x}, \hat{C}R)$ with some $q < 1$. Further we note that $\frac{\hat{C}_1+\hat{C}_2}{2} = \frac{\bar{x}(0)}{2\, e^{2\sigma}\, b}$, and therefore our regularized equation (3.18) has a unique solution x_R^δ in $\mathcal{B}(\bar{x}, \hat{C}R)$ for $\hat{C}$ satisfying $\hat{C}_1 < \hat{C} < \frac{\bar{x}(0)}{2\, e^{2\sigma}\, b}$.

Finally, fix $R > 0$ sufficiently small, and let $f_1^\delta, f_2^\delta \in \mathcal{F}$ satisfy

$$\|f_i^\delta - f\|_{\mathcal{F}} \le \delta \le k_1 R^{1/\tau_{\rm data}}, \quad i = 1, 2,$$

for the value of k_1 given in (3.27), where $\mathcal{F}$ is either $C[0, 1+\bar{R}]$ or $L_2(0, 1+\bar{R})$. Let $\hat{C}$, $\mathcal{C}$, σ, and q be as specified above, and let $f_{R,i}^\delta$ and $H_{R,i}$ be defined for $i = 1, 2$ as in (2.7) and (3.19), respectively, using f_i^δ instead of f^δ as data. Then for $i = 1, 2$ there exists a unique solution $x_{R,i}^\delta \in \mathcal{B}(\bar{x}, \hat{C}R) \subset L_2^{\sigma,R}(0,1)$ of the equation $x = H_{R,i}x$. Further,

$$\begin{aligned}
\|x_{R,1}^\delta - x_{R,2}^\delta\|_{\sigma,R} &= \|H_{R,1}x_{R,1}^\delta - H_{R,2}x_{R,2}^\delta\|_{\sigma,R} \\
&= \| \left(\alpha_R(\bar{x})I + B_R(\bar{x})\right)^{-1} \{ \left(f_{R,1}^\delta - f_{R,2}^\delta\right) \\
&\qquad + \mathcal{R}_R(\bar{x}, x_{R,2}^\delta - \bar{x}) - \mathcal{R}_R(\bar{x}, x_{R,1}^\delta - \bar{x}) + E_R(\bar{x}, x_{R,1}^\delta - x_{R,2}^\delta) \\
&\qquad - [(\alpha_R(x_{R,1}^\delta) - \alpha_R(\bar{x}))(x_{R,1}^\delta - \bar{x}) - (\alpha_R(x_{R,2}^\delta) - \alpha_R(\bar{x}))(x_{R,2}^\delta - \bar{x})] \} \|_{\sigma,R} \\
&\le \frac{1}{\alpha_R(\bar{x})} \|f_{R,1}^\delta - f_{R,2}^\delta\|_{\sigma,R} + T_R(x_{R,1}^\delta, x_{R,2}^\delta) \\
&\le \frac{\mathcal{C}\, \|f_1^\delta - f_2^\delta\|_{\mathcal{F}}}{\bar{x}(0)} \frac{2\bar{\omega}}{\underline{\omega}} R^p + q \|x_{R,1}^\delta - x_{R,2}^\delta\|_{\sigma,R},
\end{aligned}$$

where $p = -1$ in the case of $\mathcal{F} = C[0, 1+\bar{R}]$ and $p = -3/2$ in the case of $\mathcal{F} = L_2(0, 1+\bar{R})$ and where we have used arguments such as those needed to verify (3.20) in the first term above. Therefore

$$\|x_{R,1}^\delta - x_{R,2}^\delta\|_{\sigma,R} \le \frac{\mathcal{C}\, b}{(1-q)\bar{x}(0)}\, R^p\, \|f_1^\delta - f_2^\delta\|_{\mathcal{F}},$$

and continuous dependence of solutions on data is obtained for the regularized equation (3.18). ∎

COROLLARY 3.8. *Assume that the autoconvolution problem* (1.1) *has a positive solution* $\bar{x} \in W^{2,\infty}[0, 1+\bar{R}]$ *and that the measure* η *satisfies* (2.8)–(2.9) *with constants* $\bar{\omega}$, $\underline{\omega}$ *in* (2.9) *satisfying* (3.4) *for all* $R > 0$ *sufficiently small, where* $\tilde{K}$ *is independent of* R *and* $\bar{x}$. *Let* f^δ *satisfy the* $\mathcal{F}$*-data condition, and let* $\tau_{\rm data} = 1/2$ *in the case of* $\mathcal{F} = C[0, 1+\bar{R}]$ *and* $\tau_{\rm data} = 2/5$ *in the case of* $\mathcal{F} = L_2(0, 1+\bar{R})$.

Then there exist constants $k_1, \hat{C} > 0$ *independent of* R *such that, if* $\delta = \delta(R) > 0$ *satisfies*

$$\delta \le k_1 R^{1/\tau_{\rm data}}, \tag{3.31}$$

then for all $R > 0$ *sufficiently small the regularized equation* (3.18) *has a unique solution* $x_R^\delta \in L_2^{\sigma,R}(0,1)$ *satisfying*

$$\|x_R^\delta - \bar{x}\|_{\sigma,R}^2 \le \hat{C}^2 R^2,$$

where $\|\cdot\|_{\sigma,R}$ *is defined from* (2.11) *using the value of* $\mathcal{C} = \sqrt{2}$. *Further,* $x_R^\delta \in L_2^{\sigma,R}(0,1)$ *depends continuously on* $f^\delta \in \mathcal{F}$ *for all* $R > 0$ *sufficiently small.*

Proof. Let $\mathcal{C} = \sqrt{2}$, and define $\tilde{b} = 2\tilde{K}$, where $\tilde{K}$, given by (3.4), is independent of R and $\bar{x}$. We observe that without loss of generality we may assume that $\bar{x}$ satisfies

$$\bar{x}(0)\|\bar{x}'\|_\infty < \frac{\sqrt{2}\, e^{-2\sigma}}{16\,\tilde{b}^2}, \tag{3.32}$$

because if this does not already hold, we may rescale (1.1) via division by some $\kappa^2 > 1$ to obtain a new equation with data f/κ^2 and solution $\bar{y} = \bar{x}/\kappa$, where $\bar{y}$ satisfies (3.32). Note that a similar rescaling of η will not change the value of $\tilde{K}$, i.e., of $\tilde{b}$.

The proof of the corollary follows the proof of Theorem 3.7 with only a few changes. First we note that, if we let $b = \tilde{b}\bar{x}(0)$, then $b \geq 2\bar{\omega}/\underline{\omega}$, and (3.26) still holds from the proof of Theorem 3.7. Writing (3.26) in terms of $\tilde{b}$, we then have the following sufficient condition ensuring that $\|H_R x - \bar{x}\|_{\sigma,R} < \hat{C}R$ for some $\hat{C} > 0$:

$$k_1\tilde{b}\,\mathcal{C} + \|\bar{x}'\|_\infty\,\bar{x}(0)\,\tilde{b}\,\mathcal{C} + e^{2\sigma}\,\tilde{b}\,\hat{C}^2 < \hat{C}. \tag{3.33}$$

Defining k_1 as in (3.27), the inequality in (3.33) now becomes

$$L(\hat{C}) = e^{2\sigma}\,\tilde{b}\,\hat{C}^2 - \hat{C} + 2\|\bar{x}'\|_\infty\,\bar{x}(0)\,\tilde{b}\,\mathcal{C} < 0.$$

Under condition (3.32) the quadratic L has two distinct positive roots $0 < \hat{C}_1 < \hat{C}_2$ so that we may pick $\hat{C} \in (\hat{C}_1, \hat{C}_2)$ to guarantee that $L(\hat{C}) < 0$.

The remainder of the proof of Theorem 3.7 carries over without change; we note only that the new condition (3.30) on $\hat{C}$ in that proof creates no difficulty because here our roots $\hat{C}_1$ and $\hat{C}_2$ satisfy $\frac{\hat{C}_1+\hat{C}_2}{2} = \frac{1}{2e^{2\sigma}\tilde{b}} = \frac{\bar{x}(0)}{2e^{2\sigma}b}$ as needed for the proof of Theorem 3.7. ∎

4. Discretization. In this section, we will consider a discretized version of the regularized equation (2.4), which leads to a stable method on the interval $[0,1]$; the resulting method is nonlinear and nonsequential on the initial small interval $[0,R]$ but sequential and linear on the larger interval $[R,1]$.

We will consider a collocation-based discretization of (2.4) where, for simplicity, we will assume that the measure η is given by the Lebesgue measure, i.e.,

$$\int_0^R g(\rho)\,d\eta(\rho) = \int_0^R g(\rho)\,d\rho, \tag{4.1}$$

and note that condition (2.9) is satisfied in this case. Indeed our numerical results were so good using this simple measure that there seemed little need to construct a weighted measure based on an approximate value of $\bar{x}(0)$ as was discussed as a possibility in Remark 3.3.

Let $N = 1, 2, 3, \ldots$, and define

$$t_i = i\Delta t, \quad i = 0, 1, \ldots, N, \quad \Delta t = 1/N.$$

We will also let $R = r\Delta t$, where $r \in \{1, 2, \ldots, N\}$ is fixed ($r \ll N$, in general).

For $i = 2, 3, \ldots, N$, let $\chi_i(t)$ be the indicator function on the interval $(t_{i-1}, t_i]$ and $\chi_1(t)$ the indicator function on the interval $[t_0, t_1]$, and define the approximation

space of piecewise constant functions on $[0,1]$ by $\mathcal{S}_N = \text{span}\{\,\chi_i,\, i = 1,2,\ldots,N\}$. We then seek $x \in \mathcal{S}_N$ satisfying the collocation equations

$$2\int_0^R \int_0^\rho x(s)\,ds\,d\rho\,x(t_i) + \int_0^R \int_\rho^{t_i} x(t_i+\rho-s)\,x(s)\,ds\,d\rho = \int_0^R f^\delta(t_i+\rho)\,d\rho \tag{4.2}$$

for $i = 1,2,\ldots,N$. We will use the rectangular quadrature rule to approximate integrals on $[0,R]$, i.e.,

$$\int_0^R g(\rho)\,d\rho \approx \Delta t \sum_{j=0}^{r-1} g(t_j),$$

and incorporate this approximation directly into (4.2). Looking at each term in (4.2) we then make the following replacements:

$$\int_0^R f^\delta(t_i+\rho)\,d\rho \approx \Delta t \sum_{q=0}^{r-1} f^\delta(t_{i+q}),$$

and, letting $x(t) = \sum_{p=1}^N c_p \chi_p(t)$ for $t \in [0,1]$, we have

$$\begin{aligned} 2\int_0^R \int_0^\rho x(s)\,ds\,d\rho\,x(t_i) &\approx 2\Delta t \left(\int_0^0 x(s)\,ds + \int_0^{t_1} x(s)\,ds + \cdots + \int_0^{t_{r-1}} x(s)\,ds \right) c_i \\ &= 2\,(\Delta t)^2 \left(\sum_{l=1}^{r-1} \sum_{m=1}^{l} c_m \right) c_i, \end{aligned}$$

where we notice that the coefficient of c_i involves only $c_1, c_2, \ldots, c_{r-1}$. The second term of the left-hand side of (4.2) is more complicated, and, depending on the value of i in relation to r, we come to different forms for the term. In general, for any fixed $i = 1,2,\ldots,N$, we have

$$\int_0^R \int_\rho^{t_i} x(t_i+\rho-s)\,x(s)\,ds\,d\rho \approx \Delta t \sum_{j=0}^{r-1} \int_{t_j}^{t_i} x(t_{i+j}-s)x(s)\,ds. \tag{4.3}$$

Further investigation of (4.3) leads us to the following cases:

- If $i < r-1$, then

$$\begin{aligned} &\int_0^R \int_\rho^{t_i} x(t_i+\rho-s)\,x(s)\,ds\,d\rho \\ &\quad\approx (\Delta t)^2 \left[\sum_{m=1}^{i} \sum_{l=m}^{i} c_l\,c_{i+m-l} - \sum_{m=i+1}^{r-1} \sum_{l=i+1}^{m} c_l\,c_{i+1+m-l} \right], \end{aligned}$$

  which is nonlinear in c_i and involves all values of $c_1, c_2, \ldots, c_{r-1}$.
- If $i = r-1$, then

$$\int_0^R \int_\rho^{t_i} x(t_i+\rho-s)\,x(s)\,ds\,d\rho \approx (\Delta t)^2 \sum_{m=1}^{i} \sum_{l=m}^{i} c_l\,c_{i+m-l},$$

  which is also nonlinear in c_i and involves all values of $c_1, c_2, \ldots, c_{r-1}$.

- If $i \geq r$, then

$$\int_0^R \int_\rho^{t_i} x(t_i + \rho - s)\, x(s)\, ds\, d\rho \approx (\Delta t)^2 \sum_{m=1}^{r} \sum_{l=m}^{i} c_l\, c_{i+m-l}.$$

With this result, we notice that, specifically for $i = r$, (4.3) is quadratic in c_r once $c_1, c_2, \ldots, c_{r-1}$ are found. Even better, for $i > r$, (4.3) is linear in c_i once $c_1, c_2, \ldots, c_r$ are found.

Therefore, to determine the values for c_i, $i = 1, 2, \ldots, N$, we solve the following N equations:

- For $i = 1, 2, \ldots, r-2$, we have the first $r-2$ equations.

$$\left(2(\Delta t)^2 \sum_{l=1}^{r-1} \sum_{m=1}^{l} c_m\right) c_i$$

$$+ (\Delta t)^2 \left[\sum_{m=1}^{i} \sum_{l=m}^{i} c_l\, c_{i+m-l} - \sum_{m=i+1}^{r-1} \sum_{l=i+1}^{m} c_l\, c_{i+1+m-l}\right] = \Delta t \sum_{q=0}^{r+1} f^\delta(t_{i+q}).$$

Note that all values of $c_1, c_2, \ldots, c_{r-1}$ are involved in any of these $r-2$ equations.

- For $i = r-1$, we have the $(r-1)$th equation.

$$\left(2\,(\Delta t)^2 \sum_{l=1}^{r-1} \sum_{m=1}^{l} c_m\right) c_i + (\Delta t)^2 \sum_{m=1}^{i} \sum_{l=m}^{i} c_l\, c_{i+m-l} = \Delta t \sum_{q=0}^{r+1} f^\delta(t_{i+q}).$$

Note that all values of $c_1, c_2, \ldots, c_{r-1}$ are also involved in this $(r-1)$th equation. Since we now have $r-1$ equations with $r-1$ unknowns, we are ready to solve $c_1, c_2, \ldots, c_{r-1}$. We would expect the procedure to this point to be computationally intensive since we are basically solving a system of nonlinear equations for $c_1, c_2, \ldots, c_{r-1}$.

- Once $c_1, c_2, \ldots, c_{r-1}$ are found, we can solve for $c_r, \ldots, c_N$ sequentially by

$$\left(2\,(\Delta t)^2 \sum_{l=1}^{r-1} \sum_{m=1}^{l} c_m\right) c_i + (\Delta t)^2 \sum_{m=1}^{r} \sum_{l=m}^{i} c_l\, c_{i+m-l} = \Delta t \sum_{q=0}^{r+1} f^\delta(t_{i+q}) \tag{4.4}$$

for $i = r, r+1, \ldots, N$. As noted before, we still have to solve a nonlinear equation for c_r since the rth equation is quadratic in c_r. However, once $c_1, c_2, \ldots, c_r$ have already been determined, the remaining $N-r$ equations can be sequentially solved quickly since the ith equation is linear in c_i for $i > r$.

Before turning to numerical examples for this discretization, we first examine whether the method can be modified slightly on $[0, R]$ so that the resulting numerical algorithm avoids the difficulties seen above for obtaining c_i when $i = 1, \ldots, r$.

5. Alternate methods on $[0, R]$. As we have just seen, a standard discretization of the regularization method associated with Theorem 3.1 leads to a sequential numerical method for recovering $x(t)$ for $R < t \leq 1$. Unfortunately, for the $x(t)$ values on the interval $0 \leq t \leq R$, we have to solve a nonsequential nonlinear system of equations which can be numerically expensive. This disadvantage motivates us to

look for cheaper alternatives to recover x on the interval $0 \leq t \leq R$. In this section, we will propose some alternative methods and give a theoretical basis for them.

In what follows we will show that, for any function x_R sufficiently close to $\bar{x}$ on $[0, R]$, we may find a unique $\hat{x}_R^\delta \in L_2(0,1)$ for which

$$\hat{x}_R^\delta(t) = x_R(t), \quad t \in [0, R],$$

and such that $\hat{x}_R^\delta$ satisfies (2.4) on the restricted interval $(R, 1]$. That is, $\hat{x}_R^\delta \in L_2(0,1)$ satisfies

$$\alpha_R(x)\, x(t) + F_R(x)(t) = f_R^\delta(t), \quad t \in (R, 1]. \tag{5.1}$$

Further, we will see that, under suitable conditions on the true solution $\bar{x}$ and the choice of $R = R(\delta)$, the function $\hat{x}_R^\delta$ is a good approximation of $\bar{x}$ for δ small. The advantage to this new approach is that we are free to find easier ways of determining an approximation x_R to $\bar{x}$ on $[0, R]$ than that obtained by solving (2.4) on the interval $[0, R]$.

THEOREM 5.1. *Assume that f^δ satisfies the $\mathcal{F}$-data condition, and let $\tau_{\text{data}} = 1/2$ in the case of $\mathcal{F} = C[0, 1+\bar{R}]$ and $\tau_{\text{data}} = 2/5$ in the case of $\mathcal{F} = L_2(0, 1+\bar{R})$. Assume that the measure η is given satisfying* (2.8)–(2.9) *and that the autoconvolution problem* (1.1) *has a positive solution $\bar{x} \in W^{2,\infty}[0, 1+\bar{R}]$ satisfying*

$$\bar{x}(0) > 13\, b^2\, e^{2\sigma}\, \|\bar{x}'\|_\infty \tag{5.2}$$

for $\sigma \geq \sigma_0$ and $b \geq 2\bar{\omega}/\underline{\omega}$, where $\bar{\omega}$, $\underline{\omega}$ are given in (2.9). *Assume further that $\{x_R\}_{R \in (0, \bar{R}]}$ is any family of $L_\infty(0, R)$ functions satisfying*

$$\sup_{t \in [0,R]} |x_R(t) - \bar{x}(t)| \leq \tilde{C} R \tag{5.3}$$

for $\tilde{C} > 0$ satisfying

$$\tilde{C} \leq b\, \|\bar{x}'\|_\infty. \tag{5.4}$$

Then there exist constants $k_1 > 0$ and $\hat{C} > 0$, all independent of R, such that, if $\delta = \delta(R) > 0$ satisfies

$$\delta \leq k_1 R^{1/\tau_{\text{data}}}, \tag{5.5}$$

then for all $R > 0$ sufficiently small there exists a unique $\hat{x}_R^\delta \in L_2(0,1)$, with $\hat{x}_R^\delta = x_R(t)$, $t \in (0, R]$, with $\hat{x}_R^\delta$ satisfying (2.4) *for $t \in (R, 1]$ and for which*

$$\|\hat{x}_R^\delta - \bar{x}\|_{\sigma,R}^2 \leq \hat{C}^2 R^2,$$

where $\|\cdot\|_{\sigma,R}$ is defined from (2.11) *using $\mathcal{C} \in (1, \frac{13}{12}]$. Finally, if $x_R \in L_\infty(0, R)$ depends continuously on data $f^\delta \in \mathcal{F}$, then so does $\hat{x}_R^\delta \in L_2^{\sigma,R}(0,1)$ for all $R > 0$ sufficiently small.*

Proof. To begin, let $\hat{C} < \frac{\bar{x}(0)}{2\, e^{2\sigma}\, b}$, and define a new ball around $\bar{x}$ by

$$\hat{\mathcal{B}}_R(\bar{x}) \equiv \{x \in L_2^{\sigma,R}(0,1) : x|_{\,[0,R]} = x_R|_{\,[0,R]}, \|x - \bar{x}\|_{\sigma,R} \leq \hat{C} R\}, \tag{5.6}$$

noting that $\hat{\mathcal{B}}_R(\bar{x}) \subseteq \mathcal{B}(\bar{x}, \hat{C}R)$, where $\mathcal{B}(\bar{x}, \hat{C}R)$ was fundamental to the proof of Theorem 3.7.

Defining the new operator $\bar{H}_R$ via

$$\bar{H}_R(x)(t) = \begin{cases} x(t) & \text{if } t \in [0, R], \\ H_R x(t) & \text{if } t \in (R, 1], \end{cases}$$

where H_R was defined in (3.19), we will use the contraction mapping theorem to show that the operator $\bar{H}_R(x)$ has a fixed point in the ball $\hat{\mathcal{B}}_R(\bar{x})$. Let $x \in \hat{\mathcal{B}}_R(\bar{x})$. Then since $H_R x = x = x_R$ on $[0, R]$ and

$$\int_0^R e^{-2\sigma t} \left(x_R(t) - \bar{x}(t)\right)^2 \, dt \le \tilde{C}^2 R^3 \tag{5.7}$$

we have

$$\begin{aligned} \|\bar{H}_R x - \bar{x}\|_{\sigma,R}^2 &\le \tilde{C}^2 R^3 \left(\frac{\mathcal{C}^2 - 1}{R} + 1\right) + \int_R^1 e^{-2\sigma t} \left(H_R x(t) - \bar{x}(t)\right)^2 \, dt \\ &\le \tilde{C}^2 R^2 \left(\mathcal{C}^2 - (1 - R)\right) + \|H_R(x) - \bar{x}\|_{\sigma,R}^2 \\ &\le (\tilde{C}\mathcal{C}R)^2 + \|H_R(x) - \bar{x}\|_{\sigma,R}^2 \end{aligned}$$

for $R > 0$ sufficiently small. Using the proof of Theorem 3.7 (in particular, inequality (3.26)), we find that a sufficient condition for $\|\bar{H}_R x - \bar{x}\|_{\sigma,R} \le \hat{C}R$ is that

$$\tilde{C}\mathcal{C} + \frac{1}{\bar{x}(0)} \, k_1 b \, \mathcal{C} + \|\bar{x}'\|_\infty \, b \, \mathcal{C} + \frac{e^{2\sigma}}{\bar{x}(0)} \, b \, \hat{C}^2 < \hat{C}, \tag{5.8}$$

so that, letting k_1 be defined by (3.27) and using condition (5.4) on $\tilde{C}$, we find that inequality (5.8) holds if

$$L(\hat{C}) \equiv \frac{e^{2\sigma} \, b}{\bar{x}(0)} \, \hat{C}^2 - \hat{C} + 3 \, b \, \|\bar{x}'\|_\infty \, \mathcal{C} < 0.$$

In fact, L has two distinct positive real roots $0 < \hat{C}_1 < \hat{C}_2$ under assumption (5.2) and $\mathcal{C} \in (1, \frac{13}{12}]$. Thus any value of $\hat{C}$ in the interval $(\hat{C}_1, (\hat{C}_1 + \hat{C}_2)/2)$ will give $L(\hat{C}) < 0$ and also ensure that $\hat{C} < \frac{\bar{x}(0)}{2 \, e^{2\sigma} \, b}$ as prescribed at the beginning of the proof. Thus $\bar{H}_R$ takes $\hat{\mathcal{B}}_R(\bar{x})$ into itself.

Let $x_1, x_2 \in \hat{\mathcal{B}}_R(\bar{x})$. Then

$$\begin{aligned} \|\bar{H}_R(x_1) - \bar{H}_R(x_2)\|_{\sigma,R}^2 &= 0 + \int_R^1 e^{-2\sigma t} \left(H_R(x_1)(t) - H_R(x_2)(t)\right)^2 dt \\ &\le \|H_R(x_1) - H_R(x_2)\|_{\sigma,R}^2, \end{aligned}$$

and thus from the proof of Theorem 3.7 we have that the conditions already placed on $\hat{C}$ are sufficient for $\bar{H}_R$ to be a contraction in the ball $\hat{\mathcal{B}}_R(\bar{x})$.

Finally, we refine our definition of $\hat{C}$ further to ensure that the ball $\hat{\mathcal{B}}_R(\bar{x})$ is nonempty. We note from (5.4) and (5.2) that

$$\tilde{C} \le \frac{\bar{x}(0)}{13 \, b \, e^{2\sigma}} = \frac{2}{13} \frac{\hat{C}_1 + \hat{C}_2}{2},$$

so that, if we further restrict $\hat{C}$ to lie in the (nonempty) intersection of the intervals $(\frac{2}{13}(\hat{C}_1+\hat{C}_2)/2, (\hat{C}_1+\hat{C}_2)/2)$ and $(\hat{C}_1, (\hat{C}_1+\hat{C}_2)/2)$, we have that

$$\tilde{C} \leq \hat{C}.$$

Thus if we define $\hat{x}_R \in L_2^{\sigma,R}(0,1)$ via

$$\hat{x}_R = \begin{cases} x_R(t) & \text{if } t \in [0,R], \\ \bar{x}(t) & \text{if } t \in (R,1], \end{cases}$$

we have from (5.7) that

$$\|\hat{x}_R - \bar{x}\|_{\sigma,R}^2 \leq \hat{C}^2 R^2 \left(\mathcal{C}^2 - 1 + R\right) + 0 \leq \hat{C}^2 R^2 \tag{5.9}$$

for $R > 0$ sufficiently small. Thus $\hat{x}_R \in \hat{\mathcal{B}}_R(\bar{x})$.

Continuous dependence of $\hat{x}_R^\delta$ on data f^δ follows arguments similar to those in Theorem 3.7. ∎

As in Corollary 3.8 we may dispense with condition (5.2) provided that we have access to an appropriate approximation $\hat{x}(0;R)$ of $\bar{x}(0)$ and that we make use of this information in the construction of our measure η. See Remark 3.3 for more information on the construction of such a measure.

COROLLARY 5.2. *Assume that f^δ satisfies the $\mathcal{F}$-data condition, and let $\tau_{\text{data}} = 1/2$ in the case of $\mathcal{F} = C[0,1+\bar{R}]$ and $\tau_{\text{data}} = 2/5$ in the case of $\mathcal{F} = L_2(0,1+\bar{R})$. Let the solution $\bar{x} \in W^{2,\infty}[0,1+\bar{R}]$ of (1.1) be positive. Assume that the measure η satisfies (2.8)–(2.9) with $\bar{\omega}$, $\underline{\omega}$ in (2.9) satisfying (3.4) for all $R > 0$ sufficiently small, where $\tilde{K}$ is independent of R and $\bar{x}$. Assume further that $\{x_R\}_{R\in(0,\bar{R}]}$ is any family of $L_\infty(0,1)$ functions satisfying*

$$\sup_{t\in[0,R]} |x_R(t) - \bar{x}(t)| \leq \tilde{C}R \tag{5.10}$$

for some $\tilde{C} > 0$.

Then there exist constants $k_1 > 0$ and $\hat{C} > 0$, all independent of R, such that, if $\delta = \delta(R) > 0$ satisfies

$$\delta \leq k_1 R^{1/\tau_{\text{data}}}, \tag{5.11}$$

then there exists a unique $\hat{x}_R^\delta \in L_2(0,1)$, with $\hat{x}_R^\delta = x_R(t)$, $t \in (0,R]$, with $\hat{x}_R^\delta$ satisfying (2.4) for $t \in (R,1]$ and for which

$$\|\hat{x}_R^\delta - \bar{x}\|_{\sigma,R}^2 \leq \hat{C}^2 R^2,$$

where $\|\cdot\|_{\sigma,R}$ is defined from (2.11) using $\mathcal{C} = \sqrt{2}$ and $\sigma \geq \sigma_0$. Finally, if it is known that $x_R \in L_\infty(0,R)$ depends continuously on data $f^\delta \in \mathcal{F}$, then so does $\hat{x}_R^\delta \in L_2^{\sigma,R}(0,1)$ for all $R > 0$ sufficiently small.

Proof. Define b satisfying both $b \geq 2\bar{\omega}/\underline{\omega}$ and $\tilde{C} \leq b\|\bar{x}'\|_\infty$, and let $\tilde{b} = b/\bar{x}(0)$. As in Corollary 3.8 we may assume without loss of generality that the original equation (1.1) has been rescaled so that

$$\|\bar{x}'\|_\infty \bar{x}(0) < \frac{\sqrt{2}\, e^{-2\sigma}}{24\,\tilde{b}^2}. \tag{5.12}$$

Then following the proof of Theorem 5.1 a sufficient condition for $\|\bar{H}_R x - \bar{x}\|_{\sigma,R} < \hat{C}R$ is that

$$L(\hat{C}) \equiv e^{2\sigma}\,\tilde{b}\,\hat{C}^2 - \hat{C} + 3\,\|\bar{x}'\|_\infty \bar{x}(0)\,\tilde{b}\,\mathcal{C} < 0,$$

so for $\mathcal{C} = \sqrt{2}$ and under the rescaling in (5.12) we have that L has two distinct positive roots $0 < \hat{C}_1 < \hat{C}_2$. The remainder of the proof of Theorem 5.1 then carries over exactly as before, with the exception that we now have from (5.12)

$$\tilde{C} \le \tilde{b}\|\bar{x}'\|_\infty \bar{x}(0) < \frac{\sqrt{2}}{12}\frac{1}{2e^{2\sigma}\tilde{b}} = \frac{\sqrt{2}}{12}\left(\frac{\hat{C}_1 + \hat{C}_2}{2}\right),$$

so that the arguments in the proof of Theorem 5.1 remain valid if we select $\hat{C}$ in the (nonempty) intersection of the intervals $(\frac{\sqrt{2}}{12}(\hat{C}_1+\hat{C}_2)/2,\ (\hat{C}_1+\hat{C}_2)/2)$ and $(\hat{C}_1,\ (\hat{C}_1+\hat{C}_2)/2)$. ∎

Remark 5.3. The results of Theorem 5.1 and Corollary 5.2 indicate that, as before, under the a priori conditions on $R = R(\delta)$ that

$$(\delta/k_1)^{\tau_{\text{data}}} \le R(\delta) \le c\,\delta^p$$

for any $c > 0$ and $p \in (0, \tau_{\text{data}}]$ ($c \ge k_1^{-\tau_{\text{data}}}$ if $p = \tau_{\text{data}}$) as $\delta \to 0$, we obtain

$$\|x^\delta_{R(\delta)} - \bar{x}\|_{L_2(0,1)} = \mathcal{O}\left(\delta^p\right) \ \text{ as } \ \delta \to 0,$$

with an optimal rate obtained for $p = \tau_{\text{data}}$. Thus the optimal convergence rate for continuous data is $\mathcal{O}(\delta^{1/2})$, and the optimal convergence rate for L_2 data is $\mathcal{O}(\delta^{2/5})$.

As mentioned earlier, Theorem 5.1 and Corollary 5.2 provide us the freedom of finding easier ways of determining an approximation x_R to $\bar{x}$ on $[0, R]$ than that obtained by solving the local regularization equation (2.4) on the interval $[0, R]$. One way to find such an x_R (and to find an $\mathcal{O}(R)$ approximation $\hat{x}(0; R)$ of $\bar{x}(0)$, which is of particular use in Corollaries 3.8 and 5.2 via Remark 3.3) is by a standard discretization of the unregularized problem

$$G(x)(t) = \int_0^t x(t-s)x(s)\,ds = f^\delta(t), \quad t \in [0, R] \tag{5.13}$$

(with noisy data f^δ), on the interval $[0, R]$. As is well known, coarse discretization is a form of regularization; however, the ill-posedness of (5.13) can lead to significant errors in $x_R(t)$ for large t, even more so when the discretization step size is small. It is for this reason that the standard discretization will be used only on a small interval of length R.

We will describe here one simple discretization of (5.13) which is suitable for use with Theorem 5.1 and Corollary 5.2. Let $K = K(R) \ge 1$ be an integer, and partition the interval $[0, R]$ into K equal-length subintervals; i.e., let

$$t_i = i\Delta t, \quad i = 0, 1, \ldots, K, \quad \Delta t = R/K.$$

Defining $\chi_i(t)$ as in the last section (only now for $i = 1, 2, \ldots, K$), we define $\mathcal{S}_K = \text{span}\{\,\chi_i,\ i = 1, 2, \ldots, K\}$ and seek $x_R \in \mathcal{S}_K$, i.e.,

$$x_R(t) = \sum_{l=1}^{K} c_l\,\chi_l(t), \tag{5.14}$$

satisfying the collocation equations

$$G(x)(t_i) = \int_0^{t_i} x(t_i - s)x(s)ds = f^\delta(t_i), \quad i = 1, 2, \ldots, K. \tag{5.15}$$

Therefore,

$$\sum_{\gamma=1}^{i} \int_{t_{\gamma-1}}^{t_\gamma} \left[\sum_{l=1}^{K} c_l\, \chi_l(t_i - s)\right] \left[\sum_{p=1}^{K} c_p\, \chi_p(s)\right] ds = f^\delta(t_i), \quad i = 1, 2, \ldots, K,$$

or

$$\sum_{\gamma=1}^{i} c_{i-\gamma+1}\, c_\gamma = \frac{f^\delta(t_i)}{\Delta t}, \quad i = 1, 2, \ldots, K. \tag{5.16}$$

The collocation equations (5.16) allow us to explicitly solve for c_i, $i = 1, 2, \ldots, K$, provided that $f^\delta(t_1) > 0$. That is,

$$c_1 = \sqrt{\frac{f^\delta(t_1)}{\Delta t}}, \tag{5.17}$$

and if $c_1, \ldots, c_{i-1}$ have been found already, c_i is determined by

$$c_i = \frac{\frac{f^\delta(t_i)}{\Delta t} - (c_{i-1}c_2 + \cdots + c_2 c_{i-1})}{2c_1}, \tag{5.18}$$

and therefore $c_2, c_3, \ldots, c_K$ can be found uniquely, explicitly, and sequentially from (5.18).

The following result is proved using standard techniques (see the details in [8]). If the constant $\tilde{C}$ in (5.21) is not too large (a condition needed for Theorem 5.1 only, not for Corollary 5.2), it follows that the collocation-based approximation x_R is suitable for use in methods based on Theorem 5.1 or Corollary 5.2. That is, we may use x_R as an approximation for $\bar{x}$ on $[0, R]$ and then use the local regularization equation (2.4) to find an approximation for $\bar{x}$ on $[R, 1]$. We note that the condition $K = K(R) \le \tilde{M}$ for all $R > 0$ sufficiently small is not a serious constraint as we generally expect $K \ge 1$ to be a fixed integer so that the step size $\Delta t = R/K$ decreases as $R \to 0$. However, not surprisingly since we are now dealing with a discretization of the original problem, we do need pointwise values of f^δ on $[0, R]$, and we require that they be within $\mathcal{O}(R^2)$ of the corresponding values of f on that interval.

COROLLARY 5.4. *Assume the autoconvolution problem* (1.1) *has a positive solution* $\bar{x} \in C^1[0, 1]$. *Let* $x_R = \sum_{l=1}^{K} c_l\, \chi_l(t)$ *be the unique solution of the discrete autoconvolution equation* (5.15) *on the interval* $[0, R]$, *where the constants* c_i, $i = 1, 2, \ldots, K$, *are specified in* (5.17) *and* (5.18) *and* $K = K(R) \ge 1$ *is an integer. Then if there exists a constant* $\tilde{M} > 0$, *such that* $K = K(R) \le \tilde{M}$ *uniformly in* R *and* f^δ *is piecewise continuous on* $[0, R]$ *with*

$$|f^\delta(t) - f(t)| \le \delta \le k_1 R^2, \qquad t \in [0, R], \tag{5.19}$$

convergence of $x_R(t)$ *to the true solution* $\bar{x}(t)$ *for* $t \in [0, R]$ *occurs at the collocation points as* $R \to 0$, *i.e.,*

$$|x_R(t_i) - \bar{x}(t_i)| \sim \mathcal{O}(R) \quad \text{for } i = 1, 2, \ldots, K \tag{5.20}$$

for R sufficiently small. Further, we have a constant $\tilde{C}$ depending on $\bar{x}$ but independent of R such that

$$|x_R(t) - \bar{x}(t)| \le \tilde{C}R \tag{5.21}$$

for R sufficiently small and all $t \in [0, R]$.

Remark 5.5. It is worth noting that the results of Theorem 5.1 are improved if we require that x_R be an $\mathcal{O}(R^p)$ approximation to $\bar{x}$ on $[0, R]$ for some $p > 1$ [8]. This result is of use, for example, if we know both $\bar{x}(0)$ and $\bar{x}'(0)$ so that we form an $\mathcal{O}(R^2)$ approximation of $\bar{x}$ on $[0, R]$ via $x_R(t) = \bar{x}(0) + \bar{x}'(0)t$. However, as can be seen in [8], such approximations on $[0, R]$ actually perform worse (locally) in numerical examples than do approximations based on the simple $\mathcal{O}(R)$ collocation-based discretization of (5.13) given above. The reason for this is likely due to the fact that the approximation based on a Taylor expansion does not make good use of all of the available data f^δ on $[0, R]$, as does the collocation-based discretization.

5.1. Operation count for the alternative method. If we let $r = K$ in (5.17)–(5.18), then the cost of computing $c_1, c_2, \ldots, c_r$ from (5.17)–(5.18) is $\mathcal{O}(r^2/2)$ flops. The remaining steps in the alternative approach require the sequential solution of (4.4) for $c_{r+1}, \ldots c_N$. We can rewrite (4.4) for c_i as

$$(D_1 + D_2)c_i + D_3(i) = D_4(i), \quad i = r + 1, \ldots, N,$$

where D_j, $j = 1, \ldots, 4$, are constants which are independent of i in the case of D_1 and D_2,

$$D_1 = \left(2 \sum_{l=1}^{r-1} \sum_{m=1}^{l} c_m \right), \quad D_2 = 2 \sum_{m=1}^{r} c_m,$$

$$D_3(i) = \sum_{m=1}^{r} \sum_{l=m+1}^{i-1} c_l \, c_{i+m-l}, \quad D_4(i) = \frac{1}{\Delta t} \sum_{q=0}^{r+1} f^\delta(t_{i+q}).$$

The computation of D_1 and D_2 require one-time costs of order $\mathcal{O}(r)$ flops each, while the cost of $D_3(i)$ and $D_4(i)$ are $\mathcal{O}(2ir - r^2 - 3r)$ and $\mathcal{O}(r)$ flops, respectively, computations which are done for $i = r + 1$ to $i = N$. Thus the total cost of the alternative method described in this section (assuming $r \ll N$) is $\mathcal{O}(rN^2 - r^2N)$.

Operation counts for the Lavrent'ev method are likely to be similar to those found above (i.e., $\mathcal{O}(cN^2)$ for some constant $c > 0$); however, fairly accurate knowledge of the value of $\bar{x}(0)$ is of importance if one is to make use of this method; in addition, one requires an initial guess $x^\star$ for implementation of Lavrent'ev. For comparison with methods which do not preserve the causal nature of the nonlinear autoconvolution problem, we have been unable to find operation counts for Tikhonov regularization applied to this nonlinear problem (and of course there are many different approaches—methods using Newton or Levenberg–Marquardt, descent methods such as steepest descent, etc.—that one might take to solve the nonlinear Tikhonov problem). Without a suitable comparison, we will simply note that cost for a direct implementation of Tikhonov regularization on a general *linear* problem is $\mathcal{O}(N^3)$ flops [10]. Some actual computational costs (in seconds) are provided in Examples 3 and 4 below.

6. Numerical results. The examples in this section provide evidence of the effectiveness of the local regularization methods applied to the autoconvolution equation. The local regularization methods have advantages over existing methods in that they do not require an initial guess $x^\star$ and they maintain the causal nature of the problem at least for the majority of the domain, leading to very rapid solution methods. One drawback, of course, is that the theory requires that true solutions $\bar{x}$ should not cross the t-axis. However, as Examples 3 and 4 below illustrate, the method appears to perform perfectly well in this situation.

Henceforth we will refer to *method* 1 as the collocation-based discretization derived in section 4 for the local regularization equation (2.4), i.e., a discretization of the continuous local regularization method which is rigorously justified by Theorem 3.7. Recall that method 1 requires solving nonsequential nonlinear equations for $c_1, \dots, c_r$ and sequential linear equations for $c_{r+1}, \dots, c_N$.

In contrast, *method* 2 will refer to the method of determining $c_1, \dots, c_r$ via (5.17)–(5.18) (where we set $K = r$) and $c_{r+1}, \dots, c_N$ via (4.4). In method 2, all coefficients are found using a procedure that is fully sequential and linear in the unknown coefficient at each step. This algorithm is a discretization of the continuous local regularization method which is rigorously justified by Theorem 5.1. It is interesting to note in our numerical examples below that the fully sequential linear method 2, which is significantly easier to implement and considerably faster in terms of computational times, actually performs as well as the (slower) partly nonlinear method 1.

In order for easy comparison with numerical tests in the literature for existing methods (a Tikhonov regularization approach using total variation constraints in [11] and a Lavrent'ev approach in [16]), we demonstrate the recovery of both a continuous $\bar{x}$ and a discontinuous $\bar{x}$ using methods 1 and 2 above. We select our true solution $\bar{x}$ ahead of time and then generate the data function f by integration: $f(t) = \int_0^t \bar{x}(t-s)\bar{x}(s)\,ds$ for $t \in [0,1]$. The perturbed data f^δ was then produced by adding uniformly distributed noise from the interval $[-\delta f(t), \delta f(t)]$ to the discrete values of $f(t)$ for $t = t_i$, where $i = 1, 2, \dots, N$. Each example within the same noise level δ was found using the same sampled data f^δ.

In the figures below, we display the recovered solution against the true solution $\bar{x}$, first without any special regularization (except for that due to discretization alone) and then with regularization using methods 1 and 2. In all figures presented below, the true solution $\bar{x}$ is plotted as a dashed line, while the solid line expresses the approximate solution computed according to the specified method. The regularized solutions end before reaching $t = 1$ in Examples 1 and 2 because we have chosen not to extend our data past the interval $[0,1]$ in those examples; to have regularized solutions on all of $[0,1]$, one needs to make use of data f^δ on $[0, 1+R]$.

6.1. Example 1: Continuous true solution $\bar{x}$. The true solution in this example is a continuous function $\bar{x}(t) = 1 - 3(t-1/2)^2$, $0 \le t \le 1$, with the true data then given by $f(t) = \frac{3}{10}t^5 - \frac{3}{2}t^4 + t^3 + \frac{3}{4}t^2 + \frac{1}{16}t$ for $0 \le t \le 1$. For Figures 6.1–6.3, we use relative noise level $\delta = 10^{-3}$, $N = 100$, and $r = 4$ for the regularized problems. In Figure 6.1, we show the unregularized numerical solution, while, in Figures 6.2–6.3, we show the regularized solution using methods 1 and 2 as introduced in the last section.

For Figures 6.4–6.6, we repeat the above example using a larger relative noise level $\delta = 10^{-2}$, $N = 100$, and $r = 4$ for method 1 and $r = 6$ for method 2. (We selected the value of r small, which gave the best results in each case visually.)

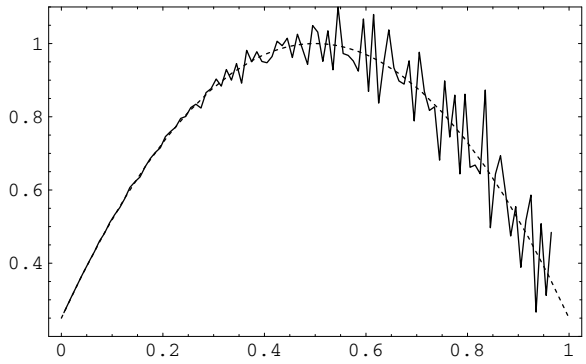


FIG. 6.1. *Example* 1. *Solution without regularization,* $\delta = 10^{-3}$, $N = 100$.

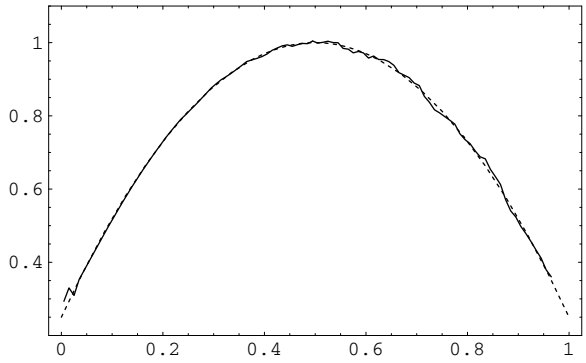


FIG. 6.2. *Example* 1. *Solution obtained by method* 1, $\delta = 10^{-3}$, $N = 100$, $r = 4$.

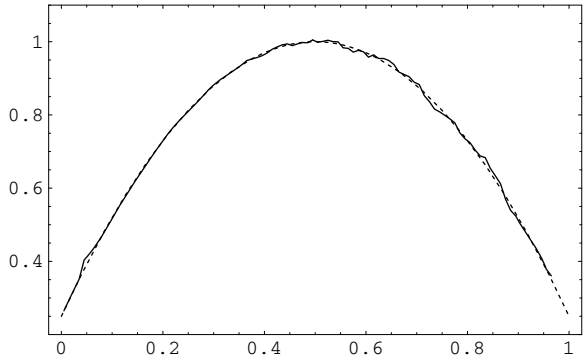


FIG. 6.3. *Example* 1. *Solution obtained by method* 2, $\delta = 10^{-3}$, $N = 100$, $r = 4$.

6.2. Example 2: Discontinuous true solution $\bar{x}$. In this example, we will use a step function as our true solution; $\bar{x}$ and true data f are given by

$$\bar{x}(t) = \begin{cases} 0.5 & \text{if } t \in [0, 0.5], \\ 0.25 & \text{if } t \in (0.5, 0.8], \\ 0.75 & \text{if } t \in (0.8, 1], \end{cases} \qquad f(t) = \begin{cases} 0.25t & \text{if } t \in [0, 0.5], \\ 0.125 & \text{if } t \in (0.5, 0.8], \\ 0.5t - 0.275 & \text{if } t \in (0.8, 1]. \end{cases}$$

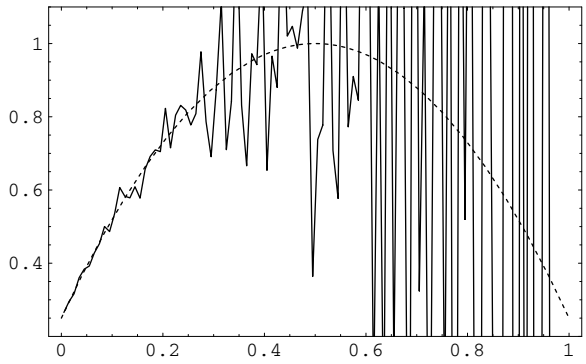


FIG. 6.4. *Example* 1. *Solution without regularization,* $\delta = 10^{-2}$, $N = 100$.

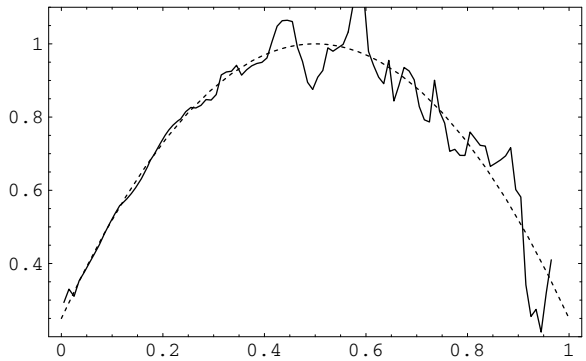


FIG. 6.5. *Example* 1. *Solution obtained by method* 1, $\delta = 10^{-2}$, $N = 100$, $r = 4$.

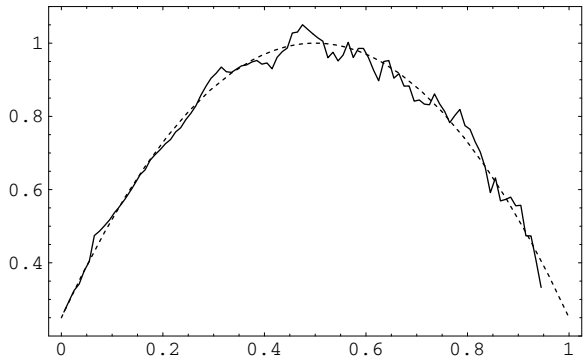


FIG. 6.6. *Example* 1. *Solution obtained by method* 2, $\delta = 10^{-2}$, $N = 100$, $r = 6$.

For all of the figures shown in this example, we use relative noise level $\delta = 10^{-3}$, $N = 200$, and $r = 4$ for the regularized problem. In Figure 6.7, we show the solution without any special regularization, while Figures 6.8–6.9 show the regularized solution using methods 1 and 2. The results for both methods are much better than was presented in [11, 16]. It is worth noting that the true solution $\bar{x}$ here does not satisfy the assumptions of Theorem 3.7 (which are only sufficient conditions for convergence anyway).

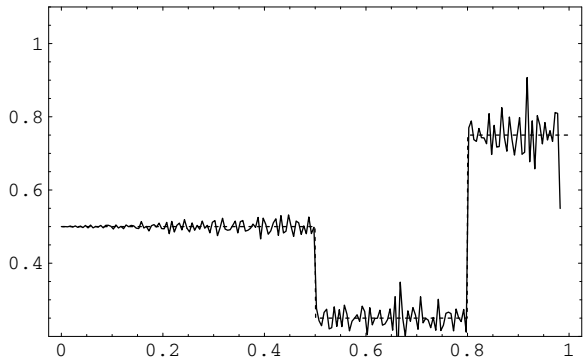


FIG. 6.7. *Example* 2. *Solution without regularization,* $\delta = 10^{-3}$, $N = 200$.

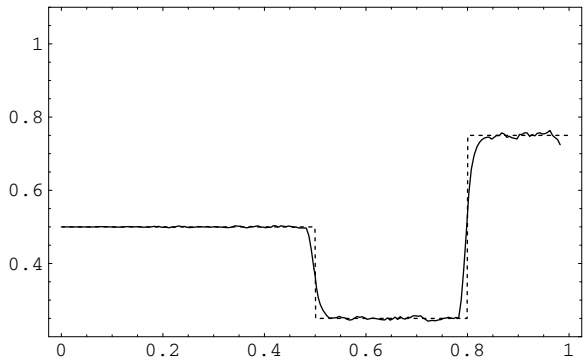


FIG. 6.8. *Example* 2. *Solution obtained by method* 1, $\delta = 10^{-3}$, $N = 200$, $r = 4$.

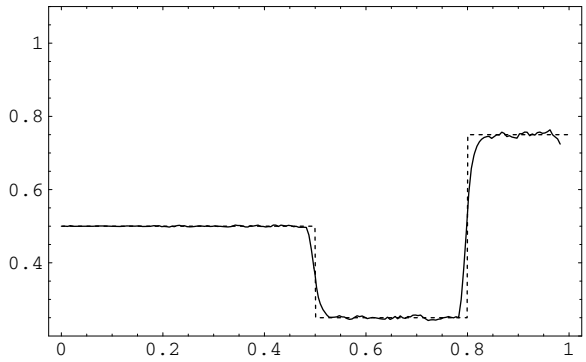


FIG. 6.9. *Example* 2. *Solution obtained by method* 2, $\delta = 10^{-3}$, $N = 200$, $r = 4$.

For comparison with local regularization, we show in Figure 6.10 the results of Tikhonov regularization and Lavrent'ev regularization applied to this same problem. In each case, regularization parameters were selected as those providing the best visual fit between approximate and true solutions. For our implementation of Lavrent'ev, the true value of $\bar{x}(0)$ was incorporated into the approximating equation, and the regularization parameter $\alpha = .01555$ was used. For Tikhonov regularization, we used the parameter $\alpha = 10^{-22}$ with a starting guess of $x(t) = \sqrt{f^{\delta}(t_1)/\Delta t}$; for more

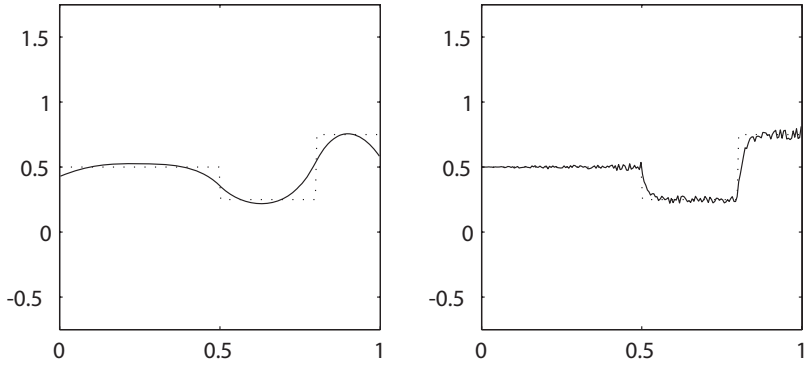


FIG. 6.10. *Example* 2, *continued. Left: Tikhonov solution. Right: Lavrent'ev solution.*

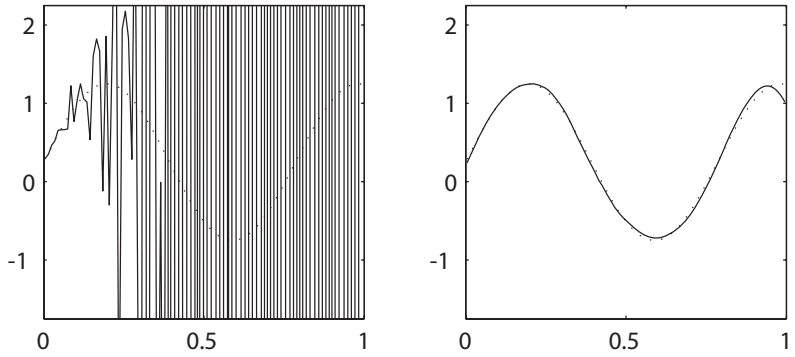


FIG. 6.11. *Example* 3 *with the nonpositive true solution,* $\delta = 10^{-2}$*, and* $N = 100$*. Left: Unregularized solution (computing time:* 0.01 *sec). Right: Tikhonov solution (*1.66 *sec).*

information about our implementation of the nonlinear Tikhonov scheme, refer to Example 3 below.

6.3. Example 3: Nonpositive true solution $\bar{x}$. Although the theory developed here requires that $\bar{x}$ be strictly positive (resp., strictly negative), in practice the numerical methods developed in this paper can still be applied to situations where the unknown $\bar{x}$ crosses the t-axis. We illustrate with an example where the true solution is given by $\bar{x}(t) = \sin 8t + (1/4)$, $0 \le t \le 1$, and the associated true data are $f(t) = \frac{5}{80}(1 + t - \cos 8t + \sin 8t) - \frac{1}{2}t \cos 8t$, $0 \le t \le 1$. We use a relative noise level $\delta = .01$ and $N = 100$ and in Figures 6.11–6.12 show the results of no regularization as well as those for Tikhonov regularization, Lavrent'ev regularization, and local regularization, all applied to the same data f^δ.

In the case of local regularization we use method 2 and $r = 7$ and in so doing make use of data on the extended interval $[0, 1.07]$; for a fair comparison, we also allow Tikhonov regularization to make use of the same extended interval for its computations; however, only the results on $[0, 1]$ are displayed for both methods. We note that the Lavrent'ev method is solved sequentially without the use of future data, so additional data past the interval $[0, 1]$ do not change its result. For the Tikhonov

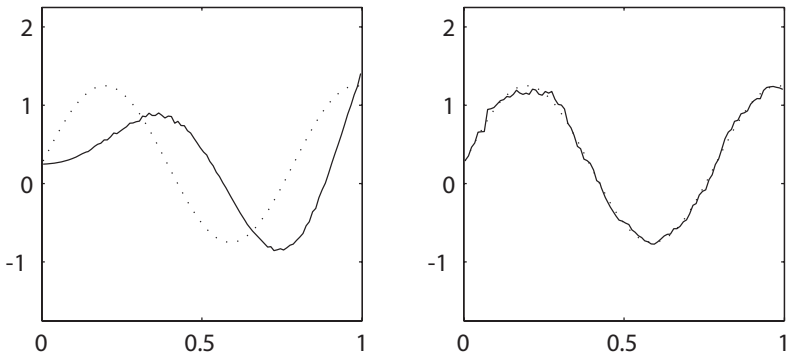


FIG. 6.12. *Example* 3, *continued. Left: Lavrent'ev method (computing time* 0.01 *sec). Right: Local regularization (method* 2*) with* $r = 7$ *(*0.01 *sec).*

and Lavrent'ev methods we use the regularization parameters $\alpha = 10^{-6}$ and $\alpha = 0.4$, respectively, incorporating the true value of $\bar{x}(0)$ into the Lavrent'ev approximating equation. Minimization of the nonlinear Tikhonov regularization functional (with regularization term $\alpha\|x\|^2$) is accomplished in these examples using the MATLAB function `fminunc` with the initial guess provided by the constant function $x(t) = c_1$, where c_1 is the first value of the approximation found in method 2 for local regularization, i.e., $c_1 = \sqrt{f^\delta(t_1)/\Delta t}$. The regularization parameters used for all methods are the best ones found by visual comparison with the true solution.

We note that Lavrent'ev provides a shifted approximation; decreasing α does reduce the shift but at the same time causes large distortions in the shape of the solution. The fact that the Tikhonov method gives a smoother solution than local regularization can be expected due to the global versus local smoothing associated with the two methods. The computing time for Lavrent'ev and local regularization is about the same as that for the unregularized solution (0.01 sec on a 2.21 GHz dual core machine), while Tikhonov regularization requires 166 times that much computing time (1.66 sec).

6.4. Example 4: Nonpositive true solution $\bar{x}$. In this example we take the true $\bar{x}$ from Example 1 above and shift it down so that it crosses the t-axis in two places. That is, we use $\bar{x}(t) = (1/2) - 3(t - 1/2)^2$ for $0 \le t \le 1$, with true data $f(t) = \frac{3}{10}t^5 - \frac{3}{2}t^4 + 2t^3 - \frac{3}{4}t^2 + \frac{1}{16}t$ for $0 \le t \le 1$. In this example we use a relative noise level $\delta = .015$ and $N = 200$, and in Figures 6.13–6.14 we again show the results of no regularization, Tikhonov regularization, Lavrent'ev regularization, and local regularization for the same data f^δ.

The true data f is graphed in Figure 6.15. The difference between the data f in Example 3 (Figure 6.16) and the data f in Figure 6.15 for this example is important in understanding the difference between the way Tikhonov regularization and local regularization perform in this particular case. Note that here the true data f (graphed in Figure 6.15) is nearly zero for the first third to half of the interval. This means that, with even small amounts of added error, the signal-to-noise ratio is likely to be so small as to make the solution unrecoverable if only the data from the first third of the interval are used. Now Tikhonov regularization uses all of the data to compute its regularized approximation, while local regularization constructs a sequential method making use of future data only on the interval $[t, t + r\Delta t]$ to compute the value of the

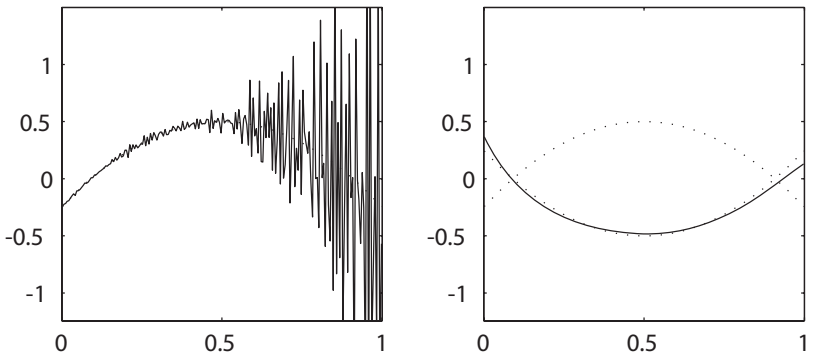


FIG. 6.13. *Example* 4 *with the nonpositive true solution,* $\delta = 1.5 \times 10^{-2}$*, and* $N = 200$*. Left: Unregularized solution (computing time:* 0.003 *sec). Right: Tikhonov solution (*5.43 *sec) finds an approximation of the negative solution* $-\bar{x}$ *(both positive and negative true solutions are graphed with dotted lines above).*

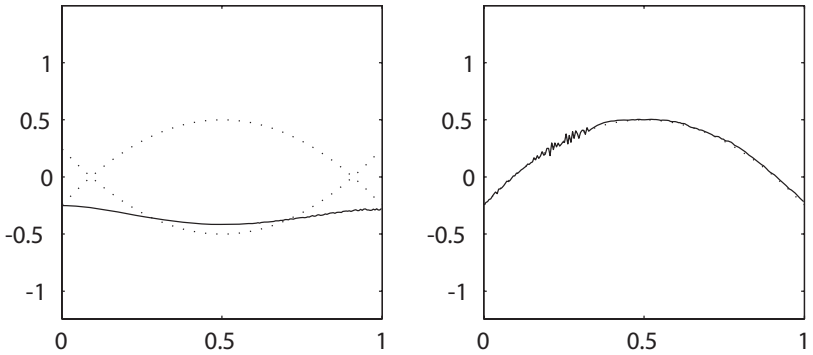


FIG. 6.14. *Example* 4*, continued. Left: Lavrent'ev method (computing time* 0.003 *sec). Right: Local regularization (method* 2*) with* $r = 65$ *(*0.03 *sec).*

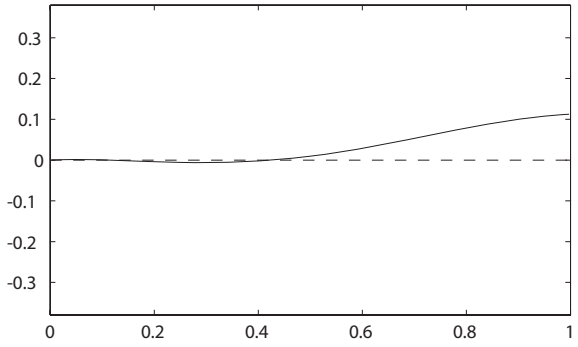


FIG. 6.15. *Example* 4*. True data* f*.*

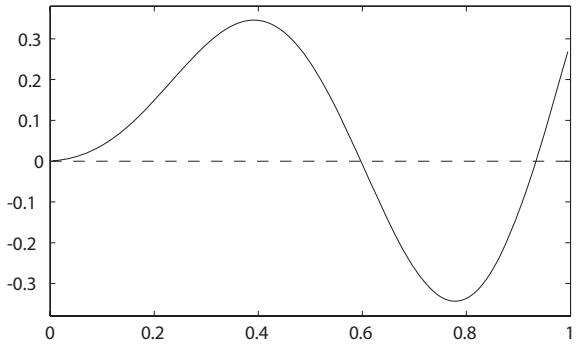


FIG. 6.16. *Example* 3. *True data* f.

approximate solution at t. Thus r will need to be large enough to allow for meaningful information about the data to be obtained in the case of local regularization. In this case, a value of $r = 65$ (nearly 1/3 of the interval) is used in order to get good results with local regularization. In contrast, the data in Example 3 do not exhibit this same difficulty with small signal-to-noise data ratios over large regions, so a small value of r ($r = 7$) is acceptable for that example. Fortunately, the need for large r can be determined by a simple inspection of the given (perturbed) data f^δ prior to use of the local regularization method.

Note further that the Lavrent'ev approximation, also found sequentially, uses no future data information in order to compute the value of the approximation at the current time t. Thus Lavrent'ev can be expected to fail badly in situations where the signal-to-noise ratio is small over large regions of the domain.

As in Example 3, we allow Tikhonov regularization to make use of the same extended data set as that for local regularization, i.e., of data on the interval $[0, 1.325]$. Although the starting value for Tikhonov is the constant function $x(t) = c_1$, where now $c_1 = -\sqrt{f^\delta(t_1)/\Delta t}$ (since $\bar{x}(0) < 0$ in this case), Tikhonov regularization actually finds the negative solution $-\bar{x}$ for this problem. (Recall that there are two solutions $\bar{x}$ and $-\bar{x}$ for every autoconvolution problem; by specifying a positive or negative starting value for each method, we are giving some indication of which of the two solutions to find.)

As expected from the data, Lavrent'ev performs quite poorly in this example. It is possible that the approximation is that for the negative solution of the autoconvolution equation, although the results are not very convincing. On the other hand, local regularization performs quite well provided r is large enough to get a meaningful local sample of the data. We note that for $r \leq 58$ the local regularization approximation blows up at about $t = 1/3$, while acceptable results are found for a large range of values of $r \geq 59$. Thus, a simple inspection of the noisy data is very important in order to obtain information about the smallest of r, which is likely to gain any meaningful information from data on an interval of length $[t, t + r\Delta t]$.

It is worth observing that, with the large value of r used in this example ($r = 65$), the computing time (.03 sec) for local regularization becomes about 10 times that for the unregularized problem and for Lavrent'ev regularization (.003 sec each); on the other hand, the computing time for Tikhonov regularization (at 5.43 sec) is about 180 times that of local regularization. Such differences in computational times

become important if discrepancy principles are to be used in the determination of the correct regularization parameters, since such methods require repeated solution of the approximation for different choices of the parameters.

7. Selection of the regularization parameter. One of the most commonly asked questions with regard to the local regularization methods is how one picks the regularization parameter r (or R in the continuous version of the regularized equation). This question has been answered for the linear convolution problem with the development of a modified discrepancy principle [3, 4] but remains an open problem (other than the need for minimal values of r in some cases, as discussed in Example 4 above) for nonlinear problems, including the nonlinear autoconvolution equation under study here.

The *discrepancy principle* is one of the most successful criteria in determining the regularization parameter α for the method of Tikhonov regularization. To summarize the discrepancy principle on the Hilbert space F, we assume that the perturbed data f^δ have an absolute noise level $\hat{\delta}$, i.e., $\|f^\delta - f\|_F \leq \hat{\delta}$. The Tikhonov theory states that for every choice of $\alpha > 0$ the unique solution u_α^δ of the Tikhonov problem (see, e.g., [10]) associated with the perturbed data f^δ is such that the discrepancy

$$\delta_d \equiv \|\mathcal{A}u_\alpha^\delta - f^\delta\|_F$$

is monotone in α; here $\mathcal{A}$ denotes a linear integral operator. The discrepancy principle is then used to pick α so

$$\delta_d = \tau\hat{\delta},$$

with $\tau \geq 1$ some constant. In practice, τ is often picked as $\sqrt{2}$. Development of the discrepancy principle for Tikhonov regularization of nonlinear operator equations can be found in [30, 32].

In the following, we use the discrepancy principle to select an "optimal" r for a given relative noise level δ and fixed discretization parameter N. Method 2 is used for all numerical experiments that follow. For local regularization the discrepancy δ_d is not exactly a monotone increasing function of the regularization parameter r. Therefore, we modified the principle to select the smallest $r > 0$ for which the discrepancy δ_d exceeds the absolute data noise $\hat{\delta}$.

7.1. Example 1a: A repeat of Example 1, now using the discrepancy principle to select r. In Table 7.1 we illustrate how the modified discrepancy principle can be used to determine r for the two numerical tests (associated with the two different noise levels $\delta = 10^{-3}$ and $\delta = 10^{-2}$) from Example 1 in the last section. We

TABLE 7.1
Example 1a *with* $\delta = 10^{-3}$ *and* $\delta = 10^{-2}$*, respectively.*

r	$\hat{\delta}$	δ_d
4	.00270788	.00265035
• 5	.00269173	.00371632
6	.00262241	.00508314
7	.00261181	.00663777
8	.00249610	.00821011

r	$\hat{\delta}$	δ_d
8	.0249610	.0188627
9	.0236252	.0182476
10	.0231554	.0194804
11	.0226741	.0197377
12	.0226101	.0211548
• 13	.0219453	.0228909
14	.0219144	.0182378

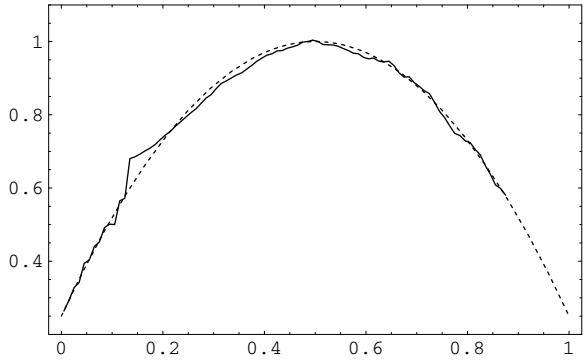


FIG. 7.1. *Example* 1a. *Solution obtained for Example* 1 *using method* 2 *with* $\delta = 10^{-2}$, $N = 100$, $r = 13$. *For comparison, see Figure* 6.6.

TABLE 7.2
Example 2a *with relative noise level* $\delta = 10^{-3}$, $N = 200$, *and discontinuous* $\bar{x}$.

r	$\hat{\delta}$	δ_d
2	0.00133155	0.00113062
• 3	0.0013285	0.00202835
4	0.00131227	0.00308485

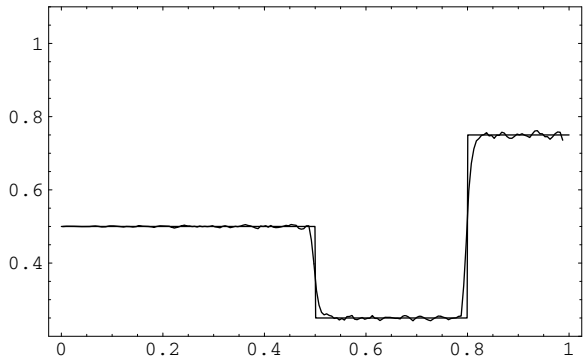


FIG. 7.2. *Example* 2a. *Solution obtained for Example* 2 *using method* 2 *with* $\delta = 10^{-3}$, $N = 200$, $r = 3$. *For comparison, see Figure* 6.9.

see that $r = 5$ is selected for the case of $\delta = 10^{-3}$, while $r = 13$ is selected for the case of $\delta = 10^{-2}$. As expected, the suggested value of r increases as the noise level in the data increases. We show in Figure 7.1 the outcome when using the predicted $r = 13$ at noise level of $\delta = .01$; indeed, in comparing with Figure 6.6 (where $r = 6$ was used), the relative root-mean-square (rms) error of the reconstructed x is 0.0258251 with $r = 13$, which is just about half of the rms error (0.0502236) using $r = 6$, when the same set of noisy data is used. We note also that the predicted $r = 13$ gives $R = 1.3 \times 10^{-1}$, a value which might be seen as more in line with the suggestion made by Theorem 3.1 than the values of R used in Example 1.

7.2. Example 2a: A repeat of Example 2, now using the discrepancy principle to select r. We show in Table 7.2 the results of applying the modified discrepancy principle to Example 2 of the last section; here the value $r = 3$ is selected by the principle, and the associated numerical solution is shown in Figure 7.2. While

Figure 7.2 looks quite similar to Figure 6.9 (where $r = 4$ was used), we can quantitatively conclude that $r = 3$ is a better choice since the relative rms error is 0.072, which is slightly better than the rms error (0.086) in the case of $r = 4$ when the same set of noisy data is used.

Other examples of the use of the modified discrepancy principle for the local regularization of the autoconvolution problem can be found in [8].

8. Conclusion. In this paper we have developed a local regularization theory for the autoconvolution equation, an ill-posed nonlinear Volterra problem. Several local regularization methods have been established, all of which provide stable solutions for the autoconvolution problem. The methods also preserve the causal nature of the autoconvolution equation, allowing for numerically fast sequential solution.

Using the underlying idea of local regularization, we formulated the local regularized equation for the autoconvolution problem and proved the convergence of the solution produced by this regularized equation to the true solution of the autoconvolution equation as the noise level in the data shrinks to zero. The convergence rate of $\mathcal{O}(\delta^{2/5})$ that we obtain for L_2 data is not as good as that $\mathcal{O}(\delta^{1/2})$ obtained by Tikhonov regularization and Lavrent'ev regularization. Only with continuous data are we able to reach the $\mathcal{O}(\delta^{1/2})$ rate of these methods. However, in contrast to these other methods, the theory for local regularization does not rely on an a priori guess for the value of the true solution or of the solution's value at $t = 0$, although a source condition does require a particular relationship between $\bar{x}(0)$ and $\|\bar{x}'\|_\infty$. The theory for both the local regularization and the Lavrent'ev methods requires a positive true solution $\bar{x}$, a condition not needed for Tikhonov regularization. However, as was shown in Examples 3 and 4, local regularization still performs quite well in numerical testing with nonpositive $\bar{x}$ provided a simple inspection of the noisy data has been made in order to determine a minimal suitable value of the local regularization parameter.

We have also presented two discretization algorithms which illustrate the theory developed in this paper; one algorithm involves the solution of nonsequential nonlinear equations on $[0, R]$; the second algorithm is an improvement in both speed and ease of implementation in that the approximate solution can be determined in a (linear) sequential manner for the entire interval $[0, 1]$. The cost for the second algorithm is $\mathcal{O}(rN^2 - r^2N)$, where $r \ll N$, so that it is of the order of the cost of solving the autoconvolution problem without any regularization whatsoever.

Finally, we have shown numerical results which provide evidence that the local regularization methods developed in this work compare favorably to the other numerical methods for the autoconvolution equation in the literature [11, 16], especially in capturing sharp features in the solution. In fact, the numerical results confirm the effectiveness of local regularization methods even in cases not completely falling under the assumptions of the general theory we developed here. Numerical examples also seem to confirm the feasibility of a modified discrepancy principle for the selection of the regularization parameter r. The theoretical analysis of this principle is currently under study.

Acknowledgments. Part of the research was completed while both authors were at the Institute for Pure and Applied Mathematics (IPAM) at UCLA. We thank IPAM for providing the financial support and hospitable environment conducive to productive research. We also thank the referees for helpful comments which improved the exposition of this paper.

REFERENCES

[1] J. Baumeister, *Deconvolution of appearance potential spectra*, in Direct and Inverse Boundary Value Problems, R. Kleinman, R. Kress, adn E. Martensen, eds., Lang, Frankfurt am Main, Germany, 1991, pp. 1–13.

[2] J. V. Beck, B. Blackwell, and C. R. St. Clair, Jr., *Inverse Heat Conduction*, Wiley-Interscience, New York, 1985.

[3] C. D. Brooks, *A Discrepancy Principle for Parameter selection in the Local Regularization of Linear Volterra Inverse Problems*, Ph.D. thesis, Department of Mathematics, Michigan State University, East Lansing, MI, 2007.

[4] C. D. Brooks and P. K. Lamm, *A Discrepancy Principle for Parameter Selection in the Local Regularization of Linear Volterra Inverse Problems*, preprint, 2007.

[5] A. C. Cinzori, *Continuous future polynomial regularization of* 1*-smoothing Volterra problems*, Inverse Problems, 20 (2004), pp. 1791–1806.

[6] A. C. Cinzori and P. K. Lamm, *Future polynomial regularization of ill-posed Volterra equations*, SIAM J. Numer. Anal., 37 (2000), pp. 949–979.

[7] C. Cui, *Local Regularization Methods for n-Dimensional First-Kind Integral Equations*, Ph.D. thesis, Michigan State University, East Lansing, MI, 2005.

[8] Z. Dai, *Local Regularization for the Autoconvolution Problem*, Ph.D. thesis, Department of Mathematics, Michigan State University, East Lansing, MI, 2005.

[9] H. W. Engl, K. Kunisch, and A. Neubauer, *Convergence rates for Tikhonov regularization of nonlinear ill-posed problems*, Inverse Problems, 5 (1989), pp. 524–540.

[10] H. W. Engl, M. Hanke, and A. Neubauer, *Regularization of Inverse Problems*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1996.

[11] G. Fleischer, R. Gorenflo, and B. Hofmann, *On the autoconvolution equation and total variation constraints*, ZAMM Z. Angew. Math. Mech., 79 (1999), pp. 149–159.

[12] G. Fleischer and B. Hofmann, *On inversion rates for the autoconvolution equation*, Inverse Problems, 12 (1996), pp. 419–435.

[13] G. Fleischer and B. Hofmann, *The local degree of ill-posedness and the autoconvolution equation*, Nonlinear Anal., 30 (1997), pp. 3323–3332.

[14] R. Gorenflo and B. Hofmann, *On autoconvolution and regularization*, Inverse Problems, 10 (1994), pp. 353–373.

[15] J. Janno, *On a regularization method for the autoconvolution equation*, Z. Angew. Math. Mech., 77 (1997), pp. 393–394.

[16] J. Janno, *Lavrent'ev regularization of ill-posed problems containing nonlinear near-to-monotone operators with application to autoconvolution equation*, Inverse Problems, 16 (2000), pp. 333–348.

[17] P. K. Lamm, *Approximation of ill-posed Volterra problems via predictor–corrector regularization methods*, SIAM J. Appl. Math., 56 (1996), pp. 524–541.

[18] P. K. Lamm, *Future-sequential regularization methods for ill-posed Volterra equations: Applications to the inverse heat conduction problem*, J. Math. Anal. Appl., 195 (1995), pp. 469–494.

[19] P. K. Lamm, *Regularized inversion of finitely smoothing Volterra operators: Predictor-corrector regularization methods*, Inverse Problems, 13 (1997), pp. 375–402.

[20] P. K. Lamm, *Variable-smoothing regularization methods for inverse problems*, in Theory and Practice of Control and Systems, A. Conte and A. M. Perdon, eds., World Scientific, Singapore, 1999.

[21] P. K. Lamm, *A survey of regularization methods for first-kind Volterra equations*, in Surveys on Solution Methods for Inverse Problems, D. Colton, H. W. Engl, A. Louis, J. R. McLaughlin, and W. Rundell, eds., Springer, Vienna, New York, 2000, pp. 53–82.

[22] P. K. Lamm, *Variable-smoothing local regularization methods for first-kind integral equations*, Inverse Problems, 19 (2003), pp. 195–216.

[23] P. K. Lamm, *Full convergence of sequential local regularization methods for Volterra inverse problems*, Inverse Problems, 21 (2005), pp. 785–803.

[24] P. K. Lamm and Z. Dai, *On local regularization methods for linear Volterra problems and nonlinear equations of Hammerstein type*, Inverse Problems, 21 (2005), pp. 1773–1790.

[25] P. K. Lamm and L. Eldén, *Numerical solution of first-kind Volterra equations by sequential Tikhonov regularization*, SIAM J. Numer. Anal., 34 (1997), pp. 1432–1450.

[26] P. K. Lamm and X. Luo, *Local Regularization Methods for Nonlinear Hammerstein Equations*, preprint, 2006.

[27] P. K. Lamm and T. Scofield, *Sequential predictor-corrector methods for the variable regularization of Volterra inverse problems*, Inverse Problems, 16 (2000), pp. 373–399.

[28] P. K. Lamm and T. Scofield, *Local regularization methods for the stabilization for ill-posed Volterra problems*, Numer. Funct. Anal. Optim., 23 (2001), pp. 913–940.
[29] R. Miller, *Nonlinear Volterra Integral Equations*, W. A. Benjamin, New York, 1971.
[30] R. Ramlau, *Morozov's discrepancy principle for Tikhonov regularization of nonlinear operators*, Numer. Funct. Anal. Optim., 23 (2002), pp. 147–172.
[31] W. Ring and J. Prix, *Sequential predictor-corrector regularization methods and their limitations*, Inverse Problems, 16 (2000), pp. 619–634.
[32] O. Scherzer, *The use of Morozov's discrepancy principle for Tikhkonov regularization for solving nonlinear ill-posed problems*, Computing, 51 (1993), pp. 45–60.
[33] H. Tanabe, *Equations of Evolution*, Pitman, London, 1979.

SIAM J. NUMER. ANAL.
Vol. 46, No. 2, pp. 869–894
© 2008 Society for Industrial and Applied Mathematics

ASYMPTOTIC HIGH-ORDER SCHEMES FOR 2 × 2 DISSIPATIVE HYPERBOLIC SYSTEMS*

DENISE AREGBA-DRIOLLET†, MAYA BRIANI‡, AND ROBERTO NATALINI§

Abstract. We investigate finite difference schemes which approximate 2 × 2 one-dimensional linear dissipative hyperbolic systems. We show that it is possible to introduce some suitable modifications in standard upwinding schemes, which keep into account the long-time behavior of the solutions, to yield numerical approximations which are increasingly accurate for large times when computing small perturbations of stable asymptotic states, respectively, around stationary solutions and in the diffusion (Chapman–Enskog) limit.

Key words. finite difference methods, dissipative hyperbolic problems, asymptotic behavior, asymptotic high-order schemes, Chapman–Enskog expansion

AMS subject classifications. 65M12, 35L65

DOI. 10.1137/060678373

1. Introduction.

Let us consider a one-dimensional 2 × 2 linear system with real coefficients

$$\begin{cases} u_t + \alpha u_x + \beta v_x = 0, \\ v_t + \gamma u_x + \delta v_x = \zeta u - \eta v \end{cases} \tag{1}$$

for $x \in \mathbb{R}$ and $t > 0$, with the initial conditions

$$u(x,0) = u_0(x),\ v(x,0) = v_0(x) \quad \text{for } x \in \mathbb{R}. \tag{2}$$

We assume that the system is strictly hyperbolic: The eigenvalues of the matrix $A = \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix}$,

$$\lambda_\pm = \frac{1}{2}\left[(\alpha+\delta) \pm \sqrt{(\alpha-\delta)^2 + 4\beta\gamma}\right],$$

are real and distinct, which is simply equivalent to the condition

$$(\alpha-\delta)^2 + 4\beta\gamma > 0. \tag{3}$$

Clearly, according to the general theory of linear hyperbolic problems (see, for instance, [16]) for every initial datum in some suitable functional class, there exists a

*Received by the editors December 22, 2006; accepted for publication (in revised form) July 25, 2007; published electronically February 20, 2008. The research activity reported in this paper was partially conducted with the support of the project "Problemi iperbolici non lineari" for the year 2006 of INDAM-GNAMPA.

http://www.siam.org/journals/sinum/46-2/67837.html

†IMB, Mathématiques Appliquées de Bordeaux, Université Bordeaux 1, 351 cours de la Libération, F-33405 Talence, France (Denise.Aregba@math.u-bordeaux1.fr).

‡Istituto per le Applicazioni del Calcolo "M. Picone," Consiglio Nazionale delle Ricerche, Viale del Policlinico 137, I-00161, Rome, Italy, and Università LUISS - Guido Carli, Rome, Italy (m.briani@iac.cnr.it).

§Istituto per le Applicazioni del Calcolo "M. Picone," Consiglio Nazionale delle Ricerche, Viale del Policlinico 137, I-00161, Rome, Italy (r.natalini@iac.cnr.it).

corresponding global solution in the same class. However, the stability of this solution, i.e., its global boundedness in time, is more involved, and mostly connected to the behavior of the equilibrium problem, when the source term vanishes, namely, the problem restricted on the manifold $v = \frac{\zeta}{\eta}u$, with u solving

$$u_t + \alpha^* u_x = 0 \tag{4}$$

for $\alpha^* = \alpha + \frac{\zeta}{\eta}\beta$. It is well known that a necessary and sufficient stability condition for problem (1) is given by the so-called subcharacteristic condition; see [17, 12]: The coefficient η is positive, and the speed of the equilibrium problem has to be bounded by the speeds of problem (1), i.e.,

$$\eta > 0, \qquad \lambda_- \le \alpha^* \le \lambda_+. \tag{5}$$

This simple example is a useful prototype for more general nonlinear relaxation systems which arise in various contexts; see again [17, 12] for examples and references. Recently, in [9, 2], a quite complete theory of global existence and of the asymptotic behavior of smooth solutions for this type of systems was developed, actually in a nonlinear and fully multidimensional framework. This theory, which in the case of our special example will be briefly summarized in section 2, needs for an extra assumption the so-called Shizuta–Kawashima condition [15, 9], which guarantees a sufficient coupling between the source and the advection terms. In the present example this condition is just equivalent to imposing the strict inequalities in (5). Roughly speaking, under these assumptions, it is possible to prove that, for every perturbation of a given stationary solution to problem (1), the corresponding solution decays in the L^p-norm to its unperturbed state as $\mathcal{O}(t^{-\frac{1}{2}(1-\frac{1}{p})})$ for $p \in [1, \infty]$. Moreover, when the asymptotic state is given by a constant state, it is possible to compare this perturbed solution with the solution $\hat{u}$ of the Chapman–Enskog expansion, which in the present case is given by

$$\hat{u}_t + \alpha^* \hat{u}_x = \nu \hat{u}_{xx}, \tag{6}$$

with $\nu = \frac{\beta}{\eta^2}[\gamma\eta - \frac{\beta}{\eta}\zeta^2 - (\alpha - \delta)\zeta]$. Notice that the strict subcharacteristic inequalities are equivalent to the condition $\nu > 0$. So (6) enjoys well-posedness of the Cauchy problem, and, again under the previous assumptions, it is actually possible to prove that the difference between u and the diffusive profile $\hat{u}$ decays in the L^p-norm as $\mathcal{O}(t^{-\frac{1}{2}(2-\frac{1}{p})})$ for $p \in [1, \infty]$, so a factor $\frac{1}{2}$ better than the simple decay to the steady state. Clearly, if the Shizuta–Kawashima condition is violated, no decay is in general expected, as follows easily, for instance, in the case $\beta = 0$.

In this paper, numerical approximations related to these asymptotic results will be introduced and discussed. Following the ideas in [5], where the case of a scalar hyperbolic equation was considered, we are looking for *asymptotic high-order* (AHO) schemes, i.e., schemes which are high-order accurate, with respect to the local truncation error, when restricted to every element of a given family of stable asymptotic states. Here we show that, for 2×2 dissipative hyperbolic systems, it is possible to introduce AHO schemes which are compatible with the behavior predicted by the qualitative analysis, respectively, for the long-time asymptotic and in the Chapman–Enskog regimes. The main idea we are going to apply is to modify standard upwinding schemes to keep into account the long-time behavior of the solutions. In section 3, we present some results of convergence and error estimates for general upwinding finite difference schemes for systems of the form (1), under the strict subcharacteristic

condition. Even if these results can be proven for more general weakly coupled systems (see, for instance, [12]), we prefer to present a precise statement to make the paper self-contained. Section 4 contains the description of our new schemes, which are AHO, respectively, around the perturbation of general steady states and in the diffusion limit. CFL estimates are presented to guarantee the monotonicity of these schemes. Some numerical tests are presented and discussed in section 5 to show the better performance of our schemes with respect to the usual pointwise approximation of the source term and even with the classical upwinding of the source proposed by Roe in 1986 [14]. In particular, we show by numerical tests that the L^∞ global error of our main diffusive adapted AHO2p-scheme decays as $\mathcal{O}(1/t)$, in agreement with the decay given by inequality (28), for a given fixed space step, against the decay as $\mathcal{O}(1/\sqrt{t})$ of the other schemes.

Let us notice that our schemes can be in principle extended to most of the dissipative nonlinear hyperbolic systems arising in applications (see examples reported in [9, 2]), even if the computation of the form of the schemes could be not trivial. Notice that the AHO modifications occur mainly in the pointwise discretization of the source term and do not require the explicit knowledge of the asymptotic solution. However, stability analysis should be much more involved in the general case, and so this problem will be addressed in a future paper.

Finally, let us remark that our approach is somewhat related to the ideas motivating the so-called well-balanced schemes; see, for instance, [1, 7, 10, 3, 13], and the book [4], which, however, are more relevant on different time scales and contexts, as, for instance, for inhomogeneous source terms.

2. Analytical backgrounds. Let us consider system (1), which first will be rewritten in a vector form, by setting $U = (u, v)$ and

$$A = \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix}, \qquad B = \begin{pmatrix} 0 & 0 \\ \zeta & -\eta \end{pmatrix}.$$

System (1) reads now

$$U_t + AU_x = BU. \tag{7}$$

The first-order system (7) is said to be in conservative-dissipative form (CD form) if the matrix A is symmetric and the coefficient ζ, in the matrix B, vanishes. Our first goal is to reduce every 2×2 system, which verifies some suitable assumptions, to the CD form, to enter the framework in [2].

PROPOSITION 2.1. *Let $\eta > 0$, and assume the strict subcharacteristic condition*

$$\lambda_- < \alpha^* < \lambda_+. \tag{8}$$

Then there exists a nonsingular linear change of variable such that system (1) *can be set in the CD form.*

Proof. First let us remark that, by an elementary computation, condition (8) is equivalent to the single inequality

$$\beta \left[\gamma\eta - \frac{\beta}{\eta}\zeta^2 - (\alpha - \delta)\zeta \right] > 0. \tag{9}$$

Let us introduce the following change of variable:

$$\begin{pmatrix} \tilde{u} \\ \tilde{v} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ q & r \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix},$$

with $q = -\frac{\zeta}{\eta} r$ and r given by

$$r = \left[\beta\eta \left(\gamma\eta - \frac{\beta}{\eta}\zeta^2 - (\alpha - \delta)\zeta \right)^{-1} \right]^{1/2}.$$

By condition (9), this quantity is well-defined, real, and positive. The new unknowns $(\tilde{u}, \tilde{v})$ solve the system

$$\begin{cases} \tilde{u}_t + (\alpha + \frac{\zeta}{\eta}\beta)\tilde{u}_x + \frac{\beta}{r}\tilde{v}_x = 0, \\ \\ \tilde{v}_t + \frac{\beta}{r}\tilde{u}_x + (\delta - \frac{\zeta}{\eta}\beta)\tilde{v}_x = -\eta\tilde{v}, \end{cases} \tag{10}$$

which is in CD form. □

Therefore, according to the previous result, from now on and without loss of generality, we shall restrict our attention to the following model problem:

$$\begin{cases} u_t + au_x + bv_x = 0, \\ \\ v_t + bu_x + cv_x = -dv, \end{cases} \tag{11}$$

where the constants a, b, c, and d are real, $b \neq 0$, and $d > 0$.

We want to obtain some estimates for the L^1-, L^∞- norms and total variation, so we need to diagonalize the system. First let us rewrite system (11) in the vector form (7), with

$$A = \begin{pmatrix} a & b \\ b & c \end{pmatrix}, \qquad B = -\begin{pmatrix} 0 & 0 \\ 0 & d \end{pmatrix}.$$

Now we can diagonalize the matrix A:

$$A = R\Lambda R^{-1},$$

where $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2)$ and $R = (r^{(1)}, r^{(2)})$ is the column matrix of right eigenvectors, i.e., $Ar^{(i)} = \lambda_i r^{(i)}$, $i = 1, 2$. Introducing the notation

$$\mathrm{w} = R^{-1}U, \tag{12}$$

problem (11) becomes

$$\begin{cases} \mathrm{w}_t + \Lambda \mathrm{w}_x = \tilde{B}\mathrm{w}, \\ \\ \mathrm{w}(x, 0) = \mathrm{w}^0(x) = R^{-1}\mathrm{u}^0(x), \end{cases} \tag{13}$$

where $\tilde{B} = R^{-1}BR$.

If we choose the matrix of the diagonalization of the system (7) as

$$R = \begin{pmatrix} 1 & 1 \\ -\dfrac{(a - \lambda_1)}{b} & \dfrac{(\lambda_2 - a)}{b} \end{pmatrix}, \tag{14}$$

the corresponding source term is then given by

$$\tilde{B} = \frac{d}{\lambda_2 - \lambda_1} \begin{pmatrix} -(a - \lambda_1) & (\lambda_2 - a) \\ (a - \lambda_1) & -(\lambda_2 - a) \end{pmatrix}. \tag{15}$$

Now we can prove the main result about monotonicity and L^1-contraction properties of system (11). Let us set

$$L = L(\lambda) = \max_{i=1,2} |\lambda_i|.$$

For any given vectors $C, D \in \mathbb{R}^2$, set $C \leq D$ if $c_i \leq d_i$ $(i = 1, 2)$; if $C \leq D$, let us denote the corresponding interval by $[C, D]$. Also, for each interval $I = (x_1, x_2) \subseteq \mathbb{R}$, we set $\tau = (x_2 - x_1)/2L$.

PROPOSITION 2.2. *Let* w *and* $\tilde{\mathrm{w}}$ *be two weak solutions of the Cauchy problem* (13) *in* $\mathbb{R} \times (0, T)$ $(T > 0)$ *for the initial data* w^0 *and* $\tilde{\mathrm{w}}^0$*, respectively. Let* $[C, D] \subseteq \mathbb{R}^2$ *be an interval with nonempty interior such that* w, $\tilde{\mathrm{w}} \in [C, D]$*. If the matrix* B *is given by* (15)*, then for each interval* $(x_1, x_2) \subseteq \mathbb{R}$ *and for all* $t \in (0, \min(\tau, T))$ *we have*

$$\sum_{i=1,2} \int_{x_1+tL}^{x_2-tL} [w_i - \tilde{w}_i]_+ dx \leq \sum_{i=1,2} \int_{x_1}^{x_2} [w_i^0 - \tilde{w}_i^0]_+ dx. \tag{16}$$

Proof. Define $\mathrm{v} = \mathrm{w} - \tilde{\mathrm{w}}$. We multiply the i component of the system (13) by the Heaviside function $H(v_i)$, where $v_i = w_i - \tilde{w}_i$, $i \in \{1, 2\}$, and we add both equations. We get

$$\sum_{i=1}^{2} \partial_t [v_i]_+ + \sum_{i=1}^{2} \lambda_i \partial_x [v_i]_+ = \sum_{i,j \in \{1,2\}} \tilde{b}_{ij} H(v_i) v_j, \tag{17}$$

where we used the relations $H(\mathrm{v})\mathrm{v} = [\mathrm{v}]_+$ and $H(\mathrm{v})\partial_\cdot \mathrm{v} = \partial_\cdot [\mathrm{v}]_+$. We have

$$\sum_{i,j \in \{1,2\}} \tilde{b}_{ij} H(v_i) v_j \leq \sum_{i \in \{1,2\}} \tilde{b}_{ii} [v_i]_+ + \sum_{i \in \{1,2\}} \sum_{j \neq i} \tilde{b}_{ij} [v_j]_+ \tag{18}$$

$$= \sum_{i \in \{1,2\}} \left(\tilde{b}_{ii} + \sum_{j \neq i} \tilde{b}_{ji} \right) [v_i]_+. \tag{19}$$

By considering (15), we can therefore conclude that the right-hand side of (17) is nonpositive. Then we argue as in [8] and obtain the desired inequality. ∎

Clearly, from Proposition 2.2 we can easily deduce, by standard arguments, L^1- and L^∞-estimates, as well as obvious comparison properties of solutions, and, as the system (11) is linear, the same estimates hold for the derivatives of solutions. More precisely we use the following result in what follows.

COROLLARY 2.3. *Let* $\mathrm{w}^0 \in C^m(\mathbb{R})$ *such that*

$$\frac{d^k \mathrm{w}^0}{dx^k} \in L^1(\mathbb{R}) \cap L^\infty(\mathbb{R}) \quad \textit{for } 0 \leq k \leq m. \tag{20}$$

Then the Cauchy problem (13) *has a unique solution* $\mathrm{w} \in C^0([0, +\infty[, L^1(\mathbb{R}))$*, and there exists a constant* $K > 0$*, not depending on* w^0*, such that for* $0 \leq k \leq m$*,* $p = 1$*, and* $p = \infty$*,*

$$\|\partial_x^k \mathrm{w}(., t)\|_p \leq K \left\| \frac{d^k \mathrm{w}^0}{dx^k} \right\|_p, \quad \|\partial_t \mathrm{w}(., t)\|_p \leq K \left(\left\| \frac{d\mathrm{w}^0}{dx} \right\|_p + \|\mathrm{w}^0\|_p \right). \tag{21}$$

Let us also notice that for system (11) it is possible to prove an energy inequality, as in [9], but in the present framework this estimate is not immediately relevant.

Now we are in position to investigate the asymptotic behavior of solutions to problem (11). Here we refer to the results in [9, 2] about the Cauchy problem for a general $n \times n$ hyperbolic symmetrizable one-dimensional system of balance laws

$$U_t + f(U)_x = g(U), \tag{22}$$

with the initial condition

$$U(x,0) = U_0(x), \tag{23}$$

where $U = (u_1, u_2) \in \Omega \subseteq \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$, with $n_1 + n_2 = n$. It is also assumed that there are n_1 conservation laws in the system, namely, that we can take

$$g(U) = \begin{pmatrix} 0 \\ q(U) \end{pmatrix}, \quad \text{with } q(U) \in \mathbb{R}^{n_2}. \tag{24}$$

In [9], a general framework of sufficient conditions was proposed which guarantee the global existence in time of smooth solutions, by showing that dissipation, at least for small data, can prevent shock formation. Actually, for the systems which are endowed with a strictly convex entropy function $\mathcal{E} = \mathcal{E}(U)$, a first natural assumption is the *entropy dissipation condition* (see [6]); namely, for every U, $\overline{U} \in \Omega$, with $g(\overline{U}) = 0$,

$$\big(\mathcal{E}'(U) - \mathcal{E}'(\overline{U})\big) \cdot g(U) \le 0.$$

Unfortunately, it is easy to see that this condition is too weak to prevent the formation of singularities. A quite natural supplementary condition can be imposed to entropy dissipative systems, following the approach first proposed by Kawashima [11] and Shizuta and Kawashima [15], which for system (22) reads

$$\text{Ker } Dg(\overline{U}) \cap \{\text{eigenspaces of } Df(\overline{U})\} = \{0\} \tag{25}$$

for every $\overline{U} \in \Omega$, with $g(\overline{U}) = 0$. It is possible to prove that this condition, which is satisfied in many interesting examples, is also sufficient to establish a result of global existence for small perturbations of equilibrium constant states; see [9]. Clearly also system (11) fits this framework. In fact the system is entropy dissipative for the entropy function $\mathcal{E}(u,v) = \frac{1}{2}(u^2 + v^2)$ if and only if $d > 0$, and condition (25) is satisfied if $b \neq 0$. Therefore we can apply all of the results proved in [2] about the asymptotic behavior of dissipative hyperbolic systems to system (11). Here we just state two main results of time decay. More details and proofs are contained in [2]. For $s > 0$, set $E_s = \max\{\|U(0)\|_{L^1}, \|U(0)\|_{H^s}\}$.

THEOREM 2.4. *Let $(u(t), v(t))$ be a smooth global solution to problem (11)–(2), with $b \neq 0$ and $d > 0$. Let $p \in [1, \infty]$. The following decay estimate holds for some positive constant $C > 0$:*

$$\|D^\beta(u(t), v(t))\|_{L^p} \le C \min\Big\{1, t^{-\frac{1}{2}(1-\frac{1}{p})-|\beta|/2}\Big\} E_{|\beta|+2}. \tag{26}$$

For the dissipative part $v(t)$, we have a more precise estimate for some positive constant $C > 0$:

$$\|D^\beta v(t)\|_{L^p} \le C \min\Big\{1, t^{-\frac{1}{2}(1-\frac{1}{p})-1/2-|\beta|/2}\Big\} E_{|\beta|+2}. \tag{27}$$

Another interesting result concerns the convergence to the Chapman–Enskog expansion of problem (11), which is given by the parabolic equation (6).

THEOREM 2.5. *Let $\hat{u}$ be the solution of problem* (6)*, with $\alpha^* = a$ and $\nu = \frac{b^2}{d}$, and the initial condition given by* (2)*. Under the assumptions of Theorem* 2.4*, we have the following more accurate decay estimate for some positive constant $C > 0$:*

$$\|D^\beta(u(t) - \hat{u}(t))\|_{L^p} \le C \min\left\{1, t^{-\frac{1}{2}(1-\frac{1}{p})-|\beta|/2-1/2}\right\} E_{|\beta|+2}. \tag{28}$$

3. Numerical approximation. Here we introduce and discuss general finite difference approximations for system (11), and we prove some convergence properties and error estimates.

3.1. Construction of the schemes. We approximate the differential part following the direction of the characteristic velocities, so first we study the methods for the system in diagonal form (13) obtained by the choice (14). We denote by $\mathrm{w} = (w^1, w^2)$ the exact solution.

We denote by h the uniform mesh length and by $x_l = l\,h$ the spatial grid points for all $l \in \mathbb{Z}$. The time levels t_n, with $t_0 = 0$, are also spaced uniformly with mesh length $\Delta t = t_{n+1} - t_n$ for $n \in \mathbb{N}$. We denote by δ the CFL ratio $\delta = \Delta t/h$, which is taken constant through the paper.

We consider the Cauchy problem (13). The initial data w^0 is supposed to be smooth and is approximated by its node values. The approximate solution $W_l^n = (w_{l,n}^1, w_{l,n}^2)^T$ is given by

$$\begin{aligned} &\frac{W_l^{n+1} - W_l^n}{\Delta t} + \frac{\Lambda}{2h}\left(W_{l+1}^n - W_{l-1}^n\right) - \frac{\tilde{Q}}{2h}\left(W_{l+1}^n - 2W_l^n + W_{l-1}^n\right) \\ &= \tilde{\mathcal{B}}_{-1} W_{l-1}^n + \tilde{\mathcal{B}}_0 W_l^n + \tilde{\mathcal{B}}_1 W_{l+1}^n, \quad l \in \mathbb{Z},\ n \in \mathbb{N}, \\ &W_l^0 = \mathrm{w}^0(x_l), \quad l \in \mathbb{Z}, \end{aligned} \tag{29}$$

where $\tilde{Q} = \mathrm{diag}(\tilde{q}_1, \tilde{q}_2)$ is the diagonal matrix of the artificial diffusion terms $\tilde{q}_i \ge 0$ ($i \in \{1, 2\}$), and $\tilde{\mathcal{B}}_{-1} = (\tilde{\beta}_{ij}^{-1})_{i,j=1,2}$, $\tilde{\mathcal{B}}_0 = (\tilde{\beta}_{ij}^0)_{i,j=1,2}$, and $\tilde{\mathcal{B}}_1 = (\tilde{\beta}_{ij}^1)_{i,j=1,2}$ are 2×2 constant matrices that define the source approximation. Those matrices may depend on h.

We set

$$W_\Delta^n(x) = W_l^n \quad \text{if } x \in [x_l, x_{l+1}[$$

and

$$W_\Delta(x,t) = W_\Delta^n(x) \quad \text{if } t \in [t_n, t_{n+1}[.$$

The scheme (29) can be seen as a linear function

$$W_\Delta^{n+1} = \mathbf{S}_{\Delta t}(W_\Delta^n). \tag{30}$$

More precisely we have

$$w_{l,n+1}^i = S^i(W_{l-1}^n, W_l^n, W_{l+1}^n), \quad i \in \{1, 2\}, \quad l \in \mathbb{Z}, \tag{31}$$

with

$$\begin{aligned} S^i(W_{l-1}^n, W_l^n, W_{l+1}^n) &= \alpha_{-1}^i w_{l-1,n}^i + \alpha_0^i w_{l,n}^i + \alpha_1^i w_{l+1,n}^i \\ &+ \Delta t \sum_{j \in \{1,2\}} \left(\tilde{\beta}_{ij}^{-1} w_{l-1,n}^j + \tilde{\beta}_{ij}^0 w_{l,n}^j + \tilde{\beta}_{ij}^1 w_{l+1,n}^j\right), \end{aligned} \tag{32}$$

where, for all $i \in \{1,2\}$,

$$\alpha_{-1}^i = \frac{\delta}{2}\lambda_i + \frac{\delta}{2}\tilde{q}_i, \quad \alpha_0^i = 1 - \delta\tilde{q}_i, \quad \alpha_1^i = -\frac{\delta}{2}\lambda_i + \frac{\delta}{2}\tilde{q}_i \,. \tag{33}$$

For any function ψ defined on $\mathbb{R}$ with values in $\mathbb{R}^2$, we set

$$\overline{\psi}(x) = (\psi(x-h), \psi(x), \psi(x+h)). \tag{34}$$

The function $\mathbf{S}_{\Delta t}$ is then defined by

$$\mathbf{S}_{\Delta t}(\psi)(x) = \left(S^1(\overline{\psi}(x)), S^2(\overline{\psi}(x))\right). \tag{35}$$

3.2. Convergence for smooth data. We assume that the scheme satisfies the two following properties:

1. *Consistency.* The scheme (29) is consistent with problem (13), i.e.,

$$\tilde{\mathcal{B}}_{-1} + \tilde{\mathcal{B}}_0 + \tilde{\mathcal{B}}_1 = \tilde{B} + h\tilde{C}, \tag{36}$$

   where $\tilde{C} = (\tilde{c}_{ij})_{i,j=1,2}$ is a 2×2 constant matrix not depending on h and Δt.
2. *Monotonicity.* For all $i \in \{1,2\}$, the operators S^i, defined in (31), are monotone nondecreasing in all of their components, i.e.,

$$\tilde{\beta}_{ij}^{-1}, \tilde{\beta}_{ij}^0, \tilde{\beta}_{ij}^1 \geq 0 \quad \forall\, i \neq j, \tag{37}$$

$$A_{i,0} = 1 - \delta\tilde{q}_i + \Delta t\tilde{\beta}_{ii}^0 \geq 0 \quad \forall\ i \in \{1,2\}, \tag{38}$$

$$\begin{cases} A_{i,-1} = \dfrac{\delta\lambda_i}{2} + \dfrac{\delta\tilde{q}_i}{2} + \Delta t\tilde{\beta}_{ii}^{-1} \geq 0, \\ \\ A_{i,1} = -\dfrac{\delta\lambda_i}{2} + \dfrac{\delta\tilde{q}_i}{2} + \Delta t\tilde{\beta}_{ii}^1 \geq 0 \end{cases} \quad \forall\ i \in \{1,2\}. \tag{39}$$

For $t \geq 0$ and $x \in \mathbb{R}$, by using the notation (34) for w, the local truncation error $T(x,t)$ is defined by

$$T(x,t) = \frac{\mathrm{w}(x, t+\Delta t) - \mathbf{S}_{\Delta t}\mathrm{w}(.,t)(x)}{\Delta t},$$

which can also be written

$$T^i(x,t) = \frac{w^i(x,t+\Delta t) - S^i(\overline{\mathrm{w}}(x,t))}{\Delta t}, \quad i \in \{1,2\}.$$

It is obtained straightforwardly by Taylor expansions:

$$\begin{aligned} T^i(x,t) \;=\; & \frac{\Delta t}{2}\partial_{tt}w^i(x,t+\theta_i\Delta t) - h\frac{\tilde{q}_i}{2}\partial_{xx}w^i(x,t) - h\sum_{j\in\{1,2\}}\tilde{c}_{ij}w^j(x,t) \\ & -h\sum_{j\in\{1,2\}}(\tilde{\beta}_{ij}^1 - \tilde{\beta}_{ij}^{-1})\partial_x w^j(x,t) + \mathcal{O}(h^2). \end{aligned} \tag{40}$$

In this formula, $\theta_i \in [0,1]$, and the last term depends on the spatial derivatives of w, up to third order.

In view of proving convergence, we need the discrete truncation error, which is defined as

$$T_\Delta(x,t) = T(x_l,t) \quad \text{for} \quad x \in [x_l, x_{l+1}[.$$

The scheme is first-order in the following sense.

PROPOSITION 3.1 (accuracy for smooth data). *Consider a smooth initial datum* w^0 *such that*

$$\frac{d^k \mathrm{w}^0}{dx^k} \in L^1(\mathbb{R}) \cap L^\infty(\mathbb{R}) \quad for \;\; 0 \le k \le 4. \tag{41}$$

Then the discrete local truncation error $T_\Delta(x,t)$ *for the scheme* (29) *satisfies*

$$\sum_{i\in\{1,2\}} \left[\|T^i_\Delta(.,t)\|_1 + \|T^i_\Delta(.,t)\|_\infty \right] \le C(\Delta t + h). \tag{42}$$

Here

$$C \le C_1 \sup_{0\le k\le 4} \left(\left\| \frac{d^k \mathrm{w}^0}{dx^k} \right\|_1 + \left\| \frac{d^k \mathrm{w}^0}{dx^k} \right\|_\infty \right), \tag{43}$$

and C_1 *is a uniform constant.*

Proof. We can express $\partial_{tt}\mathrm{w}$ as a linear function of w, $\partial_x \mathrm{w}$, $\partial_{xx}\mathrm{w}$. As a consequence, there exists a constant C_0,

$$C_0 \le C_0' \sup_{k\le 3} \sup_{t\in[0,+\infty[} (\|\partial_x^k \mathrm{w}(.,t)\|_1 + \|\partial_x^k \mathrm{w}(.,t)\|_\infty),$$

such that

$$\sum_{i\in\{1,2\}} \left[\|T^i(.,t)\|_1 + \|T^i(.,t)\|_\infty \right] \le C_0(\Delta t + h). \tag{44}$$

Moreover, it is easy to see that for $p = 1$ or $p = \infty$

$$\|T^i(.,t) - T^i_\Delta(.,t)\|_p \le h \|\partial_x T^i(.,t)\|_p, \quad i \in \{1,2\},$$

which, with Corollary 2.3, gives the estimate (42). ∎

We prove now that the scheme is stable.

PROPOSITION 3.2 (L^1 stability). *For the matrix* $\tilde{C}$ *defined in* (36), *we introduce the quantity*

$$\tilde{\gamma} = \left[\sup_{i\in\{1,2\}} \left\{ \tilde{c}_{ii} + \sum_{j\neq i} \tilde{c}_{ji} \right\} \right]_+ . \tag{45}$$

Under the monotonicity assumptions (37)–(39), *the scheme* (29) *is* L^1*-stable: For all* $W^0_\Delta \in L^1(\mathbb{R})$ *it holds that*

$$\|W^n_\Delta\|_1 \le e^{\tilde{\gamma} h t_n} \|W^0_\Delta\|_1 \quad \forall n \ge 1. \tag{46}$$

Proof. The norm of W_Δ^{n+1} can be written as

$$\|W_\Delta^{n+1}\|_1 = h \sum_{i=1,2} \sum_{l\in\mathbb{Z}} \left| w_{l,n}^i - \delta \frac{\lambda_i}{2}(w_{l+1,n}^i - w_{l-1,n}^i) + \delta \frac{\tilde{q}_i}{2}(w_{l+1,n}^i - 2w_{l,n}^i + w_{l-1,n}^i) \right.$$

$$\left. + \Delta t \sum_{j=1,2} \left(\tilde{\beta}_{ij}^{-1} w_{l-1,n}^j + \tilde{\beta}_{ij}^0 w_{l,n}^j + \tilde{\beta}_{ij}^1 w_{l+1,n}^j \right) \right| \tag{47}$$

$$\leq h \sum_{i=1,2} \sum_{l\in\mathbb{Z}} |w_{l,n}^i| A_i.$$

Here A_i is defined by

$$A_i = |A_{i,0}| + |A_{i,-1}| + |A_{i,1}| + \Delta t \sum_{j\neq i} (|\tilde{\beta}_{ji}^{-1}| + |\tilde{\beta}_{ji}^0| + |\tilde{\beta}_{ji}^1|),$$

with quantities $A_{i,k}$ defined in (37)–(39). By those assumptions and by the definition of $\tilde{C}$ we obtain that

$$A_i \leq 1 + \tilde{\gamma} h \, \Delta t.$$

Hence

$$\|W_\Delta^{n+1}\|_1 \leq (1 + \tilde{\gamma} h \, \Delta t) \|W_\Delta^n\|_1,$$

and this gives (46). ∎

We have also a uniform bound for the scheme.

PROPOSITION 3.3 (L^∞ stability). *Under the monotonicity assumptions* (37)–(39)*, the scheme* (29) *is L^∞-stable: For all $W_\Delta^0 \in L^\infty(\mathbb{R})$ it holds that*

$$\|W_\Delta^n\|_\infty \leq (1 + \tilde{\gamma} h \, \Delta t) \|W_\Delta^{n-1}\|_\infty \leq e^{\tilde{\gamma} h t_n} \|W_\Delta^0\|_\infty \quad \forall n \geq 1. \tag{48}$$

The proof is similar to the one of Proposition (3.2), so we omit it.

We have proved consistency and stability, so we have convergence.

THEOREM 3.4. *Let* w^0 *be a smooth initial datum satisfying* (41)*. We suppose that the monotonicity assumptions* (37)–(39) *are verified and that $\delta = \Delta t/h$ is fixed.*

For all $T > 0$ the scheme (29) *converges in $L^\infty([0,T], L^1(\mathbb{R}) \cap L^\infty(\mathbb{R}))$ towards the solution of the Cauchy problem* (13)*: There exists $h_0(T) > 0$ such that for all $h < h_0(T)$*

$$\sup_{t\in[0,T]} (\|W_\Delta(.,t) - \mathrm{w}(.,t)\|_1 + \|W_\Delta(.,t) - \mathrm{w}(.,t)\|_\infty) \leq (2T+1)\, C\, (\Delta t + h), \tag{49}$$

where C is defined in (43)*.*

Proof. Let us fix $t \in [0,T]$. There exists an integer n such that $t_n \leq t < t_{n+1}$. We set

$$\mathrm{w}_\Delta(x,t) = \mathrm{w}(x_l, t) \quad \text{for} \quad x \in [x_l, x_{l+1}[.$$

For $p = 1$ and $p = \infty$, we have

$$\begin{aligned} \|W_\Delta(.,t) - \mathrm{w}(.,t)\|_p \leq{}& \|W_\Delta^n - \mathrm{w}_\Delta(.,t_n)\|_p \\ &+ \|\mathrm{w}_\Delta(.,t_n) - \mathrm{w}(.,t_n)\|_p \\ &+ \|\mathrm{w}(.,t_n) - \mathrm{w}(.,t)\|_p \,. \end{aligned} \tag{50}$$

The two last terms of this inequality are easily estimated:

$$\|\mathrm{w}_\Delta(.,t_n) - \mathrm{w}(.,t_n)\|_p + \|\mathrm{w}(.,t_n) - \mathrm{w}(.,t)\|_p \le C(h+\Delta t).$$

In order to study the other term, we introduce the following notation:

$$V_\Delta^m = \mathrm{w}_\Delta(.,t_m), \quad m = 0,\dots,n.$$

We have

$$\begin{aligned}\|W_\Delta^n - \mathrm{w}_\Delta(.,t_n)\|_p \le{}& \|\mathbf{S}_{\Delta t}^{n-1}\mathbf{S}_{\Delta t}(V_\Delta^0) - \mathbf{S}_{\Delta t}^{n-1}(V_\Delta^1)\|_p \\ &+ \|\mathbf{S}_{\Delta t}^{n-2}\mathbf{S}_{\Delta t}(V_\Delta^1) - \mathbf{S}_{\Delta t}^{n-2}(V_\Delta^2)\|_p \\ &+ \cdots \\ &+ \|\mathbf{S}_{\Delta t}(V_\Delta^{n-1}) - V_\Delta^n\|_p\ .\end{aligned}$$

By L^p stability,

$$\begin{aligned}\|W_\Delta^n - \mathrm{w}_\Delta(.,t_n)\|_p \le{}& \|\mathbf{S}_{\Delta t}(V_\Delta^0) - V_\Delta^1\|_p e^{\tilde\gamma h t_{n-1}} \\ &+ \|\mathbf{S}_{\Delta t}(V_\Delta^1) - V_\Delta^2\|_p e^{\tilde\gamma h t_{n-2}} \\ &+ \cdots \\ &+ \|\mathbf{S}_{\Delta t}(V_\Delta^{n-1}) - V_\Delta^n\|_p\ .\end{aligned} \tag{51}$$

We then sum over $p \in \{1,\infty\}$. We may apply the consistency result (42) to each term in the right-hand side:

$$V_\Delta^{j+1}(x) - \mathbf{S}_{\Delta t}(V_\Delta^j)(x) = \Delta t T_\Delta(x,t_j), \quad j = 0,\dots,n-1.$$

Summing up all of the terms gives

$$\|W_\Delta^n - \mathrm{w}_\Delta(.,t_n)\|_1 + \|W_\Delta^n - \mathrm{w}(.,t_n)\|_\infty \le \Delta t C(\Delta t + h)\frac{e^{\tilde\gamma h T}-1}{e^{\tilde\gamma h \Delta t}-1}\,.$$

As δ is constant we have

$$\lim_{h\to 0} \Delta t \frac{e^{\tilde\gamma h T}-1}{e^{\tilde\gamma h \Delta t}-1} = T,$$

which ends the proof. ∎

Notice that, if $\tilde\gamma = 0$, the scheme is L^1 contracting, and we have for all h and Δt

$$\sup_{t\in[0,T]} \left(\|W_\Delta(.,t) - \mathrm{w}(.,t)\|_1 + \|W_\Delta(.,t) - \mathrm{w}(.,t)\|_\infty\right) \le (T+1)\ C\ (\Delta t + h).$$

4. Asymptotic high-order schemes. In this section, we study the discretization of the source term, defined by coefficients $\tilde{\mathcal{B}}_{-1,0,1}$ introduced in (13), to present some schemes which are increasingly accurate for large times, with respect to the asymptotic behavior of solutions. This property of accuracy is required in order to get better results for large time simulations when computing perturbations of nonconstant stable states. Actually, given a stable solution z (or a family of stable solutions), we say that a scheme is *asymptotic high-order* of order q (AHOq) with respect to this stable state if the scheme is accurate of order q when restricted to this solution; see [5].

4.1. The construction of an AHO4 scheme for stationary solutions. Here we consider the case of a perturbation of a generic stationary solution. Let us first develop the formula (40) for a smooth solution w:

$$
\begin{aligned}
T(x,t) &= \frac{\Lambda}{2h}\left(\frac{h^3}{3}\partial_x^3 \mathrm{w}(x,t) + \mathcal{O}(h^5)\right)\\
&-\frac{\tilde{Q}}{2h}\left(h^2\partial_x^2 \mathrm{w}(x,t) + \frac{h^4}{12}\partial_x^4 \mathrm{w}(x,t) + \mathcal{O}(h^6)\right)\\
&-h\tilde{C}\mathrm{w}(x,t) - h\left(\tilde{\mathcal{B}}_1 - \tilde{\mathcal{B}}_{-1}\right)\partial_x \mathrm{w}(x,t) - \frac{h^2}{2}\left(\tilde{\mathcal{B}}_1 + \tilde{\mathcal{B}}_{-1}\right)\partial_x^2 \mathrm{w}(x,t)\\
&-\frac{h^3}{6}\left(\tilde{\mathcal{B}}_1 - \tilde{\mathcal{B}}_{-1}\right)\partial_x^3 \mathrm{w}(x,t) + \mathcal{O}(h^4) + \Delta t\int_0^1 (1-\theta)\partial_t^2 \mathrm{w}(x,t+\theta\Delta t)d\theta.
\end{aligned}
\tag{52}
$$

The terms $\mathcal{O}(h^k)$ involve only spatial derivatives of w. Let us now consider the stationary solution z, namely, the solution of problem

$$
\Lambda\partial_x \mathrm{z} = \tilde{B}\mathrm{z}. \tag{53}
$$

First, we remark that if $\lambda_1\lambda_2 = 0$, i.e., $ac - b^2 = 0$, then all of the stationary solutions are constant with respect to x. Hence for such solutions

$$
T(x,t) = -h\tilde{C}\mathrm{z}.
$$

Any scheme with $\tilde{C} = 0$ is an AHO-∞ scheme. In the following we suppose that

$$
ac - b^2 \neq 0. \tag{54}
$$

For smooth data we get

$$
\partial_x^k \mathrm{z} = (\Lambda^{-1}\tilde{B})^k \mathrm{z}. \tag{55}
$$

We compute the local truncation error (52) on this particular solution. It does not depend on t anymore, and we denote it by $T_{stat}(x)$:

$$
\begin{aligned}
T_{stat}(x) = \quad & -\frac{h}{2}\left(\tilde{Q}(\Lambda^{-1}\tilde{B})^2 + 2\tilde{C} + 2\big(\tilde{\mathcal{B}}_1 - \tilde{\mathcal{B}}_{-1}\big)(\Lambda^{-1}\tilde{B})\right)\mathrm{z}(x)\\
& -\frac{h^2}{6}\left(-\tilde{B} + 3\big(\tilde{\mathcal{B}}_1 + \tilde{\mathcal{B}}_{-1}\big)\right)(\Lambda^{-1}\tilde{B})^2\mathrm{z}(x)\\
& -\frac{h^3}{24}\left(\tilde{Q}(\Lambda^{-1}\tilde{B}) + 4\big(\tilde{\mathcal{B}}_1 - \tilde{\mathcal{B}}_{-1}\big)\right)(\Lambda^{-1}\tilde{B})^3\mathrm{z}(x) + \mathcal{O}(h^4).
\end{aligned}
\tag{56}
$$

Example 1 (pointwise approximation of the source term (AHO1-UP)). Fixing, for the differential terms, the upwind approximation $\tilde{Q} = \mathrm{diag}(|\lambda_1|, |\lambda_2|)$, the basic scheme gives a first-order approximation even on the stationary solution:

$$
\tilde{B}_{up}^{-1} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \quad \tilde{B}_{up}^{0} = \begin{bmatrix} \tilde{b}_{11} & \tilde{b}_{12} \\ \tilde{b}_{21} & \tilde{b}_{22} \end{bmatrix}, \quad \tilde{B}_{up}^{1} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}. \tag{57}
$$

Example 2 (upwinding of the source term (AHO2-ROE) [14]). An improvement on the previous example is given by a second-order upwinding scheme:

$$\tilde{B}^{-1}_{roe} = \frac{1}{2}\begin{bmatrix} H(\lambda_1)\tilde{b}_{11} & H(\lambda_1)\tilde{b}_{12} \\ H(\lambda_2)\tilde{b}_{21} & H(\lambda_2)\tilde{b}_{22} \end{bmatrix}, \quad \tilde{B}^{0}_{roe} = \frac{1}{2}\begin{bmatrix} \tilde{b}_{11} & \tilde{b}_{12} \\ \tilde{b}_{21} & \tilde{b}_{22} \end{bmatrix},$$
$$\tilde{B}^{1}_{roe} = \frac{1}{2}\begin{bmatrix} (1-H(\lambda_1))\tilde{b}_{11} & (1-H(\lambda_1))\tilde{b}_{12} \\ (1-H(\lambda_2))\tilde{b}_{21} & (1-H(\lambda_2))\tilde{b}_{22} \end{bmatrix}, \tag{58}$$

where $H(\cdot)$ is the Heaviside function and where, as above, $\tilde{Q} = \mathrm{diag}(|\lambda_1|, |\lambda_2|)$.

A more accurate scheme is given by our AHO4 scheme, which is fourth-order on the stationary solution, as soon as all terms between round brackets in (56) vanish. The AHO4 coefficients $\tilde{\mathcal{B}}_{-1}$, $\tilde{\mathcal{B}}_0$, $\tilde{\mathcal{B}}_1$ are then uniquely defined by

$$\begin{cases} \tilde{\mathcal{B}}_{-1} = \tilde{B}/6 + \tilde{Q}\Lambda^{-1}\tilde{B}/8, \\ \tilde{\mathcal{B}}_0 = 2\tilde{B}/3 - h\tilde{Q}(\Lambda^{-1}\tilde{B})^2/4, \\ \tilde{\mathcal{B}}_1 = \tilde{B}/6 - \tilde{Q}\Lambda^{-1}\tilde{B}/8, \end{cases} \tag{59}$$

or equivalently

$$\tilde{\mathcal{B}}^{-1}_{aho} = \frac{1}{8}\begin{bmatrix} \tilde{b}_{11}\big(4/3 + \tilde{q}_1/\lambda_1\big) & \tilde{b}_{12}\big(4/3 + \tilde{q}_1/\lambda_1\big) \\ \tilde{b}_{21}\big(4/3 + \tilde{q}_2/\lambda_2\big) & \tilde{b}_{22}\big(4/3 + \tilde{q}_2/\lambda_2\big) \end{bmatrix},$$

$$\tilde{\mathcal{B}}^{0}_{aho} = \frac{2}{3}\begin{bmatrix} \tilde{b}_{11} & \tilde{b}_{12} \\ \tilde{b}_{21} & \tilde{b}_{22} \end{bmatrix} + \frac{adh}{4\lambda_1^2\lambda_2^2}\begin{bmatrix} \lambda_2\tilde{q}_1\tilde{b}_{11} & \lambda_2\tilde{q}_1\tilde{b}_{12} \\ \lambda_1\tilde{q}_2\tilde{b}_{21} & \lambda_1\tilde{q}_2\tilde{b}_{22} \end{bmatrix},$$

$$\tilde{\mathcal{B}}^{1}_{aho} = \frac{1}{8}\begin{bmatrix} \tilde{b}_{11}\big(4/3 - \tilde{q}_1/\lambda_1\big) & \tilde{b}_{12}\big(4/3 - \tilde{q}_1/\lambda_1\big) \\ \tilde{b}_{21}\big(4/3 - \tilde{q}_2/\lambda_2\big) & \tilde{b}_{22}\big(4/3 - \tilde{q}_2/\lambda_2\big) \end{bmatrix}.$$

The scheme is of fourth order for long times, as expressed in the following proposition.

PROPOSITION 4.1 (asymptotic high order). *Let* z *be a smooth stationary solution of system* (13). *Let* w_0 *be a small and smooth perturbation of* z.

Then for any compact set $K \subset \mathbb{R}$ *there exists a constant* $C_4(K)$ *such that the local truncation error* $T(x,t)$ *of the scheme* (13)–(59) *satisfies*

$$\|T(.,t)\|_{L^\infty(K)} \le C_4(K)\min\left(\Delta t + h, h^4 + \frac{1}{\Delta t\sqrt{t}}\right). \tag{60}$$

Proof. This result is an easy consequence of the uniform boundedness property (48) and of Theorem 2.4: There exists a constant $C > 0$ such that for all $t > 0$

$$\|\mathrm{w}(.,t) - \mathrm{z}\|_\infty \le \frac{C}{\sqrt{t}}.$$

Hence, using the notation (34) for w and z, we can write

$$\|T(.,t)\|_{L^\infty(K)} \le \frac{\|\mathrm{w}(.,t+\Delta t) - \mathrm{z}\|_\infty}{\Delta t} + \|T_{stat}\|_{L^\infty(K)} + \frac{\|S_{\Delta t}(\overline{Z} - \overline{W})\|_\infty}{\Delta t}.$$

Here we have to consider $\|T_{stat}\|_{L^\infty(K)}$ because the stationary solutions are not uniformly bounded in general. So, we find

$$\|T(.,t)\|_{L^\infty(K)} \leq \frac{C}{\Delta t\sqrt{t+\Delta t}} + C'h^4 + \frac{C}{\Delta t\sqrt{t}}.$$

To conclude, it is enough to use the consistency result (44) in the proof of Proposition 3.1. ∎

We address now the problem of the monotonicity of our schemes. Clearly the schemes AHO1-UP, given by (57), and AHO2-ROE, given by (58), are easily seen to be monotone under the standard upwind CFL condition. For the scheme AHO4 we have the following result.

PROPOSITION 4.2 (monotonicity). *The scheme* (13) *coupled with* (59) *satisfies the monotonicity requirements* (37)–(39) *under the following assumptions:*

$$h < \min_{1\leq i\leq 2} \frac{3|\lambda_i|}{|\tilde{b}_{ii}|}, \tag{61}$$

$$h \leq \left|\frac{2\lambda_1\lambda_2}{ad}\right| \quad if \quad \min(a\lambda_1, a\lambda_2) < 0, \tag{62}$$

$$1 + h\,\frac{|\tilde{b}_{ii}|}{3(4|\lambda_i| + h|\tilde{b}_{ii}|)} \leq \frac{\tilde{q}_i}{|\lambda_i|} \leq \frac{4}{3}, \quad i = 1,2, \tag{63}$$

and

$$\frac{\Delta t}{h} \leq \min_{1\leq i\leq 2} \left(\tilde{q}_i + h|\tilde{b}_{ii}|\left(\frac{adh\tilde{q}_i}{4\lambda_1\lambda_2\lambda_i} + \frac{2}{3}\right)\right)^{-1}. \tag{64}$$

Moreover these conditions are not empty.

Proof. We recall that for system (13) we have $\tilde{b}_{11}, \tilde{b}_{22} < 0$ and $\tilde{b}_{12}, \tilde{b}_{21} > 0$ and that $\lambda_1\lambda_2 \neq 0$.

We begin with condition (39). It can be written as

$$\begin{cases} \dfrac{\lambda_i}{2h} + \dfrac{\tilde{b}_{ii}}{6} + \dfrac{\tilde{q}_i}{\lambda_i}\left(\dfrac{\lambda_i}{2h} + \dfrac{\tilde{b}_{ii}}{8}\right) \geq 0, \\ -\dfrac{\lambda_i}{2h} + \dfrac{\tilde{b}_{ii}}{6} + \dfrac{\tilde{q}_i}{\lambda_i}\left(\dfrac{\lambda_i}{2h} - \dfrac{\tilde{b}_{ii}}{8}\right) \geq 0, \end{cases} \quad i = 1,2. \tag{65}$$

Adding those inequalities shows that necessarily $\tilde{q}_i > 0$. Let us denote

$$r_i = \frac{|\lambda_i|}{2h}, \quad s_i = -\frac{\tilde{b}_{ii}}{2}, \quad \alpha_i = \frac{\tilde{q}_i}{|\lambda_i|}.$$

Conditions (65) read as

$$\begin{cases} r_i - \dfrac{s_i}{3} + \alpha_i\left(r_i - \dfrac{s_i}{4}\right) \geq 0, \\ -r_i - \dfrac{s_i}{3} + \alpha_i\left(r_i + \dfrac{s_i}{4}\right) \geq 0, \end{cases} \quad i = 1,2.$$

Condition (61) ensures that $r_i > \frac{s_i}{3} > \frac{s_i}{4}$, so that we have to verify that

$$\alpha_i \geq \frac{r_i + s_i/3}{r_i + s_i/4}, \quad i = 1,2,$$

which also is written

$$1 + h\,\frac{|\tilde{b}_{ii}|}{3(4|\lambda_i| + h|\tilde{b}_{ii}|)} \leq \frac{q_i}{|\lambda_i|}, \quad i = 1, 2. \tag{66}$$

Consequently, under the assumptions of the proposition, condition (39) is satisfied. Condition (37) is satisfied if and only if

$$\begin{cases} \dfrac{4}{3} \pm \dfrac{\tilde{q}_i}{\lambda_i} \geq 0, \\ \dfrac{2}{3} + \dfrac{adh}{4|\lambda_i|\lambda_j}\dfrac{\tilde{q}_i}{|\lambda_i|} \geq 0, \end{cases} \qquad i = 1, 2, \quad j \neq i. \tag{67}$$

The first inequality is ensured by assumption (63). If $a\lambda_j$ is nonnegative, the second inequality is satisfied. Otherwise, we have to ensure that

$$h \leq \frac{8|\lambda_i\lambda_j|}{3|a|d}\,\frac{|\lambda_i|}{\tilde{q}_i}.$$

As $\frac{\tilde{q}_i}{|\lambda_i|} \leq \frac{4}{3}$, assumption (62) implies that this inequality is satisfied. It remains to verify that (38) holds. Let us denote

$$\mathcal{A}_i = \frac{2}{3} + \frac{adh}{4\lambda_1\lambda_2}\frac{\tilde{q}_i}{\lambda_i}.$$

This quantity is nonnegative, and (38) can be written as

$$1 - \frac{\Delta t}{h}\left(\tilde{q}_i - h\tilde{b}_{ii}\mathcal{A}_i\right) \geq 0, \quad i = 1, 2.$$

Taking into account the facts that $\tilde{q}_i > 0$ and $\tilde{b}_{ii} < 0$, this is equivalent to the CFL restriction (64). To end the proof we remark that the sign properties ensure that the conditions (61)–(64) are not empty. ∎

4.2. An AHO scheme for the diffusion limit. In this section we construct a numerical scheme for system (11) which is high-order accurate with respect to the parabolic asymptotic problem (6). The numerical error decays in time faster than the error of other schemes.

We consider for simplicity only the case $c = -a$, looking at the adimensional problem satisfied by $U(x, b^{-1}t)$:

$$\begin{cases} u_t + \tilde{a}u_x + v_x = 0, \\ \\ v_t + u_x - \tilde{a}v_x = -\tilde{d}v, \end{cases} \tag{68}$$

where $\tilde{a} = a/b$ and $\tilde{d} = b^{-1}d$. The Chapman–Enskog limit of system (68) is given by

$$\begin{cases} \hat{u}_t + \tilde{a}\hat{u}_x = \frac{1}{\tilde{d}}\hat{u}_{xx}, \\ \\ \hat{v} = -\frac{1}{\tilde{d}}\hat{u}_x. \end{cases} \tag{69}$$

Let $U_j^n = (u_j^n, v_j^n)$ be the numerical approximation of the analytical solution $U = (u, v)$ to problem (68). Let us consider the following finite difference approximation:

$$\begin{aligned} &\frac{U_j^{n+1} - U_j^n}{\Delta t} + \frac{A}{2h}\left(U_{j+1}^n - U_{j-1}^n\right) - \frac{R\tilde{Q}R^{-1}}{2h}\left(U_{j+1}^n - 2U_j^n + U_{j-1}^n\right) \\ &\quad = R\tilde{\mathcal{B}}_{-1}R^{-1}U_{j-1}^n + R\tilde{\mathcal{B}}_0R^{-1}U_j^n + R\tilde{\mathcal{B}}_1R^{-1}U_{j+1}^n. \end{aligned} \tag{70}$$

Let $\pm\lambda = \pm\sqrt{1+\tilde{a}^2}$ be the eigenvalues of matrix A and

$$R = \begin{pmatrix} 1 & 1 \\ -(\tilde{a}+\lambda) & \lambda - \tilde{a} \end{pmatrix}.$$

Taking

$$\tilde{Q} = \begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix},$$

we have that

$$Q = R\tilde{Q}R^{-1} = \tilde{Q}.$$

We set also

$$C = R\tilde{C}R^{-1}, \quad \mathcal{B}_k = R\tilde{\mathcal{B}}_kR^{-1}, \quad k = -1, 0, 1.$$

By Taylor expansion, the numerical approximation (70) is consistent with system

$$U_t + AU_x - BU = h\left[\frac{Q}{2}U_{xx} + (\mathcal{B}_1 - \mathcal{B}_{-1})U_x + CU\right] + \mathcal{O}(h^2 + \Delta t). \tag{71}$$

Going into details, for different schemes, the Chapman–Enskog limit of system (71) gives the following:

- For the source term pointwise approximation (AHO1-UP), given by (57), we have

$$\mathcal{B}_1 - \mathcal{B}_{-1} = 0, \quad C = 0,$$

$$\begin{cases} \hat{u}_t + \tilde{a}\hat{u}_x = \left(\frac{1}{d} + \frac{h\lambda}{2}\right)\hat{u}_{xx}, \\ \hat{v} = -\frac{1}{d}\hat{u}_x. \end{cases} \tag{72}$$

- For the upwinding of the source term (AHO2-ROE), given by (58), we have

$$\mathcal{B}_1 - \mathcal{B}_{-1} = \frac{\tilde{d}}{2\lambda}\begin{pmatrix} 0 & 1 \\ 0 & -\tilde{a} \end{pmatrix}, \quad C = 0,$$

$$\begin{cases} \hat{u}_t + \tilde{a}\hat{u}_x = \left(\frac{1}{d} + h\frac{\lambda}{2} - h\frac{1}{2\lambda}\right)\hat{u}_{xx}, \\ \hat{v} = -\frac{1}{d}\hat{u}_x. \end{cases} \tag{73}$$

- For the asymptotic high-order (AHO4) given by (59), we have

$$\mathcal{B}_1 - \mathcal{B}_{-1} = \frac{\tilde{d}}{4\lambda}\begin{pmatrix} 0 & 1 \\ 0 & -\tilde{a} \end{pmatrix}, \quad C = \frac{\tilde{a}\tilde{d}^2}{4\lambda^3}\begin{pmatrix} 0 & 1 \\ 0 & -\tilde{a} \end{pmatrix},$$

(74)
$$\begin{cases} \hat{u}_t + \tilde{a}\left(1 + h\frac{\tilde{a}\tilde{d}}{\left(4\lambda^3 + h\tilde{a}^2\tilde{d}\right)}\right)\hat{u}_x = \left[\frac{4\lambda^3}{\tilde{d}\left(4\lambda^3 + h\tilde{a}^2\tilde{d}\right)} + h\frac{\lambda}{2} - h\frac{\lambda^2}{\left(4\lambda^3 + h\tilde{a}^2\tilde{d}\right)}\right]\hat{u}_{xx}, \\ \hat{v} = -\frac{4\lambda^3}{\tilde{d}\left(4\lambda^3 + h\tilde{a}^2\tilde{d}\right)}\hat{u}_x. \end{cases}$$

Notice that for all schemes the consistency with the parabolic system (69) is of order $\mathcal{O}(h)$. In all cases, the diffusive term has a perturbation of order $\mathcal{O}(h)$. In particular, the pointwise approximation is the most diffusive one. Moreover, we observe that the AHO4 approximation has a perturbation of order $\mathcal{O}(h)$ also on the drift term.

According to this discussion, it is worth looking for a scheme which is at least of order $\mathcal{O}(h^2)$ with respect to the parabolic problem (69). Consider system (71) with $Q = \tilde{Q} = \mathrm{diag}(q, q)$,

(75)
$$\begin{cases} u_t + \tilde{a}u_x + v_x = h\left[\frac{q}{2}u_{xx} + (\beta_{11}^1 - \beta_{11}^{-1})u_x + (\beta_{12}^1 - \beta_{12}^{-1})v_x + c_{11}u + c_{12}v\right] \\ \qquad\qquad + \mathcal{O}(h^2 + \Delta t), \\ v_t + u_x - \tilde{a}v_x + \tilde{d}v = h\left[\frac{q}{2}v_{xx} + (\beta_{21}^1 - \beta_{21}^{-1})u_x + (\beta_{22}^1 - \beta_{22}^{-1})v_x + c_{21}u + c_{22}v\right] \\ \qquad\qquad + \mathcal{O}(h^2 + \Delta t). \end{cases}$$

Assuming enough regularity on the data, the decay properties of the solution stated in Theorem 2.5 allow us to consider, when t goes to $+\infty$, the following reduced system:

(76)
$$\begin{cases} u_t + \tilde{a}u_x + v_x = h\left[\frac{q}{2}u_{xx} + (\beta_{11}^1 - \beta_{11}^{-1})u_x + (\beta_{12}^1 - \beta_{12}^{-1})v_x + c_{11}u + c_{12}v\right], \\ u_x + \tilde{d}v = h\left[(\beta_{21}^1 - \beta_{21}^{-1})u_x + c_{21}u + c_{22}v\right]. \end{cases}$$

Then, by choosing $c_{11} = c_{21} = 0$ and setting $\gamma_{ij} = \beta_{ij}^1 - \beta_{ij}^{-1}$, $i, j = 1, 2$, from the second equation, we have

$$v_x = \frac{1 - h\gamma_{21}}{-\tilde{d} + hc_{22}}u_{xx}.$$

By replacing this expression in the first equation and neglecting second-order terms, we get

$$u_t + \left[\tilde{a} - h\left(\gamma_{11} - c_{12}\frac{1 - h\gamma_{21}}{\tilde{d} - hc_{22}}\right)\right]u_x = \left[\frac{1 - h\gamma_{21}}{\tilde{d} - hc_{22}}(1 - h\gamma_{12}) + h\frac{q}{2}\right]u_{xx} + \mathcal{O}(h^2 + \Delta t).$$

We now look for coefficients $c_\cdot$ and $\gamma_\cdot$ such that the drift and the diffusion terms are equal to $\tilde{a}$ and $1/\tilde{d}$, respectively.

- Studying the drift term, we have

$$\tilde{a}' = \tilde{a} + h\left(c_{12}\frac{1-h\gamma_{21}}{\tilde{d}-hc_{22}} - \gamma_{11}\right) = \tilde{a} + h\frac{c_{12}(1-h\gamma_{21}) - \gamma_{11}(\tilde{d}-hc_{22})}{\tilde{d}-hc_{22}}. \tag{77}$$

  Then, to have $c_\cdot$ and $\gamma_\cdot$ independent of h, we choose

$$c_{12} = \gamma_{11}\tilde{d}, \quad c_{22} = \gamma_{21}\tilde{d}.$$

- Studying the diffusion term, we have

$$\tilde{d}' = \frac{1-h\gamma_{21}}{\tilde{d}-hc_{22}}(1-h\gamma_{12}) + h\frac{q}{2} = \frac{2(1-h\gamma_{21})(1-h\gamma_{12}) + hq(\tilde{d}-hc_{22})}{2(\tilde{d}-hc_{22})}. \tag{78}$$

  We then write

$$\tilde{d}' = \frac{1}{\tilde{d}} + \tilde{d}'',$$

  and, therefore,

$$\tilde{d}'' = \frac{h(-2\tilde{d}(\gamma_{12}+\gamma_{21}) + 2c_{22} + \tilde{d}^2 q) + h^2(2\tilde{d}\gamma_{21}\gamma_{12} - \tilde{d}qc_{22})}{2\tilde{d}(\tilde{d}-hc_{22})}.$$

  For $c_{22} = \gamma_{21}\tilde{d}$, we obtain that $\tilde{d}'' = 0$ if

$$\gamma_{12} = \frac{\tilde{d}q}{2}.$$

To summarize, a class of finite different schemes of the form (70), consistently of order $\mathcal{O}(h^2 + \Delta t)$ with respect the asymptotic problem (69), is defined by selecting

$$c_{11} = c_{21} = 0, \quad c_{12} = \gamma_{11}\tilde{d}, \quad c_{22} = \gamma_{21}\tilde{d}, \quad \gamma_{12} = \frac{\tilde{d}q}{2}. \tag{79}$$

There are still some parameters to define, which will be chosen to guarantee the monotonicity properties of the scheme. Let us focus on the particular case $q = \lambda$ (upwind approximation for the first-order derivatives), which arises if we look for the minimal artificial diffusion in the scheme.

PROPOSITION 4.3 (monotonicity for the AHO2p scheme). *Let $q = \lambda$, and set $a_\pm = \lambda \pm \tilde{a}$. The scheme* (29) *with coefficients $\tilde{\mathcal{B}}_{-1}$, $\tilde{\mathcal{B}}_0$, $\tilde{\mathcal{B}}_1$ defined by*

$$\tilde{\mathcal{B}}_0 = \frac{\tilde{d}}{4\lambda}\begin{pmatrix} -2a_+(\lambda^2+1) & a_- \\ a_+ & -2a_-(\lambda^2+1) \end{pmatrix},$$

$$\tilde{\mathcal{B}}_{\pm 1} = \tfrac{1}{2}\left(\tilde{B} - \tilde{\mathcal{B}}_0 \pm \tilde{\Gamma} + h\tilde{C}\right), \tag{80}$$

$$\tilde{\Gamma} = \tfrac{\tilde{d}q}{2}\begin{pmatrix} -a_+ & 0 \\ 0 & a_- \end{pmatrix}, \quad \tilde{C} = \tfrac{\tilde{d}^2 q}{4\lambda}\begin{pmatrix} a_+ & -a_- \\ -a_+ & a_-, \end{pmatrix}$$

satisfies the monotonicity assumptions (37)–(39) *under the conditions:*

$$h < \frac{4}{\tilde{d}}, \quad \Delta t \le \frac{2\lambda h}{2\lambda^2 + h\tilde{d}(\lambda+\tilde{a})(\lambda^2+1)}. \tag{81}$$

Proof. Let us consider system (68) in diagonal form

$$\begin{cases} w_t - \lambda w_x = \frac{\tilde{d}}{2\lambda}\left[-(\tilde{a}+\lambda)w + (\lambda-\tilde{a})z\right], \\ \\ z_t + \lambda z_x = \frac{\tilde{d}}{2\lambda}\left[(\tilde{a}+\lambda)w - (\lambda-\tilde{a})z\right] \end{cases} \tag{82}$$

and the general numerical finite difference approximation (29). Starting from relations (79), we then construct the AHO2p coefficients $\tilde{\mathcal{B}}_{-1}$, $\tilde{\mathcal{B}}_0$, $\tilde{\mathcal{B}}_1$ (80) to satisfy conditions (37)–(39).

First of all, we look at condition (37). We have to prove that

$$\tilde{\beta}_{ij}^k \geq 0 \quad \text{for} \quad i \neq j.$$

Clearly, from consistency (36) and setting $\tilde{\Gamma} = \tilde{\mathcal{B}}_1 - \tilde{\mathcal{B}}_{-1}$, we have

$$\tilde{\mathcal{B}}_{\pm 1} = \frac{1}{2}\left(\tilde{B} - \tilde{\mathcal{B}}_0 \pm \tilde{\Gamma} + h\tilde{C}\right). \tag{83}$$

Since h goes to 0, in practice we have to impose that

$$\tilde{\beta}_{ij} - |\tilde{\gamma}_{ij}| > \tilde{\beta}_{ij}^0 \geq 0 \quad \text{for} \quad i \neq j. \tag{84}$$

By setting $a_\pm = \lambda \pm \tilde{a}$, we can write

$$\tilde{B} = -\frac{\tilde{d}}{2\lambda}\begin{pmatrix} a_+ & -a_- \\ -a_+ & a_- \end{pmatrix}.$$

So, to verify the inequalities (84), it is enough to show that

$$\frac{\tilde{d}}{2\lambda}a_- > |\tilde{\gamma}_{12}|, \qquad \frac{\tilde{d}}{2\lambda}a_+ > |\tilde{\gamma}_{21}|, \tag{85}$$

taking then the off-diagonal values of $\tilde{\mathcal{B}}_0$ small enough, for instance,

$$\tilde{\beta}_{ij}^0 = \frac{1}{2}\left(\tilde{\beta}_{ij} - |\tilde{\gamma}_{ij}|\right) \text{ for } i \neq j. \tag{86}$$

Now we can compute $\tilde{\Gamma}$. We have

$$\tilde{\Gamma} = R^{-1}\Gamma R = \frac{1}{2\lambda}\begin{pmatrix} a_- & -1 \\ a_+ & 1 \end{pmatrix}\begin{pmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{21} & \gamma_{22} \end{pmatrix}\begin{pmatrix} 1 & 1 \\ -a_+ & a_- \end{pmatrix}$$

$$= \frac{1}{2\lambda}\begin{pmatrix} a_-(\gamma_{11} - a_+\gamma_{12}) - (\gamma_{21} - a_+\gamma_{22}) & a_-(\gamma_{11} + a_-\gamma_{12}) - (\gamma_{21} + a_-\gamma_{22}) \\ \\ a_+(\gamma_{11} - a_+\gamma_{12}) + (\gamma_{21} - a_+\gamma_{22}) & a_+(\gamma_{11} + a_-\gamma_{12}) + (\gamma_{21} + a_-\gamma_{22}) \end{pmatrix}.$$

So inequalities (85) read now

$$\begin{aligned} \tilde{d}a_- &> \left|a_-(\gamma_{11} + a_-\gamma_{12}) - (\gamma_{21} + a_-\gamma_{22})\right|, \\ \tilde{d}a_+ &> \left|a_+(\gamma_{11} - a_+\gamma_{12}) + (\gamma_{21} - a_+\gamma_{22})\right|. \end{aligned} \tag{87}$$

The best we can do is to put to 0 the right-hand side in both inequalities, which gives, recalling that $\gamma_{12} = \frac{\tilde{d}q}{2}$,

$$-a_-(\gamma_{11} - \gamma_{22}) + \gamma_{21} = a_-^2 \frac{\tilde{d}q}{2},$$

$$a_+(\gamma_{11} - \gamma_{22}) + \gamma_{21} = a_+^2 \frac{\tilde{d}q}{2}.$$

By solving the linear system, we find

$$\gamma_{11} - \gamma_{22} = \tilde{d}q\tilde{a}, \qquad \gamma_{21} = \frac{\tilde{d}q}{2} = \gamma_{12}. \tag{88}$$

We can quantify the influence of the matrix $\tilde{C}$ in the monotonicity. Let us just consider the parabolic case. It is clear that, since the limit $h \to 0$ is relevant, we can start from conditions (88). Now we have that

$$\tilde{C} = R^{-1}CR = \frac{\tilde{d}}{2\lambda} \begin{pmatrix} a_- & -1 \\ a_+ & 1 \end{pmatrix} \begin{pmatrix} 0 & \gamma_{11} \\ 0 & \gamma_{21} \end{pmatrix} \begin{pmatrix} 1 & 1 \\ -a_+ & a_- \end{pmatrix}$$

$$= \frac{\tilde{d}}{2\lambda} \begin{pmatrix} -a_+a_-\gamma_{11} + a_+\gamma_{21} & a_-^2\gamma_{11} - a_-\gamma_{21} \\ -a_+^2\gamma_{11} - a_+\gamma_{21} & a_+a_-\gamma_{11} + a_-\gamma_{21} \end{pmatrix}.$$

Then inequalities (84) can be refined to

$$\tilde{\beta}_{ij} - |\tilde{\gamma}_{ij}| + h\tilde{c}_{ij} > \tilde{\beta}^0_{ij} \geq 0 \quad \text{for} \quad i \neq j. \tag{89}$$

Using conditions (88), we need to verify both

$$1 + \frac{h}{2\lambda}(a_-\gamma_{11} - \gamma_{21}) > 0 \quad \text{and} \quad 1 - \frac{h}{2\lambda}(a_+\gamma_{11} + \gamma_{21}) > 0, \tag{90}$$

which are equivalent to

$$1 - \frac{h}{4\lambda}\tilde{d}q \pm \frac{h}{2\lambda}a_\mp\gamma_{11} > 0. \tag{91}$$

So the optimal choice is given by

$$\gamma_{11} = 0, \qquad \gamma_{22} = -\tilde{d}q\tilde{a}, \qquad \gamma_{21} = \gamma_{12} = \frac{\tilde{d}q}{2}, \tag{92}$$

with the restriction on the space step

$$h < \frac{4\lambda}{\tilde{d}q}. \tag{93}$$

To summarize, up to now we have defined

$$\Gamma = \frac{\tilde{d}q}{2} \begin{pmatrix} 0 & 1 \\ 1 & -2\tilde{a} \end{pmatrix}, \quad C = \begin{pmatrix} 0 & 0 \\ 0 & \gamma_{21}\tilde{d}, \end{pmatrix}.$$

From these two relations we obtain in diagonal variables $\tilde{\Gamma}$ and $\tilde{C}$ defined by (80).

To complete the construction of the scheme and the study of monotonicity, we need to verify conditions (38) and (39). In this case, they are given by

$$1 - \frac{\Delta t}{h} q + \Delta t \tilde{\beta}^0_{ii} \geq 0, \quad i = 1, 2, \tag{94}$$

$$\frac{\lambda}{2h} + \frac{q}{2h} + \tilde{\beta}^{-1}_{ii} \geq 0, \qquad -\frac{\lambda}{2h} + \frac{q}{2h} + \tilde{\beta}^{1}_{ii} \geq 0, \quad i = 1, 2. \tag{95}$$

From (83), we have

$$\tilde{\beta}^{-1}_{11} = \frac{1}{2}\left(\tilde{\beta}_{11} - \tilde{\beta}^0_{11} - \tilde{\gamma}_{11} + h\tilde{c}_{11}\right) = \frac{\tilde{d}}{4} a_+ \left(q - \frac{1}{\lambda}\right) - \frac{\tilde{\beta}^0_{11}}{2},$$

$$\tilde{\beta}^{-1}_{22} = \frac{1}{2}\left(\tilde{\beta}_{22} - \tilde{\beta}^0_{22} - \tilde{\gamma}_{22} + h\tilde{c}_{22}\right) = -\frac{\tilde{d}}{4} a_- \left(q + \frac{1}{\lambda}\right) - \frac{\tilde{\beta}^0_{22}}{2},$$

$$\tilde{\beta}^{1}_{11} = \frac{1}{2}\left(\tilde{\beta}_{11} - \tilde{\beta}^0_{11} + \tilde{\gamma}_{11} + h\tilde{c}_{11}\right) = -\frac{\tilde{d}}{4} a_+ \left(q + \frac{1}{\lambda}\right) - \frac{\tilde{\beta}^0_{11}}{2},$$

$$\tilde{\beta}^{1}_{22} = \frac{1}{2}\left(\tilde{\beta}_{22} - \tilde{\beta}^0_{22} + \tilde{\gamma}_{22} + h\tilde{c}_{22}\right) = \frac{\tilde{d}}{4} a_- \left(q - \frac{1}{\lambda}\right) - \frac{\tilde{\beta}^0_{22}}{2}.$$

From (95), the coefficient $\tilde{\beta}^0_{11}$ has to verify the relations

$$\begin{aligned} &\frac{1}{h}(q + \lambda) + \frac{\tilde{d}}{2} a_+ \left(q - \frac{1}{\lambda}\right) - \tilde{\beta}^0_{11} \geq 0, \\ &\frac{1}{h}(q - \lambda) - \frac{\tilde{d}}{2} a_+ \left(q + \frac{1}{\lambda}\right) - \tilde{\beta}^0_{11} \geq 0, \end{aligned} \tag{96}$$

and $\tilde{\beta}^0_{22}$

$$\begin{aligned} &\frac{1}{h}(q + \lambda) - \frac{\tilde{d}}{2} a_- \left(q + \frac{1}{\lambda}\right) - \tilde{\beta}^0_{22} \geq 0, \\ &\frac{1}{h}(q - \lambda) + \frac{\tilde{d}}{2} a_- \left(q - \frac{1}{\lambda}\right) - \tilde{\beta}^0_{22} \geq 0. \end{aligned} \tag{97}$$

For $q = \lambda$, a good choice for $\tilde{\beta}^0_{11}$ and $\tilde{\beta}^0_{22}$ should be given by

$$\tilde{\beta}^0_{11} = -\frac{\tilde{d}}{2} a_+ \left(\frac{\lambda^2 + 1}{\lambda}\right), \quad \tilde{\beta}^0_{22} = -\frac{\tilde{d}}{2} a_- \left(\frac{\lambda^2 + 1}{\lambda}\right). \tag{98}$$

Substituting these two expressions in condition (94), we get the following restriction on the time step Δt:

$$\begin{aligned} &1 - \frac{\Delta t}{h}\lambda - \frac{\Delta t}{2}\tilde{d} a_+ \left(\frac{\lambda^2 + 1}{\lambda}\right) \geq 0, \\ &1 - \frac{\Delta t}{h}\lambda - \frac{\Delta t}{2}\tilde{d} a_- \left(\frac{\lambda^2 + 1}{\lambda}\right) \geq 0. \end{aligned} \tag{99}$$

Therefore,

$$\Delta t \leq \frac{2\lambda h}{2\lambda^2 + h\tilde{d}a_+(\lambda^2+1)}. \tag{100}$$

To conclude, from relations (86) and (98) we get the AHO2p coefficients $\tilde{\mathcal{B}}_{-1}$, $\tilde{\mathcal{B}}_0$, $\tilde{\mathcal{B}}_1$ defined by (80), under the monotonicity assumptions (93) with $q = \lambda$ and (100). ∎

5. Numerical tests. In this section our aim is to show how, for large time simulations, AHO schemes give better numerical results than standard approximations. Therefore, we focus our attention on the numerical error as a function of time: For all tests, we fix the grid steps $h = \mathcal{O}(10^{-2})$ and Δt satisfying the CFL conditions, and we plot the error as the time $t = n\Delta t$ increases.

For all schemes under consideration we set $\tilde{Q} = \mathrm{diag}(|\lambda_1|, |\lambda_2|)$. The reference solution is obtained by the Roe-type scheme AHO2-ROE (58), with $h = \mathcal{O}(10^{-3})$.

5.1. Test n.1. Let us consider system (11) around the stationary and time-asymptotic stable solutions given by

$$u = u_0 - bv_0 e^{px}/a, \quad v = v_0 e^{px} \tag{101}$$

for some constants u_0, v_0 and for $p = -ad/(ac - b^2)$.

First, we compare the AHO4 scheme, with the standard first-order pointwise upwind scheme (57), which is actually just an AHO1-UP scheme, and with the AHO2-ROE scheme (58). The parameters in the test are $a = 1$, $c = -a$, $b = 1$, and $d = 5$. In Figure 1, we plot the l^∞-error for the problem as a function of time. Remark that, even if the total time of simulation is $T = 5$, this interval is divided into eight subintervals to plot the error curve.

In Figures 1(a) and 1(b) we have the evolution of the error, when the initial data are taken to be the stationary solution. In this case all of the schemes do not exactly preserve this stable state, but anyway AHO schemes show a better accuracy, which is proportional to their formal asymptotic order. In Figures 1(c) and 1(d) we show the evolution of the error when the initial data are given by small, compactly supported perturbations of the stationary solutions. In that case, the errors evolve in time, and the AHO2-ROE and the AHO4 improve clearly their accuracy for large times. Moreover, we show that the AHO4 scheme gives a good approximation also at intermediate times, when the perturbation is still present in the solution. In Figure 2, we plot the solution U as obtained by the AHO4 scheme at time $T = 2.5$ (see the bottom of Figure 2). To see the properties of the AHO4 scheme, we plot, at the top of these plots, the pointwise difference between the approximate solutions, given by schemes AHO1-UP, AHO2-ROE, and AHO4, and the reference solution, which is obtained by the Roe-type scheme AHO2-ROE (58), with $h = \mathcal{O}(10^{-3})$. Again, the AHO4 scheme is better performing.

5.2. Test n.2. Let us consider for system (11) the constant equilibrium state $u = 1$ and $v = 0$. The parameters in the test are again $a = 1$, $c = -a$, $b = 1$, and $d = 5$. We consider a small, compactly supported perturbation of this constant solution as the initial data, and we expect a diffusive behavior of the solution near the corresponding solution of problem (69). Actually what we look for is always a better approximation of the hyperbolic problem, but, to obtain that, we need to force the high-order consistency with the main terms in the asymptotic behavior, which are given by the diffusive expansion.

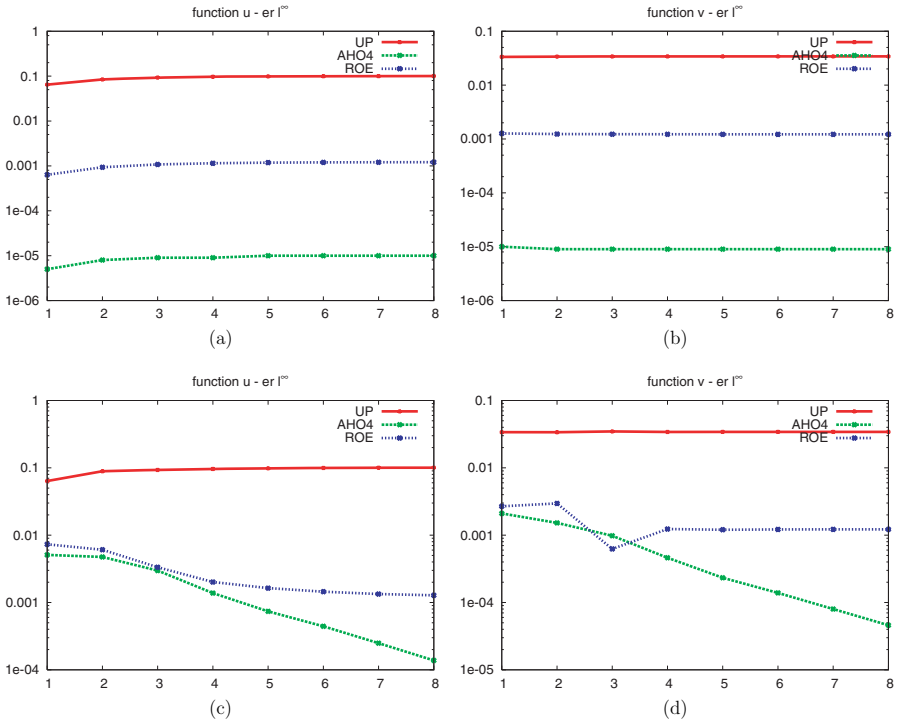


FIG. 1. *Test n.*1. (a)–(b) *Evolution of the* l^∞*-error, when the initial data are taken to be the stationary solutions* (101). (c)–(d) *Evolution of the error, when the initial data are given by a small, compactly supported perturbation of a stationary solution.*

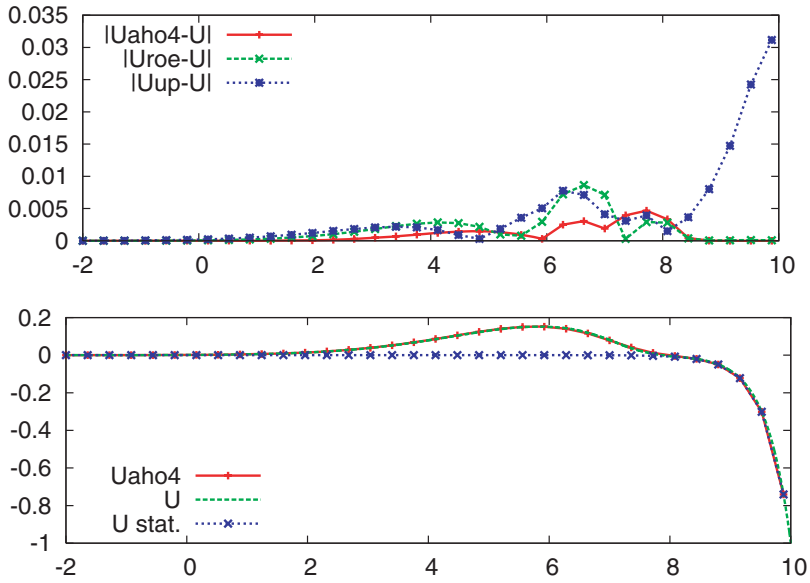


FIG. 2. *Test n.*1. *Bottom: Plot of the solution U as obtained by the AHO*4 *scheme at time* $T = 2.5$*, starting from a small perturbation of the stationary solution. Top: Pointwise difference between the approximates solutions given, respectively, by schemes AHO*1*-UP, AHO*2*-ROE, and AHO*4 *and the reference solution.*

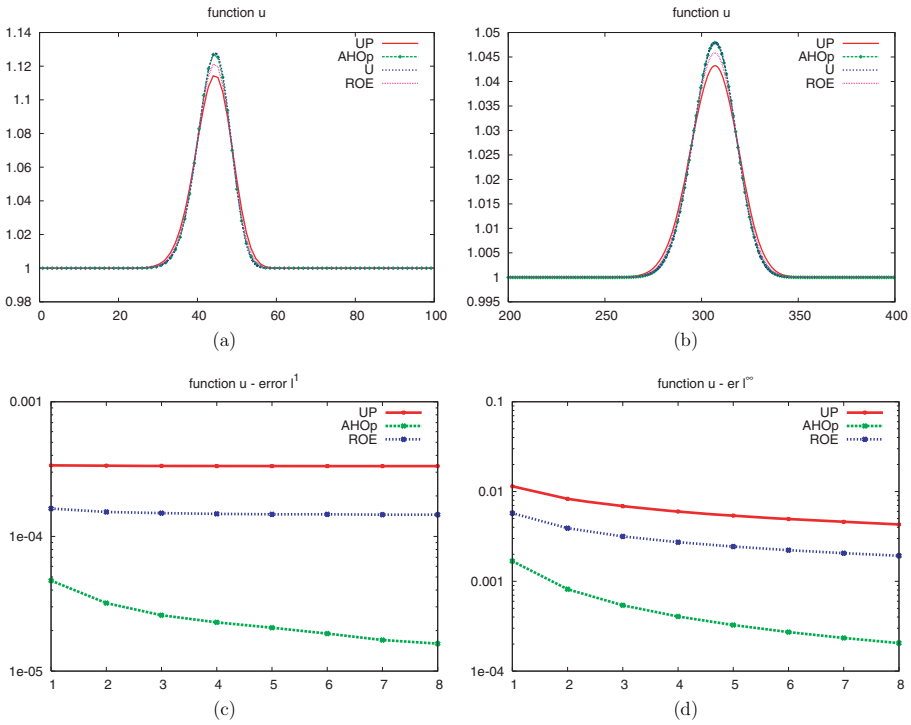


FIG. 3. *Test n.*2. (a)–(b) *Plot of different approximations of the function* u *at times* $T = 43$ *and* $T = 350$. (c)–(d) *Time evolution of the* l^1*- and* l^∞*-errors for the different approximations.*

More specifically, here we compare the AHO2p scheme (80) with the standard first-order pointwise upwind scheme (57) and with the AHO2-ROE scheme (58). The reference solution is always obtained by the AHO2-ROE scheme with $h = \mathcal{O}(10^{-3})$.

In Figures 3(a) and 3(b), we plot the different approximations of the function u at times $T = 43$ and $T = 350$. In Figures 3(c) and 3(d), we plot, respectively, the l^1- and l^∞-errors as a function of time. Again, even if the total time of simulation is $T = 350$, this interval is divided into eight subintervals to plot the error curve.

The numerical results show a clear better performance of the AHO2p scheme, in particular, for large times, as expected by our asymptotic analysis. To end, in Figure 4 we compare all of the considered schemes with the solution of the diffusion limit problem (69).

Actually this better performance is not surprising and can be explained in terms of a better time decay of the error for the AHO2p scheme. We denote by $\hat{u}^P$ the analytical solution of the parabolic problem (69) and by u^H the solution of (68). From Theorem 2.5, we know that

$$\|(\hat{u}^P - u^H)(t)\|_{L^\infty} \leq Ct^{-1} \tag{102}$$

for a positive constant C.

By construction, for large time, the scheme AHO2p closely follows the parabolic asymptotic problem (69), and so we expect that the same decay property (102) should then hold for the error function, namely,

$$e(t) = \|(U^h_{AHO2p} - u^H)(t)\|_{L^\infty} \leq Ct^{-1}, \tag{103}$$

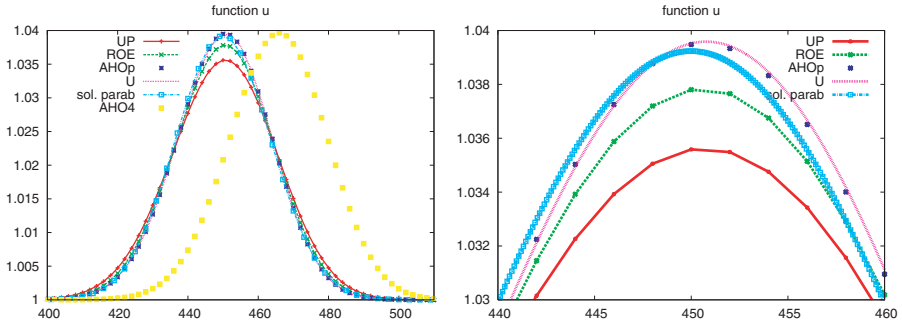


FIG. 4. *Test n.2.* (a) *Different approximations of the u component, against the reference solution of the hyperbolic problem* (11) *and of the parabolic problem* (69). (b) *Zoom on the top of the solution, to magnify the different performances.*

TABLE 1

Evaluation of constants γ and C for relation (104)*, where the solution u is given, first by the approximation of $\hat{u}^P$ computed by Crank–Nicolson scheme with a fine grid and then, respectively, by the schemes AHO2p* (80)*, AHO1-UP* (57)*, and AHO2-ROE* 58.

	$\|(u-u^H)(t)\|_{L^\infty} = Ct^{-\gamma}$	
u	γ	C
$\hat{u}^P$	0.942868	0.385312
AHO2p	0.934315	0.050421
UP	0.480781	0.071943
ROE	0.497644	0.035540

where U^h_{AHO2p} is the numerical solution of (68) obtained with the AHO2p scheme (80).

Then we look, from numerical tests, for a constant C and an exponent of decay γ which best fit the equality

$$e(t) = \|(u^H - U^h)(t)\|_{L^\infty} = Ct^{-\gamma} \tag{104}$$

for different numerical approximations U^h.

Using two instants of time t_1 and t_2, to evaluate γ, we have the following estimate:

$$\gamma \approx \frac{\ln\left[e(t_1)/e(t_2)\right]}{\ln[t_2/t_1]},$$

and C is estimated by

$$C \approx t_1^\gamma e(t_1) \approx t_2^\gamma e(t_2).$$

In Table 1, we show the values obtained for different schemes, where u^H is our reference solution of (68), always obtained by the AHO2-ROE scheme with a fine grid, say, the $h = \mathcal{O}(10^{-3})$, and where the solution $\hat{u}^P$ of (69) has been computed by the Crank–Nicolson scheme with the same fine grid. The table shows that the AHO2p scheme, unlike all of the other schemes, closely follows the theoretical time decay behavior of the analytical solution.

Acknowledgment. We thank warmly Bernard Hanouzet, in particular, for some useful discussions and his sometimes irritating questions, which as usual have been decisive to develop the scheme AHO2p.

REFERENCES

[1] A. Bermudez and M. E. Vasquez, *Upwind methods for hyperbolic conservation laws with source terms*, Comput. & Fluids, 23 (1994), pp. 1049–1071.

[2] S. Bianchini, B. Hanouzet, and R. Natalini, *Asymptotic behavior of smooth solutions for partially dissipative hyperbolic systems with a convex entropy*, Comm. Pure Appl. Math., 60 (2007), pp. 1559–1622.

[3] R. Botchorishvili, B. Perthame, and A. Vasseur, *Equilibrium schemes for scalar conservation laws with stiff sources*, Math. Comp., 72 (2003), pp. 131–157.

[4] F. Bouchut, *Nonlinear Stability of Finite Volume Methods for Hyperbolic Conservation Laws and Well-Balanced Schemes for Sources*, Front. Math., Birkhäuser, Basel, 2004.

[5] M. Briani and R. Natalini, *Asymptotic high-order schemes for integro-differential problems arising in markets with jumps*, Commun. Math. Sci., 4 (2006), pp. 81–96.

[6] G.-Q. Chen, C. D. Levermore, and T.-P. Liu, *Hyperbolic conservation laws with stiff relaxation terms and entropy*, Comm. Pure Appl. Math., 47 (1994), pp. 787–830.

[7] J. M. Greenberg and A. Y. Le Roux, *A well-balanced scheme for the numerical processing of source terms in hyperbolic equations*, SIAM J. Numer. Anal., 33 (1996), pp. 1–16.

[8] B. Hanouzet and R. Natalini, *Weakly coupled systems of quasilinear hyperbolic equations*, Differential Integral Equations, 9 (1996), pp. 1279–1292.

[9] B. Hanouzet and R. Natalini, *Global existence of smooth solutions for partially dissipative hyperbolic systems with a convex entropy*, Arch. Ration. Mech. Anal., 169 (2003), pp. 89–117.

[10] S. Jin, *A steady-state capturing method for hyperbolic systems with geometrical source terms*, M2AN Math. Model. Numer. Anal., 35 (2001), pp. 631–645.

[11] S. Kawashima, *Large-time behaviour of solutions to hyperbolic-parabolic systems of conservation laws and applications*, Proc. Roy. Soc. Edinburgh Sect. A, 106 (1987), pp. 169–194.

[12] R. Natalini, *Recent results on hyperbolic relaxation problems*, in Analysis of Systems of Conservation Laws (Aachen, 1997), Chapman & Hall/CRC, Boca Raton, FL, 1999, pp. 128–198.

[13] B. Perthame and C. Simeoni, *A kinetic scheme for the Saint-Venant system with a source term*, Calcolo, 38 (2001), pp. 201–231.

[14] P. L. Roe, *Upwind differencing schemes for hyperbolic conservation laws with source term*, in Nonlinear Hyperbolic Problems, C. Carasso, P. A. Raviart, and D. Serre, eds., Lecture Notes in Math. 1270, Springer, Berlin, 1987, pp. 41–51.

[15] Y. Shizuta and S. Kawashima, *Systems of equations of hyperbolic-parabolic type with applications to the discrete Boltzmann equation*, Hokkaido Math. J., 14 (1985), pp. 249–275.

[16] F. Trèves, *Basic Linear Partial Differential Equations*, Pure Appl. Math. (Amst.) 62, Academic Press, New York, 1975.

[17] G. B. Whitham, *Linear and Nonlinear Waves*, Wiley & Sons, New York, 1974.

SIAM J. NUMER. ANAL.
Vol. 46, No. 2, pp. 895–919
© 2008 Society for Industrial and Applied Mathematics

A NUMERICAL SCHEME FOR THE PORE-SCALE SIMULATION OF CRYSTAL DISSOLUTION AND PRECIPITATION IN POROUS MEDIA*

V. M. DEVIGNE†‡, I. S. POP§, C. J. VAN DUIJN§, AND T. CLOPEAU†

Abstract. In this paper we analyze a numerical scheme for a dissolution and precipitation model in porous media. We focus here on the chemistry, which is modeled by a parabolic problem that is coupled through the boundary conditions to an ordinary differential inclusion defined on the boundary. We use a regularization approach for constructing a semi-implicit scheme that is stable and convergent. For dealing with the emerging time discrete nonlinear problems, we propose a simple fixed-point iterative procedure. The paper is concluded by numerical results.

Key words. dissolution, precipitation, porous media, convergence, linearization, coupled system, differential inclusion

AMS subject classifications. 65M12, 65N12, 35K57, 76S05

DOI. 10.1137/060673485

1. Introduction.

In this paper we consider a pore-scale model for the dissolution and precipitation of crystals in a porous medium. This model is studied in [7] and represents the pore-scale analogue of the macroscopic (core-scale) model proposed in [17]. The particularity of the model is in the description of the precipitation processes taking place on the surface of the grains Γ_G, involving a multivalued function. Models of similar type are analyzed in a homogenization context in [23, 5, 12, 13, 24].

Without going into details, we briefly recall the background of the model. A fluid in which cations and anions are dissolved occupies the pores of a porous medium. The boundary of the void space consists of two disjoint parts: the surface of the porous skeleton (the grains), named from now on the *internal boundary*, and the external part, which is the outer boundary of the domain. Under certain conditions, the ions transported by the fluid can precipitate and form a crystalline solid, which is attached to the internal boundary and thus is immobile. The reverse reaction of dissolution is also possible.

The model proposed in [7] consists of several components: the Stokes flow in the pores, the transport of dissolved ions by convection and diffusion, and dissolution/precipitation reactions on the surface of the porous skeleton (grains). It is assumed that the flow geometry as well as the fluid properties are not affected by

*Received by the editors October 27, 2006; accepted for publication (in revised form) August 10, 2007; published electronically February 22, 2008.
http://www.siam.org/journals/sinum/46-2/67348.html

†Département de Mathématiques, Institut Camille Jordan, Université Claude Bernard Lyon 1, Bât A, bur. 1304, 50, Av. Tony Garnier, 69367 Lyon Cedex 07, France (Vincent.Devigne@univ-lyon1.fr, Thierry.Clopeau@univ-lyon1.fr). This research was initiated while the first author spent six months at the Technische Universiteit Eindhoven, supported through a European Community Marie Curie Fellowship (contract HPMT-CT-2001-00422). The research of the fourth author was partially supported by the GDR MOMAS (PACEN/CNRS).

‡Current address: Centre de Mise en Forme des matériaux, Ecole des Mines de Paris 1, rue Claude Daunesse BP207, F-06904 Sophia Antipolis Cedex, France (vincent.devigne@ensmp.fr).

§Department of Mathematics and Computer Science, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands (I.Pop@tue.nl, C.J.v.Duijn@tue.nl). The work of these two authors was supported by the Dutch government through the national program BSIK: knowledge and research capacity, in the ICT project BRICKS (http://www.bsik-bricks.nl), theme MSV1.

the chemical processes. Therefore the flow component can be completely decoupled from the remaining part of the model. This situation appears, for example, if the crystal size is negligible when compared to the typical grain size or the pore scale. We mention [25] for a preliminary investigation of the case where dissolution and precipitation lead to changes in the pore space.

Another simplification follows by considering the total electric charge in the fluid. This is a linear combination of the concentrations of the ions and can be determined by solving a linear parabolic problem that depends only on the flow and not on the chemistry. Therefore once the fluid flow is determined, the total charge can be decoupled from the chemistry (see [6] for details).

Our main interest is focused on the chemistry, this being the challenging part of the model. Here we investigate an appropriate numerical scheme for the dissolution and precipitation component of the model in [7]. The present work is closely related to [11]. The numerical scheme proposed there also involves a special treatment of the diffusion on the grain boundary. The coupled system is solved by iterating between the equations in the pores and on the solid matrix.

Numerical methods for the macroscale equations modeling the dissolution and precipitation in porous media are considered in [1, 8, 9, 21, 22]. The time stepping is performed by various first order implicit schemes, and finite elements or finite volumes are employed for the spatial discretization. For the upscaled version of the present model, a numerical algorithm for computing traveling wave solutions is proposed in [17].

We denote the flow domain by $\Omega \subset \mathbb{R}^d$ $(d > 1)$, which is assumed to be open, connected, and bounded. Further, the boundary $\partial\Omega$ is assumed to be Lipschitz continuous. It consists of two disjoint parts: an internal and an external one. The internal boundary Γ_G is the surface of the grains, and the external boundary Γ_D is the outer boundary of the domain. Let $\vec{\nu}$ denote the outer normal to $\partial\Omega$ and $T > 0$ be a fixed but arbitrarily chosen value of time. For X being Ω, Γ_G, or Γ_D, and any $0 < t \leq T$, we define

$$X^t = (0, t] \times X.$$

To simplify the presentation, we assume that the initial data are compatible in the sense of [6, 17]. Essentially this means that initially the system is in equilibrium. Moreover, the boundary data are assumed such that the total charge remains constant in time and space. In this way, and after an appropriate scaling, the model can be reduced to

$$\begin{cases} \partial_t u + \nabla \cdot (\vec{q}u - D\nabla u) = 0 & \text{in } \Omega^T, \\ -D\vec{\nu} \cdot \nabla u = \varepsilon\tilde{n}\partial_t v & \text{on } \Gamma_G^T, \\ u = 0 & \text{on } \Gamma_D^T, \\ u = u_I & \text{in } \Omega, \text{ for } t = 0, \end{cases} \tag{1.1}$$

for the ion transport and

$$\begin{cases} \partial_t v = D_a(r(u) - w) & \text{on } \Gamma_G^T, \\ w \in H(v) & \text{on } \Gamma_G^T, \\ v = v_I & \text{on } \Gamma_G, \text{ for } t = 0, \end{cases} \tag{1.2}$$

for the precipitation and dissolution. Here u denotes the cation concentration and is defined in the entire void space Ω, whereas v stands for the concentration of the

precipitate, which is defined only on the interior boundary Γ_G. These two concentrations, together with w modeling the actual value of the dissolution rate, are the unknown quantities.

In the above $\vec{q}$ is the divergence-free fluid velocity, which is assumed to be known and to have a zero trace along the internal grain boundary Γ_G:

$$\vec{q} \in [H^1(\Omega)]^d, \qquad \nabla \cdot \vec{q} = 0 \text{ in } \Omega, \qquad \vec{q} = \vec{0} \text{ on } \Gamma_G. \tag{1.3}$$

By $H^1(\Omega)$ we mean the space of functions defined on Ω and having L^2 generalized derivatives. To avoid unnecessary technical complications we also assume that the fluid velocity $\vec{q}$ is essentially bounded, $\vec{q} \in [L^\infty(\Omega)]^d$, and define

$$M_q = \|\vec{q}\|_{\infty,\Omega} < \infty. \tag{1.4}$$

For the Stokes model with homogeneous Dirichlet boundary conditions, the essential boundedness of $\vec{q}$ holds if, for example, the domain is polygonal (see [16] or [20]).

The above equations are employing several functions, which are model-specific. Here we assume them to be given. By r we denote the precipitation rate. Assuming mass action kinetics, with $[\cdot]_+$ denoting the nonnegative cut,

$$[u]_+ = \begin{cases} 0 & \text{if } u < 0, \\ u & \text{if } u > 0, \end{cases} \tag{1.5}$$

the precipitation rate is defined by

$$r(u) = [u]_+^{\tilde{m}} [(\tilde{m}u - c)/\tilde{n}]_+^{\tilde{n}}, \tag{1.6}$$

where $\tilde{m}$ and $\tilde{n}$ are the valences of the cation and the anion, respectively. As mentioned before, the total negative charge c is assumed to be known and constant in time and space. In a generalized framework, r satisfies the following:

(A_r) (i) $r : \mathbb{R} \to [0, \infty)$ is locally Lipschitz in $\mathbb{R}$;
(ii) there exists a unique $u_* \geq 0$, such that

$$r(u) = \begin{cases} 0 \qquad \text{for} \quad u \leq u_*, \\ \text{strictly increasing for } u > u_* \text{ with } r(\infty) = \infty. \end{cases}$$

Remark 1.1. In this setting, a unique u^* exists for which $r(u^*) = 1$. If $u = u^*$ for all t and x, then the system is in equilibrium: No precipitation or dissolution occurs, since the precipitation rate is balanced by the dissolution rate regardless of the presence or absence of crystals.

In (1.2_2), H stands for the Heaviside graph,

$$H(v) = \begin{cases} \{0\} & \text{if } v < 0, \\ [0,1] & \text{if } v = 0, \\ \{1\} & \text{if } v > 0. \end{cases}$$

This implies that the dissolution rate is constant (scaled to 1) in the presence of crystals, i.e., for $v > 0$ somewhere on Γ_G. In the absence of crystals ($v = 0$), the overall rate is either zero, if $r(u) \leq 1$, or positive. In the first case we have $u < u^*$, and we speak about an undersaturated fluid since there are not sufficient ions for an effective

gain in the precipitate. This second situation is appearing in the oversaturated regime, when $u > u^*$. Then we take $w = 1$, and the overall rate is $r(u) - 1 > 0$. Following the detailed discussion in [6], this can be summarized as

$$w = \begin{cases} 0 & \text{if } v < 0, \\ \min\{r(u), 1\} & \text{if } v = 0, \\ 1 & \text{if } v > 0. \end{cases} \tag{1.7}$$

Finally, D in (1.1)–(1.2) denotes the diffusion coefficient. D_a represents the ratio of the characteristic precipitation/dissolution time scale and the characteristic transport time scale—the Damköhler number. Both D and D_a are assumed to be of moderate order $O(1)$. By ε we mean the ratio of the characteristic pore scale and the reference (macroscopic) length scale. In an appropriate scaling (see Remark 1.2 of [7]), this gives

$$\varepsilon \operatorname{meas}(\Gamma_G) \approx \operatorname{meas}(\Omega). \tag{1.8}$$

This balance is natural for a porous medium, where $\operatorname{meas}(\Gamma_G)$ denotes the total surface of the porous skeleton and $\operatorname{meas}(\Omega)$ the total void volume. Throughout this paper we keep the value of ε fixed. However, the convergence results are uniform with respect to ε and thus also for arbitrarily small values of ε.

The present setting is a simplification of the dissolution/precipitation model in [7]. However, the main difficulties that are associated with that model are still present here: The system (1.1)–(1.2) consists of a parabolic equation that is coupled through the boundary conditions to a differential inclusion defined on a lower-dimensional manifold. For ease of presentation we have considered here only the case of homogeneous Dirichlet boundary conditions on the external boundary Γ_D, but the results can be extended to more general cases. With $H^1_{0,\Gamma_D}(\Omega)$ being the subspace of $H^1(\Omega)$ functions having a zero trace on the external boundary Γ_D, for the initial data we assume the following:

(A_D) The initial data are assumed essentially bounded and nonnegative. Further, $u_I \in H^1_{0,\Gamma_D}(\Omega)$ and $v_I \in L^2(\Gamma_G)$.

Due to the occurrence of the multivalued dissolution rate, classical solutions do not exist, except for some particular cases. For defining a weak solution we consider the following sets:

$$\begin{aligned} \mathcal{U} &:= \{u \in L^2(0,T; H^1_{0,\Gamma_D}(\Omega)) : \partial_t u \in L^2(0,T; H^{-1}(\Omega))\}, \\ \mathcal{V} &:= \{v \in H^1(0,T; L^2(\Gamma_G))\}, \\ \mathcal{W} &:= \{w \in L^\infty(\Gamma_G^T), : 0 \le w \le 1\}, \end{aligned}$$

where by $H^{-1}(\Omega)$ we mean the dual of $H^1_{0,\Gamma_D}(\Omega)$. Here we have used standard notations in the functional analysis. In what follows $(\cdot,\cdot)_X$ stands for the usual scalar product in a Hilbert space X, or the duality pairing between H^{-1} and H^1. With $t \in (0,T]$, by $(\cdot,\cdot)_{X^t}$ we mean the integration in time of $(\cdot,\cdot)_X$ on $(0,t)$.

DEFINITION 1.1. *A triple $(u,v,w) \in \mathcal{U} \times \mathcal{V} \times \mathcal{W}$ is called a weak solution of* (1.1) *and* (1.2) *if $\{u(0), v(0)\} = \{u_I, v_I\}$ and if*

$$(\partial_t u, \varphi)_{\Omega^T} + D(\nabla u, \nabla \varphi)_{\Omega^T} - (\vec{q} u, \nabla \varphi)_{\Omega^T} = -\varepsilon \tilde{n} (\partial_t v, \varphi)_{\Gamma_G^T}, \tag{1.9}$$

$$(\partial_t v, \theta)_{\Gamma_G^T} = D_a (r(u) - w, \theta)_{\Gamma_G^T}, \tag{1.10}$$

$$w \in H(v) \quad \text{a.e. in } \Gamma_G^T$$

for all $(\varphi,\theta)\in L^2(0,T;H^1_{0,\Gamma_D}(\Omega))\times L^2(\Gamma_G^T)$.

In the above definition the initial condition should be understood in the sense of L^2 functions. This makes sense since $\mathcal{U}\subset C([0,T];L^2(\Omega))$ and $\mathcal{V}\subset C([0,T];L^2(\Gamma_G))$ (see, for example, [31] or [36]). The existence of a weak solution is proven in [7, Theorem 2.21]. Moreover, with

$$M_u := \max\{\|u_I\|_{\infty,\Omega}, u^*\}, \tag{1.11}$$

$$M_v := \max\{\|v_I\|_{\infty,\Omega}, 1\}, \quad C_v := \frac{r(M_u)D_a}{M_v}, \tag{1.12}$$

a weak solution satisfies

$$\begin{aligned} 0\le u\le M_u &\quad \text{a.e. in } \Omega^T,\\ 0\le v(t,\cdot)\le M_v e^{C_v t} &\quad \text{for all } t\in[0,T] \text{ and a.e. on } \Gamma_G,\\ 0\le w\le 1 &\quad \text{a.e. on } \Gamma_G^T, \end{aligned} \tag{1.13}$$

and

$$\begin{aligned} \|u(t)\|_\Omega^2 + \|\nabla u\|_{\Omega^T}^2 + \|\partial_t u\|^2_{L^2(0,T;H^{-1}(\Omega))}&\\ + \varepsilon\|v(t)\|^2_{\Gamma_G} + \varepsilon\|\partial_t v\|^2_{\Gamma_G^T} \le C& \end{aligned} \tag{1.14}$$

for all $0\le t\le T$. Here $C>0$ is a constant not depending on u, v, w, or ε. The proof is based on regularization arguments and provides a solution for which, in addition, we have

$$w = r(u) \qquad \text{a.e. in } \{v=0\}\cap\Gamma_G^T.$$

This is in good agreement with the definition in (1.7).

Finally we mention that (1.1) and (1.2) have a unique weak solution, as proven in [26].

2. The time discrete numerical scheme. In this section we analyze a semi-implicit numerical scheme for the system (1.1)–(1.2). To overcome the difficulties that are due to the multivalued dissolution rate, we approximate the Heaviside graph by

$$H_\delta(v) := \begin{cases} 0 & \text{if } v\le 0,\\ v/\delta & \text{if } v\in(0,\delta),\\ 1 & \text{if } v\ge\delta, \end{cases} \tag{2.1}$$

where $\delta>0$ is a small regularization parameter.

Next we consider a time stepping that is implicit in u and explicit in v. Though possible here as well, an implicit discretization of v would involve an additional nonlinearity in v without bringing any significant improvement of the results.

With $N\in\mathbb{N}$, $\tau = T/N$, and $t_n = n\tau$ $(n=0,\dots,N)$, the approximation pair (u^n,v^n) of $(u(t_n),v(t_n))$ is the solution of the following problem.

Problem P^n_δ. Given u^{n-1} and v^{n-1} compute $u^n\in H^1_{0,\Gamma_D}(\Omega)$ and $v^n\in L^2(\Gamma_G)$ such that

$$\begin{aligned} (u^n-u^{n-1},\phi)_\Omega + \tau D(\nabla u^n,\nabla\phi)_\Omega - \tau(\vec{q}u^n,\nabla\phi)_\Omega&\\ + \epsilon\tilde{n}(v^n-v^{n-1},\phi)_{\Gamma_G} = 0,& \end{aligned} \tag{2.2}$$

$$(v^n,\theta)_{\Gamma_G} = (v^{n-1},\theta)_{\Gamma_G} + \tau D_a(r(u^n) - H_\delta(v^{n-1}),\theta)_{\Gamma_G} \tag{2.3}$$

for all $\phi \in H^1_{0,\Gamma_D}(\Omega)$ and $\theta \in L^2(\Gamma_G)$.

Here $n = 1, \ldots, N$, while $u^0 = u_I$ and $v^0 = v_I$. For consistency with the original setting, in (2.3) we approximate the dissolution rate $w(t_n)$ by

$$w^n := H_\delta(v^n). \tag{2.4}$$

To simplify the notations, we have given up the subscript δ for the solution triple (u^n, v^n, w^n).

Remark 2.1. As we will see later, for guaranteeing the stability of the scheme the regularization parameter δ should be chosen such that $\delta \geq \tau D_a$. This is the only restriction that is related to the explicit discretization of v. Further, the convergence result is obtained under the assumption that the ratio τ/δ approaches 0 as $\tau \searrow 0$. This happens, for example, if $\delta = O(\tau^\alpha)$, with some $\alpha \in (0, 1)$, which is consistent with the previous restriction. To simplify the presentation, from now on we assume $\delta = D_a\sqrt{\tau}$. Then the stability condition $\delta \geq \tau D_a$ is ensured whenever $\tau \leq 1$, which is a mild restriction.

Due to the explicit discretization of v, the ion transport equation in Problem P^n_δ can be decoupled from the time discrete precipitation/dissolution equation. Replacing the Γ_G-scalar product in (2.2) by the last term in (2.3), we end up with an elliptic problem having a nonlinear boundary condition on Γ_G. Specifically, given u^{n-1} and v^{n-1}, we seek for $u^n \in H^1_{0,\Gamma_D}(\Omega)$ such that

$$\begin{aligned}(u^n - u^{n-1}, \phi)_\Omega + \tau D(\nabla u^n, \nabla\phi)_\Omega - \tau(\vec{q}u^n, \nabla\phi)_\Omega \\ + \tau\epsilon\tilde{n}D_a(r(u^n) - H_\delta(v^{n-1}), \theta)_{\Gamma_G} = 0\end{aligned}$$

for all $\phi \in H^1_{0,\Gamma_D}(\Omega)$. Since r is monotone and Lipschitz, standard monotonicity methods provide the existence and uniqueness for the above problem (see, e.g., [34, Chapter 10]). Alternatively, a contraction argument is described in section 3. This writing is used there as the starting point for constructing a linear iterative scheme for solving Problem P^n_δ. The fixed-point approach proving the convergence of the iteration procedure discussed in section 3 can also be employed for the existence and uniqueness of a u^n solving the nonlinear problem above. Having u^n, v^n can be determined straightforwardly from (2.3). In this way we obtain the following.

LEMMA 2.2. *Problem P^n_δ has a unique solution pair (u^n, v^n).*

2.1. Stability in L^∞. All of the estimates in this section should be interpreted in the a.e. sense. As follows from (1.13), the concentrations u and v as well as the dissolution rate w are nonnegative and bounded. Here we prove similar results for the time discrete concentrations u^n and v^n. The bounds for w^n follow straightforwardly from (2.4).

LEMMA 2.3. *Assume $\tau \leq 1$ and that u^{n-1} and v^{n-1} are nonnegative. Then u^n and v^n are nonnegative as well.*

Proof. We start with the estimate in v^n. With $[\cdot]_-$ denoting the nonpositive cut (see also (1.5)), we test (2.3) with $\theta := [v^n]_-$ and obtain

$$\|[v^n]_-\|^2_{\Gamma_G} = \tau D_a\left(r(u^n), [v^n]_-\right)_{\Gamma_G} + \left(v^{n-1} - \tau D_a H_\delta(v^{n-1}), [v^n]_-\right)_{\Gamma_G}.$$

In view of (A_r), the first term on the right is nonpositive. Further, since $v^{n-1} \geq 0$ and $\delta = D_a\sqrt{\tau}$, by the construction of H_δ we have

$$v^{n-1} - \tau D_a H_\delta(v^{n-1}) \geq v^{n-1}(1 - \tau D_a/\delta) \geq 0$$

a.e. on Γ_G. Hence the second term on the right is nonpositive as well. This yields

$$\|[v^n]_-\|^2_{\Gamma_G} \le 0,$$

implying the assertion for v^n.

For proving that u^n is nonnegative we proceed in a similar manner. Testing (2.2) with $\phi := [u^n]_-$ gives

$$\|[u^n]_-\|^2_\Omega + \tau D\,\|\nabla[u^n]_-\|^2_\Omega - \tau\,(\vec{q}u^n, \nabla[u^n]_-)_\Omega + \epsilon\tilde{n}\left(v^n - v^{n-1}, [u^n]_-\right)_{\Gamma_G} = (u^{n-1}, [u^n]_-)_\Omega. \tag{2.5}$$

The first two terms in the above are nonnegative, whereas the third one vanishes. This follows from

$$\begin{aligned}(\vec{q}u^n, \nabla[u^n]_-)_\Omega &= \tfrac{1}{2}(\vec{q}, \nabla[u^n]^2_-)_\Omega \\ &= \tfrac{1}{2}(\vec{\nu}\cdot\vec{q}, [u^n]^2_-)_{\Gamma_D\cup\Gamma_G} - \tfrac{1}{2}(\nabla\cdot\vec{q}, [u^n]^2_-)_\Omega\end{aligned}$$

and the boundary conditions on $\partial\Omega$, since $\nabla\cdot\vec{q} = 0$ in Ω.

Further, since $[u^n]_- \le 0$ a.e. and it belongs to $H^1_{0,\Gamma_D}(\Omega)$, its trace $[u^n]_-|_{\Gamma_G}$ is a nonpositive $L^2(\Gamma_G)$ function. Testing (2.3) with $[u^n]_-|_{\Gamma_G}$ gives

$$\begin{aligned}(v^n - v^{n-1}, [u^n]_-)_{\Gamma_G} &= \tau D_a(r(u^n) - H_\delta(v^{n-1}), [u^n]_-)_{\Gamma_G} \\ &= -\tau D_a(H_\delta(v^{n-1}), [u^n]_-)_{\Gamma_G} \ge 0,\end{aligned}$$

where we have used (A_r) and the positivity of H_δ.

Finally, the term on the right in (2.5) is nonpositive, since $u^{n-1} \ge 0$. In this way we obtain $[u^n]_- = 0$ a.e. in Ω, implying the result. ∎

Now we turn our attention to the upper bounds for u^n and v^n. First, with M_u defined in (1.11) we have the following.

LEMMA 2.4. *If $u^{n-1} \le M_u$, then the same holds for u^n.*

Proof. We test (2.2) with $\phi := [u^n - M_u]_+$, the nonnegative part of $u^n - M_u$. This gives

$$\begin{aligned}&\|[u^n - M_u]_+\|^2_\Omega + \tau D\|\nabla[u^n - M_u]_+\|^2_\Omega - \tau(\vec{q}u^n, \nabla[u^n - M_u]_+)_\Omega \\ &\quad = (u^{n-1} - M_u, [u^n - M_u]_+)_\Omega - \epsilon\tilde{n}(v^n - v^{n-1}, [u^n - M_u]_+)_{\Gamma_G}.\end{aligned}$$

Arguing as in the proof of Lemma 2.3, we first observe that the convection term vanishes. Further, since $u^{n-1} \le M_u$, the first term on the right is nonpositive. Finally, for the last term we have

$$(v^n - v^{n-1}, [u^n - M_u]_+)_{\Gamma_G} = \tau D_a(r(u^n) - H_\delta(v^{n-1}), [u^n - M_u]_+)_{\Gamma_G}.$$

By the definition of M_u, whenever $u^n \ge M_u$ we have $r(u^n) \ge 1 \ge H_\delta(v^{n-1})$. This implies the positivity of the above scalar product. We are therefore left with

$$\|[u^n - M_u]_+\|^2_\Omega + \tau D\|\nabla[u^n - M_u]_+\|^2_\Omega \le 0,$$

implying that $u^n \le M_u$. ∎

Remark 2.5. Since for the initial data we assume $0 \le u_I \le M_u$ and $0 \le v_I$, Lemmas 2.3 and 2.4 show that $0 \le u^n \le M_u$ and $0 \le v^n$ for all $n = 0, \dots, N$.

Upper estimates for v^n can be obtained as for u^n.

LEMMA 2.6. *With M_v and C_v defined in* (1.12), *assume that* $v^{n-1} \le M_v e^{C_v(n-1)\tau}$. *Then* $v^n \le M_v e^{C_v n\tau}$.

Proof. Testing (2.3) with $\theta := [v^n - M_v e^{C_v n\tau}]_+$ gives

$$\begin{aligned}(2.6)\qquad \|[v^n - M_v e^{C_v n\tau}]_+\|^2_{\Gamma_G} &= (v^{n-1} - M_v e^{C_v(n-1)\tau}, [v^n - M_v e^{C_v n\tau}]_+)_{\Gamma_G} \\ &\quad + M_v(e^{C_v(n-1)\tau} - e^{C_v n\tau}, [v^n - M_v e^{C_v n\tau}]_+)_{\Gamma_G} \\ &\quad + \tau D_a(r(u^n) - H_\delta(v^{n-1}), [v^n - M_v e^{C_v n\tau}]_+)_{\Gamma_G}.\end{aligned}$$

We denote the terms on the right by I_1, I_2, and I_3. We first notice that the assumption on v^{n-1} implies $I_1 \le 0$. Further, since $0 \le u^n \le M_u$, $H_\delta(v^{n-1}) \ge 0$, and due to the monotonicity of r we obtain

$$I_3 \le \tau D_a(r(M_u), [v^n - M_v e^{C_v n\tau}]_+)_{\Gamma_G}.$$

Recalling (1.12), this gives

$$I_2 + I_3 \le M_v(\tau C_v + e^{C_v(n-1)\tau}(1 - e^{\tau C_v}), [v^n - M_v e^{Cn\tau}]_+)_{\Gamma_G} \le 0.$$

Here we have used the elementary inequality $e^x \ge 1 + x$, as well as $e^{C_v(n-1)\tau} \ge 1$. In this way (2.6) becomes

$$\|[v^n - M_v e^{C_v n\tau}]_+\|^2_{\Gamma_G} \le 0,$$

implying the upper bounds for v^n. ∎

Remark 2.7. As before, since $v_I \le M_v$, it follows that $v^n \le M_v e^{Cn\tau}$ for all $n = 1, \dots, N$.

Remark 2.8. The essential bounds provided by Lemmas 2.3, 2.4, and 2.6 are uniform in δ, whereas τ appears only in the upper bounds for v^n. But for any $0 \le n \le N$ one has $n\tau = t_n \le T$, and we have $v^n \le M_v e^{C_v t_n} \le M_v e^{C_v T}$, which is τ-independent as well.

2.2. A priori estimates. We continue the analysis of the numerical scheme (2.2)–(2.3) by giving some energy estimates for the sequence of time discrete concentrations $\{(u^n, v^n),\ n = 0, \dots, N\}$. We start with the estimates in v, which are depending on meas(Γ_G).

LEMMA 2.9. *For any $n \ge 1$ we have*

$$(2.7)\qquad \left\|v^n - v^{n-1}\right\|_{\Gamma_G} \le \tau D_a\ r(M_u)\ meas(\Gamma_G)^{1/2} \qquad \text{and}$$

$$(2.8)\qquad \|v^n\|_{\Gamma_G} \le \|v_I\|_{\Gamma_G} + n\tau D_a\ r(M_u)\ meas(\Gamma_G)^{1/2}.$$

Proof. Testing (2.3) with $\theta := [v^n - v^{n-1}] \in L^2(\Gamma_G)$ and applying Cauchy's inequality gives

$$(2.9)\qquad \|v^n - v^{n-1}\|^2_{\Gamma_G} \le \tau D_a \|r(u^n) - H_\delta(v^{n-1})\|_{\Gamma_G} \|v^n - v^{n-1}\|_{\Gamma_G}.$$

The essential bounds on u^n and v^n, together with the assumptions on r and H_δ, imply $-1 \le r(u^n) - H_\delta(v^{n-1}) \le r(M_u)$. By (1.11) we have $r(M_u) \ge 1$, implying the first estimates.

In a similar manner, by taking $\theta := v^n \in L^2(\Gamma_G)$ in (2.3) and applying the Cauchy inequality we obtain

$$(2.10)\qquad \|v^n\|^2_{\Gamma_G} \le \|v^{n-1}\|_{\Gamma_G} \|v^n\|_{\Gamma_G} + \tau D_a\ r(M_u)\ \text{meas}(\Gamma_G)^{1/2} \|v^n\|_{\Gamma_G}.$$

By dividing by $\|v^n\|_{\Gamma_G}$ and applying the inequality backward we immediately obtain the estimate (2.8). □

Remark 2.10. Notice that the estimates in Lemma 2.9 are δ-independent. Next, for $n \leq N$, the term on the right in (2.8) is bounded uniformly in τ as well. Finally, due to (1.8) we can replace meas(Γ_G) by C/ϵ, where C does not depend on δ, τ, or ϵ.

Now we proceed by the estimates for u.

LEMMA 2.11. *Assume $\tau \leq \tau_0$, with $\tau_0 > 0$ a fixed value that will be given below. For the time discrete solute concentrations we have*

$$\tau \sum_{n=1}^{N} \|\nabla u^n\|_\Omega^2 \leq C, \tag{2.11}$$

$$\sum_{n=1}^{N} \|u^n - u^{n-1}\|_\Omega^2 \leq C\sqrt{\tau}, \tag{2.12}$$

$$\tau \sum_{n=1}^{N} \|\nabla(u^n - u^{n-1})\|_\Omega^2 \leq C\sqrt{\tau}, \tag{2.13}$$

$$\sum_{n=1}^{N} \|u^n - u^{n-1}\|_{\Gamma_G}^2 \leq C. \tag{2.14}$$

Here C is a constant that does not depend on δ and τ.

Proof. We start by testing (2.2) with $\phi = u^n$. This gives

$$\begin{aligned}(u^n - u^{n-1}, u^n)_\Omega + \tau D\|\nabla u^n\|_\Omega^2 \\ + \tau(\vec{q}u^n, \nabla u^n)_\Omega + \varepsilon\tilde{n}(v^n - v^{n-1}, u^n)_{\Gamma_G} = 0.\end{aligned} \tag{2.15}$$

We denote the terms on the right by $T_1, \ldots, T_4$. We have

$$T_1 = \frac{1}{2}(\|u^n\|_\Omega^2 - \|u^{n-1}\|_\Omega^2 + \|u^n - u^{n-1}\|_\Omega^2).$$

Furthermore, T_2 is nonnegative, and T_3 vanishes as argued in the proof of Lemma 2.3.

Before estimating the last term we notice the existence of positive constants C_1 and C_2 such that

$$\|\varphi\|_{\Gamma_G}^2 \leq C_1\|\varphi\|_\Omega^2 + C_2\|\varphi\|_\Omega\|\nabla\varphi\|_\Omega \tag{2.16}$$

for all $\varphi \in H^1_{0,\Gamma_D}(\Omega)$. This can be obtained by following the proof of the trace theorem (e.g., see [10, Theorem 1.5.1.10]). By the inequality of means

$$ab \leq \frac{1}{4\rho}a^2 + \rho b^2 \qquad \text{for all } a, b, \text{ and } \rho > 0, \tag{2.17}$$

we get

$$\|\varphi\|_{\Gamma_G}^2 \leq \left(C_1 + \frac{C_2^2}{4\rho}\right)\|\varphi\|_\Omega^2 + \rho\|\nabla\varphi\|_\Omega^2. \tag{2.18}$$

Here $\rho > 0$ can be taken arbitrarily small.

With $M := \varepsilon\tilde{n} D_a \; r(M_u) \; \mathrm{meas}(\Gamma_G)^{1/2}$, we use (2.7) to estimate T_4:

$$|T_4| \leq \tau M \|u^n\|_{\Gamma_G} \leq \frac{\tau M^2}{4} + \tau \|u^n\|^2_{\Gamma_G}.$$

Now we use (2.18) with $\rho = D/2$ and obtain

$$|T_4| \leq \frac{\tau M^2}{4} + \tau \left(C_1 + \frac{C_2^2}{2D} \right) \|u^n\|^2_\Omega + \frac{\tau D}{2} \|\nabla u^n\|^2_\Omega.$$

Using the estimates above into (2.15), summing for $n = 1, \ldots, N$, and multiplying the result by 2 yields

$$\|u^N\|^2_\Omega + \sum_{n=1}^{N} \|u^n - u^{n-1}\|^2_\Omega + \tau D \sum_{n=1}^{N} \|\nabla u^n\|^2_\Omega \leq \|u_I\|^2_\Omega + \frac{TM^2}{2} + C,$$

where $C = T\big(C_1 + C_2^2/(2D)\big) M_u^2 \; \mathrm{meas}(\Omega)$. This estimate follows by the essential bounds on u^n. An alternative proof can be given based on the discrete Gronwall lemma. In this way we have proven (2.11). Notice that we have also obtained

$$\sum_{n=1}^{N} \|u^n - u^{n-1}\|^2_\Omega \leq C,$$

which is not as good as (2.12).

To proceed with (2.12) and (2.13) we notice that, since $u_I \in H^1_{0,\Gamma_D}$, $u^n - u^{n-1}$ is a H^1_{0,Γ_D} function for all $n \geq 1$. Testing (2.2) with $u^n - u^{n-1}$ gives

$$\begin{aligned} (2.19) \qquad & \|u^n - u^{n-1}\|^2_\Omega + \tau D (\nabla u^n, \nabla (u^n - u^{n-1}))_\Omega \\ & - \tau (\vec{q} u^n, \nabla (u^n - u^{n-1}))_\Omega + \epsilon \tilde{n} \left(v^n - v^{n-1}, u^n - u^{n-1} \right)_\Omega = 0. \end{aligned}$$

Denoting the terms in the above by $I_1, \ldots, I_4$, we have

$$I_2 = \frac{\tau D}{2} (\|\nabla u^n\|^2_\Omega - \|\nabla u^{n-1}\|^2_\Omega + \|\nabla (u^n - u^{n-1})\|^2_\Omega).$$

Recalling (1.4), the inequality of means gives

$$\begin{aligned} |I_3| &= \tau (\nabla \cdot (\vec{q} u^n), u^n - u^{n-1})_\Omega \leq \tau M_q \|\nabla u^n\|_\Omega \|u^n - u^{n-1}\|_\Omega \\ &\leq \frac{1}{2} \|u^n - u^{n-1}\|^2_\Omega + \frac{\tau^2 M_q^2}{2} \|\nabla u^n\|^2_\Omega. \end{aligned}$$

With M defined above, for I_4 we use the estimate (2.7) and obtain

$$|I_4| \leq \varepsilon \tilde{n} \|v^n - v^{n-1}\|_{\Gamma_G} \|u^n - u^{n-1}\|_{\Gamma_G} \leq \frac{(\tau M)^2}{4\mu_1} + \mu_1 \|u^n - u^{n-1}\|^2_{\Gamma_G},$$

where $\mu_1 > 0$ will be chosen below. Now we can take $\rho = \tau D/(4\mu_1)$ into (2.18) to obtain

$$|I_4| \leq \frac{(\tau M)^2}{4\mu_1} + \mu_1 \left(C_1 + \frac{\mu_1 C_2^2}{\tau D} \right) \|u^n - u^{n-1}\|^2_\Omega + \frac{\tau D}{4} \|\nabla (u^n - u^{n-1})\|^2_\Omega.$$

Using the above estimates into (2.19), taking $\mu_1 = \sqrt{\tau D}/(2C_2)$, multiplying the result by 2, and summing up for $n = 1, \dots, N$, (2.11) gives

$$\left(\frac{1}{2} - \frac{C_1\sqrt{\tau D}}{C_2}\right) \sum_{n=1}^{N} \|u^n - u^{n-1}\|_\Omega^2 + \frac{\tau D}{2} \sum_{n=1}^{N} \|\nabla(u^n - u^{n-1})\|_\Omega^2$$
$$\leq \tau D \|\nabla u_I\|_\Omega^2 + \frac{TM^2C_2}{\sqrt{D}}\tau^{1/2} + \tau C.$$

For $\tau_0 := C_2^2/(16C_1^2D)$ and $\tau \leq \tau_0$, since $u_I \in H^1_{0,\Gamma_D}$, the inequality above immediately implies (2.12) and (2.13).

Finally, using (2.18) with $\rho = \tau^{1/2}$, (2.14) is a direct consequence of (2.12) and (2.13). ∎

Remark 2.12. As in Remark 2.10, if the medium is ϵ-periodic, the constant C in Lemma 2.11 does not depend on ϵ. To see this we recall (1.8), and Lemma 3 of [12], saying that there exists a constant $C > 0$, independent of ε, such that

$$\varepsilon \|\varphi\|_{\Gamma_G}^2 \leq C\left(\|\varphi\|_\Omega^2 + \varepsilon^2 \|\nabla\varphi\|_\Omega^2\right)$$

for all $\varphi \in H^1(\Omega)$. Then the boundary term in (2.19) yields a constant C which does not depend on ε as well.

2.3. Convergence. In this part we proceed by proving the convergence of the numerical scheme defined in (2.2)–(2.3) to a weak solution of the system (1.1)–(1.2), as given in Definition 1.1. The multivalued dissolution rate hinders us in obtaining useful error estimates. Therefore convergence will be achieved by compactness arguments. In doing so we mainly follow the ideas in [7].

By considering the sequence of time discrete triples $\{(u^n, v^n, w^n), n = 1, \dots, N\}$ solving Problem P_δ^n, where w^n is defined in (2.4), we construct an approximation of the solution of (1.1) and (1.2) for all times $t \in [0, T]$. Specifically, define for any $n = 1, \dots, N$ and $t \in (t_{n-1}, t_n]$

$$Z^\tau(t) := z^n \frac{(t - t_{n-1})}{\tau} + z^{n-1} \frac{(t_n - t)}{\tau}, \tag{2.20}$$

where (z, Z) stands for (u, U) or (v, V), and define

$$W^\tau(t) := H_\delta(V^\tau(t)). \tag{2.21}$$

Notice that U_τ, V_τ, and W_τ depend not only on τ but also on the regularization parameter δ.

For the time continuous triple $\{U^\tau, V^\tau, W^\tau)\}$ we can use the uniform L^∞ bounds, as well as the a priori estimates in Lemmas 2.9 and 2.11, to derive δ-independent estimates that are similar to those for the solution defined in Definition 1.1 (see [7]).

LEMMA 2.13. *Assume $\delta \geq \tau D_a$. Then for (U^τ, V^τ, W^τ) we have*

$$0 \leq U^\tau \leq M_u \quad \textit{a.e. in } \Omega^T, \tag{2.22}$$

$$0 \leq V^\tau \leq M_v e^{CT},\ 0 \leq W^\tau \leq 1 \quad \textit{a.e. in } \Gamma_G^T, \tag{2.23}$$

$$\|U^\tau(t)\|_\Omega + \|V^\tau(t)\|_{\Gamma_G}^2 \leq C \quad \textit{for all } 0 \leq t \leq T, \tag{2.24}$$

$$\|\partial_t U^\tau\|^2_{L^2(0,T;H^{-1}(\Omega))} + \|\nabla U^\tau\|^2_{\Omega^T} + \|\partial_t V^\tau\|^2_{\Gamma_G^T} \leq C. \tag{2.25}$$

Here $C > 0$ is a constant that does not depend on τ or δ.

Proof. The essential bounds in (2.22) and (2.23) are a direct consequence of Lemmas 2.3, 2.4, and 2.6 and (2.4). The same holds for (2.24), for which we employ the stability estimates in Lemmas 2.9 and 2.11.

To proceed with the gradient estimates in (2.25) we notice that

$$\begin{aligned}\int_0^T \|\nabla U^\tau(t)\|_\Omega^2 dt &\le 2\sum_{n=1}^N \tau \|\nabla u^{n-1}\|_\Omega^2 \\ &\quad + \int_{t_{n-1}}^{t_n} 2\frac{(t-t_{n-1})^2}{\tau^2}\|\nabla(u^n - u^{n-1})\|_\Omega^2 dt \\ &\le 2\tau \sum_{n=1}^N \|\nabla u^{n-1}\|_\Omega^2 + \frac{2\tau}{3}\sum_{n=1}^N \|\nabla(u^n - u^{n-1})\|_\Omega^2 \le C.\end{aligned}$$

Here we have used the estimate in (2.11) and (2.13).

The estimate on $\partial_t V^\tau$ follows by (2.7):

$$\int_0^T \|\partial_t V^\tau(t)\|_{\Gamma_G}^2 dt = \sum_{n=1}^N \int_{t_{n-1}}^{t_n} \left\| \frac{v^n - v^{n-1}}{\tau} \right\|_{\Gamma_G}^2 dt \le CN\tau = CT.$$

Finally, for estimating $\partial_t U^\tau$ we notice that

$$(\partial_t U^\tau(t), \phi) = \left(\frac{u^n - u^{n-1}}{\tau}, \phi \right)$$

for all $\phi \in H^1_{0,\Gamma_D}(\Omega)$ and all $t \in (t_{n-1}, t_n]$. By (2.2) this gives

$$\begin{aligned}|(\partial_t U^\tau, \phi)| &\le D\ \|\nabla u^n\|_\Omega\ \|\nabla\phi\|_\Omega + M_q\ \|\nabla u^n\|_\Omega\ \|\phi\|_\Omega \\ &\quad + \frac{\epsilon\tilde{n}}{\tau}\left\|v^n - v^{n-1}\right\|_{\Gamma_G}\ \|\phi\|_{\Gamma_G}.\end{aligned}$$

By using the trace theorem we obtain

$$\begin{aligned}|(\partial_t U^\tau, \phi)| &\le (D + M_q)\|\nabla u^n\|_\Omega \|\phi\|_{H^1(\Omega)} \\ &\quad + C(\Omega)\frac{\epsilon\tilde{n}}{\tau}\|v^n - v^{n-1}\|_{\Gamma_G}\|\phi\|_{H^1(\Omega)}\end{aligned} \tag{2.26}$$

for any $\phi \in H^1_{0,\Gamma_D}(\Omega)$. This implies that

$$\|\partial_t U^\tau(t)\|_{H^{-1}(\Omega)} \le C\left(\|\nabla u^n\|_\Omega + \frac{\epsilon\tilde{n}}{\tau}\|v^n - v^{n-1}\|_{\Gamma_G} \right) \tag{2.27}$$

for all $t \in (t_{n-1}, t_n]$, and the remaining part is a direct consequence of the L^∞ and stability estimates. ∎

Remark 2.14. *For an ϵ-periodic medium, the estimates in Lemma* 2.13 *can be made ϵ-independent. This follows from Remarks* 2.10 *and* 2.12. *In this case the estimates* (2.24) *and* (2.25) *become*

$$\begin{aligned}&\|U^\tau(t)\|_\Omega + \|\partial_t U^\tau\|^2_{L^2(0,T;H^{-1}(\Omega))} + \|\nabla U^\tau\|^2_{\Omega^T} \\ &\quad + \epsilon\left(\|V^\tau(t)\|^2_{\Gamma_G} + \|\partial_t V^\tau\|^2_{\Gamma_G^T} \right) \le C\end{aligned}$$

for all $0 \le t \le T$*, where* C *does not depend on* τ*,* δ*, or* ε*.*

Having the τ and δ uniform estimates in Lemma 2.13, we can proceed by sending τ and δ to 0. For any $\tau > 0$ and $\delta = \sqrt{\tau} D_a \ge \tau D_a$ we have $(U^\tau, V^\tau, W^\tau) \in \mathcal{U} \times \mathcal{V} \times L^\infty(\Gamma_G^T)$. Obviously $\tau \searrow 0$ implies the same for δ, as well as for the ratio τ/δ. Compactness arguments give the existence of a triple $(u, v, w) \in \mathcal{U} \times \mathcal{V} \times L^\infty(\Gamma_G^T)$ and a subsequence $\tau \searrow 0$ such that

(a) $U^\tau \to u$ weakly in $L^2((0,T); H^1_{0,\Gamma_D}(\Omega))$,
(b) $\partial_t U^\tau \to \partial_t u$ weakly in $L^2((0,T); H^{-1}(\Omega))$,
(c) $V^\tau \to v$ weakly in $L^2((0,T); L^2(\Gamma_G))$,
(d) $\partial_t V^\tau \to \partial_t v$ weakly in $L^2((0,T); L^2(\Gamma_G))$,
(e) $W^\tau \to w$ weak star in $L^\infty(\Gamma_G^T)$.

It remains to show that the limit triple (u, v, w) solves (1.1)–(1.2) weakly. By the uniqueness of the solution, this would actually imply that, along any sequence $\tau \searrow 0$, the approximating triple (U^τ, V^τ, W^τ) converges to (u, v, w). This result is proven in the following.

THEOREM 2.15. *The limit triple* (u, v, w) *is the weak solution of* (1.1)–(1.2) *in the sense of Definition* 1.1*. Moreover, for the dissolution rate we have*

$$w = r(u) \qquad \text{a.e. in } \{v = 0\} \cap \Gamma_G^T,$$

meaning that w *satisfies* (1.7).

Proof. By the weak convergence, all of the estimates stated in Lemma 2.13 hold for the limit triple (u, v, w). Furthermore, for any $t \in (t_{n-1}, t_n]$, by (2.2) we have

$$\begin{aligned}(\partial_t U^\tau(t), \phi)_\Omega + D(\nabla U^\tau(t), \nabla\phi)_\Omega \\ + (\nabla \cdot (\vec{q} U^\tau(t)), \phi)_\Omega + \epsilon\tilde{n}(\partial_t V^\tau(t), \phi)_{\Gamma_G} \\ = D(\nabla(U^\tau(t) - u^n), \nabla\phi)_\Omega + (\nabla \cdot (\vec{q}(U^\tau(t) - u^n)), \phi)_\Omega\end{aligned} \tag{2.28}$$

for all $\phi \in H^1_{0,\Gamma_D}(\Omega)$. Denoting the terms on the right by $I_1(t)$ and $I_2(t)$, taking $\phi \in L^2(0,T; H^1_{0,\Gamma_D}(\Omega))$, and integrating (2.28) in time gives

$$\begin{aligned}(\partial_t U^\tau, \phi)_{\Omega^T} + D(\nabla U^\tau, \nabla\phi)_{\Omega^T} \\ + (\nabla \cdot (\vec{q}\, U^\tau), \phi)_{\Omega^T} + \epsilon\tilde{n}(\partial_t V^\tau, \phi)_{\Gamma_G^T} \\ = \sum_{n=1}^{N} \int_{t_{n-1}}^{t^n} I_1(t) + I_2(t) dt.\end{aligned} \tag{2.29}$$

The definition (2.20) of U^τ implies for almost all $0 \le t \le T$

$$|I_1(t)| \le D \frac{t_n - t}{\tau} \|\nabla(u^n - u^{n-1})\|_\Omega \|\nabla\phi\|_\Omega \tag{2.30}$$

and

$$|I_2(t)| \le M_q \frac{t_n - t}{\tau} \|u^n - u^{n-1}\|_\Omega \|\nabla\phi\|_\Omega. \tag{2.31}$$

In (2.31) we have integrated by parts and used the essential bounds (1.4) on $\vec{q}$. Now

we can proceed by estimating the terms on the right in (2.29). For I_1 we get

$$\begin{aligned}\left|\sum_{n=1}^{N}\int_{t_{n-1}}^{t^n} I_1(t)dt\right| &\le \sum_{n=1}^{N}\int_{t_{n-1}}^{t^n}\frac{t_n-t}{\tau}D\|\nabla(u^n-u^{n-1})\|_\Omega\|\nabla\phi(t)\|_\Omega dt\\ &\le \sum_{n=1}^{N}\left(\frac{\tau}{3}D^2\|\nabla(u^n-u^{n-1})\|_\Omega^2\right)^{\frac{1}{2}}\left(\int_{t_{n-1}}^{t^n}\|\nabla\phi(t)\|_\Omega^2 dt\right)^{\frac{1}{2}}\\ &\le \frac{1}{2\sqrt{3}}\tau^{\frac{1}{4}}\int_0^T\|\nabla\phi(t)\|_\Omega^2 dt+\frac{\tau^{\frac{3}{4}}}{2\sqrt{3}}\sum_{n=1}^{N}D^2\|\nabla(u^n-u^{n-1})\|_\Omega^2\\ &\le \frac{1}{2\sqrt{3}}\left(\int_0^T\|\nabla\phi(t)\|_\Omega^2 dt+C\right)\tau^{\frac{1}{4}}.\end{aligned}$$

In the above we have used the inequality of means (2.17) with $\rho=\tau^{1/4}$, as well as the estimates (2.12) and (2.13). In a similar manner, for I_2 we get

$$\left|\sum_{n=1}^{N}\int_{t_{n-1}}^{t^n} I_1(t)dt\right| \le \left(\frac{1}{2\sqrt{3}}\int_0^T\|\nabla\phi(t)\|_\Omega^2 dt+C\tau^{\frac{1}{2}}\right)\tau^{\frac{1}{4}}.$$

Letting $\tau\searrow 0$, the weak convergence of U^τ and V^τ as well as the estimates above imply that u and v satisfy (1.9).

For the dissolution-precipitation equation (1.10) we start by analyzing the behavior of U^τ on Γ_G^T. The a priori estimates, together with Lemma 9 and Corollary 4 of [31], imply

$$U^\tau\to u \quad \text{strongly in} \quad C([0,T];H^{-s}(\Omega))\cap L^2(0,T;H^s(\Omega))$$

for any $s\in(0,1)$. Then the trace theorem (see, for example, Satz 8.7 of [36]) gives

$$U^\tau\to u \quad \text{strongly in} \quad L^2(\Gamma_G^T).$$

Taking into account the Lipschitz continuity of r, this yields

$$r(U^\tau)\to r(u) \quad \text{strongly in} \quad L^2(\Gamma_G^T).$$

Further we proceed as for (1.9). For any $t_{n-1}<t\le t_n$ and $\theta\in L^2(\Gamma_G)$ we rewrite (2.3) as

$$\begin{aligned}(\partial_t V^\tau(t),\theta)_{\Gamma_G} &= D_a(r(U^\tau(t))-W^\tau(t),\theta)_{\Gamma_G}\\ &\quad + D_a(r(u^n)-r(U^\tau(t)),\theta)_{\Gamma_G}\\ &\quad + D_a(W^\tau(t)-H_\delta(v^{n-1}),\theta)_{\Gamma_G}.\end{aligned}\tag{2.32}$$

Denoting the last two terms on the right by $I_3(t)$ and $I_4(t)$, with $\theta\in L^2(\Gamma_G^T)$, we integrate (2.32) in time and obtain

$$\begin{aligned}(\partial_t V^\tau,\theta)_{\Gamma_G^T} &= D_a(r(U^\tau)-W^\tau,\theta)_{\Gamma_G^T}\\ &\quad +\sum_{n=1}^{N}\int_{t_{n-1}}^{t^n} I_3(t)+I_4(t)dt.\end{aligned}\tag{2.33}$$

For almost all $0 \le t \le T$, since r and H_δ are Lipschitz, we use the definition of V^τ and W^τ in (2.20) and (2.21) to obtain

$$|I_3(t)| \le D_a L_r \, \frac{(t_n - t)}{\tau} \|u^n - u^{n-1}\|_{\Gamma_G} \|\theta\|_{\Gamma_G} \tag{2.34}$$

and

$$|I_4(t)| \le \frac{D_a}{\delta} \, \frac{(t_n - t)}{\tau} \|v^n - v^{n-1}\|_{\Gamma_G} \|\theta\|_{\Gamma_G}, \tag{2.35}$$

respectively. Using the estimate in (2.14), the first sum in (2.33) is bounded by

$$\begin{aligned}
\left| \sum_{n=1}^{N} \int_{t_{n-1}}^{t^n} I_3(t) dt \right| &\le D_a L_r \sum_{n=1}^{N} \int_{t_{n-1}}^{t^n} \frac{t_n - t}{\tau} \|u^n - u^{n-1}\|_{\Gamma_G} \|\theta(t)\|_{\Gamma_G} dt \\
&\le D_a L_r \sum_{n=1}^{N} \left(\frac{\tau}{3} \|u^n - u^{n-1}\|^2_{\Gamma_G} \right)^{\frac{1}{2}} \left(\int_{t_{n-1}}^{t^n} \|\theta(t)\|^2_{\Gamma_G} dt \right)^{\frac{1}{2}} \\
&\le \frac{D_a L_r}{2\sqrt{3}} \left(\|\theta\|^2_{\Gamma_G^T} + \sum_{n=1}^{N} \|u^n - u^{n-1}\|^2_{\Gamma_G} \right) \tau^{\frac{1}{2}} \\
&\le \frac{D_a L_r}{2\sqrt{3}} \left(\|\theta\|^2_{\Gamma_G^T} + C \right) \tau^{\frac{1}{2}}.
\end{aligned}$$

In the above we have use the inequality of means (2.17) with $\rho = \tau^{1/2}$. Similarly, for the last sum in (2.33), by (2.7) we get

$$\begin{aligned}
\left| \sum_{n=1}^{N} \int_{t_{n-1}}^{t^n} I_4(t) dt \right| &\le \frac{D_a}{\delta} \sum_{n=1}^{N} \int_{t_{n-1}}^{t^n} \frac{t_n - t}{\tau} \|v^n - v^{n-1}\|_{\Gamma_G} \|\theta(t)\|_{\Gamma_G} dt \\
&\le \frac{D_a}{\delta} \sum_{n=1}^{N} \left(\frac{\tau}{3} \|v^n - v^{n-1}\|^2_{\Gamma_G} \right)^{\frac{1}{2}} \left(\int_{t_{n-1}}^{t^n} \|\theta(t)\|^2_{\Gamma_G} dt \right)^{\frac{1}{2}} \\
&\le \frac{1}{2\sqrt{3}} \left(\tau {D_a}^2 \|\theta\|^2_{\Gamma_G^T} + \sum_{n=1}^{N} \|v^n - v^{n-1})\|^2_{\Gamma_G} \right) \frac{1}{\delta} \\
&\le \frac{{D_a}^2}{2\sqrt{3}} \left(\|\theta\|^2_{\Gamma_G^T} + r(M_u)^2 \, \mathrm{meas}(\Gamma_G) \right) \frac{\tau}{\delta}.
\end{aligned}$$

We now let $\tau \searrow 0$. Since $\delta = \sqrt{\tau} D_a$, the above estimates together with the weak convergence of $\partial_t V^\tau$, the strong convergence of $r(U^\tau)$, and the weak-star convergence of W^τ imply that u, v, and w satisfy (1.10_1).

It remains only to show that the dissolution rate satisfies (1.10_2). To this aim we improve the previously stated weak $L^2(\Gamma_G^T)$ convergence of V^τ to a strong one, implying also a pointwise convergence a.e. This is done by employing the Riesz–Fréchet–Kolmogorov compactness criterion, requiring estimates on translations in time and space.

We first notice that V^τ is bounded uniformly in $L^2(\Gamma_G^T)$. Moreover, by (2.7) it immediately follows that, as $h \searrow 0$, the time translations of V^τ converge to 0 in $L^2(\Gamma_G^{T-h})$, and the convergence is uniform with respect to τ. So we only have

to consider estimates for translations in space. To avoid an excess of technicalities, these are proven in a simplified setting: Decomposing $x \in \mathbb{R}^n$ into $x = (y, x_n)$, with $y \in \mathbb{R}^{n-1}$, we assume Γ_G to be open and bounded in the hyperplane $\{x_n = 0\}$. The proof for general (compact and Lipschitz) boundaries follows similarly and involves a finite number of homeomorphisms, mapping Ω locally inside the half-space $\{x_n > 0\}$.

Given a $\zeta \in \mathbb{R}^{n-1}$, we consider an arbitrary $\gamma \subset \Gamma_G$ such that $y + \zeta \in \Omega$ for any $y \in \gamma$. With Δ_ζ denoting the translation operator

$$\Delta_\zeta f(y) = f(y) - f(y + \zeta)$$

and fixing a $t \in (t_{n-1}, t_n]$ $(0 < n \leq N)$ and a $y \in \gamma$, by (2.20) we get

$$\Delta_\zeta V^\tau(t, y) = \Delta_\zeta v^{n-1}(y) + (t - t_{n-1}) D_a \big[\Delta_\zeta r(u^n)(y, x_n) - \Delta_\zeta H_\delta(v^{n-1})(y)\big].$$

Since $\delta = \sqrt{\tau} D_a$ and because r is Lipschitz continuous, this yields

$$\big|\Delta_\zeta V^\tau(t, y)\big| \leq \big|\Delta_\zeta v^{n-1}(y)\big| + D_a L_r (t - t_{n-1}) \big|\Delta_\zeta u^n(y)\big|.$$

Continuing the argument for $t = t_{n-1}, \ldots, t_1$ we obtain

$$\big|\Delta_\zeta V^\tau(t, y)\big| \leq \big|\Delta_\zeta v_I(y)\big| + \tau D_a L_r \sum_{p=1}^{n} \big|\Delta_\zeta u^p(y)\big|.$$

Since $n\tau \leq T$, standard arithmetic inequalities as well as integration over γ yield

$$\|\Delta_\zeta V^\tau(t, \cdot)\|^2_{L^2(\gamma)} \leq 2\|\Delta_\zeta v_I\|^2_{L^2(\gamma)} + 2\tau T D_a^2 L_r^2 \sum_{p=1}^{N} \|\Delta_\zeta u^p\|^2_{L^2(\gamma)}.$$

With $\bar{U}^\tau$ denoting the piecewise constant time interpolation of the time discrete approximations $\{u^n, n = 1, \ldots, N\}$,

$$\bar{U}^\tau(t) = u^n \qquad \text{whenever} \qquad t_{n-1} < t \leq t_n,$$

the inequality above becomes

$$\|\Delta_\zeta V^\tau(t, \cdot)\|^2_{L^2(\gamma)} \leq 2\|\Delta_\zeta v_I\|^2_{L^2(\gamma)} + 2T D_a^2 L_r^2 \|\Delta_\zeta \bar{U}^\tau\|^2_{L^2(\gamma^T)}. \tag{2.36}$$

Since v_I is L^2, the first term on the right vanishes as $|\zeta| \to 0$. It remains only to deal with the last term. In doing so we notice that

$$(\bar{U}^\tau - U^\tau)(t) = \frac{t_n - t}{\tau}(u^n - u^{n-1}).$$

By the estimates in (2.14), this gives

$$\big\|\bar{U}^\tau - U^\tau\big\|^2_{L^2(\Gamma_G^T)} = \sum_{n=1}^{N} \int_{t_{n-1}}^{t_n} \frac{(t_n - t)^2}{\tau^2} \|u^n - u^{n-1}\|^2_{L^2(\Gamma_G)} dt \leq C\tau.$$

Since $U^\tau \to u$ strongly in $L^2(\Gamma_G^T)$ along a sequence $\tau \searrow 0$, the same follows for $\bar{U}^\tau$. Therefore the last term in (2.36) approaches 0 uniformly with respect to τ (see also Corollary IV.26 in [2] and its converse).

Once the strong $L^2(\Gamma_G^T)$ convergence $V^\tau \to v$ is proven, we can proceed as follows. For almost every pair $(t,x) \in \Gamma_G^T$ we have either $v(t,x) > 0$ or $v(t,x) = 0$. In the first case and with $\mu := v(t,x)/2 > 0$, the pointwise convergence of V^τ to v yields the existence of a $\tau_\mu > 0$ so that $V^\tau(t,x) > \mu$ for all $\tau < \tau_\mu$. Then for any $\tau \le \min\{\tau_\mu, (\mu/D_a)^2\}$ we have $W^\tau(t,x) = 1$, implying $w(t,x) = 1$. For the second case we let $S_0 = \{v = 0\}$. Since $\partial_t v \in L^2(\Gamma_G^T)$, it follows that $\partial_t v = 0$ a.e. in the interior of S_0, and therefore $w = r(u)$ with $0 \le w \le 1$ there. □

Remark 2.16. Notice that in the proof of Theorem 2.15 we have improved the convergence of V^τ stated before. Specifically, we have shown that $V^\tau \to v$ not only weakly but also strongly in $L^2((0,T); L^2(\Gamma_G))$.

Remark 2.17. Besides the convergence of the numerical scheme, Theorem 2.15 also provides an alternative existence proof for a weak solution of (1.1)–(1.2). In [7] this is obtained by fixed-point arguments, whereas here the solution is the limit of a time discrete approximating sequence.

Remark 2.18. In this section we have considered only the time discretization of the system (1.1)–(1.2). For performing the numerical calculations that are presented in section 5, we have employed piecewise linear finite elements on Ω, whereas the crystal concentration is determined at the nodes located on Γ_G. Since the solutions of the time discrete problems are sufficiently regular, the convergence of the fully discrete scheme can be proven by standard techniques (see, e.g., [35]). In the case of convection-dominated regimes, numerical instabilities may occur as the result of the lacking maximum principle for the standard finite element method discretization. Such situations require appropriate stabilization techniques, as proposed, for example, in the book [14] or more recent works [3, 15, 19].

3. A fixed-point iteration for the time discrete problems. In this section we analyze a linear iteration scheme for solving the nonlinear time discrete Problem P_δ^n. The nonlinearity appears in the boundary condition (2.3). Numerical experiments based on a Newton-type iteration led to instabilities in the form of negative concentrations or to a precipitation in the undersaturated regime ($u \le u^*$). Moreover, there is no guarantee of convergence, unless the time stepping is not small enough. Here we discuss an alternative fixed-point approach that provides stable results. Moreover, the scheme converges linearly in H^1, regardless of the initial iteration or of the parameters ε, τ, and δ.

Assume u^{n-1} and v^{n-1} to be given and to satisfy the bounds in Lemmas 2.3, 2.4, and 2.6. To construct the iteration scheme we proceed as discussed in the beginning of section 2 and decouple the ion transport equation from the dissolution/precipitation equation on the boundary. Using (2.3), (2.2) can be rewritten as

$$(u^n - u^{n-1}, \phi)_\Omega + \tau D(\nabla u^n, \nabla\phi)_\Omega - \tau(\vec{q}u^n, \nabla\phi)_\Omega + \tau\epsilon\tilde{n}D_a(r(u^n) - H_\delta(v^{n-1}), \theta)_{\Gamma_G} = 0 \tag{3.1}$$

for all $\phi \in H^1_{0,\Gamma_D}(\Omega)$. This is a scalar elliptic equation with nonlinear boundary conditions on Γ_G. We first construct a sequence $\{u^{n,i}, i \ge 0\}$ approximating the solution u^n of (3.1). Once this is computed, we use (2.3) for determining v^n directly.

Let L_r be the Lipschitz constant of the precipitation rate r on the interval $[0, M_u]$. With

$$\mathcal{K} := \{u \in H^1_{0,\Gamma_D}(\Omega) / \ 0 \le u \le M_u \text{ a.e. in } \Omega\}, \tag{3.2}$$

and for a given $u^{n,i-1} \in \mathcal{K}$, we define $u^{n,i}$ as the solution of the linear elliptic equation

$$
\begin{aligned}
(u^{n,i} - u^{n-1}, \phi)_\Omega + \tau D(\nabla u^{n,i}, \nabla\phi)_\Omega - \tau(\vec{q}u^{n,i}, \nabla\phi)_\Omega \\
= \tau\epsilon\tilde{n}D_a L_r(u^{n,i-1} - u^{n,i}, \phi)_{\Gamma_G} \\
-\tau\epsilon\tilde{n}D_a(r(u^{n,i-1}) - H_\delta(v^{n-1}), \theta)_{\Gamma_G}
\end{aligned} \tag{3.3}
$$

for all $\phi \in H^1_{0,\Gamma_D}(\Omega)$. The starting point of the iteration can be chosen arbitrarily in $\mathcal{K}$. However, a good initial guess is $u^{n,0} = u^{n-1}$.

Comparing the above to (3.1), disregarding the superscripts $i-1$ and i, the only difference is in the appearance of the first term on the right in (3.3). In the case of convergence, this term vanishes, so $u^{n,i}$ approaches u^n. Before making this sentence more precise we mention that the above construction is common in the analysis of nonlinear elliptic problems, in particular when sub- or supersolutions are sought (see, e.g., [34, p. 96]). In [29], this approach is placed in a fixed-point context, for approximating the solution of an elliptic problem with a nonlinear and possibly unbounded source term (see also [37]). The same ideas are followed in [32, 28], where similar schemes are considered for the implicit discretization of a degenerate (fast diffusion) problem in both conformal and mixed formulation. We also mention here [33] for a related work on nonlinear elliptic equations.

Since (3.3) defines a fixed-point iteration, only a linear convergence rate is to be expected. This drawback is compensated by the stability of the approximation sequence, as well as its guaranteed convergence. These statements are made precise below.

LEMMA 3.1. *Assume* $0 \le u^{n,i-1} \le M_u$ *a.e. on* Ω. *Then* $u^{n,i}$ *solving* (3.3) *satisfies the same bounds.*

Proof. This can be shown by following the ideas used in proving Lemmas 2.3 and 2.4. We omit the details here. □

Starting the iteration with $u^{n,0} \in \mathcal{K}$, a straightforward mathematical induction argument shows the stability of the entire sequence $\{u^{n,i}, i \ge 0\}$. To prove the convergence of the scheme we let

$$
e^{n,i} := u^n - u^{n,i} \tag{3.4}
$$

denote the error at iteration i and define the H^1-equivalent norm

$$
|||f|||^2 := \|f\|^2_\Omega + \tau D\|\nabla f\|^2_\Omega + \zeta\|f\|^2_{\Gamma_G}. \tag{3.5}
$$

Here f is any function in $H^1_{0,\Gamma_D}(\Omega)$, and the constant $\zeta > 0$ is defined as

$$
\zeta := \frac{\tau}{2}\epsilon\tilde{n}D_a L_r. \tag{3.6}
$$

Notice that if τ is reasonably small, $\zeta < 1$.

The lemma below shows that the iteration error defined in (3.4) is a contraction in the norm (3.5).

LEMMA 3.2. *For* $\tau < 2/(\epsilon\tilde{n}D_a L_r)$, *an* i*-independent constant* $0 < \gamma < 1$ *exists such that*

$$
|||e^{n,i}|||^2 \le \gamma|||e^{n,i-1}|||^2,
$$

provided $u^{n,i-1}$ *satisfies the bounds in Lemma* 3.1.

Proof. With ζ given above, we start by adding $2\zeta(u^n,\phi)_{\Gamma_G}$ on both sides of (2.2). Subtracting (3.3) from the resulting equation gives

$$\begin{aligned}\left(e^{n,i},\phi\right)_\Omega + \tau D\left(\nabla e^{n,i},\nabla\phi\right)_\Omega - \tau\left(qe^{n,i},\nabla\phi\right)_\Omega + 2\zeta\left(e^{n,i},\phi\right)_{\Gamma_G}\\ = 2\zeta\left(e^{n,i-1},\phi\right)_{\Gamma_G} + \tau\epsilon\tilde{n}D_a\left(r(u^n) - r(u^{n,i-1}),\phi\right)_{\Gamma_G}.\end{aligned}$$

Since $\nabla\cdot q = 0$ and $e^{n,i}$ has a zero trace on Γ_D, taking $\phi = e^{n,i}$ into the above yields

$$\begin{aligned}&\|e^{n,i}\|_\Omega^2 + \tau D\|\nabla e^{n,i}\|_\Omega^2 + 2\zeta\|e^{n,i}\|_{\Gamma_G}^2\\ &\quad\le \tau\epsilon\tilde{n}D_a\|L_r e^{n,i-1} - \left(r(u^n) - r(u^{n,i-1})\right)\|_{\Gamma_G}\|e^{n,i}\|_{\Gamma_G}\\ &\quad\le 2\zeta\|e^{n,i-1}\|_{\Gamma_G}\|e^{n,i}\|_{\Gamma_G}.\end{aligned}$$

We then immediately get

$$|||e^{n,i}|||^2 \le \zeta\|e^{n,i-1}\|_{\Gamma_G}^2 = \zeta(1-\alpha)\|e^{n,i-1}\|_{\Gamma_G}^2 + \zeta\alpha\|e^{n,i-1}\|_{\Gamma_G}^2, \tag{3.7}$$

where $|||\cdot|||$ is introduced in (3.5) and α is an arbitrary constant in $(0,1)$ to be chosen below.

Using the trace estimate (2.16) and the inequality of means, for any $\beta > 0$ we have

$$\|e^{n,i-1}\|_{\Gamma_G}^2 \le \left(C_1 + \frac{C_2^2}{4\beta}\right)\|e^{n,i-1}\|_\Omega^2 + \beta\|\nabla e^{n,i-1}\|_\Omega^2.$$

By using this in (3.7) and taking $\alpha\in(0,1)$, $\beta>0$, and $\gamma>0$ satisfying the constraints

$$\begin{aligned}\zeta\alpha\left(C_1 + \frac{C_2^2}{4\beta}\right) &\le \gamma,\\ \zeta\alpha\beta &\le \tau D\gamma,\\ \zeta(1-\alpha) &\le \gamma,\end{aligned} \tag{3.8}$$

we obtain

$$\begin{aligned}|||e^{n,i}|||^2 &\le (1-\alpha)\zeta\|e^{n,i-1}\|_{\Gamma_G}^2 + \alpha\beta\zeta\|\nabla e^{n,i-1}\|_\Omega^2\\ &\quad + \alpha\zeta\left(C_1 + \frac{C_2^2}{4\beta}\right)\|e^{n,i-1}\|_\Omega^2 \le \gamma|||e^{n,i-1}|||^2.\end{aligned} \tag{3.9}$$

With

$$\beta = \frac{\tau D}{2}\left(C_1 + \sqrt{C_1^2 + \frac{C_2^2}{\tau D}}\right) \quad\text{and}\quad \alpha = \left(1 + \frac{C_1}{2} + \frac{1}{2}\sqrt{C_1^2 + \frac{C_2^2}{\tau D}}\right)^{-1},$$

the restrictions on α and β are fulfilled. Further, this choice also gives identical lower bounds for γ in (3.8), for which we can now take

$$\gamma = \zeta\left(C_1 + \sqrt{C_1^2 + \frac{C_2^2}{\tau D}}\right)\left(2 + C_1 + \sqrt{C_1^2 + \frac{C_2^2}{\tau D}}\right)^{-1}. \tag{3.10}$$

By the assumptions on τ we have $\gamma < 1$, and hence the iteration error is contractive in the norm defined in (3.5). ∎

Remark 3.3. The iteration (3.3) defines an operator $\mathcal{T} : \mathcal{K} \to \mathcal{K}$. Following the steps in the proof of Lemma 3.2, we can show that $\mathcal{T}$ is a contraction with respect to the norm defined in (3.5). Therefore $\mathcal{T}$ has a unique fixed point, yielding the existence and uniqueness of a solution for the nonlinear equation (3.1). This immediately implies the existence and uniqueness of a solution for Problem P_δ^n, as stated in Lemma 2.2.

Lemma 3.2 implies the linear convergence in H^1 of the iteration sequence $\{u^{n,i}, i \geq 0\}$. Moreover, its limit is the solution u^n of (3.1).

THEOREM 3.4. *With $u^{n,0} \in H^1_{0,\Gamma_D}(\Omega)$ bounded essentially by 0 and M_u, if $\tau < 2/(\epsilon\tilde{n}D_aL_r)$, the iteration* (3.3) *is convergent. Specifically, for all $i > 0$ we have*

$$|||e^{n,i}|||^2 \leq \gamma^i |||e^{n,0}|||^2.$$

Remark 3.5. The contraction constant γ in the above is bounded from above by $\zeta = \tau\epsilon\tilde{n}D_aL_r/2$. Notice that the first term in the H^1-equivalent norm in (3.5) does not depend on τ. As $\tau \searrow 0$, γ approaches 0, implying a fast convergence of the iteration at least in the L^2 sense. In the numerical computations presented in the following section, 3–4 iterations were enough for obtaining a good numerical approximation of the time discrete solution.

As stated in Theorem 3.4, convergence is achieved as $i \to \infty$. In practice we stop this procedure after a finite (small) number of iterations. This means that at each time step we are adding an iteration error to the time discrete approximation. This error depends on the number of iterations performed per time step, as well as on the initial iteration error. As mentioned before, the solution at the previous time step can be used for initiating the iteration. In the remaining part of this section we show that by this choice the total error is vanishing as $\tau \searrow 0$.

To make this statement rigorous we assume that i iterations are performed at each time step n. The computed solution $\tilde{u}^n = u^{n,i}$ will be only an approximation of u^n, the solution of (3.1). Let now $\tilde{e}^n$ denote the error at the time step n:

$$\tilde{e}^n := u^n - \tilde{u}^n.$$

Based on the stability estimates in Lemma 2.11, we can estimate the total error that is accumulated in the numerical approximation of the time discrete sequence $\{u^n, n = 1, \ldots, N\}$.

LEMMA 3.6. *Assume that, for each $n = 1, \ldots, N$, i iterations* (3.3) *are performed by starting with $u^{n,0} = \tilde{u}^{n-1} = u^{n-1,i}$. We have*

$$\sum_{n=1}^{N} |||\tilde{e}^n||| \leq C\frac{\gamma^{i/2}}{\tau^{1/4}}.$$

Remark 3.7. Since $\gamma = O(\tau)$, the total error vanishes as $\tau \searrow 0$.

Proof. With $u^{n,0} = \tilde{u}^{n-1}$, the initial error at the time step n is given by

$$e^{n,0} = u^n - \tilde{u}^{n-1} = u^n - u^{n-1} + \tilde{e}^{n-1}.$$

By Theorem 3.4 the error at the time step n can be estimated as

$$|||\tilde{e}^n||| \leq \gamma^{i/2}|||e^{n,0}||| \leq \gamma^{i/2}\left(|||u^n - u^{n-1}||| + |||\tilde{e}^{n-1}|||\right).$$

Repeating this estimates inductively for $n = 1, \dots, N$, since $\tilde{e}^0 = 0$ we obtain

$$|||\tilde{e}^n||| \le \sum_{k=1}^{n} \gamma^{(n+1-k)i/2} |||u^k - u^{k-1}|||. \tag{3.11}$$

Adding the above for $n = 1$ up to N gives

$$\sum_{n=1}^{N} |||\tilde{e}^n||| \le \frac{\gamma^{i/2}}{1-\gamma^{i/2}} \sum_{k=1}^{N} \left(1 - \gamma^{(N+1-k)i/2}\right) |||u^k - u^{k-1}|||.$$

By the definition of γ in (3.10), if τ is small enough, the fraction in the above can be bounded by $C\gamma^{i/2}$ for some constant $C > 0$. Now the proof can be completed straightforwardly by applying the stability estimates in Lemma 2.11. ∎

4. Numerical results. In this section we present some numerical simulations obtained for the undersaturated regime, when only dissolution is possible. Extensive numerical results for both dissolution and precipitation, and for high or low Damköhler numbers, will be presented in a forthcoming paper. All of the computations are implemented in the research software *SciFEM* (Scilab finite element method [38]; see also [4]).

Here we consider a two-dimensional domain Ω, obtained by removing the disc $B_R(0,0)$ of radius $R = 1/4$ and centered in the origin form the square $(-1/2, 1/2)^2$. The impermeable grain is represented by the disc, and flow takes place around the grain, from the left boundary to the right one. The present experiments are similar to those presented in [7], where a simple geometry—a two-dimensional strip—has been considered. There the occurrence of a dissolution front has been investigated both analytically and numerically. After a waiting time t^*, the front started moving in the flow direction. As we will see below, the present computations reveal similar features.

We have used the following parameters and rate function:

$$D = 1,\ D_a = 1,\ \varepsilon = 1,\ \tilde{m} = \tilde{n} = 1.0, \text{ and } r(u,c) = \frac{10}{9}[u]_+[u - 0.1]_+.$$

This gives $u_* = 0.1$ and $u^* = 1.0$. The initial and (external) boundary conditions are

$$\begin{aligned} v_I &= 0.1 && \text{on } \Gamma_G, \\ u_I &= 1.0 && \text{in } \Omega, \\ u &= u_* && \text{if } x = -1/2, t > 0, \\ \partial_\nu u &= 0 && \text{on } \partial\Omega \backslash \{\Gamma_G \cup \{x = -1/2\}\}. \end{aligned}$$

With the data above, the system is initially in equilibrium: No precipitation or dissolution is taking place. This equilibrium is perturbed by injecting an undersaturated fluid at the inflow boundary $x = -1/2$. We start with a uniformly distributed layer of precipitate. As resulting from the L^∞ estimates, only dissolution is possible.

The fluid velocity $\vec{q}$ is determined by solving numerically the Stokes system

$$\begin{cases} \mu \triangle \vec{q} = \nabla p & \text{in } \Omega, \\ \nabla \cdot \vec{q} = 0 & \text{in } \Omega, \end{cases}$$

where $\mu = 0.01$. At both the in- and outflow boundaries $x = \pm 1/2$ we take $\vec{q} = (2, 0)$. On the upper and lower boundaries $y = \pm 1/2$ we take $q_y = 0$, and $\partial_y q_x = 0$. For

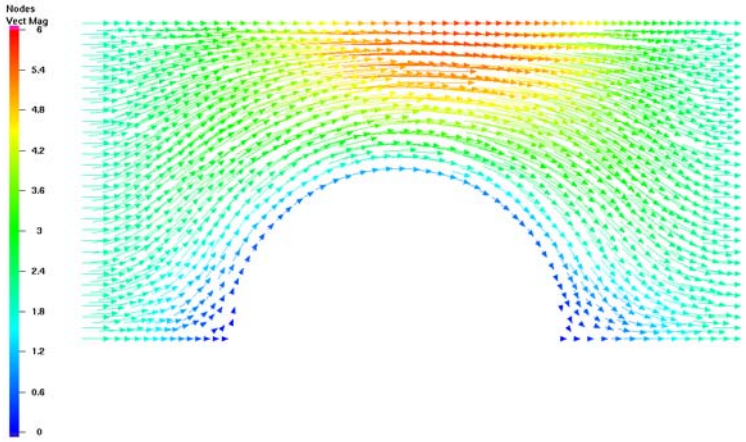


FIG. 1. *The computed velocity field* $\vec{q}$.

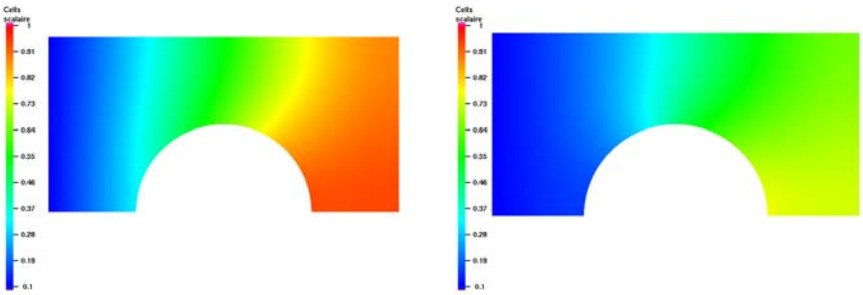


FIG. 2. *Concentration of the cation* u *at* $t_1 = 0.1$ *(left) and* $t_2 = 0.2$ *(right).*

symmetry reasons, the computations are restricted to the upper half of the domain. The numerical approximation of $\vec{q}$ is obtained by employing the bubble stabilized finite element method proposed in [30] (see also [18]). The computed velocity field is presented in Figure 1.

For computing the cation concentration u and the precipitate concentration v we have applied the scheme (2.2)–(2.3) with a fixed time step $\tau = 10^{-4}$ and the regularization parameter $\delta = 10^{-4}$. As mentioned in the appendix of [6] (see also Theorem 2.15), we set $w = \min\{r(u), 1\}$ whenever $v = 0$. The emerging nonlinear time discrete problems are solved by the linearization (3.3), with $u^{n,0} = u^{n-1}$. The procedure is stopped once the maximal difference between two successive iterations is reduced below 10^{-3}. In our experiments 3–4 iterates were sufficient to fulfill this stopping criterion. We have used piecewise linear finite elements for the spatial discretization of the time discrete problems in Ω. The crystal concentration is determined at the corresponding nodes on Γ_G. In this specific example the Peclet number is moderate, so no special stabilizing techniques were needed.

The computations are performed up to $T = 0.4$. Figure 2 displays the cation

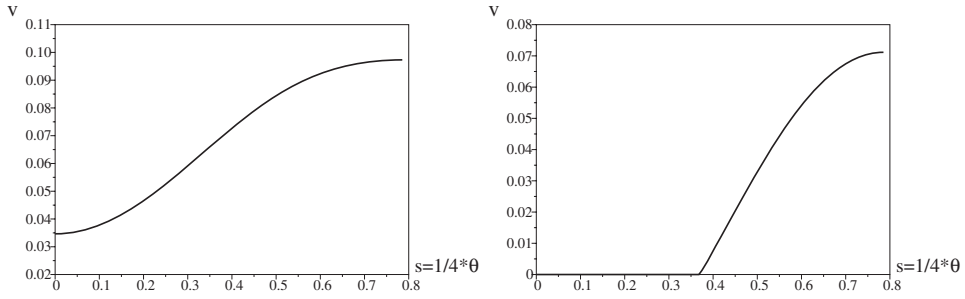


FIG. 3. *Concentration of the precipitate v at $t_1 = 0.1$ (left) and $t_2 = 0.2$ (right) as a function of the arc length.*

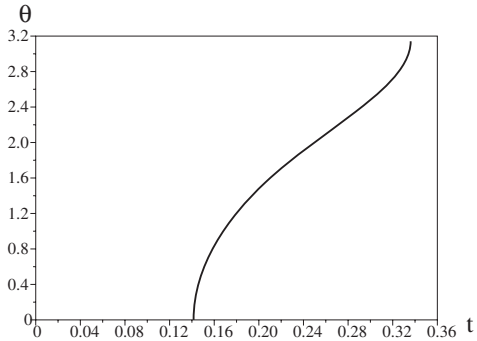


FIG. 4. *The position of the dissolution front as a function of time: At $t^* \approx 0.1415$ precipitate is present everywhere on the grain boundary; after $\tilde{t} \approx 0.3359$ the entire precipitate has been dissolved.*

concentration u at two different times $t_1 = 0.1$ and $t_2 = 0.2$. As expected, u is higher at the outflow than at the inflow and stays bounded by u_* and u^*. This excludes the possibility of precipitation. Moreover, the amount of cations is decreasing in time. Figure 3 displays the precipitate concentration v as a function of the arc length $s = R\theta$. Here $R = 1/4$, while θ is the angle between the radii through the leftmost point $(-R, 0)$ through the location of interest on Γ_G. Notice that v is monotone with respect to s. A major difference can be noticed in the v profiles for $t_1 = 0.1$ and $t_2 = 0.2$. In the former case we have $v > 0$ for any s, and thus precipitate is present everywhere. For t_2 we identify two different regions on Γ_G. In the left one the precipitate has been completely dissolved, whereas crystals can be found everywhere in the region on the right. This shows the occurrence of a dissolution front, located at the free boundary separating the two regions on Γ_G. At t_2 the dissolution front is located close to $(0, R)$, the upper point of Γ_G.

The evolution of the dissolution front is presented in Figure 4, where the angle θ is given as a function of time. Up to $t^* \approx 0.1415$, the precipitate is present everywhere on the grain surface Γ_G. For any time between t^* and $\tilde{t} \approx 0.3359$ the dissolution front moves towards the rightmost point on Γ_G, $(R, 0)$. After $\tilde{t}$ the entire precipitate has been dissolved. Notice the fast movement of the free boundary as t is greater than but close to the waiting time t^*. This is due to the fact that the grain boundary Γ_G is nearly transversal to the flow. Since the inflow boundary conditions for u and $\vec{q}$ are constant, a minimal variation of u in the vertical direction can be expected. This implies a similar behavior for v if s (or θ) is close to 0, and therefore the complete dissolution along this region occurs nearly simultaneously.

5. Conclusion. We have analyzed a numerical scheme for the time discretization of (1.1)–(1.2). This is a simplification of the pore-scale model for crystal dissolution and precipitation in porous media. The main difficulties associated with the model, a parabolic problem that is coupled through the boundary to a differential inclusion, are still present in this setting.

The numerical scheme is implicit in u and explicit in v. A regularization step is employed for dealing with the multivalued dissolution rate. We prove the stability of the scheme in both L^∞ as well as energy norms. Further, compactness arguments are used for showing that the scheme is convergent.

A fixed-point iteration is proposed for solving the nonlinear time discrete problems. This linearization is stable and converges linearly, regardless of the discretization parameters and the starting point.

Acknowledgments. We thank Prof. A. Mikelić (Lyon) and the members of the Applied Analysis group in Eindhoven for their suggestions.

REFERENCES

[1] N. Bouillard, R. Eymard, R. Herbin, and Ph. Montarnal, *Diffusion with dissolution and precipitation in a porous media: Mathematical analysis and numerical approximation of a simplified model*, M2AN Math. Model. Numer. Anal., to appear.

[2] H. Brezis, *Analyse fonctionnelle. Théorie et applications*, Masson, Paris, 1983.

[3] E. Burman and A. Ern, *Stabilized Galerkin approximation of convection-diffusion-reaction equations: Discrete maximum principle and convergence*, Math. Comp., 74 (2005), pp. 1637–1652.

[4] V. M. Devigne, *Ecoulements et Conditions aux Limites Particuliéres Appliquées en Hydrogéologie et Théorie Mathématique des Processus de Dissolution/Précipitation en Milieux Poreux*, Ph.D. Thesis, Ecole Nationale Supérieure des Mines de Saint-Etienne, Saint-Étienne, France, 2006.

[5] C. J. van Duijn, A. Mikelić, I. S. Pop, and C. Rosier, *Effective Dispersion Equations for Reactive Flows with Dominant Peclet and Damkohler Numbers*, CASA report 07-20, Eindhoven University of Technology, Eindhoven, The Netherlands, 2007.

[6] C. J. van Duijn and P. Knabner, *Travelling wave behaviour of crystal dissolution in porous media flow*, European J. Appl. Math., 8 (1997), pp. 49–72.

[7] C. J. van Duijn and I. S. Pop, *Crystal dissolution and precipitation in porous media: Pore scale analysis*, J. Reine Angew. Math., 577 (2004), pp. 171–211.

[8] R. Eymard, T. Gallouët, R. Herbin, D. Hilhorst, and M. Mainguy, *Instantaneous and noninstantaneous dissolution: Approximation by the finite volume method*, in Actes du 30ème Congrès d'Analyse Numérique: CANum '98 (Arles, 1998), R. Boyer et al., eds., ESAIM Proc. 6, Soc. Math. Appl. Indust., Paris, 1999, pp. 41–55.

[9] B. Faugeras, J. Pousin, and F. Fontvieille, *An efficient numerical scheme for precise time integration of a diffusion-dissolution/precipitation chemical system*, Math. Comp., 75 (2006), pp. 209–222.

[10] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Monogr. Stud. Math. 24, Pitman, Boston, MA, 1985.

[11] U. Hornung and W. Jäger, *A model for chemical reactions in porous media*, in Complex Chemical Reaction Systems. Mathematical Modeling and Simulation, J. Warnatz and W. Jäger, eds., Chemical Physics 47, Springer, Berlin, 1987, pp. 318–334.

[12] U. Hornung and W. Jäger, *Diffusion, convection, adsorption, and reaction of chemicals in porous media*, J. Differential Equations, 92 (2001), pp. 199–225.

[13] U. Hornung, W. Jäger, and A. Mikelić, *Reactive transport through an array of cells with semipermeable membranes*, RAIRO Modél. Math. Anal. Numér., 28 (1994), pp. 59–94.

[14] T. Ikeda, *Maximum Principle in Finite Element Models for Convection-Diffusion Phenomena*, Lecture Notes in Numer. Appl. Anal. 4, North-Holland Mathematics Studies, Amsterdam, 1983.

[15] J. Karátson and S. Korotov, *Discrete maximum principles for finite element solutions of nonlinear elliptic problems with mixed boundary conditions*, Numer. Math., 99 (2005), pp. 669–698.

[16] R. B. Kellogg and J. E. Osborn, *A regularity result for the Stokes problem in a convex polygon*, J. Funct. Anal., 21 (1976), pp. 397–431.
[17] P. Knabner, C. J. van Duijn, and S. Hengst, *An analysis of crystal dissolution fronts in flows through porous media. Part* 1: *Compatible boundary conditions*, Adv. Water Res., 18 (1995), pp. 171–185.
[18] P. Knobloch, *On the application of the p_1^{mod} element to incompressible flow problems*, Comput. Vis. Sci., 6 (2004), pp. 185–195.
[19] P. Knobloch, *Improvements of the Mizukami-Hughes method for convection-diffusion equations*, Comput. Methods Appl. Mech. Engrg., 196 (2006), pp. 579–594.
[20] J. R. Kweona and R. B. Kellogg, *Compressible Stokes problem on nonconvex polygonal domains*, J. Differential Equations, 176 (2001), pp. 290–314.
[21] E. Maisse and J. Pousin, *Diffusion and dissolution/precipitation in an open porous reactive medium*, J. Comput. Appl. Math., 82 (1997), pp. 279–290.
[22] E. Maisse and J. Pousin, *Finite element approximation of mass transfer in a porous medium with non equilibrium phase change*, J. Numer. Math., 12 (2004), pp. 207–231.
[23] A. Mikelić and C. Rosier, *Rigorous upscaling of the infinite adsorption rate reactive flow under dominant Peclet number through a pore*, Ann. Univ. Ferrara Sez. VII Sci. Mat., 53 (2007), pp. 333–359.
[24] A. Mikelić, V. Devigne, and C. J. van Duijn, *Rigorous upscaling of the reactive flow through a pore, under dominant Peclet and Damkohler numbers*, SIAM J. Math. Anal., 38 (2006), pp. 1262–1287.
[25] T. L. van Noorden and I. S. Pop, *A Stefan problem modelling dissolution and precipitation in porous media*, IMA J. Appl. Math., to appear.
[26] T. L. van Noorden, I. S. Pop, and M. Röger, *Crystal dissolution and precipitation in porous media: L^1-contraction and uniqueness*, Discrete Contin. Dyn. Syst., supplement (2007), pp. 1013–1020.
[27] I. S. Pop, V. Devigne, C. J. van Duijn, and T. Clopeau, *A numerical scheme for the micro scale dissolution and precipitation in porous media*, in Numerical Mathematics and Advanced Applications, A. Bermudez de Castro, D. Gomez, P. Quintela, and P. Salgado, eds., Springer-Verlag, Heidelberg, 2006, pp. 362–370.
[28] I. S. Pop, F. Radu, and P. Knabner, *Mixed finite elements for the Richards' equation: Linearization procedure*, J. Comput. Appl. Math., 168 (2004), pp. 365–373.
[29] I. S. Pop and W. A. Yong, *On the existence and uniqueness of a solution for an elliptic problem*, Studia Univ. Babeş-Bolyai Math., 45 (2000), pp. 97–107.
[30] A. Russo, *Bubble stabilization of finite element methods for the linearized incompressible Navier-Stokes equations*, Comput. Methods Appl. Mech. Engrg., 132 (1996), pp. 335–343.
[31] J. Simon, *Compact sets in the space $L^p(0,T;B)$*, Ann. Mat. Pura Appl. (4), 146 (1987), pp. 65–96.
[32] M. Slodička, *A robust and efficient linearization scheme for doubly nonlinear and degenerate parabolic problems arising in flow in porous media*, SIAM J. Sci. Comput., 23 (2002), pp. 1593–1614.
[33] M. Slodička, *A robust linearisation scheme for a nonlinear elliptic boundary value problem: Error estimates*, ANZIAM J., 46 (2005), pp. 449–470.
[34] J. Smoller, *Shock Waves and Reaction-Diffusion Equations*, Springer-Verlag, New York, 1983.
[35] V. Thomée, *Galerkin Finite Element Methods for Parabolic Problems*, 2nd ed., Springer-Verlag, Berlin, 2006.
[36] J. Wloka, *Partielle Differentialgleichungen*, B. G. Teubner, Stuttgart, 1982.
[37] W. A. Yong and I. S. Pop, *A Numerical Approach to Porous Medium Equations*, preprint 96-50 (SFB 359), IWR, University of Heidelberg, 1996.
[38] *An Introduction to Scilab*, available online at http://lapcs.univ-lyon1.fr/∼clopeau/scilab/index.html (in French).

SIAM J. NUMER. ANAL.
Vol. 46, No. 2, pp. 920–929
© 2008 Society for Industrial and Applied Mathematics

CONVEXITY-PRESERVING PIECEWISE RATIONAL QUARTIC INTERPOLATION*

XULI HAN†

Abstract. A family of piecewise rational quartic interpolants is given. Identified uniquely by the value of a tension parameter λ_i, each interpolant of the family can be C^2 spline without solving a linear or nonlinear system of consistency equations for the derivative values at the knots. The interpolant can preserve the local convexity/concavity properties of the given data. A proper choice of λ_i to guarantee shape preservation is given. A convergence analysis establishes an error bound in terms of λ_i and shows that $O(h^3)$ accuracy is obtained for C^2 continuity. Several examples are supplied to support the practical value of the method.

Key words. rational interpolation, tension spline, convexity-preserving

AMS subject classifications. 65D05, 41A05

DOI. 10.1137/060671577

1. Introduction.

In industrial design and manufacturing it is often required to generate a smooth function passing through a given set of data which preserves certain shape properties of the data such as positivity, monotonicity, or convexity, that is, a smooth shape-preserving interpolant.

Different polynomial methods have received attention in the study of shape-preserving interpolation. Some necessary and sufficient conditions for shape-preserving properties have been discussed. For brevity, the reader is referred to [1, 2, 3, 4]. These methods have a common feature in that no additional knots need be supplied. In contrast, the papers [5, 6, 7, 8] discussed the methods by adding one or two additional knots on the subinterval so that the monotonicity and convexity of the data are preserved.

Some rational spline methods have been presented for shape-preserving interpolation. In the papers [9, 10], a piecewise rational quadratic function was developed which produces a monotonic interpolant to monotonic data. In the papers [11, 12, 13, 14, 15], shape-preserving interpolations to convex data were shown. The given rational interpolants are C^1 continuous. For C^2 continuity, the solution of the consistency equations is concerned; see [10, 11, 12, 13, 14, 15]. For some tension methods, such as those mentioned in [15], the proper choice of the tension parameter to guarantee shape preservation is still an unsettled question.

This paper presents a piecewise rational quartic interpolant that preserves local convexity/concavity properties of the given data and is of C^2 continuity without solving a system of consistency equations for the derivative values at the knots.

The present paper is organized as follows. In section 2, the piecewise expression of the rational quartic interpolant is described and the C^2 continuity is shown. The shape-preserving properties are shown and the proper tension values are given in section 3. In section 4, error estimates are shown. Some numerical examples are given in section 5, and conclusions are given in section 6.

*Received by the editors October 5, 2006; accepted for publication (in revised form) August 7, 2007; published electronically March 5, 2008.
http://www.siam.org/journals/sinum/46-2/67157.html
†School of Mathematical Sciences and Computing Technology, Central South University, Changsha, 410083, People's Republic of China (xlhan@mail.csu.edu.cn).

2. Piecewise rational quartic interpolant. Let (x_i, f_i) be given real data and d_i be chosen derivative values at knots x_i, $i = 1, 2, \ldots, n$, where $a = x_1 < x_2 < \cdots < x_n = b$ is a partition of the interval $[a, b]$. We put

$$h_i = x_{i+1} - x_i, \quad \Delta_i = (f_{i+1} - f_i)/h_i, \quad A_i = \Delta_i - d_i, \quad B_i = d_{i+1} - \Delta_i$$

for $i = 1, 2, \ldots, n-1$.

For $[x_i, x_{i+1}]$, $i = 1, 2, \ldots, n-1$, we construct the interpolant

$$\begin{aligned} s(x) &= (1-t)f_i + tf_{i+1} \\ &\quad - \frac{h_i(1-t)t}{Q_i(t)}\big[(1-t)^2 A_i + \lambda_i(1-t)tA_iB_i + t^2 B_i\big], \end{aligned} \tag{1}$$

where $t = (x - x_i)/h_i$, λ_i are tension parameters, and

$$Q_i(t) = (1-t)^2 + \lambda_i(1-t)t\big[(1-t)B_i + tA_i\big] + t^2. \tag{2}$$

In the following section, we will suggest the choice of λ_i for the subinterval.

Clearly the interpolatory properties $s(x_i) = f_i, s(x_{i+1}) = f_{i+1}$ hold. Further,

$$\lim_{\lambda_i \to \infty} s(x) = (1-t)f_i + tf_{i+1} - \frac{h_i(1-t)tA_iB_i}{(1-t)B_i + tA_i}. \tag{3}$$

This form is the interpolant given in [12], where it was shown that the form (3) is also a special form given in [11].

THEOREM 1. *For any given d_i, $i = 2, 3, \ldots, n-1$, $s(x) \in C^1[a, b]$. If*

$$d_i = \frac{1}{h_{i-1} + h_i}(h_{i-1}\Delta_i + h_i\Delta_{i-1}), \ i = 2, 3, \ldots, n-1, \tag{4}$$

then $s(x) \in C^2[a, b]$.

Proof. For $[x_i, x_{i+1}], i = 1, 2, \ldots, n-1$, after much manipulation, the derivatives $s'(x), s''(x)$ are

$$s'(x) = \Delta_i - \frac{1}{Q_i^2(t)} \sum_{j=0}^{6} (1-t)^{6-j} t^j \bar{A}_{ij}, \tag{5}$$

$$s''(x) = \frac{1}{h_i Q_i^3(t)} \sum_{j=0}^{8} (1-t)^{8-j} t^j \bar{B}_{ij}, \tag{6}$$

where $\bar{A}_{i0} = A_i$, $\bar{A}_{i6} = -B_i$, $\bar{A}_{i1} = 2\lambda_i A_i B_i$, $\bar{A}_{i5} = -2\lambda_i A_i B_i$,

$$\begin{aligned} \bar{A}_{i2} &= A_i B_i^2 \lambda_i^2 + A_i(B_i - A_i)\lambda_i - 2A_i + 3B_i, \\ \bar{A}_{i3} &= 2(B_i^2 - A_i^2)\lambda_i - 4A_i + 4B_i, \\ \bar{A}_{i4} &= -A_i^2 B_i \lambda_i^2 - (A_i - B_i)B_i\lambda_i - 3A_i + 2B_i, \end{aligned}$$

and $\bar{B}_{i0} = 2A_i$, $\bar{B}_{i8} = 2B_i$,

$$\begin{aligned} \bar{B}_{i1} &= 6A_i^2\lambda_i + 10A_i - 6B_i, \\ \bar{B}_{i2} &= 6A_i^2 B_i \lambda_i^2 + 6(A_iB_i + 2A_i^2 - B_i^2)\lambda_i + 26A_i - 18B_i, \\ \bar{B}_{i3} &= 2A_i^2 B_i^2 \lambda_i^3 + 2(7A_i^2 B_i + A_i B_i^2 - A_i^3 - B_i^3)\lambda_i^2 \\ &\quad + 2(16A_iB_i + A_i^2 - 8B_i^2)\lambda_i + 30A_i - 18B_i, \end{aligned}$$

$$\begin{aligned}
\bar{B}_{i4} &= 4A_i^2B_i^2\lambda_i^3 + 2(5A_i^2B_i + 5A_iB_i^2 - 2A_i^3 - 2B_i^3)\lambda_i^2 \\
&\quad + 2(26A_iB_i - 7A_i^2 - 7B_i^2)\lambda_i + 6A_i + 6B_i, \\
\bar{B}_{i5} &= 2A_i^2B_i^2\lambda_i^3 + 2(A_i^2B_i + 7A_iB_i^2 - A_i^3 - B_i^3)\lambda_i^2 \\
&\quad + 2(16A_iB_i - 8A_i^2 + B_i^2)\lambda_i - 18A_i + 30B_i, \\
\bar{B}_{i6} &= 6A_iB_i^2\lambda_i^2 + 6(A_iB_i - A_i^2 + 2B_i^2)\lambda_i - 18A_i + 26B_i, \\
\bar{B}_{i7} &= 6B_i^2\lambda_i - 6A_i + 10B_i.
\end{aligned}$$

From (5) we have

$$s'(x_i^+) = d_i,\ s'(x_{i+1}^-) = d_{i+1}.$$

These imply $s(x) \in C^1[a,b]$.

From (4) and (6) we have

$$s''(x_i^+) = \frac{2A_i}{h_i} = \frac{2}{h_{i-1}+h_i}(\Delta_i - \Delta_{i-1}),\ \ s''(x_{i+1}^-) = \frac{2B_i}{h_i} = \frac{2}{h_i+h_{i+1}}(\Delta_{i+1} - \Delta_i).$$

These imply $s(x) \in C^2[a,b]$. □

Based on Theorem 1, the interpolant (1) can be C^2 continuous without solving a system of consistency equations for the derivative values at the knots. Actually, the derivatives given by (4) are widely used; see [9, 12]. In [11] and [14], it has been shown that the limits of the derivatives at the knots are given by (4).

3. Shape-preserving properties. At first, we establish some properties of $s(x)$ on strictly convex/concave sets of data.

THEOREM 2. *For $x \in [x_i, x_{i+1}]$, $i = 1, 2, \ldots, n-1$, if $A_i > 0, B_i > 0$, then $s''(x) > 0$ when*

$$\lambda_i \geq \max\left\{\frac{A_i}{B_i^2}, \frac{B_i}{A_i^2}\right\}; \tag{7}$$

if $A_i < 0, B_i < 0$, then $s''(x) < 0$ when

$$\lambda_i \leq \min\left\{\frac{A_i}{B_i^2}, \frac{B_i}{A_i^2}\right\}. \tag{8}$$

Proof. From (6), the expression of $s''(x)$ can be reduced to the following form:

$$s''(x) = \frac{1}{h_iQ_i^3(t)}\sum_{j=0}^{6}(1-t)^{6-j}t^jB_{ij}, \tag{9}$$

where $B_{i0} = 2A_i$, $B_{i6} = 2B_i$,

$$\begin{aligned}
B_{i1} &= 6A_i^2\lambda_i + 6A_i - 6B_i, \\
B_{i2} &= 6A_i^2B_i\lambda_i^2 + 6(A_i - B_i)B_i\lambda_i + 12A_i - 6B_i, \\
B_{i3} &= 2A_i^2B_i^2\lambda_i^3 + 2(A_i^2B_i + A_iB_i^2 - A_i^3 - B_i^3)\lambda_i^2 + 4(5A_iB_i - A_i^2 - B_i^2)\lambda_i, \\
B_{i4} &= 6A_iB_i^2\lambda_i^2 + 6A_i(B_i - A_i)\lambda_i - 6A_i + 12B_i, \\
B_{i5} &= 6B_i^2\lambda_i - 6A_i + 6B_i.
\end{aligned}$$

If $A_i > 0, B_i > 0, \lambda_i \geq \max\{A_i/B_i^2, B_i/A_i^2\}$, and $A_i \leq B_i$, then

$$\lambda_i \geq \max\left\{\frac{A_i}{B_i^2}, \frac{B_i}{A_i^2}\right\} = \frac{B_i}{A_i^2} \geq \frac{1}{A_i} \geq \frac{1}{B_i} \geq \frac{A_i}{B_i^2}.$$

Since $B_{i0} = 2A_i, B_{i6} = 2B_i$,

$$\begin{aligned}
B_{i1} &= 6(A_i^2\lambda_i - B_i) + 6A_i, \quad B_{i5} = 6(B_i^2\lambda_i - A_i) + 6B_i,\\
B_{i2} &= 6B_i(A_i^2\lambda_i - B_i)\lambda_i + 6(A_i\lambda_i - 1)B_i + 12A_i,\\
B_{i3} &= 2(A_i^2\lambda_i - B_i)B_i^2\lambda_i^2 + 2A_i^2(B_i - A_i)\lambda_i^2\\
&\quad + 2(A_iB_i\lambda_i + 8A_i - 2B_i)B_i\lambda_i + 4A_i(B_i - A_i)\lambda_i,\\
&= 2(A_i^2\lambda_i - B_i)B_i^2\lambda_i^2 + 2A_i^2(B_i - A_i)\lambda_i^2\\
&\quad + \frac{2B_i}{A_i}[(A_i^2\lambda_i - B_i)B_i + 8A_i^2 + B_i^2 - 2A_iB_i]\lambda_i + 4A_i(B_i - A_i)\lambda_i,\\
B_{i4} &= 6A_i(B_i^2\lambda_i - A_i)\lambda_i + 6A_i(B_i\lambda_i - 1) + 12B_i,
\end{aligned}$$

we can see that $B_{ij} > 0$, $j = 0, 1, \ldots, 6$, and then $s''(x) > 0$.

In the same way, we can show that $s''(x) > 0$ when $A_i > 0$, $B_i > 0$, $\lambda_i \geq \max\{A_i/B_i^2, B_i/A_i^2\}$, and $A_i > B_i$. The proof for the case $A_i < 0$, $B_i < 0$ is similar. □

At the end knots x_1, x_n, if we set derivative values

$$d_1 = \Delta_1 - \frac{h_1}{h_1 + h_2}(\Delta_2 - \Delta_1), \quad d_n = \Delta_{n-1} + \frac{h_{n-1}}{h_{n-2} + h_{n-1}}(\Delta_{n-1} - \Delta_{n-2}), \tag{10}$$

then

$$A_1 = \frac{h_1}{h_1 + h_2}(\Delta_2 - \Delta_1), \quad B_{n-1} = \frac{h_{n-1}}{h_{n-2} + h_{n-1}}(\Delta_{n-1} - \Delta_{n-2}).$$

From (4) we have

$$A_i = \frac{h_i}{h_{i-1} + h_i}(\Delta_i - \Delta_{i-1}), \quad B_{i-1} = \frac{h_{i-1}}{h_{i-1} + h_i}(\Delta_i - \Delta_{i-1})$$

for $i = 2, 3, \ldots, n-1$. Therefore, if we take d_i as defined by (4) and (10), then the given interpolant is convexity-preserving when the λ_i satisfy (7) or (8). For $x \in [x_i, x_{i+1}]$, when the differences Δ_k $(k = i-1, i, i+1)$ satisfy the inequalities $\Delta_{i-1} < \Delta_i < \Delta_{i+1}$ or $\Delta_{i-1} > \Delta_i > \Delta_{i+1}$, the interpolant is convexity-preserving.

Remark. If $A_iB_i > 0$, then $s(x)$ is also monotonicity-preserving for $x \in [x_i, x_{i+1}]$ when λ_i is large enough. In fact, (5) can be written as

$$s'(x) = \frac{1}{Q_i^2(t)}\sum_{j=0}^{6}(1-t)^{6-j}t^jA_{ij},$$

where $A_{i0} = d_i$, $A_{i6} = d_{i+1}$,

$$\begin{aligned}
A_{i1} &= 2B_id_i\lambda_i + 2\Delta_i, \quad A_{i5} = 2A_id_{i+1}\lambda_i + 2\Delta_i,\\
A_{i2} &= B_i^2d_i\lambda_i^2 + [2(A_i + B_i)\Delta_i + A_i^2 - A_iB_i]\lambda_i + 3\Delta_i + 2A_i - 3B_i,\\
A_{i3} &= 2A_iB_i\Delta_i\lambda_i^2 + 2[(A_i + B_i)\Delta_i + A_i^2 - B_i^2]\lambda_i + 4\Delta_i + 4A_i - 4B_i,
\end{aligned}$$

$$A_{i4} = A_i^2 d_{i+1}\lambda_i^2 + [2(A_i + B_i)\Delta_i + A_iB_i - B_i^2]\lambda_i + 3\Delta_i + 3A_i - 2B_i.$$

Therefore, if we choose λ_i so that $sgn(\lambda_i) = sgn(A_i)$, then $sgn(A_{ij}) = sgn(d_i)(j = 0, 1, 2)$, $sgn(A_{i3}) = sgn(\Delta_i)$, $sgn(A_{ij}) = sgn(d_{i+1})$ $(j = 4, 5, 6)$ when λ_i is large enough.

Now we give an inflection property that is satisfied by the curve.

THEOREM 3. *For $i = 1, 2, \ldots, n-1$, if $A_iB_i < 0$, then for $x \in [x_i, x_{i+1}]$, the curve $s(x)$ has exactly one inflection in $[x_i + 0.4573h_i,\ x_{i+1} - 0.4573h_i]$ when $\lambda_i = 0$.*

Proof. Let $A_iB_i < 0$, $\lambda_i = 0$. For $x \in [x_i, x_{i+1}]$, from (9) we have $s''(x) = 0$ when $t = (x - x_i)/h_i$ satisfies

$$\begin{aligned}&[(1-t)^6 + 3(1-t)^5 t + 6(1-t)^4 t^2 - 3(1-t)^2 t^4 - 3(1-t)t^5]A_i\\ &= [3(1-t)^5 t + 3(1-t)^4 t^2 - 6(1-t)^2 t^4 - 3(1-t)t^5 - t^6]B_i.\end{aligned}$$

If we write

$$g(t) = -3(1-t)^5 t - 3(1-t)^4 t^2 + 6(1-t)^2 t^4 + 3(1-t)t^5 + t^6,$$

then $s''(x) = 0$ can be written as $g(t) = g(1-t)|A_i/B_i|$. Thus, we have

$$\begin{aligned}g'(t) &= 3(8t^5 - 20t^4 + 24t^3 - 16t^2 + 8t - 1),\\ g''(t) &= 24[(5t^2 - 5t + 4)(1-t)t + 1],\end{aligned}$$

where the derivatives represent differentiation with respect to t. Obviously, for $t \in [0, 1]$, $g''(t) > 0$ and then $g'(t)$ has only one zero point $t = \underline{t} \approx 0.170013$. Thus, $g(t)$ is decreasing for $t \in [0, \underline{t}]$ and increasing for $t \in [\underline{t}, 1]$. Further $g(0) = 0, g(t^*) = 0, g(t) < 0$ for $t \in (0, t^*)$ and $g(t) > 0$ for $t \in (t^*, 1]$, where $t^* \approx 0.457312$. Symmetrically, $g(1-t) > 0$ for $t \in [0, 1-t^*), g(1-t) < 0$ for $t \in (1-t^*, 1)$, and $g(1-t)$ is deceasing for $t \in [0, 1-t^*]$.

For $t \in [0, 1]$, since $g(t^*) = 0$, $g(1-t^*)|A_i/B_i| > 0$, and $g(1-t^*) > 0$, $g(t^*)|A_i/B_i| = 0$, we can know that $g(t) = g(1-t)|A_i/B_i|$ holds only for one point in $[0.4573, 1 - 0.4573]$. ∎

When $A_iB_i < 0$, we also want to set $\lambda_i = 0$ based on the denominator of $s(x)$. Thus, $Q_i(t) = (1-t)^2 + t^2 \geq 1/2$.

The remaining case we need to discuss is $A_iB_i = 0$. When $A_i = B_i = 0$, $s(x)$ is linear interpolant to (x_i, f_i) and (x_{i+1}, f_{i+1}).

THEOREM 4. *If $A_i \neq 0, B_i = 0$ or $A_i = 0, B_i \neq 0$, then when $sgn(\lambda_i) = sgn(A_i + B_i)$, we have*

$$\lim_{\lambda_i \to \infty} s(x) = (1-t)f_i + tf_{i+1} \tag{11}$$

for $x \in [x_i, x_{i+1}]$, $t = (x - x_i)/h_i$.

Proof. Let $A_i \neq 0, B_i = 0$. Then, for $x \in [x_i, x_{i+1}]$,

$$s(x) = (1-t)f_i + tf_{i+1} - \frac{h_i(1-t)^3 t A_i}{Q_i(t)}.$$

Obviously,

$$\Big|\frac{(1-t)^3 t A_i}{Q_i(t)}\Big| \leq \frac{(1-t)^3 t|A_i|}{(1-t)^3 + (1-t)t^2\lambda_i A_i} \leq \frac{1-t}{2}\sqrt{\frac{A_i}{\lambda_i}}.$$

From this we obtain (11). In the same way, we can obtain (11) when $A_i = 0, B_i \neq 0$. ∎

According to (4), $A_i = 0$ implies $\Delta_{i-1} = \Delta_i$ and $B_i = 0$ implies $\Delta_i = \Delta_{i+1}$. In these cases, we can choose the tension parameter λ_i so that the interpolant (1) approximates the linear interpolant.

Based on Theorems 2, 3, and 4, we give a proper choice of the tension parameter as follows:

$$\lambda_i = \begin{cases} A_i/B_i^2 + B_i/A_i^2 & \text{for } A_iB_i > 0, \\ (A_i + B_i)c_i & \text{for } A_iB_i = 0, \\ 0 & \text{for } A_iB_i < 0, \end{cases} \tag{12}$$

where $c_i > 0$ could be left as user input.

By (12), we can treat data sets for all cases. Often a data set will not be globally convex/concave, but instead switches back and forth between convexity and concavity. Thus, expression (12) is practical.

4. Convergence analysis. Let $f(x)$ be a given function with which $s(x)$ is compared. We have $f(x_i) = f_i$ and write $f_i' = f'(x_i)$, $i = 1, 2, \ldots, n$,

$$\|f^{(r)}\|_{i,j} = \max_{x_i \le x \le x_j} |f^{(r)}(x)|, \quad \delta_i = \min_{x_{i-1} \le x \le x_{i+2}} |f^{(2)}(x)|$$

for all possible r, i, j.

THEOREM 5. *Let $f \in C^4[a, b]$. For $x \in [x_i, x_{i+1}], i = 1, 2, \ldots, n-1$, if $\lambda_i A_i \ge 0$, $\lambda_i B_i \ge 0$, then*

$$\begin{aligned} |f(x) - s(x)| \le{}& \frac{h_i^4}{384}\|f^{(4)}\|_{i,i+1} + \frac{h_i}{4}\max\{|f_i' - d_i|, |f_{i+1}' - d_{i+1}|\} \\ &+ \frac{h_i}{2 + \min\{\lambda_i A_i, \lambda_i B_i\}}|A_i - B_i|\left[\frac{\sqrt{3}}{18} + \frac{1}{16}|\lambda_i(A_i - B_i)|\right]. \end{aligned} \tag{13}$$

Proof. For $x \in [x_i, x_{i+1}]$, $i = 1, 2, \ldots, n-1$, let $t = (x - x_i)/h_i$ and

$$H_3(x) = (1-t)f_i + tf_{i+1} - h_i(1-t)t[(1-t)A_i + tB_i].$$

After some manipulations, we obtain

$$s(x) = H_3(x) + \frac{h_i(1-t)^2t^2}{Q_i(t)}(A_i - B_i)[2t - 1 + \lambda_i(1-t)t(A_i - B_i)].$$

By $f_i' = f'(x_i)$, $i = 1, 2, \ldots, n$, let

$$H_3^*(x) = (1-t)f_i + tf_{i+1} - h_i(1-t)t[(1-t)(\Delta_i - f_i') + t(f_{i+1}' - \Delta_i)].$$

Since $H_3^*(x)$ is the cubic Hermite interpolant to $f(x)$, we have

$$|f(x) - H_3^*(x)| \le \frac{h_i^4}{384}\|f^{(4)}\|_{i,i+1}.$$

It is easy to see that

$$|H_3^*(x) - H_3(x)| \le \frac{h_i}{4}\max\{|f_i' - d_i|, |f_{i+1}' - d_{i+1}|\},$$

$$Q_i(t) = (1-t)^2 + \lambda_i(1-t)t[(1-t)B_i + tA_i] + t^2$$
$$\geq (1-t)t(2 + \min\{\lambda_i A_i, \lambda_i B_i\}),$$

and $|(1-t)t(2t-1)| \leq \sqrt{3}/18$. From these, we obtain (13). ∎

THEOREM 6. *Let $f \in C^4[a,b]$, and let d_i be given by (4) and (10). For $x \in [x_i, x_{i+1}]$, $i = 1, 2, \ldots, n-1$, if $\lambda_i(\Delta_i - \Delta_{i-1}) \geq 0, \lambda_i(\Delta_{i+1} - \Delta_i) \geq 0$, then*

$$|f(x) - s(x)| \leq \frac{h_1^4}{384}\|f^{(4)}\|_{1,2} + \frac{h_1^2}{24}(h_1 + h_2)\|f^{(3)}\|_{1,3} \tag{14}$$

for $x \in [x_1, x_2]$,

$$\begin{aligned} |f(x) - s(x)| \leq{}& \frac{h_i^4}{384}\|f^{(4)}\|_{i,i+1} + \frac{h_i^2}{24}\max\{h_{i-1}, h_{i+1}\}\|f^{(3)}\|_{i-1,i+1} \\ &+ \frac{h_i^2(h_{i-1} + h_i + h_{i+1})}{18(4 + h_i\delta_i|\lambda_i|)}\|f^{(3)}\|_{i-1,i+2}\left[\frac{\sqrt{3}}{3}\right. \\ &\left. + \frac{h_i}{16}(h_{i-1} + h_i + h_{i+1})|\lambda_i|\|f^{(3)}\|_{i-1,i+2}\right] \end{aligned} \tag{15}$$

for $x \in [x_i, x_{i+1}]$, $i = 2, 3, \ldots, n-2$, and

$$|f(x) - s(x)| \leq \frac{h_{n-1}^4}{384}\|f^{(4)}\|_{n-1,n} + \frac{h_{n-1}^2}{24}(h_{n-2} + h_{n-1})\|f^{(3)}\|_{n-2,n} \tag{16}$$

for $x \in [x_{n-1}, x_n]$.

Proof. For (4) and (10), by using a Peano kernel analysis, it can be shown that

$$\begin{aligned} &f_1' - d_1 = \frac{1}{6}h_1(h_1 + h_2)f^{(3)}(\xi_1), \quad \xi_1 \in (x_1, x_3), \\ &f_n' - d_n = \frac{1}{6}h_{n-1}(h_{n-2} + h_{n-1})f^{(3)}(\xi_n), \quad \xi_n \in (x_{n-2}, x_n), \\ &d_1 + d_2 - 2\Delta_1 = 0, \quad d_{n-1} + d_n - 2\Delta_{n-1} = 0, \\ &d_i - f_i' = \frac{1}{6}h_{i-1}h_i f^{(3)}(\xi_i), \quad \xi_i \in (x_{i-1}, x_{i+1}), \\ &d_i + d_{i+1} - 2\Delta_i = \frac{1}{6}h_i(h_{i-1} + h_i + h_{i+1})f^{(3)}(\eta_i), \quad \eta_i \in (x_{i-1}, x_{i+2}), \\ &\Delta_i - \Delta_{i-1} = \frac{1}{2}(h_{i-1} + h_i)f^{(2)}(\zeta_i), \quad \zeta_i \in (x_{i-1}, x_{i+1}), \end{aligned}$$

for $i = 2, 3, \ldots, n-1$. From these, (13), and $A_i = h_i(\Delta_i - \Delta_{i-1})/(h_{i-1} + h_i)$, $B_i = h_i(\Delta_{i+1} - \Delta_i)/(h_i + h_{i+1})$, $A_i - B_i = 2\Delta_i - d_i - d_{i+1}$, it is easy to obtain (14), (15), and (16). ∎

It can be seen from Theorem 6 that $f(x) - s(x) = O(h^3)$ for $x \in [x_1, x_n]$, where $h = \max_{1 \leq i \leq n-1} h_i$.

5. Numerical results. Numerical experiments are conducted by using (4), (10), and the values of λ_i given by (12). Since the influence of c_i in (12) is little, we give the experimental results by taking $c_i = 1$. The algorithm is easy to implement and fast to compute.

We first consider two sets of data to show the convexity-preserving property. The first set of data may be found in [12], and the interpolation curve is shown in Figure 1.

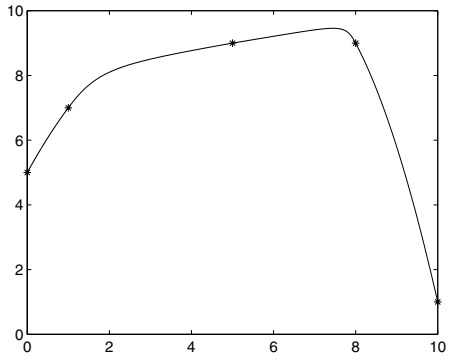


FIG. 1. *Curve on strictly convex data.*

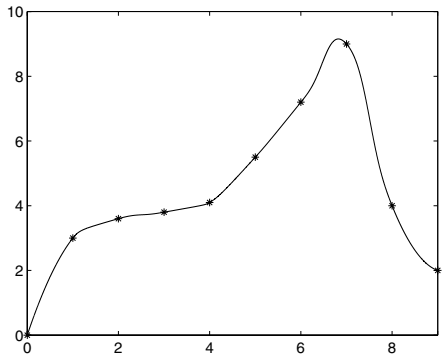


FIG. 2. *Curve on nonstrictly convex data.*

The second set of data may be found in [14], and the interpolation curve is shown in Figure 2. We conclude from the figures that the given method appears to produce visually pleasing curves.

We next consider two sets of data to show the order of convergence of the interpolation method. The first set of data is the set of the points of the curve $f_1(x) = 2\exp(-(x-1.5)^2)$, $0 \le x \le 3$. The second set of data is the set of the points of the curve $f_2(x) = \cos^2 x$, $0 \le x \le \pi/2$. The knots x_i of two examples are taken to be equally spaced. Figures 3 and 4 show the interpolation curves for $n = 10$ for the first and second sets, respectively. The errors are evaluated by $e = \max_i |f(\bar{x}_i) - s(\bar{x}_i)|$ for $\bar{x}_i = x_1 + i(x_n - x_1)10^{-4}, i = 0, 1, \ldots, 10^4$. Table 1 shows the interpolation errors with $n = 11$, 101, 1001, respectively. The errors confirm the expected convergence rates.

6. Conclusions. An explicit representation of a piecewise rational quartic interpolant has been developed, which can preserve the local convexity/concavity of the interpolation data. The given interpolant can be C^2 continuous without solving a linear or nonlinear system of consistency equations for the derivative values at the knots. For C^2 methods that concern the solution of the consistency equations, see [6, 10, 11, 12, 14, 15]. The given interpolant contains a parameter, λ_i, which serves as a tension factor. The proper choice of λ_i to guarantee shape preservation is given.

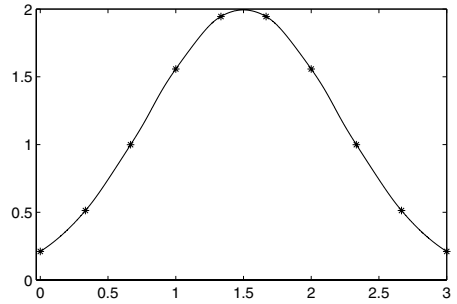


FIG. 3. *Interpolation curve for* $f_1(x)$.

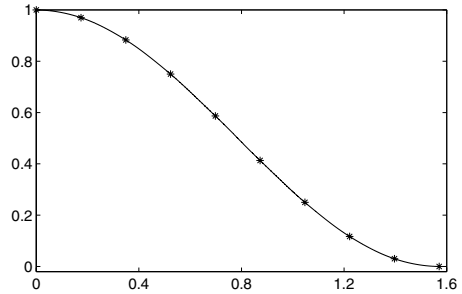


FIG. 4. *Interpolation curve for* $f_2(x)$.

TABLE 1
Interpolation errors for $f_1(x)$ *and* $f_2(x)$.

n	11	101	1001
$f_1(x)$	7.15900×10^{-3}	6.03609×10^{-6}	6.87149×10^{-9}
$f_2(x)$	9.88001×10^{-4}	9.94430×10^{-7}	9.92200×10^{-10}

A large value of λ_i will give approximately the interpolant in [12] or the special form in [11]. For C^2 interpolation, the interpolant has $O(h^3)$ convergence. Numerical examples indicate that the method produces visually pleasing curves.

Acknowledgment. The author thanks the referees for their valuable advice.

REFERENCES

[1] E. PASSOW AND J. A. ROULIER, *Monotone and convex spline interpolation*, SIAM J. Numer. Anal., 14 (1977), pp. 904–909.
[2] F. N. FRITSCH AND R. E. CARLSON, *Monotone piecewise cubic interpolation*, SIAM J. Numer. Anal., 17 (1980), pp. 238–246.
[3] P. COSTANTINI, *On monotone and convex spline interpolation*, Math. Comp., 46 (1986), pp. 203–214.
[4] R. L. DOUGHERTY, A. EDELMAN, AND J. M. HYMAN, *Nonnegativity-, monotonicity-, convexity-preserving cubic and quintic Hermite interpolation*, Math. Comp., 52 (1989), pp. 471–494.

[5] L. L. Schumaker, *On shape preserving quadratic spline interpolation*, SIAM J. Numer. Anal., 20 (1983), pp. 854–864.
[6] J. C. Fiorot and J. Tabka, *Shape-preserving C^2 cubic polynomial interpolating splines*, Math. Comp., 57 (1991), pp. 291–298.
[7] J. W. Schmidt and W. Heß, *An always successful method in univariate convex C^2 interpolation*, Numer. Math., 71 (1995), pp. 237–252.
[8] C. Manni and P. Sablonnière, *Monotone interpolation of order 3 by C^2 cubic splines*, IMA J. Numer. Anal., 17 (1997), pp. 305–320.
[9] J. A. Gregory and R. Delbourgo, *Piecewise rational quadratic interpolation to monotonic data*, IMA J. Numer. Anal., 2 (1982), pp. 123–130.
[10] R. Delbourgo and J. A. Gregory, *C^2 rational quadratic spline interpolation to monotonic data*, IMA J. Numer. Anal., 3 (1983), pp. 141–152.
[11] R. Delbourgo and J. A. Gregory, *Shape preserving piecewise rational interpolation*, SIAM J. Sci. Statist. Comput., 6 (1985), pp. 967–976.
[12] R. Delbourgo, *Shape preserving interpolation to convex data by rational functions with quadratic numerator and linear denominator*, IMA J. Numer. Anal., 9 (1989), pp. 123–136.
[13] C. E. Frost and G. L. Kinzel, *An automatic adjustment procedure for rational spline*, Computers and Graphics, 6 (1986), pp. 171–176.
[14] J. C. Clements, *Convexity-preserving piecewise rational cubic interpolation*, SIAM J. Numer. Anal., 27 (1990), pp. 1016–1023.
[15] R. Delbourgo, *Accurate C^2 rational interpolants in tension*, SIAM J. Numer. Anal., 30 (1993), pp. 595–607.

SIAM J. NUMER. ANAL.
Vol. 46, No. 2, pp. 930–948
© 2008 Society for Industrial and Applied Mathematics

NEW QUADRATURE FORMULAS FROM CONFORMAL MAPS*

NICHOLAS HALE† AND LLOYD N. TREFETHEN†

Abstract. Gauss and Clenshaw–Curtis quadrature, like Legendre and Chebyshev spectral methods, make use of grids strongly clustered at boundaries. From the viewpoint of polynomial approximation this seems necessary and indeed in certain respects optimal. Nevertheless such methods may "waste" a factor of $\pi/2$ with respect to each space dimension. We propose new nonpolynomial quadrature methods that avoid this effect by conformally mapping the usual ellipse of convergence to an infinite strip or another approximately straight-sided domain. The new methods are compared with related ideas of Bakhvalov, Kosloff and Tal-Ezer, Rokhlin and Alpert, and others. An advantage of the conformal mapping approach is that it leads to theorems guaranteeing geometric rates of convergence for analytic integrands. For example, one of the formulas presented is proved to converge 50% faster than Gauss quadrature for functions analytic in an ε-neighborhood of $[-1, 1]$.

Key words. Gauss quadrature, Clenshaw–Curtis quadrature, spectral methods, conformal mapping

AMS subject classifications. 65D32, 30C20

DOI. 10.1137/07068607X

1. Introduction. The Gauss and Clenshaw–Curtis quadrature formulas, as well as other related numerical methods for integration of nonperiodic functions or spectral solution of nonperiodic ODEs or PDEs, all cluster the grid points near the boundaries. Indeed, for any convergent numerical method derived from polynomial interpolation in the grid points, the clustering will be asymptotically the same: on $[-1, 1]$, n points will be distributed with density $\sim n/(\pi\sqrt{1-x^2})$ as $n \to \infty$ [36, Thm. 12.7].

It is well known that this clustering may cause problems. The high density of points near the boundaries may necessitate very small time steps in an explicit time-stepping spectral method or make the matrices involved in implicit time-stepping terribly ill conditioned [12, 49]. At the same time the low density of points in the middle of the domain may force one to use more points than "ought" to be needed to resolve the solution. Compared with equally spaced grids, these clustered grids have $\pi/2$ times coarser spacing in the middle, implying that, for many calculations, n must be about $\pi/2$ times larger than one might expect. In a three-dimensional calculation, a discretization may need $(\pi/2)^3 \approx 4$ times as many grid points as one would like. Such a factor may have a large impact on the cost of linear algebra operations.

In this article we focus on this $\pi/2$ problem and mainly on quadrature, not spectral, methods. The problem arises in purest form if we consider the integral from -1 to 1 of a function f analytic in a narrow strip about the real axis. If f is periodic, we can use the trapezoid rule and get geometric convergence, i.e., convergence at a rate e^{-Cn} for some $C > 0$. If f is not periodic, the trapezoid rule loses its speed, but Gauss quadrature still converges geometrically. However, the constant C is $\pi/2$ times smaller, and thus the convergence rate is $\pi/2$ times slower.

Various authors have proposed methods for countering this effect. We shall suggest a new approach based on conformal mapping, which leads to formulas that

*Received by the editors March 22, 2007; accepted for publication (in revised form) August 31, 2007; published electronically March 5, 2008.
http://www.siam.org/journals/sinum/46-2/68607.html
†Oxford University Computing Laboratory, Wolfson Bldg., Parks Rd., Oxford OX1 3QD, UK (nich@comlab.ox.ac.uk, LNT@comlab.ox.ac.uk).

converge more quickly than Gauss quadrature for many integrands. The practical performance of our methods is not much different from that of other methods in the literature, such as those of Alpert [3] and Kosloff and Tal-Ezer [33] (see section 5). Nevertheless, the conformal mapping idea has advantages. First, it is conceptually clear, simple, and flexible, explaining in a precise way the nature of the $\pi/2$ effect and exactly what improvements may be possible for integrands analytic in specified domains. Second, it leads immediately to theorems on geometric convergence rates for analytic integrands. In the existing literature of related methods, it is hard to find such theorems. An additional advantage, which we shall not discuss further, is that the conformal mapping approach connects these simple quadrature problems to more complicated problems where conformal maps have also proved useful, such as the double exponential quadrature formula and Tee's adaptive rational spectral methods [22, 38, 43, 44, 46, 47, 48].

An outline of the article is as follows. In section 2 we present the conformal transplantation idea. We apply it in section 3 to the particular case of a map to an infinite strip and in section 4 to a simpler variant that may be equally useful in practice. Related work is surveyed in section 5, and theorems about geometric convergence rates are presented in section 6. The Clenshaw–Curtis variant of our formulas is considered in section 7, and we close in section 8 with a summary discussion. The appendix includes a 14-line FFT-based MATLAB code that runs in less than a second for n as high as 10^5 yet usually delivers more accurate integrals than the Gauss rule for each value of n.

2. Transplanted quadrature formulas. Let f be an analytic function on $[-1,1]$ whose integral

$$I = I(f) = \int_{-1}^{1} f(x)\,dx \tag{2.1}$$

we seek to calculate. A general interval $[a,b]$ can be handled by a linear change of variables. All of the methods we shall consider approximate I by sums

$$I_n = I_n(f) = \sum_{k=1}^{n} w_k f(x_k) \tag{2.2}$$

defined by *nodes* $\{x_k\}$ and *weights* $\{w_k\}$. *Polynomial methods* of this kind are based on the following principle: given nodes $\{x_k\}$, the weights $\{w_k\}$ are chosen so that I_n is equal to the integral of the unique degree $n-1$ polynomial interpolant through the data points. In particular, Newton–Cotes quadrature is obtained from equally spaced nodes, Gauss quadrature from the roots of the Legendre polynomial P_n, and Clenshaw–Curtis quadrature from the Chebyshev points $x_k = \cos((k-1)\pi/(n-1))$. Any polynomial method that converges as $n \to \infty$ for all analytic f must have its nodes distributed asymptotically with density $\sim n/(\pi\sqrt{1-x^2})$ as $n \to \infty$. The Gauss and Clenshaw–Curtis formulas have this property and converge for any continuous f. The Newton–Cotes formulas do not and diverge in general even if f is analytic [5, 36, 41].

The standard theorems about convergence of polynomial methods for analytic functions assume that the integrand is analytic in an elliptical region. For any $\rho > 1$, let us define E_ρ to be the open set in the complex plane bounded by the ellipse with foci ± 1 with semiminor and semimajor axis lengths summing to ρ. This ellipse can also be described as the image of the circle of radius ρ about the origin under the

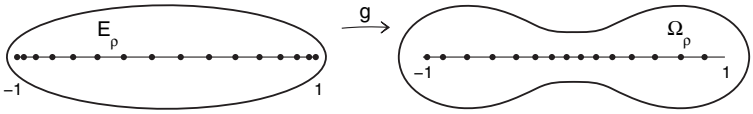


FIG. 2.1. *A transplanted quadrature formula* (2.6) *is defined by an analytic function* g *that maps the elliptical region* E_ρ *into a region* Ω_ρ *containing* $[-1,1]$. *In this illustration the map is* $g(s) = 0.5s + 0.2s^3 + 0.3s^5$ *with* $\rho = 1.4$. *The dots show Gauss points* $x_k \in E_\rho$ *for* $n = 16$ *and their images* $\tilde{x}_k \in \Omega_\rho$.

map $w = (z + z^{-1})/2$. The general idea of the following convergence theorem goes back to Bernstein in 1919 [8], but such results do not appear in many textbooks or monographs, and there is not much uniformity in the constants that one finds on the right-hand side [25, p. 114]. The particular result presented here is due to Rabinowitz [42, eq. (18)]; see also [11, Thm. 90] and [50, Thm. 4.5].

THEOREM 2.1. *If* f *is analytic in* E_ρ, *with* $|f(z)| \le M$ *for some* $\rho > 1$, *then Gauss quadrature converges geometrically with the bound*

$$|I_n - I| \le \frac{64M}{15(1-\rho^{-2})\rho^{2n}} \qquad (n \ge 0). \tag{2.3}$$

When we don't care about constants, we may simply note that Theorem 2.1 ensures geometric convergence at the rate $O(\rho^{-2n})$ as $n \to \infty$.

To experts in approximation theory, the assumption of analyticity in an elliptical region is so familiar as to seem almost beyond question. Nevertheless, the appearance of ellipses in this analysis is driven not by the quadrature formula (2.2) per se but by the decision to derive weights from polynomial interpolation. If we do not insist on polynomials, ellipses cease to have any special status.

We have already observed from the uneven distribution of Gauss and Clenshaw–Curtis nodes that there may be a reason to move beyond polynomials. Here is another argument based on the shape of E_ρ. From an applications point of view, the assumption that f is analytic in E_ρ is unbalanced, for it permits f to be "less analytic" near the ends of the interval, where the ellipse is narrow, than in the middle, where it is wide. Specifically, the Taylor series of f at a point $x \approx \pm 1$ is allowed to have more rapidly increasing coefficients than are permitted at a point $x \approx 0$. This nonuniform analyticity condition leads to Theorem 2.1 and related results for other polynomial quadrature formulas, but it has no intrinsic justification. Further consequences of the same nonuniformity reverberate throughout the field of polynomial approximation theory—for example, in the book by Ditzian and Totik [17]. Of course, there are applications where the functions of interest do have less smoothness near the boundary than in the interior, such as fluid mechanics problems with boundary layers. Even here, however, there is no reason to expect that an ellipse should be exactly the right region to consider. Weak boundary layers might benefit from less grid clustering at boundaries and strong ones from even more.

Our plan is to derive new quadrature formulas (2.2) for $[-1,1]$ by conformally mapping E_ρ to a region Ω_ρ that has straighter sides. We describe the procedure first in general terms; see Figure 2.1.

Let Ω_ρ be an open set in $\mathbb{C}$ containing $[-1,1]$ inside of which the function f is analytic. Let g be an analytic function in E_ρ satisfying

$$g(E_\rho) \subseteq \Omega_\rho, \qquad g(-1) = -1, \quad g(1) = 1. \tag{2.4}$$

Then $g([-1,1])$ is an analytic curve in Ω_ρ, parametrized by $s \in [-1,1]$, that connects -1 to 1. By Cauchy's theorem for analytic functions, the integral of f over this curve is the same as the integral of f over $[-1,1]$ itself. Thus (2.1) can be rewritten

$$I(f) = \int_{-1}^{1} g'(s)\, f(g(s))\, ds. \tag{2.5}$$

If we apply (2.2) to this integral in the s variable, we obtain

$$I_n(\, g' \cdot (f \circ g)) = \sum_{k=1}^{n} w_k\, g'(x_k)\, f(g(x_k)),$$

which can be rewritten in the form (2.2) as

$$\boxed{\tilde{I}_n = \tilde{I}_n(f) = \sum_{k=1}^{n} \tilde{w}_k\, f(\tilde{x}_k), \quad \tilde{w}_k = w_k\, g'(x_k), \quad \tilde{x}_k = g(x_k).} \tag{2.6}$$

TRANSPLANTED QUADRATURE FORMULA

In this article we generally speak of g as a conformal map, which is defined as an analytic function with nonvanishing derivative that maps one region bijectively onto another. In fact, for the theorems we present, g does not actually have to be conformal, merely analytic. Nor does g have to map $[-1,1]$ into itself; it could take complex values. For the formulas of practical interest, however, these extra conditions will usually be satisfied, and if they are, then we have the additional properties that the nodes $\tilde{x}_k$ are real and the weights $\tilde{w}_k$ are real and positive. In some applications to spectral methods, not considered here, conformality is genuinely required.

The following theorem on the convergence of transplanted Gauss quadrature is a corollary of Theorem 2.1.

THEOREM 2.2. *Let f be analytic in a region Ω_ρ containing $[-1,1]$ with $|f(z)| \le M$ for some $\rho > 1$, and let a transplanted Gauss quadrature formula* (2.6) *be defined by some map g satisfying* (2.4). *Then*

$$|\tilde{I}_n - I\,| \le \frac{64 M \gamma}{15(1-\rho^{-2})\rho^{2n}} \qquad (n \ge 0),$$

where $\gamma = \sup_{s \in E_\rho} |g'(s)| \le \infty$.

If $\gamma = \infty$ in this estimate, we can shrink ρ a little bit to make it finite. Thus transplanted Gauss quadrature always converges geometrically if f is analytic on $[-1,1]$.

3. Map to an infinite strip. We now turn to particular choices of the map g that may be useful in practice. Since our aim is to make the sides of Ω_ρ straighter than those of E_ρ, we begin by taking this notion as far as possible and considering the map g that takes E_ρ to an infinite strip Ω_ρ symmetric about the real axis with $g(\pm 1) = \pm 1$ and the ends of the ellipse mapping to $\pm\infty$ (Figure 3.1). These conditions determine g fully. The half-width a of Ω_ρ is not adjustable but is determined by the value of ρ ((3.5) below).

The map of the ellipse to the strip can be constructed in three steps (Figure 3.2). Let F_s denote the upper half of E_ρ, with distinguished boundary points $s_A = 0$, $s_B = 1$, and $s_C = (\rho + \rho^{-1})/2$ (the right end point of E_ρ). The function $u = \sin^{-1} s$ maps F_s to the rectangle F_u of width π and height $\log \rho$ with corresponding boundary

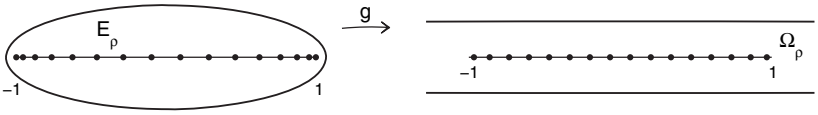


FIG. 3.1. *One choice of Ω_ρ is an infinite strip, shown here again for $\rho = 1.4$. The semiminor axis length of E_ρ is $\sim (\rho - 1)$, whereas the half-width of Ω_ρ is $\sim (2/\pi)(\rho - 1)$, reflecting the $\pi/2$ times weaker analyticity requirement of the transplanted quadrature formula. Note how the Gauss nodes in E_ρ map to nodes in Ω_ρ that are so close to equally spaced that they appear almost exactly even. For more on this spacing see Figure* 5.2.

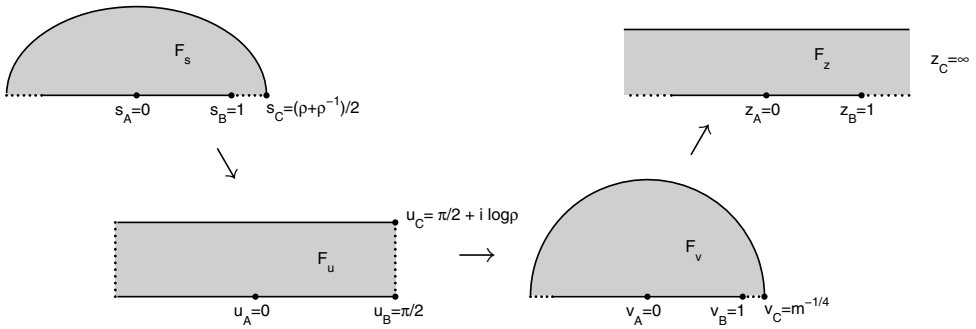


FIG. 3.2. *Construction of the strip map in three steps $s \to u \to v \to z$. The upper half of each region is shown. A large value of ρ has been used for clarity ($\rho = 2.3$), since, as ρ decreases to* 1*, the segment from v_B to v_C shrinks exponentially.*

points $u_A = 0$, $u_B = \pi/2$, and $u_C = \pi/2 + i \log \rho$. We now map F_u to a half-disk by a Jacobi elliptic sine function sn with parameter m [1, 16, 20]. Specifically, $v = \mathrm{sn}(2Ku/\pi \,|\, m)$ maps F_u to the upper half-disk F_v about the origin of radius $m^{-1/4}$, with $v_A = 0$, $v_B = 1$, and $v_C = m^{-1/4}$. Here, following standard notation, the parameter m with $0 < m < 1$ is chosen such that the associated elliptic integrals K' and K satisfy $K'/K = 4 \log(\rho)/\pi$. Given ρ, the appropriate value of m can be calculated by the rapidly convergent series expression given in [24]:

$$m^{1/4} = 2\sum_{j=1}^{\infty} \rho^{-4(j-\frac{1}{2})^2} \Big/ \left(1 + 2\sum_{j=1}^{\infty} \rho^{-4j^2}\right). \tag{3.1}$$

Finally $z = \tanh^{-1}(m^{1/4}v)/\tanh^{-1}(m^{1/4})$ maps F_v to F_z, the upper half of an infinite strip about the real axis with $z_A = 0$, $z_B = 1$, and $z_C = +\infty$. By combining the steps, we have the ellipse $\to$ strip map

$$g(s) = \tanh^{-1}\left(m^{1/4}\,\mathrm{sn}\left(\frac{2K}{\pi \sin^{-1}(s) \,|\, m}\right)\right) \Big/ \tanh^{-1}(m^{1/4}) \,. \tag{3.2}$$

The same formula is valid also below the real axis (by the Schwarz reflection principle), and thus g as given by (3.2) maps all of E_ρ to all of Ω_ρ. For the transplanted quadrature formula (2.6) we also require the derivative of the map function, which for $s \in (-1, 1)$ is given by

$$g'(s) = \frac{2Km^{1/4}}{\pi\sqrt{1-s^2}} \frac{\mathrm{cn}(\omega|m)\,\mathrm{dn}(\omega|m)}{(1 - m^{1/2}\,\mathrm{sn}^2(\omega|m))} \Big/ \tanh^{-1}(m^{1/4}), \tag{3.3}$$

where $\omega = 2K \sin^{-1}(s)/\pi$. For the end points $s = \pm 1$, with the aid of L'Hopital's rule

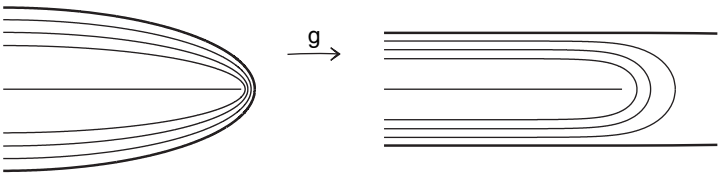


FIG. 3.3. *Though g maps E_ρ to the unbounded strip Ω_ρ, it maps smaller ellipses within E_ρ to regions that hug the interval $[-1,1]$ closely. Because of this property, f does not have to be analytic far down the strip for transplanted Gauss quadrature to be effective. Also, since the subregions have nearly straight sides, a value of ρ chosen for the strip of a particular half-width will also be effective for integrands whose true regions of analyticity are narrower.*

one can derive the formula

$$g'(\pm 1) = 4K^2\pi^{-2}m^{1/4}(1+m^{1/2}) \Big/ \tanh^{-1}(m^{1/4}) \,, \tag{3.4}$$

which we shall need for the Clenshaw–Curtis variant in section 7.

The half-width a of Ω_ρ is

$$a = \frac{\pi}{4\tanh^{-1}(m^{1/4})} < \frac{2}{\pi}(\rho-1), \tag{3.5}$$

with $a \sim (2/\pi)(\rho-1)$ as $\rho \to 1$ (we have verified this inequality numerically, and no doubt it could be proved analytically). The half-width of E_ρ in the same limit is $\sim(\rho-1)$. Thus the transplanted formula needs only $2/\pi$ times as wide a strip of analyticity to achieve the same convergence rate, as is confirmed by the following theorem. Further convergence results for this method are given in section 6.

THEOREM 3.1. *Let f be analytic in the strip Ω_ρ about $\mathbb{R}$ of half-width $(2/\pi)(\rho-1)$ for some $\rho > 1$. Let f be integrated by the transplanted Gauss quadrature formula* (2.6) *associated with the map* (3.2) *from E_ρ to Ω_ρ. Then, for any $\tilde\rho < \rho$,*

$$\tilde I_n - I = O(\tilde\rho^{-2n}) \qquad (n \to \infty). \tag{3.6}$$

Proof. The inequality (3.5) implies that $f(x)$ is analytic in the strip of half-width a, and therefore $f(g(s))$ is analytic in E_ρ. The constant γ of Theorem 2.2 is infinite for this map g, so we do not quite get $O(\rho^{-2n})$ convergence, and for this reason we have not assumed that f is bounded either. For any $\tilde\rho < \rho$, however, Theorem 2.1 still applies to the integrand $g'(s)f(g(s))$ of (2.5), which will be analytic and bounded in this smaller ellipse. This implies (3.6). ∎

To make our domain narrower in the imaginary direction by a factor of $2/\pi$, we have lengthened it in the real direction by a factor of ∞! This may seem a dubious improvement, and, indeed, it is a very strong condition that f must be analytic throughout an infinite strip. However, this is not a serious issue in practice because if we consider slightly smaller $\tilde\rho$-ellipses within E_ρ, we find that they map under g to domains about $[-1,1]$ that are much shorter. Thus little is lost if f is analytic near $[-1,1]$ but not very far down the real axis. Figure 3.3 illustrates this effect.

Let us see how strip-transplanted Gauss quadrature performs in practice. We do not normally recommend tuning ρ to the integrand at hand, for we are hardly likely to beat adaptive quadrature methods at their own game. Instead the aim is to derive a fixed family of formulas that perform well, if not quite optimally, and with this in mind we start by fixing the value $\rho = 1.4$ arbitrarily, as in Figures 2.1–3.1. Figure 3.4 shows

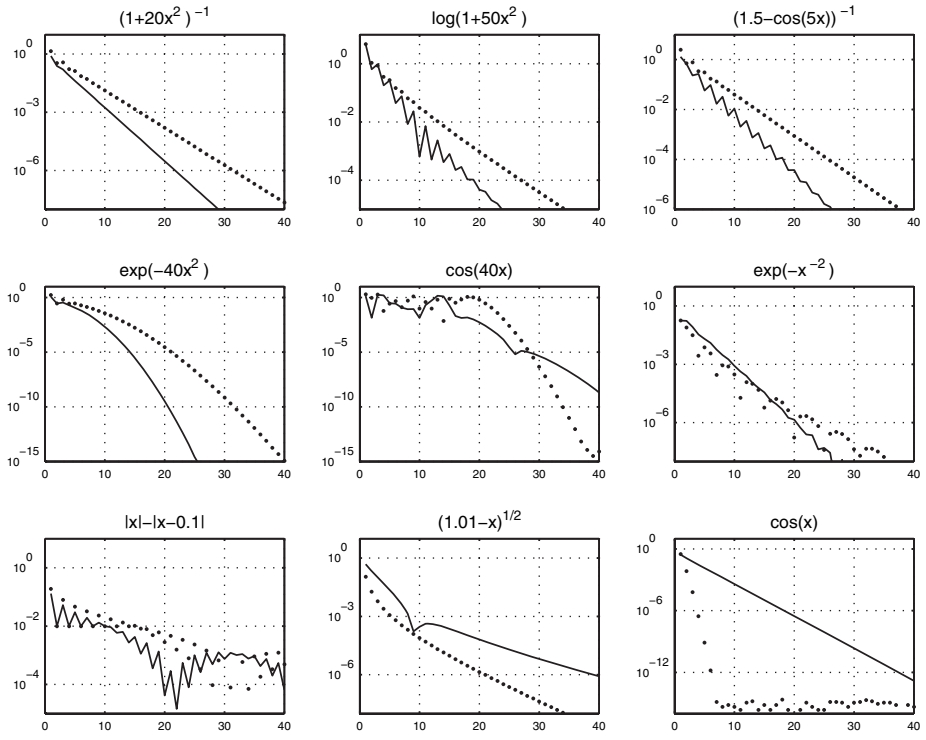


FIG. 3.4. *Comparison of Gauss quadrature (dots) against Gauss quadrature transplanted to an infinite strip (solid line) for nine integrands on* $[-1,1]$*, with* $\rho = 1.4$*. In a number of cases the transplanted formulas converge about* 1.4 *times faster.*

that, for many integrands, the result is a clear improvement over Gauss quadrature. (For this relatively large value of ρ, the improvement factor is about 1.42 rather than $\pi/2 \approx 1.57$.) This occurs, for example, for the first three integrands, which have poles or branch points close to $[-1,1]$ in the complex plane. A similar speedup is observed for the fourth function $\exp(-40x^2)$, which has no singularities but grows very rapidly as x moves away from the real axis. For the fifth function $\cos(40x)$, the transplanted quadrature formula begins to converge sooner than the untransplanted one since it resolves the oscillations sooner; once the error falls below 10^{-5}, the untransplanted formula overtakes it because the function is entire. The sixth integrand $\exp(-x^{-2})$ is C^∞ but not analytic; here we see comparable performance of the two methods. The same is observed for $|x| - |x - 0.1|$, which is continuous but not differentiable. For the eighth integrand $\sqrt{1.01 - x}$, the untransplanted Gauss rule wins because of its weaker analyticity requirement near ± 1. Finally, for the entire function $\cos(x)$, the transplanted rule loses by a substantial factor because it is stuck at the particular convergence rate $O(\rho^{-2n})$, with $\rho = 1.4$, since g is singular at the end points of this ellipse. The standard Gauss formula, by contrast, is not tied to any particular ellipse of analyticity and thus does very well in a case like this where the integrand is analytic and not too large in a sizable region around $[-1,1]$.

In fact, the convergence rate for the transplanted Gauss formula applied to $\cos(x)$ in Figure 3.4 has rather little to do with $\cos(x)$—the rate would be almost exactly the same for the integrand $f(x) = 1$! In general, nonpolynomial quadrature formulas do not integrate constants exactly. This perhaps startling property can be removed

by a scaling adjustment to our formula (2.6), whereby we multiply the weights $\tilde{w}_k$ by a constant so as to make them sum to 2. Since the transplanted quadrature formula integrates the constant function with exponential accuracy, the adjustment factor will be exponentially close to 1, and there will be no effect on the asymptotic convergence rates of Theorem 3.6 or Theorems 6.1 and 6.2 to follow. As a practical matter, the adjustment sometimes improves the behavior of our formulas slightly, and perhaps it is worth making.[1]

To implement the transplanted formula (2.6), we need to be able to compute g and g'. The following MATLAB code segment does this for the strip map (3.2). The standard MATLAB functions `ellipke` and `ellipj` used here to compute Jacobi elliptic functions are restricted to real arguments; for the complex values needed in Figures 3.1 and 3.3, we replaced `ellipj` by the function `ellipjc` from Driscoll's Schwarz–Christoffel Toolbox [19]. The function `gauss`, from [49], computes Gauss quadrature nodes and weights.

```
f = @(x) 1./(1+20*x.^2);                % change this for other integrands
[s,w] = gauss(n);                       % Gauss nodes and weights
[g,gprime] = strip(s);                  % g and g'
In = (w.*gprime')*f(g);                 % the integral

function [g,gprime] = strip(s)          % change this for a different map g
rho = 1.4;                              % this can be adjusted
num = 0; den = 0;
for j = 1:round(.5+sqrt(10/log(rho)))   % given rho, find m
  num = num + rho^(-4*(j-.5)^2);
  den = den + rho^(-4*j^2);
end
m4 = 2*num/(1+2*den); m = m4^4;         % m^{1/4} and m
K = ellipke(m);                         % Jacobi elliptic parameter
u = asin(s);
[sn,cn,dn] = ellipj(2*K*u/pi,m);        % Jacobi eliptic function
duds = 1./sqrt(1-s.^2);
dvdu = (2*K/pi)*cn.*dn;
dgdv = (m4./(1-m4.^2*sn.^2))/atanh(m4);
g = atanh(m4*sn)/atanh(m4);             % g
gprime = dgdv.*dvdu.*duds;              % g'
```

If ρ is close to 1 (smaller than about 1.1), the code just given suffers from numerical instability. The appendix outlines what can be done in such cases and offers an elementary map slightly different from (3.2) that as a practical matter may be superior.

4. A simpler conformal map. We now put aside the infinite strip and consider another idea for selecting the function g. Suppose we set out to cancel the clustering of Gauss or Clenshaw–Curtis points exactly. The function $g(s) = (2/\pi)\sin^{-1}(s)$ would achieve this, and the factor $2/\pi$ makes it plain that the result would be a grid denser by that factor in the middle of the interval. This choice of g would be useless, however, because $\sin^{-1}(s)$ has singularities at ± 1, so we would have no ellipse of analyticity at all for the transplanted integrand.

[1] Relatedly, one of the anonymous referees of this paper as originally submitted for publication pointed out that our transplanted n-point formulas, unlike polynomial methods, could in theory have arbitrarily large errors even for functions with a bounded kth derivative for any $k > 0$.

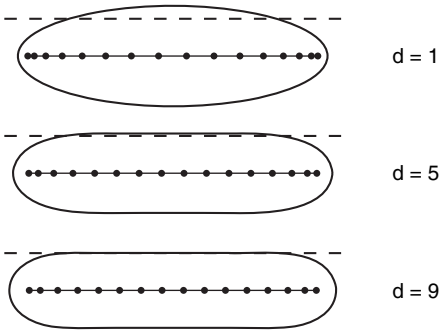


FIG. 4.1. *"Sausages" Ω_ρ corresponding to the maps* (4.1), (4.2), *and* (4.3), *with* $\rho = 1.4$. *The dashed line in each image marks the target height* $(2/\pi)(\rho - 1)$.

Instead we propose taking the terms through degree d of the Taylor series

$$\sin^{-1} s = s + \frac{1}{6}s^3 + \frac{3}{40}s^5 + \frac{5}{112}s^7 + \frac{35}{1152}s^9 + \cdots$$

and then normalizing so that $g(\pm 1) = \pm 1$ as required by (2.4). Taking $d = 1$, 5, or 9 gives the following choices, respectively:

$$g(s) = s, \tag{4.1}$$

$$g(s) = \frac{1}{149}(120s + 20s^3 + 9s^5), \tag{4.2}$$

$$g(s) = \frac{1}{53089}(40320s + 6720s^3 + 3024s^5 + 1800s^7 + 1225s^9). \tag{4.3}$$

Figure 4.1 shows the corresponding "sausage domains" Ω_ρ for $\rho = 1.4$.[2]

To implement transplanted Gauss quadrature of degree d (odd), we can replace the function `strip` in the code above by the following function `sausage`. The line with the `cumprod` command computes the Taylor series coefficients of $\sin^{-1}(x)$.

```
function [g,gprime] = sausage(s)
d = 9;                                          % this can be adjusted
c = zeros(1,d+1);
c(d:-2:1) = [1 cumprod(1:2:d-2)./cumprod(2:2:d-1)]./(1:2:d);
c = c/sum(c); g = polyval(c,s);
cp = c(1:d).*(d:-1:1); gprime = polyval(cp,s);
```

Figure 4.2 shows the performance of this method with $d = 9$ for various integrands. On the whole the convergence is much like that with the strip map, though somewhat slower for the first few examples and faster for the last few. With $f(x) = \cos x$, for example, we do much better now since g has no singularity to limit the convergence rate. Although it is entire, however, g still grows rapidly away from $[-1, 1]$, and that is why this map still does not match untransplanted Gauss quadrature for $\cos(x)$.

In section 6 we give a theorem that establishes that Gauss quadrature transplanted by (4.3) is 30% faster than standard Gauss quadrature for functions analytic in an ε-neighborhood of $[-1, 1]$.

[2]Previous authors have investigated the conformal mapping of a "bratwurst," but their maps concern exteriors rather than interiors and the applications are different [32].

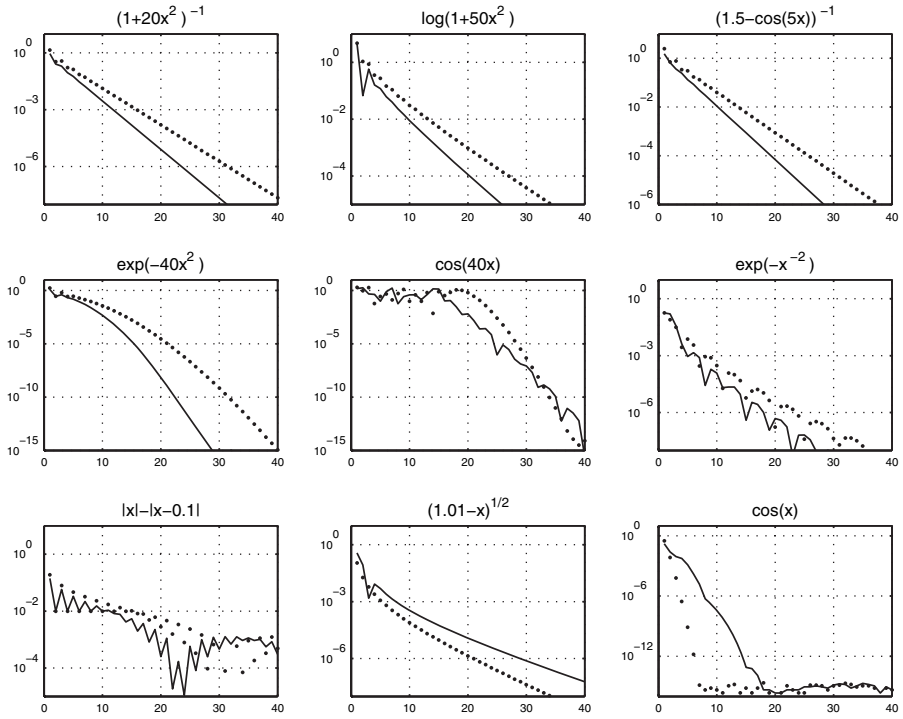


FIG. 4.2. *Comparison of Gauss quadrature (dots) against transplanted Gauss quadrature for the degree* 9 *polynomial* (4.3) *(solid line), for the same nine integrands as in Figure* 3.4. *Again in many cases the transplanted formulas converge faster.*

5. Related work. A number of ideas have been proposed over the years for combating the $\pi/2$ problem, for both quadrature and spectral methods, which might be classified as follows: (i) end point corrections, (ii) alternatives to polynomials, (iii) transplantation. Here is a brief summary of some of these developments.

The end point corrections idea starts from the observation that the trapezoid rule would achieve the ideal convergence rate if only the integrand were periodic. Accordingly, one might make some adjustments at the boundaries to achieve this effect, at least up to a prescribed order. The Euler–Maclaurin formula suggests one approach: one can add to the quadrature formula certain linear combinations of odd-order derivatives of the integrand at the end points, chosen to annihilate the error up to any prescribed power of the grid size. To get a quadrature rule based on function values only (i.e., without derivatives), one can then replace these terms by one-sided finite difference approximations of the appropriate orders. The resulting quadrature rules are known as *Gregory's formulas* [11, 30, 35], and they are effective up to medium orders of accuracy, at which point they encounter the difficulties caused by exponentially large coefficients of alternating signs. An improvement on Gregory formulas capable of achieving exponential convergence has been proposed by Alpert [3], building upon ideas of Kapur and Rokhlin [31]. Alpert's method goes beyond the equispaced grid to include a small number of extra Gauss-like nodes near the end points, now with weights guaranteed to be positive. High-precision computations are required in advance to obtain the nodes and weights, but once they are known the method is highly effective.

By alternatives to polynomials we mean the explicit construction of orthogonal bases that are nonpolynomial and have more uniform behavior over the interval of interest. An important set of functions in this connection is the *prolate spheroidal wave functions*, which were introduced in the 1960s by Slepian and Pollak [45]. These functions have excellent resolution properties, and more recent authors have shown their power for a variety of computational applications [9, 10, 13, 53]. In this literature, rather than derive theorems about convergence for functions analytic in specified domains as we are about to do, it is customary to quantify the matter of resolution by considering applications to band-limited functions.

Finally there is the transplantation idea, the basis of the present article. A notable contribution in this area is a forty-year-old theoretical paper by Bakhvalov [4]. What is the optimal family of quadrature formulas, Bakhvalov asks, for the set of functions analytic and bounded by M in a given complex region Ω containing $[-1,1]$? His first step is to transplant the problem by a conformal map g to an ellipse E_ρ. This is very close to what we have done but with a difference. In the ellipse, the new quadrature problem has a weight g'. Since his aim is to investigate optimal formulas, Bakhvalov now considers the Gauss formulas associated with this weight, which he shows are in a sense optimal. By contrast we have used the unweighted Gauss formula, including g' instead as part of the integrand. We presume that this simpler procedure does not hurt the convergence rate much in practice, but we have not investigated this matter. Bakhvalov's paper is full of interesting ideas, and it has led to subsequent developments in the theoretical quadrature literature by Petras [40] and other authors [23, 28, 34]. So far as we are aware, however, this collection of publications has not been concerned with the $\pi/2$ phenomenon, nor with particular choices of Ω, and does not propose actual quadrature formulas for numerical use.

A more practically oriented transplantation idea was introduced in the spectral methods literature by Kosloff and Tal-Ezer in 1993 [33]. Like the conformal maps proposed in the last section, the Kosloff–Tal-Ezer map is derived from the inverse sine function. Their transformation is

$$g(s) = \frac{\sin^{-1}(\alpha s)}{\sin^{-1}(\alpha)} \tag{5.1}$$

for a value of α slightly less than 1. Various methods for choosing α are considered both in their original paper and in various subsequent works by other authors [2, 7, 15, 18, 29, 37].

The Kosloff–Tal-Ezer map is quite different from those of sections 3 and 4 in original concept but rather similar in practice. These authors, and their successors, do not consider g as a conformal map, and they do not present theorems about geometric rates of convergence for analytic functions. Nevertheless the function (5.1) has the familiar effect of mapping E_ρ to a region with straighter sides, with the usual consequence that the nodes are distributed more evenly in the interior and clustered less near ± 1. To investigate a Kosloff–Tal-Ezer map in the framework established in this paper, we might begin by setting $\rho = 1.4$ and drawing a plot analogous to those of Figures 2.1, 3.1, and 4.1. The largest choice of α for which g will be analytic in E_ρ is $\alpha = 2/(\rho + \rho^{-1}) \approx 0.9459$. Figure 5.1 shows the result.[3]

The node spacings for the various transformations we have considered are shown (for $n = 24$) in Figure 5.2.

[3]In [29] Hesthaven, Dinesen, and Lynov consider a spectral method with $\alpha = \cos(1/2) \approx 0.88$, corresponding to $\rho \approx 1.69$. In most other spectral methods papers that make use of the Kosloff–Tal-Ezer mapping, α is chosen to increase toward 1 as n increases.

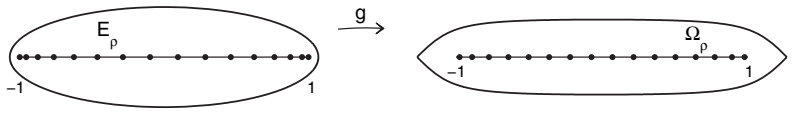


FIG. 5.1. *The Kosloff–Tal-Ezer map* (5.1) *for* $\rho = 1.4$ *and* $\alpha = 2/(\rho + \rho^{-1}) \approx 0.9459$. *For the given choice of* ρ, *this is the largest possible* α *for which* g *is analytic in* E_ρ.

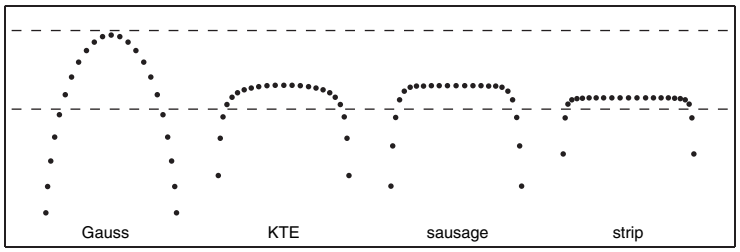


FIG. 5.2. *Comparison of node spacings with* $n = 24$ *for Gauss quadrature, the Kosloff–Tal-Ezer (KTE) map* (5.1) *with* $\rho = 1.4$ *and* $\alpha = 2/(\rho + \rho^{-1})$, *the polynomial* (4.3) *of degree* 9, *and the strip map* (3.2) *with* $\rho = 1.4$. *Dots appear at horizontal positions* $(x_j + x_{j+1})/2$ *and vertical positions* $(x_{j+1} - x_j)/2$. *The dashed lines correspond to equally spaced points and the same times* $\pi/2$.

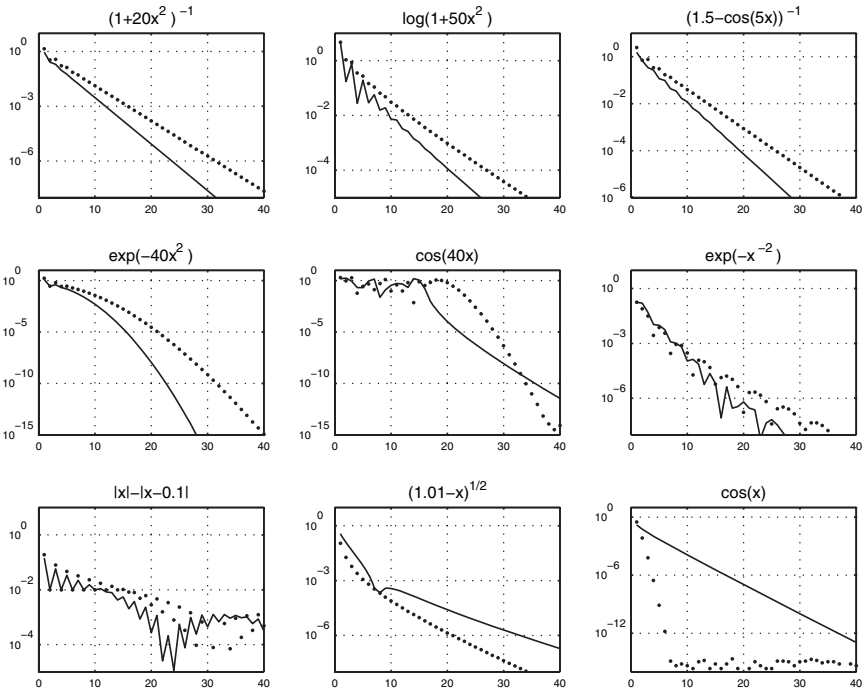


FIG. 5.3. *Another comparison of Gauss (dots) and transplanted Gauss (solid line) quadrature for the integrands of Figures* 3.4 *and* 4.2, *now for the Kosloff–Tal-Ezer map* (5.1) *with* $\rho = 1.4$ *and* $\alpha = 2/(\rho + \rho^{-1}) \approx 0.9459$.

If we apply the Kosloff–Tal-Ezer map to the same nine test integrands as before, we get the curves shown in Figure 5.3. These results are good. Nevertheless, the strip-transplanted formula of Figure 3.4 does better for the first four functions, and

the sausage-transplanted formula of Figure 4.2 does as well or better for the last five. There are parameters in all of these methods that might be tuned to enhance their performance, so one cannot draw fine judgments from these simple experiments. The proper lesson is that all of these methods are capable of outperforming Gauss quadrature by 20%–50% for many integrands. Let us now make this observation more precise.

6. Convergence theorems for analytic integrands. An advantage of the conformal mapping point of view is that it leads readily to convergence theorems for analytic integrands. The trick is to decide which analyticity assumptions and conformal maps g to consider, for the number of possibilities is unlimited. Rather than explore this terrain thoroughly, we shall offer a few representative choices. Our class of integrands will be as follows. For any $\varepsilon > 0$:

$$\mathcal{A}(\varepsilon) = \text{set of functions analytic in the open } \varepsilon\text{-neighborhood of } [-1,1]. \tag{6.1}$$

Now the ellipse with foci ± 1 whose semiminor axis length is ε has parameter $\rho = \varepsilon + \sqrt{1+\varepsilon^2} = 1 + \varepsilon + O(\varepsilon^2)$. Since $\rho > 1+\varepsilon$, any $f \in \mathcal{A}(\varepsilon)$ is bounded in the ellipse $E_{1+\varepsilon}$, and thus by Theorem 2.1, Gauss quadrature will achieve $I_n - I = O((1+\varepsilon)^{-2n})$ as $n \to \infty$ for any $f \in \mathcal{A}(\varepsilon)$. On the other hand, since $\rho \sim 1+\varepsilon$ as $\varepsilon \to 0$, it will do no better than this in general as $\varepsilon \to 0$. By contrast, the methods of the last three sections, with the parameters used there, converge at least 40%, 30%, and 30% faster, respectively.

THEOREM 6.1. *Let the transplanted Gauss quadrature formula* (2.6) *be applied to a function $f \in \mathcal{A}(\varepsilon)$. The following statements pertain to the limit $n \to \infty$:*

For the strip map (3.2) *with $\rho = 1.4$, for any $\varepsilon \le 0.24$: $\tilde{I}_n - I = O((1+1.4\varepsilon)^{-2n})$.*

For the polynomial (4.3) *with $d = 9$, for any $\varepsilon \le 0.8$: $\tilde{I}_n - I = O((1+1.3\varepsilon)^{-2n})$.*

For the KTE map (5.1) *with $\rho = 1.4$, for any $\varepsilon \le 0.3$: $\tilde{I}_n - I = O((1+1.3\varepsilon)^{-2n})$.*

Proof. Consider, for example, the first assertion, that the map (3.2) with $\rho = 1.4$ has convergence rate $O((1+1.4\varepsilon)^{-2n})$. By Theorem 2.1, this conclusion will be valid if for all $\varepsilon \le 0.24$ the function g is analytic in the elliptical region $E_{1+1.4\varepsilon}$ and maps the ellipse of parameter $1+1.4\varepsilon$ into the open ε-neighborhood of $[-1,1]$. We know that g is analytic throughout $E_{1.4}$, so the first condition holds since $1 + (1.4)(0.24) < 1.4$. The second condition can be verified numerically by plotting the image of the ellipse and the boundary of the ε-neighborhood for various ε and verifying that the first is inside the second; of course, we have chosen parameters in the theorem to make this true. ∎

The speedups of Theorem 6.1 are not particularly close to the limiting value of $\pi/2$ that can be achieved as $\varepsilon \to 0$. We now give another result that comes closer to this limit. This time we modify the Gauss quadrature estimate $O((1+\varepsilon)^{-2n})$ by improving the exponent rather than the factor multiplying ε.

THEOREM 6.2. *Let the transplanted Gauss quadrature formula* (2.6) *be applied to a function $f \in \mathcal{A}(\varepsilon)$ for any $\varepsilon \le 0.05$. For the strip map* (3.2) *with $\rho = 1.1$, $\tilde{I}_n - I = O((1+\varepsilon)^{-3n})$ as $n \to \infty$.*

The proof is as before, combining Theorem 2.1 with a numerical verification for the particular map g. Similarly, it can be shown that one gets $\tilde{I}_n - I = O((1+\varepsilon)^{-2.7n})$ for $\varepsilon \le 0.1$ for the strip map (3.2) with $\rho = 1.4$, the higher-degree analogue of the polynomial (4.3) with $d = 17$, or the KTE map (5.1) with $\rho = 1.2$, and $\tilde{I}_n - I = O((1+\varepsilon)^{-2.5n})$ for $\varepsilon \le 0.3$ for the strip map (3.2) with $\rho = 1.5$, the polynomial (4.3) with $d = 9$, or the KTE map (5.1) with $\rho = 1.4$.

7. Clenshaw–Curtis variant. Among polynomial quadrature methods, Gauss quadrature is optimal from the point of view of degree of polynomials integrated exactly, namely, $2n-1$ for the formula (2.2). It also has an elegant convergence theory for analytic integrands, as represented by Theorem 2.1 and the further theorems we have derived from it. Thus it is natural that, in the experiments of this paper until now, we have considered Gauss quadrature and its transplanted alternatives.

Gauss quadrature has the disadvantage, however, that it takes $O(n^2)$ work and memory to compute Gauss nodes and weights by the standard algorithm of Golub and Welsch [27]. In some applications this is not an issue, because either n is small or the nodes and weights are precomputed. In others, it is troublesome indeed. Certainly one would rarely use a Gauss quadrature formula with thousands of points.

An easy alternative is Clenshaw–Curtis quadrature, which is readily implemented via the fast Fourier transform in $O(n \log n)$ operations [14, 26]; now there is no problem if n is 10^4 or 10^5 [6, 50]. All of the conformal transplantation ideas discussed in this article can be applied to Clenshaw–Curtis as well as Gauss formulas, with comparable effect: a speedup for many integrands by a factor approaching $\pi/2$. A Clenshaw–Curtis code segment for $(n+1)$-point integration of f over $[-1,1]$ can be written like this [50]:

```
s = cos((0:n)'*pi/n);
[g,gprime] = strip(s);
fx = f(g).*gprime/(2*n);
h = real(fft(fx([1:n+1 n:-1:2])));
a = [h(1); h(2:n)+h(2*n:-1:n+2); h(n+1)];
w = 0*a'; w(1:2:end) = 2./(1-(0:2:n).^2);
In = w*a;
```

One's first expectation is that Clenshaw–Curtis convergence rates should generally be about half those of Gauss, for both the pure and the transplanted variants. For example, the n-point Clenshaw–Curtis formula exactly integrates polynomials only up to degree n, not $2n-1$, and similarly, a result like Theorem 2.1 holds with $O(\rho^{-2n})$ reduced to $O(\rho^{-n})$. Thus one might expect that a transplanted Clenshaw–Curtis formula should converge about $2 \times 2/\pi = 4/\pi$ times more slowly than untransplanted Gauss quadrature.

In actuality, however, Clenshaw–Curtis formulas are often about as accurate as Gauss, for the values of n of interest, despite what the standard theorems might lead one to expect. This effect was observed by O'Hara and Smith in the 1960s [39] and has recently been investigated further [21, 50, 52]. We find the same happy surprise with transplanted Clenshaw–Curtis quadrature. We shall make no attempt to add more to the theoretical discussion of this effect but just present two computational illustrations. The first is Figure 7.1, another repetition of Figure 3.4 but now comparing pure Gauss quadrature against strip-transplanted Clenshaw–Curtis, again for $\rho = 1.4$. The latter does very well—about as well as transplanted Gauss. Indeed it is only for the function $\exp(-40x^2)$ where we have lost much in moving from transplanted Gauss to transplanted Clenshaw–Curtis: the convergence rate starts out as before but then cuts abruptly in half at an error level of about 10^{-5}. This kink in the Clenshaw–Curtis convergence curve for analytic functions is explained in [21, 50, 52].

Our second example is a complicated function $f(x)$ on $[-1,1]$ considered in [51], defined by the initial conditions $f = \beta = \sin(10x)$ followed by 15 steps of the iteration $\beta = 3(1-2\beta^4)/4$, $f = f+\beta$. (In [51] the initial condition is $\sin(\pi x)$, but we change this to $\sin(10x)$ to break the periodicity.) Figure 7.2 illustrates f and the convergence of

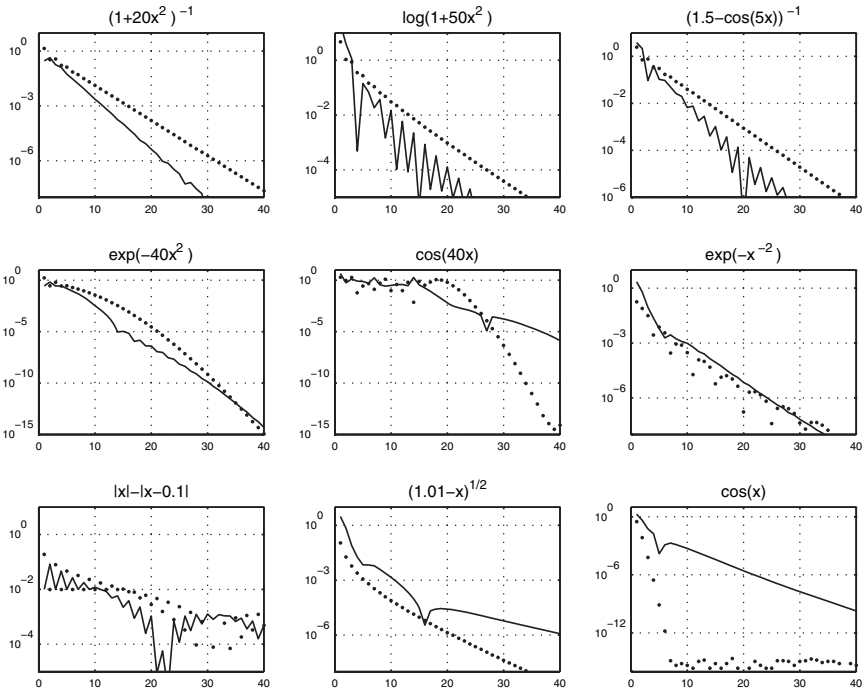


FIG. 7.1. *Another repetition of Figure* 3.4 *but with the solid curve corresponding now to strip-transplanted Clenshaw–Curtis rather than Gauss quadrature. The results are much the same as in Figure* 3.4*, the main exception being* $\exp(-40x^2)$*, where we now see a convergence rate that cuts in half at around* $n = 14$*. Evidently transplanted Clenshaw–Curtis quadrature, quite apart from its much greater speed of implementation, is often more accurate than Gauss quadrature as well.*

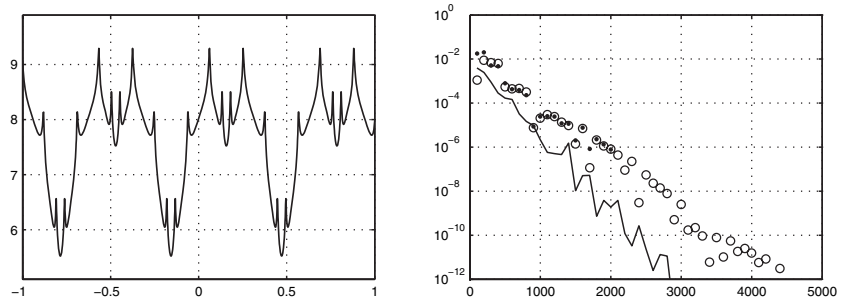


FIG. 7.2. *On the left, the function* $f(x)$ *adapted from* [51]*. On the right, convergence of Gauss (dots), Clenshaw–Curtis (circles), and strip-transplanted Clenshaw–Curtis formulas (solid line) with* $\rho = 1.1$ *for this function. Data are plotted corresponding to values of* n *equal to multiples of* 100 *(up to* $n = 2000$ *for Gauss quadrature). There is no significant difference between the Gauss and Clenshaw–Curtis formulas, and the transplanted formula beats them both by about* 50%*, as one would expect for Gauss quadrature from Theorem* 6.2*.*

quadrature formulas to $I \approx 15.3198135546173$. Strip-transplanted Clenshaw–Curtis quadrature with $\rho = 1.1$ beats untransplanted Clenshaw–Curtis, which is more or less the same as Gauss. This function f is entire, but it is very ill behaved outside a small region surrounding $[-1, 1]$, as illustrated in Figure 7.3. For example, the maximum value of $|f(z)|$ for $\operatorname{Im} z = 0.004$ is about 9.5, for $\operatorname{Im} z = 0.006$ it is about 3236, for $\operatorname{Im} z = 0.008$ it is about 10^{41}, and for $\operatorname{Im} z = 0.01$ it is about 10^{263}.

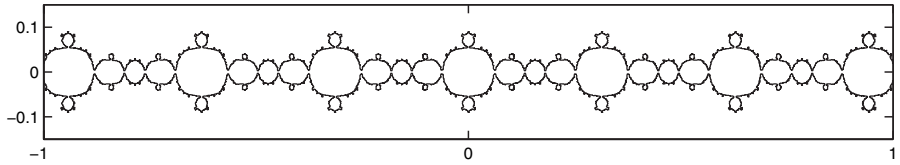


FIG. 7.3. *Although entire, the function f of Figure* 7.2 *is ill-behaved away from a strip of width about* 0.003 *surrounding* $[-1,1]$. *This plot shows the region around* $[-1,1]$ *in which* $|f(x)| < 1000$. *The boundary has a fractal structure* 15 *levels deep.*

8. Discussion. The idea of transforming the integrand in a numerical quadrature rule is an old one. What is new in this paper has been to choose transformations based on explicit conformal maps and to thereby derive specific quadrature formulas that often converge faster than Gauss quadrature. Numerical experiments justifying this claim were presented in sections 3–5 and 7 and theorems in sections 3 and 6. The theorems apply to analytic integrands f, but experiments show that the modified formulas often outperform Gauss quadrature for nonanalytic integrands, too.

All of the methods we have discussed contain adjustable parameters, notably ρ for (3.2), the degree d for the maps of section 4, and α for (5.1). Rather than try to optimize these values, we have taken somewhat arbitrary fixed choices such as $\rho = 1.4$ or 1.1 and $d = 9$ or 17, which are enough to deliver fairly robust quadrature improvements by 30%–50%. These figures could be increased closer to the 57% associated with the ratio $\pi/2$ by decreasing ρ and increasing d and α, perhaps in a fashion dependent on the number of grid points n, and, in some applications, such adjustments may be worthwhile. There is a literature on how to choose α for the application of the Kosloff–Tal-Ezer map (5.1) to spectral methods [2, 15, 18, 29, 37]. Preliminary experiments show that the speedups we have described for quadrature appear equally for spectral methods, and we hope to give details in a later publication. For spectral methods there is the additional consideration that the diminished clustering near ± 1 permits larger time steps or faster iterative matrix solves, making the choice of effective parameters a more complicated matter.

To compute a single integral, or a small number of integrals, one would normally use an adaptive quadrature program, and we do not recommend Gauss or Clenshaw–Curtis formulas, or their transplanted improvements, for such applications. Indeed, even for the complicated integrand of Figure 7.2, transplanted Clenshaw–Curtis quadrature is less than ten times faster than MATLAB's general-purpose adaptive codes `quad` and `quadl`. Rather, the potential practical relevance of the methods investigated in this paper is to situations where large numbers of integrals are embedded inside a bigger computation. We close with an artificial example of a multidimensional integral. Suppose we compute

$$\int_{-1}^{1}\int_{-1}^{1}\int_{-1}^{1}\int_{-1}^{1} \cos(100(w+x+y+z))\,dw\,dx\,dy\,dz = \left(\frac{\sin(100)}{50}\right)^4 \approx 1.05\times 10^{-8}$$

numerically by a quadrature formula applied in four dimensions, exploiting none of the symmetries or simple structure of this particular integrand. To get three digits of relative accuracy, Gauss quadrature requires $63^4 = 15{,}752{,}961$ points, whereas strip-transplanted Gauss quadrature with $\rho = 1.4$ needs $52^4 = 7{,}311{,}616$ points.

Appendix. Computation of the strip transformation in extreme cases and a MATLAB code. Figure 3.2 shows how the conformal map g of (3.2) from the ρ-ellipse E_ρ to the infinite strip Ω_ρ can be constructed in three steps: ellipse$\to$periodic strip$\to$disk$\to$infinite strip, with the middle step involving a Jacobi elliptic function. All of this is mathematically straightforward, but, when ρ is smaller than about 1.1, it is numerically troublesome. The problem is that the right end of the periodic strip gets mapped in the disk domain to the interval $[1, m^{-1/4}]$, and this interval shrinks exponentially as ρ decreases to 1 (similarly at the left end). For example, $\rho = 1.4, 1.2, 1.1$ correspond to $m^{-1/4} \approx 1.0026, 1.0000053, 1.000000000023$, and for $\rho = 1.05$ one calculates $m^{-1/4} = 1$ on a computer to machine precision. This "crowding phenomenon" is a familiar challenge in numerical conformal mapping, making it impossible to compute g accurately unless the computation is reformulated [20].

It would be possible to consider systematically the development of an algorithm to evaluate g to as close to machine precision as possible. Rather than attempt this, we propose a method that works very well for our quadrature application and is much simpler. When ρ is small, the two ends of the rectangle$\to$strip map of Figure 3.2 are exponentially decoupled from one another. In fact, for ρ less than about 1.2, the map near one end of the rectangle is indistinguishable in 16-digit arithmetic from what it would be if the rectangle were a semi-infinite strip. By following this line of reasoning, after quite of bit of algebra, one is led to consider the function g defined by $u = \sin^{-1} s$, $\tau = \pi/\log\rho$, and

$$g(s) = C\left[\log(1+e^{-\tau(\pi/2+u)}) - \log(1+e^{-\tau(\pi/2-u)}) + \left(\frac{1}{2} + \frac{1}{e^{\tau\pi}+1}\right)\tau u\right], \tag{A.1}$$

where the constant C is fixed so that $g(\pm 1) = \pm 1$. To 16-digit precision, for $\rho < 1.2$, this function is the same as the strip map (3.2); and it is numerically reliable down to about $\rho = 1.02$, which is closer to 1 than one would ever need to go in practice. For $s \in (-1, 1)$ the derivative is

$$g'(s) = \frac{-\tau C}{\sqrt{1-s^2}}\left[\frac{1}{e^{\tau(\pi/2+u)}+1} + \frac{1}{e^{\tau(\pi/2-u)}+1} - \left(\frac{1}{2} + \frac{1}{e^{\tau\pi}+1}\right)\right], \tag{A.2}$$

and at the end points we have

$$g'(\pm 1) = \frac{C\tau^2}{4}\tanh^2\left(\frac{\tau\pi}{2}\right). \tag{A.3}$$

Notice how much more elementary (A.1)–(A.2) are than (3.1)–(3.4). In principle g is not analytic in the ellipse, having branch points at $s = \pm 1$. The singularities are so weak, however, that for $\rho < 1.2$ they are undetectable in 16-digit arithmetic.

Although the map (A.1) was derived as an approximation to the ideal (3.2) for small values of ρ, it is surprisingly effective even for larger values of ρ. If (3.2) is replaced by (A.1) in the experiments of Figure 3.4, for example, it makes no difference.

The following terse code combines all of the elements we have discussed, implementing the Clenshaw–Curtis quadrature formula for the modified strip transformation (A.1). Since such a code will be especially useful for large n, the listing gives a value of ρ close to 1. If this code is applied to the function of Figure 7.2 with $n = 1800$, it gets the right answer to 15 digits in about 0.01 seconds on a 2006 workstation. (With $n = 10^6$ it gets the same answer in 5 seconds.) Untransformed Gauss quadrature with $n = 1800$ achieves 11 digits of accuracy and takes far longer.

```
function I = CCstrip(f,n)                                % transplanted C-C quad.
rho = 1.1; t = pi/log(rho);                              % rho (adjustable) and tau
d = .5+1/(exp(t*pi)+1); p2 = pi/2;                       % convenient abbreviations
up = pi*(0:n)/n; u = up-p2; um = p2-u;                   % u and shifts by +-pi/2
C = 1/(log(1+exp(-t*pi))-log(2)+p2*t*d);
g = C*(log(1+exp(-t*up))-log(1+exp(-t*um))+u*t*d); % map g
gp = 1./(exp(t*up)+1)+1./(exp(t*um)+1)-d;
gp(2:n) = -t*C*gp(2:n)./cos(u(2:n));
gp([1 n+1]) = C*(t*tanh(p2*t)/2)^2;                      % derivative g'
fx = f(g).*gp/(2*n);                                     % transplanted integrand
h = real(fft(fx([1:n+1 n:-1:2])));                       % Chebyshev coefficients
a = [h(1) h(2:n)+h(2*n:-1:n+2) h(n+1)];
w = 0*a; w(1:2:n+1) = 2./(1-(0:2:n).^2);                 % Clenshaw-Curtis weights
I = w*a';                                                % the result
```

Acknowledgments. We are grateful for helpful suggestions to Gregory Beylkin, Adam Chandler, Toby Driscoll, Christian Lubich, Ricardo Pachón, and André Weideman.

REFERENCES

[1] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions,* Dover, New York, 1972 (originally published in 1964).
[2] M. R. Abril-Raymundo and B. García-Archilla, *Approximation properties of a mapped Chebyshev method*, Appl. Numer. Math., 32 (2000), pp. 119–136.
[3] B. K. Alpert, *Hybrid Gauss-trapezoidal quadrature rules*, SIAM J. Sci. Comput., 20 (1999), pp. 1551–1584.
[4] N. S. Bakhvalov, *On the optimal speed of integrating analytic functions*, Comput. Math. Math. Phys., 7 (1967), pp. 63–75.
[5] W. Barrett, *On the convergence of Cotes' quadrature formulae*, J. London Math. Soc., 39 (1964), pp. 296–302.
[6] Z. Battles and L. N. Trefethen, *An extension of MATLAB to continuous functions and operators*, SIAM J. Sci. Comput., 25 (2004), pp. 1743–1770.
[7] A. Bayliss and E. Turkel, *Mappings and accuracy for Chebyshev pseudo-spectral approximations*, J. Comput. Phys., 101 (1992), pp. 349–359.
[8] S. Bernstein, *Quelques remarques sur l'interpolation*, Math. Ann., 79 (1919), pp. 1–12.
[9] G. Beylkin and K. Sandberg, *Wave propagation using bases for bandlimited functions*, Wave Motion, 41 (2005), pp. 263–291.
[10] J. P. Boyd, *Prolate spheroidal wavefunctions as an alternative to Chebyshev and Legendre polynomials for spectral element and pseudospectral algorithms*, J. Comput. Phys., 199 (2004), pp. 688–716.
[11] H. Brass, *Quadraturverfahren*, Vandenhoeck and Ruprecht, Göttingen, 1977.
[12] C. Canuto, M. Y. Hussaini, A. Quarteroni, and T. A. Zang, *Spectral Methods: Fundamentals in Single Domains*, Springer, New York, 2006.
[13] Q. Y. Chen, D. Gottlieb, and J. S. Hesthaven, *Spectral methods based on prolate spheroidal wave functions for hyperbolic PDEs*, SIAM J. Numer. Anal., 43 (2005), pp. 1912–1933.
[14] C. W. Clenshaw and A. R. Curtis, *A method for numerical integration on an automatic computer*, Numer. Math., 2 (1960), pp. 197–205.
[15] B. Costa, W. S. Don, and A. Simas, *Exponential Convergence of Mapped Chebyshev Methods*, manuscript, 2007.
[16] P. J. Davis and P. Rabinowitz, *Methods of Numerical Integration,* 2nd ed., Academic Press, New York, 1984.
[17] Z. Ditzian and V. Totik, *Moduli of Smoothness*, Springer, New York, 1987.
[18] W.-S. Don and A. Solomonoff, *Accuracy enhancement for higher derivatives using Chebyshev collocation and a mapping technique*, SIAM J. Sci. Comput., 18 (1997), pp. 1040–1055.
[19] T. A. Driscoll, *Algorithm* 843: *Improvements to the MATLAB toolbox for Schwarz–Christoffel mapping*, ACM Trans. Math. Software, 31 (2005), pp. 239–251.
[20] T. A. Driscoll and L. N. Trefethen, *Schwarz–Christoffel Mapping*, Cambridge University Press, London, 2002.

[21] D. Elliott, B. M. Johnston, and P. R. Johnston, *Clenshaw–Curtis and Gauss-Legendre Quadrature for Certain Boundary Element Integrals*, manuscript, 2006.
[22] H. Engels, *Numerical Quadrature and Cubature,* Academic Press, London, 1980.
[23] P. Favati, G. Lotti, and F. Romani, *Bounds on the error of Fejér and Clenshaw–Curtis type quadrature for analytic functions*, Appl. Math. Lett., 6 (1993), pp. 3–8.
[24] H. E. Fettis, *Note on the computation of Jacobi's nome and its inverse*, Computing, 4 (1969), pp. 202–206.
[25] W. Gautschi, *A Survey of Gauss–Christoffel Quadrature Formulae*, E. B. Christoffel, P. L. Butzer and F. Fehér, eds., Birkhäuser, Basel, 1981, pp. 72–147.
[26] W. M. Gentleman, *Implementing Clenshaw–Curtis quadrature* I *and* II, Comm. ACM, 15 (1972), pp. 337–346, 353.
[27] G. H. Golub and J. H. Welsch, *Calculation of Gauss quadrature rules*, Math. Comp., 23 (1969), pp. 221–230.
[28] M. Götz, *Optimal quadrature for analytic functions*, J. Comput. Appl. Math., 137 (2001), pp. 123–133.
[29] J. S. Hesthaven, P. G. Dinesen, and J. P. Lynov, *Spectral collocation time-domain modelling of diffractive optical elements*, J. Comput. Phys., 155 (1999), pp. 287–306.
[30] F. B. Hildebrand, *Introduction to Numerical Analysis*, 2nd ed., Dover, New York, 1987 (originally published in 1974).
[31] S. Kapur and V. Rokhlin, *High-order corrected trapezoidal quadrature rules for singular functions*, SIAM J. Numer. Anal., 34 (1997), pp. 1331–1356.
[32] T. Koch and J. Liesen, *The conformal 'bratwurst' maps and associated Faber polynomials*, Numer. Math., 86 (2000), pp. 173–191.
[33] D. Kosloff and H. Tal-Ezer, *A modified Chebyshev pseudospectral method with an $O(N^{-1})$ time step restriction*, J. Comput. Phys., 104 (1993), pp. 457–469.
[34] M. A. Kowalski, A. G. Werschulz, and H. Woźniakowski, *Is Gauss quadrature optimal for analytic functions?*, Numer. Math., 47 (1985), pp. 89–98.
[35] E. Martensen, *Optimale Fehlerschranken füer die Quadraturformel von Gregory*, Z. Angew. Math. Mech., 44 (1964), pp. 159–168.
[36] V. I. Krylov, *Approximate Calculation of Integrals*, Dover, New York, 2005 (originally published in 1962).
[37] J. L. Mead and R. A. Renaut, *Accuracy, resolution, and stability properties of a modified Chebyshev method*, SIAM J. Sci. Comput., 24 (2002), pp. 143–160.
[38] M. Mori and M. Sugihara, *The double-exponential transformation in numerical analysis*, J. Comput. Appl. Math., 127 (2001), pp. 287–296.
[39] H. O'Hara and F. J. Smith, *Error estimation in the Clenshaw–Curtis quadrature formula*, Computer J., 11 (1968), pp. 213–219.
[40] K. Petras, *Gaussian versus optimal integration of analytic functions*, Constr. Approx., 14 (1998), pp. 231–245.
[41] G. Pólya, *Über die Konvergenz von Quadraturverfahren*, Math. Z., 37 (1933), pp. 264–286.
[42] P. Rabinowitz, *Rough and ready error estimates in Gaussian integration of analytic functions*, Comm. ACM, 12 (1969), pp. 268–270.
[43] T. W. Sag and G. Szekeres, *Numerical evaluation of high-dimensional integrals*, Math. Comp., 18 (1964), pp. 245–253.
[44] C. Schwartz, *Numerical integration of analytic functions*, J. Comput. Phys., 4 (1969), pp. 19–29.
[45] D. Slepian and H. O. Pollak, *Prolate spheroidal wave functions, Fourier analysis and uncertainty.* I, Bell System Tech. J., 40 (1961), pp. 43–63.
[46] F. Stenger, *Numerical Methods Based on Sinc and Analytic Functions*, Springer, New York, 1993.
[47] H. Takahasi and M. Mori, *Double exponential formulas for numerical integration*, Publ. Res. Inst. Math. Sci., 9 (1974), pp. 721–741.
[48] T. W. Tee and L. N. Trefethen, *A rational spectral collocation method with adaptively transformed Chebyshev grid points*, SIAM J. Sci. Comput., 28 (2006), pp. 1798–1811.
[49] L. N. Trefethen, *Spectral Methods in MATLAB*, SIAM, Philadelphia, 2001.
[50] L. N. Trefethen, *Is Gauss quadrature better than Clenshaw–Curtis?*, SIAM Rev., 50 (2008), pp. 67–87.
[51] L. N. Trefethen, *Computing numerically with functions instead of numbers*, Math. Comput. Sci., 1 (2007), pp. 9–19.
[52] J. A. C. Weideman and L. N. Trefethen, *The kink phenomenon in Fejér and Clenshaw–Curtis quadrature*, Numer. Math., 107 (2007), pp. 707–727.
[53] H. Xiao, V. Rokhlin, and N. Yarvin, *Prolate spheroidal wavefunctions, quadrature and interpolation*, Inverse Problems, 17 (2001), pp. 805–838.

SIAM J. NUMER. ANAL.
Vol. 46, No. 2, pp. 949–979
© 2008 Society for Industrial and Applied Mathematics

COUPLING DISCONTINUOUS GALERKIN AND MIXED FINITE ELEMENT DISCRETIZATIONS USING MORTAR FINITE ELEMENTS*

VIVETTE GIRAULT†, SHUYU SUN‡, MARY F. WHEELER§, AND IVAN YOTOV¶

Abstract. Discontinuous Galerkin (DG) and mixed finite element (MFE) methods are two popular methods that possess local mass conservation. In this paper we investigate DG-DG and DG-MFE domain decomposition couplings using mortar finite elements to impose weak continuity of fluxes and pressures on the interface. The subdomain grids need not match, and the mortar grid may be much coarser, giving a two-scale method. Convergence results in terms of the fine subdomain scale h and the coarse mortar scale H are established for both types of couplings. In addition, a nonoverlapping parallel domain decomposition algorithm is developed, which reduces the coupled system to an interface mortar problem. The properties of the interface operator are analyzed.

Key words. discontinuous Galerkin, mixed finite element, flow in porous media, mortar finite element, interface problem

AMS subject classifications. 65N12, 65N15, 65N22, 65N30, 65N55, 76S05

DOI. 10.1137/060671620

1. Introduction.

In modeling of flow and reactive transport in porous media, it is important to employ algorithms that preserve mathematical properties of physical systems, such as local mass conservation and continuity of fluxes. In addition, geological media such as aquifers and petroleum reservoirs exhibit a high level of spatial variability at a multiplicity of scales, from the size of individual grains or pores to facies, stratigraphic, and hydrologic units, up to sizes of formations. Two methods that are well-suited to subsurface modeling are the mixed finite element (MFE) and discontinuous Galerkin (DG) methods. Common features of these methods are local conservation of mass and accurate treatment of rough coefficients and grids.

MFE and related methods have been very popular in the porous media modeling community. They provide accurate approximation for both the pressure and the velocity. A number of approaches have been developed to eliminate the velocity and reduce the MFE method to a cell-centered or face-centered algebraic pressure system with a substantially smaller dimension; see, e.g., [40, 8, 6]. For single phase flow, the reduced system in most cases is symmetric and positive definite, allowing for the use

*Received by the editors October 6, 2006; accepted for publication (in revised form) September 10, 2007; published electronically March 5, 2008.
http://www.siam.org/journals/sinum/46-2/67162.html

†Universite Pierre et Marie Curie, Paris VI, Laboratoire Jacques-Louis Lions, 4 place Jussieu, F-75230 Paris Cedex 05, France (girault@ann.jussieu.fr). This author's research was partially supported by the J. Tinsley Oden Faculty Fellowship, ICES, The University of Texas at Austin.

‡The Institute for Computational Engineering and Sciences (ICES), The University of Texas at Austin, Austin, TX 78712. Current address: Department of Mathematical Sciences, Clemson University, Clemson, SC 29634 (shuyu@clemson.edu). This author's work was partially supported by DOE grant DE-FGO2-04ER25617.

§The Institute for Computational Engineering and Sciences (ICES), The University of Texas at Austin, Austin, TX 78712 (mfw@tican.utexas.edu). This author's research was partially supported by DOE grant DE-FGO2-04ER25617.

¶Department of Mathematics, University of Pittsburgh, 301 Thackeray Hall, Pittsburgh, PA 15260 (yotov@math.pitt.edu). This author's research was partially supported by the J. Tinsley Oden Faculty Fellowship, ICES, The University of Texas at Austin, by NSF grants DMS 0411694 and by DMS 0620402, and DOE grant DE-FG02-04ER25618.

of efficient solvers. These reduction techniques, however, apply in most cases to low order MFE methods on relatively structured grids and may lead to deterioration in the accuracy on highly irregular or unstructured grids.

DG methods are finite element methods that use discontinuous approximations. Examples of these schemes include the Bassi–Rebay method [9], the local discontinuous Galerkin (LDG) [3, 21] methods, the Oden–Babuška–Baumann (OBB-DG) [32] method, and interior penalty Galerkin methods [23, 38, 45]. DG methods are of particular interest for multiscale problems because they (1) support local approximations of high order and are capable of delivering exponential rates of convergence, (2) are robust and nonoscillatory in the presence of high gradients, and (3) are implementable on unstructured and even nonmatching grids and can thus treat highly heterogeneous porous media. On the negative side, because of the number of unknowns, DG solvers can be expensive.

Nonoverlapping domain decomposition is a useful approach for spatial coupling and/or decoupling. A subsurface flow example is the multiblock mortar MFE methodology described in [4, 34, 35, 44]. The governing equations hold locally on the subdomains, and physically driven matching conditions are imposed on block interfaces in a numerically stable and accurate way by using mortar finite element spaces. References on the mortar approach for other discretizations include [12, 11, 46, 10] for conforming Galerkin and [24] for finite volume elements. Domain decomposition solvers and preconditioners for mortar discretizations have been developed in [28, 2, 1, 27, 33].

Some computational advantages of the multiblock approach are as follows: (1) *Multiphysics:* different physical processes/mathematical models in different parts of the domain may be coupled in a single simulation; (2) *multinumerics:* different numerical techniques may be employed on different subdomains; (3) *multiscale resolution and adaptivity:* highly refined regions or fine-scale models may be coupled with more coarsely discretized regions, and dynamic grid adaptivity may be performed locally on each block; (4) *multidomains:* highly irregular domains may be described as unions of more regular and locally discretized subdomains with the possibility of having interfaces with nonmatching grids; and (5) *parallelism:* the approach leads to domain decomposition algorithms with near-optimal computational load balance and minimal communication overhead.

Couplings of DG and MFE methods have been previously studied in the literature. In [37], a DG-MFE coupling is introduced, which uses two Lagrange multipliers to impose continuity of fluxes and pressures. A method for coupling LDG and MFE is developed in [20] by choosing appropriate numerical fluxes on interface edges.

In [5], a multiscale mortar mixed finite element method was introduced for modeling Darcy flow. There, the continuity of the flux is imposed via mortar finite elements on a coarse grid scale, while the equations in the coarse elements (or subdomains) are discretized on a fine grid scale.

In this paper we develop mortar couplings of DG with DG or MFE methods, using possibly different scales in the mortar and subdomain grids. Such couplings allow for (1) the flexibility of applying DG to subdomains where general grids are required for treating pinchouts, discrete faults and fractures, and highly variable full permeability tensors; (2) developing a mortar domain decomposition parallel DG solver via reduction to an interface problem and employing conjugate gradient or GMRES for its solution; efficient interface preconditioners such as balancing could be developed [30, 22, 33]; (3) applying the MFE method, which has substantially fewer unknowns than DG, in regions with relatively smooth or structured grids; and (4) achieving model reduction through multiscale approximations.

We study mortar couplings of type DG-DG and DG-MFE, based on four different DG formulations, the OBB-DG [32], the nonsymmetric interior penalty Galerkin (NIPG) [39], the symmetric interior penalty Galerkin (SIPG) [7, 45, 43, 42], and the incomplete interior penalty Galerkin (IIPG) [43, 23, 42]. The mortar variable has a meaning of pressure, and it is used as a Lagrange multiplier to impose weak continuity of normal velocities and subdomain pressures on the interface. This is achieved via a Robin-type matching condition, which involves a flux jump term and a penalized pressure jump term. Our approach differs from the one in [37], where two Lagrange multipliers are used and the method cannot be reduced to an interface problem.

The paper is organized as follows. In the next section we introduce the model problem and set up some notation. In section 3 we develop and analyze DG-DG mortar couplings. In particular, we establish equivalence between the DG weak formulation and the partial differential equation, existence and uniqueness for the discrete solution, and convergence estimates. The error estimates are derived in terms of h and H, the discretization parameters for the subdomain and mortar spaces, respectively. We also develop a parallel nonoverlapping domain decomposition algorithm for the solution of the algebraic system based on a reduction of the algebraic system to an interface mortar problem. Similar results are obtained in section 4 for DG-MFE mortar couplings. We end with some conclusions in section 5.

2. Problem statement and notation.

2.1. Model equations. Let the domain be $\Omega = \Omega_1 \cup \Omega_2 \cup \Gamma_{12} \subset \mathbb{R}^d$, $d = 1$, 2, or 3. Although for simplicity we present only the method for two subdomains, our results easily extend to geometrically nonconforming domain decompositions with a finite number of subdomains. Similarly, our results can be generalized to more general boundary conditions than (2.2) below, such as Dirichlet or mixed Dirichlet–Neumann boundary conditions. We consider the following equation, which can be used to model a single-phase flow process in porous media:

$$-\nabla \cdot \mathbf{K}\nabla p = f \quad \text{in } \Omega, \tag{2.1}$$

$$-\mathbf{K}\nabla p \cdot \mathbf{n} = g \quad \text{on } \partial\Omega. \tag{2.2}$$

The above system can also be written in a mixed form:

$$\begin{aligned} \mathbf{u} &= -\mathbf{K}\nabla p \quad \text{in}\, \Omega, \\ \nabla \cdot \mathbf{u} &= f \quad \text{in}\, \Omega, \\ \mathbf{u} \cdot \mathbf{n} &= g \quad \text{on}\, \partial\Omega. \end{aligned}$$

Here $\mathbf{n}$ denotes the unit outward normal vector to $\partial\Omega$. We assume that the data f belong to $L^2(\Omega)$ and g belong to $L^2(\partial\Omega)$ and they both satisfy the compatibility condition

$$\int_\Omega f dx = \int_{\partial\Omega} g d\sigma. \tag{2.3}$$

With (2.3), system (2.1)–(2.2) defines p uniquely up to an additive constant. The conductivity $\mathbf{K}$ is assumed to be uniformly symmetric positive definite and bounded from above. In Ω_1, we use a DG formulation; in Ω_2, we use either a DG or a mixed formulation, with the matching on the interface being achieved by a multiplier.

Throughout the paper we will use the following standard notation. For $D \subset \mathbb{R}^d$, we denote the norm in the Hilbert space $H^s(D)$ by $\|\cdot\|_{s,D}$. Consequently, $\|\cdot\|_{0,D}$ will

denote the norm in $L^2(D)$. We may omit the subscript D if $D = \Omega$. We denote by C a generic positive constant, independent of h and H, that may not have the same value at different occurrences. In addition, we denote by ϵ a fixed positive constant that may be chosen arbitrarily small.

Let $\mathcal{E}_h(\Omega_i)$ be a nondegenerate partition of Ω_i, $i = 1, 2$, composed of line segments if $d = 1$, triangles or quadrilaterals if $d = 2$, or tetrahedra, prisms, or hexahedra if $d = 3$. If MFE discretization is used in Ω_2, we consider only affine elements there. The partitions do not need to match on the interface Γ_{12}. If SIPG, NIPG, or IIPG is used in Ω_i, we allow $\mathcal{E}_h(\Omega_i)$ to be nonconforming by refining the mesh in some of the elements. Let $\mathcal{E}_h(\Omega) = \mathcal{E}_h(\Omega_1) \cup \mathcal{E}_h(\Omega_2)$.

Here h is the maximum element diameter for the mesh. The nondegeneracy requirement (also called regularity) is that the element is convex and that there exists $\rho > 0$ such that, if h_j is the diameter of $E_j \in \mathcal{E}_h(\Omega)$, each of the subtriangles (for $d = 2$) or subtetrahedra (for $d = 3$) of element E_j contains a ball of radius ρh_j in its interior. If E_j is a triangle (or a tetrahedron), then its subtriangles (or subtetrahedra) coincide with E_j. The set of all interior points ($d = 1$), edges ($d = 2$), or faces ($d = 3$) within $\mathcal{E}_h(\Omega_i)$ is denoted by $\Gamma_h(\Omega_i)$. Let $\Gamma_h(\Omega) = \Gamma_h(\Omega_1) \cup \Gamma_h(\Omega_2)$. On each element face $\gamma \in \Gamma_h(\Omega)$, a unit normal vector $\mathbf{n}$ is chosen once and for all. This could be done by numbering the elements of $\mathcal{E}_h(\Omega_i)$ and directing $\mathbf{n}$ from E_k to E_l if E_k and E_l are adjacent and $k < l$. On $\partial\Omega$, the normal vector $\mathbf{n}$ coincides with the outward unit normal vector $\mathbf{n}_{\partial\Omega}$. On Γ_{12}, the unit normal vector $\mathbf{n}$ is chosen as $\mathbf{n} = \mathbf{n}_{\partial\Omega_1} = -\mathbf{n}_{\partial\Omega_2}$. On Γ_{12}, we introduce a mortar finite element partition Γ_H, where H is the maximum diameter of mortar elements. We allow for Γ_H to be different from the traces of the subdomain grids on Γ_{12}.

Let E_i and E_j be two adjacent elements in $\mathcal{E}_h(\Omega)$, with $i < j$, and let $\gamma = \partial E_i \cap \partial E_j \in \Gamma_h(\Omega)$; then $\mathbf{n}$ is exterior to E_i. We denote the average and jump on γ for an elementwise smooth function ϕ by, respectively,

$$\{\phi\} := \frac{1}{2}((\phi|_{E_i})|_\gamma + (\phi|_{E_j})|_\gamma), \qquad [\phi] := (\phi|_{E_i})|_\gamma - (\phi|_{E_j})|_\gamma .$$

The following functional spaces will be used in weak formulations of our problem:

$$X(\Omega_i) := \left\{q \in L^2(\Omega_i) : \forall E \in \mathcal{E}_h(\Omega_i),\, q|_E \in H^s(E)\right\}, \quad s > \frac{3}{2}, \quad i = 1 \text{ and } 2,$$
$$\mathbf{V}_0(\Omega_2) := \{\mathbf{v} \in H(\mathrm{div}; \Omega_2) : \mathbf{v} \cdot \mathbf{n} = 0 \text{ on } \partial\Omega_2 \setminus \Gamma_{12}\},$$
$$W(\Omega_2) := L^2(\Omega_2), \quad \Lambda := H^{\frac{1}{2}}(\Gamma_{12}).$$

Here $\mathbf{v}\cdot\mathbf{n}$ is defined in a weak sense, i.e., in the dual of $H_{00}^{\frac{1}{2}}(\partial\Omega_2\setminus\Gamma_{12})$, where $H_{00}^{\frac{1}{2}}(D)$ is the interpolation space between $L^2(D)$ and $H_0^1(D)$ [29]. The space $X(\Omega_i)$ is equipped with the norm

$$\|\!|\!| \cdot |\!|\!|_{s,\Omega_i} = \left(\sum_{E\in\mathcal{E}_h(\Omega_i)} \|\cdot\|_{s,E}^2\right)^{\frac{1}{2}}. \tag{2.4}$$

For the DG discretization we will use the finite element spaces

$$X_h(\Omega_i) := \left\{q_h \in L^2(\Omega_i) : \forall E \in \mathcal{E}_h(\Omega_i),\, q_h|_E \in \mathbb{P}_r(E)\right\}, \quad i = 1, 2, \quad r \geq 1.$$

For the MFE discretization in Ω_2 we will use any of the usual mixed spaces, including the Raviart–Thomas–Nedelec (RTN) spaces [36, 31], Brezzi–Douglas–Marini (BDM) spaces [16], Brezzi–Douglas–Fortin–Marini (BDFM) spaces [15], Brezzi–Douglas–

Duràn–Fortin (BDDF) spaces [14], or Chen–Douglas (CD) spaces [18]. We denote these spaces by $\mathbf{V}_h(\Omega_2) \times W_h(\Omega_2)$, where $\mathbf{V}_h(\Omega_2) \times W_h(\Omega_2) \subset H(\mathrm{div};\Omega_2) \times L^2(\Omega_2)$. To enforce the boundary condition on $\partial\Omega_2 \setminus \Gamma_{12}$, we set

$$\mathbf{V}_{h,0}(\Omega_2) = \mathbf{V}_h(\Omega_2) \cap \mathbf{V}_0(\Omega_2).$$

On an element E, the restriction of $\mathbf{V}_h(\Omega_2)$ is denoted by $\mathbf{V}_h(E)$. We assume that the velocity space $\mathbf{V}_h(E)$ contains $(\mathbb{P}_m(E))^d$, $m \geq 0$, with normal components on each edge (face) in $\mathbb{P}_m(\gamma)$ and that the pressure space $W_h(E)$ contains $\mathbb{P}_l(E)$. In all cases $l = m$ or $l = m-1$, when $m \geq 1$.

On the interface we will use a mortar finite element space to approximate the pressure and impose weakly the continuity of flux and pressure:

$$\Lambda_H := \left\{ \mu_H \in L^2(\Gamma_{12}) : \forall \tau \in \Gamma_H,\, \mu_H|_\tau \in \mathbb{P}_{\bar{r}}(\tau) \right\}, \quad \bar{r} \geq 1.$$

In the above, r, m, and $\bar{r}$ are possibly different constants. We note that all results in this paper hold if Λ_H is replaced by its continuous version

$$\Lambda_H^c := \left\{ \mu_H \in C^0(\Gamma_{12}) : \forall \tau \in \Gamma_H,\, \mu_H|_\tau \in \mathbb{P}_{\bar{r}}(\tau) \right\}, \quad \bar{r} \geq 1.$$

3. Coupling DG with DG using a mortar space. From now on, we assume that the tensor $\mathbf{K}$ is sufficiently smooth in each element, so that the trace $\mathbf{K}\nabla p \cdot \mathbf{n}$ is well-defined on element faces. In this section we consider coupled schemes involving DG discretizations in both Ω_1 and Ω_2 and matching conditions on the interface imposed through a mortar finite element space.

3.1. Weak formulation. We define bilinear forms and linear functionals for the DG scheme in Ω_i, $i = 1, 2$:

(3.1)
$$\begin{aligned} B_i(p,q) := & \sum_{E \in \mathcal{E}_h(\Omega_i)} \int_E \mathbf{K}\nabla p \cdot \nabla q \, dx - \sum_{\gamma \in \Gamma_h(\Omega_i)} \int_\gamma \{\mathbf{K}\nabla p \cdot \mathbf{n}\} [q] \, d\sigma \\ & - s_{\mathrm{form}} \sum_{\gamma \in \Gamma_h(\Omega_i)} \int_\gamma \{\mathbf{K}\nabla q \cdot \mathbf{n}\} [p] \, d\sigma - \int_{\Gamma_{12}} \mathbf{K}\nabla p \cdot \mathbf{n}_{\partial\Omega_i} q|_{\Omega_i} \, d\sigma \\ & - \bar{s}_{\mathrm{form}} \int_{\Gamma_{12}} \mathbf{K}\nabla q \cdot \mathbf{n}_{\partial\Omega_i} p|_{\Omega_i} \, d\sigma + \sum_{\gamma \in \Gamma_h(\Omega_i)} \frac{\sigma_\gamma}{h_\gamma} \int_\gamma [p] [q] \, d\sigma + \sum_{\tau \in \Gamma_H} \frac{\sigma_\tau}{H_\tau} \int_\tau p|_{\Omega_i} q|_{\Omega_i} d\sigma, \end{aligned}$$

(3.2)
$$L_i(q;\lambda) := \int_{\Omega_i} f\, q \, dx - \int_{\partial\Omega_i \setminus \Gamma_{12}} g\, q \, d\sigma - \bar{s}_{\mathrm{form}} \int_{\Gamma_{12}} \mathbf{K}\nabla q \cdot \mathbf{n}_{\partial\Omega_i} \lambda \, d\sigma + \sum_{\tau \in \Gamma_H} \frac{\sigma_\tau}{H_\tau} \int_\tau q|_{\Omega_i} \, \lambda \, d\sigma.$$

Here $s_{\mathrm{form}} = -1$ for NIPG or OBB-DG, $s_{\mathrm{form}} = 1$ for SIPG, and $s_{\mathrm{form}} = 0$ for IIPG. The penalty parameter is a discrete positive function that takes the constant value σ_γ on an interior element face γ and σ_τ on a mortar element τ. We let $\sigma_\gamma \equiv 0$ for OBB-DG and assume that $0 < \sigma_\gamma^0 \leq \sigma_\gamma \leq \sigma_\gamma^1$ for SIPG, NIPG, and IIPG. For all methods we assume that $0 < \sigma_\tau^0 \leq \sigma_\tau \leq \sigma_\tau^1$. We will also show solvability for all schemes if $\sigma_\tau \equiv 0$, assuming that condition (A.1) holds for either $X_h(\Omega_1)$ or $X_h(\Omega_2)$, but will provide no convergence analysis in this case.

We take $\bar{s}_{\mathrm{form}} = -1$ for all methods. This leads to an easy control of the terms involving integrals on Γ_{12}. The choices 0 or 1 for $\bar{s}_{\mathrm{form}}$ are also possible. However,

in these cases, the weights in the last terms of $B_i(\cdot,\cdot)$ and $L_i(\cdot;\cdot)$ must be modified. More details are given in Remarks 3.1 and 3.3.

The weak formulation is: Find $p \in L^2(\Omega)$ with $p|_{\Omega_i} \in X(\Omega_i)$ for $i = 1, 2$, and $\lambda \in \Lambda$ such that

$$B_i(p,q) = L_i(q;\lambda) \quad \forall\, q \in X(\Omega_i), \quad i = 1,\, 2, \tag{3.3}$$

$$-\int_{\Gamma_{12}} [\mathbf{K}\nabla p \cdot \mathbf{n}]\, \mu d\sigma + \sum_{\tau\in\Gamma_H} \frac{\sigma_\tau}{H_\tau} \int_\tau \sum_{i=1,2} (p|_{\Omega_i} - \lambda)\mu d\sigma = 0 \quad \forall \mu \in \Lambda. \tag{3.4}$$

3.2. Equivalence. We next show that any solution of the mortar DG weak formulation satisfies the original problem. It is easy to check that the converse is true, provided the solution of the original problem is sufficiently smooth.

THEOREM 3.1. *If (p,λ) is a solution of* (3.3)–(3.4)*, then p satisfies* (2.1)–(2.2) *in the sense of distributions.*

Proof. We first consider the domain Ω_1. For any fixed $E \in \mathcal{E}_h(\Omega_1)$, by taking $q \in C_0^\infty(E)$ in (3.3), we easily see that (2.1) holds within E. We next consider two adjacent elements $E_1 \subset \Omega_1$ and $E_2 \subset \Omega_1$ with an interface γ. By letting $q \in H_0^2(E_1 \cup \gamma \cup E_2)$ in (3.3), we have

$$\begin{aligned}
&\int_{E_1} \mathbf{K}\nabla p \cdot \nabla q dx + \int_{E_2} \mathbf{K}\nabla p \cdot \nabla q dx - s_{\text{form}} \int_\gamma \{\mathbf{K}\nabla q \cdot \mathbf{n}\}\, [p]\, d\sigma \\
&= \int_{E_1\cup E_2} f q dx = -\int_{E_1} \nabla\cdot(\mathbf{K}\nabla p) q dx - \int_{E_2} \nabla\cdot(\mathbf{K}\nabla p) q dx \\
&= \sum_{i=1,2} \left(\int_{E_i} \mathbf{K}\nabla p \cdot \nabla q dx - \int_{\partial E_i} \mathbf{K}\nabla p \cdot \mathbf{n}_{\partial E_i} q d\sigma \right),
\end{aligned}$$

where we have used the fact that $-\nabla\cdot(\mathbf{K}\nabla p) = f$ in E for all $E \in \mathcal{E}_h(\Omega_1)$. Therefore

$$s_{\text{form}} \int_\gamma \{\mathbf{K}\nabla q \cdot \mathbf{n}\}\, [p]\, d\sigma = \int_\gamma [\mathbf{K}\nabla p \cdot \mathbf{n}]\, q d\sigma. \tag{3.5}$$

For OBB-DG, NIPG, and SIPG, we have $s_{\text{form}} \neq 0$, and we can choose $q \in H_0^2(E_1 \cup \gamma \cup E_2)$ such that q is zero on γ and $\{\mathbf{K}\nabla q \cdot \mathbf{n}\}$ is arbitrary in $H_{00}^{\frac{1}{2}}(\gamma)$. Then (3.5) implies that $[p] = 0$ on γ. When running over all interior faces of Ω_1, this means that p belongs to $H^1(\Omega_1)$. In turn, we can choose $q \in H_0^2(E_1 \cup \gamma \cup E_2)$ such that $q|_\gamma$ is arbitrary in $H_0^{\frac{3}{2}}(\gamma)$; then (3.5) implies that the jump $[\mathbf{K}\nabla p \cdot \mathbf{n}] = 0$ on γ. Thus $\mathbf{K}\nabla p$ belongs to $H(\text{div};\Omega_1)$, and, therefore, the interior equation $-\nabla\cdot\mathbf{K}\nabla p = f$ holds globally in Ω_1. The same conclusion holds in Ω_2.

If $s_{\text{form}} = 0$ (i.e., if we use IIPG), then (3.5) directly implies that $[\mathbf{K}\nabla p \cdot \mathbf{n}] = 0$ on γ. Next, by choosing for $i = 1, 2$, q in (3.3) such that $q|_{E_i} \in H^2(E_i)$ with q and $\nabla q \cdot \mathbf{n}$ both zero on $\partial(E_i) \setminus \gamma$, q arbitrary on $H_0^{\frac{3}{2}}(\gamma)$, and $q = 0$ elsewhere, we see that (3.3) reduces to

$$\frac{\sigma_\gamma}{h_\gamma} \int_\gamma [p]\, [q]\, d\sigma = 0,$$

which implies that $[p] = 0$ on γ. Then we conclude as above that p belongs to $H^1(\Omega_1)$ and (2.1) is satisfied in Ω_1 for the four DG versions. Clearly, the same result is true in Ω_2.

Now, by substituting this information into (3.3) with $i = 1, 2$, we obtain that p and λ satisfy for any $q \in X(\Omega_i)$

$$\begin{aligned}\int_{\Gamma_{12}} \mathbf{K}\nabla q \cdot \mathbf{n}_{\partial\Omega_i} p|_{\Omega_i}\, d\sigma + \sum_{\tau\in\Gamma_H} \frac{\sigma_\tau}{H_\tau}\int_\tau (pq)|_{\Omega_i}\, d\sigma = \sum_{\tau\in\Gamma_H} \frac{\sigma_\tau}{H_\tau}\int_\tau q|_{\Omega_i}\,\lambda\, d\sigma \\ - \int_{\partial\Omega_i\setminus\Gamma_{12}} (g + \mathbf{K}\nabla p\cdot\mathbf{n}_{\partial\Omega_i}) q\, d\sigma + \int_{\Gamma_{12}} \mathbf{K}\nabla q\cdot\mathbf{n}_{\partial\Omega_i}\lambda\, d\sigma.\end{aligned} \tag{3.6}$$

To recover the boundary condition (2.2), let E be an element of $\mathcal{E}_h(\Omega_i)$ adjacent to $\partial\Omega_i \setminus \Gamma_{12}$, and $\gamma = \partial E \cap (\partial\Omega_i \setminus \Gamma_{12})$. By taking $q|_{\Omega_i\setminus\overline{E}} = 0$ and $q|_E \in H^2(E)$ with $q = 0$ and $\nabla q\cdot\mathbf{n} = 0$ on $\partial E\setminus\gamma$, (3.6) reduces to

$$-\int_\gamma \mathbf{K}\nabla p\cdot\mathbf{n}\, q\, d\sigma = \int_\gamma g\, q\, d\sigma.$$

Since the trace of q is arbitrary in $H_0^{\frac{3}{2}}(\gamma)$, we conclude that $-\mathbf{K}\nabla p\cdot\mathbf{n} = g$ on γ. Therefore (2.2) is satisfied on $\partial\Omega_i\setminus\Gamma_{12}$, and (3.6) becomes

$$\begin{aligned}\int_{\Gamma_{12}} \mathbf{K}\nabla q \cdot \mathbf{n}_{\partial\Omega_i} p|_{\Omega_i}\, d\sigma + \sum_{\tau\in\Gamma_H} \frac{\sigma_\tau}{H_\tau}\int_\tau (pq)|_{\Omega_i}\, d\sigma = \sum_{\tau\in\Gamma_H} \frac{\sigma_\tau}{H_\tau}\int_\tau q|_{\Omega_i}\,\lambda\, d\sigma \\ + \int_{\Gamma_{12}} \mathbf{K}\nabla q\cdot\mathbf{n}_{\partial\Omega_i}\lambda\, d\sigma.\end{aligned} \tag{3.7}$$

Finally, we turn to the interface Γ_{12}. For $i = 1, 2$, let E be an element of $\mathcal{E}_h(\Omega_i)$ adjacent to Γ_{12}, and $\gamma = \partial E \cap \Gamma_{12}$. By taking $q|_{\Omega_i\setminus\overline{E}} = 0$ and $q|_E \in H^2(E)$ with $q = 0$ and $\nabla q\cdot\mathbf{n} = 0$ on $\partial E\setminus\gamma$, (3.7) becomes

$$\sum_{\tau\in\Gamma_H,\tau\cap\gamma\neq\emptyset} \left(\int_\tau \mathbf{K}\nabla q\cdot\mathbf{n}_{\partial\mathbf{\Omega_i}}(p|_{\Omega_i} - \lambda) d\sigma + \frac{\sigma_\tau}{H_\tau}\int_\tau (p|_{\Omega_i} - \lambda) q|_{\Omega_i}\, d\sigma\right) = 0.$$

By choosing $\mathbf{K}\nabla q\cdot\mathbf{n}$ and q arbitrarily in the interior of γ, we show that $p|_{\Omega_i} = \lambda$ on γ, and consequently for $i = 1, 2$, $p|_{\Omega_i} = \lambda$ on Γ_{12}. This implies in particular that $p \in H^1(\Omega)$. The matching condition (3.4) now becomes $-\int_{\Gamma_{12}} [\mathbf{K}\nabla p\cdot\mathbf{n}]\,\mu d\sigma = 0$, which implies that $\mathbf{K}\nabla p\cdot\mathbf{n}$ is continuous across Γ_{12}. Thus we have (2.1) over the entire domain Ω. ∎

3.3. Discretization. The mortar DG-DG finite element scheme is as follows: Find $(p_h|_{\Omega_1}, p_h|_{\Omega_2}, \lambda_H) \in X_h(\Omega_1)\times X_h(\Omega_2)\times\Lambda_H$ such that

$$B_i(p_h, q_h) = L_i(q_h;\lambda_H) \quad \forall\, q_h \in X_h(\Omega_i), \quad i = 1,\, 2, \tag{3.8}$$

$$\int_{\Gamma_{12}} [\mathbf{K}\nabla p_h\cdot\mathbf{n}]\,\mu_H d\sigma = \sum_{\tau\in\Gamma_H}\frac{\sigma_\tau}{H_\tau}\int_\tau \sum_{i=1}^{2}(p_h|_{\Omega_i} - \lambda_H)\mu_H d\sigma \quad \forall\mu_H \in \Lambda_H. \tag{3.9}$$

In the analysis we shall use the following inequalities, which hold if the penalty parameter σ_γ^0 is chosen to be sufficiently large.

LEMMA 3.1. *For $i = 1, 2$, let $\mathcal{E}_h(\Omega_i)$ be nondegenerate. Then, with each γ in $\Gamma_h(\Omega_i)$, we can associate a positive number σ_γ such that the following inequality holds for all $q_h \in X_h(\Omega_i)$ and all $q \in X(\Omega_i)$:*

$$\left|\sum_{\gamma\in\Gamma_h(\Omega_i)}\int_\gamma \{\mathbf{K}\nabla q_h\cdot\mathbf{n}\}\,[q]\, d\sigma\right| \le \frac{1}{8}\|\mathbf{K}^{\frac{1}{2}}\nabla q_h\|_{0,\Omega_i}^2 + \frac{1}{8}\sum_{\gamma\in\Gamma_h(\Omega_i)}\frac{\sigma_\gamma}{h_\gamma}\int_\gamma [q]^2\, d\sigma. \tag{3.10}$$

Proof. For each E in $\mathcal{E}_h(\Omega_i)$, let $\lambda_E^{\max}$ (resp., $\lambda_E^{\min}$) denote the maximum (resp., minimum) of the eigenvalues of $\mathbf{K}$ on E. By assumption, $\lambda_E^{\max} \le k_{\max}$ and $\lambda_E^{\min} \ge k_{\min} > 0$, with $k_{\max}$ and $k_{\min}$ independent of h and E. Let $\gamma = \partial E_1 \cap \partial E_2$ belong to $\Gamma_h(\Omega_i)$, and let us expand

$$\|\{\mathbf{K}\nabla q_h \cdot \mathbf{n}\}\|_{0,\gamma} \le \frac{1}{2} \sum_{j=1,2} \|\mathbf{K}\nabla q_h|_{E_j}\|_{0,\gamma} \le \frac{1}{2} \sum_{j=1,2} \lambda_{E_j}^{\max} \|\nabla q_h|_{E_j}\|_{0,\gamma}.$$

By reverting to the reference element $\hat{E}$, and using the equivalence of norms on a finite-dimensional space on $\hat{E}$, there exists a constant $\hat{c}$, independent of h, such that

$$\begin{aligned} \|\{\mathbf{K}\nabla q_h \cdot \mathbf{n}\}\|_{0,\gamma} &\le \frac{\hat{c}}{2} \sum_{j=1,2} \lambda_{E_j}^{\max} \left(\frac{|\gamma|}{|E_j|}\right)^{\frac{1}{2}} \|\nabla q_h\|_{0,E_j} \\ &\le \frac{\hat{c}}{2} \sum_{j=1,2} \frac{\lambda_{E_j}^{\max}}{\sqrt{\lambda_{E_j}^{\min}}} \left(\frac{|\gamma|}{|E_j|}\right)^{\frac{1}{2}} \|\mathbf{K}^{\frac{1}{2}}\nabla q_h\|_{0,E_j}. \end{aligned} \tag{3.11}$$

Therefore

$$\left| \sum_{\gamma \in \Gamma_h(\Omega_i)} \int_\gamma \big\{\mathbf{K}\nabla q_h \cdot \mathbf{n}\big\}[q] d\sigma \right| \le \frac{\hat{c}}{2} \sum_{\gamma \in \Gamma_h(\Omega_i)} \sum_{j=1,2} \frac{\lambda_{E_j}^{\max}}{\sqrt{\lambda_{E_j}^{\min}}} \left(\frac{|\gamma|}{|E_j|}\right)^{\frac{1}{2}} \|\mathbf{K}^{\frac{1}{2}}\nabla q_h\|_{0,E_j} \|[q]\|_{0,\gamma}.$$

By applying Young's inequality with parameter $\epsilon > 0$, this becomes

$$\begin{aligned} \left| \sum_{\gamma \in \Gamma_h(\Omega_i)} \int_\gamma \big\{\mathbf{K}\nabla q_h \cdot \mathbf{n}\big\}[q] d\sigma \right| &\le \frac{\epsilon}{4} \sum_{\gamma \in \Gamma_h(\Omega_i)} \sum_{j=1,2} \|\mathbf{K}^{\frac{1}{2}}\nabla q_h\|_{0,E_j}^2 \\ &+ \frac{\hat{c}^2}{4\epsilon} \sum_{\gamma \in \Gamma_h(\Omega_i)} \sum_{j=1,2} \frac{(\lambda_{E_j}^{\max})^2}{\lambda_{E_j}^{\min}} \frac{|\gamma|}{|E_j|} \|[q]\|_{0,\gamma}^2. \end{aligned} \tag{3.12}$$

It is easy to see that in the first sum an element E is counted at most L times, where L is a fixed number that depends on the type of elements used; for instance, $L = 4$ in the case of tetrahedra. Therefore, by choosing $\epsilon = \frac{1}{2L}$, we can take

$$\sigma_\gamma = 4L\hat{c}^2 \sum_{j=1,2} \frac{|\gamma| h_\gamma}{|E_j|} \frac{(\lambda_{E_j}^{\max})^2}{\lambda_{E_j}^{\min}}, \tag{3.13}$$

a quantity that is bounded above and below independently of h, owing to the nondegeneracy of the mesh. With this choice, (3.12) implies (3.10). ∎

We next analyze the solvability of the system (3.8)–(3.9). This is a square finite-dimensional system, and existence is equivalent to uniqueness. Let $f = 0$ and $g = 0$. Take $q_h = p_h$ in (3.8), and sum over the two subdomains Ω_1 and Ω_2 to obtain

$$\begin{aligned} &\sum_{E \in \mathcal{E}_h(\Omega)} \int_E \mathbf{K}\nabla p_h \cdot \nabla p_h dx - (1 + s_{\text{form}}) \sum_{\gamma \in \Gamma_h(\Omega)} \int_\gamma \{\mathbf{K}\nabla p_h \cdot \mathbf{n}\} [p_h] \, d\sigma \\ &+ \sum_{\gamma \in \Gamma_h(\Omega)} \frac{\sigma_\gamma}{h_\gamma} \int_\gamma [p_h]^2 \, d\sigma + \sum_{\tau \in \Gamma_H} \frac{\sigma_\tau}{H_\tau} \int_\tau \sum_{i=1}^{2} p_h|_{\Omega_i}^2 \, d\sigma \\ &= \int_{\Gamma_{12}} [\mathbf{K}\nabla p_h \cdot \mathbf{n}] \, \lambda_H d\sigma + \sum_{\tau \in \Gamma_H} \frac{\sigma_\tau}{H_\tau} \int_\tau \sum_{i=1}^{2} p_h|_{\Omega_i} \, \lambda_H d\sigma. \end{aligned} \tag{3.14}$$

Summation of (3.14) and (3.9) with $\mu_H = \lambda_H$ leads to

$$\begin{aligned}(3.15)\qquad &\|\mathbf{K}^{\frac{1}{2}}\nabla p_h\|_{0,\Omega}^2 - (1+s_{\text{form}})\sum_{\gamma\in\Gamma_h(\Omega)}\int_\gamma \{\mathbf{K}\nabla p_h\cdot\mathbf{n}\}\,[p_h]\,d\sigma\\ &+\sum_{\gamma\in\Gamma_h(\Omega)}\frac{\sigma_\gamma}{h_\gamma}\int_\gamma [p_h]^2\,d\sigma+\sum_{\tau\in\Gamma_H}\frac{\sigma_\tau}{H_\tau}\int_\tau\sum_{i=1}^{2}\left(p_h|_{\Omega_i}-\lambda_H\right)^2 d\sigma=0.\end{aligned}$$

First consider the case $0<\sigma_\tau^0\le\sigma_\tau\le\sigma_\tau^1$. For OBB-DG and NIPG, we have

$$(3.16)\quad \|\mathbf{K}^{\frac{1}{2}}\nabla p_h\|_{0,\Omega}^2+\sum_{\gamma\in\Gamma_h(\Omega)}\frac{\sigma_\gamma}{h_\gamma}\int_\gamma [p_h]^2\,d\sigma+\sum_{\tau\in\Gamma_H}\frac{\sigma_\tau}{H_\tau}\int_\tau\sum_{i=1}^{2}\left(p_h|_{\Omega_i}-\lambda_H\right)^2 d\sigma=0.$$

Since $\mathbf{K}$ is positive definite in Ω, the above equation implies that $\nabla p_h=0$ in each $E\in\mathcal{E}_h(\Omega)$ and $p_h|_{\Omega_i}=\lambda_H$ on Γ_{12}, i.e., p_h is continuous across Γ_{12}. For NIPG, we see, in addition, that p_h is continuous in Ω_1 and Ω_2; therefore p_h is a constant over the entire domain Ω, and λ_H is the same constant on Γ_{12}.

For OBB-DG and $i=1,2$, (3.8) now implies that

$$(3.17)\qquad \sum_{\gamma\in\Gamma_h(\Omega_i)}\int_\gamma\{\mathbf{K}\nabla q_h\cdot\mathbf{n}\}\,[p_h]\,d\sigma=0.$$

Let $\gamma\in\Gamma_h(\Omega_i)$, and let $\gamma\subset\partial E$, $E\in\mathcal{E}_h(\Omega_i)$. If $r\ge 2$, we can construct q_h such that $q_h|_{\Omega\setminus\overline{E}}=0$, $q_h|_E\in\mathbb{P}_r$, $\int_\gamma\mathbf{K}\nabla q_h\cdot\mathbf{n}d\sigma=1$, and $\int_{\gamma'}\mathbf{K}\nabla q_h\cdot\mathbf{n}d\sigma=0$ for $\gamma'\subset\partial E\setminus\gamma$ (see [39]). Equation (3.17) implies that $[p_h]=0$ on γ. Hence p_h is continuous in Ω_i, $i=1$ and 2, and therefore p_h is a constant over the entire domain Ω and λ_H is the same constant on Γ_{12}.

For SIPG and IIPG, assuming that σ_γ^0 is sufficiently large, we employ the inequality (3.10) in Ω_1 and Ω_2 with $q_h=q=p_h$ and conclude from (3.15) that

$$(3.18)\qquad 0\ge\frac{1}{2}\|\mathbf{K}^{\frac{1}{2}}\nabla p_h\|_{0,\Omega}^2+\frac{1}{2}\sum_{\gamma\in\Gamma_h(\Omega)}\frac{\sigma_\gamma}{h_\gamma}\int_\gamma[p_h]^2\,d\sigma+\sum_{\tau\in\Gamma_H}\frac{\sigma_\tau}{H_\tau}\int_\tau\sum_{i=1}^{2}\left(p_h|_{\Omega_i}-\lambda_H\right)^2 d\sigma.$$

We then conclude that p_h is a constant over the entire Ω and λ_H is the same constant on Γ_{12}.

Let us now consider the case $\sigma_\tau=0$; we see from (3.8), using (3.17), (3.16), or (3.18), that for all schemes we have

$$(3.19)\qquad \int_{\Gamma_{12}}\mathbf{K}\nabla q_h|_{\Omega_i}\cdot\mathbf{n}_{\partial\Omega_i}(p_h|_{\Omega_i}-\lambda_H)d\sigma=0,\quad i=1\text{ and }2.$$

If the mortar compatibility condition (A.1) is satisfied for $X_h(\Omega_1)$, we have $\lambda_H=p_h|_{\Omega_1}$ from (3.19) with $i=1$. Since λ_H and $p_h|_{\Omega_2}$ are both constants, we further conclude, from (3.19) with $i=2$, that $p_h|_{\Omega_1}$, $p_h|_{\Omega_2}$, and λ_H must be the same constant. This concludes the argument for $\sigma_\tau=0$.

We have shown for all schemes that the null space of the linear system (3.8)–(3.9) is the constant vector. Owing to the compatibility condition (2.3), the right-hand side for $q_h=1$ is $\int_\Omega f\,dx-\int_{\partial\Omega}g\,d\sigma=0$. Hence the solution exists and is unique up to an additive constant. We therefore have proved the following solvability theorem.

THEOREM 3.2. *For OBB-DG, we assume that $r \geq 2$. For SIPG and IIPG, we assume that σ_γ^0 is sufficiently large. We make no assumption for NIPG. Then the scheme* (3.8)–(3.9) *possesses a solution (p_h, λ_H) unique up to an additive constant that is the same for p_h and λ_H. The same conclusion holds if $\sigma_\tau = 0$, assuming that the compatibility condition* (A.1) *holds for either $i = 1$ or 2.*

3.4. Convergence of the DG-DG schemes. Now, we use an interpolant $\hat{p}$ of p that has particular properties on the elements adjacent to the interface Γ_{12}. On the other elements E, we take $\hat{p}$ to be the interpolant constructed in [39] in the case of OBB-DG or simply take $\hat{p}$ to be the $L^2(E)$-projection of p for the DG methods with interior penalties. More precisely, $\hat{p}|_{\Omega_i} \in X_h(\Omega_i)$, $i = 1, 2$, for all $E \in \mathcal{E}_h(\Omega)$, $\hat{p}|_E$ is exact on $\mathbb{P}_r$, and for all E adjacent to Γ_{12}

$$\int_\gamma \mathbf{K}\nabla\hat{p}\cdot\mathbf{n} = \int_\gamma \mathbf{K}\nabla p\cdot\mathbf{n} \quad \forall\gamma \subset \partial E \cap \Gamma_{12}. \tag{3.20}$$

In the case where $r \geq 2$ such an interpolant is constructed in [39]. In fact, that interpolant satisfies (3.20) on all element sides $\gamma \subset \partial E$. In the case where $r = 1$, we augment (3.20) with the conditions

$$\int_{\gamma_i} \mathbf{K}\nabla\hat{p}\cdot\mathbf{n}d\sigma = \int_{\gamma_i} \mathbf{K}\nabla p\cdot\mathbf{n}d\sigma, \quad \gamma_i \subset \partial E \setminus \Gamma_{12},\ i = 1, \ldots, d-k, \tag{3.21}$$

$$\int_E \hat{p}dx = \int_E p dx, \tag{3.22}$$

where k is the number of sides that E shares with Γ_{12}. Note that condition (3.21) is empty if $k = d$. It is easy to see that, if $\mathbf{K}$ is a constant on E, (3.20)–(3.22) define $\hat{p}$ uniquely and that $\hat{p}$ is exact for linears. If $\mathbf{K}$ is not a constant on E, an extension similar to the one in [39] can be used. In all cases, $\hat{p}$ has optimal approximation properties, namely, on each $E \in \mathcal{E}_h(\Omega)$,

$$|p - \hat{p}|_{k,E} \leq C h_E^{r+1-k}|p|_{r+1,E}, \quad k = 0,\ 1,\ 2. \tag{3.23}$$

Note that, for $\gamma \subset \partial E$, the trace inequality [7] $|\phi|_{k,\gamma} \leq C(h_E^{\frac{1}{2}}|\phi|_{k+1,E} + h_E^{-\frac{1}{2}}|\phi|_{k,E})$ implies that

$$|p - \hat{p}|_{k,\gamma} \leq C h_E^{r+\frac{1}{2}-k}|p|_{r+1,E}, \quad k = 0,\ 1. \tag{3.24}$$

More generally, we have the following lemma.

LEMMA 3.2. *Let $\mathcal{E}_h(\Omega_i)$, $i = 1, 2$, be nondegenerate. Then there exists a constant C, independent of h, such that for all p in $H^s(\Omega_i)$, $s > \frac{3}{2}$,*

$$\left(\sum_{\gamma\in\Gamma_h(\Omega_i)} h_\gamma \| \{\mathbf{K}\nabla(p - \hat{p})\cdot\mathbf{n}\} \|_{0,\gamma}^2 \right)^{\frac{1}{2}} \leq C\, h^{\mu-1}|p|_{\mu,\Omega_i}, \quad \mu = \min(r+1, s). \tag{3.25}$$

Proof. Consider a side γ adjacent to an element E. The result follows by switching to the reference element, applying a trace theorem, and using (3.23):

$$\begin{aligned} h_\gamma \|\mathbf{K}\nabla(p - \hat{p})|_E\cdot\mathbf{n}\|_{0,\gamma}^2 &\leq C\frac{h_\gamma|\gamma|}{|E|}(\lambda_E^{\max})^2 \left(|p - \hat{p}|_{1,E}^2 + h_E^{2(s-1)}|p - \hat{p}|_{s,E}^2 \right) \\ &\leq C\, h_E^{2(\mu-1)}|p|_{\mu,E}^2. \qquad \square \end{aligned}$$

THEOREM 3.3. *Let p be a solution of (2.1)–(2.2). Let (p_h, λ_H) be a solution of (3.8)–(3.9). We assume that $p \in H^s(\Omega)$ for some real number $s > \frac{3}{2}$ and that σ_γ^0 is sufficiently large for SIPG and IIPG. Then there exists a constant C, independent of h and H, such that*

$$\begin{aligned}&\left\|\mathbf{K}^{\frac{1}{2}}\nabla\left(p_h-p\right)\right\|_{0,\Omega}+\sqrt{\sum_{\gamma\in\Gamma_h(\Omega)}\frac{\sigma_\gamma}{h_\gamma}\left\|[p_h]\right\|_{0,\gamma}^2}+\sqrt{\sum_{\tau\in\Gamma_H}\frac{\sigma_\tau}{H_\tau}\sum_{i=1,2}\left\|p_h|_{\Omega_i}-\lambda_H\right\|_{0,\tau}^2}\\&\leq C\left(h^{\mu-1}\left(\frac{H}{h}\right)^{\frac{1}{2}}+H^{\bar{\mu}-\frac{1}{2}}\right),\quad \mu=\min\left(r+1,s\right),\ \bar{\mu}=\min\left(\bar{r}+1,s-\frac{1}{2}\right).\end{aligned}$$

Proof. Since by assumption p is smooth enough, we let $\bar{p} \in \Lambda_H$ be the continuous nodal interpolant of p and define

$$\eta := \lambda_H - p, \quad \eta^I := p - \bar{p}, \quad \eta^A := \lambda_H - \bar{p} = \eta + \eta^I.$$

Define

$$\xi := p_h - p, \quad \xi^I := p - \hat{p}, \quad \xi^A := p_h - \hat{p} = \xi + \xi^I.$$

The interpolant $\bar{p}$ satisfies [19]

$$|p - \bar{p}|_{k,\Gamma_{12}} \leq C H^{\bar{\mu}-k} |p|_{\bar{\mu}+1/2,\Omega_i},\ 0 \leq k \leq 1, \tag{3.26}$$

where the bound for fractional k is obtained by interpolation between $L^2(\Gamma_{12})$ and $H^1(\Gamma_{12})$.

By subtracting the weak formulation (3.3) from the finite element scheme (3.8) and choosing $q_h = \xi^A$, we obtain

$$B_i(\xi^A, \xi^A) = L_i(\xi^A; \lambda_H) - L_i(\xi^A; p) + B_i(\xi^I, \xi^A), \quad i = 1, 2.$$

Summation over the two subdomains leads to

$$\begin{aligned}\sum_{i=1}^{2} B_i(\xi^A, \xi^A) &= \sum_{i=1}^{2}\big(B_i(\xi^I, \xi^A) + L_i(\xi^A; \lambda_H) - L_i(\xi^A; p)\big)\\&= \sum_{i=1}^{2} B_i(\xi^I, \xi^A) + \int_{\Gamma_{12}} \left[\mathbf{K}\nabla\xi^A \cdot \mathbf{n}\right] \eta d\sigma + \sum_{i=1}^{2}\sum_{\tau\in\Gamma_H}\frac{\sigma_\tau}{H_\tau}\int_\tau \xi^A|_{\Omega_i}\,\eta d\sigma.\end{aligned} \tag{3.27}$$

Similarly, by subtracting (3.4) from (3.9), with $\mu = \mu_H = \eta^A$, and noting that p and $\mathbf{K}\nabla p \cdot \mathbf{n}$ are continuous across the interface Γ_{12}, we obtain

$$\begin{aligned}&\int_{\Gamma_{12}} \left[\mathbf{K}\nabla\xi^A \cdot \mathbf{n}\right] \eta^A d\sigma - \sum_{i=1,2}\sum_{\tau\in\Gamma_H}\frac{\sigma_\tau}{H_\tau}\int_\tau (\xi^A|_{\Omega_i} - \eta^A)\eta^A d\sigma\\&= \int_{\Gamma_{12}} \left[\mathbf{K}\nabla\xi^I \cdot \mathbf{n}\right] \eta^A d\sigma - \sum_{i=1,2}\sum_{\tau\in\Gamma_H}\frac{\sigma_\tau}{H_\tau}\int_\tau (\xi^I|_{\Omega_i} - \eta^I)\eta^A d\sigma.\end{aligned} \tag{3.28}$$

Summing (3.27) and (3.28) results in

(3.29)

$$\sum_{i=1}^{2} B_i(\xi^A,\xi^A) - \sum_{i=1,2}\sum_{\tau\in\Gamma_H}\frac{\sigma_\tau}{H_\tau}\int_\tau \xi^A\big|_{\Omega_i}\eta^A d\sigma - \sum_{i=1,2}\sum_{\tau\in\Gamma_H}\frac{\sigma_\tau}{H_\tau}\int_\tau (\xi^A\big|_{\Omega_i}-\eta^A)\eta^A d\sigma$$
$$= \sum_{i=1}^{2} B_i(\xi^I,\xi^A) - \int_{\Gamma_{12}}\left[\mathbf{K}\nabla\xi^A\cdot\mathbf{n}\right]\eta^I d\sigma + \int_{\Gamma_{12}}\left[\mathbf{K}\nabla\xi^I\cdot\mathbf{n}\right]\eta^A d\sigma$$
$$- \sum_{i=1,2}\sum_{\tau\in\Gamma_H}\frac{\sigma_\tau}{H_\tau}\int_\tau \xi^A\big|_{\Omega_i}\eta^I d\sigma - \sum_{i=1,2}\sum_{\tau\in\Gamma_H}\frac{\sigma_\tau}{H_\tau}\int_\tau (\xi^I\big|_{\Omega_i}-\eta^I)\eta^A d\sigma.$$

We expand the first term on the left-hand side of (3.29):

$$\sum_{i=1}^{2} B_i(\xi^A,\xi^A) = \|\mathbf{K}^{\frac{1}{2}}\nabla\xi^A\|_{0,\Omega}^2 - (1+s_{\text{form}})\sum_{\gamma\in\Gamma_h(\Omega)}\int_\gamma\left\{\mathbf{K}\nabla\xi^A\cdot\mathbf{n}\right\}\left[\xi^A\right]d\sigma$$
$$+\sum_{\gamma\in\Gamma_h(\Omega)}\frac{\sigma_\gamma}{h_\gamma}\int_\gamma\left[\xi^A\right]^2 d\sigma + \sum_{i=1,2}\sum_{\tau\in\Gamma_H}\frac{\sigma_\tau}{H_\tau}\int_\tau \xi^A\big|_{\Omega_i}^2 d\sigma.$$

We denote by L_{ErrEqu} and R_{ErrEqu} the left-hand and right-hand sides of (3.29), respectively. An algebraic manipulation yields

$$L_{\text{ErrEqu}} = \|\mathbf{K}^{\frac{1}{2}}\nabla\xi^A\|_{0,\Omega}^2 - (1+s_{\text{form}})\sum_{\gamma\in\Gamma_h(\Omega)}\int_\gamma\left\{\mathbf{K}\nabla\xi^A\cdot\mathbf{n}\right\}\left[\xi^A\right]d\sigma$$
$$+\sum_{\gamma\in\Gamma_h(\Omega)}\frac{\sigma_\gamma}{h_\gamma}\int_\gamma\left[\xi^A\right]^2 d\sigma + \sum_{i=1,2}\sum_{\tau\in\Gamma_H}\frac{\sigma_\tau}{H_\tau}\int_\tau\left(\xi^A\big|_{\Omega_i}-\eta^A\right)^2 d\sigma.$$

For NIPG and OBB-DG, the second term in L_{ErrEqu} vanishes, leaving only the coercive terms. For SIPG and IIPG, we employ the inequality (3.10) with $q_h = q = \xi^A$ to conclude that

$$L_{\text{ErrEqu}} \geq \frac{1}{2}\|\mathbf{K}^{\frac{1}{2}}\nabla\xi^A\|_{0,\Omega}^2 + \frac{1}{2}\sum_{\gamma\in\Gamma_h(\Omega)}\frac{\sigma_\gamma}{h_\gamma}\int_\gamma\left[\xi^A\right]^2 d\sigma$$
$$+\frac{1}{2}\sum_{i=1,2}\sum_{\tau\in\Gamma_H}\frac{\sigma_\tau}{H_\tau}\int_\tau\left(\xi^A\big|_{\Omega_i}-\eta^A\right)^2 d\sigma.$$

We now consider the right-hand side of (3.29). By expanding its first term as

$$\sum_{i=1}^{2} B_i(\xi^I,\xi^A) = \sum_{E\in\mathcal{E}_h(\Omega)}\int_E \mathbf{K}\nabla\xi^I\cdot\nabla\xi^A dx - \sum_{\gamma\in\Gamma_h(\Omega)}\int_\gamma\left\{\mathbf{K}\nabla\xi^I\cdot\mathbf{n}\right\}\left[\xi^A\right]d\sigma$$
$$- s_{\text{form}}\sum_{\gamma\in\Gamma_h(\Omega)}\int_\gamma\left\{\mathbf{K}\nabla\xi^A\cdot\mathbf{n}\right\}\left[\xi^I\right]d\sigma - \int_{\Gamma_{12}}\left[\mathbf{K}\nabla\xi^I\cdot\mathbf{n}\xi^A\right]d\sigma$$
$$+\int_{\Gamma_{12}}\left[\mathbf{K}\nabla\xi^A\cdot\mathbf{n}\xi^I\right]d\sigma + \sum_{\gamma\in\Gamma_h(\Omega)}\frac{\sigma_\gamma}{h_\gamma}\int_\gamma\left[\xi^I\right]\left[\xi^A\right]d\sigma + \sum_{i=1,2}\sum_{\tau\in\Gamma_H}\frac{\sigma_\tau}{H_\tau}\int_\tau(\xi^I\xi^A)|_{\Omega_i}d\sigma$$

and using the fact that η^I and η^A are uniquely defined on Γ_{12}, we have

$$\begin{aligned}
R_{\text{ErrEqu}} &= \sum_{E\in\mathcal{E}_h(\Omega)}\int_E \mathbf{K}\nabla\xi^I\cdot\nabla\xi^A dx - \sum_{\gamma\in\Gamma_h(\Omega)}\int_\gamma \left\{\mathbf{K}\nabla\xi^I\cdot\mathbf{n}\right\}\left[\xi^A\right]d\sigma \\
&- s_{\text{form}}\sum_{\gamma\in\Gamma_h(\Omega)}\int_\gamma \left\{\mathbf{K}\nabla\xi^A\cdot\mathbf{n}\right\}\left[\xi^I\right]d\sigma + \sum_{\gamma\in\Gamma_h(\Omega)}\frac{\sigma_\gamma}{h_\gamma}\int_\gamma\left[\xi^I\right]\left[\xi^A\right]d\sigma \\
&- \sum_{\tau\in\Gamma_H}\frac{\sigma_\tau}{H_\tau}\int_\tau\sum_{i=1,2}(\xi^I|_{\Omega_i}-\eta^I)(\xi^A|_{\Omega_i}-\eta^A)d\sigma \\
&- \int_{\Gamma_{12}}\left[\mathbf{K}\nabla\xi^A\cdot\mathbf{n}(\eta^I-\xi^I)\right]d\sigma + \int_{\Gamma_{12}}\left[\mathbf{K}\nabla\xi^I\cdot\mathbf{n}(\eta^A-\xi^A)\right]d\sigma =: \sum_{i=1}^7 T_i.
\end{aligned}$$

We now bound each term T_i of R_{ErrEqu}. We first bound T_2 for NIPG, SIPG, and IIPG:

$$|T_2| \le \epsilon\sum_{\gamma\in\Gamma_h(\Omega)}\frac{\sigma_\gamma}{h_\gamma}\int_\gamma\left[\xi^A\right]^2 d\sigma + Ch^{2\mu-2}.$$

For OBB-DG we use an argument from [39] and the property (3.20) of $\hat{p}$, which holds for all element sides. On each side $\gamma=\partial E_1\cap\partial E_2$, let $c_\gamma = c_\gamma^1 - c_\gamma^2$, where c_γ^i is the mean value of ξ^A on E_i. We have

$$\begin{aligned}
|T_2| &= \left|\sum_{\gamma\in\Gamma_h(\Omega)}\int_\gamma\left\{\mathbf{K}\nabla\xi^I\cdot\mathbf{n}\right\}\left(\left[\xi^A\right]-c_\gamma\right)d\sigma\right| \\
&\le \sum_{\gamma\in\Gamma_h(\Omega)}\left(\frac{\tilde{\epsilon}}{h_\gamma}\|[\xi^A]-c_\gamma\|_{0,\gamma}^2 + Ch_\gamma\|\{\mathbf{K}\nabla\xi^I\cdot\mathbf{n}\}\|_{0,\gamma}^2\right) \le \epsilon|\!|\!|\mathbf{K}^{\frac12}\nabla\xi^A|\!|\!|_{0,\Omega}^2 + Ch^{2\mu-2},
\end{aligned}$$

where we used the fact that

$$\|[\xi^A]-c_\gamma\|_{0,\gamma} \le \|\xi^A|_{E_1}-c_\gamma^1\|_{0,\gamma} + \|\xi^A|_{E_2}-c_\gamma^2\|_{0,\gamma} \le Ch_\gamma^{\frac12}(|\xi^A|_{1,E_1}+|\xi^A|_{1,E_2}).$$

We continue with bounds on the rest of the terms T_i of R_{ErrEqu} for all methods:

$$\begin{aligned}
|T_1| &\le \epsilon|\!|\!|\mathbf{K}^{\frac12}\nabla\xi^A|\!|\!|_{0,\Omega}^2 + C|\!|\!|\mathbf{K}^{\frac12}\nabla\xi^I|\!|\!|_{0,\Omega}^2 \le \epsilon|\!|\!|\mathbf{K}^{\frac12}\nabla\xi^A|\!|\!|_{0,\Omega}^2 + Ch^{2\mu-2}, \\
|T_3| &\le \tilde{\epsilon}\sum_{\gamma\in\Gamma_h(\Omega)}h_\gamma\int_\gamma\left\{\mathbf{K}\nabla\xi^A\cdot\mathbf{n}\right\}^2 d\sigma + C\sum_{\gamma\in\Gamma_h(\Omega)}\frac{1}{h_\gamma}\int_\gamma\left[\xi^I\right]^2 d\sigma \\
&\le \epsilon|\!|\!|\mathbf{K}^{\frac12}\nabla\xi^A|\!|\!|_{0,\Omega}^2 + Ch^{2\mu-2}, \\
|T_4| &\le \epsilon\sum_{\gamma\in\Gamma_h(\Omega)}\frac{\sigma_\gamma}{h_\gamma}\int_\gamma\left[\xi^A\right]^2 d\sigma + Ch^{2\mu-2}, \\
|T_5| &\le \epsilon\sum_{\tau\in\Gamma_H}\frac{\sigma_\tau}{H_\tau}\int_\tau\sum_{i=1,2}\left(\xi^A|_{\Omega_i}-\eta^A\right)^2 d\sigma + C\sum_{\tau\in\Gamma_H}\frac{\sigma_\tau}{H_\tau}\int_\tau\sum_{i=1,2}\xi^I|_{\Omega_i}^2\, d\sigma \\
&\quad + C\sum_{\tau\in\Gamma_H}\frac{\sigma_\tau}{H_\tau}\int_\tau\sum_{i=1,2}\left(\eta^I\right)^2 d\sigma \\
&\le \epsilon\sum_{\tau\in\Gamma_H}\frac{\sigma_\tau}{H_\tau}\int_\tau\sum_{i=1,2}\left(\xi^A|_{\Omega_i}-\eta^A\right)^2 d\sigma + Ch^{2\mu-1}H^{-1} + CH^{2\bar{\mu}}H^{-1} \\
&\le \epsilon\sum_{\tau\in\Gamma_H}\frac{\sigma_\tau}{H_\tau}\int_\tau\sum_{i=1,2}\left(\xi^A|_{\Omega_i}-\eta^A\right)^2 d\sigma + Ch^{2\mu-2} + CH^{2\bar{\mu}-1}.
\end{aligned}$$

To handle term T_6, we use the special properties (3.20)–(3.22) of the interpolant $\hat{p}$ on the interface Γ_{12}. We define

$$T_{6,i} := -\int_{\Gamma_{12}} \mathbf{K}\nabla\xi^A \cdot \mathbf{n}_{\partial\Omega_i}(\eta^I - \xi^I)\big|_{\Omega_i}\, d\sigma, \quad i = 1, 2.$$

Obviously, we have $T_6 = T_{6,1} + T_{6,2}$. We bound $T_{6,1}$, suppressing the subscript $\partial\Omega_1$ for simplicity; the bound for $T_{6,2}$ is similar. By using the same argument as for bounding T_3, it is easy to see that

$$|T_{6,1}| \le T_A + T_B + \tilde{\epsilon}\|\mathbf{K}^{\frac{1}{2}}\nabla\xi^A\|^2_{0,\Omega_1} + Ch^{2\mu-2},$$

where

$$T_A := \int_{\Gamma_{12}} \left|\mathbf{K}\nabla\xi^I \cdot \mathbf{n}\eta^I\right| d\sigma, \quad T_B := \int_{\Gamma_{12}} \left|\mathbf{K}\nabla\xi \cdot \mathbf{n}\eta^I\right| d\sigma.$$

To bound term T_A, let $P_{h,1}$ be the L^2-projection onto the space of piecewise constants on $\mathcal{E}_h(\Omega_1) \cap \Gamma_{12}$, use (3.20), and revert to the reference element to recover the $H^{\frac{1}{2}}$ norm in the second factor below; this gives

$$\begin{aligned} T_A &= \int_{\Gamma_{12}} \left|\mathbf{K}\nabla(p - \hat{p}) \cdot \mathbf{n}(\eta^I - P_{h,1}\eta^I)\right| d\sigma \\ &\le C\left(\sum_{\gamma\in\Gamma_{12}} h_\gamma \|\mathbf{K}\nabla(p - \hat{p}) \cdot \mathbf{n}\|^2_{0,\gamma}\right)^{\frac{1}{2}} \left(\sum_{\gamma\in\Gamma_{12}} \left\|\eta^I\right\|^2_{\frac{1}{2},\gamma}\right)^{\frac{1}{2}}. \end{aligned}$$

Then we apply (3.25) to the first factor and (3.26) with $k = \frac{1}{2}$ to the second factor; this gives

$$T_A \le Ch^{\mu-1}|p|_{\mu,\Omega_1} H^{\bar{\mu}-\frac{1}{2}}|p|_{\bar{\mu}+\frac{1}{2},\Omega_1} \le Ch^{2\mu-2} + CH^{2\bar{\mu}-1}.$$

To bound term T_B, we first subtract (3.3) from (3.8) and take the test function to be a piecewise constant on $\mathcal{E}_h(\Omega_1)$, to obtain the discrete mass balance equation

$$\begin{aligned} &-\sum_{\gamma\in\Gamma_h(\Omega_1)} \int_\gamma \{\mathbf{K}\nabla(p_h - p) \cdot \mathbf{n}\}\,[q]\, d\sigma - \int_{\Gamma_{12}} \mathbf{K}\nabla(p_h - p) \cdot \mathbf{n}q|_{\Omega_1} d\sigma \\ &+\sum_{\gamma\in\Gamma_h(\Omega_1)} \frac{\sigma_\gamma}{h_\gamma} \int_\gamma [p_h]\,[q]\, d\sigma + \sum_{\tau\in\Gamma_H} \frac{\sigma_\tau}{H_\tau} \int_\tau \left(p_h|_{\Omega_1} - \lambda_H\right) q d\sigma = 0. \end{aligned} \tag{3.30}$$

Let $\tilde{\eta}^I = P_1(\eta^I)$, where $P_1 \in \mathcal{L}(H^{\frac{1}{2}}(\Gamma_{12}); H^1(\Omega_1))$ is an extension operator, and let $\mathcal{Q}_h$ be the L^2-projection onto the space of piecewise constants on $\mathcal{E}_h(\Omega_1)$. Using (3.30) and the continuity of the trace of $\tilde{\eta}^I$ across interior interfaces, the term T_B has the following expression:

$$\begin{aligned} &\int_{\Gamma_{12}} \mathbf{K}\nabla\xi \cdot \mathbf{n}\eta^I d\sigma \\ &= \sum_{\gamma\in\Gamma_h(\Omega_1)} \int_\gamma \{\mathbf{K}\nabla(p_h - p) \cdot \mathbf{n}\}\left[\tilde{\eta}^I\right] d\sigma + \int_{\Gamma_{12}} \mathbf{K}\nabla(p_h - p) \cdot \mathbf{n}\tilde{\eta}^I d\sigma \\ &= \sum_{\gamma\in\Gamma_h(\Omega_1)} \int_\gamma \{\mathbf{K}\nabla(p_h - p) \cdot \mathbf{n}\}\left[\tilde{\eta}^I - \mathcal{Q}_h\tilde{\eta}^I\right] d\sigma + \int_{\Gamma_{12}} \mathbf{K}\nabla(p_h - p) \cdot \mathbf{n}(\tilde{\eta}^I - \mathcal{Q}_h\tilde{\eta}^I) d\sigma \\ &\quad + \sum_{\gamma\in\Gamma_h(\Omega_1)} \frac{\sigma_\gamma}{h_\gamma} \int_\gamma [p_h]\left[\mathcal{Q}_h\tilde{\eta}^I\right] d\sigma + \sum_{\tau\in\Gamma_H} \frac{\sigma_\tau}{H_\tau} \int_\tau \left(p_h|_{\Omega_1} - \lambda_H\right) \mathcal{Q}_h\tilde{\eta}^I d\sigma =: \sum_{i=1}^4 T_{B,i}. \end{aligned}$$

To bound the first term, observe that, by mapping to the reference element, using a trace theorem and the regularity of $\mathcal{E}_h(\Omega_1)$, we have on any segment γ adjacent to an element E

$$\|\tilde{\eta}^I - \mathcal{Q}_h\tilde{\eta}^I\|_{0,\gamma} \le Ch_\gamma^{\frac{1}{2}}|\tilde{\eta}^I|_{1,E}.$$

Applying the continuity of the extension operator P_1 gives

$$|T_{B,1}| \le C\left(\sum_{\gamma\in\Gamma_h(\Omega_1)} h_\gamma\|\mathbf{K}\nabla(p_h-p)\|_{0,\gamma}^2\right)^{\frac{1}{2}} \|\eta^I\|_{\frac{1}{2},\Gamma_{12}}.$$

For the first factor we write $p_h - p = \xi^A + \xi^I$ and apply (3.11), (3.25), and (3.26), thus deriving

$$\begin{aligned}|T_{B,1}| &\le CH^{\bar{\mu}-\frac{1}{2}}|p|_{\bar{\mu}+\frac{1}{2},\Omega_1}\left(h^{\mu-1}|p|_{\mu,\Omega_1} + \|\mathbf{K}^{\frac{1}{2}}\nabla\xi^A\|_{0,\Omega_1}\right)\\ &\le \epsilon\|\mathbf{K}^{\frac{1}{2}}\nabla\xi^A\|_{0,\Omega_1}^2 + C(h^{2\mu-2}+H^{2\bar{\mu}-1}),\end{aligned}$$

with a similar bound for $|T_{B,2}|$. The remaining terms are bounded as follows:

$$|T_{B,3}| = \left|\sum_{\gamma\in\Gamma_h(\Omega_1)}\frac{\sigma_\gamma}{h_\gamma}\int_\gamma [p_h]\left[\mathcal{Q}_h\tilde{\eta}^I - \tilde{\eta}^I\right]d\sigma\right| \le \epsilon\sum_{\gamma\in\Gamma_h(\Omega_1)}\frac{\sigma_\gamma}{h_\gamma}\|[p_h]\|_{0,\gamma}^2 + CH^{2\bar{\mu}-1},$$

$$\begin{aligned}|T_{B,4}| &\le \left|\sum_{\tau\in\Gamma_H}\frac{\sigma_\tau}{H_\tau}\int_\tau\left(p_h|_{\Omega_1}-\lambda_H\right)\left(\mathcal{Q}_h\tilde{\eta}^I-\tilde{\eta}^I\right)d\sigma\right| + \left|\sum_{\tau\in\Gamma_H}\frac{\sigma_\tau}{H_\tau}\int_\tau\left(p_h|_{\Omega_1}-\lambda_H\right)\tilde{\eta}^I d\sigma\right|\\ &\le \epsilon\sum_{\tau\in\Gamma_H}\frac{\sigma_\tau}{H_\tau}\left\|p_h|_{\Omega_1}-\lambda_H\right\|_{0,\tau}^2 + C(H^{2\bar{\mu}-2}h + H^{2\bar{\mu}-1}).\end{aligned}$$

By combining the results for T_A and $T_{B,i}$, we obtain

$$\begin{aligned}|T_{6,1}| \le \epsilon\|\mathbf{K}^{\frac{1}{2}}\nabla\xi^A\|_{0,\Omega_1}^2 &+ \epsilon\sum_{\gamma\in\Gamma_h(\Omega_1)}\frac{\sigma_\gamma}{h_\gamma}\|[p_h]\|_{0,\gamma}^2 + \epsilon\sum_{\tau\in\Gamma_H}\frac{\sigma_\tau}{H_\tau}\left\|\xi^A|_{\Omega_1}-\eta^A\right\|_{0,\tau}^2\\ &+ C(h^{2\mu-2}+H^{2\bar{\mu}-1}),\end{aligned}$$

where we have also used (3.24) and (3.26). The last term T_7 can be bounded as follows:

$$\begin{aligned}|T_7| &\le \left(\sum_{i=1,2}\sum_{\tau\in\Gamma_H}\frac{\sigma_\tau}{H_\tau}\|\xi^A|_{\Omega_i}-\eta^A\|_{0,\tau}^2\right)^{\frac{1}{2}}\left(\sum_{i=1,2}\sum_{\tau\in\Gamma_H}\frac{H_\tau}{\sigma_\tau}\|\mathbf{K}\nabla\xi^I|_{\Omega_i}\cdot\mathbf{n}\|_{0,\tau}^2\right)^{\frac{1}{2}}\\ &\le \epsilon\sum_{i=1,2}\sum_{\tau\in\Gamma_H}\frac{\sigma_\tau}{H_\tau}\|\,\xi^A\big|_{\Omega_i}-\eta^A\|_{0,\tau}^2 + Ch^{2\mu-3}H.\end{aligned}$$

Thus

$$\begin{aligned}\|\mathbf{K}^{\frac{1}{2}}\nabla\xi^A\|_{0,\Omega_1}^2 &+ \sum_{\gamma\in\Gamma_h(\Omega)}\frac{\sigma_\gamma}{h_\gamma}\left\|\left[\xi^A\right]\right\|_{0,\gamma}^2 + \sum_{\tau\in\Gamma_H}\frac{\sigma_\tau}{H_\tau}\sum_{i=1,2}\left\|\xi^A|_{\Omega_i}-\eta^A\right\|_{0,\tau}^2\\ &\le Ch^{2\mu-2}\left(\frac{H}{h}\right) + CH^{2\bar{\mu}-1}.\end{aligned}$$

The proof is completed by using the triangle inequality. ∎

Remark 3.1. It is also possible to choose $\bar{s}_{\text{form}} = 0$ or 1, independently of s_{form}. For example, taking $\bar{s}_{\text{form}} = 1$ in SIPG preserves the symmetry of the method. These choices, however, require fine-scale related weights in the interface penalty terms of $B_i(\cdot,\cdot)$ and $L_i(\cdot;\cdot)$, in order to control the terms involving integrals on Γ_{12}. More precisely, the last terms of $B_i(\cdot,\cdot)$ and $L_i(\cdot;\cdot)$ in (3.8)–(3.9) need to be replaced by

$$\sum_{\gamma\in\Gamma_{h,i}} \frac{\sigma_\gamma}{h_\gamma}\int_\gamma p_h|_{\Omega_i} q_h|_{\Omega_i} d\sigma \quad \text{and} \quad \sum_{\gamma\in\Gamma_{h,i}} \frac{\sigma_\gamma}{h_\gamma}\int_\gamma q_h|_{\Omega_i}\lambda_H\, d\sigma, \tag{3.31}$$

respectively. Here $\Gamma_{h,i}$ denotes the trace of $\mathcal{E}_h(\Omega_i)$ on Γ_{12}. Similarly, the interface pressure continuity term in (3.4) should be replaced by

$$\sum_{i=1,2}\sum_{\gamma\in\Gamma_{h,i}} \frac{\sigma_\gamma}{h_\gamma}\int_\gamma (p_h|_{\Omega_i} - \lambda_H)\mu_H d\sigma.$$

Then the terms involving integrals on Γ_{12} can be controlled by the terms in (3.31) and $|\!|\!| \cdot |\!|\!|_{0,\Omega_i}$, via the inequality, for all $q_h \in X_h(\Omega_i)$ and all $\mu \in L^2(\Gamma_{12})$,

$$\left|\int_{\Gamma_{12}} \mathbf{K}\nabla q_h\cdot\mathbf{n}\,\mu d\sigma\right| \le \frac{1}{8}|\!|\!|\mathbf{K}^{\frac{1}{2}}\nabla q_h|\!|\!|^2_{0,\Omega_i} + \frac{1}{8}\sum_{\gamma\in\Gamma_{h,i}}\frac{\sigma_\gamma}{h_\gamma}\int_\gamma \mu^2 d\sigma, \tag{3.32}$$

assuming the weights σ_γ are sufficiently large. Note that (3.32) can be shown in a way similar to (3.10). It can then be seen that all modified mortar DG methods are well-posed; i.e., Theorem 3.2 holds for $\bar{s}_{\text{form}} = 0$, 1, or -1. Moreover, under the assumptions of Theorem 3.3, following its argument, we can show that there exists a constant C, independent of h and H, such that

$$|\!|\!|\mathbf{K}^{\frac{1}{2}}\nabla\left(p_h - p\right)|\!|\!|_{0,\Omega} + \sqrt{\sum_{\gamma\in\Gamma_h(\Omega)}\frac{\sigma_\gamma}{h_\gamma}\left\|[p_h]\right\|^2_{0,\gamma}} + \sqrt{\sum_{i=1,2}\sum_{\gamma\in\Gamma_{h,i}}\frac{\sigma_\gamma}{h_\gamma}\left\|p_h|_{\Omega_i} - \lambda_H\right\|^2_{0,\gamma}}$$
$$\le C\left(h^{\mu-1} + H^{\bar{\mu}-\frac{1}{2}}\left(\frac{H}{h}\right)^{\frac{1}{2}}\right), \quad \mu = \min\left(r+1, s\right),\ \bar{\mu} = \min\left(\bar{r}+1, s-\frac{1}{2}\right).$$

Note that, if $H = O(h)$, the above bound and the estimate from Theorem 3.3 provide the same (optimal) asymptotic convergence. However, in the multiscale case $H = O(h^\alpha)$, $0 < \alpha < 1$, the bound from Theorem 3.3 is better.

3.5. The interface operator and the reduced problem. In the following we present a nonoverlapping domain decomposition algorithm, which involves the reduction of the coupled system to an interface problem in the mortar space. We are motivated by the algorithms developed in [26, 4].

Let us split $L_i(q,\lambda)$ into a sum of two terms:

$$L_i(q;\lambda) = l_i(q) + b_i(\lambda, q),$$

where

$$l_i(q) := \int_{\Omega_i} fq dx - \int_{\partial\Omega_i\setminus\Gamma_{12}} gq d\sigma, \tag{3.33}$$

$$\begin{aligned} b_i(\lambda,q) &:= -\bar{s}_{\text{form}}\int_{\Gamma_{12}}\mathbf{K}\nabla q\cdot\mathbf{n}_{\partial\Omega_i}\lambda d\sigma + \sum_{\tau\in\Gamma_H}\frac{\sigma_\tau}{H_\tau}\int_\tau q|_{\Omega_i}\lambda\, d\sigma \\ &:= \int_{\Gamma_{12}}\mathbf{K}\nabla q\cdot\mathbf{n}_{\partial\Omega_i}\lambda d\sigma + \sum_{\tau\in\Gamma_H}\frac{\sigma_\tau}{H_\tau}\int_\tau q|_{\Omega_i}\lambda\, d\sigma, \end{aligned} \tag{3.34}$$

since we take $\bar{s}_{\text{form}} = -1$. We comment on the choices $\bar{s}_{\text{form}} = 0$ or 1 in Remark 3.3.

Define a bilinear form $d_H : L^2(\Gamma_{12}) \times L^2(\Gamma_{12}) \mapsto \mathbb{R}$ for $\lambda, \mu \in L^2(\Gamma_{12})$ by

(3.35)
$$d_H(\lambda,\mu) := \sum_{i=1}^{2}\left(\int_{\Gamma_{12}} \mathbf{K}\,\nabla p_h^*(\lambda)|_{\Omega_i}\cdot \mathbf{n}_{\partial\Omega_i}\mu d\sigma - \sum_{\tau\in\Gamma_H}\frac{\sigma_\tau}{H_\tau}\int_\tau (p_h^*(\lambda)|_{\Omega_i}-\lambda)\mu d\sigma\right),$$

where $p_h^*(\lambda)|_{\Omega_i} \in X_h(\Omega_i)$, $i = 1, 2$, is the solution of

$$B_i\left(p_h^*(\lambda), q_h\right) = b_i\left(\lambda, q_h\right), \quad q_h \in X_h(\Omega_i),\ i = 1 \text{ and } 2. \tag{3.36}$$

Define a linear functional $g_H : L^2(\Gamma_{12}) \mapsto \mathbb{R}$ by

$$g_H(\mu) := -\sum_{i=1}^{2}\left(\int_{\Gamma_{12}} \mathbf{K}\,\nabla \bar{p}_h|_{\Omega_i}\cdot \mathbf{n}_{\partial\Omega_i}\mu d\sigma - \sum_{\tau\in\Gamma_H}\frac{\sigma_\tau}{H_\tau}\int_\tau \bar{p}_h|_{\Omega_i}\mu d\sigma\right),$$

where $\bar{p}_h|_{\Omega_i} \in X_h(\Omega_i)$, $i = 1, 2$, solves

$$B_i\left(\bar{p}_h, q_h\right) = l_i\left(q_h\right), \quad q_h \in X_h(\Omega_i),\ i = 1 \text{ and } 2.$$

It is easy to see that the solution (p_h, λ_H) of the DG-DG scheme (3.8)–(3.9) satisfies

$$d_H\left(\lambda_H, \mu_H\right) = g_H\left(\mu_H\right), \quad \mu_H \in \Lambda_H, \tag{3.37}$$

with

$$p_h = p_h^*(\lambda_H) + \bar{p}_h.$$

We now analyze the properties of the bilinear form $d_H(\cdot,\cdot)$ for the various DG schemes. We first let $q_h = p_h^*(\mu)$ in (3.36) for some $\mu \in L^2(\Gamma_{12})$ to obtain

$$B_i\left(p_h^*(\lambda), p_h^*(\mu)\right) = b_i\left(\lambda, p_h^*(\mu)\right), \quad i = 1 \text{ and } 2. \tag{3.38}$$

By using (3.35), (3.34), and (3.38), we have

(3.39)
$$\begin{aligned} d_H\left(\lambda,\mu\right) &= \sum_{i=1}^{2} b_i(\mu, p_h^*(\lambda)) - 2\sum_{i=1}^{2}\sum_{\tau\in\Gamma_H}\frac{\sigma_\tau}{H_\tau}\int_\tau p_h^*(\lambda)|_{\Omega_i}\mu\, d\sigma + 2\sum_{\tau\in\Gamma_H}\frac{\sigma_\tau}{H_\tau}\int_\tau \lambda\mu d\sigma \\ &= \sum_{i=1}^{2} B_i\left(p_h^*(\mu), p_h^*(\lambda)\right) - 2\sum_{i=1}^{2}\sum_{\tau\in\Gamma_H}\frac{\sigma_\tau}{H_\tau}\int_\tau p_h^*(\lambda)|_{\Omega_i}\mu\, d\sigma + 2\sum_{\tau\in\Gamma_H}\frac{\sigma_\tau}{H_\tau}\int_\tau \lambda\mu d\sigma. \end{aligned}$$

Note that the above representation implies that the interface bilinear form $d_H(\cdot,\cdot)$ is nonsymmetric for all DG versions.

To show coercivity, using (3.39) and (3.1), we obtain

$$\begin{aligned} d_H\left(\lambda,\lambda\right) &= \|\mathbf{K}^{\frac{1}{2}}\nabla p_h^*(\lambda)\|_{0,\Omega}^2 - (1+s_{\text{form}})\sum_{\gamma\in\Gamma_h(\Omega)}\int_\gamma \{\mathbf{K}\nabla p_h^*(\lambda)\cdot\mathbf{n}\}\left[p_h^*(\lambda)\right] d\sigma \\ &\quad + \sum_{\gamma\in\Gamma_h(\Omega)}\frac{\sigma_\gamma}{h_\gamma}\int_\gamma \left[p_h^*(\lambda)\right]^2 d\sigma + \sum_{i=1,2}\sum_{\tau\in\Gamma_H}\frac{\sigma_\tau}{H_\tau}\int_\tau \left(p_h^*(\lambda)|_{\Omega_i}-\lambda\right)^2 d\sigma \\ &\geq \frac{1}{2}\|\mathbf{K}^{\frac{1}{2}}\nabla p_h^*(\lambda)\|_{0,\Omega}^2 + \frac{1}{2}\sum_{\gamma\in\Gamma_h(\Omega)}\frac{\sigma_\gamma}{h_\gamma}\int_\gamma \left[p_h^*(\lambda)\right]^2 d\sigma \\ &\quad + \sum_{i=1,2}\sum_{\tau\in\Gamma_H}\frac{\sigma_\tau}{H_\tau}\int_\tau \left(p_h^*(\lambda)|_{\Omega_i}-\lambda\right)^2 d\sigma, \end{aligned} \tag{3.40}$$

where we have used (3.10) for the inequality when $s_{\text{form}} = 0$ or 1. Therefore the interface bilinear form $d_H(\cdot,\cdot)$ is positive semidefinite on $L^2(\Gamma_{12})$. The argument in the solvability Theorem 3.2 implies that $d_H(\lambda,\lambda) = 0$ only if λ = constant. We summarize our results below.

THEOREM 3.4. *Let the assumptions of Theorem* 3.2 *hold. For all DG versions, the interface bilinear form $d_H(\cdot,\cdot)$ is nonsymmetric and positive semidefinite on $L^2(\Gamma_{12})$, with the kernel consisting of the constant functions.*

Remark 3.2. If the Dirichlet boundary condition is imposed on a part of $\partial\Omega$, Γ_D, such that $|\Gamma_D| > 0$, then $d_H(\cdot,\cdot)$ is positive definite on $L^2(\Gamma_{12})$.

Remark 3.3. For a general choice of $\bar{s}_{\text{form}}$, coercivity of $d_H(\cdot,\cdot)$ can be shown for all modified schemes introduced in Remark 3.1, by using inequality (3.32). Recall that taking $\bar{s}_{\text{form}} = 1$ in SIPG gives symmetric forms $B_i(\cdot,\cdot)$. It is easy to see that in this case $d_H(\cdot,\cdot)$ is also symmetric. In particular, by using (3.35) and (3.34) with modified penalty terms and (3.38), we obtain

$$\begin{aligned} d_H(\lambda,\mu) &= -\sum_{i=1}^{2}\left(b_i(\mu,p_h^*(\lambda)) + \sum_{\gamma\in\Gamma_{h,i}}\frac{\sigma_\gamma}{h_\gamma}\int_\gamma \lambda\mu d\sigma\right) \\ &= -\sum_{i=1}^{2}\left(B_i(p_h^*(\mu),p_h^*(\lambda)) + \sum_{\gamma\in\Gamma_{h,i}}\frac{\sigma_\gamma}{h_\gamma}\int_\gamma \lambda\mu d\sigma\right) = d_H(\mu,\lambda). \end{aligned}$$

Therefore in this case the conjugate gradient method can be employed for solving (3.37).

4. Coupling DG with MFE using a mortar space.

4.1. Weak formulation. We now consider the case where DG formulation is used in Ω_1 and a mixed formulation is used in Ω_2, the matching at the interface being achieved by a mortar multiplier. Recall that $W(\Omega_2) = L^2(\Omega_2)$, $\Lambda = H^{\frac{1}{2}}(\Gamma_{12})$, and $B_i(\cdot,\cdot)$ and $L_i(\cdot;\cdot)$ are defined in (3.1) and (3.2), respectively. In this section, we use only the definitions in Ω_1 and drop the subscripts for simplicity. That is, we denote $B_1(\cdot,\cdot)$ and $L_1(\cdot;\cdot)$ simply by $B(\cdot,\cdot)$ and $L(\cdot;\cdot)$, respectively.

Let $\mathbf{u}_g$ be the restriction to Ω_2 of an extension of g satisfying $\mathbf{u}_g \in H(\text{div};\Omega)$ and $\mathbf{u}_g\cdot\mathbf{n} = g$ on $\partial\Omega$. The coupled DG-mixed weak formulation is as follows: Find $p \in L^2(\Omega)$, such that $p|_{\Omega_1} \in X(\Omega_1)$, $p|_{\Omega_2} \in W(\Omega_2)$, $\mathbf{u} \in \mathbf{V}_0(\Omega_2) + \mathbf{u}_g$, and $\lambda \in \Lambda$ such that

$$B(p,q) = L(q;\lambda) \quad \forall q \in X(\Omega_1), \tag{4.1}$$

$$\int_{\Omega_2}\mathbf{K}^{-1}\mathbf{u}\cdot\mathbf{v}dx - \int_{\Omega_2} p\,\nabla\cdot\mathbf{v}dx = -\int_{\Gamma_{12}}\mathbf{v}\cdot\mathbf{n}_{\partial\Omega_2}\lambda d\sigma \quad \forall\,\mathbf{v}\in\mathbf{V}_0(\Omega_2), \tag{4.2}$$

$$\int_{\Omega_2}\nabla\cdot\mathbf{u}\,q\,dx = \int_{\Omega_2} f\,q\,dx \quad \forall\,q\in W(\Omega_2), \tag{4.3}$$

$$-\int_{\Gamma_{12}}(\mathbf{K}\nabla p|_{\Omega_1}\cdot\mathbf{n} + \mathbf{u}\cdot\mathbf{n})\mu d\sigma + \sum_{\tau\in\Gamma_H}\int_\tau\frac{\sigma_\tau}{H_\tau}(p|_{\Omega_1}-\lambda)\mu d\sigma = 0 \quad \forall\mu\in\Lambda. \tag{4.4}$$

As in section 3, we take $\bar{s}_{\text{form}} = -1$. The choices $\bar{s}_{\text{form}} = 0$ or 1 are discussed in Remarks 4.2 and 4.4.

4.2. Equivalence.

THEOREM 4.1. *If $(\mathbf{u},p,\lambda)$ is a solution of* (4.1)–(4.4)*, then p satisfies* (2.1)–(2.2) *in the sense of distributions. Conversely, if p is a sufficiently smooth solution of* (2.1)–(2.2)*, then there exists $\mathbf{u}$ and λ such that $(\mathbf{u},p,\lambda)$ solves* (4.1)–(4.4)*.*

Proof. By using the same arguments as in the DG-DG case, we conclude that $p|_{\Omega_1} \in H^1(\Omega_1)$ satisfies (2.1) in Ω_1 and (2.2) on $\partial\Omega_1 \setminus \Gamma_{12}$ and that $p|_{\Omega_1} = \lambda$ on Γ_{12}.

We now take $\mathbf{v} \in (C_0^\infty(\Omega_2))^d$, and (4.2) becomes

$$\int_{\Omega_2} \mathbf{K}^{-1}\mathbf{u} \cdot \mathbf{v} dx - \int_{\Omega_2} p\nabla \cdot \mathbf{v} dx = 0,$$

which implies that $\mathbf{u} = -\mathbf{K}\nabla p$ in Ω_2. With $q \in C_0^\infty(\Omega_2)$, (4.3) implies that $\nabla \cdot \mathbf{u} = f$ in Ω_2. Hence (2.1) is satisfied in Ω_2. We have forced that $\mathbf{u} \cdot \mathbf{n} = g$ on $\partial\Omega_2 \setminus \Gamma_{12}$, which results in the satisfaction of (2.2) on $\partial\Omega_2 \setminus \Gamma_{12}$.

By taking $\mathbf{v} \in \left(H^1(\Omega_2)\right)^d$, with $\mathbf{v} = 0$ on $\partial\Omega_2 \setminus \Gamma_{12}$, (4.2) becomes

$$\begin{aligned} -\int_{\Gamma_{12}} \mathbf{v} \cdot \mathbf{n}_{\partial\Omega_2} \lambda d\sigma &= -\int_{\Omega_2} \nabla p \cdot \mathbf{v} dx - \int_{\Omega_2} p\nabla \cdot \mathbf{v} dx \\ &= -\int_{\partial\Omega_2} \mathbf{v} \cdot \mathbf{n}_{\partial\Omega_2} p d\sigma = -\int_{\Gamma_{12}} \mathbf{v} \cdot \mathbf{n}_{\partial\Omega_2} p d\sigma. \end{aligned}$$

Since the trace of $\mathbf{v}$ can be arbitrarily chosen in $H_{00}^{\frac{1}{2}}(\Gamma_{12})^d$, we conclude that $p|_{\Omega_2} = \lambda$ on Γ_{12}. Hence p has the same trace on both sides of Γ_{12}, and therefore p belongs to $H^1(\Omega)$.

Finally, on Γ_{12}, (4.4) gives $-\mathbf{K}\nabla p|_{\Omega_1} \cdot \mathbf{n} = \mathbf{u} \cdot \mathbf{n} = -\mathbf{K}\nabla p|_{\Omega_2} \cdot \mathbf{n}$; that is, $\mathbf{K}\nabla p \cdot \mathbf{n}$ has the same trace on both sides of Γ_{12}. Therefore (2.1) is satisfied in the entire domain Ω. ∎

4.3. Discretization. We recall that the DG and MFE approximation spaces were introduced in section 2. Let $\Pi_h \in \mathcal{L}(H^1(\Omega_2)^d; \mathbf{V}_h(\Omega_2))$ be the standard MFE interpolation operator satisfying on any $E \in \mathcal{E}_h(\Omega_2)$ [17]

$$\int_E \nabla \cdot (\mathbf{u} - \Pi_h\mathbf{u})\, q_h \, dx = 0 \quad \forall\, q_h \in W_h(E), \tag{4.5}$$

$$\int_\gamma (\mathbf{u} - \Pi_h\mathbf{u}) \cdot \mathbf{n}\ \mathbf{v}_h \cdot \mathbf{n}\, d\sigma \quad \forall\, \mathbf{v}_h \in \mathbf{V}_h(E), \quad \forall \gamma \in \partial E, \tag{4.6}$$

$$\|\Pi_h\mathbf{u}\|_{H(\mathrm{div};E)} \le C\|\mathbf{u}\|_{1,E}, \tag{4.7}$$

$$\|\mathbf{u} - \Pi_h\mathbf{u}\|_{0,E} \le Ch^{m+1}|\mathbf{u}|_{m+1,E}. \tag{4.8}$$

These properties imply that for all MFE spaces under consideration [17]

$$\nabla \cdot \mathbf{V}_{h,0}(\Omega_2) = W_h(\Omega_2). \tag{4.9}$$

More precisely, we have the following inf-sup condition.

LEMMA 4.1. *Let $\mathcal{E}_h(\Omega_2)$ be nondegenerate. For any p_h in $W_h(\Omega_2)$, there exists $\mathbf{v}_h$ in $\mathbf{V}_{h,0}(\Omega_2)$ such that*

$$\nabla \cdot \mathbf{v}_h = p_h \quad \text{in } \Omega_2$$

and a constant C independent of $\mathbf{v}_h$, p_h, and h such that

$$\|\mathbf{v}_h\|_{H(\mathrm{div};\Omega_2)} + \|\mathbf{v}_h \cdot \mathbf{n}\|_{0,\Gamma_{12}} \le C\|p_h\|_{0,\Omega_2}. \tag{4.10}$$

Proof. First, we extend p_h by a constant function in Ω_1 so that its mean value is zero in Ω. Let $\tilde{p}_h$ denote the extended function. Then

$$\|\tilde{p}_h\|_{0,\Omega} \le C\|p_h\|_{0,\Omega_2}.$$

As $\tilde{p}_h$ has mean value zero in Ω, there exists $\mathbf{v}$ in $H_0^1(\Omega)^d$ such that (cf. [25])

$$\nabla \cdot \mathbf{v} = \tilde{p}_h \quad \text{in } \Omega$$

and

$$|\mathbf{v}|_{1,\Omega} \leq C\|\tilde{p}_h\|_{0,\Omega} \leq C\|p_h\|_{0,\Omega_2}. \tag{4.11}$$

Take $\mathbf{v}_h = \Pi_h \mathbf{v}$. Then we easily derive from (4.5)–(4.7) and the regularity of $\mathbf{v}$ that the restriction of $\mathbf{v}_h$ to Ω_2 belongs to $\mathbf{V}_{h,0}(\Omega_2)$ and satisfies (4.10). ∎

Let $\mathbf{u}_{h,g}$ be an adequate approximation of $\mathbf{u}_g$ in $\mathbf{V}_h(\Omega_2)$. The finite element mortar DG-MFE discretization is to find $(p_h|_{\Omega_1}, p_h|_{\Omega_2}, \mathbf{u}_h, \lambda_H)$ in $X_h(\Omega_1)\times W_h(\Omega_2)\times (\mathbf{V}_{h,0}(\Omega_2)+\mathbf{u}_{h,g})\times \Lambda_H$ such that the following equations hold for all $(q_h|_{\Omega_1}, q_h|_{\Omega_2}, \mathbf{v}_h, \mu_H)$ in $X_h(\Omega_1)\times W_h(\Omega_2)\times \mathbf{V}_{h,0}(\Omega_2)\times \Lambda_H$:

$$B(p_h, q_h) = L(q_h; \lambda_H), \tag{4.12}$$

$$\int_{\Omega_2} \mathbf{K}^{-1}\mathbf{u}_h \cdot \mathbf{v}_h dx = \int_{\Omega_2} p_h \nabla \cdot \mathbf{v}_h dx - \int_{\Gamma_{12}} \mathbf{v}_h \cdot \mathbf{n}_{\partial\Omega_2} \lambda_H d\sigma, \tag{4.13}$$

$$\int_{\Omega_2} \nabla \cdot \mathbf{u}_h q_h dx = \int_{\Omega_2} f q_h dx, \tag{4.14}$$

$$\int_{\Gamma_{12}} \mathbf{u}_h \cdot \mathbf{n} \mu_H d\sigma = -\int_{\Gamma_{12}} \mathbf{K}\nabla p_h|_{\Omega_1} \cdot \mathbf{n}\mu_H d\sigma + \sum_{\tau\in\Gamma_H} \frac{\sigma_\tau}{H_\tau} \int_\tau (p_h|_{\Omega_1} - \lambda_H)\mu_H d\sigma. \tag{4.15}$$

Note that both $\mathbf{u}_g$ and $\mathbf{u}_{h,g}$ are introduced only for theoretical reasons and in practice we need only to approximate g.

We now address the existence and uniqueness of the solution of the above system. This is a square finite-dimensional system, and existence is equivalent to uniqueness. Let $f = 0$ and $\mathbf{u}_{h,g} = \mathbf{0}$. By taking $q_h = p_h$ in (4.12), we have

$$\begin{aligned} &\sum_{E\in\mathcal{E}_h(\Omega_1)} \int_E \mathbf{K}\nabla p_h \cdot \nabla p_h dx - (1+s_{\text{form}}) \sum_{\gamma\in\Gamma_h(\Omega_1)} \int_\gamma \{\mathbf{K}\nabla p_h \cdot \mathbf{n}\}\,[p_h]\,d\sigma \\ &+ \sum_{\gamma\in\Gamma_h(\Omega_1)} \frac{\sigma_\gamma}{h_\gamma} \int_\gamma [p_h]^2\, d\sigma + \sum_{\tau\in\Gamma_H} \frac{\sigma_\tau}{H_\tau} \int_\tau p_h|_{\Omega_1}^2 d\sigma \\ &= \int_{\Gamma_{12}} \mathbf{K}\nabla p_h|_{\Omega_1} \cdot \mathbf{n}\lambda_H d\sigma + \sum_{\tau\in\Gamma_H} \frac{\sigma_\tau}{H_\tau} \int_\tau p_h|_{\Omega_1}\lambda_H d\sigma. \end{aligned} \tag{4.16}$$

By taking $\mathbf{v}_h = \mathbf{u}_h$ in (4.13) and $q_h = p_h$ in (4.14), we have

$$\int_{\Omega_2} \mathbf{K}^{-1}\mathbf{u}_h \cdot \mathbf{u}_h dx = -\int_{\Gamma_{12}} \mathbf{u}_h \cdot \mathbf{n}_{\partial\Omega_2} \lambda_H d\sigma = \int_{\Gamma_{12}} \mathbf{u}_h \cdot \mathbf{n}\lambda_H d\sigma. \tag{4.17}$$

By taking $\mu_H = \lambda_H$ in (4.15), we obtain

$$\int_{\Gamma_{12}} \mathbf{u}_h \cdot \mathbf{n}\lambda_H d\sigma = -\int_{\Gamma_{12}} \mathbf{K}\nabla p_h|_{\Omega_1} \cdot \mathbf{n}\lambda_H d\sigma + \sum_{\tau\in\Gamma_H} \frac{\sigma_\tau}{H_\tau} \int_\tau (p_h|_{\Omega_1} - \lambda_H)\lambda_H d\sigma. \tag{4.18}$$

Summation of (4.16), (4.17), and (4.18) leads to

$$\begin{aligned} &|\!|\!|\mathbf{K}^{\frac{1}{2}}\nabla p_h|\!|\!|_{0,\Omega_1}^2 - (1+s_{\text{form}}) \sum_{\gamma\in\Gamma_h(\Omega_1)} \int_\gamma \{\mathbf{K}\nabla p_h \cdot \mathbf{n}\}\,[p_h]\,d\sigma + \|\mathbf{K}^{-\frac{1}{2}}\mathbf{u}_h\|_{0,\Omega_2}^2 \\ &+ \sum_{\gamma\in\Gamma_h(\Omega_1)} \frac{\sigma_\gamma}{h_\gamma} \int_\gamma [p_h]^2\, d\sigma + \sum_{\tau\in\Gamma_H} \frac{\sigma_\tau}{H_\tau} \int_\tau (p_h|_{\Omega_1} - \lambda_H)^2 d\sigma = 0. \end{aligned} \tag{4.19}$$

First consider the case $0 < \sigma_\tau^0 \leq \sigma_\tau \leq \sigma_\tau^1$. As in section 3.3, we easily derive from (4.19) that for OBB-DG and NIPG, $\mathbf{u}_h = \mathbf{0}$ in Ω_2, p_h is a constant in Ω_1 and λ_H is the same constant on Γ_{12}. The same conclusion holds for SIPG and IIPG, by applying inequality (3.10).

Now as $\mathbf{u}_h$ is zero, (4.13) becomes

$$\forall \mathbf{v}_h \in \mathbf{V}_{h,0}(\Omega_2)\,, \int_{\Omega_2} p_h \nabla \cdot \mathbf{v}_h dx = \int_{\Gamma_{12}} \mathbf{v}_h \cdot \mathbf{n}_{\partial\Omega_2} \lambda_H d\sigma. \tag{4.20}$$

Let $\overline{p_h}$ denote the mean value of p_h in Ω_2: $\overline{p_h} = \frac{1}{|\Omega_2|}\int_{\Omega_2} p_h dx$. Then (4.20) implies that for all $\mathbf{v}_h$ in $\mathbf{V}_{h,0}(\Omega_2)$, with $\mathbf{v}_h \cdot \mathbf{n} = 0$ on $\partial\Omega_2$,

$$\int_{\Omega_2} (p_h - \overline{p_h}) \nabla \cdot \mathbf{v}_h dx = 0.$$

As $p_h - \overline{p_h}$ belongs to $W_h(\Omega_2)$ and has mean value zero, the argument of Lemma 4.1 implies that there exists $\mathbf{v}_h$ in $\mathbf{V}_h(\Omega_2)$, with $\mathbf{v}_h \cdot \mathbf{n} = 0$ on $\partial\Omega_2$, such that

$$\int_{\Omega_2} (p_h - \overline{p_h}) \nabla \cdot \mathbf{v}_h dx = \|p_h - \overline{p_h}\|_{0,\Omega_2}^2.$$

Therefore p_h is also constant in Ω_2, and (4.20) implies that

$$\forall \mathbf{v}_h \in \mathbf{V}_{h,0}(\Omega_2)\,, \int_{\Gamma_{12}} \mathbf{v}_h \cdot \mathbf{n} (\lambda_H - p_h|_{\Omega_2}) d\sigma = 0. \tag{4.21}$$

Since λ_H is constant and coincides with the trace of p_h coming from Ω_1, we infer from (4.21) that it also coincides with the trace of p_h coming from Ω_2. Thus p_h is constant in Ω.

There remains the case $\sigma_\tau = 0$. With the information we have so far, (4.12) implies that

$$\int_{\Gamma_{12}} \mathbf{K} \nabla q_h|_{\Omega_1} \cdot \mathbf{n} (\lambda_H - p_h|_{\Omega_1}) d\sigma = 0. \tag{4.22}$$

If the mortar compatibility condition (A.2) is satisfied, since p_h is constant in Ω_2, (4.21) implies that $\lambda_H = p_h|_{\Omega_2}$. Thus, λ_H is constant, and (4.22) implies readily that p_h has the same trace on both sides of Γ_{12}. If the mortar compatibility condition (A.1) holds, as p_h is constant in Ω_1, then (4.22) implies that $\lambda_H = p_h|_{\Omega_1}$. Hence λ_H is constant, and (4.21) implies again that $\lambda_H = p_h|_{\Omega_2}$; therefore p_h is constant in Ω.

We have shown for all schemes that $\mathbf{u}_h$ is unique and that the null space of the linear system (4.12)–(4.15) for p_h and λ_H is the constant vector. The compatibility condition (2.3) implies that the right-hand side is orthogonal to the null space, and therefore the solution exists and is unique up to an additive constant for p_h and λ_H. We have proved the following solvability theorem.

THEOREM 4.2. *For OBB-DG, we assume that $r \geq 2$. For SIPG and IIPG, we assume that σ_γ^0 is sufficiently large. No assumption is needed for NIPG. Then the scheme* (4.12)–(4.15) *possesses a solution $(p_h, \mathbf{u}_h, \lambda_H)$ unique up to an additive constant that is the same for p_h and λ_H. The same conclusion holds if $\sigma_\tau = 0$, assuming that either the compatibility condition* (A.1) *for $i = 1$ or* (A.2) *holds.*

4.4. Convergence. We define the interpolant $\hat{p}$ of p such that $\hat{p}$ in Ω_2 is the L^2-projection and $\hat{p}$ in Ω_1 is defined as in the previous section. Then, on any $E \in \mathcal{E}_h(\Omega_2)$, $\hat{p}$ satisfies

$$\|p - \hat{p}\|_{0,E} \leq C h_E^{l+1} |p|_{l+1,E}. \tag{4.23}$$

For $\bar{p}$, we take again the continuous nodal interpolant of p in Λ_H. Now we choose $\mathbf{u}_{h,g}$. On any $\gamma \subset \partial E \cap (\partial\Omega_2 \setminus \Gamma_{12})$, considering that $\mathbf{u} \cdot \mathbf{n} = g$, we define $P_h g$ by

$$P_h g = (\Pi_h \mathbf{u} \cdot \mathbf{n})|_\gamma.$$

Since, by construction, $(\Pi_h \mathbf{u} \cdot \mathbf{n})|_\gamma$ does not depend on the interior values of $\mathbf{u}$, $P_h g$ depends only on g. Then we can take for $\mathbf{u}_{h,g}$ any function in $\mathbf{V}_h(\Omega_2)$ such that $\mathbf{u}_{h,g} = P_h g$ on $\partial\Omega_2 \setminus \Gamma_{12}$. All results derived below are independent of this choice and depend only on $P_h g$. To prove the convergence theorem below, we need the following trace inequality. The proof is a variant of those derived by Brenner in [13]; see also [41]. It is stated in Ω_1, but it is valid in any connected Lipschitz domain.

THEOREM 4.3. *Let $\mathcal{E}_h(\Omega_1)$ be nondegenerate, and let Γ be any portion of $\partial\Omega_1$ with positive measure. Assume that $0 < \sigma_\gamma^0 \leq \sigma_\gamma \leq \sigma_\gamma^1$. Then there exists a constant C, independent of h such that, for all functions q_h in $X_h(\Omega_1)$, the following trace inequality holds:*

$$\int_{\Gamma_{12}} q_h^2 d\sigma \leq C \left(\|\mathbf{K}^{\frac{1}{2}} \nabla q_h\|_{0,\Omega_1}^2 + \sum_{\gamma \in \Gamma_h(\Omega_1)} \frac{\sigma_\gamma}{h_\gamma} \int_\gamma [q_h]^2 \, d\sigma + \left| \int_\Gamma q_h \, d\sigma \right|^2 \right). \tag{4.24}$$

We proceed with the convergence analysis of the coupled DG-MFE methods. We consider only the cases of NIPG, SIPG, or IIPG, since Theorem 4.3 does not apply in the case of OBB-DG. We comment on that case in Remark 4.1.

THEOREM 4.4. *Let $\mathcal{E}_h(\Omega_i)$ $i = 1, 2$, be nondegenerate. Let $p \in H^s(\Omega)$, $s \geq 2$, be a solution of (2.1)–(2.2), and let $\mathbf{u} = -\mathbf{K}\nabla p$. Let $(p_h, \mathbf{u}_h, \lambda_H)$ be a solution of (4.12)–(4.15), where NIPG, SIPG, or IIPG is used in (4.12). We assume that σ_γ^0 is sufficiently large for SIPG and IIPG. Then there exists a constant C, independent of h and H, such that*

$$\|\mathbf{K}^{\frac{1}{2}} \nabla (p_h - p)\|_{0,\Omega_1} + \sqrt{\sum_{\gamma \in \Gamma_h(\Omega_1)} \frac{\sigma_\gamma}{h_\gamma} \|[p_h]\|_{0,\gamma}^2} + \left\| \mathbf{K}^{-\frac{1}{2}} (\mathbf{u}_h - \mathbf{u}) \right\|_{0,\Omega_2}$$
$$+ \sqrt{\sum_{\tau \in \Gamma_H} \frac{\sigma_\tau}{H_\tau} \left\| p_h|_{\Omega_1} - \lambda_H \right\|_{0,\tau}^2} \leq C \left(h^{\mu-1} \left(\frac{H}{h} \right)^{\frac{1}{2}} + H^{\bar{\mu} - \frac{1}{2}} + h^\nu \right),$$

where $\mu = \min(r+1, s)$, $\bar{\mu} = \min\left(\bar{r}+1, s - \frac{1}{2}\right)$, and $\nu = \min\left(m+1, s - \frac{3}{2}\right)$.

Proof. In addition to the error variables η, η^I, η^A, ξ, ξ^I, and ξ^A introduced in section 3.4, we define

$$\chi := \mathbf{u}_h - \mathbf{u}, \quad \chi^I := \mathbf{u} - \Pi_h \mathbf{u}, \quad \chi^A := \mathbf{u}_h - \Pi_h \mathbf{u} = \chi + \chi^I,$$

and observe that, owing to the choice of $\mathbf{u}_{h,g}$, χ^A belongs to $\mathbf{V}_{h,0}(\Omega_2)$. By subtracting the weak formulation (4.1) from the finite element scheme (4.12), and choosing $q_h = \xi^A$, we obtain

$$\begin{aligned} B(\xi^A, \xi^A) &= B(\xi^I, \xi^A) + L(\xi^A; \lambda_H) - L(\xi^A; p) \\ &= B(\xi^I, \xi^A) + \int_{\Gamma_{12}} \mathbf{K}\nabla \xi^A|_{\Omega_1} \cdot \mathbf{n} \eta \, d\sigma + \sum_{\tau \in \Gamma_H} \frac{\sigma_\tau}{H_\tau} \int_\tau \xi^A|_{\Omega_1} \eta \, d\sigma. \end{aligned} \tag{4.25}$$

By subtracting from the mixed finite element scheme (4.13)–(4.14) the corresponding weak formulation (4.2)–(4.3), we obtain the error equations for all $(\mathbf{v}_h, q_h)$ in $\mathbf{V}_{h,0}(\Omega_2) \times W_h(\Omega_2)$:

$$\int_{\Omega_2} \mathbf{K}^{-1}\chi \cdot \mathbf{v}_h \, dx = \int_{\Omega_2} \xi^A \, \nabla \cdot \mathbf{v}_h \, dx - \int_{\Gamma_{12}} \mathbf{v}_h \cdot \mathbf{n}_{\partial\Omega_2} \eta d\sigma, \tag{4.26}$$

$$\int_{\Omega_2} \nabla \cdot \chi^A \, q_h \, dx = 0, \tag{4.27}$$

where we have used the properties (4.9) and (4.5). Note that (4.27) and (4.9) imply that

$$\nabla \cdot \chi^A = 0. \tag{4.28}$$

By taking $\mathbf{v}_h = \chi^A$ in the equations above and noting that $\mathbf{n} = -\mathbf{n}_{\partial\Omega_2}$, we have

$$\int_{\Omega_2} \mathbf{K}^{-1}\chi \cdot \chi^A dx - \int_{\Gamma_{12}} \chi^A \cdot \mathbf{n} \, \eta \, d\sigma = 0. \tag{4.29}$$

Similarly, by subtracting the matching condition (4.4) from its finite element formulation (4.15), and taking $\mu_H = \eta^A$, we obtain

$$\int_{\Gamma_{12}} \chi \cdot \mathbf{n} \eta^A d\sigma = -\int_{\Gamma_{12}} \mathbf{K}\nabla\xi|_{\Omega_1} \cdot \mathbf{n}\eta^A d\sigma + \sum_{\tau \in \Gamma_H} \frac{\sigma_\tau}{H_\tau} \int_\tau (\xi|_{\Omega_1} - \eta)\eta^A d\sigma. \tag{4.30}$$

Summation of (4.25), (4.29), and (4.30) yields

$$\begin{aligned} B(\xi^A, \xi^A) &+ \int_{\Omega_2} \mathbf{K}^{-1}\chi \cdot \chi^A dx - \int_{\Gamma_{12}} \chi^A \cdot \mathbf{n}\eta d\sigma + \int_{\Gamma_{12}} \chi \cdot \mathbf{n}\eta^A d\sigma \\ &= B(\xi^I, \xi^A) + \int_{\Gamma_{12}} \mathbf{K}\nabla\xi^A|_{\Omega_1} \cdot \mathbf{n}\eta d\sigma + \sum_{\tau \in \Gamma_H} \frac{\sigma_\tau}{H_\tau} \int_\tau \xi^A\big|_{\Omega_1} \eta d\sigma \\ &\quad - \int_{\Gamma_{12}} \mathbf{K}\nabla\xi|_{\Omega_1} \cdot \mathbf{n}\eta^A d\sigma + \sum_{\tau \in \Gamma_H} \frac{\sigma_\tau}{H_\tau} \int_\tau (\xi|_{\Omega_1} - \eta)\eta^A d\sigma. \end{aligned}$$

Rearranging terms results in

$$\begin{aligned} &B(\xi^A, \xi^A) + \|\mathbf{K}^{-\frac{1}{2}}\chi^A\|^2_{0,\Omega_2} + \sum_{\tau \subset \Gamma_H} \frac{\sigma_\tau}{H_\tau} \int_\tau (\xi^A\big|_{\Omega_1} - \eta^A)^2 d\sigma - \sum_{\tau \in \Gamma_H} \frac{\sigma_\tau}{H_\tau} \int_\tau (\xi^A\big|_{\Omega_1})^2 d\sigma \\ &= B(\xi^I, \xi^A) - \int_{\Gamma_{12}} \mathbf{K}\nabla\xi^A|_{\Omega_1} \cdot \mathbf{n}\eta^I d\sigma - \sum_{\tau \in \Gamma_H} \frac{\sigma_\tau}{H_\tau} \int_\tau \xi^A\big|_{\Omega_1} \eta^I d\sigma \\ &\quad - \sum_{\tau \in \Gamma_H} \frac{\sigma_\tau}{H_\tau} \int_\tau (\xi^I\big|_{\Omega_1} - \eta^I)\eta^A d\sigma + \int_{\Gamma_{12}} \mathbf{K}\nabla\xi^I\big|_{\Omega_1} \cdot \mathbf{n}\eta^A d\sigma \\ &\quad + \int_{\Omega_2} \mathbf{K}^{-1}\chi^I \cdot \chi^A dx - \int_{\Gamma_{12}} \chi^A \cdot \mathbf{n}\eta^I d\sigma + \int_{\Gamma_{12}} \chi^I \cdot \mathbf{n}\eta^A d\sigma. \end{aligned} \tag{4.31}$$

We denote by L_{ErrEqu} the left-hand side of (4.31) and apply an algebraic manipulation to obtain

$$\begin{aligned} L_{\mathrm{ErrEqu}} &= \|\mathbf{K}^{\frac{1}{2}}\nabla\xi^A\|^2_{0,\Omega_1} - (1 + s_{\mathrm{form}}) \sum_{\gamma \in \Gamma_h(\Omega_1)} \int_\gamma \left\{\mathbf{K}\nabla\xi^A \cdot \mathbf{n}\right\} \left[\xi^A\right] d\sigma \\ &\quad + \sum_{\gamma \in \Gamma_h(\Omega_1)} \frac{\sigma_\gamma}{h_\gamma} \int_\gamma \left[\xi^A\right]^2 d\sigma + \|\mathbf{K}^{-\frac{1}{2}}\chi^A\|^2_{0,\Omega_2} + \sum_{\tau \in \Gamma_H} \frac{\sigma_\tau}{H_\tau} \int_\tau (\xi^A\big|_{\Omega_1} - \eta^A)^2 d\sigma. \end{aligned}$$

For NIPG and OBB-DG, the second term in L_{ErrEqu} vanishes, leaving only the coercive terms. For SIPG and IIPG, we employ the inequality (3.10) with $q_h = q = \xi^A$ to conclude that

$$L_{\text{ErrEqu}} \geq \frac{3}{4}\|\mathbf{K}^{\frac{1}{2}}\nabla\xi^A\|^2_{0,\Omega_1} + \frac{3}{4}\sum_{\gamma\in\Gamma_h(\Omega_1)}\frac{\sigma_\gamma}{h_\gamma}\int_\gamma \left[\xi^A\right]^2 d\sigma + \left\|\mathbf{K}^{-\frac{1}{2}}\chi^A\right\|^2_{0,\Omega_2} + \sum_{\tau\in\Gamma_H}\frac{\sigma_\tau}{H_\tau}\int_\tau \left(\xi^A|_{\Omega_1} - \eta^A\right)^2 d\sigma.$$

We now consider the right-hand side of (4.31), which is denoted by R_{ErrEqu}. By expanding the first term as

$$\begin{aligned} B(\xi^I,\xi^A) &= \sum_{E\in\mathcal{E}_h(\Omega_1)}\int_E \mathbf{K}\nabla\xi^I\cdot\nabla\xi^A dx - \sum_{\gamma\in\Gamma_h(\Omega_1)}\int_\gamma \left\{\mathbf{K}\nabla\xi^I\cdot\mathbf{n}\right\}\left[\xi^A\right] d\sigma \\ &- s_{\text{form}}\sum_{\gamma\in\Gamma_h(\Omega_1)}\int_\gamma \left\{\mathbf{K}\nabla\xi^A\cdot\mathbf{n}\right\}\left[\xi^I\right] d\sigma - \int_{\Gamma_{12}}\mathbf{K}\nabla\xi^I|_{\Omega_1}\cdot\mathbf{n}\xi^A|_{\Omega_1} d\sigma \\ &+ \int_{\Gamma_{12}}\mathbf{K}\nabla\xi^A|_{\Omega_1}\cdot\mathbf{n}\xi^I|_{\Omega_1} d\sigma + \sum_{\gamma\in\Gamma_h(\Omega_1)}\frac{\sigma_\gamma}{h_\gamma}\int_\gamma\left[\xi^I\right]\left[\xi^A\right] d\sigma + \sum_{\tau\in\Gamma_H}\frac{\sigma_\tau}{H_\tau}\int_\tau \xi^I|_{\Omega_1}\xi^A|_{\Omega_1} d\sigma, \end{aligned}$$

we have

$$\begin{aligned} R_{\text{ErrEqu}} &= \sum_{E\in\mathcal{E}_h(\Omega_1)}\int_E \mathbf{K}\nabla\xi^I\cdot\nabla\xi^A dx - \sum_{\gamma\in\Gamma_h(\Omega_1)}\int_\gamma \left\{\mathbf{K}\nabla\xi^I\cdot\mathbf{n}\right\}\left[\xi^A\right] d\sigma \\ &- s_{\text{form}}\sum_{\gamma\in\Gamma_h(\Omega_1)}\int_\gamma \left\{\mathbf{K}\nabla\xi^A\cdot\mathbf{n}\right\}\left[\xi^I\right] d\sigma + \sum_{\gamma\in\Gamma_h(\Omega_1)}\frac{\sigma_\gamma}{h_\gamma}\int_\gamma\left[\xi^I\right]\left[\xi^A\right] d\sigma \\ &+ \sum_{\tau\in\Gamma_H}\frac{\sigma_\tau}{H_\tau}\int_\tau (\xi^I|_{\Omega_1} - \eta^I)(\xi^A|_{\Omega_1} - \eta^A) d\sigma \\ &+ \int_{\Omega_2}\mathbf{K}^{-1}\chi^I\cdot\chi^A dx - \int_{\Gamma_{12}}\chi^A\cdot\mathbf{n}\eta^I d\sigma + \int_{\Gamma_{12}}\chi^I\cdot\mathbf{n}\eta^A d\sigma \\ &- \int_{\Gamma_{12}}\mathbf{K}\nabla\xi^A|_{\Omega_1}\cdot\mathbf{n}\left(\eta^I - \xi^I|_{\Omega_1}\right) d\sigma + \int_{\Gamma_{12}}\mathbf{K}\nabla\xi^I|_{\Omega_1}\cdot\mathbf{n}\left(\eta^A - \xi^A|_{\Omega_1}\right) d\sigma =: \sum_{i=1}^{10} T_i. \end{aligned}$$

We now bound each term in R_{ErrEqu}. We skip terms T_1–T_5 because they have the same bounds as in the proof of Theorem 3.3. Next, (4.8) implies that

$$|T_6| \leq \epsilon\left\|\mathbf{K}^{-\frac{1}{2}}\chi^A\right\|^2_{0,\Omega_2} + C\left\|\mathbf{K}^{-\frac{1}{2}}\chi^I\right\|^2_{0,\Omega_2} \leq \epsilon\left\|\mathbf{K}^{-\frac{1}{2}}\chi^A\right\|^2_{0,\Omega_2} + Ch^{2\nu}.$$

To bound term T_7, let $\tilde{\eta}^I = P_2(\eta^I)$, where $P_2 \in \mathcal{L}(H^{\frac{1}{2}}(\Gamma_{12}); H^1(\Omega_2))$ is the analogue of P_1. By considering that $\chi^A\cdot\mathbf{n}$ vanishes on $\partial\Omega_2\setminus\Gamma_{12}$, we can write

$$|T_7| \leq \left|\int_{\partial\Omega_2}(\chi^A\cdot\mathbf{n})\tilde{\eta}^I d\sigma\right| \leq \|\chi^A\cdot\mathbf{n}\|_{-\frac{1}{2},\partial\Omega_2}\|\tilde{\eta}^I\|_{\frac{1}{2},\partial\Omega_2} \leq C\left\|\chi^A\right\|_{H(\text{div};\Omega_2)}\left\|\eta^I\right\|_{\frac{1}{2},\Gamma_{12}}.$$

Then, by using (4.28) and (3.26), we obtain

$$|T_7| \leq \epsilon\left\|\mathbf{K}^{-\frac{1}{2}}\chi^A\right\|^2_{0,\Omega_2} + CH^{2\bar{\mu}-1}.$$

To estimate T_8, we split it into

$$\int_{\Gamma_{12}} (\chi^I \cdot \mathbf{n})\eta^A d\sigma = \int_{\Gamma_{12}} (\chi^I \cdot \mathbf{n})(\eta^A - \xi^A) d\sigma + \int_{\Gamma_{12}} (\chi^I \cdot \mathbf{n})\xi^A d\sigma,$$

and we consider first the last term. Let $\overline{\xi^A} = \frac{1}{|\Gamma_{12}|}\int_{\Gamma_{12}} \xi^A d\sigma$. The approximation property (4.6) of Π_h implies that

$$\int_{\Gamma_{12}} (\chi^I \cdot \mathbf{n})\xi^A d\sigma = \int_{\Gamma_{12}} (\chi^I \cdot \mathbf{n})(\xi^A - \overline{\xi^A}) d\sigma.$$

As $\xi^A - \overline{\xi^A}$ has mean value zero on Γ_{12} and $\overline{\xi^A}$ is a constant, the trace Theorem 4.3 yields:

$$\left| \int_{\Gamma_{12}} (\chi^I \cdot \mathbf{n})\xi^A d\sigma \right| \leq C \|\chi^I \cdot \mathbf{n}\|_{0,\Gamma_{12}} \left(|\!|\!| \mathbf{K}^{\frac{1}{2}} \nabla \xi^A |\!|\!|^2_{0,\Omega_1} + \sum_{\gamma \in \Gamma_h(\Omega_1)} \frac{\sigma_\gamma}{h_\gamma} \|[\xi^A]\|^2_{0,\gamma} \right)^{\frac{1}{2}}.$$

As far as the first term is concerned, we write

$$\left| \int_{\Gamma_{12}} (\chi^I \cdot \mathbf{n})(\eta^A - \xi^A) d\sigma \right| \leq \left(\sum_{\tau \in \Gamma_H} \frac{\sigma_\tau}{H_\tau} \|\eta^A - \xi^A\|^2_{0,\tau} \right)^{\frac{1}{2}} \left(\sum_{\tau \in \Gamma_H} \frac{H_\tau}{\sigma_\tau} \|\chi^I \cdot \mathbf{n}\|^2_{0,\tau} \right)^{\frac{1}{2}}.$$

By collecting these inequalities and using the approximation properties of Π_h, we derive:

$$|T_8| \leq \epsilon \sum_{\gamma \in \Gamma_h(\Omega_1)} \frac{\sigma_\gamma}{h_\gamma} \int_\gamma \left[\xi^A\right]^2 d\sigma + \epsilon |\!|\!| \mathbf{K}^{\frac{1}{2}} \nabla \xi^A |\!|\!|^2_{\Omega_1} + \epsilon \sum_{\tau \in \Gamma_H} \frac{\sigma_\tau}{H_\tau} \int_\tau \left(\xi^A\big|_{\Omega_1} - \eta^A \right)^2 d\sigma + C h^{2\nu}.$$

Term T_9 can be bounded by using the argument for $T_{6,1}$ in Theorem 3.3:

$$|T_9| \leq \epsilon |\!|\!| \mathbf{K}^{\frac{1}{2}} \nabla \xi^A |\!|\!|^2_{0,\Omega_1} + \epsilon \sum_{\gamma \in \Gamma_h(\Omega_1)} \frac{\sigma_\gamma}{h_\gamma} \|[p_h]\|^2_{0,\gamma} + \epsilon \sum_{\tau \in \Gamma_H} \frac{\sigma_\tau}{H_\tau} \left\| \xi^A\big|_{\Omega_1} - \eta^A \right\|^2_{0,\tau} + C(h^{2\mu-2} + H^{2\bar{\mu}-1}).$$

The estimate of the last term is the same as for T_7 in Theorem 3.3:

$$|T_{10}| \leq \epsilon \sum_{\tau \in \Gamma_H} \frac{\sigma_\tau}{H_\tau} \left\| \eta^A - \xi^A|_{\Omega_1} \right\|^2_{0,\tau} + C h^{2\mu-3} H.$$

Finally, we combine all terms to conclude that

$$|\!|\!| \mathbf{K}^{\frac{1}{2}} \nabla \xi^A |\!|\!|^2_{0,\Omega_1} + \sum_{\gamma \in \Gamma_h(\Omega_1)} \frac{\sigma_\gamma}{h_\gamma} \left\| \left[\xi^A\right] \right\|^2_{0,\gamma} + \left\| \mathbf{K}^{-\frac{1}{2}} \chi^A \right\|^2_{0,\Omega_2} + \sum_{\tau \in \Gamma_H} \frac{\sigma_\tau}{H_\tau} \left\| \xi^A\big|_{\Omega_1} - \eta^A \right\|^2_{0,\tau} \leq C \left(h^{2\mu-2} \left(\frac{H}{h} \right) + H^{2\bar{\mu}-1} + h^{2\nu} \right).$$

An application of the triangle inequality completes the proof. ∎

Remark 4.1. The difficulty in coupling OBB-DG with MFE in the above proof lies in the estimate of term T_8. The term is bounded by using the special trace inequality

from Theorem 4.3, which involves penalized jump terms on interior edges (faces). An alternative approach to handle T_8 is to construct a special MFE interpolant $\tilde{\Pi}\mathbf{u}$ satisfying

$$\forall \mu_H \in \Lambda_H\,, \int_{\Gamma_{12}} (\mathbf{u} - \tilde{\Pi}\mathbf{u}) \cdot \mathbf{n} \mu_H \, d\sigma = 0.$$

This can be done by assuming a specific relation between the mortar grid and the MFE grid. Due to lack of space, we do not pursue this approach further here.

Remark 4.2. The cases $\bar{s}_{\text{form}} = 0$ or 1 can be treated as in Remark 3.1. It can be shown for the modified DG-MFE methods that, under the assumptions of Theorem 4.4, there exists a constant C, independent of h and H, such that

$$\begin{aligned}
&\left|\!\left|\!\left| \mathbf{K}^{\frac{1}{2}} \nabla (p_h - p) \right|\!\right|\!\right|_{\Omega_1} + \sqrt{\sum_{\gamma \in \Gamma_h(\Omega_1)} \frac{\sigma_\gamma}{h_\gamma} \left\| [p_h] \right\|_{0,\gamma}^2} + \left\| \mathbf{K}^{-\frac{1}{2}} (\mathbf{u}_h - \mathbf{u}) \right\|_{0,\Omega_2} \\
&\quad + \sqrt{\sum_{\gamma \in \Gamma_h(\Omega_1)} \frac{\sigma_\gamma}{h_\gamma} \left\| p_h|_{\Omega_1} - \lambda_H \right\|_{0,\gamma}^2} \le C \left(h^{\mu-1} + H^{\bar{\mu}-\frac{1}{2}} \left(\frac{H}{h} \right)^{\frac{1}{2}} + h^{\nu} \right), \\
&\qquad \|p_h - p\|_{0,\Omega_2} \le C \left(h^{\mu-1} + H^{\bar{\mu}-\frac{1}{2}} \left(\frac{H}{h} \right)^{\frac{1}{2}} + h^{\nu} + h^{\bar{\nu}} \right),
\end{aligned}$$

where $\mu = \min(r+1, s)$, $\bar{\mu} = \min\left(\bar{r}+1, s-\frac{1}{2}\right)$, $\nu = \min\left(m+1, s-\frac{3}{2}\right)$, and $\bar{\nu} = \min(l+1, s)$.

4.4.1. Convergence for the MFE pressure. To estimate the error on the pressure computed by the MFE method in Ω_2, we start again with the error equation (4.26). By virtue of the inf-sup condition in Lemma 4.1, we can choose in (4.26) $\mathbf{v}_h \in \mathbf{V}_{h,0}(\Omega_2)$ such that

$$\int_{\Omega_2} (p_h - \hat{p}) \nabla \cdot \mathbf{v}_h \, dx = \|p_h - \hat{p}\|_{0,\Omega_2}^2,$$

$$\|\mathbf{v}_h\|_{H(\text{div};\Omega_2)} + \|\mathbf{v}_h \cdot \mathbf{n}\|_{0,\Gamma_{12}} \le C \|p_h - \hat{p}\|_{0,\Omega_2}.$$

Therefore,

$$\|p_h - \hat{p}\|_{0,\Omega_2} \le C \left(\|\mathbf{K}^{-\frac{1}{2}} (\mathbf{u}_h - \mathbf{u})\|_{0,\Omega_2} + \|\lambda_H - p\|_{0,\Gamma_{12}} \right).$$

The first term on the right is bounded in Theorem 4.4. For the second term on the right, we have

$$\|\lambda_H - p\|_{0,\Gamma_{12}} \le \|\lambda_H - p_h|_{\Omega_1}\|_{0,\Gamma_{12}} + \|p_h|_{\Omega_1} - p\|_{0,\Gamma_{12}}.$$

The first term on right above is bounded in Theorem 4.4. Finally, by choosing the undetermined coefficient in p_h such that $\int_{\Gamma_{12}} (p - p_h|_{\Omega_1}) \, d\sigma = 0$, and applying Theorem 4.3, the last term above can be controlled by terms bounded in Theorem 4.4.

Thus, by using a triangle inequality and (4.23), we derive the following theorem.

THEOREM 4.5. *Let the assumptions of Theorem* 4.4 *hold, and let* $\int_{\Gamma_{12}} (p - p_h) \, d\sigma = 0$. *Then there exists a constant* C*, independent of* h *and* H*, such that*

$$\|p_h - p\|_{0,\Omega_2} \le C \left(h^{\mu-1} \left(\frac{H}{h} \right)^{\frac{1}{2}} + H^{\bar{\mu}-\frac{1}{2}} + h^{\nu} + h^{\bar{\nu}} \right),$$

where $\mu = \min(r+1, s)$, $\bar{\mu} = \min\left(\bar{r}+1, s-\frac{1}{2}\right)$, $\nu = \min\left(m+1, s-\frac{3}{2}\right)$, *and* $\bar{\nu} = \min(l+1, s)$.

Remark 4.3. When Ω is decomposed into several subdomains, then the last step in the proof of Theorem 4.5 involves the estimate of $p - p_h$ over the union of several interfaces, say, $\Gamma = \cup_{1\le i\le \ell}\Gamma_i$. The statement of Theorem 4.3 can be extended to this case. Therefore, it suffices to choose the undetermined constant in p_h such that $\int_\Gamma (p_h - p)d\sigma = 0$.

4.5. The interface operator and the reduced problem. Recall the definition of $l_i(\cdot)$ and $b_i(\cdot,\cdot)$ in (3.33) and (3.34), respectively. We now define a bilinear form $d_H : L^2(\Gamma_{12}) \times L^2(\Gamma_{12}) \mapsto \mathbb{R}$ for $\lambda, \mu \in L^2(\Gamma_{12})$ by $d_H(\lambda,\mu) := \sum_{i=1}^2 d_{H,i}(\lambda,\mu)$:

$$d_{H,1}(\lambda,\mu) := \int_{\Gamma_{12}} \mathbf{K}\,\nabla p_h^*(\lambda)|_{\Omega_1}\cdot \mathbf{n}_{\partial\Omega_1}\mu d\sigma - \sum_{\tau\in\Gamma_H}\frac{\sigma_\tau}{H_\tau}\int_\tau (p_h^*(\lambda)|_{\Omega_1} - \lambda)\mu d\sigma,$$

$$d_{H,2}(\lambda,\mu) := -\int_{\Gamma_{12}} \mathbf{u}_h^*(\lambda)|_{\Omega_2}\cdot \mathbf{n}_{\partial\Omega_2}\mu d\sigma,$$

where $p_h^*(\lambda) \in X_h(\Omega_1)$ is the solution of

$$B_1\left(p_h^*(\lambda), q_h\right) = b_1\left(\lambda, q_h\right) \quad \forall q_h \in X_h(\Omega_1) \tag{4.32}$$

and $\mathbf{u}_h^*(\lambda)$ is the first component of the solution $(\mathbf{u}_h^*(\lambda), p_h^*(\lambda)) \in \mathbf{V}_{h,0}(\Omega_2)\times W_h(\Omega_2)$ of

(4.33)

$$\int_{\Omega_2}\mathbf{K}^{-1}\mathbf{u}_h^*(\lambda)\cdot\mathbf{v}_h dx = \int_{\Omega_2} p_h^*(\lambda)\nabla\cdot\mathbf{v}_h dx - \int_{\Gamma_{12}}\mathbf{v}_h\cdot\mathbf{n}_{\partial\Omega_2}\lambda d\sigma \quad \forall \mathbf{v}_h\in\mathbf{V}_{h,0}(\Omega_2),$$

$$\int_{\Omega_2}\nabla\cdot\mathbf{u}_h^*(\lambda) q_h dx = 0 \quad \forall q_h \in W_h(\Omega_2).$$

Define a linear functional $g_H : L^2(\Gamma_{12}) \mapsto \mathbb{R}$ by $g_H(\mu) := \sum_{i=1}^2 g_{H,i}(\mu)$:

$$g_{H,1}(\mu) := -\int_{\Gamma_{12}} \mathbf{K}\,\nabla\bar{p}_h|_{\Omega_1}\cdot\mathbf{n}_{\partial\Omega_1}\mu d\sigma + \sum_{\tau\in\Gamma_H}\frac{\sigma_\tau}{H_\tau}\int_\tau \bar{p}_h|_{\Omega_1}\mu d\sigma,$$

$$g_{H,2}(\mu) := \int_{\Gamma_{12}}\bar{\mathbf{u}}_h|_{\Omega_2}\cdot\mathbf{n}_{\partial\Omega_2}\mu d\sigma,$$

where $\bar{p}_h \in X_h(\Omega_1)$ solves

$$B_1\left(\bar{p}_h, q_h\right) = l_1\left(q_h\right) \quad \forall q_h \in W_h(\Omega_1)$$

and $\bar{\mathbf{u}}_h$ is the first component of the solution $(\bar{\mathbf{u}}_h, \bar{p}_h) \in (\mathbf{V}_{h,0}(\Omega_2) + \mathbf{u}_{h,g})\times W_h(\Omega_2)$ of

$$\int_{\Omega_2}\mathbf{K}^{-1}\bar{\mathbf{u}}_h\cdot\mathbf{v}_h dx = \int_{\Omega_2}\bar{p}_h\nabla\cdot\mathbf{v}_h dx \quad \forall\mathbf{v}_h\in\mathbf{V}_{h,0}(\Omega_2),$$

$$\int_{\Omega_2}\nabla\cdot\bar{\mathbf{u}}_h\, q_h dx = \int_{\Omega_2} f\, q_h\, dx \quad \forall q_h\in W_h(\Omega_2).$$

It is easy to see that the solution $(p_h, \mathbf{u}_h, \lambda_H)$ of the DG-MFE scheme (4.12)–(4.15) satisfies

$$d_H\left(\lambda_H, \mu_H\right) = g_H\left(\mu_H\right), \quad \mu_H \in \lambda_H,$$

with

$$p_h = p_h^*(\lambda_H) + \bar{p}_h \quad \text{on } \Omega_1 \text{ and } \Omega_2; \qquad \mathbf{u}_h = \mathbf{u}_h^*(\lambda_H) + \bar{\mathbf{u}}_h \quad \text{on } \Omega_2.$$

We now analyze the properties of the bilinear form $d_H(\cdot,\cdot)$ for the various DG-MFE schemes. By taking $q_h = p_h^*(\mu)$ in (4.32) and $\mathbf{v}_h = \mathbf{u}_h^*(\mu)$ in (4.33) for some $\mu \in L^2(\Gamma_{12})$, we obtain

$$B_1\left(p_h^*(\lambda), p_h^*(\mu)\right) = b_1\left(\lambda, p_h^*(\mu)\right) \tag{4.34}$$

and

$$\int_{\Omega_2} \mathbf{K}^{-1}\mathbf{u}_h^*(\lambda)\cdot\mathbf{u}_h^*(\mu)\,dx = -\int_{\Gamma_{12}} \mathbf{u}_h^*(\mu)\cdot\mathbf{n}_{\partial\Omega_2}\lambda\,d\sigma. \tag{4.35}$$

By using (3.34), (4.34), and (4.35), we have the following representation of $d_H(\cdot,\cdot)$:

$$\begin{aligned} d_H(\lambda,\mu) &= b_1(\mu, p_h^*(\lambda)) - 2\sum_{\tau\in\Gamma_H}\frac{\sigma_\tau}{H_\tau}\int_\tau p_h^*(\lambda)|_{\Omega_1}\mu\,d\sigma \\ &\quad + \sum_{\tau\in\Gamma_H}\frac{\sigma_\tau}{H_\tau}\int_\tau \lambda\,\mu\,d\sigma - \int_{\Gamma_{12}} \mathbf{u}_h^*(\lambda)|_{\Omega_2}\cdot\mathbf{n}_{\partial\Omega_2}\mu\,d\sigma \\ &= B_1(p_h^*(\mu), p_h^*(\lambda)) - 2\sum_{\tau\in\Gamma_H}\frac{\sigma_\tau}{H_\tau}\int_\tau p_h^*(\lambda)|_{\Omega_1}\mu\,d\sigma \\ &\quad + \sum_{\tau\in\Gamma_H}\frac{\sigma_\tau}{H_\tau}\int_\tau \lambda\,\mu\,d\sigma + \int_{\Omega_2}\mathbf{K}^{-1}\mathbf{u}_h^*(\mu)\mathbf{u}_h^*(\lambda)\,dx. \end{aligned}$$

By an argument similar to that used in proving (3.40), using the above representation and (3.1), we obtain

$$\begin{aligned} d_H(\lambda,\lambda) &\geq \frac{3}{4}\|\mathbf{K}^{\frac{1}{2}}\nabla p_h^*(\lambda)\|_{\Omega_1}^2 + \frac{3}{4}\sum_{\gamma\in\Gamma_h(\Omega_1)}\frac{\sigma_\gamma}{h_\gamma}\int_\gamma \left[p_h^*(\lambda)\right]^2 d\sigma \\ &\quad + \sum_{\tau\in\Gamma_H}\frac{\sigma_\tau}{H_\tau}\int_\tau \left(p_h^*(\lambda)|_{\Omega_1} - \lambda\right)^2 d\sigma + \|\mathbf{K}^{-\frac{1}{2}}\mathbf{u}_h^*(\lambda)\|_{0,\Omega_2}^2, \end{aligned}$$

implying that $d_H(\cdot,\cdot)$ is positive semidefinite on $L^2(\Gamma_{12})$. The argument in the solvability Theorem 4.2 implies that the kernel of $d_H(\cdot,\cdot)$ consists of the constant functions. We have obtained the following result.

THEOREM 4.6. *Let the assumptions of Theorem* 4.2 *hold. For all four versions of coupled DG-MFE methods, the interface bilinear form $d_H(\cdot,\cdot)$ is nonsymmetric and positive semidefinite on $L^2(\Gamma_{12})$, with the kernel consisting of the constant functions.*

Remark 4.4. In the case of $\bar{s}_{\text{form}} = 0$ or 1, coercivity of $d_H(\cdot,\cdot)$ can be shown for all modified DG-MFE schemes introduced in Remark 4.2, by using inequality (3.32). It is easy to see that $d_H(\cdot,\cdot)$ is symmetric for SIPG with $\bar{s}_{\text{form}} = 1$.

5. Discussion and conclusions. We have developed a multiscale formulation for coupling DG with DG and DG with MFE using mortar spaces. The method is based on imposing weak continuity of flux and pressure via a Robin-type matching condition with a penalized pressure jump. Although the formulations described in this paper are for two subdomains, the results can be extended to geometrically nonconforming domain decompositions with a finite number of subdomains.

Our mortar formulation can be viewed as a two-level domain decomposition solver via reduction to an interface problem. By choosing the special case of continuous approximating functions in the subdomains, this approach allows the coupling of

continuous Galerkin (CG) with DG, CG with MFE, and CG with CG. The latter represents a new mortar domain decomposition algorithm for CG.

Appendix A. Mortar compatibility conditions. Here we define the mortar compatibility conditions needed for solvability of the methods with $\sigma_\tau = 0$.

DEFINITION (mortar compatibility conditions). *We say that a DG space $X_h(\Omega_i)$ is compatible with a mortar space Λ_H if, for any $\mu_H \in \Lambda_H$,*

$$\int_{\Gamma_{12}} \mathbf{K}\nabla q_h \cdot \mathbf{n}\, \mu_H d\sigma = 0 \quad \forall q_h \in X_h(\Omega_i) \;\Rightarrow \mu_H = 0. \tag{A.1}$$

We say that the MFE space $\mathbf{V}_{h,0}(\Omega_2)$ is compatible with the mortar space Λ_H if, for any $\mu_H \in \Lambda_H$,

$$\int_{\Gamma_{12}} \mathbf{v}_h \cdot \mathbf{n}\, \mu_H d\sigma = 0 \quad \forall \mathbf{v}_h \in \mathbf{V}_{h,0}(\Omega_2) \;\Rightarrow \mu_H = 0. \tag{A.2}$$

Note that (A.1) is imposed only if $\sigma_\tau = 0$. For DG-MFE methods, either (A.1) or (A.2) is needed. For matching meshes, one can choose Γ_H to be the trace of the subdomain grids. In this case a sufficient condition for (A.1) is $r \geq \bar{r} + 1$, and a sufficient condition for (A.2) is $m \geq \bar{r}$. For nonmatching meshes, if a compatibility condition is needed, it is imposed only on one of the subdomains, allowing flexibility for the mesh and the finite element space in the other subdomain.

The compatibility condition (A.2) limits the richness of the mortar space. It has been studied in [47, 4, 33], and it has been shown to hold for very general nonmatching configurations of mortar and subdomain grids and spaces.

Below, we give some examples of spaces on nonmatching meshes satisfying the compatibility condition (A.1) for $d = 2$. For simplicity we assume that $\mathbf{K}$ is a constant tensor in each element.

PROPOSITION A.1. *Let $\mathbf{K}$ be constant in each element. Assume that $d = 2$, $r = 2$, and each element of Γ_H contains at least two element faces from $\mathcal{E}_h(\Omega_1) \cap \Gamma_{12}$. Then the DG space $X_h(\Omega_1)$ and the mortar space Λ_H with $\bar{r} = 2$ or $\bar{r} = 3$ satisfy* (A.1).

Proof. Consider a mortar element $\tau \in \Gamma_H$, and assume that $\tau \supset \gamma_1 \cup \gamma_2$, where γ_1 and γ_2 are the edges of two distinct elements $E_1 \subset \mathcal{E}_h(\Omega_1)$ and $E_2 \subset \mathcal{E}_h(\Omega_1)$. Since $\mathbf{K}$ is nonsingular (because $\mathbf{K}$ is positive definite), $\mathbf{Kn}$ is not zero. Noting that $\mathbf{K}\nabla q_h \cdot \mathbf{n} = \nabla q_h \cdot \mathbf{Kn}$ (because $\mathbf{K}$ is symmetric) and q_h ranges over $\mathbb{P}_2(\gamma_i)$, we conclude that $\mathbf{K}\nabla q_h \cdot \mathbf{n}$ ranges over $\mathbb{P}_1(\gamma_i)$.

First let us consider $\bar{r} = 2$ for Λ_H and take $\mu_H \in \mathbb{P}_2(\tau)$. The orthogonality condition in (A.1) yields

$$\int_{\gamma_i} p_1 \mu_H d\sigma = 0 \quad \forall p_1 \in \mathbb{P}_1(\gamma_i)\,,\; i = 1, 2.$$

In particular, $\int_{\gamma_1} \mu_H d\sigma = \int_{\gamma_2} \mu_H d\sigma = 0$. Therefore, μ_H has a root inside γ_1, say, α_1, and a root inside γ_2, say, α_2. Therefore $\mu_H = C(x - \alpha_1)(x - \alpha_2)$. By choosing $p_1 = (x - \alpha_1)$ in γ_1, we have $\int_{\gamma_1} C(x - \alpha_1)^2(x - \alpha_2) d\sigma = 0$. Since $(x - \alpha_2)$ cannot change sign in γ_1, we must have $C = 0$; consequently $\mu_H = 0$, and (A.1) holds.

We now let $\bar{r} = 3$ and take $\mu_H \in \mathbb{P}_3(\tau)$. Then, as before, we have two distinct roots $\alpha_1 \in \gamma_1$ and $\alpha_2 \in \gamma_2$. Since $\mu_H \in \mathbb{P}_3(\tau)$, we know that μ_H has another real root, say, α_3, and then $\mu_H = C(x - \alpha_1)(x - \alpha_2)(x - \alpha_3)$. By taking $p_1 = (x - \alpha_1)$

in γ_1, we have $\int_{\gamma_1} C(x-\alpha_1)^2(x-\alpha_2)(x-\alpha_3)d\sigma = 0$. Since $(x-\alpha_2)$ cannot change sign in γ_1, we must have either $C=0$ or $\alpha_3 \in \gamma_1$. By taking $p_1 = (x-\alpha_2)$ in γ_2, we conclude similarly that either $C=0$ or $\alpha_3 \in \gamma_2$. As γ_1 and γ_2 are disjoint, we must have $C=0$; consequently $\mu_H = 0$, and (A.1) holds. ∎

REFERENCES

[1] Y. Achdou, Y. Maday, and O. Widlund, *Iterative substructuring preconditioners for mortar element methods in two dimensions*, SIAM J. Numer. Anal., 36 (1999), pp. 551–580.

[2] Y. Achdou, Y. A. Kuznetsov, and O. Pironneau, *Substructuring preconditioners for the Q_1 mortar element method*, Numer. Math., 71 (1995), pp. 419–449.

[3] V. Aizinger, C. N. Dawson, B. Cockburn, and P. Castillo, *The local discontinuous Galerkin method for contaminant transport*, Advances in Water Resources, 24 (2000), pp. 73–87.

[4] T. Arbogast, L. C. Cowsar, M. F. Wheeler, and I. Yotov, *Mixed finite element methods on nonmatching multiblock grids*, SIAM J. Numer. Anal., 37 (2000), pp. 1295–1315.

[5] T. Arbogast, G. Pencheva, M. F. Wheeler, and I. Yotov, *A multiscale mortar mixed finite element method*, Multiscale Model. Simul., 6 (2007), pp. 319–346.

[6] T. Arbogast, M. F. Wheeler, and I. Yotov, *Mixed finite elements for elliptic problems with tensor coefficients as cell-centered finite differences*, SIAM J. Numer. Anal., 34 (1997), pp. 828–852.

[7] D. N. Arnold, *An interior penalty finite element method with discontinuous elements*, SIAM J. Numer. Anal., 19 (1982), pp. 742–760.

[8] D. N. Arnold and F. Brezzi, *Mixed and nonconforming finite element methods: Implementation, postprocessing and error estimates*, RAIRO Modèl. Math. Anal. Numèr., 19 (1985), pp. 7–32.

[9] F. Bassi and S. Rebay, *A high-order accurate discontinuous finite element method for the numerical solution of the compressible Navier-Stokes equations*, J. Comput. Phys., 131 (1997), pp. 267–279.

[10] R. Becker, P. Hansbo, and R. Stenberg, *A finite element method for domain decomposition with non-matching grids*, M2AN Math. Model. Numer. Anal., 37 (2003), pp. 209–225.

[11] F. B. Belgacem, *The mortar finite element method with Lagrange multiliers*, Numer. Math., 84 (1999), pp. 173–197.

[12] C. Bernardi, Y. Maday, and A. T. Patera, *A new nonconforming approach to domain decomposition: The mortar element method*, in Nonlinear Partial Differential Equations and Their Applications, Collège de France Seminar, Vol. XI (Paris, 1989–1991), Pitman Res. Notes Math. Ser. 299, Longman Sci. Tech., Harlow, 1994, pp. 13–51.

[13] S. Brenner, *Poincaré–Friedrichs inequalities for piecewise H^1 functions*, SIAM J. Numer. Anal., 41 (2003), pp. 306–324.

[14] F. Brezzi, J. Douglas, Jr., R. Durán, and M. Fortin, *Mixed finite elements for second order elliptic problems in three variables*, Numer. Math., 51 (1987), pp. 237–250.

[15] F. Brezzi, J. Douglas, Jr., M. Fortin, and L. D. Marini, *Efficient rectangular mixed finite elements in two and three space variables*, RAIRO Modèl. Math. Anal. Numèr., 21 (1987), pp. 581–604.

[16] F. Brezzi, J. Douglas, Jr., and L. D. Marini, *Two families of mixed elements for second order elliptic problems*, Numer. Math., 88 (1985), pp. 217–235.

[17] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer-Verlag, New York, 1991.

[18] Z. Chen and J. Douglas, Jr., *Prismatic mixed finite elements for second order elliptic problems*, Calcolo, 26 (1989), pp. 135–148.

[19] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North-Holland, New York, 1978.

[20] B. Cockburn and C. Dawson, *Approximation of the velocity by coupling discontinuous Galerkin and mixed finite element methods for flow problems*, Comput. Geosci., 6 (2002), pp. 505–522.

[21] B. Cockburn and C.-W. Shu, *The local discontinuous Galerkin method for time-dependent convection-diffusion systems*, SIAM J. Numer. Anal., 35 (1998), pp. 2440–2463.

[22] L. C. Cowsar, J. Mandel, and M. F. Wheeler, *Balancing domain decomposition for mixed finite elements*, Math. Comp., 64 (1995), pp. 989–1015.

[23] C. Dawson, S. Sun, and M. F. Wheeler, *Compatible algorithms for coupled flow and transport*, Comput. Methods Appl. Mech. Engrg., 193 (2004), pp. 2565–2580.

[24] R. Ewing, R. Lazarov, T. Lin, and Y. Lin, *Mortar finite volume element approximations of second order elliptic problems*, East-West J. Numer. Math., 8 (2000), pp. 93–110.
[25] V. Girault and P.-A. Raviart, *Finite Element Methods for Navier-Stokes Equations*, Springer-Verlag, Berlin, 1986.
[26] R. Glowinski and M. F. Wheeler, *Domain decomposition and mixed finite element methods for elliptic problems*, in First International Symposium on Domain Decomposition Methods for Partial Differential Equations, R. Glowinski, G. H. Golub, G. A. Meurant, and J. Periaux, eds., SIAM, Philadelphia, 1988, pp. 144–172.
[27] J. Gopalakrishnan and J. E. Pasciak, *Multigrid for the mortar finite element method*, SIAM J. Numer. Anal., 37 (2000), pp. 1029–1052.
[28] Y. A. Kuznetsov and M. F. Wheeler, *Optimal order substructuring preconditioners for mixed finite element methods on non-matching grids*, East-West J. Numer. Math., 3 (1995), pp. 127–143.
[29] J. L. Lions and E. Magenes, *Non-Homogeneous Boundary Value Problems and Applications*, Vol. 1, Springer-Verlag, Berlin, 1972.
[30] J. Mandel and M. Brezina, *Balancing domain decomposition for problems with large jumps in coefficients*, Math. Comp., 65 (1996), pp. 1387–1401.
[31] J. C. Nedelec, *Mixed finite elements in* $\mathbf{R}^3$, Numer. Math., 35 (1980), pp. 315–341.
[32] J. T. Oden, I. Babuška, and C. E. Baumann, *A discontinuous hp finite element method for diffusion problems*, J. Comput. Phys., 146 (1998), pp. 491–516.
[33] G. Pencheva and I. Yotov, *Balancing domain decomposition for mortar mixed finite element methods on non-matching grids*, Numer. Linear Algebra Appl., 10 (2003), pp. 159–180.
[34] M. Peszynska, Q. Lu, and M. F. Wheeler, *Coupling different numerical algorithms for two phase fluid flow*, in MAFELAP Proceedings of Mathematics of Finite Elements and Applications, Uxbridge, U.K., 1999, J. R. Whiteman, ed., pp. 205–214.
[35] M. Peszyńska, M. F. Wheeler, and I. Yotov, *Mortar upscaling for multiphase flow in porous media*, Comput. Geosci., 6 (2002), pp. 73–100.
[36] R. A. Raviart and J. M. Thomas, *A mixed finite element method for 2nd order elliptic problems*, in Mathematical Aspects of the Finite Element Method, Lecture Notes in Math. 606, Springer-Verlag, New York, 1977, pp. 292–315.
[37] B. Rivière and M. Wheeler, *Coupling locally conservative methods for single phase flow*, Comput. Geosci., 6 (2002), pp. 269–284.
[38] B. Rivière, M. F. Wheeler, and V. Girault, *Part* I: *Improved energy estimates for interior penalty, constrained and discontinuous Galerkin methods for elliptic problems*, Comput. Geosci., 3 (1999), pp. 337–360.
[39] B. Rivière, M. F. Wheeler, and V. Girault, *A priori error estimates for finite element methods based on discontinuous approximation spaces for elliptic problems*, SIAM J. Numer. Anal., 39 (2001), pp. 902–931.
[40] T. F. Russell and M. F. Wheeler, *Finite Element and Finite Difference Methods for Continuous Flows in Porous Media*, in The Mathematics of Reservoir Simulation, R. E. Ewing, ed., SIAM, Philadelphia, 1983, pp. 35–106.
[41] S. Sun, B. Rivière, and M. F. Wheeler, *A combined mixed finite element and discontinuous Galerkin method for miscible displacement problems in porous media*, in Proceedings for the International Conference Recent Progress on Computational and Applied PDEs, Zhangjiaje, China, 2001, pp. 321–348.
[42] S. Sun and M. F. Wheeler, *Symmetric and nonsymmetric discontinuous Galerkin methods for reactive transport in porous media*, SIAM J. Numer. Anal., 43 (2005), pp. 195–219.
[43] S. Sun, *Discontinuous Galerkin Methods for Reactive Transport in Porous Media*, Ph.D. thesis, The University of Texas at Austin, 2003.
[44] M. F. Wheeler and I. Yotov, *Physical and computational domain decompositions for modeling subsurface flows*, in Domain Decomposition Methods 10 (Boulder, CO, 1997), Contemp. Math. 218, AMS, Providence, RI, 1998, pp. 217–228.
[45] M. F. Wheeler, *An elliptic collocation-finite element method with interior penalties*, SIAM J. Numer. Anal., 15 (1978), pp. 152–161.
[46] B. I. Wohlmuth, *A mortar finite element method using dual spaces for the Lagrange multiplier*, SIAM J. Numer. Anal., 38 (2000), pp. 989–1012.
[47] I. Yotov, *Mixed Finite Element Methods for Flow in Porous Media*, Ph.D. thesis, Rice University, Houston, Texas, 1996.

SIAM J. NUMER. ANAL.
Vol. 46, No. 2, pp. 980–995
© 2008 Society for Industrial and Applied Mathematics

STABILITY ANALYSIS OF THE GURTIN–MACCAMY MODEL*

D. BREDA†, M. IANNELLI‡, S. MASET§, AND R. VERMIGLIO†

Abstract. In this paper we propose a numerical scheme to investigate the stability of steady states of the nonlinear Gurtin–MacCamy system, which is a basic model in population dynamics. In fact the analysis of stability is usually performed by the study of transcendental characteristic equations that are too difficult to approach by analytical methods. The method is based on the discretization of the infinitesimal generator associated to the semigroup of the solution operator by using pseudospectral differencing techniques. The method computes the rightmost characteristic roots, and it is shown to converge with spectral accuracy behavior.

Key words. age-structured population, asymptotic stability, characteristic roots, eigenvalue problem

AMS subject classifications. 92D25, 65J10, 65Q05

DOI. 10.1137/070685658

1. Introduction.

The nonlinear Gurtin–MacCamy model represents a milestone in the study of the dynamics of an age-structured population (see the original work [11]), and the qualitative analysis of its solutions is of central interest. Within this context, the investigation of the behavior of the solutions near an equilibrium requires two important steps: the determination of the steady state solutions and the stability analysis of the linearized system. In both cases the numerical methods represent powerful tools for a complete analysis. In particular, it is known that stability of steady states is governed by transcendental characteristic equations that are not easy to solve analytically. Thus efficient numerical methods to compute the characteristic roots allow one to explore the behavior of the models determining asymptotic stability. Moreover, we are interested in examining the displacement of the characteristic roots versus some significant parameter, in order to observe the possible crossing of the imaginary axis from the left to the right implying Hopf bifurcations and the existence of periodic solutions.

Within the framework described above, several papers have been devoted to the analysis of population models based on specific interactions between different ages. In particular adults-juveniles competition and cannibalism have been investigated [5], [6], [7], [8], [12], by both performing local stability analysis and simulating the solutions by special numerical methods [13], [14].

Numerical approaches for the computation of characteristic roots of linear autonomous delay differential equations have been recently proposed (see [10], [2], [3]). They are based on the discretization of either the solution operator or the infinitesimal generator relevant to the semigroup associated to the original equation. Since the spectrum of the infinitesimal generator consists of the characteristic roots and

*Received by the editors March 19, 2007; accepted for publication (in revised form) October 1, 2007; published electronically March 5, 2008.
http://www.siam.org/journals/sinum/46-2/68565.html
†Dipartimento di Matematica e Informatica, Università degli Studi di Udine, via delle Scienze 208, I-33100 Udine, Italy (dimitri.breda@dimi.uniud.it, vermiglio@dimi.uniud.it).
‡Dipartimento di Matematica, Università degli Studi di Trento, via Sommarive 14, I-38050 Povo (Trento), Italy (iannelli@science.unitn.it).
§Dipartimento di Matematica e Informatica, Università degli Studi di Trieste, via Valerio 12, I-34127 Trieste, Italy (maset@univ.trieste.it).

the spectrum of the solution operator is related to the characteristic roots by an exponential transformation, such roots can be computed as the eigenvalues of suitable matrices approximating the infinitesimal generator or the solution operator. In a recent paper [1], this technique has been proposed to analyze the stability properties for age-structured population models. Actually, the population model considered in [1] is a particular case of the Gurtin–MacCamy problem, because nonlinear effects are supposed to influence fertility only. The full problem is mathematically more complex, and the numerical method set up in [1] must be adapted and implemented to cover the full case.

Here we propose a general numerical approach for determining the asymptotic stability of solutions of a rather general linear problem which includes, as a particular case, the full Gurtin–MacCamy model linearized at a stationary solution. As a worked example, we consider a general age-structured logistic model and analyze its stability by showing Hopf bifurcation and the existence of periodic solutions.

2. The Gurtin–MacCamy model: Stationary solutions and stability analysis. Let us consider a general nonlinear Gurtin–MacCamy model, describing the dynamics of an age-structured population, here represented by its age density $p(a,t)$ at the age a and time t:

$$\frac{\partial p}{\partial t}(a,t)+\frac{\partial p}{\partial a}(a,t)+\mu(a,S(t))p(a,t)=0,\ a\in[0,a_\dagger],\ t\geq 0, \tag{2.1}$$

$$p(0,t)=\int_0^{a_\dagger}\beta(a,S(t))p(a,t)da,\ t\geq 0, \tag{2.2}$$

$$S(t)=\int_0^{a_\dagger}\gamma(a)p(a,t)da,\ t\geq 0, \tag{2.3}$$

where $\gamma(t)=(\gamma_1(t),\dots,\gamma_n(t))^T\in\mathbb{R}^n$ and then $S(t)=(S_1(t),\dots,S_n(t))^T\in\mathbb{R}^n$, with

$$S_i(t)=\int_0^{a_\dagger}\gamma_i(a)p(a,t)da,\ t\geq 0.$$

The n variables $S_i(t)$, $i=1,\dots,n$, are called *sizes* and represent different ways of weighting the age distribution. Note that both *fertility* β and *mortality* μ depend on the age and on the sizes. We will assume that the functions

$$\begin{aligned}&\beta:[0,a_\dagger]\times\mathbb{R}^n\to\mathbb{R},\\&\mu:[0,a_\dagger]\times\mathbb{R}^n\to\mathbb{R},\\&\gamma_i:[0,a_\dagger]\to\mathbb{R}\quad(i=1,\dots,n)\end{aligned}$$

are measurable and satisfy the following general conditions:

$$0\leq\beta(a,x)\leq\beta_+\ \text{a.e. in}\ [0,a_\dagger]\times\mathbb{R}^n,$$

$$0\leq\gamma_i(a)\leq\gamma_+\ \text{a.e. in}\ [0,a_\dagger],$$

$$0\leq\mu(a,x)=\mu_0(a)+\mathcal{M}(a,x)\quad\text{a.e. in}\ [0,a_\dagger]\times\mathbb{R}^n,$$

where $\mu_0(a)$ satisfies

$$\mu_0(\cdot)\in L^1_{loc}([0,a_\dagger)),\mu_0(a)\geq 0\quad\text{a.e. in}\ [0,a_\dagger],$$

$$\int_0^{a_\dagger}\mu_0(\sigma)d\sigma=+\infty,$$

and

$$\mathcal{M}(a,0) = 0 \quad \text{a.e. in } [0,a_\dagger].$$

Moreover, we assume that $\beta, \mathcal{M}$ are locally Lipschitz with respect to the second argument. These assumptions ensure that an initial condition

$$p(a,0) = p_0(a),\ a \in [0,a_\dagger],$$

uniquely fixes a solution for (2.1), (2.2), and (2.3).

A stationary solution $p^*(a)$ of the model (2.1), (2.2), and (2.3) satisfies the following equations:

$$\frac{d}{da}p^*(a) + \mu(a,S^*)p^*(a) = 0,\ a \in [0,a_\dagger], \tag{2.4}$$

$$p^*(0) = \int_0^{a_\dagger} \beta(a,S^*)p^*(a)da,$$

$$S^* = \int_0^{a_\dagger} \gamma(a)p^*(a)da.$$

By defining the function (*survival probability* at S^*)

$$\Pi(a,S^*) = \exp\left(-\int_0^a \mu(\sigma,S^*)d\sigma\right),\ a \in [0,a_\dagger], \tag{2.5}$$

a solution of (2.4) is given by

$$p^*(a) = p^*(0)\Pi(a,S^*),\ a \in [0,a_\dagger], \tag{2.6}$$

and so it is completely described by the values $p^*(0)$ and S^*. Besides the trivial stationary solution $p^*(a) = 0$, $a \in [0,a_\dagger]$, nontrivial solutions are given by positive S^* satisfying

$$1 = \int_0^{a_\dagger} \beta(a,S^*)\Pi(a,S^*)da \tag{2.7}$$

and

$$\frac{S_1^*}{\int_0^{a_\dagger} \gamma_1(a)\Pi(a,S^*)da} = \frac{S_2^*}{\int_0^{a_\dagger} \gamma_2(a)\Pi(a,S^*)da} = \cdots = \frac{S_n^*}{\int_0^{a_\dagger} \gamma_n(a)\Pi(a,S^*)da}, \tag{2.8}$$

i.e., a nonlinear algebraic system of n equations in the n unknowns S_i^*, $i = 1,\dots,n$. The value $p^*(0)$ is then given by

$$p^*(0) = \frac{S_i^*}{\int_0^{a_\dagger} \gamma_i(a)\Pi(a,S^*)da},\ i = 1,\dots,n.$$

To investigate the stability of a stationary solution p^*, we linearize the model (2.1), (2.2), and (2.3) around p^*, and we get for the function $u = p - p^*$ the following linear equations:

$$\frac{\partial u}{\partial t}(a,t) + \frac{\partial u}{\partial a}(a,t) + \mu(a,S^*)u(a,t) + p^*(a)\frac{\partial \mu}{\partial S}(a,S^*)U(t) = 0, \tag{2.9}$$

$$u(0,t) = \int_0^{a_\dagger} \beta(a,S^*)u(a,t)da + \int_0^{a_\dagger} p^*(a)\frac{\partial \beta}{\partial S}(a,S^*)U(t)da, \tag{2.10}$$

$$U(t) = (U_1(t),\dots,U_n(t))^T,\ U_i(t) = \int_0^{a_\dagger} \gamma_i(a)u(a,t)da, \tag{2.11}$$

where we used the notation $\frac{\partial f}{\partial X}(a,X)$, with $X = (x_1,\dots,x_n)^T$ for the row-vector gradient $(\frac{\partial f}{\partial x_1}(a,X),\dots,\frac{\partial f}{\partial x_n}(a,X))$ and

$$\frac{\partial f}{\partial X}(a,X)Y = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(a,X)Y_i.$$

We note that, by setting

$$u(a,t) = \Pi(a,S^*)v(a,t),\ a \in [0,a_\dagger],\ t \geq 0,$$

where $\Pi(a,S^*)$ is the survival probability defined in (2.5), from (2.9), (2.10), and (2.11) we get the following simplified equations for the function v:

$$\frac{\partial v}{\partial t}(a,t) + \frac{\partial v}{\partial a}(a,t) + p^*(0)\frac{\partial \mu}{\partial S}(a,S^*)V(t) = 0, \tag{2.12}$$

$$v(0,t) = \int_0^{a_\dagger} \beta(a,S^*)\Pi(a,S^*)v(a,t)da + \int_0^{a_\dagger} p^*(a)\frac{\partial \beta}{\partial S}(a,S^*)V(t)da, \tag{2.13}$$

$$V(t) = (V_1(t),\dots,V_n(t))^T,\ V_i(t) = \int_0^{a_\dagger} \gamma_i(a)\Pi(a,S^*)v(a,t)da. \tag{2.14}$$

It is important to observe that the zero solution of (2.9), (2.10), and (2.11) is asymptotically stable if and only if the zero solution of (2.12), (2.13), and (2.14) is asymptotically stable. Hence, the stability analysis of the linearized model can be equivalently performed with the equations for the function v.

3. Abstract formulation and characteristic equation. We are interested in studying the stability of the steady states by analyzing the stability of the zero solution of the linearized model (2.12), (2.13), and (2.14). In order to present and develop a numerical approach we rewrite the general model as an abstract evolution equation, and we derive the characteristic equation. This last step is essential to introduce the numerical approach in the next section.

In order to present our numerical method in a more general setting, we enlarge our attention to complex-valued functions, and we consider the Banach space $X = L^1([0,a_\dagger],\mathbb{C})$. Then, by introducing the linear bounded functional $K_0 : X \to \mathbb{C}$ and the linear bounded operator $\mathcal{H} : X \to X$ of finite rank, i.e.,

$$\operatorname{ran}(\mathcal{H}) = \operatorname{span}\{h_1,\dots,h_n\} \tag{3.1}$$

for some $h_i \in X$, $i = 1,\dots,n$, we propose to consider the following general linear equation:

$$\frac{\partial v}{\partial t}(a,t) + \frac{\partial v}{\partial a}(a,t) + (\mathcal{H}v(\cdot,t))(a) = 0,\ a \in [0,a_\dagger],\ t \geq 0, \tag{3.2}$$

with boundary condition

$$v(0,t) = K_0 v(\cdot,t),\ t \geq 0. \tag{3.3}$$

Observe that (3.1) implies that

$$(\mathcal{H}\varphi)(a) = H(a)K\varphi = \sum_{i=1}^{n} h_i(a)K_i\varphi,\ \varphi \in X, \tag{3.4}$$

where $H(a) = (h_1(a), \dots, h_n(a))$ and $K\varphi = (K_1\varphi, \dots, K_n\varphi)^T$, with $K_i : X \to \mathbb{C}, i = 1, \dots, n$, suitable linear bounded functionals.

The model (3.2) is a generalized version of problem (2.12), (2.13), (2.14), which fits into it by choosing the linear functionals K_0, K in (3.3) and (3.4), respectively, as follows:

$$K_0\varphi = \int_0^{a_\dagger} k_0(s)\varphi(s)ds,\ \varphi \in X, \tag{3.5}$$

with kernel $k_0 \in L^\infty([0, a_\dagger], \mathbb{C})$ given by

$$k_0(a) = \Pi(a, S^*)\left(\beta(a, S^*) + \sum_{i=1}^{n} \gamma_i(a) \int_0^{a_\dagger} p^*(s) \frac{\partial \beta}{\partial S_i}(s, S^*)\, ds\right),$$

$$K\varphi = \int_0^{a_\dagger} k(s)\varphi(s)ds = \left(\int_0^{a_\dagger} k_1(s)\varphi(s)ds, \dots, \int_0^{a_\dagger} k_n(s)\varphi(s)ds\right),\ \varphi \in X, \tag{3.6}$$

with a vector kernel $k(a) = (k_1(a), \dots, k_n(a))^T$, $k_i \in L^\infty([0, a_\dagger], \mathbb{C})$ for $i = 1, \dots, n$ given by

$$k(a) = \Pi(a, S^*)\gamma(a),$$

and

$$H(a) = p^*(0)\frac{\partial \mu}{\partial S}(a, S^*).$$

Problem (3.2) can be restated as the following abstract evolution equation on the state space X:

$$\frac{d}{dt}U(t) = \mathcal{A}U(t),\ t \geq 0,$$

where $\mathcal{A} : \mathrm{dom}(\mathcal{A}) \subseteq X \to X$ is the unbounded linear operator given by

$$\mathcal{A}\varphi = -\varphi' - \mathcal{H}\varphi,\ \varphi \in \mathrm{dom}(\mathcal{A}), \tag{3.7}$$

with domain

$$\mathrm{dom}(\mathcal{A}) = \{\varphi \in X \mid \varphi' \in X \text{ and } \varphi(0) = K_0\varphi\}. \tag{3.8}$$

Observe that $\mathcal{A}$ in (3.7) is the infinitesimal generator of the C_0-semigroup of the solution operators $\{T(t)\}_{t\geq 0}$ defined by

$$T(t)\varphi = v(\cdot, t),\ \varphi \in X,$$

where $v(a, t)$ is the solution of (3.2) fulfilling the initial condition

$$v(a, 0) = \varphi(a),\ a \in [0, a_\dagger]$$

(see [9]). It can be shown that such a semigroup is eventually compact, and so the spectrum of the infinitesimal generator $\mathcal{A}$ consists of eigenvalues only.

The asymptotic stability of the zero solution of (3.2) is determined by analyzing the position in the complex plane of the eigenvalues of the infinitesimal generator $\mathcal{A}$ defined in (3.7) and (3.8). In particular the (hyperbolic) asymptotic stability holds if and only if all of the eigenvalues lie in the left half-plane $\Re(z) < 0$. Now we show that the eigenvalues of $\mathcal{A}$ satisfy a characteristic equation.

Let $\lambda \in \mathbb{C}$ and $\varphi \in D(\mathcal{A})$. The condition

$$\mathcal{A}\varphi = \lambda\varphi$$

means

$$\begin{cases} \varphi'(a) = -\lambda\varphi(a) - H(a)K\varphi, \ a \in [0, a_\dagger], \\ \varphi(0) = K_0\varphi. \end{cases} \tag{3.9}$$

By defining $x_0 = K_0\varphi$, $x = K\varphi$, and

$$\hat{H}(a,\lambda) = \int_0^a e^{-\lambda(a-s)} H(s)ds, \ a \in [0, a_\dagger], \tag{3.10}$$

we can equivalently rewrite (3.9) as

$$\varphi(a) = e^{-\lambda a}x_0 - \hat{H}(a,\lambda)x, \ a \in [0, a_\dagger].$$

Hence λ is eigenvalue of $\mathcal{A}$ if and only if there exists a nonzero solution $\overline{x} = \left(x_0, x^T\right)^T \in \mathbb{C}^{n+1}$ of the system

$$\begin{cases} x_0 = K_0 e^{-\lambda \cdot} x_0 - K_0 \hat{H}(\cdot,\lambda)x, \\ x = K e^{-\lambda \cdot} x_0 - K \hat{H}(\cdot,\lambda)x. \end{cases}$$

By introducing the matrix

$$\overline{A}(\lambda) = \begin{pmatrix} K_0 e^{-\lambda \cdot} & -K_0 \hat{H}(\cdot,\lambda) \\ K e^{-\lambda \cdot} & -K \hat{H}(\cdot,\lambda) \end{pmatrix} \in \mathbb{C}^{(n+1)\times(n+1)}, \tag{3.11}$$

we can conclude that λ is an eigenvalue of $\mathcal{A}$ if and only if

$$\det\left(I_{n+1} - \overline{A}(\lambda)\right) = 0, \tag{3.12}$$

which represents the characteristic equation for (3.2). We note that, since the entries of $\overline{A}(\lambda)$ are entire functions, the roots of (3.12) are a discrete set of the complex plane, and any right half-plane contains only a finite number of roots.

These characteristic roots are the eigenvalues of the infinitesimal generator $\mathcal{A}$ and can be approximated by the eigenvalues of a suitable matrix which is a discrete approximation of $\mathcal{A}$. For this task the pseudospectral technique guarantees a highly accurate discretization, and recently it has been proposed to investigate the behavior of solutions for different classes of mathematical models [2], [3]. In particular, this method has been applied to the Gurtin–MacCamy equation (2.1), (2.2), and (2.3) in the special case of the nonlinearity restricted to the fertility rate [1]. In order to treat a nonlinear mortality rate the method has to be extended and improved. Such a discretization technique is presented in the next section.

4. The numerical approach. We construct a discrete approximation of the infinitesimal generator $\mathcal{A}$ defined in (3.7) and in (3.8), under the assumption that H is continuous.

For a given positive integer N, let Ω_N be the mesh of Chebyshev extremal nodes on $[0, a_\dagger]$:

$$\Omega_N = \left\{ \theta_i = \frac{a_\dagger}{2} \cos\left(\frac{N-i}{N}\pi \right) + \frac{a_\dagger}{2},\ i = 0, \dots, N \right\}.$$

We replace the continuous space X by the space $X_N = (\mathbb{C})^{\Omega_N \setminus \{0\}} \cong \mathbb{C}^N$ of the discrete functions defined on $\Omega_N \setminus \{0\}$; i.e., any $\varphi \in X$ is discretized into the N-vector $y = (y_1, \dots, y_N)^T \in X_N$ of components $y_i = \varphi(\theta_i) \in \mathbb{C}$, $i = 1, \dots, N$.

Now we replace the operator $\mathcal{A}$ by the finite-dimensional linear operator $\mathcal{A}_N : X_N \to X_N$ given by

$$\mathcal{A}_N y = -z,\ y \in X_N,$$

where

$$\begin{aligned} z_i &= \varphi_N'(\theta_i) + (\mathcal{H}\varphi_N)(\theta_i) \\ &= \varphi_N'(\theta_i) + H(\theta_i) K \varphi_N,\ i = 1, \dots, N, \end{aligned}$$

and φ_N is the N-degree polynomial interpolating the values y_i at the nodes θ_i, $i = 1, \dots, N$, and such that

$$\varphi_N(0) = K_0 \varphi_N. \tag{4.1}$$

By using the representation

$$\varphi_N(a) = \sum_{i=0}^{N} l_{N,i}(a) y_i,\ a \in [0, a_\dagger],$$

where $y_0 = \varphi_N(0)$ and $l_{N,i}$ are the Lagrange coefficients relevant to the nodes θ_i, $i = 0, 1, \dots, N$, we obtain

$$z_i = \sum_{j=0}^{N} \left(l_{N,j}'(\theta_i) + H(\theta_i) K l_{N,j} \right) y_j,\ i = 1, \dots, N,$$

and, by the condition (4.1),

$$y_0 = \sum_{i=1}^{N} (1 - K_0 l_{N,0})^{-1} K_0 l_{N,i} y_i.$$

For K_0 defined in (3.5), we have that $(1 - K_0 l_{N,0})^{-1} \to 1$ for $N \to \infty$, since $\int_0^{a_\dagger} |l_{N,0}(s)| ds \to 0$ for $N \to \infty$. Moreover we remark that, if the functionals K_0, K are given by (3.5) and (3.6), respectively, and the integrals cannot be exactly computed, we approximate them by the Clenshaw–Curtis formula, which uses Chebyshev extremal points as quadrature nodes (see, for instance, [15]). Hence, in the case of kernels k_0, k which are piecewise smooth, it is advisable to use a piecewise polynomial φ_N as now explained. Moreover the use of piecewise polynomials is necessary when H is piecewise continuous. Assume there exist $0 = a_0 < a_1 < \cdots < a_d = a_\dagger$ such

that the kernels k_0 and k and the function H are smooth on each interval $[a_l, a_{l+1}]$, $l = 0, \dots, d-1$. We choose the mesh

$$\Omega_N = \bigcup_{l=1}^{d} \Omega_N^{(l)}$$

on the interval $[0, a_\dagger]$, where $\Omega_N^{(l)}$ is the mesh of Chebyshev extremal nodes $\theta_i^{(l)}, i = 0, \dots, N$, on $[a_{l-1}, a_l]$, and consider the space $X_N = (\mathbb{C})^{\Omega_N \setminus \{0\}} \cong \mathbb{C}^{dN}$ of the discrete functions defined on the mesh $\Omega_N \setminus \{0\}$; i.e., any $\varphi \in X$ is discretized into the dN-vector

$$y = \left(y_1^{(1)}, \dots, y_N^{(1)}, y_1^{(2)}, \dots, y_N^{(2)}, \dots, y_1^{(d)}, \dots, y_N^{(d)}\right)^T \in X_N$$

of components $y_i^{(l)} = \varphi(\theta_i^{(l)}) \in \mathbb{C}$, $i = 1, \dots, N$, $l = 1, \dots, d$. Note that $y_N^{(l-1)} = \varphi(a_{l-1})$ for $l = 2, \dots, d$.

Given a vector $y \in X_N$, the piecewise polynomial φ_N is then defined by the following conditions:

1. φ_N is continuous;
2. φ_N is an N-degree polynomial on every interval $[a_{l-1}, a_l]$, $l = 1, \dots, d$;
3. $\varphi_N(\theta_i^{(l)}) = y_i^{(l)}$, $i = 1, \dots, N, l = 1, \dots, d$.

The details of the discrete operator $\mathcal{A}_N$ can be easily obtained by proceeding as described above. For simplicity in this paper we do not consider the piecewise polynomial case, since it can be treated by similar arguments.

5. Convergence analysis. As a first step we derive the discrete analogous of the characteristic equation (3.12).

For $\lambda \in \mathbb{C}$ and $\overline{x} = (x_0, x^T)^T \in \mathbb{C}^{n+1}$, where $x_0 \in \mathbb{C}$ and $x \in \mathbb{C}^n$, let $y(\cdot; \lambda, \overline{x})$ be the solution of the initial value problem

$$\begin{cases} y'(a; \lambda, \overline{x}) = -\lambda y(a; \lambda, \overline{x}) - H(a)x, \ a \in [0, a_\dagger], \\ y(0; \lambda, \overline{x}) = x_0. \end{cases} \tag{5.1}$$

It follows that

$$y(a; \lambda, \overline{x}) = e^{-\lambda a} x_0 - \hat{H}(a, \lambda)x, \ a \in [0, a_\dagger],$$

where $\hat{H}(a, \lambda)$ is defined in (3.10). Let us observe that the matrix $\overline{A}(\lambda)$ given in (3.11) is such that

$$\overline{A}(\lambda)\overline{x} = \overline{K} y(\cdot; \lambda, \overline{x}), \ \overline{x} \in \mathbb{C}^{n+1}, \tag{5.2}$$

where $\overline{K}\varphi$ denotes $(K_0\varphi, (K\varphi)^T)^T \in \mathbb{C}^{n+1}$. Moreover let $p_N(\cdot; \lambda, \overline{x})$ be the N-degree collocation polynomial relevant to the nodes in Ω_N for the initial value problem (5.1); i.e., it satisfies

$$\begin{cases} p_N'(\theta_i; \lambda, \overline{x}) = -\lambda p_N(\theta_i; \lambda, \overline{x}) - H(\theta_i)x, \ i = 1, \dots, N, \\ p_N(0; \lambda, \overline{x}) = x_0. \end{cases} \tag{5.3}$$

Similarly to (5.2) we introduce the matrix $\overline{A}_N(\lambda)$ given by

$$\overline{A}_N(\lambda)\overline{x} = \overline{K} p_N(\cdot; \lambda, \overline{x}), \ \overline{x} \in \mathbb{C}^{n+1}. \tag{5.4}$$

Then we get that λ is an eigenvalue of $\mathcal{A}_N$ if and only if

$$\det\left(I_{n+1} - \overline{A}_N(\lambda)\right) = 0, \tag{5.5}$$

which is the discrete characteristic equation.

The order of convergence depends on the smoothness properties of the function $H : [0, a_\dagger] \to \mathbb{C}^n$ in (3.4). So we do the following assumption.

ASSUMPTION 5.1. *Assume that the ran$(\mathcal{H}) \subseteq C^\infty([0, a_\dagger])$, i.e., the function $H : [0, a_\dagger] \to \mathbb{C}^n$ in* (3.4)*, is of class C^∞ and, moreover, that derivatives of H satisfy the following bounds:*

$$\left\|H^{(k)}\right\|_1 := \sum_{i=1}^{n} \left\|h_i^{(k)}\right\|_\infty \leq \beta D^k,\ k = 0, 1, \ldots, \tag{5.6}$$

where $\|\cdot\|_\infty$ is the maximum norm and β and D are positive constants.

Let us denote by $B(\lambda, \rho)$ the closed ball in $\mathbb{C}$ of center λ and radius ρ. In the next lemma we give an upper bound of the error of the collocation polynomial $p_N(\cdot; \lambda, \overline{x})$ for (5.1).

LEMMA 5.2. *Let $\overline{\lambda} \in \mathbb{C}$ and $\rho_0 > 0$. Under Assumption* 5.1*, there exists $N_0 \in \mathbb{N}$ such that for $N \geq N_0$, $\lambda \in B(\overline{\lambda}, \rho_0)$, and $\overline{x} \in \mathbb{C}^{n+1}$ the collocation polynomial $p_N(\cdot; \lambda, \overline{x})$ for* (5.1) *and based on Ω_N exists and is unique. Moreover*

$$\|p_N(\cdot; \lambda, \overline{x}) - y(\cdot; \lambda, \overline{x})\|_\infty \leq \frac{C_0}{\sqrt{N}} \left(\frac{C_1}{N}\right)^N \|\overline{x}\|_\infty$$

holds, where $C_0 = C_0(\overline{\lambda}, \rho_0)$ and $C_1 = C_1(\overline{\lambda}, \rho_0)$ are constants independent of N.

Proof. Let $\mathcal{C}$ be the subspace $C([0, a_\dagger], \mathbb{C})$ of X endowed with the maximum norm $\|\cdot\|_\infty$, and let $\|\cdot\|$ be the induced operator norm. We introduce the integral Volterra operator $V : \mathcal{C} \to \mathcal{C}$ as

$$(V\varphi)(a) = -\int_0^a \varphi(s)ds,\ \varphi \in \mathcal{C},\ a \in [0, a_\dagger],$$

and the Lagrange interpolation operator $L_{N-1} : \mathcal{C} \to \mathcal{C}$ relevant to the nodes in $\Omega_N \setminus \{0\}$.

We can rewrite (5.1) and its collocation equations (5.3) as the equations on the subspace $\mathcal{C}$

$$y = x_0 + \lambda V y + VHx \tag{5.7}$$

and

$$p_N = x_0 + \lambda V L_{N-1} p_N + V L_{N-1} Hx, \tag{5.8}$$

respectively, where $y = y(\cdot; \lambda, \overline{x})$, $p_N = p_N(\cdot; \lambda, \overline{x})$, and x_0 denotes the function of $\mathcal{C}$ of constant value x_0. By subtracting (5.7) from (5.8) and setting $e_N = p_N - y$ we get for e_N the following equation:

$$e_N = \lambda V L_{N-1} e_N + V r_N, \tag{5.9}$$

where

$$r_N = (L_{N-1} - I)(\lambda y + Hx). \tag{5.10}$$

It can be directly shown that e_N is a solution of (5.9) if and only if $e_N = V\hat{e}_N$, where $\hat{e}_N$ satisfies

$$\hat{e}_N = \lambda L_{N-1} V \hat{e}_N + r_N.$$

Since $I - \lambda V$ is invertible and, by the known estimate for the Lagrange interpolation at Chebyshev nodes, $\|(L_{N-1} - I)V\| \to 0$ as $N \to \infty$, it follows that, for sufficiently large N (and independently of $\lambda \in B(\overline{\lambda}, \rho_0)$), the operator $I - \lambda L_{N-1} V$ is invertible, and

$$\left\|(I - \lambda L_{N-1} V)^{-1}\right\| \leq 2 \left\|(I - \lambda V)^{-1}\right\|$$

holds. Hence $e_N = V(I - \lambda L_{N-1} V)^{-1} r_N$ is the unique solution of (5.9), and

$$\|e_N\|_\infty \leq 2\|V\|\|(I - \lambda V)^{-1}\|\|r_N\|_\infty$$

holds.

The function r_N in (5.10) is the error of the Lagrange interpolation relevant to the nodes in Ω_N of $y' = -\lambda y - Hx$. By (5.6) we get

$$\left\|y^{(k+1)}\right\|_\infty \leq |\lambda| \left\|y^{(k)}\right\|_\infty + \beta D^k \left\|\overline{x}\right\|_\infty, \ k = 0, 1, 2, \ldots,$$

and

$$\|y\|_\infty \leq \alpha(\lambda) \left\|\overline{x}\right\|_\infty,$$

where

$$\alpha(\lambda) = \max\left\{1, e^{-\Re(\lambda) a_\dagger}\right\} (1 + \beta a_\dagger).$$

Hence we can conclude that

$$\left\|y^{(k)}\right\|_\infty \leq \alpha(\lambda) E(\lambda)^k \left\|\overline{x}\right\|_\infty, \ k = 0, 1, 2, \ldots,$$

where

$$E(\lambda) = \max\left\{|\lambda| + \frac{\beta}{\alpha(\lambda)}, D\right\}.$$

Thus

$$\|r_N\|_\infty \leq \alpha(\lambda) E(\lambda) \frac{(E(\lambda) a_\dagger)^N}{N!} \left\|\overline{x}\right\|_\infty,$$

and by the Stirling formula the thesis follows. □

The above result is stated under Assumption 5.1, but we note that, under less regularity hypotheses such as $H : [0, a_\dagger] \to \mathbb{C}^n$ in (3.4) of class C^m, $m \geq 1$, by standard interpolation results, we get that

$$\|p_N(\cdot; \lambda, \overline{x}) - y(\cdot; \lambda, \overline{x})\|_\infty \leq \frac{C(\lambda)}{N^m} \left\|\overline{x}\right\|_\infty,$$

where $C(\lambda)$ is independent of N.

In our remark at the end of section 4 we have seen that, when the functionals K_0, K are given by (3.5) and (3.6) and the integrals cannot be exactly computed, they are approximated by the Clenshaw–Curtis formula. So we assume that the integral functionals K_0 and K in (3.5) and (3.6) are replaced by suitable functionals $K_{0,N}$ and K_N, respectively (with $K_{0,N} = K_0$ and $K_N = K$ when the integrals are exactly computed). Moreover we set

$$\overline{K}_N\varphi = (K_{0,N}\varphi, (K_N\varphi)^T)^T \in \mathbb{C}^{n+1} \tag{5.11}$$

and do the following assumption that is fulfilled by Clenshaw–Curtis formulas.

ASSUMPTION 5.3. *We assume that the sequence* $\left\{\overline{K}_N\right\}$ *of functionals in* (5.11) *satisfies*

$$\sup_{N\in\mathbb{N}} \left\|\overline{K}_N\right\| < +\infty,$$

where

$$\left\|\overline{K}_N\right\| = \sup_{\varphi\in C([0,a_\dagger],\mathbb{C})} \frac{\left\|\overline{K}_N\varphi\right\|_\infty}{\|\varphi\|_\infty}.$$

Now let us set the following notation relevant to the continuous and discrete characteristic equations (3.12) and (5.5), respectively:

$$d(\lambda) = \det\left(I_{n+1} - \overline{A}(\lambda)\right),$$
$$d_N(\lambda) = \det\left(I_{n+1} - \overline{A}_N(\lambda)\right),$$

with $\overline{A}(\lambda)$ and $\overline{A}_N(\lambda)$ defined in (3.11) and (5.4), respectively. The next two lemmas are necessary steps to state the final theorem of convergence.

LEMMA 5.4. *Let* $\overline{\lambda} \in \mathbb{C}$ *and* $\rho_0 > 0$. *Under Assumptions* 5.1 *and* 5.3, *there exists* $N_0 \in \mathbb{N}$ *such that for* $N \geq N_0$ *and* $\lambda \in B(\overline{\lambda}, \rho_0)$ *we have*

$$|d_N(\lambda) - d(\lambda)| \leq C_2\left(\varepsilon_N + \frac{1}{\sqrt{N}}\left(\frac{C_1}{N}\right)^N\right),$$

where

$$\varepsilon_N = \sup_{\substack{\lambda\in B(\overline{\lambda},\rho_0)\\ \overline{x}\in\mathbb{C}^{n+1}}} \frac{\left\|\overline{K}_N\left(y(\cdot;\lambda,\overline{x})\right) - \overline{K}\left(y(\cdot;\lambda,\overline{x})\right)\right\|_\infty}{\|\overline{x}\|_\infty},$$

$C_1 = C_1(\overline{\lambda}, \rho_0)$ *as defined in Lemma* 5.2, *and* $C_2 = C_2(\overline{\lambda}, \rho_0)$ *is a constant independent of* N.

Proof. By Lemma 5.2 we have

$$\begin{aligned}\left\|\overline{A}_N(\lambda)\overline{x} - \overline{A}(\lambda)\overline{x}\right\|_\infty &\leq \left\|\overline{K}_N\left(p_N(\cdot;\lambda,\overline{x})\right) - \overline{K}_N\left(y(\cdot;\lambda,\overline{x})\right)\right\|_\infty \\ &\quad + \left\|\overline{K}_N\left(y(\cdot;\lambda,\overline{x})\right) - \overline{K}\left(y(\cdot;\lambda,\overline{x})\right)\right\|_\infty \\ &\leq \left(\sup_{N\in\mathbb{N}}\|\overline{K}_N\|\frac{C_0}{\sqrt{N}}\left(\frac{C_1}{N}\right)^N + \varepsilon_N\right)\|\overline{x}\|_\infty\end{aligned}$$

for $N \geq N_0$ and $\lambda \in B(\overline{\lambda}, \rho_0)$.

Since the derivative of $\det(A)$ with respect to $A \in \mathbb{C}^{(n+1)(n+1)}$ is continuous, the assertion follows. ∎

LEMMA 5.5. *If the function $d(\lambda)$ has a zero $\overline{\lambda} \in \mathbb{C}$ with multiplicity ν, then there exists $\rho_1 > 0$ such that for $\lambda \in B(\overline{\lambda}, \rho_1) \setminus \{\overline{\lambda}\}$*

$$|d(\lambda)| > C_3 \left\|\lambda - \overline{\lambda}\right\|^{\nu},$$

where $C_3 = C_3(\overline{\lambda}, \rho_1)$.

Proof. By expanding $d(\lambda)$ around $\overline{\lambda}$ and taking into account the multiplicity ν we obtain

$$d(\lambda) = \frac{d^{(\nu)}(\overline{\lambda})}{\nu!}(\lambda - \overline{\lambda})^{\nu} + O\left(\left\|\lambda - \overline{\lambda}\right\|^{\nu+1}\right),$$

with $d^{(\nu)}(\overline{\lambda}) \neq 0$. The result follows. ∎

We are now able to state the convergence result.

THEOREM 5.6 (convergence). *Under Assumptions* 5.1 *and* 5.3, *if the function $d(\lambda)$ has a zero $\overline{\lambda} \in \mathbb{C}$ with multiplicity ν and if*

$$\left(\frac{C_2}{C_3}\right)^{1/\nu} \left(\varepsilon_N + \frac{1}{\sqrt{N}}\left(\frac{C_1}{N}\right)^{N}\right)^{1/\nu} \leq \rho_1$$

for $N \geq N_0$, where C_3 and ρ_1 are defined in Lemma 5.5 *and N_0, C_1, and C_2 are defined in Lemma* 5.4, *with $\rho_0 = \rho_1$, then the function $d_N(\lambda)$ has ν zeros λ_i, $i = 1, \ldots, \nu$, (taking into account multiplicities) such that*

$$\max_{1 \leq i \leq \nu} |\overline{\lambda} - \lambda_i| \leq \left(\frac{C_2}{C_3}\right)^{1/\nu} \left(\varepsilon_N + \frac{1}{\sqrt{N}}\left(\frac{C_1}{N}\right)^{N}\right)^{1/\nu}. \tag{5.12}$$

Proof. Let

$$\rho_N = \left(\frac{C_2}{C_3}\right)^{1/\nu} \left(\varepsilon_N + \frac{1}{\sqrt{N}}\left(\frac{C_1}{N}\right)^{N}\right)^{1/\nu}.$$

Since $\rho_N \leq \rho_1 = \rho_0$, Lemmas 5.4 and 5.5 yield

$$|d(\lambda)| > |d_N(\lambda) - d(\lambda)|$$

for $\left|\lambda - \overline{\lambda}\right| = \rho_N$. Hence, by Rouché's theorem [4, section 5, Theorem 3.8], d and d_N have exactly the same number of zeros counted with their multiplicities in $B(\overline{\lambda}, \rho_N)$. ∎

It is important to point out that the spectral accuracy of the collocation error stated in Lemma 5.2 under Assumption 5.1 is preserved in the behavior of the convergence error (5.12) in Theorem 5.6, if the functionals K_0, K are given by (3.5) and (3.6) and the integrals in (3.5) and (3.6) can be exactly computed, i.e., $\varepsilon_N = 0$. On the other hand, we also maintain the spectral accuracy in the convergence if we make the further assumptions that:

1. the kernels k_0 and k in (3.5) and (3.6) are of class C^{∞} and there exist constants $c, F > 0$ such that

$$\max_{i=0,1,\ldots,n} \left\|k_i^{(q)}\right\|_{\infty} \leq cF^q, \ q = 0, 1, 2, \ldots;$$

2. a simple interpolatory quadrature rule based on $N+1$ distinct nodes is used for the integrals in (3.5) and (3.6).

Finally, we observe that if the quadrature error is independent of N, i.e., $\varepsilon_N \leq$ TOL (e.g., machine precision), the error bound (5.12) in Theorem 5.6 is still valid with ε_N replaced by TOL, and thus spectral accuracy holds down to the tolerance TOL.

6. A case study: Logistic population growth. Here we consider the particular case of the Gurtin–MacCamy model endowed with a general form of the logistic mechanism designed to include the role of age structure. Namely, in system (2.1), (2.2), and (2.3) we adopt the following specific form of the vital rates:

$$\beta(a, S(t)) = R_0 \beta_0(a) e^{-S(t)}, \tag{6.1}$$

$$\mu(a, S(t)) = \mu_0(a) + m(a)\frac{S(t)}{1+S(t)}, \tag{6.2}$$

where we assume that the occupation of the niche is described through a single size

$$S(t) = \int_0^{a_\dagger} \gamma(a) p(a,t) da.$$

Moreover we assume the following specific form for the functions above:

$$\beta_0(a) = \frac{\alpha+1}{a_\dagger}, \tag{6.3}$$

$$\mu_0(a) = \frac{\alpha}{(a_\dagger - a)}, \tag{6.4}$$

$$m(a) = m_0 \chi_{[a_1,a_2]}(a), \tag{6.5}$$

$$\gamma(a) = \gamma_1 e^{-\delta_1 a} + \gamma_2 e^{\delta_2 (a - a_\dagger)}. \tag{6.6}$$

With the assumptions above we assume a simple intrinsic maternity function. We allow selection of the age interval $[a_1, a_2]$ and the intensity m_0 relative to the impact of the logistic effect on the survival probability, and, by the four parameters in $\gamma(a)$, we allow a weighted selection of adults and juveniles.

By these specific assumptions, as we have seen in (2.6), (2.7), and (2.8), a stationary solution is given by

$$p^*(a) = \frac{S^* \left(a_\dagger - a\right)^\alpha e^{-\frac{m_0 S^* M(a)}{1+S^*}}}{\int_0^{a_\dagger} \gamma(\sigma) \left(a_\dagger - \sigma\right)^\alpha e^{-\frac{m_0 S^* M(\sigma)}{1+S^*}} d\sigma}, \ a \in [0, a_\dagger],$$

where

$$M(a) = \begin{cases} 0 & \text{for } a < a_1, \\ a - a_1 & \text{for } a_1 \leq a \leq a_2, \\ a_2 - a_1 & \text{for } a > a_2, \end{cases}$$

and $S^* > 0$ is the (unique) positive solution to the equation

$$(\alpha+1) R_0 e^{-x} \int_0^1 (1-\sigma)^\alpha e^{-\frac{x M(a_\dagger \sigma)}{1+x}} d\sigma = 1, \tag{6.7}$$

which exists if and only if $R_0 > 1$. In fact the left-hand side of (6.7) is a decreasing function of x with value R_0 at $x = 0$ and vanishing as $x \to +\infty$.

When we linearize the model at such a stationary solution p^* we obtain (3.2) with the following specific kernels:

$$k_0(a) = R_0 \frac{\alpha+1}{a_\dagger}\left(1-\frac{a}{a_\dagger}\right)^\alpha e^{-S^*\left(1+\frac{m_0 M(a)}{1+S^*}\right)}\left(1-\int_0^{a_\dagger} p^*(s)ds\gamma(a)\right), \tag{6.8}$$

$$k(a) = \gamma(a)\left(1-\frac{a}{a_\dagger}\right)^\alpha e^{-\frac{m_0 S^* M(a)}{1+S^*}}, \tag{6.9}$$

and

$$H(a) = \frac{p^*(0)m_0\chi_{[a_1,a_2]}(a)}{(1+S^*)^2}. \tag{6.10}$$

The next section is devoted to describing some simulations showing stability and bifurcations.

7. Numerical exploration of the logistic model. In this section we apply the numerical method developed in this work to the logistic model presented in the previous section for the values reported in Table 7.1. In particular, we consider R_0 as a parameter of nominal value $R_0 = 1$, and then we increase its value above one analyzing the asymptotic behavior of the nontrivial equilibrium. This latter is computed by using a standard zero-finder routine as applied to (6.7). We observe that the functions (6.8), (6.9), and (6.10) are clearly piecewise defined because of the discontinuity of the function m in a_1 and a_2. Figure 7.1 shows part of the rightmost characteristic roots of the nominal system. Results are obtained with $N = 200$, and hence high accuracy is ensured as it turns out from the following convergence analysis.

The convergence behavior can be analyzed by comparing approximated and exact characteristic roots for increasing N. In the case of the rightmost one, i.e., $\lambda = 0$, convergence is so fast that we get accuracy to machine precision already with $N = 1$. This is due to the error constant C_1 in Lemma 5.2 and Theorem 5.6, which is proportional to $|\lambda| = 0$. Therefore, to better show the spectral accuracy typical of this approach, we follow the convergence of the first five rightmost roots (computed with $N = 200$ and assumed to be exact to the first 11 digits). Results are shown in Figure 7.2 for the nominal system for $N = 1, 2, \ldots, 50$, from which it can be seen that the convergence is faster for smaller $|\lambda|$.

The characteristic roots of the logistic model clearly depend on R_0. The variation of the spectrum is shown in Figure 7.3 (only for the nonnegative imaginary part, the negative one being symmetric) when R_0 increases from 1 to 20 with variable steps in order to capture the main features of this analysis. Moreover, for better clarity, only the first five righmost roots are considered, the behavior of the others being qualitatively similar. These roots are marked with $\circ$ for the nominal value $R_0 = 1$, while they are marked with $\bullet$ for greater values.

As R_0 is increased, the imaginary roots move from the right to the left, turning back to the right towards the imaginary axis. For the real root, i.e., the zero eigenvalue for $R_0 = 1$, when R_0 is increased, it moves leftward as the complex ones do. In the

TABLE 7.1
Values of the parameters in (6.1)–(6.6).

Parameter	R_0	α	a_1	a_2	$a_\dagger$	m	γ_1	δ_1	γ_2	δ_2
Value	1	1	0.6	0.85	1	0.3	0.2	0.5	7	4

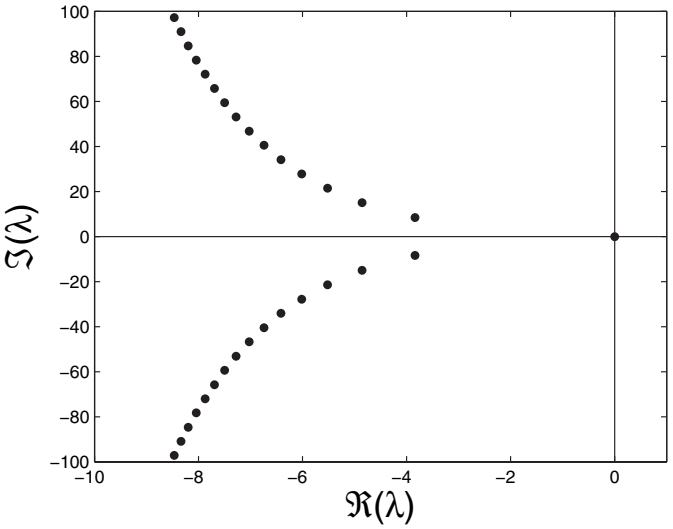


FIG. 7.1. *Partial characteristic spectrum for the logistic model with* $R_0 = 1$ *(*$N = 200$*).*

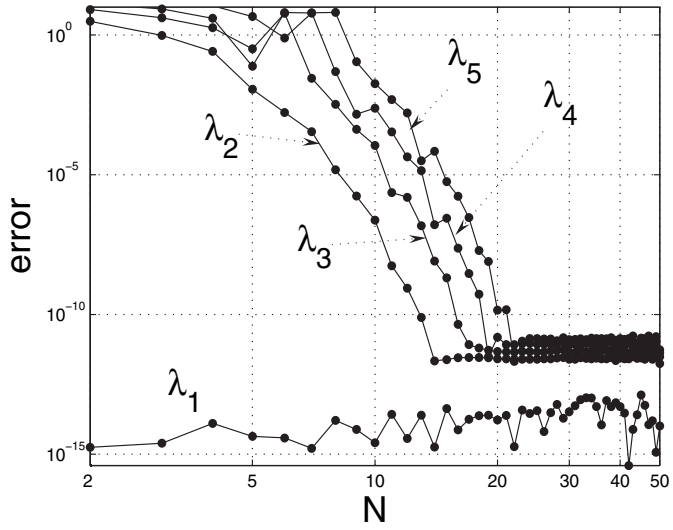


FIG. 7.2. *Absolute error analysis with respect to the first five rightmost roots of the logistic model with* $R_0 = 1$ *for varying* N*.*

meanwhile, a second real root is coming rightward from $-\infty$. At $R_0 = R_{0,1} \simeq 1.493$, these two real roots collapse and then split into a complex-conjugate pair. This latter moves rightward to the imaginary axis for increasing R_0 above $R_{0,1}$ as all the other complex branches. In particular, when $R_0 = R_{0,2} \simeq 11.903$, a Hopf bifurcation occurs.

This analysis can be performed similarly for other models by choosing different functions (6.8), (6.9), and (6.10), i.e., basically for different choices of the vital rates and of the sizes. In such a way, this numerical method represents a fast and reliable tool to investigate the dynamics of the Gurtin–MacCamy model as treated here and without the need of computing the solution.

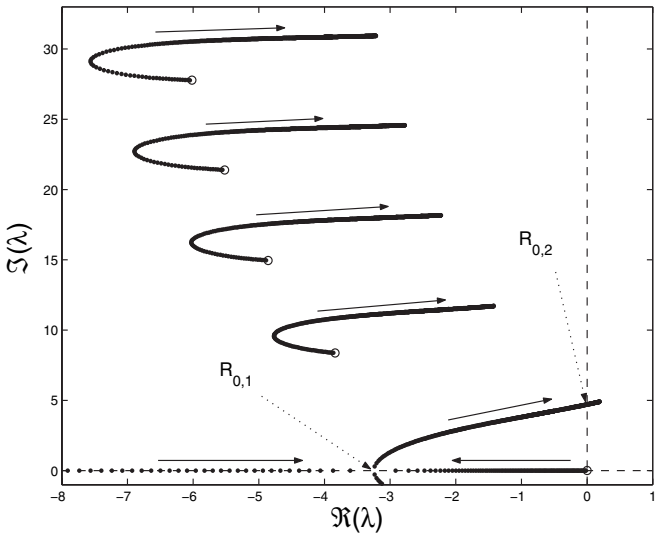


FIG. 7.3. *Some roots of the logistic model when* $R_0 = 1$ *(*$\circ$*, nominal value) and* $1 < R_0 \leq 20$ *(*$\bullet$*) computed with* $N = 20$.

REFERENCES

[1] D. BREDA, C. CUSULIN, M. IANNELLI, S. MASET, AND R. VERMIGLIO, *Stability analysis of age-structured population equations by pseudospectral differencing methods*, J. Math. Biol., published online DOI: 10.1007/s00285-006-0064-4 (2006).

[2] D. BREDA, S. MASET, AND R. VERMIGLIO, *Computing the characteristic roots for delay differential equations*, IMA J. Numer. Anal., 24 (2004), pp. 1–19.

[3] D. BREDA, S. MASET, AND R. VERMIGLIO, *Pseudospectral differencing methods for characteristic roots of delay differential equations*, SIAM J. Sci. Comput., 27 (2005), pp. 482–495.

[4] J. B. CONWAY, *Functions of One Complex Variable*, 2nd ed., Grad. Texts in Math., Springer-Verlag, New York, 1978.

[5] J. M. CUSHING AND J. LI, *On Ebenman's model for the dynamics of a population with competing juveniles and adults*, Bull. Math. Biol., 51 (1989), pp. 687–713.

[6] J. M. CUSHING AND J. LI, *Juveniles versus adults competition*, J. Math. Biol., 29 (1991), pp. 457–473.

[7] J. M. CUSHING AND J. LI, *Intra-specific competition and density dependent juvenile growth*, Bull. Math. Biol., 54 (1992), pp. 503–519.

[8] O. DIEKMANN, R. M. NISBET, W. S. C. GURNEY, AND F. VAN DEN BOSCH, *Simple mathematical models for cannibalism: A critique and a new approach*, Math. Biosci., 78 (1986), pp. 21–46.

[9] O. DIEKMANN, S. A. VAN GILS, S. M. VERDUYN LUNEL, AND H. O. WALTHER, *Delay Equations - Functional, Complex and Nonlinear Analysis*, Appl. Math. Sci. 110, Springer-Verlag, New York, 1995.

[10] K. ENGELBORGHS AND D. ROOSE, *On stability of LMS methods and characteristic roots of delay differential equations*, SIAM J. Numer. Anal., 40 (2002), pp. 629–650.

[11] M. E. GURTIN AND R. C. MACCAMY, *Non-linear age-dependent population dynamics*, Arch. Ration. Mech. Anal., 54 (1974), pp. 281–300.

[12] S. M. HENSON, *A continuous, age-structured insect population model*, J. Math. Biol., 39 (1999), pp. 217–243.

[13] T. KOSTOVA, *Numerical solutions of a hyperbolic differential-integral equation*, Comput. Math. Appl., 15 (1988), pp. 427–436.

[14] T. KOSTOVA AND M. MARCHEVA, *Numerical solutions to the Gurtin-MacCamy equation*, Math. Balkanica, 3 (1989), pp. 264–277.

[15] L. N. TREFETHEN, *Spectral Methods in MATLAB*, Software Environ. Tools, SIAM, Philadelphia, 2000.

SIAM J. NUMER. ANAL.
Vol. 46, No. 2, pp. 996–1011
© 2008 Society for Industrial and Applied Mathematics

ON STABILITY OF STAGGERED SCHEMES*

AMY L. BAUER†, RAPHAËL LOUBÈRE‡, AND BURTON WENDROFF§

Abstract. This paper investigates the theoretical stability bound of a Lagrangian staggered scheme used to solve hydrodynamics equations. We present the two-dimensional (2D) wave equation as a possible model for this study and, by using the numerical radius of the amplification matrix, we prove that the family of schemes defined with two time-centering parameters is limited by a non-classical stability bound limit defined with an analytical curve. We further show that 2D numerical experiments agree with this theoretical result.

Key words. staggered schemes, stability, wave equations, Euler equations

AMS subject classifications. 65N12, 65N22, 65M12, 65J10, 65J15

DOI. 10.1137/060660151

1. Introduction.

The concept of stability was first introduced in the seminal paper [1] of Courant, Friedrichs, and Lewy in 1928 where they discussed finite-difference methods for solving partial differential equations. Amazingly in the very same paper they studied one of the first staggered numerical schemes for the one-dimensional (1D) wave equation $v_t = cw_x, w_t = cv_x$, which was further developed in 1967 by Richtmyer and Morton in [6] and cast in its "modern" form as:

$$\frac{v_j^{n+1} - v_j^n}{\Delta t} = c\,\frac{w_{j+1/2}^n - w_{j-1/2}^n}{\Delta x}, \quad \frac{w_{j-1/2}^{n+1} - w_{j-1/2}^n}{\Delta t} = c\,\frac{v_j^{n+1} - v_{j-1}^{n+1}}{\Delta x}.$$

By considering these works as the genesis for numerical analysis and computational fluid dynamics, one would expect the stability of staggered numerical schemes to be at present well understood. However, this is far from being the case. Although staggered schemes have since been widely employed computationally, not much is known analytically about the stability of staggered schemes.

The use of staggered schemes also dates back to the inception of Los Alamos National Laboratory where the "calculation of certain time-dependent fluid flows played an important part in the wartime work of the laboratory" (preface to the first edition of [6]). Staggered schemes have frequently been used to solve the compressible fluid dynamics equations in their Lagrangian formulation. In this case, a staggered spatial placement of variables is used, where the position and velocity are defined at grid points and density, internal energy, and the pressure are defined at cell centers. Since this time, other Lagrangian hydrodynamics staggered numerical codes have been de-

*Received by the editors May 17, 2006; accepted for publication (in revised form) October 11, 2007; published electronically March 5, 2008. This work was performed under the auspices of the U.S. Department of Energy at Los Alamos National Laboratory, under contract W-7405-ENG-36. The authors acknowledge the partial support of the DOE/ASCR Program in the Applied Mathematical Sciences and the Laboratory Directed Research and Development program (LDRD). The authors also acknowledge the partial support of DOE's Accelerated Strategic Computing Initiative (ASCI).
http://www.siam.org/journals/sinum/46-2/66015.html

†Theoretical Division, T-14, Los Alamos National Laboratory MS-B262, Los Alamos, NM 87545 (albauer@lanl.gov).

‡Institut de Mathematiques et CNRS (UMR-5219), Université Paul-Sabatier Toulouse 3, 31062 Toulouse Cedex 9, France (loubere@mip.ups-tlse.fr).

§Theoretical Division, T-7, Los Alamos National Laboratory MS-B284, Los Alamos, NM 87545 (bbw@lanl.gov).

veloped around the globe, but again little advancement has been made towards a deeper understanding and analytical proof of the stability of such schemes.

Some years ago, Shashkov proposed a simple problem for the purpose of testing the stability of Lagrangian hydrodynamics codes. The initial data are given as zero velocity, constant pressure, and constant density. One then computes the total kinetic energy, which should be zero for all time. Caramana recently tried this problem with his two-dimensional (2D) staggered grid Lagrangian code and discovered that it was unstable for certain mesh ratios that "folklore" indicated should be otherwise—the instability manifested itself as an explosive growth of kinetic energy. The surprise here is not that a 1D symmetry-preserving 2D code could be unstable for 1D data; the surprise lies in the contradiction of the folklore.

The precise numerical scheme we are referring to and which is the focus of our investigation is the staggered-grid predictor-corrector compatible Lagrangian hydrodynamics scheme. We refer the reader to [3], [4] for the details of this scheme which we omit here. Briefly, there are two parameters α and β such that, if $\alpha = \beta = 1/2$, the scheme resembles a time-centered (Crank–Nicolson) scheme, while if $\alpha = \beta = 1$, it looks like backward Euler. For the Courant–Friedrichs–Lewy (CFL) condition CFL $= 1$, the scheme is stable for $\alpha = \beta = 1/2$. However, for $\alpha = 1$, $\beta = 1/2$ the scheme is unstable for CFL > 0.71, indicating that, as α increases, the CFL limit decreases. Bauer et al. then created a 1D problem by using the same data; they seed their 1D staggered-grid predictor-corrector compatible hydrodynamics Lagrangian code with a small random perturbation of the pressure. Moreover, they created the multidimension version of the problem and ran their 2D and Caramana's 3D code as well. They observed the same phenomenon in 1D, 2D, and 3D [3]. The stability limit was conjectured to be CFL $= 1/(2\sqrt{\alpha\beta})$, and numerical tests using the full nonlinear equations show that this is true provided $\alpha \geq 1/2$ and $\beta \geq 1/2$ [3].

We prove the above conjecture in 2D, which contains 1D as a special case. Of course, there is no possibility of doing this for the full nonlinear problem—the Euler equations. Ultimately, one applies a von Neumann analysis to the linearized system. Bauer et al. [3] almost succeeded in doing this theoretical analysis in 1D, resorting to numerical sampling only at the very end of the proof. We have succeeded in 2D, by using the numerical radius of the amplification matrix as a tool, an idea apparently first applied in [2]. This proof is the goal of this paper, which is articulated as follows: First we present the Lagrangian coordinates and justify the use of the wave equation as a possible linear model for the Euler equations in 1D and 2D; second, we introduce the notation; third, we show that the staggered implicit scheme is unconditionally stable for $\alpha \geq 1/2$, $\beta \geq 1/2$; fourth, we show that the predictor-corrector staggered scheme is stable for $\alpha \geq 1/2$, $\beta \geq 1/2$, and CFL $\leq 1/(2\sqrt{\alpha\beta})$. In a paragraph we show on a specific 1D example that the schemes for $\alpha < 1/2$ or $\beta < 1/2$ cannot be stable as some Fourier components are amplified. Finally, we show 2D numerical results for the wave and Euler equations using a compatible Lagrangian hydrodynamics code.

2. Lagrangian coordinates and wave equation models.

2.1. 1D. The hydrodynamic equations in 1D Lagrangian coordinates, expressed in mass variables m, are

$$\frac{du}{dt} = -\frac{\partial p}{\partial m}, \qquad \frac{d\tau}{dt} = \frac{\partial u}{\partial m}, \qquad \frac{de}{dt} = -p\frac{\partial u}{\partial m},$$

where u is the velocity, τ is the specific volume, e is the specific internal energy, and $p = p(\tau, e)$ is the pressure.

Total energy E and entropy S are conserved:

$$\begin{aligned}\frac{dE}{dt} &= \frac{1}{2}\frac{du^2}{dt} + \frac{de}{dt} \\ &= -u\frac{\partial p}{\partial m} - p\frac{\partial u}{\partial m} = -\frac{\partial pu}{\partial m},\end{aligned}$$

and

$$T\,\frac{dS}{dt} = \frac{de}{dt} + p\frac{d\tau}{dt} = 0.$$

Therefore for smooth flows the exact differential equations for u, p are

$$\begin{aligned}\frac{du}{dt} &= -\frac{\partial p}{\partial m}, \\ \frac{dp}{dt} &= \frac{dp}{d\tau}\frac{d\tau}{dt} + \frac{dp}{de}\frac{de}{dt} = \left(\frac{dp}{d\tau} - p\frac{dp}{de}\right)\frac{\partial u}{\partial m},\end{aligned}$$

and taking the coefficients p, $\frac{dp}{d\tau}$, and $\frac{dp}{de}$ to be constant yields a linear system with constant coefficients. Now assume that $\frac{dp}{d\tau} < 0$ and $\frac{dp}{de} > 0$. As

$$c = \sqrt{\left|\frac{dp}{d\tau}\right| + p\frac{dp}{de}}$$

is the sound speed, the linearization of the system with constant coefficients is obtained by freezing the sound speed; that is, c is assumed to be constant in space and time. Consequently, the resulting system is just the wave equation

$$\begin{aligned}\frac{du}{dt} &= -\frac{\partial p}{\partial m}, \\ \frac{dp}{dt} &= -c^2\frac{\partial u}{\partial m}.\end{aligned}$$

By using the new variables $u = u'/c$ and $dm/c = -da$ and dropping the prime notation, this linearized system can be written as

$$\frac{du}{dt} = \frac{\partial p}{\partial a}, \tag{2.1}$$

$$\frac{dp}{dt} = \frac{\partial u}{\partial a}. \tag{2.2}$$

2.2. 2D. The same linearized reduction can be performed for Lagrangian coordinates in 2D, but it is not quite as straightforward. In 2D the Euler equation system

is

$$\begin{aligned}
\rho\frac{du}{dt} &= -\frac{\partial p}{\partial x},\\
\rho\frac{dv}{dt} &= -\frac{\partial p}{\partial y},\\
\rho\frac{d\tau}{dt} &= \frac{\partial u}{\partial x}+\frac{\partial v}{\partial y},\\
\rho\frac{de}{dt} &= -p\left(\frac{\partial u}{\partial x}+\frac{\partial v}{\partial y}\right),\\
\frac{dx}{dt} &= u,\\
\frac{dy}{dt} &= v.
\end{aligned}$$

By following [7], [8] we make the change of variables:

$$\begin{aligned}
x &= x(a,b,t),\\
y &= y(a,b,t),
\end{aligned}$$

with the choice

$$\begin{aligned}
a &= x(a,b,0),\\
b &= y(a,b,0).
\end{aligned}$$

The Jacobian matrix of the transformation $(x,y)\leftrightarrow(a,b)$ is given by

$$\mathbf{J}=\begin{pmatrix} \frac{\partial x}{\partial a} & \frac{\partial x}{\partial b}\\ \frac{\partial y}{\partial a} & \frac{\partial y}{\partial b}\end{pmatrix},$$

the determinant of which is the Jacobian

$$J=\frac{\partial x}{\partial a}\frac{\partial y}{\partial b}-\frac{\partial y}{\partial a}\frac{\partial x}{\partial b}.$$

For the two velocity equations we now have

$$\begin{aligned}
J\rho\frac{du}{dt} &= \frac{\partial y}{\partial a}\frac{\partial p}{\partial b}-\frac{\partial y}{\partial b}\frac{\partial p}{\partial a},\\
J\rho\frac{dv}{dt} &= \frac{\partial x}{\partial b}\frac{\partial p}{\partial a}-\frac{\partial x}{\partial a}\frac{\partial p}{\partial b}.
\end{aligned}$$

But then, by starting the linearization by freezing the entries of the Jacobian, we obtain

$$\begin{aligned}
J\rho\frac{d}{dt}\left(u\frac{\partial x}{\partial a}+v\frac{\partial y}{\partial a}\right) &= -J\frac{\partial p}{\partial a},\\
J\rho\frac{d}{dt}\left(u\frac{\partial x}{\partial b}+v\frac{\partial y}{\partial b}\right) &= -J\frac{\partial p}{\partial b}.
\end{aligned}$$

Thus, by letting

$$\begin{aligned}
u' &= u\frac{\partial x}{\partial a}+v\frac{\partial y}{\partial a},\\
v' &= u\frac{\partial x}{\partial b}+v\frac{\partial y}{\partial b},
\end{aligned}$$

we obtain

$$\rho\frac{du'}{dt} = -\frac{\partial p}{\partial a},$$
$$\rho\frac{dv'}{dt} = -\frac{\partial p}{\partial b}.$$

We next need to evaluate $\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y}$ in the energy equation. First,

$$Ju = \frac{\partial y}{\partial b}u' - \frac{\partial y}{\partial a}v',$$
$$-Jv = \frac{\partial x}{\partial b}u' - \frac{\partial x}{\partial a}v'.$$

Then taking the derivative of u and v yields

$$J^2\frac{\partial u}{\partial x} = \frac{\partial y}{\partial b}\left(\frac{\partial y}{\partial b}\frac{\partial u'}{\partial a} - \frac{\partial y}{\partial a}\frac{\partial u'}{\partial b}\right) - \frac{\partial y}{\partial a}\left(\frac{\partial y}{\partial b}\frac{\partial v'}{\partial a} - \frac{\partial y}{\partial a}\frac{\partial v'}{\partial b}\right),$$
$$J^2\frac{\partial v}{\partial y} = \frac{\partial x}{\partial b}\left(\frac{\partial x}{\partial b}\frac{\partial u'}{\partial a} - \frac{\partial x}{\partial a}\frac{\partial u'}{\partial b}\right) - \frac{\partial x}{\partial a}\left(\frac{\partial x}{\partial b}\frac{\partial v'}{\partial a} - \frac{\partial x}{\partial a}\frac{\partial v'}{\partial b}\right).$$

Summing the previous equations gives

$$J^2\frac{\partial u}{\partial x} + J^2\frac{\partial v}{\partial y} = \left(\left(\frac{\partial x}{\partial b}\right)^2 + \left(\frac{\partial y}{\partial b}\right)^2\right)\frac{\partial u'}{\partial a} - \left(\frac{\partial x}{\partial a}\frac{\partial x}{\partial b} + \frac{\partial y}{\partial a}\frac{\partial y}{\partial b}\right)\frac{\partial u'}{\partial b}$$
$$- \left(\frac{\partial x}{\partial a}\frac{\partial x}{\partial b} + \frac{\partial y}{\partial a}\frac{\partial y}{\partial b}\right)\frac{\partial v'}{\partial a} + \left(\left(\frac{\partial x}{\partial a}\right)^2 + \left(\frac{\partial y}{\partial a}\right)^2\right)\frac{\partial v'}{\partial b},$$

and in order to cancel the cross terms, we are forced to assume here that

$$\frac{\partial x}{\partial a}\frac{\partial x}{\partial b} + \frac{\partial y}{\partial a}\frac{\partial y}{\partial b} = 0, \tag{2.3}$$

so that

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = d_1\frac{\partial u'}{\partial a} + d_2\frac{\partial v'}{\partial b},$$
$$d_1 = \frac{1}{J^2}\left(\left(\frac{\partial x}{\partial b}\right)^2 + \left(\frac{\partial y}{\partial b}\right)^2\right),$$
$$d_2 = \frac{1}{J^2}\left(\left(\frac{\partial x}{\partial a}\right)^2 + \left(\frac{\partial y}{\partial a}\right)^2\right).$$

So

$$\rho\frac{du'}{dt} = -\frac{\partial p}{\partial a},$$
$$\rho\frac{dv'}{dt} = -\frac{\partial p}{\partial b},$$
$$\rho\frac{d\tau}{dt} = d_1\frac{\partial u'}{\partial a} + d_2\frac{\partial v'}{\partial b},$$
$$\rho\frac{de}{dt} = -p\left(d_1\frac{\partial u'}{\partial a} + d_2\frac{\partial v'}{\partial b}\right).$$

Now ρ can be scaled into a, b, and as before, a simple substitution yields

$$\begin{aligned}\frac{du'}{dt} &= -\frac{\partial p}{\partial a},\\ \frac{dv'}{dt} &= -\frac{\partial p}{\partial b},\\ \frac{dp}{dt} &= -c^2\left(d_1\frac{\partial u'}{\partial a} + d_2\frac{\partial v'}{\partial b}\right).\end{aligned}$$

Finally, returning to (u, v, a, b), set

$$\begin{aligned} u &= -c\sqrt{d_1}u', & v &= -c\sqrt{d_2}v',\\ a &= c\sqrt{d_1}\,\widetilde{a}, & b &= c\sqrt{d_2}\,\widetilde{b},\end{aligned}$$

to obtain the 2D wave equation as a model:

$$\frac{du}{dt} = \frac{\partial p}{\partial \widetilde{a}}, \qquad \frac{dv}{dt} = \frac{\partial p}{\partial \widetilde{b}}, \qquad \frac{dp}{dt} = \frac{\partial u}{\partial \widetilde{a}} + \frac{\partial v}{\partial \widetilde{b}}. \tag{2.4}$$

However, this formulation critically depends on (2.3), so ultimately we present this as only a potential model.[1]

3. Notation. This section defines the notation that will be used throughout the rest of this paper.

The standard inner product on a complex vector space of dimension N is $\langle \mathbf{f}, \mathbf{g}\rangle = \sum_{i=1}^{N} f_i\,\bar{g}_i$ for two complex vectors $\mathbf{f} = (f_i)_{i=1,\dots,N}$ and $\mathbf{g} = (g_i)_{i=1,\dots,N}$.

Time. We will assume equal time steps; that is, the temporal interval $[0, T > 0]$ is discretized into equal intervals $[t_n, t_{n+1}]$, with $t_{n+1} = t_n + \Delta t$.

Space. In 2D we define uniform rectangles with vertices $x_{i,j}$, $y_{i,j}$, where $x_{i+1,j} = x_{i,j} + \Delta x$ and $y_{i,j+1} = y_{i,j} + \Delta y$. Nodal quantities are indexed by (i, j), while cell quantities are indexed by $(i + \frac{1}{2}, j + \frac{1}{2})$.

Discretization. The ratio of time step to space step in each dimension is given by $\lambda_x = \frac{\Delta t}{\Delta x}$, $\lambda_y = \frac{\Delta t}{\Delta y}$.

Hence, any quantity P at some point (i, j) at time t_n is represented as $P^n_{i,j}$. Similarly, any cell-based quantity C at time t_{n+1} is represented as $C^{n+1}_{i+\frac{1}{2},j+\frac{1}{2}}$.

For any variable w defined at two time levels $t_{n+1} > t_n$ on a point or in a cell, we define its interpolated value at an intermediate time $n + \kappa$ as:

$$w^{n+\kappa} = \kappa w^{n+1} + (1-\kappa)\,w^n, \qquad 0 \le \kappa \le 1.$$

We further define a vector $\mathbf{w} = (\mathbf{u}, \mathbf{v}, \mathbf{p})^t$ and $\mathbf{w}^{\alpha,\beta} = (\mathbf{u}^{n+\alpha}, \mathbf{v}^{n+\alpha}, \mathbf{p}^{n+\beta})^t$, where

$$\mathbf{u} = \{u_{i,j} : \ \forall\,(i,j)\}, \quad \mathbf{v} = \{v_{i,j} : \ \forall\,(i,j)\}, \quad \text{and} \quad \mathbf{p} = \left\{p_{i+\frac{1}{2},j+\frac{1}{2}} : \ \forall\,(i,j)\right\}.$$

[1] The Jacobian matrix $\mathbf{J}$ contains information relating to the volume, shape, and orientation of an element after the transformation $(a, b) \to (x, y)$. $\mathbf{\Lambda} = \mathbf{J}^{\mathrm{T}}\mathbf{J}$ is the associated metric tensor:

$$\mathbf{\Lambda} = \begin{pmatrix} \lambda_{11} & \lambda_{21} \\ \lambda_{12} & \lambda_{22} \end{pmatrix} = \begin{pmatrix} \left(\frac{\partial x}{\partial a}\right)^2 + \left(\frac{\partial x}{\partial b}\right)^2 & \frac{\partial x}{\partial a}\frac{\partial x}{\partial b} + \frac{\partial y}{\partial a}\frac{\partial y}{\partial b} \\ \frac{\partial x}{\partial a}\frac{\partial x}{\partial b} + \frac{\partial y}{\partial a}\frac{\partial y}{\partial b} & \left(\frac{\partial y}{\partial a}\right)^2 + \left(\frac{\partial y}{\partial b}\right)^2 \end{pmatrix},$$

$\mathbf{\Lambda}$ is a symmetric matrix, and moreover the constraint (2.3) states that $\lambda_{12} = 0$: The metric tensor $\mathbf{\Lambda}$ is diagonal, meaning that the implied system of coordinates is orthogonal.

Boundary conditions. We assume throughout that all sums exist and that Fourier transforms can be taken. Also, in the usual way, it is supposed that effectively there are no boundaries and therefore that indices can be shifted in a sum without changing its value. For example, for all integers k

$$\sum_i u_{i,j}\ p_{i+\frac{1}{2},j} = \sum_i u_{i+k,j}\ p_{i+\frac{1}{2}+k,j}.$$

Operators. There are two sets of data: nodal quantities and cell quantities. Consider the operator Q_x that maps cell data to nodal data given by

$$(Q_x p)_{i,j} = \frac{1}{2}\left(p_{i+\frac{1}{2},j+\frac{1}{2}} + p_{i+\frac{1}{2},j-\frac{1}{2}} - p_{i-\frac{1}{2},j+\frac{1}{2}} - p_{i-\frac{1}{2},j-\frac{1}{2}}\right).$$

By using the usual scalar product in both nodal and cell spaces, the adjoint Q_x^* maps nodal data to cell data and is defined by

$$\begin{aligned}\langle Q_x^* u, p\rangle = \langle u, Q_x p\rangle &= \frac{1}{2}\sum u_{i,j}\left(p_{i+\frac{1}{2},j+\frac{1}{2}} + p_{i+\frac{1}{2},j-\frac{1}{2}} - p_{i-\frac{1}{2},j+\frac{1}{2}} - p_{i-\frac{1}{2},j-\frac{1}{2}}\right)\\ &= \frac{1}{2}\sum p_{i+\frac{1}{2},j+\frac{1}{2}}\Big(u_{i,j} + u_{i,j+1} - u_{i+1,j} - u_{i+1,j+1}\Big);\end{aligned}$$

that is,

$$(Q_x^* u)_{i+\frac{1}{2},j+\frac{1}{2}} = \frac{1}{2}\Big(u_{i,j} + u_{i,j+1} - u_{i+1,j} - u_{i+1,j+1}\Big).$$

Similar definitions exist for Q_y and Q_y^*.

The notation for a staggered scheme on a rectangular grid for the 2D wave equation is depicted in Figure 3.1, where p is defined at the center of each cell and u and v are defined at nodes. For this scheme p has fractional indices, whereas u and v have integer indices. Moreover, one defines interpolated values at each midedge point; for example, for a midedge point defined with indices $i+\frac{1}{2}, j+1$ one has $p_{i+\frac{1}{2},j+1} = \frac{1}{2}(p_{i+\frac{1}{2},j+\frac{3}{2}} + p_{i+\frac{1}{2},j-\frac{1}{2}})$ and $u_{i+\frac{1}{2},j+1} = \frac{1}{2}(u_{i,j+1} + u_{i+1,j+1})$, whereas for a midedge point $i, j+\frac{1}{2}$ one has $p_{i,j+\frac{1}{2}} = \frac{1}{2}(p_{i+\frac{1}{2},j+\frac{1}{2}} + p_{i+\frac{3}{2},j+\frac{1}{2}})$ and $u_{i,j+\frac{1}{2}} = \frac{1}{2}(u_{i,j+1} + u_{i+1,j})$.

By using this notation, a fully implicit staggered scheme applied to system (2.4) can be written as

$$\begin{aligned} u_{i,j}^{n+1} &= u_{i,j}^n + \lambda_x\left(p_{i+\frac{1}{2},j}^{n+\alpha} - p_{i-\frac{1}{2},j}^{n+\alpha}\right),\\ v_{i,j}^{n+1} &= v_{i,j}^n + \lambda_y\left(p_{i,j+\frac{1}{2}}^{n+\alpha} - p_{i,j-\frac{1}{2}}^{n+\alpha}\right),\\ p_{i+\frac{1}{2},j+\frac{1}{2}}^{n+1} &= p_{i+\frac{1}{2},j+\frac{1}{2}}^n + \lambda_x\left(u_{i+1,j+\frac{1}{2}}^{n+\beta} - u_{i,j+\frac{1}{2}}^{n+\beta}\right) + \lambda_y\left(v_{i+\frac{1}{2},j+1}^{n+\beta} - v_{i+\frac{1}{2},j}^{n+\beta}\right).\end{aligned} \tag{3.1}$$

The difference operator is given by the matrix

$$\mathbf{M} = \begin{pmatrix} 0 & 0 & Q_x \\ 0 & 0 & Q_y \\ -Q_x^* & -Q_y^* & 0 \end{pmatrix},$$

together with

$$\mathbf{\Lambda} = \begin{pmatrix} \lambda_x & 0 & 0 \\ 0 & \lambda_y & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

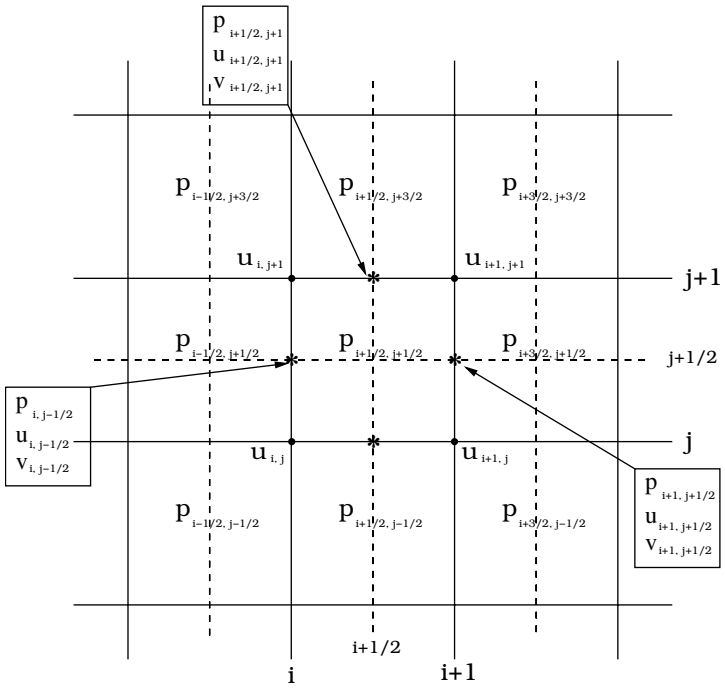


FIG. 3.1. *Rectangular staggered scheme. p is defined at cell centers with fractional indices: $p_{i+\frac{1}{2},j+\frac{1}{2}}$, whereas u, v are defined at nodes with integer indices: $u_{i,j}$. Moreover u, v, p are interpolated at each midedge point $*$. For example, $p_{i+\frac{1}{2},j+1} = \frac{1}{2}(p_{i+\frac{1}{2},j+\frac{3}{2}} + p_{i+\frac{1}{2},j-\frac{1}{2}})$ and $u_{i+\frac{1}{2},j+1} = \frac{1}{2}(u_{i,j+1} + u_{i+1,j+1})$, and similarly for v.*

Thus, the implicit difference scheme (3.1) has the form

$$\mathbf{w}^{n+1} = \mathbf{w}^n + \mathbf{\Lambda M \Lambda w}^{\alpha,\beta}.$$

THEOREM 3.1. *The staggered implicit scheme is stable for any λ_x, λ_y if $\alpha \geq \frac{1}{2}$ and $\beta \geq \frac{1}{2}$.*

Proof. Throughout the proofs, all sums are taken over both indices i and j. Applying the energy method yields

$$H(\alpha,\beta) = \sum \left(u_{i,j}^{n+1} - u_{i,j}^n\right) u_{i,j}^{n+\beta} + \sum \left(v_{i,j}^{n+1} - v_{i,j}^n\right) v_{i,j}^{n+\beta} + \sum \left(p_{i+\frac{1}{2},j+\frac{1}{2}}^{n+1} - p_{i+\frac{1}{2},j+\frac{1}{2}}^n\right) p_{i+\frac{1}{2},j+\frac{1}{2}}^{n+\alpha} = 0.$$

This follows from the fact that

$$\mathbf{M}^* = -\mathbf{M}$$

and therefore for real $\mathbf{w}$

$$\langle \mathbf{Mw}, \mathbf{w} \rangle = 0\,.$$

Now let us define $U(\beta), V(\beta), P(\alpha)$ such that $H(\alpha,\beta) = U(\beta) + V(\beta) + P(\alpha)$ and

$$\begin{aligned}
U(\beta) &= \sum \left(u_{i,j}^{n+1} - u_{i,j}^{n}\right) u_{i,j}^{n+\beta} \\
&= \beta \sum \left(u_{i,j}^{n+1}\right)^2 - (1-\beta) \sum \left(u_{i,j}^{n}\right)^2 - (2\beta - 1) \sum \left(u_{i,j}^{n+1} u_{i,j}^{n}\right), \\
V(\beta) &= \sum \left(v_{i,j}^{n+1} - v_{i,j}^{n}\right) v_{i,j}^{n+\beta} \\
&= \beta \sum \left(v_{i,j}^{n+1}\right)^2 - (1-\beta) \sum \left(v_{i,j}^{n}\right)^2 - (2\beta - 1) \sum \left(v_{i,j}^{n+1} v_{i,j}^{n}\right), \\
P(\alpha) &= \sum \left(p_{i+\frac{1}{2},j+\frac{1}{2}}^{n+1} - p_{i+\frac{1}{2},j+\frac{1}{2}}^{n}\right) p_{i+\frac{1}{2},j+\frac{1}{2}}^{n+\alpha} \\
&= \alpha \sum \left(p_{i+\frac{1}{2},j+\frac{1}{2}}^{n+1}\right)^2 - (1-\alpha) \sum \left(p_{i+\frac{1}{2},j+\frac{1}{2}}^{n}\right)^2 \\
&\quad - (2\alpha - 1) \sum \left(p_{i+\frac{1}{2},j+\frac{1}{2}}^{n+1} p_{i+\frac{1}{2},j+\frac{1}{2}}^{n}\right).
\end{aligned}$$

We detail the steps in the proof by using U and comment that the steps in the proof for V and P are similarly obtained.

Differentiating U with respect to β gives

$$U'(\beta) = \sum \left(u_{i,j}^{n+1}\right)^2 + \sum \left(u_{i,j}^{n}\right)^2 - 2 \sum \left(u_{i,j}^{n+1} u_{i,j}^{n}\right).$$

Then by the Schwarz inequality

$$U'(\beta) > 0$$

($U' = 0$ only in the trivial case that the initial data are constant). Since

$$U\left(\frac{1}{2}\right) = \frac{1}{2}\left(\sum \left(u_{i,j}^{n+1}\right)^2 - \sum \left(u_{i,j}^{n}\right)^2\right),$$

it follows that for $\beta \geq \frac{1}{2}$

$$U(\beta) \geq \frac{1}{2}\left(\sum \left(u_{i,j}^{n+1}\right)^2 - \sum \left(u_{i,j}^{n}\right)^2\right),$$

while for $\beta < \frac{1}{2}$

$$U(\beta) < \frac{1}{2}\left(\sum \left(u_{i,j}^{n+1}\right)^2 - \sum \left(u_{i,j}^{n}\right)^2\right).$$

By applying the same reasoning to V and P we see that for $\beta \geq \frac{1}{2}$ and $\alpha \geq \frac{1}{2}$

$$\begin{aligned}
\sum \left(\left(u_{i,j}^{n+1}\right)^2 - \left(u_{i,j}^{n}\right)^2\right) + \sum \left(\left(v_{i,j}^{n+1}\right)^2 - \left(v_{i,j}^{n}\right)^2\right) \\
+ \sum \left(\left(p_{i+\frac{1}{2},j+\frac{1}{2}}^{n+1}\right)^2 - \left(p_{i+\frac{1}{2},j+\frac{1}{2}}^{n}\right)^2\right) \leq 0;
\end{aligned}$$

that is, we have stability because

$$\begin{aligned}
\sum \left(u_{i,j}^{n+1}\right)^2 + \sum \left(v_{i,j}^{n+1}\right)^2 + \sum \left(p_{i+\frac{1}{2},j+\frac{1}{2}}^{n+1}\right)^2 \\
\leq \sum \left(u_{i,j}^{n}\right)^2 + \sum \left(v_{i,j}^{n}\right)^2 + \sum \left(p_{i+\frac{1}{2},j+\frac{1}{2}}^{n}\right)^2. \qquad \square
\end{aligned}$$

4. Rectangular predictor-corrector scheme. The predictor-corrector scheme in 2D [4] for the wave equation can now be described.

Predictor step.

$$\begin{aligned}
\widetilde{u}_{i,j}^{n+1} &= u_{i,j}^{n} + \lambda_x \left(Q_x p^n\right)_{i,j}, \\
\widetilde{v}_{i,j}^{n+1} &= v_{i,j}^{n} + \lambda_y \left(Q_y p^n\right)_{i,j}, \\
\widetilde{p}_{i+\frac{1}{2},j+\frac{1}{2}}^{n+1} &= p_{i+\frac{1}{2},j+\frac{1}{2}}^{n} - \lambda_x \left(Q_x^* u^{n+\beta}\right)_{i+\frac{1}{2},j+\frac{1}{2}} - \lambda_y \left(Q_y^* v^{n+\beta}\right)_{i+\frac{1}{2},j+\frac{1}{2}};
\end{aligned}$$

recall that $w^{n+\beta} = \beta \widetilde{w}^{\,n+1} + (1-\beta) w^n$ for $w = u, v$.

Corrector step.

$$\begin{aligned}
u_{i,j}^{n+1} &= u_{i,j}^{n} + \lambda_x \left(Q_x p^{n+\alpha}\right)_{i,j}, \\
v_{i,j}^{n+1} &= v_{i,j}^{n} + \lambda_y \left(Q_y p^{n+\alpha}\right)_{i,j}, \\
p_{i+\frac{1}{2},j+\frac{1}{2}}^{n+1} &= p_{i+\frac{1}{2},j+\frac{1}{2}}^{n} - \lambda_x \left(Q_x^* u^{n+\beta}\right)_{i+\frac{1}{2},j+\frac{1}{2}} - \lambda_y \left(Q_y^* v^{n+\beta}\right)_{i+\frac{1}{2},j+\frac{1}{2}};
\end{aligned}$$

recall that $p^{n+\alpha} = \alpha \widetilde{p}^{\,n+1} + (1-\alpha) p^n$ and now $w^{n+\beta} = \beta w^{n+1} + (1-\beta) w^n$ for $w = u, v$.

By substituting the equation resulting from the predictor step into the corrector step, we obtain the following difference scheme:

$$\mathbf{w}^{n+1} = \mathbf{S}\mathbf{w}^n.$$

Rather than write this out in terms of the difference operators, we immediately move to the Fourier transforms of the variables and operators.

The Fourier transform employs a substitution of variables; for instance, for p we have

$$p_{i+\frac{1}{2},j+\frac{1}{2}}^{n} \longmapsto p_0 e^{\theta(n\Delta t) + \mathrm{i}\left(2\delta\left((i+\frac{1}{2})\Delta x\right) + 2\gamma\left((j+\frac{1}{2})\Delta y\right)\right)},$$

where θ is complex and δ, γ are real. After factoring we obtain:

$$\begin{aligned}
\left(\widehat{Q}_x p\right)_{i,j} &= p_0 e^{\theta(n\Delta t) + \mathrm{i}(2\delta(i\Delta x) + 2\gamma(j\Delta y))} \frac{1}{2} \left(2\mathrm{i}\sin(\delta\Delta x)\right) \left(e^{\mathrm{i}\gamma\Delta y} + e^{-\mathrm{i}\gamma\Delta y}\right) \\
&= p_0 e^{\theta(n\Delta t) + \mathrm{i}(2\delta(i\Delta x) + 2\gamma(j\Delta y))} \left(2\mathrm{i}\sin(\delta\Delta x)\cos(\gamma\Delta y)\right),
\end{aligned}$$

with $\mathrm{i} = \sqrt{-1}$. Then if one denotes $\xi = \delta\Delta x$, and $\eta = \gamma\Delta y$, the dimensionless wave numbers in the x and y directions, respectively, on an uniform mesh, the Fourier transforms of the operator Q_x, Q_y are

$$\left(\widehat{Q}_x p\right)_{i,j} = p_0 e^{\theta(n\Delta t) + 2\mathrm{i}(i\xi + j\eta)} \left(2\mathrm{i}\sin\xi\cos\eta\right), \tag{4.1}$$

$$\left(\widehat{Q}_y p\right)_{i,j} = p_0 e^{\theta(n\Delta t) + 2\mathrm{i}(i\xi + j\eta)} \left(2\mathrm{i}\cos\xi\sin\eta\right). \tag{4.2}$$

The same formulas can be similarly obtained for Q^*.

By setting

$$\Phi_x = 2\lambda_x \sin\xi\cos\eta \quad \text{and} \quad \Phi_y = 2\lambda_y \sin\eta\cos\xi,$$

and dropping the hat notation, we obtain

$$\mathbf{S} = \begin{pmatrix} 1-\alpha\Phi_x^2 & -\alpha\Phi_x\Phi_y & \mathrm{i}\Phi_x\left(1-\alpha\beta(\Phi_x^2+\Phi_y^2)\right) \\ -\alpha\Phi_x\Phi_y & 1-\alpha\Phi_y^2 & \mathrm{i}\Phi_y\left(1-\alpha\beta\left(\Phi_x^2+\Phi_y^2\right)\right) \\ \mathrm{i}\Phi_x\left(1-\alpha\beta(\Phi_x^2+\Phi_y^2)\right) & \mathrm{i}\Phi_y\left(1-\alpha\beta(\Phi_x^2+\Phi_y^2)\right) & 1+\alpha\beta^2(\Phi_x^2+\Phi_y^2)^2-\beta\left(\Phi_x^2+\Phi_y^2\right) \end{pmatrix}.$$

The Lax–Richtmyer stability theory [6] tells us that we must show that the amplification factor $\mathbf{S}$ has uniformly bounded powers. First, observe that $\mathbf{S}\mathbf{S}^* \neq \mathbf{S}^*\mathbf{S}$, so it is not sufficient to show that the eigenvalues of $\mathbf{S}$ are less than 1. We shall show instead that the numerical radius of $\mathbf{S}$ is bounded by 1, since this implies that the powers of $\mathbf{S}$ are uniformly bounded by 2. A survey of results about the numerical radius is given in [5] along with a direct proof that, if the numerical radius is bounded by one, the powers are bounded by 2. We remark that the Kreiss matrix theorem implies a weaker result, namely, that the powers are bounded by a constant that depends on the size of the matrix [6]. The numerical radius was apparently first used as a stability analysis tool in [2], and here we use the basic ideas of that work.

We henceforth use the same symbols for grid functions and their Fourier transforms, realizing that this is an abuse of the notation, but hoping that no confusion will occur. The numerical radius of $\mathbf{S}$, denoted $R(\mathbf{S})$, is

$$R(\mathbf{S}) = \sup_{\mathbf{w}} |\langle \mathbf{S}\mathbf{w}, \mathbf{w}\rangle|, \quad \text{with} \quad \langle \mathbf{w}, \mathbf{w}\rangle = 1.$$

The matrix $\mathbf{S}$ can be split into real and imaginary parts as

$$\mathbf{S} = \mathbf{A} + \mathrm{i}\mathbf{B},$$

where

$$\mathbf{A} = \begin{pmatrix} 1-\alpha\Phi_x^2 & -\alpha\Phi_x\Phi_y & 0 \\ -\alpha\Phi_x\Phi_y & 1-\alpha\Phi_y^2 & 0 \\ 0 & 0 & 1+\alpha\beta^2(\Phi_x^2+\Phi_y^2)^2-\beta\left(\Phi_x^2+\Phi_y^2\right) \end{pmatrix},$$

$$\mathbf{B} = \left(1-\alpha\beta(\Phi_x^2+\Phi_y^2)\right)\begin{pmatrix} 0 & 0 & \Phi_x \\ 0 & 0 & \Phi_y \\ \Phi_x & \Phi_y & 0 \end{pmatrix}.$$

Now let

$$r = \langle \mathbf{A}\mathbf{w}, \mathbf{w}\rangle, \quad j = \langle \mathbf{B}\mathbf{w}, \mathbf{w}\rangle.$$

Since $\mathbf{A}$ and $\mathbf{B}$ are real and symmetric, r and j are real. Then

$$|\langle \mathbf{S}\mathbf{w}, \mathbf{w}\rangle|^2 = |\langle \mathbf{A}\mathbf{w}, \mathbf{w}\rangle + \mathrm{i}\langle \mathbf{B}\mathbf{w}, \mathbf{w}\rangle|^2 = |r+\mathrm{i}j|^2 = r^2+j^2.$$

THEOREM 4.1. *The* 2D *staggered rectangular scheme is stable if* $\alpha \geq \frac{1}{2}$, $\beta \geq \frac{1}{2}$, *and* $4\alpha\beta\max(\lambda_x^2, \lambda_y^2) \leq 1$.

Proof. Suppose that $\alpha \geq \frac{1}{2}$, $\beta \geq \frac{1}{2}$, and $4\alpha\beta\max(\lambda_x^2, \lambda_y^2) \leq 1$. Let

$$\overline{\mathrm{x}} = \sqrt{\alpha\beta}\,\Phi_x, \quad \overline{\mathrm{y}} = \sqrt{\alpha\beta}\,\Phi_y.$$

First note that

$$z^2 = \overline{\mathrm{x}}^2 + \overline{\mathrm{y}}^2 \leq 1,$$

since

$$\begin{aligned} z^2 = \overline{\mathrm{x}}^2 + \overline{\mathrm{y}}^2 &= \alpha\beta\left(\Phi_x^2 + \Phi_y^2\right) \\ &= \alpha\beta\left(4\lambda_x^2 \sin^2\xi\cos^2\eta + 4\lambda_y^2 \sin^2\eta\cos^2\xi\right) \\ &\le \underbrace{4\alpha\beta\max\left(\lambda_x^2, \lambda_y^2\right)}_{\le 1}\left[(\cos\eta\sin\xi)^2 + (\cos\xi\sin\eta)^2\right] \\ &\le (\cos\eta\sin\xi)^2 + (\cos\xi\sin\eta)^2 \le 1. \end{aligned}$$

Recall that $|\mathbf{u}|^2 + |\mathbf{v}|^2 + |\mathbf{p}|^2 = 1$, and note that

$$|\overline{\mathrm{x}}u + \overline{\mathrm{y}}v|^2 \le \left(1 - |p|^2\right)\left(\overline{\mathrm{x}}^2 + \overline{\mathrm{y}}^2\right), \tag{4.3}$$

since

$$\begin{aligned} |\overline{\mathrm{x}}u + \overline{\mathrm{y}}v|^2 = \left(\overline{\mathrm{x}}u + \overline{\mathrm{y}}v\right)\left(\overline{\mathrm{x}}\overline{u} + \overline{\mathrm{y}}\overline{v}\right) &= \overline{\mathrm{x}}^2 u\overline{u} + \overline{\mathrm{y}}^2 v\overline{v} + \overline{\mathrm{x}}\overline{\mathrm{y}}\left(\overline{u}v + \overline{v}u\right) \\ &\le \overline{\mathrm{x}}^2 u\overline{u} + \overline{\mathrm{y}}^2 v\overline{v} + \overline{\mathrm{x}}^2|v|^2 + \overline{\mathrm{y}}^2|u|^2 \\ &= \left(|u|^2 + |v|^2\right)\left(\overline{\mathrm{x}}^2 + \overline{\mathrm{y}}^2\right) \\ &\le \left(1 - |p|^2\right)\left(\overline{\mathrm{x}}^2 + \overline{\mathrm{y}}^2\right). \end{aligned} \tag{4.4}$$

Also, by setting

$$\overline{\mathrm{a}} = 1/\alpha, \quad \overline{\mathrm{b}} = 1/\beta,$$

we have that $\overline{\mathrm{a}}, \overline{\mathrm{b}}$ are positive and bounded by 2. Then

$$\begin{aligned} r &= 1 - \overline{\mathrm{b}}|\overline{\mathrm{x}}u + \overline{\mathrm{y}}v|^2 - \overline{\mathrm{a}}\left(\overline{\mathrm{x}}^2 + \overline{\mathrm{y}}^2\right)|p|^2 + \overline{\mathrm{a}}\left(\overline{\mathrm{x}}^2 + \overline{\mathrm{y}}^2\right)^2|p|^2, \\ j &= 2\sqrt{\overline{\mathrm{a}}\overline{\mathrm{b}}}\,\mathcal{R}\left(\overline{p}\left(\overline{\mathrm{x}}u + \overline{\mathrm{y}}v\right)\right)\left(1 - \left(\overline{\mathrm{x}}^2 + \overline{\mathrm{y}}^2\right)\right), \end{aligned}$$

where $\mathcal{R}(q)$ is the real part of q, and

$$\begin{aligned} r^2 &= \left(1 - \overline{\mathrm{b}}|\overline{\mathrm{x}}u + \overline{\mathrm{y}}v|^2 - \overline{\mathrm{a}}z^2\left(1 - z^2\right)|p|^2\right)^2, \\ j^2 &\le 4\overline{\mathrm{a}}\overline{\mathrm{b}}\left(1 - z^2\right)^2|p|^2|\overline{\mathrm{x}}u + \overline{\mathrm{y}}v|^2. \end{aligned}$$

Now set

$$|\overline{\mathrm{x}}u + \overline{\mathrm{y}}v|^2 = z^2\left(1 - |p|^2\right)\gamma.$$

By (4.3), $0 \le \gamma \le 1$. We thus have that

$$\begin{aligned} r^2 &= \left(1 - \gamma\overline{\mathrm{b}}z^2\left(1 - |p|^2\right) - \overline{\mathrm{a}}z^2\left(1 - z^2\right)|p|^2\right)^2, \\ j^2 &\le 4\overline{\mathrm{a}}\left(\gamma\overline{\mathrm{b}}\right)z^2\left(1 - z^2\right)^2|p|^2\left(1 - |p|^2\right). \end{aligned}$$

Also note that, for all $0 \le p \le 1$,

$$|p|^2\left(1 - |p|^2\right) \le \frac{1}{4}.$$

Since $\gamma\overline{\mathrm{b}}$ has the same range as $\overline{\mathrm{b}}$, we can replace $\gamma\overline{\mathrm{b}}$ by $\overline{\mathrm{b}}$. What we need to show then is that

$$g\left(z, p, \overline{\mathrm{a}}, \overline{\mathrm{b}}\right) = \left[1 - \overline{\mathrm{b}}z^2\left(1 - |p|^2\right) - \overline{\mathrm{a}}z^2\left(1 - z^2\right)|p|^2\right]^2 + \overline{\mathrm{a}}\overline{\mathrm{b}}z^2\left(1 - z^2\right)^2 \le 1$$

for $0 \le \overline{\mathrm{a}}, \overline{\mathrm{b}} \le 2$, and $0 \le z, p \le 1$.

LEMMA 4.2. *The function* $g(z,p,\bar{\mathrm{a}},\bar{\mathrm{b}}) \le 1$ *for* $0 \le \bar{\mathrm{a}},\bar{\mathrm{b}} \le 2$, *and* $0 \le z,p \le 1$.
Proof: Let

$$q = |p|^2, \quad t = z^2.$$

Then

$$g = \left[1 - \bar{\mathrm{b}}t\,(1-q) - \bar{\mathrm{a}}t\,(1-t)\,q\right]^2 + \bar{\mathrm{a}}\bar{\mathrm{b}}t\,(1-t)^2\,.$$

We first show that $g \le 1$ on the boundary of the (t,q) domain. Consider the t boundary values:

$$g\,(t=0) = 1,$$

and since $\bar{\mathrm{b}}(1-q) \le 2$,

$$g\,(t=1) \le 1.$$

For the q boundary values:

$$\begin{aligned} g\,(q=0) - 1 = \bar{\mathrm{b}}t\left[-2 + \bar{\mathrm{b}}t + \bar{\mathrm{a}}\,(1-t)^2\right] &\le \bar{\mathrm{b}}t\Big[-2 + \bar{\mathrm{b}}t + \bar{\mathrm{a}}\,(1-t)\Big] \\ &\le \bar{\mathrm{b}}t\Big[-2 + 2\,(t+1-t)\Big] = 0, \end{aligned}$$

while

$$g\,(q=1) - 1 = \bar{\mathrm{a}}t\,(1-t)\left[-2 + \bar{\mathrm{a}}t\,(1-t) + \bar{\mathrm{b}}\,(1-t)\right] \le \left[-2 + \bar{\mathrm{a}}t + \bar{\mathrm{b}}\,(1-t)\right] \le 0.$$

To finish the proof, we need only to show that, for each fixed t, g (a quadratic function in q) does not have a maximum for $0 < q < 1$. This is achieved by showing that either $\frac{\partial g}{\partial q}(q=1) \ge 0$, in which case g is either monotone nondecreasing or has a minimum in $0 < q < 1$, or that $\frac{\partial g}{\partial q}(q=0) \le 0$ and $\frac{\partial g}{\partial q}(q=1) \le 0$, in which case g is monotone nonincreasing in $0 < q < 1$. We proceed as follows:

$$\frac{\partial g}{\partial q} = 2\kappa\big[1 - \bar{\mathrm{b}}t + \kappa q\big],$$

where

$$\kappa = \bar{\mathrm{b}}t - \bar{\mathrm{a}}t + \bar{\mathrm{a}}t^2.$$

Note that

$$1 - \bar{\mathrm{b}}t + \kappa = 1 - \bar{\mathrm{a}}t(1-t) \ge 1 - \bar{\mathrm{a}}\frac{1}{4} \ge \frac{1}{2}. \tag{4.5}$$

- Case $\bar{\mathrm{a}} \le \bar{\mathrm{b}}$ $(\alpha \ge \beta)$:[2] In this case $\kappa \ge 0$, so by (4.5)

$$\frac{\partial g}{\partial q}\,(q=1) = 2\kappa\left(1 - \bar{\mathrm{b}}t + \kappa\right) \ge 2\kappa\,\frac{1}{2} \ge 0.$$

- Case $\bar{\mathrm{a}} \ge \bar{\mathrm{b}}$ $(\alpha \le \beta)$: Then $\kappa \le 0$ if $t \le t_0 = 1 - \frac{\bar{\mathrm{b}}}{\bar{\mathrm{a}}}$, and $\kappa \ge 0$ if $t \ge t_0$.
  - If $t \ge t_0$ $(k \ge 0)$. Then as in the case above, $\frac{\partial g}{\partial q}(q=1) > 0$.
  - If $t \le t_0$ $(k \le 0)$. Then by (4.5), both $\frac{\partial g}{\partial q}(q=1) < 0$ and $\frac{\partial g}{\partial q}(q=0) < 0$.

Thus ends the proof of the lemma and completes the proof of the theorem. ∎

[2] This case is the most important, since for the compatible Lagrangian hydrodynamics scheme applied to the Euler equations one must necessarily have $\beta = \frac{1}{2}$ for the scheme to conserve total energy.

5. Instability for $\alpha < 1/2$ or $\beta < 1/2$. Both the fully implicit and the predictor-corrector schemes are unstable if either $\alpha < 1/2$ or $\beta < 1/2$, for any mesh ratios. We just need to show this in 1D for the special case of $p = u$, for then the schemes are being applied to $\frac{du}{dt} = \frac{\partial u}{\partial x}$, with one set of difference equations used to advance u at half-integer indexes $i + 1/2$ and a different set at integer indexes i. Instability occurs if some Fourier component is amplified. So by setting $\Phi_y = 0$ in the amplification factor $\mathbf{S}$ for the predictor-corrector factor, we need only to show first that $|1 - \alpha\Phi_x^2 + \mathrm{i}\Phi_x(1 - \alpha\beta(\Phi_x^2))| > 1$ if $\alpha < 1/2$. This is clearly the case if $|\Phi_x|$ is positive and sufficiently small since

$$\left|1 - \alpha\Phi_x^2 + \mathrm{i}\Phi_x\left(1 - \alpha\beta(\Phi_x^2)\right)\right|^2 = 1 + \Phi_x^2 \underbrace{(1 - 2\alpha)}_{>0} + \underbrace{\Phi_x^4\ \alpha^2\beta^2\ \left(\frac{1}{\beta^2} + \Phi_x^2 - \frac{2}{\alpha\beta}\right)}_{\sim o(\Phi_x^4)}.$$

A similar argument works for the second row of $\mathbf{S}$ if $\beta < 1/2$. The fully implicit scheme yields to the same kind of analysis; we omit the details.

6. Numerical results. We solve the simple problem already described in [3] where, on a unit domain using a quadrangular mesh of 100×100, velocity is set to zero, density and sound speed are set to unity by using a perfect equation of state with γ, the ratio of specific heats, equal to $5/3$ and reflective boundary conditions.

Since for this problem no velocity should develop, simulations are run for a very large number of time cycles, typically 10^5, and with varying values for the CFL number $\lambda = \lambda_x = \lambda_y$ and parameters α, β. A sensitive gauge of instability is to keep track of the total kinetic energy $K^\lambda(t^n) = \frac{1}{2}\sum[(u_{i,j}^n)^2 + (v_{i,j}^n)^2]$ for a given CFL number λ at a given time t_n. Since the density and domain size are scaled to unity, this number should remain at the square of machine precision, about 10^{-28}–10^{-30} in our case. For an unstable scheme, it is observed that $K^\lambda(t^n)$ grows by several orders of magnitude long before the 10^5 cycle limit is reached. In this way, one can accurately identify the stability boundary in the CFL number and α space for a given β.

2D wave equations. We solve the 2D wave equations with the rectangular predictor-corrector scheme. From Theorem 4.1 we have that

$$4\alpha\beta \max\left(\lambda_x^2, \lambda_y^2\right) \le 1 \quad \Longrightarrow \quad \lambda \le \frac{1}{2\sqrt{\alpha\beta}},$$

provided that $\lambda = \lambda_x = \lambda_y$. Thus, the maximum CFL number achievable, denoted $\mathrm{CFL}_{\max}$, therefore depends on α and β and is described by the following function:

$$\mathrm{CFL}_{\max} = \mathrm{CFL}_{\max}(\alpha, \beta) = \frac{1}{2\sqrt{\alpha\beta}}. \tag{6.1}$$

In Figure 6.1, we plot the maximal CFL as a function of α for a sample of five β values: $\beta = \{0.5, 0.6, 0.75, 0.9, 1.0\}$. The maximum CFL numbers obtained numerically with the code for a pair of α, β are superimposed; $+$ for $\beta = 0.5$, $*$ for $\beta = 0.6$, $\square$ for $\beta = 0.75$, $\blacksquare$ for $\beta = 0.9$, and $\circ$ for $\beta = 1.0$. As expected they perfectly match the theoretical CFL limit (6.1) given by Theorem 4.1.

2D Euler equations. We present the results obtained with our 2D Lagrangian predictor-corrector code for the Euler equations by using Lagrangian coordinates. In this context (see [3], for example) the conservation of total energy is ensured provided that $\beta = \frac{1}{2}$. In Figure 6.2, we plot seven data points (symbol $+$) obtained by the code which represent the maximum CFL number that can be used for a given α for which

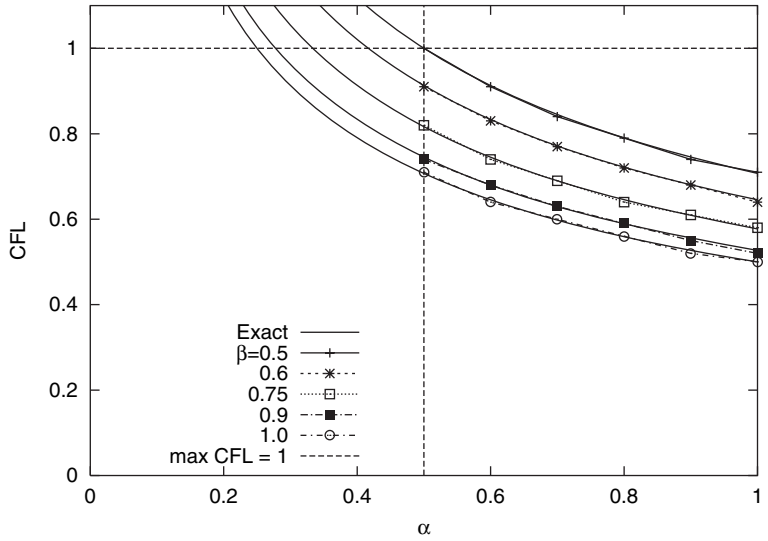


FIG. 6.1. 2D *wave equations—maximum CFL number* λ *as a function of* α *(*β *fixed) for the kinetic energy to remain on the order of machine precision for the test case. The vertical dashed line is the* $\alpha = 1/2$ *limit, and the horizontal one is the CFL* $= 1$ *limit. The theorem predicts the continuous thick lines, and the code produces the data:* $+$ *for* $\beta = 0.5$, $*$ *for* $\beta = 0.6$, $\square$ *for* $\beta = 0.75$, $\blacksquare$ *for* $\beta = 0.9$, *and* $\circ$ *for* $\beta = 1.0$. *Any scheme defined by a value* $\alpha \geq 1/2$ *and* $\beta \geq 1/2$ *is stable with the following CFL number:* $\lambda \leq 1/(2\sqrt{\alpha\beta})$.

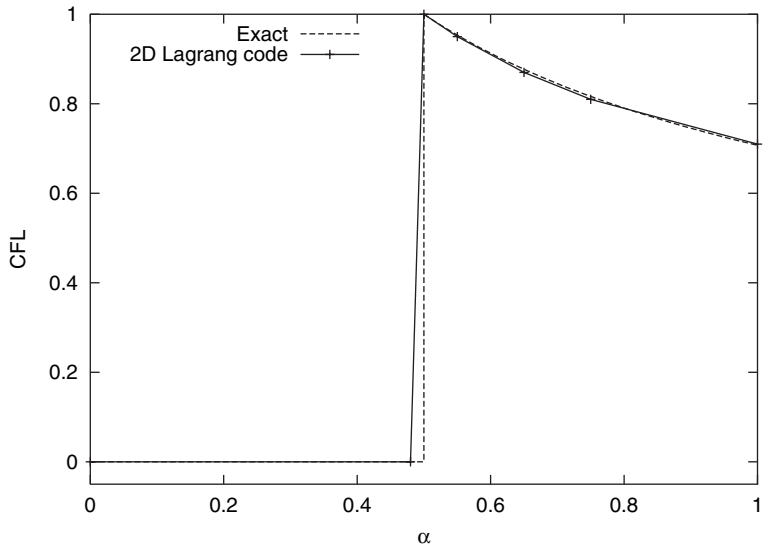


FIG. 6.2. *Hydrodynamics equations—maximum CFL number* λ *as a function of* α *(*$\beta = 1/2$*) for the kinetic energy to remain on the order of machine precision for the test case. The theorem predicts the continuous line; the* 2*D code produces the dashed line. Any scheme defined by a value* $\alpha \geq 1/2$ *is stable with the CFL number* $\lambda \leq 1/\sqrt{2\alpha}$ *and is unstable otherwise.*

numerical stability is preserved. The exact function CFL (dashed line) predicted by (6.1) is superimposed.[3] We can easily see that any scheme defined by a value $\alpha \geq 1/2$ is stable with the CFL number $\lambda \leq 1/\sqrt{2\alpha}$ and is unstable otherwise. We comment that for $\alpha = 1/2$ one exactly reaches the maximum CFL number $\lambda = 1$. The scheme with $\alpha = \beta = 1/2$ is the only staggered predictor-corrector scheme that reaches the maximum CFL number. Therefore the choice of using this scheme in several Lagrangian simulation codes, at least for this "optimal stability reason," seems justified. The results of our simulations perfectly match the predicted curve; this is also true of our 1D and 3D codes.

7. Conclusion. We have proposed the 2D wave equation as a linear constant coefficient model for 2D Lagrangian hydrodynamics, and we have established a sufficient stability condition for the standard staggered-grid numerical scheme. The 1D case is concurrently obtained as a special case of the 2D analysis.

This stability boundary has been tested with the 2D compatible Lagrangian hydrodynamics scheme for the wave equations and the Euler equations in Lagrangian coordinates. The numerical results fit the theoretical curve, showing that

- the general belief that for such numerical schemes "the more implicit, the more stable" is not always true,
- the classical choice made by several generations of Lagrangian code developers to use the $\alpha = \beta = 1/2$ scheme is reasonable as it leads to optimal stability results (at least for the artificial test case presented).

For the Euler equations these numerical results have been obtained in 1D, 2D, and 3D.

Acknowledgment. The authors thank Misha Shashkov for many fruitful discussions.

REFERENCES

[1] R. Courant, K. O. Friedrichs, and H. Lewy, *Uber die partiellen Differenzengleichungen der mathematischen Physik*, Math. Ann., 100 (1928), pp. 32–74.

[2] P. D. Lax and B. Wendroff, *Difference schemes for hyperbolic equations with high order of accuracy*, Comm. Pure Appl. Math., XVII (1964), pp. 381–398.

[3] A. L. Bauer, D. E. Burton, E. J. Caramana, R. Loubère, M. J. Shashkov, and P. P. Whalen, *The internal consistency, stability, and accuracy of the discrete, compatible formulation of Lagrangian hydrodynamics*, J. Comput. Phys., 218 (2006), pp. 572–593.

[4] E. J. Caramana, D. E. Burton, M. J. Shashkov, and P. P. Whalen, *The construction of compatible hydrodynamics algorithms utilizing conservation of total energy*, J. Comput. Phys., 146 (1998), pp. 227–262.

[5] M. Goldberg and E. Tadmor, *On the numerical radius and its applications*, Linear Algebra Appl., 42 (1982), pp. 263–284.

[6] R. D. Richtmyer and K. W. Morton, *Difference Methods for Initial Value Problems*, 2nd ed., Interscience, 1967.

[7] W. H. Hui, P. Y. Li, and Z. W. Li, *A unified coordinate system for solving the two-dimensional Euler equations*, J. Comput. Phys., 153 (1999), pp. 596–637.

[8] W. H. Hui, *The unified coordinate system in computational fluid dynamics*, Commun. Comput. Phys., 2 (2007), pp. 577–610.

[3] In this case CFL $= 1/2\sqrt{\alpha\beta} = 1/\sqrt{2\alpha}$, because $\beta = 1/2$.

SIAM J. NUMER. ANAL.
Vol. 46, No. 2, pp. 1012–1039
© 2008 Society for Industrial and Applied Mathematics

WELL-BALANCED HIGH ORDER EXTENSIONS OF GODUNOV'S METHOD FOR SEMILINEAR BALANCE LAWS*

MANUEL CASTRO†, JOSÉ M. GALLARDO†, JUAN A. LÓPEZ-GARCÍA†, AND CARLOS PARÉS†

Abstract. This paper is concerned with the development of well-balanced high order numerical schemes for systems of balance laws with a linear flux function, whose coefficients may be variable. First, well-balanced first order numerical schemes are obtained based on the use of exact solvers of Riemann problems that include both the flux and the source terms. Godunov's methods so obtained are extended to higher order schemes by using a technique of reconstruction of states. The main contribution of this paper is to introduce a reconstruction technique that preserves the well-balanced property of Godunov's methods. Some numerical experiments are presented to verify in practice the properties of the developed numerical schemes.

Key words. balance laws, well-balanced schemes, high order methods, finite volume methods, Godunov's methods

AMS subject classifications. Primary, 74S10, 65M06; Secondary, 35L60, 35L65

DOI. 10.1137/060674879

1. Introduction.

This paper is concerned with the numerical approximation of Cauchy problems for one-dimensional semilinear strictly hyperbolic systems of the form

$$W_t + A(\sigma)W_x = G(\sigma, W)\sigma_x, \quad x \in \mathbb{R},\ t > 0. \tag{1.1}$$

Here $W(x,t) \in \Omega \subset \mathbb{R}^N$ is the vector of unknowns; $\sigma(x)$ is a known function; $A(\sigma)$ is a smooth matrix-valued function; $G(\sigma, W)$ is also a smooth function from $\mathbb{R} \times \Omega$ to $\mathbb{R}^N$. The matrix $A(\sigma)$ is supposed to have N different real nonvanishing eigenvalues. Notice that systems of the form

$$W_t + A(x)W_x = G(x, W), \quad x \in \mathbb{R},\ t > 0,$$

can be considered as a particular case of (1.1).

The problems to be considered are thus systems of balance laws with a linear flux function, whose coefficients may be variable. Systems of this kind frequently arise in computational fluid dynamics, when models involving chemical reactions, external forces, or geometrical source terms are linearized; they also appear in relaxation systems (see [19] for some examples).

Besides the interest by itself of solving system (1.1), the development of numerical schemes with good properties for this kind of system may be helpful in the design of numerical methods for the more general case of a system of balance laws:

$$W_t + F(\sigma, W)_x = G(\sigma, W)\sigma_x, \quad x \in \mathbb{R},\ t > 0. \tag{1.2}$$

*Received by the editors November 14, 2006; accepted for publication (in revised form) October 24, 2007; published electronically March 5, 2008. This research was partially supported by Spanish Government Research projects BFM2003-07530-C02-02 and MTM2006-08075.
http://www.siam.org/journals/sinum/46-2/67487.html
†Departamento Análisis Matemático, Universidad de Málaga, 29071-Málaga, Spain (castro@anamat.cie.uma.es, gallardo@anamat.cie.uma.es, lopez@anamat.cie.uma.es, pares@anamat.cie.uma.es). The third author's research was supported by a FPU predoctoral grant (MEC).

In fact, the numerical difficulties due to the presence of a source term are similar in both cases, but considering a linear flux function makes easier the development and analysis of methods based on exact Riemann solvers. Nevertheless, the numerical schemes developed here can be extended to the more general case (1.2) by replacing, if necessary, the exact Riemann solver by an approximate one, provided that stationary solutions of the system can be exactly calculated.

A first numerical difficulty in solving (1.1) is related to stiff and singular source terms: In some problems the source term is concentrated over a very small region (even at a single point), or, due to the structure of the solutions, it acts only on a small region, where it can be very large. Appropriate fractional step methods, in which the problem is split into two subproblems (one related to the conservation law and the other one to the source term) can correctly handle these situations. Nevertheless, a second difficulty arises when solving steady-state solutions or solutions close to equilibrium. In this case, if the numerical source term does not balance the flux term in an adequate manner, the numerical solutions may exhibit nonphysical oscillations. In general, it is not easy to solve this difficulty with numerical methods based on a fractional step approach (see [19]).

The development and analysis of well-balanced numerical schemes for systems of balance laws is a very active front of research, and there is a huge amount of recent literature devoted to this topic. An important part of this literature is concerned with the approximation of shallow water systems with geometrical source terms. We refer to [5], [30], and the references therein for a review on the subject. In particular, the well-balanced properties of high order numerical schemes for shallow water systems have been studied in [6], [24], [25], [29], and [33].

The goal of this paper is to develop high order well-balanced numerical schemes for systems of the form (1.1). The idea is to consider finite volume methods based on the use of exact solvers of Riemann problems that include both the flux and the source terms. To do this, by following a technique introduced in [16] for the Euler equation in a nozzle, σ is interpreted as an artificial unknown of the system, whose value is fixed by the initial condition. Then (1.1) is rewritten under the form

$$\widetilde{W}_t + \widetilde{A}(\widetilde{W})\widetilde{W}_x = 0, \quad x \in \mathbb{R},\ t > 0, \tag{1.3}$$

where $\widetilde{W} = [W, \sigma]^T$. Once the system has been rewritten under this form, well-balanced first order numerical schemes are obtained by considering piecewise constant approximations of the solutions and applying Godunov's method. This approach has been also pursued in [1], [2], [7], [13], [14], [9], [15], [17].

When (1.1) is written under the form (1.3), a new linearly degenerate characteristic field is added to the system, whose corresponding eigenvalue is zero. Therefore, when one of the eigenvalues of $A(\sigma)$ vanishes, the system loses the strictly hyperbolic character: The problem becomes *resonant*, and the Riemann problems may not have a unique solution. This difficulty will not be considered here.

The organization of the article is as follows: In section 2 we consider the easier particular case

$$W_t + AW_x = CW\sigma_x, \quad x \in \mathbb{R},\ t > 0,$$

A and C being two given matrices. The solution of the Riemann problems of the system is discussed, and the expression of the corresponding Godunov's method is obtained, which is a particular case of path-conservative numerical scheme for (1.1), as

defined in [27]. In this reference, the general writing of a high order numerical scheme based on the use of a reconstruction operator and a first order path-conservative numerical scheme has also been presented. In section 3, we adapt this general methodology to problem (1.3) in order to obtain high order extensions of Godunov's methods.

Standard reconstruction operators do not preserve, in general, the well-balanced property of Godunov's methods. The main contribution of this paper is introduced in section 4: We present a technique to modify a standard reconstruction operator to solve this difficulty. The idea is to use, in the reconstruction procedure, the information given by the characteristic curves of the linearly degenerate field associated to the null eigenvalue. These curves, as will be seen, are related both to the stationary contact discontinuities and to the regular stationary solutions of (1.3).

In section 5 we consider first the extension of the numerical schemes developed to systems of the form

$$W_t + AW_x = G(W)\sigma_x, \quad x \in \mathbb{R},\ t > 0,$$

and, next, to the more general case (1.1).

Finally, in section 6 the numerical schemes are applied to some test cases.

2. Linear case: Godunov's method. We consider the system

$$W_t + AW_x = CW\sigma_x, \tag{2.1}$$

where $W \in \mathbb{R}^N$, $A, C \in \mathcal{M}_N(\mathbb{R})$, and $\sigma(x)$ is a known function. This system can be interpreted as a linear system of conservation laws with a source term.

Assume that the system is strictly hyperbolic, that is, the matrix A has N real distinct eigenvalues $\lambda_1 < \cdots < \lambda_N$, and let $R_1, \ldots, R_N$ be the associated eigenvectors.

Notice that systems of the form

$$W_t + AW_x = CW$$

can be considered as a particular case of (2.1), corresponding to the choice $\sigma(x) = x$.

If we add to system (2.1) the equation

$$\sigma_t = 0,$$

we obtain the following quasi-linear system in nonconservative form:

$$\widetilde{W}_t + \widetilde{\mathcal{A}}(\widetilde{W})\widetilde{W}_x = 0, \tag{2.2}$$

where

$$\widetilde{W} = \begin{bmatrix} W \\ \sigma \end{bmatrix}, \quad \widetilde{\mathcal{A}}(\widetilde{W}) = \begin{bmatrix} A & -CW \\ 0 & 0 \end{bmatrix}.$$

Assume that the eigenvalues λ_i of the matrix A satisfy $\lambda_i \neq 0,\ i = 1, \ldots, N$. Then, for every $\widetilde{W}$, $\widetilde{\mathcal{A}}(\widetilde{W})$ has $N+1$ real distinct eigenvalues $\lambda_1, \ldots, \lambda_N, \lambda_{N+1} = 0$, with associated eigenvectors $\widetilde{R}_1(\widetilde{W}), \ldots, \widetilde{R}_N(\widetilde{W}), \widetilde{R}_{N+1}(\widetilde{W})$ given by

$$\widetilde{R}_i(\widetilde{W}) = \begin{bmatrix} R_i \\ 0 \end{bmatrix}, \quad i = 1, \ldots, N; \qquad \widetilde{R}_{N+1}(\widetilde{W}) = \begin{bmatrix} A^{-1}CW \\ 1 \end{bmatrix}.$$

By assumption, the first N fields are linearly degenerate. Since λ_{N+1} is constant, the additional field is linearly degenerate as well. The integral curves of the characteristic fields are those of the ODE systems

$$\frac{d\widetilde{W}}{ds}(s) = \widetilde{R}_i(\widetilde{W}(s)), \quad i = 1, \ldots, N+1.$$

The solutions of these systems corresponding to initial conditions

$$\widetilde{W}(0) = \begin{bmatrix} W_i \\ \sigma_i \end{bmatrix}, \quad i = 1, \dots, N+1,$$

are, respectively,

$$\widetilde{W}(s) = \begin{bmatrix} sR_i + W_i \\ \sigma_i \end{bmatrix}, \quad i = 1, \dots, N,$$

and

$$\widetilde{W}(s) = \begin{bmatrix} e^{A^{-1}Cs} W_{N+1} \\ s + \sigma_{N+1} \end{bmatrix}, \quad i = N+1. \tag{2.3}$$

As all of the characteristic fields are linearly degenerate, the simple waves of the system are contact discontinuities linking two states belonging to the same integral curve of one of the characteristic fields. Moreover, these contact discontinuities propagate at speed λ_i if they are related to the ith characteristic field, for $i = 1, \dots, N$, or they are stationary if related to the $(N+1)$th characteristic field.

Let us consider now the Riemann problem

$$\begin{cases} \widetilde{W}_t + \widetilde{\mathcal{A}}(\widetilde{W})\widetilde{W}_x = 0, \\ \widetilde{W}(x,0) = \begin{cases} \widetilde{W}_L & \text{if } x < 0, \\ \widetilde{W}_R & \text{if } x > 0. \end{cases} \end{cases} \tag{2.4}$$

A general nonconservative theory of quasi-linear Riemann problems has been presented in [8]. We give hereafter a result of the existence and uniqueness of the solution for our particular case, whose constructive proof will play a central role in the development of the numerical schemes.

PROPOSITION 2.1. *For $[\![\sigma]\!] = \sigma_R - \sigma_L$ small enough, the Riemann problem* (2.4) *has a unique solution consisting of at most $N+2$ intermediate states linked by contact discontinuities.*

Proof. Let I be the maximum value of i for which $\lambda_i < 0$. If all of the eigenvalues are positive, we take $I = 0$, and if they are all negative, $I = N$.

Let us consider first the case $0 < I < N$. The structure of the solution must be as follows:

$$\widetilde{W}(x,t) = \begin{cases} \widetilde{W}_0 = \widetilde{W}_L & \text{if } \ x < \lambda_1 t, \\ \widetilde{W}_1 & \text{if } \ \lambda_1 t < x < \lambda_2 t, \\ \vdots & \\ \widetilde{W}_{I-1} & \text{if } \ \lambda_{I-1} t < x < \lambda_I t, \\ \widetilde{W}_I^- & \text{if } \ \lambda_I t < x < 0, \\ \widetilde{W}_I^+ & \text{if } \ 0 < x < \lambda_{I+1} t, \\ \widetilde{W}_{I+1} & \text{if } \ \lambda_{I+1} t < x < \lambda_{I+2} t, \\ \vdots & \\ \widetilde{W}_{N-1} & \text{if } \ \lambda_{N-1} t < x < \lambda_N t, \\ \widetilde{W}_N = \widetilde{W}_R & \text{if } \ \lambda_N t < x, \end{cases} \tag{2.5}$$

where the following apply:

1. The states $\widetilde{W}_i$, $i = 1, \dots, I-1$, and $\widetilde{W}_I^-$ must have the form

$$\widetilde{W}_i = \begin{bmatrix} W_L + \sum_{j=1}^{i} s_j R_j \\ \sigma_L \end{bmatrix}, \quad i = 1, \dots, I-1,$$
$$\widetilde{W}_I^- = \begin{bmatrix} W_I^- \\ \sigma_I^- \end{bmatrix} = \begin{bmatrix} W_L + \sum_{j=1}^{I} s_j R_j \\ \sigma_L \end{bmatrix}$$

for some $s_1, \dots, s_I \in \mathbb{R}$.

2. The states $\widetilde{W}_I^-$ and $\widetilde{W}_I^+$ must verify

$$\widetilde{W}_I^+ = \begin{bmatrix} W_I^+ \\ \sigma_I^+ \end{bmatrix} = \begin{bmatrix} e^{A^{-1}Cs_*} W_I^- \\ \sigma_I^- + s_* \end{bmatrix} = \begin{bmatrix} e^{A^{-1}Cs_*} \left(W_L + \sum_{j=1}^{I} s_j R_j \right) \\ \sigma_L + s_* \end{bmatrix}$$

for some $s_* \in \mathbb{R}$.

3. Finally, the states $\widetilde{W}_i$, $i = I+1, \dots, N$, must have the form

$$\widetilde{W}_i = \begin{bmatrix} W_I^+ + \sum_{j=I+1}^{i} s_j R_j \\ \sigma_I^+ \end{bmatrix}, \quad i = I+1, \dots, N,$$

for some $s_{I+1}, \dots, s_N \in \mathbb{R}$. In particular,

$$\widetilde{W}_N = \begin{bmatrix} W_I^+ + \sum_{j=I+1}^{N} s_j R_j \\ \sigma_I^+ \end{bmatrix} = \begin{bmatrix} W_R \\ \sigma_R \end{bmatrix},$$

and thus

$$s_* = \sigma_R - \sigma_L = [\![\sigma]\!],$$
$$W_I^+ = e^{A^{-1}C[\![\sigma]\!]} \left(W_L + \sum_{j=1}^{I} s_j R_j \right) = W_R - \sum_{j=I+1}^{N} s_j R_j.$$

Therefore, to find the solution of the Riemann problem, the values of $s_1, \dots, s_N$ have to be determined so that the following identity holds:

$$W_R - e^{A^{-1}C[\![\sigma]\!]} W_L = e^{A^{-1}C[\![\sigma]\!]} \sum_{j=1}^{I} s_j R_j + \sum_{j=I+1}^{N} s_j R_j. \tag{2.6}$$

Observe that (2.6) is an $N \times N$ linear system for the unknowns $s_1, \dots, s_N$. When $[\![\sigma]\!] = 0$ this system reduces to

$$W_R - W_L = \sum_{j=1}^{N} s_j R_j,$$

whose solution is given by the coordinates of $[\![W]\!] = W_R - W_L$ in the basis of eigenvectors R_j. Consequently, by continuity, system (2.6) has also a unique solution $s_1, \dots, s_N$ for small values of $[\![\sigma]\!]$.

The cases $I = 0$ and $I = N$ are similar. Notice that, in these situations, system (2.6) has always a unique solution: For $I = 0$ (respectively, $I = N$), $\{s_1, \dots, s_N\}$ are the coordinates of $W_R - e^{A^{-1}C[\![\sigma]\!]} W_L$ (respectively, $e^{-A^{-1}C[\![\sigma]\!]} W_R - W_L$) in the basis $\{R_1, \dots, R_N\}$. ∎

The hypothesis of smallness of $\sigma_R - \sigma_L$ cannot be removed in general: Consider, for instance, the particular case corresponding to

$$A = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 2 \\ 1 & 2 \end{bmatrix}.$$

Given $[\![\sigma]\!] = \sigma_R - \sigma_L$, the linear system (2.6) is written in this case as follows:

$$e^{A^{-1}C[\![\sigma]\!]}(W_L + s_1 R_1) = W_R + s_2 R_2,$$

where $R_1 = [1, 0]^T$ and $R_2 = [0, 1]^T$. A straightforward calculation shows that

$$e^{A^{-1}C[\![\sigma]\!]} = \begin{bmatrix} 2 - e^{[\![\sigma]\!]} & 2(1 - e^{[\![\sigma]\!]}) \\ e^{[\![\sigma]\!]} - 1 & 2e^{[\![\sigma]\!]} - 1 \end{bmatrix}.$$

Therefore, the linear system can be expressed as follows:

$$\begin{cases} (2 - e^{[\![\sigma]\!]})s_1 = w_1^R - 2(1 - e^{[\![\sigma]\!]})w_2^L - (2 - e^{[\![\sigma]\!]})w_1^L, \\ (e^{[\![\sigma]\!]} - 1)s_1 + s_2 = -(2e^{[\![\sigma]\!]} - 1)w_2^L - (e^{[\![\sigma]\!]} - 1)w_1^L, \end{cases}$$

where

$$W_L = \begin{bmatrix} w_1^L \\ w_2^L \end{bmatrix}, \quad W_R = \begin{bmatrix} w_1^R \\ w_2^R \end{bmatrix}.$$

Now this system has a unique solution if $[\![\sigma]\!] < \log(2)$. But when $[\![\sigma]\!] = \log(2)$, there is no solution if

$$w_1^R + 2w_2^L \neq 0.$$

Otherwise, it has infinitely many solutions.

However, in some particular cases the conclusion of Proposition 2.1 holds for arbitrary values of $[\![\sigma]\!]$.

COROLLARY 2.2. *The Riemann problem* (2.4) *has always a unique solution if one of the following conditions holds:*

(i) *All of the eigenvalues of A have the same sign.*

(ii) *The matrix C is diagonalizable and commutes with A.*

Proof. If (i) holds, then $I = 0$ or $I = N$ in the proof of Proposition 2.1, and (2.6) has always a unique solution. Notice that this is always the case for $N = 1$.

If (ii) holds, we have

$$Ae^{A^{-1}C[\![\sigma]\!]} = e^{A^{-1}C[\![\sigma]\!]}A,$$

and then

$$Ae^{A^{-1}C[\![\sigma]\!]}R_j = e^{A^{-1}C[\![\sigma]\!]}AR_j = \lambda_j e^{A^{-1}C[\![\sigma]\!]}R_j;$$

that is, $e^{A^{-1}C[\![\sigma]\!]}R_j$ is an eigenvector of A associated to λ_j. Therefore,

$$e^{A^{-1}C[\![\sigma]\!]}R_j = \beta_j R_j, \quad j = 1, \ldots, N,$$

for some $\beta_j \neq 0$. If α_j are the coordinates of $W_R - e^{A^{-1}C[\![\sigma]\!]}W_L$ in the basis composed by the eigenvectors R_j, that is,

$$W_R - e^{A^{-1}C[\![\sigma]\!]}W_L = \sum_{j=1}^{N} \alpha_j R_j,$$

the unique solution of (2.6) is given by

$$\begin{cases} s_j = \dfrac{\alpha_j}{\beta_j} & \text{for } \ j = 1, \dots, I, \\ s_j = \alpha_j & \text{for } \ j = I+1, \dots, N. \end{cases} \qquad \square$$

Remark 2.3. Notice that, in every case, the following equalities hold:

$$W_I^- - W_L = \sum_{j=1}^{I} s_j R_j, \quad W_R - W_I^+ = \sum_{j=I+1}^{N} s_j R_j.$$

Remark 2.4. The solutions of the Riemann problem found in Proposition 2.1 are not solutions in the distributional sense, as the product $CW\sigma_x$ doesn't make sense as a distribution when both W and σ are discontinuous. Nevertheless, the theory developed by Dal Maso, LeFloch, and Murat in [8] can be used to give a weaker sense to this nonconservative product as Borel measures. See also [18].

Once the Riemann problem has been analyzed, we study the application of Godunov's method to Cauchy's problem consisting of system (2.1) with initial condition $W(x,0) = W_0(x),\ x \in \mathbb{R}$, or equivalently, system (2.2) with initial condition $\widetilde{W}(x,0) = \widetilde{W}_0(x),\ x \in \mathbb{R}$, where

$$\widetilde{W}_0 = \begin{bmatrix} W_0 \\ \sigma \end{bmatrix}.$$

For the discretization of the system, computing cells $I_i = [x_{i-1/2}, x_{i+1/2}]$ are considered. For simplicity, we suppose that these cells have constant size Δx and that $x_{i+\frac{1}{2}} = i\Delta x$. Define $x_i = (i - 1/2)\Delta x$, the center of the cell I_i. Let Δt be the constant time step, and define $t^n = n\Delta t$.

We denote by $\widetilde{W}_i^n$ the approximation of the cell averages of the exact solution provided by the numerical scheme:

$$\widetilde{W}_i^n = \begin{bmatrix} W_i^n \\ \sigma_i^n \end{bmatrix} \cong \frac{1}{\Delta x} \int_{x_{i-1/2}}^{x_{i+1/2}} \widetilde{W}(x, t^n)\, dx.$$

In particular, at time $t = 0$ we consider

$$\widetilde{W}_i^0 = \begin{bmatrix} W_i^0 \\ \sigma_i \end{bmatrix},$$

where

$$W_i^0 = \frac{1}{\Delta x} \int_{x_{i-1/2}}^{x_{i+1/2}} W_0(x)\, dx, \quad \sigma_i = \frac{1}{\Delta x} \int_{x_{i-1/2}}^{x_{i+1/2}} \sigma(x)\, dx.$$

Suppose that the averages $\widetilde{W}_i^n$ at time $t = t^n$ are known. As is usual in Godunov's methods, to advance in time we consider the Riemann problems

$$\begin{cases} \widetilde{W}_t^{i+1/2} + \widetilde{\mathcal{A}}(\widetilde{W}^{i+1/2})\widetilde{W}_x^{i+1/2} = 0, \\ \widetilde{W}^{i+1/2}(x,0) = \begin{cases} \widetilde{W}_i^n & \text{if } x < 0, \\ \widetilde{W}_{i+1}^n & \text{if } x > 0, \end{cases} \end{cases} \tag{2.7}$$

and we approximate the solution at time t^{n+1} by

$$\begin{aligned} \widetilde{W}_i^{n+1} = \frac{1}{\Delta x}\Bigg(& \int_{x_{i-1/2}}^{x_i} \widetilde{W}^{i-1/2}(x - x_{i-1/2}, \Delta t)\, dx \\ & + \int_{x_i}^{x_{i+1/2}} \widetilde{W}^{i+1/2}(x - x_{i+1/2}, \Delta t)\, dx \Bigg). \end{aligned} \tag{2.8}$$

If a CFL 1/2 condition is assumed and the $(N+1)$th components are dropped, some easy calculations allow us to write the numerical scheme as follows:

$$W_i^{n+1} = W_i^n + \frac{\Delta t}{\Delta x}\left(A\big(W_{i-1/2}^{n,+} - W_i^n\big) + A(W_i^n - W_{i+1/2}^{n,-})\big)\right), \tag{2.9}$$

or, equivalently,

$$W_i^{n+1} = W_i^n + \frac{\Delta t}{\Delta x} A\big(W_{i-1/2}^{n,+} - W_{i+1/2}^{n,-}\big), \tag{2.10}$$

where $W_{i+1/2}^{n,\pm}$ are the limits of the solution of problem (2.7) to the left and to the right of $x = 0$.

Finally, by making use of the equalities in Remark 2.3, the numerical scheme can also be written as follows:

$$W_i^{n+1} = W_i^n + \frac{\Delta t}{\Delta x}\big(F_{i-1/2}^+ - F_{i+1/2}^-\big), \tag{2.11}$$

with

$$F_{i+1/2}^- = A\sum_{j=1}^{I} s_j^{i+1/2} R_j, \quad F_{i+1/2}^+ = -A\sum_{j=I+1}^{N} s_j^{i+1/2} R_j,$$

I being the maximum value of i for which $\lambda_i < 0$, and $s_j^{i+1/2}$, $j = 1, \ldots, N$, the solutions of

$$W_{i+1}^n - e^{A^{-1}C(\sigma_{i+1} - \sigma_i)} W_i^n = e^{A^{-1}C(\sigma_{i+1} - \sigma_i)} \sum_{j=1}^{I} s_j^{i+1/2} R_j + \sum_{j=I+1}^{N} s_j^{i+1/2} R_j. \tag{2.12}$$

Remark 2.5.

1. In practice, a CFL 1 condition can be used, as it ensures the stability of the scheme.
2. This numerical scheme is path-conservative in the sense defined in [27]. Consequently, it is consistent in the usual sense for regular solutions.
3. In [23] a convergence result of Lax–Wendroff type has been proved for Godunov's methods in the context of a more general class of hyperbolic nonconservative systems.

4. In the case $N = 1$ Godunov's method studied here is the particular case of the well-balanced numerical scheme introduced in [9] corresponding to a linear flux function. This is also a particular case of Godunov's method for scalar balance laws studied in [15], where the convergence of the numerical method has been proved.
5. The Goldstein–Taylor model of the Boltzmann equations (see [20]) is the particular case of (2.1) corresponding to the choice

$$W = \begin{bmatrix} v \\ u \end{bmatrix}, \quad \sigma(x) = x,$$
$$A = \frac{1}{\varepsilon}\begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}, \quad C = \frac{1}{\varepsilon}\begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix}.$$

In this case, it can be shown that Godunov's method (2.10) coincides with the well-balanced numerical scheme introduced in [11].

3. High order extensions. The goal of this section is to construct high order extensions of Godunov's methods introduced in section 2, based on the use of a reconstruction operator of order s, i.e., an operator that associates to a given sequence $\{\widetilde{W}_i\}$ two new sequences $\{\widetilde{W}^-_{i+1/2}\}$ and $\{\widetilde{W}^+_{i+1/2}\}$ in such a way that, whenever

$$\widetilde{W}_i = \frac{1}{\Delta x}\int_{I_i} \widetilde{W}(x)\,dx \quad \forall\, i \in \mathbb{Z}$$

for some smooth function $\widetilde{W}$, then

$$\widetilde{W}^{\pm}_{i+1/2} = \widetilde{W}(x_{i+1/2}) + O\left(\Delta x^s\right) \quad \forall\, i \in \mathbb{Z}.$$

We follow [6] or [27] to construct a semidiscrete high order extension of Godunov's methods. Let us introduce some notation:

$$\widetilde{W}_i(t) = \begin{bmatrix} W_i(t) \\ \sigma_i \end{bmatrix}$$

is the approximation given by the numerical scheme of the average of the solution $\widetilde{W}$ at the ith cell at time t;

$$\widetilde{W}^{\pm}_{i+1/2}(t) = \begin{bmatrix} W^{\pm}_{i+1/2}(t) \\ \sigma^{\pm}_{i+1/2}(t) \end{bmatrix}$$

are the reconstructed states at the intercells associated to the sequence $\{\widetilde{W}_i(t)\}$;

$$\widetilde{W}^{\pm}_{i+1/2,0}(t) = \begin{bmatrix} W^{\pm}_{i+1/2,0}(t) \\ \sigma^{\pm}_{i+1/2}(t) \end{bmatrix}$$

represent the limits to the left and to the right of $x = 0$ of the solution of the Riemann problem (2.7) with initial data $\widetilde{W}^{\pm}_{i+1/2}(t)$; finally,

$$\widetilde{P}^t_i = \begin{bmatrix} P^t_i \\ p^t_{N+1,i} \end{bmatrix} = \begin{bmatrix} p^t_{1,i} \\ \vdots \\ p^t_{N,i} \\ p^t_{N+1,i} \end{bmatrix}$$

is a function defined at the ith cell with values in $\mathbb{R}^{N+1}$, satisfying

$$\lim_{x\to x^+_{i-1/2}} \widetilde{P}^t_i(x) = \widetilde{W}^+_{i-1/2}(t), \qquad \lim_{x\to x^-_{i+1/2}} \widetilde{P}^t_i(x) = \widetilde{W}^-_{i+1/2}(t).$$

In practice, the definition of the reconstruction operator gives the natural choice of the functions $\widetilde{P}^t_i$, as the usual procedure is the following: Given a sequence $\{\widetilde{W}_i\}$ of values at the cells, an approximation function is calculated at every cell I_i by using the values $\widetilde{W}_j$ in a stencil,

$$\widetilde{P}_i(x; \widetilde{W}_{i-l}, \dots, \widetilde{W}_{i+r}),$$

with l, r being two natural numbers. The reconstructions $\widetilde{W}^{\pm}_{i+1/2}$ are then calculated by taking the limits of these functions at the intercells. These approximations functions are usually calculated by means of interpolation or approximation techniques. The natural choice of $\widetilde{P}^t_i$ is thus

$$\widetilde{P}^t_i(x) = \widetilde{P}_i(x; \widetilde{W}_{i-l}(t), \dots, \widetilde{W}_{i+r}(t)).$$

With this notation, the general expression of a semidiscrete numerical scheme proposed here is the following:

$$W_i' = \frac{1}{\Delta x}\left(A W^+_{i-1/2,0}(t) - A W^-_{i+1/2,0}(t) + \int_{x_{i-1/2}}^{x_{i+1/2}} C P^t_i(x) \frac{d}{dx} p^t_{N+1,i}(x)\, dx \right). \tag{3.1}$$

The following result can be shown (see [27]).

THEOREM 3.1. *Let us suppose that the reconstruction operator is of order s and that, given the sequence defined by*

$$\widetilde{W}_i = \frac{1}{\Delta x} \int_{I_i} \widetilde{W}(x)\, dx,$$

for any smooth function $\widetilde{W}$, the following approximation properties are satisfied:

$$\widetilde{P}_i(x; \widetilde{W}_{i-l}, \dots, \widetilde{W}_{i+r}) = \widetilde{W}(x) + O\left(\Delta x^{s_1}\right) \quad \forall\, x \in I_i,$$
$$\frac{d}{dx}\widetilde{P}_i(x; \widetilde{W}_{i-l}, \dots, \widetilde{W}_{i+r}) = \widetilde{W}'(x) + O\left(\Delta x^{s_2}\right) \quad \forall\, x \in I_i.$$

Then (3.1) *is of order at least $\bar{s} = \min(s, s_1 + 1, s_2 + 1)$.*

Remark 3.2. While in the case of a system of conservation laws, i.e., if $C = 0$, the numerical scheme is obviously independent of the function $\widetilde{P}^t_i$ chosen at the cells, this is not the case in the presence of source terms. As a consequence, while the numerical scheme has order s for systems of conservation laws, when $C \neq 0$ the order depends on the choice of the functions $\widetilde{P}^t_i$. For the usual reconstruction techniques one has $s_2 < s_1 \leqslant s$, and the order of (3.1) is thus $s_2 + 1$ for nonconservative systems and s for systems of conservation laws. Therefore a loss of accuracy can be observed when a technique of reconstruction is applied to a nonconservative problem. This effect has been detected and verified numerically in [6].

4. Well-balanced high order extensions. Well-balancing is related to the numerical approximation of equilibria, i.e., steady-state solutions, which are in turn related to the family of integral curves of the $(N+1)$th characteristic field (2.3):

$$\Gamma_{W^*} = \{[X(\sigma, W^*), \sigma]^T,\ \sigma \in \mathbb{R}\}, \quad W^* \in \mathbb{R}^N,$$

where

$$X(\sigma, W^*) = e^{A^{-1}C\sigma} W^*.$$

In effect, the stationary solutions of (2.2) always take values on a curve of this family. To see this, consider first a discontinuous stationary solution of the form

$$\widetilde{W}(x,t) = \begin{cases} \widetilde{W}_L & \text{if } x < 0, \\ \widetilde{W}_R & \text{if } x > 0. \end{cases}$$

By taking into account the proof of Proposition 2.1, such a solution is possible only if there exists W^* such that both states $\widetilde{W}_L$ and $\widetilde{W}_R$ belong to Γ_{W^*}.

Next, if $\widetilde{W}(x) = [W(x), \sigma(x)]^T$ is a regular stationary solution, then

$$\widetilde{A}(\widetilde{W}(x))\widetilde{W}'(x) = 0 \quad \forall\, x.$$

If $\widetilde{W}'(x) \neq 0$ for every x, $\widetilde{W}'(x)$ is an eigenvector of the matrix $\widetilde{A}(\widetilde{W}(x))$ associated to the eigenvalue 0; i.e., the function can be interpreted as a parameterization of an arc of an integral curve of the $(N+1)$th characteristic field. As a consequence, there exists W^* such that

$$W(x) = X(\sigma(x), W^*) \quad \forall\, x.$$

The following definition of a well-balanced numerical scheme can be introduced (see [26], [27]).

DEFINITION 4.1. *A numerical scheme is said to be* exactly well-balanced *for an integral curve Γ_{W^*} if it exactly solves every stationary solution taking values in this curve. The scheme is said to be exactly well-balanced if it is exactly well-balanced for every integral curve Γ_{W^*}.*

By using (2.10) and the proof of Proposition 2.1, the following result can be easily proved.

PROPOSITION 4.2. *Godunov's method* (2.8) *is exactly well-balanced.*

The definition of well-balanced numerical schemes can be extended to semidiscrete numerical schemes such as those introduced in the previous section.

DEFINITION 4.3. *A semidiscrete numerical method* (3.1) *is said to be* exactly well-balanced *for an integral curve Γ_{W^*} if, given a stationary solution $\widetilde{W}(x) = [W(x), \sigma(x)]^T$ taking values on this curve, the vector $\{W_i\}$ defined by*

$$W_i = \frac{1}{\Delta x} \int_{x_{i-1/2}}^{x_{i+1/2}} W(x)\, dx \quad \forall\, i \tag{4.1}$$

is a critical point of the system of differential equations defined by (3.1).

We introduce also the definition of well-balanced reconstruction operators. Let us suppose that the reconstructions are calculated on the basis of some approximation functions at the cells:

$$\widetilde{P}_i(x; \widetilde{W}_{i-l}, \ldots, \widetilde{W}_{i+r}).$$

DEFINITION 4.4. *The reconstruction operator is said to be* well-balanced *for an integral curve Γ_{W^*} if, given a stationary solution $\widetilde{W}(x) = [W(x), \sigma(x)]^T$ taking values*

on this curve, the approximation functions corresponding to the cell averages of $\widetilde{W}(x)$ *satisfy*

$$\widetilde{P}_i(x;\widetilde{W}_{i-l},\dots,\widetilde{W}_{i+r}) \in \Gamma_{W^*} \quad \forall\, x \in [x_{i-1/2}, x_{i+1/2}].$$

PROPOSITION 4.5. *If the reconstruction operator is well-balanced for a curve* Γ_{W^*}*, the numerical scheme* (3.1) *is exactly well-balanced for* Γ_{W^*}*.*

Proof. Let us apply the numerical scheme to the initial condition $\{W_i\}$ given by (4.1), $\widetilde{W}(x)$ being a stationary solution taking values on Γ_{W^*}. Due to the well-balanced character of the reconstruction operator, we have that, for every i,

$$x \in [x_{i-1/2}, x_{i+1/2}] \mapsto \widetilde{P}_i^0(x)$$

is a parameterization of an arc of an integral curve. Therefore,

$$\int_{x_{i-1/2}}^{x_{i+1/2}} CP_i^0(x)\frac{d}{dx}p_{N+1,i}^0(x)\,dx = A\big(W_{i+1/2}^-(0) - W_{i-1/2}^+(0)\big).$$

As a consequence, at time $t = 0$ the right-hand side of (3.1) reduces to

$$\frac{1}{\Delta x}\left(A\big(W_{i-1/2,0}^+(0) - W_{i-1/2}^+(0)\big) - A\big(W_{i+1/2,0}^-(0) - W_{i+1/2}^-(0)\big)\right).$$

Finally, as $W_{i+1/2}^{\pm}(0)$ belong to the same integral curve of the $(N+1)$th characteristic field, the solution of the corresponding Riemann problem is a stationary discontinuity. Then

$$W_{i-1/2,0}^+(0) = W_{i-1/2}^+(0), \quad W_{i+1/2,0}^-(0) = W_{i+1/2}^-(0).$$

Thus, the right-hand side of (3.1) vanishes, as we wanted to prove. ∎

Let us remark that the usual reconstruction operators are based on standard interpolation techniques, as polynomial interpolation: This is the case of essentially nonoscillatory (ENO) or weighted ENO (WENO) reconstructions, for instance. But in this case, there is not any particular reason to expect the operator to be well-balanced for a given integral curve Γ_{W^*}, except for the trivial case $W^* = 0$: Notice that the expression of the integral curve involves exponential terms that cannot be exactly approached by means of polynomial interpolators.

We propose here a strategy to obtain well-balanced reconstructions based on a modification of a standard reconstruction operator in $\mathbb{R}^N$ of order s

$$Q_i(x; W_{i-l},\dots,W_{i+r}). \tag{4.2}$$

The only requirement for (4.2) is that it has to be exact for the null function, i.e.,

$$Q_i(x;0,\dots,0) = 0 \quad \forall\, x.$$

Given a sequence $\{\widetilde{W}_i = [W_i, \sigma_i]^T\}$ of values at the cells, the reconstruction function $\widetilde{P}_i$ corresponding to the ith cell, I_i, is calculated as follows:

1. Look for W_i^* such that

$$W_i = \frac{1}{\Delta x}\int_{x_{i-1/2}}^{x_{i+1/2}} X(\sigma(x), W_i^*)\,dx. \tag{4.3}$$

2. For $j = i-l, \dots, i+r$, define V_j by

$$V_j = W_j - \frac{1}{\Delta x}\int_{x_{j-1/2}}^{x_{j+1/2}} X(\sigma(x), W_i^*)\, dx.$$

3. Set

$$Q_i(x) = Q_i(x; V_{i-l}, \dots, V_{i+r}), \quad x \in I_i. \tag{4.4}$$

4. Finally, define

$$\widetilde{P}_i(x) = \begin{bmatrix} X(\sigma(x), W_i^*) + Q_i(x) \\ \sigma(x) \end{bmatrix}, \quad x \in I_i. \tag{4.5}$$

5. The reconstructions at the intercells are thus given by

$$\widetilde{W}_{i-1/2}^{+} = \begin{bmatrix} X(\sigma(x_{i-1/2}), W_i^*) + Q_i(x_{i-1/2}) \\ \sigma(x_{i-1/2}) \end{bmatrix}, \tag{4.6}$$

$$\widetilde{W}_{i+1/2}^{-} = \begin{bmatrix} X(\sigma(x_{i+1/2}), W_i^*) + Q_i(x_{i+1/2}) \\ \sigma(x_{i+1/2}) \end{bmatrix}. \tag{4.7}$$

Notice that, following this construction, P_i is not a polynomial: It is the sum of the polynomial Q_i and the stationary solution $X(\sigma(x_{i+1/2}), W_i^*)$ whose average is W_i.

PROPOSITION 4.6. *The reconstruction operator given by* (4.6)–(4.7) *is well-balanced and of order* s.

Proof. Let us consider a vector $\{\widetilde{W}_i\}$ defined by (4.1) where $\widetilde{W}(x)$ is a stationary solution taking values on Γ_{W^*}, for some W^*. In this case, the solution of (4.3) is $W_i^* = W^*$ for every i. Moreover,

$$V_j = 0, \quad j = i-l, \dots, i+r,$$

and thus $Q_i \equiv 0$. Therefore,

$$\widetilde{P}_i(x) = [X(\sigma(x); W^*), \sigma(x)]^T \in \Gamma_{W^*} \quad \forall\, x,$$

and the reconstruction operator is well-balanced.

Next, let us suppose that the values W_i are the averages at the cells of a smooth function W and that σ and, consequently, $X(\sigma(x), W^*)$ are also smooth. Then, by construction:

$$\begin{aligned} Q_i(x_{i+1/2}) &= W(x_{i+1/2}) - X(x_{i+1/2}, W_i^*) + O(\Delta x^s), \\ Q_i(x_{i-1/2}) &= W(x_{i-1/2}) - X(x_{i-1/2}, W_i^*) + O(\Delta x^s). \end{aligned}$$

From these equalities and the definition of the functions $\widetilde{P}_i$, it can be easily deduced that the reconstruction operator is of order s. ∎

By making use of the definition of the functions $\widetilde{P}_i$, a reformulation of the well-balanced high order numerical scheme (3.1) can be given that makes appear explicitly the stationary solutions and the operator Q used in the reconstructions of states:

$$\begin{aligned} W_i' = \frac{1}{\Delta x}\Big(& A\big(W_{i-1/2,0}^{+}(t) - X(\sigma(x_{i-1/2}), W_i^*(t))\big) \\ & -A\big(W_{i+1/2,0}^{-}(t) - X(\sigma(x_{i+1/2}), W_i^*(t))\big) + \int_{x_{i-1/2}}^{x_{i+1/2}} CQ_i^t(x)\sigma_x(x)\, dx\Big). \end{aligned} \tag{4.8}$$

Remark 4.7.

1. Notice that problems (4.3) to be solved at the cells are linear:

$$\left(\frac{1}{\Delta x}\int_{x_{i-1/2}}^{x_{i+1/2}} e^{A^{-1}C\sigma(x)}\,dx\right)W_i^*(t) = W_i(t),$$

but the coefficients of the matrix can be difficult to calculate. The matrix of this linear system is regular if Δx is small enough, as its limit as Δx tends to 0 is

$$e^{A^{-1}C\sigma(x_i)},$$

which is regular.

2. Let us suppose that the following approximation hypotheses hold:

$$Q_i(x; V_{i-l}, \dots, V_{i+r}) = V(x) + O\left(\Delta x^{s_1}\right) \quad \forall\, x \in I_i,$$
$$\frac{d}{dx}Q_i(x; V_{i-l}, \dots, V_{i+r}) = V'(x) + O\left(\Delta x^{s_2}\right) \quad \forall\, x \in I_i,$$

for any given sequence defined by

$$V_i = \frac{1}{\Delta x}\int_{I_i} V(x)\,dx,$$

V being any smooth function. Then, if σ is smooth enough, it can be easily shown that the approximation functions $\widetilde{P}_i$ satisfy the hypotheses of Theorem 3.1.

3. If the integrals appearing in (4.8) are approached by a quadrature formula which is exact for the null function, the numerical scheme is still exactly well-balanced.

4. If the cell averages $\{W_i\}$ at time $t = 0$ are calculated by means of a quadrature formula, the introduced strategy still produces a well-balanced reconstruction operator if the integral in (4.3) is approached by using this same quadrature formula.

5. Extensions to more general problems.

5.1. Linear conservation laws with nonlinear source terms. We consider now the case of systems of the form

$$W_t + AW_x = G(W)\sigma_x, \tag{5.1}$$

where G is a smooth function from $\mathbb{R}^N$ to $\mathbb{R}$. The system can still be written under the form (2.2), with

$$\widetilde{\mathcal{A}}(\widetilde{W}) = \begin{bmatrix} A & -G(W) \\ 0 & 0 \end{bmatrix}.$$

Again, the system is strictly hyperbolic if A has N real distinct nonvanishing eigenvalues.

The main difference with the linear case comes from the expression of the integral curves of the $(N+1)$th characteristic field associated to the null eigenvalue. In this case, these integral curves are those of the autonomous ODE system

$$\frac{dW}{d\sigma} = A^{-1}G(W).$$

Let us denote by $X(\sigma; W^*)$ the maximal solution of this system with initial condition

$$W(0) = W^* \tag{5.2}$$

and by Γ_{W^*} the corresponding curve in $\mathbb{R}^{N+1}$.

The solution of the Riemann problem is again constructed as in the proof of Proposition 2.1, but now the following nonlinear system has to be solved: Find $s_1,\dots,$ s_N such that

$$X\left([\![\sigma]\!], W_L + \sum_{j=1}^{I} s_j R_j\right) = W_R - \sum_{j=I+1}^{N} s_j R_j, \tag{5.3}$$

where I denotes the last index of a negative eigenvalue. The Riemann problem has then a unique solution composed by $N+1$ contact discontinuities if and only if this nonlinear system has a unique solution $\{s_1,\dots,s_N\}$. In this case, the solution is still given by (2.5), with

$$\widetilde{W_0^+} = \begin{bmatrix} X\big([\![\sigma]\!], W_L + \sum_{j=1}^{I} s_j R_j\big) \\ \sigma_R \end{bmatrix}.$$

Obviously, the solution of the nonlinear equation (5.3) can be hard or even impossible to calculate in an explicit way. Nevertheless, it is easier to solve if all of the eigenvalues have the same sign, i.e., if $I=0$ or $I=N$.

Godunov's method can still be written under any of the forms (2.9), (2.10), or (2.11), by just replacing the linear equation (2.12) by the nonlinear one (5.3).

By using the notations of section 3, the expression of a high order extension is now

$$W_i' = \frac{1}{\Delta x}\left(AW^+_{i-1/2,0}(t) - AW^-_{i+1/2,0}(t) + \int_{x_{i-1/2}}^{x_{i+1/2}} G(P_i^t(x))\frac{d}{dx}p^t_{N+1,i}(x)\,dx\right),$$

and the expression of the well-balanced high order extension is as follows:

$$\begin{aligned} W_i' = \frac{1}{\Delta x}\Bigg(& A\big(W^+_{i-1/2,0}(t) - X(\sigma(x_{i-1/2}), W_i^*(t))\big) \\ & - A\big(W^-_{i+1/2,0}(t) - X(\sigma(x_{i+1/2}), W_i^*(t))\big) \\ & + \int_{x_{i-1/2}}^{x_{i+1/2}} \big(G(P_i^t(x)) - G(X(\sigma, W_i^*))\big)\sigma_x(x)\,dx\Bigg). \end{aligned}$$

The main difficulty to apply such a numerical scheme comes from the calculation of the reconstruction operator: On the one hand, the calculation of $X(\sigma; W)$ can be very difficult; on the other hand, even if it is possible to calculate these functions, the equation (4.3) is now nonlinear, and it can be difficult to solve.

5.2. Linear problems with nonconstant coefficients. We finally consider systems of the form

$$W_t + A(\sigma)W_x = G(\sigma, W)\sigma_x;$$

i.e., the coefficients of the matrix A and the function G are allowed to depend on σ and thus on x.

This system is strictly hyperbolic if the eigenvalues of $A(\sigma)$ are real, distinct, and different from zero. In this case, the integral curves of the characteristic field related to the eigenvalue zero are the solutions of the ODE system

$$\frac{dW}{d\sigma} = A(\sigma)^{-1}G(\sigma, W).$$

With the notation used in the previous subsection, the nonlinear systems to be solved in order to calculate the solution of the Riemann problem are again (5.3). Once more, these systems can be very difficult to solve.

The expression of Godunov's method is now

$$W_i^{n+1} = W_i^n + \frac{\Delta t}{\Delta x} A(\sigma_i)\big(W_{i-1/2}^{n,+} - W_{i+1/2}^{n,-}\big).$$

The general expression (3.1) of a high order extension can be written in this case as follows:

$$W_i' = \frac{1}{\Delta x}\Bigg(A(\sigma_{i-1/2}^+)W_{i-1/2,0}^+(t) - A(\sigma_{i+1/2}^-)W_{i+1/2,0}^-(t) + \int_{x_{i-1/2}}^{x_{i+1/2}} \widetilde{G}(\widetilde{P}_i^t(x))\frac{d}{dx}p_{N+1,i}^t(x)\,dx\Bigg),$$

where

$$\widetilde{G}(\widetilde{W}) = G(\sigma, W) + A'(\sigma)W,$$

where $A'(\sigma)$ represents the derivative of $A(\sigma)$. Finally, the well-balanced high order extension should be written as follows:

$$W_i' = \frac{1}{\Delta x}\Bigg(A(\sigma(x_{i-1/2}))\big(W_{i-1/2,0}^+(t) - X(\sigma(x_{i-1/2}), W_i^*(t))\big) - A(\sigma(x_{i+1/2}))\big(W_{i+1/2,0}^-(t) - X(\sigma(x_{i+1/2}), W_i^*(t))\big) + \int_{x_{i-1/2}}^{x_{i+1/2}} \big(\widetilde{G}(\widetilde{P}_i^t(x)) - \widetilde{G}(X(\sigma(x), W_i^*(t)))\big)\sigma_x(x)\,dx\Bigg).$$

Again, (4.3) can be difficult or even impossible to solve.

6. Numerical tests.

6.1. A linear scalar problem. Let us begin by considering an easy scalar problem, in order to illustrate the methodology introduced in previous sections. We consider the linear balance law

$$\begin{cases} w_t + \lambda w_x = cw, \\ w(x,0) = w_0(x), \end{cases} \tag{6.1}$$

with $\lambda > 0$. First, the problem is reformulated as follows:

$$\begin{cases} \widetilde{w}_t + \widetilde{A}(\widetilde{w})\widetilde{w}_x = 0, \\ \widetilde{w}(x_0) = \widetilde{w}_0(x), \end{cases} \tag{6.2}$$

where

$$\widetilde{w} = \begin{bmatrix} w \\ \sigma \end{bmatrix}, \quad \widetilde{A}(\widetilde{w}) = \begin{bmatrix} \lambda & -cw \\ 0 & 0 \end{bmatrix}, \quad \widetilde{w}_0(x) = \begin{bmatrix} w_0(x) \\ x \end{bmatrix}, \quad \sigma(x) = x.$$

System (6.2) is strictly hyperbolic: The matrix has two real eigenvalues λ and 0, with associated eigenvectors

$$r_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad r_2 = \begin{bmatrix} \lambda^{-1}cw \\ 1 \end{bmatrix}.$$

The integral curves of the second characteristic field are given by

$$w = ke^{\lambda^{-1}c\sigma}, \quad \sigma \in \mathbb{R},$$

where k is an arbitrary constant.

A major advantage of reformulating the problem as a quasi-linear system comes from the fact that the solutions of Riemann problems for (6.2) are self-similar and they consist of constant states linked by simple waves, which is not the case for (6.1).

After some straightforward calculations, Godunov's method (2.10) for this particular case can be written as follows:

$$w_i^{n+1} = w_i^n + \frac{\Delta t}{\Delta x}\lambda\big(w_{i-1}^n e^{\lambda^{-1}c\Delta x} - w_i^n\big),$$

which can also be reformulated in the following form:

$$w_i^{n+1} = w_i^n + \frac{\Delta t}{\Delta x}\lambda(w_{i-1}^n - w_i^n) + \frac{\Delta t}{\Delta x}\lambda\big(e^{\lambda^{-1}c\Delta x} - 1\big)w_{i-1}^n. \tag{6.3}$$

Notice that the second addend on the right-hand side is a consistent approximation of the source term, as

$$\lambda\big(e^{\lambda^{-1}c\Delta x} - 1\big) = c\Delta x + \mathcal{O}(\Delta x^2).$$

The numerical treatment of the source term in (6.3) can be interpreted as an approximation by means of a sum of Dirac masses, as was done in [11] and [10]: Once the approximations at time t^n have been obtained, the source term $cw(x, t^n)$ is approximated by the measure

$$\sum_i \lambda\big(e^{\lambda^{-1}c\Delta x} - 1\big)w_{i-1}^n \delta_{x=x_{i+1/2}},$$

and every Dirac mass in the above sum is taken into account in the Riemann problem solved at the corresponding intercells.

The wave-propagation algorithm proposed in [3] to approximate quasi-steady solutions of balance laws is based on a similar approach: At every intercell, a Riemann problem is considered where the source term is approached by a delta function. For instance, in this easy example, a Riemann problem would be considered at the intercell $x_{i+1/2}$ for an equation of the form

$$w_t + \lambda w_x = cw_{i+1/2}^n \delta_{x=x_{i+1/2}}\Delta x,$$

where $w_{i+1/2}^n$ is an average of w_i^n and w_{i+1}^n to be chosen. The numerical scheme (6.3) can thus be interpreted as the particular case of the wave-propagation algorithm corresponding to the choice

$$w_{i+1/2}^n = \frac{\lambda}{c\Delta x}\big(e^{\lambda^{-1}c\Delta x} - 1\big)w_i^n.$$

This choice, given by the solution of the Riemann problem for the nonconservative system (6.2), ensures the well-balanced character of the numerical scheme for every stationary solution.

If, instead, the choice

$$w_{i+1/2}^n = \frac{1}{2}\left(w_i^n + w_{i+1}^n\right)$$

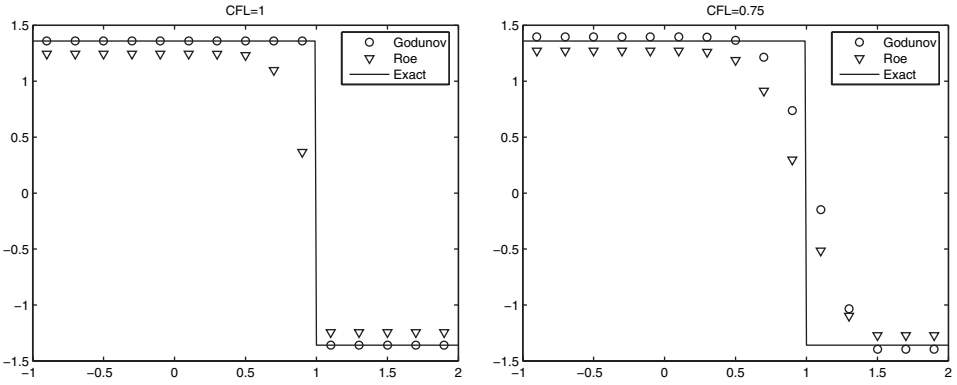


FIG. 1. *Solution of problem* (6.1) *at time* $t = 1$ *with initial condition* (6.5). *Left: CFL* $= 1$*; right: CFL* $= 0.75$.

is done, the following wave-propagation algorithm is obtained:

$$w_i^{n+1} = w_i^n + \lambda \frac{\Delta t}{\Delta x} \left(w_{i-1}^n - w_i^n \right) + \Delta t c \frac{w_{i-1}^n + w_i^n}{2}. \tag{6.4}$$

This is the classical scheme upwinding the source term (see [28] or [4]). It can also be obtained by applying to system (6.2) a generalized Roe's method of the family introduced by Toumi in [31]. This family of methods is based on the choice of a family of paths which, in the particular case of (6.4), is the family of segments: See [26] for details. In what follows, the numerical scheme (6.4) will be referred to as Roe's method. The general results shown in [26] allow one to prove that this Roe's method is not exactly well-balanced but it solves the stationary solutions up to second order.

In order to compare the numerical schemes (6.3) and (6.4), we consider (6.1) on the interval $-1 < x < 2$ with $\lambda = 1$ and $c = 1$. We consider the Riemann problem corresponding to the initial condition

$$w(x,0) = \begin{cases} 0.5 & \text{if } x < 0, \\ -0.5 & \text{if } x > 0. \end{cases} \tag{6.5}$$

In Figure 1 we compare the numerical results obtained with both numerical schemes by using $\Delta x = 0.2$ and CFL $= 1$. Observe that Godunov's method (6.3) gives the exact solution, even if the Riemann problems solved at the intercells are not associated to the PDE in (6.1) but to an approximation of it, due to the artificial piecewise constant approximation of the identity function performed. On the other hand, the numerical scheme (6.4) correctly approximates the location of the discontinuity, but the exponential growth of the states at both sides cannot be captured. The error in the approximation of these states increases with time. Nevertheless, the solutions provided by (6.4) converge to the exact solution as Δx tends to 0.

For CFL $= 0.75$, (6.3) still gives a better approximation of the growth of the states at both sides of the discontinuity; see Figure 1.

To test the well-balanced property of the numerical schemes, we consider the stationary solution $w(x) = e^x$ as the initial condition on the interval $[-3, 3]$: As expected, Godunov's method is exact, while Roe's method solves the stationary solution with order 2 (see Table 1).

TABLE 1
Error in L^1 norm for Roe's scheme (6.4) *and Godunov's scheme* (6.3) *with initial condition* $w(x,0) = e^x$ *at time* $t = 2$. *CFL* $= 0.9$, $c = \lambda = 1$.

No. cells	L^1 error (Roe)	Order (Roe)	L^1 error (Godunov)
15	1.4677	-	2.9403E-16
30	3.4590E-01	2.08	6.0211E-16
60	8.4829E-02	2.03	3.0355E-15
120	2.1053E-02	2.01	5.8025E-15
240	5.2472E-03	2.01	2.1552E-15

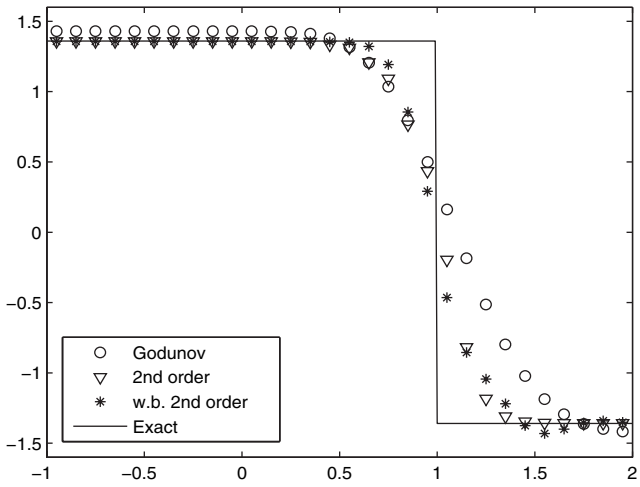


FIG. 2. *Solution of problem* (6.1) *at time* $t = 1$ *with initial condition* (6.5). *CFL* $= 0.75$.

By following section 3, the form of a high order extension of (6.3) is as follows:

$$w_i' = \frac{1}{\Delta x}\left(\lambda w^-_{i-1/2}(t)e^{\lambda^{-1}c[\sigma^t_{i+1/2}]} - \lambda w^-_{i+1/2}(t) + c\int_{x_{i-1/2}}^{x_{i+1/2}} p^t_{1,i}(x)\frac{d}{dx}p^t_{2,i}(x)\,dx\right).$$

Notice that, as $\sigma_i = x_i$ for every i, if the approximation functions are exact for polynomials of degree 1, the numerical scheme reduces to

$$w_i' = \frac{1}{\Delta x}\left(\lambda w^-_{i-1/2}(t) - \lambda w^-_{i+1/2}(t) + c\int_{x_{i-1/2}}^{x_{i+1/2}} p^t_{1,i}(x)\,dx\right). \tag{6.6}$$

We consider again the Riemann problem (6.5), and we apply a second order extension of Godunov's method based on a monotone upstream-centered schemes for conservation laws (MUSCL) reconstruction with a *minmod* slope limiter (see [32]). The discretization in time is performed with an explicit Runge–Kutta TVD second order method (see [12]). In Figure 2 we compare the results obtained for CFL $= 0.75$ with those corresponding to Godunov's method.

Unlike Godunov's method, this higher order extension is not exactly well-balanced. The errors and computed order corresponding to the stationary solution $w(x) = e^x$ as initial data on the interval $[-3, 3]$ are shown in Table 2. As expected, the solution is approximated with the order of the numerical scheme.

To obtain a well-balanced second order extension of (6.3), the techniques developed in section 4 can be applied: The form of a well-balanced high order extension

TABLE 2
Error in L^1 norm and order for the scheme (6.6) *with initial condition* $w(x,0) = e^x$ *at time* $t = 2$. *CFL* $= 0.9$, $c = \lambda = 1$.

No. cells	L^1 error	Order
100	3.9976E-03	-
200	1.0606E-03	1.91
400	2.7299E-04	1.96
800	6.9239E-05	1.98
1600	1.7434E-05	1.99

TABLE 3
Error in L^1 norm for the scheme (6.7) *with initial condition* $w(x,0) = e^x$ *at time* $t = 2$. *CFL* $= 0.9$, $c = \lambda = 1$.

No. cells	L^1 error
15	1.2888E-15
30	1.5069E-15
60	8.4222E-15
120	6.5241E-15

(4.8) applied to the particular case (6.1) is as follows:

$$w_i' = \frac{1}{\Delta x}\left(\lambda\big(w^-_{i-1/2}(t) - e^{\lambda^{-1}cx_{i-1/2}}w^*_i(t)\big) - \lambda\big(w^-_{i+1/2}(t) - e^{\lambda^{-1}cx_{i+1/2}}w^*_i(t)\big) + c\int_{x_{i-1/2}}^{x_{i+1/2}} q^t_i(x)\,dx\right), \tag{6.7}$$

where

$$w^*_i(t) = \frac{c\lambda^{-1}\Delta x}{e^{\lambda^{-1}cx_{i+1/2}} - e^{\lambda^{-1}cx_{i-1/2}}}w_i(t),$$
$$q^t_i = q_i(\cdot; v_{i-l}(t), \dots, v_{i+r}(t)),$$

q_i being the approximation functions of a reconstruction operator in $\mathbb{R}$, and

$$v_j(t) = w_j(t) - \frac{\lambda}{c\Delta x}\big(e^{\lambda^{-1}c(j+1/2)\Delta x} - e^{\lambda^{-1}c(j-1/2)\Delta x}\big)w^*_i(t), \quad j = i-l, \dots, i+r.$$

Finally,

$$w^+_{i+1/2}(t) = q^t_{i+1}(x_{i+1/2}) + e^{\lambda^{-1}cx_{i+1/2}}w^*_{i+1}(t),$$
$$w^-_{i+1/2}(t) = q^t_i(x_{i+1/2}) + e^{\lambda^{-1}cx_{i+1/2}}w^*_i(t).$$

Again, MUSCL reconstructions and an explicit Runge–Kutta TVD method for the discretization in time are considered. In Figure 2 we compare the results obtained for the considered Riemann problem above for CFL $= 0.75$ with Godunov's method, its second order extension, and the well-balanced second order extension. As can be seen, the results obtained with both second order numerical schemes are similar.

Table 3 shows the committed errors solving the stationary solution $w(x) = e^x$ on the interval $[-3, 3]$ with the well-balanced second order method. As expected, the solution is exactly calculated.

Finally, we include a test to verify the second order accuracy of the schemes for time-dependent solutions. We solve (6.1) with $w_0(x) = e^{-10x^2}$ on the interval $[-3, 3]$. In Table 4 we show the L^1 errors and the order obtained with the schemes (6.6) and (6.7).

TABLE 4
Error in L^1 norm for the schemes (6.6) and (6.7) with initial condition $w(x,0) = e^{-10x^2}$ at time $t = 1$. CFL $= 0.9$, $c = \lambda = 1$.

No. cells	L^1 error (sch. (6.6))	Order (sch. (6.6))	L^1 error (sch. (6.7))	Order (sch. (6.7))
100	2.246E-02	-	2.260E-02	-
200	6.273E-03	1.84	6.310E-03	1.84
400	1.626E-03	1.94	1.635E-03	1.94
800	4.131E-04	1.97	4.154E-04	1.97

6.2. A linear scalar problem with a singular source term. In the previous paragraph, it has been shown that numerical results obtained with Roe's scheme (6.4) are worse than those obtained with Godunov's method (6.3) but that, nevertheless, they converge to the right solution of problem (6.1) as Δx tends to 0. In this example, we show that the convergence of Roe's scheme may fail when $\sigma(x)$ is discontinuous.

We consider the scalar equation

$$\begin{cases} w_t + 2w_x = w\sigma'(x), \\ w(x,0) = w_0(x), \end{cases} \tag{6.8}$$

with $\lambda > 0$,

$$\sigma(x) = \begin{cases} -1 & \text{if } x < 0, \\ 1 & \text{if } x > 0, \end{cases}$$

and

$$w_0(x) = \begin{cases} 0.5 & \text{if } x < 0, \\ -0.5 & \text{if } x > 0. \end{cases} \tag{6.9}$$

Godunov's method (2.10) can be expressed now as follows:

$$w_i^{n+1} = w_i^n + \frac{\Delta t}{\Delta x}\lambda\big(w_{i-1}^n e^{\lambda^{-1}c(\sigma_i - \sigma_{i-1})} - w_i^n\big),$$

while Roe's method upwinding the source term is given by

$$w_i^{n+1} = w_i^n + \lambda\frac{\Delta t}{\Delta x}\left(w_{i-1}^n - w_i^n\right) + c\frac{\Delta t}{\Delta x}\frac{w_{i-1}^n + w_i^n}{2}(\sigma_i - \sigma_{i-1}). \tag{6.10}$$

We have solved problem (6.8) with initial condition (6.9) on the interval $-1 < x < 1$ with $\Delta x = 0.02$ and CFL $= 1$. The results obtained at time $t = 0.4$ have been represented in Figure 3. While Godunov's method gives again the exact solution, Roe's method (6.10) produces an incorrect intermediate state between the discontinuities. Table 5 shows the errors in L^∞ norm for CFL $= 1$ and decreasing values of Δx. In fact, in this case the solutions provided by (6.10) converge to a function which is a solution of (6.8) corresponding to a different definition of the nonconservative product. More precisely, it is a solution of the problem if the family of segments is chosen as the family of paths.

6.3. A relaxation system. We consider the following relaxation system (see [21]):

$$\begin{cases} u_t + v_x = 0, \\ v_t + au_x = \frac{1}{\varepsilon}(f(u) - v), \end{cases} \tag{6.11}$$

with $a > 0$ and $f(u) = \sqrt{a}u$.

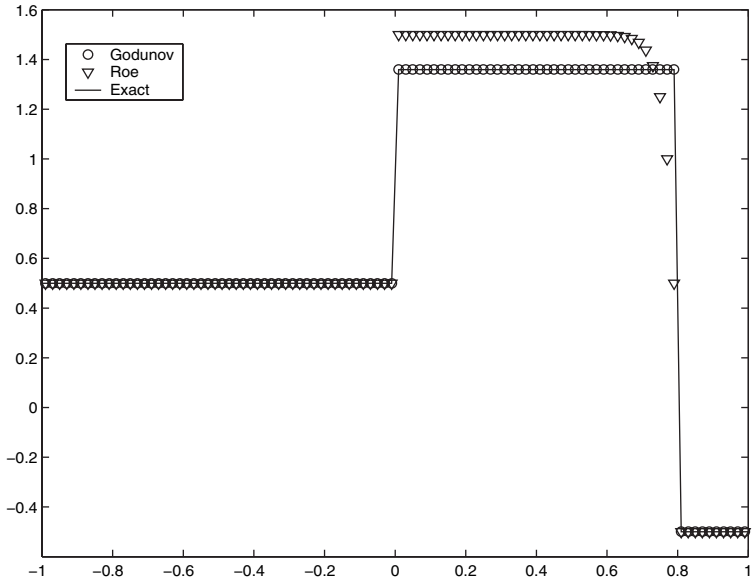


FIG. 3. *Solution of problem* (6.8) *with initial condition* (6.9) *at time* $t = 0.4$.

TABLE 5
Errors in L^∞ *norm for Roe's method* (6.10).

Δx	Roe L^∞ errors
0.2	0.13891
0.05	0.14086
0.02	0.14086
0.005	0.14086
0.002	0.14086

This system can be expressed under the formulation (5.1) with

$$W = \begin{bmatrix} u \\ v \end{bmatrix}, \quad A = \begin{bmatrix} 0 & 1 \\ a & 0 \end{bmatrix}, \quad G(W) = \begin{bmatrix} 0 \\ (f(u) - v)/\varepsilon \end{bmatrix}, \quad \sigma(x) = x.$$

The eigenvalues of the system are $\lambda_1 = -\sqrt{a}$, $\lambda_2 = \sqrt{a}$, and $\lambda_3 = 0$, with associated eigenvectors $R_1 = [1, -\sqrt{a}, 0]^T$, $R_2 = [1, \sqrt{a}, 0]^T$, and $R_3 = [(f(u) - v)/(\varepsilon a), 0, 1]^T$.

The integral curves associated to the zero eigenvalue are of the form

$$v = \overline{v}, \quad u = \frac{v}{\sqrt{a}} + Ce^{\frac{\sigma}{\varepsilon\sqrt{a}}},$$

where $\overline{v}$ and C are real constants.

Godunov's method (2.10) can be written as

$$u_i^{n+1} = u_i^n + \frac{\Delta t}{\Delta x}\big(- \sqrt{a}\ s_{2,i+1/2}^n + \sqrt{a}\ s_{1,i-1/2}^n\big),$$
$$v_i^{n+1} = v_i^n + \frac{\Delta t}{\Delta x}\big(- a\ s_{2,i-1/2}^n - a\ s_{1,i+1/2}^n\big),$$

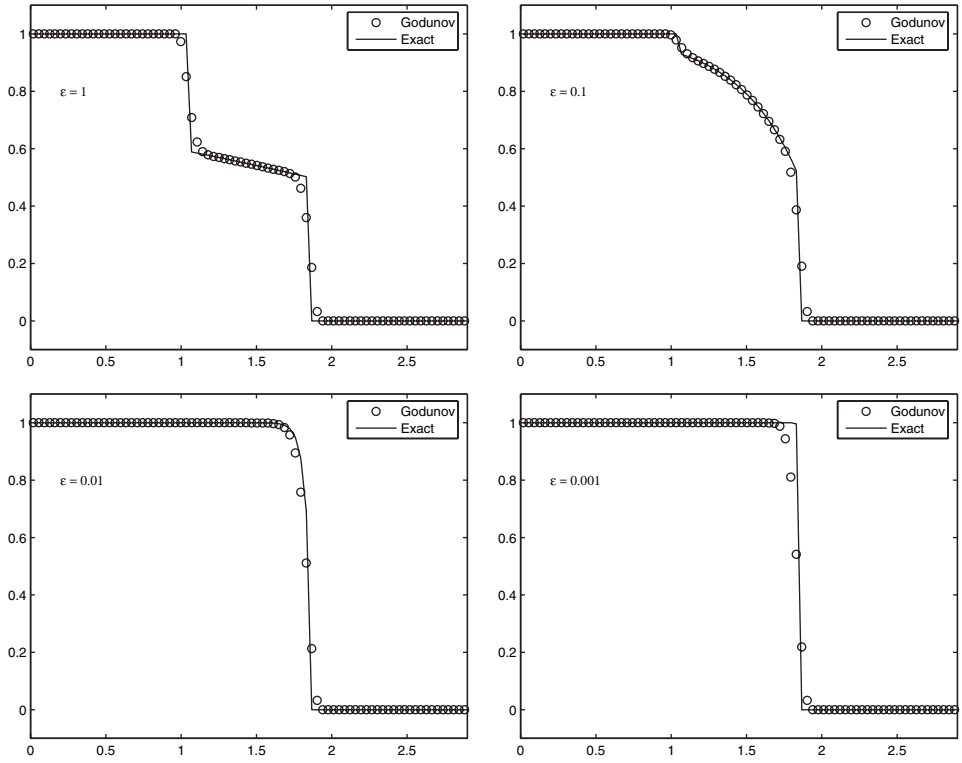


FIG. 4. *Solution u of problem* (6.11) *with initial condition* (6.12) *at time $t = 0.2$. CFL $= 0.9$,* 80 *grid cells.*

where $s^n_{1,i+1/2}$ and $s^n_{2,i+1/2}$ are given by, respectively,

$$s^n_{1,i+1/2} = \frac{1}{2}\left(u^n_{i+1} - \frac{v^n_{i+1}}{\sqrt{a}} e^{-\frac{\Delta x}{\varepsilon\sqrt{a}}}\right) - \frac{1}{2}\left(u^n_i - \frac{v^n_i}{\sqrt{a}}\right),$$

$$s^n_{2,i+1/2} = s^n_{1,i+1/2} - \frac{v^n_i - v^n_{i+1}}{\sqrt{a}}.$$

We consider a test proposed in [21] on the interval $[0, 2.9]$ with the following initial conditions:

$$u(x,0) = \begin{cases} 1 & \text{if } x < 2.9/2, \\ 0 & \text{if } x \geq 2.9/2, \end{cases}$$
$$v(x,0) = 0, \quad x \in [0, 2.9]. \tag{6.12}$$

In Figure 4 we plot the solution at time $t = 0.2$ for different values of the parameter ε. In [21] a finite volume method based on a splitting procedure, called the NM-scheme, was proposed. Numerical results shown there are more diffusive than those obtained here due to the small CFL number required to guarantee stability of the NM-scheme.

6.4. A linearized shallow water system. We consider the system obtained by linearization of the shallow water equations around a steady state $[h_0(x), q_0(x)]^T$

(see [3]):

$$\frac{\partial W}{\partial t} + \frac{\partial}{\partial x}(A(x)W) = G(x)W, \tag{6.13}$$

where $W = [h, q]^T$ represents the perturbation to the steady state and whose components correspond to the fluid depth and discharge, respectively. $A(x)$ and $G(x)$ are given by, respectively,

$$A(x) = \begin{bmatrix} 0 & 1 \\ -\dfrac{q_0(x)^2}{h_0(x)^2} + gh_0(x) & 2\dfrac{q_0(x)}{h_0(x)} \end{bmatrix}, \quad G(x) = \begin{bmatrix} 0 & 0 \\ gH'(x) & 0 \end{bmatrix},$$

where $H(x)$ is the bottom topography and g is the gravity acceleration.

The eigenvalues of the matrix A are $\lambda_1(x) = q_0/h_0 - \sqrt{gh_0}$ and $\lambda_2(x) = q_0/h_0 + \sqrt{gh_0}$; $R_1 = [1, \lambda_1]^T$ and $R_2 = [1, \lambda_2]^T$ are associated eigenvectors. We will assume in what follows that the flow is subcritical, i.e., $\lambda_1(x) < 0 < \lambda_2(x)$ for every x.

Some calculations yield that the stationary solutions of (6.13) are of the form

$$X(x, W^*) = \begin{bmatrix} h \\ q \end{bmatrix} = \begin{bmatrix} \dfrac{C_1 h_0^3 - q_0 h_0 q}{-q_0^2 + g h_0^3} \\ C_2 \end{bmatrix}, \tag{6.14}$$

C_1, C_2 being two arbitrary constants.

The expression of Godunov's method (2.10) is as follows:

$$W_i^{n+1} = W_i^n + \frac{\Delta t}{\Delta x} A(x_i)\big(- s_{2,i-1/2}^n R_{1,i} - s_{1,i+1/2}^n R_{2,i}\big), \tag{6.15}$$

where $s_{1,i+1/2}^n$ and $s_{2,i+1/2}^n$ are given by, respectively,

$$s_{1,i+1/2}^n = \frac{q_{i+1}^n - q_i^n - \lambda_{2,i+1}\beta_{i+1/2}^n}{\lambda_{1,i} - \alpha_{i+1/2}\lambda_{2,i+1}},$$
$$s_{2,i+1/2}^n = \beta_{i+1/2}^n - \alpha\ s_{1,i+1/2}^n,$$

with

$$\alpha_{i+1/2} = c_{i+1}(a_i + b_i\lambda_{1,i}) - d_{i+1}\lambda_{1,i},$$
$$\beta_{i+1/2}^n = h_{i+1}^n - c_{i+1}(a_i h_i^n + b_i q_i^n) + d_{i+1} q_i^n,$$
$$a_i = \frac{-q_{0,i}^2 + g h_{0,i}^3}{h_{0,i}^3},$$
$$b_i = \frac{q_{0,i}}{h_{0,i}^2},$$
$$c_i = \frac{1}{a_i},$$
$$d_i = \frac{q_{0,i} h_{0,i}}{-q_{0,i}^2 + g h_{0,i}^3}.$$

By following [3], we consider the depth function

$$H(x) = \begin{cases} -0.25\big(\cos(\pi(x - 0.5)/0.1) + 1\big) & \text{if } |x - 0.5| < 0.1, \\ 0 & \text{otherwise,} \end{cases} \tag{6.16}$$

TABLE 6

Error in L^1 norm for Godunov's method (6.15) *with initial condition* (6.14) *and* $C_1 = C_2 = 1$ *at time* $t = 1$. *CFL* $= 1$.

No. cells	L^1 error h	L^1 error q
75	8.2970E-015	1.2383E-014
150	1.7715E-014	2.4436E-014
300	3.9978E-014	4.6426E-014
600	8.0624E-014	9.2901E-014

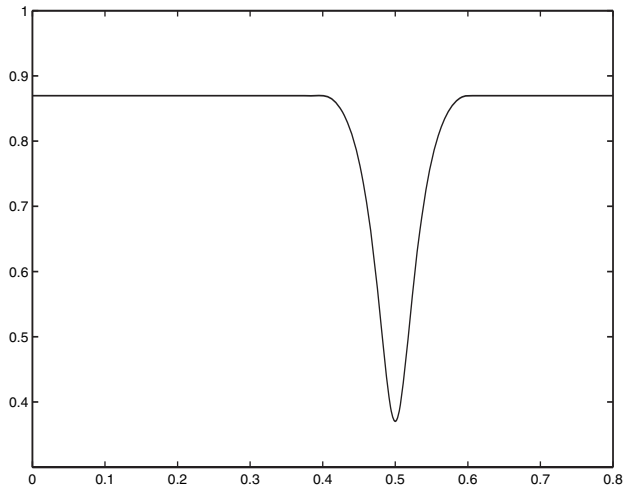


FIG. 5. *Stationary solution* (6.14) *for problem* (6.13) $C_1 = C_2 = 1$.

and consider the steady state corresponding to $q_0 = 0.15$ everywhere and $h_0 = 1$ away from the hump defined in (6.16). The value of g has been fixed to 1.

Table 6 shows that method (6.15) is well-balanced for the stationary solution (6.14) corresponding to $C_1 = C_2 = 1$ on the interval $[0, 0.8]$ (see Figure 5).

Now let us apply a well-balanced high order scheme. In this case, the expression of the scheme is the following:

$$
\begin{aligned}
W_i' = \frac{1}{\Delta x}\Bigg(& A(x_{i-1/2})\big(W^+_{i-1/2,0}(t) - X(x_{i-1/2}, W_i^*(t))\big) \\
& -A(x_{i+1/2})\big(W^-_{i+1/2,0}(t) - X(x_{i+1/2}, W_i^*(t))\big) \\
& + \int_{x_{i-1/2}}^{x_{i+1/2}} G(x) Q_i^t(x)\, dx\Bigg),
\end{aligned} \tag{6.17}
$$

where

$$
\begin{aligned}
W^+_{i+1/2,0} &= W_{i+1} - s_{2,i+1/2} R_{2,i+1/2}, \\
W^-_{i+1/2,0} &= W_i + s_{1,i+1/2} R_{1,i+1/2},
\end{aligned}
$$

with

$$
\begin{aligned}
s_{1,i+1/2} &= \frac{q_{i+1} - q_i - \lambda_{2,i+1/2}(h_{i+1} - h_i)}{\lambda_{1,i+1/2} - \lambda_{2,i+1/2}}, \\
s_{2,i+1/2} &= h_{i+1} - h_i - s_{1,i+1/2},
\end{aligned}
$$

TABLE 7

Error in L^1 norm for the third order well-balanced method (6.17) *with initial condition* (6.14) *and $C_1 = C_2 = 1$ at time $t = 1$. CFL $= 1$.*

No. cells	L^1 error h	L^1 error q
75	1.9123E-14	2.0211E-14
150	3.3747E-14	3.9726E-14
300	7.3356E-14	7.7903E-14
600	1.4558E-13	1.5481E-14

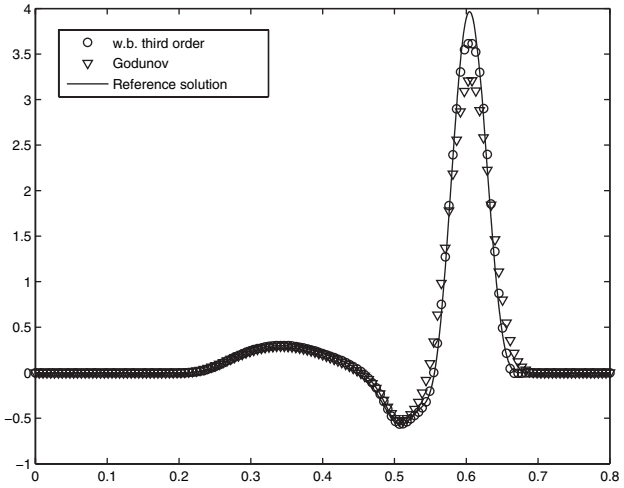


FIG. 6. *Solution of problem* (6.13) *with initial condition* (6.18) *at time $t = 0.3$. CFL $= 0.9$, 150 grid cells. The reference solution has been obtained with scheme* (6.17) *and* 1000 *grid cells.*

and, for any t, $W_i^*(t)$ is calculated so that the average of the stationary solution $X(x, W_i^*(t))$ at the ith cell coincides with $W_i(t)$. In practice, the averages are approximated by means of a quadrature rule with order at least that of the reconstruction operator.

We consider the third order reconstruction operator based on hyperbolas introduced in [22]. The time discretization is performed with a third order TVD explicit Runge–Kutta method (see [12]). Table 7 shows the committed errors solving the stationary solution (6.14) corresponding to $C_1 = C_2 = 1$.

Finally, in Figure 6 we compare the schemes (6.15) and (6.17) for the following initial perturbation proposed in [3]:

$$
\begin{aligned}
h(x,0) &= \begin{cases} \big(1 - \cos(2\pi(x-0.2)/0.15)\big)^2 & \text{if } 0.2 < x < 0.35, \\ 0 & \text{otherwise,} \end{cases} \\
q(x,0) &= (q_0/h_0 + \sqrt{gh_0})h(x,0).
\end{aligned}
\tag{6.18}
$$

The numerical solutions are compared with a reference solution obtained with a finer mesh. Notice the improved resolution obtained with the third order numerical scheme. Results are similar to those shown in [3].

REFERENCES

[1] F. Alcrudo and F. Benkhaldom, *Solution to Riemann problem of the shallow water equation with a bottom step*, Comput. & Fluids, 30 (2001), pp. 643–671.

[2] N. Andrianov, *Performance of numerical methods on the non-unique solution to the Riemann problem for shallow water equations*, Internat. J. Numer. Methods Fluids, 47 (2005), pp. 825–831.

[3] D. Bale, R. J. LeVeque, S. Mitran, and J. Rossmanith, *A wave propagation method for conservation laws and balance laws with spatially varying flux functions*, SIAM J. Sci. Comput., 24 (2002), pp. 955–978.

[4] A. Bermúdez and M. E. Vázquez-Cendón, *Upwind methods for hyperbolic conservation laws with source terms*, Comput. & Fluids, 23 (1994), pp. 1049–1071.

[5] F. Bouchut, *Nonlinear Stability of Finite Volume Methods for Hyperbolic Conservation Laws and Well-Balanced Schemes for Sources*, Birkhäuser, Boston, 2004.

[6] M. J. Castro, J. M. Gallardo, and C. Parés, *Finite volume schemes based on WENO reconstruction of states for solving nonconservative hyperbolic systems. Applications to shallow water systems*, Math. Comp., 75 (2006), pp. 1103–1134.

[7] A. Chinnaya, A. Y. LeRoux, and N. Séguin, *A well-balanced numerical scheme for the approximation of the shallow-water equations with topgraphy: The resonance phenomenon*, Internat. J. Finite Vol. (electronic), 1 (2004), pp. 1–33.

[8] G. Dal Maso, P. G. LeFloch, and F. Murat, *Definition and weak stability of nonconservative products*, J. Math. Pures Appl., 74 (1995), pp. 483–548.

[9] L. Gosse, *A well-balanced scheme using non-conservative products designed for hyperbolic system of conservation laws with source terms*, Math. Models Methods Appl. Sci., 11 (2001), pp. 339–365.

[10] L. Gosse, *Localization effects and measure source terms in numerical schemes for balance laws*, Math. Comp., 71 (2002), pp. 553–582.

[11] L. Gosse and G. Toscani, *An asymptotic-preserving well-balanced scheme for the hyperbolic heat equations*, C. R. Acad. Sci. Paris, Ser. I, 334 (2002), pp. 1–6.

[12] S. Gottlieb and C.-W. Shu, *Total variation diminishing Runge-Kutta schemes*, Math. Comp., 67 (1998), pp. 73–85.

[13] J. M. Greenberg and A. Y. LeRoux, *A well-balanced scheme for the numerical processing of source terms in hyperbolic equations*, SIAM J. Numer. Anal., 33 (1996), pp. 1–16.

[14] J. M. Greenberg, A. Y. LeRoux, R. Baraille, and A. Noussair, *Analysis and approximation of conservation laws with source terms*, SIAM J. Numer. Anal., 34 (1997), pp. 1980–2007.

[15] E. Isaacson and B. Temple, *Convergence of the* 2×2 *Godunov method for a general resonant nonlinear balance law*, SIAM J. Appl. Math., 55 (1995), pp. 625–640.

[16] P. G. LeFloch, *Shock Waves for Nonlinear Hyperbolic Systems in Nonconservative Form*, Preprint 593, Institute for Mathematics and its Application, Minneapolis, 1989.

[17] P. G. LeFloch and M. D. Thanh, *The Riemann problem for fluid flows in a nozzle with discontinuous cross-section*, Commun. Math. Sci., 4 (2003), pp. 763–797.

[18] P. G. LeFloch, *Graph solutions of nonlinear hyperbolic systems*, J. Hyperbolic Differ. Equ., 2 (2004), pp. 643–689.

[19] R. J. LeVeque, *Finite Volume Methods for Hyperbolic Problems*, Cambridge University Press, London, 2002.

[20] P. L. Lions and G. Toscani, *Diffusive limit for finite velocity Boltzmann kinetic models*, Rev. Mat. Iberoamericana, 13 (1997), pp. 473–513.

[21] E. Lorin and V. Seignole, *Convection systems with stiff source terms*, Math. Models Methods Appl. Sci., 13 (2003), pp. 971–1018.

[22] A. Marquina, *Local piecewise hyperbolic reconstuction of numerical fluxes for nonlinear scalar conservation laws*, J. Sci. Comput., 15 (1994), pp. 892–915.

[23] M. L. Muñoz and C. Parés, *Godunov's method for nonconservative hyperbolic systems*, M2AN Math. Model. Numer. Anal., 41 (2007), pp. 169–185.

[24] S. Noelle, N. Pankratz, G. Puppo, and J. R. Natvig, *Well-balanced finite volume schemes of arbitrary order of accuracy for shallow water flows*, J. Comput. Phys., 213 (2006), pp. 474–499.

[25] S. Noelle, Y. Xing, and C.-W. Shu, *High order well-balanced finite volume WENO schemes for shallow water equation with moving water*, J. Comput. Phys., 226 (2007), pp. 29–58.

[26] C. Parés and M. J. Castro, *On the well-balance property of Roe's method for nonconservative hyperbolic systems. Applications to shallow water systems*, M2AN Math. Model. Numer. Anal., 38 (2004), pp. 821–852.

[27] C. Parés, *Numerical methods for nonconservative hyperbolic systems: A theoretical framework*, SIAM J. Numer. Anal., 44 (2006), pp. 300–321.
[28] P. L. Roe, *Upwinding difference schemes for hyperbolic conservation laws with source terms*, in Nonlinear Hyperbolic Problems, Lecture Notes in Math. 1270, C. Carasso, P. A. Raviart, and D. Serre, eds., Springer, New York, 1986, pp. 41–51.
[29] H. Tang, T. Tang, and K. Xu, *A gas-kinetic scheme for shallow-water equations with source terms*, Angew. Math. Phys., 55 (2004), pp. 365–382.
[30] E. F. Toro, *Shock-Capturing Methods for Free-Surface Shallow Flows*, John Wiley & Sons, New York, 2001.
[31] I. Toumi, *A weak formulation of Roe's approximate Riemann solver*, J. Comput. Phys., 102 (1992), pp. 360–373.
[32] B. Van Leer, *Towards the ultimate conservative difference scheme.* V. *A second-order sequel to Godunov's method*, J. Comput. Phys., 32 (1979), pp. 101–136.
[33] Y. Xing and C.-W. Shu, *A new approach of high order well-balanced finite volume WENO schemes and discontinuous Galerkin methods for a class of hyperbolic systems with source terms*, Comm. Comput. Phys., 1 (2006), pp. 100–134.

SIAM J. NUMER. ANAL.
Vol. 46, No. 2, pp. 1040–1059
© 2008 Society for Industrial and Applied Mathematics

MOLLIFIED IMPULSE METHODS FOR HIGHLY OSCILLATORY DIFFERENTIAL EQUATIONS*

J. M. SANZ-SERNA†

Abstract. We introduce a family of impulselike methods for the integration of highly oscillatory second-order differential equations whose forces can be split into a fast part and a slow part. Methods of this family are specified by two weight functions ϕ, ψ; one is used to average positions and the other to mollify the force. When the fast forces are conservative and $\phi = \psi$, the methods here coincide with the mollified impulse methods introduced by García-Archilla, Sanz-Serna, and Skeel. On the other hand, the methods here extend to nonlinear situations a well-known class of exponential integrators introduced by Hairer and Lubich for cases of linear fast forces. A convergence analysis is presented that provides insight into the role played by the processes of mollification and averaging in avoiding order reduction. A simple condition on the weight functions is shown to be both necessary and sufficient to avoid order reduction.

Key words. impulse methods, highly oscillatory problem, Fourier analysis

AMS subject classifications. 65L20, 65M12, 70F10, 70H05

DOI. 10.1137/070681636

1. Introduction.

We consider numerical methods for the integration of multiple-time-scale problems of the form

$$M\frac{d^2}{dt^2}q = f(q) + g(q), \tag{1}$$

where M is a symmetric positive-definite mass matrix and the forces are split into a *soft* or *slow* part g that does not contribute any fast modes to the solution and a *strong* or *fast* part f such that the *reduced* problem

$$M\frac{d^2}{dt^2}q = f(q)$$

possesses highly oscillatory modes (and possibly slow modes as well). These problems arise in many situations including molecular dynamics, astronomy, and the time integration of partial differential equations describing wave phenomena. The fast/slow splitting is of interest in situations where the computational cost of integrating the reduced system is much smaller than that of integrating the full system. For instance, in an N-body problem, the evaluation of the fast force is cheaper than that of the slow force if the former is composed only of next-neighbor interactions and the latter includes interactions between all particles. A second example is provided by cases where the reduced problem can be solved analytically with a cost that is independent (or almost independent) of the step length. In such situations the aim is to sample the slow forces as sparingly as possible and, certainly, at a rate that is not determined by the periods of the fast motions, i.e., by the stiffness of the reduced problem.

*Received by the editors February 2, 2007; accepted for publication (in revised form) October 29, 2007; published electronically March 5, 2008. This research was supported by DGI-MCYT under project MTM2004-07194 cofinanced by FEDER funds.
http://www.siam.org/journals/sinum/46-2/68163.html
†Departamento de Matemática Aplicada, Universidad de Valladolid, 47011 Valladolid, Spain (sanzsern@mac.uva.es, http://hermite.mac.cie.uva.es/sanzserna).

The impulse method was devised to cater to these situations. The time stepping of the impulse method intersperses evaluations of the slow forces at intervals of length h with integrations of the reduced system that are taken to be exact, either because an analytic integration is feasible or because, in a multiple time-step approach [7], [8], [17], a numerical integration is performed with a time step much smaller than h. Unfortunately, the impulse method, which is formally of second-order accuracy, leads only to errors that behave $O(h^2)$ if h is small as compared with the smallest period of the reduced system. An alternative family of methods, the mollified impulse methods, was introduced in [5]; there is one mollified impulse method for each choice of a so-called weight function. Proved in [5] is the fact that, for several choices of weight functions, the mollified impulse method yields errors in q that behave as $O(h^2)$, where the implied constant is independent of the size of the fast forces and depends only on the energy of the solution and on bounds for g and its derivatives. It is then possible to apply the mollified impulse methods with a given time step h to faster and faster problems without impairing the accuracy. This is feasible because, for a given oscillation amplitude, the energy grows with the frequency; therefore, in solutions of limited energy, fast oscillations are present with small amplitudes and need not be accurately resolved. In conclusion, for problems with fast oscillations, the scenario here is then similar to that found in the development of stiff solvers [3], [12] for problems with fast rates of decay: Good stiff solvers perform satisfactorily with step lengths independent of the stiffness of the problem and determined only by the smoothness of the solution, and this is possible because they need not accurately resolve fast components. As in the (nonoscillatory) stiff situation [3], [12], methods for highly oscillatory problems have a conventional order that is brought out in simulations where the step length h is small when compared to the stiffness of the problem and a *stiff order* that manifests itself in the cases where h is not small. In general the stiff order is smaller than the conventional order, and one speaks of an *order reduction*.

In this paper we reconsider the ideas behind the mollified impulse method. In section 2 we introduce a large family of methods, the (ϕ, ψ)-methods, which extends in two ways the family in [5]. In the first place, (ϕ, ψ)-methods are specified by *two* independent weight functions: one, ψ, that defines the so-called mollification and one, ϕ, that is employed in an averaging. On the other hand, our description here of the mollification process depends on the Alekseev–Groebner formula [11] and applies to cases where the fast forces are nonconservative. This should be compared with the methods in [5] that depend on one weight function and apply only to conservative fast forces (Hamiltonian reduced problems). In section 3, it is shown that, if ϕ and ψ are taken to coincide and the reduced problem is Hamiltonian, the family of methods considered here coincides with the mollified methods of [5].

When the fast forces are linear, the reduced problem may in principle be integrated analytically, and we show in section 4 that the (ϕ, ψ)-method can be implemented in terms of two *filters* defined in terms of the Fourier transforms of the weights ϕ and ψ. In this case of linear fast forces, the family of (ϕ, ψ)-methods essentially mimics a by now well known family of exponential integrators first described by Hairer and Lubich [9] (see also, among others, [2] and Chapter XIII in the monograph [10]; an error analysis has been provided in [6]). The relation between weights and filters is explored in further detail in section 7, where filters are characterized via a classical Fourier analysis result of Paley and Wiener.

Sections 5 and 6 contain an analysis of the accuracy of (ϕ, ψ)-methods as applied to problems with linear fast forces. The approach here resembles in a way that in [5], because it is based on the weight functions rather than on the corresponding filters.

The analysis here is then completely different from those in the convergence proofs presented in [6] and (for a particular model problem with only one fast frequency) in Chapter XIII of [10]; both proofs are carried out in terms of the filters. We show how a study in terms of weights makes it possible to identify in a natural way simple conditions on the methods that are both necessary and sufficient to ensure that no order reduction takes place. Essentially, no order reduction is equivalent to the requirement that the translation of the weight function are a basis set for a consistent interpolation scheme. On the other hand, our analysis differs from that in [5] in a number of points that we believe to be of some importance. We cater here to weight functions of arbitrary bounded support; in [5] the support had to be contained in $[-1, 1]$. Our analysis applies to cases where the mollification filter ψ is different from the averaging filter ϕ; this not only means that more methods are covered but also tells apart the effects of the averaging from those of the mollifier. We finally mention that section XIV.2.4 of [10] presents an analysis for the long-average mollified impulse method that covers fast forces more general than those considered by us.

Extensive numerical experiments based on the new methods will be provided in a forthcoming paper.

2. Description of the methods. For numerical work, we rewrite the system (1) in the first-order format

$$\frac{d}{dt}p = f(q) + g(q), \qquad \frac{d}{dt}q = M^{-1}p, \tag{2}$$

with the reduced version

$$\frac{d}{dt}p = f(q), \qquad \frac{d}{dt}q = M^{-1}p. \tag{3}$$

If we denote by P_n and Q_n the numerical approximations to the true solution values of the momenta $p_n = p(t_n)$ and coordinates $q_n = q(t_n)$, respectively, at the step point $t_n = nh$, a step $n \to n + 1$ of any of the methods considered in this paper may be fit into the following composition or split-step pattern.

Kick. Compute an approximation $\bar{G}_n$ to the value of the slow force $g(q_n)$, and then set $P_n^+ = P_n + \frac{h}{2}\bar{G}_n$.

Oscillation. Use the h-flow of the reduced problem (3) to advance from (P_n^+, Q_n) to (P_{n+1}^-, Q_{n+1}). In other words, (P_{n+1}^-, Q_{n+1}) is obtained by integrating, over a time interval of length h, the system (3) with initial conditions (P_n^+, Q_n).

Kick. Set $P_{n+1} = P_{n+1}^- + \frac{h}{2}\bar{G}_{n+1}$.

Of course, the force $\bar{G}_{n+1}$ at the second kick of the current step coincides with the force at the first kick of the next step, so that the methods essentially consist of a sequence of oscillations $(P_n^+, Q_n) \to (P_{n+1}^-, Q_{n+1})$ followed by kicks $P_{n+1}^- = P_{n+1}^+ + h\bar{G}_{n+1}$; it is not necessary to compute the vector P_{n+1} unless output at t_{n+1} is required.

Different methods within the class considered here differ in the way in which the forces $\bar{G}_n$ are computed. The simplest and earliest choice, that of the original impulse method, is given by $\bar{G}_n = g(Q_n)$ and, unfortunately, suffers from an order reduction in p from 2 to 0 and in q from 2 to 1 in the stiff oscillatory situations considered here [5]. An improvement that may ensure (stiff) order 1 in p and (stiff) order 2 in q was presented in [5]. The family of mollified impulse methods introduced there have

$$\bar{G}_n = \mathcal{M}(Q_n, h)g(\mathcal{A}(Q_n, h)), \tag{4}$$

where, given Q_n and h, $\mathcal{A}(Q_n, h)$ represents an average of values of q and $\mathcal{M}$ is a so-called mollifier matrix. The use of $\mathcal{A}$ avoids [5] the dangers associated with sampling at grid points a quickly varying $g(q(t))$, and the role of $\mathcal{M}$ is to mimic the way in which extra external forces contribute to build up the momentum in highly oscillatory problems (see the analysis and counterexamples in sections 5 and 6 below).

As discussed in [5], there is much freedom within the class of mollified impulse methods, both in how $\mathcal{M}$ and $\mathcal{A}$ may be mathematically specified and in the analytic or numerical technique employed to actually compute them. Here we add to this freedom by further enlarging the class of possible choices in (4).

Before we proceed, we need some notation. We denote by $\mathcal{P}(p, q, t)$, $\mathcal{Q}(p, q, t)$ the flow of the reduced system (3), so that $p(t) = \mathcal{P}(p_0, q_0, t)$, $q(t) = \mathcal{Q}(p_0, q_0, t)$ is the value at time t of the solution of (3) with initial conditions $p(0) = p_0$, $q(0) = q_0$. Thus $P_{n+1}^- = \mathcal{P}(P_n^+, Q_n, h)$, and $Q_{n+1} = \mathcal{Q}(P_n^+, Q_n, h)$. For compactness, it is useful to combine P_n and Q_n into a single vector $Y_n = (P_n, Q_n)$ and similarly write $y(t) = (p(t), q(t))$, $\mathcal{Y} = (\mathcal{P}, \mathcal{Q})$, etc. If d represents the number of degrees of freedom in the problem, we will also use the $(2d) \times (2d)$ Jacobian matrix $\mathcal{Y}'(\alpha, t)$ of $\mathcal{Y}(\alpha, t)$ with respect to its argument α, which can be found by integrating the variational equation

$$\frac{\partial}{\partial t}\mathcal{Y}'(\alpha, t) = \begin{bmatrix} 0 & f'(\mathcal{Q}(\alpha, t)) \\ M^{-1} & 0 \end{bmatrix} \mathcal{Y}'(\alpha, t), \tag{5}$$

with initial condition $\mathcal{Y}'(\alpha, 0) = I_{2d}$. It is clear that $\mathcal{Y}'$ possesses a natural block structure corresponding to the partition of y:

$$\mathcal{Y}' = \begin{bmatrix} \mathcal{P}_p & \mathcal{P}_q \\ \mathcal{Q}_p & \mathcal{Q}_q \end{bmatrix}.$$

Each of the methods considered here is defined by two *weight* functions. For the purpose of this paper, a weight function is a bounded, integrable real-valued function $\chi(t)$ that is assumed to be even $\chi(-t) \equiv \chi(t)$ and to satisfy

$$\int_{-\infty}^{\infty} \chi(s) ds = 1. \tag{6}$$

(Note that we do not require $\chi \geq 0$.)

We may now describe the new methods. As in [5], to construct $\mathcal{A}$, we start by choosing a weight function ϕ and then define

$$\mathcal{A}(Q_n, h) = \frac{1}{h} \int_{-\infty}^{\infty} q^*(t) \phi\left(\frac{t}{h}\right) dt = \int_{-\infty}^{\infty} q^*(hs) \phi(s)\, ds, \tag{7}$$

where $q^*(t)$ is obtained by solving the reduced problem (3) with initial conditions $q = Q_n$, $p = 0$, i.e., $q^*(t) = \mathcal{Q}(0, Q_n, t)$. Since q^* is an even function of t, the integrals in (7) are in practice replaced by twice their value over $(0, \infty)$.

To define the mollifier, we choose a second weight function ψ, and, at each time step, we find a matrix-valued function Y' of the independent variable t

$$Y'(t) = \begin{bmatrix} P_p(t) & P_q(t) \\ Q_p(t) & Q_q(t) \end{bmatrix}$$

by integrating the variational problem ($q^*(t)$ is as above)

$$\frac{d}{dt}Y' = \begin{bmatrix} 0 & f'(q^*(t)) \\ M^{-1} & 0 \end{bmatrix} Y', \tag{8}$$

with initial condition $Y'(0) = I_{2d}$, and then set

$$\mathcal{M}(Q_n, h) = \frac{1}{h}\int_{-\infty}^{\infty} R(t)\psi\left(\frac{t}{h}\right)\,dt = \int_{-\infty}^{\infty} R(hs)\psi(s)\,ds, \tag{9}$$

where $R(t)$ is the upper left $d \times d$ block of the inverse matrix $Y'(t)^{-1}$, i.e.,

$$R(t) = [I_d, 0_d]Y'(t)^{-1}[I_d, 0_d]^T.$$

Note that q^*, Y', R depend on Q_n, but this dependence is not reflected in the notation, and that again, in practice, the required integrals may be replaced by twice their value over $(0, \infty)$. (Reversing the sign of t in (8) shows that the $(1, 1)$ and $(2, 2)$ diagonal blocks of $Y'(t)$ are even functions of t, while the off-diagonal blocks are odd functions of t. The inverse $Y'(t)^{-1}$ inherits this odd/even structure, and therefore $R(t)$ is even.)

To sum up, for any pair of weight functions ϕ and ψ, we have introduced a numerical scheme, to be referred to as the (ϕ, ψ)-*method*, that uses the kick/oscillation/kick structure with kicking forces $\bar{G}_n$ given by (4) and $\mathcal{A}$ and $\mathcal{M}$ defined in (7) and (9). We will stretch somehow the notation and understand that (δ, ψ), with δ the standard Dirac function, refers to a method with mollification but no averaging; (ϕ, δ) refers to a method with averaging but no mollification, and (δ, δ) is the original impulse method. All methods are symmetric or time-reversible [16], [10] and therefore convergent of the second order (in nonstiff situations). Furthermore, since individual kicks and oscillations preserve volume in phase space, all methods are volume-preserving.

The methods require at each step two integrations of the reduced problem: one to carry out the oscillation $(P_n^+, Q_n) \to (P_{n+1}^-, Q_{n+1})$ and one to compute the auxiliary $q^*(t)$ that features in (7) and (9). Furthermore the latter is to be carried out in tandem with the integration of the variational problem (8); see [5]. In some applications (see section 4 below) these integrations may be performed analytically, and in other cases one has to resort to a numerical technique with a small time step. If numerical integration is used, the weight functions ϕ and ψ have to possess bounded support so that the auxiliary problems are integrated only over a bounded time interval. The integrals in (7) and (9) can be computed along the integrations for q^* and Y' if one appends extra unknown functions b, c that evolve in time according to $db/dt = \phi(t/h)q^*(t)$ and $dc/dt = \psi(t/h)R(t)$; cf. [5].

We end this section by presenting the motivation for the definition of the mollifier. The original impulse method can be expressed [18] in terms of the Dirac delta function as an exact integration of the approximation

$$\frac{d^2}{dt^2}q = f(q) + \sum_n \delta\left(\frac{t - t_n}{h}\right) g(Q_n),$$

and the idea behind the mollification is to employ less abrupt versions [5]

$$\frac{d^2}{dt^2}q = f(q) + \sum_n \psi\left(\frac{t - t_n}{h}\right) G_n^*, \tag{10}$$

where G_n^* is the force $g(\mathcal{A}(Q_n, h))$ to be mollified.

In order to incorporate the forces $\psi((t-t_n)/h)G_n^*$ into the solution of the reduced problem, we use here the Alekseev–Groebner (AG)/nonlinear variation of constants formula [11] as distinct from the approach used in section 5 of [5], which is limited

to the case where the reduced system is Hamiltonian. According to the AG formula, solutions of (2) can be expressed in terms of the reduced flow $\mathcal{Y}$ as

$$y(t_b) = \mathcal{Y}(y(t_a), t_b - t_a) + \int_{t_a}^{t_b} \mathcal{Y}'(y(s), t_b - s) \begin{bmatrix} g(y(s)) \\ 0 \end{bmatrix} ds.$$

Assume then that, at the step $n-1 \to n$, we have performed the corresponding oscillation and we wish to additionally incorporate the effect of $\psi((t-t_n)/h)G_n^*$, acting while $t \le t_n$. (The effect of the force acting in the interval $t \ge t_n$ is incorporated at the first kick of the next step in a symmetric way.) We set $t_b = t_n$ and let $t_a \downarrow -\infty$, to conclude that

$$\left(\int_{-\infty}^{t_n} \mathcal{Y}'(y(s), t_n - s) \psi \left(\frac{s - t_n}{h} \right) ds \right) \begin{bmatrix} G_n^* \\ 0 \end{bmatrix} \tag{11}$$

is the change in solution caused by the extra force. Therefore, the extra momentum is

$$\left(\int_{-\infty}^{t_n} \mathcal{P}_p(y(s), t_n - s) \psi \left(\frac{s - t_n}{h} \right) ds \right) G_n^*,$$

and we would like to add this expression to P_n^- to obtain P_n. There are two difficulties to be tackled. The first stems from the fact that (11) contains the solution $y(t)$ being sought; it is readily circumvented by using an approximation to $y(t)$. The second is that present in (11) is a *family* of Jacobian matrices, one for each vector $y(s)$, so that, in principle, we may fear that, for each s, in order to find $\mathcal{Y}'(y(s), t_n - s)$ we would have to solve a different initial value problem for (5). However, differentiation of the identity $\mathcal{Y}(\mathcal{Y}(\alpha, t), -t) \equiv \alpha$ (group property of the flow) yields for the corresponding Jacobian matrices the well-known relation

$$\mathcal{Y}'(\mathcal{Y}(\alpha, t), -t) = \mathcal{Y}'(\alpha, t)^{-1}. \tag{12}$$

We then approximate the unknown $y(s)$ by $\mathcal{Y}((0, Q_n), s - t_n)$, use the relation (12) with $\alpha = (0, Q_n)$, $t = s - t_n$), and rewrite (11) as

$$\left(\int_{-\infty}^{t_n} \mathcal{Y}'((0, Q_n), s - t_n)^{-1} \psi \left(\frac{s - t_n}{h} \right) ds \right) \begin{bmatrix} G_n^* \\ 0 \end{bmatrix},$$

an expression that, after shifting the integration variable and multiplying by 2 to account for the effect of ψG_n^* acting while $t \ge t_n$, leads to (9).

3. Conservative fast forces. When the fast forces are conservative, $f(q) = -\nabla W(q)$, with W a scalar potential function, the system (3) is Hamiltonian [16], [10], [14], and the symmetries associated with Hamiltonian flows allow a simplification of the mollification procedure. The key observation is as follows.

LEMMA 1. *Assume that the reduced problem* (3) *is Hamiltonian. Then*

$$\mathcal{Y}'(\alpha, t)^{-1} = J^{-1} \mathcal{Y}'(\alpha, t)^T J,$$

where J is the matrix that defines the canonical symplectic form of the Hamiltonian formalism. In particular, the upper left block of the inverse of $\mathcal{Y}'(\alpha, t)$ coincides with $\mathcal{Q}_q(\alpha, t)^T$ (transposed lower right block).

Proof. Due to the symplecticness of Hamiltonian flows, $K = \mathcal{Y}'(\alpha, t)$ satisfies the relation $K^T J K = J$, or $K^{-1} = J^{-1} K^T J$. ∎

After this result, in the Hamiltonian case, the matrix R averaged in (9) can be found without inverting the $(2d) \times (2d)$ Jacobian Y'; instead we set

$$R(t) = Q_q(t)^T, \tag{13}$$

with Q_q found from solving

$$\frac{d}{dt}\begin{bmatrix} Q_p(t) \\ Q_q(t) \end{bmatrix} = \begin{bmatrix} 0 & f'(q^*(t)) \\ M^{-1} & 0 \end{bmatrix} \begin{bmatrix} Q_p(t) \\ Q_q(t) \end{bmatrix},$$

with initial condition $Q_p(t) = 0_d$, $Q_q(t) = I_d$. The left upper and lower blocks of Y' have been discarded from the variational equation (8) as they are not coupled to the lower right block Q_q to be used in (13).

The mollified impulse method was conceived by Skeel [5] in a context where much relevance was attached to the symplectic character [16], [10], [14] of the numerical schemes considered, and, for this reason, the paper [5] deals only with the case of conservative fast forces. In [5] the weight function ϕ may be freely chosen, but the mollifier is determined by

$$\mathcal{M}(Q_n, h) = \mathcal{A}'(Q_n, h)^T \tag{14}$$

(a prime denotes differentiation with respect to Q), a choice that ensures that, whenever the force g is conservative with $g(q) = -\nabla U(q)$, the force $\bar{G}_n$ used in the kicks is itself conservative as the negative gradient of the averaged potential $U(\mathcal{A}(Q_{n+1}, h))$. In the present work, fast forces need not be conservative, and there are two weight functions ϕ and ψ to be chosen. Nevertheless, if the fast forces are conservative and if ψ is chosen to coincide with ϕ, the (ϕ, ϕ)-methods defined here coincide with the methods introduced in [5]: Differentiating the average of q^* and transposing as in (14) is equivalent to averaging the transposed of the corresponding Jacobian as in (9) and (13).

4. Linear fast forces. We now address the case where the fast forces are linear: $f(q) = -Sq$, with S a stiffness matrix. We do not assume S to be symmetric, but, in order to deal with oscillatory problems, we suppose that $M^{-1/2} S M^{-1/2}$ possesses only real nonnegative eigenvalues and can be diagonalized. The material below would remain valid even if the conditions on S may be relaxed considerably; for instance, $M^{-1/2} S M^{-1/2}$ could be allowed to have complex eigenvalues close to the positive real axis. An abstract situation where S is the generator of a cosine operator [1] in a Banach space could also be catered to, but we will not concern ourselves with such extensions.

Under the hypothesis above, there exists a unique (in general, nonsymmetric) matrix Ω such that $\Omega^2 = M^{-1/2} S M^{-1/2}$ and all eigenvalues of Ω are nonnegative; the solution flow of the reduced problem is given in terms of Ω by $y(t) = \mathcal{R}(t) y(0)$, where $\mathcal{R}(t)$ is the rotation matrix

$$\mathcal{R}(t) = \begin{bmatrix} M^{1/2} \cos t\Omega \, M^{-1/2} & -M^{1/2} \Omega \sin t\Omega \, M^{1/2} \\ M^{-1/2} \Omega^{-1} \sin t\Omega \; M^{-1/2} & M^{-1/2} \cos t\Omega \, M^{1/2} \end{bmatrix}$$

or, in the spirit of the theory of cosine operators [1],

$$\mathcal{R}(t) = \begin{bmatrix} M^{1/2} C(t) M^{-1/2} & M^{1/2} \frac{d}{dt} C(t) M^{1/2} \\ M^{-1/2} \int_0^t C(s) ds \; M^{-1/2} & M^{-1/2} C(t) M^{1/2} \end{bmatrix},$$

with $C(t) = \cos t\Omega$. The second formula for $\mathcal{R}$ possesses the merit of deemphasizing the dependence of the material presented here on special properties of the trigonometric functions; the occurrence of the derivative and the integral is a direct consequence of the fact that p is essentially the time derivative of q. The upper left block of $\mathcal{R}(t)$ is bounded by a constant, and therefore the lower left block possesses an estimate of the form $O(t)$; the analysis below is based on these facts. On the other hand, from the group properties of the flow, $\mathcal{R}(t)\mathcal{R}(s) = \mathcal{R}(t+s)$, so that, in particular, $\mathcal{R}(-t)$ is the inverse of $\mathcal{R}(t)$; these properties are not contingent on trigonometric identities and will also be crucial below. It is perhaps of some interest to point out that, in agreement with section 3, if the problem is Hamiltonian, then Ω is symmetric and the upper left block of $\mathcal{R}(-t)$ coincides with the transpose of the lower right block of $\mathcal{R}(t)$.

The definition (7) now yields, since $\mathcal{R}(t)^{-1} = \mathcal{R}(-t)$,

$$\begin{aligned}\mathcal{A}(Q_n, h) &= M^{-1/2}\left(\frac{1}{h}\int_{-\infty}^{\infty}\cos t\Omega\,\phi\left(\frac{t}{h}\right)\,dt\right)M^{1/2}\,Q_n \\ &= M^{-1/2}\left(\int_{-\infty}^{\infty}\cos sh\Omega\,\phi(s)\,ds\right)M^{1/2}\,Q_n,\end{aligned}$$

an expression that, after introducing the Fourier transform of the even function ϕ

$$\hat{\phi}(\omega) = \int_{-\infty}^{\infty}\exp(-i\omega t)\phi(t)\,dt = \int_{-\infty}^{\infty}\cos\omega t\phi(t)\,dt, \tag{15}$$

can be rewritten as

$$\mathcal{A}(Q_n, h) = \mathcal{A}_h Q_n = \left(M^{-1/2}\hat{\phi}(h\Omega)M^{1/2}\right)Q_n.$$

Similarly, (9) now reads

$$\begin{aligned}\mathcal{M}_h &= M^{-1/2}\left(\frac{1}{h}\int_{-\infty}^{\infty}\cos t\Omega\,\psi\left(\frac{t}{h}\right)\,dt\right)M^{1/2} \\ &= M^{-1/2}\left(\int_{-\infty}^{\infty}\cos sh\Omega\,\psi(s)\,ds\right)M^{1/2}\end{aligned}$$

or

$$\mathcal{M}_h = M^{-1/2}\hat{\psi}(h\Omega)M^{1/2},$$

with

$$\hat{\psi}(\omega) = \int_{-\infty}^{\infty}\exp(-i\omega t)\psi(t)\,dt = \int_{-\infty}^{\infty}\cos\omega t\psi(t)\,dt, \tag{16}$$

and we conclude that the kicking force is given by

$$\bar{G}_n = M^{-1/2}\hat{\psi}(h\Omega)M^{1/2}g\left(M^{-1/2}\hat{\phi}(h\Omega)M^{1/2}Q_n\right). \tag{17}$$

For cases of linear fast forces where the stiffness matrix S is symmetric, a number of recent contributions starting from [9] have considered a family of kick-oscillate-kick algorithms whose kicking force includes *four* independent filtering functions (see

Chapter XIII of [10], [6], and references therein). If the attention is restricted to symmetric (time reversible) methods within that family, then the number of independent filters is reduced to two and the format for the kicking force is precisely (17) (see section 7).

Some final remarks: The real filter functions $\hat{\phi}$, $\hat{\psi}$ are even. Since weight functions are assumed throughout to satisfy the normalizing condition (6), we have $\hat{\phi}(0) = \hat{\psi}(0) = 1$. Furthermore the integrability of the weights implies that $\hat{\phi}$ and $\hat{\psi}$ are continuous, vanish at ∞, and possess the estimates $\|\hat{\phi}\|_\infty \leq \|\phi\|_1$, $\|\hat{\psi}\|_\infty \leq \|\psi\|_1$. If in particular ϕ or ψ are nonnegative, then $\|\hat{\phi}\|_\infty = 1$ or $\|\hat{\psi}\|_\infty = 1$. When ϕ or ψ are the Dirac function, then the Fourier transform is the unit function and the average or mollification filters become the identity, as they should.

5. Analysis: Mollification. Throughout this section and the next, we assume that the mass matrix M is the identity. What follows may easily be extended to the case $M \neq I$ (but the formulas soon become cluttered with powers of M); alternatively, a problem with $M \neq I$ may be reduced to the case $M = I$ by a linear change of variables [5]. Similarly, we assume that the stiffness matrix S is symmetric and positive-semidefinite, so that, in the Euclidean norm, $\|\cos t\Omega\| \leq 1$; again the analysis below may be easily extended to cover the more relaxed hypotheses on S considered in section 4, but this brings a bound for $\|\cos t\Omega\|$ as an additional constant in many of the formulas.

The analysis begins with the following result, which is Lemma 3 of [5] (but there is a square root missing from some formulas in that paper). To simplify the writing, we denote by $\bar{g}(\cdot) = \mathcal{M}_h g(\mathcal{A}_h \cdot)$ the function that provides the kicking force and furthermore set $\bar{g}_n = \bar{g}(q_n)$. Similarly we set $g^*(\cdot) = g(\mathcal{A}_h \cdot)$ and $g^*_n = g^*(q_n)$. Furthermore we denote by L_0, L_1, and L_2 a bound for g, a Lipschitz constant for g, and a Lipschitz constant for the derivative of g, respectively, when they exist; it follows from the boundedness of the filters that, if g is Lipschitz continuous, $\bar{g}$ and g^* are then also Lipschitz continuous with constants $\bar{L}_1 = \|\psi\|_1 L_1 \|\phi\|_1$ and $L^* = L_1 \|\phi\|_1$, respectively.

THEOREM 1. *Assume that the problem* (2) *with linear fast forces satisfies the hypotheses above, with g Lipschitz continuous, and is solved with $P_0 = p(0)$ and $Q_0 = q(0)$ by means of a (ϕ, ψ)-method (ϕ and/or ψ may be taken to be δ). Then the global error satisfies*

$$\|Q_n - q_n\| \leq \cosh(t_n \sqrt{\bar{L}_1}) \cdot \max_{1 \leq j \leq n} \|\sigma_{q,j}\|$$

and

$$\|P_n - p_n\| \leq \|\sigma_{p,n}\| + \bar{L}_1 t_n \cosh(t_n \sqrt{\bar{L}_1}) \cdot \max_{1 \leq j \leq n} \|\sigma_{q,j}\|,$$

where $\sigma_{p,n}$ $\sigma_{q,n}$ are the quadrature errors

$$\begin{bmatrix} \sigma_{p,n} \\ \sigma_{q,n} \end{bmatrix} = \sum_{j=0}^{n} h 1_j \begin{bmatrix} \cos(t_n - t_j)\Omega \\ \Omega^{-1} \sin(t_n - t_j)\Omega \end{bmatrix} \bar{g}_j - \int_0^{t_n} \begin{bmatrix} \cos(t_n - t)\Omega \\ \Omega^{-1} \sin(t_n - t)\Omega \end{bmatrix} g(q(t))\, dt \tag{18}$$

(1_j is defined to be 1 except for $1_0 = 1_n = 1/2$).

This theorem, based on a Gronwall lemma argument, provides a basic stability result for all of the methods in the class considered in this paper. To be noted is

the fact that the truncation error perpetrated at each step does not explicitly feature in the global error bound; instead the stepwise truncation errors up to t_n have been compounded into a single quadrature error σ_n. Integrated forms of the truncation error may lead to sharper error bounds than a stepwise approach whenever error cancellations are to be expected. This is an old theme in the analysis of numerical methods that, for ordinary differential equations, goes back at least to Spijker's thesis (see the discussion in [15]).

If $\bar{g}_n = g(Q_n)$ (original impulse method), σ coincides with the error in the trapezoidal quadrature rule. In the highly oscillatory scenario considered here where $h\Omega$ is not assumed to be small, the function $\cos(t_n - t)\Omega$ has an unbounded first derivative, and one cannot expect to derive $O(h)$ error bounds for σ_p that are uniform in Ω. A counterexample is provided by the case of a scalar harmonic oscillator of frequency ω subject to a constant external force g with $h\omega = 2\pi$ (the situation presented in [5] to prove that the global error—not the quadrature error—of the impulse method has stiff order 0 in the p variables). In the p component, the integral in (18) vanishes, while the trapezoidal quadrature rule takes the $O(1)$ value $t_n g$. For the q component the situation is not as bad: $\Omega^{-1}\sin(t_n - t)\Omega$ has a bounded first derivative, and standard theory leads to $O(h)$ error bounds for σ_q and hence for $Q_n - q_n$ in the application of the original impulse method to stiff oscillatory situations.

The proof of the next lemma essentially reverses the construction of the mollifier in section 2: There we incorporated into the kicking force $\bar{G}_n$ the "distributed" force $\psi((t - t_n)/h)G_n^*$ (see (10)) by way of the mollification matrix; we now recover the distributed force from the mollification matrix. The meaning of the lemma is clear: While the impulse method approximates the oscillatory integral in (18) by interpolation of the whole integrand, mollified methods interpolate only the soft force $g(q(t))$. This is precisely the idea behind Filon-type methods for the quadrature of oscillatory functions; see, e.g., [13].

LEMMA 2. *The quadrature error* (18) *may be rewritten as*

$$\begin{bmatrix} \sigma_{p,n} \\ \sigma_{q,n} \end{bmatrix} = \int_{-\infty}^{\infty} \begin{bmatrix} \cos(t_n - t)\Omega \\ \Omega^{-1}\sin(t_n - t)\Omega \end{bmatrix} \sum_{j=0}^{n} 1_j \psi\left(\frac{t - t_j}{h}\right) g_j^*\, dt - \int_0^{t_n} \begin{bmatrix} \cos(t_n - t)\Omega \\ \Omega^{-1}\sin(t_n - t)\Omega \end{bmatrix} g(q(t))\, dt. \tag{19}$$

Furthermore, for a mollified method (ϕ, ψ) (ϕ may be the Dirac function) in which ψ has bounded support and vanishes for $|\ t\ | > \mu > 0$, the quadrature error is of the form

$$\begin{bmatrix} \sigma_{p,n} \\ \sigma_{q,n} \end{bmatrix} = \int_0^{t_n} \begin{bmatrix} \cos(t_n - t)\Omega \\ \Omega^{-1}\sin(t_n - t)\Omega \end{bmatrix} I(t)\, dt + \beta_n, \tag{20}$$

where $I(t)$ is the interpolation error

$$I(t) = \left(\sum_{j=0}^{n} \psi\left(\frac{t - t_j}{h}\right) g_j^*\right) - g(q(t)) \tag{21}$$

and β_n represents boundary effects and can be estimated as

$$\|\beta_n\| \le 2\mu(1 + t_n^2)^{1/2}\|\psi\|_1 L_0 h.$$

Proof. By the definition of the mollifier and since the integral of $\sin(t_j - t)\Omega\psi((t - t_j)/h)$ vanishes, the jth term T_j of the sum in (18) can be written as

$$\begin{aligned} h1_j\mathcal{R}(t_n - t_j) &\begin{bmatrix} \hat{\psi}(h\Omega)g_j^* \\ 0 \end{bmatrix} \\ &= 1_j \int_{-\infty}^{\infty} \mathcal{R}(t_n - t_j) \begin{bmatrix} \cos(t_j - t)\Omega \\ \Omega^{-1}\sin(t_j - t)\Omega \end{bmatrix} \psi\left(\frac{t - t_j}{h}\right) g_j^*\, dt \\ &= 1_j \int_{-\infty}^{\infty} \begin{bmatrix} \cos(t_n - t)\Omega \\ \Omega^{-1}\sin(t_n - t)\Omega \end{bmatrix} \psi\left(\frac{t - t_j}{h}\right) g_j^*. \end{aligned}$$

Away from the boundary, when j is such that $0 \le t_j - \mu h$, $t_j + \mu h \le t_n$, the last integral may be taken between the limits 0 and t_n and incorporated into the integral in (18). There are at most 2μ boundary values of j (μ near 0 and μ near t_n). For those values, the corresponding term T_j is incorporated only in part to the integral. The lemma follows after noticing that each T_j can be bounded by $h(1 + t_n^2)^{1/2}\|\psi\|_1 L_0$. ∎

After Theorem 1 and Lemma 2, global error estimates depend on bounds for the interpolation error (21), and we now turn to these. It is well known that a minimum requirement for any linear interpolation scheme is that it reproduces with no error the constant functions:

$$\sum_{j=-\infty}^{\infty} \psi(t - j) \equiv 1, \tag{22}$$

a condition that was featured in [5]. We recall from the standard theory behind the Poisson summation formula that the 1-periodic function

$$\Psi(t) = \sum_{j=-\infty}^{\infty} \psi(t - j)$$

(the sum is well defined if ψ has bounded support) has a Fourier series

$$\Psi(t) = \sum_n c_n \exp(i2\pi nt)$$

whose coefficients are values of the Fourier transform of ψ:

$$c_n = \int_0^1 \exp(-i2\pi nt)\Psi(t)\, dt = \int_{-\infty}^{\infty} \exp(-i2\pi nt)\psi(t)\, dt = \hat{\psi}(2\pi n).$$

Therefore (22) is equivalent to the condition

$$\hat{\psi}(2\pi n) = 0, \quad n = \pm 1, \pm 2, \ldots \tag{23}$$

that appears in the analysis of [10], [6]. We now prove that these conditions are essentially necessary and sufficient to achieve stiff order of convergence 1.

THEOREM 2. *With the hypotheses of Theorem* 1, *assume that* g *has a bounded derivative and that the weight function* ψ *has bounded support and satisfies* (22) *or* (23) *(*ϕ *may be the Dirac function). Then the global error possesses a bound*

$$\|P_n - p_n\| + \|Q_n - q_n\| \le Ch,$$

where the constant C depends on ψ, ϕ, t_n, L_0, and L_1 and also on a bound E for the reduced energy of the true solution

$$E = \max_{-\mu h \le t \le t_n + \mu h} \left(\frac{1}{2}\|p(t)\|^2 + \frac{1}{2}\|\Omega q(t)\|^2 \right).$$

Conversely, if the (ψ, ϕ)-method, where ψ is boundedly supported and ϕ may be δ, possesses a global error bound of this form, then (22) *and* (23) *hold true.*

Proof. Consider a method such that $\hat{\psi}(2k\pi) \neq 0$ for $k \neq 0$ an integer. In the case of a scalar harmonic oscillator with frequency $\omega = 2k\pi/h$ forced by a constant g, the global errors can be found explicitly, and $P_n - p_n$ turns out to possess an $O(1)$ behavior as $h \to 0$.

Let us next derive the error bound when $\phi = \delta$. It is standard to show (by approximating $m(t)$ by a constant) that, when interpolating a function m, the interpolation error

$$\left(\sum_{j=\infty}^{\infty} \psi\left(\frac{t - t_j}{h} \right) m(t_j) \right) - m(t)$$

can be bounded by $K_\psi K_1 h$, where K_ψ depends only on ψ and K_1 is a bound for the derivative $(d/dt)m(t)$. In our situation, where $m(t) = g(q(t))$ and $(d/dt)m(t) = g'p$, this yields an upper bound of the form $K_\psi L_1 \sqrt{E} h$ for the interpolation error in (21) for t away from the boundary $\mu h \le t \le t_n - \mu h$. The remaining intervals $0 \le t \le \mu h$ and $t_n - \mu h \le t \le t_n$ have length $O(h)$, and in them $I(t)$ is bounded. This, along with Lemma 2 and Theorem 1, yields the error bound for the global error.

The effect of an averaging operation with $\phi \neq \delta$ is to perturb the interpolation by bringing in approximate values $g_n^* = g(\mathcal{A}_h q_n)$ for the function values $g_n = g(q_n)$. Since $(I - \mathcal{A}_h)(h\Omega)^{-1}$ can be bounded independently of h and Ω (see the proof of (28) below), the differences $g_n^* - g_n$ are $O(h)$, with the implied constant depending on L_1, ϕ, and a bound for $\|\Omega q_n\|$. Then a global error bound $O(h^2)$ also holds in the presence of averaging. ∎

6. Analysis: Averaging. In this section we prove that, for suitable averaging operators, the (ϕ, ψ)-method yields $O(h^2)$ errors in q uniformly in the stiffness of the problem. After Theorem 2, we consider only situations where ψ is of bounded support and (22) holds.

The analysis begins by observing that, from the Alekseev–Groebner variation of constants formula:

$$p(t) = \mathcal{P}(p(0), q(0)) + \int_0^t \cos s\Omega \, g(q(s))\, ds,$$

and

$$q(t) = \mathcal{Q}(p(0), q(0)) + \int_0^t \Omega^{-1} \sin s\Omega \, g(q(s))\, ds,$$

or, after integration by parts,

$$q(t) = \mathcal{Q}(p(0), q(0)) + \int_0^t \cos s\Omega \int_s^t g(q(u))\, du\, ds.$$

Thus, there is one quadrature of the force to build up the momentum and *two* in the evolution of q. In numerical methods, global errors are built up from local errors in a way analogous to that in which forcing terms build up the true solution. We showed in section 5 how the global error for p and q is obtained after one quadrature, and we shall presently show how the global error for q may be obtained by two quadratures. To avoid order reduction, each of these quadratures has to employ a weight function, and this essentially explains why methods that include mollification but no averaging cannot achieve stiff order 2 in q.

Led by the preceding insight, the idea now is to integrate by parts in (19) (note that (20) is of no use because the boundary terms β_n are $O(h)$). The analysis is straightforward except for problems with the boundary, and to soften these we proceed as follows. The functions $\psi((t-t_j)/h)$, $0 \le j \le n$, do not add up to 1 near the boundaries of the interval $I_n = [0, t_n]$ (this is because, for t_j near the boundary, $\psi((t-t_j)/h)$ "spreads" out of I_n, and, correspondingly, there are basis functions $\psi((t-t_j)/h)$, with $t_j < 0$ or $t_j > t_n$, that "intrude" into I_n). For fixed h and t_n, a system of $n+1$ functions λ_j with the partition of unity property

$$\sum_{j=0}^{n} \lambda_j(t) = 1, \quad 0 \le t \le t_n,$$

is easily constructed by setting, for $0 \le t \le t_n$,

$$\begin{aligned}
\lambda_0(t) &= \psi\left(\frac{t-t_0}{h}\right), \\
\lambda_j(t) &= \psi\left(\frac{t-t_j}{h}\right) + \psi\left(\frac{-t-t_j}{h}\right), \quad 0 < j < \mu, \\
\lambda_j(t) &= \psi\left(\frac{t-t_j}{h}\right), \qquad \mu \le j \le n-\mu, \\
\lambda_j(t) &= \psi\left(\frac{t-t_j}{h}\right) + \psi\left(\frac{-t+2t_n-t_j}{h}\right), \quad n-\mu < j < n, \\
\lambda_n(t) &= \psi\left(\frac{t-t_n}{h}\right).
\end{aligned}$$

(Near the boundaries the basis function $\psi((t-t_j)/h)$ have been "reflected" about 0 or t_n to compensate for the functions with $j = -1, -2, \ldots$ or $j = n+1, n+2, \ldots$ missing from the new basis.) The introduction of this corrected set of basis functions makes it possible to extend to the common interval $0 \le t \le t_n$ the integrals in (19).

LEMMA 3. *If ψ is of bounded support as in Lemma 2, the quadrature error $\sigma_{q,n}$ satisfies*

$$\begin{aligned}
\sigma_{q,n} = &\int_0^{t_n} \Omega^{-1} \sin(t_n - t)\Omega \sum_{j=0}^{n} \lambda_j(t) g_j^* \, dt \\
&- \int_0^{t_n} \Omega^{-1} \sin(t_n - t)\Omega \, g(q(t)) \, dt + \beta_n^*,
\end{aligned} \tag{24}$$

where β_n^* represent boundary contributions, with

$$\|\beta_n^*\| \le 2\mu^2 \|\psi\|_1 L_0 h^2,$$

so that, after integration by parts,

$$\sigma_{q,n} = \int_0^{t_n} \cos(t_n - t)\Omega\, I^*(t)\, dt + \beta_n^*,$$

with I^* equal to the integrated interpolation error

$$I^*(t) = \int_0^t \left[\left(\sum_{j=0}^{n} \lambda_j(s) g_j^* \right) - g(q(ts)) \right] ds. \tag{25}$$

Proof. The q-component of the first integral in (19) differs from the first integral in (24) only because of the contributions of values of j in the summation corresponding to t_j near the boundary. For $0 \le j < \mu$ there is a difference

$$1_j \int_{t_j - \mu h}^{0} \Omega^{-1} \sin(t_n - t)\Omega\ \psi\left(\frac{t - t_j}{h}\right) g_j^*\, dt$$
$$-1_j \int_{0}^{\mu h - t_j} \Omega^{-1} \sin(t_n - t)\Omega\ \psi\left(\frac{-t - t_j}{h}\right) g_j^*\, dt$$

or

$$1_j \int_{t_j - \mu h}^{0} \left[\Omega^{-1} \sin(t_n - t)\Omega - \Omega^{-1} \sin(t_n + t)\Omega\right]\ \psi\left(\frac{t - t_j}{h}\right) g_j^*\, dt,$$

whose Euclidean norm is $\le 2\mu h \times (h/2)\|\psi\|_1 \times L_0$ (note that $\Omega^{-1}\sin(t_n - t)\Omega$ is Lipschitz continuous in t with Lipschitz constant 1 so that the term in square brackets has norm $\le 2\mu h$). The estimate for β_n^* follows readily by summing over the contributions of j near the boundary at 0 and by treating in a similar way the boundary at t_n. ∎

In particular, at $t = t_n$,

$$I^*(t_n) = \sum_{j=0}^{n} h 1_j g_j^* - \int_0^{t_n} g(q(s))\, ds, \tag{26}$$

and we find once more the trapezoidal quadrature rule. For methods without averaging, $g_n^* = g(q_n)$ and second-order accuracy cannot be achieved in stiff cases where g oscillates with periods of the order of h. A simple counterexample is as follows. We take $d = 2$ and (the subindex refers to the components of the vectors)

$$\frac{d^2}{dt^2} q_{(1)} = -\omega^2 q_{(1)}, \quad \frac{d^2}{dt^2} q_{(2)} = -q_{(1)}; \tag{27}$$

the force $-\omega^2 q_{(1)}$ driving the first component is taken to be fast, and the force $q_{(1)}$ driving the second component is taken to be soft. We focus on the initial data $q_{(1)}(0) = \omega^{-1}$, $q_{(2)}(0) = \omega^{-3}$, $p_{(1)}(0) = 0$, $p_{(2)}(0) = 0$, leading to the solution $q_{(1)}(t) = \omega^{-1}\cos\omega t$, $q_{(2)}(t) = \omega^{-3}\cos\omega t$, that has reduced energy bounded independently of $\omega \ge 1$. If $h\omega = 2\pi$, then, in (26), the integral vanishes, while the value of the quadrature rule is the $O(h)$ vector $[0, \omega^{-1}t_n]$. Computation of the actual global error shows that in this instance $Q_n - q_n$ is also $O(h)$ and not $O(h^2)$.

The next and last lemma is relevant in understanding why averaging works: While g_j^* is only an $O(h)$ approximation to the true value $g(q(t_j))$ of the function being integrated in (25) or (26), hg_j^* provides a higher-order approximation to the integral

of $g(q(t))$ in the neighborhood of t_j. Note that below we use, without loss of generality, the same upper bound μ for the supports of ϕ and ψ.

LEMMA 4. *Under the hypotheses of Theorem* 2*, the following estimate holds:*

$$\|q(t) - \mathcal{A}_h q_j\| \leq (\mid t - t_j \mid + \|\phi\|_1)\sqrt{2E}\, h, \quad -\mu h \leq t \leq t_n + \mu h, \tag{28}$$

so that, in particular,

$$\|g_j - g_j^*\| = \|g(q(t_j)) - g(\mathcal{A}_h q(t_j))\| \leq L_1 \|\phi\|_1 \sqrt{2E}\, h.$$

Furthermore, if g possesses a Lipschitz continuous derivative and ϕ is boundedly supported with $\phi = 0$ for $\mid t \mid > \mu > 0$, then

$$h g_j^* = \int_{t_j - \mu h}^{t_j + \mu h} \phi\left(\frac{t - t_j}{h}\right) g(q(t))\, dt + h r_j, \tag{29}$$

where r_j can be estimated as

$$\|r_j\| \leq \|\phi\|_1 \left[L_2(\mu + \|\phi\|_1)^2 E + \frac{1}{2}\mu^2 L_0 L_1 \right] h^2.$$

Proof. To bound $q(t) - \mathcal{A}_h q_j$, add and subtract q_j. Then $q(t) - q_j$ may be written as the integral of p, and $(I - \mathcal{A}_h)q_j$ is given, by the definition of $\mathcal{A}_h$, as an integral involving the expression $(1 - \cos t\Omega)q_j$, which can be bounded by $\|h\Omega q_j\|$.

Next, by expanding g around $\mathcal{A}_h q_j$ we get

$$g(q(t)) = g_j^* + J_j[q(t) - \mathcal{A}_h q_j] + d_j, \quad t_j - \mu h \leq t \leq t_j + \mu h,$$

where J_j is the relevant Jacobian and the residual d_j may be bounded, according to (28) as $(1/2)L_2[(\mu + \|\phi\|_1)\sqrt{2E}\, h]^2$. Integration yields

$$-r_j = J_j \frac{1}{h} \int_{t_j - \mu h}^{t_j + \mu h} \phi\left(\frac{t - t_j}{h}\right) [q(t) - \mathcal{A}_h q_j]\, dt + e_j,$$

where the integrated residual e_j is bounded by $\|\phi\|_1$ times the bound for d_j. We are therefore left with the task of estimating the last integral. We note that, by the definition of averaging,

$$\frac{1}{h} \int_{t_j - \mu h}^{t_j + \mu h} \phi\left(\frac{t - t_j}{h}\right) \mathcal{A}_h q_j\, dt = \mathcal{A}_h q_j = \frac{1}{h} \int_{t_j - \mu h}^{t_j + \mu h} \phi\left(\frac{t - t_j}{h}\right) \cos(t - t_j)\Omega\, q_j dt,$$

while, by using the form of the true flow,

$$\frac{1}{h}\int_{t_j-\mu h}^{t_j+\mu h}\phi\left(\frac{t-t_j}{h}\right)q(t)\,dt = \frac{1}{h}\int_{t_j-\mu h}^{t_j+\mu h}\phi\left(\frac{t-t_j}{h}\right)[\cos(t-t_j)\Omega q_j+\sin(t-t_j)\Omega p_j+\Delta_j]\,dt, \tag{30}$$

where $\Delta_j=\Delta_j(t)$ is the integral contribution from the variation of constant formula

$$\Delta_j(t)=\int_{t_j}^{t}\Omega^{-1}\sin(t-s)\Omega\, g(q(s))\,ds.$$

In (30) the integral of the sine term vanishes by symmetry and that of the cosine term is cancelled with the same contribution from $\mathcal{A}_h$. This leaves us the term with Δ_j, and we note that, for $t_j-\mu h\le t\le t_j+\mu h$,

$$\|\Delta_j\|\le\int_{t_j}^{t}\mid t-s\mid L_0\,ds\le\frac{1}{2}\mu^2L_0h^2.$$

After combining all of these estimates, we obtain the bound being sought. ∎

Equipped with (29), we return to (25). Away from the integration limits, the integral of $\lambda_j(s)$ is h so that the jth term in the summation matches, except for an $O(h^3)$ residual, the integral of $\phi((s-t_j)/h)g(q(s))$. Therefore it is appropriate to impose the condition

$$\sum_{j=-\infty}^{\infty}\phi(t-j)\equiv 1, \tag{31}$$

as this would ensure that, except for boundary effects, the sum matches the whole integral. This leads us to the next theorem, where, for completeness, we refer to the equivalent formulation

$$\hat{\phi}(2\pi n)=0,\quad n=\pm1,\pm2,\ldots \tag{32}$$

of the last condition.

THEOREM 3. *With the hypotheses of Theorem* 1*, assume that g has a bounded, Lipschitz continuous derivative and that the weight functions ϕ and ψ have bounded support and satisfy* (22)–(23) *and* (31)–(32)*. Then the global error possesses a bound*

$$h\|P_n-p_n\|+\|Q_n-q_n\|\le Ch^2,$$

where the constant C depends on ψ, ϕ, t_n, L_0, L_1, L_2, and E.

Conversely, if the (ψ,ϕ)-method with boundedly supported ϕ and ψ possesses a bound of this form, then (22)–(23) *and* (31)–(32) *hold.*

Proof. The necessity of the condition for ϕ is proved by considering the counterexample (27).

To prove the bound we construct a set of interpolation basis functions κ_j, $j=0,1,\ldots,n$, by applying to the functions $\phi((t-t_j)/h)$ the construction that was used to derive the λ_j's from the $\psi((t-t_j)/h)$'s. Then, the κ_j's add to 1 and

$$I^*(t)=\sum_{j=0}^{n}\left[\int_0^t\lambda_j(s)g_j^*\,ds-\int_0^t\kappa(s)g(q(ts))\,ds\right].$$

If k is the largest j such that $t_j \le t$, we split I^* as $I^* = I_1^* + I_2^*$, with

$$I_1^* = \sum_{j=0}^{n} \left[\int_0^{t_k} \lambda_j(s) g_j^* \, ds - \int_0^{t_k} \kappa(s) g(q(ts)) \, ds \right],$$
$$I_2^* = \sum_{j=0}^{n} \left(\int_{t_k}^{t} \lambda_j(s) \right) g_j^* - \int_{t_k}^{t} g(q(s)) \, ds.$$

Both expressions being subtracted in I_2^* can be approximated with $O(h^2)$ errors by $(t - t_k)g_k$. For I_1^*, and as noted above, if t_j is away from the integration limits, the jth term in the sum is $O(h^3)$. For $j < \mu$, the term is only $O(h^2)$, because there κ_j does not quite match $\phi((t - t_j)/h)$. The terms with t_j near t can be dealt with in a similar way. ∎

7. Filters and weights. As discussed in section 4, when applied to cases with linear fast forces, the methods in this paper can be specified and implemented in terms of the two filter functions $\hat{\phi}$ and $\hat{\psi}$ in (17). However, both the implementation of the methods in nonlinear cases and the analysis in sections 5 and 6 require the weight functions themselves ϕ and ψ. Let us then address the question of whether given two real, even filter functions there exist corresponding weight functions that generate them as in (15)–(16). We furthermore recall that weights are assumed to be bounded and that the interest should be restricted to weights with bounded support if one is to compute numerically the averaging and mollifier operators.

The answer to our question is well known in Fourier analysis: According to a classical result by Paley and Wiener, a square integrable function $\hat{\chi}(\omega)$ of the real variable ω is the Fourier transform of a square integrable function χ supported in $[-\nu, \nu]$, $\nu > 0$, if and only if $\hat{\chi}$ can be extended to a holomorphic function of ω in the whole complex plane such that for all complex ω

$$| \hat{\chi}(\omega) | \le C \exp(\nu \, | \, \omega \, |).$$

For given ν, the space of all such $\hat{\chi}$ of *exponential type* is referred to as the Paley–Wiener space $PW_{[-\nu,\nu]}$. Hence any reversible method for linear fast forces of the family considered in [10], [6] for which the filter functions belong to some $PW_{[-\nu,\nu]}$ space can be extended to a (ϕ, ψ)-method applicable to nonlinear fast forces simply by taking as weights the inverse Fourier transforms of the filters. An example will be provided at the end of this section.

From the classical theorems of Weierstraß and Hadamard, it is well known that entire functions (functions holomorphic in the whole complex plane) are not determined by their zeros. However, a result by Titchmarsh (see, e.g., [4]) implies that any function $\hat{\chi}$ in $PW_{[-\nu,\nu]}$ can be expressed in terms of its infinitely many zeros ω_n as

$$\hat{\chi}(\omega) = \hat{\chi}(0) \prod_n \left(1 - \frac{\omega}{\omega_n} \right);$$

for the optimal convergence of the numerical method, $\hat{\chi}(0) = 1$ and the set of zeros must at least contain all nonzero even multiples of π. In this sense, the choice called *short* in [5],

$$\hat{\chi}_s(\omega) = \prod_{k=1} \left(1 - \frac{\omega^2}{4k^2\pi^2} \right) = \frac{\sin(\omega/2)}{\omega/2},$$

appears as a *minimal* filter: Filters of all potentially useful (ϕ, ψ)-methods can be obtained by multiplying $\hat{\chi}_s$ by suitable factors that bring in additional zeros. The corresponding weight function χ_s equals 1 for $-1/2 \le t \le 1/2$ and vanishes elsewhere. Obviously, no weight function may have support smaller than $[-1/2, 1/2]$ if its translations are to add to 1 (see (22), (31)).

Two other weight functions were explicitly discussed in [5]. The *long* weight χ_l, with value $1/2$ for $-1 \le t \le 1$ and 0 elsewhere, is a dilation of the short weight: $\chi_l(t) = (1/2)\chi_s(t/2)$. The *linear* weight function is the convolution of the short weight with itself, $\chi_{lin} = \chi_s * \chi_s$; it is the well-known hat basis function for piecewise linear interpolation with $\chi_{lin}(0) = 1$ and $\chi_{lin}(t) = 0$ for $| t | \ge 1$. The filter functions are, respectively, given by $\hat{\chi}_l(\omega) = \hat{\chi}_s(2\omega)$ (zeros at $k\pi$, $k \ne 0$) and $\hat{\chi}_{lin}(\omega) = \hat{\chi}_s(\omega)^2$ (double zeros at $2k\pi$, $k \ne 0$). Of course, convolution of the weights means product of the filters, and dilations translate into rescalings of ω.

The way to further generalizations by using convolutions and dilations is clear; these generalizations include the spline weights $\chi_s * \cdots * \chi_s$ (or $\chi_l * \cdots * \chi_l$) obtained by iterating the convolution of χ_s (or χ_l) with itself; the filters are the powers of $\hat{\chi}_s$ (or $\hat{\chi}_l$). More factors in the convolution mean higher-order zeros in the filter, and hopefully a better filtering, at the price of a longer support that entails higher computational work. The uncertainty principle implies that there is a limit on how much one can simultaneously concentrate near $t = 0$ the mass of χ and near $\omega = 0$ the mass of $\hat{\chi}$: Thus low computational cost and the drastic suppression of high frequencies are contradictory goals.

It is not our purpose here to settle the question as to the best choice of weight functions. We feel that the answer would very much depend both on the problem being solved and on the relation between the computational costs of integrating the reduced problem and of sampling the soft forces.

We shall finish the paper by presenting an example that illustrates how methods for linear fast forces based on filters can be extended to nonlinear fast forces by using weights. Grimm and Hochbruck have suggested [6] the filters $\omega^{-1} \sin \omega$ for the average and $\omega^{-2} \sin^2 \omega$ for the mollifier. These functions have exponential type 1 and 2, respectively, and, via the Paley–Wiener theorem, the method introduced in [6] can be extended to nonlinear fast forces by the (ϕ, ψ)-method that uses as weights the corresponding inverse Fourier transforms with supports $[-1, 1]$ and $[-2, 2]$. In fact, the weights are readily seen to be $\phi = \chi_l$ and $\psi = \chi_l * \chi_l$.

For a numerical test of this method, we use the nonlinear fast forces problem in section 4 of [5] and consider a system consisting of two unit point masses in the plane. The first mass is attached to one end of a (potentially strong) spring of unit length with stiffness ω^2 (in the experiments $0 \le \omega^2 \le 900$). The other end of this spring is fixed at the origin. The second mass is attached to the first through a (soft) spring of unit length and stiffness $1/2$. Initially the first mass is at $(1, 0)$ and the second at $(2, 0)$ so that the potential energy of the system is 0. The corresponding initial velocities are $(\sqrt{2}/4, \sqrt{2}/4)$ and $(-\sqrt{2}/4, \sqrt{2}/4)$, and the integration takes place for $0 \le t \le 16$, with step lengths $h = 1/2$ and $h = 1/4$.

Numerical results are presented for the $(\chi_l, \chi_l * \chi_l)$-method and, for comparison, also for the original impulse (δ, δ)-method and for the short filter (χ_s, χ_s)-method of [5]. The figure plots as a function of ω, $0 \le \omega \le 30$, the maximum over $0 \le t \le 16$ of the Euclidean norm of the error in the 4-dimensional vector of positions of the masses. For small values of ω the first spring is not really strong, and $(\chi_l, \chi_l * \chi_l)$ performs worse than the other two methods. For ω larger than, say, 10, the results of the original impulse method are clearly inferior to those of the methods that use

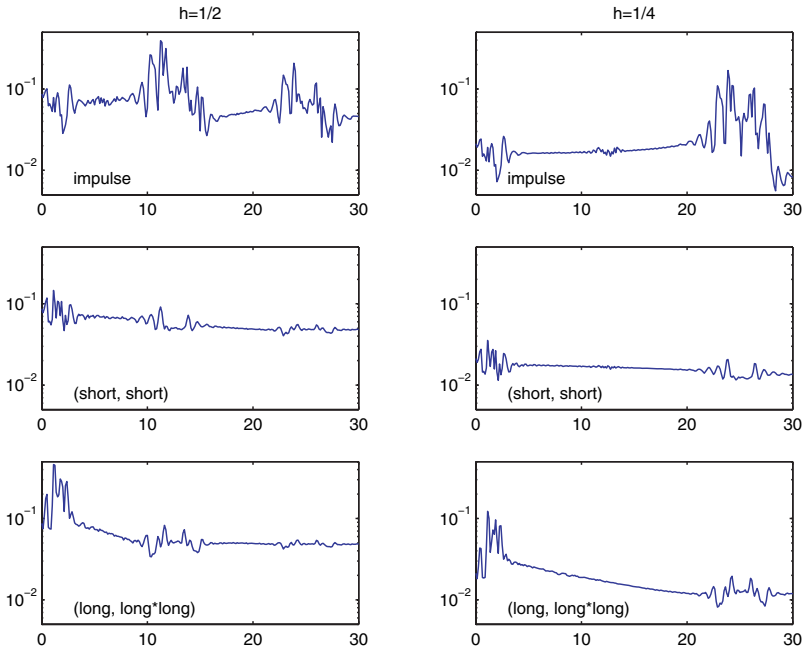


FIG. 7.1. *Two spring nonlinear problem: error in positions against* ω.

averaging and mollification. For the (χ_s, χ_s)-method the maximum over ω of the error is 0.1461 when $h = 1/2$ and 0.0354 when $h = 1/4$; for $(\chi_l, \chi_l * \chi_l)$ the corresponding figures are 0.4618 and 0.1227; thus both methods show in this nonlinear problem an $O(h^2)$ error behavior (as that ensured by Theorem 3 for the linear case). For the impulse method halving h brings down the error from 0.3931 to 0.1686, so that an order reduction manifests itself. In Figure 7.1 it is apparent that the order reduction is more pronounced if the attention is restricted to ω larger than, say, 20: Near $\omega = 8\pi$ the errors for $h = 1/4$ are hardly smaller than those for $h = 1/2$.

Acknowledgments. The author is thankful to D. Cohen, A. G. García, A. Iserles, Ch. Lubich, R. D. Skeel, and A. M. Stuart for their comments on the manuscript.

REFERENCES

[1] W. Arendt, Ch. J. K. Batty, M. Hieber, and F. Neubrander, *Vector-Valued Laplace Transforms and Cauchy Problems*, Birkhäuser, Basel, 2001.
[2] D. Cohen, E. Hairer, and Ch. Lubich, *Numerical energy conservation for multi-frequency oscillatory differential equations*, BIT, 45 (2005), pp. 287–305.
[3] K. Dekker and J. G. Verwer, *Stability of Runge-Kutta Methods for Stiff Nonlinear Differential Equations*, North-Holland, Amsterdam, 1984.
[4] A. G. García, *A brief walk through sampling theory*, Adv. Imag. Electron Phys., 124 (2002), pp. 63–137.
[5] B. García-Archilla, J. M. Sanz-Serna, and R. D. Skeel, *Long-time-step methods for oscillatory differential equations*, SIAM J. Sci. Comput., 20 (1998), pp. 930–963.
[6] V. Grimm and M. Hochbruck, *Error analysis of exponential integrators for oscillatory second-order differential equations*, J. Phys. A, 39 (2006), pp. 5495–5507.
[7] H. Grubmüller, *Dynamiksimulation sehr großer Makromoleküle auf einem Parallelrechner*, Master's thesis, Physik-Dept. der Tech. Univ. München, Munich, 1989.

[8] H. Grubmüller, H. Heller, A. Windemuth, and K. Schulten, *Generalized Verlet algorithm for efficient molecular dynamics simulations with long-range intereactions*, Mol. Sim., 6 (1991), pp. 121–142.

[9] E. Hairer and Ch. Lubich, *Long-time energy conservation of numerical methods for oscillatory differential equations*, SIAM J. Numer. Anal., 38 (2000), pp. 414–441.

[10] E. Hairer, Ch. Lubich, and G. Wanner, *Geometric Numerical Integration*, 2nd ed., Springer, Berlin, 2006.

[11] E. Hairer, S. P. Nørsett, and G. Wanner, *Solving Ordinary Differential Equations* I*, Nonstiff Problems*, 2nd. ed., Springer, Berlin, 1993.

[12] E. Hairer and G. Wanner, *Solving Ordinary Differential Equations* II*, Stiff and Differential-Algebraic Problems*, 2nd ed., Springer, Berlin, 1996.

[13] A. Iserles, *On the numerical quadrature of highly-oscillating integrals* I: *Fourier transforms,* IMA J. Numer. Anal., 24 (2004), pp. 365–391.

[14] B. Leimkuhler and S. Reich, *Simulating Hamiltonian Dynamics*, Cambridge University Press, Cambridge, 2005.

[15] J. M. Sanz-Serna, *Stability and convergence in numerical analysis* I: *Linear problems, a simple, comprehensive account*, in Nonlinear Differential Equations, J. K. Hale and P. Martínez-Amores, eds., Pitman, Boston, 1985, pp. 64–113.

[16] J. M. Sanz-Serna and M. P. Calvo, *Numerical Hamiltonian Problems*, Chapman and Hall, London, 1994.

[17] M. Tuckerman, B. J. Berne, and G. J. Martyna, *Reversible multiple time scale molecular dynamics,* J. Phys. Chem., 99 (1995), pp. 1990–2001.

[18] J. Wisdom, *The origin of the Kirkwood gaps: A mapping for asteroidal motion near the* 3/1 *commensurability*, Astrophysics J., 87 (1982), pp. 577–593.

SIAM J. NUMER. ANAL.
Vol. 46, No. 2, pp. 1060–1078
© 2008 Society for Industrial and Applied Mathematics

A SPECTRAL VISCOSITY METHOD BASED ON HERMITE FUNCTIONS FOR NONLINEAR CONSERVATION LAWS*

JULIÁN AGUIRRE† AND JUDITH RIVAS†

Abstract. We consider the approximation by a spectral method of the solution of the Cauchy problem for a scalar conservation law in one dimension posed in the whole real line. We analyze a spectral viscosity method in which the orthogonal basis considered is the one of Hermite functions. We prove the convergence of the approximate solution to the unique entropy solution of the problem by using compensated compactness arguments.

Key words. hyperbolic conservation laws, Hermite functions, spectral methods, spectral viscosity, compensated compactness

AMS subject classifications. 33C45, 65M70, 35L65

DOI. 10.1137/070684082

1. Introduction.

We consider the Cauchy problem for a genuinely nonlinear scalar conservation law in one space dimension posed in the whole real line:

$$\begin{cases} \dfrac{\partial u}{\partial t} + \dfrac{\partial f(u)}{\partial x} = 0, & x \in \mathbb{R}, t > 0, \\ u(x,0) = \varphi(x), & x \in \mathbb{R}, \end{cases} \tag{1.1}$$

where $f \in C^1$ is a smooth nonlinear function and $\varphi \in L^\infty(\mathbb{R})$. It is well known that, in general, there are no classical solutions to this problem. Moreover, weak solutions are not unique. To isolate the physically relevant solution, one has to impose the entropy condition

$$\frac{\partial U(u)}{\partial t} + \frac{\partial F(u)}{\partial x} \le 0 \tag{1.2}$$

in the sense of distributions, for all entropy pairs (U, F), with $U \in C^2$ convex and $F'(u) = U'(u)f'(u)$; see [7], [11].

Our aim is to approximate the unique entropy solution of (1.1), that is, the only weak solution verifying (1.2), by means of a spectral viscosity method based on Hermite functions.

The idea of a spectral method [8] is to approximate the solution of a PDE by a truncated series of the form $\sum_{k=0}^{N} \tilde{u}_k(t)\phi_k(x)$, where $\{\phi_k\}$ is an orthogonal basis of some Hilbert space. When applied to a hyperbolic conservation law, the spectral method must be supplied with an appropriate amount of artificial viscosity to avoid the possible instabilities that can appear due to the nonsmoothness of the exact solution. The so-called spectral viscosity method was first introduced in [21] and then developed in [15], [14], and [12] among others.

On the other hand, numerical methods such as finite differences, finite elements, or Fourier spectral methods always assume a finite domain of computation. When

*Received by the editors March 1, 2007; accepted for publication (in revised form) October 30, 2007; published electronically March 5, 2008. This work was supported by grant UPV 127.310-15969/2004.

http://www.siam.org/journals/sinum/46-2/68408.html

†Departamento de Matemáticas, Facultad de Ciencia y Tecnología, Universidad del País Vasco, Apdo. 644, 48080 Bilbao, Spain (julian.aguirre@ehu.es, judith.rivas@ehu.es).

dealing with a problem posed in an unbounded domain, the usual techniques to avoid this difficulty are the following:

(i) choose a large bounded computational domain and impose artificial boundary conditions;

(ii) assume the solution is periodic;

(iii) make a change of variables transforming the original domain into a bounded one.

Our approach respects the unboundedness of the domain considering as orthogonal basis for the spectral method the one consisting of Hermite functions.

The use of Hermite functions for the approximation of solutions of partial differential equations posed in the whole real line was proposed by Funaro and Kavian [6] and developed in [16], [13] for the case of viscous scalar conservation laws, in [5] for the Fokker–Planck equation, and in [9] for the Dirac equation.

One of the advantages of spectral methods is that they enjoy spectral accuracy; this is, if the function to be approximated is very regular, the approximation converges faster than any negative power of N. They are, therefore, very appropriate in the case of elliptic and parabolic equations thanks to the regularization properties of these operators. However, solutions of hyperbolic nonlinear problems may develop discontinuities in finite time, as mentioned before, leading to at most first order accurate spectral approximations. In [18], a postprocessing technique has been developed to enhance the convergence rate of the approximations, at least in points not too close to the shock.

The paper is organized as follows. In section 2 we describe the elements involved in the definition of our spectral method and recall the basic properties that will be needed in what follows. In section 3 we describe the method. Section 4 is devoted to finding a priori estimates for the approximate solutions that will be used in section 5 to prove the convergence of the method by compensated compactness arguments, supplemented by an L^∞ uniform boundedness assumption on the numerical solution. Finally, in section 6, we present some numerical experiments

2. Preliminaries. In this section we will describe the elements that take part in our spectral method as well as the basic properties that will be used in the proof of convergence.

As we have mentioned, we will consider the Hermite functions as an orthogonal basis. These functions are defined as

$$h_k(x) = e^{-x^2} H_k(x), \quad k = 0, 1, 2 \ldots,$$

where H_k is the Hermite polynomial of degree k and leading coefficient 2^k, the solution of the second order differential equation

$$y'' - 2xy' + 2ky = 0. \tag{2.1}$$

From the properties of Hermite polynomials [20], it follows that

$$h_k'(x) = -h_{k+1}(x), \quad k = 0, 1, 2 \ldots. \tag{2.2}$$

Hermite functions satisfy an orthogonality relation with respect to the weight function $w(x) = e^{x^2}$:

$$\int_{-\infty}^{\infty} h_n(x) h_m(x)\, w(x) dx = \delta_{m,n} 2^n n! \sqrt{\pi} \quad \forall\, n, m = 0, 1, 2, \ldots. \tag{2.3}$$

Hence, they form an orthogonal basis of the weighted L^2 space

$$L^2_w = \left\{ \varphi \colon \mathbb{R} \to \mathbb{R} \ \text{ measurable } \quad \Big| \quad \int_{-\infty}^{\infty} \Big| \varphi(x)|^2 \, w(x) dx < \infty \right\}.$$

We shall denote the norm in this space by

$$\|\varphi\|_{0,w} = \left(\int_{-\infty}^{\infty} |\varphi(x)|^2 \, w(x) dx \right)^{1/2}.$$

The weighted Sobolev spaces associated to L^2_w are

$$H^k_w = \left\{ \varphi \colon \mathbb{R} \to \mathbb{R} \quad | \quad \varphi, \varphi', \dots, \varphi^{(k)} \in L^2_w \right\}, \quad k = 0, 1, 2, \dots.$$

In [6] one can find a characterization of the functions of H^1_w that we recover in the following proposition, as well as some inequalities regarding the norms of functions in this space.

PROPOSITION 1. *Let $\varphi \in L^2_w$. $\varphi \in H^1_w$ if and only if $w^{1/2}\varphi \in H^1(\mathbb{R})$ and $x\varphi \in L^2_w$. Besides,*

$$\|(w^{1/2}\varphi)'\|_{L^2} \le \|\varphi'\|_{0,w}, \tag{2.4}$$

$$\|\varphi\|_{0,w} \le \|\varphi'\|_{0,w}, \tag{2.5}$$

$$\|x\varphi\|_{0,w} \le \|\varphi'\|_{0,w}. \tag{2.6}$$

Inequality (2.5), analogous to Poincare's one in bounded domains, implies that $\|\varphi^{(k)}\|_{0,w}$ is equivalent to the usual norm $\sum_{j=0}^{k} \|\varphi^{(j)}\|_{0,w}$ on H^k_w. Therefore, we define the norm in H^k_w as

$$\|\varphi\|_{k,w} = \|\varphi^{(k)}\|_{0,w}.$$

For noninteger $s > 0$, H^s_w is defined by interpolation, and, for $s < 0$, H^s_w is the dual of H^{-s}_w.

Since Hermite functions form an orthogonal basis of L^2_w, any $\varphi \in L^2_w$ can be expressed in a unique way as a Fourier–Hermite series of the form

$$\varphi(x) = \sum_{k=0}^{\infty} \widehat{\varphi}_k h_k(x), \qquad \text{where} \qquad \widehat{\varphi}_k = \frac{1}{\sqrt{\pi}\, 2^k \, k!} \int_{-\infty}^{\infty} \varphi(x) h_k(x) w(x) dx.$$

If we denote by V_N the subspace of L^2_w generated by the first $N+1$ Hermite functions, $h_0, h_1, \dots, h_N$, given $\varphi \in L^2_w$, its best approximation in the L^2_w norm by functions of V_N is the orthogonal or spectral projection, that is,

$$\pi_N \varphi(x) = \sum_{k=0}^{N} \widehat{\varphi}_k h_k(x).$$

From the definition of the operator π_N, one has

$$\|\pi_N \varphi\|_{0,w} \le \|\varphi\|_{0,w} \quad \forall \varphi \in L^2_w.$$

Besides, the following estimate for the weighted Sobolev norms of the error of the approximation of φ by $\pi_N \varphi$ is given in [6].

PROPOSITION 2. *Let $0 \le \mu \le \sigma$. There exists a positive constant C, independent of N, such that, for all $\varphi \in H_w^\sigma$,*

$$\|\varphi - \pi_N \varphi\|_{\mu,w} \le C N^{\frac{\mu-\sigma}{2}} \|\varphi\|_{\sigma,w}. \tag{2.7}$$

In practice, the spectral projection $\pi_N\varphi$ is of limited value since it is difficult to compute accurately. There is another way of approximating a continuous function $\varphi \in L_w^2$ by an element of V_N. Let $z_0, z_1, \ldots, z_N$ be the zeros of H_{N+1}, which are all real and distinct. Given $\varphi \in C(\mathbb{R}) \cap L_w^2$, its Nth pseudospectral projection is the unique function $I_N\varphi \in V_N$ such that

$$I_N\varphi(z_j) = \varphi(z_j), \quad j = 0, 1, \ldots, N.$$

If $I_N\varphi(x) = \sum_{k=0}^N \tilde{\varphi}_k h_k(x)$, the kth coefficient $\tilde{\varphi}_k$ is the result of approximating the exact Fourier–Hermite coefficient $\widehat{\varphi}_k$ by means of a Gaussian quadrature formula. More precisely,

$$\tilde{\varphi}_k = \sum_{j=0}^N w_{k,j}\varphi(z_j), \quad \text{where } w_{k,j} = \frac{2^{N-k} N! h_k(z_j)}{(N+1)k!(h_N(z_j))^2}. \tag{2.8}$$

In [1], the following estimate for the weighted Sobolev norms of the error of the pseudospectral projection is given.

PROPOSITION 3. *Let $\sigma \ge 1$ and $0 \le \mu \le \sigma$. There exists a positive constant C, independent of N, such that, for all $\varphi \in H_w^\sigma$,*

$$\|\varphi - I_N\varphi\|_{\mu,w} \le C N^{\frac{1}{6}+\frac{\mu-\sigma}{2}} \|\varphi\|_{\sigma,w}. \tag{2.9}$$

Finally, we introduce a differential operator which will play the role of the Laplacian in the usual viscosity spectral methods. Given $\varphi \in H_w^2$,

$$\mathcal{L}\varphi(x) = -e^{-x^2}\left(e^{x^2}\varphi'(x)\right)' = -(\varphi''(x) + 2x\varphi'(x)).$$

From (2.1) and the definition of Hermite functions it follows that the eigenfunctions of $\mathcal{L}$ are the Hermite functions h_k with eigenvalues $\lambda_k = 2(k+1)$; that is,

$$\mathcal{L}h_k(x) = 2(k+1)h_k(x), \quad k = 0, 1, 2, \ldots.$$

In [3] it is proved that $\mathcal{L}$ is a positive definite self-adjoint operator with compact inverse in L_w^2 that generates an analytic semigroup of contractions $\{e^{-t\mathcal{L}}\}$.

3. The Hermite spectral and pseudospectral viscosity methods. It is well known that the entropy solution of (1.1) can be obtained as the limit when ϵ tends to 0 of the solution of the parabolic problem obtained when one introduces in the right-hand side of (1.1) a vanishing viscosity term of the form ϵu_{xx}. Based on this fact, Tadmor defined the spectral viscosity approximation of the entropy solution u as a truncated Fourier series that is the solution of the approximated problem

$$\begin{cases} \dfrac{\partial u_N}{\partial t} + P_N \dfrac{\partial f(u_N)}{\partial x} = \epsilon_N \dfrac{\partial}{\partial x}\left(Q_{m_N} \dfrac{\partial u_N}{\partial x}\right), & x \in I, t > 0, \\ u_N(x,0) = P_N u(x,0), & x \in I, \end{cases}$$

where P_N stands for the Fourier spectral or pseudospectral projection. Since the exponentials e^{ikx} are eigenfunctions of the Laplacian, the term on the right-hand side can be easily implemented. However, when one takes the Hermite functions as an orthogonal basis, the Laplacian is not an appropriate way to introduce the artificial viscosity in the numerical scheme.

By using the properties of the operator $\mathcal{L}$ as well as compensated compactness arguments similar to those to be used in the proof of convergence in section 5, one can prove the following result about the entropy solution of (1.1).

THEOREM 1. *Let $f \in C^1$ be a nonlinear function such that $f'(0) = 0$, $\varphi \in L^2_w \cap L^\infty(\mathbb{R})$ and u_ϵ the solution of*

$$\begin{cases} \dfrac{\partial u_\epsilon}{\partial t} + \dfrac{\partial f(u_\epsilon)}{\partial x} + \epsilon \mathcal{L} u_\epsilon = 0, & x \in \mathbb{R},\ t > 0, \\ u_\epsilon(x,0) = \varphi(x), & x \in \mathbb{R}. \end{cases}$$

Then, for any $p \geq 1$ and any $\Omega \subset \mathbb{R} \times (0,\infty)$ open and bounded, u_ϵ converges in $L^p(\Omega)$ to the unique entropy solution of (1.1) *when ϵ tends to* 0.

By taking this into account, we define the spectral (or pseudospectral) viscosity approximation to the unique entropy solution of (1.1) as $u_N(x,t) = \sum_{k=0}^N \widetilde{u}_k(t)\, h_k(x)$ such that

$$\begin{cases} \dfrac{\partial u_N}{\partial t} + \dfrac{\partial}{\partial x}\left(P_{N-1} f(u_N)\right) + \epsilon_N \mathcal{L}(Q_{m_N} u_N) = 0, & x \in \mathbb{R},\ t \in (0,T), \\ u_N(x,0) = P_N \varphi(x), & x \in \mathbb{R}. \end{cases}$$

P_N stands for π_N in the spectral viscosity approximation and for I_N in the pseudospectral one. Q_{m_N} is a viscosity operator which modifies only the high modes of the Fourier–Hermite expansion, that is,

$$Q_{m_N}\left(\sum_{k=0}^N \widehat{\varphi}_k\, h_k(x)\right) = \sum_{k=0}^N \hat{q}_k\, \widehat{\varphi}_k\, h_k(x), \tag{3.1}$$

where

$$\begin{cases} \hat{q}_k = 0 & \text{if } k \leq m_N, \\ 1 - \dfrac{m_N}{k} \leq \hat{q}_k \leq 1 & \text{if } k > m_N, \end{cases} \tag{3.2}$$

$m_N < N$ being a positive integer that tends to ∞ with N. Finally, ϵ_N is a positive parameter depending on N that will tend to 0 as N tends to infinity.

We will give conditions on m_N and ϵ_N to ensure the convergence of u_N to the unique entropy solution of (1.1).

4. A priori estimates. In this section we will find two kinds of a priori estimates, following the lines of [14]. The first kind of estimate is related to the viscosity operator Q_{m_N}, while the second one is concerned with the approximate solution u_N.

LEMMA 1. *Let R be a linear operator defined in V_N in the following way:*

$$R\left(\sum_{k=0}^N \widehat{\varphi}_k\, h_k(x)\right) = \sum_{k=0}^N \hat{r}_k\, \widehat{\varphi}_k\, h_k(x),$$

where $\hat{r}_0, \ldots, \hat{r}_N$ are real numbers. Then, for all $\varphi \in V_N$,

$$\|R\varphi\|_{1,w}^2 \leq 2\left(\sum_{k=0}^{N} \hat{r}_k^2(k+1)\right)\|\varphi\|_{0,w}^2 .$$

Proof. Let $\varphi(x) = \sum_{k=0}^{N} \widehat{\varphi}_k h_k(x)$. From (2.2), $(R\varphi)'(x) = -\sum_{k=0}^{N} \hat{r}_k\, \widehat{\varphi}_k\, h_{k+1}(x)$, so that, by taking norms and recalling (2.3), one has

$$\begin{aligned}\|R\varphi\|_{1,w}^2 &= \sum_{k=0}^{N} \hat{r}_k^2\, \widehat{\varphi}_k^2\, 2^{k+1}(k+1)!\,\sqrt{\pi} \\ &\leq 2\left(\sum_{k=0}^{N} \hat{r}_k^2(k+1)\right)\left(\sum_{k=0}^{N} \widehat{\varphi}_k^2\, 2^k\, k!\,\sqrt{\pi}\right) \\ &= 2\left(\sum_{k=0}^{N} \hat{r}_k^2\,(k+1)\right)\|\varphi\|_{0,w}{}^2. \qquad \square\end{aligned}$$

By using this result we can now bound the norm of the derivative of a function of V_N by the sum of the H_w^1 norms of its high frequencies modified by Q_{m_N} and the L_w^2 norm of the function itself, multiplied by a constant that grows with N.

LEMMA 2. *Let Q_{m_N} be defined as in* (3.1), (3.2). *Then there exist positive constants C_1 and C_2, independent of N, such that*

$$\|\varphi\|_{1,w}^2 \leq C_1\,\|Q_{m_N}\varphi\|_{1,w}^2 + C_2 m_N^2\,\|\varphi\|_{0,w}^2 \quad \forall \varphi \in V_N.$$

Proof. Let $\varphi(x) = \sum_{k=0}^{N} \widehat{\varphi}_k h_k(x)$, and let $R_{m_N} = I - Q_{m_N}$, where I is the identity operator. Then

$$\|\varphi\|_{1,w}^2 \leq C\left(\|Q_{m_N}\varphi\|_{1,w}^2 + \|R_{m_N}\varphi\|_{1,w}^2\right).$$

We split φ in dyadic parts $\varphi(x) = \sum_{k=0}^{m_N} \widehat{\varphi}_k\, h_k(x) + \sum_{j=1}^{J} \varphi^j(x)$, where

$$\varphi^j(x) = \sum_{k>2^{j-1}m_N}^{2^j m_N} \widehat{\varphi}_k\, h_k(x), \qquad j = 1, \ldots, J.$$

Here $J = [\log_2 \frac{N}{m_N}] + 1$ and $\widehat{\varphi}_k = 0$ for $k = N+1, \ldots, 2^J m_N$.

From the orthogonality relation (2.3), one has

$$\|R_{m_N}\varphi\|_{1,w}^2 = \left\|R_{m_N}\sum_{k=0}^{m_N} \widehat{\varphi}_k\, h_k\right\|_{1,w}^2 + \sum_{j=1}^{J} \left\|R_{m_N}\varphi^j\right\|_{1,w}^2 . \tag{4.1}$$

We bound each summand by using the result obtained in Lemma 1. Since $\hat{q}_k = 0$ for $k \leq m_N$,

$$\begin{aligned}\left\|R_{m_N}\sum_{k=0}^{m_N} \widehat{\varphi}_k\, h_k\right\|_{1,w}^2 &\leq 2\left(\sum_{k=0}^{m_N}(1-\hat{q}_k)^2(k+1)\right)\left\|\sum_{k=0}^{m_N} \widehat{\varphi}_k\, h_k\right\|_{0,w}^2 \\ &\leq C m_N^2 \left\|\sum_{k=0}^{m_N} \widehat{\varphi}_k\, h_k\right\|_{0,w}^2 .\end{aligned}$$

For the terms in the summatory, since $\hat{q}_k \geq 1 - \frac{m_N}{k}$, we get, for any $j = 1, \ldots, J$,

$$\begin{aligned}
\left\|R_{m_N}\varphi^j\right\|_{1,w}^2 &\leq C\left(\sum_{k>2^{j-1}m_N}^{2^j m_N}(1-\hat{q}_k)^2(k+1)\right)\left\|\varphi^j\right\|_{0,w}^2 \\
&\leq Cm_N^2\sum_{k>2^{j-1}m_N}^{2^j m_N}\frac{1}{k}\left\|\varphi^j\right\|_{0,w}^2 \\
&\leq Cm_N^2\frac{1}{2^{j-1}m_N}(2^j m_N - 2^{j-1}m_N)\left\|\varphi^j\right\|_{0,w}^2 \\
&= Cm_N^2\left\|\varphi^j\right\|_{0,w}^2.
\end{aligned}$$

Hence,

$$\begin{aligned}
\|R_{m_N}\varphi\|_{1,w}^2 &\leq Cm_N^2\left(\left\|\sum_{k=0}^{m_N}\widehat{\varphi}_k\,h_k\right\|_{0,w}^2 + \sum_{j=1}^{J}\left\|\varphi^j\right\|_{0,w}^2\right) \\
&\leq Cm_N^2\left\|\varphi\right\|_{0,w}^2,
\end{aligned}$$

and by substituting in (4.1) we get the desired result. □

Now we will find some a priori estimates for the solution of the approximate problem u_N. We will use the following notation for norms in $\mathbb{R}\times[0,T]$, where $T>0$:

$$\|\varphi\|_{k,w,T}^2 = \int_0^T \|\varphi(\cdot,t)\|_{k,w}^2\,dt \quad \forall\varphi\in L^2(\mathbb{R},H_w^k).$$

There are some differences between the spectral and the pseudospectral cases; therefore we will distinguish both cases. We start with the spectral approximation.

LEMMA 3. *Let $f\in C^2(\mathbb{R})$ be such that $f'(0)=0$, $\varphi\in L_w^2$, $T>0$, $m_N=[O(N^\beta)]$, $\epsilon_N = O(N^{-\theta})$, with $0<2\beta<\theta<1/2$, Q_{m_N} given by (3.1), (3.2), and $u_N\colon[0,T]\to V_N$ the solution of*

$$\text{(4.2)}\qquad \begin{cases} \dfrac{\partial u_N}{\partial t} + \dfrac{\partial}{\partial x}\left(\pi_{N-1}f(u_N)\right) + \epsilon_N\mathcal{L}(Q_{m_N}u_N) = 0, & x\in\mathbb{R},\ t\in(0,T), \\ u_N(x,0) = \pi_N\varphi(x), & x\in\mathbb{R}. \end{cases}$$

Let us assume that there exists a positive constant c, independent of N, such that

$$\text{(4.3)}\qquad \|(1+|x|)u_N\|_{L^\infty(\mathbb{R}\times(0,T))}\leq c.$$

Then there exist positive constants k and $C(T)$, independent of N, such that

$$\text{(4.4)}\qquad \|u_N(\cdot,t)\|_{0,w}\leq e^{kt}\|u_N(\cdot,0)\|_{0,w}\quad\forall\,t\in(0,T),$$

$$\text{(4.5)}\qquad \|Q_{m_N}u_N\|_{1,w,T}\leq\frac{C(T)}{\sqrt{\epsilon_N}},$$

$$\text{(4.6)}\qquad \|u_N\|_{1,w,T}\leq\frac{C(T)}{\sqrt{\epsilon_N}}.$$

Proof. We multiply (4.2) by u_N and integrate in space with respect to the weight $w(x)$ to yield

$$\begin{aligned}\int_{-\infty}^{\infty} \frac{\partial u_N}{\partial t} u_N\, w\, dx + \int_{-\infty}^{\infty} \frac{\partial}{\partial x}\left(\pi_{N-1} f(u_N)\right) u_N\, w\, dx \\ - \epsilon_N \int_{-\infty}^{\infty} e^{-x^2} \frac{\partial}{\partial x}\left(e^{x^2} \frac{\partial}{\partial x}(Q_{m_N} u_N)\right) u_N\, w\, dx = 0.\end{aligned} \tag{4.7}$$

By taking into account that $\frac{d}{dx}\pi_{N-1} = \pi_N \frac{d}{dx}$ and the orthogonality of Hermite functions, we get

$$\int_{-\infty}^{\infty} \frac{\partial}{\partial x}\left(\pi_{N-1} f(u_N)\right) u_N\, w\, dx = \int_{-\infty}^{\infty} \frac{\partial}{\partial x}\left(f(u_N)\right) u_N\, w\, dx.$$

Let F be a primitive of $uf'(u)$ such that $F(0) = 0$. Then by integrating by parts above

$$\int_{-\infty}^{\infty} \frac{\partial}{\partial x}\left(f(u_N)\right) u_N\, w\, dx = \int_{-\infty}^{\infty} \frac{\partial}{\partial x}\left(F(u_N)\right)\, w\, dx = -\int_{-\infty}^{\infty} 2x\, F(u_N)\, w\, dx.$$

From the mean value theorem, for each (x,t) there exists η_N between 0 and u_N such that

$$F(u_N) = F'(\eta_N)\, u_N = \eta_N\, f'(\eta_N) u_N.$$

By using again the mean value theorem, this time with f', for each (x,t) there exists ξ_N between 0 and η_N such that

$$F(u_N) = \eta_N\, \left(f'(\eta_N) - f'(0)\right)\, u_N = \eta_N\, f''(\xi_N)\, \eta_N\, u_N = f''(\xi_N)\, \eta_N^2\, u_N$$

and hence

$$\begin{aligned}\left|\int_{-\infty}^{\infty} 2x\, F(u_N)\, w\, dx\right| &= \left|\int_{-\infty}^{\infty} 2x\, \eta_N^2\, f''(\xi_N)\, u_N\, w\, dx\right| \\ &\le \|2\,|x|\, u_N\|_{L^\infty(\mathbb{R}\times(0,T))} \sup_{|\xi|\le c} |f''(\xi)|\, \|\eta_N(\cdot,t)\|_{0,w}^2 \\ &\le k\, \|u_N(\cdot,t)\|_{0,w}^2\,.\end{aligned} \tag{4.8}$$

In the third integral in (4.7) we integrate by parts to yield

$$-\epsilon_N \int_{-\infty}^{\infty} e^{-x^2} \frac{\partial}{\partial x}\left(e^{x^2} \frac{\partial}{\partial x}\left(Q_{m_N} u_N\right)\right) u_N\, w\, dx = \epsilon_N \int_{-\infty}^{\infty} \frac{\partial}{\partial x}\left(Q_{m_N} u_N\right) \frac{\partial u_N}{\partial x}\, w\, dx.$$

Let $u_N(x,t) = \sum_{k=0}^{N} \widetilde{u}_k(t) h_k(x)$. From the definition of Q_{m_N} and the orthogonality of Hermite functions,

$$\begin{aligned}\int_{-\infty}^{\infty} \frac{\partial}{\partial x}\left(Q_{m_N} u_N\right) \frac{\partial u_N}{\partial x}\, w\, dx &= \sum_{k=0}^{N} \widehat{q}_k \left(\widetilde{u}_k(t)\right)^2 \|h_{k+1}\|_{0,w}^2 \\ &\ge \sum_{k=0}^{N} \widehat{q}_k^2 \left(\widetilde{u}_k(t)\right)^2 \|h_{k+1}\|_{0,w}^2 \\ &= \left\| \frac{\partial}{\partial x}(Q_{m_N} u_N)(\cdot,t)\right\|_{0,w}^2.\end{aligned} \tag{4.9}$$

By substituting (4.8) and (4.9) in (4.7), we get

$$\frac{1}{2}\frac{d}{dt}\left\|u_N(\cdot,t)\right\|_{0,w}^2+\epsilon_N\left\|\frac{\partial}{\partial x}(Q_{m_N}u_N)(\cdot,t)\right\|_{0,w}^2\leq k\left\|u_N(\cdot,t)\right\|_{0,w}^2. \tag{4.10}$$

In particular, $\frac{d}{dt}\|u_N(\cdot,t)\|_{0,w}^2\leq 2k\|u_N(\cdot,t)\|_{0,w}^2$; hence

$$\left\|u_N(\cdot,t)\right\|_{0,w}^2\leq e^{2kt}\left\|u_N(\cdot,0)\right\|_{0,w}^2,$$

and (4.4) follows. Besides, by integrating (4.10) in time between 0 and T,

$$\begin{aligned}\left\|u_N(\cdot,T)\right\|_{0,w}^2+2\epsilon_N\left\|\frac{\partial}{\partial x}(Q_{m_N}u_N)\right\|_{0,w,T}^2&\leq\left\|u_N(\cdot,0)\right\|_{0,w}^2+2k\int_0^T\left\|u_N(\cdot,t)\right\|_{0,w}^2\,dt\\ &\leq C(T)\left\|u_N(\cdot,0)\right\|_{0,w}^2.\end{aligned} \tag{4.11}$$

Since $u_N(x,0)=\pi_N\varphi(x)$ and $\varphi\in L^2_w$, we know that $\|u_N(\cdot,t)\|_{0,w}\leq\|\varphi\|_{0,w}$. Therefore, from (4.11) one deduces the estimate (4.5). Finally, by Lemma 2 one can obtain

$$\begin{aligned}\left\|\frac{\partial u_N}{\partial x}\right\|_{0,w,T}^2&\leq C\left\|\frac{\partial}{\partial x}(Q_{m_N}u_N)\right\|_{0,w,T}^2+Cm_N^2\int_0^T\left\|u_N(\cdot,t)\right\|_{0,w}^2\,dt\\ &\leq\frac{C(T)}{\epsilon_N}+C(T)m_N^2\\ &\leq C(T)\frac{1+\epsilon_N m_N^2}{\epsilon_N}.\end{aligned}$$

The inequality $2\beta<\theta$ implies that $\epsilon_N m_N^2=o(1)$ and, hence, the estimate (4.6) is verified. □

In the case of the pseudospectral approximation, more restrictive conditions must be imposed due to the fact that the approximation properties of I_N are not as good as those of π_N.

LEMMA 4. *Let $f\in C^2(\mathbb{R})$ be such that $f'(0)=0$, $\varphi\in H^1_w$, $T>0$, $m_N=[O(N^\beta]$, $\epsilon_N=O(N^{-\theta})$, with $0<2\beta<\theta<1/3$, Q_{m_N} given by (3.1), (3.2), and $u_N\colon[0,T]\to V_N$ the solution of*

$$\begin{cases}\dfrac{\partial u_N}{\partial t}+\dfrac{\partial}{\partial x}\left(I_{N-1}f(u_N)\right)+\epsilon_N\mathcal{L}(Q_{m_N}u_N)=0, & x\in\mathbb{R},\ t\in(0,T),\\ u_N(x,0)=I_N\varphi(x), & x\in\mathbb{R}.\end{cases} \tag{4.12}$$

Let us assume that there exists a positive constant c, independent of N, such that

$$\|(1+|x|)u_N\|_{L^\infty(\mathbb{R}\times(0,T))}\leq c. \tag{4.13}$$

Then there exist positive constants k and $C(T)$, independent of N, such that

$$\|u_N(\cdot,t)\|_{0,w}\leq e^{kt}\|u_N(\cdot,0)\|_{0,w}\quad\forall\,t\in(0,T), \tag{4.14}$$

$$\|Q_{m_N}u_N\|_{1,w,T}\leq\frac{C(T)}{\sqrt{\epsilon_N}}, \tag{4.15}$$

$$\|u_N\|_{1,w,T}\leq\frac{C(T)}{\sqrt{\epsilon_N}}. \tag{4.16}$$

Proof. Since the proof of these results is very similar to the one of Lemma 3, we will detail only the steps that present some difference. By multiplying (4.12) by u_N and integrating in space with respect to w, we get

$$\begin{aligned}\int_{-\infty}^{\infty} \frac{\partial u_N}{\partial t} u_N\, w\, dx + \int_{-\infty}^{\infty} \frac{\partial}{\partial x}\left(I_{N-1} f(u_N)\right) u_N\, w\, dx \\ - \epsilon_N \int_{-\infty}^{\infty} e^{-x^2} \frac{\partial}{\partial x}\left(e^{x^2} \frac{\partial}{\partial x}(Q_{m_N} u_N)\right) u_N\, w\, dx = 0.\end{aligned} \tag{4.17}$$

While the first and third integrals are rewritten and bounded as in Lemma 3, the second integral needs a different analysis, since $\frac{d}{dx} I_{N-1} \neq I_N \frac{d}{dx}$. We decompose this integral as follows:

$$\begin{aligned}\int_{-\infty}^{\infty} \frac{\partial}{\partial x}\left(I_{N-1} f(u_N)\right) u_N\, w\, dx = \int_{-\infty}^{\infty} \frac{\partial}{\partial x}\left(f(u_N)\right) u_N\, w\, dx \\ - \int_{-\infty}^{\infty} \frac{\partial}{\partial x}\left((I - I_{N-1}) f(u_N)\right) u_N\, w\, dx.\end{aligned}$$

The first summand is bounded as for u_N:

$$\left| \int_{-\infty}^{\infty} \frac{\partial}{\partial x}\left(f(u_N)\right) u_N\, w\, dx \right| \leq C \left\| u_N(\cdot, t) \right\|_{0,w}^2. \tag{4.18}$$

In the second term, we integrate by parts:

$$-\int_{-\infty}^{\infty} \frac{\partial}{\partial x}\left((I - I_{N-1}) f(u_N)\right) u_N\, w\, dx = \int_{-\infty}^{\infty} (I - I_{N-1}) f(u_N) \left(\frac{\partial u_N}{\partial x} + 2\, x\, u_N\right) w\, dx.$$

By the Cauchy–Schwartz inequality, followed by (2.9) and (2.6), one can deduce that

$$\begin{aligned}&\left| \int_{-\infty}^{\infty} \frac{\partial}{\partial x}\left((I - I_{N-1}) f(u_N)\right) u_N\, w\, dx \right| \\ &\quad \leq \left\| (I - I_{N-1}) f(u_N(\cdot, t)) \right\|_{0,w} \left\| 2\, x\, u_N(\cdot, t) + \frac{\partial u_N}{\partial x}(\cdot, t) \right\|_{0,w} \\ &\quad \leq C\, N^{-1/3} \left\| \frac{\partial}{\partial x}\left(f(u_N(\cdot, t))\right) \right\|_{0,w} \left\| \frac{\partial u_N}{\partial x}(\cdot, t) \right\|_{0,w} \\ &\quad \leq C\, N^{-1/3} \left\| \frac{\partial u_N}{\partial x}(\cdot, t) \right\|_{0,w}^2,\end{aligned}$$

where one has to take into account that $|f'(u_N)| \leq C$ due to the uniform boundedness assumed in (4.13).

Finally, we apply the estimate in Lemma 2 and obtain

$$\begin{aligned}&\left| \int_{-\infty}^{\infty} \frac{\partial}{\partial x}\left((I - I_{N-1}) f(u_N)\right) u_N\, w\, dx \right| \\ &\quad \leq C\, N^{-1/3} \left(\left\| \frac{\partial}{\partial x}(Q_{m_N} u_N)(\cdot, t) \right\|_{0,w}^2 + m_N^2 \left\| u_N(\cdot, t) \right\|_{0,w}^2 \right).\end{aligned} \tag{4.19}$$

By substituting (4.18) and (4.19) in (4.17) we obtain

$$\begin{aligned}\frac{1}{2} \frac{d}{dt} \left\| u_N(\cdot, t) \right\|_{0,w}^2 + \left(\epsilon_N - C\, N^{-1/3}\right) \left\| \frac{\partial}{\partial x}(Q_{m_N} u_N)(\cdot, t) \right\|_{0,w}^2 \\ \leq C \left(1 + N^{-1/3} m_N^2\right) \left\| u_N(\cdot, t) \right\|_{0,w}.\end{aligned}$$

Since $\epsilon_N = O(N^{-\theta})$ with $\theta < 1/3$ and $m_N = [O(N^{\beta})]$ with $2\beta < \theta$, $\epsilon_N - CN^{-1/3} = O(\epsilon_N)$ and $N^{-1/3}m_N^2 = o(1)$, so that

$$\frac{1}{2}\frac{d}{dt}\,\|u_N(\cdot,t)\|_{0,w}^2 + C\,\epsilon_N \Big\| \frac{\partial}{\partial x}(Q_{m_N}u_N)(\cdot,t)\Big\|_{0,w}^2 \le C\,\|u_N(\cdot,t)\|_{0,w}^2\,.$$

By arguing as in Lemma 3, we get (4.14) as well as

$$\epsilon_N \Big\| \frac{\partial}{\partial x}(Q_{m_N}u_N)\Big\|_{0,w,T} \le C(T)\,\|u_N(\cdot,0)\|_{0,w}^2\,.$$

By recalling that $\varphi \in H_w^1$, we use (2.9) to deduce that

$$\|u_N(\cdot,0)\|_{0,w}^2 = \|I_N\varphi\|_{0,w}^2 \le \|\varphi\|_{0,w}^2 + C\,N^{-1/3}\,\|\varphi\|_{1,w}^2 \le C\,\|\varphi\|_{1,w}^2\,,$$

and inequalities (4.15) and (4.16) are satisfied. □

The uniform boundedness hypothesis on $\{u_N\}$ is common in spectral viscosity approximations as can be seen in [21], [15], [14], although in some cases it is deduced from the approximate equation itself, as in [2]. In (4.3) and (4.13) an L^∞ bound for xu_N is also imposed. The factor x is related to the weight $2x = w'w^{-1}$. Another consequence of the presence of the weight is the more involved arguments needed to deduce the a priori estimates of Lemmas 3 and 4, compared with the case of Fourier or Legendre approximations, where the basis functions are orthogonal in the unweighted L^2 space.

5. Convergence to the unique entropy solution. In the proof of convergence of u_N to the unique entropy solution of (1.1), we will make use of compensated compactness arguments [4], [10], [19], in which the estimates just proved in the previous section will be determinant. Since both the spectral and the pseudospectral cases are analogous, we will detail only the proof of convergence of the first one and state the result for the pseudospectral case under the more restrictive conditions mentioned earlier.

The proof that, for any entropy pair (U,F) associated to (1.1), $\{\frac{\partial u_N}{\partial t} + \frac{\partial f(u_N)}{\partial x}\}$ and $\{\frac{\partial U(u_N)}{\partial t} + \frac{\partial F(u_N)}{\partial x}\}$ satisfy the hypothesis of Murat's lemma will be split into several lemmas.

LEMMA 5. *Under the hypothesis of Lemma* 3 *the sequence* $\{\epsilon_N \mathcal{L}(Q_{m_N}u_N)\}_{N\in\mathbb{N}}$ *tends to* 0 *in* $H^{-1}_{\mathrm{loc}}(\mathbb{R}\times(0,T))$.

Proof. From the definition of $\mathcal{L}$,

$$\mathcal{L}(Q_{m_N}u_N) = 2Q_{m_N}u_N - \frac{\partial}{\partial x}(2xQ_{m_N}u_N) - \frac{\partial}{\partial x}\left(\frac{\partial}{\partial x}(Q_{m_N}u_N)\right).$$

Let $K \subset \mathbb{R}\times(0,T)$ be compact. Since $0 \le \widehat{q}_k \le 1$ for $k = 0,\dots,N$, and by taking into account the estimate (4.4),

$$\|2Q_{m_N}u_N\|_{L^2(K)} \le 2\|Q_{m_N}u_N\|_{0,w,T} \le 2\|u_N\|_{0,w,T} \le C(T)\|\varphi\|_{0,w}.$$

By using now property (2.6) and estimate (4.5), we get

$$\|2xQ_{m_N}u_N\|_{L^2(K)} \le 2\|xQ_{m_N}u_N\|_{0,w,T} \le 2\|Q_{m_N}u_N\|_{1,w,T} \le \frac{C(T)}{\sqrt{\epsilon_N}}.$$

Finally, by applying again estimate (4.5) we can deduce that

$$\left\| \frac{\partial}{\partial x}(Q_{m_N} u_N) \right\|_{L^2(K)} \leq \|Q_{m_N} u_N\|_{1,w,T} \leq \frac{C(T)}{\sqrt{\epsilon_N}}.$$

Hence, $\epsilon_N \mathcal{L}(Q_{m_N} u_N)$ is the sum of a function tending to 0 in $L^2(K)$ and the derivatives of two functions also tending to 0 in $L^2(K)$, K being any compact subset, and, therefore, $\epsilon_N \mathcal{L}(Q_{m_N} u_N)$ tends to 0 in $H^{-1}_{\text{loc}}(\mathbb{R} \times (0,T))$. ∎

LEMMA 6. *Under the same hypothesis of Lemma* 3, $\{\frac{\partial}{\partial x}((I - \pi_{N-1}) f(u_N))\}_{N \in \mathbb{N}}$ *tends to* 0 *in* $H^{-1}_{\text{loc}}(\mathbb{R} \times (0,T))$.

Proof. Let $K \subset \mathbb{R} \times (0,T)$ be compact. From (2.7) we have

$$\|(I - \pi_{N-1}) f(u_N)\|_{L^2(K)} \leq \|(I - \pi_{N-1}) f(u_N)\|_{0,w,T} \leq C N^{-1/2} \|f(u_N)\|_{1,w,T}.$$

Without loss of generality, we can suppose that $f(0) = 0$. The uniform boundedness of u_N together with the continuity of f' allows us to deduce that if $C = \max_{|\xi| \leq c} |f'(\xi)|$, then $|f(u_N)| \leq C|u_N|$, so that, by (4.6), we obtain

$$\|(I - \pi_{N-1}) f(u_N)\|_{L^2(K)} \leq C N^{-1/2} \|u_N\|_{1,w,T} \leq \frac{C(T)}{\sqrt{N \epsilon_N}},$$

and $\frac{\partial}{\partial x}((I - \pi_{N-1}) f(u_N))$ is the derivative of a function that tends to 0 in $L^2_{\text{loc}}(\mathbb{R} \times (0,T))$, since $\epsilon_N = O(N^{-\theta})$ with $\theta < 1/2$. Therefore, the result follows. ∎

LEMMA 7. *Let* $U \in C^2(\mathbb{R})$ *be an entropy function of* (1.1)*, and assume that the hypotheses of Lemma* 3 *are satisfied. Then* $\epsilon_N U'(u_N) \mathcal{L}(Q_{m_N} u_N)$ *can be written as the sum of two terms, one of them tending to* 0 *in* $H^{-1}_{\text{loc}}(\mathbb{R} \times (0,T))$ *and the other one bounded in* $L^1(\mathbb{R} \times (0,T))$.

Proof. From the definition of $\mathcal{L}$, we can write

$$\epsilon_N U'(u_N) \mathcal{L}(Q_{m_N} u_N) = -\epsilon_N U'(u_N) \left(\frac{\partial^2}{\partial x^2}(Q_{m_N} u_N) + 2x \frac{\partial}{\partial x}(Q_{m_N} u_N) \right) = I + II,$$

where

$$I = \epsilon_N \left(2U'(u_N) Q_{m_N} u_N - \frac{\partial}{\partial x} \left(2x U'(u_N) Q_{m_N} u_N \right) - \frac{\partial}{\partial x} \left(U'(u_N) \frac{\partial}{\partial x}(Q_{m_N} u_N) \right) \right),$$

$$II = \epsilon_N U''(u_N) \frac{\partial u_N}{\partial x} \left(2x Q_{m_N} u_N + \frac{\partial}{\partial x}(Q_{m_N} u_N) \right).$$

U' is continuous and u_N uniformly bounded in L^∞ so that there exists a constant $C > 0$, independent of N, such that $|U'(u_N)| \leq C$. By arguing as in Lemma 5, $2U'(u_N) Q_{m_N} u_N$, $2xU'(u_N) Q_{m_N} u_N$, and $U'(u_N) \frac{\partial}{\partial x}(Q_{m_N} u_N)$ tend to 0 in $L^2_{\text{loc}}(\mathbb{R} \times (0,T))$, and hence I tends to 0 in $H^{-1}_{\text{loc}}(\mathbb{R} \times (0,T))$.

On the other hand, U'' is also continuous; therefore, by the Cauchy–Schwartz inequality, followed by property (2.6) and estimates (4.5) and (4.6), we get

$$\begin{aligned}
\|II\|_{L^1(\mathbb{R} \times (0,T))} &= \epsilon_N \left\| U''(u_N) \frac{\partial u_N}{\partial x} \left(2x Q_{m_N} u_N + \frac{\partial}{\partial x}(Q_{m_N} u_N) \right) \right\|_{L^1(\mathbb{R} \times (0,T))} \\
&\leq C \epsilon_N \left\| \frac{\partial u_N}{\partial x} \right\|_{L^2(\mathbb{R} \times (0,T))} \left\| 2x Q_{m_N} u_N + \frac{\partial}{\partial x}(Q_{m_N} u_N) \right\|_{L^2(\mathbb{R} \times (0,T))} \\
&\leq C \epsilon_N \|u_N\|_{1,w,T} \|Q_{m_N} u_N\|_{1,w,T} \\
&\leq C(T);
\end{aligned}$$

that is, II is bounded in $L^1(\mathbb{R} \times (0,T))$. ∎

LEMMA 8. *Let $U \in C^2(\mathbb{R})$ be an entropy function of (1.1), and assume that the hypotheses of Lemma 3 are satisfied. Then $U'(u_N)\frac{\partial}{\partial x}((I-\pi_{N-1})f(u_N))$ can be written as the sum of two terms, one of them tending to 0 in $H^{-1}_{\mathrm{loc}}(\mathbb{R}\times(0,T))$ and the other one bounded in $L^1(\mathbb{R}\times(0,T))$.*

Proof. We write $U'(u_N)\frac{\partial}{\partial x}((I-\pi_{N-1})f(u_N))$ as a difference in the following way:

$$U'(u_N)\frac{\partial}{\partial x}\left((I-\pi_{N-1})f(u_N)\right) = \frac{\partial}{\partial x}\left(U'(u_N)(I-\pi_{N-1})f(u_N)\right) - U''(u_N)\frac{\partial u_N}{\partial x}\left((I-\pi_{N-1})f(u_N)\right).$$

The first term is the derivative of a function that tends to 0 in $L^2(\mathbb{R}\times(0,T))$ since, by arguing as in Lemma 6,

$$\begin{aligned}\|U'(u_N)(I-\pi_{N-1})f(u_N)\|_{L^2(\mathbb{R}\times(0,T))} &\le C\|(I-\pi_{N-1})f(u_N)\|_{0,w,T}\\ &\le \frac{C(T)}{\sqrt{N\epsilon_N}}.\end{aligned}$$

Therefore, $\frac{\partial}{\partial x}\left(U'(u_N)(I-\pi_{N-1})f(u_N)\right)$ tends to 0 in $H^{-1}_{\mathrm{loc}}(\mathbb{R}\times(0,T))$.

To bound the L^1 norm of the second term we use the Cauchy–Schwartz inequality, followed by (2.7) and the estimates obtained in Lemma 3, yielding

$$\begin{aligned}&\left\|U''(u_N)\frac{\partial u_N}{\partial x}\left((I-\pi_{N-1})f(u_N)\right)\right\|_{L^1(\mathbb{R}\times(0,T))}\\ &\le C\left\|\frac{\partial u_N}{\partial x}\right\|_{L^2(\mathbb{R}\times(0,T))}\left\|(I-\pi_{N-1})f(u_N)\right\|_{L^2(\mathbb{R}\times(0,T))}\\ &\le C\|u_N\|_{1,w,T}N^{-1/2}\|u_N\|_{1,w,T}\\ &\le \frac{C(T)}{N^{1/2}\epsilon_N}.\end{aligned}$$

By recalling that $\epsilon_N = O(N^{-\theta})$ with $\theta < 1/2$, the result follows. ∎

We are now ready to prove the convergence of u_N, by making use of Murat's and Tartar's lemmas.

THEOREM 2. *Let $f \in C^2(\mathbb{R})$ be a nonlinear function such that $f'(0) = 0$, $\varphi \in L^2_w \cap L^\infty(\mathbb{R})$, $m_N = [O(N^\beta)]$, $\epsilon_N = O(N^{-\theta})$, with $0 < 2\beta < \theta < 1/2$, and assume that, for a given $T > 0$, the solution u_N of (4.2) verifies the uniform bound (4.3). Then $\{u_N\}$ converges in $L^p(\Omega)$ to the unique entropy solution of (1.1) for any $\Omega \subset \mathbb{R}\times[0,T]$ open and bounded and any $p \ge 1$.*

Proof. The uniform boundedness of $\{u_N\}$ in $L^\infty(\mathbb{R}\times(0,T))$ ensures that there exists a subsequence that we will still denote by $\{u_N\}$, which converges in the weak-* topology of L^∞. Let u be its limit. We will prove that u is the unique entropy solution of (1.1), and we will also show that the whole sequence tends to u in $L^p(\Omega)$ for every $p \ge 1$ and every $\Omega \subset \mathbb{R}\times(0,T)$ open and bounded.

Let (U,F) be an entropy pair associated to (1.1). We first prove that $\frac{\partial u_N}{\partial t} + \frac{\partial f(u_N)}{\partial x}$ and $\frac{\partial U(u_N)}{\partial t} + \frac{\partial F(u_N)}{\partial x}$ are in a compact set of $H^{-1}_{\mathrm{loc}}(\mathbb{R}\times(0,T))$. From (4.2), we can write

$$\begin{aligned}\frac{\partial u_N}{\partial t} + \frac{\partial f(u_N)}{\partial x} &= \frac{\partial u_N}{\partial t} + \frac{\partial}{\partial x}\left(\pi_{N-1}f(u_N)\right) + \frac{\partial}{\partial x}\left((I-\pi_{N-1})f(u_N)\right)\\ &= -\epsilon_N\mathcal{L}(Q_{m_N}u_N) + \frac{\partial}{\partial x}\left((I-\pi_{N-1})f(u_N)\right). \qquad (5.1)\end{aligned}$$

We have shown in Lemmas 5 and 6 that both summands in the right-hand side tend to 0 in $H^{-1}_{\mathrm{loc}}(\mathbb{R}\times(0,T))$, and therefore $\frac{\partial u_N}{\partial t} + \frac{\partial f(u_N)}{\partial x}$ is in a compact set of this space.

On the other hand, since $F'(u) = f'(u)U'(u)$,

$$\begin{aligned}\frac{\partial U(u_N)}{\partial t} + \frac{\partial F(u_N)}{\partial x} &= U'(u_N)\frac{\partial u_N}{\partial t} + F'(u_N)\frac{\partial u_N}{\partial x}\\ &= -\epsilon_N U'(u_N)\mathcal{L}(Q_{m_N}u_N) + U'(u_N)\frac{\partial}{\partial x}\left((I-\pi_{N-1})f(u_N)\right).\end{aligned}$$

From Lemmas 7 and 8 it follows that $\frac{\partial U(u_N)}{\partial t} + \frac{\partial F(u_N)}{\partial x}$ is the sum of two terms, one of them being in a compact set of $H^{-1}_{\mathrm{loc}}(\mathbb{R}\times(0,T))$ and the other one being bounded in the set of measures $\mathcal{M}(\mathbb{R}\times(0,T))$. Besides, $\frac{\partial U(u_N)}{\partial t} + \frac{\partial F(u_N)}{\partial x}$ is in $W^{-1,p}_{\mathrm{loc}}(\mathbb{R}\times(0,T))$ for any $p>2$, since U and F are continuous and u_N uniformly bounded in $L^\infty(\mathbb{R}\times(0,T))$. Therefore, by Murat's lemma, $\frac{\partial U(u_N)}{\partial t} + \frac{\partial F(u_N)}{\partial x}$ is in a compact set of $H^{-1}_{\mathrm{loc}}(\mathbb{R}\times(0,T))$.

We apply now Tartar's div-culr lemma [22] to conclude that u is in fact a weak solution of (1.1). Let us now turn to the analysis of the entropy condition.

$$\begin{aligned}\frac{\partial U(u_N)}{\partial t} + \frac{\partial F(u_N)}{\partial x} &= \epsilon_N e^{-x^2}\frac{\partial}{\partial x}\left(e^{x^2}U'(u_N)\frac{\partial}{\partial x}(Q_{m_N}u_N)\right)\\ &\quad -\epsilon_N U''(u_N)\frac{\partial u_N}{\partial x}\frac{\partial}{\partial x}(Q_{m_N}u_N)\\ &\quad +\frac{\partial}{\partial x}\left(U'(u_N)(I-\pi_{N-1})f(u_N)\right)\\ &\quad -U''(u_N)\frac{\partial u_N}{\partial x}\left((I-\pi_{N-1})f(u_N)\right).\end{aligned}$$

We multiply this equality by a nonnegative function test $\phi \in C^1_0(\mathbb{R}\times(0,T))$ and integrate by parts to obtain

$$\begin{aligned}&\int_0^\infty\int_{-\infty}^\infty \left(U(u_N)\frac{\partial\phi}{\partial t} + F(u_N)\frac{\partial\phi}{\partial x}\right)dx\,dt\\ &= \epsilon_N\int_0^\infty\int_{-\infty}^\infty U'(u_N)\frac{\partial}{\partial x}(Q_{m_N}u_N)\left(-2x\phi + \frac{\partial\phi}{\partial x}\right)dx\,dt\\ &\quad +\epsilon_N\int_0^\infty\int_{-\infty}^\infty U''(u_N)\frac{\partial u_N}{\partial x}\frac{\partial}{\partial x}(Q_{m_N}u_N)\phi\,dx\,dt\\ &\quad +\int_0^\infty\int_{-\infty}^\infty U'(u_N)\left((I-\pi_{N-1})f(u_N)\right)\frac{\partial\phi}{\partial x}\,dx\,dt\\ &\quad +\int_0^\infty\int_{-\infty}^\infty U''(u_N)\frac{\partial u_N}{\partial x}\left((I-\pi_{N-1})f(u_N)\right)\phi\,dx\,dt.\end{aligned}$$

Estimates (4.5) and (4.6) and property (2.7) imply that the first, third, and fourth integrals on the right-hand side tend to 0 when N tends to infinity. Besides, if $R_{m_N} = I - Q_{m_N}$, then

$$\frac{\partial u_N}{\partial x}\frac{\partial}{\partial x}(Q_{m_N}u_N) = \left(\frac{\partial u_N}{\partial x}\right)^2 - \frac{\partial u_N}{\partial x}\frac{\partial}{\partial x}(R_{m_N}u_N),$$

and, by taking into account the estimate obtained in the proof of Lemma 2 for $\|R_{m_N}\varphi\|_{1,w}$ and the bound (4.6), we have

$$\begin{aligned}\left\|\frac{\partial u_N}{\partial x}\frac{\partial}{\partial x}(R_{m_N}u_N)\right\|_{L^1(\mathbb{R}\times(0,T))} &\le \left\|\frac{\partial u_N}{\partial x}\right\|_{0,w,T}\left\|\frac{\partial}{\partial x}(R_{m_N}u_N)\right\|_{0,w,T}\\ &\le C(T)m_N\epsilon_N^{-1/2}\|\varphi\|_{0,w}.\end{aligned}$$

By taking this into account, we split the second integral in the right-hand side above into two terms, and we deduce that

$$\lim_{N\to\infty} \epsilon_N \int_0^\infty \int_{-\infty}^\infty U''(u_N)\frac{\partial u_N}{\partial x}\frac{\partial}{\partial x}(R_{m_N}u_N)\phi\,dx\,dt = 0,$$

while

$$\epsilon_N \int_0^\infty \int_{-\infty}^\infty U''(u_N)\left(\frac{\partial u_N}{\partial x}\right)^2 \phi \geq 0,$$

thanks to the convexity of U. We have proved, therefore, that u verifies the entropy condition. The uniqueness of the entropy solution allows us to assure that not only a subsequence but the whole sequence $\{u_N\}$ converges to u and, by recalling results about Young measures, the convergence in $L^p(\Omega)$ for any $p \geq 1$ and any $\Omega \subset \mathbb{R}\times(0,T)$ open and bounded is also fulfilled. ∎

The same convergence result is obtained for the pseudspectral viscosity approximation, under the more restrictive conditions of Lemma 4.

THEOREM 3. *Let $f \in C^2(\mathbb{R})$ be a nonlinear function such that $f'(0) = 0$, $\varphi \in H^1_w$, $m_N = [O(N^\beta)]$, $\epsilon_N = O(N^{-\theta})$, with $0 < 2\beta < \theta < 1/3$, and assume that, given $T > 0$, the solution u_N of (4.12) verifies (4.13). Then $\{u_N\}$ converges in $L^p(\Omega)$ to the unique entropy solution of (1.1) for any $\Omega \subset \mathbb{R}\times[0,T]$ open and bounded and any $p \geq 1$.*

6. Numerical experiments. In this section we present the numerical results obtained when applying our spectral viscosity method to Burgers' equation with initial condition $u(x,0) = h_0(x)$, that is, to the Cauchy problem

$$\begin{cases} \dfrac{\partial u}{\partial t} + \dfrac{\partial}{\partial x}\left(\dfrac{u^2}{2}\right) = 0, & x \in \mathbb{R}, t > 0, \\ u(x,0) = e^{-x^2}, & x \in \mathbb{R}. \end{cases} \tag{6.1}$$

The solution to this problem presents a shock at time $T^* = (e/2)^{1/2} \sim 1.1658$. All of the numerical results presented in this section have been recorded at time $t = 1.5 > T^*$.

Although we have proved the convergence of the spectral and pseudospectral viscosity methods, the implementation of the spectral one is expensive since the Fourier–Hermite coefficients defined as an integral must be approximated sufficiently accurately. Because of this, we will consider only the pseudospectral case. Hence our approximation $u_N(x,t) = \sum_{k=0}^N \widetilde{u}_k(t)h_k(x)$ will be the solution of

$$\begin{cases} \dfrac{\partial u_N}{\partial t} + \dfrac{1}{2}\dfrac{\partial}{\partial x}\left(I_{N-1}u_N^2\right) + \epsilon_N \mathcal{L}\left(Q_{m_N}u_N\right) = 0, \\ u_N(x,0) = I_N\varphi(x) = e^{-x^2}. \end{cases} \tag{6.2}$$

As stated in (2.8), the coefficients of the pseudospectral projection are obtained by a Gaussian quadrature formula, whose weights have been computed for different values of N by using *Mathematica*® and stored, so that there is no need to compute them each time the method is applied. To obtain the values of u_N^2 at the nodes z_j, we have used Clenshaw's formula [17] that gives an efficient and stable algorithm to compute the values of sums of the form $\sum_{k=0}^N \hat{\varphi}_k h_k(x)$. Besides, from (2.2), one has $(\sum_{k=0}^{N-1} \hat{\varphi}_k h_k(x))' = -\sum_{k=1}^N \hat{\varphi}_{k-1} h_k(x)$. Therfore, the coefficients of $\frac{\partial}{\partial x}(I_{N-1}u_N^2)$ are a nonlinear combination of those of u_N.

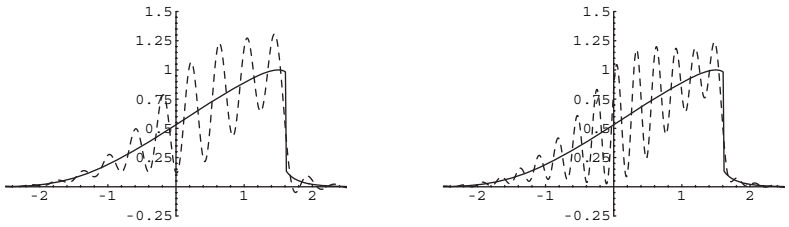


FIG. 6.1. *Graphs of the exact solution of Burgers' equation (solid line) and of the pseudospectral approximation without viscosity (dashed line), with* $N = 129$ *and* $N = 257$.

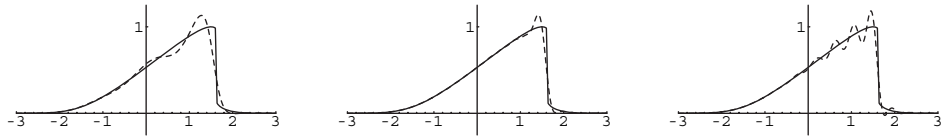


FIG. 6.2. *Graphs of the exact solution of Burgers' equation (solid line) and the viscosity pseudospectral approximation (dashed line) at time* $t = 1.5$, *for* $N = 257$, $\epsilon_N = 0.5N^{-0.33}$, *and* $m_N = [5N^{0.16}]$, *by taking* $\widehat{q}_k = \widehat{q}_k^1$, $\widehat{q}_k = \widehat{q}_k^2$, *and* $\widehat{q}_k = \widehat{q}_k^3$, *respectively, for* $k > m_N$.

On the other hand, from the definition of Q_{m_N} and the fact that h_k is an eigenfunction of $\mathcal{L}$, the coefficients of the viscosity term are obtained by multiplying those of u_N by $2(k+1)\hat{q}_k$.

Hence, the coefficients $\widetilde{u}_k(t)$, $k = 0, \ldots, N$, are the solution of a nonlinear system of ordinary differential equations that has been solved by using a fourth order Runge–Kutta method with an adaptive time step.

In Figure 6.1 we show the result of approximating u by a Hermite pseudospectral method without viscosity ($\epsilon_N = 0$) for $N = 129$ and $N = 257$. The appearance of instabilities prevents the convergence of the approximation.

The viscosity introduced by the numerical scheme (6.2) depends on the parameters ϵ_N and m_N and the operator Q_{m_N}. Convergence is ensured if

$$\epsilon_N = O(N^{-\theta}), \quad 0 < \theta < 1/3, \tag{6.3}$$

$$m_N = [O(N^{\beta})], \quad 0 < \beta < \theta/2, \tag{6.4}$$

$$1 - \frac{m_N}{k} \leq \hat{q}_k \leq 1, \quad m_N < k \leq N. \tag{6.5}$$

The bigger the coefficients $\hat{q}_k$ and ϵ_N are, the stronger the viscosity introduced is, while high values of m_N mean less viscosity since fewer coefficients are present in the viscosity term. Too much viscosity yields a worse resolution of the shock, but too little viscosity allows more oscillations that can lead to lack of convergence. By taking this into account, we have chosen $\epsilon_N = 0.5N^{-0.33}$ and $m_N = [5N^{0.16}]$, so that conditions (6.3) and (6.4) are satisfied but viscosity is not too strong.

In Figure 6.2 we show the results obtained, for $N = 257$ and $t = 1.5$, with different operators Q_{m_N} that verify in any case that $\hat{q}_N = 1$.

The first figure corresponds to taking

$$\widehat{q}_k^1 = \frac{N}{N - m_N}\left(1 - \frac{m_N}{k}\right) \quad \text{for} \quad k > m_N$$

which satisfies (6.5). The approximate solution presents few oscillations, but the shock is not very well approximated.

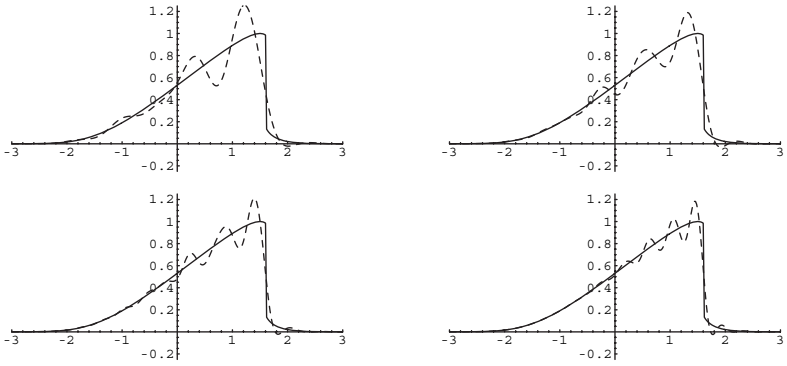


FIG. 6.3. *Graphs of the exact solution to Burgers' equation (solid line) and of the pseudospectral viscosity approximation (dashed line), by taking* $\widehat{q}_k = \widehat{q}_k^3$, $\epsilon_N = 0.5N^{-0.33}$, *and* $m_N = [5N^{0.16}]$, *for* $N = 33$, $N = 65$, $N = 129$, *and* $N = 257$.

In the second figure we present the solution obtained when

$$\widehat{q}_k^2 = \frac{k - m_N}{N - m_N} \quad \text{for } k > m_N,$$

which corresponds to coefficients laying on the straight line that joins the points $(m_N, 0)$ and $(N, 1)$, in a similar fashion as in one of the examples presented in [15]. The resulting approximation is more accurate near the shock, and there are very few oscillations.

Finally, we have taken

$$\widehat{q}_k^3 = \exp\left(-\left(\frac{k - N}{k - m_N}\right)^2\right) \quad \text{for} \quad k > m_N,$$

since in [15] the author suggests that coefficients that can be written as $\hat{q}_k = q(k/N)$, with $q \in C^\infty$, may lead to better results. This approximation is more accurate near the discontinuity, at the expense of more oscillations all over the domain.

In the three cases, the results can be slightly improved by properly choosing ϵ_N and m_N, but the third viscosity operator considered gives the best resolution of the shock.

The approximations obtained with this last operator for different values of N are shown in Figure 6.3. The improvement in the resolution of the shock when N increases is clear. However, the overshoot near the discontinuity does not decrease, and oscillations are present all along the domain. These are the main features of the so-called Gibbs phenomenon that appears when a nonregular function is approximated by its spectral or pseudospectral projection. Associated to this Gibbs phenomenon there is also a poor pointwise convergence rate even at points far away from the discontinuity.

In [18], a filter has been developed to enhance the convergence rate of the spectral projection of nonsmooth functions of L_w^2. Given $\varphi \in L_w^2$, the new approximation is defined as

$$\mathcal{F}^{\theta,p}[\pi_N \varphi](x) = \pi_p F^{N,\theta,x}(x),$$

where $F^{N,\theta,x}$ is a localization of the Nth spectral projection of φ:

$$F^{N,\theta,x}(y) = \pi_N \varphi(y) \rho\left(\frac{x - y}{\theta}\right).$$

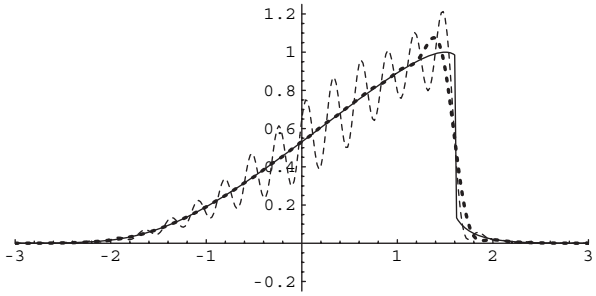


FIG. 6.4. *Graphs of the solution of Burgers' equation (solid line) and its pseudospectral viscosity approximation before filtering (dashed line) and after filtering (dotted line).*

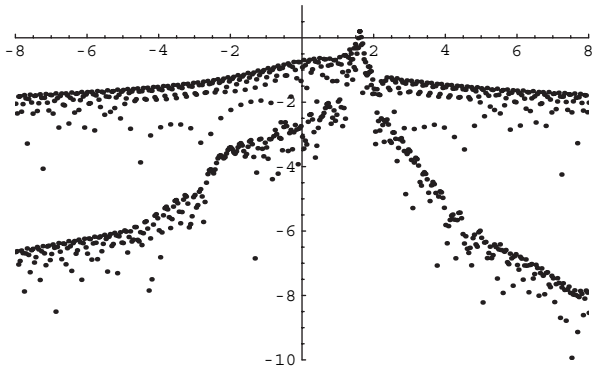


FIG. 6.5. *Logarithms of the errors multiplied by $w^{1/2}$ for the pseudospectral viscosity approximation of the solution of Burgers' equation before filtering and after filtering.*

ρ is an even smooth function with compact support, θ is a positive number, and p is an integer that will typically be taken as $p = [O(N^\alpha)]$.

If φ is piecewise C^∞ and has a jump discontinuity at $x_0 \in \mathbb{R}$, it is proven in [18] that for $\rho \in C^m$, $\alpha = \frac{1}{2} - \frac{1}{3m+1}$, and $|x - x_0| > \theta$ one has

$$e^{x^2/2} \left| \varphi(x) - \mathcal{F}^{\theta,p}[\pi_N \varphi](x) \right| = O(N^\gamma), \qquad \text{with } \gamma = -\frac{m}{4} + \frac{m}{3m+1}.$$

Spectral convergence is thus recovered, away from the discontinuity, by choosing $\rho \in C^\infty$ and $p = [O(\sqrt{N})]$.

When this filtering procedure is applied to the pseudospectral viscosity approximation, smaller artificial viscosity can be allowed. Although more oscillations will appear, the shock will be better resolved, and it is the task of the filter to diminish the oscillatory behavior and improve the convergence rate away from the discontinuity.

From the definition of $\tilde{q}_k^3$, increasing m_N makes almost no difference in the amount of viscosity introduced. Therefore we have modified only the parameter ϵ_N.

In Figure 6.4 we present the pseudospectral approximation u_N, with $\tilde{q}_k = \tilde{q}_k^3$, $m_N = [5N^{0.16}]$, and $\epsilon_N = 0.05N^{0.33}$, as well as the filtered approximation $\mathcal{F}^{\theta,p}[u_N]$ with parameters $p = [5\sqrt{N}]$, $\rho(x) = \exp(-\frac{5x^2}{1-x^2})\chi_{[-1,1]}(x)$, and $\theta = 0.9$, for $N = 257$. Figure 6.5 corresponds to the graphs of the logarithms of

$$e^{x^2/2}|u(x,t) - u_{257}(x,t)| \quad \text{and} \quad e^{x^2/2}|u(x,t) - \mathcal{F}^{\theta,p}[u_{257}](x,t)|.$$

The postprocessed solution gives a clear improvement in the convergence rate at points not too close to the discontinuity. However, the resolution of the shock is quite poor, and a better postprocessing technique should be developed in order to retain the good properties of the viscosity approximation near discontinuities.

REFERENCES

[1] J. Aguirre and J. Rivas, *Hermite pseudospectral approximations. An error estimate*, J. Math. Anal. Appl., 304 (2005), pp. 189–197.

[2] G. Q. Chen, Q. Du, and E. Tadmor, *Spectral viscosity approximations to multidimensional scalar conservation laws*, Math. Comp., 61 (1993), pp. 629–643.

[3] M. Escobedo and O. Kavian, *Variational problems related to self-similar solutions of the heat equation*, Nonlinear Anal., 11 (1987), pp. 1103–1133.

[4] L. C. Evans, *Weak Convergence Methods for Nonlinear Partial Differential Equations*, American Mathematical Society, Providence, RI, 1988.

[5] J. C. Fok, B.-Y. Guo, and T. Tang, *Combined Hermite spectral-finite difference method for the Fokker-Planck equation*, Math. Comp., 71 (2002), pp. 1497–1528.

[6] D. Funaro and O. Kavian, *Approximation of some diffusion evolution equations in unbounded domains by Hermite functions*, Math. Comp., 57 (1991), pp. 597–619.

[7] E. Godlewski and P. A. Raviart, *Hyperbolic Systems of Conservation Laws*, Societé de Mathématiques Appliquées et Industrielles, Paris, 1991.

[8] D. Gottlieb and S. A. Orszag, *Numerical Analysis of Spectral Methods: Theory and Applications*, SIAM-CBMS, Philadelphia, 1977.

[9] B.-Y. Guo, J. Shen, and C. L. Xu, *Spectral and pseudospectral approximations using Hermite functions: Application to the Dirac equation*, Adv. Comput. Math., 19 (2003), pp. 35–55.

[10] L. Hörmander, *Lectures on Nonlinear Hyperbolic Differential Equations*, Springer–Verlag, Berlin, Heidelberg, 1997.

[11] R. J. Leveque, *Numerical methods for conservation laws*, in Lectures in Mathematics ETH Zürich, Birkhäuser-Verlag, Basel, 1992.

[12] H. Ma, *Chebyshev–Legendre spectral viscosity method for nonlinear conservation laws*, SIAM J. Numer. Anal., 35 (1998), pp. 869–892.

[13] H. Ma, W. Sun, and T. Tang, *Hermite spectral methods with a time-dependent scaling for parabolic equations in unbounded domains*, SIAM J. Numer. Anal., 43 (2005), pp. 58–75.

[14] Y. Maday, S. M. Ould Kaber, and E. Tadmor, *Legendre pseudospectral viscosity method for nonlinear conservation laws*, SIAM J. Numer. Anal., 30 (1993), pp. 321–342.

[15] Y. Maday and E. Tadmor, *Analysis of the spectral vanishing viscosity method for periodic conservation laws*, SIAM J. Numer. Anal., 26 (1989), pp. 854–870.

[16] M. B. Pérez Pinilla, *Un método espectral basado en funciones de Hermite para la resolución numérica de algunas ecuaciones del calor semilineales*, Ph.D. thesis, Universidad del País Vasco, Bilbao, 1997.

[17] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C*, Cambridge University Press, Cambridge, 1992.

[18] J. Rivas, *Filtered Hermite Approximations*, preprint.

[19] D. Serre, *Systèmes de Lois de Conservation* II: *Structures Géométriques, Oscillations et Problèmes Mixtes*, Diderot Editeur, Paris, 1996.

[20] G. Szegö, *Orthogonal Polynomials*, American Mathematical Society, Providence, RI, 1939.

[21] E. Tadmor, *Convergence of spectral methods for nonlinear conservation laws*, SIAM J. Numer. Anal., 26 (1989), pp. 30–44.

[22] L. Tartar, *The compensated compactness method applied to systems of conservation laws*, NATO Sci. Ser. C Math. Phys. Sci., 111 (1983), pp. 263–285.

SIAM J. NUMER. ANAL.
Vol. 46, No. 2, pp. 1079–1095
© 2008 Society for Industrial and Applied Mathematics

NEW SOLUTION AND ANALYTICAL TECHNIQUES OF THE IMPLICIT NUMERICAL METHOD FOR THE ANOMALOUS SUBDIFFUSION EQUATION*

P. ZHUANG†, F. LIU‡, V. ANH§, AND I. TURNER§

Abstract. A physical-mathematical approach to anomalous diffusion is based on a generalized diffusion equation containing derivatives of fractional order. In this paper, an anomalous subdiffusion equation (ASub-DE) is considered. A new implicit numerical method (INM) and two solution techniques for improving the order of convergence of the INM for solving the ASub-DE are proposed. The stability and convergence of the INM are investigated by the energy method. Some numerical examples are given. The numerical results demonstrate the effectiveness of theoretical analysis. These methods and supporting theoretical results can also be applied to other fractional integro-differential equations and higher-dimensional problems.

Key words. anomalous subdiffusion equation, implicit numerical method, stability, convergence, fractional integro-differential equation

AMS subject classifications. 26A33, 45K05, 35K37, 65M12

DOI. 10.1137/060673114

1. Introduction.

Fractional differential equations have attracted considerable interest because of their ability to model many phenomena, for example, in fractal media, mathematical biology, chemistry, statistical mechanics, and biochemistry due to anomalous diffusion effects in constrained environments. A physical-mathematical approach to anomalous diffusion may be based on generalized partial differential equations containing derivatives of fractional order in space, or time, or space-time [1, 2]. These fractional diffusion equations can be derived asymptotically from basic random walk models or from generalized master equations. Anomalous diffusion, known since Richardson's treatise on turbulent diffusion in 1926, is ubiquitous in physical and biological systems where trapping and binding of particles can occur. In particular, the trapping of Brownian particles by static traps randomly distributed over either a Euclidean or a disordered substrate is a fundamental problem of nonequilibrium statistical mechanics and chemistry with a very wide range of applications [3]. Fractional kinetic equations have proved particularly useful in the context of anomalous subdiffusion [1]. The mean square displacement of the particles from the original starting site is no longer linear in time but verifies a generalized Fick's second law. Subdiffusive motion is characterized by an asymptotic long-time behavior of the mean square displacement of the form

$$\langle x^2(t)\rangle \sim \frac{2K_\gamma}{\Gamma(1+\gamma)}t^\gamma, \quad t \to \infty, \tag{1.1}$$

*Received by the editors October 24, 2006; accepted for publication (in revised form) November 7, 2007; published electronically March 5, 2008. This research was supported by Australian Research Council grant LP0348653, National Natural Science Foundation of China grant 10271098, and Natural Science Foundation of Fujian province grant Z0511009.

http://www.siam.org/journals/sinum/46-2/67311.html

†School of Mathematical Sciences, Xiamen University, Xiamen 361006, China (zxy1104@xmu.edu.cn).

‡School of Mathematical Sciences, Queensland University of Technology, GPO Box 2434, Brisbane, QLD 4001, Australia and School of Mathematical Sciences, South China University of Technology, Guangzhou 510640, China (fwliu@xmu.edu.cn).

§School of Mathematical Sciences, Queensland University of Technology, GPO Box 2434, Brisbane, QLD 4001, Australia (f.liu@qut.edu.au, v.anh@qut.edu.au, i.turner@qut.edu.au).

where γ (with $0 < \gamma < 1$) is the anomalous diffusion exponent and K_γ is the generalized diffusion coefficient. Ordinary (or Brownian) diffusion corresponds to $\gamma = 1$ with $K_1 = D$ (the ordinary diffusion coefficient).

Subdiffusive motion is particularly important in the context of complex systems such as glassy and disordered materials, in which pathways are constrained for geometric or energetic reasons. For anomalous subdiffusive random walkers, the continuum description via the ordinary diffusion equation is replaced by the fractional diffusion equation. It has been suggested that the probability density function $u(x,t)$ that describes anomalous subdiffusion particles follows the anomalous subdiffusion equation (ASub-DE) [1, 3, 4, 5]:

$$\frac{\partial u}{\partial t} = \frac{\partial^{1-\gamma}}{\partial t^{1-\gamma}}\left[K_\gamma \frac{\partial^2 u}{\partial x^2}\right], \quad 0 \le x \le a, \quad 0 < t \le T, \tag{1.2}$$

where $u(x,t)$ is the probability density that the particle that started at 0 at time 0 is at x at time t and $\frac{\partial^{1-\gamma} u}{\partial t^{1-\gamma}}$ denotes the Riemann–Liouville fractional derivative of order $1-\gamma$ defined by

$$\frac{\partial^{1-\gamma}}{\partial t^{1-\gamma}} u(x,t) = \frac{1}{\Gamma(\gamma)} \frac{\partial}{\partial t} \int_0^t \frac{u(x,\eta)}{(t-\eta)^{1-\gamma}} d\eta, \tag{1.3}$$

with $0 \le \gamma \le 1$. For $\gamma = 1$ one recovers the identity operator and for $\gamma = 0$ the ordinary first-order derivative. The integro-differential nature of the Riemann–Liouville fractional operator according to (1.3) ensures the non-Markovian nature of the subdiffusive process defined by (1.2). Equation (1.2) can be rewritten as the following equivalent time fractional diffusion equation:

$$\frac{\partial^\gamma}{\partial t^\gamma} u(x,t) - \frac{t^{-\gamma}}{\Gamma(1-\gamma)} u_0(x) = K_\gamma \frac{\partial^2}{\partial x^2} u(x,t), \tag{1.4}$$

where $u_0(x)$ is the initial value.

Anomalous subdiffusion equations have been widely studied. Schneider and Wyss [6] considered the following time fractional diffusion and wave equations:

$$\frac{\partial^\gamma u(\mathbf{x},t)}{\partial t^\gamma} = \frac{\partial^2 u(\mathbf{x},t)}{\partial \mathbf{x}^2}, \quad 0 < \gamma \le 2. \tag{1.5}$$

They derived the corresponding Green functions in closed form for arbitrary space dimensions in terms of Fox functions. Huang and Liu [7] considered the time fractional diffusion equation in n-dimensional whole space and half-space. They investigated the explicit relationships between the problems in whole space with the corresponding problems in half-space by the Fourier–Laplace transform. Mainardi [8] derived the fundamental solutions for the fractional diffusion-wave equation. Anh and Leonenko [9] presented spectral analysis of the fractional kinetic equation with random data; Gorenflo, Iskenderov, and Luchko [10] investigated mappings between solutions of fractional diffusion-wave equations; Kilbas, Saigo, and Saxena [11] presented solutions of Volterra integro-differential equations with kernels given in the form of generalized Mittag–Leffler functions.

Different numerical methods for solving the space, or time, or space-time fractional partial differential equations have been proposed. Liu et al. [12, 13] transformed the space fractional partial differential equation into a system of ordinary differential

equations (method of lines) that was then solved by using backward differentiation formulas. McLean, Sloan, and Thomée [14] considered time discretization via a Laplace transformation of an inhomogeneous integro-differential equation of parabolic type. The method is combined with a finite element discretization in the spatial variables to yield a fully discrete method. McLean, Mustrapha, and Thomee [15] also proposed numerical methods for some fractional differential equations and gave a convergence and stability analysis. Roop [16] investigated the numerical approximation of the variational solution to the fractional advection dispersion equation. Meerschaert and Tadjeran [17] examined finite difference approximations for fractional advection-dispersion flow equations. Shen and Liu [18] proposed an explicit finite difference approximation for the space fractional diffusion equation and gave an error analysis. Liu et al. [19] discussed an approximation of the Lévy–Feller advection-dispersion process by a random walk and finite difference method. Zhang, Liu, and Anh [20] proposed a numerical approximation for the Lévy–Feller fractional diffusion equation. They used the finite difference method to discretize the integro-differential equation and gave an analysis of stability and convergence in the L_1-norm. Liu et al. [21] derived an analysis of a discrete non-Markovian random walk approximation for the time fractional diffusion equation. Zhuang and Liu [22] analyzed an implicit difference approximation for the time fractional diffusion equation, and the stability and convergence of the method were also discussed. Liu et al. [23] also discussed the stability and convergence of the difference methods for the space-time fractional advection-diffusion equation. Lin and Liu [24] proposed the high order (2–6) approximations of the fractional ordinary differential equation and discussed the consistency, convergence, and stability of these fractional high order methods.

There have been some attempts on deriving numerical methods and analysis techniques for the ASub-DE (1.2). Cao, Burrage, and Abdullah [25] presented a variable coefficient fractional derivative approximation scheme and used embedding techniques to develop a variable step-size implementation for solving time fractional differential equations. They gave a stability and order analysis for a fractional Euler method and a fractional trapezoidal method. Yuste and Acedo [4] proposed an explicit finite difference method and a new von Neumann-type stability analysis for the ASub-DE (1.2). Langlands and Henry [5] also investigated this problem and proposed an implicit numerical scheme (L1 approximation). However, effective numerical methods and supporting error analyses for the ASub-DE (1.2) are still limited. The main purpose of this paper is to solve and analyze this problem by introducing an implicit numerical method and new solution techniques.

In section 2, an implicit numerical method (INM) is proposed. In sections 3 and 4, stability and convergence analyses are given for the INM. In section 5, two solution techniques to improve the order of convergence are presented. Finally, some numerical results are given in section 6 to evaluate the method.

2. An implicit numerical method for the ASub-DE. In this section, we construct a new implicit numerical method for the ASub-DE

$$\frac{\partial u}{\partial t} = \frac{\partial^{1-\gamma}}{\partial t^{1-\gamma}}\left[K_\gamma \frac{\partial^2 u}{\partial x^2} + f(x,t)\right], \quad 0 \le x \le a,\ 0 < t \le T, \tag{2.1}$$

with initial and boundary conditions:

$$u(x,0) = \phi(x), \quad 0 \le x \le a, \tag{2.2}$$

$$u(0,t) = \varphi_1(t), \quad u(a,t) = \varphi_2(t), \quad 0 \le t \le T, \tag{2.3}$$

where $0 < \gamma < 1$.

Yuste and Acedo [4] combined the forward time centered space (FTCS) method, which is well known for the numerical integration of ordinary diffusion equations, with the Grünwald–Letnikov discretization of the Riemann–Liouville derivative to obtain an explicit FTCS scheme for the ASub-DE (1.2). However, they did not give the convergence analysis and pointed out that this is not such an easy task when implicit methods are considered. Langlands and Henry [5] proposed an implicit numerical scheme (L1 approximation) for the ASub-DE (1.2) and discussed the accuracy and stability of this scheme. However, the global accuracy of the implicit numerical scheme has not been derived, and it seems that the unconditional stability for all γ in the range $0 < \gamma \leq 1$ has not been established.

Let $\Omega = [0, a] \times [0, T]$. We define the function space

$$G(\Omega) = \left\{ w(x,t) | \frac{\partial^2 w}{\partial x^2} \in C^2(\Omega) \text{ and } \frac{\partial^5 w}{\partial x^4 \partial t} \in C(\Omega) \right\}.$$

In this paper, we suppose the continuous problem (2.1)–(2.3) has a smooth solution $u(x,t) \in G(\Omega)$.

Define $t_k = k\tau, k = 0, 1, 2, \ldots, n, x_i = ih, i = 0, 1, 2, \ldots, m$, where $\tau = \frac{T}{n}$ and $h = \frac{a}{m}$ are the space and time step sizes, respectively. For convenience, we introduce the following notations:

$$Lu(x,t) = K_\gamma \frac{\partial^2 u(x,t)}{\partial x^2},$$
$$\delta_x^2 u(x,t) = u(x+h,t) - 2u(x,t) + u(x-h,t),$$

and

$$b_j = (j+1)^\gamma - j^\gamma, \quad j = 0, 1, 2, \ldots, n. \tag{2.4}$$

By integrating both sides of (2.1), we obtain

$$\begin{aligned} u(x_i, t_{k+1}) &= u(x_i, t_k) + \frac{1}{\Gamma(\gamma)} \int_0^{t_{k+1}} \frac{Lu(x_i,\eta) + f(x_i,\eta)}{(t_{k+1} - \eta)^{1-\gamma}} d\eta \\ &\quad - \frac{1}{\Gamma(\gamma)} \int_0^{t_k} \frac{Lu(x_i,\eta) + f(x_i,\eta)}{(t_k - \eta)^{1-\gamma}} d\eta \\ &= u(x_i, t_k) + \frac{1}{\Gamma(\gamma)} \int_0^{\tau} \frac{Lu(x_i,\eta) + f(x_i,\eta)}{(t_{k+1} - \eta)^{1-\gamma}} d\eta \\ &\quad + \frac{1}{\Gamma(\gamma)} \int_0^{t_k} \frac{Lv(x_i,\eta) + g(x_i,\eta)}{(t_k - \eta)^{1-\gamma}} d\eta, \end{aligned} \tag{2.5}$$

where $v(x,t) = u(x, t+\tau) - u(x,t)$ and $g(x,t) = f(x, t+\tau) - f(x,t)$. By letting

$$I_1 = \frac{1}{\Gamma(\gamma)} \int_0^{\tau} \frac{Lu(x_i,\eta) + f(x_i,\eta)}{(t_{k+1} - \eta)^{1-\gamma}} d\eta$$

and

$$I_2 = \frac{1}{\Gamma(\gamma)} \int_0^{t_k} \frac{Lv(x_i,\eta) + g(x_i,\eta)}{(t_k - \eta)^{1-\gamma}} d\eta,$$

(2.5) can be written as

$$u(x_i, t_{k+1}) = u(x_i, t_k) + I_1 + I_2.$$

For I_1, we can use the following approximation:

$$\begin{aligned} I_1 &= \frac{1}{\Gamma(\gamma)} \int_0^\tau \frac{Lu(x_i,\tau) + f(x_i,\tau)}{(t_{k+1}-\eta)^{1-\gamma}} d\eta + R_{11} \\ &= \frac{\tau^\gamma}{\Gamma(\gamma+1)} \left[(k+1)^\gamma - k^\gamma\right] \left[Lu(x_i,\tau) + f(x_i,\tau)\right] + R_{11} \\ &= \frac{\tau^\gamma}{\Gamma(\gamma+1)} b_k \left[K_\gamma \frac{1}{h^2} \delta_x^2 u(x_i,\tau) + R_{12} + f(x_i,\tau)\right] + R_{11} \\ &= \frac{\tau^\gamma}{\Gamma(\gamma+1)} b_k \left[K_\gamma \frac{1}{h^2} \delta_x^2 u(x_i,\tau) + f(x_i,\tau)\right] + R_1, \end{aligned}$$

where

$$\begin{aligned} R_{11} &= \frac{1}{\Gamma(\gamma)} \int_0^\tau \frac{Lu(x_i,\eta) - Lu(x_i,\tau) + f(x_i,\eta) - f(x_i,\tau)}{(t_{k+1}-\eta)^{1-\gamma}} d\eta, \\ R_{12} &= K_\gamma \left[\frac{\partial^2 u(x_i,\tau)}{\partial x^2} - \frac{1}{h^2} \delta_x^2 u(x_i,\tau)\right], \end{aligned}$$

and $R_1 = \frac{b_k \tau^\gamma}{\Gamma(\gamma+1)} R_{12} + R_{11}$. By noting that

$$\begin{aligned} Lu(x_i,\eta) + f(x_i,\eta) &= K_\gamma \frac{\partial^2 u(x_i,\eta)}{\partial x^2} + f(x_i,\eta) \\ &= K_\gamma \frac{\partial^2 u(x_i,\tau)}{\partial x^2} + f(x_i,\tau) + \left[K_\gamma \frac{\partial^3 u(x_i,\xi_1)}{\partial x^2 \partial t} + \frac{\partial f(x_i,\xi_1)}{\partial t}\right](\eta - \tau), \end{aligned}$$

where $0 \le \eta \le \xi_1 \le \tau$, we obtain

$$|R_{11}| \le C_1 \tau \frac{1}{\Gamma(\gamma)} \int_0^\tau \frac{1}{(t_{k+1}-\eta)^{1-\gamma}} d\eta \le \frac{C_1 \tau^{1+\gamma}}{\Gamma(\gamma+1)} b_k.$$

Again, it is apparent that

$$|R_{12}| \le C_2 h^2.$$

Hence, we have $|R_1| \le C b_k \tau^\gamma (\tau + h^2)$.

As for I_2, we can obtain the following approximation:

$$\begin{aligned} I_2 &= \frac{1}{\Gamma(\gamma)} \sum_{j=0}^{k-1} \int_{t_j}^{t_{j+1}} \frac{Lv(x_i,\eta) + g(x_i,\eta)}{(t_k-\eta)^{1-\gamma}} d\eta \\ &= \frac{1}{\Gamma(\gamma)} \sum_{j=0}^{k-1} \int_{t_j}^{t_{j+1}} \frac{Lv(x_i,t_{j+1}) + g(x_i,t_{j+1})}{(t_k-\eta)^{1-\gamma}} d\eta + R_{21} \\ &= \frac{\tau^\gamma}{\Gamma(\gamma+1)} \sum_{j=0}^{k-1} b_{k-j-1} \left[Lv(x_i,t_{j+1}) + g(x_i,t_{j+1})\right] + R_{21} \\ &= \frac{\tau^\gamma}{\Gamma(\gamma+1)} \sum_{j=0}^{k-1} b_{k-j-1} \left[L_h v(x_i,t_{j+1}) + g(x_i,t_{j+1})\right] + R_{22} + R_{21}, \end{aligned}$$

where

$$R_{21} = \frac{1}{\Gamma(\gamma)} \sum_{j=0}^{k-1} \int_{t_j}^{t_{j+1}} \frac{Lv(x_i,\eta) - Lv(x_i,t_{j+1}) + g(x_i,\eta) - g(x_i,t_{j+1})}{(t_k-\eta)^{1-\gamma}} d\eta,$$

$$R_{22} = \frac{\tau^\gamma}{\Gamma(\gamma+1)} \sum_{j=0}^{k-1} b_{k-j-1} \left[Lv(x_i,t_{j+1}) - L_h v(x_i,t_{j+1}) \right],$$

and $L_h u(x_i,t_j) = K_\gamma \frac{1}{h^2} \delta_x^2 u(x_i,t_j)$.

LEMMA 2.1. *Let $u(x,t) \in G(\Omega)$ be the solution of (2.1)–(2.3). Then we have*

(1) $|R_{21}| \le C\tau^2$;

(2) $|R_{22}| \le C\tau h^2$.

Proof. (1) When $t_j \le \eta \le t_{j+1}$, we have

$$\begin{aligned} Lv(x_i,\eta+\tau) + g(x_i,\eta+\tau) &= K_\gamma \frac{\partial^2 v(x_i,\eta+\tau)}{\partial x^2} + g(x_i,\eta+\tau) \\ &= K_\gamma \frac{\partial^2 v(x_i,t_{j+2})}{\partial x^2} + g(x_i,t_{j+2}) \\ &\quad + K_\gamma \frac{\partial^3 v(x_i,\widetilde{\eta}_1)}{\partial x^2 \partial t}(\eta - t_{j+1}) + \frac{\partial g(x_i,\widetilde{\eta}_1)}{\partial t}(\eta - t_{j+1}), \end{aligned}$$

where $\eta+\tau \le \widetilde{\eta}_1 \le t_{j+2}$.

Again,

$$\frac{\partial^3 v(x_i,\widetilde{\eta}_1)}{\partial x^2 \partial t} = \frac{\partial^3 u(x_i,\widetilde{\eta}_1+\tau)}{\partial x^2 \partial t} - \frac{\partial^3 u(x_i,\widetilde{\eta}_1)}{\partial x^2 \partial t} = \frac{\partial^4 u(x_i,\widetilde{\eta}_2)}{\partial x^2 \partial t^2}\tau,$$

$$\frac{\partial g(x_i,\widetilde{\eta}_1)}{\partial t} = \frac{\partial f(x_i,\widetilde{\eta}_1+\tau)}{\partial t} - \frac{\partial f(x_i,\widetilde{\eta}_1)}{\partial t} = \frac{\partial^2 f(x_i,\widetilde{\eta}_3)}{\partial t^2}\tau,$$

where $\widetilde{\eta}_1 \le \widetilde{\eta}_2, \widetilde{\eta}_3 \le \widetilde{\eta}_1 + \tau$. Consequently,

$$|Lu(x_i,\eta+\tau) - Lu(x_i,\eta) - [Lu(x_i,t_{j+2}) - Lu(x_i,t_{j+1})]\,| \le C_3\tau^2.$$

Thus, we have

$$|R_{21}| \le C_3\tau^2 \frac{1}{\Gamma(\gamma)} \int_0^{t_k} \frac{d\eta}{(t_k-\eta)^{1-\gamma}} \le C\tau^2.$$

(2) By using Taylor's formula, we obtain

$$\begin{aligned} Lv(x_i,t_{j+1}) &= L_h v(x_i,t_{j+1}) + \frac{h^2}{12}\frac{\partial^4 v(\xi_2,t_{j+1})}{\partial x^4} \\ &= L_h v(x_i,t_{j+1}) + \frac{h^2}{12}\left[\frac{\partial^4 u(\xi_2,t_{j+2})}{\partial x^4} - \frac{\partial^4 u(\xi_2,t_{j+1})}{\partial x^4}\right] \\ &= L_h v(x_i,t_{j+1}) + \frac{h^2\tau}{12}\frac{\partial^5 u(\xi_2,\widetilde{\eta}_4)}{\partial x^4 \partial t}, \\ &|Lv(x_i,t_{j+1}) - L_h v(x_i,t_{j+1})| \le C_3\tau h^2 \tau. \end{aligned}$$

Thus, we have

$$|R_{22}| \le C_3\tau h^2 \frac{\tau^\gamma}{\Gamma(\gamma+1)} \sum_{j=0}^{k-1} b_{k-j+1} \le C\tau h^2. \qquad \square$$

From the above results, we obtain

$$\begin{aligned} u(x_i,t_{k+1}) &= u(x_i,t_k) + r_1 b_k \delta_x^2 u(x_i,\tau) + r_2 b_k f(x_i,\tau) \\ &\quad + r_1 \sum_{j=0}^{k-1} b_{k-j-1} \left[\delta_x^2 u(x_i,t_{j+2}) - \delta_x^2 u(x_i,t_{j+1})\right] \\ &\quad + r_2 \sum_{j=0}^{k-1} b_{k-j-1} \left[f(x_i,t_{j+2}) - f(x_i,t_{j+1})\right] + R_i^{k+1}, \end{aligned}$$

where $r_1 = K_\gamma \frac{\tau^\gamma}{\Gamma(\gamma+1)h^2}, r_2 = \frac{\tau^\gamma}{\Gamma(\gamma+1)}$, and

$$|R_i^{k+1}| \le C(b_k\tau^\gamma + \tau)(\tau + h^2). \tag{2.6}$$

LEMMA 2.2. *The coefficients $b_k (k = 0, 1, 2, \ldots)$ defined by (2.4) satisfy:*
(1) $b_0 = 1, b_k > 0, k = 0, 1, 2, \ldots$;
(2) $b_k > b_{k+1}, k = 0, 1, 2, \ldots$;
(3) *there exists a positive constant $C > 0$ such that*

$$\tau \le C b_k \tau^\gamma, \quad k = 1, 2, \ldots.$$

Proof. Let $\psi_1(x) = x^\gamma$ and $\psi_2(x) = (x+1)^\gamma - x^\gamma$. For $x > 0$, it can be seen that $\psi_1(x)$ is monotone increasing and $\psi_2(x)$ is monotone decreasing. Thus, (1) and (2) in Lemma 2.2 hold.

For (3) in Lemma 2.2, by using

$$\lim_{n\to\infty} \frac{n^{\gamma-1}}{b_n} = \lim_{n\to\infty} \frac{n^{-1}}{(1+n^{-1})^\gamma - 1} = \frac{1}{\gamma},$$

we have

$$\frac{n^{\gamma-1}}{b_n} \le C_1,$$

or

$$n^{-1} \le C_1 b_n n^{-\gamma} \le C_1 b_k n^{-\gamma}.$$

Thus, from $\tau = \frac{T}{n}$, the inequality (3) can be obtained. ∎

Let $\mathbf{R}^k = (R_1^k, R_2^k, \ldots, R_{m-1}^k)^T$. By applying Lemma 2.2 with (2.6), we obtain the following lemma.

LEMMA 2.3. *Suppose that $\|\mathbf{R}^k\|_2 = \sqrt{\sum_{i=1}^{m-1} |R_i^k|^2}$. If $u(x,t) \in G(\Omega)$ is the solution of (2.1)–(2.3), then we have*

$$\|\mathbf{R}^{k+1}\|_2 \le C b_k \tau^\gamma (\tau + h^2).$$

Let u_i^k be the numerical approximation to $u(x_i,t_k)$, $f_i^k = f(x_i,t_k)$, and introduce the following notations:

$$\delta_x^2 u_i^k = u_{i+1}^k - 2u_i^k + u_{i-1}^k, \quad \Delta_x u_i^k = u_{i+1}^k - u_i^k. \tag{2.7}$$

We obtain the following implicit difference scheme:

$$\begin{aligned} u_i^{k+1} &= u_i^k + r_1 b_k \delta_x^2 u_i^1 + r_2 b_k f_i^1 \\ &\quad + r_1 \sum_{j=0}^{k-1} b_{k-j-1} \left[\delta_x^2 u_i^{j+2} - \delta_x^2 u_i^{j+1}\right] + r_2 \sum_{j=0}^{k-1} b_{k-j-1} \left[f_i^{j+2} - f_i^{j+1}\right]. \end{aligned}$$

The INM can be rewritten in the following form:

$$
\begin{aligned}
u_i^{k+1} = u_i^k + r_1\delta_x^2 u_i^{k+1} + r_1\sum_{j=0}^{k-1}(b_{j+1}-b_j)\delta_x^2 u_i^{k-j} \\
+ r_2 f_i^{k+1} + r_2\sum_{j=0}^{k-1}(b_{j+1}-b_j) f_i^{k-j},
\end{aligned} \tag{2.8}
$$

for $i = 1, 2, \ldots, m-1, k = 0, 1, 2, \ldots, n-1$.

The initial and boundary conditions are

$$
\begin{aligned}
&u_i^0 = \phi(ih), \quad i = 0, 1, 2, \ldots, m, \\
&u_0^k = \varphi_1(k\tau), \quad u_m^k = \varphi_2(k\tau), \quad k = 1, 2, \ldots, n.
\end{aligned} \tag{2.9}
$$

Equations (2.8) and (2.9) can be rewritten as the following matrix form:

$$
\begin{cases}
A\mathbf{u}^1 = \mathbf{u}^0 + r_1\mathbf{v}^1 + r_2\mathbf{f}^{\mathbf{1}}, \\
A\mathbf{u}^{k+1} = \mathbf{u}^k + r_1\mathbf{v}^{k+1} + r_1\mathbf{w}^{k+1} + r_2\mathbf{g}^{\mathbf{k+1}}, k > 0, \\
\mathbf{u}^0 = \mathbf{\Phi},
\end{cases} \tag{2.10}
$$

where

$$
A = \begin{bmatrix}
1+2r_1 & -r_1 & 0 & \cdots & 0 & 0 \\
-r_1 & 1+2r_1 & -r_1 & \cdots & 0 & 0 \\
0 & -r_1 & 1+2r_1 & \cdots & 0 & 0 \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
0 & 0 & 0 & \ldots & 1+2r_1 & -r_1 \\
0 & 0 & 0 & \ldots & -r_1 & 1+2r_1
\end{bmatrix}, \tag{2.11}
$$

$$
\mathbf{u}^k = \begin{bmatrix} u_1^k \\ u_2^k \\ \vdots \\ u_{m-2}^k \\ u_{m-1}^k \end{bmatrix}, \quad
\mathbf{f}^k = \begin{bmatrix} f_1^k \\ f_2^k \\ \vdots \\ f_{m-2}^k \\ f_{m-1}^k \end{bmatrix}, \quad
\mathbf{v}^{k+1} = \begin{bmatrix} u_0^{k+1} \\ 0 \\ \vdots \\ 0 \\ u_m^{k+1} \end{bmatrix}, \quad
\mathbf{\Phi} = \begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_{m-2} \\ \phi_{m-1} \end{bmatrix}, \tag{2.12}
$$

and $\mathbf{w}^{k+1} = \sum_{j=0}^{k-1}(b_{j+1}-b_j)\delta_x^2\mathbf{u}^{k-j}$, $\mathbf{g}^{k+1} = \mathbf{f}^{k+1} + \sum_{j=0}^{k-1}(b_{j+1}-b_j)\mathbf{f}^{k-j}$, $\phi_i = \phi(ih)$.

In (2.11) we see that the matrix A is strictly diagonally dominant with positive diagonal terms and nonpositive off-diagonal terms. Hence, the following theorem can be obtained.

THEOREM 2.4. *The discretization matrix A is invertible. Furthermore, the system* (2.8) *and* (2.9) *has a unique solution.*

3. Stability of the INM. For $\mathbf{u} = (u_1, u_2, \ldots, u_{m-1})^T, \mathbf{v} = (v_1, v_2, \ldots, v_{m-1})^T$, we define

$$
(\mathbf{u}, \mathbf{v}) = \sum_{j=1}^{m-1} u_j v_j h, \ \|\mathbf{u}\|_2 = \sqrt{(\mathbf{u}, \mathbf{u})} = \left(\sum_{j=1}^{m-1} u_j^2 h\right)^{\frac{1}{2}}. \tag{3.1}
$$

We suppose that $\widetilde{u}_i^k, \ i = 0, 1, 2, \ldots, m; \ j = 0, 1, 2, \ldots, n$ is the approximate solution of (2.8) and (2.9). The rounding error $\varepsilon_i^k = \widetilde{u}_i^k - u_i^k$ satisfies

$$
\varepsilon_i^{k+1} = \varepsilon_i^k + r_1\delta_x^2\varepsilon_i^{k+1} + r_1\sum_{j=0}^{k-1}(b_{j+1}-b_j)\delta_x^2\varepsilon_i^{k-j} \tag{3.2}
$$

and

$$\varepsilon_0^k = 0, \quad \varepsilon_m^k = 0, \quad k = 1, 2, \ldots, n. \tag{3.3}$$

Let $\mathbf{E}^k = (\varepsilon_1^k, \varepsilon_2^k, \ldots, \varepsilon_{m-1}^k)^T$. By multiplying (3.2) by $h\varepsilon_i^{k+1}$ and summing i from 1 to $m-1$, we obtain

$$\begin{aligned}
\|\mathbf{E}^{k+1}\|_2^2 &= (\mathbf{E}^{k+1}, \mathbf{E}^k) + r_1(\delta_x^2\mathbf{E}^{k+1}, \mathbf{E}^{k+1}) + r_1\sum_{j=0}^{k-1}(b_{j+1}-b_j)(\delta_x^2\mathbf{E}^{k-j}, \mathbf{E}^{k+1}) \\
&= (\mathbf{E}^{k+1}, \mathbf{E}^k) - r_1[(\varepsilon_1^{k+1})^2 h + \|\Delta_x\mathbf{E}^{k+1}\|_2^2] \\
&\quad + r_1\sum_{j=0}^{k-1}(b_{j+1}-b_j)[-\varepsilon_1^{k-j}\varepsilon_1^{k+1}h - (\Delta_x\mathbf{E}^{k-j}, \Delta_x\mathbf{E}^{k+1})] \\
&\le \frac{1}{2}[\|\mathbf{E}^{k+1}\|_2^2 + \|\mathbf{E}^k\|_2^2] - r_1|\varepsilon_1^{k+1}|^2 h - r_1\|\Delta_x\mathbf{E}^{k+1}\|_2^2 \\
&\quad + \frac{r_1}{2}\sum_{j=0}^{k-1}(b_j - b_{j+1})[|\varepsilon_1^{k+1}|^2 h + |\varepsilon_1^{k-j}|^2 h] \\
&\quad + \frac{r_1}{2}\sum_{j=0}^{k-1}(b_j - b_{j+1})[\|\Delta_x\mathbf{E}^{k-j}\|_2^2 + \|\Delta_x\mathbf{E}^{k+1}\|_2^2].
\end{aligned}$$

By noting that $\sum_{j=0}^{k-1}(b_j - b_{j+1}) = b_0 - b_k = 1 - b_k$ and $b_k > 0$, we have

$$\begin{aligned}
\|\mathbf{E}^{k+1}\|_2^2 &\le \frac{1}{2}[\|\mathbf{E^{k+1}}\|_2^2 + \|\mathbf{E^k}\|_2^2] - \frac{r_1}{2}(1+b_k)[|\varepsilon_1^{k+1}|^2 h + \|\Delta_x\mathbf{E}^{k+1}\|_2^2] \\
&\quad + \frac{r_1}{2}\sum_{j=0}^{k-1}(b_j - b_{j+1})(|\varepsilon_1^{k-j}|^2 h + \|\Delta_x\mathbf{E}^{k-j}\|_2^2) \\
&\le \frac{1}{2}[\|\mathbf{E^{k+1}}\|_2^2 + \|\mathbf{E^k}\|_2^2] - \frac{r_1}{2}[|\varepsilon_1^{k+1}|^2 h + \|\Delta_x\mathbf{E}^{k+1}\|_2^2] \\
&\quad + \frac{r_1}{2}\sum_{j=0}^{k-1}(b_j - b_{j+1})(|\varepsilon_1^{k-j}|^2 h + \|\Delta_x\mathbf{E}^{k-j}\|_2^2),
\end{aligned}$$

i.e.,

$$\begin{aligned}
&\|\mathbf{E}^{k+1}\|_2^2 + r_1\sum_{j=0}^{k} b_j\{|\varepsilon_1^{k+1-j}|^2 h + \|\Delta_x\mathbf{E}^{k+1-j}\|_2^2\} \\
&\qquad \le \|\mathbf{E}^k\|_2^2 + r_1\sum_{j=0}^{k-1} b_j\{|\varepsilon_1^{k-j}|^2 h + \|\Delta_x\mathbf{E}^{k-j}\|_2^2\}.
\end{aligned} \tag{3.4}$$

By defining the energy norm $\|\mathbf{E}^k\|_E^2 = \|\mathbf{E}^k\|_2^2 + r_1\sum_{j=0}^{k-1} b_j\{|\varepsilon_1^{k-j}|^2 h + \|\Delta_x\mathbf{E}^{k-j}\|_2^2\}$, we have

$$\|\mathbf{E}^{k+1}\|_2^2 \le \|\mathbf{E}^{k+1}\|_E^2 \le \|\mathbf{E}^k\|_E^2 \le \cdots \le \|\mathbf{E}^1\|_E^2.$$

Because $\varepsilon_i^1 = \varepsilon_i^0 + r_1\delta_x^2\varepsilon_i^1$, we obtain

$$\begin{aligned}
\|\mathbf{E}^1\|_2^2 &= (\mathbf{E}^1, \mathbf{E}^0) + r_1(\delta_x^2\mathbf{E}^1, \mathbf{E}^1) \\
&= (\mathbf{E}^1, \mathbf{E}^0) - r_1[(\varepsilon_1^1)^2 h + \|\Delta_x\mathbf{E}^1\|_2^2] \\
&\le \tfrac{1}{2}[\|\mathbf{E}^1\|_2^2 + \|\mathbf{E}^0\|_2^2] - r_1|\varepsilon_1^1|^2 h - r_1\|\Delta_x\mathbf{E}^1\|_2^2,
\end{aligned}$$

i.e.,

$$\|\mathbf{E}^1\|_E^2 = \|\mathbf{E}^1\|_2^2 + r_1 \sum_{j=0}^{k} b_j \{|\varepsilon_1^{1-j}|^2 h + \|\Delta_x \mathbf{E}^{1-j}\|_2^2\} \le \|\mathbf{E}^0\|_2^2, \tag{3.5}$$

so that $\|\mathbf{E}^{k+1}\|_2^2 \le \|\mathbf{E}^0\|_2^2$.

Hence, the following theorem of stability is obtained.

THEOREM 3.1. *The fractional implicit numerical method defined by* (2.8) *is unconditionally stable.*

4. Convergence of the INM. In this section, we discuss the convergence of the INM. Let $u(x_i, t_k), i = 0, 1, 2, \ldots, m; k = 0, 1, 2, \ldots, n$ be the exact solution of the ASub-DE (2.1)–(2.3) at mesh point (x_i, t_k). Suppose that $u(x, t)$ is the solution of (2.1)–(2.3) and $u(x, t) \in G(\Omega)$.

Define

$$\eta_i^k = u(x_i, t_k) - u_i^k, \quad i = 0, 1, 2, \ldots, m; \quad k = 0, 1, 2, \ldots, n$$

and $\mathbf{Y}^k = (\eta_1^k, \eta_2^k, \ldots, \eta_{m-1}^k)^T$. By using $u_i^k = u(x_i, t_k) - \eta_i^k$, substitution into (2.8) leads to

$$\eta_i^{k+1} = \eta_i^k + r_1 \delta_x^2 \eta_i^{k+1} + r_1 \sum_{j=0}^{k-1} (b_{j+1} - b_j) \delta_x^2 \eta_i^{k-j} + R_i^{k+1}, \tag{4.1}$$

where $i = 1, 2, \ldots, m-1; k = 0, 1, 2, \ldots, n-1$, and

$$\eta_i^0 = 0, \quad i = 0, 1, \ldots, m,$$
$$\eta_0^k = \eta_m^k = 0, \quad k = 0, 1, \ldots, n.$$

By multiplying (4.1) by $h\eta_i^{k+1}$ and summing i from 1 to $m-1$, we obtain

$$\begin{aligned} \|\mathbf{Y}^{k+1}\|_2^2 &= (\mathbf{Y}^{k+1}, \mathbf{Y}^k) + r_1 (\delta_x^2 \mathbf{Y}^{k+1}, \mathbf{Y}^{k+1}) \\ &\quad + r_1 \sum_{j=0}^{k-1} (b_{j+1} - b_j)(\delta_x^2 \mathbf{Y}^{k-j}, \mathbf{Y}^{k+1}) + (\mathbf{R}^{k+1}, \mathbf{Y}^{k+1}). \end{aligned} \tag{4.2}$$

For $j = 0, 1, \ldots, k+1$, we have

$$(\delta_x^2 \mathbf{Y}^j, \mathbf{Y}^{k+1}) = -\eta_1^j \eta_1^{k+1} - (\Delta_x \mathbf{Y}^j, \Delta_x \mathbf{Y}^{k+1}).$$

By using $|v \cdot w| \le \sigma v^2 + \frac{1}{4\sigma} w^2$, $\sigma > 0$, we obtain

$$(\mathbf{Y}^{k+1}, \mathbf{Y}^j) \le \frac{1}{2} \left[\|\mathbf{Y}^{k+1}\|^2 + \|\mathbf{Y}^j\|^2 \right], \tag{4.3}$$

$$\eta_1^{k+1} \eta_1^j \le \frac{1}{2} \left[|\eta_1^{k+1}|^2 + |\eta_1^j|^2 \right], \tag{4.4}$$

and

$$|(\mathbf{R}^{k+1}, \mathbf{Y}^{k+1})| \le \frac{r_1 h^2 b_k}{a^2} \|\mathbf{Y}^{k+1}\|_2^2 + \frac{a^2}{4 r_1 h^2 b_k} \|\mathbf{R}^{k+1}\|_2^2. \tag{4.5}$$

Noting that $\sum_{j=0}^{k-1}(b_j - b_{j+1}) = b_0 - b_k = 1 - b_k$ and $b_k > 0$, similarly to the proof of the stability, we have

$$\begin{aligned}\|\mathbf{Y}^{k+1}\|_2^2 \le{}& \frac{1}{2}[\|\mathbf{Y}^{\mathbf{k+1}}\|_2^2 + \|\mathbf{Y}^k\|_2^2] - \frac{r_1}{2}(1+b_k)[|\eta_1^{k+1}|^2 h + \|\Delta_x \mathbf{Y}^{k+1}\|_2^2] \\ &+ \frac{r_1}{2}\sum_{j=0}^{k-1}(b_j - b_{j+1})(|\eta_1^{k-j}|^2 h + \|\Delta_x \mathbf{Y}^{k-j}\|_2^2) \\ &+ \frac{r_1 h^2 b_k}{a^2}\|\mathbf{Y}^{k+1}\|_2^2 + \frac{a^2}{4r_1 h^2 b_k}\|\mathbf{R}^{k+1}\|_2^2.\end{aligned} \tag{4.6}$$

LEMMA 4.1. *Given* $\|\mathbf{Y}^k\|_\infty = \max_{1\le i\le m-1}|\eta_i^k|$, *then*

$$\|\mathbf{Y}^k\|_2^2 \le a\|\mathbf{Y}^k\|_\infty^2 \le \frac{a^2}{2h^2}\left[h|\eta_1^k|^2 + \|\Delta_x \mathbf{Y}^k\|_2^2\right].$$

Proof. The first inequality is apparent.
For the second inequality, let $|\eta_{i_0}^k| = \max_{1\le i\le m-1}|\eta_i^k|$, and then

$$\eta_{i_0}^k = \eta_1^k + \sum_{j=1}^{i_0-1}\Delta_x \eta_j^k, \quad \eta_{i_0}^k = -\sum_{j=i_0}^{m-1}\Delta_x \eta_j^k.$$

Thus

$$2|\eta_{i_0}^k| \le |\eta_1^k| + \sum_{j=1}^{m-1}|\Delta_x \eta_j^k|.$$

By using the Cauchy–Schwarz inequality, we have

$$\begin{aligned}4|\eta_{i_0}^k|^2 &\le 2m\left[|\eta_1^k| + \sum_{j=1}^{m-1}|\Delta_x \eta_j^k|\right]^2 \\ &\le \frac{2a}{h^2}\left[h|\eta_1^k|^2 + \|\Delta_x \mathbf{Y}^k\|_2^2\right].\end{aligned}$$

Therefore,

$$\|\mathbf{Y}^k\|_\infty^2 \le \frac{a}{2h^2}\left[h|\eta_1^k|^2 + \|\Delta_x \mathbf{Y}^k\|_2^2\right]. \qquad \square$$

By applying Lemma 4.1, we have

$$\frac{r_1 h^2 b_k}{a^2}\|\mathbf{Y}^{k+1}\|_2^2 \le \frac{r_1 h^2 b_k}{a^2}\cdot a\|\mathbf{Y}^{k+1}\|_\infty^2 \le \frac{b_k}{2}r_1\left[|\eta_1^{k+1}|^2 h + \|\Delta_x \mathbf{Y}^{k+1}\|_2^2\right]. \tag{4.7}$$

Hence, from (4.6) and (4.7), we have

$$\begin{aligned}\|\mathbf{Y}^{k+1}\|_2^2 \le{}& \frac{1}{2}[\|\mathbf{Y}^{k+1}\|_2^2 + \|\mathbf{Y}^{\mathbf{k}}\|_2^2] - \frac{r_1}{2}[|\eta_1^{k+1}|^2 h + \|\Delta_x \mathbf{Y}^{k+1}\|_2^2] \\ &+ \frac{r_1}{2}\sum_{j=0}^{k-1}(b_j - b_{j+1})(|\eta_1^{k-j}|^2 h + \|\Delta_x \mathbf{Y}^{k-j}\|_2^2) + \frac{a^2}{4r_1 h^2 b_k}\|\mathbf{R}^{k+1}\|_2^2.\end{aligned}$$

Let $\rho_k = \|\mathbf{Y}^k\|_2^2 + r_1 \sum_{j=0}^{k-1} b_j \{|\eta_1^{k-j}|^2 h + \|\Delta_x \mathbf{Y}^{k-j}\|_2^2\}$, and then

$$\rho_{k+1} \le \rho_k + \frac{a^2}{4r_1 h^2 b_k} \|\mathbf{R}^{k+1}\|_2^2, \tag{4.8}$$

$$\rho_0 = 0. \tag{4.9}$$

According to Lemma 2.3, and $r_1 = K_r \frac{\tau^\gamma}{\Gamma(\gamma+1)h^2}$, we have

$$\rho_{k+1} \le \rho_k + C b_k \tau^\gamma (\tau + h^2)^2. \tag{4.10}$$

Hence, we obtain

$$\rho_{k+1} \le C \sum_{j=0}^{k} b_j \tau^\gamma (\tau + h^2)^2. \tag{4.11}$$

By noting that $\sum_{j=0}^{k} b_j \tau^\gamma = (k+1)^\gamma \tau^\gamma \le T^\gamma$ and $\|\mathbf{Y}^{k+1}\|_2^2 \le \rho_{k+1}$, we have

$$\|\mathbf{Y}^{k+1}\|_2^2 \le C T^\gamma (\tau + h^2)^2.$$

Consequently, the following theorem of convergence is obtained.

THEOREM 4.2. *Let $u(x,t) \in G(\Omega)$ be the solution of (2.1)–(2.3). Then the fractional implicit difference scheme defined by (2.8) is convergent, and there exists a positive constant $C > 0$ such that*

$$\|\mathbf{Y}^{k+1}\|_2 \le C(\tau + h^2), \quad k = 0, 1, 2, \ldots, n-1. \tag{4.12}$$

5. Improving the order of convergence of the INM. In this section, we propose two techniques to improve the rate of convergence of the INM.

5.1. The extrapolation method (EM). According to Theorem 4.2, we know that the INM converges at the rate $O(\tau + h^2)$. In order to improve the order of convergence, we apply the INM on a (coarse) grid $\Delta t = \tau$ and then on a finer grid of size $\tau/2$ with the same Δx. The extrapolated solution is then computed from $u(x_i, t) \approx 2u_i^{2k}(\tau/2) - u_i^k(\tau)$, $i = 1, 2, \ldots, m-1$, where $t = t_k$ on the coarse grid, while $t = t_{2k}$ on the fine grid. Here $u_i^k(\tau)$ and $u_i^{2k}(\tau/2)$ are numerical solutions on the coarse grid and the fine grid, respectively. Thus, the extrapolation method can be used to obtain a solution with convergence order $O(\tau^2 + h^2)$.

5.2. Improved difference scheme (IDS). We use a technique of increasing the order of local truncation error to obtain an IDS. By integrating both sides of (2.1) for t, we obtain

$$u(x_i, t_{k+1}) = u(x_i, 0) + \frac{1}{\Gamma(\gamma)} \int_0^{t_{k+1}} \frac{K_\gamma \frac{\partial^2 u(x_i,\eta)}{\partial x^2} + f(x_i, \eta)}{(t_{k+1} - \eta)^{1-\gamma}} d\eta,$$

or

$$u(x_i, t_{k+1}) = u(x_i, 0) + \frac{1}{\Gamma(\gamma)} \sum_{j=0}^{k} \int_{t_j}^{t_{j+1}} \frac{K_\gamma \frac{\partial^2 u(x_i,\eta)}{\partial x^2} + f(x_i, \eta)}{(t_{k+1} - \eta)^{1-\gamma}} d\eta.$$

LEMMA 5.1. *If the function $v(x,t)$ is sufficiently smooth, then*

$$v(x_i, \eta) = \frac{(t_{j+1} - \eta) v(x_i, t_j) + (\eta - t_j) v(x_i, t_{j+1})}{\tau} + O(\tau^2). \tag{5.1}$$

Proof. The estimate (5.1) can be obtained from the following formulas:

$$v(x_i,t_j)=v(x_i,\eta)+\frac{\partial v(x_i,\eta)}{\partial t}(t_j-\eta)+\frac{1}{2}\frac{\partial^2 v(x_i,\xi_1)}{\partial t}(t_j-\eta)^2,$$

$$v(x_i,t_{j+1})=v(x_i,\eta)+\frac{\partial v(x_i,\eta)}{\partial t}(t_{j+1}-\eta)+\frac{1}{2}\frac{\partial^2 v(x_i,\xi_2)}{\partial t}(t_{j+1}-\eta)^2,$$

where $t_j\le\xi_1\le\eta\le\xi_2\le t_{j+1}$. ∎

By applying Lemma 5.1, we have

$$\begin{aligned}u(x_i,t_{k+1})=u(x_i,0)&+r_1\sum_{j=0}^{k}\left[c_j^{(1)}\delta_x^2u(x_i,t_{k-j+1})+c_j^{(2)}\delta_x^2u(x_i,t_{k-j})\right]\\&+r_2\sum_{j=0}^{k}\left[c_j^{(1)}f(x_i,t_{k-j+1})+c_j^{(2)}f(x_i,t_{k-j})\right]+R_i^{k+1},\end{aligned}\tag{5.2}$$

where $r_1=\frac{K_\gamma\tau^\gamma}{\Gamma(\gamma+1)h^2}$, $r_2=\frac{\tau^\gamma}{\Gamma(\gamma+1)}$, and

$$c_j^{(1)}=(j+1)^\gamma-\frac{1}{\gamma+1}\left[(j+1)^{\gamma+1}-j^{\gamma+1}\right],\tag{5.3}$$

$$c_j^{(2)}=\frac{1}{\gamma+1}\left[(j+1)^{\gamma+1}-j^{\gamma+1}\right]-j^\gamma,\tag{5.4}$$

$$|R_i^{k+1}|\le Ct_{k+1}^{\gamma+1}(\tau^2+h^2).\tag{5.5}$$

Thus, we obtain the following improved difference scheme:

$$u_i^{k+1}=u_i^0+r_1\sum_{j=0}^{k}\left[c_j^{(1)}\delta_x^2u_i^{k-j+1}+c_j^{(2)}\delta_x^2u_i^{k-j}\right]+r_2\sum_{j=0}^{k}\left[c_j^{(1)}f_i^{k-j+1}+c_j^{(2)}f_i^{k-j}\right].$$

6. Numerical results. In order to demonstrate the effectiveness of our theoretical analysis, two examples are now presented. To simplify the computation, (2.8) is rewritten as

$$\begin{aligned}u_i^{l+1}-u_i^l=r_1\delta_x^2u_i^{l+1}&+r_1\sum_{j=0}^{l-1}(b_{j+1}-b_j)\delta_x^2u_i^{l-j}\\&+r_2f_i^{l+1}+r_2\sum_{j=0}^{l-1}(b_{j+1}-b_j)f_i^{l-j}.\end{aligned}\tag{6.1}$$

By summing l from 0 to $k-1$, we obtain

$$u_i^k=u_i^0+r_1\sum_{j=0}^{k-1}b_j\delta_x^2u_i^{k-j}+r_2\sum_{j=0}^{k-1}b_jf_i^{k-j},$$

where $i=1,2,\dots,m-1$; $k=1,2,\dots,n$.

Example 1. Consider the following ASub-DE:

$$\frac{\partial u(x,t)}{\partial t}=\frac{\partial^{1-\gamma}}{\partial t^{1-\gamma}}\left[0.5\Gamma(\gamma+3)\frac{\partial^2u(x,t)}{\partial x^2}+0.5\Gamma(\gamma+3)(t^2-t^{2+\gamma})e^x\right],\tag{6.2}$$

TABLE 6.1
The error $\|\mathbf{Y}^n\|_2$ for the INM, EM, and IDS and the effect of the grid size reduction at time $t = 1.0$ ($\gamma = 0.9$).

Δx	Δt	INM	EM	IDS
$\frac{1}{10}$	$\frac{1}{10}$	1.4252E-2	3.5158E-4	3.5663E-5
$\frac{1}{15}$	$\frac{1}{15}$	9.6200E-3	1.6191E-4	8.4894E-6
$\frac{1}{20}$	$\frac{1}{20}$	7.2613E-3	1.2173E-4	8.1641E-6
$\frac{1}{25}$	$\frac{1}{25}$	5.8285E-3	1.1443E-4	6.4540E-6

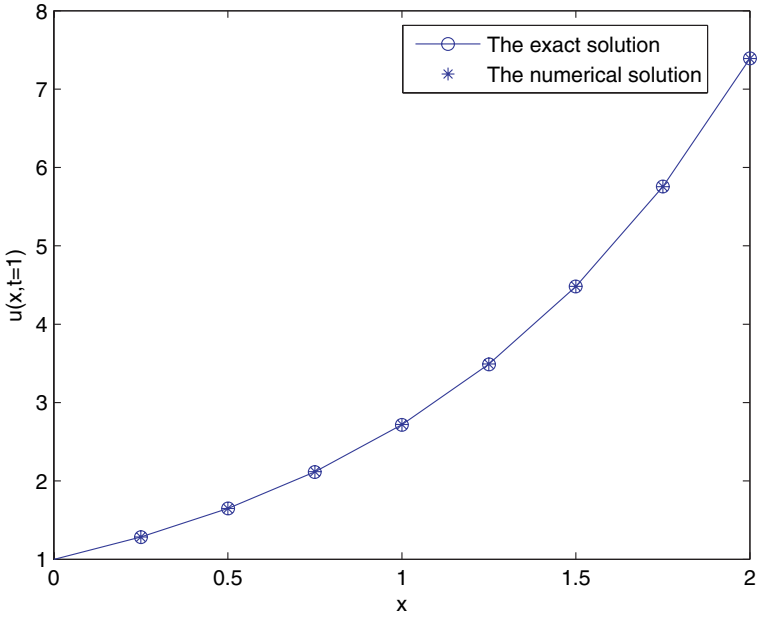


FIG. 6.1. *The comparison of the exact solution and the numerical solution (INM) when $\gamma = 0.9$ and $t = 1$.*

with the boundary and initial conditions

$$u(0,t) = t^{2+\gamma}, \quad u(1,t) = et^{2+\gamma}, \quad u(x,0) = 0. \tag{6.3}$$

The exact solution of (6.2) and (6.3) is $u(x,t) = t^{2+\gamma}e^x$, which can be obtained by evaluating the fractional derivative formula [26, 27].

Table 6.1 shows the numerical errors at $t = 1$ between the exact solution and the numerical solution obtained by the INM, the EM, and the IDS. From Table 6.1, all three methods are in excellent agreement with the exact solution, and the convergence order of the EM and IDS is improved significantly.

A comparison of the exact solution and the numerical solution by using INM with various time and space steps for ASub-DE at $t = 1.0$ and $\gamma = 0.9$ is shown in Figure 6.1. It is apparent that the numerical solution (INM) is in excellent agreement with the exact solution.

Example 2. Consider the following ASub-DE:

$$\frac{\partial u(x,t)}{\partial t} = \frac{\partial^{1-\gamma}}{\partial t^{1-\gamma}}\left[\frac{\partial^2 u(x,t)}{\partial x^2}\right], \tag{6.4}$$

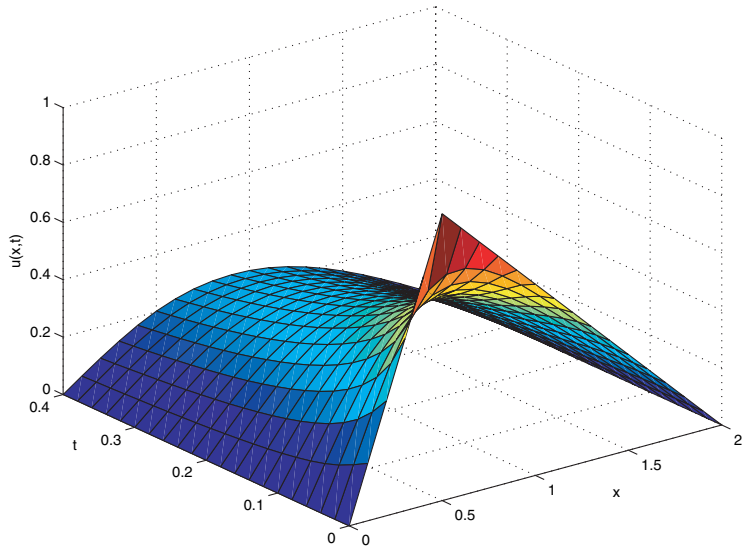


FIG. 6.2. *The numerical approximation when* $\gamma = 0.8$.

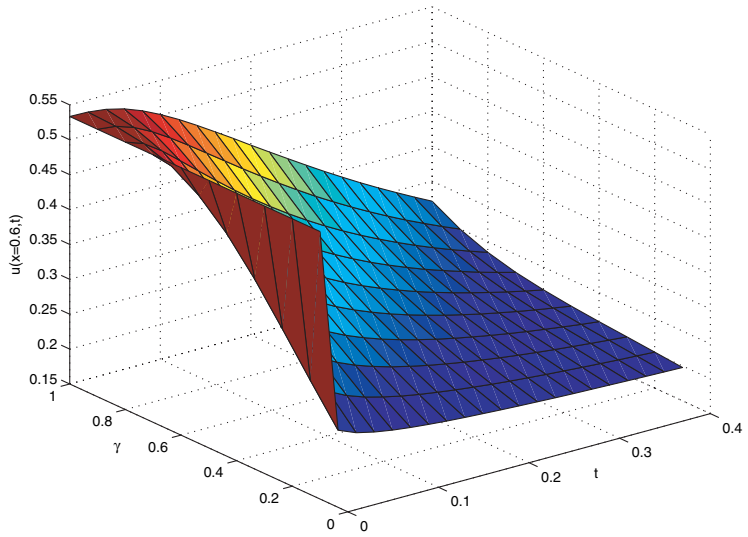


FIG. 6.3. *The numerical approximation* $u(x,t)$ *for various* γ *when* $x = 0.6$.

with initial and boundary conditions

$$u(x,0) = \begin{cases} 2x, & 0 \le x \le 0.5, \\ \frac{4-2x}{3}, & 0.5 \le x \le 2, \end{cases}$$
$$u(0,t) = u(2,t) = 0, \quad 0 \le t \le 0.4.$$

The evolution results for the INM when $\gamma = 0.8$, $0 \le t \le 0.4$, $0 \le x \le 2$, $0.1 \le \gamma \le 1$, $0 \le t \le 0.4$, $x = 0.6$, and $0.1 \le \gamma \le 1$, $t = 0.4$, $0 \le x \le 2$ are shown in Figures 6.2, 6.3, and 6.4, respectively. Figures 6.2–6.4 show that the system exhibits subdiffusion behaviors. From Figure 6.4, it can be seen that the solution continuously depends on the time fractional derivative.

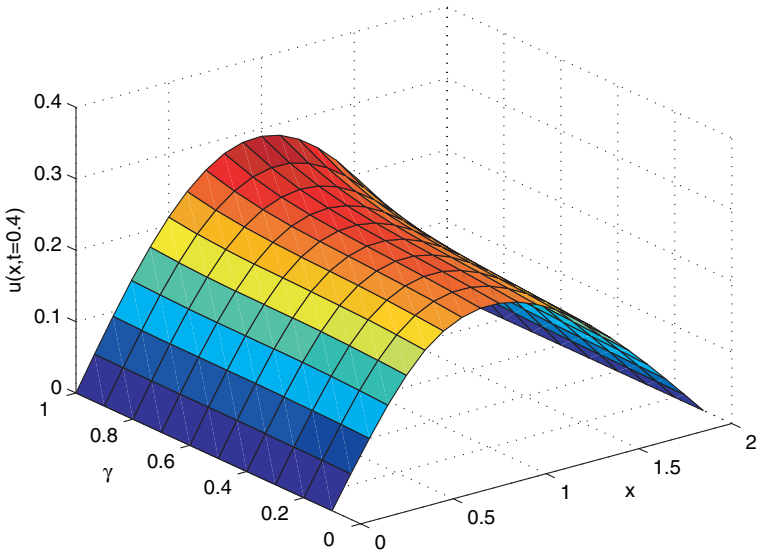


FIG. 6.4. *The numerical approximation $u(x,t)$ for various γ when $t = 0.4$.*

7. Conclusions. In this paper, an INM for the fractional subdiffusion equation in a bounded domain has been described and demonstrated. We prove that the INM is unconditionally stable and convergent. Two techniques to improve the convergence rate are presented. Both the EM and the IDS techniques provide computationally effective tools for simulating the behavior of the solution of the ASub-DE. These methods and analytical techniques can also be extended to any fractional integro-differential equations and higher-dimensional problems.

Acknowledgment. The authors thank the referees for their constructive comments and suggestions

REFERENCES

[1] R. METZLER AND J. KLAFTER, *The random walk's guide to anomalous diffusion: A fractional dynamics approach*, Phys. Rep., 339 (2000) pp. 1–77.
[2] R. KUTNER, A. PEKALSKI, AND K. SZNAJD-WERON, *Anomalous Diffusion, Basics to Application*, Springer, Berlin, 1999.
[3] S.B. YUSTE AND L. ACEDO, *Some exact results for the trapping of subdiffusive particles in one dimension*, Phys. A, 336 (2004), pp. 334–346.
[4] S.B. YUSTE AND L. ACEDO, *An explicit finite difference method and a new von Neumann-type stability analysis for fractional diffusion equations*, SIAM J. Numer. Anal., 42 (2005), pp. 1862–1874.
[5] T.A.M. LANGLANDS AND B.I. HENRY, *The accuracy and stability of an implicit solution method for the fractional diffusion equation*, J. Comput. Phys., 205 (2005), pp. 719–736.
[6] W.R. SCHNEIDER AND W. WYSS, *Fractional diffusion and wave equations*, J. Math. Phys., 30 (1989), pp. 134–144.
[7] F. HUANG AND F. LIU, *The time fractional diffusion and advection-dispersion equation*, ANZIAM J., 46 (2005), pp. 317–330.
[8] F. MAINARDI, *The fundamental solutions for the fractional diffusion-wave equation*, Appl. Math. Lett., 9 (1996), pp. 23-28
[9] V.V. ANH AND N.N. LEONENKO, *Spectral analysis of fractional kinetic equations with random data*, J. Stat. Phys., 104 (2001), pp. 1349–1387.
[10] R. GORENFLO, A. ISKENDEROV, AND YU. LUCHKO, *Mapping between solutions of fractional diffusion-wave equations*, Fract. Calc. Appl. Math., 3 (2000), pp. 75–86.

[11] A.A. Kilbas, M. Saigo, and R.K. Saxena, *Solutions of Volterra integro-differential equations with generalised Mittag-Leffler function in the kernels*, J. Integral Equations, 14 (2002), pp. 377–396.

[12] F. Liu, V. Anh, and I. Turner, *Numerical Solution of the Space Fractional Fokker-Planck Equation*, J. Comput. Appl. Math., 166 (2004), pp. 209–219.

[13] F. Liu, V. Anh, I. Turner, and P. Zhuang, *Numerical simulation for solute transport in fractal porous media*, ANZIAM J., 45 (2004), pp. 461–473.

[14] W. McLean, I. H. Sloan, and V. Thomée, *Time discretization via Laplace transformation of an integro-differential equation of parabolic type*, Numer. Math., 102 (2006), pp. 497–522.

[15] W. McLean, K. Mustrapha, and V. Thomee, *Numerical methods for some fractional-order partial differential equations*, in Proceedings of the International Symposium on Fractional Calculus, Dunedin, University of Otago, 2006, pp. 9–13.

[16] J.P. Roop, *Computational aspects of FEM approximation of fractional advection dispersion equation on bounded domains in* $\mathtt{R}^2$, J. Comput. Appl. Math., 193 (2006), pp. 243–268.

[17] M. Meerschaert and C. Tadjeran, *Finite difference approximations for fractional advection-dispersion flow equations*, J. Comput. Appl. Math., 172 (2004), pp. 65–77.

[18] S. Shen and F. Liu, *Error analysis of an explicit finite difference approximation for the space fractional diffusion*, ANZIAM J., 46 (2005), pp. 871–887.

[19] Q. Liu, F. Liu, I. Turner, and V. Anh, *Approximation of the Levy-Feller advection-dispersion process by random walk and finite difference method*, J. Phys. Comp., 222 (2007), pp. 57–70.

[20] H. Zhang, F. Liu, and V. Anh, *Numerical approximation of Levy-Feller diffusion equation and its probability interpretation*, J. Comput. Appl. Math., 206 (2007), pp. 1098–1115.

[21] F. Liu, S. Shen, V. Anh, and I. Turner, *Analysis of a discrete non-Markovian random walk approximation for the time fractional diffusion equation*, ANZIAM J., 46 (2005), pp. 488–504.

[22] P. Zhuang and F. Liu, *Implicit difference approximation for the time fractional diffusion equation*, J. Appl. Math. Comput., 22 (2006), pp. 87–99.

[23] F. Liu, P. Zhuang, V. Anh, I. Turner, and K. Burrage, *Stability and convergence of the difference methods for the space-time fractional advection-diffusion equation*, Appl. Math. Comput., 191 (2007), pp. 12–20.

[24] R. Lin and F. Liu, *Fractional high order methods for the nonlinear fractional ordinary differential equation*, Nonlinear Anal., 66 (2007), pp. 856–869.

[25] X. Cao, K. Burrage, and F. Abdullah, *A variable stepsize implementation for fractional differential equations*, BIT, submitted.

[26] K.B. Oldham and J. Spanier, *The Fractional Calculus*, Academic Press, New York and London, 1974.

[27] I. Sokolov, J. Klafter, and A. Blumen, *Fractional Kinetics*, Physics Today, November, (2002), pp. 48–54; also available online from http://www.physicstoday.org.

SIAM J. NUMER. ANAL.
Vol. 46, No. 3, pp. 1097–1115
© 2008 Society for Industrial and Applied Mathematics

STABLE AND CONVERGENT UNSYMMETRIC MESHLESS COLLOCATION METHODS*

LEEVAN LING† AND ROBERT SCHABACK‡

Abstract. In the theoretical part of this paper, we introduce a simplified proof technique for error bounds and convergence of a variation of Kansa's well-known unsymmetric meshless collocation method. For a numerical implementation of the convergent variation, a previously proposed greedy technique is coupled with linear optimization. This algorithm allows a fully adaptive on-the-fly data-dependent meshless selection of test and trial spaces. The new method satisfies the assumptions of the background theory, and numerical experiments demonstrate its stability.

Key words. Kansa's method, convergence, error bounds, linear optimization

AMS subject classifications. 65N12, 65N15, 65N70

DOI. 10.1137/06067300X

1. Introduction.

A general framework for solving PDE problems in strong or weak form by kernel-based meshless methods was outlined in [13]. It writes the PDE problem as uncountably many simultaneous scalar equations

$$\lambda(u) = f_\lambda \in \mathbb{R} \quad \text{for all} \quad \lambda \in \Lambda. \tag{1.1}$$

We call Λ the *test set*, because applying all functionals λ from Λ to a function u tests whether the function u solves the given problem. If several differential or boundary operators are involved, we simply put everything into a single set Λ of functionals of various types.

Testing can be done in *strong* and *weak* forms. For testing in *strong form*, the set Λ consists of infinitely many linear real-valued functionals λ that usually take the form of point evaluations of functions or their derivatives at points inside a domain or on some boundary or interface layer.

Testing in *weak* form uses functionals

$$\lambda_{v,a}(u) := a(u, v),$$

which evaluate bilinear forms $a(.,.)$ after inserting a *test function* v. These bilinear forms may contain derivatives, and they always require an integration.

No matter how testing is done, we call (1.1) a *generalized interpolation problem*, and we assume it to be solvable by a function u^* that generates the data via $f_\lambda := \lambda(u^*)$ for all $\lambda \in \Lambda$.

To make generalized interpolation problems numerically accessible, we have to *discretize* them. Discretization on the *trial* side is done by choosing a finite-dimensional *trial space* $\mathcal{U}$ of functions which can approximate the exact solution u^* well. It will be the space of functions from which the numerical solution u will be constructed, and

*Received by the editors Ocober 23, 2006; accepted for publication (in revised form) October 29, 2007; published electronically March 7, 2008.
http://www.siam.org/journals/sinum/46-3/67300.html
†Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong (lling@hkbu.edu.hk). The research of this author was supported in part by HKBU FRGs.
‡Institut für Numerische und Angewandte Mathematik, Georg-August-Universität Göttingen, Lotzestraße 16-18, D-37083 Göttingen, Germany (schaback@math.uni-goettingen.de).

we assume it to be spanned by a basis $\{u_1, \ldots, u_m\}$ as

$$u := \sum_{j=1}^{m} \alpha_j u_j \ \in \mathcal{U} := \mathrm{span}\{u_1, \ldots, u_m\}. \tag{1.2}$$

Discretization on the *test* side just consists in replacing the infinite set Λ by some finite unstructured subset $\Lambda_n := \{\lambda_1, \ldots, \lambda_n\}$. The space spanned by these functionals can be called the *test space*.

Then the discretized problem reads as

$$\lambda_i(u) = \sum_{j=1}^{m} \alpha_j \lambda_i(u_j) = f_{\lambda_i} = \lambda_i(u^*), \quad 1 \le i \le n, \tag{1.3}$$

when written as an unsymmetric system of linear equations for a function u of the trial space $\mathcal{U}$. In most cases, the system will have $n \ge m$, but it is always a hard problem to study the matrix with entries $\lambda_i(u_j)$, e.g., to assert that it has rank m.

Weak formulations use functionals of the form $\lambda_i(u) := a_i(u, v_i)$, with certain bilinear forms a_i and *test functions* $v_1, \ldots, v_n$ such that the discretized problem takes the familiar form

$$\lambda_i(u) = a_i(u, v_i) = \sum_{j=1}^{m} \alpha_j a_i(u_j, v_i) = f_{\lambda_i} = \lambda_i(u^*), \quad 1 \le i \le n,$$

of meshless Petrov–Galerkin schemes [1]. If trial functions u_j and test functions v_i coincide, and if there is only a single bilinear form involved, this reduces to the well-known Galerkin method. It has a strong built-in connection of test functionals to trial functions and yields a linear system with a positive semidefinite symmetric Gramian matrix. This makes it easy to handle, and it forms the core of the finite element method and some of its generalizations. However, for unsymmetric weak Petrov–Galerkin schemes, the theoretical and numerical analysis is rather hard [15].

To make problems in strong formulation symmetric, the connection between test functionals and trial functions must be established differently. To get a truly meshless technique and to allow very general problems, the connection is made via a symmetric positive definite *kernel* $\Phi : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ if the underlying problem is posed for functions of d variables on a domain $\Omega \subset \mathbb{R}^d$. It takes the discretized set Λ_m of test functionals and defines the trial functions in (1.2) as $u_j := \lambda_j^y \Phi(\cdot, y)$ for $1 \le j \le m$, where the superscript of λ indicates the variable of Φ on which the functional operates. Then the collocation matrix (1.3) is symmetric with entries $\lambda_i(u_j) = \lambda_i^x \lambda_j^y \Phi(x, y)$ for $1 \le i, j \le n$. This *symmetric collocation* technique dates back to [17] and has a solid mathematical basis [3, 4]. Like in the standard (non-Petrov) Galerkin scheme, the trial and test functions or functionals are closely related in order to maintain symmetry.

But this paper will deal with *unsymmetric meshless collocation*. By following an early idea of Kansa [8, 9] used by many authors afterwards (see an overview in [5]), one takes a set $X_m := \{x_1, \ldots, x_m\} \subset \mathbb{R}^d$ of m scattered points and uses a kernel Φ to define the trial space $\mathcal{U}$ spanned by the trial functions

$$u_j = \Phi(\cdot, x_j) = \lambda_{\delta_{x_j}}^y \Phi(\cdot, y), \quad 1 \le j \le m, \tag{1.4}$$

associated to simple point evaluation functionals $\lambda_{\delta_{x_j}}$. Often, but not necessarily, the points x_j are irregularly placed within Ω. Since they determine the trial functions in

(1.4), we can call them *trial centers*. The resulting unsymmetric collocation matrix has the entries

$$\lambda_i(u_j) = \lambda_i^y \Phi(y, x_j), \quad 1 \le i \le n, \quad 1 \le j \le m,$$

and can be singular in exceptional cases [6]. Consequently, there are very limited mathematical results on this technique, though it gives very good results in plenty of applications in science and engineering.

To overcome these solvability problems, one has to modify the setting. To get solvability and error bounds, there should at least be a unique solution to the modified discretized system that converges to the true solution if the discretization is refined. The first question requires that, if $n \gg m$ test functionals are fixed and are linearly independent, the system has rank m, if the trial functions are chosen properly. This was proven in [11], while the earlier paper [13] contained an asymptotic analysis for sufficiently many test functionals and trial functions. By a rather complicated and abstract machinery, a fairly general theory leading to error bounds and convergence results for the unsymmetric meshless collocation method was finally provided in [14].

This paper is independent of the theory in [14] and will prove a much more elementary convergence result for a variation of meshless unsymmetric collocation. Furthermore, since [14] does not focus on numerical techniques, we show how certain algorithms allow one to satisfy the requirements needed for convergence, in the context of both this paper and [14]. In particular, we combine an adaptive greedy method with linear programming, and some numerical examples show that our new adaptive on-the-fly algorithm works stably in spite of large condition numbers.

2. Well-posed problems. We now go back to the infinite problem (1.1) and proceed to define a variation of the unsymmetric collocation technique that will have a convergence proof. To this end, we have to fix the mathematical background.

The problem (1.1) is posed in a Hilbert space $\mathcal{U}$ via a test set $\Lambda \subset \mathcal{U}^*$ of continuous linear functionals on $\mathcal{U}$. The functionals can clearly be normalized to satisfy $\|\lambda\|_{\mathcal{U}^*} = 1$ for all $\lambda \in \Lambda$. The specific solution to our problem is called u^* from now on. This way, we assume solvability within the Hilbert space $\mathcal{U}$, even if the space $\mathcal{U}$ is smaller than the "existence space," i.e., the space in which minimal assumptions guarantee the existence of a solution by mathematical analysis. For instance, a Poisson problem on a domain $\Omega \subset \mathbb{R}^d$ can be stated in Sobolev "existence" space $W_2^1(\Omega)$ in weak formulation, but we can also take a Hilbert space $\mathcal{U}$ of at least twice differentiable functions and pose the Poisson problem there in strong form, by using the data from a smooth function $u^* \in \mathcal{U}$. Note that standard regularity theory [10] allows this, if the data and the domain are smooth. The above general framework allows both settings, if applied in weak form, since the given data always are of the form $\lambda(u^*),\ \lambda \in \Lambda$, for a specific function $u^* \in \mathcal{U}$. But here we want to stick to problems posed in strong form, and we assume sufficient additional regularity.

Furthermore, we need the problem to be *well-posed*, and this is expressed by the requirement that

$$\|u\|_\Lambda := \sup_{\lambda \in \Lambda} |\lambda(u)| \tag{2.1}$$

is a norm on $\mathcal{U}$. This definition will always provide a seminorm without further assumptions, but we get a norm if we can prove unique solvability by additional analytic arguments like a maximum principle. Note that this norm will be weaker

than the norm in $\mathcal{U}$ because of

$$\|u\|_\Lambda \le \|u\|_{\mathcal{U}} \quad \text{for all} \quad u \in \mathcal{U} \tag{2.2}$$

following from

$$|\lambda(u)| \le \|\lambda\|_{\mathcal{U}^*}\|u\|_{\mathcal{U}} = \|u\|_{\mathcal{U}} \quad \text{for all} \quad u \in \mathcal{U},\ \lambda \in \Lambda,$$

and normalization of the test functionals. The well-posedness of the problem is often alternatively expressed in the form

$$\|u\|_A \le C_A\|u\|_\Lambda \le C_A\|u\|_{\mathcal{U}} \quad \text{for all} \quad u \in \mathcal{U}, \tag{2.3}$$

where $\mathcal{U}_A$ with norm $\|.\|_A$ is a data-independent function space like $\mathcal{U}$ but with a weaker norm. This also means that $\mathcal{U}$ is continuously embedded in $\mathcal{U}_A$. The upshot here is that the space $\mathcal{U}_A$ and its norm should not be dependent on the test set Λ anymore. We call (2.3) the basic *analytic a priori inequality* for well-posedness of the problem in the space $\mathcal{U}_A$. The condition (2.3) looks rather abstract, but it can be explained by the following.

Example 2.1. Let a Poisson problem in two dimensions be given in strong form as

$$\begin{aligned} \Delta u &= f \quad && \text{in } \Omega, \\ u &= \varphi \quad && \text{on } \partial\Omega, \end{aligned}$$

with f continuous in $\overline{\Omega}$ and φ continuous on $\partial\Omega$. Furthermore, let Ω be a bounded domain in $\mathbb{R}^2$. The domain should allow the standard maximum principle and admit the standard a priori inequality

$$\|u\|_{W^2(\Omega)} \le C_A\|\Delta u\|_{L_2(\Omega)}$$

for all u in $W_0^2(\Omega)$ (see [10, p. 66, eq. (6.8)]). Since we are in two dimensions, there is a Sobolev embedding inequality

$$\|u\|_{\infty,\Omega} \le C_S\|u\|_{W^2(\Omega)}$$

for all $u \in W^2(\Omega)$. We pose a strong problem by taking the test set Λ to consist of all functionals

$$\begin{aligned} u &\mapsto (\Delta u)(x), \quad && x \in \overline{\Omega}, \\ u &\mapsto u(y), && y \in \partial\Omega. \end{aligned}$$

We consider the space $\mathcal{U} = C^2(\overline{\Omega})$ of functions on $\overline{\Omega}$. For all functions $u \in \mathcal{U}$ we define

$$\begin{aligned} \|u\|_\Lambda &:= \max(\|\Delta u\|_{\infty,\Omega}, \|u\|_{\infty,\partial\Omega}), \\ \|u\|_{\mathcal{U}} &:= 2 \cdot \max_{|\alpha|\le 2} \|D^\alpha u\|_{\infty,\overline{\Omega}} \end{aligned}$$

to get $\|u\|_\Lambda \le \|u\|_{\mathcal{U}}$. Let there be a $W^2(\Omega)$-regular true solution u^*, and assume that we have an approximate solution $v \in \mathcal{U} = C^2(\overline{\Omega})$, with

$$\begin{aligned} \Delta v &\approx f \quad && \text{in } \Omega, \\ v &\approx \quad && \varphi \text{ on } \partial\Omega \end{aligned}$$

such that Δv is still continuous in $\overline{\Omega}$ and v is continuous on $\partial\Omega$. Let w be the solution of

$$\begin{aligned} \Delta w &= 0 && \text{in } \Omega, \\ w &= v - \varphi && \text{on } \partial\Omega. \end{aligned}$$

Then, by the maximum principle, we have

$$\|w\|_{\infty,\Omega} \le \|w\|_{\infty,\partial\Omega} = \|v - \varphi\|_{\infty,\partial\Omega},$$

and we get the error bound

$$\|u^* - v\|_{\infty,\Omega} \le \|u^* - v + w\|_{\infty,\Omega} + \|w\|_{\infty,\Omega} \le \|u^* - v + w\|_{\infty,\Omega} + \|v - \varphi\|_{\infty,\partial\Omega}.$$

The function $u^* - v + w$ has vanishing boundary values and thus lies in $W_0^2(\Omega)$. With the Sobolev embedding inequality we then get

$$\begin{aligned} \|u^* - v + w\|_{\infty,\Omega} &\le C_S \|u^* - v + w\|_{W^2(\Omega)} \\ &\le C_S C_A \|\Delta(u^* - v + w)\|_{L_2(\Omega)} \\ &= C_S C_A \|\Delta(u^* - v)\|_{L_2(\Omega)} \\ &= C_S C_A \|f - \Delta v\|_{L_2(\Omega)} \\ &\le C_S C_A \sqrt{vol(\Omega)} \|f - \Delta v\|_{\infty,\Omega} \end{aligned}$$

and finally

$$\begin{aligned} \|u^* - v\|_{\infty,\Omega} &\le \|v - \varphi\|_{\infty,\partial\Omega} + C_S C_A \sqrt{vol(\Omega)} \|f - \Delta v\|_{\infty,\Omega} \\ &\le \left(1 + C_S C_A \sqrt{vol(\Omega)}\right) \max(\|v - \varphi\|_{\infty,\partial\Omega}, \|f - \Delta v\|_{\infty,\Omega}) \\ &= \left(1 + C_S C_A \sqrt{vol(\Omega)}\right) \|v - u^*\|_\Lambda. \end{aligned}$$

Under the above assumptions, there is a constant C, independent of $u^* \in W^2(\Omega)$ and $v \in C^2(\Omega)$, such that the following holds:

$$\|v - u^*\|_{\infty,\Omega} \le C \|v - u^*\|_\Lambda.$$

This altogether proves (2.3) for $\mathcal{U}_A = C(\overline{\Omega})$ with norm $\|.\|_A = \|.\|_\infty$.

Note that our assumption of well-posedness replaces standard assumptions such as uniform ellipticity of the underlying differential operator. Thus we can handle much more general situations.

3. Approximation in trial spaces. Now let $\mathcal{U}_\epsilon$ be a subspace of $\mathcal{U}$ such that for all $u \in \mathcal{U}$ there is some approximation $v_{u,\epsilon} \in \mathcal{U}_\epsilon$, with

$$\|u - v_{u,\epsilon}\|_\Lambda \le \epsilon \|u\|_{\mathcal{U}} \quad \text{for all} \quad u \in \mathcal{U}, \tag{3.1}$$

with a small number $\epsilon > 0$. This is satisfied for subspaces that come from a sequence of subspaces getting dense in $\mathcal{U}$, e.g., for Kansa-type trial spaces with sufficiently dense trial centers. One will, however, need a little more regularity in $\mathcal{U}$ than the regularity to define $\|u\|_\Lambda$ to get such a bound. Let us explain how to derive (3.1) in special cases. Details and proof techniques are in the recent book [16].

Example 3.1. Take a smooth positive definite radial basis function ϕ, and define a translation-invariant kernel Φ on $\mathbb{R}^d$ via

$$\Phi(x - y) := \phi(\|x - y\|_2), \quad x, y \in \mathbb{R}^d.$$

Its smoothness is measured via Fourier transforms and a positive real number β satisfying

$$\hat{\Phi}(\omega) \leq C(1+\|\omega\|_2)^{-d-\beta} \quad \text{for all} \quad \omega \in \mathbb{R}^d.$$

Then there is a unique "native" Hilbert space $\mathcal{U}$ isomorphic to $W^{(d+\beta)/2}(\mathbb{R}^d)$ in which Φ is a reproducing kernel. Now take a large set X of scattered centers in a bounded domain $\Omega \subset \mathbb{R}^d$ such that the *fill distance* or *mesh norm*

$$h := \sup_{y\in\Omega} \min_{x\in X} \|x-y\|_2$$

is small. For each $u \in \mathcal{U}$ consider the standard interpolant $s_{u,X}$ based on translates $\Phi(\cdot - x_j)$ with respect to the data in X. Then there is a pointwise error bound

$$|(D^\alpha u)(x) - (D^\alpha s_{u,X})(x)| \leq C_\alpha h^{\beta/2-|\alpha|}\|u\|_{\mathcal{U}} \quad \text{for all} \quad u \in \mathcal{U},\ x \in \Omega,\ 0 \leq |\alpha| < \beta/2,$$

where C is a constant depending only on $\mathcal{U}$ and the domain but not on X, x, h, α, and u. We apply this for all functionals $\lambda \in \Lambda$ if they are in strong form, i.e., evaluations of operators acting on u and evaluated at a point, like $\mu(u) := D^\alpha_x u := (D^\alpha u)(x)$. Then it is clear that (3.1) holds, provided that regularity of the solution and smoothness of the kernel are good enough to guarantee $\beta/2 - |\alpha| > 0$ for all occurring derivative orders $|\alpha|$ in the strong functionals of the test set Λ.

Again, our assumptions are much more general than in standard theories, because we just require a reasonably good approximation of the true solution by the trial space, no matter how it is defined.

4. Convergence. Now we describe the basic convergence argument, but we start in a rather abstract setting and do not even specify whether we apply weak or strong testing or how we do the numerical calculations. By using the subspace $\mathcal{U}_\epsilon \subset \mathcal{U}$ provided by the previous section, we would like to construct a function $v_\epsilon \in \mathcal{U}_\epsilon$ that solves

$$v_\epsilon = \arg\min_{v\in\mathcal{U}_\epsilon} \|v-u^*\|_\Lambda = \arg\min_{v\in\mathcal{U}_\epsilon} \sup_{\lambda\in\Lambda} |\lambda(v)-\lambda(u^*)| \tag{4.1}$$

over all $v \in \mathcal{U}_\epsilon$. We do not know whether the minimum is attained or how to calculate it, but we know that there is a function $v^*_\epsilon \in \mathcal{U}_\epsilon$, with

$$\|v^*_\epsilon - u^*\|_\Lambda \leq 2\|v_{u^*,\epsilon} - u^*\|_\Lambda \leq 2\epsilon\|u^*\|_{\mathcal{U}}, \tag{4.2}$$

which is sufficient for our purpose. Though almost trivial, we state the following.

THEOREM 4.1. *Let $\mathcal{U}$ be a normed linear space with norm $\|.\|_{\mathcal{U}}$, dual space $\mathcal{U}^*$, and dual unit sphere $\mathcal{U}^*_1 := \{\lambda \in \mathcal{U}^* : \|\lambda\|_{\mathcal{U}^*} = 1\}$. Let a test set $\Lambda \subset \mathcal{U}^*_1$ be given such that $\|.\|_\Lambda$ is defined on $\mathcal{U}$ with* (2.1) *and* (2.2)*. Assume further that the general interpolation problem* (1.1) *is well-posed, i.e., that $\|.\|_\Lambda$ is a norm. Let $\{\mathcal{U}_\epsilon\}_\epsilon$ be a scale of subspaces of $\mathcal{U}$ for $\epsilon \to 0$ such that for all $u \in \mathcal{U}$ there is a $v_{u,\epsilon} \in \mathcal{U}_\epsilon$ with* (3.1)*. For all $\epsilon \to 0$, take a function v^*_ϵ satisfying* (4.2)*. Then the error bound* (4.2) *holds, and there is convergence $\|v^*_\epsilon - u^*\|_\Lambda \to 0$. If well-posedness holds via* (2.3) *for a space $\mathcal{U}_A$, then the convergence and the error bound can both be rewritten in terms of $\|.\|_A$. The same results hold for solutions v_ϵ of* (4.1)*.*

Note that convergence rates of the approximation in $\mathcal{U}_\epsilon$ spaces carry over to convergence rates of the approximate solutions v^*_ϵ, precisely as in the finite element method (FEM) case. In other words, the rate of convergence of the approximate solutions is exactly the one of the best approximation in $\mathcal{U}_\epsilon$ with respect to the norm $\|.\|_\Lambda$ to functions in $\mathcal{U}$. Details can be worked out along the lines of the above examples.

5. Abstract greedy adaptive algorithm. Now we want to set Theorem 4.1 to work. For ϵ fixed, we want to show how (4.2) can be solved. The space $\mathcal{U}_\epsilon$ will be a standard finite-dimensional Kansa trial space along the lines of section 3. Following the previous section, we have to solve the semi-infinite linear optimization problem P_Λ defined by

$$\begin{aligned} &\textbf{Minimize} \quad \eta \text{ such that (s.t.)} \\ &\lambda(v) - \lambda(u^*) \le \eta \qquad \text{for all } \lambda \in \Lambda, \\ &-\lambda(v) + \lambda(u^*) \le \eta \qquad \text{for all } \lambda \in \Lambda, \\ &v \in \mathcal{U}_\epsilon, \\ &\eta \in \mathbb{R} \end{aligned} \tag{5.1}$$

to obtain the solution v_ϵ in (4.1) or at least a suboptimal function v_ϵ^* satisfying (4.2). But we want to *discretize* the test side of this optimization problem, replacing Λ above by a finite subset $\Lambda_n := \{\lambda_1, \dots, \lambda_n\}$ to get a standard finite-dimensional linear optimization problem P_{Λ_n}.

To this end, we define the following abstract *greedy adaptive* algorithm:

- For $n := 0$ define $v_0 := 0$ and $\Lambda_0 := \emptyset$.
- 1. If $\|v_n - u^*\|_\Lambda = 0$ holds, stop.
  2. Else take $\lambda_{n+1} \in \Lambda$, with

$$\frac{2}{3}\|v_n - u^*\|_\Lambda \le |\lambda_{n+1}(v_n - u^*)| \le \|v_n - u^*\|_\Lambda, \tag{5.2}$$

  and set $\Lambda_{n+1} = \Lambda_n \cup \{\lambda_{n+1}\}$.
  3. Solve the finite-dimensional linear optimization problem $P_{\Lambda_{n+1}}$ by a function $v_{n+1} \in \mathcal{U}_\epsilon$.
  4. Set $n \Leftarrow n + 1$, and repeat.

We now show how close this algorithm comes to solving (4.1) or (5.1) with the optimal value

$$\eta^* = \min_{v \in \mathcal{U}_\epsilon} \|v - u^*\|_\Lambda = \min_{v \in \mathcal{U}_\epsilon} \sup_{\lambda \in \Lambda} |\lambda(v) - \lambda(u^*)|.$$

THEOREM 5.1. *If the algorithm is finite, then $u^* \in \mathcal{U}_\epsilon$ holds, and the true solution of the problem is reached in a finite number of steps. Otherwise,*

$$\eta^* \le \|v_n - u^*\|_\Lambda \le 2\|v_n - u^*\|_{\Lambda_n} =: 2\eta_n \le 2\eta^*$$

holds for infinitely many n. After a finite number of steps, the sufficient criterion

$$\|v_n - u^*\|_\Lambda \le 2\eta_n \tag{5.3}$$

for v_n being a solution of (4.2) *is satisfied.*

Proof. The first statement is trivial. If the algorithm generates an infinite sequence with $\|v_n - u^*\|_\Lambda > 0$, we renormalize and get convergence of a subsequence to

$$\frac{v_n - u^*}{\|v_n - u^*\|_\Lambda} \to w \in \mathcal{U}_\epsilon,$$

with $\|w\|_\Lambda = 1$, because $\mathcal{U}_\epsilon$ and u^* span a finite-dimensional space. By our selection rule,

$$\|v_n - u^*\|_\Lambda \ge \|v_n - u^*\|_{\Lambda_{n+1}} \ge |\lambda_{n+1}(v_n - u^*)| \ge \frac{2}{3}\|v_n - u^*\|_\Lambda$$

and

$$\begin{aligned}\|w\|_{\Lambda_{n+1}} &\geq \left\| \frac{v_n - u^*}{\|v_n - u^*\|_\Lambda} \right\|_{\Lambda_{n+1}} - \left\| w - \frac{v_n - u^*}{\|v_n - u^*\|_\Lambda} \right\|_{\Lambda_{n+1}} \\ &\geq \frac{2}{3} - \left\| w - \frac{v_n - u^*}{\|v_n - u^*\|_\Lambda} \right\|_\Lambda \\ &\geq \frac{1}{2}\end{aligned}$$

for sufficiently large n from our subsequence. Similarly,

$$\begin{aligned}\frac{\eta_n}{\|v_n - u^*\|_\Lambda} &= \frac{\|v_n - u^*\|_{\Lambda_n}}{\|v_n - u^*\|_\Lambda} \\ &\geq \|w\|_{\Lambda_n} - \left\| w - \frac{v_n - u^*}{\|v_n - u^*\|_\Lambda} \right\|_{\Lambda_n} \\ &\geq \|w\|_{\Lambda_n} - \left\| w - \frac{v_n - u^*}{\|v_n - u^*\|_\Lambda} \right\|_\Lambda\end{aligned}$$

implies that we have

$$\frac{\eta_n}{\|v_n - u^*\|_\Lambda} \geq \frac{1}{2}$$

and

$$\frac{1}{2}\|v_n - u^*\|_\Lambda \leq \eta_n \leq \|v_n - u^*\|_\Lambda$$

for infinitely many sufficiently large n from our subsequence. Since

$$0 \leq \eta_1 \leq \cdots \leq \eta_n \leq \lim_{j\to\infty} \eta_j =: \tilde{\eta} \leq \eta^* := \inf_{v\in\mathcal{U}_\epsilon} \|v - u^*\|_\Lambda < \infty,$$

we can conclude that

$$\eta^* \leq \|v_n - u^*\|_\Lambda \leq 2\eta_n \leq 2\tilde{\eta} \leq 2\eta^*$$

holds for sufficiently large n of a subsequence. If (5.3) holds, the above inequality is satisfied. Then we have (4.2) due to

$$\|v_n - u^*\|_\Lambda \leq 2\eta^* \leq 2\|v_{u^*,\epsilon} - u^*\|_\Lambda \leq 2\epsilon\|u^*\|_{\mathcal{U}}. \qquad \square$$

Note that the criterion (5.3) is rather close to being numerically available.

Theorem 5.1 shows that the selection of useful finite subsets Λ_n can be done by an adaptive greedy method, by taking sequences of finite linear programs, checking their solutions on the remaining test functionals, and taking in those functionals where large errors occur. The next sections will address the implementation of the above abstract algorithm.

6. Linear optimization. With the theoretical background of the previous section, we now proceed towards the implementation of a greedy adaptive on-the-fly variant of a linear optimization algorithm, but we have to postpone adaptivity to the next section. Assume that we have a selection $\Lambda_n := \{\lambda_1, \ldots, \lambda_n\}$ of test functionals

and a selection $\mathcal{U}_m := \text{ span } \{u_1, \ldots, u_m\}$ of a space of trial functions such that the linear optimization problem, i.e., the discretized version of (5.1), reads as

$$\begin{aligned} &\textbf{Minimize} \quad \eta \text{ s.t.} \\ \sum_{k=1}^{m} \alpha_k \lambda_j(u_k) - \lambda_j(u^*) &\le \eta, \quad 1 \le j \le n, \\ -\sum_{k=1}^{m} \alpha_k \lambda_j(u_k) + \lambda_j(u^*) &\le \eta, \quad 1 \le j \le n, \\ \alpha_k &\in \mathbb{R}, \quad 1 \le k \le m, \\ \eta &\in \mathbb{R}. \end{aligned}$$

This is a linear discrete Chebyshev approximation problem of the form

$$\begin{aligned} &\textbf{Minimize} \quad \eta \text{ s.t.} \\ Ax - b &\le \eta \cdot \mathbf{1}, \\ -Ax + b &\le \eta \cdot \mathbf{1} \end{aligned}$$

written out componentwise with a matrix $A \in \mathbb{R}^{n \times m}$, $m \le n$, and vectors $b \in \mathbb{R}^n$ and $\mathbf{1} \in \mathbb{R}^n$. We explicitly reformulate the standard reduction to a dual problem in canonical form [2] because we want to add adaptivity later. First, we rewrite the constraints as

$$\begin{pmatrix} A & -\mathbf{1} \\ -A & -\mathbf{1} \end{pmatrix} \cdot \begin{pmatrix} x \\ \eta \end{pmatrix} \le \begin{pmatrix} b \\ -b \end{pmatrix}$$

and the objective function as

$$-\eta = (\mathbf{0}, -1) \cdot \begin{pmatrix} x \\ \eta \end{pmatrix}$$

to be *maximized*. The standard dual of such a problem is

$$\begin{aligned} &\textbf{Minimize } (u^T, v^T) \cdot \begin{pmatrix} b \\ -b \end{pmatrix} \text{ s.t.} \\ &\begin{pmatrix} A^T & -A^T \\ -\mathbf{1}^T & -\mathbf{1}^T \end{pmatrix} \cdot \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ -1 \end{pmatrix}, \\ &u, v \ge 0 \end{aligned}$$

or, equivalently,

$$\begin{aligned} &\textbf{Minimize } b^T(u - v) \text{ s.t.} \\ A^T(u - v) &= 0, \\ \mathbf{1}^T(u + v) &= 1, \\ u, v &\ge 0. \end{aligned} \tag{6.1}$$

Since the original problem is solvable, the weak and strong duality theorems hold. In addition, the complementary slackness conditions at the optimal primal and dual solutions, denoted by x^*, η^*, u^*, v^*, will be satisfied. They are

$$\begin{aligned} u_j^* \cdot (Ax^* - b - \eta^* \cdot \mathbf{1})_j &= 0, \\ v_j^* \cdot (-Ax^* + b - \eta^* \cdot \mathbf{1})_j &= 0, \end{aligned}$$

while the factors are nonnegative. The upshot is that there are *active* indices j which can be grouped into two sets

$$\begin{array}{lllll} j \in J^+ & \text{if} & u_j^* > 0 & \text{and} & (Ax^* - b)_j = \eta, \\ j \in J^- & \text{if} & v_j^* > 0 & \text{and} & (Ax^* - b)_j = -\eta. \end{array}$$

If $\eta > 0$, these index sets are disjoint. This implies that the index sets are related to the signs of the nonzero components of w^*, if there are no degenerations.

We work with the revised simplex method based on (6.1). This means that (except for the startup phase, which is postponed here) we always keep the inverse C of a $(m+2) \times (m+2)$ matrix

$$C^{-1} = B := \begin{pmatrix} A_s^T & 0 \\ \mathbf{1}^T & 0 \\ b_s^T & -1 \end{pmatrix},$$

with a sign vector $s \in \{-1, +1\}^{m+1}$ such that B contains a nonsingular selection of $m+1$ columns of

$$\mathcal{B} := \begin{pmatrix} A^T & -A^T & 0 \\ \mathbf{1}^T & \mathbf{1}^T & 1 \\ b^T & -b^T & -1 \end{pmatrix}$$

with the appropriate signs. In addition, one must store the numbers of the selected columns of $\mathcal{B}$ or, equivalently, the numbers of the selected test functionals. We call the selection of $m+1$ signs and functionals occurring in B the *dual basis*. The current dual vertex is a vector $w_s \in \mathbb{R}^{m+1}$ such that

$$\begin{array}{rcl} A_s^T w_s & = & 0, \\ \mathbf{1}^T w_s & = & 1. \end{array} \tag{6.2}$$

This means that

$$\begin{pmatrix} A_s^T & 0 \\ \mathbf{1}^T & 0 \\ b_s^T & -1 \end{pmatrix} \cdot \begin{pmatrix} w_s \\ b_s^T w_s \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \text{ and } \begin{pmatrix} w_s \\ b_s^T w_s \end{pmatrix} = C \cdot \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$$

hold. Thus the dual vertex and the value of the dual objective function can be read off the penultimate column of the matrix C. The corresponding primal vertex is described by

$$\begin{pmatrix} A_s & \mathbf{1} & b_s \\ 0 & 0 & -1 \end{pmatrix} \cdot \begin{pmatrix} x_s \\ -\eta \\ -1 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \text{ or } \begin{pmatrix} x_s \\ -\eta \\ -1 \end{pmatrix} = C^T \cdot \begin{pmatrix} 0 \\ 1 \end{pmatrix} \tag{6.3}$$

and can be read off the last column of C^T which is the last row of C.

The revised simplex method based on (6.1) already is an on-the-fly technique, provided that new columns of the constraint matrix are successively added, while the rows are fixed [2]. For the original problem, this means that new test functionals are introduced, while the trial space is fixed. So there are no problems when adding new test functionals to a solver running as a revised simplex method for the dual problem (6.1). We shall explain this in detail in the following section. The actual calculation matrix C (i.e., the inverse of a square part B of the full data matrix $\mathcal{B}$) will be of constant size for a fixed trial space.

For completeness and later use, let us describe the on-the-fly recipe. In theory, the standard full simplex tableau is the matrix $C \cdot \mathcal{B}$ which is of the form

$$\begin{pmatrix} S_+ & S_- & w_s \\ z_+^T & z_-^T & b_s^T w_s \end{pmatrix}$$

due to the factorization

$$\begin{pmatrix} A_s^T & 0 \\ \mathbf{1}^T & 0 \\ b_s^T & -1 \end{pmatrix} \cdot \begin{pmatrix} S_+ & S_- & w_s \\ z_+^T & z_-^T & b_s^T w_s \end{pmatrix} = \begin{pmatrix} A^T & -A^T & 0 \\ \mathbf{1}^T & \mathbf{1}^T & 1 \\ b^T & -b^T & 0 \end{pmatrix}. \tag{6.4}$$

Note that this tableau is never stored. The crucial row for the simplex method is the final one, containing the vectors

$$z_+^T = b_s S_+ - b^T, \quad z_-^T = b_s S_- + b^T.$$

It is representable as $e_{n+2}^T C\mathcal{B}$, but we know that $e_{n+2}^T C$ contains the primal data $(x_s^T, -\eta, -1)$. This means that

$$(z_+^T, z_-^T, b_s^T w_s) = (x_s^T, -\eta, -1) \cdot \begin{pmatrix} A^T & -A^T & 0 \\ \mathbf{1}^T & \mathbf{1}^T & 1 \\ b^T & -b^T & 0 \end{pmatrix}$$

holds; i.e.,

$$\begin{aligned} Ax_s - b - \eta\mathbf{1} &= z_+, \\ -Ax_s + b - \eta\mathbf{1} &= z_-, \\ -\eta &= b_s^T w_s. \end{aligned} \tag{6.5}$$

The tableau is optimal iff all components of the z vectors are nonpositive. This means $\|Ax_s - b\|_\infty \le \eta$, as expected.

7. Adaptive linear optimization. We now know that we have to look for a test functional $\lambda \in \Lambda$ with a large positive value of

$$|\lambda(U_m)^T \cdot x_s - f_\lambda| - \eta,$$

with $(\lambda(u_1), \dots, \lambda(u_m)) =: \lambda(U_m)^T$. This can be done without storing any matrix data, and the search for a good test functional can be made on an extremely large set of test functionals. Going for the largest possible value can be called a "greedy" strategy. This is the common choice in linear optimization, and it has been used favorably [11] also in meshless adaptive PDE solvers based on unsymmetric collocation. But note that (5.2) is satisfied only if we make sure that our new functional λ has the property

$$|\lambda(U_m)^T \cdot x_s - f_\lambda| \ge \frac{2}{3} \sup_{\mu \in \Lambda} |\mu(U_m)^T \cdot x_s - f_\mu|;$$

i.e., it suffices to find a reasonably good suboptimal functional to make our error bounds and convergence results valid.

Once we have such a functional, we generate the corresponding column of $\mathcal{B}$, multiply it by C, and get the corresponding column of the full tableau. We have to

make sure that the coresponding entry of z_- or z_+ is positive. In view of (6.5) we go for the left-hand part of $\mathcal{B}$, if

$$\lambda(U_m)^T \cdot x_s - f_\lambda - \eta$$

is positive. This sign information must be stored together with the identification or index for λ. Up to this point, we have found the column of the pivot for the next step of the simplex algorithm. The pivot row is now generated in the standard way, by comparing entries of the pivot column to the final column of the tableau, which is the penultimate column of C. Such a pivot row must exist, because otherwise our discrete linear optimization would be unsolvable. This determines a Gauss–Jordan transformation, which is carried out on C and *not* on the tableau. In fact, since the standard Gauss–Jordan transformation can be written as multiplication of the tableau from the left by an $(m+2) \times (m+2)$ matrix J, we can use (6.4) to see that we need only to multiply C^{-1} with J^{-1} from the right. This means that we can update C by forming $J \cdot C$. This finishes the step of the revised simplex method.

The above algorithm runs on a fixed trial space and can handle extremely large amounts of test functionals. It does not need to arrive at a full minimum, in particular if the current value of $\eta = -b_s^T w_s$ is too large to be tolerated for the given problem. Note that for any intermediate situation we have the inequality

$$\eta \le \inf_{u \in U_m} \sup_{\lambda \in \Lambda} |\lambda(u) - f_\lambda|. \tag{7.1}$$

This follows from the weak duality theorem in linear optimization. In fact, if u is any function from the trial space with coefficient vector $x \in \mathbb{R}^m$ in the basis $u_1, \ldots, u_m$, we have $(A_s x - b_s)^T w_s = -b_s^T w_s = \eta$ and find

$$\eta \le \|w_s\|_1 \|A_s x - b_s\|_\infty = \|A_s x - b_s\|_\infty \le \sup_{\lambda \in \Lambda} |\lambda(u) - f_\lambda|$$

because $A_s x - b_s$ is a signed collection of values $\lambda(u) - f_\lambda$ with those functionals λ which constitute the actual rows of A_s.

In (7.1) one would finally get equality if the revised simplex algorithm is carried out indefinitely with $\mathcal{U}_m$ fixed. But if η is at some stage already rather large to be tolerated, it does not make sense to work all of the way towards a full optimization with respect to all test functionals, because η will increase all of the time. It is more advisable to enlarge the trial space to make smaller values of η possible at all. This is to be investigated in the next section.

8. Adaptivity of trial spaces. We focus now on a situation where there are no new test functionals to be introduced while the error level η is still too large. This calls for an enlargement of the trial space. Adding a new trial function means adding a column to A, and at this point we do not assume anything about that choice. One can use radial basis functions of any shape or scale, finite elements, or singular functions to cope with boundary singularities. However, adding a column to A means adding a row to A^T, and consequently the actual vertex solution of the dual problem (6.1) cannot be used any more. A complete restart should be avoided. We thus look for an on-the-fly technique that makes use of the current optimal vertex of (6.1) while adding a new row to A^T and which continues effectively.

We first comment on the selection of a new row a^T of A^T. The current dual vector w_s has the property $A_s w_s = 0$, and a^T would be useless if $a^T w_s = 0$. This quantity

has the form

$$a^T w_s = \sum_{j=1}^{n} w_j s_j \lambda_j(u)$$

if a comes out of a new trial function u, where λ_j and s_j are the functionals and signs of the current dual basis. In this way, one can ask for an "optimal" new trial function u that maximizes this quantity under some normalization of u, e.g., $\|u\|_\Lambda := \sup_{\lambda\in\Lambda} |\lambda(u)| = 1$. In the following, we assume only $a^T w_s \neq 0$ and denote our new trial function by u_{m+1}.

At this point, one could restart the whole process, by using the above test functionals first before introducing others. This would mean a new startup of the whole algorithm.

To avoid a restart, we have to produce a new enlarged version of the square matrices B and C. But then we also need an additional column. This means that we have to add another test functional, too. But since we already have a good mechanism for finding good test functionals, we add the zero functional as a dummy that will be eliminated in the next simplex step. The enlarged B matrix will be

$$B' := \begin{pmatrix} A_s^T & 0 & 0 \\ a^T & 0 & 0 \\ \mathbf{1}^T & 1 & 0 \\ b_s^T & 0 & -1 \end{pmatrix},$$

which is nonsingular because $a^T w_s \neq 0$, and $A_s^T w_s = 0$ ensures that a^T is not a linear combination of the rows of A_s^T. The inverse C of B has the form

$$\begin{pmatrix} C_s & w_s & 0 \\ b_s^T C_s & b_s^T w_s & -1 \end{pmatrix},$$

with

$$\begin{aligned} A_s C_s &= I, \\ \mathbf{1}^T C_s &= 0^T. \end{aligned}$$

Then the inverse of B' is

$$\begin{pmatrix} C_s' & \alpha \cdot w_s & 0 & 0 \\ z^T & -\alpha & 1 & 0 \\ b_s^T C_s' & \alpha \cdot b_s^T w_s & 0 & -1 \end{pmatrix},$$

with

$$\begin{aligned} \alpha &= \frac{1}{a^T w_s}, \\ z^T &= -\alpha a^T C_s, \\ C_s' &= C_s + w_s z^T. \end{aligned}$$

This is an easy $\mathcal{O}(m^2)$ recipe to upgrade C.

Inspection of the new matrix C' via (6.3) shows that the new value of η will be zero and that the old primal coefficients are the same, the new coefficient being $\alpha b_s^T w_s$. If u and u' are the old and new primal solutions, respectively, written as trial functions, then

$$\lambda(u') = \lambda(u) + \alpha b_s^T w_s \lambda(u_{m+1})$$

can be used as a simplified technique for evaluating new test functionals on the new solution, if the $\lambda(u)$ were stored.

By looking at the system (6.2) and using Cramer's rule, we see that

$$|\det B'| = \left|\det \begin{pmatrix} A_s^T \\ a^T \end{pmatrix}\right| = \left|\det \begin{pmatrix} A_s^T \\ \mathbf{1}^T \end{pmatrix}\right| \cdot |a^T w_s| = |\det B| \cdot |a^T w_s|.$$

This gives a hint to stabilize the full process. One can keep track of the determinant of B and select a (or, equivalently, the new trial function) such that $|\det B'|$ is as close to 1 as possible. This strategy was quite successful in the adaptive greedy method of [11].

We also suggest the method of [11] for startup. It will make a selection of n test functionals $\lambda_1, \ldots, \lambda_n$ and n trial functions $u_1, \ldots, u_n$ to provide exact generalized interpolation for a function u from the trial space. The inverse $\tilde{C}$ of the $n \times n$ matrix $\tilde{A} = (\lambda_j(u_k))$ will be available when the algorithm of [11] stops.

We then find another functional λ_{n+1} such that $\lambda_{n+1}(u) \neq f_{\lambda_{n+1}}$ and define the $(n+1) \times n$ matrix $A = (\lambda_j(u_k))$ of the standard form of this paper. Then $A^T = (\tilde{A}^T, z)$ is known, where z is the vector with components $\lambda_{n+1}(u_j)$. With $v := \tilde{C}^T z$ we have $\tilde{A}^T v - z = 0$, and $w := (v^T, -1)^T$ is in the nullspace of A^T. From this vector, we derive a sign vector s by taking the signs of its components, and we renormalize the vector to generate a vector w_s of nonnegative components such that $\|w_s\|_1 = 1$ and $A_s^T w_s = 0$. This means that the signs of the components of w now occur as signs associated to columns of A_s^T and that we can write $A_s^T = (\tilde{A}^T, z) \cdot D_s$ with a diagonal matrix D_s carrying the signs in its diagonal. This way, we get the system (6.2). By using the sign vector, we can form b_s and have the matrix B that we need.

To get $C = B^{-1}$ cheaply, it suffices to get the inverse of

$$\begin{pmatrix} A_s^T \\ \mathbf{1}^T \end{pmatrix} = \begin{pmatrix} (\tilde{A}^T, z) \\ s^T \end{pmatrix} \cdot D_s$$

cheaply. This can be done at $\mathcal{O}(n^2)$ cost from the inverse $\tilde{C}$ of the $n \times n$ matrix $\tilde{A}$ in the standard rank-one update way.

9. Numerical demonstrations. In this section, we show some numerical examples that demonstrate the efficiency of our proposed algorithm. In all presented examples, we have used the multiquadric kernel

$$\Phi_c(x, y) = \sqrt{1 + \frac{||x - y||^2}{c^2}},$$

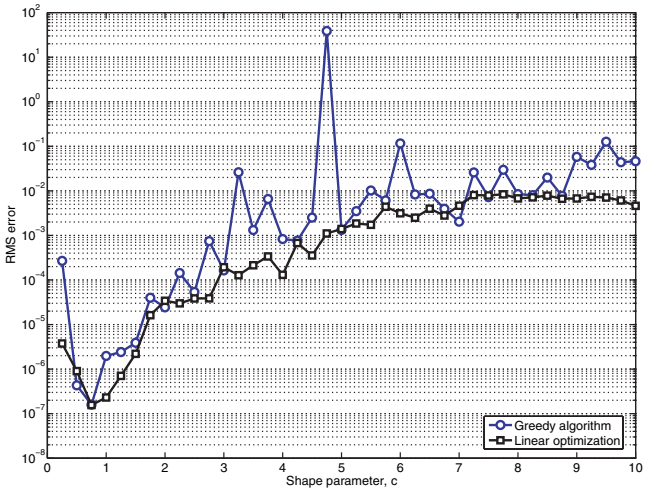
where $x, y \in \mathbb{R}^2$ and $c > 0$ is the *scaling parameter*. As a test equation, we solve the Poisson problem with Dirichlet boundary conditions on $\Omega = [-1, 1]^2$, i.e.,

$$\begin{aligned} \triangle u(x) &= f(x) \quad \text{for } x \in \Omega \subset \mathbb{R}^d, \\ u(x) &= \varphi(x) \quad \text{for } x \in \partial\Omega. \end{aligned} \tag{9.1}$$

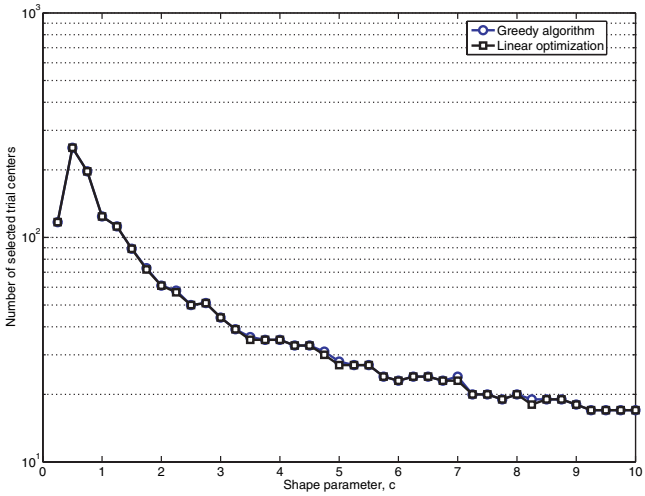
The functions f and φ in (9.1) are generated by the exact solution

$$u^*(x, y) = \frac{1}{2} \log\left((x-2)^2 + (x-2)^2\right).$$

First, the test problem is run on $\Omega = [-1, 1]^2$ with different shape parameters c in order to demonstrate the performance of the adaptive greedy scheme in [11] and the proposed adaptive greedy linear optimization scheme of this paper. The number of trial functions and test functionals offered to the algorithms for selection are $M = 8822$ and $N = 2129$, respectively, but note that the algorithms select much smaller subsets adaptively. In Figure 9.1(a), the root mean square (RMS) errors of both schemes are given. Most of the tested c values are unusual for the original radial basis functions (RBF) collocation method, since in the



(a) *Error profile.*



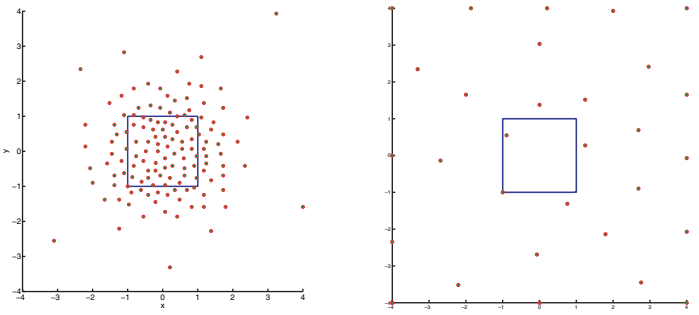
(b) *Degrees of freedom.*

FIG. 9.1. *Solving a Poisson equation with the multiquadric (MQ) kernel on a square domain: Error profile and degrees of freedom as a function of shape parameter c.*

standard setup the shape parameters are usually chosen to have roughly the same order as the mesh norm of the test functionals in order to circumvent the problem of ill-conditioning [7].

It turns out that the new linear optimization scheme is more stable in comparison to the adaptive greedy scheme. In particular, when $c = 4.75$, the adaptive greedy scheme fails to provide an acceptable solution, whereas the new linear optimization scheme performs reasonably well. We observe, from Figure 9.1(a), that there is an optimal shape parameter value of $c = 0.75$ for this particular example,

Furthermore, we show the number m of selected trial centers in Figure 9.1(b). Recall that the new adaptive linear optimization scheme uses the selected m trial centers and finally performs an ℓ^∞ fit to *all* $N \gg m$ available test functionals. In



(a) $c = 1$. (b) $c = 5$.

FIG. 9.2. *Solving a Poisson equation with the MQ kernel on a square domain: Selected trial centers for different shape parameters c.*

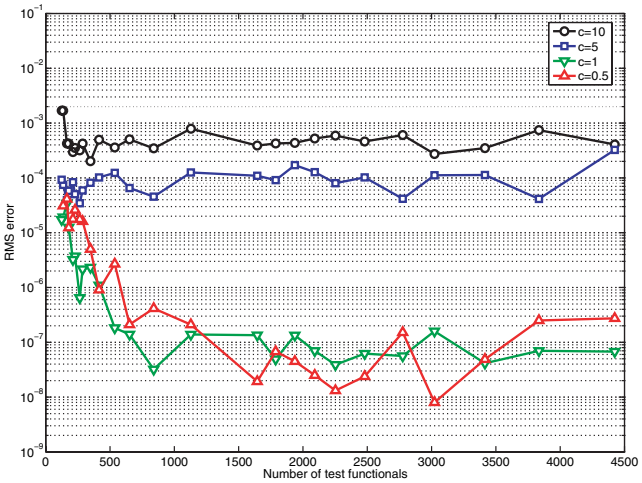


FIG. 9.3. *Solving a Poisson equation with the MQ kernel on an L-shaped domain: Error profile as a function of numbers of test functionals.*

contrast, the original adaptive greedy scheme of [11] calculates an exact interpolation by using the selected m trial functions and $m = n \ll N$ selected test functionals.

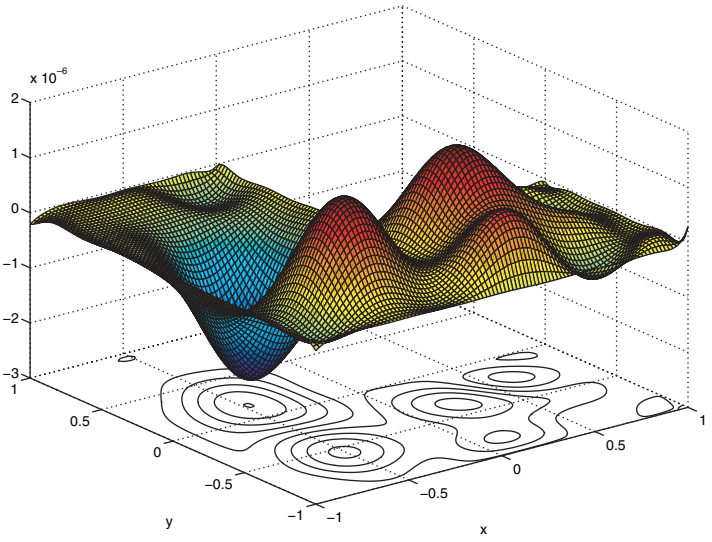
When the shape parameter c increases, the number of trial centers decreases. Moreover, the distributions of the selected trial centers change. As shown in Figure 9.2, in which we show the trial center distributions of two cases when $c = 1$ and $c = 5$, the adaptive algorithm tends to select trial centers away from the domain when the shape parameter is large.

Our next test demonstrates the effect of increasing the number of offered test functionals. This test is run on an L-shaped domain

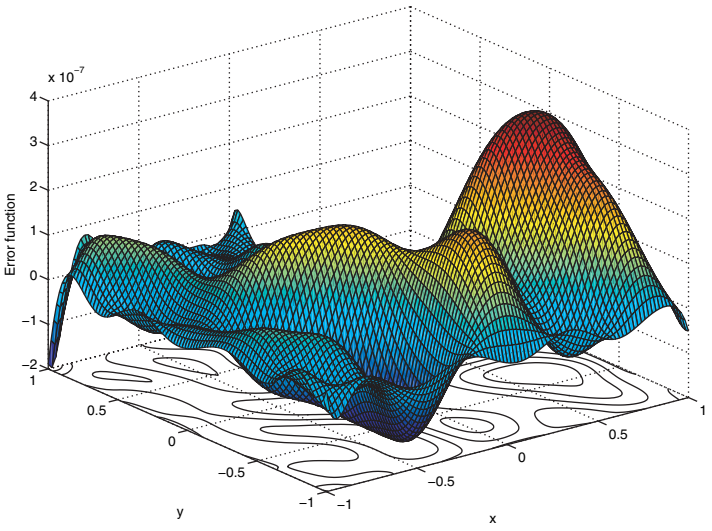
$$\Omega = \{(x, y) : \ -1 \le x, y \le 1, \, \mathrm{sign}(x) + \mathrm{sign}(y) \le 0\}.$$

For all runs, the trial space is adaptively selected from a fixed set of $M = 34692$ trial functions. Four shape parameters $c = 0.5$, $c = 1$, $c = 5$, and $c = 10$ are tested.

Figure 9.3 shows the corresponding RMS errors versus different numbers of test functionals. In all cases, large numbers of test functionals have no significant influence



(a) *Adaptive greedy.*



(b) *Linear optimization.*

FIG. 9.4. *Solving a Poisson equation with the MQ kernel on an L-shaped domain: Error functions.*

on the error. This is in line with our theory, since the error is dominated by the approximation properties of the trial space alone. Enlarging the set of test functionals will stabilize the algorithm. In fact, the error profiles for all tested c values remain stable even though the nonsquare background matrix increases in size, which usually causes ill-conditioning in the original formulation.

Last, Figures 9.4(a) and 9.4(b) show typical error functions of the adaptive greedy method and the proposed linear optimization method, respectively. Since the adaptive

greedy method solves the selected small square submatrix system in the sense of exact interpolation, its error function shows clear peaks. In contrast, the linear optimization method handles the overdetermined system with the selected trial functions and *all* test functionals in such a way that the maximal error in the differential equation and the boundary values is minimized on very many points. Thus the error profile usually is smaller in magnitude and more oscillatory with less obvious peaks.

10. Conclusion. On the theoretical side, we prove that the unsymmetric meshless collocation method, if carried out with smooth kernel-based trial functions and sufficiently many test functionals, converges at the same rate as interpolation of the solution. In particular, trial spaces formed by multiquadrics and Gaussian basis would result in exponential convergence [12] if the solution is analytic. However, from the computational point of view, exponential convergence is difficult to observe under double precision computation. Our proof techniques are much simpler than those of [14] because we focus on minimizing the maximum of the residuals via semi-infinite linear optimization.

On the practical side, we couple the previously proposed greedy technique of [11] with linear optimization to provide a fully adaptive on-the-fly data-dependent method guaranteeing solvability and fitting seamlessly into our theory. Numerical results show that the proposed coupled method is very stable, even though serious ill conditioning of the resultant systems must be expected.

In terms of computational cost, the saving of meshing the domain has shifted to solving the linear optimization problem. The complexity of the linear optimization may be by no means trivial. Interested readers are encouraged to employ the greedy algorithm with least-squares optimization despite the lack of theoretical results.

REFERENCES

[1] S. N. Atluri and S. P. Shen, *The Meshless Local Petrov-Galerkin (MLPG) Method*, Tech Science Press, Encino, CA, 2002.

[2] I. Barrodale and A. Young, *Algorithms for best L_1 and L_∞ linear approximations on a discrete set*, Numer. Math., 8 (1966), pp. 295–306.

[3] C. Franke and R. Schaback, *Convergence order estimates of meshless collocation methods using radial basis functions*, Adv. Comput. Math., 8 (1998), pp. 381–399.

[4] C. Franke and R. Schaback, *Solving partial differential equations by collocation using radial basis functions*, Appl. Math. Comput., 93 (1998), pp. 73–82.

[5] M. A. Golberg and C. S. Chen, eds., *The dual reciprocity method and radial basis functions*, Eng. Anal. Boundary Elements, 24 (2000), pp. 7–8.

[6] Y. C. Hon and R. Schaback, *On unsymmetric collocation by radial basis functions*, Appl. Math. Comput., 119 (2001), pp. 177–186.

[7] E. J. Kansa and Y. C. Hon, *Circumventing the ill-conditioning problem with multiquadric radial basis functions: Applications to elliptic partial differential equations*, Comput. Math. Appl., 39 (2000), pp. 123–137.

[8] E. J. Kansa, *Multiquadrics—A scattered data approximation scheme with applications to computational fluid-dynamics.* I. *Surface approximations and partial derivative estimates*, Comput. Math. Appl., 19 (1990), pp. 127–145.

[9] E. J. Kansa, *Multiquadrics—A scattered data approximation scheme with applications to computational fluid-dynamics.* II. *Solutions to parabolic, hyperbolic and elliptic partial differential equations*, Comput. Math. Appl., 19 (1990), pp. 147–161.

[10] O. A. Ladyzhenskaya, *The boundary value problems of mathematical physics*, Appl. Math. Sci. 49, Springer-Verlag, New York, 1985 (translated from Russian by Jack Lohwater [Arthur J. Lohwater]).

[11] L. Ling, R. Opfer, and R. Schaback, *Results on meshless collocation techniques*, Eng. Anal. Boundary Elements, 30 (2006), pp. 247–253.

[12] W. R. Madych, *Miscellaneous error bounds for multiquadric and related interpolators*, Comput. Math. Appl., 24 (1992), pp. 121–138.

[13] R. Schaback, *On the versatility of meshless kernel methods*, in Advances in Computational & Experimental Engineering & Sciences 428, S.N. Atluri, D.E. Beskos, and D. Polyzos, eds., Universität Göttingen, 2003.
[14] R. Schaback, *Convergence of unsymmetric kernel-based meshless collocation methods*, SIAM J. Numer. Anal., 45 (2005), pp. 333–351.
[15] R. Schaback, *Unsymmetric Meshless Methods for Operator Equations*, 2006, http://www.num.math.uni-goettingen.de/schaback.
[16] H. Wendland, *Scattered Data Approximation*, Cambridge University Press, London, 2005.
[17] Z. M. Wu, *Hermite-Birkhoff interpolation of scattered data by radial basis functions*, Approx. Theory Appl., 8 (1992), pp. 1–10.

SIAM J. NUMER. ANAL.
Vol. 46, No. 3, pp. 1116–1146
© 2008 Society for Industrial and Applied Mathematics

AN A POSTERIORI–A PRIORI ANALYSIS OF MULTISCALE OPERATOR SPLITTING*

D. ESTEP†, V. GINTING‡, D. ROPP§, J. N. SHADID¶, AND S. TAVENER‡

Abstract. In this paper, we analyze a multiscale operator splitting method for solving systems of ordinary differential equations such as those that result upon space discretization of a reaction-diffusion equation. Our goal is to analyze and accurately estimate the error of the numerical solution, including the effects of any instabilities that can result from multiscale operator splitting. We present both an a priori error analysis and a new type of hybrid a priori–a posteriori error analysis for an operator splitting discontinuous Galerkin finite element method. Both analyses clearly distinguish between the effects of the operator splitting and the discretization of each component of the decomposed problem. The hybrid analysis has the form of a computable a posteriori leading order expression and a provably higher order a priori expression. The hybrid analysis takes into account the fact that the adjoint problems for the original problem and a multiscale operator splitting discretization differ in significant ways. In particular, this provides the means to monitor global instabilities that can arise from operator splitting.

Key words. a posteriori error analysis, adjoint problem, discontinuous Galerkin method, generalized Green's function, goal oriented error estimates, multiscale method, operator decomposition, operator splitting, reaction-diffusion equations, residual

AMS subject classifications. 65N15, 65N30, 65N50

DOI. 10.1137/07068237X

1. Introduction.

Operator decomposition is perhaps the most widely used technique for solving multiscale, multiphysics problems. The general approach is to decompose a model into components involving simpler physics over a relatively limited range of scales, and then to seek the solution of the entire system by using numerical solutions of the individual components. This approach has many appealing aspects. For example, it capitalizes on the significant progress that has been made on the stable, accurate, and efficient solution of a broad spectrum of single-physics problems. It provides an avenue to use highly evolved legacy codes to tackle multiphysics problems. It also provides a natural way to tackle problems encompassing multiple time and length scales.

The classic example of operator decomposition is *operator splitting* for reaction-diffusion equations. The generic picture is a relatively fast, destabilizing reaction component interacting with a relatively slow, stabilizing diffusion component. Accuracy considerations dictate the use of relatively small steps to integrate the reaction component. On the other hand, stability considerations over moderate to long time

*Received by the editors February 12, 2007; accepted for publication (in revised form) November 12, 2007; published electronically March 7, 2008.
http://www.siam.org/journals/sinum/46-3/68237.html

†Department of Mathematics and Department of Statistics, Colorado State University, Fort Collins, CO 80523 (estep@math.colostate.edu). This author's work was supported in part by the Department of Energy (DE-FG02-04ER25620, DE-FG02-05ER25699), the National Aeronautics and Space Administration (NNG04GH63G), the National Science Foundation (DMS-0107832, DGE-0221595003, MSPA-CSE-0434354), and the Sandia Corporation (PO299784).

‡Department of Mathematics, Colorado State University, Fort Collins, CO 80523 (ginting@math.colostate.edu, tavener@math.colostate.edu). The work of these authors was supported in part by the Department of Energy (DE-FG02-04ER25620).

§SAIC, 4001 Fairfax Dr., Arlington, VA 22203 (DAVID.L.ROPP@saic.com).

¶Computational Sciences R&D Group, Sandia National Laboratories, P.O. Box 5800, MS0316, Albuquerque, NM 87185 (jnshadi@sandia.gov).

intervals suggest the use of implicit, dissipative numerical methods for integrating diffusion problems. Such methods are expensive to use per step, but relatively large steps can be used on a purely dissipative problem. If the reaction and diffusion components are integrated together, then the small steps required for accurate resolution of the reaction lead to an expensive computation.

In a multiscale operator splitting approach, the reaction and diffusion components are integrated independently inside each time interval of a discretization of time and "synchronized" in some fashion only at the nodes of the interval. The reaction component is often integrated by using significantly smaller substeps (e.g., 10^{-5} smaller is not uncommon) than those used to integrate the diffusion component, which can lead to a tremendous computational savings.

However, operator decomposition presents an entirely new set of accuracy and stability issues, some obvious, some subtle, and all difficult to correct. In the case of operator splitting, the instantaneous interaction between reaction and diffusion is discretized and this has a strong effect on accuracy and stability, even if each component is solved exactly. For example, consider the instability observed in the Brusselator problem [28, 27].

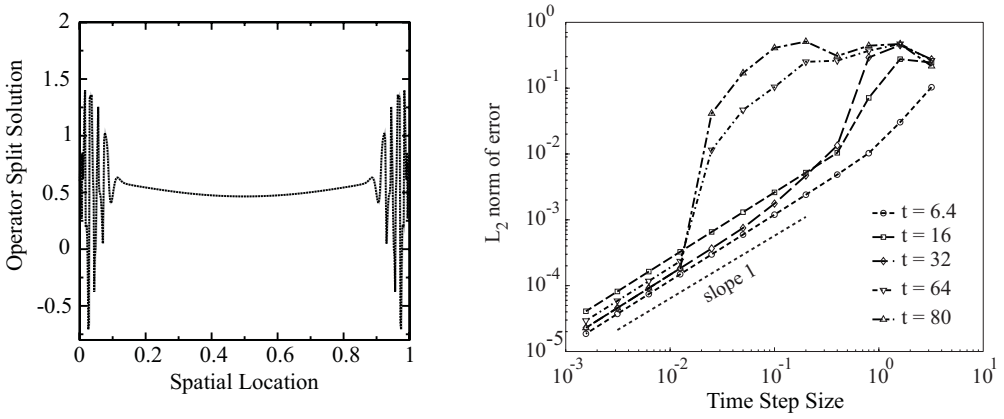


FIG. 1.1. *The left-hand plot illustrates typical instability that can arise from multiscale operator splitting applied to the Brusselator problem. Solution is shown at time* 80. *On the right, we show plots of the error in the* L_2 *norm versus time step size at different times.*

Example 1.1. We illustrate the instability of operator splitting applied to the Brusselator problem in Figure 1.1. We apply a standard first order splitting scheme to a space discretization of the Brusselator model with 500 discrete points (see (6.6) in section 6 with $\alpha = .6$, $\beta = 2$, $k_1 = k_2 = .025$). We integrate using the trapezoidal rule with time step of .2 for the diffusion and backward Euler with time step of .004 for the reaction. On the left of Figure 1.1, we show a numerical solution that exhibits nonphysical oscillations that developed after some time. On the right, we show plots of the error versus time steps at different times. There is a critical time step above which the instability develops. Moreover, changing the space discretization does not improve the accuracy. In [25, 28], it is demonstrated that a finer spatial discretization for a constant time step size leads to significantly more error in the long time solution. The observed instability is a direct consequence of the operator splitting, which in effect separates the stabilizing effect of the diffusion component from the destabilizing reaction component.

Beginning with the classic work of Marchuk and Strang, there is a highly developed literature devoted to a priori convergence analysis of operator splitting for reaction-diffusion equations; see, for example, [32, 23, 5, 20, 31, 34, 8] and the references therein. Almost all of the classic literature assumes that the time steps for the reaction and diffusion components are comparably sized, though Dawson and Wheeler consider multiscale operator splitting for convection-reaction-diffusion problems in [5]. There are also studies of stability properties of various operator splitting procedures; see, for example, [21].

All of the existing convergence analysis is carried out under assumptions that, if strictly enforced, would prevent instabilities arising from the operator splitting discretization. However, enforcing such requirements presents serious problems on both theoretical and practical grounds. This provides the motivation to seek an a posteriori error analysis that devises a computational basis for detecting and correcting instabilities and inaccuracies arising from operator splitting as a particular computation proceeds.

To simplify the presentation, we consider a simple example of operator splitting applied to a model system of ordinary differential equations of the following form: find $y \in \mathbb{R}^l$, $l \geq 1$, such that

$$\begin{cases} \dot{y} = Ay(t) + F(y(t)), & 0 < t \leq T, \\ y(0) = y_0, \end{cases} \tag{1.1}$$

where A is an $l \times l$ constant matrix representing a "diffusion component" and $F(y) = (F_1(y), F_2(y), \ldots, F_l(y))^\top$ is a vector of nonlinear functions representing a "reaction component."

Example 1.2. Such a system arises after semidiscretization of a reaction-diffusion governed by the initial boundary value problem,

$$\begin{cases} \dfrac{\partial u}{\partial t} = \epsilon \, \Delta u + f(u), & x \in \Omega, 0 < t, \\ \text{suitable boundary conditions}, & x \in \partial\Omega, 0 < t, \\ u(\cdot, 0) = u_0(\cdot), \end{cases} \tag{1.2}$$

where $\Omega \subset \mathbb{R}^d$ is a spatial domain. For example, using a continuous, piecewise linear finite element method for a Dirichlet problem with N_e elements, we obtain the initial value problem (1.1). The unknown y consists of the vector of nodal values of the finite element solution with dimension $N_e - 1$, and A is a symmetric negative definite sparse matrix.

Our results extend in a straightforward way to different operator splitting schemes and differential equations with a nonlinear diffusion component, e.g., $\dot{y} = F_D(y) + F_R(y)$, albeit at the cost of complicated notation. Formally, our approach also extends directly to reaction-diffusion problems (1.2) where the splitting is carried out at the continuous differential equation level followed by the discretization independently for each component. A technical issue is dealing with the finite-dimensional representation of the reaction component, which is an ordinary differential equation in a Banach space when the reaction is decoupled from the diffusion.

Our main goal is to derive a computable a posteriori expression that *accurately* estimates the error in a specified quantity of interest computed from a multiscale operator splitting approximate solution of (1.1). The a posteriori analysis is based on duality, adjoint operators, and variational analysis [14, 15, 13, 10, 11, 17, 24, 29].

To deal with operator splitting, we distinguish the effects of operator splitting from the effects of numerical discretization of the components. The latter can be treated with the standard a posteriori analysis. Estimating the error arising from operator splitting requires a new approach.

A main technical issue is the definition of a suitable adjoint problem. The standard approach in error analysis is to form an implicit equation for the error, linearize this problem around an average of the true and approximate solutions using the integral mean value theorem, and then employ the unique adjoint to this linearized error equation [17]. This approach generally works for implicit discretizations, for which the differential operator and its discretization are close in some sense. This approach generally fails or requires significant modification for other discretizations, e.g., explicit, multistep, and implicit/explicit schemes. Indeed, the solution operator for many classes of numerical schemes is associated with *different* adjoint operators than the solution operator for the original problem. This is particularly true for operator decomposition methods, which are generally associated with radically different adjoint operators than the forward problem (see below and [2, 18]). This is one important reason, perhaps unrecognized, why previous a posteriori analyses of evolution problems have focused on fully implicit time integration methods.

Because an adjoint problem carries the global stability information about the quantity of interest computed from the solution, accounting for the differences between adjoint problems associated with the original problem and a numerical discretization is critical for obtaining accurate error estimates. In the estimate derived below, this difference takes the form of "residuals" between certain adjoint operators associated with the fully coupled problem and an analytic operator split version. A practical difficulty with such a result is that solving the adjoint for the fully coupled problem poses the same multiphysics challenges as solving the original forward problem. We therefore develop a new hybrid a priori–a posteriori estimate that combines a computable leading order expression obtained using a posteriori arguments with a provably higher order bound obtained using a priori convergence result.

The rest of the paper is organized as follows. In section 2, we formulate the analytic operator splitting procedure for (1.1) and conduct a preliminary investigation of the instability in operator splitting using an illuminating "blow up" problem. We present a multiscale operator splitting Galerkin finite element method for (1.1) in section 3. We begin the analysis by presenting the results of an a priori convergence analysis of the finite element method in section 4. The main result of the a posteriori analysis of the operator splitting finite element method is presented in section 5, followed by several numerical examples in section 6. In section 7, we give the details of proof of the a priori result. Finally, in section 8, we present a conclusion.

2. Analytic operator splitting. In this section, we define an *analytic* operator splitting version of (1.1). We first discretize $[0,T]$ into $0 = t_0 < t_1 < t_2 < \cdots < t_N = T$ with *diffusion* time steps $\{\Delta t_n\}_{n=1}^N$, $\Delta t_n = t_n - t_{n-1}$, and $\Delta t = \max_{1\le n\le N}(\Delta t_n)$.

For the first discretization, we introduce a theoretical discretization in which each component is solved exactly. We define a piecewise continuous approximate solution

$$\tilde{y}(t) = \frac{t_n - t}{\Delta t_n}\,\tilde{y}_{n-1} + \frac{t - t_{n-1}}{\Delta t_n}\,\tilde{y}_n, \quad t_{n-1} \le t \le t_n, \tag{2.1}$$

with nodal values $\tilde{y}_n$ obtained from the following procedure.

ALGORITHM 2.1 (analytic operator splitting).

- *Set* $\tilde{y}_0 = y_0$.
- *For* $n = 1$ *to* N
  - *Compute* $y^r(t_n^-)$ *satisfying the reaction component*

$$\begin{cases} \dot{y}^r = f(y^r(t)), \quad t_{n-1} < t \le t_n, \\ y^r(t_{n-1}^+) = \tilde{y}_{n-1}. \end{cases} \tag{2.2}$$

  - *Compute* $y^d(t_n^-)$ *satisfying the diffusion component*

$$\begin{cases} \dot{y}^d = Ay^d(t), \quad t_{n-1} < t \le t_n, \\ y^d(t_{n-1}^+) = y^r(t_n^-). \end{cases} \tag{2.3}$$

  - *Set* $\tilde{y}_n = y^d(t_n^-)$.

  end for

Example 2.2. To provide motivation, we consider a problem in which the reaction component exhibits finite time blow up when undamped by the diffusion component. The problem is

$$\begin{cases} \dot{y} + \lambda y = y^2, \qquad t > 0, \\ y(0) = y_0 \in \mathbb{R}, \end{cases} \tag{2.4}$$

which has the exact solution

$$y(t) = \frac{\lambda y_0}{y_0 - (y_0 - \lambda)\, e^{\lambda t}} \tag{2.5}$$

when $\lambda \neq 0$. The exact solution exists for all time and tends to zero as $t \to \infty$ when $\lambda > y_0$. On the other hand, there is finite time blow up if $\lambda < y_0$.

Applying the analytic operator splitting (2.2), (2.3) to (2.4), the solutions of the two components and the true solution are

$$y^r(t) = \frac{y_{n-1}^{d-}}{1 - y_{n-1}^{d-}(t - t_{n-1})}, \quad y^d(t) = e^{-\lambda(t-t_{n-1})}\, y_n^{r-}, \quad \tilde{y}_n = \frac{e^{-\lambda \Delta t_n}\, \tilde{y}_{n-1}}{1 - \Delta t_n\, \tilde{y}_{n-1}}$$

when the reaction component is defined. We see that splitting off the smoothing effect provided by instantaneous interaction with the diffusion component means that the reaction component can blow up in finite time.

The different behavior of the reaction and diffusion components introduces a difference in scales. Consider $\lambda = 2$ with initial condition $y_0 = 1$, so $y(t) = 2/(1+e^{2t})$. This solution has an *asymptotic* decay rate proportional to e^{-2t}, i.e., an *asymptotic* decay time scale of $1/2$. This same asymptotic behavior is observed for any $y_0 < 2$. The solution to the diffusion equation, $y^d(t) = e^{-2(t-t_{n-1})} y_n^{r-}$, has a fixed decay rate proportional to $e^{-2(t-t_{n-1})}$, i.e., a fixed decay time scale $\tau_D = 1/2$. However, the solution to the reaction equation, $y^r(t) = y_{n-1}^{d-}/(1 - y_{n-1}^{d-}(t - t_{n-1}))$, becomes unbounded at $t = t_{n-1} + 1/y_{n-1}^{d-}$ for all $y_{n-1}^{d-} > 0$, and has a doubling time equal to $t_{n-1} + 1/(2y_{n-1}^{d-})$, suggesting a reaction time scale $\tau_R = 1/y_{n-1}^{d-}$. If $y_{n-1}^{d-} \ll 1$, $\tau_R > \tau_D$. If $y_{n-1}^{d-} \approx \lambda$, $\tau_R \approx \tau_D$, but the reaction time scale is not constant and decreases rapidly as the reaction progresses independently undamped by diffusion. Operator splitting has created two uncoupled problems with very different stability properties and with very different time scales. Furthermore, the time scales are not clearly separated nor do they remain in constant proportion. The time scale τ_D is

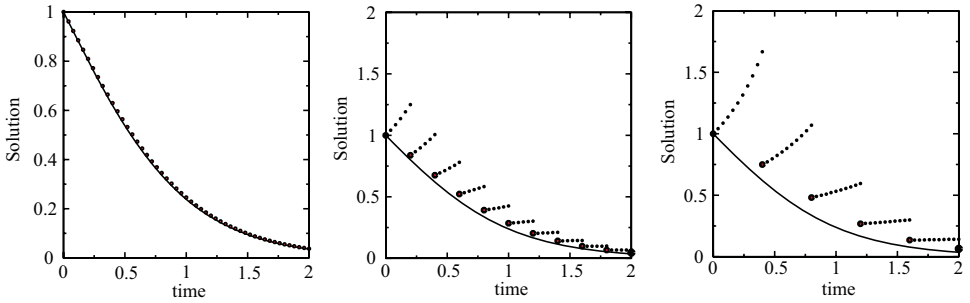


FIG. 2.1. *Plots of the solution $\tilde{y}$ and the true solution. Left: $N = 50$, $M = 1$. Middle: $N = 10$, $M = 5$. Right: $N = 5$, $M = 10$. The nodal values of $\tilde{y}$ are denoted by the larger points, while the smaller points denote node values of the reaction components on the mesh used for the reaction.*

related to the time scale of the coupled process in the limit of small solutions, but the time scale τ_R is completely unrelated.

On the left side of Figure 2.1, we plot the true solution and the nodal values of the approximation $\tilde{y}$ for $N = 50$ diffusion steps and $M = 1$ reaction step per diffusion step. The approximation is reasonably accurate.

Next, we increase the diffusion step by choosing $N = 10$ and, in order to maintain the same resolution, we correspondingly increase to $M = 5$. In Figure 2.1, we plot $\tilde{y}$, the reaction component y^r, and the true solution.

The node values of $\tilde{y}$ are relatively close to those of y. The subsequent nodal values of the reaction component solution y^r inside each step move away from the true solution. This large departure is somewhat counteracted by application of the diffusion operator. The multiscale reaction components exhibit significant growth inside each diffusion step, which severely affects accuracy.

If we increase the diffusion step by taking $N = 5$ and maintain resolution in the reaction component by taking $M = 10$, the approximation becomes even less accurate. In Figure 2.1, we plot $\tilde{y}$, the reaction component y^r, and the true solution. If we increase the diffusion step further, then the reaction component actually blows up inside a diffusion step.

3. A multiscale operator splitting finite element method. We first discretize $[0, T]$ into $0 = t_0 < t_1 < t_2 < \cdots < t_N = T$ with *diffusion* time steps $\{\Delta t_n\}_{n=1}^N$, $\Delta t_n = t_n - t_{n-1}$, and $\Delta t = \max_{1\le n\le N}(\Delta t_n)$. For each diffusion step, we choose a (small) time step $\Delta s_n = \Delta t_n / M_n$ with $\Delta s = \max_{1\le n\le N}(\Delta s_n)$, and the nodes $t_{n-1} = s_{0,n} < s_{1,n} < \cdots < s_{M_n,n} = t_n$ (see Figure 3.1). We associate the time intervals $I_n = [t_{n-1}, t_n]$ and $I_{m,n} = [s_{m-1,n}, s_{m,n}]$ with these discretizations.

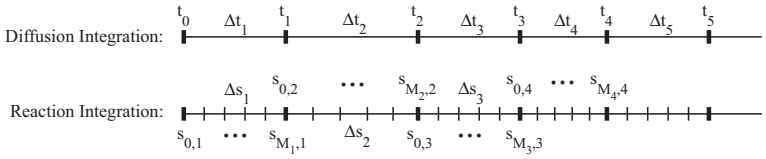


FIG. 3.1. *Discretization of time used for multiscale operator splitting.*

We observe that an operator splitting discretization is actually a consistent discretization of the analytic operator split problem, that is, a consistent discretization

of (2.2) followed by a consistent discretization of (2.3). Since the a posteriori analysis is based on the framework of variational formulation and adjoint problems, we formulate the discretization as a discontinuous Galerkin (dG) finite element method in time [7, 6, 13, 16, 17].

The variational formulation of (1.1) reads, for $n = 1, \ldots, N$, find $y^d \in C^1(I_n)$ such that

$$\begin{cases} \displaystyle\int_{I_n} (\dot{y^d}, v)\, dt = \int_{I_n} (Ay^d, v)\, dt \quad \text{for all } v \in C^1(I_n), \\ y^d(t_{n-1}^+) = y^r(t_n^-), \end{cases} \tag{3.1}$$

where $y^r \in C^1(I_n)$ satisfies

$$\begin{cases} \displaystyle\int_{I_n} (\dot{y^r}, w)\, dt = \int_{I_n} (F(y^r), w)\, dt \quad \text{for all } w \in C^1(I_n), \\ y^r(t_{n-1}^+) = \tilde{y}_{n-1}, \end{cases} \tag{3.2}$$

and we set $\tilde{y}_n = y_n^{d-}$. The finite element approximate solutions are sought in piecewise polynomial spaces,

$$\mathcal{V}^{(q_d)} = \left\{ U : U|_{I_n} \in \mathcal{P}^{(q_d)}(I_n), 1 \le n \le N \right\},$$

$$\mathcal{V}^{(q_r)}(I_n) = \left\{ U : U|_{I_{m,n}} \in \mathcal{P}^{(q_r)}(I_{m,n}), 1 \le m \le M_n \right\}$$

for $n = 1, \ldots, N$, and $I_n = [t_{n-1}, t_n]$ and $I_{m,n} = [s_{m-1,n}, s_{m,n}]$. $\mathcal{P}^{(q_d)}(I_n)$ denotes the space of polynomials in $\mathbb{R}^l$ of degree q_d on I_n. A similar definition holds for $\mathcal{P}^{(q_r)}(I_{m,n})$. We let $U_n^{+,-}$ denote the left- and right-hand limits of U at t_n and $[U]_n = U_n^+ - U_n^-$ the jump value of U at t_n.

Let $\tilde{Y}(t)$ be the piecewise continuous finite element approximation of the operator splitting with

$$\tilde{Y}(t) = \frac{t_n - t}{\Delta t_n}\, \tilde{Y}_{n-1} + \frac{t - t_{n-1}}{\Delta t_n}\, \tilde{Y}_n, \quad t_{n-1} \le t \le t_n.$$

The nodal values $\tilde{Y}_n$ are obtained from the following procedure.

ALGORITHM 3.1 (multiscale operator splitting finite element method).

- *Set* $\tilde{Y}_0 = y_0$.
- *For* $n = 1$ *to* N
  - *Set* $Y_{0,n}^{r-} = \tilde{Y}_{n-1}$.
  - *For* $m = 1$ *to* M_n *compute* $Y^r|_{I_{m,n}} \in \mathcal{P}^{(q_r)}(I_{m,n})$ *satisfying*

$$\int_{I_{m,n}} \left(\dot{Y^r}, W \right) dt + \left([Y^r]_{m-1,n}, W_{m-1}^+ \right) = \int_{I_{m,n}} (F(Y^r), W)\, dt \quad \textit{for all } W \in \mathcal{P}^{(q_r)}(I_{m,n}). \tag{3.3}$$

  end for
  - *Set* $Y_{n-1}^{d-} = Y_{M_n,n}^{r-}$.

– *Compute* $Y^d|_{I_n} \in \mathcal{P}^{(q_d)}(I_n)$ *satisfying*

$$\int_{I_n} (\dot{Y^d}, V)\,dt + \left([Y^d]_{n-1}, V_{n-1}^+\right) = \int_{I_n} (AY^d, V)\,dt \quad \textit{for all } V \in \mathcal{P}^{(q_d)}(I_n). \tag{3.4}$$

– *Set* $\tilde{Y}_n = Y_n^{d-}$.

end for

In general, the dG method using polynomials of degree q converges with order up to $q+1$ at all points t, while its nodal values from the left converge with order up to $2q+1$ under certain conditions; see [11, 13].

Example 3.2. For $q_r = q_d = 0$,

$$\int_{I_{m,n}} F(Y^r)\,dt \to F(Y_{m-1,n}^{r-})\,\Delta s_n, \quad \int_{I_n} AY^d\,dt \to AY_n^{d-}\,\Delta t_n.$$

The dG approximations for the two components are

$$Y_{m,n}^{r-} = Y_{m-1,n}^{r-} + F(Y_{m-1,n}^{r-})\,\Delta s_n, \quad Y_n^{d-} = Y_{n-1}^{d-} + AY_n^{d-}\,\Delta t_n.$$

The former is equivalent to the forward Euler scheme, while the latter is equivalent to the backward Euler scheme.

By employing quadrature formulas to evaluate the integrals in (3.3) and (3.4), many popular finite difference schemes can be described using this variational framework; see [7, 6, 16, 17]. Using quadrature requires a straightforward extension of the a posteriori analysis presented below. The results also extend easily to higher order dG methods as well as the continuous Galerkin (cG) method [15, 17].

Example 3.3. We can use the dG $q_d = 0$ for the diffusion component and the cG $q_r = 1$ implemented with the trapezoidal rule to obtain the reaction component approximation

$$Y_{m,n}^r = Y_{m-1,n}^r + \frac{1}{2}\,\Delta s_n\,\left(F(Y_{m-1,n}^r) + F(Y_{m,n}^r)\right),$$

which is equivalent to the Crank–Nicolson scheme.

4. An a priori convergence analysis. In this section, we carry out an a priori convergence analysis for the multiscale operator splitting dG finite element method. We require this convergence result for the hybrid a posteriori–a priori error analysis. The convergence analysis uses familiar tools from standard results for difference schemes. However, we carry out the analysis in an unusual way by using the analytic operator split problem (2.2) and (2.3) directly.

The goal is to estimate the global error $(y - \tilde{Y}) = (y - \tilde{y}) + (\tilde{y} - \tilde{Y})$, which we decompose as the sum of the error due to the analytical operator splitting and the error due to the numerical discretization of the components of the problem. We give estimates of these two errors in Theorems 4.2 and 4.4, respectively, and combine these in Theorem 4.5.

For $y \in \mathbb{R}^l$, we define the Euclidean norm as $|y|^2 = \sum_{i=1}^l |y_i|^2$.

4.1. A priori error analysis of analytic operator splitting. With $y_n = y(t_n)$, we write the Taylor expansion of $y(t)$ solving (1.1) around t_{n-1} as

$$y(t) = y_{n-1} + (t-t_{n-1})\,(Ay_{n-1} + F(y_{n-1})) + \frac{1}{2}\,(t-t_{n-1})^2\,(A + F'(y_{n-1}))(Ay_{n-1} + F(y_{n-1})) + \mathbf{O}(\Delta t_n^3), \tag{4.1}$$

where $F'(y)$ denotes the Jacobian of $F(y)$. Similarly,

$$y_n^{r-} = y_{n-1}^{r+} + \Delta t_n\,F(y_{n-1}^{r+}) + \frac{1}{2}\,\Delta t_n^2\,F'(y_{n-1}^{r+})\,F(y_{n-1}^{r+}) + \mathbf{O}(\Delta t_n^3)$$

and

$$y_n^{d-} = \left(I + \Delta t_n\,A + \frac{1}{2}\,\Delta t_n^2\,A^2 + \mathbf{O}(\Delta t_n^3)\right)\,y_{n-1}^{d+}.$$

Since $y_{n-1}^{d+} = y_n^{r-}$, $y_{n-1}^{r+} = \tilde{y}_{n-1}$, and $\tilde{y}_n = y_n^{d-}$, we have

$$\tilde{y}_n = \tilde{y}_{n-1} + \Delta t_n\,(A\tilde{y}_{n-1} + F(\tilde{y}_{n-1})) + \frac{1}{2}\,\Delta t_n^2\,\left(A^2\tilde{y}_{n-1} + 2AF(\tilde{y}_{n-1}) + F'(\tilde{y}_{n-1})\,F(\tilde{y}_{n-1})\right) + \mathbf{O}(\Delta t_n^3). \tag{4.2}$$

Furthermore, substitution of (4.2) in (2.1) yields a representation of $\tilde{y}(t)$ with $t_{n-1} \le t \le t_n$:

$$\tilde{y}(t) = \tilde{y}_{n-1} + (t-t_{n-1})\,(A\tilde{y}_{n-1} + F(\tilde{y}_{n-1})) + \frac{1}{2}\,\Delta t_n\,(t-t_{n-1})\,\left(A^2\tilde{y}_{n-1} + F'(\tilde{y}_{n-1})\,F(\tilde{y}_{n-1}) + 2AF(\tilde{y}_{n-1})\right) + \mathbf{O}(\Delta t_n^3). \tag{4.3}$$

LEMMA 4.1. *In the case of* $\tilde{y}_{n-1} = y_{n-1}$,

$$|y_n - \tilde{y}_n| \approx \mathbf{O}(\Delta t_n^2) \quad \textit{and} \quad \int_{I_n} |y(t) - \tilde{y}(t)|\,dt \approx \mathbf{O}(\Delta t_n^3).$$

Proof. With $\tilde{y}_{n-1} = y_{n-1}$, (4.2) yields

$$y_n - \tilde{y}_n = \frac{1}{2}\Delta t_n^2\Big(F'(y_{n-1})Ay_{n-1} - AF(y_{n-1})\Big) + \mathbf{O}(\Delta t_n^3),$$

from which we get the first part of the lemma. Moreover, subtraction of (4.3) from (4.1) with $\tilde{y}_{n-1} = y_{n-1}$ yields

$$y(t) - \tilde{y}(t) = \frac{1}{2}(t-t_{n-1})^2\,(A + F'(y_{n-1}))(Ay_{n-1} + F(y_{n-1})) - \frac{1}{2}\,\Delta t_n\,(t-t_{n-1})\,\left(A^2y_{n-1} + F'(y_{n-1})\,F(y_{n-1}) + 2AF(y_{n-1})\right) + \mathbf{O}(\Delta t_n^3),$$

and the second estimate in the lemma results from integration over I_n. ∎

THEOREM 4.2 (global analytic operator splitting error). *Let y be the solution of* (1.1) *and let $\tilde{y}$ be the solution of the analytical operator splitting* (2.2) *and* (2.3). *Then* $|y_N - \tilde{y}_N| \approx \mathbf{O}(\Delta t)$.

Remark 4.1. If we use higher order splitting, analogous analysis yields a higher order accuracy.

Proof. Subtraction of (4.3) from (4.1) gives

$$\begin{aligned} y_n - \tilde{y}_n &= (y_{n-1} - \tilde{y}_{n-1}) + \Delta t_n \left(F(y_{n-1}) - F(\tilde{y}_{n-1}) + A(y_{n-1} - \tilde{y}_{n-1})\right) + \mathbf{O}(\Delta t_n^2) \\ &\leq C\,|y_{n-1} - \tilde{y}_{n-1}| + \mathbf{O}(\Delta t_n^2), \end{aligned}$$

where C is a generic constant independent of Δt. Summing,

$$|y_N - \tilde{y}_N| \leq C \sum_{n=1}^{N} \mathbf{O}(\Delta t_n^2) \leq C\,\Delta t. \qquad \square$$

Example 4.3. We consider the blow up example in section 2. Assuming a uniform diffusion time step, induction yields

$$\tilde{y}_N = \frac{y_0\, e^{-\lambda t_N}\,(1 - e^{-\lambda \Delta t})}{(1 - e^{-\lambda \Delta t}) - y_0\,\Delta t\,(1 - e^{-\lambda t_N})}.$$

Using (2.5), we obtain an estimate of the analytic operator splitting error

$$|y_N - \tilde{y}_N| = \frac{\left(1 - e^{-2t_N}\right)\,\left(2\Delta t - (1 - e^{-2\Delta t})\right)}{(1 + e^{2t_N})\,((1 - e^{-2\Delta t}) - (1 - e^{-2t_N})\,\Delta t)} \leq C\Delta t.$$

Table 4.1 shows the error versus the time step.

TABLE 4.1
Analytic operator splitting errors at $T = 2.0$ for the blow up example in section 2.

Δt	Error	Order
0.050000	0.00185	1.102
0.025000	0.00089	1.048
0.012500	0.00044	1.023
0.006250	0.00021	1.011
0.003125	0.00010	1.005

4.2. Analysis of the multiscale operator splitting dG finite element method. We now turn to the analysis of the numerical solution of the multiscale operator splitting problem (2.3) and (2.2). The unusual feature of this problem is the numerical solution of an alternating sequence of diffusion and reaction components. The analysis is carried out using the analogue of the standard local error analysis for a finite difference scheme. For each component solve on each interval, we decompose the error as a sum of the error in the initial condition inherited from the previous component solve and the error of the numerical solution of the component assuming exact initial conditions on the current interval. We describe the main results below and give the detailed proof in section 7.

THEOREM 4.4 (numerical discretization error). *Let $\tilde{y}$ be the solution of analytic operator splitting and let $\tilde{Y}$ be the multiscale operator splitting dG finite element solution, which approximates $\tilde{y}$. Assume that there exists a positive constant L such that $|F(u) - F(v)| \leq L\,|u - v|$ for all u, v. Then for $q_d = 0, 1$ and $q_r = 0, 1$,*

$$|\tilde{y}_N - \tilde{Y}_N| \approx \mathbf{O}(\Delta t^{q_d+1}) + \mathbf{O}(\Delta s^{q_r+1}). \tag{4.4}$$

Proof. We set $e = \tilde{y} - \tilde{Y}$ and write

$$e_n = \tilde{y}_n - \tilde{Y}_n = y_n^{d-} - Y_n^{d-} = (y_n^{d-} - X_n^-) + (X_n^- - Y_n^{d-}) = \theta_n^- + \zeta_n^-, \tag{4.5}$$

where X and θ are defined in Lemma 7.5. Subtracting (3.4) from (7.6) yields

$$\int_{I_n} (\dot{\zeta}, V)\, dt - \int_{I_n} (A\zeta, V)\, dt + (\zeta_{n-1}^+, V_{n-1}^+) = (e_n^{r-}, V_{n-1}^+)$$

for all $V \in \mathcal{P}^{(q_d)}(I_n)$, where $e_n^{r-} = y_n^{r-} - Y_n^{r-}$. Now, $\zeta \in \mathcal{V}^{(q_d)}$. Arguing as in the proof of Lemma 7.1, we find

$$|\zeta_n^-| \le \exp(8C_A \Delta t_n)\, |e_n^{r-}|,$$

where $C_A = |A|$. By Lemma 7.2, this inequality together with (4.5) and Lemma 7.5 give the recursive relation

$$|e_n| \le C\left(\Delta t_n^{q_d+2} + \exp(\tau_n)\Delta s_n^{q_r+1}\, \Delta t_n + \exp(2\tau_n)\, |e_{n-1}|\right),$$

where $\tau_n = C\Delta t_n$. From this recursive inequality, we have

$$|e_N| \le C\, \Delta t^{q_d+1} \sum_{n=1}^{N} \exp(2(n-1)\tau_n)\Delta t_n + C\, \Delta s^{q_r+1} \sum_{n=1}^{N} \exp((2n-1)\tau_n)\, \Delta t_n,$$

yielding the desired result. ∎

Combining Theorems 4.2 and 4.4 we have the following global a priori error bound.

THEOREM 4.5 (global a priori error bound). *Let $\tilde{Y}$ be a multiscale operator splitting dG finite element solution. Assume that there exists a positive constant L such that $|F(u) - F(v)| \le L\,|u - v|$ for all u, v. Then for $q_d = 0, 1$ and $q_r = 0, 1$, there exists constants C_1, C_2, C_3 such that*

$$|y_N - \tilde{Y}_N| \le C_1 \Delta t + C_2 \Delta t^{q_d+1} + C_3 \Delta s^{q_r+1}.$$

5. An a posteriori–a priori analysis. We now turn to the hybrid a posteriori–a priori analysis of the multiscale operator splitting dG finite element method. As mentioned in section 1, the analysis differs from the standard a posteriori analysis for an evolution problem in several ways. Mainly, we have to account for the fact that different adjoint operators are associated with the fully coupled problem (1.1) and the analytic operator split version (2.2)–(2.3). The difference between these adjoints provides the means to estimate the effects of operator splitting on the stability properties of a quantity of interest computed from a solution.

5.1. Defining adjoint problems. To define adjoint operators for the forward problems, we seek analogues of the classic representation formula involving the Green's function of a linear elliptic problem. In order to obtain unique adjoint problems from this condition, we use linearization. We assume that $y = 0$ is a steady state solution of (1.1), which can be achieved by assuming that

$$\textit{Homogeneity assumption:} \quad F(0) = 0,$$

and we linearize in a region around 0. In terms of applications to reaction-diffusion problems, there are mathematical reasons for making the homogeneity assumption [30], and it is satisfied in a great many cases. We can modify the analysis to allow for linearization around any other constant steady state solution and more generally around any known function of time. The selection of a suitable "point" for linearization requires some insight into the behavior of the solution being approximated. We illustrate below in the bistable example presented in section 6.

On time interval (t_{n-1}, t_n), we consider the linearized problem

$$\begin{cases} \dot{y} = A\, y(t) + \overline{F'(y)}\, y(t), & t_{n-1} < t \le t_n, \\ y(t_{n-1}) = y_{n-1}, \end{cases}$$

where $\overline{F'(y)}$ is a matrix whose ijth entry is expressed as

$$\overline{F'(y)}_{ij} = \int_0^1 \frac{\partial F_i(sy)}{\partial y_j}\, ds.$$

We note that $\overline{F'(y)}y = F(y)$ because $F(0) = 0$. The generalized Green's function φ satisfies the adjoint problem

$$\begin{cases} -\dot{\varphi} = A^\top \varphi(t) + \overline{F'(y)}^\top \varphi(t), & t_n > t \ge t_{n-1}, \\ \varphi(t_n) = \psi_n, \end{cases} \tag{5.1}$$

where ψ_n determines the quantity of interest $(y(t_n), \psi_n)$, and $A^\top$ and $\overline{F'(y)}^\top$ denote the transpose of A and $\overline{F'(y)}$, respectively.

Remark 5.1. The analysis extends to cover quantities of interest of the form $\int_0^T (y, \psi)\, dt$ in a straightforward way.

We choose $\psi_n = \varphi(t_n^+)$, which couples the local adjoint problems (5.1) to form a global adjoint problem. This definition yields a simple representation of the solution value over one time step. Taking the Euclidean inner product of (5.1) with y and integrating over I_n gives

$$\begin{aligned} 0 &= \int_{I_n} \left(-\dot{\varphi} - A^\top \varphi(t) - \overline{F'(y)}^\top \varphi(t), y \right) dt \\ &= -\int_{I_n} \frac{d}{dt}(\varphi, y)\, dt + \int_{I_n} (\dot{y} - Ay - \overline{F'(y)}y, \varphi) dt. \end{aligned} \tag{5.2}$$

The second term on the right-hand side vanishes by (1.1), and thus integrating the first term yields

$$(y_n, \psi_n) = (y_{n-1}, \varphi_{n-1}). \tag{5.3}$$

Using the analogous approach for each component of the analytic operator splitting (2.2) and (2.3), we define the two adjoint problems for $n = 1, \ldots, N$,

$$\begin{cases} -\dot{\varphi^r} = \overline{F'(y^r)}^\top \varphi^r(t), & t_n > t \ge t_{n-1}, \\ \varphi^r(t_n^-) = \psi_n^r, \end{cases} \tag{5.4}$$

$$\begin{cases} -\dot{\varphi^d} = A^\top \varphi^d(t), & t_n > t \ge t_{n-1}, \\ \varphi^d(t_n^-) = \psi_n^d. \end{cases} \tag{5.5}$$

Note that (5.4) is linearized around y^r. We obtain the following representations for the component solutions:

$$(y_n^{r-}, \psi_n^r) = (y_{n-1}^{r+}, \varphi_{n-1}^{r+}) = (\tilde{y}_{n-1}, \varphi_{n-1}^{r+}), \tag{5.6}$$

$$(y_n^{d-}, \psi_n^d) = (y_{n-1}^{d+}, \varphi_{n-1}^{d+}) = (y_n^{r-}, \varphi_{n-1}^{d+}). \tag{5.7}$$

5.2. A local representation of the analytic operator splitting error. As with the a priori analysis, we begin with the decomposition

$$y - \tilde{Y} = (y - \tilde{y}) + (\tilde{y} - \tilde{Y}). \tag{5.8}$$

In order to construct a representation of the nodal value $y_n - \tilde{y}_n$, where $\tilde{y}_n = y_n^{d-}$, it is natural to set $\psi_n^r = \varphi_{n-1}^{d+}$. Thus, taken together, the adjoint components (5.4)–(5.5) are analogous to an analytic operator split version of the adjoint problem (5.1), except that the linearization is taken around y^r instead of y. Note that the adjoint component problems (5.4)–(5.5) are solved "in the reverse order," which mirrors the fact that the adjoint of a composition of linear transformations is equal to the composition of the adjoints of the transformations computed in reverse order.

Now we combine (5.6) and (5.7) to get

$$(\tilde{y}_n, \psi_n^d) = (y_n^{d-}, \psi_n^d) = (\tilde{y}_{n-1}, \varphi_{n-1}^{r+}). \tag{5.9}$$

To obtain a representation of the local error of operator splitting, we assume that the adjoint problems (5.1) and (5.5) and (5.4) have the same data $\psi_n^d = \psi_n$ at time t_n. Using (5.3) and (5.9) we obtain the error representation

$$(y_n - \tilde{y}_n, \psi_n) = (y_{n-1}, \varphi_{n-1}) - (\tilde{y}_{n-1}, \varphi_{n-1}^{r+}). \tag{5.10}$$

At this point, it is convenient to change notation and introduce the solution operators of the various adjoint problems. Let $\Phi_n(y)$ be the solution operator for the full adjoint problem (5.1) so $\Phi_n(y)\psi_n = \varphi_{n-1}$. For a problem of the form (1.1),

$$\Phi_n(y) = \exp\left(\Delta t_n A^\top + \int_{I_n} \overline{F'(y)}^\top \, dt\right).$$

Similarly, we set $\Phi_n^r(y^r)$ and Φ_n^d to be the solution operators for (5.4) and (5.5), respectively, so

$$\Phi_n^r(y^r)\varphi_n^{r-} = \varphi_{n-1}^{r+} \quad \text{and} \quad \Phi_n^d \psi_n^d = \varphi_{n-1}^{d+}.$$

For (1.1),

$$\Phi_n^r(y^r) = \exp\left(\int_{I_n} \overline{F'(y^r)}^\top \, dt\right) \quad \text{and} \quad \Phi_n^d = \exp\left(\Delta t_n A^\top\right).$$

Under the assumptions above, $\varphi_{n-1}^{r+} = \Phi_n^r(y^r)\Phi_n^d\psi_n$, and (5.10) reads

$$(y_n - \tilde{y}_n, \psi_n) = (y_{n-1}, \Phi_n(y)\psi_n) - (\tilde{y}_{n-1}, \Phi_n^r(y^r)\Phi_n^d\psi_n).$$

To conduct a local analysis, we assume that the forward problems have the correct data $\tilde{y}_{n-1} = y_{n-1}$. Thus, we obtain the following.

THEOREM 5.1 (local analytical splitting error). *The splitting error over a single diffusion time step is represented by*

$$(y_n - \tilde{y}_n, \psi_n) = (\tilde{y}_{n-1}, \Delta\Phi_n \, \psi_n), \tag{5.11}$$

with $\Delta\Phi_n = \Phi_n(y) - \Phi_n^r(y^r)\Phi_n^d$.

We note that the first term on the right-hand side of (5.11) includes the effects associated with the linearization of the nonlinear reaction function around y and y^r, respectively. The goal now is to use (5.11) to derive a computable estimate.

5.3. An error representation for the multiscale operator splitting dG finite element solution. We now derive the error estimate for the numerical discretization of the analytic operator splitting. Let ϑ^d define the adjoint solution associated with (3.4) (diffusion component) satisfying

$$\begin{cases} -\dot{\vartheta^d} = A^\top \vartheta^d(t), & t_n > t \geq t_{n-1}, \\ \vartheta^d(t_n^-) = \psi_n. \end{cases}$$

Furthermore, let ϑ^r define the adjoint solution associated with (3.3) (reaction component) satisfying

$$\begin{cases} -\dot{\vartheta^r} = (\hat{F}'(y^r, Y^r))^\top \vartheta^r(t), & s_{m,n} > t \geq s_{m-1,n}, \\ \vartheta^r(s_{m,n}) = \psi^r_{m,n} \end{cases}$$

for $m = M_n, \ldots, 1$, with $\psi^r_{M_n,n} = \vartheta^{d+}_{n-1}$ and $\psi^r_{m,n} = \vartheta^r_{m,n}$ for $m < M_n$. Thus ϑ^r is continuous across the internal reaction time nodes $s_{m,n}$, $m = 1, \ldots, M_n - 1$. Here $\hat{F}'(y^r, Y^r)$ is the Jacobian matrix linearized along the trajectory $zy^r + (1-z)Y^r$, with entries

$$\hat{F}'_{ij} = \int_0^1 \frac{\partial F_i}{\partial y_j}(zy^r + (1-z)Y^r)\, dz.$$

We set $e^d = y^d - Y^d$ and $e^r = y^r - Y^r$. Obviously on time interval I_n,

$$\int_{I_n} (e^d, \dot{\vartheta^d} + A^\top \vartheta^d)\, dt = 0 \quad \text{and} \quad \int_{I_n} (\dot{y^d} - Ay^d, \vartheta^d)\, dt = 0.$$

Using integration by parts

$$\int_{I_n} (e^d, \dot{\vartheta^d})\, dt = (e^{d-}_n, \psi_n) - (e^{d+}_{n-1}, \vartheta^{d+}_{n-1}) - \int_{I_n} (\dot{e^d}, \vartheta^d)\, dt,$$

we obtain

$$(e^{d-}_n, \psi_n) = (e^{d+}_{n-1}, \vartheta^{d+}_{n-1}) - \int_{I_n} (\dot{Y^d} - AY^d, \vartheta^d)\, dt.$$

Furthermore, taking $V = \Pi\vartheta^d \in \mathcal{V}^{(q_d)}$ in (3.4) and using the fact that $e^{d+}_{n-1} = e^{r-}_{M_n,n} - [Y^d]_{n-1}$ yields

$$(e^{d-}_n, \psi_n) = (e^{r-}_n, \vartheta^{d+}_{n-1}) - \int_{I_n} (\dot{Y^d} - AY^d, \vartheta^d - \Pi\vartheta^d)\, dt - ([Y^d]_{n-1}, \vartheta^{d+}_{n-1} - \Pi\vartheta^{d+}_{n-1}).$$

To get a representation for e^{r-}_n, we use similar arguments to obtain

$$0 = \sum_{m=1}^{M_n} \left((e^{r-}_{m,n}, \vartheta^{r-}_{m,n}) - (e^{r+}_{m-1,n}, \vartheta^{r+}_{m-1,n}) - \int_{I_{m,n}} (\dot{e^r} - (\hat{F}'(y^r, Y^r))e^r, \vartheta^r)\, dt \right).$$

Since $(\hat{F}'(y^r, Y^r))e^r = F(y^r) - F(Y^r)$, and y^r satisfies (2.2), this reduces to

$$0 = \sum_{m=1}^{M_n} \left((e^{r-}_{m,n}, \vartheta^{r-}_{m,n}) - (e^{r+}_{m-1,n}, \vartheta^{r+}_{m-1,n}) + \int_{I_{m,n}} (\dot{Y}^r - F(Y^r), \vartheta^r)\, dt \right).$$

Taking $V = \Pi\vartheta^r \in \mathcal{V}^{(q_r)}(I_n)$ in (3.3), we then have

$$0 = \sum_{m=1}^{M_n} (e^{r-}_{m,n}, \vartheta^{r-}_{m,n}) - (e^{r-}_{m-1,n}, \vartheta^{r+}_{m-1,n}) + \int_{I_{m,n}} (\dot{Y}^r - F(Y^r), \vartheta^r - \Pi\vartheta^r)\, dt + ([Y^r]_{m-1,n}, \vartheta^{r+}_{m-1,n} - \Pi\vartheta^{r+}_{m-1,n}).$$

Now since $\vartheta^{r-}_{m,n} = \vartheta^{r+}_{m,n}$, $e^{r-}_{0,n} = \tilde{y}_{n-1} - \tilde{Y}_{n-1}$, we get

$$(e^{r-}_{M_n,n}, \vartheta^{d+}_{n-1}) = (\tilde{y}_{n-1} - \tilde{Y}_{n-1}, \vartheta^{r+}_{0,n}) - \sum_{m=1}^{M_n} \left(\int_{I_{m,n}} (\dot{Y}^r - f(Y^r), \vartheta^r - \Pi\vartheta^r)\, dt + ([Y^r]_{m-1,n}, \vartheta^{r+}_{m-1,n} - \Pi\vartheta^{r+}_{m-1,n}) \right).$$

Note that $e^{d-}_n = \tilde{y}_n - \tilde{Y}_n$, and thus we can combine all these representations into the following result.

THEOREM 5.2 (local discretization error). *The discretization error over a single diffusion step is represented by*

$$(\tilde{y}_n - \tilde{Y}_n, \psi_n) = (\tilde{y}_{n-1} - \tilde{Y}_{n-1}, \vartheta^{r+}_{0,n}) - \sum_{m=1}^{M_n} \left(\int_{I_{m,n}} (\dot{Y}^r - F(Y^r), \vartheta^r - \Pi\vartheta^r)\, dt + ([Y^r]_{m-1,n}, \vartheta^{r+}_{m-1,n} - \Pi\vartheta^{r+}_{m-1,n}) \right) - \int_{I_n} (\dot{Y}^d - AY^d, \vartheta^d - \Pi\vartheta^d)\, dt - ([Y^d]_{n-1}, \vartheta^{d+}_{n-1} - \Pi\vartheta^{d+}_{n-1}). \tag{5.12}$$

5.4. A computable error representation. In light of (5.12) and (5.11), which represent the errors over a single diffusion step, and using $\tilde{y}_{n-1} = y_{n-1}$, we obtain the recursive relation

$$(y_n - \tilde{Y}_n, \psi_n) = (y_{n-1} - \tilde{Y}_{n-1}, \vartheta^{r+}_{0,n}) + (y_{n-1}, \Delta\Phi_n\, \psi_n) - \sum_{m=1}^{M_n} \left(\int_{I_{m,n}} (\dot{Y}^r - F(Y^r), \vartheta^r - \Pi\vartheta^r)\, dt + ([Y^r]_{m-1,n}, \vartheta^{r+}_{m-1,n} - \Pi\vartheta^{r+}_{m-1,n}) \right) - \int_{I_n} (\dot{Y}^d - AY^d, \vartheta^d - \Pi\vartheta^d)\, dt - ([Y^d]_{n-1}, \vartheta^{d+}_{n-1} - \Pi\vartheta^{d+}_{n-1}). \tag{5.13}$$

Undoing (5.13) after choosing the adjoint data $\psi_{n-1} = \vartheta^{r+}_{0,n}$ for $n = 2, \ldots, N-1$ and summing the resulting expressions up to N yields the following result.

THEOREM 5.3. *Given an adjoint data ψ_N, the operator splitting error at the final time can be represented as*

$$
\begin{aligned}
(y_N - \tilde{Y}_N, \psi_N) &= \sum_{n=1}^{N} (y_{n-1}, \Delta\Phi_n\, \psi_n) \\
&- \sum_{n=1}^{N} \sum_{m=1}^{M_n} \left(\int_{I_{m,n}} \left(\dot{Y^r} - F(Y^r), \vartheta^r - \Pi\vartheta^r \right) dt + \left([Y^r]_{m-1,n}, \vartheta^{r+}_{m-1,n} - \Pi\vartheta^{r+}_{m-1,n} \right) \right) \\
&- \sum_{n=1}^{N} \left(\int_{I_n} \left(\dot{Y^d} - AY^d, \vartheta^d - \Pi\vartheta^d \right) dt + \left([Y^d]_{n-1}, \vartheta^{d+}_{n-1} - \Pi\vartheta^{d+}_{n-1} \right) \right) \\
&= \tilde{Q}_1 + Q_2 + Q_3,
\end{aligned} \tag{5.14}
$$

where $\psi_{n-1} = \vartheta^{r+}_{0,n}$ for $n = 2, \ldots, N-1$.

$\tilde{Q}_1$ represents the accumulated effects of operator splitting measured using the *adjoint residual* $\Delta\Phi_n\, \psi_n$ weighted by the solution values y_{n-1}. Note that if y tends to zero as time passes, then the accumulated effects of operator splitting decrease correspondingly. On the other hand, if y is increasing, then the effects of operator splitting also increase. The quantities Q_2 and Q_3 are typical a posteriori error representations for the dG finite element approximate solutions of the component problems and are computable up to the linearization of the error equation (which typically does not cause difficulty [17, 9].) Note that we expect the dG approximations to be much closer to the analytic operator split solutions, $\mathbf{O}(\Delta s^{q_r+1})$ and $\mathbf{O}(\Delta t^{q_d+1})$, respectively, than to the solution of the full problem.

Unfortunately, $\tilde{Q}_1$ contains quantities that are not computable, namely, the exact solution of the fully coupled problem (1.1) and the associated adjoint operator. Even if we have the true solution, we cannot expect to solve the adjoint of the full problem since this nominally presents the same difficulties as the original forward problem. We now decompose $\tilde{Q}_1$ into a computable leading order expression and a remainder that is provably higher order. As a first step,

$$
\begin{aligned}
(y_N - \tilde{Y}_N, \psi_N) &= \sum_{n=1}^{N} \left(\tilde{Y}_{n-1}, \Delta\Phi_n\, \psi_n \right) + Q_2 + Q_3 + \sum_{n=1}^{N} \left(y_{n-1} - \tilde{Y}_{n-1}, \Delta\Phi_n\, \psi_n \right) \\
&= Q_1 + Q_2 + Q_3 + Q_4,
\end{aligned} \tag{5.15}
$$

where Q_2 and Q_3 are the same as for (5.14), and the new expression Q_1 does not depend directly on the solution values. Next, we derive a computable estimate for the factor $\Delta\Phi_n\, \psi_n$ in Q_1 and then use asymptotic analysis to prove that after substituting the estimate the remainders from Q_1 as well as from Q_4 are higher order.

Example 5.4. To motivate (5.15) as a natural representation, we consider the blow up example in section 2 once again. Since there is no numerical solution involved, the expressions Q_2 and Q_3 drop out. The adjoint associated with the blow up problem is

$$
\varphi(t) = \Phi_n(t)\, \psi_n, \quad t_n > t \geq t_{n-1}, \quad \Phi_n(t) = \frac{y_{n-1} - (y_{n-1} - \lambda)e^{\lambda(t-t_{n-1})}}{y_{n-1} - (y_{n-1} - \lambda)e^{\lambda\Delta t_n}}.
$$

Since the diffusion component is linear, the associated adjoint is readily seen to be

$$
\varphi^d(t) = \Phi^d_n(t)\, \psi^d_n, \quad t_n > t \geq t_{n-1}, \quad \Phi^d_n(t) = e^{\lambda(t-t_n)}.
$$

For the reaction component, the associated adjoint is

$$\varphi(t) = \Phi_n^r(t)\,\psi_n, \quad t_n > t \geq t_{n-1}, \quad \Phi_n^r(t) = \frac{1 - y_{n-1}^{d-}(t - t_{n-1})}{1 - y_{n-1}^{d-}\Delta t_n}$$

for $t \in I_n$. At the final time level t_N, (5.15) reads

$$y_N - \tilde{y}_N = \sum_{n=1}^{N} y_{n-1}^{d-}\,\Delta\Phi_n(t_{n-1})\,\psi_n + \sum_{n=1}^{N} \left(y_{n-1} - y_{n-1}^{d-}\right)\,\Delta\Phi_n(t_{n-1})\psi_n,$$

where $\psi_N = 1$ and $\psi_n = \Phi_{n+1}^r(t_n)\,\Phi_{n+1}^d(t_n)\,\psi_{n+1}$ for $1 \leq n < N$.

Table 5.1 illustrates the performance of the a posteriori error representation. We use $\lambda = 0.9$ and initial condition $y_0 = 1$. The column with title "order" shows the computed order of convergence of Q_4 to zero.

TABLE 5.1
Operator splitting estimated errors at $T = 2.0$ for blow up example in section 2.

Δt	Exact	Q_1	Q_4	Order
1/80	−0.176863	−0.183464	0.006600	2.176
1/160	−0.085044	−0.086602	0.001557	2.083
1/320	−0.041723	−0.042101	0.000378	2.040
1/640	−0.020667	−0.020760	0.000093	2.020
1/1280	−0.010285	−0.010309	0.000023	2.010

This example suggests that

$$\Delta\Phi_n = \Phi_n(y) - \Phi_n^r(y^r)\Phi_n^d \approx \mathbf{O}(\Delta t_n^2). \tag{5.16}$$

If this is true, then Q_4 is $\mathbf{O}(\Delta t^{q_d+2})$ and can be ignored in the asymptotic limit when computing an estimate.

It helps to decompose (5.16) into two separate quantities,

$$\Phi_n(y) - \Phi_n^r(y^r)\Phi_n^d = \left(\Phi_n(y) - \Phi_n^r(y)\Phi_n^d\right) + \left((\Phi_n^r(y) - \Phi_n^r(y^r))\;\Phi_n^d\right) = E_1 + E_2.$$

E_1 measures the splitting error around the true solution, while E_2 measures the effect of switching the linearization from around the true solution y to around y^r in the nonlinear reaction.

We next derive an asymptotic representation for E_1.

LEMMA 5.5. *$E_1 = \overline{E}_1 + \mathbf{O}(\Delta t_n^3)$, with*

$$\overline{E}_1 = \frac{1}{2}\Delta t_n \left(A^\top R(\tilde{y}) - R(\tilde{y})\,A^\top\right) \quad \textit{and} \quad R(y) = \int_{I_n} \overline{F'(y)}^\top\, dt.$$

Proof. Using the expansion $\exp(B) = \sum_{i=0}^{\infty} \frac{1}{i!} B^i$ for a square matrix B, E_1 can be expressed as

$$E_1 = \frac{1}{2}\left(\Delta t_n A^\top\, R(y) - R(y)\,\Delta t_n A^\top\right) + \mathbf{O}(\Delta t_n^3). \tag{5.17}$$

E_1 can be approximated by the first term in (5.17) up to $\mathbf{O}(\Delta t_n^3)$. For practical reasons, the dependence on y has to be avoided. A natural candidate to replace y is the operator splitting solution $\tilde{y}$, which is approximated to high order by the operator

splitting dG finite element method. By adding and subtracting appropriate terms, we write

(5.18)
$$E_1 \approx \frac{1}{2}\Delta t_n \left(A^\top R(\tilde{y}) - R(\tilde{y})\, A^\top\right) + \frac{1}{2}\Delta t_n A^\top \left(R(y) - R(\tilde{y})\right) + \frac{1}{2}\Delta t_n (R(\tilde{y}) - R(y))\, A^\top.$$

It is obvious that the first term is computable. The task is to demonstrate that the remaining terms in (5.18) are higher order. Specifically, we show that

$$\left(\tilde{Y}_{n-1}, A^\top \left(R(y) - R(\tilde{y})\right) \psi_n\right) \approx \mathbf{O}(\Delta t_n^2) \tag{5.19}$$

and

$$\left(\tilde{Y}_{n-1}, (R(\tilde{y}) - R(y))\, A^\top \psi_n\right) \approx \mathbf{O}(\Delta t_n^2). \tag{5.20}$$

For (5.19), we may write

$$\begin{aligned}\left(\tilde{Y}_{n-1}, A^\top (R(y) - R(\tilde{y}))\psi_n\right) &= \left(A\tilde{Y}_{n-1}, (R(y) - R(\tilde{y}))\psi_n\right) \\ &\le |A|_{l_\infty} |\tilde{Y}_{n-1}|_{l_\infty} |R(y) - R(\tilde{y})|_{l_1} |\psi_n|_{l_1},\end{aligned}$$

where we have used the usual vector norms and the corresponding subordinate matrix norms. In particular, since $R(y)$ involves the matrix transpose,

$$|R(y) - R(\tilde{y})|_{l_1} = \max_{1\le i\le l} \sum_{j=1}^{l} \left| \int_{I_n} (\overline{F'}_{ij}(y) - \overline{F'}_{ij}(\tilde{y}))\, dt \right| \le C \int_{I_n} |y - \tilde{y}|\, dt,$$

where we have used the Lipschitz continuity of $\overline{F'}_{ij}$. By Lemma 4.1 this yields (5.19). Similar arguments hold for (5.20) and the proof is complete. ∎

Next, we seek an approximate estimate for E_2 that does not depend on the unknown solution y.

LEMMA 5.6. *Assume that* $\frac{\partial F_i}{\partial y_j}$ *and* $\frac{\partial^2 F_i}{\partial y_j \partial y_k}$ *are Lipschitz continuous. Then* $E_2 = \overline{E}_2 + \mathbf{O}(\Delta t_n^3)$, *with* $\overline{E}_2 = \left(\Phi_n^r(\tilde{y}) - \Phi_n^r(y^r)\right) \Phi_n^d$.

Proof. We write

$$E_2 = \left(\Phi_n^r(\tilde{y}) - \Phi_n^r(y^r)\right) \Phi_n^d + \left(\Phi_n^r(y) - \Phi_n^r(\tilde{y})\right) \Phi_n^d.$$

We want to show that

$$\left(\Phi_n^r(y) - \Phi_n^r(\tilde{y})\right) \Phi_n^d \approx \mathbf{O}(\Delta t_n^3).$$

This requires estimating the difference between the solution operators of the reaction component adjoints corresponding to linearization around y and $\tilde{y}$. The solution operator $\Phi_n^r(y)$ is associated with the problem

$$\begin{cases} -\dot{\omega} = \overline{F'(y)}^\top \omega(t), & t_n > t \ge t_{n-1}, \\ \omega(t_n) = \Phi_n^d \psi_n, \end{cases} \tag{5.21}$$

while the solution operator $\Phi_n^r(\tilde{y})$ is associated with

$$\begin{cases} -\dot{\tilde{\omega}} = \overline{F'(\tilde{y})}^\top \tilde{\omega}(t), & t_n > t \ge t_{n-1}, \\ \tilde{\omega}(t_n) = \Phi_n^d \psi_n. \end{cases} \tag{5.22}$$

We expect that if $\tilde{y}$ closely approximates y, then the solutions of these differential equations are close as well. In fact, Lemma 4.1 implies that $|y_n - \tilde{y}_n| \approx \mathbf{O}(\Delta t_n^2)$ for $\tilde{y}_{n-1} = y_{n-1}$.

Using a Taylor expansion, we write

$$\begin{aligned}\omega(t_{n-1}) &= \omega(t_n) - \Delta t_n \frac{d\omega(t_n)}{dt} + \frac{1}{2}\Delta t_n^2 \frac{d^2\omega(t_n)}{dt^2} + \mathbf{O}(\Delta t_n^3) \\ &= \left(I + \Delta t_n \overline{F'(y_n)}^\top - \frac{1}{2}\Delta t_n^2 \left(\frac{d\overline{F'(y_n)}^\top}{dt} + [\overline{F'(y_n)}^\top]^2 \right) \right) \Phi_n^d \psi_n + \mathbf{O}(\Delta t_n^3).\end{aligned}$$

Similarly,

$$\tilde{\omega}(t_{n-1}) = \left(I + \Delta t_n \overline{F'(\tilde{y_n})}^\top - \frac{1}{2}\Delta t_n^2 \left(\frac{d\overline{F'(\tilde{y_n})}^\top}{dt} + [\overline{F'(\tilde{y_n})}^\top]^2 \right) \right) \Phi_n^d \psi_n + \mathbf{O}(\Delta t_n^3).$$

In these two expansions, we have applied the initial condition $\omega(t_n) = \tilde{\omega}(t_n) = \Phi_n^d \psi_n$. Taking the difference between the expansions,

$$\begin{aligned}&\omega(t_{n-1}) - \tilde{\omega}(t_{n-1}) \\ &= \left(\Delta t_n (\overline{F'(y_n)}^\top - \overline{F'(\tilde{y}_n)}^\top) + \frac{1}{2}\Delta t_n^2 \frac{d}{dt} \left(\overline{F'(\tilde{y}_n)}^\top - \overline{F'(y_n)}^\top \right) \right. \\ &\qquad \left. + \frac{1}{2}\Delta t_n^2 \left(\left[\overline{F'(\tilde{y_n})}^\top \right]^2 - \left[\overline{F'(y_n)}^\top \right]^2 \right) \right) \Phi_n^d \psi_n + \mathbf{O}(\Delta t_n^3).\end{aligned} \tag{5.23}$$

Now, we show that $\overline{F'(y_n)}^\top - \overline{F'(\tilde{y}_n)}^\top$ is $\mathbf{O}(\Delta t_n^2)$, and that $([\overline{F'(\tilde{y_n})}^\top]^2 - [\overline{F'(y_n)}^\top]^2)$ and $d/dt(\overline{F'(\tilde{y}_n)}^\top - \overline{F'(y_n)}^\top)$ are at least $\mathbf{O}(\Delta t_n)$. In elemental form, we may write

$$\overline{F'(y_n)}_{ij} - \overline{F'(\tilde{y}_n)}_{ij} = \int_0^1 \left(\frac{\partial F_i(sy_n)}{\partial y_j} - \frac{\partial F_i(s\tilde{y}_n)}{\partial y_j} \right) ds \leq C|y_n - \tilde{y}_n| \approx \mathbf{O}(\Delta t_n^2),$$

where we have used the assumption that $\partial F_i(y)/\partial y_j$ is Lipschitz continuous. Moreover, we may write

$$\begin{aligned}\left([\overline{F'(\tilde{y_n})}^\top]^2 - [\overline{F'(y_n)}]^2 \right) &= \left(\overline{F'(\tilde{y_n})}^\top - \overline{F'(y_n)}^\top \right) \overline{F'(\tilde{y_n})}^\top \\ &\quad + \overline{F'(y_n)} \left(\overline{F'(\tilde{y_n})}^\top - \overline{F'(y_n)}^\top \right).\end{aligned}$$

Arguing as above shows this term is $\mathbf{O}(\Delta t_n^2)$, and thus the contribution of the third term in (5.23) is $\mathbf{O}(\Delta t_n^4)$. The same conclusion about the second term in (5.23) follows using the assumption that $\frac{\partial^2 F_i}{\partial y_j \partial y_k}$ is Lipschitz continuous. This finally leads to the conclusion that $\omega(t_{n-1}) - \tilde{\omega}(t_{n-1}) = \mathbf{O}(\Delta t_n^3)$. ∎

Finally, Lemmas 5.5 and 5.6 involve $\tilde{y}$ and y^r, which we do not have in practice. We use the same kinds of arguments to replace them by the approximations $\tilde{Y}$ and Y^r at the cost of further higher order expressions. The result is contained in Theorem 5.7.

THEOREM 5.7 (computable a posteriori error estimate). *A computable error representation for the multiscale operator splitting dG finite element method is*

$$
\begin{aligned}
(y_N - \tilde{Y}_N, \psi_N) &= \sum_{n=1}^{N} (\tilde{Y}_{n-1}, (\overline{E}_1 + \overline{E}_2)\psi_n) \\
&- \sum_{n=1}^{N} \sum_{m=1}^{M_n} \left(\int_{I_{m,n}} (\dot{Y^r} - F(Y^r), \vartheta^r - \Pi\vartheta^r)\, dt + ([Y^r]_{m-1,n}, \vartheta^{r+}_{m-1,n} - \Pi\vartheta^{r+}_{m-1,n}) \right) \\
&- \sum_{n=1}^{N} \left(\int_{I_n} (\dot{Y^d} - AY^d, \vartheta^d - \Pi\vartheta^d)\, dt + ([Y^d]_{n-1}, \vartheta^{d+}_{n-1} - \Pi\vartheta^{d+}_{n-1}) \right) \\
&+ \mathbf{O}(\Delta t^{q_d+2}) + \mathbf{O}(\Delta t\, \Delta s^{q_r+1}),
\end{aligned}
$$

where

$$
\overline{E}_1 = \frac{1}{2}\Delta t_n \left(A^\top R(\tilde{Y}) - R(\tilde{Y})\, A^\top \right) \quad \textit{and} \quad \overline{E}_2 = \left(\Phi^r_n(\tilde{Y}) - \Phi^r_n(Y^r) \right) \Phi^d_n.
$$

We have finally obtained the hybrid a posteriori–a priori error estimate, where the leading order expressions are computable and the remainder is provably higher order. Recall that the term $\overline{E}_1$ is a matrix that consists of the diffusion matrix $A^\top$ and the Jacobian of the nonlinear reaction, $R(\tilde{Y}) = \int_{I_n} \overline{F'(\tilde{Y})} dt$. These are computable and do not require the solution of an adjoint problem. On the other hand, the term E_2 involves the computation of three adjoint solutions, namely, one adjoint solution of the diffusion component and the adjoint solutions of the reaction components linearized about $\tilde{Y}$ and Y^r, respectively.

5.5. Remark on the definition of the adjoint problem. The homogeneity assumption $F(y) = 0$ implies that $y = 0$ is a steady state solution of (1.1). We can also define adjoint problems using other steady state solutions or a given function of time. We illustrate by assuming that c is a vector such that

$$
\textit{General homogeneity assumption:} \quad F(c) = 0.
$$

We define the variable $z = y - c$ and set

$$
\overline{F'(z)} = \int_0^1 F'(sz + c)\, ds,
$$

so that $\overline{F'(z)}z = F(y) - F(c) = F(y)$. We can define a differential equation governing z,

$$
\dot{z} = A\, z(t) + F(y) + Ac = A\, z(t) + \overline{F'(z)}\, z(t) + Ac. \tag{5.24}
$$

The associated generalized Green's function φ satisfies the adjoint differential equation

$$
-\dot{\varphi} = A^\top \varphi(t) + \overline{F'(z)}^\top \varphi(t). \tag{5.25}
$$

Arguing as in section 5.1 yields

$$
\begin{aligned}
0 &= \int_{I_n} \left(-\dot{\varphi} - A^\top \varphi(t) - (\overline{f'}(z))^\top \varphi(t), z \right) dt \\
&= -\int_{I_n} \frac{d}{dt}(\varphi, z)\, dt + \int_{I_n} (\dot{z} - Az - \overline{F'(z)}z - Ac, \varphi) dt + \int_{I_n} (Ac, \varphi) dt.
\end{aligned} \tag{5.26}
$$

The second term on the right-hand side vanishes by (5.24), and thus integrating the first term yields

$$(z_n, \psi_n) = (z_{n-1}, \varphi_{n-1}) + \int_{I_n} (Ac, \varphi) dt. \tag{5.27}$$

Using the analogous approach for each component of the analytic operator splitting (2.2) and (2.3), we define, respectively, $z^r = y^r - c$ and $z^d = y^d - c$, and the associated differential equations

$$\dot{z^r} = F(y^r) = \overline{F'(z^r)} z^r, \quad \dot{z^d} = Az^d + Ac.$$

To these differential equations, we define the respective adjoint problems

$$-\dot{\varphi^r} = \overline{F'(z^r)}^\top \varphi^r(t), \quad -\dot{\varphi^d} = A^\top \varphi^d(t).$$

Again, arguing as in section 5.1, we obtain the following representations for the component solutions:

$$(z_n^{r-}, \psi_n^r) = (z_{n-1}^{r+}, \varphi_{n-1}^{r+}),$$
$$(z_n^{d-}, \psi_n^d) = (z_{n-1}^{d+}, \varphi_{n-1}^{d+}) + \int_{I_n} \left(Ac, \varphi^d\right) dt.$$

Next we can follow the analysis in section 5.2 to get an error representation analogous to Theorem 5.1.

THEOREM 5.8. *Let c be a vector such that $F(c) = 0$. Then the splitting error over a single diffusion time step can be represented by*

$$(y_n - \tilde{y}_n, \psi_n) = (\tilde{y}_{n-1} - c, \Delta\Phi_n \psi_n) + \int_{I_n} \left(Ac, \varphi - \varphi^d\right) dt$$

with $\Delta\Phi_n = \Phi_n(y) - \Phi_n^r(y^r)\Phi_n^d$.

We note that when $c = 0$, the original error representation in Theorem 5.1 is recovered. Also, we showed above that the first term involving $\Delta\Phi_n \psi_n$ can be cast into expressions that are computable. Similar analysis can be carried out for the last term involving $\varphi - \varphi^d$. To see this, we write

$$\varphi - \varphi^d = (\varphi - \omega) + (\omega - \tilde{\omega}) + (\tilde{\omega} - \varphi^d), \tag{5.28}$$

where ω and $\tilde{\omega}$ are the functions governed by differential equations (5.21) and (5.22). Using the expansion of φ and ω gives

$$\begin{aligned} \varphi(t) - \omega(t) &= (t_n - t)\, A^\top \psi_n - \Delta t_n\, A^\top \psi_n + \mathbf{O}\left((t_n - t)^2 + \Delta t_n^2\right) \\ &= (t_{n-1} - t)\, A^\top \psi_n + \mathbf{O}\left((t_n - t)^2 + \Delta t_n^2\right), \end{aligned} \tag{5.29}$$

from which we obtain

$$\int_{I_n} (Ac, \varphi - \omega)\, dt = -\frac{1}{2}\Delta t_n^2 \left(Ac, A^\top \psi_n\right) + \mathbf{O}(\Delta t_n^3). \tag{5.30}$$

Using a similar argument to that in the proof of Lemma 5.6, we can deduce that

$$\int_{I_n} (Ac, \omega - \tilde{\omega})\, dt = \mathbf{O}(\Delta t_n^3). \tag{5.31}$$

Finally, the function $\tilde{\omega}$ is actually computed in the error estimator and thus we may write

$$\int_{I_n} \left(Ac, \varphi - \varphi^d\right)\, dt = -\frac{1}{2}\Delta t_n^2 \left(Ac, A^\top \psi_n\right) + \int_{I_n} \left(Ac, \tilde{\omega} - \varphi^d\right)\, dt + \mathbf{O}(\Delta t_n^3). \tag{5.32}$$

We note that in this setting, the rest of the error estimator components remains as before.

6. Numerical examples. In this section, we present several numerical examples that show the performance of the error estimates. All of these problems originate as reaction-diffusion initial boundary value problems. Unless otherwise noted, we use a continuous, piecewise linear Galerkin finite element method with N_e elements to discretize space. All forward problems are solved using the lowest order, piecewise constant dG method, which is equivalent to backward Euler scheme, while the adjoint solutions are computed using a second order, piecewise linear, cG method, which is equivalent to the Crank–Nicolson scheme.

6.1. Assessment of the operator splitting exact error. To gain some insights into the splitting procedure, we solve a nonlinear initial boundary value problem described in (1.2) with the following specification:

$$f(x,u) = \frac{u^2}{\sin(\pi x)}, \quad u_d = 0, \quad u_0(x) = \sin(\pi x).$$

After spatial discretization, the vector of initial conditions g has entries $g_j = \sin(\pi x_j)$ $= \sin(j\pi/N_e)$, while using the trapezoidal rule to approximate the integrals involving the forcing terms gives

$$F_i(y) \approx \frac{1}{2}\left(0 + \frac{y_i^2}{\sin(\pi x_i)}\right) + \frac{1}{2}\left(\frac{y_i^2}{\sin(\pi x_i)} + 0\right) = \frac{y_i^2}{\sin(\pi x_i)}. \tag{6.1}$$

The eigenvalues of the matrix A are

$$\lambda_i = \frac{2k}{h^2} - \frac{2k}{h^2}\cos(i\pi/N_e), \tag{6.2}$$

with the associated eigenvectors V_i whose entries are

$$V_{ij} = \sin(ij\pi/N_e). \tag{6.3}$$

We note that $g = V_1$ and assume that the exact solution of (1.1) has entries

$$y(t) = \xi(t) V_1, \tag{6.4}$$

where $\xi(t)$ is a scalar function to be determined later and V_1 is the eigenvector of matrix D associated with the eigenvalue λ_1. Consequently, we have

$$F_i(y) = F_i(\xi V_1) = \frac{(\xi V_{1i})^2}{\sin(\pi x_i)} = \xi^2 V_{1i}, \tag{6.5}$$

using the fact that $V_{1i} = \sin(i\pi/N_e) = \sin(\pi x_i)$. Substitution of (6.5) and (6.4) into (1.1) yields the ordinary differential equation (noting that $AV_1 = \lambda_1 V_1$)

$$\dot{\xi}(t) + \lambda_1\, \xi(t) = \xi^2(t),$$

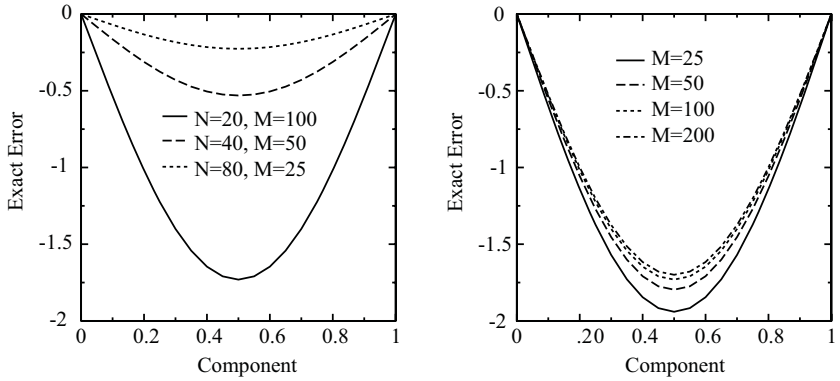


FIG. 6.1. *Comparison of exact errors at $T = 2$ against spatial location: reaction time step is kept constant (left), diffusion time step is kept constant (right).*

with $\xi_i(0) = 1$. This has the analytical solution

$$\xi(t) = \frac{\lambda_1}{1 - (1 - \lambda_1)e^{\lambda_1 t}}.$$

Figure 6.1 shows exact errors of the splitting procedure for this problem. The plot on the left shows the exact error at $T = 2$ plotted against the spatial location while the reaction time step is kept constant. This demonstrates that the error behaves like $\mathbf{O}(\Delta t)$. The plot on the right shows the exact error at $T = 2$ while the diffusion time step is kept constant. As the reaction time step is decreased, the error also decreases until it reaches a condition where the dominating error comes from the diffusion time step. After this, there is no improvement in accuracy. This result confirms the a priori analysis established in the previous section that the error in the operator splitting is dominated by the component that is of $\mathbf{O}(\Delta t)$.

6.2. Performance of the a posteriori error representation. The main purpose of the next four examples is to test the accuracy of the computable parts of the hybrid a posteriori–a priori error estimate where we drop the uncomputable higher order terms. Since we do not have true solutions for most of these problems, we approximate exact errors by computing the difference between the operator splitting solution with numerical solution of the fully coupled problems computed on a very fine mesh. We refer to this approximated error as the "error."

The first three examples fall into the category of stable diffusion interacting with unstable reaction. The reaction components in the three examples present a range of instability. The last example is radically different in that it represents competition between stable reaction and diffusion components.

6.2.1. A blow up problem. The first example is the "blow up" problem obtained from

$$\begin{cases} \dfrac{\partial u}{\partial t} - 0.05\,\dfrac{\partial^2 u}{\partial x^2} = u^2, & x \in (0,1), t > 0, \\ u(0,t) = u(1,t) = 0, & t > 0, \\ u(x,0) = 4x(1-x), & x \in (0,1). \end{cases}$$

TABLE 6.1
Operator splitting error estimate for the blow up problem at $T = 1$, reaction time step $= 10^{-3}$.

Δt	M	Exact error (%)	Effectivity
10^{-1}	100	11.07	1.0286
10^{-2}	10	1.35	1.0067
10^{-3}	1	0.45	1.0020

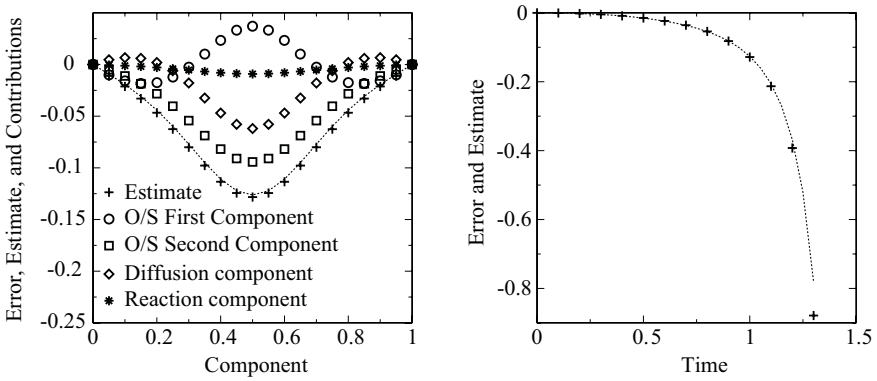


FIG. 6.2. *Blow up results: Comparison of errors against the spatial location (left), time history of errors at the midpoint location (right). The dotted line is the exact error and the + is the estimated error.*

As mentioned in section 2, the solution of the reaction component exhibits finite time blow up when undamped by the diffusion component. This is perhaps the most extreme form of instability. For this computation, we use 20 spatial finite elements. Table 6.1 shows the effectivity index of the error estimate computed at the final time $T = 1$. In this computation, we keep the reaction time step constant and varied the diffusion time step and number of reaction time steps.

Figure 6.2 shows a comparison of the errors computed using $\Delta t = 0.05$ and $M = 50$ reaction time steps. The left plot in Figure 6.2 compares the errors at $T = 1$ plotted against the spatial location. In this plot, we plot each of the expressions contributing to the overall estimate. In the plot on the right, the time history up to $T = 1.3$ of the error at the midpoint location is shown.

6.2.2. Chemical dynamics: The Brusselator problem. The Brusselator problem is represented by a coupled set of equations first introduced by Prigogine and Lefever [26] as a model of chemical dynamics:

$$
\begin{cases}
\dfrac{\partial u_1}{\partial t} - k_1 \dfrac{\partial^2 u_1}{\partial x^2} = \alpha - (\beta + 1)u_1 + u_1^2 u_2, & x \in (0,1), t > 0, \\
\dfrac{\partial u_2}{\partial t} - k_2 \dfrac{\partial^2 u_2}{\partial x^2} = \beta u_1 - u_1^2 u_2, & x \in (0,1), t > 0, \\
u_1(0,t) = u_1(1,t) = \alpha,\ u_2(0,t) = u_2(1,t) = \beta/\alpha, & t > 0, \\
u_1(x,0) = u_{1,0}(x),\ u_2(x,0) = u_{2,0}(x), & x \in (0,1),
\end{cases}
\tag{6.6}
$$

where u_1 and u_2 are the concentration of species 1 and 2, respectively. We use $\alpha = 2$, $\beta = 5.45$, $k_1 = 0.008$, $k_2 = 0.004$ and initial conditions $u_1(x,0) = \alpha + 0.1\sin(\pi x)$

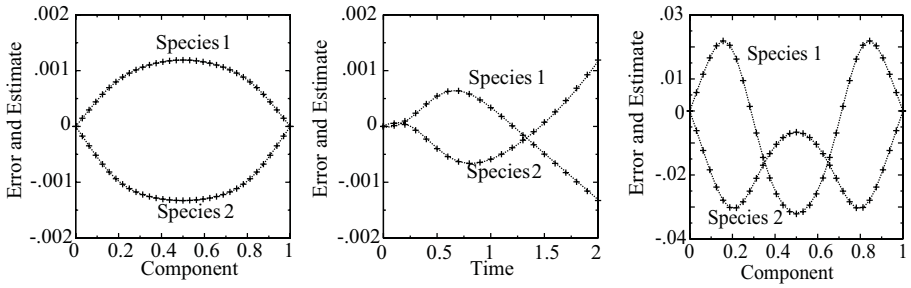


FIG. 6.3. *Brusselator results. Left: comparison of errors against the spatial location at $T = 2$. Middle: time history of errors at the midpoint location on $[0, 2]$. Right: comparison of errors against the spatial location at $T = 40$. The dotted line is the exact error and the + is the estimated error.*

and $u_2(x, 0) = \beta/\alpha + 0.1\sin(\pi x)$, which yields an oscillatory solution. In this case, the reaction is very mildly unstable, with at most a polynomial rate accumulation of perturbations as time passes. We use a 32 node spatial finite element discretization, resulting in an ordinary differential equation system with dimension 62. We note that in original form, the reaction terms do not satisfy the requirement $F(0) = 0$. Instead, we use a vector c as discussed in section 5.5 with $c_i = \alpha$ for $i = 1, \ldots, N_e - 1$ and $c_i = \beta/\alpha$ for $i = N_e, \ldots, 2N_e - 2$, so that $F(c) = 0$.

Figure 6.3 compares the errors computed using $\Delta t = 0.01$ and $M = 10$ reaction time steps to the hybrid a posteriori error estimates. We show results for $[0, 2]$ when the solution is still in a transient stage, and at $T = 40$ when the solution has become periodic. All the results show that the exact and estimated errors are in remarkable agreement.

6.2.3. Chaotic dynamics: The Lorenz problem. In this example, we treat the chaotic Lorenz equations as a reaction component of a reaction-diffusion problem:

$$
\begin{cases}
\dfrac{\partial u_1}{\partial t} - 0.5\,\dfrac{\partial^2 u_1}{\partial x^2} = 10(u_2 - u_1), & x \in (0,1), t > 0, \\
\dfrac{\partial u_2}{\partial t} - 0.5\,\dfrac{\partial^2 u_2}{\partial x^2} = u_1(28 - u_3) - u_2, & x \in (0,1), t > 0, \\
\dfrac{\partial u_3}{\partial t} - 0.5\,\dfrac{\partial^2 u_3}{\partial x^2} = u_1 u_2 - \dfrac{8}{3}u_3, & x \in (0,1), t > 0, \\
u_i(0,t) = u_i(1,t) = 0, & t > 0, i = 1,2,3, \\
u_i(x,0) = .1\sin(\pi x), & x \in (0,1), i = 1,2,3.
\end{cases}
\tag{6.7}
$$

The chaotic nature of the Lorenz equations implies that errors accumulate locally at an exponential rate in an average sense, placing this instability between that of the blow up and the Brusselator problems. Here, we use 20 spatial finite elements, resulting in a system of ordinary differential equations with dimension 57. Figure 6.4 shows typical error estimates for the three species. Across the top row, we compare estimates and errors at the final time $T = 2$ using $\Delta t = 5 \times 10^{-3}$ and $M = 10$ reaction time steps. We show the time history at the midpoint across the bottom row. Again the figures show good agreement between exact and estimated errors.

We illustrate the accuracy at a longer time of $T = 16$ in Figure 6.5.

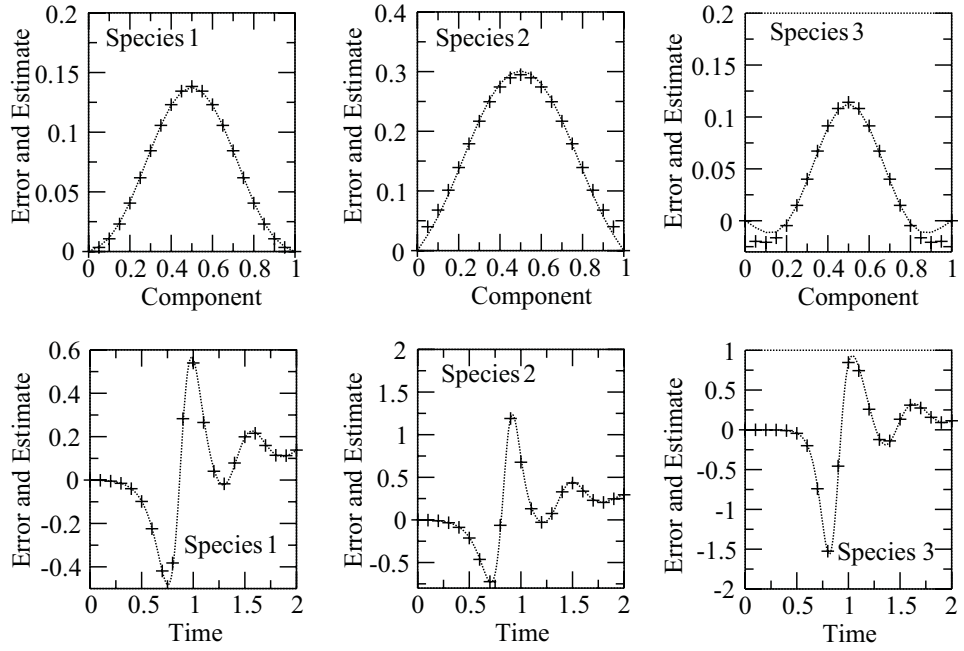


FIG. 6.4. *Lorenz results. Top row: comparison of errors against the spatial location. Bottom row: time history of errors at the midpoint location. The dotted line is the exact error and the + is the estimated error.*

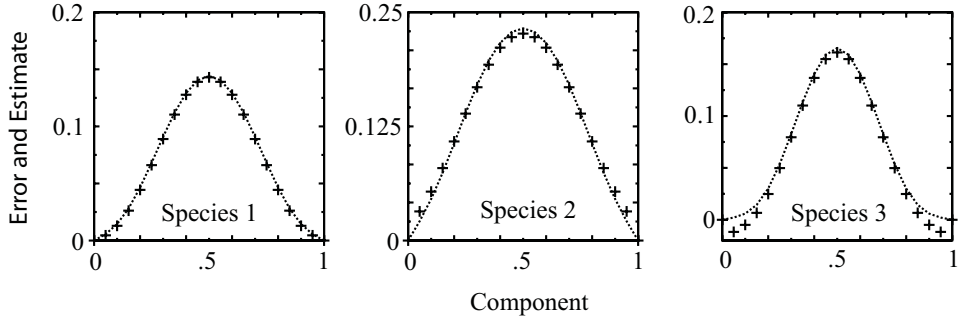


FIG. 6.5. *Accuracy of the error estimate for the Lorenz problem at time $T = 16$. The dotted line is the exact error and the + is the estimated error.*

6.2.4. The bistable problem. The bistable problem,

$$\begin{cases} \dfrac{\partial u}{\partial t} - \epsilon \dfrac{\partial^2 u}{\partial x^2} = u - u^3, & x \in (0,1), t > 0, \\ u_x(0,t) = u_x(1,t) = 0, & t > 0, \\ u(x,0) = u_0(x), & x \in (0,1), \end{cases} \tag{6.8}$$

is a well-studied example of nonlinear relaxation to equilibrium in the presence of competing stable steady states [3, 19, 1, 12, 17]. The stable steady states are $u = 1$ and $u = -1$, while $u = 0$ is an unstable steady state. Here, both the diffusion and the

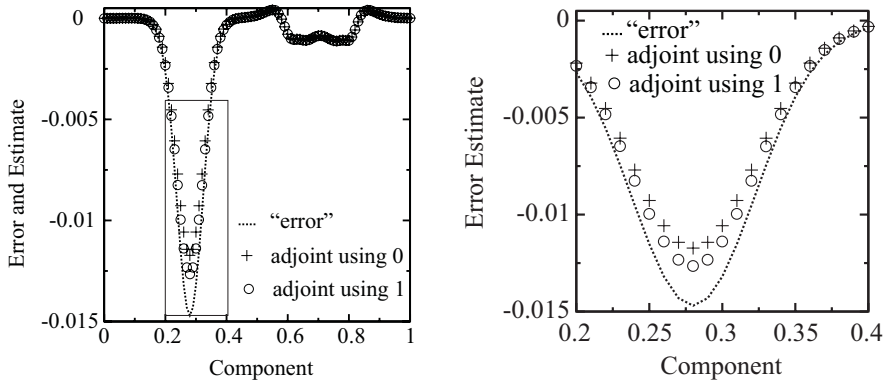


FIG. 6.6. *Bistable problem error estimate at $T = 41$; the right plot is the zoomed view of the boxed portion on the left plot.*

reaction are stabilizing, while instability arises because of the competition between these two components. Nonequilibrium solutions are characterized by long periods of "metastability" during which the solution is nearly stationary over periods of time on the order of $\exp(1/\epsilon)$ that are punctuated by rapid transients. The spatial profile of a metastable solution consists of narrow layers between regions where the solution has values of 1 and -1. We emphasize that the interaction between reaction and diffusion is very delicate in a metastable solution, and it is by no means clear that operator splitting is a reasonable approach.

For the computation we use 100 spatial finite elements. We compute the solution until final time level $T = 41$. Figure 6.6 shows comparison of the error. Here, we use $c = 0$ and $c = 1$ as the linearization points. The figure shows an observable improvement when computing the error using $c = 1$ than when using $c = 0$. Note that the solution eventually converges to 1.

7. Details of the a priori analysis. The following section contains the sequence of lemmas required for the proof of Theorem 4.4.

Lemma 7.1 determines the error in the numerical solution of the reaction component arising from errors in the initial conditions.

LEMMA 7.1. *Let $Z \in \mathcal{V}^{(q_r)}(I_n)$ satisfy*

$$\int_{I_{m,n}} (\dot{Z}, W)\, dt - \int_{I_{m,n}} (F(Z), W)\, dt + (Z^+_{m-1,n}, W^+_{m-1}) = (Z^-_{m-1,n}, W^+_{m-1}) \tag{7.1}$$

for all $W \in \mathcal{P}^{(q_r)}(I_{m,n})$, $m = 1, 2, \ldots, M_n$, and $Z^-_{0,n} = y^{r+}_{n-1}$. With $\xi = Z - Y^r$,

$$|\xi^-_{M_n,n}|^2 \le \exp(16L\Delta t_n)\, |\tilde{y}_{n-1} - \tilde{Y}_{n-1}|^2.$$

Proof. By construction, $\xi \in \mathcal{V}^{(q_r)}(I_n)$. It is obvious that for $q_r = 0, 1$,

$$\int_{I_{m,n}} |\xi|^2\, dt \le 2\Delta s_n\, |\xi^-_{m,n}|^2 + \frac{2}{3}\, \Delta s_n^3\, |\dot{\xi}|^2_{I_{m,n}}. \tag{7.2}$$

Subtracting (3.3) from (7.1) yields

$$\int_{I_{m,n}} (\dot{\xi}, W)\, dt - \int_{I_{m,n}} (F(Z) - F(Y^r), W)\, dt + (\xi^+_{m-1,n}, W^+_{m-1}) = (\xi^-_{m-1,n}, W^+_{m-1}) \tag{7.3}$$

for every $W \in \mathcal{P}^{(q_r)}(I_{m,n})$. Setting $W = (t - s_{m-1,n})\dot{\xi}$ in (7.3) and estimating yields

$$\frac{1}{2}\Delta s_n^2 |\dot{\xi}|^2_{I_{m,n}} \leq \frac{\epsilon^2}{2} L^2 \Delta s_n \int_{I_{m,n}} |\xi|^2 \, dt + \frac{1}{2\epsilon^2} \Delta s_n^2 \, |\dot{\xi}|^2_{I_{m,n}}$$

for some $\epsilon > 0$. Thus,

$$\Delta s_n^2 |\dot{\xi}|^2_{I_{m,n}} \leq C_\epsilon \, L^2 \Delta s_n \int_{I_{m,n}} |\xi|^2 \, dt, \tag{7.4}$$

where $C_\epsilon = \epsilon^4/(\epsilon^2 - 1)$. Substitution of (7.4) into (7.2) gives

$$\left(1 - \frac{2}{3} C_\epsilon \Delta s_n^2 L^2\right) \int_{I_{m,n}} |\xi|^2 \, dt \leq 2\Delta s_n \, |\xi^-_{m,n}|^2,$$

from which we obtain

$$\int_{I_{m,n}} |\xi|^2 \, dt \leq 4\Delta s_n \, |\xi^-_{m,n}|^2, \tag{7.5}$$

provided $1 - \frac{2}{3} C_\epsilon \Delta s_n^2 L^2 > 1/2$. Next, we choose $W = \xi$ in (7.3) and use the Lipschitz continuity of F to get

$$\frac{1}{2}|\xi^-_{m,n}|^2 + \frac{1}{2}|\xi^+_{m-1,n}|^2 \leq L \int_{I_{m,n}} |\xi|^2 \, dt + |\xi^-_{m-1,n}| \, |\xi^+_{m-1,n}|.$$

This gives

$$|\xi^-_{m,n}|^2 \leq |\xi^-_{m-1,n}|^2 + 2L \int_{I_{m,n}} |\xi|^2 \, dt.$$

Using (7.5), this implies

$$|\xi^-_{m,n}|^2 \leq \frac{1}{1 - 8\Delta s_n L} \, |\xi^-_{m-1,n}|^2 \leq \exp(16\Delta s_n L) \, |\xi^-_{m-1,n}|^2,$$

provided $1 - 8\Delta s_n L > 1/2$. Applying this recursive relation M_n times yields the desired result. ∎

Lemma 7.2 determines the error in the numerical solution of the reaction component due to discretization and the inherited initial conditions.

LEMMA 7.2. *Let y^r be the solution of analytical reaction component and let Y^r be the dG numerical solution approximating y^r. For $n = 1, \ldots, N$,*

$$|y^{r-}_n - Y^{r-}_{M,n}| \leq C \, \Delta t_n \, \Delta s_n^{q_r+1} + \exp(8L\Delta t_n) \, |\tilde{y}_{n-1} - \tilde{Y}_{n-1}|.$$

Proof. We write

$$y^{r-}_n - Y^{r-}_{M,n} = (y^{r-}_n - Z^-_{M,n}) + (Z^-_{M,n} - Y^{r-}_{M,n}),$$

where $Z \in \mathcal{V}^{(q_r)}(I_n)$ is defined as in Lemma 7.1. Moreover, it has been established (cf. Delfour and Dubean [6], Johnson [22], and Estep [13]) that

$$|y^{r-}_n - Z^-_{M,n}| \leq C \, \Delta t_n \, \Delta s_n^{q_r+1}.$$

Using Lemma 7.1, the proof is complete. ∎

We also use the following result from Delfour, Hager, and Trochu [7].

LEMMA 7.3. *Let $w \in H^1(I_n)$ satisfying $\dot{w} + A^\top w = 0$ in I_n. Then there exists a constant C independent of $s \in I_n$, such that $\|w\|_{r+1,n} \le C|w(s)|$, where the norm $\|\cdot\|_{r,n}$ is associated with the Sobolev space*

$$H^r(I_n) = \left\{ w : \|w\|_{r,n}^2 = \sum_{i=0}^{r} \left\| \frac{d^i w}{dt^i} \right\|_{L^2(I_n)}^2 < \infty \right\}.$$

Lemma 7.4 bounds the residual to a linear constant coefficient differential equation when its exact solution is projected into a piecewise polynomial space.

LEMMA 7.4. *Let $\Pi : L^2(I_n) \to \mathcal{P}^{(q)}(I_n)$ denote an interpolation operator. Let w be as defined in Lemma 7.3, and $\Pi w \in \mathcal{P}^q(I_n)$, $q \ge 1$, its interpolation function. Then there is a positive constant C such that for all $s \in I_n$,*

$$\left\| \frac{d}{dt}\Pi w + A^\top \Pi w \right\|_{L^2(I_n)} \le C\, \Delta t_n^q\, |w(s)|.$$

Proof. Since w in Lemma 7.3 is homogeneous, the interpolation property (cf. Strang and Fix [33] and Ciarlet [4]) and Lemma 7.3 imply that for $t_{n-1} \le s \le t_n$,

$$\begin{aligned} \left\| \frac{d}{dt}\Pi w + A^\top \Pi w \right\| &= \left\| \frac{d}{dt}(\Pi w - w) + A^\top (\Pi w - w) \right\| \\ &\le C\, \|w - \Pi w\|_{1,n} \le C\, \Delta t_n^q\, \|w\|_{q+1,n} \le C\, \Delta t_n^q\, |w(s)|. \end{aligned}$$ ∎

Lemma 7.5 bounds the error in the numerical solution to a constant coefficient differential equation.

LEMMA 7.5. *Let $X \in \mathcal{P}^{q_d}(I_n)$ satisfy*

$$\int_{I_n} (\dot{X}, V)\, dt - \int_{I_n} (AX, V)\, dt + (X_{n-1}^+, V_{n-1}^+) = (y_n^{r-}, V_{n-1}^+) \tag{7.6}$$

for all $V \in \mathcal{P}^{q_d}(I_n)$, where $X_{n-1}^- = y^{r-}{}_n$. With $\theta = y^d - X$, $|\theta_n^-| \le C\, \Delta t_n^{q_d+2}$ for $q_d = 0, 1$.

Proof. Obviously y^d defined in (2.3) satisfies (7.6). Let $U \in \mathcal{P}^{q_d}(I_n)$ satisfy

$$\int_{I_n} (\dot{U}, V)\, dt - \int_{I_n} (Ay^d, V)\, dt + (U_{n-1}^+, V_{n-1}^+) = (y_n^{r-}, V_{n-1}^+) \quad \text{for all } V \in \mathcal{P}^{q_d}(I_n). \tag{7.7}$$

For sufficiently small Δt_n, the following estimate holds (see [13]):

$$|\theta|_{I_n} \le |y^d - U|_{I_n} + |U - X|_{I_n} \le C\Delta t_n^{q_d+1} |y^{(q_d+1)}|_{I_n} + C\Delta t_n^{q_d+3/2} |y^{(q_d+1)}|_{I_n}. \tag{7.8}$$

Subtracting (7.6) from the variational formulation of y^d gives

$$\int_{I_n} (\dot{\theta}, V)\, dt - \int_{I_n} (A\theta, V)\, dt + (\theta_{n-1}^+, V_{n-1}^+) = 0. \tag{7.9}$$

Integration by parts gives

$$(\theta_n^-, V_n^-) = \int_{I_n} (\theta, \dot{V} + A^\top V)\, dt. \tag{7.10}$$

For $q_d = 0$, we choose $V = V_n^- = \theta_n^-$ in (7.9) and use (7.8) to get $|\theta_n^-| \leq C\Delta t_n^2$. For $q_d = 1$, we choose $V = \Pi w$ in (7.9), with w defined as in Lemma 7.3, and $w(t_n) = \theta_n^-$. Then using (7.8) and Lemma 7.4,

$$|\theta_n^-|^2 \leq \int_{I_n} |\theta|\, |\dot{\Pi w} + A^\top \Pi w|\, dt \leq C\Delta t_n^3 \sqrt{\Delta t_n}\, |\theta_n^-|, \tag{7.11}$$

which gives the desired estimate. □

8. Conclusion. In this paper, we derive both an a priori convergence estimate and a hybrid a posteriori–a priori estimate for the multiscale operator splitting discontinuous Galerkin finite element method for a system of ordinary differential equations. The system of equations is assumed to have a form that typically arises from space discretization of a reaction-diffusion problem. The a priori analysis uses the fact that an operator splitting approximation is a consistent approximation of an analytic operator splitting problem. The a posteriori analysis takes into account the fact that the original problem and an analytic operator split version are associated with different adjoint problems. The differences in the adjoints provides the means to accurately understand the effects of operator splitting on the stability properties of a quantity of interest computed from the solution. To produce a computable a posteriori estimate of the error, we manipulate the original a posteriori error representation to obtain a computable expression plus uncomputable quantities that are provably higher order in an asymptotic sense. We conclude with some examples that demonstrate the accuracy of the computable parts of the hybrid a posteriori–a priori estimate.

Our results can be extended in a straightforward way to differential equations with a more general nonlinear form, higher order splitting schemes, and different time finite element methods. The results also extend formally, and rigorously for the most part, to operator splitting finite element schemes for reaction-diffusion problems where the splitting occurs at the partial differential equation level. We will pursue the complete extension to the latter in future work.

The general approach to analyzing operator splitting can be used to analyze operator decomposition in a wide variety of settings; see, e.g., [2, 18].

REFERENCES

[1] L. Bronsard and R. Kohn, *On the slowness of phase-boundary motion in one space dimension*, Comm. Pure Appl. Math., 43 (1990), pp. 983–997.

[2] V. Carey, D. Estep, and S. Tavener, *A posteriori analysis and adaptive error control for multiscale operator decomposition solution of elliptic systems* I: *Triangular systems*, SIAM J. Numer. Anal., submitted (2006).

[3] J. Carr and R. Pego, *Metastable patterns in $u_t = \epsilon^2 \Delta u - f(u)$*, Comm. Pure Appl. Math., 42 (1989), pp. 523–576.

[4] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, Stud. Math. Appl. 4, North–Holland, Amsterdam, 1978.

[5] C. N. Dawson and M. F. Wheeler, *Time-splitting methods for advection-diffusion-reaction equations arising in contaminant transport*, in ICIAM 91 (Washington, DC, 1991), SIAM, Philadelphia, 1992, pp. 71–82.

[6] M. C. Delfour and F. Dubeau, *Discontinuous polynomial approximations in the theory of one-step, hybrid and multistep methods for nonlinear ordinary differential equations*, Math. Comp., 47 (1986), pp. 169–189; supplement, pp. S1–S8.

[7] M. Delfour, W. Hager, and F. Trochu, *Discontinuous Galerkin methods for ordinary differential equations*, Math. Comp., 36 (1981), pp. 455–473.

[8] S. Descombes, *Convergence of a splitting method of high order for reaction-diffusion systems*, Math. Comp., 70 (2001), pp. 1481–1501.

[9] S. Eastman, *Analysis and Application of the Nonlinear Power Method*, Ph.D. thesis, Colorado State University, Fort Collins, CO, 2005.
[10] K. Eriksson, D. Estep, P. Hansbo, and C. Johnson, *Introduction to adaptive methods for differential equations*, in Acta Numerica, 1995, Cambridge University Press, Cambridge, UK, 1995, pp. 105–158.
[11] K. Eriksson, D. Estep, P. Hansbo, and C. Johnson, *Computational Differential Equations*, Cambridge University Press, Cambridge, UK, 1996.
[12] D. Estep, *An analysis of numerical approximations of metastable solutions of the bistable equation*, Nonlinearity, 7 (1994), pp. 1445–1462.
[13] D. Estep, *A posteriori error bounds and global error control for approximation of ordinary differential equations*, SIAM J. Numer. Anal., 32 (1995), pp. 1–48.
[14] D. Estep, *A Short Course on Duality, Adjoint Operators, Green's Functions, and A Posteriori Error Analysis*, Sandia National Laboratories, Albuquerque, NM, 2004. Notes can be downloaded from http://math.colostate.edu/~estep.
[15] D. Estep and D. French, *Global error control for the continuous Galerkin finite element method for ordinary differential equations*, RAIRO Modél. Math. Anal. Numér., 28 (1994), pp. 815–852.
[16] D. Estep and A. M. Stuart, *The dynamical behavior of the discontinuous Galerkin method and related difference schemes*, Math. Comp., 71 (2002), pp. 1075–1103.
[17] D. Estep, M. G. Larson, and R. D. Williams, *Estimating the error of numerical solutions of systems of reaction-diffusion equations*, Mem. Amer. Math. Soc., 146 (2000).
[18] D. Estep, S. Tavener, and T. Wildey, *A posteriori analysis and improved accuracy for an operator decomposition solution of a conjugate heat transfer problem*, SIAM J. Numer. Anal., to appear.
[19] G. Fusco and J. Hale, *Slow-motion manifolds, dormant instability, and singular perturbations*, J. Dynam. Differential Equations, 1 (1989), pp. 75–94.
[20] W. Hundsdorfer and J. Verwer, *Numerical Solution of Time-Dependent Advection-Diffusion-Reaction Equations*, Springer Ser. Comput. Math. 33, Springer-Verlag, Berlin, 2003.
[21] W. H. Hundsdorfer and J. G. Verwer, *Stability and convergence of the Peaceman-Rachford ADI method for initial-boundary value problems*, Math. Comp., 53 (1989), pp. 81–101.
[22] C. Johnson, *Error estimates and adaptive time-step control for a class of one-step methods for stiff ordinary differential equations*, SIAM J. Numer. Anal., 25 (1988), pp. 908–926.
[23] G. I. Marchuk, *On the theory of the splitting-up method*, in Proceedings of the Second Symposium on Numerical Solution of Partial Differential Equations, SVNSPADE, Academic Press, New York, 1970, pp. 469–500.
[24] G. I. Marchuk, V. I. Agoshkov, and V. P. Shutyaev, *Adjoint Equations and Perturbation Algorithms in Nonlinear Problems*, CRC Press, Boca Raton, FL, 1996.
[25] C. C. Ober and J. N. Shadid, *Studies on the accuracy of time-integration methods for the radiation-diffusion equations*, J. Comput. Phys., 195 (2004), pp. 743–772.
[26] I. Prigogine and R. Lefever, *Symmetry breaking instabilities in dissipative systems*, J. Chem. Phys., 48 (1968), pp. 1695–1700.
[27] D. L. Ropp and J. N. Shadid, *Stability of operator splitting methods for systems with indefinite operators: Reaction-diffusion systems*, J. Comput. Phys., 203 (2005), pp. 449–466.
[28] D. L. Ropp, J. N. Shadid, and C. C. Ober, *Studies of the accuracy of time integration methods for reaction-diffusion equations*, J. Comput. Phys., 194 (2004), pp. 544–574.
[29] J. Sandelin, *Global Estimate and Control of Model, Numerical, and Parameter Error*, Ph.D. thesis, Department of Mathematics, Colorado State University, Fort Collins, CO, 2006.
[30] J. A. Smoller, *Shock Waves and Reaction-Diffusion Equations*, 2nd ed., Springer-Verlag, New York, 1983.
[31] B. Sportisse, *An analysis of operator splitting techniques in the stiff case*, J. Comput. Phys., 161 (2000), pp. 140–168.
[32] G. Strang, *On the construction and comparison of difference schemes*, SIAM J. Numer. Anal., 5 (1968), pp. 506–517.
[33] G. Strang and G. J. Fix, *An Analysis of the Finite Element Method*, Prentice–Hall Ser. Automat. Comput., Prentice–Hall, Englewood Cliffs, NJ, 1973.
[34] J. G. Verwer and B. Sportisse, *A Note on Operator Splitting Analysis in the Stiff Linear Case*, CWI report MAS-R9801, Amsterdam, The Netherlands, 1998.

SIAM J. NUMER. ANAL.
Vol. 46, No. 3, pp. 1147–1165
© 2008 Society for Industrial and Applied Mathematics

HIGH FREQUENCY INDUCED INSTABILITY IN NYSTRÖM METHODS FOR THE VAN DER POL EQUATION*

S. W. SCHOOMBIE† AND E. MARÉ‡

Abstract. In this paper several Nyström methods for the van der Pol equation are considered. In an earlier study by Cai, Aoyagi, and Abe it was shown that the second order Nyström, or leapfrog, method fails to approximate the limit cycle of the van der Pol equation, exhibiting a periodic modulation of the amplitude and sporadic high frequency noise instead. Cai et al. did a linear analysis and concluded that the spurious behavior was due to the interaction of the main part of the solution with a high frequency computational mode. In this paper we also apply a third and fourth order Nyström method to the van der Pol equation. Numerical experiments show that in these cases the high frequency mode causes blowup after some time. The onset of the instability can be delayed by decreasing the time step. We also improve on their analysis of the second order scheme by doing a nonlinear analysis, to wit a discrete multiple scales analysis. By this means we are able to explain the spurious behavior of this system completely.

Key words. van der Pol equation, spurious solutions, Nyström, leapfrog, multiple scales, discrete approximation, difference equations

AMS subject classifications. 74H15, 74H55, 93C70, 65L07, 65P40

DOI. 10.1137/050644148

1. Introduction.

The van der Pol equation

$$\frac{d^2x}{dt^2} - \epsilon(1-x^2)\frac{dx}{dt} + \omega^2 x = 0, \tag{1.1}$$

with

$$0 < \epsilon \ll 1, \tag{1.2}$$

was first formulated by van der Pol [12] in 1922 in the context of electronic circuits in vacuum tubes.

This equation is well known to have an unstable equilibrium at the origin of its phase plane, surrounded by a stable limit cycle. It is a standard classroom exercise to find an approximate analytic expression for the solution of the equation, using a multiple scales method [3, 6], namely,

$$x = \frac{2\cos(\omega t + \phi)}{\sqrt{1 + (4/a_0^2 - 1)e^{-\epsilon t}}} + O(\epsilon), \tag{1.3}$$

where a_0 is the initial amplitude.

Thus the limit cycle has a period which is very near to $2\pi/\omega$ and an amplitude which is very near to 2. To find the limit cycle more precisely, the van der Pol equation needs to be solved numerically. Various approaches have been used for this purpose.

*Received by the editors November 3, 2005; accepted for publication (in revised form) November 16, 2007; published electronically March 7, 2008.
http://www.siam.org/journals/sinum/46-3/64414.html
†Department of Mathematics and Applied Mathematics, University of the Free State, P.O. Box 339, Bloemfontein 9300, Republic of South Africa (schoomsw.sci@ufs.ac.za).
‡Department Mathematics and Applied Mathematics, University of Pretoria, Pretoria 0002, Republic of South Africa (eben.mare@up.ac.za). This work was carried out while this author was affiliated with the University of the Free State.

These include extrapolation-based methods with step size control [2], Runge–Kutta Nyström methods [2], the implicit Adams method of order up to twelve, and Gear's method for stiff equations [4], among others.

That care must be taken when designing a high order finite difference method for the van der Pol equation was shown by Cai, Aoyagi, and Abe [1]. They considered a second order Nyström method which showed a peculiar type of spurious behavior. Its solution periodically approaches the correct limit cycle, but then moves away from it again. They essentially used a linear analysis to explain this phenomenon, ascribing the spurious behavior to a spurious mode which interferes with the numerical solution.

We investigated this phenomenon further by also applying third and fourth order Nyström methods to the van der Pol equation. In these cases we found that the solution behaved reasonably well for some time, after which there was rapid growth of a high frequency mode, followed by blowup. By decreasing the time step the onset of this nonlinear instability may be delayed.

To analyze this type of instability, we considered only the simplest case, namely, the second order Nyström, or leapfrog, method used by Cai et al. [1]. We felt that their linear analysis provides only part of the explanation, and that a nonlinear analysis would give a more complete picture. This paper attempts to do just that. We use a discrete version of the multiple scales technique mentioned above to analyze the difference equation, and finally show that the full explanation is to be found in the phase plane.

In section 2 we will briefly review the difference scheme considered by Cai et al. [1] and demonstrate the nature of its spurious behavior, as well as show the behavior of the third and fourth order Nyström schemes.

In sections 3 and 4 we then proceed with a nonlinear analysis to explain the observed phenomena in the case of the second order scheme. The results of the nonlinear analysis is then related to some numerical results in section 5. Section 6 contains some conclusive remarks.

2. Nyström time discretizations. Following [1], we rewrite (1.1) as a system of two first order equations, namely,

$$\frac{dx}{dt} = y, \tag{2.1}$$

$$\frac{dy}{dt} = \epsilon(1 - x^2)y - \omega^2 x. \tag{2.2}$$

To discretize these equations, we introduce the time step τ, the finite difference approximations $x^n \approx x(n\tau)$ and $y^n \approx y(n\tau)$, and the shift operators E and E^{-1}, defined for any sequence x^n by

$$Ex^n = x^{n+1} \quad \text{and} \quad E^{-1}x^n = x^{n-1}. \tag{2.3}$$

Finally, we define the central divided difference operator

$$\delta = (E - E^{-1})/(2\tau). \tag{2.4}$$

2.1. Second order Nyström discretization (leapfrog). Following Cai et al. [1], we first discretize the system (2.1), (2.2) by means of a second order Nyström method:

$$\delta x^n = y^n, \tag{2.5}$$

$$\delta y^n = \epsilon(1 - (x^n)^2)y^n - \omega^2 x^n. \tag{2.6}$$

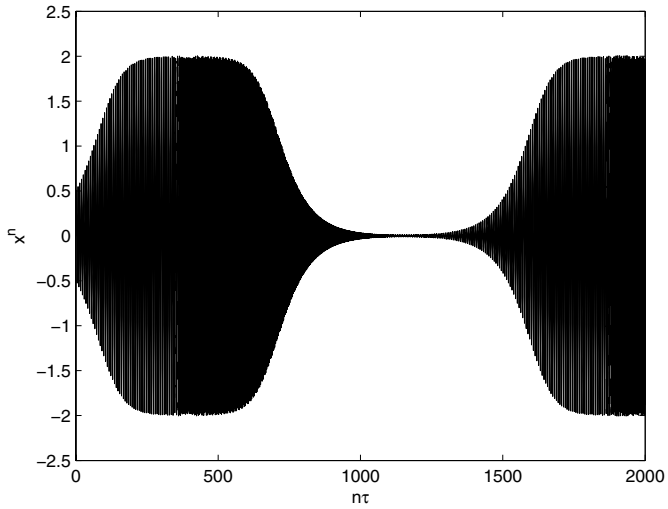


FIG. 1. *Results when running the difference system* (2.5), (2.6) *with* $\epsilon = 0.025$, $\omega = 1$, $\tau = 0.2$, *and* $x^0 = 0$, $y^0 = 0.5$.

This discretization is often referred to as a leapfrog method.

The system of difference equations (2.5), (2.6) can easily be seen to be equivalent to a single fourth order difference equation,

$$(E^2 - 2 + E^{-2})x^n/(4\tau^2) - \epsilon(1 - (x^n)^2)(E - E^{-1})x^n/(2\tau) + \omega^2 x^n = 0. \tag{2.7}$$

Using the following notation for the forward and backward divided difference operators, respectively,

$$\Delta = (E - 1)/\tau, \quad \nabla = (1 - E^{-1})/\tau, \tag{2.8}$$

we can rewrite (2.7) in the more concise form

$$\left[\Delta\nabla + \frac{\tau^2}{4}\Delta^2\nabla^2\right] x^n - \epsilon(1 - (x^n)^2)\delta x^n + \omega^2 x^n = 0. \tag{2.9}$$

To compute a numerical solution from (2.5) and (2.6), we need to specify initial values x^0 and y^0. Values for x^1 and y^1 are also required, and these are calculated by means of a second order Runge–Kutta starter with time increment τ. Figure 1 shows the result of such a computation with $\epsilon = 0.025$, $\omega = 1$, $\tau = 0.2$, and $x^0 = 0$, $y^0 = 0.5$. Initially everything seems to go well, and the solution approaches the limit cycle with its amplitude at around 2 and remains with it for a while. However, after some time it leaves the limit cycle and eventually returns to it again. What seems to be a periodic amplitude modulation is imposed, which is quite spurious.

Cai et al. [1] showed that this spurious behavior is due to the presence of a high frequency computational mode. They performed a linear stability analysis and showed that the computational mode starts growing when the amplitude of the numerical solution reaches a value of about $\sqrt{0.8}$.

The onset of the growth of the high frequency mode can be delayed by decreasing the time step, as shown in Figure 2, where the time step was decreased to $\tau = 0.1$.

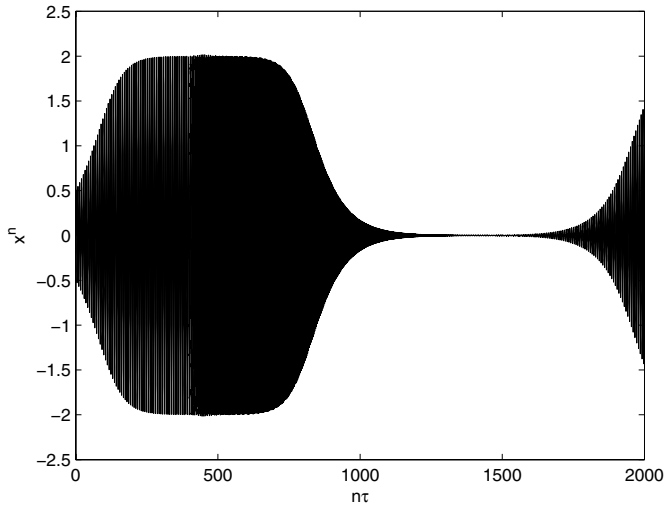


FIG. 2. *Results when running the difference system* (2.5), (2.6) *with* $\epsilon = 0.025$, $\omega = 1$, $\tau = 0.1$, *and* $x^0 = 0$, $y^0 = 0.5$.

2.2. Third order Nyström time discretization. The order of convergence of the numerical method can be increased by using a higher order Nyström discretization. For a third order scheme the discretization of the system becomes

$$\delta x^n = \frac{1}{6}[7y^n - 2y^{n-1} + y^{n-2}], \tag{2.10}$$

$$\delta y^n = \frac{1}{6}[7f(x^n, y^n) - 2f(x^{n-1}, y^{n-1}) + f(x^{n-2}, y^{n-2})], \tag{2.11}$$

where

$$f(x, y) = \epsilon(1 - x^2)y - \omega^2 x. \tag{2.12}$$

This time we need to supply two extra initial values, namely, for both $n = 1$ and $n = 2$. Because of the increased order of this method we used a fourth order Runge–Kutta method to calculate these initial values.

Using the same parameters as in Figure 1, we find that the numerical solution does not stay finite though spuriously modulated as with the leapfrog discretization, but that it suddenly blows up. The reason for this instability is the same, however: a sudden growth of a high frequency mode. The result is shown in Figure 3.

Since this is a convergent method, like the leapfrog method above, the onset of the instability can again be delayed by using a smaller time step. In order to reach the limit cycle before the high frequency mode grows significantly, we found that a time step of at most $\tau = 0.05$ has to be used.

Thus, although a higher order method is used, a much smaller time step is required in order to find the limit cycle numerically. The high frequency induced (HFI) instability observed here enforces an upper limit to the time step.

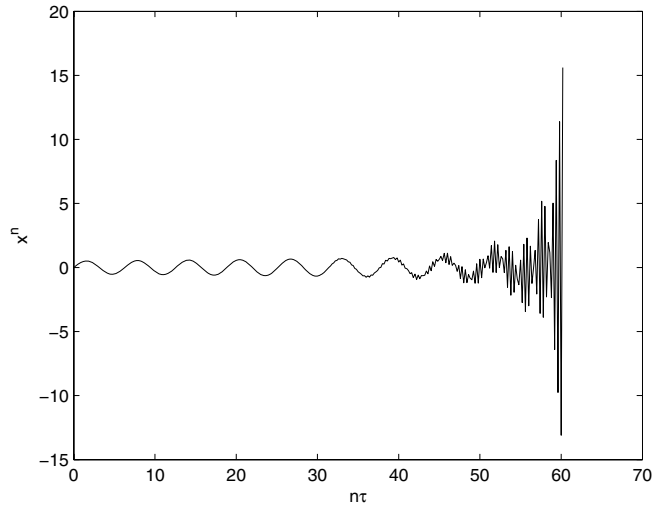


FIG. 3. *Results when running the difference system* (2.10), (2.11) *with* $\epsilon = 0.025$, $\omega = 1$, $\tau = 0.2$, *and* $x^0 = 0$, $y^0 = 0.5$.

2.3. Fourth order Nyström time discretization. Applying the fourth order Nyström discretization to the system (2.1), (2.2), we get

$$\delta x^n = \frac{1}{6}[8y^n - 5y^{n-1} + 4y^{n-2} - y^{n-3}], \tag{2.13}$$

$$\delta y^n = \frac{1}{6}[8f(x^n, y^n) - 5f(x^{n-1}, y^{n-1}) + 4f(x^{n-2}, y^{n-2}) - f(x^{n-3}, y^{n-3})] \tag{2.14}$$

with f as in (2.12).

The three extra initial values are again computed with a fourth order Runge–Kutta method. The results are similar to those of the third order method, except that the high frequency mode now appears much sooner, again leading to blowup. Figure 4 shows what happens when $\tau = 0.05$, and it turns out that the limit cycle can be reached fairly accurately only for time steps lower than about $\tau = 0.01$. Thus the upper bound for the time step is even lower than before.

Having observed the havoc caused by the high frequency spurious mode, the next thing to be attempted is an analysis of this phenomenon. This is done in the next section for the leapfrog case.

3. A discrete multiple scales analysis. Since the behavior of the leapfrog time discretization is the most complicated (the others simply blow up), and since this is also the simplest difference scheme, we shall attempt an analysis of this case only.

Since the spurious behavior of (2.9) consists of an amplitude modulation, a multiple scales analysis would seem to be an appropriate tool to investigate it.

To perform a multiple scales analysis of (2.9) we shall work in direct analogy with the well-known continuous multiple scales analysis performed for the van der Pol equation, e.g., [6]. We shall use two discrete time scales, namely,

$$T_p = \epsilon^p n\tau, \quad p = 0, 1, \tag{3.1}$$

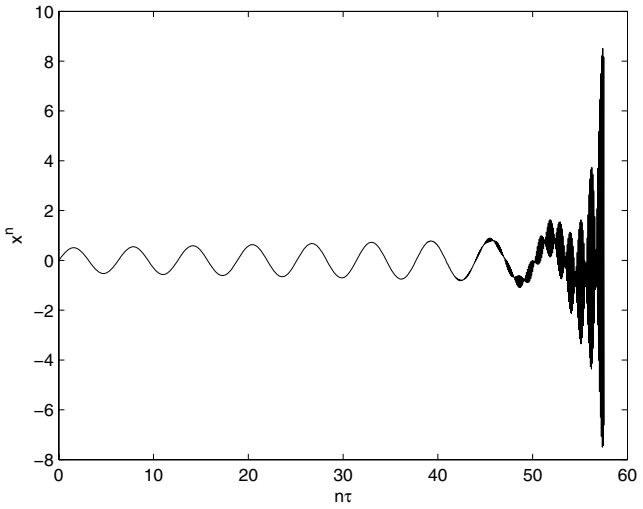


FIG. 4. *Results when running the difference system* (2.13), (2.14) *with* $\epsilon = 0.025$, $\omega = 1$, $\tau = 0.05$, *and* $x^0 = 0$, $y^0 = 0.5$.

and consider the expansion [9]

$$x^n = x_0(T_0, T_1) + \epsilon x_1(T_0, T_1) + O(\epsilon^2). \tag{3.2}$$

We shall also use partial shift operators E_{T_p} which we define as

$$E_{T_p} f(\ldots, T_p, \ldots) := f(\ldots, T_p + \epsilon^p \tau, \ldots) \tag{3.3}$$

in analogy with the temporal shift operator in (2.3). Using the partial shift operator (3.3) above we also define partial divided difference operators for T_p, namely,

$$\Delta_{T_p} := (E_{T_p} - 1)/(\epsilon^p \tau), \quad \nabla_{T_p} := (1 - E_{T_p}^{-1})/(\epsilon^p \tau). \tag{3.4}$$

The analogue of the central difference operator δ, defined in (2.4), follows similarly:

$$\delta_{T_p} \equiv (\Delta_{T_p} + \nabla_{T_p})/2. \tag{3.5}$$

Following Schoombie [9, 10] we note the following discrete multiple scales expansion of the differential approximations Δ and ∇ given in (2.8), namely,

$$\Delta = \Delta_{T_0} + \epsilon\, \Delta_{T_1} E_{T_0}, \quad \nabla = \nabla_{T_0} + \epsilon \nabla_{T_1} E_{T_0}^{-1}. \tag{3.6}$$

By making use of the definition of δ in (2.4) and substitution of (3.6) therein, we also obtain that

$$\delta = \delta_{T_0} + \epsilon/2(\Delta_{T_1} E_{T_0} + \nabla_{T_1} E_{T_0}^{-1}). \tag{3.7}$$

The equations above are essentially the discrete analogues of the chain rule for differentiation as used in a normal continuous multiple scales analysis.

We now proceed with our discrete multiple scales analysis. We substitute (3.6) into (2.9). Note the expansion of the first operator in (2.9):

$$\begin{aligned}&\Delta\nabla + (\tau^2/4)\Delta^2\nabla^2\\ &= \Delta_{T_0}\nabla_{T_0} + (\tau^2/4)\Delta_{T_0}^2\nabla_{T_0}^2\\ &+ \epsilon(\Delta_{T_0}\Delta_{T_1} + \nabla_{T_0}\nabla_{T_1} + (\tau^2/2)\Delta_{T_0}^2\nabla_{T_0}\Delta_{T_1} + (\tau^2/2)\nabla_{T_0}^2\Delta_{T_0}\nabla_{T_1}).\end{aligned} \tag{3.8}$$

We proceed by collecting terms with equal powers of ϵ. In this way we generate a set of equations given by

$$Lx_0 = 0, \tag{3.9}$$

relevant to $O(1)$, and for $O(\epsilon)$,

$$Lx_1(T_0, T_1) = R(T_0, T_1), \tag{3.10}$$

where the difference operator L is defined by

$$Lu(T_0, T_1) = (\Delta_{T_0}\nabla_{T_0} + (\tau^2/4)\Delta_{T_0}^2\nabla_{T_0}^2)u(T_0, T_1) + \omega^2 u(T_0, T_1) \tag{3.11}$$

for any discrete variable $u(T_0, T_1)$, and where

$$\begin{aligned}R(T_0, T_1) = &-(\Delta_{T_0}\Delta_{T_1} + \nabla_{T_0}\nabla_{T_1} + (\tau^2/2)\Delta_{T_0}^2\nabla_{T_0}\Delta_{T_1} + (\tau^2/2)\nabla_{T_0}^2\Delta_{T_0}\nabla_{T_1})x_0\\ &+ \beta(1 - x_0^2)\delta_{T_0}x_0.\end{aligned} \tag{3.12}$$

It is easy to see that any real-valued solution of (3.9) must be of the form

$$x_0(T_0, T_1) = a(T_1)e^{i\Omega n\tau/2} + (-1)^n b(T_1)e^{i\Omega n\tau/2} + \text{c.c.}, \tag{3.13}$$

where c.c. denotes the complex conjugate, and where

$$\cos(\Omega\tau) = 1 - 2\omega^2\tau^2. \tag{3.14}$$

The second term in (3.13) represents the computational wave considered by Cai et al. [1], the presence of which causes spurious behavior.

If we substitute this expression for x_0 into the right-hand side of (3.10), we eventually find that (neglecting $O(\epsilon^2)$ terms)

$$\begin{aligned}\frac{\tau}{i}R(T_0, T_1) = &\{-\sin(\Omega\tau)\delta_{T_1}a + \sin(\Omega\tau/2)(a - a|a|^2 + 3b^2\bar{a} - 2a|b|^2)\}e^{i\Omega n\tau/2}\\ &+ \{-\sin(\Omega\tau)\delta_{T_1}b + \sin(\Omega\tau/2)(-b + b|b|^2 - 3a^2\bar{b} + 2b|a|^2)\}(-1)^n e^{i\Omega n\tau/2}\\ &+ \sin(\Omega\tau/2)(-a^3 + ab^2)e^{3i\Omega n\tau/2}\\ &+ \sin(\Omega\tau/2)(b^3 - a^2b)(-1)^n e^{3i\Omega n\tau/2} + \text{c.c.},\end{aligned} \tag{3.15}$$

where the bar indicates the complex conjugate.

To ensure a bounded solution for x_1, we need to remove secular terms, which is done by requiring that the coefficients of both $e^{i\Omega n\tau/2}$ and $(-1)^n e^{i\Omega n\tau/2}$ be equated to zero. This leads to the two ordinary differential equations

$$a' = r(a - a|a|^2 + 3b^2\bar{a} - 2a|b|^2), \tag{3.16}$$

$$b' = -r(b - b|b|^2 + 3a^2\bar{b} - 2b|a|^2), \tag{3.17}$$

where

$$r = \frac{1}{2\cos(\Omega\tau/2)}, \tag{3.18}$$

and we have noted from the definition of δ_{T_1} that we can write

$$\delta_{T_1} = (E_{T_1} - E_{T_1}^{-1})/(2\epsilon\tau) = \partial_{T_1} + O(\epsilon^2). \tag{3.19}$$

Next we put

$$a = \frac{1}{2}\alpha(T_1)e^{i\theta(T_1)}, \quad b = \frac{1}{2}\beta(T_1)e^{i\phi(T_1)}, \tag{3.20}$$

where α and β are real and nonnegative, and θ and ϕ are real. This enables us to write x_0 in the real form

$$x_0(T_0, T_1) = \alpha\cos(\Omega n\tau/2 + \theta) + (-1)^n\beta\cos(\Omega n\tau/2 + \phi). \tag{3.21}$$

When we put (3.20) into (3.16) and (3.17), and separate real and imaginary parts, we finally obtain the following coupled system of nonlinear ordinary differential equations:

$$\alpha' = r\alpha(1 - \alpha^2/4 - \beta^2/2) + \frac{3}{4}r\alpha\beta^2\cos(2\phi - 2\theta), \tag{3.22}$$

$$\beta' = -r\beta(1 - \beta^2/4 - \alpha^2/2) - \frac{3}{4}r\beta\alpha^2\cos(2\phi - 2\theta), \tag{3.23}$$

$$\theta' = \frac{3}{4}r\beta^2\sin(2\phi - 2\theta), \tag{3.24}$$

$$\phi' = \frac{3}{4}r\alpha^2\sin(2\phi - 2\theta). \tag{3.25}$$

4. Amplitude dynamics. The system (3.22) through (3.25) can be reduced to three equations by introducing the variable $\gamma = \cos(2\phi - 2\theta)$:

$$\alpha' = r\alpha(1 - \alpha^2/4 - \beta^2/2) + \frac{3}{4}r\alpha\beta^2\gamma, \tag{4.1}$$

$$\beta' = -r\beta(1 - \beta^2/4 - \alpha^2/2) - \frac{3}{4}r\beta\alpha^2\gamma, \tag{4.2}$$

$$\gamma' = \frac{3}{2}r(\beta^2 - \alpha^2)(1 - \gamma^2). \tag{4.3}$$

As far as the orbits of this system are concerned, we will be interested only in the orbits within the region $\alpha \geq 0$, $\beta \geq 0$, $-1 \leq \gamma \leq 1$.

First consider the orbits for which the projections on the $\alpha\beta$-plane are along the α-axis (i.e., $\beta = 0$). For these the system (4.1)–(4.3) reduces to the two equations

$$\alpha' = r\alpha(1 - \alpha^2/4), \tag{4.4}$$

$$\gamma' = -(3/2)r\alpha^2(1 - \gamma^2). \tag{4.5}$$

Equation (4.4) has the solution

$$\alpha = \frac{2}{\sqrt{1 + (4/\alpha_0^2 - 1)e^{-2rT_1}}} = \frac{2}{\sqrt{1 + (4/\alpha_0^2 - 1)e^{-2r\epsilon n\tau}}}, \tag{4.6}$$

where

$$\alpha(0) = \alpha_0, \tag{4.7}$$

which is very close to the amplitude in (1.3), since $2r = 1 + O(\tau^2)$ from (3.18).

Equation (4.5) shows that if $|\gamma(0)| < 1$, γ will decrease monotonically and eventually approach the limiting value of $\gamma = -1$. This orbit therefore ends at the unstable equilibrium point $(2, 0, -1)$.

For the rest of the orbits in the quarter cylinder, it is convenient to transform to cylindrical coordinates, i.e., the coordinates (R, Θ, γ), where

$$\alpha = R\cos\Theta, \tag{4.8}$$

$$\beta = R\sin\Theta. \tag{4.9}$$

In terms of these coordinates, the system (4.1)–(4.3) then becomes

$$R' = rR(1 - R^2/4)\cos(2\Theta), \tag{4.10}$$

$$\Theta' = -r\left[1 - \frac{3}{8}R^2(1-\gamma)\right]\sin(2\Theta), \tag{4.11}$$

$$\gamma' = -\frac{3}{2}rR^2(1-\gamma^2)\cos(2\Theta). \tag{4.12}$$

This system can be solved analytically. First we eliminate Θ by dividing (4.12) by (4.10) to get the differential equation

$$\frac{d\gamma}{dR} = -\frac{3}{2}\frac{R(1-\gamma^2)}{1 - R^2/4}. \tag{4.13}$$

This can easily be solved by separation of variables to give the following relationship between γ and R:

$$\gamma = \frac{C(1 - R^2/4)^6 - 1}{C(1 - R^2/4)^6 + 1}, \tag{4.14}$$

where C is a constant defined by

$$C = \frac{(1+\gamma_0)}{(1-\gamma_0)(1 - R_0^2/4)^6}, \tag{4.15}$$

which is positive since $-1 < \gamma_0 = \gamma(0) < 1$. We note that in the limit if $R \to 2$, then $\gamma \to -1$.

Note that (4.14) defines a family of nonintersecting surfaces, which are all tangent to the curve

$$R = 2, \quad \gamma = -1, \quad 0 \le \Theta \le \pi/2.$$

Figure 5 shows some of these surfaces.

If we now substitute the relationship (4.14) into (4.10) and (4.11), we get the system

$$R' = rR(1 - R^2/4)\cos(2\Theta), \tag{4.16}$$

$$\Theta' = -r\left[1 - \frac{3R^2}{4[C(1 - R^2/4)^6 + 1]}\right]\sin(2\Theta). \tag{4.17}$$

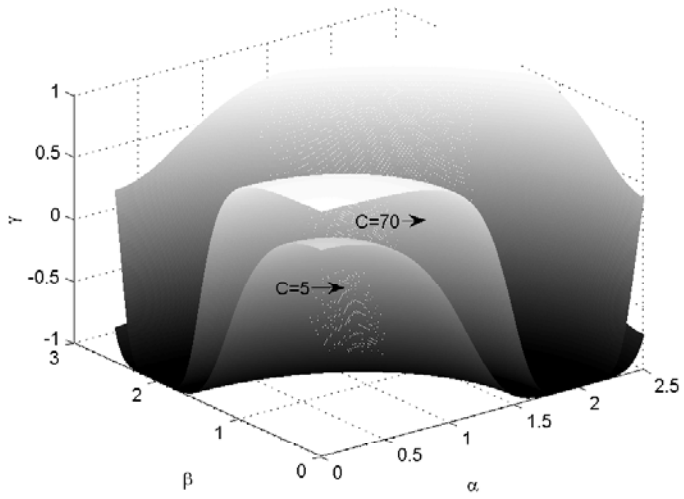


FIG. 5. *Two of the surfaces defined by* (4.14)*, namely, for* $C = 5$ *and for* $C = 70$.

This system describes the projection on the $\alpha\beta$-plane of all orbits intersecting the circle arc

$$R = R_0, \quad 0 \le \Theta \le \pi/2, \quad \gamma = \gamma_0, \tag{4.18}$$

defining C in (4.15).

Since we are primarily interested in the interaction of α and β (and therefore that of R and Θ), this dynamical system is of particular interest to us, and it is worthwhile to examine its dynamics in some detail.

In the quadrant $0 \le \Theta \le \pi/2$ the system has the equilibrium points

$$(R, \Theta) = (0, 0), \ \ (2, 0), \ \ (2, \pi/2),$$

as well as the point where $\Theta = \pi/4$ and

$$3R^2 = 4[C(1 - R^2/4)^6 + 1]. \tag{4.19}$$

If we put $u = 1 - R^2/4$, where $u \le 1$, (4.19) can be written in the simpler form

$$Cu^6 = 2 - 3u. \tag{4.20}$$

As can be seen clearly from Figure 6, the curve representing the left-hand side of (4.20) and the line representing the right-hand side have two real intersections for each value of C, namely, $u_1 > 0$ and $u_2 < 0$. This corresponds to two equilibrium points $(R_1, \pi/4)$ and $(R_2, \pi/4)$, such that $R_1 < 2$ and $R_2 > 2$. Clearly, if $C \to 0$, then $u_1 \to 2/3$ and $R_1 \to 2/\sqrt{3}$, while $u_2 \to -\infty$ and $R_2 \to \infty$. Also, if $C \to \infty$, then both $u_1, u_2 \to 0$ and $R_1, R_2 \to 2$. Thus

$$2/\sqrt{3} \le R_1 \le 2, \quad 2 \le R_2 < \infty, \tag{4.21}$$

with the equalities achieved when $\gamma_0 = -1$ (so that $C = 0$) and $\gamma_0 = 1$ (so that $C = \infty$), respectively.

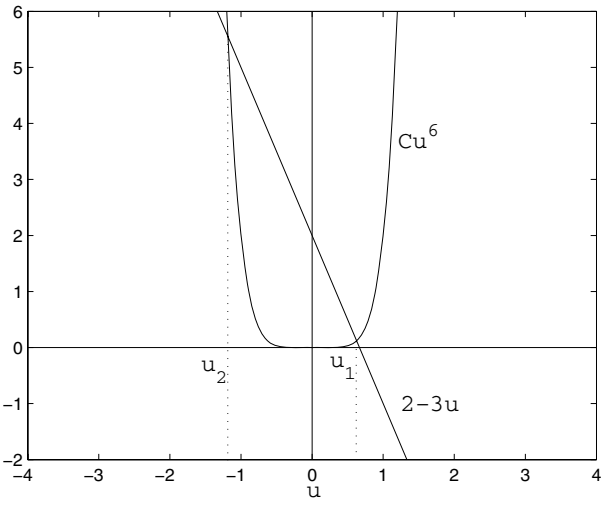


FIG. 6. *The two real solutions* $u_2 < 0$ *and* $u_1 > 0$ *of* (4.20)*, shown here for* $C = 2$.

The points $(R, \Theta) = (0, 0), (2, 0)$, and $(2, \pi/2)$ are easily seen to be saddle points. We shall show that the remaining two equilibrium points of the projected system are both surrounded by closed orbits, and are therefore centers.

It is also easy to see that the lines $\Theta = 0$ and $\Theta = \pi/2$ are orbits, directed away from and towards the origin, respectively, as well as the circle segment $R = 2$, $0 \le \Theta \le \pi/2$, in a counterclockwise direction.

The rest of the orbits can be found from the differential equation

$$\frac{du}{d\Theta} = \frac{2u(1-u)(Cu^6+1)\cot(2\Theta)}{Cu^6+1-3(1-u)}, \tag{4.22}$$

obtained by dividing (4.16) by (4.17) and putting

$$u = 1 - R^2/4.$$

Integrating this by separation of variables yields the orbit equation

$$\ln(\sin 2\Theta) = \ln\frac{u}{1-u} - \frac{1}{2}\ln\left(\frac{Cu^6}{1+Cu^6}\right) + \text{ constant}, \tag{4.23}$$

or

$$L\sqrt{1+Cu^6} = u^2(1-u)\sin(2\Theta), \tag{4.24}$$

where L is a constant which is nonnegative for orbits in the region under consideration. (Note that (4.24) is also true for $C = 0$, i.e., $\gamma = -1$. According to (4.12) these are planar orbits.)

To investigate the nature of these orbits, we first note that (4.22) shows that for $u < 1$ and $C > 0$,

$$\frac{du}{d\Theta} = 0 \text{ when } \Theta = \pi/4. \tag{4.25}$$

The extremes of u are therefore on the line $\Theta = \pi/4$ and are defined by

$$u^2(1-u) = L\sqrt{1+Cu^6} \tag{4.26}$$

or, equivalently, are given by those zeros of the polynomial

$$f(u) = (L^2C-1)u^6 + 2u^5 - u^4 + L^2 \tag{4.27}$$

which are less than one. Now since

$$f'(u) = 2u^3[3(L^2C-1)u^2 + 5u - 2], \tag{4.28}$$

the critical points of f are at $u = 0$ and $u = u_\pm$, where

$$u_\pm = \frac{-5 \pm \sqrt{1+24L^2C}}{6(L^2C-1)}. \tag{4.29}$$

For all $L \geq 0$, u_+ is always a local minimum, with

$$0 \leq u_+ = \frac{4}{(5+\sqrt{1+24L^2C})} \leq 2/3 < 1. \tag{4.30}$$

On the other hand, when $L^2C > 1$, u_- is a local minimum with

$$u_- = \frac{5+\sqrt{1+24L^2C}}{6(1-L^2C)} < 0, \tag{4.31}$$

while if $L^2C < 1$, u_- is a local maximum and $u_- > 1$.

Thus the nature of f, and therefore of its zeros, depends very much on whether $L^2C > 1$ or $L^2C < 1$. If $L^2C < 1$, i.e., for $0 < L < 1/\sqrt{C}$, the graph of f is as in Figure 7.

There is then always one and only one negative zero, signifying a single extreme value for u and hence an open orbit in the region $R > 2$. There are no positive zeros at all, unless $L \leq L_1$, where L_1 is that value of L for which f is tangent to the horizontal axis at u_+ (see Figure 7(b)). For $L < L_1$ there are two positive zeros of f less than one (Figure 7(a)), signifying two extremes in u in the region $R < 2$, and thus a closed orbit. When $L > L_1$, the function f is positive for all positive u (Figure 7(c)), and hence for all $R < 2$. Since this also implies that $\sin(2\Theta) > 1$, there are no orbits in this region for these values of L. Hence all orbits inside the quarter circle $R < 2$, $0 < \Theta < \pi/2$ are closed. However, there are open orbits in the region $R > 2$ for the entire range of values of L between 0 and $1/\sqrt{C}$.

When $L^2C > 1$, i.e., $L > 1/\sqrt{C}$, the graph of f is as in Figure 8.

There are no positive zeros of f less than one at all, and negative zeros only if $L < L_2$, where L_2 is that value of L for which f is tangent to the horizontal axis at u_- (Figure 8(b)). If $1/\sqrt{C} < L < L_2$, there are two negative zeros, signifying closed orbits in the region $R > 2$ (Figure 8(a)). For $L > L_2$, the function f is positive for all u (Figure 8(c)), which means that there are no orbits at all for these values of L.

To find L_1 and L_2, we return to (4.26). The function f will be tangent to the horizontal axis when the functions $f_1 = u^2(1-u)$ and $f_2 = L\sqrt{1+Cu^6}$ are tangent to each other. If we suppose that the tangent point is at $u = \nu$, corresponding to $L = L^*$, then ν and L^* may be found by solving the two equations

$$\nu^2(1-\nu) = L^*\sqrt{1+C\nu^6}, \tag{4.32}$$

$$2\nu - 3\nu^2 = \frac{3CL^*\nu^5}{\sqrt{1+C\nu^6}}, \tag{4.33}$$

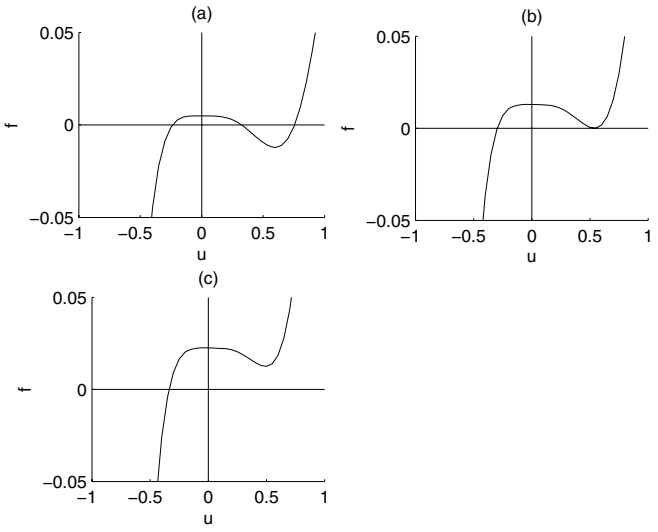


FIG. 7. *The function f as defined in* (4.27) *for* $C = 16$ *and for* (a) $L = 0.07 < L_1$, (b) $L = 0.1135 = L_1$, *and* (c) $L = 0.15 > L_1$. *In all three cases* $L^2C < 1$.

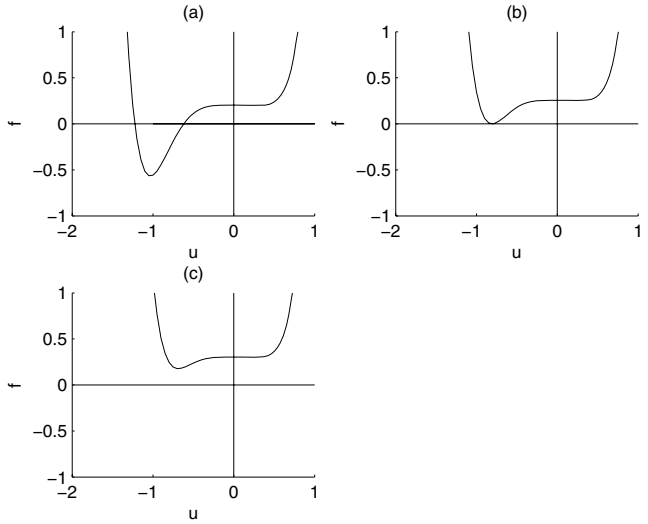


FIG. 8. *The function f as defined in* (4.27) *for* $C = 16$ *and for* (a) $L = 0.4 < L_2$, (b) $L = 0.5055 = L_2$, *and* (c) $L = 0.55 > L_2$. *In all three cases* $L^2C > 1$.

from which we find, by eliminating L^*, that ν satisfies

$$2 - 3\nu = C\nu^6, \tag{4.34}$$

which is just (4.20). Thus $\nu = u_1, u_2$ and $L^* = L_1, L_2$, respectively, and the tangent points are the centers $(R, \Theta) = (R_i, \pi/4)$, $i = 1, 2$, of the projected system. Also

$$L_i = \frac{(u_i)^2\sqrt{1 - u_i}}{\sqrt{3}}. \tag{4.35}$$

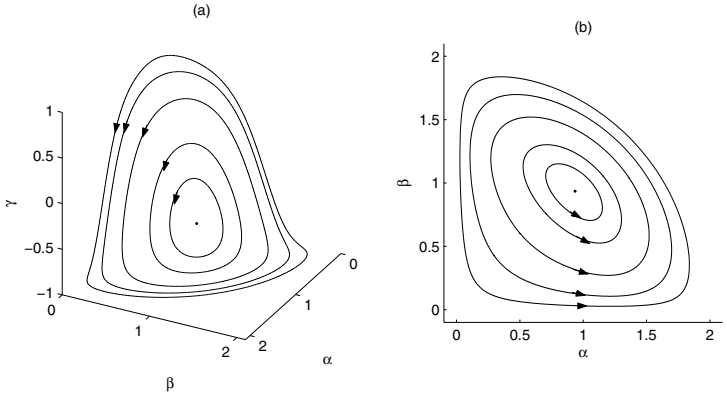


FIG. 9. *Three-dimensional phase diagram* (a) *as well as its projection* (b) *on the* $\alpha\beta$*-plane for* $C = 10$.

Thus the orbits may be classified as follows:

- $L = 0$: The saddle points $(R, \Theta) = (0, 0), (2, 0), (2, \pi/2)$, and the orbit $R = 2$, $0 < \Theta < \pi/2$. (Note that from (4.15), whenever $R = 2$, $\gamma = -1$, so that these saddle points will correspond to the unstable equilibrium points $(\alpha, \beta, \gamma) = (0, 0, (C-1)/(C+1)), (0, 2, -1), (2, 0, -1)$ of the three-dimensional system (4.1)–(4.3).)
- $0 < L < L_1$: Closed orbits for $R < 2$, open orbits for $R > 2$.
- $L = L_1$: Center at $R = R_1, \Theta = \pi/4$ and an open orbit in the region $R > 2$.
- $L_1 < L \leq 1/\sqrt{C}$: Only open orbits in the region $R > 2$.
- $1/\sqrt{C} < L < L_2$: Only closed orbits in the region $R > 2$.
- $L = L_2$: Center at $R = R_2$, $\Theta = \pi/4$.
- $L > L_2$: No orbits.

From (4.22) we can also see that $\frac{d\Theta}{du}$ will be zero if $Cu^6 + 3u - 2 = 0$ or $u = u_1, u_2$. Then the extreme values of Θ on the orbit are given by

$$\sin 2\Theta_e = L/L_i, \quad i = 1, 2, \tag{4.36}$$

i.e.,

$$\Theta_{\max} = \pi/2 - \frac{1}{2}\arcsin(L/L_i), \tag{4.37}$$

$$\Theta_{\min} = \frac{1}{2}\arcsin(L/L_i). \tag{4.38}$$

A typical phase diagram, both in three-dimensional phase space and as projected on the $\alpha\beta$-plane, is shown in Figure 9 for $C = 10$.

5. Application to the leapfrog method. Having studied the dynamics of the amplitudes of the two interacting waves in general, we can now relate it to the actual numerical results obtained when solving (2.5) and (2.6). It should be remembered that the values of y^1 and x^1 must be found by using some one-step method like the Euler method or one of the Runge–Kutta methods before the leapfrog method can start. Because of this, the computational wave is present from the outset, as we will demonstrate below.

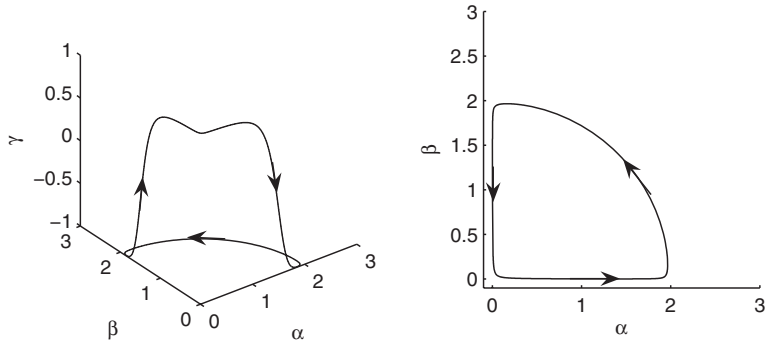


FIG. 10. *Orbit in three dimensions, and projected on the $\alpha\beta$-plane, for an Euler starter and $\tau = 0.2$, $\epsilon = 0.025$, $\omega = 1$, $x^0 = 0$, $y^0 = 0.5$.*

To find the initial amplitudes $\alpha(0) = \alpha_0$ and $\beta(0) = \beta_0$, as well as the initial phase angles θ_0 and ϕ_0 (and hence also γ_0), we put

$$x^n = \alpha_0 \cos\left(\frac{\Omega n\tau}{2} + \theta_0\right) + (-1)^n \beta_0 \cos\left(\frac{\Omega n\tau}{2} + \phi_0\right) \tag{5.1}$$

for $n = 0, 1, 2$, and 3. (Here we assume that ϵ is small enough so that the functions depending on T_1, i.e., α, β, θ, and ϕ, do not change significantly over three time steps.) Note that x^0 and y^0 are the given initial conditions, and x^1, x^2, and x^3 depend on the actual starter scheme used. By solving for $\alpha_0 \cos\theta_0$, $\alpha_0 \sin\theta_0$, $\beta_0 \cos\phi_0$, and $\beta_0 \sin\phi_0$ first, we need only solve a linear algebraic system. Eventually we can then find R_0, Θ_0, and γ_0, and hence the orbit parameters C and L.

For $\tau = 0.2$, $\epsilon = 0.025$, $\omega = 1$, $x^0 = 0$, $y^0 = 0.5$, and an Euler starter, the initial amplitudes and phases were

$$\alpha_0 = 0.5089, \tag{5.2}$$

$$\beta_0 = 0.0040, \tag{5.3}$$

$$\theta_0 = -1.5718 \text{ radians}, \tag{5.4}$$

$$\phi_0 = -1.4428 \text{ radians}. \tag{5.5}$$

Thus the computational wave has a small but significant amplitude β_0 initially. The initial orbit polar coordinates are then

$$R_0 = 0.5089, \tag{5.6}$$

$$\Theta_0 = 0.0079 \text{ radians}, \tag{5.7}$$

$$\gamma_0 = 0.9669, \tag{5.8}$$

from which the orbital parameters of this numerical solution can be calculated as $C = 88.8002$ and $L = 0.00011453$. The resulting orbit and its projection in the $\alpha\beta$-plane are shown in Figure 10.

We also used both a second order and a fourth order Runge–Kutta starter. The initial phases and amplitudes for these are given in Table 1.

The corresponding orbits are shown in Figures 11 and 12.

When comparing Figures 10, 11, and 12, we note that the projection of the orbit on the $\alpha\beta$-plane looks virtually the same in all three cases, but that the full

TABLE 1
Orbital parameters for different starter schemes.

Starter	α_0	β_0	θ_0	ϕ_0	R_0	Θ_0	γ_0	C	L
2nd order RK	0.5038	0.00117	−1.5715	1.2499	0.5038	0.0023	0.8019	13.4768	0.000083
4th order RK	0.5038	0.00132	−1.5722	1.0015	0.5038	0.0026	0.4215	3.6407	0.000160

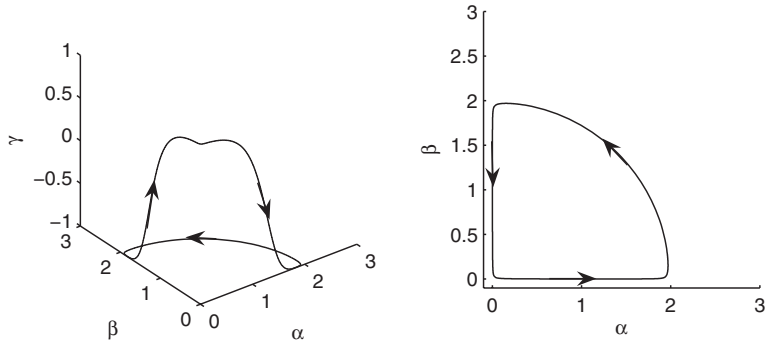


FIG. 11. *Orbit in three dimensions, and projected on the $\alpha\beta$-plane, for a second order Runge–Kutta starter and $\tau = 0.2$, $\epsilon = 0.025$, $\omega = 1$, $x^0 = 0$, $y^0 = 0.5$.*

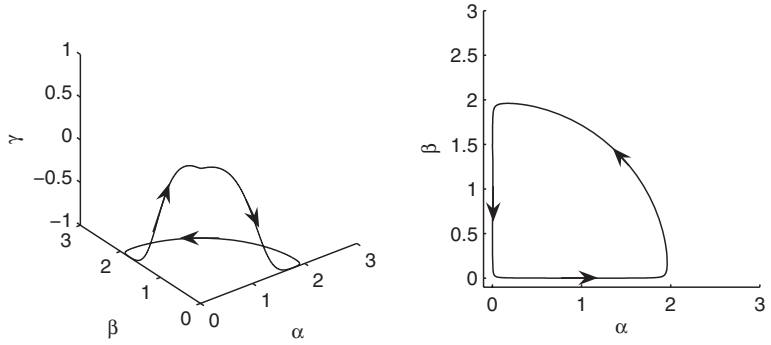


FIG. 12. *Orbit in three dimensions, and projected on the $\alpha\beta$-plane, for a fourth order Runge–Kutta starter and $\tau = 0.2$, $\epsilon = 0.025$, $\omega = 1$, $x^0 = 0$, $y^0 = 0.5$.*

three-dimensional orbit varies significantly. The more inaccurate the starter scheme is, the more variation there seems to be in γ, and hence in the phase angles θ and ϕ.

From the actual orbits it is easy to find the point where the computational wave starts to grow. Cai et al. [1] calculated that this happens when $\alpha = \sqrt{0.8} = 0.8944$. Our nonlinear analysis allows us to be more precise by finding the lowest point of a particular orbit. In fact, we can already see that the computational wave is initially damped, until the damping eventually becomes an amplification. In Table 2 we list, for each of the three starter schemes and the parameter values and initial conditions as before, the point $\alpha = \alpha_1^*$, where the amplitude β of the computational wave reaches a minimum value of β^*. At this point the computational wave starts growing, albeit very slowly at first. We also list the point $\alpha = \alpha_2^*$ where the amplitude β first exceeds the value of 0.01.

TABLE 2

The point $\alpha = \alpha_1^*$ on the orbit where β reaches its minimum β^*, together with the value α_2^* of α, where β first exceeds 0.01 for the three different starter schemes, and $\tau = 0.2$, $\epsilon = 0.025$, $\omega = 1$, $x^0 = 0$, $y^0 = 0.5$.

Starter	α_1^*	α_2^*	β^*
Euler	1.4450	1.8866	0.000996
2nd order RK	1.2507	1.9043	0.000465
4th order RK	1.1065	1.8655	0.000713

We see that α_1^* is sensitive for the particular starter scheme used, whereas α_2^* does not vary very much. From Table 1 we see that C decreases with the accuracy of the initial values x^1 and y^1. It is worthwhile to note at this point that if $C = 0$ (and therefore $\gamma = -1$ throughout), then if we consider an orbit close to $\beta = 0$, by putting $\beta = \epsilon \ll 1$, then (4.2) becomes

$$\epsilon' = r\epsilon(5\alpha^2/4 - 1), \tag{5.9}$$

showing that α_1^* would take on the value of $\sqrt{0.8}$ predicted by Cai et al. Their value is therefore a limiting value. In practice, growth of the computational wave starts a bit later. In practical terms, for the type of accuracy expected from a second order method and a time step of 0.2, the computational wave will start to cause trouble only when its amplitude becomes larger than about 0.01, i.e., about 0.5% of the limit circle amplitude. This does not happen before the solution gets very close to the limit cycle, i.e., at about $\alpha = 1.9$. Up to then the effect of β in (3.22) and (3.24) is negligible, so that the phase angle θ is virtually constant, and the amplitude of the solution is given by (4.6), which is as near to the analytic solution as can be expected from a second order method. In other words, the numerical solution is quite sound and useful until it very nearly reaches the limit cycle.

We may also note from the phase diagram in Figure 9 that when we start with initial conditions which leads to an initial amplitude larger than that of the limit cycle, the open orbits in the region $R > 2$ imply eventual blowup. To test this, we used the parameter values we used before, but changed the initial values to $x^0 = 5$, $y^0 = 0$. With a second order Runge–Kutta starter, the numerical results are shown in Figure 13.

The computational wave is clearly seen to become significantly large quite early, causing a blowup of the solution at about 50 time units. The orbit parameters for these parameter values and initial conditions turn out to be $C = 3.1070 \times 10^{-6}$ and $L = 5.2642$. For this low value of C, closed orbits in the region $R > 2$ start only at about $L = 567$, so that this orbit is definitely open. Figure 14 shows this orbit, both in three dimensions and projected on the $\alpha\beta$-plane. The point on this orbit where β is a minimum is at $\alpha = \alpha_1^* = 6.82$. Since the initial point on this orbit is further along, the computational wave is amplified from the very start, where it already has an amplitude of 0.08. Thus, for these particular initial conditions, the numerical solution is useless.

Finally (4.12) implies that there is another set of planar orbits besides $\gamma = -1$, namely, those on the plane $\gamma = 1$ ($C = \infty$). Then, according to (4.11), $\Theta' < 0$ on $(0, \pi/2)$, and hence β, monotonically decreases for all orbits on this plane. Thus the computational wave will cause no problems as long as $\gamma = 1$ (i.e., $\theta = \phi$). Unfortunately these orbits are structurally unstable. The smallest deviation from $\gamma = 1$ will move the solution to one of the other C-surfaces, and hence lead to growth of β.

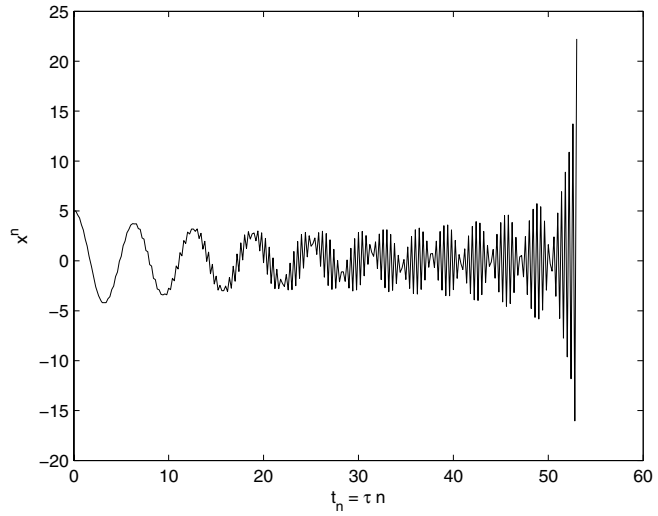


FIG. 13. *Results when running the difference system* (2.5), (2.6) *with* $\epsilon = 0.025$, $\tau = 0.2$, *and* $x^0 = 5$, $y^0 = 0$.

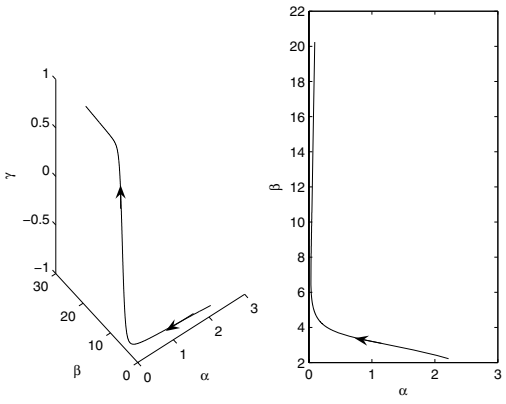


FIG. 14. *Orbit in three dimensions, and projected on the* $\alpha\beta$*-plane, for a second order Runge–Kutta starter and* $\tau = 0.2$, $\epsilon = 0.025$, $\omega = 1$, $x^0 = 5$, $y^0 = 0$.

6. Discussion. In this paper we demonstrated that attempts to calculate the solution of the van der Pol equation by means of several Nyström methods of different orders lead to nonlinear instabilities caused by sudden growth of a high frequency mode. This high frequency induced (HFI) instability can be delayed simply by decreasing the time step. Thus it is always possible to calculate the limit cycle accurately, provided a small enough time step is used. Time step restrictions get more severe as the order of the method is increased.

As far as the analysis is concerned, we first improved upon the work done by Cai et al. [1], in the sense that by doing a full nonlinear analysis of the amplitude dynamics of the second order Nyström difference scheme (2.5), we were able to fully explain the observed spurious amplitude modulation imposed by it, and also to assess the extent to which this scheme could still be useful as a second order numerical method.

We also demonstrated the use of a discrete multiple scales technique combined with a phase plane analysis to analyze such a nonlinear difference scheme. These techniques will certainly also be applicable to other numerical methods for the van der Pol and other nonlinear differential equations, and indeed to nonlinear models which are discrete to start with.

The fact that leapfrog time discretizations could cause spurious behavior has been known for some time [7, 8, 11], and Sanz-Serna [7] and Sanz-Serna and Vadillo [8] did a phase plane analysis in some simpler cases. Of course, in the case of (2.5) the problem is much more complicated, since it is not the behavior of the solution itself which needs to be analyzed (as in the cases studied by Sanz-Serna and his coworkers), but rather that of its amplitude, which makes it necessary to use a discrete multiple scales method first.

In a case such as Duffing's equation, which models oscillations but does not have a limit cycle, we found that the HFI instabilities were less severe, in the sense that it seems to be totally absent in the case of the second order Nyström scheme, although still present in the case of the higher order methods.

We also observed HFI instability where Nyström methods were applied to other cases where limit cycles occurred in the dynamics of the solutions. One such case was a predator-prey system described in [5, Chapter 3, section 3.3]. An analysis of these and other cases is currently under way and will be discussed in forthcoming papers.

REFERENCES

[1] D. Cai, A. Aoyagi, and K. Abe, *Parametric excitation of computational mode of the Leapfrog scheme applied to the van der Pol equation*, J. Comput. Phys., 107 (1993), pp. 146–151.

[2] E. Hairer, S. P. Nørsett, and G. Wanner, *Solving Ordinary Differential Equations* I, Springer-Verlag, Berlin, 1988.

[3] D. W. Jordan and P. Smith, *Nonlinear Ordinary Differential Equations*, 3rd ed., Oxford University Press, Oxford, UK, 1999.

[4] G. M. Moremedi and D. P. Mason, *Analysis of the limit cycle of a generalized van der Pol equation by a time transformation method*, Quaest. Math., 17 (1994), pp. 349–380.

[5] J. D. Murray, *Mathematical Biology.* I, 3rd ed., Springer-Verlag, New York, 2002.

[6] A. H. Nayfeh, *Perturbation Methods*, Wiley, New York, 1973.

[7] J. M. Sanz-Serna, *Studies in numerical nonlinear instability* I. *Why do leapfrog schemes go unstable?*, SIAM J. Sci. Comput., 6 (1985), pp. 923–938.

[8] J. M. Sanz-Serna and F. Vadillo, *Studies in numerical nonlinear instability* III. *Augmented Hamiltonian systems*, SIAM J. Appl. Math., 47 (1987), pp. 92–108.

[9] S. W. Schoombie, *A Discrete Multiple Scales Analysis of a Discrete Version of the Korteweg–de Vries Equation*, Report NA/109, University of Dundee, 1987.

[10] S. W. Schoombie, *A discrete multiple scales analysis of a discrete version of the Korteweg–de Vries equation*, J. Comput. Phys., 101 (1992), pp. 55–70.

[11] D. M. Sloan and A. R. Mitchell, *On nonlinear instabilities in leap-frog finite difference schemes*, J. Comput. Phys., 67 (1986), pp. 372–395.

[12] B. van der Pol, *On a type of oscillation hysteresis in a simple triode generator*, Philos. Mag., 43 (1922), pp. 177–193.

SIAM J. NUMER. ANAL.
Vol. 46, No. 3, pp. 1166–1189
© 2008 Society for Industrial and Applied Mathematics

ON THE ERROR IN QR INTEGRATION*

LUCA DIECI† AND ERIK S. VAN VLECK‡

Abstract. An important change of variables for a linear time varying system $\dot{x} = A(t)x,\ t \geq 0$, is that induced by the QR-factorization of the underlying fundamental matrix solution: $X = QR$, with Q orthogonal and R upper triangular (with positive diagonal). To find this change of variable, one needs to solve a nonlinear matrix differential equation for Q. Practically, this means finding a numerical approximation to Q by using some appropriate discretization scheme, whereby one attempts to control the local error during the integration. Our contribution in this work is to obtain global error bounds for the numerically computed Q. These bounds depend on the local error tolerance used to integrate for Q, and on structural properties of the problem itself, but not on the length of the interval over which we integrate. This is particularly important, since—in principle—Q may need to be found on the half-line $t \geq 0$.

Key words. QR methods, orthogonal integration, Lyapunov exponents, integral separation

AMS subject classification. 65L

DOI. 10.1137/06067818X

Notation. An $(n \times n)$ real matrix X is indicated by $X \in \mathbb{R}^{n\times n}$. $\mathrm{diag}(X)$ is the matrix comprising the diagonal part of X, the rest being all 0's; $\mathrm{upp}(X)$ is the matrix comprising the upper triangular part of X, the rest being all 0's; and $\mathrm{low}(X)$ is the matrix comprising the strictly lower triangular part of X, the rest being all 0's. The default norm we consider is the 2-norm of vectors and the induced norm for matrices.

1. Introduction.

Consider the homogeneous nonautonomous linear differential equation

$$\dot{x}(t) = A(t)x(t),\ t \geq 0, \tag{1.1}$$

where A is a bounded function taking values in $\mathbb{R}^{n\times n}$. Equation (1.1) appears pervasively in the study of dynamical systems. For example, it is the equation we end up with when we study variation with respect to the initial conditions, or parameters, of a nonlinear system. Therefore, it is the problem we have to face when we do general stability analyses for trajectories of a dynamical system, e.g., for periodic or for chaotic trajectories. Moreover, (1.1) is also the problem at hand during a Newton process to solve general nonlinear differential systems, a process often advocated for solving boundary value problems. Alas, in spite of its apparent simplicity, numerical investigation of (1.1) is extremely hard, since the solution structure depends on the fundamental matrix solution. Unquestionably, the problem is certainly conceptually and computationally simpler if A happened to be triangular. For this reason, techniques which find an orthogonal change of variable to triangular structure have been studied by several researchers for a long time; e.g., see [6, 12, 18]. Our own interest in these techniques originates with methods to approximate Lyapunov exponents of

*Received by the editors December 20, 2006; accepted for publication (in revised form) November 19, 2007; published electronically March 7, 2008. This work was supported in part under NSF grants DMS-0139895, DMS-0139824, and DMS-0513438.
http://www.siam.org/journals/sinum/46-3/67818.html
†School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332 (dieci@math.gatech.edu).
‡Department of Mathematics, University of Kansas, Lawrence, KS 66045 (evanvleck@math.ku.edu).

dynamical systems, a feat which is greatly simplified when the system is brought in upper triangular structure; e.g., see [13, 8, 9].

Of course, the factor Q which performs the change of variables has to be found numerically, and this itself is not easy since Q satisfies a nonlinear matrix differential equation. Thus, to find Q, one must approximate the solution of this nonlinear matrix equation in some appropriate way. In practice, this means that we will control the local errors while approximating Q, a fact which generally does not guarantee that Q will be approximated accurately, i.e., that the global error in our approximation will stay small. Our contribution in this work is to provide accurate bounds on the *global* error when finding Q: Our bounds will depend on the *local* error tolerance and on the coefficient function A, but not on the length of the interval over which we approximate Q. Our result is somewhat atypical and is important. It is atypical because, even though Q lies in a compact space, usually one does not obtain accurate global error bounds (on arbitrarily long intervals) except for contractive problems, and our problem is not contractive. It is important, because—used in conjunction with standard techniques to approximate Lyapunov exponents—it can be used to obtain global error bounds on the computed Lyapunov exponents of a linear time varying system, as well as global errors on other spectral quantities.

The way we will obtain global error bounds for the computed Q is in itself interesting and apparently new. Our main idea is to combine two types of error analyses: A backward error analysis guaranteeing that the computed Q factor gives a transformation to nearly triangular form, and a forward error analysis guaranteeing that for this nearly triangular problem there is a near-the-identity orthogonal transformation reducing it further to a triangular structure. Combining these two ingredients, we will obtain the sought result. Oversimplifying it, let us sketch the basic idea which has guided us:

- We want to express $X = QR$, Q orthogonal, R upper triangular with positive diagonal.
  - If we had Q, then R would satisfy a triangular system $\dot{R} = \widetilde{B}R$.
  - Suppose that instead of Q we compute (*backward error result*) an orthogonal Q_c, which gives $X = Q_c\widehat{R}$, with $\dot{\widehat{R}} = (B + F)\widehat{R}$, with B triangular and F of small norm (F not triangular).
  - Suppose also that we write $\widehat{R} = VU$, with V orthogonal and U upper triangular with positive diagonal. Then we have $X = Q_cVU$, and so, by uniqueness, $R = U$ and $Q_cV = Q$.
- If we now show that $V \approx I$ (*forward error result*), then we will infer that $Q_c \approx Q$ (*global error result*).

An outline of the paper is as follows. In the remainder of this introduction we review the basic change of variables $X = QR$ and the differential equations satisfied by Q and R. In section 2 we recall the key backward error statement which we proved in [10]. In section 3 we give in a concise way the global error statement result, and in section 4 we give details of a systematic way to obtain sharp bounds on the quantities appearing in the error bound. In section 5 we illustrate our results in an example. Conclusions are in section 6, which include a remark on the modifications needed to handle the case in which we only have a "reduced" QR-factorization, that is, when X comprises only a subset of columns of the fundamental matrix solution.

We now consider the differential equations governing the evolution of the Q and R factors in the QR-factorization of X. Presently, X is a fundamental matrix solution

for (1.1): $\dot{X} = A(t)X$, $X(0) = X_0$ invertible. Let $X_0 = Q_0R_0$ be the unique QR-factorization of X_0 with the diagonal of R_0 being positive.

Differentiating the relation $X = QR$ one obtains $\dot{Q}R + Q\dot{R} = A(t)QR$, and multiplying by Q^T on the left we obtain the equation for R:

$$\dot{R} = \widetilde{B}(t)R\ , \quad R(0) = R_0, \tag{1.2}$$

where we have set

$$\widetilde{B}(t) := Q^T(t)A(t)Q(t) - Q^T(t)\dot{Q}(t). \tag{1.3}$$

Let us formally set $S := Q^T\dot{Q}$. Since R has to be upper triangular, we must have $\widetilde{B}$ upper triangular, which leads to

$$\dot{Q} = QS(Q, A(t)),\ \ t \geq 0, \tag{1.4}$$

where

$$S(Q(t), A(t))_{ij} = \begin{cases} (Q^T(t)A(t)Q(t))_{ij}, & i > j, \\ 0, & i = j, \\ -(Q^T(t)A(t)Q(t))_{ji}, & i < j. \end{cases} \tag{1.5}$$

In particular, we notice that if Q is known, then R satisfies (1.2), and we also notice that S is linear in A. Furthermore, in light of (1.5), for the entries of $\widetilde{B}$ we have $\widetilde{B}_{ij} = \left(Q^TAQ\right)_{ij} + \left(Q^TAQ\right)_{ji}$ for $i < j$ and $\widetilde{B}_{ii} = \left(Q^TAQ\right)_{ii}$, that is, $\widetilde{B} = \mathrm{upp}(Q^TAQ) + (\mathrm{low}(Q^TAQ))^T$.

The above derivation of the equations for the QR-factorization of X has been obtained many times before, and specific attention has been paid in recent years to techniques which maintain orthogonality while approximating the factor Q. A sample of relevant references includes [2, 3, 14, 12, 7, 16]. We are not going to review these works in detail, because the precise way in which the approximation for Q is obtained is not relevant to our main scope here, which is to derive global error bounds for the obtained approximations to Q. What is relevant is that the obtained approximations be orthogonal at the grid-points found during numerical integration of (1.4), a fact which the schemes proposed in the above cited works do achieve.

2. Background. Suppose we are seeking the factorization $X(t_k) = Q(t_k)R(t_k)$, $k = 0, 1, 2, \ldots$. In other words, we are looking for the change of variables, the factor Q in the QR-factorization of X, at the grid-points $0 = t_0 < t_1 < \cdots$. Practically, the grid-points $\{t_k\}$ may have been found during numerical integration of (1.4) by any of the schemes in the previously cited works. Alternatively, we can always think of indirectly having found approximations to Q by directly seeking the QR-factorization of $X(t_k)$ as follows. Write

$$X(t_k) = \Phi(t_k, t_{k-1}) \ldots \Phi(t_2, t_1)\Phi(t_1, 0)X_0, \tag{2.1}$$

where

$$\dot{\Phi}(t, t_{j-1}) = A(t)\Phi(t, t_{j-1}),\ \Phi(t_{j-1}, t_{j-1}) = I, \quad t_{j-1} \leq t \leq t_j,\ j = 1, 2, \ldots, k. \tag{2.2}$$

Then, for $j = 1, 2, \ldots, k$, recursively consider (*discrete QR technique*)

$$\begin{aligned} &\dot{\Psi}(t, t_{j-1}) = A(t)\Psi(t, t_{j-1}),\ \ \Psi(t_{j-1}, t_{j-1}) = Q(t_{j-1}), \\ &\text{and factor } \Psi(t_j, t_{j-1}) = Q(t_j)R(t_j, t_{j-1}), \end{aligned} \tag{2.3}$$

where $Q(t_j)$ are orthogonal and $R(t_j, t_{j-1})$ are upper triangular with positive diagonal. So, we have the QR-factorization of $X(t_k)$,

$$X(t_k) = Q(t_k)R(t_k, t_{k-1}) \dots R(t_2, t_1)R(t_1, t_0)R(t_0). \tag{2.4}$$

If we adopt this point of view, the error we commit in finding Q is inherited from the error we do when approximating the transition matrices $\Phi(t_j, t_{j-1})$, $j = 1, 2, \dots, k$.

Notice that taking this point of view, we have expressed $R(t_k)$ as the product of local triangular transition matrices:

$$R(t_k) = R(t_k, t_{k-1}) \cdots R(t_2, t_1)R(t_1, t_0),$$

where each of these triangular transition matrices is the same as the solution of

$$\dot{R}(t, t_{j-1}) = \widetilde{B}(t)R(t, t_{j-1}),\ R(t_{j-1}, t_{j-1}) = I, \quad t_{j-1} \le t \le t_j,\ j = 1, 2, \dots, k,$$

where $\widetilde{B}$ is given in (1.3).

Now, we cannot hope to be able to obtain the exact factors $Q(t_k)$ (and $R(t_k)$). Still, let us assume that the obtained numerical approximations to the $Q(t_k)$'s, call them Q_k's, are orthogonal. The key fact, which we proved in [10], is the following: "By using either direct integration of (1.4) or having indirectly approximated Q via the discrete QR technique, with a numerical realization of the change of variables $X = QR$, we are obtaining a numerical approximation to $X(t_k)$, call it X_k, and to the triangular transition matrices $R(t_k, t_{k-1})$, call these R_k, so that we have

$$X_k = Q_k R_k R_{k-1} \dots R_2 R_1 R(t_0)\,,\ \ k = 1, 2, \dots, \tag{2.5}$$

and at the same time

$$X_k = Q(t_k)[R(t_k, t_{k-1}) + E_k] \dots [R(t_2, t_1) + E_2][R(t_1, t_0) + E_1]R(t_0)\,,\ \ k = 1, 2, \dots, \tag{2.6}$$

where $Q(t_k)$ is the exact Q-factor at t_k and the triangular transitions $R(t_j, t_{j-1})$ are also the exact ones. Moreover, the factors E_j, $j = 1, \dots, k$, are bounded in norm by the local error committed during integration of the relevant differential equations; see Theorems 3.1 and 3.16."

We will henceforth simply write

$$\|E_j\| \ \le\ \eta,\ \ j = 1, 2, \dots, \tag{2.7}$$

and stress that η is computable, in fact controllable, in terms of local error tolerances.

Furthermore, close inspection of the error terms E_j, $j = 1, \dots, k$, $k = 1, 2, \dots$, allowed us to obtain a backward error result, which we summarize below. For details, we refer to the original work (see [10, Theorem 3.12]); here we are content with a useful rephrasing of this result.

SUMMARY 2.1. *With a numerical realization of the QR change of variables, either having directly integrated* (1.4) *or indirectly through the numerical realization of the discrete QR technique, we do not obtain the exact transformation to the triangular form* (1.2)–(1.3)*, but rather find an orthogonal change of variable to the perturbed triangular system*

$$\dot{\widehat{R}} = \big(B(t) + F(t)\big)\widehat{R},\ \ t \ge 0, \tag{2.8}$$

where B is the piecewise constant (and triangular) function given by

$$B(t) := \frac{1}{t_{j+1}-t_j}\log\big(R(t_{j+1},t_j)\big), \quad t_j \le t < t_{j+1}, \tag{2.9}$$

and F is bounded as

$$\sup_{t\ge 0}\|F\| \le c\eta + O(\eta^2) =: \delta. \tag{2.10}$$

For the sake of completeness, we remark that in (2.10) the bounds on the norm of F are obtained locally, on each subinterval $[t_{j-1},t_j]$, $j=1,2,\dots$, so that one really has $\sup_{t_{j-1}\le t\le t_j}\|F\| \le c_j\eta + O(\eta^2)$, and the main contribution to the magnification factor c_j is given by the departure from normality of the exact triangular transition factors $R(t_j,t_{j-1})$. Indeed, at first order in $\mathtt{TOL}$, we have

$$\sup_{t_{j-1}\le t\le t_j}\|F\| \;\approx\; \mathtt{TOL}(1+\kappa_{j-1}h_{j-1}),\;\; h_{j-1}=t_j-t_{j-1}, \tag{2.11}$$

where $\mathtt{TOL}$ is the local error for the obtained approximation to the transition matrix, and κ_{j-1} is the departure from normality of $R(t_j,t_{j-1})$. In any case, we stress once more that the bounds on the norm of F are computable.

Remark 2.1. In order to obtain (2.10), in [10], we needed to have a certain condition satisfied; see [10, Assumption 3.5]. This amounted to the requirement that

$$\mathtt{TOL}\left[\min\left(1,\min_{1\le i\le n}\exp\left(\int_{t_{j-1}}^{t_j}B_{ii}(s)ds\right)\right)\right]^{-1} < 1.$$

In practice, this means that one may need to have the stepsizes $h_j := t_j - t_{j-1}$, $j=1,2,\dots$, sufficiently small.

3. Global error bounds for Q. Next, consider the unperturbed and perturbed triangular systems

$$\dot R = B(t)R,\; R(0)=R_0 \qquad \text{and} \qquad \dot{\widehat R} = [B(t)+F(t)]\widehat R,\; \widehat R(0)=R_0, \tag{3.1}$$

where we can assume that $\sup_{t\ge0}\|F(t)\|\le\delta$.

In [11], we proved (see Lemma 3.1 below) that there is an orthogonal change of variables, close to the identity, taking the perturbed triangular system to triangular form. The proof we gave used global (and fairly crude) norm estimates and proceeded as follows.

First, write $R = R_D + R_U$, where $R_D = \mathrm{diag}(R)$ and $R_U = \mathrm{upp}(R)$, so that $R = (I + R_U R_D^{-1})R_D =: ZR_D$. Accordingly, we have the unperturbed and perturbed diagonal systems

$$\dot R_D = D(t)R_D,\;\; \dot{\widehat R}_D = [D(t)+E(t)]\widehat R_D, \tag{3.2}$$

where $D(t)=\mathrm{diag}(B(t))$, $E = Z^{-1}FZ$, and $\widehat R = Z\widehat R_D$. Define

$$\mathrm{cond}(Z) = \sup_{t\ge0}\|Z(t)\|\cdot\|Z^{-1}(t)\|, \tag{3.3}$$

and assume that $\mathrm{cond}(Z)$ is bounded. In other words, we are assuming that Z is a Lyapunov transformation; e.g., this is certainly the case if the triangular system $\dot R = BR$ is integrally separated. This last assertion follows from [8, Theorem 5.1].

Recall that R is an *integrally separated* fundamental matrix solution if there exist $a > 0$ and $d \geq 1$ such that

$$\frac{||R(t)e_i||}{||R(s)e_i||} \cdot \frac{||R(s)e_{i+1}||}{||R(t)e_{i+1}||} \geq de^{a(t-s)} \tag{3.4}$$

for all $t, s,\ t \geq s \geq 0$, and $i = 1, 2, \ldots, n-1$. We also recall that integral separation is a generic property for linear systems (see [19]) and is a necessary and sufficient condition for stability of the Lyapunov exponents when they are distinct [1].

Next, let $\omega := \sup_{t\geq 0} \|E(t)\|$, and observe that

$$\omega \ \leq \ \sup_{t\geq 0} \|F\| \ \mathrm{cond}(Z) \ \leq \ \delta \ \mathrm{cond}(Z). \tag{3.5}$$

We make note here that the integral separation constants used in Lemmas 3.1 and 4.1 below are the integral separation constants for the piecewise constant upper triangular system that results from (2.4); see (2.9). That is, we write

$$R(t_{j+1}, t_j) = e^{(t_{j+1}-t_j)B(t)}, \quad t_j \leq t < t_{j+1},$$

where in fact

$$B(t) := \frac{1}{t_{j+1} - t_j} \log\big(R(t_{j+1}, t_j)\big), \quad t_j \leq t < t_{j+1}.$$

This piecewise constant triangular system produces the same upper triangular fundamental matrix solution as the exact upper triangular system when evaluated at mesh-points. Therefore,

$$\int_{t_j}^{t_{j+1}} B_{ii}(\tau)d\tau = \int_{t_j}^{t_{j+1}} \widetilde{B}_{ii}(\tau)d\tau, \tag{3.6}$$

where B denotes the piecewise constant triangular coefficient matrix function and $\widetilde{B}$ the exact triangular coefficient matrix function of (1.3). Thus, if $\widetilde{B}$ has integral separation with constants $\widetilde{a} > 0$ and $\widetilde{d} \geq 0$ so that for $t \geq s$ (take logarithms in (3.4))

$$\int_s^t \big(\widetilde{B}_{ii}(\tau) - \widetilde{B}_{i+1,i+1}(\tau)\big)d\tau \geq \widetilde{a}(t-s) - \widetilde{d} \tag{3.7}$$

for $i = 1, \ldots, n-1$, then for $t_{j-1} < s < t_j$ and $t_k < t < t_{k+1}$,

$$\begin{aligned} &\int_s^t \big(B_{ii}(\tau) - B_{i+1,i+1}(\tau)\big)d\tau = \int_s^t \big(\widetilde{B}_{ii}(\tau) - \widetilde{B}_{i+1,i+1}(\tau)\big)d\tau \\ &+\int_s^{t_j} \big[(B_{ii}(\tau) - \widetilde{B}_{ii}(\tau)) - (B_{i+1,i+1}(\tau) - \widetilde{B}_{i+1,i+1}(\tau))\big]d\tau \\ &+\int_{t_k}^t \big[(B_{ii}(\tau) - \widetilde{B}_{ii}(\tau)) - (B_{i+1,i+1}(\tau) - \widetilde{B}_{i+1,i+1}(\tau))\big]d\tau \geq a(t-s) - d, \end{aligned} \tag{3.8}$$

where $a = \widetilde{a}$, $d \leq \widetilde{d} + 4M_{i,i+1}h_{max}$, $M_{i,i+1} = \sup_{t\geq 0} |B_{ii}(t) - B_{i+1,i+1}(t)|$, and $h_{max} = \sup_j (t_{j+1} - t_j)$. In other words, the problem with B also has integral separation with constants $a > 0$ and $d \geq 0$.

Then the following result shows the existence of a near identity orthogonal change of variables which brings the perturbed diagonal system to upper triangular, provided

that ω is small enough. We proved this result in [11] under the assumption of integral separation of both unperturbed and perturbed triangular systems.

LEMMA 3.1 (see [11]). *Let* $\sup_{t\geq 0}\|B(t)\| = M$, *and let* a *and* d *be as defined in* (3.4). *Let* $\widehat{Q}$ *be the orthogonal factor in the QR-factorization of* $\widehat{R}_D$.

If $\omega < \omega_+(\alpha, K, M)$, *then* $|\widehat{Q}_{ij}(t)| \leq \overline{\rho}$ *for* $i \neq j$ *and all* $t \geq 0$. *Here,* $\overline{\rho} = \beta \cdot \omega$, $\beta = \alpha K$, $\alpha > 1$, $K = e^d/a$, *and*

$$\omega_+(\alpha, K, M) := \left(\sqrt{a_1^2 + 4(\alpha-1)a_2} - a_1\right)/(2a_2), \tag{3.9}$$

where $a_2 = n^2\beta^2[M\beta + 2]$ *and* $a_1 = n\beta[2M\beta + 1]$.

As an immediate consequence we have the following.

COROLLARY 3.1. *If* $\omega < \omega_+(\alpha, K, M)$ *and* $(n-1)\overline{\rho}^2 \leq 1$, *then* $\|\widehat{Q}(t) - I\| \leq \rho \equiv (n-1)(\overline{\rho} + \overline{\rho}^2)$ *and* $\|\widehat{Q}(t) - I\|_F \leq \rho_F \equiv \overline{\rho}\sqrt{2(n^2-n)}$, *where* $\|\cdot\|_F$ *is the Frobenius norm.*

Perhaps surprisingly, we already have all the ingredients to obtain global error bounds on $Q_k - Q(t_k)$.

First of all, let us look again at (2.6). In the notation of Corollary 3.1, if $\omega < \omega_+(\alpha, K, M)$, then (2.6) can be rewritten as

$$X_k = Q(t_k)Z(t_k)\widehat{Q}(t_k)U(t_k), \tag{3.10}$$

where $U(t_k)$ is upper triangular with positive diagonal elements, and $\|\widehat{Q}(t_k) - I\| \leq \rho$.

Next, let $\widehat{Z} \equiv Z(t_k)$ and $\widehat{Q} \equiv \widehat{Q}(t_k)$. Then

$$\widehat{Z}\widehat{Q} = \widehat{Z}(I + \Delta\widehat{Q}) = \widehat{Z} + \widehat{Z}\Delta\widehat{Q} =: \widehat{Z} + \Delta\widehat{Z}, \tag{3.11}$$

where $\|\Delta\widehat{Z}\| \leq \|\widehat{Z}\| \cdot \|\Delta\widehat{Q}\| \leq \|\widehat{Z}\|\rho$.

THEOREM 3.2. *With the previous notation, assume that*

1. $\omega < \omega_+(\alpha, K, M)$,
2. $\mathrm{cond}(Z)\rho < 1/2$, *and*
3. $\|\widehat{Q}(t_k) - I\| \leq \rho$.

Then we have

$$\|Q_k - Q(t_k)\| \leq \epsilon := \frac{3\mathrm{cond}(Z)\rho}{1 - 2\mathrm{cond}(Z)\rho},\ k = 0, 1, 2\ldots. \tag{3.12}$$

Proof. By the perturbation theory for the QR-factorization (e.g., see [22, Theorem 3.1]),

$$\widehat{Z} + \Delta\widehat{Z} = (I + W)(\widehat{Z} + G), \tag{3.13}$$

where $I + W$ is orthogonal, $\widehat{Z} + G$ is upper triangular with positive diagonal elements, and

$$\|W\| \leq \frac{3\|\widehat{Z}^{-1}\| \cdot \|\Delta\widehat{Z}\|}{1 - 2\|\widehat{Z}^{-1}\| \cdot \|\Delta\widehat{Z}\|}. \tag{3.14}$$

Thus, (3.10) may be written as

$$X_k = Q(t_k)(I + W)(\widehat{Z} + G)U(t_k), \tag{3.15}$$

and by the uniqueness of the QR-factorization, from (2.5), $Q_k = Q(t_k)(I + W)$ and therefore

$$\|Q_k - Q(t_k)\| = \|Q^T(t_k)Q_k - I\| = \|W\| \le \epsilon. \qquad \square \tag{3.16}$$

It is possible to improve the perturbation bounds on the QR-factors of nearby matrices; see, e.g., [5, 4] and the references therein. However, the real drawback of the global error bound in (3.12) is actually due to the fact that we have used a global transformation (via Z) to diagonal form and are thus penalized by $\mathrm{cond}(Z)$. The optimal situation of course is if $Z = I$, which occurs for instance when the upper triangular problem is in fact diagonal. In this case, we can take $\omega = \delta$ in (3.5).

However, aside from this case of $Z = I$, it is probably best to avoid altogether the diagonalizing transformation Z and tackle directly the perturbed triangular problem in (3.1), thereby attempting to bring directly $\widehat{R}$ to triangular form via an orthogonal near-the-identity transformation and obtain sharper estimates. This is what we do in the next section.

4. Handling the triangular term directly. Let us consider the perturbed triangular problem (3.1), rewritten here again as

$$\dot{\widehat{R}} = [B(t) + F(t)]\widehat{R}, \;\; t \ge 0, \;\; \widehat{R}(0) = R_0, \tag{4.1}$$

with $\sup_{t\ge 0} \|F(t)\| \le \omega$. Recall that B has upper triangular structure, and ω is small. We have $\omega = \delta$ here (see (2.10)), but we chose to use ω to unify the notation to Lemma 3.1.

Below, we show that there exists an orthogonal change of variables to the upper triangular structure, that is, a change of variables $\widehat{R} = \widehat{Q}U$ with $\widehat{Q}$ orthogonal and U upper triangular with positive diagonal, such that $\widehat{Q}$ remains, under reasonable conditions, a small perturbation of the identity given the initial condition $\widehat{Q}(0) = \widehat{Q}_0 = I$. The proof of the lemma below uses a similar technique to that used to prove Lemma 3.1 (see [11]), but much more careful estimates are now employed. In the simplest sense, Lemma 4.1 is a componentwise version of Lemma 3.1 but for a perturbed triangular system as opposed to a perturbed diagonal system. In order to obtain bounds on the entries of $\widehat{Q}$, we assume bounds on the entries of B and assume integral separation constants for both consecutive and nonconsecutive diagonal diagonal elements of B. Our bounds will be of the type $|\widehat{Q}_{ij}(t)| \le \rho_{ij}$, $i \ne j$, with $\rho_{ij} = \alpha_{ij}K_{ij}\omega$; see below. The key to obtaining this result is that the α_{ij} in this bound may be found recursively starting from $|i - j| = n - 1$ down to $|i - j| = 1$ with, for instance, $\alpha_{1,n} = \alpha_{n,1} = 2$.

LEMMA 4.1. *Consider the problem* (4.1) *and write* $B(t)+F(t) = D(t)+T(t)+F(t)$ *for all* t, *where* $D = \mathrm{diag}(B)$ *and* $T = \mathrm{upp}(B)$. *Also, let* $\sup_{t\ge 0} \|F(t)\| \le \omega$. *Then there exists an orthogonal change of variables* $\widehat{Q}$, *with* $\widehat{Q}(0) = I$, *which bring* $B + F$ *to upper triangular structure* $C := \widehat{Q}^T[B + F]\widehat{Q} - \widehat{Q}^T\dot{\widehat{Q}}$.

Moreover, let $|D_{ii}(t)| \le \kappa_{ii}$ *for* $i = 1, \ldots, n$, *let* $|T_{ij}(t)| \le \kappa_{ij}$ *for* $i < j$ *for all* $t \ge 0$, *and let* K_{ij} *be such that for all* $t \ge 0$,

$$K_{ij} \ge \int_0^t e^{-\int_\tau^t (D_{ii}(r) - D_{jj}(r))dr} d\tau, \;\; i < j, \;\; \text{and} \;\; K_{ij} = K_{ji}, \;\; i > j. \tag{4.2}$$

For $|i - j| = n - 1, |i - j| = n - 2, \ldots, |i - j| = 1$, *choose* α_{ij} *such that*

$$\alpha_{ij} > 1 + \sum_{k=j+1}^{n} \kappa_{jk}K_{ik}\alpha_{ik} + \sum_{k=1}^{i-1} K_{jk}\alpha_{jk}\kappa_{ki} \quad \text{for } i < j, \tag{4.3}$$

and let $\alpha_{ij} = \alpha_{ji}$, $i > j$.

Set

$$\omega_+^{(ij)} := \left(\sqrt{(a_1^{(ij)})^2 - 4a_0^{(ij)}a_2^{(ij)}} - a_1^{(ij)}\right)/(2a_2^{(ij)}),\ \ \omega_+ := \min_{i,j}\omega_+^{(ij)}, \tag{4.4}$$

where $a_0^{(ij)}, a_1^{(ij)}, a_2^{(ij)}$ *are defined in* (4.17).

If $\omega < \omega_+(\{\alpha_{ij}\},\{K_{ij}\},\{\kappa_{ij}\})$, *then* $|\widehat{Q}_{ij}(t)| \le \rho_{ij}$ *for* $i \ne j$ *and all* $t \ge 0$, *where* $\rho_{ij} = \alpha_{ij}K_{ij}\cdot\omega$.

Proof. Recall (see (1.4)) that $\widehat{Q}$ must satisfy $\dot{\widehat{Q}} = \widehat{Q}S(\widehat{Q}, B+F)$. So, for $i<j$ we have

$$\begin{aligned}\dot{\widehat{Q}}_{ij} &= -\widehat{Q}_{ij}[D_{ii}-D_{jj}] + \left(\widehat{Q}_{ij}[D_{ii}-D_{jj}] + e_i^T(\widehat{Q}[S(\widehat{Q},D)+S(\widehat{Q},T)+S(\widehat{Q},F)])e_j\right)\\ &=: -\widehat{Q}_{ij}[D_{ii}-D_{jj}] + q_{ij}(t,\widehat{Q},\omega)\end{aligned} \tag{4.5}$$

and a similar formula for $i>j$. We want to show that if the conditions of the theorem are satisfied and $\widehat{Q}(0)=I$, then $|\widehat{Q}_{ij}(t)| \le \rho_{ij}$ for all $i\ne j$ and $t\ge 0$. The proof involves applying [15, Theorem IV.2.1].

Using the nonlinear variation of constants formula, we have for $\widehat{Q}(0)=I$ and $i<j$

$$\widehat{Q}_{ij}(t) = \int_0^t e^{-\int_\tau^t (D_{ii}(r)-D_{jj}(r))dr} q_{ij}(\tau,\widehat{Q}(\tau),\omega)d\tau. \tag{4.6}$$

Thus, $\sup_t|\widehat{Q}_{ij}(t)| \le K_{ij}\sup_t|q_{ij}(t,\widehat{Q}(t),\omega)|$. We have

$$\begin{aligned}|q_{ij}(t,\widehat{Q},\omega)| &\le |q_{ij}(t,\widehat{Q},\omega) - q_{ij}(t,I,\omega)| + |q_{ij}(t,I,\omega)|\\ &\le \eta(\{\rho_{kl}\},\omega)\rho_{ij} + N(\omega),\end{aligned} \tag{4.7}$$

where since $S(I,D)=S(I,T)=0$ and $S(I,F)=F_L-F_L^T$, where $F_L=\mathrm{low}(F)$, $N(\omega)\le\omega$. To bound $\eta(\{\rho_{kl}\},\omega)$ write

$$\begin{aligned}q_{ij}(t,\widehat{Q},\omega) &= q_{ij}^D(t,\widehat{Q},\omega) + q_{ij}^T(t,\widehat{Q},\omega) + q_{ij}^F(t,\widehat{Q},\omega)\\ &:= \left(\widehat{Q}_{ij}[D_{ii}-D_{jj}] + e_i^T\widehat{Q}S(\widehat{Q},D)e_j\right) + e_i^T\widehat{Q}S(\widehat{Q},T)e_j + e_i^T\widehat{Q}S(\widehat{Q},F)e_j\end{aligned} \tag{4.8}$$

and consider the case in which $i<j$ (the case $i>j$ is similar).

For $q_{ij}^D(t,\widehat{Q},\omega)$ we have, from (A.3) using the notation $\beta_{ij}=\alpha_{ij}K_{ij}$,

$$\begin{aligned}&|q_{ij}^D(t,\widehat{Q},\omega) - q_{ij}^D(t,I,\omega)|\\ &\quad\le \kappa_{ii}\left[\rho_{ij}\left(\rho_{ij}^2 + 2\sum_{k=j+1}^n \rho_{ik}^2\right)\right] + \kappa_{jj}\left[\rho_{ij}\sum_{k=1,k\ne j}^n \rho_{jk}^2 + 2\sum_{k=j+1}^n \rho_{ik}\rho_{jk}\right]\\ &\qquad + \sum_{l\ne i,j}\kappa_{ll}\left[\rho_{lj}\left(\sum_{k=1}^{j-1}\rho_{ik}\rho_{lk} + \sum_{k=j+1}^n \rho_{ik}\rho_{lk}\right)\right]\end{aligned}$$

$$
\begin{aligned}
&\le \kappa_{ii}\left[\rho_{ij}\left(\rho_{ij}^2+2\sum_{k=j+1}^{n}\rho_{ik}^2\right)\right]+\kappa_{jj}\left[\rho_{ij}\sum_{k=1,k\neq j}^{n}\rho_{jk}^2+2\sum_{k=j+1}^{n}\rho_{ik}\rho_{jk}\right]\\
&\quad+\sum_{l\neq i,j}\kappa_{ll}\left[\rho_{lj}\left(\rho_{li}+\rho_{il}+\sum_{k=1,k\neq i,j,l}^{n}\rho_{ik}\rho_{lk}\right)\right]\\
&=\rho_{ij}\left\{\omega^2\left[\kappa_{ii}\left(\beta_{ij}^2+2\sum_{k=j+1}^{n}\beta_{ik}^2\right)+\kappa_{jj}\sum_{k=1,k\neq j}^{n}\beta_{jk}^2+\sum_{l\neq i,j}\kappa_{ll}\frac{\beta_{lj}}{\beta_{ij}}\sum_{k=1,k\neq i,j,l}^{n}\beta_{ik}\beta_{lk}\right]\right.\\
&\quad\left.+\omega\left[\frac{2\kappa_{jj}}{\beta_{ij}}\sum_{k=j+1}^{n}\beta_{ik}\beta_{jk}+\sum_{l\neq i,j}\kappa_{ll}\frac{\beta_{lj}}{\beta_{ij}}(\beta_{li}+\beta_{il})\right]\right\}\\
&=:\rho_{ij}\eta_{ij}^{D}=:\rho_{ij}\big(\omega^2\eta_{ij}^{D,2}+\omega\eta_{ij}^{D,1}\big).
\end{aligned}
\tag{4.9}
$$

Next, we obtain, using (A.4) and again $\beta_{ij}=\alpha_{ij}K_{ij}$,

$$
\begin{aligned}
|q_{ij}^T(t,\widehat{Q},\omega)-q_{ij}^T(t,I,\omega)|&\le\sum_{m=j+1}^{n}\kappa_{jm}\rho_{im}+\sum_{l=1}^{i-1}\rho_{jl}\kappa_{li}+\sum_{l=1,l\neq j}^{n}\sum_{m=l+1,m\neq i}^{n}\rho_{jl}\kappa_{lm}\rho_{im}\\
&\quad+\sum_{k=1,k\neq i}^{j-1}\rho_{ik}\sum_{l=1}^{n}\sum_{m=l+1}^{n}\rho_{jl}\kappa_{lm}\rho_{km}+\sum_{k=j+1}^{n}\rho_{ik}\sum_{l=1}^{n}\sum_{m=l+1}^{n}\rho_{kl}\kappa_{lm}\rho_{jm}\\
&\le\sum_{m=j+1}^{n}\kappa_{jm}\rho_{im}+\sum_{l=1}^{i-1}\rho_{jl}\kappa_{li}+\sum_{l=1,l\neq j}^{n}\sum_{m=l+1,m\neq i}^{n}\rho_{jl}\kappa_{lm}\rho_{im}\\
&\quad+\sum_{k=1,k\neq i}^{j-1}\rho_{ik}\left[\sum_{m=j+1}^{n}\kappa_{jm}\rho_{km}+\sum_{l=1}^{k-1}\rho_{jl}\kappa_{lk}+\sum_{l=1,l\neq j}^{n}\sum_{m=l+1,m\neq k}^{n}\rho_{jl}\kappa_{lm}\rho_{km}\right]\\
&\quad+\sum_{k=j+1}^{n}\rho_{ik}\left[\sum_{m=k+1}^{n}\kappa_{km}\rho_{jm}+\sum_{l=1}^{j-1}\rho_{kl}\kappa_{lj}+\sum_{l=1,l\neq k}^{n}\sum_{m=l+1,m\neq j}^{n}\rho_{kl}\kappa_{lm}\rho_{jm}\right]\\
&=\rho_{ij}\left\{\omega^2\left[\sum_{k=1,k\neq i}^{j-1}\frac{\beta_{ik}}{\beta_{ij}}\sum_{l=1,l\neq j}^{n}\sum_{m=l+1,m\neq k}^{n}\beta_{jl}\beta_{km}\kappa_{lm}\right.\right.\\
&\quad\left.+\sum_{k=j+1}^{n}\frac{\beta_{ik}}{\beta_{ij}}\sum_{l=1,l\neq k}^{n}\sum_{m=l+1,m\neq j}^{n}\beta_{kl}\beta_{jm}\kappa_{lm}\right]\\
&\quad+\omega\left[\sum_{l=1,l\neq j}^{n}\sum_{m=l+1,m\neq i}^{n}\frac{\beta_{jl}\beta_{im}\kappa_{lm}}{\beta_{ij}}+\sum_{k=1,k\neq i}^{j-1}\frac{\beta_{ik}}{\beta_{ij}}\left(\sum_{m=j+1}^{n}\kappa_{jm}\beta_{km}+\sum_{l=1}^{k-1}\beta_{jl}\kappa_{lk}\right)\right.\\
&\quad\left.\left.+\sum_{k=j+1}^{n}\frac{\beta_{ik}}{\beta_{ij}}\left(\sum_{m=k+1}^{n}\kappa_{km}\beta_{jm}+\sum_{l=1}^{j-1}\beta_{kl}\kappa_{lj}\right)\right]+1\left[\sum_{m=j+1}^{n}\kappa_{jm}\frac{\beta_{im}}{\beta_{ij}}+\sum_{l=1}^{i-1}\kappa_{li}\frac{\beta_{jl}}{\beta_{ij}}\right]\right\}\\
&=:\rho_{ij}\eta_{ij}^{T}=:\rho_{ij}\eta_{ij}^{T,2}\omega^2+\eta_{ij}^{T,1}\omega+\eta_{ij}^{T,0}.
\end{aligned}
\tag{4.10}
$$

Then, using the fact that $0<\widehat{Q}_{jj}\le 1$, and so $1-\widehat{Q}_{jj}\le 1-\widehat{Q}_{jj}^2=\sum_{k=1,k\neq j}^{n}\widehat{Q}_{jk}^2$,

we obtain from (A.5)

$$
\begin{aligned}
&|q_{ij}^F(t,\widehat{Q},\omega) - q_{ij}^F(t,I,\omega)| \\
&\le \omega\left[\sum_{k\neq j,k=1}^{n}\rho_{jk}^2 + \sum_{k\neq i,k=1}^{n}\rho_{ik}^2 + \sum_{(l,m)\neq(j,i),l,m=1}^{n}\rho_{lj}\rho_{mi} + \sum_{k=1,k\neq i,j}^{n}\rho_{ik}\right] \\
&= \rho_{ij}\left\{\omega^2\left[\sum_{k\neq j,k=1}^{n}\frac{\beta_{jk}^2}{\beta_{ij}} + \sum_{k\neq i,k=1}^{n}\frac{\beta_{ik}^2}{\beta_{ij}} + \sum_{(l,m)\neq(j,i),l,m=1}^{n}\frac{\beta_{lj}\beta_{mi}}{\beta_{ij}}\right] + \omega\sum_{k=1,k\neq i,j}^{n}\frac{\beta_{ik}}{\beta_{ij}}\right\} \\
&=: \rho_{ij}\eta_{ij}^F =: \rho_{ij}\big(\omega^2\eta_{ij}^{F,2} + \omega\eta_{ij}^{F,1}\big),
\end{aligned}
\tag{4.11}
$$

where $\beta_{ij} = \alpha_{ij}K_{ij}$.

So, we have, using (4.7), (4.8), (4.9), (4.10), and (4.11),

$$\eta_{ij}(\{\rho_{kl}\},\omega) \le \eta_{ij}^D + \eta_{ij}^T + \eta_{ij}^F, \tag{4.12}$$

and finally from (4.7) we obtain

$$\sup_t |\widehat{Q}_{ij}(t)| \le K_{ij}\big(N(\omega) + \eta_{ij}(\{\rho_{kl}\},\omega)\rho_{ij}\big).$$

Since $N(\omega) \le \omega$, Theorem IV.2.1 of [15] may be applied if

$$K_{ij}[(\eta_{ij}^D + \eta_{ij}^T + \eta_{ij}^F)\rho_{ij} + \omega] \ < \ \rho_{ij}, \tag{4.13}$$

or

$$1 > K_{ij}(\eta_{ij}^D + \eta_{ij}^T + \eta_{ij}^F) + 1/\alpha_{ij}, \tag{4.14}$$

or equivalently

$$0 > \alpha_{ij}K_{ij}(\eta_{ij}^D + \eta_{ij}^T + \eta_{ij}^F) + (1 - \alpha_{ij}). \tag{4.15}$$

Then we need to have

$$
\begin{aligned}
0 > {}& \alpha_{ij}K_{ij}(\eta_{ij}^{D,2} + \eta_{ij}^{T,2} + \eta_{ij}^{F,2})\omega^2 + \alpha_{ij}K_{ij}(\eta_{ij}^{D,1} + \eta_{ij}^{T,1} + \eta_{ij}^{F,1})\omega \\
&+ [\alpha_{ij}K_{ij}\eta_{ij}^{T,0} + 1 - \alpha_{ij}],
\end{aligned}
\tag{4.16}
$$

which we rewrite as

$$f(\omega) := a_2^{(ij)}\omega^2 + a_1^{(ij)}\omega + a_0^{(ij)} < 0, \tag{4.17}$$

where in particular

$$a_0^{(ij)} = 1 - \alpha_{ij} + \sum_{l=1}^{i-1}\kappa_{li}\alpha_{jl}K_{jl} + \sum_{m=j+1}^{n}\kappa_{jm}\alpha_{im}K_{im}. \tag{4.18}$$

We notice that $a_1^{(ij)} > 0$, $a_2^{(ij)} > 0$. Since $f'(\omega) > 0$ for $\omega > 0$, we need to have $a_0^{(ij)} < 0$ in order to be sure that there are values of ω satisfying $f(\omega) < 0$. This is guaranteed by (4.3).

Thus, if $\omega < \omega_+$ with ω_+ given in (3.9), then $|\widehat{Q}_{ij}(t)| \le \rho_{ij} \equiv \alpha_{ij}K_{ij}\omega$ for $i \neq j$ and $t \ge 0$. ∎

What one expects from Lemma 4.1 are $\rho_{ij} \equiv \alpha_{ij} K_{ij} \omega$, where ρ_{ij} are smaller for $|i-j|$ large than for $|i-j|$ small, e.g., $|i-j| = 1$. This is due in part to the sharper bounds obtained by employing $K_{ji} \equiv K_{ij} = e^{d_{ij}}/a_{ij}$ for $i < j$, where for $t \geq s$

$$\int_s^t \big(B_{ii}(\tau) - B_{jj}(\tau)\big) d\tau \geq a_{ij}(t-s) - d_{ij},$$

as opposed to

$$\int_s^t \big(B_{ii}(\tau) - B_{jj}(\tau)\big) d\tau \geq \left[\sum_{k=i}^{j-1} a_{k,k+1}\right](t-s) - \sum_{k=i}^{j-1} d_{k,k+1},$$

essentially avoiding the use of a triangular inequality. It is also, due to the form of the recursion, possible to determine the α_{ij} in (4.3) in which we choose $\alpha_{1,n} = \alpha_{n,1} > 1$, then determine α_{ij} for $|i-j| = n-2$, etc. There is the potential for the α_{ij} to become large as $|i-j| \downarrow 1$, depending on the size of the off-diagonal elements of B characterized by κ_{ij} and the strength of the integral separation between diagonal elements as characterized by K_{ij}.

The bound on the perturbation F that allows for application of Lemma 4.1 is given in (4.4) using (4.17) with $a_0^{(ij)}$ given in (4.18). The coefficients $a_1^{(ij)}, a_2^{(ij)}$ may be obtained from (4.9), (4.10), and (4.11), which give for $\beta_{ij} := \alpha_{ij} K_{ij}$

$$\begin{aligned}
a_1^{(ij)} &:= \beta_{ij}(\eta_{ij}^{D,1} + \eta_{ij}^{T,1} + \eta_{ij}^{F,1}) \\
&= \left[2\kappa_{jj} \sum_{k=j+1}^{n} \beta_{ik}\beta_{jk} + \sum_{l\neq i,j} \kappa_{ll}\beta_{lj}(\beta_{li} + \beta_{il})\right] \\
&\quad + \left[\sum_{l=1,l\neq j}^{n} \sum_{m=l+1,m\neq i}^{n} \beta_{jl}\beta_{im}\kappa_{lm} + \sum_{k=1,k\neq i}^{j-1} \beta_{ik}\left(\sum_{m=j+1}^{n} \kappa_{jm}\beta_{km} + \sum_{l=1}^{k-1} \beta_{jl}\kappa_{lk}\right)\right. \\
&\quad \left. + \sum_{k=j+1}^{n} \beta_{ik}\left(\sum_{m=k+1}^{n} \kappa_{km}\beta_{jm} + \sum_{l=1}^{j-1} \beta_{kl}\kappa_{lj}\right)\right] + \sum_{k=1,k\neq i,j}^{n} \beta_{ik}
\end{aligned} \tag{4.19}$$

and

$$\begin{aligned}
a_2^{(ij)} &:= \beta_{ij}(\eta_{ij}^{D,2} + \eta_{ij}^{T,2} + \eta_{ij}^{F,2}) \\
&= \left[\beta_{ij}\left\{\kappa_{ii}\left(\beta_{ij}^2 + 2\sum_{k=j+1}^{n} \beta_{ik}^2\right) + \kappa_{jj} \sum_{k=1,k\neq j}^{n} \beta_{jk}^2\right\} + \sum_{l\neq i,j} \kappa_{ll}\beta_{lj} \sum_{k=1,k\neq i,j,l}^{n} \beta_{ik}\beta_{lk}\right] \\
&\quad + \left[\sum_{k=1,k\neq i}^{j-1} \beta_{ik} \sum_{l=1,l\neq j}^{n} \sum_{m=l+1,m\neq k}^{n} \beta_{jl}\beta_{km}\kappa_{lm}\right. \\
&\quad \left. + \sum_{k=j+1}^{n} \beta_{ik} \sum_{l=1,l\neq k}^{n} \sum_{m=l+1,m\neq j}^{n} \beta_{kl}\beta_{jm}\kappa_{lm}\right] \\
&\quad + \left[\sum_{k\neq j,k=1}^{n} \beta_{jk}^2 + \sum_{k\neq i,k=1}^{n} \beta_{ik}^2 + \sum_{(l,m)\neq(j,i),l,m=1}^{n} \beta_{lj}\beta_{mi}\right].
\end{aligned} \tag{4.20}$$

The following theorem shows that C obtained from Lemma 4.1 is a small perturbation of the piecewise constant upper triangular B and that C may be interpreted as upper triangularizing a perturbation of A in which the perturbation is not in general small.

THEOREM 4.2. *Let $\widetilde{B}$ be the upper triangular matrix function obtained for the exact Q: $\widetilde{B} = Q^T AQ - S(Q,A)$. If we write the piecewise constant upper triangular B as $B = \widetilde{B} + \widetilde{F}$, then the upper triangular C that results from having only approximated Q satisfies for $\overline{Q} = Q\widehat{Q}$*

$$\begin{aligned} C &= \overline{Q}^T[A+G]\overline{Q} - S(\overline{Q}, A+G) \\ &= \widetilde{B} + \mathrm{upp}[\overline{Q}^T(A+G)\overline{Q} - Q^TAQ] + (\mathrm{low}[\overline{Q}^T(A+G)\overline{Q} - Q^TAQ])^T \\ &= B + \mathrm{upp}[-\widetilde{F} + \overline{Q}^T(A+G)\overline{Q} - Q^TAQ] + (\mathrm{low}[-\widetilde{F} + \overline{Q}^T(A+G)\overline{Q} - Q^TAQ])^T, \end{aligned} \tag{4.21}$$

where $G = Q[\widetilde{F} + F - S(Q,A)]Q^T$.

Moreover, if Lemma 4.1 holds, then C is an $O(\widetilde{\rho})$ perturbation of B and C is an $O(\widetilde{\rho}) + O(\|\widetilde{F}\|)$ perturbation of $\widetilde{B}$ for $\widetilde{\rho} = \max_{i\neq j}\rho_{ij}$.

Proof. We have

$$\begin{aligned} C &= \widehat{Q}^T[B+F]\widehat{Q} - S(\widehat{Q}, B+F) = \widehat{Q}^T[\widetilde{B}+\widetilde{F}+F]\widehat{Q} - S(\widehat{Q}, \widetilde{B}+\widetilde{F}+F) \\ &= \overline{Q}^T[A+G]\overline{Q} - S(\overline{Q}, A+G) = \mathrm{upp}[\overline{Q}^T(A+G)\overline{Q}] + (\mathrm{low}[\overline{Q}^T(A+G)\overline{Q}])^T \\ &= \widetilde{B} + \mathrm{upp}[\overline{Q}^T(A+G)\overline{Q} - Q^TAQ] + (\mathrm{low}[\overline{Q}^T(A+G)\overline{Q} - Q^TAQ])^T \\ &= B - \widetilde{F} + \mathrm{upp}[\overline{Q}^T(A+G)\overline{Q} - Q^TAQ] + (\mathrm{low}[\overline{Q}^T(A+G)\overline{Q} - Q^TAQ])^T \\ &= B + \mathrm{upp}[-\widetilde{F} + \overline{Q}^T(A+G)\overline{Q} - Q^TAQ] + (\mathrm{low}[\overline{Q}^T(A+G)\overline{Q} - Q^TAQ])^T \\ &= B + \mathrm{upp}[-\widetilde{F} + \overline{Q}^T(A+G)\overline{Q} - Q^TAQ] + (\mathrm{low}[-\widetilde{F} + \overline{Q}^T(A+G)\overline{Q} - Q^TAQ])^T, \end{aligned} \tag{4.22}$$

where we have used that $\mathrm{upp}[\widetilde{F}] = \widetilde{F}$ and $\mathrm{low}[\widetilde{F}] = 0$. We now show that C is a small perturbation of B.

Since $\overline{Q}^T G\overline{Q} = \widehat{Q}^T[\widetilde{F}+F]\widehat{Q} - \widehat{Q}^T S(Q,A)\widehat{Q}$, we have

$$\widehat{Q}^T\widetilde{F}\widehat{Q} = \widetilde{F} + (\widehat{Q}^T - I)\widetilde{F} + \widetilde{F}(\widehat{Q} - I) + (\widehat{Q}^T - I)\widetilde{F}(\widehat{Q} - I)$$

and

$$\widehat{Q}^T S(Q,A)\widehat{Q} = S(Q,A) + (\widehat{Q}^T - I)S(Q,A) + S(Q,A)(\widehat{Q} - I) + (\widehat{Q}^T - I)S(Q,A)(\widehat{Q} - I).$$

Since $\mathrm{upp}[S(Q,A)] + (\mathrm{low}[S(Q,A)])^T = 0$, we then obtain (i), and (ii) follows from (4.22) and the definition of G. ∎

Notice that although C results from having upper triangularized $A+G$, in general G is not small. We conclude this section with a few important remarks.

Remarks 4.1.

(a) While Lemma 3.1 (see Lemma 3.1 of [11]) together with Theorem 3.2 of [11] are in a sense analogous to a classical Bauer–Fike theorem by employing a diagonalizing transformation, Lemma 4.1 together with (4.21) may obtain sharper bounds by avoiding the use of a diagonalizing transformation.

(b) We note here that if the original problem with coefficient matrix function A is integrally separated, then (see, e.g., [1, 20, 21, 8, 9]) the $\widetilde{B}$ problem is integrally separated, and hence so is the B problem. Besides the bound on F, δ, and the measure of integral separation, the K_{ij}, the bounds obtained depend on the $\kappa_{ij} = \sup_t |B_{ij}(t)|$ for $i \le j$.

(c) An important point to make here is that we can view the exact solution as a perturbation of the computed solution as opposed to the computed solution being a perturbation of the exact solution. The bound on the norm of the perturbation, δ, is the same in either case, but by considering the exact solution as a perturbation of the computed solution, the quantities employed to bound the error, e.g., K_{ij} and κ_{ij}, may be obtained from the computed solution.

4.1. Simplified bounds and approximations. Next, we derive somewhat simplified bounds on $\|\widehat{Q}(t) - I\|$ by first taking the largest of the ρ_{ij}, then using $\rho_k = \sup_{k=|i-j|} \rho_{ij}$ and $\kappa = \sup_{i,j} \kappa_{ij} = \sup_{i,j} |B_{ij}(t)|$. In addition, we determine an asymptotic approximation for ω_+. Note that we have $\rho_{ij} = \rho_{ji}$ for $i \neq j$, so the bounds we obtain on $\|\widehat{Q}(t) - I\|$ are identical in the 1-, 2-, and ∞-norms.

COROLLARY 4.1. *In the notation of Lemma* 4.1, *let* $\delta \equiv \omega < \omega_+(\{\alpha_{ij}\}, \{K_{ij}\}, \{\kappa_{ij}\})$. *Let* $\widetilde{\rho} = \max_{i \neq j} \rho_{ij}$, *and assume that* $(n-1)(\widetilde{\rho} + \widetilde{\rho}^2) \le 1$. *Then,* $\|\widehat{Q}(t) - I\| \le \rho \equiv (n-1)(\widetilde{\rho} + \widetilde{\rho}^2)$ *and* $\|\widehat{Q}(t) - I\|_F \le \rho_F \equiv \sqrt{2(n^2 - n)}\widetilde{\rho}$ *for all* $t \ge 0$. *Moreover, for* $k = 0, 1, 2, \ldots$, $\|Q_k - Q(t_k)\| \le \rho$.

We next prove a corollary that gives a more computable bound on the error in the approximate Q while taking into account the variation in ρ_{ij} as a function of i and j. Let $\rho_k = \sup_{k=|i-j|} \rho_{ij}$, $K_k = \sup_{k=|i-j|} K_{ij}$, and $\alpha_k = \sup_{k=|i-j|} \alpha_{ij}$, and assume that $\kappa = \sup_{i,j} \kappa_{ij}$.

COROLLARY 4.2. *If the assumptions of Lemma* 4.1 *are satisfied,* $K_{n-1} \le K_{n-2} \le \cdots \le K_1$, *and* $\sum_{k=1}^{n-1} \rho_k + \rho_k^2 \le 1$, *then* $\rho_j \le \alpha_j K_j \omega$, $\|\widehat{Q}(t) - I\|_F \le (2\sum_{k=1}^{n-1}(n-k)\rho_k^2 + \sum_{i=1}^{n}(\sum_{k=1}^{i-1}\rho_k^2 + \sum_{i=1}^{n-i}\rho_k^2)^2)^{1/2}$, *and*

$$\|\widehat{Q}(t) - I\| \le \begin{cases} 2\displaystyle\sum_{k=1}^{(n-1)/2} (\rho_k + \rho_k^2), & n \text{ odd}, \\ 2\displaystyle\sum_{k=1}^{(n-2)/2} (\rho_k + \rho_k^2) + (\rho_{n/2} + \rho_{n/2}^2), & n \text{ even}, \end{cases} \tag{4.23}$$

where $\alpha_{n-1} > 1$ *and*

$$\alpha_j > 1 + (n-j-1)\kappa K_{j+1}\big(1 + (n-j-2)\kappa K_{j+2}(1 + \cdots (1 + \kappa K_{n-1}\alpha_{n-1}))\cdots\big) \tag{4.24}$$

for $j = n-2, \ldots, 1$.

Proof. The proof involves recursively applying the condition (4.3) for $i < j$ and $k = |i-j| = n-1, \ldots, 1$. This means we can choose α_k so that

$$\begin{aligned} &\alpha_{n-1} > 1, \ \ \alpha_{n-2} > 1 + \kappa K_{n-1}\alpha_{n-1} > \alpha_{n-1}, \\ &\alpha_{n-3} > 1 + 2\kappa K_{n-2}\alpha_{n-2} > \alpha_{n-2}, \ldots \end{aligned} \tag{4.25}$$

or in general

$$\alpha_{j-1} > 1 + (n-j)\kappa K_j \alpha_j > \alpha_j, \ \ j = n-1, \ldots, 1. \tag{4.26}$$

The result then follows by recalling that $\rho_j \leq \alpha_j K_j \omega$ and obtaining the bounds on the matrix norms in a straightforward fashion using

$$\|\widehat{Q}(t) - I\| \leq \sqrt{\|\widehat{Q}(t) - I\|_1 \|\widehat{Q}(t) - I\|_\infty}. \qquad \square$$

In spite of their appearances, Corollaries 4.1 and 3.1 are very different, even if in Corollary 3.1 we happen to have $\omega = \delta$ (as when $Z = I$). Moreover, the factors ω_+, as well as ρ, are different in these two contexts; notice the use of $\bar{\rho}$ in Corollary 3.1 versus the use of $\widetilde{\rho}$ in Corollary 4.1.

To estimate ω_+, we reason as follows. An asymptotic analysis of the terms in (4.20), (4.19), and (4.18) that contribute to a_2, a_1, and a_0 in (4.16) and (4.17) which determine ω_+ in (4.4) suggest that

$$a_2 \approx (3\kappa_D + \kappa)\alpha_1^3 K_1^3, \;\; a_1 \approx (2\kappa_D + \kappa)\alpha_1^2 K_1^2, \;\; a_0 \approx 1 - \alpha_1 + 2\kappa\alpha_2 K_2 < 0, \tag{4.27}$$

where $\kappa_D = \sup_{i,t} |B_{ii}(t)|$.

These coefficients were determined by first observing that the dominant term for $a_2 \equiv \beta_{ij}(\eta_{ij}^{D,2} + \eta_{ij}^{T,2} + \eta_{ij}^{F,2})$ in (4.20), where $\beta_{ij} = \alpha_{ij} K_{ij}$, is proportional to $\alpha_1^3 K_1^3$. Then in (4.20) we obtain the term $3\kappa_D$ from $\kappa_{ii}\beta_{ij}^3 + \kappa_{jj}\beta_{ij}(\beta_{j,j-1} + \beta_{j,j+1})$ when $j = i+1$ and the term κ is obtained from $\beta_{ik}\beta_{jl}\beta_{km}\kappa_{lm}$ when $k = i+1, l = j-1, m = i+2$, and $j = i+2$. There are no terms in $\eta_{ij}^{F,2}$ proportional to $\alpha_1^3 K_1^3$. To determine the dominant terms in a_1 observe that the term $2\kappa_D$ is obtained from $\kappa_{ll}\beta_{lj}(\beta_{li} + \beta_{il})$ in (4.19) when $l = j-1 = i+1$ and the term κ is obtained from $\beta_{jl}\beta_{im}\kappa_{lm}$ in (4.19) when $l = j-1$, $m = i+1$, and $j = i+1$. The dominant terms in $\eta_{ij}^{F,1}$ are not proportional to $\alpha_1^2 K_1^2$. The approximation of a_0 is found by considering (4.18) when $j = i+1$.

Using the form for ω_+ in (4.4) and the approximation $\sqrt{1+x} \approx 1 + \frac{x}{2}$, we have $\omega_+ \approx -\frac{a_0}{a_1} \approx C[(2\kappa_D + \kappa)\alpha_1^2 K_1^2]^{-1}$, where $C \approx -a_0$.

Notice that the ρ_{ij} we have when treating the triangular term directly appears to decrease as $|i-j|$ grows. There is an accumulation for $|i-j|$ small, but it looks friendlier than the accumulation to find $\mathrm{cond}(Z)$ when using the diagonalizing transformation. Indeed, it is interesting to compare the bounds one obtains with the two different approaches: (i) using the diagonalizing transformation Z, and (ii) working directly with the triangular system. We do this below on a two-dimensional system.

4.2. Comparison for two-dimensional systems. Here we compare the two global error bounds obtained by the two different approaches we examined: (i) using the diagonalizing transformation Z (see section 3), and (ii) dealing directly with the triangular coefficient matrix function (call this the *triangular approach*) for a two-dimensional system. That is, we have $B = \left(\begin{smallmatrix} D_{11} & T_{12} \\ 0 & D_{22} \end{smallmatrix}\right)$.

Of course, a bound on the error in Q using the diagonalizing transformation approach is given by (3.12), while with the triangular approach it is given in Corollary 4.1. The interesting thing is to see what bounds we need for ω in the two cases.

We assume $|D_{ii}(t)| \leq M$ for $i = 1, 2$ and $|T_{12}(t)| \leq \kappa_{12}$ for all t and use the bound $||Z(t)||_F, ||Z^{-1}(t)||_F \leq \sqrt{2 + \kappa_{12}^2 K^2}$, where $K = e^d/a$ (see Lemma 3.1).

Then we have a bound $|\sin(\theta(t))| \leq \rho := 2K\omega$, where (i) $|\sin(\theta(t))| = |\widehat{Q}_{12}(t)|$ and $\omega = \delta \mathrm{cond}(Z)$ with the diagonalizing transformation approach, and (ii) $|\sin(\theta(t))| = |\widehat{Q}_{12}(t)|$ and $\omega = \delta$ with the triangular approach, provided that in these two cases we have

$$\text{(i)} \quad \omega < \omega_+^D := \frac{\sqrt{1 + 4MK} - 1}{2K(2MK + 1)} \cdot \frac{1}{\mathrm{cond}(Z)} \tag{4.28}$$

and

$$\text{(ii)} \quad \omega < \omega_+^T \equiv \frac{\sqrt{1+4M/(K\kappa_{12}^2)}-1}{8MK/\kappa_{12}}, \tag{4.29}$$

respectively.

Since for $x > 0$, $1 + x/2 - x^2/8 \le \sqrt{1+x} \le 1 + x/2$, using $\mathrm{cond}(Z) \le 2 + K^2\kappa_{12}^2$, we have

$$\frac{M(1-MK)}{(2MK+1)\mathrm{cond}(Z)} < \omega_+^D < \frac{M}{(2MK+1)\mathrm{cond}(Z)} < \frac{1}{2K\mathrm{cond}(Z)} \tag{4.30}$$

and

$$\begin{aligned} \frac{1}{4K^2\kappa_{12}}\left[1-\frac{M}{K\kappa_{12}^2}\right] &= \frac{1-M/(K\kappa_{12}^2)}{4K^2\kappa_{12}} < \omega_+^T < \frac{M/(K\kappa_{12}^2)}{4MK/\kappa_{12}} \\ &= \frac{1}{4K^2\kappa_{12}} \le \frac{1}{4K\sqrt{\mathrm{cond}(Z)-2}}. \end{aligned} \tag{4.31}$$

Quite clearly, the triangular approach gives much improved bounds.

4.3. Bounding the K_{ij}'s. In light of Remark 4.1(c), besides the size of the perturbation of B, the quantities needed to apply Lemma 4.1, the κ_{ij}'s that measure the nonnormality and the K_{ij}'s that measure the integral separation, can be obtained from the computed solution. To this end we consider now how to obtain bounds for the K_{ij}'s. We consider two approaches.

The first approach follows ideas developed in [8] using Steklov averages; see also Adrianova [1, Lemma 5.4.1]. In particular, for $i < j$ set

$$p(t) := B_{ii}(t) - B_{jj}(t)$$

and for some $H > 0$ consider finding the quantities $c(H)$ (positive) and M:

$$c(H) = \inf_t \frac{1}{H}\int_t^{t+H} p(r)dr > 0, \quad M = \inf_t p(t). \tag{4.32}$$

We have the following result.

LEMMA 4.3. *For $a := c(H) > 0$, and a positive integer N, let*

$$d := \left\{\sum_{k=1}^{N}(c(H) - c(H/2^k))\cdot(H/2^k)\right\} + (c(H) - M)\cdot(H/2^N) \ge 0.$$

Then, for $t \ge s$, we have

$$\int_s^t p(r)dr \ge a(t-s) - d. \tag{4.33}$$

Proof. That $d \ge 0$ follows from

$$c(H) \ge c(H/2) \ge c(H/4) \ge \cdots \ge c(H/2^N) \ge M.$$

If $t - s = jH$ for some positive integer j, then

$$\int_s^t p(r)dr \ge c(H)\cdot(t-s) \ge a(t-s) - d$$

since $d \geq 0$. Otherwise, $t-s = jH + (t-s-jH)$ with $0 < (t-s-jH) < H$, and for $\gamma_k \in \{0,1\}$ we can write

$$t-s-jH = \left\{ H \sum_{k=1}^{N} \gamma_k / 2^k \right\} + x,$$

where $0 \leq x < H/2^N$.

$$\begin{aligned}\int_s^t p(r)dr &\geq jH \cdot c(H) + \left\{ H \sum_{k=1}^{N} \gamma_k 2^k \cdot c(H/2^k) \right\} + x \cdot M \\ &= c(H) \cdot (t-s) + (jH - (t-s))c(H) + \left\{ H \sum_{k=1}^{N} \gamma_k / 2^k \cdot c(H/2^k) \right\} + x \cdot M \\ &\geq a(t-s) - d. \qquad \square\end{aligned}$$

(4.34)

Recalling (4.2), based on the above lemma, we may use

$$K_{ij} = e^d / a.$$

Of course, the K_{ij}'s are still functions of H: $K_{ij}(H)$. The idea now is to use for K_{ij} the minimum value of $K_{ij}(H)$ subject to maintaining $c(H) > 0$.

We also develop an alternative approach that is a simplification of Lemmas 4.1 and 4.2 of [17]. This alternative approach may yield better bounds on the K_{ij}'s.

As before, for all t, let $p(t) = B_{ii}(t) - B_{jj}(t)$, $i < j$. Consider a discretization of the interval $[0,T]$:

$$0 = t_0 < t_1 < \cdots < t_N = T.$$

LEMMA 4.4. *Let $\epsilon > 0$ be given. There exists $a_k > 0$ and $d_k \geq 0$ such that for $t_k \leq s \leq t_{k+1}$,*

$$\int_s^{t_{k+1}} p(r)dr \geq a_k(t_{k+1} - s) - d_k, \tag{4.35}$$

where for $h_k = t_{k+1} - t_k$ and for

$$Y_k = \min_{t_k \leq s \leq t_{k+1}} \frac{1}{t_{k+1} - s} \int_s^{t_{k+1}} p(r)dr, \tag{4.36}$$

$$a_k = \begin{cases} \epsilon h_k^{-1}, & h_k Y_k < \epsilon, \\ Y_k, & h_k Y_k \geq \epsilon, \end{cases} \qquad d_k = \begin{cases} \epsilon - h_k Y_k, & h_k Y_k < \epsilon, \\ 0, & h_k Y_k \geq \epsilon, \end{cases} \tag{4.37}$$

and

$$\int_0^T e^{-\int_s^T p(r)dr} ds \leq \sum_{k=0, d_k=0}^{N-1} e^{-\left[\sum_{l=k+1}^{N-1} X_l\right]} \frac{(1 - e^{-Y_k h_k})}{Y_k} + \frac{(e^\epsilon - 1)}{\epsilon} \tag{4.38}$$

$$\times \sum_{k=0, a_k = \epsilon h_k^{-1}}^{N-1} e^{-\left[\sum_{l=k+1}^{N-1} X_l\right]} h_k e^{-h_k Y_k} =: K(T), \tag{4.39}$$

where

$$X_k = \int_{t_k}^{t_{k+1}} p(r)dr.$$

Proof. If $h_k Y_k < \epsilon$, then for $a_k = \epsilon/h_k$, $d_k = \epsilon - h_k Y_k \geq 0$, and $s \in [t_k, t_{k+1}]$,

$$\int_s^{t_{k+1}} p(r)dr \geq (t_{k+1}-s)Y_k = (t_{k+1}-s)\left[\frac{\epsilon}{h_k} - \frac{(\epsilon - h_k Y_k)}{h_k}\right] \geq \frac{\epsilon}{h_k}(t_{k+1}-s) - (\epsilon - h_k Y_k). \tag{4.40}$$

If $h_k Y_k \geq \epsilon$, then for $a_k = Y_k$, $d_k = 0$, and $s \in [t_k, t_{k+1}]$,

$$\int_s^{t_{k+1}} p(r)dr \geq (t_{k+1} - s)Y_k = a_k(t_{k+1} - s). \tag{4.41}$$

The proof of (4.38) is then a direct consequence of the estimate

$$\begin{aligned}
\int_0^T e^{-\int_s^T p(r)dr} ds &= \sum_{k=0}^{N-1} \int_{t_k}^{t_{k+1}} e^{-\int_s^T p(r)dr} ds \\
&= \sum_{k=0}^{N-1} \int_{t_k}^{t_{k+1}} e^{-\left[\int_s^{t_{k+1}} p(r)dr + \sum_{l=k+1}^{N-1} \int_{t_l}^{t_{l+1}} p(r)dr\right]} ds \\
&\leq \sum_{k=0}^{N-1} e^{-\left[\sum_{l=k+1}^{N-1} X_l\right]} \int_{t_k}^{t_{k+1}} e^{-a_k(t_{k+1}-s)+d_k} ds \\
&= \sum_{k=0}^{N-1} e^{-\left[\sum_{l=k+1}^{N-1} X_l\right]} \frac{e^{d_k}}{a_k}(1 - e^{-a_k h_k}). \qquad \square
\end{aligned}$$

Observe that the bounds (4.38) can be used to obtain bounds on the K_{ij}'s in (4.2) by setting $K_{ij} = \sup_k K(t_k)$.

5. Example. Here we illustrate our results, in particular the effectiveness of the bounds on the error in the orthogonal matrix function Q, in the following example, also considered in [11]. We report only on the improved bounds obtained when handling directly the triangular term.

Let $B(t) = D(t) + U(t)$ be the upper triangular matrix function with

$$D(t) = \mathrm{diag}(D_{11}(t), D_{22}(t), D_{33}(t), D_{44}(t)), \tag{5.1}$$

where we take $D_{11}(t) = 10 + \sin(t)$, $D_{22}(t) = \zeta\cos(t)$, $D_{33}(t) = \lambda - \zeta\cos(t)$, $D_{44}(t) = -10 + \sin(t)$, $\zeta > 0$, and

$$U(t) = \kappa \begin{pmatrix} 0 & \cos(t) & \sin(t) & \cos(t) \\ 0 & 0 & \cos(t) & \sin(t) \\ 0 & 0 & 0 & \cos(t) \\ 0 & 0 & 0 & 0 \end{pmatrix}. \tag{5.2}$$

The parameter κ changes the degree to which there is nonnormality in the upper triangular part, and the parameters λ and ζ determine the degree of integral separation in the system. For simplicity, in our experiments below we fix $\lambda = -5$, and all computations refer to this case.

Form the matrix function

$$A(t) = Q(t)B(t)Q^T(t) + \dot{Q}(t)Q^T(t),$$

TABLE 5.1

Error in the approximate Q varying the degree of nonnormality and integral separation, method, and tolerance.

$T = 10^4$, $\lambda = -5$, TOL = 1.E-6.

κ	ζ	Method	Error	Method	Error
0	1	Cont QR	6E-7	Disc QR	7E-8
0	2	Cont QR	6E-7	Disc QR	1E-7
0	4	Cont QR	3E-6	Disc QR	5E-7
1	1	Cont QR	6E-7	Disc QR	7E-8
1	2	Cont QR	6E-7	Disc QR	1E-7
1	4	Cont QR	2E-6	Disc QR	5E-7
10	1	Cont QR	6E-7	Disc QR	1E-7
10	2	Cont QR	6E-7	Disc QR	2E-7
10	4	Cont QR	9E-6	Disc QR	5E-7
100	1	Cont QR	5E-6	Disc QR	4E-7
100	2	Cont QR	7E-6	Disc QR	4E-7
100	4	Cont QR	1E-4	Disc QR	2E-6

where

$$Q(t) = \mathrm{diag}(1, Q_\beta(t), 1) \cdot \mathrm{diag}(Q_\eta(t), Q_\eta(t))$$

and

$$Q_\gamma(t) = \begin{pmatrix} \cos(\gamma t) & \sin(\gamma t) \\ -\sin(\gamma t) & \cos(\gamma t) \end{pmatrix}, \ \eta = 1, \ \beta = \sqrt{2}.$$

Results for this problem were obtained using the code LESLIS, which we developed (see www.math.gatech.edu/∼dieci and www.math.ku.edu/∼evanvleck).

In particular, we employ the continuous QR method (Cont QR in Table 5.1) using the projected fifth order scheme (IPAR(8)=0 in LESLIS) with local error control on the orthogonal factor Q (IPAR(10)=1 in LESLIS), and the discrete QR method (Disc QR in Table 5.1) with a fifth order scheme (IPAR(8)=4 in LESLIS), with local error control on the Lyapunov exponents (IPAR(10)=0 in LESLIS). TOL is the value of the local error tolerance, and we used TOL = 1.E-6 throughout. Before reporting on the results, we remark that (see (2.11)) we expect to have ω to be about the same as TOL. In other words, the bound on the norm of the perturbation term F in Lemma 4.1 is essentially TOL. In Table 5.1 we tabulate the actual error for different methods and varying κ and ζ values. We report on the error in Q in the two norm (the largest singular value of the error) at grid-points. Exponential notation is used throughout.

We further compare the actual error with the error bounds obtained in the previous sections. Although the quantities needed to determine the bounds, $\rho_{ij} = \alpha_{ij} K_{ij} \omega$, on $|\widehat{Q}_{ij}(t)|$, in particular the α_{ij}, are somewhat difficult to give in closed form, the recursion (4.3) is straightforward to code. Likewise, the formula for ω_+, given by (4.4) with (4.18), (4.19), and (4.20), are functions of the K_{ij}, the measure of integral separation, and the κ_{ij}, the measure of nonnormality. For this problem, we have $\kappa_{11} = \kappa_{44} = 11$, $\kappa_{22} = \zeta$, $\kappa_{33} = |\lambda| + \zeta$, and $\kappa_{ij} = \kappa$ for $i < j$.

We have, for $\lambda = -5$ and $\zeta < \sqrt{24}$, $K_{12} = K_{24} = 1/(10 - \sqrt{1+\zeta^2})$, $K_{13} = 1/(15 - \sqrt{1+\zeta^2})$, $K_{34} = 1/(5 - \sqrt{1+\zeta^2})$, and $K_{14} = 1/20$. If $\zeta < 5/2$, then $K_{23} = 1/(5 - 2\zeta)$. We next focus on determining bounds on K_{23} using Lemmas 4.3 and 4.4 when $\zeta \geq 5/2$.

TABLE 5.2
Bounds on K_{23} obtained using Lemmas 4.3 and 4.4.

K_{23} bounds for $T = 10^4$, $\lambda = -5$, and $\zeta = 4$.

N	K_{23} w/ Lemma 4.3	K_{23} w/ Lemma 4.4 ($h = \pi/2^N, \epsilon = 10^{-8}$)
0	2.1E5	3.9E4
1	4.7E4	2.1E2
2	3.9E4	9.0E1
3	3.8E4	5.1E1
4	3.8E4	4.1E1
5	3.8E4	3.9E1
6	3.8E4	3.8E1

We have $p(t) := D_{22}(t) - D_{33}(t) = 5 + 2\zeta\cos(t)$. To employ Lemma 4.3 we have

$$K_{23}(H) = \frac{e^{d(H)}}{c(H)}, \quad d(H) = 4\zeta\left\{H/2^{N+1} + \sum_{k=1}^{N}\sin(H/2^{k+1}) - \sin(H/2)\right\} \tag{5.3}$$

for H sufficiently large so that $c(H) \equiv 5 - 4\zeta\sin(H/2)/H > 0$. For different values of N we can determine, at least approximately, optimal values of H. When using Lemma 4.4, and in the notation used there, we set $h \equiv h_k = \pi/2^N$ for integer $N \geq 0$. When $p(t)$ is decreasing, $Y_k = 5 + 2\zeta\cos(t_{k+1})$, and when $p(t)$ is increasing, $Y_k = 5 + 2\zeta[\sin(t_{k+1}) - \sin(t_k)]/h$.

In Table 5.2 we tabulate bounds on K_{23} found using Lemmas 4.3 and 4.4 for different values of N. In the case of Lemma 4.3, N refers to discretization of the Steklov window length as in (5.3), while in the case of Lemma 4.4, N refers to the fineness of the discretization of $[0, T]$ ($h = \pi/2^N$). We will use the best of the bounds found to determine bounds on the error in Q.

By Corollary 4.2 we have $\|Q(t_k) - Q_k\| \approx 2\rho_1 + \rho_2$, and we recall that $\rho_i \leq \alpha_i K_i \omega$, $i = 1, 2$. For convenience, in Table 5.3 we report on the estimates for ρ_1 and ρ_2 obtained by using these bounds and those for α_i, K_i, $i = 1, 2$, for different values of κ and ζ.

At this point, we make an important observation: *If we use $\omega =$ `TOL` in the expressions obtained for ρ_1 and ρ_2 in Table* 5.3*, and use $2\rho_1 + \rho_2$ as a measure of the true global error, then we observe that the observed true global error is always below the theoretical estimate.* This is actually an important fact, since it is not easy to know exactly the value of ω. At the same time, this is not entirely rigorous, since we can use $2\rho_1 + \rho_2$ as a measure of the global error only if $\omega < \omega_+$. For this reason, we also estimated ω_+ using the approximations to a_2, a_1, a_0 in (4.27).

For the example we are considering, $\kappa_D := \sup_{i,t}|B_{ii}(t)| = 11$. In the last column of Table 5.3 we record this estimated value of $[(2\kappa_D + \kappa)\alpha_1^2 K_1^2]^{-1}$, which serves as the estimate of ω_+ (see the discussion after (4.27)).

Now, by comparison of the true errors obtained in Table 5.1, using the estimates for ω_+ from Table 5.3, and still adopting the rationale that $\omega =$ `TOL`, we see that the results for $\kappa = 0$, $\kappa = 1$, or $\kappa = 10$ and $\zeta \neq 4$ in Table 5.3 validate that $\omega < \omega_+$. To further corroborate this fact, we notice that for these value of κ, the error bounds obtained are in good agreement with the actual errors observed and recorded in Tables 5.1 and 5.3. For $\kappa = 10^2$ or $\kappa = 10$ and $\zeta = 4$ the bounds obtained do not appear to be sharp, in the sense that it becomes increasingly difficult to satisfy the hypothesis $\omega < \omega^+$ that is necessary to apply Lemma 4.1 and Corollary 4.2.

TABLE 5.3
Bound on the error in Q is dominated by $\rho_1 = \alpha_1 K_1 \omega$.

Bounds obtained for the example with $\lambda = -5$.

κ	ζ	ρ_1/ω	ρ_2/ω	$(2\rho_1+\rho_2)/\omega$	$\omega_+ \approx [(2\kappa_D+\kappa)\alpha_1^2 K_1^2]^{-1}$
0	1	6.6E-1	2.3E-1	1.6E0	8.5E-2
0	2	2.0E0	2.6E-1	4.3E0	2.6E-2
0	4	7.4E1	3.4E-1	1.5E2	2.7E-4
1	1	8.0E-1	2.4E-1	1.9E0	6.5E-2
1	2	2.4E0	2.7E-1	5.1E0	1.9E-2
1	4	9.4E1	3.6E-1	1.9E2	1.1E-4
10	1	2.6E0	3.5E-1	5.6E0	8.3E-3
10	2	8.2E0	3.9E-1	1.7E1	1.3E-3
10	4	3.6E2	5.1E-1	7.2E2	7.6E-7
10^2	1	7.7E1	1.4E0	1.6E2	1.6E-6
10^2	2	2.5E2	1.5E0	5.0E2	1.6E-7
10^2	4	1.2E4	2.0E0	2.4E4	7.4E-11

6. Conclusions and consequences. We have provided a global error analysis for the factor Q in the change of variables $X = QR$ of a fundamental matrix solution for a nonautonomous linear system. Our basic technique consists of a combination of backward and forward error analyses, and this is seemingly a new approach in the context of numerical integration.

Several comments are in order.

(1) Among the noteworthy consequences of a global error analysis for Q, probably the most important one is that it becomes simple to obtain global error statements for all quantities which derive from the simplified triangular structure of the linear system, e.g., error bounds on the Lyapunov exponents and/or the Sacker–Sell spectrum.

(2) We have made our analysis for the entire fundamental matrix solution, although at times one is interested in computing the QR-factorization of only a few of the columns of the fundamental matrix solution: *reduced* QR-factorization. However, dealing with this case is—in principle—simple, since as a consequence of the combined backward and forward error analyses for all k we have

$$\begin{aligned} Q_k &= Q(t_k)\widetilde{Q}(t_k), \\ &\quad R_k R_{k-1} \dots R_2 R_1 R(t_0) \\ &= \widetilde{Q}^T(t_k)[R(t_k,t_{k-1}) + E_k] \dots [R(t_2,t_1) + E_2][R(t_1,t_0) + E_1]R(t_0) \\ &\equiv \widetilde{R}(t_k,t_{k-1}) \dots \widetilde{R}(t_2,t_1)\widetilde{R}(t_1,t_0)R(t_0). \end{aligned} \tag{6.1}$$

Thus, multiplying by $\left(\begin{smallmatrix} I & 0 \\ 0 & 0 \end{smallmatrix}\right)$, one obtains backward error statements and subsequent forward error estimates for the case of the reduced QR-factorization, obtaining bounds which cannot be worse than the bounds obtained on the full Q.

(3) We remark that in this work we made use of the assumption of integral separation for the linear system. This is reasonable, since this is the generic case. However, it would be interesting to obtain direct bounds also removing this assumption, in a similar way to [11, Theorem 4.3].

(4) For the general case of integration of a matrix equation of the form

$$\dot{y} = f(t,y), \;\; y(t_0) = y_0 \;\text{ orthogonal}$$

we note that in order for the solution to remain orthogonal $y^T\dot{y}$ must be skew-symmetric; hence $f(t,y) = y\ s(t,y)$, where $s(t,y)$ is skew-symmetric. Obtaining a result similar in flavor to Lemma 4.1 would require determining a splitting

$$\dot{y} = yL + y[s(t,y) - L], \;\; y(t_0) = I,$$

where L is such that $y[s(t,y) - L]$ remains small enough for $y(t) \approx I$. In the specific case considered here the splitting was motivated by integral separation, a natural property for linear nonautonomous differential equations $\dot{x} = A(t)x$, and we took advantage of the fact that the skew-symmetric matrix function $S(Q,A)$ is linear in A.

Appendix. Here we derive some technical expressions that are useful in proving Lemma 4.1. In particular, we derive expressions relative to $q_{ij}^D(t,\widehat{Q},\omega)$, $q_{ij}^T(t,\widehat{Q},\omega)$, and $q_{ij}^F(t,\widehat{Q},\omega)$ in (4.8) that are useful in obtaining the bounds (4.9), (4.10), (4.11) in the proof of Lemma 4.1.

First, consider $q_{ij}^D(t,\widehat{Q},\omega)$. We have, writing $\widehat{Q}(t) = [\widehat{Q}_1(t)|\cdots|\widehat{Q}_n(t)]$,

$$\begin{aligned}
&q_{ij}^D(t,\widehat{Q},\omega) - q_{ij}^D(t,I,\omega)\\
&= q_{ij}^D(t,\widehat{Q},\omega) = \widehat{Q}_{ij}[D_{ii} - D_{jj}] + \left[-\sum_{k=1}^{j-1} \widehat{Q}_{ik}\widehat{Q}_k^T + \sum_{k=j+1}^{n} \widehat{Q}_{ik}\widehat{Q}_k^T \right] D\widehat{Q}_j\\
&= \widehat{Q}_{ij}[D_{ii} - D_{jj}] + \sum_{l=1}^{n} D_{ll}\left[\widehat{Q}_{lj} \cdot \left\{ -\sum_{k=1}^{j-1} \widehat{Q}_{ik}\widehat{Q}_{lk} + \sum_{k=j+1}^{n} \widehat{Q}_{ik}\widehat{Q}_{lk} \right\} \right]\\
&= D_{ii}\left[\widehat{Q}_{ij} \cdot \left\{ 1 - \sum_{k=1}^{j-1} \widehat{Q}_{ik}^2 + \sum_{k=j+1}^{n} \widehat{Q}_{ik}^2 \right\} \right]\\
&\quad + D_{jj}\left[-\widehat{Q}_{ij} + \widehat{Q}_{jj} \cdot \left\{ -\sum_{k=1}^{j-1} \widehat{Q}_{ik}\widehat{Q}_{jk} + \sum_{k=j+1}^{n} \widehat{Q}_{ik}\widehat{Q}_{jk} \right\} \right]\\
&\quad + \sum_{l\neq i,j} D_{ll}\left[\widehat{Q}_{lj} \cdot \left\{ -\sum_{k=1}^{j-1} \widehat{Q}_{ik}\widehat{Q}_{lk} + \sum_{k=j+1}^{n} \widehat{Q}_{ik}\widehat{Q}_{lk} \right\} \right].
\end{aligned} \tag{A.1}$$

By orthogonality we have

$$1 = \sum_{k=1}^{n} \widehat{Q}_{ik}^2 \quad \text{and} \quad \widehat{Q}_{ij}\widehat{Q}_{jj} = -\sum_{k\neq j} \widehat{Q}_{ik}\widehat{Q}_{jk}, \tag{A.2}$$

so for $i < j$,

$$\begin{aligned}
q_{ij}^D(t,\widehat{Q},\omega) &= D_{ii}\left[\widehat{Q}_{ij} \left(\widehat{Q}_{ij}^2 + 2\sum_{k=j+1}^{n} \widehat{Q}_{ik}^2 \right) \right]\\
&\quad + D_{jj}\left[-\widehat{Q}_{ij}(1 - \widehat{Q}_{jj}^2) + 2\widehat{Q}_{jj} \sum_{k=j+1}^{n} \widehat{Q}_{ik}\widehat{Q}_{jk} \right]\\
&\quad + \sum_{l\neq i,j} D_{ll}\left[\widehat{Q}_{lj} \left(-\sum_{k=1}^{j-1} \widehat{Q}_{ik}\widehat{Q}_{lk} + \sum_{k=j+1}^{n} \widehat{Q}_{ik}\widehat{Q}_{lk} \right) \right].
\end{aligned} \tag{A.3}$$

Next, consider the term $q_{ij}^T(t,\widehat{Q},\omega)$. We have for $i<j$ (and similarly for $i>j$)

$$\begin{aligned}
q_{ij}^T(t,\widehat{Q},\omega) - q_{ij}^T(t,I,\omega) &= q_{ij}^T(t,\widehat{Q},\omega) = \sum_{k=1}^{n} \widehat{Q}_{ik}(S(\widehat{Q},T))_{kj} \\
&= -\sum_{k=1}^{j-1} \widehat{Q}_{ik}(\widehat{Q}_j^T T \widehat{Q}_k) + \sum_{k=j+1}^{n} \widehat{Q}_{ik}(\widehat{Q}_k^T T \widehat{Q}_j) \\
&= -\sum_{k=1}^{j-1} \widehat{Q}_{ik} \sum_{l=1}^{n}\sum_{m=l+1}^{n} \widehat{Q}_{jl} T_{lm} \widehat{Q}_{km} + \sum_{k=j+1}^{n} \widehat{Q}_{ik} \sum_{l=1}^{n}\sum_{m=l+1}^{n} \widehat{Q}_{kl} T_{lm} \widehat{Q}_{jm} \\
&= -\widehat{Q}_{ii} \sum_{l=1}^{n}\sum_{m=l+1}^{n} \widehat{Q}_{jl} T_{lm} \widehat{Q}_{im} \\
&\quad - \sum_{k=1,k\neq i}^{j-1} \widehat{Q}_{ik} \sum_{l=1}^{n}\sum_{m=l+1}^{n} \widehat{Q}_{jl} T_{lm} \widehat{Q}_{km} + \sum_{k=j+1}^{n} \widehat{Q}_{ik} \sum_{l=1}^{n}\sum_{m=l+1}^{n} \widehat{Q}_{kl} T_{lm} \widehat{Q}_{jm} \\
&= -\widehat{Q}_{ii}\left[\widehat{Q}_{jj} \sum_{m=j+1}^{n} T_{jm} \widehat{Q}_{im} + \widehat{Q}_{ii} \sum_{l=1}^{i-1} \widehat{Q}_{jl} T_{li} + \sum_{l=1,l\neq j}^{n}\sum_{m=l+1,m\neq i}^{n} \widehat{Q}_{jl} T_{lm} \widehat{Q}_{im} \right] \\
&\quad - \sum_{k=1,k\neq i}^{j-1} \widehat{Q}_{ik} \sum_{l=1}^{n}\sum_{m=l+1}^{n} \widehat{Q}_{jl} T_{lm} \widehat{Q}_{km} + \sum_{k=j+1}^{n} \widehat{Q}_{ik} \sum_{l=1}^{n}\sum_{m=l+1}^{n} \widehat{Q}_{kl} T_{lm} \widehat{Q}_{jm}.
\end{aligned} \tag{A.4}$$

Finally, consider the term $q_{ij}^F(t,\widehat{Q},\omega)$. Using (1.5), we have for $i<j$ (and similarly for $i>j$)

$$\begin{aligned}
q_{ij}^F(t,\widehat{Q},\omega) - q_{ij}^F(t,I,\omega) &= [\widehat{Q}S(\widehat{Q},F) - S(I,F)]_{ij} \\
&= -\sum_{k=1}^{j-1} \widehat{Q}_{ik} S(\widehat{Q},F)_{jk} + \sum_{k=j+1}^{n} \widehat{Q}_{ik} S(\widehat{Q},F)_{kj} + F_{ji} \\
&= F_{ji}\left(1 - \widehat{Q}_{jj} + \sum_{k\neq i,k=1}^{n} \widehat{Q}_{ik}^2 \widehat{Q}_{jj}\right) \\
&\quad - \widehat{Q}_{ii} \sum_{(l,m)\neq(j,i),l,m=1}^{n} \widehat{Q}_{lj} F_{lm} \widehat{Q}_{mi} + \sum_{k=1,k\neq i}^{n} \widehat{Q}_{ik} S(\widehat{Q},F)_{kj}.
\end{aligned} \tag{A.5}$$

REFERENCES

[1] L. Ya. Adrianova, *Introduction to Linear Systems of Differential Equations*, Transl. Math. Monogr. 146, AMS, Providence, RI, 1995.

[2] M. P. Calvo, A. Iserles, and A. Zanna, *Numerical solution of isospectral flows*, Math. Comp., 66 (1997), pp. 1461–1486.

[3] E. Celledoni and B. Owren, *A class of intrinsic schemes for orthogonal integration*, SIAM J. Numer. Anal., 40 (2002), pp. 2069–2084.

[4] X.-W. Chang and C. C. Paige, *Componentwise perturbation analyses for the QR factorization*, Numer. Math., 88 (2001), pp. 319–345.

[5] X.-W. Chang, C. C. Paige, and G. W. Stewart, *Perturbation analyses for the QR factorization*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 775–791.

[6] A. Davey, *An automatic orthonormalization method for solving stiff BVPs*, J. Comput. Phys., 51 (1983), pp. 343–356.
[7] L. Dieci and E. Van Vleck, *Computation of orthonormal factors for fundamental solution matrices*, Numer. Math., 83 (1999), pp. 599–620.
[8] L. Dieci and E. S. Van Vleck, *Lyapunov spectral intervals: Theory and computation*, SIAM J. Numer. Anal., 40 (2002), pp. 516–542.
[9] L. Dieci and E. S. Van Vleck, *Lyapunov and Sacker-Sell spectral intervals*, J. Dynam. Differential Equations, 19 (2007) pp. 265–293.
[10] L. Dieci and E. S. Van Vleck, *On the error in computing lyapunov exponents by QR methods*, Numer. Math., 101 (2005), pp. 619–642.
[11] L. Dieci and E. S. Van Vleck, *Perturbation theory for approximation of Lyapunov exponents by QR methods*, J. Dynam. Differential Equations, 18 (2006), pp. 815–840.
[12] L. Dieci, R. D. Russell, and E. S. Van Vleck, *Unitary integrators and applications to continuous orthonormalization techniques*, SIAM J. Numer. Anal., 31 (1994), pp. 261–281.
[13] L. Dieci, R. D. Russell, and E. S. Van Vleck, *On the computation of Lyapunov exponents for continuous dynamical systems*, SIAM J. Numer. Anal., 34 (1997), pp. 402–423.
[14] F. Diele, L. Lopez, and R. Peluso, *The Cayley transform in the numerical solution of unitary differential systems*, Adv. Comput. Math., 8 (1998), pp. 317–334.
[15] J. K. Hale, *Ordinary Differential Equations,* Robert E. Krieger, Huntington, NY, 1980.
[16] D. Higham, *Time-stepping and preserving orthonormality*, BIT, 37 (1997), pp. 24–36.
[17] W. Liu and E. S. Van Vleck, *Exponential Dichotomy for Asymptotically Hyperbolic Two-Dimensional Linear Systems*, 2007, submitted.
[18] G. H. Meyer, *Continuous orthonormalization for boundary value problems*, J. Comput. Phys., 62 (1986), pp. 248–262.
[19] K. J. Palmer, *The structurally stable systems on the half-line are those with exponential dichotomy*, J. Differential Equations, 33 (1979), pp. 16–25.
[20] K. J. Palmer, *Exponential dichotomy, integral separation and diagonalizability of linear systems of ordinary differential equations*, J. Differential Equations, 43 (1982), pp. 184–203.
[21] K. J. Palmer, *Exponential separation, exponential dichotomy and spectral theory for linear systems of ordinary differential equations*, J. Differential Equations, 46 (1982), pp. 324–345.
[22] G. W. Stewart, *Perturbation bounds for the QR factorization of a matrix*, SIAM J. Numer. Anal., 14 (1977), pp. 509–518.

SIAM J. NUMER. ANAL.
Vol. 46, No. 3, pp. 1190–1211
© 2008 Society for Industrial and Applied Mathematics

A LOCALLY DIVERGENCE-FREE INTERIOR PENALTY METHOD FOR TWO-DIMENSIONAL CURL-CURL PROBLEMS*

SUSANNE C. BRENNER†, FENGYAN LI‡, AND LI-YENG SUNG§

Abstract. An interior penalty method for certain two-dimensional curl-curl problems is investigated in this paper. This method computes the divergence-free part of the solution using locally divergence-free discontinuous P_1 vector fields on graded meshes. It has optimal order convergence (up to an arbitrarily small ϵ) for the source problem and the eigenproblem. Results of numerical experiments that corroborate the theoretical results are also presented.

Key words. curl-curl problem, Maxwell eigenproblem, locally divergence-free, interior penalty methods, graded meshes

AMS subject classifications. 65N30, 65N15, 35Q60

DOI. 10.1137/060671760

1. Introduction.

Let $\Omega \subset \mathbb{R}^2$ be a bounded polygonal domain. Consider the following weak curl-curl problem: Find $\boldsymbol{u} \in H_0(\mathrm{curl};\Omega)$ such that

$$(\nabla \times \boldsymbol{u}, \nabla \times \boldsymbol{v}) + \alpha(\boldsymbol{u}, \boldsymbol{v}) = (\boldsymbol{f}, \boldsymbol{v}) \qquad \forall\, \boldsymbol{v} \in H_0(\mathrm{curl};\Omega), \tag{1.1}$$

where α is a constant, $(\cdot,\cdot)$ denotes the inner product of $L_2(\Omega)$ (or $[L_2(\Omega)]^2$),

$$H(\mathrm{curl};\Omega) = \left\{ \boldsymbol{v} = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \in [L_2(\Omega)]^2 : \nabla \times \boldsymbol{v} = \frac{\partial v_2}{\partial x_1} - \frac{\partial v_1}{\partial x_2} \in L_2(\Omega) \right\},$$

and

$$H_0(\mathrm{curl};\Omega) = \left\{ \boldsymbol{v} = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \in H(\mathrm{curl};\Omega) : \boldsymbol{n} \times \boldsymbol{v} = n_1 v_2 - n_2 v_1 = 0 \text{ on } \partial\Omega \right\},$$

with $\boldsymbol{n} = \left[\begin{smallmatrix} n_1 \\ n_2 \end{smallmatrix}\right]$ the unit outer normal on $\partial\Omega$.

The curl-curl problem (1.1) is related to electromagnetic problems. For $\alpha \leq 0$, it is the weak form of the time-harmonic (frequency-domain) Maxwell equations. For $\alpha > 0$, it is related to the spatial problems appearing in implicit semidiscretizations of the time-dependent (time-domain) Maxwell equations.

It is well known [24, 25, 5] that H^1 conforming nodal finite element methods can lead to a wrong solution of (1.1) if Ω is not convex. Many alternative approaches have been developed which include $H(\mathrm{curl};\Omega)$ conforming edge element methods [41, 42, 13,

*Received by the editors October 8, 2006; accepted for publication (in revised form) June 4, 2007; published electronically March 21, 2008.
http://www.siam.org/journals/sinum/46-3/67176.html

†Department of Mathematics and Center for Computation and Technology, Louisiana State University, Baton Rouge, LA 70803 (brenner@math.lsu.edu). The research of this author was supported by the Alexander von Humboldt Foundation through a Humboldt Research Award and in part by National Science Foundation grant DMS-03-11790.

‡Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, NY 12180 (lif@rpi.edu). The work of this author was supported in part by National Science Foundation grant DMS-06-52481.

§Department of Mathematics, Louisiana State University, Baton Rouge, LA 70803 (sung@math.lsu.edu).

33, 38, 40], H^1 conforming nodal finite element methods with weighted regularization [27, 29], the singular complement/field method [32, 5], and interior penalty methods [43, 36, 37, 35, 34].

In this paper we take a different approach. By the Helmholtz decomposition, the solution $\boldsymbol{u} \in H_0(\mathrm{curl};\Omega)$ of (1.1) can be written uniquely as

$$\boldsymbol{u} = \mathring{\boldsymbol{u}} + \nabla\phi, \tag{1.2}$$

where $\mathring{\boldsymbol{u}} \in H_0(\mathrm{curl};\Omega) \cap H(\mathrm{div}^0;\Omega)$, $\phi \in H_0^1(\Omega)$,

$$H(\mathrm{div}^0;\Omega) = \left\{ \boldsymbol{v} = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \in [L_2(\Omega)]^2 : \nabla\cdot\boldsymbol{v} = \frac{\partial v_1}{\partial x_1} + \frac{\partial v_2}{\partial x_2} = 0 \right\}.$$

It is easy to see that ϕ satisfies

$$\alpha(\nabla\phi, \nabla\psi) = (\boldsymbol{f}, \nabla\psi) \qquad \forall\, \psi \in H_0^1(\Omega). \tag{1.3}$$

Since the Poisson problem (1.3) (when $\alpha \neq 0$) can be solved by many standard methods under the assumption that $\boldsymbol{f} \in H(\mathrm{div};\Omega)$, we will focus on the divergence-free part $\mathring{\boldsymbol{u}}$, which satisfies

$$(\nabla\times\mathring{\boldsymbol{u}}, \nabla\times\boldsymbol{v}) + \alpha(\mathring{\boldsymbol{u}}, \boldsymbol{v}) = (\boldsymbol{f}, \boldsymbol{v}) \qquad \forall\, \boldsymbol{v} \in H_0(\mathrm{curl};\Omega) \cap H(\mathrm{div}^0;\Omega). \tag{1.4}$$

We shall refer to (1.4) as the weak form of the reduced curl-curl problem and assume in the case where $\alpha \leq 0$ that (1.4) has a unique solution (i.e., $-\alpha$ is not a Maxwell eigenvalue). Note that the strong form of the reduced curl-curl problem is given by

$$\nabla\times(\nabla\times\mathring{\boldsymbol{u}}) + \alpha\mathring{\boldsymbol{u}} = Q\boldsymbol{f}, \tag{1.5}$$

where $Q : [L_2(\Omega)]^2 \longrightarrow H(\mathrm{div}^0;\Omega)$ is the orthogonal projection.

Remark 1.1. The strong form (1.5) and the well-posedness of (1.4) imply immediately

$$\|\nabla\times\mathring{\boldsymbol{u}}\|_{H^1(\Omega)} \leq C_{\Omega,\alpha}\|Q\boldsymbol{f}\|_{L_2(\Omega)} \leq C_{\Omega,\alpha}\|\boldsymbol{f}\|_{L_2(\Omega)}. \tag{1.6}$$

Here and below we use C (with or without subscripts) to denote a generic positive constant that can take different values at different occurrences.

The advantage of working with (1.4) is that it behaves like an elliptic problem, unlike the full curl-curl problem (1.1). In particular, the solution $\mathring{\boldsymbol{u}}$ of (1.4) enjoys elliptic regularity under the assumption that $\boldsymbol{f} \in L_2(\Omega)$, which greatly simplifies the duality argument needed for the analysis.

In [15] we developed a numerical scheme (in the case where $\alpha \leq 0$) for (1.4) using locally divergence-free Crouzeix–Raviart nonconforming P_1 vector fields [30]. In this paper we investigate an interior penalty method that avoids the difficulties involved in the construction of local bases for locally divergence-free Crouzeix–Raviart vector fields at the cost of a larger number of unknowns. These two methods are very closely related, which simplifies the analysis of the new scheme.

Since the solution operator of the interior penalty method converges uniformly to the solution operator of (1.4) with respect to the norm $\|\cdot\|_{L_2(\Omega)}$, the classical spectral approximation theory [7] implies that we can also apply the interior penalty method

to the following Maxwell eigenproblem: Find $(\lambda, \mathring{\boldsymbol{u}}) \in \mathbb{R} \times H_0(\mathrm{curl};\Omega) \cap H(\mathrm{div}^0;\Omega)$ such that

$$(\nabla \times \mathring{\boldsymbol{u}}, \nabla \times \boldsymbol{v}) = \lambda(\mathring{\boldsymbol{u}}, \boldsymbol{v}) \qquad \forall\, \boldsymbol{v} \in H_0(\mathrm{curl};\Omega) \cap H(\mathrm{div}^0;\Omega). \tag{1.7}$$

In particular, the spectral approximation based on our interior penalty method is free of spurious eigenvalues.

The rest of the paper is organized as follows. We introduce the finite element space and graded meshes in section 2. The interior penalty method and some preliminary estimates are presented in section 3. The convergence analysis for the source problem (1.1) is carried out in section 4. Application to the eigenproblem (1.7) is discussed in section 5, followed by numerical results in section 6, and we end with some concluding remarks in section 7.

2. Finite element spaces and graded meshes. In this section we will introduce discontinuous finite element spaces defined on meshes graded around the corners $c_1, \ldots, c_L$ of Ω.

We assume that the triangulation $\mathcal{T}_h$ of Ω satisfies the following condition:

$$h_T \approx h\Phi_\mu(T) \qquad \forall\, T \in \mathcal{T}_h, \tag{2.1}$$

where $h_T = \mathrm{diam}\, T$, h is a mesh parameter proportional to $\max_{T\in\mathcal{T}_h} h_T$, and $\mu = (\mu_1, \ldots, \mu_L)$ is the vector containing the grading parameters. The constants in the equivalence (2.1) are independent of h and depend in general on the minimum angle of the mesh. The weight $\Phi_\mu(T)$ is defined by

$$\Phi_\mu(T) = \Pi_{\ell=1}^{L} |c_\ell - c_{_T}|^{1-\mu_\ell}, \tag{2.2}$$

where $c_{_T}$ is the center of T. Note that

$$h_T \le Ch. \tag{2.3}$$

The choices of the grading parameters are dictated by the singularities [6, 28, 26] of the solution $\mathring{\boldsymbol{u}}$ of (1.4). In order to recover optimal convergence rates in both the energy norm and the L_2 norm (cf. [15]), we take

$$\begin{aligned} &\mu_\ell = 1 && \text{if} \quad \omega_\ell \le \frac{\pi}{2}, \\ &\mu_\ell < \frac{\pi}{2\omega_\ell} && \text{if} \quad \omega_\ell > \frac{\pi}{2}, \end{aligned} \tag{2.4}$$

where ω_ℓ is the interior angle at the corner c_ℓ. The construction of $\mathcal{T}_h$ satisfying (2.1) can be found, for example, in [31, 4, 1, 14]. Note that $\mathcal{T}_h$ can be constructed so that it satisfies a minimum angle condition for any fixed choice of the grading parameters.

Remark 2.1. The choice of the grading parameters in (2.4) indicates that grading is needed at any corner where the angle is larger than $\pi/2$. This is different from problems involving the Laplace operator where grading is only needed at reentrant corners and is due to the fact that the singularities of the curl-curl operator are one order worse than the singularities of the Laplace operator.

The finite element space V_h is defined by

$$V_h = \{\boldsymbol{v} \in [L_2(\Omega)]^2 : \boldsymbol{v}_{_T} = \boldsymbol{v}\big|_{_T} \in [P_1(T)]^2 \text{ and } \nabla \cdot \boldsymbol{v}_{_T} = 0\ \forall\, T \in \mathcal{T}_h\}.$$

For any $s > \frac{1}{2}$ there is a natural weak interpolation operator $\Pi_T : [H^s(T)]^2 \longrightarrow [P_1(T)]^2$ defined by

$$(\Pi_T \boldsymbol{\zeta})(m_{e_i}) = \frac{1}{|e_i|}\int_{e_i} \boldsymbol{\zeta}\, ds \quad \text{for} \quad i = 1, 2, 3, \tag{2.5}$$

where m_{e_i} is the midpoint of the edge e_i of T and $|e|$ denotes the length of an edge e.

It follows immediately from (2.5), the midpoint rule, and Green's theorem that

$$\int_T \nabla \times (\Pi_T \boldsymbol{\zeta}) dx = \int_T \nabla \times \boldsymbol{\zeta}\, dx \qquad \forall\, \boldsymbol{\zeta} \in [H^s(T)]^2, \tag{2.6}$$

$$\int_T \nabla \cdot (\Pi_T \boldsymbol{\zeta}) dx = \int_T \nabla \cdot \boldsymbol{\zeta}\, dx \qquad \forall\, \boldsymbol{\zeta} \in [H^s(T)]^2. \tag{2.7}$$

Furthermore, given $s \in (1/2, 2]$, we have the following interpolation error estimates [30]:

$$\|\boldsymbol{\zeta} - \Pi_T \boldsymbol{\zeta}\|_{L_2(T)} + h_T^{\min(s,1)} |\boldsymbol{\zeta} - \Pi_T \boldsymbol{\zeta}|_{H^{\min(s,1)}(T)} \le C_T h_T^s |\boldsymbol{\zeta}|_{H^s(T)} \tag{2.8}$$

for all $\boldsymbol{\zeta} \in [H^s(T)]^2$, where the positive constant C_T depends on the minimum angle of T (and also on s when s is close to $1/2$).

Since the solution $\mathring{\boldsymbol{u}}$ of the reduced curl-curl problem belongs to $[H^s(\Omega)]^2$ for some $s > \frac{1}{2}$ (cf. [40]), we can define a global interpolant of $\mathring{\boldsymbol{u}}$ by

$$(\Pi_h \mathring{\boldsymbol{u}})_T = \Pi_T \mathring{\boldsymbol{u}}_T \qquad \forall\, T \in \mathcal{T}_h,$$

where $\mathring{\boldsymbol{u}}_T = \mathring{\boldsymbol{u}}\big|_T$. It follows from (2.7) and $\nabla \cdot \mathring{\boldsymbol{u}} = 0$ that $\Pi_h \mathring{\boldsymbol{u}} \in V_h$.

Remark 2.2. Note that $\Pi_h \mathring{\boldsymbol{u}}$ is the same interpolation operator used in [15] and $\Pi_h \mathring{\boldsymbol{u}}$ belongs to the space of locally divergence-free Crouzeix–Raviart nonconforming P_1 vector fields [30].

Since the vector fields in V_h are discontinuous, their jumps across the edges of $\mathcal{T}_h$ play an important role in the development of interior penalty methods. We will denote by $\mathcal{E}_h$ (resp., $\mathcal{E}_h^i$) the set of the edges (resp., interior edges) of $\mathcal{T}_h$. Let $e \in \mathcal{E}_h^i$ be shared by the two triangles $T_1, T_2 \in \mathcal{T}_h$ (cf. Figure 2.1) and let $\boldsymbol{n}_1$ (resp., $\boldsymbol{n}_2$) be the unit normal of e pointing towards the outside of T_1 (resp., T_2). We define, on e,

$$[\![\boldsymbol{n} \times \boldsymbol{v}]\!] = \boldsymbol{n}_1 \times \boldsymbol{v}_{T_1}\big|_e + \boldsymbol{n}_2 \times \boldsymbol{v}_{T_2}\big|_e, \tag{2.9a}$$

$$[\![\boldsymbol{n} \cdot \boldsymbol{v}]\!] = \boldsymbol{n}_1 \cdot \boldsymbol{v}_{T_1}\big|_e + \boldsymbol{n}_2 \cdot \boldsymbol{v}_{T_2}\big|_e. \tag{2.9b}$$

For an edge e along $\partial\Omega$, we take $\boldsymbol{n}_e$ to be the unit normal of e pointing towards the outside of Ω and define

$$[\![\boldsymbol{n} \times \boldsymbol{v}]\!] = \boldsymbol{n}_e \times \boldsymbol{v}\big|_e. \tag{2.10}$$

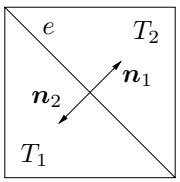


FIG. 2.1. *Triangles and normals in the definitions of* $[\![\boldsymbol{n} \times \boldsymbol{v}]\!]$ *and* $[\![\boldsymbol{n} \cdot \boldsymbol{v}]\!]$.

We will also denote the piecewise defined curl and div operators by $\nabla_h\times$ and $\nabla_h\cdot$, i.e.,

$$(\nabla_h \times \boldsymbol{v})\big|_T = \nabla \times \boldsymbol{v}_T \quad \text{and} \quad (\nabla_h \cdot \boldsymbol{v})\big|_T = \nabla \cdot \boldsymbol{v}_T.$$

3. Discretization and preliminary error estimates. In this section we present the interior penalty method for the reduced curl-curl problem and some preliminary estimates.

The discrete problem is to find $\mathring{\boldsymbol{u}}_h \in V_h$ such that

$$a_h(\mathring{\boldsymbol{u}}_h, \boldsymbol{v}) = (\boldsymbol{f}, \boldsymbol{v}) \qquad \forall\, \boldsymbol{v} \in\ V_h, \tag{3.1}$$

where

$$\begin{aligned} a_h(\boldsymbol{w}, \boldsymbol{v}) = {} & (\nabla_h \times \boldsymbol{w}, \nabla_h \times \boldsymbol{v}) + \alpha(\boldsymbol{w}, \boldsymbol{v}) \\ & + \sum_{e\in\mathcal{E}_h} \frac{[\Phi_\mu(e)]^2}{|e|} \int_e [\![\boldsymbol{n} \times \boldsymbol{w}]\!][\![\boldsymbol{n} \times \boldsymbol{v}]\!]\, ds \\ & + \sum_{e\in\mathcal{E}_h^i} \frac{[\Phi_\mu(e)]^2}{|e|} \int_e [\![\boldsymbol{n} \cdot \boldsymbol{w}]\!][\![\boldsymbol{n} \cdot \boldsymbol{v}]\!]\, ds \\ & + h^{-2} \sum_{e\in\mathcal{E}_h} \frac{1}{|e|} \int_e \big(\Pi_e^0[\![\boldsymbol{n} \times \boldsymbol{w}]\!]\big)\big(\Pi_e^0[\![\boldsymbol{n} \times \boldsymbol{v}]\!]\big)\, ds \\ & + h^{-2} \sum_{e\in\mathcal{E}_h^i} \frac{1}{|e|} \int_e \big(\Pi_e^0[\![\boldsymbol{n} \cdot \boldsymbol{w}]\!]\big)\big(\Pi_e^0[\![\boldsymbol{n} \cdot \boldsymbol{v}]\!]\big)\, ds, \end{aligned} \tag{3.2}$$

where Π_e^0 is the orthogonal projection from $L_2(e)$ onto $P_0(e)$ (the space of constant functions on e). The weight $\Phi_\mu(e)$ in (3.2) is defined by

$$\Phi_\mu(e) = \Pi_{\ell=1}^L |c_\ell - m_e|^{1-\mu_\ell},$$

where m_e is the midpoint of e.

Remark 3.1. The weight $\Phi_\mu(e)$ is closely related to the weight $\Phi_\mu(T)$ in (2.2). In fact, we have

$$\Phi_\mu(e) \approx \Phi_\mu(T) \qquad \text{if} \quad e \subset \partial T, \tag{3.3}$$

where the constants in the equivalence are independent of h and depend in general on the minimum angle of the mesh. The inclusion of $\Phi_\mu(e)$ in (3.2) is crucial for the derivation of optimal convergence rates on graded meshes.

Remark 3.2. The weak problem (1.4) for the reduced curl-curl problem can be written as

$$a_h(\mathring{\boldsymbol{u}}, \boldsymbol{v}) = (\boldsymbol{f}, \boldsymbol{v}) \qquad \forall\, \boldsymbol{v} \in H_0(\mathrm{curl};\Omega) \cap H(\mathrm{div}^0;\Omega).$$

Remark 3.3. For nonconforming Crouzeix–Raviart P_1 vector fields $\boldsymbol{w}$ and $\boldsymbol{v}$, we have

$$\begin{aligned} a_h(\boldsymbol{w},\boldsymbol{v}) &= (\nabla_h\times\boldsymbol{w},\nabla_h\times\boldsymbol{v})+\alpha(\boldsymbol{w},\boldsymbol{v}) \\ &\quad+\sum_{e\in\mathcal{E}_h}\frac{[\Phi_\mu(e)]^2}{|e|}\int_e[\![\boldsymbol{n}\times\boldsymbol{w}]\!][\![\boldsymbol{n}\times\boldsymbol{v}]\!]\,ds \\ &\quad+\sum_{e\in\mathcal{E}_h^i}\frac{[\Phi_\mu(e)]^2}{|e|}\int_e[\![\boldsymbol{n}\cdot\boldsymbol{w}]\!][\![\boldsymbol{n}\cdot\boldsymbol{v}]\!]\,ds, \end{aligned}$$

which is the variational form used in [15]. The two additional sums in (3.2) compensate for the lack of weak continuity for the vector fields in V_h.

Remark 3.4. The weak overpenalization causes the discrete system to become more ill-conditioned. For example, the condition number grows at the rate of $O(h^{-4})$ on quasi-uniform meshes. However, there exists a simple block-diagonal preconditioner that can reduce the growth of the condition number to $O(h^{-2})$ on quasi-uniform meshes and $O(h^{-2}(1+|\ln h|)$ on graded meshes [16, 17].

Let the mesh-dependent energy norm $\|\cdot\|_h$ be defined by

$$\begin{aligned} \|\boldsymbol{v}\|_h^2 &= \|\nabla_h\times\boldsymbol{v}\|_{L_2(\Omega)}^2+\|\boldsymbol{v}\|_{L_2(\Omega)}^2 \\ &\quad+\sum_{e\in\mathcal{E}_h}\frac{[\Phi_\mu(e)]^2}{|e|}\|[\![\boldsymbol{n}\times\boldsymbol{v}]\!]\|_{L_2(e)}^2+\sum_{e\in\mathcal{E}_h^i}\frac{[\Phi_\mu(e)]^2}{|e|}\|[\![\boldsymbol{n}\cdot\boldsymbol{v}]\!]\|_{L_2(e)}^2 \\ &\quad+h^{-2}\left(\sum_{e\in\mathcal{E}_h}\frac{1}{|e|}\|\Pi_e^0[\![\boldsymbol{n}\times\boldsymbol{v}]\!]\|_{L_2(e)}^2+\sum_{e\in\mathcal{E}_h^i}\frac{1}{|e|}\|\Pi_e^0[\![\boldsymbol{n}\cdot\boldsymbol{v}]\!]\|_{L_2(e)}^2\right). \end{aligned} \tag{3.4}$$

Note that

$$\|\boldsymbol{v}\|_{L_2(\Omega)}\le\|\boldsymbol{v}\|_h\qquad\forall\,\boldsymbol{v}\in H_0(\mathrm{curl};\Omega)\cap H(\mathrm{div}^0;\Omega)+V_h, \tag{3.5}$$

and from (3.2) and (3.4),

$$|a_h(\boldsymbol{w},\boldsymbol{v})|\le(|\alpha|+1)\|\boldsymbol{w}\|_h\|\boldsymbol{v}\|_h\qquad\forall\,\boldsymbol{w},\boldsymbol{v}\in[H_0(\mathrm{curl};\Omega)\cap H(\mathrm{div}^0;\Omega)]+V_h. \tag{3.6}$$

For $\alpha>0$, $a_h(\cdot,\cdot)$ is coercive with respect to $\|\cdot\|_h$, i.e.,

$$a_h(\boldsymbol{v},\boldsymbol{v})\ge\min(1,\alpha)\|\boldsymbol{v}\|_h^2\qquad\forall\,\boldsymbol{v}\in H_0(\mathrm{curl};\Omega)\cap H(\mathrm{div}^0;\Omega)+V_h. \tag{3.7}$$

In this case the discrete problem is well-posed and we have the following abstract error estimate.

LEMMA 3.5. *Let α be positive, $\beta=\min(1,\alpha)$, $\mathring{\boldsymbol{u}}$ be the solution of* (1.4)*, and $\mathring{\boldsymbol{u}}_h$ satisfy* (3.1)*. It holds that*

$$\|\mathring{\boldsymbol{u}}-\mathring{\boldsymbol{u}}_h\|_h\le\Big(\frac{1+\alpha+\beta}{\beta}\Big)\inf_{\boldsymbol{v}\in V_h}\|\mathring{\boldsymbol{u}}-\boldsymbol{v}\|_h+\frac{1}{\beta}\sup_{\boldsymbol{w}\in V_h\setminus\{\boldsymbol{0}\}}\frac{a_h(\mathring{\boldsymbol{u}}-\mathring{\boldsymbol{u}}_h,\boldsymbol{w})}{\|\boldsymbol{w}\|_h}. \tag{3.8}$$

Proof. Let $\boldsymbol{v} \in V_h$ be arbitrary. It follows from (3.6), (3.7), and the triangle inequality that

$$\begin{aligned}\|\mathring{\boldsymbol{u}} - \mathring{\boldsymbol{u}}_h\|_h &\leq \|\mathring{\boldsymbol{u}} - \boldsymbol{v}\|_h + \|\boldsymbol{v} - \mathring{\boldsymbol{u}}_h\|_h \\ &\leq \|\mathring{\boldsymbol{u}} - \boldsymbol{v}\|_h + \frac{1}{\beta} \sup_{\boldsymbol{w}\in V_h\setminus\{\boldsymbol{0}\}} \frac{a_h(\boldsymbol{v} - \mathring{\boldsymbol{u}}_h, \boldsymbol{w})}{\|\boldsymbol{w}\|_h} \\ &\leq \Big(\frac{1+\alpha+\beta}{\beta}\Big)\|\mathring{\boldsymbol{u}} - \boldsymbol{v}\|_h + \frac{1}{\beta} \sup_{\boldsymbol{w}\in V_h\setminus\{\boldsymbol{0}\}} \frac{a_h(\mathring{\boldsymbol{u}} - \mathring{\boldsymbol{u}}_h, \boldsymbol{w})}{\|\boldsymbol{w}\|_h},\end{aligned}$$

which implies (3.8). ∎

For $\alpha \leq 0$, the following Gårding (in)equality holds:

$$a_h(\boldsymbol{v},\boldsymbol{v}) + (|\alpha|+1)(\boldsymbol{v},\boldsymbol{v}) = \|\boldsymbol{v}\|_h^2 \qquad \forall\, \boldsymbol{v} \in [H_0(\mathrm{curl};\Omega)\cap H(\mathrm{div}^0;\Omega)] + V_h. \tag{3.9}$$

In this case the discrete problem is indefinite and the following lemma provides an abstract error estimate for the scheme (3.1) under the assumption that it has a solution.

LEMMA 3.6. *Let* $\mathring{\boldsymbol{u}} \in H_0(\mathrm{curl};\Omega)\cap H(\mathrm{div}^0;\Omega)$ *satisfy* (1.4) *and let* $\mathring{\boldsymbol{u}}_h$ *be a solution of* (3.1). *It holds that*

$$\begin{aligned}\|\mathring{\boldsymbol{u}} - \mathring{\boldsymbol{u}}_h\|_h \leq (2|\alpha|+3) \inf_{\boldsymbol{v}\in V_h} \|\mathring{\boldsymbol{u}} - \boldsymbol{v}\|_h + \sup_{\boldsymbol{w}\in V_h\setminus\{\boldsymbol{0}\}} \frac{a_h(\mathring{\boldsymbol{u}} - \mathring{\boldsymbol{u}}_h, \boldsymbol{w})}{\|\boldsymbol{w}\|_h} \\ + (|\alpha|+1)\|\mathring{\boldsymbol{u}} - \mathring{\boldsymbol{u}}_h\|_{L_2(\Omega)}.\end{aligned} \tag{3.10}$$

Proof. It follows from (3.5) and (3.9) that, for $\boldsymbol{v} \in V_h \setminus \{\boldsymbol{0}\}$,

$$\begin{aligned}\|\boldsymbol{v}\|_h &\leq \frac{a_h(\boldsymbol{v},\boldsymbol{v})}{\|\boldsymbol{v}\|_h} + (|\alpha|+1)\frac{(\boldsymbol{v},\boldsymbol{v})}{\|\boldsymbol{v}\|_h} \\ &\leq \sup_{\boldsymbol{w}\in V_h\setminus\{\boldsymbol{0}\}} \frac{a_h(\boldsymbol{v},\boldsymbol{w})}{\|\boldsymbol{w}\|_h} + (|\alpha|+1)\|\boldsymbol{v}\|_{L_2(\Omega)}.\end{aligned} \tag{3.11}$$

Let $\boldsymbol{v} \in V_h$ be arbitrary. We find, using (3.5), (3.6), (3.11), and the triangle inequality,

$$\begin{aligned}\|\mathring{\boldsymbol{u}} - \mathring{\boldsymbol{u}}_h\|_h &\leq \|\mathring{\boldsymbol{u}} - \boldsymbol{v}\|_h + \|\boldsymbol{v} - \mathring{\boldsymbol{u}}_h\|_h \\ &\leq \|\mathring{\boldsymbol{u}} - \boldsymbol{v}\|_h + \sup_{\boldsymbol{w}\in V_h\setminus\{\boldsymbol{0}\}} \frac{a_h(\boldsymbol{v} - \mathring{\boldsymbol{u}}_h, \boldsymbol{w})}{\|\boldsymbol{w}\|_h} + (|\alpha|+1)\|\boldsymbol{v} - \mathring{\boldsymbol{u}}_h\|_{L_2(\Omega)} \\ &\leq (2|\alpha|+3)\|\mathring{\boldsymbol{u}} - \boldsymbol{v}\|_h + \sup_{\boldsymbol{w}\in V_h\setminus\{\boldsymbol{0}\}} \frac{a_h(\mathring{\boldsymbol{u}} - \mathring{\boldsymbol{u}}_h, \boldsymbol{w})}{\|\boldsymbol{w}\|_h} + (|\alpha|+1)\|\mathring{\boldsymbol{u}} - \mathring{\boldsymbol{u}}_h\|_{L_2(\Omega)},\end{aligned}$$

which implies (3.10). ∎

From here on we consider α to be fixed and drop the dependence on α in our estimates.

Remark 3.7. The first term on the right-hand side of (3.8) and (3.10) measures the approximation property of V_h with respect to the norm $\|\cdot\|_h$. The second term

measures the consistency error. The third term on the right-hand side of (3.10) addresses the indefiniteness of the reduced curl-curl problem when $\alpha \leq 0$.

As mentioned in Remark 2.2, the interpolation operator Π_h is also the one used in [15]. Therefore in our analysis we can use the following two results from that paper (cf. Lemmas 5.1 and 5.2 of [15]), which were obtained using (2.4), (2.8), and a representation of $\mathring{\boldsymbol{u}}$ as the sum of a regular part and a singular part.

LEMMA 3.8. *Let $\mathring{\boldsymbol{u}} \in H_0(\mathrm{curl};\Omega) \cap H(\mathrm{div}^0;\Omega)$ be the solution of (1.4). We have the following interpolation error estimate:*

$$\|\mathring{\boldsymbol{u}} - \Pi_h \mathring{\boldsymbol{u}}\|_{L_2(\Omega)} \leq C_\epsilon h^{2-\epsilon} \|\boldsymbol{f}\|_{L_2(\Omega)} \qquad \text{for any} \quad \epsilon > 0. \tag{3.12}$$

LEMMA 3.9. *Let $\mathring{\boldsymbol{u}} \in H_0(\mathrm{curl};\Omega) \cap H(\mathrm{div}^0;\Omega)$ be the solution of (1.4). We have the following interpolation error estimate:*

$$\sum_{e\in\mathcal{E}_h} \frac{[\Phi_\mu(e)]^2}{|e|} \|[\![\mathring{\boldsymbol{u}} - \Pi_h \mathring{\boldsymbol{u}}]\!]\|^2_{L_2(e)} \leq C_\epsilon h^{2-\epsilon} \|\boldsymbol{f}\|^2_{L_2(\Omega)} \tag{3.13}$$

for any $\epsilon > 0$, where $[\![\mathring{\boldsymbol{u}} - \Pi_h \mathring{\boldsymbol{u}}]\!]$ is the jump of $\mathring{\boldsymbol{u}} - \Pi_h \boldsymbol{u}$ across the interior edges of $\mathcal{T}_h$ and $[\![\mathring{\boldsymbol{u}} - \Pi_h \mathring{\boldsymbol{u}}]\!] = \mathring{\boldsymbol{u}} - \Pi_h \mathring{\boldsymbol{u}}$ on the boundary edges of $\mathcal{T}_h$.

LEMMA 3.10. *It holds that*

$$\sum_{e\in\mathcal{E}_h} |e|\,[\Phi_\mu(e)]^{-2} \|\eta - \bar{\eta}_{T_e}\|^2_{L_2(e)} \leq Ch^2 |\eta|^2_{H^1(\Omega)} \qquad \forall\, \eta \in H^1(\Omega),$$

where

$$\bar{\eta}_{T_e} = \frac{1}{|T_e|} \int_{T_e} \eta \, dx \tag{3.14}$$

is the mean of η over T_e, one of the triangles in $\mathcal{T}_h$ that has e as an edge.

Proof. This is the consequence of (2.1), (3.3), the trace theorem (with scaling), and a standard interpolation error estimate [22, 18]:

$$\begin{aligned}
&\sum_{e\in\mathcal{E}_h} |e|\,[\Phi_\mu(e)]^{-2} \|\eta - \bar{\eta}_{T_e}\|^2_{L_2(e)} \\
&\qquad \leq C \sum_{e\in\mathcal{E}_h} [\Phi_\mu(T)]^{-2} \big(\|\eta - \bar{\eta}_{T_e}\|^2_{L_2(T_e)} + h_T^2 |\eta - \bar{\eta}_{T_e}|^2_{H^1(T_e)} \big) \\
&\qquad \leq C \sum_{e\in\mathcal{E}_h} [\Phi_\mu(T)]^{-2} h_T^2 |\eta|^2_{H^1(T_e)} \leq Ch^2 |\eta|^2_{H^1(\Omega)}. \qquad \square
\end{aligned}$$

Recall that Q is the L_2 orthogonal projection operator onto $H(\mathrm{div}^0;\Omega)$. The following result will be useful in addressing the consistency error caused by the appearance of Q in (1.5).

LEMMA 3.11. *The following estimate holds:*

$$\|\boldsymbol{v} - Q\boldsymbol{v}\|_{L_2(\Omega)} \leq Ch \|\boldsymbol{v}\|_h \quad \forall\, \boldsymbol{v} \in [H_0(\mathrm{curl};\Omega) \cap H(\mathrm{div}^0;\Omega)] + V_h. \tag{3.15}$$

Proof. Let $\boldsymbol{v} \in [H_0(\mathrm{curl};\Omega) \cap H(\mathrm{div}^0;\Omega)] + V_h$ be arbitrary. Since $\boldsymbol{v} - Q\boldsymbol{v}$ belongs to $\nabla H_0^1(\Omega)$, the orthogonal complement of $H(\mathrm{div}^0;\Omega)$ in $[L_2(\Omega)]^2$, we have, by duality,

$$\|\boldsymbol{v} - Q\boldsymbol{v}\|_{L_2(\Omega)} = \sup_{\eta\in H_0^1(\Omega)\setminus\{0\}} \frac{(\boldsymbol{v} - Q\boldsymbol{v}, \nabla\eta)}{\|\nabla\eta\|_{L_2(\Omega)}} = \sup_{\eta\in H_0^1(\Omega)\setminus\{0\}} \frac{(\boldsymbol{v}, \nabla\eta)}{\|\nabla\eta\|_{L_2(\Omega)}}. \tag{3.16}$$

Let $\eta \in H_0^1(\Omega)$ be arbitrary. Since $\nabla \cdot \boldsymbol{v} = 0$ on each triangle $T \in \mathcal{T}_h$, we find, using integration by parts and the fact that $\eta\big|_{\partial\Omega} = 0$,

$$(\boldsymbol{v}, \nabla\eta) = \sum_{e\in\mathcal{E}_h^i} \int_e (\eta - \bar{\eta}_{T_e}) [\![\boldsymbol{n}\cdot\boldsymbol{v}]\!]\, ds + \sum_{e\in\mathcal{E}_h^i} \int_e \bar{\eta}_{T_e} \big(\Pi_e^0 [\![\boldsymbol{n}\cdot\boldsymbol{v}]\!]\big)\, ds = S_1 + S_2, \tag{3.17}$$

where $\bar{\eta}_{T_e}$ is defined in (3.14).

By the Cauchy–Schwarz inequality, (3.4), and Lemma 3.10, we have

$$\begin{aligned} S_1 &\le \left[\sum_{e\in\mathcal{E}_h^i} |e|[\Phi_\mu(e)]^{-2} \|\eta - \bar{\eta}_{T_e}\|_{L_2(e)}^2\right]^{1/2} \left[\sum_{e\in\mathcal{E}_h^i} \frac{[\Phi_\mu(e)]^2}{|e|} \|[\![\boldsymbol{n}\cdot\boldsymbol{v}]\!]\|_{L_2(e)}^2\right]^{1/2} \\ &\le Ch\|\nabla\eta\|_{L_2(\Omega)} \|\boldsymbol{v}\|_h. \end{aligned} \tag{3.18}$$

On the other hand, from (3.4), the Cauchy–Schwarz inequality, and a Poincaré–Friedrichs inequality, we find

$$\begin{aligned} S_2 &\le \sum_{e\in\mathcal{E}_h^i} \big(|e|^{1/2} \|\bar{\eta}_{T_e}\|_{L_2(e)}\big) \big(|e|^{-1/2} \|\Pi_e^0 [\![\boldsymbol{n}\cdot\boldsymbol{v}]\!]\|_{L_2(e)}\big) \\ &\le Ch \left[\sum_{e\in\mathcal{E}_h^i} \|\bar{\eta}_{T_e}\|_{L_2(T_e)}^2\right]^{1/2} \left[h^{-2} \sum_{e\in\mathcal{E}_h^i} \frac{1}{|e|} \|\Pi_e^0 [\![\boldsymbol{n}\cdot\boldsymbol{v}]\!]\|_{L_2(e)}^2\right]^{1/2} \qquad (3.19) \\ &\le Ch\|\eta\|_{L_2(\Omega)} \|\boldsymbol{v}\|_h \le Ch\|\nabla\eta\|_{L_2(\Omega)} \|\boldsymbol{v}\|_h. \end{aligned}$$

Here we have also used the simple fact that, if e is an edge of a triangle T, then

$$|e|\, \|q\|_{L_2(e)}^2 \le C_T \|q\|_{L_2(T)}^2 \quad \text{for any constant function } q, \tag{3.20}$$

where the positive constant C_T depends only on the shape regularity of T.

The estimate (3.15) follows from (3.16)–(3.19). ∎

4. Convergence analysis. We begin with three lemmas that provide estimates to the three terms on the right-hand side of (3.8) and (3.10).

LEMMA 4.1. *Let $\mathring{\boldsymbol{u}} \in H_0(\mathrm{curl};\Omega) \cap H(\mathrm{div}^0;\Omega)$ be the solution of (1.4). It holds that*

$$\inf_{\boldsymbol{v}\in V_h} \|\mathring{\boldsymbol{u}} - \boldsymbol{v}\|_h \le \|\mathring{\boldsymbol{u}} - \Pi_h\mathring{\boldsymbol{u}}\|_h \le C_\epsilon h^{1-\epsilon} \|\boldsymbol{f}\|_{L_2(\Omega)} \tag{4.1}$$

for any $\epsilon > 0$.

Proof. Since $\Pi_e^0 [\![\boldsymbol{n}\times(\mathring{\boldsymbol{u}} - \Pi_h\mathring{\boldsymbol{u}})]\!] = 0$ for all $e \in \mathcal{E}_h$ and $\Pi_e^0 [\![\boldsymbol{n}\cdot(\mathring{\boldsymbol{u}} - \Pi_h\mathring{\boldsymbol{u}})]\!] = 0$ for all $e \in \mathcal{E}_h^i$, we have

$$\begin{aligned} \|\mathring{\boldsymbol{u}} - \Pi_h\mathring{\boldsymbol{u}}\|_h^2 &= \|\nabla_h \times (\mathring{\boldsymbol{u}} - \Pi_h\mathring{\boldsymbol{u}})\|_{L_2(\Omega)}^2 + \|\mathring{\boldsymbol{u}} - \Pi_h\mathring{\boldsymbol{u}}\|_{L_2(\Omega)}^2 \\ &\quad + \sum_{e\in\mathcal{E}_h} \frac{[\Phi_\mu(e)]^2}{|e|} \|[\![\boldsymbol{n}\times(\mathring{\boldsymbol{u}} - \Pi_h\mathring{\boldsymbol{u}})]\!]\|_{L_2(e)}^2 \qquad (4.2) \\ &\quad + \sum_{e\in\mathcal{E}_h^i} \frac{[\Phi_\mu(e)]^2}{|e|} \|[\![\boldsymbol{n}\cdot(\mathring{\boldsymbol{u}} - \Pi_h\mathring{\boldsymbol{u}})]\!]\|_{L_2(e)}^2. \end{aligned}$$

The second term on the right-hand side of (4.2) has been estimated in Lemma 3.8, and the third and fourth terms can be estimated using Lemma 3.9. Therefore it only remains to estimate the first term.

Observe that (2.6) implies

$$\nabla_h \times (\Pi_h \mathring{\boldsymbol{u}}) = \Pi_h^0(\nabla \times \mathring{\boldsymbol{u}}), \tag{4.3}$$

where Π_h^0 is the orthogonal projection from $L_2(\Omega)$ onto the space of piecewise constant functions with respect to $\mathcal{T}_h$. It then follows from (1.6), (2.3), (4.3), and a standard interpolation error estimate [22, 18] that

$$\begin{aligned}\|\nabla_h \times (\mathring{\boldsymbol{u}} - \Pi_h \mathring{\boldsymbol{u}})\|_{L_2(\Omega)}^2 &= \|\nabla \times \mathring{\boldsymbol{u}} - \Pi_h^0(\nabla \times \mathring{\boldsymbol{u}})\|_{L_2(\Omega)}^2 \\ &\le Ch^2 |\nabla \times \mathring{\boldsymbol{u}}|_{H^1(\Omega)}^2 \le Ch^2 \|\boldsymbol{f}\|_{L_2(\Omega)}^2.\end{aligned} \tag{4.4}$$

The estimate (4.1) follows from (4.2), (4.4), and Lemmas 3.8 and 3.9. ∎

LEMMA 4.2. *Let $\mathring{\boldsymbol{u}} \in H_0(\mathrm{curl};\Omega) \cap H(\mathrm{div}^0;\Omega)$ be the solution of (1.4) and let $\mathring{\boldsymbol{u}}_h \in V_h$ satisfy (3.1). It holds that*

$$\sup_{\boldsymbol{w} \in V_h \setminus \{\boldsymbol{0}\}} \frac{a_h(\mathring{\boldsymbol{u}} - \mathring{\boldsymbol{u}}_h, \boldsymbol{w})}{\|\boldsymbol{w}\|_h} \le Ch \|\boldsymbol{f}\|_{L_2(\Omega)}. \tag{4.5}$$

Proof. Let $\boldsymbol{w} \in V_h$ be arbitrary. Using integration by parts, the discrete problem (3.1), the strong form of the reduced curl-curl problem (1.5), and the fact that $\nabla_h \cdot \boldsymbol{w} = 0$, we find

$$\begin{aligned}a_h(\mathring{\boldsymbol{u}} - \mathring{\boldsymbol{u}}_h, \boldsymbol{w}) &= a_h(\mathring{\boldsymbol{u}}, \boldsymbol{w}) - (\boldsymbol{f}, \boldsymbol{w}) \\ &= (Q\boldsymbol{f} - \boldsymbol{f}, \boldsymbol{w}) + \sum_{T \in \mathcal{T}_h} \int_{\partial T} (\nabla \times \mathring{\boldsymbol{u}})(\boldsymbol{n}_T \times \boldsymbol{w}_T)\, ds,\end{aligned} \tag{4.6}$$

where $w_T = w\big|_T$ and $\boldsymbol{n}_T$ is the unit outer normal along ∂T. We see from (4.6) that there are two sources for the inconsistency of the scheme defined by (3.1), namely the projection Q that appears in (1.5) and the discontinuity of the vector fields in V_h.

Equation (4.6) can be rewritten as

$$\begin{aligned}a_h(\mathring{\boldsymbol{u}} - \mathring{\boldsymbol{u}}_h, \boldsymbol{w}) &= (\boldsymbol{f}, Q\boldsymbol{w} - \boldsymbol{w}) + \sum_{e \in \mathcal{E}_h} \int_e (\nabla \times \mathring{\boldsymbol{u}}) [\![\boldsymbol{n} \times \boldsymbol{w}]\!]\, ds \\ &= (\boldsymbol{f}, Q\boldsymbol{w} - \boldsymbol{w}) + \sum_{e \in \mathcal{E}_h} \int_e (\nabla \times \mathring{\boldsymbol{u}} - \overline{(\nabla \times \mathring{\boldsymbol{u}})}_{T_e}) [\![\boldsymbol{n} \times \boldsymbol{w}]\!]\, ds \\ &\quad + \sum_{e \in \mathcal{E}_h} \int_e \overline{(\nabla \times \mathring{\boldsymbol{u}})}_{T_e} (\Pi_e^0 [\![\boldsymbol{n} \times \boldsymbol{w}]\!])\, ds,\end{aligned} \tag{4.7}$$

where $\overline{(\nabla \times \mathring{\boldsymbol{u}})}_{T_e}$ is the mean of $\nabla \times \mathring{\boldsymbol{u}}$ on one of the triangles $\mathcal{T}_e \in \mathcal{T}_h$ that has e as an edge. The first two terms on the right-hand side of (4.7) satisfy the estimate

$$(\boldsymbol{f}, Q\boldsymbol{w} - \boldsymbol{w}) + \sum_{e \in \mathcal{E}_h} \int_e (\nabla \times \mathring{\boldsymbol{u}} - \overline{(\nabla \times \mathring{\boldsymbol{u}})}_{T_e}) [\![\boldsymbol{n} \times \boldsymbol{w}]\!]\, ds \le Ch \|\boldsymbol{f}\|_{L_2(\Omega)} \|\boldsymbol{w}\|_h. \tag{4.8}$$

The derivation of (4.8), which is based on Lemmas 3.10 and 3.11 and (1.6), can be found in the proof of Lemma 6.2 in [15].

Using the Cauchy–Schwarz inequality, (1.6), (3.4), and (3.20), the third term on the right-hand side of (4.7) can be estimated as follows:

$$\sum_{e\in\mathcal{E}_h}\int_e \overline{(\nabla\times\mathring{\boldsymbol{u}})}_{T_e}\big(\Pi_e^0[\![\boldsymbol{n}\times\boldsymbol{w}]\!]\big)\,ds$$

(4.9)

$$\begin{aligned}
&\le \sum_{e\in\mathcal{E}_h}\big(|e|^{1/2}\|\overline{(\nabla\times\mathring{\boldsymbol{u}})}_{T_e}\|_{L_2(e)}\big)\big(|e|^{-1/2}\|\Pi_e^0[\![\boldsymbol{n}\times\boldsymbol{w}]\!]\|_{L_2(e)}\big)\\
&\le Ch\left(\sum_{e\in\mathcal{E}_h}\|\overline{(\nabla\times\mathring{\boldsymbol{u}})}_{T_e}\|^2_{L_2(T_e)}\right)^{1/2}\left(h^{-2}\sum_{e\in\mathcal{E}_h}\frac{1}{|e|}\|\Pi_e^0[\![\boldsymbol{n}\times\boldsymbol{w}]\!]\|^2_{L_2(e)}\right)^{1/2}\\
&\le Ch\|\nabla\times\mathring{\boldsymbol{u}}\|_{L_2(\Omega)}\|\boldsymbol{w}\|_h \le Ch\|\boldsymbol{f}\|_{L_2(\Omega)}\|\boldsymbol{w}\|_h.
\end{aligned}$$

The estimate (4.5) follows from (4.7)–(4.9). ∎

LEMMA 4.3. *Let $\mathring{\boldsymbol{u}}\in H_0(\mathrm{curl};\Omega)\cap H(\mathrm{div}^0;\Omega)$ be the solution of (1.4) and let $\mathring{\boldsymbol{u}}_h\in V_h$ satisfy (3.1). It holds that*

$$\|\mathring{\boldsymbol{u}}-\mathring{\boldsymbol{u}}_h\|_{L_2(\Omega)}\le C_\epsilon\big(h^{2-\epsilon}\|\boldsymbol{f}\|_{L_2(\Omega)}+h^{1-\epsilon}\|\mathring{\boldsymbol{u}}-\mathring{\boldsymbol{u}}_h\|_h\big) \tag{4.10}$$

for any $\epsilon>0$.

Proof. We use a duality argument. Let $\mathring{\boldsymbol{z}}\in H_0(\mathrm{curl};\Omega)\cap H(\mathrm{div}^0;\Omega)$ satisfy the reduced curl-curl problem

$$a_h(\boldsymbol{v},\mathring{\boldsymbol{z}})=(\boldsymbol{v},\mathring{\boldsymbol{u}}-\mathring{\boldsymbol{u}}_h)\qquad\forall\,\boldsymbol{v}\in H_0(\mathrm{curl};\Omega)\cap H(\mathrm{div}^0;\Omega). \tag{4.11}$$

The strong form of (4.11) is

$$\nabla\times(\nabla\times\mathring{\boldsymbol{z}})+\alpha\mathring{\boldsymbol{z}}=Q(\mathring{\boldsymbol{u}}-\mathring{\boldsymbol{u}}_h), \tag{4.12}$$

and we have the following analogue of (1.6):

$$\|\nabla\times\mathring{\boldsymbol{z}}\|_{H^1(\Omega)}\le C_\Omega\|\mathring{\boldsymbol{u}}-\mathring{\boldsymbol{u}}_h\|_{L_2(\Omega)}. \tag{4.13}$$

It follows from (4.12) and integration by parts that

$$a_h(\mathring{\boldsymbol{u}}_h,\mathring{\boldsymbol{z}})=(\mathring{\boldsymbol{u}}_h,Q(\mathring{\boldsymbol{u}}-\mathring{\boldsymbol{u}}_h))+\sum_{e\in\mathcal{E}_h}\int_e(\nabla\times\mathring{\boldsymbol{z}})[\![\boldsymbol{n}\times\mathring{\boldsymbol{u}}_h]\!]\,ds. \tag{4.14}$$

From (4.11) and (4.14), we have

$$\begin{aligned}
\|\mathring{\boldsymbol{u}}-\mathring{\boldsymbol{u}}_h\|^2_{L_2(\Omega)}&=(\mathring{\boldsymbol{u}},\mathring{\boldsymbol{u}}-\mathring{\boldsymbol{u}}_h)-(\mathring{\boldsymbol{u}}_h,\mathring{\boldsymbol{u}}-\mathring{\boldsymbol{u}}_h)\\
&=a_h(\mathring{\boldsymbol{u}}-\mathring{\boldsymbol{u}}_h,\mathring{\boldsymbol{z}})+a_h(\mathring{\boldsymbol{u}}_h,\mathring{\boldsymbol{z}})-(\mathring{\boldsymbol{u}}_h,\mathring{\boldsymbol{u}}-\mathring{\boldsymbol{u}}_h)\\
&=a_h(\mathring{\boldsymbol{u}}-\mathring{\boldsymbol{u}}_h,\mathring{\boldsymbol{z}})-(\mathring{\boldsymbol{u}}_h,(I-Q)(\mathring{\boldsymbol{u}}-\mathring{\boldsymbol{u}}_h))\\
&\quad+\sum_{e\in\mathcal{E}_h}\int_e(\nabla\times\mathring{\boldsymbol{z}})[\![\boldsymbol{n}\times\mathring{\boldsymbol{u}}_h]\!]\,ds.
\end{aligned} \tag{4.15}$$

We will estimate the three terms on the right-hand side of (4.15) separately.

Using (4.7) and the fact that $[\![\boldsymbol{n}\times(\Pi_h\mathring{\boldsymbol{z}})]\!]$ vanishes at the midpoints of all $e\in\mathcal{E}_h$, we can rewrite the first term as

$$\begin{aligned} a_h(\mathring{\boldsymbol{u}}-\mathring{\boldsymbol{u}}_h,\mathring{\boldsymbol{z}}) &= a_h(\mathring{\boldsymbol{u}}-\mathring{\boldsymbol{u}}_h,\mathring{\boldsymbol{z}}-\Pi_h\mathring{\boldsymbol{z}})+a_h(\mathring{\boldsymbol{u}}-\mathring{\boldsymbol{u}}_h,\Pi_h\mathring{\boldsymbol{z}}) \\ &= a_h(\mathring{\boldsymbol{u}}-\mathring{\boldsymbol{u}}_h,\mathring{\boldsymbol{z}}-\Pi_h\mathring{\boldsymbol{z}})+(\boldsymbol{f},Q\Pi_h\mathring{\boldsymbol{z}}-\Pi_h\mathring{\boldsymbol{z}}) \\ &\quad+\sum_{e\in\mathcal{E}_h}\int_e(\nabla\times\mathring{\boldsymbol{u}}-\overline{(\nabla\times\mathring{\boldsymbol{u}})}_{T_e})[\![\boldsymbol{n}\times(\Pi_h\mathring{\boldsymbol{z}})]\!]\,ds, \end{aligned} \tag{4.16}$$

where $\overline{(\nabla\times\mathring{\boldsymbol{u}})}_{T_e}$ is the mean of $\nabla\times\mathring{\boldsymbol{u}}$ on one of the triangles $T_e\in\mathcal{T}_h$ that has e as an edge. Using the expression (4.16), the following estimate was obtained in the proof of Lemma 6.5 of [15]:

$$a_h(\mathring{\boldsymbol{u}}-\mathring{\boldsymbol{u}}_h,\mathring{\boldsymbol{z}})\le C_\epsilon\big(h^{2-\epsilon}\|\boldsymbol{f}\|_{L_2(\Omega)}+h^{1-\epsilon}\|\mathring{\boldsymbol{u}}-\mathring{\boldsymbol{u}}_h\|_h\big)\|\mathring{\boldsymbol{u}}-\mathring{\boldsymbol{u}}_h\|_{L_2(\Omega)}. \tag{4.17}$$

Since $(I-Q)\mathring{\boldsymbol{u}}=\boldsymbol{0}$, the second term on the right-hand side of (4.15) can be estimated by Lemma 3.11:

$$\begin{aligned} -\big(\mathring{\boldsymbol{u}}_h,(I-Q)(\mathring{\boldsymbol{u}}-\mathring{\boldsymbol{u}}_h)\big) &= \big(\mathring{\boldsymbol{u}}-\mathring{\boldsymbol{u}}_h,(I-Q)(\mathring{\boldsymbol{u}}-\mathring{\boldsymbol{u}}_h)\big) \\ &\le C\|\mathring{\boldsymbol{u}}-\mathring{\boldsymbol{u}}_h\|_{L_2(\Omega)}\big(h\|\mathring{\boldsymbol{u}}-\mathring{\boldsymbol{u}}_h\|_h\big). \end{aligned} \tag{4.18}$$

Finally we estimate the third term on the right-hand side of (4.15). We have

$$\begin{aligned} \sum_{e\in\mathcal{E}_h}\int_e(\nabla\times\mathring{\boldsymbol{z}})[\![\boldsymbol{n}\times\mathring{\boldsymbol{u}}_h]\!]\,ds &= \sum_{e\in\mathcal{E}_h}\int_e\big(\nabla\times\mathring{\boldsymbol{z}}-\overline{(\nabla\times\mathring{\boldsymbol{z}})}_{T_e}\big)[\![\boldsymbol{n}\times\mathring{\boldsymbol{u}}_h]\!]\,ds \\ &\quad+\sum_{e\in\mathcal{E}_h}\int_e\overline{(\nabla\times\mathring{\boldsymbol{z}})}_{T_e}\big(\Pi_e^0[\![\boldsymbol{n}\times\mathring{\boldsymbol{u}}_h]\!]\big)\,ds, \end{aligned} \tag{4.19}$$

where $\overline{(\nabla\times\mathring{\boldsymbol{z}})}_{T_e}$ is the mean of $\nabla\times\mathring{\boldsymbol{z}}$ on a triangle $T_e\in\mathcal{T}_h$ that has e as an edge. The following estimate can be found in the proof of Lemma 6.5 of [15]:

$$\sum_{e\in\mathcal{E}_h}\int_e\big(\nabla\times\mathring{\boldsymbol{z}}-\overline{(\nabla\times\mathring{\boldsymbol{z}})}_{T_e}\big)[\![\boldsymbol{n}\times\mathring{\boldsymbol{u}}_h]\!]\,ds\le Ch\|\mathring{\boldsymbol{u}}-\mathring{\boldsymbol{u}}_h\|_{L_2(\Omega)}\|\mathring{\boldsymbol{u}}-\mathring{\boldsymbol{u}}_h\|_h. \tag{4.20}$$

On the other hand, we obtain, by the Cauchy–Schwarz inequality, (3.4), (3.20), and (4.13),

$$\begin{aligned} &\sum_{e\in\mathcal{E}_h}\int_e\overline{(\nabla\times\mathring{\boldsymbol{z}})}_{T_e}\big(\Pi_e^0[\![\boldsymbol{n}\times\mathring{\boldsymbol{u}}_h]\!]\big)\,ds \\ &\quad=\sum_{e\in\mathcal{E}_h}\int_e\overline{(\nabla\times\mathring{\boldsymbol{z}})}_{T_e}\big(\Pi_e^0[\![\boldsymbol{n}\times(\mathring{\boldsymbol{u}}_h-\mathring{\boldsymbol{u}})]\!]\big)\,ds \\ &\quad\le Ch\|\nabla\times\mathring{\boldsymbol{z}}\|_{L_2(\Omega)}\|\mathring{\boldsymbol{u}}-\mathring{\boldsymbol{u}}_h\|_h\le Ch\|\mathring{\boldsymbol{u}}-\mathring{\boldsymbol{u}}_h\|_{L_2(\Omega)}\|\mathring{\boldsymbol{u}}-\mathring{\boldsymbol{u}}_h\|_h. \end{aligned} \tag{4.21}$$

The estimate (4.10) follows from (4.15) and (4.17)–(4.21). ∎

In the case where $\alpha>0$, the following theorem is an immediate consequence of Lemmas 3.5, 4.1, 4.2, and 4.3.

THEOREM 4.4. *Let α be positive. The following estimates hold for the solution $\mathring{\boldsymbol{u}}_h$ of* (3.1):

$$\|\mathring{\boldsymbol{u}} - \mathring{\boldsymbol{u}}_h\|_h \leq C_\epsilon h^{1-\epsilon} \|\boldsymbol{f}\|_{L_2(\Omega)} \qquad \text{for any} \quad \epsilon > 0, \tag{4.22}$$

$$\|\mathring{\boldsymbol{u}} - \mathring{\boldsymbol{u}}_h\|_{L_2(\Omega)} \leq C_\epsilon h^{2-\epsilon} \|\boldsymbol{f}\|_{L_2(\Omega)} \qquad \text{for any} \quad \epsilon > 0. \tag{4.23}$$

In the case where $\alpha \leq 0$, we have the following convergence theorem for the scheme (3.1).

THEOREM 4.5. *Assume that $-\alpha \geq 0$ is not a Maxwell eigenvalue. There exists a positive number h_* such that the discrete problem* (3.1) *is uniquely solvable for all $h \leq h_*$, in which case the following discretization error estimates are valid:*

$$\|\mathring{\boldsymbol{u}} - \mathring{\boldsymbol{u}}_h\|_h \leq C_\epsilon h^{1-\epsilon} \|\boldsymbol{f}\|_{L_2(\Omega)} \qquad \text{for any} \quad \epsilon > 0, \tag{4.24}$$

$$\|\mathring{\boldsymbol{u}} - \mathring{\boldsymbol{u}}_h\|_{L_2(\Omega)} \leq C_\epsilon h^{2-\epsilon} \|\boldsymbol{f}\|_{L_2(\Omega)} \qquad \text{for any} \quad \epsilon > 0. \tag{4.25}$$

Proof. We follow the approach of Schatz [44] for indefinite problems. Assuming $\mathring{\boldsymbol{u}}_h$ satisfies (3.1), it follows from (3.10) and Lemmas 4.1–4.3 that

$$\|\mathring{\boldsymbol{u}} - \mathring{\boldsymbol{u}}_h\|_h \leq C_\epsilon h^{1-\epsilon} \big(\|\boldsymbol{f}\|_{L_2(\Omega)} + \|\mathring{\boldsymbol{u}} - \mathring{\boldsymbol{u}}_h\|_h \big) \quad \text{for any} \quad \epsilon > 0. \tag{4.26}$$

By choosing an $\epsilon_* > 0$, we deduce from (4.26) that, for $h \leq h_* = 1/(2C_{\epsilon_*})^{1/(1-\epsilon_*)}$,

$$\begin{aligned} \|\mathring{\boldsymbol{u}} - \mathring{\boldsymbol{u}}_h\|_h &\leq C_{\epsilon_*} h^{1-\epsilon_*} \big(\|\boldsymbol{f}\|_{L_2(\Omega)} + \|\mathring{\boldsymbol{u}} - \mathring{\boldsymbol{u}}_h\|_h \big) \\ &\leq C_{\epsilon_*} h^{1-\epsilon_*} \|\boldsymbol{f}\|_{L_2(\Omega)} + C_{\epsilon_*} h_*^{1-\epsilon_*} \|\mathring{\boldsymbol{u}} - \mathring{\boldsymbol{u}}_h\|_h \\ &\leq C_{\epsilon_*} h^{1-\epsilon_*} \|\boldsymbol{f}\|_{L_2(\Omega)} + \frac{1}{2} \|\mathring{\boldsymbol{u}} - \mathring{\boldsymbol{u}}_h\|_h, \end{aligned}$$

and hence

$$\|\mathring{\boldsymbol{u}} - \mathring{\boldsymbol{u}}_h\|_h \leq 2C_{\epsilon_*} h^{1-\epsilon_*} \|\boldsymbol{f}\|_{L_2(\Omega)}. \tag{4.27}$$

Therefore, any solution $\mathring{\boldsymbol{z}}_h \in V_h$ of the homogeneous discrete problem

$$a_h(\mathring{\boldsymbol{z}}_h, \boldsymbol{v}) = 0 \qquad \forall\, \boldsymbol{v} \in V_h, \tag{4.28}$$

which corresponds to the special case where $\boldsymbol{f} = \boldsymbol{0} = \mathring{\boldsymbol{z}}$, will satisfy the following special case of (4.27):

$$\|\mathring{\boldsymbol{z}}_h\|_h \leq 0.$$

Hence the only solution of (4.28) is the trivial solution and the discrete problem (3.1) is uniquely solvable for $h \leq h_*$.

The energy error estimate (4.24) now follows from (4.26) and (4.27), and the L_2 error estimate (4.25) follows from Lemma 4.3 and (4.24). ∎

5. Application to the Maxwell eigenproblem. Given any $\boldsymbol{f} \in [L_2(\Omega)]^2$, we define $T\boldsymbol{f} \in H_0(\mathrm{curl};\Omega) \cap H(\mathrm{div}^0;\Omega)$ by the condition that

$$(\nabla \times (T\boldsymbol{f}), \nabla \times \boldsymbol{v}) + (T\boldsymbol{f}, \boldsymbol{v}) = (\boldsymbol{f}, \boldsymbol{v}) \qquad \forall\, \boldsymbol{v} \in H_0(\mathrm{curl};\Omega) \cap H(\mathrm{div}^0;\Omega). \tag{5.1}$$

Clearly T is a bounded linear operator from $[L_2(\Omega)]^2$ into $H_0(\mathrm{curl};\Omega) \cap H(\mathrm{div}^0;\Omega)$. Since $H_0(\mathrm{curl};\Omega) \cap H(\mathrm{div}^0;\Omega)$ is a compact subspace of $[L_2(\Omega)]^2$ (cf. [23]), the operator $T : [L_2(\Omega)]^2 \longrightarrow [L_2(\Omega)]^2$ is symmetric, positive, and compact. Moreover, $(\lambda, \mathring{\boldsymbol{u}})$ satisfy the Maxwell eigenproblem (1.7) if and only if

$$T\mathring{\boldsymbol{u}} = \frac{1}{1+\lambda}\mathring{\boldsymbol{u}}. \tag{5.2}$$

Similarly, let $T_h\boldsymbol{f} \in V_h$ be defined by the condition

$$a_{h,1}(T_h\boldsymbol{f}, \boldsymbol{v}) = (\boldsymbol{f}, \boldsymbol{v}) \qquad \forall\, \boldsymbol{v} \in V_h, \tag{5.3}$$

where $a_{h,1}(\cdot,\cdot)$ is the bilinear form $a_h(\cdot,\cdot)$ in (3.2) with $\alpha = 1$; then

$$T_h\mathring{\boldsymbol{u}}_h = \frac{1}{1+\lambda_h}\mathring{\boldsymbol{u}}_h \tag{5.4}$$

is equivalent to

$$a_{h,0}(\mathring{\boldsymbol{u}}_h, \boldsymbol{v}) = \lambda_h(\mathring{\boldsymbol{u}}_h, \boldsymbol{v}) \qquad \forall\, \boldsymbol{v} \in V_h, \tag{5.5}$$

where $a_{h,0}(\cdot,\cdot)$ is the bilinear form $a_h(\cdot,\cdot)$ in (3.2) with $\alpha = 0$.

From Theorem 4.4 we have

$$\|(T - T_h)\boldsymbol{f}\|_{L_2(\Omega)} \le C_\epsilon h^{2-\epsilon}\|\boldsymbol{f}\|_{L_2(\Omega)} \tag{5.6}$$

for all $\boldsymbol{f} \in [L_2(\Omega)]^2$ and

$$\|(T - T_h)\boldsymbol{f}\|_h \le C_\epsilon h^{1-\epsilon}\|\boldsymbol{f}\|_{L_2(\Omega)} \le C_\epsilon h^{1-\epsilon}\|\boldsymbol{f}\|_h \tag{5.7}$$

for all $\boldsymbol{f} \in H_0(\mathrm{curl};\Omega) \cap H(\mathrm{div}^0;\Omega) + V_h$. Thus the symmetric finite rank operator T_h converges uniformly to the symmetric positive compact operator T as $h \downarrow 0$, which implies that the classical theory of spectral approximation [39, 7] can be applied to the approximation of the eigenvalues and eigenfunctions of T by the eigenvalues and eigenfunctions of T_h. Hence we have the following result through the connections between (1.7) and (5.2) and between (5.5) and (5.4).

THEOREM 5.1. *Let $0 \le \lambda_1 \le \lambda_2 \le \cdots$ be the eigenvalues of (1.7), let $\lambda = \lambda_j = \lambda_{j+1} = \cdots = \lambda_{j+m-1}$ be an eigenvalue with multiplicity m, and let $V_\lambda \subset H_0(\mathrm{curl};\Omega) \cap H(\mathrm{div}^0;\Omega)$ be the corresponding m-dimensional eigenspace. Let $\lambda_{h,1} \le \lambda_{h,2} \le \cdots$ be the eigenvalues of (5.5). Then, as $h \downarrow 0$, we have*

$$|\lambda_{h,\ell} - \lambda| \le C_{\lambda,\epsilon}h^{2-\epsilon}, \qquad \ell = j, \ldots, j+m-1. \tag{5.8}$$

Furthermore, if $V_{h,\lambda} \subset V_h$ is the space spanned by the eigenfunctions corresponding to $\lambda_{h,j}, \ldots, \lambda_{h,j+m-1}$, then the gap between V_λ and $V_{h,\lambda}$ goes to zero at the rate of $C_\epsilon h^{2-\epsilon}$ in the L_2 norm and at the rate of $C_\epsilon h^{1-\epsilon}$ in the norm $\|\cdot\|_h$.

Proof. Let E_λ (resp., $E_{h,\lambda}$) be the L_2 orthogonal projection onto V_λ (resp., $V_{h,\lambda}$). It follows from (5.6) and (5.7) and the classical theory of spectral approximation of compact operators [39, 7] that

$$\|(E_\lambda - E_{h,\lambda})\boldsymbol{w}\|_{L_2(\Omega)} \leq C_\epsilon h^{2-\epsilon}\|\boldsymbol{w}\|_{L_2(\Omega)} \tag{5.9}$$

for all $\boldsymbol{w} \in [L_2(\Omega)]^2$ and

$$\|(E_\lambda - E_{h,\lambda})\boldsymbol{w}\|_h \leq C_\epsilon h^{1-\epsilon}\|\boldsymbol{w}\|_h \tag{5.10}$$

for all $\boldsymbol{w} \in H_0(\mathrm{curl};\Omega) \cap H(\mathrm{div}^0;\Omega) + V_h$.

Let $\boldsymbol{x} \in V_{h,\lambda}$ be a unit eigenfunction of $\lambda_{h,\ell}$ for some ℓ between j and $j+m-1$. Then we have

$$(T_h\boldsymbol{x},\boldsymbol{x}) = \mu_{h,\ell} \quad \text{and} \quad \|\boldsymbol{x}\|_{L_2(\Omega)} = 1, \tag{5.11}$$

where (cf. (5.4))

$$\mu_{h,\ell} = (1+\lambda_{h,\ell})^{-1}. \tag{5.12}$$

Let $\hat{\boldsymbol{x}} = E_\lambda\boldsymbol{x}$ and $\hat{\boldsymbol{y}} = \hat{\boldsymbol{x}} - \boldsymbol{x}$. Then $\hat{\boldsymbol{x}}$ and $\hat{\boldsymbol{y}}$ are orthogonal with respect to both the L_2 inner product and the inner product $(T\cdot,\cdot)$, and the estimate (5.9) implies that

$$\|\hat{\boldsymbol{y}}\|_{L_2(\Omega)} = \|(E_\lambda - E_{h,\lambda})\boldsymbol{x}\|_{L_2(\Omega)} \leq C_\epsilon h^{2-\epsilon}. \tag{5.13}$$

In particular, it follows from Pythagoras' theorem with respect to the L_2 inner product that

$$1 - \|\hat{\boldsymbol{x}}\|^2_{L_2(\Omega)} = \|\boldsymbol{x}\|^2_{L_2(\Omega)} - \|\hat{\boldsymbol{x}}\|^2_{L_2(\Omega)} = \|\hat{\boldsymbol{y}}\|^2_{L_2(\Omega)} \leq C_\epsilon h^{4-\epsilon}. \tag{5.14}$$

Let $\hat{\boldsymbol{e}} = \hat{\boldsymbol{x}}/\|\hat{\boldsymbol{x}}\|_{L_2(\Omega)}$. Then $\hat{\boldsymbol{e}}$ is a unit eigenfunction of T corresponding to the eigenvalue (cf. (5.2))

$$\mu = (1+\lambda)^{-1}, \tag{5.15}$$

and we have

$$|\mu - \mu_{h,\ell}| = \big|(T\hat{\boldsymbol{e}},\hat{\boldsymbol{e}}) - (T_h\boldsymbol{x},\boldsymbol{x})\big| \leq \big|((T-T_h)\boldsymbol{x},\boldsymbol{x})\big| + \big|(T\hat{\boldsymbol{e}},\hat{\boldsymbol{e}}) - (T\boldsymbol{x},\boldsymbol{x})\big|. \tag{5.16}$$

From (5.6) we have

$$\big|((T-T_h)\boldsymbol{x},\boldsymbol{x})\big| \leq C_\epsilon h^{2-\epsilon}, \tag{5.17}$$

and it follows from Pythagoras' theorem with respect to the inner product $(T\cdot,\cdot)$ that

$$\begin{aligned}\big|(T\hat{\boldsymbol{e}},\hat{\boldsymbol{e}}) - (T\boldsymbol{x},\boldsymbol{x})\big| &= \big|(T\hat{\boldsymbol{e}},\hat{\boldsymbol{e}}) - (T\hat{\boldsymbol{x}},\hat{\boldsymbol{x}}) - (T\hat{\boldsymbol{y}},\hat{\boldsymbol{y}})\big| \\ &\leq (1 - \|\hat{\boldsymbol{x}}\|^2_{L_2(\Omega)})(T\hat{\boldsymbol{e}},\hat{\boldsymbol{e}}) + (T\hat{\boldsymbol{y}},\hat{\boldsymbol{y}}).\end{aligned} \tag{5.18}$$

Combining (5.13), (5.14), and (5.18), we find

$$\big|(T\hat{\boldsymbol{e}},\hat{\boldsymbol{e}}) - (T\boldsymbol{x},\boldsymbol{x})\big| \leq C_\epsilon h^{4-\epsilon},$$

which, together with (5.16) and (5.17), implies

$$|\mu - \mu_{h,\ell}| \le C_\epsilon h^{2-\epsilon}. \tag{5.19}$$

The estimate (5.8) follows from (5.12), (5.15), and (5.19).

Recall that the gap $\hat{\delta}(M,N)$ between two subspaces M and N of a normed linear space $(X, \|\cdot\|_X)$ is defined by (cf. [39])

$$\hat{\delta}(M,N) = \max(\delta(M,N), \delta(N,M)),$$

where

$$\delta(M,N) = \sup_{\substack{x\in M \\ \|x\|_X=1}} \inf_{y\in N} \|x-y\|_X.$$

Therefore the statements about the gap between V_λ and $V_{h,\lambda}$ follow immediately from (5.9) and (5.10). ∎

Remark 5.2. The compactness of the solution operator T and the existence of the uniform estimates (5.6) and (5.7) greatly simplify the analysis of the Maxwell eigenproblem. These ingredients are absent from Maxwell spectral approximations based on the full curl-curl problem and hence their justifications are much more involved [10, 11, 8, 20, 21, 9, 19, 12].

6. Numerical results. In this section, we report a series of numerical examples that corroborate our theoretical results. Besides the L_2 error $\|\mathring{\boldsymbol{u}} - \mathring{\boldsymbol{u}}_h\|_{L_2(\Omega)}$ and the energy error $\|\mathring{\boldsymbol{u}} - \mathring{\boldsymbol{u}}_h\|_h$, we also report the errors measured in the seminorm $|\cdot|_{\mathrm{curl}}$ defined by

$$|\boldsymbol{v}|_{\mathrm{curl}} = \|\nabla_h \times \boldsymbol{v}\|_{L_2(\Omega)}.$$

6.1. Source problems. We first demonstrate the performance of our scheme for the source problems where we take α to be $-k^2$ for all the computations in this subsection.

In the first experiment, we check the convergence behavior of our numerical scheme (3.1) on the square $(0, 0.5)^2$ with uniform meshes, where the exact solution is

$$\mathring{\boldsymbol{u}} = [y(y-0.5)\sin(ky), x(x-0.5)\cos(kx)] \tag{6.1}$$

for $k = 0$, 1, and 10. The results tabulated in Table 6.1 confirm the estimates (4.24) and (4.25). Observe also that, as expected, finer meshes are needed for computing satisfactory approximate solutions when the wave numbers become larger.

Since our numerical scheme is designed for solving the reduced curl-curl problem (1.5), its performance should not be affected by the addition to the right-hand side of a gradient term ∇G where $G \in H_0^1(\Omega)$. This is demonstrated by our second experiment. Here we add ∇G with

$$G(x,y) = xy(x-0.5)(y-0.5)\sin(x+y) \tag{6.2}$$

to the right-hand side

$$\boldsymbol{f} = \nabla \times (\nabla \times \mathring{\boldsymbol{u}}) - \mathring{\boldsymbol{u}} \tag{6.3}$$

of the problem in the first experiment (with $k = 1$) and compare their performance. The results are reported in Table 6.2.

TABLE 6.1

Convergence of the scheme (3.1) *on the square* $(0,0.5)^2$ *with uniform meshes and exact solution* $\mathring{\boldsymbol{u}}$ *given by* (6.1).

h	$\frac{\|\mathring{\boldsymbol{u}}-\mathring{\boldsymbol{u}}_h\|_{L_2(\Omega)}}{\|\mathring{\boldsymbol{u}}\|_{L_2(\Omega)}}$	Order	$\frac{\|\mathring{\boldsymbol{u}}-\mathring{\boldsymbol{u}}_h\|_h}{\|\mathring{\boldsymbol{u}}\|_h}$	Order	$\frac{\|\mathring{\boldsymbol{u}}-\mathring{\boldsymbol{u}}_h\|_{\mathrm{curl}}}{\|\mathring{\boldsymbol{u}}\|_{\mathrm{curl}}}$	Order
$k=0$						
1/10	8.11E−02	—	3.61E−01	—	1.70E−01	—
1/20	1.85E−02	2.13	1.76E−01	1.04	8.52E−02	0.99
1/40	4.43E−03	2.07	8.69E−02	1.02	4.25E−02	1.00
1/80	1.08E−03	2.03	4.30E−02	1.01	2.12E−02	1.00
1/160	2.68E−04	2.02	2.15E−02	1.00	1.06E−02	1.00
$k=1$						
1/10	7.19E−02	—	3.45E−01	—	1.64E−01	—
1/20	1.65E−02	2.12	1.68E−01	1.04	8.14E−02	1.01
1/40	3.93E−03	2.07	8.32E−02	1.02	4.09E−02	1.00
1/80	9.60E−04	2.04	4.13E−02	1.01	2.05E−02	1.00
1/160	2.37E−04	2.01	2.06E−02	1.01	1.02E−02	1.00
$k=10$						
1/10	4.82E−01	—	9.11E−01	—	4.33E−01	—
1/20	7.62E−02	2.66	3.65E−01	1.32	2.00E−01	1.12
1/40	1.70E−02	2.17	1.75E−01	1.06	9.89E−02	1.02
1/80	4.07E−03	2.06	8.63E−02	1.02	4.93E−02	1.01
1/160	9.99E−04	2.03	4.32E−02	1.00	2.47E−02	1.00

TABLE 6.2

Comparison of the performance of the scheme (3.1) *on* $(0,0.5)^2$ *with* $k=1$ *for the right-hand sides* $\boldsymbol{f}$ *and* $\boldsymbol{f}+\nabla G$*, where* $\boldsymbol{f}$ *is given by* (6.3) *and* G *is given by* (6.2).

h	rhs $=\boldsymbol{f}$	rhs $=\boldsymbol{f}+\nabla G$
$\|\mathring{\boldsymbol{u}}-\mathring{\boldsymbol{u}}_h\|_{L_2(\Omega)}$		
1/10	**0.001643**092661	**0.001643**020363
1/20	**0.00037**6001607	**0.00037**5992406
1/40	**0.000089765**112	**0.000089764**564
1/80	**0.000021922**450	**0.000021922**530
1/160	**0.000005415**103	**0.000005417**204
$\|\mathring{\boldsymbol{u}}-\mathring{\boldsymbol{u}}_h\|_h$		
1/10	**0.051109**229470	**0.051109**792169
1/20	**0.024919**760832	**0.024919**911664
1/40	**0.012296**880765	**0.012296**952986
1/80	**0.006107**23566	**0.006107**362343
1/160	**0.003043**400530	**0.003043**424464

TABLE 6.3

Loss of convergence without the last two stabilizing terms in the scheme (3.1)*. The computation is on the square* $(0,0.5)^2$ *with uniform meshes and the exact solution* $\mathring{\boldsymbol{u}}$ *is given by* (6.1) *with* $k=1$.

h	$\frac{\|\mathring{\boldsymbol{u}}-\mathring{\boldsymbol{u}}_h\|_{L_2(\Omega)}}{\|\mathring{\boldsymbol{u}}\|_{L_2(\Omega)}}$	Order	$\frac{\|\mathring{\boldsymbol{u}}-\mathring{\boldsymbol{u}}_h\|_{\mathrm{curl}}}{\|\mathring{\boldsymbol{u}}\|_{\mathrm{curl}}}$	Order
1/10	5.17E−00	—	2.61E−01	—
1/20	4.91E−00	0.22	1.71E−01	0.61
1/40	4.78E−00	0.04	1.33E−01	0.37
1/80	4.69E−00	0.03	1.20E−01	0.15

Note that the scheme (3.1) differs from the one proposed in [15] because of the last two stabilizing terms. In the third experiment, we show that these two terms are necessary for the convergence of our method due to the lack of continuity of the approximating vectors across the element interfaces. See Table 6.3 for the numerical evidence.

TABLE 6.4

Convergence of the scheme (3.1) *with graded meshes on the L-shaped domain* $(-0.5,0.5)^2 \setminus [0,0.5]^2$ *with* $k=1$ *and exact solution* $\mathring{\boldsymbol{u}}$ *given by* (6.4).

h	$\frac{\|\mathring{\boldsymbol{u}}-\mathring{\boldsymbol{u}}_h\|_{L_2(\Omega)}}{\|\mathring{\boldsymbol{u}}\|_{L_2(\Omega)}}$	Order	$\frac{\|\mathring{\boldsymbol{u}}-\mathring{\boldsymbol{u}}_h\|_h}{\|\mathring{\boldsymbol{u}}\|_h}$	Order	$\frac{\lvert\mathring{\boldsymbol{u}}-\mathring{\boldsymbol{u}}_h\rvert_{\mathrm{curl}}}{\lvert\mathring{\boldsymbol{u}}\rvert_{\mathrm{curl}}}$	Order
1/4	1.77E+02	—	2.20E+01	—	9.13E−00	—
1/8	4.22E+01	2.07	7.76E−00	1.50	4.52E−00	1.02
1/16	3.42E−00	3.63	2.24E−00	1.79	7.86E−01	2.52
1/32	7.10E−01	2.27	1.09E−00	1.04	4.27E−01	0.88
1/64	1.77E−01	2.01	5.53E−01	0.98	2.34E−01	0.86

The last example in this subsection demonstrates the convergence behavior of our scheme on the L-shaped domain $(-0.5,0.5)^2 \setminus [0,0.5]^2$ with corners $(0.5,0)$, $(0,0)$, $(0,0.5)$, $(-0.5,0.5)$, $(-0.5,-0.5)$, and $(0.5,-0.5)$. We take $k=1$ and the exact solution is chosen to be

$$\mathring{\boldsymbol{u}} = \nabla \times \Big(r^{2/3} \cos\Big(\frac{2}{3}\theta - \frac{\pi}{3}\Big)\phi(r/0.5)\Big), \tag{6.4}$$

where (r,θ) are the polar coordinates at the origin and the cutoff function is given by

$$\phi(r) = \begin{cases} 1, & r \le 0.25, \\ -16(r-0.75)^3 & \\ \quad \times\big[5+15(r-0.75)+12(r-0.75)^2\big], & 0.25 \le r \le 0.75, \\ 0, & r \ge 0.75. \end{cases}$$

The meshes are graded around the reentrant corner with the grading parameter equal to 1/3. The results are tabulated in Table 6.4 and they agree with the estimates (4.24) and (4.25).

6.2. Maxwell eigenproblem. In this subsection, we report the numerical results for the Maxwell eigenproblem.

In the first experiment, the computation is carried out on the square domain $(0,\pi)^2$ with uniform meshes. In this case, the exact eigenvalues are r^2+s^2, $r,s = 0,1,2,3,4,\ldots$ with $r^2+s^2 \neq 0$. For instance, the first ten eigenvalues are 1, 1, 2, 4, 4, 5, 5, 8, 9, 9. In Figure 6.1, we plot the first twenty numerical eigenvalues versus the parameter $n=\pi/h$. The symbol "o" on the right side denotes the exact eigenvalue, and "(2)" indicates that the multiplicity of the eigenvalue is 2. All the numerical approximations converge at a second order rate. To save space, only the first five numerical eigenvalues are included in Table 6.5. Furthermore, there is no spurious eigenvalue in our results.

In the second experiment, we compute the eigenvalues for the L-shaped domain $(-0.5,0.5)^2 \setminus [0,0.5]^2$ and the meshes are graded around the reentrant corner with the grading parameter equal to 1/3. Table 6.6 contains the first five numerical eigenvalues, which show second order convergence of our method. This agrees with our analysis. In this case the "exact eigenvalues" are derived from numerical results of Dauge, which can be found at http://perso.univ-rennes1.fr/monique.dauge/core/index.html. In Figure 6.2, we plot the first ten numerical eigenvalues versus the parameter $n = 1/(2h)$. Again, there is no spurious eigenvalue.

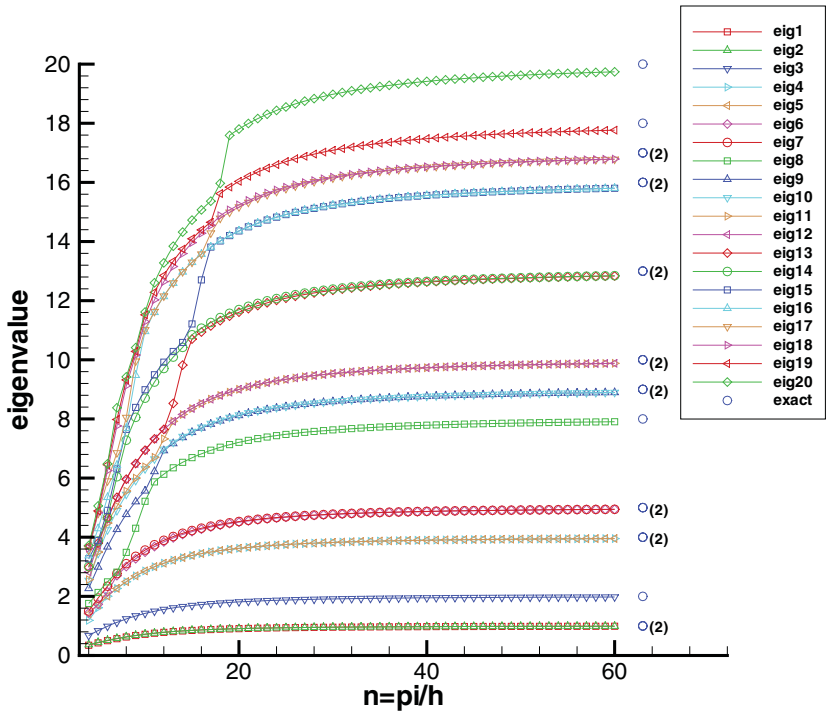


FIG. 6.1. *First twenty numerical eigenvalues versus* $n = \pi/h$ *for Maxwell operator on* $(0,\pi)^2$.

TABLE 6.5

Convergence of the first five numerical eigenvalues on $(0,\pi)^2$ *with* γ = *order of convergence and* λ = *exact eigenvalue.*

h	1st	γ	2nd	γ	3rd	γ	4th	γ	5th	γ
$\pi/4$	0.347	—	0.367	—	0.673	—	1.188	—	1.409	—
$\pi/8$	0.626	0.80	0.634	0.79	1.239	0.80	2.495	0.90	2.496	0.78
$\pi/16$	0.863	1.45	0.866	1.45	1.723	1.46	3.446	1.44	3.446	1.44
$\pi/32$	0.962	1.84	0.963	1.84	1.923	1.84	3.845	1.83	3.845	1.83
$\pi/64$	0.990	1.97	0.990	1.97	1.980	1.97	3.960	1.96	3.960	1.96
$\pi/128$	0.998	1.97	0.998	2.00	1.995	2.01	3.990	1.99	3.990	1.99
λ	1		1		2		4		4	

TABLE 6.6

Convergence of the first five numerical eigenvalues on the L-shaped domain $(-0.5,0.5)^2\backslash[0,0.5]^2$ *with* graded meshes, $\mu = 1/3$ *at the reentrant corner,* γ = *order of convergence, and* λ = *exact eigenvalue.*

h	1st	γ	2nd	γ	3rd	γ	4th	γ	5th	γ
1/4	2.626	—	3.442	—	5.087	—	5.174	—	5.656	—
1/8	4.747	1.50	9.963	1.36	11.233	0.28	11.316	0.28	12.226	0.26
1/16	5.574	1.81	13.102	2.01	30.631	1.67	36.653	3.32	38.548	2.25
1/32	5.819	1.98	13.875	1.98	37.008	1.84	38.822	2.11	44.091	2.26
1/64	5.882	2.02	14.071	2.00	38.847	1.97	39.319	2.04	45.204	2.05
λ	5.902		14.136		39.478		39.478		45.558	

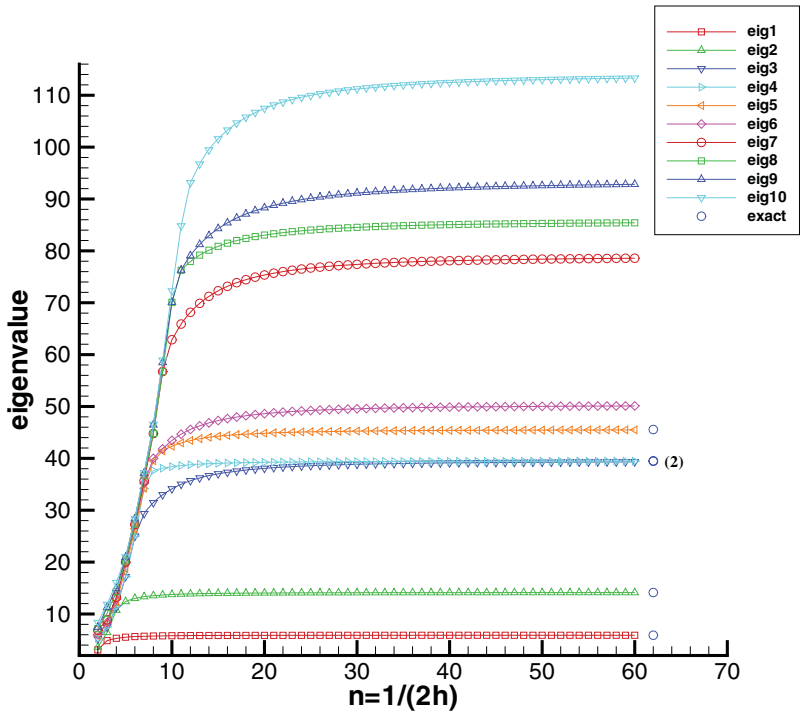


FIG. 6.2. *First ten numerical eigenvalues versus* $n = 1/(2h)$ *for the Maxwell operator on* $(-0.5, 0.5)^2 \setminus [0, 0.5]^2$.

7. Concluding remarks. We have analyzed an interior penalty method with weak overpenalization for the two-dimensional reduced curl-curl problem and the Maxwell eigenproblem. This method is inconsistent but stable without the need of tuning any penalty parameter. We expect other discontinuous Galerkin schemes will be discovered along this line. An advantage of this new approach is that fast solvers for the resulting schemes can be constructed using techniques developed for elliptic problems.

The scheme developed in this paper can be extended to three-dimensional domains. In the case of uniform meshes, our two-dimensional analysis can be easily generalized to establish convergence at a suboptimal rate. It is likely that quasi-optimal convergence rates can be recovered using graded meshes, which is of course much more complicated in three dimensions [2, 3]. This will be one of the topics of our ongoing research.

Acknowledgment. Part of the research in this paper was carried out while the first and third authors were visiting the Humboldt Universität zu Berlin, and they would like to thank the members of the Institut für Mathematik for their hospitality.

REFERENCES

[1] TH. APEL, *Anisotropic Finite Elements*, Teubner, Stuttgart, 1999.

[2] Th. Apel and S. Nicaise, *The finite element method with anisotropic mesh grading for elliptic problems in domains with corners and edges*, Math. Methods Appl. Sci., 21 (1998), pp. 519–549.

[3] Th. Apel, S. Nicaise, and J. Schöberl, *Crouzeix-Raviart type finite elements on anisotropic meshes*, Numer. Math., 89 (2001), pp. 193–223.

[4] Th. Apel, A.-M. Sändig, and J. R. Whiteman, *Graded mesh refinement and error estimates for finite element solutions of elliptic boundary value problems in non-smooth domains*, Math. Methods Appl. Sci., 19 (1996), pp. 63–85.

[5] F. Assous, P. Ciarlet, Jr., S. Labrunie, and S. Lohrengel, *The singular complement method*, in Domain Decomposition Methods in Science and Engineering, N. Debit, M. Garbey, R. Hoppe, D. Keyes, Y. Kuznetsov, and J. Périaux, eds., CIMNE, Barcelona, 2002, pp. 161–189.

[6] F. Assous, P. Ciarlet, Jr., and E. Sonnendrücker, *Resolution of the Maxwell equation in a domain with reentrant corners*, M2AN Math. Model. Numer. Anal., 32 (1998), pp. 359–389.

[7] I. Babuška and J. Osborn, *Eigenvalue problems*, in Handbook of Numerical Analysis II, P. G. Ciarlet and J. L. Lions, eds., North–Holland, Amsterdam, 1991, pp. 641–787.

[8] D. Boffi, *Fortin operators and discrete compactness for edge elements*, Numer. Math., 87 (2000), pp. 229–246.

[9] D. Boffi, *Approximation of eigenvalues in mixed form, discrete compactness property, and application to hp mixed finite elements*, Comput. Methods Appl. Mech. Engrg., 196 (2007), pp. 3672–3681.

[10] D. Boffi, F. Brezzi, and L. Gastaldi, *On the convergence of eigenvalues for mixed formulations*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 25 (1997), pp. 131–154.

[11] D. Boffi, P. Fernandes, L. Gastaldi, and I. Perugia, *Computational models of electromagnetic resonators: Analysis of edge element approximation*, SIAM J. Numer. Anal., 36 (1999), pp. 1264–1290.

[12] D. Boffi, F. Kikuchi, and J. Schöberl, *Edge element computation of Maxwell's eigenvalues on general quadrilateral meshes*, Math. Models Methods Appl. Sci., 16 (2006), pp. 265–273.

[13] A. Bossavit, *Computational Electromagnetism*, Academic Press, San Diego, 1998.

[14] S. C. Brenner and C. Carstensen, *Finite element methods*, in Encyclopedia of Computational Mechanics, E. Stein, R. de Borst, and T. J. R. Hughes, eds., Wiley, Weinheim, Germany, 2004, pp. 73–118.

[15] S. C. Brenner, F. Li, and L.-Y. Sung, *A locally divergence-free nonconforming finite element method for the time-harmonic Maxwell equations*, Math. Comp., 76 (2007), pp. 573–595.

[16] S. C. Brenner and L. Owens, *A weakly over-penalized nonsymmetric interior penalty method*, J. Numerical Analysis, Industrial and Applied Mathematics, to appear.

[17] S. C. Brenner, L. Owens, and L.-Y. Sung, *A Weakly over-penalized symmetric interior penalty method*, preprint, Department of Mathematics, Louisiana State University, LA, 2007.

[18] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, 2nd ed., Springer-Verlag, New York, Berlin, Heidelberg, 2002.

[19] A. Buffa and I. Perugia, *Discontinuous Galerkin approximation of the Maxwell eigenproblem*, SIAM J. Numer. Anal., 44 (2006), pp. 2198–2226.

[20] S. Caorsi, P. Fernandes, and M. Raffetto, *On the convergence of Galerkin finite element approximations of electromagnetic eigenproblems*, SIAM J. Numer. Anal., 38 (2000), pp. 580–607.

[21] S. Caorsi, P. Fernandes, and M. Raffetto, *Spurious-free approximations of electromagnetic eigenproblems by means of Nédélec-type elements*, M2AN Math. Model. Numer. Anal., 35 (2001), pp. 331–358.

[22] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North–Holland, Amsterdam, 1978.

[23] M. Costabel, *A remark on the regularity of solutions of Maxwell's equations on Lipschitz domains*, Math. Methods Appl. Sci., 12 (1990), pp. 365–368.

[24] M. Costabel, *A coercive bilinear form for Maxwell's equations*, J. Math. Anal. Appl., 157 (1991), pp. 527–541.

[25] M. Costabel and M. Dauge, *Maxwell and Lamé eigenvalues on polyhedra*, Math. Methods Appl. Sci., 22 (1999), pp. 243–258.

[26] M. Costabel and M. Dauge, *Singularities of electromagnetic fields in polyhedral domains*, Arch. Ration. Mech. Anal., 151 (2000), pp. 221–276.

[27] M. Costabel and M. Dauge, *Weighted regularization of Maxwell equations in polyhedral domains*, Numer. Math., 93 (2002), pp. 239–277.

[28] M. Costabel, M. Dauge, and S. Nicaise, *Singularities of Maxwell interface problems*, M2AN Math. Model. Numer. Anal., 33 (1999), pp. 627–649.
[29] M. Costabel, M. Dauge, and C. Schwab, *Exponential convergence of hp-FEM for Maxwell equations with weighted regularization in polygonal domains*, Math. Models Methods Appl. Sci., 15 (2005), pp. 575–622.
[30] M. Crouzeix and P.-A. Raviart, *Conforming and nonconforming finite element methods for solving the stationary Stokes equations* I, RAIRO Anal. Numér., 7 (1973), pp. 33–75.
[31] P. Grisvard, *Elliptic Problems in Non Smooth Domains*, Pitman, Boston, 1985.
[32] C. Hazard and S. Lohrengel, *A singular field method for Maxwell's equations: Numerical aspects for* 2*D magnetostatics*, SIAM J. Numer. Anal., 40 (2002), pp. 1021–1040.
[33] R. Hiptmair, *Finite elements in computational electromagnetism*, Acta Numer., 11 (2002), pp. 237–339.
[34] P. Houston, I. Perugia, A. Schneebeli, and D. Schötzau, *Interior penalty method for the indefinite time-harmonic Maxwell equation*, Numer. Math., 100 (2005), pp. 485–518.
[35] P. Houston, I. Perugia, A. Schneebeli, and D. Schötzau, *Mixed discontinuous Galerkin approximation of the Maxwell operator: The indefinite case*, M2AN Math. Model. Numer. Anal., 39 (2005), pp. 727–753.
[36] P. Houston, I. Perugia, and D. Schötzau, *Mixed discontinuous Galerkin approximation of the Maxwell operator*, SIAM J. Numer. Anal., 42 (2004), pp. 434–459.
[37] P. Houston, I. Perugia, and D. Schötzau, *Mixed discontinuous Galerkin approximation of the Maxwell operator: Non-stabilized formulation*, J. Sci. Comput., 22/23 (2005), pp. 315–346.
[38] J.-M. Jin, *The Finite Element Method in Electromagnetics*, 2nd ed., Wiley, New York, 2002.
[39] T. Kato, *Perturbation Theory of Linear Operators*, Springer-Verlag, Berlin, 1966.
[40] P. Monk, *Finite Element Methods for Maxwell's Equations*, Oxford University Press, New York, 2003.
[41] J.-C. Nédélec, *Mixed finite elements in* $\mathbf{R}^3$, Numer. Math., 35 (1980), pp. 315–341.
[42] J.-C. Nédélec, *A new family of mixed finite elements in* $\mathbf{R}^3$, Numer. Math., 50 (1986), pp. 57–81.
[43] I. Perugia, D. Schötzau, and P. Monk, *Stabilized interior penalty methods for the time-harmonic Maxwell equations*, Comput. Methods. Appl. Mech. Engrg., 191 (2002), pp. 4675–4697.
[44] A. Schatz, *An observation concerning Ritz-Galerkin methods with indefinite bilinear forms*, Math. Comp., 28 (1974), pp. 959–962.

SIAM J. NUMER. ANAL.
Vol. 46, No. 3, pp. 1212–1249
© 2008 Society for Industrial and Applied Mathematics

ON FINITE ELEMENT METHODS FOR FULLY NONLINEAR ELLIPTIC EQUATIONS OF SECOND ORDER*

KLAUS BÖHMER†

Dedicated to my highly esteemed teacher and friend, Prof. Dr. Karl Nickel, on the occasion of his 80th birthday

Abstract. For the first time, we present for the general case of fully nonlinear elliptic differential equations of second order a nonstandard C^1 finite element method (FEM). We consider, throughout the paper, two cases in parallel: For convex, bounded, polyhedral domains in $\mathbf{R}^n$, or for C^2 bounded domains in $\mathbf{R}^2$, we prove stability and convergence for the corresponding conforming or nonconforming C^1 FEM, respectively. The results for equations and systems of orders 2 and $2m$ and quadrature approximations appear elsewhere. The classical theory of discretization methods is applied to the differential operator or the combined differential and the boundary operator. The consistency error for satisfied or violated boundary conditions on polyhedral or curved domains has to be estimated. The stability has to be proved in an unusual way. This is the hard core of the paper. Essential tools are linearization, a compactness argument, the interplay between the weak and strong form of the linearized operator, and a new regularity result for solutions of finite element equations. An essential basis for our proofs are Davydov's results for C^1 FEs on polyhedral domains in $\mathbf{R}^n$ or of local degree 5 for C^2 domains in $\mathbf{R}^2$. Better convergence and extensions to $\mathbf{R}^n$ for C^2 domains are to be expected from his forthcoming results on curved domains. Our proof for the second case in $\mathbf{R}^n$, includes the first essentially as a special case. The method applies to quasi-linear elliptic problems not in divergence form as well. A discrete Newton method is shown to converge locally quadratically, essentially independently of the actual grid size by the mesh independence principle.

Key words. conforming, nonconforming finite element method, $\mathbf{R}^n$, variational crimes, fully nonlinear elliptic differential equations, order 2, C^1 finite elements, quasi-linear elliptic equations not in divergence form, stability, convergence, regularity for finite element solutions, discrete Newton method, locally quadratic convergence

AMS subject classifications. 35J20, 35J25, 35J60, 35J65, 41A15, 46N40, 47N40, 65D18, 65H10, 65N12, 65N22, 65N30

DOI. 10.1137/040621740

1. Introduction.

We present a finite element method (FEM) for the general case of fully nonlinear elliptic differential equations of second order in $\mathbf{R}^n, n \geq 2$. To the knowledge of many specialists, consulted by the author, this seems to be the first such paper. The generalization to equations and systems of orders 2 and $2m$ and quadrature approximations is to appear elsewhere; see Böhmer [11]. Here we give a short summary of the main ideas behind the formulation and the proofs for our "nonstandard" FEM. The standard approaches for FEMs do not work; see below. So we proceed as follows:

- Our FEM for general fully nonlinear elliptic differential equations requires a bounded, convex, polyhedral or a C^2 domain in $\mathbf{R}^n$ with conforming and nonconforming C^1 FEs. Both cases are presented in parallel.
- Only FEMs for strong, not for weak, forms make sense. This requires FEs in C^1, satisfying and violating Dirichlet boundary conditions on polyhedral and C^2 domains.

*Received by the editors December 29, 2004; accepted for publication (in revised form) June 7, 2007; published electronically March 21, 2008.
http://www.siam.org/journals/sinum/46-3/62174.html
†Philipps Universität, Fachbereich Mathematik und Informatik, 35032 Marburg, Germany (boehmer@mathematik.uni-marburg.de).

- The classical theory of discretization methods has to be applied to the differential operator and, on C^2 domains, to the differential and the boundary operator.
- The most difficult claim, the stability of the nonlinear operator, is proved via the stability of the linearized operator, with coinciding strong and weak bilinear FE forms. Compact perturbation and a new regularity result for FE solutions of weak linear FE equations lift the "weak stability" back to the "strong stability."
- For computations the weak linear forms, coinciding with the strong forms, are used. Thus the many known powerful numerical methods apply for our FEM as well.
- The mesh independence principle (MIP) guarantees a quadratically convergent solution process via the Newton method for the highly nonlinear systems of equations.

Our FEM is a natural translation of the nonstandard situation for fully nonlinear elliptic differential equations, including quasi-linear equations not in divergence form; we abbreviate this as *nondivergent quasi-linear equations.* As for fully nonlinear equations, weak forms are not possible. The solution has to be determined from the original strong form with derivatives up to second order. This situation is directly translated into our strong form of a FEM; see section 2 for more details. This FEM, based upon the classical discretization theory, strongly differs from the earlier approaches.

Difference methods for viscosity solutions are studied by Crandall and Lions [23], Crandall et al. [25], and Kuo and Trudinger [41] with discrete maximum principles.

For FEMs, there are several papers with results for the Monge–Ampère equations in $\mathbf{R}^2$ or modifications. Oliker and Prussner [44] prove, by convexity arguments, that their sequences converge monotonically to the solution. Fulton [34] studies multigrid solutions as models of nonlinear balance equations in meteorology. Dean and Glowinski [32, 33] reformulate it as a problem of calculus of variations involving the biharmonic (or a related) operator. A reinterpretation as saddle-point problem for a well-chosen augmented Lagrangian functional leads to iterative methods such as that of Uzawa, Douglas, and Rachford, applicable to the related Pucci equation.

These FE approaches are based upon either combinations of the special properties of a specific problem with the proposed numerical method or upon maximum and/or monotonicity arguments. But none of them seems to be applicable nor has any been applied to the general fully nonlinear elliptic differential equations (and systems) as treated in this paper. This seems to indicate the need for new ideas.

2. Main ideas and results for the new FEM: An extended summary. We propose an approach totally different from those in the above papers. As mentioned, a major part of the analysis for fully nonlinear differential equations is based upon the strong form of these equations; see, e.g., Gilbarg and Trudinger [35], Showalter [49], and Taylor [54, 55, 56]. This motivates a FEM, based upon the strong form of the differential equation. For proving existence, uniqueness, and related results for the exact solution, e.g., monotonicity, maximum principles, and the Schauder fixed point theorem are needed. We avoid this, but employ, as for many other numerical approaches, these results as the basis for our method.

We simplify the presentation by considering only FEMs for second order equations; see Böhmer [11] for the general case. We prove convergence in the framework of general discretization theory, based upon consistency and stability (see, e.g., Stetter

[51]), in a first step fitted to the present problem. We need the strong form for the FEM and (see below) a convex, bounded (and we assume additionally) polyhedral Ω or an $\Omega \in C^2$; both cases are discussed here. The strong form requires C^1 FEs for an L^2 evaluation of the differential equation. For a polyhedral Ω, (trivial) boundary conditions are satisfied exactly; curved domains cause violated boundary conditions.

The second step is the proof of consistency. For polyhedral Ω and their conforming FEs this is a straightforward consequence of a vanishing interpolation error for a continuous differential operator.

For smooth $\Omega \in C^2$, requiring nonconforming FEs, consistency is more complicated. The earlier approaches (see, e.g., Lenoir [42] and Brenner and Scott [13] for linear problems) do not seem to be applicable to the fully nonlinear problems. So other nonstandard, not surprisingly similarly complicated considerations are necessary. The clue is an operator with two components, the differential and the boundary operator. This requires a more complicated version of the general discretization theory. Specifically, we have to carefully introduce the necessary projectors and spaces for including the violated differential operator *and* the boundary conditions into the classical discretization theory. Then the "classical consistency error" for the differential and the boundary operator can be estimated in a simple proof; see (6.29) and Theorem 7.1 below. Essentially we split this error into the FE error for a conforming FEM on an $\Omega_c^h \approx \Omega$ and its complementary *geometric error,* due to $\partial\Omega \neq \partial\Omega_c^h$, according to a new idea of Tiihonen [57].

The most difficult part is the proof of stability. We did not see a direct possibility for the strong form. So we combine linearization with regularity of FE solutions. This is another difference of our FEM to the standard methods and those in the previous papers. Stability is first proved for the linearized weak form. Unfortunately, we lose the powerful machinery of weak bilinear forms and their FEMs, *unless* we find a kind of *connection* between the two forms. This is achieved by Proposition 8.2 and Lemma 8.5. The first shows coinciding weak and strong FE bilinear forms in our case. The latter states, similarly to the exact solution of a "smooth" linear elliptic problem, a new regularity and stability result for our FE solutions. This and the available existence, uniqueness, and regularity results require a convex, bounded, polyhedral Ω or an $\Omega \in C^2$, implying satisfied or violated boundary conditions for our FEM. Generalizing our earlier compact perturbation techniques (see Böhmer and Sassmannshausen [12] and Böhmer [9, 8]) to satisfied or violated boundary conditions allows lifting the stability from coercive to general elliptic bilinear forms. A combination of these results lifts the weak back to the strong stability and allows the proof of convergence.

We do need convex, bounded (polyhedral) or smooth $\Omega \in C^2$ for existence, convergence, and the regularity lemma, Lemma 8.5. The smooth $\Omega \in C^2$ requires and allows the extension operator in Theorem 4.5. We require FEs in C^1, since for FEs in C^0 we did not see a possibility to handle the interior variational crimes. The standard interplay between the weak and strong form is impossible for fully nonlinear problems and *excludes, e.g., discontinuous Galerkin methods.* Thus we choose the unusual C^1 FEs, conforming and nonconforming for polyhedral and C^2 domains, respectively. Davydov [27] has proved the complete approximation results on bounded, polyhedral Ω in R^n. For $\Omega \in C^2$ we need the new "stable splitting" of C^1 FEs into interior and boundary elements. Presently, these results are available for modified Argyris piecewise polynomials of degree 5 on polygonal $\Omega^h \approx \Omega \subset \mathbf{R}^2$. This restricts, for $\Omega \in C^2$, the order of convergence of our FEM to 2. Davydov is extending his results

to higher degrees and $\mathbf{R}^n$. Our proof is formulated for including these extensions as well.

The FE equations have to be set up. They are solved numerically by a quadratically converging discrete Newton method. Its local quadratic convergence is guaranteed by the MIP; see Allgower and Böhmer [2, 3], Allgower et al. [4], and section 9. It starts with a good approximation, e.g., obtained by a continuation method. Again linear problems are essential to solve the highly nonlinear systems of equations. As a welcome consequence of our coinciding weak and strong FE bilinear forms, the standard solution techniques for weak linear problems are applicable to the fully nonlinear equations as well. However, the standard C^0 FEs have to be replaced by C^1 FEs.

With the notation introduced below, let G be a uniformly elliptic operator on Ω; see (3.8), (3.10). Homogeneous Dirichlet conditions are obtained by the standard techniques. Violated boundary conditions for the nonconforming FEM below will play a dominant role. The exact solution, u_0, is determined by the differential equation and the boundary condition in the trace sense:

$$\begin{aligned} u_0 \in \mathcal{U} := H^2(\Omega): \quad & G(u_0)(x) := G^w(x, u_0(x), Du_0(x), D^2u_0(x)) = 0 \quad \forall x \in \Omega \\ & \iff (G(u_0), v)_{L^2(\Omega)} = 0 \quad \forall v \in L^2(\Omega) = (L^2(\Omega))', \\ & \text{and } u_0|_{\partial\Omega} = 0 \iff (u_0|_{\partial\Omega}, v_b)_{L^2(\partial\Omega)} = 0 \quad \forall v_b \in L^2(\partial\Omega), \end{aligned} \tag{2.1}$$

with $G^w : \Omega' \times \mathbf{R} \times \mathbf{R}^n \times \mathbf{R}^{n^2} \to \mathbf{R}$, $\overline{\Omega}$, $\overline{\Omega_c^h} \subset \Omega'$ (see (3.6)), and so $G(u^h)(x) := G^w(x, u^h(x), Du^h(x), D^2u^h(x))$ is defined as well. Since $H^{3/2}(\partial\Omega)$ is dense in $L^2(\partial\Omega)$ testing with only for all $v_b \in L^2(\partial\Omega)$ is correct. For a systematic study of the differential and the boundary operator, we abbreviate (2.1) into an equivalent form, determining

$$\begin{aligned} & u_0 \in \mathcal{U}: \quad 0 = F(u_0) := (G(u_0)|_\Omega = 0, u_0|_{\partial\Omega} = 0) \\ \iff & (G(u_0), v)_{L^2(\Omega)} + (u_0|_{\partial\Omega}, v_b)_{L^2(\partial\Omega)} = 0 \quad \forall v \in L^2(\Omega), \quad v_b \in L^2(\partial\Omega). \end{aligned} \tag{2.2}$$

We introduce the C^1 FE spaces of local degree d in $\mathbf{R}^n$, the $\mathcal{S}_d^1(\mathcal{T}_c^h)$, for polyhedral, and $d \geq 5$ for C^2 domains. For a unified notation we assume that they are defined on a triangulation $\mathcal{T}_c^h$ of the original Ω, a polyhedral Ω^h, or a curved $\Omega_c^h \approx \Omega$, with the same notation $\mathcal{T}_c^h, \Omega_c^h$ for all cases; cf. section 4 for $S^h \subset \mathcal{S}_d^1(\mathcal{T}_c^h)$:

$$\mathcal{U}^h := S^h \subset \mathcal{S}_d^1(\mathcal{T}_c^h) \text{ and } \mathcal{U}_0^h := \mathcal{V}^h := \{u^h \in S^h : u^h|_{\partial\Omega_c^h} = 0\} \subset \mathcal{U}^h \neq \mathcal{V}^h. \tag{2.3}$$

For convex polyhedral Ω, the conforming FE solution, $u_0^h \in \mathcal{U}_0^h$, is determined by testing only the differential equation w.r.t. $v^h \in \mathcal{V}^h$, again in the L^2 sense:

$$u_0^h \in \mathcal{U}_0^h = \mathcal{V}^h \text{ s.t. } (G(u_0^h), v^h)_{L^2(\Omega_c^h)} = 0 \quad \forall v^h \in \mathcal{V}^h. \tag{2.4}$$

For smooth $\Omega \in C^2$ we have to test both operators. So the FE solution satisfies

$$u_0^h \in \mathcal{U}^h \text{ s.t. } (G(u_0^h), v^h)_{L^2(\Omega_c^h)} = 0 \quad \forall v^h \in \mathcal{V}^h, \text{ and } u_0^h|_{\partial\Omega_c^h} = 0, \text{ and so } u_0^h \in \mathcal{V}^h; \tag{2.5}$$

see Theorem 8.1. The FEM (cf. (2.2)) determines $u_0^h \in \mathcal{U}^h$ and defines F^h s.t.

$$(G(u_0^h), v^h)_{L^2(\Omega_c^h)} + (u_0^h|_{\partial\Omega_c^h}, v_b^h)_{L^2(\partial\Omega_c^h)} = 0 \quad \forall v^h \in \mathcal{V}^h, \quad v_b^h \in \mathcal{V}_b^h \Leftrightarrow 0 = F^h(u_0^h), \tag{2.6}$$

with $\mathcal{V}_b^h := \{u^h|_{\partial\Omega_c^h} : u^h \in \mathcal{U}^h\}$. For $\Omega \in C^2$, comparing u_0^h and u_0 requires extension operators, e.g., $E_c : H^2(\Omega) \to H^2(\Omega \cup \Omega_c^h)$.

Our FEM transforms $F(u_0) = 0$ in (2.2) into $F^h(u_0^h) = 0$ in (2.6). It is a classical discretization method in the sense of, e.g., Stetter [51]. Related approaches for general discretization methods are discussed in Stummel [52, 53], Vainikko [59], Keller [40], Reinhardt [46], Zeidler [62, 63, 64], and Böhmer [7, 9, 8]. Here Theorems 1.2.3–1.2.5 in Stetter [51] provide the tools for showing convergence if consistency and stability can be proved. Thus sections 4–8 are the core of this paper.

For our unusual FEM, we summarize the main ideas. We mainly formulate the results for the nonconforming FEM for $\Omega \in C^2$. The corresponding reduction to the conforming FEs for polyhedral Ω is obvious.

1. *The FEMs, transforming* (2.1) *into* (2.4)*, and* (2.2) *into* (2.6)*, are general discretization methods:*

- The existence, uniqueness, and regularity results for the original problems in section 3 are the basic results for a general discretization method.
- The fully nonlinear elliptic problem (2.1), (2.2) is approximated by the *strong* form (2.4), (2.6), requiring FEs in C^1. For (2.4) identical ansatz and test functions, $\mathcal{U}_0^h = \mathcal{V}^h$, are possible. For (2.6) we choose as ansatz functions $\mathcal{U}^h = S^h$ and test the differential operator with $v^h \in \mathcal{V}^h = S^h \cap H_0^1(\Omega_c^h)$ and the boundary operator with the restrictions to the boundary, $v_b^h \in \mathcal{V}_b^h = S^h|_{\partial\Omega_c^h}$; see (2.3) and section 4 for the relevant extension results in Theorem 4.5.
- The FEM is a general linear discretization method; cf. sections 5 and 6.
- In subsections 5.1–5.3 and 6.1–6.3 we reformulate the FEM, introduce the necessary projectors, and verify the required conditions. Thus the FEMs are convergent if they are stable and consistent; see Theorem 5.1.
- For a "smooth" problem our FEM is consistent; see (5.15) and Theorem 7.1.

2. *Stability for the nonlinear* strong *equation* (2.6) *is the most difficult problem; cf. section* 8*:*

- The nonlinear stability is a consequence of the stability for the *linearized strong* problem; see (2.7) and Theorem 8.1. For given u_0^h and f, ϕ, determine the FE solution u_1^h from

$$u_1^h \in \mathcal{U}^h \text{ s.t. } (F^h)'(u_0^h)u_1^h = \big(G'(u_0^h)u_1^h, u_1^h|_{\partial\Omega_c^h}\big) = (f, \phi). \tag{2.7}$$

  For this linearized problem, two FEMs complement each other; see below:
- The first component of the linearized *strong* and *weak* bilinear forms coincide:

$$(G'(u_0^h)u^h,\ v^h)_{L^2(\Omega_c^h)} = \langle G'(u_0^h)u^h,\ v^h\rangle \quad \forall\ u^h \in \mathcal{U}^h, \quad v^h \in \mathcal{V}^h, \tag{2.8}$$

  with $\langle\cdot,\cdot\rangle = \langle\cdot,\cdot\rangle_{H^{-1}(\Omega_c^h)\times H^1(\Omega_c^h)}$; see Proposition 8.2.
- For (extended) $(f, \phi) \in H^{-1}(\Omega_c^h) \times H^{1/2}(\partial\Omega_c^h)$, the linearized *weak* problem

$$(F^h)'(u_0^h)u_1^h = (f, \phi) \Leftrightarrow \langle G'(u_0^h)u_1^h - f, v^h\rangle + ((u_1^h - \phi)|_{\partial\Omega_c^h}, v_b^h)_{L^2(\partial\Omega_c^h)} = 0$$

  for all $v^h \in \mathcal{V}^h, v_b^h \in \mathcal{V}_b^h$ is stable, essentially if u_0 is an isolated solution of $F(u_0) = 0$ with boundedly invertible $F'(u_0)$. This is a consequence of the discussion at the end of section 3 and the compactness arguments in Theorem 8.4.
- By Lemma 8.5 the regularity results for the exact solution of a linear elliptic problem are valid for the corresponding FE solutions. In the *weak linearized*

problem, $(F^h)'(u_0^h)u_1^h = (f,\phi) \in H^{-1}(\Omega_c^h) \times H^{1/2}(\partial\Omega_c^h)$, we determine $u_1^h \in \mathcal{U}^h$ for the smooth $(f,\phi) \in L^2(\Omega_c^h) \times H^{3/2}(\partial\Omega_c^h)$; see (2.6), (2.8). The stability of the *weak* problem implies the stability of the *strong* problem

$$(G'(u_0^h)u_1^h - f, v^h)_{L^2(\Omega_c^h)} + ((u_1^h - \phi)|_{\partial\Omega_c^h}, v_b^h)_{L^2(\partial\Omega_c^h)} = 0 \quad \forall v^h \in \mathcal{V}^h, \quad v_b^h \in \mathcal{V}_b^h.$$

This allows estimates for discrete solutions w.r.t. a piecewise H^2-norm instead of the usual H^1-norm. This requires smooth coefficients, necessary for good convergence, and a convex, bounded (polyhedral) domain or $\Omega \in C^2$.

3. *Summary: Essentially simultaneously with a unique* u_0*, the FE solutions* u_0^h *of* (2.4)*,* (2.5)*, and* (2.6) *uniquely exist, for small enough* h*, and converge; see* (2.9) *and Theorems* 5.1 *and* 8.7*.*

4. *Setting up the FE equations* (2.5) *and solving them; see section* 9*:*

- Setting up the nonlinear system (2.5) is formulated in (9.1).
- As a consequence of **1.**, **2.**, the MIP is valid. So the Newton method for the discrete equation (2.5), with a good enough initial guess $u_1^h = P^h u_1$, e.g., obtained by a continuation method, converges quadratically to u_0^h equally well as the Newton–Kantorovich method, started in u_1, for the original problem (2.1); see Theorem 9.1 and (2.10).
- The linearization of G is elliptic. Thus the efficient solvers for linear elliptic problems in weak form (see (2.7)) can be applied in the continuation and Newton processes, replacing C^0 FEs by C^1 FEs. Results of Davydov [28] and Davydov and Stevenson [30] for Ω^h allow extensions to nested sequences of triangulations $\mathcal{T}_1^h \sqsubset \mathcal{T}_2^h \sqsubset \cdots$, hence even multiresolution techniques.

We formulate the convergence of the FE and the discrete Newton method: Let (2.1), for smooth enough G and Ω, have a smooth isolated solution $u_0 \in H^\ell(\Omega)$ with bounded $(F'(u_0))^{-1}$ and choose FEs of local degree d in C^1 with $0 \le \ell - 1 \le d$; cf. (4.9). Then the discrete solutions u_0^h of (2.5) converge as (see Theorem 8.7)

$$\|E_c u_0 - u_0^h\|_{H^2(\Omega_c^h)}^h \le C h^{\min\{\ell-2,p\}} \|u_0\|_{H^\ell(\Omega)} \quad \text{for large enough } \ell,\ d. \tag{2.9}$$

For convex, bounded, polyhedral domains all results are known in $\mathbf{R}^n$ (cf. [27]), and we get $p = \ell - 2$. For $\Omega \in C^2$ we obtain $p = 2$ for polyhedral Ω^h and $p > 2$ for curved Ω_c^h. Davydov has fully proved the case $\mathbf{R}^2, d = 5, p = 2$ and is working on higher $\mathbf{R}^n, p > 2$. The proofs in this paper are based upon these generalizations.

The Newton method for the discrete problem (2.6) has the form (2.10) and converges locally quadratically; see Theorem 9.1:

$$\begin{aligned} &u_{\nu+1}^h \in \mathcal{U}^h : \ (F^h)'(u_\nu^h)(u_{\nu+1}^h - u_\nu^h) = -F^h(u_\nu^h) \\ &\implies \|u_{\nu+1}^h - u_\nu^h\|_{H^2(\Omega_c^h)}^h \le C(\|u_\nu^h - u_{\nu-1}^h\|_{H^2(\Omega_c^h)}^h)^2. \end{aligned} \tag{2.10}$$

Finally, we indicate, but do not explicitly formulate, two further extensions. In section 8 the linearized weak problem turns out to be a compact perturbation of a coercive bilinear form. Hence, the usual Fredholm alternative results apply. The above FEM can easily be extended to parameter dependent problems and for studying bifurcation scenarios. Some general results prove the convergence of the bifurcation characteristics of the discrete FEM to those of the original problem, starting with Brezzi, Rappaz, and Raviart in [14, 15, 16], and, e.g., Böhmer [7, 9, 8] and Cliff, Spence, and Tavener [21]. Extensions of these results to convergence of the characteristics of center manifolds for the discrete FEM to those for the original problem are, for the first time, studied in Böhmer [9, 11].

3. Fully nonlinear and general quasi-linear elliptic equations. We introduce the standard notations for derivatives

$$(3.1)\ \partial^i u = \frac{\partial u}{\partial x_i},\ Du = (\partial^i u)_{i=1}^n,\ D^2u = (\partial^i\partial^j u)_{i,j=1}^n,\ \partial^0 u := u, (-1)_{j>0} := 1,\ j = 0,$$

and $:= -1$ for $j \geq 1$. We formulate the conditions for the domain and introduce four Hilbert spaces with $\mathcal{W}_0$ for trivial Dirichlet boundary conditions:

$$(3.2)\quad \Omega \text{ is an (open) bounded domain in } \mathbf{R}^n \text{ with Lipschitz continuous } \partial\Omega \text{ and}$$
$$\mathcal{U} := H^2(\Omega) \subset \mathcal{W} := H^1(\Omega) \subset \mathcal{V} := L^2(\Omega) = \mathcal{V}', \mathcal{V}_D := H^{3/2}(\partial\Omega), \mathcal{W}_D := H^{1/2}(\partial\Omega),$$
$$\mathcal{W}_0 := H_0^1(\Omega) := \{v \in H^1(\Omega) : v = 0 \text{ on } \partial\Omega\},\ \mathcal{U}_0 := \mathcal{W}_0 \cap \mathcal{U}, \mathcal{W}' := H^{-1}(\Omega),$$

with the standard Sobolev norms, seminorms, inner products for $\mathcal{U}, \mathcal{W}, \mathcal{V}$, and Sobolev–Sloboditskii norms for $\mathcal{V}_D, \mathcal{W}_D$; cf. Hackbusch [37] and Oden and Reddy [43].

For a quasi-linear elliptic equation in divergence form a weak conforming FEM determines the FE solution $u_0^h \in \mathcal{V}^h = \mathcal{U}_0^h$ as

$$(3.3)\ \langle Gu_0^h, v^h\rangle_{\mathcal{W}'\times\mathcal{W}} := \int_\Omega \sum_{j=0}^n a_j(x, u_0^h, Du_0^h)\partial^j v^h dx = \langle f, v^h\rangle_{\mathcal{W}'\times\mathcal{W}} \quad \forall v^h \in \mathcal{V}^h.$$

For a general quasi-linear or a fully nonlinear G a weak form is impossible:

$$(3.4)\qquad u_0 \in \mathcal{U}_0 :\ Gu_0 - f = \sum_{i,j=0}^n a_{i,j}(\cdot, u_0, Du_0)\partial^i\partial^j u_0 - f(\cdot) = 0 \in \mathcal{V},\ \text{or}$$

$$(3.5)\qquad u_0 \in \mathcal{U}_0 :\ G(u_0) = G(\cdot, u_0(\cdot), Du_0(\cdot), D^2u_0(\cdot)) = 0 \in \mathcal{V}.$$

Consequently, a weak form of a FEM is impossible, thus motivating our FEM.

The general theory of discretization methods used in this paper can be applied if the original problem has two properties: It admits a locally unique solution u_0, strongly related to the stability, and u_0 is smooth enough, implying the consistency and good enough convergence of our method. So we list some results for fully nonlinear elliptic equations, based upon rather technical conditions, summarizing, e.g., Gilbarg and Trudinger [35], Chen and Wu [18], and Skrypnik [50]. We assume (3.2) for $\Omega \subset \mathbf{R}^n$, a slightly larger open $\Omega' \subset \mathbf{R}^n$ with $\overline{\Omega} \subset \Omega'$, $\overline{\Omega_c^h} \subset \Omega'$, where Ω_c^h indicates later approximations for Ω for the $u^h \in \mathcal{U}^h$. A real valued function G^w is defined s.t.

$$(3.6)\ G^w : w = (x, z, p, r) \in \mathcal{D}(G^w) = \Omega' \times \Gamma \to \mathbf{R}, \Gamma \subset \mathbf{R} \times \mathbf{R}^n \times \mathbf{R}^{n^2}, \Omega', \Gamma \text{ open}.$$

We consider the strong elliptic BVP (cf. (2.1), (3.5), (3.6)), often with $\phi = 0$ and

$$(3.7)\quad w(x) := w_u(x) := (x, u(x), Du(x), D^2u(x)), \quad u \to G(u) := G^w(w_u) \text{ for}$$
$$u \in \mathcal{D}(G) := \{u \in \mathcal{U} : w_u(x) \in \mathcal{D}(G^w) \quad \forall x \in \Omega \text{ and } G^w(u) \in \mathcal{V}\},$$
$$(3.8)\qquad \text{the solution satisfies } u_0 \in \mathcal{D}(G) :\ G(u_0) = 0 \text{ on } \Omega, \quad u_0 = \phi \text{ on } \partial\Omega.$$

Ellipticity is defined by the linear $G'(u_0)$ in $w_0 := w_{u_0}$ applied to u with $u, u_0 \in \mathcal{D}(G)$:

$$(3.9)\qquad G'(u_0)u = \frac{\partial G^w}{\partial z}(w_0)u + \sum_{i=1}^n \frac{\partial G^w}{\partial p_i}(w_0)\partial^i u + \sum_{i,j=1}^n \frac{\partial G^w}{\partial r_{ij}}(w_0)\partial^i\partial^j u.$$

This will be important below for the local uniqueness and stability. The operator G is called *uniformly elliptic* in $u_0 \in \mathcal{D}(G)$ and all of $\mathcal{D}(G)$ if its linearization $G'(u)$ is uniformly elliptic for $u = u_0$ and for all $u \in \mathcal{D}(G)$, respectively. In the first case its principal part, $a_p(u, v)$ (see (3.11)), satisfies, for $0 < \lambda < \Lambda$, and $w_0 = w_{u_0}$ in (3.7),

$$0 < \lambda|\vartheta|^2 \le -\sum_{i,j=1}^{n} \frac{\partial G^w}{\partial r_{ij}}(w_0)\vartheta^i\vartheta^j \le \Lambda|\vartheta|^2 \quad \forall 0 \ne \vartheta = (\vartheta_1, \dots, \vartheta_n) \in \mathbf{R}^n, \tag{3.10}$$

sometimes only for a.e. $x \in \Omega$ in $w_0(x)$. For fully nonlinear second order elliptic equations this implies the $\mathcal{W}$-coercivity of the bilinear form induced by the principal part, $a_p(u, v)$ of $G'(u_0)$. By choosing $\vartheta^i = \partial^i u(x)$, integrating over Ω, and using partial integration we obtain, with the equivalence in $\mathcal{W}_0$ of the energy norm $(\sum_{i,j=1}^n \|\partial^i u\|_{\mathcal{V}}^2)^{1/2}$ and $\|\cdot\|_{H^1(\Omega)}$ and again with $0 < \lambda_1 < \Lambda_1 \in \mathbf{R}$,

(3.11)

$$\lambda_1\|u\|_{H^1(\Omega)}^2 \le a_p(u, u) := \int_\Omega \sum_{i,j=1}^{n} \frac{\partial G^w}{\partial r_{ij}}(w)\partial^i u\partial^j u dx \ \le \Lambda_1\|u\|_{H^1(\Omega)}^2 \quad \forall u \in \mathcal{W}_0.$$

Therefore we chose the Hilbert space setting, $\mathcal{W} = H^1(\Omega)$. A Banach space setting, $W^{1,p}(\Omega)$, might be appropriate if in addition to (3.10) some monotonicity conditions would be considered. Furthermore, since $W^{2,p}(\Omega)$ is continuously embedded into $H^2(\Omega)$, for $p \ge 2$, our new FEM even works for fully nonlinear equations in $W^{2,p}(\Omega)$, e.g., $G(u) = (\Delta u)^p + f(u)$; however, it yields only discrete $H^2(\Omega)$ estimates.

We formulate some of the most important examples.

EXAMPLE 3.1 (Monge–Ampère equation). *This is, in particular for* $n = 2$,

$$0 = G(u) := \det D^2u - f(x, u, Du) = u_{xx}u_{yy} - u_{xy}^2 - f(x, u, Du) \quad \textit{for } n = 2. \tag{3.12}$$

This fully nonlinear $G(u)$ is uniformly elliptic in u_0 for a symmetric strictly positive matrix $r = D^2u_0$ in Ω, hence for a function $u_0 \in C^2(\Omega)$, strictly convex for all $x \in \Omega$ on a convex Ω. With $\det(D^2u) > 0$, uniformly convex solutions are possible only for positive f [35, pp. 441ff.]. For partially nonelliptic generalizations, see [35, pp. 467ff.] and Courant and Hilbert [22, p. 324]. According to Haltiner, Kasahara, and Fulton the leading term in the balance equation in dynamical meteorology has the form (3.12) [38, 39, 34]. In Westcott [60] it is related to geometric optics.

Among the many applications in geometry we mention the following equation.

Equation for a surface with prescribed Gauss curvature $K = K(x)$ at x. It is closely related to the Monge–Ampère equation. Equation (3.12) has to be modified as

$$G(u) := \det D^2u - K(1 + |Du|^2)^{(n+2)/2} = 0, \quad K(x) > 0. \tag{3.13}$$

The existence of surfaces with prescribed Gauss curvature is known for appropriate given curvature, K. Specific properties would become more flexibly available via our FEMs than by finite difference methods, e.g., by Crandall and Lions [24]: Our FEM allows C^1 FEs on quasi-uniform grids and yields, for convex polyhedral Ω or the forthcoming better boundary approximations, higher orders of convergence than the difference methods in [24], which are defined only on equidistant grids. □

To guarantee existence and uniqueness results we have to impose *growth conditions*. We distinguish the cases $n = 2$ and $n \ge 2$, evaluate the following derivatives in

$w=(x,z,p,r)\in\mathcal{D}(G^w), x\in\partial\Omega$, and require, for $G^w_x, G^w_z, G^w_p, G^w_r$ (see (3.9), (3.10)),

$$\begin{aligned}&\text{for } n=2: |G^w_z(w)|, |G^w_p(w)|\le\lambda\mu,\\ &\qquad |G^w_x(w)|\le\mu\lambda(1+|p|+|r|) \text{ and}\end{aligned}\tag{3.14}$$

$$\begin{aligned}&\text{for } n\ge 2: |G^w_z(w)|, |G^w_p(w)|, |G^w_{rx}(w)|, |G^w_{px}(w)|, |G^w_{zx}(w)|\le\lambda\mu,\\ &\qquad |G^w_x(w)|, |G^w_{xx}(w)|\le\mu\lambda(1+|p|+|r|).\end{aligned}\tag{3.15}$$

The nonnegative $\lambda=\lambda(\cdot)$ and $\Lambda=\Lambda(\cdot), \mu=\mu(\cdot)$ in (3.10), (3.14)–(3.15) depend upon $|z|$ and have to be nonincreasing and nondecreasing in $z\in[0,\infty)$, respectively.

THEOREM 3.1 (see [35, Theorem 17.12]). *Assume* $n=2,\ \partial\Omega\in C^3,\ \phi\in C^3(\overline{\Omega})$; *see* (3.8). *Let* G *satisfy the conditions* (3.10), (3.14), *and* $G^w_z\le 0$ *in* Γ; *see* (3.6). *Then the classical Dirichlet problem* (3.8) *has a unique solution* $u_0\in C^{2,\alpha}(\overline{\Omega})$ *for all* $0<\alpha<1$.

THEOREM 3.2 (see [35, Theorem 17.17]). *Assume* $n\ge 2, \phi\in C(\partial\Omega)$. *Let* Ω *satisfy an exterior sphere condition for every* $\xi\in\partial\Omega$ *and* G *satisfy the conditions* (3.10), (3.15), *and* $G^w\in C^2(\Gamma)$ *is concave (or convex) w.r.t.* (z,p,r), *nonincreasing w.r.t.* z. *Then the classical Dirichlet problem* (3.8) *has a unique solution* $u_0\in C^2(\Omega)\cap C(\overline{\Omega})$.

Remark 3.1. Ω satisfies an exterior sphere condition, for every $\xi\in\partial\Omega$, if there exists a ball $B=B_R(y)\subset\mathbf{R}^n$ s.t. $\overline{B}\cap\overline{\Omega}=\{\xi\}$. Mind the difference in these theorems. For $n=2$ a $\partial\Omega\in C^3$ implies the global $u_0\in C^{2,\alpha}(\overline{\Omega})$ for all $0<\alpha<1$, in contrast to the interior estimate $u_0\in C^2(\Omega)\cap C(\overline{\Omega})$ for $n\ge 2$ and no condition for $\partial\Omega$.

More detailed results for the above special cases in Example 3.1 are presented in [35] and Skrypnik [50]. Reference [50] requires for the Monge–Ampère equation (3.12) on $\Omega=B_R(0)\subset\mathbf{R}^2$ with center 0 functions ϕ,ψ,f_0 s.t.

$$(u_0-\phi)|_{\partial B_R(0)}=0,\ \psi: B_R(0)\subset\mathbf{R}^2\to\mathbf{R}_+,\ f_0:\mathbf{R}^2\to\mathbf{R}_+,\ m:=\max_{x\in B_R(0)}\phi(x),$$

$$f(x,u,p)\le\psi(\|x\|^2)f_0(\|p\|^2)\quad \text{for } x,p,s\in\mathbf{R}^2,\quad \mathbf{R}\ni u\le m \text{ and let}\tag{3.16}$$

$$\int_{B_R(0)}\psi(\|x\|^2)dx\le\int_{\mathbf{R}^2}\inf_{\|s-p\|^2<M_H}1/f_0(\|s\|^2)ds,$$

where M_H is the lower winding of $(u_0-\phi)|_{\partial B_R(0)}=0$. Finally, let $\partial^1\partial^1 u(x)>0$:

$$H^4_+(\Omega):=\{u\in H^4(\Omega): \det D^2u(x)>0, \partial^1\partial^1u(x)>0\quad\forall x\in\overline{B_R(0)}\}.$$

THEOREM 3.3 (regular solutions [50, Theorem 3.2.3]). *Let* (3.16) *be satisfied and* $f\in C^{2,\gamma}(\Omega), \phi\in C^{4,\gamma}(\partial\Omega)$. *Then* (3.12) *has at least one solution in* $H^4_+(\Omega)$.

Regularity results for $u\in C^{k,\gamma}(\Omega), k\ge 2$, can be deduced from Chen and Wu [18, Chapter 7] and are extended in Böhmer [10].

Existence and uniqueness results are obtained by methods, usually independent of linearization. However, we assume additionally a locally unique solution u_0 in (2.2), with boundedly invertible Fréchet derivative, similarly for G on convex polyhedral Ω:

$$F:\mathcal{U}\to\mathcal{V}\times\mathcal{V}_D,\quad \text{locally } \exists_1 u_0: F(u_0)=(0,\phi),\quad (F'(u_0))^{-1}\in\mathcal{L}(\mathcal{V}\times\mathcal{V}_D,\mathcal{U}).\tag{3.17}$$

We defend this condition by the close relation to the numerically necessary condition of a (locally) well-conditioned problem $F(u)=0$. For an existing locally unique solution the Fréchet differentiability is satisfied for many practically relevant problems. By the Taylor formula, a boundedly invertible $F'(u_0)\in\mathcal{L}(\mathcal{U},\mathcal{V}\times\mathcal{V}_D)$ is very

probable but not necessarily satisfied. But there will be exceptions. For our numerical approach we detect this *exceptional case,* e.g., the discrete Newton method below. The *convergence* for our numerical method *is not proved.* Unless $F'(u_0) \in \mathcal{L}(\mathcal{U}, \mathcal{V} \times \mathcal{V}_D)$ is (nearly) singular, $(F^h)'(u_0^h)$ is stable; see Theorem 8.6. If (3.17) would be violated, small perturbations of an original parameter dependent problem would usually have a bifurcation point and numerical methods for bifurcation would have to be applied.

4. Smooth FEs and extension theorems. We have emphasized above that the proof of the "strong stability" will require

(4.1) a bounded either convex polyhedral Ω with conforming FEs or

(4.2) $\Omega \in C^2$, with a polyhedral or curved $\Omega_c^h \approx \Omega$ with nonconforming FEs.

We give a short summary of the Davydov results in [27]: C^1 FEs, defined on polyhedral domains Ω, are conforming, satisfying the boundary conditions. All necessary approximation theoretic results are known for $n \geq 2, d \geq 2^n + 1$. For $\Omega \in C^2$, C^1 FEs on polyhedral or curved approximating domains $\Omega_c^h \approx \Omega$ usually are nonconforming by violating the boundary conditions. The results are proved for modified Argyris FEs in $\mathbf{R}^2$ on Ω_c^h. We indicate the forthcoming generalization of [27] to curved approximations of Ω, allowing higher accuracy, and to $\mathbf{R}^n$. These FEs are split into the sum of FEs vanishing and living on the boundary. For comparing an original function and its interpolating FE we need extension operators from Ω and Ω_c^h to $\Omega' \supset \Omega_c^h \cup \Omega$.

4.1. Smooth FEs on polyhedral and curved domains. Fully nonlinear weak forms are impossible. So the FEs u^h fit to our discrete equations (2.5), (2.6) if $G(u^h) \in L^2(\Omega_c^h)$, hence for $u^h \in C^1(\Omega_c^h)$. In fact, the standard FEs $u^h \in C(\Omega_c^h)$, with $G(u^h) \in L^2(\mathcal{T}_c^h)$, would imply interior variational crimes by the nondifferentiability of the $u^h \notin H^2(\Omega_c^h)$ between neighboring simplices $T \in \mathcal{T}_c^h$. The interplay of strong and weak forms, impossible for fully nonlinear operators, seemed to leave no other choice than FEs in C^1.

We combine the Davydov results [27] with a new idea of Tiihonen [57]. He decomposes, for linear sample problems, the error for FEMs by introducing an auxiliary problem defined in a polyhedral $\Omega^h \approx \Omega$ for a *smooth* domain. Then a standard FEM on Ω^h with known error estimates is combined with the "geometric part of the error", caused by $\Omega^h \approx \Omega$. We generalize his approach to our equations; see Theorem 4.4.

Let $\mathcal{T}^h$ be a simplicial subdivision or triangulation of a polyhedral Ω, or a polyhedral approximation $\overline{\Omega^h} := \cup_{T \in \mathcal{T}^h} \overline{T} \approx \overline{\Omega} \in C^2$. In the final $\mathcal{T}_c^h$ we maintain all "interior" facets of $\mathcal{T}^h$ with at most one boundary point. As boundary facets (see Figure 4.1), we choose either plane ∂T or curved facets ∂T_c, approximating $\partial\Omega$, e.g., between P_i and P_e. Finally, let $\overline{\Omega_c^h} := \cup_{T \in \mathcal{T}_c^h} \overline{T} \approx \overline{\Omega}$ with $h := \max_{T \in \mathcal{T}_c^h} \{\text{diam } T\}$. $\mathcal{T}_c^h$ satisfies the standard conditions; see [27]: It is a finite set of nondegenerate open pairwise disjoint n-simplices s.t. $\Omega \approx \Omega_c^h$. Each facet of a simplex $T \in \mathcal{T}_c^h$ either is a common face of, in $\mathbf{R}^2$ exactly two, for $\mathbf{R}^n, n > 2$, only few simplices in $\mathcal{T}_c^h$ or it approximates, as plane or curved surface, a part of $\partial\Omega$. We use the unifying notations Ω_c^h and $\mathcal{T}_c^h$ and $\mathcal{S}_d^1(\mathcal{T}_c^h)$, $\mathbf{R}^n$ for all cases.

For any $d \geq 0$, we consider FEs, $S_d(\mathcal{T}_c^h)$, and $S_d^1(\mathcal{T}_c^h)$, the spaces of all[1] piecewise polynomial functions of total degree $\leq d$ w.r.t. $\mathcal{T}_c^h$ of n variables, s.t.

$$(4.3)\quad s \in S_d(\mathcal{T}_c^h) \iff s|_T \in \Pi_d^n \quad \forall T \in \mathcal{T}_c^h \text{ and } S_d^1(\mathcal{T}_c^h) = S_d(\mathcal{T}_c^h) \cap C^1(\Omega^h).$$

[1] We do not want to overformalize the notation and have chosen $h \in H \subset \mathbf{R}, \inf_{h \in H} h = 0$.

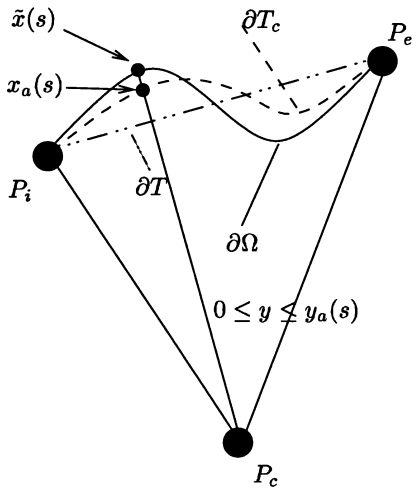


FIG. 4.1. *Approximation of the* $\partial\Omega$.

We need $d \ge 2^n + 1, n \ge 2$, and so $d \ge 5$ in $\mathbf{R}^2$ and $d \ge 9$ in $\mathbf{R}^3$ for $S^1_d(\mathcal{T}^h_c)$.

Davydov uses *quasi-uniform* and *nondegenerate* FEs, the *shape regularity constant* $\omega_{\mathcal{T}^h_c}$ with $1/\omega_{\mathcal{T}^h_c} := \min_{T\in\mathcal{T}^h} \operatorname{diam} B_T/\operatorname{diam} T \ge \chi > 0$, where B_T is the largest ball in $\overline{T}$. The *star,* $\operatorname{star}(v)$, of a vertex, v, of $\mathcal{T}^h_c$ is the union of all n-simplices $\overline{T} \in \mathcal{T}^h_c$ attached to v. We set $\operatorname{star}^1(v) := \operatorname{star}(v)$ and define $\operatorname{star}^\gamma(v)$, $\gamma \ge 2$, recursively as the union of the stars of all vertices of $\mathcal{T}^h_c$ contained in $\operatorname{star}^{\gamma-1}(v)$.

The $\mathcal{S}^1_d(\mathcal{T}^h_c) \not\subset H^k(\Omega^h_c), 1 < k-1$, are subsets of the broken Sobolev spaces, $H^k(\mathcal{T}^h_c)$. We choose three *sequences of linear FE subspaces*

$$\mathcal{U}^h = \mathcal{W}^h = S^h \subset \mathcal{S}^1_d(\mathcal{T}^h_c) \supset \mathcal{V}^h = S^h_0 \subset \mathcal{S}^1_d(\mathcal{T}^h_c) \cap H^1_0(\Omega^h_c), \text{ with } d \ge 2^n + 1$$

(4.4) and $\|u^h\|_{\mathcal{U}^h} := \|u^h\|_{H^2(\mathcal{T}^h_c)}, \quad \|w^h\|_{\mathcal{W}^h} := \|w^h\|_{H^1(\mathcal{T}^h_c)}, \quad \|v^h\|_{\mathcal{V}^h} := \|v^h\|_{L^2(\mathcal{T}^h_c)}.$

The $\|u^h\|_{\mathcal{U}^h}$, $\|w^h\|_{\mathcal{W}^h}$ are nonequivalent norms for $h \to 0$, essential for convergence.

4.2. Inverse and error estimates for polyhedral Ω in $\mathbf{R}^n$. For polyhedral Ω and the FEMs below we choose the subspaces $\mathcal{U}^h = \mathcal{V}^h = S^h_0$ of $\mathcal{S}^1_d(\mathcal{T}^h_c)$, exactly vanishing on $\partial\Omega^h_c$. For this case and its conforming FEs we need only the "inverse" and "error estimates"; cf. Theorems 4.1 and 4.2 with (4.5) and (4.11). They are proved in [27] for polyhedral $\Omega \subset \mathbf{R}^n, n \ge 2, d \ge 2^n + 1$, or polyhedral approximations $\Omega^h_c \approx \Omega$.

THEOREM 4.1. *For* $0 \le k < \mu \le d$, *we have*

$$|s|_{H^\mu(T)} \le \frac{K}{(\operatorname{diam} T)^{\mu-k}} |s|_{H^k(T)} \quad \forall s \in S_d(\mathcal{T}^h_c), \quad T \in \mathcal{T}^h_c. \tag{4.5}$$

The constant K *depends only on* $n, d, \omega_{\mathcal{T}^h_c}$. *For a quasi-uniform triangulation* $\mathcal{T}^h_c$ *the* $|s|_{H^j(T)}, j = \mu, k$, *can be replaced by* $|s|_{H^j(\mathcal{T}^h_c)}$.

We introduce a *stable and local basis* for S^h and its dual basis for $S^{h'}$:

$$\mathcal{S}_d(\mathcal{T}^h_c) \supset \mathcal{S}^1_d(\mathcal{T}^h_c) \supset S^h := \operatorname{span}[s^1, \dots, s^N] \text{ and } S^{h'} := \operatorname{span}[\lambda^1, \dots, \lambda^N]. \tag{4.6}$$

This means, for each $s^k, k = 1, \dots, N$, there exist a set $E_k \subset \Omega^h_c$ and a vertex v_k, s.t.

$$\operatorname{supp} s^k \subset E_k, \text{ with a vertex } v_k, \text{ and } E_k \subset \operatorname{star}^\gamma(v_k), \text{ and with} \tag{4.7}$$

$$\|s^k\|_{L_\infty(\Omega^h_c)} \le C_1, \text{ and } |\lambda^k s| \le C_2 \|s\|_{L_\infty(E_k)} \ \ \forall s \in S^h, \ \lambda^k \text{ in (4.6)}, \tag{4.8}$$

for some C_1, C_2 and γ; see (3.1)–(3.4) and (3.13) in [27]. Hence, after renorming the s^μ, the norms

$$\|(c_\mu)_{\mu=1}^N\|_p \text{ and } \Big\|\sum_{\mu=1}^N c_\mu s^\mu\Big\|_p \quad \forall 1 \le p \le \infty \text{ are equivalent.}$$

This property is important for stable interpolations. It is proved for $S_d^1(\mathcal{T}_c^h)$ and subspaces $S^h \subset S_d^1(\mathcal{T}_c^h)$ if $d \ge 2^n + 1$ in Davydov [28, 27] and the references therein; for FE subspaces, see, e.g., [13]. Generalizing (4.3), let the S^h satisfy the global property

$$\Pi_{\ell-1}^n \subset S^h \subset S_d(\mathcal{T}_c^h) \quad \text{for some } 0 \le \ell - 1 \le d. \tag{4.9}$$

THEOREM 4.2. *Under the conditions* (4.6)–(4.9)*, there exists a linear operator* $I^h : L_1(\Omega_c^h) \to S^h$*, s.t. for any* $T \in \mathcal{T}_c^h$*,* $0 \le |\alpha| \le \ell - 1$*,* $u \in L_1(\Omega_c^h)$ *with* $|u|_{H^\ell(\Omega_T^\gamma)} < \infty$*,*

$$\|D^\alpha(u - I^h u)\|_{L^2(T)} \le K h_T^{\ell-|\alpha|} |u|_{H^\ell(\Omega_T^\gamma)}, \text{ with } \Omega_T^\gamma := \cup_{v\in T} \mathrm{star}^{2\gamma-1}(v), \tag{4.10}$$

with the Lipschitz constant $L_{\partial\Omega_c^h}$ *of* $\partial\Omega_c^h$ *and* $K = K(n, d, \omega_{\mathcal{T}_c^h}, \gamma, C_1, C_2, L_{\partial\Omega_c^h})$*. If* $u \in H^\ell(\Omega_c^h)$*, then for all* $0 \le |\alpha|, k \le \ell - 1$*,*

$$\|D^\alpha(u - I^h u)\|_{L^2(\Omega_c^h)} \le K h^{\ell-|\alpha|} |u|_{H^\ell(\Omega_c^h)} \Rightarrow \lim_{h\to 0} \|I^h u\|_{H^k(\mathcal{T}_c^h)} = \|u\|_{H^k(\Omega_c^h)}, \tag{4.11}$$

$$\|u - I^h u\|_{H^k(\mathcal{T}_c^h)} \le K h^{\ell-k} |u|_{H^\ell(\Omega_c^h)} \quad \textit{and} \tag{4.12}$$

$$\|u - I^h u\|_{H^k(\partial\Omega_c^h)} \le K h^{\ell-k-1/2} |u|_{H^\ell(\Omega_c^h)} \Longrightarrow \lim_{h\to 0} \|I^h u\|_{H^k(\partial\Omega_c^h)} = \|u\|_{H^k(\partial\Omega_c^h)},$$

with the broken boundary norm, $\|u^h\|_{H^k(\partial\Omega_c^h)} = \|u^h\|_{H^k(\partial\mathcal{T}_c^h)}$*; cf. the trace theorem.*

4.3. Stable splitting of $S^h = S_0^h \oplus S_b^h$ for $\Omega \in C^2$. The S^h and $S^{h'}$, constructed for an $\Omega_c^h \approx \Omega$, and their above local stable basis for S^h have to be determined in a special way. They have to allow a stable splitting for the FEs, S_0^h, vanishing on $\partial\Omega_c^h$, and the complementary set, S_b^h, with two local stable bases and their duals satisfying

$$\begin{aligned} &S_d^1(\mathcal{T}_c^h) \supset \mathrm{span}\{s^1, \dots, s^N\} = S^h = S_0^h \cup S_b^h, S_0^h \cap S_b^h = \{0\}, \text{ with} \\ &S_0^h = \mathrm{span}\{s^1, \dots, s^{N_0}\}, \quad S_b^h = \mathrm{span}\{s^{1+N_0}, \dots, s^N\}, \quad N_0 + N_b = N. \end{aligned} \tag{4.13}$$

Analogously we obtain, with $\lambda^i s^j = \delta_{i,j}$, for the dual bases

$$\begin{aligned} &\mathrm{span}\{\lambda^1, \dots, \lambda^{N_0}, \lambda^{1+N_0}, \dots, \lambda^N\} = S^{h'} = S_0^{h'} \cup S_b^{h'}, \quad S_0^{h'} \cap S_b^{h'} = \{0\}, \text{ s.t. for} \\ &i = 1, \dots, N: \quad s^j \in S_0^h, \quad j = 1, \dots, N_0: \ \lambda^i s^j = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{otherwise} \end{cases} \text{ and} \\ &s_b^j = s^{j+N_0} \in S_b^h, \quad \lambda_b^j = \lambda^{j+N_0}, \quad j = 1, \dots, N_b: \quad \lambda^i s_b^j = \begin{cases} 1 & \text{if } j + N_0 = i, \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \tag{4.14}$$

We apply, for simplicity, the following interpolation operators, I^h, only to functions

$$\lambda^i, \quad i = 1, \dots, N, \text{ are defined, e.g., for } f \in C^{k_0}(\Omega_c^h). \tag{4.15}$$

Otherwise averaged Taylor polynomials or extension theorems could be used; see [27].

The above I^h (cf. (4.11)) and the new I_b^h have the form

$$(4.16)\, I^h, I_b^h : C^2(\Omega_c^h) \to S^h,\ I^h(f) = \sum_{i=1}^{N} \lambda^i(f) s^i \in S^h,\ I_b^h(f) = \sum_{i=1+N_0}^{N} \lambda^i(f) s^i \in S_b^h.$$

A stable splitting with the properties in Theorem 4.4 is possible only if S^h satisfies certain unexpected conditions. For the usual Argyris FEs in $\mathcal{S}_5^1(\mathcal{T}_c^h)$, the standard continuous second derivatives in all vertices are required for S^h only for interior vertices. So the chosen degree, smoothness properties, and interpolation conditions at $\partial\Omega_c^h$ are related; cf. the subtle discussion for the modified Argyris FEs in [27].

Presently, the following results are fully proved in [27] only for modified Argyris FEs in $\mathbf{R}^2$ on a polygonal $\Omega_c^h \approx \Omega$ with $\operatorname{dist}(\partial\Omega_c^h, \partial\Omega) \le Ch^2$. This restricts, for $\Omega \in C^2$, the convergence of our FEM to the order $p = 2$. Davydov is generalizing these results to FEs $\mathcal{S}_d^1(\mathcal{T}_c^h)$ on piecewise conical $\partial\Omega_c^h \approx \partial\Omega$. This allows FEMs of order $p > 2$ for $\Omega \in C^p$, probably even in $\mathbf{R}^n, n \ge 2$.

THEOREM 4.3. *Let $\partial\Omega \in C^p, p \ge 2$. Then the polyhedral Ω^h with $p = 2$ and piecewise curved approximation Ω_c^h for Ω with $p > 2$ yield*

$$(4.17)\ \operatorname{dist}\{\partial\Omega, \partial\Omega_c^h\} \le Ch^2 \|\partial\Omega\|_{C^p} \text{ and } \operatorname{dist}\{\partial\Omega, \partial\Omega_c^h\} \le Ch^p \|\partial\Omega\|_{C^p},\ \textit{respectively.}$$

THEOREM 4.4. *We assume $n \ge 2, d \ge 2^n + 1$, $\partial\Omega \in C^p, 2 \le p \le 5$ and choose the piecewise polyhedral, $p = 2$, or conical $\partial\Omega_c^h \approx \partial\Omega, 2 < p \le 5$ from Theorem 4.3. Davydov [27] proves and remarks for $n = 2, d = 5, p = 2$ and $n \ge 2, d \ge 2^n - 1$:*

1. *There exists a subspace S^h of $\mathcal{S}_d^1(\mathcal{T}_c^h)$ in $\mathbf{R}^n$ with a stable splitting of a local stable basis in (4.6)–(4.9) into two complementary stable subbases in (4.13), (4.14).*

2. *The above* interpolation operators $I : C^{k_0}(\Omega_c^h) \to S$ *and* $I_b : C^{k_0}(\Omega_c^h) \to S_b$ *(see (4.15) and (4.16)) have the following properties.*

   a. *If $f|_{\partial\Omega_c^h} = 0$, then $I(f) \in S_0$.*

   b. *For $f \in C^{k_0}(\Omega_c^h)$ and $\ell \le d + 1$, let $H^\ell(\Omega_c^h) \subset C^{k_0}(\Omega_c^h)$. Then*

$$(4.18)\ \|D^\alpha(f - I(f))\|_{L_2(\Omega_c^h)} \le Kh^{\ell-|\alpha|} |f|_{H^\ell(\Omega_c^h)} 0 \quad \forall 0 \le |\alpha| \le \ell - 1;\ \textit{moreover,}$$
$$I_b(f)|_{\partial\Omega_c^h} = I(f)|_{\partial\Omega_c^h} \text{ and } \|D^\alpha(f - I_b(f))\|_{L_p(\partial\Omega_c^h)} \le Kh^{\ell-|\alpha|-1/2} |f|_{H^\ell(\Omega_c^h)}.$$

The constant K depends only on p, $\omega_{\mathcal{T}_c^h}$, and the Lipschitz constant $L_{\partial\Omega_c^h}$.

3. *Restricting $u^h \in S^h$ to the boundary $\partial\Omega_c^h$ yields*

$$(4.19)\quad u^h \in S^h \text{ satisfies } u^h|_{\Omega_c^h} = 0 \Longleftrightarrow (u^h|_{\Omega_c^h}, v^h|_{\Omega_c^h})_{L^2(\partial\Omega_c^h)} = 0 \quad \forall v^h \in S_b^h.$$

Davydov [27] states in his Remark 5.2 that 1.–3. remain correct for $n = 2$, some curved $\partial\Omega_c^h \approx \partial\Omega, 2 < p \le 5$, and modified S^h.

The convergence proofs of our FEM below are formulated on the basis of the results in Theorem 4.4. So future updates of our FEM are straightforward by these general results, based upon Theorem 5.1 and Remark 5.2 of [27].

4.4. An extension theorem and error estimates for $\Omega \in C^2$. The preceding FE spaces are defined on $\Omega_c^h \ne \Omega$. Similarly to Crouzeix and Raviart [26], Ciarlet [19], Lenoir [42], and Bernadou [5, 6], we apply extension operators, E_c (see Theorem 4.5), to the original functions. Then Theorem 4.6 correctly formulates error estimates as $\|E_c u - P^h u\|_{H^\ell(\Omega_c^h)}$. Correspondingly, the boundary conditions $B_D u = u|_{\partial\Omega} = 0$ are violated into $B_D^h u = (E_c u)|_{\partial\Omega_c^h} = \mathcal{O}(h^p)$ and for $\Omega \in C^p$, with $p = 2$ for polygonal Ω^h and $p > 2$ for curved Ω_c^h.

Different types of extension operators are studied in the literature; see, e.g., Gilbarg and Trudinger [35, pp. 136ff.], Zeidler [63, pp. 305ff.], and Brenner and Scott [13, p. 31]. Treves [58] proves Theorem 26.A.3 for functions in $u \in C^k(\overline{\Omega})$ and extends it continuously to $u \in W^{k,p}(\Omega), 1 \le p < \infty$, in Corollary 26.A.1, yielding (4.20), (4.21). This technique can obviously be extended to show (4.22).

THEOREM 4.5 (extension operator). *Let $\Omega \in C^k$, Ω_c^h, $\Omega_0 \supset \overline{(\Omega \cup \Omega_c^h)}$, and $k \in \mathbf{N}$ be given. Then there exist a linear bounded extension $E_c : H^k(\Omega) \to H^k(\Omega \cup \Omega_c^h) \to H_0^k(\Omega_0)$ and a constant, $C = C(k, \Omega, \Omega_c^h, \Omega_0)$, s.t.*

$$E_c u|_\Omega = u \text{ on } \Omega \text{ and } \|E_c u\|_{H^k(\Omega \cup \Omega_c^h)} \le C \|u\|_{H^k(\Omega)} \quad \forall u \in H^k(\Omega). \tag{4.20}$$

The traces of the partials of u and $E_c u \in H^k(\Omega \cup \Omega_c^h)$ to the boundary $\partial\Omega$ coincide:

$$H^{k-|\beta|-1/2}(\partial\Omega) \ni \partial^\beta (u - E_c u)|_{\partial\Omega} = 0 \quad \forall |\beta| \le k, \forall u \in H^k(\Omega) \tag{4.21}$$

$$\Longrightarrow \|\partial^\beta \big(E_c u(x)\big) - E_c\big(\partial^\beta u(x)\big)\|_{L^2(\Omega \cup \Omega_c^h)} \le Ch \|u\|_{H^{|\beta|}(\Omega)} = 0 \quad \forall u \in H_0^k(\Omega). \tag{4.22}$$

THEOREM 4.6 (interpolation errors). *Under the conditions* (4.6)–(4.9)*, those of Theorems* 4.4 and 4.5*, and for Ω piecewise in $C^p, 2 \le p \le 5$, a positive constant $K = K(n, d, \omega_{\mathcal{T}_c^h}, \gamma, C := C_1 C_2, L_{\partial\Omega^h})$ exists, s.t. for $u \in H^\ell(\Omega)$, and $0 \le s \le \ell - 1$ (cf.* (4.9), (4.15)*), we obtain for the operator*

$$P^h = I^h E_c : \quad \|E_c u - P^h u\|_{H^s(\mathcal{T}_c^h)} \le C\ h^{\ell - s} |u|_{H^\ell(\Omega)}, \text{ and } \|P^h u\|_{H^s(\mathcal{T}_c^h)} \le C \|u\|_{H^\ell(\Omega)}$$

$$\|E_c u - P^h u\|_{H^s(\partial\Omega_c^h)} \le C\ h^{\ell - s - 1/2} |u|_{H^\ell(\Omega)}, \text{ and } \|P^h u\|_{H^s(\partial\Omega_c^h)} \le C \|u\|_{H^\ell(\Omega)}. \tag{4.23}$$

The $\mathcal{U}^h$ approximate the $H^s(\Omega), \ell \ge s$, and $H^s(\partial\Omega)$, $\ell \ge s + 1/2$, e.g., in the sense that

$$dist(E_c u, \mathcal{U}^h) := \inf_{u^h \in \mathcal{U}^h} ||E_c u - u^h||_{H^s(\mathcal{T}_c^h)} \to 0 \quad \forall u \in H^s(\Omega) \quad \text{for } h \to 0. \tag{4.24}$$

5. An abstract convergence theorem. We want to apply the general theory of discretization methods, e.g., in Stetter [51, Definition 1.1.2] and Böhmer [9, 10], with its standard "consistency and stability implies convergence" result to our FEM, here the *conforming and the nonconforming FEMs* in the next section. We formulated the new FEM for second order equations with trivial Dirichlet boundary conditions; see (2.6). We use the notations in (3.1), (3.6), and (3.7). Let (3.8) have a locally unique solution (see section 3) with boundedly invertible derivative $G'(u_0)$.

We recall that our program for proving convergence requires several steps, summarized in section 2, **1.**–**4.**: Until now we have worked through the first two bullets of **1.**, and we turn, in this and the next section, to the last three bullets of **1.**.

The *conforming FEs,* defined on a triangulation with straight edges, $\mathcal{T}^h$, require

$$\Omega \text{ convex, bounded and polyhedral with } \overline{\Omega} = \cup_{T \in \mathcal{T}^h} \overline{T}. \tag{5.1}$$

In section 6 we will return to *nonconforming FEs,* defined on $\Omega \in C^2$. This yields the more complicated case of violated differential and boundary equations.

5.1. The original problem and its discretization. We need the following spaces:

$$\mathcal{U} = H^2 \subset \mathcal{V} = \mathcal{V}' = L^2, \quad \mathcal{U}_0 = \mathcal{U} \cap H_0^1 \text{ all defined on } \Omega. \tag{5.2}$$

G and its exact solution u_0 satisfy

$$G : \mathcal{D}(G) \subset \mathcal{U} \to \mathcal{V} \text{ and } G(u_0) = 0 \in \mathcal{V} \Longleftrightarrow (G(u_0), v)_{\mathcal{V}} = 0 \quad \forall v \in \mathcal{V}. \tag{5.3}$$

We choose the discrete G^h for G as in (2.4). The *sequences of approximating subspaces* $\mathcal{U}^h = \mathcal{V}^h = S_0^h$ in (4.4), defined on the original Ω by (5.1), satisfy the boundary conditions on the exact boundary. This FEM is a so-called *conforming Petrov–Galerkin method.* It yields the standard "consistency and stability implies convergence" result. More precisely we need

$$\{\mathcal{U}^h, \mathcal{V}^h, P^h, Q^h, G^h\}_{h\in H}, \text{ with } \inf\{0 < h \in H\} = 0, \quad \mathcal{S}_d^1(\mathcal{T}^h) \supset \mathcal{U}^h = \mathcal{V}^h = S_0^h. \tag{5.4}$$

For these spaces with identical elements we require (asymptotically) *nonequivalent norms and scalar products,* defined as those of the original spaces (cf. (5.2)), e.g.,

$$\|u^h\|_{\mathcal{U}^h} := \|u^h\|_{\mathcal{U}}, \quad \|v^h\|_{\mathcal{V}^h} := \|v^h\|_{\mathcal{V}}. \tag{5.5}$$

By Theorems 4.2 and 4.6 these sequences *approximate the Banach spaces* $\mathcal{U}, \mathcal{V}$.

For the final FE version of (2.4) we need $G^w, w_{u^h}(x), \mathcal{D}(G^w)$ in (3.6), (3.8), (5.3):

$$\mathcal{D}(G^h) := \{u^h \in \mathcal{U}^h : w_{u^h}(x) \in \mathcal{D}(G^w) \quad \forall x \in \Omega, G(u^h) \in \mathcal{V}\}. \tag{5.6}$$

Then we define the FE solution $u_0^h \in \mathcal{U}^h \subset \mathcal{U}$ s.t.

$$G^h : \mathcal{D}(G^h) \to \mathcal{V}^h, \text{ and } G^h(u_0^h) = 0 \Longleftrightarrow (G(u_0^h), v^h)_{\mathcal{V}} = 0 \quad \forall v^h \in \mathcal{V}^h \subset \mathcal{V}. \tag{5.7}$$

5.2. The necessary projectors. Hence, we introduce *sequences* of projectors, $\{P^h, Q^h\}_{h\in H}$, as appropriate tools for our FEM. The P^h have to satisfy

$$P^h \in \mathcal{L}(\mathcal{U}, \mathcal{U}^h) : \ \forall u \in \mathcal{U} \ \lim_{h\to 0} \|P^h u\|_{\mathcal{U}^h} = \|u\|_{\mathcal{U}}, \text{ we choose } P^h u := I^h u. \tag{5.8}$$

Instead of interpolating values of functions or derivatives as I^h in (4.10), the P^h may be based upon some kind of quasi- or averaging-interpolation or best approximation operator, sometimes called Clément interpolation; cf., e.g., Clément [20], de Boor [31], Scott and Zhang [48], and Davydov [27].

More delicate than P^h are the projectors to the image spaces, the $Q^h \in \mathcal{L}(\mathcal{V}, \mathcal{V}^{h'})$, motivated by G^h. We test $G(u_0) = 0$ and $G^h(u_0^h) = 0$ as in (5.3) and in (5.7). We keep the concepts for the original and the discrete problem as close as possible and use the function spaces and norms in (5.2), (5.4), (5.5). In (5.3) and (5.7) we have satisfied the differential equation and its discrete counterpart by testing with $\mathcal{V}$ and $\mathcal{V}^h$. We define (see (4.20)–(4.22) and Theorem 4.6), for all $f \in \mathcal{V}$,[2]

$$Q^h \in \mathcal{L}(\mathcal{V}, \mathcal{V}^{h'}) : \ \langle Q^h f - f, v^h\rangle_{(\mathcal{V}^{h'}\cup\mathcal{V})\times\mathcal{V}^h} = 0 \quad \forall v^h \in \mathcal{V}^h; \ \lim_{h\to 0} \|Q^h f\|_{\mathcal{V}^{h'}} = \|f\|_{\mathcal{V}}. \tag{5.9}$$

The last limit is obtained as a consequence of (4.11). By (4.13), (4.14) with $\mathcal{V}^h \neq \mathcal{V}^{h'}$ it makes sense that for a function, e.g., $f \in \mathcal{V}$ or $G(u^h) \notin \mathcal{V}^h$, the $Q^h f \in \mathcal{V}^{h'}$ is not a function. We did not identify $\mathcal{V}^h \neq \mathcal{V}^{h'}$ for emphasizing this situation more clearly.

The choice of P^h, Q^h, for the given $\mathcal{U}^h, \mathcal{V}^h$, and problem (5.3) certainly is not unique. The combination $\mathcal{U}, \mathcal{V}, \mathcal{U}^h, \mathcal{V}^h, P^h, Q^h, G^h$ should be chosen appropriately to yield classical consistency (see below) with the highest possible order; see (5.13).

[2]Systematically we use the following identification: For $\mathcal{U} \subset \mathcal{V} = \mathcal{V}'$, for all $u \in \mathcal{U}$, for all $v \in \mathcal{V} \subset \mathcal{U}'$, and the corresponding discrete terms, we obtain $(u, v)_{\mathcal{V}} = \langle u, v\rangle_{\mathcal{U}\times\mathcal{U}'} = \langle u, v\rangle_{\mathcal{U}'\times\mathcal{U}}$, similarly for $\mathcal{V}_D \subset \mathcal{V}_b = \mathcal{V}_b' \subset \mathcal{V}_D', \ldots$, and $\mathcal{W} \subset \mathcal{V} = \mathcal{V}' \subset \mathcal{W}'$ below. This is applied for the following projectors $Q^h, \ldots$ as well.

$$\begin{array}{ccccc} \mathcal{U} & \xrightarrow{G} & \mathcal{V} & \overset{tested\ by}{\longleftrightarrow} & \mathcal{V}=\mathcal{V}' \\ P^h\big\downarrow & \Phi^h\big\downarrow & Q^h\big\downarrow & & \\ \mathcal{U}^h & \xrightarrow{G^h} & \mathcal{V}^{h'} & \overset{tested\ by}{\longleftrightarrow} & \mathcal{V}^h\neq\mathcal{V}^{h'} \end{array}$$

FIG. 5.1. *Conforming FEMs for fully nonlinear problems.*

5.3. Projectors and discretization. For the nonlinear operator, $G : \mathcal{D}(G) \subset \mathcal{U} \to \mathcal{V}$, in (5.3), the previously introduced projectors P^h, Q^h allow a shorthand notation of the corresponding discrete operator, $G^h : \mathcal{D}(G^h) \subset \mathcal{U}^h \to \mathcal{V}^{h'}$; see (5.7). So we determine the exact solution $u_0^h \in \mathcal{U}^h$ s.t.

$$\langle G(u_0^h), v^h\rangle_{\mathcal{V}^{h'}\times\mathcal{V}^h} = 0 \quad \forall v^h \in \mathcal{V}^h \Longleftrightarrow G^h(u_0^h) = 0 \text{ with } G^h := Q^h G|_{\mathcal{U}^h}. \tag{5.10}$$

This brings about a mapping

$$\mathcal{D}(\Phi^h) := \{G \text{ in (5.3)}\} \text{ and } \Phi^h \text{ defined by } \Phi^h(G) := G^h := Q^h G|_{\mathcal{U}^h} \text{ in (5.10)}. \tag{5.11}$$

This Φ^h is even linear s.t.

$$\Phi^h(c_1G^1 + c_2G^2) = c_1\Phi^h(G^1) + c_2\Phi^h(G^2) \quad \forall c_1, c_2 \in \mathbf{R},\ \forall G^1, G^2 \in \mathcal{D}(\Phi^h). \tag{5.12}$$

Linearity is essential for the MIP in section 9, allowing the proof of a quadratically convergent discrete Newton method.

Our modification of Stetter [51] and Vainikko [59] emphasizes the testing with the $\mathcal{V},\ \mathcal{V}^h$: Summarizing yields Figure 5.1 with uniformly bounded P^h, Q^h.

A method is called *convergent* and *convergent of order k* if the solutions u_0 and u_0^h of $Gu_0 = 0$ and $G^h u_0^h = 0$ in (6.5) and (5.10), satisfy[3]

$$\|u_0^h - P^h u_0\|_{\mathcal{U}^h} \to 0 \text{ and } \|u_0^h - P^h u_0\|_{\mathcal{U}^h} = \mathcal{O}(h^k) \text{ for } h \to 0, \text{ respectively.}$$

These $\mathcal{O}(h^k)$ are usually estimated by $Ch^k\|u_0\|_{H^{k'}(\Omega)}$ with appropriate k, k'.

For convergence, we need the *classical consistency errors* for general u (see (5.3)):

$$\|G^h P^h u - Q^h G u\|_{\mathcal{V}^{h'}} \quad \text{for } u_0 : \|G^h P^h u_0 - Q^h G u_0\|_{\mathcal{V}^{h'}} = \|Q^h G P^h u_0\|_{\mathcal{V}^{h'}}, \tag{5.13}$$

where these residuals can be used for adaptive methods. A discretization method for $G(u) = 0$ is called *consistent* and *consistent of order k* in u, if, respectively,

$$\|Q^h G P^h u - Q^h G u\|_{\mathcal{V}^{h'}} \to 0 \text{ and } \|Q^h G P^h u - Q^h G u\|_{\mathcal{V}^{h'}} = \mathcal{O}(h^k) \text{ for } h \to 0. \tag{5.14}$$

Our conforming FEM is consistent for a continuous or Lipschitz continuous G, e.g.,

$$\|Q^h G P^h u - Q^h G u\|_{\mathcal{V}^{h'}} \le \|Q^h\|_{\mathcal{V}^{h'}\hookleftarrow\mathcal{V}} L \|P^h u - u\|_{\mathcal{U}}. \tag{5.15}$$

For the *stability,* we start with a fixed sequence $\{u^h \in \mathcal{D}(G^h)\}_{h\in H}$. Assume there exist constants, $h_0, r, S \in \mathbf{R}_+$, *independent of h,* s.t., *uniformly* for all $h \in H, h < h_0$,

$$\forall\ u_i^h \in B_r(u^h) \subset \mathcal{D}(G^h), \quad i = 1, 2, \Longrightarrow \|u_1^h - u_2^h\|_{\mathcal{U}^h} \le S\|G^h(u_1^h) - G^h(u_2^h)\|_{\mathcal{V}^{h'}}, \tag{5.16}$$

[3] Although we could use $\|u_0^h - u_0\|_{\mathcal{U}}$ in our case, we use the notation $\|u_0^h - P^h u_0\|_{\mathcal{U}^h}$, applicable to the nonconforming case below.

with $B_r(u^h) = \{v^h \in \mathcal{U}^h : \|v^h - u^h\|_{\mathcal{U}^h} < r\}$, similarly for $B_r(u)$. We always assume $B_r(u), B_r(u^h) \subset \mathcal{D}(G)$. G^h is called *stable in the sequence* $\{u^h\}_{h\in H}$ or for short *stable in* u^h, and the S and r are called *stability bound and stability threshold,* respectively. Often we choose $u^h := P^h u_0$. Combining Theorems 1.2.3 and 1.2.4 in Stetter [51] we get the following.

THEOREM 5.1. *Let the original problem* (5.3) *have the exact solution* $u_0 \in \mathcal{U}$. *Let its discretization* $G^h = Q^h G|_{\mathcal{U}^h} : \mathcal{U}^h \to \mathcal{V}^{h'}$ *in* (5.10) *satisfy the following:*

1. *The projectors* P^h *and* $Q^h_{c,\Pi}$ *are (equi-)bounded and satisfy* (5.8) *and* (5.9).
2. $G^h : \mathcal{U}^h \to \mathcal{V}^{h'}$ *is defined and continuous in* $B_r(P^h u_0),\ r > 0, h$*-independent.*
3. G^h *is consistent with* G *in* $P^h u_0$.
4. G^h *is stable for* $P^h u_0$.

Then the discrete problem $G^h(u^h) = 0$ *possesses a unique solution* $u^h_0 \in \mathcal{U}^h$ *near* u_0 *for all sufficiently small* $h \in H$, *and* u^h_0 *converges to* u_0. *If* G^h *is consistent and consistent of order* p, *then* u^h_0 *converges and converges of order* p, *respectively:*

$$\|u^h_0 - P^h u_0\|_{\mathcal{U}^h} \le S\|Q^h G P^h u_0\|_{\mathcal{V}^{h'}} \to 0 \text{ and } \|u^h_0 - P^h u_0\|_{\mathcal{U}^h} \le \mathcal{O}(h^p). \tag{5.17}$$

We will discuss the significance of the conditions 1.–4. in Remark 6.1 below.

6. An abstract convergence theorem for nonconforming FEs. Now we return to *nonconforming FEs.* So we require

$$\Omega \in C^2 \text{ and a triangulation, } \mathcal{T}^h_c, \text{ with curved edges, s.t. } \overline{\Omega} = \cup_{T\in\mathcal{T}^h_c}\overline{T}. \tag{6.1}$$

Again, the first two paragraphs in section 5 remain valid. The $\mathcal{U}, \mathcal{W}, \mathcal{V} = \mathcal{V}'$, defined on the original Ω, now are approximated by *sequences of nonconforming subspaces* $\mathcal{U}^h, \mathcal{W}^h, \mathcal{V}^h$, defined on the approximate Ω^h_c. This implies *violated boundary conditions* on the exact boundary; cf. (6.2). The FEM changes into a so-called *nonconforming Petrov–Galerkin method,* requiring a more complicated treatment than in section 5.

6.1. The original problem and its nonconforming discretization. We need (cf. (5.2))

$$\begin{aligned}&\mathcal{U} = H^2 \subset \mathcal{W} = H^1 \subset \mathcal{V} = \mathcal{V}' = L^2 \subset \mathcal{W}' = H^{-1},\ \mathcal{W}_0 = H^1_0 \text{ all defined on } \Omega \text{ versus}\\ &\mathcal{U}_c = H^2 \subset \mathcal{W}_c = H^1 \subset \mathcal{V}_c = L^2 \subset \mathcal{W}'_c = H^{-1},\ \mathcal{W}_{c,0} = H^1_0 \text{ all defined on } \Omega^h_c.\end{aligned} \tag{6.2}$$

We have indicated above that the realization of the differential equation *and* the boundary condition and their violation is a deciding step of this method. Hence, the discretization method has to care about both aspects. This is achieved by the new operator F with the differential operator and the boundary operator as its two components; see (2.1), (2.2). The boundary operator, B_D, maps, by the trace theorem,

$$B_D : \mathcal{U} \to \mathcal{V}_D := H^{3/2}(\partial\Omega) = B_D\mathcal{U}, B_D u := u|_{\partial\Omega}, \quad \mathcal{V}_D \subset \mathcal{V}'_b = \mathcal{V}_b := L^2(\partial\Omega). \tag{6.3}$$

We consider here mainly trivial Dirichlet conditions. Thus we have to characterize

$$u|_{\partial\Omega} = 0 \Leftrightarrow \langle v', u|_{\partial\Omega}\rangle_{\mathcal{V}'_D\times\mathcal{V}_D} = 0 \quad \forall v' \in \mathcal{V}'_D \Leftrightarrow (u|_{\partial\Omega}, v)_{L^2(\partial\Omega)} = 0 \quad \forall v \in \mathcal{V}_b \tag{6.4}$$

for $u \in \mathcal{U}$, with $\mathcal{V}_D = H^{3/2}(\partial\Omega) \hookrightarrow \mathcal{V}_b = \mathcal{V}'_b \hookrightarrow \mathcal{V}'_D$, a dense (compact) embedding.

The transformation from nontrivial to trivial Dirichlet conditions will be combined in Theorems 7.1, 8.4, and 8.7 obtaining stability and convergence for the inhomogeneous case as well.

$F = (G, B_D)$ is called *elliptic* simultaneously with G. F and its Fréchet derivative $F'(u)$ map $\mathcal{U}$ into and onto, respectively, the Cartesian product $\mathcal{V} \times \mathcal{V}_D$:

$$F := (F_1, F_2) := (G, B_D) : \mathcal{U} \to \mathcal{V}_\Pi := \mathcal{V} \times \mathcal{V}_D \subset \hat{\mathcal{V}}_\Pi := \mathcal{V} \times \mathcal{V}_b = \hat{\mathcal{V}}_\Pi'; \tag{6.5}$$

see (3.7). Since $\mathcal{V}_D$ is dense in $\mathcal{V}_b$, the exact solution u_0 is hence characterized by

$$\begin{aligned} F(u_0) = 0 \in \mathcal{V}_\Pi &\Longleftrightarrow (F(u_0), v_\Pi)_{\hat{\mathcal{V}}_\Pi} = 0 \quad \forall v_\Pi := (v, v_b) \in \hat{\mathcal{V}}_\Pi \\ &\Longleftrightarrow (G(u_0), v)_{\mathcal{V}} = 0 \ \forall v \in \mathcal{V}, \text{ and } (B_D u_0, v_b)_{\mathcal{V}_b} = 0 \quad \forall v_b \in \mathcal{V}_b, \\ \text{abbreviated } &\Longleftrightarrow (G(u_0), v)_{\mathcal{V}} + (B_D u_0, v_b)_{\mathcal{V}_b} = 0 \quad \forall v \in \mathcal{V}, \quad v_b \in \mathcal{V}_b. \end{aligned} \tag{6.6}$$

For the standard weak form for $F'(u)$ in section 8, we need (cf. (6.3), (6.4))

$$\begin{aligned} &B_D : \mathcal{W} \to \mathcal{W}_D := H^{1/2}(\partial\Omega) = B_D \mathcal{W}, \quad u \to B_D u = u|_{\partial\Omega}, \text{ with} \\ &u \in \mathcal{W} \text{ and } B_D u = 0 \Longleftrightarrow (u|_{\partial\Omega}, v)_{L^2(\partial\Omega)} = 0 \quad \forall v \in \mathcal{V}_b \text{ dense in } \mathcal{W}_D'. \end{aligned} \tag{6.7}$$

$B_D : \mathcal{U} \to \mathcal{V}_D$ and $B_D : \mathcal{W} \to \mathcal{W}_D$ are continuous w.r.t. the corresponding Sobolev–Sloboditskii norms [37, 43]. Similarly to (6.2), we introduce

$$\begin{aligned} &\mathcal{V}_{c,D} := H^{3/2}(\partial\Omega_c^h) \subset \mathcal{W}_{c,D} := H^{1/2}(\partial\Omega_c^h) \subset \mathcal{V}_{c,b} := L^2(\partial\Omega_c^h) \text{ and} \\ &B_{c,D} : \mathcal{U}_c \to \mathcal{V}_{c,D}, u_c \to B_{c,D} u_c := u_c|_{\partial\Omega_c^h}; \text{ similarly, } B_{c,D} : \mathcal{W}_c \to \mathcal{W}_{c,D}. \end{aligned} \tag{6.8}$$

To apply the general discretization theory [51, 10], we introduce the discrete F^h for F, the corresponding projectors $P^h \in \mathcal{L}(\mathcal{U}, \mathcal{U}^h)$, and $Q_\Pi^h \in \mathcal{L}(\mathcal{V}_\Pi, \mathcal{V}_\Pi^{h'})$, $Q_{c,\Pi}^h \in \mathcal{L}(\mathcal{V}_{c,\Pi}, \mathcal{V}_\Pi^{h'})$; see (6.26). This yields the standard "consistency and stability implies convergence" result, (6.1); mind $S_0^h = S^h \cap H_0^1(\mathcal{T}_c^h)$:

$$\mathcal{S}_d^1(\mathcal{T}_c^h) \supset \mathcal{U}^h = \mathcal{W}^h = S^h = S_0^h \oplus S_b^h, \quad \mathcal{V}^h = S_0^h, \quad \mathcal{V}_D^h = \mathcal{W}_D^h = \mathcal{V}_b^h := S_b^h|_{\partial\Omega_c^h}. \tag{6.9}$$

Again these spaces require *nonequivalent norms and scalar products.* With the norms of the original spaces, now defined on Ω_c^h and $\partial\Omega_c^h$ (cf. (6.2), (6.8)), let

$$\begin{aligned} &\|u^h\|_{\mathcal{U}^h} := \|u^h\|_{\mathcal{U}_c}, \quad \|w^h\|_{\mathcal{W}^h} := \|w^h\|_{\mathcal{W}_c}, \quad \|v^h\|_{\mathcal{V}^h} := \|v^h\|_{\mathcal{V}_c}, \quad \|u^h|_{\partial\Omega_c^h}\|_{\mathcal{V}_D^h} \\ &:= \|u^h|_{\partial\Omega_c^h}\|_{\mathcal{V}_{c,D}}, \quad \|w^h|_{\partial\Omega_c^h}\|_{\mathcal{W}_D^h} := \|w^h|_{\partial\Omega_c^h}\|_{\mathcal{W}_{c,D}}, \quad \|v^h|_{\partial\Omega_c^h}\|_{\mathcal{V}_b^h} := \|v^h|_{\partial\Omega_c^h}\|_{\mathcal{V}_{c,b}}. \end{aligned} \tag{6.10}$$

By Theorems 4.2 and 4.6 these sequences *approximate the Banach spaces* $E_c\mathcal{U}, E_c\mathcal{W}$, $E_c\mathcal{V}, E_c\mathcal{V}_D, E_c\mathcal{W}_D, E_c\mathcal{V}_b$ defined on Ω_c^h and $\partial\Omega_c^h$ w.r.t. the corresponding norms. The $u^h \in S^h$, and $v^h \in \mathcal{V}_b^h$, violate the original boundary conditions.

The product operator $F : \mathcal{U} \to \mathcal{V} \times \mathcal{V}_D$ and the corresponding $F^h : \mathcal{U}^h \to \mathcal{V}^{h'} \times \mathcal{V}_D^h$ cause some unpleasant technicalities in the definition of Q_Π^h, $Q_{c,\Pi}^h$, compared to section 5. Subsequently, the consistency theory becomes relatively simple again; see Figure 6.1 and section 7. These technicalities are not surprising compared to some other approaches for violated boundary conditions, e.g., in Lenoir [42] and Bernadou [5, 6]. The unavoidable complications are piled up in section 8.

The terms u^h, $G(u^h)$, and $u^h|_{\partial\Omega_c^h}$ are, by (3.7), defined on Ω_c^h and $\partial\Omega_c^h$ instead of the original Ω and $\partial\Omega$. Thus we start updating the preliminary FE version (2.6). With $G^w, \mathcal{D}(G^w), w_{u^h}(x)$ in (3.6), (3.7), (3.8), and $\mathcal{U}^h$ in (4.4), let

$$\mathcal{D}(G^h) := \{u^h \in \mathcal{U}^h : w_{u^h}(x) \in \mathcal{D}(G^w) \quad \forall x \in \Omega \cup \Omega_c^h, G(u^h) := G^w(w_{u^h}) \in \mathcal{V}_c\}. \tag{6.11}$$

This extended definition of G^w was the reason for introducing the Ω' in (3.6).

We replace the differential equation $G(u_0) = 0 \in \mathcal{V}, u_0 \in \mathcal{D}(G)$, in (6.5) by the FE equations in equivalent formulations; see (2.5), (2.6), (6.5): Define (cf. (6.10))

$$\text{(6.12) the sequence } F_1^h(u_0^h) = G^h(u_0^h) = 0 \Longleftrightarrow (G(u_0^h), v^h)_{\mathcal{V}_c} = 0 \quad \forall v^h \in \mathcal{V}^h \subset \mathcal{V}_c.$$

Similarly, the boundary conditions (6.4) are tested (see the equivalence (4.19)) as

$$\text{(6.13)} \quad u^h \in \mathcal{U}^h : F_2^h(u^h) = B_D^h u^h = u^h|_{\partial\Omega_c^h} = 0 \in \mathcal{V}_D^h \Leftrightarrow (v_b^h, u^h)_{\mathcal{V}_b^h} = 0 \quad \forall v_b^h \in S_b^h,$$

essentially with S_b^h dense in $\mathcal{V}'_{c,D}$. Consequently we introduce (cf. F in (6.5))

$$F^h = (F_1^h, F_2^h) = (G^h, B_D^h) \text{ with } \mathcal{D}(F^h) := \mathcal{D}(G^h). \tag{6.14}$$

6.2. The nonconforming projectors. Again, projectors represent an appropriate tool for $F^h = (F_1^h, F_2^h)$. They have to handle the discrepancy between u^h, $G(u^h)$, and $u^h|_{\partial\Omega_c^h}$, by (3.7), defined on Ω_c^h and $\partial\Omega_c^h$ instead of the original Ω and $\partial\Omega$. We combine the extension operator E_c introduced in Theorem 4.5 with inverse estimates and the estimates for interpolation errors on Ω in Theorems 4.1–4.6. The *sequence* of P^h has to satisfy

$$P^h \in \mathcal{L}(\mathcal{U},\mathcal{U}^h) : \ \forall u \in \mathcal{U} \ \lim_{h\to 0} \|P^h u\|_{\mathcal{U}^h} = \|u\|_{\mathcal{U}}, \text{ we choose } P^h u := I^h E_c u. \tag{6.15}$$

E_c is unnecessary for the Argyris FEs but possibly needed for other schemes.

Again more delicate than P^h are the projectors on the product space, $Q_\Pi^h \in \mathcal{L}(\mathcal{V}_\Pi, \mathcal{V}_\Pi^{h'})$. They live on the original and discrete image spaces, listed in (6.16)–(6.18), with their duals, scalar products, and pairings; cf. (6.5), (6.6). For the reader's convenience, we summarize and introduce the spaces defined on Ω and Ω_c^h; cf. (4.13), (4.14), Theorem 4.4, (6.2), (6.4), (6.9), (6.12), (6.13). We mark functions and functionals, according to their definition, by obvious indices. For functions defined on Ω_c^h we use $_c$, for $\partial\Omega$ we use $_b$, for $\partial\Omega_c^h$ we use $_{c,b}$, for the images of B_D on $\partial\Omega_c^h$ we use $_{c,D}$, and on the product of domain and boundary on Ω and Ω_c^h we use $_\Pi$ or $_{c,\Pi}$:

$$\begin{aligned} &\mathcal{V}_\Pi := \mathcal{V} \times \mathcal{V}_D \subset \hat{\mathcal{V}}_\Pi := \mathcal{V} \times \mathcal{V}_b, \text{ with } \mathcal{U} = H^2(\Omega), \mathcal{V} = L^2(\Omega) = \mathcal{V}', \\ &\mathcal{V}_b = L^2(\partial\Omega) = \mathcal{V}_b' \supset \mathcal{V}_D = B_B\mathcal{U} = H^{3/2}(\partial\Omega), \text{ defined on } \Omega, \text{ with} \\ &\big((u,u_b),(v,v_b)\big)_{\mathcal{V}_\Pi} := (u,v)_{\mathcal{V}} + (u_b,v_b)_{\mathcal{V}_D}, \ \big((u,u_b),(v,v_b)\big)_{\hat{\mathcal{V}}_\Pi} := (u,v)_{\mathcal{V}} + (u_b,v_b)_{\mathcal{V}_b}. \end{aligned} \tag{6.16}$$

Similarly, we define the spaces and scalar products on the curved approximations

$$\mathcal{V}_c = L^2(\Omega_c^h), \quad \mathcal{V}_{c,D} = B_B\mathcal{U}_c = H^{3/2}(\partial\Omega_c^h), \quad \mathcal{V}_{c,\Pi}, \mathcal{V}_{c,b}, \dots, \text{ defined on } \Omega_c. \tag{6.17}$$

Correspondingly, the discrete analog spaces and their scalar products are

$$\begin{aligned} &\mathcal{V}_\Pi^h := \mathcal{V}^h \times \mathcal{V}_D^h := S_0^h \times S_b^h|_{\partial\Omega_c^h}, \ \mathcal{V}^{h'} \neq \mathcal{V}^h, \ \mathcal{V}_\Pi^{h'} = \mathcal{V}^{h'} \times \mathcal{V}_D^h := S_0^{h'} \times S_b^h|_{\partial\Omega_c^h} \neq \mathcal{V}_\Pi^h, \\ &\hat{\mathcal{V}}_\Pi^h := \mathcal{V}^h \times \mathcal{V}_b^h, \quad \langle (u^{h'},u_b^h),(v^h,v_b^h)\rangle_{\mathcal{V}_\Pi^{h'}\times\mathcal{V}_\Pi^h} := \langle u^{h'},v^h\rangle_{\mathcal{V}^{h'}\times\mathcal{V}^h} + (u_b^h,v_b^h)_{\mathcal{V}_D^h}. \end{aligned} \tag{6.18}$$

In (6.6), (6.12) and (6.4), (6.13) we have satisfied the differential equation and boundary condition or their discrete counterparts by testing with $\mathcal{V}$ and $\mathcal{V}_b$ or (see the equivalence (4.19)) with $\mathcal{V}^h$ and $\mathcal{V}_b^h$. The two components (6.6), (6.4) and (6.12), (6.13) of F and F^h, defined on Ω and Ω_c^h, respectively, are now combined with the

extension operator E_c in (4.20) for motivating the appropriate definitions of the projection operators. For the discrete spaces $\mathcal{V}^{h'} \neq \mathcal{V}^h$ and $\mathcal{V}_b^{h'} = \mathcal{V}_b^h$ the FE equations always are tested with $v^h \in \mathcal{V}^h$ and $v_b^h \in \mathcal{V}_b^h$, but we use for the boundary condition the equivalence (6.13). Thus corresponding to the product spaces $\mathcal{V}_\Pi$ and $\mathcal{V}_{c,\Pi}$ we have to introduce two projectors $Q_\Pi^h := (Q^h, Q_b^h) \in \mathcal{L}(\mathcal{V}_\Pi, \mathcal{V}_\Pi^{h'})$ and $Q_{c,\Pi}^h := (Q_c^h, Q_{c,b}^h) \in \mathcal{L}(\mathcal{V}_{c,\Pi}, \mathcal{V}_\Pi^{h'})$. We start with the component of the differential operator, (6.12) (mind $\mathcal{V} = \mathcal{V}', \mathcal{V}_b = \mathcal{V}_b'$), and define (see (4.20)–(4.22) and Theorem 4.6)

$$(6.19)\quad Q^h \in \mathcal{L}(\mathcal{V}, \mathcal{V}^{h'}) \text{ for } f \in \mathcal{V} \text{ by } \langle Q^h f, v^h\rangle_{\mathcal{V}^{h'}\times\mathcal{V}^h} - (E_c f, v^h)_{\mathcal{V}_c} = 0 \quad \forall v^h \in \mathcal{V}^h.$$

The approximate domains require

$$Q_c^h \in \mathcal{L}(\mathcal{V}_c, \mathcal{V}^{h'}) \text{ for } f_c \in \mathcal{V}_c:\ \langle Q_c^h f_c, v^h\rangle_{\mathcal{V}^{h'}\times\mathcal{V}^h} - (f_c, v^h)_{\mathcal{V}_c} = 0 \quad \forall v^h \in \mathcal{V}^h.$$

Both projectors yield the norm limits, the last limit as a consequence of (4.11):

$$(6.20)\qquad \lim_{h\to 0} \|Q^h f\|_{\mathcal{V}^{h'}} = \|f\|_{\mathcal{V}}, \text{ and } \lim_{h\to 0} \|Q_c^h f_c\|_{\mathcal{V}^{h'}} = \lim_{h\to 0} \|f_c\|_{\mathcal{V}_c}.$$

Since the S_0^h and $S_0^{h'}$ form a pair of dual spaces (see Theorem 4.4), we alternatively formulate corresponding projectors (see Theorem 4.6) and use them, e.g., in (8.31):

$$(6.21)\qquad Q_d^h \in \mathcal{L}(\mathcal{V}, \mathcal{V}^h), \text{ by } f \in \mathcal{V}:\ Q_d^h f := \sum_{j=1}^{m_0} \langle Q^h f, s^j\rangle_{\mathcal{V}^{h'}\times\mathcal{V}^h} s^j \in \mathcal{V}^h$$

$$\text{again with } \lim_{h\to 0} \|Q_d^h f\|_{\mathcal{V}^h} = \|f\|_{\mathcal{V}};\ \text{similarly, e.g., } \lim_{h\to 0} \|Q_{c,d}^h f_c\|_{\mathcal{V}^h} = \lim_{h\to 0} \|f_c\|_{\mathcal{V}_c}.$$

We turn to the projectors for the traces on the boundary. We always have to recall that the $u_b \in \mathcal{V}_D$ and $u_b^h \in \mathcal{V}_b^h \subset \mathcal{V}_b$ have the forms $u_b = u|_{\partial\Omega}, u \in \mathcal{U}$ and $u_b^h = u^h|_{\partial\Omega_c^h}, u^h \in \mathcal{V}^h$, respectively. It suffices testing vanishing boundary conditions w.r.t. the L^2 scalar product $(\cdot,\cdot)_{\mathcal{V}_{c,b}}$. This and the dense embeddings $\mathcal{V}_D \subset \mathcal{V}_b$ and $\mathcal{V}_{c,D} \subset \mathcal{V}_{c,b}$ motivate the following projectors:

$$(6.22)\quad Q_b^h \in \mathcal{L}(\mathcal{V}_D, \mathcal{V}_D^h) \quad \text{for } u_b := u|_{\partial\Omega} \in \mathcal{V}_D \subset \mathcal{V}_b : (Q_b^h u_b - (E_c u)|_{\partial\Omega_c^h}, v_b^h)_{\mathcal{V}_{c,b}} = 0,$$
$$Q_{c,b}^h \in \mathcal{L}(\mathcal{V}_{c,D}, \mathcal{V}_D^h), \quad u_{c,b} := u|_{\partial\Omega_c^h} \in \mathcal{V}_{c,D} : (Q_{c,b}^h u_{c,b} - u_{c,b}, v_b^h)_{\mathcal{V}_{c,b}} = 0 \quad \forall v_b^h \in \mathcal{V}_b^h,$$

and by (6.9) we obtain for all $v_b^h \in \mathcal{V}_b^h = \mathcal{V}_D^h$ that $v_b^h = v^h|_{\partial\Omega_c^h}$ for an appropriate $v^h \in \mathcal{V}^h$ and $v_b^h = Q_{c,b}^h v^h|_{\partial\Omega_c^h}$. The convergence results along $\partial\Omega_c^h$ in (4.12), (4.19), and Theorems 4.4 and 4.6 imply the limits of the boundary norms

$$(6.23)\qquad \lim_{h\to 0} \|Q_{c,b}^h u_{c,b}\|_{\mathcal{V}_D^h} = \lim_{h\to 0} \|u_{c,b}\|_{\mathcal{V}_{c,D}}, \quad \lim_{h\to 0} \|Q_b^h u_b\|_{\mathcal{V}_D^h} = \lim_{h\to 0} \|u_b\|_{\mathcal{V}_D}.$$

We summarize the product projectors and norms in (6.19)–(6.23) obtaining

$$(6.24)\qquad Q_\Pi^h := (Q^h, Q_b^h) \in \mathcal{L}\big(\mathcal{V}\times\mathcal{V}_D, \mathcal{V}^{h'}\times\mathcal{V}_D^h\big) = \mathcal{L}\big(\mathcal{V}_\Pi, \mathcal{V}_\Pi^{h'}\big),$$
$$Q_{d,\Pi}^h := (Q_d^h, Q_b^h) \in \mathcal{L}\big(\mathcal{V}_\Pi, \mathcal{V}_\Pi^h\big), \text{with } \|(f, u|_{\partial\Omega})\|_{\mathcal{V}_\Pi} = \|f\|_{\mathcal{V}} + \|u|_{\partial\Omega}\|_{\mathcal{V}_D}, \text{ and}$$
$$\lim_{h\to 0} \|Q_\Pi^h (f, u|_{\partial\Omega})\|_{\mathcal{V}_\Pi^{h'}} = \lim_{h\to 0} \|Q_{d,\Pi}^h (f, u|_{\partial\Omega})\|_{\mathcal{V}_\Pi^h} = \|(f, u|_{\partial\Omega})\|_{\mathcal{V}_\Pi}.$$

The approximate domains yield

$$(6.25)\qquad Q_{c,\Pi}^h := (Q_c^h, Q_{c,b}^h) \in \mathcal{L}\big(\mathcal{V}_c\times\mathcal{V}_{c,D}, \mathcal{V}_\Pi^{h'}\big) = \mathcal{L}\big(\mathcal{V}_{c,\Pi}, \mathcal{V}_\Pi^{h'}\big),$$
$$\lim_{h\to 0} \|Q_{c,\Pi}^h (f_c, E_c u|_{\partial\Omega_c^h})\|_{\mathcal{V}_\Pi^{h'}} = \lim_{h\to 0} \|(f_c, E_c u|_{\partial\Omega_c^h})\|_{\mathcal{V}_{c,\Pi}}.$$

Again, the choice of $P^h, Q_\Pi^h, Q_{c,\Pi}^h, \ldots$ is not unique.

$$\begin{array}{ccccc} \mathcal{U} & \xrightarrow{F} & \mathcal{V}_\Pi = \mathcal{V}\times\mathcal{V}_D & \overset{tested\ by}{\longleftrightarrow} & \hat{\mathcal{V}}_\Pi = \mathcal{V}\times\mathcal{V}_b, \text{ cf. (6.6)} \\ P^h\big\downarrow & \Phi^h\big\downarrow & Q^h_\Pi\big\downarrow & & \\ \mathcal{U}^h & \xrightarrow{F^h} & \mathcal{V}^{h'}_\Pi = \mathcal{V}^{h'}\times\mathcal{V}^h_D & \overset{tested\ by}{\longleftrightarrow} & \hat{\mathcal{V}}^h_\Pi = \mathcal{V}^h\times\mathcal{V}^h_b, \text{ cf. (6.26)} \end{array}$$

FIG. 6.1. *Nonconforming FEMs for fully nonlinear problems.*

6.3. Nonconforming projectors and discretization. We need only a few changes compared to subsection 5.3. The operator, spaces, and projectors there, $G:\mathcal{D}(G)\subset\mathcal{U}\to\mathcal{V}, Q^h$, have to be replaced by the product terms, $F=(F_1,F_2)=(G,B_D):\mathcal{D}(F)\subset\mathcal{U}\to\mathcal{V}_\Pi=\mathcal{V}\times\mathcal{V}_D$, $Q^h_\Pi, Q^h_{c,\Pi}$. Again we obtain a short definition of the corresponding discrete operator, $F^h:\mathcal{U}^h\to\mathcal{V}^{h'}\times\mathcal{V}^h_b$, and a linear mapping, Φ^h; see (6.11), (6.12) and (6.4), (6.13), (6.16)–(6.18), (6.24), (5.11). We finally get

$$F^h(u^h)=(G^h(u^h),u^h|_{\partial\Omega^h_c})=(Q^h_cG(u^h),Q^h_{c,b}u^h|_{\partial\Omega^h_c}). \tag{6.26}$$

We determine the approximate solution u^h_0 from

$$\begin{aligned} &u^h_0\in\mathcal{U}^h \text{ s.t. } F^h(u^h_0)=0\Longleftrightarrow(F(u^h_0),v^h_\Pi)_{\mathcal{V}_{c,\Pi}}=0\quad\forall v^h_\Pi=(v^h,v^h_b)\in\mathcal{V}^h_\Pi \\ &\Longleftrightarrow(G(u^h_0),v^h)_{\mathcal{V}^h}+(u^h_0,v^h_b)_{\mathcal{V}^h_b}=0\quad\forall v^h\in S^h_0=\mathcal{V}^h\subset\mathcal{V}_c\quad\forall v^h_b\in S^h_b=\mathcal{V}^h_b. \end{aligned} \tag{6.27}$$

By (6.26), (6.27) the new FEM is defined for the $F=(G,B_D)$ in (6.5) (cf. (6.8)) as

$$\mathcal{D}(\Phi^h):=\{F=(G,B_D)\}:\Phi^h(F):=F^h=Q^h_{c,\Pi}F|_{\mathcal{U}^h}=(Q^h_cG|_{\mathcal{U}^h},Q^h_{c,b}B_{c,D}|_{\mathcal{U}^h}). \tag{6.28}$$

We have thus defined a *nonconforming Petrov–Galerkin method applicable* to every F in (6.5). With uniformly bounded $P^h, Q^h_\Pi, Q^h_{c,\Pi}$ we modify the above Figure 5.1 into Figure 6.1 We reformulate only the not quite trivial consistency:

$$\|F^hP^hu-Q^h_\Pi Fu\|_{\mathcal{V}^{h'}_\Pi}=\|\big((Q^h_cG(P^hu)-Q^hG(u)),(P^hu|_{\partial\Omega^h_c}-E_cu|_{\partial\Omega^h_c})\big)\|_{\mathcal{V}^{h'}_\Pi}. \tag{6.29}$$

Remark 6.1. 1. With the above changes, *convergence, consistency, stability, and the results in Theorem* 5.1 remain valid.

2. Stetter [51] requires for his theory linear bounded operators P^h and $Q^h_{c,\Pi}:=(Q^h_c,Q^h_{c,b})$ in (6.15) and (6.24). But in fact he does need only equicontinuous operators and the limits of the norms listed above; see also (6.19), (6.23). The convergence properties in (6.9) are not required here, since usually 3. is impossible without them.

3. Condition 2. is correct if F or G are continuous in $B_r(P^hu_0)$ or $B_r(u_0)$.

4. The consistency, 3., will be verified in Theorem 7.1. Due to the careful preparations this is a relatively simple proof.

5. Then 4. is the only missing and hard condition in Theorem 5.1. In our approach the stability of the linearized problem and its regularity are the essential tools for the proof of stability; see Theorem 8.1. This makes sense only if the original F is continuously differentiable. Then F is Lipschitz continuous as well.

7. Consistency for the nonconforming FEM. Verifying the conditions of Theorem 5.1 we start with the easier consistency. Similarly to the Ciarlet, Raviart, Lenoir, and Bernadou approaches (see [26, 19, 42, 5, 6]) we have already and will furthermore use the extension operator, E_c (cf. Theorem 4.5). It allows estimates of

the interpolation errors on Ω and $\partial\Omega$, as in (4.23),(4.24), and thus are essential for the following proof.

THEOREM 7.1. *We require the conditions for Theorems* 4.2, 4.4, *mind* Ω *in* $C^p, 2 \le p \le 5$, *an exact solution* $u_0 \in H^\ell(\Omega), \ell > 2$, *a Lipschitz continuous* G^w *in* (3.6) *with a global constant* L,

$$G^w(\cdot) \in C_L\big(\ \overline{(\Omega_c^h \cup \Omega) \times \Gamma'}\ \big) \text{ with } \overline{(\Omega_c^h \cup \Omega) \times \Gamma'} := \{(x,z,p,r) : \tag{7.1}$$
$$x \in \overline{\Omega_c^h \cup \Omega}, (z,p,r) = \big((E_c u)(x), D(E_c u)(x), D^2(E_c u)(x)\big) \text{ with } u \in \overline{B_{r,H^2(\Omega)}(u_0)}\}\big],$$

and $d \ge n^2 + 1$. *Then* F^h *is consistent with* F *in* $P^h u_0$ *and*

$$\|F^h P^h u_0 - Q_\Pi^h F u_0\|_{\mathcal{V}_\Pi^{h\prime}} = \|\big(Q_c^h G(P^h u_0), -E_c u_0|_{\partial\Omega_c^h}\big)\|_{\mathcal{V}_\Pi^{h\prime}} \tag{7.2}$$
$$\le CLh^{\min\{\ell-2,p\}}\|u_0\|_{H^\ell(\Omega)}.$$

Remark 7.1. $d \ge n^2 + 1$ (see Theorem 4.4), and ℓ are related by (4.9). For inhomogeneous boundary conditions as in (6.3), the estimate (7.2) remains valid.

Proof. We estimate the first component in (7.2) with (6.19), L in (7.1), $G(u_0) \equiv 0$, and (4.23):

$$\|Q_c^h G(P^h u_0) - Q^h G(u_0)\|_{\mathcal{V}^{h\prime}} = \|Q_c^h G(P^h u_0) - 0\|_{\mathcal{V}^{h\prime}} \tag{7.3}$$
$$\le \|Q_c^h G(P^h u_0) - Q_c^h G(E_c u_0)\|_{\mathcal{V}^{h\prime}} + \|Q_c^h G(E_c u_0)\|_{\mathcal{V}^{h\prime}}$$
$$\le CL\|P^h u_0 - E_c u_0\|_{\mathcal{U}^h} + \|Q_c^h G(E_c u_0)\|_{\mathcal{V}^{h\prime}} \text{ by (4.18)}$$
$$\le CLh^{\ell-2}\|u_0\|_{H^\ell(\Omega)} + \|Q_c^h G(E_c u_0)\|_{\mathcal{V}^{h\prime}}.$$

We test $Q_c^h G(E_c u_0) - Q^h G(u_0)$ with $v^h \in \mathcal{V}^h$, combine it with (3.6), (6.19), and get

$$\big|\langle Q_c^h G(E_c u_0), v^h\rangle_{\mathcal{V}^{h\prime}\times\mathcal{V}^h}\big| = \Big|\int_{\Omega_c^h} \big[G(E_c u_0)(x)\big] v^h(x)dx\Big| \quad \forall v^h \in \mathcal{V}^h \tag{7.4}$$
$$\le \Big|\int_{\Omega_c^h} \big[G(E_c u_0)(x)\big] v^h(x)dx - \int_\Omega \big[0 \equiv G(u_0)(x)\big] E_c v^h(x)dx\Big|$$
$$\le \int_{\Omega_c^h\setminus\Omega} \Big|G(E_c u_0)(x) v^h(x)\Big|dx, \text{ since } \Big(\int_{\Omega_c^h} - \int_\Omega\Big)\Big|_{\Omega_c^h\cap\Omega} = 0,$$

where $G(E_c u_0), v^h \in L^2(\Omega_c^h)$ are well defined; cf. (6.11). By $\big[0 \equiv G(u_0)(x)\big] E_c v^h(x)$, we do not need to explicitly define the extension $E_c v^h$. Now (7.1) and $u_0 \in H^\ell(\Omega)$, $\ell > 2$, imply $|G(E_c u_0)(x)| \le CL\|u_0\|_{H^\ell(\Omega)}$. This is combined with $\operatorname{dist}(\Omega_c^h, \Omega) \le C' h^p$ (cf. (4.17)) and the bounded $S_{\partial\Omega} :=$ surface $(\partial\Omega) < \infty$ yielding

$$G(E_c u_0)(x) \text{ satisfies } G(E_c u_0) \equiv 0 \text{ in } \Omega \Longrightarrow \|G(E_c u_0)\|_{L^2(\Omega_c^h\setminus\Omega)} \tag{7.5}$$
$$\le CL\|u_0\|_{H^\ell(\Omega)}\|1\|_{L^2(\Omega_c^h\setminus\Omega)} \le CL\|u_0\|_{H^\ell(\Omega)} S_{\partial\Omega} C' h^p \le C^* h^p \|u_0\|_{H^\ell(\Omega)}.$$

We combine (7.3), (7.4), (7.5) with the Cauchy–Schwarz inequality

$$\|Q_c^h G(P^h u_0)\|_{\mathcal{V}^h} \le CLh^{\ell-2}\|u_0\|_{H^\ell(\Omega)} + \|Q_c^h G(E_c u_0)\|_{\mathcal{V}^h} \tag{7.6}$$
$$\le CLh^{\ell-2}\|u_0\|_{H^\ell(\Omega)} + C'' h^p \|u_0\|_{H^\ell(\Omega)} \le \hat{C} h^{\min\{\ell-2,p\}}\|u_0\|_{H^\ell(\Omega)};$$

thus we have estimated the first component in (7.2).

To simplify estimating the second component in (7.2), we assume $n < 2(\ell - 1)$. Then the Sobolev embedding theorem implies the compact embedding of $H^\ell(\Omega)$ into

$C^1(\overline{\Omega})$. Then $u_0 \in H^\ell(\Omega) \cap H^1_0(\Omega)$ implies $C(\partial\Omega) \ni u_0|_{\partial\Omega} \equiv 0$. Now choose any $P \in \partial\Omega^h_c$, a ray through $P \perp \partial\Omega^h_c$, the intersection point, P_b of this ray with $\partial\Omega$, and the directional derivative $\partial_P E_c u_0$ of $E_c u_0$ from P_b to P. Then

$$|E_c u_0(P)| = \Big|0 + \int_{P_b}^{P} \partial_P E_c u_0\Big| \le \|u_0\|_{H^\ell(\Omega)} C' h^p.$$

So we obtain for the second component in (7.2)

$$\|E_c u_0\|_{L^2(\partial\Omega^h_c)} \le C'' h^p \|u_0\|_{H^\ell(\Omega)}.$$

For inhomogeneous boundary conditions (cf. (6.3)), we employ the extension of ϕ, necessarily here $\phi \in H^{\ell-1/2}(\partial\Omega)$, to $\phi \in H^\ell(\Omega)$; we still have $G(u_0) = 0$, and so (7.3), (7.4) remain correct. Boundary conditions for the exact and the discrete problem are $(u_0 - \phi)|_{\partial\Omega} \equiv 0$ and $(u^h_0 - P^h\phi)|_{\partial\Omega^h_c} \equiv 0$ with $P^h\phi$ extended to Ω. We replace the exact boundary conditions by $(u_0 - P^h\phi)|_{\partial\Omega} \equiv 0$. Then the error estimates in Theorem 4.6 and the stability results for linear elliptic operators w.r.t. the right-hand side $(f, \phi) \in \mathcal{V} \times \mathcal{V}_D$, e.g., in Hackbusch [37, Theorem 9.1.16], yield again (7.2). ∎

8. Stability for the linearized operator and convergence. Fortunately, there is a standard result strongly simplifying the proof of stability; see, e.g., Stetter [51, Theorem 1.2.5], Stummel [52, 53], and Keller [40]: We need a simple corollary of Stetter's theorem, Theorem 1.2.5. It replaces the condition of a continuous and differentiable discretization, F^h, in $B_r(u^h)$ by the continuous differentiability in $B_r(u^h)$. This slightly enforced condition is usually satisfied by the equations considered here. It is even simplified in our case, since $F \in C^1(B_r(u))$ implies $F^h \in C^1(B_{r-\delta}(P^h u))$ with a small $\delta > 0$ for $h_0 > h \in H$.

THEOREM 8.1. *Let the nonlinear F be continuously differentiable in $B_r(u)$, and let the sequence of linearized $(F^h)'$ be stable at $P^h u$, satisfying*

$$\|\big((F^h)'(P^h u)\big)^{-1}\|_{\mathcal{U}^h \hookleftarrow \mathcal{V}^{h'}_\Pi} \le S \ \text{uniformly} \quad \forall h \in H, \quad h_0 > h. \tag{8.1}$$

Then the nonlinear F^h is stable at $P^h u$ as well with stability bound and threshold, $2S$ and $r_0 = r/(2S)$, respectively.

According to section 2, **3.**, Theorems 5.1, 7.1, and 8.1 and Remark 6.1 show that the only missing claim for convergence is the stability for the linearized operator for $u^h = P^h u$, here $u = u_0$. So we have to elaborate the program in section 2 in **3.** We indicate only the conforming FEM on convex polyhedral domains in (4.1) and mainly discuss the nonconforming FEM on C^2 domains in (4.2). For the conforming FEs it suffices to discuss the differential operator G. Then the boundary operators, B_D, the extension, E_c, and all the product machinery, in particular in subsections 8.3 and 8.4, can be avoided. This would considerably simplify the proofs, presented here only for the nonconforming FEMs. The interested reader easily can verify this claim. For our $F(u)$ with $F'(u_1)u = (G'(u_1)u, u|_{\partial\Omega})$ the linearization of G is essential. Thus a major part of the following discussion is concerned with $G'(u_1)$. We obtain the claimed stability, $\|\big((F^h)'(u^h_1)|_{\mathcal{U}^h \to \mathcal{V}^{h'}_\Pi}\big)^{-1}\|_{\mathcal{V}^{h'}_\Pi \hookleftarrow \mathcal{U}^h} \le C$, by the end of this section.

8.1. The linearized operator. We start with G assuming a continuously differentiable operator G or its defining function G^w; see (3.7), (3.8). Here we sometimes integrate the boundary conditions into the spaces $\mathcal{U}_0, \mathcal{W}_0$ and $\mathcal{U}^h_0, \mathcal{W}^h_0$. In the literature the same notation $G'(u_0) : \mathcal{U} \to \mathcal{V}$ and $G'(u_0) : \mathcal{W} \to \mathcal{W}'$ is used for the strong

and weak forms of the linearized operators, $G'_s(u_0) = G'(u_0)$ and $G'(u_0)$, respectively. For a smooth enough situation, the strong form is just the restriction of the weak form, and so $G'(u_0)|_{\mathcal{U}} = G'_s(u_0) : \mathcal{U} \to \mathcal{V}$ under the condition (8.2). The proofs in this section fundamentally depend upon a systematic interplay between these strong and weak forms of the linearized operators. To avoid misunderstandings, it is necessary to distinguish at least the strong and weak bilinear forms as $a_s(\cdot,\cdot)$ and $a(\cdot,\cdot)$. Later on, we even use the notations $A = G'(u_0), A_s = G'_s(u_0)$ and A^h, A^h_s for the weak and strong linear operators. Correspondingly, we formulate two different FEMs for the linearized problem. To its strong form, $a_s(\cdot,\cdot)$, we apply the FEM (6.12), (8.7), to its weak form, $a(\cdot,\cdot)$, the standard FEM; see (8.8) below. We will see that for the linear case both bilinear forms coincide, and so $a_s(u^h, v^h) = a(u^h, v^h)$ for all $u^h \in \mathcal{U}^h = \mathcal{W}^h \supset \mathcal{U}^h_0 = \mathcal{V}^h \ni v^h$; see (8.6) and Proposition 8.2.

We called the original nonlinear problem G *uniformly elliptic* in u_0 if its derivative $G'(u_0)u$ in (3.9) satisfies (3.10). In contrast to G, it is possible to transform the linearization in (3.9) into the standard weak form. We start with the strong operator $G'(u_0) : \mathcal{U} \to \mathcal{V}$ (see (3.2)) and introduce new coefficients a_{ij}. This yields the *strong bilinear form* $a_s(\cdot,\cdot)$ and conditions for the partials of G^w; see (3.2), (3.6), (3.7):

$$
\begin{aligned}
(8.2)\qquad & G'(u_0) : \mathcal{U} \to \mathcal{V} \text{ and } a_s(u,v) := (G'(u_0)u, v)_{\mathcal{V}} \quad \forall u \in \mathcal{U}, v \in \mathcal{V} \text{ with} \\
& G'(u_0)u = \sum_{i,j=0}^{n} (-1)_{j>0} \partial^j (a_{ij} \partial^i\ u) \text{ and } w_0(\cdot) = (\cdot, u_0(\cdot), Du_0(\cdot), D^2u_0(\cdot)) : \\
& a_{ij} = a_{ij}(u_0) = \frac{\partial G^w}{\partial r_{ij}}(w_0) \text{ for } i,j \geq 1, \quad a_{00} = a_{00}(u_0) = \frac{\partial G^w}{\partial z}(w_0), \\
& a_{i0} = a_{i0}(u_0) = \frac{\partial G^w}{\partial p_i}(w_0) + \sum_{j=1}^{n} \partial^j \frac{\partial G^w}{\partial r_{ij}}(w_0), \quad a_{0i} = 0 \text{ for } i \geq 1; \\
& \text{this strong form } G'(u_0) \text{ is bounded if } a_{ij} = a_{ij}(w_0(\cdot)) \in W^{1-\delta_{0j},\infty}(\Omega).
\end{aligned}
$$

Partial integration or the Green's formula now combined with $v \in \mathcal{W} \cap H^1_0(\Omega)$ yields the (standard) weak problem or the *coinciding bounded weak and strong bilinear form,* $a(\cdot,\cdot) : \mathcal{W} \times \mathcal{W} \to \mathbf{R}$, and the induced weak, $G'(u_0) : \mathcal{W} \to \mathcal{W}'$ (see (3.2), (3.3)), as

$$
\begin{aligned}
(8.3)\qquad & a(u,v) := \int_\Omega \sum_{i,j=0}^{n} a_{ij} \partial^i u \partial^j v dx\ =: \langle G'(u_0)u, v\rangle_{\mathcal{W}' \times \mathcal{W}} \quad \forall u \in \mathcal{W},\ \forall v \in \mathcal{W}_0, \\
& a(\cdot,\cdot), G'(u_0) \text{ bounded for } a_{ij} = a_{ij}(u_0) = a_{ij}(w_0(\cdot)) \in L^\infty(\Omega) \text{ for } i,j = 0,\dots,n, \\
& \text{and with } a_s(u,v) = a(u,v) \quad \forall u \in \mathcal{U},\ \forall v \in \mathcal{W}_0.
\end{aligned}
$$

The principal part $a_p(u,v)$ of $a(u,v)$ is $\mathcal{W}_0$-*coercive*, by (3.10), and, since $\|u\|_{H^1_0(\Omega)}$ and $\|u\|_{\mathcal{W}}$ are equivalent norms on $\mathcal{W}_0$,

$$
(8.4)\qquad a_p(u,u) \geq \lambda \int_\Omega \sum_{i=1}^{n} (\partial^i u)^2 dx = \lambda \|u\|^2_{H^1_0(\Omega)} \geq \alpha \|u\|^2_{\mathcal{W}} \quad \forall u \in \mathcal{W}_0.
$$

The original $a(u,v)$, inducing $A = G'(u_0)$, is the sum of $a_p(u,v)$, inducing B, and its complement $c(u,v)$, inducing C, a compact perturbation of $A = B + C$, with $A, B, C \in \mathcal{L}(\mathcal{W}, \mathcal{W}')$:

$$
(8.5)\qquad a(u,v) = a_p(u,v) + c(u,v) = \langle G'(u_0)u, v\rangle_{\mathcal{W}' \times \mathcal{W}} = \langle (B+C)u, v\rangle_{\mathcal{W}' \times \mathcal{W}}.
$$

B is boundedly invertible on $\mathcal{W}_0$ by (8.4). Thus $G'(u_0)$ satisfies the *Fredholm alternative* on $\mathcal{W}_0$. This compact perturbation $C \in \mathcal{L}(\mathcal{W}, \mathcal{W}')$ will play a dominant role for the following stability results.

In addition to the above approximating spaces (see (3.1), (4.4)), we need for $\mathcal{W}_0$

$$\mathcal{U}^h = \mathcal{W}^h \supset \mathcal{V}^h \text{ and } \mathcal{U}_0^h = \mathcal{W}_0^h := \mathcal{V}^h = \mathcal{U}^h \cap H_0^1(\Omega_c^h) \quad \text{for } \mathcal{U}_0 \subset \mathcal{W}_0, \tag{8.6}$$

again with the corresponding nonequivalent norms $\|\cdot\|_{\mathcal{U}^h}, \|\cdot\|_{\mathcal{W}^h}, \|\cdot\|_{\mathcal{V}^h}$.

We want to contrast the two different FEMs for the linear problem. We start with (6.12) and then turn to the standard FEM based upon the *weak bilinear form* $a(\cdot,\cdot)$. In both cases we have to determine $u_1^h \in \mathcal{U}_0^h$ and $u_1^h \in \mathcal{W}_0^h$ s.t.

$$\begin{aligned} u_1^h \in \mathcal{U}_0^h : a_s(u_1^h, v^h) &= \big(G'(P^h u_0)u_1^h, v^h\big)_{\mathcal{V}_c} := \Big(\sum_{i,j=0}^n (-1)_{j>0}\partial^j(a_{ij}(P^h u_0)\partial^i\ u_1^h), v^h\Big)_{\mathcal{V}_c} \\ &= (E_c f, v^h)_{\mathcal{V}_c} \quad \forall v^h \in \mathcal{V}^h, \ \text{ with } f \in \mathcal{V}, a_{ij} \in W^{1-\delta_{0j},\infty}(\Omega); \text{ see (8.2).} \end{aligned} \tag{8.7}$$

Similarly as G^w we assume the $a_{ij}(P^h u_0)$ to be well defined on the extended $\Omega \cup \Omega_c^h$. The standard FEM for the *weak bilinear form* $a(\cdot,\cdot)$ with its solution $u_1^h \in \mathcal{W}_0^h$ is

$$\begin{aligned} u_1^h \in \mathcal{W}_0^h : a(u_1^h, v^h) &= \int_{\Omega_c^h} \sum_{i,j=0}^n a_{ij}(P^h u_0)\partial^i u_1^h \partial^j v^h dx \ = \langle G'(P^h u_0)u_1^h, v^h\rangle_{\mathcal{W}_c' \times \mathcal{W}_c} \\ &= \langle E_c f, v^h\rangle_{\mathcal{W}_c' \times \mathcal{W}_c} \text{ or } = \langle E_c f, v^h\rangle_{\mathcal{V}_c} \text{ for } f \in \mathcal{V} \ \ \forall v^h \in \mathcal{W}_0^h, \text{ with } a_{ij} \in L^\infty(\Omega). \end{aligned} \tag{8.8}$$

As in (6.12) for G, we use here the same notation for the discrete $a_s(u^h, v^h), a(u^h, v^h)$ and the original $a_s(u,v), a(u,v)$. These bilinear forms are, for small enough h, simultaneously bounded and elliptic. Equations (8.7), (8.8) can and will be similarly reinterpreted by using the above projectors and defining new ones.

8.2. Weak and strong linear operators coincide on $\mathcal{U}^h \times \mathcal{V}^h$. The techniques in this and subsection 8.4 are strongly related to the standard approach in spectral methods for collocation (see Canuto et al. [17]) and to regularity for difference equations; see, e.g., Hackbusch [37, Theorem 9.2.26].

PROPOSITION 8.2. *Choose the above $\mathcal{U}^h, \dots, \subset \mathcal{S}_d^1(\mathcal{T}_c^h)$ in (8.6) for $d \geq 2^n + 1$ as approximating spaces for $\mathcal{U}, \dots$, and let $a_{ij} \in W^{1-\delta_{0j},\infty}(\Omega)$. Then for all $u^h \in \mathcal{U}^h, v^h \in \mathcal{V}^h$,*

$$\begin{aligned} a_s(u^h, v^h) &= (G'(P^h u_0)u^h, v^h)_{\mathcal{V}_c} \\ = a(u^h, v^h) &= \langle G'(P^h u_0)u^h, v^h\rangle_{\mathcal{W}_c' \times \mathcal{W}_c} \quad \forall u^h \in \mathcal{U}^h = \mathcal{W}^h, \quad v^h \in \mathcal{V}^h = \mathcal{U}_0^h. \end{aligned} \tag{8.9}$$

Proof. We apply partial integration for every $T \in \mathcal{T}_c^h$ with $\overline{\Omega_c^h} = \cup_{T \in \mathcal{T}_c^h} \overline{T}$. We use the notations $\nu = \nu_T$ for the outer unit normal vector of T and B_a and u for the boundary operator, induced by $A = G'(u_0)$; see, e.g., [37]:

$$B_a\ u = \sum_{i,j=1}^n \nu_j a_{ij}\partial^i\ u + \sum_{j=1}^n \nu_j a_{0j}\ u, \text{ e.g., } B_a\ u = \partial u/\partial\nu \quad \text{for } A_s u = -\Delta u. \tag{8.10}$$

All ∂T have a *positive orientation* w.r.t. T. Hence in (8.12) every interior edge $e \in \overline{T}$ will be obtained twice in opposite directions. Edges $e \subset \partial\Omega_c^h$ will appear once. Let $\nu = \nu_T$ be one of the normal vectors for an interior edge $e \subset \overline{T_r}$. It is oppositely

oriented for neighboring $T_r\ , T_l \in \mathcal{T}_c^h$ with $e \subset \overline{T_r} \cap \overline{T_l}$ and ν the outer normal[4] for $e \subset \partial\Omega_c^h$. To consider the transition from a triangle T_l to its neighboring T_r, we introduce the restriction of v to $\overline{T_l}, \overline{T_r}, \partial\Omega_c^h$ and the standard notation; see [61]:

$$v_l = v|_{\overline{T_l}}, \quad v_r = v|_{\overline{T_r}}, \quad [v] := v_l|_e - v_r|_e, \quad \text{and } \{v\} := (v_l|_e + v_r|_e)/2, \tag{8.11}$$

for the corresponding jumps and arithmetic means of v across an interior e, and

$$[v] := \{v\} := v|_{\partial\Omega_c^h} \text{ along } e \in \partial\Omega_c^h, \quad \text{and} \quad v \ \text{ arbitrary in } \ \mathbf{R}^n \backslash \Omega_c^h.$$

The next transformation via the n-dimensional Green's formula is the deciding reason for the of C^1 FEs. Only for this choice do the following jump contributions, the $\big[B_a u^h\big] = 0$, vanish (the $[v^h]$ vanish already for C FEs v^h):

$$\begin{aligned} a(u^h, v^h) = \langle G'(E_c u_0) u^h, v^h \rangle_{\mathcal{W}_c' \times \mathcal{W}_c} \ &= \ \sum_{T \in \mathcal{T}_c^h} \Big(\int_T \sum_{i,j=0}^{n} a_{ij} \partial^i \ u^h \partial^j \ v^h dx \Big) \\ &= \sum_{T \in \mathcal{T}_c^h} \Big(\int_T \sum_{i,j=0}^{n} (-1)_{j>0} \partial^j (a_{ij} \partial^i \ u^h) v^h dx \Big) + \int_{\partial\Omega_c^h} v_l^h B_a u^h \, ds \\ + \sum_{e \in \mathcal{T}_c^h \backslash \partial\Omega_c^h} \int_e \big(\{v^h\} \big[B_a u^h\big] + [v^h] \{B_a u^h\} \big) ds &= \int_{\Omega_c^h} (G'(E_c u_0)_s \ u^h) v^h dx \end{aligned} \tag{8.12}$$

$$= a_s(u^h, v^h) \text{ since } u^h \in \mathcal{U}^h \subset C^1(\Omega_c^h), v^h \in \mathcal{V}^h \subset H_0^1(\Omega_c^h); \text{ hence,}$$
$$a(u^h, v^h) = a_s(u^h, v^h) \quad \forall u^h \in \mathcal{U}^h, v^h \in \mathcal{V}^h.$$

Discontinuous Galerkin methods would certainly be possible for $a(u^h, v^h)$. But they would violate the necessary $a(u^h, v^h) = a_s(u^h, v^h)$ by their penalty terms. ∎

8.3. Stability of the weak linear operator. The previous subsections play the role of preparing this stability proof. We essentially had confined the discussion there to the differential equation and indicated the boundary conditions either by the index $_0$ or using $v^h \in \mathcal{V}^h$. Instead of the previous strong and weak forms, $A_s : \mathcal{U} \to \mathcal{V}$ and $A : \mathcal{W} \to \mathcal{W}'$ (see (6.2), (6.5)–(6.7)), we need their two component versions:

$$\begin{aligned} &A_{\mathcal{U}} := (A_s, B_D) : \mathcal{U} \to \mathcal{V} \times \mathcal{V}_D \subset \hat{\mathcal{V}}_\Pi = \mathcal{V} \times \mathcal{V}_b, \ A_{\mathcal{U}} \in \mathcal{L}(\mathcal{U}, \mathcal{V} \times \mathcal{V}_D) \text{ and} \\ &A_{\mathcal{W}} := (A, B_D) : \mathcal{W} \to \mathcal{W}' \times \mathcal{W}_D \subset \mathcal{W}_\Pi' = \mathcal{W}' \times \mathcal{V}_b, \ A_{\mathcal{W}} \in \mathcal{L}(\mathcal{W}, \mathcal{W}' \times \mathcal{W}_D). \end{aligned} \tag{8.13}$$

With the dense (compact) embedding $\mathcal{W}_D \hookrightarrow \mathcal{V}_b = L^2(\partial\Omega)$ we test the $\mathcal{W}' \times \mathcal{W}_D \subset \mathcal{W}_\Pi'$ by the following $\mathcal{W}_\Pi$. We emphasize that the stable splitting $S^h = S_0^h + S_b^h$ in Theorem 4.4 plays a dominant role. We choose (cf. (4.18), (4.19), (6.16)–(6.18))

$$\begin{aligned} &\mathcal{W} \times \mathcal{W}_D \subset \mathcal{W}_\Pi := \mathcal{W} \times \mathcal{V}_b \neq \mathcal{W}_\Pi' = \mathcal{W}' \times \mathcal{V}_b, \text{ with } \mathcal{V}_b \supset \mathcal{W}_D \text{ and} \\ &\qquad \langle (u', u_b), (v, v_b) \rangle_{\mathcal{W}_\Pi' \times \mathcal{W}_\Pi} := \langle u', v \rangle_{\mathcal{W}' \times \mathcal{W}} + (u_b, v_b)_{\mathcal{V}_b}, \text{ and on} \\ &\Omega_c^h \text{ the } \mathcal{W}_{c,\Pi} = \mathcal{W}_c \times \mathcal{V}_{c,b} \neq \mathcal{W}_{c,\Pi}' \text{ with } \langle (u', u_b), (v, v_b) \rangle_{\mathcal{W}_{c,\Pi}' \times \mathcal{W}_{c,\Pi}}. \end{aligned} \tag{8.14}$$

The discrete counterparts are the

$$\begin{aligned} &\mathcal{W}_\Pi^h = \mathcal{W}^h \times \mathcal{V}_b^h := S^h \times S_b^h|_{\partial\Omega_c^h} \neq \mathcal{V}_\Pi^h, \quad \mathcal{W}_\Pi^{h'} := \mathcal{W}^{h'} \times \mathcal{V}_b^h := S^{h'} \times S_b^h|_{\partial\Omega_c^h} \neq \mathcal{V}_\Pi^{h'}, \\ &\text{and } \langle (u^{h'}, u_b^h), (v^h, v_b^h) \rangle_{\mathcal{W}_\Pi^{h'} \times \mathcal{W}_\Pi^h} := \langle u^{h'}, v^h \rangle_{\mathcal{W}^{h'} \times \mathcal{W}^h} + (u_b^h, v_b^h)_{\mathcal{V}_b^h}. \end{aligned} \tag{8.15}$$

[4]We will use the notation $e \in \partial\Omega_c^h$, although mostly $e \subset \partial\Omega_c^h$; similarly, $e \in \mathcal{T}_c^h \setminus \partial\Omega_c^h$.

$$\begin{array}{ccccc} \mathcal{W} & \xrightarrow{A_{\mathcal{W}}} & \mathcal{W}'_\Pi = \mathcal{W}' \times \mathcal{W}_D & \overset{tested\ by}{\longleftrightarrow} & \mathcal{W}_\Pi = \mathcal{W} \times \mathcal{V}_b \\ P^h \big\downarrow & \Phi^h_{\mathcal{W}} \big\downarrow & Q^{h'}_\Pi \big\downarrow & & \\ \mathcal{W}^h & \xrightarrow{A^h_{\mathcal{W}}} & \mathcal{W}^{h'}_\Pi = \mathcal{W}^{h'} \times \mathcal{W}^h_D & \overset{tested\ by}{\longleftrightarrow} & \mathcal{W}^h_\Pi = \mathcal{W}^h \times \mathcal{V}^h_b \end{array}$$

FIG. 8.1. *Nonconforming FEMs for linear problems.*

The exact and discrete linear boundary conditions remain nearly unchanged, except replacing $\mathcal{V}_D$ by $\mathcal{W}_D$, and so they can be treated as in (6.4), (6.13); see (6.22), (6.23).

For the weak differential operator, we introduce a modified Ritz operator. The above extension operator E_c in Theorem 4.5 is continuously extended to $f \in H^k(\Omega), k < 0$. We define the projectors, $Q^{h'}, Q^{h'}_c$, and so on, indicating the weak form by the exponent $^{h'}$, in contrast to the strong projectors Q^h and so on with the exponent ^{h} in (6.24); cf. (4.11), (4.12), (4.18), (4.19):

(8.16)

$$\begin{aligned} &Q^{h'} \in \mathcal{L}(\mathcal{W}', \mathcal{W}^{h'}) \quad \text{for } f \in \mathcal{W}' \text{ by } \langle Q^{h'} f, v^h\rangle_{\mathcal{W}^{h'}\times\mathcal{W}^h} - \langle (E_c f), v^h\rangle_{\mathcal{W}'_c\times\mathcal{W}_c} = 0, \\ &Q^{h'}_c \in \mathcal{L}(\mathcal{W}'_c, \mathcal{W}^{h'}) : f_c \in \mathcal{W}'_c : \ \langle Q^{h'}_c f_c, v^h\rangle_{\mathcal{W}^{h'}\times\mathcal{W}^h} - \langle f_c, v^h\rangle_{\mathcal{W}'_c\times\mathcal{W}_c} = 0 \quad \forall v^h \in \mathcal{W}^h. \end{aligned}$$

This allows formulating the product projectors, and we obtain

$$Q^{h'}_\Pi := (Q^{h'}, Q^h_b) \in \mathcal{L}\big(\mathcal{W}'_\Pi, \mathcal{W}^{h'}_\Pi\big), \quad Q^{h'}_{c,\Pi} := (Q^{h'}_c, Q^h_{c,b}) \in \mathcal{L}\big(\mathcal{W}'_{c,\Pi}, \mathcal{W}^{h'}_\Pi\big). \tag{8.17}$$

For both types of projectors we obtain the necessary norm limits:

$$\begin{aligned} &\lim_{h\to 0} \|Q^{h'} f\|_{\mathcal{W}^{h'}} = \|f\|_{\mathcal{W}'}, \text{and } \lim_{h\to 0} \|Q^{h'}_c f_c\|_{\mathcal{W}^{h'}} = \lim_{h\to 0} \|f_c\|_{\mathcal{W}'_c} = \lim_{h\to 0} \|E_c f_c\|_{\mathcal{W}'}, \\ &\lim_{h\to 0} \|Q^{h'}_\Pi (f, u_1|_{\partial\Omega})\|_{\mathcal{W}^{h'}_\Pi} = \|(f, u_1|_{\partial\Omega})\|_{\mathcal{W}'_\Pi}, \text{ and } \lim_{h\to 0} \|Q^{h'}_{c,\Pi}(f_c, (E_c u)|_{\partial\Omega^h_c})\|_{\mathcal{W}^{h'}_\Pi} \\ &\qquad = \lim_{h\to 0} \|(f_c, (E_c u)|_{\partial\Omega^h_c})\|_{\mathcal{W}'_{c,\Pi}} = \lim_{h\to 0} \|(E_c f_c, (E_c u)|_{\partial\Omega})\|_{\mathcal{W}'_\Pi}. \end{aligned} \tag{8.18}$$

Similarly to the above $F^h = Q^h_{c,\Pi} F|_{\mathcal{U}^h}$ we use the projectors for reformulating (8.7) and (8.8) (cf. (8.13)) for $a_{ij} \in C_L(\Omega \cup \Omega^h_c)$. For the strong $A_{\mathcal{U}}$ and $A^h_{\mathcal{U}}$ we get

$$A^h_{\mathcal{U}} = Q^h_{c,\Pi} A_{\mathcal{U}}|_{\mathcal{U}^h} : u^h_0 \in \mathcal{U}^h : A^h_{\mathcal{U}} u^h_0 = Q^h_{c,\Pi}(A_s u^h_0, u^h_0|_{\partial\Omega^h_c}) = Q^h_\Pi(f, \phi). \tag{8.19}$$

Similarly, the weak linear operators $A_{\mathcal{W}}$ and $A^h_{\mathcal{W}}$ yield, for $(f, \phi) \in \mathcal{W}'_\Pi$,

$$A^h_{\mathcal{W}} = Q^{h'}_{c,\Pi} A_{\mathcal{W}}|_{\mathcal{W}^h} : u^h_0 \in \mathcal{W}^h : A^h_{\mathcal{W}} u^h_0 = Q^{h'}_{c,\Pi}(A u^h_0, u^h_0|_{\partial\Omega^h_c}) = Q^{h'}_\Pi(f, \phi). \tag{8.20}$$

We summarize the new situation with uniformly bounded $A^h, P^h, Q^{h'}_{c,\Pi}$ in Figure 8.1. For proving the stability for $A^h_{\mathcal{W}}$ in Theorem 8.4 assume the principal part induces the ($\mathcal{W}_0$-coercive) operator B; cf. (8.5). We use the notations

$$A_{\mathcal{W}} = (A, B_D),\ B_{\mathcal{W}} = (B, B_D),\ C_{\mathcal{W}} = (C, B_D),\ C \text{ compact},\ A^h_{\mathcal{W}}, B^h_{\mathcal{W}}, C^h_{\mathcal{W}}. \tag{8.21}$$

THEOREM 8.3 (stability and convergence for the principal part). *Assume* (8.3), *the ellipticity* (8.4), *the approximate spaces in* (8.6) *with the norm limits in* (8.18) *and* (6.24), *and* $B_{\mathcal{W}}$ *and* $B^h_{\mathcal{W}}$ *as in* (8.13), (8.20). *Then* $B_{\mathcal{W}}$ *is boundedly invertible,*

and $B^h_{\mathcal{W}}$ is stable. Then unique solutions u_1 of $B_{\mathcal{W}}u_1 = (f,\phi) \in \mathcal{W}'_\Pi$ and u_1^h of $B^h_{\mathcal{W}}u_1^h = Q^{h'}_\Pi(f,\phi) \in \mathcal{W}^{h'}_\Pi$ exist and u_1^h converge s.t. $\|u_1^h - E_c u_1\|_{\mathcal{W}^h} \to 0$ for $h \to 0$.

Proof. Equation (8.3) shows that $a_p^h(u^h, v^h)$ is bounded on $\mathcal{W}^h$. The ellipticity condition (8.4) implies with the usual arguments (cf. (8.4)) that

$$a_p^h(u^h,u^h) \geq \lambda \int_{\Omega_c^h} \sum_{i=1}^n (\partial^i u^h)^2 dx = \lambda \|u^h\|^2_{H^1_0(\Omega_c^h)} \geq \alpha \|u^h\|^2_{\mathcal{W}^h} \quad \forall u^h \in \mathcal{W}^h_0, \tag{8.22}$$

hence the stability of $B^h_{\mathcal{W}}$ in $\mathcal{W}^h_0$. We reduce the inhomogeneous problem

$$B^h_{\mathcal{W}}u_1^h = Q^{h'}_\Pi(f,\phi) = (Q^{h'}f \in \mathcal{W}^{h'}, Q_b^h\phi \in \mathcal{V}_b^h) \in \mathcal{W}^{h'}_\Pi$$

to $\mathcal{W}^h_0$: As usual, we determine $u_2^h \in \mathcal{W}^h$ with $(u_2^h - Q_b^h\phi)|_{\partial\Omega_c^h} = 0 \in \mathcal{V}_b^h$. Then

$$B^h_{\mathcal{W}}(u_1^h - u_2^h) = (Q^{h'}f - B^h u_2^h \in \mathcal{W}^{h'}, 0|_{\partial\Omega_c^h} \in \mathcal{V}_b^h) \in \mathcal{W}^{h'}_\Pi.$$

This guarantees, by (8.22) and Theorems 5.1 and 7.1, the stability, consistency, and hence existence of the unique discrete solution $u_1^h - u_2^h$ and its convergence to the corresponding exact solution $u_1 - u_2$, and hence of u_1^h as well. ∎

THEOREM 8.4 (stability of $A^h_{\mathcal{W}} = (A = G'(u_0), B_D)^h \in \mathcal{L}(\mathcal{W}^h, \mathcal{W}^{h'} \times \mathcal{V}_b^h)$). *Under the conditions of Theorem* 8.3*, assume a boundedly invertible $A_{\mathcal{W}}$ in* (8.13)*. Then $A^h_{\mathcal{W}}$ in* (8.20) *is stable.*

Proof. This is a generalization to nonconforming FEMs of the Böhmer and Sassmannshausen results (see [47, 9, 8]) for conforming FEMs. For an arbitrary $u \in \mathcal{W}$ choose $v_w := C_{\mathcal{W}}u \in \mathcal{W}'_\Pi = \mathcal{W}' \times \mathcal{W}_D$. By assumption, unique exact and discrete solutions, $\hat{u}$ and $\hat{u}^h$, exist for the equations

$$B_{\mathcal{W}}\hat{u} = v_w \text{ and } B^h_{\mathcal{W}}\hat{u}^h = Q^{h'}_{c,\Pi}B_{\mathcal{W}}|_{\mathcal{W}^h}\hat{u}^h = Q^{h'}_\Pi B_{\mathcal{W}}\hat{u} = Q^{h'}_\Pi v_w \in \mathcal{W}^{h'}_\Pi. \tag{8.23}$$

With the notations

$$T_w := B^{-1}_{\mathcal{W}} \in \mathcal{L}(\mathcal{W}'_\Pi, \mathcal{W}) \text{ and } T_w^h := (B^h_{\mathcal{W}})^{-1}Q^{h'}_\Pi \in \mathcal{L}(\mathcal{W}'_\Pi, \mathcal{W}^h).$$

Theorem 8.3 implies that

$$\|E_c\hat{u} - \hat{u}^h\|_{\mathcal{W}^h} \;=\; \|(E_cT_w - T_w^h)C_{\mathcal{W}}u\|_{\mathcal{W}^h} \to 0 \quad \text{for } h \to 0 \text{ and } \forall u \in \mathcal{W}.$$

Since, by assumption, $C_{\mathcal{W}}$ is compact and $(E_cT_w - T_w^h)$ are equibounded, this implies

$$\|(E_cT_w - T_w^h)C_{\mathcal{W}}\|_{\mathcal{W}\hookleftarrow\mathcal{W}} \to 0 \quad \text{for } h \to 0. \tag{8.24}$$

Now let $u^h \in \mathcal{W}^h \subset \mathcal{W}_c$. Because $A_{\mathcal{W}}$ has been extended to yield

$$A_{\mathcal{W}} : \mathcal{W}_c \to \mathcal{W}'_{c,\Pi} = \mathcal{W}'_c \times \mathcal{W}_{c,D} \text{ still boundedly invertible with } \mathcal{W}^h \subset \mathcal{W}_c, \tag{8.25}$$

we can estimate

$$\begin{aligned}\|u^h\|_{\mathcal{W}^h} &\leq \|A^{-1}_{\mathcal{W}}\|_{\mathcal{W}^h\hookleftarrow\mathcal{W}'_{c,\Pi}}\|A_{\mathcal{W}}u^h\|_{\mathcal{W}'_{c,\Pi}} = \|A^{-1}_{\mathcal{W}}\|_{\mathcal{W}^h\hookleftarrow\mathcal{W}'_{c,\Pi}}\|B_{\mathcal{W}}(I + T_wC_{\mathcal{W}})u^h\|_{\mathcal{W}'_{c,\Pi}} \\ &\leq \|A^{-1}_{\mathcal{W}}\|_{\mathcal{W}^h\hookleftarrow\mathcal{W}'_{c,\Pi}}\|B_{\mathcal{W}}\|_{\mathcal{W}'_{c,\Pi}\hookleftarrow\mathcal{W}^h}\|(I + T_wC_{\mathcal{W}})u^h\|_{\mathcal{W}^h}; \text{ hence,} \\ \|(I + T_wC_{\mathcal{W}})u^h\|_{\mathcal{W}^h} &\geq (\|A^{-1}_{\mathcal{W}}\|_{\mathcal{W}^h\hookleftarrow\mathcal{W}'_{c,\Pi}}\|B_{\mathcal{W}}\|_{\mathcal{W}'_{c,\Pi}\hookleftarrow\mathcal{W}^h})^{-1}\|u^h\|_{\mathcal{W}^h}.\end{aligned} \tag{8.26}$$

The $Q^{h'}_{\Pi} B_{\mathcal{W}} u$ is defined for every $u \in \mathcal{W}$. We apply the stability of $B^h_{\mathcal{W}}$ to $w^h := B^h_{\mathcal{W}} u^h$:

$$\|u^h\|_{\mathcal{W}^h} \leq \|(B^h_{\mathcal{W}})^{-1}\|_{\mathcal{W}^h \hookleftarrow \mathcal{W}^{h'}_{\Pi}} \cdot \|w^h\|_{\mathcal{W}^{h'}_{\Pi}}.$$

Furthermore

$$\begin{aligned}\|(I + T^h_w C_{\mathcal{W}})u^h\|_{\mathcal{W}^h} &= \|(B^h_{\mathcal{W}})^{-1} B^h_{\mathcal{W}}(I + T^h_w C_{\mathcal{W}})u^h\|_{\mathcal{W}^h} \\ &\leq \|(B^h_{\mathcal{W}})^{-1}\|_{\mathcal{W}^h \hookleftarrow \mathcal{W}^{h'}_{\Pi}} \|B^h_{\mathcal{W}}(I + T^h_w C_{\mathcal{W}})u^h\|_{\mathcal{W}^{h'}_{\Pi}}\end{aligned}$$

implies

$$\|B^h_{\mathcal{W}}(I + T^h_w C_{\mathcal{W}})u^h\|_{\mathcal{W}^{h'}_{\Pi}} \geq \frac{1}{\|(B^h_{\mathcal{W}})^{-1}\|_{\mathcal{W}^h \hookleftarrow \mathcal{W}^{h'}_{\Pi}}} \cdot \|(I + T^h_w C_{\mathcal{W}})u^h\|_{\mathcal{W}^h}. \tag{8.27}$$

We combine (8.23)–(8.27) and use the fact that $A_{\mathcal{W}}, B_{\mathcal{W}}, C_{\mathcal{W}}$ are extended to $\mathcal{W}^h$, s.t. $Q^{h'}_{c,\Pi} A_{\mathcal{W}} u^h$ is well defined for estimating

$$\begin{aligned}\|A^h_{\mathcal{W}} u^h\|_{\mathcal{W}^{h'}_{\Pi}} &= \|Q^{h'}_{c,\Pi} A_{\mathcal{W}}|_{\mathcal{W}^h} u^h\|_{\mathcal{W}^{h'}_{\Pi}} = \|Q^{h'}_{c,\Pi} A_{\mathcal{W}} u^h\|_{\mathcal{W}^{h'}_{\Pi}} \quad \forall u^h \in \mathcal{W}^h \\ &= \|Q^{h'}_{c,\Pi} B_{\mathcal{W}}(I + T_w C_{\mathcal{W}})u^h\|_{\mathcal{W}^{h'}_{\Pi}} \overset{(8.23)}{=} \|B^h_{\mathcal{W}}(I + T_w C_{\mathcal{W}})u^h\|_{\mathcal{W}^{h'}_{\Pi}} \\ &\geq \|B^h_{\mathcal{W}}(I + T^h_w C_{\mathcal{W}})u^h\|_{\mathcal{W}^{h'}} - \|B^h_{\mathcal{W}}(E_c T_w - T^h_w) C_{\mathcal{W}} u^h\|_{\mathcal{W}^{h'}_{\Pi}} \\ &\overset{(8.27)}{\geq} \|(I + T^h_w C_{\mathcal{W}})u^h\|_{\mathcal{W}^h} / \|(B^h_{\mathcal{W}})^{-1}\|_{\mathcal{W}^h \hookleftarrow \mathcal{W}^{h'}_{\Pi}} - \|B^h_{\mathcal{W}}(E_c T_w - T^h_w) C_{\mathcal{W}} u^h\|_{\mathcal{W}^{h'}_{\Pi}} \\ &\geq (\|(I + T_w C_{\mathcal{W}})u^h\|_{\mathcal{W}^h} - \|(E_c T_w - T^h_w) C_{\mathcal{W}} u^h\|_{\mathcal{W}^h}) / \|(B^h_{\mathcal{W}})^{-1}\|_{\mathcal{W}^h \hookleftarrow \mathcal{W}^{h'}_{\Pi}} \\ &\quad - \|B^h_{\mathcal{W}}(E_c T_w - T^h_w) C_{\mathcal{W}}\|_{\mathcal{W}^{h'}_{\Pi} \hookleftarrow \mathcal{W}^h} \|u^h\|_{\mathcal{W}^h} \\ &\overset{(8.26)}{\geq} \Big(\frac{1}{(\|A^{-1}_{\mathcal{W}}\|_{\mathcal{W}^h \hookleftarrow \mathcal{W}'_{c,\Pi}} \|B^h_{\mathcal{W}}\|_{\mathcal{W}^{h'}_{\Pi} \hookleftarrow \mathcal{W}^h} \|(B^h_{\mathcal{W}})^{-1}\|_{\mathcal{W}^h \hookleftarrow \mathcal{W}^{h'}_{\Pi}})} \\ &\quad - \mathcal{O}(\|(E_c T_w - T^h_w) C_{\mathcal{W}}\|_{\mathcal{W}^{h'}_{\Pi} \hookleftarrow \mathcal{W}^h}) \Big) \|u^h\|_{\mathcal{W}^h}.\end{aligned}$$

The last estimate shows the following: The $\|A^{-1}_{\mathcal{W}}\|_{\mathcal{W}^h \hookleftarrow \mathcal{W}'_{c,\Pi}}$ is the deciding factor if $B^h_{\mathcal{W}}$ has a moderate condition number. Because of (8.24) and the stability of $(B^h_{\mathcal{W}})_{h \in H}$ there exists a positive constant K, independent of h, such that for all $h \leq h_0$,

$$\|A^h_{\mathcal{W}} u^h\|_{\mathcal{W}^h} = \|Q^{h'}_{c,\Pi} A_{\mathcal{W}} u^h\|_{\mathcal{W}^h} \geq K \|u^h\|_{\mathcal{W}^h} \quad \forall u^h \in \mathcal{W}^h.$$

This and $\dim \mathcal{W}^h < \infty$ show that $Q^{h'}_{c,\Pi} A_{\mathcal{W}}|_{\mathcal{W}^h}$ is invertible for $h \leq h_0$. Moreover, we obtain $\|(Q^{h'}_{c,\Pi} A_{\mathcal{W}}|_{\mathcal{W}^h})^{-1}\|_{\mathcal{W}^h \hookleftarrow \mathcal{W}^{h'}} \leq 1/K$; i.e., $A^h_{\mathcal{W}}|_{\mathcal{W}^h}$ for all $h \in H$ is stable. □

8.4. Regularity for FE solutions. Our next task is lifting the preceding stability result w.r.t. the weak forms back to the strong form. This is achieved by the regularity for FE solutions in this subsection. Until now, *regularity results for the solutions of FEMs* seem to be known only as oscillation results in the sense of De Giorgi, Nash, and Moser; see Aguilera and Caffarelli [1]. They are not applicable to our problem of stability. Here we need, for the solution of a FEM applied to a differential equation of order 2, results analogous to those for the exact solution. These are estimates of the $H^2(\Omega)$ instead of the usual $H^1(\Omega)$ norm for a right-hand

side in $L^2(\Omega) \times H^{3/2}(\Omega)$ instead of the usual $H^{-1}(\Omega) \times H^{1/2}(\Omega)$. To the author's knowledge, the following type of lemma was not known for FEMs. It is a transformation to our FEM of Hackbusch's regularity result [37, Theorem 9.2.26] for difference methods. We employ inverse and interpolation error estimates (4.5), (4.11), (4.23), valid under the conditions of Theorems 4.1, 4.2, and 4.6. The lemma is based upon a regularity result for the exact solution u_1 of the corresponding linear problem $A_{\mathcal{W}}u_1 = (f,\phi) \in \mathcal{V} \times \mathcal{V}_D \subset \hat{\mathcal{V}}_\Pi, u_1 \in \mathcal{W}_0$. Indeed, known regularity results, e.g., in Hackbusch [37, Theorems 9.1.22, 9.1.16], show that, for a convex bounded or smooth domain and smooth coefficients, the restriction $A_{\mathcal{U}} = A_{\mathcal{W}}|_{\mathcal{U}} \in \mathcal{L}(\mathcal{U}, \mathcal{V} \times \mathcal{V}_D)$ is boundedly invertible for a boundedly invertible $A_{\mathcal{W}} \in \mathcal{L}(\mathcal{W}, \mathcal{W}' \times \mathcal{W}_D)$; hence we assume

$$\Omega \text{ convex bounded, polyhedral or } \Omega \in C^2, \tag{8.28}$$
$$(A_{\mathcal{W}})^{-1} \in \mathcal{L}(\mathcal{W}' \times \mathcal{W}_D, \mathcal{W}), \quad a_{i,j} \in W^{1-\delta_{j0},\infty}(\Omega) \quad \forall 0 \le i,j \le n.$$

LEMMA 8.5 (regularity of solutions of FEMs). *We assume the bounded elliptic $A_{\mathcal{W}} \in \mathcal{L}(\mathcal{W}, \mathcal{W}' \times \mathcal{W}_D)$ and $A_{\mathcal{U}} \in \mathcal{L}(\mathcal{U}, \mathcal{V} \times \mathcal{V}_D)$ to be boundedly invertible, e.g., by* (8.28). *Choose the above FEs,* (8.19) *and* (8.20), *of local degree $d \ge 2$, on a quasi-uniform triangulation; see* (4.4), (8.6). *Then $A^h_{\mathcal{W}}$ is regular w.r.t. stability and convergence; hence, there exist unique solutions u_1 for $A_{\mathcal{W}}u_1 = (f,\phi)$, u_1^h, and a $C' > 0$, independent of h, s.t. (see* (4.11), (4.23)*)*

(8.29)
$$A^h_{\mathcal{W}}u_1^h = Q^{h'}_{c,\Pi}(f,\phi), \quad (f,\phi) \in \mathcal{V} \times \mathcal{V}_D \implies \|u_1^h\|_{\mathcal{U}^h} \le C'(\|f\|_{\mathcal{V}} + \|\phi\|_{\mathcal{V}_D}) \quad \textit{with}$$
$$\|E_c u_1 - u_1^h\|_{\mathcal{U}^h} \le C'(\|E_c u_1 - P^h u_1\|_{\mathcal{U}^h} + \|E_c\phi - P^h u_1\|_{\mathcal{V}_D^h}). \tag{8.30}$$

$A_{\mathcal{W}} \in \mathcal{L}(\mathcal{W}, \mathcal{W}'_\Pi)$ and $A_{\mathcal{U}} \in \mathcal{L}(\mathcal{U}, \hat{\mathcal{V}}_\Pi)$ simultaneously satisfy the Fredholm alternative.

Remark 8.1. Hence, the FE error in the strong norm $\|E_c u_1 - u_1^h\|_{\mathcal{U}^h}$ is estimated by the interpolation errors of u_1 in $\mathcal{U}^h$ and $\mathcal{V}_D^h$. The FE spaces considered here need a quasi-uniform triangulation guaranteeing inverse estimates in the global form, $\|v^h\|^h_{H^j(\Omega_c^h)} \le C\ h^{l-j}\|v^h\|^h_{H^l(\Omega_c^h)}$; see (4.5). If inverse estimates will be available for $\mathcal{S}^1_d(\mathcal{T}_c^h)$ on nondegenerate triangulations, e.g., by applying the techniques in Graham, Hackbusch, and Sauter [36], the lemma remains valid for this case. Under the conditions of the lemma, (8.29) even holds for noninvertible $A_{\mathcal{W}}$ with $\|(f,\phi)\|_{\hat{\mathcal{V}}_\Pi}$ replaced by $\|(f,\phi)\|_{\hat{\mathcal{V}}_\Pi} + \|u_1\|_{\mathcal{W}}$; see [37].

For Figure 8.2 and the following proof we contrast the strong and weak operators $A_{\mathcal{U}} \in \mathcal{L}(\mathcal{U} \cup \mathcal{U}_c, \hat{\mathcal{V}}_\Pi)$ and $A_{\mathcal{W}} \in \mathcal{L}(\mathcal{W} \cup \mathcal{W}_c, \mathcal{W}'_\Pi)$, extended as in (8.25); see (6.16)–(6.18), (8.14), (8.15) with the FE versions in (6.24), (8.17). The horizontal arrows in the first line of Figure 8.2 indicate compact embeddings and in the second line the decreasingly weaker norms in the spaces $\mathcal{U}^h = \mathcal{W}^h$ and $\hat{\mathcal{V}}_\Pi^h \subset \mathcal{W}_\Pi^{h'}$. The $\longrightarrow$ between $\mathcal{W}, \hat{\mathcal{V}}_\Pi$ are missing, since only $\mathcal{W} \longrightarrow \mathcal{V}$ would be appropriate.

Proof. For a boundedly invertible $A_{\mathcal{W}} \in \mathcal{L}(\mathcal{W}, \mathcal{W}'_\Pi)$, Theorem 8.4 implies the stability and thus the bounded invertibility of $A^h_{\mathcal{W}} = Q^{h'}_{c,\Pi}A_{\mathcal{W}}|_{\mathcal{W}^h} \in \mathcal{L}(\mathcal{W}^h, \mathcal{W}_\Pi^{h'})$. This yields with the consistency in Theorem 7.1 the existence of a discrete solution u_1^h for $A^h_{\mathcal{W}}u_1^h = Q^{h'}_\Pi(f,\phi) \in \mathcal{W}_\Pi^{h'}$ by Theorem 5.1. The solution $u_1^h \in \mathcal{W}^h$ for $(f,\phi) \in \mathcal{W}'_\Pi$ satisfies the estimate $\|u_1^h\|_{\mathcal{W}^h} \le C\|(f,\phi)\|_{\mathcal{W}'_\Pi}$. Now we use the same technique as in the proof of Theorem 8.3 for reducing the inhomogeneous to trivial Dirichlet boundary conditions. We have assumed $\phi \in \mathcal{V}_D = H^{3/2}(\partial\Omega)$ as a restriction of $\phi \in \mathcal{V} = H^2(\Omega)$.

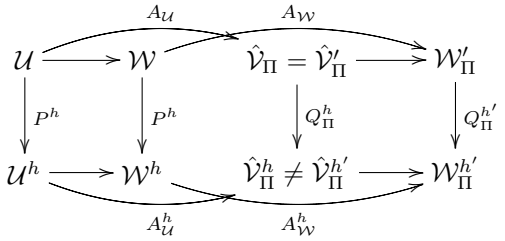


FIG. 8.2. *Strong and weak FEMs for linear problems.*

This is possible by combining Theorem 4.5 with Zeidler [64, p. 1030]. The last lines of the proof of Theorem 7.1 can then be updated to obtain the (8.29), (8.30) results for $\phi \not\equiv 0$ if they are correct for $\phi \equiv 0$. So we restrict the discussion for the remaining proof to trivial Dirichlet boundary conditions.

We find for $P^h = I^h E_c$, $A_{\mathcal{U}}$, and the $v^h_\Pi = (v^h, 0) \in \hat{\mathcal{V}}^h_\Pi$ (cf. (8.25), (6.16)–(6.18))

$$P^h A^{-1}_{\mathcal{U}} : \mathcal{V}^h_\Pi \subset \mathcal{V}_{c,\Pi} \to \mathcal{U}^h \text{ or } P^h A^{-1}_{\mathcal{U}} v^h_\Pi \text{ is well defined } \ \forall v^h_\Pi = (v^h, 0) \in \hat{\mathcal{V}}^h_\Pi.$$

$\mathcal{V}^h \subset C^1(\ \overline{\Omega^h_c}\ ) \subset C^\gamma(\ \overline{\Omega^h_c}\ )$ for a $0 < \gamma < 1$ implies that $A^{-1}_{\mathcal{U}} v^h_\Pi \in C^{2,\gamma}(\ \overline{\Omega^h_c}\ )$ by standard regularity results, e.g., [63, Chapter 6.3, Problem 6.8, p. 259] and so the $P^h A^{-1}_{\mathcal{U}} v^h_\Pi$ is well defined for all $v^h_\Pi = (v^h, 0 \in \hat{\mathcal{V}}^h_\Pi$. Alternatively, we choose interpolation via the Bramble–Hilbert lemma. Thus $P^h A^{-1}_{\mathcal{U}} v^h_\Pi$ is well defined for derivatives up to second order for P^h in (4.14), e.g., for the Argyris FEs. Higher derivatives defining P^h require quasi interpolants as, e.g., in Theorem 4.2, Scott and Zhang [48], and Davydov [27].

For the estimate in (8.35) we need the embedding $I_{WU} : \mathcal{W}^h \to \mathcal{U}^h$, the bounded transformation (cf. (6.21)), and [27]:

(8.31)
$$I^h_{\mathcal{V}^0} : \mathcal{V}^{h',0}_\Pi := \mathcal{V}^{h'} \times \{0\} \subset \mathcal{V}^{h'}_\Pi \to \mathcal{V}^0_{c,\Pi} := \mathcal{V}_c \times \{0\} \subset \mathcal{V}_{c,\Pi},\ (v^{h'}, 0) \to (Q^h_d v^{h'}, 0).$$

We assumed a smooth $(f \in \mathcal{V}, \phi \equiv 0) \in \hat{\mathcal{V}}_\Pi$. By (6.22) and $\phi \equiv 0$ we restrict the discussion to $u^h \in \mathcal{W}^h \cap S^h_0 = \mathcal{U}^h_0 = \mathcal{W}^h_0 = \mathcal{V}^h$. For the equalities between weak and strong problems we use special spaces and (6.16)–(6.19), (8.9), (8.14)–(8.18):

$$\forall u^h, v^h \in \mathcal{U}^h_0 = \mathcal{W}^h_0 = \mathcal{V}^h,\ \forall v^h_\Pi := (v^h, 0) \in \hat{\mathcal{V}}^{h,0}_\Pi := \mathcal{V}^h \times \{0\},\ \hat{\mathcal{V}}^{h',0}_\Pi,\ \hat{\mathcal{W}}^{h,0}_\Pi,\ \mathcal{W}^{h',0}_\Pi.$$

We find for $f \in \mathcal{V}$ and the definition of A, A_s in $\Omega \cup \Omega^h_c$ that

$$\langle A^h_{\mathcal{W}} u^h, v^h_\Pi \rangle_{\mathcal{W}^{h',0}_\Pi \times \mathcal{W}^{h,0}_\Pi} = \langle A u^h, v^h \rangle_{\mathcal{W}'_c \times \mathcal{W}_c} = (A_s u^h, v^h)_{\mathcal{V}_c} = \langle A^h_{\mathcal{U}} u^h, v^h_\Pi \rangle_{\mathcal{V}^{h',0}_\Pi \times \mathcal{V}^{h,0}_\Pi} \text{ and}$$

$$(E_c f, v^h)_{\mathcal{V}_c} = \langle Q^{h',0}_\Pi (E_c f, 0), v^h_\Pi \rangle_{\mathcal{W}^{h',0}_\Pi \times \mathcal{W}^{h,0}_\Pi} = \langle Q^{h,0}_\Pi (E_c f, 0), v^h_\Pi \rangle_{\mathcal{V}^{h',0}_\Pi \times \mathcal{V}^{h,0}_\Pi}. \tag{8.32}$$

This implies coinciding weak and strong bilinear forms, since for all $v^h \in \mathcal{V}^h$

$$\langle A^h_{\mathcal{W}} u^h_1, v^h_\Pi \rangle_{\mathcal{W}^{h',0}_\Pi \times \mathcal{W}^{h,0}_\Pi} = (A_s u^h_1, v^h)_{\mathcal{V}_c} = \langle A^h_{\mathcal{U}} u^h_1, v^h_\Pi \rangle_{\mathcal{V}^{h',0}_\Pi \times \mathcal{V}^{h,0}_\Pi} = (E_c f, v^h)_{\mathcal{V}_c}. \tag{8.33}$$

Hence, a weak solution u^h_1 solves the strong problem as well:

$$A^h_{\mathcal{W}} u^h_1 = Q^{h',0}_\Pi (f, 0), \quad (f, 0) \in \mathcal{V}^0_\Pi := \mathcal{V} \times \{0\} \Longrightarrow A^h_{\mathcal{U}} u^h_1 = Q^{h,0}_\Pi (f, 0). \tag{8.34}$$

For proving stability and consistency for the strong form, we start with the identity

$$(A^h_{\mathcal{U}})^{-1}_{\mathcal{U}^h_0 \hookleftarrow \mathcal{V}^{h',0}_{\Pi}} = \big[P^h_{\mathcal{U}^h_0 \hookleftarrow \mathcal{U}_{c,0}}(A^{-1}_{\mathcal{U}})_{\mathcal{U}_{c,0} \hookleftarrow \mathcal{V}^0_{c,\Pi}} - (I_{WU})_{\mathcal{U}^h_0 \hookleftarrow \mathcal{W}^h_0} (A^h_{\mathcal{U}})^{-1}_{\mathcal{W}^h_0 \hookleftarrow \mathcal{W}^{h',0}_{\Pi}} [(A^h_{\mathcal{U}} P^h - Q^h_{c,\Pi} A_{\mathcal{U}})]_{\mathcal{W}^{h',0}_{\Pi} \hookleftarrow \mathcal{U}_{c,0}} (A^{-1}_{\mathcal{U}})_{\mathcal{U}_{c,0} \hookleftarrow \mathcal{V}^0_{c,\Pi}}\big] I_{\mathcal{V}^0_{c,\Pi}, \mathcal{V}^{h',0}_{\Pi}}. \tag{8.35}$$

The inverse and the error estimates (4.5) and (4.11), (4.23), the approximation property (4.23), for the local degree $d \geq 2$, show the three inequalities

$$\|I_{WU}\|_{\mathcal{U}^h_0 \hookleftarrow \mathcal{W}^h_0} \leq C_1 h^{-1},\ \|P^h\|_{\mathcal{U}^h_0 \hookleftarrow \mathcal{U}_{c,0}} \leq C_2,\ \|P^h u_1 - E_c u_1\|_{\mathcal{W}^h} \leq C_3 h \|u_1\|_{\mathcal{U}}. \tag{8.36}$$

With (4.23), the solution of $A_{\mathcal{U}} u_1 = (f, 0) \in \mathcal{V}_{\Pi}$, equal weak and strong bilinear forms, $a_{i,j}$ (cf. (8.2)), defined in $\Omega \cup \Omega^h_c$, and the special $v^h_{\Pi} = (v^h, 0) \in \mathcal{V}^{h,0}_{\Pi}$, we get by (4.22)

$$\begin{aligned}
|\langle A^h_{\mathcal{U}} P^h u_1 - Q^h_{\Pi} A_{\mathcal{U}} u_1, v^h_{\Pi} \rangle_{\mathcal{W}^{h',0}_{\Pi} \times \mathcal{W}^{h,0}_{\Pi}}| &= |(A_s P^h u_1 - Q^h f, v^h)_{\mathcal{V}_c}| \\
&= |(A_s P^h u_1 - E_c f, v^h)_{\mathcal{V}_c}| = |(A_s P^h u_1 - E_c A_s u_1, v^h)_{\mathcal{V}_c}| \\
&= \Big| \Big(\sum_{i,j=0}^{n} a_{ij} (\partial^i\ P^h u_1 -\ E_c \partial^i u_1), \partial^j v^h \Big)_{\mathcal{V}_c} \Big| + Ch \|u_1\|_{\mathcal{U}} \|v^h\|_{\mathcal{W}^h} \\
&\leq (C_3 + C) h \|u_1\|_{\mathcal{U}} \|v^h\|_{\mathcal{W}^h} \quad \forall v^h_{\Pi} = (v^h, 0) \in \mathcal{V}^{h,0}_{\Pi}.
\end{aligned}$$

With (8.34), this consistency, and the stability of $A^h_{\mathcal{W}}$, the $\|(A^h_{\mathcal{W}})^{-1}_{\mathcal{W}^h_0 \hookleftarrow \mathcal{W}^{h',0}_{\Pi}}\| \leq C_4$, we estimate the first term in (8.35):

$$\begin{aligned}
&\|(A^h_{\mathcal{U}})^{-1}_{\mathcal{W}^h_0 \hookleftarrow \mathcal{W}^{h',0}_{\Pi}} [(A^h_{\mathcal{U}} P^h - Q^h_{c,\Pi} A_{\mathcal{U}})]_{\mathcal{W}^{h',0}_{\Pi} \hookleftarrow \mathcal{U}_{c,0}}\|_{\mathcal{W}^h_0 \hookleftarrow \mathcal{U}_{c,0}} \\
&= \|(A^h_{\mathcal{W}})^{-1}_{\mathcal{W}^h_0 \hookleftarrow \mathcal{W}^{h',0}_{\Pi}} [(A^h_{\mathcal{U}} P^h - Q^h_{c,\Pi} A_{\mathcal{U}})]_{\mathcal{W}^{h',0}_{\Pi} \hookleftarrow \mathcal{U}_{c,0}}\|_{\mathcal{W}^h_0 \hookleftarrow \mathcal{U}_{c,0}} \\
&\leq C_4 \|(A^h_{\mathcal{U}} P^h - Q^h_{\Pi} A_{\mathcal{U}})\|_{\mathcal{W}^{h',0}_{\Pi} \hookleftarrow \mathcal{U}_{c,0}} \\
&\leq C_4 \sup_{0 \neq u \in \mathcal{U}_{c,0}} \sup_{0 \neq v^h \in \mathcal{W}^h_0} \frac{|\langle A^h_{\mathcal{U}} P^h u - A_{\mathcal{U}} u, v^h_{\Pi} \rangle_{\mathcal{W}^{h',0}_{\Pi} \times \mathcal{W}^{h,0}_{\Pi}}|}{\|v^h\|_{\mathcal{W}^h} \|u\|_{\mathcal{U}}} \leq C_3 C_4 h.
\end{aligned}$$

With (8.36), (8.35) and a boundedly invertible $\|(A^{-1}_{\mathcal{U}}\|_{\mathcal{U}_{c,0} \hookleftarrow \mathcal{V}^0_{c,\Pi}} \leq C_5$ and a bound for $\|I_{\mathcal{V}^0_{c,\Pi}, \mathcal{V}^{h',0}_{\Pi}}\|_{\mathcal{V}^0_{c,\Pi} \hookleftarrow \mathcal{V}^{h',0}_{\Pi}} \leq C_6$, this yields the stability of $A^h_{\mathcal{U}} : \mathcal{U}^h_0 \to \mathcal{V}^{h',0}_{\Pi}$, since

$$\begin{aligned}
\|(A^h_{\mathcal{U}})^{-1}\|_{\mathcal{U}^h_0 \hookleftarrow \mathcal{V}^{h',0}_{\Pi}} &\leq \big[\|P^h\|_{\mathcal{U}^h_0 \hookleftarrow \mathcal{U}_{c,0}} \|(A^{-1}_{\mathcal{U}})\|_{\mathcal{U}_{c,0} \hookleftarrow \mathcal{V}^0_{c,\Pi}} + \|(I_{WU})_{\mathcal{U}^h_0 \hookleftarrow \mathcal{W}^h_0}\| \\
&\|(A^h_{\mathcal{U}})^{-1}_{\mathcal{W}^h_0 \hookleftarrow \mathcal{W}^{h',0}_{\Pi}} [(A^h_{\mathcal{U}} P^h - Q^h_{c,\Pi} A_{\mathcal{U}})]_{\mathcal{W}^{h',0}_{\Pi} \hookleftarrow \mathcal{U}_{c,0}}\|_{\mathcal{W}^h_0 \hookleftarrow \mathcal{U}_{c,0}} \|A^{-1}_{\mathcal{U}}\|_{\mathcal{U}_{c,0} \hookleftarrow \mathcal{V}^0_{c,\Pi}}\big], \\
\|I_{\mathcal{V}^0_{c,\Pi}, \mathcal{V}^{h',0}_{\Pi}}\|_{\mathcal{V}^0_{c,\Pi} \hookleftarrow \mathcal{V}^{h',0}_{\Pi}} &\leq \big[C_2 C_5 + C_1 h^{-1} C_3 C_4 h C_5\big] C_6 \leq C.
\end{aligned}$$

The combination of these inequalities yields the claim, since

$$\|u^h_1\|_{\mathcal{U}^h} = \|(A^h_{\mathcal{U}})^{-1} Q^h_{\Pi}(f, 0)\|_{\mathcal{U}^h} \leq \|(A^h_{\mathcal{U}})^{-1}\|_{\mathcal{U}^h_0 \hookleftarrow \mathcal{V}^{h',0}_{\Pi}} \|Q^h_{\Pi}\|_{\mathcal{V}^{h',0}_{\Pi} \hookleftarrow \mathcal{V}^0_{\Pi}} \|f\|_{\mathcal{V}} \leq C \|f\|_{\mathcal{V}}.$$

The convergence in (8.30) is, by Theorem 5.1, an immediate consequence of the stability in (8.29) and the consistency in Theorem 7.1. ∎

8.5. Stability for the strong linear operator and convergence. Combining Theorem 8.4, Proposition 8.2, and Lemma 8.5 we obtain the following.

THEOREM 8.6. *For the linearized operator* $A_s = G'(u_0) \in \mathcal{L}(\mathcal{U}, \mathcal{V})$ *in* (8.2)*, assume* (3.10)*, for the domain* (8.28)*, a quasi-uniform* $\mathcal{S}^1_d(\mathcal{T}^h), d \geq 2^n + 1$ *(see* (4.4)*), and let the corresponding* $A_s \in \mathcal{L}(\mathcal{U}_0, \mathcal{V})$ *and, hence,* $A_{\mathcal{U}} := (A_s, B_D) \in \mathcal{L}(\mathcal{U}, \mathcal{V} \times \mathcal{V}_D)$ *be boundedly invertible. Then the conforming FEM* (8.7) *for the linear* $A^h_s u^h_1 = Q^h A_s|_{\mathcal{U}_0} u^h_1 = Q^h f, f \in \mathcal{V}$*, and the nonconforming FEM* (8.19)*,* $A^h_{\mathcal{U}} u^h_1 = Q^h_{c,\Pi} A_{\mathcal{U}} u^h_1 = Q^h_\Pi(f, \phi)$*,* $(f, \phi) \in \mathcal{V} \times \mathcal{V}_D$*, are stable as well.*

We discussed the relation between the bounded invertibility of $A_s = G'(u_0) \in \mathcal{L}(\mathcal{U}_0, \mathcal{V})$ and $A_{\mathcal{U}} = (A_s, B_D) \in \mathcal{L}(\mathcal{U}, \mathcal{V} \times \mathcal{V}_D)$ and the existence of a locally unique solution of $G(u_0) = 0$ and $F(u_0) = 0$ already at the end of section 3.

Proof. The boundedness of the strong $A_{\mathcal{U}} := (A_s, B_D) \in \mathcal{L}(\mathcal{U}, \mathcal{V} \times \mathcal{V}_D)$ and its weak counterpart $A_{\mathcal{W}} \in \mathcal{L}(\mathcal{W}, \mathcal{W}' \times \mathcal{W}_D)$ under the conditions in (8.28), (8.2), (8.3) is straightforward. We want to show that the bounded invertibility of $A_{\mathcal{U}}$ implies that of $A_{\mathcal{W}} \in \mathcal{L}(\mathcal{W}, \mathcal{W}' \times \mathcal{W}_D)$ as well. Indeed, by (8.2)–(8.3), the corresponding $G'(u_0) \in \mathcal{L}(\mathcal{U}, \mathcal{V}')$ induces a strong bounded bilinear form, by Proposition 8.2 simultaneously a $\mathcal{W}$-elliptic bilinear form, and thus finally $A_{\mathcal{W}} \in \mathcal{L}(\mathcal{W}, \mathcal{W}' \times \mathcal{W}_D)$. Thus the Fredholm alternative applies: If $A_{\mathcal{W}} \in \mathcal{L}(\mathcal{W}, \mathcal{W}' \times \mathcal{W}_D)$ would not be boundedly invertible, each of the infinitely many solutions of $A_{\mathcal{W}} u_1 = (f, \phi) \in \mathcal{V} \times \mathcal{V}_D$ for appropriate (f, ϕ) would be $u_1 \in \mathcal{U}$ by (8.28) and the well-known regularity results; see, e.g., Hackbusch [37, Chapter 9]. So $A_{\mathcal{U}}$ could not be boundedly invertible either.

Now we apply Theorem 8.4 to guarantee the stability for the strong form: The bounded invertibility of $Q^{h'}_\Pi A_{\mathcal{W}}|_{\mathcal{W}^h} \in \mathcal{L}(\mathcal{W}^h, \mathcal{W}^{h'}_\Pi)$ and Lemma 8.5 guarantee the stability of $Q^h_\pi A_{\mathcal{U}}|_{\mathcal{U}^h} \in \mathcal{L}(\mathcal{U}^h, \mathcal{V}^h_\Pi)$ and thus the claim. ∎

Theorems 5.1, 7.1, and 8.1 and the last section combined with the last proof show that we have proved stability and convergence for the FEM in (6.26) if G is continuously differentiable near a locally unique solution u_0 with boundedly invertible $F'(u_0)$. This is guaranteed, again in the interplay between weak and strong form of the derivative, e.g., with $B_{r,\mathcal{W}}(u_0)$ the ball w.r.t. $\|u - u_0\|_{\mathcal{W}} < r$, by

$$\begin{aligned} &a_{ij}(w(\cdot)) \in C^{1-\delta_{0j},\infty}(B_{r,\mathcal{W}}(u_0)) \text{ and } a_{ij}(w(\cdot)) \in C^{1-\delta_{0j},\infty}(B_{r,\mathcal{U}}(u_0)), \\ &\text{yield } C(B_{r,\mathcal{W}}(u_0)) \ni A = G'(\cdot) : \mathcal{W} \to \mathcal{W}' \text{ and } C(B_{r,\mathcal{U}}(u_0)) \ni A_s = G'(\cdot) : \mathcal{U} \to \mathcal{V}', \end{aligned} \tag{8.37}$$

where $i, j = 0, \ldots, n$, $a_{ij}(w(\cdot))$ are defined in (3.7). Alternatively, we might require $a_{ij}(\cdot) \in C^{1-\delta_{0j}}(\mathcal{D}(G^w))$, with $w, w_0 \in \mathcal{D}(G^w)$ as in (3.6).

THEOREM 8.7. *Let the nonlinear elliptic problem in* (6.6)*,* $G(u_0) = 0, u_0 \in \mathcal{U}_0$*, or* $F(u_0) = (G, B_D)(u_0) = 0, u_0 \in \mathcal{U}$*, have an isolated solution, with boundedly invertible* $G'(u_0) : \mathcal{U}_0 \to \mathcal{V}$*, or* $F'(u_0) : \mathcal{U} \to \hat{\mathcal{V}}_\Pi$*, let* G *be continuously differentiable in* $B_r(u_0)$*, e.g., by* (8.37)*, and let the domain satisfy* (8.28)*. Choose* $\mathcal{U}^h = \mathcal{S}^1_d(\mathcal{T}^h_c), d \geq 2^n + 1$ *(see* (4.4)*), on a quasi-uniform triangulation. Then the conforming FEM* (5.10) *is*

$$G^h = Q^h_c G|_{\mathcal{U}^h_0} : \mathcal{U}^h_0 \to \mathcal{V}^{h'}, \quad u^h_0 \in \mathcal{D}(G^h) \subset \mathcal{U}^h_0 \text{ s.t. } G^h u^h_0 = 0.$$

The nonconforming FEM (6.26) *is*

$$F^h = Q^h_{c,\Pi} F|_{\mathcal{U}^h} : \mathcal{U}^h \to \mathcal{V}^{h'} \times \mathcal{V}^h_b, \quad u^h_0 \in \mathcal{D}(F^h) \subset \mathcal{U}^h \text{ s.t. } F^h u^h_0 = 0.$$

Both methods are stable in $P^h u_0$*, consistent, and convergent. Both have, for small enough* h*, a unique solution* u^h_0 *near* u_0*. We obtain (cf.* (7.2)*), for* $u_0 \in H^\ell(\Omega)$*,* $\ell > 2$*, for the conforming and nonconforming FEM, for polyhedral and* C^2 *domains,*

$$\|P^h u_0 - u^h_0\|_{\mathcal{U}^h} \leq C h^{\ell-2} \|u_0\|_{H^\ell(\Omega)} \text{ and } \leq C h^{\min\{\ell-2,p\}} \|u_0\|_{H^\ell(\Omega)}, \tag{8.38}$$

respectively, with $p = 2$ *for polyhedral* Ω^h *and* $p > 2$ *for polyhedral or curved* $\Omega_c^h \approx \Omega \in C^2$. $F'(u_0) \in \mathcal{L}(\mathcal{W}, \mathcal{W}' \times \mathcal{W}_D)$ *and* $F'(u_0) \in \mathcal{L}(\mathcal{U}, \mathcal{V} \times \mathcal{V}_D)$ *simultaneously satisfy the Fredholm alternative.*

Remark 8.2. 1. For the above F (see (3.8), (3.10), (8.2)), ℓ and d are related by (4.9). With the ellipticity assumption (3.10), the previous (8.2), (8.3), (8.28) are implied by (8.37), except Ω convex polyhedral or $\in C^2$. The latter are necessary for the regularity and the extension operators.

2. For quadrature approximations, exactly reproducing polynomials of degree $< k$ with $k > 2(d-1) > d > 2$, the solution $\tilde{u}_0^h$ of an approximate $\tilde{F}^h(\tilde{u}_0^h) = 0$ satisfies the same estimate as in (8.38); see Böhmer [11].

9. Numerical solution of the FE equations with Newton's method. We have indicated this problem for the (highly) nonlinear system (cf. (2.2), (2.6)) already in section 2, **4.**: We have to determine the FE $u_0^h \in \mathcal{U}^h \subset C^1(\Omega_c^h)$ s.t.

$$\langle F^h(u_0^h), v_\Pi^h \rangle = 0 \Longleftrightarrow (G(u_0^h), v^h)_{\mathcal{V}_c} = 0 \quad \forall v^h \in \mathcal{V}^h, (u_0^h, v_b^h)_{\mathcal{V}_{c,b}} = 0 \quad \forall v_b^h \in \mathcal{V}_b^h. \tag{9.1}$$

In this context again *local and stable bases* and *stable splittings* $S^h = S_0^h + S_b^h$ are important; see Theorems 4.2 and 4.4. We insert $u_0^h = \sum_{\mu=1}^m c_\mu s^\mu$ into (9.1) and test $G(u_0^h)$ with $v^h = s^j, j = 1, \dots, m_0$, and the boundary term with $v_b^h = s_b^i = s^{i+m_0}, i = 1, \dots, m - m_0$. So we obtain a nonlinear system of m equations for the m unknowns $c_\mu, \mu = 1, \dots, m$. For computing the terms in (9.1) results of Davydov [28] and Davydov and Stevenson [30], with multiresolution techniques quadrature [11], have to be used.

Equation (9.1) is solved by the well-known continuation and Newton methods. We formulate the discrete Newton method and its quadratic convergence via the less-known mesh independence principle (MIP). Both methods require the solution of the corresponding *linearized sparse FE equations,* with the defect or residuum, $F^h(u_\nu^h), u_\nu^h \approx u_0^h$. The $u^h|_{\partial\Omega_c^h} = 0$ and Proposition 8.2 allow the standard efficient solvers with FEs in $C^0(\Omega^h)$ replaced by $C^1(\Omega_c^h)$. Efficient strategies combining continuation with the MIP on different discretization levels have been studied in Allgower and Böhmer [2, 3]. Presently, multiresolution techniques are available for polyhedral Ω in $\mathbf{R}^n$ and $d \geq 2^n + 1$; see Davydov's papers, e.g., [28, 29, 30]. For $\Omega \in C^2$ and the necessary Ω_c^h for general $\mathcal{S}_d^1(\mathcal{T}_c^h)$ this is still open.

For specific cases, not discussed here, (9.1) represents a *sparse nonlinear system.*

The MIP, the Newton–Kantorowich method, and the convergence of our FEM require a unique solution u_0, s.t. $F'(u_0)$ is boundedly invertible. So we assume

$$F(u_0) = 0 \text{ with } F'(\cdot) \in C_L(B_r(u_0)), \quad (F'(u_0))^{-1} \in \mathcal{L}(\mathcal{V}, \mathcal{U}). \tag{9.2}$$

Starting near $u_0^h \approx u_0$, we apply Newton's method to F^h. By Theorem 8.6, the $(F^h(u_i^h))' : \mathcal{U}^h \to \mathcal{V}^h$ are equiboundedly invertible; hence

$$u_{i+1}^h := u_i^h - (F^h(u_i^h)')^{-1} F^h(u_i^h), \quad i = 1, \dots, \text{ is well defined.} \tag{9.3}$$

In addition to (9.2), the MIP (see Allgower and Böhmer [2, 3] and Allgower et al. [4]) requires a linear FEM; cf. (5.12). Then the Newton method converges quadratically; cf. Theorem 9.1.

Equation (8.38) shows that for $u_0 \in \mathcal{U}_s = H^\ell(\Omega), \ell > 2$ (see (8.38)), the FEM converges of orders $p' = \ell - 2 > 0$ and $p' := \min\{\ell - 2, p\} > 0$. Then, with the above

projectors $P^h = I^h E_c : \mathcal{U} \to \mathcal{U}^h$, $Q^h_\Pi : \hat{\mathcal{V}}_\Pi \to \mathcal{V}^{h'}_\Pi$ (see (6.15), (6.24)), we impose the condition

$$F|_{\mathcal{U}_s} : \mathcal{D}(F) \cap (\mathcal{U}_s := H^\ell(\Omega)) \to H^{\ell-2}(\Omega). \tag{9.4}$$

THEOREM 9.1 (MIP; see [2, 4]). *In addition to the conditions in Theorem* 8.7 *with convergence in* (8.38) *of order* $p' = \ell - 2 > 0$ *or* $p' = \min\{\ell-2, p\} > 0$ *we assume* (9.2), (9.4). *Start the Newton processes for* F, *by omitting all exponents* ^{h} *in* (9.3), *and for* F^h *with the sufficiently good starting values* $u_1 \in B_r(u_0) \cap \mathcal{D}(F) \cap \mathcal{U}_s$ *and* $u^h_1 := P^h u_1$. *Then, for* $h \le h_0 = (\min\{r, (SCL)^{-1}\}/(2Sc_0))^{1/p'}$ *(cf. the following proof), both Newton methods converge quadratically to* u_0 *and* u^h_0, *respectively, and*

$$\begin{aligned}
&\|u^h_{\nu+1} - u^h_\nu\|_{\mathcal{U}^h} \le C(\|u^h_\nu - u^h_{\nu-1}\|_{\mathcal{U}^h})^2 \text{ and}\\
&\|u^h_i - P^h u_i\|_{\mathcal{U}^h} \le Ch^{p'}\|u_i\|_{\mathcal{U}_s}, \quad i = 1, \dots,\\
&\|F^h(u^h_i) - Q^h_\Pi F(u_i)\|_{\mathcal{V}} \le Ch^{p'}\|u_i\|_{\mathcal{U}_s}, \quad i = 1, \dots,\\
&\|(u^h_i - u^h_0) - (P^h(u_i - u_0))\|_{\mathcal{U}^h} \le Ch^{p'}\|u_i\|_{\mathcal{U}_s}, \quad i = 1, \dots.
\end{aligned} \tag{9.5}$$

Under these conditions there exists $\overline{h} = \overline{h}(\epsilon, u_0)$ *s.t. for all* $h \in (0, \overline{h}]$

$$\big|\min\{i \ge 1 : \|u^h_i - P^h u_0\|_{\mathcal{U}^h} < \epsilon\} - \min\{i \ge 1 : \|u_i - u_0\|_{\mathcal{U}} < \epsilon\}\big| \le 1. \tag{9.6}$$

This last result is the reason for the notation as MIP and our remark concerning the essential h independence of the convergence.

Proof. We have to show that the conditions in [4] are satisfied. By section 6, $P^h = I^h E_c \in \mathcal{L}(\mathcal{U}, \mathcal{U}^h)$, $Q^h_\Pi \in \mathcal{L}(\hat{\mathcal{V}}_\Pi = \hat{\mathcal{V}}'_\Pi,\ \mathcal{V}^{h'}_\Pi \ne \mathcal{V}^h_\Pi)$ are bounded and linear, and $F^h = Q^h_{c,\Pi} F|_{\mathcal{U}^h} : \mathcal{D}(F^h) = \mathcal{D}(F) \cap \mathcal{U}^h \to \mathcal{V}^{h'}_\Pi$. As a consequence of (9.4) and $\mathcal{U}^h$, we obtain the uniform Lipschitz continuity of our $(F^h)'$:

$$\|(F^h)'(u^h) - (F^h)'(v^h)\|_{\mathcal{V}^h_\Pi \leftarrow \mathcal{U}^h} = \|Q^h_{c,\Pi}(F'(u^h) - F'(v^h))\|_{\mathcal{V}^h_\Pi \leftarrow \mathcal{U}^h} \le CL\|u^h - v^h\|_{\mathcal{U}^h}$$

for all $u^h, v^h \in B_r(P^h u_0)$. As a consequence of $F'(\cdot) \in C_L(B_r(u_0))$ and the isolated solution u_0 of $F(u) = 0$ with boundedly invertible $F'(u_0)$ we obtain stability of $(F^h(u^h))'$, and hence $\|((F^h)'(u^h))^{-1}\|_{\mathcal{U}^h \leftarrow \mathcal{V}^h_\Pi} \le S$ for all $u^h \in B_{r/2}(P^h u_0)$ with a slightly reduced $r/2$; see Theorem 8.1. Finally we need the *consistent differentiability*

$$\begin{aligned}
&\forall u, v \in B_r(u_0) \cap \mathcal{U}_s : \quad \|Q^h_\Pi F(u) - Q^h_{c,\Pi} F(P^h u)\|_{\mathcal{V}^h_\Pi} \le Ch^{p'}\|u\|_{\mathcal{U}_s} \le c_0 h^{p'},\\
&\text{and } \|Q^h_\Pi F'(u)v - Q^h_{c,\Pi} F'(P^h u)P^h v\|_{\mathcal{V}^h_\Pi} \le Ch^{p'}(1 + \|u\|_{\mathcal{U}_s})\|v\|_{\mathcal{U}_s} \le c_1 h^{p'}\|v\|_{\mathcal{U}_s},
\end{aligned} \tag{9.7}$$

an immediate consequence of $F'(\cdot) \in C_L(B_r(u_0))$ and of (9.4).

For (9.6), the only missing conditions (see (3.25) in [4]) require the existence of a constant $\delta > 0$ s.t.

$$\liminf_{h>0} \|P^h u\|_{\mathcal{U}^h} \ge \delta\|u\|_{\mathcal{U}} \quad \forall u \in \mathcal{U}_s.$$

This is an immediate consequence of (4.11). ∎

Acknowledgments. I want to thank the two referees. They have very constructively helped to improve the original version of this paper. Prof. Wolfgang Hackbusch and Prof. Oleg Davydov helped with stimulating discussions. Davydov introduced me to his stable local bases for the $\mathcal{S}^1_d(\mathcal{T}^h_c)$ splines on polygonal domains. It was an exciting joint process extending his and my results as the necessary tools in his [27] for the formulation and proofs of my paper.

REFERENCES

[1] N. E. Aguilera and L. A. Caffarelli, *Regularity results for discrete solutions of second order elliptic problems in the finite element method*, Calcolo, 23 (1986), pp. 327–353.

[2] E. L. Allgower and K. Böhmer, *A Mesh Independence Principle for Operator Equations and Their Discretization*, Technical report, Department of Mathematics, Colorado State University, Fort Collins, CO, 1984.

[3] E. L. Allgower and K. Böhmer, *Application of the mesh independence principle to mesh refinement strategies*, SIAM J. Numer. Anal., 24 (1987), pp. 1335–1351.

[4] E. L. Allgower, K. Böhmer, F. A. Potra, and W. C. Rheinboldt, *A mesh-independence principle for operator equations and their discretization*, SIAM J. Numer. Anal., 23 (1986), pp. 160–169.

[5] M. Bernadou, *$\mathcal{C}^1$-curved finite elements with numerical integration for thin plate and thin shell problems.* I: *Construction and interpolation properties of curved $\mathcal{C}^1$ finite elements*, Comput. Methods Appl. Mech. Engrg., 102 (1993), pp. 255–289.

[6] M. Bernadou, *β-curved finite elements with numerical itegration for thin plate and thin shell problems.* II: *Approximation of thin plate and thin shell problems*, Comput. Methods Appl. Mech. Engrg., 102 (1993), pp. 389–421.

[7] K. Böhmer, *On a numerical Liapunov-Schmidt method for operator equations*, Computing, 51 (1993), pp. 237–269.

[8] K. Böhmer, *On numerical bifurcation studies for general operator equations*, in International Conference on Differential Equations, Vol. 2, J. Sprekels, B. Fiedler, and K. Gröger, eds., 1999, World Scientific, Singapore, 2000, pp. 877–883.

[9] K. Böhmer, *On hybrid methods for bifurcation studies for general operator equations*, in Ergodic theory, Analysis, and Efficient Simulation of Dynamical Systems, B. Fiedler, ed., Springer-Verlag, Berlin, Heidelberg, New York, 2001, pp. 73–107.

[10] K. Böhmer, *Numerical Methods for Bifurcation and Center Manifolds in Elliptic and Parabolic Problems*, Preliminary Version, 2006.

[11] K. Böhmer, *On Finite Element Methods for Fully Nonlinear Elliptic Equations and Systems of Orders* 2 *and* $2m$, Bericht Fachbereich Mathematik und Informatik 3, Philipps Universität Marburg, Marburg, Germany, 2008.

[12] K. Böhmer and N. Sassmannshausen, *Numerical Liapunov-Schmidt spectral method for k-determined problems*, Comput. Methods Appl. Mech. Engrg., 170 (1999), pp. 277–312.

[13] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, Springer-Verlag, New York, 2002.

[14] F. Brezzi, J. Rappaz, and P. A. Raviart, *Finite-dimensional approximation of nonlinear problems,* I. *Branches of nonsingular solutions*, Numer. Math., 36 (1980), pp. 1–25.

[15] F. Brezzi, J. Rappaz, and P. A. Raviart, *Finite-dimensional approximation of nonlinear problems,* II. *Limit points*, Numer. Math., 37 (1981), pp. 1–28.

[16] F. Brezzi, J. Rappaz, and P. A. Raviart, *Finite-dimensional approximation of nonlinear problems,* III. *Simple bifurcation points*, Numer. Math., 38 (1981), pp. 1–30.

[17] C. Canuto, M. Y. Hussaini, A. Quarteroni, and T. A. Zang, *Spectral Methods in Fluid Dynamics*, 3rd ed., Springer, Berlin, 1997.

[18] Y. Z. Chen and L. C. Wu, *Second Order Elliptic Equations and Elliptic Systems*, Transl. Math. Monogr. 174, AMS, Providence, RI, 1998.

[19] P.-G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North–Holland, Amsterdam, New York, 1978.

[20] P. Clément, *Approximation by finite element functions using local regularization*, RAIRO Anal. Numer., 9 (1975), pp. 77–84.

[21] K. A. Cliff, A. Spence, and S. J. Tavener, *The numerical analysis of bifurcation problems with application to fluid mechanics*, in Acta Numerica, Cambridge University Press, Cambridge, UK, 2000, pp. 39–131.

[22] R. Courant and D. Hilbert, *Methods of Mathematical Physics. Volume* II: *Partial Differential Equations*, John Wiley & Sons, New York, 1989.

[23] M. G. Crandall and P.-L. Lions, *Hamilton-Jacobi equations in infinite dimensions.* I: *Uniqueness of viscosity solutions*, J. Funct. Anal., 62 (1985), pp. 379–396.

[24] M. G. Crandall and P.-L. Lions, *Convergent difference schemes for nonlinear parabolic equations and mean curvature motion*, Numer. Math., 75 (1996), pp. 17–41.

[25] M. G. Crandall, M. Kocan, P. L. Lions, and A. Swiech, *Existence results for boundary problems for uniformly elliptic and parabolic fully nonlinear equations*, Electron. J. Differential Equations, 1999.

[26] M. Crouziex and P. A. Raviart, *Conforming and nonconforming finite element methods for*

solving the stationary Stokes equations. I., Rev. Française Automat. Informat. Recherche Opérationnelle Sér. Rouge, 3 (1973), pp. 33–75.

[27] O. Davydov, *Smooth Finite Elements and Stable Splitting*, Bericht Fachbereich Mathematik und Informatik 3, Philipps Universität Marburg, Marburg, Germany, 2007.

[28] O. Davydov, *Stable local bases for multivariate spline spaces*, J. Approx. Theory, 111 (2001), pp. 267–297.

[29] O. Davydov and L. L. Schumaker, *On stable local bases for bivariate polynomial spline spaces*, Constr. Approx., 18 (2002), pp. 87–116.

[30] O. Davydov and R. Stevenson, *Hierarchical Riesz bases for* $h^s(\omega)$*,* $1 < s < \frac{5}{2}$, Constr. Approx., 43 (2005), pp. 365–394.

[31] C. de Boor, *A Practical Guide to Splines*, Springer-Verlag, New York, Heidelberg, Berlin, 1978.

[32] E. J. Dean and R. Glowinski, *Numerical solution of the two-dimensional elliptic Monge–Ampère equation with Dirichlet boundary conditions: An augmented Lagrangian approach*, C. R. Math. Acad. Sci. Paris, 336 (2003), pp. 779–784.

[33] E. J. Dean and R. Glowinski, *An augmented Lagrangian approach to the numerical solution of the Dirichlet problem for the elliptic Monge-Ampère equation in two dimensions*, Electron. Trans. Numer. Anal., 22 (2006), pp. 71–96.

[34] S. R. Fulton, *Multigrid solution of non-linear balance equations in meteorology*, Commun. Appl. Numer. Methods, 8 (1992), pp. 695–706.

[35] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin, 2001.

[36] I. G. Graham, W. Hackbusch, and S. A. Sauter, *Hybrid Galerkin boundary elements: Theory and implementation*, Numer. Math, 86 (2000), pp. 139–172.

[37] W. Hackbusch, *Elliptic Differential Equations: Theory and Numerical Treatment*, Springer-Verlag, Berlin, 1992.

[38] G. J. Haltiner, *Numerical Weather Prediction*, Wiley, New York, 1971.

[39] A. Kasahara, *Significance of Non-Elliptic Regions in Balanced Flows of the Tropical Atmosphere*, preprint, National Center for Atmospheric Research, University of Boulder, Boulder, CO, 1981.

[40] H. B. Keller, *Approximation methods for nonlinear problems with application to two-point boundary value problems*, Math. Comp., 29 (1975), pp. 464–474.

[41] H.-J. Kuo and N. S. Trudinger, *Linear elliptic difference inequalities with random coefficients*, Math. Comp., 55 (1990), pp. 37–53.

[42] M. Lenoir, *Optimal isoparametric finite elements and error estimates for domains involving curved boundaries*, SIAM J. Numer. Anal., 23 (1986), pp. 562–580.

[43] J. T. Oden and J. N. Reddy, *An Introduction to the Mathematical Theory of Finite Elements*, Wiley, New York, 1976.

[44] V. I. Oliker and L. D. Prussner, *On the numerical solution of the equation* $\frac{\partial^2 z}{\partial x^2}\frac{\partial^2 z}{\partial y^2} - (\frac{\partial^2 z}{\partial x \partial y})^2 = f$ *and its discretizations.* I., Numer. Math., 54 (1988), pp. 271–293.

[45] J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.

[46] H.-J. Reinhardt, *Analysis of Approximation Methods for Differential and Integral Equations*, Springer-Verlag, Berlin, Heidelberg, New York, 1985.

[47] N. Sassmannshausen and K. Böhmer, *Petrov-Galerkin Methods for Projected linear operator equation, stability and convergence*, preprint, DFG-Schwerpunktprogramm Danse, 1998.

[48] L. R. Scott and S. Zhang, *Finite element interpolation of nonsmooth functions satisfying boundary conditions*, Math. Comp., 54 (1990), pp. 483–493.

[49] R. E. Showalter, *Monotone Operators in Banach Spaces and Nonlinear Partial Differential Equations*, AMS, Providence, RI, 1997.

[50] I. Skrypnik, *Nonlinear Elliptic Boundary Value Problems*, Teubner Verlag, Leipzig, 1986.

[51] H. Stetter, *Analysis of Discretization Methods for Ordinary Differential Equations*, Springer-Verlag, Berlin, Heidelberg, New York, 1973.

[52] F. Stummel, *Diskrete Konvergenz linearer Operatoren* I, Math. Ann., 190 (1970), pp. 45–92.

[53] F. Stummel, *Diskrete Konvergenz linearer Operatoren* II, Math. Z., 120 (1971), pp. 231–264.

[54] M. E. Taylor, *Partial Differential Equations* I*, Basic Theory*, Springer-Verlag, New York, Berlin, Heidelberg, 1996.

[55] M. E. Taylor, *Partial Differential Equations* II*, Qualitative Studies of Linear Equations*, Springer-Verlag, New York, Berlin, Heidelberg, 1996.

[56] M. E. Taylor, *Partial Differential Equations* III*, Nonlinear Equations*, Springer-Verlag, New York, Berlin, Heidelberg, 1996.

[57] T. Tiihonen, *Shape calculus and finite element method in smooth domains*, Math. Comp., 70 (2001), pp. 1–15.
[58] F. Treves, *Basic Linear Partial Differential Equations*, Academic Press, London, 1975.
[59] G. Vainikko, *Funktionsanalysis der Diskretisierungsmethoden*, Teubner-Texte zur Mathematik, Teubner Verlag, Leipzig, 1976.
[60] B. S. Westcott, *Shaped Reflector Antenna Design*, Research Studies Press, Letchworth, Herdforedshire, England, 1983.
[61] M. F. Wheeler, *An elliptic collocation-finite element method with interior penalties*, SIAM J. Numer. Anal., 15 (1978), pp. 152–161.
[62] E. Zeidler, *Nonlinear Functional Analysis and Its Applications* I, *Fixed-Point Theorems*, Springer-Verlag, New York, Berlin, Heidelberg, London, Paris, Tokyo, 1986, 1990.
[63] E. Zeidler, *Nonlinear Functional Analysis and Its Applications* II/A, Springer-Verlag, New York, Berlin, Heidelberg, London, Paris, Tokyo, 1986, 1990.
[64] E. Zeidler, *Nonlinear Functional Analysis and Its Applications* II/B, Springer-Verlag, New York, Berlin, Heidelberg, London, Paris, Tokyo, 1986, 1990.

SIAM J. NUMER. ANAL.
Vol. 46, No. 3, pp. 1250–1265
© 2008 Society for Industrial and Applied Mathematics

OPTIMAL CONVERGENCE OF THE ORIGINAL DG METHOD FOR THE TRANSPORT-REACTION EQUATION ON SPECIAL MESHES*

BERNARDO COCKBURN†, BO DONG†, AND JOHNNY GUZMÁN†

Abstract. We show that the approximation given by the original discontinuous Galerkin method for the transport-reaction equation in d space dimensions is optimal provided the meshes are suitably chosen: the L^2-norm of the error is of order $k+1$ when the method uses polynomials of degree k. These meshes are not necessarily conforming and do not satisfy any uniformity condition; they are required only to be made of simplexes, each of which has a unique *outflow* face. We also find a new, element-by-element postprocessing of the derivative in the direction of the flow which superconverges with order $k+1$.

Key words. discontinuous Galerkin methods, transport-reaction equation, error estimates

AMS subject classifications. 65N30, 65M60

DOI. 10.1137/060677215

1. Introduction.

We show that the original discontinuous Galerkin (DG) [15, 13] method can approximate in an *optimal* fashion the solution of the convection-reaction problem

$$\boldsymbol{\beta}\cdot\nabla u + c\,u = f \quad \text{in } \Omega, \tag{1.1a}$$

$$u = g \quad \text{on } \Gamma^-. \tag{1.1b}$$

Here $\Omega \subset R^d$ is a polyhedral domain, $\Gamma^- := \{x \in \partial\Omega : \boldsymbol{\beta}\cdot\boldsymbol{n}(x) < 0\}$, and $\boldsymbol{n}(x)$ is the outward unit normal at the point $x \in \partial\Omega$. The functions f and g are smooth, $\boldsymbol{\beta}$ is a nonzero constant unit vector, and c is a bounded function. Indeed, if u_h denotes the approximation given by the DG method using polynomials of degree k, we prove that, for a *special* class of triangulations $\mathcal{T}_h$, we have

$$\|\,u-u_h\,\|_{L^2(\mathcal{T}_h)} + \|\,\mathsf{P}(\partial_{\boldsymbol{\beta}}u) - \partial_{\boldsymbol{\beta},h}u_h\,\|_{L^2(\mathcal{T}_h)} \le C\,|\,u\,|_{H^{k+1}(\mathcal{T}_h)}\,h^{k+1},$$

where $\partial_{\boldsymbol{\beta}} := \boldsymbol{\beta}\cdot\nabla$, P is the L^2-projection into the finite element space, and $\partial_{\boldsymbol{\beta},h}u_h$ is an approximation to $\partial_{\boldsymbol{\beta}}\,u$ obtained by using an element-by-element postprocessing of u_h. Note that the above approximation result is optimal for the quantity $\|\,u-u_h\,\|_{L^2(\mathcal{T}_h)}$ in the order of convergence in h as well as in the regularity of the exact solution; the estimate of the quantity $\|\,\mathsf{P}(\partial_{\boldsymbol{\beta}}u) - \partial_{\boldsymbol{\beta},h}u_h\,\|_{L^2(\mathcal{T}_h)}$ is clearly a superconvergence result. This has to be contrasted with the estimate for general triangulations

$$\|\,u-u_h\,\|_{L^2(\mathcal{T}_h)} + h^{1/2}\,\|\,\partial_{\boldsymbol{\beta}}u - \partial_{\boldsymbol{\beta}}u_h\,\|_{L^2(\mathcal{T}_h)} \le\ C\,|\,u\,|_{H^{k+1}(\mathcal{T}_h)}\,h^{k+1/2},$$

which follows from results obtained back in 1986 in [12].

*Received by the editors December 11, 2006; accepted for publication (in revised form) September 18, 2007; published electronically March 21, 2008.
http://www.siam.org/journals/sinum/46-3/67721.html

†School of Mathematics, University of Minnesota, Vincent Hall, Minneapolis, MN 55455 (cockburn@math.umn.edu, bdong@dam.brown.edu, guzma033@umn.edu). The first author's research was supported in part by National Science Foundation grant DMS-0411254 and by the University of Minnesota Supercomputing Institute. The third author's research was supported by an NSF Mathematical Science Postdoctoral Research Fellowship (DMS-0503050).

The mechanisms that induce the loss of $h^{1/2}$ in the order of convergence of the L^2-norm of the error are not very well known yet. In 1988, it was shown [16] that, in the two-dimensional case, the estimate

$$\| u - u_h \|_{L^2(\mathcal{T}_h)} \leq C \, | u |_{H^{k+2}(\mathcal{T}_h)} \, h^{k+1}$$

holds for conforming triangulations obtained by using slabs of parallelograms divided into two triangles always in the same way. The triangles were chosen so that their sides are uniformly not aligned with the convection direction $\boldsymbol{\beta}$, that is, so that they satisfy what we could call the *transversality* condition

$$|\boldsymbol{\beta} \cdot \boldsymbol{n}_K| \geq \gamma > 0 \quad \text{on } \partial K \quad \text{for all } K \in \mathcal{T}_h, \tag{1.2}$$

where $\boldsymbol{n}_K$ is the outward unit normal of the simplex K and γ is a fixed constant. In 1991, the rate of convergence of $h^{k+1/2}$ for $\| u - u_h \|_{L^2(\mathcal{T}_h)}$ was shown to be sharp in [14]; a rigorous proof was given for the case $k = 0$, and convincing numerical evidence was shown for the case $k = 1$, also in the two-dimensional case. As the reader might expect, in those numerical experiments, triangulations violating the transversality condition played a central role which led the author to conjecture that this condition "*may be a natural condition under which to seek improved estimates.*" However, a consequence of our main result is that improved estimates can be obtained even if the transversality condition is not satisfied in *any* triangle. Indeed, our improved estimates hold for triangulations $\mathcal{T}_h$ made of simplexes K satisfying the following simple *flow* conditions with respect to $\boldsymbol{\beta}$:

(1.3a) Each simplex K has a unique *outflow* face with respect to $\boldsymbol{\beta}$, e_K^+.

(1.3b) Each interior face e_K^+ is included in an *inflow* face with respect to $\boldsymbol{\beta}$ of another simplex.

We say that the face e of the simplex K is an outflow (inflow) face with respect to $\boldsymbol{\beta}$ if $\boldsymbol{\beta} \cdot \boldsymbol{n}_K|_e > (<)\ 0$. We say that a face is interior if it is not included in $\partial\Omega$. Note that the second condition allows the triangulation to be nonconforming. In two dimensions, this means that hanging nodes in a simplex K are allowed provided they are *not* in its outflow edge; an example of such a triangulation is given in Figure 1.1. Note that this triangulation is not uniform or translation invariant. In the appendix, we show how to construct triangulations satisfying the flow conditions in any number of space dimensions.

As usual, the families of triangulations we consider also satisfy the classical assumption of shape regularity (see [6]); namely, there is a constant $\sigma > 0$ such that

$$\text{for each simplex } K \in \mathcal{T}_h : h_K / \rho_K \geq \sigma, \tag{1.4}$$

where h_K denotes the diameter of the simplex K and ρ_K the diameter of the biggest ball included in K.

The main idea behind the devising of these meshes, and the associated analysis giving rise to optimal error estimates, is a suitable adaptation to the convection-reaction equation under consideration of the projection $(\boldsymbol{\Pi}, \mathbb{P})$ recently introduced for the analysis of superconvergent DG methods for second-order elliptic problems in [7]. The first component of such projection, $\boldsymbol{\Pi}$, was previously used in the analysis of the so-called minimal-dissipation local discontinuous Galerikin method for convection-diffusion problems. Here, we use the second component, $\mathbb{P}$, to render the analysis

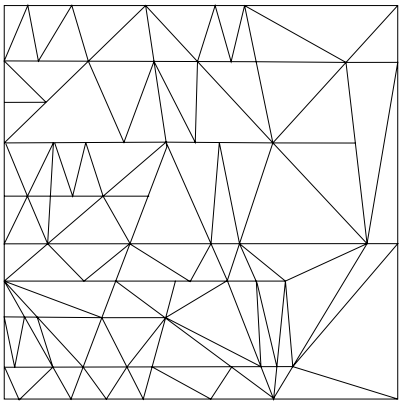


FIG. 1.1. *A triangulation satisfying the flow conditions with* $\boldsymbol{\beta} = (1, 0)$.

of error in u not only extremely simple and optimal but also capable of handling nonconforming meshes that are not uniform in any way.

Let us emphasize that the role of the special meshes for the construction of the approximation $\partial_{\boldsymbol{\beta},h} u_h$ to the directional derivative $\partial_{\boldsymbol{\beta}} u$ is *not* essential for two reasons. The first is that such an approximation can be defined on any mesh of simplexes. The second is that, for the general meshes considered in [12], we can still obtain the estimate

$$\| \mathsf{P}(\partial_{\boldsymbol{\beta}} u) - \partial_{\boldsymbol{\beta},h} u_h \|_{L^2(\mathcal{T}_h)} \leq C \, | \, u \, |_{H^{k+1}(\mathcal{T}_h)} \, h^{k+1/2},$$

which shows that even in this case, $\partial_{\boldsymbol{\beta},h} u_h$ is a better approximation than $\partial_{\boldsymbol{\beta}} u_h$. Postprocessings similar to the one giving rise to $\partial_{\boldsymbol{\beta},h} u_h$ have been used before in the context of flow through porous media [4], in the context of the Navier–Stokes equations [8], and in the context of linear elastic incompressible materials [9]. In all the above-mentioned references, they have been used to construct exactly divergence-free approximations to the velocity or exactly divergence-free stresses.

It is interesting to note that, when the triangulations $\mathcal{T}_h$ satisfy the flow conditions (1.3) with respect to $\boldsymbol{\beta}$ *and* with respect to $-\boldsymbol{\beta}$, the case in which the triangulation consists of tubes aligned with $\boldsymbol{\beta}$, we can obtain error estimates of the numerical trace $\widehat{u}_h$ in *each* outflow face e, namely,

$$\|\mathsf{P}_\partial u - \widehat{u}_h\|_{H^{-s}(e)} \leq \mathsf{C}(c, s) h^{k+s+1} | \, u \, |_{H^{k+1}(\Omega_e)}$$

for $s \in [0, k]$, where P_∂ is the L^2-projection into the space of polynomials of degree at most k on each face, Ω_e is a suitably chosen subset of $\mathcal{T}_h$, and $c \in W^{s,\infty}(\Omega_e)$. In particular, this implies that the average of $\widehat{u}_h$ on the face e converges to the average of u on that face with order $2\,k+1$. This is the only result of this type in the current available scientific literature, to the knowledge of the authors.

Finally, we extend our results in two directions. First, we consider approximations that are polynomials of different degrees on different elements. With the condition that the degrees of polynomials are nonincreasing in the direction of $\boldsymbol{\beta}$, we show that the estimates of approximations of u and $\partial_{\boldsymbol{\beta},h} u_h$ still hold. Then we consider the

singularly perturbed problem in $\Omega \subset \mathbb{R}^2$,

$$-\epsilon \Delta u + \boldsymbol{\beta} \cdot \nabla u + cu = f \quad \text{in } \Omega,$$

$$u = g \quad \text{on } \partial\Omega,$$

where $0 < \epsilon \ll 1$ is a constant, and we prove optimal *local* error estimates for $\epsilon \leq h^2$ on quasi-uniform meshes. We also show that internal numerical layers have width $\max(h, \epsilon^{1/2})$ on triangulations $\mathcal{T}_h$ that satisfy the flow conditions (1.3) with respect to both $\boldsymbol{\beta}$ and $-\boldsymbol{\beta}$. This improves the best known result of $h^{1/2}$; see [11]. A similar result for the streamline-diffusion method is contained in [17].

The paper is organized as follows. In section 2 we state and prove our main results. In section 3, we extend our results to variable-degree versions of the method; we show the results for the singularly perturbed problem. Our theoretical results are verified in section 4 by numerical experiments. We end in section 5 with some concluding remarks.

2. The main results.

2.1. The DG method. Suppose we have a family of triangulations $\{\mathcal{T}_h\}$ of Ω satisfying the flow conditions (1.3). To each triangulation $\mathcal{T}_h$, we associate the number $h = \sup_{K \in \mathcal{T}_h} h_K$, where $h_K = diam(K)$, and the finite-dimensional space V_h^k, which is composed of functions that are polynomials of degree at most k on each simplex $K \in \mathcal{T}_h$.

The DG approximation $u_h \in V_h^k$ of the solution of (1.1) satisfies

$$B(u_h, v_h) = (f, v_h)_{\mathcal{T}_h} - \langle g, v_h \, \boldsymbol{\beta} \cdot \boldsymbol{n} \rangle_{\Gamma^-} \quad \text{for all } v_h \in V_h^k, \tag{2.1a}$$

where

$$B(w, v) = -(w, \partial_{\boldsymbol{\beta}} v)_{\mathcal{T}_h} + \langle \widehat{w}, v \, \boldsymbol{\beta} \cdot \boldsymbol{n} \rangle_{\partial \mathcal{T}_h \setminus \Gamma^-} + (c\, w, v)_{\mathcal{T}_h} \tag{2.1b}$$

for any w, v in $H^1(\mathcal{T}_h)$. Note that we need only to define the numerical trace $\widehat{w}$ on faces that are not parallel to the direction $\boldsymbol{\beta}$ and do not belong to the inflow part of the boundary Γ^-. Therefore, the numerical trace of a function w on a simplex K for such faces e is given by

$$\widehat{w} := w^-, \tag{2.1c}$$

where $w^{\pm}(z) = \lim_{\delta \downarrow 0} w(z \pm \delta \boldsymbol{\beta})$, where $z \in e$. We are using the notation

$$(\boldsymbol{\sigma}, \boldsymbol{v})_{\mathcal{T}_h} := \sum_{K \in \mathcal{T}_h} \int_K \boldsymbol{\sigma}(x) \cdot \boldsymbol{v}(x)\, dx, \qquad (\zeta, \omega)_{\mathcal{T}_h} := \sum_{K \in \mathcal{T}_h} \int_K \zeta(x)\, \omega(x)\, dx,$$

$$\langle \zeta, \boldsymbol{v} \cdot \boldsymbol{n} \rangle_{\partial \mathcal{T}_h} := \sum_{K \in \mathcal{T}_h} \int_{\partial K} \zeta(\gamma)\, \boldsymbol{v}(\gamma) \cdot \boldsymbol{n}\, d\gamma$$

for any functions $\boldsymbol{\sigma}, \boldsymbol{v}$ in $\boldsymbol{H}^1(\mathcal{T}_h) := [H^1(\mathcal{T}_h)]^d$ and ζ, ω in $H^1(\mathcal{T}_h)$. The outward normal unit vector to ∂K is denoted by $\boldsymbol{n}$.

Notice that the exact solution u of (1.1) also satisfies the weak formulation (2.1), so we have the error equation

$$B(u - u_h, v_h) = 0 \qquad \text{for all } v_h \in V_h^k. \tag{2.2}$$

2.2. The approximation of u. To state our result on the approximation of u, we need to introduce a special projection, $\mathbb{P}$, defined on triangulations $\mathcal{T}_h$ satisfying the flow condition (1.3a). The function $\mathbb{P}u \in V_h^k$ restricted to $K \in \mathcal{T}_h$ is given by

$$(\mathbb{P}u - u, v)_K = 0 \qquad \text{for all } v \in \mathcal{P}^{k-1}(K), \tag{2.3a}$$

$$\langle \mathbb{P}u - u, w \rangle_{e_K^+} = 0 \qquad \text{for all } w \in \mathcal{P}^k(e_K^+), \tag{2.3b}$$

where $\mathcal{P}^\ell(D)$ stands for the space of polynomials of total degree at most ℓ defined on the set D.

The following lemma was proved in [7].

LEMMA 2.1. *If the triangulation $\mathcal{T}_h$ satisfies the flow condition* (1.3a)*, the projection $\mathbb{P}$ given by* (2.3) *is well defined. Moreover, if the triangulation $\mathcal{T}_h$ is shape-regular (see* (1.4)*), then, on each simplex $K \in \mathcal{T}_h$, we have*

$$\|\mathbb{P}u - u\|_{L^2(K)} \le Ch^{k+1} |\, u \,|_{H^{k+1}(K)}, \tag{2.4}$$

where C depends only on k and the shape regularity constant σ.

Now we can state our first error estimate.

THEOREM 2.2. *If $\mathcal{T}_h$ satisfies the flow conditions* (1.3) *and the shape-regularity condition* (1.4)*, then for $h^{1/2} \| c \|_{L^\infty(\Omega)}$ small enough, the error between u_h given by the discontinuous Galerkin method* (2.1) *and the exact solution u of the equations* (1.1) *is bounded as follows:*

$$\|\mathbb{P}u - u_h\|_{L^2(\mathcal{T}_h)} \le C \|c(u - \mathbb{P}u)\|_{L^2(\mathcal{T}_h)},$$

where C depends on $\| c \|_{L^\infty(\Omega)}$ and the diameter of Ω. In particular, if $c \equiv 0$, then

$$u_h = \mathbb{P}u.$$

Note that, after a straightforward application of the triangle inequality, we get

$$\|u - u_h\|_{L^2(\mathcal{T}_h)} \le \|u - \mathbb{P}u\|_{L^2(\mathcal{T}_h)} + C \|c(u - \mathbb{P}u)\|_{L^2(\mathcal{T}_h)},$$

and, if we assume the shape-regularity condition (1.4) on the triangulation $\mathcal{T}_h$, by the approximation property of the projection $\mathbb{P}$, (2.4), we obtain that

$$\| u - u_h \|_{L^2(\mathcal{T}_h)} \le C \, | \, u \, |_{H^{k+1}(\mathcal{T}_h)} \, h^{k+1}$$

whenever $u \in H^{k+1}(\mathcal{T}_h)$, as claimed in the introduction. Let us prove the above theorem.

Proof. Set $\mathbb{E} = u_h - \mathbb{P}u$. By the error equation (2.2), we have that for all $v \in V_h^k$,

$$B(\mathbb{E}, v) = B(u - \mathbb{P}u, v) = \sum_{i=1}^{3} T_i,$$

where, by definition of the bilinear form $B(\cdot, \cdot)$, (2.1b),

$$T_1 := - (u - \mathbb{P}u, \partial_{\boldsymbol{\beta}} v)_{\mathcal{T}_h},$$

$$T_2 := \langle u - \widehat{\mathbb{P}u}, v \, \boldsymbol{\beta} \cdot \boldsymbol{n} \rangle_{\partial \mathcal{T}_h \setminus \Gamma^-},$$

$$T_3 := (c \, (u - \mathbb{P}u), v)_{\mathcal{T}_h}.$$

Now, by the definition of the projection $\mathbb{P}$, (2.3a),

$$T_1 = 0.$$

Moreover,

$$\begin{aligned} T_2 =& \langle u - \widehat{\mathbb{P}u}, v\,\boldsymbol{\beta}\cdot\boldsymbol{n}\rangle_{\partial\mathcal{T}_h\setminus\Gamma^-} \\ =& \sum_{K\in\mathcal{T}_h} \langle u - \widehat{\mathbb{P}u}, v\,\boldsymbol{\beta}\cdot\boldsymbol{n}\rangle_{\partial K\setminus\Gamma^-} \\ =& \sum_{K\in\mathcal{T}_h} \langle u - \widehat{\mathbb{P}u}, (v^- - v^+)\,\boldsymbol{\beta}\cdot\boldsymbol{n}\rangle_{e_K^+} \\ =& \sum_{K\in\mathcal{T}_h} \langle u - \mathbb{P}u, (v^- - v^+)\,\boldsymbol{\beta}\cdot\boldsymbol{n}\rangle_{e_K^+} \end{aligned}$$

by the definition of the numerical trace $\widehat{\mathbb{P}\,u}$, (2.1c). But, since by the second flow condition (1.3b), $(v^- - v^+)|_{e_K^+} \in \mathcal{P}^k(e_K^+)$, we can conclude that

$$T_2 = 0,$$

by the definition of the projection $\mathbb{P}$, (2.3b). Therefore, we have that

$$B(\mathbb{E}, v) = (c(u - \mathbb{P}u), v)_{\mathcal{T}_h} \quad \text{for all } v \in V_h^k.$$

We claim that this implies

$$\|\mathbb{E}\|_{L^2(\mathcal{T}_h)} \le C\|c\,(u - \mathbb{P}u)\|_{L^2(\mathcal{T}_h)}$$

for $h^{1/2}\,\|\,c\,\|_{L^\infty(\Omega)}$ small enough. This claim is a straightforward consequence of the stability result Theorem 2.1 in [12], which implies that the DG approximation u_h defined by (2.1) satisfies the inequality

$$\|\,u_h\,\|_{L^2(\Omega)} \le C\left(\|\,f\,\|_{L^2(\Omega)} + \|\,\sqrt{|\boldsymbol{\beta}\cdot\boldsymbol{n}|}\,g\,\|_{L^2(\Gamma^-)}\right)$$

for $h^{1/2}\,\|\,c\,\|_{L^\infty(\Omega)}$ small enough. The claim follows by replacing u_h by $\mathbb{E}$, f by $c\,(u - \mathbb{P}u)$, and g by 0. Let us emphasize the fact that although this stability result was obtained for the two-dimensional case and conforming triangulations, it can be easily extended to multidimensional cases and nonconforming triangulations. This completes the proof. □

2.3. Postprocessing: The approximation to $\partial_{\boldsymbol{\beta}}\boldsymbol{u}$. Next we show how to postprocess u_h in order to get a superconvergent approximation of $\partial_{\boldsymbol{\beta}}u$. To this end, for each simplex K we define $\boldsymbol{q}_h \in \boldsymbol{\mathcal{P}}^k(K) + \boldsymbol{x}\,\mathcal{P}^k(K)$ to be the solution of

$$(\boldsymbol{q}_h - \boldsymbol{\beta}u_h, \boldsymbol{v})_K = 0 \qquad \text{for all } \boldsymbol{v} \in \boldsymbol{\mathcal{P}}^{k-1}(K), \tag{2.5a}$$

$$\langle(\boldsymbol{q}_h - \boldsymbol{\beta}\lambda_h)\cdot\boldsymbol{n}, w\rangle_e = 0 \qquad \text{for all } w \in \mathcal{P}^k(e), \text{ for all faces } e \text{ of } K, \tag{2.5b}$$

where $\lambda_h = \mathsf{P}_\partial g$ on Γ^- and $\lambda_h = \widehat{u}_h$ otherwise; here $\boldsymbol{\mathcal{P}}^k(K) := [\mathcal{P}^k(K)]^d$. The existence and uniqueness of $\boldsymbol{q}_h$ are well known; see, for example, [5]. We then define

$$\partial_{\boldsymbol{\beta},h} u_h := \nabla \cdot \boldsymbol{q}_h \qquad \text{in } \mathcal{T}_h.$$

We can now state the error estimate between $\partial_{\boldsymbol{\beta},h} u_h$ and $\partial_{\boldsymbol{\beta}} u$.

THEOREM 2.3. *If $\mathcal{T}_h$ is an* arbitrary *triangulation of Ω, then*

$$\|\partial_{\boldsymbol{\beta},h} u_h - \mathsf{P}(\partial_{\boldsymbol{\beta}} u)\|_{L^2(\mathcal{T}_h)} \le C \|c(u - u_h)\|_{L^2(\mathcal{T}_h)}.$$

In particular, if $c \equiv 0$, then

$$\partial_{\boldsymbol{\beta},h} u_h = \mathsf{P}(\partial_{\boldsymbol{\beta}} u).$$

Here P is the L^2-projection into V_h^k.

We can thus see that, if we use the estimates for arbitrary triangulations $\mathcal{T}_h$ obtained in [12], which assume the shape-regularity condition (1.4), we obtain

$$\| \mathsf{P}(\partial_{\boldsymbol{\beta}} u) - \partial_{\boldsymbol{\beta},h} u_h \|_{L^2(\mathcal{T}_h)} \le C \, | \, u \, |_{H^{k+1}(\mathcal{T}_h)} \, h^{k+1/2},$$

and if we assume that $\mathcal{T}_h$ satisfies the flow conditions (1.3), we obtain, by Theorem 2.2,

$$\| \mathsf{P}(\partial_{\boldsymbol{\beta}} u) - \partial_{\boldsymbol{\beta},h} u_h \|_{L^2(\mathcal{T}_h)} \le C \, | \, u \, |_{H^{k+1}(\mathcal{T}_h)} \, h^{k+1},$$

as claimed in the introduction. This implies that, if u is smooth enough, the quantity $\| \partial_{\boldsymbol{\beta}} u - \partial_{\boldsymbol{\beta},h} u_h \|_{L^2(\mathcal{T}_h)}$ is of order $h^{k+1/2}$ for arbitrary meshes and of order h^{k+1} for the special meshes under consideration.

Proof. For $v \in V_h^k$ we have

$$\begin{aligned}
(\partial_{\boldsymbol{\beta},h} u_h - \mathsf{P}(\partial_{\boldsymbol{\beta}} u), v)_{\mathcal{T}_h} &= (\nabla \cdot \boldsymbol{q}_h - \partial_{\boldsymbol{\beta}} u, v)_{\mathcal{T}_h} \\
&= -(\boldsymbol{q}_h, \nabla v)_{\mathcal{T}_h} + \langle \boldsymbol{q}_h \cdot \boldsymbol{n}, v \rangle_{\partial \mathcal{T}_h} - (\partial_{\boldsymbol{\beta}} u, v)_{\mathcal{T}_h} \\
&= -(\boldsymbol{q}_h, \nabla v)_{\mathcal{T}_h} + \langle \boldsymbol{q}_h \cdot \boldsymbol{n}, v \rangle_{\partial \mathcal{T}_h} - (f, v)_{\mathcal{T}_h} + (c\, u, v)_{\mathcal{T}_h},
\end{aligned}$$

by the definition of the exact solution u of (1.1a). By the definition of $\boldsymbol{q}_h$, (2.5),

$$\begin{aligned}
(\partial_{\boldsymbol{\beta},h} u_h - \mathsf{P}(\partial_{\boldsymbol{\beta}} u), v)_{\mathcal{T}_h} &= -(u_h, \partial_{\boldsymbol{\beta}} v)_{\mathcal{T}_h} + \langle \widehat{u}_h, v\, \boldsymbol{\beta} \cdot \boldsymbol{n} \rangle_{\mathcal{T}_h \setminus \Gamma^-} + \langle g, v \boldsymbol{\beta} \cdot \boldsymbol{n} \rangle_{\Gamma^-} \\
&\quad - (f, v)_{\mathcal{T}_h} + (c\, u, v)_{\mathcal{T}_h} \\
&= (c\,(u - u_h), v)_{\mathcal{T}_h},
\end{aligned}$$

by the definition of the approximate solution u_h, (2.1). The proof is complete once we take $v = \partial_{\boldsymbol{\beta},h} u_h - \mathsf{P}(\partial_{\boldsymbol{\beta}} u)$. ∎

2.4. The approximation of the numerical trace $\widehat{u}_h$ on outflow faces. To state the approximation result, let us introduce some notation. We are going to use the negative-order norm

$$\| \gamma \|_{H^{-s}(e)} := \sup_{\varphi \in \mathcal{C}^\infty(e)} \frac{\langle \gamma, \varphi \rangle_e}{\| \varphi \|_{H^s(e)}}, \tag{2.6}$$

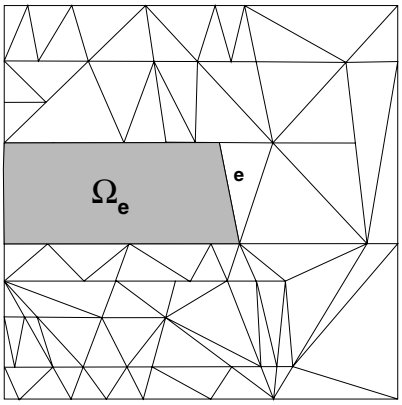


FIG. 2.1. *An example of the set* Ω_e *for* $\boldsymbol{\beta} = (1, 0)$.

where e is the outflow face of the simplex $K \in \mathcal{T}_h$. To this face, we associate the subset of $\mathcal{T}_h$ defined by

$$\Omega_e := \{K' \in \mathcal{T}_h : \text{for all } x \in K', x + \boldsymbol{\beta} t \text{ lies on } e \text{ for some } t \geq 0\}.$$

An example is displayed in Figure 2.1.

THEOREM 2.4. *Let* $\mathcal{T}_h$ *be a triangulation of* Ω *satisfying the shape-regularity condition* (1.4) *and the flow conditions* (1.3)*, with respect to* $\boldsymbol{\beta}$ *and* $-\boldsymbol{\beta}$*. Then on any outflow face* e

$$\|\mathsf{P}_\partial u - \widehat{u}_h\|_{H^{-s}(e)} \leq \mathsf{C}(c, s)\, h^{k+s+1} |\, u \,|_{H^{k+1}(\Omega_e)}$$

for $s \in [0, k]$*, where* $\mathsf{C}(c, s) := C(\|c\|_{W^{\max\{s,0\},\infty}(\Omega_e)}, \Omega_e)\, \|c\|^2_{L^\infty(\Omega_e)} / |\, \boldsymbol{\beta} \cdot \boldsymbol{n}_e \,|$*. In particular, if* $c \equiv 0$*, then*

$$\mathsf{P}_\partial u = \widehat{u}_h.$$

Note that if the triangulation $\mathcal{T}_h$ satisfies the flow conditions (1.3) with respect to $\boldsymbol{\beta}$ and $-\boldsymbol{\beta}$, then each of its simplexes K has exactly one outflow face e_K^+ and one inflow face e_K^-. Moreover, the outflow face e_K^+ lies on $\partial\Omega \setminus \Gamma^-$ or coincides with an inflow face, and e_K^- lies on Γ^- or coincides with an outflow face.

This result implies, in particular, that the average of the numerical trace $\widehat{u}_h$ on each outflow face e converges to the average of the exact solution u on that face with order $2k + 1$ for general c and with order $2k + 2$ in the case $c = 0$. To see this, let us begin by noting that if $\Upsilon \in H^{k+1}(e)$, we have that

$$\begin{aligned}\langle \widehat{u}_h - u, \Upsilon \rangle_e &= \langle \widehat{u}_h - \mathsf{P}_\partial u, \Upsilon \rangle_e + \langle \mathsf{P}_\partial u - u, \Upsilon \rangle_e \\ &= \langle \widehat{u}_h - \mathsf{P}_\partial u, \Upsilon \rangle_e + \langle \mathsf{P}_\partial u - u, \Upsilon - \mathsf{P}_\partial \Upsilon \rangle_e,\end{aligned}$$

and so

$$\begin{aligned}|\,\langle \widehat{u}_h - u, \Upsilon\rangle_e\,| &\le \|\,\widehat{u}_h - \mathsf{P}_\partial u\,\|_{H^{-k}(e)}\,\|\,\Upsilon\,\|_{H^k(e)}\\ &\quad + \|\,\mathsf{P}_\partial u - u\,\|_{L^2(e)}\,\|\,\Upsilon - \mathsf{P}_\partial\Upsilon\,\|_{L^2(e)}\\ &\le C\,h^{2\,k+1},\end{aligned}$$

where

$$C = \mathsf{C}(c,s)\,|\,u\,|_{H^k(\Omega_e)}\,\|\,\Upsilon\,\|_{H^{k+1}(e)} + C\,h\,|\,u\,|_{H^{k+1}(e)}\,|\,\Upsilon\,|_{H^{k+1}(e)}.$$

The claim follows by simply taking $\Upsilon = 1$.

Let us emphasize that the order of convergence of the approximation of linear functionals of the form $\langle u\,,\,\Upsilon\rangle_{\partial\Omega\setminus\Gamma^-}$ can be proven to be of order $2\,k+1$; see [1] for the case $c=0$ and $\Upsilon = 1$. This result holds for functions Υ that are independent of the mesh; as a consequence, Υ cannot be taken to have support in a single element face. To the knowledge of the authors, Theorem 2.4 is the only result that allows this.

Next, we give a proof of the result. To do that, we need to introduce some notation. We begin by introducing the so-called corresponding dual problem, namely,

$$-\boldsymbol{\beta}\cdot\nabla\psi + c\,\psi = 0 \qquad \text{in } \Omega_e, \tag{2.7a}$$

$$\psi = \varphi/\boldsymbol{\beta}\cdot\boldsymbol{n} \qquad \text{on } e. \tag{2.7b}$$

We are also going to use two auxiliary projections. We set $\mathbb{P}^+ := \mathbb{P}$, which is defined in (2.3), and we let $\mathbb{P}^-$ be the projection which satisfies (2.3a) and imposes (2.3b) on e_K^- instead of e_K^+.

We are now ready to prove Theorem 2.4.

Proof. To prove this result, we begin by noting that

$$\langle \mathsf{P}_\partial u - \widehat{u}_h, \varphi\rangle_e = \langle \mathsf{P}_\partial u - \widehat{u}_h, \psi\boldsymbol{\beta}\cdot\boldsymbol{n}\rangle_e,$$

by the boundary condition of the dual problem, (2.7b), and that

$$\langle \mathsf{P}_\partial u - \widehat{u}_h, \varphi\rangle_e = \langle \mathsf{P}_\partial u - \widehat{u}_h, \psi\boldsymbol{\beta}\cdot\boldsymbol{n}\rangle_{\partial\Omega_e\setminus\Gamma^-},$$

by our assumptions on the triangulation $\mathcal{T}_h$. Then

$$\begin{aligned}\langle \mathsf{P}_\partial u - \widehat{u}_h, \varphi\rangle_e &= \langle \mathsf{P}_\partial u - \widehat{u}_h, (\psi - \mathbb{P}^-\psi)\,\boldsymbol{\beta}\cdot\boldsymbol{n}\rangle_{\partial\Omega_e\setminus\Gamma^-}\\ &\quad + \langle \mathsf{P}_\partial u - \widehat{u}_h, \mathbb{P}^-\psi\,\boldsymbol{\beta}\cdot\boldsymbol{n}\rangle_{\partial\Omega_e\setminus\Gamma^-}.\end{aligned}$$

Since we can rewrite the error equation (2.2) in terms of the projections $\mathbb{P}^+$ and P_∂ as

$$-(\mathbb{P}^+u - u_h, \partial_{\boldsymbol{\beta}} v)_{\Omega_e} + \langle \mathsf{P}_\partial u - \widehat{u}_h, v\,\boldsymbol{\beta}\cdot\boldsymbol{n}\rangle_{\partial\Omega_e\setminus\Gamma^-} + (c\,(u-u_h), v)_{\Omega_e} = 0,$$

taking $v := \mathbb{P}^- \psi$, we get that

$$\begin{aligned}
\langle \mathsf{P}_\partial u - \widehat{u}_h, \varphi \rangle_e = \; & \langle \mathsf{P}_\partial u - \widehat{u}_h, (\psi - \mathbb{P}^- \psi)\, \boldsymbol{\beta} \cdot \boldsymbol{n} \rangle_{\partial \Omega_e \setminus \Gamma^-} \\
& + (\mathbb{P}^+ u - u_h, \partial_{\boldsymbol{\beta}} \mathbb{P}^- \psi)_{\Omega_e} \\
& - (c\,(u - u_h), \mathbb{P}^- \psi)_{\Omega_e} \\
= \; & \langle \mathsf{P}_\partial u - \widehat{u}_h, (\psi - \mathbb{P}^- \psi)\, \boldsymbol{\beta} \cdot \boldsymbol{n} \rangle_{\partial \Omega_e \setminus \Gamma^-} \\
& + (\mathbb{P}^+ u - u_h, \partial_{\boldsymbol{\beta}} \mathbb{P}^- \psi)_{\Omega_e} \\
& - (c\,(\mathbb{P}^+ u - u_h), \mathbb{P}^- \psi)_{\Omega_e} \\
& - (c\,(u - \mathbb{P}^+ u), \mathbb{P}^- \psi)_{\Omega_e}.
\end{aligned}$$

Hence, by the dual equation (2.7a), we can write

$$\langle \mathsf{P}_\partial u - \widehat{u}_h, \varphi \rangle_e = \sum_{i=1}^{4} T_i,$$

where

$$\begin{aligned}
T_1 := \; & \langle \mathsf{P}_\partial u - \widehat{u}_h, (\psi - \mathbb{P}^- \psi)\, \boldsymbol{\beta} \cdot \boldsymbol{n} \rangle_{\partial \Omega_e \setminus \Gamma^-}, \\
T_2 := \; & (\mathbb{P}^+ u - u_h, \partial_{\boldsymbol{\beta}} (\mathbb{P}^- \psi - \psi))_{\Omega_e}, \\
T_3 := \; & - (c\,(\mathbb{P}^+ u - u_h), \mathbb{P}^- \psi - \psi)_{\Omega_e}, \\
T_4 := \; & - (c\,(u - \mathbb{P}^+ u), \mathbb{P}^- \psi)_{\Omega_e}.
\end{aligned}$$

Let us estimate the terms $T_i, i = 1, 2, 3, 4$. We begin by showing that $T_1 + T_2 = 0$. After a simple integration by parts, we get that

$$\begin{aligned}
T_2 = & - \langle \mathbb{P}^+ u - u_h, (\psi - \mathbb{P}^- \psi)\, \boldsymbol{\beta} \cdot \boldsymbol{n} \rangle_{\partial \Omega_e} + (\partial_{\boldsymbol{\beta}} (\mathbb{P}^+ u - u_h), \psi - \mathbb{P}^- \psi)_{\Omega_e} \\
= & - \langle \mathbb{P}^+ u - u_h, (\psi - \mathbb{P}^- \psi)\, \boldsymbol{\beta} \cdot \boldsymbol{n} \rangle_{\partial \Omega_e}
\end{aligned}$$

by the orthogonality property of the projection $\mathbb{P}^-$, (2.3a). Hence,

$$\begin{aligned}
T_1 + T_2 = & \langle \mathsf{P}_\partial u - \widehat{u}_h, (\psi - \mathbb{P}^- \psi)\, \boldsymbol{\beta} \cdot \boldsymbol{n} \rangle_{\partial \Omega_e \setminus \Gamma^-} - \langle \mathbb{P}^+ u - u_h, (\psi - \mathbb{P}^- \psi)\, \boldsymbol{\beta} \cdot \boldsymbol{n} \rangle_{\partial \Omega_e} \\
= & \; 0
\end{aligned}$$

since, by the definition of $\mathbb{P}^+$ and $\widehat{u}_h$, we have that $\mathsf{P}_\partial u = \mathbb{P}^+ u$ and $\widehat{u}_h = u_h$ on any outflow face, and since, by the definition of $\mathbb{P}^-$, we have that

$$\langle \omega, (\psi - \mathbb{P}^- \psi)\, \boldsymbol{\beta} \cdot \boldsymbol{n} \rangle_{\tilde{e}} = 0 \quad \text{for all } \omega \in \mathcal{P}^k(\tilde{e})$$

on any inflow face $\tilde{e}$.

It remains to estimate T_3 and T_4. To do this, we are going to use the following stability estimate for the solution ψ of the dual problem (2.7):

$$|\psi|_{H^s(\Omega_e)} \le C\|\varphi\|_{H^s(e)}/\boldsymbol{\beta}\cdot\boldsymbol{n}, \tag{2.8}$$

where $C = C(\|c\|_{W^{\max\{s,0\},\infty}(\Omega_e)}, \Omega_e)$. It follows from the expression

$$\psi(x) = \varphi(x_0)e^{-\int_0^t c(x_0-\tau\boldsymbol{\beta}))d\tau}/\boldsymbol{\beta}\cdot\boldsymbol{n},$$

where $x = x_0 - \boldsymbol{\beta}\, t \in \Omega_e$ for some $t \ge 0$ and $x_0 \in e$, by successive differentiation.

The estimate of T_3 follows after a straightforward application of the Cauchy–Schwarz inequality, the estimate of Theorem 2.2, the approximation property of the projection $\mathbb{P}^-$, (2.4), and the stability estimate (2.8). Indeed, we easily get that

$$T_3 \le \mathsf{C}(c,s)\, h^{k+s+1}|\, u\,|_{H^{k+1}(\Omega_e)}\|\varphi\|_{H^s(e)}$$

for $s \in [0, k+1]$. Let us estimate T_4. If we denote by P^{k-1} the L^2-projection onto V_h^{k-1}, we have that

$$\begin{aligned} T_4 &= (c\,(u-\mathbb{P}^+u), \psi - \mathbb{P}^-\psi)_{\Omega_e} - (u-\mathbb{P}^+u, c\psi)_{\Omega_e} \\ &= (c\,(u-\mathbb{P}^+u), \psi - \mathbb{P}^-\psi)_{\Omega_e} - (u-\mathbb{P}^+u, c\psi - \mathsf{P}^{k-1}c\psi)_{\Omega_e}, \end{aligned}$$

by the orthogonality property of the projection $\mathbb{P}^+$, (2.3a). Then we get

$$T_4 \le\ \mathsf{C}(c,s)\, h^{k+s+1}\|\varphi\|_{H^s(e)}|\, u\,|_{H^{k+1}(\Omega_e)}$$

for $s \in [0,k]$, by the approximation properties of the projections $\mathbb{P}^+$, $\mathbb{P}^-$, and P^{k-1} and the stability estimate (2.8). This completes the proof. ∎

3. Extensions.

3.1. The variable-degree DG method. The variable-degree DG approximation u_h belongs to the finite dimensional space $V_h^{\boldsymbol{k}}$ which is given by

$$V_h^{\boldsymbol{k}} = \{v \in L^2(\Omega), v|_K \in \mathcal{P}^{k(K)}(K) \quad \text{for all } K \in \mathcal{T}_h\},$$

where $\boldsymbol{k} = \{k(K) \text{ for all } K \in \mathcal{T}_h\}$.

The definition of $\mathbb{P}$ can easily be generalized for the variable-degree space. We define $\mathbb{P}u \in V_h^{\boldsymbol{k}}$ using the definition of the previous section but with the space $\mathcal{P}^{k-1}(K)$ replaced with $\mathcal{P}^{k(K)-1}(K)$ and the space $\mathcal{P}^k(e_K^+)$ replaced with $\mathcal{P}^{k(K)}(e_K^+)$.

We need to impose a condition on the space $V_h^{\boldsymbol{k}}$:

$$k(K) \ge k(K') \text{ whenever } e_K^+ \subseteq e_{K'}^-. \tag{3.1}$$

The following result can be proven in the same way as Theorem 2.2.

THEOREM 3.1. *Let $u_h \in V_h^{\boldsymbol{k}}$ be the variable degree DG approximation given by* (2.1) *with V_h^k replaced with $V_h^{\boldsymbol{k}}$. If $\mathcal{T}_h$ satisfies the flow conditions* (1.3) *with respect to β and the shape-regularity condition* (1.4)*, and $V_h^{\boldsymbol{k}}$ satisfies the condition of the polynomial degrees* (3.1)*, then the results of Theorem* 2.2 *hold for the variable-degree DG method.*

3.2. Singularly perturbed problem. We present local error estimates of a DG method for the following singular perturbed problem in $\Omega \subset \mathbb{R}^2$:

$$-\epsilon \triangle u + \boldsymbol{\beta} \cdot \nabla u + c\, u = f \quad \text{in } \Omega, \tag{3.2a}$$

$$u = g \quad \text{on } \partial\Omega. \tag{3.2b}$$

We use the interior penalty method to discretize the viscosity term $-\epsilon\triangle u$; see [2] and [10]. Any other DG method which is consistent and stable could be used to discretize this term; see [3].

Thus, the DG approximation $u_h \in V_h^k$, where $k \geq 1$, of (3.2) solves

$$\epsilon A(u_h, v_h) + B(u_h, v_h) = F(v_h) \quad \text{for all } v_h \in V_h^k, \tag{3.3}$$

where

$$A(\omega, v) = (\nabla\omega, \nabla v)_{\mathcal{T}_h} - \sum_{e\in\mathcal{E}_h} \Big(\langle \{\!\{\nabla\omega\}\!\}, [\![v]\!] \rangle_e + \langle \{\!\{\nabla v\}\!\}, [\![\omega]\!] \rangle_e - \frac{\eta}{h} \langle [\![\omega]\!], [\![v]\!] \rangle_e \Big),$$

$$F(v) = (f, v)_{\mathcal{T}_h} + \langle g, v\boldsymbol{\beta}\cdot\boldsymbol{n}\rangle_{\partial\Omega} + \epsilon \langle \nabla v \cdot \boldsymbol{n}, g\rangle_{\partial\Omega} - \epsilon \frac{\eta}{h} \langle g, v\rangle_{\partial\Omega}.$$

The set $\mathcal{E}_h$ is the collection of edges of the triangulation $\mathcal{T}_h$. The parameter η is large enough in order to ensure stability.

The average $\{\!\{\cdot\}\!\}$ and jump $[\![\cdot]\!]$ operators are defined as follows: For an interior edge e, we set

$$\{\!\{\boldsymbol{q}\}\!\} = \frac{1}{2}(\boldsymbol{q}_1 + \boldsymbol{q}_2), \qquad [\![\boldsymbol{q}]\!] = \boldsymbol{q}_1 \cdot \boldsymbol{n}_1 + \boldsymbol{q}_2 \cdot \boldsymbol{n}_2,$$

$$\{\!\{\varphi\}\!\} = \frac{1}{2}(\varphi_1 + \varphi_2), \qquad [\![\varphi]\!] = \varphi_1\boldsymbol{n}_1 + \varphi_2\boldsymbol{n}_2,$$

where $e = K_1 \cap K_2$, $\boldsymbol{q}_i = \boldsymbol{q}|_{K_i}$, $\varphi_i = \varphi|_{K_i}$, and $\boldsymbol{n}_i$ is the exterior unit normal to K_i, $i = 1, 2$. For a boundary edge e, we set

$$\{\!\{\boldsymbol{q}\}\!\} = \boldsymbol{q}, \quad [\![\varphi]\!] = \varphi\boldsymbol{n},$$

where $\boldsymbol{n}$ is the outward unit normal. The quantities $[\![\boldsymbol{q}]\!]$ and $\{\!\{\varphi\}\!\}$ on boundary edges are not required, so they are left undefined.

In order to state the result, we need to introduce a proper weight function. For simplicity we assume $\boldsymbol{\beta} = (1, 0)$. Accordingly, we set, for fixed x_0, y_1, and y_2,

$$\Omega_0 = \big((-\infty, x_0] \times [y_1, y_2]\big) \cap \Omega,$$

and construct a function ω satisfying

$$C_1 \leq \omega(x, y) \leq C_2 \qquad \text{for } (x, y) \in \Omega_0,$$

$$|\omega(x, y)| \leq C_2 e^{-(x-x_0)/M\rho} \qquad \text{for } x \geq x_0 + h,$$

$$|\omega(x, y)| \leq C_2 e^{-(y-y_2)/M\sigma} \qquad \text{for } y \geq y_2 + h,$$

$$|\omega(x, y)| \leq C_2 e^{-(y_1-y)/M\sigma} \qquad \text{for } y \leq y_1 - h.$$

Here $\rho \geq 0$, $\sigma \geq 0$ are parameters that will depend on the mesh size h and ϵ. We say that ρ is the size of the upwind layer and σ is the size of the crosswind layer. The positive constants C_1, C_2, and M are fixed.

THEOREM 3.2. *Assume that $\epsilon \leq h$ and that the triangulation $\mathcal{T}_h$ is quasi-uniform; that is, assume that there is a parameter $\kappa > 0$ such that*

$$\max_{K\in\mathcal{T}_h}\{h_K\} \leq \kappa \min_{K\in\mathcal{T}_h}\{h_K\}.$$

If $\mathcal{T}_h$ satisfies the flow conditions (1.3) *with respect to $\boldsymbol{\beta}$. Then the error between u_h given by* (3.3) *and u given by* (3.2) *is*

$$\|\omega(u-u_h)\|_{L^2(\mathcal{T}_h)} \leq C\bigg(\left(1+\frac{\epsilon^{1/2}}{h}\right)\|\omega(u-\mathbb{P}u)\|_{L^2(\mathcal{T}_h)} + \epsilon^{1/2}\|\omega\nabla(u-\mathbb{P}u)\|_{L^2(\mathcal{T}_h)}$$

$$+ h\epsilon^{1/2}\|\omega D^2(u-\mathbb{P}u)\|_{L^2(\mathcal{T}_h)}\bigg),$$

where ω is given above with $\rho = \log(\frac{1}{h})h$, $\sigma = h^{1/2}$, and M is a sufficiently large fixed constant. Moreover, if $\mathcal{T}_h$ also satisfies the flow conditions (1.3) *with respect to $-\boldsymbol{\beta}$, then we can choose $\sigma = \max(\epsilon^{1/2}, h)$.*

From this result, we immediately get optimal weighted error estimates if $\epsilon \leq h^2$. The size of the crosswind layer σ is typically $\sigma = h^{1/2}$ for general triangulations; see [11]. However, we see that the size of the crosswind layer is reduced to $\sigma = h$ if the flow conditions (1.3) for both $\boldsymbol{\beta}$ and $-\boldsymbol{\beta}$ are satisfied. This is exactly the result that was obtained for the streamline-diffusion method in [17]. However, in [17] the following almost-uniform condition on the mesh was imposed: two adjacent elements K and K' that share an edge that is not aligned with $\boldsymbol{\beta}$ must satisfy

$$|h_{\boldsymbol{\beta},K} - h_{\boldsymbol{\beta},K'}| \leq Ch_{\boldsymbol{\beta}}^2,$$

where $h_{\boldsymbol{\beta},K}$ is the length of the edge of K that is parallel to $\boldsymbol{\beta}$ and $h_{\boldsymbol{\beta}} = \max_K h_{\boldsymbol{\beta},K}$; see the inequality (3.8) in [17]. In contrast, in our results, we assume only quasi uniformity. Moreover, our results also hold for high-order elements.

The proof of Theorem 3.2 is very similar to the proof of the local estimates given in [11]. However, instead of using the L^2-projection, one must use the projections $\mathbb{P} = \mathbb{P}^+$ and $\mathbb{P}^-$ used in this paper.

4. Numerical results. In this section we present numerical experiments which validate our theoretical results. The domain is $\Omega = (-.5, .5) \times (-.5, .5)$. The coefficients are $\boldsymbol{\beta} \equiv (1, 0)$ and $c \equiv 1$. We choose the right-hand side f so that the solution is $u(x, y) = (x + 1/2)\sin(x)\sin(y)$.

We start with a uniform mesh of size h which satisfies the flow condition with respect to $\boldsymbol{\beta}$; see Figure 4.1, left. We then perturb the coordinates of the interior nodes randomly by at most $\frac{2h}{5}$ in such a way that the resulting mesh is *no longer uniform but still satisfies the flow condition with respect to $\boldsymbol{\beta}$*; see Figure 4.1, right. A mesh ℓ is a perturbation of a uniform mesh of size $h = \frac{1}{2^\ell}$.

In Table 4.1 we display the error and orders of convergence for approximate solutions using polynomials of degree $k = 0, 1, 2$. We see that order $k + 1$ is observed for $\|u - u_h\|_{L^2(\Omega)}$ and $\|\partial_{\boldsymbol{\beta}}u - \partial_{\boldsymbol{\beta},h}u_h\|_{L^2(\Omega)}$ as the theory predicts. Moreover, we see that $\max_{e\in\mathcal{E}_h} |\text{avg}_e(u - \widehat{u}_h)|$ converges with order $2k + 1$ as expected.

FIG. 4.1. *Meshes satisfying the flow condition with respect to* $\boldsymbol{\beta} = (1, 0)$*: Uniform mesh* ($h = \frac{1}{2^3}$) *(left) and its nonuniform perturbation* ($\ell = 3$) *(right).*

TABLE 4.1
History of convergence.

	Mesh	$\|u - u_h\|_{L^2(\Omega)}$		$\|\partial_{\boldsymbol{\beta}} u - \partial_{\boldsymbol{\beta},h} u_h\|_{L^2(\Omega)}$		$\max_{e \in \mathscr{E}_h} \|\text{avg}_e(u - \widehat{u}_h)\|$	
k	ℓ	error	order	error	order	error	order
0	1	.27e-1	-	.11e-0	-	.16e-1	-
	2	.19e-1	0.50	.54e-1	1.05	.88e-2	0.83
	3	.11e-1	0.80	.29e-1	0.93	.61e-2	0.54
	4	.58e-2	0.92	.15e-1	0.92	.35e-2	0.82
	5	.29e-2	1.01	.75e-2	1.02	.16e-2	1.11
	6	.15e-2	0.98	.38e-2	0.97	.82e-3	0.96
	7	.74e-3	0.99	.19e-2	1.00	.42e-3	0.98
1	1	.12e-1	-	.27e-1	-	.82e-3	-
	2	.34e-2	1.81	.69e-2	1.98	.92e-4	3.16
	3	.86e-3	1.97	.15e-2	2.17	.18e-4	2.36
	4	.23e-3	1.90	.43e-3	1.84	.25e-5	2.83
	5	.58e-4	1.99	.11e-3	2.01	.30e-6	3.08
	6	.15e-4	1.95	.28e-4	1.94	.44e-7	2.74
	7	.38e-5	1.99	.69e-5	2.00	.60e-8	2.90
2	1	.16e-2	-	.14e-2	-	.49e-5	-
	2	.19e-3	3.10	.17e-3	3.08	.53e-6	3.20
	3	.26e-4	2.88	.24e-4	2.82	.20e-7	4.74
	4	.33e-5	2.99	.30e-5	3.00	.66e-9	4.92
	5	.43e-6	2.92	.40e-6	2.91	.25e-10	4.74
	6	.54e-7	2.99	.51e-7	2.99	.93e-12	4.73

5. Concluding remarks. This paper contains the first instance in which the approximation error $\| u - u_h \|_{L^2(\mathcal{T}_h)}$, where u is the solution of the transport-reaction equation and u_h is given by the original DG method, is proven to be optimal in the mesh size h *and* in the regularity of the exact solution. Unexpectedly, this happens with meshes whose main feature is to be, roughly speaking, *aligned* with the flow.

The fact that the direction of the flow $\boldsymbol{\beta}$ is a constant and the use of simplexes to define the triangulations $\mathcal{T}_h$ seem to play a major role in the result. The case of variable $\boldsymbol{\beta}$ and simplexes with curved boundaries is the subject of ongoing research.

Appendix. The construction of triangulations satisfying the flow conditions. Let us show that it is always possible to construct a triangulation of the domain $\Omega \subset \mathbb{R}^d$ satisfying the flow conditions (1.3).

We can do this as follows. First, we triangulate the inflow border of Ω, Γ^-, by using $(d-1)$-dimensional simplexes T; let $\Gamma_h^- := \{T\}$ be such a triangulation. Next, for each T, we construct the d-dimensional *prism*

$$\mathcal{T}_T := (T \oplus \{\lambda\, \boldsymbol{\beta} : \lambda \in \mathbb{R}\}) \cap \Omega,$$

which we are going to triangulate by using conforming d-dimensional simplexes. Note that the triangulations of the prisms are completely independent of each other since no conformity between them is required thanks to the flow conditions (1.3).

To triangulate the prism $\mathcal{T}_T$, we have only to show that, if T and T' are two $(d-1)$-dimensional simplexes whose vertices are on the border of $\mathcal{T}_T$, we can triangulate the convex hull of T and T', $CH(T, T')$, by using exactly d d-dimensional simplexes, each of which has one inflow face and one outflow face. Those simplexes are $CH(T_{i-1}, T_i)$, $i = 1, \ldots, d$, where the $(d-1)$-dimensional simplexes T_i, $i = 0, \ldots, d$, are constructed in such a way that $T_0 := T$, $T_d := T'$, and, assuming that T is an inflow face, that T_{i-1} is the only inflow face of the simplex $CH(T_{i-1}, T_i)$ and T_i is the only outflow face of the simplex $CH(T_{i-1}, T_i)$.

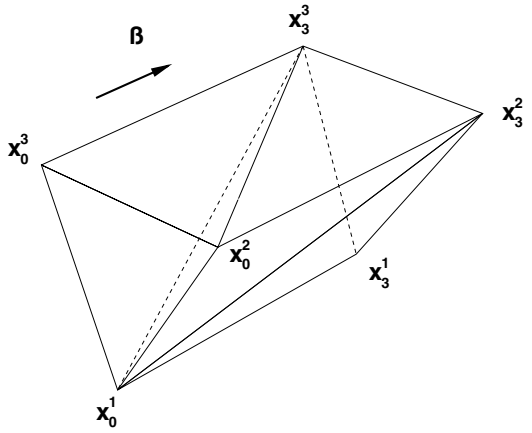


FIG. A.1. *Detail of the construction of a three-dimensional triangulation satisfying the flow conditions.*

If we identify the $(d-1)$-dimensional simplex T_i with its set of vertices, $\{\boldsymbol{x}_i^j\}_{j=1}^d$,

it is not difficult to see that we can take

$$T_0 \equiv \{\boldsymbol{x}_0^1, \boldsymbol{x}_0^2, \ldots, \boldsymbol{x}_0^{d-2}, \boldsymbol{x}_0^{d-1}, \boldsymbol{x}_0^d\},$$

$$T_1 \equiv \{\boldsymbol{x}_0^1, \boldsymbol{x}_0^2, \ldots, \boldsymbol{x}_0^{d-2}, \boldsymbol{x}_0^{d-1}, \boldsymbol{x}_d^d\},$$

$$T_2 \equiv \{\boldsymbol{x}_0^1, \boldsymbol{x}_0^2, \ldots, \boldsymbol{x}_0^{d-2}, \boldsymbol{x}_d^{d-1}, \boldsymbol{x}_d^d\},$$

$$T_3 \equiv \{\boldsymbol{x}_0^1, \boldsymbol{x}_0^2, \ldots, \boldsymbol{x}_d^{d-2}, \boldsymbol{x}_d^{d-1}, \boldsymbol{x}_d^d\},$$

$$\ldots$$

$$T_d \equiv \{\boldsymbol{x}_d^1, \boldsymbol{x}_d^2, \ldots, \boldsymbol{x}_d^{d-2}, \boldsymbol{x}_d^{d-1}, \boldsymbol{x}_d^d\},$$

respectively. An example of this construction is illustrated in Figure A.1.

REFERENCES

[1] S. Adjerid and T. C. Massey, *A posteriori discontinuous finite element error estimation for two-dimensional hyperbolic problems*, Comput. Methods Appl. Mech. Engrg., 191 (2002), pp. 5877–5897.

[2] D. N. Arnold, *An interior penalty finite element method with discontinuous elements*, SIAM J. Numer. Anal., 19 (1982), pp. 742–760.

[3] D. N. Arnold, F. Brezzi, B. Cockburn, and L. D. Marini, *Unified analysis of discontinuous Galerkin methods for elliptic problems*, SIAM J. Numer. Anal., 39 (2002), pp. 1749–1779.

[4] P. Bastian and B. Rivière, *Superconvergence and $H(\mathrm{div})$ projection for discontinuous Galerkin methods*, Internat. J. Numer. Methods Fluids, 42 (2003), pp. 1043–1057.

[5] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer-Verlag, New York, 1991.

[6] P. Ciarlet, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1978.

[7] B. Cockburn, B. Dong, and J. Guzmán, *A superconvergent LDG-hybridizable Galerkin method for second-order elliptic problems*, Math. Comp., accepted.

[8] B. Cockburn, G. Kanschat, and D. Schötzau, *A locally conservative LDG method for the incompressible Navier-Stokes equations*, Math. Comp., 74 (2005), pp. 1067–1095.

[9] B. Cockburn, D. Schötzau, and J. Wang, *Discontinuous Galerkin methods for incompressible elastic materials*, Comput. Methods Appl. Mech. Engrg., 195 (2006), pp. 3184–3204.

[10] J. Gopalakrishnan and G. Kanschat, *A multilevel discontinuous Galerkin method*, Numer. Math., 95 (2003), pp. 527–550.

[11] J. Guzmán, *Local analysis of discontinuous Galerkin methods applied to singularly perturbed problems*, J. Numer. Math., 14 (2006), pp. 41–56.

[12] C. Johnson and J. Pitkäranta, *An analysis of the discontinuous Galerkin method for a scalar hyperbolic equation*, Math. Comp., 46 (1986), pp. 1–26.

[13] P. Lesaint and P.-A. Raviart, *On a finite element method for solving the neutron transport equation*, in Mathematical Aspects for Finite Elements in Partial Differential Equations, C. de Boor, ed., Academic Press, New York, 1974, pp. 89–123.

[14] T. E. Peterson, *A note on the convergence of the discontinuous Galerkin method for a scalar hyperbolic equation*, SIAM J. Numer. Anal., 28 (1991), pp. 133–140.

[15] W. Reed and T. Hill, *Triangular Mesh Methods for the Neutron Transport Equation*, Technical report LA-UR-73-479, Los Alamos National Laboratory, Los Alamos, NM, 1973.

[16] G. R. Richter, *An optimal-order error estimate for the discontinuous Galerkin method*, Math. Comp., 50 (1988), pp. 75–88.

[17] G. Zhou and R. Rannacher, *Pointwise superconvergence of the streamline diffusion finite-element method*, Numer. Methods Partial Differential Equations, 12 (1996), pp. 123–145.

SIAM J. NUMER. ANAL.
Vol. 46, No. 3, pp. 1266–1280
© 2008 Society for Industrial and Applied Mathematics

IMPLICIT RUNGE–KUTTA METHODS FOR LIPSCHITZ CONTINUOUS ORDINARY DIFFERENTIAL EQUATIONS*

XIAOJUN CHEN† AND SAYED MAHMOUD‡

Abstract. Implicit Runge–Kutta (IRK) methods for solving the nonsmooth ordinary differential equation (ODE) involve a system of nonsmooth equations. We show superlinear convergence of the slanting Newton method for solving the system of nonsmooth equations. We prove the slanting differentiability and give a slanting function for the involved function. We develop a new code based on the slanting Newton method and the IRK method for nonsmooth ODEs arising from structural oscillation and pounding. We show that the new code is efficient for solving a nonsmooth ODE model for the collapse of the Tacoma Narrows suspension bridge and simulating 13 different earthquakes.

Key words. implicit Runge–Kutta method, slanting Newton method, nonsmooth equations, ordinary differential equation

AMS subject classifications. 65H10, 65L05, 65L06

DOI. 10.1137/070688699

1. Introduction.

Let $D \subset R^n$ be a domain, $p : D \to R^n$ be a continuously differentiable function, and $f : D \to R^n$ be a continuous function. The function f is not necessarily differentiable.

We consider the following system of nonsmooth ordinary differential equations of the first order:

$$\begin{aligned} &\frac{dp(u)}{dt} = f(u), \qquad t \geq 0, \\ &u(0) = u_0, \end{aligned} \tag{1.1}$$

where $u \in R^n$. Here smoothness refers to continuous differentiability. It is known that if f is Lipschitz continuous, then (1.1) has a unique solution in a certain interval $[0, T]$ [9]. Moreover, it is shown in [2] that f is Lipschitz continuous if and only if f is slantly differentiable. A slanting Newton method can be defined in appropriate function spaces. The concept of slanting differentiability of f means that f has a slanting function f^o such that

$$f(u + \triangle u) - f(u) - f^o(u + \triangle u)\triangle u = o(\|\triangle u\|).$$

This property ensures that the slanting Newton method

$$u^{k+1} = u^k - f^o(u^k)^{-1} f(u^k)$$

has a superlinear convergence rate [2].

Note that, though the function f is nonsmooth, the solution of (1.1) may be smooth (continuously differentiable). See the example of the collapse of the Tacoma

*Received by the editors April 19, 2007; accepted for publication (in revised form) November 2, 2007; published electronically March 21, 2008. This work was partly supported by a Grant-in-Aid from the Japan Society for the Promotion of Science and a scholarship from the Egyptian Government.

http://www.siam.org/journals/sinum/46-3/68869.html

†Department of Applied Mathematics, The Hong Kong Polytechnic University, Hung Hom Kowloon, Hong Kong (maxjchen@polyu.edu.hk).

‡Department of Mathematical Sciences, Hirosaki University, Hirosaki 036-8561, Japan (elseedy@hotmail.com).

Narrows suspension bridge in section 4. In this paper, we consider numerical solution of the nonsmooth ODEs (1.1).

One step of an s-stage implicit Runge–Kutta (IRK) method for solving (1.1) has the following version [8].

IRK method. Given a step size h, a coefficient matrix $A \in R^{s\times s}$, and a weight vector $b \in R^s$, let $U_0 = u_0$. For $k \geq 0$ do the following.

Step 1. Solve the $s \times n$-dimensional system of nonlinear equations

$$H(x) := \begin{pmatrix} p(x_1) - p(U_k) - h\sum_{j=1}^{s} a_{1j} f(x_j) \\ \vdots \\ p(x_s) - p(U_k) - h\sum_{j=1}^{s} a_{sj} f(x_j) \end{pmatrix} = 0 \tag{1.2}$$

to get a solution $x^k = (x_1^k, x_2^k, \ldots, x_s^k)^T \in R^{s\times n}$.

Step 2. Solve the n-dimensional system of nonlinear equations

$$\tilde{H}(U) := p(U) - p(U_k) - h\sum_{j=1}^{s} b_j f(x_j^k) = 0, \tag{1.3}$$

and let the solution be U_{k+1}.

A practical IRK method can be defined, by choosing appropriate matrix A and vector b, such as coefficients of Gauss, Radau IA & IIA, Lobatto IIIA, Burrage, etc. [1, 8]. Recently, much attention has been paid to choosing A, b such that the IRK method has the best properties in some sense of stability [5]. Moreover, Jay [8] showed that for various choices of A, b the use of inexact simplified Newton methods is efficient for solving the system of nonlinear equations (1.2) under the condition that H is continuously differentiable. In this paper, we focus our attention on how to solve the system of nonlinear equations (1.2) efficiently when f is not differentiable. Such nonsmooth problems arise from mathematical models of structural pounding earthquake and structural oscillations [7, 10, 11, 14]. We apply the slanting Newton method [2,6] to solve (1.2). In section 2, we show that H is slantly differentiable if f is piecewise continuously differentiable. Moreover, we give a simple method to compute a slanting function for H. In section 3, we discuss the convergence order of IRK methods for nonsmooth ODEs. In section 4, we illustrate the slanting Newton method by using a simple model for the collapse of the Tacoma Narrows suspension bridge [14]. Furthermore, we develop a code based on the slanting Newton method and the IRK method for nonsmooth ODEs arising from structural oscillation and pounding. A suite of 27 ground motion records from 12 different earthquakes and a record from the Kobe earthquake are used to show that the code is efficient for simulation of a structural pounding earthquake. All data used in the numerical experiments were taken from the PEER Strong Motion Database [15].

2. Slanting Newton iterations for IRK methods. Let X and Y be Banach spaces and D be an open domain in X. Let $L(X, Y)$ denote the set of all bounded linear operators on X into Y.

DEFINITION 2.1 (see [2]). *A function $H : D \subset X \to Y$ is said to be slantly differentiable at $x \in D$ if there exists a mapping $H^o : D \to L(X, Y)$ such that the*

family $\{H^o(x+\triangle x)\}$ *of bounded linear operators is uniformly bounded in the operator norm for* h *sufficiently small and*

$$\lim_{\triangle x\to 0}\frac{H(x+\triangle x)-H(x)-H^o(x+\triangle x)\triangle x}{\|\triangle x\|}=0.$$

The function H^o *is called a slanting function for* H *at* x.

DEFINITION 2.2 (see [2, 4]). *A function* $H: X\to Y$ *is said to be Lipschitz continuous at* x *if there is a constant* K *such that for all sufficiently small* $\triangle x$

$$\|H(x+\triangle x)-H(x)\|\le K\|\triangle x\|.$$

LEMMA 2.1 (see [2]). *A function* $H: X\to Y$ *is slantly differentiable at* x *if and only if* H *is Lipschitz continuous at* x.

LEMMA 2.2 (see [2]). *Suppose that* H *is slantly differentiable at a solution* x^* *of* $H(x)=0$. *Let* H^o *be a slanting function for* H *at* x^* *and* $\|H^o(x)^{-1}\|\le\Gamma$ *in a neighborhood* $\mathcal{N}$ *of* x^*, *where* Γ *is a positive constant. Then the iterative sequence* $\{x^m\}$ *generated by the slanting Newton method*

$$x^{m+1}=x^m-H^o(x^m)^{-1}H(x^m) \tag{2.1}$$

superlinearly converges to x^* *in a neighborhood of* x^*.

DEFINITION 2.3 (see [12]). *We say* $f: R^n\to R^n$ *is piecewise continuously differentiable if it is continuous and there is a finite collection of continuously differentiable functions* $\phi_l: R^n\to R, l\in\mathcal{L}:=\{1,2,\dots,L\}$, *such that*

$$f_i(u)\in\{\phi_l(u),\ l\in\mathcal{L}\},\ \textit{for any } u\in R^n,\quad i\in\{1,2,\dots,n\}.$$

Such a collection is called a representation for f *on* R^n.

THEOREM 2.1. *The function* $H: R^{s\times n}\to R^{s\times n}$ *defined in* (1.2) *is slantly differentiable at any point* $x\in R^{s\times n}$ *if* f *is piecewise continuously differentiable.*

Proof. We first show that, for a fixed point $\bar u\in R^n$ and a fixed index $i\in\{1,2,\dots,n\}$, f_i is Lipschitz continuous at $\bar u\in R^n$. Let $\mathcal{I}$ and $\mathcal{J}$ be two index sets defined by

$$\mathcal{I}=\{l\,|\,f_i(\bar u)=\phi_l(\bar u),\ l\in\mathcal{L}\}$$

and

$$\mathcal{J}=\{i\,|\,f_i(\bar u)\ne\phi_l(\bar u),\ l\in\mathcal{L}\},$$

respectively. Since $\mathcal{J}$ is a finite set, we have

$$r=\min_{l\in\mathcal{J}}\|\phi_l(\bar u)-f_i(\bar u)\|>0. \tag{2.2}$$

Since f_i is continuous in R^n, there are positive constants δ_0 and $\alpha<1$ such that if $\|u-\bar u\|<\delta_0$, then

$$\|f_i(u)-f_i(\bar u)\|\le\alpha r. \tag{2.3}$$

Furthermore, since every $\phi_l, l\in\mathcal{L}$, is continuously differentiable, there are positive constants K_l, δ_l such that if $\|u-\bar u\|<\delta_l$, then

$$\|\phi_l(u)-\phi_l(\bar u)\|\le K_l\|u-\bar u\|,\quad l\in\mathcal{L}. \tag{2.4}$$

Let

$$K = \max_{l\in\mathcal{L}} K_l \quad \text{and} \quad \delta = \min\left\{\delta_0, \frac{1-\alpha}{K}r, \min_{l\in\mathcal{L}}\delta_l\right\}.$$

For any u satisfying $\|u-\bar{u}\| < \delta$, if $f_i(u) = \phi_l(u), l \in \mathcal{I}$, then we have

$$\|f_i(u) - f_i(\bar{u})\| = \|\phi_l(u) - \phi_l(\bar{u})\| \le K\|u-\bar{u}\|.$$

Otherwise, if $f_i(u) = \phi_j(u), j \in \mathcal{J}$, we apply (2.2)–(2.4) and obtain

$$\begin{aligned}\alpha r &\ge \|f_i(u) - f_i(\bar{u})\| \\ &= \|\phi_j(u) - f_i(\bar{u})\| \\ &\ge \|\phi_j(\bar{u}) - f_i(\bar{u})\| - \|\phi_j(u) - \phi_j(\bar{u})\| \\ &\ge r - K\|u-\bar{u}\| > \alpha r,\end{aligned}$$

which implies that there is no $j \in \mathcal{J}$ such that $f_i(u) = \phi_j(u), j \in \mathcal{J}$. Hence

$$\|f_i(u) - f_i(\bar{u})\| \le K\|u-\bar{u}\|. \tag{2.5}$$

Now we show that H is Lipschitz continuous at a point $x^* = (x_1^*, \ldots, x_s^*) \in R^{s\times n}$ for the norm $\|\cdot\|_\infty$.

Note that p is continuously differentiable in R^n, and the point $\bar{u} \in R^n$, the index $i \in \{1, 2, \ldots, n\}$, and the norm in (2.2)–(2.5) are arbitrarily chosen. Without loss of generality, we may assume that there are a constant K and a neighborhood $\mathcal{N}^* := \mathcal{N}_1^* \times \mathcal{N}_2^* \ldots \mathcal{N}_s^*$ of x^* such that, for any $x \in \mathcal{N}^*$,

$$\|f(x_i) - f(x_i^*)\|_\infty \le K\|x_i - x_i^*\|_\infty$$

and

$$\|p(x_i) - p(x_i^*)\|_\infty \le K\|x_i - x_i^*\|_\infty, \quad i = 1, 2 \ldots, s.$$

Therefore, by the definition of H, we find

$$\begin{aligned}\|H(x) - H(x^*)\|_\infty &\le \max_{1\le i\le s}\|p(x_i) - p(x_i^*)\|_\infty + h\|A\|_\infty s \max_{1\le i\le s}\|f(x_i) - f(x_i^*)\|_\infty \\ &\le \max_{1\le i\le s} K\|x_i - x_i^*\|_\infty + hs\|A\|_\infty \max_{1\le i\le s} K\|x_i - x_i^*\|_\infty \\ &\le (1 + hs\|A\|_\infty)K \max_{1\le i\le s}\|x_i - x_i^*\|_\infty \\ &\le (1 + hs\|A\|_\infty)K\|x - x^*\|_\infty.\end{aligned}$$

By Lemma 2.1, H is slantly differentiable at x^*. □

THEOREM 2.2. *Let $f^o \in R^{n\times n}$ be a slanting function for f at x_i^*, $i = 1, 2, \ldots, s$. Then*

$$H^o(x) = \operatorname{diag}(p'(x_1), \ldots, p'(x_s)) - h(A \otimes I_n)\operatorname{diag}(f^o(x_1), \ldots, f^o(x_s)) \tag{2.6}$$

is a slanting function for H at x^*, where diag *denotes block diagonal, $I_n \in R^{n\times n}$ is the identity matrix, and the symbol $\otimes$ denotes the tensor product.*

Proof. Since p is continuously differentiable and f^o is a slanting function for f at x_i^*, $i = 1, 2, \ldots, s$, we have that, if $\|x - x^*\|$ is sufficiently small,

$$p(x_i) - p(x_i^*) = p'(x_i)(x_i - x_i^*) + o(\|x_i - x_i^*\|), \quad i = 1, 2, \ldots s,$$

and

$$f(x_i) - f(x_i^*) = f^o(x_i)(x_i - x_i^*) + o(\|x_i - x_i^*\|), \quad i = 1, 2, \ldots, s.$$

Therefore, we get

$$\begin{aligned}
&H(x) - H(x^*) \\
&= \begin{pmatrix} p(x_1) - p(x_1^*) \\ \vdots \\ p(x_s) - p(x_s^*) \end{pmatrix} - h \begin{pmatrix} \sum_{j=1}^{s} a_{1j}(f(x_j) - f(x_j^*)) \\ \vdots \\ \sum_{j=1}^{s} a_{sj}(f(x_j) - f(x_j^*)) \end{pmatrix} \\
&= \begin{pmatrix} p'(x_1)(x_1 - x_1^*) + o(\|x_1 - x_1^*\|) \\ \vdots \\ p'(x_s)(x_s - x_s^*) + o(\|x_s - x_s^*\|) \end{pmatrix} - h \begin{pmatrix} \sum_{j=1}^{s} a_{1j}(f^o(x_j)(x_j - x_j^*) + o(\|x_j - x_j^*\|)) \\ \vdots \\ \sum_{j=1}^{s} a_{sj}(f^o(x_j)(x_j - x_j^*) + o(\|x_j - x_j^*\|)) \end{pmatrix} \\
&= \begin{pmatrix} p'(x_1) & & \\ & \ddots & \\ & & p'(x_s) \end{pmatrix} \begin{pmatrix} x_1 - x_1^* \\ \vdots \\ x_s - x_s^* \end{pmatrix} - h \begin{pmatrix} a_{11}f^o(x_1) & \ldots & a_{1s}f^o(x_s) \\ \vdots & & \vdots \\ a_{s1}f^o(x_1) & \ldots & a_{ss}f^o(x_s) \end{pmatrix} \begin{pmatrix} x_1 - x_1^* \\ \vdots \\ x_s - x_s^* \end{pmatrix} \\
&\qquad + o(\|x - x^*\|) \\
&= (\mathrm{diag}(p'(x_1), \ldots, p'(x_s)) - h(A \otimes I_n)\mathrm{diag}(f^o(x_1), \ldots, f^o(x_s)))(x - x^*) + o(\|x - x^*\|).
\end{aligned}$$

Hence H^o defined in (2.6) is a slanting function for H at x^*. ∎

Now we consider how to compute a slanting function f^o for a piecewise continuously differentiable function f. From Theorem 2.1, f is a locally Lipschitzian function. By the Rademacher theorem, f is differentiable almost everywhere. Hence, we can define the Clarke generalized Jacobian [3]

$$\partial f(y) = \mathrm{co}\{ \lim_{\substack{y^k \to y \\ y^k \in D_f}} f'(y^k)\},$$

where the symbol co denotes the convex hull and D_f is the set of points where f is differentiable. In [2], it is shown that any single valued selection of $\partial f(y)$ is a slanting function of f at y. Furthermore, from Lemma 2 in [12], if f_i is differentiable at y, then there exists ϕ_l in the representation for f such that

$$f_i(y) = \phi_l(y) \quad \text{and} \quad f_i'(y) = \phi_l'(y).$$

Therefore, we can define a slanting function of f as

$$(f^o(y))_i = \phi_l'(y), \qquad \text{where } f_i(y) = \phi_l(y).$$

For such a slanting function, by Theorems 2.1 and 2.2 and Lemma 2.1, we can easily obtain the following corollary.

COROLLARY 2.1. *Let x^* be a solution of $H(x) = 0$. If f is piecewise continuously differentiable, and $p'(x_i^*)$, $i = 1, 2, \ldots, s$, are nonsingular, then for small h the slanting Newton method* (2.1) *is well defined and superlinearly converges to x^*.*

Proof. By the continuity of p' and the nonsingularity of $p'(x_1^*)$, there are a positive constant γ_1 and a neighborhood $\mathcal{N}_1$ of x^* such that, for any $x \in \mathcal{N}_1$, $p'(x_i)$ are all nonsingular and $\|p'(x_i)^{-1}\|_\infty \le \gamma_1$, $i = 1, 2, \ldots, s$. Moreover, by the locally Lipschitzian continuity of f, there are a positive constant γ_2 and a neighborhood $\mathcal{N}_2$ of x^* such that, for any $x \in \mathcal{N}_2$, $\|f^o(x_i)\|_\infty \le \gamma_2$, $i = 1, 2, \ldots, s$. Let

$$B = h\mathrm{diag}(p'(x_1)^{-1}, \ldots, p'(x_s)^{-1})(A \otimes I)\mathrm{diag}(f^o(x_1), \ldots, f^o(x_s)),$$

$$\gamma = s\gamma_1\gamma_2\|A\|_\infty,$$

and $\mathcal{N} = \mathcal{N}_1 \cap \mathcal{N}_2$. Then, for $h \le \lambda/\gamma$ ($\lambda < 1$) and $x \in \mathcal{N}$, we have

$$\|B\|_\infty \le \lambda < 1.$$

Hence, for $h \le \lambda/\gamma$ ($\lambda < 1$) and $x \in \mathcal{N}$, $H^o(x)$ is nonsingular and

$$\begin{aligned}\|H^o(x)^{-1}\| &= \|\mathrm{diag}(p'(x_i)^{-1})(I - B)\|_\infty \\ &\le \|\mathrm{diag}(p'(x_i)^{-1})\|_\infty \|I + B + B^2 + \cdots\|_\infty \\ &\le \frac{\gamma}{1-\lambda} \\ &=: \Gamma.\end{aligned}$$

By Lemma 2.2, the slanting Newton method (2.1) is well defined and superlinearly converges to x^*. ∎

Remark 2.1. When H is continuously differentiable, Jay [8] suggested to use the simplified Newton method

$$x^{m+1} = x^m - H'(x^0)^{-1}H(x^m) \tag{2.7}$$

for solving the system of nonlinear equations in IRK methods. If x^0 is sufficiently close to a solution x^* of $H(x) = 0$,

$$\Phi(x) := x - H'(x^0)^{-1}H(x)$$

is locally contractive in a neighborhood of x^* containing x^0. By using the contractive property, Jay proved the convergence of an inexact simplified Newton method. Note that, when H is continuously differentiable, $H'(x)$ is a slanting function $H^o(x)$ for H. However, if H is not differentiable at x^*,

$$\Psi(x) := x - H^o(x^0)^{-1}H(x)$$

is not locally contractive in any neighborhood of x^* containing x^0. For example,

$$H(x) = \begin{cases} x^2 - 1, & x \ge 1, \\ 1 - x, & x < 1. \end{cases}$$

The function H is not differentiable at the solution $x^* = 1$ of $H(x) = 0$. It is easy to get a slanting function for H:

$$H^o(x) = \begin{cases} 2x, & x \ge 1, \\ -1, & x < 1. \end{cases}$$

For $x^0 < 1$ and $x \ge 1$,

$$|\Psi(x) - \Psi(x^*)| = |x - H^o(x^0)^{-1}H(x) - x^*| = |x + (x^2 - 1) - 1| = |x + 2||x - x^*| > |x - x^*|.$$

For $x^0 \geq 1$ and $0 < x < 1$,

$$|\Psi(x) - \Psi(x^*)| = |x - H^o(x^0)^{-1}H(x) - x^*|$$
$$= \left|x - \frac{1-x}{2x^0} - 1\right| = \left|1 + \frac{1}{2x^0}\right| |x - x^*| > |x - x^*|.$$

Therefore, the simplified Newton method cannot be applied to solve (1.2) when H is not differentiable.

Remark 2.2. Analysis in this section can be easily generalized to the system of nonautonomous ordinary differential equations

$$\begin{aligned} &\frac{dp(t,u)}{dt} = f(t,u), \qquad t \geq 0, \\ &u(0) = u_0. \end{aligned} \tag{2.8}$$

3. Convergence order of IRK methods. We have tested various IRK methods with the slanting Newton method and the bisection method on numerous problems in structural oscillation and pounding. From our numerical experiments, we observe that the order of the IRK methods for Lipschitz continuous ODEs can be preserved if there are finite discontinuous times, and we can find these discontinuous times sufficiently accurately. However, in many cases the order of convergence may drop to one. In theory, we can show that the order of convergence is at least one for the Lipschitz continuous ODEs under mild conditions. This is done for the solution $u(t)$ to (1.1) in a fixed interval $[0, T]$ with the number n of steps chosen such that $t_n = nh = T$. Let

$$e_k(h) = U_k - u(t_k), \qquad k = 0, 1, \ldots, n,$$

and

$$E(h) = \max_{k=0,\ldots,n} \|e_k(h)\|.$$

We assume that the IRK method is well defined; that is, for any U_k there are x^k, U_{k+1} such that $H(x^k) = 0$ and $\tilde{H}(U_{k+1}) = 0$. For simplicity, we consider $p(u) = u$. Let

$$G(x,U) := \begin{pmatrix} x_1 - U - h\sum_{j=1}^{s} a_{1j} f(x_j) \\ \vdots \\ x_s - U - h\sum_{j=1}^{s} a_{sj} f(x_j) \end{pmatrix}.$$

Since f is Lipschitz continuous, the function G is Lipschitz continuous. By the Rademacher theorem, G is differentiable almost everywhere. Hence we can define the Clarke generalized Jacobian [3]

$$\partial G(x,U) = [\pi_x \partial G(x,U), \pi_U \partial G(x,U)],$$

where $\pi_x \partial G(x,U)$ signifies the set of all $(s \times n) \times (s \times n)$ matrices M such that, for some $(s \times n) \times n$ matrix N, the $(s \times n) \times (s \times n + n)$ matrix $[M, N]$ belongs to $\partial G(x,U)$.

For small h, $\pi_x \partial G(x^k, U_k)$ is of maximal rank; i.e., every matrix M in $\pi_x \partial G(x^k, U_k)$ is nonsingular. By the implicit function theorem for Lipschitz continuous function in [3], there exist a neighborhood $\mathcal{U}_k$ of U_k and a Lispchitz function $\psi_k(\cdot; h) : \mathcal{U}_k \to R^{s\times n}$ such that $x^k = \psi_k(U_k; h)$ and, for every $U \in \mathcal{U}_k$, $G(\psi_k(U; h), U) = 0$. Therefore, we have

$$U_{k+1} = U_k + h\sum_{j=1}^{s} b_j f((\psi_k(U_k; h))_j) =: U_k + h\phi_k(U_k; h). \tag{3.1}$$

Obviously $\phi_k(\cdot; h) : R^n \to R^n$ is a locally Lipschitz continuous function.

THEOREM 3.1. *Suppose that there are positive constants K_1 and K_2 such that*

$$\|\phi_k(u(t_k); h) - \phi_k(U_k; h)\| \le K_1\|u(t_k) - U_k\| \tag{3.2}$$

and

$$\|\phi_k(u(t_k); h) - u'(t_k)\| \le K_2 h. \tag{3.3}$$

Then there is a constant $\alpha > 0$ such that

$$E(h) \le \alpha h.$$

Proof. By the Lipschitz continuity of f, u' is Lipschitz continuous. This implies that there is a constant $K_3 > 0$ such that

$$\left\| \frac{u(t_k + h) - u(t_k)}{h} - u'(t_k) \right\| \le K_3 h. \tag{3.4}$$

By (3.1), we have

$$e_{k+1} - e_k = U_{k+1} - U_k - (u(t_{k+1}) - u(t_k)) = h\left(\phi_k(U_k; h) - \frac{u(t_k + h) - u(t_k)}{h}\right).$$

Hence from (3.2)–(3.4), we obtain

$$\begin{aligned}\|e_{k+1}\| &\le (1 + hK_1)\|e_k\| + (K_2 + K_3)h^2 \\ &= (1 + hK_1)^{k+1}\|e_0\| + \frac{K_2 + K_3}{K_1}\left(K_1 h\sum_{j=0}^{k}(1 + K_1 h)^j\right)h,\end{aligned}$$

which, together with $\|e_0\| = 0$, implies that

$$E(h) \le \frac{K_2 + K_3}{K_1}(e^{K_1 T} - 1)h. \qquad \square$$

If $s = 1$, $a_{11} = 0$, and $b_1 = 1$, then $\psi_k(U_k; h) = U_k$, $\phi_k(U_k; h) = f(U_k)$, and (3.1) reduces to the Euler method which has convergence order one. Obviously, (3.2) and (3.3) hold for $\phi_k(U_k; h) = f(U_k)$ and $\phi_k(u(t_k); h) = f(u(t_k))$. Theorem 3.1 shows that the IRK methods for Lispchitz continuous ODEs have convergence order one as the Euler method if $u'(t)$ is not differentiable at some points in the interval $[t_k, t_{k+1}]$ for some k. It is worth noting that the implicit RK methods are numerically stable but require more computational time to solve a system of nonsmooth equations at each step, while explicit methods are faster but may cause a numerical stability problem for stiff ODEs. It will be interesting to study a hybrid method that takes advantage of both implicit and explicit RK methods for solving Lipschitz continuous ODEs.

4. Numerical experiment. Jay [8] developed a code based on 3-stage IRK methods for the numerical solution of (1.1) where f is continuously differentiable. To solve (1.1) where f is not differentiable, we developed a new code based on the slanting Newton method and Jay's code. In our code, the function f can be defined by a finite collection of continuously differentiable functions $\{\phi_l, l \in \mathcal{L}\}$. See Definition 2.3 and Theorem 2.1. The set of points where f is not differentiable is

$$\mathcal{S}_f = \{u \in D \,|\, \phi_i(u) - \phi_j(u) = 0, \phi_i'(u) - \phi_j'(u) \neq 0, i, j \in \mathcal{L}\}.$$

If a point in $\mathcal{S}_f$ is contained within an integration interval, then the order of the IRK method may drop to one. If we can compute exactly the discontinuity times, then the order of the IRK method can be preserved. See the example in subsection 4.1. In our code, a root-finding process is used to find the discontinuities in time. We tested the new code by using many nondifferentiable ODEs. Numerical results show that the new code is efficient. In this section, we report numerical results of IRK methods with the 2-stage Burrage coefficient [1]

$$A = \begin{pmatrix} 1/4 & 0 \\ 1/2 & 1/4 \end{pmatrix} \quad \text{and} \quad b = \begin{pmatrix} 1/2 \\ 1/2 \end{pmatrix},$$

the 2-stage Radau IA coefficient [13]

$$A = \begin{pmatrix} 1/4 & -1/4 \\ 1/4 & 5/12 \end{pmatrix} \quad \text{and} \quad b = \begin{pmatrix} 1/4 \\ 3/4 \end{pmatrix},$$

the 2-stage Radau IIA coefficient [13]

$$A = \begin{pmatrix} 5/12 & -1/12 \\ 3/4 & 1/4 \end{pmatrix} \quad \text{and} \quad b = \begin{pmatrix} 3/4 \\ 1/4 \end{pmatrix},$$

and the 3-stage Lobatto IIIA coefficient [8]

$$A = \begin{pmatrix} 0 & 0 & 0 \\ 5/24 & 1/3 & -1/24 \\ 1/6 & 2/3 & 1/6 \end{pmatrix} \quad \text{and} \quad b = \begin{pmatrix} 1/6 \\ 2/3 \\ 1/6 \end{pmatrix}$$

for solving a nondifferentiable ODE model for the collapse of the Tacoma Narrows suspension bridge and simulating 13 different earthquakes with a suite of 28 ground motion records.

4.1. Bridge collapse. The collapse of the Tacoma Narrows suspension bridge in 1940 left many open questions about the collapse. Lazer and McKenna [10] contend that the nonlinear effects were the main factors leading to the large oscillations of the bridge. The following ODE is a simple version of their model [14]:

$$m\ddot{u} + q(u) = g(t), \qquad u(0) = 0, \quad \dot{u}(0) = \gamma > 0, \tag{4.1}$$

where

$$q(u) = \begin{cases} \alpha u, & u \geq 0, \\ \beta u, & u < 0. \end{cases}$$

Here m is the mass of the section of the roadway, g is the applied force, q is an upward restoring force when $u \geq 0$ and a downward restoring force when $u < 0$, and α and

β are Hooke's constants for the tension and compression, respectively. Note that the function $q(u)$ can be written as

$$q(u) = \alpha \max(0, u) - \beta \max(0, -u).$$

It is shown in [6] that the mapping $\max(0, \cdot) : L^q(\Omega) \to L^p(\Omega)$, with $1 \le p < q \le \infty$, is slantly differentiable on $L^q(\Omega)$ and

$$G(u)(t) = \begin{cases} 1, & u(t) > 0, \\ 0, & u(t) < 0, \\ \delta, & u(t) = 0, \end{cases}$$

is a slanting function for the mapping $\max(0, \cdot)$, where δ is a fixed arbitrary number. By the linear combination [2], the function $q : L^q(\Omega) \to L^p(\Omega)$, with $1 \le p < q \le \infty$, is slantly differentiable on $L^q(\Omega)$ and

$$q^o(u)(t) = \begin{cases} \alpha, & u(t) > 0, \\ \beta, & u(t) < 0, \\ \delta, & u(t) = 0, \end{cases}$$

is a slanting function for q. Furthermore, this problem can be equivalently written as (2.8), by resetting

$$u = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}, \qquad p(u) = \begin{pmatrix} u_1 \\ mu_2 \end{pmatrix}, \quad \text{and} \quad f(t, u) = \begin{pmatrix} u_2 + \gamma \\ g(t) - q(u_1) \end{pmatrix}.$$

Hence, we can apply the IRK method with the slanting Newton method to obtain numerical solution $U^k, k \ge 0$, of this problem. Moreover, if $q(U^k)q(U^{k-1}) < 0$, then there may be a $\tilde{t} \in [t_{k-1}, t_k]$ such that $q(u(\tilde{t})) = 0$. We used the bisection method to find the discontinuity time $\tilde{t}$ and recomputed U^k by the IRK method with $\tilde{t}$ and $\tilde{h} = \tilde{t} - t_{k-1}$.

For $\alpha = 4, \beta = 1, \gamma = 1$, and $g(t) = \sin 4t$ (Example 1, pp. 264 in [14]), the exact solution in the interval $[0, 3\pi]$ is

$$u^*(t) = \begin{cases} (\frac{2}{3} - \frac{1}{6}\cos 2t)\sin 2t, & 0 \le t \le \frac{\pi}{2}, \\ (\frac{7}{5} - \frac{4}{15}\sin t \cos 2t)\cos t, & \frac{\pi}{2} \le t \le \frac{3\pi}{2}, \\ (-\frac{11}{15} - \frac{1}{6}\cos 2t)\sin 2t, & \frac{3\pi}{2} \le t \le 2\pi, \\ (-\frac{23}{15} - \frac{4}{15}\cos t \cos 2t)\sin t, & 2\pi \le t \le 3\pi, \end{cases}$$

which is twice continuously differentiable in $[0, 3\pi]$.

We chose step sizes

$$h_l = \frac{3\pi}{N_l}, \quad N_l = 100 \times 2^l, \quad l = 0, \ldots, 6,$$

so that

$$\log h_{l-1} - \log h_l = \log 2, \quad l = 1, \ldots, 6.$$

We obtain approximate solutions $\{U^k\}_{k=0}^{N_l}$, $l = 0, \ldots, 6$.

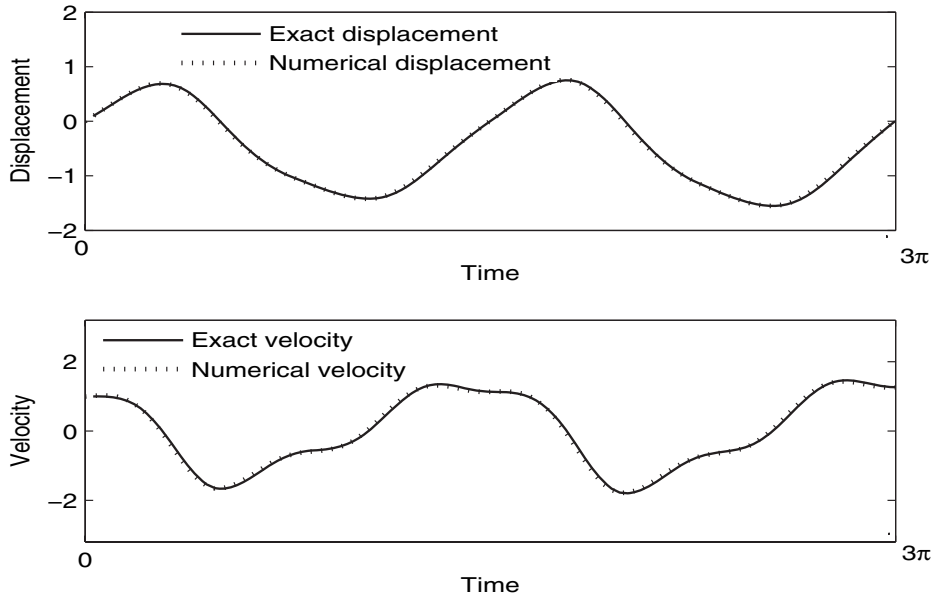


FIG. 4.1. *Exact solution and numerical solution for the collapse of the Tacoma Narrows suspension bridge.*

Figure 4.1 shows what the exact solution and the velocity look like and the numerical solution obtained by the IRK method with the Radau IA and $h_l = \frac{3\pi}{100}$. Figure 4.2 shows the error of the approximate solution

$$\|U_{N_l} - u^*(t_{N_l})\|_2, \quad l = 0, \ldots, 6,$$

at the end point $t_{N_l} = 3\pi$ to compare the 2-stage Burrage, Radau IA, and Radau IIA coefficients. The logarithmic scale is used for the step sizes and errors so that the order of convergence can be seen clearly, when $\|U_{N_l} - u^*(t_{N_l})\|_2 = O(h^l)$. Figure 4.2 shows that the IRK method has order two of convergence for this nonsmooth ODE.

4.2. Seismic pounding. We consider the following nonlinear model of seismic pounding between two adjacent structures called $B1$ and $B2$.

For $i = 1, 2$, let m_i be the masses, r_i be the viscous damping coefficients and k_i be the initial stiffness for $B1$ and $B2$, respectively. The coupling equation of motion for two adjacent buildings subjected to horizontal ground motion $\ddot{u}_g$ has the following version [7, 11]:

$$\begin{array}{rcl} m_1\ddot{u}_1 + r_1\dot{u}_1 + k_1u_1 + q(u_1, u_2, \dot{u}_1, \dot{u}_2) & = & -m_1\ddot{u}_g, \\ m_2\ddot{u}_2 + r_2\dot{u}_2 + k_2u_2 - q(u_1, u_2, \dot{u}_1, \dot{u}_2) & = & -m_2\ddot{u}_g, \end{array} \tag{4.2}$$

where q is the pounding force

$$q(u_1, u_2, \dot{u}_1, \dot{u}_2) = \begin{cases} \alpha(u_1 - u_2 - d)^\gamma + \beta(\dot{u}_1 - \dot{u}_2) & \text{if } u_1 - u_2 > d,\ \dot{u}_1 - \dot{u}_2 > 0, \\ \alpha(u_1 - u_2 - d)^\gamma & \text{if } u_1 - u_2 > d,\ \dot{u}_1 - \dot{u}_2 \le 0, \\ 0 & \text{if } u_1 - u_2 \le d. \end{cases}$$

Here d is the initial separation distance between $B1$ and $B2$, and $\alpha > 0$, $\beta > 0$, and $\gamma > 1$ are fixed parameters.

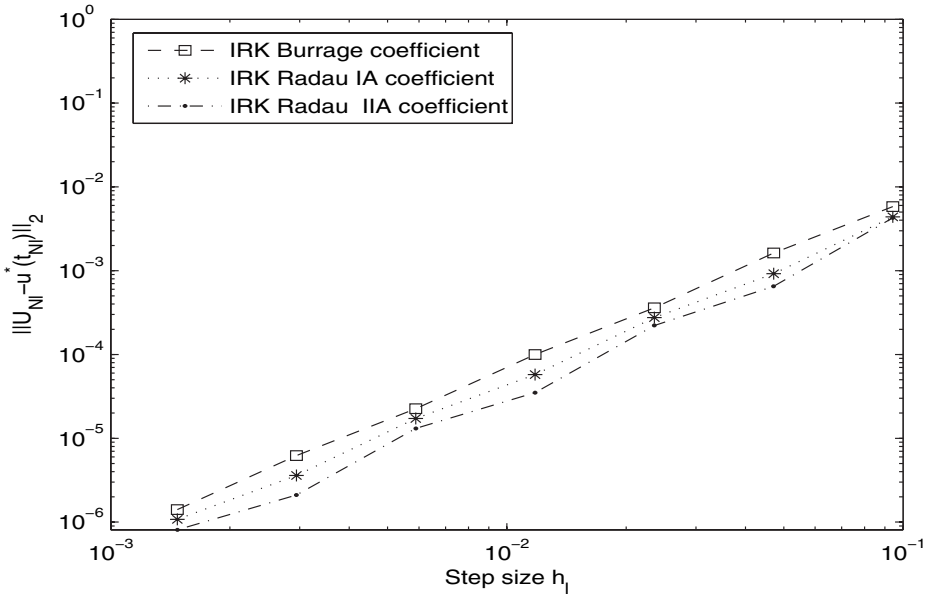


FIG. 4.2. *Error at $t_{N_l} = 3\pi$ of the IRK method with Burrage and Radau coefficients and various step sizes h_l for the collapse of the Tacoma Narrows suspension bridge.*

The ordinary differential equation (4.2) is equivalent to the following system:

$$\begin{aligned} \dot{u}_1 &= u_3, \\ \dot{u}_2 &= u_4, \\ m_1\dot{u}_3 &= -r_1u_3 - k_1u_1 - q(u_1,u_2,u_3,u_4) - m_1\ddot{u}_g, \\ m_2\dot{u}_4 &= -r_2u_4 - k_2u_2 + q(u_1,u_2,u_3,u_4) - m_2\ddot{u}_g. \end{aligned} \tag{4.3}$$

Let us denote

$$u = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{pmatrix}, \quad p(u) = \begin{pmatrix} u_1 \\ u_2 \\ m_1u_3 \\ m_2u_4 \end{pmatrix}, \quad \text{and} \quad f(t,u) = \begin{pmatrix} u_3 \\ u_4 \\ -r_1u_3 - k_1u_1 - q(u) - m_1\ddot{u}_g \\ -r_2u_4 - k_2u_2 + q(u) - m_2\ddot{u}_g \end{pmatrix}.$$

Together with the initial condition $u(0) = 0$, we obtain a system of nonsmooth ordinary differentiable equations (2.8).

We choose

$$m_1 = m_2 = 7.8,\ r_1 = 16.34,\ r_2 = 8.17,\ k_1 = 3421.5,\ k_2 = 855.4,$$
$$\alpha = 25000,\ \gamma = 3/2,\ d = 0.5,\ T = 10.$$

Table 4.1 reports numerical results of the use of a suite of 27 ground motion records with different peak ground accelerations (PGA) from 12 different earthquakes and difference parameters β. The values

$$e(h_1) = ||U^{h_1}(T) - U^{h_4}(T)||_2 \quad \text{and} \quad e(h_3) = ||U^{h_3}(T) - U^{h_4}(T)||_2$$

show the difference of numerical solutions at the end point, for approximate solutions U^{h_l}, with different step sizes

$$h_1 = 2 \times 10^{-3}, \quad h_3 = 2 \times 10^{-4}, \quad h_4 = 10^{-4}.$$

TABLE 4.1
Numerical results for a suite of 27 ground motion records.

	Earthquake	PGA	β	Burrage		Lobatto IIIA	
				$e(h_1)$	$e(h_3)$	$e(h_1)$	$e(h_3)$
I	Northridge, 1994	0.11	113.77	0.0430	0.0010	0.0676	0.0010
	Imperial valley, 1979	0.10	118.90	0.0183	0.0015	0.0564	0.0004
	San Fernando, 1971	0.09	104.88	0.0286	0.0032	0.1902	0.0017
	Loma Prieta, 1989	0.19	119.17	0.0380	0.0030	0.2140	0.0012
	Morgan Hill, 1984	0.19	103.95	0.0337	0.0027	0.1407	0.0004
	N.Palm Springs, 1986	0.21	126.02	0.0276	0.0030	0.1184	0.0003
	Whittier Narrows, 1987	0.30	143.88	0.0104	0.0017	0.0229	0.0005
	Landers, 1992	0.28	141.05	0.0210	0.0008	0.1645	0.0061
	Morgan Hill, 1984	0.29	138.12	0.0341	0.0034	0.1951	0.0022
II	Loma Prieta, 1989	0.37	121.05	0.0870	0.0077	0.1180	0.0074
	Northridge, 1994	0.42	130.49	0.1980	0.0066	0.0871	0.0088
	Cape Mendocino, 1992	0.39	128.33	0.0183	0.0032	0.3519	0.0032
	Northridge, 1994	0.51	161.44	0.1408	0.0128	0.3350	0.0311
	Loma Prieta, 1989	0.48	150.10	0.0485	0.0053	0.5086	0.0049
	Northridge, 1994	0.48	163.14	0.1893	0.0111	0.4833	0.0104
	Loma Prieta, 1989	0.51	146.00	0.0883	0.0036	0.5095	0.0027
	N.Palm Springs, 1986	0.59	146.00	0.0392	0.0028	0.1813	0.0005
	Cape Mendocino, 1992	0.59	166.23	0.0683	0.0022	0.4249	0.0013
III	Loma Prieta, 1989	0.61	153.37	0.3305	0.0194	0.3815	0.0106
	Coalinga, 1983	0.60	145.43	0.0570	0.0070	0.3204	0.0084
	Northridge, 1994	0.59	160.51	0.3654	0.0380	0.7291	0.0387
	Cape Mendocino, 1999	0.66	157.76	0.0726	0.0048	0.8559	0.0059
	Duzce, 1983	0.82	149.00	0.0525	0.0026	0.1905	0.0032
	Coalinga, 1983	0.84	154.89	0.0127	0.0016	0.1759	0.0007
	Northridge, 1994	0.84	140.00	0.0021	0.0002	0.3798	0.0005
	Superstition Hills, 1987	0.89	141.05	0.0941	0.0038	0.3936	0.0084
	Cape Mendocino, 1992	1.04	142.30	0.0469	0.0033	0.0721	0.0023
CPU time (sec) for the 27 records				514	33532	573	34017
I: Low intensity II: Moderate intensity III: High intensity							

Figure 4.3 reports numerical results of the 3-stage Lobatto IIIA coefficient with parameter $\beta = 112$ in a simulation of the 1995 Great Hanshin earthquake, commonly referred to as the Kobe earthquake. The exact solution of this problem is unknown. In the first subplot, we show the difference of simulation solutions at end point $T = 20$:

$$\|U^{h_l}(T) - U^{h_5}(T)\|_2, \quad l = 0, \ldots, 4,$$

where

$$h_l = \frac{20}{N_l}, \quad N_l = 1000 \times 2^l, \quad l = 0, \ldots, 5.$$

We use a logarithmic scale for the step sizes and the difference. The 3-stage Lobatto IIIA method seems to maintain its fourth order of convergence for this problem. The remainder of Figure 4.3 uses step size $h = 0.002$, which shows the displacement u_1, u_2 and the force q for the first 20 seconds of the earthquake. Moreover, we enlarge the plot around the maximum force in the first 10 seconds and the following 10 seconds to show what the pounding force looks like.

All of the ground motion records used in the numerical experiments were taken from the PEER Strong Motion Database [15]. The numerical experiments were performed by using MATLAB 7.0 on a Dell PC with 2 MB memory and 800 MHz.

5. Conclusion. Systems of slantly differentiable ODEs arise from many applications in earthquake engineering and structural dynamics. We proposed IRK methods

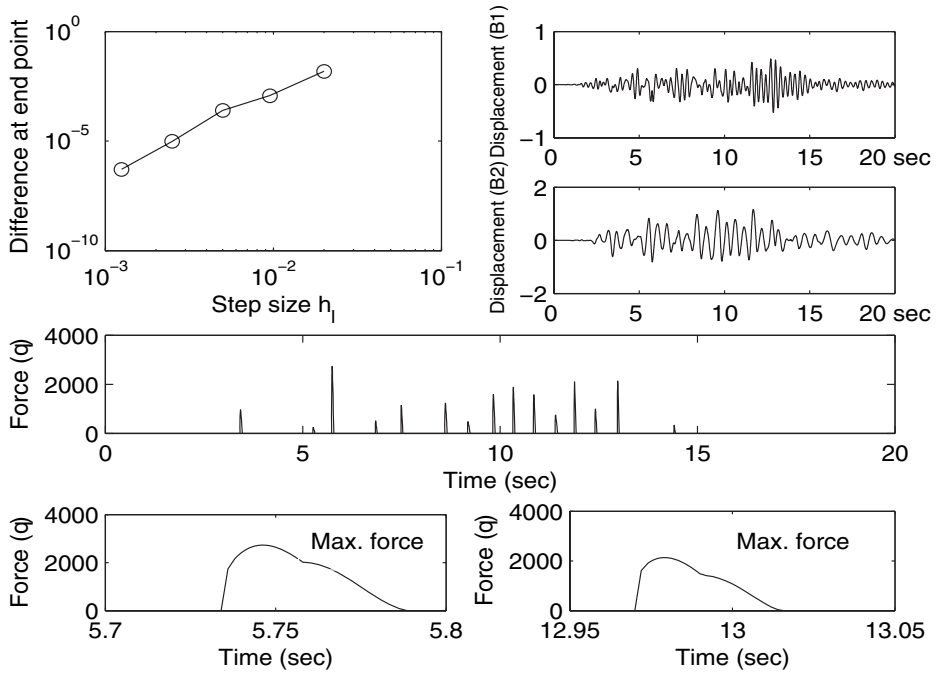


FIG. 4.3. *Difference of simulation solutions. Displacement u_1, u_2, force q, and maximum force during the first* 10 *seconds and the following* 10 *seconds.*

with the slanting Newton method to solve the system of nonsmooth ODEs. We studied the slanting differentiability of the nonsmooth functions involved in the ODEs and proved the superlinear convergence of the slanting Newton method for solving the system of nonsmooth equations in the IRK method. Moreover, we developed a code for nonsmooth ODEs based on Jay's code [8] and the analysis in section 2. Numerical experiments show that the code is efficient. In section 4, we reported some numerical results for the collapse of the Tacoma Narrows suspension bridge and simulating 13 different earthquakes.

Acknowledgments. The authors are grateful to Professor Robert Russell and two anonymous referees for their many helpful comments.

REFERENCES

[1] K. BURRAGE, *Efficiently implementable algebraically stable Runge–Kutta methods*, SIAM J. Numer. Anal., 19 (1982), pp. 245–258.

[2] X. CHEN, Z. NASHED, AND L. QI, *Smoothing methods and semismooth methods for nondifferentiable operator equations*, SIAM J. Numer. Anal., 38 (2000), pp. 1200–1216.

[3] F.H. CLARKE, *Optimization and Nonsmooth Analysis*, Wiley, New York, 1983.

[4] J.E. DENNIS AND R.B. SCHNABEL, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, SIAM, Philadelphia, 1996.

[5] L. FERRACINA AND M.N. SPLIKER, *An extension and analysis of the Shu-Osher representation of Runge-Kutta methods*, Math. Comp., 74 (2004), pp. 201–219.

[6] M. HINTERMÜLLER, K. ITO, AND K. KUNISCH, *The primal-dual active set strategy as a semismooth Newton method*, SIAM J. Optim., 13 (2003), pp. 865–888.

[7] R. JANKOWSKI, *Non-linear viscoelastic modelling of earthquake-induced structural pounding*, Earthquake Engng. Struct. Dyn., 34 (2005), pp. 595–611.

[8] L.O. Jay, *Inexact simplified Newton iterations for implicit Runge-Kutta methods*, SIAM J. Numer. Anal., 38 (2000), pp. 1369–1388.
[9] R. Kress, *Numerical Analysis*, Springer-Verlag, New York, 1998.
[10] A.C. Lazer and P.J. McKenna, *Large-amplitude periodic oscillations in suspension bridges: Some new connections with nonlinear analysis*, SIAM Rev., 32 (1990), pp. 537–578.
[11] S. Muthukumar and R. DesRoches, *A Hertz contact model with non-linear damping for pounding simulation*, Earthquake Engng. Struct. Dyn., 35 (2006), pp. 811–828.
[12] R.T. Rockafellar, *A property of piecewise smooth functions*, Comput. Optim. Appl., 25 (2003), pp. 247–250.
[13] B. Tasic, *Numerical Methods for Solving ODE Flow*, Eindhoven University Press, The Netherlands, 2004.
[14] D.G. Zill, *Differential Equations with Modeling Applications*, 7th ed., Brooks/Cole Thompson Learning, Pacific Grove, CA, 2001.
[15] http://peer.berkeley.edu/smcat/.

SIAM J. NUMER. ANAL.
Vol. 46, No. 3, pp. 1281–1297
© 2008 Society for Industrial and Applied Mathematics

NATURAL SUPERCONVERGENCE POINTS IN THREE-DIMENSIONAL FINITE ELEMENTS*

RUNCHANG LIN† AND ZHIMIN ZHANG‡

Abstract. A systematic and analytic process is conducted to identify natural superconvergence points of high degree polynomial C^0 finite elements in a three-dimensional setting. This identification is based upon explicitly constructing an orthogonal decomposition of local finite element spaces. Derivative and function value superconvergence points are investigated for both the Poisson and the Laplace equations. Superconvergence results are reported for hexahedral, pentahedral, and tetrahedral elements up to certain degrees.

Key words. finite element methods, natural superconvergence, hexahedral, pentahedral (triangular prism), tetrahedral elements polynomial

AMS subject classifications. 65N30, 65M60, 74S05, 41A10

DOI. 10.1137/070681168

1. Introduction.

In this paper, we study *natural superconvergence* points of the finite element (FE) approximation u_h to the Poisson equation

$$\begin{cases} -\Delta u = f & \text{in} \quad \Omega, \\ u = g & \text{on} \quad \partial\Omega, \end{cases} \tag{1.1}$$

where $\Omega \subset \mathbb{R}^3$ is an open bounded domain with Lipschitz boundary $\partial\Omega$ and f and g are sufficiently smooth given functions. When $f \equiv 0$, the Poisson equation (1.1) reduces to the Laplace equation.

Superconvergence for FE methods has been a research topic for about 35 years. For the literature, the reader is referred to [1, 2, 4, 9, 10, 14, 15, 16, 22, 26, 27] and references therein. Here, by "natural superconvergence," we mean that the superconvergence points are obtained without employing any postprocessing in FE solutions. The authors are aware that many superconvergence results can be obtained from postprocessing or recovery techniques. Nevertheless, our investigation is limited to the natural superconvergence in this paper. Moreover, *natural superconvergence* is abbreviated as *superconvergence* when no confusion arises.

Superconvergence points for the three-dimensional (3D) tensor-product FE can be determined from one-dimensional (1D) and/or two-dimensional (2D) results by the tensor-product technique. When the element is not tensor-product due to either geometry or construction, the problem becomes tricky. In [6] Brandts and Křížek surveyed techniques and typical difficulties of superconvergence problems in 3D FEs. Three different methods were reviewed, namely, the *element orthogonality analysis* introduced by Chen [9], Zhu and Lin [26], et al.; the *symmetry theory* due to Schatz, Sloan, and Wahlbin [19]; and *computer-based methods* proposed by Babuška et al. [5].

*Received by the editors January 29, 2007; accepted for publication (in revised form) December 6, 2007; published electronically March 21, 2008.
http://www.siam.org/journals/sinum/46-3/68116.html
†Department of Mathematical and Physical Sciences, Texas A&M International University, Laredo, Texas 78041-2569 (rlin@tamiu.edu).
‡Department of Mathematics, Wayne State University, Detroit, Michigan 48202-3622 (zzhang@math.wayne.edu). This author's research was supported in part by the National Science Foundation grants DMS-0311807 and DMS-0612908.

Following the reasoning of computer-based methods, Zhang [23, 24] proposed an analytic approach to locate derivative superconvergence points for some 2D rectangular and 3D hexahedral elements. In this paper, we systematically extend this analytic approach to identify superconvergence points for 3D higher order FEs of different geometries, including pentahedral (triangular prism) and tetrahedral elements. Specifically, Lagrangian and Serendipity elements of both a hexahedron and a pentahedron are considered; two patterns of tetrahedral elements are also investigated. Superconvergence results for function values and derivatives are provided for elements up to certain degrees. The reader is referred to [25] as a supplement for computational aspects. To the best of our knowledge, many results in this paper are reported for the first time in the literature.

The article is organized as follows. Section 2 contains some preliminaries. Hierarchic shape functions for the elements considered in this paper are constructed in section 3. In section 4, we introduce a decomposition of the FE spaces, which provides us a way to specify the polynomial spaces needed in locating superconvergence points. The structures of the spaces are illustrated in section 5. Section 6 is devoted to presenting the main superconvergence results in different elements. Some concluding remarks are made in the last section.

2. Preliminaries.

2.1. Master cell and reference cell. Consider superconvergence near $\boldsymbol{x}^0 \in \Omega$. We assume that the local FE triangulation is based on a uniform hexahedral mesh (which may be further decomposed into pentahedral or tetrahedral elements) near $\boldsymbol{x}^0$. Without loss of generality, we may assume that the point $\boldsymbol{x}^0$ is the center of one cell in the mesh, which is referred to as the *master cell*. Denote a cell centered at $\boldsymbol{x} = (x_1, x_2, x_3) \in \Omega$ with side $2l$ by

$$c(\boldsymbol{x}, l) = \{\boldsymbol{y} = (y_1, y_2, y_3) \in \Omega \mid |y_i - x_i| \le l,\ i = 1, 2, 3\}.$$

The master cell is hence $c(\boldsymbol{x}^0, h)$, where h is the *mesh parameter*. Let $\Omega_1 = c(\boldsymbol{x}^0, H)$ and $\Omega_0 = c(\boldsymbol{x}^0, 2H) \subset\subset \Omega$ with $H = h^\delta$, $0 < 3\delta/2 < 1$.

Using the reference element technique, the study in the master cell $c(\boldsymbol{x}^0, h)$ can be equivalently conducted in the *reference cell* $K = [-1, 1]^3$. In this context, we will use (x_1, x_2, x_3) for the standard Euclidean coordinates in the master cell (or physical domain Ω), and use (ξ, η, ζ) for those in the reference cell.

2.2. Function spaces. In this paper, we use standard notation for Lebesgue and Sobolev spaces and the corresponding norms.

We denote by Π_n the space of polynomials in three variables with degree not exceeding n. Associated with a particular partition of the reference cell K, we denote by $\Pi_n^w(K)$ ($\Pi_n^\pi(K)$, respectively) the set of *piecewise continuous* (*periodic piecewise continuous*, respectively) polynomials of degree not greater than n. Here, we say a function f is *periodic with respect to* K, or simply *periodic*, if $f(1, \eta, \zeta) = f(-1, \eta, \zeta)$, $f(\xi, 1, \zeta) = f(\xi, -1, \zeta)$, and $f(\xi, \eta, 1) = f(\xi, \eta, -1)$ for any $\xi, \eta, \zeta \in [-1, 1]$. Functions *periodic with respect to the master cell* can be analogously defined. These functions can be periodically extended to Ω_0, which are referred to as *$2h$-periodic functions*.

Let $\mathcal{H}_{n+1}$ be the space of *homogeneous harmonic polynomials* in three variables of degree $n+1$. For $\boldsymbol{x} \in \mathbb{R}^3$, denote its Euclidean norm by $|\boldsymbol{x}|$. It is proved in [3, p. 76] that every $p \in \Pi_{n+1} \backslash \Pi_n$ can be expressed as $p = p_1 + |\boldsymbol{x}|^2 p_2$ with $p_1 \in \mathcal{H}_{n+1}$ and $p_2 \in \Pi_{n-1} \backslash \Pi_{n-2}$. We have the following proposition [3].

PROPOSITION 2.1. *For $n \ge 1$, the space $\mathcal{H}_{n+1}$ has dimension* $\dim \mathcal{H}_{n+1} = 2n+3$.

Furthermore, with help of the *Kelvin transform*, we can find an explicit basis of $\mathcal{H}_{n+1}$. For the detailed approach, see [3, p. 92].

We denote by V_n the polynomial FE space of degree n, and by V_n^π the collection of $2h$-periodic functions in V_n. We denote also by $V_n(K)$ and $V_n^\pi(K)$ the corresponding FE spaces on K.

2.3. Auxiliary functions. The following auxiliary functions will be used later. Let P_k be the *Legendre polynomial* of degree k on $[-1,1]$. Define

$$\phi_k(x) = \int_{-1}^{x} P_{k-1}(t)\ \mathrm{d}t, \quad k = 2,3,\ldots. \tag{2.1}$$

We define also

$$\begin{aligned} \tilde{\phi}_2 &= \phi_2 + \frac{1}{3}, \\ \tilde{\phi}_k &= \phi_k \quad \forall k = 3,4,\ldots. \end{aligned} \tag{2.2}$$

Note that ϕ_k and $\tilde{\phi}_k$ are polynomials of degree k for $k \geq 2$.

2.4. Residue estimates. Let u be the weak solution of (1.1), and u_h be the FE approximation of u in V_n. It follows that

$$a(u - u_h, v) = 0 \qquad \forall v \in V_n^{comp}, \tag{2.3}$$

where V_n^{comp} is the collection of functions in V_n with compact support in Ω, and the bilinear form $a(\cdot,\cdot)$ is defined as

$$a(v,w) = \int_\Omega \nabla v \cdot \nabla w\ \mathrm{d}\Omega.$$

Under various conditions given in [20], the following L_∞ estimates of $u - u_h$ and its derivatives are obtained.

LEMMA 2.2. *Let u and u_h be the weak solution of* (1.1) *and its FE approximation, respectively. Assume that $s \geq 0$ and $1 \leq q \leq \infty$. Then there exists a constant C independent of u, u_h, h, H, and $\boldsymbol{x}^0$ such that*

$$\|u - u_h\|_{L_\infty(\Omega_1)} \leq C \left(\ln \frac{H}{h}\right)^{\bar{n}} \min_{v \in V_n} \|u - v\|_{L_\infty(\Omega_0)} + \ CH^{-s-3/q} \|u - u_h\|_{W_q^{-s}(\Omega_0)}.$$

Here $\bar{n} = 1$ if $n = 1$, and $\bar{n} = 0$ otherwise.

LEMMA 2.3. *Under the same hypotheses as in Lemma* 2.2, *there exists a constant C independent of u, u_h, h, H, and $\boldsymbol{x}^0$ such that*

$$\begin{aligned} \|u - u_h\|_{W_\infty^1(\Omega_1)} \leq C \min_{v \in V_n} \left(|u - v|_{W_\infty^1(\Omega_0)} + H^{-1}\|u - v\|_{L_\infty(\Omega_0)}\right) \\ + \ CH^{-1-s-\frac{3}{q}} \|u - u_h\|_{W_q^{-s}(\Omega_0)}. \end{aligned}$$

The proofs of Lemmas 2.2 and 2.3 can be found in [20].

For $n \geq 1$, we assume that

$$\|u - u_h\|_{L_\infty(\Omega_0)} \leq Ch^{n+1-L}, \tag{2.4}$$

with $L \geq 0$ and $L + 3\delta/2 < 1$. This assumption implies that pollution effects from outside of the domain Ω_0 have been properly controlled and the error loss is of order h^L. Moreover, for $n > 1$, we assume that

$$\|u - u_h\|_{W_\infty^{-1}(\Omega_0)} \leq Ch^{n+2-\Lambda}, \tag{2.5}$$

with $\Lambda \geq 0$ and $\Lambda + \delta < 1$.

The following theorems are analogous to the 2D results in [5] and [18]. We refer the reader also to [22].

THEOREM 2.4. *Assume that the FE mesh over Ω is locally translation invariant on Ω_0 and (2.4) holds. Then, for $n \geq 1$, we have*

$$\frac{\partial}{\partial x_i}(u - u_h)(\boldsymbol{x}) = \frac{\partial \psi}{\partial x_i}(\boldsymbol{x}) + R_{x_i}(\boldsymbol{x}), \qquad i = 1, 2, 3, \quad \boldsymbol{x} \in \Omega_1,$$

where

$$\|R_{x_i}\|_{L_\infty(\Omega_1)} \leq Ch^n \left(h^\delta + h^{1-L-\frac{3}{2}\delta}\right).$$

Here, ψ is the periodic extension of a piecewisely defined polynomial of degree $n+1$, which is periodic with respect to the master cell.

Proof. Let Q be the $n+1$st order Taylor expansion of u at $\boldsymbol{x}_0$. By approximation theory one has

$$\|u - Q\|_{W_\infty^s(\Omega_0)} \leq CH^{n+2-s}, \qquad 0 \leq s \leq n+2. \tag{2.6}$$

Let $N_h Q$ be the *Neumann projection* of Q into $V_n(\Omega_0)$ in the sense that

$$a(Q - N_h Q, v) = 0 \quad \forall v \in V_n(\Omega_0), \quad \text{and} \quad \int_{\Omega_0} (Q - N_h Q)\, \mathrm{d}\Omega = 0. \tag{2.7}$$

Let $R_Q = (u - u_h) - (Q - N_h Q)$. From (2.3) and (2.7) one has

$$a(R_Q, v) = 0 \qquad \forall v \in V_n^{comp}(\Omega_0).$$

By Lemma 2.3 with $s = 0$ and $q = 2$, it follows that, for $i = 1, 2$, and 3,

$$\begin{aligned} \left\| \frac{\partial R_Q}{\partial x_i} \right\|_{L_\infty(\Omega_1)} \leq C \min_{v \in V_n(\Omega_0)} & |(u - Q) - v|_{W_\infty^1(\Omega_0)} \\ & + H^{-1}\|(u - Q) - v\|_{L_\infty(\Omega_0)}) + CH^{-5/2}\|R_Q\|_{L_2(\Omega_0)}. \end{aligned} \tag{2.8}$$

Let $v \in V_n(\Omega_0)$ be the interpolation of $u - Q$. By approximation theory and (2.6) with $s = n + 1$,

$$|(u - Q) - v|_{W_\infty^1(\Omega_0)} \leq Ch^n \|u - Q\|_{W_\infty^{n+1}(\Omega_0)} \leq Ch^n H. \tag{2.9}$$

Similarly,

$$H^{-1}\|(u - Q) - v\|_{L_\infty(\Omega_0)} \leq Ch^{n+1}. \tag{2.10}$$

Recalling (2.4), one has

$$\|u - u_h\|_{L_2(\Omega_0)} \leq CH\|u - u_h\|_{L_\infty(\Omega_0)} \leq CHh^{n+1-L}. \tag{2.11}$$

It follows, by the standard L_2 estimate, that

$$\begin{aligned}\|Q - N_h Q\|_{L_2(\Omega_0)} &\le Ch^{n+1}\|Q\|_{W_2^{n+1}(\Omega_0)} \\ &\le Ch^{n+1}H^{3/2}\|Q\|_{W_\infty^{n+1}(\Omega_0)} \le CH^{3/2}h^{n+1-L}.\end{aligned} \tag{2.12}$$

Hence, from (2.8)–(2.12), one gets

$$\left\| \frac{\partial R_Q}{\partial x_i} \right\|_{L_\infty(\Omega_1)} \le Ch^n(H + H^{-3/2}h^{1-L} + H^{-1}h^{1-L}). \tag{2.13}$$

Let $\rho = Q - I_h Q$, where $I_h Q$ is an interpolation of Q in $V_n(\Omega_0)$. Since, as indicated in [5], $I_h Q$ is translation invariant by $2h$, it can be shown that ρ is $2h$-periodic; cf. [5, 22]. Set $\psi = \rho - N_h^\pi \rho$, where $N_h^\pi \rho$ is the Neumann projection of ρ into $V_n^\pi(\Omega_0)$; i.e., $N_h^\pi \rho$ is the $2h$-periodic version of $N_h \rho$. It follows immediately from (2.7) that

$$a(\psi, v) = 0 \qquad \forall v \in V_n^{comp}(\Omega_0). \tag{2.14}$$

Rewrite $Q - N_h Q = \psi + [Q - N_h Q - \psi]$. From (2.7) and (2.14), one gets

$$a(Q - N_h Q - \psi, v) = 0 \qquad \forall v \in V_n^{comp}(\Omega_0).$$

Notice that $v = Q - N_h Q - \psi = I_h Q + N_h^\pi \rho - N_h Q \in V_n(\Omega_0)$. It follows from Lemma 2.3 with $s = 0, q = 2$ that

$$\begin{aligned}\|Q - N_h Q - \psi\|_{W_\infty^1(\Omega_1)} &\le CH^{-5/2}\|Q - N_h Q - \psi\|_{L_2(\Omega_0)} \\ &\le CH^{-5/2}\left(\|Q - N_h Q\|_{L_2(\Omega_0)} + \|\psi\|_{L_2(\Omega_0)}\right).\end{aligned} \tag{2.15}$$

Another application of the L_2 estimate gives

$$\|Q - N_h Q\|_{L_2(\Omega_0)} \le Ch^{n+1-L}H^{3/2}. \tag{2.16}$$

By a duality argument (see, e.g., [8]),

$$\begin{aligned}\|\psi\|_{L_2(\Omega_0)} &\le Ch\|\rho\|_{W_2^1(\Omega_0)} \le Ch^{n+1}\|Q\|_{W_2^{n+1}(\Omega_0)} \\ &\le Ch^{n+1}H^{3/2}\|Q\|_{W_\infty^{n+1}(\Omega_0)} \le Ch^{n+1-L}H^{3/2}.\end{aligned} \tag{2.17}$$

From (2.15)–(2.17), one concludes that

$$\|Q - N_h Q - \psi\|_{W_\infty^1(\Omega_1)} \le Ch^{n+1-L}H^{-1}. \tag{2.18}$$

Finally, set $R = (u - u_h) - \psi$ and $R_{x_i} = \partial R/\partial x_i$, $i = 1, 2, 3$. From (2.13) and (2.18), one gets

$$\|R_{x_i}\|_{L_\infty(\Omega_1)} \le Ch^n(H + H^{-3/2}h^{1-L} + H^{-1}h^{1-L}).$$

Note that $H = h^\delta < 1$, and the desired result follows. ∎

THEOREM 2.5. *Assume that the FE mesh on Ω is locally translation invariant over Ω_0. Assume also that* (2.5) *holds. Then, for $n > 1$, we have*

$$(u - u_h)(\boldsymbol{x}) = \psi(\boldsymbol{x}) + R(\boldsymbol{x}), \quad \boldsymbol{x} \in \Omega_1,$$

where

$$\|R\|_{L_\infty(\Omega_1)} \le Ch^{n+1}\left(h^\delta + h^{1-\Lambda-\delta}\right),$$

and ψ is as described in Theorem 2.4.

Proof. Define N_hQ, R_Q, and R as in Theorem 2.4. By Lemma 2.2 with $q = \infty, s = 1$,

$$\|R_Q\|_{L_\infty(\Omega_1)} \le C \min_{v\in V_n(\Omega_0)} \|(u-Q)-v\|_{L_\infty(\Omega_0)} + CH^{-1}\|R_Q\|_{W_\infty^{-1}(\Omega_0)}. \tag{2.19}$$

Let $v \in V_n(\Omega_0)$ be the interpolation of $u - Q$. From (2.6) with $s = n + 1$, one gets

$$\|(u-Q)-v\|_{L_\infty(\Omega_0)} \le Ch^{n+1}\|u-Q\|_{W_\infty^{n+1}(\Omega_0)} \le Ch^{n+1}H. \tag{2.20}$$

By assumption (2.5)

$$\|R_Q\|_{W_\infty^{-1}(\Omega_0)} \le \|u-u_h\|_{W_\infty^{-1}(\Omega_0)} + \|Q-N_hQ\|_{W_\infty^{-1}(\Omega_0)} \le Ch^{n+2-\Lambda}. \tag{2.21}$$

Combining (2.19)–(2.21) yields

$$\|R_Q\|_{L_\infty(\Omega_1)} \le Ch^{n+1}H + CH^{-1}h^{n+2-\Lambda}. \tag{2.22}$$

Since $Q - N_hQ - \psi \in V_n(\Omega_0)$, from Lemma 2.2,

$$\begin{aligned}\|Q-N_hQ-\psi\|_{L_\infty(\Omega_1)} &\le CH^{-1}\|Q-N_hQ-\psi\|_{W_\infty^{-1}(\Omega_0)}\\ &\le CH^{-1}\left(\|Q-N_hQ\|_{W_\infty^{-1}(\Omega_0)} + \|\psi\|_{W_\infty^{-1}(\Omega_0)}\right).\end{aligned} \tag{2.23}$$

By assumption (2.5)

$$\|Q-N_hQ\|_{W_\infty^{-1}(\Omega_0)} \le Ch^{n+2-\Lambda}. \tag{2.24}$$

By a duality argument,

$$\|\psi\|_{W_\infty^{-1}(\Omega_0)} \le \|\rho\|_{W_\infty^{-1}(\Omega_0)} \le Ch^{n+2}\|Q\|_{W_\infty^{n+1}(\Omega_0)} \le Ch^{n+2-\Lambda}. \tag{2.25}$$

Therefore, (2.23)–(2.25) give

$$\|Q-N_hQ-\psi\|_{L_\infty(\Omega_1)} \le CH^{-1}h^{n+2-\Lambda}. \tag{2.26}$$

From (2.22) and (2.26), one gets

$$\|R\|_{L_\infty(\Omega_1)} \le Ch^{n+1}(H + H^{-1}h^{1-\Lambda}), \tag{2.27}$$

which completes the proof of the theorem. □

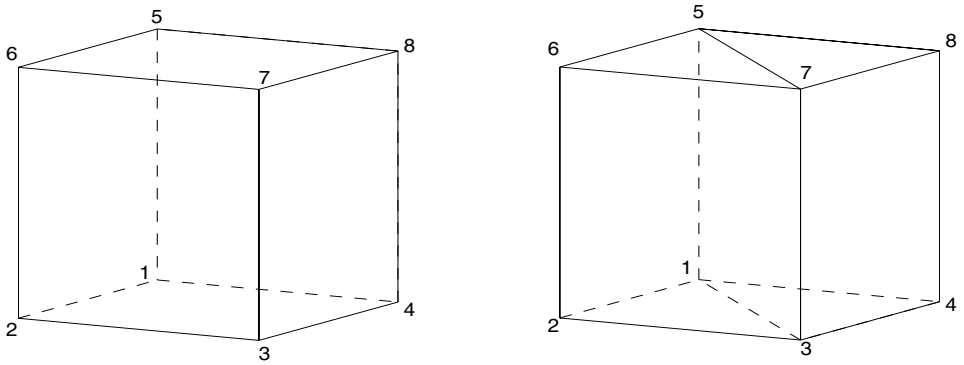
Remark 2.6. Theorem 2.5 may fail when $n = 1$. In fact, even for the one-dimensional case, there is no superconvergence point for function values when piecewise linear elements are employed [22].

Remark 2.7. Notice that ψ is a piecewise polynomial on the master cell and is $2h$-periodic in Ω_0. Theorems 2.4 and 2.5 tell us that the FE local errors in derivatives and function values can be majorized by $\partial\psi/\partial x_i$ and ψ, respectively. Therefore, the task of finding superconvergence points in Ω_1 can be narrowed down to the master cell or, equivalently, to the reference cell under polynomial inputs. The study of superconvergence in a locally translation invariant mesh is from now on carried out in K.

3. Three-dimensional FEs. In this section, hexahedral, pentahedral, and tetrahedral elements are considered. In a hexahedral mesh, a cell in the physical domain serves as an element, which can be mapped to the reference cell K illustrated in Figure 1(a) by the affine mapping

$$(\xi, \eta, \zeta) = \left(\frac{x_1 - x_1^*}{h}, \frac{x_2 - x_2^*}{h}, \frac{x_3 - x_3^*}{h} \right), \tag{3.1}$$

where $\boldsymbol{x}^* = (x_1^*, x_2^*, x_3^*)$ is the center of the physical cell.



(a) Hexahedral Element (b) Pentahedral Elements

FIG. 1. *Partition of K for hexahedral and pentahedral elements.*

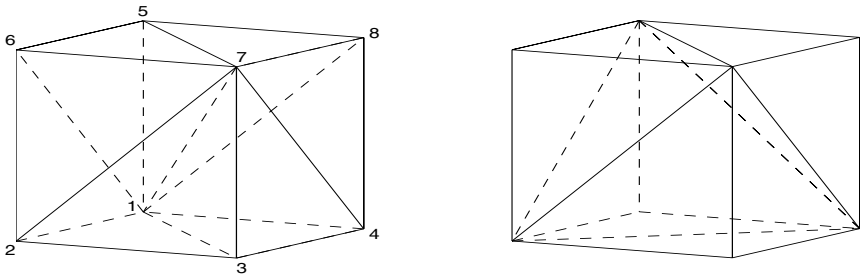
For pentahedral elements, our consideration is for a mesh in which each cell in the physical domain is mapped to the reference cell illustrated in Figure 1(b) by (3.1). The cross-section of this 3D mesh in the x_1x_2-plane is a 2D triangular mesh in a regular pattern [18].

For tetrahedral meshes, two schemes are usually used to partition a unit cell. See Figure 2. Scheme 1 is called *Kuhn partition* of the unit cell, which produces six isomorphic tetrahedra. The master cell can be mapped to the reference cell K by the mapping (3.1). Superconvergence results in [7, 9, 12, 13] are all for tetrahedral elements under Kuhn partition.

Scheme 2 divides a cell into five tetrahedra, four of which are isomorphic. Unlike the previous partitions, to get a translation invariant patch under this scheme, at least eight cubes are needed. As illustrated in Figure 3, there are two candidates for a master cell, which are essentially the same. We consider the first patch in this paper. See Figure 4 for partition of the reference cell. Few superconvergence results are available in the literature for tetrahedral elements under this partition.

We next construct FE spaces $V_n(K)$ and $V_n^\pi(K)$. Hierarchic basis functions are used for $V_n(K)$, which are organized into four categories: nodal shape functions, edge modes, face modes, and internal modes [21]. The nodal shape functions are the standard shape functions, each of which has value 1 at the corresponding node and value 0 at the other nodes. The modal shape functions are, in principle, products of the nodal shape functions of the nodes in the edges, faces, or polyhedra.

Periodic bases of $V_n^\pi(K)$ can be obtained from the basis of $V_n(K)$. In particular, the sum of the nodal shape functions is periodic. The sums of the same order edge modes associated with the three groups of parallel sides, and the sums of the same



Scheme 1 Scheme 2

FIG. 2. *Schemes partitioning unit cell into tetrahedra.*

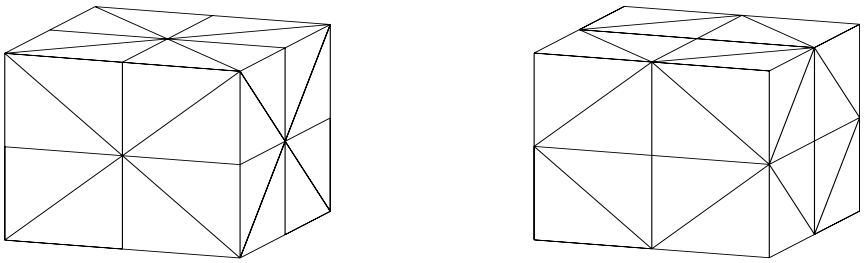
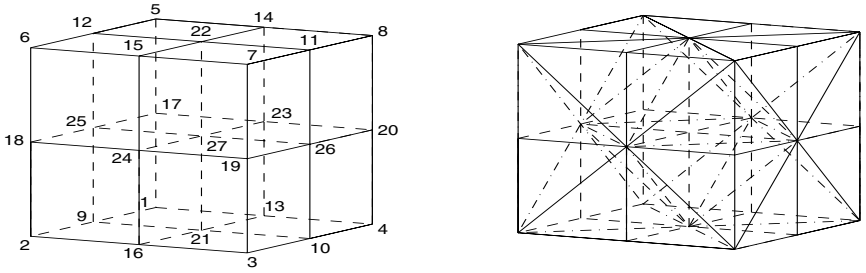


FIG. 3. *Translation invariant patches under scheme* 2.



Nodes in K Partition of K

FIG. 4. *Partition of K for tetrahedral elements under scheme* 2.

order face modes for the opposite faces, are all periodic functions. The internal modes are automatically periodic. For more details, see [17].

A basis of Lagrangian pentahedral FE space and a basis of Serendipity pentahedral FE space of degree n are constructed as an example. In Figure 1(b), $\{(\xi,\eta,\zeta)|\xi \geq \eta\}$ is referred to as the first element and the other the second element. We consider Lagrangian elements first.

(1) *Nodal shape functions.* There are 8 nodal shape functions, which are defined

as

$$N^p(\xi,\eta,\zeta)=\begin{cases}\frac{1}{4}(1-\xi)(1+\zeta_p\zeta),\\ \frac{1}{4}(1-\eta)(1+\zeta_p\zeta),\end{cases}\quad p=1,5,$$

$$N^p(\xi,\eta,\zeta)=\begin{cases}\frac{1}{4}(\xi-\eta)(1+\zeta_p\zeta),\\ 0,\end{cases}\quad p=2,6,$$

$$N^p(\xi,\eta,\zeta)=\begin{cases}\frac{1}{4}(1+\eta)(1+\zeta_p\zeta),\\ \frac{1}{4}(1+\xi)(1+\zeta_p\zeta),\end{cases}\quad p=3,7,$$

$$N^p(\xi,\eta,\zeta)=\begin{cases}0,\\ \frac{1}{4}(\eta-\xi)(1+\zeta_p\zeta),\end{cases}\quad p=4,8,$$

where ζ_p is the ζ-coordinate of the pth node. Hereafter, the two expressions of each basis function are its definitions in elements one and two, respectively.

(2) *Edge modes.* There are $(n-1)$ edge modes on each side with index $i=2,\dots,n$. In particular, the edge modes on $l_{1,2}$ are defined as

$$E_i^{1,2}=\begin{cases}\frac{1}{8}(1-\xi)(\xi-\eta)(1-\zeta)\varphi_{i-2}(\xi-\frac{1}{2}\eta-\frac{1}{2}),\\ 0,\end{cases}$$

where φ_i's are defined from the auxiliary functions in (2.1) by

$$\phi_{i+2}(x)=\frac{1}{4}(1-x^2)\varphi_i(x).$$

Note that each $\phi_k(x)$ has a factor $1-x^2$. Analogously, we define basis functions on sides parallel to the ξ and η axes such as $E_i^{1,4},E_i^{5,6},E_i^{5,8},E_i^{2,3},E_i^{3,4},E_i^{6,7},E_i^{7,8}$. On edges $l_{1,5}$ and $l_{2,6}$, we have

$$E_i^{1,5}=\begin{cases}\frac{1}{8}(1-\xi)(1-\zeta^2)\varphi_{i-2}(\zeta),\\ \frac{1}{8}(1-\eta)(1-\zeta^2)\varphi_{i-2}(\zeta),\end{cases}$$

$$E_i^{2,6}=\begin{cases}\frac{1}{8}(\xi-\eta)(1-\zeta^2)\varphi_{i-2}(\zeta),\\ 0.\end{cases}$$

The edge modes associated with $l_{3,7}$ and $l_{4,8}$ are similarly defined. Finally, for the diagonal edge $l_{1,3}$, we define

$$E_i^{1,3}=\begin{cases}\frac{1}{8}(1-\xi)(1+\eta)(1-\zeta)\varphi_{i-2}(\frac{\xi+\eta}{2}),\\ \frac{1}{8}(1+\xi)(1-\eta)(1-\zeta)\varphi_{i-2}(\frac{\xi+\eta}{2}),\end{cases}$$

and we define $E_i^{5,7}$ similarly.

(3) *Face modes.* On each rectangular face, $(n-1)^2$ shape functions are defined when $n\ge 2$. For instance, on faces 2-3-6-7 and 1-3-5-7, the face modes are

$$F_{i,j}^{2,3,6,7}=\begin{cases}\frac{1}{16}(1+\eta)(\xi-\eta)(1-\zeta^2)\varphi_{i-2}(\eta-\frac{\xi}{2}+\frac{1}{2})\varphi_{j-2}(\zeta),\\ 0,\end{cases}$$

$$F_{i,j}^{1,3,5,7}=\begin{cases}\frac{1}{16}(1-\xi)(1+\eta)(1-\zeta^2)\varphi_{i-2}(\frac{\xi+\eta}{2})\varphi_{j-2}(\zeta),\\ \frac{1}{16}(1+\xi)(1-\eta)(1-\zeta^2)\varphi_{i-2}(\frac{\xi+\eta}{2})\varphi_{j-2}(\zeta),\end{cases}$$

with $i, j = 2, \ldots, n$. On each triangular face, $(n-1)(n-2)/2$ shape functions are defined when $n \geq 3$. For example, on face 1-2-3, the face modes are

$$F_{i,j}^{1,2,3} = \begin{cases} \frac{1}{16}(1-\xi)(\xi-\eta)(1+\eta)(1-\zeta)\xi^i\eta^j, \\ 0, \end{cases}$$

with $i, j = 0, \ldots, n-3$ and $i + j = 0, \ldots, n-3$.

(4) *Internal modes.* There are $(n-1)^2(n-2)/2$ internal modes in every element. For $i, j = 0, \ldots, n-3$, $i+j = 0, \ldots, n-3$, and $k = 2, \ldots, n$, we define internal modes of the first element as

$$I_{i,j,k}^{e_1} = \begin{cases} \frac{1}{32}(1-\xi)(\xi-\eta)(1+\eta)(1-\zeta^2)\xi^i\eta^j\varphi_{k-2}(\zeta), \\ 0. \end{cases}$$

Internal modes of the second elements are analogous.

For Serendipity elements, the nodal shape functions, edge modes, and triangular face modes are congruent as above. However, rectangular face modes are defined as above only when $n \geq 4$, with $i, j = 2, \ldots, n-2$ and $i + j = 4, \ldots, n$. Internal modes are involved when $n \geq 5$; these are defined as above with $i, j = 0, \ldots, n-5$, $k = 2, \ldots, n-3$, and $i + j + k = 2, \ldots, n-3$.

For hierarchic basis functions of hexahedral and tetrahedral elements, the reader is referred to [4, 17, 25].

4. Periodic FE approximation. Let $\Phi_{n+1}(K)$ be a subset of $\Pi_{n+1}^{\pi}(K)$, so that each function ψ in $\Phi_{n+1}(K)$ can be decomposed into the polynomial space $\Pi_{n+1}(K)$ and the piecewise continuous polynomial space $\Pi_n^w(K)$, and

$$\int_K \nabla\psi \cdot \nabla v = 0 \quad \forall v \in \Pi_n^{\pi}(K) \quad \text{and} \quad \int_K \psi = 0. \tag{4.1}$$

That is, for any $\psi \in \Phi_{n+1}(K)$, we have $\psi = \chi + r$ satisfying (4.1), where $\chi \in \Pi_{n+1}(K)$ and $r \in \Pi_n^w(K)$. It is straightforward to show the following result [17, 18].

PROPOSITION 4.1. $\dim \Phi_{n+1}(K) = \dim \Pi_{n+1}(K) - \dim \Pi_n(K)$.

The proposition says that, for each monomial of degree $n+1$, there is a unique function in $\Phi_{n+1}(K)$ corresponding to it. Similarly, for each homogeneous harmonic polynomial of degree $n+1$, there is a unique function in $\Phi_{n+1}(K)$ corresponding to it. In particular, we define the following set, which will be used to study superconvergence of FE solutions for Laplace equations:

$$\Phi_{n+1}^{\mathcal{H}}(K) = \{\, \psi \in \Phi_{n+1}(K) \mid \psi = \chi + r,\ \chi \in \mathcal{H}_{n+1}(K),\ r \in \Pi_n^w(K) \,\}. \tag{4.2}$$

We next introduce a process for determining a basis for $\Phi_{n+1}(K)$ and $\Phi_{n+1}^{\mathcal{H}}(K)$. Here we use tetrahedral elements under scheme 1 as an example; see Figure 2. Let u be a polynomial of degree $n+1$ and define an interpolation operator I_n into $V_n(K)$, such that the following hold:

(1) $I_n u(\xi_p, \eta_p, \zeta_p) = u(\xi_p, \eta_p, \zeta_p)$, $p = 1, \ldots, 8$, where (ξ_p, η_p, ζ_p) is the pth node of K.

(2) Along each edge l in the partition of K,

$$\int_l (u - I_n u) r^j \, \mathrm{d}r = 0, \quad j = 0, 1, \ldots, n-2. \tag{4.3}$$

(3) On each face S in the partition of K,

$$\int_S (u - I_n u) r^j s^k \, \mathrm{d}r \, \mathrm{d}s = 0, \quad j,k \geq 0, \quad j+k = 0,1,\ldots,n-3. \tag{4.4}$$

(4) In each tetrahedron T in the partition of K,

$$\int_T (u - I_n u)\xi^j \eta^k \zeta^l \, \mathrm{d}\xi \, \mathrm{d}\eta \, \mathrm{d}\zeta = 0, \quad j,k,l \geq 0, \quad j+k+l = 0,1,\ldots,n-4. \tag{4.5}$$

Note that $u - I_n u$ is a periodic function in K [4, 22]. Thus, we may find a periodic FE approximation z^π of $u - I_n u$ by solving

$$\int_K \nabla(u - I_n u - z^\pi) \cdot \nabla v = 0 \quad \forall v \in V_n^\pi(K). \tag{4.6}$$

To determine z^π uniquely, we also need

$$\int_K (u - I_n u - z^\pi) = 0. \tag{4.7}$$

Then $\psi|_u = u - I_n u - z^\pi \in \Phi_{n+1}(K)$. See [17, 25] for examples and more details. Letting u run through all monomials of degree $n+1$, we obtain a basis of space $\Phi_{n+1}(K)$ by this approach. Using the process described above, we can also have a unique correspondence of $u \in \mathcal{H}_{n+1}$ in $\Phi_{n+1}^{\mathcal{H}}(K)$.

With a symbolic computation software like Maple, the basis functions of $\Phi_{n+1}(K)$ and $\Phi_{n+1}^{\mathcal{H}}(K)$ can be explicitly determined. Therefore, the process is analytic. Here we list only results for tetrahedral elements of scheme 1. As for results in other geometric shapes, such as tetrahedral elements of scheme 2 and pentahedral elements, the reader is referred to [17].

Notice that the process developed above provides functions ψ which satisfy conditions in Theorems 2.4 and 2.5. The following theorems are thus consequences of Theorems 2.4 and 2.5.

THEOREM 4.2. *Let the assumptions of Theorems* 2.4 *and* 2.5 *hold. Then the following hold:*

(i) *For $n > 1$, the function value superconvergence points of $V_n(K)$ for the Poisson equation are the intersections of the following collection of curves:*

$$\{\, \psi = 0 \mid \psi \in \Phi_{n+1}(K) \backslash V_n(K) \,\}.$$

(ii) *For $n \geq 1$, the derivative superconvergence points of $V_n(K)$ in the ξ-direction for the Poisson equation are the intersections of the following collection of curves:*

$$\left\{ \frac{\partial \psi}{\partial \xi} = 0 \mid \psi \in \Phi_{n+1}(K) \backslash V_n(K) \right\}.$$

A similar result holds for derivatives in other directions.

THEOREM 4.3. *Let the assumptions of Theorems* 2.4 *and* 2.5 *hold. Then the following hold:*

(i) *For $n > 1$, the function value superconvergence points of $V_n(K)$ for the Laplace equation are the intersections of the following collection of curves:*

$$\{\, \psi = 0 \mid \psi \in \Phi_{n+1}^{\mathcal{H}}(K) \backslash V_n(K) \,\}.$$

(ii) *For $n \geq 1$, the derivative superconvergence points of $V_n(K)$ in the ξ-direction for the Laplace equation are the intersections of the following collection of curves:*

$$\left\{ \frac{\partial \psi}{\partial \xi} = 0 \mid \psi \in \Phi_{n+1}^{\mathcal{H}}(K) \backslash V_n(K) \right\}.$$

A similar result holds for derivatives in other directions.

Remark 4.4. From Theorems 4.2 and 4.3, one concludes that, if bases of $\Phi_{n+1}(K)\backslash V_n(K)$ and $\Phi_{n+1}^{\mathcal{H}}(K)\backslash V_n(K)$ are determined, then the superconvergence points can be located by solving a system of polynomial equations, which is a routine exercise of numerical computation.

5. Structures of $\boldsymbol{\Phi_{n+1}(K)\backslash V_n(K)}$ and $\boldsymbol{\Phi_{n+1}^{\mathcal{H}}(K)\backslash V_n(K)}$. To determine the spaces $\Phi_{n+1}(K)\backslash V_n(K)$ and $\Phi_{n+1}^{\mathcal{H}}(K)\backslash V_n(K)$, one may first determine a basis of $\Pi_{n+1}(K) \setminus V_n(K)$ and $\mathcal{H}_{n+1}(K) \setminus V_n(K)$, respectively, and then use the approach developed in section 4 to find the corresponding basis functions in the desired spaces.

5.1. Hexahedral elements. Consider Lagrangian hexahedral elements. We note that the nth degree local FE space $V_n(K)$ contains not only functions in $\Pi_n(K)$, but also polynomials of degree higher than n. In particular,

$$\Pi_{n+1}(K) \setminus V_n(K) = \operatorname{Span}\left\{\xi^{n+1}, \eta^{n+1}, \zeta^{n+1}\right\}.$$

It is straightforward to verify that

$$\Phi_{n+1}(K)\backslash V_n(K) = \operatorname{Span}\left\{\tilde{\phi}_{n+1}(\xi), \tilde{\phi}_{n+1}(\eta), \tilde{\phi}_{n+1}(\zeta)\right\},$$

where $\tilde{\phi}_n$ is defined in (2.2). $\Phi_{n+1}^{\mathcal{H}}(K)\backslash V_n(K)$ can be specified analogously.

For Serendipity hexahedral elements, there are fewer basis functions in $V_n(K)$. In particular, we have

$$\Pi_{n+1}(K) \setminus V_n(K) = \operatorname{Span}\left\{\xi^i \eta^j \zeta^k \mid i+j+k = n+1,\ i,j,k \neq 1\right\}.$$

Then $\Phi_{n+1}(K)\backslash V_n(K)$ is determined accordingly, and so is $\Phi_{n+1}^{\mathcal{H}}(K)\backslash V_n(K)$. See [23] for similar formulations obtained from a different approach.

5.2. Pentahedral elements. Consider Lagrangian pentahedral elements. We have

$$\Pi_{n+1}(K) \setminus V_n(K) = \operatorname{Span}\left\{\zeta^{n+1}, \xi^{n+1-i}\eta^i \mid i = 0, \ldots, n+1\right\}.$$

Then $\Phi_{n+1}(K)\backslash V_n(K)$ can be determined. It turns out to be spanned by $\tilde{\phi}_{n+1}(\zeta)$ and the basis of the triangular elements for a regular pattern in [18]. In addition, $\Phi_{n+1}^{\mathcal{H}}(K)\backslash V_n(K)$ has a basis consisting of $\tilde{\phi}_{n+1}(\zeta)$ and two harmonic basis functions of triangular regular pattern in [18].

For Serendipity pentahedral elements, we have

$$\Pi_{n+1}(K) \setminus V_n(K) = \operatorname{Span}\left\{\xi^{n+1-i-j}\eta^i\zeta^j \mid i,j,i+j = 0, \ldots, n+1,\ j \neq 1, n\right\}.$$

Using the process introduced in section 4, we can determine $\Phi_{n+1}(K)\backslash V_n(K)$ and $\Phi_{n+1}^{\mathcal{H}}(K)\backslash V_n(K)$ accordingly. Note that when $n = 1$ and 2, the spaces for Serendipity elements are the same as those for Lagrangian elements. When $n \geq 3$, the spaces are different.

TABLE 1
Function value superconvergence results of hexahedral FE.

	Lagrangian hexahedral FE	Serendipity hexahedral FE	
n	Poisson & Laplace	Poisson	Laplace
2	HF_2	SPH	SPH
3	HF_3	None	HSet1
4	HF_4	SPH	SPH
5	HF_5	None	None
6	HF_6	SPH	SPH

TABLE 2
ξ-derivative superconvergence results of hexahedral FE.

	Lagrangian hexahedral FE	Serendipity hexahedral FE	
n	Poisson & Laplace	Poisson	Laplace
1	HD_1	HD_1	HD_1
2	HD_2	HD_2	HD_2
3	HD_3	HSet2	HSet4
4	HD_4	None	None
5	HD_5	HSet3	HSet3
6	HD_6	None	None

5.3. Tetrahedral elements. For tetrahedral elements, we have

$$\Pi_n(K) \subset \Pi_n^w(K) = V_n(K).$$

Note that $V_n(K)$ does not contain any polynomial of degree greater than n; thus,

$$\Pi_{n+1}(K) \setminus V_n(K) = \Pi_{n+1}(K) \setminus \Pi_n(K).$$

Therefore, $\Phi_{n+1}(K)\backslash V_n(K) = \Phi_{n+1}(K)$. A basis of $\Phi_{n+1}^{\mathcal{H}}(K)$ can be obtained from a basis of $\mathcal{H}_{n+1}$.

6. Superconvergence results. Once the bases of spaces $\Phi_{n+1}(K)\backslash V_n(K)$ and $\Phi_{n+1}^{\mathcal{H}}(K)\backslash V_n(K)$ are determined, one may apply Theorems 4.2 and 4.3 to the basis functions and obtain superconvergence points for the Poisson and Laplace equations, respectively.

6.1. Hexahedral elements. In Tables 1 and 2, we collect superconvergence results for Lagrangian and Serendipity hexahedral elements of degree up to 6, though our method can produce superconvergence results for elements of higher degrees. Here, "None" means no superconvergence results. The other sets are defined below.

$\text{SPH} = \{\ (\xi,\eta,\zeta) \in K \mid \xi,\eta,\zeta = -1, 0, 1\ \};$

$\text{HF}_n = \{\ (\xi,\eta,\zeta) \in K \mid \phi_{n+1}(\xi) = 0,\ \phi_{n+1}(\eta) = 0,\ \phi_{n+1}(\zeta) = 0\ \};$

$\text{HD}_n = \{\ (\xi,\eta,\zeta) \in K \mid P_n(\xi) = 0\ \};$

$\text{HSet1} = \{\ (\pm\frac{\sqrt{-9+6\sqrt{3}}}{3}, \pm\frac{\sqrt{-9+6\sqrt{3}}}{3}, \pm\frac{\sqrt{-9+6\sqrt{3}}}{3}),$
$(\pm\frac{\sqrt{-9+6\sqrt{3}}}{3}, \pm\frac{\sqrt{-9+6\sqrt{3}}}{3}, \pm\frac{\sqrt{-225+210\sqrt{3}}}{15}),$
$(\pm\frac{\sqrt{-9+6\sqrt{3}}}{3}, \pm\frac{\sqrt{-225+210\sqrt{3}}}{15}, \pm\frac{\sqrt{-9+6\sqrt{3}}}{3}),$
$(\pm\frac{\sqrt{-225+210\sqrt{3}}}{15}, \pm\frac{\sqrt{-9+6\sqrt{3}}}{3}, \pm\frac{\sqrt{-9+6\sqrt{3}}}{3})\ \};$

$\text{HSet2} = \{\ (\xi,\eta,\zeta) \in K \mid \xi = 0\ \} \cup \{\ (\pm\sqrt{3/5}, \pm\sqrt{1/3}, \pm\sqrt{1/3})\ \};$

$\text{HSet3} = \{\ (0,0,0), (0,\pm1,0), (0,0,\pm1), (0,\pm1,\pm1)\ \};$

$\text{HSet4} = \{\ (\xi,\eta,\zeta) \in K \mid \xi = 0, \quad \text{or } \xi = \pm\sqrt{\tfrac{1+6\eta^2}{5}} \text{ and } \zeta = \pm\eta\ \};$

where P_n is the Legendre polynomial of degree n, and ϕ_n is defined in (2.1). Superconvergence results of the η- and ζ-derivatives can be obtained analogously.

6.2. Pentahedral elements. Tables 3 and 4 are devoted to superconvergence results for Lagrangian and Serendipity pentahedral elements. Elements of degree up to 5 are studied.

TABLE 3
Function value superconvergence results of pentahedral FE.

	Lagrangian pentahedral FE		Serendipity pentahedral FE	
n	Poisson	Laplace	Poisson	Laplace
2	SPH	PF_2	SPH	PF_2
3	None	PF_3	None	None
4	PSet1	PF_4	SPH	SPH
5	None	PF_5	None	None

TABLE 4
Derivative superconvergence results of pentahedral FE.

	Lagrangian pentahedral FE		Serendipity pentahedral FE	
n	Poisson	Laplace	Poisson	Laplace
1	ξ- : PSet2 ζ- : $PD1_1$	ξ- : $PD2_1$ ζ- : $PD1_1$	ξ- : PSet2 ζ- : $PD1_1$	ξ- : PSet2 ζ- : $PD1_1$
2	ξ- : PSet3 ζ- : $PD1_2$	ξ- : $PD2_2$ ζ- : $PD1_2$	ξ- : PSet3 ζ- : $PD1_2$	ξ- : $PD2_2$ ζ- : $PD1_2$
3	ξ- : PSet2 ζ- : $PD1_3$	ξ- : $PD2_3$ ζ- : $PD1_3$	ξ- : PSet2 ζ- : $\{\zeta = 0\}$	ξ- : PSet2 ζ- : PSet6
4	ξ- : None ζ- : $PD1_4$	ξ- : $PD2_4$ ζ- : $PD1_4$	ξ- : None ζ- : None	ξ- : None ζ- : None
5	ξ- : PSet2 ζ- : $PD1_5$	ξ- : $PD2_5$ ζ- : $PD1_5$	ξ- : PSet4 ζ- : PSet5	ξ- : PSet4 ζ- : PSet5

The sets in the tables are defined as

$\mathrm{PF}_n = \{(\xi,\eta,\zeta) \mid (\xi,\eta) \in \mathrm{FSR}_n, \quad \phi_{n+1}(\zeta) = 0\};$

$\mathrm{PD1}_n = \{(\xi,\eta,\zeta) \mid P_n(\zeta) = 0\};$

$\mathrm{PD2}_n = \{(\xi,\eta,\zeta) \mid (\xi,\eta) \in \mathrm{DSR}_{\xi,n}\};$

$\mathrm{PSet1} = \{(\xi,\eta,\zeta) \mid \xi,\eta = \pm 1, 0, \quad \phi_5(\zeta) = 0\};$

$\mathrm{PSet2} = \{(\xi,\eta,\zeta) \mid \xi = 0, \ \eta = \pm 1\};$

$\mathrm{PSet3} = \{(\xi,\eta,\zeta) \mid \xi = \pm\sqrt{3}/3, \ \eta = \pm 1\};$

$\mathrm{PSet4} = \{(0,\pm 1,\pm 1), \ (0,\pm 1,0)\};$

$\mathrm{PSet5} = \{(0,0,0), \ (\pm 1,0,0), \ (0,\pm 1,0), \ (\pm 1,\pm 1,0)\};$

$\mathrm{PSet6} = \{(\xi,\eta,\zeta) \mid \zeta = 0\} \cup \{\pm(1 - \frac{\sqrt{3}}{3}, -1 + \frac{\sqrt{3}}{3}, \pm\frac{\sqrt{90-50\sqrt{3}}}{5})\};$

where FSR_n is the collection of the nth order 2D regular pattern function value superconvergence points of the Laplace equation, and $\mathrm{DSR}_{\xi,n}$ is the set of nth order 2D regular pattern ξ-derivative superconvergence points of the Laplace equation [18]. Superconvergence results of the η-derivatives are analogous to those of the ξ-derivatives.

6.3. Tetrahedral elements. Superconvergence results for tetrahedral elements of degree up to 4 are collected in Tables 5 and 6.

Here, TSet1 is the set of vertices and midpoints of edges of K; TSet2 is the set of vertices of K and midpoints of diagonal edges of K (e.g., edges $l_{1,21}$, $l_{1,23}$, $l_{1,25}$ in Figure 4); TSet3 is the set of midpoints of the edges of K which are parallel to the ξ-axis; and TSet4 is the set of second order Gaussian points along the edges of

TABLE 5
Function value superconvergence results of tetrahedral FE.

	Scheme 1	Scheme 2	
n	Poisson & Laplace	Poisson	Laplace
2	TSet1	TSet2	TSet1*
3	None	None	None
4	TSet1	TSet2	TSet1*

TABLE 6
ξ-derivative superconvergence results of tetrahedral FE.

	Scheme 1	Scheme 2	
n	Poisson & Laplace	Poisson	Laplace
1	TSet3	None	TSet3*
2	TSet4	None	TSet4*
3	TSet3	None	TSet3*
4	None	None	None

K which are parallel to the ξ-axis. Note that TSet$m$ and TSet$m$* are different due to the different partition schemes. For η- and ζ-derivatives, superconvergence results can be obtained analogously.

In addition, under partition scheme 2, if we consider derivatives along diagonal directions (e.g., directions of edges $l_{1,21}$, $l_{1,23}$, $l_{1,25}$ in Figure 4), there are superconvergence results. In particular, when $n = 1$ and 3, the midpoints of the edges parallel to the tangential direction are superconvergence points; when $n = 2$, the second order Gaussian points of the tangential direction edges are superconvergence points; when $n = 4$, there are no superconvergence points. The results hold for both the Poisson and the Laplace equations.

7. Concluding remarks. In this article, a systematic and analytic approach in determining 3D FE superconvergence points is presented. We investigate hexahedral, pentahedral, and tetrahedral elements, which are widely used in engineering applications. The results for the Poisson equation can be applied to many problems which have second order elliptic operators, including some nonlinear problems. The approach can be used to locate superconvergence points of any degree and provides conclusive results. We conclude this article by the following remarks.

Remark 7.1. In most 3D elements, superconvergence points of both the Poisson and Laplace equations are the same, as demonstrated in section 6. This is in contrast with the 2D cases, where there are many more superconvergence points of the Laplace equation than that of the Poisson equation for higher order elements. The reason is that, when $n \geq 2$, there are only two basis functions in the space of harmonic polynomials for each degree n in two dimensions, while there are $2n+3$ basis functions in the space of harmonic polynomials for each degree n in three dimensions. The likelihood of $2n + 3$ polynomials passing a common point is much smaller than the probability of two polynomials intersecting.

Remark 7.2. Although some superconvergence points can also be predicted by the symmetry theory and the tensor-product technique, we nevertheless provide conclusive results and demonstrate that in some cases symmetry points and/or points obtained by tensor-product are the only superconvergence points, while in other cases there are more superconvergence points which are not these types of points.

Remark 7.3. As indicated in [27, Chap. 14], many recovery techniques utilize a priori knowledge of natural superconvergence points. Therefore, the results in this paper are not only interesting by themselves but also good "hints" for more flexible and robust recovery methods.

Acknowledgment. The authors are grateful to the anonymous referees for their helpful comments and suggestions, which improved and polished the presentation of this article.

REFERENCES

[1] M. Ainsworth and J. T. Oden, *A Posteriori Error Estimation in Finite Element Analysis*, Wiley Interscience, New York, 2000.

[2] A. B. Andreev and R. D. Lazarov, *Superconvergence of the gradient for quadratic triangular finite elements*, Numer. Methods Partial Differential Equations, 4 (1988), pp. 15–32.

[3] S. Axler, P. Bourdon, and W. Ramey, *Harmonic Function Theory*, 2nd ed., Springer-Verlag, New York, 2001.

[4] I. Babuška and T. Strouboulis, *The Finite Element Method and Its Reliability*, Oxford University Press, Oxford, UK, 2001.

[5] I. Babuška, T. Strouboulis, C. S. Upadhyay, and S. K. Gangaraj, *Computer-based proof of the existence of superconvergence points in the finite element method; superconvergence of the derivatives in finite element solutions of Laplace's, Poisson's, and the elasticity equations*, Numer. Methods Partial Differential Equations, 12 (1996), pp. 347–392.

[6] J. H. Brandts and M. Křížek, *History and future of superconvergence in three-dimensional finite element methods*, in Proceedings of the Conference on Finite Element Methods: Three-Dimensional Problems, GAKUTO Internat. Ser. Math. Sci. Appl. 15, Gakkōtosho, Tokyo, 2001, pp. 24–35.

[7] J. H. Brandts and M. Křížek, *Superconvergence of tetrahedral quadratic finite elements*, J. Comput. Math., 23 (2005), pp. 27–36.

[8] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, Springer-Verlag, New York, 1994.

[9] C. M. Chen, *Structure Theory of Superconvergence of Finite Elements*, Hunan Science Press, Changsha, China, 2001 (in Chinese).

[10] C. M. Chen and Y. Q. Huang, *High Accuracy Theory of Finite Element Methods*, Hunan Science Press, Changsha, China, 1995 (in Chinese).

[11] R. E. Ewing, R. D. Lazarov, and J. Wang, *Superconvergence of the velocity along the Gauss lines in mixed finite element methods*, SIAM J. Numer. Anal., 28 (1991), pp. 1015–1029.

[12] G. Goodsell, *Pointwise superconvergence of the gradient for the linear tetrahedral element*, Numer. Methods Partial Differential Equations, 10 (1994), pp. 651–666.

[13] V. Kantchev and R. D. Lazarov, *Superconvergence of the Gradient of Linear Finite Elements for 3D Poisson Equation. Optimal Algorithms*, Publ. Bulg. Acad. Sci., Sofia, 1986, pp. 172–182.

[14] M. Křížek and P. Neittaanmäki, *On superconvergence techniques*, Acta Appl. Math., 9 (1987), pp. 175–198.

[15] M. Křížek, P. Neittaanmäki, and R. Stenberg, eds., *Finite Element Methods: Superconvergence, Post-Processing, and A Posteriori Estimates*, Lecture Notes in Pure and Applied Mathematics 196, Marcel Dekker, New York, 1997.

[16] Q. Lin and N. Yan, *Construction and Analysis of High Efficient Finite Elements*, Hebei University Press, Baoding, China, 1996 (in Chinese).

[17] R. Lin, *Natural Superconvergence in Two- and Three-Dimensional Finite Element Methods*, Ph.D. dissertation, Department of Mathematics, Wayne State University, Detroit, MI, 2005.

[18] R. Lin and Z. Zhang, *Natural superconvergence points of triangular finite elements*, Numer. Methods Partial Differential Equations, 20 (2004), pp. 864–906.

[19] A. H. Schatz, I. H. Sloan, and L. B. Wahlbin, *Superconvergence in finite element methods and meshes that are locally symmetric with respect to a point*, SIAM J. Numer. Anal., 33 (1996), pp. 505–521.

[20] A. H. Schatz and L. B. Wahlbin, *Interior maximum-norm estimates for finite element methods* II, Math. Comp., 64 (1995), pp. 907–928.

[21] B. Szabó and I. Babuška, *Finite Element Analysis*, John Wiley & Sons, New York, 1991.

[22] L. B. Wahlbin, *Superconvergence in Galerkin Finite Element Methods*, Lecture Notes in Math. 1605, Springer-Verlag, Berlin, 1995.
[23] Z. Zhang, *Derivative superconvergence points in finite element solutions of Poisson's equation for the Serendipity and intermediate families—a theoretical justification*, Math. Comp., 67 (1998), pp. 541–552.
[24] Z. Zhang, *Derivative superconvergence points in finite element solutions of harmonic functions—a theoretical justification*, Math. Comp., 71 (2002), pp. 1421–1430.
[25] Z. Zhang and R. Lin, *Locating natural superconvergent points of finite element methods in* 3*D*, Internat. J. Numer. Anal. Model., 2 (2005), pp. 19–30.
[26] Q. D. Zhu and Q. Lin, *Superconvergence Theory of the Finite Element Method*, Hunan Science Press, China, Changsha, 1989 (in Chinese).
[27] O. C. Zienkiewicz and R. L. Taylor, *The Finite Element Method. Vol.* 1. *The Basis*, 5th ed., Butterworth-Heinemann, Oxford, UK, 2000.

SIAM J. NUMER. ANAL.
Vol. 46, No. 3, pp. 1298–1322
© 2008 Society for Industrial and Applied Mathematics

ANALYSIS OF AN ASYMPTOTIC PRESERVING SCHEME FOR THE EULER–POISSON SYSTEM IN THE QUASINEUTRAL LIMIT*

PIERRE DEGOND†, JIAN-GUO LIU‡, AND MARIE-HÉLÈNE VIGNAL†

Abstract. In a previous work [P. Crispel, P. Degond, and M.-H. Vignal, *J. Comput. Phys.*, 223 (2007), pp. 208–234], a new numerical discretization of the Euler–Poisson system was proposed. This scheme is "asymptotic preserving" in the quasineutral limit (i.e., when the Debye length ε tends to zero), which means that it becomes consistent with the limit model when $\varepsilon \to 0$. In the present work, we show that the stability domain of the present scheme is independent of ε. This stability analysis is performed on the Fourier transformed (with respect to the space variable) linearized system. We show that the stability property is more robust when a space-decentered scheme is used (which brings in some numerical dissipation) rather than a space-centered scheme. The linearization is first performed about a zero mean velocity and then about a nonzero mean velocity. At the various stages of the analysis, our scheme is compared with more classical schemes and its improved stability property is outlined. The analysis of a fully discrete (in space and time) version of the scheme is also given. Finally, some considerations about a model nonlinear problem, the Burgers–Poisson problem, are also discussed.

Key words. stiffness, Debye length, electron plasma period, Burgers–Poisson, sheath problem, Klein–Gordon

AMS subject classifications. 82D10, 76W05, 76X05, 76N10, 76N20, 76L05

DOI. 10.1137/070690584

1. Introduction.

In a previous work [9] (see also [7]), a new numerical discretization of the Euler–Poisson system was proposed. The Euler–Poisson system under consideration consists of the isentropic Euler equations for the particle and momentum densities coupled with the Poisson equation through a source term modeling the electrostatic force. In dimensionless units, the coupling constant can be expressed in terms of a parameter ε which represents the scaled Debye length. When ε is small, the coupling is strong. In this situation, the particle density is constrained to be close to the background density of the oppositely charged particle, which we suppose to be uniform and equal to 1 in scaled units. The velocity then evolves according to the incompressible Euler equation. The limit $\varepsilon \to 0$ is called the quasineutral limit, since the charge density almost vanishes identically. When two or more particle species are considered, the limit $\varepsilon \to 0$ leads to a more complex model usually referred to in the physics literature as a quasineutral model. In the present work, we restrict ourselves to the case of a single particle species as described above.

The scheme which has been proposed in [9] and [7] is "asymptotic preserving" (AP) in the quasineutral limit, which means that it becomes consistent with the limit model when $\varepsilon \to 0$. The goal of the present work is to analyze the stability properties of this scheme in a one-dimensional framework and to show that its stability domain is independent of ε. This stability analysis is performed on the Fourier transformed (with

*Received by the editors May 7, 2007; accepted for publication (in revised form) December 6, 2007; published electronically March 21, 2008.
http://www.siam.org/journals/sinum/46-3/69058.html

†Institut de Mathématiques de Toulouse, groupe "Mathématiques pour l'Industrie et la Physique," UMR 5219 (CNRS-UPS-INSA-UT1-UT2), Université Paul Sabatier, 118 route de Narbonne, 31062 Toulouse cedex 9, France (degond@mip.ups-tlse.fr, mhvignal@mip.ups-tlse.fr).

‡Department of Mathematics and Institute for Physical Science and Technology, University of Maryland, College Park, MD 20742 (jliu@math.umd.edu). The research of this author was partially supported by NSF grant DMS 05-12176.

respect to the space variable) linearized system. We show that the stability property is more robust when a space-decentered scheme is used (which brings in some numerical dissipation) rather than a space-centered scheme. The linearization is first performed about a zero mean velocity and then about a nonzero mean velocity. At the various stages of the analysis, our scheme is compared with more classical schemes and its improved stability property is outlined. The analysis of a fully discrete (in space and time) version of the scheme is also given. Finally, some considerations about a model nonlinear problem, the Burgers–Poisson problem, are also discussed.

The Euler–Poisson model is one of the most widely used fluid models in plasma and semiconductor physics (see, e.g., [21], [1] for plasmas and [25] for semiconductors). It can be derived from a moment expansion of kinetic models such as the Vlasov or Boltzmann–Poisson equations supplemented with a convenient closure assumption (see the references cited above).

There are two important physical length and time scales associated with this model (see, e.g., [1], [21]): the Debye length and the electron plasma period. These two scales are related one to each other by the thermal speed, which is an order one quantity. We are interested in the quasineutral regime where both parameters can be very small compared with typical macroscopic length and time scales. A standard explicit scheme must resolve these microscale phenomena in order to remain stable. The satisfaction of these constraints requires huge computational resources which make the use of explicit methods almost impracticable.

The search for schemes free of such constraints has been the subject of a vast literature. A number of works deal with particle-in-cell (PIC) methods, which are specific discretizations of the kinetic model, the Vlasov–Poisson system (see, e.g., the direct implicit method [3], [23] and the implicit moment method [26], [27]). These methods have proved extremely efficient in a number of situations, but there are still regions where short time steps must be used (see, e.g., the recent discussion in [19]). So the improvement of classical time-stepping strategies, if possible, would offer attractive perspectives. More references can be found in [9]. We also mention that the strategy developed in the present paper has been applied to PIC codes in [10].

The present work deals with fluid models. For fluid models the literature is comparatively less abundant. We can refer the reader to the pioneering work [16] and the more recent works [2], [4], [28], [29]. When the fluid models are drift-diffusion models, implicit strategies have been proposed in [31], [32], [24].

To cancel the fast scales associated with electron plasma frequency, quasineutral models have been very frequently considered [15] (see also references in [9]). Recently, two-fluid quasineutral models have been studied [6], [8], [11], [12], [13], [14].

However, in situations where quasineutral and nonquasineutral regions coexist, a specific treatment is needed to connect the quasineutral model with a nonquasineutral model across the interface. Such situations arise in sheath problems (see, e.g., [18], [20], [12]; other references can be found in [9]). In such problems, one has often to deal with a dynamic interface—the tracking of which gives rise to a complex numerical problem. Additionally, the interface dynamics is not a priori known and either must be derived from an asymptotic analysis like in [12], [13], or [20] or must be inferred from physical considerations. In both cases, great care is required to ensure that the proper dynamics is implemented. Another problem is related to the fact that the quasineutral to nonquasineutral transition may not be a sharp transition, but rather a fairly diffuse one, and its approximation into a sharp interface may actually lead to some unphysical behavior.

For these reasons, it is highly desirable to develop numerical methods which automatically shift from a quasineutral to a nonquasineutral model across the transition region when such a transition is encountered. The scheme proposed in [9] serves this purpose. Additionally, it has been shown that the numerical cost of this scheme is the same as the standard strategy (we refer the reader to [9] for more details). The goal of the present work is to provide elements toward a rigorous analysis of its stability properties. We shall mostly restrict ourselves to a linearized stability analysis (except in section 5), although linearized stability analysis is insufficient when the solution is not close to the reference state.

The paper is organized as follows. In section 2, we present the model and our AP strategy. Our AP scheme consists in an implicit time-differencing of both the flux term in the mass balance equation and the source term in the momentum equation. For this reason, we shall refer to it as an implicit-implicit II scheme. It will be compared to schemes in which either one or the other or both terms are treated explicitly (referred to as the IE, EI, or EE schemes). For instance, the IE scheme is implicit in the mass flux term and explicit in the momentum source term. The classical strategy is the EI scheme, where the mass flux term is explicit and the momentum source term is implicit [16]. The EE strategy is unconditionally unstable as was shown by Fabre in [16] and as reviewed below. We will always consider an explicit time-differencing for the momentum flux term, so there will be no mention of it in our terminology.

In section 3, the Euler–Poisson system is linearized about zero mean velocity. The linearized Euler–Poisson LEP system has a wave-equation–like formulation, which is indeed a Klein–Gordon equation with an $O(\varepsilon)$ mass term, as shown in section 3.1.

The Fourier transformed (with respect to the space variable) system is considered in section 3.2, and the stability of the time-stepping strategy is analyzed on this model system. This analysis is a model for the analysis of a fully discrete scheme (in space and time) using a centered space-differencing. Indeed, the symbol of the continuous space derivative and that of a centered space-differencing are both purely imaginary and therefore behave in a similar way for stability considerations. Using the symbol of the continuous operator instead of the discrete one makes the computations a little bit simpler (we shall see that they are already quite complex) while still providing a significant insight into the properties of the fully discrete scheme. Of course, to mimic the fact that a fully discrete scheme cannot resolve large wave-numbers, the range of admissible wave-numbers is limited to a value $\pi/\Delta x$, where Δx is the value of the underlying space discretization. To support this approach, the fully discrete scheme will be analyzed in a particular situation.

In this section, we show that the AP strategy (i.e., the II scheme) is stable uniformly with respect to ε while the IE scheme is limited by a stability constraint of the form $\delta \leq O(\varepsilon)$, where δ is the time step. The IE scheme corresponds to an explicit time discretization of the wave-equation formulation of the LEP system, while the II is an implicit one. In the case of the II scheme, a stability analysis of a fully discrete scheme is given. This analysis shows that the scheme is stable provided that ε is small enough (compared to Δx). This can be easily understood since, when ε is large, the system is close to the standard Euler system for which a centered space-differencing is known to be unstable (even though the mass flux term is treated implicitly, the momentum flux is treated explicitly, and the standard instability of the centered scheme manifests itself).

For this reason, in section 3.3, the influence of a decentered space-differencing is investigated. The numerical diffusion is modeled by a diffusion term added to the mass

and momentum balance equations in the continuous model, leading to a linearized viscous Euler–Poisson (LVEP) model. Energy estimates for this system and for a space discretization of this system are given. The similarity of these estimates encourages us to discard the space discretization, and instead, we consider the Fourier transform (with respect to the space variable) of the continuous system, and we perform a stability analysis of the various time discrete schemes under consideration. The AP strategy (i.e., the II scheme) is compared with the EI scheme (which corresponds to the standard approach; see, e.g., [16]) and the EE scheme. Only the II scheme has a uniform stability with respect to ε. The EI scheme has a stability constraint of the form $\delta \leq O(\varepsilon)$, while the EE scheme is unconditionally unstable, as already noted by Fabre [16].

Only the II scheme has the desired asymptotic stability property. It remains to check if this property remains true when the system is linearized about a nonzero velocity u_0. This is performed in section 4. In this section, we restrict ourselves to the II scheme, but we investigate the influence of the space discretization by looking first at the Fourier transformed nonviscous LEP system (mimicking a space-centered differencing) and then at the LVEP system. As already pointed out, the II scheme for the nonviscous LEP system is only "marginally" uniformly stable in the sense that it is stable if ε is small enough (depending on u_0 and other parameters such as Δx), while the II scheme for the LVEP system is uniformly stable with respect to ε: there exists a stability region for δ which is independent of ε for ε ranging in the whole interval $[0, \infty)$. This result shows that the II scheme associated with a decentered space-differencing has the requested uniform stability property with respect to ε. This has been observed numerically in the simulations shown in [9].

There remains the question of knowing whether these results which are proved for the linearized system translate to the fully nonlinear system. In section 5, we consider a toy model for the Euler–Poisson system: the Burgers–Poisson system. We show that this system has an entropy-entropy flux pair and that a space semidiscrete system shares this property with the continuous model. This leads to a stability result for this semidiscrete scheme which allows us to prove a convergence estimate (at least as long as the solution of the Burgers–Poisson problem remains regular). The question whether an II time discretization of this model would show the same stability property is still an open question which will be investigated in future work. Another open problem is to show the stability of this scheme for weak solutions in the presence of shocks. This is a difficult question which has also been put aside in the present work (see, e.g., [5], [22] for the convergence analysis of numerical schemes for entropy solutions in the case of a standard hyperbolic problem).

The extension of the present analysis to two or more space dimensions is still an open problem. For example, in section 4, we restrict ourselves to waves propagating in the same direction as the constant velocity u_0 about which the linearization is performed. A more general setting would be a propagation at an oblique angle with respect to u_0. However, the analysis of section 4 is already quite complicated, and we shall leave this extension to future work. Let us just mention that the numerical simulations reported in [9] suggest that the scheme is actually uniformly stable also in the two-dimensional case.

A short conclusion is drawn in section 6.

A last short comment is in order to address the issue of convergence. Here, we are seeking stability for large time steps compared with ε. Therefore, the scheme cannot be a good approximation of the original (ε perturbed) problem but should

be an approximation of the limit quasineutral model. In the present one-dimensional linearized setting, the quasineutral limit is trivial (the linearized solution is identically zero). It is readily seen that taking the limit ε and Δt to 0 keeping $\varepsilon \ll \Delta t$ in all the AP schemes developed below leads to this limit, showing that they are all consistent with the quasineutral model. Using this remark, and the Wendroff technique combining stability and consistency, it is certainly possible to show the convergence of the AP scheme toward the solution of the quasineutral model. However, we shall not pursue this route in this paper.

2. Quasineutral limit of the Euler–Poisson system and asymptotically stable schemes.

2.1. Quasineutral limit. In this work, we are mainly concerned with the Euler–Poisson model. For the sake of simplicity, we shall restrict ourselves to a one-species model, but our scheme construction can be easily extended to any number of charged-particle species (see [7], [9]). The scaled one-species Euler–Poisson model (EP) is written

$$\partial_t n + \nabla \cdot q = 0, \tag{2.1}$$

$$\partial_t q + \nabla \cdot \left(\frac{q \otimes q}{n} \right) + \nabla p(n) = n \nabla \phi, \tag{2.2}$$

$$\varepsilon^2 \Delta \phi = n - 1, \tag{2.3}$$

where $n = n(x,t)$ is the particle number density, $p(n)$ is the hydrodynamic pressure, $q = q(x,t)$ is the momentum (i.e., $q = nu$, where $u(x,t)$ is the average velocity), $p(n) = n^\gamma$ is the pressure law with $\gamma \geq 1$, and $\phi = \phi(x,t)$ is the potential, where $x \in \mathbb{R}^d$ is the position and $t > 0$ is the time. The scheme construction can be easily extended to a model including the full energy equation in place of the isentropic assumption on the pressure.

Here, we consider negatively charged electrons (with scaled charge equal to -1), and the right-hand side of the Poisson equation involves a uniform ion background density equal to 1 in the present scaled variables. Again, this simplification is for the sake of clarity only and a space-dependent ion background density or a model with simultaneously evolving electrons and ions could be considered as well (see [7], [9]).

The dimensionless parameter $\varepsilon = \lambda_D / L$ is the scaled Debye length, i.e., the ratio of the actual Debye length λ_D to the macroscopic length scale L. We recall that the Debye length has the following expression in physical variables:

$$\lambda_D = \left(\frac{\epsilon_0 k_B T}{e^2 n_0} \right)^{1/2},$$

where ϵ_0 is the vacuum electric permittivity, k_B the Boltzmann constant, T the average plasma temperature, e the absolute electron charge, and n_0 the average plasma density.

In many instances, the Debye length is very small compared to the macroscopic length and thus $\varepsilon \ll 1$. In this case, the quasineutral limit $\varepsilon \to 0$ of system (2.1)–(2.3) can be considered. It leads to $n = 1$; i.e., the electron density must be everywhere equal to the background ion density and to the following quasineutral model (QN)

for the flux q:

$$\nabla \cdot q = 0, \tag{2.4}$$

$$\partial_t q + \nabla \cdot (q \otimes q) = \nabla \phi. \tag{2.5}$$

We note that ϕ is now the Lagrange multiplier of the incompressibility constraint (2.4) and, upon changing ϕ into $-\pi$, that the quasineutral model is nothing but the incompressible Euler equations with hydrostatic pressure π.

In the present paper, our goal is to analyze a class of numerical schemes for EP which remains uniformly stable in the limit $\varepsilon \to 0$. Such schemes have first been proposed and tested in [7], [9], where their uniform stability with respect to ε has been demonstrated.

The design of this class of scheme is based on the following remark. In the passage from EP to QN, the equation for the electric potential ϕ changes dramatically, from the Poisson equation (2.3) into

$$\Delta \phi = \nabla^2 : (q \otimes q), \tag{2.6}$$

where ∇^2 denotes the Hessian matrix and ":" the contracted product of rank two tensors. Any attempt to design a uniformly stable scheme in the limit $\varepsilon \to 0$ requires finding a different (and equivalent) formulation of the Poisson equation (2.3) in which the transition toward (2.6) when $\varepsilon \to 0$ appears explicitly.

Such a formulation is provided by the following equation:

$$-\nabla \cdot ((n + \varepsilon^2 \partial_{tt}^2) \nabla \phi) + \nabla^2 : f(n, q) = 0, \tag{2.7}$$

where the momentum flux tensor $f(n, q)$ is given by

$$f(n, q) = \frac{q \otimes q}{n} + p(n)\mathrm{Id}. \tag{2.8}$$

Equation (2.7) is formally equivalent to the Poisson equation provided (n, q) satisfies the mass and momentum equations (2.1), (2.2) and that the Poisson equation and its time derivative are satisfied at $t = 0$:

$$(\varepsilon^2 \Delta \phi - n + 1)|_{t=0} = 0, \quad (\partial_t (\varepsilon^2 \Delta \phi - n + 1))|_{t=0} = 0.$$

We refer the reader to [7], [9] for a proof of this equivalence. Essentially, it relies on a second order (wave-like) formulation of the Euler equations and replacing the second time derivative of the density in this formulation by the second time derivative of the Laplacian of the potential (thanks to the Poisson equation). We shall see below the importance of using a wave-like formulation of the Euler equations in this matter.

We remark that solving the elliptic equation (2.7) (after time discretization) is no more costly than solving the Poisson equation (2.3) since both of them are elliptic equations. However, one important advantage of (2.7) over (2.3) is that it does not degenerate when $\varepsilon \to 0$ and moreover reduces to the QN elliptic equation (2.6) for ϕ when $\varepsilon = 0$. Therefore, we can expect that the asymptotic stability with respect to ε is obtained if a suitable time discretization is used.

2.2. Asymptotically stable scheme in the quasineutral limit. The question of finding an asymptotically stable scheme in the QN limit is in the first place a question of time discretization. We propose a time discretization which formally satisfies this asymptotic stability property. Below, we shall suppose that the symbol D stands for space discretizations of the ∇ operator, and we shall postpone the question of finding suitable expressions of such operators to forthcoming sections.

First, we describe the most standard time marching procedure for solving the EP system. Suppose n^m, q^m, ϕ^m are approximations at time t^m of n, q, and ϕ, and let $\delta = t^{m+1} - t^m$. Then, the standard procedure consists in discretizing (2.1)–(2.3) by

$$\delta^{-1}\,(n^{m+1} - n^m) + D \cdot q^m = 0, \tag{2.9}$$

$$\delta^{-1}\,(q^{m+1} - q^m) + Df(n^m, q^m) = n^{m+1} D\phi^{m+1}, \tag{2.10}$$

$$\varepsilon^2 D^2 \phi^{m+1} = n^{m+1} - 1. \tag{2.11}$$

We note that the scheme involves an implicit computation of the electric force in the momentum equation but the cost stays the same as that of a fully explicit scheme. Indeed, n^{m+1} can be computed first by the mass equation (2.9). Then, ϕ^{m+1} is obtained by the Poisson equation (2.11). Finally, the momentum q^{m+1} can be advanced using (2.10). It is common wisdom in plasma physics that the stability of this scheme is constrained by the condition

$$\delta \lesssim \varepsilon, \tag{2.12}$$

which is very penalizing in the QN regime $\varepsilon \ll 1$.

Let us note that if an explicit electric force discretization is used at the right-hand side of (2.10), the scheme is simply unconditionally unstable [16].

Our new AP strategy consists in computing the mass flux in (2.9) implicitly. More precisely, we consider the following scheme:

$$\delta^{-1}\,(n^{m+1} - n^m) + D \cdot q^{m+1} = 0, \tag{2.13}$$

$$\delta^{-1}\,(q^{m+1} - q^m) + Df(n^m, q^m) = n^m D\phi^{m+1}, \tag{2.14}$$

$$\varepsilon^2 D^2 \phi^{m+1} = n^{m+1} - 1. \tag{2.15}$$

By reproducing the procedure which, in the continuous case, allowed us to pass from the standard formulation of the Poisson equation (2.3) to the reformulated one (2.7), we can transform the discrete Poisson equation (2.15) into a reformulated one:

$$\begin{aligned} &-D \cdot ((\delta^{-2}\varepsilon^2 + n^m) D\phi^{m+1}) \\ &\quad = -D^2 : f(n^m, q^m) - \delta^{-2}(2n^m - n^{m-1} - 1) := G(n^m, q^m, n^{m-1}). \end{aligned} \tag{2.16}$$

This is a new elliptic equation for ϕ^{m+1} which is formally equivalent to the original one (2.15), is as easy to solve, and does not degenerate when $\varepsilon \to 0$.

The first important remark about this scheme is that it provides a discretization of the QN Euler system (2.4), (2.5) when $\varepsilon = 0$; i.e., it is an AP scheme when $\varepsilon \to 0$. To see this, we can use either of the formulations (2.15) or (2.16) of the Poisson equation, coupled with (2.13)–(2.14). Of course, we suppose that at step m, the data n^m and q^m satisfy $n^m = 1$ and $D \cdot q^m = 0$. If we let $\varepsilon \to 0$ in (2.15), we first find

that $n^{m+1} = 1$. Then, given that $n^m = 1$, (2.13) yields $D \cdot q^{m+1} = 0$. Finally, (2.14) leads to

$$\delta^{-1}\,(q^{m+1} - q^m) + Df(1, q^m) = D\phi^{m+1},$$

which clearly gives a discretization of the QN Euler system (2.4), (2.5). Note that the standard scheme (2.9)–(2.11), which has an explicit mass flux, would lead to $D \cdot q^m = 0$, the other limiting equations remaining unchanged, and that this would not allow us to compute ϕ^{m+1}.

If one uses (2.16) instead of (2.15), then one first lets $\varepsilon \to 0$ in (2.16) leading to

$$-D \cdot (D\phi^{m+1}) = -D^2 : f(1, q^m)$$

(since $n^m = 1$), which is a discretization of the quasineutral potential equation (2.6). Then, by acting operator $D\cdot$ on (2.14) and using that $D \cdot q^m = 0$, we deduce that $D \cdot q^{m+1} = 0$. Finally, inserting this into (2.13), we get that $n^{m+1} = n^m = 1$, and again, we get a discretization of the QN Euler system (2.4), (2.5). Here, using the standard scheme (2.9), (2.10) (which has an explicit mass flux) in conjunction with the modified Poisson equation (2.16) would also lead to an AP scheme. However, it has been numerically experienced that this scheme is unstable, and so this possibility will be disregarded.

The second important remark is that the reformulated Poisson equation (2.16) allows us to compute ϕ^{m+1} in terms of known quantities $G(n^m, q^m, n^{m-1})$ from the previous time step. Indeed, once ϕ^{m+1} is known after solving (2.16), q^{m+1} can be computed by advancing the momentum equation (2.14), and then, n^{m+1} in turn is obtained by using the density equation (2.13). Therefore, the time marching procedure of the AP scheme (2.13)–(2.15) involves the same computational cost as that of the standard scheme (2.9)–(2.11).

Finally, even more importantly, we claim that, by contrast with the standard scheme, the AP scheme (2.13)–(2.15) is not constrained by the stability condition (2.12) but rather has a stability condition of the form

$$\delta = O(1) \quad \text{as } \varepsilon \to 0. \tag{2.17}$$

We stress, however, that the scheme is still constrained by the CFL condition of the hydrodynamic system (which is an $O(1)$ quantity in our scaled units), and this can be penalizing for electrons due to their very small mass. It is possible, however, to extend this strategy into a scheme which is asymptotically stable also in the small mass limit (see remark in [9]). This strategy will be developed in future works. We also note that similar ideas can be applied to the low Mach number limit of compressible Euler equations and to the drift approximation of highly magnetized plasmas (works also in progress).

The stability estimate (2.17) is so far formal. It has been verified in numerical computations using different types of space discretizations in [7], [9]. However, no rigorous arguments have been available so far. The goal of the present paper is to provide some rigorous support to the estimate (2.17) by analyzing the linearized system.

Two types of analyses will be performed. The first one uses Fourier analysis and keeps the time semidiscrete framework, mimicking space discretization by suitable Fourier symbols. The second one considers both a space semidiscrete setting (the so-called method of lines) and a full time and space discretization and develops energy estimates. Both methods show that the AP strategy gives rise to a stability constraint of the type (2.17) while the standard strategy has the more severe stability constraint (2.12).

3. The linearized system about zero velocity.

3.1. The linearization of the Euler–Poisson system. We linearize system (2.1)–(2.3) about the stationary homogeneous solution $n_0 = 1$, $q_0 = 0$, $\partial_x \phi_0 = 0$. We denote by $\zeta \ll 1$ the size of the initial perturbation to this stationary state. Then, we can formally expand the solution as follows: $n = 1 + \zeta n' + O(\zeta^2)$, $q = \zeta u' + O(\zeta^2)$, $\phi = \zeta \phi' + O(\zeta^2)$. Retaining only terms of order ζ, we are led to the following linearized Euler–Poisson LEP system (where the primes have been dropped for the sake of clarity):

$$\partial_t n + \partial_x u = 0, \tag{3.1}$$

$$\partial_t u + c_s^2 \partial_x n = \partial_x \phi, \tag{3.2}$$

$$\varepsilon^2 \partial_{xx} \phi = n, \tag{3.3}$$

where $c_s = \sqrt{p'(1)}$ is the speed of sound.

We note that both ϕ and u can be eliminated and the LEP system can be rephrased into a wave-like formulation (specifically a Klein–Gordon equation with mass equal to $1/\varepsilon^2$):

$$\partial_{tt}^2 n - c_s^2 \partial_{xx}^2 n + \frac{1}{\varepsilon^2} n = 0. \tag{3.4}$$

By a partial Fourier transform with respect to the x-variable, and denoting $\hat{n}(\xi, t)$, $\hat{u}(\xi, t)$, $\hat{\phi}(\xi, t)$ the transforms of n, u, and ϕ, we obtain

$$\partial_t \hat{n} + i\xi \hat{u} = 0, \tag{3.5}$$

$$\partial_t \hat{u} + i\xi c_s^2 \hat{n} = i\xi \hat{\phi}, \tag{3.6}$$

$$-\varepsilon^2 \xi^2 \hat{\phi} = \hat{n}. \tag{3.7}$$

After elimination of $\hat{\phi}$, the system reduces to a 2×2 system of the form

$$\begin{pmatrix} \partial_t \hat{n} \\ \partial_t \hat{u} \end{pmatrix} + \begin{pmatrix} 0 & i\xi \\ i\xi c_s^2 + \frac{i}{\varepsilon^2 \xi} & 0 \end{pmatrix} \begin{pmatrix} \hat{n} \\ \hat{u} \end{pmatrix} = 0. \tag{3.8}$$

Its solutions are of the form

$$\begin{pmatrix} \hat{n}(\xi, t) \\ \hat{u}(\xi, t) \end{pmatrix} = \begin{pmatrix} \hat{n}_+(\xi) \\ \hat{u}_+(\xi) \end{pmatrix} e^{i\theta^\varepsilon(\xi)t} + \begin{pmatrix} \hat{n}_-(\xi) \\ \hat{u}_-(\xi) \end{pmatrix} e^{-i\theta^\varepsilon(\xi)t}, \tag{3.9}$$

where

$$\theta^\varepsilon(\xi) = \left(c_s^2 \xi^2 + \frac{1}{\varepsilon^2} \right)^{1/2}. \tag{3.10}$$

The quantities $s_\pm = \pm i\theta^\varepsilon(\xi)$ are the two eigenvalues of the matrix A at the left-hand side of (3.8). The vectors $(\hat{n}_\pm(\xi), \hat{u}_\pm(\xi))^T$ are obtained as the projections of the initial data $(\hat{n}_0, \hat{u}_0)^T$ onto the associated eigenspaces of A (here the exponent T denotes the transpose).

From (3.9) and the fact that $|e^{\pm i\theta(\xi)t}| = 1$, we see that the L^2 norm of the solution in Fourier space stays bounded in time, which by Plancherel's identity shows that the corresponding L^2 norm in physical space also stays bounded in time. Additionally, when $\varepsilon \to 0$, we have

$$\theta^\varepsilon \sim \frac{1}{\varepsilon} \to \infty,$$

which shows that the solution has high frequency time oscillations of frequency $\sim \frac{1}{\varepsilon}$, irrespective of the value of the wave-number ξ.

Therefore, the quasineutral limit $\varepsilon \to 0$ gives rise to high frequency oscillations which are nothing but the ubiquitous plasma oscillations of the physics literature. Explicit schemes are constrained to resolve these oscillations as a condition for their stability, as we shall see below.

Another viewpoint which emphasizes the wave-like behavior of this system is that of energy estimates. Multiplying (3.4) by $2n_t$, one has

$$\partial_t\Big((\partial_t n)^2 + c_s^2(\partial_x n)^2 + \frac{1}{\varepsilon^2}n^2\Big) = 2c_s^2\partial_x\Big(\partial_x n\partial_t n\Big). \tag{3.11}$$

Integrating over the computational domain, say, (a, b), to fix the idea, we get

$$\partial_t \int_a^b \Big((\partial_t n)^2 + c_s^2(\partial_x n)^2 + \frac{1}{\varepsilon^2}n^2\Big)\,dx = 2c_s^2\Big(\partial_x n\partial_t n\Big)\Big|_a^b. \tag{3.12}$$

If periodic or homogeneous Dirichlet or homogeneous Neumann conditions are imposed at the boundary points, the right-hand side of the above equation vanishes and conservation of energy follows:

$$\partial_t \int_a^b \Big((\partial_t n)^2 + c_s^2(\partial_x^2 n)^2 + \frac{1}{\varepsilon^2}n^2\Big)\,dx = 0. \tag{3.13}$$

This energy estimate indicates a wave-like behavior. When $\varepsilon \to 0$, more and more energy is stored in the last term, which can be interpreted as a potential.

Below, we state stability criteria for various kinds of time discretizations of the LEP system.

3.2. AP strategy with centered-like space-differencing. For the LEP system in Fourier variables (3.5)–(3.7), our scheme (2.13)–(2.15) reads

$$\frac{\hat{n}^{m+1} - \hat{n}^m}{\delta} + i\xi\hat{u}^{m+1} = 0, \tag{3.14}$$

$$\frac{\hat{u}^{m+1} - \hat{u}^m}{\delta} + i\xi c_s^2\hat{n}^m + \frac{i}{\varepsilon^2\xi}\hat{n}^{m+1} = 0. \tag{3.15}$$

It is easy to see that this scheme can be recast into an implicit scheme for the wave-equation (3.4) expressed in Fourier variables, i.e.,

$$\partial_{tt}^2\hat{n} + c_s^2\xi^2\hat{n} + \frac{1}{\varepsilon^2}\hat{n} = 0. \tag{3.16}$$

Indeed, taking the time difference of (3.14) and combining it with (3.15) to eliminate the velocity leads to

$$\frac{\hat{n}^{m+1} - 2\hat{n}^m + \hat{n}^{m-1}}{\delta^2} + c_s^2\xi^2\hat{n}^m + \frac{1}{\varepsilon^2}\hat{n}^{m+1} = 0. \tag{3.17}$$

By contrast, let us consider the following explicit scheme for the wave-equation formulation:

$$\frac{\hat{n}^{m+1} - 2\hat{n}^m + \hat{n}^{m-1}}{\delta^2} + c_s^2 \xi^2 \hat{n}^m + \frac{1}{\varepsilon^2}\hat{n}^m = 0. \tag{3.18}$$

This scheme can be obtained by eliminating the velocity from the original formulation of the LEP system. Two schemes for the original LEP system lead to the same scheme (3.18). The first one is implicit in the mass flux and explicit in the momentum flux and in the source term:

$$\frac{\hat{n}^{m+1} - \hat{n}^m}{\delta} + i\xi \hat{u}^{m+1} = 0, \tag{3.19}$$

$$\frac{\hat{u}^{m+1} - \hat{u}^m}{\delta} + i\xi c_s^2 \hat{n}^m + \frac{i}{\varepsilon^2 \xi}\hat{n}^m = 0. \tag{3.20}$$

The second one is explicit in the mass flux and implicit in the momentum flux and the source term:

$$\frac{\hat{n}^{m+1} - \hat{n}^m}{\delta} + i\xi \hat{u}^m = 0, \tag{3.21}$$

$$\frac{\hat{u}^{m+1} - \hat{u}^m}{\delta} + i\xi c_s^2 \hat{n}^{m+1} + \frac{i}{\varepsilon^2 \xi}\hat{n}^{m+1} = 0. \tag{3.22}$$

We see that the schemes (3.14)–(3.15) and (3.19)–(3.20) are both implicit in the mass flux. The first one is also implicit in the source term while the second one is explicit. For these reasons, we shall refer to the first scheme as the II scheme (for implicit-implicit) while the second one is the IE scheme (for implicit-explicit). What the equivalence of scheme (3.21), (3.22) with the scheme (3.18) shows is that if the mass flux is taken to be explicit, there is no way to obtain a uniformly stable scheme, no matter how implicit one takes the momentum flux or the source term to be.

If additionally the momentum flux is taken as explicit and only the source term is taken as implicit, one is led to the following scheme:

$$\frac{\hat{n}^{m+1} - \hat{n}^m}{\delta} + i\xi \hat{u}^m = 0, \tag{3.23}$$

$$\frac{\hat{u}^{m+1} - \hat{u}^m}{\delta} + i\xi c_s^2 \hat{n}^m + \frac{i}{\varepsilon^2 \xi}\hat{n}^{m+1} = 0, \tag{3.24}$$

which is equivalently written in the wave-equation formulation as follows:

$$\frac{\hat{n}^{m+1} - 2\hat{n}^m + \hat{n}^{m-1}}{\delta^2} + c_s^2 \xi^2 \hat{n}^{m-1} + \frac{1}{\varepsilon^2}\hat{n}^m = 0. \tag{3.25}$$

In this scheme the mass flux is taken as explicit and the source term as implicit. For this reason, we call this scheme the EI scheme. We shall analyze the stability of this scheme in the next section in relation to decentered space-differencing. Indeed, we know already that with a centered space-differencing, the hydrodynamic part of this scheme will be unconditionally unstable and that it is unlikely that the implicit source term restores stability, and if it does so, it will do it only marginally.

For all the reasons previously exposed, in the remainder of this paper, we will always treat the momentum flux explicitly. We will consider various combinations of

explicit or implicit schemes for the mass flux and the source term, and the terminology will refer to these terms. For instance, the IE scheme refers to a scheme that is implicit in the mass flux but explicit in the source term.

Looking at the stability of a time-semidiscretized scheme in Fourier space gives a good indication of the behavior of a fully discretized scheme associated with a space-centered discretization. Indeed, the symbol of a space-centered discretization is purely imaginary, like the symbol of the continuous space derivative. Of course, there might exist some quantitative differences between the fully discrete scheme and the semidiscrete scheme in Fourier space, but we believe that this analysis gives a correct rough picture. Using the space continuous Fourier transform allows us also to considerably simplify the computations and to express the results in a more compact form, which makes them easier to interpret.

In view of a space discretization of mesh size Δx, we assume that the range of admissible wave-numbers ξ is $|\xi| \le \xi^* = \pi/\Delta x$.

We can show (the proof is technical and is omitted) that a sufficient stability condition for the II scheme for a given wave-number ξ is

$$\delta \le \delta_{ii}(\xi) := \frac{2(\frac{1}{\varepsilon^2} + c_s^2\xi^2)^{1/2}}{c_s^2\xi^2}. \tag{3.26}$$

The II scheme is stable over the range of admissible wave-numbers ξ as soon as

$$\delta \le \delta_{ii}^* := \delta_{ii}(\xi^*) = 2\left(\frac{\Delta x}{\pi c_s}\right)^2\left(\frac{1}{\varepsilon^2} + \left(\frac{\pi c_s}{\Delta x}\right)^2\right)^{1/2}. \tag{3.27}$$

By contrast, the stability domain of the IE scheme for a given wave-number ξ is given by

$$\delta \le \delta_{ie}(\xi) := \frac{2}{(\frac{1}{\varepsilon^2} + c_s^2\xi^2)^{1/2}}. \tag{3.28}$$

The IE scheme is stable over the range of admissible wave-numbers iff

$$\delta \le \delta_{ie}^* := \delta_{ie}(\xi^*) = \frac{2}{(\frac{1}{\varepsilon^2} + \left(\frac{\pi c_s}{\Delta x}\right)^2)^{1/2}}. \tag{3.29}$$

The stability domain of the IE scheme shrinks when $\varepsilon \to 0$, while that of the II scheme turns out to be uniformly stable with respect to ε. Furthermore, the II scheme is also consistent with the limit model when $\varepsilon \to 0$ (AP property) as we have seen in section 2.2 in the case of the fully nonlinear equation (the proof is the same for the linearized equation and is left to the reader).

Indeed, we see that $\delta_{ie}^* = O(\varepsilon)$ as $\varepsilon \to 0$ with a fixed Δx, while $\delta_{ii}^* = O(\varepsilon^{-1})$, showing the announced behavior. More precisely, $\delta_{ii}^* \sim 2(\frac{\Delta x}{2\pi c_s})^2\frac{1}{\varepsilon}$. We note that this increased stability with respect to ε is paid for by a more restrictive CFL condition of diffusion type with respect to the acoustic waves. By contrast, in the absence of a source term or, equivalently, when $\Delta x \to 0$ for a fixed ε, the stability constraints for both the IE and II schemes reduce to $\delta \le 2\frac{\Delta x}{\pi c_s}$, which is similar to a CFL condition. We can also show that the II scheme damps the plasma oscillations. At each time step, the amplitude of the oscillation is multiplied by a factor ε/δ (in the limit $\varepsilon \to 0$ for fixed δ). When ε is very small, the damping is very strong. At each time step, the phase of the solution is multiplied by the factor $e^{\pm i\theta}$ with

$$\theta = \cos^{-1}\frac{(1 - \frac{1}{2}\delta^2 c_s^2\xi^2)}{(1 + \frac{\delta^2}{\varepsilon^2})^{1/2}} \quad \overset{\varepsilon\to 0}{\longrightarrow} \quad \frac{\pi}{2}, \tag{3.30}$$

which means that, in the limit $\varepsilon \to 0$, a shift of $\pi/2$ is added to the phase of the oscillation at each time step. In this limit, the period of the plasma oscillations stays bounded, equal to 4δ.

To support the conclusions of this Fourier analysis of the scheme, we show below that the same stability and dissipation properties are still valid when we use a fully discretized scheme in both time and space.

We denote by $D^2u = \frac{u_{i+1}-2u_i+u_{i-1}}{h^2}$ the standard second order difference operator and by $D_-u = \frac{u_i-u_{i-1}}{h}$ and $D_+u = \frac{u_{i+1}-u_i}{h}$ the standard one-sided difference operator. Then $D^2 = D_+D_- = D_-D_+$. Denote $\tilde{D} = (D_+ + D_-)/2$. Here h is the space step.

We consider a space discretization of (3.4) by replacing the continuous second order space derivative by the discrete operator D^2:

$$\partial_{tt}^2 n - c_s^2 D^2 n + \frac{1}{\varepsilon^2} n = 0. \tag{3.31}$$

Applying the II time-difference scheme (3.17) to this equation leads to

$$\frac{n^{m+1} - 2n^m + n^{m-1}}{\delta^2} - c_s^2 D^2 n^m + \frac{1}{\varepsilon^2} n^{m+1} = 0. \tag{3.32}$$

This scheme can be viewed as a wave-type version of the following scheme for the original LEP system (3.1)–(3.3):

$$\frac{n^{m+1} - n^m}{\delta} + D_- u^{m+1} = 0, \tag{3.33}$$

$$\frac{u^{m+1} - u^m}{\delta} + c_s^2 D_+ n^m = D_+ \phi^{m+1}, \tag{3.34}$$

$$\varepsilon^2 D^2 \phi^{m+1} = n^{m+1}. \tag{3.35}$$

Indeed, eliminating u between (3.33) and (3.34), one has

$$\frac{n^{m+1} - 2n^m + n^{m-1}}{\delta^2} - c_s^2 D^2 n^m + D^2 \phi^{m+1} = 0, \tag{3.36}$$

and using (3.35) we get (3.32). We can also eliminate n^{m+1} from (3.36) using (3.35) and get a discretization of the reformulated Poisson equation (2.7) (in linearized form):

$$\left(\frac{\varepsilon^2}{\delta^2} + 1\right) D^2 \phi^{m+1} = \frac{2n^m - n^{m-1}}{\delta^2} + c_s^2 D^2 n^m. \tag{3.37}$$

We can show (the proof is technical and is omitted) that this scheme is stable provided that $2\varepsilon c_s \le h$. We observe that the scheme is asymptotically stable in the limit $\varepsilon \to 0$ but is not stable if ε is large. This feature can easily be explained. When $\varepsilon \to \infty$ the electric field term disappears and we are left with the standard LEP system discretized via centered differences. A time explicit discretization will be unconditionally unstable. Here, the time-differencing is implicit in the mass flux term but still explicit in the momentum flux. This semi-implicit discretization is not implicit enough to stabilize the space-centered discretization. It is a remarkable fact that the electric field contributes to stabilize the instability of the space-centered scheme. However, this scheme cannot be used in a far from quasineutrality regime,

which is a severe drawback in situations close to quasineutrality in some regions but far from it in others. This unpleasant feature will also be found via Fourier analysis when linearizing the Euler–Poisson system about a nonzero velocity (see section 4). This drawback will be overcome by means of a space-decentered discretization.

We also note that the implicit treatment of the last term in (3.32) gives a strong dissipation (like in the Fourier analysis case). This can be seen from the use of the so-called equivalent equation (i.e., the equation obtained by retaining the leading order term in the Taylor expansion of the time semidiscrete scheme). This equivalent equation is written

$$\partial_{tt}^2 n - c_s^2 \partial_{xx}^2 n + \frac{1}{\varepsilon^2} n + \frac{\delta}{\varepsilon^2} n_t = 0, \tag{3.38}$$

which can be recast as an LEP system with a damping term

$$\partial_t n + \partial_x u = -\tfrac{\delta}{\varepsilon^2} n, \tag{3.39}$$

$$\partial_t u + c_s^2 \partial_x n = \partial_x \phi, \tag{3.40}$$

$$\varepsilon^2 \partial_{xx}^2 \phi = n. \tag{3.41}$$

The energy estimate (3.13) for (3.38) becomes

$$\partial_t \int_a^b \left((\partial_t n)^2 + c_s^2 (\partial_x n)^2 + \frac{1}{\varepsilon^2} n^2 \right) dx + \frac{2\delta}{\varepsilon^2} \int_a^b (\partial_t n)^2 \, dx = 0, \tag{3.42}$$

and we observe the strong damping of the energy with a rate which tends to infinity as $\varepsilon \to 0$ (for fixed δ) as fast as $1/\varepsilon^2$.

The use of centered differencing associated with only a partial implicit time-differencing may lead to stability problems when the coupling with the field term is weak, i.e., ε is large. This will be particularly clear with the analysis of the LEP system about a nonzero velocity in section 4. For this reason, we now analyze schemes associated with decentered space-differencing, which we shall mimic by looking at the LEP system with diffusion.

3.3. AP strategy associated with upwind space-decentered differencing. In order to mimic the influence of a decentered space-difference, we consider the following two viscous perturbation of the LEP system. The linearized viscous Euler–Poisson (LVEP) system reads either as

$$\partial_t n + \partial_x u = 0, \tag{3.43}$$

$$\partial_t u + c_s^2 \partial_x n = \nu \partial_{xx}^2 u + \partial_x \phi, \tag{3.44}$$

$$\varepsilon^2 \partial_{xx} \phi = n \tag{3.45}$$

if numerical diffusion is added only to the momentum balance equation or as

$$\partial_t n + \partial_x u = \beta \partial_{xx}^2 n, \tag{3.46}$$

$$\partial_t u + c_s^2 \partial_x n = \nu \partial_{xx}^2 u + \partial_x \phi, \tag{3.47}$$

$$\varepsilon^2 \partial_{xx} \phi = n \tag{3.48}$$

if we also consider numerical diffusion in the mass conservation equation. Here, β and ν are artificial viscosity coefficients. These systems will be referred to below, respectively, as the LVEP-1 and LVEP-2 systems. The LVEP-1 system is of course a particular case of the LVEP-2 system by taking $\beta = 0$.

Eliminating u and ϕ, the LVEP-1 system reduces to the following Klein–Gordon equation with a viscoelastic term:

$$\partial_{tt}^2 n = c_s^2 \partial_{xx}^2 n - \frac{1}{\varepsilon^2} n + \nu \partial_{txx}^3 n. \tag{3.49}$$

For the LVEP-2 system, following Slemrod, we introduce $v = u - \beta \partial_x n$. Then the above system is equivalent to

$$\partial_t n + \partial_x v = 0, \tag{3.50}$$

$$\partial_t v + c_s^2 \partial_x n = (\nu + \beta) \partial_{xx}^2 u + \nu\beta \partial_{xxx}^3 n + \partial_x \phi, \tag{3.51}$$

$$\varepsilon^2 \partial_{xx}^2 \phi = n. \tag{3.52}$$

Eliminating v and ϕ, we are led to the following Klein–Gordon equation with viscoelastic and capillary-type terms:

$$\partial_{tt}^2 n = c_s^2 \partial_{xx}^2 n - \frac{1}{\varepsilon^2} n + (\nu + \beta) \partial_{txx}^3 n - \nu\beta \partial_{xxxx}^4 n. \tag{3.53}$$

We now perform Fourier and energy analyses for the LVEP-2 system which includes the LVEP-1 system as a particular case. After a partial Fourier transform with respect to x, the system reads

$$\partial_t \hat{n} + i\xi \hat{u} + \beta \xi^2 \hat{n} = 0, \tag{3.54}$$

$$\partial_t \hat{u} + i\xi c_s^2 \hat{n} + \nu \xi^2 \hat{u} = i\xi \hat{\phi}, \tag{3.55}$$

$$-\varepsilon^2 \xi^2 \hat{\phi} = \hat{n}, \tag{3.56}$$

or, equivalently, after elimination of $\hat{\phi}$,

$$\begin{pmatrix} \partial_t \hat{n} \\ \partial_t \hat{u} \end{pmatrix} + \begin{pmatrix} \beta\xi^2 & i\xi \\ i\xi c_s^2 + \frac{i}{\varepsilon^2 \xi} & \nu\xi^2 \end{pmatrix} \begin{pmatrix} \hat{n} \\ \hat{u} \end{pmatrix} = 0. \tag{3.57}$$

For $|\xi|$ not too large, the eigenmodes of the LVEP-2 system are given by

$$s^{\pm} = r_{\beta,\nu}(\xi) \pm i\theta_\varepsilon(\xi) := \frac{\beta + \nu}{2} \xi^2 \pm i \left(\frac{1}{\varepsilon^2} + c_s^2 \xi^2 - \frac{(\beta + \nu)^2}{4} \xi^4 \right)^{1/2}. \tag{3.58}$$

Now the oscillations are damped by an exponential factor $e^{-r_{\beta,\nu}t}$ which is independent of ε and proportional to $\beta+\nu$. Moreover, the phase oscillates with a frequency $O(\frac{1}{\varepsilon})$ when $\varepsilon \to 0$, independently of ξ.

In order to derive energy estimates, we multiply (3.53) with $\partial_t n$ and get

$$\partial_t\Big(\tfrac{1}{2}(\partial_t n)^2\Big) = c_s^2\,\partial_t n\,\partial_{xx}^2 n - \tfrac{1}{\varepsilon^2}\partial_t\Big(\tfrac{1}{2}n^2\Big) + (\nu+\beta)\,\partial_t n\,\partial_{txx}^3 n - \nu\beta\,\partial_t n\,\partial_{xxxx}^4 n.$$

Simple computations give

$$\begin{aligned}
\partial_t n\,\partial_{xx}^2 n &= \partial_x(\partial_x n\,\partial_t n) - \partial_x n\,\partial_{tx}^2 n = \partial_x(\partial_x n\,\partial_t n) - \partial_t\Big(\tfrac{1}{2}(\partial_x n)^2\Big),\\
\partial_t n\,\partial_{txx}^3 n &= \partial_x(\partial_{tx}^2 n\,\partial_t n) - (\partial_{tx}^2 n)^2,\\
\partial_t n\,\partial_{xxxx}^4 n &= \partial_x(\partial_{xxx}^3 n\,\partial_t n - \partial_{tx}^2 n\,\partial_{xx}^2 n) + \partial_{txx}^3 n\,\partial_{xx}^2 n\\
&= \partial_x(\partial_{xxx}^3 n\,\partial_t n - \partial_{tx}^2 n\,\partial_{xx}^2 n) + \partial_t\Big(\tfrac{1}{2}(\partial_{xx}^2 n)^2\Big).
\end{aligned}$$

Summing up, we get

$$\begin{aligned}
&\partial_t\tfrac{1}{2}\Big((\partial_t n)^2 + c_s^2(\partial_x n)^2 + \tfrac{1}{\varepsilon^2}n^2 + \nu\beta(\partial_{xx}^2 n)^2\Big) + (\beta+\nu)(\partial_{tx}^2 n)^2\\
&= \partial_x\Big(c_s^2 n_x\partial_t n + (\nu+\beta)\partial_{tx}^2 n\,\partial_t n - \nu\beta(\partial_{xxx}^3 n\,\partial_t n - \partial_{tx}^2 n\,\partial_{xx}^2 n)\Big).
\end{aligned}$$

For periodic or homogeneous Neumann boundary conditions, we obtain

$$\partial_t\tfrac{1}{2}\int_a^b \Big((\partial_t n)^2 + c_s^2(\partial_x n)^2 + \tfrac{1}{\varepsilon^2}n^2 + \nu\beta(\partial_{xx}^2 n)^2\Big)\,dx + (\beta+\nu)\int_a^b (\partial_{tx}^2 n)^2\,dx = 0.$$

Again we see that the second order dissipative coefficient is dependent only on $\nu+\beta$.

This energy estimate is still true for a space-discrete version of the LVEP-2 system with $\beta=\nu$ (simply referred to as the LVEP system). Indeed, the natural finite difference space discretization for the LVEP system is given by (with the same notation as above):

$$\partial_t n + \tilde{D}u = \beta D^2 n, \tag{3.59}$$

$$\partial_t u + c_s^2\tilde{D}n = \beta D^2 u + \tilde{D}\phi, \tag{3.60}$$

$$\varepsilon^2 D^2\phi = n, \tag{3.61}$$

where $\tilde{D} = (D_+ + D_-)/2$ and $D^2 = D_+D_-$.

Both the upwinding and Lax–Friedrichs schemes for LEP (3.1)–(3.3) take the same form with $\beta = c_s h/2$ and $\beta = \frac{h^2}{2\delta}$, respectively, as we will explain below. This is also true for the modified Lax–Friedrichs scheme used in [7] and [9], because in the linearized case, the modified Lax–Friedrichs scheme and the upwind scheme coincide. Indeed, for the upwinding scheme, we first diagonalize the LEP system (3.1)–(3.3) and then take upwind difference as follows:

$$\begin{aligned}
\partial_t(u+c_s n) + c_s D_-(u+c_s n) &= \tilde{D}\phi,\\
\partial_t(u-c_s n) - c_s D_+(u-c_s n) &= \tilde{D}\phi,\\
\varepsilon^2 D^2\phi &= n.
\end{aligned}$$

Recast to the original form, it can be written

$$\partial_t n + \tilde{D}u = \frac{c_s}{2}(D_+ - D_-)n,$$

$$\partial_t u + c_s^2 \tilde{D}n = \frac{c_s}{2}(D_+ - D_-)u + \tilde{D}\phi,$$

$$\varepsilon^2 D^2 \phi = n.$$

Note that $D_+ - D_- = hD^2$. The above is the same as (3.59)–(3.61) with $\beta = c_s h/2$. For the Lax–Friedrichs scheme we first write the scheme in its usual form:

$$(n^{m+1} - n^m)/\delta + \tilde{D}u^m = \frac{h^2}{2\delta} D^2 n^m,$$

$$(u^{m+1} - u^m)/\delta + c_s^2 \tilde{D}n^m = \frac{h^2}{2\delta} D^2 u^m + \tilde{D}\phi^{m+1},$$

$$\varepsilon^2 D^2 \phi^{m+1} = n^{m+1}.$$

We can regard it as a time discretization scheme of (3.59)–(3.61) with $\beta = \frac{h^2}{2\delta}$.

We can show that the time continuous scheme (3.59)–(3.61) has the following energy estimate for the periodic boundary condition case:

$$\partial_t \tfrac{1}{2}\left(c_s^2 \|D_- n\|^2 + \|D_- u\|^2 + \frac{1}{\varepsilon^2}\|n\|^2\right) + \beta(c_s^2 \|D^2 n\|^2 + \|D^2 u\|^2) + \frac{1}{\varepsilon^2}\|D_- n\|^2 = 0. \tag{3.62}$$

This indicates that the space discretization has the same behavior as the space-continuous model. For this reason, from now on, we shall consider only the problem of the stability of the time discretization and replace the influence of the space discretization by a Fourier transform, limiting the admissible wave-numbers to a bounded interval (the latter being related to the space step Δx).

In the remainder of this analysis, we will be using the LVEP-2 system (3.46)–(3.48) with $\beta = \nu$ and will simply refer to it as the LVEP system. The coefficient β is linked with numerical diffusion. For most schemes,

$$\beta = O(c_s \Delta x). \tag{3.63}$$

Again, we suppose that $|\xi| \le \xi^* = \pi/\Delta x$, having in mind that this mimics a space discretization of mesh step Δx. Therefore, assuming (3.63), there exists a constant $C^* > 0$ such that

$$\frac{\beta|\xi|}{c_s} \le C^*. \tag{3.64}$$

The AP strategy (II scheme) for the LVEP system reads as follows:

$$\frac{\hat{n}^{m+1} - \hat{n}^m}{\delta} + i\xi \hat{u}^{m+1} + \beta \xi^2 \hat{n}^m = 0, \tag{3.65}$$

$$\frac{\hat{u}^{m+1} - \hat{u}^m}{\delta} + i\xi c_s^2 \hat{n}^m + \beta \xi^2 \hat{u}^m + \frac{i}{\varepsilon^2 \xi}\hat{n}^{m+1} = 0. \tag{3.66}$$

We shall compare this strategy with an explicit time-differencing of the mass flux term that is still implicit in the Poisson source term. Such a scheme (EI scheme) reads

$$\frac{\hat{n}^{m+1}-\hat{n}^m}{\delta}+i\xi\hat{u}^m+\beta\xi^2\hat{n}^m=0, \tag{3.67}$$

$$\frac{\hat{u}^{m+1}-\hat{u}^m}{\delta}+i\xi c_s^2\hat{n}^m+\beta\xi^2\hat{u}^m+\frac{i}{\varepsilon^2\xi}\hat{n}^{m+1}=0. \tag{3.68}$$

Note that in section 3.2, we mentioned the converse strategy (implicit in the mass flux, and explicit in the Poisson source term; see system (3.19), (3.20)) called the IE scheme and we showed that in the nonviscous case, the stability requirement scales down to 0 as $\varepsilon\to 0$. It is unlikely that the use of an artificial viscosity can cure this instability, since the cause of this instability is clearly linked to the coupling with the field (because of its dependence with respect to ε) while the motivation of the artificial viscosity is to compensate for the instability of the space-centered scheme of the purely hydrodynamic part of the model. For this reason and for the sake of conciseness, we do not consider the IE scheme in the viscous case.

For completeness, we mention the full explicit scheme, in both mass flux and Poisson source term. This scheme (EE scheme) reads as follows:

$$\frac{\hat{n}^{m+1}-\hat{n}^m}{\delta}+i\xi\hat{u}^m+\beta\xi^2\hat{n}^m=0, \tag{3.69}$$

$$\frac{\hat{u}^{m+1}-\hat{u}^m}{\delta}+i\xi c_s^2\hat{n}^m+\beta\xi^2\hat{u}^m+\frac{i}{\varepsilon^2\xi}\hat{n}^m=0. \tag{3.70}$$

We mention that the EE and EI schemes for the nonviscous system would be unconditionally unstable because of the well-known instability of space-centered schemes for hydrodynamics. It is easy to check this instability directly. The computation is left to the reader.

For these schemes, a stability analysis (omitted) leads to the following conclusions. The AP II scheme shows a uniform stability condition with respect to the parameter ε, as expected. More precisely, a sufficient stability condition is

$$\delta\le\frac{2\beta}{c_s^2}F(C^*), \tag{3.71}$$

where $F(C^*)$ is a convenient function of the constant C^* (see (3.64)). Again, this looks like a CFL condition, which depends only on the purely hydrodynamic part of the scheme and not on ε.

By contrast, the EI scheme has a stability domain which shrinks down to 0 as $\varepsilon\to 0$. The analysis is simplified by making the following assumption:

$$\delta\le\frac{2\beta}{c_s^2}\,\frac{1}{1+C^{*2}}:=\delta^*, \tag{3.72}$$

where C^* is given by (3.64). This condition is again similar to a CFL condition (see (3.63)). Under this condition, a sufficient stability constraint for this scheme is of the form

$$\delta\le\frac{2}{\sqrt{1+C^{*2}}}\,\varepsilon\quad\forall\varepsilon\in\left[0,\frac{\beta}{c_s^2\sqrt{1+C^{*2}}}\right]. \tag{3.73}$$

Additionally if δ satisfying (3.72) is such that $\delta > 2\varepsilon$, the scheme is unstable. Therefore, the stability domain is an interval of size exactly $O(\varepsilon)$. This scheme is the linearized version of the standard scheme used for the Euler–Poisson problem. The conditional stability of this scheme in a fully discrete version was proved earlier by Fabre in [16].

Finally, the fully explicit EE scheme has disastrous behavior. Its stability condition for a given wave-number ξ is given by

$$\delta \leq \delta_{ee}(\xi) := \frac{2\beta\xi^2}{\beta^2\xi^4 + c_s^2\xi^2 + \frac{1}{\varepsilon^2}}. \tag{3.74}$$

Over the range of admissible wave-numbers $[0, \xi^*]$, the minimal value of $\delta_{ee}(\xi)$ is zero, showing that the scheme is actually unstable. Therefore, numerical diffusion is not sufficient to stabilize the EE scheme. This was proved by Fabre for a fully discrete version of the scheme in [16].

To summarize, the EE scheme is unconditionally unstable, even used in conjunction with a decentered space discretization of the hydrodynamic fluxes. The EI scheme is conditionally stable: the time step must be kept proportional to ε. Finally, the II scheme is asymptotically stable when $\varepsilon \to 0$. Its stability requirement involves only a CFL condition related to the space discretization of the hydrodynamic part of the model.

In the next section, we investigate if these conclusions remain true when the LEP system is linearized about a nonzero velocity. We show that the II scheme remains uniformly stable in the viscous case (i.e., for a space-decentered discretization) but not in the nonviscous case (for a space-centered scheme).

4. The LEP system about nonzero velocity. Linearizing system (2.1)–(2.3) about the stationary homogeneous solution $n_0 = 1$, $u_0 = a$, a being constant (but not equal to 0), and $\partial_x\phi_0 = 0$, leads to the following LEP system:

$$\partial_t n + a\partial_x n + \partial_x u = 0, \tag{4.1}$$

$$\partial_t u + a\partial_x u + c_s^2\partial_x n = \partial_x\phi, \tag{4.2}$$

$$\varepsilon^2\partial_{xx}^2\phi = n. \tag{4.3}$$

When $|a| < |c_s|$, it can be reduced to a wave-like formulation (specifically a Klein–Gordon equation with mass equal to $1/\varepsilon^2$):

$$\partial_{tt}^2 n + 2a\partial_{tx}^2 n - (c_s^2 - a^2)\partial_{xx}^2 n + \frac{1}{\varepsilon^2}n = 0. \tag{4.4}$$

We can get an energy estimate for this system by multiplying by $\partial_t n$. We have

$$\partial_t\Big((\partial_t n)^2 + (c_s^2 - a^2)(\partial_x n)^2 + \frac{1}{\varepsilon^2}n^2\Big) = 2(c_s^2 - a^2)\partial_x\Big(\partial_x n\,\partial_t n\Big) - 2a\partial_x\Big((\partial_t n)^2\Big). \tag{4.5}$$

If periodic or homogeneous Dirichlet boundary conditions are imposed at the boundary points, conservation of energy follows:

$$\partial_t \int_a^b \Big((\partial_t n)^2 + (c_s^2 - a^2)(\partial_x n)^2 + \frac{1}{\varepsilon^2}n^2\Big)\,dx = 0. \tag{4.6}$$

We consider a space discretization of (4.4) by replacing the continuous second order space derivative by the discrete operator D^2, with the same meaning as above for $\tilde{D}$:

$$\partial_{tt}^2 n + 2a\partial_t \tilde{D} n - (c_s^2 - a^2) D^2 n + \frac{1}{\varepsilon^2} n = 0. \tag{4.7}$$

Applying the II time-difference scheme (3.17) to this equation leads to

$$\frac{n^{m+1} - 2n^m + n^{m-1}}{\delta^2} + 2a \frac{\tilde{D} n^m - \tilde{D} n^{m-1}}{\delta} + a^2 D^2 n^{m-1} - c_s^2 D^2 n^m + \frac{1}{\varepsilon^2} n^{m+1} = 0. \tag{4.8}$$

This scheme can be viewed as a wave-type version of the following scheme for the original LEP system (4.1)–(4.3):

$$\frac{n^{m+1} - n^m}{\delta} + a D_- n^m + D_- u^{m+1} = 0, \tag{4.9}$$

$$\frac{u^{m+1} - u^m}{\delta} + a D_+ u^m + c_s^2 D_+ n^m = D_+ \phi^{m+1}, \tag{4.10}$$

$$\varepsilon^2 D^2 \phi^{m+1} = n^{m+1}. \tag{4.11}$$

Indeed, eliminating u between (4.9) and (4.10), one has

$$\frac{n^{m+1} - 2n^m + n^{m-1}}{\delta^2} + 2a \frac{\tilde{D} n^m - \tilde{D} n^{m-1}}{\delta} + a^2 D^2 n^{m-1} - c_s^2 D^2 n^m + D^2 \phi^{m+1} = 0, \tag{4.12}$$

and using (4.11), we get (4.8). We can also eliminate n^{m+1} from (4.12) using (4.11) and get a discretization of the reformulated Poisson equation (2.7) (in linearized form):

$$\left(\frac{\varepsilon^2}{\delta^2} + 1 \right) D^2 \phi^{m+1} = \frac{2n^m - n^{m-1}}{\delta^2} + c_s^2 D^2 n^m - 2a \frac{\tilde{D} n^m - \tilde{D} n^{m-1}}{\delta} - a^2 D^2 n^{m-1}. \tag{4.13}$$

We also note that the implicit treatment of the last term in (4.8) gives a strong dissipation (like in the Fourier analysis case). This can be seen from the use of the so-called equivalent equation (i.e., the equation obtained by retaining the leading order term in the Taylor expansion of the time semidiscrete scheme). This equivalent equation is written

$$\partial_{tt}^2 n + 2a\partial_{tx}^2 n - (c_s^2 - a^2)\partial_{xx}^2 n + \frac{1}{\varepsilon^2} n = -\frac{\delta}{\varepsilon^2} \partial_t n + a^2 \delta \partial_{txx}^3 n + a\delta \partial_{ttx}^3 n. \tag{4.14}$$

The first two terms on the right-hand side are dissipative terms, while the nature of the last term is less clear.

By examining the LEP system about zero velocity, we have shown that schemes that are implicit in both the mass flux term and the electric field source term are asymptotically stable in the small Debye length limit $\varepsilon \to 0$.

Two schemes have been shown to have this property. A first scheme is associated with centered space-differencing: the II scheme for the LEP system without diffusion. A second scheme is associated with a decentered shock-capturing–type space discretization: the II scheme for the viscous linearized Euler–Poisson (LVEP) system (see section 3.3). We have seen in section 3.2 that the II scheme for the inviscid LEP system is not uniformly stable with respect to ε when $\varepsilon \to \infty$ and that this lack of uniformity is related to the space-centered differencing.

We are now concerned with the stability of these two schemes for the LEP problem about nonzero velocity. This analysis will confirm that the II scheme without diffusion is uniformly stable in the limit $\varepsilon \to 0$ but not in the limit $\varepsilon \to \infty$ while the II scheme with diffusion is uniformly stable for all values of ε, however small or large.

We linearize system (2.1)–(2.3) about the stationary homogeneous solution $n_0 = 1$, $q_0 =$ Constant (but not equal to 0), and $\partial_x \phi_0 = 0$. We denote by $\zeta \ll 1$ the size of the initial perturbation to this stationary state. Then, we can formally expand the solution as follows: $n = 1+\zeta n' + O(\zeta^2)$, $q = q_0 + \zeta(q_0 n' + u') + O(\zeta^2)$, $\phi = \zeta\phi' + O(\zeta^2)$. Retaining only terms of order ζ, we are led to the following LEP system (where the primes have been dropped for the sake of clarity):

$$\partial_t n + \partial_x u = 0, \tag{4.15}$$

$$\partial_t u + \partial_x(2q_0 u - q_0^2 n) + c_s^2 \partial_x n = \partial_x \phi, \tag{4.16}$$

$$\varepsilon^2 \partial_{xx}\phi = n. \tag{4.17}$$

After a Fourier transform in space, the system reads

$$\partial_t \hat{n} + i\xi \hat{u} = 0, \tag{4.18}$$

$$\partial_t \hat{u} + i\xi(2q_0 u - q_0^2 n) + i\xi c_s^2 \hat{n} = i\xi\hat{\phi}, \tag{4.19}$$

$$-\varepsilon^2 \xi^2 \hat{\phi} = \hat{n}. \tag{4.20}$$

After elimination of $\hat{\phi}$, the system reduces to a 2×2 system of the form

$$\begin{pmatrix} \partial_t \hat{n} \\ \partial_t \hat{u} \end{pmatrix} + \begin{pmatrix} 0 & i\xi \\ -iq_0^2\xi + i\xi c_s^2 + \frac{i}{\varepsilon^2 \xi} & 2iq_0\xi \end{pmatrix} \begin{pmatrix} \hat{n} \\ \hat{u} \end{pmatrix} = 0. \tag{4.21}$$

Its solutions are of the form

$$\begin{pmatrix} \hat{n}(\xi,t) \\ \hat{u}(\xi,t) \end{pmatrix} = \begin{pmatrix} \hat{n}_+(\xi) \\ \hat{u}_+(\xi) \end{pmatrix} e^{i\theta_+^\varepsilon(\xi)t} + \begin{pmatrix} \hat{n}_-(\xi) \\ \hat{u}_-(\xi) \end{pmatrix} e^{i\theta_-^\varepsilon(\xi)t}, \tag{4.22}$$

where

$$\theta_\pm^\varepsilon(\xi) = -q_0\xi \pm \left(c_s^2\xi^2 + \frac{1}{\varepsilon^2}\right)^{1/2}. \tag{4.23}$$

These eigenmodes are obtained by shifting the phase by the quantity $-q_0\xi$. Indeed, the linearization about a uniform nonzero velocity can be deduced from a linearization about a zero velocity by a Galilean transformation $x' = x + q_0 t$, $v' = v$ thanks to the Galilean invariance of the nonlinear Euler problem. A Galilean shift translates in Fourier space into a constant phase shift.

It should not be inferred, however, that the stability of the schemes for the linearization about a nonzero velocity can be deduced from considerations concerning the linearized system about a zero velocity. Indeed, the mass and momentum flux terms are treated differently (one is implicit while the other one is explicit), which breaks this Galilean invariance.

Now, the II scheme for the inviscid Euler–Poisson system linearized about nonzero velocity reads

$$\frac{\hat{n}^{m+1}-\hat{n}^m}{\delta}+i\xi\hat{u}^{m+1}=0, \tag{4.24}$$

$$\frac{\hat{u}^{m+1}-\hat{u}^m}{\delta}+i\xi(2q_0u^m-q_0^2n^m)+i\xi c_s^2\hat{n}^m+\frac{i}{\varepsilon^2\xi}\hat{n}^{m+1}=0, \tag{4.25}$$

while the II scheme for the viscous system is written

$$\frac{\hat{n}^{m+1}-\hat{n}^m}{\delta}+i\xi\hat{u}^{m+1}+\beta\xi^2\hat{n}^m=0, \tag{4.26}$$

$$\frac{\hat{u}^{m+1}-\hat{u}^m}{\delta}+i\xi(2q_0u^m-q_0^2n^m)+i\xi c_s^2\hat{n}^m+\beta\xi^2\hat{u}^m+\frac{i}{\varepsilon^2\xi}\hat{n}^{m+1}=0. \tag{4.27}$$

We recall that the inviscid linearized system in Fourier space mimics a centered spatial discretization while the viscous system mimics a decentered spatial discretization associated with a numerical diffusion equal to β.

We can prove that the II scheme for the inviscid system is uniformly stable in the limit $\varepsilon \to 0$. But ε must be large enough. To state a more precise result, we recall that the range of admissible wave-numbers ξ is limited to the interval $|\xi| \le \xi^* = \frac{\pi}{\Delta x}$ in order to mimic the influence of a space discretization Δx. Then, for all $K < 1$, there exists a $\delta^* > 0$ such that for all $\delta < \delta^*$ and all ε such that $\varepsilon < K\delta$, the II scheme for the inviscid system is stable. We note that as K gets closer to 1, δ^* tends to 0. In particular, the scheme is not stable when $\varepsilon = O(1)$ and $\delta \to 0$, which would be the situation if we wanted to resolve the plasma period scale. We can show that if $\varepsilon = 0$, the scheme is unstable. Therefore, the II scheme for the inviscid system cannot be of practical usefulness.

We now turn to the II scheme for the viscous system. We assume that the diffusion constant β is proportional to the time step: $\beta = B\delta$. Then, the result we can prove is that if B is large enough (depending only on q_0 and c_s), then there exists an interval of time steps $[0, \delta^*]$ (again, depending only on q_0 and c_s) such that for all δ's in this interval, all admissible wave-numbers ξ (i.e., such that $|\xi| \le \xi^*$) and all $\varepsilon \ge 0$, the scheme is stable. Usually, the diffusion constant is taken proportional to Δx (or to $\Delta x^2/\delta$ in the case of the Lax–Friedrichs scheme) rather than to δ. The result would also remain true in this case, but the proof would be more complex. Indeed, supposing that $\beta = O(\delta)$ allows us to express the problem in terms of the CFL variable, $X = \xi\delta$ only, which simplifies the (already quite technical) proof. Also, the result in itself is interesting as it shows that the scheme is stable even with a vanishingly small diffusion proportional to δ.

These results show that, at least for the linearized problem about any uniform state (of zero or nonzero velocity), the II scheme associated with a decentered space-differencing is uniformly stable with respect to all values of ε small or large.

5. Burgers–Poisson equation. In a preliminary attempt to analyze the nonlinear effects, we consider the following model consisting of a nonlinear Burgers–Poisson system:

$$\partial_t u + u\partial_x u = -\partial_x \phi, \tag{5.1}$$

$$\varepsilon^2 \partial_{xx}^2 \phi = u. \tag{5.2}$$

This equation appears as a model describing high-frequency wave propagation in relaxing media [30, 17]. In [30], special soliton solutions called "loop solitons" are constructed using inverse scattering theory. However, these solutions are multivalued and cannot be computed by the method developed below. Note that the quasineutral limit has been investigated theoretically in [17].

This system owns the entropy pair $(\frac{u^2}{2}, \frac{u^3}{3} + \frac{\varepsilon^2 \phi_x^2}{2})$. Indeed, simple computations show that

$$\partial_t \left(\frac{u^2}{2} \right) + \partial_x \left(\frac{u^3}{3} + \frac{\varepsilon^2 (\partial_x \phi)^2}{2} \right) = 0. \tag{5.3}$$

The theoretical study of this equation will be reported in a future paper.

We consider the conservative scheme

$$\partial_t u_j + \frac{u_{j+1} + u_j + u_{j-1}}{3} \frac{u_{j+1} - u_{j-1}}{2h} = -\frac{\phi_{j+1} - \phi_{j-1}}{2h}, \tag{5.4}$$

$$\varepsilon^2 \frac{\phi_{j+1} - 2\phi_j + \phi_{j-1}}{h^2} = u_j, \tag{5.5}$$

where $h = \Delta x$ is the grid size. The above scheme may generate some oscillations near shocks due to the dispersive nature of the scheme. We can recast the above scheme in a conservative form,

$$\partial_t u_j + \frac{f_{j+1/2} - f_{j-1/2}}{h} = -\frac{\phi_{j+1} - \phi_{j-1}}{2h}, \tag{5.6}$$

$$\varepsilon^2 \frac{\phi_{j+1} - 2\phi_j + \phi_{j-1}}{h^2} = u_j, \tag{5.7}$$

where $f_{j+1/2} = \frac{1}{6}(u_j^2 + u_j u_{j+1} + u_{j+1}^2)$. We have a discrete entropy pair

$$\partial_t \left(\frac{u_j^2}{2} \right) + \frac{g_{j+1/2} - g_{j-1/2}}{h} = 0, \tag{5.8}$$

where $g_{j+1/2} = \frac{1}{6}(u_j^2 u_{j+1} + u_j u_{j+1}^2) + \frac{\varepsilon^2}{2}(\frac{\phi_{j+1} - \phi_j}{h})^2$. As a direct consequence, we can prove a stability result for the space semidiscretization of this model. We can also get some convergence results assuming the solution is smooth.

More precisely, we consider only the periodic boundary condition case for simplicity. Let $\phi(x), u(x)$ be the exact solution. Let $\Phi_j = \phi(x_j)$, $U_j = \varepsilon^2 \frac{\Phi_{j+1} - 2\Phi_j + \Phi_{j-1}}{h^2}$. Assuming that the initial data are approximated up to second order, we have

$$\|u_h - u\| = \sqrt{\sum_j (u_j - u(x_j))^2 h} \le Ch^2$$

and

$$\varepsilon\|\phi_h - \phi\| = \varepsilon\sqrt{\sum_j (\phi_j - \phi(x_j))^2 h} \le Ch^2$$

for all $\varepsilon \le \varepsilon_0$, with arbitrary $\varepsilon_0 > 0$ and where $C = C(u, \phi, \varepsilon_0)$.

This analysis indicates that a similar implicit time discretization as in the LVEP system will provide asymptotic stability with respect to $\varepsilon \to 0$. The investigation of this problem will be the topic of future work.

6. Conclusion. In this paper, we have analyzed the stability property of a recently proposed numerical scheme for the Euler–Poisson system [9]. For this analysis, we have considered the linearized Euler–Poisson system about both zero and nonzero velocities, and, after a Fourier transform in space, we have applied the time discretization under scrutiny. We have mimicked the influence of a space-decentered discretization by adding viscosity terms in both the mass and momentum conservation equations. We have shown that the resulting scheme is uniformly stable with respect to the Debye parameter ε. By contrast, other schemes with less implicit treatment or with centered space-differencing are shown to lose the uniform stability property. Elements of a nonlinear stability analysis have been given by considering a model Burgers–Poisson problem. The analysis of this strategy for the kinetic version of this model [10] is in progress.

REFERENCES

[1] F. F. Chen, *Introduction to Plasma Physics and Controlled Fusion*, 2nd ed., Plenum, New York, 1984.

[2] H.-H. Choe, N. S. Yoon, S. S. Kim, and D.-I. Choi, *A new unconditionally stable algorithm for steady-state fluid simulation of high density plasma discharge*, J. Comput. Phys., 170 (2001), pp. 550–561.

[3] B. I. Cohen, A. B. Langdon, and A. Friedman, *Implicit time integration for plasma simulation*, J. Comput. Phys., 46 (1982), pp. 15–53.

[4] Ph. Colella, M. R. Dorr, and D. D. Wake, *A conservative finite difference method for the numerical solution of plasma fluid equations*, J. Comput. Phys., 149 (1999), pp. 168–193.

[5] M. G. Crandall and A. Majda, *Monotone difference approximations for scalar conservation laws*, Math. Comp., 34 (1980), pp. 1–21.

[6] P. Crispel, P. Degond, C. Parzani, and M.-H. Vignal, *Trois formulations d'un modèle de plasma quasi-neutre avec courant non-nul*, C. R. Acad. Sci. Paris Sér. I Math., 338 (2004), pp. 327–332.

[7] P. Crispel, P. Degond, and M.-H. Vignal, *An asymptotically stable discretization for the Euler-Poisson system in the quasineutral limit*, C. R. Acad. Sci. Paris Sér. I Math., 341 (2005), pp. 323–328.

[8] P. Crispel, P. Degond, and M.-H. Vignal, *Quasi-neutral fluid models for current carrying plasmas*, J. Comput. Phys., 205 (2005), pp. 408–438.

[9] P. Crispel, P. Degond, and M.-H. Vignal, *An asymptotically preserving scheme for the two-fluid Euler-Poisson model in the quasi-neutral limit*, J. Comput. Phys., 223 (2007), pp. 208–234.

[10] P. Degond, F. Deluzet, and L. Navoret, *An asymptotically stable Particle-in-Cell (PIC) scheme for collisionless plasma simulations near quasineutrality*, C. R. Acad. Sci. Paris Sér. I Math., 343 (2006), pp. 613–618.

[11] P. Degond, C. Parzani, and M.-H. Vignal, *Un modèle d'expansion de plasma dans le vide*, C. R. Acad. Sci. Paris Sér. I Math., 335 (2002), pp. 399–404.

[12] P. Degond, C. Parzani, and M.-H. Vignal, *A one-dimensional model of plasma expansion*, Math. Comput. Modelling, 38 (2003), pp. 1093–1099.

[13] P. Degond, C. Parzani, and M.-H. Vignal, *Plasma expansion in vacuum: Modeling the breakdown of quasineutrality*, Multiscale Model. Simul., 2 (2003), pp. 158–178.

[14] P. Degond, C. Parzani, and M.-H. Vignal, *On plasma expansion in vacuum*, in Free Boundary Problems: Theory and Applications, P. Colli, C. Verdi, and A. Visintin, eds., Internat. Ser. Numer. Math. 147, Birkhäuser Verlag, Basel, 2004, pp. 103–112.

[15] G. DiPeso, D. W. Hewett, and G. F. Simonson, *Extension of the streamlined Darwin model to quasineutral plasmas*, J. Comput. Phys., 111 (1994), pp. 237–247.

[16] S. Fabre, *Stability analysis of the Euler-Poisson equations*, J. Comput. Phys., 101 (1992), pp. 445–451.

[17] K. Fellner and C. Schmeiser, *Burgers–Poisson: A nonlinear dispersive model equation*, SIAM J. Appl. Math., 64 (2004), pp. 1509–1525.

[18] R. N. Franklin and J. R. Ockendon, *Asymptotic matching of plasma and sheath in an active low pressure discharge*, J. Plasma Physics, 4 (1970), pp. 371–385.

[19] A. Friedman, *Implicit multiscale PIC and related topics*, in Workshop on Multiscale Processes in Fusion Plasmas, Institute for Pure and Applied Mathematics, UCLA, Los Angeles, CA, 2005, http://www.ipam.ucla.edu/publications/fus2005/fus2005_5425.pdf.

[20] S. Y. Ha and M. Slemrod, *Global existence of plasma ion sheaths and their dynamics*, Comm. Math. Phys., 238 (2003), pp. 149–186.

[21] N. A. Krall and A. W. Trivelpiece, *Principles of Plasma Physics*, San Francisco Press, San Fransico, CA, 1986.

[22] N. N. Kuznecov and S. A. Volosin, *On monotone difference approximations for a first-order quasi-linear equation*, Soviet Math. Dokl., 17 (1976), pp. 1203–1206.

[23] A. B. Langdon, B. I. Cohen, and A. Friedman, *Direct implicit large time-step particle simulation of plasmas*, J. Comput. Phys., 51 (1983), pp. 107–138.

[24] G. Lapenta, F. Iinoya, and J. U. Brackbill, *Particle-in-cell simulations of glow discharges in complex geometries*, IEEE Trans. Plasma Sci., 23 (1995), pp. 769–779.

[25] P. A. Markowich, C. Ringhofer, and C. Schmeiser, *Semiconductor Equations*, Springer, Wien, 1990.

[26] R. J. Mason, *Implicit moment particle simulation of plasmas*, J. Comput. Phys., 41 (1981), pp. 233–244.

[27] R. J. Mason, *Implicit moment PIC-hybrid simulation of collisional plasmas*, J. Comput. Phys., 51 (1983), pp. 484–501.

[28] R. Schneider and C.-D. Munz, *The approximation of two-fluid plasma flow with explicit upwind schemes*, Internat. J. Numer. Model., 8 (2005), pp. 399–416.

[29] U. Shumlak and J. Loverich, *Approximate Riemann solver for the two-fluid plasma model*, J. Comput. Phys., 187 (2003), pp. 620–638.

[30] V. Vakhnenko and J. Parkes, *A novel nonlinear evolution equation integrable by the inverse scattering method*, Proceedings of the Institute of Mathematics of NAS of Ukraine, 43 (2002), pp. 384–391.

[31] P. L. G. Ventzek, T. J. Sommerer, R. J. Hoekstra, and M. J. Kushner, *Two-dimensional hybrid model of inductively coupled plasma sources for etching*, Appl. Phys. Lett., 63 (1993), pp. 605–607.

[32] P. L. G. Ventzek, R. J. Hoekstra, and M. J. Kushner, *Two-dimensional modeling of high plasma density inductively coupled sources for materials processing*, J. Vac. Sci. Technol. B, 12 (1994), pp. 461–477.

SIAM J. NUMER. ANAL.
Vol. 46, No. 3, pp. 1323–1337
© 2008 Society for Industrial and Applied Mathematics

THE FABER–MANTEUFFEL THEOREM FOR LINEAR OPERATORS*

V. FABER†, J. LIESEN‡, AND P. TICHÝ§

Abstract. A short recurrence for orthogonalizing Krylov subspace bases for a matrix A exists if and only if the adjoint of A is a low-degree polynomial in A (i.e., A is normal of low degree). In the area of iterative methods, this result is known as the Faber–Manteuffel theorem [V. Faber and T. Manteuffel, *SIAM J. Numer. Anal.*, 21 (1984), pp. 352–362]. Motivated by the description by J. Liesen and Z. Strakoš, we formulate here this theorem in terms of linear operators on finite dimensional Hilbert spaces and give two new proofs of the necessity part. We have chosen the linear operator rather than the matrix formulation because we found that a matrix-free proof is less technical. Of course, the linear operator result contains the Faber–Manteuffel theorem for matrices.

Key words. cyclic subspaces, Krylov subspaces, orthogonal bases, orthogonalization, short recurrences, normal matrices

AMS subject classifications. 65F10, 65F25

DOI. 10.1137/060678087

1. Introduction.

At the Householder Symposium VIII held in Oxford in July 1981, Golub posed as an open question to characterize necessary and sufficient conditions on a matrix A for the existence of a three-term conjugate gradient–type method for solving linear systems with A (cf. *SIGNUM Newsletter*, vol. 16, no. 4, 1981). This important question was answered by Faber and Manteuffel in 1984 [4]. They showed that an $(s+2)$-term conjugate gradient type method for A, based on some given inner product, exists if and only if the adjoint of A with respect to the inner product is a polynomial of degree s in A (i.e., A is normal of degree s). In the area of iterative methods this result is known as the Faber–Manteuffel theorem; see, e.g., [7, Chapter 6] or [13, Chapter 6.10].

The theory of [4] and some further developments have recently been surveyed in [12]. There the Faber–Manteuffel theorem is formulated independently of the conjugate gradient context and solely as a result on the existence of a short recurrence for generating orthogonal bases for Krylov subspaces of the matrix A. A new proof of the sufficiency part is given, and the normality condition on A is thoroughly characterized. For the proof of the (significantly more difficult) necessity part, however, the authors refer to [4]. In particular, they suggest that, in light of the fundamental nature of the result, it is desirable to find an alternative, and possibly simpler, proof. Note that a proof similar to the one of Faber and Manteuffel but for other classes of matrices has been given in [14].

Motivated by the description in [12], we here take a new approach to formulate and prove the necessity part of the Faber–Manteuffel theorem. Instead of a matrix

*Received by the editors December 20, 2006; accepted for publication (in revised form) September 10, 2007; published electronically March 26, 2008.
http://www.siam.org/journals/sinum/46-3/67808.html

†BD Biosciences-Bioimaging Systems (vance_faber@bd.com).

‡Institute of Mathematics, Technical University of Berlin, Straße des 17. Juni 136, 10623 Berlin, Germany (liesen@math.tu-berlin.de). The work of this author was supported by the Emmy Noether-Programm of the Deutsche Forschungsgemeinschaft.

§Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod vodárenskou věží 2, 18207 Prague, Czech Republic (tichy@cs.cas.cz). The work of this author was supported by the Emmy Noether-Programm of the Deutsche Forschungsgemeinschaft and by the Czech National Program of Research Information Society under project 1ET400300415.

we consider a given linear operator A on a finite dimensional Hilbert space V. By the cyclic decomposition theorem, the space V decomposes into cyclic invariant subspaces, i.e., Krylov subspaces, of A (see section 2 for details). The Faber–Manteuffel theorem then gives a necessary (and sufficient) condition on A, so that the standard Gram–Schmidt algorithm for generating orthogonal bases of the cyclic subspaces reduces from a full to a short recurrence.

We have chosen this setting because we believe that the proof of necessity is easier to follow when we use linear operators rather than matrices. In this paper we give two different proofs of the necessity part, both based on restriction of the linear operator A to certain cyclic invariant subspaces. The resulting technicalities in the matrix formulation would obstruct rather than help the understanding. Moreover, our formulation may serve as a starting point for extending the results to infinite dimensional spaces. We are not aware that any such extensions have been obtained yet.

The paper is organized as follows. In section 2 we introduce the notation and the required background from the theory of linear operators. In section 3 we translate the matrix concepts introduced in [12] into the language of linear operators. In section 4 we state and prove several technical lemmas that are required in the proof of the main result, which is given in section 5. In section 6 we give an alternative proof, which we consider elementary and constructive. This proof involves structure-preserving orthogonal transformations of Hessenberg matrices, which may be of interest beyond our context here. In section 7 we discuss our rather theoretical analysis in the preceeding sections. This discussion includes a matrix formulation of the Faber–Manteuffel theorem, a high-level description of the strategies of our two proofs of the necessity part, and our reasoning why necessity is more difficult to prove than sufficiency. For obtaining a more detailed overview of the results in this paper, section 7 may be read before the other sections.

2. Notation and background. In this section we introduce the notation and recall some basic results from the theory of linear operators; see Gantmacher's book [6, Chapters VII and IX] for more details.

Let V be a finite dimensional Hilbert space, i.e., a complex vector space equipped with a (fixed) inner product $(\cdot,\cdot)$. Let $A : V \to V$ be a given invertible linear operator. For any vector $v \in V$, we can form the sequence

$$v, Av, A^2v, \ldots . \tag{2.1}$$

Since V is finite dimensional, there exists an integer $d = d(A, v)$ such that the vectors $v, Av, \ldots, A^{d-1}v$ are linearly independent, while $A^d v$ is a linear combination of them. This means that there exist scalars, $\alpha_1, \ldots, \alpha_{d-1}$, not all equal to zero, such that

$$A^d v \;=\; -\sum_{j=0}^{d-1} \alpha_j A^j v\,. \tag{2.2}$$

Defining the monic polynomial $\phi(z) = z^d + \alpha_{d-1} z^{d-1} + \cdots + \alpha_0$, we can rewrite (2.2) as

$$\phi(A)v \;=\; 0\,. \tag{2.3}$$

We say that ϕ *annihilates* v. It would be more accurate to say "ϕ annihilates v with respect to A," but when it is clear which operator A is meant, the reference to A is

omitted for the sake of brevity. The monic polynomial ϕ is the uniquely determined monic polynomial of smallest degree that annihilates v, and it is called the *minimal polynomial of* v. Its degree, equal to $d(A, v)$, is called the *grade of* v, and v is said to be of grade $d(A, v)$.

Consider any basis of V, and define the polynomial Φ as the least common multiple of the minimal polynomials of the basis vectors. Then Φ is the uniquely defined (independent of the choice of the basis!) monic polynomial of smallest degree that annihilates all vectors $v \in V$, and it is called the *minimal polynomial of* A. We denote its degree by $d_{\min}(A)$. Apparently, $d_{\min}(A) \geq d(A, v)$ for all $v \in V$, and Φ is divisible by the minimal polynomial of every vector $v \in V$.

If $v \in V$ is any vector of grade d, then

$$\text{span}\{v, \dots\ A^{d-1}v\} \ \equiv\ \mathcal{K}_d(A, v) \tag{2.4}$$

is a d-dimensional invariant subspace of A. Because of (2.2) and the special character of the basis vectors, the subspace $\mathcal{K}_d(A, v)$ is called *cyclic*. The letter $\mathcal{K}$ has been chosen because this space is often called the *Krylov subspace* of A and v. The vector v is called the *generator* of this subspace.

A central result in the theory of linear operators on finite dimensional vector spaces is that the space V can be decomposed into cyclic subspaces. This result has several equivalent formulations, and in this paper we will use the one from [6, Chapter VII, section 4, Theorem 3]: there exist vectors $w_1, \dots, w_j \in V$ of respective grades $d_1, \dots, d_j$ such that

$$V \ =\ \mathcal{K}_{d_1}(A, w_1) \oplus \cdots \oplus \mathcal{K}_{d_j}(A, w_j)\,, \tag{2.5}$$

where the minimal polynomial of w_1 is equal to the minimal polynomial of A, and for $k = 1, \dots, j-1$, the minimal polynomial of w_k is divisible by the minimal polynomial of w_{k+1}.

Since the decomposition (2.5) is an important tool in this paper, we illustrate it by a simple example (adapted from [9, section 7.2]; also see [10] for a short and self-contained proof of the decomposition (2.5)). Suppose that A is the linear operator on $V = \mathbb{R}^3$ whose matrix representation in the canonical basis of $\mathbb{R}^3$ is

$$\begin{bmatrix} 2 & -3 & -3 \\ -3 & 2 & 3 \\ 3 & -3 & -4 \end{bmatrix}.$$

The characteristic polynomial of A is $(z-2)(z+1)^2$, while the minimal polynomial is $\Phi = (z-2)(z+1)$, so that $d_{\min}(A) = 2$. Any nonzero vector in $\mathbb{R}^3$ is either of grade one (and hence is an eigenvector) or of grade two. It is easy to see that the first canonical basis vector is not an eigenvector. Thus, $w_1 \equiv [1,\ 0,\ 0]^T$ is of grade $d_1 = 2$, i.e., $\mathcal{K}_{d_1}(A, w_1)$ has dimension two, and the minimal polynomial of w_1 is Φ. Note that

$$\mathcal{K}_{d_1}(A, w_1) \ =\ \text{span}\left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 2 \\ -3 \\ 3 \end{bmatrix} \right\} \ =\ \left\{ \begin{bmatrix} \alpha \\ \beta \\ -\beta \end{bmatrix} \,:\, \alpha, \beta \in \mathbb{R} \right\}.$$

Since $V = \mathbb{R}^3$ has dimension three, it remains to find a vector $w_2 \notin \mathcal{K}_{d_1}(A, w_1)$ that is of grade one and has minimal polynomial $z+1$, i.e., w_2 is an eigenvector with respect to the eigenvalue -1, that is not contained in $\mathcal{K}_{d_1}(A, w_1)$. These requirements are satisfied by $w_2 \equiv [1,\ 0,\ 1]^T$, giving

$$\mathbb{R}^3 \ =\ \mathcal{K}_{d_1}(A, w_1) \oplus \mathcal{K}_{d_2}(A, w_2) \ =\ \text{span}\,\{w_1,\ Aw_1\} \oplus \text{span}\,\{w_2\}\,.$$

In the basis of $\mathbb{R}^3$ given by w_1, Aw_1, w_2, the linear operator A has the matrix representation

$$\left[\begin{array}{cc|c} 0 & 2 & \\ 1 & 1 & \\ \hline & & -1 \end{array}\right].$$

Here the two diagonal blocks correspond to the decomposition (2.5), where each block is the companion matrix of the minimal polynomial of the respective cyclic subspace generators. This matrix representation is sometimes called the *rational canonical form*. When this canonical form consists of a single diagonal block in companion form, A is called *nonderogatory*. Hence in our example A is derogatory, but the restriction of A to the cyclic subspace generated by w_1 is nonderogatory. Loosely speaking, this restriction is the "largest nonderogatory part" of A.

3. Orthogonalization of a cyclic subspace basis. Let $v \in V$ be a vector of grade d. For theoretical as well as practical purposes it is often convenient to *orthogonalize* the basis $v, \ldots, A^{d-1}v$ of the cyclic subspace $\mathcal{K}_d(A, v)$. The classical approach to orthogonalization, which appears in different mathematical areas (see, e.g., [2, p. 15], [5, p. 74]) is the Gram–Schmidt algorithm:

$$v_1 = v\,, \tag{3.1}$$

$$v_{n+1} = Av_n - \sum_{m=1}^{n} h_{m,n} v_m\,, \tag{3.2}$$

$$h_{m,n} = \frac{(Av_n, v_m)}{(v_m, v_m)}\,, \quad m = 1, \ldots, n\,, \quad n = 1, \ldots, d-1\,. \tag{3.3}$$

The resulting vectors $v_1, \ldots, v_d$ are mutually orthogonal, and for $n = 1, \ldots, d$ they satisfy $\operatorname{span}\{v_1, \ldots, v_n\} = \operatorname{span}\{v, \ldots, A^{n-1}v\}$. We call v (or v_1) the *initial vector* of the algorithm (3.1)–(3.3). When A is a (square) matrix, this algorithm is usually referred to as Arnoldi's method [1]. It can be equivalently written as

$$v_1 = v\,, \tag{3.4}$$

$$A \underbrace{[v_1, \ldots, v_{d-1}]}_{\equiv V_{d-1}} = \underbrace{[v_1, \ldots, v_d]}_{\equiv V_d} \underbrace{\begin{bmatrix} h_{1,1} & \cdots & h_{1,d-1} \\ 1 & \ddots & \vdots \\ & \ddots & h_{d-1,d-1} \\ & & 1 \end{bmatrix}}_{\equiv H_{d,d-1}}, \tag{3.5}$$

$$(v_i, v_j) = 0 \quad \text{for } i \neq j\,, \quad i, j = 1, \ldots, d\,. \tag{3.6}$$

The matrix $H_{d,d-1}$ is an unreduced upper Hessenberg matrix of size $d \times (d-1)$. Its band structure determines the length of the recurrence (3.2) that generates the orthogonal basis. To state this formally, we need the following definition [12, Definition 2.1].

DEFINITION 3.1. *An unreduced upper Hessenberg matrix is called $(s+2)$-band Hessenberg when its sth superdiagonal contains at least one nonzero entry and all its entries above its sth superdiagonal are zero.*

If $H_{d,d-1}$ is $(s+2)$-band Hessenberg, then for $n = 1, \ldots, d-1$, the recurrence (3.2) reduces to

$$v_{n+1} \;=\; Av_n \;-\; \sum_{m=n-s}^{n} h_{m,n} v_m \,, \tag{3.7}$$

and thus the orthogonal basis is generated by an $(s+2)$-*term recurrence*. Since precisely the latest $s+1$ basis vectors $v_n, \ldots, v_{n-s}$ are required to determine v_{n+1}, and only one operation with A is performed, an $(s+2)$-term recurrence of the form (3.7) is called *optimal*.

DEFINITION 3.2 (linear operator version of [12, Definition 2.4]). *Let A be an invertible linear operator with minimal polynomial degree $d_{\min}(A)$ on a finite dimensional Hilbert space, and let s be a nonnegative integer, $s+2 \le d_{\min}(A)$.*

1. *If for an initial vector the matrix $H_{d,d-1}$ in (3.4)–(3.6) is $(s+2)$-band Hessenberg, then we say that A admits for the given initial vector an optimal $(s+2)$-term recurrence.*
2. *If A admits for any given initial vector an optimal recurrence of length at most $s+2$, while it admits for at least one given initial vector an optimal $(s+2)$-term recurrence, then we say that A admits an optimal $(s+2)$-term recurrence.*

We denote the *adjoint of* A by A^*. This is the uniquely determined operator that satisfies $(Av, w) = (v, A^*w)$ for all vectors v and w in the given Hilbert space. The operator A is called *normal* if it commutes with its adjoint, $AA^* = A^*A$. This holds if and only if A has a complete orthonormal system of eigenvectors. Equivalently, A^* can be written as a polynomial in A, $A^* = p(A)$, where p is completely determined by the condition that $p(\lambda_j) = \overline{\lambda}_j$ for all eigenvalues λ_j of A (cf. [6, Chapter IX, section 10]). We will be particularly interested in the degree of this polynomial.

DEFINITION 3.3. *Let A be an invertible linear operator on a finite dimensional Hilbert space. If the adjoint of A satisfies $A^* = p(A)$, where p is a polynomial of smallest degree s having this property, then A is called normal of degree s, or, shortly, normal(s).*

The condition that A is normal(s) is *sufficient* for A to admit an optimal $(s+2)$-term recurrence. The precise formulation of this statement is the following.

THEOREM 3.4. *Let A be an invertible linear operator with minimal polynomial degree $d_{\min}(A)$ on a finite dimensional Hilbert space. Let s be a nonnegative integer, $s+2 < d_{\min}(A)$. If A is normal(s), then A admits an optimal $(s+2)$-term recurrence.*

Proof. A matrix version of this result is given in [12, Theorem 2.9], and the proof given there can be easily adapted from matrices to linear operators. ∎

The main result we will prove in this paper is that the condition that A is normal(s) also is *necessary*.

THEOREM 3.5. *Let A be an invertible linear operator with minimal polynomial degree $d_{\min}(A)$ on a finite dimensional Hilbert space. Let s be a nonnegative integer, $s+2 < d_{\min}(A)$. If A admits an optimal $(s+2)$-term recurrence, then A is normal(s).*

4. Technical lemmas. We prove Theorem 3.5 in section 5. To do so, we need several technical lemmas that are stated and proved in this section.

LEMMA 4.1. *Let A be an invertible linear operator with minimal polynomial degree $d_{\min}(A)$ on a finite dimensional Hilbert space. If $1 < i < n \le d_{\min}(A)$ and $(u_1, Au_i) = 0$ for every initial vector u_1 of grade n, then $(v_1, Av_i) = 0$ for every initial vector v_1 of grade m, where $i \le m \le n$.*

(Here u_i, v_i are the ith basis vectors generated by (3.1)–(3.3) *with initial vectors u_1, v_1, respectively.)*

Proof. If $m = n$, there is nothing to prove. Hence, suppose that $m < n$, and let v_1 be a vector of grade m, and u_1 be a vector of grade n, such that the minimal polynomial of v_1 divides the minimal polynomial of u_1. Define

$$x_1 \equiv x_1(\gamma) \equiv v_1 + \gamma u_1 , \tag{4.1}$$

where γ is a scalar parameter. It is clear that, except for finitely many choices of γ, the vector x_1 is of grade n.

Suppose that γ has been chosen so that x_1 is of grade n, and consider the corresponding ith basis vector x_i, where $1 < i \le m$. By construction, $x_i = p(A)x_1$, where p is a polynomial of (exact) degree $i-1$. The vector x_i is defined uniquely (up to scaling) by the conditions

$$(A^j x_1, x_i) = (A^j x_1, p(A)x_1) = 0 , \quad j = 0, \ldots, i-2 .$$

The hypothesis

$$(x_1, Ax_i) = (x_1, Ap(A)x_1) = 0$$

gives one additional condition. We thus have i conditions that translate into i homogeneous linear equations for the i coefficients of the polynomial p. The existence of x_i implies that the determinant of the matrix $M(x_1)$ of these equations must be zero, where

$$M(x_1) = \begin{bmatrix} (x_1, x_1) & (x_1, Ax_1) & \cdots & (x_1, A^{i-1}x_1) \\ (Ax_1, x_1) & (Ax_1, Ax_1) & \cdots & (Ax_1, A^{i-1}x_1) \\ \vdots & \vdots & \vdots & \vdots \\ (A^{i-2}x_1, x_1) & (A^{i-2}x_1, Ax_1) & \cdots & (A^{i-2}x_1, A^{i-1}x_1) \\ (x_1, Ax_1) & (x_1, A^2x_1) & \cdots & (x_1, A^i x_1) \end{bmatrix} .$$

Now note that $\det M(x_1)$ is a continuous function of γ. By construction, this function is zero for all but a finite number of choices of γ. Therefore $\det M(x_1) = 0$ for all γ, and in particular, $\det M(v_1) = 0$. Consequently, there exists a nontrivial solution of the linear system with $M(v_1)$, defining a vector $w = p(A)v_1$, where p is a polynomial of degree at most $i-1$. The first $i-1$ rows mean that w is orthogonal to $v_1, \ldots, A^{i-2}v_1$, so w is a multiple of v_i. The last row means that Aw and hence Av_i is orthogonal to v_1. ∎

The decomposition (2.5) shows that for any linear operator A on a finite dimensional Hilbert space V, there exists a vector in V whose minimal polynomial coincides with the minimal polynomial of A. The following result shows that there in fact exists a basis of V consisting of vectors with this property.

LEMMA 4.2. *Let A be an invertible linear operator with minimal polynomial degree $d_{\min}(A)$ on a finite dimensional Hilbert space V. Then there exists a basis of V consisting of vectors that all are of grade $d_{\min}(A)$.*

Proof. From the cyclic decomposition theorem, cf. (2.5), we know that there exist vectors $w_1, \ldots, w_j$ with minimal polynomials $\phi_1, \ldots, \phi_j$ of respective degrees $d_1, \ldots, d_j$, such that the space V can be decomposed as

$$V = \mathcal{K}_{d_1}(A, w_1) \oplus \cdots \oplus \mathcal{K}_{d_j}(A, w_j) ,$$

where ϕ_1 equals the minimal polynomial of A, and ϕ_k is divisible by ϕ_{k+1} for $k = 1, \ldots, j-1$. In particular, $d_1 = d_{\min}(A)$. Consequently, a basis of V is given by

$$w_1, \ldots, A^{d_1-1}w_1, \quad w_2, \ldots, A^{d_2-1}w_2, \quad \ldots, \quad w_j, \ldots, A^{d_j-1}w_j\,.$$

But then it is easy to see that

$$w_1, \ldots, A^{d_1-1}w_1, \quad w_2+w_1, \ldots, A^{d_2-1}w_2+w_1, \quad \ldots, \quad w_j+w_1, \ldots, A^{d_j-1}w_j+w_1$$

is a basis of V consisting of vectors that all are of grade d_1. ∎

The following result is a generalization of [11, Lemma 4.1], which in turn can be considered a (considerably) strengthened version of [4, Lemma 2].

LEMMA 4.3. *Let A be an invertible linear operator with minimal polynomial degree $d_{\min}(A)$ on a finite dimensional Hilbert space. Let B be a linear operator on the same space, and let s be a nonnegative integer, $s+2 \le d_{\min}(A)$. If*

$$Bv \in \operatorname{span}\{v, \ldots, A^s v\} \quad \textit{for all vectors } v \textit{ of grade } d_{\min}(A)\,, \tag{4.2}$$

then $AB = BA$. In particular, if $B = A^*$, then A is normal(t) for some $t \le s$.

Proof. For notational convenience, we denote $\delta = d_{\min}(A)$. Let v be any vector of grade δ. Since A is invertible, $\mathcal{K}_\delta(A, v) = \mathcal{K}_\delta(A, Av)$. In addition, except possibly when γ is an eigenvalue of A, the vector $w = (A-\gamma I)v$ satisfies $\mathcal{K}_\delta(A, w) = \mathcal{K}_\delta(A, v)$. In the following, we exclude those values of γ. By assumption, there exist polynomials p_γ, q, and r of degree at most s, which satisfy

$$Bw = p_\gamma(A)w, \quad B(Av) = q(A)(Av), \quad Bv = r(A)v,$$

where p_γ depends on γ, but q and r do not. We can then write Bw as

$$Bw \;=\; p_\gamma(A)w \;=\; p_\gamma(A)Av - \gamma p_\gamma(A)v$$

and

$$Bw \;=\; BAv - \gamma Bv \;=\; q(A)Av - \gamma r(A)v\,.$$

Combining these two identities yields

$$t_\gamma(A)v \;=\; 0\,, \quad \text{where} \quad t_\gamma(z) = z(p_\gamma(z) - q(z)) - \gamma(p_\gamma(z) - r(z))\,.$$

The polynomial t_γ is of degree at most $s+1 < s+2 \le \delta$. Thus, except for finitely many γ, $t_\gamma = 0$. Some straightforward algebraic manipulation gives, for all but these γ,

$$\gamma(q(z) - r(z)) \;=\; (z-\gamma)\widehat{p}_\gamma(z)\,,$$

where $\widehat{p}_\gamma \equiv p_\gamma - q$ is of degree at most $s-1$. Therefore, every γ that is not an eigenvalue of A is a root of the polynomial $r-q$, which consequently must be identically zero.

But then

$$B(Av) \;=\; q(A)(Av) \;=\; Aq(A)v \;=\; Ar(A)v \;=\; A(Bv)\,.$$

By Lemma 4.2, there exists a basis consisting of vectors of grade δ. Hence $BAv = ABv$ for a basis of vectors v, so that $BA = AB$.

Finally, if $B = A^*$, then $AA^* = A^*A$, so that A is normal and hence $A^* = p(A)$ for some polynomial. From (4.2) we see that the degree of p is at most s, so that A is normal(t) for some $t \le s$. ∎

5. Proof of Theorem 3.5. Let A be an invertible linear operator on a finite dimensional Hilbert space, and let s be a nonnegative integer, $s+2 < d_{\min}(A)$. Suppose that A admits an optimal $(s+2)$-term recurrence.

Step 1. *Restriction to a cyclic subspace of dimension* $s+2$.

If u_1 is any vector of grade $s+3$, then (with the obvious meaning of u_{s+2})

$$0 \;=\; h_{1,s+2} \;=\; (u_1, Au_{s+2})\,. \tag{5.1}$$

Consider any v_1 of grade $s+2$, and the corresponding cyclic subspace $\mathcal{K}_{s+2}(A, v_1)$. Let $\widehat{A}$ be the restriction of A to $\mathcal{K}_{s+2}(A, v_1)$, i.e., the invertible linear operator

$$\widehat{A} : \mathcal{K}_{s+2}(A, v_1) \to \mathcal{K}_{s+2}(A, v_1)\,, \qquad v \mapsto Av \;\text{ for } v \in \mathcal{K}_{s+2}(A, v_1)\,.$$

Clearly, $d_{\min}(\widehat{A}) = s+2$. Let $\mathcal{K}_{s+2}(A, v_1)$ be equipped with the same inner product as the whole space.

Let $y_1 \in \mathcal{K}_{s+2}(A, v_1)$ be any vector of grade $s+2$. Obviously, the grade of y_1 with respect to A is the same as the grade of y_1 with respect to $\widehat{A}$. Since (5.1) holds for any u_1 of grade $s+3$ (with respect to A), Lemma 4.1 (with $i = m = s+2$ and $n = s+3$) implies that (with the obvious meaning of y_{s+2})

$$0 \;=\; (y_1, Ay_{s+2}) \;=\; (y_1, \widehat{A}y_{s+2}) \;=\; (\widehat{A}^* y_1, y_{s+2})\,, \tag{5.2}$$

where $\widehat{A}^* : \mathcal{K}_{s+2}(A, v_1) \to \mathcal{K}_{s+2}(A, v_1)$ is the adjoint operator of $\widehat{A}$. But this means that

$$\widehat{A}^* y_1 \;\in\; \operatorname{span}\{y_1, \ldots, \widehat{A}^s y_1\}\,. \tag{5.3}$$

Since this holds for any vector $y_1 \in \mathcal{K}_{s+2}(A, v_1) = \mathcal{K}_{s+2}(\widehat{A}, v_1)$ of grade $s+2 = d_{\min}(\widehat{A})$, Lemma 4.3 implies that $\widehat{A}$ is normal(t) for some $t \le s$. In particular, $\widehat{A}$ is normal, and has $s+2$ distinct eigenvalues, λ_k, $k = 1, \ldots, s+2$, with corresponding eigenvectors that are mutually orthogonal. Moreover, there exists a polynomial of degree at most s such that $p(\lambda_k) = \overline{\lambda}_k$, $k = 1, \ldots, s+2$. By definition, any eigenpair of $\widehat{A}$ is an eigenpair of A. Therefore, A acting on *any* vector of grade $s+2$ has $s+2$ distinct eigenvalues, and the corresponding eigenvectors are mutually orthogonal in the given inner product.

Step 2. *Extension to the whole space.*

Consider the cyclic decomposition of the whole space as in (2.5). Then the cyclic subspace $\mathcal{K}_{d_1}(A, w_1)$, where w_1 has the same minimal polynomial as A, can be further decomposed into

$$\mathcal{K}_{d_1}(A, w_1) \;=\; \mathcal{K}_{c_1}(A, z_1) \oplus \cdots \oplus \mathcal{K}_{c_\ell}(A, z_\ell)\,,$$

where the minimal polynomial of z_k is $(z - \lambda_k)^{c_k}$, $k = 1, \ldots, \ell$, and $\lambda_1, \ldots, \lambda_\ell$ are the distinct eigenvalues of A (see, e.g., [6, Chapter VII, section 2, Theorem 1]). In other words, $\mathcal{K}_{d_1}(A, w_1)$ is decomposed into ℓ cyclic invariant subspaces of A, where each of these corresponds to one of the ℓ distinct eigenvalues of A. (Recall that the restriction of A to $\mathcal{K}_{d_1}(A, w_1)$ is nonderogatory; see the example at the end of section 2.) In particular, if A is diagonalizable, then $\ell = d_{\min}(A)$, and $c_1 = \cdots = c_\ell = 1$, and $z_1, \ldots, z_\ell$ are eigenvectors of A corresponding to $\lambda_1, \ldots, \lambda_\ell$, respectively. In general, we can assume that $c_1 \ge c_2 \ge \cdots \ge c_\ell$. If $c_1 \ge s+2$, we can determine a vector v_1 of grade $s+2$ in $\mathcal{K}_{c_1}(A, z_1)$. But then the above implies that A acting on v_1 has $s+2$

distinct eigenvalues, which is a contradiction. Hence $c_1 < s+2$. We therefore can find an index m so that $c_1 + \cdots + c_{m-1} + \tilde{c}_m = s+2$, $0 \le \tilde{c}_m \le c_m$. Let $\tilde{z}_m$ be any vector of grade $\tilde{c}_m$ in $\mathcal{K}_{c_m}(A, z_m)$; then $w = z_1 + \cdots + z_{m-1} + \tilde{z}_m$ is of grade $s+2$. Hence A acting on w has $s+2$ distinct eigenvalues, which shows that $c_1 = c_2 = \cdots = c_\ell = 1$. To these eigenvalues correspond $s+2$ eigenvectors that are mutually orthogonal in the given inner product.

In the cyclic decomposition (2.5), the minimal polynomial of w_k is divisible by the minimal polynomial of w_{k+1}. Therefore the whole space completely decomposes into one-dimensional cyclic subspaces of A, i.e., A has a complete system of eigenvectors. We know that any $s+2$ of these corresponding to distinct eigenvalues of A must be mutually orthogonal. In the subspaces corresponding to a multiple eigenvalue we can find an orthogonal basis. Therefore A has a complete orthonormal system of eigenvectors, and hence A is normal. For every subset of $s+2$ distinct eigenvalues there exists a polynomial p of degree at most s that satisfies $p(\lambda_k) = \overline{\lambda}_k$ for all eigenvalues λ_k in the subset. If we take any two subsets having $s+1$ eigenvalues in common, the two corresponding polynomials must be identical. Thus all the polynomials are identical, so that A is normal(t) for some $t \le s$.

If $t < s$, then by the sufficiency result in Theorem 3.4, A admits an optimal $(t+2)$-term recurrence, which contradicts our initial assumption. Hence $t = s$, so that A is normal(s), which concludes the proof. ∎

6. Another proof based on the Rotation Lemma. In this section we discuss an elementary and more constructive approach to proving Theorem 3.5, which is based on orthogonal transformations ("rotations") of upper Hessenberg matrices. With this approach, we can prove Theorem 3.5 with the assumption $s+2 < d_{\min}(A)$ replaced by $s+3 < d_{\min}(A)$. We discuss the missing case $s+3 = d_{\min}(A)$ in section 7.

As above, let A be an invertible linear operator with minimal polynomial degree $d_{\min}(A)$ on a finite dimensional Hilbert space. Let s be a given nonnegative integer, $s+3 < d_{\min}(A)$. We assume that

(6.1) A admits an $(s+2)$-term recurrence, but A is *not* normal(s),

and derive a contradiction.

For deriving the contradiction we need some notation. Suppose that the space is decomposed into cyclic invariant subspaces of A as in (2.5). Let $\widehat{A}$ be the restriction of A to $\mathcal{K}_{d_1}(A, w_1)$, i.e., the invertible linear operator defined by

$$\widehat{A} \,:\, \mathcal{K}_{d_1}(A, w_1) \to \mathcal{K}_{d_1}(A, w_1)\,, \qquad v \mapsto Av \ \text{ for } v \in \mathcal{K}_{d_1}(A, w_1)\,.$$

The operator $\widehat{A}$ depends on the choice of w_1, which we consider fixed here, so $\widehat{A}$ is fixed as well. It is clear that $d_1 = d_{\min}(A) = d_{\min}(\widehat{A})$. We denote $d = d_1$ for simplicity. Now let $v_1 \in \mathcal{K}_d(\widehat{A}, w_1)$ be any initial vector of grade d, and let $v_1, \ldots, v_d$ be the corresponding orthogonal basis of $\mathcal{K}_d(\widehat{A}, v_1) = \mathcal{K}_d(\widehat{A}, w_1)$ generated by (3.1)–(3.3). Then the matrix representation of the operator $\widehat{A}$ with respect to this particular basis is a $d \times d$ unreduced upper Hessenberg matrix H_d, which is defined by the equation

$$\widehat{A}\,[v_1, \ldots, v_d] \;=\; [v_1, \ldots, v_d]\, H_d\,. \tag{6.2}$$

The matrix formed by the first $d-1$ columns of H_d coincides with the $d \times (d-1)$ upper Hessenberg matrix generated by (3.1)–(3.3) with $\widehat{A}$ and the initial vector v_1,

while the last column of H_d is given by the vector

$$h_d \;=\; \begin{bmatrix} h_{1,d} \\ \vdots \\ h_{d,d} \end{bmatrix}, \quad \text{where} \quad h_{m,d} \;=\; \frac{(\widehat{A}v_d, v_m)}{(v_m, v_m)}, \quad m = 1, \ldots, d\,. \tag{6.3}$$

In short, $H_d = [H_{d,d-1}, h_d]$. We now proceed in two steps.

Step 1. *Show that there exists a basis for which* $h_{1,d} \neq 0$.

We first show that under assumption (6.1) there exists an initial vector $v_1 \in \mathcal{K}_d(\widehat{A}, w_1)$ of grade $d = d_{\min}(\widehat{A})$ for which the matrix representation H_d of $\widehat{A}$ has $h_{1,d} \neq 0$. Suppose not, i.e., for all $v_1 \in \mathcal{K}_d(\widehat{A}, w_1)$ of grade $d_{\min}(\widehat{A})$, we have for the resulting entry $h_{1,d}$,

$$0 \;=\; h_{1,d} \;=\; \frac{(\widehat{A}v_d, v_1)}{(v_1, v_1)} \;=\; \frac{(v_d, \widehat{A}^* v_1)}{(v_1, v_1)}\,,$$

where $\widehat{A}^*$ is the adjoint of $\widehat{A}$. In particular, this implies that for all vectors $v_1 \in \mathcal{K}_d(\widehat{A}, w_1)$ of grade $d = d_{\min}(\widehat{A})$,

$$\widehat{A}^* v_1 \;\in\; \{v_1, \ldots, \widehat{A}^{d-2} v_1\}\,.$$

By Lemma 4.3, $\widehat{A}$ is normal(t) for some $t \leq d_{\min}(\widehat{A}) - 2$. Therefore, A acting on any vector of grade $d_{\min}(A)$ has $d_{\min}(A)$ distinct eigenvalues and corresponding eigenvectors that are mutually orthogonal. From this it is easy to see that A is normal(t). By the sufficiency result in Theorem 3.4, A admits an optimal $(t+2)$-term recurrence. However, we have assumed in (6.1) that A admits an optimal $(s+2)$-term recurrence, so $t = s$. But then A is normal(s), which contradicts the second part of the assumption. In summary, there exists an initial vector v_1 of grade $d = d_{\min}(A)$, such that (6.2) holds with $H_d = [H_{d,d-1}, h_d]$, where $H_{d,d-1}$ is $(s+2)$-band Hessenberg (this follows from the first part of our assumption), while $h_{1,d} \neq 0$.

Step 2. *Rotation of the nonzero entry* $h_{1,d}$.

The following result is called the Rotation Lemma for reasons apparent from its proof.

LEMMA 6.1 (Rotation Lemma). *Let* s, d *be nonnegative integers,* $s + 3 < d$. *Let* H_d *be a* $d \times d$ *unreduced upper Hessenberg matrix with* $h_{1,d} \neq 0$ *and* $H_{d,d-1}$, *the matrix formed by the first* $d-1$ *columns of* H_d, *being an* $(s+2)$*-band Hessenberg matrix. Then there exists a unitary matrix* G *such that* $\widetilde{H}_d \equiv G^* H_d G$ *is a* $d \times d$ *unreduced upper Hessenberg matrix with* $[\widetilde{h}_{1,d-1}, \widetilde{h}_{2,d-1}] \neq [0, 0]$.

Proof. The main idea of this proof is to find $d-1$ (complex) Givens rotations of the form

$$G_i \equiv \begin{bmatrix} I_{d-1-i} & & & \\ & c_i & \overline{s}_i & \\ & -s_i & c_i & \\ & & & I_{i-1} \end{bmatrix}, \quad c_i^2 + |s_i|^2 = 1, \quad c_i \in \mathbb{R}, \quad i = 1, \ldots, d-1, \tag{6.4}$$

which, applied symmetrically to H_d, "rotate" the nonzero entry $h_{1,d}$ to the $(d-1)$st column of the resulting matrix $\widetilde{H}_d = (G_1 \cdots G_{d-1})^* H_d (G_1 \cdots G_{d-1})$. To prove the assertion it suffices to show the following. First, $\widetilde{H}_d$ must be an unreduced upper Hessenberg matrix, and, second, at least one of its entries $\widetilde{h}_{1,d-1}, \widetilde{h}_{2,d-1}$ is nonzero. See Figure 6.1 for an illustration of this idea.

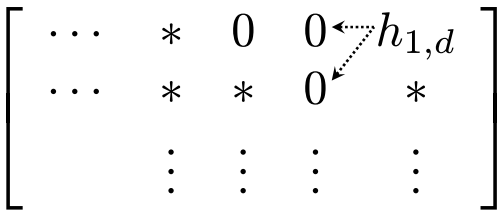


FIG. 6.1. *Graphical illustration of the Rotation Lemma. Shown is the upper-right-hand corner of* $H_d = [H_{d,d-1}, h_d]$. *We know that* $H_{d,d-1}$ *is* $(s+2)$*-band Hessenberg with* $s+3 < d$ *and that* $h_{1,d} \neq 0$. *We construct an orthogonal transformation* G *such that the matrix* $\widetilde{H}_d = G^* H_d G$ *remains unreduced upper Hessenberg, while the nonzero entry* $h_{1,d} \neq 0$ *is "rotated" to the last column of* $\widetilde{H}_{d,d-1}$, *so that at least one of its entries* $\widetilde{h}_{1,d-1}$ *and* $\widetilde{h}_{2,d-1}$ *is nonzero.*

Proceeding in an inductive manner, we denote $H^{(0)} \equiv H_d$. To start, choose $s_1 \in \mathbb{R} \setminus \{0\}$ and $c_1 \in \mathbb{R}$ such that $c_1^2 + s_1^2 = 1$. We have explicitly chosen real parameters s_1, c_1 since this simplifies our arguments below. These two parameters determine our first Givens rotation G_1 of the form (6.4). By construction, the matrix $H^{(1)} \equiv G_1^* H^{(0)} G_1$ is upper Hessenberg except for its entry

$$h_{d,d-2}^{(1)} = s_1 h_{d-1,d-2}^{(0)}.$$

Since $s_1 \neq 0$ and $h_{d-1,d-2}^{(0)} \neq 0$ ($H^{(0)}$ is unreduced), we have $h_{d,d-2}^{(1)} \neq 0$. The transformation by G_1 modifies only the last two rows and columns of $H^{(0)}$, so that the entries on the subdiagonal of $H^{(1)}$ satisfy $h_{i+1,i}^{(1)} = h_{i+1,i}^{(0)} \neq 0$, $i = 1, \ldots, d-3$. Next, we determine G_2 such that its application from the right to $H^{(1)}$ eliminates the nonzero entry in position $(d, d-2)$. Application of G_2^* from the left then introduces a nonzero entry in position $(d-1, d-3)$, which we will subsequently eliminate using G_3, and so forth.

In a general step $j = 2, \ldots, d-1$, suppose that $s_{j-1} \neq 0$, $h_{i+1,i}^{(j-1)} = h_{i+1,i}^{(0)} \neq 0$, $i = 1, \ldots, d-j-1$, and $h_{i+1,i}^{(j-1)} \neq 0$ for $i = d-j+2, \ldots, d-1$. Next suppose that

$$H^{(j-1)} \equiv G_{j-1}^* H^{(j-2)} G_{j-1}$$

is an upper Hessenberg matrix except for its entry

$$h_{d-j+2,d-j}^{(j-1)} = s_{j-1} h_{d-j+1,d-j}^{(0)} \neq 0.$$

The next Givens rotation G_j is (uniquely) determined to eliminate this nonzero entry, i.e., we determine c_j and s_j by the equation

$$[h_{d-j+2,d-j}^{(j-1)}, h_{d-j+2,d-j+1}^{(j-1)}] \begin{bmatrix} c_j & \overline{s}_j \\ -s_j & c_j \end{bmatrix} = [0, h_{d-j+2,d-j+1}^{(j)}]. \tag{6.5}$$

Since $h_{d-j+2,d-j}^{(j-1)} \neq 0$, it is clear that $s_j \neq 0$ and $h_{d-j+2,d-j+1}^{(j)} \neq 0$. As a result, the matrix

$$H^{(j)} \equiv G_j^* H^{(j-1)} G_j$$

is an upper Hessenberg except for its entry

$$h_{d-j+1,d-j-1}^{(j)} = s_j h_{d-j,d-j-1}^{(0)} \neq 0.$$

The unitary transformation determined by G_j modifies only $(d-j)$th and $(d-j+1)$st rows and columns of $H^{(j-1)}$. Therefore, the subdiagonal entries of $H^{(j)}$ satisfy $h^{(j)}_{i+1,i} = h^{(0)}_{i+1,i} \neq 0$ for $i = 1, \ldots, d-j-2$, and, since $h^{(j)}_{d-j+2,d-j+1} \neq 0$, we have shown inductively that indeed $h^{(j)}_{i+1,i} \neq 0$ for $i = d-j+1, \ldots, d-1$. In the end, we receive the unitary matrix $G = G_1 \cdots G_{d-1}$ and the upper Hessenberg matrix $H^{(d-1)} = G^* H^{(0)} G$ with $h^{(d-1)}_{i+1,i} \neq 0$ for $i = 2, \ldots, d-1$. To complete the proof we need to show that the initial parameters s_1, c_1 can be chosen so that, first, $h^{(d-1)}_{2,1} \neq 0$ ($H^{(d-1)}$ is unreduced), and, second, $[\, h^{(d-1)}_{1,d-1}, h^{(d-1)}_{2,d-1}\,] \neq [0,0]$.

First, if $h^{(d-1)}_{2,1} = 0$, then we must have $h^{(d-1)}_{1,1} \neq 0$, for if otherwise $H^{(d-1)}$ would be singular. From $H^{(d-1)} = G^* H^{(0)} G$ we receive $H^{(0)} G = G H^{(d-1)}$, and thus the first column of G is an eigenvector of $H^{(0)}$ corresponding to the eigenvalue $h^{(d-1)}_{1,1}$. Note that the first column of G depends on our choice of s_1, while the matrix $H^{(0)}$ is fixed and has at most d linearly independent eigenvectors. Apparently, the case $h^{(d-1)}_{2,1} = 0$ happens only for a finite number of values of s_1 (if any); almost every initial choice of s_1 will yield $h^{(d-1)}_{2,1} \neq 0$.

Second, we have assumed that the first $d-1$ columns of $H^{(0)}$ form an unreduced $(s+2)$-band Hessenberg matrix with $s+3 < d$, and therefore $h^{(0)}_{1,d-2} = h^{(0)}_{1,d-1} = 0$ (see Figure 6.1). Denote the entries of the (lower Hessenberg) matrix G by $g_{i,j}$. It is easy to see that $g_{d,d-1} = -c_2 s_1$. Again consider the matrix equation $H^{(0)} G = G H^{(d-1)}$. Comparing the entries in position $(1, d-1)$ on both sides shows that

$$-c_2 s_1 h^{(0)}_{1,d} \;=\; g_{1,1} h^{(d-1)}_{1,d-1} + g_{1,2} h^{(d-1)}_{1,d-1}\,, \tag{6.6}$$

where $h^{(0)}_{1,d} \neq 0$ and $s_1 \neq 0$. Therefore, to show that $[h^{(d-1)}_{1,d-1}, h^{(d-1)}_{2,d-1}] \neq [0,0]$, it suffices to show that $c_2 \neq 0$. For c_2 it holds that (cf. (6.5))

$$h^{(1)}_{d,d-2} c_2 - h^{(1)}_{d,d-1} s_2 \;=\; 0.$$

We know that $h^{(1)}_{d,d-2} \neq 0 \neq s_2$. Thus, $c_2 = 0$ if and only if $h^{(1)}_{d,d-1} = 0$, which holds if and only if

$$c_1 s_1 h^{(0)}_{d-1,d-1} + c_1^2 h^{(0)}_{d,d-1} - s_1^2 h^{(0)}_{d-1,d} - c_1 s_1 h^{(0)}_{d,d} \;=\; 0.$$

We write $s_1 = \sin(\theta)$, $c_1 = \cos(\theta)$ and apply standard identities for trigonometric functions to see that the above equation is equivalent with

$$\left(h^{(0)}_{d-1,d-1} - h^{(0)}_{d,d}\right) \sin(2\theta) + \left(h^{(0)}_{d,d-1} + h^{(0)}_{d-1,d}\right) \cos(2\theta) + \left(h^{(0)}_{d,d-1} - h^{(0)}_{d-1,d}\right) \;=\; 0\,.$$

The left-hand side in this equation is a nontrivial trigonometric polynomial of degree two, which has at most two roots in the interval $[0, 2\pi)$. Consequently, for almost all choices of s_1 we receive $c_2 \neq 0$, giving a nonzero right-hand side in (6.6). Hence, for almost all choices of s_1, we must have $[h^{(d-1)}_{1,d-1}, h^{(d-1)}_{2,d-1}] \neq [0,0]$. ∎

We can now derive the contradiction to (6.1). Consider the relation (6.2), where H_d is of the form assumed in the Lemma 6.1. Without loss of generality we may assume that the columns of V_d are normalized (normalization does not alter the nonzero pattern of H_d). By Lemma 6.1, there exists a unitary matrix G such that $\widetilde{H}_d = G^* H_d G$ is unreduced upper Hessenberg with either $\widetilde{h}_{1,d-1}$ or $\widetilde{h}_{2,d-1}$ nonzero. Then (6.2) is equivalent with

$$\widehat{A}(V_d G) \;=\; (V_d G) \widetilde{H}_d\,. \tag{6.7}$$

Denote the entries of G by $g_{i,j}$, and let $V_d G \equiv [y_1, \ldots, y_d]$. Then, since the basis $v_1, \ldots, v_d$ is orthonormal and the matrix G is unitary, the basis $y_1, \ldots, y_d$ is orthonormal,

$$(y_i, y_j) \;=\; \left(\sum_{k=1}^{d} v_k g_{k,i}, \sum_{k=1}^{d} v_k g_{k,j} \right) \;=\; \sum_{k=1}^{d} \overline{g}_{k,j} g_{k,i} \;=\; \delta_{i,j}\,,$$

where $\delta_{i,j}$ is the Kronecker delta. By (6.7), the vectors $y_1, \ldots, y_d$ form the unique (up to scaling) basis of $\mathcal{K}_d(\widehat{A}, y_1)$ generated by (3.1)–(3.3) with $\widehat{A}$ and starting vector y_1. But since $[\widetilde{h}_{1,d-1}, \widetilde{h}_{2,d-1}] \neq [0, 0]$, we see that $\widehat{A}$ (and hence A) admits for the given y_1 an optimal recurrence of length at least $d-1$. Since we have assumed that A admits an optimal $(s+2)$-term recurrence, we must have $d - 1 \leq s + 2$, or, equivalently, $d = d_{\min}(A) \leq s + 3$. This is a contradiction since $s + 3 < d_{\min}(A)$.

As claimed at the beginning of this section, we now have shown Theorem 3.5, with the assumption $s + 2 < d_{\min}(A)$ replaced by $s + 3 < d_{\min}(A)$.

7. Concluding discussion. In this section we discuss our rather theoretical analysis above.

1. *Matrix formulation and the Faber–Manteuffel theorem.*

When formulated in terms of matrices rather than linear operators, Theorems 3.4 and 3.5 make up the Faber–Manteuffel theorem [4] in the formulation given in [12, section 2]. We state this result here for completeness.

THEOREM 7.1. *Let A be an $N \times N$ nonsingular matrix with minimal polynomial degree $d_{\min}(A)$. Let B be an $N \times N$ Hermitian positive definite matrix, and let s be a nonnegative integer, $s + 2 < d_{\min}(A)$. Then A admits for the given B an optimal $(s+2)$-term recurrence if and only if A is B-normal(s).*

In this formulation, the Hilbert space from Theorems 3.4 and 3.5 is $\mathbb{C}^N$, equipped with the inner product generated by the Hermitian positive definite matrix B. (In case A is real, we consider B to be real as well, and the adjoint A^* is the regular transpose A^T.) The matrix A is B-normal(s) if its B-adjoint, i.e., the matrix $A^+ \equiv B^{-1} A^* B$, is a polynomial of degree s in A, and s is the smallest degree for which this is true. A complete characterization of the matrices A and B for which A is B-normal(s) is given in [12, section 3].

In this paper we have chosen the linear operator rather than the matrix formulation, because it appears to be a natural generalization. Moreover, both proofs we have given use the restriction of the linear operator A to certain cyclic invariant subspaces. In the matrix formulation, such restrictions lead to nonsquare as well as square but singular matrices. This involves a more complicated notation, which obstructs rather than helps the theoretical understanding. For instance, the restriction $\widehat{A}$ of a nonsingular $N \times N$ matrix A to a cyclic invariant subspace of A with (orthonormal) basis $v_1, \ldots, v_d$ can be represented as $\widehat{A} = VHV^*$, where $V = [v_1, \ldots, v_d]$ and H is a $d \times d$ nonsingular matrix. If $d < N$, $\widehat{A}$ is a singular $N \times N$ matrix (more precisely, $\widehat{A}$ has rank $d < N$). Any vector w in the cyclic invariant subspace can be represented as $w = V\omega$, where ω is a vector of length d containing the coefficients of w in the basis, so that $Aw = \widehat{A}w = VH\omega$, where VH is a (nonsquare) matrix of size $N \times d$. On the other hand, in the linear operator formulation, $\widehat{A}$ is invertible, and we may simply write $\widehat{A}w$ for the application of $\widehat{A}$ to any vector w in the space.

2. *On the strategies of the two different proofs of Theorem* 3.5.

The two different proofs of Theorem 3.5 given in this paper (with the second one excluding the case $s + 3 = d_{\min}(A)$; see below) follow two different strategies.

The first proof, given in section 5, is based on vectors of grade $s+2$, and works its way up to vectors of full grade $d_{\min}(A)$. This general strategy is similar to the one in the original paper of Faber and Manteuffel [4]. The details of our proof here, however, are quite different from [4]. In particular, simple arguments about the number of roots of certain polynomials (particularly in Lemmas 4.1 and 4.3) have replaced the continuity and topology arguments in the proof of [4]. We therefore consider this a simpler proof than the one given in [4].

The second proof, given in section 6, works immediately with vectors of full grade $d_{\min}(A)$. We consider this approach more elementary than our first proof. We assume that the assertion of Theorem 3.5 is false, i.e., that A admits an optimal $(s+2)$-term recurrence but is not normal(s). We show that if A is not normal(s), there must exist at least one initial vector v_1 of full grade $d = d_{\min}(A)$, for which the corresponding matrix H_d has a nonzero entry above its sth superdiagonal. If this nonzero entry already is in $H_{d,d-1}$, we are done. However, we cannot guarantee this, and therefore we need the Rotation Lemma to rotate a nonzero from the dth column of H_d into the $(d-1)$st column. This shows that A cannot admit an optimal $(s+2)$-term recurrence, contradicting our initial assumption.

3. *The Rotation Lemma and the missing case* $s+3 = d_{\min}(A)$.

In the Rotation Lemma we rotate the nonzero entry $h_{1,d}$, where $d = d_{\min}(A)$, to give $\widetilde{h}_{1,d-1} \neq 0$ *or* $\widetilde{h}_{2,d-1} \neq 0$; see Figure 6.1. Therefore, the matrix $\widetilde{H}_{d,d-1}$ is *at least* $(d-1)$-band Hessenberg. The shortest possible optimal recurrence that A may admit hence is of length $d-1$, or $s+2$ for $s = d-3$. The assumption that A admits an optimal recurrence of length $s+3 < d_{\min}(A)$ then leads to a contradiction.

To prove also the missing case $s+3 = d_{\min}(A)$, we need to guarantee that there exists a choice of s_1 so that $\widetilde{h}_{1,d-1} \neq 0$, giving a d-band Hessenberg matrix $\widetilde{H}_{d,d-1}$. Since Theorem 3.5 also holds for the case $s+3 = d_{\min}(A)$, we know that such s_1 *must exist*, but we were unable to prove the existence without using Theorem 3.5. Note, however, that in practical applications we are interested in recurrences of length $s+2 \ll d_{\min}(A)$. Therefore the missing case of the Rotation Lemma is only of rather theoretical interest.

We point out that the construction given in the Rotation Lemma, namely, the structure-preserving unitary transformation of an upper Hessenberg matrix, may be of interest beyond its application in our current context. To state this idea in a more general way, we introduce some notation. Let Ω_d be the set of the $d \times d$ unreduced upper Hessenberg matrices, and let $\Omega_d(s+2)$ be the subset consisting of the $(s+2)$-band Hessenberg matrices (these are unreduced by assumption; cf. Definition 3.1). Consider a *fixed* $H \in \Omega_d$, and define the set

$$\mathcal{R}_H \equiv \{\, G^* H G \in \Omega_d \;:\; G \text{ is unitary} \,\}.$$

Hence $\mathcal{R}_H$ is the set of all unitary transformations of H that are unreduced upper Hessenberg. Note that since $H \in \mathcal{R}_H$, the set $\mathcal{R}_H$ is nonempty. Using the Rotation Lemma (for $s+3 < d$) and Theorem 3.5 (for $s+3 = d$) the following result can be proved.

THEOREM 7.2. *Let s, d be given nonnegative integers, $s+2 < d$. For any $H \in \Omega_d$, the following assertions are equivalent:*

(1) *H is I-normal(s), i.e., $H^* = p(H)$ for a polynomial of (smallest possible) degree s;*
(2) *$\mathcal{R}_H \subset \Omega_d(s+2)$.*

This result means that an unreduced upper Hessenberg matrix H is I-normal(s) if and only if H is $(s+2)$-band Hessenberg, and all unitary transformations that preserve the unreduced upper Hessenberg structure of H also preserve the $(s+2)$-band structure of H.

4. *What distinguishes Theorem* 3.5 *from other results about normal operators.*

Theorem 3.5 gives a *necessary* condition when an operator A is normal (of some degree s). This condition is also *sufficient*, as shown by Theorem 3.4. Hence this condition might be taken as a *definition* of normality, and it might be included among the numerous equivalent definitions in [8, 3]. We believe, however, that the nature of the result distinguishes it from the many other equivalent ones. This distinction is clear from the second proof given in section 6.

Consider the linear operator A, and any cyclic invariant subspace $\mathcal{K}_d(A, v_1)$. Then the matrix representation of A with respect to the orthogonal basis $v_1, \ldots, v_d$ of $\mathcal{K}_d(A, v_1)$ generated by (3.1)–(3.3) is a $d \times d$ unreduced upper Hessenberg matrix H_d (cf. (6.2), where this is shown for the restriction of A to $\mathcal{K}_d(A, v_1)$). Typically, equivalent results for normality are derived using knowledge of *the whole matrix*, H_d in this case. But Theorem 3.5 is based only on knowledge of *a part of the matrix*, namely, the first $d-1$ columns of H_d. Our experience in this area shows that this difference also is the reason why Theorem 3.5 is rather difficult to prove, particularly when compared with other results about normal matrices or operators.

Acknowledgments. We thank Tom Manteuffel and an anonymous referee for suggestions that helped us to improve the presentation of the results.

REFERENCES

[1] W. E. Arnoldi, *The principle of minimized iteration in the solution of the matrix eigenvalue problem*, Quart. Appl. Math., 9 (1951), pp. 17–29.
[2] E. W. Cheney, *Introduction to Approximation Theory*, McGraw-Hill, New York, 1966.
[3] L. Elsner and K. D. Ikramov, *Normal matrices: An update*, Linear Algebra Appl., 285 (1998), pp. 291–303.
[4] V. Faber and T. Manteuffel, *Necessary and sufficient conditions for the existence of a conjugate gradient method*, SIAM J. Numer. Anal., 21 (1984), pp. 352–362.
[5] D. K. Faddeev and V. N. Faddeeva, *Computational Methods of Linear Algebra*, W. H. Freeman, San Francisco, 1963.
[6] F. R. Gantmacher, *The Theory of Matrices*. Vols. 1, 2, Chelsea, New York, 1959.
[7] A. Greenbaum, *Iterative Methods for Solving Linear Systems*, Frontiers in Appl. Math. 17, SIAM, Philadelphia, 1997.
[8] R. Grone, C. R. Johnson, E. M. de Sá, and H. Wolkowicz, *Normal matrices*, Linear Algebra Appl., 87 (1987), pp. 213–225.
[9] K. Hoffman and R. Kunze, *Linear Algebra*, 2nd ed., Prentice-Hall, Englewood Cliffs, NJ, 1971.
[10] A. Kleppner, *The cyclic decomposition theorem*, Integral Equations Oper. Theory, 25 (1996), pp. 490–495.
[11] J. Liesen and P. E. Saylor, *Orthogonal Hessenberg reduction and orthogonal Krylov subspace bases*, SIAM J. Numer. Anal., 42 (2005), pp. 2148–2158 (electronic).
[12] J. Liesen and Z. Strakoš, *On optimal short recurrences for generating orthogonal Krylov subspace bases*, SIAM Rev., to appear.
[13] Y. Saad, *Iterative Methods for Sparse Linear Systems*, 2nd ed., SIAM, Philadelphia, 2003.
[14] H. Zha and Z. Zhang, *The Arnoldi process, short recurrences and displacement ranks*, Linear Algebra Appl., 249 (1996), pp. 169–188.

SIAM J. NUMER. ANAL.
Vol. 46, No. 3, pp. 1338–1363
© 2008 Society for Industrial and Applied Mathematics

A FINITE ELEMENT METHOD FOR THE MAGNETOSTATIC PROBLEM IN TERMS OF SCALAR POTENTIALS*

ALFREDO BERMÚDEZ†, RODOLFO RODRÍGUEZ‡, AND PILAR SALGADO†

Abstract. The aim of this paper is to analyze a numerical method for solving the magnetostatic problem in a three-dimensional bounded domain containing prescribed currents and magnetic materials. The method discretizes a well-known formulation of this problem based on two scalar potentials: the total potential, defined in magnetic materials, and the reduced potential, defined in dielectric media and in nonmagnetic conductors carrying currents. The topology of the domain of each material is not assumed to be trivial. The resulting variational problem is proved to be well posed and is discretized by means of standard piecewise linear finite elements. Transmission conditions are imposed by means of a piecewise linear Lagrange multiplier on the surface separating the domains of both potentials. Error estimates for the numerical method are proved and the results of some numerical tests are reported to assess the performance of the method.

Key words. magnetostatic problem, finite element computational electromagnetism, scalar potentials

AMS subject classifications. 78M10, 65N30

DOI. 10.1137/06067568X

1. Introduction.

The goal of this paper is to analyze a finite element method to solve the magnetostatic problem in terms of scalar potentials in a three-dimensional bounded domain.

Several finite element methods have been developed to solve the magnetostatic problem in the last decades; see, for instance, [1, 5, 6, 10, 15, 16, 17, 21, 25, 26]. The main difference in the numerical methods lies in the choice of the primary unknowns. Indeed, formulations based on the magnetic vector potential [5, 10], on the magnetic field [15], on the magnetic and the magnetic induction fields [1, 20], or on different scalar potentials [6, 12, 17, 25, 26] can be found in engineering and mathematical journals.

The published numerical results show that the scalar potential formulations are specially efficient. In particular, the combination of two different potentials, the so-called *reduced scalar potential* and *total scalar potential* (see [16, 21]) seems to be the most effective in terms of accuracy and computer cost. In this approach, the total potential is defined in magnetic materials without current sources, while the reduced potential is defined in dielectric media and in nonmagnetic materials carrying currents.

This formulation, which was introduced by Simkin and Trowbridge in [25] for two-dimensional domains and extended to three-dimensional domains in [26], is very well known in the engineering literature and has been used in linear and nonlinear

*Received by the editors December 22, 2006; accepted for publication (in revised form) October 1, 2007; published electronically March 26, 2008.
http://www.siam.org/journals/sinum/46-3/67568.html

†Departamento de Matemática Aplicada, Universidade de Santiago de Compostela, 15706, Santiago de Compostela, Spain (mabermud@usc.es, mpilar@usc.es). The research of these authors was partially supported by MEC research project DPI2003-01316 (Spain) and Xunta de Galicia project PGIDIT06PXIB207052PR (Spain).

‡GI²MA, Departamento de Ingeniería Matemática, Universidad de Concepción, Casilla 160-C, Concepción, Chile (rodolfo@ing-mat.udec.cl). This author's research was partially supported by FONDAP in Applied Mathematics (Chile) and Xunta de Galicia project PGIDIT06PXIB207052PR (Spain).

cases. There exist several papers dealing with this formulation in bounded domains by using finite element methods (FEMs) (see [12, 16, 21, 31]) and others in unbounded domains by using boundary element methods (BEMs) or coupled BEM-FEM (see [3, 22, 23]).

However, to the best of the authors' knowledge, the approximation of the total scalar potential/reduced scalar potential formulation in bounded domains by standard finite elements has not been analyzed yet from a mathematical point of view. This is the aim of the present paper in the context of three-dimensional domains.

We notice that the magnetostatic problem in terms of scalar potentials leads to Poisson problems defined in different domains with suitable interface conditions. At this point, we will describe an amenable procedure to impose these conditions by means of a Lagrange multiplier. On the other hand, the topological structure of the region occupied by the magnetic materials plays a very important role, because the total potential is defined in this set from a curl-free magnetic field. We will allow for multiply connected domains, which leads to consider a total scalar potential with constant jumps on certain "cut" surfaces (see [2, 31]).

The outline of the paper is as follows. In section 2 we introduce the magnetostatic problem and the geometrical framework for our analysis. We recall the Biot–Savart law and deduce the differential equations of the total potential/reduced potential formulation on a bounded domain. In section 3 we obtain a weak formulation of the problem. We prove that it is well posed and that its solution satisfies the classical magnetostatic equations in a weak form. In section 4 we introduce a discretization based on standard finite elements, with the interface constraint between both potentials being imposed as an essential condition. This approach, although unfeasible from the implementation viewpoint, allows us to prove optimal-order error estimates. In section 5 we introduce a Lagrange multiplier to impose conveniently the interface constraint at the discrete level. We prove that the resulting mixed problem is equivalent to the previously analyzed discrete problem. Finally, in section 6, we report some numerical results which confirm the order of convergence predicted by the theory and allow us to assess the performance of the method.

2. The magnetostatic problem in terms of scalar potentials. The classical magnetostatic model is obtained by neglecting the time derivatives in Maxwell's equations. Thus, given a divergence-free stationary source current density $\boldsymbol{J}$, the magnetic field $\boldsymbol{H}$ and the magnetic induction $\boldsymbol{B}$ satisfy the following equations:

$$\mathbf{curl}\,\boldsymbol{H} = \boldsymbol{J}, \tag{2.1}$$

$$\operatorname{div}\boldsymbol{B} = 0. \tag{2.2}$$

The vector fields $\boldsymbol{H}$ and $\boldsymbol{B}$ are linked by the constitutive relation

$$\boldsymbol{B} = \mu\boldsymbol{H}, \tag{2.3}$$

where μ is the magnetic permeability which satisfies

$$0 < \mu_{\min} \le \mu \le \mu_{\max}. \tag{2.4}$$

Throughout the paper we use boldface letters to denote vector fields and variables, as well as vector-valued operators. We also use standard notation for Sobolev spaces and norms.

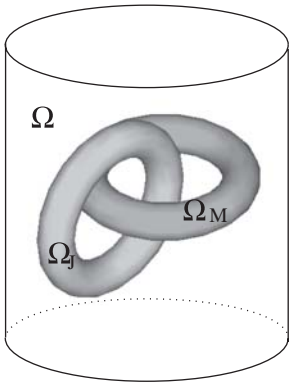


FIG. 2.1. *Sketch of the domains. Case* $\overline{\Omega}_{\rm J} \subset \Omega$.

The typical magnetostatic problem involves magnetic materials with permeability $\mu \neq \mu_0$ (where μ_0 is the magnetic permeability of vacuum), current sources, and eventually magnets. In this paper we focus on solving the magnetostatic problem in a bounded three-dimensional domain Ω containing prescribed currents and linear magnetic materials, but not magnets. This domain is assumed to be simply connected with a Lipschitz-continuous connected boundary Γ.

We denote by $\Omega_{\rm M}$ an open subset of Ω containing all the magnetic materials (i.e., all the points of Ω where $\mu \neq \mu_0$) and by $\Omega_{\rm J}$ another open subset of Ω such that $\overline{\Omega}_{\rm J}$ contains the support of the current source $\boldsymbol{J}$ in $\overline{\Omega}$. We assume that $\Omega_{\rm M} \cap \Omega_{\rm J} = \emptyset$, so that $\boldsymbol{J}|_{\Omega_{\rm M}} = \boldsymbol{0}$. This is a very typical case, for instance, in electromagnets like that of the second numerical test in section 6.2. The set $\Omega_{\rm M}$ is assumed to be connected although in general not simply connected. Moreover, we assume for simplicity that the boundary of $\Omega_{\rm M}$ is connected. The analysis of this paper extends with minimal changes to sets $\Omega_{\rm M}$ with disconnected boundaries; we discuss briefly this point in Remark 5.2 (see also Remark 3.1).

We further assume that $\overline{\Omega}_{\rm M} \subset \Omega$, which in practice is not restrictive at all. However, concerning $\Omega_{\rm J}$, it is useful not to impose a similar restriction, in order to allow for eventual current flows through the boundary of the domain Ω. Our analysis covers problems in which the domain Ω contains all the source currents (*closed circuits*), as that shown in Figure 2.1, and also problems in which there is a current flow through a part $\Gamma_{\rm J} := \partial\Omega_{\rm J} \cap \Gamma$ of the boundary (*open circuits*), as that shown in Figure 2.2. In any case, since the source current is divergence free, the net current flux through the boundary of Ω has to vanish.

To pose the magnetostatic problem in the bounded domain Ω, we have to add adequate boundary conditions to (2.1)–(2.3). In particular, we consider the following one:

$$\boldsymbol{B} \cdot \boldsymbol{n} = g \qquad \text{on } \Gamma,$$

where g is a given data function and $\boldsymbol{n}$ the outward unit normal vector to Γ. Notice that, by virtue of (2.2), the data g must have zero mean, i.e., $\int_\Gamma g = 0$. Let us remark that mixed boundary conditions, including $\boldsymbol{H} \times \boldsymbol{n} = \boldsymbol{0}$ on part of the boundary, can be dealt with the techniques of this paper, as well. However, for simplicity, we restrict our analysis to the simpler case above. We refer to Remark 5.1 for a brief discussion about how to proceed to deal with the boundary condition $\boldsymbol{H} \times \boldsymbol{n} = \boldsymbol{0}$.

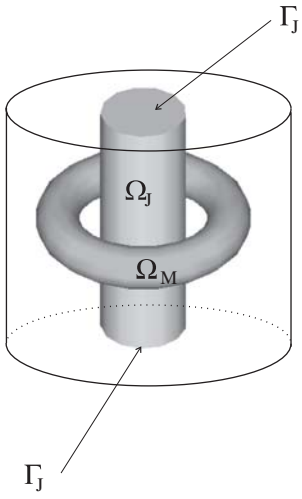


FIG. 2.2. *Sketch of the domains. Case* $\Gamma_J := \partial\Omega_J \cap \Gamma \neq \emptyset$.

Therefore, the magnetostatic problem that we are going to analyze, written in terms of the magnetic field, consists of finding $\boldsymbol{H}$ satisfying

$$\mathbf{curl}\,\boldsymbol{H} = \boldsymbol{J} \qquad \text{in } \Omega, \tag{2.5}$$

$$\operatorname{div}(\mu\boldsymbol{H}) = 0 \qquad \text{in } \Omega, \tag{2.6}$$

$$\mu\boldsymbol{H}\cdot\boldsymbol{n} = g \qquad \text{on } \Gamma. \tag{2.7}$$

This problem has a unique solution. In fact, let $\varphi \in \mathrm{H}^1(\Omega)/\mathbb{R}$ be the unique solution of $\operatorname{div}(\mu\,\mathbf{grad}\,\varphi) = 0$ in Ω with $\mu\frac{\partial\varphi}{\partial\boldsymbol{n}} = g$ on Γ. Clearly $\boldsymbol{H}$ is a solution of (2.5)–(2.7) if and only if $\hat{\boldsymbol{H}} := \boldsymbol{H} - \mathbf{grad}\,\varphi$ is a solution of the analogue of (2.5)–(2.7) with $g = 0$, and the latter has a unique solution (see [10, Remark p. 295], for instance).

Simkin and Trowbridge have introduced in [25] a formulation of this problem in the two-dimensional case based on two scalar potentials defined in different regions of the domain. The same authors have extended this formulation to the three-dimensional case in [26]. The aim of this paper is to analyze this formulation and a finite element method to compute its solution in the three-dimensional case.

Before introducing these potentials, we recall the Biot–Savart law. Given a current density $\widehat{\boldsymbol{J}} \in \mathrm{L}^2(\mathbb{R}^3)^3$ such that $\operatorname{div}\widehat{\boldsymbol{J}} = 0$ in $\mathbb{R}^3$, let $\boldsymbol{T}$ be defined as follows:

$$\boldsymbol{T}(\boldsymbol{x}) := \frac{1}{4\pi}\int_{\mathbb{R}^3} \widehat{\boldsymbol{J}}(\boldsymbol{y}) \times \frac{\boldsymbol{x}-\boldsymbol{y}}{|\boldsymbol{x}-\boldsymbol{y}|^3}\,\mathrm{d}\boldsymbol{y}\,, \qquad \boldsymbol{x} \in \mathbb{R}^3.$$

So defined, $\boldsymbol{T}$ turns out to be a divergence-free vector potential of $\widehat{\boldsymbol{J}}$ (see, for instance, [18, Lemma 2.2(b)]), namely,

$$\mathbf{curl}\,\boldsymbol{T} = \widehat{\boldsymbol{J}} \qquad \text{in } \mathbb{R}^3,$$

$$\operatorname{div}\boldsymbol{T} = 0 \qquad \text{in } \mathbb{R}^3.$$

Moreover, if the support of $\widehat{\boldsymbol{J}}$ is compact, then $\boldsymbol{T} \in \mathrm{H}^1(\mathbb{R}^3)^3$ and

$$\|\boldsymbol{T}\|_{1,\mathbb{R}^3} \leq C\,\|\widehat{\boldsymbol{J}}\|_{0,\mathbb{R}^3}$$

with C a strictly positive constant independent of $\widehat{\boldsymbol{J}}$ (see again [18, Lemma 2.2(b)]).

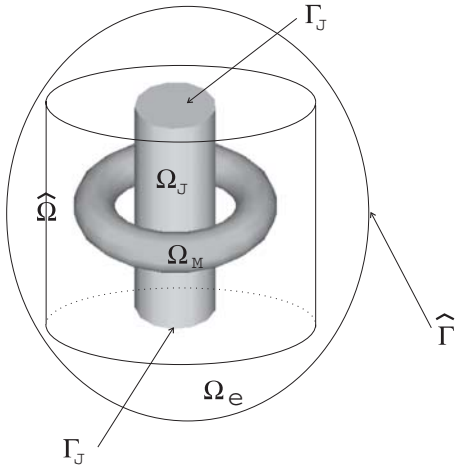


FIG. 2.3. *Auxiliary domains* $\widehat{\Omega}$ *and* $\Omega_{\rm e}$.

In our problem, the current density $\boldsymbol{J}$ is a divergence-free vector field defined only in Ω. In case that $\boldsymbol{J}\cdot\boldsymbol{n}=0$ on Γ (which necessarily happens if $\overline{\Omega}_{\rm J}\subset\Omega$, for instance), $\boldsymbol{J}$ can be extended by zero to the whole $\mathbb{R}^3$ in order to apply the previous results. Indeed, in such a case, if we denote this extension by $\widehat{\boldsymbol{J}}$, then $\operatorname{div}\widehat{\boldsymbol{J}}=0$ in $\mathbb{R}^3$. Thus, all the conclusions above hold true for the vector field $\boldsymbol{T}$ defined by the Biot–Savart law, which in this case reads

$$\boldsymbol{T}(\boldsymbol{x}):=\frac{1}{4\pi}\int_{\Omega_{\rm J}}\boldsymbol{J}(\boldsymbol{y})\times\frac{\boldsymbol{x}-\boldsymbol{y}}{|\boldsymbol{x}-\boldsymbol{y}|^3}\,\mathrm{d}\boldsymbol{y}\,,\qquad \boldsymbol{x}\in\mathbb{R}^3. \tag{2.8}$$

However, we are interested in solving the magnetostatic problem also in cases where $\boldsymbol{J}\cdot\boldsymbol{n}$ does not vanish on the whole Γ. In such a case we consider an auxiliary larger bounded domain $\widehat{\Omega}\supset\overline{\Omega}$ to obtain a vector potential for $\boldsymbol{J}$ (see Figure 2.3).

Let $\Omega_{\rm e}:=\widehat{\Omega}\setminus\overline{\Omega}$ and consider the following Neumann problem:

$$\begin{aligned}-\operatorname{div}(\mathbf{grad}\,\zeta)&=0 && \text{in } \Omega_{\rm e},\\ \mathbf{grad}\,\zeta\cdot\boldsymbol{n}&=\boldsymbol{J}\cdot\boldsymbol{n} && \text{on } \Gamma,\\ \mathbf{grad}\,\zeta\cdot\widehat{\boldsymbol{n}}&=0 && \text{on } \widehat{\Gamma}\end{aligned}$$

with $\widehat{\boldsymbol{n}}$ the outward unit normal vector to $\widehat{\Gamma}:=\partial\widehat{\Omega}$. This problem is well posed because $\int_\Gamma\boldsymbol{J}\cdot\boldsymbol{n}=\int_\Omega\operatorname{div}\boldsymbol{J}=0$. Moreover, it is straightforward to show that $\|\mathbf{grad}\,\zeta\|_{0,\Omega_{\rm e}}\leq C\,\|\boldsymbol{J}\|_{0,\Omega}$, with $C>0$ independent of $\boldsymbol{J}$.

Next, we define

$$\widehat{\boldsymbol{J}}:=\begin{cases}\boldsymbol{J} & \text{in } \Omega,\\ \mathbf{grad}\,\zeta & \text{in } \Omega_{\rm e},\\ \boldsymbol{0} & \text{in } \mathbb{R}^3\setminus\Omega_{\rm e},\end{cases}$$

so that $\operatorname{div}\widehat{\boldsymbol{J}}=0$ in $\mathbb{R}^3$. Thus, all the conclusions above also hold true in this case for the vector potential $\boldsymbol{T}$ defined by the Biot–Savart law, which now reads

$$\boldsymbol{T}(\boldsymbol{x})=\frac{1}{4\pi}\int_{\Omega_{\rm J}}\boldsymbol{J}(\boldsymbol{y})\times\frac{\boldsymbol{x}-\boldsymbol{y}}{|\boldsymbol{x}-\boldsymbol{y}|^3}\,\mathrm{d}\boldsymbol{y}+\frac{1}{4\pi}\int_{\Omega_{\rm e}}\mathbf{grad}\,\zeta(\boldsymbol{y})\times\frac{\boldsymbol{x}-\boldsymbol{y}}{|\boldsymbol{x}-\boldsymbol{y}|^3}\,\mathrm{d}\boldsymbol{y}\,,\qquad \boldsymbol{x}\in\mathbb{R}^3. \tag{2.9}$$

We collect these results in the following lemma.

LEMMA 2.1. *Let* $\boldsymbol{J} \in \mathrm{L}^2(\Omega)^3$ *be such that* $\operatorname{div} \boldsymbol{J} = 0$ *in* Ω. *Let* $\boldsymbol{T}$ *be defined by the Biot–Savart law* (2.8) *if* $\boldsymbol{J} \cdot \boldsymbol{n} = 0$ *on* Γ, *or by* (2.9) *otherwise. Then,* $\boldsymbol{T}$ *satisfies*

$$\mathbf{curl}\, \boldsymbol{T} = \boldsymbol{J} \qquad \text{in } \Omega, \tag{2.10}$$

$$\operatorname{div} \boldsymbol{T} = 0 \qquad \text{in } \Omega. \tag{2.11}$$

Moreover, $\boldsymbol{T}|_\Omega \in \mathrm{H}^1(\Omega)^3$ *with*

$$\|\boldsymbol{T}\|_{1,\Omega} \leq C \, \|\boldsymbol{J}\|_{0,\Omega}\,, \tag{2.12}$$

where C *is a strictly positive constant independent of* $\boldsymbol{J}$.

We can use $\boldsymbol{T}$ to obtain a potential formulation of problem (2.5)–(2.7). For this purpose notice that $\mathbf{curl}\,(\boldsymbol{H} - \boldsymbol{T}) = \boldsymbol{0}$ because of (2.5) and (2.10). Therefore, since Ω is simply connected, there exists a scalar potential ϕ^{R} such that

$$\boldsymbol{H} = \boldsymbol{T} - \mathbf{grad}\, \phi^{\mathrm{R}} \qquad \text{in } \Omega. \tag{2.13}$$

The scalar field ϕ^{R} is known as the *reduced scalar potential*.

In principle the magnetostatic problem (2.5)–(2.7) can be written in terms of ϕ^{R}, which leads to an important saving in computational effort. However, this possibility is not recommended in the literature because its numerical approximation leads to the so-called *cancellation error*. The origin of this error is that, often, the two terms $\boldsymbol{T}$ and $\mathbf{grad}\, \phi^{\mathrm{R}}$ in (2.13) are almost of the same size and opposite direction within the magnetic materials, which leads to large cancellation errors when computing the magnetic field $\boldsymbol{H}$ (see, for instance, [16, 17]). An alternative proposed by several authors consists of introducing an additional scalar potential in the domain Ω_{M} containing the magnetic materials (see [16, 25]).

To do this, let $\Omega_{\mathrm{R}} := \Omega \setminus \overline{\Omega}_{\mathrm{M}}$ and recall that $\mu = \mu_0$ (constant) in Ω_{R}. Also, let $\Gamma_{\mathrm{I}} := \partial\Omega_{\mathrm{R}} \cap \partial\Omega_{\mathrm{M}}$ be the interface between Ω_{R} and Ω_{M}, and $\boldsymbol{\nu}$ the unit normal vector to Γ_{I} pointing outwards Ω_{M}. Notice that $\Gamma_{\mathrm{I}} = \partial\Omega_{\mathrm{M}}$, because we have chosen $\Omega \supset \overline{\Omega}_{\mathrm{M}}$. Another consequence of this and the connectedness of $\partial\Omega_{\mathrm{M}}$ is that Ω_{R} is connected, too. As stated above, we restrict our analysis to domains Ω_{M} with a connected boundary, only for simplicity. If $\partial\Omega_{\mathrm{M}}$ were disconnected, then Ω_{R} would be disconnected, as well. We discuss briefly in Remarks 3.2 and 4.1 the main differences arising in such a case.

The domain Ω_{M} is supposed to be connected although in general not simply connected. However, we assume that there exists a finite number of open connected surfaces (so-called *cuts*) Σ_j, $j = 1, \ldots, J$, such that

- $\Sigma_j \subset \Omega_{\mathrm{M}}$ and $\partial\Sigma_j \subset \partial\Omega_{\mathrm{M}}$,
- $\overline{\Sigma}_j \cap \overline{\Sigma}_k = \emptyset$ for $j \neq k$,
- the open set $\widetilde{\Omega}_{\mathrm{M}} := \Omega_{\mathrm{M}} \setminus \bigcup_{j=1}^{J} \Sigma_j$ is pseudo-Lipschitz and simply connected.

Any set of surfaces satisfying these assumptions is called "an admissible set of cuts" (see [2], for instance, where the proper definition of *pseudo-Lipschitz* domains also appears in Definition 3.1). Figure 2.4 shows a particular case for $J = 1$.

We assume that for each of these cuts Σ_j there exists a simple closed curve $\gamma_j \subset \partial\Omega_{\mathrm{M}}$, crossing $\overline{\Sigma}_j$ once and only once, and such that γ_j is the boundary of an open surface $S_j \subset \Omega_{\mathrm{R}}$ (see again Figure 2.4).

Let $\widetilde{\Omega}_{\mathrm{M}}^j := \Omega_{\mathrm{M}} \setminus \Sigma_j$, $j = 1, \ldots, J$. We fix a unit normal vector $\boldsymbol{n}_j$ on Σ_j and denote its two faces by Σ_j^- and Σ_j^+, with $\boldsymbol{n}_j$ being the "outer" normal to $\widetilde{\Omega}_{\mathrm{M}}^j$ along

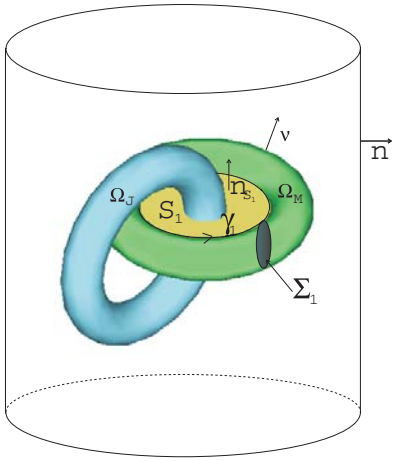


FIG. 2.4. *Cut surface and geometry description.*

Σ_j^+. We choose an orientation for each γ_j by taking its initial and end points on $\overline{\Sigma}_j^-$ and $\overline{\Sigma}_j^+$, respectively. We denote by $\boldsymbol{t}_j$ the unit vector tangent to γ_j.

For any function $\widetilde{\psi} \in \mathrm{H}^1(\widetilde{\Omega}_{\mathrm{M}})$, we denote by

$$[\![\widetilde{\psi}]\!]_{\Sigma_j} := \widetilde{\psi}|_{\Sigma_j^-} - \widetilde{\psi}|_{\Sigma_j^+}$$

the jump of $\widetilde{\psi}$ through Σ_j along $\boldsymbol{n}_j$. The gradient of $\widetilde{\psi}$ in $\mathcal{D}'(\widetilde{\Omega}_{\mathrm{M}})$ can be extended to $\mathrm{L}^2(\Omega_{\mathrm{M}})^3$ and will be denoted by $\widetilde{\mathbf{grad}}\,\widetilde{\psi}$.

Let Θ be the subspace of $\mathrm{H}^1(\widetilde{\Omega}_{\mathrm{M}})$ defined by

$$\Theta := \left\{ \widetilde{\psi} \in \mathrm{H}^1(\widetilde{\Omega}_{\mathrm{M}}) : \ [\![\widetilde{\psi}]\!]_{\Sigma_j} = \text{constant}, \ j = 1, \ldots, J \right\}.$$

For all $\widetilde{\psi} \in \mathrm{H}^1(\widetilde{\Omega}_{\mathrm{M}})$, we have that $\widetilde{\psi} \in \Theta$ if and only if $\mathbf{curl}\left(\widetilde{\mathbf{grad}}\,\widetilde{\psi}\right) = \mathbf{0}$ in Ω_{M} (see [2, Lemma 3.11]). More precisely, the operator $\mathbf{curl}: \ \mathrm{H}(\mathbf{curl}, \Omega_{\mathrm{M}}) \longrightarrow \mathrm{L}^2(\Omega_{\mathrm{M}})^3$ has a null space given by

$$\mathrm{Ker}\,(\mathbf{curl}) = \widetilde{\mathbf{grad}}\,\Theta. \tag{2.14}$$

Moreover, given $\widetilde{\psi} \in \Theta$, we have that $\widetilde{\psi} \in \mathrm{H}^1(\Omega_{\mathrm{M}})$ if and only if $[\![\widetilde{\psi}]\!]_{\Sigma_j} = 0$, $j = 1, \ldots, J$ (see [4]).

Now we are in a position to introduce the additional scalar potential that will be used to solve problem (2.5)–(2.7). Indeed, since $\boldsymbol{J}$ vanishes in Ω_{M}, by virtue of (2.14) the solution of this problem can be written in this subdomain as follows:

$$\boldsymbol{H}|_{\Omega_{\mathrm{M}}} = -\,\widetilde{\mathbf{grad}}\,\widetilde{\phi}, \qquad \widetilde{\phi} \in \Theta.$$

The scalar function $\widetilde{\phi}$ is known as the *total scalar potential*.

Remark 2.1. The jumps of this potential have a precise physical meaning. Indeed, provided $\boldsymbol{H}$ is sufficiently smooth, by using the Stokes theorem, (2.5), and the fact that $\gamma_j \subset \Gamma_{\mathrm{I}}$, we have that

$$[\![\widetilde{\phi}]\!]_{\Sigma_j} = -\int_{\gamma_j} \widetilde{\mathbf{grad}}\,\widetilde{\phi} \cdot \boldsymbol{t}_j = \int_{\gamma_j} \boldsymbol{H} \cdot \boldsymbol{t}_j = \int_{S_j} \mathbf{curl}\,\boldsymbol{H} \cdot \boldsymbol{n}_{S_j} = \int_{S_j} \boldsymbol{J} \cdot \boldsymbol{n}_{S_j} =: I_j,$$

with $\boldsymbol{n}_{S_j}$ being a unit vector normal to S_j, as shown in Figure 2.4. Thus, the jump of the total scalar potential $\widetilde{\phi}$ across each cut surface Σ_j is completely determined by the net current intensity I_j flowing through the surface S_j (as defined in the equation above). Notice that, in particular, if $S_j \cap \Omega_{\mathrm{J}} = \emptyset$, then $[\![\widetilde{\phi}]\!]_{\Sigma_j} = 0$.

Our next step is to write the magnetostatic problem (2.5)–(2.7) in terms of the reduced and the total scalar potentials, $\widetilde{\phi} \in \Theta$ and $\phi^{\mathrm{R}} \in \mathrm{H}^1(\Omega_{\mathrm{R}})$, so that

$$\boldsymbol{H} = \begin{cases} -\,\widetilde{\mathbf{grad}}\,\widetilde{\phi} & \text{in } \Omega_{\mathrm{M}}, \\ \boldsymbol{T} - \mathbf{grad}\,\phi^{\mathrm{R}} & \text{in } \Omega_{\mathrm{R}}. \end{cases} \tag{2.15}$$

Boundary and interface conditions must be written in terms of these potentials, as well. With this purpose, notice that for $\boldsymbol{J} \in \mathrm{L}^2(\Omega_{\mathrm{J}})^3$ we have from (2.10) that $\boldsymbol{H} \in \mathrm{H}(\mathbf{curl}, \Omega)$ and, hence, $\boldsymbol{H} \times \boldsymbol{\nu}$ does not have jumps across Γ_{I}. Analogously, $\mu \boldsymbol{H} \cdot \boldsymbol{\nu}$ does not have jumps across Γ_{I} either, because of (2.6). Consequently, we have to search for $\widetilde{\phi} \in \Theta$ and $\phi^{\mathrm{R}} \in \mathrm{H}^1(\Omega_{\mathrm{R}})$ satisfying the following equations:

$$-\operatorname{div}\left(\mu_0\,\mathbf{grad}\,\phi^{\mathrm{R}}\right) = 0 \qquad \text{in } \Omega_{\mathrm{R}}, \tag{2.16}$$

$$-\operatorname{div}\left(\mu\,\widetilde{\mathbf{grad}}\,\widetilde{\phi}\right) = 0 \qquad \text{in } \Omega_{\mathrm{M}}, \tag{2.17}$$

$$-\mu\,\mathbf{grad}\,\phi^{\mathrm{R}} \cdot \boldsymbol{n} = g - \mu_0\,\boldsymbol{T} \cdot \boldsymbol{n} \qquad \text{on } \Gamma, \tag{2.18}$$

$$\mu_0\,\mathbf{grad}\,\phi^{\mathrm{R}} \cdot \boldsymbol{\nu} - \mu\,\widetilde{\mathbf{grad}}\,\widetilde{\phi} \cdot \boldsymbol{\nu} = \mu_0\,\boldsymbol{T} \cdot \boldsymbol{\nu} \qquad \text{on } \Gamma_{\mathrm{I}}, \tag{2.19}$$

$$\mathbf{grad}\,\phi^{\mathrm{R}} \times \boldsymbol{\nu} - \widetilde{\mathbf{grad}}\,\widetilde{\phi} \times \boldsymbol{\nu} = \boldsymbol{T} \times \boldsymbol{\nu} \qquad \text{on } \Gamma_{\mathrm{I}}. \tag{2.20}$$

3. Well-posedness of the reduced/total potential formulation of the magnetostatic problem. The aim of this section is to obtain a weak formulation of the magnetostatic problem in terms of the total and the reduced scalar potentials and to prove its well-posedness.

To obtain the weak formulation, we multiply (2.16) and (2.17) by a sufficiently smooth test function ψ, integrate by parts, and use (2.18) and (2.19). Thus, we obtain

$$\int_{\Omega_{\mathrm{M}}} \mu\,\widetilde{\mathbf{grad}}\,\widetilde{\phi} \cdot \mathbf{grad}\,\psi + \int_{\Omega_{\mathrm{R}}} \mu_0\,\mathbf{grad}\,\phi^{\mathrm{R}} \cdot \mathbf{grad}\,\psi = \int_{\Gamma} \mu_0 \boldsymbol{T} \cdot \boldsymbol{n}\,\psi - \int_{\Gamma_{\mathrm{I}}} \mu_0 \boldsymbol{T} \cdot \boldsymbol{\nu}\,\psi - \int_{\Gamma} g\psi$$

as long as g is sufficiently smooth for the last integral to make sense.

For the subsequent analysis, (2.20) will be imposed as an essential condition. We introduce the space

$$\mathcal{X} := \Theta/\mathbb{R} \times \mathrm{H}^1(\Omega_{\mathrm{R}})/\mathbb{R},$$

endowed with the norm

$$\left\|(\widetilde{\psi}, \psi^{\mathrm{R}})\right\|_{\mathcal{X}} := \left(\left\|\,\widetilde{\mathbf{grad}}\,\widetilde{\psi}\,\right\|^2_{0,\Omega_{\mathrm{M}}} + \left\|\mathbf{grad}\,\psi^{\mathrm{R}}\right\|^2_{0,\Omega_{\mathrm{R}}}\right)^{1/2},$$

and the closed linear manifold

$$\mathcal{V}(\boldsymbol{T}) := \left\{(\widetilde{\psi}, \psi^{\mathrm{R}}) \in \mathcal{X} :\ \mathbf{grad}\,\psi^{\mathrm{R}} \times \boldsymbol{\nu} - \widetilde{\mathbf{grad}}\,\widetilde{\psi} \times \boldsymbol{\nu} = \boldsymbol{T} \times \boldsymbol{\nu} \text{ on } \Gamma_{\mathrm{I}}\right\}.$$

In what follows we prove a couple of lemmas involving this manifold and the corresponding subspace

$$\mathcal{V}(\boldsymbol{0}) := \left\{(\widetilde{\psi}, \psi^{\mathrm{R}}) \in \mathcal{X} :\ \mathbf{grad}\,\psi^{\mathrm{R}} \times \boldsymbol{\nu} - \widetilde{\mathbf{grad}}\,\widetilde{\psi} \times \boldsymbol{\nu} = \boldsymbol{0} \text{ on } \Gamma_{\mathrm{I}}\right\}.$$

LEMMA 3.1. $\mathcal{V}(\boldsymbol{T}) \cap \left[\mathrm{H}^2(\widetilde{\Omega}_{\mathrm{M}})/\mathbb{R} \times \mathrm{H}^2(\Omega_{\mathrm{R}})/\mathbb{R}\right] \neq \emptyset$.

Proof. Because of Lemma 2.1, since $\boldsymbol{J}$ vanishes in Ω_{M}, $\mathbf{curl}\,\boldsymbol{T} = \mathbf{0}$ in Ω_{M}. Hence, according to (2.14), there exists $\widetilde{\xi} \in \Theta$ such that $\boldsymbol{T}|_{\Omega_{\mathrm{M}}} = -\,\widetilde{\mathbf{grad}}\,\widetilde{\xi}$. Moreover, $\widetilde{\xi} \in \mathrm{H}^2(\widetilde{\Omega}_{\mathrm{M}})/\mathbb{R}$, because $\boldsymbol{T} \in \mathrm{H}^1(\Omega_{\mathrm{M}})^3$ (by virtue of Lemma 2.1, again). Consequently, the pair $\left(\widetilde{\xi}, 0\right)$ belongs to $\mathcal{V}(\boldsymbol{T}) \cap \left[\mathrm{H}^2(\widetilde{\Omega}_{\mathrm{M}})/\mathbb{R} \times \mathrm{H}^2(\Omega_{\mathrm{R}})/\mathbb{R}\right]$. ∎

LEMMA 3.2. $\left(\widetilde{\psi}, \psi^{\mathrm{R}}\right) \in \mathcal{V}(\mathbf{0})$ *if and only if there exists* $\psi \in \mathrm{H}^1(\Omega)/\mathbb{R}$ *such that* $\widetilde{\psi} = \psi|_{\Omega_{\mathrm{M}}}$ *and* $\psi^{\mathrm{R}} = \psi|_{\Omega_{\mathrm{R}}}$.

Proof. Given $(\widetilde{\psi}, \psi^{\mathrm{R}}) \in \mathcal{V}(\mathbf{0})$, let ψ^{R}_* be an H^1-extension to the whole Ω of a representative of ψ^{R} (that we also denote by ψ^{R}); i.e., $\psi^{\mathrm{R}}_* \in \mathrm{H}^1(\Omega)$ and $\psi^{\mathrm{R}}_*|_{\Omega_{\mathrm{R}}} = \psi^{\mathrm{R}}$. Hence, $\widetilde{\mathbf{grad}}\left(\widetilde{\psi} - \psi^{\mathrm{R}}_*|_{\Omega_{\mathrm{M}}}\right) \in \mathrm{H}_0(\mathbf{curl}^{\mathbf{0}}, \Omega_{\mathrm{M}})$, where

$$\mathrm{H}_0(\mathbf{curl}^{\mathbf{0}}, \Omega_{\mathrm{M}}) := \{\boldsymbol{G} \in \mathrm{H}(\mathbf{curl}, \Omega_{\mathrm{M}}) :\ \boldsymbol{G} \times \boldsymbol{\nu} = \mathbf{0} \text{ on } \partial\Omega_{\mathrm{M}} \text{ and } \mathbf{curl}\,\boldsymbol{G} = \mathbf{0} \text{ in } \Omega_{\mathrm{M}}\}.$$

As we show below, $\mathrm{H}_0(\mathbf{curl}^{\mathbf{0}}, \Omega_{\mathrm{M}}) \subset \mathbf{grad}\left(\mathrm{H}^1(\Omega_{\mathrm{M}})\right)$. Consequently, we have that $\widetilde{\mathbf{grad}}\,\widetilde{\psi} = \mathbf{grad}\left(\psi^{\mathrm{R}}_*|_{\Omega_{\mathrm{M}}}\right) + \widetilde{\mathbf{grad}}\left(\widetilde{\psi} - \psi^{\mathrm{R}}_*|_{\Omega_{\mathrm{M}}}\right) \in \mathbf{grad}\left(\mathrm{H}^1(\Omega_{\mathrm{M}})\right)$. Thus $\widetilde{\psi} \in \mathrm{H}^1(\Omega_{\mathrm{M}})/\mathbb{R}$ and $\widetilde{\mathbf{grad}}\,\widetilde{\psi}$ is the actual gradient of $\widetilde{\psi}$ in the whole Ω_{M} and not only in $\widetilde{\Omega}_{\mathrm{M}}$.

Now, from this and the definition of $\mathcal{V}(\mathbf{0})$, $\mathbf{grad}\,\psi^{\mathrm{R}} \times \boldsymbol{\nu} = \mathbf{grad}\,\widetilde{\psi} \times \boldsymbol{\nu}$ on Γ_{I}. Therefore, any pair of representatives of $\psi^{\mathrm{R}} \in \mathrm{H}^1(\Omega_{\mathrm{R}})/\mathbb{R}$ and $\widetilde{\psi} \in \mathrm{H}^1(\Omega_{\mathrm{M}})/\mathbb{R}$ only differ by a constant on Γ_{I}. Hence, we may choose these representatives (which we also denote by ψ^{R} and $\widetilde{\psi}$) coinciding on Γ_{I}. Thus, if we define ψ in Ω by $\psi|_{\Omega_{\mathrm{M}}} = \widetilde{\psi} \in \mathrm{H}^1(\Omega_{\mathrm{M}})$ and $\psi|_{\Omega_{\mathrm{R}}} = \psi^{\mathrm{R}} \in \mathrm{H}^1(\Omega_{\mathrm{R}})$, then $\psi \in \mathrm{H}^1(\Omega)$. The converse is clearly true.

There only remains to prove that $\mathrm{H}_0(\mathbf{curl}^{\mathbf{0}}, \Omega_{\mathrm{M}}) \subset \mathbf{grad}\left(\mathrm{H}^1(\Omega_{\mathrm{M}})\right)$. Indeed, given $\boldsymbol{G} \in \mathrm{H}_0(\mathbf{curl}^{\mathbf{0}}, \Omega_{\mathrm{M}})$, let $\xi \in \mathrm{H}^1_0(\Omega_{\mathrm{M}})$ be the unique solution of

$$\int_{\Omega_{\mathrm{M}}} \mathbf{grad}\,\xi \cdot \mathbf{grad}\,\eta = \int_{\Omega_{\mathrm{M}}} \boldsymbol{G} \cdot \mathbf{grad}\,\eta \qquad \forall \eta \in \mathrm{H}^1_0(\Omega_{\mathrm{M}}).$$

Then $\boldsymbol{G} - \mathbf{grad}\,\xi$ belongs to the space of *Dirichlet harmonic fields* in Ω_{M}, $\boldsymbol{\mathcal{H}} := \left\{\boldsymbol{F} \in \mathrm{H}_0(\mathbf{curl}^{\mathbf{0}}, \Omega_{\mathrm{M}}) :\ \mathrm{div}\,\boldsymbol{F} = 0 \text{ in } \Omega_{\mathrm{M}}\right\}$, which is trivial for $\partial\Omega_{\mathrm{M}}$ connected. Hence $\boldsymbol{G} = \mathbf{grad}\,\xi \in \mathbf{grad}\left(\mathrm{H}^1_0(\Omega_{\mathrm{M}})\right)$. ∎

Remark 3.1. This lemma remains true even for domains Ω_{M} with a not connected boundary. Indeed, in any case $\boldsymbol{\mathcal{H}} \subset \mathbf{grad}\left(\mathrm{H}^1(\Omega_{\mathrm{M}})\right)$ (see [2, Proposition 3.18]). Hence, $\boldsymbol{G} \in \mathbf{grad}\left(\mathrm{H}^1(\Omega_{\mathrm{M}})\right)$ in this case too.

The last lemma allows us to write in a more symmetric form the variational formulation of the magnetostatic problem in terms of the total and the reduced scalar potentials as follows.

PROBLEM P. *Find* $\left(\widetilde{\phi}, \phi^{\mathrm{R}}\right) \in \mathcal{V}(\boldsymbol{T})$ *such that*

$$\begin{aligned}(3.1)\qquad &\int_{\Omega_{\mathrm{M}}} \mu\,\widetilde{\mathbf{grad}}\,\widetilde{\phi} \cdot \widetilde{\mathbf{grad}}\,\widetilde{\psi} + \int_{\Omega_{\mathrm{R}}} \mu_0\,\mathbf{grad}\,\phi^{\mathrm{R}} \cdot \mathbf{grad}\,\psi^{\mathrm{R}} \\ &\qquad = \int_{\Gamma} \mu_0 \boldsymbol{T} \cdot \boldsymbol{n}\,\psi^{\mathrm{R}} - \int_{\Gamma_{\mathrm{I}}} \mu_0 \boldsymbol{T} \cdot \boldsymbol{\nu}\,\psi^{\mathrm{R}} - \left\langle g, \psi^{\mathrm{R}}\right\rangle_{\Gamma} \qquad \forall \left(\widetilde{\psi}, \psi^{\mathrm{R}}\right) \in \mathcal{V}(\mathbf{0}).\end{aligned}$$

Let us remark that according to Lemma 3.2, for all test functions $\left(\widetilde{\psi}, \psi^{\mathrm{R}}\right) \in \mathcal{V}(\mathbf{0})$, the total potential $\widetilde{\psi} \in \mathrm{H}^1(\Omega_{\mathrm{M}})$ and, consequently, the jumps $[\![\widetilde{\psi}]\!]_{\Sigma_j}$ actually vanish. (We denote $\widetilde{\mathbf{grad}}\,\widetilde{\psi}$ instead of $\mathbf{grad}\,\psi$ in the problem above, just to emphasize the symmetry of the formulation.) This does not happen for the trial functions $\left(\widetilde{\psi}, \psi^{\mathrm{R}}\right) \in$

$\mathcal{V}(\boldsymbol{T})$. However, the jumps of such $\widetilde{\psi}$ are completely determined through $\boldsymbol{T}$ by the intensities I_j defined in Remark 2.1. In fact, provided $\boldsymbol{J}$ is smooth enough,

$$\begin{aligned}[\![\widetilde{\psi}]\!]_{\Sigma_j} &= -\int_{\gamma_j} \widetilde{\mathbf{grad}}\,\widetilde{\psi}\cdot\boldsymbol{t}_j = \int_{\gamma_j}\left(\mathbf{grad}\,\psi^{\mathrm{R}} - \widetilde{\mathbf{grad}}\,\widetilde{\psi}\right)\cdot\boldsymbol{t}_j = \int_{\gamma_j}\boldsymbol{T}\cdot\boldsymbol{t}_j \\ &= \int_{S_j}\mathbf{curl}\,\boldsymbol{T}\cdot\boldsymbol{n}_{S_j} = \int_{S_j}\boldsymbol{J}\cdot\boldsymbol{n}_{S_j} = I_j,\end{aligned}$$

where we have used the Stokes theorem and (2.10).

Note that, according to the definition of $\mathcal{X}$, the test functions in Problem P are determined up to two arbitrary additive constants, whereas, according to Lemma 3.2, the trial functions seem to be determined up to one constant. However, there is no contradiction. Indeed, the equalities $\widetilde{\psi} = \psi|_{\Omega_{\mathrm{M}}}$ and $\psi^{\mathrm{R}} = \psi|_{\Omega_{\mathrm{R}}}$ from Lemma 3.2 have to be understood in $\mathrm{H}^1(\Omega_{\mathrm{R}})/\mathbb{R}$ and $\mathrm{H}^1(\Omega_{\mathrm{M}})/\mathbb{R}$, respectively, since $\widetilde{\psi}$ and ψ^{R} belong to these quotient spaces. Therefore, two arbitrary constants appear anyway.

The expression $\langle\cdot,\cdot\rangle_\Gamma$ in Problem P refers to the standard duality pairing between $\mathrm{H}^{1/2}(\Gamma)$ and its dual space $\mathrm{H}^{-1/2}(\Gamma)$. The data g is only assumed to belong to $\mathrm{H}^{-1/2}(\Gamma)$ and to satisfy $\langle g, 1\rangle_\Gamma = 0$, so that $\langle g, \psi^{\mathrm{R}}\rangle_\Gamma$ is well defined for $\psi^{\mathrm{R}} \in \mathrm{H}^1(\Omega_{\mathrm{R}})/\mathbb{R}$. Recall that this last assumption is actually necessary from the very beginning for (2.6) and (2.7) to hold true. On the other hand, $\int_\Gamma \mu_0\boldsymbol{T}\cdot\boldsymbol{n}\,\psi^{\mathrm{R}} - \int_{\Gamma_{\mathrm{I}}}\mu_0\boldsymbol{T}\cdot\boldsymbol{\nu}\,\psi^{\mathrm{R}}$ is also well defined for $\psi^{\mathrm{R}} \in \mathrm{H}^1(\Omega_{\mathrm{R}})/\mathbb{R}$, because $\int_\Gamma \mu_0\boldsymbol{T}\cdot\boldsymbol{n} - \int_{\Gamma_{\mathrm{I}}}\mu_0\boldsymbol{T}\cdot\boldsymbol{\nu} = \int_{\Omega_{\mathrm{R}}}\mathrm{div}\,(\mu_0\boldsymbol{T}) = 0$, as a consequence of (2.11) (recall that μ_0 is constant). Finally, the bilinear form in the left-hand side above,

$$a\big((\widetilde{\phi},\phi^{\mathrm{R}}),(\widetilde{\psi},\psi^{\mathrm{R}})\big) := \int_{\Omega_{\mathrm{M}}}\mu\,\widetilde{\mathbf{grad}}\,\widetilde{\phi}\cdot\widetilde{\mathbf{grad}}\,\widetilde{\psi} + \int_{\Omega_{\mathrm{R}}}\mu_0\,\mathbf{grad}\,\phi^{\mathrm{R}}\cdot\mathbf{grad}\,\psi^{\mathrm{R}},$$

is clearly continuous and elliptic in $\mathcal{X}$, as a consequence of (2.4). Therefore, we have the following result.

THEOREM 3.3. *Problem* P *has a unique solution.*

Proof. The result follows from Lax–Milgram's lemma, because a is $\mathcal{X}$-elliptic in $\mathcal{V}(\boldsymbol{0})$, the linear functional in the left-hand side of (3.1) is bounded in the same subspace, and $\mathcal{V}(\boldsymbol{T})$ is a nonempty closed linear manifold of $\mathcal{X}$ (Lemma 3.1). ∎

Once the total and the reduced scalar potentials are known, the magnetic field $\boldsymbol{H}$ can be readily computed by means of (2.15). In the following theorem we show that this vector field actually satisfies the equations of the magnetostatic problem (2.5)–(2.7).

THEOREM 3.4. *Let* $\big(\widetilde{\phi},\phi^{\mathrm{R}}\big)$ *be the solution of Problem* P *and let* $\boldsymbol{H}$ *be defined by* (2.15)*; namely,* $\boldsymbol{H} := -\,\widetilde{\mathbf{grad}}\,\widetilde{\phi}$ *in* Ω_{M} *and* $\boldsymbol{H} := \boldsymbol{T} - \mathbf{grad}\,\phi^{\mathrm{R}}$ *in* Ω_{R}*. Then,*

$$\mathbf{curl}\,\boldsymbol{H} = \boldsymbol{J} \qquad \text{in } \Omega, \tag{3.2}$$

$$\mathrm{div}\,(\mu\boldsymbol{H}) = 0 \qquad \text{in } \Omega, \tag{3.3}$$

$$\mu\boldsymbol{H}\cdot\boldsymbol{n} = g \qquad \text{on } \Gamma_{\mathrm{I}}. \tag{3.4}$$

Proof. For $\boldsymbol{H}$ defined by (2.15), $\boldsymbol{H}|_{\Omega_{\mathrm{M}}} \in \mathrm{H}(\mathbf{curl},\Omega_{\mathrm{M}})$ because of (2.14), and $\boldsymbol{H}|_{\Omega_{\mathrm{R}}} \in \mathrm{H}(\mathbf{curl},\Omega_{\mathrm{R}})$ because of Lemma 2.1. Moreover, from the definition of $\mathcal{V}(\boldsymbol{T})$, $\boldsymbol{H}|_{\Omega_{\mathrm{M}}}\times\boldsymbol{\nu} = \boldsymbol{H}|_{\Omega_{\mathrm{R}}}\times\boldsymbol{\nu}$ on Γ_{I}. Hence $\boldsymbol{H} \in \mathrm{H}(\mathbf{curl},\Omega)$ and (3.2) holds true as a consequence of (2.10).

For the second equation, let $\psi \in \mathcal{D}(\Omega)$. From Lemma 3.2, $(\psi|_{\Omega_{\rm M}}, \psi|_{\Omega_{\rm R}}) \in \mathcal{V}(\mathbf{0})$. Then

$$\begin{aligned}
-\int_\Omega \mu \boldsymbol{H} \cdot \mathbf{grad}\, \psi &= \int_{\Omega_{\rm M}} \mu\, \widetilde{\mathbf{grad}}\, \widetilde{\phi} \cdot \mathbf{grad}\, \psi - \int_{\Omega_{\rm R}} \left(\mu_0 \boldsymbol{T} - \mu_0\, \mathbf{grad}\, \phi^{\rm R}\right) \cdot \mathbf{grad}\, \psi \\
&= \int_{\Omega_{\rm M}} \mu\, \widetilde{\mathbf{grad}}\, \widetilde{\phi} \cdot \mathbf{grad}\, \psi + \int_{\Omega_{\rm R}} \mu_0\, \mathbf{grad}\, \phi^{\rm R} \cdot \mathbf{grad}\, \psi \\
&\quad + \int_{\Omega_{\rm R}} \mathrm{div}\, (\mu_0 \boldsymbol{T})\, \psi - \int_{\Gamma_{\rm I}} \mu_0 \boldsymbol{T} \cdot \boldsymbol{n}\, \psi \\
&= 0,
\end{aligned}$$

the latter holds because of (3.1) and (2.11). Hence, $\mu \boldsymbol{H} \in \mathrm{H}(\mathrm{div}, \Omega)$ and (3.3) holds true.

Finally, to prove the last equation, let $v \in \mathrm{H}^{1/2}(\Gamma)$ and $\psi \in \mathrm{H}^1(\Omega)$ be such that $\psi|_\Gamma = v$. According to Lemma 3.2, $(\psi|_{\Omega_{\rm M}}, \psi|_{\Omega_{\rm R}}) \in \mathcal{V}(\mathbf{0})$ and, hence, it can be used as a test function in (3.1). Therefore,

$$\begin{aligned}
\langle g, v \rangle_\Gamma &= -\int_{\Omega_{\rm M}} \mu\, \widetilde{\mathbf{grad}}\, \widetilde{\phi} \cdot \mathbf{grad}\, \psi - \int_{\Omega_{\rm R}} \mu_0\, \mathbf{grad}\, \phi^{\rm R} \cdot \mathbf{grad}\, \psi \\
&\quad + \int_\Gamma \mu_0 \boldsymbol{T} \cdot \boldsymbol{n}\, \psi - \int_{\Gamma_{\rm I}} \mu_0 \boldsymbol{T} \cdot \boldsymbol{\nu}\, \psi \\
&= -\int_{\Omega_{\rm M}} \mu\, \widetilde{\mathbf{grad}}\, \widetilde{\phi} \cdot \mathbf{grad}\, \psi - \int_{\Omega_{\rm R}} \mu_0\, \mathbf{grad}\, \phi^{\rm R} \cdot \mathbf{grad}\, \psi \\
&\quad + \int_{\Omega_{\rm R}} \mathrm{div}\, (\mu_0 \boldsymbol{T})\, \psi + \int_{\Omega_{\rm R}} \mu_0 \boldsymbol{T} \cdot \mathbf{grad}\, \psi \\
&= -\int_{\Omega_{\rm M}} \mu\, \widetilde{\mathbf{grad}}\, \widetilde{\phi} \cdot \mathbf{grad}\, \psi + \int_{\Omega_{\rm R}} \left(\mu_0 \boldsymbol{T} - \mu_0\, \mathbf{grad}\, \phi^{\rm R}\right) \cdot \mathbf{grad}\, \psi \\
&= \int_\Omega \mu \boldsymbol{H} \cdot \mathbf{grad}\, \psi \\
&= -\int_\Omega \mathrm{div}\, (\mu \boldsymbol{H})\, \psi + \langle \mu \boldsymbol{H} \cdot \boldsymbol{n}, \psi \rangle_\Gamma \\
&= \langle \mu \boldsymbol{H} \cdot \boldsymbol{n}, v \rangle_\Gamma\,,
\end{aligned}$$

where we have used (2.11) and (3.3). Hence (3.4) holds true in $\mathrm{H}^{-1/2}(\Gamma)$, and we conclude the proof. ∎

Remark 3.2. We have restricted the analysis of this section to connected sets $\Omega_{\rm R}$. However, it extends with minor modifications to a disconnected $\Omega_{\rm R}$. In such a case, let us denote by $\Omega_{\rm R}^1, \ldots, \Omega_{\rm R}^K$ its connected components. The reduced potential $\phi^{\rm R}$ is uniquely defined up to an additive constant for each of these connected components. Therefore, the function space for the pair of scalar potentials must be defined as follows: $\mathcal{X} := \Theta/\mathbb{R} \times \Pi_{k=1}^K \left[\mathrm{H}^1(\Omega_{\rm R}^k)/\mathbb{R}\right]$.

4. Finite element discretization. In this section we introduce a discretization of Problem P and prove an error estimate.

In what follows we assume that Ω, $\Omega_{\rm M}$, and $\Omega_{\rm R}$ are Lipschitz polyhedra and consider a family $\{\mathcal{T}_h\}$ of regular tetrahedral meshes of Ω such that each element $K \in \mathcal{T}_h$ is contained either in $\overline{\Omega}_{\rm M}$ or in $\overline{\Omega}_{\rm R}$ (h stands as usual for the corresponding mesh size). We define $\mathcal{T}_h^{\Omega_{\rm R}} := \left\{K \in \mathcal{T}_h :\ K \subset \overline{\Omega}_{\rm R}\right\}$ and $\mathcal{T}_h^{\Omega_{\rm M}} := \left\{K \in \mathcal{T}_h :\ K \subset \overline{\Omega}_{\rm M}\right\}$.

We assume that the cut surfaces Σ_j are polyhedral and that the meshes are compatible with them, in the sense that each $\overline{\Sigma}_j$ is a union of faces of tetrahedra for each mesh $\mathcal{T}_h$. Therefore, $\mathcal{T}_h^{\Omega_{\rm M}}$ can also be seen as a mesh of $\widetilde{\Omega}_{\rm M}$.

We introduce the following finite element spaces:

$$\begin{aligned}
\mathcal{L}_h(\widetilde{\Omega}_{\rm M}) &:= \left\{ \widetilde{\psi}_h \in \mathcal{C}(\widetilde{\Omega}_{\rm M}) :\ \widetilde{\psi}_h|_K \in \mathcal{P}_1(K)\ \forall K \in \mathcal{T}_h^{\Omega_{\rm M}} \right\},\\
\Theta_h &:= \left\{ \widetilde{\psi}_h \in \mathcal{L}_h(\widetilde{\Omega}_{\rm M}) :\ [\![\widetilde{\psi}_h]\!]_{\Sigma_j} = \text{constant},\ j = 1, \dots, J \right\},\\
\mathcal{L}_h(\Omega_{\rm R}) &:= \left\{ \psi_h^{\rm R} \in \mathcal{C}(\Omega_{\rm R}) :\ \psi_h^{\rm R}|_K \in \mathcal{P}_1(K)\ \forall K \in \mathcal{T}_h^{\Omega_{\rm R}} \right\},\\
\mathcal{X}_h &:= \Theta_h/\mathbb{R} \times \mathcal{L}_h(\Omega_{\rm R})/\mathbb{R}.
\end{aligned}$$

Definitions analogous to that of $\mathcal{L}_h(\Omega_{\rm R})$ hold for $\mathcal{L}_h(\Omega_{\rm M})$, $\mathcal{L}_h(\Omega)$, etc.

We also have to introduce suitable approximations of $\boldsymbol{T}$ to be used on $\Gamma_{\rm I}$ and Γ for the discrete problem. Notice that only the normal component of $\boldsymbol{T}$ is needed on Γ, whereas the tangential and normal components are needed on $\Gamma_{\rm I}$. We will use the Nédélec interpolant of $\boldsymbol{T}$ for the former and its Raviart–Thomas interpolant for the latter. In what follows we recall their definitions and some basic properties (see, for instance, [19] for more details and proofs).

The lowest-order Nédélec edge finite element space is defined as follows:

$$\boldsymbol{\mathcal{N}}_h(\Omega) := \{ \boldsymbol{G}_h \in {\rm H}(\mathbf{curl}, \Omega) :\ \boldsymbol{G}_h|_K \in \boldsymbol{\mathcal{N}}(K)\ \forall K \in \mathcal{T}_h \},$$

where $\boldsymbol{\mathcal{N}}(K) := \left\{ \boldsymbol{G}_h \in \mathcal{P}_1(K)^3 :\ \boldsymbol{G}_h(\boldsymbol{x}) = \boldsymbol{a} \times \boldsymbol{x} + \boldsymbol{b},\ \boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^3,\ \boldsymbol{x} \in K \right\}$. If $\boldsymbol{G}$ is smooth enough, then its Nédélec interpolant $\boldsymbol{G}_{\rm N}$ is well defined by

$$\boldsymbol{G}_{\rm N} \in \boldsymbol{\mathcal{N}}_h(\Omega) :\quad \int_\ell \boldsymbol{G}_{\rm N} \cdot \boldsymbol{t}_\ell = \int_\ell \boldsymbol{G} \cdot \boldsymbol{t}_\ell \quad \forall \ell \text{ edge of } \mathcal{T}_h,$$

where $\boldsymbol{t}_\ell$ denotes a unit vector tangent to ℓ. The results of [2] allow extending the above definition to the space ${\rm H}^r(\mathbf{curl}, \Omega) := \{ \boldsymbol{G} \in {\rm H}^r(\Omega) :\ \mathbf{curl}\, \boldsymbol{G} \in {\rm H}^r(\Omega) \}$ with $r > \frac{1}{2}$.

On the other hand, the lowest-order Raviart–Thomas finite element space is defined as follows:

$$\boldsymbol{\mathcal{R}}_h(\Omega) := \{ \boldsymbol{G}_h \in {\rm H}(\mathrm{div}, \Omega) :\ \boldsymbol{G}_h|_K \in \boldsymbol{\mathcal{R}}(K)\ \forall K \in \mathcal{T}_h \},$$

where $\boldsymbol{\mathcal{R}}(K) := \left\{ \boldsymbol{G}_h \in \mathcal{P}_1(K)^3 :\ \boldsymbol{G}_h(\boldsymbol{x}) = a\boldsymbol{x} + \boldsymbol{b};\ a \in \mathbb{R},\ \boldsymbol{b} \in \mathbb{R}^3,\ \boldsymbol{x} \in K \right\}$. For $\boldsymbol{G}$ sufficiently smooth, its Raviart–Thomas interpolant $\boldsymbol{G}_{\rm RT}$ is defined by

$$\boldsymbol{G}_{\rm RT} \in \boldsymbol{\mathcal{R}}_h(\Omega) :\quad \int_T \boldsymbol{G}_{\rm RT} \cdot \boldsymbol{n}_T = \int_T \boldsymbol{G} \cdot \boldsymbol{n}_T \quad \forall T \text{ face of } \mathcal{T}_h,$$

where $\boldsymbol{n}_T$ denotes a unit vector normal to T. The following error estimate holds for all $\boldsymbol{G} \in {\rm H}^1(\Omega)^3$:

$$\|\boldsymbol{G} - \boldsymbol{G}_{\rm RT}\|_{0,\Omega} \le Ch\, \|\boldsymbol{G}\|_{1,\Omega}\,. \tag{4.1}$$

Here and thereafter, C denotes a strictly positive constant, not necessarily the same at each occurrence, but always independent of h.

One basic property that will be used in what follows is that the Raviart–Thomas interpolant of a divergence free field is divergence free:

$$\mathrm{div}\, \boldsymbol{G} = 0 \quad \text{in } \Omega \qquad \Longrightarrow \qquad \mathrm{div}\, \boldsymbol{G}_{\rm RT} = 0 \quad \text{in } \Omega. \tag{4.2}$$

Next, we introduce the following finite-dimensional approximations of $\mathcal{V}(\boldsymbol{T})$ and $\mathcal{V}(\boldsymbol{0})$:

$$\mathcal{V}_h(\boldsymbol{T}) := \left\{ \left(\widetilde{\psi}_h, \psi_h^{\mathrm{R}}\right) \in \mathcal{X}_h : \ \mathbf{grad}\, \psi_h^{\mathrm{R}} \times \boldsymbol{\nu} - \widetilde{\mathbf{grad}}\, \widetilde{\psi}_h \times \boldsymbol{\nu} = \boldsymbol{T}_{\mathrm{N}} \times \boldsymbol{\nu} \text{ on } \Gamma_{\mathrm{I}} \right\},$$

$$\mathcal{V}_h(\boldsymbol{0}) := \left\{ \left(\widetilde{\psi}_h, \psi_h^{\mathrm{R}}\right) \in \mathcal{X}_h : \ \mathbf{grad}\, \psi_h^{\mathrm{R}} \times \boldsymbol{\nu} - \widetilde{\mathbf{grad}}\, \widetilde{\psi}_h \times \boldsymbol{\nu} = \boldsymbol{0} \text{ on } \Gamma_{\mathrm{I}} \right\}.$$

Let us remark that the Nédélec interpolant $\boldsymbol{T}_{\mathrm{N}}$ in the definition of $\mathcal{V}_h(\boldsymbol{T})$ is well defined, because $\boldsymbol{T} \in \mathrm{H}^1(\Omega_{\mathrm{M}})$ and $\mathbf{curl}\, \boldsymbol{T} = \boldsymbol{0}$ in Ω_{M} (Lemma 2.1).

Thus we arrive to the following discrete version of Problem P.

Problem DP. *Find* $\left(\widetilde{\phi}_h, \phi_h^{\mathrm{R}}\right) \in \mathcal{V}_h(\boldsymbol{T})$ *such that*

$$\begin{aligned} (4.3) \quad & \int_{\Omega_{\mathrm{M}}} \mu\, \widetilde{\mathbf{grad}}\, \widetilde{\phi}_h \cdot \widetilde{\mathbf{grad}}\, \widetilde{\psi}_h + \int_{\Omega_{\mathrm{R}}} \mu_0\, \mathbf{grad}\, \phi_h^{\mathrm{R}} \cdot \mathbf{grad}\, \psi_h^{\mathrm{R}} \\ & \quad = \int_{\Gamma} \mu_0 \boldsymbol{T}_{\mathrm{RT}} \cdot \boldsymbol{n}\, \psi_h^{\mathrm{R}} - \int_{\Gamma_{\mathrm{I}}} \mu_0 \boldsymbol{T}_{\mathrm{RT}} \cdot \boldsymbol{\nu}\, \psi_h^{\mathrm{R}} - \left\langle g, \psi_h^{\mathrm{R}} \right\rangle_{\Gamma} \qquad \forall \left(\widetilde{\psi}_h, \psi_h^{\mathrm{R}}\right) \in \mathcal{V}_h(\boldsymbol{0}). \end{aligned}$$

Notice that the right-hand side above is well defined for $\psi_h^{\mathrm{R}} \in \mathcal{L}_h(\Omega_{\mathrm{R}})/\mathbb{R}$. Indeed, $\int_{\Gamma} \mu_0 \boldsymbol{T}_{\mathrm{RT}} \cdot \boldsymbol{n} - \int_{\Gamma_{\mathrm{I}}} \mu_0 \boldsymbol{T}_{\mathrm{RT}} \cdot \boldsymbol{\nu} = \int_{\Omega_{\mathrm{R}}} \operatorname{div}(\mu_0 \boldsymbol{T}_{\mathrm{RT}}) = 0$, because of (4.2), since $\operatorname{div} \boldsymbol{T} = 0$ in Ω_{R} and μ_0 is constant. On the other hand, g has been assumed to satisfy $\langle g, 1 \rangle_{\Gamma} = 0$.

The following lemmas will be used to prove the existence and uniqueness of solution for this problem, as well as an error estimate.

Lemma 4.1. $\mathcal{V}_h(\boldsymbol{T}) \neq \emptyset$.

Proof. According to Lemma 3.1, there exists $\left(\widetilde{\psi}, \psi^{\mathrm{R}}\right) \in \mathcal{V}(\boldsymbol{T})$ such that $\widetilde{\psi} \in \mathrm{H}^2(\widetilde{\Omega}_{\mathrm{M}})/\mathbb{R}$ and $\psi^{\mathrm{R}} \in \mathrm{H}^2(\Omega_{\mathrm{R}})/\mathbb{R}$. Thus, the Lagrange interpolants of these functions, $\widetilde{\psi}_{\mathrm{L}}$ and $\psi_{\mathrm{L}}^{\mathrm{R}}$, are well defined and $\left(\widetilde{\psi}_{\mathrm{L}}, \psi_{\mathrm{L}}^{\mathrm{R}}\right) \in \mathcal{X}_h$. Moreover, since the Nédélec interpolant of a gradient is the gradient of the Lagrange interpolant (see, for instance, again [19]), we have

$$\mathbf{grad}\, \psi_{\mathrm{L}}^{\mathrm{R}} \times \boldsymbol{\nu} - \widetilde{\mathbf{grad}}\, \widetilde{\psi}_{\mathrm{L}} \times \boldsymbol{\nu} = \left(\mathbf{grad}\, \psi^{\mathrm{R}}\right)_{\mathrm{N}} \times \boldsymbol{\nu} - \left(\widetilde{\mathbf{grad}}\, \widetilde{\psi}\right)_{\mathrm{N}} \times \boldsymbol{\nu} = \boldsymbol{T}_{\mathrm{N}} \times \boldsymbol{\nu} \qquad \text{on } \Gamma_{\mathrm{I}}.$$

Thus, $\left(\widetilde{\psi}_{\mathrm{L}}, \psi_{\mathrm{L}}^{\mathrm{R}}\right) \in \mathcal{V}_h(\boldsymbol{T})$. ∎

Lemma 4.2. *The following properties hold true:*

(i) *there exists* $\widetilde{\xi} \in \Theta \cap \mathrm{H}^2(\widetilde{\Omega}_{\mathrm{M}})$ *such that* $\boldsymbol{T} = -\widetilde{\mathbf{grad}}\, \widetilde{\xi}$ *in* Ω_{M}*;*

(ii) *there exists* $\widetilde{\xi}_{\mathrm{L}} \in \Theta_h$ *such that* $\boldsymbol{T}_{\mathrm{N}} = -\widetilde{\mathbf{grad}}\, \widetilde{\xi}_{\mathrm{L}}$ *in* Ω_{M}*;*

(iii) $\left\| \widetilde{\mathbf{grad}}\, \widetilde{\xi} - \widetilde{\mathbf{grad}}\, \widetilde{\xi}_{\mathrm{L}} \right\|_{0,\Omega_{\mathrm{M}}} \leq Ch \left\| \boldsymbol{T} \right\|_{1,\Omega_{\mathrm{M}}}$.

Proof. Property (i) has been already shown in the proof of Lemma 3.1. To prove (ii), let $\widetilde{\xi}_{\mathrm{L}}$ be the Lagrange interpolant of $\widetilde{\xi}$, which is well defined because $\widetilde{\xi} \in \mathrm{H}^2(\widetilde{\Omega}_{\mathrm{M}})$. For all the vertices P of the mesh lying on a given cut surface Σ_j,

$$[\![\widetilde{\xi}_{\mathrm{L}}(P)]\!]_{\Sigma_j} = [\![\widetilde{\xi}(P)]\!]_{\Sigma_j} = \text{constant (the same for all such } P).$$

Then $[\![\widetilde{\xi}_{\mathrm{L}}]\!]_{\Sigma_j}$ is constant too and, hence, $\widetilde{\xi}_{\mathrm{L}} \in \Theta_h$. Now, using again the relation between the Nédélec and the Lagrange interpolants, we have

$$-\widetilde{\mathbf{grad}}\, \widetilde{\xi}_{\mathrm{L}} = -\left(\widetilde{\mathbf{grad}}\, \widetilde{\xi}\right)_{\mathrm{N}} = \boldsymbol{T}_{\mathrm{N}} \qquad \text{in } \Omega_{\mathrm{M}},$$

and we conclude (ii).

Finally (iii) is a consequence of the standard approximation properties of the Lagrange interpolant (see, for instance, [19]) and of (i). ∎

LEMMA 4.3. *Let* $\left(\widetilde{\phi}, \phi^{\mathrm{R}}\right) \in \mathcal{V}(\boldsymbol{T})$ *be such that* $\widetilde{\phi} \in \mathrm{H}^{1+r}(\widetilde{\Omega}_{\mathrm{M}})/\mathbb{R}$ *and* $\phi^{\mathrm{R}} \in \mathrm{H}^{1+r}(\Omega_{\mathrm{R}})/\mathbb{R}$ *with* $0 < r \leq 1$. *Then, there exists* $\left(\widetilde{\phi}_{\mathrm{I}}, \phi^{\mathrm{R}}_{\mathrm{I}}\right) \in \mathcal{V}_h(\boldsymbol{T})$ *such that*

$$\left\|\left(\widetilde{\phi} - \widetilde{\phi}_{\mathrm{I}}, \phi^{\mathrm{R}} - \phi^{\mathrm{R}}_{\mathrm{I}}\right)\right\|_{\mathcal{X}} \leq Ch^r \left(\left\| \widetilde{\mathbf{grad}}\, \widetilde{\phi} \right\|_{r,\widetilde{\Omega}_{\mathrm{M}}} + \left\|\mathbf{grad}\, \phi^{\mathrm{R}}\right\|_{r,\Omega_{\mathrm{R}}} + \|\boldsymbol{J}\|_{0,\Omega}\right).$$

Proof. Let $\phi^{\mathrm{R}}_{*} \in \mathrm{H}^{1+r}(\Omega_{\mathrm{R}})$ be an extension to the whole Ω of a representative of $\phi^{\mathrm{R}} \in \mathrm{H}^{1+r}(\Omega_{\mathrm{R}})/\mathbb{R}$ such that $\left\|\mathbf{grad}\, \phi^{\mathrm{R}}_{*}\right\|_{r,\Omega} \leq C \left\|\mathbf{grad}\, \phi^{\mathrm{R}}\right\|_{r,\Omega_{\mathrm{R}}}$. From now on, we will not distinguish between classes modulo $\mathbb{R}$ and their representatives.

Let $\widetilde{\xi} \in \Theta \cap \mathrm{H}^2(\widetilde{\Omega}_{\mathrm{M}})$ be as in Lemma 4.2(i) and let $\widetilde{\eta} := \widetilde{\phi} - \phi^{\mathrm{R}}_{*}|_{\Omega_{\mathrm{M}}} + \widetilde{\xi} \in \mathrm{H}^{1+r}(\widetilde{\Omega}_{\mathrm{M}})$. For $\left(\widetilde{\phi}, \phi^{\mathrm{R}}\right) \in \mathcal{V}(\boldsymbol{T})$, we have $\widetilde{\mathbf{grad}}\, \widetilde{\eta} \times \boldsymbol{\nu} = \widetilde{\mathbf{grad}}\, \widetilde{\phi} \times \boldsymbol{\nu} - \mathbf{grad}\, \phi^{\mathrm{R}} \times \boldsymbol{\nu} - \boldsymbol{T} \times \boldsymbol{\nu} = \mathbf{0}$ on the connected set Γ_{I}; hence, $\widetilde{\eta}|_{\Gamma_{\mathrm{I}}}$ is constant. Therefore, for each cut surface Σ_j, $\widetilde{\eta}|_{\partial\Sigma_j^+}$ and $\widetilde{\eta}|_{\partial\Sigma_j^-}$ are identical functions in $\mathrm{H}^r(\partial\Sigma_j)$. Consequently, $[\![\widetilde{\eta}]\!]_{\Sigma_j} = [\![\widetilde{\eta}]\!]_{\partial\Sigma_j} = 0$, so that $\widetilde{\eta}$ lies in $\mathrm{H}^{1+r}(\Omega_{\mathrm{M}})$, rather than in $\mathrm{H}^{1+r}(\widetilde{\Omega}_{\mathrm{M}})$. Moreover,

$$\left\| \mathbf{grad}\, \widetilde{\eta} \right\|_{r,\Omega_{\mathrm{M}}} \leq \left\| \widetilde{\mathbf{grad}}\, \widetilde{\phi} \right\|_{r,\Omega_{\mathrm{M}}} + C \left\|\mathbf{grad}\, \phi^{\mathrm{R}}\right\|_{r,\Omega_{\mathrm{M}}} + \|\boldsymbol{T}\|_{r,\Omega_{\mathrm{M}}}.$$

Let $\widetilde{\eta}_{\mathrm{C}} \in \mathcal{L}_h(\Omega_{\mathrm{M}})$ be a Clément-like interpolant of $\widetilde{\eta} \in \mathrm{H}^{1+r}(\Omega_{\mathrm{M}})$ such that $\widetilde{\eta}_{\mathrm{C}}|_{\Gamma_{\mathrm{I}}} = \widetilde{\eta}|_{\Gamma_{\mathrm{I}}}$ (constant) and

$$\left\| \mathbf{grad}\, \widetilde{\eta} - \mathbf{grad}\, \widetilde{\eta}_{\mathrm{C}} \right\|_{0,\Omega_{\mathrm{M}}} \leq Ch^r \left\| \mathbf{grad}\, \widetilde{\eta} \right\|_{r,\Omega_{\mathrm{M}}}$$

(see [27]). Let $\phi^{\mathrm{R}}_{*\mathrm{C}} \in \mathcal{L}_h(\Omega)$ be a Clément interpolant of $\phi^{\mathrm{R}}_{*} \in \mathrm{H}^{1+r}(\Omega_{\mathrm{R}})$ such that

$$\left\|\mathbf{grad}\, \phi^{\mathrm{R}}_{*} - \mathbf{grad}\, \phi^{\mathrm{R}}_{*\mathrm{C}}\right\|_{0,\Omega} \leq Ch^r \left\|\mathbf{grad}\, \phi^{\mathrm{R}}\right\|_{r,\Omega_{\mathrm{R}}}.$$

Let $\widetilde{\xi}_{\mathrm{L}} \in \Theta_h$ be as in Lemma 4.2(ii), so that from property (iii) of the same lemma

$$\left\| \widetilde{\mathbf{grad}}\, \widetilde{\xi} - \widetilde{\mathbf{grad}}\, \widetilde{\xi}_{\mathrm{L}} \right\|_{0,\Omega_{\mathrm{M}}} \leq Ch \left\|\boldsymbol{T}\right\|_{1,\Omega_{\mathrm{M}}}.$$

Let $\widetilde{\phi}_{\mathrm{I}} := \widetilde{\eta}_{\mathrm{C}} + \phi^{\mathrm{R}}_{*\mathrm{C}}|_{\Omega_{\mathrm{M}}} - \widetilde{\xi}_{\mathrm{L}} \in \Theta_h$ and $\phi^{\mathrm{R}}_{\mathrm{I}} := \phi^{\mathrm{R}}_{*\mathrm{C}}|_{\Omega_{\mathrm{R}}} \in \mathcal{L}_h(\Omega_{\mathrm{R}})$. Hence

$$\widetilde{\mathbf{grad}}\, \widetilde{\phi}_{\mathrm{I}} \times \boldsymbol{\nu} - \mathbf{grad}\, \phi^{\mathrm{R}}_{\mathrm{I}} \times \boldsymbol{\nu} = \widetilde{\mathbf{grad}}\, \widetilde{\eta}_{\mathrm{C}} \times \boldsymbol{\nu} - \widetilde{\mathbf{grad}}\, \widetilde{\xi}_{\mathrm{L}} \times \boldsymbol{\nu} = \boldsymbol{T}_{\mathrm{N}} \times \boldsymbol{\nu} \qquad \text{on } \Gamma_{\mathrm{I}},$$

the last equality follows because $\widetilde{\mathbf{grad}}\, \widetilde{\eta}_{\mathrm{C}} \times \boldsymbol{\nu} = 0$ on Γ_{I}, since the Clément-like interpolant $\widetilde{\eta}_{\mathrm{C}}$ was defined as constant on Γ_{I}. Thus, $\left(\widetilde{\phi}_{\mathrm{I}}, \phi^{\mathrm{R}}_{\mathrm{I}}\right) \in \mathcal{V}_h(\boldsymbol{T})$.

On the other hand,

$$\begin{aligned}\left\| \widetilde{\mathbf{grad}}\, \left(\widetilde{\phi} - \widetilde{\phi}_{\mathrm{I}}\right) \right\|_{0,\Omega_{\mathrm{M}}} &= \left\| \widetilde{\mathbf{grad}}\, \left(\widetilde{\eta} + \phi^{\mathrm{R}}_{*} - \widetilde{\xi}\right) - \widetilde{\mathbf{grad}}\, \left(\widetilde{\eta}_{\mathrm{C}} + \phi^{\mathrm{R}}_{*\mathrm{C}} - \widetilde{\xi}_{\mathrm{L}}\right) \right\|_{0,\Omega_{\mathrm{M}}} \\ &\leq Ch^r \left(\left\| \widetilde{\mathbf{grad}}\, \widetilde{\phi} \right\|_{r,\widetilde{\Omega}_{\mathrm{M}}} + \left\|\mathbf{grad}\, \phi^{\mathrm{R}}\right\|_{r,\Omega_{\mathrm{R}}} + \|\boldsymbol{T}\|_{1,\Omega}\right)\end{aligned}$$

and

$$\left\|\mathbf{grad}\, \left(\phi^{\mathrm{R}} - \phi^{\mathrm{R}}_{\mathrm{I}}\right)\right\|_{0,\Omega_{\mathrm{R}}} \leq \left\|\mathbf{grad}\, \phi^{\mathrm{R}}_{*} - \mathbf{grad}\, \phi^{\mathrm{R}}_{*\mathrm{C}}\right\|_{0,\Omega} \leq Ch^r \left\|\mathbf{grad}\, \phi^{\mathrm{R}}\right\|_{r,\Omega_{\mathrm{R}}}.$$

Finally, the last two inequalities and (2.12) allow us to conclude the proof. ∎

Now, we are in a position to prove the main result of this section.

THEOREM 4.4. *Problem* DP *has a unique solution* $\left(\widetilde{\phi}_h, \phi^{\mathrm{R}}_h\right)$. *Moreover, if the solution* $(\widetilde{\phi}, \phi^{\mathrm{R}})$ *of Problem* P *is such that* $\boldsymbol{H}|_{\Omega_{\mathrm{M}}} = -\widetilde{\mathbf{grad}}\, \widetilde{\phi} \in \mathrm{H}^r(\Omega_{\mathrm{M}})^3$ *and* $\boldsymbol{H}|_{\Omega_{\mathrm{R}}} =$

$\boldsymbol{T}|_{\Omega_{\mathrm{R}}} - \mathbf{grad}\,\phi^{\mathrm{R}} \in \mathrm{H}^r(\Omega_{\mathrm{R}})^3$, with $0 < r \leq 1$, then there exists a strictly positive constant C such that

$$\left\|\left(\widetilde{\phi} - \widetilde{\phi}_h, \phi^{\mathrm{R}} - \phi_h^{\mathrm{R}}\right)\right\|_{\mathcal{X}} \leq Ch^r \left(\left\|\boldsymbol{H}\right\|_{r,\Omega_{\mathrm{M}}} + \left\|\boldsymbol{H}\right\|_{r,\Omega_{\mathrm{R}}} + \|\boldsymbol{J}\|_{0,\Omega}\right).$$

Proof. The existence and uniqueness of the solution follows from Lax–Milgram's lemma, since the bilinear form a is $\mathcal{X}$-elliptic in $\mathcal{V}_h(\mathbf{0})$ and $\mathcal{V}_h(\boldsymbol{T})$ is a nonempty closed linear manifold of $\mathcal{X}_h$ (Lemma 4.1).

To prove the error estimate, we start noticing that, as a consequence of the assumptions of this theorem and Lemma 2.1, the solution $\left(\widetilde{\phi}, \phi^{\mathrm{R}}\right)$ of Problem P satisfies the hypotheses of Lemma 4.3. Then, let $\left(\widetilde{\phi}_{\mathrm{I}}, \phi_{\mathrm{I}}^{\mathrm{R}}\right)$ be as in that lemma. By using the ellipticity of a, we have that

$$\begin{aligned}\mu_{\min}\left\|\left(\widetilde{\phi} - \widetilde{\phi}_h, \phi^{\mathrm{R}} - \phi_h^{\mathrm{R}}\right)\right\|_{\mathcal{X}}^2 &\leq a\big((\widetilde{\phi} - \widetilde{\phi}_h, \phi^{\mathrm{R}} - \phi_h^{\mathrm{R}}), (\widetilde{\phi} - \widetilde{\phi}_h, \phi^{\mathrm{R}} - \phi_h^{\mathrm{R}})\big)\\ &= a\big((\widetilde{\phi} - \widetilde{\phi}_h, \phi^{\mathrm{R}} - \phi_h^{\mathrm{R}}), (\widetilde{\phi} - \widetilde{\phi}_{\mathrm{I}}, \phi^{\mathrm{R}} - \phi_{\mathrm{I}}^{\mathrm{R}})\big)\\ &\quad + a\big((\widetilde{\phi} - \widetilde{\phi}_h, \phi^{\mathrm{R}} - \phi_h^{\mathrm{R}}), (\widetilde{\phi}_{\mathrm{I}} - \widetilde{\phi}_h, \phi_{\mathrm{I}}^{\mathrm{R}} - \phi_h^{\mathrm{R}})\big).\end{aligned}$$

We estimate the first term on the right-hand side above by using the continuity of a and Lemma 4.3:

$$\begin{aligned}&a\big((\widetilde{\phi} - \widetilde{\phi}_h, \phi^{\mathrm{R}} - \phi_h^{\mathrm{R}}), (\widetilde{\phi} - \widetilde{\phi}_{\mathrm{I}}, \phi^{\mathrm{R}} - \phi_{\mathrm{I}}^{\mathrm{R}})\big)\\ &\quad\leq \mu_{\max}\left\|\left(\widetilde{\phi} - \widetilde{\phi}_h, \phi^{\mathrm{R}} - \phi_h^{\mathrm{R}}\right)\right\|_{\mathcal{X}}\left\|\left(\widetilde{\phi} - \widetilde{\phi}_{\mathrm{I}}, \phi^{\mathrm{R}} - \phi_{\mathrm{I}}^{\mathrm{R}}\right)\right\|_{\mathcal{X}}\\ &\quad\leq Ch^r\left\|\left(\widetilde{\phi} - \widetilde{\phi}_h, \phi^{\mathrm{R}} - \phi_h^{\mathrm{R}}\right)\right\|_{\mathcal{X}}\left(\left\|\,\widetilde{\mathbf{grad}}\,\widetilde{\phi}\,\right\|_{r,\widetilde{\Omega}_{\mathrm{M}}} + \left\|\mathbf{grad}\,\phi^{\mathrm{R}}\right\|_{r,\Omega_{\mathrm{R}}} + \|\boldsymbol{J}\|_{0,\Omega_{\mathrm{J}}}\right).\end{aligned}$$

For the second term, notice that $\left(\widetilde{\phi}_{\mathrm{I}} - \widetilde{\phi}_h, \phi_{\mathrm{I}}^{\mathrm{R}} - \phi_h^{\mathrm{R}}\right) \in \mathcal{V}_h(\mathbf{0})$, because $\left(\widetilde{\phi}_{\mathrm{I}}, \phi_{\mathrm{I}}^{\mathrm{R}}\right)$ and $\left(\widetilde{\phi}_h, \phi_h^{\mathrm{R}}\right)$ both belong to $\mathcal{V}_h(\boldsymbol{T})$. Then, from (3.1) and (4.3) we obtain

$$\begin{aligned}&a\big((\widetilde{\phi} - \widetilde{\phi}_h, \phi^{\mathrm{R}} - \phi_h^{\mathrm{R}}), (\widetilde{\phi}_{\mathrm{I}} - \widetilde{\phi}_h, \phi_{\mathrm{I}}^{\mathrm{R}} - \phi_h^{\mathrm{R}})\big)\\ &\quad= \int_{\Gamma} \mu_0\,(\boldsymbol{T} - \boldsymbol{T}_{\mathrm{RT}}) \cdot \boldsymbol{n}\,\left(\phi_{\mathrm{I}}^{\mathrm{R}} - \phi_h^{\mathrm{R}}\right) - \int_{\Gamma_{\mathrm{I}}} \mu_0\,(\boldsymbol{T} - \boldsymbol{T}_{\mathrm{RT}}) \cdot \boldsymbol{\nu}\,\left(\phi_{\mathrm{I}}^{\mathrm{R}} - \phi_h^{\mathrm{R}}\right)\\ &\quad= \int_{\Omega_{\mathrm{R}}} \mu_0\,(\boldsymbol{T} - \boldsymbol{T}_{\mathrm{RT}}) \cdot \mathbf{grad}\,\left(\phi_{\mathrm{I}}^{\mathrm{R}} - \phi_h^{\mathrm{R}}\right)\\ &\quad\leq \mu_0\,\|\boldsymbol{T} - \boldsymbol{T}_{\mathrm{RT}}\|_{0,\Omega_{\mathrm{R}}}\left\|\mathbf{grad}\,\left(\phi_{\mathrm{I}}^{\mathrm{R}} - \phi_h^{\mathrm{R}}\right)\right\|_{0,\Omega_{\mathrm{R}}}.\end{aligned}$$

For the last equality, we have used integration by parts and that $\operatorname{div}\boldsymbol{T} = \operatorname{div}\boldsymbol{T}_{\mathrm{RT}} = 0$ in Ω_{R}, because of (2.11) and (4.2), respectively. Next, using (4.1) and (2.12), we have

$$\|\boldsymbol{T} - \boldsymbol{T}_{\mathrm{RT}}\|_{0,\Omega_{\mathrm{R}}} \leq Ch\,\|\boldsymbol{T}\|_{1,\Omega_{\mathrm{R}}} \leq Ch\,\|\boldsymbol{J}\|_{0,\Omega_{\mathrm{J}}}.$$

On the other hand, from Lemma 4.3, there holds

$$\begin{aligned}\left\|\mathbf{grad}\,\left(\phi_{\mathrm{I}}^{\mathrm{R}} - \phi_h^{\mathrm{R}}\right)\right\|_{0,\Omega_{\mathrm{R}}} &\leq \left\|\mathbf{grad}\,\left(\phi_{\mathrm{I}}^{\mathrm{R}} - \phi^{\mathrm{R}}\right)\right\|_{0,\Omega_{\mathrm{R}}} + \left\|\mathbf{grad}\,\left(\phi^{\mathrm{R}} - \phi_h^{\mathrm{R}}\right)\right\|_{0,\Omega_{\mathrm{R}}}\\ &\leq Ch^r\left(\left\|\,\widetilde{\mathbf{grad}}\,\widetilde{\phi}\,\right\|_{r,\widetilde{\Omega}_{\mathrm{M}}} + \left\|\mathbf{grad}\,\phi^{\mathrm{R}}\right\|_{r,\Omega_{\mathrm{R}}} + \|\boldsymbol{J}\|_{0,\Omega_{\mathrm{J}}}\right)\\ &\quad + \left\|\left(\widetilde{\phi} - \widetilde{\phi}_h, \phi^{\mathrm{R}} - \phi_h^{\mathrm{R}}\right)\right\|_{\mathcal{X}}.\end{aligned}$$

Finally, the estimate of the theorem follows from all these inequalities, basic algebraic manipulations, and the fact that $\left\|\,\widetilde{\mathbf{grad}}\,\widetilde{\phi}\,\right\|_{r,\widetilde{\Omega}_{\mathrm{M}}} + \left\|\mathbf{grad}\,\phi^{\mathrm{R}}\right\|_{r,\Omega_{\mathrm{R}}} = \|\boldsymbol{H}\|_{r,\Omega_{\mathrm{M}}} + \|\boldsymbol{H} - \boldsymbol{T}\|_{r,\Omega_{\mathrm{R}}} \leq \|\boldsymbol{H}\|_{r,\Omega_{\mathrm{M}}} + \|\boldsymbol{H}\|_{r,\Omega_{\mathrm{R}}} + C\,\|\boldsymbol{J}\|_{0,\Omega_{\mathrm{J}}}$ (Lemma 2.1). ∎

Remark 4.1. As in the case of the continuous problem (see Remark 3.2), the analysis of this section also extends with minor modifications to a disconnected set Ω_{R}. In particular, if $\Omega_{\mathrm{R}}^1, \ldots, \Omega_{\mathrm{R}}^K$ are its connected components, then the discrete space for the pair of scalar potentials must be $\mathcal{X}_h := \Theta_h/\mathbb{R} \times \Pi_{k=1}^K \left[\mathcal{L}_h(\Omega_{\mathrm{R}}^k)/\mathbb{R}\right]$.

5. Implementation of the interface constraint by means of a Lagrange multiplier. The numerical solution of Problem DP requires to impose somehow the constraint on Γ_{I} appearing in the definition of $\mathcal{V}_h(\boldsymbol{T})$. For this purpose one could use standard techniques proposed in the literature (see, for instance, [7, 31]). As an alternative, in what follows we will introduce a Lagrange multiplier to handle this interface condition in a weak sense. While this approach would be in principle more expensive in terms of degrees of freedom, it has the advantage that its computer implementation is quite straightforward and only needs the usual data structures for FEMs.

Let us denote by $\mathcal{T}_h^{\Gamma_{\mathrm{I}}}$ the mesh induced by $\mathcal{T}_h$ on the polyhedral surface Γ_{I}: $\mathcal{T}_h^{\Gamma_{\mathrm{I}}} := \{T \subset \Gamma_{\mathrm{I}} :\ T \text{ face of } K \in \mathcal{T}_h\}$. Let $\widetilde{\Gamma}_{\mathrm{I}} := \Gamma_{\mathrm{I}} \setminus \bigcup_{j=1}^J \partial\Sigma_j$. Notice that, since each $\overline{\Sigma}_j$ is a union of faces of $K \in \mathcal{T}_h$, $\mathcal{T}_h^{\Gamma_{\mathrm{I}}}$ can also be seen as a mesh on $\widetilde{\Gamma}_{\mathrm{I}}$. Let

$$\mathcal{L}_h(\Gamma_{\mathrm{I}}) := \left\{\eta_h \in \mathcal{C}(\Gamma_{\mathrm{I}}) :\ \eta_h|_T \in \mathcal{P}_1(T)\ \forall T \in \mathcal{T}_h^{\Gamma_{\mathrm{I}}}\right\},$$
$$\mathcal{L}_h(\widetilde{\Gamma}_{\mathrm{I}}) := \left\{\widetilde{\eta}_h \in \mathcal{C}(\widetilde{\Gamma}_{\mathrm{I}}) :\ \widetilde{\eta}_h|_T \in \mathcal{P}_1(T)\ \forall T \in \mathcal{T}_h^{\Gamma_{\mathrm{I}}}\right\}.$$

Given $\widetilde{\eta}_h \in \mathcal{L}_h(\widetilde{\Gamma}_{\mathrm{I}})$, clearly $\widetilde{\eta}_h \in \mathcal{L}_h(\Gamma_{\mathrm{I}})$ if and only if $[\![\widetilde{\eta}_h]\!]_{\partial\Sigma_j} = 0$, $j = 1, \ldots, J$.

To define the Lagrange multipliers, it is useful to introduce the surface gradient operator. Since we will use it acting only on piecewise linear functions defined on polyhedral surfaces, we give a simple definition valid in this case (see [9] for a more general definition):

$$\widetilde{\mathbf{grad}}_\Gamma :\ \mathcal{L}_h(\widetilde{\Gamma}_{\mathrm{I}}) \longrightarrow \mathrm{L}^2_{\mathbf{t}}(\Gamma_{\mathrm{I}})^3 := \left\{\boldsymbol{G} \in \mathrm{L}^2(\Gamma_{\mathrm{I}})^3 :\ \boldsymbol{G} \cdot \boldsymbol{\nu} = 0 \text{ on } \Gamma_{\mathrm{I}}\right\}$$

is defined on each element $T \in \mathcal{T}_h^{\Gamma_{\mathrm{I}}}$ by

$$\left(\widetilde{\mathbf{grad}}_\Gamma\, \widetilde{\eta}_h\right)|_T := \nabla\left(\widetilde{\eta}_h|_T\right),$$

where ∇ denotes the usual gradient of a function of two variables. More precisely, if $\{\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{\nu}\}$ is an orthonormal coordinate system (i.e., $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$ are orthonormal tangent vectors to T) and $(\zeta_1, \zeta_2, \zeta_3)$ denote the coordinates in this system, then

$$\left(\widetilde{\mathbf{grad}}_\Gamma\, \widetilde{\eta}_h\right)|_T := \frac{\partial\left(\widetilde{\eta}_h|_T\right)}{\partial\zeta_1}\,\boldsymbol{e}_1 + \frac{\partial\left(\widetilde{\eta}_h|_T\right)}{\partial\zeta_2}\,\boldsymbol{e}_2.$$

Let us remark that the tangential component of $\widetilde{\mathbf{grad}}_\Gamma\, \widetilde{\eta}_h$ is continuous along all the edges of triangles on Γ_{I}, except eventually for those lying on the boundary $\partial\Sigma_j$ of a cut surface. Of course, the same definition also holds for $\eta_h \in \mathcal{L}_h(\Gamma_{\mathrm{I}}) \subset \mathcal{L}_h(\widetilde{\Gamma}_{\mathrm{I}})$; moreover, in this case, the tangential components are continuous even along the edges on $\partial\Sigma_j$. To help remember this, we write $\mathbf{grad}_\Gamma\, \eta_h$ instead of $\widetilde{\mathbf{grad}}_\Gamma\, \eta_h$, when $\eta_h \in \mathcal{L}_h(\Gamma_{\mathrm{I}})$.

It is straightforward to show that, for all $(\widetilde{\psi}_h, \psi_h^{\mathrm{R}}) \in \mathcal{X}_h$, there holds

$$\widetilde{\psi}_h|_{\Gamma_{\mathrm{I}}} \in \mathcal{L}_h(\widetilde{\Gamma}_{\mathrm{I}}) \quad \text{and} \quad \widetilde{\mathbf{grad}}_\Gamma\left(\widetilde{\psi}_h|_{\Gamma_{\mathrm{I}}}\right) = \boldsymbol{\nu} \times \left(\widetilde{\mathbf{grad}}\,\widetilde{\psi}_h \times \boldsymbol{\nu}\right) \quad \text{on } \widetilde{\Gamma}_{\mathrm{I}},$$
$$\psi_h^{\mathrm{R}}|_{\Gamma_{\mathrm{I}}} \in \mathcal{L}_h(\Gamma_{\mathrm{I}}) \quad \text{and} \quad \mathbf{grad}_\Gamma\left(\psi_h^{\mathrm{R}}|_{\Gamma_{\mathrm{I}}}\right) = \boldsymbol{\nu} \times \left(\mathbf{grad}\,\psi_h^{\mathrm{R}} \times \boldsymbol{\nu}\right) \quad \text{on } \Gamma_{\mathrm{I}}.$$

From now on, for simplicity we write $\widetilde{\mathbf{grad}}_\Gamma\,\widetilde{\psi}_h := \widetilde{\mathbf{grad}}_\Gamma\,(\widetilde{\psi}_h|_{\Gamma_{\rm I}})$ and $\mathbf{grad}_\Gamma\,\psi_h^{\rm R} := \mathbf{grad}_\Gamma\,(\psi_h^{\rm R}|_{\Gamma_{\rm I}})$. Thus, the constraint in the definition of $\mathcal{V}_h(\boldsymbol{T})$ can be equivalently written as follows:

$$\mathbf{grad}_\Gamma\,\psi_h^{\rm R} - \widetilde{\mathbf{grad}}_\Gamma\,\widetilde{\psi}_h = \boldsymbol{\nu}\times(\boldsymbol{T}_{\rm N}\times\boldsymbol{\nu}) \qquad \text{on } \Gamma_{\rm I}.$$

The following lemma will allow us to write an equivalent weak form of this constraint. We recall that

$$I_j := \int_{S_j} \boldsymbol{J}\cdot\boldsymbol{n}_{S_j}, \qquad j=1,\dots,J,$$

was already defined in Remark 2.1 and it corresponds to the total current flux through S_j in the direction of $\boldsymbol{n}_{S_j}$ (see Figure 2.4). Note that, since $\operatorname{div}\boldsymbol{J}=0$ in Ω, the integral in the definition has a weak sense as long as $\boldsymbol{J}\in \mathrm{H}^s(\Omega)^3$ for any $s>0$ (see [8, section III.3.3]). From now on, we assume this with $0<s<\frac{1}{2}$.

LEMMA 5.1. *Let $\widetilde{\psi}_h\in\mathcal{L}_h(\widetilde{\Gamma}_{\rm I})$ be such that*

$$[\![\widetilde{\psi}_h]\!]_{\partial\Sigma_j} = I_j, \qquad j=1,\dots,J. \tag{5.1}$$

Then, $\widetilde{\mathbf{grad}}_\Gamma\,\widetilde{\psi}_h = \boldsymbol{\nu}\times(\boldsymbol{T}_{\rm N}\times\boldsymbol{\nu})$ on $\Gamma_{\rm I}$ if and only if

$$\int_{\Gamma_{\rm I}} \widetilde{\mathbf{grad}}_\Gamma\,\widetilde{\psi}_h\cdot\mathbf{grad}_\Gamma\,\eta_h = \int_{\Gamma_{\rm I}} \boldsymbol{\nu}\times(\boldsymbol{T}_{\rm N}\times\boldsymbol{\nu})\cdot\mathbf{grad}_\Gamma\,\eta_h \qquad \forall\eta_h\in\mathcal{L}_h(\Gamma_{\rm I}). \tag{5.2}$$

Proof. Clearly, for all $\widetilde{\psi}_h\in\mathcal{L}_h(\widetilde{\Gamma}_{\rm I})$, $\widetilde{\mathbf{grad}}_\Gamma\,\widetilde{\psi}_h = \boldsymbol{\nu}\times(\boldsymbol{T}_{\rm N}\times\boldsymbol{\nu})$ on $\Gamma_{\rm I}$ implies (5.2).

Conversely, let $\widetilde{\psi}_h\in\mathcal{L}_h(\widetilde{\Gamma}_{\rm I})$ be such that (5.1) and (5.2) hold true. We are going to prove that there exists $\eta_h\in\mathcal{L}_h(\Gamma_{\rm I})$ such that $\mathbf{grad}_\Gamma\,\eta_h = \widetilde{\mathbf{grad}}_\Gamma\,\widetilde{\psi}_h - \nu\times(\boldsymbol{T}_{\rm N}\times\nu)$, so that we can take it as a test function in (5.2) to conclude the lemma.

Let $\widetilde{\xi}_{\rm L}\in\Theta_h$ be as in Lemma 4.2(ii). Hence,

$$-\widetilde{\mathbf{grad}}_\Gamma\,\widetilde{\xi}_{\rm L} = -\boldsymbol{\nu}\times\big(\widetilde{\mathbf{grad}}\,\widetilde{\xi}_{\rm L}\times\boldsymbol{\nu}\big) = \boldsymbol{\nu}\times\big(\boldsymbol{T}_{\rm N}\times\boldsymbol{\nu}\big) \qquad \text{on } \Gamma_{\rm I},$$

and therefore

$$[\![\widetilde{\xi}_{\rm L}]\!]_{\partial\Sigma_j} = -\int_{\gamma_j}\widetilde{\mathbf{grad}}_\Gamma\,\widetilde{\xi}_{\rm L}\cdot\boldsymbol{t}_j = \int_{\gamma_j}\boldsymbol{\nu}\times\big(\boldsymbol{T}_{\rm N}\times\boldsymbol{\nu}\big)\cdot\boldsymbol{t}_j = \int_{\gamma_j}\boldsymbol{T}_{\rm N}\cdot\boldsymbol{t}_j.$$

On the other hand, from (5.1), because of (2.10) and Stokes' theorem we have

$$[\![\widetilde{\psi}_h]\!]_{\partial\Sigma_j} = I_j = \int_{S_j}\boldsymbol{J}\cdot\boldsymbol{n}_{S_j} = \int_{S_j}\mathbf{curl}\,\boldsymbol{T}\cdot\boldsymbol{n}_{S_j} = \int_{\gamma_j}\boldsymbol{T}\cdot\boldsymbol{t}_j.$$

The integrals above have a precise weak sense since we are assuming $\boldsymbol{J}\in\mathrm{H}^s(\Omega)^3$, $0<s<\frac{1}{2}$, and consequently $\mathbf{curl}\,\boldsymbol{T}=\boldsymbol{J}\in\mathrm{H}^s(\Omega)^3$ as well (see [2]). The curve γ_j joining $\overline{\Sigma}_j^-$ with $\overline{\Sigma}_j^+$ can always be chosen formed by edges of the mesh. In such a case, from the definition of the Nédélec interpolant,

$$\int_{\gamma_j}\boldsymbol{T}\cdot\boldsymbol{t}_j = \int_{\gamma_j}\boldsymbol{T}_{\rm N}\cdot\boldsymbol{t}_j, \qquad j=1,\dots,J.$$

Now, let $\eta_h := \widetilde{\psi}_h - \widetilde{\xi}_{\rm L}|_{\widetilde{\Gamma}_{\rm I}}\in\mathcal{L}_h(\widetilde{\Gamma}_{\rm I})$. From the equations above,

$$[\![\eta_h]\!]_{\partial\Sigma_j} = [\![\widetilde{\psi}_h]\!]_{\partial\Sigma_j} - [\![\widetilde{\xi}_{\rm L}]\!]_{\partial\Sigma_j} = 0, \qquad j=1,\dots,J,$$

and consequently $\eta_h \in \mathcal{L}_h(\Gamma_{\rm I})$. Moreover,

$$\mathbf{grad}_\Gamma\, \eta_h = \widetilde{\mathbf{grad}}_\Gamma\, \widetilde{\psi}_h - \boldsymbol{\nu} \times (\boldsymbol{T}_{\rm N} \times \boldsymbol{\nu}) \qquad \text{on } \Gamma_{\rm I}.$$

Thus, using η_h in (5.2), we obtain

$$\int_{\Gamma_{\rm I}} \left| \widetilde{\mathbf{grad}}_\Gamma\, \widetilde{\psi}_h - \boldsymbol{\nu} \times (\boldsymbol{T}_{\rm N} \times \boldsymbol{\nu}) \right|^2 = 0,$$

which yields

$$\widetilde{\mathbf{grad}}_\Gamma\, \widetilde{\psi}_h = \boldsymbol{\nu} \times (\boldsymbol{T}_{\rm N} \times \boldsymbol{\nu}) \qquad \text{on } \Gamma_{\rm I},$$

and we conclude the proof. ∎

Next, we impose the constraint in the definition of $\mathcal{V}_h(\boldsymbol{T})$ by using the previous lemma to rewrite Problem DP as follows.

PROBLEM DLP. *Find* $\big((\widetilde{\phi}_h, \phi_h^{\rm R}), \lambda_h\big) \in \mathcal{X}_h \times \mathcal{L}_h(\Gamma_{\rm I})/\mathbb{R}$ *such that*

$$[\![\widetilde{\phi}_h]\!]_{\Sigma_j} = I_j, \qquad j = 1, \dots, J, \tag{5.3}$$

$$\begin{aligned} &\int_{\Omega_{\rm M}} \mu\, \widetilde{\mathbf{grad}}\, \widetilde{\phi}_h \cdot \widetilde{\mathbf{grad}}\, \widetilde{\psi}_h + \int_{\Omega_{\rm R}} \mu_0\, \mathbf{grad}\, \phi_h^{\rm R} \cdot \mathbf{grad}\, \psi_h^{\rm R} \\ &\quad + \int_{\Gamma_{\rm I}} \big(\mathbf{grad}_\Gamma\, \psi_h^{\rm R} - \widetilde{\mathbf{grad}}_\Gamma\, \widetilde{\psi}_h\big) \cdot \mathbf{grad}_\Gamma\, \lambda_h \\ &\qquad = \int_\Gamma \mu_0 \boldsymbol{T}_{\rm RT} \cdot \boldsymbol{n}\, \psi_h^{\rm R} - \int_{\Gamma_{\rm I}} \mu_0 \boldsymbol{T}_{\rm RT} \cdot \boldsymbol{\nu}\, \psi_h^{\rm R} - \big\langle g, \psi_h^{\rm R} \big\rangle_\Gamma \qquad \forall \big(\widetilde{\psi}_h, \psi_h^{\rm R}\big) \in \mathcal{X}_h, \end{aligned} \tag{5.4}$$

$$\begin{aligned} &\int_{\Gamma_{\rm I}} \big(\mathbf{grad}_\Gamma\, \phi_h^{\rm R} - \widetilde{\mathbf{grad}}_\Gamma\, \widetilde{\phi}_h\big) \cdot \mathbf{grad}_\Gamma\, \eta_h \\ &\qquad = \int_{\Gamma_{\rm I}} \boldsymbol{\nu} \times (\boldsymbol{T}_{\rm N}|_{\Gamma_{\rm I}} \times \boldsymbol{\nu}) \cdot \mathbf{grad}_\Gamma\, \eta_h \qquad \forall \eta_h \in \mathcal{L}_h(\Gamma_{\rm I})/\mathbb{R}. \end{aligned} \tag{5.5}$$

THEOREM 5.2. *Problem* DLP *has a unique solution.*

Proof. Problem DLP reduces to a linear system with the same number of equations and unknowns, so that it is enough to prove that, for vanishing data, this problem attains only the null solution. Let $\big((\widetilde{\phi}_o, \phi_o^{\rm R}), \lambda_o\big) \in \mathcal{X}_h \times \mathcal{L}_h(\Gamma_{\rm I})/\mathbb{R}$ be a solution of (5.3)–(5.5) with $I_j = 0$, $j = 1, \dots, J$, $\boldsymbol{T} = \boldsymbol{0}$, and $g = 0$.

In such a case $[\![\widetilde{\phi}_o]\!]_{\Sigma_j} = 0$, $j = 1, \dots, J$, because of (5.3). Thus, if we define $\eta_h := \big(\phi_o^{\rm R} - \widetilde{\phi}_o\big)|_{\widetilde{\Gamma}_{\rm I}}$, then $[\![\eta_h]\!]_{\partial\Sigma_j} = 0$, $j = 1, \dots, J$, and hence $\eta_h \in \mathcal{L}_h(\Gamma_{\rm I})/\mathbb{R}$. Then, by testing (5.5) with η_h, we obtain

$$\mathbf{grad}_\Gamma\, \phi_o^{\rm R} - \widetilde{\mathbf{grad}}_\Gamma\, \widetilde{\phi}_o = \boldsymbol{0}.$$

Therefore, by testing (5.4) with $\big(\widetilde{\phi}_o, \phi_o^{\rm R}\big)$, we have

$$\int_{\Omega_{\rm M}} \mu \left| \widetilde{\mathbf{grad}}\, \widetilde{\phi}_o \right|^2 + \int_{\Omega_{\rm R}} \mu_0 \left| \mathbf{grad}\, \phi_o^{\rm R} \right|^2 = 0,$$

which implies that $\widetilde{\phi}_o = 0$ in $\Theta_h/\mathbb{R}$ and $\phi_o^{\rm R} = 0$ in $\mathcal{L}_h(\Omega_{\rm R})/\mathbb{R}$.

Thus, there only remains to prove that $\lambda_o = 0$. For this purpose, let ψ_h^{R} be any function in $\mathcal{L}_h(\Omega_{\mathrm{R}})/\mathbb{R}$ such that $\psi_h^{\mathrm{R}}|_{\Gamma_{\mathrm{I}}} = \lambda_o$, and let $\widetilde{\psi}_h = 0$ in $\Theta_h/\mathbb{R}$. Then $\mathbf{grad}_\Gamma\, \psi_h^{\mathrm{R}} - \widetilde{\mathbf{grad}}_\Gamma\, \widetilde{\psi}_h = \mathbf{grad}_\Gamma\, \lambda_o$ on Γ_{I}, and by testing (5.4) with $\big(\widetilde{\psi}_h, \psi_h^{\mathrm{R}}\big)$ we obtain

$$\int_{\Gamma_{\mathrm{I}}} |\mathbf{grad}_\Gamma\, \lambda_o|^2 = 0.$$

Therefore, $\lambda_o = 0$ in $\mathcal{L}_h(\Gamma_{\mathrm{I}})/\mathbb{R}$, and we conclude the proof. ∎

The final result of this section is that Problems DP and DLP are equivalent.

THEOREM 5.3. *If* $\big((\widetilde{\phi}_h, \phi_h^{\mathrm{R}}), \lambda_h\big)$ *is the solution of Problem* DLP*, then* $\big(\widetilde{\phi}_h, \phi_h^{\mathrm{R}}\big)$ *is the solution of Problem* DP*.*

Conversely, if $\big(\widetilde{\phi}_h, \phi_h^{\mathrm{R}}\big)$ *is the solution of Problem* DP*, then there exists* λ_h *such that* $\big((\widetilde{\phi}_h, \phi_h^{\mathrm{R}}), \lambda_h\big)$ *is the solution of Problem* DLP*.*

Proof. Let $\big((\widetilde{\phi}_h, \phi_h^{\mathrm{R}}), \lambda_h\big)$ be the solution of Problem DLP. Because of (5.3), (5.5), and Lemma 5.1, we have that $\big(\widetilde{\phi}_h, \phi_h^{\mathrm{R}}\big) \in \mathcal{V}_h(\boldsymbol{T})$. Now, let $\big(\widetilde{\psi}_h, \psi_h^{\mathrm{R}}\big) \in \mathcal{V}_h(\mathbf{0})$. Since $\mathbf{grad}_\Gamma\, \psi_h^{\mathrm{R}} - \widetilde{\mathbf{grad}}_\Gamma\, \widetilde{\psi}_h = \mathbf{0}$, by testing (5.4) with such $\big(\widetilde{\psi}_h, \psi_h^{\mathrm{R}}\big)$, we obtain (4.3). Thus, $\big(\widetilde{\phi}_h, \phi_h^{\mathrm{R}}\big)$ is the solution of Problem DP.

Conversely, let $\big(\widetilde{\phi}_h, \phi_h^{\mathrm{R}}\big)$ be the solution of Problem DP and $\big((\widetilde{\phi}'_h, \phi_h^{\mathrm{R}\prime})), \lambda_h\big)$ the solution of Problem DLP. We have just proved that $\big(\widetilde{\phi}'_h, \phi_h^{\mathrm{R}\prime}\big)$ is also a solution of Problem DP. Since this solution is unique because of Theorem 4.4, then $\big(\widetilde{\phi}'_h, \phi_h^{\mathrm{R}\prime}\big) = \big(\widetilde{\phi}_h, \phi_h^{\mathrm{R}}\big)$, and we conclude the proof. ∎

Remark 5.1. We have restricted our analysis to the magnetostatic problem with a boundary condition $\mu \boldsymbol{H} \cdot \boldsymbol{n} = g$ on the whole Γ. However, as stated above, the method can deal with a condition $\boldsymbol{H} \times \boldsymbol{n} = \mathbf{0}$ on a piece Γ_{D} of the boundary, as well. In such a case, an additional Lagrange multiplier must be used to impose

$$\mathbf{grad}\, \phi_h^{\mathrm{R}} \times \boldsymbol{n} - \boldsymbol{T}_{\mathrm{N}} \times \boldsymbol{n} = \mathbf{0} \qquad \text{on } \Gamma_{\mathrm{D}}.$$

Remark 5.2. We have also restricted the analysis of this section to domains Ω_{M} with a connected boundary, and consequently to Ω_{R} a connected set. However, it also extends with minor modifications to the case $\Gamma_{\mathrm{I}} = \partial\Omega_{\mathrm{M}}$ disconnected (and hence Ω_{R} disconnected too). We have already pointed out in Remark 4.1 how to deal with a disconnected set Ω_{R}. Regarding Γ_{I}, if $\Gamma_1, \ldots, \Gamma_K$ are its connected components, then the constraint in the definition of $\mathcal{V}_h(\boldsymbol{T})$, $\mathbf{grad}\, \psi_h^{\mathrm{R}} \times \boldsymbol{\nu} - \widetilde{\mathbf{grad}}\, \widetilde{\psi}_h \times \boldsymbol{\nu} = \boldsymbol{T}_{\mathrm{N}} \times \boldsymbol{\nu}$ on Γ_{I}, must be imposed by means of a Lagrange multiplier on each connected component. Therefore, the functional space for λ_h and η_h in Problem DLP must be $\Pi_{k=1}^K [\mathcal{L}_h(\Gamma_k)/\mathbb{R}]$.

6. Numerical results. In this section we present some numerical results obtained with a MATLAB code which implements the method described above.

Notice that the implementation requires the computation of $\boldsymbol{T}_{\mathrm{N}}$ and $\boldsymbol{T}_{\mathrm{RT}}$ on Γ_{I}. Since we have assumed that $\Omega_{\mathrm{M}} \cap \Omega_{\mathrm{J}} = \emptyset$, the integral involved in the definition of $\boldsymbol{T}$ does not include singular terms on this boundary. However, to compute the magnetic field $\boldsymbol{H} = \boldsymbol{T} - \mathbf{grad}\, \phi^{\mathrm{R}}$ in the source region, we have to deal with singular integrals. There exists an important list of papers devoted to the numerical evaluation of the Biot–Savart integral; some of these techniques are based on assumptions on the shape of the source region or on the current density $\boldsymbol{J}$ (see, for instance, [13, 29, 30] and the references therein), while other works are based on the transformation of the volume integral into line integrals [28]. Specifically, we have used the techniques developed in [13] in the numerical test in section 6.2.

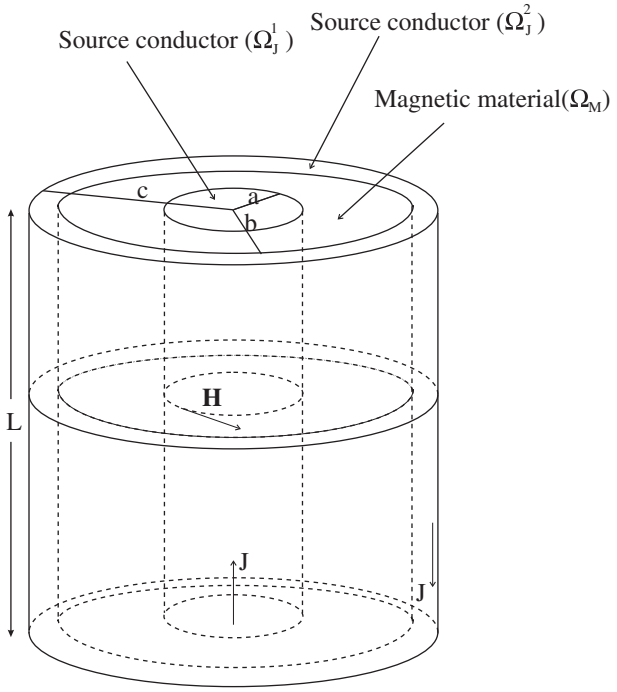


FIG. 6.1. *Geometric data for the verification test.*

6.1. Verification of the FEM. First, in order to test the convergence properties settled in the previous sections, we report the results obtained by applying the method to a simple problem with a known analytical solution.

Consider the magnetostatic problem in a cylindrical domain Ω, which consists of a current source region with two concentric connected components, Ω_J^1 and Ω_J^2, separated by a magnetic material Ω_M, as shown in Figure 6.1. These sets are assumed to be sections of height L of respective infinite cylinders. This assumption allows us to obtain the analytical solution of the problem.

The magnetic permeability of $\Omega_R := \Omega_J^1 \cup \Omega_J^2$ is taken equal to that of vacuum ($\mu_0 = 4\pi\,10^{-7}\,\mathrm{Hm}^{-1}$), whereas that of Ω_M is taken constant and equal to $10^4\mu_0$. The conductors Ω_J^1 and Ω_J^2 carry uniformly distributed currents $\boldsymbol{J}$ of the same intensity $I = 70000\,\mathrm{A}$, but in opposite directions (see Figure 6.1).

Consider a cylindrical coordinate system (r, θ, z), with the z-axis coinciding with the common axis of the cylinders, and denote by $\boldsymbol{e}_r$, $\boldsymbol{e}_\theta$, and $\boldsymbol{e}_z$ the unit vectors in the corresponding coordinate directions. Simple calculus (see [14, Example 5.1, p. 260]) shows that

$$\boldsymbol{H}(r,\theta,z) = \begin{cases} \dfrac{rI}{2\pi a^2}\,\boldsymbol{e}_\theta, & 0 \le r \le a, \\ \dfrac{I}{2\pi r}\,\boldsymbol{e}_\theta, & a \le r \le b, \\ \left\{ \dfrac{rI}{2\pi\,(c^2-b^2)} + \dfrac{1}{r}\left[\dfrac{I}{2\pi} - \dfrac{b^2 I}{2\pi\,(c^2-b^2)} \right] \right\} \boldsymbol{e}_\theta, & b < r < c. \end{cases}$$

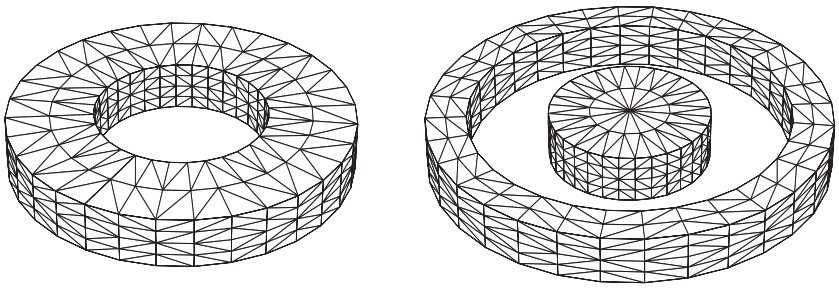


FIG. 6.2. *Initial mesh $\mathcal{T}_h$ for the verification test: $\mathcal{T}_h^{\Omega_M}$ (left) and $\mathcal{T}_h^{\Omega_R}$ (right).*

TABLE 6.1
Verification test. $L^2(\Omega)$*-norms of computed solution and error for the magnetic field.*

Mesh size	d.o.f.	$\|\boldsymbol{H}_h\|_{0,\Omega}$	$\|\boldsymbol{H}-\boldsymbol{H}_h\|_{0,\Omega}$	Relative error (%)
h	1220	23746.15	2757.19	11.61
$h/2$	7254	23704.27	1390.96	5.87
$h/3$	21944	23696.28	930.31	3.93
$h/4$	49130	23693.46	698.50	2.95
$h/5$	92652	23692.13	559.18	2.36

Taking into account the expression of the gradient operator of a scalar field in cylindrical coordinates, we have that

$$\boldsymbol{H} = -\,\widetilde{\mathbf{grad}}\,\widetilde{\phi} \qquad \text{in } \Omega_M$$

with

$$\widetilde{\phi}\,(r,\theta,z) = -\frac{I}{2\pi}\,\theta, \qquad r \in (a,b)\,, \quad \theta \in [0,2\pi]\,, \quad z \in \mathbb{R}.$$

Notice that this scalar potential has a jump of magnitude I across the cut surface Σ on the half-plane $\theta = 0$.

In this example, we have taken $\boldsymbol{T} = \boldsymbol{H}$ in Ω_R, so that ϕ^R is constant, and the boundary condition

$$\mu\boldsymbol{H}\cdot\boldsymbol{n} = 0 \qquad \text{on } \partial\Omega,$$

which is fulfilled by the exact solution on the whole boundary.

The geometrical data has been taken as follows: $a = 0.5\,\text{m}$, $b = 1\,\text{m}$, $c = 1.25\,\text{m}$, and $L = 0.5\,\text{m}$.

The numerical method has been used with a sequence of meshes obtained by uniformly refining the mesh $\mathcal{T}_h$ shown in Figure 6.2. We denote by h, $h/2$, $h/3$, etc. the corresponding mesh sizes.

The computed solutions have been compared with the analytical one to determine the orders of convergence. Table 6.1 shows the $L^2(\Omega)$-norms of the computed magnetic field $\boldsymbol{H}_h$ and the corresponding errors. The total number of degrees of freedom (d.o.f.) for the potentials (i.e., total number of vertices) is also included for each mesh.

Figure 6.3 shows a log-log plot of the relative error of the computed magnetic field in $L^2(\Omega)$-norm versus the number of d.o.f. of each mesh.

A linear dependence on the mesh size can be clearly observed from Figure 6.3. Let us remark that this norm coincides with the $\mathcal{X}$-norm of the potentials, for which

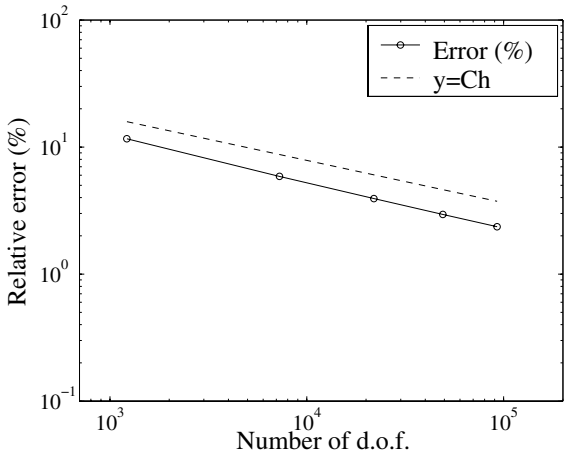


FIG. 6.3. *Verification test.* $\|\boldsymbol{H} - \boldsymbol{H}_h\|_{0,\Omega} / \|\boldsymbol{H}_h\|_{0,\Omega}$ *versus number of d.o.f. (log-log scale).*

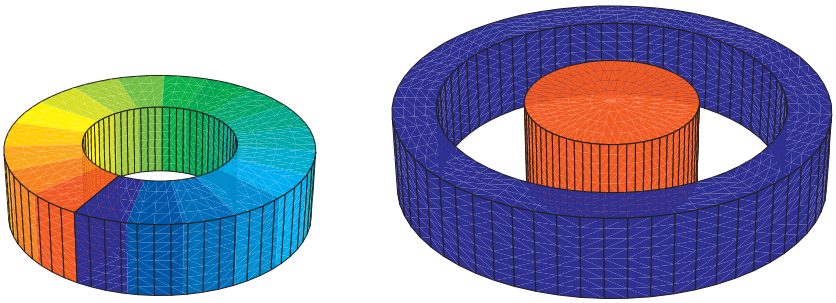


FIG. 6.4. *Verification test. Total and reduced potential in* Ω_{M} *(left) and* Ω_{J} *(right).*

Theorem 4.4 provides error estimates. The observed linear dependence agrees with the theoretically predicted order of convergence, even though this problem does not fully satisfy the geometrical assumptions of the previous sections.

Finally, Figure 6.4 shows the computed total and reduced potentials in the sets Ω_{M} and Ω_{J}, respectively.

6.2. The magnetostatic problem in an electromagnet. The second test corresponds to an axisymmetric electromagnet. This example has been taken from [12], where the magnetostatic problem is solved also in terms of total and reduced scalar potentials by using an adaptive finite element code.

The problem consists of a ferromagnetic cylindrical core (iron) surrounded by a toroidal coil (copper) with a rectangular cross section. A two-dimensional section of the geometry and the dimensions are shown in Figure 6.5.

A stationary uniform current with intensity $I = 1\,\mathrm{A}$ flows through the coil. The magnetic permeabilities have been taken as follows: $\mu_0 = 4\pi\,10^{-7}\,\mathrm{Hm}^{-1}$ for the air and the copper coil (Ω_{R}) and $10^4\mu_0$ for the iron core (Ω_{M}). We have used a homogeneous boundary condition $\mu\boldsymbol{H}\cdot\boldsymbol{n} = 0$ on the whole boundary. The vector potential $\boldsymbol{T}$ has been computed as proposed in [13].

Figure 6.6 shows the coarser meshes used to solve the problem.

Figure 6.7 shows the total and reduced scalar potentials on a plane containing the symmetry axis.

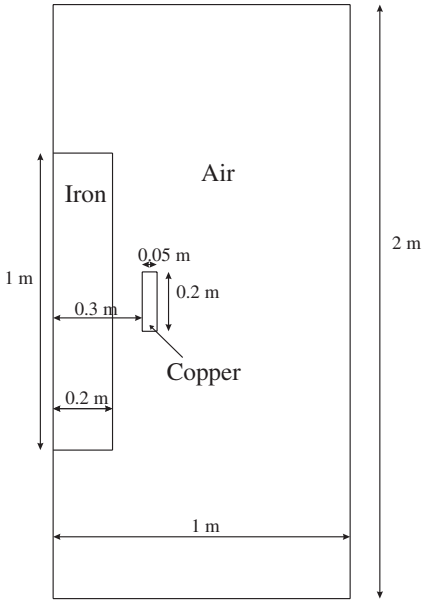


FIG. 6.5. *Electromagnet. Axial section of the domain.*

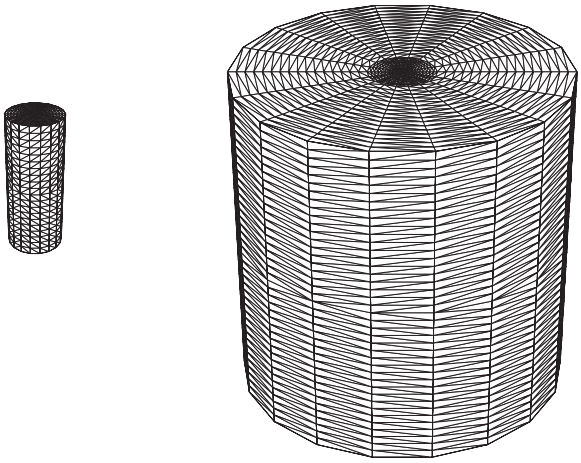


FIG. 6.6. *Electromagnet. Meshes in* Ω_{M} *(left) and* Ω_{R} *(right).*

Figure 6.8 shows the intensity of the magnetic induction field, $|\boldsymbol{B}| = |\mu \boldsymbol{H}|$, on the same plane.

To validate the proposed method, we have also solved this magnetostatic problem by means of the commercial code *FLUX2D* (see [24]), which takes advantage of the axial symmetry of the problem. We have computed the magnetic energy (see, for instance, [14])

$$W = \frac{1}{2} \int_D \boldsymbol{B} \cdot \boldsymbol{H} = \frac{1}{2} \int_D \mu \, |\boldsymbol{H}|^2 \,,$$

where D represents a generic domain occupied by any of the three media: air, copper, or iron.

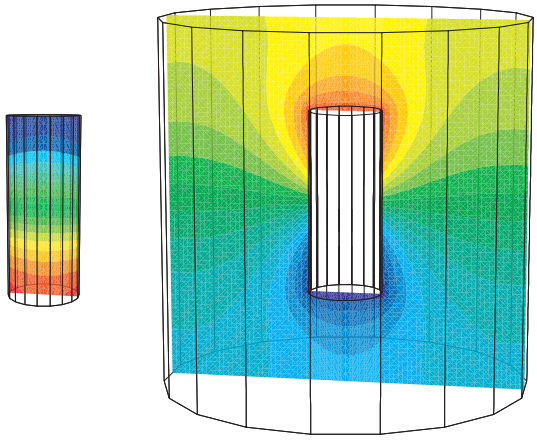


FIG. 6.7. *Electromagnet. Total potential* $\widetilde{\phi}_h$ *in* Ω_{M} *(left) and reduced potential* ϕ_h^{R} *in* Ω_{R} *(right).*

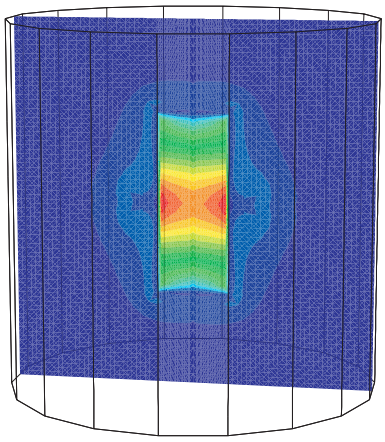


FIG. 6.8. *Electromagnet. Intensity of the magnetic induction field,* $|\boldsymbol{B}|$ *in* Ω.

TABLE 6.2
Electromagnet. Computed magnetic energies and relative errors.

	Medium	Air	Iron	Copper
Mesh 1	Magnetic Energy (J)	9.13e-07	4.93e-10	2.60e-08
19933 d.o.f.	Relative error (%)	0.43	4.23	27.07
Mesh 2	Magnetic Energy (J)	9.10e-07	4.83e-10	3.37e-08
37159 d.o.f.	Relative error (%)	0.15	2.11	6.65
FLUX2D	Magnetic Energy (J)	9.09e-07	4.73e-10	3.61e-08

We report in Table 6.2 the values of the magnetic energy computed with our method on two different meshes and the corresponding values computed with *FLUX2D* on a highly refined two-dimensional mesh. We include the total numbers of d.o.f. for each three-dimensional mesh and the relative errors obtained by using the values computed with *FLUX2D* as "exact."

Table 6.2 shows that the error of the computed magnetic energy in each medium goes to zero as the mesh is refined. On the other hand, the agreement between the values obtained with the proposed method on the coarser mesh and *FLUX2D* is excellent in the magnetic material (which is typically the domain of interest), very good in the air, but not so good in the conducting domain. A similar behavior has been already reported in [11].

Acknowledgment. The numerical computations have been performed on the system HPC320 at "Centro de Supercomputación de Galicia" (CESGA).

REFERENCES

[1] P. Alotto and I. Perugia, *A field-based finite element method for magnetostatics derived from an error minimization approach*, Internat. J. Numer. Methods Engrg., 49 (2000), pp. 573–598.

[2] C. Amrouche, C. Bernardi, M. Dauge, and V. Girault, *Vector potentials in three-dimensional nonsmooth domains*, Math. Methods Appl. Sci., 21 (1998), pp. 823–864.

[3] M. Ayoub, F. Roy, F. Bouillault, and A. Razek, *Numerical modelling of 3D magnetostatic saturated structures with a hybrid FEM-BEM technique*, IEEE Trans. Magnetics, 28 (1992), pp. 1052–1055.

[4] A. Bermúdez, R. Rodríguez, and P. Salgado, *A finite element method with Lagrange multipliers for low-frequency harmonic Maxwell equations*, SIAM J. Numer. Anal., 40 (2002), pp. 1823–1849.

[5] O. Biro, K. Preis, and K. R. Richter, *On the use of the magnetic vector potential in the nodal and edge finite element analysis of 3D magnetostatic problems*, IEEE Trans. Magnetics, 32 (1996), pp. 651–653.

[6] O. Biro, K. Preis, G. Vrisk, K. R. Richter, and I. Ticar, *Computation of 3D magnetostatic fields using a reduced scalar potential*, IEEE Trans. Magnetics, 29 (1993), pp. 1329–1332.

[7] A. Bossavit, *Computational Electromagnetism: Variational Formulations, Complementarity, Edge Elements*, Academic Press, New York, 1998.

[8] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer-Verlag, New York, 1991.

[9] A. Buffa, M. Costabel, and D. Sheen, *On traces for* $\mathrm{H}(\mathbf{curl}, \Omega)$ *in Lipschitz domains*, J. Math. Anal. Appl., 276 (2002), pp. 845–867.

[10] P. Fernandes and I. Perugia, *Vector potential formulation of magnetostatic and modelling of permanent magnets*, IMA J. Appl. Math., 66 (2001), pp. 293–318.

[11] M. Fontana, *Integration Methods for the Calculation of the Magnetostatic Field Due to Coils*, Preprint 2001-7, Chalmers Finite Element Center, University of Chalmers, Göteborg, 2001.

[12] M. Fontana, *Adaptive Finite Element Computation of 3D Magnetostatic Problems in Potential Formulation*, Preprint 2001-8, Chalmers Finite Element Center, University of Chalmers, Göteborg, 2001.

[13] L. K. Forbes, S. Crozier, and D. Doddrell, *Rapid computation of static fields produced by thick circular solenoids*, IEEE Trans. Magnetics, 33 (1997), pp. 4405–4410.

[14] C. T. A. Jhonk, *Engineering Electromagnetic Fields and Waves*, John Wiley & Sons, New York, 1988.

[15] F. Kikuchi, *Mixed formulations for finite element analysis of magnetostatic and electrostatic problems*, Japan J. Appl. Math., 6 (1989), pp. 209–221.

[16] Ch. Magele, H. Stögner, and K. Preis, *Comparison of different finite element formulations for 3D magnetostatic problems*, IEEE Trans. Magnetics, 24 (1988), pp. 31–34.

[17] I. D. Mayergoyz, M. V. C. Chari, and J. D'Angelo, *A new scalar potential formulation for three dimensional magnetostatic problems*, IEEE Trans. Magnetics, MAG-23 (1987), pp. 3889–3894.

[18] A. J. Meir and P. G. Schmidt, *Variational methods for stationary MHD flow under natural interface conditions*, Nonlinear Anal., 26 (1996), pp. 659–689.

[19] P. Monk, *Finite Element Methods for Maxwell's Equations*, Oxford University Press, New York, 2003.

[20] I. Perugia, *A mixed formulation for 3D magnetostatic problems: Theoretical analysis and face-edge finite element approximation*, Numer. Math., 84 (1999), pp. 305–326.

[21] K. Preis, I. Bardi, O. Biro, C. Magele, W. Rhenhart, K. R. Ritcher, and G. Vrisk, *Numerical analysis of 3D magnetostatic fields*, IEEE Trans. Magnetics, 27 (1991), pp. 3798–3803.

[22] W. M. Rucker and K. R. Richter, *Three-dimensional magnetostatic field calculation using boundary element method*, IEEE Trans. Magnetics, 24 (1988), pp. 23–26.
[23] W. M. Rucker, Ch. Magele, E. Schlemmer, and K. R. Richter, *Boundary element analysis of* 3*-D magnetostatic problems using scalar potentials*, IEEE Trans. Magnetics, 28 (1992), pp. 1099–1102
[24] *FLUX*2*D, Electromagnetic* 2*D Software, Version* 8.1, provided by Cedrat, France.
[25] J. Simkin and C. W. Trowbridge, *Use of the total scalar potential in the numerical solution of field problems in electromagnetics*, Internat. J. Numer. Methods Engrg., 14 (1979), pp. 423–440.
[26] J. Simkin and C. W. Trowbridge, *Three-dimensional nonlinear electromagnetic field computations, using scalar potentials*, IEE Proc. B, 127 (1980), pp. 368–374.
[27] L. R. Scott and S. Zhang, *Finite element interpolation of nonsmooth functions satisfying boundary conditions*, Math. Comp., 54 (1990), pp. 483–493.
[28] J.-C. Suh, *The evaluation of the Biot–Savart integral*, J. Engrg. Math., 37 (2000), pp. 375–395.
[29] L. K. Urankar, *Vector potential and magnetic field of current-carrying finite arc segment in analytical form, part* III: *Exact computation for rectangular cross section*, IEEE Trans. Magnetics, MAG-18 (1982), pp. 1860–1867.
[30] L. K. Urankar, *Vector potential and magnetic field of current-carrying finite arc segment in analytical form, part* IV: *General three-dimensional current density*, IEEE Trans. Magnetics, MAG-20 (1984), pp. 2145–2150.
[31] A. Vourdas, *Magnetostatics with scalar potentials in multiply connected regions*, IEEE Proc. A, 126 (1989), pp. 49–54.

SIAM J. NUMER. ANAL.
Vol. 46, No. 3, pp. 1364–1398
© 2008 Society for Industrial and Applied Mathematics

ERROR ESTIMATES FOR THE RUNGE–KUTTA DISCONTINUOUS GALERKIN METHOD FOR THE TRANSPORT EQUATION WITH DISCONTINUOUS INITIAL DATA*

BERNARDO COCKBURN† AND JOHNNY GUZMÁN†

Abstract. We study the approximation of nonsmooth solutions of the transport equation in one space dimension by approximations given by a Runge–Kutta discontinuous Galerkin method of order two. We take an initial datum, which has compact support and is smooth except at a discontinuity, and show that, if the ratio of the time step size to the grid size is less than 1/3, the error at the time T in the $L^2(\mathbb{R}\setminus\mathcal{R}_T)$-*norm* is the optimal order two when $\mathcal{R}_T$ is a region of size $O(T^{1/2}\,h^{1/2}\,\log 1/h)$ to the right of the discontinuity and of size $O(T^{1/3}\,h^{2/3}\,\log 1/h)$ to the left. Numerical experiments validating these results are presented.

Key words. discontinuous Galerkin methods, error estimates, hyperbolic problems

AMS subject classification. 65N30

DOI. 10.1137/060668936

1. Introduction.

In this paper, we present the first error estimates for the Runge–Kutta discontinuous Galerkin (RKDG) method for the transport equation with *discontinuous* initial data. These results are obtained for a formally second-order accurate RKDG method applied to the model problem

$$U_t + U_x = 0 \qquad \text{in } \mathbb{R}\times(0,T), \tag{1.1a}$$

$$U(\cdot,0) = U_0(\cdot) \quad \text{on } \mathbb{R}, \tag{1.1b}$$

where the initial condition U_0 has compact support; it has a discontinuity at $x = 0$ and is smooth everywhere else. Roughly speaking, we show that the quality of the approximation at time T is of second order in the size of the mesh h *outside* a region of size $O(T^{1/2}\,h^{1/2}\,\log 1/h)$ to the right of the discontinuity and of size $O(T^{1/3}\,h^{2/3}\,\log 1/h)$ to the left. An illustration of this result can be seen in Figure 1.1.

The RKDG method was introduced by Cockburn, Shu, and co-workers in a series of papers [10, 9, 8, 6, 11]; see also the monographs [4, 5] and the review [12]. Most a priori and a posteriori error analyzes of discontinuous Galerkin (DG) methods for hyperbolic problems have been carried out either for the semidiscrete version of the method or for DG methods using space-time elements; see [7], where the development of the DG methods up to the end of past century is described. To the knowledge of the authors, the only a priori error analysis for the RKDG method is due to Shu and Zhang [19], who proved, among other things, that the same method considered here (but applied to nonlinear scalar conservation laws in several space dimensions) converges with order two in the $L^\infty(0,T;L^2(\mathbb{R}^d))$-norm, provided that the solution is *smooth*. In this paper, we continue this effort to understand the RKDG method and analyze it

*Received by the editors September 5, 2006; accepted for publication (in revised form) October 3, 2007; published electronically March 26, 2008.
http://www.siam.org/journals/sinum/46-3/66893.html

†School of Mathematics, University of Minnesota, Vincent Hall, Minneapolis, MN 55455 (cockburn@math.umn.edu, guzma033@.umn.edu). The first author's research was supported in part by National Science Foundation grant DMS-0411254. The second author's research was supported by a National Science Foundation Mathematical Science PostDoctoral Research Fellowship DMS-0503050.

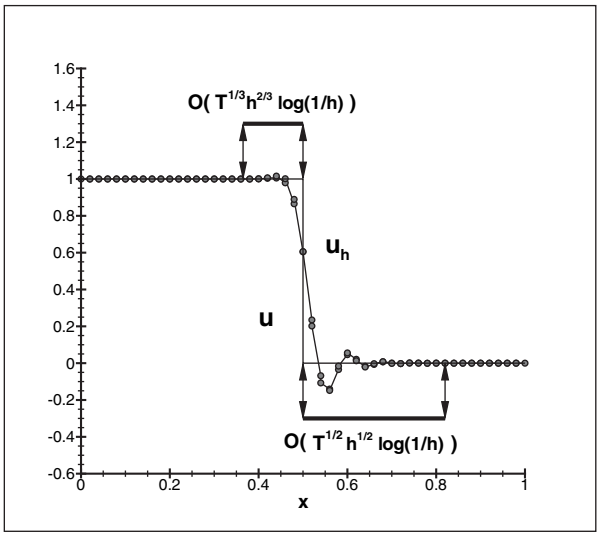


FIG. 1.1. *The two parts of the region $\mathcal{R}_T$ containing the numerical layer of the RKDG method: to the left (top) and to the right (bottom) of the discontinuity of the exact solution u. The approximate solution u_h was obtained by using $h = 1/100$ and $k/h = 0.3$. The time T is* 0.5.

in the case of solutions that have *discontinuities*. As a stepping stone towards the goal of solving the much more complicated case of nonlinear scalar conservation laws in several space dimensions, we consider here the simpler model problem (1.1) and find, for each time T, the region $\mathcal{R}_T$ around the discontinuity of the exact solution $U(\cdot, T)$ such that the approximate solution $u_h(\cdot, T)$ given by the RKDG method converges with the optimal order of two in the $L^2(\mathbb{R} \setminus \mathcal{R}_T)$-norm.

To do that, we use an approach which is a modification of the classical L^2-argument to obtain error estimates in the $L^\infty(0, T; L^2(\mathbb{R}^d))$-norm; see, for example, subsection 2.7 of [5], where the semidiscrete case is treated. The modification has three main features. The first is the use of the decomposition of the error of the approximation given by the RKDG method proposed by Shu and Zhang [19]. The second is the use of special projections that allow us to obtain the full order of convergence of the approximation; see [3]. In our technique, without these projections, the order of convergence is reduced by 1/2. The third is the introduction of suitably chosen *weights* thanks to which we can localize the estimates and make the difference between the region to the left of the discontinuity and that to the right of it.

Similar weights were originally used by Johnson, Nävert, and Pitkäranta [16] to prove local L^2 error estimates for a singularly perturbed reaction-convection-diffusion problem approximated by the streamline diffusion (SD) method; see [17] for L^∞ results. Recently, Guzmán [13] proved similar results for a DG method. Moreover, if one approximates (1.1) with either the standard SD or the DG method and linear space-time elements, one can show by using techniques in [16, 13] that the numerical layer resulting from discontinuous initial data is contained in a region whose size is at most $O(\log(1/h)h^{1/2})$ from either side of the discontinuity. In this article we show that, in one side of the discontinuity, the size of the numerical layer can be reduced to $O(\log(1/h)h^{2/3})$ for the RKDG method. We accomplish this by taking advantage of the monotonicity of one of the two weight functions that we use; see Theorem 4.11 below.

Results of this type were obtained many years ago for finite difference methods for the model problem (1.1) with discontinuous initial data; see [1, 14, 2, 15]. Indeed,

by using Fourier techniques, the size of the numerical layer to the left of the discontinuity was shown to be different to the size of the numerical layer to its right for some schemes; see (1.13) in [15]. In particular, the results concerning the size of the numerical layer for second-order accurate finite difference methods with fourth-order dissipation in [15] are exactly the same result we prove in this paper for the RKDG method.

The organization of the paper is as follows. In section 2, we introduce the RKDG method under consideration and state and discuss the main result, which is Theorem 2.1. In section 4, we prove the theorem, and, in section 5, we display numerical results showing that the result is sharp. We end in section 6 with some concluding remarks.

2. The main result. To state our main result, we first give a precise definition of the RKDG under consideration. The RKDG method is obtained by discretizing the equations in space by means of the DG method and then discretizing the resulting system of ODEs by a second-order TVD Runge–Kutta time stepping [18]; see [9].

We take uniform grids in both space and time. Let $\{I_j\}_j$ be a uniform partition of the real line, where $I_j = (x_{j-1/2}, x_{j+1/2})$. Denote the mesh size by $h = x_{j+1/2} - x_{j-1/2}$ and the midpoint by $x_j = (x_{j+1/2} + x_{j-1/2})/2$. Accordingly, the RKDG approximation at each time $t^n = k\,n$, where k denotes the time step, is taken in the space

$$V_h = \{v \in L^2(\mathbb{R}), v|_{I_j} \in P^1(I_j) \ \ \forall j \in \mathbb{Z}\}.$$

For any function $v \in V_h$, we define the jump in v at the nodal point $x_{j+1/2}$ by

$$[v]_{j+1/2} = v(x^+_{j+1/2}) - v(x^-_{j+1/2}).$$

The RKDG approximation $\{\, u_h^n \,\}_{n=0}^{\infty}$ is defined as follows. For $n \geq 0$, u_h^{n+1} is the element of V_h such that

$$\sum_j (u_h^{n+1}, v_h)_j = \sum_j \left\{ \frac{1}{2}(u_h^n, v_h)_j + \frac{1}{2}(w_h^n, v_h)_j \right\} + k\, H(w_h^n, v_h), \tag{2.1a}$$

where $w_h^n \in V_h$ is given by

$$\sum_j (w_h^n, v_h)_j = \sum_j (u_h^n, v_h)_j + kH(u_h^n, v_h) \tag{2.1b}$$

and

$$H(p, q) = \sum_j (p, q_x)_j + p(x^-_{j+1/2})[q]_{j+1/2}. \tag{2.1c}$$

Here $(w, v)_j = \int_{I_j} w\, v\, dx$. For $n = 0$, we set

$$u_h^0 = P_-(U_0). \tag{2.1d}$$

Here the projection $P_-(v)$ of a function v in the space

$$H_h^1(\mathbb{R}) = \{v : \mathbb{R} \mapsto \mathbb{R} \text{ such that } v|_{I_j} \in H^1(I_j)\ \forall j \in \mathbb{Z}\}$$

is defined as the element of V_h satisfying, on each interval I_j, $j \in \mathbb{Z}$,

$$(P_-(v), 1)_j = (v, 1)_j \quad \text{ and } \quad P_-(v)(x^-_{j+1/2}) = v(x^-_{j+1/2}).$$

This projection was introduced in [20] in the framework of DG methods for ODEs and was later used in [3] to obtain optimal error estimates for DG methods for one-dimensional convection-diffusion problems.

We can now state our main result.

THEOREM 2.1. *Let U be the solution of the initial value problem* (1.1)*, and let u_h^N be the solution of the* RKDG *method* (2.1) *at time $T = N\,k$. Suppose that $\lambda := \frac{k}{h} \le 1/3 - \epsilon$ for some fixed $\epsilon \in (0, 1/3)$, and then there exist constants $\beta > 0$ and $C > 0$ such that for any $s \ge 4$*

$$\| U(T) - u_h^N \|_{L^2(\mathbb{R}\setminus\mathcal{R}_T)} \le C\,T\,h^2 \left(1 + \left(\frac{k}{T\,\epsilon}\right)^{1/2}\right) + Ch^s + ||U(T) - P_-(U(T))||_{L^2(\mathbb{R}\setminus\mathcal{R}_T)},$$

where the interval $\mathcal{R}_T$ is given by

$$\mathcal{R}_T = T + \beta\,s\,\log(1/h)\epsilon^{-1}\,(-\lambda^{-7/3}T^{1/3}\,h^{2/3}, \lambda^{-1/2}T^{1/2}\,h^{1/2})$$

and C is independent of h, k, T, and ϵ.

A couple of remarks are in order. The first is that our hypothesis that the CFL condition $\lambda \le 1/3 - \epsilon$ is reasonable, since it is well known that the RKDG under consideration is L^2-stable under the CFL condition

$$\lambda \le 1/3;$$

see, for example, [9, 12]. Taking $\epsilon > 0$ allows us to take advantage of the damping properties of the RKDG method. We would have to significantly modify our approach to deal with the case $\epsilon = 0$, and the result might be very different.

The second remark is that this error estimate extends the result by Shu and Zhang [19] for smooth solutions to nonsmooth solutions, in the special case considered here, of course. Indeed, in our case such a result would read

$$\| U(T) - u_h^N \|_{L^2(\mathbb{R})} \le C\,h^2,$$

provided the initial condition U_0 is smooth. Our results say that the same rate of convergence in h holds if the region $\mathcal{R}_T$ around the discontinuity $x = t$ is removed from $\mathbb{R}$.

3. Idea of the proof. Since the proof of Theorem 2.1 is very technical, we explain here the idea upon which it is based. Suppose that we are interested in establishing error estimates to the left of the line $t = x$; that is, suppose that we want to prove

$$\begin{aligned} \| U(T) - u_h^N \|_{L^2(\mathbb{R}\setminus\mathcal{R}_T^-)} &\le C||U(T) - P_-(U(T))||_{L^2(\mathbb{R}\setminus\mathcal{R}_T^-)} \\ &\quad + C\,T\,h^2 \left(1 + \left(\frac{k}{T\,\epsilon}\right)^{1/2}\right) + Ch^s, \end{aligned} \tag{3.1}$$

where $\mathcal{R}_T^- = (T - \beta\,s\,\log(1/h)\epsilon^{-1}\lambda^{-7/3}T^{1/3}\,h^{2/3}, -\infty)$.

To do this, we do not bound $U(T) - u_h^N$ directly. Instead, we compare u_h^N with a smooth approximation u of U so that

$$\| U(T) - u_h^N \|_{L^2(\mathbb{R}\setminus\mathcal{R}_T^-)} = \|u(T) - u_h^N\|_{L^2(\mathbb{R}\setminus\mathcal{R}_T^-)}.$$

We then obtain a bound of the form

$$\|u(T) - u_h^N\|_{L^2(\mathbb{R}\setminus\mathcal{R}_T^-)} \le C\|\phi_1(T)(u(T) - u_h^N)\|_{L^2(\mathbb{R})},$$

where ϕ_1 is a suitably chosen positive weight function.

The function ϕ_1 will be of order one in the region of interest and will satisfy

$$(\phi_1)_t + (\phi_1)_x = 0.$$

It will be a decreasing function in x which is exponentially small near the line $x = t$; as a consequence, it will have an internal layer. Moreover, ϕ_1 will satisfy

$$\left| \frac{D_x^l(\phi_1(x,t))}{\phi_1(x,t)} \right| \leq \frac{C}{M^l} \quad \text{for all } (x,t), l = 1,2,3,4. \tag{3.2}$$

It is clear that the parameter M is nothing but the size of the above-mentioned layer.

In this way, we reduce the original problem to that of showing a weighted stability result for the error. In solving this problem, we will try to match the size of the layer of the weight function M with that of the internal layer produced by the numerical scheme to the left of the discontinuity. Indeed, roughly speaking, we will show that, if we take $M = \epsilon^{-1}\lambda^{-7/3}T^{1/3}\,h^{2/3}$, we obtain

(3.3)
$$\|\phi_1(T)(u(T) - u_h^N)\|_{L^2(\mathbb{R})} \leq C\|\phi_1(0)(u(0) - u_h^0)\|_{L^2(\mathbb{R})} + C\,T\,h^2\left(1 + \left(\frac{k}{T\,\epsilon}\right)^{1/2}\right).$$

Finally, if we take $u_0 = U_0$ in the region of interest, and since $\phi_1(0)$ is exponentially small when $u_0 \neq U_0$, we would arrive at (3.1).

A similar argument is used to estimate the error to the right of the discontinuity. In this case, however, it turns out that we cannot choose the corresponding weight function ϕ_2 to be a decreasing function in x. This leads to a larger numerical layer to the right.

Most of this paper is dedicated to proving a result similar to (3.3) for the two different weights ϕ_1 and ϕ_2. Next we illustrate one of the main ideas behind the proof of such an estimate by proving a similar result for a continuous problem; the techniques will be similar but more simple.

To this end, we have to find a continuous model for the equation satisfied by the error $\boldsymbol{e}$. Although the model equation for stable second-order accurate finite difference methods with constant coefficients is of the form

$$\boldsymbol{e}_t + \boldsymbol{e}_x = c_2\,h^2\boldsymbol{e}_{xxx} - c_3\,h^3\boldsymbol{e}_{xxxx},$$

where $c_3 > 0$ (see [15]) we are going to work with a simpler equation in order to simplify the computations; the final result will be the same. Thus, let $\boldsymbol{e}$ be the solution of

$$\boldsymbol{e}_t + \boldsymbol{e}_x = h^2\boldsymbol{e}_{xxx} \quad \text{in } \mathbb{R}\times(0,T), \tag{3.4a}$$
$$\boldsymbol{e}(x,0) = \boldsymbol{e}_0(x) \quad \text{on } \mathbb{R}, \tag{3.4b}$$

where $\boldsymbol{e}_0$ is smooth with compact support, and suppose that we are interested in proving a weighted stability result of the form

$$\|\phi(T)\boldsymbol{e}(T)\|_{L^2(\mathbb{R})} \leq C\|\phi(0)\boldsymbol{e}_0\|, \tag{3.5}$$

where ϕ solves $\phi_t + \phi_x = 0$, is a decreasing function in x, and satisfies (3.2).

In obtaining the estimate (3.5), one important question is how to pick M in an optimal way. To answer this question, we proceed as follows. First, we multiply both sides of (3.4a) by $\phi^2 \boldsymbol{e}$, integrate in space and time, and obtain

$$||\phi(T)\boldsymbol{e}(T)||^2_{L^2(\mathbb{R})} = ||\phi(0)\boldsymbol{e}(0)||^2_{L^2(\mathbb{R})} + h^2 \int_0^T (\boldsymbol{e}_{xxx}(t), \phi^2(t)\boldsymbol{e})\, dt.$$

By using the fact that $\boldsymbol{e}_{xxx}\boldsymbol{e} = \frac{1}{2}(\boldsymbol{e}^2)_{xxx} - \frac{3}{2}((\boldsymbol{e}_x)^2)_x$,

$$\begin{aligned} h^2 \int_0^T (\boldsymbol{e}_{xxx}(t), \phi^2(t)\boldsymbol{e})\, dt &= \frac{h^2}{2} \int_0^T (\phi^2(t), (\boldsymbol{e}^2(t))_{xxx})\, dt - \frac{3h^2}{2} \int_0^T ((\phi(t))^2, ((\boldsymbol{e}_x)^2)_x)\, dt \\ &= \frac{h^2}{2} \int_0^T ((\phi^2(t))_{xxx}, \boldsymbol{e}^2(t))\, dt + \frac{3h^2}{2} \int_0^T ((\phi^2(t))_x, (\boldsymbol{e}_x)^2)\, dt. \end{aligned}$$

In the last equation we used integration by parts and the fact that $\boldsymbol{e}$ has compact support. Now we use the fact that $(\phi^2(t))_x < 0$ to get

$$\begin{aligned} ||\phi(T)\boldsymbol{e}(T)||^2_{L^2(\mathbb{R})} &\le ||\phi(0)\boldsymbol{e}(0)||^2_{L^2(\mathbb{R})} + \frac{h^2}{2} \int_0^T ((\phi^2(t))_{xxx}, \boldsymbol{e}^2(t))\, dt \\ &\le ||\phi(0)\boldsymbol{e}(0)||^2_{L^2(\mathbb{R})} + \frac{Ch^2}{M^3} \int_0^T \|\phi(t)\boldsymbol{e}^2(t)\|^2_{L^2(\mathbb{R})}\, dt. \end{aligned}$$

In the last equation we used (3.4a). Finally, by Gronwall's inequality, we have

$$||\phi(T)\boldsymbol{e}(T)||^2_{L^2(\mathbb{R})} \le ||\phi(0)\boldsymbol{e}_0||^2_{L^2(\mathbb{R})} \left(1 + \frac{CTh^2}{M^3} e^{\frac{CTh^2}{M^3}}\right),$$

and we see that we must choose $M = CT^{1/3}h^{2/3}$ to obtain the wanted estimate.

Note that if ϕ is not a decreasing function in x, then M will have to be larger in order to prove stability. This actually happens when dealing with the numerical layer to the right of the discontinuity. This is why the layer to the left of the discontinuity is smaller.

4. Proof. In this section we give a detailed proof of Theorem 2.1. We proceed in several steps.

4.1. Step 1: The error equations. First we define a suitable approximation u to U. The smooth approximation u will satisfy

$$u_t + u_x = 0 \qquad \text{in } \mathbb{R} \times (0,T), \tag{4.1a}$$

$$u(x,0) = u_0(x) \quad \text{on } \mathbb{R}, \tag{4.1b}$$

where u_0 is a smooth function which agrees with U_0 on $(-\infty, -h) \cup (h, \infty)$ such that

$$|D^l u_0(x)| \le Ch^{-l} \quad \text{for } x \in [-h,h],\ l = 1,2,3,4. \tag{4.2}$$

We then obtain the equations satisfied by the error $e_u^n = u(t^n) - u_h^n$. To capture the two-step nature of the RKDG method under consideration, we follow [19] and introduce the function

$$w(x,t) = u(x,t) - k\, u_x(x,t) \tag{4.3}$$

and the corresponding error, namely, $e_w^n = w(t^n) - w_h^n$. Finally, we write, for $p = u$ and $p = w$,

$$e_p^n = \xi_p^n - \eta_p^n, \tag{4.4a}$$

$$\xi_p^n = P_-(p(t^n)) - p_h^n \qquad \text{(projection of the error } e_p^n\text{)}, \tag{4.4b}$$

$$\eta_p^n = P_-(p(t^n)) - p(t^n) \quad \text{(interpolation error)}. \tag{4.4c}$$

In addition to the projection P_-, we also have to introduce the similar projection P_+ given on I_j by

$$(P_+(v), 1)_j = (v, 1)_j \quad \text{and} \quad P_+(v)(x_{j-1/2}^+) = v(x_{j-1/2}^+).$$

The projections $P_\pm$ are strongly related to the bilinear form $H(\cdot,\cdot)$ defined by (2.1c), as we can see in the next result.

LEMMA 4.1. *We have, for any $w \in H_h^1(\mathbb{R})$ and any $v_h \in V_h$,*

$$H(w - P_-(w), v_h) = 0, \tag{4.5a}$$

$$H(v_h, w - P_+(w)) = 0. \tag{4.5b}$$

Proof. The identity (4.5a) follows from the definitions of H and P_-. By using integration by parts we can rewrite $H(p, q)$ as

$$H(p, q) = -\sum_j \{(p_x, q)_j + q(x_{j+1/2}^+)[p]_{j+1/2}\}.$$

The identity (4.5b) now follows from the above equation and the definition of P_+. ∎

The equations for the error are contained in the following result.

LEMMA 4.2. *We have, for any $v \in H_h^1(\mathbb{R})$,*

$$\sum_j (\xi_w^n, v)_j = \sum_j (\xi_u^n, v)_j + k\, H(\xi_u^n, v) + E_2(v) + E_3(v), \tag{4.6a}$$

$$\begin{aligned} \sum_j (\xi_u^{n+1}, v)_j = &\sum_j (\xi_u^n, v)_j + \frac{k}{2} H(\xi_w^n, v) + \frac{k}{2} H(\xi_u^n, v) \\ &+ E_1(P_+(v)) + \frac{1}{2} E_2(v) + \frac{1}{2} E_3(v) + E_4(v) + E_5(v), \end{aligned} \tag{4.6b}$$

where

$$E_1(v) = \frac{k^3}{6} \sum_j (u_{ttt}(\zeta^n), v)_j,$$

$$E_2(v) = \sum_j (\xi_w^n - \xi_u^n, v - P_+(v))_j,$$

$$E_3(v) = \sum_j (\eta_w^n - \eta_u^n, P_+(v))_j,$$

$$E_4(v) = \sum_j (\xi_u^{n+1} - (\xi_u^n + \xi_w^n)/2, v - P_+(v))_j,$$

$$E_5(v) = \sum_j (\eta_u^{n+1} - (\eta_u^n + \eta_w^n)/2, P_+(v))_j.$$

Here ζ^n is a function of x lying in the interval (t^n, t^{n+1}).

Proof. Let us begin by proving the identity (4.6a). For any function v in $H_h^1(\mathbb{R})$, we have, by the definition of ξ_p and η_p—(4.4)—

$$\begin{aligned}\sum_j (\xi_w^n, v)_j &= \sum_j \{(e_w^n, v)_j + (\eta_w^n, v)_j\} \\ &= \sum_j \{(e_w^n, P_+(v))_j + (e_w^n, v - P_+(v))_j + (\eta_w^n, v)_j\} \\ &= \sum_j \{(e_w^n, P_+(v))_j + (\xi_w^n, v - P_+(v))_j + (\eta_w^n, P_+ v)_j\}.\end{aligned}$$

To suitably rewrite the first term of the right-hand side, we notice that, by the definition of w—(4.3)—and since u is smooth, we have

$$\sum_j (w(t^n), v_h)_j = \sum_j (u(t^n), v_h)_j + k\, H(u(t^n), v_h)$$

for any $v_h \in V_h$. Hence, by subtracting the equation defining w_h^n—(2.1b)—we obtain

$$\sum_j (e_w^n, v_h)_j = \sum_j (e_u^n, v_h)_j + k\, H(e_u^n, v_h). \tag{4.7}$$

Now, by taking $v_h = P_+(v)$, we immediately get

$$\begin{aligned}\sum_j (e_w^n, P_+(v))_j &= \sum_j (e_u^n, P_+(v))_j + k\, H(e_u^n, P_+(v)) \\ &= \sum_j (e_u^n, P_+(v))_j + k\, H(\xi_u^n, v),\end{aligned}$$

since, by the properties (4.5a) and (4.5b), we have

$$H(e_u^n, P_+(v)) \;=\; H(P_-(e_u^n), P_+(v)) \;=\; H(\xi_u^n, P_+(v)) \;=\; H(\xi_u^n, v).$$

As a consequence, we get

$$\begin{aligned}\sum_j (\xi_w^n, v)_j &= \sum_j (e_u^n, P_+(v))_j + k\, H(\xi_u^n, v) \\ &\quad + \sum_j \{(\xi_w^n, v - P_+(v))_j + (\eta_w^n, P_+(v))_j\} \\ &= \sum_j (\xi_u^n, v)_j + k\, H(\xi_u^n, v) \\ &\quad + \sum_j \{(\xi_w^n - \xi_u^n, v - P_+(v))_j + (\eta_w^n - \eta_u^n, P_+(v))_j\} \\ &= \sum_j (\xi_u^n, v)_j + k\, H(\xi_u^n, v) + E_2(v) + E_3(v)\end{aligned}$$

by the definition of $E_2(v)$ and $E_3(v)$. This proves the identity (4.6a).

It remains to prove the identity (4.6b). For any function v in $H_h^1(\mathbb{R})$ we have

$$\begin{aligned}\sum_j (\xi_u^{n+1}, v)_j &= \sum_j \{(\xi_u^{n+1}, P_+(v))_j + (\xi_u^{n+1}, v - P_+(v))_j\} \\ &= \sum_j \{(e_u^{n+1}, P_+(v))_j + (\eta_u^{n+1}, P_+(v))_j + (\xi_u^{n+1}, v - P_+(v))_j\}\end{aligned}$$

by the definition of ξ_p and η_p—(4.4).

Next, we rewrite the first term of the right-hand side. By Taylor's theorem

$$u(x,t+k)=\frac{1}{2}u(x,t)+\frac{1}{2}w(x,t)-\frac{k}{2}w_x(x,t)+\frac{k^3}{6}u_{ttt}(x,\zeta),$$

where ζ depends on x and is $t\le\zeta\le t+k$. After a simple integration by parts, we have

$$\sum_j(u(t^{n+1}),v)_j=\sum_j\frac{1}{2}(u(t^n),v)_j+\frac{1}{2}(w(t^n),v)_j+\frac{k}{2}H(w(t^n),v)+E_1(v).$$

By subtracting the equation defining u_h^{n+1}—(2.1a)—we get for any $v_h\in V_h$

$$\sum_j(e_u^{n+1},v_h)_j=\sum_j\frac{1}{2}(e_u^n,v_h)_j+\frac{1}{2}(e_w^n,v_h)_j+\frac{k}{2}H(e_w^n,v_h)+E_1(v_h),$$

and, by using (4.7), we obtain

$$\sum_j(e_u^{n+1},v_h)_j=\sum_j(e_u^n,v_h)_j+\frac{k}{2}H(e_w^n,v_h)+\frac{k}{2}H(e_u^n,v_h)+E_1(v_h).$$

Finally, by taking $v_h=P_+(v)$ and using (4.5b), this equation becomes

$$\sum_j(e_u^{n+1},P_+(v))_j=\sum_j(e_u^n,P_+(v))_j+\frac{k}{2}H(\xi_w^n,v)_j+\frac{k}{2}H(\xi_u^n,v)+E_1(P_+(v)).$$

This implies that

$$\begin{aligned}\sum_j(\xi_u^{n+1},v)_j&=\sum_j(e_u^n,P_+(v))_j+\frac{k}{2}H(\xi_w^n,v)+\frac{k}{2}H(\xi_u^n,v)+E_1(P_+(v))\\&\quad+\sum_j\{(\eta_u^{n+1},P_+(v))_j+(\xi_u^{n+1},v-P_+(v))_j\}\\&=\sum_j(\xi_u^n,v)_j+\frac{k}{2}H(\xi_w^n,v)+\frac{k}{2}H(\xi_u^n,v)+E_1(P_+(v))\\&\quad+\sum_j\{(\eta_u^{n+1}-\eta_u^n,P_+(v))_j+(\xi_u^{n+1}-\xi_u^n,v-P_+(v))_j\}\\&=\sum_j(\xi_u^n,v)_j+\frac{k}{2}H(\xi_w^n,v)+\frac{k}{2}H(\xi_u^n,v)+E_1(P_+(v))\\&\quad+\frac{1}{2}E_2(v)+\frac{1}{2}E_3(v)+E_4(v)+E_5(v).\qquad\square\end{aligned}$$

4.2. Step 2: The weights. Theorem 2.1 will follow from estimates of quantities of the form

$$\|\phi(\cdot,t^n)(u(t^n)-u_h^n)\|_{L^2(\mathbb{R})},$$

where $\phi(x,t)$ is a suitably chosen weight function. Here we describe the weights with which we are going to work.

Since, as we stated in the introduction, we have different results on the left and on the right of the discontinuity, we consider two slightly different weight functions

ϕ_1 and ϕ_2. The weight ϕ_1 will be of order one in the region $x < t + x_1$ for some number $x_1 < 0$ and "small" in the region $x > t + x_1 + d_1$ for some $d_1 > 0$, whereas the weight ϕ_2 is of order one on the region $x > t + x_2$ for some number $x_2 > 0$ and small in the region $x < t + x_2 - d_2$ for some $d_2 > 0$. A very important property that will allow better results to the left of the discontinuity is that $\phi_1(x,t)$ can be chosen as a decreasing function in x.

We take the functions ϕ_i, for $i = 1, 2$, as solutions of

$$\begin{aligned} \phi_{i,t} + \phi_{i,x} &= 0 \qquad \text{in } R \times (0,T), \\ \phi_i(x,0) &= b_i(x) \quad \text{on } R \end{aligned}$$

for some initial conditions b_1 and b_2. We choose the initial conditions b_1 and b_2 so that ϕ_1 and ϕ_2 satisfy the following.

PROPOSITION 4.3. *We have*

$$c \le \phi_i(x,t) \le C \qquad \text{for } (x,t) \in \Omega_i(0), \tag{4.8a}$$

$$\phi_i(x,t) \le h^s \qquad \text{for } (x,t) \in \mathbb{R}^2 \backslash \Omega_i(d_i), \tag{4.8b}$$

$$(-1)^{i+1}\phi_{i,x}(x,t) < 0 \qquad \text{for all } (x,t), \tag{4.8c}$$

$$\left|\frac{D_x^{l+1}(\phi_i(x,t))}{(\phi_i(x,t))_x}\right| + \left|\frac{D_x^{l}(\phi_i(x,t))}{\phi_i(x,t)}\right| \le \frac{C}{(\tilde{K}_i \tilde{T}^{1-\gamma_i} h^{\gamma_i})^l} \qquad \text{for all } (x,t), l = 1,2,3, \tag{4.8d}$$

$$RO(S,\phi_i) + RO(S,(\phi_i)_x) \le C \qquad \text{for all squares } S \text{ with sides of length } \tilde{T}^{1-\gamma_i} h^{\gamma_i}, \tag{4.8e}$$

where

$$\begin{aligned} \Omega_1(d) &= \{(x,t) : x \le t + x_1 + d\}, \\ \Omega_2(d) &= \{(x,t) : x \ge t + x_2 - d\}, \\ d_i &= s \log(1/h) \tilde{K}_i \tilde{T}^{1-\gamma_i} h^{\gamma_i}, \end{aligned}$$

$RO(D,\chi) = \max_{r\in D} |\chi(r)| \,/\, \min_{r\in D} |\chi(r)|$ *for any domain* D, *and* c *and* C *are positive constants independent of* h, λ, *and* T. *Here* $\tilde{T} = \frac{T}{\lambda}$ *and* $\tilde{K}_i = \frac{K_i}{\lambda^{m_i}}$. *The parameters* K_i, γ_i, *and* m_i *satisfy* $K_i \ge 1$, $0 \le \gamma_i \le 1$, *and* $m_i \ge 0$ *for* $i = 1, 2$.

In the rest of this paper we assume that the CFL number satisfies $\lambda \le 1/3$. Therefore, often we will use the fact that $\frac{1}{\tilde{K}_i} \le \frac{1}{K_i} \le 1$. Also, since $\tilde{T}^{1-\gamma_i} h^{\gamma_i} = (\frac{T}{k})^{1-\gamma_i} h \ge h$, we see that $RO(S,\phi_i)$ and $RO(S,(\phi_i)_x)$ are bounded for squares S with sides of size h.

The construction of b_1 and b_2 is very similar to the construction of weight functions used in [16]. We include it in the appendix for completeness.

4.3. Step 3: The error in one time step. Next, we find how the weighted error changes in a single time step. This information is captured in a key identity contained in the following result.

LEMMA 4.4. *Let* ϕ *be any solution of the equation* $\phi_t + \phi_x = 0$. *Then we have*

$$\|\phi(t^{n+1})\xi_u^{n+1}\|^2_{L^2(\mathbb{R})} + \mathbb{J}_h = \|\phi(t^n)\xi_u^n\|^2_{L^2(\mathbb{R})} + \frac{k}{2}\Theta((\phi^2)_x) + \Psi_h(\phi) + E_h, \tag{4.9}$$

where

$$\begin{aligned}\mathbb{J}_h :=& \frac{k}{2}\sum_j \{[\phi(t^n)\xi_u^n]^2_{j+1/2} + [\phi(t^{n+1})\xi_w^n]^2_{j+1/2}\}\\ \Theta(v) :=& \sum_j \Theta_j(v), \quad \Theta_j(v) := (\xi_w^n - \xi_u^n, v(t^n)(\xi_w^n - \xi_u^n))_j,\\ \Psi_h(\phi) :=& \|\phi(t^{n+1})(\xi_u^{n+1} - \xi_w^n)\|^2_{L^2(\mathbb{R})}.\end{aligned}$$

Moreover,

$$\begin{aligned}E_h :=& 2\,E_1(P_+(\phi^2(t^{n+1})\xi_w^n)) + 2\,E_4(\phi^2(t^{n+1})\xi_w^n) + 2\,E_5(\phi^2(t^{n+1})\xi_w^n)\\ &+ E_2(\phi^2(t^n)\xi_u^n) + E_3(\phi^2(t^n)\xi_u^n) - \frac{k^2}{2}\sum_j (\xi_w^n - \xi_u^n, (\phi^2(t^n))_{xx}\xi_w^n)_j\\ &+ S_1 + S_2,\end{aligned}$$

where

$$S_1 = -\frac{k^3}{6}\sum_j (\xi_u^n, (\phi^2)_{xxx}(\theta_1^n)\xi_w^n)_j \quad \textit{and} \quad S_2 = \frac{k^3}{4}\sum_j (\xi_w^n, (\phi^2)_{xxx}(\theta_2^n)\xi_w^n)_j.$$

Here θ_i^n, $i = 1, 2$, depend on x and $t^n \le \theta_i^n \le t^{n+1}$.

Before proving this result, let us briefly discuss some of its salient features. Notice the following:

• This result is completely independent of the fact that the approximate solution is piecewise linear in space. It takes into account only the nature of the Runge–Kutta time-stepping method we are considering.

• The term $\mathbb{J}_h$ containing the jumps across interelement boundaries captures the dissipative nature of the DG-space discretization. In the analysis, it will allow us to control the terms of the right-hand side.

• The term $\Theta((\phi^2)_x)$ will allow us to distinguish between the behavior of the error to the right and that to the left of the discontinuity. Since $\Theta((\phi_1^2)_x) \le 0$, this term enhances the damping properties of the method to the left of the discontinuity. This property does not hold for $\Theta((\phi_2^2)_x)$, which will have to be controlled by the jumps in $\mathbb{J}_h$. This is the technicality that captures the fact that the approximation properties to the left and to the right of the discontinuity of the exact solution are very different. This results in a region $\mathcal{R}_T$ whose size is significantly *smaller* to the left of the discontinuity than to its right.

• If the initial condition U_0 were smooth, we could then take $\phi \equiv 1$, and the above result would be the first step in the L^2-error analysis; see [19]. In such a case, the estimates of the terms of the right-hand side play a crucial role, with the exception of the term $\Theta((\phi^2)_x)$, which is identically equal to zero.

• Note that this lemma also contains the first step of the L^2-stability analysis of the RKDG method under consideration, which we obtain by setting $\phi \equiv 1$ and $u = 0$. In this case, (4.9) becomes

$$||u_h^{n+1}||^2_{L^2(\mathbb{R})} + \frac{k}{2}\sum_j \{[u_h^n]_j^2 + [w_h^n]_j^2\} = ||u_h^n||^2_{L^2(\mathbb{R})} + ||u_h^{n+1} - w_h^n||^2_{L^2(\mathbb{R})}. \tag{4.10}$$

The second term of the left-hand side reflects the dissipative nature of the DG-space discretization of the method, whereas the second term of the right-hand side captures the *antidissipative* nature typical of explicit schemes. The condition that the dissipative term dominates the antidissipative is nothing but the CFL condition. As we are going to see, a sharp estimate of the term $||u_h^{n+1} - w_h^n||_{L^2(\mathbb{R})}$ allows us to see that the method is L^2-stable, provided $\lambda \leq 1/3$.

Let us now prove Lemma 4.4.

Proof. We begin by considering the trivial identity

$$\begin{aligned}\|\phi(t^{n+1})\xi_u^{n+1}\|^2_{L^2(\mathbb{R})} =& \|\phi(t^{n+1})(\xi_u^{n+1} - \xi_w^n)\|^2_{L^2(\mathbb{R})} \\ &+ 2\sum_j (\xi_u^{n+1}, \phi^2(t^{n+1})\xi_w^n)_j - \|\phi(t^{n+1})\xi_w^n\|^2_{L^2(\mathbb{R})}.\end{aligned}$$

If we subtract (4.6a) from two times (4.6b), we get

$$\begin{aligned}2\sum_j (\xi_u^{n+1}, v)_j = &\sum_j \{(\xi_w^n, v)_j + (\xi_u^n, v)_j\} + kH(\xi_w^n, v) \\ &+ 2E_1(P_+(v)) + 2E_4(v) + 2E_5(v).\end{aligned}$$

By taking $v = \phi^2(t^{n+1})\, \xi_w^n$ in this equation and $v = \phi^2(t^n)\, \xi_u^n$ in (4.6a), we obtain

$$\begin{aligned}2\sum_j (\xi_u^{n+1}, \phi^2(t^{n+1})\xi_w^n)_j = &\|\phi(t^n)\xi_u^n\|^2_{L^2(\mathbb{R})} + \|\phi(t^{n+1})\xi_w^n\|^2_{L^2(\mathbb{R})} \\ &+ kH(\xi_u^n, \phi^2(t^n)\xi_u^n) + kH(\xi_w^n, \phi^2(t^{n+1})\xi_w^n) \\ &+ 2E_1(P_+(\phi^2(t^{n+1})\xi_w^n)) + 2E_4(\phi^2(t^{n+1})\xi_w^n) \\ &+ 2E_5(\phi^2(t^{n+1})\xi_w^n) + E_2(\phi^2(t^n)\xi_u^n) \\ &+ E_3(\phi^2(t^n)\xi_u^n) + \sum_j (\xi_w^n, (\phi^2(t^{n+1}) - \phi^2(t^n))\xi_u^n)_j,\end{aligned}$$

and hence

$$\begin{aligned}\|\phi(t^{n+1})\xi_u^{n+1}\|^2_{L^2(\mathbb{R})} = &\|\phi(t^n)\xi_u^n\|^2_{L^2(\mathbb{R})} + \|\phi(t^{n+1})(\xi_u^{n+1} - \xi_w^n)\|^2_{L^2(\mathbb{R})} \\ &+ kH(\xi_u^n, \phi^2(t^n)\xi_u^n) + kH(\xi_w^n, \phi^2(t^{n+1})\xi_w^n) \\ &+ 2E_1(P_+(\phi^2(t^{n+1})\xi_w^n)) + 2E_4(\phi^2(t^{n+1})\xi_w^n) \\ &+ 2E_5(\phi^2(t^{n+1})\xi_w^n) + E_2(\phi^2(t^n)\xi_u^n) \\ &+ E_3(\phi^2(t^n)\xi_u^n) + \sum_j (\xi_u^n, (\phi^2(t^{n+1}) - \phi^2(t^n))\xi_w^n)_j.\end{aligned}$$

Since, for any $v \in H_h^1(\mathbb{R})$,

$$\begin{aligned}H(v, \phi^2(t)v) =& H(\phi(t)v, \phi(t)v) + \frac{1}{2}\sum_j (v, (\phi^2(t))_x v)_j \\ =& -\frac{1}{2}\sum_j [\phi(t)v]^2_{j+1/2} + \frac{1}{2}\sum_j (v, (\phi^2(t))_x v)_j,\end{aligned}$$

we get

$$\begin{aligned}\|\phi(t^{n+1})\xi_u^{n+1}\|^2_{L^2(\mathbb{R})} + \mathbb{J}_h = \|\phi(t^n)\xi_u^n\|^2_{L^2(\mathbb{R})} + \Psi_h(\phi)\\ +\frac{k}{2}\sum_j\{(\xi_u^n,(\phi^2(t^n))_x\xi_u^n)_j + (\xi_w^n,(\phi^2(t^{n+1}))_x\xi_w^n)_j\}\\ +2E_1(P_+(\phi^2(t^{n+1})\xi_w^n)) + 2E_4(\phi^2(t^{n+1})\xi_w^n)\\ +2E_5(\phi^2(t^{n+1})\xi_w^n) + E_2(\phi^2(t^n)\xi_u^n)\\ +E_3(\phi^2(t^n)\xi_u^n) + \sum_j(\xi_u^n,(\phi^2(t^{n+1})-\phi^2(t^n))\xi_w^n)_j.\end{aligned} \tag{4.11}$$

By using Taylor's expansion, we see that

$$\begin{aligned}\sum_j(\xi_u^n,(\phi^2(t^{n+1})-\phi^2(t^n))\xi_w^n)_j = -k\sum_j(\xi_u^n,(\phi^2(t^n))_x\xi_w^n)_j\\ +\frac{k^2}{2}\sum_j(\xi_u^n,(\phi^2(t^n))_{xx}\xi_w^n)_j + S_1.\end{aligned}$$

Here we used that $(\phi^2)_t = -(\phi^2)_x$. Similarly,

$$\begin{aligned}\frac{k}{2}\sum_j(\xi_w^n,(\phi^2(t^{n+1}))_x\xi_w^n)_j = \frac{k}{2}\sum_j(\xi_w^n,(\phi^2(t^n))_x\xi_w^n)_j\\ -\frac{k^2}{2}\sum_j(\xi_w^n,(\phi^2(t^n))_{xx}\xi_w^n)_j + S_2.\end{aligned}$$

Therefore,

$$\begin{aligned}&\frac{k}{2}\sum_j\{(\xi_u^n,(\phi^2(t^n))_x\xi_u^n)_j + (\xi_w^n,(\phi^2(t^{n+1}))_x\xi_w^n)_j\}\\ &\quad+\sum_j(\xi_u^n,(\phi^2(t^{n+1})-\phi^2(t^n))\xi_w^n)_j\\ &=\frac{k}{2}\sum_j\{(\xi_u^n,(\phi^2(t^n))_x\xi_u^n)_j + (\xi_w^n,(\phi^2(t^n))_x\xi_w^n)_j\}\\ &\quad-k\sum_j(\xi_u^n,(\phi^2(t^n))_x\xi_w^n)_j - \frac{k^2}{2}\sum_j(\xi_w^n-\xi_u^n,(\phi^2(t^n))_{xx}\xi_w^n) + S_1 + S_2\\ &=\frac{k}{2}\Theta((\phi^2)_x) - \frac{k^2}{2}\sum_j(\xi_w^n-\xi_u^n,(\phi^2(t^n))_{xx}\xi_w^n) + S_1 + S_2.\end{aligned}$$

The result follows after inserting the above identity into (4.11). This completes the proof. ∎

4.4. Step 4: Bound for Ψ_h. In order to bound Ψ_h we first need a bound for $||\xi_u^{n+1}-\xi_w^n||_{I_j}$ for each interval I_j. By exploiting the fact that $\xi_u^{n+1}-\xi_w^n$ is linear on each element, we express $\xi_u^{n+1}-\xi_w^n$ in terms of its mean and derivative on each element. More precisely, if $v\in V_h$, then for $x\in I_j$ we have

$$v(x) = \bar{v}(x_j) + v_x(x_j)(x-x_j), \tag{4.12}$$

where the average of v on I_j is $\bar{v}(x_j) = \frac{1}{h}(v,1)_j$. The representation we are seeking is contained in the following lemma, which is proved in the appendix. For the rest of

this paper θ^n will denote a number satisfying $t^n \le \theta^n \le t^{n+1}$ which depends on x; the values of θ^n may be different in separate occurrences. It will appear in different places when we use Taylor's theorem.

LEMMA 4.5. *For $x \in I_j$, we have*

$$(\xi_u^{n+1} - \xi_w^n)(x) = \frac{-3k^2}{h^2}[\xi_u^n]_{j-1/2} - \frac{k}{2h}[\xi_w^n - \xi_u^n]_{j-1/2} + \frac{k^3}{6h}R_{3,j} - \frac{k}{2h^3}R_{1,j}$$
$$+ \left(\frac{3k}{h^2}[\xi_w^n - \xi_u^n]_{j-1/2} + \frac{k^3}{6h^3}R_{4,j} - \frac{6}{h^3}R_{1,j} + \frac{6}{h^3}R_{2,j}\right)(x - x_j) \tag{4.13a}$$

and

$$(\xi_w^n - \xi_u^n)(x) = -k(\xi_u^n)_x(x_j) - \frac{k}{h}[\xi_u^n]_{j-1/2} + \left(6\frac{k}{h^2}[\xi_u^n]_{j-1/2} + \frac{12}{h^3}R_{1,j}\right)(x - x_j), \tag{4.13b}$$

where

$$R_{1,j} = \int_{I_j} (\eta_w^n - \eta_u^n)(x - x_j)dx, \quad R_{2,j} = \int_{I_j} (\eta_u^{n+1} - (\eta_w^n + \eta_u^n)/2)(x - x_j)dx,$$
$$R_{3,j} = \int_{I_j} u_{ttt}(\theta^n)dx, \qquad R_{4,j} = \int_{I_j} u_{ttt}(\theta^n)(x - x_j)dx.$$

Now we calculate the term $\|(\xi_u^{n+1} - \xi_w^n)\|^2_{L^2(I_j)}$. We are going to express it in terms of the vector $[\boldsymbol{\xi}^n]_{j-1/2} := ([\xi_u^n]_{j-1/2}, [\xi_w^n]_{j-1/2})$.

LEMMA 4.6. *If $\lambda = \frac{k}{h}$, then*

$$\|\xi_u^{n+1} - \xi_w^n\|^2_{L^2(I_j)} = k\,[\boldsymbol{\xi}^n]_{j-1/2}\mathbb{A}[\boldsymbol{\xi}^n]^t_{j-1/2} + Y_j, \tag{4.14}$$

where

$$\mathbb{A} := \begin{pmatrix} Q(\lambda) & \frac{1}{2}Z(\lambda) \\ \frac{1}{2}Z(\lambda) & \lambda \end{pmatrix}, \quad Q(\lambda) = 9\lambda^3 - 3\lambda^2 + \lambda, \quad Z(\lambda) = 3\lambda^2 - 2\lambda,$$

and

$$Y_j = h\left(\frac{k^3}{6h}R_{3,j} - \frac{k}{2h^3}R_{1,j}\right)^2 + \frac{h^3}{12}\left(\frac{k^3}{6h^3}R_{4,j} - \frac{6}{h^3}R_{1,j} + \frac{6}{h^3}R_{2,j}\right)^2$$
$$+ 2h\left(\frac{-3k^2}{h^2}[\xi_u^n]_{j-1/2} - \frac{k}{2h}[\xi_w^n - \xi_u^n]_{j-1/2}\right)\left(\frac{k^3}{6h}R_{3,j} - \frac{k}{2h^3}R_{1,j}\right)$$
$$+ 2\frac{h^3}{12}\left(\frac{3k}{h^2}[\xi_w^n - \xi_u^n]_{j-1/2}\right)\left(\frac{k^3}{6h^3}R_{4,j} - \frac{6}{h^3}R_{1,j} + \frac{6}{h^3}R_{2,j}\right).$$

Proof. By using (4.13a) and the fact that $\int_{I_j}(x - x_j)^2 = \frac{h^3}{12}$ and $\int_{I_j}(x - x_j)dx = 0$, we get

$$\|\xi_u^{n+1} - \xi_w^n\|^2_{L^2(I_j)} = h\,(\overline{(\xi_u^{n+1} - \xi_w^n)}(x_j))^2 + \frac{h^3}{12}((\xi_u^{n+1} - \xi_w^n)_x(x_j))^2$$
$$= h\left(\frac{-3k^2}{h^2}[\xi_u^n]_{j-1/2} - \frac{k}{2h}[\xi_w^n - \xi_u^n]_{j-1/2} + \frac{k^3}{6h}R_{3,j} - \frac{k}{2h^3}R_{1,j}\right)^2$$
$$+ \frac{h^3}{12}\left(\frac{3k}{h^2}[\xi_w^n - \xi_u^n]_{j-1/2} + \frac{k^3}{6h^3}R_{4,j} - \frac{6}{h^3}R_{1,j} + \frac{6}{h^3}R_{2,j}\right)^2,$$

and the result follows after a few simple algebraic manipulations. ∎

Note that, by simply setting $\phi = 1$ and $u = 0$, we obtain an identity which is used for the L^2-stability analysis, namely,

$$\|u_h^{n+1} - w_h^n\|^2_{L^2(I_j)} = k \sum_j [\boldsymbol{u}^n]_{j-1/2} \, \mathbb{A} [\boldsymbol{u}^n]^t_{j-1/2},$$

where $[\boldsymbol{u}^n]_{j-1/2} := ([u_h^n]_{j-1/2}, [w_h^n]_{j-1/2})$. Indeed, by inserting this expression in the identity (4.10), we get

$$\|u_h^{n+1}\|^2_{L^2(\mathbb{R})} + k \sum_j [\boldsymbol{u}^n]_{j-1/2} \left(\frac{1}{2} \mathbb{I} - \mathbb{A} \right) [\boldsymbol{u}^n]^t_{j-1/2} = \| \, u_h^n \, \|^2_{L^2(\mathbb{R})},$$

where $\mathbb{I}$ is the 2×2 identity matrix. Therefore, the method is L^2-stable if the matrix $\frac{1}{2}\mathbb{I} - \mathbb{A}$ is positive semidefinite. This occurs precisely for $\lambda \leq 1/3$.

Note also that, although we do not need to take $\lambda < 1/3$ to achieve L^2-stability, we need to take this choice in our error analysis. As we are going to see, taking $\lambda \leq 1/3 - \epsilon$, with $\epsilon > 0$, ensures that the matrix $\frac{1}{2}\mathbb{I} - \mathbb{A}$ is positive definite, and, as a consequence, the terms involving the jumps of ξ_u^n and ξ_w^n in the right-hand side of the identity (4.9) of Lemma 4.4 can be controlled by the term $\mathbb{J}_h$.

We now state and prove the bound for $\Psi_h(\phi_i)$. We will need the following notation:

$$[\boldsymbol{\phi}\boldsymbol{\xi}^n]_{j-1/2} := ([\phi(t^n)\, \xi_u^n]_{j-1/2}, \, [\phi(t^{n+1})\, \xi_w^n]_{j-1/2}).$$

LEMMA 4.7. *We have*

$$\begin{aligned}
\Psi_h(\phi_i) \leq k \sum_j & [\boldsymbol{\phi}_i \boldsymbol{\xi}^n]_{j-1/2} \, \mathbb{B}_i [\boldsymbol{\phi}_i \boldsymbol{\xi}^n]^t_{j-1/2} \\
& + C \, K_i \|\phi_i(t^n)(\eta_w^n - \eta_u^n)\|^2_{L^2(I_j)} \\
& + C \, K_i \|\phi_i(t^n)(\eta_u^{n+1} - (\eta_u^n + \eta_w^n)/2)\|^2_{L^2(I_j)} \\
& + C \, K_i k^6 \|\phi_i(t^n) u_{ttt}(\theta^n)\|^2_{L^2(\mathbb{R})},
\end{aligned}$$

where

$$\mathbb{B}_i = \mathbb{A} + \frac{C}{K_i} \, \mathbb{I}.$$

Proof. To simplify the notation, we drop the subindex i. We have

$$\Psi_h(\phi) = T_1 + T_2,$$

where

$$\begin{aligned}
T_1 &= \sum_j \phi^2(x_{j-1/2}, t^{n+1}) \|(\xi_u^{n+1} - \xi_w^n)\|^2_{L^2(I_j)}, \\
T_2 &= \sum_j \int_{I_j} (\phi^2(x, t^{n+1}) - \phi^2(x_{j-1/2}, t^{n+1}))(\xi_u^{n+1} - \xi_w^n)^2(x) \, dx.
\end{aligned}$$

We first estimate the term T_1. We multiply (4.14) by $\phi^2(x_{j-1/2}, t^{n+1})$, and we bound the terms of the resulting right-hand side separately. Since

$$\begin{aligned}
\phi^2(x_{j-1/2}, t^{n+1}) &= \phi^2(x_{j-1/2}, t^n) + k(\phi^2)_t(x_{j-1/2}, \theta^n), \\
\phi(x_{j-1/2}, t^{n+1}) &= \phi(x_{j-1/2}, t^n) + k\phi_t(x_{j-1/2}, \theta^n),
\end{aligned}$$

by using inequalities (4.8d) and (4.8e), we have

$$k\phi^2(x_{j-1/2},t^{n+1})[\boldsymbol{\xi}^n]_{j-1/2}\mathbb{A}[\boldsymbol{\xi}^n]^t_{j-1/2} \le k[\boldsymbol{\phi\xi}^n]_{j-1/2}\mathbb{A}[\boldsymbol{\phi\xi}^n]^t_{j-1/2} + k\left(\frac{Ch^{1-\gamma}}{\tilde{T}^{1-\gamma}\tilde{K}}\right)[\boldsymbol{\phi\xi}^n]_{j-1/2}\mathbb{I}[\boldsymbol{\phi\xi}^n]^t_{j-1/2}.$$

Here we used the fact that the entries of $\mathbb{A}$ are bounded for $\lambda \in [0, 1/3]$.

By using (4.8e) and the Cauchy–Schwarz inequality, we get, from the definition of $R_{l,j}$ for $l = 1, \cdots, 4$,

$$\begin{aligned}&\phi(x_{j-1/2},t^n)\left(\frac{k^6}{h}|R_{3,j}|^2 + \frac{1}{h^3}|R_{1,j}|^2 + \frac{k^6}{h^3}|R_{4,j}|^2 + \frac{1}{h^3}|R_{2,j}|^2\right)\\ &\le C(\|\phi(t^n)(\eta^n_w - \eta^n_u)\|^2_{L^2(I_j)} + \|\phi(t^n)(\eta^{n+1}_u - (\eta^n_u + \eta^n_w)/2)\|^2_{L^2(I_j)})\\ &\quad + Ck^6\|\phi(t^n)u_{ttt}(\theta^n)\|^2_{L^2(I_j)}.\end{aligned}$$

Therefore, if we again apply (4.8e), multiply and divide each term of Y_j by $K^{1/2}$, and apply Young's inequality, we obtain

$$\begin{aligned}|\phi^2(x_{j-1/2},t^{n+1})Y_j| \le{}& \frac{1}{K}k[\boldsymbol{\phi\xi}^n]_{j-1/2}\mathbb{I}[\boldsymbol{\phi\xi}^n]^t_{j-1/2}\\ &+ C\,K\|\phi(t^n)(\eta^n_w - \eta^n_u)\|^2_{L^2(I_j)}\\ &+ C\,K\|\phi(t^n)(\eta^{n+1}_u - (\eta^n_u + \eta^n_w)/2)\|^2_{L^2(I_j)}\\ &+ C\,Kk^6\|\phi(t^n)u_{ttt}(\theta^n)\|^2_{L^2(I_j)}.\end{aligned}$$

Hence,

$$\begin{aligned}T_1 \le{}& k\sum_j[\boldsymbol{\phi\xi}^n]_{j-1/2}\left(\mathbb{A} + \left(\left(\frac{h}{\tilde{T}}\right)^{1-\gamma}\frac{C}{\tilde{K}} + \frac{C}{K}\right)\mathbb{I}\right)[\boldsymbol{\phi\xi}^n]^t_{j-1/2}\\ &+ C\,K\|\phi(t^n)(\eta^n_w - \eta^n_u)\|^2_{L^2(I_j)}\\ &+ C\,K\|\phi(t^n)(\eta^{n+1}_u - (\eta^n_u + \eta^n_w)/2)\|^2_{L^2(I_j)} + C\,K\,k^6\|\phi(t^n)u_{ttt}(\theta^n)\|^2_{L^2(\mathbb{R})}.\end{aligned}$$

To bound T_2 we see from (4.14) that

$$T_2 \le \sum_j \|\phi^2(x_{j-1/2},t^{n+1}) - \phi^2(t^{n+1})\|_{L^\infty(I_j)}(k[\boldsymbol{\phi\xi}^n]_{j-1/2}\mathbb{A}[\boldsymbol{\phi\xi}^n]^t_{j-1/2} + Y_j).$$

By using the inequality

$$\|\phi^2(x_{j-1/2},t^{n+1}) - \phi^2(t^{n+1})\|_{L^\infty(I_j)} \le \frac{Ch^{1-\gamma}}{\tilde{T}^{1-\gamma}\tilde{K}}\|\phi^2(t^{n+1})\|_{L^\infty(I_j)},$$

the inequality (4.8e), and Young's inequality, we obtain

$$\begin{aligned}&\|\phi^2(x_{j-1/2},t^{n+1}) - \phi^2(t^{n+1})\|_{L^\infty(I_j)}(k[\boldsymbol{\phi\xi}^n]_{j-1/2}\mathbb{A}[\boldsymbol{\phi\xi}^n]^t_{j-1/2})\\ &\le \frac{Ch^{1-\gamma}}{\tilde{T}^{1-\gamma}\tilde{K}}k[\boldsymbol{\phi\xi}^n]_{j-1/2}\mathbb{I}[\boldsymbol{\phi\xi}^n]^t_{j-1/2}.\end{aligned}$$

By the triangle inequality and (4.8e) we have

$$\|\phi^2(x_{j-1/2},t^{n+1}) - \phi^2(t^{n+1})\|_{L^\infty(I_j)}|Y_j| \le C|\phi^2(x_{j-1/2},t^{n+1})Y_j|.$$

Therefore,

$$T_2 \le k\sum_j [\boldsymbol{\phi\xi}^n]_{j-1/2}\left(\left(\frac{h}{\bar{\bar{T}}}\right)^{1-\gamma}\frac{C}{\bar{\bar{K}}}+\frac{C}{K}\right)\mathbb{I}[\boldsymbol{\phi\xi}^n]^t_{j-1/2}$$
$$+C\,K\|\phi(t^n)(\eta_w^n-\eta_u^n)\|^2_{L^2(I_j)}$$
$$+C\,K\|\phi(t^n)(\eta_u^{n+1}-(\eta_u^n+\eta_w^n)/2)\|^2_{L^2(I_j)}+M(\epsilon)k^6\|\phi(t^n)u_{ttt}(\theta^n)\|^2_{L^2(\mathbb{R})}.$$

Combining the estimates of T_1 and T_2 and using $\frac{h}{\bar{\bar{T}}}\le 1$, $\gamma\le 1$, and $\frac{1}{\bar{\bar{K}}}\le\frac{1}{K}$ concludes the proof. ∎

4.5. Step 6: Estimates of E_h. We state two slightly different estimates for $E_h(\phi_i)$. One would be applied with the weight ϕ_1 and the other for ϕ_2. To simplify notation, for the rest of this section we let $\phi=\phi_i$, $K=K_i$, $m=m_i$, and $\gamma=\gamma_i$ for $i=1$ or $i=2$.

LEMMA 4.8. *If K is sufficiently large, then*

$$E_h(\phi)\le(1+\lambda^{m-2})\frac{Ck}{K}\sum_j[\boldsymbol{\phi\xi}^n]_{j-1/2}\mathbb{I}[\boldsymbol{\phi\xi}^n]^t_{j-1/2}+\frac{k}{4}\Theta(|(\phi^2)_x|)$$
$$+\left\{\frac{k}{4T}+\left(1+\left(\frac{h}{\bar{\bar{T}}}\right)^{(3-4\gamma)}+\left(\frac{h}{\bar{\bar{T}}}\right)^{2-3\gamma}\right)\frac{C\,k}{K\,T}\right\}||\phi(t^n)\xi_u^n||^2_{L^2(\mathbb{R})}$$
$$+\left(1+\left(\frac{h}{\bar{\bar{T}}}\right)^{2(3/2-2\gamma)}+\left(\frac{h}{\bar{\bar{T}}}\right)^{2-3\gamma}\right)\frac{C\,T}{k}||\phi(t^n)(\eta_w^n-\eta_u^n)||^2_{L^2(\mathbb{R})}$$
$$+\frac{C\,T}{k}(||\phi(t^n)(\eta_u^{n+1}-\eta_u^n)||^2_{L^2(\mathbb{R})}+k^6||\phi(t^n)u_{ttt}(\theta^n)||^2_{L^2(\mathbb{R})})$$

and

$$E_h(\phi)\le\frac{Ck}{K}\sum_j[\boldsymbol{\phi\xi}^n]_{j-1/2}\mathbb{I}[\boldsymbol{\phi\xi}^n]^t_{j-1/2}+\frac{k}{4}\Theta(|(\phi^2)_x|)$$
$$+\left\{\frac{k}{4T}+\left(1+\left(\frac{h}{\bar{\bar{T}}}\right)^{2(3/2-2\gamma)}+\left(\frac{h}{\bar{\bar{T}}}\right)^{2-3\gamma}+\left(\frac{h}{\bar{\bar{T}}}\right)^{1-2\gamma}\right)\frac{C\,k}{K\,T}\right\}||\phi(t^n)\xi_u^n||^2_{L^2(\mathbb{R})}$$
$$+\left(1+\left(\frac{h}{\bar{\bar{T}}}\right)^{2(3/2-2\gamma)}+\left(\frac{h}{\bar{\bar{T}}}\right)^{2-3\gamma}+\left(\frac{h}{\bar{\bar{T}}}\right)^{1-2\gamma}\right)\frac{C\,T}{k}||\phi(t^n)(\eta_w^n-\eta_u^n)||^2_{L^2(\mathbb{R})}$$
$$+\frac{C\,T}{k}(||\phi(t^n)(\eta_u^{n+1}-\eta_u^n)||^2_{L^2(\mathbb{R})}+k^6||\phi(t^n)u_{ttt}(\theta^n)||^2_{L^2(\mathbb{R})}).$$

The proof of this lemma is contained in the appendix. Here we would like to point out that the main tool in proving this result is superapproximation. The superapproximation result involving weights are similar to the ones used in [16, 13]. For the sake of completeness we include the proof of the following superapproximation lemma in the appendix.

LEMMA 4.9. *If $P = P_-$ or $P = P_+$ and $l = 0, 1, 2$, then there exists $C > 0$ independent of h such that for all $v \in V_h$ the following estimates hold:*

$$\|\phi^{-1}(t)(\phi^2(t)v - P(\phi^2(t)v))\|_{L^2(I_j)} + h\|\phi^{-1}(t)(\phi^2(t)v - P(\phi^2(t)v))_x\|_{L^2(I_j)}$$
$$\leq \frac{Ch^{2-2\gamma}}{\tilde{T}^{2(1-\gamma)}\tilde{K}^2}\|\phi(t)v\|_{L^2(I_j)}$$

(4.15a)
$$+ \frac{Ch^{2-\gamma/2}}{\tilde{T}^{(1-\gamma)/2}\tilde{K}^{1/2}}\|(|\phi(t)\phi_x(t)|)^{1/2}v_x\|_{L^2(I_j)}, \|\phi^{-1}(t)(D_x^l(\phi^2(t))v$$
$$- P(D_x^l(\phi^2(t))v))\|_{L^2(I_j)} + h\|\phi^{-1}(t)(D_x^l(\phi^2(t))v - P(D_x^l(\phi^2(t))v))_x\|_{L^2(I_j)}$$

(4.15b)
$$\leq \left(\frac{Ch^{2-(l+2)\gamma}}{\tilde{T}^{(1-\gamma)(l+2)}\tilde{K}^{2+l}} + \frac{Ch^{1-(l+1)\gamma}}{\tilde{T}^{(1-\gamma)(l+1)}\tilde{K}^{l+1}}\right)\|\phi(t)v\|_{L^2(I_j)}.$$

Let us illustrate how we can use Lemma 4.9 to prove Lemma 4.8 by showing how we can bound one of the terms appearing in the definition of E_h.

LEMMA 4.10. *Let $t^n \leq t \leq t^{n+1}$ and $v_h \in V_h$, and then for $l = 0, 1, 2$,*

$$h^l|E_2(D_x^l(\phi^2(t))v_h)| + h^l|E_4(D_x^l(\phi^2(t))v_h)| \leq \frac{C\,k}{K^{l+1}}\sum_j [\boldsymbol{\phi}\boldsymbol{\xi}^n]_{j-1/2}\mathbb{I}[\boldsymbol{\phi}\boldsymbol{\xi}^n]_{j-1/2}^t$$
$$+ \left(1 + \left(\frac{h}{\tilde{T}}\right)^{2(l+1/2-(l+1)\gamma)} + \left(\frac{h}{\tilde{T}}\right)^{2(l+3/2-(l+2)\gamma)}\right)\frac{C\,k}{K^{l+1}\,T}\|\phi(t^n)v_h\|_{L^2(\mathbb{R})}^2$$

$$+ \frac{CT}{k}(\|\phi(t^n)(\eta_w^n - \eta_u^n)\|_{L^2(\mathbb{R})}^2 + \|\phi(t^n)(\eta_u^{n+1} - \eta_u^n)\|_{L^2(\mathbb{R})}^2 + k^6\|u_{ttt}(\theta^n)\|_{L^2(\mathbb{R})}^2). \tag{4.16}$$

If we take $l = 0$, $v_h = \xi_u^n$, and $t = t^n$, then we get the E_2 term appearing in the definition of E_h.

Proof. By using (4.13b) and the properties of P_+ we get

$$h^l|E_2(D_x^l(\phi^2(t))v_h)| = h^l\left|\sum_j(\xi_w^n - \xi_u^n, D_x^l(\phi^2(t))v_h - P_+(D_x^l(\phi^2(t))v_h))_j\right|$$
$$= h^l\left|\sum_j\left(\frac{6k}{h^2}[\xi_u^n]_{j-1/2} + \frac{12}{h^3}R_{1,j}\right)\right.$$
$$\left.\times\,(x - x_j, D_x^l(\phi^2(t))v_h - P_+(D_x^l(\phi^2(t))v_h))_j\right|.$$

By multiplying and dividing the last equation by $\phi(t^n, x_{j-1/2})$ and using (4.8e), we get

$$h^l|E_2(D_x^l(\phi^2(t))v_h)| \leq h^{l+3/2}\sum_j\left(\frac{6k}{h^2}|[\phi(t^n)\xi_u^n]_{j-1/2}| + \frac{12}{h^3}|R_{1,j}\phi(t^n, x_{j-/2})|\right)$$
$$\times\,||\phi^{-1}(t)(D_x^l(\phi^2(t))v_h - P_+(D_x^l(\phi^2(t))v_h))||_{L^2(I_j)}, \tag{4.17}$$

where we used the fact that $||x-x_j||_{L^2(I_j)} \le Ch^{3/2}$. If we use (4.15b), we see that

$$\begin{aligned}
&kh^{l-1/2}\sum_j |[\phi(t^n)\xi_u^n]_{j-1/2}|\ ||\phi^{-1}(t)(D_x^l(\phi^2(t))v_h - P_+(D_x^l(\phi^2(t))v_h))||_{L^2(I_j)}\\
&\le \left(\frac{C}{K^{l+1}}+\frac{C}{K^{l+2}}\right)k\sum_j [\phi(t^n)\xi_u^n]_{j-1/2}^2\\
&\left(\left(\frac{h}{\bar{\bar{T}}}\right)^{2(l+1/2-(l+1)\gamma)}\frac{C}{K^{l+1}}+\left(\frac{h}{\bar{\bar{T}}}\right)^{2(l+3/2-(l+2)\gamma)}\frac{C}{K^{l+2}}\right)\frac{k}{T}\|\phi(t^n)v_h\|_{L^2(\mathbb{R})}^2. \qquad (4.18)
\end{aligned}$$

By the Cauchy–Schwarz inequality we get $|R_{1,j}| \le Ch^{3/2}\|\eta_w^n - \eta_u^n\|_{L^2(I_j)}$. Hence,

$$\begin{aligned}
&h^{l-3/2}\sum_j |R_{1,j}\phi(t^n,x_{j-1/2})|\ ||\phi^{-1}(t)(D_x^l(\phi^2(t))v_h - P_+(D_x^l(\phi^2(t))v_h))||_{L^2(I_j)}\\
&\le \left(\frac{C}{K^{2(l+1)}}+\frac{C}{K^{2(l+2)}}\right)\frac{k}{T}\|\phi(t^n)v_h\|_{L^2(\mathbb{R})}^2\\
&\quad+\left(\left(\frac{h}{\bar{\bar{T}}}\right)^{2(1-\gamma)(l+1)}+\left(\frac{h}{\bar{\bar{T}}}\right)^{2(1-\gamma)(l+2)}\right)\frac{CT}{k}\|\eta_w^n-\eta_u^n\|_{L^2(I_j)}^2. \qquad (4.19)
\end{aligned}$$

Therefore, if we plug (4.18) and (4.19) into (4.17) and use $\frac{h}{\bar{\bar{T}}} \le 1$, $\gamma \le 1$, and $K \ge 1$, we can bound $h^l|E_2(D_x^l(\phi^2(t))v_h)|$ by the right-hand side of (4.16). In a similar fashion we can bound $h^l|E_4(D_x^l(\phi^2(t))v_h)|$ to conclude the proof. ∎

4.6. Step 7: Estimates for $\Theta((\phi^2)_x)$ and the final estimates for the error in one time step.

4.6.1. Case 1: The weight ϕ_1. In this case, the only estimate we need for $\Theta((\phi_1^2)_x)$ is simply

$$\Theta((\phi_1^2)_x) < 0, \qquad (4.20)$$

which follows from (4.8c). By collecting the estimates for Ψ_h and E_h and $\Theta((\phi_1^2)_x)$, we can give the right bound for the weighted ϕ_1 error produced after one time step.

THEOREM 4.11. *Let $0 < \epsilon < \frac{1}{3}$, and suppose that $\lambda \le 1/3 - \epsilon$. If $\gamma_1 = 2/3$, $m_1 = 2$, and*

$$\frac{1}{K_1} = c\,\epsilon \qquad (4.21)$$

for a sufficiently small fixed constant $c > 0$, then

$$\begin{aligned}
\|\phi_1(t^{n+1})\xi_u^{n+1}\|_{L^2(\mathbb{R})}^2 \le{}& \left(1+\frac{k}{T}\right)\|\phi_1(t^n)\xi_u^n\|_{L^2(\mathbb{R})}^2\\
&+C\left(\frac{T}{k}+\epsilon^{-1}\right)\|\phi_1(t^n)(\eta_u^{n+1}-\eta_u^n)\|_{L^2(\mathbb{R})}^2\\
&+C\left(\frac{T}{k}+\epsilon^{-1}\right)\|\phi_1(t^n)(\eta_w^n-\eta_u^n)\|_{L^2(\mathbb{R})}^2\\
&+C\left(\frac{T}{k}+\epsilon^{-1}\right)k^6\,\|\phi_1(t^n)u_{ttt}(\theta^n)\|_{L^2(\mathbb{R})}^2.
\end{aligned}$$

Proof. To simplify notation we let $\phi = \phi_1$, $\gamma = \gamma_1$, $m_1 = m$, and $K = K_1$. If we insert (4.20), the result of Lemma 4.7, and (4.15) into Lemma 4.4, we get

$$\begin{aligned}
&\|\phi(t^{n+1})\xi_u^{n+1}\|_{L^2(\mathbb{R})} + \frac{k}{4}\Theta(|(\phi^2)_x|) \\
&+ k\sum_j [\boldsymbol{\phi}\boldsymbol{\xi}^n]_{j-1/2}\left(\left(\frac{1}{2} - \frac{C}{K}(1+\lambda^{m-2})\right)\mathbb{I} - \mathbb{A}\right)[\boldsymbol{\phi}\boldsymbol{\xi}^n]^t_{j-1/2} \\
&+ \left\{1 + \frac{k}{4T} + \left(1 + \left(\frac{h}{\bar{\bar{T}}}\right)^{2(3/2-2\gamma)} + \left(\frac{h}{\bar{\bar{T}}}\right)^{2-3\gamma}\right)\frac{C\,k}{K\,T}\right\}||\phi(t^n)\xi_u^n||^2_{L^2(\mathbb{R})} \\
&+ \left\{CK + \left(1 + \left(\frac{h}{\bar{\bar{T}}}\right)^{2(3/2-2\gamma)} + \left(\frac{h}{\bar{\bar{T}}}\right)^{2-3\gamma}\right)\frac{C\,T}{k}\right\}||\phi(t^n)(\eta_w^n - \eta_u^n)||^2_{L^2(\mathbb{R})} \\
&+ \left\{CK + \frac{C\,T}{k}\right\}(||\phi(t^n)(\eta_u^{n+1} - \eta_u^n)||^2_{L^2(\mathbb{R})} + k^6||\phi(t^n)u_{ttt}(\theta^n)||^2_{L^2(\mathbb{R})}),
\end{aligned}$$

where we used the fact that $|(\phi^2)_x| = -(\phi^2)_x$ in this case where $\phi = \phi_1$. We see from the term

$$\left(\frac{h}{\bar{\bar{T}}}\right)^{2-3\gamma}\frac{C\,k}{K\,T}\|\phi(t^n)\xi_u^n\|^2_{L^2(\mathbb{R})}$$

that we are forced to take $\gamma = 2/3$. By substituting $\gamma = 2/3$ and $m = 2$ and using our hypothesis (4.21), we see that

$$\begin{aligned}
\|\phi(t^{n+1})\xi_u^{n+1}\|_{L^2(\mathbb{R})} + k\sum_j [\boldsymbol{\phi}\boldsymbol{\xi}^n]_{j-1/2}\mathbb{D}[\boldsymbol{\phi}\boldsymbol{\xi}^n]^t_{j-1/2} &\le \left(1 + \frac{k}{T}\right)\|\phi(t^n)\xi_u^n\|^2_{L^2(\mathbb{R})} \\
+ \left(\frac{CT}{k}1 + C\epsilon^{-1}\right)(\|\phi(t^n)(\eta_w^n - \eta_u^n)\|^2_{L^2(\mathbb{R})} &+ \|\phi(t^n)(\eta_u^{n+1} - (\eta_u^n + \eta_w^n)/2)\|^2_{L^2(\mathbb{R})}) \\
+ \left(\frac{CT}{k}1 + C\epsilon^{-1}\right)k^6\,\|\phi(t^n)u_{ttt}(\theta^n)\|^2_{L^2(\mathbb{R})},&
\end{aligned}$$

where

$$\mathbb{D} = \left(\left(\frac{1}{2} - \epsilon\right)\mathbb{I} - \mathbb{A}\right). \tag{4.22}$$

Here we used the fact that $\frac{h}{\bar{\bar{T}}} \le 1$.

The proof will be complete if we can show that

$$k\sum_j [\boldsymbol{\phi}\boldsymbol{\xi}^n]_{j-1/2}\mathbb{D}[\boldsymbol{\phi}\boldsymbol{\xi}^n]^t_{j-1/2} \ge 0.$$

This occurs when $\mathbb{D}$ is positive semidefinite. This is guaranteed by our hypothesis $\lambda \le 1/3 - \epsilon$. ∎

4.6.2. Case 2: The weight ϕ_2. In this case we do not have $\Theta((\phi_2^2)_x) < 0$, and therefore we must find an appropriate bound for this term. The bound actually is true for both ϕ_1 and ϕ_2.

LEMMA 4.12. *Let $\phi=\phi_i$, $K=K_i$, and $\gamma=\gamma_i$ for either $i=1$ or $i=2$. Then*

$$
\begin{aligned}
k\Theta((\phi^2)_x) &\le \frac{Ck}{K}\sum_j [\boldsymbol{\phi\xi}^n]_{j-1/2}\mathbb{I}[\boldsymbol{\phi\xi}^n]_{j-1/2}^t\\
&+\left\{\frac{k}{4\,T}+\left(1+\left(\frac{h}{\tilde{\bar{T}}}\right)^{(1-2\gamma)}+\left(\frac{h}{\tilde{T}}\right)^{(3-4\gamma)}+\left(\frac{h}{\tilde{\bar{T}}}\right)^{(5-6\gamma)}\right)\frac{C\,k}{K\,T}\right\}\|\phi(t^n)\xi_u^n\|^2_{L^2(\mathbb{R})}\\
&+\left(1+\left(\frac{h}{\bar{\tilde{T}}}\right)^{2(1/2-\gamma)}+\left(\frac{h}{\tilde{T}}\right)^{(3-4\gamma)}+\left(\frac{h}{\bar{\tilde{T}}}\right)^{(5-6\gamma)}\right)\frac{C\,T}{k}\|\phi(t^n)(\eta_w^n-\eta_u^n)\|^2_{L^2(\mathbb{R})}\\
&+\frac{C\,T}{k}(\|\phi(t^n)(\eta_u^{n+1}-\eta_u^n)\|^2_{L^2(\mathbb{R})}+\|\phi(t^n)u_{ttt}(\theta^n)\|^2_{L^2(\mathbb{R})}).
\end{aligned}
$$

Proof. By using (4.6a), we have

$$
\begin{aligned}
\frac{k}{2}\sum_j(\xi_w^n-\xi_u^n,(\phi^2(t^n))_x(\xi_w^n-\xi_u^n))_j &= \frac{k^2}{2}H(\xi_u^n,(\phi^2(t^n))_x(\xi_w^n-\xi_u^n))\\
&+\frac{k}{2}E_2((\phi^2(t^n))_x(\xi_w^n-\xi_u^n))+\frac{k}{2}E_3((\phi^2(t^n))_x(\xi_w^n-\xi_u^n)).
\end{aligned}
$$

The result now follows if we apply Lemmas 7.9, 4.10, 7.7, and 7.3. ∎

Now we can bound the weighted ϕ_2 error in one time step.

THEOREM 4.13. *Let $0<\epsilon<\frac{1}{3}$, and suppose that $\lambda\le 1/3-\epsilon$. If $\gamma_2=1/2$, $m_2=0$, and*

$$
\frac{1}{K_2}=c\,\epsilon \tag{4.23}
$$

for a sufficiently small fixed constant $c>0$, then

$$
\begin{aligned}
\|\phi_2(t^{n+1})\xi_u^{n+1}\|^2_{L^2(\mathbb{R})} &\le \left(1+\frac{k}{T}\right)\|\phi_2(t^n)\xi_u^n\|^2_{L^2(\mathbb{R})}\\
&+C\left(\frac{T}{k}+\epsilon^{-1}\right)\|\phi_2(t^n)(\eta_u^{n+1}-\eta_u^n)\|^2_{L^2(\mathbb{R})}\\
&+C\left(\frac{T}{k}+\epsilon^{-1}\right)\|\phi_2(t^n)(\eta_w^n-\eta_u^n)\|^2_{L^2(\mathbb{R})}\\
&+C\left(\frac{T}{k}+\epsilon^{-1}\right)k^6\,\|\phi_2(t^n)u_{ttt}(\theta^n)\|^2_{L^2(\mathbb{R})}.
\end{aligned}
$$

Proof. Again to simplify notation we let $\phi=\phi_2$, $\gamma=\gamma_2$, $m=m_2$, and $K=K_2$. If we insert (4.15) and the results of Lemmas 4.12 and 4.7 into Lemma 4.4, we get

$$
\begin{aligned}
&\|\phi(t^{n+1})\xi_u^{n+1}\|_{L^2(\mathbb{R})}+k\sum_j[\boldsymbol{\phi\xi}^n]_{j-1/2}\left(\left(\frac{1}{2}-\frac{C}{K}\right)\mathbb{I}-\mathbb{A}\right)[\boldsymbol{\phi\xi}^n]_{j-1/2}^t\\
&\le\left(1+\frac{k}{2\,T}+\frac{C\,k}{K\,T}\right)\|\phi(t^n)\xi_u^n\|^2_{L^2(\mathbb{R})}\\
&\quad+\left(C\,K+\frac{C\,T}{k}\right)(\|\phi(t^n)(\eta_w^n-\eta_u^n)\|^2_{L^2(\mathbb{R})}+\|\phi(t^n)(\eta_u^{n+1}-\eta_u^n)\|^2_{L^2(\mathbb{R})})\\
&\quad+\left(C\,K+\frac{C\,T}{k}\right)k^6\|\phi(t^n)u_{ttt}(\theta^n)\|^2_{L^2(\mathbb{R})},
\end{aligned}
$$

where we used $\frac{h}{T} \le 1$ and our hypothesis $\gamma = 1/2$ and $m = 0$. In fact, we needed only our original assumption $m \ge 0$ for this inequality. We choose $m = 0$, which minimizes the numerical layer. If we now use our hypothesis (4.23), we see that

$$\begin{aligned}\|\phi_2(t^{n+1})\xi_u^{n+1}\|^2_{L^2(\mathbb{R})} + k\sum_j [\boldsymbol{\phi\xi}^n]_{j-1/2}\mathbb{D}[\boldsymbol{\phi\xi}^n]^t_{j-1/2} &\le \left(1+\frac{k}{T}\right)\|\phi_2(t^n)\xi_u^n\|^2_{L^2(\mathbb{R})}\\ &+ \left(\frac{C\,T}{k} + C\epsilon^{-1}\right)\|\phi(t^n)(\eta_u^{n+1}-\eta_u^n)\|^2_{L^2(\mathbb{R})}\\ &+ \left(\frac{C\,T}{k} + C\epsilon^{-1}\right)\|\phi(t^n)(\eta_w^{n}-\eta_u^n)\|^2_{L^2(\mathbb{R})}\\ &+ \left(\frac{C\,T}{k} + C\epsilon^{-1}\right)k^6\,\|\phi(t^n)u_{ttt}(\theta^n)\|^2_{L^2(\mathbb{R})},\end{aligned}$$

where $\mathbb{D}$ is defined in (4.22). By our hypothesis $\lambda \le 1/3 - \epsilon$, $\mathbb{D}$ is positive definite. This proves Theorem 4.13. ∎

4.7. Step 9: Conclusion of the proof of Theorem 2.1. Let K_1 and ϕ_1 be as in Theorem 4.11. Choose $x_1 = -2sK_1\log(1/h)\lambda^{-7/3}T^{1/3}h^{2/3}$ for some $s > 0$, where x_1 appears in the definition of ϕ_1. We choose s sufficiently large below. Since $x_1 < -h$,

$$\begin{aligned}&\|U(T) - u_h^N\|_{L^2(-\infty,T+x_1)} = \|u(T) - u_h^N\|_{L^2(-\infty,T+x_1)}\\ &\le \|\eta_u^N\|_{L^2(-\infty,T+x_1)} + \|\xi_u^N\|^2_{L^2(-\infty,T+x_1)}\\ &= ||P_-(U(T)) - U(T)||_{L^2(-\infty,T+x_1)} + \|\xi_u^N\|_{L^2(-\infty,T+x_1)}.\end{aligned}$$

By (4.8a), we have $\|\xi_u^N\|_{L^2(-\infty,T+x_1)} \le C\|\phi_1(t^N)(\xi_u^N)\|_{L^2(\mathbb{R})}$. Hence, we need only to bound $\|\phi_1(t^N)\xi_u^N\|_{L^2(\mathbb{R})}$. Now we apply Theorem 4.11 for $n \le N$, and

$$\begin{aligned}\|\phi_1(t^n)\xi_u^n\|^2_{L^2(\mathbb{R})} &\le \left(1+\frac{k}{T}\right)\|\phi_1(t^{n-1})\xi_u^{n-1}\|^2_{L^2(\mathbb{R})}\\ &\quad+\left(\frac{C\,T}{k} + C\,\epsilon^{-1}\right)(\|\phi_1(t^{n-1})(\eta_u^n-\eta_u^{n-1})\|^2_{L^2(\mathbb{R})}\\ &\quad+\|\phi_1(t^{n-1})(\eta_w^{n-1}-\eta_u^{n-1})\|^2_{L^2(\mathbb{R})})\\ &\quad+\left(\frac{C\,T}{k} + C\,\epsilon^{-1}\right)k^6\,\|\phi_1(t^{n-1})u_{ttt}(\theta^{n-1})\|^2_{L^2(\mathbb{R})}.\end{aligned}$$

We decompose the real line as $\mathbb{R} = S_n \cup S_n^c$, where $S_n = (-\infty,\ t^n + x_1 + sK_1\log(1/h)\lambda^{-7/3}T^{1/3}h^{2/3}) = (-\infty,\ t^n - sK_1\log(1/h)\lambda^{-7/3}T^{1/3}h^{2/3})$. Notice that $u(\cdot,t^n) = U(\cdot,t^n)$ in S_n. Hence, by the approximation properties of the projection operator P_-, we have

$$\begin{aligned}&\left(\frac{C\,T}{k} + C\,\epsilon^{-1}\right)\|\phi_1(t^{n-1})(\eta_u^n-\eta_u^{n-1})\|^2_{L^2(S_{n-1})}\\ &\quad+\left(\frac{C\,T}{k} + C\,\epsilon^{-1}\right)\|\phi_1(t^{n-1})(\eta_w^{n-1}-\eta_u^{n-1})\|^2_{L^2(S_{n-1})}\\ &\quad+\left(\frac{C\,T}{k} + C\,\epsilon^{-1}\right)k^6\,\|\phi_1(t^{n-1})u_{ttt}(\theta^n)\|^2_{L^2(S_{n-1})} \le \left(\frac{C\,T}{k} + C\,\epsilon^{-1}\right)k^2\,h^4.\end{aligned}$$

It follows from (4.8b) that $\phi(x,t^{n-1}) \le h^s$ for $x \in \mathbb{R}\backslash S_{n-1}$. By using (4.2) and choosing $s \ge 4$, we can easily show that

$$\left(\frac{C\,T}{k} + C\,\epsilon^{-1}\right) \|\phi(t^{n-1})(\eta_u^n - \eta_u^{n-1})\|^2_{L^2(\mathbb{R}\backslash S_{n-1})}$$
$$+ \left(\frac{C\,T}{k} + C\,\epsilon^{-1}\right) \|\phi(t^{n-1})(\eta_w^{n-1} - \eta_u^{n-1})\|^2_{L^2(\mathbb{R}\backslash S_{n-1})}$$
$$+ \left(\frac{C\,T}{k} + C\,\epsilon^{-1}\right) k^6\, \|\phi(t^{n-1})u_{ttt}(\theta^n)\|^2_{L^2(\mathbb{R}\backslash S_{n-1})} \le \left(\frac{C\,T}{k} + C\,\epsilon^{-1}\right) k^2\, h^4.$$

Hence,

$$\|\phi_1(t^n)\xi_u^n\|^2_{L^2(\mathbb{R})} \le \left(1 + \frac{k}{T}\right) \|\phi_1(t^{n-1})\xi_u^{n-1}\|^2_{L^2(\mathbb{R})} + \left(\frac{C\,T}{k} + C\,\epsilon^{-1}\right) k^2\, h^4,$$

and it follows by Gronwall's inequality that

$$\|\phi_1(T)\xi_u^N\|^2_{L^2(\mathbb{R})} \le (1+e)\|\phi_1(0)\xi_u^0\|^2_{L^2(\mathbb{R})} + \left(\frac{C\,T}{k} + C\,\epsilon^{-1}\right) T\,k\,h^4$$
$$= (1+e)\|\phi_1(0)(P_-(U_0) - P_-(u_0))\|^2_{L^2(\mathbb{R})} + \left(\frac{C\,T}{k} + C\,\epsilon^{-1}\right) T\,k\,h^4.$$

Again, by using the fact that $U_0 = u_0$ in $(-\infty, -h)$ and using the decay properties of $\phi_1(0)$, we have

$$\|\phi_1(t^N)\xi_u^N\|^2_{L^2(\mathbb{R})} \le \left(\frac{C\,T}{k} + C\,\epsilon^{-1}\right) T\,k\,h^4 + Ch^s.$$

Therefore, by using this result and (4.8a) we get

$$\|U(T) - u_h^N\|_{L^2(-\infty,\, T - 2\,s\,K_1\,\log(1/h)\lambda^{-7/3}T^{1/3}h^{2/3})} \le CTh^2 + C\left(\frac{Tk}{\epsilon}\right)^{1/2} h^2 + Ch^s$$
$$+ \|P_-(U(T)) - U(T)\|_{L^2(-\infty,\, T - 2\,s\,K_1\,\log(1/h)\lambda^{-7/3}T^{1/3}h^{2/3})}.$$

Since $K_1 = \frac{1}{c\,\epsilon}$, we can choose $\beta = \frac{2}{c}$ to obtain

$$\|U(T) - u_h^N\|_{L^2(-\infty,\, T - \beta\,s\log(1/h)\lambda^{-7/3}\epsilon^{-1}T^{1/3}h^{2/3})} \le CTh^2 + C\left(\frac{Tk}{\epsilon}\right)^{1/2} h^2 + Ch^s$$
$$+ \|P_-(U(T)) - U(T)\|_{L^2(-\infty,\, T - \beta\,s\log(1/h)\lambda^{-7/3}\epsilon^{-1}T^{1/3}h^{2/3})}.$$

The estimate in the region $x > T + \beta\,s\log(1/h)\lambda^{-1/2}\epsilon^{-1}T^{1/2}h^{1/2}$ can be established in a similar fashion.

5. Numerical eperiments. The purpose of this section is to verify our main result (Theorem 2.1) and to present numerical evidence suggesting that it is sharp.

5.1. L^2-errors. To verify the estimate of Theorem 2.1, we consider the problem (1.1) with periodic initial condition $U_0(x) = \sin(2\pi x) + \chi(x)$ for $x \in [0,1]$, where χ is the characteristic function of the interval $(\frac{1}{4}, \frac{3}{4})$. We use uniform spatial meshes and uniform time stepping and take $CFL = \lambda = k/h = 0.33$.

TABLE 5.1
Error to the left and to the right of the singularity.

ℓ	eL_ℓ	oL_ℓ	eR_ℓ	oR_ℓ
1	.44e-8	-	.58e-8	-
2	.11e-8	1.98	.15e-8	1.96
3	.28e-9	1.99	.38e-9	1.98
4	.70e-10	1.99	.95e-10	1.99
5	.18e-10	2.00	.24e-10	1.99
6	.44e-11	2.00	.60e-11	1.99
7	.20e-11	2.00	.15e-11	2.00

We will investigate the error at time $T = 1$. The mesh ℓ will denote a mesh with size $h_l = \frac{2^{-\ell}}{1000}$. We compute the L^2-error to the left and right of the singularity $x = \frac{1}{4}$ and $T = 1$. Specifically, we compute the following errors:

$$eL_\ell := \|U(T) - u_h(T)\|_{L^2(0,\frac{1}{4}-5h_l^{2/3})},$$
$$eR_\ell := \|U(T) - u_h(T)\|_{L^2(\frac{1}{4}+5h_l^{1/2},\frac{7}{10})}.$$

We then compute the orders of convergence defined by

$$oL_\ell := \frac{\log(\frac{eL_\ell}{eL_{\ell+1}})}{\log(2)},$$
$$oR_\ell := \frac{\log(\frac{eR_\ell}{eR_{\ell+1}})}{\log(2)}.$$

We list the results in Table 5.1. As we can see, the optimal orders of convergence are realized; this confirms the prediction of Theorem 2.1. Note that we did not use a logarithmic factor in this computational experiment. We simply took the region around the discontinuity as $\mathcal{R}_T = \frac{1}{4} + (-5\,h^{2/3}, 5\,h^{1/2})$.

5.2. Behavior of the error near discontinuities. To show that our results are sharp, we explore a very precise form of the error near the discontinuity for the special periodic initial condition $U_0(x) = \chi(x)$ for $x \in [0, 1]$.

We would study the error *near* the discontinuity $x = 1/4$ and $T = 1$. We do this by scaling the error near the discontinuity and plotting the results for different h. To the left of the discontinuity, for each fix mesh size h we plot the scaled error

$$\text{errorL}(y, h) := \left|U(1/4 + h^{2/3}\, y, T) - u_h(1/4 + h^{2/3}\, y, T)\right|.$$

Note that here $y \le 0$. The results are given in the top panel of Figure 5.1. Note that as we decrease h the graphs seem to be converging. This reflects the fact that the scaling $h^{2/3}$ is the correct one and that our results are sharp.

Similarly, to the right of the discontinuity we define, for $y \ge 0$,

$$\text{errorR}(y, h) := \left|U(1/4 + h^{1/2}\, y, T) - u_h(1/4 + h^{1/2}\, y, T)\right|.$$

Since the above function is very oscillatory, we are going to consider its "envelope" which we denote by "envelopeR(y, h)"; see the middle panel of Figure 5.1. In the bottom panel of Figure 5.1, we plot those envelopes for various h in Figure 5.1. Notice that the graphs seem to be converging as we decrease h. Again, this reflects the fact that the scaling $h^{1/2}$ is the correct one and that our results are sharp.

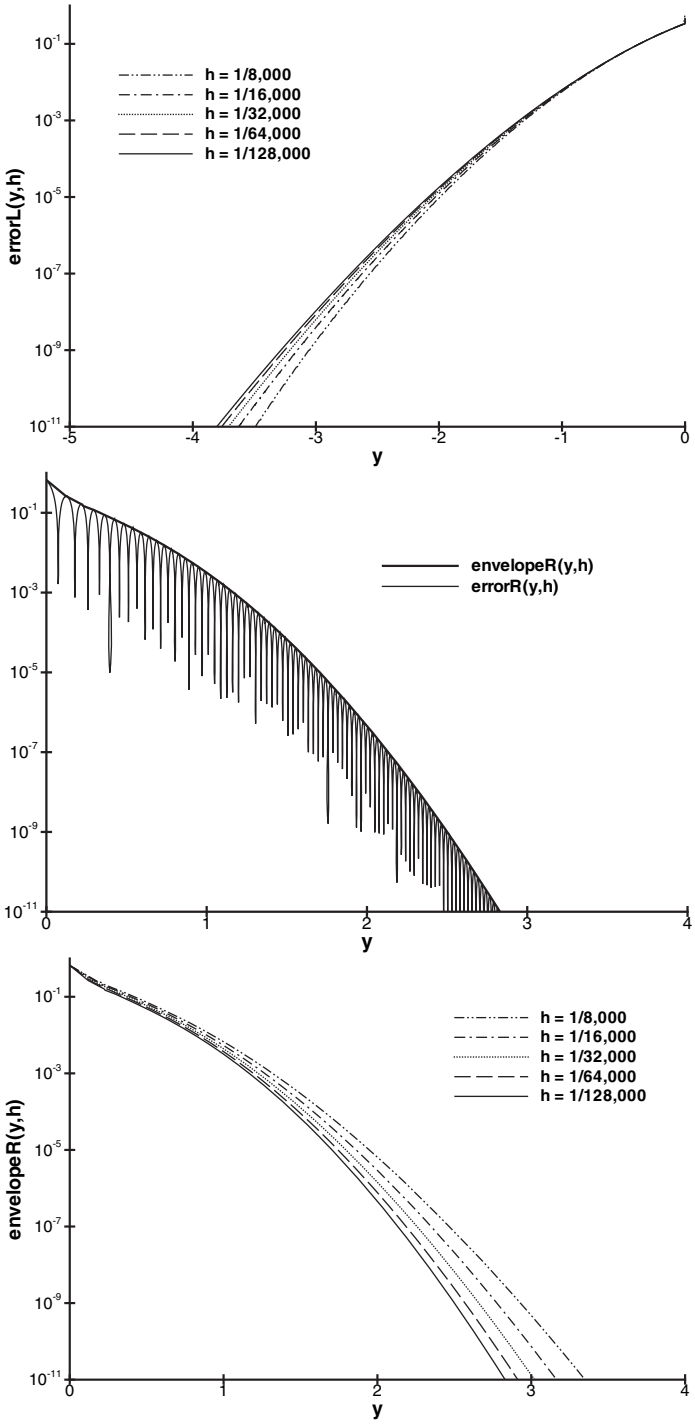


FIG. 5.1. *History of convergence of the scaled error to the left of the discontinuity (top) and of the envelope of the scaled error to the right of the discontinuity (bottom) for different values of the mesh size h. In the middle, we see an example of the scaled error to the right of the discontinuity and its envelope.*

6. Concluding remarks. The approach developed in this paper could be applied to the study of general RKDG methods for the model problem under consideration. The only difficulty in carrying out such a task is that the use of high-order accurate Runge–Kutta methods quickly complicates the calculations. If we solve exactly in time, however, a characterization of the region containing the numerical layer should be fairly easy to obtain for any given polynomial degree p for the DG discretization in space. This constitutes the subject of ongoing work.

7. Appendix.

7.1. Construction of weights. To finish the construction of the weights ϕ_1 and ϕ_2, we need only to construct the initial conditions b_1 and b_2:

$$b_1(x) = \psi\left((x - x_1)/\tilde{K}_1\tilde{T}^{1-\gamma_1}\, h^{\gamma_1}\right) \quad \text{and} \quad b_2(x) = \psi\left((x_2 - x)/\tilde{K}_2\tilde{T}^{1-\gamma_2}\, h^{\gamma_2}\right).$$

The function ψ is given by

$$\psi(r) = \int_r^\infty e^{-g(s)} ds,$$

where $g \in C^5(\mathbb{R})$ and is such that $g(s) = |s|$ for $|s| \geq 1$.

Now we can prove Proposition 4.3.

Proof. Since we have the following representation of ϕ for $i = 1, 2$:

$$\phi_1(x,t) = b_1(x-t) = \psi\left(((x-t) - x_1)/\tilde{K}_1\tilde{T}^{1-\gamma_1}\, h^{\gamma_1}\right),$$
$$\phi_2(x,t) = b_2(x-t) = \psi\left((x_2 - (x-t))/\tilde{K}_2\tilde{T}^{1-\gamma_2}\, h^{\gamma_2}\right),$$

it is not difficult to show that these results can be easily obtained by using the following properties of ψ:

$$\begin{aligned} c \leq \psi(r) \leq C \ \text{ for } \ r \leq 1, \\ \psi'(x) < 0 \ \text{ for } \ x \in \mathbb{R} \\ \psi(r) = e^{-r} \quad \text{for } \ r > 1, \\ |\psi'(r)| + |\psi''(r)| + |\psi'''(r)| \leq C|\psi(r)| \quad \text{for } \ r \in \mathbb{R}, \\ |\psi''(r)| + |\psi'''(r)| + |\psi''''(r)| \leq C|\psi'(r)| \quad \text{for } \ r \in \mathbb{R} \end{aligned}$$

and that, for any interval I of unit length, $RO(I,\psi) + RO(I,\psi') \leq C$. This completes the sketch of the proof. ∎

7.2. Proof of Lemma 4.9. The proof is very similar to the superapproximation proofs contained in [16, 13].

Proof. To simplify notation we drop t. By the approximation properties of P, we have

$$\begin{aligned} &\|D_x^l(\phi^2)v - P(D_x^l(\phi^2)v)\|_{L^2(I_j)} + h\|(D_x^l(\phi^2)v)_x - P(D_x^l(\phi^2)v)_x\|_{L^2(I_j)} \\ &\leq Ch^2\|(D_x^l(\phi^2)v)_{xx}\|_{L^2(I_j)}. \end{aligned}$$

Since $v_{xx} = 0$, we have

$$h^2\|(D_x^l(\phi^2)v)_{xx}\|_{L^2(I_j)} \leq h^2\|D_x^{l+2}(\phi^2)v\|_{L^2(I_j)} + 2h^2\|D_x^{l+1}(\phi^2)v_x\|_{L^2(I_j)}.$$

By using (4.8d), we obtain

$$\begin{aligned}&\|D_x^l(\phi^2)v - P(D_x^l(\phi^2)v)\|_{L_h^2(I_j)} + h\|(D_x^l(\phi^2)v)_x - P(D_x^l(\phi^2)v)_x\|_{L^2(I_j)}\\ &\le \frac{Ch^{2-(l+1)\gamma}}{\tilde{T}^{(1-\gamma)(l+2)}\tilde{K}^{l+2}}\|\phi^2 v\|_{L^2(I_j)} + \frac{h^{2-\frac{1}{2}(l+1)\gamma}}{\bar{T}^{(1-\gamma)(l+1/2)}K^{(l+1/2)}}\|\phi|\phi\phi_x|^{1/2}v_x\|_{L^2(I_j)}.\end{aligned} \tag{7.1}$$

If we multiply both sides by ϕ^{-1} (here we use (4.8e)) and set $l = 0$, we arrive at (4.15a). The inequality (4.15b) follows from (7.1), (4.8d), (4.8e), and (7.4). ∎

7.3. Proof of Lemma 4.5. We start by proving (4.13b). If χ_j is the characteristic function of I_j, then

$$\overline{\xi_w^n - \xi_u^n}(x_j) = \frac{1}{h}(\xi_w^n - \xi_u^n, 1)_j = \frac{k}{h}H(\xi_u^n, \chi_j) + \frac{1}{h}(E_2(\chi_j) + E_3(\chi_j)).$$

Since $\chi_j \in V_h$, $E_2(\chi_j) = 0$. By the properties of P_- we have $E_3(\chi_j) = 0$. Also, from the definition of H we have $H(\xi_u^n, \chi_j) = -h(\xi_u^n)_x(x_j) - [\xi_u^n]_{j-1/2}$. This shows that

$$\overline{\xi_w^n - \xi_u^n}(x_j) = -k(\xi_u^n)_x(x_j) - \frac{k}{h}[\xi_u^n]_{j-1/2}. \tag{7.2}$$

By using (4.12), we have

$$\begin{aligned}(\xi_w^n - \xi_u^n)_x(x_j) &= (\xi_w^n - \xi_u^n, x - x_j)_j / \int_{I_j}(x - x_j)^2 dx\\ &= \frac{12}{h^3}\{kH(\xi_u^n, \chi_j(x - x_j)) + E_2(\chi_j(x - x_j)) + E_3(\chi_j(x - x_j))\}.\end{aligned}$$

We used the fact that $\int_{I_j}(x - x_j)^2 = \frac{h^3}{12}$.

Again, since $\chi_j(x - x_j) \in V_h$, we see that $E_2(\chi_j(x - x_j)) = 0$. By the definition of E_3, we have $E_3(\chi_j(x - x_j)) = R_{1,j}$. Also, $H(\xi_u^n, \chi_j(x - x_j)) = -\int_{I_j}(x - x_j)(\xi_u^n)_x dx - (x - x_j)(x_{j-1/2}^+)[\xi_u^n]_{j-1/2} = \frac{h}{2}[\xi_u^n]_{j-1/2}$, since $(\xi_u^n)_x$ is constant on I_j and $\int_{I_j}(x - x_j)dx = 0$.

Therefore,

$$(\xi_w^n - \xi_u^n)_x(x_j) = \frac{6k}{h^2}[\xi_u^n]_{j-1/2} + \frac{12}{h^3}R_{1,j}.$$

Equation (4.13b) now follows from (4.12), (7.2), and the above identity.

By subtracting (4.6a) from (4.6b) we have

$$\begin{aligned}(\xi_u^{n+1} - \xi_w^n, v) = \frac{k}{2}H(\xi_w^n - \xi_u^n, v) + E_1(P_+(v)) - E_2(v)/2 - E_3(v)/2\\ + E_4(v) + E_5(v).\end{aligned}$$

By using this representation and following the techniques used to prove (4.13b), we can easily show (4.13a).

7.4. Proof of Lemma 4.8. In order to prove Lemma 4.8 we will state and prove a series of lemmas. Most of these lemmas are bounds for the terms of E_h. We first state standard inverse estimates.

LEMMA 7.1. *There exists a $C > 0$ such that for every $v \in P^1(I_j)$*

$$|v(x_{j+1/2})| + |v(x_{j-1/2})| \le Ch^{-1/2}\|v\|_{L^2(I_j)} \tag{7.3}$$

and

$$\|v'\|_{L^2(I_j)} \le Ch^{-1}\|v\|_{L^2(I_j)}. \tag{7.4}$$

We also need the following preliminary estimates.

LEMMA 7.2. *Let $t^n \le t \le t^{n+1}$; then for $v_h \in V_h$*

$$\|\phi^{-1}(t)(P_+(\phi^2(t)v_h) - \phi^2(t)v_h)\|_{L^2(I_j)} \le \frac{C}{K}\|\phi(t^n)v_h\|_{L^2(I_j)}, \tag{7.5}$$

and for K sufficiently large

$$\|\phi^{-1}(t)P_+(\phi^2(t)v_h)\|_{L^2(I_j)} \le 2\|\phi(t^n)v_h\|_{L^2(I_j)}. \tag{7.6}$$

Proof. By (4.15b) we have

$$\|\phi^{-1}(t)(P_+(\phi^2(t)v_h) - \phi^2(t)v_h)\|_{L^2(I_j)} \le \left(\left(\frac{h}{\bar{T}}\right)^{1-\gamma}\frac{C}{K} + \left(\frac{h}{\bar{T}}\right)^{2(1-\gamma)}\frac{C}{K^2}\right)\|\phi(t^n)v_h\|_{L^2(I_j)}.$$

The inequality (7.5) now follows from $K \ge 1$ and $\frac{h}{\bar{T}} \le 1$. The inequality (7.6) easily follows from (7.5). ∎

The following result compares ξ_w^n to ξ_u^n.

LEMMA 7.3. *If $t^n \le t \le t^{n+1}$ and K is sufficiently large, then*

$$\|\phi(t)\xi_w^n\|^2_{L^2(\mathbb{R})} \le C\|\phi(t^n)\xi_u^n\|^2_{L^2(\mathbb{R})} + C\|\phi(t^n)(\eta_w^n - \eta_u^n)\|^2_{L^2(\mathbb{R})}.$$

Proof. By using (4.6a), we have

$$\|\phi(t)\xi_w^n\|^2_{L^2(\mathbb{R})} = \sum_j (\xi_u^n, \phi^2(t)\xi_w^n)_j + kH(\xi_u^n, \phi^2(t)\xi_w^n) + E_2(\phi^2(t)\xi_w^n) + E_3(\phi^2(t)\xi_w^n).$$

By applying (4.8e) and the Cauchy–Schwarz inequality, we get

$$\sum_j (\xi_u^n, \phi^2(t)\xi_w^n)_j \le C\|\phi(t^n)\xi_u^n\|^2_{L^2(\mathbb{R})} + \frac{1}{8}\|\phi(t)\xi_w^n\|^2_{L^2(\mathbb{R})}.$$

If we use (4.8e), (7.3), and (7.4), we get

$$kH(\xi_u^n, \phi^2(t)\xi_w^n) \le C\|\phi(t^n)\xi_u^n\|^2_{L^2(\mathbb{R})} + \frac{1}{8}\|\phi(t)\xi_w^n\|^2_{L^2(\mathbb{R})}.$$

By using (7.5) and taking K sufficiently large, we get

$$E_2(\phi^2(t^n)\xi_w^n) \le C\|\phi(t^n)\xi_u^n\|^2_{L^2(\mathbb{R})} + \frac{1}{8}\|\phi(t)\xi_w^n\|^2_{L^2(\mathbb{R})}.$$

If we use (7.6), we get

$$E_3(\phi^2(t)\xi_w^n) \le C\|\phi(t^n)\xi_u^n\|^2_{L^2(\mathbb{R})} + \frac{1}{8}\|\phi(t)\xi_w^n\|^2_{L^2(\mathbb{R})}.$$

Combining the last four inequalities proves Lemma 7.3. ∎

We will need a different bound than what is given in Lemma 4.10 for the case $l = 0$ and $v_h = \xi_u^n$ or $v_h = \xi_w^n$.

LEMMA 7.4. *Let $t^n \le t \le t^{n+1}$, and then*

$$\begin{aligned}
&|E_2(\phi^2(t)\xi_u^n)| + |E_4(\phi^2(t)\xi_w^n)| \\
&\le (1+\lambda^{m-2})\frac{Ck}{K}\sum_j [\boldsymbol{\phi\xi}^n]_{j-1/2}\mathbb{I}[\boldsymbol{\phi\xi}^n]_{j-1/2}^t + \frac{k}{6}\Theta(|(\phi^2)_x|) \\
&\quad + \left(1+\left(\frac{h}{\tilde{T}}\right)^{2(3/2-2\gamma)}\right)\frac{C\,k}{K\,T}||\phi(t^n)\xi_u^n||_{L^2(\mathbb{R})}^2 \\
&\quad + \left(1+\left(\frac{h}{\tilde{T}}\right)^{2(3/2-2\gamma)}\right)\frac{T}{k}||\phi(t^n)(\eta_w^n-\eta_u^n)||_{L^2(\mathbb{R})}^2 \\
&\quad + \frac{C\,T}{k}(||\phi(t^n)(\eta_u^{n+1}-\eta_u^n)||_{L^2(\mathbb{R})}^2 + k^6||\phi(t^n)u_{ttt}(\theta^n)||_{L^2(\mathbb{R})}^2).
\end{aligned}$$

Proof. As was done in the proof of Lemma 4.10, we see that

$$\begin{aligned}
|E_2(\phi^2(t)\xi_u^n)| &= \left|\sum_j\left(\frac{6k}{h^2}[\xi_u^n]_{j-1/2}(\phi(t)(x-x_j),\phi^{-1}(t)(\phi^2(t)\xi_u^n - P_+(\phi^2(t)\xi_u^n)))_j\right)\right. \\
&\qquad \left. + \sum_j \frac{12}{h^3}R_{1,j}(\phi(t)(x-x_j),\phi^{-1}(t)(\phi^2(t)\xi_u^n - P_+(\phi^2(t)\xi_u^n)))_j\right| \\
&\le M_1 + M_2,
\end{aligned}$$

where

$$\begin{aligned}
M_1 &:= \frac{C\,k}{h^{1/2}}\sum_j |[\phi(t^n)\xi_u^n]_{j-1/2}|\,\|\phi^{-1}(t)(\phi^2(t)\xi_u^n - P_+(\phi^2(t)\xi_u^n))\|_{L^2(I_j)}, \\
M_2 &:= \frac{C}{h^{3/2}}\sum_j |R_{1,j}\phi(t^n,x_{j-1/2})|\,\|\phi^{-1}(t)(\phi^2(t)\xi_u^n - P_+(\phi^2(t)\xi_u^n))\|_{L^2(I_j)}.
\end{aligned}$$

We first bound M_2. By using (4.19) with $v_h = \xi_u^n$ and $l = 0$ combined with the inequalities $K \ge 1$, $\frac{h}{\tilde{T}} \le 1$, and $\gamma \le 1$, we have

$$M_2 \le \frac{C\,k}{K\,T}\|\phi(t^n)\xi_u^n\|_{L^2(\mathbb{R})}^2 + \frac{C\,T}{k}\|\phi(t^n)(\eta_w^n-\eta_u^n)\|_{L^2(I_j)}^2. \tag{7.7}$$

For the next term we use (4.8e) and (4.15a) to obtain

$$\begin{aligned}
M_1 &\le \frac{Ckh^{3/2-2\gamma}}{\tilde{T}^{2(1-\gamma)}\tilde{K}^2}\sum_j |[\phi(t^n)(\xi_u^n)]_{j-1/2}|\,\|\phi(t^n)\xi_u^n\|_{L^2(I_j)} \\
&\quad + \frac{Ckh^{3/2-\gamma/2}}{\tilde{T}^{(1-\gamma)/2}\tilde{K}^{1/2}}\sum_j |[\phi(t^n)(\xi_u^n)]_{j-1/2}|\,\|(|\phi(t^n)(\phi)_x(t^n)|)^{1/2}(\xi_u^n)_x\|_{L^2(I_j)}.
\end{aligned}$$

By using Young's inequality, we have

$$\begin{aligned}
&\frac{Ckh^{3/2-2\gamma}}{\tilde{T}^{2(1-\gamma)}\tilde{K}^2}\sum_j |[\phi(t^n)(\xi_u^n)]_{j-1/2}|\,\|\phi(t^n)\xi_u^n\|_{L^2(I_j)} \\
&\le \frac{C}{K^2}k\sum_j [\phi(t^n)\xi_u^n]_{j+1/2}^2 + \left(\frac{h}{\tilde{T}}\right)^{2(3/2-2\gamma)}\frac{C\,k}{K^2\,T}\|\phi(t^n)\xi_u^n\|_{L^2(\mathbb{R})}^2.
\end{aligned} \tag{7.8}$$

If we use (4.13b), the triangle inequality, and the fact that $|x-x_j| \le h$ for $x \in I_j$, then

$$\|(|\phi(t^n)(\phi)_x(t^n)|)^{1/2}(\xi_u^n)_x\|_{L^2(I_j)} \le \frac{1}{k}\Theta_j(|(\phi^2)_x|)^{1/2} + \frac{Ch^{-(1+\gamma)/2}}{\tilde{K}^{1/2}\tilde{T}^{(1-\gamma)/2}}|[\phi(t^n)\xi_u^n]_{j-1/2}| + \frac{Ch^{-\gamma/2}}{k\tilde{K}^{1/2}\tilde{T}^{(1-\gamma)/2}}||\phi(t^n)(\eta_w^n-\eta_u^n)||_{L^2(I_j)},$$

where we also used $|R_{1,j}| \le h^{3/2}||\eta_w^n-\eta_u^n||_{L^2(I_j)}$, (4.8d), and (4.8e).

Therefore,

$$\begin{aligned}
&\frac{Ckh^{3/2-\gamma/2}}{\tilde{T}^{(1-\gamma)/2}\tilde{K}^{1/2}}\sum_j |[\phi(t^n)(\xi_u^n)]_{j-1/2}|\ \|(|\phi(t^n)(\phi)_x(t^n)|)^{1/2}(\xi_u^n)_x\|_{L^2(I_j)}\\
&\le \left(\frac{C\lambda^{-2}}{\tilde{K}}\left(\frac{h}{\tilde{T}}\right)^{1-\gamma} + \frac{C}{\tilde{K}}\left(\frac{h}{\tilde{T}}\right)^{2(1-\gamma)} + \frac{C\lambda^{-1}}{\tilde{K}^2}\left(\frac{h}{\tilde{T}}\right)^{2(3/2-\gamma)}\right)k\sum_j[\phi(t^n)\xi_u^n]_{j-1/2}^2\\
&\quad + \frac{k}{12}\Theta(|(\phi^2)_x|) + \frac{T}{k}||\eta_w^n-\eta_u^n||_{L^2(\mathbb{R})}^2\\
&\le (\lambda^{m-2}+\lambda^m+\lambda^{2m-1})\frac{Ck}{K}\sum_j[\phi(t^n)\xi_u^n]_{j-1/2}^2 + \frac{k}{12}\Theta(|(\phi^2)_x|)\\
&\quad + \frac{CT}{k}||\phi(t^n)(\eta_w^n-\eta_u^n)||_{L^2(\mathbb{R})}^2.
\end{aligned}$$

Here we used the fact that $\tilde{K} = \lambda^{-m}K$, $\frac{h}{\tilde{T}} \le 1$, $\gamma \le 1$, and $K \ge 1$. By combining this inequality with (7.8) we get

$$\begin{aligned}
M_1 &\le (1+\lambda^{m-2})\frac{Ck}{K}\sum_j[\phi(t^n)\xi_u^n]_{j-1/2}^2 + \frac{k}{12}\Theta(|(\phi^2)_x|)\\
&\quad + \left(\frac{h}{\tilde{T}}\right)^{2(3/2-2\gamma)}\frac{Ck}{KT}\|\phi(t^n)\xi_u^n\|_{L^2(\mathbb{R})}^2\\
&\quad + \frac{CT}{k}||\phi(t^n)(\eta_w^n-\eta_u^n)||_{L^2(\mathbb{R})}^2,
\end{aligned}$$

where we used the fact that $K \ge 1$, $\lambda \le 1$, and $m \ge 0$. Therefore, combining the estimate of M_1 and M_2 gives the right bound for $|E_2(\phi^2(t)\xi_u^n)|$. We are left to bound $|E_4(\phi^2(t)\xi_w^n)|$. To this end, we write

$$|E_4(\phi^2(t)\xi_w^n)| = |E_4(\phi^2(t)\xi_u^n) + E_4(\phi^2(t)(\xi_w^n-\xi_u^n))|.$$

We can follow the proof of the bound for $E_2(\phi^2(t)\xi_u^n)$ to obtain

$$\begin{aligned}
|E_4(\phi^2(t)\xi_u^n)| &\le (1+\lambda^{m-2})\frac{Ck}{K}\sum_j[\phi(t^n)\xi_w^n]_{j-1/2}^2\\
&\quad + \frac{k}{24}\Theta(|(\phi^2)_x|) + \left(1+\left(\frac{h}{\tilde{T}}\right)^{2(3/2-2\gamma)}\right)\frac{Ck}{TK}||\phi(t^n)\xi_u^n||_{L^2(\mathbb{R})}^2\\
&\quad + \left(1+\left(\frac{h}{\tilde{T}}\right)^{2(3/2-2\gamma)}\right)\frac{CT}{k}||\phi(t^n)(\eta_w^n-\eta_u^n)||_{L^2(\mathbb{R})}^2\\
&\quad + \frac{CT}{Kk}(||\phi(t^n)(\eta_u^{n+1}-\eta_u^n)||_{L^2(\mathbb{R})}^2 + k^6||\phi(t^n)u_{ttt}(\theta^n)||_{L^2(\mathbb{R})}^2).
\end{aligned}$$

If we use (4.13a), (4.13b), and the properties of P_+, we have

$$\begin{aligned}|E_4(\phi^2(t)(\xi_w^n-\xi_u^n))| &= \sum_j \left(\frac{3k}{h^2}[\xi_w^n]_{j-1/2}+\frac{k^3}{6h^3}R_{4,j}+\frac{12}{h^3}R_{2,j}\right)\\ &\qquad \times (x-x_j,\ \phi^2(t)(\xi_w^n-\xi_u^n)-P_+(\phi^2(t)(\xi_w^n-\xi_u^n)))_j\\ &\le Q_1+Q_2+Q_3,\end{aligned}$$

where

$$\begin{aligned}Q_1 &:= \frac{C}{h^{3/2}}\sum_j |R_{2,j}\phi(t^n,x_{j-1/2})|\,D_j,\\ Q_2 &:= \frac{C\,k^3}{h^{3/2}}\sum_j |R_{4,j}\phi(t^n,x_{j-1/2})|\,D_j,\\ Q_3 &:= \frac{C\,k}{h^{1/2}}\sum_j |[\phi(t^n)\xi_w^n]_{j-1/2}|\,D_j,\end{aligned}$$

and

$$D_j := \|\phi^{-1}(t)(\phi^2(t)(\xi_w^n-\xi_u^n)-P_+(\phi^2(t)(\xi_w^n-\xi_u^n)))\|_{L^2(I_j)}.$$

By following the proof of the bound for M_2, we can show that

$$Q_1 \le \frac{C\,k}{K\,T}\|\phi(t^n)\xi_u^n\|^2_{L^2(\mathbb{R})}+\frac{C\,T}{k}(\|\phi(t^n)(\eta_w^n-\eta_u^n)\|^2_{L^2(\mathbb{R})}+\|\phi(t^n)(\eta_u^{n+1}-\eta_u^n)\|^2_{L^2(\mathbb{R})}),$$

where in addition we used Lemma 7.3. Similarly, we can show that

$$Q_2 \le \frac{C\,k}{K\,T}\|\phi(t^n)\xi_u^n\|^2_{L^2(\mathbb{R})}+\frac{C\,T}{k}(\|\phi(t^n)(\eta_w^n-\eta_u^n)\|^2_{L^2(\mathbb{R})}+k^6\|\phi(t^n)u_{ttt}(\theta^n)\|^2_{L^2(\mathbb{R})}).$$

In order to bound Q_3 we first bound D_j by using (4.15a):

$$\begin{aligned}|D_j| &\le \frac{Ch^{2-2\gamma}}{\tilde{T}^{2(1-\gamma)}\tilde{K}^2}(||\phi(t^n)\xi_u^n||_{L^2(I_j)}+||\phi(t^n)\xi_w^n||_{L^2(I_j)})\\ &\quad+\frac{Ch^{2-\gamma/2}}{\tilde{T}^{(1-\gamma)/2}\tilde{K}^{1/2}}||(|\phi(t^n)(\phi(t^n))_x|)^{1/2}(\xi_w^n-\xi_u^n)_x||_{L^2(I_j)}\\ &\le \frac{Ch^{2-2\gamma}}{\tilde{T}^{2(1-\gamma)}\tilde{K}}(||\phi(t^n)\xi_u^n||_{L^2(I_j)}+\|\phi(t^n)(\eta_w^n-\eta_u^n)\|_{L^2(I_j)})\\ &\quad+\frac{Ch^{1-\gamma/2}}{\tilde{T}^{(1-\gamma)/2}\tilde{K}^{1/2}}\Theta_j(|(\phi^2)_x|)^{1/2}.\end{aligned}$$

In the last inequality we used Lemma 7.3 and the inverse estimate (7.4). Therefore,

$$\begin{aligned}Q_3 &\le \frac{Ck}{K}\sum_j[\phi(t^{n+1})\xi_w^n]^2_{j-1/2}+\left(\frac{h}{\tilde{T}}\right)^{2(3/2-2\gamma)}\frac{C\,k}{KT}||\phi\xi_u^n||^2_{L^2(\mathbb{R})}\\ &\quad+\left(\frac{h}{\tilde{T}}\right)^{2(3/2-2\gamma)}\frac{C\,T}{k}\|\phi(t)(\eta_w^n-\eta_u^n)\|^2_{L^2(\mathbb{R})}\\ &\quad+\frac{k}{24}\Theta(|(\phi^2)_x|),\end{aligned}$$

where we used the fact that $\gamma \le 1$, $K \ge 1$, and $\lambda \le 1$. Hence,

$$
\begin{aligned}
|E_4(\phi^2(t)(\xi_w^n - \xi_u^n))| &\le \frac{Ck}{K} \sum_j [\phi(t^{n+1})\xi_w^n]_{j-1/2}^2 \\
&+ \frac{k}{24}\Theta(|(\phi^2)_x|) + \left(1 + \left(\frac{h}{\tilde{T}}\right)^{2(3/2-2\gamma)}\right) \frac{C\,k}{KT} ||\phi(t^n)\xi_u^n||_{L^2(\mathbb{R})}^2 \\
&+ \left(1 + \left(\frac{h}{\tilde{T}}\right)^{2(3/2-2\gamma)}\right) \frac{C\,T}{k} \|\phi(t^n)(\eta_w^n - \eta_u^n)\|_{L^2(\mathbb{R})}^2 \\
&+ \frac{CT}{k}(\|\phi(t^n)(\eta_u^{n+1} - \eta_w^n)\|_{L^2(\mathbb{R})}^2 + k^6 ||\phi(t^n)u_{ttt}(\theta^n)||_{L^2(\mathbb{R})}^2).
\end{aligned}
$$

This completes the proof of Lemma 7.4. ∎

LEMMA 7.5.

$$
\begin{aligned}
\left|\frac{k^2}{2}\sum_j (\xi_w^n - \xi_u^n, (\phi^2(t^n))_{xx}\xi_w^n)_j\right| &\le \left(\frac{h}{\tilde{T}}\right)^{2-3\gamma} \frac{C\,k}{K^3\,T}\|\phi(t^n)\xi_u^n\|_{L^2(\mathbb{R})}^2 + \frac{k}{12}\Theta(|(\phi^2)_x|) \\
&+ \left(\frac{h}{\tilde{T}}\right)^{2-3\gamma} \frac{C\,T}{K^3\,k}\|\phi(t^n)(\eta_w^n - \eta_u^n)\|_{L^2(\mathbb{R})}^2
\end{aligned} \tag{7.9}
$$

and

$$
\begin{aligned}
\left|\frac{k^2}{2}\sum_j (\xi_w^n - \xi_u^n, (\phi^2(t^n))_{xx}\xi_w^n)_j\right| &\le \left(\frac{h}{\tilde{T}}\right)^{1-2\gamma} \frac{C\,k}{K^2\,T}\|\phi(t^n)\xi_u^n\|_{L^2(\mathbb{R})}^2 \\
&+ \left(\frac{h}{\tilde{T}}\right)^{1-2\gamma} \frac{C\,T}{K^2\,k}\|\phi(t^n)(\eta_w^n - \eta_u^n)\|_{L^2(\mathbb{R})}^2.
\end{aligned} \tag{7.10}
$$

Proof. The results are simple consequences of (4.8d), (4.8e), the Cauchy–Schwarz inequality, and Lemma 7.3. ∎

LEMMA 7.6. *For K sufficiently large*

$$
|E_1(P_+(\phi^2(t^{n+1})\xi_w^n))| \le \frac{k}{8\,T}\|\phi(t^n)\xi_u^n\|_{L^2(\mathbb{R})}^2 + CTk^5\|\phi(t^n)u_{ttt}(\theta^n)\|_{L^2(\mathbb{R})}^2.
$$

Proof. By applying the Cauchy–Schwarz inequality, (7.6), and Young's inequality, we obtain Lemma 7.6. ∎

LEMMA 7.7. *Let $t^n \le t \le t^{n+1}$ and $v_h = \xi_u^n$ or $v_h = \xi_w^n$. For $l = 0, 1, 2$, we have*

$$
\begin{aligned}
&h^l|E_3(D_x^l(\phi^2(t))v_h)| + h^l|E_5(D_x^l(\phi^2(t))v_h)| \\
&\le \frac{k}{16\,T}\|\phi(t^n)\xi_u^n\|_{L^2(\mathbb{R})}^2 + \frac{C\,T}{k}(\|\phi(t^n)(\eta_u^{n+1} - \eta_u^n)\|_{L^2(\mathbb{R})}^2 + \|\phi(t^n)(\eta_w^n - \eta_u^n)\|_{L^2(\mathbb{R})}^2).
\end{aligned}
$$

Proof. This result easily follows from the Cauchy–Schwarz inequality, (7.6), Lemma 7.3, and Young's inequality. ∎

LEMMA 7.8. *We have*

$$
|S_1| + |S_2| \le \left(\frac{h}{\tilde{T}}\right)^{2-3\gamma} \frac{Ck}{K^3T}\|\phi(t^n)\xi_u^n\|_{L^2(\mathbb{R})}^2 + \left(\frac{h}{\tilde{T}}\right)^{2-3\gamma} \frac{CT}{k}||\phi(t^n)(\eta_w^n - \eta_u^n)||_{L^2(\mathbb{R})}^2.
$$

Proof. This follows from (4.8d), (4.8e), the Cauchy–Schwarz inequality, and Lemma 7.3. ∎

Now we can prove Lemma 4.8, which is a simple consequence of the lemmas of this section.

Proof. We start with the proof of (4.15). If we apply Lemmas 7.6, 7.4, and 7.7, (7.9), and Lemma 7.8, we get (4.15).

To prove (4.15) we proceed as in the proof for (4.15), but instead of using Lemma 7.4 we apply Lemma 4.10 only, and we use (7.10) rather than (7.9). We also use Lemma 7.3. ∎

7.5. Auxiliary lemma. The following lemma is used to prove (4.12).

LEMMA 7.9. *Let $\phi = \phi_i$, $\gamma = \gamma_i$, and $K = K_i$ for $i = 1$ or $i = 2$. If $t^n < t < t^{n+1}$, then*

$$k^2|H(\xi_u^n,(\phi^2(t))_x(\xi_w^n-\xi_u^n))| \le \frac{Ck}{K}\sum_j [\boldsymbol{\phi\xi}^n]_{j-1/2}\mathbb{I}[\boldsymbol{\phi\xi}^n]_{j-1/2}^t$$
$$+\left(\left(\frac{h}{\tilde{\tilde{T}}}\right)^{2(1/2-\gamma)}+1\right)\frac{C\,k}{K\,T}\|\phi(t^n)\xi_u^n\|_{L^2(\mathbb{R})}^2$$
$$+\left(1+\left(\frac{h}{\tilde{\tilde{T}}}\right)^{2(1/2-\gamma)}\right)\frac{C\,T}{k}\|\phi(t^n)(\eta_w^n-\eta_u^n)\|_{L^2(\mathbb{R})}^2.$$

Proof.

$$k^2|H(\xi_u^n,(\phi^2(t))_x(\xi_w^n-\xi_u^n))|$$
$$= k^2\left|\sum_j(\xi_u^n,((\phi^2)_x(t)(\xi_w^n-\xi_u^n))_x)_j+\xi_u^n(x_{j+1/2}^-)[(\phi^2)_x(t)(\xi_w^n-\xi_u^n)]_{j+1/2}\right|.$$

By applying the Cauchy–Schwarz inequality, (4.8d), (4.8e), (7.3), and Young's inequality, we have

$$k^2\left|\sum_j \xi_u^n(x_{j+1/2}^-)[(\phi^2)_x(t)(\xi_w^n-\xi_u^n)]_{j+1/2}\right|$$
$$\le \frac{C\,k}{K}\sum_j[\boldsymbol{\phi\xi}^n]_{j-1/2}\mathbb{I}[\boldsymbol{\phi\xi}^n]_{j-1/2}^t+\left(\frac{h}{\tilde{\tilde{T}}}\right)^{2(1/2-\gamma)}\frac{C\,k}{K\,T}\|\phi(t^n)\xi_u^n\|_{L^2(\mathbb{R})}^2.$$

The product rule gives

$$\left|\sum_j(\xi_u^n,((\phi^2)_x(t)(\xi_w^n-\xi_u^n))_x)_j\right| = \left|\sum_j\{(\xi_u^n,(\phi^2)_{xx}(t)(\xi_w^n-\xi_u^n))_j\right.$$
$$\left.+(\xi_u^n,(\phi^2)_x(t)(\xi_w^n-\xi_u^n)_x)_j\}\right|.$$

By applying the Cauchy–Schwarz inequality, (4.8d), Lemma 7.3, and Young's inequality, we have

$$k^2\left|\sum_j(\xi_u^n,(\phi^2)_{xx}(t^n)(\xi_w^n-\xi_u^n))_j\right| \le \left(\frac{h}{\tilde{T}}\right)^{1-2\gamma}\frac{C\,k}{K^2T}(\|\phi(t^n)\xi_u^n\|^2_{L^2(\mathbb{R})} + \|\phi(t^n)(\eta_w^n-\eta_u^n)\|^2_{L^2(\mathbb{R})}).$$

If we use (4.13b), (4.8e), (4.8d), and Young's inequality, we see that

$$\begin{aligned}k^2\left|\sum_j(\xi_u^n,(\phi^2)_x(t)(\xi_w^n-\xi_u^n)_x)_j\right| &\le \frac{C\,k}{K}\sum_j[\phi(t^n)\xi_u^n]^2_{j-1/2}\\ &\quad+\left(\left(\frac{h}{\tilde{T}}\right)^{2(1/2-\gamma)}+1\right)\frac{C\,k}{K\,T}\|\phi(t^n)\xi_u^n\|^2_{L^2(\mathbb{R})}\\ &\quad+\frac{C\,T}{k}\|\phi(t^n)(\eta_w^n-\eta_u^n)\|^2_{L^2(\mathbb{R})}.\end{aligned}$$

This concludes the proof. ∎

REFERENCES

[1] M. Apelkrans, *On difference schemes for hyperbolic equations with discontinuous initial values*, Math. Comp., 22 (1968), pp. 525–539.

[2] P. Brenner and V. Thomée, *Estimates near discontinuities for some difference schemes*, Math. Scand., 28 (1971), pp. 329–340.

[3] P. Castillo, B. Cockburn, D. Schötzau, and C. Schwab, *Optimal a priori error estimates for the hp-version of the local discontinuous Galerkin method for convection-diffusion problems*, Math. Comp., 71 (2002), pp. 455–478.

[4] B. Cockburn, *An introduction to the discontinuous Galerkin method for convection-dominated problems*, in Advanced Numerical Approximation of Nonlinear Hyperbolic Equations, Lecture Notes in Math. 1697, Fond. C.I.M.E., A. Quarteroni, ed., Springer-Verlag, New York, 1998, pp. 151–268.

[5] B. Cockburn, *Discontinuous Galerkin methods for convection-dominated problems*, in High-Order Methods for Computational Physics, Lect. Notes Comput. Sci. Eng. 9, T. Barth and H. Deconink, eds., Springer-Verlag, Berlin, 1999, pp. 69–224.

[6] B. Cockburn, S. Hou, and C.-W. Shu, *TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws* IV: *The multidimensional case*, Math. Comp., 54 (1990), pp. 545–581.

[7] B. Cockburn, G. Karniadakis, and C.-W. Shu, *The development of discontinuous Galerkin methods*, in Discontinuous Galerkin Methods. Theory, Computation and Applications, Lect. Notes Comput. Sci. Eng. 11, B. Cockburn, G. Karniadakis, and C.-W. Shu, eds., Springer-Verlag, Berlin, 2000, pp. 3–50.

[8] B. Cockburn, S. Lin, and C.-W. Shu, *TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws* III: *One dimensional systems*, J. Comput. Phys., 84 (1989), pp. 90–113.

[9] B. Cockburn and C.-W. Shu, *TVB Runge-Kutta local projection discontinuous Galerkin finite element method for scalar conservation laws* II: *General framework*, Math. Comp., 52 (1989), pp. 411–435.

[10] B. Cockburn and C.-W. Shu, *The Runge-Kutta local projection P^1-discontinuous Galerkin method for scalar conservation laws*, RAIRO Modél. Math. Anal. Numér., 25 (1991), pp. 337–361.

[11] B. Cockburn and C.-W. Shu, *The Runge-Kutta discontinuous Galerkin finite element method for conservation laws* V*: Multidimensional systems*, J. Comput. Phys., 141 (1998), pp. 199–224.

[12] B. Cockburn and C.-W. Shu, *Runge-Kutta discontinuous Galerkin methods for convection-dominated problems*, J. Sci. Comput., 16 (2001), pp. 173–261.

[13] J. Guzmán, *Local analysis of discontinuous Galerkin methods applied to singularly perturbed problems*, J. Numer. Math., 45 (2006), pp. 41–56.
[14] G.W. Hedstrom, *The rate of convergence of some difference schemes*, SIAM J. Numer. Anal., 5 (1968), pp. 363–406.
[15] G.W. Hedstrom, *Models of difference schemes for $u_t + u_x = 0$ by partial differential equations*, Math. Comp., 29 (1975), pp. 969–977.
[16] C. Johnson, U. Nävert, and J. Pitkäranta, *Finite element methods for linear hyperbolic problems*, Comput. Methods Appl. Mech. Engrg., 45 (1984), pp. 285–312.
[17] C. Johnson, A. Schatz, and L. Walhbin, *Crosswind smear and pointwise errors in streamline diffusion finite element methods*, Math. Comp., 49 (1987), pp. 25–38.
[18] C.-W. Shu, *Total-variation-diminishing time discretizations*, SIAM J. Sci. Comput., 9 (1988), pp. 1073–1084.
[19] Q. Zhang and C.-W. Shu, *Error estimates to smooth solutions of Runge–Kutta discontinuous Galerkin methods for scalar conservation laws*, SIAM J. Numer. Anal., 42 (2004), pp. 641–666.
[20] V. Thomée, *Galerkin Finite Element Methods for Parabolic Equations*, Springer-Verlag, New York, 1997.

SIAM J. NUMER. ANAL.
Vol. 46, No. 3, pp. 1399–1422
© 2008 Society for Industrial and Applied Mathematics

A CONVERGENT IMPLICIT FINITE ELEMENT DISCRETIZATION OF THE MAXWELL–LANDAU–LIFSHITZ–GILBERT EQUATION*

L'UBOMÍR BAŇAS†, SÖREN BARTELS‡, AND ANDREAS PROHL§

Abstract. We propose an implicit, fully discrete scheme for the numerical solution of the Maxwell–Landau–Lifshitz–Gilbert equation which is based on linear finite elements and satisfies a discrete sphere constraint as well as a discrete energy law. As numerical parameters tend to zero, solutions weakly accumulate at weak solutions of the Maxwell–Landau–Lifshitz–Gilbert equation. A practical linearization of the nonlinear scheme is proposed and shown to converge for certain scalings of numerical parameters. Computational studies are presented to indicate finite-time blow-up behavior and to study combined electromagnetic phenomena in ferromagnets for benchmark problems.

Key words. ferromagnetism, Maxwell–Landau–Lifshitz–Gilbert equation, finite elements, convergence

AMS subject classifications. 35K55, 65M12, 65M15, 68U10, 94A08

DOI. 10.1137/070683064

1. Introduction.

The Maxwell–Landau–Lifschitz–Gilbert equation (MLLG) describes certain electromagnetic phenomena in a ferromagnet occupying the domain $\omega \Subset \Omega \subseteq \mathbb{R}^d$, $d = 2, 3$. For a parameter $\alpha > 0$ which serves as a damping factor, the magnetization field $\mathbf{m} : (0,T) \times \omega \to \mathbb{S}^2$, where $\mathbb{S}^2 = \{\mathbf{x} \in \mathbb{R}^3 | \, |\mathbf{x}| = 1\}$ is the unit sphere, and the electric and magnetic fields $(\mathbf{E}, \mathbf{H}) : (0,T) \times \Omega \to \mathbb{R}^3$ satisfy for all $T > 0$

$$\mathbf{m}_t + \alpha\, \mathbf{m} \times \mathbf{m}_t = (1+\alpha^2)\, \mathbf{m} \times \mathbf{H}_{\text{eff}} \quad \text{in } \omega_T := (0,T) \times \omega\,, \tag{1.1}$$

$$\varepsilon_0\, \mathbf{E}_t - \nabla \times \mathbf{H} + \sigma\, \chi_\omega \mathbf{E} = -\mathbf{J} \quad \text{in } \Omega_T := (0,T) \times \Omega\,, \tag{1.2}$$

$$\mu_0\, \mathbf{H}_t + \nabla \times \mathbf{E} = -\mu_0\, \mathbf{m}_t \qquad \text{in } \Omega_T \tag{1.3}$$

for the (simplified) effective field $\mathbf{H}_{\text{eff}} = \Delta \mathbf{m} + \mathbf{H}$. This choice of $\mathbf{H}_{\text{eff}}$ comprises the most relevant contributions to the more general version (5.1) below. The constants $\varepsilon_0, \mu_0 \geq 0$ denote, respectively, the electric and magnetic permeability of free space, while $\sigma \geq 0$ describes the conductivity of the ferromagnet. The field $\mathbf{J} : \Omega_T \to \mathbb{R}^3$ denotes an applied current density and $\chi_\omega : \Omega \to \{0,1\}$ is the characteristic function of ω. For simplicity we suppose that $\Omega \subset \mathbb{R}^3$ is a bounded cavity with a perfectly conducting outer surface $\partial\Omega$ into which the ferromagnet $\omega \Subset \Omega$ is embedded, and $\Omega \setminus \overline{\omega}$ is assumed to be a vacuum [15]. The system (1.1)–(1.3) is supplemented with

*Received by the editors February 20, 2007; accepted for publication (in revised form) November 29, 2007; published electronically March 26, 2008.
http://www.siam.org/journals/sinum/46-3/68306.html

†Department of Mathematics, Heriot-Watt University, Edinburgh, EH14 4AS, United Kingdom (L.Banas@hw.ac.uk). This author's research was supported by the EPSRC grant EP/C548973/1.

‡Institut für Numerische Simulation, Rheinische Friedrich-Wilhelms-Universität Bonn, Wegelerstraße 6, D-53115 Bonn, Germany (bartels@ins.uni-bonn.de). This author's research was supported by Deutsche Forschungsgemeinschaft through the DFG Research Center MATHEON "Mathematics for Key Technologies" in Berlin.

§Mathematisches Institut, Universität Tübingen, Auf der Morgenstelle 10, D-72076 Tübingen, Germany (prohl@na.uni-tuebingen.de).

initial conditions

$$\mathbf{m}(0,\cdot)=\mathbf{m}_0 \quad \text{in } \omega\,, \qquad \mathbf{E}(0,\cdot)=\mathbf{E}_0\,, \quad \mathbf{H}(0,\cdot)=\mathbf{H}_0 \quad \text{in } \Omega \tag{1.4}$$

and boundary conditions

$$\partial_{\mathbf{n}}\mathbf{m}=0 \quad \text{on } \partial\omega_T\,, \qquad \mathbf{E}\times\mathbf{n}=0 \quad \text{on } \partial\Omega_T\,. \tag{1.5}$$

We remark that the inclusion of a damping term is necessary to permit the magnetic field to align with the effective field, and we refer the reader to [5, 23, 12] for a more detailed discussion of the mathematical model. Interesting computational studies of the model (1.1)–(1.5) can be found in [15, 22].

The construction of convergent numerical schemes for the Maxwell–Landau-Lifshitz–Gilbert system (1.1)–(1.5) is difficult due to the nonlinear character of (1.1). Moreover, the limited flexibility of piecewise polynomial finite elements and the artificial damping of implicit time-discretization schemes make it hard to appropriately account for the constraint $|\,\mathbf{m}\,|=1$ a.e. in ω_T. The semidiscrete scheme proposed in [15] uses numerical integration to guarantee that the constraint is satisfied at the nodes of a triangulation. Stability of that scheme is proved in [15], but its convergence has not been discussed in the literature so far.

According to Landau and Lifshitz [11], damped precession ($\alpha>0$) of the magnetization $\mathbf{m}:\omega_T\to\mathbb{S}^2$ is governed by

$$\mathbf{m}_t=\mathbf{m}\times\mathbf{H}_{\text{eff}}-\alpha\,\mathbf{m}\times(\mathbf{m}\times\mathbf{H}_{\text{eff}})\,. \tag{1.6}$$

Gilbert's approach modifies the undamped precession equation by a damping term which is proportional to the rate of change of magnetization; see (1.1). Both approaches are analytically equivalent for smooth magnetization fields, and weak solutions to MLLG exist in both cases [23, 6]. The numerical analysis differs significantly, and Gilbert's approach turns out to be attractive from a numerical viewpoint. Explicit and implicit discretizations for the simplest choice $\mathbf{H}_{\text{eff}}=\Delta\mathbf{m}$ have been proposed and analyzed in [1, 3, 4], and approximations are known to respectively converge conditionally and unconditionally with respect to discretization parameters to weak solutions of (1.1). Corresponding results for (1.6) are so far only available for (locally existing) strong solutions; cf. [18, 7]. The articles [14, 21] study convergence for the simplification $\mathbf{H}_{\text{eff}}=\mathbf{H}$ for cases where smooth solutions to (1.1)–(1.5) exist. Convergence of iterates of the following scheme to weak solutions of (1.1) for $\mathbf{H}_{\text{eff}}=\Delta\mathbf{m}$ is verified in [4]; the proof relies on conservation of $|\,\mathbf{m}_h^j\,|=1$ at all mesh points and for all $j\geq 0$ as well as a discrete energy law.

ALGORITHM 1.1. *Let $\mathbf{m}_h^0\in\mathbf{V}_h$. Given $j\geq 0$ and $\mathbf{m}_h^j\in\mathbf{V}_h$, let $\mathbf{m}_h^{j+1}\in\mathbf{V}_h$ solve*

$$(d_t\mathbf{m}_h^{j+1},\boldsymbol{\phi}_h)_h+\alpha\,(\mathbf{m}_h^j\times d_t\mathbf{m}_h^{j+1},\boldsymbol{\phi}_h)_h=(1+\alpha^2)(\overline{\mathbf{m}}_h^{j+1/2}\times\tilde{\Delta}_h\overline{\mathbf{m}}_h^{j+1/2},\boldsymbol{\phi}_h)_h$$
$$\forall\,\boldsymbol{\phi}_h\in\mathbf{V}_h\,.$$

Here, $\mathbf{V}_h\subset W^{1,2}(\omega;\mathbb{R}^3)$ is the finite element space subordinate to a triangulation $\mathcal{T}_h$ of ω consisting of piecewise linear functions, and $(\cdot,\cdot)_h$ denotes a discrete version (numerical integration) of the inner product in $L^2(\omega;\mathbb{R}^3)$. Moreover, $\tilde{\Delta}_h:W^{1,2}(\omega;\mathbb{R}^3)\to\mathbf{V}_h$ is an approximation of the Laplace operator. For a time-step size $k>0$ we write $d_t\varphi^j:=k^{-1}\big(\varphi^j-\varphi^{j-1}\big)$ for $j\geq 1$, and $\overline{\varphi}^{j+1/2}:=\frac{1}{2}\big(\varphi^{j+1}+\varphi^j\big)$ for $j\geq 0$ and a sequence $\{\varphi^j\}_{j\geq 0}$; we refer the reader to section 2 for details.

The goal of this work is (i) to propose a discretization of (1.1)–(1.3) based on linear finite elements which satisfies a discrete sphere constraint and a discrete energy identity, (ii) to verify convergence of iterates toward a weak solution of (1.1)–(1.3), and (iii) to propose a reliable solver for the nonlinear system of equations in each time step. This program is motivated by possible finite-time blow-up behavior of solutions to (1.1)–(1.3) (see section 5), whose numerical simulation requires reliable numerical schemes. Our strategy is to appropriately modify Algorithm 1.1 and to extend its analysis from [4].

The discretization of (1.1)–(1.3) requires a proper choice of Ansatz spaces for the approximation of $(\mathbf{E}, \mathbf{H})$. Suitable pairs can be found in [14, 15], and we let $\mathbf{X}_h \subset \mathbf{H}_0(\mathbf{curl}; \Omega)$ and $\mathbf{Y}_h \subset L^2(\Omega, \mathbb{R}^3)$ be finite dimensional spaces satisfying $\nabla \times \mathbf{X}_h \subset \mathbf{Y}_h$. We then aim at analyzing the following algorithm for the numerical approximation of MLLG.

ALGORITHM 1.2. *Let* $(\mathbf{m}_h^0, \mathbf{E}_h^0, \mathbf{H}_h^0) \in \mathbf{V}_h \times \mathbf{X}_h \times \mathbf{Y}_h$. *For* $j \geq 0$ *and* $(\mathbf{m}_h^j, \mathbf{E}_h^j, \mathbf{H}_h^j) \in \mathbf{V}_h \times \mathbf{X}_h \times \mathbf{Y}_h$, *let* $(\mathbf{m}_h^{j+1}, \mathbf{E}_h^{j+1}, \mathbf{H}_h^{j+1}) \in \mathbf{V}_h \times \mathbf{X}_h \times \mathbf{Y}_h$ *solve*

$$(1.7)\quad (d_t \mathbf{m}_h^{j+1}, \boldsymbol{\phi}_h)_h + \alpha\, (\mathbf{m}_h^j \times d_t \mathbf{m}_h^{j+1}, \boldsymbol{\phi}_h)_h = (1+\alpha^2) \Big(\overline{\mathbf{m}}_h^{j+1/2} \times (\tilde{\Delta}_h \overline{\mathbf{m}}_h^{j+1/2} + \mathbf{P}_{\mathbf{V}_h} \overline{\mathbf{H}}_h^{j+1/2}), \boldsymbol{\phi}_h \Big)_h \quad \forall\, \boldsymbol{\phi}_h \in \mathbf{V}_h\,,$$

$$(1.8)\quad \varepsilon_0\, (d_t \mathbf{E}_h^{j+1}, \boldsymbol{\varphi}_h) - (\overline{\mathbf{H}}_h^{j+1/2}, \nabla \times \boldsymbol{\varphi}_h) + \sigma\, (\chi_\omega \overline{\mathbf{E}}_h^{j+1/2}, \boldsymbol{\varphi}_h) = -(\overline{\mathbf{J}}_h^{j+1/2}, \boldsymbol{\varphi}_h) \qquad \forall\, \boldsymbol{\varphi}_h \in \mathbf{X}_h\,,$$

$$(1.9)\quad \mu_0\, (d_t \mathbf{H}_h^{j+1}, \boldsymbol{\mathfrak{Z}}_h) + (\nabla \times \overline{\mathbf{E}}_h^{j+1/2}, \boldsymbol{\mathfrak{Z}}_h) = -\mu_0\, (d_t \mathbf{m}_h^{j+1}, \boldsymbol{\mathfrak{Z}}_h) \qquad \forall\, \boldsymbol{\mathfrak{Z}}_h \in \mathbf{Y}_h\,.$$

Here, $\mathbf{P}_{\mathbf{V}_h} : L^2(\omega, \mathbb{R}^3) \to \mathbf{V}_h$, with $(\mathbf{P}_{\mathbf{V}_h} \mathbf{u}, \boldsymbol{\varphi}_h)_h = (\mathbf{u}, \boldsymbol{\varphi}_h)$ for all $\boldsymbol{\varphi}_h \in \mathbf{V}_h$, denotes the L^2-projection into $\mathbf{V}_h$.

Lemma 3.1 below establishes conservation of $|\, \mathbf{m}_h^j\, | = 1$, $j \geq 0$, at nodes of the triangulation $\mathcal{T}_h$, and a discrete energy law for solutions to Algorithm 1.2. Our first main result is Theorem 3.1, which states unconditional convergence of subsequences of outputs of Algorithm 1.2. A simple fixed-point iteration for the approximate solution of the nonlinear system of equations in each step of Algorithm 1.2, which preserves the unit-length constraint, is proposed in Algorithm 4.1. Its convergence is proved under the mesh constraint $k = \mathcal{O}(h^2)$ in Lemma 4.1. Numerical studies which indicate possible finite-time blow-up of weak solutions of (1.1)–(1.5) are reported in section 5. Simulations of benchmark problems from [8] are also reported.

2. Preliminaries. Throughout this paper we assume that $\mathcal{T}_h$ is a regular triangulation of the polygonal or polyhedral bounded Lipschitz domain $\Omega \subset \mathbb{R}^d$ into triangles or tetrahedra of maximal mesh-size (maximal diameter of triangles or tetrahedra) $h > 0$ for $d = 2$ or $d = 3$, respectively, and $\mathcal{T}_h\big|_\omega$ denotes its restriction to $\omega \Subset \Omega$. We define the lowest order conforming finite element space $\mathbf{V}_h \subset W^{1,2}(\omega; \mathbb{R}^3)$ by

$$\mathbf{V}_h = \big\{ \boldsymbol{\phi}_h \in C(\overline{\omega}; \mathbb{R}^3) : \boldsymbol{\phi}_h|_K \in \mathcal{P}_1(K; \mathbb{R}^3) \quad \forall\, K \in \mathcal{T}_h\big|_\omega \big\}\,,$$

where $\mathcal{P}_1(K; \mathbb{R}^3)$ denotes the set of polynomials of total degree less than or equal to one if restricted to the element $K \in \mathcal{T}_h$. Given the set of nodes $\{\mathbf{x}_\ell : \ell \in L\}$ of the triangulation $\mathcal{T}_h\big|_\omega$, the nodal interpolation operator $\boldsymbol{\mathcal{I}}_{\mathbf{V}_h} : C(\overline{\omega}; \mathbb{R}^3) \to \mathbf{V}_h$ satisfies

$\boldsymbol{\mathcal{I}}_{\mathbf{V}_h}\boldsymbol{\phi}(\mathbf{x}_\ell) = \boldsymbol{\phi}(\mathbf{x}_\ell)$ for all $\ell \in L$. We let $\langle\cdot,\cdot\rangle$ denote the inner product in $\mathbb{R}^d$, and we define

$$\big(\mathbf{f},\mathbf{g}\big) = \int_\Omega \langle \mathbf{f},\mathbf{g}\rangle \,\mathrm{d}\mathbf{x} \qquad \text{and} \qquad \big(\boldsymbol{\phi},\boldsymbol{\mathfrak{Z}}\big)_h = \int_\omega \boldsymbol{\mathcal{I}}_{\mathbf{V_h}}\big(\langle\boldsymbol{\phi},\boldsymbol{\mathfrak{Z}}\rangle\big)\,\mathrm{d}\mathbf{x} = \sum_{\ell\in L}\beta_\ell\langle\boldsymbol{\phi}(\mathbf{x}_\ell),\boldsymbol{\mathfrak{Z}}(\mathbf{x}_\ell)\rangle$$

for $\mathbf{f},\mathbf{g} \in L^2(\Omega;\mathbb{R}^d)$, certain weights $\beta_\ell > 0$, $\ell \in L$, and continuous functions $\boldsymbol{\phi}, \boldsymbol{\mathfrak{Z}} \in C(\overline{\omega};\mathbb{R}^3)$. Note that we use the notation $(\cdot,\cdot)$ for the inner product in $L^2(\Omega)$ as well as in $L^2(\omega)$—it will always be clear from the context which of them it represents. For each $\ell \in L$ we let $\varphi_\ell \in C(\overline{\omega})$ denote the nodal basis function in $\mathbf{V}_h$ which is $\mathcal{T}_h$-elementwise affine and satisfies $\varphi_\ell(\mathbf{x}_m) = \delta_{\ell m}$ for all $m,\ell \in L$, and we define $\beta_\ell = \int_\omega \varphi_\ell \,\mathrm{d}\mathbf{x}$. We write $||\boldsymbol{\phi}||_h^2 = \big(\boldsymbol{\phi},\boldsymbol{\phi}\big)_h$ and notice that

$$\|\,\boldsymbol{\phi}_h\,\|_{L^2}^2 \le \|\,\boldsymbol{\phi}_h\,\|_h^2 \le (d+2)\,\|\,\boldsymbol{\phi}_h\,\|_{L^2}^2 \qquad \forall\,\boldsymbol{\phi}_h \in \mathbf{V}_h\,. \tag{2.1}$$

Basic interpolation estimates yield for all $\boldsymbol{\phi}_h, \boldsymbol{\mathfrak{Z}}_h \in \mathbf{V}_h$ that

$$\big|(\boldsymbol{\phi}_h,\boldsymbol{\mathfrak{Z}}_h)_h - (\boldsymbol{\phi}_h,\boldsymbol{\mathfrak{Z}}_h)\big| \le Ch\|\boldsymbol{\phi}_h\|\,\|\nabla\boldsymbol{\mathfrak{Z}}_h\|\,. \tag{2.2}$$

We define a discrete Laplace operator $\tilde{\Delta}_h : W^{1,2}(\omega;\mathbb{R}^3) \to \mathbf{V}_h$ by

$$(-\tilde{\Delta}_h\boldsymbol{\phi},\boldsymbol{\chi}_h)_h = (\nabla\boldsymbol{\phi},\nabla\boldsymbol{\chi}_h) \quad \forall\,\boldsymbol{\chi}_h \in \mathbf{V}_h\,. \tag{2.3}$$

It is well known that there exist constants $c_1, c_2 > 0$ such that for all $\boldsymbol{\phi}_h \in \mathbf{V}_h$ and all $\ell \in L$ we have (see, e.g., [4])

$$||\tilde{\Delta}_h\boldsymbol{\phi}_h||_h \le c_1 h_{min}^{-1}||\nabla\boldsymbol{\phi}_h||_{L^2} \le c_2 h_{min}^{-2}||\boldsymbol{\phi}_h||_{L^2}\,, \tag{2.4}$$

$$|\tilde{\Delta}_h\boldsymbol{\phi}_h(\mathbf{x}_\ell)| \le c_3 h_{min}^{-2}||\boldsymbol{\phi}_h||_{L^\infty}\,, \tag{2.5}$$

where $h_{min} > 0$ is the minimal diameter of elements in $\mathcal{T}_h$.

To discretize Maxwell's equations, we employ finite element spaces $\mathbf{X}_h \subset \mathbf{H}_0(\mathbf{curl};\Omega)$, $\mathbf{Y}_h \subset L^2(\Omega;\mathbb{R}^3)$ subordinate to $\mathcal{T}_h$ such that $\nabla\times\mathbf{X}_h \subset \mathbf{Y}_h$; common examples are Nédélec's first and second family of edge elements on tetrahedra, where we choose the latter subsequently; cf. [13, Chapter 8.5]

$$\mathbf{X}_h = \big\{\boldsymbol{\varphi}_h \in \mathbf{H}_0(\mathbf{curl};\Omega) : \boldsymbol{\varphi}_h|_K \in \mathcal{P}_1(K;\mathbb{R}^3) \quad \forall\,K \in \mathcal{T}_h\big\}$$

and

$$\mathbf{Y}_h = \big\{\boldsymbol{\mathfrak{Z}}_h \in L^2(\Omega;\mathbb{R}^3) : \boldsymbol{\mathfrak{Z}}_h|_K \in \mathcal{P}_0(K;\mathbb{R}^3) \quad \forall\,K \in \mathcal{T}_h\big\}\,,$$

where global interpolants of sufficiently smooth functions ($\delta > 0$, $p > 2$)

$$\boldsymbol{\mathcal{I}}_{\mathbf{X}_h} : W^{1/2+\delta,2}(\Omega,\mathbb{R}^3) \cap W^{1,p}(\Omega,\mathbb{R}^3) \to \mathbf{X}_h \quad \text{and} \quad \boldsymbol{\mathcal{I}}_{\mathbf{Y}_h} : W^{1/2+\delta,2}(\Omega,\mathbb{R}^3) \to \mathbf{Y}_h$$

exist and satisfy [13, Theorem 8.15, Remark 8.8]

$$\|\,\boldsymbol{\varphi} - \boldsymbol{\mathcal{I}}_{\mathbf{X}_h}\boldsymbol{\varphi}\,\|_{L^2} + h\,\|\,\nabla\times(\boldsymbol{\varphi} - \boldsymbol{\mathcal{I}}_{\mathbf{X}_h}\boldsymbol{\varphi})\,\|_{L^2} \le Ch^2\,\|\,\nabla^2\boldsymbol{\varphi}\,\|_{L^2}\,, \tag{2.6}$$

$$\|\,\boldsymbol{\mathfrak{Z}} - \boldsymbol{\mathcal{I}}_{\mathbf{Y}_h}\boldsymbol{\mathfrak{Z}}\,\|_{L^2} \le Ch\,\|\,\boldsymbol{\mathfrak{Z}}\,\|_{H^1} \tag{2.7}$$

for all $\boldsymbol{\varphi} \in W^{2,2}(\Omega,\mathbb{R}^3)$ and $\boldsymbol{\mathfrak{Z}} \in W^{1,2}(\Omega,\mathbb{R}^3)$.

To define weak solutions of (1.1)–(1.3), we assume that the given data satisfy

$$\mathbf{m}_0 \in W^{1,2}(\omega, \mathbb{S}^2)\,, \quad \mathbf{H}_0,\, \mathbf{E}_0 \in L^2(\Omega, \mathbb{R}^3)\,, \quad \mathbf{J} \in L^2(\Omega_T, \mathbb{R}^3)\,. \tag{2.8}$$

We assume that the set of initial data satisfies $\operatorname{div} \mathbf{H}_0 = 0$ and is consistent in the sense that

$$\operatorname{div}(\mathbf{H}_0 + \chi_\omega \mathbf{m}_0) = 0 \quad \text{in } \Omega\,, \qquad \langle \mathbf{H}_0 + \chi_\omega \mathbf{m}_0, \mathbf{n}\rangle = 0 \quad \text{on } \partial\Omega\,. \tag{2.9}$$

Definition 2.1. *Suppose* (2.8)–(2.9) *hold. Then* $(\mathbf{m}, \mathbf{E}, \mathbf{H})$ *is called a weak solution to MLLG if, for all* $T > 0$,

(1) $\mathbf{m} \in L^\infty\big(0,T; W^{1,2}(\omega, \mathbb{R}^3)\big)$, *such that* $\mathbf{m}_t \in L^2(\omega_T, \mathbb{R}^3)$ *and* $|\,\mathbf{m}\,| = 1$ *a.e. in* ω_T, *and* $\mathbf{E}, \mathbf{H} \in L^\infty\big(0,T; L^2(\Omega, \mathbb{R}^3)\big)$;

(2) *for all* $\boldsymbol{\phi} \in C^\infty\big(\overline{\omega_T}; \mathbb{R}^3\big)$ *and* $\boldsymbol{\mathfrak{Z}} \in \mathcal{D}\big([0,T); C^\infty(\Omega, \mathbb{R}^3) \cap \mathbf{H}_0(\mathbf{curl}, \Omega)\big)$,

$$\begin{aligned}(2.10)\quad &\int_{\omega_T} \langle \mathbf{m}_t, \boldsymbol{\phi}\rangle \,\mathrm{d}\mathbf{x}\,\mathrm{d}t + \alpha \int_{\omega_T} \langle \mathbf{m}\times\mathbf{m}_t, \boldsymbol{\phi}\rangle \,\mathrm{d}\mathbf{x}\,\mathrm{d}t \\ &\qquad = -(1+\alpha^2)\left[\int_{\omega_T} \langle \mathbf{m}\times\nabla\mathbf{m}, \nabla\boldsymbol{\phi}\rangle \,\mathrm{d}\mathbf{x}\,\mathrm{d}t - \int_{\omega_T} \langle \mathbf{m}\times\mathbf{H}, \boldsymbol{\phi}\rangle \,\mathrm{d}\mathbf{x}\,\mathrm{d}t\right],\end{aligned}$$

$$\begin{aligned}(2.11)\quad &-\varepsilon_0 \int_{\Omega_T} \langle \mathbf{E}, \boldsymbol{\mathfrak{Z}}_t\rangle \,\mathrm{d}\mathbf{x}\,\mathrm{d}t - \int_{\Omega_T} \langle \mathbf{H}, \nabla\times\boldsymbol{\mathfrak{Z}}\rangle \,\mathrm{d}\mathbf{x}\,\mathrm{d}t \\ &\qquad + \sigma \int_{\omega_T} \langle \mathbf{E}, \boldsymbol{\mathfrak{Z}}\rangle \,\mathrm{d}\mathbf{x}\,\mathrm{d}t = -\int_{\Omega_T} \langle \mathbf{J}, \boldsymbol{\mathfrak{Z}}\rangle \,\mathrm{d}\mathbf{x}\,\mathrm{d}t + \varepsilon_0 \int_\Omega \langle \mathbf{E}_0, \boldsymbol{\mathfrak{Z}}(0,\cdot)\rangle \,\mathrm{d}\mathbf{x}\,,\end{aligned}$$

$$\begin{aligned}(2.12)\quad &-\mu_0 \int_{\Omega_T} \langle \mathbf{H} + \chi_\omega \mathbf{m}, \boldsymbol{\mathfrak{Z}}_t\rangle \,\mathrm{d}\mathbf{x}\,\mathrm{d}t + \int_{\Omega_T} \langle \mathbf{E}, \nabla\times\boldsymbol{\mathfrak{Z}}\rangle \,\mathrm{d}\mathbf{x}\,\mathrm{d}t \\ &\qquad = \mu_0 \int_\Omega \langle \mathbf{H}_0, \boldsymbol{\mathfrak{Z}}(0,\cdot)\rangle \,\mathrm{d}\mathbf{x} + \mu_0 \int_\omega \langle \mathbf{m}_0, \boldsymbol{\mathfrak{Z}}(0,\cdot)\rangle \,\mathrm{d}\mathbf{x}\,;\end{aligned}$$

(3) *we have* $\mathbf{m}(0,\cdot) = \mathbf{m}_0$ *in the sense of traces;*

(4) *for almost all* $T' \in (0,T)$ *we have*

$$\begin{aligned}&\mathcal{E}_{(\mathbf{m},\mathbf{H},\mathbf{E})}(T') + \int_{\omega_{T'}} \left(\frac{\alpha\mu_0}{1+\alpha^2}\,|\,\mathbf{m}_t\,|^2 + \sigma|\,\mathbf{E}\,|^2\right) \mathrm{d}\mathbf{x}\,\mathrm{d}t \\ &\qquad \le \mathcal{E}_{(\mathbf{m},\mathbf{H},\mathbf{E})}(0) - \int_{\Omega_{T'}} (\mathbf{J}, \mathbf{E}) \,\mathrm{d}\mathbf{x}\,\mathrm{d}t\,,\end{aligned}$$

where

$$\begin{aligned}&\mathcal{E}_{(\mathbf{m},\mathbf{H},\mathbf{E})}(T') \\ &\qquad = \frac{\mu_0}{2} \int_\omega |\,\nabla\mathbf{m}(T',\cdot)\,|^2 \,\mathrm{d}\mathbf{x} + \int_\Omega \Big[\frac{\mu_0}{2}|\,\mathbf{H}(T',\cdot)\,|^2 + \frac{\varepsilon_0}{2}|\,\mathbf{E}(T',\cdot)\,|^2\Big] \,\mathrm{d}\mathbf{x}\,.\end{aligned}$$

Existence of weak solutions was first shown in [6].

3. Stability and convergence. The following lemma provides a discrete counterpart of (4) in Definition 2.1 for Algorithm 1.2. Solvability of (1.7)–(1.9) follows from a contraction argument, which employs (i) the discrete energy law (Lemma 3.1), (ii) the isomorphism property of the mapping $\mathbf{v} \mapsto \mathbf{v} - \mathbf{u} \times \mathbf{v}$ in $\mathbb{R}^3$ (cf. [2, p. 1079]), and, in particular, (iii) the isomorphism property for the linear problem (1.8)–(1.9). For a less general proof of existence of a solution, see section 4.

LEMMA 3.1. *Suppose that* $|\mathbf{m}_h^0(\mathbf{x}_\ell)| = 1$ *for all* $\ell \in L$. *Then the sequence* $\{(\mathbf{m}_h^j, \mathbf{E}_h^j, \mathbf{H}_h^j)\}_{j\geq 0}$ *produced by Algorithm* 1.2 *satisfies for all* $j \geq 0$

$$\text{(i)} \quad |\,\mathbf{m}_h^{j+1}(\mathbf{x}_\ell)\,| = 1 \quad \forall\, \ell \in L\,,$$

$$\text{(ii)} \quad \mathcal{E}_h\big(\{\mathbf{m}_h^{j+1}, \mathbf{H}_h^{j+1}, \mathbf{E}_h^{j+1}\}\big) + k\sum_{\ell=0}^{j} \frac{\alpha\mu_0}{1+\alpha^2} \| \, d_t\mathbf{m}_h^{\ell+1} \, \|_h^2 + \sigma \, \| \, \overline{\mathbf{E}}_h^{\ell+1/2} \, \|_{L^2(\omega)}^2$$

$$= \mathcal{E}_h\big(\{\mathbf{m}_h^0, \mathbf{H}_h^0, \mathbf{E}_h^0\}\big) - k\sum_{\ell=0}^{j} (\overline{\mathbf{J}}_h^{\ell+1/2}, \overline{\mathbf{E}}_h^{\ell+1/2})\,,$$

where

$$\mathcal{E}_h\big(\{\mathbf{m}_h^j, \mathbf{H}_h^j, \mathbf{E}_h^j\}\big) = \frac{\mu_0}{2}\int_\omega |\,\nabla \mathbf{m}_h^j\,|^2 \,\mathrm{d}\mathbf{x} + \int_\Omega \Big[\frac{\mu_0}{2}\,|\,\mathbf{H}_h^j\,|^2 + \frac{\varepsilon_0}{2}\,|\,\mathbf{E}_h^j\,|^2\Big]\,\mathrm{d}\mathbf{x}\,.$$

Proof. Verification of (i) follows from choosing $\boldsymbol{\phi}_h = \varphi_\ell \overline{\mathbf{m}}_h^{j+1/2}(\mathbf{x}_\ell) \in \mathbf{V}_h$ for $\ell \in L$ in (1.7): with the properties of the discrete inner product and the cross product we infer

$$\beta_\ell d_t \big|\mathbf{m}_h^{j+1}(\mathbf{x}_\ell)\big|^2 = \beta_\ell \langle d_t\mathbf{m}_h^{j+1}(\mathbf{x}_\ell), \overline{\mathbf{m}}_h^{j+1/2}(\mathbf{x}_\ell)\rangle = \big(d_t\mathbf{m}_h^{j+1}, \varphi_\ell \overline{\mathbf{m}}_h^{j+1/2}(\mathbf{x}_\ell)\big)_h = 0.$$

Hence, if $|\mathbf{m}_h^j(\mathbf{x}_\ell)| = 1$, then also $|\mathbf{m}_h^{j+1}(\mathbf{x}_\ell)| = 1$. To verify (ii), we first choose $\boldsymbol{\phi}_h = -\big(\tilde{\Delta}_h \overline{\mathbf{m}}_h^{j+1/2} + \mathbf{P}_{\mathbf{V}_h} \overline{\mathbf{H}}_h^{j+1/2}\big)$ to obtain

$$\frac{1}{2} d_t \| \, \nabla \mathbf{m}_h^{j+1} \, \|_{L^2}^2 - (d_t\mathbf{m}_h^{j+1}, \mathbf{P}_{\mathbf{V}_h} \overline{\mathbf{H}}_h^{j+1/2})_h$$

$$= \alpha\big(\overline{\mathbf{m}}_h^{j+1/2} \times d_t\mathbf{m}_h^{j+1}, \tilde{\Delta}_h \overline{\mathbf{m}}_h^{j+1/2} + \mathbf{P}_{\mathbf{V}_h} \overline{\mathbf{H}}_h^{j+1/2}\big)_h\,.$$

Choosing $\boldsymbol{\phi}_h = d_t\mathbf{m}_h^{j+1}$ yields

$$\frac{\alpha}{1+\alpha^2} \| \, d_t\mathbf{m}_h^{j+1} \, \|_h^2 = \alpha\Big(\overline{\mathbf{m}}_h^{j+1/2} \times (\tilde{\Delta}_h \overline{\mathbf{m}}_h^{j+1/2} + \mathbf{P}_{\mathbf{V}_h} \overline{\mathbf{H}}_h^{j+1/2}), d_t\mathbf{m}_h^{j+1}\Big)_h\,.$$

Adding the last two identities and using the definition of $\mathbf{P}_{\mathbf{V}_h}$ lead to

$$\frac{1}{2} d_t \| \, \nabla \mathbf{m}_h^{j+1} \, \|_{L^2}^2 + \frac{\alpha}{1+\alpha^2} \| \, d_t\mathbf{m}_h^{j+1} \, \|_h^2 = (d_t\mathbf{m}_h^{j+1}, \overline{\mathbf{H}}_h^{j+1/2})\,.$$

In a second step, we choose $(\boldsymbol{\varphi}_h, \boldsymbol{\mathfrak{Z}}_h) = (\overline{\mathbf{E}}_h^{j+1/2}, \overline{\mathbf{H}}_h^{j+1/2})$ in (1.8)–(1.9) and add the resulting identities,

$$d_t\Big(\frac{\mu_0}{2} \| \, \mathbf{H}_h^{j+1} \, \|_{L^2}^2 + \frac{\varepsilon_0}{2} \| \, \mathbf{E}_h^{j+1} \, \|_{L^2}^2\Big) + \sigma \| \, \chi_\omega \overline{\mathbf{E}}_h^{j+1/2} \, \|_{L^2}^2$$

$$= -\mu_0\,(d_t\mathbf{m}_h^{j+1}, \overline{\mathbf{H}}_h^{j+1/2}) - (\overline{\mathbf{J}}_h^{j+1/2}, \overline{\mathbf{E}}_h^{j+1/2})\,.$$

Summation of last two identities then proves assertion (ii). ∎

DEFINITION 3.1. *For* $\mathbf{x} \in \Omega$ *and* $t \in [t_j, t_{j+1})$ *define for* $\boldsymbol{\xi}^\ell = \mathbf{m}_h^\ell, \mathbf{H}_h^\ell, \mathbf{E}_h^\ell, \mathbf{J}_h^\ell$, *and* $\ell = j, j+1$,

$$\tilde{\boldsymbol{\xi}}(t,\mathbf{x}) := \frac{t-t_j}{k}\boldsymbol{\xi}_h^{j+1}(\mathbf{x}) + \frac{t_{j+1}-t}{k}\boldsymbol{\xi}_h^{j}(\mathbf{x})\,,$$

$$\tilde{\boldsymbol{\xi}}^-(t,\mathbf{x}) := \boldsymbol{\xi}_h^j(\mathbf{x})\,, \quad \tilde{\boldsymbol{\xi}}^+(t,\mathbf{x}) := \boldsymbol{\xi}_h^{j+1}(\mathbf{x})\,, \quad \overline{\tilde{\boldsymbol{\xi}}}(t,\mathbf{x}) := \overline{\boldsymbol{\xi}}_h^{j+1/2}\,.$$

Given any $T' \geq 0$, equation (ii) in Lemma 3.1 may be rewritten as

$$\mathcal{E}_{(\tilde{\mathbf{m}}^+,\tilde{\mathbf{H}}^+,\tilde{\mathbf{E}}^+)}(T') + \frac{\alpha\mu_0}{1+\alpha^2}\int_0^{T'} ||\tilde{\mathbf{m}}_t||_h^2\,\mathrm{d}t + \sigma\int_0^{T'} \|\,\overline{\tilde{\mathbf{E}}}\,\|_{L^2(\omega)}^2\,\mathrm{d}t \leq \mathcal{E}_{(\tilde{\mathbf{m}}_0,\tilde{\mathbf{H}}_0,\tilde{\mathbf{E}}_0)}(0) + \int_0^{T'} (\overline{\tilde{\mathbf{J}}}, \overline{\tilde{\mathbf{E}}})\,\mathrm{d}t\,. \tag{3.1}$$

We define $\mathbf{P}_{\mathbf{Y}_h} : L^2(\Omega,\mathbb{R}^3) \to \mathbf{Y}_h$ through $(\mathbf{u} - \mathbf{P}_{\mathbf{Y}_h}\mathbf{u}, \boldsymbol{\varphi}_h) = 0$ for all $\boldsymbol{\varphi}_h \in \mathbf{Y}_h$ and all $\mathbf{u} \in L^2(\Omega,\mathbb{R}^3)$. Letting $\big(\boldsymbol{\phi}_h, \boldsymbol{\varphi}_h, \boldsymbol{\mathfrak{Z}}_h\big)(t,\cdot) := \big(\boldsymbol{\mathcal{I}}_{\mathbf{V}_h}\boldsymbol{\phi}, \boldsymbol{\mathcal{I}}_{\mathbf{X}_h}\boldsymbol{\varphi}, \mathbf{P}_{\mathbf{Y}_h}\boldsymbol{\mathfrak{Z}}\big)(t,\cdot)$ for $\boldsymbol{\phi} \in C^\infty(\omega_T;\mathbb{R}^3)$ and $\boldsymbol{\mathfrak{Z}}, \boldsymbol{\varphi} \in \mathcal{D}\big([0,T); C^\infty(\Omega,\mathbb{R}^3) \cap \mathbf{H}_0(\mathbf{curl},\Omega)\big)$ we may rewrite Algorithm 1.2 as follows:

$$\int_0^T (\tilde{\mathbf{m}}_t, \boldsymbol{\phi}_h)_h\,\mathrm{d}t + \alpha\int_0^T (\tilde{\mathbf{m}}^- \times \tilde{\mathbf{m}}_t, \boldsymbol{\phi}_h)_h\,\mathrm{d}t = (1+\alpha^2)\int_0^T \big(\overline{\tilde{\mathbf{m}}} \times (\tilde{\Delta}_h\overline{\tilde{\mathbf{m}}} + \mathbf{P}_{\mathbf{V}_h}\overline{\tilde{\mathbf{H}}}), \boldsymbol{\phi}_h\big)_h\,\mathrm{d}t\,, \tag{3.2}$$

$$\varepsilon_0\int_0^T (\tilde{\mathbf{E}}_t, \boldsymbol{\varphi}_h)\,\mathrm{d}t - \int_0^T (\overline{\tilde{\mathbf{H}}}, \nabla\times\boldsymbol{\varphi}_h)\,\mathrm{d}t + \sigma\int_0^T (\chi_\omega\overline{\tilde{\mathbf{E}}}, \boldsymbol{\varphi}_h)\,\mathrm{d}t = \int_0^T (\overline{\tilde{\mathbf{J}}}, \boldsymbol{\varphi}_h)\,\mathrm{d}t\,, \tag{3.3}$$

$$\mu_0\int_0^T (\tilde{\mathbf{H}}_t, \boldsymbol{\mathfrak{Z}}_h)\,\mathrm{d}t + \int_0^T (\nabla\times\overline{\tilde{\mathbf{E}}}, \boldsymbol{\mathfrak{Z}}_h)\,\mathrm{d}t = -\mu_0\int_0^T (\tilde{\mathbf{m}}_t, \boldsymbol{\mathfrak{Z}}_h)\,\mathrm{d}t\,. \tag{3.4}$$

The a priori bounds in Lemma 3.1 provide the existence of a triple

$$(\,\mathbf{m},\mathbf{H},\mathbf{E}\,) \in \big[L^\infty\big(0,T;W^{1,2}(\omega,\mathbb{S}^2)\big) \cap W^{1,2}(\omega_T;\mathbb{R}^3)\big] \times \big[L^\infty\big(0,T;L^2(\Omega;\mathbb{R}^3)\big)\big]^2,$$

which is the weak limit (as $k,h \to 0$) of a subsequence $\big\{(\,\tilde{\mathbf{m}},\tilde{\mathbf{H}},\tilde{\mathbf{E}}\,)\big\}_{k,h}$, such that

$$\tilde{\mathbf{m}}, \tilde{\mathbf{m}}^\pm, \overline{\tilde{\mathbf{m}}} \overset{*}{\rightharpoonup} \mathbf{m} \ \text{ in } L^\infty\big(0,T;W^{1,2}(\omega,\mathbb{R}^3)\big)\,, \quad \tilde{\mathbf{m}} \rightharpoonup \mathbf{m} \ \text{ in } W^{1,2}(\omega_T,\mathbb{R}^3)\,,$$

$$\tilde{\mathbf{m}}, \tilde{\mathbf{m}}^\pm\,\overline{\tilde{\mathbf{m}}} \to \mathbf{m} \ \text{ in } L^2(\omega_T,\mathbb{R}^3)\,,$$

$$(\,\tilde{\mathbf{H}}, \tilde{\mathbf{H}}^\pm, \overline{\tilde{\mathbf{H}}}; \tilde{\mathbf{E}}, \tilde{\mathbf{E}}^\pm, \overline{\tilde{\mathbf{E}}}\,) \overset{*}{\rightharpoonup} (\,\mathbf{H};\mathbf{E}\,) \ \text{ in } \big[L^\infty\big((0,T);L^2(\Omega,\mathbb{R}^3)\big)\big]^2\,.$$

Since $|\,\tilde{\mathbf{m}}^+(\mathbf{x}_\ell)\,| = 1$ for all $\ell \in L$ and all $t \in (0,T)$ by Lemma 3.1(i), we deduce with a discrete Poincaré inequality that

$$\begin{aligned}\big\||\tilde{\mathbf{m}}^+|^2 - 1\big\|_{L^2(K)} &\leq Ch\,\big\|\nabla\big[|\tilde{\mathbf{m}}^+|^2 - 1\big]\big\|_{L^2(K)}\\ &\leq Ch\,\|(\tilde{\mathbf{m}}^+)^T\nabla\tilde{\mathbf{m}}^+\|_{L^2(K)}^2 \leq Ch\,\|\nabla\tilde{\mathbf{m}}^+\|_{L^2(K)}^2\end{aligned}$$

for all $K \in \mathcal{T}_h$ and hence that $|\tilde{\mathbf{m}}^+| \to 1$ a.e. in Ω_T. In particular, we deduce that $|\mathbf{m}| = 1$ a.e. in Ω_T. Owing to item (ii) of Lemma 3.1 and weak lower semicontinuity of norms we verify that $(\,\mathbf{m}, \mathbf{H}, \mathbf{E}\,)$ satisfies part (4) of Definition 2.1. If $\mathbf{m}_h^0 \to \mathbf{m}_0$ in $L^2(\omega;\mathbb{R}^3)$ as $h \to 0$, then weak continuity of the trace operator yields that $\mathbf{m}(0,\cdot) = \mathbf{m}_0$ in the sense of traces.

Identification of limits in (3.2) apart from the last term can be done as in [4], and we discuss the term which involves $\mathbf{P}_{\mathbf{V}_h}\overline{\tilde{\mathbf{H}}}$. By definition of $\mathbf{P}_{\mathbf{V}_h}$,

$$\begin{aligned}(\overline{\tilde{\mathbf{m}}} \times \mathbf{P}_{\mathbf{V}_h}\overline{\tilde{\mathbf{H}}}, \boldsymbol{\phi}_h)_h &= -(\overline{\tilde{\mathbf{m}}} \times \boldsymbol{\phi}_h, \mathbf{P}_{\mathbf{V}_h}\overline{\tilde{\mathbf{H}}})_h = -\big(\boldsymbol{\mathcal{I}}_{\mathbf{V}_h}(\overline{\tilde{\mathbf{m}}} \times \boldsymbol{\phi}_h), \mathbf{P}_{\mathbf{V}_h}\overline{\tilde{\mathbf{H}}}\big)_h \\ &= -\big(\boldsymbol{\mathcal{I}}_{\mathbf{V}_h}(\overline{\tilde{\mathbf{m}}} \times \boldsymbol{\phi}_h), \overline{\tilde{\mathbf{H}}}\big) \\ &= -\big((\boldsymbol{\mathcal{I}}_{\mathbf{V}_h} - \mathbf{Id})(\overline{\tilde{\mathbf{m}}} \times \boldsymbol{\phi}_h), \overline{\tilde{\mathbf{H}}}\big) - \big(\overline{\tilde{\mathbf{m}}} \times \boldsymbol{\phi}_h, \overline{\tilde{\mathbf{H}}}\big).\end{aligned}$$

Noting that

$$\begin{aligned}\big((\boldsymbol{\mathcal{I}}_{\mathbf{V}_h} - \mathbf{Id})(\overline{\tilde{\mathbf{m}}} \times \boldsymbol{\phi}_h), \overline{\tilde{\mathbf{H}}}\big) &\le Ch^2 \sum_{K\in\mathcal{T}_h} \| D^2(\overline{\tilde{\mathbf{m}}} \times \boldsymbol{\phi}_h) \|_{L^2(K)} \| \overline{\tilde{\mathbf{H}}} \|_{L^2(K)} \\ &\le Ch^2 \| \nabla\overline{\tilde{\mathbf{m}}} \|_{L^2} \| \nabla\boldsymbol{\phi}_h \|_{L^\infty} \| \overline{\tilde{\mathbf{H}}} \|_{L^2},\end{aligned}$$

we deduce that

$$\int_0^T \big(\overline{\tilde{\mathbf{m}}} \times \mathbf{P}_{\mathbf{V}_h}\overline{\tilde{\mathbf{H}}}, \boldsymbol{\phi}_h\big)_h \,\mathrm{d}t \to \int_0^T (\mathbf{m} \times \mathbf{H}, \boldsymbol{\phi})$$

as $h, k \to 0$. To identify the limit in (3.3), we use the identity

$$\int_0^T (\tilde{\mathbf{E}}_t, \boldsymbol{\varphi}_h)\,\mathrm{d}t = -\int_0^T \big(\tilde{\mathbf{E}}, (\boldsymbol{\varphi}_h)_t\big)\,\mathrm{d}t + \big(\mathbf{E}(T,\cdot), \boldsymbol{\varphi}_h(T,\cdot)\big) - \big(\mathbf{E}(0,\cdot), \boldsymbol{\varphi}_h(0,\cdot)\big),$$

and a passage to the limits in the terms on the right-hand side is straightforward owing to their linearity and the convergence properties of $\tilde{\mathbf{E}}$; we proceed accordingly with the leading term in (3.3). It remains to consider the second term in (3.4). Owing to $\nabla \times \mathbf{X}_h \subset \mathbf{Y}_h$, $\mathfrak{Z}_h = \mathbf{P}_{\mathbf{Y}_h}\mathfrak{Z}$, and the properties of $\overline{\tilde{\mathbf{E}}}$, we verify

$$\int_0^T \big(\nabla \times \overline{\tilde{\mathbf{E}}}, \mathfrak{Z}_h\big)dt = \int_0^T \big(\nabla \times \overline{\tilde{\mathbf{E}}}, \mathfrak{Z}\big)dt = \int_0^T (\overline{\tilde{\mathbf{E}}}, \nabla \times \mathfrak{Z})\,\mathrm{d}t \to \int_0^T (\mathbf{E}, \nabla \times \mathfrak{Z})\,\mathrm{d}t.$$

Finally, weak lower semicontinuity of norms and strong convergence of discrete initial data imply the energy inequality in item (4) of Definition 2.1 for the limit $(\,\mathbf{m}, \mathbf{E}, \mathbf{H}\,)$. We have thus proved the following theorem.

THEOREM 3.1. *Let* (2.8)–(2.9) *be valid. Suppose that we have* $|\mathbf{m}_h^0(\mathbf{x}_\ell)| = 1$ *for all* $\ell \in L$ *and let* $\{(\,\mathbf{m}_h^j, \mathbf{E}_h^j, \mathbf{H}_h^j\,)\}_{j\ge 0}$ *solve Algorithm* 1.2. *Assume that* $\mathbf{m}_h^0 \to \mathbf{m}_0$ *in* $W^{1,2}(\omega)$ *and* $(\,\tilde{\mathbf{H}}_h^0, \tilde{\mathbf{E}}_h^0\,) \to (\,\mathbf{H}_0, \mathbf{E}_0\,)$ *in* $L^2(\Omega,\mathbb{R}^3)$ *as* $h \to 0$, *and let* $T > 0$ *be a fixed constant. As* $k, h \to 0$, *a subsequence of* $(\,\tilde{\mathbf{m}}, \tilde{\mathbf{H}}, \tilde{\mathbf{E}}\,)$ *converges weakly to* $(\,\mathbf{m}, \mathbf{H}, \mathbf{E}\,)$ *in* $\big[L^\infty\big(0,T;W^{1,2}(\Omega,\mathbb{S}^2)\big) \cap W^{1,2}(\omega_T,\mathbb{R}^3)\big] \times \big[L^\infty\big((0,T);L^2(\Omega,\mathbb{R}^3)\big)\big]^2$, *and* $(\,\mathbf{m}, \mathbf{H}, \mathbf{E}\,)$ *is a weak solution of* (1.1)–(1.3).

In fact, every weak accumulation point of $(\,\tilde{\mathbf{m}}, \tilde{\mathbf{H}}, \tilde{\mathbf{E}}\,)$ solves (1.1)–(1.3). Note also that no discrete version of the compatibility assumption (2.9) is assumed for $(\,\tilde{\mathbf{H}}_h^0, \tilde{\mathbf{E}}_h^0, \mathbf{m}_h^0\,)$.

4. Solving the nonlinear system. We employ a fixed-point iteration to solve the nonlinear system in each step of Algorithm 1.2: Given $\mathbf{m}_h^j$, $\mathbf{H}_h^j$, and $\mathbf{E}_h^j$ (or approximations $\tilde{\mathbf{m}}_h^j$, $\tilde{\mathbf{H}}_h^j$, and $\tilde{\mathbf{E}}_h^j$), we aim at approximating $\mathbf{w}_h := \overline{\mathbf{m}}_h^{j+1/2}$, $\mathbf{F}_h := \overline{\mathbf{E}}_h^{j+1/2}$, and $\mathbf{G}_h := \overline{\mathbf{H}}_h^{j+1/2}$. The time derivative $d_t\mathbf{m}_h^{j+1}$ is replaced by $\frac{2}{k}(\mathbf{w}_h - \mathbf{m}_h^j)$ and similar expressions for $d_t\mathbf{E}_h^{j+1}$ and $d_t\mathbf{H}_h^{j+1}$. A linearization of the nonlinear term $\big(\mathbf{w}_h \times (\tilde\Delta_h \mathbf{w}_h + \mathbf{P}_{\mathbf{L}^2}\mathbf{G}_h), \boldsymbol{\phi}_h\big)_h$ and the identity $\mathbf{m}_h^j \times d_t\mathbf{m}_h^{j+1} = -\frac{2}{k}\overline{\mathbf{m}}_h^{j+1/2} \times \mathbf{m}_h^j$ lead to the following algorithm.

ALGORITHM 4.1. *Set* $(\mathbf{w}_h^0, \mathbf{F}_h^0, \mathbf{G}_h^0) := (\tilde{\mathbf{m}}_h^j, \tilde{\mathbf{E}}_h^j, \tilde{\mathbf{H}}_h^j)$ *and* $\ell := 0$.

(i) *Compute* $(\mathbf{w}_h^{\ell+1}, \mathbf{F}_h^{\ell+1}, \mathbf{G}_h^{\ell+1}) \in \mathbf{V}_h \times \mathbf{X}_h^0 \times \mathbf{Y}_h$ *such that for all* $(\boldsymbol{\phi}_h, \boldsymbol{\varphi}_h, \boldsymbol{\mathfrak{Z}}_h) \in \mathbf{V}_h \times \mathbf{X}_h^0 \times \mathbf{Y}_h$ *there holds*

$$
\begin{aligned}
&\frac{2}{k}(\mathbf{w}_h^{\ell+1}, \boldsymbol{\phi}_h)_h - \frac{2\alpha}{k}(\mathbf{w}_h^{\ell+1} \times \tilde{\mathbf{m}}_h^j, \boldsymbol{\phi}_h)_h \\
&-(1+\alpha^2)\big(\mathbf{w}_h^{\ell+1} \times (\tilde\Delta_h \mathbf{w}_h^\ell + \mathbf{P}_{\mathbf{V}^h}\mathbf{G}_h^\ell), \boldsymbol{\phi}_h\big)_h = \frac{2}{k}(\tilde{\mathbf{m}}_h^j, \boldsymbol{\phi}_h)_h,
\end{aligned} \tag{4.1}
$$

$$
\frac{2\varepsilon_0}{k}(\mathbf{F}_h^{\ell+1}, \boldsymbol{\varphi}_h) - (\mathbf{G}_h^{\ell+1}, \nabla \times \boldsymbol{\varphi}_h) + \sigma(\chi_\omega \mathbf{F}_h^{\ell+1}, \boldsymbol{\varphi}_h) = \frac{2\varepsilon_0}{k}(\tilde{\mathbf{E}}_h^j, \boldsymbol{\varphi}_h) - (\overline{\mathbf{J}}_h^{j+1/2}, \boldsymbol{\varphi}_h),
$$

$$
\frac{2\mu_0}{k}(\mathbf{G}_h^{\ell+1}, \boldsymbol{\mathfrak{Z}}_h) + (\nabla \times \mathbf{F}_h^{\ell+1}, \boldsymbol{\mathfrak{Z}}_h) + \frac{2\mu_0}{k}(\mathbf{w}_h^{\ell+1}, \boldsymbol{\mathfrak{Z}}_h) = \frac{2\mu_0}{k}(\tilde{\mathbf{H}}_h^j, \boldsymbol{\mathfrak{Z}}_h) + \frac{2\mu_0}{k}(\tilde{\mathbf{m}}_h^j, \boldsymbol{\mathfrak{Z}}_h).
$$

(ii) *Stop and set* $(\tilde{\mathbf{m}}_h^{j+1}, \tilde{\mathbf{E}}_h^{j+1}, \tilde{\mathbf{H}}_h^{j+1}) := 2(\mathbf{w}_h^{\ell+1}, \mathbf{F}_h^{\ell+1}, \mathbf{G}_h^{\ell+1}) - (\tilde{\mathbf{m}}_h^j, \tilde{\mathbf{E}}_h^j, \tilde{\mathbf{H}}_h^j)$ *once*

$$
||\tilde\Delta_h(\mathbf{w}_h^{\ell+1} - \mathbf{w}_h^\ell)||_h + ||\mathbf{G}_h^{\ell+1} - \mathbf{G}_h^\ell||_{L^2} \le \varepsilon. \tag{4.2}
$$

(iii) *Set* $\ell := \ell + 1$ *and go to* (i).

For $\varepsilon \to 0$, the output of the iteration converges to the solution of (1.7)–(1.9) (in case $\tilde{\mathbf{m}}_h^j = \mathbf{m}_h^j$, $\tilde{\mathbf{E}}_h^j = \mathbf{E}_h^j$, and $\tilde{\mathbf{H}}_h^j = \mathbf{H}_h^j$) provided that $k \le ch_{min}^2/(1+\alpha^2)$ with a factor $c > 0$ that depends only on the geometry of $\mathcal{T}_h$.

LEMMA 4.1. *Suppose that* $||\tilde{\mathbf{m}}_h^j||_{L^\infty} \le c_0$. *For all* $\ell \ge 0$ *there exists a unique* $(\mathbf{w}_h^{\ell+1}, \mathbf{F}_h^{\ell+1}, \mathbf{G}_h^{\ell+1})$ *solving* (4.1) *and there holds*

$$
\begin{aligned}
&\sqrt{\mu_0/2}||\mathbf{w}_h^{\ell+1} - \mathbf{w}_h^\ell||_h + \sqrt{2\varepsilon_0}||\mathbf{F}_h^{\ell+1} - \mathbf{F}_h^\ell||_{L^2} + \sqrt{\mu_0}||\mathbf{G}_h^{\ell+1} - \mathbf{G}_h^\ell||_{L^2} \\
&\qquad \le \Theta\Big(\sqrt{\mu_0/2}||\mathbf{w}_h^\ell - \mathbf{w}_h^{\ell-1}||_h + \sqrt{\mu_0}||\mathbf{G}_h^\ell - \mathbf{G}_h^{\ell-1}||_{L^2}\Big)
\end{aligned} \tag{4.3}
$$

with $\Theta = c_1^2 c_0 \sqrt{15}(1+\alpha^2)kh_{min}^{-2}$ *provided that* $c_1^2\sqrt{5}h_{min}^{-2} \ge 1$. *Let* $(\tilde{\mathbf{m}}_h^{j+1}, \tilde{\mathbf{E}}_h^{j+1}, \tilde{\mathbf{H}}_h^{j+1})$ *be the output of Algorithm* 4.1. *There exists a function* $\mathbf{R}^j \in \mathbf{V}_h$ *satisfying* $||\mathbf{R}^j||_h \le \varepsilon$ *such that for all* $(\boldsymbol{\phi}_h, \boldsymbol{\varphi}_h, \boldsymbol{\mathfrak{Z}}_h) \in \mathbf{V}_h \times \mathbf{X}_h^0 \times \mathbf{Y}_h$ *there holds*

$$
\begin{aligned}
&(d_t\tilde{\mathbf{m}}_h^{j+1}, \boldsymbol{\phi}_h)_h + \alpha(\tilde{\mathbf{m}}_h^j \times d_t\tilde{\mathbf{m}}_h^{j+1}, \boldsymbol{\phi}_h)_h \\
&\quad - (1+\alpha^2)\big(\overline{\tilde{\mathbf{m}}}_h^{j+1/2} \times (\tilde\Delta_h \overline{\tilde{\mathbf{m}}}_h^{j+1/2} + \mathbf{P}_{\mathbf{V}^h}\overline{\tilde{\mathbf{H}}}_h^{j+1/2}), \boldsymbol{\phi}_h\big)_h \\
&\quad = (1+\alpha^2)\big(\overline{\tilde{\mathbf{m}}}_h^{j+1/2} \times \mathbf{R}^j, \boldsymbol{\phi}_h\big)_h,
\end{aligned} \tag{4.4}
$$

$$
\frac{\varepsilon_0}{k}(d_t\tilde{\mathbf{E}}_h^{j+1}, \boldsymbol{\varphi}_h) - (\overline{\tilde{\mathbf{E}}}_h^{j+1/2}, \nabla \times \boldsymbol{\varphi}_h) + \sigma(\chi_\omega \overline{\tilde{\mathbf{E}}}_h^{j+1/2}, \boldsymbol{\varphi}_h) = -(\overline{\mathbf{J}}_h^{j+1/2}, \boldsymbol{\varphi}_h),
$$

$$
\frac{\mu_0}{k}(d_t\tilde{\mathbf{H}}_h^{j+1}, \boldsymbol{\mathfrak{Z}}_h) + (\nabla \times \overline{\tilde{\mathbf{E}}}_h^{j+1/2}, \boldsymbol{\mathfrak{Z}}_h) + \frac{\mu_0}{k}(d_t\tilde{\mathbf{m}}_h^{j+1}, \boldsymbol{\mathfrak{Z}}_h) = 0.
$$

Moreover, if $|\tilde{\mathbf{m}}_h^j(\mathbf{x}_m)| = 1$ *for all* $m \in L$*, then there holds* $|\tilde{\mathbf{m}}_h^{j+1}(\mathbf{x}_m)| = 1$ *for all* $m \in L$.

Proof. Step 1. For $(\boldsymbol{\phi}_h, \boldsymbol{\varphi}_h, \mathfrak{Z}_h) = (\mathbf{w}_h^{\ell+1}, \mathbf{F}_h^{\ell+1}, \mathbf{G}_h^{\ell+1})$ the sum of the left-hand sides in (4.1) (after multiplication of the first equation by μ_0) equals

$$\frac{2\mu_0}{k}||\mathbf{w}_h^{\ell+1}||_h^2 + \frac{2\varepsilon_0}{k}||\mathbf{F}_h^{\ell+1}||_{L^2}^2 + \sigma||\chi_\omega \mathbf{F}_h^{\ell+1}||_{L^2}^2 + \frac{2\mu_0}{k}||\mathbf{G}_h^{\ell+1}||_{L^2}^2 + \frac{2\mu_0}{k}(\mathbf{w}_h^{\ell+1}, \mathbf{G}_h^{\ell+1}).$$

Since $2(\mathbf{w}_h^{\ell+1}, \mathbf{G}_h^{\ell+1}) \geq -||\mathbf{w}_h^{\ell+1}||_h^2 - ||\mathbf{G}_h^{\ell+1}||_{L^2}^2$, the bilinear form defined by the left-hand side of (4.1) is positive definite on $\big[\mathbf{V}_h \times \mathbf{X}_h^0 \times \mathbf{Y}_h\big]^2$ and (4.1) admits a unique solution.

Step 2. We next control $||\mathbf{w}_h^{\ell+1}||_{L^\infty}$ uniformly for all $\ell \geq 0$: Let $m \in L$ be such that $||\mathbf{w}_h^{\ell+1}||_{L^\infty} = |\mathbf{w}_h^{\ell+1}(\mathbf{x}_m)|$. Putting $\boldsymbol{\phi}_h = \varphi_m \mathbf{w}_h^{\ell+1}(\mathbf{x}_m)$ in the first equation of (4.1) leads to

$$||\mathbf{w}_h^{\ell+1}||_{L^\infty} = |\mathbf{w}_h^{\ell+1}(\mathbf{x}_m)| \leq |\tilde{\mathbf{m}}_h^j(\mathbf{x}_m)| \leq ||\tilde{\mathbf{m}}_h^j||_{L^\infty} \leq c_0. \tag{4.5}$$

Step 3. Subtraction of two subsequent equations of iteration (4.1) shows

$$\begin{aligned}
&\frac{2}{k}(\mathbf{w}_h^{\ell+1} - \mathbf{w}_h^\ell, \boldsymbol{\phi}_h)_h - \frac{2\alpha}{k}\big((\mathbf{w}_h^{\ell+1} - \mathbf{w}_h^\ell) \times \tilde{\mathbf{m}}_h^j, \boldsymbol{\phi}_h\big)_h \\
&\quad -(1+\alpha^2)\big((\mathbf{w}_h^{\ell+1} - \mathbf{w}_h^\ell) \times (\tilde{\Delta}_h \mathbf{w}_h^\ell + \mathbf{P}_{\mathbf{V}^h}\mathbf{G}_h^\ell), \boldsymbol{\phi}_h\big)_h \\
&\quad -(1+\alpha^2)\big(\mathbf{w}_h^\ell \times [(\tilde{\Delta}_h \mathbf{w}_h^\ell + \mathbf{P}_{\mathbf{V}^h}\mathbf{G}_h^\ell) - (\tilde{\Delta}_h \mathbf{w}_h^{\ell-1} + \mathbf{P}_{\mathbf{V}^h}\mathbf{G}_h^{\ell-1})], \boldsymbol{\phi}_h\big)_h = 0, \\
&\frac{2\varepsilon_0}{k}(\mathbf{F}_h^{\ell+1} - \mathbf{F}_h^\ell, \boldsymbol{\varphi}_h) - (\mathbf{G}_h^{\ell+1} - \mathbf{G}_h^\ell, \nabla \times \boldsymbol{\varphi}_h) + \sigma\big(\chi_\omega(\mathbf{F}_h^{\ell+1} - \mathbf{F}_h^\ell), \boldsymbol{\varphi}_h\big) = 0, \\
&\frac{2\mu_0}{k}(\mathbf{G}_h^{\ell+1} - \mathbf{G}_h^\ell, \mathfrak{Z}_h) + \big(\nabla \times (\mathbf{F}_h^{\ell+1} - \mathbf{F}_h^\ell), \mathfrak{Z}_h\big) + \frac{2\mu_0}{k}(\mathbf{w}_h^{\ell+1} - \mathbf{w}_h^\ell, \mathfrak{Z}_h) = 0.
\end{aligned}$$

Adding the three identities (after multiplying the first equation by μ_0) and choosing $(\boldsymbol{\phi}_h, \boldsymbol{\varphi}_h, \mathfrak{Z}_h) = (\mathbf{w}_h^{\ell+1} - \mathbf{w}_h^\ell, \mathbf{F}_h^{\ell+1} - \mathbf{F}_h^\ell, \mathbf{G}_h^{\ell+1} - \mathbf{G}_h^\ell)$ provides

$$\begin{aligned}
&\frac{2\mu_0}{k}||\mathbf{w}_h^{\ell+1} - \mathbf{w}_h^\ell||_h^2 + \frac{2\varepsilon_0}{k}||\mathbf{F}_h^{\ell+1} - \mathbf{F}_h^\ell||_{L^2}^2 + \sigma||\chi_\omega(\mathbf{F}_h^{\ell+1} - \mathbf{F}_h^\ell)||_{L^2}^2 \\
&\quad + \frac{2\mu_0}{k}||\mathbf{G}_h^{\ell+1} - \mathbf{G}_h^\ell||_{L^2}^2 \\
&= -\frac{2\mu_0}{k}(\mathbf{w}_h^{\ell+1} - \mathbf{w}_h^\ell, \mathbf{G}_h^{\ell+1} - \mathbf{G}_h^\ell) \\
&\quad + \mu_0(1+\alpha^2)\big(\mathbf{w}_h^\ell \times [(\tilde{\Delta}_h \mathbf{w}_h^\ell + \mathbf{P}_{\mathbf{V}^h}\mathbf{G}_h^\ell) - (\tilde{\Delta}_h \mathbf{w}_h^{\ell-1} + \mathbf{P}_{\mathbf{V}^h}\mathbf{G}_h^{\ell-1})], \mathbf{w}_h^{\ell+1} - \mathbf{w}_h^\ell\big)_h \\
&\leq \frac{3\mu_0}{2k}||\mathbf{w}_h^{\ell+1} - \mathbf{w}_h^\ell||_h^2 + \frac{\mu_0}{k}||\mathbf{G}_h^{\ell+1} - \mathbf{G}_h^\ell||_{L^2}^2 \\
&\quad + k\frac{\mu_0}{2}(1+\alpha^2)^2 c_0^2\big(||\tilde{\Delta}_h(\mathbf{w}_h^\ell - \mathbf{w}_h^{\ell-1})||_h + ||\mathbf{P}_{\mathbf{V}^h}(\mathbf{G}_h^\ell - \mathbf{G}_h^{\ell-1})||_h\big)^2.
\end{aligned}$$

The estimates $||\tilde{\Delta}_h \boldsymbol{\phi}_h||_h \leq c_1^2\sqrt{5}h_{min}^{-2}||\boldsymbol{\phi}_h||_h$ and $||\mathbf{P}_{\mathbf{V}^h}\mathfrak{Z}_h||_h \leq ||\mathfrak{Z}_h||_{L^2}$ imply (4.3).

Step 4. In order to prove (4.4) we replace $d_t\tilde{\mathbf{m}}_h^{j+1} = \frac{2}{k}(\mathbf{w}_h^{\ell+1} - \tilde{\mathbf{m}}_h^j)$, $\overline{\tilde{\mathbf{m}}}_h^{j+1/2} = \mathbf{w}_h^{\ell+1}$, etc. in (4.1) to verify with $\mathbf{R}^j := (\tilde{\Delta}_h \mathbf{w}_h^\ell + \mathbf{P}_{\mathbf{V}^h}\mathbf{G}_h^\ell) - (\tilde{\Delta}_h \mathbf{w}_h^{\ell+1} + \mathbf{P}_{\mathbf{V}^h}\mathbf{G}_h^{\ell+1})$

$$
\begin{aligned}
&(d_t\tilde{\mathbf{m}}_h^{j+1}, \boldsymbol{\phi}_h)_h - \frac{2\alpha}{k}(\overline{\tilde{\mathbf{m}}}_h^{j+1/2} \times \tilde{\mathbf{m}}_h^j, \boldsymbol{\phi}_h)_h \\
&\qquad - (1+\alpha^2)\big(\overline{\tilde{\mathbf{m}}}_h^{j+1/2} \times (\tilde{\Delta}_h \overline{\tilde{\mathbf{m}}}_h^{j+1/2} + \mathbf{P}_{\mathbf{V}^h}\overline{\tilde{\mathbf{H}}}_h^{j+1/2}), \boldsymbol{\phi}_h\big)_h \\
&\qquad = (1+\alpha^2)\big(\overline{\tilde{\mathbf{m}}}_h^{j+1/2} \times \mathbf{R}^j, \boldsymbol{\phi}_h\big)_h, \\
&\frac{\varepsilon_0}{k}\,(d_t\tilde{\mathbf{E}}_h^{j+1}, \boldsymbol{\varphi}_h) - (\overline{\tilde{\mathbf{E}}}_h^{j+1/2}, \nabla \times \boldsymbol{\varphi}_h) + \sigma(\chi_\omega \overline{\tilde{\mathbf{E}}}_h^{j+1/2}, \boldsymbol{\varphi}_h) = -(\overline{\mathbf{J}}_h^{j+1/2}, \boldsymbol{\varphi}_h), \\
&\frac{\mu_0}{k}\,(d_t\tilde{\mathbf{H}}_h^{j+1}, \boldsymbol{\mathfrak{Z}}_h) + (\nabla \times \overline{\tilde{\mathbf{E}}}_h^{j+1/2}, \boldsymbol{\mathfrak{Z}}_h) + \frac{\mu_0}{k}\,(d_t\tilde{\mathbf{m}}_h^{j+1}, \boldsymbol{\mathfrak{Z}}_h) = 0.
\end{aligned}
$$

Using $\frac{2}{k}\overline{\tilde{\mathbf{m}}}_h^{j+1/2} \times \tilde{\mathbf{m}}_h^j = -\tilde{\mathbf{m}}_h^j \times d_t\tilde{\mathbf{m}}_h^{j+1}$ we verify (4.4), and the stopping criterion implies $||\mathbf{R}||_h \le \varepsilon$.

Step 5. We choose $\boldsymbol{\phi}_h = \varphi_m \overline{\tilde{\mathbf{m}}}_h^{j+1/2}(\mathbf{x}_m)$ in (4.4) to verify $d_t|\tilde{\mathbf{m}}_h^{j+1}(\mathbf{x}_m)|^2 = 0$. ∎

Remark 4.1. Convergence to a weak solution of MLLG of approximations satisfying (4.4) as $(k, h, \varepsilon) \to 0$ such that $\varepsilon = o(h_{min}^2)$ is verified as in section 3. The only difference is a perturbed energy law: instead of (ii) in Lemma 3.1 one has

$$
\begin{aligned}
&\mathcal{E}_h\big(\{\tilde{\mathbf{m}}_h^{j+1}, \tilde{\mathbf{H}}_h^{j+1}, \tilde{\mathbf{E}}_h^{j+1}\}\big) + k\sum_{\ell=0}^{j}(1-\varepsilon)\frac{\alpha\mu_0}{1+\alpha^2}\| d_t\tilde{\mathbf{m}}_h^{\ell+1} \|_h^2 + \sigma \| \overline{\tilde{\mathbf{E}}}_h^{\ell+1/2} \|_{L^2(\omega)}^2 \\
&\qquad \le \mathcal{E}_h\big(\{\tilde{\mathbf{m}}_h^0, \tilde{\mathbf{H}}_h^0, \tilde{\mathbf{E}}_h^0\}\big) - k\sum_{\ell=0}^{j}(\overline{\mathbf{J}}_h^{\ell+1/2}, \overline{\tilde{\mathbf{E}}}_h^{\ell+1/2}) \\
&\qquad + k\sum_{\ell=0}^{j}\left\{(1+\alpha^2)\| \mathbf{R}^\ell \|_h\big(c_3 h_{min}^{-2} + \| \tilde{\mathbf{H}}_h^{\ell+1/2} \|\big) + \frac{1}{4\varepsilon}(1+\alpha^2)\alpha \| \mathbf{R}^\ell \|_h^2\right\}.
\end{aligned}
$$

5. Computational experiments.

5.1. Physical model. Below, let $\mathbf{j} = \frac{\mathbf{J}}{M_s}$, $\mathbf{h} = \frac{\mathbf{H}}{M_s}$, and $\mathbf{e} = \frac{\mathbf{E}}{M_s}$ denote the scaled electric current, magnetic field, and electric field, respectively.

In practical computations, the following physical constants have to be included in the model: the permeability of vacuum μ_0^*, the permittivity of vacuum ε_0^*, the exchange constant A^*, the anisotropy constant K^*, the saturation magnetization M_s, and the gyromagnetic ratio γ. The direction of the uniaxial anisotropy is characterized by a unit vector $\mathbf{p} \in \mathbb{S}^2$. Without loss of generality, we assume that $\mathbf{p}$ is parallel to one of the coordinate axes. The effective field then becomes

$$
\mathbf{h}_{\text{eff}} = A\Delta\mathbf{m} + K\langle \mathbf{m}, \mathbf{p}\rangle \mathbf{p} + \mathbf{h}, \tag{5.1}
$$

with constants

$$
A = \frac{2A^*}{\mu_0^* M_s^2}, \quad K = \frac{2K^*}{\mu_0^* M_s^2}.
$$

The Landau–Lifshitz–Gilbert equation (LLG), which after rescaling in time takes a dimensionless form

$$\mathbf{m}_t + \alpha\,\mathbf{m}\times\mathbf{m}_t = (1+\alpha^2)\,\mathbf{m}\times\mathbf{h}_{\rm eff} \quad \text{in} \quad \omega_T\,, \tag{5.2}$$

now corresponds to a fully physical situation with time measured in units of $(\gamma M_s)^{-1}$s; cf., e.g., [20].

To have the same time scales for the whole MLLG system, the Maxwell equations have to be rescaled in time appropriately. After a change of the time variable and an additional scaling by factor M_s^{-1}, we obtain

$$\varepsilon_0\,\mathbf{e}_t + \nabla\times\mathbf{h} + \sigma\,\chi_\omega\mathbf{e} = -\mathbf{j} \quad \text{on} \quad \Omega_{T^*}, \tag{5.3}$$

$$\mu_0\,\mathbf{h}_t - \nabla\times\mathbf{e} = -\mu_0\,\mathbf{m}_t \qquad \text{on} \quad \Omega_{T^*}\,, \tag{5.4}$$

where $\varepsilon_0 = \gamma M_s\varepsilon_0^*$, $\mu_0 = \gamma M_s\mu_0^*$, and $T^* = \gamma M_s\,T$.

A counterpart of Lemma 3.1 for (5.2)–(5.4) reads as follows.

LEMMA 5.1. *Suppose that* $|\mathbf{m}_h^0(\mathbf{x}_\ell)| = 1$ *for all* $\ell \in L$. *Then the sequence* $\{(\,\mathbf{m}_h^j, \mathbf{E}_h^j, \mathbf{H}_h^j)\}_{j\ge 0}$ *produced by Algorithm* 1.2 *satisfies for all* $j \ge 0$

(i) $\quad |\,\mathbf{m}_h^{j+1}(\mathbf{x}_\ell)\,| = 1 \quad \forall\,\ell \in L\,,$

(ii) $$\mathcal{E}_T := \mathcal{E}_h^*\big(\{\mathbf{m}_h^{j+1}, \mathbf{h}_h^{j+1}, \mathbf{e}_h^{j+1}\}\big) + k\sum_{\ell=0}^{j}\frac{\alpha\mu_0 M_s}{1+\alpha^2}\|\,d_t\mathbf{m}_h^{\ell+1}\,\|_h^2 + \sigma\,\|\,\overline{\mathbf{e}}_h^{\ell+1/2}\,\|_{L^2(\omega)}^2$$

$$= \mathcal{E}_h^*\big(\{\mathbf{m}_h^0, \mathbf{h}_h^0, \mathbf{e}_h^0\}\big) - k\sum_{\ell=0}^{j}(\bar{\mathbf{j}}_h^{\ell+1/2}, \overline{\mathbf{e}}_h^{\ell+1/2})\,,$$

where

$$\begin{aligned}\mathcal{E}_h^*\big(\{\mathbf{m}_h^j, \mathbf{h}_h^j, \mathbf{e}_h^j\}\big) &= \frac{\mu_0 M_s}{2}\int_\omega\Big[A|\,\nabla\mathbf{m}_h^j\,|^2 + K\boldsymbol{\mathcal{I}}_{\mathbf{V}_h}\Big(1-\langle\mathbf{m}_h^j,\mathbf{p}\rangle^2\Big)\Big]\,d\mathbf{x}\\ &\quad + \int_\Omega\Big[\frac{\mu_0}{2}\,|\,\mathbf{h}_h^j\,|^2 + \frac{\varepsilon_0}{2}\,|\,\mathbf{e}_h^j\,|^2\Big]\,d\mathbf{x}\\ &:= \mathcal{E}_{ex} + \mathcal{E}_{anis} + \mathcal{E}_H + \mathcal{E}_E\,.\end{aligned}$$

In the following we refer to $\mathcal{E}_T$ as the total energy and to the terms $\mathcal{E}_{ex}$, $\mathcal{E}_{an}$, $\mathcal{E}_H$, and $\mathcal{E}_E$ as the exchange energy, anisotropy energy, magnetic field energy, and electric field energy, respectively.

Remark 5.1. To obtain physically relevant results, the initial condition for the LLG system should satisfy the "divergence-free" constraint from, i.e., $\mathrm{div}_h(\mathbf{h}_h^0 + \chi_\omega\mathbf{m}_h^0) = 0$. This can be achieved by taking $\mathbf{h}_h^0 = \mathbf{h}_*^0 - \chi_\omega(\mathbf{P}_h^*\mathbf{m}_h^0)$, with $\mathbf{h}_*^0 \in \mathbf{Y}_h$ such that $\mathrm{div}_h\mathbf{h}_*^0 = 0$. The projection $\mathbf{P}_h^* : \mathbf{V}_h \to \mathbf{Y}_h$ is for $\mathbf{u}_h \in \mathbf{V}_h$ defined through $(\mathbf{P}_h^*\mathbf{u}_h, \boldsymbol{\mathfrak{Z}}_h) = (\mathbf{u}_h, \boldsymbol{\mathfrak{Z}}_h)$ for all $\boldsymbol{\mathfrak{Z}}_h \in \mathbf{Y}_h$. Since $\mathbf{m}_h^j$ is piecewise linear, we have that the value of $\mathbf{P}_h^*\mathbf{m}_h^j$ on an element $K \in \mathcal{T}_h\big|_\omega$ corresponds to the value of $\mathbf{m}_h^j$ in the barycenter of K. Further, we have from (1.9) that

$$\begin{aligned}&\frac{\mu_0}{k}\,(\mathbf{h}_h^{j+1}, \boldsymbol{\mathfrak{Z}}_h) + \frac{\mu_0}{k}\,(\mathbf{P}_h^*\mathbf{m}_h^{j+1}, \boldsymbol{\mathfrak{Z}}_h)\\ &\qquad = -(\nabla\times\mathbf{e}_h^{j+1/2}, \boldsymbol{\mathfrak{Z}}_h) + \frac{\mu_0}{k}\,(\mathbf{h}_h^j, \boldsymbol{\mathfrak{Z}}_h) + \frac{\mu_0}{k}\,(\mathbf{P}_h^*\mathbf{m}_h^j, \boldsymbol{\mathfrak{Z}}_h) \quad \forall\,\boldsymbol{\mathfrak{Z}}_h \in \mathbf{Y}_h.\end{aligned}$$

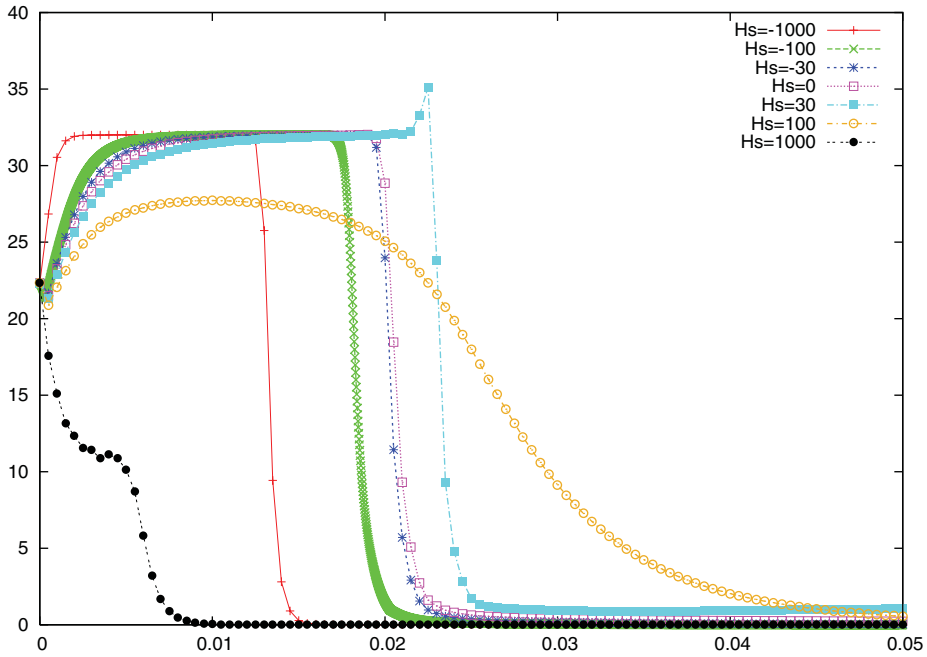


FIG. 1. *Example* 5.1: *Plot of* $t \mapsto \|\nabla \mathbf{m}_h(t,\cdot)\|_\infty$.

From the previous equation, it can be deduced by induction that $\mathrm{div}_h(\mathbf{h}_h^{j+1} + \chi_\omega(\mathbf{P}_{\mathbf{Y}_h}\mathbf{m}_h^{j+1})) = 0$ is satisfied pointwise in Ω. The above arguments remain valid for (4.1). In our experiments, we simply take $\mathbf{h}_*^0$ to be a constant vector field.

5.2. Solution of the discrete system. Without loss of generality, we consider the discrete system with $\sigma = 0$. The first equation from (4.1), corresponding to the discrete LLG equation, is efficiently solved by the biconjugate gradient stabilized (BiCGStab) method. The solution of the second and third equations from (4.1) is equivalent to solving a discrete algebraic system of the form

$$\begin{pmatrix} \mathbf{A} & -\mathbf{C}^T \\ \mathbf{C} & \mathbf{B} \end{pmatrix} \begin{pmatrix} \overline{\mathbf{e}} \\ \overline{\mathbf{h}} \end{pmatrix} = \begin{pmatrix} \overline{\mathbf{f}} \\ \overline{\mathbf{g}} \end{pmatrix}, \tag{5.5}$$

where

$$\mathbf{A}_{ij} = \frac{2\varepsilon_0}{k}(\boldsymbol{\varphi}_h^i, \boldsymbol{\varphi}_h^j), \quad \mathbf{B}_{ij} = \frac{2\mu_0}{k}(\boldsymbol{\mathfrak{Z}}_h^i, \boldsymbol{\mathfrak{Z}}_h^j), \quad \mathbf{C}_{ij} = (\nabla \times \boldsymbol{\varphi}_h^j, \boldsymbol{\mathfrak{Z}}_h^i),$$

and the vectors of unknowns $\overline{\mathbf{e}} = \{e_i\}$, $\overline{\mathbf{h}} = \{h_i\}$ are defined through

$$\mathbf{F}_h = \sum_i e_i \boldsymbol{\varphi}_h^i, \quad \mathbf{G}_h = \sum_i h_i \boldsymbol{\mathfrak{Z}}_h^i.$$

Similarly, we define $\overline{\mathbf{f}} = \{f_i\}$, $\overline{\mathbf{g}} = \{g_i\}$, where $\mathbf{f} = \sum_i f_i \boldsymbol{\varphi}_h^i$ and $\mathbf{g} = \sum_i g_i \boldsymbol{\mathfrak{Z}}_h^i$ represent the right-hand sides of the discrete problem.

The system (5.5) can be effectively solved by a preconditioned inexact Uzawa method which consists of two steps:

1. $\overline{\mathbf{h}}^n = \mathbf{B}^{-1}(\overline{\mathbf{g}} - \mathbf{C}\overline{\mathbf{e}}^n)$,

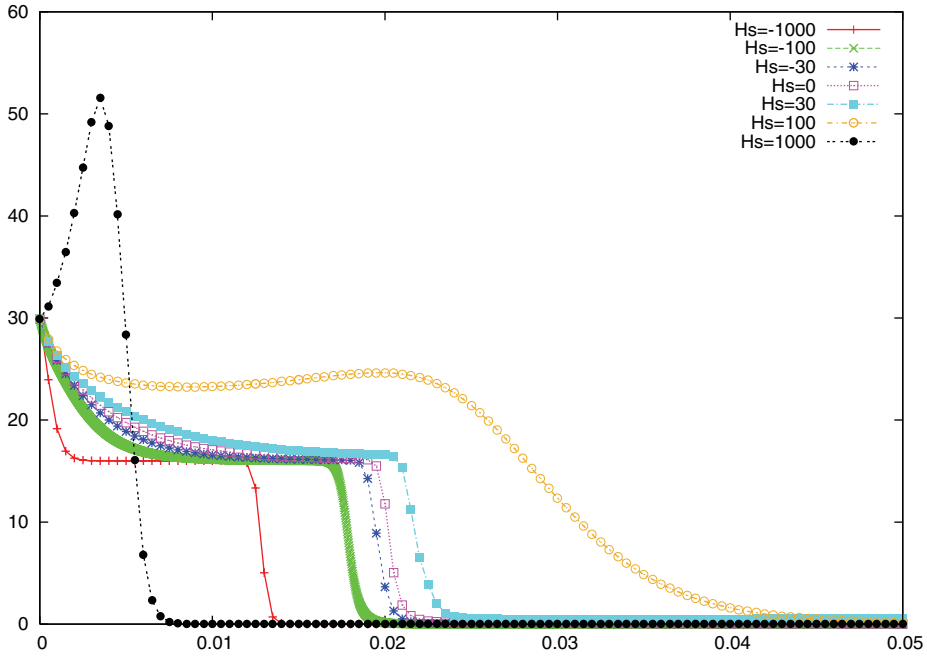


FIG. 2. *Example* 5.1: *Plot of* $t \mapsto \|\nabla \mathbf{m}_h(t,\cdot)\|_2$.

2. $\overline{\mathbf{e}}^{n+1} = \overline{\mathbf{e}}^n + \rho \mathbf{S}^{-1}(\overline{\mathbf{f}} + \mathbf{C}^T \overline{\mathbf{h}}^n - \mathbf{A}\overline{\mathbf{e}}^n)$.

Here, $\rho > 0$ is a constant (we set $\rho = 1$ below), and $\mathbf{S}^{-1}$ is a suitably chosen preconditioner that can considerably speed up the convergence of the Uzawa iterations. Note that the computation of $\mathbf{B}^{-1}$ in the first step of the above Uzawa algorithm is trivial, since the matrix $\mathbf{B}$ is diagonal, owing to the choice of piecewise constant functions.

The preconditioner $\mathbf{S}$ is an approximation of the Schur complement, i.e.,

$$\mathbf{S} \approx (\mathbf{C}^T \mathbf{B}^{-1} \mathbf{C}) + \mathbf{A}\,.$$

The construction of the preconditioner is motivated by the fact that our formulation can be (formally) considered as a mixed approximation of an eddy current problem of the form

$$\frac{2\varepsilon_0}{k}(\mathbf{F}_h^*, \boldsymbol{\varphi}) + \frac{k}{2\mu_0}(\nabla \times \mathbf{F}_h^*, \nabla \times \boldsymbol{\varphi}) = (\mathbf{f}, \boldsymbol{\varphi}) + \frac{k}{2\mu_0}(\nabla \times \mathbf{g}, \boldsymbol{\varphi}) \qquad \forall \boldsymbol{\varphi} \in \mathbf{H}_0(\mathbf{curl}; \Omega).$$

In matrix notation we have $\mathbf{S} = \mathbf{M} + \mathbf{R}$, with matrices

$$\mathbf{M}_{ij} = \frac{2\varepsilon_0}{k}(\boldsymbol{\varphi}_h^i, \boldsymbol{\varphi}_h^j)\,, \quad \mathbf{R}_{ij} = \frac{k}{2\mu_0}(\nabla \times \boldsymbol{\varphi}_h^i, \nabla \times \boldsymbol{\varphi}_h^j)\,.$$

We compute the approximation of $\mathbf{S}^{-1}$ by the BiCGStab algorithm with at most 50 iterations for every substep. A better approximation of $\mathbf{S}^{-1}$ can be obtained by using a multigrid method for eddy current equations (see, e.g., [9]), which together with the moderate number of outer Uzawa iterations gives an effective method with multigrid complexity.

It is possible to eliminate $\overline{\mathbf{h}}$ from (5.5), which leads to a system of equations in the Schur complement form

$$\big(\mathbf{C}^T \mathbf{B}^{-1} \mathbf{C} + \mathbf{A}\big)\,\overline{\mathbf{e}} = \overline{\mathbf{f}} + \mathbf{C}^T \mathbf{B}^{-1} \overline{\mathbf{g}}\,. \tag{5.6}$$

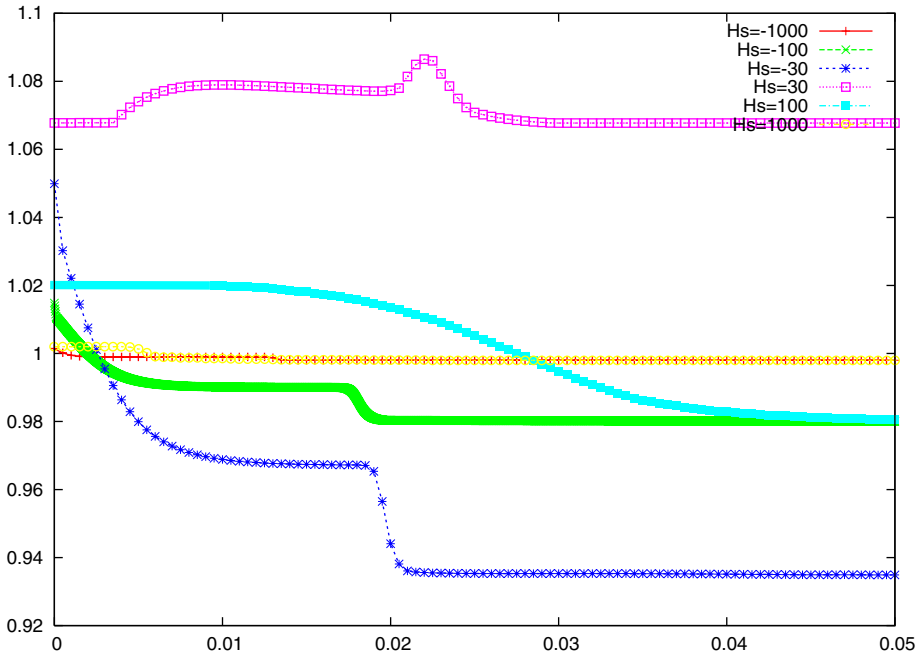


FIG. 3. *Example* 5.1: *Plot of* $t \mapsto \|\mathbf{h}_h(t,\cdot)\|_\infty/|H_s|$.

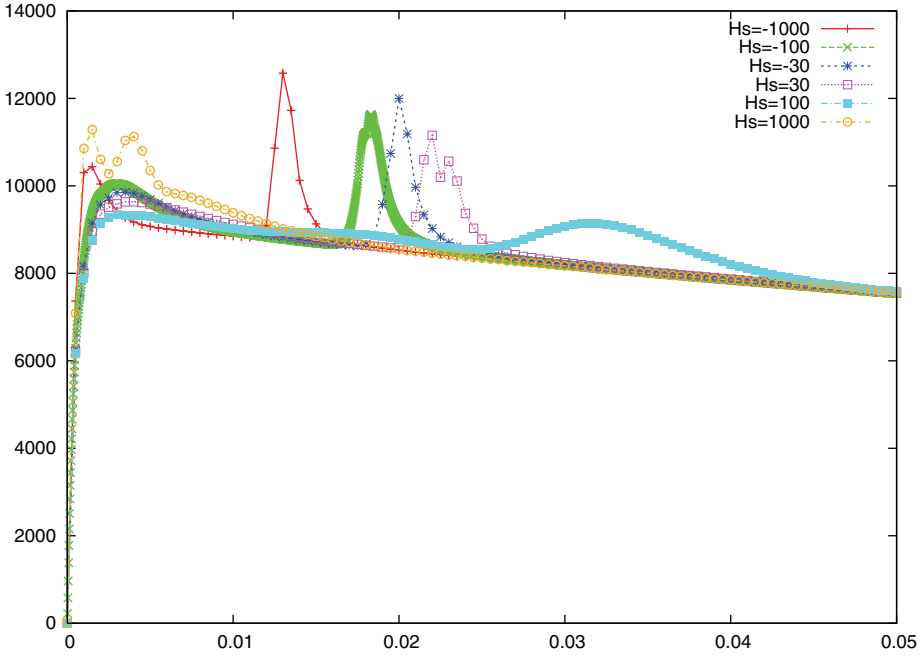


FIG. 4. *Example* 5.1: *Plot of* $t \mapsto \|\mathbf{e}_h(t,\cdot)\|_\infty$.

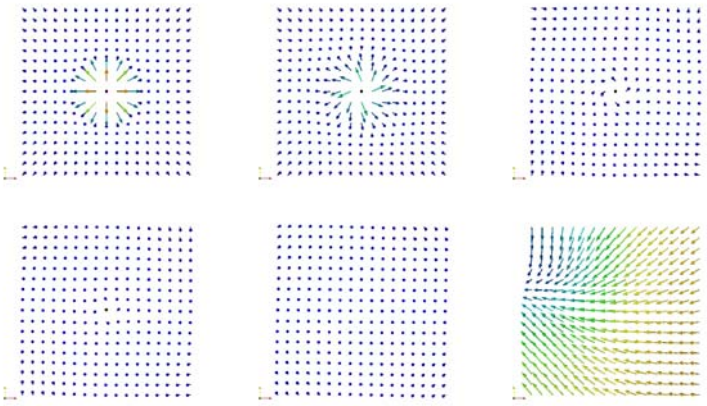


FIG. 5. *Example* 5.1: *Magnetization for* $H_s = 30$ *at times* $t = 0, 0.01, 0.015, 0.020, 0.030, 0.214$ *for* $h = 1/2^4$ *(from left to right, from top to bottom).*

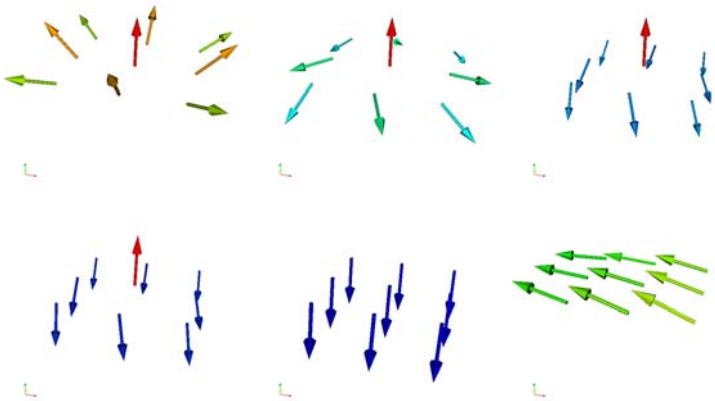


FIG. 6. *Example* 5.1: *Details of the magnetization for* $H_s = 30$ *near* $\mathbf{x} = (0.5, 0.5, 0.5)$ *at times* $t = 0, 0.01, 0.015, 0.020, 0.030, 0.214$ *for* $h = 1/2^4$ *(from left to right, from top to bottom).*

An alternative approach to the Uzawa algorithm is to solve (5.6) by the conjugate gradient method; see, e.g., [17] and the references therein. The conjugate gradient algorithm needs to evaluate $\mathbf{B}^{-1}$, which is trivial, since $\mathbf{B}$ is diagonal. We can speed up the convergence of the conjugate gradient algorithm by using a suitable preconditioner. In our case, one could use the same preconditioner as for the Uzawa algorithm. We use the conjugate gradient algorithm for (5.6) without preconditioning in our experiments, since it proves to be slightly faster than the preconditioned Uzawa algorithm. We expect both methods to have comparable performance with effective multigrid preconditioners; cf. [17].

Remark 5.2. We observe slow convergence of the algebraic solvers for domains of size $\mathcal{O}(10^{-6})$, i.e., in practical applications. The convergence properties of the solvers improve for larger values of the gyromagnetic ratio γ or for smaller sizes of the time step k. We believe that the convergence rates can be substantially improved with a suitable multigrid preconditioner. Note that similar problems with solver

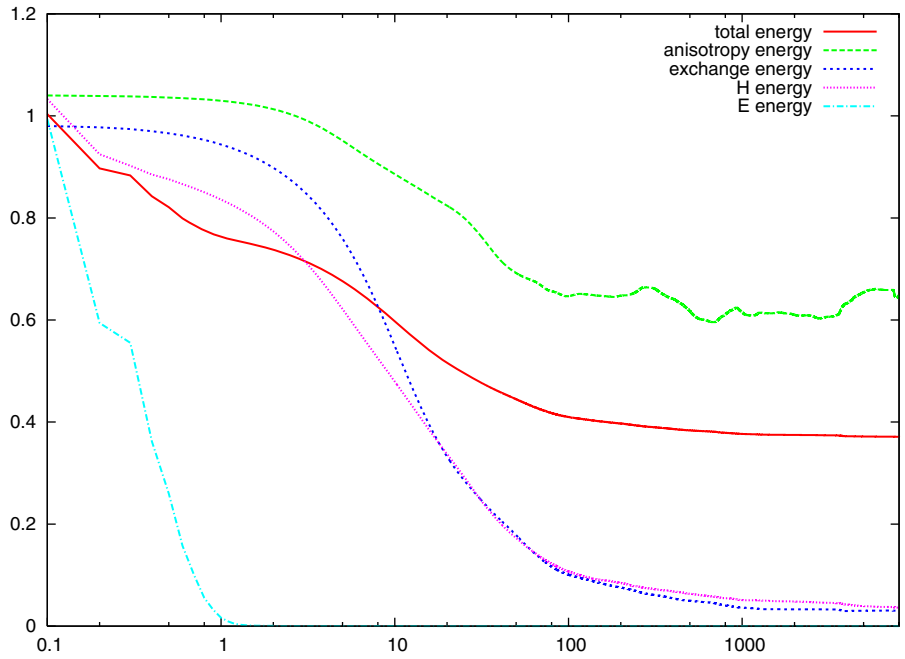


FIG. 7. *Example* 5.2: *Evolution of the energies,* $\log(t) \mapsto \mathcal{E}_T(t)/22$, $\mathcal{E}_{ex}(t)/11$, $\mathcal{E}_{an}(t)/0.04$, $\mathcal{E}_H(t)/7$, $\mathcal{E}_E(t)/4$.

convergence in micromagnetic applications are reported in [22], where a conjugate gradient algorithm is applied to the eddy current formulation of Maxwell's equations.

5.3. Computational results. Our computational code is based on the finite element package ALBERT (see [19]) with tetrahedral meshes in three dimensions. We use standard piecewise linear elements for the discretization of $\mathbf{m}_h$, with degrees of freedom (DOFs) located at the vortices of the mesh, and piecewise constant elements for $\mathbf{h}_h$, with DOFs located at the barycenters of the mesh elements. We discretize $\mathbf{e}_h$ by edge elements of the first kind, with one DOF per every edge of the mesh. The edge elements of the first kind have only slightly worse approximation properties than the edge elements of the second kind; however, the latter need two DOFs per edge. In our code we replace (4.2) in Algorithm 4.1 by a slightly more practical stopping criterion for the fixed-point iterations, i.e.,

$$\begin{aligned} ||\mathbf{w}_h^{\ell+1} - \mathbf{w}_h^{\ell}||_h &\leq h^2\varepsilon, \\ ||\mathbf{G}_h^{\ell+1} - \mathbf{G}_h^{\ell}||_{L^2} &\leq \varepsilon. \end{aligned}$$

We choose $\varepsilon = 10^{-8}$ in our computations. The convergence of the fixed-point iterations is attained after at most six steps in all presented experiments. We observe monotone decrease of the discrete energy from Lemma 5.1 in all experiments, which confirms good numerical convergence of Algorithm 4.1.

The first example is academic and studies possible finite-time blow-up behavior of weak solutions to (1.1)–(1.5) with the help of Algorithm 1.2. We say that (discrete) finite-time blow-up occurs if the sequence $\left\{||\nabla \mathbf{m}_h^j||_{L^\infty}\right\}_j$ attains the maximum value of the mapping $\mathbf{v}_h \mapsto ||\nabla \mathbf{v}_h||_{L^\infty}$ among functions $\mathbf{v}_h \in \mathbf{V}_h$ satisfying $|\mathbf{v}_h(\mathbf{x}_\ell)| = 1$ for all $\ell \in L$. Corresponding studies are reported in [4, 3] for $\mathbf{h}_{\text{eff}} = \Delta\mathbf{m}$ and $\omega \subset \mathbb{R}^2$, where blow-up is computationally evidenced for supercritical (initial) data. Here, we

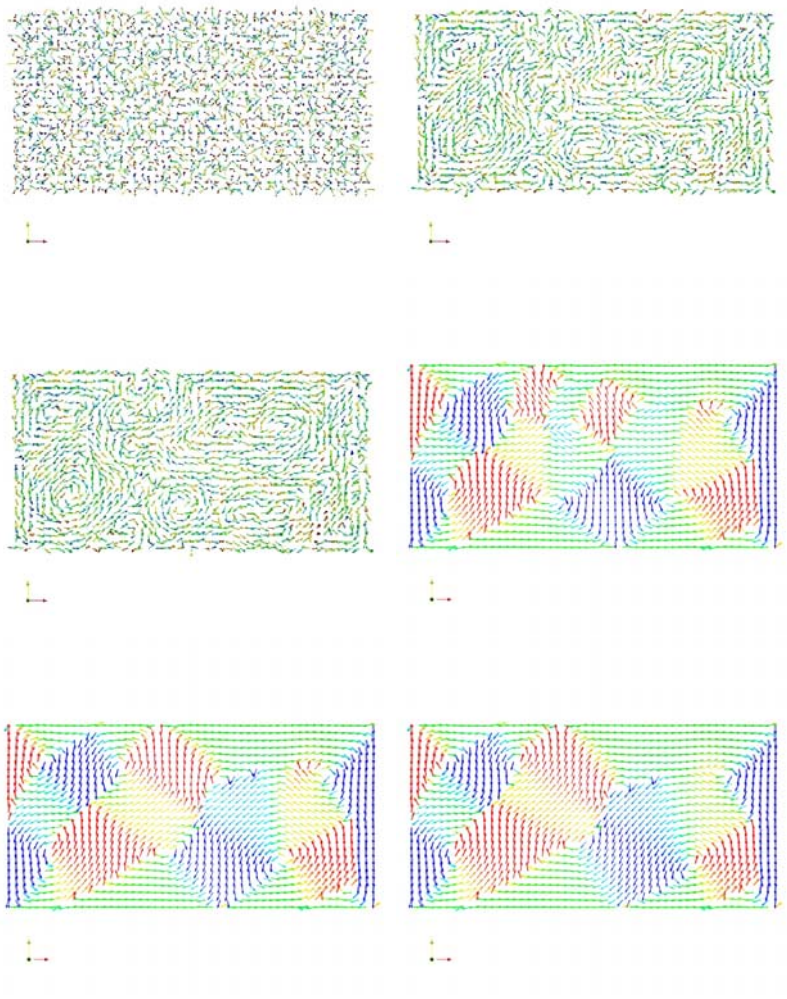


FIG. 8. *Example* 5.2: *Magnetization at times* $t = 0, 100, 200, 2500, 5000, 8000$ *(from left to right, from top to bottom).*

study the influence of the electromagnetic field $(\mathbf{e}^0, \mathbf{h}^0) : \Omega \to \mathbb{R}^3$ on the supercritical $\mathbf{m}^0 : \omega \to \mathbb{S}^2$.

Example 5.1. Let $(0,1)^3 = \omega = \Omega$ and $\mathbf{j} \equiv 0$. Let $\mathbf{m}^0 : \omega \to \mathbb{S}^2$ and $(\mathbf{e}^0, \mathbf{h}^0) : \Omega \to \left[\mathbb{R}^3\right]^2$ be defined by

$$\mathbf{m}^0(\mathbf{x}) = \begin{cases} (0,0,-1) & \text{for } |\mathbf{x}^*| \geq 1/2, \\ (2\mathbf{x}^* A, A^2 - |\mathbf{x}^*|^2)/(A^2 + |\mathbf{x}^*|^2) & \text{for } |\mathbf{x}^*| \leq 1/2 \text{ and } \mathbf{x}^* \in \omega, \end{cases}$$

$$\mathbf{h}^0_*(\mathbf{x}) = (0,0,H_s) \qquad \text{in } \Omega\,,$$

$$\mathbf{e}^0(\mathbf{x}) = (0,0,0) \qquad \text{in } \Omega,$$

with $\mathbf{x}^* = (x_1 - 0.5, x_2 - 0.5, 0)$ and $A = (1 - 2|\mathbf{x}^*|)^4/4$. The computational domain

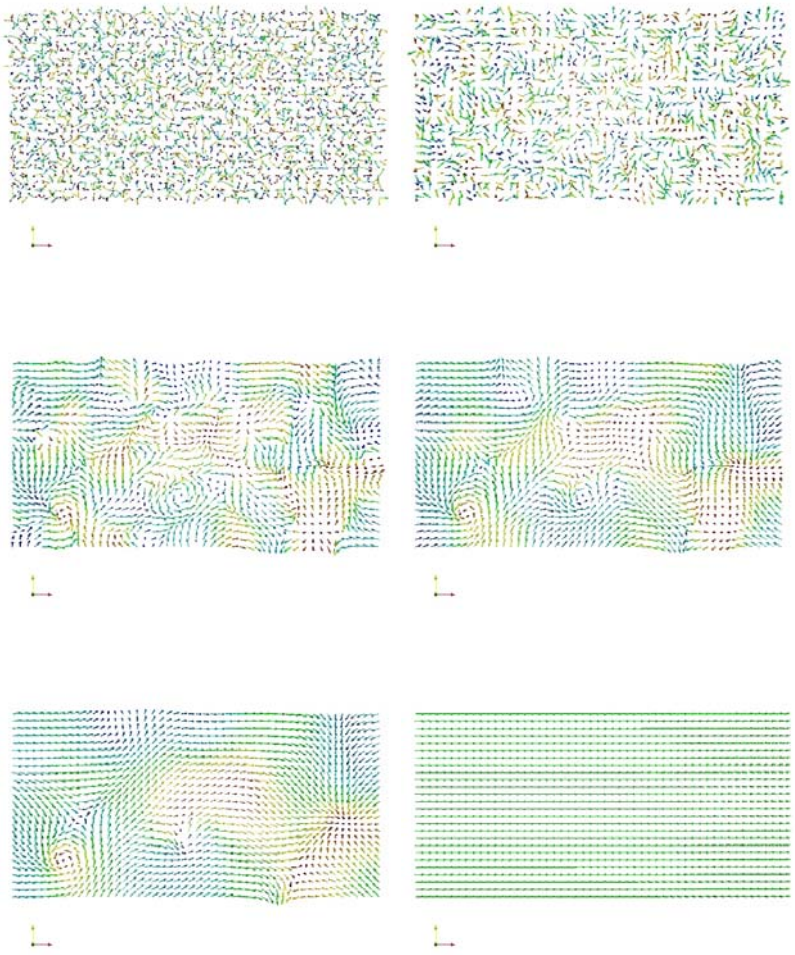


FIG. 9. *Example* 5.2: *Magnetization without the magnetic field at times* $t = 0, 10, 50, 100, 200, 2500$ *(from left to right, from top to bottom).*

$\Omega = \omega = (0,1)^3$ is partitioned into uniform cubes with side h, each cube consisting of six tetrahedra. We choose the time step $k = 10^{-5}$ and set the other parameters $\alpha = \gamma = M_s = \mu_0 = 1$, $\varepsilon_0 = 10^{-6}$, $K = 0$.

The constant H_s represents the strength of initial field in the x_3-direction. We computed the experiments for $H_s = -30, 0, 30$ on meshes with $h = 1/2^4$ and $h = 1/2^5$. The evolution of $\|\nabla \mathbf{m}_h\|_\infty$ is depicted in Figure 1. The evolution of the exchange energy can be found in Figure 2. Figure 3 displays the evolution of $\|\mathbf{h}_h\|_\infty / H_s$. We deduce from Figures 1 and 3 that variations of $\|\nabla \mathbf{m}_h\|_\infty$ close to a certain time $T^* > 0$ ("blow-up time") correspond to variations of $\|\mathbf{h}_h\|_\infty$ close to the same time T^*. Similar behavior is found for the evolution of $\|\mathbf{e}_h\|_\infty$, which is depicted in Figure 4. Further, we observe that the mesh-size not only determines the quantity $\max_t \|\nabla \mathbf{m}_h(t,\cdot)\|_\infty$ (cf. [4, 3]), but also influences the evolution of $\|\mathbf{h}_h\|_\infty$, $\|\mathbf{e}_h\|_\infty$. The computations with $H_s = -30$ show that a negative initial magnetic field accelerates the blow-up of

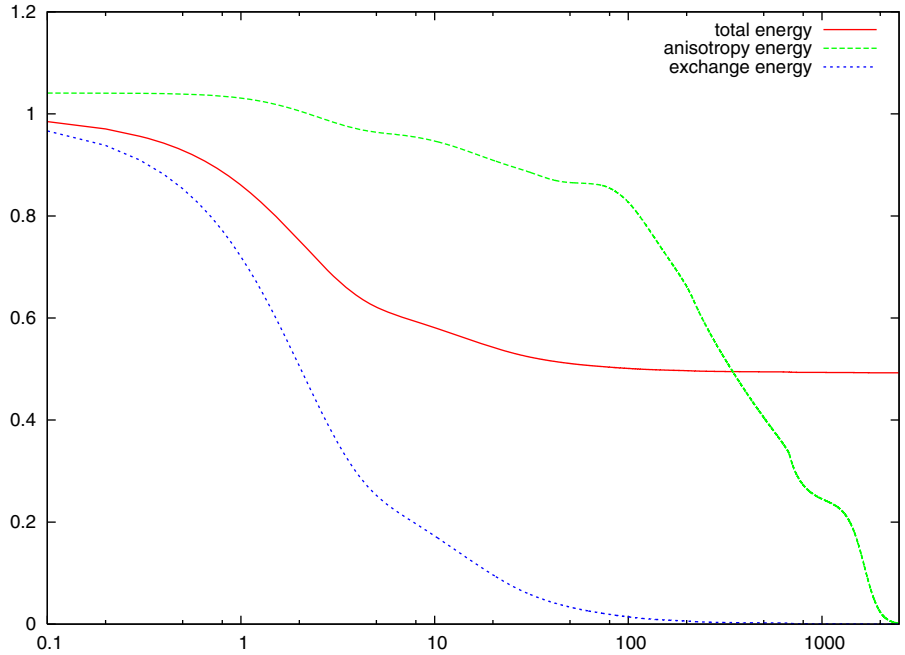


FIG. 10. *Example* 5.2: *Evolution of the energies,* $\log(t) \mapsto \mathcal{E}_T(t)/11$, $\mathcal{E}_{ex}(t)/11$, $\mathcal{E}_{an}(t)/0.04$.

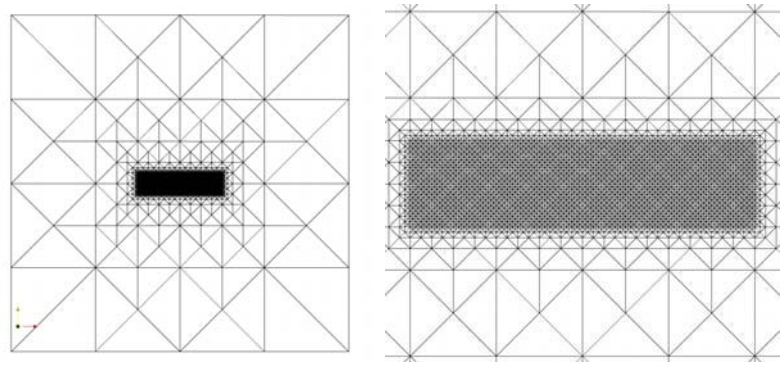


FIG. 11. *Example* 5.3: *Mesh for the domain* Ω *at* $x_3 = 0$ *(left) and zoom at the mesh for the domain* ω *at* $x_3 = 0$ *(right).*

the solution. The evolution becomes more complex for $H_s = 30$; the blow-up time is slightly delayed. Beyond the instability, when the magnetization is aligned along the $(0, 0, -1)$-direction, the influence of the magnetic field on the evolution prevails and the magnetization starts to rotate in the opposite direction, i.e., $(0, 0, 1)$; see Figure 5 for $h = 2^{-4}$. We did not compute beyond the time $t = 0.214$; however, it is reasonable to expect that the magnetization will be aligned in the $(0, 0, 1)$-direction after the steady state is reached. The details of the solution for $H_s = 30$, $h = 2^{-4}$ near $\mathbf{x} = (0.5, 0.5, 0.5)$ are depicted in Figure 6. There was only a little variation of the magnetization in the x_3-direction; we therefore present all results on the cross-cut through the domain at $x_3 = 0.5$.

For comparison, Figure 9 shows the evolution of magnetization with the same

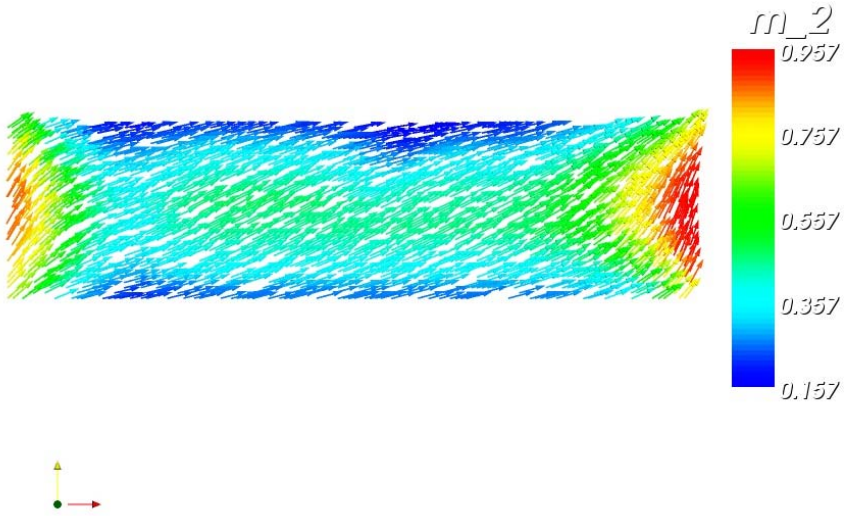


FIG. 12. *Example* 5.3: *Magnetization at time* $t = 300$*, near the S-state.*

initial condition computed without taking the coupling with Maxwell's equations into account. The evolution of the discrete energies $\mathcal{E}_T$, $\mathcal{E}_{ex}$, $\mathcal{E}_{an}$ is displayed in Figure 10 (note that $\mathcal{E}_H = \mathcal{E}_E \equiv 0$). The figures show a clear difference between the two cases.

The following example is derived from a benchmark problem [8, Problem #1] of a thin uniaxial ferromagnetic film, for which long-time dynamics is studied in [15], in the case $d = 2$. Here, we study the dynamics of the problem for $d = 3$. According to our knowledge, there are no comparable studies for $d = 3$ in the existing literature.

Example 5.2. Let $\omega = (0, 2) \times (0, 1) \times (0, 0.02)$ and $\Omega = (-0.2, 2.2) \times (-0.2, 1.2) \times (-0.04, 0.06)$ (the domain dimensions are in μm), with

$$\alpha = 0.5, \quad \gamma = 2.2 \times 10^9, \quad M_s = 8 \times 10^5, \quad K^* = 5 \times 10^2, \quad A^* = 1.3 \times 10^{-11},$$
$$\mathbf{p} = (1, 0, 0), \quad \varepsilon_0^* = 0.88422 \times 10^{-11}, \quad \mu_0^* = 1.25667 \times 10^{-6}, \quad \sigma = 0,$$

and ($\mathbf{j} \equiv \mathbf{0}$)

$$(\mathbf{e}^0, \mathbf{h}^0) = (\mathbf{0}, \mathbf{0}) \quad \text{in } \Omega.$$

The initial condition $\mathbf{m}_h^0$ is defined by assigning unit vectors with random orientation to every vortex of the mesh.

The domains Ω and ω are partitioned into bricks of dimension $0.04 \times 0.04 \times 0.02$. Subsequently, the corresponding computational meshes are obtained by subdivision of each brick into six tetrahedra. This partition results in 7956 DOFs for $\mathbf{m}_h$ on ω and

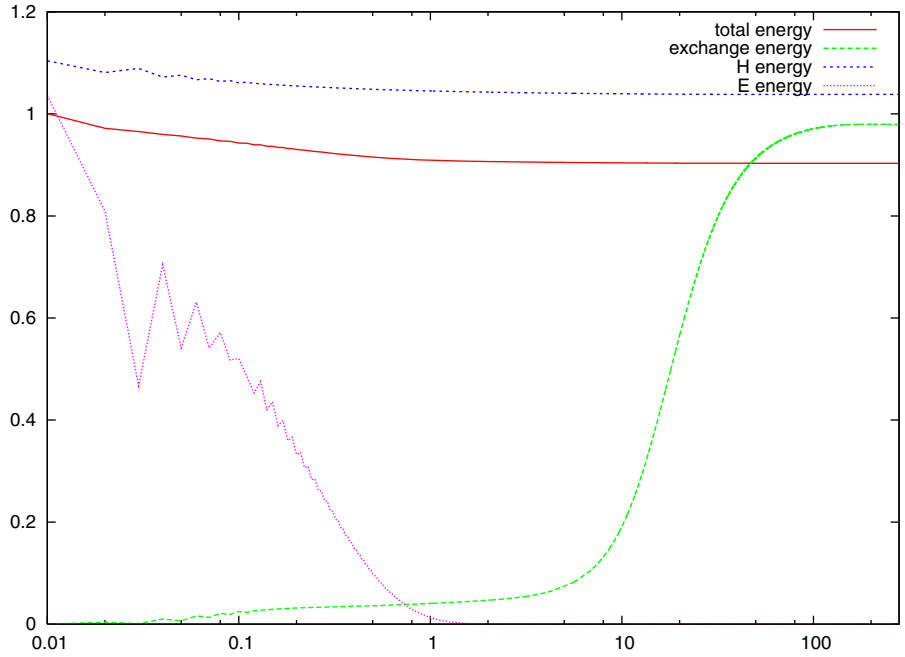


FIG. 13. *Example* 5.3: *Evolution of the energies,* $\log(t) \mapsto \mathcal{E}_T(t)/2.3$, $\mathcal{E}_{ex}(t)/0.0004$, $\mathcal{E}_H(t)/2$, $\mathcal{E}_E(t)/0.09$.

189000, 81325 DOFs for $\mathbf{e}_h$, $\mathbf{h}_h$ on Ω, respectively. We compute on a time interval $(0, 8000)$ using a uniform time step $k = 0.1$.

The evolution of the discrete energies $\mathcal{E}_T$, $\mathcal{E}_{ex}$, $\mathcal{E}_{an}$, $\mathcal{E}_H$, $\mathcal{E}_E$ (i.e., the total, exchange, anisotropy, magnetic field energy, and electric field energy) from Lemma 5.1 is depicted in Figure 7. Snapshots of the magnetization at different time levels can be found in Figure 8; the vectors are colored according to the value of the x_2-component of the magnetization vector.

The data for the last example are taken from [8, Problem #4]. Here we study the evolution of the magnetization toward a steady state, the so-called S-state; cf. [8, Problem #4].

Example 5.3. Let $\omega = (0, 0.5) \times (0, 0.125) \times (0, 0.003)$ and $\Omega = (-0.75, 1.25) \times (-0.9375, 1.0625) \times (-0.7665, 0.7695)$ (in μm), with

$$\alpha = 1, \quad \gamma = 2.2 \times 10^9, \quad M_s = 8 \times 10^5, \quad K^* = 0, \quad A^* = 1.3 \times 10^{-11},$$
$$\varepsilon_0^* = 0.88422 \times 10^{-11}, \quad \mu_0^* = 1.25667 \times 10^{-6}, \quad \sigma = 0,$$

and ($\mathbf{j} \equiv \mathbf{0}$)

$$\mathbf{m}^0 = (\,1, 0, 0\,) \quad \text{in } \omega\,,$$
$$\mathbf{h}_*^0 = (0.01, 0.01, 0.01), \quad \mathbf{e}^0 = \mathbf{0} \quad \text{in } \Omega\,.$$

The domain ω is partitioned uniformly into cubes of dimensions $0.00390625 \times 0.00390625 \times 0.003$, where each cube consists of six tetrahedra. The nonuniform mesh for the domain Ω is constructed in such a way that it is identical to the mesh for ω in the region $\Omega \cap \omega$ and the mesh-size gradually increases away from the overlapping region. A cross-cut through the mesh at $x_3 = 0$ is displayed in Figure 11.

The above discretization of the computational domains results in $25,542$ DOFs for $\mathbf{m}_h$ on ω and $138,302$, $353,568$ DOFs for $\mathbf{e}_h$, $\mathbf{h}_h$, respectively. We employ a uniform time stepping, with $k = 0.01$.

The magnetization at time $T = 300$ is displayed in Figure 12; the vectors are colored according to the x_2-component of the magnetization. Convergence toward the S-state (cf. [8, Problem #4]) can be observed from the results. No steady state has yet been reached at the final time; however, the convergence toward the steady state after the time $T = 300$ has been very slow. The evolution of the discrete energies from Lemma 5.1 is depicted in Figure 13.

6. Concluding remarks. We devised an implicit discretization of the MLLG system which is based on linear finite elements. For pairs of time-step sizes and mesh-sizes tending simultaneously but independently to zero, we showed that every accumulation point of the sequence of numerical approximations is a weak solution of the continuous equations satisfying an energy inequality. An iterative solver for the solution of the nonlinear system of equations in each time step is proposed, and its convergence is proved under the time-step restriction $k \leq Ch^2$. The convergence of the nonlinear solver is robust for small values of the damping parameters α. Owing to expected nonuniqueness and possible occurrence of singularities, strong convergence of the (whole) sequence or error estimates cannot be expected unless additional assumptions on an exact solution are made. The use of higher order finite elements for the discretization of $\mathbf{m}$ are beyond the scope of this paper and remain to be analyzed in future work. Our numerical experiments confirm the theoretical results and indicate that finite-time blow-up for MLLG is possible and hence that singular solutions can develop from smooth initial data. Two examples based on standard micromagnetic benchmark problems are computed to demonstrate the potential of the method for practical applications. Further, we constructed a Schur complement preconditioner for a saddle point system of algebraic equations which arise in every iteration of the nonlinear solver. We observed that the preconditioning dramatically reduced the number of iterations needed for the convergence of the linear solver on structured and unstructured meshes. A multigrid-type preconditioner, which is robust with respect to the time-step size, is subject to our current research.

REFERENCES

[1] F. Alouges and P. Jaisson, *Convergence of a finite elements discretization for the Landau Lifshitz equations*, Math. Models Methods Appl. Sci., 16 (2006), pp. 299–316.

[2] F. Alouges and A. Soyeur, *On global weak solutions for Landau–Lifshitz equations: Existence and nonuniqueness*, Nonlinear Anal., 18 (1992), pp. 1071–1084.

[3] S. Bartels, J. Ko, and A. Prohl, *Numerical approximation of Landau–Lifshitz equations and finite time blow up of weak solutions*, Math. Comp., 76 (2007), pp. 1847–1859.

[4] S. Bartels and A. Prohl, *Convergence of an implicit finite element method for the Landau–Lifshitz–Gilbert equation*, SIAM J. Numer. Anal., 44 (2006), pp. 1405–1419.

[5] W. Brown, *Micromagnetics*, Tracts of Physics, Wiley Interscience, New York, 1963.

[6] G. Carbou and P. Fabrie, *Time average in micromagnetism*, J. Differential Equations, 147 (1998), pp. 383–409.

[7] I. Cimrák, *Error estimates for a semi-implicit numerical scheme solving the Landau–Lifshitz equation with an exchange field*, IMA J. Numer. Anal., 25 (2005), pp. 611–634.

[8] CTCMS, *MMMG: Micromagnetic Modeling Activity Group*, http://www.ctcms.nist.gov/∼rdm/mumag.org.html.

[9] R. Hiptmair, *Multigrid method for Maxwell's equations*, SIAM J. Numer. Anal., 36 (1998), pp. 204–225.

[10] M. Kružík and A. Prohl, *Recent developments in the modeling, analysis, and numerics of ferromagnetism*, SIAM Rev., 48 (2006), pp. 439–483.

[11] L. D. Landau and E. M. Lifshitz, *On the theory of the dispersion of magnetic permeability in ferromagnetic bodies*, Phys. Z. Sowjetunion, 8 (1935), pp. 153–169.
[12] B. Lax and K. Button, *Microwave Ferrites and Ferrimagnetics*, McGraw-Hill, New York, 1962.
[13] P. B. Monk, *Finite Element Methods for Maxwell's Equations*, Oxford University Press, Oxford, UK, 2003.
[14] P. B. Monk and O. Vacus, *Error estimates for a numerical scheme for ferromagnetic problems*, SIAM J. Numer. Anal., 36 (1999), pp. 696–718.
[15] P. B. Monk and O. Vacus, *Accurate discretization of a non-linear micromagnetic problem*, Comput. Methods Appl. Mech. Engrg., 190 (2001), pp. 5243–5269.
[16] J. Nédélec, *A new family of mixed finite elements in* $\mathbb{R}^3$, Numer. Math., 50 (1986), pp. 57–81.
[17] J. Peters, V. Reichelt, and A. Reusken, *Fast iterative solvers for discrete Stokes equations*, SIAM J. Sci. Comput., 27 (2005), pp. 646–666.
[18] A. Prohl, *Computational Micromagnetism*, Teubner, Stuttgart, 2001.
[19] A. Schmidt and K. G. Siebert, *ALBERT—software for scientific computations and applications*, Acta Math. Univ. Comenian. (N.S.), 70 (2000), pp. 105–122.
[20] C. Serpico, I. D. Mayergoyz, and G. Bertotti, *Numerical technique for integration of the Landau–Lifshitz equation*, J. Appl. Phys., 89 (2001), pp. 6991–6993.
[21] M. Slodička and Ľ. Baňas, *A numerical scheme for Maxwell–Landau–Lifshitz–Gilbert system*, Appl. Math. Comput., 158 (2004), pp. 79–99.
[22] J. Sun, F. Collino, P. B. Monk, and L. Wang, *An eddy-current and micromagnetism model with applications to disk write heads*, Internat. J. Numer. Methods Engrg., 60 (2004), pp. 1673–1698.
[23] A. Visintin, *On Landau–Lifshitz' equations for ferromagnetism*, Japan J. Appl. Math., 2 (1985), pp. 69–84.

SIAM J. NUMER. ANAL.
Vol. 46, No. 3, pp. 1423–1441
© 2008 Society for Industrial and Applied Mathematics

NONDEGENERACY AND WEAK GLOBAL CONVERGENCE OF THE LLOYD ALGORITHM IN $\mathbb{R}^{D*}$

MARIA EMELIANENKO†, LILI JU‡, AND ALEXANDER RAND§

Abstract. The Lloyd algorithm originated in the context of optimal quantization and represents a fixed point iteration for computing an optimal quantizer. Reducing average distortion at every step, it constructs a Voronoi partition of the domain and replaces each generator with the centroid of the corresponding Voronoi cell. Optimal quantization is obtained in the case of a centroidal Voronoi tessellation (CVT), which is a special Voronoi tessellation of a domain $\Omega \in \mathbb{R}^d$ having the property that the generators of the Voronoi diagram are also the centers of mass, with respect to a given density function $\rho \geq 0$, of the corresponding Voronoi cells. The Lloyd iteration is currently the most popular and elegant algorithm for computing CVTs and optimal quantizers, but many questions remain about its convergence, especially in d-dimensional spaces ($d > 1$). In this paper, we prove that any limit point of the Lloyd iteration in any dimensional spaces is nondegenerate provided that Ω is a convex and bounded set and ρ belongs to $L^1(\Omega)$ and is positive almost everywhere. This ensures that the fixed point map remains closed and hence the standard theory of descent methods guarantees weak global convergence of the Lloyd iteration to the set of nondegenerate fixed point quantizers. While previously only conjectured, the convergence properties of the Lloyd iteration are rigorously justified under such minimal regularity assumptions on the density functional. The results presented in this paper go beyond existing convergence theories for CVTs and optimal quantization related algorithms and should be of interest to both the mathematical and engineering communities.

Key words. centroidal Voronoi tessellations, quantization, Lloyd algorithm, global convergence

AMS subject classifications. 65D99, 94A15, 68U05

DOI. 10.1137/070691334

1. Introduction.

Given a bounded domain $\Omega \in \mathbb{R}^d$ and a set of distinct points $\mathbf{Z} = \{\mathbf{z}_i\}_{i=1}^n \subset \Omega$, for each point $\mathbf{z}_i$, $i = 1, \ldots, n$, define the corresponding Voronoi region $V_i(\mathbf{Z})$, $i = 1, \ldots, n$, by

$$V_i(\mathbf{Z}) = \big\{\mathbf{z} \in \Omega \;\mid\; |\mathbf{z} - \mathbf{z}_i| < |\mathbf{z} - \mathbf{z}_j| \quad \text{for } j = 1, \ldots, n \quad \text{and} \quad j \neq i\big\},$$

where $|\cdot|$ denotes the Euclidean norm in $\mathbb{R}^d$. Clearly $V_i(\mathbf{Z}) \cap V_j(\mathbf{Z}) = \emptyset$ for $i \neq j$, and $\cup_{i=1}^n \overline{V}_i(\mathbf{Z}) = \overline{\Omega}$ so that $\{V_i(\mathbf{Z})\}_{i=1}^n$ is a tessellation of Ω. We refer to $\mathbf{V}(\mathbf{Z}) = \{V_i(\mathbf{Z})\}_{i=1}^n$ as the *Voronoi tessellation* of Ω associated with the point set $\mathbf{Z}$. A point $\mathbf{z}_i$ is called a *generator*; a subdomain $V_i(\mathbf{Z}) \subset \Omega$ is referred to as the *Voronoi region* corresponding to the generator $\mathbf{z}_i$. It is well known that the dual tessellation (in a graph-theoretical sense) to a Voronoi tessellation of Ω is the so-called *Delaunay triangulation* (DT). All Voronoi regions $V_i(\mathbf{Z})$'s are convex (especially convex polygons/polyhedra, except for their part on the boundary) if Ω is convex.

*Received by the editors May 14, 2007; accepted for publication (in revised form) December 7, 2007; published electronically March 26, 2008.
http://www.siam.org/journals/sinum/46-3/69133.html

†Department of Mathematical Sciences, George Mason University, Fairfax, VA 22030 (memelian@gmu.edu). The work of this author was partially supported by the Center for Nonlinear Analysis at Carnegie Mellon University under the National Science Foundation grant 0405343.

‡Department of Mathematics, University of South Carolina, Columbia, SC 29208 (ju@math.sc.edu). This author's research was supported in part by the National Science Foundation under grant DMS-0609575 and the Department of Energy under grant DE-FG02-07ER64431.

§Department of Mathematical Sciences, Carnegie Mellon University, Pittsburgh, PA 15213 (arand@andrew.cmu.edu).

Given a density function $\rho(\mathbf{z}) \geq 0$ defined on Ω, for any region $V \subset \Omega$, define $\mathbf{c}_V$, the *mass center* or *centroid* of V, by

$$\mathbf{c}_V = \frac{\int_V \mathbf{y}\rho(\mathbf{y})\,\mathrm{d}\mathbf{y}}{\int_V \rho(\mathbf{y})\,\mathrm{d}\mathbf{y}}.$$

Then a special family of Voronoi tessellations can be defined in the following.

DEFINITION 1.1 (see [6]). *We refer to a Voronoi tessellation* $(\mathbf{Z}, \mathbf{V}(\mathbf{Z}))$ *of* Ω *as a* centroidal Voronoi tessellation *(CVT) if and only if the points* $\mathbf{Z} = \{\mathbf{z}_i\}_{i=1}^n$ *which serve as the generators of the associated Voronoi regions* $\mathbf{V}(\mathbf{Z}) = \{V_i(\mathbf{Z})\}_{i=1}^n$ *are also the centroids of those regions, i.e., if and only if we have that*

$$\mathbf{z}_i = \mathbf{c}_{V_i(\mathbf{Z})} \quad \textit{for} \;\; i = 1, \ldots, n\,.$$

General Voronoi tessellations do not satisfy the CVT property. It is worth noting that for a given domain and density, at least one CVT exists [1] but may not be unique [6]. The CVT concept can be generalized to very broad settings that range from abstract spaces and distance metrics to discrete point sets [6]. In the latter setting, it can be recognized as being closely related to the k-means and h-means algorithms in clustering and quantization applications [10]. Its extension to general surfaces and manifolds has also been studied in [7, 8]. CVTs are very useful in many applications, including but not limited to image and data analysis, vector quantization, resource optimization, design of experiments, optimal placement of sensors and actuators, cell biology, territorial behavior of animals, numerical partial differential equations, point sampling, meshless computing, mesh generation and optimization, reduced-order modeling, computer graphics, and mobile sensing networks.

CVTs possess an optimization property that can be used as a basis for an alternate definition. Given any set of points $\mathbf{Z} = \{\mathbf{z}_i\}_{i=1}^n$ in Ω and any tessellation $\mathbf{V} = \{V_i\}_{i=1}^n$ of Ω, define the *energy* by

$$\mathcal{E}(\mathbf{Z}, \mathbf{V}) = \sum_{i=1}^{n} \int_{V_i} \rho(\mathbf{y})|\mathbf{y} - \mathbf{z}_i|^2\,\mathrm{d}\mathbf{y}. \tag{1.1}$$

Then it can be shown that $\mathcal{E}$ is minimized only if $(\mathbf{Z}, \mathbf{V})$ forms a CVT of Ω [6]. Notice that $(\mathbf{Z}, \mathbf{V})$ still may not be a minimizer of $\mathcal{E}$ although $(\mathbf{Z}, \mathbf{V})$ is a CVT. For some properties of the minimizers of the energy functional $\mathcal{E}$ (i.e., CVTs or optimal quantizations), see discussions in [6, 2, 21]. Although the energy $\mathcal{E}$ may not be directly identified with the energy of some physical system, it is often naturally associated with quantities such as *error distortion*, *variance*, and *cost* in many applications.

Currently the most popular and effective algorithm for construction of CVTs is the so-called Lloyd algorithm [18] that is a simple iteration between constructing Voronoi diagrams and mass centroids.

ALGORITHM 1.1 (Lloyd algorithm). *We are given a domain* Ω*, a density function* $\rho(\mathbf{x})$ *defined on* Ω*, and a positive integer* n*.*

0. *Select an initial set of* n *distinct points* $\mathbf{Z} = \{\mathbf{z}_i\}_{i=1}^n$ *on* Ω*.*
1. *Construct the Voronoi regions* $\mathbf{V}(\mathbf{Z}) = \{V_i(\mathbf{Z})\}_{i=1}^n$ *of* Ω *associated with* $\mathbf{Z}$*.*
2. *Determine the centroids of the Voronoi regions* $\mathbf{V}(\mathbf{Z})$*; these centroids form the new set of points* $\mathbf{Z}$*.*

3. *If the new points meet some convergence criterion, return* $(\mathbf{Z}, \mathbf{V}(\mathbf{Z}))$ *and terminate; otherwise, go to step* 1.

The Lloyd algorithm also coincides with the simplest version of the k-means algorithm in clustering and quantization fields [10, 12, 18, 15]. A commonly used stopping criterion is to test whether $\|\mathbf{Z}^k - \mathbf{Z}^{k-1}\| \leq \epsilon$ holds for some predefined tolerance ϵ at some iteration k. It is worth noting that the energy $\mathcal{E}(\mathbf{Z}, \mathbf{V}(\mathbf{Z}))$ decreases monotonically along the Lloyd iteration. Some acceleration schemes based on multigrid techniques are proposed in [4]. A probabilistic version of the Lloyd method for computing CVTs and its parallel implementation have also been studied in [14]. One famous generalization in the vector quantization setting is the Linde–Buzo–Gray algorithm [17], widely used in statistics and data mining applications.

Since Lloyd's pioneering work, many studies have been done on convergence properties of the Lloyd iteration [6, 5, 13, 16, 19]. Despite the algorithm's long history and wide popularity in many applications, its convergence theory is far from being complete. Many researchers addressed the question of existence and uniqueness of an optimal quantizer. Most notably, in [9] and later in [24] and [16], it was shown that the Lloyd method is a local contraction (thus *locally* convergent) in the one-dimensional space when the density function is logarithmically concave. This result was reiterated in [6] with a much simpler proof. In [25], this result was extended to all continuous and positive densities in the one-dimensional setting of scalar quantization. A different approach was used in [5], where the fixed point map was shown to be closed and a weak global convergence result was presented for all continuous and positive density functionals along with some numerical studies and precise theoretic estimates of the local convergence rate.

In higher dimensions, convergence results are much more limited. For discrete distributions, the algorithm is shown to converge and, moreover, reach a CVT in a finite number of steps [23, 11]. In the case of continuous distributions, [11, 22] show convergence of the energy functional by defining the value of the Lloyd map on degenerate points. In order to ensure that the algorithm is closed, this requires that degenerate points be *defined* as fixed points in the definition of the algorithm. So, while standard theory of descent methods ensures that this extended Lloyd algorithm will converge to a fixed point, these limit points could be degenerate. Similarly, the multidimensional global convergence of a subsequence argument presented in [5] implicitly assumed nondegeneracy of the Lloyd map and hence suffered from a similar deficiency. We seek to apply the theory of descent methods to the Lloyd algorithm without any extension to degenerate points. This makes sense as it will be shown that points of degeneracy locally maximize the energy in some sense. Moreover, this is necessary in order to ensure that limit points are in fact CVTs.

In this paper we prove that for an arbitrary, positive density function under a mild L^1 regularity assumption, the fixed point mapping remains closed, which guarantees weak global convergence of the Lloyd iteration. Although we rely on the standard global convergence theory for descent algorithms as discussed in previous investigations (see, e.g., [22, 24]), we take a more direct approach of rigorously justifying the nondegeneracy of the limiting and intermediate iteration points, an approach that much simplifies the analysis of the method and that addresses a question that escaped rigorous treatment until now. These results ensure that Lloyd iterates do not approach degenerate points and that any limit point of the algorithm is a CVT with n distinct generators.

The rest of the paper is organized as follows. In section 2, we introduce some

necessary notation and give an overview of some relevant results and properties. In subsection 3.1, we derive precise estimates on the distance from a centroid to the boundary of the domain and obtain an upper bound on the difference of two Voronoi regions. In subsection 3.2 we present the proof of nondegeneracy of the limit points of the Lloyd iteration, followed in section 3.3 by the main result of the weak global convergence of Lloyd algorithm. Conclusions and some open questions are given in section 4.

2. Preliminaries. For any set $V \subset \Omega$, let us denote by $m(V)$ the measure of V, i.e.,

$$m(V) := \int_V \mathrm{d}\mathbf{y},$$

and by $M(V)$ the mass of V, i.e.,

$$M(V) := \int_V \rho(\mathbf{y})\, \mathrm{d}\mathbf{y}.$$

The fact that ρ is positive almost everywhere means that $M(V) > 0$ if $m(V) > 0$. We also define $\hat{\Omega}^n$ as follows:

$$\hat{\Omega}^n := \{\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n) \in \overline{\Omega}^n \mid \mathbf{z}_i \neq \mathbf{z}_j \ \ \forall\, i \neq j\}.$$

The so-called Lloyd map is the function $\mathbf{T} : \hat{\Omega}^n \to \hat{\Omega}^n$ that takes a tuple of generators into the tuple of the centroids of its Voronoi regions (exactly the iteration used in the Lloyd algorithm). More precisely, let $\mathbf{T} = (\mathbf{T}_1, \ldots, \mathbf{T}_n)$; then, for any $\mathbf{Z} \in \hat{\Omega}^n$,

$$\mathbf{T}_i(\mathbf{Z}) = \frac{\displaystyle\int_{V_i(\mathbf{Z})} \mathbf{y}\rho(\mathbf{y})\, \mathrm{d}\mathbf{y}}{\displaystyle\int_{V_i(\mathbf{Z})} \rho(\mathbf{y})\, \mathrm{d}\mathbf{y}}$$

for $i = 1, \ldots, n$. The Lloyd map is continuous from $\hat{\Omega}^n$ to $\hat{\Omega}^n$ [6, 5] but not defined on $\bar{\Omega}^n \setminus \hat{\Omega}^n$. The Lloyd algorithm is then the process generating a sequence of point sets:

$$\mathbf{Z}, \mathbf{T}(\mathbf{Z}), \mathbf{T}^2(\mathbf{Z}), \ldots, \mathbf{T}^k(\mathbf{Z}), \ldots .$$

Clearly, the resulting Lloyd iterates are bounded.

For any $\mathbf{Z} \in \hat{\Omega}^n$, the *quantization energy* is defined as

$$\mathcal{G}(\mathbf{Z}) = \mathcal{E}(\mathbf{Z}, \mathbf{V}(\mathbf{Z})) = \sum_{i=1}^{n} \int_{V_i(\mathbf{Z})} \rho(\mathbf{y})|\mathbf{y} - \mathbf{z}_i|^2\, \mathrm{d}\mathbf{y}.$$

The quantization energy function is also continuous on $\hat{\Omega}^n$. Furthermore, this energy can be extended continuously to $\overline{\Omega}^n$. When $\mathbf{Z} \notin \hat{\Omega}^n$, the value of $\mathcal{G}$ is given by the same formula except we consider $\mathbf{Z}$ without duplicate components. For example, in the case of $n = 2$, if $\mathbf{Z} = (\mathbf{z}_1, \mathbf{z}_2)$ and $\mathbf{z}_1 \neq \mathbf{z}_2$, then

$$\mathcal{G}(\mathbf{Z}) = \int_{V_1(\mathbf{Z})} \rho(\mathbf{y})|\mathbf{y} - \mathbf{z}_1|^2\, \mathrm{d}\mathbf{y} + \int_{V_2(\mathbf{Z})} \rho(\mathbf{y})|\mathbf{y} - \mathbf{z}_2|^2\, \mathrm{d}\mathbf{y}.$$

In the degenerate case, we have $\mathbf{Z} = (\mathbf{z}_1, \mathbf{z}_1)$, so that

$$\mathcal{G}(\mathbf{Z}) = \mathcal{G}((\mathbf{z}_1)) = \int_\Omega \rho(\mathbf{y})|\mathbf{y} - \mathbf{z}_1|^2 \, \mathrm{d}\mathbf{y}.$$

The derivative of $\mathcal{G}$ in $\hat{\Omega}^n$ is given [6] by

$$\frac{\partial \mathcal{G}}{\partial \mathbf{z}_i}(\mathbf{Z}) = 2M(V_i(\mathbf{Z}))(\mathbf{Z}_i - \mathbf{T}_i(\mathbf{Z})), \tag{2.1}$$

and thus the set of critical points of $\mathcal{G}$ in $\hat{\Omega}^n$ is exactly the set of fixed points of the Lloyd map and thus is the set of CVTs.

For convenience of discussion, a more general energy $\mathcal{H}$ is defined in [5] as follows: for any $(\mathbf{Y}, \mathbf{Z}) \in (\hat{\Omega}^n, \hat{\Omega}^n)$,

$$\mathcal{H}(\mathbf{Y}, \mathbf{Z}) = \mathcal{E}(\mathbf{Z}, \mathbf{V}(\mathbf{Y})) = \sum_{i=1}^{n} \int_{V_i(\mathbf{Y})} \rho(\mathbf{y})|\mathbf{y} - \mathbf{z}_i|^2 \, \mathrm{d}\mathbf{y}.$$

Notice that $\mathcal{H}$ cannot be continuously extended to $(\overline{\Omega}^n, \overline{\Omega}^n)$ like the function $\mathcal{G}$. The Lloyd algorithm is in fact a dual minimization process [5]: for any $\mathbf{Z} \in \hat{\Omega}^n$,

$$\mathcal{H}(\mathbf{Z}, \mathbf{T}(\mathbf{Z})) = \min_{\mathbf{Y} \in \hat{\Omega}^n} \mathcal{H}(\mathbf{Z}, \mathbf{Y}),$$

$$\mathcal{H}(\mathbf{Z}, \mathbf{Z}) = \min_{\mathbf{Y} \in \hat{\Omega}^n} \mathcal{H}(\mathbf{Y}, \mathbf{Z}).$$

As long as ρ is positive almost everywhere, the above minimization problems on the right-hand side both have a unique solution. Thus we have that

$$\mathcal{G}(\mathbf{T}(\mathbf{Z})) = \mathcal{H}(\mathbf{T}(\mathbf{Z}), \mathbf{T}(\mathbf{Z})) \le \mathcal{H}(\mathbf{Z}, \mathbf{T}(\mathbf{Z})) \le \mathcal{H}(\mathbf{Z}, \mathbf{Z}) = \mathcal{G}(\mathbf{Z}). \tag{2.2}$$

Uniqueness of the centroid implied by the positive density in a convex set ensures that if $(\mathbf{Z}, \mathbf{V}(\mathbf{Z}))$ is not a CVT, then $\mathcal{G}(\mathbf{T}(\mathbf{Z})) < \mathcal{G}(\mathbf{Z})$.

The following definition of weak convergence will be used throughout the paper.

DEFINITION 2.1. *Denote by $\sigma(x, S)$ a distance between a point $x \in \Omega$ and a set S and by Γ the set of fixed points of $\mathbf{T}$. We will say that the sequence of Lloyd iterates $\{\mathbf{Z}^i\}_{i=0}^{\infty}$ is* weakly convergent *if $\lim_{i \to +\infty} \nabla \mathcal{G}(\mathbf{Z}^i) = \mathbf{0}$ and $\lim_{i \to +\infty} \sigma(\mathbf{Z}^i, \Gamma) = 0$.*

In other words, the sequence of Lloyd iterates converges in a weak sense if the energy values asymptotically approach some limit and the sequence of Lloyd iteration points approaches the set of fixed points of $\mathbf{T}$.

In order to analyze *global* convergence of the Lloyd algorithm in general dimensional spaces, we recall the following result.

THEOREM 2.1 (see [5, 20]). *If the iterates in the Lloyd algorithm stay in a compact set where the Lloyd map $\mathbf{T}$ is continuous, then the algorithm is weakly globally convergent, i.e., $\lim_{i \to +\infty} \nabla \mathcal{G}(\mathbf{Z}^i) = \mathbf{0}$ for the Lloyd iterates $\{\mathbf{Z}^i\}_{i=0}^{\infty}$ starting from any initial guess (in other words, any limit point of $\{\mathbf{Z}^i\}_{i=0}^{\infty}$ is a critical point of $\mathcal{G}$).*

Theorem 2.1 was proved in [5] using the properties of $\mathcal{H}$ and the (weak) global convergence theorem stated in [20]. It was also shown in [5] and [25] that the Lloyd algorithm has to converge to the set of critical points of the energy in one-dimensional space provided that ρ is strictly positive and smooth, but for higher dimensional cases, this question is still open, as discussed earlier in the introduction.

It has also never been rigorously justified that the set of limit points cannot contain more than one point, except for the case of log-concave density in the one-dimensional case, restricting the convergence in the aforementioned result to the "weak" sense as given in Definition 2.1. Although we conjecture that a stronger convergence result can be obtained in most practical cases, this question requires a separate rigorous treatment, while at present we focus our attention on extending Theorem 2.1 to higher dimensions.

More precisely, the goal of this paper is to present a proof for the weak global convergence of the Lloyd algorithm in any dimensional space. We do this by showing that limit points of the Lloyd method in $\mathbb{R}^d$ are nondegenerate, and hence the algorithm converges in the sense of Definition 2.1 for any L^1 density functions in any dimension.

3. Weak global convergence of Lloyd iteration. In the following part of this paper, we assume that the following conditions hold.

Assumption 3.1. (1) The domain $\Omega \in \mathbb{R}^d$ is a convex and bounded set with the diameter

$$\operatorname{diam}(\Omega) := \sup_{\mathbf{z},\mathbf{y}\in\Omega} |\mathbf{z}-\mathbf{y}| = R_\Omega < +\infty.$$

(2) The density function ρ belongs to $L^1(\Omega)$ and is positive almost everywhere. Consequently, we have that

$$0 < M(\Omega) = \|\rho\|_{L^1(\Omega)} = \int_\Omega \rho(\mathbf{y})\mathrm{d}\mathbf{y} < +\infty.$$

For the proof of weak global convergence of the Lloyd algorithm, we will show that any limit point of the algorithm is nondegenerate (in $\hat{\Omega}^n$) and thus the iterates stay inside a compact set in $\hat{\Omega}^n$. The weak global convergence of the algorithm will then follow from Theorem 2.1. This ensures that any limit point of the algorithm does in fact correspond to a CVT. Before these results can be shown, a number of technical intermediate lemmas must be shown. The goal of these lemmas is to estimate how the quantization energy changes between iterations of the Lloyd map when iterates approach the boundary of $\hat{\Omega}^n$. As iterates approach the boundary of $\hat{\Omega}^n$, some generator approaches the boundary of its Voronoi region. As all Voronoi regions are convex sets, it becomes important to bound the distance between the boundary and the centroid of a generic convex set in Ω with a prescribed mass.

3.1. Some lemmas. First, let us present a classic result from measure theory.

LEMMA 3.1. *Let Assumption* 3.1 *be satisfied. Then for any $\epsilon > 0$, there exists a constant $\eta > 0$ such that*

$$M(V) = \int_V \rho(\mathbf{y})\ \mathrm{d}\mathbf{y} \le \epsilon \tag{3.1}$$

for any set $V \subset \Omega$ with $m(V) \le \eta$.

The proof of this lemma is given in the appendix. For any $\epsilon \ge 0$, define

$$E_\epsilon = \{\eta \in [0, m(\Omega)] \ \mid\ M(V) \le \epsilon \ \text{ if } \ V \subset \Omega \text{ with } m(V) \le \eta\},$$

and Lemma 3.1 tells us that E_ϵ is a nonempty set, so that

$$\eta_\epsilon := \sup_{\eta\in E_\epsilon} \eta \tag{3.2}$$

is well defined and satisfies that $\eta_{\epsilon_1} \le \eta_{\epsilon_2}$ if $\epsilon_1 \le \epsilon_2$. It is also obvious that $\eta_\epsilon > 0$ if $\epsilon > 0$. For a given $\epsilon \ge 0$, η_ϵ is often difficult to compute, but here we assert only its existence.

Now we estimate the distance from the boundary to the centroid of a general convex set. The important feature of the estimate below is that the lower bound on the distance between the centroid and the boundary of a general convex set depends only on the set through the mass of the set.

For any $V \in \Omega$, let $R_V := \mathrm{diam}(V)$ denote the diameter of V (clearly $R_V \le R_\Omega$), and define the V-dependent constant δ_V by

$$\delta_V := \frac{\eta^2_{M(V)/4}}{64 R_V^{2d-1}},$$

where $\eta_{M(V)/4}$ is the constant defined in (3.2) with $\epsilon = \frac{M(V)}{4}$. Then the following result holds.

LEMMA 3.2. *Let Assumption* 3.1 *be satisfied. Then for any convex set $V \subset \Omega$, it holds that*

$$\mathrm{dist}(\mathbf{c}_V, \partial V) \ge \delta_V, \tag{3.3}$$

where $\mathrm{dist}(\mathbf{c}_V, \partial V)$ *denotes the shortest distance from the centroid of V denoted by $\mathbf{c}_V$ to the boundary of V.*

Proof. Without loss of generality, we assume that the origin is the nearest boundary point to the centroid $\mathbf{c}_V$, and we let $\mathbf{c}_V$ have the coordinate $(\mathbf{0}, x)$ for some $x > 0$, where $\mathbf{0}$ denotes the zero vector in $\mathbb{R}^{d-1}$. Since V is convex and $\sup_{\mathbf{z},\mathbf{y}\in\Omega} |\mathbf{z}-\mathbf{y}| = R_V$, we must have $R_V \ge y_d \ge 0$ for any $\mathbf{y} = (y_1, \ldots, y_d) \in V$. Further, the dth coordinate of $\mathbf{c}_V$ exactly equals $\mathrm{dist}(\mathbf{c}_V, \partial V)$. Then, from the definition of the centroid, we have that

$$\begin{aligned}
\mathrm{dist}(\mathbf{c}_V, \partial V) &= \frac{\int_V \rho(\mathbf{y}) y_d \, \mathrm{d}\mathbf{y}}{\int_V \rho(\mathbf{y}) \, \mathrm{d}\mathbf{y}} \\
&= \frac{\int_0^{R_V} \int_{A(y_d)} \rho(\tilde{\mathbf{y}}, y_d) y_d \, \mathrm{d}\tilde{\mathbf{y}} \mathrm{d}y_d}{M(V)} \\
&= \frac{\int_0^{R_V} y_d [\int_{A(y_d)} \rho(\tilde{\mathbf{y}}, y_d) \, \mathrm{d}\tilde{\mathbf{y}}] \, \mathrm{d}y_d}{M(V)}, \qquad (3.4)
\end{aligned}$$

where $\tilde{\mathbf{y}} = (y_1, \ldots, y_{d-1})$ and $A(y_d) = \{\tilde{\mathbf{y}} \in \mathbb{R}^{d-1} \mid (\tilde{\mathbf{y}}, y_d) \in V\}$. Let us define

$$f_\rho(y_d) := \int_{A(y_d)} \rho(\tilde{\mathbf{y}}, y_d) \, \mathrm{d}\tilde{\mathbf{y}}.$$

The representation above has essentially reduced the problem to a similar one-dimensional problem. To bound (3.4) from below, follow the following steps. In the first step, we find a set such that $\int_S f_\rho(y_d) \, \mathrm{d}y_d$ is sufficiently large and f_ρ is bounded below away from 0. Second, a lower bound on this Lebesgue measure of this set is found. Finally, the third step is to combine these estimates to get a lower bound on (3.4) in terms of the mass of V.

Step 1. For any number $\mu > 0$, define $S_\mu = \{y_d \mid f_\rho(y_d) > \mu R_V^{d-1}\}$ and let $L(S_\mu)$ denote the measure of S_μ. First, we have

$$\begin{aligned} M(V) &= \int_V \rho(\mathbf{y})\, \mathrm{d}\mathbf{y} \\ &= \int_0^{R_V} f_\rho(y_d)\, \mathrm{d}y_d \\ &= \int_{S_\mu} f_\rho(y_d)\, \mathrm{d}y_d + \int_{[0,R_V]-S_\mu} f_\rho(y_d)\, \mathrm{d}y_d \\ &\le \int_{S_\mu} f_\rho(y_d)\, \mathrm{d}y_d + \mu R_V^d. \end{aligned}$$

Set $\mu_1 = \frac{M(V)}{2R_V^d}$; then the above inequality gives us

$$\int_{S_{\mu_1}} f_\rho(y_d)\, \mathrm{d}y_d \ge M(V) - \mu_1 R_V^d = M(V) - \frac{M(V)}{2} = \frac{M(V)}{2}. \tag{3.5}$$

Step 2. From the fact that

$$M(V) \ge \int_{S_\mu} f_\rho(y_d)\, \mathrm{d}y_d \ge L(S_\mu)\mu R_V^{d-1},$$

it is easy to deduce that

$$L(S_\mu) \le \frac{M(V)}{\mu R_V^{d-1}},$$

and so that

$$m(V_\mu) \le L(S_\mu) R_V^{d-1} \le \frac{M(V)}{\mu},$$

where $V_\mu = \{(\tilde{\mathbf{y}}, y_d) \in V \mid y_d \in S_\mu,\ \tilde{\mathbf{y}} \in A(y_d)\}$. Setting $\mu_2 = \frac{M(V)}{\eta_{M(V)/4}}$, we have

$$m(V_{\mu_2}) \le \frac{M(V)}{\mu_2} = \eta_{M(V)/4}.$$

Then, by Lemma 3.1, we obtain that

$$\int_{S_{\mu_2}} f_\rho(y_d)\, \mathrm{d}y_d = \int_{S_{\mu_2}} \int_{A(y_d)} \rho(\tilde{\mathbf{y}}, y_d)\, \mathrm{d}\tilde{\mathbf{y}} \mathrm{d}y_d = \int_{V_{\mu_2}} \rho(\mathbf{y})\, \mathrm{d}\mathbf{y} \le \frac{M(V)}{4}. \tag{3.6}$$

According to (3.5) and (3.6), we get

$$\int_{S_{\mu_1}-S_{\mu_2}} f_\rho(y_d)\, \mathrm{d}y_d \ge \frac{M(V)}{2} - \frac{M(V)}{4} \ge \frac{M(V)}{4}. \tag{3.7}$$

Notice that we also have

$$\int_{S_{\mu_1}-S_{\mu_2}} f_\rho(y_d)\, \mathrm{d}y_d \le \mu_2 R_V^{d-1} L(S_{\mu_1}). \tag{3.8}$$

The combination of (3.7) and (3.8) gives us

$$L(S_{\mu_1}) \geq \frac{M(V)}{4\mu_2 R_V^{d-1}}. \tag{3.9}$$

Step 3. Using (3.4) and (3.9), we get

$$\begin{aligned}
\operatorname{dist}(\mathbf{c}_V, \partial V) &= \frac{\int_0^{R_V} y_d f_\rho(y_d)\, \mathrm{d}y_d}{M(V)} \\
&\geq \frac{\int_{S_{\mu_1}} y_d \mu_1 R_V^{d-1}\, \mathrm{d}y_d}{M(V)} \\
&\geq \frac{\mu_1 R_V^{d-1} \int_0^{L(S_{\mu_1})} y_d\, \mathrm{d}y_d}{M(V)} \\
&\geq \frac{\mu_1 R_V^{d-1} [L(S_{\mu_1})]^2}{2M(V)} \\
&\geq \frac{\mu_1 M(V)}{32\mu_2^2 R_V^{d-1}} \\
&\geq \frac{\eta^2_{M(V)/4}}{64 R_V^{2d-1}},
\end{aligned}$$

which completes the proof. □

Remark 3.1. If $\rho \in L^q(\Omega)$ for some $q > 1$, then δ_V can be selected such that it does not depend on the constant $\eta_{M(V)/4}$ such as

$$\delta_V := \frac{M(V)^{2q/(q-1)}}{4^{(5q-3)/(q-1)} \|\rho\|^{2q/(q-1)}_{L^q(V)} R_V^{2d-1}}.$$

This follows from that fact that applying the Hölder inequality to $\int_{S_\mu} f_\rho(y_d)\, \mathrm{d}y_d$ gives us that

$$\int_{S_\mu} f_\rho(y_d)\, \mathrm{d}y_d \leq \frac{\|\rho\|_{L^q(V)} [M(V)]^{(q-1)/q}}{\mu^{(q-1)/q}},$$

and so that

$$\int_{S_{\mu_2}} f_\rho(y_d)\, \mathrm{d}y_d \leq \frac{M(V)}{4},$$

where $\mu_2 = \frac{(4\|\rho\|_{L^q(V)})^{q/(q-1)}}{[M(V)]^{1/(q-1)}}$.

The next lemma is a local estimate on the difference in the quantization energy over a particular set with respect to different generators. For our purposes, it is important to compare the energy with respect to the centroid to the energy generated by a point near the boundary. To prove this, the previous lemma is essential; it ensures that the centroid is sufficiently far away from the boundary. This again leads to an estimate that depends on the particular convex set V through $M(V)$.

For any convex set $V \subset \Omega$, let us define the V-dependent constant γ_V by

$$\gamma_V := \frac{M(V)\delta_V^2}{4}.$$

LEMMA 3.3. *Let Assumption* 3.1 *be satisfied and $V \subset \Omega$ be a convex set. Then for any $\mathbf{z} \in V$ with $\operatorname{dist}(\mathbf{z}, \partial V) \leq \frac{\delta_V}{2}$, it holds that*

$$\int_V \rho(\mathbf{y})|\mathbf{y} - \mathbf{c}_V|^2 \, \mathrm{d}\mathbf{y} \leq \int_V \rho(\mathbf{y})|\mathbf{y} - \mathbf{z}|^2 \, \mathrm{d}\mathbf{y} - \gamma_V, \tag{3.10}$$

where $\mathbf{c}_V$ is the centroid of V.

Proof. First, from Lemma 3.2, it is obvious that

$$|\mathbf{z} - \mathbf{c}_V| \geq \operatorname{dist}(\mathbf{c}_V, \partial V) - \operatorname{dist}(\mathbf{z}, \partial V) \geq \frac{\delta_V}{2}.$$

We have

$$\begin{aligned}
&\int_V \rho(\mathbf{y})|\mathbf{y} - \mathbf{z}|^2 \, \mathrm{d}\mathbf{y} - \int_V \rho(\mathbf{y})|\mathbf{y} - c_V|^2 \, \mathrm{d}\mathbf{y} \\
&\quad = \int_V \rho(\mathbf{y})\langle \mathbf{y} - \mathbf{z}, \mathbf{y} - \mathbf{z}\rangle \mathrm{d}\mathbf{y} - \int_V \rho(\mathbf{y})\langle \mathbf{y} - \mathbf{c}_V, \mathbf{y} - \mathbf{c}_V\rangle \mathrm{d}\mathbf{y} \\
&\quad = \int_V \rho(\mathbf{y})(|\mathbf{y}|^2 + |\mathbf{z}|^2 - 2\langle \mathbf{y}, \mathbf{z}\rangle) \mathrm{d}\mathbf{y} \\
&\qquad - \int_V \rho(\mathbf{y})(|\mathbf{y}|^2 + |\mathbf{c}_V|^2 - 2\langle \mathbf{y}, \mathbf{c}_V >\rangle \mathrm{d}\mathbf{y} \\
&\quad = \int_V \rho(\mathbf{y})(|\mathbf{z}|^2 - |\mathbf{c}_V|^2) \, \mathrm{d}\mathbf{y} + 2\int_V \rho(\mathbf{y})\langle \mathbf{y}, \mathbf{c}_V - \mathbf{z}\rangle \mathrm{d}\mathbf{y}.
\end{aligned}$$

Notice that

$$\int_V \rho(\mathbf{y})(|\mathbf{z}|^2 - |\mathbf{c}_V|^2) \, \mathrm{d}\mathbf{y} = M(V)(|\mathbf{z}|^2 - |\mathbf{c}_V|^2) \tag{3.11}$$

and

$$\begin{aligned}
\int_V \rho(\mathbf{y})\langle \mathbf{y}, \mathbf{c}_V - \mathbf{z}\rangle \mathrm{d}\mathbf{y} &= \left\langle \int_V \mathbf{y}\rho(\mathbf{y}) \, \mathrm{d}\mathbf{y}, \mathbf{c}_V - \mathbf{z} \right\rangle \\
&= \langle M(V)\mathbf{c}_V, \mathbf{c}_V - \mathbf{z}\rangle \\
&= M(V)(|\mathbf{c}_V|^2 - \langle \mathbf{c}_V, \mathbf{z}\rangle). \qquad (3.12)
\end{aligned}$$

Using (3.11) and (3.12), we go further and get

$$\begin{aligned}
&\int_V \rho(\mathbf{y})|\mathbf{y} - \mathbf{z}|^2 \, \mathrm{d}\mathbf{y} - \int_V \rho(\mathbf{y})|\mathbf{y} - \mathbf{c}_V|^2 \, \mathrm{d}\mathbf{y} \\
&\quad = M(V)(|\mathbf{z}|^2 + |\mathbf{c}_V|^2 - 2\langle \mathbf{c}_V, \mathbf{z}\rangle) \\
&\quad = M(V)|\mathbf{z} - \mathbf{c}_V|^2 \\
&\quad \geq \frac{M(V)\delta_V^2}{4},
\end{aligned}$$

which deduces (3.10) and completes the proof. ∎

Before presenting the next lemma, we need to define some notation. For any point set $\mathbf{Z} \in \Omega^n$, denote by $\mathbf{Z}_i$ the ith point of $\mathbf{Z}$.

DEFINITION 3.1. *Let $k, n \in \mathbb{N}$. For any points $\mathbf{Z} \in \hat{\Omega}^k$ and $\mathbf{Y} \in \hat{\Omega}^n$, define the d_*-distance from $\mathbf{Z}$ to $\mathbf{Y}$ by*

$$d_*(\mathbf{Y}, \mathbf{Z}) = \max_{1 \le j \le k} \min_{1 \le i \le n} |\mathbf{Z}_j - \mathbf{Y}_i|.$$

Then the natural definition of the distance between $\mathbf{Z}$ and $\mathbf{Y}$ is given by

$$d(\mathbf{Y}, \mathbf{Z}) = \max(d_*(\mathbf{Y}, \mathbf{Z}), d_*(\mathbf{Z}, \mathbf{Y})).$$

This is a generalization of the regular Euclidean difference between two elements of Ω^n to the case of vectors of a different length. This definition will become useful for us when we consider a degenerate set of generators in the proof of our main theorem. Provided with this definition, let us consider the set theoretic difference of the two Voronoi tessellations generated by the vectors $\mathbf{Z}$ and $\mathbf{Y}$ as

$$\sum_{i=1}^{k} M\left(V_i(\mathbf{Z}) - \cup_{j \in S_i} V_j(\mathbf{Y})\right).$$

This definition is illustrated in Figure 3.1 below.

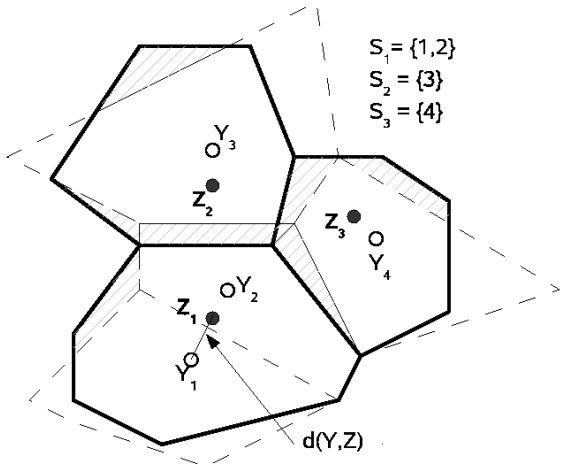


FIG. 3.1. *An illustration of the definition of $d(\mathbf{Y}, \mathbf{Z})$ and the difference of Voronoi tessellations generated by $\mathbf{Z}$ and $\mathbf{Y}$, where $\mathbf{Z} \in \hat{\Omega}_3$ and $\mathbf{Y} \in \hat{\Omega}_4$ (shaded).*

The next lemma for quantifies the fact that if two possibly degenerate points in Ω^n are close together in the sense of Definition 3.1, then the corresponding Voronoi regions of these points are also close together in some sense. The appropriate measurement of how far apart these Voronoi regions are is the total mass of the difference of corresponding Voronoi regions.

LEMMA 3.4. *Let Assumption 3.1 be satisfied. Fix $\mathbf{Z} \in \hat{\Omega}^k$ for $k \le n$. For any $\epsilon > 0$, there exists a $\delta_\epsilon > 0$ such that if $\mathbf{Y} \in \hat{\Omega}^n$ and $d(\mathbf{Z}, \mathbf{Y}) \le \delta_\epsilon$, then*

$$\sum_{i=1}^{k} M\left(V_i(\mathbf{Z}) - \cup_{j \in S_i} V_j(\mathbf{Y})\right) \le \epsilon,$$

where $\{S_i\}_{i=1}^k$ *is any partition of* $\{1,\dots,n\}$ *satisfying* $S_i \in \{j \mid |\mathbf{Z}_i - \mathbf{Y}_j| \le |\mathbf{Z}_m - \mathbf{Y}_j|$ *for all* $m \in \{1,\dots,k\}\}$.

Proof. Let us suppose that $\mathbf{Y} \in \hat{\Omega}^n$ and $d(\mathbf{Z},\mathbf{Y}) < \delta$ for some $\delta > 0$. Since $\mathbf{Z} \in \hat{\Omega}^k$ is fixed, let us define the constant $d_{\mathbf{Z}}$ by

$$d_{\mathbf{Z}} = \min_{i \neq j} \frac{|\mathbf{Z}_i - \mathbf{Z}_j|}{2}.$$

We let δ be small enough such that $\delta \le \frac{d_{\mathbf{Z}}}{2}$; then it is not difficult to verify that S_i is a nonempty set for each $1 \le i \le k$. From Definition 3.1, it is also obvious that for any $1 \le i \le k$,

$$|\mathbf{y} - \mathbf{Z}_i| - \delta \le |\mathbf{y} - \mathbf{Y}_j| \le |\mathbf{y} - \mathbf{Z}_i| + \delta \quad \forall\, \mathbf{y} \in \Omega,\ j \in S_i.$$

For any $\mathbf{y}$, if $\mathbf{y} \in V_i(\mathbf{Z}) - \cup_{j\in S_i} V_j(\mathbf{Y})$ (i.e., $\mathbf{y} \in V_i(\mathbf{Z})$ and $\mathbf{y} \notin \cup_{j\in S_i} V_j(\mathbf{Y})$), then we will argue by contradiction that $\operatorname{dist}(\mathbf{y}, \partial V_i(\mathbf{Z}))$, the shortest distance from $\mathbf{y}$ to the boundary of $V_i(\mathbf{Z})$, satisfies

$$\operatorname{dist}(\mathbf{y}, \partial V_i(\mathbf{Z})) < \sigma = \frac{3\delta}{\alpha_{\mathbf{Z}}}, \tag{3.13}$$

where $\alpha_{\mathbf{Z}} = \frac{2}{\sqrt{(1+R_\Omega/d_{\mathbf{Z}})^2 + (R_\Omega/d_{\mathbf{Z}})^2}}$ is a constant depending only on $\mathbf{Z}$ and R_Ω. Let us first assume that $\operatorname{dist}(\mathbf{y}, \partial V_i(\mathbf{Z})) \ge \sigma$. We will show that $\mathbf{y} \in \cup_{j\in S_i} V_j(\mathbf{Y})$. Take any $1 \le j \le k$ with $j \neq i$; without loss of generality, assume that $V_i(\mathbf{Z})$ and $V_j(\mathbf{Z})$ are neighbor Voronoi cells. Notice that $V_i(\mathbf{Z})$ and $V_j(\mathbf{Z})$ are both convex polygons/polyhedra except for the part belonging to the boundary of Ω; see Figure 3.2 for an illustration of two-dimensional cases.

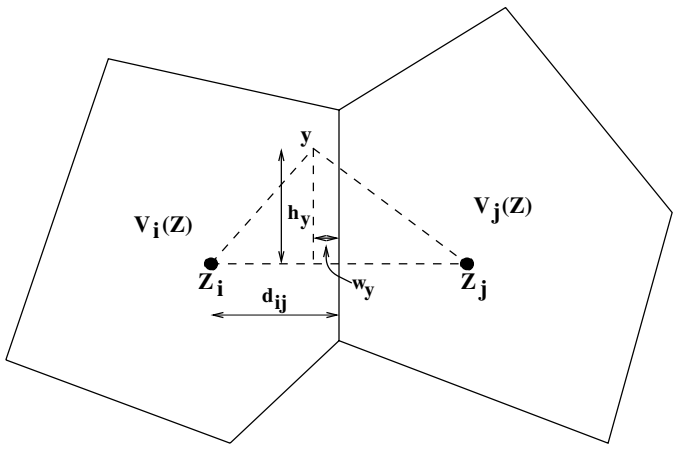


FIG. 3.2. *Two neighboring Voronoi regions.*

One can easily verify that

$$\begin{aligned}|\mathbf{y}-\mathbf{Z}_j|-|\mathbf{y}-\mathbf{Z}_i| &= \sqrt{(d_{ij}+w_{\mathbf{y}})^2+h_{\mathbf{y}}^2}-\sqrt{(d_{ij}-w_{\mathbf{y}})^2+h_{\mathbf{y}}^2}\\ &\geq \frac{2d_{ij}w_{\mathbf{y}}}{\sqrt{(d_{ij}+w_{\mathbf{y}})^2+h_{\mathbf{y}}^2}}\\ &= \frac{2w_{\mathbf{y}}}{\sqrt{(1+w_{\mathbf{y}}/d_{ij})^2+(h_{\mathbf{y}}/d_{ij})^2}}.\end{aligned}$$

Since $d_{ij}\geq d_{\mathbf{Z}}$, $R_\Omega\geq w_{\mathbf{y}}>\sigma$, and $h_{\mathbf{y}}\leq R_\Omega$, we obtain

$$\begin{aligned}|\mathbf{y}-\mathbf{Z}_j|-|\mathbf{y}-\mathbf{Z}_i| &\geq \frac{2\sigma}{\sqrt{(1+R_\Omega/d_{\mathbf{Z}})^2+(R_\Omega/d_{\mathbf{Z}})^2}}\\ &= 3\delta.\end{aligned}$$

Then, for any $j\in S_i$ and any $m\in S_l$ with $l\neq i$, it holds that

$$\begin{aligned}|\mathbf{y}-\mathbf{Y}_j| &\leq |\mathbf{y}-\mathbf{Z}_i|+\delta\\ &\leq |\mathbf{y}-\mathbf{Z}_l|-3\delta+\delta\\ &\leq |\mathbf{y}-\mathbf{Y}_m|+\delta-2\delta\\ &\leq |\mathbf{y}-\mathbf{Y}_m|-\delta\\ &< |\mathbf{y}-\mathbf{Y}_m|.\end{aligned}$$

This means that $\mathbf{y}\in\cup_{j\in S_i}V_j(\mathbf{Y})$ and gives us a contradiction with $\mathbf{y}\notin\cup_{j\in S_i}V_j(\mathbf{Y})$, so that the inequality (3.13) holds. Since $V_i(\mathbf{Z})\subset\Omega$ is convex and $\operatorname{diam}(\Omega)=R_\Omega<+\infty$, it holds that

$$m(\partial V_i(\mathbf{Z}))\leq S_d\left(\frac{R_\Omega}{2}\right)^{d-1}, \tag{3.14}$$

where $S_d=\frac{2\pi^{d/2}}{\Gamma(d/2)}$ denotes the hypersurface area of an d-sphere of unit radius; see [3] for details. Then the inequality (3.13) with (3.14) implies that

$$m(V_i(\mathbf{Z})-\cup_{j\in S_i}V_j(\mathbf{Y}))\leq\frac{\sigma S_d R_\Omega^{d-1}}{2^{d-1}}.$$

Then selecting

$$\delta=\delta_\epsilon:=\min\left(\frac{d_{\mathbf{Z}}}{2},\ \frac{2^{d-1}\eta_{\epsilon/k}\alpha_{\mathbf{Z}}}{3S_dR_\Omega^{d-1}}\right)$$

ensures that $m(V_i(\mathbf{Z})-\cup_{j\in S_i}V_j(\mathbf{Y}))\leq\eta_{\epsilon/k}$. Due to Lemma 3.1, this means that

$$\begin{aligned}\sum_{i=1}^k M\left(V_i(\mathbf{Z})-\cup_{j\in S_i}V_j(\mathbf{Y})\right) &= \sum_{i=1}^k\int_{V_i(\mathbf{Z})-\cup_{j\in S_i}V_j(\mathbf{Y})}\rho(\mathbf{y})\,\mathrm{d}\mathbf{y}\\ &\leq \sum_{i=1}^k\frac{\epsilon}{k}\\ &\leq \epsilon,\end{aligned}$$

and thus the result holds. ∎

Remark 3.2. In the case that $\rho \in L^q(\Omega)$ for some $q > 1$, the Hölder inequality again can be used to find a δ_ϵ in the above lemma which does not depend on the constant $\eta_{\epsilon/k}$ such as

$$\delta = \delta_\epsilon := \min\left(\frac{d_{\mathbf{Z}}}{2}, \frac{2^{d-1}\alpha_{\mathbf{Z}}}{3S_d R_\Omega^{d-1}}\left(\frac{\epsilon}{k\|\rho\|_{L^q(\Omega)}}\right)^{q/(q-1)}\right).$$

3.2. Nondegeneracy of limit points. Now, let us present our first important result.

THEOREM 3.5. *Let Assumption* 3.1 *be satisfied. Given $n \in \mathbb{N}$ and any initial point $\mathbf{Z}^0 \in \hat{\Omega}^n$, let $\{\mathbf{Z}^i\}_{i=0}^\infty$ be the iterates of the Lloyd algorithm starting with $\mathbf{Z}^0$. Then any limit point $\mathbf{Z}$ of these iterates is nondegenerate (i.e., $\mathbf{Z} \in \hat{\Omega}^n$).*

Proof. Let $\mathbf{Z} \in \overline{\Omega}^n$ be any limit point of these iterates $\{\mathbf{Z}^i\}_{i=0}^\infty$. We will argue by contradiction that $\mathbf{Z} \in \hat{\Omega}^n$. Let us first assume that $\mathbf{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_n)$ is degenerate, i.e., in the sense that several of its components are equal. Without loss of generality, assume that $\mathbf{Z}_1 = \cdots = \mathbf{Z}_r$ for some $1 < r \le n$.

Since $\mathbf{Z}^i$'s are Lloyd iterates, by the continuity of $\mathcal{G}$ on $\overline{\Omega}^n$ and the monotone decrease of $\mathcal{G}$ along $\mathbf{Z}^i$'s due to (2.2), we know that

$$\mathcal{G}(\mathbf{Z}^i) \searrow_{i\to+\infty} \mathcal{G}(\mathbf{Z}); \tag{3.15}$$

i.e., $\mathcal{G}(\mathbf{Z}^i)$ monotonically decreases and converges to $\mathcal{G}(\mathbf{Z})$. Let $\tilde{\mathbf{Z}} \in \hat{\Omega}^k$ ($k = n-r+1 < n$) be the set of points in $\mathbf{Z}$ without duplication; i.e.,

$$\tilde{\mathbf{Z}} = (\tilde{\mathbf{Z}}_1, \ldots, \tilde{\mathbf{Z}}_k) = (\mathbf{Z}_1, \mathbf{Z}_{r+1}, \ldots, \mathbf{Z}_n).$$

It is obvious that $\mathcal{G}(\tilde{\mathbf{Z}}) = \mathcal{G}(\mathbf{Z})$. From $\{\mathbf{Z}^i\}_{i=0}^\infty$, select a subsequence $\{\mathbf{Z}^{i_j}\}_{j=0}^\infty$ such that $\mathbf{Z}^{i_j} \to \mathbf{Z}$ as $j \to +\infty$.

Now define

$$\alpha = \frac{M(V_1(\tilde{\mathbf{Z}}))}{2r}, \quad \beta = \frac{\eta_{\alpha/4}^2}{64R_\Omega^{2d-1}}, \quad \text{and} \quad \gamma = \frac{\alpha\beta^2}{4}.$$

Obviously we have $\alpha, \beta, \gamma > 0$ since $(V_1(\tilde{\mathbf{Z}}))$ is a convex set with positive measure.

Now let us select a j^* large enough such that the following inequalities hold:

$$M(\cup_{m=1}^r V_m(\mathbf{Z}^{i_{j^*}})) > \frac{M(V_1(\tilde{\mathbf{Z}}))}{2}, \tag{3.16}$$

$$\mathcal{G}(\mathbf{Z}^{i_{j^*}}) - \mathcal{G}(\tilde{\mathbf{Z}}) < \frac{\gamma}{2}, \tag{3.17}$$

$$|\mathbf{Z}_m^{i_{j^*}} - \mathbf{Z}_m| < \frac{\beta}{4}, \quad m = 1, 2, \ldots, n. \tag{3.18}$$

Inequality (3.16) holds for large j^* by Lemma 3.4 using $\epsilon = \frac{M(V_1(\tilde{\mathbf{Z}}))}{2}$. Further, that also means that for some $m \in \{1, \ldots, r\}$, $M(V_m(\mathbf{Z}^{i_{j^*}})) > \alpha$. Without loss of generality, assume this m is 1, i.e., $M(V_1(\mathbf{Z}^{i_{j^*}})) > \alpha$. That gives us

$$\delta_{V_1(\mathbf{Z}^{i_{j^*}})} \ge \beta, \qquad \gamma_{V_1(\mathbf{Z}^{i_{j^*}})} \ge \gamma. \tag{3.19}$$

The inequalities (3.18) and (3.19) imply

$$\operatorname{dist}(\mathbf{Z}_1^{i_{j^*}}, \partial V_1(\mathbf{Z}^{i_{j^*}})) \le |\mathbf{Z}_1^{i_{j^*}} - \mathbf{Z}_1| + |\mathbf{Z}_2^{i_{j^*}} - \mathbf{Z}_1| < \frac{\beta}{2} \le \frac{\delta_{V_1(\mathbf{Z}^{i_{j^*}})}}{2}.$$

Then by Lemma 3.3 and (3.19), we immediately obtain

$$\begin{aligned}
&\int_{V_1(\mathbf{Z}^{i_{j^*}})} \rho(\mathbf{y})|\mathbf{y} - \mathbf{c}_{V_1(\mathbf{Z}^{i_{j^*}})}|^2 \, \mathrm{d}\mathbf{y} \\
&\qquad \le \int_{V_1(\mathbf{Z}^{i_{j^*}})} \rho(\mathbf{y})|\mathbf{y} - \mathbf{Z}_1^{i_{j^*}}|^2 \, \mathrm{d}\mathbf{y} - \gamma_{V_1(\mathbf{Z}^{i_{j^*}})} \\
&\qquad \le \int_{V_1(\mathbf{Z}^{i_{j^*}})} \rho(\mathbf{y})|\mathbf{y} - \mathbf{Z}_1^{i_{j^*}}|^2 \, \mathrm{d}\mathbf{y} - \gamma. \qquad (3.20)
\end{aligned}$$

Now, let us estimate the quantization energy $\mathcal{G}(\mathbf{T}(\mathbf{Z}^{i_{j^*}}))$. Using the dual minimization property, we get

$$\begin{aligned}
\mathcal{G}(\mathbf{T}(\mathbf{Z}^{i_{j^*}})) &= \mathcal{H}(\mathbf{T}(\mathbf{Z}^{i_{j^*}}), \mathbf{T}(\mathbf{Z}^{i_{j^*}})) \\
&\le \mathcal{H}(\mathbf{Z}^{i_{j^*}}, \mathbf{T}(\mathbf{Z}^{i_{j^*}})) \\
&= \int_{V_1(\mathbf{Z}^{i_{j^*}})} \rho(\mathbf{y})|\mathbf{y} - \mathbf{T}_1(\mathbf{Z}^{i_{j^*}})|^2 \, \mathrm{d}\mathbf{y} \\
&\qquad + \sum_{k=2}^{n} \int_{V_k(\mathbf{Z}^{i_{j^*}})} \rho(\mathbf{y})|\mathbf{y} - \mathbf{T}_k(\mathbf{Z}^{i_{j^*}})|^2 \, \mathrm{d}\mathbf{y}.
\end{aligned}$$

It is easy to see that

$$\sum_{k=2}^{n} \int_{V_k(\mathbf{Z}^{i_{j^*}})} \rho(\mathbf{y})|\mathbf{y} - \mathbf{T}_k(\mathbf{Z}^{i_{j^*}})|^2 \, \mathrm{d}\mathbf{y} \le \sum_{k=2}^{n} \int_{V_k(\mathbf{Z}^{i_{j^*}})} \rho(\mathbf{y})|\mathbf{y} - \mathbf{Z}_k^{i_{j^*}}|^2 \, \mathrm{d}\mathbf{y} \qquad (3.21)$$

since $\mathbf{T}_k(\mathbf{Z}^{i_{j^*}})$ is the centroid of $V_k(\mathbf{Z}^{i_{j^*}})$. With (3.17), (3.20), and (3.21), we have

$$\begin{aligned}
\mathcal{G}(\mathbf{T}(\mathbf{Z}^{i_{j^*}})) &\le \int_{V_1(\mathbf{Z}^{i_{j^*}})} \rho(\mathbf{y})|\mathbf{y} - \mathbf{Z}_1^{i_{j^*}}|^2 \, \mathrm{d}\mathbf{y} - \gamma \\
&\qquad + \sum_{k=2}^{n} \int_{V_k(\mathbf{Z}^{i_{j^*}})} \rho(\mathbf{y})|\mathbf{y} - \mathbf{Z}_k^{i_{j^*}}|^2 \, \mathrm{d}\mathbf{y} \\
&= \mathcal{G}(\mathbf{Z}^{i_{j^*}}) - \gamma \\
&< \mathcal{G}(\tilde{\mathbf{Z}}) + \frac{\gamma}{2} - \gamma \\
&= \mathcal{G}(\tilde{\mathbf{Z}}) - \frac{\gamma}{2},
\end{aligned}$$

which contradicts the fact (3.15). We conclude that $\mathbf{Z} \in \hat{\Omega}^n$; i.e., any limit point of the Lloyd iteration cannot be a degenerate point. ∎

THEOREM 3.6. *Let Assumption* 3.1 *be satisfied. Given* $n \in \mathbb{N}$ *and any initial point* $\mathbf{Z}^0 \in \hat{\Omega}^n$, *let* $\{\mathbf{Z}^i\}_{i=0}^{\infty}$ *be the iterates of the Lloyd algorithm starting with* $\mathbf{Z}^0$. *Then there exists a constant* $D_{\mathbf{Z}^0} > 0$ *depending only on* n, Ω, ρ, *and* $\mathbf{Z}^0$ *such that for any* $i \geq 0$,

$$|\mathbf{Z}_j^i - \mathbf{Z}_k^i| > D_{\mathbf{Z}^0}$$

for all $j, k \in \{1, \dots, n\}$ *and* $j \neq k$.

Proof. We will use the contradiction argument again. Let us assume that the theorem is not true, then for all $m \in \mathbb{N}$, there always exist an index i_m, some $j_m, k_m \in \{1, \dots, n\}$, and $j_m \neq k_m$ such that

$$|\mathbf{Z}_{j_m}^{i_m} - \mathbf{Z}_{k_m}^{i_m}| < \frac{1}{m}.$$

Let us consider the subsequence $\{\mathbf{Z}^{i_m}\}_{m=1}^{\infty}$. Without loss of generality, assume the index set $I = \{i_m\}_{m=1}^{\infty}$ is a monotonically increasing sequence (otherwise we can do that by adding some restriction such as $i_{k+1} > i_k$ during choosing $\{\mathbf{Z}^{i_m}\}_{m=1}^{\infty}$). Notice that the set $JK = \{(j_m, k_m)\}_{m=1}^{\infty}$ has at most $n(n-1)$ different elements, so there exists a pair $\{(j^*, k^*)\} \in JK$ that appears infinitely many times along the increasing of m, and let us record their appearances by the increasing sequence $\{m_1, m_2, \dots\} \subset \mathbb{N}$. This means that

$$|\mathbf{Z}_{j^*}^{i_{m_p}} - \mathbf{Z}_{k^*}^{i_{m_p}}| < \frac{1}{m_p} \quad \forall\, p \in \mathbb{N}. \tag{3.22}$$

Since $\{\mathbf{Z}^{i_{m_p}}\}$ are bounded, there always exists a convergent subsequence of $\{\mathbf{Z}^{i_{m_p}}\}$, and assume $\mathbf{Z}$ is the corresponding limit point. Due to (3.22), we must have $\mathbf{Z}_{j^*} = \mathbf{Z}_{k^*}$, which means that $\mathbf{Z}$ is a degenerate point. However, $\mathbf{Z}$ is also a limit point of the original Lloyd iterates $\{\mathbf{Z}^i\}_{i=0}^{\infty}$, and according to Theorem 3.5, $\mathbf{Z}$ could not be a degenerate point. We then get a contradiction, and so the theorem holds. ∎

Remark 3.3. In Theorem 3.6, we are only able to show existence of the lower bound of the distances between any pair of generators of all Lloyd iterates starting with $\mathbf{Z}^0$. For the one-dimensional case, an explicit estimate was derived in [5] when the density function is strictly positive and smooth.

A direct consequence of Theorem 3.6 is the following corollary.

COROLLARY 3.7. *Let Assumption* 3.1 *be satisfied. Given* $n \in \mathbb{N}$ *and any initial point* $\mathbf{Z}^0 \in \hat{\Omega}^n$, *let* $\{\mathbf{Z}^i\}_{i=0}^{\infty}$ *be the iterates of the Lloyd algorithm starting with* $\mathbf{Z}^0$. *Then the Lloyd map is continuous at any of the iterates* $\{\mathbf{Z}^i\}_{i=0}^{\infty}$ *and* $m(V_j(\mathbf{Z}^i)) \geq (\frac{D_{\mathbf{Z}^0}}{2})^d$ *for any* $1 \leq j \leq n$ *and* $i \geq 0$.

3.3. Main result. Finally, let us present our main result in the following theorem.

THEOREM 3.8 (weak global convergence). *Let Assumption* 3.1 *be satisfied. Given* $n \in \mathbb{N}$ *and any initial point* $\mathbf{Z}^0 \in \hat{\Omega}^n$, *let* $\{\mathbf{Z}^i\}_{i=0}^{\infty}$ *be the iterates of the Lloyd algorithm starting with* $\mathbf{Z}^0$. *Then the following hold:*

1. $\{\mathbf{Z}^i\}_{i=0}^{\infty}$ *is weakly convergent (i.e.,* $\lim_{i \to +\infty} \nabla \mathcal{G}(\mathbf{Z}^i) = 0$*) and any limit point of* $\{\mathbf{Z}^i\}_{i=0}^{\infty}$ *is also a nondegenerate critical point of the quantization energy* $\mathcal{G}$ *(and thus a CVT).*
2. *Moreover, it also holds that* $\lim_{i \to +\infty} \|\mathbf{Z}^{i+1} - \mathbf{Z}^i\| = 0$.

Proof. Using the result of Corollary 3.7, we see that we can define a compact set (away from the degenerate points) such that for any initial point $\mathbf{Z}^0 \in \hat{\Omega}^n$, the Lloyd

iterates (the images of the Lloyd maps) will stay in such a compact set. Thus, we can apply Theorem 2.1 and Corollary 3.7 to deduce the weak global convergence of the Lloyd iterations. The first part of the theorem is then complete with the help of Theorem 3.5.

From Corollary 3.7, we also know that for any $1 \le k \le n$ and $i \ge 0$,

$$m(V_k(\mathbf{Z}^i)) \ge \left(\frac{D_{\mathbf{Z}^0}}{2}\right)^d > 0.$$

Together with Lemma 3.1 and the assumption that ρ is positive almost everywhere, we know that there exists some constant $M^* > 0$ such that

$$M(V_k(\mathbf{Z}^i)) \ge M^* > 0. \tag{3.23}$$

Using (2.1), we then get

$$\begin{aligned}
\|\nabla \mathcal{G}(\mathbf{Z}^i)\| &= \left(\sum_{k=1}^{n} \Big|\frac{\partial \mathcal{G}}{\partial \mathbf{z}_k}(\mathbf{Z}^i)\Big|^2\right)^{1/2} \\
&= \left(\sum_{k=1}^{n} |2M(V_k(\mathbf{Z}^i))(\mathbf{Z}_k^i - \mathbf{T}_k(\mathbf{Z}^i))|^2\right)^{1/2} \\
&\ge 2M^* \left(\sum_{k=1}^{n} |\mathbf{Z}_k^i - \mathbf{T}_k(\mathbf{Z}^i)|^2\right)^{1/2} \\
&= 2M^* \|\mathbf{Z}^i - \mathbf{Z}^{i+1}\|. \qquad (3.24)
\end{aligned}$$

Consequently, we get

$$0 \le \lim_{i \to +\infty} \|\mathbf{Z}^{i+1} - \mathbf{Z}^i\| \le \frac{1}{2M^*} \lim_{i \to +\infty} \|\nabla \mathcal{G}(\mathbf{Z}^i)\| = 0,$$

which deduces the second part of this theorem. □

Remark 3.4. It is a consequence of part (1) of Theorem 3.8 that the generators of any limit point configuration of the Lloyd algorithm cannot lie on the boundary of the convex domain, since the limit point is a CVT and the density function ρ is positive almost everywhere.

Remark 3.5. Part (2) of Theorem 3.8 in fact tells us that the Lloyd method always terminates under general practical stopping criteria.

4. Conclusions and open questions. In this paper, we prove that any limit point of the Lloyd iteration is nondegenerate provided that Ω is a convex and bounded set and ρ belongs to $L^1(\Omega)$ and is positive almost everywhere in any dimensional space. It follows that, for every initial point set, the Lloyd iteration always approaches the set of nondegenerate critical points of the so-called quantization energy (which is exactly the set of CVTs). These results go beyond those presented in earlier papers; they enlarge the class of densities in the one-dimensional case and assign a stronger meaning to the statement of global convergence of the Lloyd iteration by eliminating the possibility of degeneracy. Although the proofs in their generality hold for all L^1 density functions, we have been able to provide a more constructive characterization

for the constants δ_V and δ_ϵ in the case of $L^q, q > 1$, which possess their own merit from a practical point of view.

There are still some open questions. Throughout our proof, the convexity of Ω is required in many places. The first question is whether this condition is really necessary or whether Ω can be a nonconvex set with some constraints. Another question is under which conditions weak convergence transforms into a strong single limit-point convergence. Also, it remains to explore whether the proof can be extended to replace the density function with a more general measure; ideally, it would be extended to a set of measures more general than those considered in [22]. Finally, it is also very meaningful and interesting to study the convergence properties of the Lloyd algorithm when the CVT is defined under some other error weighting functions, such as the very commonly used l^1 and l^∞ that arise in many engineering applications. The geometric proof provided here clearly provides room for such generalizations; however, the detailed analysis would complicate current presentation and hence is left to a separate treatment. These questions and generalizations are the focus of current research and will be addressed in subsequent publications.

Appendix. Proof of Lemma 3.1.

Proof. Suppose that the claim is false. Then there exists an $\epsilon_0 > 0$ such that for all $k \in \mathbb{N}$, $\exists A_k \subset \Omega$ with $m(A_k) \leq (1/2)^{k+1}$, it holds that $\int_{A_k} \rho(\mathbf{y})\, \mathrm{d}\mathbf{y} > \epsilon_0$. Let $B_k = \cup_{i=k}^{\infty} A_i$; clearly $m(B_k) \leq \sum_{k=1}^{\infty} m(A_i) \leq (1/2)^k$ and $\int_{B_k} \rho(\mathbf{y})\, \mathrm{d}\mathbf{y} > \epsilon_0$ for all $k \in \mathbb{N}$. Since $\{\int_{B_k} \rho(\mathbf{y})\, \mathrm{d}\mathbf{y}\}_{k=1}^{\infty}$ is a positive nonincreasing sequence, it is easy to see that

$$\lim_{k \to +\infty} \int_{B_k} \rho(\mathbf{y})\, \mathrm{d}\mathbf{y} \geq \epsilon_0 > 0.$$

On the other hand, notice that $B_1 \supset B_2 \supset \cdots$. By the dominated convergence theorem in measure theory, we have

$$\begin{aligned} \lim_{k \to +\infty} \int_{B^*} \rho(\mathbf{y})\, \mathrm{d}\mathbf{y} &= \lim_{k \to +\infty} \int_{\Omega} (\mathbf{1}_{B_k}(\mathbf{y}) \rho(\mathbf{y}))\, \mathrm{d}\mathbf{y} \\ &= \int_{\Omega} \lim_{k \to +\infty} (\mathbf{1}_{B_k}(\mathbf{y}) \rho(\mathbf{y}))\, \mathrm{d}\mathbf{y} \\ &= 0, \end{aligned}$$

which gives us a contradiction, and so this lemma holds. ∎

Acknowledgments. The authors are grateful to Noel Walkington of Carnegie Mellon University, Qiang Du of Penn State University, and Max Gunzburger of Florida State University for many stimulating discussions. Special thanks are extended to the anonymous referees for providing valuable suggestions that improved the quality of the presentation and for bringing several relevant references to the authors' attention. M. Emelianenko gratefully acknowledges the support provided by the Center for Nonlinear Analysis at Carnegie Mellon University under the National Science Foundation grant 0405343.

REFERENCES

[1] E. F. Abaya and G. L. Wise, *Convergence of vector quantizers with applications to optimal quantization*, SIAM J. Appl. Math., 44 (1984), pp. 183–189.

[2] E. S. Barnes and N. J. Sloane, *The optimal lattice quantizer in three dimensions*, SIAM J. Algebraic Discrete Methods, 4 (1983), pp. 30–41.
[3] J. H. Conway and N. J. Sloane, *Sphere Packings, Lattices and Groups*, Springer-Verlag, New York, 1988.
[4] Q. Du and M. Emelianenko, *Uniform convergence of a nonlinear energy-based multilevel quantization scheme via centroidal Voronoi tessellations*, SIAM J. Numer. Anal., to appear.
[5] Q. Du, M. Emelianenko, and L. Ju, *Convergence of the Lloyd algorithm for computing centroidal Voronoi tessellations*, SIAM J. Numer. Anal., 44 (2006), pp. 102–119.
[6] Q. Du, V. Faber, and M. Gunzburger, *Centroidal Voronoi tessellations: Applications and algorithms*, SIAM Rev., 41 (1999), pp. 637–676.
[7] Q. Du, M. D. Gunzburger, and L. Ju, *Constrained centroidal Voronoi tessellations for surfaces*, SIAM J. Sci. Comput., 24 (2003), pp. 1488–1506.
[8] Q. Du and D. Wang, *Anisotropic centroidal Voronoi tessellations and their applications*, SIAM. J. Sci. Comput., 26 (2005), pp. 737–761.
[9] P. Fleischer, *Sufficient conditions for achieving minimum distortion in a quantizer*, in IEEE International Convention Record, I, IEEE, Washington, D.C., 1964, pp. 104–111.
[10] A. Gersho and R. Gray, *Vector Quantization and Signal Compression*, Kluwer, Boston, 1992.
[11] R. Gray, J. Kieffer, and Y. Linde, *Locally optimal block quantizer design*, Inform. and Control, 45 (1980), pp. 178–198.
[12] R. Gray and D. Neuhoff, *Quantization*, IEEE Trans. Inform. Theory, 44 (1998), pp. 2325–2383.
[13] P. M. Gruber, *Optimum quantization and its applications*, Adv. Math., 186 (2004), pp. 456–497.
[14] L. Ju, Q. Du, and M. Gunzburger, *Probabilistic methods for centroidal Voronoi tessellations and their parallel implementations*, Parallel Comput., 28 (2002), pp. 1477–1500.
[15] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, *An efficient k-means clustering algorithm: Analysis and implementation*, IEEE Trans. Pattern Anal. Machine Intell., 24 (2002), pp. 881–892.
[16] J. Kieffer, *Uniqueness of locally optimal quantizer for log-concave density and convex error function*, IEEE Trans. Inform. Theory, 29 (1983), pp. 42–47.
[17] Y. Linde, A. Buzo, and R. Gray, *An algorithm for vector quantizer design*, IEEE Trans. Commun., 28 (1980), pp. 84–95.
[18] S. Lloyd, *Least squares quantization in PCM*, IEEE Trans. Inform. Theory, 28 (1982), pp. 129–137.
[19] F. Lu and G. Wise, *A further investigation of the Lloyd-Max algorithm for quantizer design*, in Proceedings of the 21st Annual Allerton Conference on Communication, Control, and Computing, University of Illinois, Monticello, IL, 1983, pp. 481–490.
[20] D. G. Luenberger, *Linear and Nonlinear Programming*, Addison-Wesley, Reading, MA, 1984.
[21] D. Newman, *The hexagon theorem*, IEEE Trans. Inform. Theory, 28 (1982), pp. 137–139.
[22] J. Sabin and R. Gray, *Global convergence and empirical consistency of the generalized Lloyd algorithm*, IEEE Trans. Inform. Theory, 32 (1986), pp. 148–155.
[23] S. Z. Selim and M. A. Ismail, *K-means-type algorithms: A generalized convergence theorem and characterization of local optimality*, IEEE Trans. Pattern Anal. Machine Intell., 6 (1984), pp. 81–87.
[24] A. V. Trushkin, *Sufficient conditions for uniqueness of a locally optimal quantizer for a class of convex error weighting functions*, IEEE Trans. Inform. Theory, 28 (1982), pp. 187–198.
[25] X. Wu, *On convergence of Lloyd's method* I, IEEE Trans. Inform. Theory, 38 (1992), pp. 171–174.

SIAM J. NUMER. ANAL.
Vol. 46, No. 3, pp. 1442–1453
© 2008 Society for Industrial and Applied Mathematics

ERROR ESTIMATES FOR THE RAVIART–THOMAS INTERPOLATION UNDER THE MAXIMUM ANGLE CONDITION*

RICARDO G. DURÁN† AND ARIEL L. LOMBARDI†

Abstract. The classical error analysis for the Raviart–Thomas interpolation on triangular elements requires the so-called regularity of the elements, or equivalently, the minimum angle condition. However, in the lowest order case, optimal order error estimates have been obtained in [G. Acosta and R. G. Durán, *SIAM J. Numer. Anal.*, 37 (2000), pp. 18–36] replacing the regularity hypothesis by the maximum angle condition, which was known to be sufficient to prove estimates for the standard Lagrange interpolation. In this paper we prove error estimates on triangular elements for the Raviart–Thomas interpolation of any order under the maximum angle condition. Also, we show how our arguments can be extended to the three-dimensional case to obtain error estimates for tetrahedral elements under the regular vertex property introduced in [G. Acosta and R. G. Durán, *SIAM J. Numer. Anal.*, 37 (2000), pp. 18–36].

Key words. mixed finite elements, Raviart–Thomas, anisotropic finite elements

AMS subject classification. 65N30

DOI. 10.1137/060665312

1. Introduction.

The classical error analysis for finite element approximations is based on the so-called regularity assumption on the elements. In other words, the constants in the error estimates obtained depend on the ratio between outer and inner diameter of the elements and blow up when this ratio goes to infinity (see, for example, [4, 5]).

However, it is well known that the regularity assumption can be relaxed for standard finite element approximations. For example, in the two-dimensional (2d) case, optimal order error estimates have been proved for triangular elements under the weaker maximum angle condition (i.e., angles bounded away from π). This condition allows the use of the so-called anisotropic elements which is of interest in several applications, for example in problems with boundary or interior layers.

Error estimates under the maximum angle condition were first obtained in [3, 9]. After these pioneering works many papers have considered different generalizations of their results (see, for example, [2] and the references therein).

The usual error analysis for mixed finite element methods also makes use of the regularity assumption [12, 14]. In view of the results for the standard method mentioned above, it is a natural question whether the regularity hypothesis on the elements can be relaxed in this case also. A positive answer for the Raviart–Thomas space of lowest order $\mathcal{R}T_0$ is given in [1], where an optimal error estimate is proved under the maximum angle condition. However, it is not straightforward to extend the arguments given in that paper to higher order approximations. In [7] it is proved that the maximum angle condition is also sufficient to obtain optimal error estimates for the Raviart–Thomas space $\mathcal{R}T_1$.

*Received by the editors July 18, 2006; accepted for publication (in revised form) September 27, 2007; published electronically April 4, 2008. This work was supported by ANPCyT under grant PICT 03-13719, Fundación Antorchas, by Universidad de Buenos Aires under grant X052, and by CONICET under grant PIP 5478. The authors are members of CONICET, Argentina.
http://www.siam.org/journals/sinum/46-3/66531.html

†Departamento de Matemática, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, 1428 Buenos Aires, Argentina (rduran@dm.uba.ar, aldoc7@dm.uba.ar).

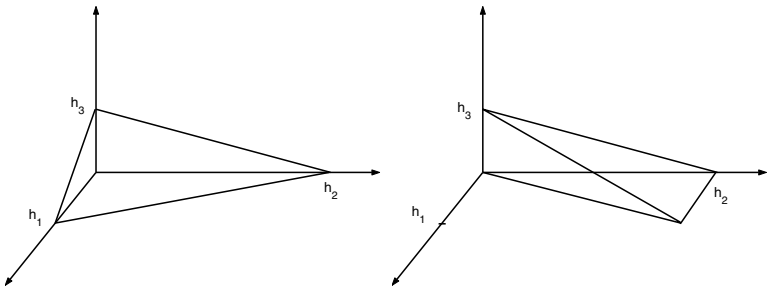


FIG. 1. *Two families of reference elements in the* 3*d case.*

The goal of this paper is to prove that the maximum angle condition is also sufficient to obtain optimal error estimates for the approximation using the Raviart–Thomas space $\mathcal{R}T_k$ for arbitrary k.

The three-dimensional (3d) case presents some important differences with the 2d case. Results for the standard finite element approximations in three dimensions were obtained in several papers (see, for example, [2, 6, 8, 10, 13]). In [10] Krízek proved optimal order error estimates for the Lagrange interpolation on a tetrahedron K for smooth functions, namely $u \in W^{2,\infty}$, provided the angles between faces and the angles in the faces are bounded away from π.

The Krízek condition seems to be a natural extension of the 2d maximum angle condition. However, there is another possible extension: the regular vertex property introduced in [1]. Roughly speaking, a family of tetrahedral elements satisfies this condition if for each element there is at least one vertex such that the unit vectors in the direction of the edges sharing that vertex are "uniformly" linearly independent, in the sense that the volume determined by them is uniformly bounded away from zero.

It is easy to see that in the 2d case the maximum angle condition and the regular vertex property are equivalent. However, in three dimensions this is not the case. Indeed, the Krízek condition allows for more general elements. To understand better these two possible extensions of the maximum angle condition, consider the two families of elements given in Figure 1, where h_1, h_2, and h_3 are arbitrary positive numbers.

It is easy to see that both families satisfy the maximum angle condition, but the second family does not satisfy the regular vertex property. Moreover, it was proved in [1] that the family of all elements satisfying the Krízek condition with a constant $\psi < \pi$ (i.e., angles between faces and angles in the faces are less than or equal to ψ) can be obtained by transforming both families in the figure by affine transformations with bounded condition number. On the other hand, the family of all elements satisfying the regular vertex property with a given constant (see section 3 for the formal definition of this condition) is obtained by transforming in the same way only the first family in the figure.

In [1], optimal error estimates under the Krízek condition were obtained for the $\mathcal{R}T_0$ spaces in three dimensions. In this paper we prove optimal error estimates for the general $\mathcal{R}T_k$ spaces under the regular vertex property. The more general case of elements satisfying the Krízek condition requires a different argument and will be the subject of future research.

To simplify the exposition we present first the proofs in the 2d case in section 2. Then, in section 3 we show how the arguments can be easily generalized to the 3d case for elements satisfying the regular vertex property.

2. Error estimates in the 2d case. We use the standard notation $\mathcal{P}_k$ for the space of polynomials of degree less than or equal to k. Then the local Raviart–Thomas space of order $k \geq 0$ on a triangle T [12] is defined by

$$\mathcal{R}T_k(T) = \mathcal{P}_k^2(T) + (x, y)\mathcal{P}_k(T). \tag{2.1}$$

It is known that there exists an operator

$$\Pi_k : H^1(T)^2 \longrightarrow \mathcal{R}T_k(T)$$

such that

$$\int_\ell \Pi_k \mathbf{v} \cdot \mathbf{n} p_k = \int_\ell \mathbf{v} \cdot \mathbf{n} p_k \quad \forall p_k \in \mathcal{P}_k(\ell) \quad \forall \ell \,\text{side of } T \tag{2.2}$$

and, if $k \geq 1$,

$$\int_T \Pi_k \mathbf{v} \cdot \mathbf{p}_{k-1} = \int_T \mathbf{v} \cdot \mathbf{p}_{k-1} \quad \forall \mathbf{p}_{k-1} \in \mathcal{P}_{k-1}^2(T). \tag{2.3}$$

Introducing the L^2 orthogonal projection

$$P_k \ : L^2(T) \longrightarrow \mathcal{P}_k(T),$$

it is not difficult to check that Π_k and P_k satisfy the following commutative diagram:

$$\begin{array}{lll} H^1(T)^2 & \xrightarrow{\text{div}} & L^2(T) \\ \Pi_k \big\downarrow & & \big\downarrow P_k \\ \mathcal{R}T_k(T) & \xrightarrow{\text{div}} & \mathcal{P}_k(T) \quad \longrightarrow 0 \end{array} \tag{2.4}$$

where the right bottom arrow in this diagram indicates that the divergence operator is surjective.

In the rest of the paper the letter C will denote a generic constant independent of the element T and of the functions involved in the estimates.

The error estimates for Π_k will be a consequence of the generalized Poincaré inequality given in the next lemma.

LEMMA 2.1. *Let T be a triangle and ξ_1 and ξ_2 unit vectors in the directions of two sides of T that have lengths h_1 and h_2. Assume that for one side ℓ of T and $f \in H^{k+1}(T)$ we have*

$$\int_\ell f\, q = 0 \qquad \forall q \in \mathcal{P}_k(\ell) \tag{2.5}$$

and, if $k \geq 1$,

$$\int_T f\, p = 0 \qquad \forall p \in \mathcal{P}_{k-1}(T). \tag{2.6}$$

Then, there exists a constant C independent of T and f such that

$$\|f\|_{L^2(T)} \leq C \sum_{i+j=k+1} h_1^i h_2^j \left\| \frac{\partial^{k+1} f}{\partial \xi_1^i \partial \xi_2^j} \right\|_{L^2(T)}. \tag{2.7}$$

Proof. First we prove the result for the reference element $\widehat{T}$ which has vertices at $(0,0)$, $(1,0)$, and $(0,1)$. Namely, we will show that if $\hat{f} \in H^{k+1}(\widehat{T})$ satisfies

$$\int_{\hat{\ell}} \hat{f}\, q = 0 \qquad \forall q \in \mathcal{P}_k(\hat{\ell}) \tag{2.8}$$

and, if $k \geq 1$,

$$\int_{T} \hat{f}\, p = 0 \qquad \forall p \in \mathcal{P}_{k-1}(\widehat{T}), \tag{2.9}$$

where $\hat{\ell}$ is one of the sides of $\widehat{T}$, then

$$\|\hat{f}\|_{L^2(\widehat{T})} \leq C \sum_{i+j=k+1} \left\| \frac{\partial^{k+1} \hat{f}}{\partial \hat{x}^i \partial \hat{y}^j} \right\|_{L^2(\widehat{T})}. \tag{2.10}$$

Let $\mathcal{S}_0 = \mathcal{P}_0(\widehat{T})$ and, for $k \geq 1$, let $\mathcal{S}_k$ be the orthogonal complement of $\mathcal{P}_{k-1}(\widehat{T})$ in $\mathcal{P}_k(\widehat{T})$. Then, $\|\,.\,\|_{L^2(\hat{\ell})}$ is a norm on $\mathcal{S}_k$. For $k = 0$ this is trivial and to see this for $k \geq 1$ it is enough to check that if $p \in \mathcal{S}_k$ satisfies $\|p\|_{L^2(\hat{\ell})} = 0$, then $p = 0$. But, $\|p\|_{L^2(\hat{\ell})} = 0$ implies that p vanishes on $\hat{\ell}$ and therefore $p = \lambda_1 q$ with $q \in \mathcal{P}_{k-1}(\widehat{T})$, where $\lambda_1 \in \mathcal{P}_1$ is the function such that $\lambda_1 = 0$ is the equation of the line containing the side $\hat{\ell}$ (i.e., the so-called barycentric coordinate). Now, since $p \in \mathcal{S}_k$ we have

$$\int_{\widehat{T}} \lambda_1\, q^2 = 0$$

and, since λ_1 does not change sign in $\hat{T}$, it follows that $q = 0$ and so $p = 0$ as we wanted to see.

In view of (2.9), the L^2 orthogonal projection of $\hat{f}$ on $\mathcal{P}_k(\widehat{T})$, namely $\widehat{P}_k \hat{f}$, belongs to $\mathcal{S}_k$. Then, since $\mathcal{S}_k$ is a finite-dimensional space, there exists a constant C depending only on k and $\widehat{T}$ such that

$$\|\widehat{P}_k \hat{f}\|_{L^2(\widehat{T})} \leq C \|\widehat{P}_k \hat{f}\|_{L^2(\hat{\ell})}. \tag{2.11}$$

On the other hand, using (2.8) we have

$$\|\hat{P}_k \hat{f}\|^2_{L^2(\hat{\ell})} = \int_{\hat{\ell}} \widehat{P}_k \hat{f}\, (\widehat{P}_k \hat{f} - \hat{f}) \leq \|\widehat{P}_k \hat{f}\|_{L^2(\hat{\ell})} \|\widehat{P}_k \hat{f} - \hat{f}\|_{L^2(\hat{\ell})}$$

which, together with (2.11), gives

$$\|\widehat{P}_k \hat{f}\|_{L^2(\widehat{T})} \leq C \|\widehat{P}_k \hat{f} - \hat{f}\|_{L^2(\hat{\ell})}$$

and then, using a standard trace theorem to bound the right-hand side and known error estimates for the L^2 projection, we obtain

$$\|\widehat{P}_k \hat{f}\|_{L^2(\widehat{T})} \leq C \|\widehat{P}_k \hat{f} - \hat{f}\|_{H^1(\widehat{T})} \leq C \sum_{i+j=k+1} \left\| \frac{\partial^{k+1} \hat{f}}{\partial \hat{x}^i \partial \hat{y}^j} \right\|_{L^2(\widehat{T})}$$

with another constant C depending only on k and $\widehat{T}$. Therefore, (2.10) follows using the triangular inequality.

Assume, without loss of generality, that the vertex shared by the sides of T in the directions ξ_1 and ξ_2 is at the origin. Then, T is the image of $\widehat{T}$ by the linear map

$$\begin{pmatrix} x \\ y \end{pmatrix} = B \begin{pmatrix} \hat{x} \\ \hat{y} \end{pmatrix},$$

where the columns of B are given by $h_1\xi_1$ and $h_2\xi_2$. Therefore, if $\hat{f}(\hat{x},\hat{y}) = f(x,y)$, then

$$\frac{\partial^{k+1}\hat{f}}{\partial\hat{x}^i\partial\hat{y}^j} = h_1^i h_2^j \frac{\partial^{k+1} f}{\partial\xi_1^i\partial\xi_2^j}.$$

Then, (2.7) follows from (2.10) by changing variables. ∎

Our next goal is to prove a result for the L^2 orthogonal projection which will be needed in our main theorem. For clarity we divide the proof in the following two lemmas.

LEMMA 2.2. *Let $\widehat{T}$ be the triangle with vertices at $(0,0)$, $(1,0)$, and $(0,1)$. Given (i,j) such that $i+j=k$ we define*

$$q_{ij} := \frac{\partial^k}{\partial x^i \partial y^j}\big(x^i y^j (1-x-y)^k\big).$$

Then, for any $f \in H^k(\widehat{T})$,

$$\int_{\widehat{T}} q_{ij}\, f = (-1)^k \int_{\widehat{T}} \big(x^i y^j (1-x-y)^k\big) \frac{\partial^k f}{\partial x^i \partial y^j}. \tag{2.12}$$

Proof. We want to prove that

$$\int_{\widehat{T}} \frac{\partial^k}{\partial x^i \partial y^j}\big(x^i y^j\ (1-x-y)^k\big)\, f = (-1)^k \int_{\widehat{T}} \big(x^i y^j (1-x-y)^k\big) \frac{\partial^k f}{\partial x^i \partial y^j}$$

so, it is enough to see that the boundary terms arising in the integration by parts vanish. For example, when we integrate by parts in the x variable the boundary terms will be of the form

$$\int_{\partial\widehat{T}} \frac{\partial^{l+j}}{\partial x^l \partial y^j}\big(x^i y^j (1-x-y)^k\big)\, \frac{\partial^{i-1-l} f}{\partial x^{i-1-l}}\, n_1,$$

where n_1 is the first component of the unit normal at the boundary and $l = 0,\dots,i-1$. Since $n_1 = 0$ on the side contained in the line $\{y=0\}$, it is enough to see that

$$\frac{\partial^{l+j}}{\partial x^l \partial y^j}\big(x^i y^j (1-x-y)^k\big) = 0 \tag{2.13}$$

on the other two sides of $\widehat{T}$. This derivative can be written as a sum of terms which, up to a multiplicative constant, are of the form

$$x^{i-m} y^{j-n} (1-x-y)^{k+m+n-l-j}$$

with $m = 0,\dots,l$ and $n = 0,\dots,j$. But, since $l < i$, we have $m < i$ and $l+j < i+j = k$ and so in all of these terms the exponents of x and $(1-x-y)$ are positive. Therefore,

(2.13) holds on the sides of $\widehat{T}$ contained in $\{x = 0\}$ and $\{1 - x - y = 0\}$ as we wanted to show. ∎

LEMMA 2.3. *If T is the right triangle with vertices at $(0,0)$, $(h_1,0)$, and $(0,h_2)$, then, for any $f \in H^k(T)$ and any i,j such that $i+j=k$,*

$$\left\| \frac{\partial^k P_k f}{\partial x^i \partial y^j} \right\|_{L^2(T)} \leq C \left\| \frac{\partial^k f}{\partial x^i \partial y^j} \right\|_{L^2(T)}$$

with a constant C independent of the right triangle T.

Proof. We prove the estimate for the reference triangle $\widehat{T}$ with vertices at $(0,0)$, $(1,0)$, and $(0,1)$. Then, the result follows by making the change of variables $(x,y) = (h_1\hat{x}, h_2\hat{y})$.

Since the result is trivial for $k = 0$ we can assume that $k \geq 1$. Fix (i,j) such that $i + j = k$ and let $\{p_1, \ldots, p_N\}$ be an orthonormal basis of the subspace $\mathcal{S} \subset \mathcal{P}_k(\widehat{T})$ expanded by

$$\{x^m y^n \,:\, m + n \leq k, (m,n) \neq (i,j)\}.$$

From the definition of q_{ij} given in the previous lemma it is easy to check that $q_{ij} \in \mathcal{P}_k$. Moreover, since $\frac{\partial^k p}{\partial x^i \partial y^j} = 0$ for all $p \in \mathcal{S}$, it follows from (2.12) that q_{ij} is orthogonal to $\mathcal{S}$.

Then, defining $p_{N+1} := q_{ij}/\|q_{ij}\|_{L^2(\widehat{T})}$, we have that $\{p_1, \ldots, p_{N+1}\}$ is an orthonormal basis of $\mathcal{P}_k(\widehat{T})$. Then

$$P_k f = \sum_{s=1}^{N+1} \left(\int_{\widehat{T}} f p_s \right) p_s$$

and, since $\frac{\partial^k p_s}{\partial x^i \partial y^j} = 0$ for $s = 1, \ldots, N$, we have

$$\frac{\partial^k P_k f}{\partial x^i \partial y^j} = \left(\int_{\widehat{T}} f p_{N+1} \right) \frac{\partial^k p_{N+1}}{\partial x^i \partial y^j}, \tag{2.14}$$

but it follows from (2.12) that

$$\int_{\widehat{T}} f p_{N+1} = \frac{(-1)^k}{\|q_{ij}\|_{L^2(\widehat{T})}} \int_{\widehat{T}} \left(x^i y^j (1-x-y)^k \right) \frac{\partial^k f}{\partial x^i \partial y^j}$$

which, together with (2.14) and an application of the Schwarz inequality, gives

$$\left\| \frac{\partial^k P_k f}{\partial x^i \partial y^j} \right\|_{L^2(\widehat{T})} \leq \widehat{C} \left\| \frac{\partial^k f}{\partial x^i \partial y^j} \right\|_{L^2(\widehat{T})}$$

as we wanted to prove. ∎

To end with the preliminary results we give in the next lemma a relation between derivatives of a function in $\mathcal{RT}_k(T)$ and derivatives of its divergence.

LEMMA 2.4. *If T is a triangle and $\mathbf{u} \in \mathcal{RT}_k(T)$, then*

$$\frac{\partial^{k+1} \mathbf{u}}{\partial x^{k+1}} = \left(\frac{k+1}{k+2} \frac{\partial^k (\operatorname{div} \mathbf{u})}{\partial x^k}, 0 \right), \qquad \frac{\partial^{k+1} \mathbf{u}}{\partial y^{k+1}} = \left(0, \frac{k+1}{k+2} \frac{\partial^k (\operatorname{div} \mathbf{u})}{\partial y^k} \right)$$

and for $i+j=k+1$, with $i>0$ and $j>0$,

$$\frac{\partial^{k+1}\mathbf{u}}{\partial x^i\partial y^j}=\left(\frac{i}{2+k}\frac{\partial^k \mathrm{div}\,\mathbf{u}}{\partial x^{i-1}\partial y^j}\,,\frac{j}{2+k}\frac{\partial^k \mathrm{div}\,\mathbf{u}}{\partial x^i\partial y^{j-1}}\right).$$

Proof. Any function $\mathbf{u}\in\mathcal{R}T_k(T)$ can be written as $\mathbf{u}=\mathbf{p}+(xq,yq)$ with $\mathbf{p}\in\mathcal{P}_k^2$ and $q\in\mathcal{P}_k$. An elementary computation shows that

$$\frac{\partial^{k+1}\mathbf{u}}{\partial x^{k+1}}=\Big((k+1)\frac{\partial^k q}{\partial x^k}\,,\,0\Big)\quad,\qquad \frac{\partial^{k+1}\mathbf{u}}{\partial y^{k+1}}=\Big(0\,,\,(k+1)\frac{\partial^k q}{\partial y^k}\Big)$$

and, for any positive i,j such that $i+j=k+1$,

$$\frac{\partial^{k+1}\mathbf{u}}{\partial x^i\partial y^j}=\Big(i\frac{\partial^k q}{\partial x^{i-1}\partial y^j}\,,j\frac{\partial^k q}{\partial x^i\partial y^{j-1}}\Big).$$

On the other hand, for $l+m=k$

$$\frac{\partial^k(\mathrm{div}\,\mathbf{u})}{\partial x^l\partial y^m}=(k+2)\frac{\partial^k q}{\partial x^l\partial y^m}\,.$$

Then the lemma is easily obtained. ∎

Now we are ready to prove our main result. Given a triangle T we call h_T the longest side of T. We will also use the notation

$$D^m f=\sum_{i+j=m}\left|\frac{\partial^m f}{\partial x^i\partial y^j}\right|$$

and a tilde over operators will indicate that the derivatives are taken with respect to $(\tilde{x},\tilde{y})$.

THEOREM 2.5. *Let T be a triangle with maximum angle α. Let h_1 and h_2 be the lengths of the sides adjacent to α and let ξ_1 and ξ_2 be unit vectors in the directions of those sides. Then, for any function $\mathbf{v}\in H^{k+1}(T)^2$,*

(2.15)
$$\|\mathbf{v}-\Pi_k\mathbf{v}\|_{L^2(T)}\le\frac{C}{\sin\alpha}\left\{\sum_{i+j=k+1}h_1^ih_2^j\left\|\frac{\partial^{k+1}\mathbf{v}}{\partial\xi_1^i\partial\xi_2^j}\right\|_{L^2(T)}+h_T^{k+1}\|D^k\mathrm{div}\,\mathbf{v}\|_{L^2(T)}\right\}$$

with a constant C independent of T.

Proof. In view of (2.2) and (2.3), for any side ℓ_s of T the function $(\mathbf{v}-\Pi_k\mathbf{v})\cdot\mathbf{n}_s$, where $\mathbf{n}_s$ denotes the unit exterior normal on ℓ_s, satisfies conditions (2.6) and (2.5) of Lemma 2.1, and therefore

$$\|(\mathbf{v}-\Pi_k\mathbf{v})\cdot\mathbf{n}_s\|_{L^2(T)}\le C\sum_{i+j=k+1}h_1^ih_2^j\left\|\frac{\partial^{k+1}\left[(\mathbf{v}-\Pi_k\mathbf{v})\cdot\mathbf{n}_s\right]}{\partial\xi_1^i\partial\xi_2^j}\right\|_{L^2(T)}.$$

But, choosing ℓ_1 and ℓ_2 as the sides with directions ξ_1 and ξ_2, it is not difficult to see that

$$\|\mathbf{v}-\Pi_k\mathbf{v}\|_{L^2(T)}\le\frac{C}{\sin\alpha}\left\{\|(\mathbf{v}-\Pi_k\mathbf{v})\cdot\mathbf{n}_1\|_{L^2(T)}+\|(\mathbf{v}-\Pi_k\mathbf{v})\cdot\mathbf{n}_2\|_{L^2(T)}\right\}$$

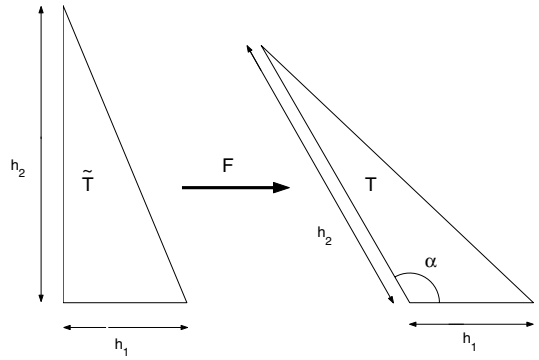


FIG. 2. *The map F applying $\widetilde{T}$ on T.*

and then,

$$\|\mathbf{v}-\Pi_k\mathbf{v}\|_{L^2(T)} \le \frac{C}{\sin\alpha}\sum_{i+j=k+1} h_1^i h_2^j \left\| \frac{\partial^{k+1}(\mathbf{v}-\Pi_k\mathbf{v})}{\partial\xi_1^i\partial\xi_2^j}\right\|_{L^2(T)}. \tag{2.16}$$

Then (2.15) is obtained from the inequality

$$\sum_{i+j=k+1}\left\| \frac{\partial^{k+1}\Pi_k\mathbf{v}}{\partial\xi_1^i\partial\xi_2^j}\right\|_{L^2(T)} \le C\left\|D^k\mathrm{div}\,\mathbf{v}\right\|_{L^2(T)}, \tag{2.17}$$

that we will prove in what follows.

Without loss of generality, we can assume that T has vertices at $(0,0)$, $(h_1,0)$, and (a,b) and that the vertex corresponding to the maximum angle is at (0,0). Then $\xi_1=(1,0)$ and $\xi_2=(\frac{a}{h_2},\frac{b}{h_2})$.

Consider the right triangle $\widetilde{T}$ with vertices at $(0,0)$, $(h_1,0)$, and $(0,h_2)$. Then, the linear transformation $(x,y)=F(\tilde{x},\tilde{y})$ defined as

$$\begin{pmatrix} x\\ y\end{pmatrix} = B\begin{pmatrix}\tilde{x}\\ \tilde{y}\end{pmatrix}$$

with

$$B=\begin{pmatrix} 1 & \frac{a}{h_2}\\ 0 & \frac{b}{h_2}\end{pmatrix}$$

maps $\widetilde{T}$ onto T (see Figure 2). We have

$$\|B\|_2 \le \sqrt{2}, \tag{2.18}$$

where $\|\,.\,\|_2$ denotes the matrix norm associated with the euclidean norm.

Now, given $\mathbf{v}\in H^{k+1}(T)^2$, we define the corresponding function $\tilde{\mathbf{v}}\in H^{k+1}(\widetilde{T})^2$ via the Piola transform, that is,

$$\mathbf{v}(x,y)=\frac{1}{|\det B|}B\,\tilde{\mathbf{v}}(\tilde{x},\tilde{y}),\quad (x,y)\in T.$$

Then, we have

$$\operatorname{div} \mathbf{v} = \frac{1}{|\det B|} \widetilde{\operatorname{div}}\, \tilde{\mathbf{v}}. \tag{2.19}$$

Denoting by $\widetilde{\Pi}_k$ the Raviart–Thomas interpolation on $\widetilde{T}$, it is proved in [12, page 303] that

$$\Pi_k \mathbf{v}(x,y) = \frac{1}{|\det B|} B\, \widetilde{\Pi}_k \tilde{\mathbf{v}}(\tilde{x},\tilde{y}), \quad (x,y) \in T.$$

Then, we have

$$\frac{\partial^{k+1} \Pi_k \mathbf{v}}{\partial \xi_1^i \partial \xi_2^j}(x,y) = \frac{1}{|\det B|} B\, \frac{\partial^{k+1} \widetilde{\Pi}_k \tilde{\mathbf{v}}}{\partial \tilde{x}^i \partial \tilde{y}^j}(\tilde{x},\tilde{y}),$$

and, therefore,

$$\left\| \frac{\partial^{k+1} \Pi_k \mathbf{v}}{\partial \xi_1^i \partial \xi_2^j} \right\|_{L^2(T)} \leq \frac{\|B\|_2}{|\det B|^{\frac{1}{2}}} \left\| \frac{\partial^{k+1} \widetilde{\Pi}_k \tilde{\mathbf{v}}}{\partial \tilde{x}^i \partial \tilde{y}^j} \right\|_{L^2(\tilde{T})}.$$

Now, since $\widetilde{\Pi}_k \tilde{\mathbf{v}} \in \mathcal{R}T_k(\widetilde{T})$, using Lemma 2.4, the commutative diagram property (2.4) for $\widetilde{\Pi}_k$, Lemma 2.3 in $\widetilde{T}$, and (2.19), we have

$$\begin{aligned}
\sum_{i+j=k+1} \left\| \frac{\partial^{k+1} \Pi_k \mathbf{v}}{\partial \xi_1^i \partial \xi_2^j} \right\|_{L^2(T)} &\leq \frac{\|B\|_2}{|\det B|^{\frac{1}{2}}} \sum_{i+j=k+1} \left\| \frac{\partial^{k+1} \widetilde{\Pi}_k \tilde{\mathbf{v}}}{\partial \tilde{x}^i \partial \tilde{y}^j} \right\|_{L^2(\tilde{T})} \\
&\leq C \frac{\|B\|_2}{|\det B|^{\frac{1}{2}}} \left\| \widetilde{D}^k \widetilde{\operatorname{div}}\, \widetilde{\Pi}_k \tilde{\mathbf{v}} \right\|_{L^2(\tilde{T})} \\
&\leq C \frac{\|B\|_2}{|\det B|^{\frac{1}{2}}} \left\| \widetilde{D}^k \widetilde{P}_k \widetilde{\operatorname{div}}\, \tilde{\mathbf{v}} \right\|_{L^2(\tilde{T})} \\
&\leq C \frac{\|B\|_2}{|\det B|^{\frac{1}{2}}} \left\| \widetilde{D}^k \widetilde{\operatorname{div}}\, \tilde{\mathbf{v}} \right\|_{L^2(\tilde{T})} \\
&\leq C \|B\|_2^{k+1} \left\| D^k \operatorname{div} \mathbf{v} \right\|_{L^2(T)},
\end{aligned}$$

then (2.17) follows using (2.18). ∎

Remark 2.1. It is possible to prove error estimates for the Raviart–Thomas interpolation under the maximum angle condition without using the Piola transform. However, the dependence of the constant on the maximum angle of T obtained in this way is worse than that in Theorem 2.5.

Indeed, using the same change of variables from a right triangle $\widetilde{T}$ onto T as in Theorem 2.5, and the estimate given in Lemma 2.3 we can prove, for any $f \in H^k(T)$,

$$\|D^k P_k f\|_{L^2(T)} \leq \frac{C}{(\sin \alpha)^k} \|D^k f\|_{L^2(T)}, \tag{2.20}$$

where α is the maximum angle of T.

Then, using Lemma 2.4, the commutative diagram property for Π_k (2.4), and (2.20) we have

$$\begin{aligned}\sum_{i+j=k+1}\left\|\frac{\partial^{k+1}\Pi_k\mathbf{v}}{\partial\xi_1^i\partial\xi_2^j}\right\|_{L^2(T)} &\le \|B\|_2^{k+1}\sum_{i+j=k+1}\left\|\frac{\partial^{k+1}\Pi_k\mathbf{v}}{\partial x^i\partial y^j}\right\|_{L^2(T)}\\ &\le C\|B\|_2^{k+1}\sum_{i+j=k+1}\left\|D^k\operatorname{div}\Pi_k\mathbf{v}\right\|_{L^2(T)}\\ &= C\|B\|_2^{k+1}\sum_{i+j=k+1}\left\|D^kP_k\operatorname{div}\mathbf{v}\right\|_{L^2(T)}\\ &\le \frac{C}{(\sin\alpha)^k}\|B\|_2^{k+1}\sum_{i+j=k+1}\left\|D^k\operatorname{div}\mathbf{v}\right\|_{L^2(T)}.\end{aligned}$$

This inequality, together with (2.16) and (2.18), gives

$$\|\mathbf{v}-\Pi_k\mathbf{v}\|_{L^2(T)} \le \frac{C}{\sin\alpha}\left\{\sum_{i+j=k+1}h_1^ih_2^j\left\|\frac{\partial^{k+1}\mathbf{v}}{\partial\xi_1^i\partial\xi_2^j}\right\|_{L^2(T)} + \frac{h_T^{k+1}}{(\sin\alpha)^k}\|D^k\operatorname{div}\mathbf{v}\|_{L^2(T)}\right\}.$$

3. The 3d case. The Raviart–Thomas spaces have been generalized to the 3d case by Nedelec [11]. In this section we briefly explain how to extend the results previously obtained to a class of tetrahedral elements satisfying a condition which is weaker than regularity.

Following [1], we say that a tetrahedron K satisfies the *regular vertex property* with a constant $\bar{c}>0$ if there exist a vertex $\mathbf{Z}$ of K, called the *regular vertex*, unit vectors ξ_i, and scalars h_i, $i=1,2,3$, such that K is the convex hull of $\{\mathbf{Z}\}\cup\{\mathbf{Z}+h_i\xi_i, i=1,2,3\}$, and the matrix B made up with ξ_i as its columns verifies $|\det B|\ge\bar{c}$.

Given a tetrahedron K the spaces introduced in [11] are given, for $k\ge 0$, by

$$\mathcal{R}T_k(K)=\mathcal{P}_k(K)^3+(x,y,z)\mathcal{P}_k(K).$$

Also in [11] it is proved that there exists an operator

$$\Pi_k : H^1(K)^3 \longrightarrow \mathcal{R}T_k(K)$$

such that

$$\int_F \Pi_k\mathbf{v}\cdot\mathbf{n}p_k=\int_F\mathbf{v}\cdot\mathbf{n}p_k \quad \forall p_k\in\mathcal{P}_k(F),\ \forall F \ \text{ face of } \ K$$

and, if $k\ge 1$,

$$\int_K \Pi_k\mathbf{v}\cdot\mathbf{p}_{k-1}=\int_K\mathbf{v}\cdot\mathbf{p}_{k-1}\quad\forall\mathbf{p}_{k-1}\in\mathcal{P}_{k-1}^3(K).$$

In what follows we state, for tetrahedral elements satisfying the regular vertex property, an error estimate for Π_k analogous to that given in Theorem 2.5. We will mention without proof the 3d versions of Lemmas 2.1 and 2.3 that can be obtained following the same arguments as in the 2d case.

We define the tetrahedron $\widetilde{K}$ with vertices at $(0,0,0)$, $(h_1,0,0)$, $(0,h_2,0)$, and $(0,0,h_3)$. We denote by $\mathbf{n}_1$, $\mathbf{n}_2$, and $\mathbf{n}_3$ the normals to the faces of $\widetilde{K}$ contained in

the planes $x = 0$, $y = 0$, and $z = 0$, respectively. For a tetrahedron K we call h_K its diameter.

THEOREM 3.1. *Let K be a tetrahedron satisfying the regular vertex property with a constant $\bar{c}$ and let h_i and ξ_i, $i = 1, 2, 3$, be the lengths and unit directions corresponding to the edges sharing the regular vertex $\mathbf{Z}$. Then, for any function $\mathbf{v} \in H^{k+1}(K)^3$,*

$$\|\mathbf{v} - \Pi_k \mathbf{v}\|_{L^2(K)} \leq C \left\{ \sum_{i+j+l=k+1} h_1^i h_2^j h_3^l \left\| \frac{\partial^{k+1} \mathbf{v}}{\partial \xi_1^i \partial \xi_2^j \partial \xi_3^l} \right\|_{L^2(K)} + h_K^{k+1} \|D^k \operatorname{div} \mathbf{v}\|_{L^2(T)} \right\}$$

with a constant C depending only on $\bar{c}$.

Proof. From [1, Lemma 5.2] there exists a constant C depending only on $\bar{c}$ such that

$$\|\mathbf{v} - \Pi_k \mathbf{v}\|_{L^2(K)} \leq C \sum_{s=1}^{3} \|(\mathbf{v} - \Pi_k \mathbf{v}) \cdot \mathbf{n}_s\|_{L^2(K)}.$$

We can apply to $(\mathbf{v} - \Pi_k \mathbf{v}) \cdot \mathbf{n}_s$ a straightforward generalization of Lemma 2.1 to obtain

$$\|\mathbf{v} - \Pi_k \mathbf{v}\|_{L^2(K)} \leq C \sum_{i+j+l=k+1} h_1^i h_2^j h_3^l \left\| \frac{\partial^{k+1} (\mathbf{v} - \Pi_k \mathbf{v})}{\partial \xi_1^i \partial \xi_2^j \partial \xi_3^l} \right\|_{L^2(K)}. \tag{3.1}$$

Hence, we have to prove

$$\sum_{i+j+l=k+1} \left\| \frac{\partial^{k+1} \Pi_k \mathbf{v}}{\partial \xi_1^i \partial \xi_2^j \partial \xi_3^l} \right\|_{L^2(K)} \leq C \left\| D^k \operatorname{div} \mathbf{v} \right\|_{L^2(K)}.$$

Without loss of generality, we can assume that K has vertices $(0, 0, 0)$, $(h_1, 0, 0)$, $(a_2, b_2, 0)$, and (a_3, b_3, c_3) and that it satisfies the definition of the regular vertex property with $\mathbf{Z} = (0, 0, 0)$, $\xi_1 = (1, 0, 0)$, $\xi_2 = \frac{1}{h_2}(a_2, b_2, 0)$, and $\xi_3 = \frac{1}{h_3}(a_3, b_3, c_3)$.

The linear transformation

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = B \begin{pmatrix} \tilde{x} \\ \tilde{y} \\ \tilde{z} \end{pmatrix}$$

with

$$B = \begin{pmatrix} 1 & \frac{a_2}{h_2} & \frac{a_3}{h_3} \\ 0 & \frac{b_2}{h_2} & \frac{b_3}{h_3} \\ 0 & 0 & \frac{c_3}{h_3} \end{pmatrix}$$

maps $\widetilde{K}$ onto K.

Moreover, it is easy to check that

$$\|B\|_2 \leq 3\sqrt{3}.$$

For $\mathbf{v} \in H^{k+1}(K)^3$ we can consider $\tilde{\mathbf{v}} \in H^{k+1}(\widetilde{K})^3$ defined by the Piola transform

$$\mathbf{v} = \frac{1}{|\det B|} B \, \tilde{\mathbf{v}}.$$

Then, as in the proof of Theorem 2.5, we have

$$\left\| \frac{\partial^{k+1}\Pi_k \mathbf{v}}{\partial\xi_1^i \partial\xi_2^j \partial\xi_3^l} \right\|_{L^2(K)} \leq \frac{\|B\|_2}{|\det B|^{\frac{1}{2}}} \left\| \frac{\partial^{k+1}\widetilde{\Pi}_k \tilde{\mathbf{v}}}{\partial\tilde{x}^i \partial\tilde{y}^j \partial\tilde{z}^l} \right\|_{L^2(\widetilde{K})}.$$

But, as in the 2d case, for any tetrahedral element T we can write the derivatives of order $k+1$ of $\mathbf{u} \in \mathcal{R}T_k(T)$ in terms of the derivatives of its divergence, namely,

$$\begin{aligned}
\frac{\partial^{k+1}\mathbf{u}}{\partial x^{k+1}} &= \left(\frac{k+1}{k+3} \frac{\partial^k(\operatorname{div}\mathbf{u})}{\partial x^k}, 0, 0 \right), \\
\frac{\partial^{k+1}\mathbf{u}}{\partial y^{k+1}} &= \left(0, \frac{k+1}{k+3} \frac{\partial^k(\operatorname{div}\mathbf{u})}{\partial y^k}, 0 \right), \\
\frac{\partial^{k+1}\mathbf{u}}{\partial z^{k+1}} &= \left(0, 0, \frac{k+1}{k+3} \frac{\partial^k(\operatorname{div}\mathbf{u})}{\partial z^k} \right)
\end{aligned}$$

and for $i+j+l = k+1$, with $i > 0$, $j > 0$, and $l > 0$,

$$\frac{\partial^{k+1}\mathbf{u}}{\partial x^i \partial y^j \partial z^l} = \left(\frac{i}{k+3} \frac{\partial^k \operatorname{div}\mathbf{u}}{\partial x^{i-1} \partial y^j \partial z^l}, \frac{j}{k+3} \frac{\partial^k \operatorname{div}\mathbf{u}}{\partial x^i \partial y^{j-1} \partial z^l}, \frac{l}{k+3} \frac{\partial^k \operatorname{div}\mathbf{u}}{\partial x^i \partial y^j \partial z^{l-1}} \right).$$

We conclude, as in the last part of the proof of Theorem 2.5, by using the commutative diagram property of $\widetilde{\Pi}_k$ and by observing that Lemma 2.3 has a straightforward extension to the element $\widetilde{K}$. ∎

REFERENCES

[1] G. Acosta and R. G. Durán, *The maximum angle condition for mixed and nonconforming elements: Application to the Stokes equations*, SIAM J. Numer. Anal., 37 (2000), pp. 18–36.

[2] T. Apel, *Anisotropic Finite Elements: Local Estimates and Applications*, Adv. Numer. Math., Teubner, Stuttgart, 1999.

[3] I. Babuska and A. K. Aziz, *On the angle condition in the finite element method*, SIAM J. Numer. Anal., 13 (1976), pp. 214–226.

[4] S. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, Springer, New York, 1994.

[5] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North–Holland, Amsterdam, 1978.

[6] R. G. Durán, *Error estimates for* 3*-d narrow finite elements*, Math. Comp. 68 (1999), pp. 187–199.

[7] R. G. Durán, *Error estimates for anisotropic finite elements and applications*, in Proceedings of the International Congress of Mathematicians, 2006.

[8] R. G. Durán and A. L. Lombardi, *Error estimates on anisotropic $\mathcal{Q}_1$ elements for functions in weighted Sobolev spaces*, Math. Comp., 74 (2005), pp. 1679–1706.

[9] P. Jamet, *Estimations d'erreur pour des éléments finis droits presque dégénérés*, RAIRO Anal. Numér., 10 (1976), pp. 46–71.

[10] M. Krízek, *On the maximum angle condition for linear tetrahedral elements*, SIAM J. Numer. Anal., 29 (1992), pp. 513–520.

[11] J. C. Nedelec, *Mixed finite elements in* $\mathbb{R}^3$, Numer. Math., 35 (1980), pp. 315–341.

[12] P. A. Raviart and J. M. Thomas, *A mixed finite element method for second order elliptic problems*, in Mathematical Aspects of the Finite Element Method, Lecture Notes in Math. 606, I. Galligani and E. Magenes, eds., Springer, Berlin, 1977, pp. 292–315.

[13] N. Al Shenk, *Uniform error estimates for certain narrow Lagrange finite elements*, Math. Comp., 63 (1994), pp. 105–119.

[14] J. M. Thomas, *Sur l'Analyse Numérique des Méthodes d'Éléments Finis Hybrides et Mixtes*, Thèse d'Etat, Université Pierre et Marie Curie, Paris, 1977.

SIAM J. NUMER. ANAL.
Vol. 46, No. 3, pp. 1454–1482
© 2008 Society for Industrial and Applied Mathematics

FINITE-DIMENSIONAL APPROXIMATION SETTINGS FOR INFINITE-DIMENSIONAL MOORE–PENROSE INVERSES*

NAILIN DU†

Abstract. After a brief survey of basic results about finite-dimensional approximation settings (such as mathematical frameworks of various projection methods, including the least-squares method, the dual least-squares method, and the Galerkin method) for infinite-dimensional Moore–Penrose inverses, this paper proceeds to a detailed study from the following aspects: For projection methods, we investigate convergence and weak convergence of their approximation setting to develop a unified theory on projection methods for infinite-dimensional Moore–Penrose inverses; this investigation yields a fundamental convergence theorem (Theorem 2.2), from which the criterion for convergence, the criterion for weak convergence, and the generalized dual least-squares method are derived. We also derive general results on the least-squares method, by which two flaws in Groestch's results are corrected. For nonprojection methods (whose approximation setting is a more general framework), we investigate weak perfect convergence of their approximation setting and provide a necessary and sufficient condition of such convergence holding (Theorem 3.2). Several examples are proposed as counterexamples to illustrate the differences between some important concepts or as concrete algorithms to show how the present work can help to analyze their behavior.

Key words. convergence, weak convergence, weak perfect convergence, stability

AMS subject classifications. 47A58, 47A55, 65J20

DOI. 10.1137/060661120

1. Introduction.

Let X and Y be Hilbert spaces over the same scalars and $T \in \mathcal{B}(X,Y)$, i.e., $T : X \to Y$ is a bounded linear operator. If X_0 and Y_0 are closed subspaces of X and Y, respectively, P_{X_0} and Q_{Y_0} will denote the orthogonal projections from X onto X_0 and from Y onto Y_0, respectively. Let $\{X_n\}$ and $\{Y_n\}$ be sequences of closed subspaces of X and Y, respectively, which satisfy the completeness conditions, that is,

$$\operatorname*{s-lim}_{n\to\infty} P_n = I_X, \qquad \operatorname*{s-lim}_{n\to\infty} Q_n = I_Y,$$

where $P_n := P_{X_n}$ and $Q_n := Q_{Y_n}$; for each $n \in \mathbb{N}$, let $T_n \in \mathcal{B}(X_n, Y_n)$ with closed range $\mathcal{R}(T_n)$. Consider the operator equation of the first kind

$$Tu = b, \quad u \in X, \tag{1.1}$$

with its approximations

$$T_n u_n = b_n, \quad u_n \in X_n, \tag{1.2}$$

where $b \in Y$, $b_n \in Y_n$, and $b_n \to b$ in Y as $n \to \infty$. A practical case of interest is that $\dim X_n < \infty$ $(n \in \mathbb{N})$, $\dim \mathcal{R}(T) = \infty$, and X_n $(n \in \mathbb{N})$ are increasing and eventually dense in X, that is,

$$X_n \subseteq X_{n+1} \quad \text{and} \quad \overline{\bigcup_{n=1}^{\infty} X_n} = X.$$

*Received by the editors May 30, 2006; accepted for publication (in revised form) December 12, 2007; published electronically April 4, 2008. This research was partially supported by China National Natural Science Foundation grant 10571068.

http://www.siam.org/journals/sinum/46-3/66112.html

†School of Mathematics and Statistics, Wuhan University, Wuhan 430072, People's Republic of China (dunailin@yahoo.com.cn).

We are interested in the problem of approximating the best-approximate solution of (1.1), i.e., the least-squares solution of the minimal norm of (1.1), which is that vector $u_\infty \in X$ of smallest norm satisfying

$$\|Tu_\infty - b\| = \inf\{\|Tx - b\| \,|\, x \in X\}.$$

The vector u_∞ exists if and only if $Q_{\overline{\mathcal{R}(T)}} b \in \mathcal{R}(T)$, and u_∞ is given by $u_\infty = T^\dagger b$, where $T^\dagger$ is the Moore–Penrose generalized inverse of T. $T^\dagger$ is, in general, a closed linear operator with

$$\mathcal{D}\left(T^\dagger\right) = \mathcal{R}(T) + \mathcal{R}(T)^\perp, \quad \mathcal{N}\left(T^\dagger\right) = \mathcal{R}(T)^\perp, \quad \text{and} \quad \mathcal{R}\left(T^\dagger\right) = \mathcal{N}(T)^\perp.$$

Note that $\mathcal{D}\left(T^\dagger\right) = Y$ if and only if $\mathcal{R}(T)$ is closed. Here we remark that, since $T_n \in \mathcal{B}(X_n, Y_n)$ is with closed range $\mathcal{R}(T_n)$ in Y_n, the generalized inverse $T_n^\dagger$ of T_n is a bounded linear operator with

$$\mathcal{D}\left(T_n^\dagger\right) = Y_n, \quad \mathcal{N}\left(T_n^\dagger\right) = \mathcal{R}(T_n)^\perp \cap Y_n, \quad \text{and} \quad \mathcal{R}\left(T_n^\dagger\right) = \mathcal{N}(T_n)^\perp \cap X_n.$$

See [11, 22, 28] for comprehensive information about generalized inverses. For infinite-dimensional problems with the approximation setting (1.1) and (1.2), the questions of finite-dimensional approximations of the infinite-dimensional Moore–Penrose inverse $T^\dagger$ and their convergence naturally arise, so we formulate the following conditions:

- (H1) (*consistency*):

$$\lim_{n\to\infty} \|Q_n (T_n - T) P_n x\| = 0 \quad (\forall\, x \in X).$$

- (H1)$^{\mathrm{u}}$ (*uniform consistency*):

$$\lim_{n\to\infty} \|Q_n (T_n - T) P_n\| = 0.$$

- (H2) (*stability*): There is a constant $C > 0$ such that

$$\left\|T_n^\dagger\right\|_n := \sup\left\{\left\|T_n^\dagger y\right\| \,\middle|\, y \in Y_n, \|y\| = 1\right\} \leqslant C \quad (\forall\, n \in \mathbb{N}).$$

- (H3)$^{\mathrm{s}}$ (*convergence*):

$$\lim_{n\to\infty} T_n^\dagger Q_n y = T^\dagger y \quad \left(\forall\, y \in \mathcal{D}\left(T^\dagger\right)\right).$$

- (H3)$^{\mathrm{w}}$ (*weak convergence*):

$$\operatorname*{w\text{-}lim}_{n\to\infty} T_n^\dagger Q_n y = T^\dagger y \quad \left(\forall\, y \in \mathcal{D}\left(T^\dagger\right)\right).$$

- (H3)$^{\mathrm{s}}_{\mathrm{p}}$ (*perfect convergence*):

$$\mathcal{D}\left(T^\dagger\right) = Y \quad \text{and} \quad \operatorname*{s\text{-}lim}_{n\to\infty} T_n^\dagger Q_n = T^\dagger.$$

- (H3)$^{\mathrm{u}}_{\mathrm{p}}$ (*uniform perfect convergence*):

$$\mathcal{D}\left(T^\dagger\right) = Y \quad \text{and} \quad \lim_{n\to\infty} T_n^\dagger Q_n = T^\dagger.$$

- (H3)$^{\mathrm{w}}_{\mathrm{p}}$ (*weak perfect convergence*):

$$\mathcal{D}\left(T^\dagger\right) = Y \quad \text{and} \quad \operatorname*{w\text{-}lim}_{n\to\infty} T_n^\dagger Q_n = T^\dagger.$$

In addition, we introduce the following notation:

$$\operatorname*{s\text{-}lim}_{n\to\infty} S_n := \{\, x \mid \text{there is a sequence } \{x_n\} \text{ such that } S_n \ni x_n \to x \},$$

$$\operatorname*{w\text{-}\widetilde{lim}}_{n\to\infty} S_n := \left\{ x \mid \text{there is a sequence } \{x_n\} \text{ such that } \bigcup_{k=n}^{\infty} S_k \ni x_n \rightharpoonup x \right\},$$

where $\{S_n\}$ is a sequence of nonempty subsets of a Banach space, and

$$A^{\perp} := \{\, x \in X \text{ (or } Y) \mid \langle x, u\rangle = 0 \ (\forall\ u \in A)\},$$
$$B^{\perp_n} := \{\, x \in X_n \text{ (or } Y_n) \mid \langle x, u\rangle = 0 \ (\forall\ u \in B)\},$$

where $A \subseteq X$ (or Y) and $B \subseteq X_n$ (or Y_n). The notation in the paper can generally be found in [4] except when noted specifically.

In practice, to solve the best-approximate solution of (1.1) numerically we naturally reduce (1.1) to the finite system of (1.2) whose best-approximate solution is available, where the approximation setting (1.1) and (1.2) is usually designed to be with (H1) or $(\mathrm{H1})^{\mathrm{u}}$, which falls into the category of projection methods (see [17, pp. 199–321] for the theory of projection methods). For projection methods, the convergence and stability of their approximation settings are usually major objectives to study, so in this work we are interested in various convergence settings, such as the setting (1.1) and (1.2) with $(\mathrm{H3})^{\mathrm{s}}$, $(\mathrm{H3})^{\mathrm{w}}$, $(\mathrm{H3})^{\mathrm{s}}_{\mathrm{p}}$, $(\mathrm{H3})^{\mathrm{w}}_{\mathrm{p}}$, or $(\mathrm{H3})^{\mathrm{u}}_{\mathrm{p}}$, etc.

Next, in this section we give a brief survey of basic results about various convergence settings and then present new work. We will discuss the approximation setting (1.1) and (1.2) with $(\mathrm{H1})^{\mathrm{u}}$ or (H1) in the following situations:

Situation 1: $\mathcal{R}(T)$ is closed; that is, (1.1) is a well-posed problem.

Situation 2: $\mathcal{R}(T)$ is nonclosed; that is, (1.1) is an ill-posed problem.

1.1. Well-posed problems. For well-posed problems with the approximation setting (1.1) and (1.2), $(\mathrm{H3})^{\mathrm{s}}_{\mathrm{p}}$, $(\mathrm{H3})^{\mathrm{u}}_{\mathrm{p}}$, and $(\mathrm{H3})^{\mathrm{w}}_{\mathrm{p}}$ are the three kinds of convergence to be considered (in this situation $(\mathrm{H3})^{\mathrm{s}}$ and $(\mathrm{H3})^{\mathrm{w}}$ are equivalent to $(\mathrm{H3})^{\mathrm{s}}_{\mathrm{p}}$ and $(\mathrm{H3})^{\mathrm{w}}_{\mathrm{p}}$, respectively). The works [2, 3, 4, 14, 19, 20, 21, 23, 25, 29] may be viewed as research involving $(\mathrm{H3})^{\mathrm{s}}_{\mathrm{p}}$, $(\mathrm{H3})^{\mathrm{u}}_{\mathrm{p}}$, and $(\mathrm{H3})^{\mathrm{w}}_{\mathrm{p}}$, in which the basic results, due to Izumino [14], are as follows.

THEOREM 1.1 (see [14, Proposition 2.3]). *Let $\{T_n\}$ be a sequence in $\mathcal{CR}(X,Y)$, and let $T_n \xrightarrow{\mathrm{u}} T \in \mathcal{CR}(X,Y)$ $[(\mathrm{H1})^{\mathrm{u}}]$, where $\mathcal{CR}(X,Y)$ is the set of all operators in $\mathcal{B}(X,Y)$ with closed range. Then the following conditions are equivalent:*

(a) $T_n^{\dagger} \xrightarrow{\mathrm{u}} T^{\dagger}$ $[(\mathrm{H3})^{\mathrm{u}}_{\mathrm{p}}]$;

(b) $\sup_n \|T_n^{\dagger}\| < \infty$ [(H2)];

(c) $T_n^{\dagger}T_n \xrightarrow{\mathrm{u}} T^{\dagger}T$;

(d) $T_nT_n^{\dagger} \xrightarrow{\mathrm{u}} TT^{\dagger}$.

THEOREM 1.2 (see [14, Proposition 2.4]). *Let $\{T_n\}$ be a sequence in $\mathcal{CR}(X,Y)$, and let $T_n \xrightarrow{\mathrm{s}} T \in \mathcal{CR}(X,Y)$ [(H1)]. Then the following conditions are equivalent:*

(a) $T_n^{\dagger} \xrightarrow{\mathrm{s}} T^{\dagger}$ $[(\mathrm{H3})^{\mathrm{s}}_{\mathrm{p}}]$;

(b) $\sup_n \|T_n^{\dagger}\| < \infty$ [(H2)], $T_n^{\dagger}T_n \xrightarrow{\mathrm{s}} T^{\dagger}T$, *and* $T_nT_n^{\dagger} \xrightarrow{\mathrm{s}} TT^{\dagger}$.

From the computational point of view the above results may be at best regarded as academic, for T_n and T are defined on the same maybe infinite-dimensional space. Let us recall two results more convenient for numerical analysis.

THEOREM 1.3 (see [4, Theorem 2.1]). *For the approximation setting* (1.1) *and* (1.2), *suppose* (H1). *Then the following propositions hold:*

(a) $(\mathrm{H3})^{\mathrm{s}}_{\mathrm{p}} \Longleftrightarrow \begin{cases} (\mathrm{H2}), \\ \mathrm{w}\text{-}\widetilde{\lim}_{n\to\infty}\mathcal{N}(T_n) \subseteq \mathcal{N}(T) + \mathrm{s}\text{-}\lim_{n\to\infty}\mathcal{N}(T_n), \\ \mathrm{w}\text{-}\widetilde{\lim}_{n\to\infty}\mathcal{R}(T_n) \subseteq \overline{\mathcal{R}(T)}. \end{cases}$

(b) $(\mathrm{H3})^{\mathrm{s}}_{\mathrm{p}} \Longrightarrow \begin{cases} \mathrm{w}\text{-}\widetilde{\lim}_{n\to\infty}\mathcal{N}(T_n) = \mathrm{s}\text{-}\lim_{n\to\infty}\mathcal{N}(T_n) = \mathcal{N}(T), \\ \mathrm{w}\text{-}\widetilde{\lim}_{n\to\infty}\mathcal{R}(T_n) = \mathrm{s}\text{-}\lim_{n\to\infty}\mathcal{R}(T_n) = \mathcal{R}(T). \end{cases}$

THEOREM 1.4 (see [4, Theorem 3.1]). *For the approximation setting* (1.1) *and* (1.2), *suppose* $(\mathrm{H1})^{\mathrm{u}}$. *Then the following propositions hold:*

(a) $(\mathrm{H2}) \Longrightarrow \begin{cases} \mathrm{w}\text{-}\widetilde{\lim}_{n\to\infty}\mathcal{N}(T_n) = \mathrm{s}\text{-}\lim_{n\to\infty}\mathcal{N}(T_n) = \mathcal{N}(T), \\ \mathrm{w}\text{-}\widetilde{\lim}_{n\to\infty}\mathcal{R}(T_n) = \mathrm{s}\text{-}\lim_{n\to\infty}\mathcal{R}(T_n) = \mathcal{R}(T). \end{cases}$

(b) $(\mathrm{H3})^{\mathrm{s}}_{\mathrm{p}} \Longleftrightarrow (\mathrm{H3})^{\mathrm{w}}_{\mathrm{p}} \Longleftrightarrow (\mathrm{H2})$.

(c) $(\mathrm{H3})^{\mathrm{u}}_{\mathrm{p}} \Longleftrightarrow \begin{cases} (\mathrm{H2}), \\ \lim_{n\to\infty}\|T - Q_nTP_n\| = 0. \end{cases}$

As a corollary of Theorem 1.4, the following practical result holds.

COROLLARY 1.5 (see [4, Theorem 3.4]). *Let* $\{X_n\}$ *and* $\{Y_n\}$ *be sequences of finite-dimensional subspaces of* X *and* Y *which are increasing and eventually dense in* X *and* Y, *respectively. Suppose that* $T \in \mathcal{B}(X,Y)$ *and*

$$T_n := Q_nTP_n : X_n \to Y_n \quad (\forall\ n \in \mathbb{N}).$$

Then Theorem 1.4 *holds under the given conditions.*

Remark 1.6. By [4, Lemmas 2.12 and 2.13], there hold

$$\mathrm{w}\text{-}\widetilde{\lim_{n\to\infty}}\mathcal{N}(T_n) = \mathrm{s}\text{-}\lim_{n\to\infty}\mathcal{N}(T_n) = \mathcal{N}(T) \Longleftrightarrow \mathrm{s}\text{-}\lim_{n\to\infty}P_{\mathcal{N}(T_n)} = P_{\mathcal{N}(T)},$$
$$\mathrm{w}\text{-}\widetilde{\lim_{n\to\infty}}\mathcal{R}(T_n) = \mathrm{s}\text{-}\lim_{n\to\infty}\mathcal{R}(T_n) = \mathcal{R}(T) \Longleftrightarrow \mathrm{s}\text{-}\lim_{n\to\infty}Q_{\mathcal{R}(T_n)} = Q_{\mathcal{R}(T)}.$$

1.2. Ill-posed problems. For ill-posed problems with the approximation setting (1.1) and (1.2), one needs to consider $(\mathrm{H3})^{\mathrm{s}}$ and $(\mathrm{H3})^{\mathrm{w}}$ of the two kinds of convergence (because $(\mathrm{H3})^{\mathrm{s}}_{\mathrm{p}}$, $(\mathrm{H3})^{\mathrm{u}}_{\mathrm{p}}$, and $(\mathrm{H3})^{\mathrm{w}}_{\mathrm{p}}$ are meaningless in this situation). There are two natural and practical approaches to approximating $T^{\dagger}$, namely, the least-squares method and the dual least-squares method, which have been studied by many authors, e.g., [6, 8, 9, 12, 13, 15, 24, 26]. Note that the two methods fit into the framework of the setting (1.1) and (1.2) with $(\mathrm{H1})^{\mathrm{u}}$. We recall some basic results about the two methods, which are the assertions concerning $(\mathrm{H3})^{\mathrm{s}}$ and $(\mathrm{H3})^{\mathrm{w}}$.

1.2.1. The basic results on the least-squares method. The least-squares method (see, e.g., [9, 12, 10]) is to find the minimum-norm solution of (1.1) in a finite-dimensional subspace of X. That is, given a sequence $\{X_n\}_{n=1}^{\infty}$ of finite-dimensional subspaces of X which satisfies that X_n $(n \in \mathbb{N})$ are increasing and eventually dense in X, let x_n be the minimal norm least-squares solution of (1.1) in the subspace X_n. Obviously,

$$x_n = T_n^{\dagger}b,$$

where $T_n := T|_{X_n} \in \mathcal{B}(X_n, Y)$. Note that, since T_n has closed range,

$$\mathcal{D}\left(T_n^{\dagger}\right) = Y, \quad T_n^{\dagger} \in \mathcal{B}(Y, X_n) \subseteq \mathcal{B}(Y, X), \quad T_n^{\dagger} = (TP_n)^{\dagger}.$$

Here we recall the basic results of Groetsch and Neubauer [9, 12].

THEOREM 1.7 (see [9, Proposition 0]). *The sequence $\{x_n\}$ converges (to $T^\dagger b$) for all $b \in \mathcal{D}\left(T^\dagger\right)$ if and only if the sequence $\{\|R_n\|\}$ is bounded. Here $R_n := T_n^\dagger Q_{\mathcal{R}(T_n)} T$.*

Remark 1.8. The approximation setting (1.1) and (1.2) with X_n being as above, $Y_n := Y$, and $T_n := T|_{X_n}$ provides the mathematical framework of the least-squares method, and the above theorem says that:

$$(\mathrm{H3})^{\mathrm{s}} \iff (\mathrm{H3})^{\mathrm{w}} \iff \sup_n \left\|T_n^\dagger T\right\| < \infty\,.$$

THEOREM 1.9 (see [12, Theorem 1]). *For a given $g \in \mathcal{D}\left(T^\dagger\right)$, let $f_n := T_n^\dagger g$. Then $\{f_n\}$ converges weakly to $f_0 := T^\dagger g$ if and only if $\{\|f_n\|\}$ is bounded.*

Remark 1.10. The above theorem is incomplete because there is an absence of a condition in its statement. We should note that, without that condition, the theorem would be false. We will explain this observation by giving a counterexample.

1.2.2. The basic result on the dual least-squares method. The dual least-squares method (see, e.g., [6, 7, 10, 26]), different from the least-squares method, is to look to the range space itself for finite-dimensional subspaces in order to form finite rank approximations to $T^\dagger$. That is, given a sequence

$$V_1 \subseteq V_2 \subseteq V_3 \subseteq \cdots$$

of finite-dimensional subspaces of $\overline{\mathcal{R}(T)} = \mathcal{N}(T^*)^\perp \subseteq Y$ whose union is dense in $\mathcal{N}(T^*)^\perp$, let x_n be now defined as the least-squares solution of minimal norm of the equation

$$T_n u = b_n, \qquad T_n := Q_{V_n} T, \quad b_n := Q_{V_n} b.$$

Note that $T_n \in \mathcal{B}(X,Y) \cap \mathcal{B}(X,V_n)$ (with closed range), and therefore

$$T_n^\dagger \in \mathcal{B}(Y,X) \cap \mathcal{B}(V_n,X)\,, \quad x_n = T_n^\dagger b = T_n^\dagger b_n.$$

The basic result on the dual least-squares method is the following (see, e.g., [7, Theorem 3.24], [10, Theorems 4.1.2 and 4.1.3], and [26, Theorem 2.1]).

THEOREM 1.11. (a) *Let $b \in \mathcal{D}\left(T^\dagger\right)$ and x_n as above. Then $x_n = P_{U_n} x^\dagger$, where $U_n := T^*(V_n)$, $x^\dagger := T^\dagger b$. Moreover, $x_n \to x^\dagger$ as $n \to \infty$.*

(b) *If $b \notin \mathcal{D}\left(T^\dagger\right)$ and x_n is as above, then $\|x_n\| \to \infty$ as $n \to \infty$.*

Remark 1.12. Set $X_n := X$, $Y_n := V_n + \mathcal{N}(T^*)$, $P_n := I_X$, $Q_n := Q_{V_n} + Q_{\mathcal{N}(T^*)}$, and $T_n := Q_{V_n} T$. Then

$$\operatorname*{s-lim}_{n\to\infty} P_n = I_X, \quad \operatorname*{s-lim}_{n\to\infty} Q_n = I_Y, \quad T_n = Q_n T,$$

so that the approximation setting (1.1) and (1.2) in this case provides the mathematical framework of the dual least-squares method. Theorem 1.11 shows that, for the dual least-squares method, $(\mathrm{H3})^{\mathrm{s}}$ as well as $(\mathrm{H3})^{\mathrm{w}}$ are always guaranteed. However, the finite-dimensional subspaces V_n, whose union is dense in $\overline{\mathcal{R}(T)}$, are sometimes not easy to come by. In fact, the subspaces V_n can be replaced by more convenient finite-dimensional subspaces of Y that might be desirable; we notice that the special choice of V_n is only to result in $\sup_n \left\|T_n^\dagger Q_n T\right\| < \infty$ (see Corollaries 2.4 and 2.5).

1.2.3. Projection methods. The least-squares method and the dual least-squares method belong to the class of projection methods. The unified approximation

setting of projection methods is the following: For the approximation setting (1.1) and (1.2), let $\{X_n\}$ and $\{Y_n\}$ be sequences of finite-dimensional subspaces of X and Y, respectively, which are increasing and eventually dense in X and Y, and let

$$T_n := Q_n T P_n : X_n \to Y_n \quad (\forall\, n \in \mathbb{N})\,;$$

for a given vector $b \in \mathcal{D}\left(T^\dagger\right)$, let

$$x^\dagger := T^\dagger b, \qquad x_n := T_n^\dagger Q_n b \quad (\forall\, n \in \mathbb{N})\,.$$

Note that, since T_n has closed range, the operator $T_n^\dagger$ is bounded, and x_n is a stable approximation of $x^\dagger$; but, without additional assumptions, it cannot be guaranteed that x_n converges to $x^\dagger$ whether strongly or weakly, as the example from Seidman [27] shows. Naturally, it is interesting for us to look for criteria for $(\mathrm{H3})^{\mathrm{s}}$ and $(\mathrm{H3})^{\mathrm{w}}$ of the unified projection approximation setting. In this direction the author hasn't seen a desirable general result; but, in the special case of "$\mathcal{N}(T_n) = \{0\}$ $(n \in \mathbb{N})$," there holds

$$T_n^\dagger Q_n T x = P_n x + T_n^\dagger Q_n T\,(I - P_n)\,x,$$

and therefore

$$\lim_{n\to\infty} x_n = x^\dagger \ \ (\forall\, b \in \mathcal{R}(T)) \iff \sup_n \left\|T_n^\dagger Q_n T\right\| < \infty \implies \mathcal{N}(T) = \{0\}\,.$$

Here we refer to [16, Theorem 3.7] and [18, Theorem 17.1].

Of course, there are many other methods for obtaining stable approximations of infinite-dimensional Moore–Penrose inverses, the important ones including iterative methods, the conjugate gradient method, the Tikhonov regularization method, and various discretization forms of the above methods. Since there has already been much research about them (see [7, 10, 11, 22, 24, 26, 28], and the references cited therein), the present paper is not going to involve those methods.

The above is a brief survey of basic results about various convergences of the finite-dimensional approximation settings for infinite-dimensional Moore–Penrose inverses.

1.3. New work. The dual least-squares method is sometimes suggested in discretization of regularization methods for ill-posed problems. The approximation setting of this method as well as the least-squares method is with the special projection scheme (see Remarks 1.12 and 1.8), and therefore $(\mathrm{H3})^{\mathrm{s}}$ and $(\mathrm{H3})^{\mathrm{w}}$ are always guaranteed (see Theorem 1.11). Naturally, one wants to generalize this characterization to more general projection schemes. In section 2, we develop a unified theory on the general projection methods for infinite-dimensional Moore–Penrose inverses, which contains:

- fundamental convergence theorems (see Theorem 2.2 and its corollaries), which are generalizations of Theorems 1.11 and 1.7, and with a wider scope of application;
- the criterion for $(\mathrm{H3})^{\mathrm{s}}$, the criterion for $(\mathrm{H3})^{\mathrm{w}}$, etc. (see Theorem 2.6);
- general results for the least-squares method (see Theorem 2.8), by which the flaws in [9, Proposition 0] and [12, Theorem 1] are corrected (see Remark 2.9).

In section 3, we investigate $(\mathrm{H3})^{\mathrm{w}}_{\mathrm{p}}$ for the approximation setting (1.1) and (1.2) with (H1). Note that, for such a general approximation setting, $(\mathrm{H3})^{\mathrm{w}}_{\mathrm{p}}$ is equivalent neither to (H2) nor to $(\mathrm{H3})^{\mathrm{s}}_{\mathrm{p}}$; the relation among the three is

$$(\mathrm{H3})^{\mathrm{s}}_{\mathrm{p}} \implies (\mathrm{H3})^{\mathrm{w}}_{\mathrm{p}} \implies (\mathrm{H2}).$$

We present a result to characterize $(\mathrm{H3})^{\mathrm{w}}_{\mathrm{p}}$ by use of (H2) with other necessary compensated conditions (see Theorem 3.2), which is further explained by four examples. A corollary of the result is also presented, which can be applied to an interesting special case.

In section 4, we summarize the conclusions of the paper. Theorems 2.2 and 3.2 are the fundamental results obtained.

2. Projection methods for Moore–Penrose inverses. In this section, the approximation setting (1.1) and (1.2) satisfies that, for each $n \in \mathbb{N}$,

$$T_n := Q_n T P_n : X_n \to Y_n, \qquad \mathcal{R}(T_n) \text{ is closed}, \qquad R_n := T_n^{\dagger} Q_n T\,. \tag{2.1}$$

In such a case, the setting (1.1) and (1.2) provides the unified mathematical framework of projection methods (including the least-squares method and the dual least-squares method as special cases), which is with $(\mathrm{H1})^{\mathrm{u}}$ automatically. For this approximation setting, we will present a general convergence theory, which can be applied to ill-posed problems.

2.1. The fundamental convergence theorem for projection methods. We prepare a lemma first.

LEMMA 2.1. *Let H be a Hilbert space and $\{H_n\}$ a sequence of closed subspaces of H.*

(a) *There holds*

$$\{P_{H_n}\} \text{ is strongly convergent} \iff \operatorname*{s\,-lim}_{n\to\infty} H_n = \operatorname*{w\,-\widetilde{lim}}_{n\to\infty} H_n\,;$$

in the case that $\{P_{H_n}\}$ is strongly convergent,

$$\operatorname*{s\,-lim}_{n\to\infty} P_{H_n} = P_M\,, \quad \text{where } M := \operatorname*{s\,-lim}_{n\to\infty} H_n\,.$$

(b) *If N is a closed subspace of H, then*

$$\begin{aligned} \operatorname*{s\,-lim}_{n\to\infty} P_{H_n} = P_N &\iff \operatorname*{w\,-lim}_{n\to\infty} P_{H_n} = P_N \\ &\iff \lim_{n\to\infty} \|P_{H_n} x\| = \|P_N x\| \ \ (\forall\, x \in X)\,. \end{aligned}$$

Proof. See [4, Lemma 2.13] or [5, Theorem 2.1] for the proof of (a).

In order to prove (b) we need only to show that

$$\lim_{n\to\infty} \|P_{H_n} x\| = \|P_N x\| \ \ (\forall\ x \in X) \implies \operatorname*{s\,-lim}_{n\to\infty} P_{H_n} = P_N. \tag{2.2}$$

Indeed, by the polarization identities (see, e.g., [1, p. 20]) and the fact of "$P_N^2 = P_N = P_N^*$," it is easy to see that

$$\lim_{n\to\infty} \|P_{H_n} x\| = \|P_N x\| \ \ (\forall\ x \in X) \iff \operatorname*{w\,-lim}_{n\to\infty} P_{H_n} = P_N. \tag{2.3}$$

Thus, (2.2) follows with (2.3) and [30, Theorem 8 on p. 124]. ∎

Now we present the fundamental convergence theorem for projection methods, which is one of the main results in the paper.

THEOREM 2.2. *Let the approximation setting* (1.1) *and* (1.2) *satisfy* (2.1). *If* $\sup_n \|R_n\| < \infty$, *then:*

(a) $\text{s-lim}_{n\to\infty} \mathcal{N}(T_n) = \text{w-}\widetilde{\lim}_{n\to\infty} \mathcal{N}(T_n) = \mathcal{N}(T)$*; that is,*

$$\operatorname*{s\,-lim}_{n\to\infty} P_{\mathcal{N}(T_n)} = P_{\mathcal{N}(T)};$$

(b) *for all* $b \in \mathcal{D}\left(T^{\dagger}\right)$,

$$\left\|T_n^{\dagger} Q_n Q_{\overline{\mathcal{R}(T)}} b - T^{\dagger} b\right\| \leqslant \left\|P_{\mathcal{N}(T_n)} T^{\dagger} b\right\| + \|R_n - I_X\| \operatorname{dist}\left(T^{\dagger} b, X_n\right) \to 0;$$

(c) $\text{s-lim}_{n\to\infty} Q_{\mathcal{R}(T_n)} Q_{\overline{\mathcal{R}(T)}} = Q_{\overline{\mathcal{R}(T)}}$, *and*

$$\lim_{n\to\infty} \left\|T_n^{\dagger} Q_n Q_{\overline{\mathcal{R}(T)}} y\right\| = \begin{cases} \left\|T^{\dagger} y\right\|, & \text{as } y \in \mathcal{D}\left(T^{\dagger}\right), \\ \infty, & \text{as } y \notin \mathcal{D}\left(T^{\dagger}\right). \end{cases}$$

Proof. The proof will be carried out in three steps.

Step 1 (to prove (a)). Let $x \in \mathcal{N}(T)$. Then

$$R_n P_n x + R_n (I - P_n) x = 0.$$

Since

$$R_n P_n x = T_n^{\dagger} T_n P_n x = P_n x - P_{\mathcal{N}(T_n)} x,$$

and since $\sup_n \|R_n\| < \infty$, we then have

$$P_{\mathcal{N}(T_n)} x = P_n x + R_n (I - P_n) x \to x \quad \text{as } n \to \infty.$$

Thus, it follows that

$$\mathcal{N}(T) \subseteq \operatorname*{s\text{-}lim}_{n\to\infty} \mathcal{N}(T_n). \tag{2.4}$$

Now let $x \in \text{w-}\widetilde{\lim}_{n\to\infty} \mathcal{N}(T_n)$. Then, by the definition of $\text{w-}\widetilde{\lim}_{n\to\infty} \mathcal{N}(T_n)$, there exists a sequence $\{x_n\}$ such that

$$x_n \in \bigcup_{k \geqslant n} \mathcal{N}(T_k) \ (\forall\, n) \quad \text{and} \quad x_n \rightharpoonup x \text{ as } n \to \infty. \tag{2.5}$$

Since for each $n \in \mathbb{N}$ there is a $k_n \in \mathbb{N}$ such that $k_n \geqslant n$ and $x_n \in \mathcal{N}(T_{k_n})$, and since $T_{k_n} = Q_{k_n} T P_{k_n}$ (see (2.1)), we then have

$$\begin{aligned} |\langle Tx, u\rangle| &\leqslant |\langle T(x - x_n), u\rangle| + |\langle T_{k_n} x_n, u\rangle| + |\langle Tx_n, (I_Y - Q_{k_n}) u\rangle| \\ &\leqslant |\langle T(x - x_n), u\rangle| + \|T\| \|x_n\| \|(I_Y - Q_{k_n}) u\| \qquad (\forall\, u \in X). \end{aligned} \tag{2.6}$$

So it follows from (2.5) and (2.6) that $x \in \mathcal{N}(T)$. Thus, we obtain

$$\operatorname*{w\text{-}\widetilde{lim}}_{n\to\infty} \mathcal{N}(T_n) \subseteq \mathcal{N}(T). \tag{2.7}$$

From (2.4) and (2.7), together with Lemma 2.1 or Remark 1.6, we conclude that (a) holds.

Step 2 (to prove (b)). Let $b \in \mathcal{D}\left(T^{\dagger}\right)$. Then

$$\begin{aligned} T_n^{\dagger} Q_n Q_{\overline{\mathcal{R}(T)}} b - T^{\dagger} b &= \left(T_n^{\dagger} Q_n Q_{\overline{\mathcal{R}(T)}} b - T_n^{\dagger} T_n P_n T^{\dagger} b\right) + \left(T_n^{\dagger} T_n P_n T^{\dagger} b - T^{\dagger} b\right) \\ &= \left(T_n^{\dagger} Q_n Q_{\overline{\mathcal{R}(T)}} b - T_n^{\dagger} Q_n T P_n T^{\dagger} b\right) + \left(T_n^{\dagger} T_n P_n T^{\dagger} b - T^{\dagger} T T^{\dagger} b\right) \\ &= T_n^{\dagger} Q_n \left(T T^{\dagger} b - T P_n T^{\dagger} b\right) + \left(T_n^{\dagger} T_n P_n - T^{\dagger} T\right) T^{\dagger} b \\ &= R_n (I_X - P_n) T^{\dagger} b + \left(P_n - P_{\mathcal{N}(T_n)} - I_X + P_{\mathcal{N}(T)}\right) T^{\dagger} b \\ &= -P_{\mathcal{N}(T_n)} T^{\dagger} b + (R_n - I_X)(I_X - P_n) T^{\dagger} b, \end{aligned}$$

so that

$$\left\|T_n^\dagger Q_n Q_{\overline{\mathcal{R}(T)}} b - T^\dagger b\right\| \leqslant \left\|P_{\mathcal{N}(T_n)} T^\dagger b\right\| + \|R_n - I_X\| \operatorname{dist}\left(T^\dagger b, X_n\right) . \tag{2.8}$$

Thus, (b) follows from (2.8) and (a).

Step 3 (to prove (c)). By (b), there holds

$$\lim_{n\to\infty} T_n^\dagger Q_n Q_{\overline{\mathcal{R}(T)}} b = T^\dagger b \quad \left(\forall\ b \in \mathcal{D}\left(T^\dagger\right)\right) .$$

Hence, for any $v \in Y$ and $b \in \mathcal{D}\left(T^\dagger\right)$,

$$\begin{aligned}\left\langle Q_{\mathcal{R}(T_n)} Q_{\overline{\mathcal{R}(T)}} b\,, v\right\rangle &= \left\langle T_n T_n^\dagger Q_n Q_{\overline{\mathcal{R}(T)}} b\,, v\right\rangle \\ &= \left\langle Q_n T T_n^\dagger Q_n Q_{\overline{\mathcal{R}(T)}} b\,, v\right\rangle \\ &= \left\langle T T_n^\dagger Q_n Q_{\overline{\mathcal{R}(T)}} b\,, Q_n v\right\rangle \to \left\langle Q_{\overline{\mathcal{R}(T)}} b\,, v\right\rangle \qquad \text{as } n \to \infty;\end{aligned}$$

that is,

$$\operatorname*{w\text{-}lim}_{n\to\infty} Q_{\mathcal{R}(T_n)} Q_{\overline{\mathcal{R}(T)}} b = Q_{\overline{\mathcal{R}(T)}} b \qquad \left(\forall\ b \in \mathcal{D}\left(T^\dagger\right)\right) . \tag{2.9}$$

Since $\overline{\lim}_{n\to\infty} \|Q_{\mathcal{R}(T_n)} Q_{\overline{\mathcal{R}(T)}} b\| \leqslant \|Q_{\overline{\mathcal{R}(T)}} b\|$ is always valid, by [30, Theorem 1 on p. 120], (2.9) implies that

$$\lim_{n\to\infty} \left\|Q_{\mathcal{R}(T_n)} Q_{\overline{\mathcal{R}(T)}} b\right\| = \left\|Q_{\overline{\mathcal{R}(T)}} b\right\| \qquad \left(\forall\ b \in \mathcal{D}\left(T^\dagger\right)\right) . \tag{2.10}$$

Now, by [30, Theorem 8 on p. 124], it follows from (2.9) and (2.10) that

$$\operatorname*{s\text{-}lim}_{n\to\infty} Q_{\mathcal{R}(T_n)} Q_{\overline{\mathcal{R}(T)}} b = Q_{\overline{\mathcal{R}(T)}} b \qquad \left(\forall\ b \in \mathcal{D}\left(T^\dagger\right)\right) .$$

Notice that $\mathcal{D}\left(T^\dagger\right)$ is a dense subspace of Y, so, by the Banach–Steinhaus theorem [1, Theorem 2.4.5 on p. 58], we obtain

$$\operatorname*{s\text{-}lim}_{n\to\infty} Q_{\mathcal{R}(T_n)} Q_{\overline{\mathcal{R}(T)}} = Q_{\overline{\mathcal{R}(T)}} . \tag{2.11}$$

Next, let $y \notin \mathcal{D}\left(T^\dagger\right)$, and we need only to show that

$$\lim_{n\to\infty} \left\|T_n^\dagger Q_n Q_{\overline{\mathcal{R}(T)}} y\right\| = \infty . \tag{2.12}$$

Assume that there is a sequence $n_k \to \infty$ $(k \to \infty)$ such that

$$\sup_k \left\|T_{n_k}^\dagger Q_{n_k} Q_{\overline{\mathcal{R}(T)}} y\right\| < \infty ,$$

so that $\{T_{n_k}^\dagger Q_{n_k} Q_{\overline{\mathcal{R}(T)}} y\}$ has a subsequence (denoted again by $\{T_{n_k}^\dagger Q_{n_k} Q_{\overline{\mathcal{R}(T)}} y\}$) which converges weakly to an $x_\infty \in X$. Since T is also weakly sequentially continuous,

$$T T_{n_k}^\dagger Q_{n_k} Q_{\overline{\mathcal{R}(T)}} y \rightharpoonup T x_\infty \quad \text{as } k \to \infty . \tag{2.13}$$

On the other hand, for any $v \in Y$ it follows from (2.11) and (2.13) that

$$\begin{aligned}\left\langle TT_{n_k}^{\dagger} Q_{n_k} Q_{\overline{\mathcal{R}(T)}} y\,, v \right\rangle &= \left\langle TT_{n_k}^{\dagger} Q_{n_k} Q_{\overline{\mathcal{R}(T)}} y\,, Q_{n_k} v \right\rangle \\ &\qquad + \left\langle TT_{n_k}^{\dagger} Q_{n_k} Q_{\overline{\mathcal{R}(T)}} y\,, (I_Y - Q_{n_k})\, v \right\rangle \\ &= \left\langle T_{n_k} T_{n_k}^{\dagger} Q_{n_k} Q_{\overline{\mathcal{R}(T)}} y\,, v \right\rangle + \left\langle TT_{n_k}^{\dagger} Q_{n_k} Q_{\overline{\mathcal{R}(T)}} y\,, (I_Y - Q_{n_k})\, v \right\rangle \\ &= \left\langle Q_{\mathcal{R}\left(T_{n_k}\right)} Q_{\overline{\mathcal{R}(T)}} y\,, v \right\rangle + \left\langle TT_{n_k}^{\dagger} Q_{n_k} Q_{\overline{\mathcal{R}(T)}} y\,, (I_Y - Q_{n_k})\, v \right\rangle \\ &\to \left\langle Q_{\overline{\mathcal{R}(T)}} y\,, v \right\rangle \qquad \text{as } k \to \infty\,.\end{aligned}$$

This implies that

$$TT_{n_k}^{\dagger} Q_{n_k} Q_{\overline{\mathcal{R}(T)}} y \rightharpoonup Q_{\overline{\mathcal{R}(T)}} y \quad \text{as } k \to \infty\,. \tag{2.14}$$

From (2.13) and (2.14) we obtain that $Q_{\overline{\mathcal{R}(T)}} y = T x_\infty \in \mathcal{R}(T)$, so that $y \in \mathcal{D}\left(T^{\dagger}\right)$. Thus, if $y \notin \mathcal{D}\left(T^{\dagger}\right)$, no bounded sequence $\{\|T_{n_k}^{\dagger} Q_{n_k} Q_{\overline{\mathcal{R}(T)}} y\|\}$ (where $n_k \to \infty$ $(k \to \infty)$) can exist; hence (2.12) holds. ∎

2.2. Corollaries of the fundamental convergence theorem. The following three results are simple corollaries of Theorem 2.2.

COROLLARY 2.3 (the least-squares method). *Let $T \in \mathcal{B}(X,Y)$, X_n $(n \in \mathbb{N})$ be finite-dimensional subspaces of X such that*

$$X_1 \subseteq X_2 \subseteq X_3 \subseteq \cdots\,, \quad \overline{\bigcup_{n=1}^{\infty} X_n} = X\,.$$

Let $T_n := T|_{X_n} : X_n \to Y$ $(n \in \mathbb{N})$. If $\sup_n \left\|T_n^{\dagger} T\right\| < \infty$, then:

(a) *there hold*

$$\mathcal{N}(T_n) = \mathcal{N}(T) \cap X_n, \qquad \overline{\bigcup_{n=1}^{\infty} \mathcal{N}(T_n)} = \mathcal{N}(T)\;;$$

(b) *for all $b \in \mathcal{D}\left(T^{\dagger}\right)$,*

$$\left\|T_n^{\dagger} b - T^{\dagger} b\right\| \leqslant \left\|T_n^{\dagger} T - I_X\right\| \operatorname{dist}\left(T^{\dagger} b, X_n\right) \to 0\,;$$

(c) *there hold*

$$\mathcal{R}(T_n) = T(X_n)\;, \qquad \overline{\bigcup_{n=1}^{\infty} \mathcal{R}(T_n)} = \overline{\mathcal{R}(T)}\,,$$

and

$$\lim_{n\to\infty} \left\|T_n^{\dagger} y\right\| = \begin{cases} \left\|T^{\dagger} y\right\|, & \text{as } y \in \mathcal{D}\left(T^{\dagger}\right), \\ \infty, & \text{as } y \notin \mathcal{D}\left(T^{\dagger}\right). \end{cases}$$

COROLLARY 2.4 (the dual least-squares method). *Let $T \in \mathcal{B}(X,Y)$, V_n $(n \in \mathbb{N})$ be finite-dimensional subspaces of $\overline{\mathcal{R}(T)}$ such that*

$$V_1 \subseteq V_2 \subseteq V_3 \subseteq \cdots\,, \quad \overline{\bigcup_{n=1}^{\infty} V_n} = \overline{\mathcal{R}(T)}\,.$$

Let $T_n := Q_{V_n} T : X \to V_n$ $(n \in \mathbb{N})$. Then:

(a) *there hold*

$$\mathcal{N}(T_n) = T^*(V_n)^\perp \quad \text{and} \quad \bigcap_{n=1}^{\infty} \mathcal{N}(T_n) = \mathcal{N}(T)\,;$$

(b) *for all $b \in \mathcal{D}\left(T^\dagger\right)$,*

$$\left\|T_n^\dagger Q_{V_n} b - T^\dagger b\right\| \leqslant \left\|P_{\mathcal{N}(T_n)} T^\dagger b\right\| = \operatorname{dist}\left(T^\dagger b,\, T^*(V_n)\right) \to 0\,,$$

and therefore

$$T_n^\dagger Q_{V_n} b = P_{T^*(V_n)} T^\dagger b\,;$$

(c) *there hold*

$$\mathcal{R}(T_n) = V_n \quad \text{and} \quad \lim_{n\to\infty} \left\|T_n^\dagger Q_{V_n} y\right\| = \begin{cases} \left\|T^\dagger y\right\|, & \text{as } y \in \mathcal{D}\left(T^\dagger\right), \\ \infty, & \text{as } y \notin \mathcal{D}\left(T^\dagger\right); \end{cases}$$

(d) *$\mathcal{R}(T)$ is closed if and only if $\sup_n \left\|T_n^\dagger\right\|_n < \infty$, where*

$$\left\|T_n^\dagger\right\|_n := \sup\left\{\left\|T_n^\dagger y\right\| \;\middle|\; y \in V_n,\, \|y\| = 1\right\}.$$

Corollary 2.5 (the generalized dual least-squares method). *Let $T \in \mathcal{B}(X,Y)$, Y_n $(n \in \mathbb{N})$ be finite-dimensional subspaces of Y such that*

$$\operatorname*{s\,-\,lim}_{n\to\infty} Q_n = I_Y, \qquad \text{where} \quad Q_n := Q_{Y_n} \;\; (n \in \mathbb{N})\,.$$

Let $T_n := Q_n T : X \to Y_n$ $(n \in \mathbb{N})$. Then:

(a) $\text{s-}\lim_{n\to\infty} \mathcal{N}(T_n) = \widetilde{\text{w-}\lim}_{n\to\infty} \mathcal{N}(T_n) = \mathcal{N}(T)$, *with* $\mathcal{N}(T_n) = T^*(Y_n)^\perp$, *and*

$$\operatorname*{s\,-\,lim}_{n\to\infty} P_{\mathcal{N}(T_n)} = P_{\mathcal{N}(T)};$$

(b) *for all $b \in \mathcal{D}\left(T^\dagger\right)$,*

$$\left\|T_n^\dagger Q_n Q_{\overline{\mathcal{R}(T)}} b - T^\dagger b\right\| \leqslant \left\|P_{\mathcal{N}(T_n)} T^\dagger b\right\| = \operatorname{dist}\left(T^\dagger b, T^*(Y_n)\right) \to 0\,,$$

and therefore

$$T_n^\dagger Q_n Q_{\overline{\mathcal{R}(T)}} b = P_{T^*(Y_n)} T^\dagger b\,;$$

(c) $\text{s-}\lim_{n\to\infty} Q_{\mathcal{R}(T_n)} Q_{\overline{\mathcal{R}(T)}} = Q_{\overline{\mathcal{R}(T)}}$ *and*

$$\lim_{n\to\infty} \left\|T_n^\dagger Q_n Q_{\overline{\mathcal{R}(T)}} y\right\| = \begin{cases} \left\|T^\dagger y\right\|, & \text{as } y \in \mathcal{D}\left(T^\dagger\right), \\ \infty, & \text{as } y \notin \mathcal{D}\left(T^\dagger\right). \end{cases}$$

Proof. Since $T_n := Q_n T$, we have

$$\mathcal{N}(T_n) = \mathcal{R}(T^* Q_n)^\perp = T^*(Y_n)^\perp\,.$$

Let $R_n := T_n^\dagger Q_n T$. Then

$$R_n = T_n^\dagger T_n = P_{T^*(Y_n)}, \qquad \|R_n\| \leqslant 1\,.$$

Therefore the corollary is a consequence of Theorem 2.2, with $X_n = X$ $(n \in \mathbb{N})$. ∎

2.3. The criteria for (H3)$^{\mathrm{s}}$ and (H3)$^{\mathrm{w}}$ on projection methods. We can obtain the following result by use of Theorem 2.2.

THEOREM 2.6. *Let the approximation setting* (1.1) *and* (1.2) *satisfy* (2.1). *Then:*

(a) $\lim_{n\to\infty} T_n^\dagger Q_n Q_{\overline{\mathcal{R}(T)}} y = T^\dagger y \ \left(\forall\ y \in \mathcal{D}\left(T^\dagger\right)\right) \iff \sup_{n\in\mathbb{N}} \left\|T_n^\dagger Q_n T\right\| < \infty$;

(b) $(\mathrm{H3})^{\mathrm{s}} \iff \begin{cases} \sup_{n\in\mathbb{N}} \left\|T_n^\dagger Q_n T\right\| < \infty, \\ T_n^\dagger Q_n y \to 0 \text{ if } T_n^* Q_n y \to 0 \end{cases}$ *(criterion for* (H3)$^{\mathrm{s}}$*);*

(c) $(\mathrm{H3})^{\mathrm{w}} \iff \begin{cases} \sup_{n\in\mathbb{N}} \left\|T_n^\dagger Q_n T\right\| < \infty, \\ T_n^\dagger Q_n y \rightharpoonup 0 \text{ if } T_n^* Q_n y \to 0 \end{cases}$ *(criterion for* (H3)$^{\mathrm{w}}$*);*

(d) $(\mathrm{H3})^{\mathrm{w}} \implies \begin{cases} \mathrm{s}\text{-}\lim_{n\to\infty} P_{\mathcal{N}(T_n)} = P_{\mathcal{N}(T)}, \\ \mathrm{s}\text{-}\lim_{n\to\infty} Q_{\mathcal{R}(T_n)} = Q_{\overline{\mathcal{R}(T)}}. \end{cases}$

Remark 2.7. The two conditions of characterizing (H3)$^{\mathrm{s}}$ or (H3)$^{\mathrm{w}}$, in (b) or (c), are independent (see Examples 3.6 and 4.2).

Proof. It is clear that (a) holds by use of the uniform boundedness principle [30, Theorem 1 on p. 68] and Theorem 2.2.

Notice that for any $b \in \mathcal{D}\left(T^\dagger\right)$ there holds

$$T_n^\dagger Q_n b - T^\dagger b = T_n^\dagger Q_n Q_{\mathcal{R}(T)^\perp} b - P_{\mathcal{N}(T_n)} T^\dagger b + \left(R_n - I_X\right)\left(I_X - P_n\right) T^\dagger b\,. \tag{2.15}$$

From (2.15), by the uniform boundedness principle and Theorem 2.2, we have

$$(\mathrm{b}') \quad (\mathrm{H3})^{\mathrm{s}} \iff \begin{cases} \sup_{n\in\mathbb{N}} \left\|T_n^\dagger Q_n T\right\| < \infty, \\ T_n^\dagger Q_n y \to 0 \text{ if } y \in \mathcal{R}\left(T\right)^\perp; \end{cases}$$

$$(\mathrm{c}') \quad (\mathrm{H3})^{\mathrm{w}} \iff \begin{cases} \sup_{n\in\mathbb{N}} \left\|T_n^\dagger Q_n T\right\| < \infty, \\ T_n^\dagger Q_n y \rightharpoonup 0 \text{ if } y \in \mathcal{R}\left(T\right)^\perp. \end{cases}$$

So, to prove (b) and (c) we need only to show that

$$\mathcal{R}\left(T\right)^\perp = \left\{\, y \in Y \mid T_n^* Q_n y \to 0 \ \ (n \to \infty)\right\}. \tag{2.16}$$

Because for any $x \in X$ and $y \in Y$ there holds

$$\left\langle x\,, T^* y\right\rangle = \left\langle x\,, T_n^* Q_n y\right\rangle + \left\langle x - P_n x\,, T^* y\right\rangle + \left\langle T P_n x\,, y - Q_n y\right\rangle,$$

this implies that

$$y \in \mathcal{N}\left(T^*\right) \iff T_n^* Q_n y \rightharpoonup 0 \ \ (n \to \infty)\,. \tag{2.17}$$

Notice that

$$T_n^* = P_n T^* Q_n : Y_n \to X_n\,.$$

This implies that

$$y \in \mathcal{N}\left(T^*\right) \implies T_n^* Q_n y = P_n T^* \left(Q_n - I_Y\right) y \to 0 \ \ (n \to \infty)\,. \tag{2.18}$$

Now, from (2.17) and (2.18), we obtain

$$y \in \mathcal{N}\left(T^*\right) \iff T_n^* Q_n y \to 0 \ \ (n \to \infty)\,.$$

Since $\mathcal{R}\left(T\right)^\perp = \mathcal{N}\left(T^*\right)$, (2.16) follows.

To prove (d), we note that (H3)$^{\mathrm{w}}$ implies that $\sup_{n\in\mathbb{N}}\left\|T_n^{\dagger}Q_nT\right\|<\infty$, and therefore (by Theorem 2.2)

$$\operatorname*{s\text{-}lim}_{n\to\infty}P_{\mathcal{N}(T_n)}=P_{\mathcal{N}(T)}\,,\quad \operatorname*{s\text{-}lim}_{n\to\infty}Q_{\mathcal{R}(T_n)}Q_{\overline{\mathcal{R}(T)}}=Q_{\overline{\mathcal{R}(T)}}\,. \tag{2.19}$$

Hence, we need only to show that

$$\operatorname*{s\text{-}lim}_{n\to\infty}Q_{\mathcal{R}(T_n)}=Q_{\overline{\mathcal{R}(T)}}\,. \tag{2.20}$$

Notice that

$$\begin{aligned}Q_{\mathcal{R}(T_n)}Q_{\mathcal{R}(T)^{\perp}}&=T_nT_n^{\dagger}Q_nQ_{\mathcal{R}(T)^{\perp}}\\&=TT_n^{\dagger}Q_nQ_{\mathcal{R}(T)^{\perp}}-\left(I_Y-Q_n\right)TT_n^{\dagger}Q_nQ_{\mathcal{R}(T)^{\perp}}\,.\end{aligned}$$

Since (H3)$^{\mathrm{w}}$ implies that

$$TT_n^{\dagger}Q_nQ_{\mathcal{R}(T)^{\perp}}\rightharpoonup TT^{\dagger}Q_{\mathcal{R}(T)^{\perp}}=0\quad\text{as}\quad n\to\infty\,,$$

by which $\sup_n\left\|TT_n^{\dagger}Q_nQ_{\mathcal{R}(T)^{\perp}}\right\|<\infty$ holds, and since this implies that

$$\left(I_Y-Q_n\right)TT_n^{\dagger}Q_nQ_{\mathcal{R}(T)^{\perp}}\rightharpoonup 0\quad\text{as}\quad n\to\infty\,,$$

we then obtain

$$Q_{\mathcal{R}(T_n)}Q_{\mathcal{R}(T)^{\perp}}\rightharpoonup 0\quad\text{as}\quad n\to\infty\,.$$

This together with (2.19) gives

$$\operatorname*{w\text{-}lim}_{n\to\infty}Q_{\mathcal{R}(T_n)}=Q_{\overline{\mathcal{R}(T)}}.$$

Thus, (2.20) follows by Lemma 2.1. □

2.4. General results for the least-squares method. We especially investigate the least-squares method and correct flaws in the Groetsch results in the present subsection.

THEOREM 2.8. *Let $T\in\mathcal{B}(X,Y)$, X_n $(n\in\mathbb{N})$ be finite-dimensional subspaces of X such that*

$$X_1\subseteq X_2\subseteq X_3\subseteq\cdots\,,\quad \overline{\bigcup_{n=1}^{\infty}X_n}=X\,.$$

Let $T_n:=T|_{X_n}:X_n\to Y$ $(n\in\mathbb{N})$. Then:

(a) *for any $y\in Y$ there holds*

$$\lim_{n\to\infty}\left\|TT_n^{\dagger}y-Q_{\overline{\mathcal{R}(T)}}y\right\|=0;$$

if $\varliminf_{n\to\infty}\left\|T_n^{\dagger}y\right\|<\infty$, then $y\in\mathcal{D}\left(T^{\dagger}\right)$ and

$$\varnothing\neq\operatorname*{w\text{-}\widetilde{lim}}_{n\to\infty}\left\{T_n^{\dagger}y\right\}\subseteq T^{-1}\left(Q_{\overline{\mathcal{R}(T)}}y\right)\,.$$

(b) *If*

$$\overline{\bigcup_{n=1}^{\infty}\left(\mathcal{N}(T)\cap X_n\right)}=\mathcal{N}(T) \tag{$\star$}$$

and $y \in Y$ is given, then

$$\left\{ y \in \mathcal{D}\left(T^{\dagger}\right),\ \operatorname*{w\text{-}lim}_{n\to\infty} T_n^{\dagger} y = T^{\dagger} y \right\} \Longleftrightarrow \sup_n \left\| T_n^{\dagger} y \right\| < \infty .$$

(c) *If* $\sup_n \|R_n\| < \infty$ *($R_n := T_n^{\dagger} T$), then condition ($\star$) holds.*
(d) $(\mathrm{H3})^{\mathrm{s}} \Longleftrightarrow (\mathrm{H3})^{\mathrm{w}} \Longleftrightarrow \sup_n \|R_n\| < \infty$.
(e) $(\mathrm{H3})^{\mathrm{s}}_{\mathrm{p}} \Longleftrightarrow (\mathrm{H3})^{\mathrm{w}}_{\mathrm{p}} \Longleftrightarrow (\mathrm{H2}) \Longleftrightarrow \{\sup_n \|R_n\| < \infty$, *and* $\mathcal{R}(T)$ *is closed*$\}$.

Remark 2.9. Assertion (b) is Theorem 1.9 (due to Groetsch–Neubauer [12, Theorem 1]). Note that condition ($\star$) is missing in the original statement of the theorem, and without that condition the theorem would be false. We will explain this observation by giving a counterexample (see Example 2.10 below).

Assertion (d) is Theorem 1.7 (due to Groetsch [9, Proposition 0]). Its proof given in [9] (or [10, Theorem 4.1.1]) applies only to the case of "$\mathcal{N}(T_n) = \{0\} (n \in \mathbb{N})$."

Proof. (a) Note that $\mathcal{R}(T_n)$ is closed (since X_n is finite-dimensional) and that

$$\mathcal{R}(T_n) \subseteq \mathcal{R}(T_{n+1}) \subseteq \mathcal{R}(T)\ \ (n \in \mathbb{N}) \quad \text{and} \quad \overline{\bigcup_{n\in\mathbb{N}} \mathcal{R}(T_n)} = \overline{\mathcal{R}(T)} .$$

This implies that

$$T T_n^{\dagger} = Q_{\mathcal{R}(T_n)} \xrightarrow{\mathrm{s}} Q_{\overline{\mathcal{R}(T)}} \quad \text{as} \quad n \to \infty . \tag{2.21}$$

Assume that $\varliminf_{n\to\infty} \left\| T_n^{\dagger} y \right\| < \infty$. Then $\left\{ T_n^{\dagger} y \right\}$ has a bounded subsequence $\left\{ T_{n_k}^{\dagger} y \right\}$ and therefore has a subsequence, again denoted by $\left\{ T_{n_k}^{\dagger} y \right\}$, converging weakly to some $u \in X$ by the Eberlein–Shmulyan theorem (see, e.g., [30, p. 141]). Since T is bounded, we have

$$Q_{\mathcal{R}\left(T_{n_k}\right)} y = T T_{n_k}^{\dagger} y \rightharpoonup T u \quad \text{as} \quad k \to \infty. \tag{2.22}$$

From (2.21) and (2.22) it follows that $Q_{\overline{\mathcal{R}(T)}} y = T u$, that is, $y \in \mathcal{D}(T^{\dagger})$ and $u \in T^{-1}(Q_{\overline{\mathcal{R}(T)}} y)$. Clearly, we can prove that

$$\varnothing \neq \operatorname*{w\text{-}\widetilde{lim}}_{n\to\infty} \left\{ T_n^{\dagger} y \right\} \subseteq T^{-1} \left(Q_{\overline{\mathcal{R}(T)}} y \right) . \tag{2.23}$$

(b) We need only to show that if condition ($\star$) is satisfied and $y \in Y$ is with $\sup_n \left\| T_n^{\dagger} y \right\| < \infty$, then

$$y \in \mathcal{D}\left(T^{\dagger}\right) \quad \text{and} \quad \operatorname*{w\text{-}lim}_{n\to\infty} T_n^{\dagger} y = T^{\dagger} y . \tag{2.24}$$

Indeed, by (a) we obtain immediately $y \in \mathcal{D}\left(T^{\dagger}\right)$ and (2.23). Since

$$\mathcal{N}(T_n) = \mathcal{N}(T) \cap X_n, \quad \mathcal{N}(T_n)^{\perp_n} = \mathcal{N}(T_n)^{\perp} \cap X_n,$$

by condition ($\star$) we have that

$$\operatorname*{w\text{-}\widetilde{lim}}_{n\to\infty} \left\{ T_n^{\dagger} y \right\} \subseteq \operatorname*{w\text{-}\widetilde{lim}}_{n\to\infty} \mathcal{N}(T_n)^{\perp_n} \subseteq \bigcap_{n\in\mathbb{N}} (\mathcal{N}(T) \cap X_n)^{\perp} = \mathcal{N}(T)^{\perp} . \tag{2.25}$$

Hence, it follows from (2.23) and (2.25) that

$$\operatorname*{w\text{-}\widetilde{lim}}_{n\to\infty} \left\{ T_n^{\dagger} y \right\} = \mathcal{N}(T)^{\perp} \cap T^{-1} \left(Q_{\overline{\mathcal{R}(T)}} y \right) .$$

This together with $\sup_n \|T_n^\dagger y\| < \infty$ implies that w-$\lim_{n\to\infty} T_n^\dagger y = T^\dagger y$, completing the proof of (2.24).

(c) Assume that $\sup_n \|R_n\| < \infty$. Then, by (a) of Corollary 2.3, condition $(\star)$ holds.

(d) We need only to prove that $\sup_n \|R_n\| < \infty$ implies $(\mathrm{H3})^{\mathrm{s}}$. This is obvious by (b) of Corollary 2.3 or by using the assertions (b) and (c).

(e) By Corollary 1.5 and (c) of Corollary 2.3, (e) holds obviously. □

Example 2.10. Let $X := l^2$, $X_n := \operatorname{span}\{e_1, e_2, \dots, e_n\}$, where

$$e_k = (0, \dots, 0, \underset{\substack{\uparrow \\ k\,\mathrm{th}}}{1}, 0, \dots) \quad (\forall\, k \in \mathbb{N}).$$

We define an operator $T : X \to X$ as follows:

$$Tx := x - \langle x, e\rangle\, e \qquad (x \in X),$$

where $\langle\cdot,\cdot\rangle$ is the inner product on X and

$$e := \frac{\sqrt{3}}{2}e_1 + \frac{\sqrt{3}}{2^2}e_2 + \cdots + \frac{\sqrt{3}}{2^k}e_k + \cdots.$$

This operator satisfies

$$T = T^2 = T^* = T^\dagger \quad \text{and} \quad \mathcal{N}(T) = \operatorname{span}\{e\};$$

that is, T is the orthogonal projection of X onto $e^\perp$. Now, let

$$T_n := T|_{X_n} : X_n \to X.$$

Then we have the following three propositions (whose proofs are in the appendix):

(A.1) $\mathcal{N}(T_n) = \mathcal{N}(T) \cap X_n = \{0\}$, $\overline{\bigcup_{n=1}^{\infty} (\mathcal{N}(T) \cap X_n)} \subsetneqq \mathcal{N}(T)$.

(A.2) $T_n^\dagger y = P_n y - 4^n \langle (I - P_n)\, y, e\rangle\, P_n e \quad (\forall\, y \in X)$.

(A.3) If $y \in X$, then

$$\sup_n \left\|T_n^\dagger y\right\| < \infty \Longleftrightarrow \sup_n 4^n \left|\langle (I - P_n)\, y, e\rangle\right| < \infty,$$

$$T_n^\dagger y \rightharpoonup T^\dagger y \Longleftrightarrow 4^n \langle (I - P_n)\, y, e\rangle \longrightarrow \langle y, e\rangle.$$

Note that by (A.1) condition $(\star)$ does not hold. We take, for example,

$$y = \left(\frac{\sqrt{3}}{4}, \frac{3\sqrt{3}}{4^2}, \dots, \underset{\substack{\uparrow \\ n\,\mathrm{th}}}{\frac{(2^n - 1)\sqrt{3}}{4^n}}, \dots\right)$$

and notice that

$$\langle y, e\rangle = 1 - \frac{3}{7} \quad \text{and} \quad \langle (I - P_n)\, y, e\rangle = \frac{1}{4^n} - \frac{3}{7 \times 8^n}.$$

Then

$$\begin{cases} \sup_n 4^n \left|\langle (I - P_n)\, y, e\rangle\right| = \sup_n 4^n \left(\frac{1}{4^n} - \frac{3}{7\times 8^n}\right) < \infty, \\ 4^n \langle (I - P_n)\, y, e\rangle = 4^n \left(\frac{1}{4^n} - \frac{3}{7\times 8^n}\right) \longrightarrow 1 \neq \langle y, e\rangle. \end{cases}$$

Therefore, by (A.3) we conclude that

$$\sup_n \left\|T_n^\dagger y\right\| < \infty \quad \text{and} \quad \operatorname*{w\text{-}lim}_{n\to\infty} T_n^\dagger y \neq T^\dagger y.$$

This indicates that Theorem 1.9 would be false without condition $(\star)$.

3. Weak perfect convergence for nonprojection methods. To make it clear, the discretization methods with the approximation setting (1.1) and (1.2) are called "generalized projection methods" or "nonprojection methods" if their approximation setting is with $(\mathrm{H1})^{\mathrm{u}}$ or (H1), respectively. For generalized projection methods, by Theorem 1.4, we have

$$(\mathrm{H3})_{\mathrm{p}}^{\mathrm{s}} \iff (\mathrm{H3})_{\mathrm{p}}^{\mathrm{w}} \iff (\mathrm{H2}).$$

But for nonprojection methods $(\mathrm{H3})_{\mathrm{p}}^{\mathrm{w}}$ is neither equivalent to (H2) nor to $(\mathrm{H3})_{\mathrm{p}}^{\mathrm{s}}$, and the relation among the three is

$$(\mathrm{H3})_{\mathrm{p}}^{\mathrm{s}} \implies (\mathrm{H3})_{\mathrm{p}}^{\mathrm{w}} \implies (\mathrm{H2}).$$

In this section, for nonprojection methods we present a criterion for $(\mathrm{H3})_{\mathrm{p}}^{\mathrm{w}}$ by (H2) with other necessary compensated conditions, which is further explained by four examples.

3.1. A criterion for weak perfect convergence. We need a lemma as follows.

LEMMA 3.1. *Suppose that the approximation setting* (1.1) *and* (1.2) *is with* (H1). *Then:*

(a) $\text{s-lim}_{n\to\infty} T_nP_n = T$, $\sup_n \|T_n\|_{\widetilde{n}} = \sup_n \|T_nP_n\| < \infty$;

(b) $\text{s-lim}_{n\to\infty} \mathcal{N}(T_n) \subseteq \mathcal{N}(T) \cap \text{w-}\widetilde{\lim}_{n\to\infty} \mathcal{N}(T_n)$;

(c) $\overline{\mathcal{R}(T)} \subseteq \text{s-lim}_{n\to\infty} \mathcal{R}(T_n) \subseteq \text{w-}\widetilde{\lim}_{n\to\infty} \mathcal{R}(T_n)$;

(d) $\mathcal{G}(T) = \text{s-lim}_{n\to\infty} \mathcal{G}(T_n)$, *where* $\mathcal{G}(\cdot)$ *stands for "graph of";*

(e) $(\mathrm{H2}) \Longrightarrow \text{s-lim}_{n\to\infty} \mathcal{N}(T_n) = \mathcal{N}(T)$;

(f) $\mathcal{N}(T^*) = \{ y \in Y \mid T_n^* Q_n y \rightharpoonup 0 \ (n \to \infty) \}$.

Proof. (a) follows from the fact that for any $x \in X$ there holds

$$\|T_nP_nx - Tx\| \leqslant \|Q_n(T_n - T)P_nx\| + \|T\|\,\|(I - P_n)x\| + \|(I - Q_n)Tx\|.$$

To prove (b), let $x \in \text{s-lim}_{n\to\infty} \mathcal{N}(T_n)$. We need only to show that $x \in \mathcal{N}(T)$. By the definition of $\text{s-lim}_{n\to\infty} \mathcal{N}(T_n)$ there exists a sequence $\{x_n\}$ such that

$$x_n \in \mathcal{N}(T_n) \ (\forall\, n) \quad \text{and} \quad x_n \to x \text{ as } n \to \infty. \tag{3.1}$$

Notice that

$$\|Tx\| \leqslant \|Tx - T_nP_nx\| + M\,\|P_nx - x_n\| \quad (M = \sup_{k\in\mathbb{N}} \|T_k\|_k). \tag{3.2}$$

By putting $n \to \infty$, by (a), it follows from (3.1) and (3.2) that $x \in \mathcal{N}(T)$.

To prove (c), we need only to show that $\mathcal{R}(T) \subseteq \text{s-lim}_{n\to\infty} \mathcal{R}(T_n)$ (note that $\text{s-lim}_{n\to\infty} \mathcal{R}(T_n)$ is closed [4, Lemma 2.7]). Let $y \in \mathcal{R}(T)$. Then there is an $x \in X$ such that $Tx = y$. By (a) we have

$$T_nP_nx \to y \text{ as } n \to \infty,$$

which implies that $y \in \text{s-lim}_{n\to\infty} \mathcal{R}(T_n)$.

(d) follows from (a), because we have the following:

$$(x,y) \in \mathcal{G}(T) \overset{\text{(a)}}{\Longrightarrow} \begin{cases} P_nx \to x \\ T_nP_nx \to y \end{cases} \implies (x,y) \in \underset{n\to\infty}{\text{s-}\lim} \mathcal{G}(T_n);$$

$$(x,y)\in \operatorname*{s\,-\,lim}_{n\to\infty}\mathcal{G}(T_n)\Longrightarrow \exists x_n\in X_n:\begin{cases} x_n\to x\\ T_nx_n\to y\end{cases}$$

$$\overset{\text{(a)}}{\Longrightarrow}\begin{cases} T_nx_n-T_nP_nx\to 0\\ T_nP_nx\to Tx\\ T_nx_n\to y\end{cases}\Longrightarrow Tx=y,\ \text{i.e.,}\ (x,y)\in\mathcal{G}(T)\ .$$

Assume that (H2) holds. Then for any $x\in\mathcal{N}(T)$ there holds

$$P_nx-T_n^{\dagger}T_nP_nx\to x\quad \text{as } n\to\infty\,.$$

Note that, for any $x\in\mathcal{N}(T)$, $P_nx-T_n^{\dagger}T_nP_nx\in\mathcal{N}(T_n)$ $(\forall\ n\in\mathbb{N})$, so that we have $x\in\text{s-lim}_{n\to\infty}\mathcal{N}(T_n)$. Thus, (e) holds.

Finally, because we have the identity

$$\begin{aligned}\langle u,T^*v\rangle=\langle u,T_n^*Q_nv\rangle+\langle u-P_nu,T^*v\rangle+\langle TP_nu,v-Q_nv\rangle\\ +\langle Q_n(T-T_n)P_nu,v\rangle\qquad(\forall u\in X,\,\forall v\in Y),\end{aligned}$$

and because (H1) holds, (f) follows. □

THEOREM 3.2. *Suppose that the approximation setting* (1.1) *and* (1.2) *is with* (H1). *Then the following propositions hold:*

(a) *If* (H2) *and* $\text{w-}\widetilde{\lim}_{n\to\infty}\mathcal{N}(T_n)\subseteq\mathcal{N}(T)+\text{s-lim}_{n\to\infty}\mathcal{N}(T_n)$ *are valid, then:*

(i) $\text{s-lim}_{n\to\infty}\mathcal{N}(T_n)=\text{w-}\widetilde{\lim}_{n\to\infty}\mathcal{N}(T_n)=\mathcal{N}(T)$*; that is,*

$$\operatorname*{s\,-\,lim}_{n\to\infty}P_{\mathcal{N}(T_n)}=P_{\mathcal{N}(T)}\,;$$

(ii) $\mathcal{R}(T)$ *is closed, and*

$$\begin{aligned}\left\|T_n^{\dagger}Q_nQ_{\mathcal{R}(T)}y-T^{\dagger}y\right\|\leqslant\left\|T_n^{\dagger}Q_nT-I\right\|\operatorname{dist}\left(T^{\dagger}y,X_n\right)+\left\|P_{\mathcal{N}(T_n)}T^{\dagger}y\right\|\\ +\left\|T_n^{\dagger}\right\|_n\left\|Q_n(T_n-T)P_nT^{\dagger}y\right\|\qquad(\forall\ y\in Y)\,;\end{aligned}$$

(iii) *there hold*

$$\operatorname*{s\,-\,lim}_{n\to\infty}T_n^{\dagger}Q_nQ_{\mathcal{R}(T)}=T^{\dagger}\,,\quad \operatorname*{s\,-\,lim}_{n\to\infty}Q_{\mathcal{R}(T_n)}Q_{\mathcal{R}(T)}=Q_{\mathcal{R}(T)}.$$

(b) *Criterion for* $(\text{H3})_{\text{p}}^{\text{w}}$*:*

$$(\text{H3})_{\text{p}}^{\text{w}}\Longleftrightarrow\begin{cases}\text{(H2) }\textit{and}\ \operatorname*{w\,-\,\widetilde{lim}}_{n\to\infty}\mathcal{N}(T_n)\subseteq\mathcal{N}(T)+\operatorname*{s\,-\,lim}_{n\to\infty}\mathcal{N}(T_n)\,,\\ \operatorname*{w\,-\,lim}_{n\to\infty}T_n^{\dagger}Q_ny=0\ \textit{if}\ \operatorname*{w\,-\,lim}_{n\to\infty}T_n^{*}Q_ny=0\,.\end{cases}$$

Proof. The proof will be carried out in two steps.

Step 1 (to prove (a)). By (e) in Lemma 3.1, (H2) together with $\text{w-}\widetilde{\lim}_{n\to\infty}\mathcal{N}(T_n)\subseteq\mathcal{N}(T)+\text{s-lim}_{n\to\infty}\mathcal{N}(T_n)$ implies that

$$\operatorname*{s\,-\,lim}_{n\to\infty}\mathcal{N}(T_n)=\operatorname*{w\,-\,\widetilde{lim}}_{n\to\infty}\mathcal{N}(T_n)=\mathcal{N}(T)\ . \tag{3.3}$$

Assume that $\{y^{(m)}\}\subseteq\mathcal{R}(T)$ satisfies $\text{s-lim}_{n\to\infty}y^{(m)}=y$. By (d) in Lemma 3.1, $\mathcal{G}(T)=\text{s-lim}_{n\to\infty}\mathcal{G}(T_n)$. Hence, for every $m\in\mathbb{N}$ there exists a sequence $\{(x_n^{(m)},y_n^{(m)})\}_{n=1}^{\infty}$ in $X\times Y$ such that

$$y_n^{(m)}=T_n\left(x_n^{(m)}\right)\quad(\forall\ n\in\mathbb{N}) \tag{3.4}$$

and

$$x_n^{(m)} \to T^\dagger\left(y^{(m)}\right), \quad y_n^{(m)} \to y^{(m)} \quad \text{as } n \to \infty\,. \tag{3.5}$$

It follows from (3.4) and (3.5) that, for every $m \in \mathbb{N}$,

$$\sup_n \left\|T_n^\dagger\left(y_n^{(m)}\right)\right\| \le \sup_n \left\|x_n^{(m)}\right\| < \infty\,. \tag{3.6}$$

For every $m \in \mathbb{N}$, because the sequence $\{T_n^\dagger(y_n^{(m)})\}_{n=1}^\infty$ is bounded (by (3.6)), $\{T_n^\dagger(y_n^{(m)})\}_{n=1}^\infty$ contains a subsequence $\{T_{n_j}^\dagger(y_{n_j}^{(m)})\}_{j=1}^\infty$ which converges weakly to an element $x^{(m)}$ of X by the Eberlein–Shmulyan theorem [30, p. 141], and therefore, with the use of (3.5), we have

$$x_{n_j}^{(m)} - T_{n_j}^\dagger\left(y_{n_j}^{(m)}\right) \rightharpoonup T^\dagger\left(y^{(m)}\right) - x^{(m)} \quad \text{as } j \to \infty \tag{3.7}$$

and (by [30, Theorem 1, p. 120])

$$\left\|x^{(m)}\right\| \le \varliminf_{j\to\infty} \left\|T_{n_j}^\dagger\left(y_{n_j}^{(m)}\right)\right\|\,. \tag{3.8}$$

Notice that $\{x_{n_j}^{(m)} - T_{n_j}^\dagger(y_{n_j}^{(m)})\}_{m=1}^\infty \subseteq \mathcal{N}(T_{n_j})$ for every $j \in \mathbb{N}$ (by (3.4)), so it follows from (3.7) that

$$\left\{T^\dagger\left(y^{(m)}\right) - x^{(m)}\right\}_{m=1}^\infty \subseteq \operatorname{w}\text{-}\widetilde{\varliminf_{n\to\infty}}\mathcal{N}\left(T_n\right).$$

This together with (3.3) implies that

$$T\left(x^{(m)}\right) = y^{(m)} \quad (\forall\ m \in \mathbb{N})\,. \tag{3.9}$$

Note that from (3.8) and (3.5) it follows that

$$\begin{aligned}\sup_m \left\|x^{(m)}\right\| &\le \sup_m \varliminf_{j\to\infty} \left\|T_{n_j}^\dagger\left(y_{n_j}^{(m)}\right)\right\| \\ &\le \sup_n \left\|T_n^\dagger\right\| \sup_m \varliminf_{j\to\infty} \left\|y_{n_j}^{(m)}\right\| \\ &= \sup_n \left\|T_n^\dagger\right\| \sup_m \left\|y^{(m)}\right\| < \infty.\end{aligned}$$

This implies that $\{x^{(m)}\}$ contains a subsequence $\{x^{(m_j)}\}_{j=1}^\infty$ which converges weakly to some $x^{(\infty)}$ in X. Furthermore, note that, by the Mazur lemma (see [30, Theorem 2 (Mazur), p. 120] or [1, Theorem 3.2.11, p. 109]), for every $j \in \mathbb{N}$ there exists a convex combination

$$x_j = \sum_{i=1}^{k_j} \alpha_i^{(j)} x^{(m_{j+i})} \quad \left(\alpha_i^{(j)} \ge 0,\ \sum_{i=1}^{k_j} \alpha_i^{(j)} = 1\right) \tag{3.10}$$

such that

$$\left\|x^{(\infty)} - x_j\right\| \le \frac{1}{j}. \tag{3.11}$$

By (3.9) and (3.10), we have

$$\begin{aligned}\|T(x_j)-y\| &= \left\|\sum_{i=1}^{k_j}\alpha_i^{(j)}T\left(x^{(m_{j+i})}\right)-y\right\| \\ &= \left\|\sum_{i=1}^{k_j}\alpha_i^{(j)}y^{(m_{j+i})}-y\right\| \\ &\le \sum_{i=1}^{k_j}\alpha_i^{(j)}\left\|y^{(m_{j+i})}-y\right\| \\ &\le \sup_{m\ge j}\left\|y^{(m)}-y\right\|.\end{aligned} \tag{3.12}$$

By putting $j\to\infty$, by (3.11) and (3.12) we conclude that $T\left(x^{(\infty)}\right)=y$ and therefore $y\in\mathcal{R}(T)$. Thus, $\mathcal{R}(T)$ is closed. This implies that $T^\dagger\in\mathcal{B}(Y,X)$.

Next, because

$$\begin{aligned}&T_n^\dagger Q_n Q_{\overline{\mathcal{R}(T)}}-T^\dagger = T_n^\dagger Q_n Q_{\overline{\mathcal{R}(T)}}-T_n^\dagger T_n P_n T^\dagger+T_n^\dagger T_n P_n T^\dagger-T^\dagger T T^\dagger \\ &= T_n^\dagger Q_n\left(Q_{\overline{\mathcal{R}(T)}}-TP_nT^\dagger\right)-T_n^\dagger Q_n(T_n-T)P_nT^\dagger+\left(T_n^\dagger T_nP_n-T^\dagger T\right)T^\dagger \\ &= T_n^\dagger Q_nT(I_X-P_n)T^\dagger-T_n^\dagger Q_n(T_n-T)P_nT^\dagger+\left(P_n-P_{\mathcal{N}(T_n)}-I_X+P_{\mathcal{N}(T)}\right)T^\dagger \\ &= \left(T_n^\dagger Q_nT-I_X\right)(I_X-P_n)T^\dagger-T_n^\dagger Q_n(T_n-T)P_nT^\dagger-P_{\mathcal{N}(T_n)}T^\dagger,\end{aligned}$$

and because $\mathcal{R}(T)$ is closed, we have

$$\begin{aligned}\left\|T_n^\dagger Q_nQ_{\mathcal{R}(T)}y-T^\dagger y\right\| \le{}& \left\|T_n^\dagger Q_nT-I\right\|\operatorname{dist}\left(T^\dagger y,X_n\right)+\left\|P_{\mathcal{N}(T_n)}T^\dagger y\right\| \\ &+\left\|T_n^\dagger\right\|_n\left\|Q_n(T_n-T)P_nT^\dagger y\right\| \qquad (\forall\, y\in Y).\end{aligned} \tag{3.13}$$

Since (3.3) gives (by Lemma 2.1 or Remark 1.6)

$$\operatorname*{s-lim}_{n\to\infty}P_{\mathcal{N}(T_n)}=P_{\mathcal{N}(T)},$$

and since (by (H1) and (H2))

$$\lim_{n\to\infty}\|P_nx-x\|=\lim_{n\to\infty}\|Q_n(T_n-T)P_nx\|=0,\quad \sup_n\left\|T_n^\dagger\right\|_n<\infty,$$

it follows from (3.13) that

$$\operatorname*{s-lim}_{n\to\infty}T_n^\dagger Q_nQ_{\mathcal{R}(T)}=T^\dagger.$$

This implies, by (a) in Lemma 3.1, that

$$\operatorname*{s-lim}_{n\to\infty}T_nT_n^\dagger Q_nQ_{\mathcal{R}(T)}=TT^\dagger,\quad \text{that is,}\quad \operatorname*{s-lim}_{n\to\infty}Q_{\mathcal{R}(T_n)}Q_{\mathcal{R}(T)}=Q_{\mathcal{R}(T)}.$$

Thus, (a) follows.

Step 2 (to prove (b)). Assume that $(\mathrm{H3})^{\mathrm{w}}_{\mathrm{p}}$ holds. Then (H2) holds by the uniform boundedness principle. Because for any $x, u \in X$ there holds

$$\begin{aligned}\left|\left\langle (T_nP_n)^\dagger (T_nP_n)\, x - T^\dagger Tx, u\right\rangle\right| &\leqslant \left|\left\langle (T_nP_n)^\dagger Q_n (T_n - T) P_n x, u\right\rangle\right| \\ &\quad + \left|\left\langle (T_nP_n)^\dagger Q_n T P_n x - T^\dagger Tx, u\right\rangle\right| \\ &\leqslant \left|\left\langle T_n^\dagger Q_n (T_n - T) P_n x, u\right\rangle\right| + \left|\left\langle T_n^\dagger Q_n T P_n x - T_n^\dagger Q_n Tx, u\right\rangle\right| \\ &\quad + \left|\left\langle T_n^\dagger Q_n Tx - T^\dagger Tx, u\right\rangle\right| \\ &\leqslant \left\|T_n^\dagger\right\|_n \|Q_n (T_n - T) P_n x\| \|u\| + \left\|T_n^\dagger Q_n\right\| \|T\| \|P_n x - x\| \|u\| \\ &\quad + \left|\left\langle \left(T_n^\dagger Q_n - T^\dagger\right) Tx, u\right\rangle\right| ,\end{aligned}$$

we obtain

$$(T_nP_n)^\dagger (T_nP_n) \rightharpoonup T^\dagger T \quad \text{as} \quad n \to \infty ,$$

so that

$$P_{\mathcal{N}(T_nP_n)} \rightharpoonup P_{\mathcal{N}(T)} \quad \text{as} \quad n \to \infty . \tag{3.14}$$

Notice that

$$P_{\mathcal{N}(T_nP_n)} = I - P_n + P_{\mathcal{N}(T_n)} .$$

By Lemma 2.1, this with (3.14) implies that

$$\operatorname*{s\text{-}lim}_{n\to\infty} P_{\mathcal{N}(T_n)} = P_{\mathcal{N}(T)}, \quad \text{i.e.,} \quad \operatorname*{s\text{-}lim}_{n\to\infty} \mathcal{N}(T_n) = \operatorname*{w\text{-}\widetilde{lim}}_{n\to\infty} \mathcal{N}(T_n) = \mathcal{N}(T) . \tag{3.15}$$

This implies that

$$\operatorname*{w\text{-}\widetilde{lim}}_{n\to\infty} \mathcal{N}(T_n) \subseteq \mathcal{N}(T) + \operatorname*{s\text{-}lim}_{n\to\infty} \mathcal{N}(T_n) . \tag{3.16}$$

Next, let $y \in Y$ satisfy $T_n^* Q_n y \rightharpoonup 0$. Then $y \in \mathcal{N}(T^*)$ by (f) in Lemma 3.1; that is, $y \in \mathcal{N}\left(T^\dagger\right)$. Therefore, it follows from $(\mathrm{H3})^{\mathrm{w}}_{\mathrm{p}}$ that

$$T_n^\dagger Q_n y \rightharpoonup 0 \quad \text{if} \quad T_n^* Q_n y \rightharpoonup 0 . \tag{3.17}$$

So $(\mathrm{H3})^{\mathrm{w}}_{\mathrm{p}}$ implies (H2), (3.16), and (3.17).

Inversely, assume (H2), (3.16), and (3.17). Then, by (a), (H2) together with (3.16) implies that $\mathcal{R}(T)$ is closed and that (3.13) and (3.15) hold. By (f) in Lemma 3.1, (3.17) implies that

$$T_n^\dagger Q_n y \rightharpoonup 0 \quad \text{if} \quad y \in \mathcal{N}(T^*) = \mathcal{R}(T)^\perp . \tag{3.18}$$

Hence, it follows from (3.13), (3.15), and (3.18) that

$$T_n^\dagger Q_n y \rightharpoonup T^\dagger y \quad \text{if} \quad y \in \mathcal{D}\left(T^\dagger\right) ,$$

that is, $(\mathrm{H3})^{\mathrm{w}}_{\mathrm{p}}$ holds. □

Corollary 3.3. *Suppose that the approximation setting* (1.1) *and* (1.1) *is with* (H1). *If* $X = Y$, $X_n = Y_n$, *and* T *and* T_n *are both self-adjoint, then*

$$(\mathrm{H3})^{\mathrm{s}}_{\mathrm{p}} \Longleftrightarrow (\mathrm{H3})^{\mathrm{w}}_{\mathrm{p}} \Longleftrightarrow \left\{(\mathrm{H2}) \text{ and } \operatorname*{w\text{-}\widetilde{lim}}_{n\to\infty} \mathcal{N}(T_n) \subseteq \mathcal{N}(T) + \operatorname*{s\text{-}lim}_{n\to\infty} \mathcal{N}(T_n)\right\} .$$

Proof. By Theorem 3.2, we need only to show that

$$\left\{\text{(H2) and } \operatorname*{w\text{-}\widetilde{lim}}_{n\to\infty}\mathcal{N}(T_n)\subseteq\mathcal{N}(T)+\operatorname*{s\text{-}lim}_{n\to\infty}\mathcal{N}(T_n)\right\}\Longrightarrow(\mathrm{H3})^{\mathrm{s}}_{\mathrm{p}}\,.$$

Assume that the left side of the above implication is valid. Then $\mathcal{R}(T)$ is closed, and (3.3) holds, that is,

$$\operatorname*{w\text{-}\widetilde{lim}}_{n\to\infty}\mathcal{N}(T_n)=\operatorname*{s\text{-}lim}_{n\to\infty}\mathcal{N}(T_n)=\mathcal{N}(T)\,.$$

Since T_n $(n\in\mathbb{N})$ and T are all self-adjoint, we hence have

$$\operatorname*{w\text{-}\widetilde{lim}}_{n\to\infty}\mathcal{R}(T_n)^{\perp_n}=\operatorname*{s\text{-}lim}_{n\to\infty}\mathcal{R}(T_n)^{\perp_n}=\mathcal{R}(T)^{\perp}\,,$$

that is (by Lemma 2.1),

$$\operatorname*{s\text{-}lim}_{n\to\infty}\left(P_n-P_{\mathcal{R}(T_n)}\right)=I-P_{\mathcal{R}(T)}\,.$$

Here we note that $P_n-P_{\mathcal{R}(T_n)}$ is the orthogonal projection from X onto $\mathcal{R}(T_n)^{\perp_n}$. Hence, we obtain

$$\operatorname*{s\text{-}lim}_{n\to\infty}P_{\mathcal{R}(T_n)}=P_{\mathcal{R}(T)}\,,$$

that is (by Lemma 2.1),

$$\operatorname*{w\text{-}\widetilde{lim}}_{n\to\infty}\mathcal{R}(T_n)=\operatorname*{s\text{-}lim}_{n\to\infty}\mathcal{R}(T_n)=\mathcal{R}(T)\,. \tag{3.19}$$

Now $(\mathrm{H3})^{\mathrm{s}}_{\mathrm{p}}$ follows from (H2), (3.3), and (3.19) by Theorem 1.3. ∎

3.2. Examples on weak perfect convergence. In this subsection we will give some examples to indicate that the three conditions of characterizing $(\mathrm{H3})^{\mathrm{w}}_{\mathrm{p}}$ in Theorem 3.2 are independent. The three conditions are

- (H2);
- $\text{w-}\widetilde{\lim}_{n\to\infty}\mathcal{N}(T_n)\subseteq\mathcal{N}(T)+\text{s-}\lim_{n\to\infty}\mathcal{N}(T_n)$, *denoted by* (CH1);
- $\text{w-}\lim_{n\to\infty}T_n^{\dagger}Q_ny=0$ if $\text{w-}\lim_{n\to\infty}T_n^{*}Q_ny=0$, *denoted by* (CH2).

In fact, for the approximation setting (1.1) and (1.2) with (H1), the following three cases arise:

$$\text{case 1: }\begin{cases}\text{(H2)},\\ \text{(CH1)},\\ \neg\text{(CH2)};\end{cases}\qquad\text{case 2: }\begin{cases}\text{(H2)},\\ \neg\text{(CH1)},\\ \text{(CH2)};\end{cases}\qquad\text{case 3: }\begin{cases}\neg\text{(H2)},\\ \text{(CH1)},\\ \text{(CH2)}.\end{cases}$$

Here the negation of a statement P is written by $\neg P$.

Example 3.4. We define $T:l^2\to l^2$ and $T_n:l^2\to l^2$ $(n\in\mathbb{N})$ as follows:

$$T(\mathbf{x}):=(0,x_1,x_2,\dots)\,,$$

$$T_n(\mathbf{x}):=(\underset{\substack{\uparrow\\1\text{st}}}{x_n},\underset{\substack{\uparrow\\2\text{nd}}}{x_1-x_n},\underset{\substack{\uparrow\\3\text{rd}}}{x_2},\dots,\underset{\substack{\uparrow\\n\text{th}}}{x_{n-1}},\underset{\substack{\uparrow\\(n+1)\text{th}}}{x_{n+1}},\dots),$$

where $\mathbf{x}=(x_1,x_2,\dots)\in l^2$. Then

(i) $T_n\xrightarrow{\mathrm{s}}T$; that is, the approximation setting is with (H1);

(ii) $\mathcal{N}(T_n) = \{\mathbf{0}\}$, $\mathcal{R}(T_n) = l^2$, and

$$T_n^\dagger(\mathbf{y}) = (y_1 + y_2,\ y_3, \dots,\ \underset{(n-1)\text{th}}{y_n},\ y_1,\ \underset{n\text{th}}{y_{n+1}}, \dots);$$

(iii) $\sup_n \left\|T_n^\dagger\right\|_n < \infty$; that is, the setting is with (H2);
(iv) $\mathcal{N}(T) = \{\mathbf{0}\}$, $\mathcal{R}(T) = \mathbf{e}_1^\perp$, where $\mathbf{e}_1 = (1, 0, 0, \dots)$, and

$$T^\dagger(\mathbf{y}) = (y_2, y_3, \dots);$$

(v) $\widetilde{\text{w-lim}}_{n\to\infty}\mathcal{N}(T_n) = \text{s-lim}_{n\to\infty}\mathcal{N}(T_n) = \mathcal{N}(T)$, and the setting is with (CH1);
(vi) $T_n^*\mathbf{y} \rightharpoonup \mathbf{0}$ if and only if $\mathbf{y} \in \text{span}\{\mathbf{e}_1\}$, where

$$T_n^*(\mathbf{y}) = (y_2, \dots, y_n, y_1 - y_2, y_{n+1}, \dots);$$

it is clear that

$$T_n^\dagger(y_1, 0, 0, \dots) = (y_1,\ \underset{1\text{st}}{0},\ \underset{2\text{nd}}{0}, \dots,\ \underset{(n-1)\text{th}}{0},\ y_1,\ \underset{n\text{th}}{0},\ \underset{(n+1)\text{th}}{0}, \dots),$$

and therefore the setting is not with (CH2);
(vii) $T_n^\dagger \rightharpoonup T^\dagger + P_{\text{span}\{\mathbf{e}_1\}}$.

Example 3.5. Let $T, T_n \in \mathcal{B}(l^2)$ $(n \in \mathbb{N})$ be defined by

$$T(\mathbf{x}) := \mathbf{x}, \quad T_n(\mathbf{x}) := (x_1 - x_{n+1}, x_2, \dots, x_n, 0, \dots),$$

where $\mathbf{x} = (x_1, x_2, \dots) \in l^2$. Then
(i) $T_n \xrightarrow{\text{s}} T$; that is, the approximation setting is with (H1);
(ii)
$$\begin{cases} \mathcal{N}(T_n) = \{\mathbf{x} \mid x_1 = x_{n+1},\ x_2 = \cdots = x_n = 0\}, \\ \mathcal{R}(T_n) = \{\mathbf{x} \mid x_k = 0\ \ (k \geqslant n+1)\}, \\ T_n^\dagger(\mathbf{x}) = (\frac{1}{2}x_1, x_2, \dots, x_n, -\frac{1}{2}x_1, 0, 0, \dots); \end{cases}$$
(iii) $\sup_n \left\|T_n^\dagger\right\|_n < \infty$; that is, the setting is with (H2);
(iv) $\mathcal{N}(T) = \{\mathbf{0}\}$, $\mathcal{R}(T) = l^2$, and $T^\dagger = T$;
(v) $\widetilde{\text{w-lim}}_{n\to\infty}\mathcal{N}(T_n) \supsetneqq \mathcal{N}(T) = \text{s-lim}_{n\to\infty}\mathcal{N}(T_n)$, and the setting is not with (CH1);
(vi) $T_n^*(\mathbf{y}) = (y_1, y_2 \dots, y_n, -y_1, 0, 0, \dots)$; hence we have

$$T_n^*\mathbf{y} \rightharpoonup \mathbf{0} \iff \mathbf{y} = \mathbf{0},$$

and the setting is with (CH2);
(vii) $T_n^\dagger \rightharpoonup T^\dagger - \frac{1}{2}P_{\text{span}\{\mathbf{e}_1\}}$, where $\mathbf{e}_1 = (1, 0, 0, \dots)$.

Example 3.6. Take Example 2.10. Then
(i) the approximation setting is with $(\text{H1})^{\text{u}}$ (uniform consistency);
(ii) $\mathcal{N}(T_n) = \{0\}$, $\mathcal{R}(T_n) = T(X_n)$, and

$$T_n^\dagger y = P_n y - 4^n \langle (I - P_n)\, y, e\rangle\, P_n e \quad (\forall y \in X);$$

(iii) $\sup_n \left\|T_n^\dagger\right\|_n = \infty$ (in fact, $\sup_n \left\|T_n^\dagger T\right\| = \infty$ since $T_n^\dagger T = T_n^\dagger$) because

$$\sup_n \left\|T_n^\dagger\right\|_n \geqslant \left\|T_m^\dagger e_{m+1}\right\| = 2^{m-1}\left(1 - \frac{1}{4^m}\right)\sqrt{3} \to \infty \quad \text{as} \quad m \to \infty;$$

hence the setting is not with (H2);

(iv) $\mathcal{N}(T) = \operatorname{span}\{e\}$, $\mathcal{R}(T) = e^{\perp}$, and $T^{\dagger} = T^{*} = T$;

(v) $\widetilde{\text{w-lim}}_{n\to\infty}\mathcal{N}(T_n) = \text{s-lim}_{n\to\infty}\mathcal{N}(T_n) \subsetneqq \mathcal{N}(T)$, and the setting is with (CH1);

(vi) $T_n^{*} = P_n T$ and therefore

$$T_n^{*} y \rightharpoonup 0 \iff y \in \mathcal{N}(T) \ (\Longleftrightarrow T_n^{*} y \to 0) \ ,$$

and the setting is with (CH2) (in fact, $T_n^{\dagger} Q_n y = 0$ if $y \in \mathcal{N}(T)$; here $Q_n = I$);

(vii) $\widetilde{\text{w-lim}}_{n\to\infty}\mathcal{R}(T_n) = \text{s-lim}_{n\to\infty}\mathcal{R}(T_n) = \mathcal{R}(T)$;

(viii) the setting is not with $(\text{H3})^{\text{w}}$.

The above examples, together with Theorem 3.2, indicate that, for characterizing $(\text{H3})_{\text{p}}^{\text{w}}$ of the approximation setting (1.1) and (1.2) with (H1), the conditions (H2), (CH1), and (CH2) are not only necessary and sufficient but also independent.

A natural problem left is: For such an approximation setting, is there a case that $(\text{H3})_{\text{p}}^{\text{w}}$ holds and $(\text{H3})_{\text{p}}^{\text{s}}$ fails? The answer is yes.

Example 3.7. Let $T, T_n \in \mathcal{B}\left(l^2\right)$ $(n \in \mathbb{N})$ be defined by

$$T\mathbf{x} = (0, x_1, x_2, \dots) \ ,$$

$$T_n\mathbf{x} = (x_n, x_1, \dots, x_{n-1}, x_{n+1}, \dots) \ ,$$

where $\mathbf{x} = (x_1, x_2, \dots) \in l^2$. Then

(i) $T_n \overset{\text{s}}{\to} T$; that is, the approximation setting is with (H1);

(ii) $\mathcal{N}(T_n) = \{\mathbf{0}\}$, $\mathcal{R}(T_n) = l^2$, and

$$T_n^{\dagger}\mathbf{y} = (\underset{\substack{\uparrow\\ 1\text{st}}}{y_2}, \underset{\substack{\uparrow\\ 2\text{nd}}}{y_3}, \dots, \underset{\substack{\uparrow\\ (n-1)\text{th}}}{y_n}, \underset{\substack{\uparrow\\ n\text{th}}}{y_1}, \underset{\substack{\uparrow\\ (n+1)\text{th}}}{y_{n+1}}, \dots);$$

(iii) $\sup_n \left\|T_n^{\dagger}\right\|_n < \infty$; that is, the setting is with (H2);

(iv) $\mathcal{N}(T) = \{\mathbf{0}\}$, $\mathcal{R}(T) = \mathbf{e}_1^{\perp}$, where $\mathbf{e}_1 = (1, 0, 0, \dots)$, and

$$T^{\dagger}\mathbf{y} = (y_2, y_3, \dots) \ ;$$

(v) $\widetilde{\text{w-lim}}_{n\to\infty}\mathcal{N}(T_n) = \text{s-lim}_{n\to\infty}\mathcal{N}(T_n) = \mathcal{N}(T)$, and the setting is with (CH1);

(vi) $T_n^{*}\mathbf{y} = T_n^{\dagger}\mathbf{y}$, and therefore the setting is with (CH2);

(vii) $\widetilde{\text{w-lim}}_{n\to\infty}\mathcal{R}(T_n) = \text{s-lim}_{n\to\infty}\mathcal{R}(T_n) \supsetneqq \overline{\mathcal{R}(T)}$.

Note: By Theorems 1.3 and 3.2, the approximation setting is not with $(\text{H3})_{\text{p}}^{\text{s}}$, but it is with $(\text{H3})_{\text{p}}^{\text{w}}$.

4. Conclusions. Fundamental results on projection methods of Moore–Penrose inverses for well- and ill-posed problems have been obtained (see Corollary 1.5 and Theorems 2.2, 2.6, and 2.8 with their corollaries). Therefore, Tables 4.1 and 4.2 are produced. We come to the following conclusions:

- Stable projection methods (whose approximation setting is with (H2)) apply only to well-posed problems, which means that the projection methods applied to ill-posed problems are unstable (whose approximation setting must be with $\lim_{n\to\infty} \left\|T_n^{\dagger}\right\|_n = \infty$ by Corollary 1.5)!
- The projection methods that can cope with ill-posed problems have first to be with scheme regularization, which means that their approximation setting has to provide a regularization scheme (on the basis of which one chooses a regular strategy). Table 4.1 indicates that the scheme regularization of projection methods should have four degrees:

TABLE 4.1
Projection methods.

$(\mathrm{H3})^{\mathrm{s}}_{\mathrm{p}}$	$\Rightarrow$	$(\mathrm{H3})^{\mathrm{s}}$	$\Rightarrow$	$T_n^\dagger Q_n Q_{\mathcal{R}(T)^\perp} \xrightarrow{\mathrm{s}} 0$
$\Updownarrow$		$\Downarrow$		
$(\mathrm{H3})^{\mathrm{w}}_{\mathrm{p}}$	$\Rightarrow$	$(\mathrm{H3})^{\mathrm{w}}$	$\Rightarrow$	$Q_{\mathcal{R}(T_n)} \xrightarrow{\mathrm{s}} Q_{\overline{\mathcal{R}(T)}}$
$\Updownarrow$		$\Downarrow$		$\Downarrow$
(H2)	$\Rightarrow$	$\sup_n \|R_n\| < \infty$	$\Rightarrow$	$Q_{\mathcal{R}(T_n)} Q_{\overline{\mathcal{R}(T)}} \xrightarrow{\mathrm{s}} Q_{\overline{\mathcal{R}(T)}}$
$\Downarrow$		$\Downarrow$		
$\overline{\mathcal{R}(T)} = \mathcal{R}(T)$		$P_{\mathcal{N}(T_n)} \xrightarrow{\mathrm{s}} P_{\mathcal{N}(T)}$	$\Rightarrow$	$P_{\mathcal{N}(T_n)} T^\dagger \xrightarrow{\mathrm{s}} 0$ on $\mathcal{D}(T^\dagger)$

Note: (1) $\sup_n \|R_n\| < \infty \Longleftrightarrow T_n^\dagger Q_n Q_{\overline{\mathcal{R}(T)}} y \to T^\dagger y \ \ (y \in \mathcal{D}(T^\dagger))$;
(2) If $\mathcal{R}(T_n) \subseteq \overline{\mathcal{R}(T)}$ $(n \in \mathbb{N})$, $(\mathrm{H3})^{\mathrm{s}} \iff (\mathrm{H3})^{\mathrm{w}} \iff \sup_n \|R_n\| < \infty$.

TABLE 4.2
Special projection methods.

LS	$(\mathrm{H3})^{\mathrm{s}}_{\mathrm{p}}$	$\Rightarrow$	$(\mathrm{H3})^{\mathrm{s}}$	$T_n^\dagger Q_{\mathcal{R}(T)^\perp} = 0$ [**T**]
	$\Updownarrow$		$\Updownarrow$	
	$(\mathrm{H3})^{\mathrm{w}}_{\mathrm{p}}$	$\Rightarrow$	$(\mathrm{H3})^{\mathrm{w}}$	$Q_{\mathcal{R}(T_n)} \xrightarrow{\mathrm{s}} Q_{\overline{\mathcal{R}(T)}}$ [**T**]
	$\Updownarrow$		$\Updownarrow$	
	(H2)	$\Rightarrow$	$\sup_n \|R_n\| < \infty$	$Q_{\mathcal{R}(T_n)} Q_{\overline{\mathcal{R}(T)}} \xrightarrow{\mathrm{s}} Q_{\overline{\mathcal{R}(T)}}$ [**T**]
	$\Downarrow$		$\Downarrow$	
	$\overline{\mathcal{R}(T)} = \mathcal{R}(T)$		$P_{\mathcal{N}(T_n)} \xrightarrow{\mathrm{s}} P_{\mathcal{N}(T)}$	$P_{\mathcal{N}(T_n)} T^\dagger = 0$ [**T**]
DLS	$(\mathrm{H3})^{\mathrm{s}}_{\mathrm{p}}$		$(\mathrm{H3})^{\mathrm{s}}$ [**T**]	$T_n^\dagger Q_n Q_{\mathcal{R}(T)^\perp} = 0$ [**T**]
	$\Updownarrow$			
	$(\mathrm{H3})^{\mathrm{w}}_{\mathrm{p}}$		$(\mathrm{H3})^{\mathrm{w}}$ [**T**]	$Q_{\mathcal{R}(T_n)} \xrightarrow{\mathrm{s}} Q_{\overline{\mathcal{R}(T)}}$ [**T**]
	$\Updownarrow$			
	(H2)		$\sup_n \|R_n\| \leqslant 1$ [**T**]	$Q_{\mathcal{R}(T_n)} Q_{\overline{\mathcal{R}(T)}} \xrightarrow{\mathrm{s}} Q_{\overline{\mathcal{R}(T)}}$ [**T**]
	$\Updownarrow$			
	$\overline{\mathcal{R}(T)} = \mathcal{R}(T)$		$P_{\mathcal{N}(T_n)} \xrightarrow{\mathrm{s}} P_{\mathcal{N}(T)}$ [**T**]	$P_{\mathcal{N}(T_n)} T^\dagger \xrightarrow{\mathrm{s}} 0$ [**T**]
GDLS	$(\mathrm{H3})^{\mathrm{s}}_{\mathrm{p}}$	$\Rightarrow$	$(\mathrm{H3})^{\mathrm{s}}$	$\Rightarrow$ $T_n^\dagger Q_n Q_{\mathcal{R}(T)^\perp} \xrightarrow{\mathrm{s}} 0$
	$\Updownarrow$		$\Downarrow$	
	$(\mathrm{H3})^{\mathrm{w}}_{\mathrm{p}}$	$\Rightarrow$	$(\mathrm{H3})^{\mathrm{w}}$	$\Rightarrow$ $Q_{\mathcal{R}(T_n)} \xrightarrow{\mathrm{s}} Q_{\overline{\mathcal{R}(T)}}$
	$\Updownarrow$			
	(H2)		$\sup_n \|R_n\| \leqslant 1$ [**T**]	$Q_{\mathcal{R}(T_n)} Q_{\overline{\mathcal{R}(T)}} \xrightarrow{\mathrm{s}} Q_{\overline{\mathcal{R}(T)}}$ [**T**]
	$\Downarrow$			
	$\overline{\mathcal{R}(T)} = \mathcal{R}(T)$		$P_{\mathcal{N}(T_n)} \xrightarrow{\mathrm{s}} P_{\mathcal{N}(T)}$ [**T**]	$P_{\mathcal{N}(T_n)} T^\dagger \xrightarrow{\mathrm{s}} 0$ [**T**]
Note	**T** := true (always holds).			

– "$P_{\mathcal{N}(T_n)} \xrightarrow{\mathrm{s}} P_{\mathcal{N}(T)}$"-regularization, *denoted by* (R0)-regularization;
– "$\sup_n \|R_n\| < \infty$"-regularization, *denoted by* (R1)-regularization;
– $(\mathrm{H3})^{\mathrm{w}}$-regularization;
– $(\mathrm{H3})^{\mathrm{s}}$-regularization.

It remains unknown whether the four degrees of scheme regularization occur. But it is certain that the case without any scheme regularization appears. See Examples 4.1, 4.2 (below), and 3.6. Besides, a conjecture is proposed (Conjecture 4.3). As special cases, by Table 4.2, the least-squares methods (LS) have two degrees of scheme regularization: (R0)- and $(\mathrm{H3})^{\mathrm{s}}$-regularization; the generalized dual least-squares methods (GDLS) have three degrees: (R1)-, $(\mathrm{H3})^{\mathrm{w}}$-, and $(\mathrm{H3})^{\mathrm{s}}$-regularization (the first one always holds);

the dual least-squares methods (DLS) are always with the highest degree: $(\mathrm{H3})^{\mathrm{s}}$-regularization. The criteria for $(\mathrm{H3})^{\mathrm{w}}$- and $(\mathrm{H3})^{\mathrm{s}}$-regularization are given by Theorem 2.6.

- Theorem 2.2 provides a general error estimate on projection methods of Moore–Penrose inverses: For all $b \in \mathcal{D}\left(T^{\dagger}\right)$,

$$\left\|T_n^{\dagger} Q_n Q_{\overline{\mathcal{R}(T)}} b - T^{\dagger} b\right\| \leqslant \left\|P_{\mathcal{N}(T_n)} T^{\dagger} b\right\| + \|R_n - I_X\| \operatorname{dist}\left(T^{\dagger} b, X_n\right). \tag{4.1}$$

  It follows that, for all $b \in \mathcal{D}\left(T^{\dagger}\right)$ and $b_n \in Y_n$,

$$\begin{aligned}\left\|T_n^{\dagger} b_n - T^{\dagger} b\right\| &\leqslant \left\|T_n^{\dagger}\right\|_n \left\|Q_n\left(b_n - b\right)\right\| + \left\|T_n^{\dagger} Q_n Q_{\mathcal{R}(T)^{\perp}} b\right\| \\ &\quad + \left\|P_{\mathcal{N}(T_n)} T^{\dagger} b\right\| + \|R_n - I_X\| \operatorname{dist}\left(T^{\dagger} b, X_n\right) \\ &= \left\|T_n^{\dagger}\right\|_n \left\|Q_n\left(b_n - b\right)\right\| + \left\|T_n^{\dagger} Q_n \left(Q_{\mathcal{R}(T_n)} - Q_{\overline{\mathcal{R}(T)}}\right) b\right\| \\ &\quad + \left\|\left(P_{\mathcal{N}(T_n)} - P_{\mathcal{N}(T)}\right) T^{\dagger} b\right\| + \|R_n - I_X\| \operatorname{dist}\left(T^{\dagger} b, X_n\right),\end{aligned}$$

  which is fundamental to design regularization schemes by projection including regular strategies for ill-posed problems or to design stable schemes by projection for well-posed problems.
- When one designs a discretization method by projection, for a well-posed problem, (H2) is the most important condition; for an ill-posed problem, scheme regularization and strategy regularization are both important conditions; here (R1)-regularization should be emphasized, which is convenient for application.
- By observing the approximation settings of various projection methods, one can see that LS and GDLS are two extreme cases of projection methods, and DLS is the best case of GDLS (in the sense of scheme regularization). We remark that, for LS, DLS, and GDLS, the error estimate (4.1) becomes

$$\left\|T_n^{\dagger} b - T^{\dagger} b\right\| \leqslant \|R_n - I_X\| \operatorname{dist}\left(T^{\dagger} b, X_n\right) \quad \left(\forall\ b \in \mathcal{D}\left(T^{\dagger}\right)\right),$$

$$\left.\begin{aligned}\left\|T_n^{\dagger} Q_n b - T^{\dagger} b\right\| \leqslant \left\|P_{\mathcal{N}(T_n)} T^{\dagger} b\right\| \quad \left(\forall\ b \in \mathcal{D}\left(T^{\dagger}\right)\right), \\ \text{where } Q_n := Q_{\mathcal{R}(T_n)} + Q_{\mathcal{R}(T)^{\perp}}\end{aligned}\right\},$$

  and

$$\left\|T_n^{\dagger} Q_n Q_{\overline{\mathcal{R}(T)}} b - T^{\dagger} b\right\| \leqslant \left\|P_{\mathcal{N}(T_n)} T^{\dagger} b\right\| \quad \left(\forall\ b \in \mathcal{D}\left(T^{\dagger}\right)\right),$$

  respectively. Tables 4.1 and 4.2 illustrate the differences and relations among the results for LS, DLS, GDLS, and general projection methods.

Example 4.1 (see [7, Example 3.19]). Let $\{e_n\}$ be any orthonormal basis of a infinite-dimensional Hilbert space X, and let $X_n := \operatorname{span}\{e_1, \ldots, e_n\}$ $(n \in \mathbb{N})$. We define an operator $T : X \to Y := X$ as follows:

$$T : \sum_{i=1}^{\infty} \xi_i e_i \mapsto \sum_{i=1}^{\infty} \left(a_i \xi_i + b_i \xi_1\right) e_i,$$

with

$$a_i := \begin{cases} i^{-1}, & i \text{ odd}, \\ i^{-\frac{5}{2}}, & i \text{ even}, \end{cases} \qquad b_i := \begin{cases} 0, & i = 1, \\ i^{-1}, & i > 1. \end{cases}$$

TABLE 4.3
Nonprojection methods.

$(\mathrm{H3})^{\mathrm{s}}_{\mathrm{p}}$	$\Rightarrow$	$(\mathrm{H3})^{\mathrm{s}}$	$\Rightarrow$	$T_n^\dagger Q_n Q_{\mathcal{R}(T)^\perp} \xrightarrow{\mathrm{s}} 0$
$\Downarrow$		$\Downarrow$		
$(\mathrm{H3})^{\mathrm{w}}_{\mathrm{p}}$	$\Rightarrow$	$(\mathrm{H3})^{\mathrm{w}}$	$\not\Rightarrow$	$Q_{\mathcal{R}(T_n)} \xrightarrow{\mathrm{s}} Q_{\overline{\mathcal{R}(T)}}$
$\Downarrow$		$\Downarrow$		
$\left\{(\mathrm{H2}) \text{ and } P_{\mathcal{N}(T_n)} \xrightarrow{\mathrm{s}} P_{\mathcal{N}(T)}\right\}$	$\Rightarrow$	$\sup_n \|R_n\| < \infty$		
$\Downarrow$				
$\overline{\mathcal{R}(T)} = \mathcal{R}(T)$				

Note: (1) $(\mathrm{H3})^{\mathrm{s}}_{\mathrm{p}} \Leftrightarrow \left\{(\mathrm{H2}),\ P_{\mathcal{N}(T_n)} \xrightarrow{\mathrm{s}} P_{\mathcal{N}(T)},\ \text{and } Q_{\mathcal{R}(T_n)} \xrightarrow{\mathrm{s}} Q_{\overline{\mathcal{R}(T)}}\right\}$;
(2) $(\mathrm{H3})^{\mathrm{w}}_{\mathrm{p}} \not\Rightarrow Q_{\mathcal{R}(T_n)} \xrightarrow{\mathrm{s}} Q_{\overline{\mathcal{R}(T)}}$ (see Example 3.7);
(3) $(\mathrm{H2}) \not\Rightarrow P_{\mathcal{N}(T_n)} \xrightarrow{\mathrm{s}} P_{\mathcal{N}(T)}$ (see Example 3.5).

This operator is linear, injective, and compact. If $T_n := T|_{X_n} : X_n \to X$, then

$$P_{\mathcal{N}(T_n)} = P_{\mathcal{N}(T)} = 0 \ (\forall\ n)\,, \qquad \sup_n \|R_n\| = \infty.$$

This gives such an example of LS whose approximation setting is with (R0)- but without (R1)-regularization.

Example 4.2. Let $T : X \to Y$ defined as Example 2.10, where $X = Y = l^2$. If $T_n := Q_n T : X \to Y_n := \operatorname{span}\{e_1, e_2, \ldots, e_n\}$, then

$$\mathcal{R}(T_n) = Y_n\,, \qquad \overline{\mathcal{R}(T)} = \mathcal{R}(T) = e^\perp\,, \qquad Q_{\mathcal{R}(T_n)} \xrightarrow{\mathrm{s}} I \neq Q_{\overline{\mathcal{R}(T)}}\,.$$

This provides such an example of GDLS whose approximation setting is, by Theorem 2.6, with (R1)- but without $(\mathrm{H3})^{\mathrm{w}}$-regularization. In addition, the example demonstrates that "$(\mathrm{H2}) \Longleftrightarrow \overline{\mathcal{R}(T)} = \mathcal{R}(T)$" is not valid for GDLS; however, this proposition holds for DLS (see Corollary 2.4).

CONJECTURE 4.3. *Let the approximation setting* (1.1) *and* (1.2) *satisfy* (2.1). *Then*

$$(\mathrm{H3})^{\mathrm{s}} \iff (\mathrm{H3})^{\mathrm{w}} \iff \begin{cases} \sup_n \|R_n\| < \infty\,, \\ Q_{\mathcal{R}(T_n)} \xrightarrow{s} Q_{\overline{\mathcal{R}(T)}}\,. \end{cases}$$

Fundamental results on nonprojection methods of Moore–Penrose inverses for well-posed problems are also given (see Theorems 1.3 and 3.2). Therefore, Table 4.3 is produced, and the following error estimates hold: For all $y \in \mathcal{D}\left(T^\dagger\right)$ and $y_n \in Y_n$,

$$\begin{aligned} \left\|T_n^\dagger Q_n Q_{\overline{\mathcal{R}(T)}} y - T^\dagger y\right\| &\leqslant \left\|T_n^\dagger\right\|_n \left\|Q_n\left(T_n - T\right) P_n T^\dagger y\right\| + \left\|P_{\mathcal{N}(T_n)} T^\dagger y\right\| \\ &\quad + \left\|T_n^\dagger Q_n T - I\right\| \operatorname{dist}\left(T^\dagger y, X_n\right), \end{aligned} \tag{4.2}$$

$$\begin{aligned} \left\|T_n^\dagger y_n - T^\dagger y\right\| &\leqslant \left\|T_n^\dagger\right\|_n \left(\left\|Q_n\left(y_n - y\right)\right\| + \left\|Q_n\left(T_n - T\right) P_n T^\dagger y\right\|\right) + \left\|P_{\mathcal{N}(T_n)} T^\dagger y\right\| \\ &\quad + \left\|T_n^\dagger Q_n Q_{\mathcal{R}(T)^\perp} y\right\| + \left\|T_n^\dagger Q_n T - I\right\| \operatorname{dist}\left(T^\dagger y, X_n\right) \\ &= \left\|T_n^\dagger\right\|_n \left(\left\|Q_n\left(y_n - y\right)\right\| + \left\|Q_n\left(T_n - T\right) P_n T^\dagger y\right\|\right) \\ &\quad + \left\|\left(P_{\mathcal{N}(T_n)} - P_{\mathcal{N}(T)}\right) T^\dagger y\right\| + \left\|T_n^\dagger Q_n \left(Q_{\mathcal{R}(T_n)} - Q_{\overline{\mathcal{R}(T)}}\right) y\right\| \\ &\quad + \left\|T_n^\dagger Q_n T - I\right\| \operatorname{dist}\left(T^\dagger y, X_n\right). \end{aligned}$$

We remark that the above estimates can be also applied to ill-posed problems, based on which regularization nonprojection schemes including regular strategies can be designed and analyzed.

Clearly, nonprojection methods are with a wider scope of application than projection methods; but the latter are more convenient for application than the former. By observing Tables 4.1 and 4.3 with the error estimates (4.1) and (4.2), one can see the difference and relation between projection methods and nonprojection methods.

Appendix. The proof of the three propositions in Example 2.10.

Let $X := l^2$, $X_n := \operatorname{span}\{e_1, e_2, \dots, e_n\}$, where

$$e_k = (0, \dots, 0, \underset{k\,\text{th}}{\overset{\uparrow}{1}}, 0, \dots) \quad (\forall\ k \in \mathbb{N}).$$

We define an operator $T : X \to X$ as follows:

$$Tx := x - \langle x, e\rangle\, e \qquad (x \in X),$$

where $\langle\cdot,\cdot\rangle$ is the inner product on X and

$$e := \frac{\sqrt{3}}{2}e_1 + \frac{\sqrt{3}}{2^2}e_2 + \cdots + \frac{\sqrt{3}}{2^k}e_k + \cdots.$$

This operator satisfies

$$T^2 = T = T^* \quad \text{and} \quad \mathcal{N}(T) = \operatorname{span}\{e\}; \tag{A.1}$$

that is, T is the orthogonal projection of X onto $e^{\perp}$. Therefore, we have

$$T^{\dagger} = T. \tag{A.2}$$

Now set $T_n := T|_{X_n} : X_n \to X$. Then the following three propositions are valid.

PROPOSITION A.1. $\mathcal{N}(T_n) = \mathcal{N}(T) \cap X_n = \{0\}$, $\overline{\bigcup_{n=1}^{\infty} (\mathcal{N}(T) \cap X_n)} \subsetneqq \mathcal{N}(T)$.

Proof. This proposition is clear by (A.1). ∎

PROPOSITION A.2. *Let P_n denote the orthogonal projection onto X_n. Then*

$$T_n^{\dagger} y = P_n y - 4^n \langle (I - P_n)\, y, e\rangle\, P_n e \quad (\forall\ y \in X).$$

Proof. Note that

$$\|e\| = 1, \quad \langle P_n e, e\rangle = \|P_n e\|^2 = 1 - \frac{1}{4^n}. \tag{A.3}$$

Define $S_n : X \to X$ as follows:

$$S_n y := y - 4^n \langle (I - P_n)\, y, e\rangle\, e \quad (\forall\ y \in X). \tag{A.4}$$

Then it follows from (A.4), (A.1), and (A.3) that

$$\mathcal{N}(T) \subseteq \mathcal{N}(S_n), \quad \mathcal{R}((I - P_n)\, S_n) \subseteq e^{\perp}. \tag{A.5}$$

By use of (A.4), (A.5), and (A.3) we have

$$S_n|_{X_n} = P_n|_{X_n}, \quad S_n T = S_n, \quad T S_n = T, \tag{A.6}$$

and

$$T(I - P_n)\, S_n = (I - P_n)\, S_n, \quad [(I - P_n)\, S_n]^* = (I - P_n)\, S_n. \tag{A.7}$$

It follows from (A.6) that the operator $P_nS_n : X \to X_n$ satisfies

$$\text{(A.8)} \qquad \begin{cases} T_n\,(P_nS_n)\,T_n = T_n\,P_n|_{X_n} = T_n, \\ (P_nS_n)\,T_n\,(P_nS_n) = P_n|_{X_n}\,(P_nS_n) = P_nS_n, \\ [(P_nS_n)\,T_n]^* = I_n^* = I_n = (P_nS_n)\,T_n. \end{cases}$$

It follows from (A.7), (A.6), and (A.1) that $P_nS_n : X \to X_n$ also satisfies

$$\text{(A.9)} \qquad [T_n\,(P_nS_n)]^* = [T - (I - P_n)\,S_n]^* = T - (I - P_n)\,S_n = T_n\,(P_nS_n)\,.$$

Therefore, by (A.8) and (A.9), we have

$$T_n^\dagger = P_nS_n.$$

Thus, the proposition is proved. ∎

PROPOSITION A.3. *If $y \in X$, then*

$$\sup_n \left\|T_n^\dagger y\right\| < \infty \Longleftrightarrow \sup_n 4^n\,|\langle (I - P_n)\,y, e\rangle| < \infty\,,$$
$$T_n^\dagger y \rightharpoonup T^\dagger y \Longleftrightarrow 4^n\,\langle (I - P_n)\,y, e\rangle \longrightarrow \langle y, e\rangle\ .$$

Proof. By Proposition A.2, it follows that

$$\begin{cases} \left\|T_n^\dagger y\right\| \leqslant 4^n\,|\langle (I - P_n)\,y, e\rangle|\,\|P_ne\| + \|P_ny\|\,, \\ 4^n\,|\langle (I - P_n)\,y, e\rangle|\,\|P_ne\| \leqslant \left\|T_n^\dagger y\right\| + \|P_ny\|\,, \end{cases}$$

so that we obtain (notice that $\frac{\sqrt{3}}{2} \leqslant \|P_ne\| \leqslant 1$ and $\|P_ny\| \leqslant \|y\|$)

$$\sup_n \left\|T_n^\dagger y\right\| < \infty \Longleftrightarrow \sup_n 4^n\,|\langle (I - P_n)\,y, e\rangle| < \infty\,.$$

By Proposition A.2 and (A.2), it follows that for any $x \in X$ there holds

$$\begin{aligned} \left\langle x, T_n^\dagger y - T^\dagger y\right\rangle &= \langle x, (P_n - I)\,y - 4^n\,\langle (I - P_n)\,y, e\rangle\,P_ne + \langle y, e\rangle\,e\rangle \\ &= \langle x, e\rangle\,\overline{\langle y, e\rangle} - 4^n\,\langle P_nx, e\rangle\,\overline{\langle (I - P_n)\,y, e\rangle} - \langle x, (I - P_n)\,y\rangle\,. \end{aligned}$$

This implies that

$$\begin{aligned} T_n^\dagger y \rightharpoonup T^\dagger y &\Longleftrightarrow 4^n\,\langle P_nx, e\rangle\,\overline{\langle (I - P_n)\,y, e\rangle} \longrightarrow \langle x, e\rangle\,\overline{\langle y, e\rangle}\ \ (\forall\ x \in X) \\ &\Longleftrightarrow 4^n\,\langle (I - P_n)\,y, e\rangle \longrightarrow \langle y, e\rangle\ . \end{aligned}$$

The proof is complete. ∎

Acknowledgment. I am glad to acknowledge the very detailed suggestions of the referee and the editor for improving my manuscript. Their comments and suggestions were very helpful to me.

REFERENCES

[1] K. ATKINSON AND W. HAN, *Theoretical Numerical Analysis*, Springer-Verlag, New York, 2001.
[2] G. CHEN, M. WEI, AND Y. XUE, *Perturbation analysis of the least squares solution in Hilbert spaces*, Linear Algebra Appl., 244 (1996), pp. 69–80.

[3] J. Ding and L. J. Huang, *On the continuity of generalized inverses of linear operators in Hilbert spaces*, Linear Algebra Appl., 262 (1997), pp. 229–242.
[4] N. Du, *The basic principles for stable approximations to orthogonal generalized inverses of linear operators in Hilbert spaces*, Numer. Funct. Anal. Optim., 26 (2005), pp. 675–708.
[5] N. L. Du, *Strong convergence criteria for sequences of orthogonal projections and Galerkin approximations for generalized inverses*, Acta Math. Sinica, 50 (2007), pp. 43–54.
[6] H. W. Engl, *On the convergence of regularization methods for ill-posed linear operator equations*, in Improperly Posed Problems and Their Numerical Treatment, G. Hämmerlin and K. H. Hoffmann, eds., Birkhäuser, Basel, 1983, pp. 81–95.
[7] H. W. Engl, M. Hanke, and A. Neubauer, *Regularization of Inverse Problems*, Kluwer, Dordrecht, 2000.
[8] H. W. Engl and C. W. Groetsch, *Projection-regularization methods for linear operator equations of the first kind*, Proc. Centre Math. Anal. (Canberra), 17 (1988), pp. 17–31.
[9] C. W. Groetsch, *On a regularization-Ritz method for Fredholm equations of the first kind*, J. Integral Equations, 4 (1982), pp. 173–182.
[10] C. W. Groetsch, *The Theory of Tikhonov Regularization for Fredholm Equations of the First Kind*, Pitman, London, 1984.
[11] C. W. Groetsch, *Generalized Inverses of Linear Operators: Representation and Approximation*, Dekker, New York, 1977.
[12] C. W. Groetsch and A. Neubauer, *Convergence of a general projection method for an operator equation of the first kind*, Houston J. Math., 14 (1988), pp. 201–208.
[13] C. W. Groetsch and A. Neubauer, *Regularization of ill-posed problems: optimal parameter choice in finite dimensions*, J. Approx. Theory, 58 (1989), pp. 184–200.
[14] S. Izumino, *Convergence of generalized inverses and spline projectors*, J. Approx. Theory, 38 (1983), pp. 269–278.
[15] J. T. King and A. Neubauer, *A variant of finite-dimensional Tikhonov regularization with a-posteriori parameter choice*, Computing, 40 (1988), pp. 91–109.
[16] A. Kirsch, *An Introduction to the Mathematical Theory of Inverse Problems*, Springer-Verlag, New York, 1996.
[17] M. A. Krasnosel'skii, G. M. Vainikko, P. P. Zabreiko, Ya. B. Rutitskii, and V. Ya. Stetsenko, *Approximate Solution of Operator Equations*, translated by D. Louvish, Wolters-Noordhoff Publishing, Groningen, 1972.
[18] R. Kress, *Linear Integral Equations*, Springer-Verlag, Berlin, 1989.
[19] J. J. Koliha, *The Drazin and Moore-Penrose inverse in C^*-algebras*, Math. Proc. R. Ir. Acad., 99A (1999), pp. 17–27.
[20] J. P. Ma, *Continuously sufficient and necessary condition for Moore-Penrose inverse $A^\dagger$*, Sci. China Ser. A, 33 (1990), pp. 1294–1302.
[21] J. Ma, W. Cao, and G. Song, *On the necessary and sufficient conditions of the continuity of M-P inverses $A^\dagger$*, Chinese Ann. Math. Ser. B, 13 (1992), pp. 251–256.
[22] M. Z. Nashed, ed., *Generalized Inverses and Applications*, Academic Press, New York, 1976.
[23] M. Z. Nashed, *Perturbations and approximations for generalized inverses and linear operator equations*, in Generalized Inverses and Applications, M. Z. Nashed, ed., Academic Press, New York, 1976, pp. 325–396.
[24] R. Plato and G. Vainikko, *On the regularization of projection methods for solving ill-posed problems*, Numer. Math., 57 (1990), pp. 63–79.
[25] V. Rakočević, *On continuity of the Moore-Penrose and Drazin inverses*, Mat. Vesnik, 49 (1997), pp. 163–172.
[26] G. Vainikko, *On the discretization and regularization of ill-posed problems with noncompact operators*, Numer. Funct. Anal. Optim., 13 (1992), pp. 381–396.
[27] T. I. Seidman, *Nonconvergence results for the application of least-squares estimation to ill-posed problems*, J. Optim. Theory Appl., 30 (1980), pp. 535–547.
[28] G. Wang, Y. Wei, and S. Qiao, *Generalized Inverses: Theory and Computations*, Science Press, Beijing, 2004.
[29] Y. Wei and G. Chen, *Perturbation of least squares problem in Hilbert spaces*, Appl. Math. Comput., 121 (2001), pp. 177–183.
[30] K. Yosida, *Functional Analysis*, 6th ed., Springer-Verlag, Berlin, 1980.

SIAM J. NUMER. ANAL.
Vol. 46, No. 3, pp. 1483–1502
© 2008 Society for Industrial and Applied Mathematics

UNIFORM CONVERGENCE OF A NONLINEAR ENERGY-BASED MULTILEVEL QUANTIZATION SCHEME*

QIANG DU† AND MARIA EMELIANENKO‡

Abstract. A popular vector quantization scheme can be constructed by centroidal Voronoi tessellations (CVTs) which also have many other applications in diverse areas of science and engineering. The development of efficient algorithms for their construction is a key to the successful applications of CVTs in practice. This paper studies the details of a new optimization-based multilevel algorithm for the numerical computation of CVTs. The rigorous proof of its uniform convergence in one space dimension and the results of computational simulations are provided. They substantiate recent claims on the significant speedup demonstrated by the new scheme in comparison with traditional methods.

Key words. optimal quantization, centroidal Voronoi tessellations, Lloyd's method, multilevel method, uniform convergence

AMS subject classifications. 65D18, 65H10, 65Y20, 68U10

DOI. 10.1137/050648699

1. Introduction.

Optimal vector quantization is used in many applications such as image and data compression, pattern recognition, and image rendering [24]. A vector quantizer maps N-dimensional vectors in the domain $\Omega \subset \mathbb{R}^N$ into a finite set of vectors $\{\mathbf{z}_i\}_{i=1}^k$. Each vector $\mathbf{z}_i$ is called a code vector or a *codeword*, and the set of all the codewords is called a codebook. A special quantization scheme is given by the Voronoi tessellation which associates with each codeword $\mathbf{z}_i$ (also called a *generator*) a nearest neighbor region that is called a Voronoi region $\{V_i\}_{i=1}^k$. That is, for each i, V_i consists of all points in the domain Ω that are closer to $\mathbf{z}_i$ than to all the other generating points, and a Voronoi tessellation refers to the tessellation of a given domain by the Voronoi regions $\{V_i\}_{i=1}^k$ associated with a set of given generating points $\{\mathbf{z}_i\}_{i=1}^k \subset \Omega$ [1, 34].

With a suitably defined distortion measure, an optimal quantizer can be described as a centroidal Voronoi tessellation. For a given density function ρ defined on Ω, we may define the centroids, or mass centers, of regions $\{V_i\}_{i=1}^k$ by

$$\mathbf{z}_i^* = \Big(\int_{V_i} \mathbf{y}\rho(\mathbf{y})\,d\mathbf{y}\Big)\Big(\int_{V_i} \rho(\mathbf{y})\,d\mathbf{y}\Big)^{-1}. \tag{1.1}$$

Then an *optimal quantization* may be constructed through a *centroidal Voronoi tessellation* (CVT) for which the generators of the Voronoi tessellation are the centroids of their respective Voronoi regions; in other words, $\mathbf{z}_i = \mathbf{z}_i^*$ for all i. Such a connection between CVTs and optimal quantization schemes has been explored extensively in the literature [13].

*Received by the editors January 3, 2006; accepted for publication (in revised form) December 17, 2007; published electronically April 4, 2008. This research was supported in part by the U.S. NSF through grants DMS-0409297 and ITR-0205232.

http://www.siam.org/journals/sinum/46-3/64869.html

†Department of Mathematics, Pennsylvania State University, University Park, PA 16802 (qdu@math.psu.edu).

‡Department of Mathematical Sciences, George Mason University, Fairfax, VA 22030 (memelian@gmu.edu).

Given a set of points $\{\mathbf{z}_i\}_{i=1}^k$ and a tessellation $\{V_i\}_{i=1}^k$ of the domain, we may define the *energy functional* or the *distortion value* for the pair $(\{\mathbf{z}_i\}_{i=1}^k, \{V_i\}_{i=1}^k)$ by

$$\mathcal{G}\Big(\{\mathbf{z}_i\}_{i=1}^k, \{V_i\}_{i=1}^k\Big) = \sum_{i=1}^k \int_{V_i} \rho(\mathbf{y})|\mathbf{y}-\mathbf{z}_i|^2\, d\mathbf{y}\,. \tag{1.2}$$

The minimizer of $\mathcal{G}$, that is, the optimal quantizer, necessarily forms a CVT which illustrates the optimization property of the CVT [13]. The terms optimal quantizer and CVT are thus to be used interchangeably in what follows. For more studies on optimal quantization schemes, we refer to [23, 24, 38]. We note that, besides providing an optimal least squares vector quantizer design in electrical engineering applications, the CVT concept also has applications in diverse areas such as astronomy, biology, image and data analysis, resource optimization, sensor networks, geometric design, and numerical partial differential equations [5, 8, 13, 14, 15, 16, 17, 18, 26, 27, 33, 39, 40]. We refer to [13] for a more comprehensive review of the mathematical theory and diverse applications of CVTs.

In the seminal work of Lloyd on least squares quantization [32], one of the algorithms proposed for computing optimal quantizers is a fixed point–type iterative algorithm consisting of the following simple steps: Starting from an initial quantization (a Voronoi tessellation corresponding to an old set of generators), a new set of generators is defined by the mass centers of the Voronoi regions. This process is continued until a certain stopping criterion is met. For algorithms on the computation of Voronoi tessellations, we refer to [1, 19, 21]. It is easy to see that the Lloyd algorithm is an energy descent iteration of the energy functional (1.2), which gives strong indications to its practical convergence. We refer to [11, 20] for some discussions on the recent development of a rigorous convergence theory.

Lloyd's algorithms and their variants have been proposed and studied in many contexts for different applications [15, 24, 31, 36]. A particular extension using parallel and probabilistic sampling was given in [27] which allows efficient and mesh-free implementation of the Lloyd algorithm. Still, the Lloyd algorithm is at best linearly convergent. Moreover, it slows down as the number of generators increases.

For modern applications of the CVT concept in large scale scientific and engineering problems such as data communication, vector quantization, and mesh generation, efficient algorithms for computing the CVTs play crucial roles.

Recently, a new multilevel approach to the optimal quantization problem has been developed [10, 12]. The new multilevel scheme offers considerable speedup over traditional methods such as the celebrated Lloyd iteration. It can be combined with other techniques to accelerate the computation of CVTs and the optimal vector quantizers. In this paper, we present a rigorous mathematical theory for the new algorithm. We focus on the main characteristics of this multilevel scheme, namely, the uniform convergence with respect to the problem size. Though the result is shown in one space dimension only, such a result is the first of its type in the vector quantization field and on the computation of CVTs. Proofs for higher-dimensional cases are more involved, and they are now being worked out in our ongoing study [12].

The rest of the paper is organized as follows. In section 2, the optimization-based nonlinear multilevel algorithm is introduced. In section 3, the uniform convergence theory is established in the one-dimensional case. Numerical results demonstrating its superiority over traditional methods are given in section 4. Final conclusions are made in section 5.

2. Optimization-based nonlinear multilevel algorithm. Since the original concept of CVTs is related to the solution of a nonlinear optimization problem, and the monotone energy descent property is preserved by the Lloyd iteration [13], we may thus investigate whether monotone energy reduction can be achieved in a multilevel procedure which would also improve the performance of the simple-minded iteration.

The problem of constructing a CVT is nonlinear in nature, and hence standard linear multigrid theory cannot be directly applied. There are still several ways one could implement a nonlinear multilevel scheme in this context (see [9, 10, 28, 29]). A Newton-type acceleration method, studied earlier in [9], is based on some global linearization as the outer loop, coupled with other fast solvers in the inner loop. Alternatively, we now study an approach that overcomes the difficulties of the nonlinearity by essentially relying on direct energy minimization without any type of global linearization.

The optimization formulation of CVTs is not new [13]. CVTs and optimal quantizers are known to be closely related to the problem of minimizing the functional $\mathcal{G}\big(\{\mathbf{z}_i\}_{i=1}^k, \{V_i\}_{i=1}^k\big)$ with $\{V_i\}_{i=1}^k$ forming the Voronoi tessellation corresponding to the generators $\{\mathbf{z}_i\}_{i=1}^k$. In fact, we may note that the vector of generators of a CVT forms a critical point of the energy, and vice versa [13]. This is one of the important characterizations of CVTs which will be used in the later discussion. In our study here, we follow the ideas presented in the literature on the extension of multigrid ideas to nonlinear optimization problems (see [6, 22, 25, 30, 37] and the references cited therein). However, we observe that a direct application of multilevel schemes to the minimization of the original energy is not the best strategy. We illustrate later in this section that an alternative minimization formulation can be introduced to make the problem of computing CVTs more amenable for the multilevel framework.

2.1. Preconditioned formulation. Let $\mathbf{V} = \{V_i\}_{i=1}^k$ denote a Voronoi tessellation corresponding to some choice of iteration points $\mathbf{Z}$ in Ω. For any Voronoi cell $V_i \subset \Omega$, let us denote by R_i the mass of V_i, i.e.,

$$R_i := \int_{V_i} \rho(\mathbf{y})\, \mathrm{d}\mathbf{y}.$$

The fact that ρ is positive almost everywhere implies that $R_i > 0$ for all $i = 1, \ldots, k$.

We define the set of iteration points $\mathbf{W}$ by

$$\mathbf{W} = \{(w_i)|_{i=0}^{k+1} | \quad 0 = w_0 \le w_i \le w_{i+1} \le w_{k+1} = 1 \quad \forall\, 0 \le i \le k\}\,, \tag{2.1}$$

which includes the generators and the end points. The so-called Lloyd map is the function $\mathbf{T} : \mathbf{W} \to \mathbf{W}$ that takes a tuple of generators into the tuple of the centroids of its Voronoi regions (exactly the iteration used in the Lloyd algorithm). More precisely, let $\mathbf{T} = (T_1, \ldots, T_k)$; then for any $z \in \mathbf{W}$,

$$T_i(z) = \frac{\int_{V_i(z)} \mathbf{y}\rho(\mathbf{y})\, \mathrm{d}\mathbf{y}}{\int_{V_i(z)} \rho(\mathbf{y})\, \mathrm{d}\mathbf{y}}$$

for $i = 1, \ldots, k$. The Lloyd map is continuous from $\mathbf{W}$ to $\mathbf{W}$ [13, 11].

For any $z \in \mathbf{W}$, the *Lloyd energy* is defined as

$$\mathcal{H}(z) = \mathcal{G}(z, \mathbf{V}(z)) = \sum_{i=0}^{n} \int_{V_i(z)} \rho(\mathbf{y}) |\mathbf{y} - z_i|^2\, \mathrm{d}\mathbf{y}.$$

The Lloyd energy function is also continuous on $\mathbf{W}$. The derivative of $\mathcal{H}$ is given [13] by

$$\frac{\partial \mathcal{H}}{\partial z_i}(z) = 2R(V_i(z))(z_i - \mathbf{T}_i(z)) \,. \tag{2.2}$$

To perform rigorous analysis of our nonlinear multigrid scheme, it is desirable if the energy functional is locally convex and satisfies some spectral equivalence properties in the appropriate norm. In general, we cannot expect these properties to hold. However, relating our problem to an equivalent optimization problem through a technique that mimics the role of a dynamic nonlinear preconditioner turns out to be very effective. More precisely, denote by P a nonsingular matrix playing the role of a preconditioner. It is easy to deduce that $P\nabla\mathcal{H} = 0$ at a CVT. Hence we arrive at an equivalent formulation of the minimization problem: $\min ||P\nabla\mathcal{H}||^2$, with respect to the standard Euclidean norm. In particular, we can choose $P = (\mathbf{I} - S)^{-1/2}R^{-1}$, where R is a diagonal matrix $R = \mathtt{diag}(R_1, R_2, \ldots, R_k)$, where, as before, the R_i's are the masses of the corresponding Voronoi cells. As we will show later, this preconditioning strategy renders the problem convex and shares some spectral equivalence properties in an appropriately chosen norm for a class of densities, say of the form $\rho(x) = 1 + \epsilon g(x)$. For such densities, the energy functional $\mathcal{H}$ may not be locally convex; however, the equivalent preconditioned strategy successfully eliminates this difficulty. Notice that for this type of density we have $\frac{\partial \mathbf{T}}{\partial z} = S + \epsilon\widetilde{S}(z)$, where

$$S = \begin{pmatrix} 1/4 & 1/4 & 0 & & 0 & 0 & 0 \\ 1/4 & 1/2 & 1/4 & \ldots & 0 & 0 & 0 \\ 0 & 1/4 & 1/2 & & 0 & 0 & 0 \\ & \ldots & & \ldots & & \ldots & \\ 0 & 0 & 0 & & 1/2 & 1/4 & 0 \\ 0 & 0 & 0 & \ldots & 1/4 & 1/2 & 1/4 \\ 0 & 0 & 0 & & 0 & 1/4 & 1/4 \end{pmatrix}$$

is the derivative of the Lloyd map corresponding to the case of the constant density. This preconditioned formulation leads to a multilevel algorithm based on the following nonlinear optimization problem:

$$\min_{z\in\mathbf{W}} \widetilde{\mathcal{H}}(z), \text{ where } \widetilde{\mathcal{H}}(z = \{z_i\}_{i=0}^{k+1}) = ||P\nabla\mathcal{H}(\{\mathbf{z}_i\}_{i=1}^{k})||^2 \,. \tag{2.3}$$

The functional $\widetilde{\mathcal{H}}$ may be regarded as a preconditioned energy, and with the above choice of the preconditioner P, we have

$$\widetilde{\mathcal{H}}(z) = ||2(\mathbf{I} - S)^{-1/2}R^{-1}R(z - \mathbf{T}(z))||^2 = 2[z - \mathbf{T}(z)]^T[(\mathbf{I} - S)^{-1}(z - \mathbf{T}(z))] \tag{2.4}$$

and the gradient can be computed as

$$\begin{aligned} \nabla\widetilde{\mathcal{H}}(z) &= 4\left(\mathbf{I} - \frac{\partial \mathbf{T}}{\partial z}\right)(\mathbf{I} - S)^{-1}(z - \mathbf{T}(z)) = 4(\mathbf{I} - \epsilon\widehat{S}(z))(z - \mathbf{T}(z)) \\ &= 4(z - \mathbf{T}(z)) - 4\epsilon\beta(z), \end{aligned} \tag{2.5}$$

where $\widehat{S}(z) = \widetilde{S}(z)(I - S)^{-1}$ and $\beta(z) = \widehat{S}(z)(z - \mathbf{T}(z))$.

A key observation is that as R varies with respect to the generators, the above transformation or *dynamic preconditioning* keeps the modified objective functional convex in a suitably large neighborhood of the minimizer and therefore makes the new formulation more amenable to analysis than the original problem.

2.2. Space decomposition. Suppose that without loss of generality we start with an odd number of generators $k = 2^{n_J} - 1$. Now, we define the set of iteration points $\mathbf{W}$ by (2.1).

While the set of iteration points will be dynamically updated at each step of the algorithm, we want to construct a domain decomposition that will not be affected by these changes. Namely, we can denote by $\mathbf{X} = \mathbf{X}^J$ the set of uniformly distributed grid nodes in the unit interval, and by $\mathbf{X}^i$ the sequence of grid nodes $\mathbf{X}^i = \{x_j^i\}_{j=0}^{n_i}$, obtained via uniform subdivision $x_j^i = j/2^{n_i}$. A set of nonoverlapping grid points at the ith level would consist of $\overline{\mathbf{X}}^i = \mathbf{X}^i/\mathbf{X}^{i-1}$. Next we define by $\mathcal{T} = \mathcal{T}_J$ a mesh corresponding to $\mathbf{X}$ and consider a sequence of nested uniform finite element meshes $\mathcal{T}_1 \subset \mathcal{T}_2 \subset \cdots \mathcal{T}_J$, where $\mathcal{T}_i$ consists of all finite element meshes $\{\tau_j^i\}_{j=1}^{n_i-1}$ with mesh parameters $h_i = h_{i-1}/2$, such that $\cup_{j=1}^{n_i} \tau_j^i = \Omega = [0,1]$ and $n_i = 2n_{i-1}$. Corresponding to each partition $\mathcal{T}_i$ $(i = 1, \ldots, J)$, there is a finite element space $\mathbf{Y}_i$ defined by

$$\mathbf{Y}_i = \{v \in H_0^1(\Omega) \ \mid \ v|_\tau \in \mathcal{P}_1(\tau) \ \forall \tau \in \mathcal{T}_i\} \ .$$

For each $\mathbf{Y}_i$, there corresponds a nodal basis $\{\psi_j^i\}_{j=1}^{n_i}$, such that $\psi_j^i(x_k^i) = \delta_{jk}$, where δ_{jk} is the usual Kronecker delta function and $\{x_k^i\}_{k=1}^{n_i}$ is the set of all nodes of the elements of $\mathcal{T}_i$ with $x_1^J = 0, x_{n_J}^J = 1$. Define the corresponding one-dimensional subspaces $\mathbf{Y}_{i,j} = \text{span}\{\psi_j^i\}$. Then the decomposition can be regarded as

$$\mathbf{Y}_J = \sum_{i=1}^{J} \sum_{j=1}^{n_i} \mathbf{Y}_{i,j} = \bigoplus_{i=1}^{J} \overline{\mathbf{Y}}_i,$$

where $\overline{\mathbf{Y}}_i = \mathbf{Y}_i/\mathbf{Y}_{i-1}$ for $i > 1$ and $\overline{\mathbf{Y}}_1 = \mathbf{Y}_1$, following the standard multigrid construction given in [4]. Now clearly for each $\psi_j^i \in \mathbf{Y}_i$, we can find a vector $w_j^i = \{\bar{\psi}_{jm}^i\}_{m=1}^{n_J-1} \in \mathbb{R}^{n_J}$, such that $\psi_j^i(x) = \sum_{m=1}^{n_J-1} \bar{\psi}_{jm}^i \psi_m^J(x)$ for $x \in \Omega$.

Figure 2.1 gives an illustration of this decomposition. Notice that this configuration is independent of the positions of generators and relies entirely on the grids $\mathcal{T}_i$ that are chosen to be uniform.

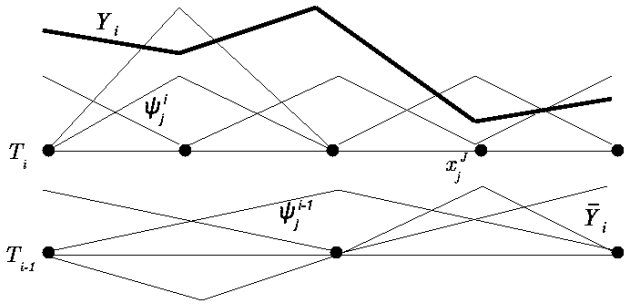


FIG. 2.1. *The structure of the space decomposition subordinate to the uniform mesh* $\mathbf{X}$ *in* $[0,1]$.

Using the linear interpolation coefficients $w_j^i = \{\bar{\psi}_{jm}^i\}_{m=1}^{n_j} \in \mathbb{R}^{n_J}$ defined above, we can decompose the space of iteration points as follows:

$$\mathbf{W} = \sum_{i=1}^{J} \mathbf{W}_i = \bigoplus_{i=1}^{J} \overline{\mathbf{W}}_i,$$

where $\mathbf{W}_i = \text{span}\{w_j^i\}$ and $\overline{\mathbf{W}}_i = \mathbf{W}_i/\mathbf{W}_{i-1}$ for $i > 1$. The vector spaces defined this way provide all the information necessary to fully describe the iteration procedure given below.

Note that once a mesh coarsening procedure is described, the above setup naturally applies to higher-dimensional cases, where $\mathbf{X}$ consists of grids points in a higher-dimensional domain and is discussed in more detail in [12]. Here we will restrict our attention to the one-dimensional case. It may then be noted that the set of basis functions $[w_1^i, \ldots, w_{n_i}^i]^T \in \mathbb{R}^{n_i \times k}$ used at each iteration can be pregenerated using the recursive procedure $[w_1^J, \ldots, w_{nJ}^J]^T = I_{k\times k}$ and $[w_1^{J-s}, \ldots, w_{nJ-s}^{J-s}]^T = (\Pi_{i=1}^s P_{J-i})$, where P_i is the basis transformation from space $\mathbf{W}_{i+1}$ to $\mathbf{W}_i$ which plays the role of a restriction operator.

2.3. Description of the algorithm. Using the above notation, we design Algorithm 2.2, which is a multilevel successive subspace correction algorithm [6, 41]. Each step of the procedure outlined below involves solving a system of nonlinear equations, which plays the role of relaxation. We can use the Newton iteration to solve this nonlinear system. Solution at current iterate is updated after each nonlinear solve by the Gauss–Seidel-type procedure, and hence the resulting scheme is successive in nature. The algorithm uses the procedure *CoarseGridSolve*, which, as the name indicates, refers to finding the solution at the coarsest level. In our implementation, this procedure consists of applying the Lloyd method to the $2^m - 1$ coarsest generators for a few steps (given by $imax$) or until saturation. In the implementation, we mostly use $m = 2$ so that there are only three coarsest generators.

ALGORITHM 2.1. $\mathbf{U}$ = CoarseGridSolve($\mathbf{U}$,tol).
Input: A set of k iteration points $\mathbf{U} = \{u_i\}_{i=1}^k$ *on* Ω.
Method:
Define the set of coarsest grid nodes: $z = \{z_j\}_{j=1}^{2^m-1}$, *where* $z_j = u_{\frac{j(k+1)}{2^m}}$.
For `iter` $= 1 : imax$
Perform Lloyd iterations with $2^m - 1$ *generators:*
1. *Construct the Voronoi regions* $\{V_j(z)\}_{j=1}^{2^m-1}$ *of* Ω *associated with z;*
2. *Determine the centroids* $\mathbf{T}(z)$ *of the Voronoi regions* $V_j(z)$*;*
these centroids form the new set of points z;
3. *If* $||z - \mathbf{T}(z)|| <$ tol, *return z and terminate the cycle;*
otherwise, goto step 1.

endfor.
Update $\mathbf{U}$ *by* $u_{\frac{j(k+1)}{2^m}} = z_j, j = 1, 2, \ldots, 2^m - 1$.

In general, other efficient optimization methods, as well as Newton's method, can be used in order to quickly damp the error, since the number of unknowns on the coarsest grid remains relatively small.

The "slash" cycle can be defined as follows.

ALGORITHM 2.2. Successive correction $V(\nu)$ scheme.
Input:
k, number of generators; $u_1 = \{z_i\}_{i=1}^k$, *the ends plus the initial set of generators.*
tol, *tolerance for the coarse grid solution accuracy.*
Output after n cycles:
$u_n = \{z_i\}_{i=0}^{k+1}$, *the ends plus the set of generators for CVT.*
Method: For $n > 1$, *given* u_n, *do*
1. *For* $i = 1 : J$
$\bar{u}_{n+\frac{i-1}{J}} = u_{n+\frac{i-1}{J}}$

Perform ν smoothings:

$\bar{u}_{n+\frac{i-1}{J}} = \bar{u}_{n+\frac{i-1}{J}} + \alpha^0_{j\,l} w^i_j \in \overline{\mathbf{W}}_i$ *sequentially for* $1 \le j \le n_i$,

such that $\widetilde{\mathcal{H}}(\bar{u}_{n+\frac{i-1}{J}} + \alpha^0_{j\,l} w^i_j) = \min_{\alpha_{jl}} \widetilde{\mathcal{H}}(\bar{u}_{n+\frac{i-1}{J}} + \alpha_{j\,l} w^i_j)$,

Update the solution at ith level:

$$u_{n+\frac{i}{J}} = \bar{u}_{n+\frac{i-1}{J}} = u_{n+\frac{i-1}{J}} + e^i_n, \text{ where } e^i_n = \sum_{l=1}^{\nu_1} \sum_{j=1}^{n_i - 1} \alpha^0_{j\,l} w^i_j$$

endfor.

2. *On the coarsest level,* $u_{n+1} \leftarrow$*CoarseGridSolve(*u_{n+1},tol*).*
3. *Set* $n = n + 1$ *repeat procedure* 1 *to* 2 *until some stopping criterion is met.*

Here, the parameter ν denotes the number of Gauss–Seidel iterations (smoothings) used at each level. Although it is enough to have $\nu = 1$ in theory, larger values are used in practice due to the inexact nonlinear solver. The values $\nu \le 3$ usually suffice for the optimization to reach saturation. In the above description, no postsmoothings are used to make the analysis more transparent. A complete $V(\nu_1, \nu_2)$ cycle with ν_1 presmoothings and ν_2 postsmoothings can be defined and analyzed in a similar fashion. The algorithm essentially depends only on the proper space decompositions and its relation to the set of generators, and thus it is applicable in any dimension. The more general forms, including detailed description of the grid coarsening procedure, will be discussed in our subsequent works.

3. The uniform convergence theory. The uniform convergence of the new multilevel scheme can be rigorously proved, at least, for a large class of density functions in one-dimensional space. This is the first step toward a more comprehensive theoretical analysis of this type of multilevel scheme in general. Here, let us first establish some important properties of the energy functional defined in (2.3). Then we introduce our main convergence results. Without further notice, all the analysis in this section is restricted to the one-dimensional case only, and we further assume without loss of generality that the domain is simply the unit interval $\Omega = (0, 1)$, which can always be transformed into any other intervals.

3.1. Notation. First, for each $u, v \in \mathbf{W}$, we adopt the following H^1 inner product:

$$\langle u, v \rangle_{\mathbf{W}} = \frac{1}{k} \sum_{i=1}^{k+1} (u_i - u_{i-1})(v_i - v_{i-1}).$$

Next, if $y = u - v$, we have $y_0 = y_{k+1} = 0$ and

$$||y||^2_{\mathbf{W}} = \frac{1}{k} \sum_{i=1}^{k+1} (y_i - y_{i-1})^2$$

defines a $\mathbf{W}$-norm that we will be using throughout the rest of the paper.

To simplify the presentation, let us introduce the following notation: for $i = 1, \ldots, k$,

$$u^-_i = \frac{u_i + u_{i-1}}{2}, \qquad u^+_{i-1} = \frac{u_{i-1} + u_i}{2}$$

with $a_i = u_i - u_{i-1}, b_i = w_i - w_{i-1}, x_i = u_i - w_i$, and $u_0 = w_0 = 0$, $u_{k+1} = w_{k+1} = 1$ being the fixed ends of the interval.

3.2. Technical lemmas. In the discussion that follows we say that a functional F satisfies the *convexity* and *continuity* properties in $\mathbf{W}$ if there exist constants $K > 0$, $L > 0$ satisfying

$$(\nabla F(w) - \nabla F(v), w - v) \geq K||w - v||^2_{\mathbf{W}} \quad \forall w, v \in \mathbf{W} \,, \tag{3.1}$$

$$(\nabla F(w) - \nabla F(v), w - v) \leq L||w - v||^2_{\mathbf{W}} \quad \forall w, v \in \mathbf{W} \,, \tag{3.2}$$

where $(\cdot, \cdot)$ denotes the regular Euclidean inner product.

Let us first turn our attention to the case of a constant density. In this simple case, with the above notation, we get the following result for the preconditioned energy functional.

PROPOSITION 3.1. *Let $\rho(x) = 1$ be the density function on $[0, 1]$. Then the following relation holds for all $u, w \in \mathbf{W}$:*

$$(\nabla\widetilde{\mathcal{H}}(u) - \nabla\widetilde{\mathcal{H}}(w), u - w) = \frac{1}{2}\sum_{i=1}^{k+1}(a_i - b_i)^2 \,.$$

Proof. Following (2.5), we have

$$\frac{\partial\widetilde{\mathcal{H}}}{\partial u_i} = 2(u_i - T_i) = 2\Big(u_i - \frac{u_i^+ + u_i^-}{2}\Big) = \frac{1}{2}(a_i - a_{i+1}) \,.$$

Thus,

$$\begin{aligned}(\nabla\widetilde{\mathcal{H}}(u) - \nabla\widetilde{\mathcal{H}}(w), u - w) &= \frac{1}{2}\sum_{i=1}^{k}(u_i - w_i)(a_i - a_{i+1} - b_i + b_{i+1}) \\ = \frac{1}{2}\sum_{i=1}^{k}(u_i - w_i)(a_i - b_i) - \frac{1}{2}\sum_{i=1}^{k}(u_{i-1} - w_{i-1})(a_i - b_i) &= \frac{1}{2}\sum_{i=1}^{k+1}(a_i - b_i)^2 \,. \qquad \square\end{aligned}$$

COROLLARY 3.1. *For a constant density, the functional $F = k^{-1}\widetilde{\mathcal{H}}$ satisfies the continuity and convexity conditions with $K = L = 1/2$ for all points in $\mathbf{W}$.*

Note that this is a simple consequence of the fact $\sum_{i=1}^{k+1}(a_i - b_i)^2 = k||u - w||^2_{\mathbf{W}}$.

One can extend the above result to a broader class of density functions. First, the following auxiliary lemma can be verified.

LEMMA 3.2. *Let $M_g = \sup_{x\in[0,1]}|g(x)|$ and $M' = \sup_{x\in[0,1]}|g'(x)|$ on $[0, 1]$, and define*

$$Q_i(u) = \frac{\int_{u_i^-}^{u_i^+}(2x - (u_i^+ + u_i^-))g(x)dx}{(u_i^+ - u_i^-) + \epsilon\int_{u_i^-}^{u_i^+} g(x)dx} \,. \tag{3.3}$$

Then

$$|Q_i(u)| \leq \frac{4M_g|u_i^+ - u_i^-|}{1 - 2\epsilon M_g} \tag{3.4}$$

and

$$|Q_i(u) - Q_i(w)| \leq C_Q\alpha_i \,, \tag{3.5}$$

where $\alpha_i = |u_i^+ - w_i^+| + |u_i^- - w_i^-|$ and $C_Q = M_g + M' + \epsilon M_g^2 + 3\epsilon M_g M'$.

Proof. Denote

$$Q_i(u) = \frac{\int_{u_i^-}^{u_i^+} (2x - (u_i^+ + u_i^-))g(x)dx}{(u_i^+ - u_i^-) + \epsilon \int_{u_i^-}^{u_i^+} g(x)dx} = \frac{N_i}{D_i} .$$

In order to represent the above expression in a more convenient form, we employ the following change of variables argument:

$$\int_\alpha^\beta f(x)\,dx = (\beta - \alpha)\int_{-1}^{1} f\Big(\frac{\alpha+\beta}{2} + \frac{s}{2}(\beta-\alpha)\Big)\,ds . \tag{3.6}$$

Introducing $-1 \le s \le 1$, we can rewrite $x_u(s) = \frac{u_i^+ + u_i^-}{2} + \frac{s}{2}(u_i^+ - u_i^-)$ and $x_w(s) = \frac{w_i^+ + w_i^-}{2} + \frac{s}{2}(w_i^+ - w_i^-)$. Then for the numerator we have

$$\begin{aligned} N_i(u) &= \int_{u_i^-}^{u_i^+} (2x - (u_i^+ + u_i^-))g(x)dx \\ &= (u_i^+ - u_i^-)\int_{-1}^{1} s(u_i^+ - u_i^-)g(x_u(s))\,ds \\ &= (u_i^+ - u_i^-)^2 \int_{-1}^{1} g(x_u(s))s\,ds, \end{aligned}$$

while the denominator is equal to

$$D_i(u) = (u_i^+ - u_i^-)\Big(1 + \epsilon\int_{-1}^{1} g(x_u(s))\,ds\Big) .$$

For the ratio $Q_i = N_i/D_i$ we have

$$Q_i(u) = \frac{N_i(u)}{D_i(u)} = \frac{(u_i^+ - u_i^-)\int_{-1}^{1} g(x_u(s))s\,ds}{1 + \epsilon\int_{-1}^{1} g(x_u(s))\,ds} .$$

We thus can use the bound on g to get (3.4). Now, for simplicity let us redefine the modified numerator and denominator as

$$\widetilde{N}_i(u) = (u_i^+ - u_i^-)\int_{-1}^{1} g(x_u(s))s\,ds , \quad \widetilde{D}_i(u) = 1 + \epsilon\int_{-1}^{1} g(x_u(s))\,ds .$$

In the new notation,

$$\begin{aligned} |Q_i(u) - Q_i(w)| &\le \frac{1}{\widetilde{D}_i(u)\widetilde{D}_i(w)}\Big|\widetilde{N}_i(u)\widetilde{D}_i(w) - \widetilde{N}_i(w)\widetilde{D}_i(u)\Big| \\ &\le \frac{1}{2}\Big|\widetilde{N}_i(u) - \widetilde{N}_i(w)\Big|\Big(\widetilde{D}_i(u) + \widetilde{D}_i(w)\Big) \\ &\quad + \frac{1}{2}\Big(\widetilde{N}_i(u) + \widetilde{N}_i(w)\Big)\Big|\widetilde{D}_i(u) - \widetilde{D}_i(w)\Big| . \end{aligned}$$

Notice further that $|x_u(s) - x_w(s)| \le |u_i^+ - w_i^+| + |u_i^- - w_i^-| = \alpha_i$. It follows that

$$\left|\widetilde{D}_i(u) - \widetilde{D}_i(w)\right| = \epsilon\left|\int_{-1}^{1} (g(x_u(s)) - g(x_w(s)))\, ds\right| \le \epsilon M_g' \alpha_i \ ,$$

$$\begin{aligned}\left|\widetilde{N}_i(u) - \widetilde{N}_i(w)\right| &= \Big|\int_{-1}^{1} ((u_i^+ - u_i^-)g(x_u(z)) - (w_i^+ - w_i^-)g(x_w(s)))s\, ds\Big| \\ &\le 2(M_g + M_g')\alpha_i \ .\end{aligned}$$

Finally, since $|\widetilde{D}_i| \le 1 + \epsilon M_g$ and $\widetilde{N}_i \le 2M_g$, we have

$$|Q_i(u) - Q_i(w)| \le \Big((M_g + M')(1 + \epsilon M_g) + 2\epsilon M_g M'\Big)\alpha_i = C_Q \alpha_i,$$

which leads to (3.5). This proves the lemma. ∎

In a similar fashion, by employing the change of variables argument, we can easily prove the following.

LEMMA 3.3. *With the same assumptions (and notation) as the previous lemma,*

$$|\mathbf{T}_i(u) - \mathbf{T}_i(w)| \le C_T \alpha_i \quad \textit{and} \quad \sum_j |\widetilde{S}_{ij}(u) - \widetilde{S}_{ij}(w)| \le C_S \alpha_i.$$

Proof. The proofs for both $\mathbf{T}$ and $\widetilde{S}$ are analogous and follow along the lines of Lemma 3.2. For instance, direct calculation yields

$$\begin{aligned}\mathbf{T}_i(u) &= u_i - \Big(\int \rho(x)dx\Big)^{-1}\Big(\int (u_i - x)dx + \epsilon\int (u_i - x)g(x)dx\Big) \\ &= u_i - \frac{(u_i^+ - u_i^-)(2x_i - u_i^+ - u_i^-) + 2\epsilon\int (u_i - x)g(x)dx}{2(u_i^+ - u_i^-) + 2\epsilon\int g(x)dx} \\ &= \frac{u_i^+ + u_i^-}{2} - \frac{2\epsilon\int (u_i - x)g(x)dx - \epsilon(u_i^+ - u_i^-)\int g(x)dx}{2(u_i^+ - u_i^-) + 2\epsilon\int g(x)dx} \\ &= \frac{1}{2}\Big(u_i^+ + u_i^-\Big) + \epsilon\Big(\frac{\int (2x - (u_i^+ + u_i^-))g(x)dx}{2(u_i^+ - u_i^-) + 2\epsilon\int g(x)dx}\Big) \ ,\end{aligned}$$

where the integrals all refer to the integral over $[u_i^-, u_i^+]$. By the definition (3.3), we have

$$\mathbf{T}_i(u) = \frac{u_i^+ + u_i^-}{2} + \frac{1}{2}\epsilon Q_i \ . \tag{3.7}$$

One may then use estimates on Q_i to derive similar estimates for $\mathbf{T}_i$. We omit the rest of the details. ∎

With the help of Lemmas 3.2 and 3.3 we can finally derive the following general result.

PROPOSITION 3.2. *For any $\rho(x) = 1 + \epsilon g(x)$ with ϵ suitably small, there exist constants $C_l > 0$ and $C_u > 0$ dependent only on k, such that for any $u, w \in \mathbf{W}$ and with the notation defined earlier*

$$kC_l(k)||u - w||_{\mathbf{W}}^2 \le (\nabla\widetilde{\mathcal{H}}(u) - \nabla\widetilde{\mathcal{H}}(w), u - w) \le kC_u(k)||u - w||_{\mathbf{W}}^2 \ .$$

Proof. Recall from (2.5) that $\frac{\partial \widetilde{\mathcal{H}}}{\partial u_i} = 4[(u_i - \mathbf{T}_i(u)) - \epsilon\beta_i(u)]$. By (3.7), we get

$$\frac{\partial \widetilde{\mathcal{H}}}{\partial u_i}(u) = \Big(a_i - a_{i+1}\Big) - 2\epsilon(Q_i(u) + 2\beta_i(u))$$

and hence

$$(\nabla\widetilde{\mathcal{H}}(u) - \nabla\widetilde{\mathcal{H}}(w), u - w) = \sum_{i=1}^{k} (a_i - b_i)^2 - 2\epsilon(Q(u) - Q(w), u - w)$$

$$- 4\epsilon(\beta(u) - \beta(w), u - w)\ . \tag{3.8}$$

The first term in (3.8) comes from the constant part of the density and hence complies with the results of the previous theorem.

For the second term, from the Cauchy inequality and Lemma 3.2, we have

$$|\epsilon(Q(u) - Q(w), u - w)| \le \sum_{i=1}^{k} \epsilon|Q_i(u) - Q_i(w)|\,|u_i - w_i| \le C_q\epsilon k\|u - w\|^2_{\mathbf{W}} \tag{3.9}$$

for some constant C_q.

Finally, for the last term in (3.8), we note that by (2.5) and (3.7)

$$\beta(u) = \widehat{S}(u)(u - \mathbf{T}(u)) = \widehat{S}(u)\left[(I - S)u - \frac{1}{2}\epsilon Q(u)\right] = \widetilde{S}(u)\left[u - \frac{\epsilon}{2}(I - S)^{-1}Q(u)\right].$$

Hence, $(u - w)^T(\beta(u) - \beta(w)) = B_1 - B_2 + B_3 - B_4$, where

$$B_1 = (u - w)^T\widetilde{S}(u)(u - w), \qquad B_2 = \frac{\epsilon}{2}(u - w)^T\widetilde{S}(u)(I - S)^{-1}[Q(u) - Q(w)],$$

$$B_3 = (u - w)^T[\widetilde{S}(u) - \widetilde{S}(w)]w, \quad B_4 = \frac{\epsilon}{2}(u - w)^T[\widetilde{S}(u) - \widetilde{S}(w)](I - S)^{-1}Q(u).$$

Using the fact that $\widetilde{S}(u)$ is uniform bounded, we get from the Cauchy inequality that

$$|B_1| \le C_{b1}k\|u - w\|^2_{\mathbf{W}}$$

for some constant C_{b1}. Similarly for B_2, we apply the Cauchy inequality along with Lemma 3.2 and the spectral bound on $(1 - S)^{-1}$ (which is of the order k^2) to get

$$|B_2| \le \epsilon C_{b2}k^3\|u - w\|^2_{\mathbf{W}}$$

for some constant C_{b2}. As for B_3, we may use the estimates on $\widetilde{S}(u) - \widetilde{S}(w)$ in Lemma 3.3 and apply the Cauchy inequality to get

$$|B_3| \le C_{b3}k\|u - w\|^2_{\mathbf{W}}$$

for some constant C_{b3}. In addition, for ϵ small, it is easy to see that $Q_i(u)$ is uniformly bounded (independent of k). Hence we use the Cauchy inequality, the estimates for $\widetilde{S}(u) - \widetilde{S}(w)$ in Lemma 3.3, along with the same spectral bound on $(1 - S)^{-1}$ we used in B_2 estimation to get

$$|B_4| \le C_{b4}\epsilon k^3\|u - w\|^2_{\mathbf{W}}$$

for some constant C_{b4}. Hence, we get for some constant C_β, independent of k,

$$|\epsilon(\beta(u)-\beta(w),u-w)| \le C_\beta k(\epsilon+\epsilon^2k^2)||u-w||^2_{\mathbf{W}} \ . \tag{3.10}$$

Combining (3.5), (3.8), (3.9), and (3.10), we get

$$\begin{aligned} kC_l||u-w||^2_{\mathbf{W}} &= k[1-2\epsilon C_q-4C_\beta(\epsilon+\epsilon^2k^2)]\|u-w\|^2_{\mathbf{W}} \\ &\le (\nabla\widetilde{\mathcal{H}}(u)-\nabla\widetilde{\mathcal{H}}(w),u-w) \\ &\le k[1+2\epsilon C_q+4C_\beta(\epsilon+\epsilon^2k^2)]\|u-w\|^2_{\mathbf{W}} \\ &= kC_u||u-w||^2_{\mathbf{W}} \ , \end{aligned}$$

with constants $C_l = 1-2\epsilon C_q-4C_\beta(\epsilon+\epsilon^2k^2)]$ and $C_u = 1+2\epsilon C_q+4C_\beta(\epsilon+\epsilon^2k^2)$.

So, if ϵ is small such that ϵk is also small, independent of k, then we have the proposition. □

Note that it follows from Proposition 3.2 that the functional $F = k^{-1}\widetilde{\mathcal{H}}$ preserves the continuity and convexity for the class of densities considered here. Consequently, the preconditioned energy is guaranteed to be convex as long as the perturbations to a constant density $\rho(x) = 1+\epsilon g(x)$ remain small enough with $\epsilon = o(k^{-1})$. In other words, we obtain the following.

COROLLARY 3.4. *For any $\rho(x) = 1+\epsilon g(x)$ with ϵ suitably small, the functional $F = k^{-1}\widetilde{\mathcal{H}}$ satisfies the continuity and convexity conditions with $K = C_l, L = C_u$ for all points in $\mathbf{W}$.*

In addition to the convexity and continuity properties, we also need the following conditions on the space decomposition to be satisfied.

Condition A. For all $v \in \mathbf{W}$, there exists $v_i \in \overline{\mathbf{W}}_i$ such that $\sum_{i=1}^J v_i = v$, and $(\sum_{i=1}^J ||v_i||^2_{\mathbf{W}})^{1/2} \le C_1||v||_{\mathbf{W}}$ for some C_1 independent of k.

Condition B. *Strengthened Cauchy–Schwarz.* For all $w_{ij} \in \mathbf{W}, u_i \in \overline{\mathbf{W}}_i, v_j \in \overline{\mathbf{W}}_j \Rightarrow$

$$\sum_{i,j=1}^J (\nabla F(w_{ij}+u_i)-\nabla F(w_{ij}),v_j) \le C_2\left(\sum_{i=1}^J ||u_i||^2_{\mathbf{W}}\right)^{1/2}\left(\sum_{j=1}^J ||v_j||^2_{\mathbf{W}}\right)^{1/2}$$

for some $C_2 \le cL$, where c is independent of k.

These conditions are verified below. First, the following result follows from a simple accounting and the properties of the uniform domain decomposition we defined in section 2.2. Here and later supp denotes the support set of a function.

LEMMA 3.5. *For the one-dimensional* "hat" *basis, for all $u \in \overline{\mathbf{W}}_j$ and all $v \in \overline{\mathbf{W}}_l$ such that $j > l$, the following scaling of the supports relationship holds:*

$$\operatorname{supp}(u) \cap \operatorname{supp}(v) = 2^{-|j-l|}\operatorname{supp}(v).$$

In addition, denote the connected set of common indices $\mathcal{I} = \{0 \le i \le k+1|\ x_i^J \in \operatorname{supp}(u)\cap\operatorname{supp}(v)\}$ and define a restricted norm to be

$$||v||^2_{\mathbf{W},\operatorname{supp}(u)\cap\operatorname{supp}(v)} = \frac{1}{k}\sum_{i\in\mathcal{I}}(v_i-v_{i-1})^2 \ .$$

Then the following norm equivalence up to scaling is satisfied:

$$||v||_{\mathbf{W},\operatorname{supp}(u)\cap\operatorname{supp}(v)} = 2^{-|j-l|}||v||_{\mathbf{W}}. \tag{3.11}$$

The result of this lemma follows by counting the number of common nodes in the intersection of the supports of functions at different levels of the hierarchical decomposition and from the properties of the piecewise linear functions.

With the help of Lemma 3.5, we can show that for the chosen decomposition, both Conditions A and B are satisfied.

THEOREM 3.6. *For the nested subspace decomposition with the choice of the "hat" basis functions,* $(\sum_{i=1}^{J} ||v_i||_{\mathbf{W}}^2)^{1/2} = ||v||_{\mathbf{W}}$, *so that* $C_1 = 1$. *Moreover, for* $F = k^{-1}\widetilde{\mathcal{H}}$, C_2 *can be estimated as* $C_2 = L \cdot \max_j (\sum_{l=1}^{J} 2^{-|j-l|}) \le 2L$.

Proof. Notice that "hat" functions form an orthogonal basis, so

$$\left(\sum_{i=1}^{J} ||v_i||_{\mathbf{W}}^2 \right)^{1/2} = ||v||_{\mathbf{W}}$$

follows easily from calculation. As for the C_2, without loss of generality fix $j > l$ and take any $w \in \mathbf{W}$ and $u \in \overline{\mathbf{W}}_j$, $v \in \overline{\mathbf{W}}_l$. Now, adopting a similar argument for those in the proofs of Propositions 3.1 and 3.2 and taking notice that the contributions are confined in the support regions of u and v, we get

$$\begin{aligned}
&(\nabla\widetilde{\mathcal{H}}(u+w) - \nabla\widetilde{\mathcal{H}}(w), v) \\
&= \sum_{i=1}^{k+1} \{a_i\, b_i - 2\epsilon[(Q_i(u+w) - Q_i(w))v_i + 2(\beta_i(u+w) - \beta_i(w))v_i]\} \\
&= \sum_{i\in\mathcal{I}} (v_i - v_{i-1})(u_i - u_{i-1}) - 2\epsilon[(Q_i(u+w) - Q_i(w))v_i + 2(\beta_i(u+w) - \beta_i(w))v_i] \\
&\le k||u||_{\mathbf{W},\mathrm{supp}(u)\cap\mathrm{supp}(v)}||v||_{\mathbf{W},\mathrm{supp}(u)\cap\mathrm{supp}(v)} \\
&\quad + 2\epsilon C_q||u||_{\mathbf{W},\mathrm{supp}(u)\cap supp(v)}||v||_{\mathbf{W},\mathrm{supp}(u)\cap\mathrm{supp}(v)} \\
&\quad + 4kC_\beta(\epsilon + \epsilon^2 k^2)||u||_{\mathbf{W},\mathrm{supp}(u)\cap\mathrm{supp}(v)}||v||_{\mathbf{W},\mathrm{supp}(u)\cap\mathrm{supp}(v)} \\
&\le kL||u||_{\mathbf{W}} \cdot ||v||_{\mathbf{W},\mathrm{supp}(u)\cap\mathrm{supp}(v)},
\end{aligned}$$

where $\mathcal{I}$ is the same as in Lemma 3.5, $L = 1 + 2\epsilon C_q + 4C_\beta(\epsilon + \epsilon^2 k^2)$ is the continuity constant defined in Corollary 3.4, and C_β, C_q are as defined in Proposition 3.2.

Now by Lemma 3.5, $||v||_{\mathbf{W},\mathrm{supp}(u)\cap\mathrm{supp}(v)} = 2^{-|j-l|}||v||_{\mathbf{W}}$. It follows that, since for any symmetric matrix $||Ax|| \le \rho(A)||x|| \le \max_i (\sum_{j=1}^{n} |a_{ij}|)||x||$,

$$\begin{aligned}
\sum_{i,j=1}^{J} (\nabla\widetilde{\mathcal{H}}(w_{ij} + u_i) - \nabla\widetilde{\mathcal{H}}(w_{ij}), v_j) &\le kL \sum_{i,j=1}^{J} 2^{-|i-j|}||u_i||_{\mathbf{W}}||v_j||_{\mathbf{W}} \\
&\le kL \left(\max_j \sum_{i=1}^{J} 2^{-|i-j|} \right) \left(\sum_{i=1}^{J} ||u_i||_{\mathbf{W}}^2 \right)^{1/2} \left(\sum_{j=1}^{J} ||v_j||_{\mathbf{W}}^2 \right)^{1/2}.
\end{aligned}$$

Henceforth, we have $C_2 = L \cdot \max_j \sum_{i=1}^{J} 2^{-|i-j|} \le 2L$.

Note that although the proof above is presented for the special case of the "hat" basis that we used in our numerical implementation, similar arguments can be used to show that Conditions A and B hold for other suitable bases. Indeed, since the energy part of the argument does not depend on the decomposition, Conditions A and B will remain true for any basis as long as

$$\left(\sum_{i=1}^{J} ||v_i||_{\mathbf{W}}^2 \right)^{1/2} \le C_1||v||_{\mathbf{W}}$$

and

$$||v||_{\mathbf{W},\mathrm{supp}(v)\cap\mathrm{supp}(u)} \le C\cdot 2^{-|j-l|}||v||_{\mathbf{W}}$$

hold for some constants C, C_1 independent of k.

3.3. Uniform convergence theorem. Finally, putting together Conditions A and B and using convexity and continuity of $F = k^{-1}\widetilde{\mathcal{H}}$ in $\mathbf{W}$, we are ready to prove the following uniform convergence result.

THEOREM 3.7. *Under Conditions* A *and* B *on space decomposition, Algorithm* 2.1 *converges uniformly in* $\mathbf{W}$ *for any density of type* $\rho(x) = 1 + \epsilon g(x)$ *with sufficiently small* ϵ. *Moreover,* $d_n = \widetilde{\mathcal{H}}(u_n) - \widetilde{\mathcal{H}}(u)$ *satisfies*

$$d_n \le r d_{n-1}, \quad r \in (0,1),$$

where $r = \frac{C}{1+C}$ *and* $C = C_1^2 C_2^2 L/K^3$.

Before proceeding to its proof, let us first state a consequence.

COROLLARY 3.8. *In the case of a "hat" basis, the constants* C_1 *and* C_2 *can be estimated as* $C_1 = 1$ *and* $C_2 = 2L$, *so, for example, when* $\rho(x) = 1$, $C = 4$.

The proof of this result is similar to the one given in [37] and relies on the following inequality:

$$F(u) - F(v) - (\nabla F(v), u - v) = \int_0^1 (u - v, \nabla F(u + t(v-u)) - \nabla F(v))dt$$

$$\ge K\int_0^1 ||(1-t)(u-v)||_{\mathbf{W}}^2 \frac{dt}{1-t} = \frac{K}{2}||u-v||_{\mathbf{W}}^2 . \tag{3.12}$$

3.4. Proof of Theorem 3.7. Denote u to be the exact solution of (2.3). Consider u_n as the approximate solution after one J-level iteration of the Algorithm 2.2. At the ith level, since the supports of the basis functions w_j^i are disjoint in $\overline{\mathbf{W}}_i$, we have

$$u_{n+\frac{i}{J}} = u_{n+\frac{i-1}{J}} + e_n^i, \quad \widetilde{\mathcal{H}}(u_{n+\frac{i-1}{J}} + e_n^i) \le \widetilde{\mathcal{H}}(u_{n+\frac{i-1}{J}} + v_i) \quad \forall v_i \in \overline{\mathbf{W}}_i,$$

where in the notation of the algorithm, $e_n^i = \sum_{l=1}^{\nu_1}\sum_{j=1}^{n_i}\alpha_{jl}^0 w_j^i \in \overline{\mathbf{W}}_i$.

First notice that since the minimizer $u_{n+\frac{i}{J}}$ satisfies $(\nabla\widetilde{\mathcal{H}}(u_{n+\frac{i}{J}}), v) = 0$ for all $v \in \overline{\mathbf{W}}_i$, it follows from (3.12) that

$$\begin{aligned}
\widetilde{\mathcal{H}}(u_n) - \widetilde{\mathcal{H}}(u_{n+1}) &= \sum_{i=1}^{J}(\widetilde{\mathcal{H}}(u_{n+\frac{i-1}{J}}) - \widetilde{\mathcal{H}}(u_{n+\frac{i}{J}})) \\
&\ge \sum_{i=1}^{J}\left((\nabla\widetilde{\mathcal{H}}(u_{n+\frac{i}{J}}), u_{n+\frac{i-1}{J}} - u_{n+\frac{i}{J}}) + \frac{K}{2}||u_{n+\frac{i-1}{J}} - u_{n+\frac{i}{J}}||_{\mathbf{W}}^2\right) \\
&= \frac{K}{2}\sum_{i=1}^{J}||e_n^i||_{\mathbf{W}}^2 .
\end{aligned}$$

Next, let us use Condition A to decompose $u_{n+1} - u = \sum_{i=1}^{J} v_i$. Then

$$\begin{aligned}(\nabla\widetilde{\mathcal{H}}(u_{n+1}) - \nabla\widetilde{\mathcal{H}}(u), u_{n+1} - u) &= (\nabla\widetilde{\mathcal{H}}(u_{n+1}), u_{n+1} - u)\\ &= \sum_{i=1}^{J} (\nabla\widetilde{\mathcal{H}}(u_{n+1}), v_i) = \sum_{i=1}^{J} (\nabla\widetilde{\mathcal{H}}(u_{n+1}) - \nabla\widetilde{\mathcal{H}}(u_{n+\frac{i}{J}}), v_i)\\ &= \sum_{i=1}^{J}\sum_{j\geq i+1}^{J} (\nabla\widetilde{\mathcal{H}}(u_{n+\frac{j}{J}}) - \nabla\widetilde{\mathcal{H}}(u_{n+\frac{j-1}{J}}), v_i)\\ &\leq C_2\Big(\sum_{j=1}^{J} ||e_n^j||_{\mathbf{W}}^2\Big)^{1/2}\Big(\sum_{i=1}^{J} ||v_i||_{\mathbf{W}}^2\Big)^{1/2}.\end{aligned}$$

Hence

$$\begin{aligned}(\nabla\widetilde{\mathcal{H}}(u_{n+1}) - \nabla\widetilde{\mathcal{H}}(u), u_{n+1} - u) &\leq C_1C_2\Big(\sum_{j=1}^{J} ||e_n^j||_{\mathbf{W}}^2\Big)^{1/2} ||u_{n+1} - u||_{\mathbf{W}}\\ &\leq C_1C_2\left(\frac{2}{K}(\widetilde{\mathcal{H}}(u_n) - \widetilde{\mathcal{H}}(u_{n+1}))\right)^{1/2} ||u_{n+1} - u||_{\mathbf{W}}\,.\end{aligned}$$

Denote $d_n = \widetilde{\mathcal{H}}(u_n) - \widetilde{\mathcal{H}}(u)$; then $\widetilde{\mathcal{H}}(u_n) - \widetilde{\mathcal{H}}(u_{n+1}) = d_n - d_{n+1}$, and it follows from the inequality above that

$$\left(\frac{2}{K}(d_n - d_{n+1})\right)^{1/2} \geq \frac{(\nabla\widetilde{\mathcal{H}}(u_{n+1}) - \nabla\widetilde{\mathcal{H}}(u), u_{n+1} - u)}{C_1C_2||u_{n+1} - u||_{\mathbf{W}}}\,.$$

Thus,

$$\begin{aligned}d_n - d_{n+1} &\geq \frac{K}{2}\left(\frac{(\nabla\widetilde{\mathcal{H}}(u_{n+1}) - \nabla\widetilde{\mathcal{H}}(u), u_{n+1} - u)}{C_1C_2||u_{n+1} - u||_{\mathbf{W}}}\right)^2\\ &\geq \frac{K}{2}(C_1C_2)^{-2}K^2||u_{n+1} - u||_{\mathbf{W}}^2 \geq \frac{K^3}{C_1^2C_2^2L}d_{n+1}\,.\end{aligned}$$

The last step of the argument uses (3.2) and an inequality similar to (3.12) for the upper bound:

$$d_{n+1} = \widetilde{\mathcal{H}}(u_{n+1}) - \widetilde{\mathcal{H}}(u) \leq \frac{L}{2}||u_{n+1} - u||_{\mathbf{W}}^2\,.$$

As a consequence, we get

$$d_{n+1} \leq \frac{C_1^2C_2^2L}{K^3}(d_n - d_{n+1}) \Rightarrow d_{n+1} \leq \frac{C}{1+C}d_n, \quad \text{where } C = \frac{C_1^2C_2^2L}{K^3}\,.$$

Since for any basis satisfying Conditions A and B, $C_1^2C_2^2L/K^3 = O(L^3/K^3)$, the convergence is uniform as long as the ratio L/K does not depend on k. By looking at the L and K estimates obtained in Proposition 3.2, it is easy to see that this condition is satisfied for any smooth perturbation of the constant density $1 + \epsilon g(x)$ with ϵ of the order of $O(1/k)$. This concludes the proof of the main theorem. ∎

It follows that, for a suitable choice of decomposition in one dimension, the asymptotic convergence factor of our multilevel algorithm is independent of the size of the problem and the number of grid levels, which gives a significant speedup compared to other methods, like the traditional Lloyd iteration. This claim is further substantiated by the following numerical examples. Although for the sake of simplicity we have presented the detailed theory only for the case of the $V(\nu)$ multigrid cycle with no postsmoothings, same conclusions can be drawn for the case of the full $V(\nu_1, \nu_2)$ cycle, and the results of the numerical experiments in both cases are outlined below.

4. Numerical examples. We now report some numerical results obtained using the new multilevel algorithm.

First we compare the results of our $V(1,1)$ multigrid implementation with the usual Lloyd iteration for the one-dimensional problem. Then we present some results for a two-dimensional test problem in a parallelogram domain. The results are obtained with the MATLAB 6.5 implementation of the new algorithm. The test runs were performed on a PC with a Pentium IV processor and 512MB RAM.

The one-dimensional implementation is very straightforward. Here, we take the unit interval and test a couple of different density functions, like $\rho(x) = 1$ and $\rho(x) = 1 + x$. We plot the convergence factor $\rho \approx |z_{n+1} - z_n|/|z_n - z_{n-1}|$ for each $V(1,1)$ cycle with respect to the total number of generators (grid points) involved.

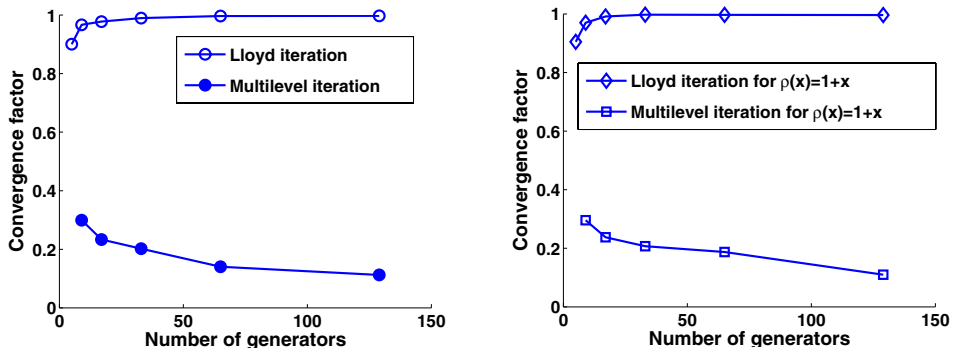


FIG. 4.1. *Plot of the convergence factor vs. the number of generators for the regular Lloyd (upper) and the multilevel (lower curves) iterations for* $\rho(x) = 1$ *(left) and* $\rho(x) = 1 + 0.1x$ *(right).*

Figure 4.1 substantiates the fact that the speed of convergence for the proposed scheme remains nearly constant as the number of generators increases. In Figure 4.2, the computational time needed for the $V(1,0)$ implementation of the multilevel method to reach 10^{-12} accuracy is given for $\rho(x) = 1$ and $\rho(x) = 1 + x$, respectively. The graph shows that in the one-dimensional case, the computational time *scales almost linearly* with the problem size.

The data in Table 4.1 shows the stabilization of the number of multigrid cycles $V(\nu_1, \nu_2)$ needed to reduce the error to $\epsilon = 10^{-12}$ in the constant density case. While there is a visible difference between the number of iterations required for $V(1,1)$ and $V(2,2)$ cycles, respectively, saturation occurs if the values of the relaxation parameters are increased, which is why $\nu_{1,2} \leq 2$ in most of our calculations. The geometric rate of the energy and error reduction asserted by Theorem 3.7 are confirmed by the experiments. Indeed, Figure 4.3 shows the convergence history of a $V(1,1)$ cycle against the total number of relaxations for the $k = 129$ case.

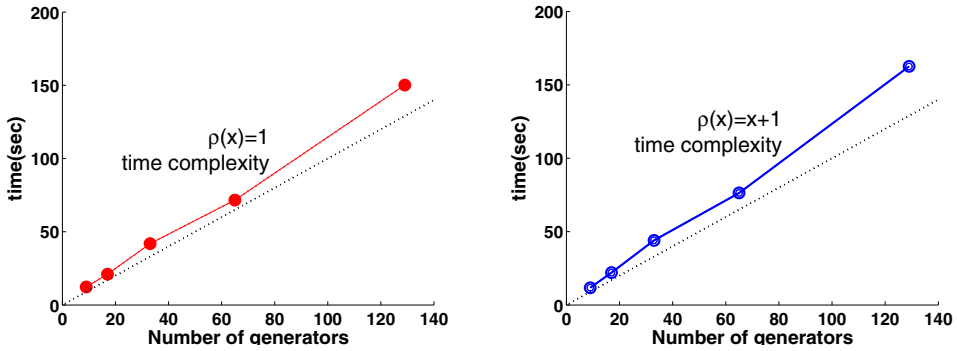


FIG. 4.2. *Computational time vs. problem size for the one-dimensional implementation.*

TABLE 4.1
The number of $V(\nu_1, \nu_2)$ cycles needed vs. the number of generators.

$k/V(\nu_1,\nu_2)$	$V(1,0)$	$V(0,1)$	$V(1,1)$	$V(2,0)$	$V(0,2)$	$V(2,2)$
3	7	8	6	6	7	4
5	11	11	8	8	8	6
9	13	14	9	9	9	7
17	18	18	12	12	12	8
33	21	20	13	12	13	8
65	21	22	12	12	12	8
129	21	21	12	12	12	8
257	20	23	12	12	13	7
513	20	22	12	11	13	7
1025	19	22	11	11	13	7

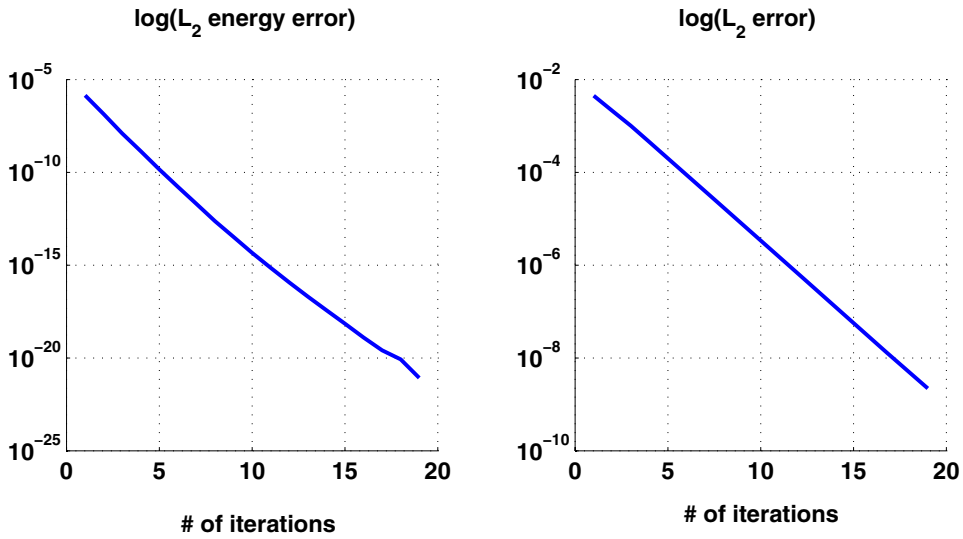


FIG. 4.3. *The energy reduction (left) and the convergence history (right) for* 129 *generators in the log-normal scale.*

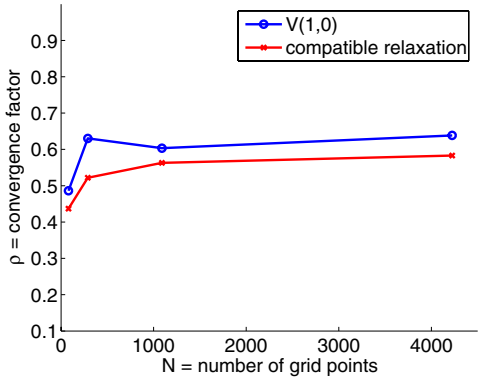


FIG. 4.4. *Comparison of the convergence factors for the compatible relaxation and a* $V(1,0)$ *cycle for* $\rho = 1$ *on a two-dimensional parallelogram-shaped domain.*

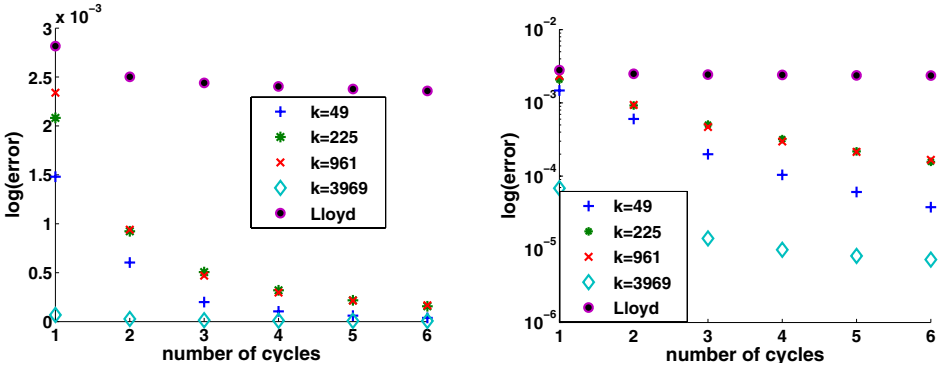


FIG. 4.5. *Convergence history for the two-dimensional multigrid scheme for* $\rho = 1$ *on a parallelogram-shaped domain compared to the Lloyd scheme shown on the top curve in the normal scale (left) and log-normal scale (right).*

Next, the convergence factors for a two-dimensional problems with $\rho = 1$ on a parallelogram domain are compared in Figure 4.4. The lower line in the graph shows the convergence factor for the compatible relaxation, that is, a relaxation performed on the grid with the exact solution given at the coarse nodes (see [2] for discussions on the compatible relaxation). This factor can serve as a lower bound on the convergence factor of the full multigrid cycle, and the quality of the coarse grid influences the distance between the two graphs. As we can see from Figure 4.4, the result of the compatible relaxation in this case is in good agreement with the convergence factor of the whole $V(1,0)$ cycle given on top, which is an indication of a good quality for the coarsening procedure.

The convergence history plots are given in Figure 4.5. The top curves in both graphs depict the error reductions given by the Lloyd iteration, while those below correspond to the convergence of the multigrid scheme for various problem sizes. One can clearly see that the slopes of error reduction on the logarithmic scale do not depend on the number of generators. Hence, even though our theoretical results here are proved only in one dimension, the proposed algorithm preserves its characteristics in certain higher-dimensional implementations.

5. Conclusion. Recently, several methods have been proposed for accelerating the convergence of the classical Lloyd iteration commonly used in the context of quantization and in the construction of centroidal Voronoi tessellations [9, 10]. These algorithmic advances are important for making the computation of optimal quantizers more efficient and for many other successful applications of CVTs. A few possible extensions that use multilevel techniques to accelerate the convergence of the CVTs have been suggested. One of such extensions uses some algebraic multigrid solvers in the spirit of [3, 7, 35] as preconditioners to accelerate the solution of the linear system at every Newton iteration [9], while the other adopts an energy-based nonlinear multigrid approach with the use of a dynamic nonlinear preconditioning [10]. In this paper, we focus on the latter approach, and for the first time, a rigorous analysis of its convergence properties is presented for a class of one-dimensional density functions, with the results of several numerical experiments given for both one- and two-dimensional cases. A more detailed analysis in the higher-dimensional cases as well as more efficient and robust implementations of the energy-based nonlinear multigrid approach are presently under investigation [12]. We conclude by commenting that there is obviously a great potential in using such multigrid methods to accelerate the optimal quantizer design and more generally the computation of CVTs.

Acknowledgment. The authors would like to thank M. Gunzburger of FSU and L. Zikatanov and J. Xu of PSU for interesting discussions on the subject, and anonymous referees for constructive criticism that helped to make current presentation more consistent.

REFERENCES

[1] F. Aurenhammer, *Voronoi diagrams. A survey of a fundamental geometric data structure*, ACM Comput. Surveys, 23 (1990), pp. 345–405.
[2] A. Brandt, *General highly accurate algebraic coarsening*, Electron. Trans. Numer. Anal., 10 (2000), pp. 1–20.
[3] M. Brezina, A. J. Cleary, R. D. Falgout, V. E. Henson, J. E. Jones, T. A. Manteuffel, S. F. McCormick, and J. W. Ruge, *Algebraic multigrid based on element interpolation (AMGe)*, SIAM J. Sci. Comput., 22 (2000), pp. 1570–1592.
[4] W. L. Briggs, V. E. Henson, and S. F. McCormick, *A Multigrid Tutorial*, 2nd ed., SIAM, Philadelphia, 2000.
[5] M. Cappellari and Y. Copin, *Adaptive spatial binning of integral-field spectroscopic data using Voronoi tessellations*, Monthly Notices Roy. Astronom. Soc., 342 (2003), pp. 345–354.
[6] T. Chan and I. Sharapov, *Subspace correction multilevel methods for elliptic eigenvalue problems*, Numer. Linear Algebra Appl., 1 (1993), pp. 1–7.
[7] Q. Chang and Z. Huang, *Efficient algebraic multigrid algorithms and their convergence*, SIAM J. Sci. Comput., 24 (2002), pp. 597–618.
[8] J. Cortes, S. Martinez, T. Karata, and F. Bullo, *Coverage control for mobile sensing networks*, IEEE Trans. Robotics Automation, 20 (2004), pp. 243–255.
[9] Q. Du and M. Emelianenko, *Acceleration of algorithms for the computation of centroidal Voronoi tessellations*, Numer. Linear Algebra Appl., 13 (2006), pp. 173–192.
[10] Q. Du and M. Emelianenko, *Uniform convergence of a multilevel energy-based quantization scheme*, in Domain Decomposition Methods in Science and Engineering XVI, Lect. Notes Comput. Sci. Eng. 55, O. B. Widlund and D. E. Keyes, eds., Springer, Berlin, 2007, pp. 533–541.
[11] Q. Du, M. Emelianenko, and L. Ju, *Convergence properties of the Lloyd algorithm for computing the centroidal Voronoi tessellations*, SIAM J. Numer. Anal., 44 (2006), pp. 102–119.
[12] Q. Du, M. Emelianenko, and L. Zikatanov, *An Energy-Based Multilevel Quantization Scheme in Multidimension*, preprint, 2007.
[13] Q. Du, V. Faber, and M. Gunzburger, *Centroidal Voronoi tessellations: Applications and algorithms*, SIAM Rev., 41 (1999), pp. 637–676.

[14] Q. Du and M. Gunzburger, *Grid generation and optimization based on centroidal Voronoi tessellations*, Appl. Math. Comput., 133 (2002), pp. 591–607.
[15] Q. Du, M. Gunzburger, and L. Ju, *Meshfree, probabilistic determination of points, support spheres, and connectivities for meshless computing*, Comput. Methods Appl. Mech. Engrg., 191 (2002), pp. 1349–1366.
[16] Q. Du, M. Gunzburger, and L. Ju, *Constrained centroidal Voronoi tessellations for surfaces*, SIAM J. Sci. Comput., 24 (2003), pp. 1488–1506.
[17] Q. Du and D. Wang, *Tetrahedral mesh generation and optimization based on centroidal Voronoi tessellations*, Internat. J. Numer. Methods Engrg., 56 (2002), pp. 1355–1373.
[18] Q. Du and D. Wang, *Recent progress in robust and quality Delaunay mesh generation*, J. Comput. Appl. Math., 195 (2006), pp. 8–23.
[19] R. Dwyer, *Higher-dimensional Voronoi diagrams in linear expected time*, Discrete Comput. Geom., 6 (1991), pp. 343–367.
[20] M. Emelianenko, L. Ju, and A. Rand, *Nondegeneracy and weak global convergence of the Lloyd algorithm in* $\mathbb{R}^d$, SIAM J. Numer. Anal., 46 (2008), pp. 1423–1441.
[21] S. Fortune, *A sweepline algorithm for Voronoi diagrams*, Algorithmica, 2 (1987), pp. 153–174.
[22] E. Gelman and J. Mandel, *On multilevel iterative methods for optimization problems*, Math. Programming, 48 (1990), pp. 1–17.
[23] A. Gersho, *Asymptotically optimal block quantization*, IEEE Trans. Inform. Theory, 25 (1979), pp. 373–380.
[24] R. Gray and D. Neuhoff, *Quantization*, IEEE Trans. Inform. Theory, 44 (1998), pp. 2325–2383.
[25] W. Hackbusch and A. Reusken, *Analysis of a damped nonlinear multilevel method*, Numer. Math., 55 (1989), pp. 225–246.
[26] S. Hiller, H. Hellwig, and O. Deussen, *Beyond stippling methods for distributing objects on the plane*, Computer Graphics Forum, 22 (2003), pp. 515–522.
[27] L. Ju, Q. Du, and M. Gunzburger, *Probabilistic methods for centroidal Voronoi tessellations and their parallel implementations*, Parallel Comput., 28 (2002), pp. 1477–1500.
[28] Y. Koren and I. Yavneh, *Adaptive multiscale redistribution for vector quantization*, SIAM J. Sci. Comput., 27 (2006), pp. 1573–1593.
[29] Y. Koren, I. Yavneh, and A. Spira, *A multigrid approach to the* 1*-D quantization problem*, IEEE Trans. Inform. Theory, 51 (200), pp. 2993–2998.
[30] R. Kornhuber, *Adaptive Monotone Multigrid Methods for Nonlinear Variational Problems*, Teubner, Stuttgart, 1997.
[31] Y. Linde, A. Buzo, and R. Gray, *An algorithms for vector quantizer design*, IEEE Trans. Commun., 28 (1980), pp. 84–95.
[32] S. Lloyd, *Least square quantization in PCM*, IEEE Trans. Inform. Theory, 28 (1982), pp. 129–137.
[33] A. Mendes and I. Themido, *Multi-outlet retail site location assessment*, Int. Trans. Oper. Res., 11 (2004), pp. 1–18.
[34] A. Okabe, B. Boots, and K. Sugihara, *Spatial Tessellations; Concepts and Applications of Voronoi Diagrams*, Wiley, Chichester, UK, 1992.
[35] J. W. Ruge and K. Stüben, *Algebraic multigrid*, in Multigrid Methods, S. F. McCormick, ed., Frontiers in Appl. Math. 3, SIAM, Philadelphia, 1987, pp. 73–130.
[36] J. Sabin and R. Gray, *Global convergence and empirical consistency of the generalized Lloyd algorithm*, IEEE Trans. Inform. Theory, 32 (1986), pp. 148–155.
[37] X. Tai and J. Xu, *Global convergence of subspace correction methods for some convex optimization problems*, Math. Comp., 71 (2002), pp. 105–124.
[38] A. Trushkin, *On the design of an optimal quantizer*, IEEE Trans. Inform. Theory, 39 (1993), pp. 1180–1194.
[39] S. Valette and J. Chassery, *Approximated centroidal Voronoi diagrams for uniform polygonal mesh coarsening*, Computer Graphics Forum, 23 (2004), pp. 381–390.
[40] C. Wager, B. Coull, and N. Lange, *Modelling spatial intensity for replicated inhomogeneous point patterns in brain imaging*, J. R. Stat. Soc. Ser. B Stat. Methodol., 66 (2004), pp. 429–446.
[41] J. Xu and L. Zikatanov, *The method of alternating projections and the method of subspace corrections in Hilbert space*, J. Amer. Math. Soc., 15 (2002), pp. 573–597.

SIAM J. NUMER. ANAL.
Vol. 46, No. 3, pp. 1503–1518
© 2008 Society for Industrial and Applied Mathematics

A HYBRID POLYNOMIAL SYSTEM SOLVING METHOD FOR MIXED TRIGONOMETRIC POLYNOMIAL SYSTEMS*

BO YU† AND BO DONG†

Abstract. Mixed trigonometric polynomial systems arise in many fields of science and engineering. Commonly, this class of systems is transformed into polynomial systems by variable substituting and adding some quadratic equations, and then solved by some polynomial system solving method. In this paper, by exploiting the special structure of the additional quadratic equations, an efficient hybrid method for solving polynomial systems coming from mixed trigonometric polynomial systems is presented. It combines the homotopy method, in which the homotopy is a combination of coefficient parameter homotopy and the random product homotopy, with decomposition, variable substitution, and reduction techniques. Numerical tests are given to show its effectiveness, and it is applied to solve a practical problem.

Key words. homotopy method, polynomial system, mixed trigonometric polynomial system, symbolic-numeric computation

AMS subject classifications. 65H10, 65H20, 68W30

DOI. 10.1137/070681740

1. Introduction.

In this paper, we consider the task of finding all solutions of mixed trigonometric polynomial systems. Commonly, this class of systems is transformed into polynomial systems by variable substituting and adding some quadratic equations, and then solved by some polynomial solving method, symbolic or numeric. In this paper, by exploiting the special structure of the additional quadratic equations, an efficient hybrid method for solving polynomial systems coming from mixed trigonometric polynomial systems is presented. It combines the homotopy method, in which the homotopy is a combination of coefficient parameter homotopy and the random product homotopy, with decomposition, variable substitution, and reduction techniques.

The paper is organized as following: The problem with some examples of applications and an introduction to homotopy methods for solving polynomial systems are given in the rest of this section. A hybrid method for solving polynomial systems coming from mixed trigonometric polynomial systems is given in section 2. The main theorem with proof is given in section 3. Numerical examples with application are given in section 4.

1.1. Mixed trigonometric polynomial systems.

Mixed trigonometric polynomial systems (MTPS) are shown as follows:

$$F(y,\theta) = (f_1(y,\theta),\ldots,f_{n+m}(y,\theta))^T = 0, \tag{1.1}$$

where $y = (y_1,\ldots,y_n)$, $\theta = (\theta_1,\ldots,\theta_m)$, and for $1 \le i \le n+m$,

$$f_i(y,\theta) = \sum_{\alpha\in S_i} b_{i\alpha} y^{\beta} \sin^{\mu}\theta \cos^{\nu}\theta + b_i, \tag{1.2}$$

*Received by the editors February 5, 2007; accepted for publication (in revised form) December 18, 2007; published electronically April 4, 2008. This research was supported by the National Natural Science Foundation of China (10671029) and the Research Fund for the Doctoral Program of Higher Education (20060141029).

http://www.siam.org/journals/sinum/46-3/68174.html

†Department of Applied Mathematics, Dalian University of Technology, Dalian, Liaoning 116024, China (yubo@dlut.edu.cn, dongbodlut@gmail.com).

where $\alpha = (\beta,\ \mu,\ \nu)$, $S_i \subset \mathcal{N}^{n+2m}$ (here, for convenience, we write the constant term of $f_i(y,\theta)$ explicitly, i.e., for all i, $0 \notin S_i$), and

$$\begin{cases} y^{\beta} = y_1^{\beta_1} \cdots y_n^{\beta_n}, \\ \sin^{\mu}\theta = \sin^{\mu_1}\theta_1 \cdots \sin^{\mu_m}\theta_m, \\ \cos^{\nu}\theta = \cos^{\nu_1}\theta_1 \cdots \cos^{\nu_m}\theta_m. \end{cases} \tag{1.3}$$

This class of systems is common in many fields of science and engineering; see, for example, its application in neurophysiology [33], the inverse position and kinematics problems [33, 8], and the PUMA robot [26, 27].

In the following we give two examples of MTPS.

Example 1.1. This is a simple example in two dimensions.

$$F(y,\theta) = \begin{pmatrix} \sin^3\theta + y^2 + 1 \\ \cos^3\theta + y\sin\theta - 2 \end{pmatrix} = 0. \tag{1.4}$$

Example 1.2. This is a practical problem we meet in signal processing of sonar and radar, which is the original motivation of this research.

$$\begin{cases} \sum\limits_{i=1}^{n} y_i = b_1, & \sum\limits_{i=1}^{n} y_i^2 = b_2, \\ \sum\limits_{i=1}^{n} y_i \sin\theta_i = b_3, & \sum\limits_{i=1}^{n} y_i \cos\theta_i = b_4, \\ \sum\limits_{i=1}^{n} y_i \sin^2\theta_i = b_5, & \sum\limits_{i=1}^{n} y_i \sin\theta_i \cos\theta_i = b_6, \\ \sum\limits_{i=1}^{n} y_i^2 \sin\theta_i = b_7, & \sum\limits_{i=1}^{n} y_i^2 \cos\theta_i = b_8, \\ \sum\limits_{i=1}^{n} y_i^2 \sin^2\theta_i = b_9, & \sum\limits_{i=1}^{n} y_i^2 \sin\theta_i \cos\theta_i = b_{10}, \end{cases} \tag{1.5}$$

where $n = 3$ to 5 (if $n < 5$, then only the first $2n$ equations are needed).

By the variable substitution

$$\begin{cases} x_i = y_i, & i = 1, \ldots, n, \\ x_{n+j} = \sin\theta_j, & j = 1, \ldots, m, \\ x_{n+m+k} = \cos\theta_k, & k = 1, \ldots, m, \end{cases} \tag{1.6}$$

and adding m quadratic equations,

$$x_{n+k}^2 + x_{n+m+k}^2 - 1 = 0, \quad k = 1, \ldots, m,$$

the MTPS $F(y,\theta) = 0$ in (1.1), (1.2), and (1.3) can be transformed into the polynomial system

$$P(x) = \begin{pmatrix} p_1(x) + b_1 \\ \cdots \\ p_{n+m}(x) + b_{n+m} \\ x_{n+1}^2 + x_{n+m+1}^2 - 1 \\ \cdots \\ x_{n+m}^2 + x_{n+2m}^2 - 1 \end{pmatrix} = 0, \tag{1.7}$$

where

$$p_i(x) = \sum_{\alpha \in S_i} b_{ij} x^{\alpha}. \tag{1.8}$$

The transformed polynomial systems from Examples 1.1 and 1.2 are as follows.

Example 1.1 (continued).

$$P(x) = \begin{pmatrix} x_2^3 + x_1^2 + 1 \\ x_3^3 + x_2 x_1 - 2 \\ x_2^2 + x_3^2 - 1 \end{pmatrix} = 0. \tag{1.9}$$

Example 1.2 (continued).

$$\begin{cases} \sum_{i=1}^{n} x_i = b_1, & \sum_{i=1}^{n} x_i^2 = b_2, \\ \sum_{i=1}^{n} x_i x_{n+i} = b_3, & \sum_{i=1}^{n} x_i x_{n+m+i} = b_4, \\ \sum_{i=1}^{n} x_i x_{n+i}^2 = b_5, & \sum_{i=1}^{n} x_i x_{n+i} x_{n+m+i} = b_6, \\ \sum_{i=1}^{n} x_i^2 x_{n+i} = b_7, & \sum_{i=1}^{n} x_i^2 x_{n+m+i} = b_8, \\ \sum_{i=1}^{n} x_i^2 x_{n+i}^2 = b_9, & \sum_{i=1}^{n} x_i^2 x_{n+i} x_{n+m+i} = b_{10}, \\ x_{n+1}^2 + x_{2n+1}^2 = 1, & \\ \cdots & \\ x_{2n}^2 + x_{3n}^2 = 1. & \end{cases} \tag{1.10}$$

1.2. Methods for solving polynomial systems. The primary methods for solving polynomial systems include some elimination methods, e.g., Gröbner methods [3]; Wu's method [45]; (sparse) resultant methods [4]; and the numerical homotopy method. The homotopy method is an efficient numerical method for solving polynomial systems suggested independently by Garcia and Zangwill [6] and Drexler [5] in 1977. The typical homotopy method for solving $P(x) = 0$ is to construct a homotopy

$$H(x,t) = (1-t)Q(x) + tP(x).$$

If the system $Q(x) = 0$, known as the start system, is chosen correctly, the following three properties hold:

Triviality. The solutions of $Q(x) = 0$ are known.

Smoothness. The solution set of $H(x,t) = 0$ for $0 \le t \le 1$ consists of a finite number of smooth paths, called homotopy paths, each parameterized by $t \in [0,1)$.

Accessibility. Every isolated solutions of $H(x,1) = P(x) = 0$ can be reached by some paths originating at a solution of $H(x,0) = Q(x) = 0$.

Tracing numerically homotopy paths for t from 0 to 1, we can find all isolated solutions of $P(x) = 0$.

Many homotopies with triviality, smoothness, and accessibility have been constructed [4, 5, 8, 9, 14, 15, 16, 17, 18, 19, 23, 24, 25, 26, 27, 32, 34, 35, 38, 39, 43, 44]. To solve a polynomial system $P(x) = (p_1(x), \ldots, p_n(x))^T = 0$ of n polynomials in n

unknowns with deg $p_i(x) = d_i$, the standard homotopies [15, 25, 44, 48], whose start system is

$$Q(x) = \begin{pmatrix} a_1 x_1^{d_1} - b_1 \\ \cdots \\ a_n x_n^{d_n} - b_n \end{pmatrix},$$

need to trace a Bézout number of paths to get all isolated solutions of the polynomial system. However, for most applications, the polynomial system is deficient; i.e., the actual number of solutions is less than the Bézout number. For deficient systems, standard homotopies generate many paths that tend to infinity and hence make wasted computation. So studies on homotopy methods for solving polynomial systems are now focused on exploiting the sparsity of the system, i.e., constructing better homotopy such that the number of homotopy paths to be traced is proportional to the actual number of isolated solutions of the target system.

In [18, 19], Li, Sauer, and Yorke presented the random product homotopy for solving deficient polynomial systems whose solutions at infinity are the union of some linear subspace. In [21], Li and Wang proposed a homotopy keeping subschemes at infinity invariant, which can serve as a general theoretical base of homotopy methods for solving deficient polynomial systems. In [26, 27], Morgan and Sommese suggested the use of multihomogeneous homotopy, i.e., keeping multihomogeneous structures. Some homotopies that keep multihomogeneous structures are given in [35, 36, 41, 46, 47]. In [46], a two-stage homotopy method is proposed for deficient polynomial systems. Verschelde derived the GBQ-algorithm in [37] by taking for different equations different partitions of the sets of unknowns. In [29], Morgan, Sommese, and Wampler presented polynomial product homotopies, in which every component of the start system is the product of some polynomials.

Polynomial systems arising in science and engineering are often parameterized. We have to solve repeatedly polynomial systems with different parameters. For this class of systems, one may first solve a system with generic parameters, then take this system as the start system and construct the homotopy, and then finally trace numerically homotopy paths to get the solution of the desired system with special parameters. In [20, 22, 28], some homotopies for parameterized systems were suggested.

In the mid-seventies, Bernstein [1], Kushnirenko [13], and Khovanskii [11] presented a root count in $(\mathcal{C}^*)^n$ $(\mathcal{C}^* = \mathcal{C} \setminus \{0\})$ for systems of Laurent polynomials. Canny and Rojas [2] named this root count the BKK bound. In [40], based on the idea of Bernstein [1], Verschelde, Verlinden, and Cools presented homotopies exploiting Newton polytopes for solving sparse polynomial systems. The polyhedral homotopy method can solve polynomial systems by applying their special structures. In [9], a more general polyhedral homotopy method was presented when some of the p_i's share the same support, and in fact, the homotopy defined in [9] can be used to establish an independent proof of Bernstein's theorem. A root count in $\mathcal{C}^n$ was presented in [23] by Li and Wang. In [10], Huber and Sturmfels gave a tighter upper bound for the root number of the polynomial system $P(x) = 0$ in $\mathcal{C}^n$. Based on this tighter upper bound, a more efficient homotopy method was presented. We refer to [16, 17] for developments of polyhedral homotopies for polynomial systems.

There are several freely available packages for solving polynomial systems. PHCpack, written by Verschelde [34], implements a variety of homotopies mentioned above, with the exception of polynomial product homotopies. In addition to isolated solutions, an algorithm for handling positive dimensional solutions is implemented. HomLab [32], written by Wampler, runs in the MATLAB environment and

implements the product homotopies and coefficient-parameter homotopy. PHoM, by Kojima et al. [12], is a C++ code and applies polyhedral homotopy to solve square systems (same number of equations as variables). HOMPACK [42, 30] and POLSYS_PLP [43] are a sequence of increasingly sophisticated continuation algorithms written in Fortran. MVOLgl, by Li et al. [14], implements efficient algorithms for the computation of mixed volumes. BERTINI [31] is a C code and features a high-level interface for parameter homotopies (including automatic differentiation) and multiple-precision routines that can adjust precision on the fly.

2. A hybrid polynomial system solving method for MTPS. For its essential complexity, the polynomial system solving problem is far from satisfactorily solved. The above-mentioned symbolic methods and homotopy methods still cannot efficiently solve practical polynomial systems of moderate size. Further effort needs to be made to solve polynomial systems more efficiently by exploiting their special structures and sparsity.

Polynomial systems coming from the mixed trigonometric polynomial system, as in (1.7), have a special structure: the last m equations are $x_{n+i}^2 + x_{n+m+i}^2 - 1 = 0, i = 1, \ldots, m$. Based on this special structure, we present a hybrid method to solve them more efficiently.

2.1. The hybrid homotopy. To solve (1.7) and (1.8), we construct the homotopy as

$$H(x,t) = \gamma(1-t)Q(x) + tP(x), \tag{2.1}$$

where

$$Q(x) = \begin{pmatrix} p_1(x) + a_1 \\ \cdots \\ p_{n+m}(x) + a_{n+m} \\ (a_{11}^1 x_{n+1} + a_{12}^1 x_{n+m+1} + a_{10}^1)(a_{11}^2 x_{n+1} + a_{12}^2 x_{n+m+1} + a_{10}^2) \\ \cdots \\ (a_{m1}^1 x_{n+m} + a_{m2}^1 x_{n+2m} + a_{m0}^1)(a_{m1}^2 x_{n+m} + a_{m2}^2 x_{n+2m} + a_{m0}^2) \end{pmatrix}, \tag{2.2}$$

and γ, $a_1, \ldots, a_{n+m}, a_{10}^1, \ldots, a_{m2}^2$ are random numbers in $\mathcal{C}^*$.

Remark 2.1. We call (2.1) a *hybrid homotopy*, because the upper part of its first $n+m$ components is a special coefficient parameter homotopy, and the lower part of its last m components is a random linear product homotopy.

Remark 2.2. The upper part of the start system keeps the sparse structure of the target system invariant to keep as much information, e.g., solution number and mixed volume, of the target system as possible, which is very useful for solving polynomial systems by the homotopy method. We set a_i $(i = 1, \ldots, n+m)$ as random numbers to ensure nonsingularity of the homotopy. If some components of the target system lack constant terms, we can keep them 0 such that the solution number of the start system is as small as possible.

2.2. The solution procedure. Using the hybrid homotopy, combined with decomposition, variable substitution, and some reduction techniques, we suggest a hybrid solution procedure for MTPS as follows.

HYBRID POLYNOMIAL SYSTEM SOLVING METHOD FOR MTPS.

Step 1. Construct the hybrid homotopy as in (2.1).

Step 2. Decompose the start system into 2^m subsystems:

$$Q_\sigma(x) = \begin{pmatrix} p_1(x) + a_1 \\ \cdots \\ p_{n+m}(x) + a_{n+m} \\ a_{11}^{\sigma_1} x_{n+1} + a_{12}^{\sigma_1} x_{n+m+1} + a_{10}^{\sigma_1} \\ \cdots \\ a_{m1}^{\sigma_m} x_{n+m} + a_{m2}^{\sigma_m} x_{n+2m} + a_{m0}^{\sigma_m} \end{pmatrix} = 0,$$

where $\sigma \in \mathfrak{S} \triangleq \{(\sigma_1, \ldots, \sigma_m) | \sigma_i \in \{1, 2\}\}$.

Step 3. By variable substitution

$$x_{n+m+i} = l_i(x_{n+i}) \triangleq -\frac{a_{11}^{\sigma_1}}{a_{12}^{\sigma_1}} x_{n+i} - \frac{a_{10}^{\sigma_1}}{a_{12}^{\sigma_1}} \tag{2.3}$$

for $i = 1, \ldots, m$, $Q_\sigma(x) = 0$ can be transformed into a new simpler system with fewer variables and equations:

$$\widehat{Q}_\sigma(\widehat{x}) = \begin{pmatrix} q_1(\widehat{x}) + a_1 \\ \cdots \\ q_{n+m}(\widehat{x}) + a_{n+m} \end{pmatrix} = 0,$$

where $\widehat{x} = (x_1, \ldots, x_{n+m})$ and for all $1 \le i \le n + m$,

$$q_i(\widehat{x}) = \sum_{\alpha \in S_i} b_{i\alpha} (x_1, \ldots, x_{n+m}, l_1(x_{n+1}), \ldots, l_m(x_{n+m}))^\alpha.$$

Hence solving $Q(x) = 0$ is equivalent to solving 2^m subsystems $\widehat{Q}_\sigma(\widehat{x}) = 0$, whose Bézout number is $\deg(p_1) \cdots \deg(p_{n+m})$.

Step 4. Reduction: reduce subsystems $\widehat{Q}_\sigma(\widehat{x}) = 0$, if possible, by using some symbolic reduction procedure (e.g., [32, 39]) or by hand, so that they will have fewer Bézout numbers or, in some extreme case, assert that some of them have no solution.

Step 5. Solve the reduced systems of $\widehat{Q}_\sigma(\widehat{x}) = 0$ by one of the homotopy methods given in [15, 16, 17, 18, 19, 20, 24, 26, 27, 36, 37, 38, 40, 48] to obtain all isolated solutions to $Q(x) = 0$. These solutions serve as start points of the homotopy (2.1).

Step 6. Trace homotopy paths generated by the homotopy (2.1) to get all isolated solutions of the target system $P(x) = 0$.

Remark 2.3. If the system $P(x) = 0$ is sparse enough, its solution number as well as the solution number of $Q(x)$ are much less than its Bézout number and multi-homogeneous Bézout number. Then the main work in the solution procedure is to solve the start system $Q(x) = 0$. In Steps 2 and 3, the start system is decomposed into some simpler subsystems and then transformed by variable substitution into systems whose dimensions are reduced by m; when m is big, the hybrid solution procedure will take much less computational time than existing homotopy methods for sparse polynomial systems, e.g., linear product homotopy or polyhedral homotopy, for the following three reasons:

(1) Tracing homotopy paths in $(n + m)$-dimensional space will be considerably faster than doing so in $(n + 2m)$-dimensional space.

(2) Finding the best multihomogeneous structure or mixed subdivision of initial subsystems with less dimension and lower total degree is much easier than that of the target system.

(3) Further potential for saving computational time is in Step 4, since sometimes the subsystems $Q_\sigma(\widehat{x})$ can be considerably reduced in Bézout number even when the target system $P(x)$ can hardly be reduced, as in the following example.

Example 2.1. We consider the system in (1.9). Its total degree and multihomogeneous Bézout number are both 18, and applying standard homotopy or multihomogeneous homotopy to solve $P(x) = 0$ requires tracing 18 homotopy paths.

Applying our new method, the start system is constructed as follows:

$$\begin{pmatrix} x_2^3 + x_1^2 + 1 \\ x_3^3 + x_1x_2 - 2 \\ (x_2 + \mathbf{i}\, x_3)(\mathbf{i}\, x_2 + x_3) \end{pmatrix} = 0. \tag{2.4}$$

Here, for simplicity of expression, without loss of generality, we have fixed random coefficients as $a_{11}^1 = 1$, $a_{12}^1 = \mathbf{i}$, $a_{10}^1 = 0$, $a_{11}^2 = \mathbf{i}$, $a_{12}^2 = 1$, $a_{10}^2 = 0$, where $\mathbf{i} = \sqrt{-1}$. The start system $Q(x) = 0$ can be decomposed into the following two subsystems:

$$Q_{\sigma_1} = \begin{pmatrix} x_2^3 + x_1^2 + 1 \\ x_3^3 + x_1x_2 - 2 \\ x_2 + \mathbf{i}\, x_3 \end{pmatrix} = 0 \tag{2.5}$$

and

$$Q_{\sigma_2} = \begin{pmatrix} x_2^3 + x_1^2 + 1 \\ x_3^3 + x_1x_2 - 2 \\ \mathbf{i}\, x_2 + x_3 \end{pmatrix} = 0. \tag{2.6}$$

By eliminating variable x_2, (2.5) can be equivalently transformed into

$$\begin{pmatrix} x_2^3 + x_1^2 + 1 \\ -\mathbf{i}\, x_2^3 + x_1x_2 - 2 \end{pmatrix} = 0. \tag{2.7}$$

By simple elimination, (2.7) is equivalent to

$$\widehat{Q}_{\sigma_1}(\widehat{x}) = \begin{pmatrix} x_1^3 + x_3^2 + 1 \\ \mathbf{i}\, x_1^2 + x_1x_2 + \mathbf{i} - 2 \end{pmatrix} = 0, \tag{2.8}$$

whose total degree is 6. Performing the same calculations on (2.6), we get

$$\widehat{Q}_{\sigma_2}(\widehat{x}) = \begin{pmatrix} x_1^3 + x_3^2 + 1 \\ \mathbf{i}\, x_1^2 - x_1x_2 + \mathbf{i} + 2 \end{pmatrix} = 0, \tag{2.9}$$

whose total degree is also 6. Hence, applying the standard homotopy to solve (2.8) and (2.9), the total number of homotopy paths that must be traced is 12, which is just the actual solution number of (1.9).

3. The homotopy theorem with proof. The following theorem shows that the hybrid homotopy (2.1) and (2.2) can be used to find all isolated solutions to polynomial systems coming from MTPS.

THEOREM 3.1. *There exists an open dense full-measure subset U of $\mathcal{C}^{n+7m}$, such that for all $\{a_1, \ldots, a_{n+m}, a_{10}^1, \ldots, a_{m2}^2\} \in U$, the solutions to $Q(x) = 0$ in (2.2) are all isolated and regular, and the homotopy map $H(x,t)$ in (2.1) possess the following properties:*

Smoothness. *The solution set to $H(x,t) = 0$ for $0 \le t \le 1$ consists of a finite number of smooth paths, each parameterized by t.*

Accessibility. *Every isolated solution to* $P(x) = 0$ *in* (1.7) *and* (1.8) *can be reached by some path originating at a solution to* $Q(x) = 0$.

The proof of Theorem 3.1 is inspired by [20]. The following preliminary results are needed.

A *linear system* $\mathcal{L}$ is a complex vector space

$$\mathcal{L} = \left\{ \sum_{i=1}^{r} c_i m_i(z) : (c_1, \ldots, c_r) \in \mathcal{C}^r \right\} \tag{3.1}$$

spanned by homogeneous polynomials $m_1(z), \ldots, m_r(z)$ of some fixed degree and parameterized by $\mathcal{C}^r$. The *base locus* of $\mathcal{L}$ is the set

$$\{z \in \mathcal{CP}^n : m_i(z) = 0, i = 1, \ldots, r\}.$$

LEMMA 3.2 (Bertini's theorem [7]). *Let* $\mathcal{Z}$ *be a nonsingular quasi-projective subvariety of* $\mathcal{CP}^n$ *of dimension* k. *Let* $\mathcal{L}$ *be a linear system as in* (3.1). *Then there exists a nonzero homogeneous polynomial* G *in* r *variables such that* $G(c_1, \ldots, c_r) \neq 0$ *implies that the intersection of* $\mathcal{Z}$ *with the hypersurface*

$$\sum_{i=1}^{r} c_i m_i(z) = 0$$

is of dimension $k-1$ *and is nonsingular at each point not in the base locus of* $\mathcal{L}$.

Let $\langle f_1, \ldots, f_n \rangle$ denote the ideal of homogeneous polynomials generated by $f_1, \ldots, f_n$, let $\langle f_1, \ldots, f_n \rangle_M$ denote the subset of $\langle f_1, \ldots, f_n \rangle$ consisting of degree M polynomials, and let $V(f_1, \ldots, f_n)$ denote the variety defined by $f_1, \ldots, f_n$.

LEMMA 3.3 (see [20]). *Let* $f_1, \ldots, f_n$ *be homogeneous polynomials in the variables* $z_0, \ldots, z_n$ *such that the set*

$$S = \{(z_0, \ldots, z_n) \in V(f_1, \ldots, f_n) : z_0 \neq 0\}$$

consists of isolated points. Then there exists a positive integer K_0 *such that for* $K \geq K_0$,

$$\langle f_1, \ldots, f_n, z_0^{K_0} \rangle_M = \langle f_1, \ldots, f_n, z_0^{K} \rangle_M$$

for sufficiently large integers M.

The key lemma we need to detect paths diverging to infinity is the following.

LEMMA 3.4 (see [19]). *Let* $H : \mathcal{C}^n \times T \to \mathcal{C}^n$, *where* $T \subset \mathcal{R}$ *is an open subset,* H *is continuous in* t, *and* $H(x,t)$ *is a polynomial map for each* t, *and* $\widetilde{h}_i$ *is the homogenization of the* i*th polynomial of* $H(x,t) = 0$. *Suppose there are integers* $K < m$ *such that* $\dim_{\mathcal{C}} \langle \widetilde{h}_1, \ldots, \widetilde{h}_n, {z_0}^K \rangle_m$ *is independent of* $t \in U$ *and* $k \in \{K, K+1\}$. *Then no path of solutions of* $H(x,t) = 0$ *parameterized by* t *tends to infinity as* $t \to t_0 \in U$.

Proof of Theorem 3.1. (1) Rewrite $H(x)$ as

$$H(x) = \begin{pmatrix} \widehat{p}_1(x) + c_1 \\ \cdots \\ \widehat{p}_{n+m}(x) + c_{n+m} \\ \widehat{p}_{n+m+1}(x) + c_{n+m+1} \\ \cdots \\ \widehat{p}_{n+2m}(x) + c_{n+2m} \end{pmatrix},$$

where

$$\begin{aligned}
c_j &= tb_j + \gamma(1-t)a_j,\\
c_{n+m+i} &= \gamma(1-t)a^1_{i0}a^2_{i0} - t,\\
\widehat{p}_j(x) &= (\gamma(1-t)+t)p_j(x),\\
\widehat{p}_{n+m+i}(x) &= (\gamma(1-t)a^1_{i1}a^2_{i1}+t)x^2_{n+i} + (\gamma(1-t)a^1_{i2}a^2_{i2}+t)x^2_{n+m+i}\\
&\quad + \gamma(1-t)(a^1_{i1}a^2_{i2}+a^1_{i2}a^2_{i1})x_{n+i}x_{n+m+i}\\
&\quad + \gamma(1-t)((a^1_{i0}a^2_{i1}+a^1_{i1}a^2_{i0})x_{n+i} + (a^1_{i0}a^2_{i2}+a^1_{i2}a^2_{i0})x_{n+m+i})
\end{aligned}$$

for all $i = 1, \ldots, m,\ j = 1, \ldots, n+m$.

Let $d_i = \deg(p_i)$ and let $\widetilde{p}_i(z)$ be the homogenization of $\widehat{p}_i(x)$ $(i = 1, \ldots, n+2m)$. For any $t \in [0, 1)$ and $i = 1, \ldots, n+2m$, define

$$g_i(z) = \xi_i \widetilde{p}_i(z) + \eta_i z_0^{d_i}.$$

Let $\mathcal{Z}_0 = \mathcal{CP}^{n+2m} \setminus V(z_0) = \mathcal{C}^{n+2m}$ be the quasi-projective variety. The base locus of the linear system

$$\mathcal{L}_1 = \{\xi_1 \widetilde{p_1}(z) + \eta_1 z_0^{d_1}, \xi_1, \eta_1 \in \mathcal{C}\}$$

is contained in $V(z_0)$. Hence, by Bertini's theorem, there exists a nonzero homogeneous polynomial $G_1(\xi_1, \eta_1)$ such that $\mathcal{Z}_0 \cap V(g_1)$ is an $(n+2m-1)$-dimensional nonsingular quasi-projective variety for $G_1(\xi_1, \eta_1) \neq 0$. G_1 is homogeneous, so the polynomial $\widehat{G}_1(\eta_1) = G_1(1, \eta_1)$ is not identically zero. $V(\widehat{G}_1)$ consists of finite complex numbers. $\mathcal{Z}_0 \cap V(\widetilde{p}_1(z) + c_1 z_0^{d_1})$ is an $(n+2m-1)$-dimensional quasi-projective variety for all $1 \neq t \in R$ and $c_1 \in U_1 = \mathcal{C} \setminus V(\widehat{G}_1)$ which is an open dense full-measure subset of $\mathcal{C}$, and hence $\mathcal{Z}_1 = \mathcal{Z}_0 \cap V(\widehat{p}_1(x) + c_1)$ is also an $(n+2m-1)$-dimensional nonsingular quasi-variety.

The process is repeated. At every step, the base locus is contained in $V(z_0)$ and so does not intersect $\mathcal{Z}_i$. Applying Bertini's theorem, there exists an open dense full-measure subset U_i of $\mathcal{C}$, and for all $c_i \in V_i$, $\mathcal{Z}_i = \mathcal{Z}_{i-1} \cap V(\widehat{p}_i(x) + c_i)$ is an $(n+2m-i)$-dimensional nonsingular quasi-projective variety. When the process is repeated $n+2m$ times, the Jacobian of $H(x,t)$ is nonsingular and the solution set to $H(x,t) = 0$ consists of isolated complex numbers for all real $t \neq 1$. Thus the smoothness is proved.

(2) Denote by S_e the complex vector space of degree e homogeneous polynomials in $z_0, \ldots, z_{n+2m}$ with complex coefficients. Let $h_i(x) = \widehat{p}_i(x) + c_i$, and $\widetilde{h}_1(z), \ldots, \widetilde{h}_{n+2m}(z)$ is the homogenization of $h_1(x), \ldots, h_{n+2m}(x)$, $\widetilde{h}_i(z) \in S_{d_i}$. For $e \geq \max\{d_i\}$, define the linear transformation

$$T_e(\widetilde{h}_1, \ldots, \widetilde{h}_{n+2m}, z_0^K) : S_{e-d_1} \times \cdots \times S_{e-d_{n+2m}} \times S_{e-K} \to S_e,$$

$$T_e(\widetilde{h}_1, \cdots, \widetilde{h}_{n+2m}, z_0^K)(r_1, \ldots, r_{n+2m+1}) = \widetilde{h}_1 r_1 + \cdots + \widetilde{h}_{n+2m} r_{n+2m} + z_0^K r_{n+2m+1}.$$

Let $\mathfrak{T}_e$ be the matrix representation of T_e using the basis of all degree e monomials for each S_e. It follows that

$$\operatorname{rank}_{\mathcal{C}} \mathfrak{T}_e(\widetilde{h}_1, \ldots, \widetilde{h}_{n+2m}, z_0^K) = \dim_{\mathcal{C}} \langle \widetilde{h}_1, \ldots, \widetilde{h}_{n+2m}, z_0^K \rangle_e. \tag{3.2}$$

The following discussion shows that the homotopy map (2.1) satisfies accessibility if choosing $a_1, \ldots, a_{n+m}, a^1_{10}, \ldots, a^2_{m2}$ from an open dense full-measure subset U of $\mathcal{C}^{n+7m}$.

For a given e, let A be a highest-rank submatrix of $\mathfrak{T}_e(\widetilde{h}_1, \ldots, \widetilde{h}_{n+2m}, z_0^K)$ $(K < e)$. Then $\det(A)$ is a polynomial $D(a_1, \ldots, a_{n+m}, a_{10}^1, \ldots, a_{m2}^2, t)$ in variables $a_1, \ldots, a_{n+m}$, $a_{10}^1, \ldots, a_{m2}^2, t$ and not identically zero. $\det(A)$, if considered as a polynomial in the variables $a_1, \ldots, a_{n+m}, a_{10}^1, \ldots, a_{m2}^2$, has leading monomial

$$b(1-t)^s {a_1}^{\alpha_1} {a_{n+m}}^{\alpha_{n+m}} {a_{10}^1}^{\alpha_{n+m+1}} \cdots {a_{m2}^2}^{\alpha_{n+7m}}$$

for some nonnegative integer $\alpha_1, \ldots, \alpha_{n+7m}$ and $b \in \mathcal{C}^*$, so $\det(A)$ is not identically zero for each $t_0 \neq 1$ and vanishes on a complex codimension 1 subset $\mathcal{Z}_{t_0}$ of $\mathcal{C}^{n+7m}$. The union $\mathcal{Z} = \cup_{t_0 \in \mathcal{R} \setminus \{1\}} \mathcal{Z}_{t_0}$ of the $\mathcal{Z}_{t_0}$ is of real codimension 1 in $\mathcal{C}^{n+7m}$, and hence the complement $U^{'}$ of the closure is an open dense full-measure subset of $\mathcal{C}^{n+7m}$. For all $(a_1, \ldots, a_{n+m}, a_{10}^1, \ldots, a_{m2}^2) \in U^{'}$,

$$\operatorname{rank}_{\mathcal{C}} \mathfrak{T}_e(\widetilde{h}_1, \ldots, \widetilde{h}_{n+2m}, z_0^K)$$

is independent of $t \in \mathcal{R} \setminus \{1\}$. Since (3.2) holds,

$$\dim_{\mathcal{C}} \langle \widetilde{h}_1, \ldots, \widetilde{h}_{n+2m}, {z_0}^K \rangle_e \tag{3.3}$$

is independent of $t \in \mathcal{R} \setminus \{1\}$.

Replacing K by $K+1$ in the above, the same argument holds for $K+1$ on an open dense full-measure subset $U^{''}$ of $\mathcal{C}^{n+3m}$.

The solution to $Q(x) = 0$ are all nonsingular, thus by Lemma 3.3,

$$\langle \widetilde{h}_1, \ldots, \widetilde{h}_{n+2m}, {z_0}^K \rangle_e = \langle \widetilde{h}_1, \ldots, \widetilde{h}_{n+2m}, {z_0}^{K+1} \rangle_e$$

holds for $t = 0$. Hence

$$\dim_{\mathcal{C}} \langle \widetilde{h}_1, \ldots, \widetilde{h}_{n+2m}, {z_0}^{K+1} \rangle_e = \dim_{\mathcal{C}} \langle \widetilde{h}_1, \ldots, \widetilde{h}_{n+2m}, {z_0}^K \rangle_e$$

also holds for $t = 0$. Because (4.2) is independent of $t \in \mathcal{R} \setminus \{1\}$,

$$\dim_{\mathcal{C}} \langle \widetilde{h}_1, \ldots, \widetilde{h}_{n+2m}, {z_0}^{K+1} \rangle_e = \dim_{\mathcal{C}} \langle \widetilde{h}_1, \ldots, \widetilde{h}_{n+2m}, {z_0}^K \rangle_e$$

holds for all $t \in \mathcal{R} \setminus \{1\}$. Therefore, for all $(a_1, \ldots, a_{n+m}, s_{10}, \ldots, s_{2m}) \in U = U^{'} \cap U^{''}$,

$$\dim_{\mathcal{C}} \langle \widetilde{h}_1, \ldots, \widetilde{h}_{n+2m}, {z_0}^k \rangle_e$$

is independent of $t \in \mathcal{R} \setminus \{1\}$ and $k \in \{K, K+1\}$. By Lemma 3.4, all homotopy paths remain finite for $t \neq 1$. Thus each path must cross $t = 0$ in a finite solution to $Q(x) = 0$. The accessibility follows.

4. Numerical examples and applications. We have done some numerical tests and solved the practical problem (1.10) in Example 1.2 using our new hybrid

TABLE 4.1

(a) Size of Example 4.1.

n	m	BN	MBN	MV	SN
0	4	128	16	16	16

(b) Elapsed time for Example 4.1.

MHH	PH	Hybrid-MHH			Hybrid-PH		
		T1	T2	T	T1	T2	T
1.3	1.7	0.6	0.5	1.1	0.8	0.5	1.3

method, and we also compared our method with some other homotopy methods, such as multihomogeneous homotopy and polyhedral homotopy, mentioned in section 1.2. In this section, we will list some numerical results for several deficient polynomial systems coming from MTPS which appeared in the literature [8, 24, 27] and the practical problem (1.10) for $n = 3, 4, 5$. All our numerical results were obtained by PHCpack [34] running on a desktop with the Windows XP Professional operating system, Intel Core2 1.86GHz processor, and 960 megabytes of memory.

Problems are listed in Examples 4.1 to 4.6. Sizes of target systems, including numbers of variables, Bézout numbers, multi-Bézout numbers, mixed volumes, and solution numbers, are listed in Tables 4.1(a) to 4.6(a). Comparison of elapsed time of our hybrid method with multihomogeneous homotopy and polyhedral homotopy are listed in Tables 4.1(b) to 4.6(b). In Tables 4.1(a) to 4.6(a), "BN" and "MBN" denote the Bézout number and multihomogeneous Bézout number of the target system, "MV" is the mixed volume, and "SN" is the number of the isolated solutions of the target systems. In Tables 4.1(b) to 4.6(b) the numbers in the "MHH" column are the elapsed time of solving the target system by the multihomogeneous homotopies, and the numbers in the "PH" column are the elapsed time of solving the target system by the polyhedral homotopies. The numbers in the "Hybrid-MHH" column are the elapsed time of solving the target system by our hybrid method, in which we apply multihomogeneous homotopy to solve the simpler subsystems coming from the start system; the numbers in the "T_1" column are the elapsed time for solving the start system, the numbers in the "T_2" column are the elapsed time for tracing homotopy paths generated by (2.1) from $t = 0$ to $t = 1$, and the numbers in the "T" column are the total elapsed time. The numbers in the "Hybrid-PH" column are the elapsed time of solving the target system by our hybrid method, in which we apply polyhedral homotopy to solve the simpler subsystems coming from the start system. All times in the tables are in seconds.

Example 4.1. The PUMA system in [24] (see Table 4.1).

$$\begin{cases} -0.7623x_1 + 0.2238x_2 + 0.3461 = 0; \\ x_6x_8 + 0.3578x_1 + 0.004731x_2 = 0; \\ 0.2238x_1x_3 + 0.7623x_2x_3 + 0.2638x_1 - 0.07745x_2 - 0.6734x_4 - 0.6022 = 0; \\ 0.004731x_1x_3 - 0.3578x_2x_3 - 0.1238x_1 - 0.001637x_2 - 0.9338x_4 + x_7 - 0.3571 = 0; \\ x_1^2 + x_2^2 - 1 = 0; \\ x_3^2 + x_4^2 - 1 = 0; \\ x_5^2 + x_6^2 - 1 = 0; \\ x_7^2 + x_8^2 - 1 = 0. \end{cases}$$

TABLE 4.2

(a) Size of Example 4.2.

n	m	BN	MBN	MV	SN
0	6	4608	320	192	48

(b) Elapsed time for Example 4.2.

MHH	PH	Hybrid-MHH			Hybrid-PH		
		T1	T2	T	T1	T2	T
65.0	37.9	31.4	3.8	35.2	25.2	3.8	29.0

Example 4.2. d1 system in [8] (see Table 4.2).

$$\begin{cases} x_1^2+x_2^2-1=0;\\ x_3^2+x_4^2-1=0;\\ x_5^2+x_6^2-1=0;\\ x_7^2+x_8^2-1=0;\\ x_9^2+x_{10}^2-1=0;\\ x_{11}^2+x_{12}^2-1=0;\\ 3x_3+2x_5+x_7-3.9701=0;\\ 3x_1x_4+2x_1x_6+x_1x_8-1.7172=0;\\ 3x_2x_4+2x_2x_6+x_2x_8-4.0616=0;\\ x_3x_9+x_5x_9+x_7x_9-1.9791=0;\\ x_2x_4x_9+x_2x_6x_9+x_2x_8x_9+x_1x_{10}-1.9115=0;\\ -x_3x_{10}x_{11}-x_5x_{10}x_{11}-x_7x_{10}x_{11}+x_4x_{12}+x_6x_{12}+x_8x_{12}-0.4077=0. \end{cases}$$

Example 4.3. ipp system in [27] (see Table 4.3).

$$\begin{cases} x_1^2+x_2^2-1=0;\\ x_3^2+x_4^2-1=0;\\ x_5^2+x_6^2-1=0;\\ x_7^2+x_8^2-1=0;\\ -0.24915068x_1x_3+1.6091354x_1x_4+0.27942343x_2x_3+1.4348016x_2x_4\\ +0.40026384x_5x_8-0.80052768x_6x_7+0.074052388x_1-0.083050031x_2\\ -0.38615961x_3-0.75526603x_4+0.50420168x_5-1.0916287x_6\\ +0.40026384x_8+0.04920729=0;\\ 1.2501635x_1x_3-6.8660736x_1x_4-1.1922812x_2x_3-7.1994047x_2x_4\\ -4.3241927x_5x_7-8.6483855x_6x_8-0.3715727x_1+0.35436896x_2\\ +0.85383482x_3-0.39251967x_5-4.3241927x_7+0.1387301=0;\\ -6.3555007x_1x_3-1.1571992x_1x_4-6.6640448x_2x_3+1.1036211x_2x_4\\ +2.9070203x_5x_7+12.587767x_5x_8-6.2938836x_6x_7+5.8140406x_6x_8\\ +1.9594662x_1-12.280342x_2-0.79034221x_4+0.26387877x_5-0.5713143x_6\\ -11.628081x_7+12.587767x_8+21.62575=0;\\ 1.4894773x_1x_3+0.23062341x_1x_4+1.3281073x_2x_3-0.25864503x_2x_4\\ +1.165172x_5x_7-0.26908494x_5x_8+0.53816987x_6x_7+0.58258598x_6x_8\\ -0.20816985x_1+2.686832x_2-0.69910317x_3+0.35744413x_4+1.2499117x_5\\ +1.467736x_6+1.165172x_7+1.10763397x_8-0.69686809=0. \end{cases}$$

TABLE 4.3

(a) Size of Example 4.3.

n	m	BN	MBN	MV	SN
0	4	256	96	64	48

(b) Elapsed time for Example 4.3.

MHH	PH	Hybrid-MHH			Hybrid-PH		
		T1	T2	T	T1	T2	T
15.7	8.4	3.0	1.3	4.3	5.6	1.3	6.9

TABLE 4.4

(a) Size of Example 4.4.

n	m	BN	MBN	MV	SN
3	3	576	192	96	24

(b) Elapsed time for Example 4.4.

MHH	PH	Hybrid-MHH			Hybrid-PH		
		T1	T2	T	T1	T2	T
60.7	37.4	23.1	0.5	23.6	24.3	0.5	24.8

TABLE 4.5

(a) Size of Example 4.5.

n	m	BN	MBN	MV	SN
4	4	10368	2464	992	192

(b) Elapsed time for Example 4.5.

MHH	PH	Hybrid-MHH			Hybrid-PH		
		T1	T2	T	T1	T2	T
2001.9	264.5	833.0	12.6	845.6	179.2	12.6	191.8

TABLE 4.6

(a) Size of Example 4.6.

n	m	BN	MBN	MV	SN
5	5	331776	73728	14592	960

(b) Elapsed time for Example 4.6.

MHH	PH	Hybrid-MHH			Hybrid-PH		
		T1	T2	T	T1	T2	T
3.2e+5	7.5e+04	1.3e+5	195.5	1.3e+5	5772.4	195.5	6164.9

Example 4.4. The system in (1.10) for $n = 3$ (see Table 4.4).

Example 4.5. The system in (1.10) for $n = 4$ (see Table 4.5).

Example 4.6. The system in (1.10) for $n = 5$ (see Table 4.6).

Remark 4.1. Numerical results show that for MTPS, the hybrid method we proposed in this paper is more efficient, especially for high dimensional and highly sparse problems, such as Example 4.6.

Remark 4.2. As mentioned above, our start system can be decomposed into some subsystems whose dimension is reduced by m. Applying another homotopy method, e.g., polyhedral homotopy, to solve the target system, computing the mixed volume, and tracing the homotopy paths have to be done in $(n + 2m)$-dimensional space. However, applying our hybrid method, our main work is proceeded in $(n+m)$-dimensional space; thus when m is larger, our method is more efficient than other homotopy methods.

Remark 4.3. As mentioned in section 1.2, there is much freely variable software for solving polynomial systems. All of this software can be used to solve target systems and subsystems coming from our start system. We can choose one of them as a tool to compare our hybrid method with other homotopy methods. Maybe there is some difference in the computational time, but the comparison result of our hybrid method with other homotopy methods will not remarkably change due to the reasons mentioned in Remark 4.2.

Remark 4.4. In Examples 4.4 to 4.6, in practice, it is necessary to solve this problem for various parameters b_i $(i = 1, \ldots, 2n)$ and in real time. To achieve this aim, we can apply coefficient parameter homotopy, or cheater's homotopy, to solve the target system. We construct a start system whose constant terms are random nonzero complex numbers, and we solve it using our hybrid method. The solutions obtained by our hybrid method were used as start points of coefficient parameter homotopy or cheater's homotopy.

5. Conclusion. A hybrid polynomial system solving method for mixed trigonometric polynomial systems is proposed, which hybridizes the homotopy method and the elimination method and uses a hybrid homotopy that combines linear product homotopy and coefficient parameter homotopy or cheater's homotopy. It is more efficient for high dimensional and highly sparse problems.

Acknowledgment. The authors are grateful to the anonymous referee for his helpful suggestions.

REFERENCES

[1] D. N. Bernstein, *The number of roots of a system of equations*, Funct. Anal. Appl., 9 (1975), pp. 183–185.
[2] J. Canny and J. M. Rojas, *An optimal condition for determining the exact number of roots of a polynomial system*, in Proceedings of the 1991 International Symposium on Symbolic and Algebraic Computation, ACM, New York, 1991, pp. 96–101.
[3] D. Cox, J. Little, and D. O'Shea, *Ideals, Varieties, and Algorithms. An Introduction to Computational Algebraic Geometry and Commutative Algebra*, 2nd ed., Undergrad. Texts Math., Springer-Verlag, New York, 1997.
[4] D. Cox, J. Little, and D. O'Shea, *Using Algebraic Geometry*, Springer-Verlag, Berlin, 1998.
[5] F. J. Drexler, *Eine Methode zur Berechnung sämtlicher Lösungen von Polynomgleichungssystemen*, Numer. Math., 29 (1977/1978), pp. 45–58.
[6] C. B. Garcia and W. I. Zangwill, *Finding all solutions to polynomial systems and other systems of equations*, Math. Programming, 16 (1979), pp. 159–176.
[7] R. Hartshorne, *Algebraic Geometry*, Grad. Texts in Math. 52, Springer-Verlag, Berlin, New York, 1977.
[8] H. Hong and V. Stahl, *Safe starting regions by fixed points and tightening*, Computing, 53 (1995), pp. 322–335.
[9] B. Huber and B. Sturmfels, *A polyhedral method for solving sparse polynomial systems*, Math. Comp., 64 (1995), pp. 1541–1555.
[10] B. Huber and B. Sturmfels, *Bernstein's theorem in affine space*, Discrete Comput. Geom., 7 (1997), pp. 137–141.
[11] A. G. Khovanskii, *Newton polyhedra, and the genus of complete intersections*, Funct. Anal. Appl., 12 (1978), pp. 38–46.
[12] M. Kojima, http://www.is.titech.ac.jp/∼kojima/PHoM/index.html.
[13] A. G. Kushnirenko, *Newton polytopes and the Bézout theorem*, Funct. Anal. Appl., 10 (1976), pp. 233–235.
[14] T. Y. Li, http://www.mth.msu.edu/∼li/.
[15] T. Y. Li, *On Chow, Mallet-Paret and Yorke homotopy for solving systems of polynomials*, Bull. Inst. Math. Acad. Sinica, 11 (1983), pp. 433–437.
[16] T. Y. Li, *Solving polynomial systems by polyhedral homotopies*, Taiwanese J. Math., 3 (1999), pp. 251–279.

[17] T. Y. Li, *Numerical solution of polynomial systems by homotopy continuation methods*, in Handbook of Numerical Analysis, Vol. XI, North–Holland, Amsterdam, 2003, pp. 209–304.
[18] T. Y. Li, T. Sauer, and J. A. Yorke, *Numerical solution of a class of deficient polynomial systems*, SIAM J. Numer. Anal., 24 (1987), pp. 435–451.
[19] T. Y. Li, T. Sauer, and J. A. Yorke, *The random product homotopy and deficient polynomial systems*, Numer. Math., 51 (1987), pp. 481–500.
[20] T. Y. Li, T. Sauer, and J. A. Yorke, *The cheater's homotopy: An efficient procedure for solving systems of polynomial equations*, SIAM J. Numer. Anal., 26 (1989), pp. 1241–1251.
[21] T. Y. Li and X. Wang, *Solving deficient polynomial systems with homotopies which keep the subschemes at infinity invariant*, Math. Comp., 56 (1991), pp. 693–710.
[22] T. Y. Li and X. Wang, *Nonlinear homotopies for solving deficient polynomial systems with parameters*, SIAM J. Numer. Anal., 29 (1992), pp. 1104–1118.
[23] T. Y. Li and X. Wang, *The BKK root count in $\mathcal{C}^n$*, Math. Comp., 65 (1996), pp. 1477–1484.
[24] A. Morgan and V. Shapiro, *Box-bisection for solving second-degree systems and the problem of clustering*, ACM Trans. Math. Software, 13 (1987), pp. 152-167.
[25] A. Morgan and A. Sommese, *A homotopy for solving polynomial systems*, Appl. Math. Comput., 18 (1986), pp. 173–177.
[26] A. Morgan and A. Sommese, *A homotopy for solving general polynomial systems that respects m-homogeneous structures*, Appl. Math. Comput., 24 (1987), pp. 101–113.
[27] A. Morgan and A. Sommese, *Computing all solutions to polynomial systems using homotopy continuation*, Appl. Math. Comput., 24 (1987), pp. 115–138.
[28] A. Morgan and A. Sommese, *Coefficient-parameter polynomial continuation*, Appl. Math. Comput., 29 (1989), pp. 123–160; errata: Appl. Math. Comput., 51 (1992), p. 207.
[29] A. P. Morgan, A. J. Sommese, and C. W. Wampler, *A product-decomposition bound for Bézout numbers*, SIAM J. Numer. Anal., 32 (1995), pp. 1308–1325.
[30] A. Morgan, A. J. Sommese, and L. T. Watson, *Finding all isolated solutions to polynomial systems using HOMPACK*, ACM Trans. Math. Software, 12 (1989), pp. 93–122.
[31] A. Sommese, http://www.nd.edu/∼sommese/.
[32] A. Sommese and C. W. Wampler, *The Numerical Solution of Systems of Polynomials Arising in Engineering and Science*, World Scientific, Hackensack, NJ, 2005.
[33] P. Van Hentenryck, D. McAllester, and D. Kapur, *Solving polynomial systems using a branch and prune approach*, SIAM J. Numer. Anal., 34 (1997), pp. 797–827.
[34] J. Verschelde, *Algorithm* 795: *PHCpack: A general-purpose solver for polynomial systems by homotopy continuation*, ACM Trans. Math. Software, 25 (1999), pp. 251–276.
[35] J. Verschelde, M. Beckers, and A. Haegemans, *A new start system for solving deficient polynomial systems using continuation*, Appl. Math. Comput., 44 (1991), pp. 225–239.
[36] J. Verschelde and A. Haegemans, *The BQ-Algorithm for the Construction of an m-Homogeneous Start System*, Report TW 147, Department of Computer Science, K.U. Leuven, 1991.
[37] J. Verschelde and A. Haegemans, *The GBQ-algorithm for constructing start systems of homotopies for polynomial systems*, SIAM J. Numer. Anal., 30 (1993), pp. 583–594.
[38] J. Verschelde and R. Cools, *Symmetric homotopy construction*, J. Comput. Appl. Math., 50 (1994), pp. 572–592.
[39] J. Verschelde and R. Cools, *Nonlinear reduction for solving deficient polynomial systems by continuation methods*, Numer. Math., 63 (1992), pp. 263–282.
[40] J. Verschelde, P. Verlinden, and R. Cools, *Homotopies exploiting Newton polytopes for solving sparse polynomial systems*, SIAM J. Numer. Anal., 31 (1994), pp. 915–930.
[41] C. W. Wampler, A. P. Morgan, and A. J. Sommese, *Numerical continuation methods for solving polynomial systems arising in kinematics*, ASME J. Mechanical Design, 112 (1990), pp. 59–68.
[42] L. T. Watson, S. C. Billups, and A. P. Morgan, *ALGORITHM* 652: *HOMPACK: A suite of codes for globally convergent homotopy algorithms*, ACM Trans. Math. Software, 13 (1987), pp. 281–310.
[43] S. M. Wise, A. J. Sommese, and L. T. Watson, *Algorithm* 801: *POLSYS_PLP: A partitioned linear product homotopy code for solving polynomial systems of equations*, ACM Trans. Math. Software, 26 (2000), pp. 176–200.
[44] A. H. Wright, *Finding all solutions to a system of polynomial equations*, Math. Comp., 44 (1985), pp. 125–133.
[45] W. T. Wu, *On the decision problem and the mechanization of theorem-proving in elementary geometry*, in Automated Theorem Proving: After 25 Years, W. Bledsoe and D. Loveland, eds., Contemp. Math. 29, AMS, Providence, RI, pp. 213–234.

[46] B. Yu, *Homotopy Methods for Solving Polynomial Systems and Multiparameter Eigenvalue Problems*, Ph.D. Dissertation, Jilin University, 1992.

[47] B. Yu and G. C. Feng, *The random product homotopy for solving polynomial systems in* $C^{k_1} \times C^{k_2} \times \cdots \times C^{k_m}$, in Computer Mathematics (Tianjin, 1991), World Scientific, River Edge, NJ, 1993, pp. 36–45.

[48] W. Zulehner, *A simple homotopy method for determining all isolated solutions to polynomial systems*, Math. Comp., 50 (1988), pp. 167–177.

SIAM J. NUMER. ANAL.
Vol. 46, No. 3, pp. 1519–1553
© 2008 Society for Industrial and Applied Mathematics

WALSH SPACES CONTAINING SMOOTH FUNCTIONS AND QUASI–MONTE CARLO RULES OF ARBITRARY HIGH ORDER*

JOSEF DICK†

Abstract. We define a Walsh space which contains all functions whose partial mixed derivatives up to order $\delta \geq 1$ exist and have finite variation. In particular, for a suitable choice of parameters, this implies that certain Sobolev spaces are contained in these Walsh spaces. For this Walsh space we then show that quasi–Monte Carlo rules based on digital (t, α, s)-sequences achieve the optimal rate of convergence of the worst-case error for numerical integration. This rate of convergence is also optimal for the subspace of smooth functions. Explicit constructions of digital (t, α, s)-sequences are given, hence providing explicit quasi–Monte Carlo rules which achieve the optimal rate of convergence of the integration error for arbitrarily smooth functions.

Key words. numerical integration, quasi–Monte Carlo, digital nets and sequences, Walsh functions

AMS subject classifications. Primary, 11K38, 11K45, 65C05; Secondary, 42C10

DOI. 10.1137/060666639

1. Introduction.

Quasi–Monte Carlo rules are quadrature rules which aim to approximate an integral $\int_{[0,1]^s} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$ by the average of the N function values $f(\boldsymbol{x}_n)$ at the quadrature points $\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{N-1} \in [0,1]^s$ (and hence are equal weight quadrature rules). The dimension s can be arbitrarily large. The task here is to find ways of choosing those quadrature points in order to obtain a fast convergence of the approximation to the integral. Until now, explicit constructions of quadrature points in arbitrary high dimensions have been available for the following two cases:

1. for sufficiently smooth periodic functions, arbitrary high convergence can be achieved using Kronecker sequences [18, Theorem 5.3] or a modification of digital nets recently introduced in [2];
2. a convergence of $\mathcal{O}(N^{-1}(\log N)^{s-1})$ can be achieved for functions of bounded variation (in this case the functions are not required to be periodic).

For nonperiodic functions no explicit constructions have been established which can fully exploit the smoothness of the integrand. This paper provides a complete solution to this problem.

Among other things we show that an explicit construction of suitable point sets and sequences can be obtained in the following way: let $d \geq 1$ be an integer and let $\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots \in [0,1)^{ds}$ be the points of a digital (t, m, ds)-net or digital (t, ds)-sequence over a finite field $\mathbb{F}_q$ in dimension ds (see [20] for the definition of digital nets and sequences and see, for example, [9, 20, 21, 23, 31] for explicit constructions of suitable digital nets and sequences). Let $\boldsymbol{x}_n = (x_{n,1}, \ldots, x_{n,ds})$ with $x_{n,j} = x_{n,j,1}q^{-1} + x_{n,j,2}q^{-2} + \cdots$ and $x_{n,j,i} \in \{0, \ldots, q-1\}$ (i.e., $x_{n,j,i}$ are the digits in the base q representation of $x_{n,j}$). Then for $n \geq 0$ we define $\boldsymbol{y}_n = (y_{n,1}, \ldots, y_{n,s})$ with

$$y_{n,j} = \sum_{i=1}^{\infty} \sum_{k=1}^{d} x_{n,(j-1)d+k,i} q^{-k-(i-1)d} \quad \text{for } j = 1, \ldots, s.$$

*Received by the editors August 2, 2006; accepted for publication (in revised form) December 19, 2007; published electronically April 4, 2008.
http://www.siam.org/journals/sinum/46-3/66663.html
†School of Mathematics and Statistics, University of New South Wales, Sydney 2052, Australia (josef.dick@unsw.edu.au).

(Note that the addition here is carried out in $\mathbb{R}$ and that the sum over i above is often finite as $x_{n,j,i} = 0$ for i large enough.) We point out here that the quality of the point set or sequence is directly related to the t-value of the underlying (t, m, ds)-net or (t, ds)-sequence; see Theorems 4.11 and 4.12.

Corollary 5.5 now shows that quasi–Monte Carlo rules using the points $\boldsymbol{y}_0, \dots, \boldsymbol{y}_{N-1}$ (with $N = q^m$ for some $m \geq 1$) achieve the optimal rate of convergence of the integration error of $\mathcal{O}(N^{-\vartheta}(\log N)^{\vartheta s})$ for functions which have partial mixed derivatives up to order ϑ which are square integrable as long as $1 \leq \vartheta \leq d$ (Corollary 5.5 is actually more general). If $\vartheta > d$, no improvement of the convergence rate is obtained compared to functions with smoothness $\vartheta = d$; i.e., we obtain a convergence of $\mathcal{O}(N^{-d}(\log N)^{ds})$. Similar but less general results for periodic functions compared to those in this paper have been shown in [2] by a different proof method. (The construction above is an example of a construction method which can be used. In section 4 we outline the general algebraical properties required for the construction of suitable point sets.)

The quasi–Monte Carlo algorithm based on digital nets and sequences proposed here has also some further useful properties. For example, our results also hold if one randomizes the point set by, say, a random digital shift (see, for example, [4, 5, 17]). (This follows easily because the worst-case error (see section 5) is invariant with respect to digital shifts in the Walsh space, and hence we obtain the same upper bounds for randomized digital nets and sequences.) In summary the quadrature rules have the following properties:

- The quadrature rules introduced in this paper are equal weight quadrature rules which achieve the optimal rate of convergence up to some $\log N$ factors, and the result holds for deterministic and randomly digitally shifted quadrature rules.
- The construction of the underlying point set is explicit, and suitable point sets are available in arbitrary high dimensions and arbitrary high number of points.
- The quadrature rules automatically adjust themselves to the optimal rate of convergence $\mathcal{O}(N^{-\vartheta}(\log N)^{s\vartheta})$ as long as $1 \leq \vartheta \leq d$.
- The underlying point set is extensible in the dimension as well as in the number of points; i.e., one can always add some coordinates or points to an existing point set such that the quality of the point set is preserved.

In the following we lay out some of the underlying principles used in this work which stem from the behavior of the Walsh coefficients of smooth functions. Walsh functions are piecewise constant wavelets which form an orthonormal set of $\mathcal{L}_2([0,1]^s)$. In their simplest form, for a nonnegative integer k with base 2 representation $k = \kappa_0 + \cdots + \kappa_{m-1}2^{m-1}$ and an $x \in [0,1)$ with base 2 representation $x = x_1 2^{-1} + x_2 2^{-2} + \cdots$, the kth Walsh function in base 2 is given by

$$\mathrm{wal}_k(x) = (-1)^{\kappa_0 x_1 + \cdots + \kappa_{m-1} x_m}.$$

(Later on we will use the more general definition of Walsh functions over groups.)

The behavior of the Fourier coefficients of smooth periodic functions is well known; i.e., the smoother the function, the faster the Fourier coefficients go to zero (see, for example, [34]). An analogous result for Walsh functions has, to the best of the author's knowledge, not been known until now (see Fine [10], who, for example, shows that the only absolute continuous functions whose kth Walsh coefficients decay faster than $1/k$ are constant functions). This will be established here and subsequently be exploited to obtain quasi–Monte Carlo rules with arbitrary high order of convergence.

To give a glimpse of how the Walsh coefficients of smooth functions behave, consider, for example, the Walsh series for $1/2 - x$:

$$1/2 - x = \sum_{k=0}^{\infty} c_k \mathrm{wal}_k(x) = \sum_{a=0}^{\infty} 2^{-a-2} \mathrm{wal}_{2^a}(x).$$

Although the function is infinitely smooth, in general the decay of the Walsh coefficient is only of order $1/k$. But note that most of the Walsh coefficients are actually 0. For example, when we consider $(1/2 - x)^2$, then typically we would have that the Walsh coefficient of $k = 2^a$ is of order 2^{-a}, the Walsh coefficient of $k = 2^{a_1} + 2^{a_2}$ $(a_1 > a_2)$ is of order $2^{-a_1-a_2}$, and for $k = 2^{a_1} + 2^{a_2} + \cdots + 2^{a_v}$ with $a_1 > \cdots > a_v$ and $v > 2$ the kth Walsh coefficient would be 0. By considering $(1/2 - x)^3, (1/2 - x)^4, \ldots$, or more generally polynomials, one can now realize that the speed of convergence of the Walsh coefficients depends on how many nonzero digits k has. This is the basic feature which we will relate to the speed of convergence of the Walsh coefficients for smooth functions.

Subsequently we will explicitly state and use the behavior of the Walsh coefficients of smooth functions. In general the Walsh functions depend on the base q digit expansion of the wavenumber k and also of the point x where the Walsh function is to be evaluated. Hence, maybe not surprisingly, the value of the kth Walsh coefficient of a smooth function depends also on the q-adic expansion of k. We show that the Walsh space $\mathcal{E}_{s,q,\vartheta,\boldsymbol{\gamma}}$ introduced in section 3 contains all functions whose partial mixed derivatives up to order $\delta < \vartheta$ exist and have finite variation, where ϑ is a parameter restricting the behavior of the Walsh coefficients of the function space $\mathcal{E}_{s,q,\vartheta,\boldsymbol{\gamma}}$. (We use a similar, though much more general, technique to that which Fine [10] used for showing that the Walsh coefficients of a differentiable function cannot decay faster than $1/k$.)

The concept of digital (t, α, β, m, s)-nets and digital (t, α, β, s)-sequences (see section 4 and also [2] for a similar concept) is now designed to yield point sets which work well for the Walsh space $\mathcal{E}_{s,q,\vartheta,\boldsymbol{\gamma}}$, just in the same way as the digital nets and sequences from [9, 19, 20, 23, 31] are designed to work well for the spaces considered, for example, in [4, 11] (or as lattice rules are designed to work well for periodic Korobov spaces). Here the power of the result that the Walsh space $\mathcal{E}_{s,q,\vartheta,\boldsymbol{\gamma}}$ contains smooth functions comes into play: it follows that we can fully exploit the smoothness of an integrand using digital $(t, \alpha, \alpha, m, s)$-nets or digital (t, α, α, s)-sequences. As the construction of the points $\boldsymbol{y}_0, \boldsymbol{y}_1, \ldots$ introduced at the beginning yields explicit examples of digital $(t, \alpha, \alpha, m, s)$-nets or digital (t, α, α, s)-sequences as shown in section 4 we therefore obtain explicit constructions of quasi–Monte Carlo rules which can achieve the optimal order of convergence for arbitrary smooth functions.

In the next section we introduce Walsh functions over groups and state some of their essential properties.

2. Walsh functions over groups. In this section we give the definition of Walsh functions over groups and present some essential properties. Walsh functions in base 2 were first introduced by Walsh [33], though a similar but noncomplete set of functions had already been studied by Rademacher [27]. Further important results were obtained in [10]. We follow [26] in our presentation.

2.1. Definition of Walsh functions over groups. An essential tool for the investigation of digital nets are Walsh functions. A very general definition, corresponding to the most general construction of digital nets over finite rings, was given

in [15]. There, Walsh functions over a finite abelian group G, using some bijection φ, were defined. Here we restrict ourselves to the additive groups of the finite fields $\mathbb{F}_{p^r}$, p prime, and $r \geq 1$. We restate the definitions for this special case here for the sake of convenience. In the following let $\mathbb{N}$ denote the set of positive integers and $\mathbb{N}_0$ the set of nonnegative integers.

DEFINITION 2.1 (Walsh functions). *Let $q = p^r$, p prime, $r \in \mathbb{N}$, and let $\mathbb{F}_q$ be the finite field with q elements. Let $\mathbb{Z}_q = \{0, 1, \ldots, q-1\} \subset \mathbb{Z}$ and let $\varphi : \mathbb{Z}_q \longrightarrow \mathbb{F}_q$ be a bijection such that $\varphi(0) = 0$, the neutral element of addition in $\mathbb{F}_q$. Moreover denote by ψ the canonical isomorphism (described below) of additive groups $\psi : \mathbb{F}_q \longrightarrow \mathbb{Z}_p^r$ and define $\eta := \psi \circ \varphi$. For $1 \leq i \leq r$ denote by π_i the projection $\pi_i : \mathbb{Z}_p^r \longrightarrow \mathbb{Z}_p$, $\pi_i(x_1, \ldots, x_r) = x_i$.*

$$\begin{array}{ccc} \mathbb{Z}_q & \xrightarrow{\varphi} & \mathbb{F}_q \\ & \searrow^{\eta} & \downarrow{\scriptstyle \psi} \\ & & \mathbb{Z}_p^r \xrightarrow{\pi_i} \mathbb{Z}_p \end{array}$$

Now let $k \in \mathbb{N}_0$ with base q representation $k = \kappa_0 + \kappa_1 q + \cdots + \kappa_{m-1} q^{m-1}$ where $\kappa_l \in \mathbb{Z}_q$ and let $x \in [0,1)$ with base q representation $x = x_1/q + x_2/q^2 + \cdots$ (unique in the sense that infinitely many x_l must be different from $q-1$). Then the kth Walsh function over the additive group of the finite field $\mathbb{F}_q$ with respect to the bijection φ is defined by

$$_{\mathbb{F}_q,\varphi}\mathrm{wal}_k(x) = \exp\left(\frac{2\pi \mathtt{i}}{p} \sum_{l=0}^{m-1} \sum_{i=1}^{r} (\pi_i \circ \eta)(\kappa_l)(\pi_i \circ \eta)(x_{l+1})\right).$$

For convenience we will in the rest of the paper omit the subscript and simply write wal_k *if there is no ambiguity.*

Multivariate Walsh functions are defined by multiplication of the univariate components, i.e., for $s > 1$, $\boldsymbol{x} = (x_1, \ldots, x_s) \in [0,1)^s$ and $\boldsymbol{k} = (k_1, \ldots, k_s) \in \mathbb{N}_0^s$, we set

$$\mathrm{wal}_{\boldsymbol{k}}(\boldsymbol{x}) = \prod_{j=1}^{s} \mathrm{wal}_{k_j}(x_j).$$

We now briefly describe the canonical isomorphism. Let $\mathbb{F}_q = \mathbb{Z}_p[\theta]$, such that $\{1, \theta, \ldots, \theta^{r-1}\}$ is a basis of $\mathbb{F}_q$ over $\mathbb{Z}_p$ as a vector space. Then the isomorphism ψ between $\mathbb{F}_q$ and $\mathbb{Z}_p^r$ shall be given by

$$\psi(x) = (x_1, \ldots, x_r)^\top \text{ for } x = \sum_{i=1}^{r} x_i \theta^{i-1},\ x_i \in \mathbb{Z}_p.$$

For more information on the Walsh functions defined above see [26].

We summarize some important properties of Walsh functions over the additive group of a finite field which will be used throughout the paper. The proofs of the subsequent results can be found, e.g., in [16, 25] (see also [1]). In the following we call $x \in [0,1)$ a q-adic rational if x can be represented by a finite base q expansion.

PROPOSITION 2.2. *Let p, q, $\mathbb{F}_q$, and φ be as in Definition 2.1. For x, y with q-adic representations $x = \sum_{i=w}^{\infty} x_i q^{-i}$ and $y = \sum_{i=w}^{\infty} y_i q^{-i}$, $w \in \mathbb{Z}$ (taking w negative, hence the following operations are also defined for integers), define $x \oplus_\varphi y :=$*

$\sum_{i=w}^{\infty} z_i q^{-i}$, *where* $z_i := \varphi^{-1}(\varphi(x_i) + \varphi(y_i))$, *and* $\ominus_\varphi x := \sum_{i=w}^{\infty} v_i q^{-i}$, *where* $v_i := \varphi^{-1}(-\varphi(x_i))$. *Further we set* $x \ominus_\varphi y := x \oplus_\varphi (\ominus_\varphi y)$. *For vectors* $\boldsymbol{x}, \boldsymbol{y}$ *we define the operations componentwise. Then we have the following:*

1. *For all* $k, l \in \mathbb{N}_0$ *and all* $x, y \in [0,1)$, *with the restriction that if* x, y *are not* q*-adic rationals, then* $x \oplus_\varphi y$ *is not allowed to be a* q*-adic rational, we have*

$$\mathrm{wal}_k(x) \cdot \mathrm{wal}_l(x) = \mathrm{wal}_{k \oplus_\varphi l}(x), \quad \mathrm{wal}_k(x) \cdot \mathrm{wal}_k(y) = \mathrm{wal}_k(x \oplus_\varphi y)$$

and, with the restriction that if x, y *are not* q*-adic rationals, then* $x \ominus_\varphi y$ *is not allowed to be a* q*-adic rational,*

$$\mathrm{wal}_k(x) \cdot \overline{\mathrm{wal}_l(x)} = \mathrm{wal}_{k \ominus_\varphi l}(x), \quad \mathrm{wal}_k(x) \cdot \overline{\mathrm{wal}_k(y)} = \mathrm{wal}_k(x \ominus_\varphi y).$$

2. *We have*

$$\sum_{k=0}^{q-1} \mathrm{wal}_l(k/q) = \begin{cases} 0 & \text{if } l \neq 0, \\ q & \text{if } l = 0. \end{cases}$$

3. *We have*

$$\int_0^1 \mathrm{wal}_0(x)\,\mathrm{d}x = 1 \quad \text{and} \quad \int_0^1 \mathrm{wal}_k(x)\,\mathrm{d}x = 0 \text{ if } k > 0.$$

4. *For all* $\boldsymbol{k}, \boldsymbol{l} \in \mathbb{N}_0^s$ *we have the following orthogonality properties:*

$$\int_{[0,1)^s} \mathrm{wal}_{\boldsymbol{k}}(\boldsymbol{x})\overline{\mathrm{wal}_{\boldsymbol{l}}(\boldsymbol{x})}\,\mathrm{d}\boldsymbol{x} = \begin{cases} 1 & \text{if } \boldsymbol{k} = \boldsymbol{l}, \\ 0 & \text{otherwise.} \end{cases}$$

5. *For any* $f \in \mathcal{L}_2([0,1)^s)$ *and any* $\boldsymbol{\sigma} \in [0,1)^s$ *we have*

$$\int_{[0,1)^s} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \int_{[0,1)^s} f(\boldsymbol{x} \oplus_\varphi \boldsymbol{\sigma})\,\mathrm{d}\boldsymbol{x}.$$

6. *For any integer* $s \geq 1$ *the system* $\{\mathrm{wal}_{\boldsymbol{k}} : \boldsymbol{k} \in \mathbb{N}_0^s\}$ *is a complete orthonormal system in* $\mathcal{L}_2([0,1)^s)$.

Remark 2.3. The restrictions in item 1 were added to exclude cases like $x = (0.010101\ldots)_2$, $y = (0.0010101\ldots)_2$, and $x \oplus y = (0.1)_2$, for which the result is of course not true. On the other hand, the result holds for $x \oplus y = (0.0111111\ldots)_2$.

Throughout the paper we will use a fixed bijection φ, and a fixed finite field $\mathbb{F}_q$ is used for Walsh functions and $\oplus_\varphi$ and $\ominus_\varphi$. Hence we will often write $\oplus$ and $\ominus$ instead of $\oplus_\varphi$ and $\ominus_\varphi$.

In the following section we will deal with Walsh series and Walsh coefficients, which we briefly describe in the following: functions $f \in \mathcal{L}_2([0,1)^s)$ have an associated Walsh series

$$f(\boldsymbol{x}) \sim \sum_{\boldsymbol{k} \in \mathbb{N}_0^s} \hat{f}(\boldsymbol{k})\mathrm{wal}_{\boldsymbol{k}}(\boldsymbol{x}),$$

where the Walsh coefficients $\hat{f}(\boldsymbol{k})$ are given by

$$\hat{f}(\boldsymbol{k}) = \int_{[0,1)^s} f(\boldsymbol{x})\mathrm{wal}_{\boldsymbol{k}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}.$$

For smooth functions the Walsh series converges to the function, which is shown in section 3.3.

3. Walsh spaces containing smooth functions. In the following we investigate how the Walsh coefficients of smooth functions decay, and subsequently we use this to define function classes based on Walsh functions which contain smooth functions. But first we introduce a suitable variation.

3.1. A generalized weighted Hardy and Krause variation. In the following we generalize the Hardy and Krause variation, which suits our purposes later on.

3.1.1. Hölder condition. A function $f : [0,1) \to \mathbb{R}$ satisfies a Hölder condition with coefficient $0 < \lambda \le 1$ if there is a constant $C_f > 0$ such that

$$|f(x) - f(y)| \le C_f |x-y|^\lambda \quad \text{for all } x, y \in [0,1).$$

The right-hand side of the above inequality forms a metric on $[0,1)$. When one considers the higher-dimensional domain $[0,1)^s$ then $|x-y|$ is changed to some other metric on $[0,1)^s$. Here we consider tensor product spaces and generalize the Hölder condition to higher dimensions in a way which is suitable for tensor product spaces in our context. Consider, for example, the function $f(\boldsymbol{x}) = \prod_{j=1}^s f_j(x_j)$, where $\boldsymbol{x} = (x_1, \dots, x_s)$ and each $f_j : [0,1) \to \mathbb{R}$ satisfies a Hölder condition with coefficient $0 < \lambda \le 1$. Then it follows that for all $\emptyset \ne u \subseteq \mathcal{S} := \{1, \dots, s\}$ we have

$$\prod_{j \in u} |f_j(x_j) - f_j(y_j)| \le \prod_{j \in u} C_{f_j} \prod_{j \in u} |x_j - y_j|^\lambda \tag{3.1}$$

for all $x_j, y_j \in [0,1)$ with $j \in u$. But here $\prod_{j=1}^s |x_j - y_j|$ is not a metric on $[0,1)^s$.

Note that we have

$$\prod_{j \in u} |f_j(x_j) - f_j(y_j)| = \left| \sum_{v \subseteq u} (-1)^{|v|-|u|} \prod_{j \in v} f_j(x_j) \prod_{j \in u \setminus v} f_j(y_j) \right|, \tag{3.2}$$

which can be described in words in the following way: for given $\emptyset \ne u \subseteq \mathcal{S}$ let $x_j, y_j \in [0,1)$ with $x_j \ne y_j$ for all $j \in u$; consider the box J with vertices $\{(a_j)_{j \in u} : a_j = x_j \text{ or } a_j = y_j \text{ for } j \in u\}$. Then (3.2) is the alternating sum of the function $\prod_{j \in u} f_j$ at the vertices of J where adjacent vertices have opposite signs. This sum can also be defined for functions on $[0,1)^s$ which are not of product form.

Indeed, for a subinterval $J = \prod_{j=1}^s [x_j, y_j)$ with $0 \le x_j < y_j \le 1$ and a function $f : [0,1)^s \to \mathbb{R}$, let the function $\Delta(f, J)$ denote the alternating sum of f at the vertices of J where adjacent vertices have opposite signs. (Hence for $f = \prod_{j=1}^s f_j$ we have $\Delta(f, J) = \prod_{j=1}^s (f_j(x_j) - f_j(y_j))$.)

3.1.2. Generalized Vitali variation. Let $\mathfrak{p} \ge 1$. Then we define the generalized variation in the sense of Vitali with coefficient $0 < \lambda \le 1$ by

$$V^{(s)}_{\lambda,\mathfrak{p}}(f) = \sup_{\mathcal{P}} \left(\sum_{J \in \mathcal{P}} \mathrm{Vol}(J) \left| \frac{\Delta(f, J)}{\mathrm{Vol}(J)^\lambda} \right|^{\mathfrak{p}} \right)^{1/\mathfrak{p}}, \tag{3.3}$$

where the supremum is extended over all partitions $\mathcal{P}$ of $[0,1]^s$ into subintervals and $\mathrm{Vol}(J)$ denotes the volume of the subinterval J.

Note that for $\lambda = 1$ and $\mathfrak{p} = 1$ one obtains the usual definition of the Vitali variation; see, for example, [20]. If we take $\mathfrak{p} = \infty$, then we obtain a condition of the

form (3.1) where $u = \mathcal{S}$ and where we can take the constant $\prod_{j=1}^{s} C_{f_j} = V^{(s)}_{\lambda,\infty}(f)$. For $s = 1$ and $\mathfrak{p} = \infty$ we obtain a Hölder condition with coefficient $0 < \lambda \le 1$. In this sense we can view (3.3) as a fractional Vitali variation of order λ.

For $\lambda = 1$ and if the partial derivatives of f are continuous on $[0,1]^s$ we also have the formula

$$V^{(s)}_{1,\mathfrak{p}}(f) = \left(\int_{[0,1]^s} \left| \frac{\partial^s f}{\partial x_1 \cdots \partial x_s} \right|^{\mathfrak{p}} \, \mathrm{d}\boldsymbol{x} \right)^{1/\mathfrak{p}} \tag{3.4}$$

for all $\mathfrak{p} \ge 1$. Indeed we have

$$|\Delta(f,J)| = \left| \int_J \frac{\partial^s f}{\partial x_1 \cdots \partial x_s}(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \right| = \mathrm{Vol}(J) \left| \frac{\partial^s f}{\partial x_1 \cdots \partial x_s}(\boldsymbol{\zeta}_J) \right|$$

for some $\boldsymbol{\zeta}_J \in \overline{J}$, which follows by applying the mean value theorem to the inequality

$$\min_{\boldsymbol{x} \in \overline{J}} \left| \frac{\partial^s f}{\partial x_1 \cdots \partial x_s}(\boldsymbol{x}) \right| \le \mathrm{Vol}(J)^{-1} \left| \int_J \frac{\partial^s f}{\partial x_1 \cdots \partial x_s}(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \right| \le \max_{\boldsymbol{x} \in \overline{J}} \left| \frac{\partial^s f}{\partial x_1 \cdots \partial x_s}(\boldsymbol{x}) \right|.$$

Therefore we have

$$\sum_{J \in \mathcal{P}} \mathrm{Vol}(J) \left| \frac{\Delta(f,J)}{\mathrm{Vol}(J)} \right|^{\mathfrak{p}} = \sum_{J \in \mathcal{P}} \mathrm{Vol}(J) \left| \frac{\partial^s f}{\partial x_1 \cdots \partial x_s}(\boldsymbol{\zeta}_J) \right|^{\mathfrak{p}},$$

which is just a Riemann sum for the integral $\int_{[0,1]^s} \left| \frac{\partial^s f}{\partial x_1 \cdots \partial x_s} \right|^{\mathfrak{p}} \mathrm{d}\boldsymbol{x}$, and thus the equality follows.

Using Hölder's inequality and the fact that $\left(\sum_{J \in \mathcal{P}} (\mathrm{Vol}(J)^{1-1/\mathfrak{p}})^{\mathfrak{p}/(\mathfrak{p}-1)} \right)^{1-1/\mathfrak{p}} = \left(\sum_{J \in \mathcal{P}} \mathrm{Vol}(J) \right)^{1-1/\mathfrak{p}} = 1$, it follows that

$$V^{(s)}_{\lambda,1}(f) \le V^{(s)}_{\lambda,\mathfrak{p}}(f) \quad \text{for all } \mathfrak{p} \ge 1.$$

3.1.3. Generalized Hardy and Krause variation. Until now we did not take projections to lower-dimensional faces into account (in (3.1) we did take projections into account as we considered all $\emptyset \ne u \subseteq \mathcal{S}$).

For $\emptyset \ne u \subseteq \mathcal{S}$, let $V^{(|u|)}_{\lambda,\mathfrak{p}}(f_u; u)$ be the generalized Vitali variation with coefficient $0 < \lambda \le 1$ of the $|u|$-dimensional function $f_u(\boldsymbol{x}_u) = \int_{[0,1)^{s-|u|}} f(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}_{\mathcal{S} \setminus u}$. For $u = \emptyset$ we have $f_\emptyset = \int_{[0,1)^s} f(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}_{\mathcal{S}}$ and we define $V^{(|\emptyset|)}_{\lambda,\mathfrak{p}}(f_\emptyset; \emptyset) = |f_\emptyset|$. Let $\mathfrak{q} \ge 1$. Then

$$V_{\lambda,\mathfrak{p},\mathfrak{q}}(f) = \left(\sum_{u \subseteq \mathcal{S}} \left(V^{(|u|)}_{\lambda,\mathfrak{p}}(f_u; u) \right)^{\mathfrak{q}} \right)^{1/\mathfrak{q}} \tag{3.5}$$

is called the generalized Hardy and Krause variation of f on $[0,1]^s$.

For $\lambda = \mathfrak{p} = \mathfrak{q} = 1$ one obtains an unanchored version of the usual definition of the Hardy and Krause variation; see [20]. A function f for which $V_{\lambda,\mathfrak{p},\mathfrak{q}}(f) < \infty$ is said to be of finite variation with coefficient λ. (We remark that in some cases it might be appropriate to leave out the term corresponding to $u = \emptyset$ in (3.5), but here this term will be needed later on, and hence we include it already in the definition of the variation.)

3.1.4. Generalized weighted Hardy and Krause variation. As first suggested in [30] (see also [6]) different coordinates might have different importance, hence we can also define a weighted variation. In the spirit of the weighted Sobolev spaces in [6], let $\boldsymbol{\gamma} = (\gamma_u)_{u \subset \mathbb{N}}$ be an indexed set of nonnegative real numbers. Then we define the weighted variation $V_{\lambda,\mathfrak{p},\mathfrak{q},\boldsymbol{\gamma}}(f)$ of f with coefficient $0 < \lambda \le 1$ by

$$V_{\lambda,\mathfrak{p},\mathfrak{q},\boldsymbol{\gamma}}(f) = \left(\sum_{u \subseteq \mathcal{S}} \gamma_u^{-1} \left(V_{\lambda,\mathfrak{p}}^{(|u|)}(f_u; u) \right)^{\mathfrak{q}} \right)^{1/\mathfrak{q}}.$$

Note that for $\lambda = 1$ and $\mathfrak{p} = \mathfrak{q} = 2$ the weighted variation $V_{\lambda,\mathfrak{p},\mathfrak{q},\boldsymbol{\gamma}}(f)$ coincides with the norm in a weighted unanchored Sobolev space for any function in this Sobolev space; i.e., we have the identity $V_{1,2,2,\boldsymbol{\gamma}}(f) = \|f\|_{\mathrm{sob}}$, where

$$\|f\|_{\mathrm{sob}} = \left(\sum_{u \subseteq \mathcal{S}} \gamma_u^{-1} \int_{[0,1)^{|u|}} \left| \int_{[0,1)^{s-|u|}} \frac{\partial^{|u|} f(\boldsymbol{x})}{\partial \boldsymbol{x}_u} \, \mathrm{d}\boldsymbol{x}_{\mathcal{S}\setminus u} \right|^2 \mathrm{d}\boldsymbol{x}_u \right)^{1/2}$$

denotes the norm in the weighted Sobolev space (see [6] for more information on this Sobolev space).

3.2. The decay of the Walsh coefficients of smooth functions. We are now ready to show how the Walsh coefficients of smooth functions decay. This behavior is essentially captured in Definition 3.5 below. But before we get there we need several lemmas to prove the result. The following lemma is needed to show how the Walsh coefficients of functions with bounded variation decay. A simpler version of it was shown in [25, Lemma 4].

LEMMA 3.1. *Let $f \in \mathcal{L}_1([0,1)^s)$ and let $\boldsymbol{k} = (k_1, \dots, k_s) \in \mathbb{N}^s$ with $k_j = \kappa_j q^{a_j - 1} + k_j'$, where $a_j \in \mathbb{N}$, $\kappa_j \in \{1, \dots, q-1\}$, $0 \le k_j' < q^{a_j-1}$, and let $0 \le c_j < q^{a_j-1}$ for $j = 1, \dots, s$. Then*

$$\left| \int_{\prod_{j=1}^s [c_j q^{-a_j+1}, (c_j+1)q^{-a_j+1})} f(\boldsymbol{x}) \, \overline{\mathrm{wal}_{\boldsymbol{k}}(\boldsymbol{x})} \, \mathrm{d}\boldsymbol{x} \right| \le q^{-\sum_{j=1}^s (a_j - 1)} \sup_J |\Delta(f, J)|,$$

where the supremum is taken over all boxes of the form

$$J = \prod_{j=1}^s [d_j, e_j) \subseteq \prod_{j=1}^s [c_j q^{-a_j+1}, (c_j+1)q^{-a_j+1})$$

with $q^{a_j} |e_j - d_j| \in \{1, \dots, q-1\}$.

Proof. We have $\overline{\mathrm{wal}_{k_j}} = \overline{\mathrm{wal}_{\kappa_j q^{a_j-1}}} \, \overline{\mathrm{wal}_{k_j'}}$ and the function $\overline{\mathrm{wal}_{k_j'}}$ is constant on each subinterval $[c_j q^{-a_j+1}, (c_j+1)q^{-a_j+1})$. Hence we have

$$\begin{aligned} &\left| \int_{\prod_{j=1}^s [c_j q^{-a_j+1}, (c_j+1)q^{-a_j+1})} f(\boldsymbol{x}) \, \overline{\mathrm{wal}_{\boldsymbol{k}}(\boldsymbol{x})} \, \mathrm{d}\boldsymbol{x} \right| \\ &= \left| \int_{\prod_{j=1}^s [c_j q^{-a_j+1}, (c_j+1)q^{-a_j+1})} f(\boldsymbol{x}) \prod_{j=1}^s \overline{\mathrm{wal}_{\kappa_j q^{a_j-1}}(x_j)} \, \mathrm{d}\boldsymbol{x} \right|. \end{aligned}$$

Note that the function $\overline{\mathrm{wal}_{\kappa_j q^{a_j-1}}}$ is constant on each of the subintervals $[r_j q^{-a_j}, (r_j+1)q^{-a_j})$ for $r_j = 0, \dots, q^{a_j} - 1$ for $j = 1, \dots, s$. Without loss of generality we

may assume that $c_j = 0$; for all other c_j the result follows by the same arguments. Thus we have

$$\int_{\prod_{j=1}^s [0,q^{-a_j+1})} f(\boldsymbol{x}) \prod_{j=1}^s \overline{\mathrm{wal}_{\kappa_j q^{a_j-1}}(x_j)}\,\mathrm{d}\boldsymbol{x}$$

$$= \sum_{r_1,\ldots,r_s=0}^{q-1} \prod_{j=1}^s \overline{\mathrm{wal}_{\kappa_j}(r_j/q)} \int_{\prod_{j=1}^s [r_j q^{-a_j},(r_j+1)q^{-a_j})} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}.$$

Now let $a_{(r_1,\ldots,r_s)} = \int_{\prod_{j=1}^s [r_j q^{-a_j},(r_j+1)q^{-a_j})} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$ and for a given $0 \le r_1,\ldots,r_s < q$ let

$$A(r_1,\ldots,r_s) = q^{-s} \sum_{t_1,\ldots,t_s=0}^{q-1} \sum_{\emptyset \ne u \subseteq \{1,\ldots,s\}} (-1)^{|u|} a_{(\boldsymbol{t}_u, \boldsymbol{r}_{\{1,\ldots,s\}\setminus u})}, \tag{3.6}$$

where $(\boldsymbol{t}_u, \boldsymbol{r}_{\{1,\ldots,s\}\setminus u})$ denotes the vector obtained by setting the jth coordinate to t_j if $j \in u$ and r_j if $j \notin u$. Further let

$$B(r_1,\ldots,r_s) = q^{-s} \sum_{t_1,\ldots,t_s=0}^{q-1} \sum_{u \subseteq \{1,\ldots,s\}} (-1)^{|u|} a_{(\boldsymbol{t}_u, \boldsymbol{r}_{\{1,\ldots,s\}\setminus u})}.$$

Then we have

$$\sum_{r_1,\ldots,r_s=0}^{q-1} \prod_{j=1}^s \overline{\mathrm{wal}_{\kappa_j}(r_j/q)} a_{(r_1,\ldots,r_s)}$$

$$= -\sum_{r_1,\ldots,r_s=0}^{q-1} \prod_{j=1}^s \overline{\mathrm{wal}_{\kappa_j}(r_j/q)} A(r_1,\ldots,r_s)$$

$$+ \sum_{r_1,\ldots,r_s=0}^{q-1} \prod_{j=1}^s \overline{\mathrm{wal}_{\kappa_j}(r_j/q)} (a_{(r_1,\ldots,r_s)} + A(r_1,\ldots,r_s)).$$

Since $\sum_{r=0}^{q-1} \overline{\mathrm{wal}_\kappa(r/q)} = 0$ and $A(r_1,\ldots,r_s)$ is a sum where each summand does not depend on at least one r_j, i.e., the case $u = \emptyset$ is excluded in (3.6), it follows that the first sum on the right-hand side above is zero. Further we have $a_{(r_1,\ldots,r_s)} + A(r_1,\ldots,r_s) = B(r_1,\ldots,r_s)$ and thus

$$\left| \sum_{r_1,\ldots,r_s=0}^{q-1} \prod_{j=1}^s \overline{\mathrm{wal}_{\kappa_j}(r_j/q)} a_{(r_1,\ldots,r_s)} \right| \le \sum_{r_1,\ldots,r_s=0}^{q-1} |B(r_1,\ldots,r_s)|.$$

We have

$$|B(r_1,\ldots,r_s)| \le \max_{\boldsymbol{t} \in \{0,\ldots,q-1\}^s} \left| \sum_{u \subseteq \{1,\ldots,s\}} (-1)^{|u|} a_{(\boldsymbol{t}_u, \boldsymbol{r}_{\{1,\ldots,s\}\setminus u})} \right|.$$

Therefore we have

$$\begin{aligned}
&\left| \int_{\prod_{j=1}^s [0,q^{-a_j+1})} f(\boldsymbol{x}) \prod_{j=1}^s \overline{\mathrm{wal}_{\kappa_j q^{a_j-1}}(x_j)}\,\mathrm{d}\boldsymbol{x} \right| \\
&\le \sum_{r_1,\ldots,r_s=0}^{q-1} \max_{\boldsymbol{t}\in\{0,\ldots,q-1\}^s} \left| \sum_{u\subseteq\{1,\ldots,s\}} (-1)^{|u|} a_{(\boldsymbol{t}_u,\boldsymbol{r}_{\{1,\ldots,s\}\setminus u})} \right| \\
&\le q^s \max_{\boldsymbol{r},\boldsymbol{t}\in\{0,\ldots,q-1\}^s} \left| \sum_{u\subseteq\{1,\ldots,s\}} (-1)^{|u|} a_{(\boldsymbol{t}_u,\boldsymbol{r}_{\{1,\ldots,s\}\setminus u})} \right|.
\end{aligned}$$

Note that if in the above maximum there is a j such that $r_j = t_j$, then it follows that

$$\left| \sum_{u\subseteq\{1,\ldots,s\}} (-1)^{|u|} a_{(\boldsymbol{t}_u,\boldsymbol{r}_{\{1,\ldots,s\}\setminus u})} \right| = 0.$$

Hence we may in the following assume without loss of generality that the maximum in the last line of the inequality above is taken on for $\boldsymbol{r} = (r_1,\ldots,r_s)$ and $\boldsymbol{t} = (t_1,\ldots,t_s)$, which satisfy $r_j \ne t_j$ for $j = 1,\ldots,s$.

We have

$$\begin{aligned}
&\left| \sum_{u\subseteq\{1,\ldots,s\}} (-1)^{|u|} a_{(\boldsymbol{t}_u,\boldsymbol{r}_{\{1,\ldots,s\}\setminus u})} \right| \\
&= \left| \int_{\prod_{j=1}^s [r_j q^{-a_j},(r_j+1)q^{-a_j})} \sum_{u\subseteq\{1,\ldots,s\}} (-1)^{|u|} f(\boldsymbol{x}+\boldsymbol{y}_u)\,\mathrm{d}\boldsymbol{x} \right|,
\end{aligned}$$

where $\boldsymbol{y}_u = (y_1,\ldots,y_s)$ with $y_j = 0$ for $j \notin u$ and $y_j = (t_j - r_j)q^{-a_j}$ for $j \in u$. We can write

$$\sum_{u\subseteq\{1,\ldots,s\}} (-1)^{|u|} f(\boldsymbol{x}+\boldsymbol{y}_u) = \Delta(f, J_{\boldsymbol{x}}),$$

where $J_{\boldsymbol{x}} = \prod_{j=1}^s [\min(x_j, x_j + (t_j - r_j)q^{-a_j}), \max(x_j, x_j + (t_j - r_j)q^{-a_j}))$. Therefore it follows that

$$\left| \sum_{u\subseteq\{1,\ldots,s\}} (-1)^{|u|} a_{(\boldsymbol{t}_u,\boldsymbol{r}_{\{1,\ldots,s\}\setminus u})} \right| \le q^{-\sum_{j=1}^s a_j} \sup_{\boldsymbol{x}\in\prod_{j=1}^s [r_j q^{-a_j},(r_j+1)q^{-a_j})} |\Delta(f, J_{\boldsymbol{x}})|.$$

The result follows. □

In the following lemma we now obtain a bound on the Walsh coefficients for functions of bounded variation. It is a generalization of [25, Proposition 6].

LEMMA 3.2. *Let $0 < \lambda \le 1$ and let $f \in \mathcal{L}_2([0,1)^s)$ satisfy $V_{\lambda,1,1,\boldsymbol{\gamma}}(f) < \infty$. Then for any $\boldsymbol{k} \in \mathbb{N}_0^s \setminus \{\boldsymbol{0}\}$ the $\boldsymbol{k}$th Walsh coefficient of f satisfies*

$$|\hat{f}(\boldsymbol{k})| \le q^{|u|-\lambda\sum_{j\in u}(a_j-1)} V_{\lambda,1}^{(|u|)}(f_u; u),$$

where $\boldsymbol{k} = (k_1, \dots, k_s)$, $u = \{1 \le j \le s : k_j \ne 0\}$, *and for* $j \in u$ *we have* $k_j = \kappa_j q^{a_j - 1} + k_j'$, *where* $\kappa_j \in \{1, \dots, q-1\}$, $a_j \in \mathbb{N}$, *and* $0 \le k_j' < q^{a_j - 1}$.

Proof. Let $f \in \mathcal{L}_2([0,1)^s)$ with $\boldsymbol{k}$th Walsh coefficient $\hat{f}(\boldsymbol{k})$. First note that it suffices to show the result for $\boldsymbol{k} \in \mathbb{N}^s$, as otherwise we only need to replace the function f with the function $f_u(\boldsymbol{x}_u) = \int_{[0,1)^{s-|u|}} f(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}_{\mathcal{S}\setminus u}$. Hence now let $\boldsymbol{k} \in \mathbb{N}^s$ be given and let $\boldsymbol{k}' = (k_1', \dots, k_s')$. Then we have

$$\begin{aligned}
&\left| \int_{[0,1)^s} f(\boldsymbol{x}) \, \overline{\mathrm{wal}_{\boldsymbol{k}}(\boldsymbol{x})} \, \mathrm{d}\boldsymbol{x} \right| \\
&\le \sum_{\substack{0 \le c_j < q^{a_j - 1} \\ 1 \le j \le s}} \left| \int_{c_1 q^{-a_1 + 1}}^{(c_1+1) q^{-a_1+1}} \cdots \int_{c_s q^{-a_s+1}}^{(c_s+1) q^{-a_s+1}} f(\boldsymbol{x}) \, \overline{\mathrm{wal}_{\boldsymbol{k}}(\boldsymbol{x})} \, \mathrm{d}\boldsymbol{x} \right|.
\end{aligned}$$

Now we use Lemma 3.1 and thereby obtain that the above sum is bounded by

$$\sum_{0 \le c_1 < q^{a_1 - 1}} \cdots \sum_{0 \le c_s < q^{a_s - 1}} q^{-\sum_{j=1}^s (a_j - 1)} \sup_J |\Delta(f, J)|,$$

where the supremum is taken over all boxes $J = \prod_{j=1}^s [d_j, e_j) \subseteq \prod_{j=1}^s [c_j q^{-a_j+1}, (c_j + 1) q^{-a_j+1})$ with $q^{a_j} |e_j - d_j| \in \{1, \dots, q-1\}$. Now we have

$$q^{-\sum_{j=1}^s (a_j - 1)} \sup_J |\Delta(f, J)| \le \sup_{\mathcal{P}_{\mathbf{c}}} \sum_{I \in \mathcal{P}_{\mathbf{c}}} \mathrm{Vol}(I)^{1-\lambda} |\Delta(f, I)| \frac{q^{-\sum_{j=1}^s (a_j - 1)}}{\mathrm{Vol}(I)^{1-\lambda}},$$

where the supremum on the right-hand side is taken over all partitions $\mathcal{P}_{\mathbf{c}}$ of the cube $\prod_{j=1}^s [c_j q^{-a_j+1}, (c_j+1) q^{-a_j+1})$ and where each $I \in \mathcal{P}_{\mathbf{c}}$ is of the form $I = \prod_{j=1}^s [x_j, y_j)$ with $q^{a_j} |y_j - x_j| \in \{1, \dots, q-1\}$. We have $q^{-\sum_{j=1}^s a_j} \le \mathrm{Vol}(I) \le q^{-\sum_{j=1}^s (a_j - 1)}$ and therefore

$$\frac{q^{-\sum_{j=1}^s (a_j - 1)}}{\mathrm{Vol}(I)^{1-\lambda}} \le \mathrm{Vol}(I)^{\lambda} q^s \le q^{s - \lambda \sum_{j=1}^s (a_j - 1)}$$

and hence

$$q^{-\sum_{j=1}^s (a_j - 1)} \sup_J |\Delta(f, J)| \le q^{s - \lambda \sum_{j=1}^s (a_j - 1)} \sup_{\mathcal{P}_{\mathbf{c}}} \sum_{I \in \mathcal{P}_{\mathbf{c}}} \mathrm{Vol}(I)^{1-\lambda} |\Delta(f, I)|.$$

Note that

$$\sum_{\substack{0 \le c_j < q^{a_j - 1} \\ 1 \le j \le s}} \sup_{\mathcal{P}_{\mathbf{c}}} \sum_{I \in \mathcal{P}_{\mathbf{c}}} \mathrm{Vol}(I)^{1-\lambda} |\Delta(f, I)| \le \sup_{\mathcal{P}} \sum_{J \in \mathcal{P}} \mathrm{Vol}(J) \frac{|\Delta(f, J)|}{\mathrm{Vol}(J)^{\lambda}},$$

where the supremum on the left-hand side is taken over all partitions $\mathcal{P}$ of the cube $\prod_{j=1}^s [c_j q^{-a_j+1}, (c_j + 1) q^{-a_j+1})$ into subintervals and the supremum on the right-hand side is taken over all partitions of $[0,1)^s$ into subintervals. Thus the result follows. ∎

For the next lemma we will need the following two functions. For $\kappa \in \{1, \dots, q-1\}$ now let

$$\upsilon_\kappa = \sum_{r=0}^{q-1} r \, \mathrm{wal}_\kappa(r/q).$$

If q is chosen to be a prime number and the bijections φ and η are chosen to be the identity, then $\upsilon_\kappa = q(\mathrm{e}^{2\pi \mathrm{i}\kappa/q}-1)^{-1}$; see [4, Appendix A].

Further for $l \in \{1,\dots,q-1\}$ we define the function $\zeta_a(x) = \sum_{r=0}^{x_a-1} \overline{\mathrm{wal}_l(r/q)}$, where $a \geq 1$ and $x = x_1 q^{-1} + x_2 q^{-2} + \cdots$ and where for $x_a = 0$ we set $\zeta_a(x) = 0$. The function ζ_a depends on x only through x_a, thus it is a step-function which is constant on the intervals $[cq^{-a}, (c+1)q^{-a})$ for $c = 0,\dots,q^a-1$. By [25, Proposition 5] it follows that ζ_a can be represented by a finite Walsh series. Indeed, there are numbers $c_0,\dots,c_{q-1}$ (which depend on l but not on a) such that

$$\zeta_a(x) = \sum_{z=0}^{q-1} c_z \overline{\mathrm{wal}_{zq^{a-1}}(x)}.$$

If q is chosen to be a prime number and the bijections φ and η are chosen to be the identity, then $\zeta_a(x) = (1-\overline{\mathrm{wal}_{lq^{a-1}}(x)})(1-\overline{\mathrm{wal}_l(1/q)})^{-1}$, i.e., $c_0 = (1-\overline{\mathrm{wal}_l(1/q)})^{-1}$, $c_l = (\overline{\mathrm{wal}_l(1/q)}-1)^{-1}$, and $c_z = 0$ for $z \neq 0, l$.

The following lemma will be used in the induction step for differentiable functions. For example, for a differentiable function $F:\mathbb{R}\to\mathbb{R}$ given by $F(x) = \int_0^x f(y)\,\mathrm{d}y$ we can calculate the Walsh coefficients using integration by parts in the following way: For $k > 0$ we have

$$\begin{aligned}\hat{F}(k) &= \int_0^1 F(x)\overline{\mathrm{wal}_k(x)}\,\mathrm{d}x = \left[\int_0^x \overline{\mathrm{wal}_k(y)}\,\mathrm{d}y F(x)\right]_0^1 - \int_0^1 f(x)\int_0^x \overline{\mathrm{wal}_k(y)}\,\mathrm{d}y\,\mathrm{d}x \\ &= -\int_0^1 f(x)\int_0^x \overline{\mathrm{wal}_k(y)}\,\mathrm{d}y\,\mathrm{d}x, \qquad (3.7)\end{aligned}$$

where we used $\int_0^0 \overline{\mathrm{wal}_k(x)}\,\mathrm{d}x = \int_0^1 \overline{\mathrm{wal}_k(x)}\,\mathrm{d}x = 0$. For $k = 0$ on other hand we obtain

$$\begin{aligned}\hat{F}(0) &= \int_0^1 F(x)\overline{\mathrm{wal}_0(x)}\,\mathrm{d}x = \left[\int_0^x \overline{\mathrm{wal}_0(y)}\,\mathrm{d}y F(x)\right]_0^1 - \int_0^1 f(x)\int_0^x \overline{\mathrm{wal}_0(y)}\,\mathrm{d}y\,\mathrm{d}x \\ &= \int_0^1 f(x)\,\mathrm{d}x - \int_0^1 f(x)\int_0^x \overline{\mathrm{wal}_0(y)}\,\mathrm{d}y\,\mathrm{d}x. \qquad (3.8)\end{aligned}$$

Thus if we know the Walsh series for f, then we can easily calculate the Walsh series for F, provided that we know the Walsh series for $\int_0^x \overline{\mathrm{wal}_k(y)}\,\mathrm{d}y$. This will be calculated in the following lemma. It appeared in a simpler form in [10].

LEMMA 3.3. *For $k \in \mathbb{N}_0$ and $x \in [0,1)$ define $J_k(x) = \int_0^x \overline{\mathrm{wal}_k(y)}\,\mathrm{d}y$. For $k \geq 1$ let $k = lq^{a-1}+k'$, where $l \in \{1,\dots,q-1\}$, $a \geq 1$, and $0 \leq k' < q^{a-1}$. Then J_k can be represented by a Walsh series which is given by*

$$J_k(x) = q^{-a}\left(\sum_{z=0}^{q-1} c_z \overline{\mathrm{wal}_{zq^{a-1}+k'}(x)} + 2^{-1}\overline{\mathrm{wal}_k(x)} + \sum_{c=1}^{\infty}\sum_{\kappa=1}^{q-1} q^{-c-1}\upsilon_\kappa \overline{\mathrm{wal}_{\kappa q^{a+c-1}+k}(x)}\right).$$

Further we have

$$J_0(x) = 1/2 + \sum_{c=1}^{\infty}\sum_{\kappa=1}^{q-1} q^{-c-1}\upsilon_\kappa\,\overline{\mathrm{wal}_{\kappa q^{c-1}}(x)}.$$

Proof. Let $k = lq^{a-1}+k'$ with $a \geq 1$, $0 \leq k' < q^{a-1}$, and $l \in \{1,\dots,q-1\}$. The function $\overline{\mathrm{wal}_{lq^{a-1}}(y)}$ is constant on each interval $[rq^{-a},(r+1)q^{-a})$ and

$\overline{\mathrm{wal}_{k'}(y)}$ is constant on each interval $[cq^{-a+1},(c+1)q^{-a+1})$. We have $\overline{\mathrm{wal}_k(y)} = \overline{\mathrm{wal}_{lq^{a-1}}(y)}\,\overline{\mathrm{wal}_{k'}(y)}$. For any $0 \le c < q^{a-1}$ we have

$$\begin{aligned}\int_{[cq^{-a+1},(c+1)q^{-a+1})} \overline{\mathrm{wal}_k(y)}\,\mathrm{d}y &= \overline{\mathrm{wal}_{k'}(cq^{-a+1})}\int_{[cq^{-a+1},(c+1)q^{-a+1})} \overline{\mathrm{wal}_{lq^{a-1}}(y)}\,\mathrm{d}y\\ &= \overline{\mathrm{wal}_{k'}(cq^{-a+1})}q^{-a}\sum_{r=0}^{q-1}\mathrm{wal}_l(r/q)\\ &= 0.\end{aligned}$$

Thus we have

$$J_k(x) = \overline{\mathrm{wal}_{k'}(x)}J_{lq^{a-1}}(x).$$

Let $x = x_1q^{-1} + x_2q^{-2} + \cdots$ and $y = x_{a+1}q^{-1} + x_{a+2}q^{-2} + \cdots$; then we have

$$J_{lq^{a-1}}(x) = q^{-a}\sum_{r=0}^{x_a-1}\overline{\mathrm{wal}_l(r/q)} + q^{-a}\overline{\mathrm{wal}_l(x_a/q)}y.$$

We now investigate the Walsh series representation of the function $J_{lq^{a-1}}(x)$. First note that $\overline{\mathrm{wal}_l(x_a/q)} = \overline{\mathrm{wal}_{lq^{a-1}}(x)}$. Further, by a slight adaption of [4, eq. (30)] we obtain

$$y = 1/2 + \sum_{c=1}^{\infty}\sum_{\kappa=1}^{q-1} q^{-c-1}\upsilon_\kappa\,\overline{\mathrm{wal}_{\kappa q^{c-1}}(y)}. \tag{3.9}$$

As $\overline{\mathrm{wal}_{\kappa q^{c-1}}(y)} = \overline{\mathrm{wal}_{\kappa q^{a+c-1}}(x)}$ we obtain

$$y = 1/2 + \sum_{c=1}^{\infty}\sum_{\kappa=1}^{q-1} q^{-c-1}\upsilon_\kappa\,\overline{\mathrm{wal}_{\kappa q^{a+c-1}}(x)}.$$

As noted above, the Walsh series of $\zeta_a(x) = \sum_{r=0}^{x_a-1}\overline{\mathrm{wal}_l(r/q)}$, where for $x_a = 0$ we set $\zeta_a(x) = 0$, can be written as

$$\zeta_a(x) = \sum_{z=0}^{q-1} c_z\overline{\mathrm{wal}_{zq^{a-1}}(x)}.$$

Altogether we obtain

$$\begin{aligned}q^aJ_{lq^{a-1}}(x) = \sum_{z=0}^{q-1} c_z\overline{\mathrm{wal}_{zq^{a-1}}(x)} &+ 2^{-1}\overline{\mathrm{wal}_{lq^{a-1}}(x)}\\ &+ \sum_{c=1}^{\infty}\sum_{\kappa=1}^{q-1} q^{-c-1}\upsilon_\kappa\overline{\mathrm{wal}_{\kappa q^{a+c-1}+lq^{a-1}}(x)}\end{aligned}$$

and therefore

$$q^aJ_k(x) = \sum_{z=0}^{q-1} c_z\overline{\mathrm{wal}_{zq^{a-1}+k'}(x)} + 2^{-1}\overline{\mathrm{wal}_k(x)} + \sum_{c=1}^{\infty}\sum_{\kappa=1}^{q-1} q^{-c-1}\upsilon_\kappa\overline{\mathrm{wal}_{\kappa q^{a+c-1}+k}(x)}.$$

The result for $k = 0$ follows easily from (3.9). □

Note that Lemma 3.3 can easily be generalized to arbitrary dimensions s, since for $\boldsymbol{k} = (k_1, \dots, k_s) \in \mathbb{N}_0^s$ we have for any $\boldsymbol{x} = (x_1, \dots, x_s) \in [0,1)^s$ that

$$J_{\boldsymbol{k}}(\boldsymbol{x}) = \int_{[0,\boldsymbol{x})} \overline{\mathrm{wal}_{\boldsymbol{k}}(\boldsymbol{y})}\, \mathrm{d}\boldsymbol{y} = \prod_{j=1}^{s} J_{k_j}(x_j),$$

where $[0, \boldsymbol{x}) = \prod_{j=1}^{s} [0, x_j)$.

The next lemma shows how the Walsh coefficients of a function $F = \int f$ can be obtained from the Walsh coefficients of f.

LEMMA 3.4. *Let $f \in \mathcal{L}_2([0,1)^s)$ and let $F(\boldsymbol{x}) = \int_{[0,\boldsymbol{x})} f(\boldsymbol{y})\, \mathrm{d}\boldsymbol{y}$, where $[0, \boldsymbol{x}) = \prod_{j=1}^{s} [0, x_j)$ with $\boldsymbol{x} = (x_1, \dots, x_s)$. Further let $\hat{F}(\boldsymbol{k})$ denote the $\boldsymbol{k}$th Walsh coefficient of F. Let $\boldsymbol{k} = (k_1, \dots, k_s) \in \mathbb{N}_0^s$ and let $U = \{1 \le j \le s : k_j \ne 0\}$. For $j \in U$ let $k_j = l_j q^{a_j - 1} + k_j'$, $0 < l_j < q$ and $0 \le k_j' < q^{a_j - 1}$, and further let $\boldsymbol{k}' = (k_1', \dots, k_s')$, where $k_j' = 0$ for $j \notin U$. Then we have*

$$\hat{F}(\boldsymbol{k}) = q^{-\sum_{j \in U} a_j} \sum_{U \subseteq v \subseteq \mathcal{S}} (-1)^{|v|} \sum_{\boldsymbol{h}_v \in \mathbb{N}_0^{|v|}} \hat{f}(\boldsymbol{k}' + (\boldsymbol{h}_v, \boldsymbol{0}))\, \chi_{U,v,\boldsymbol{k}}(\boldsymbol{h}_v),$$

where $(\boldsymbol{h}_v, \boldsymbol{0})$ denotes the s-dimensional vector whose jth component is h_j for $j \in v$ and 0 otherwise and where for $\boldsymbol{h}_v = (h_j)_{j \in v} \in \mathbb{N}_0^{|v|}$ we set

$$\chi_{U,v,\boldsymbol{k}}(\boldsymbol{h}_v) = \prod_{j \in U} \rho_{k_j}(h_j) \prod_{j \in v \setminus U} \phi(h_j).$$

Here

$$\rho_{k_j}(h_j) = \begin{cases} c_z + 2^{-1} 1_{z = l_j} & \text{for } h_j = z q^{a_j - 1}, \\ \upsilon_z q^{-i-1} & \text{for } h = z q^{a_j - 1 + i} + l_j q^{a_j - 1},\ i > 0,\ 0 < z < q, \\ 0 & \text{otherwise,} \end{cases}$$

where $1_{z = l_j} = 1$ for $z = l_j$ and 0 otherwise, and

$$\phi(h_j) = \begin{cases} 2^{-1} & \text{for } h_j = 0, \\ \upsilon_z q^{-i-1} & \text{for } h = z q^{a_j - 1 + i},\ i > 0,\ 0 < z < q, \\ 0 & \text{otherwise.} \end{cases}$$

Proof. Using integration by parts in each coordinate, Fubini's theorem, and $J_k(0) = J_k(1) = 0$ for any $k \in \mathbb{N}$ (see (3.7) and (3.8) for one-dimensional examples), it follows that

$$\hat{F}(\boldsymbol{k}) = \int_{[0,1)^s} F(\boldsymbol{x}) \overline{\mathrm{wal}_{\boldsymbol{k}}(\boldsymbol{x})}\, \mathrm{d}\boldsymbol{x} = \sum_{U \subseteq v \subseteq \mathcal{S}} (-1)^{|v|} \int_{[0,1)^s} J_{\boldsymbol{k}_v}(\boldsymbol{x}_v) f(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x},$$

where for $\boldsymbol{k} = (k_1, \dots, k_s)$ and $\boldsymbol{x} = (x_1, \dots, x_s)$ we have $J_{\boldsymbol{k}_v}(\boldsymbol{x}_v) = \prod_{j \in v} J_{k_j}(x_j)$.

Using the Walsh series expansion of $J_{\boldsymbol{k}}$ given by Lemma 3.3 we can now express the Walsh coefficient $\hat{F}(\boldsymbol{k})$ as a sum of the Walsh coefficients $\hat{f}(\boldsymbol{h})$, from which the result follows. □

The following definition now captures the essence of the decay of the Walsh coefficients of smooth functions and will be used in the statement of subsequent lemmas, theorems, and corollaries.

DEFINITION 3.5. *Let $k = k(v; a_1, \dots, a_v) = \kappa_1 q^{a_1 - 1} + \cdots + \kappa_v q^{a_v - 1}$ with $v \geq 1$, let $\kappa_1, \dots, \kappa_v \in \{1, \dots, q-1\}$, and let $1 \leq a_v < \cdots < a_1$ be a natural number. For $k = 0$ we set $v = 0$, i.e., $k(0) = 0$. A function $\mathcal{B} : \mathbb{N}_0 \to \mathbb{R}$ is called q-adically nonincreasing if $\mathcal{B}(k) = \mathcal{B}(k(v; a_1, \dots, a_v))$ is nonincreasing in v and each a_i for $i = 1, \dots, v$; that is, for any $v \geq 0$ we have*

$$\mathcal{B}(k(v; a_1, \dots, a_v)) \geq \mathcal{B}(k(v+1; a'_1, \dots, a'_{v+1}))$$

with $1 \leq a'_{v+1} < \cdots < a'_1$ and $a_1, \dots, a_v \in \{a'_1, \dots, a'_{v+1}\}$ and for an arbitrary $1 \leq i \leq v$ we have

$$\mathcal{B}(k(v; a_1, \dots, a_v)) \geq \mathcal{B}(k(v; a_1, \dots, a_{i-1}, a_i + 1, a_{i+1}, \dots, a_v)),$$

provided that $a_i + 1 < a_{i-1}$ in case $1 < i \leq v$.

In the following lemma we give a bound on the Walsh coefficients of F if f satisfies some smoothness condition.

LEMMA 3.6. *Let $\mathcal{B} : \mathbb{N}_0^s \to [0, \infty)$ be a q-adically nonincreasing function in each variable. Let $f \in \mathcal{L}_2([0,1)^s)$ and let the Walsh coefficients of f satisfy*

$$|\hat{f}(\boldsymbol{k})| \leq \mathcal{B}(\boldsymbol{k}) \quad \text{for all } \boldsymbol{k} \in \mathbb{N}_0^s.$$

Let $F(\boldsymbol{x}) = \int_{[0,\boldsymbol{x})} f(\boldsymbol{y}) \, \mathrm{d}\boldsymbol{y}$, where $[0, \boldsymbol{x}) = \prod_{j=1}^s [0, x_j)$ with $\boldsymbol{x} = (x_1, \dots, x_s)$. Further let $\hat{F}(\boldsymbol{k})$ denote the $\boldsymbol{k}$th Walsh coefficient of F. Let $\boldsymbol{k} = (k_1, \dots, k_s) \in \mathbb{N}_0^s \setminus \{\boldsymbol{0}\}$ and let $U = \{1 \leq j \leq s : k_j \neq 0\}$. For $j \in U$ let $k_j = l_j q^{a_j - 1} + k'_j$ and $0 \leq k'_j < q^{a_j - 1}$ and further let $\boldsymbol{k}' = (k'_1, \dots, k'_s)$, where $k'_j = 0$ for $j \notin U$. Then there is a constant $C_{s,U} > 0$ independent of $\boldsymbol{k}$ such that

$$|\hat{F}(\boldsymbol{k})| \leq C_{s,U} \, q^{-\sum_{j \in U} a_j} \mathcal{B}(\boldsymbol{k}').$$

Proof. Using Lemma 3.4 we obtain that

$$|\hat{F}(\boldsymbol{k})| \leq q^{-\sum_{j \in U} a_j} \mathcal{B}(\boldsymbol{k}') \sum_{U \subseteq v \subseteq \mathcal{S}} \sum_{\boldsymbol{h}_v \in \mathbb{N}_0^{|v|}} |\chi_{U,v,\boldsymbol{k}}(\boldsymbol{h}_v)|,$$

as $|\hat{f}(\boldsymbol{k}' + (\boldsymbol{h}_v, \boldsymbol{0}))| \leq \mathcal{B}(\boldsymbol{k}' + (\boldsymbol{h}_v, \boldsymbol{0})) \leq \mathcal{B}(\boldsymbol{k}')$ for all values of $\boldsymbol{h}_v \in \mathbb{N}_0^{|v|}$ for which $\chi_{U,v,\boldsymbol{k}}(\boldsymbol{h}_v) \neq 0$, since $\mathcal{B}$ is q-adically nonincreasing in each variable.

Thus it remains to bound $\sum_{U \subseteq v \subseteq \mathcal{S}} \sum_{\boldsymbol{h}_v \in \mathbb{N}_0^{|v|}} |\chi_{U,v,\boldsymbol{k}}(\boldsymbol{h}_v)|$ independently of $\boldsymbol{k}$. We only prove the case where q is chosen to be a prime number and the bijections φ and η are chosen to be the identity, as in this case we can obtain an explicit constant $C_{s,U} > 0$. The general case can be obtained by similar arguments using the result from Lemma 3.4.

Using the notation from Lemma 3.4 we have

$$\begin{aligned}\sum_{h \in \mathbb{N}_0} |\rho_{k_j}(h)| &= |1 - \omega_q^{-l_j}|^{-1} + 2^{-1} |1 + \omega_q^{-l_j}| |\omega_q^{-l_j} - 1|^{-1} \\ &\quad + \sum_{i=1}^{\infty} q^{-i} \sum_{z=1}^{q-1} |\mathrm{e}^{2\pi \mathrm{i} z/q} - 1|^{-1} \\ &\leq 3(2 - 2\cos(2\pi/q))^{-1/2}\end{aligned}$$

and

$$\sum_{h=0}^{\infty} |\phi(h)| = 2^{-1} + \sum_{i=1}^{\infty} q^{-i} \sum_{z=1}^{q-1} |\mathrm{e}^{2\pi \mathrm{i} z/q} - 1|^{-1} \leq 2^{-1} + (2 - 2\cos(2\pi/q))^{-1/2}.$$

Therefore we have

$$\sum_{U\subseteq v\subseteq S}\sum_{\boldsymbol{h}_v\in\mathbb{N}_0^{|v|}}|\chi_{U,v,\boldsymbol{k}}(\boldsymbol{h}_v)|$$

$$\le 3^{|U|}(2-2\cos(2\pi/q))^{-|U|/2}(3/2+(2-2\cos(2\pi/q))^{-1/2})^{s-|U|}$$

and hence we can choose

$$C_{s,U}=3^{|U|}(2-2\cos(2\pi/q))^{-|U|/2}(3/2+(2-2\cos(2\pi/q))^{-1/2})^{s-|U|} \tag{3.10}$$

in Lemma 3.6 for this case. □

We use the above results now to establish an upper bound on the Walsh coefficients of a polynomial. The proof will give a glimpse of how the argument will work for more general function classes.

LEMMA 3.7. *Let $k=\kappa_1q^{a_1-1}+\cdots+\kappa_vq^{a_v-1}$ with $v\ge 1$, $\kappa_1,\dots,\kappa_v\in\{1,\dots,q-1\}$ and $1\le a_v<\cdots<a_1$. For $v=0$ let $k=0$. Let $f:[0,1)\to\mathbb{R}$ be the polynomial $f(x)=f_0+f_1x+\cdots+f_ix^i$ with $f_i\ne 0$ and let $\hat f(k)$ denote the kth Walsh coefficient of f. Then for $v\ge 0$ there are constants $0<C_{f,i,v}<\infty$ such that*

$$|\hat f(k)|\le C_{f,i,v}q^{-a_1-\cdots-a_v},$$

where we can choose $C_{f,i,v}=0$ for $v>i$.

Proof. Let $f(x)=f_0+f_1x+\cdots+f_ix^i$, where $i=\deg(f)$ (that is, $f_i\ne 0$). Then we have $f^{(i)}(x)=i!f_i\ne 0$. As $f^{(i)}$ is a constant function, its Walsh series representation is simply given by $f^{(i)}(x)=i!f_i$. Now we use Lemma 3.6. The dimension s in our case is 1, and we can choose the function $\mathcal{B}_1$ by $\mathcal{B}_1(0)=i!|f_i|$ and for $k>0$ we set $\mathcal{B}_1(k)=0$. Note the function $\mathcal{B}_1$ defined this way is a q-adically nonincreasing function. Then it follows that there is a constant $C_1>0$ such that the Walsh coefficients of the function $\int_0^x f^{(i)}(t)\,\mathrm{d}t=f^{(i-1)}(x)-f^{(i-1)}(0)$ are bounded by $C_1q^{-a_1}i!|f_i|$ for all k where $v=1$ and the Walsh coefficients are 0 for $v>1$. The Walsh coefficient for $k=0$ is given by $f^{(i-2)}(1)-f^{(i-2)}(0)-f^{(i-1)}(0)$.

Now consider the function $\int_0^x f^{(i-1)}(t)\,\mathrm{d}t=f^{(i-2)}(x)-f^{(i-2)}(0)$. It follows from the above and Lemma 3.6 that the Walsh coefficients of $f^{(i-2)}(x)-f^{(i-2)}(0)$ can be bounded by a q-adically nonincreasing function $\mathcal{B}_2$. Indeed there are constants $C_2,C_3>0$ such that we can choose $\mathcal{B}_2(0)=|f^{(i-2)}(1)-f^{(i-2)}(0)-f^{(i-1)}(0)|$, $\mathcal{B}_2(k)=C_2q^{-a_1}$ for $v=1$, $\mathcal{B}_2(k)=C_3q^{-a_1-a_2}$ for $v=2$, and $\mathcal{B}_2(k)=0$ for $v>2$. Again $\mathcal{B}_2$ is a q-adically nonincreasing function and Lemma 3.6 can again be used.

By using the above argument iteratively we obtain that there is a constant $C>0$ such that $|\hat f(k)|\le Cq^{-a_1-\cdots-a_v}$ for $k=\kappa_1q^{a_1-1}+\cdots+\kappa_vq^{a_v-1}$ with $\kappa_1,\dots,\kappa_v\in\{1,\dots,q-1\}$ and $1\le a_v<\cdots<a_1$. The result thus follows. □

For the case where q is chosen to be a prime number and the bijections φ and η are chosen to be the identity and for $0\le v\le i$ we can choose

$$C_{f,i,v}=\bar C^v\sum_{l=v}^{i}C'^{i-l}l!|f_l|, \tag{3.11}$$

where $\bar C=(2-2\cos(2\pi/q))^{-1/2}$ and $C'=3/2+(2-2\cos(2\pi/q))^{-1/2}$ in Lemma 3.7.

Let $f : [0,1)^s \to \mathbb{R}$ be such that the partial mixed derivatives up to order $\delta \geq 1$ in each variable exist and are continuous. We need some further notation: Let $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_s)$ and

$$f^{(\boldsymbol{\tau})}(\boldsymbol{x}) = \frac{\partial^{\tau_1+\cdots+\tau_s}}{\partial x_1^{\tau_1} \cdots \partial x_s^{\tau_s}} f(\boldsymbol{x}).$$

For $\boldsymbol{\tau} \in \{0, \ldots, \delta\}^s$ let $u(\boldsymbol{\tau}) = \{1 \leq j \leq s : \tau_j = \delta\}$. Let $\boldsymbol{\gamma} = (\gamma_v)_{v \subset \mathbb{N}}$ be an indexed set of nonnegative real numbers. Let $v(\boldsymbol{\tau}) = \{1 \leq j \leq s : \tau_j > 0\}$. Then for $0 < \lambda \leq 1$ and $\mathfrak{p}, \mathfrak{q}, \mathfrak{r} \geq 1$ ($\mathfrak{p}, \mathfrak{q}, \mathfrak{r}$ do not appear in the subscript of N as they do not have influence on our subsequent bounds—we assume only that they are greater than or equal to 1) we define

$$N_{\delta,\lambda,\boldsymbol{\gamma}}(f) = \left(\sum_{\boldsymbol{\tau} \in \{0,\ldots,\delta\}^s} \gamma_{v(\boldsymbol{\tau})}^{-1} \left[V_{\lambda,\mathfrak{p},\mathfrak{q},\mathbf{1}}^{(|u(\boldsymbol{\tau})|)} (f^{(\boldsymbol{\tau})}(\cdot, \mathbf{0}_{\mathcal{S}\setminus u(\boldsymbol{\tau})})) \right]^{\mathfrak{r}} \right)^{1/\mathfrak{r}}, \tag{3.12}$$

where, for clarity, we introduce the additional superscript $(|u(\boldsymbol{\tau})|)$ in the Hardy and Krause variation $V_{\lambda,\mathfrak{p},\mathfrak{q},\mathbf{1}}^{(|u(\boldsymbol{\tau})|)}$ which indicates the dimension of the function and where for $u(\boldsymbol{\tau}) = \emptyset$ we set $V_{\lambda,\mathfrak{p},\mathfrak{q},\mathbf{1}}^{(|u(\boldsymbol{\tau})|}(f^{(\boldsymbol{\tau})}(\cdot, \mathbf{0}_{\mathcal{S}\setminus u(\boldsymbol{\tau})})) = |f^{(\boldsymbol{\tau})}(\mathbf{0})|$.

The weights $\boldsymbol{\gamma}$ are introduced to modify the importance of various coordinate projections and were first introduced in [30]; see also [6, 7]. If for some $v' \subseteq \mathcal{S}$ the weight $\gamma_{v'} = 0$, then we assume that the function f satisfies $V_{\lambda,\mathfrak{p},\mathfrak{q},\mathbf{1}}^{(|v'|)}(f^{(\boldsymbol{\tau})}(\cdot, \mathbf{0}_{\mathcal{S}\setminus v'})) = 0$ for all $\boldsymbol{\tau} \in \{0, \ldots, \delta\}^s$ with $v(\boldsymbol{\tau}) = v'$, and in (3.12) we formally set $0/0 = 0$.

The parameters in the definition of $N_{\delta,\lambda,\boldsymbol{\gamma}}$ have the following meaning:

- δ denotes the order of partial derivatives of f required in order for $N_{\delta,\lambda,\boldsymbol{\gamma}}(f)$ to make sense;
- λ is a Hölder-type parameter or fractional order–type parameter of the generalized Hardy and Krause variation; roughly, f needs to have partial derivatives up to order $\delta + \lambda$, where for $0 < \lambda < 1$ this means some type of fractional smoothness or in dimension one a Hölder condition of order λ;
- the Vitali variation is in $\mathfrak{p}$ norm;
- $\mathfrak{q}$ is the norm in the summation of the generalized Hardy and Krause variation;
- $\mathfrak{r}$ is the norm in the summation over the $\boldsymbol{\tau}$;
- $\boldsymbol{\gamma}$ are the weights which regulate the importance of different coordinate projections.

Note that for $\lambda = 1$ and $\mathfrak{p} = \mathfrak{q} = \mathfrak{r} = 2$ the functional $N_{\delta,\lambda,\boldsymbol{\gamma}}$ is just the norm in a weighted reproducing kernel Sobolev space with continuous partial mixed derivatives up to order $\delta + 1$ in each variable. In one dimension the unweighted norm in this reproducing kernel Sobolev space is given by

$$\begin{aligned} &\langle f, g \rangle_{\mathrm{sob,an},\delta+1} \\ &\quad = f(0)g(0) + \cdots + f^{(\delta-1)}(0)g^{(\delta-1)}(0) \\ &\qquad + \int_0^1 f^{(\delta)}(x)\,\mathrm{d}x \int_0^1 g^{(\delta)}(x)\,\mathrm{d}x + \int_0^1 f^{(\delta+1)}(x) g^{(\delta+1)}(x)\,\mathrm{d}x \end{aligned} \tag{3.13}$$

and for higher dimensions one just takes the weighted tensor product of the one-dimensional reproducing kernel Sobolev spaces (see [7] for examples of weighted tensor product reproducing kernel Sobolev spaces). Let the s-dimensional weighted inner

product be denoted by $\langle\cdot,\cdot\rangle_{\mathrm{sob,an},s,\delta+1,\boldsymbol{\gamma}}$ and the corresponding norm $\|\cdot\|_{\mathrm{sob,an},s,\delta+1,\boldsymbol{\gamma}}$ (indeed if the partial mixed derivatives up to order $\delta+1$ of f are continuous on $[0,1]^s$, then we have $\|f\|_{\mathrm{sob,an},s,\delta+1,\boldsymbol{\gamma}} = N_{\delta,1,\boldsymbol{\gamma}}(f)$).

In the following we define a function μ which will be used throughout the paper: Let $\delta \geq 1$ be an integer and let $0 < \lambda \leq 1$ be a real number. Then for $\boldsymbol{k} = (k_1, \ldots, k_s)$ we set

$$\mu_{q,\delta+\lambda}(\boldsymbol{k}) = \sum_{j=1}^{s} \mu_{q,\delta+\lambda}(k_j) \tag{3.14}$$

with

$$\mu_{q,\delta+\lambda}(k) = \begin{cases} 0 & \text{for } k = 0, \\ a_1 + \cdots + a_v & \text{for } v \leq \delta, \\ a_1 + \cdots + a_\delta + \lambda a_{\delta+1} & \text{for } v > \delta, \end{cases} \tag{3.15}$$

where for $k \in \mathbb{N}$ we write $k = \kappa_1 q^{a_1-1} + \cdots + \kappa_v q^{a_v-1}$ with $v \geq 1$, $\kappa_1, \ldots, \kappa_v \in \{1, \ldots, q-1\}$, and $1 \leq a_v < \cdots < a_1$.

THEOREM 3.8. *Let $\delta \geq 1$ be an integer, let $0 < \lambda \leq 1$, let $\mathfrak{p}, \mathfrak{q}, \mathfrak{r} \geq 1$ be real numbers, and let an indexed set $\boldsymbol{\gamma} = (\gamma_v)_{v \subset \mathbb{N}}$ of nonnegative real numbers be given. Let $f : [0,1)^s \to \mathbb{R}$ be such that the partial mixed derivatives up to order δ in each variable exist and such that $N_{\delta,\lambda,\boldsymbol{\gamma}}(f) < \infty$. Then for any $\boldsymbol{k} \in \mathbb{N}_0^s \setminus \{\mathbf{0}\}$ it follows that there is a constant $C_{f,q,s,\boldsymbol{\gamma}} > 0$ independent of $\boldsymbol{k}$ such that*

$$|\hat{f}(\boldsymbol{k})| \leq C_{f,q,s,\boldsymbol{\gamma}} q^{-\mu_{q,\delta+\lambda}(\boldsymbol{k})}.$$

Proof. In order to prove the result we use the Taylor series expansion of the function f. We have

$$\begin{aligned} f(\boldsymbol{x}) = {} & \sum_{\boldsymbol{\tau} \in \{0,\ldots,\delta-1\}^s} \frac{\boldsymbol{x}^{\boldsymbol{\tau}}}{\boldsymbol{\tau}!} f^{(\boldsymbol{\tau})}(\mathbf{0}) + \sum_{\emptyset \neq u \subseteq \mathcal{S}} \sum_{\boldsymbol{\tau}_{\mathcal{S}\setminus u} \in \{0,\ldots,\delta-1\}^{s-|u|}} ((\delta-1)!)^{-|u|} \\ & \times \frac{\prod_{j \in \mathcal{S}\setminus u} x_j^{\tau_j}}{\prod_{j \in \mathcal{S}\setminus u} \tau_j!} \int_{[\mathbf{0}_u, \boldsymbol{x}_u)} f^{(\boldsymbol{\delta}_u, \boldsymbol{\tau}_{\mathcal{S}\setminus u})}(\boldsymbol{y}_u, \mathbf{0}_{\mathcal{S}\setminus u}) \prod_{j \in u} (x_j - y_j)^{\delta-1} \, \mathrm{d}\boldsymbol{y}_u. \end{aligned} \tag{3.16}$$

First note that the first sum in (3.16) is a polynomial in $\boldsymbol{x}$ and therefore the Walsh coefficients of this polynomial satisfy the desired bound by Lemma 3.7.

Now we consider the second sum. Let $\emptyset \neq u \subseteq \mathcal{S}$ with $u = \{j_1, \ldots, j_{|u|}\}$ be given. Then for $j \notin u$ the Walsh coefficients satisfy the desired bound by Lemma 3.7. Hence it remains to consider the Walsh coefficients of

$$\begin{aligned} G_u(\boldsymbol{x}_u) &= \int_{[\mathbf{0}_u, \boldsymbol{x}_u)} f^{(\boldsymbol{\delta}_u, \boldsymbol{\tau}_{\mathcal{S}\setminus u})}(\boldsymbol{y}_u, \mathbf{0}_{\mathcal{S}\setminus u}) \prod_{j \in u} (x_j - y_j)^{\delta-1} \, \mathrm{d}\boldsymbol{y}_u \\ &= \int_{[\mathbf{0}_u, \mathbf{1}_u)} f^{(\boldsymbol{\delta}_u, \boldsymbol{\tau}_{\mathcal{S}\setminus u})}(\boldsymbol{y}_u, \mathbf{0}_{\mathcal{S}\setminus u}) \prod_{j \in u} (x_j - y_j)_+^{\delta-1} \, \mathrm{d}\boldsymbol{y}_u, \end{aligned}$$

where $(x-y)_+ = \max(0, x-y)$.

By differentiating the function G_u in each variable $0 \leq k < \delta$ times we obtain (see [32, pp. 153, 154])

$$\frac{\partial^{k|u|}}{\partial \boldsymbol{x}_u^k} G_u(\boldsymbol{x}_u) = \left(\frac{(\delta-1)!}{(\delta-1-k)!} \right)^{|u|} \int_{[\mathbf{0}_u, \boldsymbol{x}_u)} f^{(\boldsymbol{\delta}_u, \boldsymbol{\tau}_{S\setminus u})}(\boldsymbol{y}_u, \mathbf{0}_{S\setminus u}) \prod_{j \in u} (x_j - y_j)^{\delta-1-k} \, \mathrm{d}\boldsymbol{y}_u.$$

Hence $\frac{\partial^{k|u|}}{\partial \boldsymbol{x}_u^k} G_u(\boldsymbol{x}_u) = 0$ if there is at least one $j \in u$ such that $x_j = 0$. Further we have

$$\frac{\partial^{\delta|u|}}{\partial \boldsymbol{x}_u^\delta} G_u(\boldsymbol{x}_u) = ((\delta-1)!)^{|u|} f^{(\boldsymbol{\delta}_u, \boldsymbol{\tau}_{S\setminus u})}(\boldsymbol{x}_u, \mathbf{0}_{S\setminus u}).$$

From $N_{\delta,\lambda,\boldsymbol{\gamma}}(f) < \infty$ it follows that the Walsh coefficients of $f^{(\boldsymbol{\delta}_u, \boldsymbol{\tau}_{\mathcal{S}\setminus u})}(\boldsymbol{x}_u, \mathbf{0}_{\mathcal{S}\setminus u})$ decay with order $\mu_{q,0+\lambda}(k)$ in each variable. Further we have

$$G_u(\boldsymbol{x}_u) = \int_{[\mathbf{0},\boldsymbol{x}_u)} \int_{[\mathbf{0},\boldsymbol{y}_1)} \cdots \int_{[\mathbf{0},\boldsymbol{y}_{\delta-1})} G_u^{(\boldsymbol{\delta}_u)}(\boldsymbol{y}_\delta)\, \mathrm{d}\boldsymbol{y}_\delta \cdots \mathrm{d}\boldsymbol{y}_1$$

as the function G_u, and its derivatives are 0 if at least one $x_j = 0$ for $j \in u$; i.e., we have

$$\int_{[\mathbf{0},\boldsymbol{x}_u)} G_u^{(\boldsymbol{\tau})}(\boldsymbol{y})\, \mathrm{d}\boldsymbol{y} = \sum_{v \subseteq u} (-1)^{u\setminus v} G_u^{(\boldsymbol{\tau}-\mathbf{1})}(\boldsymbol{x}_v, \mathbf{0}_{u\setminus v}) = G_u^{(\boldsymbol{\tau}-\mathbf{1})}(\boldsymbol{x}_u).$$

Hence it follows by repeated use of Lemma 3.6 that the desired bound holds for G_u and thus the result follows from (3.16). □

For the case where q is chosen to be a prime number and the bijections φ and η are chosen to be the identity, we can also obtain an explicit constant in Theorem 3.8. Indeed, using Lemmas 3.2 and 3.6 together with the explicit constant (3.10) and Lemma 3.7 together with the explicit constant (3.11), we obtain that the constant $C_{q,\boldsymbol{\gamma}}$ can be chosen as

$$\begin{aligned} C_{f,q,s,\boldsymbol{\gamma}} = {} & \sum_{\substack{\boldsymbol{\tau} \in \{0,\ldots,\delta-1\}^s \\ \gamma_{v(\boldsymbol{\tau})} \neq 0}} |f^{(\boldsymbol{\tau})}(\mathbf{0})| \hat{C}^{\tau_1+\cdots+\tau_s} + \sum_{\emptyset \neq u \subseteq \mathcal{S}} q^{|u|} C_{s,u}^\delta \\ & \times \sum_{\substack{\boldsymbol{\tau}_{\mathcal{S}\setminus u} \in \{0,\ldots,\delta-1\}^{s-|u|} \\ \gamma_{u\cup v(\boldsymbol{\tau}_{\mathcal{S}\setminus u})} \neq 0}} \hat{C}^{\sum_{j\in\mathcal{S}\setminus u} \tau_j} V_{\lambda,1}^{(|u|)}(f^{(\boldsymbol{\delta}_u, \boldsymbol{\tau}_{\mathcal{S}\setminus u})}(\cdot, \mathbf{0}_{\mathcal{S}\setminus u})), \end{aligned}$$

where $\hat{C} = 1$ for $2 \le q < 6$ and $\hat{C} = (2 - 2\cos(2\pi/q))^{-1/2}$ for $q > 6$ (note that for $q > 6$ we have $\hat{C} > 1$) and

$$C_{s,u} = 3^{|u|}(2 - 2\cos(2\pi/q))^{-|u|/2}(3/2 + (2 - 2\cos(2\pi/q))^{-1/2})^{s-|u|}.$$

As noted above, under certain conditions we can write $V_{\lambda,1}^{(|u|)}$ also as an integral; see (3.4).

We give a further useful estimation of the constant by separating the dependence of the function from the constants. This way we obtain

$$C_{f,q,s,\boldsymbol{\gamma}} \le C_{\delta,q,s,\boldsymbol{\gamma}} N_{\delta,\lambda,\boldsymbol{\gamma}}(f),$$

where

$$\begin{aligned} C_{\delta,q,s,\boldsymbol{\gamma}} = {} & \sum_{\boldsymbol{\tau} \in \{0,\ldots,\delta-1\}^s} \gamma_{v(\boldsymbol{\tau})} \hat{C}^{\tau_1+\cdots+\tau_s} \\ & + \sum_{\emptyset \neq u \subseteq \mathcal{S}} q^{|u|} C_{s,u}^\delta \sum_{\boldsymbol{\tau}_{\mathcal{S}\setminus u} \in \{0,\ldots,\delta-1\}^{s-|u|}} \gamma_{u\cup v(\boldsymbol{\tau}_{\mathcal{S}\setminus u})} \hat{C}^{\sum_{j\in\mathcal{S}\setminus u} \tau_j}. \end{aligned} \tag{3.17}$$

Consider now the case where the weights are of product form (see [7]); i.e., there is a sequence of positive real numbers $(\gamma_j)_{j\in\mathbb{N}}$ such that $\gamma_v = \prod_{j\in v}\gamma_j$ for all $v \subset \mathbb{N}$ and for $v = \emptyset$ we set $\gamma_v = 1$. If q is now prime with $2 \le q < 6$, then

$$C_{\delta,q,s,\boldsymbol{\gamma}} = \prod_{j=1}^{s}(1+\gamma_j(\delta-1)) + \prod_{j=1}^{s}\left[(1+\gamma_j(\delta-1))(3/2+\hat{C}) + \gamma_j 3q\hat{C}\right]$$
$$- \prod_{j=1}^{s}\left[(1+\gamma_j(\delta-1))(3/2+\hat{C})\right].$$

For example, for $q = 2$, $\delta = 1$ and product weights we obtain

$$C_{1,2,s,\boldsymbol{\gamma}} = 1 - 2^s + 2^s\prod_{j=1}^{s}(1+6\gamma_j).$$

The approach used here for prime q and φ the identity map can also be used for arbitrary prime powers q and arbitrary mappings φ with $\varphi(0) = 0$. Hence we obtain the following corollary.

COROLLARY 3.9. *Under the assumptions of Theorem* 3.8 *there exists a constant* $C_{\delta,q,s,\boldsymbol{\gamma}} > 0$ *independent of* $\boldsymbol{k}$ *and* f *such that*

$$|\hat{f}(\boldsymbol{k})| \le C_{\delta,q,s,\boldsymbol{\gamma}} N_{\delta,\lambda,\boldsymbol{\gamma}}(f) q^{-\mu_{q,\delta+\lambda}(\boldsymbol{k})} \quad \text{for all } \boldsymbol{k} \in \mathbb{N}_0^s,$$

where $\mu_{q,\delta+\lambda}$ *is given by* (3.14) *and* (3.15).

Remark 3.10. The results in this section also hold for the following generalization. In the definition of $N_{\delta,\lambda,\boldsymbol{\gamma}}$ we anchored the function and its derivatives at 0, i.e., we used $V^{(|u(\boldsymbol{\tau})|)}_{\lambda,\mathfrak{p},\mathfrak{q},\mathbf{1}}(f(\cdot,\mathbf{0}_{\mathcal{S}\setminus u(\boldsymbol{\tau})}))$. This can be generalized by choosing an arbitrary $\boldsymbol{a} \in [0,1]^s$ and using $V^{(|u(\boldsymbol{\tau})|)}_{\lambda,\mathfrak{p},\mathfrak{q},\mathbf{1}}(f(\cdot,\boldsymbol{a}_{S\setminus u(\boldsymbol{\tau})}))$ in the definition of $N_{\delta,\lambda,\boldsymbol{\gamma}}$. It can be shown that in this case we also have Theorem 3.8 and Corollary 3.9.

3.3. Convergence of the Walsh series. For our purposes here we need strong assumptions on the convergence of the Walsh series $S(f)(\boldsymbol{x}) = \sum_{\boldsymbol{k}\in\mathbb{N}_0^s}\hat{f}(\boldsymbol{k})\mathrm{wal}_{\boldsymbol{k}}(\boldsymbol{x})$ to the function f, i.e., we require that the partial series $S_m(f)(\boldsymbol{x}) = \sum_{\substack{\boldsymbol{k}\in\mathbb{N}_0^s\\ k_j<m}}\hat{f}(\boldsymbol{k})\mathrm{wal}_{\boldsymbol{k}}(\boldsymbol{x})$ converge to $f(\boldsymbol{x})$ at every point $\boldsymbol{x} \in [0,1)^s$ as $m \to \infty$. (Note that the Walsh series $S(f)$ for the functions considered in this paper is always absolutely convergent, i.e., $\sum_{\boldsymbol{k}\in\mathbb{N}_0^s}|\hat{f}(\boldsymbol{k})| < \infty$; hence the Walsh series $S(f)(\boldsymbol{x})$ is uniformly bounded by $\sum_{\boldsymbol{k}\in\mathbb{N}_0^s}|\hat{f}(\boldsymbol{k})|$ and therefore $S(f)(\boldsymbol{x})$ itself converges at every point $\boldsymbol{x} \in [0,1)^s$.) This is necessary as we want to approximate the integral at function values $\boldsymbol{x}_n$, and for our analysis we deal with the Walsh series rather than the function itself; hence it is paramount that the function and its Walsh series coincide at every point $\boldsymbol{x} \in [0,1)^s$.

As the functions considered here are at least differentiable it follows that they are continuous, and using the argument in [10, p. 373] it follows that the Walsh series really converges at every point $\boldsymbol{x} \in [0,1)^s$ to the function value $f(\boldsymbol{x})$. Indeed, for a given $\boldsymbol{x} \in [0,1)^s$ we have

$$S_{q^m}(f)(\boldsymbol{x}) = \sum_{\boldsymbol{k}\in\{0,\dots,q^m-1\}^s}\hat{f}(\boldsymbol{k})\mathrm{wal}_{\boldsymbol{k}}(\boldsymbol{x}) = \mathrm{Vol}(J_{\boldsymbol{x}})^{-1}\int_{J_{\boldsymbol{x}}} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x},$$

where $J_{\boldsymbol{x}} = \prod_{j=1}^{s}[q^{-m}\lfloor q^m x_j\rfloor, q^{-m}\lfloor q^m x_j\rfloor + q^{-m})$. The last equality follows from

$$\sum_{\boldsymbol{k}\in\{0,\dots,q^m-1\}^s} \hat{f}(\boldsymbol{k})\mathrm{wal}_{\boldsymbol{k}}(\boldsymbol{x}) = \int_{[0,1)^s} f(\boldsymbol{y}) \sum_{\boldsymbol{k}\in\{0,\dots,q^m-1\}^s} \mathrm{wal}_{\boldsymbol{k}}(\boldsymbol{x})\overline{\mathrm{wal}_{\boldsymbol{k}}(\boldsymbol{y})}\,\mathrm{d}\boldsymbol{y}$$

$$= \mathrm{Vol}(J_{\boldsymbol{x}})^{-1} \int_{J_{\boldsymbol{x}}} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}.$$

As the function f is continuous, it immediately follows that $S_{q^m}(f)(\boldsymbol{x})$ converges to $f(\boldsymbol{x})$ as m goes to infinity, and the result follows.

3.4. A function space based on Walsh functions containing smooth functions. In this section we use the above results to define a function space based on Walsh functions which contains smooth functions for smoothness conditions considered in the previous section.

Let $\vartheta > 1$ be a real number and q a prime power. Then for $\boldsymbol{k} \in \mathbb{N}_0^s$ we set $r_{q,\vartheta}(\boldsymbol{k}) = q^{-\mu_{q,\vartheta}(\boldsymbol{k})}$, where $\mu_{q,\vartheta}$ is given by (3.14) and (3.15) (if ϑ is an integer, then choose $\lambda = 1$ and $\delta = \vartheta - 1$ and otherwise $\delta = \lfloor\vartheta\rfloor$ and $\lambda = \vartheta - \lfloor\vartheta\rfloor$).

Now we define a function space $\mathcal{E}_{s,q,\vartheta,\boldsymbol{\gamma}} \subseteq \mathcal{L}_2([0,1)^s)$ with norm $\|\cdot\|_{\mathcal{E}_{s,q,\vartheta,\boldsymbol{\gamma}}}$ given by

$$\|f\|_{\mathcal{E}_{s,q,\vartheta,\boldsymbol{\gamma}}} = \max_{\substack{u\subseteq\mathcal{S}\\ \gamma_u\neq 0}} \gamma_u^{-1} \sup_{\boldsymbol{k}_u\in\mathbb{N}^{|u|}} \frac{|\hat{f}(\boldsymbol{k}_u, \boldsymbol{0}_{\mathcal{S}\setminus u})|}{r_{q,\vartheta}(\boldsymbol{k}_u)},$$

where again for $\gamma_u = 0$ we assume that $\hat{f}(\boldsymbol{k}_u, \boldsymbol{0}_{\mathcal{S}\setminus u}) = 0$ for all $\boldsymbol{k}_u \in \mathbb{N}^{|u|}$.

The following result follows now directly from Corollary 3.9.

COROLLARY 3.11. *Let $\delta \geq 1$, let $0 < \lambda \leq 1$, let $\mathfrak{p}, \mathfrak{q}, \mathfrak{r} \geq 1$, and let an indexed set $\boldsymbol{\gamma} = (\gamma_v)_{v\subset\mathbb{N}}$ of nonnegative real numbers be given. Then there exists a constant $C_{\delta,q,s,\boldsymbol{\gamma}} > 0$ such that for every function $f : [0,1)^s \to \mathbb{R}$, whose partial mixed derivatives up to order δ exist, we have*

$$\|f\|_{\mathcal{E}_{s,q,\delta+\lambda,\boldsymbol{\gamma}}} \leq C_{\delta,q,s,\boldsymbol{\gamma}} N_{\delta,\lambda,\boldsymbol{\gamma}}(f),$$

where $\mu_{q,\delta+\lambda}$ is given by (3.14) *and* (3.15).

Again, using (3.17), an explicit constant in Corollary 3.11 can be obtained for q prime and φ the identity map. For all other cases (i.e., arbitrary prime powers q and mappings φ with $\varphi(0) = 0$) explicit constants can be obtained as well, but in this case the constant may also depend on the particular choice of q and φ. Further, as noted above, for $\lambda = 1$ and $\mathfrak{p} = \mathfrak{q} = \mathfrak{r} = 2$ the functional $N_{\delta,\lambda,\boldsymbol{\gamma}}$ coincides with the norm in a certain Sobolev space (the one-dimensional inner product for this Sobolev space is given by (3.13), and for higher dimensions one just considers tensor products of the one-dimensional space), and hence it follows that $\mathcal{E}_{s,q,\delta+1,\boldsymbol{\gamma}}$ contains certain Sobolev spaces. Hence Corollary 3.11 shows that if we want to prove results for smooth functions, it is enough to consider $\mathcal{E}_{s,q,\vartheta,\boldsymbol{\gamma}}$. (In the following we design quasi–Monte Carlo rules which work well for $\mathcal{E}_{s,q,\vartheta,\boldsymbol{\gamma}}$ rather than directly for smooth functions, so the results for smooth functions come as a byproduct.)

A function $f \in \mathcal{E}_{s,q,\vartheta,\boldsymbol{\gamma}}$ can be written as a sum of their anova terms $f = \sum_{u\subseteq\mathcal{S}} f_u$ (see [8]). For a function $f \in \mathcal{E}_{s,q,\vartheta,\boldsymbol{\gamma}}$ given by $f(\boldsymbol{x}) = \sum_{\boldsymbol{k}\in\mathbb{N}_0^s} \hat{f}(\boldsymbol{k})\mathrm{wal}_{\boldsymbol{k}}(\boldsymbol{x})$ the anova term f_u corresponding to a subset $u \subseteq \mathcal{S}$ is simply given by

$$f_u(\boldsymbol{x}_u) = \sum_{\boldsymbol{k}_u\in\mathbb{N}^{|u|}} \hat{f}(\boldsymbol{k}_u, \boldsymbol{0}_{\mathcal{S}\setminus u})\mathrm{wal}_{\boldsymbol{k}_u}(\boldsymbol{x}_u).$$

If for some $u \subseteq \mathcal{S}$ we have $\gamma_u = 0$, then this implies that the anova term corresponding to u satisfies $f_u \equiv 0$. Hence the Walsh space $\mathcal{E}_{s,q,\vartheta,\boldsymbol{\gamma}}$ consists only of functions whose anova term belonging to a subset u is zero for all subsets u with $\gamma_u = 0$ (see also [7]).

4. Digital $(t, \alpha, \beta, n \times m, s)$-nets and digital $(t, \alpha, \beta, \sigma, s)$-sequences. In this section we give the definition of digital $(t, \alpha, \beta, n \times m, s)$-nets and digital $(t, \alpha, \beta, \sigma, s)$-sequences. Similar point sets were introduced in [2].

4.1. The digital construction scheme. The construction of the point set used here is a slight generalization of the digital construction scheme introduced by Niederreiter [20] by breaking with the tradition of having square generating matrices.

DEFINITION 4.1. *Let q be a prime power and let $n, m, s \geq 1$ be integers. Let $C_1, \dots, C_s$ be $n \times m$ matrices over the finite field $\mathbb{F}_q$ of order q. Now we construct q^m points in $[0,1)^s$: for $0 \leq h \leq q^m - 1$ let $h = h_0 + h_1 q + \cdots + h_{m-1} q^{m-1}$ be the q-adic expansion of h. Consider an arbitrary but fixed bijection $\varphi : \{0, 1, \dots, q-1\} \longrightarrow \mathbb{F}_q$. Identify h with the vector $\vec{h} = (\varphi(h_0), \dots, \varphi(h_{m-1}))^\top \in \mathbb{F}_q^m$, where $\top$ means the transpose of the vector (note that we write $\vec{h}$ for vectors in the finite field $\mathbb{F}_q^m$ and $\boldsymbol{h}$ for vectors of integers or real numbers). For $1 \leq j \leq s$ multiply the matrix C_j by $\vec{h}$, i.e.,*

$$C_j \vec{h} =: (y_{j,1}(h), \dots, y_{j,n}(h))^\top \in \mathbb{F}_q^n,$$

and set

$$x_{h,j} := \frac{\varphi^{-1}(y_{j,1}(h))}{q} + \cdots + \frac{\varphi^{-1}(y_{j,n}(h))}{q^n}.$$

The point set $\{\boldsymbol{x}_0, \dots, \boldsymbol{x}_{q^m-1}\}$ is called a digital net (over $\mathbb{F}_q$) (with generating matrices $C_1, \dots, C_s$).

For $n, m = \infty$ we obtain a sequence $\{\boldsymbol{x}_0, \boldsymbol{x}_1, \dots\}$, which is called a digital sequence (over $\mathbb{F}_q$) (with generating matrices $C_1, \dots, C_s$).

Niederreiter's concept of a digital (t, m, s)-net and a digital (t, s)-sequence will appear as a special case in the subsequent section. Further, the digital nets considered below all satisfy $n \geq m$.

For a digital net with generating matrices $C_1, \dots, C_s$ let $\mathcal{D} = \mathcal{D}(C_1, \dots, C_s)$ be the dual net given by

$$\mathcal{D} = \{\boldsymbol{k} \in \mathbb{N}_0^s \setminus \{\boldsymbol{0}\} : C_1^\top \vec{k}_1 + \cdots + C_s^\top \vec{k}_s = \vec{0}\},$$

where for $\boldsymbol{k} = (k_1, \dots, k_s)$ with $k_j = \kappa_{j,0} + \kappa_{j,1} q + \cdots$ and $\kappa_{j,i} \in \{0, \dots, q-1\}$ let $\vec{k}_j = (\varphi(\kappa_{j,0}), \dots, \varphi(\kappa_{j,n-1}))^\top$. Further, for $\emptyset \neq u \subseteq \mathcal{S}$ let $\mathcal{D}_u = \mathcal{D}((C_j)_{j \in u})$ and $\mathcal{D}_u^* = \mathcal{D}_u \cap \mathbb{N}^{|u|}$.

Note that throughout the paper Walsh functions and digital nets are defined using the same finite field $\mathbb{F}_q$ and the same bijection φ.

The following lemma is a slight generalization of [26, Lemma 2.5].

LEMMA 4.2. *Let $\{\boldsymbol{x}_0, \dots, \boldsymbol{x}_{q^m-1}\}$ be a digital net over $\mathbb{F}_q$ with bijection φ, where $\varphi(0) = 0$, generated by the $n \times m$ matrices $C_1, \dots, C_s$ over $\mathbb{F}_q$, $n, m \geq 1$. Then for any vector $\boldsymbol{k} = (k_1, \dots, k_s)$ of nonnegative integers $0 \leq k_1, \dots, k_s < q^n$ we have*

$$\sum_{h=0}^{q^m-1} {}_{\mathbb{F}_q, \varphi}\mathrm{wal}_{\boldsymbol{k}}(\boldsymbol{x}_h) = \begin{cases} q^m & \text{if } \boldsymbol{k} \in \mathcal{D} \cup \{\boldsymbol{0}\}, \\ 0 & \text{else}, \end{cases}$$

where $\boldsymbol{0}$ is the zero vector in $\mathbb{N}_0^s$.

4.2. $(t, \alpha, \beta, n \times m, s)$-nets and $(t, \alpha, \beta, \sigma, s)$-sequences. Digital (t, α, β, m, s)-nets and digital (t, α, β, s)-sequences were first introduced in [2]. Those point sets were used for quasi–Monte Carlo rules which achieve the optimal rate of convergence of the worst-case error in Korobov spaces (which are reproducing kernel Hilbert spaces of smooth periodic functions). By a slight generalization of digital (t, α, β, m, s)-nets we will show that those digital nets also achieve the optimal convergence of the worst-case error in the space $\mathcal{E}_{s,q,\vartheta,\boldsymbol{\gamma}}$ for all $1 < \vartheta \le \alpha$.

The t-value of a (t, m, s)-net is a quality parameter for the distribution properties of the net. A low t-value yields well-distributed point sets, and it has been shown (see, for example, [4, 20]) that a small t-value also guarantees a small worst-case error for integration in Sobolev spaces for which the partial first derivatives are square integrable. In [2] it was shown how the definition of the t-value needs to be modified in order to obtain faster convergence rates for periodic Sobolev spaces for which the partial derivatives up to order $\delta \le \beta$ are square integrable. Here we extend those result in several ways. First we generalize the digital (t, α, β, m, s)-nets used in [2] to digital $(t, \alpha, \beta, n \times m, s)$-nets and show that we then can remove the periodicity assumption necessary in [2]. Further, if the derivatives up to order δ also have bounded variation with coefficient $0 < \lambda \le 1$, then we have shown that such functions are in $\mathcal{E}_{s,q,\delta+\lambda,\boldsymbol{\gamma}}$.

In the following we repeat some definitions and results from [2] and give the definition of digital $(t, \alpha, \beta, n \times m, s)$-nets and digital $(t, \alpha, \beta, \sigma, s)$-sequences.

For a real number $\vartheta > 1$ the definition of the Walsh space $\mathcal{E}_{s,q,\vartheta,\boldsymbol{\gamma}}$ suggests that we define the metric $\mu_{q,\vartheta}(\boldsymbol{k}, \boldsymbol{l}) = \mu_{q,\vartheta}(\boldsymbol{k} \ominus \boldsymbol{l})$ on $\mathbb{N}_0^s$, where $\mu_{q,\vartheta}(\boldsymbol{k} \ominus \boldsymbol{l})$ is given by (3.14) and (3.15), which is an extension of the metric introduced in [19]; see also [28] (the metric for $\vartheta = 1$ can be used for Walsh spaces, for example, as considered in [4]; for this case one basically obtains the metric in [19, 28]). As we will see later, in order to obtain a small worst-case error in the Walsh space $\mathcal{E}_{s,q,\vartheta,\boldsymbol{\gamma}}$ we need digital nets for which $\min\{\mu_{q,\vartheta}(\boldsymbol{k}) : \boldsymbol{k} \in \mathcal{D}\}$ is large. By translating this property into a linear independence property of the row vectors of the generating matrices $C_1, \dots, C_s$ we arrive at the following definition.

DEFINITION 4.3. *Let $n, m, \alpha \ge 1$ be natural numbers, let $0 < \beta \le \alpha m/n$ be a real number, and let $0 \le t \le \beta n$ be a natural number. Let $\mathbb{F}_q$ be the finite field of prime power order q and let $C_1, \dots, C_s \in \mathbb{F}_q^{n \times m}$ with $C_j = (c_{j,1}, \dots, c_{j,n})^\top$. If for all $1 \le i_{j,\nu_j} < \cdots < i_{j,1} \le n$, where $0 \le \nu_j \le m$ for all $j = 1, \dots, s$, with*

$$\sum_{j=1}^{s} \sum_{l=1}^{\min(\nu_j, \alpha)} i_{j,l} \le \beta n - t$$

the vectors

$$c_{1,i_{1,\nu_1}}, \dots, c_{1,i_{1,1}}, \dots, c_{s,i_{s,\nu_s}}, \dots, c_{s,i_{s,1}}$$

are linearly independent over $\mathbb{F}_q$, then the digital net with generating matrices $C_1, \dots, C_s$ is called a digital $(t, \alpha, \beta, n \times m, s)$-net over $\mathbb{F}_q$. Further we call a digital $(t, \alpha, \beta, n \times m, s)$-net over $\mathbb{F}_q$ with the largest possible value of β, i.e., $\beta = \alpha m/n$, a digital $(t, \alpha, n \times m, s)$-net over $\mathbb{F}_q$.

If t is the smallest nonnegative integer such that the digital net generated by $C_1, \dots, C_s$ is a digital $(t, \alpha, \beta, n \times m, s)$-net, then we call the digital net a strict digital $(t, \alpha, \beta, n \times m, s)$-net or a strict digital $(t, \alpha, n \times m, s)$-net if $\beta = \alpha m/n$.

Remark 4.4. Using duality theory (see [22]) it follows that for a digital $(t, \alpha, \beta, n \times m, s)$-net we have $\min_{\boldsymbol{k} \in \mathcal{D}} \mu_{q,\alpha}(\boldsymbol{k}) > \beta n - t$ and for a strict digital $(t, \alpha, \beta, n \times m, s)$-net

we have $\min_{\boldsymbol{k}\in\mathcal{D}} \mu_{q,\alpha}(\boldsymbol{k}) = \beta n - t + 1$. Hence digital $(t, \alpha, \beta, n \times m, s)$-nets with high quality have a large value of $\beta n - t$.

Remark 4.5. In summary the parameters $t, \alpha, \beta, n, m, s$ have the following meaning:

- s denotes the dimension of the point set.
- n and m denote the size of the generating matrices for digital nets, i.e., the generating matrices are of size $n \times m$; in particular this means the point set has q^m points.
- t denotes the quality parameter of the point set; a low t-value means high quality. In the upper bound, t is a quality parameter related to the constant in the upper bound.
- β is also a quality parameter. We will see later that the integration error is roughly q^{-n}. This is of course true only within boundaries, which is the reason for the parameter β, i.e., the integration error is roughly $q^{-\beta n}$. Hence β is a quality parameter related to the convergence rate.
- α is the smoothness parameter of the point set.

We can group the parameters also in the following way:

- m, n, s are fixed parameters, i.e., they specify the number and size of the generating matrices.
- α is a variable parameter, i.e., given (fixed) generating matrices, we can, for example, generate a $(t_1, 1, \beta_1, 10 \times 5, 5)$-net, a $(t_2, 2, \beta_2, 10 \times 5, 5)$-net, and so on (note the point set is always the same in each instance; the values $t_1, t_2, \ldots, \beta_1, \beta_2, \ldots$ may differ). This is necessary as in the upper bounds α will be the smoothness of the integrand, which may not be known explicitly.
- t and β are dependent parameters; they will depend on the generating matrices and on α. For given generating matrices, it is desirable to know the values of β and t for each value of $\alpha \in \mathbb{N}$.

Digital $(t, \alpha, \beta, n \times m, s)$-nets do not exist for arbitrary choices of the parameters $t, \alpha, \beta, n, m, s$; see [2]. The digital nets considered in [2] had the restriction that $n = m$, and special attention was paid to those digital nets with high quality, i.e., where $\alpha = \beta$. In this paper, a special role will be played by those digital nets for which $n = \alpha m$ and $\beta = 1$. The restriction on the linear independence of the digital nets comprises now $n - t = \alpha m - t$ row vectors, which is the same as in [2], with the only difference being that the size of the generating matrices is now bigger since now each generating matrix has $n = \alpha m$ rows. Because those digital nets play a special role in this work, we have the following definition.

DEFINITION 4.6. *A digital $(t, \alpha, 1, \alpha m \times m, s)$-net over $\mathbb{F}_q$ is called a digital $(t, \alpha, \alpha m \times m, s)$-net over $\mathbb{F}_q$. A strict digital $(t, \alpha, 1, \alpha m \times m, s)$-net over $\mathbb{F}_q$ is called a strict digital $(t, \alpha, \alpha m \times m, s)$-net over $\mathbb{F}_q$.*

Remark 4.7. For practical purposes we would like to explicitly know digital $(t, \alpha, \alpha m \times m, s)$-nets for all $\alpha, m, s \geq 1$ with t as small as possible. (As will be shown later, they achieve the optimal rate of convergence of the integration error of integrands for which all mixed partial derivatives of order α are, for example, square integrable, thus their usefulness.)

Further, for given $\alpha, m, s \geq 1$ and a given digital $(t, \alpha, \alpha m \times m, s)$-net P, we would then also like to know the t'- and β'-values of this point set P when viewed as a digital $(t', \delta, \beta', \alpha m \times m, s)$-net for all values $\delta \in \mathbb{N}$, i.e., t' and β' are functions of δ. (This is because we would also like to know how well such a digital net P performs if the integrand has partial mixed derivatives of order up to δ, because we might not

know the smoothness of the integrand, but still would wish that P performs as well as possible.)

We can also define sequences of points for which the first q^m points form a digital $(t, \alpha, \beta, n \times m, s)$-net. In the classical case [20] one can just consider the left-upper $m \times m$ submatrices of the generating matrices of a digital sequence and determine the net properties of these for each $m \in \mathbb{N}$. Here, on the other hand, we are considering digital nets whose generating matrices are $n \times m$ matrices. So we would have to consider the left-upper $n_m \times m$ submatrices of the generating matrices of the digital sequence for each $m \in \mathbb{N}$ and where $(n_m)_{m\in\mathbb{N}}$ is a sequence of natural numbers. For our purposes here it is enough to consider only n_m of the form σm for some given $\sigma \in \mathbb{N}$.

DEFINITION 4.8. *Let $\alpha, \sigma \geq 1$ and $t \geq 0$ be integers and let $0 < \beta \leq \alpha/\sigma$ be a real number. Let $\mathbb{F}_q$ be the finite field of prime power order q and let $C_1, \ldots, C_s \in \mathbb{F}_q^{\infty\times\infty}$ with $C_j = (c_{j,1}, c_{j,2}, \ldots)^\top$. Further let $C_{j,\sigma m\times m}$ denote the left upper $\sigma m \times m$ submatrix of C_j. If for all $m > t/(\beta\sigma)$ the matrices $C_{1,\sigma m\times m}, \ldots, C_{s,\sigma m\times m}$ generate a digital $(t, \alpha, \beta, \sigma m \times m, s)$-net, then the digital sequence with generating matrices $C_1, \ldots, C_s$ is called a digital $(t, \alpha, \beta, \sigma, s)$-sequence over $\mathbb{F}_q$. Further we call a digital $(t, \alpha, 1, \alpha, s)$-sequence over $\mathbb{F}_q$ a digital (t, α, s)-sequence over $\mathbb{F}_q$.*

If t is the smallest nonnegative integer such that the digital sequence generated by $C_1, \ldots, C_s$ is a digital $(t, \alpha, \beta, \sigma, s)$-sequence, then we call the digital sequence a strict digital $(t, \alpha, \beta, \sigma, s)$-sequence or a strict digital (t, α, s)-sequence if $\alpha = \sigma$ and $\beta = 1$.

For brevity we will often write $(t, \alpha, \beta, n \times m, s)$-net instead of digital $(t, \alpha, \beta, n \times m, s)$-net over $\mathbb{F}_q$. The same applies to the other notions defined above.

Remark 4.9. Note that the definition of a digital $(t, 1, m\times m, s)$-net coincides with the definition of a digital (t, m, s)-net and the definition of a digital $(t, 1, s)$-sequence coincides with the definition of a digital (t, s)-sequence as defined by Niederreiter [20]. Further note that the t-value depends on α, β, and σ, i.e., $t = t(\alpha, \beta, \sigma)$ or $t = t(\alpha)$ if $\alpha = \sigma$ and $\beta = 1$.

The definition of (t, α, s)-sequences here differs slightly from the definition in [2]. Indeed the definition of a (t, α, s)-sequence in [2] corresponds to a $(t, \alpha, \alpha, 1, s)$-sequence in the terminology of this paper, whereas here we call a $(t, \alpha, 1, \alpha, s)$-sequence a (t, α, s)-sequence. On the other hand note that the condition of linear independence in Definition 4.3 is the same in both cases; i.e., the sum $i_{1,1} + \cdots + i_{1,\min(\nu_1,\alpha)} + \cdots + i_{s,1} + \cdots + i_{s,\min(\nu_s,\alpha)}$ needs to be bounded by $\alpha m - t$ for all m for $(t, \alpha, 1, \alpha, s)$-sequences and also for $(t, \alpha, \alpha, 1, s)$-sequences.

4.3. Some properties of $(t, \alpha, \beta, n \times m, s)$-nets and $(t, \alpha, \beta, \sigma, s)$-sequences. The properties of such digital nets and sequences shown in [2] also hold here. For example, it was shown there that a digital (t, α, m, s)-net is also a digital $(\lceil t\alpha'/\alpha\rceil, \alpha', m, s)$-net for all $1 \leq \alpha' \leq \alpha$ and every digital (t, α, s)-sequence is also a digital $(\lceil t\alpha'/\alpha\rceil, \alpha', s)$-sequence for all $1 \leq \alpha' \leq \alpha$. In the same way we have the following theorem.

THEOREM 4.10. *Let P be a digital $(t, \alpha, \beta, n \times m, s)$-net over $\mathbb{F}_q$ and let S be a digital $(t, \alpha, \beta, \sigma, s)$-sequence over $\mathbb{F}_q$. Then we have the following:*

(i) *P is a digital $(t', \alpha, \beta', n \times m, s)$-net for all $1 \leq \beta' \leq \beta$ and all $t \leq t' \leq \beta' m$, and S is a digital $(t', \alpha, \beta', \sigma, s)$-sequence for all $1 \leq \beta' \leq \beta$ and all $t \leq t'$.*

(ii) *P is a digital $(t', \alpha', \beta', n \times m, s)$-net for all $1 \leq \alpha' \leq n$ where $\beta' = \beta \min(\alpha, \alpha')/\alpha$ and $t' = \lceil t \min(\alpha, \alpha')/\alpha\rceil$, and S is a digital $(t', \alpha', \beta', \sigma, s)$-sequence for all $\alpha' \geq 1$ where $\beta' = \beta \min(\alpha, \alpha')/\alpha$ and where $t' = \lceil t \min(\alpha, \alpha')/\alpha\rceil$.*

(iii) *Any digital* $(t, \alpha, n \times m, s)$*-net is a digital* $(\lceil t\alpha'/\alpha \rceil, \alpha', n \times m, s)$*-net for all* $1 \le \alpha' \le \alpha$ *and every digital* (t, α, σ, s)*-sequence is a digital* $(\lceil t\alpha'/\alpha \rceil, \alpha', \sigma, s)$*-sequence for all* $1 \le \alpha' \le \alpha$.

(iv) *If* $C_1, \dots, C_s \in \mathbb{Z}_b^{n \times m}$ *are the generating matrices of a digital* $(t, \alpha, \beta, n \times m, s)$*-net, then the matrices* $C_1^{(n')}, \dots, C_s^{(n')}$*, where* $C_j^{(n')}$ *consists of the first* n' *rows of* C_j*, generate a digital* $(t, \alpha, \beta, n' \times m, s)$*-net for all* $1 \le n' \le n$.

(v) *Any digital* $(t, \alpha, \beta, \sigma, s)$*-sequence is a digital* $(t, \alpha, \beta, \sigma', s)$*-sequence for all* $1 \le \sigma' \le \sigma$.

4.4. Constructions of $(t, \alpha, \beta, n \times m, s)$-nets and (t, α, σ, s)-sequences. In this section we show how explicit examples of $(t, \alpha, \beta, n \times m, s)$-nets and $(t, \alpha, \beta, \sigma, s)$-sequences can be constructed. The idea for the construction is based on the construction method presented in [2].

Let $d \ge 1$ and let $C_1, \dots, C_{sd}$ be the generating matrices of a digital (t, m, sd)-net. Note that many explicit examples of such generating matrices are known; see, for example, [9, 20, 23, 31] and the references therein. For the construction of a (t, α, m, s)-net any of the above-mentioned explicit constructions can be used, but as will be shown below the quality of the (t, α, m, s)-net obtained depends on the quality of the underlying digital (t, m, sd)-net on which our construction is based on.

Let $C_j = (c_{j,1}, \dots, c_{j,m})^\top$ for $j = 1, \dots, sd$; i.e., $c_{j,l}$ are the row vectors of C_j. Now let the matrix $C_j^{(d)}$ be made of the first rows of the matrices $C_{(j-1)d+1}, \dots, C_{jd}$, then the second rows of $C_{(j-1)d+1}, \dots, C_{jd}$, and so on. The matrix $C_j^{(d)}$ is then a $dm \times m$ matrix; i.e., $C_j^{(d)} = (c_{j,1}^{(d)}, \dots, c_{j,dm}^{(d)})^\top$, where $c_{j,l}^{(d)} = c_{u,v}$ with $l = (v-j)d + u$, $1 \le v \le m$, and $(j-1)d < u \le jd$ for $l = 1, \dots, dm$ and $j = 1, \dots, s$. The following result is a slight generalization of [2, Theorem 3] and can be obtained using the same proof technique.

THEOREM 4.11. *Let* $d \ge 1$ *be a natural number and let* $C_1, \dots, C_{sd}$ *be the generating matrices of a digital* (t', m, sd)*-net over the finite field* $\mathbb{F}_q$ *of prime power order* q*. Let* $C_1^{(d)}, \dots, C_s^{(d)}$ *be defined as above. Then for any* $\alpha \ge 1$ *the matrices* $C_1^{(d)}, \dots, C_s^{(d)}$ *are generating matrices of a digital* $(t, \alpha, \min(1, \alpha/d), dm \times m, s)$*-net over* $\mathbb{F}_q$ *with*

$$t = \min(\alpha, d)\, t' + \left\lceil \frac{s(d-1)\min(\alpha, d)}{2} \right\rceil.$$

The above construction and Theorem 4.11 can easily be extended to $(t, \alpha, \beta, \sigma, s)$-sequences. Indeed, let $d \ge 1$ and let $C_1, \dots, C_{sd}$ be the generating matrices of a digital (t, sd)-sequence. Again many explicit generating matrices are known; see, for example, [9, 20, 23, 31]. Let $C_j = (c_{j,1}, c_{j,2}, \dots)^\top$ for $j = 1, \dots, sd$; i.e., $c_{j,l}$ are the row vectors of C_j. Now let the matrix $C_j^{(d)}$ be made of the first rows of the matrices $C_{(j-1)d+1}, \dots, C_{jd}$, then the second rows of $C_{(j-1)d+1}, \dots, C_{jd}$, and so on, i.e.,

$$C_j^{(d)} = (c_{(j-1)d+1,1}, \dots, c_{jd,1}, c_{(j-1)d+1,2}, \dots, c_{jd,2}, \dots)^\top.$$

The following theorem states that the matrices $C_1^{(d)}, \dots, C_s^{(d)}$ are the generating matrices of a digital $(t, \alpha, \min(1, \alpha/d), d, s)$-sequence; cf. [2, Theorem 4].

THEOREM 4.12. *Let* $d \ge 1$ *be a natural number and let* $C_1, \dots, C_{sd}$ *be the generating matrices of a digital* (t', sd)*-sequence over the finite field* $\mathbb{F}_q$ *of prime power order* q*. Let* $C_1^{(d)}, \dots, C_s^{(d)}$ *be defined as above. Then for any* $\alpha \ge 1$ *the matrices* $C_1^{(d)}, \dots, C_s^{(d)}$ *are generating matrices of a digital* $(t, \alpha, \min(1, \alpha/d), d, s)$*-sequence over* $\mathbb{F}_q$ *with*

$$t = \min(\alpha, d)\, t' + \left\lceil \frac{s(d-1)\min(\alpha, d)}{2} \right\rceil.$$

The last result shows that $(t, \alpha, \beta, \sigma m \times m, s)$-nets indeed exist for $\beta = 1$ and any $0 < \sigma \leq \alpha$ and for m arbitrarily large. We have even shown that digital $(t, \alpha, \beta, \alpha m \times m, s)$-nets exist which are extensible in m and s. This can be achieved by using an underlying (t', sd)-sequence which is itself extensible in m and s. If the t'-value of the original (t', m, s)-net or (t', s)-sequence is known explicitly, then we also know the t-value of the digital $(t, \alpha, \beta, \alpha m \times m, s)$-net or $(t, \alpha, \beta, \sigma, s)$-sequence. Furthermore it has also been shown how such digital nets can be constructed in practice. Further results on such sequences are established in [2].

5. Numerical integration in the Walsh space $\mathcal{E}_{s,q,\vartheta,\boldsymbol{\gamma}}$. In this section we investigate numerical integration in the Walsh space $\mathcal{E}_{s,q,\vartheta,\boldsymbol{\gamma}}$ using quasi–Monte Carlo rules,

$$Q_{q^m,s}(f) = \frac{1}{q^m} \sum_{n=0}^{q^m-1} f(\boldsymbol{x}_n),$$

where $\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{q^m-1}$ are the points of a digital (t, α, β, m, s)-net over $\mathbb{F}_q$. More precisely, we want to approximate the integral

$$I_s(f) = \int_{[0,1]^s} f(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}$$

by the quasi–Monte Carlo rule $Q_{q^m,s}(f)$. As a quality measure for our rule we introduce the worst-case error in the next section.

5.1. The worst-case error in the Walsh space $\mathcal{E}_{s,q,\vartheta,\boldsymbol{\gamma}}$. The worst-case error for the Walsh space $\mathcal{E}_{s,q,\vartheta,\boldsymbol{\gamma}}$ using the quasi–Monte Carlo rule $Q_{q^m,s}$ is given by

$$e(Q_{q^m,s}, \mathcal{E}_{s,q,\vartheta,\boldsymbol{\gamma}}) = \sup_{\substack{f \in \mathcal{E}_{s,q,\vartheta,\boldsymbol{\gamma}} \\ \|f\|_{\mathcal{E}_{s,q,\vartheta,\boldsymbol{\gamma}}} \leq 1}} |I_s(f) - Q_{q^m,s}(f)|.$$

The initial error is given by

$$e(Q_{0,s}, \mathcal{E}_{s,q,\vartheta,\boldsymbol{\gamma}}) = \sup_{\substack{f \in \mathcal{E}_{s,q,\vartheta,\boldsymbol{\gamma}} \\ \|f\|_{\mathcal{E}_{s,q,\vartheta,\boldsymbol{\gamma}}} \leq 1}} |I_s(f)|.$$

In the following we use digital nets generated by the matrices $C_1, \ldots, C_s$ as quadrature points for the quadrature rule $Q_{q^m,s}$. Let $f \in \mathcal{E}_{s,q,\vartheta,\boldsymbol{\gamma}}$. Using Lemma 4.2 it follows that

$$\begin{aligned} |I_s(f) - Q_{q^m,s}(f)| &= \left| \sum_{\boldsymbol{k} \in \mathcal{D}} \hat{f}(\boldsymbol{k}) \right| \\ &\leq \sum_{\boldsymbol{k} \in \mathcal{D}} |\hat{f}(\boldsymbol{k})| = \sum_{\emptyset \neq u \subseteq \mathcal{S}} \sum_{\boldsymbol{k}_u \in \mathcal{D}_u^*} |\hat{f}(\boldsymbol{k}_u, \boldsymbol{0}_{\mathcal{S}\setminus u})|. \end{aligned}$$

Now we have $|\hat{f}(\boldsymbol{k}_u, \boldsymbol{0}_{\mathcal{S}\setminus u})| \leq \gamma_u r_{q,\vartheta}(\boldsymbol{k}_u) \|f\|_{\mathcal{E}_{s,q,\vartheta,\boldsymbol{\gamma}}}$ and thus we obtain

$$|I_s(f) - Q_{q^m,s}(f)| \leq \|f\|_{\mathcal{E}_{s,q,\vartheta,\boldsymbol{\gamma}}} \sum_{\emptyset \neq u \subseteq \mathcal{S}} \gamma_u \sum_{\boldsymbol{k}_u \in \mathcal{D}_u^*} r_{q,\vartheta}(\boldsymbol{k}_u). \tag{5.1}$$

By choosing $\hat{f}(\boldsymbol{k}_u, \mathbf{0}_{\mathcal{S}\setminus u}) = \gamma_u r_{q,\vartheta}(\boldsymbol{k}_u)$ for all u and $\boldsymbol{k}_u$ we can also obtain equality in (5.1). Thus we have

$$e(Q_{q^m,s}, \mathcal{E}_{s,q,\vartheta,\boldsymbol{\gamma}}) = \sum_{\emptyset\neq u\subseteq\mathcal{S}} \gamma_u \sum_{\boldsymbol{k}_u\in\mathcal{D}_u^*} r_{q,\vartheta}(\boldsymbol{k}_u). \tag{5.2}$$

From the last formula we can now see that essentially a large value of $\min\{\mu_{q,\vartheta}(\boldsymbol{k}) : \boldsymbol{k}\in\mathcal{D}\}$ guarantees a small worst-case error. Further it can be shown that

$$e(Q_{0,s}, \mathcal{E}_{s,q,\vartheta,\boldsymbol{\gamma}}) = \gamma_\emptyset. \tag{5.3}$$

We have shown the following theorem.

THEOREM 5.1. *The initial error for multivariate integration in the Walsh space $\mathcal{E}_{s,q,\vartheta,\boldsymbol{\gamma}}$ is given by* (5.3)*, and the worst-case error for multivariate integration in the Walsh space $\mathcal{E}_{s,q,\vartheta,\boldsymbol{\gamma}}$ using a digital net as quadrature points is given by* (5.2).

In the following lemma we establish an upper bound on the sum $\sum_{\boldsymbol{k}_u\in\mathcal{D}_u^*} r_{q,\vartheta}(\boldsymbol{k}_u)$ for digital $(t,\alpha,\beta,n\times m,s)$-nets over $\mathbb{F}_q$. The proof is similar to that of [2, Lemma 6].

LEMMA 5.2. *Let $\vartheta > 1$ be a real number, let $q \geq 2$ be a prime power, let $C_1,\ldots,C_s \in \mathbb{F}_q^{n\times m}$ be the generating matrices of a digital $(t,\lceil\vartheta\rceil,\beta,n\times m,s)$-net over $\mathbb{F}_q$ with $0<\beta\leq 1$, and let $\mathcal{D}_u^* = \mathcal{D}_u^*((C_j)_{j\in u})$. For all $\emptyset\neq u\subseteq\mathcal{S}$ we have that if ϑ is not an integer, it follows that*

$$\sum_{\boldsymbol{k}_u\in\mathcal{D}_u^*} r_{q,\vartheta}(\boldsymbol{k}_u) \leq C_{|u|,q,\vartheta}(\beta n - t + \lceil\vartheta\rceil)^{|u|\lceil\vartheta\rceil-1} q^{-\vartheta\lfloor(\beta n-t)/\lceil\vartheta\rceil\rfloor},$$

where

$$C_{|u|,q,\vartheta} = q^{|u|\lceil\vartheta\rceil}((q-q^{\vartheta-\lfloor\vartheta\rfloor})^{-1} + (1-q^{(1-\vartheta)/\lceil\vartheta\rceil})^{-|u|\lceil\vartheta\rceil}),$$

and if ϑ is an integer, it follows that

$$\sum_{\boldsymbol{k}_u\in\mathcal{D}_u^*} r_{q,\vartheta}(\boldsymbol{k}_u) \leq C'_{|u|,q,\vartheta}(\beta n - t + \vartheta)^{|u|\vartheta} q^{-(\beta n-t)},$$

where

$$C'_{|u|,q,\vartheta} = q^{|u|\vartheta}(q^{-1} + (1-q^{1/\vartheta-1})^{-|u|\vartheta}).$$

Proof. To simplify the notation we prove the result only for $u=\mathcal{S}$. For all other subsets the result follows by the same arguments.

We first consider the case where $\vartheta>1$ is not an integer. We partition the set $\mathcal{D}_{\mathcal{S}}^*$ into parts where the highest digits of k_j are prescribed, and we count the number of solutions of $C_1^\top\vec{k}_1+\cdots+C_s^\top\vec{k}_s=\vec{0}$. For $j=1,\ldots,s$ now let $i_{j,\lceil\vartheta\rceil}<\cdots<i_{j,1}$ with $i_{j,1}\geq 1$. Note that we now allow $i_{j,l}<1$, in which case the contributions of those $i_{j,l}$ are to be ignored. This notation is adopted in order to avoid considering many special cases. Further we write $\boldsymbol{i}_{s,\lceil\vartheta\rceil} = (i_{1,1},\ldots,i_{1,\lceil\vartheta\rceil},\ldots,i_{s,1},\ldots,i_{s,\lceil\vartheta\rceil})$ and define

$$\begin{aligned}\mathcal{D}_{\mathcal{S}}^*(\boldsymbol{i}_{s,\lceil\vartheta\rceil}) = \{\boldsymbol{k}\in\mathcal{D}_{\mathcal{S}}^* : k_j = \lfloor \kappa_{j,1}q^{i_{j,1}-1}+\cdots+\kappa_{j,\lceil\vartheta\rceil}q^{i_{j,\lceil\vartheta\rceil}-1}+l_j\rfloor \\ \text{with } 0\leq l_j<q^{i_{j,\lceil\vartheta\rceil}-1}\text{and } 1\leq\kappa_{j,l}<q \text{ for } j=1,\ldots,s\},\end{aligned}$$

where $\lfloor\cdot\rfloor$ just means that the contributions of $i_{j,l}<1$ are to be ignored. Let $\mu(\boldsymbol{i}_{s,\lceil\vartheta\rceil}) = i_{1,1}+\cdots+i_{1,\lceil\vartheta\rceil-1}+(\vartheta-\lfloor\vartheta\rfloor)i_{1,\lceil\vartheta\rceil}+\cdots+i_{s,1}+\cdots+i_{s,\lceil\vartheta\rceil-1}+(\vartheta-\lfloor\vartheta\rfloor)i_{s,\lceil\vartheta\rceil}$.

Then we have

$$\sum_{\boldsymbol{k}_{\mathcal{S}}\in\mathcal{D}^*_{\mathcal{S}}} r_{q,\vartheta}(\boldsymbol{k}_{\mathcal{S}}) = \sum_{i_{1,1}=1}^{\infty}\cdots\sum_{i_{1,\lceil\vartheta\rceil}=1}^{i_{1,\lceil\vartheta\rceil-1}-1}\cdots\sum_{i_{s,1}=1}^{\infty}\cdots\sum_{i_{s,\lceil\vartheta\rceil}=1}^{i_{s,\lceil\vartheta\rceil-1}-1}\frac{|\mathcal{D}^*_{\mathcal{S}}(\boldsymbol{i}_{s,\lceil\vartheta\rceil})|}{q^{\mu(\boldsymbol{i}_{s,\lceil\vartheta\rceil})}}. \tag{5.4}$$

Some of the sums above can be empty, in which case we just set the corresponding summation index $i_{j,l}=0$.

Note that by the $(t,\lceil\vartheta\rceil,\beta,n\times m,s)$-net property we have that $|\mathcal{D}^*_{\mathcal{S}}(\boldsymbol{i}_{s,\lceil\vartheta\rceil})|=0$ as long as $i_{1,1}+\cdots+i_{1,\lceil\vartheta\rceil}+\cdots+i_{s,1}+\cdots+i_{s,\lceil\vartheta\rceil}\le\beta n-t$. Hence now let $0\le i_{1,1},\dots,i_{s,\lceil\vartheta\rceil}$ be given such that $i_{1,1},\dots,i_{s,1}\ge 1$, $i_{j,\lceil\vartheta\rceil}<\cdots<i_{j,1}$ for $j=1,\dots,s$ and where if $i_{j,l}<1$, we set $i_{j,l}=0$ (in which case we also have $i_{j,l+1}=i_{j,l+2}=\cdots=0$ and the inequalities $i_{j,l}>\cdots>i_{j,\lceil\vartheta\rceil}$ are ignored) and $i_{1,1}+\cdots+i_{1,\lceil\vartheta\rceil}+\cdots+i_{s,1}+\cdots+i_{s,\lceil\vartheta\rceil}>\beta n-t$. We now need to estimate $|\mathcal{D}^*_{\mathcal{S}}(\boldsymbol{i}_{s,\lceil\vartheta\rceil})|$, that is, we need to count the number of $\boldsymbol{k}\in\mathcal{D}^*_{\mathcal{S}}$ with $k_j=\lfloor\kappa_{j,1}b^{i_{j,1}-1}+\cdots+\kappa_{j,\lceil\vartheta\rceil}b^{i_{j,\lceil\vartheta\rceil}-1}+l_j\rfloor$.

There are at most $(q-1)^{\lceil\vartheta\rceil s}$ choices for $\kappa_{1,1},\dots,\kappa_{s,\lceil\vartheta\rceil}$ (we say "at most" because if $i_{j,l}<1$, then the corresponding $\kappa_{j,l}$ does not have any effect and therefore need not to be included).

Now let $1\le\kappa_{1,1},\dots,\kappa_{s,\lceil\vartheta\rceil}<q$ be given and define

$$\vec{g}=\kappa_{1,1}c^{\top}_{1,i_{1,1}}+\cdots+\kappa_{1,\lceil\vartheta\rceil}c^{\top}_{1,i_{1,\lceil\vartheta\rceil}}+\cdots+\kappa_{s,1}c^{\top}_{s,i_{s,1}}+\cdots+\kappa_{s,\lceil\vartheta\rceil}c^{\top}_{s,i_{s,\lceil\vartheta\rceil}},$$

where we set $c^{\top}_{j,l}=0$ if $l<1$ or $l>n$. Further let

$$B=(c^{\top}_{1,1},\dots,c^{\top}_{1,i_{1,\lceil\vartheta\rceil}-1},\dots,c^{\top}_{s,1},\dots,c^{\top}_{s,i_{s,\lceil\vartheta\rceil}-1}).$$

Now the task is to count the number of solutions $\vec{l}$ of $B\vec{l}=\vec{g}$.

As long as the columns of B are linearly independent the number of solutions can at most be 1. By the $(t,\lceil\vartheta\rceil,\beta,n\times m,s)$-net property this is certainly the case if (we write $(x)_+=\max(x,0)$)

$$\begin{aligned}(i_{1,\lceil\vartheta\rceil}-1)_+&+\cdots+(i_{1,\lceil\vartheta\rceil}-\lceil\vartheta\rceil)_+\\ &+(i_{s,\lceil\vartheta\rceil}-1)_+\cdots+(i_{s,\lceil\vartheta\rceil}-\lceil\vartheta\rceil)_+\le\lceil\vartheta\rceil(i_{1,\lceil\vartheta\rceil}+\cdots+i_{s,\lceil\vartheta\rceil})\\ &\le\beta n-t,\end{aligned}$$

that is, as long as

$$i_{1,\lceil\vartheta\rceil}+\cdots+i_{s,\lceil\vartheta\rceil}\le\frac{\beta n-t}{\lceil\vartheta\rceil}.$$

Now let $i_{1,\lceil\vartheta\rceil}+\cdots+i_{s,\lceil\vartheta\rceil}>\frac{\beta n-t}{\lceil\vartheta\rceil}$. Then by considering the rank of the matrix B and the dimension of the space of solutions of $B\vec{l}=\vec{0}$ it follows that the number of solutions of $B\vec{l}=\vec{g}$ is smaller than or equal to $q^{i_{1,\lceil\vartheta\rceil}+\cdots+i_{s,\lceil\vartheta\rceil}-\lfloor(\beta n-t)/\lceil\vartheta\rceil\rfloor}$. Thus we have

$$|\mathcal{D}^*_{\mathcal{S}}(\boldsymbol{i}_{s,\lceil\vartheta\rceil})|=0$$

if $\sum_{j=1}^{s}\sum_{l=1}^{\lceil\vartheta\rceil}i_{j,l}\le\beta n-t$, we have

$$|\mathcal{D}^*_{\mathcal{S}}(\boldsymbol{i}_{s,\lceil\vartheta\rceil})|=(q-1)^{s\lceil\vartheta\rceil}$$

if $\sum_{j=1}^{s}\sum_{l=1}^{\lceil\vartheta\rceil}i_{j,l}>\beta n-t$ and $\sum_{j=1}^{s}i_{j,\lceil\vartheta\rceil}\le\frac{\beta n-t}{\lceil\vartheta\rceil}$, and finally we have

$$|\mathcal{D}^*_{\mathcal{S}}(\boldsymbol{i}_{s,\lceil\vartheta\rceil})| \le (q-1)^{s\lceil\vartheta\rceil} q^{i_{1,\lceil\vartheta\rceil}+\cdots+i_{s,\lceil\vartheta\rceil}-\lfloor(\beta n-t)/\lceil\vartheta\rceil\rfloor}$$

if $\sum_{j=1}^{s}\sum_{l=1}^{\lceil\vartheta\rceil} i_{j,l} > \beta n - t$ and $\sum_{j=1}^{s} i_{j,\lceil\vartheta\rceil} > \frac{\beta n-t}{\lceil\vartheta\rceil}$.

We estimate the sum (5.4) now. Let S_1 be the sum in (5.4) where $i_{1,1}+\cdots+i_{s,\lceil\vartheta\rceil} > \beta n - t$ and $i_{1,\lceil\vartheta\rceil}+\cdots+i_{s,\lceil\vartheta\rceil} \le \frac{\beta n-t}{\lceil\vartheta\rceil}$. Let $l_1 = i_{1,1}+\cdots+i_{1,\lceil\vartheta\rceil-1}+\cdots+i_{s,1}+\cdots+i_{s,\lceil\vartheta\rceil-1}$ and let $l_2 = i_{1,\lceil\vartheta\rceil}+\cdots+i_{s,\lceil\vartheta\rceil}$. Let $A(l_1+l_2)$ denote the number of admissible choices of $i_{1,1},\ldots,i_{s,\lceil\vartheta\rceil}$ such that $l_1+l_2 = i_{1,1}+\cdots+i_{s,\lceil\vartheta\rceil}$. Then we have

$$S_1 = (q-1)^{s\lceil\vartheta\rceil} \sum_{l_2=0}^{\lfloor\frac{\beta n-t}{\lceil\vartheta\rceil}\rfloor} \frac{1}{q^{(\vartheta-\lfloor\vartheta\rfloor)l_2}} \sum_{l_1=\beta n-t+1-l_2}^{\infty} \frac{A(l_1+l_2)}{b^{l_1}}.$$

We have $A(l_1+l_2) \le \binom{l_1+l_2+s\lceil\vartheta\rceil-1}{s\lceil\vartheta\rceil-1}$ and hence we obtain

$$S_1 \le (q-1)^{s\lceil\vartheta\rceil} \sum_{l_2=0}^{\lfloor\frac{\beta n-t}{\lceil\vartheta\rceil}\rfloor} \frac{1}{q^{(\vartheta-\lfloor\vartheta\rfloor)l_2}} \sum_{l_1=\beta n-t+1-l_2}^{\infty} \frac{1}{q^{l_1}} \binom{l_1+l_2+s\lceil\vartheta\rceil-1}{s\lceil\vartheta\rceil-1}.$$

From a result by Matoušek [17, Lemma 2.18] (see also [4, Lemma 6]), we have

$$\begin{aligned}(q-1)^{s\lceil\vartheta\rceil} &\sum_{l_1=\beta n-t+1-l_2}^{\infty} \frac{1}{q^{l_1}} \binom{l_1+l_2+s\lceil\vartheta\rceil-1}{s\lceil\vartheta\rceil-1} \\ &\le q^{l_2-\beta n+t-1+s\lceil\vartheta\rceil} \binom{\beta n-t+s\lceil\vartheta\rceil}{s\lceil\vartheta\rceil-1}\end{aligned}$$

and further we have

$$\sum_{l_2=0}^{\lfloor\frac{\beta n-t}{\lceil\vartheta\rceil}\rfloor} \frac{q^{l_2}}{q^{(\vartheta-\lfloor\vartheta\rfloor)l_2}} = \sum_{l_2=0}^{\lfloor\frac{\beta n-t}{\lceil\vartheta\rceil}\rfloor} q^{l_2(\lceil\vartheta\rceil-\vartheta)} = \frac{q^{(\lceil\vartheta\rceil-\vartheta)(\lfloor(\beta n-t)/\lceil\vartheta\rceil\rfloor+1)}-1}{q^{\lceil\vartheta\rceil-\vartheta}-1}.$$

Thus we obtain

$$\begin{aligned}S_1 &\le \frac{q^{(\lceil\vartheta\rceil-\vartheta)(\lfloor(\beta n-t)/\lceil\vartheta\rceil\rfloor+1)}-1}{q^{\lceil\vartheta\rceil-\vartheta}-1} q^{-\beta n+t-1+s\lceil\vartheta\rceil} \binom{\beta n-t+s\lceil\vartheta\rceil}{s\lceil\vartheta\rceil-1} \\ &\le \frac{q^{s\lceil\vartheta\rceil-1}}{1-q^{\vartheta-\lceil\vartheta\rceil}} \binom{\beta n-t+s\lceil\vartheta\rceil}{s\lceil\vartheta\rceil-1} q^{-\vartheta\lfloor(\beta n-t)/\lceil\vartheta\rceil\rfloor}.\end{aligned}$$

Let S_2 be the part of (5.4) for which $i_{1,1}+\cdots+i_{s,\lceil\vartheta\rceil} > \beta n - t$ and $i_{1,\lceil\vartheta\rceil}+\cdots+i_{s,\lceil\vartheta\rceil} > \frac{\beta n-t}{\lceil\vartheta\rceil}$, i.e., we have

$$\begin{aligned}S_2 \le (q-1)^{s\lceil\vartheta\rceil} &\sum_{i_{1,1}=1}^{\infty} \cdots \sum_{i_{1,\lceil\vartheta\rceil}=1}^{i_{1,\lceil\vartheta\rceil-1}-1} \cdots \\ &\sum_{i_{s,1}=1}^{\infty} \cdots \sum_{i_{s,\lceil\vartheta\rceil}=1}^{i_{s,\lceil\vartheta\rceil-1}-1} \frac{q^{-\lfloor(\beta n-t)/\lceil\vartheta\rceil\rfloor} q^{(i_{1,\lceil\vartheta\rceil}+\cdots+i_{s,\lceil\vartheta\rceil})(\lceil\vartheta\rceil-\vartheta)}}{q^{i_{1,1}+\cdots+i_{1,\lceil\vartheta\rceil-1}+\cdots+i_{s,1}+\cdots+i_{s,\lceil\vartheta\rceil-1}}},\end{aligned}$$

where we have the additional conditions $i_{1,1}+\cdots+i_{s,\lceil\vartheta\rceil} > \beta n - t$ and $i_{1,\lceil\vartheta\rceil}+\cdots+i_{s,\lceil\vartheta\rceil} > \frac{\beta n-t}{\lceil\vartheta\rceil}$. As above let $l_1 = i_{1,1}+\cdots+i_{1,\lceil\vartheta\rceil-1}+\cdots+i_{s,1}+\cdots+i_{s,\lceil\vartheta\rceil-1}$ and

let $l_2 = i_{1,\lceil\vartheta\rceil} + \cdots + i_{s,\lceil\vartheta\rceil}$. Let $A(l_1+l_2)$ denote the number of admissible choices of $i_{1,1},\dots,i_{s,\lceil\vartheta\rceil}$ such that $l_1+l_2 = i_{1,1}+\cdots+i_{s,\lceil\vartheta\rceil}$. Note that $l_1 > \lfloor\vartheta\rfloor l_2$. Then we have $A(l_1+l_2) \le \binom{l_1+l_2+s\lceil\vartheta\rceil-1}{s\lceil\vartheta\rceil-1}$ and hence obtain

$$\begin{aligned}
S_2 &\le (q-1)^{s\lceil\vartheta\rceil} q^{-\lfloor(\beta n-t)/\lceil\vartheta\rceil\rfloor} \\
&\quad\times \sum_{l_2=\lfloor\frac{\beta n-t}{\lceil\vartheta\rceil}\rfloor+1}^{\infty} q^{(\lceil\vartheta\rceil-\vartheta)l_2} \sum_{l_1=\lfloor\vartheta\rfloor l_2+1}^{\infty} \frac{1}{q^{l_1}} \binom{l_1+l_2+s\lceil\vartheta\rceil-1}{s\lceil\vartheta\rceil-1} \\
&= (q-1)^{s\lceil\vartheta\rceil} q^{-\lfloor(\beta n-t)/\lceil\vartheta\rceil\rfloor} \\
&\quad\times \sum_{l_2=\lfloor\frac{\beta n-t}{\lceil\vartheta\rceil}\rfloor+1}^{\infty} \sum_{l_1=0}^{\infty} q^{-l_1+l_2-1-l_2\vartheta} \binom{l_1+l_2+\lfloor\vartheta\rfloor l_2-1+s\lceil\vartheta\rceil-1}{s\lceil\vartheta\rceil-1}.
\end{aligned}$$

By using again Matoušek [17, Lemma 2.18] (see also [4, Lemma 6]), we have

$$\begin{aligned}
&(q-1)^{s\lceil\vartheta\rceil} \sum_{l_1=0}^{\infty} q^{-l_1+l_2-1-l_2\vartheta} \binom{l_1+l_2+\lfloor\vartheta\rfloor l_2-1+s\lceil\vartheta\rceil-1}{s\lceil\vartheta\rceil-1} \\
&\qquad\qquad \le q^{s\lceil\vartheta\rceil} q^{l_2(1-\vartheta)-1} \binom{l_2\lceil\vartheta\rceil-1+s\lceil\vartheta\rceil-1}{s\lceil\vartheta\rceil-1}
\end{aligned}$$

and also

$$\begin{aligned}
&q^{s\lceil\vartheta\rceil-1-\lfloor(\beta n-t)/\lceil\vartheta\rceil\rfloor} \sum_{l_2=\lfloor\frac{\beta n-t}{\lceil\vartheta\rceil}\rfloor+1}^{\infty} q^{l_2(1-\vartheta)} \binom{l_2\lceil\vartheta\rceil-1+s\lceil\vartheta\rceil-1}{s\lceil\vartheta\rceil-1} \\
&\le q^{s\lceil\vartheta\rceil-1-\lfloor(\beta n-t)/\lceil\vartheta\rceil\rfloor} \sum_{l_2=\beta n-t}^{\infty} q^{l_2(1-\vartheta)/\lceil\vartheta\rceil} \binom{l_2+\lceil\vartheta\rceil-1+s\lceil\vartheta\rceil-1}{s\lceil\vartheta\rceil-1} \\
&\le q^{s\lceil\vartheta\rceil} (1-q^{(1-\vartheta)/\lceil\vartheta\rceil})^{-s\lceil\vartheta\rceil} \binom{\beta n-t+\lceil\vartheta\rceil-2+s\lceil\vartheta\rceil}{s\lceil\vartheta\rceil-1} q^{-\vartheta(\beta n-t)/\lceil\vartheta\rceil}.
\end{aligned}$$

Hence we have

$$S_2 \le q^{s\lceil\vartheta\rceil} (1-q^{(1-\vartheta)/\lceil\vartheta\rceil})^{-s\lceil\vartheta\rceil} \binom{\beta n-t+\lceil\vartheta\rceil-2+s\lceil\vartheta\rceil}{s\lceil\vartheta\rceil-1} q^{-\vartheta(\beta n-t)/\lceil\vartheta\rceil}.$$

Note that we have $\sum_{\boldsymbol{k}_S\in\mathcal{D}_S^*} r_{q,\vartheta}(\boldsymbol{k}_S) = S_1 + S_2$. Let $a \ge 1$ and $b \ge 0$ be integers; then we have

$$\binom{a+b}{b} = \prod_{i=1}^{b} \left(1+\frac{a}{i}\right) \le (1+a)^b.$$

Therefore we obtain

$$S_1 \le \frac{q^{s\lceil\vartheta\rceil-1}}{1-q^{\vartheta-\lceil\vartheta\rceil}} (\beta n-t+2)^{s\lceil\vartheta\rceil-1} q^{-\vartheta\lfloor(\beta n-t)/\lceil\vartheta\rceil\rfloor}$$

and

$$S_2 \le q^{s\lceil\vartheta\rceil} (1-q^{(1-\vartheta)/\lceil\vartheta\rceil})^{-s\lceil\vartheta\rceil} (\beta n-t+\lceil\vartheta\rceil)^{s\lceil\vartheta\rceil-1} q^{-\vartheta(\beta n-t)/\lceil\vartheta\rceil}.$$

Thus we have

$$\sum_{\boldsymbol{k}_{\mathcal{S}} \in \mathcal{D}^*_{\mathcal{S}}} r_{q,\vartheta}(\boldsymbol{k}_{\mathcal{S}}) \le C_{s,q,\vartheta}(\beta n - t + \lceil \vartheta \rceil)^{s\lceil \vartheta \rceil - 1} q^{-\vartheta \lfloor (\beta n - t)/\lceil \vartheta \rceil \rfloor},$$

where

$$C_{s,q,\vartheta} = q^{s\lceil \vartheta \rceil}((q - q^{\vartheta - \lfloor \vartheta \rfloor})^{-1} + (1 - q^{(1-\vartheta)/\lceil \vartheta \rceil})^{-s\lceil \vartheta \rceil}).$$

The result follows for the case $0 < \vartheta - \lfloor \vartheta \rfloor < 1$.

Now let $\vartheta > 1$ be an integer. Then using the same arguments as above it can be shown that

$$S_1 \le (\beta n - t + 2)^{s\vartheta} q^{-(\beta n - t) - 1 + s\vartheta}$$

and

$$S_2 \le q^{s\vartheta}(1 - q^{1/\vartheta - 1})^{-s\vartheta}(\beta n - t + \vartheta)^{s\vartheta - 1} q^{-(\beta n - t)}.$$

Thus we have

$$\sum_{\boldsymbol{k}_{\mathcal{S}} \in \mathcal{D}^*_{\mathcal{S}}} r_{q,\vartheta}(\boldsymbol{k}_{\mathcal{S}}) \le C'_{s,q,\vartheta}(\beta n - t + \vartheta)^{s\vartheta} q^{-(\beta n - t)},$$

where

$$C'_{s,q,\vartheta} = q^{s\vartheta}(q^{-1} + (1 - q^{1/\vartheta - 1})^{-s\vartheta}).$$

The result now follows. □

Remark 5.3. We note that the above lemma does not hold for $\beta > 1$ in general. Indeed, take, for example, $u = \{1\}$; then $\boldsymbol{k}_u = (k_1)$ and choose $k_1 = q^n$. Then the digit vector of the first n digits of q^n is $(0, \dots, 0)^\top$; therefore $C_1^\top \vec{k}_1 = \vec{0}$ and hence $\boldsymbol{k}_{(1)} \in \mathcal{D}^*_{(1)}$. Thus

$$\sum_{\boldsymbol{k}_{(1)} \in \mathcal{D}^*_{(1)}} r_{q,\vartheta}(\boldsymbol{k}_{(1)}) \ge q^{-n-1}$$

and hence a counterexample can be obtained for some choices of n, β, ϑ.

In [2] we did allow $\beta > 1$, but therein we had the additional assumption that the functions are periodic. In this case we were able to show that the Walsh coefficients $r_{q,\alpha}(\boldsymbol{k}, \boldsymbol{l}) = \prod_{j=1}^s r_{q,\alpha}(k_j, l_j)$ of the reproducing kernel also satisfy the additional property that $r_{q,\alpha}(q^m k_j, q^m k_j) = r_{q,\alpha}(q^m k_j) = q^{-2\alpha m} r_{q,\alpha}(k_j, k_j)$ for all $k_j, m \in \mathbb{N}$; see [2, Lemma 15]. Similarly, if we would also assume here that $r_{q,\vartheta}(q^n k) = q^{-\vartheta n} r_{q,\vartheta}(k)$ and $r_{q,\vartheta}(k)$ given as above if $q \not| k$, then the above counterexample would fail as then $r_{q,\vartheta}(q^n) = r_{q,\vartheta}(1q^n) = q^{-\vartheta(n+1)} r_{q,\vartheta}(1)$.

Using the above lemma we can now obtain an upper bound on the worst-case error.

THEOREM 5.4. *Let $\vartheta > 1$ be a real number and let $q \ge 2$ be a prime power. The worst-case error for multivariate integration in the Walsh space $\mathcal{E}_{s,q,\vartheta,\boldsymbol{\gamma}}$ using a digital $(t, \lceil \vartheta \rceil, \beta, n \times m, s)$-net over $\mathbb{F}_q$, with $0 < \beta \le 1$, as quadrature points is for nonintegers ϑ bounded by*

$$e(Q_{q^m,s}, \mathcal{E}_{s,q,\vartheta,\boldsymbol{\gamma}}) \le q^{-\vartheta \lfloor (\beta n - t)/\lceil \vartheta \rceil \rfloor} \sum_{\emptyset \ne u \subseteq \mathcal{S}} \gamma_u C_{|u|,q,\vartheta}(\beta n - t + \lceil \vartheta \rceil)^{|u|\lceil \vartheta \rceil - 1},$$

where

$$C_{|u|,q,\vartheta} = q^{|u|\lceil\vartheta\rceil}((q - q^{\vartheta-\lfloor\vartheta\rfloor})^{-1} + (1 - q^{(1-\vartheta)/\lceil\vartheta\rceil})^{-|u|\lceil\vartheta\rceil}),$$

and if ϑ is an integer, the worst-case error is bounded by

$$e(Q_{q^m,s}, \mathcal{E}_{s,q,\vartheta,\boldsymbol{\gamma}}) \le q^{-(\beta n - t)} \sum_{\emptyset \ne u \subseteq \mathcal{S}} \gamma_u C'_{|u|,q,\vartheta} (\beta n - t + \vartheta)^{|u|\vartheta},$$

where

$$C'_{|u|,q,\vartheta} = q^{|u|\vartheta}(q^{-1} + (1 - q^{1/\vartheta - 1})^{-|u|\vartheta}).$$

As a direct consequence of Corollary 3.11 we obtain the following result.

COROLLARY 5.5. *Let $\delta \ge 1$ be an integer, let $0 < \lambda \le 1$, and let $q \ge 2$ be a prime power. Then for any function $f : [0,1)^s \to \mathbb{R}$ whose partial mixed derivatives up to order δ exist it follows that the integration error using a digital $(t, \delta+1, \beta, n \times m, s)$-net over $\mathbb{F}_q$ with $0 < \beta \le 1$ as quadrature points is for $0 < \lambda < 1$ bounded by*

$$\begin{aligned} |I_s(f) - Q_{q^m,s}(f)| &\le q^{-(\delta+\lambda)\lfloor(\beta n - t)/(\delta+1)\rfloor} C_{\delta,s,q,\boldsymbol{\gamma}} N_{\delta,\lambda,\boldsymbol{\gamma}}(f) \\ &\quad \times \sum_{\emptyset \ne u \subseteq \mathcal{S}} \gamma_u C_{|u|,q,\delta+\lambda} (\beta n - t + \delta + 1)^{|u|(\delta+1)-1} \end{aligned}$$

and for $\lambda = 1$ the integration error is bounded by

$$\begin{aligned} |I_s(f) - Q_{q^m,s}(f)| &\le q^{-(\beta n - t)} C_{\delta,s,q,\boldsymbol{\gamma}} N_{\delta,\lambda,\boldsymbol{\gamma}}(f) \\ &\quad \times \sum_{\emptyset \ne u \subseteq \mathcal{S}} \gamma_u C'_{|u|,q,\delta+1} (\beta n - t + \delta + 1)^{|u|\delta+1}, \end{aligned}$$

where the constant $C_{\delta,s,q,\boldsymbol{\gamma}}$ is given in Corollary 3.11 and the constants $C_{|u|,q,\delta+\lambda}$ and $C'_{|u|,q,\delta+1}$ are given in Theorem 5.4.

Explicit constructions of digital $(t, \alpha, \min(1, \alpha/d), dm \times m, s)$-nets over $\mathbb{F}_q$ for all prime powers q and integers $\alpha, d, m, s > 1$ are given in section 4.4. By choosing $d = \alpha = \lceil\vartheta\rceil = \delta + 1$, by Theorem 5.4 and Corollary 5.5 we obtain a convergence of $\mathcal{O}(q^{-\vartheta m} m^{s\lceil\vartheta\rceil+1})$, which is optimal even for the smooth functions contained in the Walsh space $\mathcal{E}_{s,q,\vartheta,\boldsymbol{\gamma}}$; see [29], where a lower bound for smooth periodic functions was shown.

Remark 5.6. In [2, Remark 4] it was noted that if $m = n$ and $\beta > \alpha$, the t-value must grow with m and hence the restriction $\beta \le \alpha$ was added. A similar argument yields in our case that the t-value must grow with n if $\beta n > \alpha m$, as Theorem 5.4 shows a convergence of $\mathcal{O}(q^{-\beta n + t})$, but the best possible convergence rate is $q^{-\alpha m}$; hence the restriction $\beta \le \alpha m/n$ was added.

In case the smoothness of the function is not known, our constructions adjust themselves automatically up to a certain degree in the following way: for the construction of the digital net we choose some value of $d \ge 1$ and construct a digital $(t, \alpha, \min(1, \alpha/d), dm \times m, s)$-net or a digital $(t, \alpha, \min(1, \alpha/d), d, s)$-sequence for all $\alpha \ge 1$. The values $\delta \ge 1$ and $0 < \lambda \le 1$ determine the real smoothness of the function, which we now assume is not known. The value of α is the smoothness analogue for the digital net, i.e., we need to choose $\alpha = \delta + 1$. First assume that $\delta + \lambda \le d$; then $\min(1, \alpha/d) = (\delta + 1)/d$, and therefore we have $\beta = (\delta + 1)/d$. As $n = dm$ it follows

that $\beta n = (\delta+1)m$, and therefore Corollary 5.5 shows that we achieve a convergence of $\mathcal{O}(q^{-(\delta+\lambda)m} m^{s(\delta+1)+1})$, which is optimal. Now assume on the other hand that $\delta + \lambda > d$; then $\min(1, \alpha/d) = 1$ and therefore $\beta = 1$. Again we have $n = dm$ and hence $\beta n = dm$. In this case Corollary 5.5 shows that our construction achieves a convergence of $\mathcal{O}(q^{-dm} m^{s(\delta+1)+1})$.

Note that numerical integration of functions with less smoothness, i.e., for example functions with partial mixed derivatives up to degree 1 in $\mathcal{L}_2([0,1)^s)$ or functions with bounded variation, has been considered in many papers and monographs; see, for example, [3, 4, 6, 7, 14, 20, 30, 31]. Using the notation from above, basically those results are concerned with the case where $\delta = 0$ and $\lambda = 1$, and hence the results here are a direct continuation of what was previously known. The construction of digital nets proposed here for $d = 1$ yields obviously digital (t, m, s)-nets and (t, s)-sequences as, for example, defined in [20]. In view of Corollary 5.5 and the explanation which followed it is hence not surprising that the classical examples and theory (see, for example, [3, 4, 12, 13, 14, 20, 26, 31]) yielded only a convergence of $\mathcal{O}(q^{m(-1+\varepsilon)})$ for any $\varepsilon > 0$ (the ε here is used to hide the powers of m).

Note that the worst-case error in the Walsh space $\mathcal{E}_{s,q,\vartheta,\boldsymbol{\gamma}}$ is invariant with respect to a digital shift (see [4]); hence Corollary 5.5 also holds for digitally shifted digital nets. Thus, if one wants to use randomized digital nets, one can also use randomly digitally shifted digital nets. The root mean square worst-case error for this case would of course be bounded by the bound in Corollary 5.5, as this bound holds for any digital shift; i.e., our result here is even stronger in that we have shown that even for the worst digital shift we still have the bound of Corollary 5.5. From this, it follows that for our situation here, there is, in some sense, no bad digital shift. Other more sophisticated scrambling methods which do not destroy the essential properties of the point set can be used as well (for example, a digital shift of depth m; see [5, 17]); see [24] for some ideas in this direction.

REFERENCES

[1] H. E. Chrestenson, *A class of generalized Walsh functions*, Pacific J. Math., 5 (1955), pp. 17–31.

[2] J. Dick, *Explicit constructions of quasi-Monte Carlo rules for the numerical integration of high-dimensional periodic functions*, SIAM J. Numer. Anal., 45 (2007), pp. 2141–2176.

[3] J. Dick, F. Y. Kuo, F. Pillichshammer, and I. H. Sloan, *Construction algorithms for polynomial lattice rules for multivariate integration*, Math. Comp., 74 (2005), pp. 1895–1921.

[4] J. Dick and F. Pillichshammer, *Multivariate integration in weighted Hilbert spaces based on Walsh functions and weighted Sobolev spaces*, J. Complexity, 21 (2005), pp. 149–195.

[5] J. Dick and F. Pillichshammer, *On the mean square weighted L_2 discrepancy of randomized digital (t, m, s)-nets over $\mathbb{Z}_2$*, Acta Arith., 117 (2005), pp. 371–403.

[6] J. Dick, I. H. Sloan, X. Wang, and H. Woźniakowksi, *Liberating the weights*, J. Complexity, 20 (2004), pp. 593–623.

[7] J. Dick, I. H. Sloan, X. Wang, and H. Woźniakowski, *Good lattice rules in weighted Korobov spaces with general weights*, Numer. Math., 103 (2006), pp. 63–97.

[8] B. Efron and C. Stein, *The jackknife estimate of variance*, Ann. Statist., 9 (1981), pp. 586–598.

[9] H. Faure, *Discrèpances de suites associèes á un système de numèration (en dimension s)*, Acta Arith., 41 (1982), pp. 337–351.

[10] N. J. Fine, *On the Walsh functions*, Trans. Amer. Math. Soc., 65 (1949), pp. 372–414.

[11] S. Heinrich, F. J. Hickernell, and R. X. Yue, *Optimal quadrature for Haar wavelet spaces*, Math. Comp., 73 (2004), pp. 259–277.

[12] E. Hlawka, *Funktionen von beschränkter Variation in der Theorie der Gleichverteilung*, Ann. Mat. Pura Appl., 54 (1961), pp. 325–333.

[13] J. F. Koksma, *Een algemeene stelling uit de theorie der gelijkmatige verdeeling modulo* 1, Mathematica B (Zutphen), 11 (1942/1943), pp. 7–11.
[14] L. Kuipers and H. Niederreiter, *Uniform Distribution of Sequences*, Pure Appl. Math. (N.Y.), John Wiley, New York, 1974.
[15] G. Larcher, H. Niederreiter, and W. Ch. Schmid, *Digital nets and sequences constructed over finite rings and their application to quasi-Monte Carlo integration*, Monatsh. Math., 121 (1996), pp. 231–253.
[16] G. Larcher and G. Pirsic, *Base change problems for generalized Walsh series and multivariate numerical integration*, Pacific J. Math., 189 (1999), pp. 75–105.
[17] J. Matoušek, *Geometric Discrepancy*, Algorithms Combin. 18, Springer, Berlin, 1999.
[18] H. Niederreiter, *Quasi-Monte Carlo methods and pseudo-random numbers*, Bull. Amer. Math. Soc., 84 (1978), pp. 957–1041.
[19] H. Niederreiter, *Low-discrepancy point sets*, Monatsh. Math., 102 (1986), pp. 155–167.
[20] H. Niederreiter, *Random Number Generation and Quasi-Monte Carlo Methods*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 63, SIAM, Philadelphia, 1992.
[21] H. Niederreiter, *Constructions of* (t, m, s)*-nets and* (t, s)*-sequences*, Finite Fields Appl., 11 (2005), pp. 578–600.
[22] H. Niederreiter and G. Pirsic, *Duality for digital nets and its applications*, Acta Arith., 97 (2001), pp. 173–182.
[23] H. Niederreiter and C. P. Xing, *Quasirandom points and global function fields*, in Finite Fields and Applications, S. Cohen and H. Niederreiter, eds., London Math. Soc. Lecture Note Ser. 233, Cambridge University Press, Cambridge, UK, 1996, pp. 269–296.
[24] A. B. Owen, *Monte Carlo, quasi-Monte Carlo, and randomized quasi-Monte Carlo*, in Monte Carlo and Quasi-Monte Carlo Methods 1998 (Claremont, CA), Springer, Berlin, 2000, pp. 86–97.
[25] G. Pirsic, *Embedding Theorems and Numerical Integration of Walsh Series over Groups*, Ph.D. thesis, University of Salzburg, 1997.
[26] G. Pirsic, J. Dick, and F. Pillichshammer, *Cyclic digital nets, hyperplane nets, and multivariate integration in Sobolev spaces*, SIAM J. Numer. Anal., 44 (2006), pp. 385–411.
[27] H. A. Rademacher, *Einige Sätze über Reihen von allgemeinen Orthogonalfunktionen*, Math. Ann., 87 (1922), pp. 112–138.
[28] M. Yu. Rosenbloom and M. A. Tsfasman, *Codes in the m-metric*, Problemi Peredachi Inf., 33 (1997), pp. 45–52.
[29] I. F. Sharygin, *A lower estimate for the error of quadrature formulas for certain classes of functions*, Zh. Vychisl. Mat. i Mat. Fiz., 3 (1963), pp. 370–376.
[30] I. H. Sloan and H. Woźniakowski, *When are quasi-Monte Carlo algorithms efficient for high dimensional integrals?*, J. Complexity, 14 (1998), pp. 1–33.
[31] I. M. Sobol, *The distribution of points in a cube and the approximate evaluation of integrals*, Zh. Vychisl. Mat. i Mat. Fiz., 7 (1967), pp. 784–802.
[32] J. Stoer and R. Bulirsch, *Introduction to Numerical Analysis*, Texts Appl. Math. 12, Springer, New York, 2002.
[33] J. L. Walsh, *A closed set of normal orthogonal functions*, Amer. J. Math., 55 (1923), pp. 5–24.
[34] A. Zygmund, *Trigonometric Series*, Cambridge University Press, Cambridge, UK, 1959.

SIAM J. NUMER. ANAL.
Vol. 46, No. 3, pp. 1554–1579
© 2008 Society for Industrial and Applied Mathematics

A CONVERGENT FINITE DIFFERENCE SCHEME FOR THE CAMASSA–HOLM EQUATION WITH GENERAL H^1 INITIAL DATA*

G. M. COCLITE†, K. H. KARLSEN‡, AND N. H. RISEBRO‡

Abstract. We suggest a finite dfference scheme for the Camassa–Holm equation that can handle general H^1 initial data. The form of the difference scheme is judiciously chosen to ensure that it satisfies a total energy inequality. We prove that the difference scheme converges strongly in H^1 towards a dissipative weak solution of the Camassa–Holm equation.

Key words. Camassa–Holm equation, hyperbolic-elliptic system, weak solution, finite difference scheme, convergence

AMS subject classifications. 35G25, 35L05, 65M06, 65M12

DOI. 10.1137/060673242

1. Introduction.

In this paper we present and analyze a finite difference scheme for the Camassa–Holm partial differential equation [7]

$$\partial_t u - \partial^3_{txx} u + 3u\partial_x u = 2\partial_x u \partial^2_{xx} u + u\partial^3_{xxx} u, \qquad (t,x) \in \mathbb{R}_+ \times \mathbb{R}, \tag{1.1}$$

which we augment with an initial condition:

$$u|_{t=0} = u_0 \in H^1(\mathbb{R}). \tag{1.2}$$

By rewriting (1.1) as

$$(1-\partial^2_{xx})\left[\partial_t u + u\partial_x u\right] + \partial_x\left(u^2 + \frac{1}{2}(\partial_x u)^2\right) = 0,$$

we see that (for smooth solutions) (1.1) is equivalent to the elliptic-hyperbolic system

$$\partial_t u + u\partial_x u + \partial_x P = 0, \qquad -\partial^2_{xx} P + P = u^2 + \frac{1}{2}(\partial_x u)^2. \tag{1.3}$$

By recalling that $e^{-|x|}/2$ is the Green's function of the operator $1-\partial^2_{xx}$, (1.3) can be written as

$$\partial_t u + \partial_x F(u, \partial_x u) = 0, \quad F(u,\partial_x u) = \frac{1}{2}\left[u^2 + e^{-|x|} \star \left(u^2 + \frac{1}{2}(\partial_x u)^2\right)\right], \tag{1.4}$$

which can be viewed as a conservation law with a nonlocal flux function. In this paper the relevant formulation of the Camassa–Holm equation (1.1) is the one provided by the hyperbolic-elliptic system (1.3).

*Received by the editors October 25, 2006; accepted for publication (in revised form) December 26, 2007; published electronically April 4, 2008. This work was supported by the Research Council of Norway through the project WaveMaker.

http://www.siam.org/journals/sinum/46-3/67324.html

†Dipartimento di Matematica, Università degli Studi di Bari, Via E. Orabona 4, 70125 Bari, Italy (coclitegm@dm.uniba.it).

‡Centre of Mathematics for Applications (CMA), University of Oslo, P.O. Box 1053, Blindern, N–0316 Oslo, Norway (kennethk@math.uio.no, http://folk.uio.no/kennethk/, nilshr@math.uio.no, http://folk.uio.no/nilshr/). The second author's research was supported by the Research Council of Norway through an Outstanding Young Investigators Award.

The Camassa–Holm equation (1.1) can be viewed as a model for the propagation of unidirectional shallow water waves [7, 32]. The equation is a member of the class of weakly nonlinear and weakly dispersive shallow water models, a class which already contains the Korteweg–de Vries (KdV) and Benjamin–Bona–Mahony (BBM) equations. The Camassa–Holm equation contains higher order nonlinear dispersive/nonlocal balances not present in the KdV and BBM equations. As is the case with the BBM equation but not in the KdV equation, the linear dispersion relation in the Camassa–Holm equation remains bounded for large wave numbers.

In another interpretation the Camassa–Holm equation models finite length, small-amplitude radial deformation waves in cylindrical compressible hyperelastic rods [21]. It arises also in the context of differential geometry as an equation for geodesics of the H^1-metric on the diffeomorphism group; see [17, 18, 30, 36].

The Camassa–Holm equation possesses several extraordinary properties such as an infinite number of conserved integrals, a bi-Hamiltonian structure, and complete integrability [2, 7, 19, 13, 26]. Moreover, it enjoys an infinite number of nonsmooth solitary wave solutions, called peakons, which have to be interpreted as weak solutions of (1.4).

From a mathematical point of view the Camassa–Holm equation has by now become rather well studied. While it is impossible to give a complete overview of the mathematical literature, we shall here mention a few typical results, starting with local(-in-time) existence results [14, 35, 37]. For $u_0 \in H^s(\mathbb{R})$, with $s > \frac{3}{2}$, there exists a unique solution $u \in C([0,T]; H^s(\mathbb{R})) \cap C^1([0,T]; H^{s-1}(\mathbb{R}))$ of (1.1)–(1.2) for some T that depends on $\|u_0\|_{H^s(\mathbb{R})}$. Furthermore, the flow map is continuous from $H^s(\mathbb{R})$ to the class defined above. The proof of this result is based on the "momentum" formulation of the Camassa–Holm equation,

$$\partial_t m + u\partial_x m + 2m\partial_x u = 0, \qquad m := (1 - \partial_{xx}^2)u. \tag{1.5}$$

to which one applies Kato's theory for quasi-linear hyperbolic equations. For local well-posedness results based on Besov spaces, see [23, 22].

The Camassa–Holm equation posseses an infinite number of conservation laws, but none of them control the H^s-norm for $s > 1$. Hence these local existence results cannot (in general) be turned into global ones. Indeed, it is well known that global (strong) solutions do not exist and wave breaking occurs [7]. Wave breaking means that the solution itself stays bounded while the spatial derivative $\partial_x u$ tends to $-\infty$ as $t \uparrow T^*$, where T^* denotes the maximal time of existence. More precisely, the following results are proved in [14, 16]. Assume that $u_0 \in H^3(\mathbb{R})$ is odd with $\partial_x u_0(0) < 0$. Then the solution of of (1.1)–(1.2) does not exist globally, and T^* is estimated above by $1/(2\,|\partial_x u_0(0)|)$. Another result says that if the initial function $u_0 \in H^3(\mathbb{R})$ has at some point a slope which is less than $-(1/\sqrt{2})\,\|u_0\|_{H^1(\mathbb{R})}$, then T^* is finite and wave breaking occurs. It was observed in [14] that the solutions are global if $m_0 := (1 - \partial_{xx}^2)u_0$ (cf. (1.5)) is a bounded measure with a definitive sign.

In view of what we have said so far (peakon solutions/wave breaking), it is clear that a theory based on weak solutions is essential. In the literature there are a number of results on weak solutions of the Camassa–Holm equation. Here we will mention only a few of them, starting with the results obtained in [15, 20]. Suppose that $u_0 \in H^1(\mathbb{R})$ with $m_0 := (1 - \partial_{xx}^2)u_0 \in \mathcal{M}(\mathbb{R})$. Then the authors prove that there exists a final time $T = T(\|m_0\|_{\mathcal{M}}) > 0$ and a unique weak solution $u \in C([0,T]; H^1(\mathbb{R})) \cap L^\infty(0,T; W^{1,1}(\mathbb{R}))$, with $\partial_x u \in L^\infty(0,T; BV(\mathbb{R}))$, of (1.1)–(1.2); i.e., u is a distributional solution of (1.4)–(1.2). Additionally, the following

two time-dependent quantities remain constant: $E(u) := \int_{\mathbb{R}} \left[u^2 + (\partial_x u)^2\right] dx$ and $F(u) := \int_{\mathbb{R}} \left[u^3 + u(\partial_x u)^2\right] dx$. In particular, this weak solution is total energy conserving; i.e., $E(u(t,\cdot)) = E(u_0)$. Finally, if m_0 has a definite sign, then u is global in time. The sign of m_0 is maintained by $m(t,\cdot)$ at all times t. It is possible to prove the existence of local weak solutions without the sign assumption on m_0; see [22]. The proofs in [15, 20] are based on the momentum formulation (1.5).

For other approaches to conservative weak solutions, we refer to [4, 5, 29].

More relevant from the point of view of the present paper is the result of Xin and Zhang [38], which states the existence of a global (dissipative) weak solution for any H^1 initial data (see [11, 12] for similar results for a generalized Camassa–Holm equation). These solutions are global in the sense that they are defined even past the blow-up time (wave breaking). More precisely, suppose that $u_0 \in H^1(\mathbb{R})$. Then there exists a global weak solution $u \in C(\mathbb{R}_+ \times \mathbb{R}) \cap L^\infty(0,T;H^1(\mathbb{R}))$ of (1.1)–(1.2), satisfying the following properties:

$$\begin{aligned} &\|u(t,\cdot)\|_{H^1(\mathbb{R})} \le \|u_0\|_{H^1(\mathbb{R})}, \\ &\partial_x u \in L^p_{\mathrm{loc}}(\mathbb{R}_+ \times \mathbb{R}), \quad p < 3, \\ &\partial_x u(t,x) \le \frac{2}{t} + C, \quad t > 0, \end{aligned} \tag{1.6}$$

where C is a positive constant that depends only on $\|u_0\|_{H^1(\mathbb{R})}$. The proof of the existence result is based on the vanishing viscosity method; see [38].

We remark that the last item in (1.6) serves as an "entropy condition" that singles out a (presumably) unique weak solution after the occurrence of wave breaking. This solution is often referred to as a *dissipative* weak solution as the total energy is merely nonincreasing in time: $E(u(t,\cdot)) \le E(u_0)$. The entropy condition is (formally) seen to hold by inspecting the equation satisfied by the spatial derivative $q := \partial_x u$ (cf. [38] for details), which reads

$$\partial_t q + u\partial_x q + \frac{q^2}{2} - u^2 + P = 0, \qquad -\partial_{xx}^2 P + P = u^2 + \frac{q^2}{2}. \tag{1.7}$$

Currently there is no uniqueness result for weak solutions of the type constructed in [38]. The problem appears to be connected to a lack of temporal integrability (of the L^∞-norm) of the spatial derivative. Indeed, if one furthermore knows the existence of a function $b \in L^2_{\mathrm{loc}}(\mathbb{R}_+)$ such that $\|\partial_x u(t,\cdot)\|_{L^\infty(\mathbb{R})} \le b(t)$, then the weak solution of Xin and Zhang is unique (in a particular class) [39]. For example, if m_0 is a positive bounded Radon measure, then $\partial_x u$ is pointwise bounded [20], and uniqueness thus holds.

For a different approach to dissipative weak solutions, see the recent work [3].

Let us now turn to the main topic of the present paper, namely, convergent numerical schemes for the Camassa–Holm equation. Although there are few works on convergent numerical schemes, there are several authors that employ numerical schemes to obtain approximate solutions. The first numerical results are presented in [8], where a pseudospectral scheme is utilized. Additional numerical simulations with pesudospectral schemes are reported in [25, 31]. Numerical schemes based on multi-peakons (thereby exploiting the Hamiltonian structure of the Camassa–Holm equation) are examined in [6, 9, 10]. In a different direction, an adaptive high-resolution finite volume scheme is developed and used in [1].

Regarding works that provide numerical schemes with some sort of theoretical foundation, we know only of the papers [27, 28, 33]. In [28], the authors prove that

the multipeakon algorithm from [9, 10] converges to the solution of the Camassa–Holm equation (1.1) as the number of peakons tends to infinity (in an appropriate way). This convergence result applies to the situation where the initial function $u_0 \in H^1$ is such that $(1 - \partial_{xx}^2)u_0$ is a positive measure. In [33], the authors establish error estimates for a spectral projection scheme, though under the (unrealistic) assumption of smooth solutions.

It seems rather difficult to construct numerical schemes for which one can prove rigorously the convergence to a solution of the Camassa–Holm equation, a fact that is related to the nonlinear and nonlocal features of the equation. It has been observed in [27] that certain "natural" schemes either diverge or converge to a wrong solution. Indeed, a priori it is not even clear which one of the three formulations of the Camassa–Holm equation—(1.1), (1.3), or (1.5)—should be used as a starting point for discretization. Nevertheless, in [27] the authors commence from the momentum formulation (1.5) and thereby restrict themselves to initial data u_0 in H^1 for which $m_0 = (1-\partial_{xx}^2)u_0$ is a positive measure, in which case also $m(t,\cdot)$ remains positive and consequently so does u. They prove that the following difference scheme converges strongly in H^1 to the weak solution identified in [15, 20]:

$$\frac{d}{dt}m_j + D_-(m_j u_j) + m_j D u_j = 0, \quad m_j = u_j - D_- D_+ u_j, \quad t > 0,\ j \in \mathbb{Z},$$

where D_-, D, and D_+ denote, respectively, the backward, central, and forward difference operators and $m_j(t) \approx m(t, x_j)$, $u_j(t) \approx u(t, x_j)$, $x_j = j\Delta x$, and $\Delta x > 0$.

The main aim of this paper is to provide a convergent finite difference scheme that works for any H^1 initial data and not merely the subclass considered in [27]. Neither the scheme nor the analysis presented in [27] works in the general case.

At variance with [27], we shall herein take as a starting point the hyperbolic-elliptic formulation (1.3). From the point of view of conservation laws (e.g., the inviscid Burgers equation) and their shock wave (discontinuous) solutions, it might seem natural to employ a conservative finite difference scheme of the upwind type [34] to the u-equation in (1.3). As is well known, the upwinding will render a scheme stable since the difference stencil utilizes information only from the side from which the (discontinuous) waves are coming. However, here one should keep in mind that solutions to the Camassa–Holm equation are continuous and that prospective discontinuities occur only in the variable $q = \partial_x u$, which satisfies the transport equation in (1.7). Thus, herein we will not opt for this strategy.

Instead we will devise a tailored difference scheme for the u-equation in (1.3) that yields an upwind difference scheme for the q-equation in (1.7). A key feature of the scheme is that it satisfies a total energy inequality in which only the q-part of the total energy is dissipated (not the u-part!). To avoid complicating further the convergence analysis, we restrict our attention to a semidiscrete finite difference scheme. To turn the difference scheme into a fully discrete one, we can rely on a variety of different time-discretization techniques.

Now we outline the finite difference scheme (here only briefly since the details can be found in section 3). To this end, we start with discretizing the spatial domain $\mathbb{R}$ by specifying the mesh points $x_j = j\Delta x$, $x_{j+1/2} = (j + 1/2)\Delta x$ for $j = 0, \pm 1, \pm 2, \ldots$, where $\Delta x > 0$ is the length between two consecutive mesh points (the mesh size). Our numerical scheme will generate approximations $u_{j+1/2}(t) \approx u(t, x_{j+1/2})$, $P_j(t) \approx P(t, x_j)$ for $t \geq 0$ and $j \in \mathbb{Z}$, where we remark that the discretization of P is shifted (staggered) one half-cell compared to that of u. Our finite difference scheme for

$\{u_{j+1/2}(t)\}_{j\in\mathbb{Z}}$ reads

$$\frac{d}{dt}u_{j+1/2} + \left(u_{j+1/2}\vee 0\right) D_- u_{j+1/2} + \left(u_{j+1/2}\wedge 0\right) D_+ u_{j+1/2} + D_+ P_j = 0, \tag{1.8}$$

while the difference scheme for $\{P_j(t)\}_{j\in\mathbb{Z}}$ takes the form

$$-D_-D_+P_j + P_j = \left(u_{j+1/2}\vee 0\right)^2 + \left(u_{j-1/2}\wedge 0\right)^2 + \frac{1}{2}\left(D_- u_{j+1/2}\right)^2.$$

Of course, as we have already alluded to above, from the point of view of the inviscid Burgers equation, (1.8) is not a reasonable discretization. However, the quantity $q_j := D_- u_{j+1/2}$ automatically satisfies the difference scheme

$$\begin{aligned} q_j' + \left(u_{j-1/2}\vee 0\right) D_- q_j + \left(u_{j+1/2}\wedge 0\right) D_+ q_j + \frac{q_j^2}{2} \\ - \left(u_{j+1/2}\vee 0\right)^2 - \left(u_{j-1/2}\wedge 0\right)^2 + P_j = 0, \end{aligned} \tag{1.9}$$

which contains proper upwinding of the transport term in (1.7). In our situation (compare with [27]) u does not have a definite sign, hence the splitting of u into positive and negative parts. As with the "pressure" P, the discretization of q is staggered compared to that of the "velocity" u.

By properly extending $\{u_{j+1/2}\}_{j\in\mathbb{Z}}$, $\{q_j\}_{j\in\mathbb{Z}}$ to functions $u_{\Delta x}$, $q_{\Delta x}$ defined at all points (t,x) in the domain, we prove that $u_{\Delta x}$ converges strongly in H^1 to a dissipative weak solution of the Camassa–Holm equation, which constitutes the main result of the present paper. Regarding the proof, we derive several a priori estimates in Lebesgue and Sobolev spaces as well as a uniform upper bound on q_j serving as a discrete version of the "entropy condition," among which a discrete total energy inequality constitutes the key building block. The total energy inequality ensures only weak compactness of the sequence $\left\{q_{\Delta x}^2\right\}_{\Delta x>0}$. However, it is crucial to know that this sequence is strongly compact. Strong compactness is needed if we want to recover the original equation (1.4) when sending $\Delta x \downarrow 0$ in the finite difference scheme. To establish the strong compactness property we apply ideas from the theory of renormalized solutions (in the sense of DiPerna and Lions) to the finite difference scheme (1.9). As a part of establishing strong compactness, a higher integrability estimate for $q_{\Delta x}$ is needed to ensure that weak limit points of $q_{\Delta x}^2$ do not contain singular measures. Our convergence proof can be best understood as a discrete variant of the proof used in [38] for the vanishing viscosity method.

This paper is organized as follows: In section 2 we state some relevant notations. The difference scheme and its main convergence theorem are presented in section 3. This theorem is a consequence of the results proved in sections 4–9.

Throughout this paper we use C to denote a generic constant; the actual value of C may change from one line to the next in a calculation.

2. Preliminaries. In this section we introduce some notations to be used throughout this paper and a few basic mathematical results that will be relevant to the convergence analysis of the numerical scheme.

The notations $a\vee 0 = \max\{a,0\} = \frac{a+|a|}{2}$ and $a\wedge 0 = \min\{a,0\} = \frac{a-|a|}{2}$ will be used frequently. In what follows, unless otherwise stated, the index j will run over $\mathbb{Z}$. For such an index we set $x_{j+1/2} = (j+1/2)\Delta x$ and introduce the grid cells $I_j = [x_{j-1/2}, x_{j+1/2})$, where Δx is a small positive number ("the discretization

parameter"). The grid cells I_j are centered around the points $x_j = j\Delta x$. For any sequence $\{v_j\}_{j\in\mathbb{Z}}$ we introduce the following difference operators:

$$D_+v_j := \frac{v_{j+1}-v_j}{\Delta x},\ D_-v_j := \frac{v_j - v_{j-1}}{\Delta x},\ Dv_j := \frac{D_+v_j + D_-v_j}{2} = \frac{v_{j+1}-v_{j-1}}{2\Delta x}.$$

We also use the notations

$$\left\|\{v_j\}_j\right\|_{\ell^p} := \left(\Delta x \sum_{j\in\mathbb{Z}} |v_j|^p\right)^{\frac{1}{p}},\ 1<p<\infty, \quad \left\|\{v_j\}_j\right\|_{\ell^\infty} := \sup_j |v_j|,$$

$$\left\|\{v_j\}_j\right\|_{h^1} := \left(\Delta x \sum_{j\in\mathbb{Z}} \left[v_j^2 + (D_-v_j)^2\right]\right)^{\frac{1}{2}}.$$

Let $\{v_j\}_{j\in\mathbb{Z}}$ be a sequence such that $\|\{v_j\}_j\|_{\ell^1\cap\ell^2} < \infty$. Then $\|\{v_j\}_j\|_{\ell^\infty} \le \frac{1}{\sqrt{\Delta x}}\|\{v_j\}_j\|_{\ell^2} \le \frac{1}{\Delta x}\|\{v_j\}_j\|_{\ell^1}$. Let $\{v_j\}_{j\in\mathbb{Z}}$ be such that $\|\{v_j\}_j\|_{h^1} < \infty$. It is easy to see that the following discrete Sobolev inequality holds:

$$\left\|\{v_j\}_j\right\|_{\ell^\infty} \le \frac{1}{\sqrt{2}}\left\|\{v_j\}_j\right\|_{h^1}. \tag{2.1}$$

Let $\{v_j\}_{j\in\mathbb{Z}}, \{w_j\}_{j\in\mathbb{Z}}$ be two sequences. Then the discrete Leibniz rule reads

$$D_\pm(v_jw_j) = v_jD_\pm w_j + D_\pm v_j w_{j\pm1}, \tag{2.2}$$

while the discrete chain rule states that for any C^2 function f there holds

$$D_\pm f(v_j) = f'(v_j)D_\pm v_j \pm \frac{\Delta x}{2} f''(\xi_j^\pm)(D_\pm v_j)^2 \tag{2.3}$$

for some $\xi_j^\pm$ between $v_{j\pm1}$ and v_j. A key difficulty in designing converging difference schemes for nonlinear equations is that there is no exact chain rule for discrete derivatives but merely the formula (2.3) showing that the chain rule holds only up to a certain error term.

3. Difference scheme and convergence theorem. In this section we present the semidiscrete upwind difference scheme for the Camassa–Holm equation and state the main convergence result.

For $t>0$, we let $\{u_{j+1/2}(t)\}_{j\in\mathbb{Z}}$, where $u_{j+1/2}(t) \approx u(t, x_{j+1/2})$, solve the following system of ODEs:

$$\frac{d}{dt}u_{j+1/2} + \left(u_{j+1/2}\vee 0\right)D_-u_{j+1/2} + \left(u_{j+1/2}\wedge 0\right)D_+u_{j+1/2} + D_+P_j = 0, \tag{3.1}$$

where we specify the initial values as follows:

$$u_{j+1/2}(0) = u_0(x_{j+1/2}). \tag{3.2}$$

For $t\ge 0$, we let $\{P_j(t)\}_{j\in\mathbb{Z}}$, where $P_j(t) \approx P(t, x_j)$, solve

$$-D_-D_+P_j + P_j = \left(u_{j+1/2}\vee 0\right)^2 + \left(u_{j-1/2}\wedge 0\right)^2 + \frac{1}{2}\left(D_-u_{j+1/2}\right)^2. \tag{3.3}$$

By introducing $f_j := (u_{j+1/2} \vee 0)^2 + (u_{j-1/2} \wedge 0)^2 + \frac{q_j^2}{2}$, $h := (1 + 2\frac{1-e^{-\kappa}}{(\Delta x)^2})^{-1}$, and $\kappa := \ln(1 + \frac{\Delta x^2}{2} + \frac{\Delta x}{2}\sqrt{4 + \Delta x^2})$, it is not difficult to verify that the solution of (3.3) has the form

$$P_j = h \sum_i e^{-\kappa|j-i|} f_i, \qquad j \in \mathbb{Z}. \tag{3.4}$$

For later use, one should observe that

$$h = \frac{\Delta x}{2} + \mathcal{O}(\Delta x^2), \quad \frac{|e^{\kappa} - 1|}{\Delta x} = 1 + \mathcal{O}(\Delta x), \quad \frac{|e^{-\kappa} - 1|}{\Delta x} = 1 + \mathcal{O}(\Delta x).$$

Since $\{P_j\}_{j\in\mathbb{Z}}$ can be expressed solely in terms of $\{u_{j+1/2}(t)\}_{j\in\mathbb{Z}}$, we see that (3.1) constitutes an infinite-dimensional system of ODEs of the form

$$\frac{d}{dt} u_{j+1/2}(t) = F\left(\left\{u_{j+1/2}(t)\right\}_{j\in\mathbb{Z}}\right). \tag{3.5}$$

The proof of the next lemma is standard and is therefore omitted.

LEMMA 3.1. *For each fixed $\Delta x > 0$, the ODE system* (3.1)*, or* (3.5)*, has a continuously differentiable solution defined for all $t > 0$.*

Next, let us derive the difference scheme satisfied by

$$q_j = D_- u_{j+1/2}. \tag{3.6}$$

This will be done by applying the difference operator D_- to the u-equation (3.1). To this end, by applying the discrete Leibniz rule we get

$$D_-\left[\left(u_{j+1/2} \vee 0\right) D_- u_{j+1/2}\right] = \left(u_{j-1/2} \vee 0\right) D_- q_j + D_-\left(u_{j+1/2} \vee 0\right) q_j$$

and

$$D_-\left[\left(u_{j+1/2} \wedge 0\right) D_+ u_{j+1/2}\right] = \left(u_{j+1/2} \wedge 0\right) D_+ q_j + D_-\left(u_{j+1/2} \wedge 0\right) q_j,$$

so that

$$\begin{aligned} D_-\left[\left(u_{j+1/2} \vee 0\right) D_- u_{j+1/2} + \left(u_{j+1/2} \wedge 0\right) D_+ u_{j+1/2}\right] \\ = \left(u_{j-1/2} \vee 0\right) D_- q_j + \left(u_{j+1/2} \wedge 0\right) D_+ q_j + q_j^2. \end{aligned} \tag{3.7}$$

The P-equation (3.3) rephrased in terms of q reads

$$-D_+ D_- P_j + P_j = \left(u_{j+1/2} \vee 0\right)^2 + \left(u_{j-1/2} \wedge 0\right)^2 + \frac{1}{2} q_j^2. \tag{3.8}$$

Employing (3.7) and (3.8) when applying D_- to the u-equation in (3.1) yields

$$\begin{aligned} q_j' + \left(u_{j-1/2} \vee 0\right) D_- q_j + \left(u_{j+1/2} \wedge 0\right) D_+ q_j \\ + \frac{q_j^2}{2} + P_j - \left(u_{j+1/2} \vee 0\right)^2 - \left(u_{j-1/2} \wedge 0\right)^2 = 0. \end{aligned} \tag{3.9}$$

Inasmuch as q can be discontinuous, (3.9) is a natural discretization of the q-equation in (1.7). Regarding the initial values, it easy to see that

$$q_j(0) = \frac{1}{\Delta x} \int_{I_j} \partial_x u_0(x)\, dx, \quad j \in \mathbb{Z}. \tag{3.10}$$

The main aim of this paper is to prove that the numerical scheme defined in section 3 converges to a solution of the Camassa–Holm equation. Before we can do that, we need to define what we mean by a "solution."

DEFINITION 3.1. *We call a function* $u = u(t,x) : [0,\infty)\times\mathbb{R} \to \mathbb{R}$ *a weak solution of the Cauchy problem* (1.1)–(1.2) *provided*

(i) $u \in L^\infty(\mathbb{R}_+; H^1(\mathbb{R})) \cap C([0,\infty)\times\mathbb{R})$*;*

(ii) u *satisfies* (1.4) *in the sense of distributions; that is,* $\forall \phi \in C_c^\infty(\mathbb{R}_+\times\mathbb{R})$

$$\int_{\mathbb{R}_+}\int_{\mathbb{R}} u\partial_t\phi + F(u,\partial_x u)\partial_x\phi\, dxdt = 0; \tag{3.11}$$

(iii) $u(0,x) = u_0(x)$ *for every* $x \in \mathbb{R}$*;*

(iv) $\|u(t,\cdot)\|_{H^1(\mathbb{R})} \le \|u_0\|_{H^1(\mathbb{R})}$ *for each* $t > 0$.

If, in addition, there is a constant $C \ge 0$ *depending only on* $\|u_0\|_{H^1(\mathbb{R})}$ *such that*

$$u_x(t,x) \le \frac{2}{t} + C, \qquad (t,x) \in (0,\infty)\times\mathbb{R}, \tag{3.12}$$

then we call u *a dissipative weak solution of the Cauchy problem* (1.1)–(1.2).

The existence of a dissipative solution has been previously established in [38]; see also [39] for a partial uniqueness result.

Supplied with the sequences $\{u_{j+1/2}(t)\}_{j\in\mathbb{Z}}$, $\{q_j(t)\}_{j\in\mathbb{Z}}$ defined by (3.1)–(3.3) and (3.6), we introduce the function

$$u_{\Delta x}(t,x) = q_j(t)(x - x_{j+1/2}) + u_{j-1/2}(t), \quad t \ge 0,\ x \in I_j,\ j \in \mathbb{Z}. \tag{3.13}$$

Observe that $u_{\Delta x}(t,\cdot)$ is piecewise linear and continuous. Besides,

$$\begin{aligned} &u_{\Delta x}(t, x_{j\pm1/2}) = u_{j\pm1/2}(t), \quad t \ge 0,\ j \in \mathbb{Z},\\ &\partial_x u_{\Delta x}(t,x) = q_j(t), \qquad t \ge 0,\ x \in I_j\ j \in \mathbb{Z}. \end{aligned}$$

We are now in a position to state our main result.

THEOREM 3.1. *Suppose that* (1.2) *holds. Let* $\{u_{\Delta x}\}_{\Delta x>0}$ *be a sequence of difference solutions defined by* (3.13), (3.1)–(3.3), *and* (3.6). *Then, along a subsequence as* $\Delta x \downarrow 0$,

$$u_{\Delta x} \to u \text{ in } H^1_{\mathrm{loc}}(\mathbb{R}_+\times\mathbb{R}),$$

where u *is a dissipative weak solution of the Cauchy problem* (1.1)–(1.2).

This theorem is a consequence of the results stated and proved in sections 4–9.

4. Discrete total energy estimate. The finite difference scheme (3.1)–(3.6) is designed to admit the discrete total energy estimate stated below, which contains a dissipation term resulting from the upwind nature of the scheme for the q-variable (notice that there is no dissipation associated with the u-variable).

LEMMA 4.1. *For each* $t \ge 0$,

$$\begin{aligned} &\left\|\left\{u_{j+1/2}(t)\right\}_j\right\|_{h^1}^2 + \Delta x^2 \sum_j \int_0^t \left|u_{j+1/2}(s)\right| \left(D_+D_- u_{j+1/2}(s)\right)^2 ds \\ &\qquad = \left\|\left\{u_{j+1/2}(0)\right\}_j\right\|_{h^1}^2. \end{aligned} \tag{4.1}$$

Proof. We multiply the u-equation in (3.1) by $u_{j+1/2}$ and use $q_j = D_- u_{j+1/2}$ to obtain

$$\begin{aligned}\frac{d}{dt}\left(\frac{u_{j+1/2}^2}{2}\right) &+ \left(u_{j+1/2}\vee 0\right)^2 q_j \\ &+ \left(u_{j+1/2}\wedge 0\right)^2 q_{j+1} + \left(D_+ P_j\right) u_{j+1/2} = 0,\end{aligned} \tag{4.2}$$

while multiplying the q-equation in (3.9) by q_j yields

$$\begin{aligned}\frac{d}{dt}\left(\frac{q_j^2}{2}\right) &+ \left(u_{j-1/2}\vee 0\right)\left(D_- q_j\right) q_j + \left(u_{j+1/2}\wedge 0\right)\left(D_+ q_j\right) q_j \\ &+ \frac{q_j^3}{2} - \left(u_{j+1/2}\vee 0\right)^2 q_j - \left(u_{j-1/2}\wedge 0\right)^2 q_j + P_j q_j = 0.\end{aligned} \tag{4.3}$$

Adding (4.2) and (4.3), multiplying the result with Δx, and summing over j yields

$$\frac{d}{dt}\left[\Delta x \sum_j \left(\frac{u_{j+1/2}^2}{2} + \frac{q_j^2}{2}\right)\right] + \mathrm{I} + \mathrm{II} + \mathrm{III} = 0,$$

where

$$\begin{aligned}\mathrm{I} &= \Delta x \sum_j \left(u_{j-1/2}\vee 0\right)\left(D_- q_j\right) q_j + \Delta x \sum_j \left(u_{j+1/2}\wedge 0\right)\left(D_+ q_j\right) q_j + \Delta x \sum_j \frac{q_j^3}{2}, \\ \mathrm{II} &= \Delta x \sum_j \left(u_{j+1/2}\vee 0\right)^2 q_j + \Delta x \sum_j \left(u_{j+1/2}\wedge 0\right)^2 q_{j+1} \\ &\quad - \Delta x \sum_j \left(u_{j+1/2}\vee 0\right)^2 q_j - \Delta x \sum_j \left(u_{j-1/2}\wedge 0\right)^2 q_j \equiv 0 \quad \text{(by shifting indices)}, \\ \mathrm{III} &= \Delta x \sum_j D_+ P_j u_{j+1/2} + \Delta x \sum_j P_j q_j \equiv 0 \quad \text{(by summation by parts)}.\end{aligned}$$

Let us now deal with term I. The discrete chain rule implies that

$$\left(D_\pm q_j\right) q_j = D_\pm\left(\frac{q_j^2}{2}\right) \mp \frac{\Delta x}{2}\left(D_\pm q_j\right)^2.$$

Hence

$$\begin{aligned}\mathrm{I} &= \Delta x \sum_j \left(u_{j-1/2}\vee 0\right)\left[D_-\left(\frac{q_j^2}{2}\right) + \frac{\Delta x}{2}\left(D_- q_j\right)^2\right] \\ &\quad + \Delta x \sum_j \left(u_{j+1/2}\wedge 0\right)\left[D_+\left(\frac{q_j^2}{2}\right) - \frac{\Delta x}{2}\left(D_+ q_j\right)^2\right] + \Delta x \sum_j \frac{q_j^3}{2} \\ &= \mathrm{I}_1 + \mathrm{I}_2,\end{aligned}$$

where

$$\begin{aligned}
\mathrm{I}_1 &= \Delta x \sum_j \left(u_{j-1/2} \vee 0\right) D_- \left(\frac{q_j^2}{2}\right) + \Delta x \sum_j \left(u_{j+1/2} \wedge 0\right) D_+ \left(\frac{q_j^2}{2}\right) + \Delta x \sum_j \frac{q_j^3}{2}, \\
\mathrm{I}_2 &= \frac{\Delta x^2}{2} \sum_j \Big[\left(u_{j-1/2} \vee 0\right) (D_- q_j)^2 - \left(u_{j+1/2} \wedge 0\right) (D_+ q_j)^2\Big] \\
&= \frac{\Delta x^2}{2} \sum_j \Big[\left(u_{j+1/2} \vee 0\right) (D_- q_{j+1})^2 - \left(u_{j+1/2} \wedge 0\right) (D_+ q_j)^2\Big] \\
&= \frac{\Delta x^2}{2} \sum_j |u_{j+1/2}| (D_+ q_j)^2 \geq 0.
\end{aligned}$$

To handle the I_1-term, we use the discrete Leibniz rule, which implies that

$$\begin{aligned}
D_- \left[\left(u_{j+1/2} \vee 0\right) \frac{q_j^2}{2}\right] &= \left(u_{j-1/2} \vee 0\right) D_- \left(\frac{q_j^2}{2}\right) + D_- \left(u_{j+1/2} \vee 0\right) \frac{q_j^2}{2}, \\
D_+ \left[\left(u_{j-1/2} \wedge 0\right) \frac{q_j^2}{2}\right] &= \left(u_{j+1/2} \wedge 0\right) D_+ \left(\frac{q_j^2}{2}\right) + D_+ \left(u_{j-1/2} \vee 0\right) \frac{q_j^2}{2} \\
&= \left(u_{j+1/2} \wedge 0\right) D_+ \left(\frac{q_j^2}{2}\right) + D_- \left(u_{j+1/2} \vee 0\right) \frac{q_j^2}{2},
\end{aligned}$$

to obtain

$$\begin{aligned}
\mathrm{I}_1 = {} & \Delta x \sum_j D_- \left[\left(u_{j+1/2} \vee 0\right) \frac{q_j^2}{2}\right] - \Delta x \sum_j D_- \left(u_{j+1/2} \vee 0\right) \frac{q_j^2}{2} \\
& + \Delta x \sum_j D_+ \left[\left(u_{j-1/2} \wedge 0\right) \frac{q_j^2}{2}\right] - \Delta x \sum_j D_- \left(u_{j+1/2} \vee 0\right) \frac{q_j^2}{2} + \Delta x \sum_j \frac{q_j^3}{2} \\
= {} & - \Delta x \sum_j D_- u_{j+1/2} \frac{q_j^2}{2} + \Delta x \sum_j \frac{q_j^3}{2} - \Delta x \sum_j \frac{q_j^3}{2} + \Delta x \sum_j \frac{q_j^3}{2} = 0.
\end{aligned}$$

Summarizing our findings, the following discrete energy estimate holds:

$$\frac{d}{dt}\left[\Delta x \sum_j \left(\frac{u_{j+1/2}^2}{2} + \frac{q_j^2}{2}\right)\right] + \frac{\Delta x^2}{2} \sum_j \left|u_{j+1/2}\right| \left(D_+ q_j\right)^2 = 0.$$

Finally, by integrating over $[0,t]$ we get (4.1). □

Remark 4.1. In view of (4.1) and (2.1)

$$\left\| \left\{u_{j+1/2}(t)\right\}_j \right\|_{\ell^\infty} \leq C \left\|u_0\right\|_{H^1(\mathbb{R})}, \qquad t \geq 0, \tag{4.4}$$

where $C > 0$ is a constant that is independent of Δx.

Later we will need the strong L^p convergence of the numerical approximations of the "pressure" P. We now collect some estimates on $\{P_j\}_{j\in\mathbb{Z}}$ implying this convergence; the estimates are all straightforward consequences of (4.1) and (3.4).

LEMMA 4.2. *For each $t \geq 0$,*

$$\left\| \{P_j(t)\}_j \right\|_{\ell^\infty}, \left\| \{P_j(t)\}_j \right\|_{\ell^1} \leq C \left\| u_0 \right\|^2_{H^1(\mathbb{R})}, \tag{4.5}$$

$$\left\| \{D_+P_j(t)\}_j \right\|_{\ell^\infty}, \left\| \{D_+P_j(t)\}_j \right\|_{\ell^1} \leq C \left\| u_0 \right\|^2_{H^1(\mathbb{R})}, \tag{4.6}$$

where $C > 0$ is a constant independent of Δx.

5. Discrete Oleĭnik estimate. The aim of this section is to prove that the quantity $q_j = D_- u_{j+1/2}$ is uniformly upper bounded on $\{t > 0\}$, thereby revealing the dissipative nature of our scheme.

LEMMA 5.1. *For $t > 0$, $j \in \mathbb{Z}$,*

$$q_j(t) \leq \frac{2}{t} + C \tag{5.1}$$

for some positive constant C that is independent of Δx.

Proof. By (4.4) and (4.5), it follows from (1.9) that

$$q_j' + \frac{q_j^2}{2} \leq L - \left[\left(u_{j-1/2} \vee 0 \right) D_- q_j + \left(u_{j+1/2} \wedge 0 \right) D_+ q_j \right], \quad j \in \mathbb{Z},\ t > 0, \tag{5.2}$$

for some constant $L > 0$. Since $\lim_{j \to \pm\infty} q_j(t) = 0$ there is an index $i(t) \in \mathbb{Z}$ such that

$$q_{i(t)}(t) = \sup_{j \in \mathbb{Z}} q_j(t), \qquad t > 0. \tag{5.3}$$

At $j = i(t)$ for $t > 0$ there holds $D_+ q_{i(t)}(t) \leq 0 \leq D_- q_{i(t)}(t)$, so that

$$\left(u_{i(t)-1/2}(t) \vee 0 \right) D_- q_{i(t)}(t) + \left(u_{i(t)+1/2}(t) \wedge 0 \right) D_+ q_{i(t)}(t) \geq 0, \qquad t > 0,$$

which inserted into (5.2) yields

$$q_{i(t)}'(t) + \frac{q_{i(t)}^2(t)}{2} \leq L, \qquad t > 0. \tag{5.4}$$

One can check that $f(t) := \frac{2}{t} + \sqrt{2L}$ is a supersolution of the ODE $y' + \frac{y^2}{2} = L$ on $\{t > 0\}$, while (5.4) shows that $q_{i(t)}(t)$ is a subsolution. Hence, by the comparison principle for ODEs and (5.3), $q_j(t) \leq q_{i(t)}(t) \leq \frac{2}{t} + \sqrt{2L}$ for $j \in \mathbb{Z}$, $t > 0$. ∎

6. Discrete higher integrability estimate. In view of (4.1) we infer that $(\partial_x u_{\Delta x})^2$ converges (in the sense of measures) along a subsequence as $\Delta x \downarrow 0$. To ensure that the limit does not contain concentration effects (singular measures), we shall derive a discrete higher integrability estimate.

To prepare for the derivation of this estimate (but also for later use), we will derive a "renormalized form" of the finite difference scheme for q_j. To this end, let f be a nonlinear function (renormalization) of appropriate regularity and growth. By multiplying (3.9) by $f'(q_j)$ and using the discrete chain rule, which in the present context reads

$$(D_\pm q_j) f'(q_j) = D_\pm f(q_j) \mp \frac{\Delta x}{2} f''(\xi_j^\pm) \left(D_\pm q_j \right)^2,$$

for some numbers $\xi_j^{\pm}$ between q_j and $q_{j\pm1}$, we obtain the following renormalized difference scheme:

$$\begin{aligned}\frac{d}{dt}f(q_j)+\big(u_{j-1/2}\vee 0\big)\,D_-f(q_j)+\big(u_{j+1/2}\wedge 0\big)\,D_+f(q_j)+\frac{q_j^2}{2}f'(q_j)\\ +\Big[P_j-\big(u_{j+1/2}\vee 0\big)^2-\big(u_{j-1/2}\wedge 0\big)^2\Big]f'(q_j)+I_{\Delta x,f'',j}=0,\end{aligned}\tag{6.1}$$

where

$$I_{\Delta x,f'',j}:=\frac{\Delta x}{2}\Big\{\big(u_{j-1/2}\vee 0\big)\,f''(\xi_j^-)(D_-q_j)^2-\big(u_{j+1/2}\wedge 0\big)\,f''(\xi_j^+)(D_+q_j)^2\Big\}.$$

Let us now write (6.1) in divergence form. To this end, observe that the discrete Leibniz rule (2.2) implies the following relations:

$$\begin{aligned}D_-\left[\big(u_{j+1/2}\vee 0\big)\,f(q_j)\right]&=\big(u_{j-1/2}\vee 0\big)\,D_-f(q_j)+D_-\big(u_{j+1/2}\vee 0\big)\,f(q_j),\\ D_+\left[\big(u_{j-1/2}\wedge 0\big)\,f(q_j)\right]&=\big(u_{j+1/2}\wedge 0\big)\,D_+f(q_j)+D_+\big(u_{j-1/2}\wedge 0\big)\,f(q_j)\\ &=\big(u_{j+1/2}\wedge 0\big)\,D_+f(q_j)+D_-\big(u_{j+1/2}\wedge 0\big)\,f(q_j),\end{aligned}$$

and therefore, by using that $q_j=D_-u_{j+1/2}$,

$$\begin{aligned}&\big(u_{j-1/2}\vee 0\big)\,D_-f(q_j)+\big(u_{j+1/2}\wedge 0\big)\,D_+f(q_j)\\ &\qquad=D_-\left[\big(u_{j+1/2}\vee 0\big)\,f(q_j)\right]+D_+\left[\big(u_{j-1/2}\wedge 0\big)\,f(q_j)\right]-q_jf(q_j).\end{aligned}$$

Hence, we end up with the following divergence-form variant of the renormalized difference scheme (6.1):

$$\begin{aligned}&\frac{d}{dt}f(q_j)+D_-\left[\big(u_{j+1/2}\vee 0\big)\,f(q_j)\right]+D_+\left[\big(u_{j-1/2}\wedge 0\big)\,f(q_j)\right]\\ &\qquad+\frac{q_j^2}{2}f'(q_j)-q_jf(q_j)+\Big[P_j-\big(u_{j+1/2}\vee 0\big)^2-\big(u_{j-1/2}\wedge 0\big)^2\Big]f'(q_j)\\ &\qquad\qquad+I_{\Delta x,f'',j}=0.\end{aligned}\tag{6.2}$$

We are now in a position to prove the following lemma.

LEMMA 6.1. *Let $\alpha\in(0,1)$, $T>0$, and j_a,j_b be integers such that $j_a<j_b$. Set $a:=j_a\Delta x$ and $b:=j_b\Delta x$. There exists a positive constant C, depending only on u_0,α,T,a,b, such that*

$$\int_0^T\Delta x\sum_{j=j_a}^{j_b}|q_j(t)|^{2+\alpha}\,dt\le C.\tag{6.3}$$

Proof. Our proof exploits (5.1). We start by introducing the notations

$$\begin{gathered}\mathcal{J}=\{j_a,\dots,j_b\},\\ \mathcal{N}(t)=\big\{j\in\mathcal{J}\;\big|\;q_j(t)<0\big\},\quad \mathcal{P}(t)=\big\{j\in\mathcal{J}\;\big|\;q_j(t)\ge 0\big\},\\ I=\int_0^T\Delta x\sum_{j\in\mathcal{J}}|q_j(t)|^{2+\alpha}\,dt,\\ I_-=\int_0^T\Delta x\sum_{j\in\mathcal{N}(t)}|q_j(t)|^{2+\alpha}\,dt,\quad I_+=\int_0^T\Delta x\sum_{j\in\mathcal{P}(t)}|q_j(t)|^{2+\alpha}\,dt\end{gathered}$$

and observing that

$$\mathcal{J} = \mathcal{N}(t) \cup \mathcal{P}(t), \quad I = I_+ + I_-.$$

By (4.1) and (5.1), and since $\alpha < 1$,

$$I_+ \le \int_0^T \Delta x \sum_{j \in \mathcal{P}(t)} |q_j(t)|^2 \left(\frac{2}{t} + C\right)^\alpha \, dt \le C(T,\alpha) \left\| \{u_j(0)\}_j \right\|_{h^1}^2 .$$

We have to estimate I_-. With $f(\xi) = |\xi|^{1+\alpha}$, (6.2) reads

$$\begin{aligned}
&\frac{d}{dt} |q_j|^{1+\alpha} + D_- \left[\left(u_{j+1/2} \vee 0\right) |q_j|^{1+\alpha} \right] + D_+ \left[\left(u_{j-1/2} \wedge 0\right) |q_j|^{1+\alpha} \right] \\
&\quad + \frac{\alpha - 1}{2} \operatorname{sign}(q_j) |q_j|^{2+\alpha} + (1+\alpha) P_j \operatorname{sign}(q_j) |q_j|^\alpha \\
&\quad - (1+\alpha) \left[\left(u_{j+1/2} \vee 0\right)^2 + \left(u_{j-1/2} \wedge 0\right)^2 \right] \operatorname{sign}(q_j) |q_j|^\alpha = -I_{\Delta x, f'', j} \le 0,
\end{aligned}$$

where we used the convexity of f to conclude the inequality. Let χ be a smooth cutoff function such that

$$0 \le \chi \le 1, \quad \xi \in [a, b+1] \Longrightarrow \chi(\xi) = 1, \quad \xi \notin [a-1, b+2] \Longrightarrow \chi(\xi) = 0.$$

By multiplying by $\Delta x \chi(j\Delta x)$, summing over $j \in \mathbb{Z}$, and integrating over $t \in (0,T)$, we arrive at

$$\begin{aligned}
0 \le\ & \Delta x \sum_j \left(|q_j(0)|^{1+\alpha} - |q_j(T)|^{1+\alpha} \right) \chi(j\Delta x) \\
& + \frac{1-\alpha}{2} \int_0^T \Delta x \sum_j \operatorname{sign}(q_j) |q_j|^{2+\alpha} \chi(j\Delta x) \, dt \\
& - \int_0^T \Delta x \sum_j D_- \left[\left(u_{j+1/2} \vee 0\right) |q_j|^{1+\alpha} \right] \chi(j\Delta x) \, dt \\
& - \int_0^T \Delta x \sum_j D_+ \left[\left(u_{j-1/2} \wedge 0\right) |q_j|^{1+\alpha} \right] \chi(j\Delta x) \, dt \\
& - (1+\alpha) \int_0^T \Delta x \sum_j |P_j| |q_j|^\alpha \chi(j\Delta x) \, dt \\
& + (1+\alpha) \int_0^T \Delta x \sum_j \left(u_{j+1/2} \vee 0\right)^2 |q_j|^\alpha \chi(j\Delta x) \, dt \\
& + (1+\alpha) \int_0^T \Delta x \sum_j \left(u_{j-1/2} \wedge 0\right)^2 |q_j|^\alpha \chi(j\Delta x) \, dt.
\end{aligned} \tag{6.4}$$

Next, we introduce the notations

$$\widetilde{\mathcal{N}}(t) = \left\{ j \in \mathbb{Z} \;\middle|\; q_j(t) < 0 \right\}, \qquad \widetilde{\mathcal{P}}(t) = \left\{ j \in \mathbb{Z} \;\middle|\; q_j(t) \ge 0 \right\}$$

and observe that, since $\mathcal{N}(t) \subset \widetilde{\mathcal{N}}(t)$ and $\widetilde{\mathcal{N}}(t) \cup \widetilde{\mathcal{P}}(t) = \mathbb{Z}$, there holds

$$\begin{aligned}
&\int_0^T \Delta x \sum_j \operatorname{sign}(q_j)\, |q_j|^{2+\alpha}\, \chi(j\Delta x)\, dt \\
&\quad = \Delta x \int_0^T \sum_{j \in \mathcal{P}(t)} |q_j|^{2+\alpha}\, dt - \int_0^T \Delta x \sum_{j \in \mathcal{N}(t)} |q_j|^{2+\alpha}\, dt \\
&\quad \le \int_0^T \Delta x \sum_{j \in \mathcal{P}(t)} |q_j|^{2+\alpha}\, dt - I_-.
\end{aligned}$$

Therefore, from (6.4),

$$\frac{1-\alpha}{2} I_- \le I_1 + I_2 + I_3 + I_4, \tag{6.5}$$

where

$$\begin{aligned}
I_1 &= \frac{1-\alpha}{2} \int_0^T \Delta x \sum_{j \in \mathcal{P}(t)} |q_j|^{2+\alpha}\, dt, \\
I_2 &= \Delta x \sum_j \left(|q_j(0)|^{1+\alpha} - |q_j(T)|^{1+\alpha} \right) \chi(j\Delta x), \\
I_3 &= -\int_0^T \Delta x \sum_j D_- \left[\left(u_{j+1/2} \vee 0 \right) |q_j|^{1+\alpha} \right] \chi(j\Delta x)\, dt \\
&\qquad - \int_0^T Dx \sum_j D_+ \left[\left(u_{j-1/2} \wedge 0 \right) |q_j|^{1+\alpha} \right] \chi(j\Delta x)\, dt, \\
I_4 &= -(1+\alpha) \int_0^T \Delta x \sum_j |P_j|\, |q_j|^{\alpha}\, \chi(j\Delta x)\, dt \\
&\qquad + (1+\alpha) \int_0^T \Delta x \sum_j \left(u_{j+1/2} \vee 0 \right)^2 |q_j|^{\alpha}\, \chi(j\Delta x)\, dt \\
&\qquad + (1+\alpha) \int_0^T \Delta x \sum_j \left(u_{j-1/2} \wedge 0 \right)^2 |q_j|^{\alpha}\, \chi(j\Delta x)\, dt.
\end{aligned}$$

For I_1 we repeat what we did for I_+. Indeed, due to (4.1), (5.1), and $\alpha < 1$,

$$I_1 \le \frac{1}{2} \int_0^T \Delta x \sum_{j \in \mathcal{P}(t)} |q_j|^2 \left(\frac{2}{t} + C \right)^{\alpha} dt \le C(T, \alpha) \left\| \{u_{j+1/2}(0)\}_j \right\|_{h^1}^2.$$

For the other terms we use Hölder's inequality for sums, the discrete Leibniz rule,

(4.1), (4.4), and (4.5):

$$
\begin{aligned}
I_2 &\le \left(\left\| \left\{ |q_j(0)|^{1+\alpha} \right\}_j \right\|_{\ell^{\frac{2}{1+\alpha}}} + \left\| \left\{ |q_j(T)|^{1+\alpha} \right\}_j \right\|_{\ell^{\frac{2}{1+\alpha}}} \right) \left\| \{\chi(j\Delta x)\}_j \right\|_{\ell^{\frac{2}{1-\alpha}}} \\
&= \left(\left\| \{q_j(0)\}_j \right\|_{\ell^2}^{1+\alpha} + \left\| \{q_j(T)\}_j \right\|_{\ell^2}^{1+\alpha} \right) \left\| \{\chi(j\Delta x)\}_j \right\|_{\ell^{\frac{2}{1-\alpha}}} \\
&\le C_1 \left\| \{u_j(0)\}_j \right\|_{h^1}^{1+\alpha} \left\| \{\chi(j\Delta x)\}_j \right\|_{\ell^{\frac{2}{1-\alpha}}}, \\
I_3 &= \int_0^T \Delta x \sum_j \left(u_{j+1/2} \vee 0 \right) |q_j|^{1+\alpha} D_- \chi((j-1)\Delta x)\, dt \\
&\qquad + \int_0^T \Delta x \sum_j \left(u_{j-1/2} \wedge 0 \right) |q_j|^{1+\alpha} D_+ \chi((j+1)\Delta x)\, dt \\
&\le \int_0^T \left\| \left\{ u_{j+1/2}(t) \right\}_j \right\|_{\ell^\infty} \left\| \{q_j(t)\}_j \right\|_{\ell^2}^{1+\alpha} \\
&\qquad \times \left(\left\| \{D_-\chi((j-1)\Delta x)\}_j \right\|_{\ell^{\frac{2}{1+\alpha}}} + \left\| \{D_+\chi((j+1)\Delta x) dt\}_j \right\|_{\ell^{\frac{2}{1+\alpha}}} \right) dt \\
&\le C_2 T \left\| \{u_j(0)\}_j \right\|_{h^1}^{2+\alpha} \\
&\qquad \times \left(\left\| \{D_-\chi((j-1)\Delta x)\}_j \right\|_{\ell^{\frac{2}{1+\alpha}}} + \left\| \{D_+\chi((j+1)\Delta x)\}_j \right\|_{\ell^{\frac{2}{1+\alpha}}} \right), \\
I_4 &\le (1+\alpha) \int_0^T \Delta x \sum_j \left(\left\| \{P_j\}_j \right\|_{\ell^\infty} + 2 \left\| \left\{ u_{j+1/2} \right\}_j \right\|_{\ell^\infty}^2 \right) |q_j|^\alpha \chi(j\Delta x)\, dt \\
&\le C_3 \left\| \{u_j(0)\}_j \right\|_{h^1}^2 \int_0^T \left\| \{q_j(t)_j\}_j \right\|_{\ell^2}^\alpha \left\| \{\chi(j\Delta x)\}_j \right\|_{\ell^{\frac{2}{2-\alpha}}} dt \\
&\le C_4 \left\| \{u_j(0)\}_j \right\|_{h^1}^{2+\alpha} T \left\| \{\chi(j\Delta x)\}_j \right\|_{\ell^{\frac{2}{2-\alpha}}},
\end{aligned}
$$

where the constants $C_1, \dots, C_4$ are independent of Δx. Now a bound on I_- follows from (6.5), and thereby the proof is concluded. ∎

7. Basic convergence results. In this section we present some convergence results that are straightforward consequences of the a priori estimates established earlier.

LEMMA 7.1. *There exists a limit function $u \in L^\infty(\mathbb{R}_+; H^1(\mathbb{R})) \cap C([0,\infty) \times \mathbb{R})$ such that along a subsequence as $\Delta x \downarrow 0$*

$$
u_{\Delta x} \overset{\star}{\rightharpoonup} u \text{ in } L^\infty(\mathbb{R}_+; H^1(\mathbb{R})), \tag{7.1}
$$

$$
u_{\Delta x} \to u \text{ uniformly on } [0,T] \times [a,b] \tag{7.2}
$$

for each set $[0,T] \times [a,b] \subset \mathbb{R}^2$. Additionally,

$$
u(t,x) \overset{t \downarrow 0}{\to} u_0(x) \text{ for each } x \in \mathbb{R}. \tag{7.3}
$$

Proof. It is not hard to see that (3.13) implies that

$$
\left| \int_{\mathbb{R}} u_{\Delta x}^2\, dx - \Delta x \sum_j u_{j+1/2}^2 \right| \le \left(\Delta x \sum_j q_j^2 \right) \Delta x, \qquad t \ge 0,
$$

and

$$\int_{\mathbb{R}} (\partial_x u_{\Delta x})^2 \, dx = \Delta x \sum_j q_j^2, \qquad t \geq 0,$$

so, by (4.1),

$$\begin{aligned} &\int_{\mathbb{R}} \Big[(u_{\Delta x})^2 + (\partial_x u_{\Delta x})^2 \Big] \, dx \\ &\leq \int_{\mathbb{R}} \Big[(u_{\Delta x}(0,x))^2 + (\partial_x u_{\Delta x}(0,x))^2 \Big] \, dx + \mathcal{O}(\Delta x) \\ &\leq C \left(\|u_0\|^2_{H^1(\mathbb{R})} + \Delta x \right), \qquad t \geq 0, \end{aligned} \tag{7.4}$$

where C is independent of Δx. Claim (7.1) is an outcome of (7.4).

To verify claim (7.2) we will show that

$$\{u_{\Delta x}\}_{\Delta x > 0} \text{ is bounded in } W^{1,2+\alpha}((0,T) \times (a,b)) \tag{7.5}$$

for any (fixed) set $(0,T) \times (a,b) \subset \mathbb{R}^2$, where α is provided by Lemma 6.1.

Without loss generality, let us assume that $a = j_a \Delta x$ and $b = j_b \Delta x$ for some integers j_a and j_b. Then Lemma 6.1 tells us that

$$\int_0^T \int_a^b |\partial_x u_{\Delta x}|^{2+\alpha} \, dx \, dt = \Delta x \sum_{j=j_a}^{j_b} \int_0^T |q_j(t)|^{2+\alpha} \, dt \leq C \tag{7.6}$$

for some constant $C = C(u_0, \alpha, T, a, b)$ independent of Δx.

By taking into account (3.13), (3.1), and (3.9), there holds for any $x \in I_j$, $j \in \mathbb{Z}$,

$$\begin{aligned} |\partial_t u_{\Delta x}(t,x)| &= \Big| q_j'(t)(x - x_{j-1/2}) + u_{j-1/2}'(t) \Big| \\ &\leq \Big| u_{j-1/2}'(t) \Big| + \big| q_j'(t) \big| \, \Delta x \\ &\leq \big| u_{j+1/2} q_j \big| + \big| u_{j+1/2} q_{j+1} \big| + |D_+ P_j| \\ &\quad + \Delta x \Big(\big| u_{j-1/2} D_- q_j \big| + \big| u_{j+1/2} D_+ q_j \big| \\ &\qquad + \frac{q_j^2}{2} + u_{j+1/2}^2 + u_{j-1/2}^2 + |P_j| \Big). \end{aligned} \tag{7.7}$$

Observe that $\Delta x \, q_j = \mathcal{O}(1)$ for all j, which is clearly true thanks to (3.6) and (4.4). By using this and (4.4) and (4.6) in (7.7) we acquire the pointwise estimate

$$|\partial_t u_{\Delta x}| \leq C \left(1 + |q_{j-1}| + |q_j| + |q_{j+1}|\right) \quad \text{for each } x \in I_j, \, j \in \mathbb{Z}, \tag{7.8}$$

for some constant C independent of Δx. Consequently, in view of (6.3),

$$\int_0^T \int_a^b |\partial_t u_{\Delta x}|^{2+\alpha} \, dx \, dt \leq C \left(1 + \Delta x \sum_{j=j_a}^{j_b} \int_0^T |q_j|^{2+\alpha} \, dt \right) \leq C$$

for some constant $C = C(u_0, \alpha, T, a, b)$ independent of Δx.

In summary, we have proved that (7.5) holds. Since $W^{1,2+\alpha}((0,T)\times(a,b))$ is compactly embedded into $C^{0,\ell}([0,T]\times[a,b])$, with $\ell = 1-2/(2+\alpha)$, there exists a continuous function $u:[0,\infty)\times\mathbb{R}\to\mathbb{R}$ such that along a subsequence

$$u_{\Delta x}\to u \text{ uniformly on } [0,T]\times[a,b] \text{ and pointwise in } \mathbb{R}_+\times\mathbb{R} \text{ as } \Delta x\downarrow 0.$$

By combining this with a diagonal argument we conclude that claim (7.2) is true.

Finally, let us prove that the limit u satisfies the initial condition. We fix an arbitrary point $x_0\in\mathbb{R}$ and let $t\in(0,1)$. Then we proceed as follows:

$$\begin{aligned}|u(t,x_0)-u_0(x_0)| &\le |u(t,x_0)-u_{\Delta x}(t,x_0)|\\ &\quad + |u_{\Delta x}(t,x_0)-u_{\Delta x}(0,x_0)| + |u_{\Delta x}(0,x_0)-u_0(x_0)|\\ &\le |u(t,x_0)-u_{\Delta x}(t,x_0)|\\ &\quad + Ct^{\ell} + |u_{\Delta x}(0,x_0)-u_0(x_0)|\,,\quad \ell = 1-\frac{2}{2+\alpha},\end{aligned}$$

where we used (7.5) to derive the last inequality (C does not depend on Δx). Equipped with (7.2), (1.2), and (3.2), we deduce (7.3) by sending first $\Delta x\downarrow 0$ and second $t\downarrow 0$. This concludes the proof of the lemma. ∎

Equipped with the sequence $\{P_j\}_{j\in\mathbb{Z}}$ defined by (3.1)–(3.6), we introduce the piecewise linear and continuous function

$$P_{\Delta x}(t,x) = P_j(t) + (x-x_j)D_+P_j(t),\quad t\ge 0,\ x\in I_{j+1/2},\ j\in\mathbb{Z}. \tag{7.9}$$

Observe that $\partial_x P_{\Delta x}(t,x) = D_+P_j(t)$ for $t\ge 0$, $x\in I_{j+1/2}$, $j\in\mathbb{Z}$.

LEMMA 7.2. *There exists a limit $P\in L^\infty(\mathbb{R}_+;W^{1,\infty}(\mathbb{R}))\cap L^\infty(\mathbb{R}_+;W^{1,1}(\mathbb{R}))$ such that along a subsequence as $\Delta x\downarrow 0$*

$$P_{\Delta x}\to P \text{ in } L^p_{\mathrm{loc}}(\mathbb{R}_+\times\mathbb{R}) \text{ for each } 1\le p<\infty. \tag{7.10}$$

Proof. First of all, it is not difficult to see from (7.9) and Lemma 4.2 that

$$\{P_{\Delta x}\}_{\Delta x>0} \text{ is bounded in } L^\infty(\mathbb{R}_+;W^{1,\infty}(\mathbb{R}))\cap L^\infty(\mathbb{R}_+;W^{1,1}(\mathbb{R})).$$

Next, we prove that $\{\partial_t P_{\Delta x}\}_{\Delta x>0}$ is bounded in $L^1([0,T]\times\mathbb{R})$ for each fixed $T>0$. To this end, we write $P_j = P_{1,j}+P_{2,j}$ (cf. (3.4)) where

$$\begin{aligned}P_{1,j} &= h\sum_{i\in\mathbb{Z}}\left(e^{-\kappa|j-i|}\right)\left((u_{i+1/2}\vee 0)^2 + (u_{i-1/2}\wedge 0)^2\right),\\ P_{2,j} &= \frac{h}{2}\sum_{i\in\mathbb{Z}}\left(e^{-\kappa|j-i|}\right)q_i^2.\end{aligned}$$

We shall prove that there is a constant C, independent of Δx, such that

$$\int_0^T \Delta x\sum_j\left|\frac{d}{dt}P_{1,j}\right|\,dt\le C, \tag{7.11}$$

$$\int_0^T \Delta x\sum_j\left|\Delta x\,D_+\left(\frac{d}{dt}P_{1,j}\right)\right|\,dt\le C, \tag{7.12}$$

$$\int_0^T \Delta x\sum_j\left|\frac{d}{dt}P_{2,j}\right|\,dt\le C, \tag{7.13}$$

$$\int_0^T \Delta x\sum_j\left|\Delta x\,D_+\left(\frac{d}{dt}P_{2,j}\right)\right|\,dt\le C. \tag{7.14}$$

Note that (7.12) and (7.14) follow from (7.11) and (7.13), respectively, since if $|a_j| \le C$ for all j, then $|\Delta x D_\pm a_j| \le |a_j| + |a_{j\pm1}| \le 2C$ for all j.

To prove (7.11) observe that

$$\begin{aligned} P'_{1,j} &= h\sum_{i\in\mathbb{Z}} \left(e^{-\kappa|j-i|}\right)\left((u_{i+1/2}\vee 0)u'_{i+1/2} + (u_{i-1/2}\wedge 0)u'_{i-1/2}\right) \\ &\le \frac{h}{2}\sum_{i\in\mathbb{Z}} \left(e^{-\kappa|j-i|}\right)\left(u^2_{i+1/2} + (u'_{i+1/2})^2 + u^2_{i-1/2} + (u'_{i-1/2})^2\right), \end{aligned}$$

and thus

$$\begin{aligned} \Delta x\sum_j |P'_{1,j}| &\le h\Delta x\sum_{i,j} \left(e^{-\kappa|j-i|}\right)\left(u^2_{i+1/2} + (u'_{i+1/2})^2 + u^2_{i-1/2} + (u'_{i-1/2})^2\right) \\ &\le C\left(\left\|\left\{e^{-\kappa|j|}\right\}_j\right\|_{\ell^1} + 1\right)\left(\left\|\left\{u_{j+1/2}\right\}_j\right\|^2_{\ell^2} + \left\|\left\{u'_{j+1/2}\right\}_j\right\|^2_{\ell^2}\right), \end{aligned}$$

where $C > 0$ is a constant independent of Δx. From (7.8) and (4.1) it follows that $\{u'_{j+1/2}\}_{j\in\mathbb{Z}}$ is bounded in $L^2(0,T;\ell^2)$, which implies (7.11).

To prove (7.13) we use (6.2) with $f(q) = \frac{q^2}{2}$ to obtain

$$\begin{aligned} P'_{2,j} &= h\sum_{i\in\mathbb{Z}} \left(e^{-\kappa|j-i|}\right)\frac{d}{dt}\left(\frac{q_i^2}{2}\right) \\ &= -h\sum_{i\in\mathbb{Z}} \left(e^{-\kappa|j-i|}\right)\left(D_-\left[\left(u_{i+1/2}\vee 0\right)\frac{q_i^2}{2}\right] + D_+\left[\left(u_{i-1/2}\wedge 0\right)\frac{q_i^2}{2}\right]\right) \\ &\quad - h\sum_{i\in\mathbb{Z}} \left(e^{-\kappa|j-i|}\right)\left[P_i q_i - \left(u_{j+1/2}\vee 0\right)^2 q_i - \left(u_{i-1/2}\wedge 0\right)^2 q_i \right. \\ &\qquad\qquad \left. + \frac{\Delta x}{2}\left[\left(u_{i-1/2}\vee 0\right)(D_- q_i)^2 - \left(u_{i+1/2}\wedge 0\right)(D_+ q_i)^2\right]\right] \\ &= -h\sum_{i\in\mathbb{Z}} \left(D_-\left(e^{-\kappa|j-i|}\right)\right)\left(u_{i-1/2}\vee 0\right)\frac{q_{i-1}^2}{2} \\ &\quad - h\sum_{i\in\mathbb{Z}} \left(D_+\left(e^{-\kappa|j-i|}\right)\right)\left(u_{i+1/2}\wedge 0\right)\frac{q_{i+1}^2}{2} \\ &\quad - h\sum_{i\in\mathbb{Z}} \left(e^{-\kappa|j-i|}\right)\left[P_i q_i - \left(u_{i+1/2}\vee 0\right)^2 q_i - \left(u_{i-1/2}\wedge 0\right)^2 q_i \right. \\ &\qquad\qquad \left. + \frac{\Delta x}{2}\left[\left(u_{i-1/2}\vee 0\right)(D_- q_i)^2 - \left(u_{i+1/2}\wedge 0\right)(D_+ q_i)^2\right]\right], \end{aligned}$$

and hence (7.13) follows from (4.1) and (4.4).

Since $\partial_t P_{\Delta x}(t,x) = P'_j(t) + (x - x_j)D_+ P'_j(t)$, $t \ge 0$, $x \in I_{j+1/2}$, the bound on $\{\partial_t P_{\Delta x}\}_{\Delta x>0}$ follows from (7.11)–(7.14). As a result the sequence $\{P_{\Delta x}\}_{\Delta x>0}$ is bounded in $W^{1,1}_{\mathrm{loc}}(\mathbb{R}_+\times\mathbb{R})$. Combining this with the L^∞-bound in Lemma 4.2 yields the existence of a subsequence that converges as claimed in (7.10). ∎

8. Strong convergence result. Endowed with the sequence $\{q_j(t)\}_{j\in\mathbb{Z}}$ defined in terms of (3.1)–(3.3) and (3.6), we introduce the function

$$q_{\Delta x}(t,x) = q_j(t), \quad t \ge 0,\ x \in I_j,\ j \in \mathbb{Z}. \tag{8.1}$$

Observe that

$$\partial_x u_{\Delta x}(t,x) = q_{\Delta x}(t,x), \quad t \geq 0,\, x \in I_j,\, j \in \mathbb{Z}. \tag{8.2}$$

Additionally, by (3.10) (cf. also (3.13))

$$\|q_0 - q_{\Delta x,0}\|_{L^2(\mathbb{R})} \to 0 \quad \text{as } \Delta x \to 0 \quad (q_0 := \partial_x u_0,\, q_{\Delta x,0} := \partial_x u_{\Delta x}|_{t=0}). \tag{8.3}$$

The ensuing lemma is a simple consequence of the main estimates established in earlier sections.

LEMMA 8.1. *Fix any $1 \leq p < 3$ and $1 \leq r < \frac{3}{2}$. Then there exist two functions $q \in L^p_{\mathrm{loc}}(\mathbb{R}_+ \times \mathbb{R})$, $\overline{q^2} \in L^r_{\mathrm{loc}}(\mathbb{R}_+ \times \mathbb{R})$ such that along a subsequence as $\Delta x \downarrow 0$*

$$q_{\Delta x} \overset{\star}{\rightharpoonup} q \text{ in } L^\infty_{\mathrm{loc}}(\mathbb{R}_+; L^2(\mathbb{R})), \qquad q_{\Delta x} \rightharpoonup q \text{ in } L^p_{\mathrm{loc}}(\mathbb{R}_+ \times \mathbb{R}), \tag{8.4}$$

$$q^2_{\Delta x} \rightharpoonup \overline{q^2} \text{ in } L^r_{\mathrm{loc}}(\mathbb{R}_+ \times \mathbb{R}). \tag{8.5}$$

Moreover,

$$q^2(t,x) \leq \overline{q^2}(t,x) \text{ for a.e. } (t,x) \in \mathbb{R}_+ \times \mathbb{R} \tag{8.6}$$

and

$$\partial_x u = q \quad \text{in the sense of distributions on } [0,\infty) \times \mathbb{R}. \tag{8.7}$$

Finally, there is a positive constant C depending only on $\|u_0\|_{H^1(\mathbb{R})}$ such that

$$q(t,x) \leq \frac{2}{t} + C, \qquad t > 0,\, x \in \mathbb{R}. \tag{8.8}$$

Remark 8.1. Throughout the paper we use overbars to denote weak limits. Below and later we routinely use some well known results related to weak convergence, which are listed, for example, in [24, Chapter 2]).

Proof. Claims (8.4) and (8.5) are direct consequences of (8.1), (3.6), and Lemmas 4.1 and 6.1. Claim (8.6) is true thanks to (8.5), while (8.7) is a consequence of (8.2) and Lemma 7.1. Finally, by (5.1), $q_{\Delta x}(t,x) \leq \frac{2}{t} + C$, $t \geq 0$, $x \in \mathbb{R}$, and hence, because of (8.4), claim (8.8) follows. □

In view of the weak convergences stated in (8.4), we can assume that for any function $f \in C^1(\mathbb{R})$ with f' bounded

$$\begin{aligned} &f(q_{\Delta x}) \overset{\star}{\rightharpoonup} \overline{f(q)} \quad \text{in } L^\infty_{\mathrm{loc}}(\mathbb{R}_+; L^2(\mathbb{R})), \\ &f(q_{\Delta x}) \rightharpoonup \overline{f(q)} \text{ in } L^p_{\mathrm{loc}}(\mathbb{R}_+ \times \mathbb{R}),\, 1 \leq p < 3, \end{aligned} \tag{8.9}$$

where the same subsequence of $\Delta x \downarrow 0$ applies to any f from the specified class.

In what follows, we let $\overline{qf(q)}$ and $\overline{f'(q)q^2}$ denote the weak limits of $q_{\Delta x} f(q_{\Delta x})$ and $f'(q_{\Delta x}) q^2_{\Delta x}$, respectively, in $L^r_{\mathrm{loc}}(\mathbb{R}_+ \times \mathbb{R})$, $1 \leq r < \frac{3}{2}$.

LEMMA 8.2. *For any convex function $f \in C^1(\mathbb{R})$ with f' bounded we have*

$$\begin{aligned} &\iint_{\mathbb{R}_+ \times \mathbb{R}} \left(\overline{f(q)} \partial_t \varphi + u \overline{f(q)} \partial_x \varphi \right) dx\, dt + \int_{\mathbb{R}} f(q_0(x)) \varphi(x,0)\, dx \\ &\qquad \geq \iint_{\mathbb{R}_+ \times \mathbb{R}} \left(\frac{1}{2} \overline{f'(q) q^2} - \overline{qf(q)} + \left(P - u^2\right) \overline{f'(q)} \right) \varphi\, dx\, dt \end{aligned} \tag{8.10}$$

for any nonnegative $\varphi \in C^\infty_c([0,\infty) \times \mathbb{R})$.

Proof. Set $\varphi_j(t) = \frac{1}{\Delta x}\int_{I_j}\varphi(x,t)\,dx$. We multiply (6.2) by $\Delta x\varphi_j$, sum over $j\in\mathbb{Z}$, integrate over $t\in\mathbb{R}_+$, and take into account the convexity of f. After a partial integration and a partial summation, the final result reads

$$\begin{aligned}
&\Delta x\sum_j f(q_j(0))\varphi_j(0) + \int_{\mathbb{R}_+}\Delta x\sum_j f(q_j)\varphi_j'\,dt\\
&+\int_{\mathbb{R}_+}\Delta x\sum_j\Big[\big(u_{j+1/2}\vee 0\big)\,f(q_j)D_+\varphi_j + \big(u_{j-1/2}\wedge 0\big)\,f(q_j)D_-\varphi_j\Big]\,dt\\
&+\int_{\mathbb{R}_+}\Delta x\sum_j\Big[q_jf(q_j) - \frac{q_j^2}{2}f'(q_j)\Big]\varphi_j\,dt\\
&+\int_{\mathbb{R}_+}\Delta x\sum_j\Big[\big(u_{j+1/2}\vee 0\big)^2 + \big(u_{j-1/2}\wedge 0\big)^2 - P_j\Big]f'(q_j)\varphi_j\,dt\ge 0.
\end{aligned}$$

We write this inequality as

$$E_0+E_1+E_2+E_3+E_4\ge 0.$$

By (8.3),

$$\Delta x\sum_j f(q_j(0))\varphi_j(0)\to\int_{\mathbb{R}} f(q_0(x))\varphi(x,0)\,dx\quad\text{as }\Delta x\downarrow 0.$$

Clearly, by the choice of φ_j and (8.9),

$$E_1=\int_{\mathbb{R}_+}\int_{\mathbb{R}} f(q_{\Delta x})\partial_t\varphi\,dx\,dt\to\int_{\mathbb{R}_+}\int_{\mathbb{R}}\overline{f(q)}\partial_t\varphi\,dxdt\quad\text{as }\Delta x\downarrow 0.$$

Next,

$$E_2=\int_{\mathbb{R}_+}\int_{\mathbb{R}} u_{\Delta x}f(q_{\Delta x})\partial_x\varphi\,dx\,dt+E_{2,1}+E_{2,2},$$

where

$$E_{2,1}=\int_{\mathbb{R}_+}\sum_j\int_{I_j}\Big[\big(u_{j+1/2}\vee 0\big)\,f(q_j)D_+\varphi_j-(u_{\Delta x}\vee 0)f(q_{\Delta x})\partial_x\varphi\Big]\,dx\,dt$$

and

$$E_{2,2}=\int_{\mathbb{R}_+}\sum_j\int_{I_j}\Big[\big(u_{j-1/2}\wedge 0\big)\,f(q_j)D_-\varphi_j-(u_{\Delta x}\wedge 0)f(q_{\Delta x})\partial_x\varphi\Big]\,dx\,dt.$$

We decompose $E_{2,1}$ as follows:

$$\begin{aligned}
E_{2,1}=&\int_{\mathbb{R}_+}\sum_j\int_{I_j}\Big[\big(u_{j+1/2}\vee 0\big)-(u_{\Delta x}\vee 0)\Big]f(q_{\Delta x})\partial_x\varphi\,dx\,dt\\
&+\int_{\mathbb{R}_+}\sum_j\int_{I_j}\big(u_{j+1/2}\vee 0\big)\,f(q_{\Delta x})\Big[D_+\phi-\partial_x\varphi\Big]\,dx\,dt=:E_{2,1,1}+E_{2,1,2}.
\end{aligned}$$

For $x \in I_j$, we have by (3.13)

$$\left|\left(u_{j+1/2} \vee 0\right) - \left(u_{\Delta x} \vee 0\right)\right| \leq \left|u_{j+1/2} - u_{\Delta x}\right| \leq C_1 \Delta x \left|q_j\right|. \tag{8.11}$$

Hence, by keeping in mind that $|f(q)| = \mathcal{O}(1+|q|)$ for all $q \in \mathbb{R}$ and using (8.1) and (4.1),

$$|E_{2,1,1}| \leq C_2 \Delta x \left\|\partial_x \varphi\right\|_{L^\infty(\mathbb{R}_+ \times \mathbb{R})} \iint_{\mathrm{supp}(\varphi)} q_{\Delta x}^2 \, dx \, dt \leq C_2 \Delta x,$$

where C_2 depends on φ but not on Δx. For $x \in I_j$, $|D_+\phi - \partial_x \varphi| \leq C_3 \Delta x$, where C_3 depends on φ but not Δx, and thus by (8.1) and (4.1),

$$|E_{2,1,2}| \leq C_4 \Delta x \iint_{\mathrm{supp}(\varphi)} q_{\Delta x} \, dx \, dt \leq C_5 \Delta x,$$

where the final constant C_5 depends on φ but not on Δx. To summarize, we have proved that $|E_{2,1}| = \mathcal{O}(\Delta x) \to 0$ as $\Delta x \downarrow 0$. Similarly, we can prove that $|E_{2,2}| = \mathcal{O}(\Delta x) \to 0$ as $\Delta x \downarrow 0$. Consequently, by (7.2) and (8.9),

$$E_2 \to \int_{\mathbb{R}_+} \int_{\mathbb{R}} u \overline{f(q)} \partial_x \varphi \, dx \, dt \quad \text{as } \Delta x \downarrow 0.$$

We can take the limit in E_3 directly:

$$\begin{aligned} E_3 &= \iint_{\mathbb{R}_+ \times \mathbb{R}} \left[q_{\Delta x} f(q_{\Delta x}) - \frac{q_{\Delta x}^2}{2} f'(q_{\Delta x}) \right] dx \, dt \\ &\overset{\Delta x \downarrow 0}{\longrightarrow} \iint_{\mathbb{R}_+ \times \mathbb{R}} \left[\overline{q f(q)} - \overline{f'(q) q^2} \right] dx \, dt. \end{aligned}$$

Finally, let us analyze E_4, which we write as the sum of four terms:

$$E_4 = \iint_{\mathbb{R}_+ \times \mathbb{R}} \left[u_{\Delta x}^2 - P_{\Delta x} \right] f'(q_{\Delta x}) \varphi \, dx \, dt + E_{4,1} + E_{4,2} - E_{4,3},$$

where

$$\begin{aligned} E_{4,1} &= \int_{\mathbb{R}_+} \sum_j \int_{I_j} \left[\left(u_{j+1/2} \vee 0\right)^2 f'(q_j) \varphi_j - \left(u_{\Delta x} \vee 0\right)^2 f'(q_{\Delta x}) \varphi \right] dx \, dt, \\ E_{4,2} &= \int_{\mathbb{R}_+} \sum_j \int_{I_j} \left[\left(u_{j-1/2} \wedge 0\right)^2 f'(q_j) \varphi_j - \left(u_{\Delta x} \wedge 0\right)^2 f'(q_{\Delta x}) \varphi \right] dx \, dt, \\ E_{4,3} &= \int_{\mathbb{R}_+} \sum_j \int_{I_j} \left[P_j f'(q_j) \varphi_j - P_{\Delta x} f'(q_{\Delta x}) \varphi \right] dx \, dt. \end{aligned}$$

Let us write

$$\begin{aligned} E_{4,1} = &\underbrace{\int_{\mathbb{R}_+} \sum_j \int_{I_j} \left[\left(u_{j+1/2} \vee 0\right)^2 - \left(u_{\Delta x} \vee 0\right)^2 \right] f'(q_{\Delta x}) \varphi \, dx \, dt}_{E_{4,1,1}} \\ &+ \underbrace{\int_{\mathbb{R}_+} \sum_j \int_{I_j} \left(u_{j+1/2} \vee 0\right)^2 f'(q_{\Delta x}) \left[\varphi_j - \varphi \right] dx \, dt}_{E_{4,1,2}}. \end{aligned}$$

In view of (4.4) and (8.11), we have the following estimate for $x \in I_j$:

$$\left|\left(u_{j+1/2} \vee 0\right)^2 - (u_{\Delta x} \vee 0)^2\right| \leq \left|u_{j+1/2} - u_{\Delta x}\right| \, \left|u_{j+1/2} + u_{\Delta x}\right| \leq C\Delta x \, |q_j| \, ,$$

where the constant C does not depend on Δx. Hence we infer that $|E_{4,1,1}| = \mathcal{O}(\Delta x)$ (cf. the treatment of $E_{2,1,1}$). As $|\varphi_j - \varphi| = \mathcal{O}(\Delta x)$, we can argue as we did with $E_{2,1,2}$ to reach the conclusion that $E_{4,1,2} = O(\Delta x)$. Therefore, $E_{4,1} = \mathcal{O}(\Delta x)$.

Along the same lines we can prove that $E_{4,2} = O(\Delta x)$.

Similarly to the estimates of $E_{4,1}$ and $E_{4,2}$ we can show that $E_{4,3} = \mathcal{O}(\Delta x)$, by exploiting (7.9) and (4.6) to conclude that $|P_j - P_{\Delta x}| = \mathcal{O}(\Delta x)$ for $x \in I_{j+1/2}$.

By the previous calculations,

$$E_4 \to \iint_{\mathbb{R}_+ \times \mathbb{R}} \left[u^2 - P\right] \overline{f'(q)} \varphi \, dx \, dt \quad \text{as } \Delta x \downarrow 0.$$

This concludes the proof of (8.10). □

We know that $\{q^2_{\Delta x}\}_{\Delta x > 0}$ is a subset of $L^\infty(\mathbb{R}_+; L^1(\mathbb{R})) \cap L^r_{\mathrm{loc}}(\mathbb{R}_+ \times \mathbb{R})$ for $1 \leq r < \frac{3}{2}$. Additionally, from (6.2) with $f(q) = \frac{q^2}{2}$ (cf. the proof of Lemma 8.2), it follows that $\{\partial_t q^2_{\Delta x}\}_{\Delta x > 0}$ is bounded in $L^1(0, T; W^{-1,1}(\mathbb{R}))$ for each $T > 0$. The Ascoli–Arzelà theorem then implies the following convergence for each $\varphi \in C_c^\infty(\mathbb{R})$:

$$\int_{\mathbb{R}} q^2_{\Delta x} \varphi \, dx \to \int_{\mathbb{R}} \overline{q^2} \varphi \, dx \quad \text{uniformly on compact subsets of } [0, \infty), \tag{8.12}$$

and thus

$$[0, \infty) \ni t \mapsto \int_{\mathbb{R}} \overline{q^2} \varphi \, dx \quad \text{is continuous on } [0, \infty). \tag{8.13}$$

The statements (8.12) and (8.13) hold with $q^2_{\Delta x}$ and $\overline{q^2}$ replaced, respectively, by $f(q_{\Delta x})$ and $\overline{f(q_{\Delta x})}$ for any convex function $f \in C^1(\mathbb{R})$ with f' bounded.

LEMMA 8.3. *Let q and $\overline{q^2}$ be the weak limits identified in Lemma* 8.1. *Then*

$$\begin{aligned} &\iint_{\mathbb{R}_+ \times \mathbb{R}} (q \partial_t \varphi + u q \partial_x \varphi) \, dx \, dt + \int_{\mathbb{R}} q_0(x) \varphi(x, 0) \, dx \\ &\quad = \iint_{\mathbb{R}_+ \times \mathbb{R}} \left(-\frac{1}{2} \overline{q^2} + \left(P - u^2\right) \right) \varphi \, dx \, dt, \quad \varphi \in C_c^\infty([0, \infty) \times \mathbb{R}). \end{aligned} \tag{8.14}$$

Proof. By starting off from (6.2) with $f(q) = q$, we argue as in the proof of Lemma 8.2 to conclude the validity of (8.14). □

The succeeding lemma tells us in which sense the weak limits singled out in Lemma 8.1 satisfy the initial data.

LEMMA 8.4. *Let q and $\overline{q^2}$ be the weak limits identified in Lemma* 8.1. *Then*

$$\lim_{t \downarrow 0} \int_{\mathbb{R}} q^2(t, x) \, dx = \int_{\mathbb{R}} (\partial_x u_0)^2 \, dx, \quad \lim_{t \downarrow 0} \int_{\mathbb{R}} \overline{q^2}(t, x) \, dx = \int_{\mathbb{R}} (\partial_x u_0)^2 \, dx. \tag{8.15}$$

Proof. In view of previous results, $q \in C\left(0, T; L^2_{\mathrm{weak}}(\mathbb{R})\right)$. Hence, it follows from (8.14) that $q(t, \cdot) \rightharpoonup q_0$ $(= \partial_x u_0)$ in $L^2(\mathbb{R})$ as $t \downarrow 0$. Consequently,

$$\int_{\mathbb{R}} (\partial_x u_0)^2 \, dx \leq \liminf_{t \downarrow 0} \int_{\mathbb{R}} q^2(t, x) \, dx. \tag{8.16}$$

On the other hand, (7.1) tells us that $u_{\Delta x}(t,\cdot) \rightharpoonup u(t,\cdot)$ in $H^1(\mathbb{R})$ for a.e. $t>0$, and thereby, by using also (7.2), (7.4), and (8.5),

$$\int_{\mathbb{R}} (u(t,x))^2\,dx + \int_{\mathbb{R}} \overline{q^2}(t,x)\,dx \le \int_{\mathbb{R}} u_0^2\,dx + \int_{\mathbb{R}} (\partial_x u_0)^2\,dx. \tag{8.17}$$

Since (8.13) holds, this inequality is valid for all $t>0$. By exploiting the continuity of u (see Lemma 7.1), (8.17) yields

$$\limsup_{t\downarrow 0} \int_{\mathbb{R}} \overline{q^2}(t,x)\,dx \le \int_{\mathbb{R}} \left(\partial_x u_0\right)^2\,dx. \tag{8.18}$$

Clearly, (8.6), (8.16), and (8.18) imply (8.15). ∎

We are now in a position to conclude the strong convergence of $\{q_{\Delta x}\}_{\Delta x>0}$.

LEMMA 8.5. *Let q and $\overline{q^2}$ be the weak limits identified in Lemma* 8.1. *Then*

$$\overline{q^2}(t,x) = q^2(t,x) \text{ for a.e. } (t,x)\in\mathbb{R}_+\times\mathbb{R}. \tag{8.19}$$

Consequently, as $\Delta x\downarrow 0$,

$$q_{\Delta x}\to q \text{ in } L^2_{\mathrm{loc}}(\mathbb{R}_+\times\mathbb{R}) \text{ and a.e. in } \mathbb{R}_+\times\mathbb{R}. \tag{8.20}$$

Proof. Lemma 8.2 tells us that for any convex function $f\in C^1(\mathbb{R})$ with f' bounded there holds

$$\partial_t \overline{f(q)} + \partial_x\left(u\overline{f(q)}\right) \le \overline{qf(q)} - \frac{1}{2}\overline{f'(q)q^2} + \left(u^2-P\right)\overline{f'(q)}, \tag{8.21}$$

in the sense of distributions on $\mathbb{R}_+\times\mathbb{R}$. Moreover, by Lemma 8.3,

$$\partial_t q + \partial_x(uq) = \frac{1}{2}\overline{q^2} + u^2 - P, \tag{8.22}$$

in the sense of distributions on $\mathbb{R}_+\times\mathbb{R}$. Equipped with (8.21), (8.22), Lemma 8.15, and (8.8), we can argue exactly as in Xin and Zhang [38] to arrive at (8.19). In view of standard functional analysis considerations, claim (8.20) now follows immediately from (8.19) and (7.6). ∎

9. Concluding the proof of Theorem 3.1. Lemmas 7.1 and 8.1 and (3.12) show that the strong H^1-limit u satisfies conditions (i), (iii), (iv), and (3.12) of Definition 3.1. It remains to prove that u satisfies condition (ii), i.e., the weak formulation (3.1).

We start by deriving a divergence-form version of the scheme (3.1). To this end, introduce the functions $f_\vee$, $f_\wedge$ defined by $f_\vee'(u)=u\vee 0$, $f_\vee(0)=0$, $f_\wedge'(u)=u\wedge 0$, $f_\wedge(0)=0$, i.e., $f_\vee(u)=\frac{1}{2}(u\vee 0)^2$ and $f_\wedge(u)=\frac{1}{2}(u\wedge 0)^2$. Observe that $f_\vee, f_\wedge$ are piecewise C^2, and the absolute values of the second derivatives are bounded (by 1). By the discrete chain rule,

$$\left(u_{j+1/2}\vee 0\right) D_- u_{j+1/2} = D_- f_\vee(u_{j+1/2}) + \mathcal{O}\left(\Delta x\left(D_- u_{j+1/2}\right)^2\right)$$

and

$$\left(u_{j+1/2}\wedge 0\right) D_+ u_{j+1/2} = D_+ f_\wedge(u_{j+1/2}) + \mathcal{O}\left(\Delta x\left(D_+ u_{j+1/2}\right)^2\right).$$

Consequently, we can replace (3.1) by

$$\begin{aligned} &u'_{j+1/2} + D_- f_\vee(u_{j+1/2}) + D_+ f_\wedge(u_{j+1/2}) + D_+ P_j \\ &\qquad = \mathcal{O}\left(\Delta x \left\{ \left(D_- u_{j+1/2}\right)^2 + \left(D_+ u_{j+1/2}\right)^2 \right\}\right). \end{aligned} \tag{9.1}$$

We can restate (9.1) as

$$\begin{aligned} &u'_{j+1/2} + D_- \left[\frac{u_{j+1/2}^2}{4} + \frac{1}{2}\left(f_\vee(u_{j+1/2}) - f_\wedge(u_{j+1/2}) \right) \right] \\ &\quad + D_+ \left[\frac{u_{j+1/2}^2}{4} + \frac{1}{2}\left(f_\wedge(u_{j+1/2}) - f_\vee(u_{j+1/2}) \right) \right] + D_+ P_j \\ &\qquad = \mathcal{O}\left(\Delta x \left\{ \left(D_- u_{j+1/2}\right)^2 + \left(D_+ u_{j+1/2}\right)^2 \right\}\right). \end{aligned} \tag{9.2}$$

By using the relation

$$\begin{aligned} &D_-\left(f_\vee(u_{j+1/2}) - f_\wedge(u_{j+1/2}) \right) + D_+\left(f_\wedge(u_{j+1/2}) - f_\vee(u_{j+1/2}) \right) \\ &\quad = \Delta x D_- D_+ f_\wedge(u_{j+1/2}) - \Delta x D_- D_+ f_\vee(u_{j+1/2}), \end{aligned}$$

(9.2) becomes

$$\begin{aligned} &u'_{j+1/2} + D\left(\frac{u_{j+1/2}^2}{2} \right) + D_+ P_j \\ &\qquad = \mathcal{O}\left(\Delta x \left\{ \left(D_- u_{j+1/2}\right)^2 + \left(D_+ u_{j+1/2}\right)^2 \right\}\right) \\ &\qquad\quad + \Delta x \left\{ D_- D_+ f_\vee(u_{j+1/2}) - D_- D_+ f_\wedge(u_{j+1/2}) \right\}. \end{aligned} \tag{9.3}$$

Fix $\varphi \in C_c^\infty(\mathbb{R}_+ \times \mathbb{R})$, and set $\varphi_j(t) = \frac{1}{\Delta x}\int_{I_j} \varphi(x,t)\,dx$. We multiply (9.3) by $\Delta x \varphi_j$, sum over $j \in \mathbb{Z}$, and integrate over $t \in \mathbb{R}_+$. After a partial integration and a partial summation, the final result reads

$$\begin{aligned} &\underbrace{\int_{\mathbb{R}_+} \Delta x \sum_j u_{j+1/2} \varphi'_j \,dt}_{=:E_1} + \underbrace{\int_{\mathbb{R}_+} \Delta x \sum_j \frac{u_{j+1/2}^2}{2} D\varphi_j \,dt}_{=:E_2} \\ &\qquad + \underbrace{\int_{\mathbb{R}_+} \Delta x \sum_j P_j D_- \varphi_j \,dt}_{=:E_3} = \mathcal{O}(\Delta x), \end{aligned} \tag{9.4}$$

where the right-hand side is a consequence of (4.1).

First, since $\left|u_{j+1/2} - u_{\Delta x}\right| \le C\Delta x\,|q_j|$ (cf. (3.13)) and by Lemmas 4.1 and 7.1,

$$\begin{aligned} E_1 &= \int_{\mathbb{R}_+}\int_{\mathbb{R}} u_{\Delta x} \partial_t \varphi \,dt + \int_{\mathbb{R}_+} \sum_j \int_{I_j} \left(u_{j+1/2} - u_{\Delta x}\right) \partial_t \varphi \,dx\,dt \\ &= \iint_{\mathbb{R}_+\times\mathbb{R}} u_{\Delta x} \partial_t \varphi \,dx\,dt + \mathcal{O}(\Delta x) \overset{\Delta x \downarrow 0}{\longrightarrow} \iint_{\mathbb{R}_+\times\mathbb{R}} u \partial_t \varphi \,dx\,dt. \end{aligned}$$

Next,

$$E_2 = \underbrace{\iint_{\mathbb{R}_+\times\mathbb{R}} \frac{u_{\Delta x}^2}{2} u_{\Delta x} \partial_t \varphi \, dt}_{E_{2,1}} + \underbrace{\int_{\mathbb{R}_+} \sum_j \int_{I_j} \left(\frac{u_{j+1/2}^2}{2} D\varphi_j - \frac{u_{\Delta x}^2}{2} \partial_x \varphi \right) dx \, dt}_{E_{2,2}}.$$

Let us analyze the term $E_{2,2}$. We have $E_{2,2} = E_{2,2,1} + E_{2,2,2}$, where

$$E_{2,2,1} = \int_{\mathbb{R}_+} \sum_j \int_{I_j} \left(\frac{u_{j+1/2}^2}{2} - \frac{u_{\Delta x}^2}{2} \right) \partial_x \varphi \, dx \, dt,$$

$$E_{2,2,2} = \int_{\mathbb{R}_+} \sum_j \int_{I_j} \frac{u_{j+1/2}^2}{2} \left(D\varphi_j - \partial_x \varphi \right) dx \, dt.$$

Since, by (4.4),

$$\left| \frac{u_{j+1/2}^2}{2} - \frac{u_{\Delta x}^2}{2} \right| \le C \Delta x \left| q_j \right|, \quad \left| D\varphi_j - \partial_x \varphi \right| \le C \Delta x, \quad x \in I_j, \, j \in \mathbb{Z},$$

where C does not depend on Δx, we use again (4.1) to conclude that $|E_{2,2}| = \mathcal{O}(\Delta x)$.

It remains to analyze E_3. By (3.4), we have

$$E_3 = \int_{\mathbb{R}_+} h \Delta x \sum_{j,i} \left(e^{-\kappa |j-i|} \right) \left((u_{i+1/2} \vee 0)^2 + (u_{i-1/2} \wedge 0)^2 + \frac{q_i^2}{2} \right) D_- \varphi_j \, dt.$$

By Lemmas 7.1, 8.5, and 8.1,

$$\lim_{\Delta x \downarrow 0} E_3 = \frac{1}{2} \iint_{\mathbb{R}_+\times\mathbb{R}} \left[\int_{\mathbb{R}} e^{-|x-y|} \left((u(t,y))^2 + \frac{1}{2} (q(t,y))^2 \right) dy \right] \partial_x \varphi(t,x) \, dx \, dt.$$

In summary, our calculations show that by sending $\Delta x \downarrow 0$ in (9.4) we obtain (3.11). This concludes the proof of Theorem 3.1. □

REFERENCES

[1] R. Artebrant and H. J. Schroll, *Numerical simulation of Camassa-Holm peakons by adaptive upwinding*, Appl. Numer. Math., 56 (2006), pp. 695–711.

[2] R. Beals, D. H. Sattinger, and J. Szmigielski, *Multipeakons and the classical moment problem*, Adv. Math., 154 (2000), pp. 229–257.

[3] A. Bressan and A. Constantin, *Global dissipative solutions of the Camassa-Holm equation*, Anal. Appl. (Singap.), 5 (2007), pp. 1–27.

[4] A. Bressan and A. Constantin, *Global conservative solutions of the Camassa-Holm equation*, Arch. Ration. Mech. Anal., 183 (2007), pp. 215–239.

[5] A. Bressan and M. Fonte, *An optimal transportation metric for solutions of the Camassa-Holm equation*, Methods Appl. Anal., 12 (2005), pp. 191–219.

[6] R. Camassa, *Characteristics and the initial value problem of a completely integrable shallow water equation*, Discrete Contin. Dyn. Syst. Ser. B, 3 (2003), pp. 115–139.

[7] R. Camassa and D. D. Holm, *An integrable shallow water equation with peaked solitons*, Phys. Rev. Lett., 71 (1993), pp. 1661–1664.

[8] R. Camassa, D. D. Holm, and J. Hyman, *A new integrable shallow water equation*, Adv. Appl. Mech, 31 (1994), pp. 1–33.

[9] R. Camassa, J. Huang, and L. Lee, *On a completely integrable numerical scheme for a nonlinear shallow-water wave equation*, J. Nonlinear Math. Phys., 12 (2005), pp. 146–162.

[10] R. Camassa, J. Huang, and L. Lee, *Integral and integrable algorithms for a nonlinear shallow-water wave equation*, J. Comput. Phys., to appear.
[11] G. M. Coclite, H. Holden, and K. H. Karlsen, *Global weak solutions to a generalized hyperelastic-rod wave equation*, SIAM J. Math. Anal., 37 (2005), pp. 1044–1069.
[12] G. M. Coclite, H. Holden, and K. H. Karlsen, *Wellposedness for a parabolic-elliptic system*, Discrete Contin. Dyn. Syst., 13 (2005), pp. 659–682.
[13] A. Constantin, *On the scattering problem for the Camassa-Holm equation*, Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 457 (2001), pp. 953–970.
[14] A. Constantin and J. Escher, *Global existence and blow-up for a shallow water equation*, Ann. Sc. Norm. Super. Pisa Cl. Sci. (4), 26 (1998), pp. 303–328.
[15] A. Constantin and J. Escher, *Global weak solutions for a shallow water equation*, Indiana Univ. Math. J., 47 (1998), pp. 1527–1545.
[16] A. Constantin and J. Escher, *Wave breaking for nonlinear nonlocal shallow water equations*, Acta Math., 181 (1998), pp. 229–243.
[17] A. Constantin and B. Kolev, *On the geometric approach to the motion of inertial mechanical systems*, J. Phys. A, 35 (2002), pp. R51–R79.
[18] A. Constantin and B. Kolev, *Geodesic flow on the diffeomorphism group of the circle*, Comment. Math. Helv., 78 (2003), pp. 787–804.
[19] A. Constantin and H. P. McKean, *A shallow water equation on the circle*, Comm. Pure Appl. Math., 52 (1999), pp. 949–982.
[20] A. Constantin and L. Molinet, *Global weak solutions for a shallow water equation*, Comm. Math. Phys., 211 (2000), pp. 45–61.
[21] H. H. Dai, *Model equations for nonlinear dispersive waves in a compressible Mooney–Rivlin rod*, Acta Mech., 127 (1998), pp. 193–207.
[22] R. Danchin, *A few remarks on the Camassa-Holm equation*, Differential Integral Equations, 14 (2001), pp. 953–988.
[23] R. Danchin, *A note on well-posedness for Camassa-Holm equation*, J. Differential Equations, 192 (2003), pp. 429–444.
[24] E. Feireisl, *Dynamics of Viscous Compressible Fluids*, Oxford Lecture Ser. Math. Appl. 26, Oxford University Press, Oxford, 2004.
[25] O. B. Fringer and D. D. Holm, *Integrable vs. nonintegrable geodesic soliton behavior*, Phys. D, 150 (2001), pp. 237–263.
[26] B. Fuchssteiner and A. S. Fokas, *Symplectic structures, their Bäcklund transformations and hereditary symmetries*, Phys. D, 4 (1981/82), pp. 47–66.
[27] H. Holden and X. Raynaud, *Convergence of a finite difference scheme for the Camassa–Holm equation*, SIAM J. Numer. Anal., 44 (2006), pp. 1655–1680.
[28] H. Holden and X. Raynaud, *A convergent numerical scheme for the Camassa-Holm equation based on multipeakons*, Discrete Contin. Dyn. Syst., 14 (2006), pp. 505–523.
[29] H. Holden and X. Raynaud, *Global conservative solutions of the Camassa-Holm equation - A Lagrangian point of view*, Comm. Partial Differential Equations, to appear.
[30] D. D. Holm, J. E. Marsden, and T. S. Ratiu, *The Euler-Poincaré equations and semidirect products with applications to continuum theories*, Adv. Math., 137 (1998), pp. 1–81.
[31] D. D. Holm and M. F. Staley, *Wave structure and nonlinear balances in a family of evolutionary PDEs*, SIAM J. Appl. Dyn. Syst., 2 (2003), pp. 323–380.
[32] R. S. Johnson, *Camassa-Holm, Korteweg-de Vries and related models for water waves*, J. Fluid Mech., 455 (2002), pp. 63–82.
[33] H. Kalisch and X. Raynaud, *Convergence of a spectral projection of the Camassa-Holm equation*, Numer. Methods Partial Differential Equations, to appear.
[34] R. J. LeVeque, *Finite Volume Methods for Hyperbolic Problems*, Cambridge Texts Appl. Math., Cambridge University Press, Cambridge, 2002.
[35] Y. A. Li and P. J. Olver, *Well-posedness and blow-up solutions for an integrable nonlinearly dispersive model wave equation*, J. Differential Equations, 162 (2000), pp. 27–63.
[36] G. Misiołek, *A shallow water equation as a geodesic flow on the Bott-Virasoro group*, J. Geom. Phys., 24 (1998), pp. 203–208.
[37] G. Rodríguez-Blanco, *On the Cauchy problem for the Camassa-Holm equation*, Nonlinear Anal., 46 (2001), pp. 309–327.
[38] Z. Xin and P. Zhang, *On the weak solutions to a shallow water equation*, Comm. Pure Appl. Math., 53 (2000), pp. 1411–1433.
[39] Z. Xin and P. Zhang, *On the uniqueness and large time behavior of the weak solutions to a shallow water equation*, Comm. Partial Differential Equations, 27 (2002), pp. 1815–1844.

SIAM J. NUMER. ANAL.
Vol. 46, No. 3, pp. 1580–1601
© 2008 Society for Industrial and Applied Mathematics

MAXIMAL USE OF CENTRAL DIFFERENCING FOR HAMILTON–JACOBI–BELLMAN PDEs IN FINANCE*

J. WANG† AND P. A. FORSYTH†

Abstract. In order to ensure convergence to the viscosity solution, the standard method for discretizing Hamilton–Jacobi–Bellman partial differential equations uses forward/backward differencing for the drift term. In this paper, we devise a monotone method which uses central weighting as much as possible. In order to solve the discretized algebraic equations, we have to maximize a possibly discontinuous objective function at each node. Nevertheless, convergence of the overall iteration can be guaranteed. Numerical experiments on two examples from the finance literature show higher rates of convergence for this approach compared to the use of forward/backward differencing only.

Key words. stochastic control, nonlinear Hamilton–Jacobi–Bellman partial differential equations, central differencing, monotone scheme, finance

AMS subject classifications. 65M12, 65M60, 91B28

DOI. 10.1137/060675186

1. Introduction.

There are a number of financial models which result in nonlinear Hamilton–Jacobi–Bellman (HJB) partial differential equations (PDEs). These problems usually arise in the context of optimal stochastic control. Some examples of these HJB-type equations include transaction cost/uncertain volatility models [21, 3, 27], passport options [2, 29], unequal borrowing/lending costs [12], large investor effects [1], risk control in reinsurance [23], pricing options and insurance in incomplete markets using an instantaneous Sharpe ratio [31, 22, 10], minimizing ruin probability in insurance [28, 13], and optimal consumption [11, 15]. A recent survey article on the theoretical aspects of this topic is given in [25].

In many cases, classical solutions to these PDEs do not exist, and we seek to find the viscosity solution of the HJB equation [17]. It is important to ensure that the discrete solution converges to the viscosity solution as the mesh size and timestep are reduced (see [27] for an example where seemingly reasonable discretization methods can converge to nonviscosity solutions). In order to guarantee convergence to the viscosity solution, the discrete scheme must be pointwise consistent, l_∞ stable, and monotone [9, 4].

A monotone scheme is usually constructed by using a positive coefficient method [20, 24, 7, 18]. Typically, a positive coefficient method is developed using forward or backward differencing for the drift term. The choice of forward or backward differencing depends on the control variable. This has the disadvantage that the truncation error in the space-like direction is only first order.

If implicit timestepping is used, then the nonlinear discretized algebraic equations are solved by an iterative method. The usual iterative approach [20, 18] requires solution of a local optimization problem for the optimal control at each grid node and at every iteration. Since the discretization at each node is a function of the control

*Received by the editors November 17, 2006; accepted for publication (in revised form) August 17, 2007; published electronically April 11, 2008. This work was supported by the Natural Sciences and Engineering Research Council of Canada.
http://www.siam.org/journals/sinum/46-3/67518.html
†Cheriton School of Computer Science, University of Waterloo, Waterloo, ON, N2L 3G1, Canada (j27wang@uwaterloo.ca, paforsyt@uwaterloo.ca).

variable at that node, the type of discretization (i.e., forward or backward differencing) may change at each iteration. Use of forward/backward differencing means that the local objective function at each node is a continuous function of the control variable but nonsmooth.

In this paper, we take a slightly different approach compared to the standard technique. We will use a combination of central/forward/backward differencing at each node. Given a value of the control variable at a node, we use the following criteria to select the differencing method:

- Central differencing is used if the discretization is a positive coefficient method (for this particular choice of control).
- Forward/backward differencing is used only if central differencing does not result in a positive coefficient method. One of forward or backward differencing must satisfy the positive coefficient condition.

This method has the advantage that central differencing is used as much as possible, so that use of a locally second order method is maximized. However, in general, the local objective function at each grid node is now a discontinuous function of the control. It would appear that iterative solution of the discretized equations would be problematic in this case.

In this paper we note that the proof of convergence of the iterative scheme for solution of the discretized algebraic equations does not, in fact, require continuity of the local objective function. Hence, convergence of the iterative method for solution of the discrete equations can be guaranteed, even if the local objective function is a discontinuous function of the control. Nevertheless, it is not clear that, in practice,

- the use of a locally second order method as much as possible will result in improved convergence as the mesh is refined for practical parameter values;
- the rate of convergence of the nonlinear iteration will be acceptable if the local objective function is a discontinuous function of the control.

We report the results of several numerical experiments for passport options [2] and optimal asset allocation [14]. These experiments show that we can often obtain higher rates of convergence using central differencing as much as possible, although at some additional cost compared to the standard approach.

2. General form for the example problems. To avoid repetition, we will carry out our analysis for a general form for the example problems.

As is typically the case with finance problems, we solve backwards in time from the expiry date of the contract $t = T$ to $t = 0$ by use of the variable $\tau = T - t$. Set

$$\mathcal{L}^Q V \equiv a(S,\tau,Q)V_{SS} + b(S,\tau,Q)V_S - c(S,\tau,Q)V, \tag{2.1}$$

where the control parameter Q is in general a vector, that is, $Q = (q_1, q_2, \dots)'$. We write our problems in the general form

$$V_\tau = \sup_{Q\in\hat{Q}} \left\{ \mathcal{L}^Q V + d(S,\tau,Q) \right\},$$

$$S \in [S_{\min}, S_{\max}], \quad 0 \le \tau \le T. \tag{2.2}$$

Here we include the $d(S,t,Q)$ term in (2.2) for generality, although in the examples in this paper, we will always have $d \equiv 0$. We can also replace the sup in (2.2) by an inf, and all the results of this paper hold in this case as well.

2.1. Boundary conditions. We assume the boundary and payoff conditions satisfy the following conditions.

Assumption 2.1 (payoff and boundary conditions). At $\tau = 0$, we set $V(S, 0)$ to the specified contract payoff. As $S \to S_{\min}, S \to S_{\max}$, we assume that either

- a Dirichlet boundary condition is specified and $a(S, \tau, Q) > 0$;
- the coefficient $a(S, \tau, Q)$ vanishes, and the sign of $b(S, \tau, Q)$ is such that no boundary condition is required [8]; i.e., the boundary condition is the limit of the PDE from the interior.

Note that it may be the case that the original problem has $S_{\max} = +\infty$ or $S_{\min} = -\infty$. In these cases, we will use a finite computational domain, and we assume that financial reasoning can be used to determine an appropriate Dirichlet condition. This is clearly an approximation and introduces a *localization error*. However, as pointed out in [6], we can expect any errors incurred by imposing approximate boundary conditions at finite values of $S_{\min}, S_{\max}$ to be small in areas of interest if $|S_{\min}|$, $|S_{\max}|$ are selected sufficiently large. We will assume in the following that the original problem has been localized to a finite domain.

An alternative approach is introduced by [32] for handling the infinite domain problem, in which reversion conditions for stochastic models are defined. At $S = S_{\min}$, a reversion condition would be $a(S_{\min}, \tau, Q) = 0$ and $b(S_{\min}, \tau, Q) \geq 0$. The authors in [32] prove that if the models satisfy reversion conditions, then the financial PDEs have a unique solution and their numerical solutions can be obtained without using any artificial boundary conditions. However, as discussed in [32], in general this may require modification of the PDE coefficients for extreme values of S.

Assumption 2.2 (properties of the HJB PDE). We make the assumption that the coefficients a, b, c, d are Lipschitz continuous functions of (S, τ, Q), with $a \geq 0$, and $c \geq 0$, that a, b, c, d are bounded on $S_{\min} \leq S \leq S_{max}$, and that the set of admissible controls $\hat{Q}$ is compact.

Remark 2.1. Since we restrict ourselves to a finite computational domain $S_{\min} \leq S \leq S_{max}$, we avoid difficulties associated with coefficients that grow with S as $|S| \to \infty$. It follows from [16, 8] that solutions to (2.2), satisfying Assumptions 2.1 and 2.2, satisfy the strong comparison property. Hence, a unique viscosity solution exists for (2.2).

3. Implicit controls. Define a grid $\{S_0, S_1, \ldots, S_p\}$ with $S_0 = S_{\min}$, $S_p = S_{\max}$, and let V_i^n be a discrete approximation to $V(S_i, \tau^n)$. Let $V^n = [V_0^n, \ldots, V_p^n]'$, and let $(\mathcal{L}_h^Q V^n)_i$ denote the discrete form of the differential operator (2.1) at node (S_i, τ^n). The operator (2.1) can be discretized using forward, backward, or central differencing in the S direction to give

$$(\mathcal{L}_h^Q V^{n+1})_i = \alpha_i^{n+1} V_{i-1}^{n+1} + \beta_i^{n+1} V_{i+1}^{n+1} - (\alpha_i^{n+1} + \beta_i^{n+1} + c_i^{n+1}) V_i^{n+1}. \tag{3.1}$$

Here α_i, β_i are defined in Appendix A.

It is important that central, forward, or backward discretizations be used to ensure that (3.3) is a positive coefficient discretization. To be more precise, this condition is given as follows.

Condition 3.1 (positive coefficient condition).

$$\alpha_i^{n+1} \geq 0, \quad \beta_i^{n+1} \geq 0, \quad c_i^{n+1} \geq 0, \quad i = 0, \ldots, p-1. \tag{3.2}$$

We will assume that all models have $c_i^{n+1} \geq 0$. Consequently, we choose central, forward, or backward differencing at each node to ensure that $\alpha_i^{n+1},\ \beta_i^{n+1} \geq 0$. Note that different nodes can have different discretization schemes.

Equation (2.2) can now be discretized using fully implicit timestepping along with the discretization (3.1) to give

$$\frac{V_i^{n+1} - V_i^n}{\Delta\tau} = \sup_{Q^{n+1} \in \hat{Q}} \left\{ (\mathcal{L}_h^{Q^{n+1}} V^{n+1})_i + d_i^{n+1} \right\}. \tag{3.3}$$

Note that $\alpha_i^{n+1} = \alpha_i^{n+1}(Q_i^{n+1})$, $\beta_i^{n+1} = \beta_i^{n+1}(Q_i^{n+1})$, $c_i^{n+1} = c_i^{n+1}(Q_i^{n+1})$, and $d_i^{n+1} = d_i^{n+1}(Q_i^{n+1})$; that is, the discrete equation coefficients are functions of the local optimal control Q_i^{n+1}. This makes (3.3) highly nonlinear in general. We refer to methods which use an implicit timestepping method where the control is handled implicitly as an *implicit control* method in the following.

3.1. Matrix form of the discrete equations. It will be convenient to use matrix notation for (3.3), coupled with boundary conditions.

If a Dirichlet condition is specified at $S = S_{\min}, \tau = \tau^n$ $(i = 0)$, then we denote this value by G_0^n. If a Dirichlet boundary condition is specified at $S = S_{\max}, \tau = \tau^n$ $(i = p)$, then we denote this value by G_p^n. Set $Q^n = [Q_0^n, Q_1^n, \ldots, Q_p^n]'$, with each Q_i^n a local optimal control. We can write the discrete operator $(\mathcal{L}_h^Q V^n)_i$ as

$$\begin{aligned} (\mathcal{L}_h^Q V^n)_i &= [A^n(Q^n)V^n]_i \\ &= \left[\alpha_i^n V_{i-1}^n + \beta_i^n V_{i+1}^n - (\alpha_i^n + \beta_i^n + c_i^n) V_i^n\right], \qquad 1 < i < p. \end{aligned} \tag{3.4}$$

The first and last rows of A are modified as needed to handle the boundary conditions. The boundary conditions at $S = S_{\min}, S_{\max}$ can be enforced by specifying a boundary condition vector $G^n = [G_0^n, 0, \ldots, 0, G_p^n]'$. If a Dirichlet condition is specified at $i = p$, we set G_p^n to the appropriate value and set the last row in A^n to be zero. With a slight abuse of notation, we denote this last row in this case as $(A^n)_p \equiv 0$. Conversely, if no boundary condition is required at $i = p$, then we use backward differencing at node $i = p$ (which means that $\beta_p = 0$) and set $G_p^n = 0$. The boundary condition at $S = S_{\min}$, $i = 0$, is handled in a similar fashion. Let D^n be the vector with entries

$$[D]_i^n = \begin{cases} d_i^n, & 1 < i < p, \\ 0, & i = 0, p \quad \text{if a Dirichlet condition is specified,} \\ d_i^n, & i = 0, p \quad \text{if no boundary condition is required.} \end{cases}$$

Let

$$F^n(Q, V^n) = A^n(Q)V^n + D^n(Q). \tag{3.5}$$

As we shall see in section 6.3, $[F^n(Q, V^n)]_i$, regarded as a function of Q, can have a finite number of discontinuities at points where the differencing scheme changes. Recall that an upper semicontinuous ($\mathcal{USC}$) function $g(x)$ has the property that

$$\limsup_{x \to y} g(x) \leq g(y). \tag{3.6}$$

In particular, we will define $F(Q, V^n)$ so that it is an $\mathcal{USC}$ function of Q, i.e.,

$$\limsup_{X \to Q} [F^n(X, V^n)]_i \ \leq [F^n(Q, V^n)]_i, \tag{3.7}$$

where

$$F^n(Q, V^n) = A^n(Q, V^n)V^n + D^n(Q), \tag{3.8}$$

and we now note that the definition of A^n may depend on V^n. We will give an algorithm for determining $A^n(Q, V^n)$ for a specific case so that $F^n(Q, V^n)$ is $\mathcal{USC}$ in section 6.2.

Since $F^n(Q, V^n)$ is $\mathcal{USC}$, and $\hat{Q}$ is compact, the discrete equations (3.3) can then be written as

$$\left[I - \Delta\tau A^{n+1}(Q^{n+1}, V^{n+1})\right] V^{n+1} = V^n + \Delta\tau D^{n+1} + (G^{n+1} - G^n),$$

$$\text{where } Q_i^{n+1} = \arg\max_{Q\in\hat{Q}} \left\{ \left[F^{n+1}(Q, V^{n+1})\right]_i \right\}. \tag{3.9}$$

Here the term $(G^{n+1} - G^n)$ enforces possible Dirichlet boundary conditions at $S = S_0, S_p$. Note also that the discrete equations (3.9) are nonlinear since $Q^{n+1} = Q^{n+1}(V^{n+1})$.

4. Convergence to the viscosity solution. In [27], examples were given in which seemingly reasonable discretizations of nonlinear option pricing PDEs were unstable or converged to the incorrect solution. It is important to ensure that we can generate discretizations which are guaranteed to converge to the viscosity solution [4, 17]. Assuming that (2.2) satisfies the strong comparison property [5, 8, 16], then, from [9, 4], a numerical scheme converges to the viscosity solution if the method is pointwise consistent, stable (in the l_∞ norm), and monotone.

It is straightforward, using the methods in [7, 18], to show that scheme (3.3) is monotone, pointwise consistent, and stable.

THEOREM 4.1 (convergence to the viscosity solution). *Provided that the original HJB satisfies Assumptions* 2.1 *and* 2.2 *and discretization* (3.9) *satisfies the positive coefficient Condition* 3.1*, then scheme* (3.9) *converges to the viscosity solution of* (2.2).

Proof. Using the methods in [7, 18], this can be shown to follow from results in [9, 4]. We give a brief overview of the proof.

Monotonicity: From the positive coefficient Condition 3.1 and following the same method as in [7, 18], we can show that scheme (3.9) is monotone.

Pointwise consistency: From Appendix A, a simple Taylor series verifies consistency.

Stability: Using the same technique as in [18], we can show that scheme (3.9) is l_∞ stable by a maximum analysis.

As shown in [9], if a discretization scheme is monotone, pointwise consistent, and l_∞ stable, then that scheme converges to the viscosity solution. ∎

Remark 4.1 (rate of convergence). If $\Delta S = \max_i(S_{i+1} - S_i)$ and $\Delta\tau = C_1 h$, $\Delta S = C_2 h$, where C_1, C_2 are positive constants, then there has been considerable effort in recent years to determine rates of convergence for monotone finite difference

schemes for HJB equations. Typically, one obtains estimates of the error of the form $O(h^\rho)$, where ρ varies from 1/27 to 1/2 depending on assumptions about regularity of the solution and the PDE coefficients. See [7] for an overview of recent work along these lines. These results seem generally pessimistic when compared with numerical experiments.

It is also useful to note the following property of the matrix $[I - \Delta\tau A^n(Q, V^n)]$.

LEMMA 4.2 (M-matrix). *If the positive coefficient condition* (3.2) *is satisfied, and either Dirichlet boundary conditions are specified or no boundary condition is required, then* $[I - \Delta\tau A^n(Q, V^n)]$ *is an M-matrix for any* $Q \in \hat{Q}$.

Proof. The positive coefficient condition (3.2) implies that $\alpha_i^n, \beta_i^n, c_i^n$ in (3.4) are nonnegative $\forall Q \in \hat{Q}$. Hence $[I - \Delta\tau A^n(Q, U)]$ has positive diagonals, has nonpositive off diagonals, and is diagonally dominant. As a result, it is an M-matrix. ∎

4.1. Discretization of the control. Suppose we have a single control $q \in \hat{Q}$, where $\hat{Q} = [q_{\min}, q_{\max}]$, with $q_{\min}, q_{\max}$ bounded. It is sometimes convenient to discretize the control; i.e., we replace $\hat{Q}$ by $\hat{Y}$, where $\hat{Y} = [y_0, y_1, y_2, \ldots, y_k]$, with $y_0 = q_{\min}, y_k = q_{\max}$, and $y_{i+1} > y_i$. Let $\max_i(y_{i+1} - y_i) = C_3 h$, where C_3 is a positive constant. Then we have the following lemma.

LEMMA 4.3 (consistency of discrete control approximation). *If the HJB equation satisfies Assumption* 2.2, *then the discretized control problem with* $\max_i(y_{i+1} - y_i) = C_3 h$

$$V_\tau = \sup_{Q \in \hat{Y}} \left\{ \mathcal{L}^Q V + d(S, \tau, Q) \right\} \tag{4.1}$$

is consistent with (2.2) *as* $h \to 0$.

Proof. Let $\phi(S, \tau)$ be a smooth test function possessing bounded derivatives of appropriate order; then, in view of the fact that the coefficients of (2.2) are assumed to be Lipschitz continuous, bounded functions of Q,

$$\left| \phi_\tau - \sup_{Q \in \hat{Q}} \left\{ \mathcal{L}^Q \phi + d(S, \tau, Q) \right\} - \left(\phi_\tau - \max_{Q \in \hat{Y}} \left\{ \mathcal{L}^Q \phi + d(S, \tau, Q) \right\} \right) \right|$$

$$= O(h), \quad h \to 0. \tag{4.2}$$ ∎

PROPOSITION 4.4 (discrete control approximation: convergence to the viscosity solution). *Let* $\Delta\tau = C_1 h$, $\max_i(S_{i+1} - S_i) = C_2 h$, $\max_i(y_{i+1} - y_i) = C_3 h$, *with* C_i *being positive constants. Provided the conditions for Theorem* 4.1 *and Lemma* 4.3 *are satisfied, then the discretization* (3.9) *with* $\hat{Q}$ *replaced by the discrete control set* $\hat{Y}$ *converges to the viscosity solution of* (2.2) *as* $h \to 0$.

Proof. This follows immediately from Theorem 4.1 and Lemma 4.3. ∎

5. Solution of algebraic discrete equations. Although we have established that discretization (3.9) is consistent, l_∞ stable, and monotone, it is not obvious that this is a practical scheme, since the implicit timestepping method requires solution of highly nonlinear algebraic equations at each timestep.

5.1. Iterative method. Recall the definitions of $F^n(Q, V^n)$ and $A^n(Q, V^n)$ in (3.8). Consider the following iteration scheme:

ITERATIVE SOLUTION OF THE DISCRETE EQUATIONS

Let $(V^{n+1})^0 = V^n$

Let $\hat{V}^k = (V^{n+1})^k$

For $k = 0, 1, 2, \ldots$ until convergence

Solve

$$\left[I - \Delta\tau A^{n+1}(Q^k, \hat{V}^k)\right] \hat{V}^{k+1} = V^n + (G^{n+1} - G^n) + \Delta\tau D^{n+1}(Q^k) \tag{5.1}$$

$$Q_i^k = \arg\max_{Q \in \hat{Q}} \left\{ \left[F^{n+1}(Q, \hat{V}^k) \right]_i \right\}$$

If $(k > 0)$ and $\left(\max_i \dfrac{\left| \hat{V}_i^{k+1} - \hat{V}_i^k \right|}{\max\left(scale, \left| \hat{V}_i^{k+1} \right| \right)} < tolerance \right)$, then quit

EndFor

The term *scale* in scheme (5.1) is used to ensure that unrealistic levels of accuracy are not required when the value is very small. Typically, $scale = 1$ for financial problems priced in dollars.

Some manipulation of algorithm (5.1) results in

$$\begin{aligned} &\left[I - \Delta\tau A^{n+1}(Q^k, \hat{V}^k)\right] (\hat{V}^{k+1} - \hat{V}^k) \\ &= \Delta\tau \Big[\left(A^{n+1}(Q^k, \hat{V}^k)\hat{V}^k + D^{n+1}(Q^k) \right) \\ &\qquad - \left(A^{n+1}(Q^{k-1}, \hat{V}^{k-1})\hat{V}^k + D^{n+1}(Q^{k-1}) \right) \Big]. \end{aligned} \tag{5.2}$$

The proof of convergence of the iteration scheme (5.1) is given in [26, 18]. Since scheme (5.1) can be regarded as a variant of policy iteration for infinite horizon Markov chains, the convergence proof is similar to the proof of convergence for policy iteration [20]. For the convenience of the reader, we sketch the proof below.

In order to prove the convergence of algorithm (5.1), we first need an intermediate result.

LEMMA 5.1 (sign of right-hand side (5.2)). *If* $A^{n+1}(Q^k, \hat{V}^k)$ *is given by* (3.8), (3.9), *with the control parameter determined by*

$$Q_i^k = \arg\max_{Q \in \hat{Q}} \left\{ \left[F^{n+1}(Q, \hat{V}^k) \right]_i \right\}, \tag{5.3}$$

then every element of the right-hand side of (5.2) *is nonnegative, that is,*

$$\left[\left(A^{n+1}(Q^k, \hat{V}^k)\hat{V}^k + D^{n+1}(Q^k) \right) - \left(A^{n+1}(Q^{k-1}, \hat{V}^{k-1})\hat{V}^k + D^{n+1}(Q^{k-1}) \right) \right]_i \geq 0. \tag{5.4}$$

Proof. Recall that Q^k is selected to maximize $A^{n+1}(Q^k, \hat{V}^k)\hat{V}^k + D^{n+1}(Q^k)$ for given $\hat{V}^k$. Hence, any other choice of coefficients, for example $A^{n+1}(Q^{k-1}, \hat{V}^{k-1})\hat{V}^k + D^{n+1}(Q^{k-1})$, cannot exceed $A^{n+1}(Q^k, \hat{V}^k)\hat{V}^k + D^{n+1}(Q^k)$. ∎

It is now easy to show that iteration (5.1) always converges.

THEOREM 5.2 (convergence of iteration (5.1)). *Provided that the conditions for Lemmas* 4.2 *and* 5.1 *are satisfied, then the iteration* (5.1) *converges to the unique solution of* (3.9) *for any initial iterate* $\hat{V}^0$. *Moreover, the iterates converge monotonically.*

Proof. Given Lemmas 5.1 and 4.2, the proof of this result is similar to the proof of convergence given in [27]. We give a brief outline of the steps in this proof and refer the reader to [27] for details. A straightforward maximum analysis of scheme (5.1) can be used to bound $\|\hat{V}^k\|_\infty$ independent of iteration k. From Lemma 5.1, we have that the right-hand side of (5.2) is nonnegative. Noting that $[I - \Delta\tau A^{n+1}(Q^k, \hat{V}^k)]$ is an M-matrix (from Lemma 4.2) and hence $[I - \Delta\tau A^{n+1}(Q^k, \hat{V}^k)]^{-1} \geq 0$, it is easily seen that the iterates form a bounded nondecreasing sequence. In addition, if $\hat{V}^{k+1} = \hat{V}^k$, the residual is zero. Hence the iteration converges to a solution. It follows from the M-matrix property of $[I - \Delta\tau A^{n+1}(Q^k, \hat{V}^k)]$ that the solution is unique. ∎

Remark 5.1 (Q dependent discretizations). Note that we obtain convergence of iteration (5.1) even if the discrete equations, regarded as a function of the control Q, are discontinuous. We do, however, require that the coefficients a, b, c, d in (2.1)–(2.2) are bounded functions of the control Q, in order to ensure that A^{n+1}, D^{n+1} are bounded.

Remark 5.2 (uniqueness of solution). The above argument shows that the solution for V^{n+1} is unique. However, this does not imply that the controls Q^{n+1} are unique. As a simple counterexample, consider the case where a, b, c, d in (2.1)–(2.2) are independent of Q, in which case the solution for V is unique for any choice of $Q \in \hat{Q}$.

6. Passport options. Passport options are financial derivative contracts which allow the holder to take profit from a trading account while obligating the writer to cover losses [2, 29]. The holder is allowed to trade an underlying asset S at any time during the option lifetime, say T. Let q denote the number of shares of the underlying the holder holds at time t, $0 \leq t \leq T$. q is limited to an amount C, i.e., $|q| \leq C$. At the maturity T, the holder keeps any net gain, while any loss is covered by the writer.

In [2], this problem is solved using central weighting and Crank–Nicolson timestepping. While the results appear to converge to the correct solution, convergence to the viscosity solution cannot be guaranteed, since this is a nonmonotone method. In [26], it is shown that, for passport options, it is not possible to preselect central, forward, or backward differencing at a node, independent of the control, and guarantee a positive coefficient scheme. In other words, the scheme must depend on the control.

6.1. The pricing model for passport options. Let S be the underlying asset price which follows the stochastic process

$$dS = \mu S\, dt + \sigma S\, dZ, \tag{6.1}$$

where dZ is the increment of a Wiener process, σ is volatility, and μ is the drift rate. Let $V(S, W, t)$ denote the option value at time t with underlying price S and

wealth W. Under the process (6.1), the pricing PDE for passport options can be written as [30]

$$(6.2)\quad -V_t = -rV + (r-\gamma)SV_S + \sup_{|q|\le 1}\left[-((\gamma - r + r_c)qS - r_tW)V_W + \frac{\sigma^2S^2}{2}(V_{SS} + 2qV_{SW} + q^2V_{WW})\right],$$

where the following are defined:

W is the accumulated wealth of the underlying trading account.

r is the risk-free interest rate.

γ is the dividend rate on the underlying asset S.

r_c is a cost of carry rate.

r_t is an interest rate for the trading account.

q is the number of shares of S that an investor holds, which is also termed a trading strategy. q is limited to $|q| \le 1$ in (6.2). Different position limits can be handled by scaling [19].

We consider two types of payoff at $t = T$. The standard payoff is

$$(6.3)\qquad V(S, W, t = T) = \max(W, 0),$$

and the *asset or nothing* payoff [2] is

$$(6.4)\qquad V(S, W, t = T) = \begin{cases} S & \text{if } W \ge 0, \\ 0 & \text{otherwise.} \end{cases}$$

The above payoff can be generalized to specify a nonzero strike, but we assume the form (6.4) in this paper.

For both of these payoffs, we can reduce the problem to solving for $V(S, W, t) = Su(x, \tau)$, where $x = W/S$ and $\tau = T - t$ [2], so that (6.2) can be reduced to a one-dimensional problem for u:

$$(6.5)\quad u_\tau = -\gamma u + \sup_{|q|\le 1}\left[((r - \gamma - r_c)q - (r - \gamma - r_t)x)u_x + \frac{\sigma^2}{2}(x - q)^2u_{xx}\right],$$

where $x \in [-\infty, +\infty]$. The standard payoff becomes

$$(6.6)\qquad u(x, \tau = 0) = \max(x, 0),$$

with boundary conditions

$$(6.7)\qquad u(x \to -\infty, \tau) = 0, \quad u(x \to \infty, \tau) = x,$$

while the asset or nothing payoff is

$$(6.8)\qquad u(x, \tau = 0) = \begin{cases} 1 & \text{if } x \ge 0, \\ 0 & \text{otherwise} \end{cases}$$

with boundary conditions

$$u(x \to -\infty, \tau) = 0, \quad u(x \to \infty, \tau) = \exp(-\gamma\tau). \tag{6.9}$$

For computational purposes, we truncate the domain to $x \in [x_{\min}, x_{\max}]$ and apply the boundary conditions (6.7) and (6.9) at $x_{\min}, x_{\max}$.

6.2. Discretization. Passport option valuation is a special case of the general HJB equation (2.2) if we note that in this case

$$Q = (q), \quad \hat{Q} = [-1, +1], \quad a(x,\tau,Q) = \frac{\sigma^2}{2}(x-q)^2, \tag{6.10}$$

$$b(x,\tau,Q) = (r-\gamma-r_c)q - (r-\gamma-r_t)x, \quad c(x,\tau,Q) = \gamma, \quad d(x,\tau,Q) = 0,$$

where $Q,\ \hat{Q},\ a,\ b,\ c,\ d$ are defined in section 2. Let the discrete approximation for $u(x_i, \tau^n)$ be denoted by u_i^n with $U^n = [u_0^n, u_1^n, \ldots, u_{imax}^n]'$. Let $\hat{U}^k$ be the $k'th$ estimate for U^n; then the local objective function which must be maximized in algorithm (5.1) is

(6.11)

$$[F^{n+1}(Q, \hat{U}^k)]_i = [A^{n+1}(Q, \hat{U}^k)\hat{U}^k + D^{n+1}(Q)]_i$$

$$= ((r-\gamma-r_c)q - (r-\gamma-r_t)x_i)[(\hat{u}^k)_x]_i + \frac{\sigma^2}{2}(x_i-q)^2[(\hat{u}^k)_{xx}]_i^{n+1},$$

where $[\cdot]_i^{n+1}$ refers to the discrete form for $[\cdot]$. We need a positive coefficient scheme to solve the pricing PDE (6.5). Given node $x = x_i$, with current solution estimate $\hat{U}^k$, suppose the sets of q's, which give a positive coefficient scheme for central, forward, and backward differencing, respectively, are $P_i^{cent},\ P_i^{fwd},\ P_i^{bwd}$. Since central differencing is the most accurate, it should be used as much as possible. Consequently, given a q, if central differencing satisfies positive coefficient conditions, central differencing will be used for that q (with the constraint that F^{n+1} is $\mathcal{USC}$). In other words, the proper ranges of q for various differencings are $Range_i^{cent} = P_i^{cent}$, $Range_i^{fwd} = P_i^{fwd} - (P_i^{cent} \cap P_i^{fwd})$, and $Range_i^{bwd} = P_i^{bwd} - (P_i^{cent} \cap P_i^{bwd})$. Note that

$$Range_i^j \cap Range_i^k = \emptyset, \quad \text{where } j,k \in \{cent, bwd, fwd\} \text{ and } j \neq k,$$

$$Range_i^{bwd} \cup Range_i^{fwd} \cup Range_i^{cent} = \hat{Q}. \tag{6.12}$$

Let $\overline{Range}_i^j$ denote the closure of $Range_i^j$. For a given asset grid and the option values at the current iteration, algorithm (6.13) is used to determine the optimal control and to decide which differencing should be applied. This choice of differencing method uses central differencing as much as possible and determines $A^{n+1}(Q, \hat{U}^k)$ so that $F^{n+1}(Q, \hat{U}^k)$ is $\mathcal{USC}$.

For a given differencing method, the range of possible values of the control is divided into closed line segments where the objective function is smooth. Standard methods are then used to determine the maximum within each segment. If an analytic form for the local objective function is not available, then an alternate approach is discussed in section 7.4.

DETERMINING THE OPTIMAL CONTROL AND THE DIFFERENCING METHOD

(6.13)

Apply boundary conditions at the first node x_0 and the last node x_p
For each $x_0 < x_i < x_p$
  Compute the positive coefficient sets $Range_i^{cent}$, $Range_i^{fwd}$, $Range_i^{bwd}$
  diff = cent, $q^* = 0$, $F_{\max} = -\infty$
  For j = cent, fwd, bwd
    Compute $[A^{n+1}(q, \hat{V}^k)]_i$ with scheme $= j$
    Solve $q_j^* = \underset{q \in \overline{Range_i^j}}{\arg\max} \{[F^{n+1}(q, \hat{V}^k)]_i\}$
    If $[F^{n+1}(q_j^*, \hat{V}^k)]_i > F_{\max}$
      diff $= j$, $q^* = q_j^*$
      $F_{\max} = [F^{n+1}(q_j^*, \hat{V}^k)]_i$
    EndIf
  EndFor
EndFor

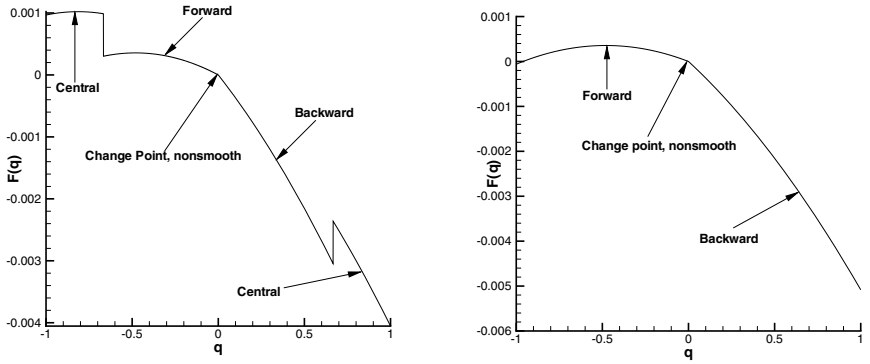
6.3. Discontinuity of the objective function. When we use central differencing as much as possible, the local objective function at each node is in general a discontinuous function of the control q. However, the proof of convergence of the iterative scheme for solution of the discretized algebraic equations (Theorem 5.2) does not require continuity of the local objective function. If forward/backward differencing are applied, the local objective function is continuous but not smooth. Figure 1 shows these features for a given grid and solution values.

Note that algorithm (6.13) requires determination of the maximum of the local objective function for each set of points where central, forward, or backward differencing is used. For example, as shown in Figure 1, central differencing can be used on two disjoint intervals of the control space. On each of the subintervals, the objective function is a smooth function of the control, and hence we can use standard methods to maximize the objective function. Determination of the range of controls where central, forward, and backward differencing give rise to a positive coefficient method is generally only possible if we have an analytic expression for the objective function. If this is not available, we can discretize the control and use a linear search to find the maximum as described in section 7.4. This is, of course, much more computationally expensive compared to analytic maximization.

6.4. Numerical results. In this section, we will examine the convergence as the grid and timesteps are refined for various differencing methods.

We use a convergence tolerance of 10^{-7} in algorithm (5.1), and we truncate the computational domain to $[x_{\min}, x_{\max}] = [-3, 4]$. Numerical experiments show that increasing the size of the computational domain does not affect solution values to six digits.

In [2, 26], input parameters in Table 1 are used for testing. An unequally spaced grid in the x direction is used, and new fine grid nodes are added between each two



(a) **(b)**

FIG. 1. *Local objective function* (6.11) *for the passport option at* $x_i = 0$. (a) *Using central differencing as much as possible;* (b) *using forward/backward differencing only. Parameters:* $r = 0.05$, $\sigma = 0.03$, *dividend* $= 0.04$, $r_c = 0.07$, $r_t = 0.03$. *Nodes:* $x_{i-1} = -0.01$, $u_{i-1} = 0.173298$; $x_i = 0.$, $u_i = 0.173888$; $x_{i+1} = 0.01$, $u_{i+1} = 0.174135$.

coarse grid nodes at each level of refinement. In Table 2, *ratio* refers to the ratio of successive changes in the solution as the grid is refined by a factor of two, and the timestep sizes are reduced by a factor of four. Since fully implicit timestepping is used, this allows us to isolate the effect of the use of central weighting as much as possible, compared to forward/backward differencing only. Local second order convergence (in terms of x node spacing) would be consistent with a ratio of four, while first order convergence would be consistent with a ratio of two.

For this particular set of parameters an analytical solution is known [2], $V(S_0, 0, \tau = T) = \$13.1381$ (to six figures). Table 2 shows that our numerical solution

TABLE 1
Parameters used in [2, 26], *passport option.*

r	0	σ	0.3
dividend	0	r_c	0
r_t	0	S_0	\$100
payoff	call	strike	\$ 0
initial wealth	0	time to expiry T	1 year

TABLE 2
Numerical example used in [2, 26], *passport option, convex payoff. Fully implicit timestepping is applied, using constant timesteps. On each refinement, new nodes are inserted between each coarse grid node, and the timestep is divided by four. Parameters are given in Table* 1. *Ratio is the ratio of successive changes in computed solution as the discretization parameters are reduced. Analytic price is* \$13.1381.

Nodes	Timesteps	Nonlinear iterations	CPU time (Sec)	Option value	Ratio
Central differencing as much as possible					
67	25	52	0.01	13.0553	
133	100	201	0.06	13.1173	
265	400	801	0.56	13.1329	3.97
529	1600	3201	4.43	13.1368	3.99
1057	6400	12801	34.84	13.1378	4.03

TABLE 3
Parameters for the convex payoff, passport option.

r	0.08	σ	0.2
dividend	0.035	r_c	0.1
r_t	0.04	S_0	\$100
payoff	call	strike	\$ 0
initial wealth	0	time to expiry T	1 year

converges to the analytical solution. The numerical solutions in [2, 26] also converge to that value. However, the discretization schemes in [2, 26] are not monotone so that convergence to the viscosity solution is not guaranteed. As shown in Theorem 4.1, our numerical method guarantees convergence to the viscosity solution.

We now compare the results from two differencing methods—using central differencing as much as possible and using forward/backward differencing only. Using the input parameters in Table 3, Table 4 presents a convergence study, which also reports the actual initial option values, i.e., $V = S_0\, u(x = W_0/S_0, \tau = T)$.

The payoff type is a call (convex payoff). As expected, Table 4 indicates that quadratic convergence is obtained by using central differencing as much as possible, and first order convergence is obtained by using forward/backward differencing.

For a convex payoff, it is always optimal to choose $q = -1$ or 1 [26]. But for a nonconvex payoff, q can be any value in $[-1, 1]$. Table 5 presents a convergence study using the parameters in Table 3, but the payoff type is an asset or nothing, nonconvex payoff (this is a digital call in terms of u; see (6.8)). For the asset or nothing payoff, if $W_0 \geq 0$, the option value will be very high (close to the initial stock value) and insensitive to the grid refinement, and so it is difficult to carry out a convergence study. In this example, option values are reported at $W_0 = -25$ (initial wealth is $-\$25$). When central differencing as much as possible is applied, the convergence rate is close to second order. First order convergence is obtained by using forward/backward differencing.

From the numerical results, we can conclude that generally we can obtain higher rates of convergence using central weighting as much as possible, compared to forward/backward differencing only. Of course, we cannot guarantee that this will always

TABLE 4
Convergence study, passport option, convex payoff. Fully implicit timestepping is applied, using constant timesteps. On each refinement, new nodes are inserted between each coarse grid node, and the timestep is divided by four. Parameters are given in Table 3. Ratio is the ratio of successive changes in computed solution as the discretization parameters are reduced.

Nodes	Timesteps	Nonlinear iterations	CPU time (Sec)	Option value	Ratio
Central differencing as much as possible					
133	100	224	0.05	10.6651	
265	400	851	0.42	10.6768	
529	1600	3277	3.17	10.6798	3.95
1057	6400	12848	25.55	10.6805	3.96
2113	25600	51223	196.83	10.6807	3.95
Forward/backward differencing only					
133	100	224	0.03	10.6945	
265	400	851	0.32	10.6916	
529	1600	3278	2.17	10.6872	0.66
1057	6400	12849	18.19	10.6842	1.49
2113	25600	51224	131.47	10.6826	1.77

TABLE 5

Convergence study, passport option, nonconvex payoff. Fully implicit timestepping is applied, using constant timesteps. On each refinement, new nodes are inserted between each coarse grid node, and the timestep is divided by four. Parameters are given in Table 3 except that $W_0 = -25$, $K = 0$, and the payoff is an asset or nothing. Ratio is the ratio of successive changes in computed solution as the discretization parameters are reduced.

Nodes	Timesteps	Nonlinear iterations	CPU time (Sec)	Option value	Ratio
Central differencing as much as possible, $W_0 = -25$					
133	100	267	0.06	25.8812	
265	400	865	0.46	26.0794	
529	1600	3300	3.32	26.1315	3.80
1057	6400	12887	25.81	26.1452	3.82
2113	25600	51274	202.68	26.1488	3.76
Forward/backward differencing only, $W_0 = -25$					
133	100	266	0.04	26.2905	
265	400	863	0.29	26.2902	
529	1600	3298	2.15	26.2384	.005
1057	6400	12890	17.40	26.1990	1.31
2113	25600	51271	132.27	26.1758	1.70

occur, but we can rarely obtain second order convergence using forward/backward differencing. In all our numerical experiments, we have never seen a case where central weighting as much as possible converges at a slower rate compared to forward/backward differencing only.

Note that in both these examples, forward/backward differencing requires only about 60% of the CPU time compared to central differencing as much as possible. This is simply because of the additional tests required to determine the ranges of possible central weighting in algorithm (6.13).

Both Tables 4 and 5 show that the number of nonlinear iterations per timestep is about two, indicating that algorithm (5.1) converges rapidly, in spite of the discontinuous objective function that is maximized at each node.

7. Defined contribution pension plan. The second example in this paper concerns an optimal dynamic asset allocation strategy for a defined contribution pension plan. A traditional asset allocation strategy for a defined contribution pension plan is *deterministic lifestyling*. Initially, the contributions of the plan are invested entirely in equities. Beginning on a predetermined date, say N years prior to retirement, the contributions are switched into bonds at a rate of $1/N$ per year. Then all assets are invested in bonds by the date of retirement. Deterministic lifestyling can reduce the losses of the plan in case of a sudden fall in the stock market just before the date of retirement. This strategy is simple and widely used. However, obviously it is not the optimal strategy.

We will follow the approach given in [14] to find the optimal dynamic asset allocation strategy for a defined contribution pension plan. The objective of the strategy is to maximize the plan member's utility at retirement. It is assumed that the utility is a function of the plan member's wealth to yearly income ratio [14].

We give a brief derivation of the model equations; for details we refer the reader to [14].

7.1. Stochastic model. Suppose there are two underlying assets in the pension plan: one is risk-free (e.g., a government bond) and the other is risky (e.g., a stock

market fund). The risky asset S follows the stochastic process

$$dS = (r + \xi_1\sigma_1)S\ dt + \sigma_1 S\ dZ_1, \tag{7.1}$$

where dZ_1 is the increment of a Wiener process, σ_1 is volatility, r is the interest rate, and ξ_1 is the market price of risk. Suppose that the plan member continuously pays into the pension plan at a fraction π of his/her yearly salary Y, which follows the process

$$dY = (r + \mu_Y)Y\ dt + \sigma_{Y_0} Y\ dZ_0 + \sigma_{Y_1} Y\ dZ_1, \tag{7.2}$$

where μ_Y, σ_{Y_0}, and σ_{Y_1} are constants, and dZ_0 is another increment of a Wiener process, which is independent of dZ_1. Let $W(t)$ denote the wealth accumulated in the pension plan, let p denote the proportion of this wealth invested in the risky asset S, and let $1 - p$ denote the fraction of wealth invested in the risk-free asset. Then

$$dW = (r + p\xi_1\sigma_1)W\ dt + p\sigma_1 W dZ_1 + \pi Y dt. \tag{7.3}$$

Assume the plan member has a power utility function $u(W(T), Y(T))$ at retirement time T, which is defined as a function of the wealth-income ratio,

$$u(W(T), Y(T)) = \begin{cases} \frac{1}{\gamma}\left(\frac{W(T)}{Y(T)}\right)^{\gamma}, & \text{where } \gamma < 1 \text{ and } \gamma \neq 0, \\ \log\left(\frac{W(T)}{Y(T)}\right) & \text{when } \gamma = 0. \end{cases} \tag{7.4}$$

Our goal is to find the optimal asset allocation strategy to maximize the expected terminal utility. Define a new state variable $X(t) = W(t)/Y(t)$; then by Itô's lemma, we obtain

$$\begin{aligned} dX = [\pi + X(-\mu_Y + p\sigma_1(\xi_1 - \sigma_{Y_1}) + \sigma_{Y_0}^2 + \sigma_{Y_1}^2)]dt \\ - \sigma_{Y_0} X dZ_0 + X(p\sigma_1 - \sigma_{Y_1})dZ_1. \end{aligned} \tag{7.5}$$

Let $J(t, x, p) = E[u(X_p(T))|X(t) = x]$, where $X(t)$ is the path of X given the asset allocation strategy $p = p(t, x)$, and $E[\cdot]$ is the expectation operator. We define

$$V(x, \tau) = \sup_{p \in \hat{P}} E[u(X_p(T))|X(T - \tau) = x] = \sup_{p \in \hat{P}} J(T - \tau, x, p), \tag{7.6}$$

where $\hat{P}$ is the set of all admissible asset allocation strategies, and $\tau = T - t$. Then $V(x, \tau)$ satisfies the HJB equation

$$V_\tau = \sup_{p \in \hat{P}} \left\{ \mu_X^p V_x + \frac{1}{2}(\sigma_X^p)^2 V_{xx} \right\}, \quad x \in [0, \infty], \tag{7.7}$$

with terminal condition

$$V(x, \tau = 0) = \begin{cases} \gamma^{-1} x^{\gamma}, & \text{where } \gamma < 1 \text{ and } \gamma \neq 0, \\ \log(x) & \text{when } \gamma = 0, \end{cases} \tag{7.8}$$

and where

$$\begin{aligned} \mu_X^p &= \pi + x(-\mu_Y + p\sigma_1(\xi_1 - \sigma_{Y_1}) + \sigma_{Y_0}^2 + \sigma_{Y_1}^2), \\ (\sigma_X^p)^2 &= x^2(\sigma_{Y_0}^2 + (p\sigma_1 - \sigma_{Y_1})^2), \end{aligned} \tag{7.9}$$

TABLE 6

Computational parameters, pension plan. Convergence tolerance used in algorithm (5.1). ϵ *is used to adjust the terminal condition* (7.12). $p_{\max}$ *is the maximum value of the equity proportion* (7.11). $x_{\max}$ *is the maximum* x *value in the finite computational domain.*

convergence tolerance	10^{-7}
ϵ	10^{-3}
$p_{\max}$	200
$x_{\max}$	80

with boundary conditions

$$V_\tau(x=0,\tau) = \pi V_x, \quad V(x\to\infty,\tau) = 0. \tag{7.10}$$

For computational purposes, we truncate the domain to $[0, x_{\max}]$ and impose the boundary conditions (7.10) on this finite domain. In order to ensure that Assumption 2.2 holds, we define the range of controls to be

$$\hat{P} = [0, p_{\max}]. \tag{7.11}$$

Note that in the original problem in [14], $\hat{P} = [0,\infty]$. A value of $p > 1$ indicates that the holder borrows to invest in risky assets. As a practical matter, it is unlikely that anyone could borrow an unlimited amount relative to his/her wealth to invest in risky assets. We will choose $p_{\max}$ sufficiently large so that the computed solution is insensitive to $p_{\max}$.

The terminal condition (7.4) is undefined for $x = 0$ if, for example, $\gamma < 0$. We adopt the simple expedient of replacing condition (7.4) by

$$V(x,\tau=0) = \begin{cases} \frac{1}{\gamma}\max(x,\epsilon)^\gamma & \text{if } \gamma < 0, \\ \log(\max(x,\epsilon)) & \text{if } \gamma = 0, \end{cases} \tag{7.12}$$

where $\epsilon > 0, \epsilon \ll 1$. We choose ϵ sufficiently small so that the computed results are insensitive to this value. Table 6 shows the computational parameters used in our numerical tests.

Note that the HJB equation (7.7) becomes independent of p at $x = 0$. This nonuniqueness of p does not affect the solution value u. It will be understood in the following that if we refer to a value of, say, $(p\ x)$ at $x = 0$, then we are really referring to

$$\lim_{x\to 0}(p\ x). \tag{7.13}$$

7.2. Discretization. The pension plan asset allocation model is a special case of the general HJB equation (2.2) if we make the identification

$$Q = (p)\ , \quad \hat{Q} = [0, p_{\max}],$$

$$a(x,\tau,Q) = \frac{1}{2}(\sigma_X^p)^2, \quad b(x,\tau,Q) = \mu_X^p, \quad c(x,\tau,Q) = 0, \quad d(x,\tau,Q) = 0, \tag{7.14}$$

where Q, a, b, c, d are defined in (2.2). Given node $x = x_i$, with specified solution estimate $\hat{V}^k = [\hat{v}_0^k, \ldots, \hat{v}_{imax}^k]'$, the objective function which is maximized at each node in algorithm (5.1) is

$$\begin{aligned}[F^{n+1}(Q,\hat{V}^k)]_i &= [A^{n+1}(Q,\hat{V}^k)\hat{V}^k + D^{n+1}(Q)]_i \\ &= [\mu_X^p]_i[(\hat{v}^k)_x]_i + \frac{1}{2}([\sigma_X^p]_i)^2[(\hat{v}^k)_{xx}]_i^{n+1}, \end{aligned} \tag{7.15}$$

TABLE 7

Parameters used in the pension plan examples. The time units in this problem are years, so that the ratio of wealth to salary x has the units of years.

μ_y	0.	ξ_1	0.2
σ_1	0.2	σ_{Y1}	0.05
σ_{Y0}	0.05	π	0.1
T	20 years	γ	-5

where $\hat{V}^k$ is the vector containing the current estimate of the discrete solution values. Similar to the passport option case, if we want to apply central differencing as much as possible, algorithm (6.13) is used to decide which differencing scheme is used (which depends on Q and $\hat{V}^k$).

7.3. Numerical results. Given the parameters in Table 7, Table 8 shows the numerical results. Recall that as we refine the grid, by inserting a fine grid node between two coarse grid nodes, we reduce the timestep size by four. Since fully implicit timestepping is used (which guarantees a monotone scheme), then the ratio of successive changes in the solution, as the grid is refined, should be four for quadratic convergence and two for linear convergence. As expected, Table 8 shows that quadratic convergence is obtained by using central differencing as much as possible, and first order convergence is obtained by using forward/backward differencing. As for the passport option case, convergence of algorithm (5.1) is rapid.

TABLE 8

Convergence study, pension plan example. Fully implicit timestepping is applied, using constant timesteps. On each refinement, new fine grid nodes are inserted between each coarse grid node, and the timestep size is reduced by four. Parameters are given in Table 7. The utility values are given at $x = 1$ and $x = 0$. Ratio is the ratio of successive changes in computed solution as the discretization parameters are reduced.

Nodes	Timesteps	Nonlinear iterations	CPU time (sec)	Utility	Ratio
Central differencing as much as possible, $x = 0$					
87	160	331	0.04	-4.06482×10^{-3}	
173	640	1280	0.36	-3.65131×10^{-3}	
345	2560	5120	2.75	-3.58063×10^{-3}	5.85
689	10240	20480	21.31	-3.56354×10^{-3}	4.13
1377	40960	81920	168.07	-3.55922×10^{-3}	3.96
Forward/backward differencing only, $x = 0$					
87	160	399	0.03	-6.73472×10^{-3}	
173	640	1296	0.22	-4.68055×10^{-3}	
345	2560	5135	1.68	-4.04828×10^{-3}	3.25
689	10240	20480	13.06	-3.79150×10^{-3}	2.46
1377	40960	81920	103.09	-3.67543×10^{-3}	2.21
Central differencing as much as possible, $x = 1$					
87	160	331	0.04	-4.68742×10^{-4}	
173	640	1280	0.36	-4.31528×10^{-4}	
345	2560	5120	2.75	-4.26814×10^{-4}	7.89
689	10240	20480	21.31	-4.25611×10^{-4}	3.92
1377	40960	81920	168.07	-4.25305×10^{-4}	3.92
Forward/backward differencing only, $x = 1$					
87	160	399	0.03	-7.14415×10^{-4}	
173	640	1296	0.22	-5.55931×10^{-4}	
345	2560	5135	1.68	-4.87660×10^{-4}	2.32
689	10240	20480	13.06	-4.55786×10^{-4}	2.14
1377	40960	81920	103.09	-4.40348×10^{-4}	2.06

Numerical tests with the parameters in Table 6 indicated that increasing the truncated domain size $x_{\max}$, increasing the maximum value of the control $p_{\max}$, and decreasing the convergence tolerance and ϵ resulted in no change to the results in Table 8 to six figures.

In the passport option case, particularly with a convex payoff, numerical experiments indicate that using central differencing only does converge to the viscosity solution [2, 26]. However, this cannot be guaranteed. In contrast, in the pension plan case, the numerical scheme does not appear to converge at all using central differencing only. In this respect, the pension plan problem appears to be more challenging than passport option valuation.

Given the parameters in Table 7, Figure 2 shows the expected terminal utility at $t = 0$ and the corresponding optimal asset allocation strategy p as a function of the wealth to salary ratio x. Note that as $x \to 0$, the proportion invested in the risky asset becomes very large. However, as noted in [14], the amount actually invested in the risky asset $(p\ x)$, tends to zero as $x \to 0$. This is clearly illustrated Figure 2(d). Moreover, [14] points out that $(p\ x)$ converges to 0 at the same rate as $\sqrt{x}$. Using

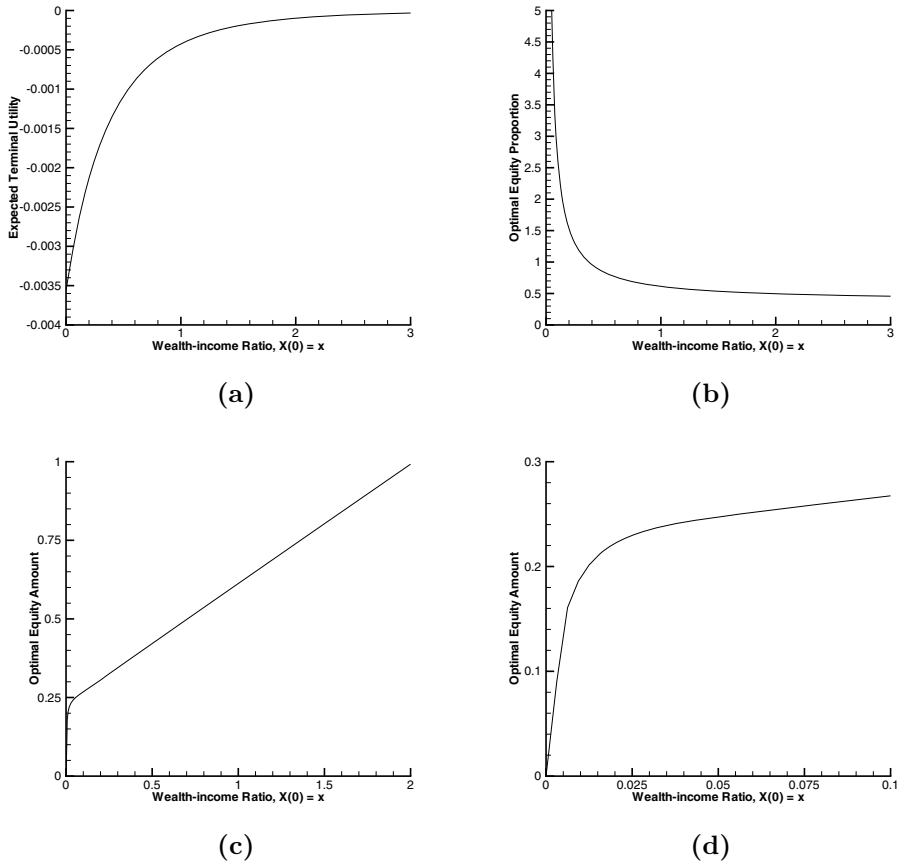


FIG. 2. *Utility and optimal asset allocation strategy at* $t = 0$*, pension plan example. Parameters are given in Table* 7. (a) *Expected terminal utility;* (b) *optimal asset allocation strategy;* (c) *optimal equity amount* $(p\ x)$*;* (d) *magnified graph of* (c).

the parameters in Table 7, we find that $\lim_{x\to 0} \frac{px}{\sqrt{x}} \simeq 1.59$. In [14], the authors do not give details about their numerical method and present only graphical results. We can comment only that the results in Figure 2 are qualitatively similar to the results in [14].

7.4. Discretization of the control. In some cases, if the form of the HJB equation is complex, then it may be difficult to implement algorithm (6.13). In this case, a simpler approach is desirable. Suppose there is one control q at each node, and we discretize the possible control values as described in section 4.1. From Lemma 4.3, we have that a scheme using discrete controls will converge to the viscosity solution of the original HJB equation. To determine the optimal control at each node, as required in algorithm (5.1), we simply perform a linear search of the discrete control values. For a given q, we use central weighting if this results in a positive coefficient method; otherwise, forward/backward differencing is used.

Note that since we cannot assume that the objective function is a continuous function of the control, linear search is the only way to find the optimal value of q. This method has the obvious advantage that it is very easy to implement, especially in the case where central differencing is used as much as possible.

The numerical results obtained using this method for the pension plan problem are given in Table 9. The results are very close to the results reported in Table 8. Of course, this method requires much more CPU time compared to algorithm (6.13). This is simply due to the comparatively crude method used to find the optimal control at each grid node.

In an effort to do better than linear search, we experimented with various approximate methods for finding the optimal control (assuming a discrete set of controls). Seemingly reasonable methods based on smooth approximations to the objective function were very unreliable, and algorithm (5.1) typically failed to converge. This is simply because the smooth approximation may not find the global maximum of the local objective function, and hence the argument used to prove the convergence of the iteration (Theorem 5.2) breaks down.

TABLE 9

Convergence study, pension plan example, discretized control. x-nodes refers to the number of nodes in the x grid. p-nodes refers to the number of nodes in the discretization of the range of control values. Fully implicit timestepping is used with constant timesteps. On each refinement, new fine grid nodes are inserted between each coarse grid node, and the timestep is reduced by four. Central differencing is used as much as possible. Ratio is the ratio of successive changes in computed solution as the discretization parameters are reduced. Problem data is given in Table 7.

x-nodes	p-nodes	Timesteps	Nonlinear iterations	CPU (sec)	Utility	Ratio
			$x = 0$			
173	113	640	1317	1.9	-3.65307×10^{-3}	
345	225	2560	5146	29.4	-3.58083×10^{-3}	
689	449	10240	20511	457	-3.56358×10^{-3}	4.19
1377	897	40961	82016	7240	-3.55923×10^{-3}	3.97
			$x = 1.0$			
173	113	640	1317	1.9	-4.31662×10^{-4}	
345	225	2560	5146	29.4	-4.26845×10^{-4}	
689	449	10240	20511	457	-4.25619×10^{-4}	3.93
1377	897	40961	82016	7240	-4.25306×10^{-4}	3.92

8. Extension to multidimensional models. We have fully discussed the solution of one-factor cases. Many financial problems are modelled as multifactor HJB equations. Our scheme can, in principal, be extended to multifactor problems. For example, for two-factor models, the operator $\mathcal{L}$ (defined by (2.1)) changes to

$$\begin{aligned}\mathcal{L}^Q V \equiv\ & a_1(X,Y,\tau,Q)V_{XX} + a_2(X,Y,\tau,Q)V_{YY} + a_3(X,Y,\tau,Q)V_{XY} \\ & + a_4(X,Y,\tau,Q)V_X + a_5(X,Y,\tau,Q)V_Y - a_6(X,Y,\tau,Q)V, \end{aligned} \tag{8.1}$$

where

$$\begin{bmatrix} a_1 & a_3/2 \\ a_3/2 & a_2 \end{bmatrix} \tag{8.2}$$

is positive semidefinite. Consequently, the general form of our pricing PDE changes to

$$\begin{aligned} V_\tau = \sup_{Q\in\hat{Q}} \left\{ \mathcal{L}^Q V + d(X,Y,\tau,Q) \right\}, \\ X \in [X_{\min}, X_{\max}], \quad Y \in [Y_{\min}, Y_{\max}], \quad 0 \le \tau \le T. \end{aligned} \tag{8.3}$$

If we can find a positive coefficient discretization scheme for (8.3), we can solve this problem using the same methods as described for the single factor case. All the proofs follow without change.

However, the difficulty results from the fact that it is not trivial to find a positive coefficient discretization for the term $a_3 V_{XY}$. One possible approach is discussed in [7], where if certain grid spacing conditions are satisfied, then a positive coefficient discretization can be obtained if a seven point finite difference method is used to approximate the V_{XY} term (the orientation of the seven point operator depends on the sign of a_3). Note that forward/backward differencing for the first order terms is suggested in [7].

9. Conclusions. Many financial models result in nonlinear HJB PDEs. Classical solutions to these PDEs do not usually exist. In order to ensure convergence to the viscosity solution, monotone difference methods must be used. The standard approach simply uses forward/backward differencing to ensure monotonicity. Clearly, this method can be at most first order accurate.

However, in many financial applications, it is often the case that central differencing can be used at many nodes. This possibility seems to have been ignored in previous work. In this paper, we use central differencing as much as possible. When we use central differencing as much as possible, the local objective function at each node is a discontinuous function of the control. However, the proof of convergence of the iterative scheme for solution of the fully implicit discretized algebraic equations does not require continuity of the local objective function. Hence convergence of the iterative algorithm for solving the nonlinear discretized equations is guaranteed.

We have reported numerical experiments for pricing passport options and optimal asset allocation for defined contribution pension plans. In all cases, use of central differencing as much as possible converges at a higher rate than use of forward/backward differencing only. We have seen this same effect in many numerical experiments. Use of central differencing as much as possible is never worse (in terms of convergence rate) and almost always superior to forward/backward differencing only. Note that

these higher rates cannot be guaranteed, but convergence to the viscosity solution is guaranteed.

However, use of central differencing as much as possible is more costly than simply using forward/backward differencing. This is due to the fact that the possible discontinuity points of the objective function must be identified, so that standard methods can be used to maximize the objective function on intervals where the objective function is smooth. This also adds somewhat to implementation complexity.

In cases where it is not possible to analytically maximize the objective function, a very straightforward approach is to discretize the control and to maximize the objective function using a linear search. This method is very easy to implement and is also guaranteed to converge to the viscosity solution, but this technique is much more computationally expensive. In these cases, it is clearly advantageous to use central differencing as much as possible, so that higher rates of convergence may compensate for the increased computational cost.

The methods developed in this article can be extended to multifactor problems if a monotone scheme is used to discretize the differential operator. However, in this case it may be nontrivial to satisfy the monotonicity requirement, due to the presence of a cross-derivative term.

Appendix A. Discrete equation coefficients. Let Q_i^n denote the vector of optimal controls at node i and time level n, and set

$$a_i^n = a(S_i, \tau^n, Q_i^n), \quad b_i^n = b(S_i, \tau^n, Q_i^n), \quad c_i^n = c(S_i, \tau^n, Q_i^n). \tag{A.1}$$

Then we can use central, forward, or backward differencing at any node.

Central differencing:

$$\alpha^n_{i,central} = \left[\frac{2a_i^n}{(S_i - S_{i-1})(S_{i+1} - S_{i-1})} - \frac{b_i^n}{S_{i+1} - S_{i-1}} \right],$$

$$\beta^n_{i,central} = \left[\frac{2a_i^n}{(S_{i+1} - S_i)(S_{i+1} - S_{i-1})} + \frac{b_i^n}{S_{i+1} - S_{i-1}} \right]. \tag{A.2}$$

Forward/backward differencing ($b_i^n > 0/\ b_i^n < 0$):

$$\alpha^n_{i,forward/backward} = \left[\frac{2a_i^n}{(S_i - S_{i-1})(S_{i+1} - S_{i-1})} + \max\left(0, \frac{-b_i^n}{S_i - S_{i-1}}\right) \right],$$

$$\beta^n_{i,forward/backward} = \left[\frac{2a_i^n}{(S_{i+1} - S_i)(S_{i+1} - S_{i-1})} + \max\left(0, \frac{b_i^n}{S_{i+1} - S_i}\right) \right]. \tag{A.3}$$

REFERENCES

[1] A. L. Amadori, *Nonlinear integro-differential evolution problems arising in option pricing: A viscosity solution approach*, Differential Integral Equations, 16 (2003), pp. 787–811.

[2] L. Andersen, J. Andreasen, and R. Brotherton-Ratcliffe, *The passport option*, J. Computational Finance, 1 (1998), pp. 15–36.

[3] M. Avellaneda, A. Levy, and A. Parás, *Pricing and hedging derivative securities in markets with uncertain volatilities*, Appl. Math. Finance, 2 (1995), pp. 73–88.

[4] G. Barles, *Convergence of numerical schemes for degenerate parabolic equations arising in finance*, in Numerical Methods in Finance, L. C. G. Rogers and D. Talay, eds., Cambridge University Press, Cambridge, UK, 1997, pp. 1–21.

[5] G. Barles and J. Burdeau, *The Dirichlet problem for semilinear second-order degenerate elliptic equations and applications to stochastic exit time control problems*, Comm. Partial Differential Equations, 20 (1995), pp. 129–178.

[6] G. Barles, C. Daher, and M. Romano, *Convergence of numerical shemes for parabolic equations arising in finance theory*, Math. Models Methods Appl. Sci., 5 (1995), pp. 125–143.

[7] G. Barles and E. Jakobsen, *Error bounds for monotone approximation schemes for parabolic Hamilton-Jacobi-Bellman equations*, Math. Comp., 76 (2007), pp. 1861–1893.

[8] G. Barles and E. Rouy, *A strong comparison result for the Bellman equation arising in stochastic exit time control problems and its applications*, Comm. Partial Differential Equations, 23 (1998), pp. 1945–2033.

[9] G. Barles and P. Souganidis, *Convergence of approximation schemes for fully nonlinear equations*, Asymptotic Anal., 4 (1991), pp. 271–283.

[10] E. Bayraktar and V. Young, *Pricing Options in Incomplete Markets via the Instantaneous Sharpe Ratio*. Working paper, University of Michigan, Ann Arbor, MI, 2005.

[11] F. E. Benth, K. Karlsen, and K. Reikvam, *Optimal portfolio management management rules in a non-Gaussian market with durability and intertemporal substitution*, Finance Stoch., 5 (2001), pp. 447–467.

[12] Y. Bergman, *Option pricing with differential interest rates*, Review Financial Studies, 8 (1995), pp. 475–500.

[13] J. Cai and H. Yang, *Ruin in the perturbed compound Poisson risk process under interest force*, Adv. in Appl. Probab., 37 (2005), pp. 819–835.

[14] A. Cairns, D. Blake, and K. Dowd, *Stochastic lifestyling: Optimal dynamic asset allocation for defined contribution pension plans*, J. Econom. Dynam. Control, 30 (2006), pp. 843–877.

[15] J. P. Chancellor, B. Oksendal, and A. Sulem, *Combined stochastic control and optimal stopping and application to numerical approximation of combined stochastic and impulse control*, Stochastic Financial Math., 237 (2002), pp. 149–173.

[16] S. Chaumont, *A Strong Comparison Result for Viscosity Solutions to Hamilton-Jacobi-Bellman Equations with Dirichlet Conditions on a Non-Smooth Boundary*, Working paper, Institute Elie Cartan, Université Nancy I, Nancy Cedex, France, 2004.

[17] M. G. Crandall, H. Ishii, and P. L. Lions, *User's guide to viscosity solutions of second order partial differential equations*, Bull. Amer. Math. Soc. (N.S.), 27 (1992), pp. 1–67.

[18] P. Forsyth and G. Labahn, *Numerical methods for controlled Hamilton-Jacobi-Bellman PDEs in finance*, J. Computational Finance, 11 (2007/2008, Winter), pp. 1–43.

[19] T. Hyer, A. Lipton-Lifschitz, and D. Pugachevsky, *Passport to success*, Risk Magazine, 10 (1997), pp. 127–131.

[20] H. Kushner and P. Dupuis, *Numerical Methods for Stochastic Control Problems in Continuous Time*, Springer-Verlag, New York, 1991.

[21] H. E. Leland, *Option pricing and replication with transaction costs*, J. Finance, 40 (1985), pp. 1283–1301.

[22] M. Milevsky, S. Promislow, and V. Young, *Financial Valuation of Mortality Risk via the Instantaneous Sharpe Ratio*, Working paper, York University, Toronto, ON, Canada, 2005.

[23] M. Mnif and A. Sulem, *Optimal risk control and dividend policies under excess of loss reinsurance*, Stoch. Stoch. Rep., 77 (2005), pp. 455–476.

[24] B. Oksendal and A. Sulem, *Applied Stochastic Control of Jump Diffusions*, Springer-Verlag, Berlin, 2005.

[25] H. Pham, *On some recent aspects of stochastic control and their applications*, Probab. Surv., 2 (2005), pp. 506–549.

[26] D. M. Pooley, *Numerical Methods for Nonlinear Equations in Option Pricing*, Ph.D. thesis, School of Computer Science, University of Waterloo, Waterloo, ON, Canada, 2003.

[27] D. Pooley, P. Forsyth, and K. Vetzal, *Numerical convergence properties of option pricing PDEs with uncertain volatility*, IMA J. Numer. Anal., 23 (2003), pp. 241–267.

[28] H. Schmidli, *On minimizing the ruin probability by investment and reinsurance*, Ann. Appl. Probab., 12 (2002), pp. 890–907.

[29] S. Shreve and J. Vecer, *Options on a traded account: Vacation calls, vacation puts, and passport options*, Finance Stoch., 4 (2000), pp. 255–274.

[30] J. Topper, *A Finite Element Implementation of Passport Options*, M.Sc. thesis, University of Oxford, Oxford, UK, 2003.

[31] H. Windcliff, J. Wang, P. Forsyth, and K. Vetzal, *Hedging with a correlated asset: Solution of a nonlinear pricing PDE*, J. Comput. Appl. Math., 200 (2007), pp. 86–115.

[32] Y.-L. Zhu and J. Li, *Multi-factor financial derivatives on finite domains*, Commun. Math. Sci., 1 (2003), pp. 343–359.

SIAM J. NUMER. ANAL.
Vol. 46, No. 3, pp. 1602–1618
© 2008 Society for Industrial and Applied Mathematics

MAXIMUM NORM A POSTERIORI ERROR ESTIMATE FOR A 2D SINGULARLY PERTURBED SEMILINEAR REACTION-DIFFUSION PROBLEM*

NATALIA KOPTEVA†

Abstract. A singularly perturbed semilinear reaction-diffusion equation, posed in the unit square, is discretized on arbitrary nonuniform tensor-product meshes. We establish a second-order maximum norm a posteriori error estimate that holds true uniformly in the small diffusion parameter. No mesh aspect ratio assumption is made. Numerical results are presented that support our theoretical estimate.

Key words. reaction-diffusion, singular perturbation, a posteriori error estimate, maximum norm, no mesh aspect ratio condition, finite differences, layer-adapted mesh

AMS subject classifications. 65N06, 65N15, 65N50

DOI. 10.1137/060677616

1. Introduction.

Solutions of singularly perturbed differential equations typically exhibit sharp boundary and interior layers, which are narrow regions where solutions change rapidly. To obtain reliable numerical approximations of layer solutions in an efficient way, one has to use locally refined meshes that are fine in layer regions and standard outside. Furthermore, as is shown in [20, 19, 7, 13] by the numerical analysis of model problems, for which the location and width of the layers are known a priori, optimal layer-adapted meshes have extremely high maximum aspect ratios (typically $O(\varepsilon^{-1})$, where ε is the layer width).

In contrast, a posteriori error estimates, which underlie any reliable a posteriori mesh construction, are typically obtained under the minimum angle condition, which is equivalent to the bounded-mesh-aspect-ratio condition; see, e.g., [2, 21]. But the minimum angle condition seems rather restrictive and makes a posteriori error estimates less practical for layer solutions, for which a posteriori mesh generation is most needed.

The aim of the present paper is to establish an a posteriori error estimate for one singularly perturbed problem under *no mesh aspect ratio condition*. Note that our error estimate is in the maximum norm, which is sufficiently strong to capture layers and hence seems most appropriate for singularly perturbed problems.

Consider the singularly perturbed semilinear reaction-diffusion problem posed in the unit square:

$$\begin{aligned} Tu := -\varepsilon^2 \triangle u + b(x,y,u) = 0, \qquad & (x,y) \in \Omega = (0,1)\times(0,1), \\ u(x,y) = 0, \qquad & (x,y) \in \partial\Omega. \end{aligned} \tag{1.1}$$

*Received by the editors December 14, 2006; accepted for publication (in revised form) August 21, 2007; published electronically April 11, 2008. This publication emanated from research conducted with the financial support of Science Foundation Ireland under the Basic Research Grant Programme 2004, grant 04/BR/M0055.

http://www.siam.org/journals/sinum/46-3/67761.html

†Department of Mathematics and Statistics, University of Limerick, Limerick, Ireland (natalia.kopteva@ul.ie).

Here ε is a small positive parameter, $\triangle = \partial^2/\partial x^2 + \partial^2/\partial y^2$ is the Laplace operator, the function b is sufficiently smooth, and

$$0 < \beta < b_u(x, y, u) \le \bar{\beta} \qquad \text{for all } (x, y, u) \in [0, 1]^2 \times \mathbb{R}. \tag{1.2}$$

Under condition (1.2), problem (1.1) has a unique solution, which exhibits sharp boundary layers of width $O(\varepsilon|\ln\varepsilon|)$ along the boundary $\partial\Omega$.

We discretize (1.1) using the standard second-order five-point difference scheme—see (2.2) for details—on an arbitrary tensor-product mesh $\{(x_i, y_j)\}$, where $0 = x_0 < x_1 < \cdots < x_N = 1$ and $0 = y_0 < y_1 < \cdots < y_M = 1$, while $h_i = x_i - x_{i-1}$ and $k_j = y_j - y_{j-1}$ are the local mesh sizes.

This is an idealized situation in the a posteriori mesh construction context since an irregular mesh, rather than a tensor-product mesh, seems more suitable for a practical a posteriori mesh construction algorithm. Therefore the error estimate, which we present, is more interesting from a theoretical point of view. In particular, it shows that the bounded-mesh-aspect-ratio condition/minimal-angle condition is not essential in the a posteriori error estimation. Furthermore, if tensor-product meshes are used at least in crucial layer regions, where the mesh adaptation is most needed, one might conjecture that in such regions, local analogues of our a posteriori error estimate would apply.

Our main result is the following maximum norm *a posteriori error estimate*, in which the error is understood as the difference between the exact solution and the bilinear interpolant of the computed solution:

$$\|U^{\mathrm{B}} - u\|_\infty \le C_0 \left[\max_{\substack{i=1,\dots,N \\ j=0,\dots,M}} \left\{ h_i^2 M_{1,ij} \right\} + \max_{\substack{i=0,\dots,N \\ j=1,\dots,M}} \left\{ k_j^2 M_{2,ij} \right\} \right] \tag{1.3}$$

—see Theorem 2.1—where, roughly speaking,

$$M_{1,ij} \approx |D_x^2 U_{ij}| \ln\bigl(2 + \varepsilon/\kappa\bigr) + 1, \qquad M_{2,ij} \approx |D_y^2 U_{ij}| \ln\bigl(2 + \varepsilon/\kappa\bigr) + 1, \tag{1.4}$$

with $\kappa := \min\{\min_i\{h_i\}, \min_j\{k_j\}\}$. Here U^{B} is the bilinear interpolant of the computed solution U (the finite difference computed solution is originally defined at the mesh nodes only; hence to measure the error in the entire domain, one first has to interpolate the computed solution there). The quantities $D_x^2 U_{ij}$ and $D_y^2 U_{ij}$ are the standard discrete approximations of $\partial^2 u/\partial x^2$ and $\partial^2 u/\partial y^2$ defined in (2.3). In (1.4), a few terms are skipped, for which the one-dimensional analysis [12] and the numerical results of section 6 show that they are less important; see Theorem 2.1 for the precise definitions of $M_{1,ij}$ and $M_{2,ij}$.

The error constant C_0 in (1.3) is independent of ε, the mesh, and aspect ratios of its elements, although this constant is not specified. In a posteriori error estimation, much attention focuses on specifying the error constants. Note that for singularly perturbed problems, the error constant might blow up as ε becomes small, and hence the existence of an ε-uniform error constant is more significant than its precise value.

Note that, roughly speaking, the a posteriori error estimate (1.3), (1.4) might be viewed as a discrete analogue of the linear interpolation error estimates [8, 6], which imply that a suitable anisotropic mesh should be quasi-uniform under the metric induced by the Hessian matrix. An example of this idea being exploited for mesh generation is given, e.g., in [10], where anisotropic meshes, defined as mappings of regular uniform grids, are obtained via functional minimization.

The present paper follows [11] and in particular [12], where certain maximum norm a posteriori error estimates were derived for one-dimensional singularly perturbed convection-diffusion and reaction-diffusion problems; see also a more recent paper [18] for a similar one-dimensional a posteriori error estimate. Note that the papers [11, 12, 18] report one-dimensional error estimates, while now we extend the one-dimensional analysis [12] to a two-dimensional case.

Problem (1.1) has often been addressed in the numerical analysis literature. In particular, we refer the reader to [5], where iterative techniques for the standard finite difference discretization of (1.1) are developed, and [15, 16], where certain a posteriori error estimates for a linear version of (1.1) are obtained on anisotropic meshes in the energy norm.

The paper is organized as follows. In section 2, we describe the numerical method, present our a posteriori error estimate in Theorem 2.1, and outline its proof. Next, in section 3, we establish some estimates for the Green's function of a linearized version of (1.1). They imply certain stability properties of the differential operator T from (1.1), which are presented in section 4. Then in section 5, we complete the proof of Theorem 2.1. Finally, in section 6, numerical results are given that support our theoretical estimate.

Notation. Let $\|\cdot\|_{p;\tilde\Omega}$, where $1 \le p \le \infty$, denote the standard $L_p(\tilde\Omega)$ norm for any domain $\tilde\Omega$. Furthermore, the standard notation $W^{k,p}(\tilde\Omega)$ is used for the Sobolev spaces with the norm $\|\cdot\|_{k,p;\tilde\Omega}$ defined, for a function $v(x,y)$ in a domain $\tilde\Omega$, by

$$\|v\|_{k,p;\tilde\Omega} = \|v\|_{p;\tilde\Omega} + \sum_{l=1}^{k} |v|_{l,p;\tilde\Omega}, \qquad k = 1, 2,$$

$$|v|_{1,p;\tilde\Omega} = \|v_x\|_{p;\tilde\Omega} + \|v_y\|_{p;\tilde\Omega}\,, \qquad |v|_{2,p;\tilde\Omega} = \|v_{xx}\|_{p;\tilde\Omega} + \|v_{xy}\|_{p;\tilde\Omega} + \|v_{yy}\|_{p;\tilde\Omega}\,;$$

see, e.g., [9]. We shall use the notation $\|\cdot\|_p$ and $\|\cdot\|_{k,p}$ for $\|\cdot\|_{p;\Omega}$ and $\|\cdot\|_{k,p;\Omega}$ when there is no ambiguity. Sometimes the domain of interest will be an open ball $B(a,b;\rho) = \{(x,y) : (x-a)^2 + (y-b)^2 < \rho^2\}$ centered at (a,b) of radius ρ.

Throughout the paper we let C denote a generic positive constant that may take different values in different formulas but is always independent of the mesh and ε. A subscripted C (e.g., C_1) denotes a positive constant that is independent of h and ε and takes a fixed value.

Remark 1.1. The assumption $b_u(x,y,u) \le \bar\beta$ in (1.2) can be omitted since it follows, for some constant $\bar\beta$, from $0 < \beta < b_u(x,y,u)$ and u being a unique and bounded solution of (1.1); see, e.g., [22, section 12]. Note that assumption (1.2) enables us to linearize (1.1) and then invoke the Green's function in our analysis. On the other hand, this assumption implies that our problem has a unique solution and thus excludes the possibility of multiple solutions. Hence, strictly speaking, our results are not applicable to the multiple-solution case. Having said this, we still believe that the present paper provides some insight into the numerical solution of a more general multiple-solution version of (1.1), considered, e.g., in [10].

2. Numerical method. Main result. Let our problem (1.1) satisfy the standard compatibility conditions at the corners of the domain Ω:

$$b(0,0,0) = b(0,1,0) = b(1,0,0) = b(1,1,0) = 0, \tag{2.1}$$

which guarantee that $u \in C^3(\bar\Omega)$.

Numerical method. We require the computed solution U to satisfy the standard five-point finite difference discretization of problem (1.1):

$$-\varepsilon^2 D_x^2 U_{ij} - \varepsilon^2 D_y^2 U_{ij} + b(x_i, y_j, U_{ij}) = 0, \tag{2.2}$$

for $i = 1, \dots N-1$, $j = 1, \dots, M-1$, where $U_{0,j} = U_{N,j} = U_{i,0} = U_{i,M} = 0$. Here, as usual, U_{ij} is associated with the mesh node (x_i, y_j), and we use the standard finite difference operators, defined for a discrete function V_{ij} by

$$\begin{aligned} D_x^- V_{ij} &= \frac{V_{ij} - V_{i-1,j}}{h_i}, \qquad D_x^2 V_{ij} = \frac{D_x^- V_{i+1,j} - D_x^- V_{i,j}}{(h_i + h_{i+1})/2}, \\ D_y^- V_{ij} &= \frac{V_{ij} - V_{i,j-1}}{k_j}, \qquad D_y^2 V_{ij} = \frac{D_y^- V_{i,j+1} - D_y^- V_{i,j}}{(k_j + k_{j+1})/2}. \end{aligned} \tag{2.3}$$

By condition (1.2), there exists a unique solution of the discrete problem (2.2) on an arbitrary mesh $\{(x_i, y_j)\}$; see, e.g., [5].

Clearly, $D_x^2 U_{ij}$ is defined for $i = 1, \dots, N-1$, $j = 0, \dots, M$, while $D_y^2 U_{ij}$ is defined for $i = 0, \dots, N$, $j = 1, \dots, M-1$. We now extend $D_x^2 U_{ij}$ to the mesh nodes $i = 0, N$ as follows. First, formally extend the discrete equation (2.2) to $i = 0$ and $i = N$, in which, using the zero boundary conditions, we set $D_y^2 U_{0,j} = D_y^2 U_{N,j} = 0$. This yields

$$D_x^2 U_{0,j} := \varepsilon^{-2} b(0, y_j, 0), \qquad D_x^2 U_{N,j} := \varepsilon^{-2} b(1, y_j, 0), \qquad j = 0, \dots, M. \tag{2.4a}$$

Similarly, extend $D_x^2 U_{ij}$ to $j = 0, M$ by

$$D_y^2 U_{i,0} := \varepsilon^{-2} b(x_i, 0, 0), \qquad D_y^2 U_{i,M} := \varepsilon^{-2} b(x_i, 1, 0), \qquad i = 0, \dots, N. \tag{2.4b}$$

Note that by (2.1), the above relations (2.4) imply that $D_x^2 U_{ij} = D_y^2 U_{ij} = 0$ if (x_i, y_j) is $(0,0)$, $(0,1)$, $(1,0)$, or $(1,1)$, which is consistent with the boundary condition in (1.1).

Remark 2.1. Now that $D_x^2 U_{ij}$ and $D_y^2 U_{ij}$ are extended by (2.4) to all i, j, our discrete equation (2.2) holds true for all $i = 0, \dots, N$ and $j = 0, \dots, M$.

Bilinear interpolation notation. Let $U^{\mathrm{B}} = U^{\mathrm{B}}(x, y)$ be the standard bilinear interpolant of the computed solution U_{ij}; i.e., U^{B} is continuous in $\bar{\Omega}$, bilinear on each $[x_{i-1}, x_i] \times [y_{j-1}, y_j]$, and equal to U_{ij} at the mesh nodes:

$$U^{\mathrm{B}}(x_i, y_j) = U_{ij} \qquad \text{for} \quad i = 0, \dots, N, \ j = 0, \dots, M. \tag{2.5}$$

Similarly, we introduce the bilinear interpolant $v^{\mathrm{B}}(x, y)$ for any discrete function v_{ij} or any continuous function $v(x, y)$.

Furthermore, we shall use the standard one-dimensional linear interpolants v^{I} and v^{J} with respect to x and y, respectively, that are defined, for any function v, as follows. For each fixed y in the domain of v, we have $v^{\mathrm{I}}(x_i, y) = v(x_i, y)$, and $v^{\mathrm{I}}(x, y)$ is linear on each $[x_{i-1}, x_i]$. Similarly, for each fixed x in the domain of v, we have $v^{\mathrm{J}}(x, y_j) = v(x, y_j)$, and $v^{\mathrm{J}}(x, y)$ is linear on each $[y_{j-1}, y_j]$.

In particular, consider the bilinear interpolant U^{B} of U_{ij} and the linear one-dimensional interpolants $U^{\mathrm{I}}(x, y_j)$ and $U^{\mathrm{J}}(x_i, y)$. Clearly,

$$[U^{\mathrm{I}}(x, y_j)]^{\mathrm{J}} = [U^{\mathrm{J}}(x_i, y)]^{\mathrm{I}} = U^{\mathrm{B}}(x, y). \tag{2.6}$$

Now we state a maximum norm a posteriori error estimate, which is the main result of the present paper.

THEOREM 2.1. *Let $u(x,y)$ be a solution of problem* (1.1), (1.2), (2.1), *U_{ij} a solution of problem* (2.2) *on an arbitrary mesh $\{(x_i,y_j)\}$, and $U^{\rm B}(x,y)$ its piecewise bilinear interpolant* (2.5). *Then*

$$\|U^{\rm B}-u\|_\infty \le C_0 \left[\max_{\substack{i=1,\dots,N\\ j=0,\dots,M}} \bigl\{h_i^2 M_{1,ij}\bigr\} + \max_{\substack{i=0,\dots,N\\ j=1,\dots,M}} \bigl\{k_j^2 M_{2,ij}\bigr\} \right], \tag{2.7}$$

where

$$M_{1,ij} := \min\bigl\{|D_x^2 U_{i-1,j}|, |D_x^2 U_{ij}|\bigr\} \ln\bigl(2+\varepsilon/\kappa\bigr) + \varepsilon |D_x^- D_x^2 U_{ij}| + |D_x^- U_{ij}|^2 + 1,$$
$$M_{2,ij} := \min\bigl\{|D_y^2 U_{i,j-1}|, |D_y^2 U_{ij}|\bigr\} \ln\bigl(2+\varepsilon/\kappa\bigr) + \varepsilon |D_y^- D_y^2 U_{ij}| + |D_y^- U_{ij}|^2 + 1,$$

with $\kappa := \min\{\min_i\{h_i\}, \min_j\{k_j\}\}$, while the constant C is independent of ε and the mesh.

Proof outline. By (1.1), we have

$$TU^{\rm B} - Tu = -\varepsilon^2 \left[\partial^2/\partial x^2 + \partial^2/\partial y^2\right] U^{\rm B} + b(x,y,U^{\rm B}),$$

where $\partial^2 U^{\rm B}/\partial x^2$ and $\partial^2 U^{\rm B}/\partial y^2$ are understood in the sense of distributions. Define an auxiliary function

$$q(x,y) := b(x,y,U^{\rm B}(x,y))$$

and let $q^{\rm B}$ denote its piecewise bilinear interpolant on the mesh $\{(x_i,y_j)\}$. Hence

$$TU^{\rm B} - Tu = -\varepsilon^2 \left[\partial^2/\partial x^2 + \partial^2/\partial y^2\right] U^{\rm B} + q^{\rm B} + \bigl[q - q^{\rm B}\bigr].$$

Noting that $q_{ij} := q(x_i,y_j) = b(x_i,y_j,U_{ij})$ and recalling the discrete equation (2.2) combined with Remark 2.1 yields $q_{ij} = \varepsilon^2 D_x^2 U_{ij} + \varepsilon^2 D_y^2 U_{ij}$ for $i=0,\dots,N$, $j=0,\dots,M$. Next, decompose this as $q_{ij} = q_{1,ij} + q_{2,ij}$, where

$$q_{1,ij} := \varepsilon^2 D_x^2 U_{ij}, \qquad q_{2,ij} := \varepsilon^2 D_y^2 U_{ij}, \qquad i=0,\dots,N,\ j=0,\dots,M. \tag{2.8}$$

Furthermore, using analogues of (2.6) for q_1 and q_2, we get

$$q^{\rm B}(x,y) = q_1^{\rm B}(x,y) + q_2^{\rm B}(x,y) = \bigl[q_1^{\rm I}(x,y_j)\bigr]^{\rm J} + \bigl[q_2^{\rm J}(x_i,y)\bigr]^{\rm I}, \qquad (x,y)\in\bar\Omega.$$

Therefore

$$TU^{\rm B} - Tu = \left[-\varepsilon^2 \frac{\partial^2 U^{\rm I}(x,y_j)}{\partial x^2} + q_1^{\rm I}(x,y_j)\right]^{\rm J} + \left[-\varepsilon^2 \frac{\partial^2 U^{\rm J}(x_i,y)}{\partial y^2} + q_2^{\rm J}(x_i,y)\right]^{\rm I} + \bigl[q - q^{\rm B}\bigr].$$

Here we used

$$\frac{\partial^2 U^{\rm B}(x,y)}{\partial x^2} = \left[\frac{\partial^2 U^{\rm I}(x,y_j)}{\partial x^2}\right]^{\rm J}, \qquad \frac{\partial^2 U^{\rm B}(x,y)}{\partial y^2} = \left[\frac{\partial^2 U^{\rm J}(x_i,y)}{\partial y^2}\right]^{\rm I}, \tag{2.9}$$

which follow from (2.6); see also Remark 2.2.

The proof is completed in section 5. First, the residual $TU^{\rm B} - Tu$ is represented as

$$TU^{\rm B} - Tu = \frac{\partial}{\partial x} F_1(x,y) + \frac{\partial}{\partial y} F_2(x,y) + \bigl[q - q^{\rm B}\bigr], \tag{2.10}$$

where F_1 and F_2 are functions of the current mesh and computed solution. This will enable us to estimate the error $U^{\mathrm{B}} - u$ in the maximum norm by linearizing the operator T and invoking its stability properties, which are obtained in section 4 using sharp estimates of the Green's function of section 3.

Remark 2.2. We understand $\partial^2 U^{\mathrm{B}}/\partial x^2$ and $\partial^2 U^{\mathrm{B}}/\partial y^2$ in the sense of distributions. To be more precise, in (2.9) we use $\partial^2 U^{\mathrm{I}}(x, y_j)/\partial x^2 = \sum_{i=1}^{N-1} \upsilon_i(y_j)\delta(x - x_i)$, and $[\partial^2 U^{\mathrm{I}}(x, y_j)/\partial x^2]^{\mathrm{J}} = \sum_{i=1}^{N-1} \upsilon_i^{\mathrm{J}}(y)\delta(x - x_i)$, where $\delta(\cdot)$ is the Dirac δ-distribution, while $\upsilon_i(y_j) := D^2 U_{ij}\,(h_i + h_{i+1})/2$.

3. Green's function. Assumption (1.2) enables us to linearize (1.1) and then invoke the Green's function in our analysis. Hence we start with a linear case of (1.1), where $b(x, y, u) := p(x, y)u - f(x, y)$:

$$Lu := -\varepsilon^2 \triangle u + p(x, y)u = f(x, y) \quad \text{in } \Omega, \qquad u = 0 \quad \text{on } \partial\Omega. \tag{3.1}$$

Here $p \in L_\infty(\Omega)$ and, in accordance with (1.2),

$$0 < \beta \le p(x, y) \le \bar{\beta}. \tag{3.2}$$

Introduce the Green's function $G(x, y; \xi, \eta)$ of the linear self-adjoint operator L that, for each $(x, y) \in \Omega$, satisfies

$$\begin{aligned} LG = -\varepsilon^2 \big[G_{\xi\xi} + G_{\eta\eta}\big] + p(\xi, \eta)G = \delta(x - \xi)\,\delta(y - \eta), \qquad (\xi, \eta) \in \Omega, \\ G(x, y; \xi, \eta) = 0, \qquad (\xi, \eta) \in \partial\Omega, \end{aligned} \tag{3.3}$$

where $\delta(\cdot)$ is the Dirac δ-distribution. Then the unique solution u of problem (3.1) is

$$u(x, y) = \int_\Omega G(x, y; \xi, \eta)\, f(\xi, \eta)\, d\xi\, d\eta. \tag{3.4}$$

Starting from (3.3), throughout the present section, the differential operators L and $\bar{L}$ are understood as differential operators in the variables (ξ, η). Furthermore, all norms are understood as norms of functions of (ξ, η).

THEOREM 3.1. *The Green's function $G(x, y; \xi, \eta)$ from* (3.3) *satisfies*

$$\big|G(x, y; \cdot)\big|_{1,1\,;\Omega} \le C\varepsilon^{-1}. \tag{3.5a}$$

Furthermore, for any ball $B(a, b\,; \rho)$ of radius ρ we have

$$\big|\,G(x, y; \cdot)\big|_{1,1\,;B(a,b\,;\rho)\cap\Omega} \le C\varepsilon^{-2}\rho, \tag{3.5b}$$

while for the ball $B(x, y\,; \rho)$ centered at (x, y) of radius ρ we have

$$\big|\,G(x, y; \cdot)\big|_{2,1\,;\Omega\setminus B(x,y\,;\rho)} \le C\varepsilon^{-2} \ln(2 + \varepsilon/\rho). \tag{3.5c}$$

3.1. Constant-coefficient case. First, we shall establish a particular case of Theorem 3.1. Let $p := \gamma^2$, where $\gamma = \text{const} > 0$, and let Ω be the quarter plane $\mathbb{R}^2_+ = \{x, y > 0\}$. In this particular case we denote the differential operator by $\bar{L}$ and the Green's function by $\bar{G}$, and for each (x, y) we have

$$\bar{L}\bar{G}(x, y; \xi, \eta) := -\varepsilon^2 \big[\bar{G}_{\xi\xi} + \bar{G}_{\eta\eta}\big] + \gamma^2 \bar{G} = \delta(x - \xi)\,\delta(y - \eta), \qquad \xi, \eta > 0. \tag{3.6}$$

The fundamental solution for the differential operator $\bar{L}$ is

$$g(x, y; \xi, \eta) := \frac{1}{2\pi\varepsilon^2} K_0\left(\frac{\gamma r}{\varepsilon}\right), \qquad r := \sqrt{(x - \xi)^2 + (y - \eta)^2}, \tag{3.7}$$

where K_0 is the modified Bessel function of the second kind of order zero [1]. Hence the Green's function for this differential operator over the quarter plane is

$$\bar{G}(x,y;\xi,\eta) = g(x,y;\xi,\eta) - g(-x,y;\xi,\eta) - g(x,-y;\xi,\eta) + g(-x,-y;\xi,\eta). \tag{3.8}$$

LEMMA 3.2. *For $\bar{G}(x,y;\xi,\eta)$ of* (3.8)*, estimates* (3.5) *hold true, in which G is replaced by $\bar{G}$ and Ω is replaced by $\mathbb{R}^2_+$.*

Proof. We shall prove estimates (3.5) only for $\|\bar{G}_\xi\|_1$ and $\|\bar{G}_{\xi\xi}\|_1$ here, since the estimates for $\|\bar{G}_\eta\|_1$ and $\|\bar{G}_{\eta\eta}\|_1$ with $\|\bar{G}_{\xi\eta}\|_1$ are obtained similarly. Furthermore, it suffices to show (3.5) with $\bar{G}$ replaced by the first term $g(x,y;\xi,\eta)$ of the representation (3.8) of $\bar{G}$, as the estimates for the other three terms are similar.

Since $K_0' = -K_1$, where K_1 is the modified Bessel function of the second kind of order one, and $\partial r/\partial\xi = -(x-\xi)/r$, we get

$$g_\xi(x,y;\xi,\eta) = \frac{\gamma}{2\pi\varepsilon^3} K_1\left(\frac{\gamma r}{\varepsilon}\right)\frac{x-\xi}{r}.$$

Furthermore, a similar calculation invoking $K_1'(s) = -K_0(s) - s^{-1}K_1(s)$ yields

$$g_{\xi\xi}(x,y;\xi,\eta) = \frac{\gamma^2}{2\pi\varepsilon^4}\left[K_0\left(\frac{\gamma r}{\varepsilon}\right) + \frac{K_1(\gamma r/\varepsilon)}{\gamma r/\varepsilon}\right]\frac{(x-\xi)^2}{r^2} - \frac{\gamma}{2\pi\varepsilon^3}K_1\left(\frac{\gamma r}{\varepsilon}\right)\frac{(y-\eta)^2}{r^3}.$$

Hence we have

$$|g_\xi| \le C\varepsilon^{-3}K_1(s), \qquad |g_{\xi\xi}(x,y;\xi,\eta)| \le C\varepsilon^{-4}\left[K_0(s) + \frac{K_1(s)}{s}\right], \qquad s := \frac{\gamma r}{\varepsilon}.$$

To obtain the desired bounds for $\|g_\xi\|_1$ and $\|g_{\xi\xi}\|_1$, we represent these integral norms in polar coordinates and then substitute $r = \varepsilon\gamma^{-1}s$; note that $d\xi\,d\eta = r\,dr\,d\varphi = C\varepsilon^2 s\,ds\,d\varphi$. For $\|g_\xi\|_{1;\mathbb{R}^2}$, this yields

$$\big\|g_\xi(x,y;\cdot)\big\|_{1;\mathbb{R}^2} \le C\varepsilon^{-3}\int_{\mathbb{R}^2} K_1(s)\,d\xi\,d\eta \le C\varepsilon^{-1}\int_0^\infty sK_1(s)\,ds \le C\varepsilon^{-1},$$

where we also used $0 < K_1(s) < Cs^{-1}e^{-s}$ [1]. Similarly,

$$\big\|g_\xi(x,y;\cdot)\big\|_{1;B(a,b;\rho)} \le C\varepsilon^{-3}\int_{B(x,y;\rho)} K_1(s)\,d\xi\,d\eta \le C\varepsilon^{-1}\int_0^{\gamma\rho/\varepsilon} sK_1(s)\,ds \le C\varepsilon^{-2}\rho.$$

Here replacing the integral over $B(a,b;\rho)$ by the integral over $B(x,y;\rho)$ yields an upper bound since K_1 is a positive decreasing function. Thus we obtained the desired estimates for $\|g_\xi\|_1$, which imply estimates (3.5a) and (3.5b) for $\|\bar{G}_\xi\|_1$.

Next, in a similar manner, we estimate $\|g_{\xi\xi}\|_1$:

$$\begin{aligned}\big\|g_{\xi\xi}(x,y;\cdot)\big\|_{1;\mathbb{R}^2\setminus B(x,y;\rho)} &\le C\varepsilon^{-2}\int_{\gamma\rho/\varepsilon}^\infty s\left[K_0(s) + \frac{K_1(s)}{s}\right]ds \le C\varepsilon^{-2}\left[1 + K_0\left(\frac{\rho}{\varepsilon}\right)\right]\\ &\le C\varepsilon^{-2}\ln\left(2 + \frac{\varepsilon}{\rho}\right),\end{aligned}$$

where, to get the second inequality, we used $0 < K_0(s) < Cs^{-1}e^{-s}$ and $K_1 = -K_0'$, while to get the final bound, we invoked $0 < K_0(s) < C\ln(2+s^{-1})$ [1]. This implies estimate (3.5c) for $\|\bar{G}_{\xi\xi}\|_1$. ∎

Remark 3.1. An inspection of the proof of Lemma 3.2, in which we used the explicit representation (3.8), (3.7) of the Green's function in the constant-coefficient case, shows that the estimates of the Green's function in Theorem 3.1 are sharp.

3.2. Smooth-coefficient case. Next, we shall use Lemma 3.2 to establish a variable-coefficient case of Theorem 3.1 under the additional assumption that

$$\|p\|_{C^{0,1}(\bar{\Omega})} \le C, \tag{3.9}$$

where $C^{0,1}(\bar{\Omega})$ is a standard Hölder space.

LEMMA 3.3. *Under condition* (3.9)*, the Green's function* $G(x,y\,;\xi,\eta)$ *from* (3.3) *satisfies estimates* (3.5a) *and* (3.5c)*.*

Proof. Fix $(x,y) \in \Omega$. Without loss of generality consider only the case of $x \le 1/2$ and $y \le 1/2$, as the other cases are similar. Let $\bar{G}$ be defined by (3.7), (3.8) with the frozen coefficient $\gamma^2 := p(x,y)$. By Lemma 3.2, estimates (3.5) hold true for $\bar{G}$. Hence is suffices to show that the function

$$v = v(x,y;\xi,\eta) := G(x,y;\xi,\eta) - \omega(\xi,\eta)\,\bar{G}(x,y;\xi,\eta), \tag{3.10}$$

where $\omega(\xi,\eta)$ is a smooth cut-off function that equals 1 on $[0,3/4] \times [0,3/4]$ and vanishes on the boundaries $\xi = 1$ and $\eta = 1$, satisfies

$$\varepsilon^2 \big|v(x,y;\cdot)\big|_{2,1\,;\Omega} + \varepsilon \big|v(x,y;\cdot)\big|_{1,1\,;\Omega} \le C. \tag{3.11}$$

Clearly, $v = 0$ for $(\xi,\eta) \in \partial\Omega$. Furthermore, using $LG = \bar{L}\bar{G}$ and (3.6), it is easy to check that $Lv = (\gamma^2 - p)\bar{G} + L[(\bar{G} - \omega\bar{G}]$, i.e.,

$$Lv = -\varepsilon^2[v_{\xi\xi} + v_{\eta\eta}] + p(\xi,\eta)\,v = \phi, \tag{3.12a}$$

$$\phi = \phi_1 + \phi_2, \quad \phi_1 := \big[\gamma^2 - p(\xi,\eta)\big]\bar{G}, \quad \phi_2 := L\big[(1-\omega)\bar{G}\big]. \tag{3.12b}$$

Let the stretching transformation from (ξ,η) to the new coordinates $\hat{\xi} := (\xi - x)/\varepsilon$ and $\hat{\eta} := (\eta - y)/\varepsilon$ map the original domain Ω into the domain $\hat{\Omega}$. Using the notation $\hat{w}(\hat{\xi},\hat{\eta}) := w(x,y;\xi,\eta)$ for any function w, rewrite (3.12a) as $-\triangle\hat{v} + \hat{p}\,\hat{v} = \hat{\phi}$. Combining this with $\hat{v} = 0$ on $\partial\hat{\Omega}$ yields $\|\hat{v}\|_{2;\hat{\Omega}} \le \beta^{-1/2}\|\hat{\phi}\|_{2;\hat{\Omega}}$ and, more importantly,

$$\big\|\hat{v}\big\|_{2,2\,;\hat{\Omega}} \le C_1\Big(\big\|\hat{\phi}\big\|_{2;\hat{\Omega}} + \big\|\hat{v}\big\|_{2;\hat{\Omega}}\Big) \le C_1\big(1 + \beta^{-1/2}\big)\,\big\|\hat{\phi}\big\|_{2;\hat{\Omega}}\,, \tag{3.13}$$

where the constant C_1 is independent of the size of the domain $\hat{\Omega}$ [17, Chapter 3, Lemma 8.1, p. 175]. Estimate (3.13), rewritten in the original variables (ξ,η), implies that

$$\varepsilon^2 \big|v(x,y;\cdot)\big|_{2,2\,;\Omega} + \varepsilon \big|v(x,y;\cdot)\big|_{1,2\,;\Omega} \le C\big\|\phi(x,y;\cdot)\big\|_{2\,;\Omega}\,. \tag{3.14}$$

Next, we claim that for ϕ from (3.12b) we have $\|\phi(x,y;\cdot)\|_{2\,;\Omega} \le C$. Indeed, $\phi_2 = 0$ for $(\xi,\eta) \in [0,3/4] \times [0,3/4]$ and $|\phi_2| \le C$ outside. Furthermore, condition (3.9) implies that $|\gamma^2 - p(\xi,\eta)| \le Cr$, where $\gamma^2 = p(x,y)$. Hence $|\phi_1| \le C\varepsilon^{-2} r K_0(\gamma r/\varepsilon)$, while

$$\big\|\phi_1\big\|_2^2 \le C\,\varepsilon^{-2}\int_\Omega (r/\varepsilon)^2 K_0^2(\gamma r/\varepsilon)\,d\xi\,d\eta \le C\int_0^\infty s^3\,K_0^2(s)\,ds \le C.$$

Thus we have established that $\|\phi\|_{L_2(\Omega)} \le C$. Combining this with (3.14) and $|v|_{k,1} \le C|v|_{k,2}$ for $k = 1,2$, we obtain (3.11) and complete the proof. ∎

Note that condition (3.9), under which we now proved Theorem 3.1, is suitable for the particular linear case (3.1) of problem (1.1). However, if we consider the semilinear case (1.1) and linearize (2.10) to the form (3.1), then the coefficient p depends on u and U and assumption (3.9) becomes unrealistic. Hence, we still have to prove Theorem 3.1 for the general case of p satisfying only (3.2).

3.3. General case. Proof of Theorem 3.1. Let $\bar G$ be defined by (3.8), (3.7) with the coefficient $\gamma^2 := \beta$. Then, by the maximum/comparison principle, $L[\bar G - G] = [p(\xi,\eta) - \beta]\,\bar G \ge 0$ combined with $\bar G - G \ge 0$ on $\partial\Omega$ implies that $0 \le G \le \bar G$.

Imitating the proof of Lemma 3.3, we arrive at (3.14), in which again $\|\phi_2\|_2 \le C$, while $|\gamma^2 - p(\xi,\eta)| \le C$ implies only that $|\phi_1| \le C\varepsilon^{-2}K_0(\gamma r/\varepsilon)$. Hence $\|\phi_1\|_2 \le C\varepsilon^{-1}$ and $\|\phi\|_2 \le C\varepsilon^{-1}$. Combining this with (3.14) and $|v|_{k,1\,;B(a,b\,;\rho)} \le C\rho|v|_{k,2\,;\Omega}$ or $|v|_{k,1\,;B(x,y\,;\varepsilon)} \le C\varepsilon|v|_{k,2\,;\Omega}$ for $k = 1, 2$, we obtain

$$\varepsilon\big|v(x,y;\cdot)\big|_{1,1\,;B(a,b\,;\rho)} \le C\varepsilon^{-1}\rho \tag{3.15}$$

and

$$\varepsilon^2\big|v(x,y;\cdot)\big|_{2,1\,;B(x,y\,;\varepsilon)} + \varepsilon\big|v(x,y;\cdot)\big|_{1,1\,;B(x,y\,;\varepsilon)} \le C\,; \tag{3.16}$$

compare with (3.11).

Combining (3.15) and (3.16) with Lemma 3.2, we observe that $G = v + \omega\bar G$ satisfies (3.5b), and furthermore, G satisfies (3.5a) and (3.5c), with Ω replaced by $B(x,y\,;\varepsilon)$. Hence, to complete the proof, it suffices to show that

$$\varepsilon^2\big|G(x,y;\cdot)\big|_{2,1\,;\Omega\setminus B(x,y\,;\varepsilon)} + \varepsilon\big|G(x,y;\cdot)\big|_{1,1\,;\Omega\setminus B(x,y\,;\varepsilon)} \le C\,. \tag{3.17}$$

Next, divide the domain $\Omega\setminus B(x,y\,;\varepsilon)$ into the subdomains $\mathcal{D}_j := \{(\xi,\eta) \in \Omega : \rho_j < r < \rho_{j+1}\}$, where $j = 0, 1, \dots$ and $\rho_j := 2^j\varepsilon$. Furthermore, $\mathcal{D}_j \subset \mathcal{D}'_j := \mathcal{D}_{j-1} \cup \bar{\mathcal{D}}_j \cup \mathcal{D}_{j+1}$ so that $\operatorname{dist}(\partial\mathcal{D}_j\setminus\partial\Omega\,,\partial\mathcal{D}'_j\setminus\partial\Omega) \ge \varepsilon/2$.

Let the stretching transformation from (ξ,η) to the new coordinates $\hat\xi := (\xi - x)/\varepsilon$ and $\hat\eta := (\eta - y)/\varepsilon$ map $\mathcal{D}_j$ into $\hat{\mathcal{D}}_j$. Imitating the proof of Lemma 3.3, we use the notation $\hat G(\hat\xi,\hat\eta) := G(x,y;\xi,\eta)$ and rewrite the equation from (3.3) as $-\triangle\hat G + \hat p\,\hat G = 0$ to get the local estimate

$$\big\|\hat G\big\|_{2,2\,;\hat{\mathcal{D}}_j} \le C_2\big\|\hat G\big\|_{2;\hat{\mathcal{D}}'_j} \tag{3.18}$$

[17, Chapter 3, estimate (8.6), p. 171]; compare with the global estimate (3.13). Here the constant C_2 is independent of ε since $\operatorname{dist}(\partial\hat{\mathcal{D}}_j\setminus\partial\hat\Omega\,,\partial\hat{\mathcal{D}}'_j\setminus\partial\hat\Omega) \ge 1/2$. Rewriting the above estimate in the variables (ξ,η) yields

$$\varepsilon^2\big|G(x,y;\cdot)\big|_{2,2;\mathcal{D}_j} + \varepsilon\big|G(x,y;\cdot)\big|_{1,2;\mathcal{D}_j} \le C\big\|G(x,y;\cdot)\big\|_{2,\mathcal{D}'_j} \le C\|g(x,y;\cdot)\|_{2,\mathcal{D}'_j}.$$

Here we also used $|G| \le \bar G \le 4g$; see (3.7), (3.8). Finally, since $|G|_{k,1\,;\mathcal{D}_j} \le C\rho_j|G|_{k,2\,;\mathcal{D}_j}$ for $k = 1, 2$, we arrive at

$$\varepsilon^2\big|G(x,y;\cdot)\big|_{2,1;\mathcal{D}_j} + \varepsilon\big|G(x,y;\cdot)\big|_{1,1;\mathcal{D}_j} \le C\rho_j\big\|g(x,y;\cdot)\big\|_{2,\mathcal{D}'_j}\,. \tag{3.19}$$

Note that

$$\big\|g(x,y;\cdot)\big\|^2_{2,\mathcal{D}'_j} \le C\varepsilon^{-2}\int_{\gamma\rho_{j-1}/\varepsilon}^{\gamma\rho_{j+1}/\varepsilon} sK_0^2(s)\,ds \le C\varepsilon^{-2}e^{-2s}\Big|_{\gamma\rho_{j+1}/\varepsilon}^{\gamma\rho_{j-1}/\varepsilon} \le C\varepsilon^{-2}e^{-\gamma\rho_j/\varepsilon}$$

since for $r \ge \varepsilon/2$ we have $g \le C\varepsilon^{-2}s^{-1/2}e^{-s}$, where $s = \gamma r/\varepsilon$. Furthermore,

$$\sum_{j=0}^{\infty} \rho_j \big\| g(x,y;\cdot) \big\|_{2,\mathcal{D}'_j} \le C \sum_{j=0}^{\infty} \frac{\gamma(\rho_j - \rho_{j-1})}{2\varepsilon}\, e^{-\gamma\rho_j/(2\varepsilon)} \le C \int_{\gamma/4}^{\infty} s e^{-s}\, ds \le C \tag{3.20}$$

since $\rho_j = 2(\rho_j - \rho_{j-1})$, while e^{-s} is decreasing. Combining (3.20) with (3.19), we get (3.17) and thus complete the proof. ∎

4. Stability properties of differential operators. The main result of the present section is the following stability theorem for the semilinear differential operator T from (1.1), which we shall further apply to relation (2.10).

Consider the right-hand side f in the special form

$$f(x,y) = -\frac{\partial}{\partial x}\big[F_1(x,y) + \bar{F}_1(x,y)\big] - \frac{\partial}{\partial y}\big[F_2(x,y) + \bar{F}_2(x,y)\big] + \bar{f}(x,y), \tag{4.1a}$$

where F_1, $\bar{F}_1$, F_2, $\bar{F}_2$, $f \in L_\infty(\Omega)$ and

$$F_1(x,y) = A_i(y)(x - x_{i-1/2}) \quad \text{for } (x,y) \in (x_{i-1}, x_i) \times [0,1], \quad i = 1, \dots, N, \tag{4.1b}$$

$$F_2(x,y) = B_j(x)(y - y_{j-1/2}) \quad \text{for } (x,y) \in [0,1] \times (y_{j-1}, y_j), \quad j = 1, \dots, M, \tag{4.1c}$$

while $x_{i-1/2} := x_i - h_i/2$ and $y_{j-1/2} := y_j - k_j/2$.

THEOREM 4.1. *Let the function b in* (1.1) *satisfy* (1.2). *Then, for any v, $w \in W^{1,2}(\Omega)$ such that $Tv(x,y) - Tw(x,y) = f(x,y)$, where f is defined by* (4.1), *and $v(x,y) = w(x,y)$ for $(x,y) \in \partial\Omega$, we have*

$$\begin{aligned} \|v - w\|_\infty \le{}& C\varepsilon^{-1}\big(\|\bar{F}_1\|_\infty + \|\bar{F}_2\|_\infty\big) + \beta^{-1}\|\bar{f}\|_\infty \\ &+ C\left[\max_{i=1,\dots,N}\left\{\frac{h_i^2}{\varepsilon^2}\max_{y\in[0,1]}\big|A_i(y)\big|\right\} + \max_{j=1,\dots,M}\left\{\frac{k_j^2}{\varepsilon^2}\max_{x\in[0,1]}\big|B_j(x)\big|\right\}\right]\ln\left(2 + \frac{\varepsilon}{\kappa}\right). \end{aligned}$$

The proof is deferred to section 4.2

4.1. Linear reaction-diffusion. First, we address the linear problem (3.1), (3.2) with the right-hand side (4.1). Since the differential operator L is linear, it is convenient to establish stability of u with respect to various components of f separately.

LEMMA 4.2. *There exists a unique solution $u \in L_\infty(\Omega)$ of problem* (3.1), (3.2) *with the right-hand side* (4.1a). *Furthermore, if $F_1 = F_2 := 0$, then*

$$\|u\|_\infty \le \beta^{-1}\|\bar{f}\|_\infty + C\varepsilon^{-1}\big(\|\bar{F}_1\|_\infty + \|\bar{F}_2\|_\infty\big). \tag{4.2}$$

Proof. Since L is linear, it suffices to establish the desired estimate in the following two cases. Case A: $f = \bar{f}$, while $\bar{F}_1 = \bar{F}_2 := 0$. Then estimate (4.2) is well known and follows from the maximum/comparison principle extended to functions in the Sobolev space $W^{1,2}$ [9, section 8.1]. Case B: $\bar{f} := 0$. Now the desired estimate (4.2) follows from (3.4) combined with estimate (3.5a). ∎

LEMMA 4.3. *Let $\bar{F}_1 = \bar{F}_2 = \bar{f} := 0$ in* (4.1a). *Then the solution $u \in L_\infty(\Omega)$ of problem* (3.1), (3.2) *with the right-hand side* (4.1) *satisfies*

$$\|u\|_\infty \le C\left[\max_{i=1,\dots,N}\left\{\frac{h_i^2}{\varepsilon^2}\max_{y\in[0,1]}\big|A_i(y)\big|\right\} + \max_{j=1,\dots,M}\left\{\frac{k_j^2}{\varepsilon^2}\max_{x\in[0,1]}\big|B_j(x)\big|\right\}\right]\ln\left(2 + \frac{\varepsilon}{\kappa}\right),$$

where $\kappa := \min\{\min_i\{h_i\}, \min_j\{k_j\}\}$.

Proof. It suffices to consider only the case of $f := -\partial F_1/\partial x$, i.e., $F_2 := 0$, as the case of $f := -\partial F_2/\partial y$ is considered similarly, while our differential operator L is linear.

Fix (x, y) and denote $v(\xi, \eta) := G(x, y; \xi, \eta)$. Then, by (3.4), we have

$$u(x,y) = \int_\Omega F_1(\xi,\eta) v_\xi(\xi,\eta)\, d\xi\, d\eta = \sum_{i=1}^N \int_{\Omega_i} A_i(\eta)(\xi - x_{i-1/2})\, v_\xi(\xi)\, d\xi\, d\eta, \tag{4.3}$$

where $\Omega_i := (x_{i-1}, x_i) \times (0,1),\ i = 1, \dots, N$. Furthermore, let $x \in [x_{m-1/2}, x_{m+1/2}]$ for some $0 < m < N$ (the cases of $x \in [0, x_{1/2}]$ and $x \in [x_{N-1/2}, 1]$ are similar) and introduce the rectangular domain

$$\Omega' := (x_{m-1}, x_{m+1}) \times (y - \tilde h_m, y + \tilde h_m), \quad \text{where } \tilde h_m := \min\{h_m, h_{m+1}\}/2,$$

so that

$$B(x, y\,; \tilde h_m) \subset \Omega' \subset B(x_{m-1/2}, y\,; h_m) \cup B(x_{m+1/2}, y; h_{m+1}). \tag{4.4}$$

Clearly, (4.3) can be written as $u(x, y) = S_1 + S_2$, where

$$S_1 = \sum_{i=1}^N \int_{\Omega_i} A_i(\eta)(\xi - x_{i-1/2})\, \tilde v_\xi(\xi,\eta)\, d\xi\, d\eta\,,$$

with $\tilde v := v$ in $\Omega_i \setminus \Omega'$ and $\tilde v := 0$ in Ω', while

$$S_2 = \sum_{i=m}^{m+1} \int_{\Omega_i \cap \Omega'} A_i(\eta)(\xi - x_{i-1/2})\, v_\xi(\xi,\eta)\, d\xi\, d\eta\,.$$

To estimate S_1, note that

$$\tilde v_\xi(\xi,\eta) = \tilde v_\xi(x_{i-1},\eta) + \int_{x_{i-1}}^\xi \tilde v_{\xi\xi}(s,\eta)\, ds \qquad \text{for } \ \xi \in (x_{i-1}, x_i). \tag{4.5}$$

Here $\tilde v_{\xi\xi}$ is well defined, since the singularity of v occurs at (x, y), which is inside Ω', and furthermore, for each i and η either $\tilde v_{\xi\xi} = v_{\xi\xi}$ or $\tilde v_{\xi\xi} = 0$ for all $\xi \in (x_{i-1}, x_i)$. Combining (4.5) with $\int_{x_{i-1}}^{x_i} (\xi - x_{i-1/2})\, d\xi = 0$ yields

$$\begin{aligned}\left| \int_{x_{i-1}}^{x_i} (\xi - x_{i-1/2})\, \tilde v_\xi(\xi,\eta)\, d\xi \right| &= \left| \int_{x_{i-1}}^{x_i} d\xi\, (\xi - x_{i-1/2}) \int_{x_{i-1}}^\xi \tilde v_{\xi\xi}(s,\eta)\, ds \right| \\ &\le \frac{h_i^2}{4} \int_{x_{i-1}}^{x_i} \left|\tilde v_{\xi\xi}(s,\eta)\right| ds\,.\end{aligned}$$

Hence

$$\begin{aligned}|S_1| &\le \sum_{i=1}^N \frac{h_i^2}{4} \int_0^1 |A_i(\eta)| d\eta \int_{x_{i-1}}^{x_i} |\tilde v_{\xi\xi}(s,\eta)|\, ds \\ &\le \max_{i=1,\dots,N} \left\{ \frac{h_i^2}{4} \max_{\eta\in[0,1]} |A_i(\eta)| \right\} \int_\Omega |\tilde v_{\xi\xi}(\xi,\eta)|\, d\xi\, d\eta\,.\end{aligned}$$

Finally, recalling (4.4) and estimate (3.5c) for $v_{\xi\xi} = G_{\xi\xi}$, we get

$$\int_\Omega |\tilde{v}_{\xi\xi}(\xi,\eta)|\,d\xi\,d\eta \le \int_{\Omega\setminus B(x,y;\tilde{h}_m)} |v_{\xi\xi}(\xi,\eta)|\,d\xi\,d\eta \le C\varepsilon^{-2}\ln(2+\varepsilon/\tilde{h}_m)$$

and thus the desired estimate for S_1:

$$|S_1| \le C \max_{i=1,\dots,N}\left\{\frac{h_i^2}{\varepsilon^2}\max_{\eta\in[0,1]}\bigl|A_i(\eta)\bigr|\right\}\ln\left(2+\frac{\varepsilon}{\kappa}\right).$$

It remains to obtain a similar estimate for S_2:

$$\begin{aligned} S_2 &\le \sum_{i=m}^{m+1} h_i \max_{\eta\in[0,1]}|A_i(\eta)| \int_{B(x_{i-1/2},y;h_i)} |v_\xi(\xi,\eta)|\,d\xi\,d\eta \le \sum_{i=m}^{m+1}\max_{\eta\in[0,1]}|A_i(\eta)|h_i\frac{Ch_i}{\varepsilon^2} \\ &\le C\max_i\left\{\frac{h_i^2}{\varepsilon^2}\max_{\eta\in[0,1]}|A_i(\eta)|\right\}. \end{aligned}$$

Here we used (4.4) and estimate (3.5b) for $v_\xi = G_\xi$. ∎

4.2. Semilinear reaction-diffusion. Proof of Theorem 4.1. Using the standard linearization technique, we have $Tv(x,y) - Tw(x,y) = L\bigl[v(x,y) - w(x,y)\bigr]$, where the operator L is linear and defined by (3.1) with the coefficient $p(x,y) = \int_0^1 b_u(x,y,w(x,y)+s[v(x,y)-w(x,y)])\,ds$, which, by (1.2), satisfies condition (3.2). Hence the desired estimate follows from Lemmas 4.2 and 4.3. ∎

5. Analysis of the numerical method. Proof of Theorem 2.1. To complete the proof of Theorem 2.1, which we started in section 2, we shall invoke the following lemma.

LEMMA 5.1. *We have*

$$-\varepsilon^2\frac{\partial^2 U^{\mathrm{I}}(x,y_j)}{\partial x^2} + q_1^{\mathrm{I}}(x,y_j) = \frac{\partial}{\partial x}F_1(x,y_j), \quad -\varepsilon^2\frac{\partial^2 U^{\mathrm{J}}(x_i,y)}{\partial y^2} + q_2^{\mathrm{J}}(x_i,y) = \frac{\partial}{\partial y}F_2(x_i,y),$$

where

$$F_1(x,y_j) := q_{1,ij}(x-x_{i-1/2}) + \frac{D_x^- q_{1,ij}}{2}(x_i-x)^2, \qquad x\in(x_{i-1},x_i), \tag{5.1a}$$

for $i = 1,\dots,N$, $j = 0,\dots,M$, *and*

$$F_2(x_i,y) := q_{2,ij}(y-y_{j-1/2}) + \frac{D_y^- q_{2,ij}}{2}(y_j-y)^2, \qquad y\in(y_{i-1},y_i), \tag{5.1b}$$

for $i = 0,\dots,N$, $j = 1,\dots,M$.

Proof. We closely imitate the one-dimensional argument used in the proof of [12, Theorem 3.3] and give the proof here only for completeness.

It suffices to obtain the first desired relation, as the other one is similar. To simplify the presentation, within this proof, fix y_j and use the notation $U^{\mathrm{I}}(x) := U^{\mathrm{I}}(x,y_j)$, $q_1^{\mathrm{I}}(x) := q_1^{\mathrm{I}}(x,y_j)$, $q_{1,i} := q_{1,ij}$, and $F_1(x) := F(x,y_j)$. Furthermore, for any function v, let $v' := \partial v/\partial x$. Thus we intend to show that $-\varepsilon^2(U^{\mathrm{I}})'' + q_1^{\mathrm{I}} = F_1'$.

First, note that

$$-\varepsilon^2(U^{\mathrm{I}})'' + q_1^I = -\left[\varepsilon^2(U^{\mathrm{I}})' + \int_x^1 q_1^I(s)\,\mathrm{d}s\right]'. \tag{5.2}$$

Recalling (2.3) and (2.8), we observe that

$$\varepsilon^2(U^{\mathrm{I}})' = \varepsilon^2 D_x^- U_i^{\mathrm{I}} = \varepsilon^2 D_x^- U_N^{\mathrm{I}} - \sum_{m=i}^{N-1} q_{1,m} \frac{h_m + h_{m+1}}{2}\,, \quad x \in (x_{i-1}, x_i), \quad i = 1, \dots, N.$$

Now, substituting the above representation in (5.2) and omitting the derivative of the constant $\varepsilon^2 D^- U_N^{\mathrm{I}}$, we arrive at $-\varepsilon^2(U^{\mathrm{I}})'' + q_1^I = \tilde{F}_1'$, where

$$\tilde{F}_1(x) := \sum_{m=i}^{N-1} q_{1,m} \frac{h_m + h_{m+1}}{2} - \int_x^1 q_1^I(s)\,\mathrm{d}s, \quad x \in (x_{i-1}, x_i), \quad i = 1, \dots, N.$$

A calculation shows that

$$\sum_{m=i}^{N-1} q_{1,m} \frac{h_m + h_{m+1}}{2} = q_{1,i} \frac{h_i}{2} + \int_{x_i}^1 q_1^I(s)\,\mathrm{d}s - q_{1,N} \frac{h_N}{2},$$

and, omitting the derivative of another constant $q_{1,N} h_N/2$, we obtain $\tilde{F}_1' = F_1'$, where

$$F_1(x) := q_{1,i} \frac{h_i}{2} - \int_x^{x_i} q_1^I(s)\,\mathrm{d}s, \quad x \in (x_{i-1}, x_i), \quad i = 1, \dots, N. \tag{5.3}$$

Thus we have obtained the desired relation $-\varepsilon^2(U^{\mathrm{I}})'' + q_1^I = F_1'$, in which F_1 is defined by (5.3), and it remains to show that the definition of F_1 in (5.1a) is equivalent to (5.3). Indeed, by computing $\int_x^{x_i} q^I(s)\,\mathrm{d}s$ in (5.3) using $q_1^I(x) = q_{1,i} - (x_i - x)\, D^- q_{1,i}$, we get (5.1a). □

Remark 5.1. One can easily check that F_1 and F_2 or (5.1) allow an alternative representation:

$$F_1(x, y_j) := q_{1,i-1,j}(x - x_{i-1/2}) + \frac{D_x^- q_{1,ij}}{2}\, O(h_i^2), \qquad x \in (x_{i-1}, x_i),$$

$$F_2(x_i, y) := q_{2,i,j-1}(y - y_{j-1/2}) + \frac{D_y^- q_{2,ij}}{2}\, O(k_j^2), \qquad y \in (y_{i-1}, y_i).$$

E.g., the new representation of F_1 follows from $q_{1,ij} = q_{1,i-1,j} + h_i\, D_x^- q_{1,ij}$.

Proof of Theorem 2.1 (continued from section 2). Extending F_1 and F_2 of (5.1) onto the whole domain $\bar{\Omega}$ by linear interpolation,

$$F_1(x, y) := \big[F_1(x, y_j)\big]^{\mathrm{J}}, \qquad F_2(x, y) := \big[F_2(x_i, y)\big]^{\mathrm{I}},$$

we obtain (2.10). Now, invoking Theorem 4.1 yields

$$\begin{aligned}\|U^{\mathrm{B}} - u\|_\infty \le C &\left[\max_{\substack{i=1,\dots,N\\ j=0,\dots,M}} \left\{ \frac{h_i^2}{\varepsilon^2} |q_{1,ij}| \right\} + \max_{\substack{i=0,\dots,N\\ j=1,\dots,M}} \left\{ \frac{k_j^2}{\varepsilon^2} |q_{2,ij}| \right\} \right] \ln\Big(2 + \frac{\varepsilon}{\kappa}\Big) \\ &+ C \left[\max_{\substack{i=1,\dots,N\\ j=0,\dots,M}} \left\{ \frac{h_i^2}{\varepsilon} |D_x^- q_{1,ij}| \right\} + \max_{\substack{i=0,\dots,N\\ j=1,\dots,M}} \left\{ \frac{k_j^2}{\varepsilon} |D_y^- q_{2,ij}| \right\} \right] + \beta^{-1} \|q - q^{\mathrm{B}}\|_\infty.\end{aligned} \tag{5.4}$$

Combining this with (2.8) and the bilinear interpolation estimate

$$\|q - q^{\mathrm{B}}\|_\infty \le C \left[\max_{\substack{i=1,\dots,N\\ j=0,\dots,M}} \big\{ h_i^2 (1 + |D_x^- U_{ij}|^2) \big\} + \max_{\substack{i=0,\dots,N\\ j=1,\dots,M}} \big\{ k_j^2 (1 + |D_y^- U_{ij}|^2) \big\} \right], \tag{5.5}$$

we obtain a version of the desired a posteriori error estimate (2.7) in which the quantities $\min\{|D_x^2 U_{i-1,j}|, |D_x^2 U_{ij}|\}$ and $\min\{|D_y^2 U_{i,j-1}|, |D_y^2 U_{ij}|\}$ are replaced by $|D_x^2 U_{ij}|$ and $|D_y^2 U_{ij}|$, respectively.

By Remark 5.1, the quantities $|q_{1,ij}|$ and $|q_{2,ij}|$ in (5.4) can be replaced by $\min\{|q_{1,i-1,j}|, |q_{1,ij}|\}$ and $\min\{|q_{2,i,j-1}|, |q_{2,ij}|\}$, respectively. Combining this sharper version of (5.4) with (2.8) and (5.5) yields the desired estimate (2.7).

Finally, note that the interpolation error estimate (5.5), which we used, follows from $q - q^B = [q - q^I] + [q^I - (q^I)^J]$ combined with $|\partial^2 q(x, y_j)/\partial x^2| \le C(1 + |D_x^- U_{ij}|^2)$ for $x \in (x_{i-1}, x_i)$ and $|\partial^2 q(x_i, y)/\partial y^2| \le C(1 + |D_y^- U_{ij}|^2)$ for $y \in (y_{j-1}, y_j)$; see [3, Comment 2.15] for a similar argument. ∎

6. Numerical results. The maximum norm a posteriori error estimate of Theorem 2.1 implies that

$$e := \|U^{\mathrm{B}} - u\|_\infty \le \tilde{C}\eta, \qquad \eta := \max\{\eta_0, \eta_1, \eta_2, \eta_3\}, \tag{6.1}$$

$$\eta_l := \max\left\{ \max_{\substack{i=1,\dots,N \\ j=0,\dots,M}} \{h_i^2 M_{1,ij}^{(l)}\} \; ; \; \max_{\substack{i=0,\dots,N \\ j=1,\dots,M}} \{k_j^2 M_{2,ij}^{(l)}\} \right\}, \qquad l = 0, 1, 2, 3,$$

where $\tilde{C} = C \ln\bigl(2 + \varepsilon/\kappa\bigr)$, while $M_{1,ij}^{(0)} = M_{2,ij}^{(0)} = 1$,

$$M_{1,ij}^{(1)} := |D_x^- U_{ij}|^2, \quad M_{1,ij}^{(2)} := \min\{|D_x^2 U_{i-1,j}|\,, |D_x^2 U_{ij}|\}, \quad M_{1,ij}^{(3)} := \varepsilon |D_x^- D_x^2 U_{ij}|,$$

$$M_{2,ij}^{(1)} := |D_y^- U_{ij}|^2, \quad M_{2,ij}^{(2)} := \min\{|D_y^2 U_{i,j-1}|\,, |D_y^2 U_{ij}|\}, \quad M_{2,ij}^{(3)} := \varepsilon |D_y^- D_y^2 U_{ij}|.$$

Here η_l and $M^{(l)}$, $l = 1, 2, 3$, involve discrete analogues of lth-order derivatives.

In this section we present numerical results on a priori chosen meshes to investigate the efficiency of the upper maximum norm error estimator η in (6.1) and its components η_l. It is also of interest which of η_l is the principal component in η if any. We shall examine the errors e and, more importantly, the quantities η, e/η, η_l, e/η_l and their dependence on ε, $N = M$, and particular mesh choices.

We consider $\varepsilon = 10^{-k}$, $k = 1, \dots, 10$, and two tensor-product meshes with $M = N$: a variant of the layer-adapted mesh by Bakhvalov [4] and a simple uniform mesh; see Tables 6.1–6.5. Note that a Bakhvalov-type layer-adapted mesh was chosen for the numerical experiments, since it yields ε-uniform second-order accuracy [4, 13]. Furthermore, we expect a robust adaptive algorithm to generate a mesh that is very close to a Bakhvalov mesh, as in [14, section 6 and Figure 2].

To be precise, if $\varepsilon \le \bar{\varepsilon}$, our Bakhvalov-type mesh is given by $x_i = y_i := x(i/N)$ for $i = 0, 1, \dots, N$, where $x(\xi) := \varepsilon\lambda \ln[b/(b - \xi)]$ for $\xi \in [0, \theta]$, $x(1) := 1$, and $x(\xi)$ is continuous on $[0, 1]$ and linear on $[\theta, 1]$. We use the constants $b = 1/2$, $\bar{\varepsilon} = b/\lambda$, and $\theta = b - \varepsilon\lambda$. The constant λ will be specified later. For $\varepsilon > \bar{\varepsilon}$, the Bakhvalov mesh is defined to be a simple uniform mesh.

Our test problem is the linear problem (3.1), in which $p(x, y) := 1$, and whose exact solution

$$u(x, y) := \left(\cos(\pi x) - \frac{e^{-x/\varepsilon} - e^{-1/\varepsilon}}{1 - e^{-1/\varepsilon}}\right)\left(1 - y - \frac{e^{-y/\varepsilon} - e^{-1/\varepsilon}}{1 - e^{-1/\varepsilon}}\right)$$

exhibits boundary layers and a corner layer.

TABLE 6.1

Bakhvalov mesh, $\lambda = 3$: Maximum norm error e and the efficiency constant e/η for the upper maximum norm error estimator η.

	$\varepsilon = 10^{-1}$		$\varepsilon = 10^{-2}$		$\varepsilon = 10^{-3}$		$\varepsilon = 10^{-k}$, $k = 4, \dots, 10$	
N	e	e/η	e	e/η	e	e/η	e	e/η
32	3.60e-3	1.25e-1	4.04e-3	1.19e-1	4.21e-3	1.23e-1	4.23e-3	1.24e-1
64	9.16e-4	1.25e-1	1.02e-3	1.19e-1	1.07e-3	1.24e-1	1.08e-3	1.24e-1
128	2.32e-4	1.26e-1	2.58e-4	1.19e-1	2.70e-4	1.24e-1	2.72e-4	1.25e-1
256	5.85e-5	1.27e-1	6.46e-5	1.19e-1	6.78e-5	1.24e-1	6.82e-5	1.25e-1
512	1.47e-5	1.27e-1	1.62e-5	1.19e-1	1.70e-5	1.24e-1	1.71e-5	1.25e-1

TABLE 6.2

Bakhvalov mesh, $\lambda = 3$: Upper maximum norm error estimator η, its components η_1, η_2, η_3, and its efficiency constant e/η.

	$\varepsilon = 10^{-1}$				$\varepsilon = 10^{-k}$, $k = 4, \dots, 10$			
N	η_1	η_2	$\eta_3 = \eta$	e/η	η_1	η_2	$\eta_3 = \eta$	e/η
32	1.75e-2	2.61e-2	2.88e-2	1.25e-1	3.10e-2	3.09e-2	3.41e-2	1.24e-1
64	4.72e-3	6.96e-3	7.31e-3	1.25e-1	8.26e-3	8.25e-3	8.65e-3	1.24e-1
128	1.22e-3	1.80e-3	1.84e-3	1.26e-1	2.13e-3	2.13e-3	2.18e-3	1.25e-1
256	3.11e-4	4.57e-4	4.62e-4	1.27e-1	5.41e-4	5.41e-4	5.47e-4	1.25e-1
512	7.86e-5	1.15e-4	1.16e-4	1.27e-1	1.36e-4	1.36e-4	1.37e-4	1.25e-1

TABLE 6.3

Uniform mesh: Maximum norm error e and the efficiency constant e/η_2 for the component η_2 of the upper maximum norm error estimator η.

	$\varepsilon = 10^{-2}$		$\varepsilon = 10^{-3}$		$\varepsilon = 10^{-4}$		$\varepsilon = 10^{-k}$, $k = 5, \dots, 10$	
N	e	e/η_2	e	e/η_2	e	e/η_2	e	e/η_2
32	3.45e-1	4.17e-1	5.68e-1	5.71e-1	5.77e-1	5.78e-1	5.78e-1	5.79e-1
64	1.60e-1	2.77e-1	5.54e-1	5.59e-1	5.77e-1	5.77e-1	5.78e-1	5.78e-1
128	5.40e-2	1.91e-1	5.11e-1	5.28e-1	5.76e-1	5.77e-1	5.78e-1	5.78e-1
256	1.57e-2	1.53e-1	4.04e-1	4.56e-1	5.73e-1	5.74e-1	5.78e-1	5.78e-1
512	4.21e-3	1.35e-1	2.20e-1	3.24e-1	5.63e-1	5.66e-1	5.78e-1	5.78e-1

TABLE 6.4

Uniform mesh: The components η_2 and η_3 of the upper maximum norm error estimator η and the efficiency constant e/η_2 for η_2.

	$\varepsilon = 10^{-4}$			$\varepsilon = 10^{-7}$			$\varepsilon = 10^{-10}$		
N	η_2	$\eta_3 = \eta$	e/η_2	η_2	$\eta_3 = \eta$	e/η_2	η_2	$\eta_3 = \eta$	e/η_2
32	9.99e-1	3.12e+2	5.78e-1	9.99e-1	3.12e+5	5.79e-1	9.99e-1	3.12e+8	5.79e-1
64	1.00e+0	1.56e+2	5.77e-1	1.00e+0	1.56e+5	5.78e-1	1.00e+0	1.56e+8	5.78e-1
128	9.99e-1	7.81e+1	5.77e-1	1.00e+0	7.81e+4	5.78e-1	1.00e+0	7.81e+7	5.78e-1
256	9.99e-1	3.90e+1	5.74e-1	1.00e+0	3.91e+4	5.78e-1	1.00e+0	3.91e+7	5.78e-1
512	9.95e-1	1.95e+1	5.66e-1	1.00e+0	1.95e+4	5.78e-1	1.00e+0	1.95e+7	5.78e-1

TABLE 6.5

Bakhvalov mesh, $\lambda = 1$: Maximum norm error e, upper maximum norm error estimator η, its components η_1, η_2, η_3, and its efficiency constant e/η.

	$\varepsilon = 10^{-5}$					$\varepsilon = 10^{-10}$				
N	e	η_1	η_2	$\eta_3 = \eta$	e/η	e	η_1	η_2	$\eta_3 = \eta$	e/η
32	7.50e-2	9.60e-3	1.67e-2	2.13e-1	3.52e-1	9.48e-2	9.61e-3	1.58e-2	2.41e-1	3.93e-1
64	3.81e-2	2.41e-3	8.37e-3	1.05e-1	3.64e-1	4.91e-2	2.41e-3	7.93e-3	1.20e-1	4.09e-1
128	1.89e-2	6.02e-4	4.22e-3	5.15e-2	3.66e-1	2.49e-2	6.02e-4	3.98e-3	5.94e-2	4.19e-1
256	9.17e-3	1.51e-4	2.13e-3	2.52e-2	3.64e-1	1.25e-2	1.51e-4	1.99e-3	2.95e-2	4.23e-1
512	4.40e-3	3.76e-5	1.07e-3	1.23e-2	3.58e-1	6.22e-3	3.76e-5	9.96e-4	1.46e-2	4.25e-1

Tables 6.1 and 6.2 present numerical results for the Bakhvalov mesh with $\lambda = 3$. This mesh yields ε-uniform second-order accuracy in the maximum norm; i.e., ultimately, we would like to be able to construct a similar adaptive mesh. We observe agreement with our theoretical estimate (6.1). Not only does e/η stabilize—see Table 6.1—but it becomes very close to the linear interpolation error constant $1/8 = 1.25e - 1$. Table 6.2 is given to compare the components η_l of η. We observe that $\eta_2 \approx \eta_3 = \eta$. Furthermore, for $\varepsilon \leq 10^{-2}$ we have $\eta_1 \approx \eta_2 \approx \eta_3$, while for $\varepsilon = 10^{-1}$ the quantity η_1 is dominated by η_2 and η_3. The quantity η_0 is not presented, since it is negligible and, furthermore, known a priori.

When uniform meshes are used—see Tables 6.3 and 6.4—the boundary layers are not resolved and $e = O(1)$. This is indicated by $\eta = \eta_3$ blowing up even more significantly than e. Unlike η_3 the component η_2 remains bounded. Thus both η_2 and η_3 not being small correctly indicates that the method is inaccurate. But η_2 better reflects the actual errors since $e/\eta_2 \approx \text{const} = 0.58$ in Table 6.4.

Finally we consider the Bakhvalov mesh with $\lambda = 1$; see Table 6.5. Since the condition $\lambda > 2$, which implies ε-uniform second-order accuracy for our test problem [4, 13], is violated, the errors slightly decrease as $\varepsilon \to 0$. We observe that η_1 is too small compared to η and e.

In summary, for our test problem on the meshes considered, the error estimator η indicates correctly whether or not the method is ε-uniformly accurate. Furthermore, we observe that the quantity $\eta = \eta_3$ might blow up (see Table 6.4), while the component η_1 is sometimes too optimistic (see Table 6.5). The component η_2 seems the most relevant estimator for the actual error e. Besides, η_2 does not blow up, like η_3, and hence seems a suitable error indicator for a posteriori mesh construction.

REFERENCES

[1] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, National Bureau of Standards Applied Mathematics Series 55, U.S. Government Printing Office, Washington, D.C., 1964.

[2] M. Ainsworth and I. Babuška, *Reliable and robust a posteriori error estimation for singularly perturbed reaction-diffusion problems*, SIAM J. Numer. Anal., 36 (1999), pp. 331–353.

[3] T. Apel, *Anisotropic Finite Elements: Local Estimates and Applications*, Teubner, Stuttgart, 1999.

[4] N. S. Bakhvalov, *Towards optimization of methods for solving boundary value problems in the presence of a boundary layer*, Zh. Vychisl. Mat. i Mat. Fiz., 9 (1969), pp. 841–859 (in Russian).

[5] I. Boglaev, *On monotone iterative methods for a nonlinear singularly perturbed reaction-diffusion problem*, J. Comput. Appl. Math., 162 (2004), pp. 445–466.

[6] L. Chen, P. Sun, and J. Xu, *Optimal anisotropic simplicial meshes for minimizing interpolation errors in L_p-norm*, Math. Comp., 76 (2007), pp. 179–204.

[7] C. Clavero, J. L. Gracia, and E. O'Riordan, *A parameter robust numerical method for a two dimensional reaction-diffusion problem*, Math. Comp., 74 (2005), pp. 1743–1758.

[8] E. F. D'Azevedo, *Optimal triangular mesh generation by coordinate transformation*, SIAM J. Sci. Statist. Comput., 12 (1991), pp. 755–786.

[9] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin, 1998.

[10] W. Huang, *Variational mesh adaptation: Isotropy and equidistribution*, J. Comput. Phys., 174 (2001), pp. 903–924.

[11] N. Kopteva, *Maximum norm a posteriori error estimates for a one-dimensional convection-diffusion problem*, SIAM J. Numer. Anal., 39 (2001), pp. 423–441.

[12] N. Kopteva, *Maximum norm a posteriori error estimates for a 1D singularly perturbed semilinear reaction-diffusion problem*, IMA J. Numer. Anal., 27 (2007), pp. 576–592.

[13] N. Kopteva, *Maximum norm error analysis of a 2d singularly perturbed semilinear reaction-diffusion problem*, Math. Comp., 76 (2007), pp. 631–646.

[14] N. Kopteva and M. Stynes, *A robust adaptive method for a quasi-linear one-dimensional convection-diffusion problem*, SIAM J. Numer. Anal., 39 (2001), pp. 1446–1467.

[15] G. Kunert and R. Verfürth, *Edge residuals dominate a posteriori error estimates for linear finite element methods on anisotropic triangular and tetrahedral meshes*, Numer. Math., 86 (2000), pp. 283–303.

[16] G. Kunert, *Robust a posteriori error estimation for a singularly perturbed reaction-diffusion equation on anisotropic tetrahedral meshes. A posteriori error estimation and adaptive computational methods*, Adv. Comput. Math., 15 (2001), pp. 237–259.

[17] O. A. Ladyzhenskaya and N. N. Ural'tseva, *Linear and Quasilinear Elliptic Equations*, Academic Press, New York, 1968.

[18] T. Linß, *Maximum-norm error analysis of a non-monotone FEM for a singularly perturbed reaction-diffusion problem*, BIT, 47 (2007), pp. 379–391.

[19] T. Linß and M. Stynes, *The SDFEM on Shishkin meshes for linear convection-diffusion problems*, Numer. Math., 87 (2001), pp. 457–484.

[20] J. J. H. Miller, E. O'Riordan, and G. I. Shishkin, *Solution of Singularly Perturbed Problems with ε-uniform Numerical Methods—Introduction to the Theory of Linear Problems in One and Two Dimensions*, World Scientific, Singapore, 1996.

[21] R. H. Nochetto, *Pointwise a posteriori error estimates for elliptic problems on highly graded meshes*, Math. Comp., 64 (1995), pp. 1–22.

[22] A. H. Schatz and L. B. Wahlbin, *On the finite element method for singularly perturbed reaction-diffusion problems in two and one dimensions*, Math. Comp., 40 (1983), pp. 47–88.

SIAM J. NUMER. ANAL.
Vol. 46, No. 3, pp. 1619–1639
© 2008 Society for Industrial and Applied Mathematics

WEIGHTED-NORM FIRST-ORDER SYSTEM LEAST-SQUARES (FOSLS) FOR DIV/CURL SYSTEMS WITH THREE DIMENSIONAL EDGE SINGULARITIES*

E. LEE†, T. A. MANTEUFFEL‡, AND C. R. WESTPHAL§

Abstract. A weighted-norm first-order system least-squares (FOSLS) method for div/curl problems with edge singularities is presented. Traditional finite element methods, including least-squares methods, often suffer from a global loss of accuracy due to the influence of a nonsmooth solution near polyhedral edges. By minimizing a modified least-squares functional, optimal accuracy in weighted and nonweighted norms is recovered. Error estimates with and without mesh refinements are presented, and numerical results are given to confirm the theory.

Key words. singularities, weighted Sobolev spaces, least-squares, finite element methods

AMS subject classification. 65N30

DOI. 10.1137/06067345X

1. Introduction. In this paper, a weighted-norm first-order system least-squares (FOSLS) method for problems with edge singularities is studied. The least-squares finite element method is well suited for many problems with first-order differential operators. Such systems often arise naturally from physical laws or can be formulated from a higher-order system (cf. [18]). The numerical solution is found by minimizing the norm of the residual of the system over an appropriate finite element space. Thus, to approximate a solution to the first-order system, the goal of a least-squares method is largely to choose the correct norm and finite element space. Under sufficient smoothness assumptions, many least-squares functionals induce a norm that can be shown to be equivalent to the product H^1-norm, indicating that H^1-conforming finite elements and multigrid can be used to solve the equations efficiently (cf. [9], [10]). However, the presence of boundary singularities may cause a loss of H^1-equivalence, and the use of H^1-finite element spaces may cause a loss of global accuracy in the approximate solution.

We focus on problems posed in polyhedral domains where the boundary contains edges with an inner angle larger than π or edges upon which different types of boundary conditions meet with an inner angle larger than $\pi/2$. Such problems generally require special consideration and have been the focus of study in both Galerkin and least-squares methods. One of the most common approaches is to use $H(div)$- or $H(curl)$-conforming finite elements, for example, Raviart–Thomas or Nédélec edge elements (see [29]). In papers [15] and [7], finite element spaces having a grid decomposition property and $H(div)$-conforming finite element spaces to get optimal

*Received by the editors October 27, 2006; accepted for publication (in revised form) August 30, 2007; published electronically April 11, 2008. This work was supported by the National Science Foundation under grant DMS-0420873 and the Department of Energy under grants DE-FC02-01ER25479 and DE-FG02-03ER25574.

http://www.siam.org/journals/sinum/46-3/67345.html

†School of Computer Science, Florida State University, Tallahassee, FL 32306-4120 (eunjung@scs.fsu.edu).

‡Department of Applied Mathematics, Campus Box 526, University of Colorado at Boulder, Boulder, CO 80309-0526 (tmanteuf@colorado.edu).

§Department of Mathematics and Computer Science, Wabash College, P.O. Box 352, Crawfordsville, IN 47933 (westphac@wabash.edu).

convergence were investigated. In the classical paper [14], singular functions were added to standard finite element spaces, admitting optimal approximations at minimal additional cost. This approach in the context of a FOSLS formulation was explained in [6] and [5]. A weighted regularization in which weight functions were used in the divergence integral was presented in [13]. In [8], the H^{-1}-norm least-squares approach in discrete space was introduced in a weak variational formulation. In [24] and [22], two different types of modified first-order system LL* methods were developed that allow an accurate approximation using H^1-conforming finite elements for equations having singular boundaries in two and three dimensions, respectively. The latter of these employed partially weighted functionals.

In [23], two dimensional div/curl problems with boundary singularities are solved by minimizing a functional weighted according to the distance from the singular point. Here, we extend our earlier work to three dimensional problems with edge singularities and further consider local mesh refinement. Since many properties do not follow the two dimensional ones, this extension requires the investigation of the three dimensional singular solutions in the proper weighted Sobolev spaces and the establishment of certain new Poincaré-type bounds in scalar and vector forms. The goal of this paper is retaining the global L^2-accuracy and achieving the optimal order discretization accuracy in the induced weighted Sobolev spaces. In section 2, we introduce notation, definitions, and our model problem. Some regularity results in weighted Sobolev spaces are described in section 3. In section 4, we investigate Poincaré inequalities in weighted Sobolev spaces. We prove optimal L^2- and H^1-error convergence away from the singularities and the optimal L^2-error convergence globally by error estimates in weighted norms and L^2-norms in section 5. Establishing error estimates in weighted L^2- and H^1-norms with graded mesh refinement are presented in section 6. In section 7, we report several numerical examples. The theory applies for a range of values for the power of the weight function. Our numerical results show remarkable agreement with theory within this range and somewhat outside this range as well, indicating the possibility that the theory can be extended.

2. Weighted-norm least squares. Let Q be a polygon in $\mathbb{R}^2$ and $I \subset \mathbb{R}$ be a bounded interval. Then, consider the prototype domain

$$\Omega := Q \times I \subset \mathbb{R}^3, \tag{2.1}$$

which is a polyhedral cylinder. In this paper, we restrict ourselves to the case where the domain has one singular edge, which means

- the polygon Q has a corner with inner angle in $(\pi/2, \pi]$, or
- the polygon Q has a corner with inner angle $> \pi$, or
- different types of boundary condition meet at the edge with inner angle $> \pi/2$;

however, the general case follows by considering our approach as a local result. In this paper, we focus on resolving the difficulties from the second and the third case above. The first case gives more regular solution than the second or the third case, but it can still reduce the rate of convergence, and the treatment will be similar to the other two cases. We refer to the edge that causes the boundary singularity as E and the inner angle of the edge E as ω. From now on, the value λ indicates π/ω. We assume that $E = (x_0, y_0) \times I \subset \partial\Omega$. By translation and rotation, we may suppose that E coincides with the z-axis. Without loss of generality, we further assume that $diam(\Omega) \leq 1$.

We denote $\langle \cdot, \cdot \rangle$ the L^2-inner product defined by $\langle u, v \rangle = \int_\Omega u\ v\ d\Omega$, for $u, v \in L^2(\Omega)$, and denote $\| \cdot \|$ the L^2-norm defined by $||u|| = \langle u, u \rangle^{\frac{1}{2}}$. Let $H^k(\Omega)$ be the

standard Sobolev space with the norm $\|\cdot\|_k$ and the seminorm $|\cdot|_k$. Denote by $H^k_\beta(\Omega)$ the weighted Sobolev space of functions u such that

$$\|u\|^2_{k,\beta} = \sum_{|m|=0}^{k} \int_\Omega r^{2(\beta+|m|-k)} \left|D^m u\right|^2 d\Omega < \infty,$$

where $r := r(\mathbf{x})$ is the distance of $\mathbf{x} \in \Omega$ from the singular edge, E, and $|D^m u|^2 = \sum_{a_1+a_2+a_3=m} |\partial_x^{a_1}\partial_y^{a_2}\partial_z^{a_3} u|^2$. The corresponding seminorm is $|\cdot|_{k,\beta} := \|D^k \cdot\|_{0,\beta}$. Characters in bold represent vector functions, and components are subscripted by integers. For example, $\mathbf{u} = (u_1, u_2, u_3)^t$, and the vector L^2-norm is given by $\|\mathbf{u}\| = (\|u_1\|^2 + \|u_2\|^2 + \|u_3\|^2)^{\frac{1}{2}}$. Define Hilbert spaces $H(\nabla\cdot)$ and $H(\nabla\times)$ as the spaces of L^2 vector functions, $\mathbf{u}$, satisfying $\nabla \cdot \mathbf{u} \in L^2(\Omega)$ and $\nabla \times \mathbf{u} \in L^2(\Omega)^3$, respectively.

For convenience, we omit the superscript t for vector transpose. Throughout this paper, c is a generic constant that is used to denote various constants, and its dependence on other quantities will be indicated if necessary.

Consider the div/curl system

$$\begin{aligned} \nabla \times \mathbf{u} &= \mathbf{f} \quad && \text{in} \quad \Omega, \\ \nabla \cdot \mathbf{u} &= g \quad && \text{in} \quad \Omega, \\ \mathbf{n} \times \mathbf{u} &= \mathbf{0} \quad && \text{on} \quad \Gamma_D, \\ \mathbf{n} \cdot \mathbf{u} &= 0 \quad && \text{on} \quad \Gamma_N, \end{aligned} \tag{2.2}$$

where $\mathbf{n}$ is the outward unit normal and Γ_D, Γ_N are a finite number of connected pieces satisfying $\Gamma_D \cap \Gamma_N = \emptyset$, $\Gamma_D \cup \Gamma_N = \partial\Omega$. The above system can be rewritten as

$$L\mathbf{u} = \begin{bmatrix} \nabla\times \\ \nabla\cdot \end{bmatrix} \mathbf{u} = \begin{bmatrix} \mathbf{f} \\ g \end{bmatrix} = \mathbf{F}, \tag{2.3}$$

where L is a linear operator from $\mathcal{U}$ to $L^2(\Omega)^4$ with

$$\mathcal{U} = \{\mathbf{v} \in H(\nabla\cdot) \cap H(\nabla\times) : \mathbf{n} \times \mathbf{v} = 0 \text{ on } \Gamma_D,\ \mathbf{n} \cdot \mathbf{v} = 0 \text{ on } \Gamma_N\}.$$

This type of div/curl system appears in many applications, including electromagnetics, porous media flow, and solid and fluid mechanics. Thus, solving $L\mathbf{u} = \mathbf{F}$ in nonsmooth domains is of wide interest. The traditional least-squares method solves system (2.3) by minimizing the residual functional

$$G(\mathbf{v}; \mathbf{F}) = \|L\mathbf{v} - \mathbf{F}\|^2 = \|\nabla \times \mathbf{v} - \mathbf{f}\|^2 + \|\nabla \cdot \mathbf{v} - g\|^2$$

over $\mathcal{U}$ in the weak sense: find $\mathbf{u} \in \mathcal{U}$ satisfying

$$\langle L\mathbf{u}, L\mathbf{v}\rangle = \langle \mathbf{F}, L\mathbf{v}\rangle \qquad \text{for all } \mathbf{v} \in \mathcal{U}.$$

The FOSLS method includes the design of functionals whose bilinear part is equivalent to an appropriate norm, often the H^1-norm when possible. As briefly mentioned in the introduction, if the domain is not convex and the boundary is not $\mathcal{C}^{1,1}$ or different types of boundary conditions meet at an edge with an inner angle larger than $\pi/2$, then the space $\mathcal{U}$ is not continuously embedded into $H^1(\Omega)^3$; therefore, the solution may not be in $H^1(\Omega)^3$. This would seem to preclude the use of H^1-conforming finite element spaces even though the solution is smooth away from the edge. To overcome this difficulty, we introduce the following weighted-norm functional:

$$G_w(\mathbf{u}; \mathbf{F}) = \|w(L\mathbf{u} - \mathbf{F})\|^2 = \|w\nabla \times (\mathbf{u} - \mathbf{f})\|^2 + \|w(\nabla \cdot \mathbf{u} - g)\|^2, \tag{2.4}$$

where the weight function has the form $w = r^\beta$ for some $\beta > 0$, and we define

$$\|L\mathbf{u}\|_{0,\beta}^2 = \|r^\beta \nabla \times \mathbf{u}\|^2 + \|r^\beta \nabla \cdot \mathbf{u}\|^2.$$

Of course, one can restrict the weight function to some small region by use of scaling or a smooth cut-off function; that is, one can define a weight function local to each of the singular edges of an arbitrary polyhedral domain. It is reasonable to provide a local treatment since while the pollution effect produces global errors, its cause is a local loss of smoothness. In this presentation, we assumed $diam(\Omega) \leq 1$ and use the weight function as above for convenience of presentation.

3. Coercivity of L in weighted Sobolev spaces. In this section, we show that the operator L defined in (2.3) is coercive in certain weighted Sobolev spaces, leading to the norm equivalence between $\|L\mathbf{u}\|_{0,\beta}$ and $\|\nabla \mathbf{u}\|_{0,\beta}$.

LEMMA 3.1. *Let the operator L be defined as in* (2.3) *and* $(\mathbf{f}, g) \in H_\beta^0(\Omega)^4$.

(i) *If* $\mathbf{f} = \mathbf{0}$, *then* $L : H_\beta^1(\Omega)^3 \cap \mathcal{U} \to H_\beta^0(\Omega)^4$ *is injective for*

$$\beta \in \begin{cases} (1-\lambda, 1+\lambda) & \text{when } \partial\Omega = \Gamma_D, \\ (1-\lambda, 1) & \text{when } \partial\Omega = \Gamma_N, \\ (1-\lambda/2, 1+\lambda/2) & \text{when } \Gamma_D \neq \emptyset \text{ and } \Gamma_N \neq \emptyset. \end{cases} \tag{3.1}$$

(ii) *If* $g = 0$, *then* $L : H_\beta^1(\Omega)^3 \cap \mathcal{U} \to H_\beta^0(\Omega)^4$ *is injective for*

$$\beta \in \begin{cases} (1-\lambda, 1+\lambda) & \text{when } \partial\Omega = \Gamma_D, \\ (1-\lambda, 1) & \text{when } \partial\Omega = \Gamma_N \text{ and the sides of the domain} \\ & \quad \text{are parallel to the coordinate axes}, \\ (1-\lambda/2, 1+\lambda/2) & \text{when } \Gamma_D \neq \emptyset \text{ and } \Gamma_N \neq \emptyset \text{ and the sides of the} \\ & \quad \text{domain are parallel to the coordinate axes.} \end{cases} \tag{3.2}$$

(iii) *If* $\mathbf{f} \neq \mathbf{0}$ *and* $g \neq 0$, *then* $L : H_\beta^1(\Omega)^3 \cap \mathcal{U} \to H_\beta^0(\Omega)^4$ *is injective for*

$$\beta \in \begin{cases} (1-\lambda, 1+\lambda) & \text{when } \partial\Omega = \Gamma_D, \\ (1-\lambda, 1) & \text{when } \partial\Omega = \Gamma_N \text{ and the sides of the domain} \\ & \quad \text{are parallel to the coordinate axes.} \end{cases} \tag{3.3}$$

Proof. The result is proved by establishing a decomposition of $\mathbf{u} \in H_\beta^1(\Omega)^3 \cap \mathcal{U}$ and then demonstrating that components of the decomposition each satisfy a Poisson equation, for which known results apply (cf. [25], [26], [28]). The two dimensional orthogonal decomposition of $\mathbf{u}$ with boundary conditions in (2.2) naturally provides Poisson equations with Dirichlet or Neumann boundary conditions [23]. But, it is not true that each component of the orthogonal decomposition of $\mathbf{u}$ can be easily induced to Dirichlet or Neumann Poisson equations in three dimensional space. So, we consider things in several cases as presented in Lemma 3.1.

(i) If $\mathbf{f} = \mathbf{0}$, then there exists a unique $p \in H^1(\Omega)$ satisfying $\mathbf{u} = \nabla p$ and

$$\begin{cases} \Delta p \;=\; \nabla \cdot \mathbf{u} \;\text{ in } \Omega, \\ p \;=\; c \;\text{ on } \Gamma_D, \quad \mathbf{n} \cdot \nabla p \;=\; 0 \;\text{ on } \Gamma_N, \end{cases} \tag{3.4}$$

where c is any constant. The above Poisson problem (3.4) is an isomorphism from $H_\beta^2(\Omega)$ to $H_\beta^0(\Omega)$ for $\beta \in (1-\lambda, 1+\lambda)$ if $\partial\Omega = \Gamma_D$ [26], for $\beta \in (1-\lambda, 1)$ if $\partial\Omega = \Gamma_N$ [26], and for $\beta \in (1-\lambda/2, 1+\lambda/2)$ if $\Gamma_D \neq \emptyset$ and $\Gamma_N \neq \emptyset$ [25], [28]. Also, there exists a constant C satisfying

$$\|p\|_{2,\beta} \;\leq\; C\|\nabla \cdot \mathbf{u}\|_{0,\beta} \tag{3.5}$$

under the above assumptions on β and boundary conditions. The inequality (3.5) implies that $\|\mathbf{u}\|_{1,\beta} \le c\|L\mathbf{u}\|_{0,\beta}$; therefore, L is injective.

(ii) If $g = 0$, then there exists a vector potential $\phi = (\phi_1, \phi_2, \phi_3) \in H(\nabla\times)$ such that $\mathbf{u} = \nabla \times \phi$ and $\nabla \cdot \phi = 0$. Suppose $\partial\Omega = \Gamma_D$; then $\nabla \times \mathbf{u} = \nabla \times \nabla \times \phi = -\Delta\phi + \nabla\nabla \cdot \phi = -\Delta\phi$ and the boundary conditions $\mathbf{n} \times \phi = \mathbf{0}$, $\mathbf{n} \times \nabla \times \phi = \mathbf{0}$ yield, for $i = 1, 2$,

$$\begin{cases} -\Delta\phi_i = (\nabla \times \mathbf{u})_i & \text{in } \Omega, \\ \phi_i = 0 & \text{on } \partial\Omega \end{cases} \quad \text{and} \quad \begin{cases} -\Delta\phi_3 = (\nabla \times \mathbf{u})_3 & \text{in } \Omega, \\ \phi_3 = 0 & \text{on the side walls,} \\ \mathbf{n} \cdot \nabla\phi_3 = 0 & \text{on top and bottom,} \end{cases}$$

where $(\nabla \times \mathbf{u})_i$ is the ith component of $\nabla \times \mathbf{u}$. Since the inner angle between top and side walls is $\pi/2$, there are no singularities around those edges. Therefore, L is injective for $\beta \in (1 - \lambda, 1 + \lambda)$. Analogously, L is injective for $\beta \in (1 - \lambda, 1)$ when $\partial\Omega = \Gamma_N$ and the sides of the domain are parallel to the coordinate axes, in which $\mathbf{n}$ is of the form $(n_1, 0, 0), (0, n_2, 0)$, and $(0, 0, n_3)$. For the case when $\Gamma_D \neq \emptyset$ and $\Gamma_N \neq \emptyset$, we use the following decomposition in [1]: a function $\mathbf{u}$ satisfies $\nabla \cdot \mathbf{u} = 0$ and $\mathbf{n} \cdot \mathbf{u} = 0$ on Γ_N if and only if there exists a function $\phi \in H(\nabla\times)$ such that

$$\begin{cases} \mathbf{u} = \nabla \times \phi, \quad \nabla \cdot \phi = 0 \ \text{ in } \Omega, \\ \mathbf{n} \times \phi = \mathbf{0} \text{ on } \Gamma_N, \quad \text{and} \quad \mathbf{n} \cdot \phi = 0 \text{ on } \Gamma_D. \end{cases} \tag{3.6}$$

If the sides of the domain are parallel to the coordinate axes, then we can derive a Poisson equation with mixed boundary conditions from (3.6). Hence, L is injective for $\beta \in (1 - \lambda/2, 1 + \lambda/2)$.

(iii) Let $\mathbf{u} \in H^1_\beta(\Omega)^3 \cap \mathcal{U}$; then $\mathbf{u}$ has the orthogonal decomposition [16]

$$\mathbf{u} = \nabla \times \phi + \nabla p,$$

where $\phi \in H(\nabla\times)$ with $\nabla \cdot \phi = 0$ and $\mathbf{n} \times \phi = \mathbf{0}$, and $p \in H^1_0(\Omega)$ is the solution of

$$\langle \nabla p, \nabla\xi \rangle = \langle \mathbf{u}, \nabla\xi \rangle \quad \text{for all } \xi \in H^1_0(\Omega).$$

Similarly to (i) and (ii), one can establish the results. ∎

Remark 3.2. Note that the above restrictions on β are sufficient but may not be necessary. In the remainder of the paper, we establish results that depend on L being injective. While the above theorem establishes sufficient conditions, the operator L may be injective for a much wider range of values for the weight β. Extension of the range is beyond the scope of this paper, but we remark that numerical results seem to indicate that it may be possible.

If the div/curl system is derived from the Poisson equation, then the problem easily matches to the first case in Lemma 3.1.

LEMMA 3.3. *Let $\mathbf{u} \in H^1_\beta(\Omega)^3 \cap \mathcal{U}$; then there exists a constant c satisfying*

$$\|\mathbf{u}\|_{1,\beta} \le c \left(\|L\mathbf{u}\|_{0,\beta} + \|\mathbf{u}\|_{0,\beta-1} \right) \tag{3.7}$$

for any $\beta > 0$.

Proof. Since the norms $\|u\|_{1,\beta}$ and $\|r^\beta u\|_1$ are equivalent (see [25]) and $\|r^\beta u\| \le \|u\|_{0,\beta-1}$, it is enough to show that $\|\nabla(r^\beta \mathbf{u})\|$ is less than the right-hand side of (3.7). Since $r^\beta \mathbf{u} \in H^1(\Omega)^3$, the result in [12] and the triangle inequality yield $\|\nabla(r^\beta \mathbf{u})\| \le c\, \|L(r^\beta \mathbf{u})\| \le c\, (\|L\mathbf{u}\|_{0,\beta} + \|\mathbf{u}\|_{0,\beta-1})$. ∎

Based on Lemmas 3.1 and 3.3, we achieve the goal of this section in the following theorem by using a modified compactness argument.

THEOREM 3.4. *Let $L : H^1_\beta(\Omega)^3 \cap \mathcal{U} \to H^0_\beta(\Omega)^4$ be injective and $\mathbf{u} \in H^1_\beta(\Omega)^3 \cap \mathcal{U}$. Then, there exists a constant $c = c(\Omega, \beta)$ such that*

$$\|\mathbf{u}\|_{1,\beta} \leq c \, \|L\mathbf{u}\|_{0,\beta}. \tag{3.8}$$

Proof. First, we let

$$\Omega_R = \{\mathbf{x} = (r, \theta, z) \in \Omega \ : \ r > R\}$$

and define a norm $\|\mathbf{u}\|^2_{1,\beta,\Omega_R} = |\mathbf{u}|^2_{1,\beta,\Omega_R} + \|\mathbf{u}\|^2_{0,\beta-1,\Omega_R}$ for any $R \geq 0$. Define Ω_{2R} and a norm $\|\mathbf{u}\|_{1,\beta,\Omega_{2R}}$ in the same way by replacing R with $2R$. Let $\delta(r)$ be a smooth cut-off function,

$$\delta(r) = \begin{cases} 0 & \text{if} \quad r \leq R, \\ 1 & \text{if} \quad r \geq 2R, \end{cases}$$

where $|\delta| \leq 1$, $|\delta'| \leq cR^{-1}$, and δ' has support only for $r \in (R, 2R)$. By the fact that, for any $\mathbf{u} \in H^1_\beta(\Omega)^3$, we have $\delta r^\beta \mathbf{u} \in H^1(\Omega)^3$ and the triangle inequality, we have the first (cf. [12]) and the second inequalities, respectively, of the following:

$$\begin{aligned} \|\nabla(\delta r^\beta \mathbf{u})\| &\leq c\|L(\delta r^\beta \mathbf{u})\| \leq c(\|\delta r^\beta L\mathbf{u}\| + \|\nabla(\delta r^\beta) \cdot \mathbf{u}\|) \\ &\leq c(\|L\mathbf{u}\|_{0,\beta,\Omega_R} + c\|\mathbf{u}\|_{0,\beta-1,\Omega_R}), \end{aligned} \tag{3.9}$$

where c is independent of R. We also have

$$\begin{aligned} \|\nabla \mathbf{u}\|_{0,\beta,\Omega_{2R}} &\leq \|\delta r^\beta \nabla \mathbf{u}\| \leq \|\delta r^\beta \nabla \mathbf{u} + \nabla(\delta r^\beta) \cdot \mathbf{u}\| + \|\nabla(\delta r^\beta) \cdot \mathbf{u})\| \\ &\leq \|\nabla(\delta r^\beta \mathbf{u})\| + c\|\mathbf{u}\|_{0,\beta-1,\Omega_R}, \end{aligned} \tag{3.10}$$

where c is independent of R. Putting (3.9) and (3.10) together we have

$$\|\nabla \mathbf{u}\|_{0,\beta,\Omega_{2R}} \leq c(\|L\mathbf{u}\|_{0,\beta,\Omega_R} + \|\mathbf{u}\|_{0,\beta-1,\Omega_R}). \tag{3.11}$$

Here, we use a modified compactness argument: Assume that (3.8) is not true; that is, there is a sequence $\{\mathbf{u}_j\}_{j=1,\infty} \subset H^1_\beta(\Omega)^3 \cap \mathcal{U}$ such that

$$\|\mathbf{u}_j\|_{1,\beta} = 1 \quad \text{and} \quad \|L\mathbf{u}_j\|_{0,\beta} \to 0 \ \text{ as } n \to \infty. \tag{3.12}$$

Since $H^1_\beta(\Omega)$ is compactly embedded into $H^0_\beta(\Omega)$ [25], there is a subsequence $\{\mathbf{u}_{j_\ell}\}_{\ell=1,\infty}$, which we shall denote by $\{\mathbf{u}_\ell\}_{\ell=1,\infty}$, that is Cauchy in $H^0_\beta(\Omega)^3$ and, thus, has limit $\tilde{\mathbf{u}} \in H^0_\beta(\Omega)^3$. That is, $\|\mathbf{u}_\ell - \tilde{\mathbf{u}}\|_{0,\beta} \to 0$. Now, for any $R > 0$, the norms $\|\cdot\|_{0,\beta,\Omega_R}$ and $\|\cdot\|_{0,\beta-1,\Omega_R}$ are equivalent. This implies that $\{\mathbf{u}_\ell\}_{\ell=1,\infty}$ is Cauchy in $H^0_{\beta-1}(\Omega_R)$ and

$$\|\mathbf{u}_\ell - \tilde{\mathbf{u}}\|_{0,\beta-1,\Omega_R} \to 0.$$

Using (3.11) we have

$$|\mathbf{u}_\ell - \mathbf{u}_m|_{1,\beta,\Omega_{2R}} \leq c\left(\|L(\mathbf{u}_\ell - \mathbf{u}_m)\|_{0,\beta,\Omega_R} + \|\mathbf{u}_\ell - \mathbf{u}_m\|_{0,\beta-1,\Omega_R}\right), \tag{3.13}$$

which implies that $\{\mathbf{u}_\ell\}_{\ell=1,\infty}$ is Cauchy in $H^1_\beta(\Omega_{2R})^3$ and, thus, $\tilde{\mathbf{u}} \in H^1_\beta(\Omega_{2R})^3$ and

$$\|\mathbf{u}_\ell - \tilde{\mathbf{u}}\|_{1,\beta,\Omega_{2R}} \to 0. \tag{3.14}$$

The triangle inequality yields

$$\begin{aligned}\|L\tilde{\mathbf{u}}\|_{0,\beta,\Omega_{2R}} &\le c\|L(\tilde{\mathbf{u}}-\mathbf{u}_\ell)\|_{0,\beta,\Omega_{2R}} + c\|L\mathbf{u}_\ell\|_{0,\beta,\Omega_{2R}}\\ &\le c\|(\tilde{\mathbf{u}}-\mathbf{u}_\ell)\|_{1,\beta,\Omega_{2R}} + c\|L\mathbf{u}_\ell\|_{0,\beta,\Omega_{2R}}\end{aligned}$$

for every ℓ, which implies that $L\tilde{\mathbf{u}}=0$ on Ω_{2R}. Since this is true for every $R>0$, we have

$$L\tilde{\mathbf{u}}=0 \qquad \text{on} \quad \Omega. \tag{3.15}$$

To summarize what we now know, $\tilde{\mathbf{u}}\in H^0_\beta(\Omega)^3$, $\tilde{\mathbf{u}}\in H^1_\beta(\Omega_R)^3$ for every $R>0$, and $L\tilde{\mathbf{u}}=0$ on Ω. If $\tilde{\mathbf{u}}\in H^0_{\beta-1}(\Omega)^3$, using Lemma 3.3 and the fact that $L\tilde{\mathbf{u}}=0$, we have $\tilde{\mathbf{u}}\in H^1_\beta(\Omega)^3$, which would contradict the assumption that L is injective on $H^1_\beta(\Omega)^3\cap\mathcal{U}$. So, we assume that $\tilde{\mathbf{u}}\notin H^0_{\beta-1}(\Omega)^3$, that is,

$$\lim_{R\to 0}\|\tilde{\mathbf{u}}\|_{0,\beta-1,\Omega_R}=\infty.$$

However,

$$\begin{aligned}\|\tilde{\mathbf{u}}\|_{0,\beta-1,\Omega_R} &\le \|\tilde{\mathbf{u}}-\mathbf{u}_\ell\|_{0,\beta-1,\Omega_R} + \|\mathbf{u}_\ell\|_{0,\beta-1,\Omega_R} \le \|\tilde{\mathbf{u}}-\mathbf{u}_\ell\|_{0,\beta-1,\Omega_R} + \|\mathbf{u}_\ell\|_{0,\beta-1,\Omega}\\ &\le \|\tilde{\mathbf{u}}-\mathbf{u}_\ell\|_{0,\beta-1,\Omega_R} + \|\mathbf{u}_\ell\|_{1,\beta,\Omega} \le \|\tilde{\mathbf{u}}-\mathbf{u}_\ell\|_{0,\beta-1,\Omega_R}+1\end{aligned}$$

for every ℓ independent of R. Thus, $\|\tilde{\mathbf{u}}\|_{0,\beta-1,\Omega_R}\le 1$ for all $R>0$. This is a contradiction, and the result is proved. ∎

The above theorem implies that the corresponding weak formulation of the least-squares method is coercive, whenever L is injective, in the proper weighted Sobolev space. The continuity is obtained easily by the triangle inequality. Since the solution is in $H^1_\beta(\Omega)$, that is, the weighted L^2-norm of gradient of the solution is finite, it is reasonable to minimize the residual functional in the weighted norm to approximately solve the problem when using H^1-conforming finite elements.

4. Poincaré inequalities. In this section, we develop several types of Poincaré inequalities, which are useful in various settings. Recall that E is the singular edge and $diam(\Omega)<1$. From now on, we consider the domain in the cylindrical coordinate system for convenience. The domain, Ω, can be rewritten as

$$\Omega=\{\mathbf{x}=(r,\theta,z): 0<r<R(\theta),\ 0<\theta<\omega<2\pi,\ a<z<b\},$$

where $R(\theta)$ is the distance from the singular edge, E, to boundary points depending on the angle θ. Then, $\partial\Omega=\{(r,\theta,z): r=R(\theta), 0\le\theta\le\omega\}\cup\{(r,\theta,z):\theta=0, 0\le r\le R(0)\}\cup\{(r,\theta,z):\theta=\omega, 0\le r\le R(\omega)\}\cup\{(r,\theta,z): z=a \text{ or } z=b\}$. The following lemma is known from [20].

LEMMA 4.1. *If $q\in H^1_\beta(\Omega)$ vanishes on $\partial\Omega$, then, for any β,*

$$\|q\|_{0,\beta-1} \ \le\ c\ \|\nabla q\|_{0,\beta}.$$

Now, we show several generalized Poincaré-type inequalities.

LEMMA 4.2. *For $p\in H^1_\beta(\Omega)$, there exists a constant c such that*

$$\|p\|_{0,\beta-1}\ \le\ c\ (\|p\|_{0,\beta}+\|\nabla p\|_{0,\beta})$$

when $\beta>0$.

Proof. Let $R_0 = (\min_{0<\theta<\omega} R(\theta))/4$ and let χ be a smooth function defined in Ω such that $\chi(r) = 1$ when $r < R_0$ and $\chi(r) = 0$ when $r > 2R_0$ and $|\chi'| \le cR_0^{-1}$ for some constant c. Since $1 = \chi + 1 - \chi$,

$$\int_0^{R(\theta)} r^{2\beta-2}|p|^2 r dr = \int_0^{R(\theta)} r^{2\beta-2}\left|\chi p + (1-\chi)p\right|^2 r dr \le 2\int_0^{R(\theta)} r^{2\beta-2}\left(|\chi p|^2 + |(1-\chi)p|^2\right) r dr.$$

By the modified Hardy inequality in [19], for $\beta > 0$,

$$\int_0^{R(\theta)} r^{2\beta-2}|\chi p|^2 r dr \le c\int_0^{R(\theta)} r^{2\beta}\left|\frac{\partial(\chi p)}{\partial r}\right|^2 r dr \le c\int_0^{2R_0} r^{2\beta}\left(\frac{1}{R_0^2}|p|^2 + \left|\frac{\partial p}{\partial r}\right|^2\right) r dr.$$

Since $(1-\chi)p$ has nonzero values only on $(R_0, R(\theta))$,

$$\begin{aligned}\int_0^{R(\theta)} r^{2\beta-2}|(1-\chi)p|^2 r dr &= \int_{R_0}^{R(\theta)} r^{2\beta-2}|(1-\chi)p|^2 r dr \\ &\le {R_0}^{-2}\int_{R_0}^{R(\theta)} r^{2\beta}|(1-\chi)p|^2 r dr \le {R_0}^{-2}\int_0^{R(\theta)} r^{2\beta}|p|^2 r dr.\end{aligned}$$

Hence, by Fubini's theorem,

$$\int_\Omega r^{2\beta-2}|p|^2 d\Omega \le c\ R_0^{-2}\int_\Omega r^{2\beta}|p|^2 d\Omega + c\int_\Omega r^{2\beta}|\nabla p|^2 d\Omega. \qquad \square$$

The above lemma is useful in establishing weighted-norm Poincaré inequalities since they are guaranteed if we can show that

$$\|p\|_{0,\beta} \le c\|\nabla p\|_{0,\beta}. \tag{4.1}$$

Combining Lemma 4.2 with the following lemma provides a Poincaré inequality when there are no given boundary conditions.

LEMMA 4.3. *Let $\epsilon > 0$, $\beta > -1$, and $\beta + 1 - \epsilon > 0$. If $p \in H^1_{\beta+1-\epsilon}(\Omega)$, then there exist constants b and $c = c(\Omega, \beta, \epsilon)$ such that*

$$\|p - b\|_{0,\beta} \ \le\ c\ \|\nabla p\|_{0,\beta+1-\epsilon}.$$

Proof. Here, we show an outline of the proof. The details can be found in [21]. If $p \in H^1_{\beta+1-\epsilon}(\Omega)$ when $\beta + 1 - \epsilon > 0$, then $p \in H^1(S)$, where $S \subset \Omega$ and $\bar{S} \cap E = \emptyset$. Then, we can consider the following expression for p: for $(r, \theta, z), (r_0, \theta_0, z_0) \in \Omega$,

$$\begin{aligned}&p(r,\theta,z) - p(r_0,\theta_0,z_0)\\ &= p(r,\theta,z) - p(r,\theta_0,z) + p(r,\theta_0,z) - p(r_0,\theta_0,z) + p(r_0,\theta_0,z) - p(r_0,\theta_0,z_0)\\ &= \int_{\theta_0}^{\theta} \frac{\partial p}{\partial\tilde{\theta}}(r,\tilde{\theta},z)\ d\tilde{\theta}\ +\ \int_{r_0}^{r} \frac{\partial p}{\partial\tilde{r}}(\tilde{r},\theta_0,z)\ d\tilde{r}\ +\ \int_{z_0}^{z} \frac{\partial p}{\partial\tilde{z}}(r_0,\theta_0,\tilde{z})\ d\tilde{z}.\end{aligned}$$

Multiply by $r_0^{\beta+\frac{1}{2}}$ and perform the integration $\int_\Omega r_0 dr_0 d\theta_0 dz_0$ on both sides:

$$\begin{aligned}c_1 p(r,\theta,z) = &\int_\Omega r_0^{\beta+\frac{1}{2}} p(r_0,\theta_0,z_0) r_0 dr_0 d\theta_0 dz_0 + \int_\Omega r_0^{\beta+\frac{1}{2}}\left\{\int_{t_0}^{\theta} \frac{\partial p}{\partial\tilde{\theta}}(r,\tilde{\theta},z)\ d\tilde{\theta}\right.\\ &\left. + \int_{r_0}^{r} \frac{\partial p}{\partial\tilde{r}}(\tilde{r},\theta_0,z)\ d\tilde{r}\ +\ \int_{z_0}^{z} \frac{\partial p}{\partial\tilde{z}}(r_0,\theta_0,\tilde{z})\ d\tilde{z}\right\} r_0 dr_0 d\theta_0 dz_0,\end{aligned} \tag{4.2}$$

where $c_1 = \int_\Omega r_0^{\beta+\frac{1}{2}} r_0 dr_0 d\theta_0 dz_0$. Let

$$b = \frac{1}{c_1} \int_\Omega r_0^{\beta+\frac{1}{2}} p(r_0, \theta_0, z_0)\ r_0 dr_0 d\theta_0 dz_0;$$

then $|b| \le c\|p\| < \infty$. Subtracting b from both sides in (4.2), applying Fubini's theorem, inserting $\tilde{r}^{\frac{-1+2\epsilon}{2}} \cdot \tilde{r}^{\frac{1-2\epsilon}{2}} = 1$ into group $\tilde{r}^{\frac{1-2\epsilon}{2}}$ with the $\frac{\partial p}{\partial \tilde{r}}$ term, using the Cauchy–Schwarz inequality, and squaring both sides yields

$$|p(r,\theta,z) - b|^2 \le c \left\{ \int_0^\omega \left| \frac{\partial p}{\partial \tilde{\theta}}(r, \tilde{\theta}, z) \right|^2 d\tilde{\theta} + \int_0^\omega \int_0^R \tilde{r}^{2\beta+3} \left| \frac{\partial p}{\partial \tilde{r}}(\tilde{r}, \theta_0, z) \right|^2 d\tilde{r} d\theta_0 \right.$$
$$\left. + \int_0^\omega \left\{ \frac{R^{2\epsilon}}{\epsilon} \int_r^R \tilde{r}^{1-2\epsilon} \left| \frac{\partial p}{\partial \tilde{r}}(\tilde{r}, \theta_0, z) \right|^2 d\tilde{r} + \int_0^R r_0^{2\beta+3} \int_a^b \left| \frac{\partial p}{\partial \tilde{z}}(r_0, \theta_0, \tilde{z}) \right|^2 d\tilde{z} dr_0 \right\} d\theta_0 \right\}.$$

To establish the weighted L^2-norm of $|p-b|$, multiply by $r^{2\beta}$ and take an integration over Ω. Then, we have

$$\int_\Omega r^{2\beta} |p(r,\theta,z) - b|^2 d\Omega \le c \int_\Omega r^{2\beta+2} \left(\left| \frac{1}{r} \frac{\partial p}{\partial \theta} \right|^2 + \left| \frac{\partial p}{\partial r} \right|^2 + \left| \frac{\partial p}{\partial z} \right|^2 \right) + r^{2\beta+2-2\epsilon} \left| \frac{\partial p}{\partial r} \right|^2 d\Omega$$
$$\le c \int_\Omega r^{2\beta+2-2\epsilon} |\nabla p\ |^2 d\Omega,$$

where $c = c(\Omega, \beta, \epsilon, (\beta+1)^{-1}, \epsilon^{-1}) \to \infty$ as $\epsilon \to 0$ and $\beta \to -1$. ∎

The next result follows from setting $\epsilon = 1$ in Lemma 4.3.

THEOREM 4.4. *Let $\beta > 0$ and $p \in H^1_\beta(\Omega)$; then there exist constants b, c satisfying*

$$\|p - b\|_{0,\beta-1} \le c \|\nabla p\|_{0,\beta}.$$

Proof. From Lemmas 4.2 and 4.3 (set $\epsilon = 1$), we deduce the result. ∎

COROLLARY 4.5. *If $p \in H^1_\beta(\Omega) \cap H^1(\Omega)/\mathbb{R}$ is the solution of Poisson equation with Neumann boundary condition, then for $\beta > 0$, by Theorem 4.4, p satisfies*

$$\|p\|_{0,\beta-1} \le c \|\nabla p\|_{0,\beta}. \tag{4.3}$$

We now consider functions with zero boundary conditions given on a nonempty portion of the boundary. The following theorem shows that the weighted-norm Poincaré inequality holds wherever the zero boundary conditions are located.

THEOREM 4.6. *Let $\beta > 0$ and $p \in H^1_\beta(\Omega)$. If $p = 0$ on $\Gamma_0 \neq \emptyset$, then*

$$\|p\|_{0,\beta-1} \ \le \ c\ \|\nabla p\|_{0,\beta}. \tag{4.4}$$

Proof. In order to prove (4.4), by Lemma 4.2, it is enough to show (4.1). Here, we show (4.1) for p satisfying $p = 0$ on $\Gamma_0 \subset \{\mathbf{x} \in \partial\Omega : z = a\}$; however, the other cases can be proved in the same manner. Consider the following expression for p:

$$p(r,\theta,z) = \int_{r_0}^r \frac{\partial p(\tilde{r}, \theta, z)}{\partial \tilde{r}} d\tilde{r} + \int_{\theta_0}^\theta \frac{\partial p(r_0, \tilde{\theta}, z)}{\partial \tilde{\theta}} d\tilde{\theta} + \int_a^z \frac{\partial p(r_0, \theta_0, \tilde{z})}{\partial \tilde{z}} d\tilde{z},$$

where $(r_0, \theta_0, a) \in \Gamma_0$. If $r \geq r_0$, then squaring both sides, applying the Cauchy inequality, and multiplying by $r_0^{2\beta+1}$ yields

$$\begin{aligned} r_0^{2\beta+1}|p(r,\theta,z)|^2 &\leq c\int_{r_0}^{r} \tilde{r}^{2\beta+1}\left|\frac{\partial p(\tilde{r},\theta,z)}{\partial \tilde{r}}\right|^2 d\tilde{r} + c\int_{\theta_0}^{\theta} r_0^{2\beta+1}\left|\frac{1}{r_0}\frac{\partial p(r_0,\tilde{\theta},z)}{\partial \tilde{\theta}}\right|^2 d\tilde{\theta} \\ &\quad + c\int_{a}^{z} r_0^{2\beta+1}\left|\frac{\partial p(r_0,\theta_0,\tilde{z})}{\partial \tilde{z}}\right|^2 d\tilde{z} \\ &\leq c\int_{0}^{R(\theta)} \tilde{r}^{2\beta+1}\left|\frac{\partial p(\tilde{r},\theta,z)}{\partial \tilde{r}}\right|^2 d\tilde{r} + c\int_{0}^{\omega} r_0^{2\beta+1}\left|\frac{1}{r_0}\frac{\partial p(r_0,\tilde{\theta},z)}{\partial \tilde{\theta}}\right|^2 d\tilde{\theta} \\ &\quad + c\int_{a}^{b} r_0^{2\beta+1}\left|\frac{\partial p(r_0,\theta_0,\tilde{z})}{\partial \tilde{z}}\right|^2 d\tilde{z}. \end{aligned}$$

First, integrate with respect to r_0, θ_0 and then multiply by $r^{2\beta+1}$ on both sides and integrate over Ω with respect to r, θ, and z, respectively. Then, by Fubini's theorem,

$$\int_\Omega r^{2\beta}|p|^2 d\Omega \leq c\left(\int_\Omega \tilde{r}^{2\beta}\left|\frac{\partial p}{\partial \tilde{r}}\right|^2 d\Omega + \int_\Omega r_0^{2\beta}\left|\frac{1}{r_0}\frac{\partial p}{\partial \tilde{\theta}}\right|^2 d\Omega + \int_\Omega r_0^{2\beta}\left|\frac{\partial p}{\partial \tilde{z}}\right|^2 d\Omega\right),$$

which yields (4.1).

If $r < r_0$, similarly, we square both sides, use the Cauchy inequality, and multiply by $r_0^{2\beta+1} r^{2\beta+1}$. Then, we have

$$\begin{aligned} &r_0^{2\beta+1} r^{2\beta+1}|p(r,\theta,z)|^2 \leq c r_0^{2\beta+1}\int_{r}^{r_0} \tilde{r}^{2\beta+1}\left|\frac{\partial p(\tilde{r},\theta,z)}{\partial \tilde{r}}\right|^2 d\tilde{r} \\ &+ c r^{2\beta+1}\int_{\theta_0}^{\theta} r_0^{2\beta+1}\left|\frac{1}{r_0}\frac{\partial p(r_0,\tilde{\theta},z)}{\partial \tilde{\theta}}\right|^2 d\tilde{\theta} + c r^{2\beta+1}\int_{a}^{z} r_0^{2\beta+1}\left|\frac{\partial p(r_0,\theta_0,\tilde{z})}{\partial \tilde{z}}\right|^2 d\tilde{z} \\ &\leq c r_0^{2\beta+1}\int_{0}^{R(\theta)} \tilde{r}^{2\beta+1}\left|\frac{\partial p}{\partial \tilde{r}}\right|^2 d\tilde{r} + c r^{2\beta+1}\int_{0}^{\omega} r_0^{2\beta+1}\left|\frac{1}{r_0}\frac{\partial p}{\partial \tilde{\theta}}\right|^2 d\tilde{\theta} + c r^{2\beta+1}\int_{a}^{b} r_0^{2\beta+1}\left|\frac{\partial p}{\partial \tilde{z}}\right|^2 d\tilde{z}. \end{aligned}$$

Taking integrations with respect to r_0, θ_0, r, θ, and z implies (4.1). ∎

So far, Poincaré inequalities in weighted Sobolev spaces for scalar functions have been established. We now extend these to vector functions.

THEOREM 4.7. *Let $\beta > 0$ and $\mathbf{u} \in H^1_\beta(\Omega)^3$ with $\mathbf{n} \times \mathbf{u} = \mathbf{0}$ on Γ_D and $\mathbf{n} \cdot \mathbf{u} = 0$ on Γ_N; then there exists a constant c such that*

$$\|\mathbf{u}\|_{0,\beta-1} \leq c\, \|\nabla \mathbf{u}\|_{0,\beta} \tag{4.5}$$

except for the case with $\Gamma_D = \{\mathbf{x} \in \partial\Omega : z = a \text{ and } b\}$ and $\Gamma_N = \partial\Omega \backslash \Gamma_D$.

Proof. Here, we prove (4.5) only for the case $\partial\Omega = \Gamma_N$. However, all the other cases can be proved easily by using the same methodology.

First, we rotate the domain so that the side $\theta = 0$ lies on the x-axis. Since $\mathbf{n}\cdot\mathbf{u} = 0$ on $\partial\Omega$, $\mathbf{n} \cdot \mathbf{u} = (n_1, n_2, n_3) \cdot (u_1, u_2, u_3) = n_3 u_3 = 0$ on the top and bottom of the domain, that is, where $z = a$ and b. Then, by Theorem 4.6, $\|u_3\|_{0,\beta-1} \leq c\|\nabla u_3\|_{0,\beta}$. From $\mathbf{n} = (0, n_2, 0)$ on $\theta = 0$, we have $\mathbf{n} \cdot \mathbf{u} = n_2 u_2 = 0$, which yields $u_2 = 0$; therefore, Theorem 4.6 yields $\|u_2\|_{0,\beta-1} \leq c\|\nabla u_2\|_{0,\beta}$. On the boundary that has

nonzero n_1-component, we have $n_1u_1 + n_2u_2 = 0$. Then, the triangle inequality leads to

$$\begin{aligned}\|\mathbf{u}\|^2_{0,\beta-1} &= \|u_1\|^2_{0,\beta-1} + \|u_2\|^2_{0,\beta-1} + \|u_3\|^2_{0,\beta-1}\\ &\le c\frac{1}{n_1^2}\|n_1u_1 + n_2u_2\|^2_{0,\beta-1} + c\frac{n_2^2}{n_1^2}\|u_2\|^2_{0,\beta-1} + \|u_2\|^2_{0,\beta-1} + \|u_3\|^2_{0,\beta-1}\\ &\le c\frac{1}{n_1^2}\|\nabla(n_1u_1 + n_2u_2)\|^2_{0,\beta} + c\|u_2\|^2_{0,\beta-1} + \|u_3\|^2_{0,\beta-1}\\ &\le c(\|\nabla u_1\|^2_{0,\beta} + \|\nabla u_2\|^2_{0,\beta} + \|\nabla u_3\|^2_{0,\beta}) = c\|\nabla\mathbf{u}\|^2_{0,\beta}. \qquad (4.6)\end{aligned}$$

If $\Gamma_D = \{\mathbf{x} \in \partial\Omega : z = a,\ b\}$ and $\Gamma_N = \partial\Omega\backslash\{\mathbf{x} \in \partial\Omega : z = a,\ b\}$, then $\mathbf{n} \times \mathbf{u} = \mathbf{0}$ on Γ_D and $\mathbf{n} \cdot \mathbf{u} = 0$ on Γ_N do not provide any information about boundary condition for u_3. Therefore, we exclude this case. ∎

5. Error bounds. Let $\mathcal{T}_h$ be a quasi-uniform partition of the domain $\bar{\Omega} = \cup_{\tau\in\mathcal{T}_h}\tau$ and each finite element $\tau \in \mathcal{T}_h$ be a closed subset of $\bar{\Omega}$ with $h := \max\{h_\tau : h_\tau = diam(\tau)\}$. Assume that the partition $\mathcal{T}_h$ is regular so that we may choose a finite element basis that is conforming and satisfies the standard approximation properties (see [11]). We also assume that there exists a constant, ρ, satisfying $h \le \rho \min_\tau h_\tau$. Let $\mathcal{P}^h$ denote the space of $\mathcal{C}^0$ piecewise polynomials on each finite element and $\mathcal{W}^h$ be a subspace of $\mathcal{P}^h$ such that

$$\mathcal{W}^h = \{\mathbf{v}^h \in \mathcal{P}^h : \mathbf{n} \times \mathbf{v}^h = \mathbf{0} \text{ on } \Gamma_D,\ \ \mathbf{n} \cdot \mathbf{v}^h = 0 \text{ on } \Gamma_N\}.$$

Here, the normal vector is defined to be normal to the corresponding boundary face. On the edge where the Dirichlet and Neumann boundary conditions meet, for example, $\mathbf{n} \times \mathbf{v}^h = \mathbf{0}$ on Γ_1, $\mathbf{n} \cdot \mathbf{v}^h = 0$ on Γ_2, $\Gamma_1 \cap \Gamma_2 \neq \emptyset$, and Γ_1 includes the edge, the normal vector on the edge can be defined as a normal to the boundary face Γ_1 or the average of two normal vectors corresponding to Γ_1 and Γ_2. Either case provides the same numerical results. The discrete weighted-norm least-squares approximation minimizes the functional

$$G_w(\mathbf{u}^h; f) = \min_{\mathbf{v}^h\in\mathcal{W}^h} G_w(\mathbf{v}^h; f). \qquad (5.1)$$

For ease of discussion, we will consider cubes. However, arbitrary triangulations can be handled in a similar manner. Let $\mathcal{I}^h$ be a standard polynomial interpolation operator such that $\mathcal{I}^h p = p$, for any $p \in \mathbb{Q}_k(\tau)$, where $\mathbb{Q}_k(\tau)$ is a set of kth order polynomials with respect to each variable on τ. Define $\mathcal{I}^h_0$ by

$$\mathcal{I}^h_0 u|_\tau \ = \ \begin{cases} \mathcal{I}^h u|_\tau = \sum_{i=1}^{(k+1)^3} u(a_i)\phi_i & \text{if } \tau \text{ does not meet the singular edge, } E,\\ \sum_{a_i\notin N} u(a_i)\phi_i & \text{if } \tau \text{ meets the singular edge, } E,\end{cases}$$

where ϕ_i are the basis functions, a_i are the nodal points corresponding to ϕ_i, and $N = \{a_j : a_j = (0, 0, z_j)\}$ is the set of $k + 1$ nodes seated along the singular edge, E. Here, it is easy to see that $\mathcal{I}^h_0\mathbf{u} \in \mathcal{W}^h$ for $\mathbf{u} \in H^1_\beta(\Omega)^3$ with $\mathbf{n} \times \mathbf{u} = \mathbf{0}$ on Γ_D and $\mathbf{n} \cdot \mathbf{u} = 0$ on Γ_N since we can consider $\mathcal{I}^h_0\mathbf{u}$ as $\mathcal{I}^h(\xi\mathbf{u})$, where the smooth function ξ satisfies $\xi = 0$ when $r < \epsilon$ and $\xi = 1$ when $r \ge 2\epsilon$ for sufficiently small positive ϵ.

LEMMA 5.1. *Let $u \in H^m_\beta(\Omega)$; then*

$$\|u - \mathcal{I}^h_0 u\|_{1,\beta} \ \le \ c\ h^{m-1}\|u\|_{m,\beta}, \qquad (5.2)$$

for $3/2 < m$ and any $0 < \beta$, where $\mathcal{I}_0^h$ is the modified interpolation operator into piecewise polynomials of degree $m-1 \le k$.

Proof. We rewrite

$$\|u - \mathcal{I}_0^h u\|_{1,\beta}^2 \;=\; \sum_{\tau \in \mathcal{T}_h} \|u - \mathcal{I}_0^h u\|_{1,\beta,\tau}^2,$$

where $\|u\|_{1,\beta,\tau} = \left(\int_\tau r^{2\beta-2}|u|^2 + r^{2\beta}|\nabla u|^2 d\tau\right)^{\frac{1}{2}}$. Consider $\|u - \mathcal{I}_0^h u\|_{1,\beta,\tau}^2$ on each element τ. If τ does not touch the singular edge, then, by the definition of $\mathcal{I}_0^h$, $\mathcal{I}_0^h u|_\tau = \mathcal{I}^h u|_\tau$. Let $r_{min} = \inf\{r : (x,y,z) \in \tau\}$ and $r_{max} = \sup\{r : (x,y,z) \in \tau\}$ in τ. Then, $ch \le r_{min} \le r = (x^2+y^2)^{\frac{1}{2}} \le r_{max} \le r_{min} + \sqrt{3}h$ and the standard interpolation property yield

$$\begin{aligned}
\|u - \mathcal{I}_0^h u\|_{1,\beta,\tau}^2 &= \|u - \mathcal{I}^h u\|_{1,\beta,\tau}^2 = \int_\tau r^{2\beta}|\nabla(u - \mathcal{I}^h u)|^2 + r^{2(\beta-1)}|u - \mathcal{I}^h u|^2 d\tau \\
&\le r_{max}^{2\beta} \int_\tau |\nabla(u - \mathcal{I}^h u)|^2 d\tau + r_{max}^{2\beta} r_{min}^{-2} \int_\tau |u - \mathcal{I}^h u|^2 d\tau \\
&\le c r_{max}^{2\beta} h^{2(m-1)} |u|_{m,\tau}^2 + c r_{max}^{2\beta} r_{min}^{-2} h^{2m} |u|_{m,\tau}^2 = c r_{max}^{2\beta} h^{2(m-1)} (1 + r_{min}^{-2} h^2) |u|_{m,\tau}^2 \\
&\le c r_{max}^{2\beta} h^{2(m-1)} |u|_{m,\tau}^2 \le c h^{2(m-1)} r_{max}^{2\beta} r_{min}^{-2\beta} \int_\tau r^{2\beta} |D^m u|^2 d\tau \\
(5.3) &\le c h^{2(m-1)} \left(\frac{r_{min} + \sqrt{3}h}{r_{min}}\right)^{2\beta} \int_\tau r^{2\beta} |D^m u|^2 d\tau \le c h^{2(m-1)} \|D^m u\|_{0,\beta,\tau}^2.
\end{aligned}$$

Next, consider the case in which $\tau \cap E \neq \emptyset$. Let $\delta \in \mathcal{C}^\infty$ be a cut-off function defined by

$$\delta(r) = \begin{cases} 1 & \text{if} \quad r \le \frac{h}{3k}, \\ 0 & \text{if} \quad r > \frac{h}{2k}, \end{cases} \tag{5.4}$$

with $|\delta^{(m)}| \le ch^{-m}$, where $\delta^{(m)}$ is the mth derivative of δ. By the triangle inequality,

$$\|u - \mathcal{I}_0^h u\|_{1,\beta,\tau}^2 \le c(\|\delta u - \mathcal{I}_0^h(\delta u)\|_{1,\beta,\tau}^2 + \|(1-\delta)u - \mathcal{I}_0^h((1-\delta)u)\|_{1,\beta,\tau}^2). \tag{5.5}$$

On these τ, $\mathcal{I}_0^h(\delta u) = 0$. Thus, for the first term in (5.5), the properties in δ and Fubini's theorem imply that

$$\begin{aligned}
&\|\delta u - \mathcal{I}_0^h(\delta u)\|_{1,\beta,\tau}^2 = \|\delta u\|_{1,\beta,\tau}^2 = \int_\tau r^{2\beta}|\nabla(\delta u)|^2 + r^{2(\beta-1)}|\delta u|^2 d\tau \\
&\le c \int_\tau r^{2\beta}(|\nabla\delta \cdot u|^2 + |\delta \nabla u|^2) + r^{2(\beta-1)}|\delta u|^2 d\tau \\
&\le c \iiint_{\frac{h}{3k}}^{\frac{h}{2k}} r^{2\beta} h^{-2} |u|^2 d\tau + c \iiint_0^{\frac{h}{3k}} r^{2\beta}|\nabla u|^2 + r^{2(\beta-1)}|u|^2 d\tau \\
&\le c \iiint_{\frac{h}{3k}}^{\frac{h}{2k}} r^{2(\beta-1)} |u|^2 d\tau + c \iiint_0^{\frac{h}{3k}} r^{2\beta}|\nabla u|^2 + r^{2(\beta-1)}|u|^2 d\tau \\
&\le c \int_\tau r^{2(m-1)}(r^{2(\beta-m+1)}|\nabla u|^2 + r^{2(\beta-m)}|u|^2) d\tau \le c h^{2(m-1)} \|u\|_{m,\beta,\tau}^2.
\end{aligned}$$

For the second term in (5.5), use $\mathcal{I}_0^h((1-\delta)u) = \mathcal{I}^h((1-\delta)u)$ and Theorem 4.6 to obtain

$$\|(1-\delta)u - \mathcal{I}_0^h((1-\delta)u)\|_{1,\beta,\tau}^2 \le c|(1-\delta)u - \mathcal{I}^h((1-\delta)u)|_{1,\beta,\tau}^2.$$

Since $(1-\delta)u \in H^m(\tau)$, we use the standard interpolation property, Fubini's theorem, and the properties of δ to obtain

$$\begin{aligned}
&|(1-\delta)u - \mathcal{I}^h((1-\delta)u)|_{1,\beta,\tau}^2 \le ch^{2\beta}\int_\tau |\nabla((1-\delta)u - \mathcal{I}^h((1-\delta)u))|^2 d\tau \\
&\le ch^{2\beta}h^{2(m-1)}\int_\tau |D^m((1-\delta)u)|^2 d\tau \le ch^{2(\beta+m-1)}\int_\tau \sum_{j=0}^{m} |D^{m-j}(1-\delta)D^j u|^2 d\tau \\
&\le ch^{2(\beta+m-1)}\left(\iiint_{\frac{h}{3k}}^{\frac{h}{2k}} \sum_{j=0}^{m-1} |h^{j-m}D^j u|^2 d\tau + \iiint_{\frac{h}{3k}}^{r(\theta)} |(1-\delta)D^m u|^2 d\tau\right) \\
&\le ch^{2(\beta+m-1)}\left(\sum_{j=0}^{m-1}\iiint_{\frac{h}{3k}}^{\frac{h}{2k}} h^{-2\beta}r^{2(\beta+j-m)}|D^j u|^2 d\tau + \iiint_{\frac{h}{3k}}^{r(\theta)} h^{-2\beta}r^{2\beta}|D^m u|^2 d\tau\right) \\
&= ch^{2(m-1)}\sum_{j=0}^{m}\int_\tau r^{2(\beta+j-m)}|D^j u|^2 d\tau = ch^{2(m-1)}\|u\|_{m,\beta,\tau}^2.
\end{aligned}$$

Hence, we have

$$\|u - \mathcal{I}_0^h u\|_{1,\beta}^2 = \sum_{\tau\in\mathcal{T}_h} \|u - \mathcal{I}_0^h u\|_{1,\beta,\tau}^2 \le ch^{2(m-1)}\sum_{\tau\in\mathcal{T}_h}\|u\|_{m,\beta,\tau}^2 \le ch^{2(m-1)}\|u\|_{m,\beta}^2. \qquad \square$$

LEMMA 5.2. *Let $\mathbf{u}^h \in \mathcal{W}^h$; then there exists a constant c such that*

$$\|\mathbf{u}^h\|_{0,\gamma} \le ch^{-\eta}\|\mathbf{u}^h\|_{0,\gamma+\eta} \tag{5.6}$$

for any real $\gamma \ge 0$ and any $\eta > 0$.

Proof. In [23], inequality (5.6) is shown in two dimensions. The analogous three dimensional result can be proved in the same manner. $\square$

Lemma 5.1 shows that there exists an approximation of the solution in the finite dimensional subspace $\mathcal{W}^h$ that yields an optimal weighted H^1-convergence. By using this, we can show the following error estimates.

THEOREM 5.3. *Let $\mathbf{u} \in \mathcal{U}$ be the solution of* (2.3) *and $\mathbf{u}^h \in \mathcal{W}^h$ be the solution of* (5.1). *Assume $\mathbf{u} \in H_\beta^{\alpha+\beta}(\Omega)^3$ and L satisfies Theorem* 3.4. *Then,*

$$\|\mathbf{u} - \mathbf{u}^h\|_{l,\beta} \le c\, h^{\alpha+\beta-l}\|\mathbf{u}\|_{\alpha+\beta,\beta} \quad \text{for } l = 0, 1, \tag{5.7}$$

$$\|\mathbf{u} - \mathbf{u}^h\| \le c\, h^{\alpha}\|\mathbf{u}\|_{\alpha+\beta,\beta}, \tag{5.8}$$

where $3/2 < \alpha + \beta \le k + 1$ and k is the degree of the piecewise polynomials in $\mathcal{W}^h$.

Proof. First, we prove (5.7) for $l = 1$. Since $\mathbf{u}^h$ is the minimizer of (5.1), by Theorem 3.4, the triangle inequality, and Lemma 5.1, we have

$$\begin{aligned}
\|\mathbf{u} - \mathbf{u}^h\|_{1,\beta} &\le c\|L(\mathbf{u} - \mathbf{u}^h)\|_{0,\beta} \le c\|L(\mathbf{u} - \mathcal{I}_0^h\mathbf{u})\|_{0,\beta} \le c\|\mathbf{u} - \mathcal{I}_0^h\mathbf{u}\|_{1,\beta} \\
&\le ch^{\alpha+\beta-1}\|\mathbf{u}\|_{\alpha+\beta,\beta}.
\end{aligned} \tag{5.9}$$

In order to prove the result for $l = 0$, we consider the weighted L^2-norm on each element. We have

$$\|\mathbf{u} - \mathbf{u}^h\|_{0,\beta}^2 = \sum_{\tau \in \mathcal{T}_h} \|\mathbf{u} - \mathbf{u}^h\|_{0,\beta,\tau}^2.$$

By the Cauchy inequality, we have

$$\begin{aligned}\|\mathbf{u} - \mathbf{u}^h\|_{0,\beta,\tau}^2 &= \int_\tau r^{2\beta} |\mathbf{u} - \mathbf{u}^h|^2 d\tau \le \left(\int_\tau 1^{\frac{3}{2}} d\tau \right)^{\frac{2}{3}} \left(\int_\tau \left(r^{2\beta} |\mathbf{u} - \mathbf{u}^h|^2 \right)^3 d\tau \right)^{\frac{1}{3}} \\ &\le c\, h^{3 \cdot \frac{2}{3}} \left(\int_\tau \left(r^\beta |\mathbf{u} - \mathbf{u}^h| \right)^6 d\tau \right)^{\frac{1}{6} \cdot 2} = c\, h^2 \|r^\beta (\mathbf{u} - \mathbf{u}^h)\|_{L^6(\tau)}^2. \end{aligned} \tag{5.10}$$

Since $H^1(\Omega)$ is continuously embedded into $L^6(\Omega)$ [11] and $r^\beta \mathbf{u} \in H^1(\Omega)^3$,

$$\begin{aligned}\|\mathbf{u} - \mathbf{u}^h\|_{0,\beta}^2 &\le c\, h^2 \sum_{\tau \in \mathcal{T}_h} \|r^\beta (\mathbf{u} - \mathbf{u}^h)\|_{L^6(\tau)}^2 \le c\, h^2 \|r^\beta (\mathbf{u} - \mathbf{u}^h)\|_{L^6(\Omega)}^2 \\ &\le c\, h^2 \|r^\beta (\mathbf{u} - \mathbf{u}^h)\|_1^2 \le c\, h^2 \|\mathbf{u} - \mathbf{u}^h\|_{1,\beta}^2.\end{aligned}$$

Applying (5.9) to the above implies that

$$\|\mathbf{u} - \mathbf{u}^h\|_{0,\beta} \le ch \|\mathbf{u} - \mathbf{u}^h\|_{1,\beta} \le chh^{\alpha+\beta-1} \|\mathbf{u}\|_{\alpha+\beta,\beta} = ch^{\alpha+\beta} \|\mathbf{u}\|_{\alpha+\beta,\beta}.$$

To establish (5.8), we let $K = \{\tau \in \mathcal{T}_h : \tau \cap E \neq \emptyset\}$ and separate the norm $\|\mathbf{u} - \mathbf{u}^h\|$ into two parts:

$$\|\mathbf{u} - \mathbf{u}^h\|^2 = \sum_{\tau \in K} \|\mathbf{u} - \mathbf{u}^h\|_{0,\tau}^2 + \sum_{\tau \in \mathcal{T}_h \backslash K} \|\mathbf{u} - \mathbf{u}^h\|_{0,\tau}^2.$$

First, we consider the case $\beta < 1$. We have $r \le ch$ when $\tau \in K$; therefore,

$$\begin{aligned}\|\mathbf{u} - \mathbf{u}^h\|^2 &= \sum_{\tau \in K} \|\mathbf{u} - \mathbf{u}^h\|_{0,\tau}^2 + \sum_{\tau \in \mathcal{T}_h \backslash K} \|\mathbf{u} - \mathbf{u}^h\|_{0,\tau}^2 \\ &\le \sum_{\tau \in K} h^{2(1-\beta)} \|\mathbf{u} - \mathbf{u}^h\|_{0,\beta-1,\tau}^2 + \sum_{\tau \in \mathcal{T}_h \backslash K} \|\mathbf{u} - \mathbf{u}^h\|_{0,\tau}^2. \end{aligned} \tag{5.11}$$

When $\tau \in \mathcal{T}_h \backslash K$, similarly to (5.10), we have

$$\|\mathbf{u} - \mathbf{u}^h\|_{0,\tau}^2 \le c\, h^2 \|\mathbf{u} - \mathbf{u}^h\|_{L^6(\tau)}^2.$$

Then, the continuous embedding from H^1 into L^6 and $h \le cr$ when $\tau \in \mathcal{T}_h \backslash K$ yields

$$\begin{aligned}&\sum_{\tau \in \mathcal{T}_h \backslash K} \|\mathbf{u} - \mathbf{u}^h\|_{0,\tau}^2 \le c\, h^2 \sum_{\tau \in \mathcal{T}_h \backslash K} \|\mathbf{u} - \mathbf{u}^h\|_{L^6(\tau)}^2 \le c\, h^2 \|\mathbf{u} - \mathbf{u}^h\|_{L^6(\mathcal{T}_h \backslash K)}^2 \\ &\le ch^2 \|\mathbf{u} - \mathbf{u}^h\|_{H^1(\mathcal{T}_h \backslash K)}^2 \le c\, h^2 h^{-2\beta} \int_{\mathcal{T}_h \backslash K} r^{2\beta} \left(|\mathbf{u} - \mathbf{u}^h|^2 + |\nabla(\mathbf{u} - \mathbf{u}^h)|^2 \right) d\Omega \\ &\le\ ch^{2(1-\beta)} \|\mathbf{u} - \mathbf{u}^h\|_{1,\beta}^2.\end{aligned}$$

Combining the above with (5.11) and using (5.9) gives

$$\|\mathbf{u} - \mathbf{u}^h\| \le ch^{1-\beta} \|\mathbf{u} - \mathbf{u}^h\|_{1,\beta} \le ch^\alpha \|\mathbf{u}\|_{\alpha+\beta,\beta}.$$

For $\beta \geq 1$, by the triangle inequality, Lemma 5.2, and (5.9), we have

$$\begin{aligned}\|\mathbf{u}-\mathbf{u}^h\|^2 &\leq c\|\mathbf{u}-\mathcal{I}_0^h\mathbf{u}\|^2 + c\|\mathcal{I}_0^h\mathbf{u}-\mathbf{u}^h\|^2 \leq c\|\mathbf{u}-\mathcal{I}_0^h\mathbf{u}\|^2 + ch^{2(1-\beta)}\|\mathcal{I}_0^h\mathbf{u}-\mathbf{u}^h\|_{0,\beta-1}^2 \\ &\leq c\sum_{\tau\in\mathcal{T}_h}\|\mathbf{u}-\mathcal{I}_0^h\mathbf{u}\|_{0,\tau}^2 + ch^{2(1-\beta)}\left(\|\mathbf{u}-\mathcal{I}_0^h\mathbf{u}\|_{0,\beta-1}^2 + \|\mathbf{u}-\mathbf{u}^h\|_{0,\beta-1}^2\right) \\ &\leq c\sum_{\tau\in\mathcal{T}_h}\|\mathbf{u}-\mathcal{I}_0^h\mathbf{u}\|_{0,\tau}^2 + ch^{2(1-\beta)}\|\mathbf{u}-\mathcal{I}_0^h\mathbf{u}\|_{1,\beta}^2. \qquad (5.12)\end{aligned}$$

If $\tau \in \mathcal{T}_h \backslash K$, then we simply use $h \leq cr$ to obtain

$$\|\mathbf{u}-\mathcal{I}_0^h\mathbf{u}\|_{0,\tau}^2 \leq ch^{2(1-\beta)}\|\mathbf{u}-\mathcal{I}_0^h\mathbf{u}\|_{0,\beta-1,\tau}^2. \qquad (5.13)$$

If $\tau \in K$, we recall the definition of the operator $\mathcal{I}_0^h$ and let δ be a smooth cut-off function defined in (5.4). By the triangle inequality,

$$\|\mathbf{u}-\mathcal{I}_0^h\mathbf{u}\|_{0,\tau}^2 \leq c\|(1-\delta)\mathbf{u}-\mathcal{I}_0^h((1-\delta)\mathbf{u})\|_{0,\tau}^2 + c\|\delta\mathbf{u}-\mathcal{I}_0^h(\delta\mathbf{u})\|_{0,\tau}^2.$$

Since $(1-\delta)\mathbf{u} \in H^{\alpha+\beta}(\tau)^3$ and $\mathcal{I}_0^h((1-\delta)\mathbf{u}) = \mathcal{I}^h((1-\delta)\mathbf{u})$ from the definition, the analogous calculation in the proof of Lemma 5.1 provides

$$\begin{aligned}\|(1-\delta)\mathbf{u}-\mathcal{I}_0^h((1-\delta)\mathbf{u})\|_{0,\tau}^2 &= \|(1-\delta)\mathbf{u}-\mathcal{I}^h((1-\delta)\mathbf{u})\|_{0,\tau}^2 \\ &\leq ch^{2(\alpha+\beta)}\int_\tau \left|D^{\alpha+\beta}((1-\delta)\mathbf{u})\right|^2 d\tau = ch^{2(\alpha+\beta)}\iiint_{\frac{h}{3k}}^{r(\theta)}\left|D^{\alpha+\beta}((1-\delta)\mathbf{u})\right|^2 d\tau \\ &\leq ch^{2(\alpha+\beta)}h^{-2\beta}\iiint_{\frac{h}{3k}}^{r(\theta)} r^{2\beta}\left|D^{\alpha+\beta}((1-\delta)\mathbf{u})\right|^2 d\tau \\ &\leq ch^{2\alpha}\sum_{l=0}^{\alpha+\beta}\int_\tau r^{2(\beta-(\alpha+\beta)+l)}\left|D^l\mathbf{u}\right|^2 d\tau = ch^{2\alpha}\|\mathbf{u}\|_{\alpha+\beta,\beta,\tau}^2. \qquad (5.14)\end{aligned}$$

By the definition of $\mathcal{I}_0^h$ and $\|\mathbf{u}\|_{0,\beta-(\alpha+\beta)} \leq \|\mathbf{u}\|_{\alpha+\beta,\beta}$, we have

$$\begin{aligned}\|\delta\mathbf{u}-\mathcal{I}_0^h(\delta\mathbf{u})\|_{0,\tau}^2 = \|\delta\mathbf{u}\|_{0,\tau}^2 &\leq \int_\tau r^{2\alpha}r^{-2\alpha}|\mathbf{u}|^2 d\tau \leq ch^{2\alpha}\int_\tau r^{-2\alpha}|\mathbf{u}|^2 d\tau \\ &= ch^{2\alpha}\int_\tau r^{2(\beta-(\alpha+\beta))}|\mathbf{u}|^2 d\tau = ch^{2\alpha}\|\mathbf{u}\|_{0,\beta-(\alpha+\beta),\tau}^2 \leq ch^{2\alpha}\|\mathbf{u}\|_{\alpha+\beta,\beta,\tau}^2. \qquad (5.15)\end{aligned}$$

Substituting (5.13), (5.14), and (5.15) into (5.12) and applying Lemma 5.1 yields

$$\|\mathbf{u}-\mathbf{u}^h\|^2 \leq ch^{2\alpha}\sum_{\tau\in K}\|\mathbf{u}\|_{\alpha+\beta,\beta,\tau}^2 + ch^{2(1-\beta)}\|\mathbf{u}-\mathcal{I}_0^h\mathbf{u}\|_{1,\beta}^2 \leq ch^{2\alpha}\|\mathbf{u}\|_{\alpha+\beta,\beta}. \qquad \square$$

Minimizing the least-squares functional in the weighted norm using H^1-conforming elements and appropriate values of the weight allows the approximate solution to converge. The above theorem also implies optimal L^2-error convergence with H^1-conforming finite elements. In the next section, we consider error bounds using graded mesh refinements.

6. Graded mesh refinement. The paper [4] is one of the earliest papers to examine graded mesh refinement in the context of weighted Sobolev spaces. Following that, many have used similar mesh refinements when weighted Sobolev spaces were

considered to overcome the difficulties associated with singularities (see [2], [3]). In this section we show some error estimate results using graded mesh refinements in our methodology.

Assume $\beta < 1$. For the global mesh parameter, h $(0 < h < 1)$, let $\mathcal{T}_h^m$ be a triangulation of the domain $\overline{\Omega} = \cup_{\tau \in \mathcal{T}_h^m} \tau$, where the diameter, $h_\tau = diam(\tau)$, of each tetrahedron, τ, is defined by

$$h_\tau \sim \begin{cases} h^{\frac{1}{1-\beta}} & \text{if} \quad \tau \cap E \neq \emptyset, \\ h r_\tau^\beta & \text{if} \quad \tau \cap E = \emptyset, \end{cases} \tag{6.1}$$

where $r_\tau = \min\{r = \sqrt{x^2+y^2} : (x,y,z) \in \tau\}$. Define $\mathcal{V}^h$ to be a space of $\mathcal{C}^0$ piecewise polynomials on each finite element satisfying $\mathbf{n} \times \mathbf{v}^h = \mathbf{0}$ on Γ_D and $\mathbf{n} \cdot \mathbf{v}^h = 0$ on Γ_N and let $\mathbf{u}^h$ satisfy

$$G_w(\mathbf{u}^h; f) = \min_{\mathbf{v}^h \in \mathcal{V}^h} G_w(\mathbf{v}^h; f). \tag{6.2}$$

THEOREM 6.1. *Let $\mathbf{u} \in H_\beta^{\alpha+\beta}(\Omega)^3 \cap \mathcal{U}$ and $\mathbf{u}^h \in \mathcal{V}^h$ be the solutions of* (2.3) *and* (6.2)*, respectively. Suppose that $\beta < 1$ and the hypotheses of Lemma* 3.1 *are satisfied. Then,*

$$\|\mathbf{u} - \mathbf{u}^h\| \leq c\ h^{\alpha+\beta} \|\mathbf{u}\|_{\alpha+\beta,\beta}, \tag{6.3}$$

where $3/2 < \alpha + \beta \leq k+1$ and k is the degree of the piecewise polynomials in $\mathcal{V}^h$.

Proof. Let $K = \{\tau \in \mathcal{T}_h^m : \tau \cap E \neq \emptyset\}$. By the Cauchy inequality and (6.1), we have

$$\begin{aligned}
\|\mathbf{u} - \mathbf{u}^h\|^2 &= \sum_{\tau \in K} \|\mathbf{u} - \mathbf{u}^h\|_{0,\tau}^2 + \sum_{\tau \in \mathcal{T}_h^m \setminus K} \|\mathbf{u} - \mathbf{u}^h\|_{0,\tau}^2 \\
&\leq \sum_{\tau \in K} c h_\tau^{2(1-\beta)} \|\mathbf{u} - \mathbf{u}^h\|_{0,\beta-1,\tau}^2 + \sum_{\tau \in \mathcal{T}_h^m \setminus K} \left(\int_\tau 1 d\tau \right)^{\frac{2}{3}} \left(\int_\tau |\mathbf{u} - \mathbf{u}^h|^6 d\tau \right)^{\frac{1}{3}} \\
&\leq \sum_{\tau \in K} c h^2 \|\mathbf{u} - \mathbf{u}^h\|_{0,\beta-1,\tau}^2 + \sum_{\tau \in \mathcal{T}_h^m \setminus K} c h_\tau^2 ||\mathbf{u} - \mathbf{u}^h||_{L^6(\tau)}^2 \\
&\leq \sum_{\tau \in K} c h^2 \|\mathbf{u} - \mathbf{u}^h\|_{0,\beta-1,\tau}^2 + \sum_{\tau \in \mathcal{T}_h^m \setminus K} c h^2 ||r^\beta(\mathbf{u} - \mathbf{u}^h)||_{L^6(\tau)}^2. \qquad (6.4)
\end{aligned}$$

The continuous embedding of $H^1(\Omega)$ into $L^6(\Omega)$ and $r^\beta \mathbf{u} \in H^1(\Omega)^3$ leads to

$$\begin{aligned}
\sum_{\tau \in \mathcal{T}_h^m \setminus K} c h^2 ||r^\beta(\mathbf{u} - \mathbf{u}^h)||_{L^6(\tau)}^2 &\leq c h^2 ||r^\beta(\mathbf{u} - \mathbf{u}^h)||_{L^6(\Omega)}^2 \leq c h^2 ||r^\beta(\mathbf{u} - \mathbf{u}^h)||_1^2 \\
&\leq c h^2 ||\mathbf{u} - \mathbf{u}^h||_{1,\beta}^2. \qquad (6.5)
\end{aligned}$$

We define $\mathcal{I}_0^h$ in the same manner as in section 5 on each element $\tau \in \mathcal{T}_h^m$. Then, a calculation analogous to the proof of Lemma 5.1 yields

$$\begin{aligned}
\|\mathbf{u} - \mathcal{I}_0^h \mathbf{u}\|_{1,\beta} &\leq c h^{\alpha+\beta-1} |\mathbf{u}|_{\alpha+\beta,(\alpha+\beta)\beta} + c h^{\frac{\alpha+\beta-1}{1-\beta}} \|\mathbf{u}\|_{\alpha+\beta,\beta} \\
&= c h^{\alpha+\beta-1} \left(|\mathbf{u}|_{\alpha+\beta,(\alpha+\beta)\beta} + c h^{\frac{\alpha+\beta-1}{1-\beta}\beta} \|\mathbf{u}\|_{\alpha+\beta,\beta} \right). \qquad (6.6)
\end{aligned}$$

In (6.6), since $r^{(\alpha+\beta)\beta} < r^{\beta}$ and $h^{\frac{\alpha+\beta-1}{1-\beta}\beta} < 1$, we have

$$\|\mathbf{u} - \mathcal{I}_0^h \mathbf{u}\|_{1,\beta} \leq c h^{\alpha+\beta-1} \|\mathbf{u}\|_{\alpha+\beta,\beta}. \tag{6.7}$$

Since $\mathbf{u}^h$ is the minimizer of (6.2), Theorem 3.4 and the triangle inequality yield

$$\|\mathbf{u} - \mathbf{u}^h\|_{1,\beta} \leq c\|L(\mathbf{u} - \mathbf{u}^h)\|_{0,\beta} \leq c\|L(\mathbf{u} - \mathcal{I}_0^h \mathbf{u})\|_{0,\beta} \leq c\|\mathbf{u} - \mathcal{I}_0^h \mathbf{u}\|_{1,\beta}. \tag{6.8}$$

Therefore, (6.4), (6.5), (6.8), and (6.7) imply the error bound (6.3). ∎

In section 5, we demonstrated optimal L^2-error convergence. Now, the above theorem shows that graded mesh refinement provides faster L^2-convergence by doing extra work around the singular area. Graded mesh refinement does the same work locally so that the finer mesh helps the order of convergence. However, in this paper, we want to observe the error convergence without any mesh refinement. Therefore, in the following section, we present numerical tests on uniform grids.

7. Computational results. In this section, we present some numerical examples. By minimizing the weighted-norm least-squares functional, we verify the optimal error convergence given in section 5. As mentioned in the introduction, if the internal angle of the edge, ω, is bigger than π or different types of boundary conditions meet at an edge with the internal angle bigger than $\pi/2$, then the boundary singularities occur. Thus, the solution of the Poisson equation is in H^{1+s}, where $s < \pi/\omega$ if the problem has Dirichlet or Neumann boundary conditions at a reentrant edge or $s < \pi/(2\omega)$ if Dirichlet and Neumann boundary conditions meet at an edge with inner angle ω [17]. Here, we construct our test problems based on the leading term of the singular part of the solution of Poisson's equation.

The software package FOSPACK (cf. [30]) was used to build the discrete system and to solve it by a conjugate gradient iterative method preconditioned by algebraic multigrid using W(1,1)-cycles. The stopping criterion for the iteration was a residual reduction 10^{-8}. This unnecessarily large reduction was used to remove the algebraic error from the calculation of the convergence of the discrete solution. No graded mesh refinement is applied.

The domain of Examples 7.1 and 7.2 is an L-shaped polyhedral cylinder:

$$\Omega = (-0.5, 0.5) \times (-0.5, 0.5) \times (0, 1) \backslash [0, 0.5) \times (-0.5, 0] \times (0, 1).$$

Example 7.1. We choose $\mathbf{u} = \nabla p$, where p has the form

$$p = \delta(r) r^{\frac{2}{3}} \sin(2\theta/3) \sin(\pi z)$$

in the cylindrical coordinate system with $\delta(r)$ a smooth cut-off function satisfying $\delta(r) = 1$ when $r \leq 0.25$ and $\delta(r) = 0$ when $r \geq 0.375$, and choose f by Δp. Then, p satisfies

$$\Delta p = f \quad \text{in } \Omega \tag{7.1}$$

with $p = 0$ on $\partial\Omega$, and $\mathbf{u} \in H^{\frac{2}{3}}(\Omega)^3$ satisfies

$$\nabla \times \mathbf{u} = \mathbf{0}, \quad \nabla \cdot \mathbf{u} = f \quad \text{in } \Omega \tag{7.2}$$

$$\mathbf{n} \times \mathbf{u} = \mathbf{0} \quad \text{on } \partial\Omega. \tag{7.3}$$

TABLE 7.1
Example 1 *with* $\beta = 4/3$.

	$\|\|\mathbf{u}-\mathbf{u}^h\|\|$		$\|\|\mathbf{u}-\mathbf{u}^h\|\|_{0,\beta}$		$\|\mathbf{u}-\mathbf{u}^h\|_{1,\beta}$	
1/8	2.398e-01	rates	3.734e-02	rates	7.775e-01	rates
1/16	1.074e-01	1.16	1.252e-02	1.58	2.223e-01	1.81
1/32	5.532e-02	0.96	3.116e-03	2.01	6.177e-02	1.85
1/64	3.350e-02	0.72	7.891e-04	1.98	1.875e-02	1.72
1/128	2.098e-02	0.68	1.978e-04	2.00	6.855e-03	1.45

It is easy to see that $\mathbf{u}$ in the above example is not in $H^1(\Omega)^3$. Since $\alpha = 2/3$ in this case, our theory holds for any β between 5/6 and 5/3. Here, we choose $\beta = 4/3$ so that $3/2 < \alpha + \beta = 2$ and observe the errors in L^2-, weighted L^2-, and weighted H^1-norms in Table 7.1. The error estimates in Theorem 5.3 predict the asymptotic L^2-error convergence rate to be approximately $O(h^{\frac{2}{3}})$ and the weighted L^2- and weighted H^1-rates at $O(h^2)$ and $O(h)$, respectively. We report the numbers in Table 7.1 to support our theory. The convergence rate of the weighted H^1-seminorm in Table 7.1 shows better convergence than expected at the resolution we are able to compute. However, based on two dimensional test results, it is likely that on increasingly fine meshes the asymptotic rate will approach what theory predicted.

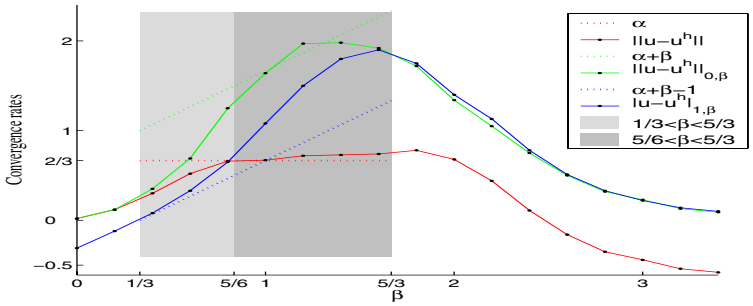


FIG. 7.1. *Finite element convergence rate as a function of* β *for Example* 7.1.

In Figure 7.1, we present the results with more β values when mesh size decreases from 1/64 to 1/128. It shows that the errors in L^2- and weighted H^1-norms behave as expected and better than expected for weighted L^2-norm errors. The darker shaded region is the area where Theorem 5.3 holds.

Example 7.2. In this example, we choose $\mathbf{u} = \nabla p$ and $f = \Delta p$, where

$$p = \delta(r) r^{\frac{2}{3}} \cos(2\theta/3) \cos(\pi z)$$

with the same smooth cut-off function, $\delta(r)$, defined in Example 7.1. Then, p satisfies a Poisson problem with Neumann boundary condition, that is, (7.1) with $\mathbf{n} \cdot \nabla p = 0$ on $\partial\Omega = \Gamma_N$, and $\mathbf{u} \in H^{\frac{2}{3}}(\Omega)^3$ satisfies (7.2) but with $\mathbf{n} \cdot \mathbf{u} = 0$ on $\partial\Omega$.

Example 7.2 is similar to Example 7.1 but with a different type of boundary condition. The range of β satisfying Theorem 5.3 is the interval $(5/6, 1)$, but numerical tests show that $\beta = 1$ gives the best convergence results. Here, we report the numerical results with $\beta = 1$ (see Table 7.2). With $\beta = 1$, we expect the L^2-, weighted L^2-, and weighted H^1-error convergence rates to be approximately $O(h^{\frac{2}{3}})$, $O(h^{\frac{5}{3}})$, and $O(h^{\frac{2}{3}})$, respectively. Figure 7.2 shows behavior similar to Figure 7.1 when the mesh h moves from 1/64 to 1/128. It shows that the convergence rates follow the theory for awhile after β becomes bigger than 1, but soon the rates break down. As briefly mentioned

TABLE 7.2
Example 2 *with* $\beta = 1$.

	$\|\|\mathbf{u}-\mathbf{u}^h\|\|$		$\|\|\mathbf{u}-\mathbf{u}^h\|\|_{0,\beta}$		$\|\mathbf{u}-\mathbf{u}^h\|_{1,\beta}$	
1/8	2.192e-01	rates	5.197e-02	rates	1.099e+00	rates
1/16	1.122e-01	0.97	1.842e-02	1.50	3.288e-01	1.74
1/32	6.028e-02	0.90	4.986e-03	1.89	1.131e-01	1.54
1/64	3.683e-02	0.71	1.461e-03	1.77	5.425e-02	1.06
1/128	2.310e-02	0.67	4.733e-04	1.63	3.201e-02	0.76

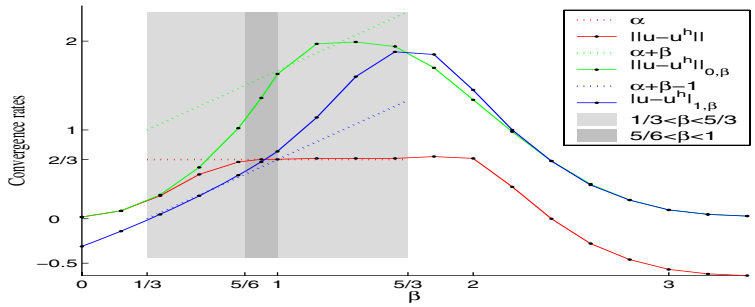


FIG. 7.2. *Finite element convergence rate as a function of* β *for Example* 7.2.

TABLE 7.3
Example 3 *with* $\beta = 1.4$.

	$\|\|\mathbf{u}-\mathbf{u}^h\|\|$		$\|\|\mathbf{u}-\mathbf{u}^h\|\|_{0,\beta}$		$\|\mathbf{u}-\mathbf{u}^h\|_{1,\beta}$	
1/6	3.702E-01	rates	3.332E-02	rates	6.694E-01	rates
1/12	1.168E-01	1.66	1.106E-02	1.59	2.608E-01	1.36
1/24	6.542E-02	0.84	3.553E-03	1.64	7.157E-02	1.87
1/48	4.447E-02	0.56	8.975E-04	1.99	2.213E-02	1.69
1/96	3.125E-02	0.51	2.301E-04	1.96	8.405E-03	1.40
1/192	2.204E-02	0.50	6.179E-05	1.90	3.975E-03	1.08

in Remark 3.2, the restriction on β is sufficient, and Figure 7.2 shows the possibility that the bounds hold beyond the restriction. However, again, we do not address that issue here.

As the last example, we introduce mixed boundary conditions and the domain $\Omega = (-0.5, 0.5) \times (0, 0.5) \times (0, 0.5)$.

Example 7.3. Let $\mathbf{u} = \nabla p$ and $f = \Delta p$, where p has the form

$$p = \delta(r) r^{\frac{1}{2}} \sin(\theta/2) \sin(2\pi z).$$

Then, $\mathbf{u} \in H^{\frac{1}{2}}(\Omega)^3$ satisfies (7.2) and boundary conditions

$$\mathbf{n} \cdot \mathbf{u} = 0 \quad \text{on } \Gamma_N = \{(r, \theta, z) \in \partial\Omega : \theta = \pi\},$$
$$\mathbf{n} \times \mathbf{u} = \mathbf{0} \quad \text{on } \Gamma_D = \partial\Omega \backslash \Gamma_N.$$

In Example 7.3, different types of boundary conditions meet at an edge with inner angle π. Still, boundary singularity occurs along this edge, since the internal angle is bigger than $\pi/2$. In this case, the condition $1 < \beta < 3/2$ guarantees that Theorem 5.3 holds. According to Theorem 5.3, asymptotic convergence rates are $O(h^{\frac{1}{2}})$ for L^2-error, $O(h^{\frac{1}{2}+\beta})$ for weighted L^2-error, and $O(h^{\beta-\frac{1}{2}})$ for weighted H^1-error. Table 7.3 shows the error convergence in L^2-, weighted L^2-, and weighted H^1-norms with

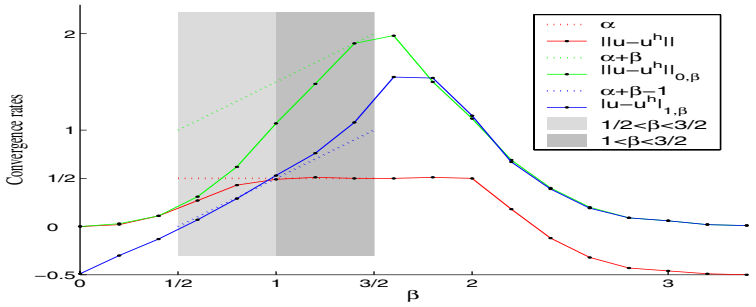


FIG. 7.3. *Finite element convergence rate as a function of* β *for Example* 7.3.

$\beta = 1.4$, where $3/2 < 1/2 + \beta$. In Figure 7.3, the results with more β values when mesh size decreases from 1/96 to 1/192 are presented.

8. Conclusion. Low regularity of the solution caused by singularities on the boundary usually precludes the use of H^1-conforming finite elements in the standard L^2 FOSLS functional. To overcome these difficulties, we use weighted norms in the least-squares functional. This is based on the idea of unweighting the area where the singularity occurs by multiplying by the weight functions. In this paper, we showed that minimization of the least-squares functional in the weighted norm allows us to approximately solve the div/curl system in H^1-conforming finite element spaces with optimal error convergence in L^2-, weighted L^2-, and weighted H^1-norms without using graded mesh refinement. Here, we focused on the case when we have one singular edge. However, in the case when we have several singular edges, we can have the weight function be multi-indexed with r^β at any edge, giving a vector $\beta = (\beta_1, \beta_2, \ldots, \beta_m)$, for a domain with m edges with each of them having inner angle $> \pi/2$. In this paper, we assumed an appropriate weight at each edge and assumed that the analysis is simpler to consider as a "local" result, the way we do to generalize.

REFERENCES

[1] C. AMROUCHE, C. BERNARDI, M. DAUGE, AND V. GIRAULT, *Vector potentials in three-dimensional nonsmooth domains*, Math. Methods Appl. Sci., 21 (1998), pp. 823–864.

[2] TH. APEL AND B. HEINRICH, *Mesh refinement and windowing near edges for some elliptic problem*, SIAM J. Numer. Anal., 31 (1994), pp. 695–708.

[3] TH. APEL AND S. NICAISE, *The finite element method with anisotropic mesh grading for elliptic problems in domains with corners and edges*, Math. Methods Appl. Sci., 21 (1998), pp. 519–549.

[4] I. BABUŠKA, R. B. KELLOGG, AND J. PITKÄRANTA, *Direct and inverse error estimates for finite elements with mesh refinements*, Numer. Math., 33 (1979), pp. 447–471.

[5] M. BERNDT, T. A. MANTEUFFEL, S. F. MCCORMICK, AND G. STARKE, *Analysis of first-order system least squares (FOSLS) for elliptic problems with discontinuous coefficients: Part* I, SIAM J. Numer. Anal., 43 (2005), pp. 386–408.

[6] M. BERNDT, T. A. MANTEUFFEL, AND S. F. MCCORMICK, *Analysis of first-order system least squares (FOSLS) for elliptic problems with discontinuous coefficients: Part* II, SIAM J. Numer. Anal., 43 (2005), pp. 409–436.

[7] P. BOCHEV AND M. GUNZBURGER, *On least-squares finite element methods for the Poisson equation and their connection to the Dirichlet and Kelvin principles*, SIAM J. Numer. Anal., 43 (2005), pp. 340–362.

[8] J. H. BRAMBLE, T. V. KOLEV, AND J. E. PASCIAK, *A least-squares approximation method for the time-harmonic Maxwell equations*, J. Numer. Math., 13 (2005), pp. 237–263.

[9] Z. Cai, R. Lazarov, T. A. Manteuffel, and S. F. McCormick, *First-order system least squares for second-order partial differential equations: Part* I, SIAM J. Numer. Anal., 31 (1994), pp. 1785–1799.

[10] Z. Cai, T. A. Manteuffel, and S. F. McCormick, *First-order system least squares for second-order partial differential equations: Part* II, SIAM J. Numer. Anal., 34 (1997), pp. 425–454.

[11] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, Classics Appl. Math. 40, SIAM, Philadelphia, 2002.

[12] M. Costabel, *A coercive bilinear form for Maxwell's equations*, J. Math. Anal. Appl., 157 (1991), pp. 527–541.

[13] M. Costabel and M. Dauge, *Weighted regularization of Maxwell equations in polyhedral domains. A rehabilitation of nodal finite elements*, Numer. Math., 93 (2002), pp. 239–277.

[14] G. J. Fix, S. Gulati, and G. I. Wakoff, *On the use of singular functions with finite element approximations*, J. Comput. Phys., 13 (1973), pp. 209–228.

[15] G. J. Fix, M. D. Gunzburger, and R. A. Nicolaides, *On mixed finite element methods for first-order elliptic systems*, Numer. Math., 37 (1981), pp. 29–48.

[16] V. Girault and P.-A. Raviart, *Finite Element Methods for Navier-Stokes Equations. Theory and Algorithms*, Springer Ser. Comput. Math. 5, Springer-Verlag, Berlin, 1986.

[17] P. Grisvard, *Singularities in Boundary Value Problems*, Rech. Math. Appl. 22, Masson, Paris, Springer-Verlag, Berlin, 1992.

[18] B. Jiang, *The Least-Squares Finite Element Method. Theory and Applications in Computational Fluid Dynamics and Electromagnetics*, Sci. Comput., Springer-Verlag, Berlin, 1998.

[19] V. A. Kondraťev, *Boundary value problems for elliptic equations in domains with conical or angular points*, Trudy Moskov. Mat. Obšč., 16 (1967), pp. 209–292.

[20] V. A. Kondraťev, *The smoothness of a solution of Dirichlet's problem for second-order elliptic equations in a region with a piecewise-smooth boundary*, Differential Equations, 6 (1970), pp. 1392–1401.

[21] E. Lee, *FOSLL* for Eddy Current Problems with Three-Dimensional Edge Singularities*, Ph.D. thesis, University of Colorado, Boulder, CO, 2005.

[22] E. Lee and T. A. Manteuffel, *FOSLL* method for the eddy current problem with three-dimensional edge singularities*, SIAM J. Numer. Anal., 45 (2007), pp. 787–809.

[23] E. Lee, T. A. Manteuffel, and C. R. Westphal, *Weighted-norm first-order system least squares (FOSLS) for problems with corner singularities*, SIAM J. Numer. Anal., 44 (2006), pp. 1974–1996.

[24] T. A. Manteuffel, S. F. McCormick, J. Ruge, and J. G. Schmidt, *First-order system $\mathcal{LL}^*$ (FOSLL*) for general scalar elliptic problems in the plane*, SIAM J. Numer. Anal., 43 (2005), pp. 2098–2120.

[25] V. G. Maźja and B. A. Plamenevskiĭ, *L_p-estimates of solutions of elliptic boundary value problems in domains with ribs*, Trudy Moskov. Mat. Obshch., 37 (1978), pp. 49–93, 270.

[26] V. Maźya, S. Nazarov, and B. Plamenevskij, *Asymptotic Theory of Elliptic Boundary Value Problems in Singularly Perturbed Domains*, Vol. II, Oper. Theory Adv. Appl. 112, Birkhäuser Verlag, Basel, 2000.

[27] V. G. Maźya and B. A. Plamenevskii, *Problems with oblique derivatives in regions with piecewise smooth boundaries*, Funct. Anal. Appl., 5 (1971), pp. 256–258.

[28] V. G. Maźya and J. Rossmann, *Weighted L_p estimates of solutions to boundary value problems for second order elliptic systems in polyhedral domains*, ZAMM Z. Angew. Math. Mech., 83 (2003), pp. 435–467.

[29] S. Nicaise, *Edge elements on anisotropic meshes and approximation of the Maxwell equations*, SIAM J. Numer. Anal., 39 (2001), pp. 784–816.

[30] J. Ruge, *FOSPACK: A First-Order System Least-Squares (FOSLS) Code*.

SIAM J. NUMER. ANAL.
Vol. 46, No. 3, pp. 1640–1651
© 2008 Society for Industrial and Applied Mathematics

A WEIGHTED $H(\mathrm{div})$ LEAST-SQUARES METHOD FOR SECOND-ORDER ELLIPTIC PROBLEMS*

Z. CAI† AND C. R. WESTPHAL‡

Abstract. This paper presents analysis of a weighted-norm least squares finite element method for elliptic problems with boundary singularities. We use $H(\mathrm{div})$ conforming Raviart–Thomas elements and continuous piecewise polynomial elements. With only a rough estimate of the power of the singularity, we employ a simple, locally weighted L^2 norm to eliminate the pollution effect and recover better rates of convergence. Theoretical results are carried out in weighted Sobolev spaces and include ellipticity bounds of the homogeneous least-squares functional, new weighted Raviart–Thomas interpolation results, and error estimates in both weighted and nonweighted norms. Numerical tests are given to confirm the theoretical estimates and to illustrate the practicality of the method

Key words. boundary singularities, least squares, finite element methods, weighted Sobolev spaces

AMS subject classifications. 65N30, 65N12, 35B65

DOI. 10.1137/070698531

1. Introduction.

The loss of regularity of second-order partial differential equations at corners/edges of polygonal/polyhedral domains is well known. Numerical methods for approximating solutions to these problems often suffer either reduced rates of accuracy or loss of convergence due to the nonsmooth solutions. For finite element methods, the rate of convergence generally depends on both the smoothness of the solution and the approximation properties of the finite element space used. When the rates are diminished due to these boundary singularities, standard rates of convergence can only be recovered with care.

In this paper we present a weighted-norm least-squares method for scalar elliptic problems of the form

$$\begin{cases} -\nabla \cdot A\nabla u + \mathbf{b} \cdot \nabla u + cu = f & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega, \end{cases} \tag{1.1}$$

where A is symmetric, positive definite and $\mathbf{b}$ and c are smooth, bounded functions. For the analysis in this paper we assume homogeneous Dirichlet boundary conditions, though the method may easily be applied to other boundary conditions. It is well known that if Ω is a polygonal/polyhedral domain in $\mathbb{R}^2/\mathbb{R}^3$, then the regularity of system (1.1) may be limited by the geometry of Ω. To illustrate, assume that $\Omega \subseteq \mathbb{R}^2$ has one reentrant corner, which has interior angle ω, and (r, θ) are polar coordinates centered at the corner. For $f \in L^2(\Omega)$, solutions to (1.1) may be decomposed by

$$u = \tilde{u} + r^{\lambda} w,$$

*Received by the editors July 30, 2007; accepted for publication (in revised form) December 21, 2007; published electronically April 11, 2008. This work was sponsored by the National Science Foundation under grant DMS–0511430.

http://www.siam.org/journals/sinum/46-3/69853.html

†Department of Mathematics, Purdue University, West Lafayette, IN 47907 (zcai@math.purdue.edu).

‡Department of Mathematics and Computer Science, Wabash College, P.O. Box 352, Crawfordsville, IN 47933 (westphac@wabash.edu).

where $\tilde{u} \in H^2(\Omega)$ is the smooth component of u, and $r^\lambda w \in H^{\alpha+1}(\Omega)$ for $\alpha < \lambda$ is the singular component. The function $w = w(\theta)$ and parameter α are characteristic of the operator and the domain. For Poisson's equation the singular functions are multiples of $r^{\pi/\omega}\sin(\pi\theta/\omega)$. Edge singularities in $\mathbb{R}^3$ may be considered analogously, where $w = w(\theta, z)$.

The least-squares finite element approach for a variety of problems affords several unique features. Among these is the freedom to independently choose finite element spaces for different unknowns. In [5], system (1.1) is reformulated as a first-order system by introducing the flux variable $\boldsymbol{\sigma} = -A\nabla u$, and $(u, \boldsymbol{\sigma}) \in H^1 \times H(\mathrm{div})$ are approximated using piecewise polynomial and $H(\mathrm{div})$-conforming Raviart–Thomas elements. See also [4, 9] for results concerning least-squares minimization of this system. For polygonal/polyhedral domains this method results in $O(h^\alpha)$ approximations, even for smooth data or higher-order approximation spaces. Thus, for problems with $\alpha < 1$, the convergence is suboptimal. Moreover, the reduction in convergence rate is generally a global effect—that is, convergence throughout the entire domain is suboptimal. This well-known *pollution effect* is also seen in Galerkin discretizations as well as in div-curl least-squares formulations. We overcome this difficulty by minimizing an appropriately weighted functional. Our approach eliminates the pollution effect with only a trivial modification of the equations, requires only crude a priori knowledge of the power of the singularity (not the singular function itself), and can be used in conjunction with adaptive mesh refinement. A least-squares formulation provides a natural error estimator that can be used in an adaptive mesh refinement strategy. We note that, while many problems are effectively treated with only mesh refinement near boundary singularities, some problems admit solutions much less smooth than considered here. If the approximation space used is not dense in the solution space, then global errors may persist, despite any amount of mesh refinement. In these cases, a weighted-norm method in the spirit of this paper can be effectively used to recover convergence, allowing mesh refinement to be much less aggressive.

In [6], Cox and Fix use a weighted-norm approach similar to our approach, but achieve optimal convergence with continuous P_1 finite elements and specialized graded meshes. In [10, 11], the weighted-norm least-squares approach is presented for div-curl functionals in two and three dimensions. An aggressive weighting is used to recover optimal convergence without relying on mesh refinement. In this paper, using mixed affine finite element spaces allows a less aggressive weighting to achieve optimal results, even with uniform meshes.

The organization of this paper is as follows. We define notation and detail the problem formulation in section 2. In section 3, we show equivalence of the homogeneous least-squares functional to an appropriate weighted norm, present new interpolation bounds for Raviart–Thomas finite element spaces in weighted norms, and show the resulting error bounds. Numerical results are given in section 4 that confirm the theoretical error estimates and demonstrate the practicality of the method.

2. Problem formulation. We use standard notation for the L^2 norm, $\|\cdot\|$, and inner product, $\langle\cdot,\cdot\rangle$, and use $\|\cdot\|_{k;\Omega}$ to denote the norm corresponding to the Sobolev space $H^k(\Omega)^d$, omitting subscript Ω and superscript d when the domain and dimension are clear by context. For noninteger k, $H^k(\Omega)$ represents the standard interpolation space.

For this paper, we assume that Ω is a bounded domain in $\mathbb{R}^d$, $d = 2, 3$, with one boundary corner in $\mathbb{R}^2$ or one boundary edge in $\mathbb{R}^3$ that admits a singularity with $\alpha < 1$. Assume further that this singular point is located at the origin for $d = 2$ or

along the z-axis for $d = 3$, and denote this part of $\partial\Omega$ by $\mathcal{E}$. For simplicity, we consider only one singular point although techniques here may be applied locally to more than one singular point. In section 4 we discuss practical implementation in such cases.

Define the weighted Sobolev norm by

$$\|v\|_{k,\beta;\Omega} = \left(\sum_{|j|\le k} \|r^{\beta-k+|j|} D^j v\|_{0;\Omega}^2 \right)^{1/2}$$

and seminorm by

$$|v|_{k,\beta;\Omega} = \left(\sum_{|j|=k} \|r^{\beta} D^j v\|_{0;\Omega}^2 \right)^{1/2}.$$

For $k = 0$, 1, we have

$$\|v\|_{0,\beta;\Omega} = \|r^{\beta} v\|_{0;\Omega}, \quad \|v\|_{1,\beta;\Omega} = \left(\|r^{\beta}\nabla v\|_{0;\Omega}^2 + \|r^{\beta-1} v\|_{0;\Omega}^2 \right)^{1/2},$$
$$\text{and} \quad |v|_{1,\beta;\Omega} = \|r^{\beta}\nabla v\|_{0;\Omega}.$$

Then the weighted Sobolev space $H_{\beta}^{k}(\Omega)$ is defined by

$$H_{\beta}^{k}(\Omega) = \{v \in L^2(\Omega) : \|v\|_{k,\beta;\Omega} < +\infty\}.$$

The dual spaces are defined by duality in the natural way. Let

$$D_{-\beta}^{1} = \{v \in H_{-\beta}^{1}(\Omega) : v = 0 \text{ on } \partial\Omega\},$$

and define its dual, $H_{\beta}^{-1}(\Omega)$, by the norm

$$\|v\|_{-1,\beta} = \sup_{\phi \in D_{-\beta}^{1}} \frac{\langle v, \phi\rangle}{\|\phi\|_{1,-\beta}}.$$

By introducing the new unknown $\boldsymbol{\sigma} = -A\nabla u$ and denoting the lower-order terms by $Xu = \mathbf{b}\cdot\nabla u + cu$, we may expand system (1.1) to the first-order system

$$\begin{cases} \boldsymbol{\sigma} + A\nabla u = 0 & \text{in } \Omega, \\ \nabla\cdot\boldsymbol{\sigma} + Xu = f & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega. \end{cases} \tag{2.1}$$

In this paper, we consider the weighted least-squares functional

$$G(u, \boldsymbol{\sigma}; f) = \|A^{-1/2}(\boldsymbol{\sigma} + A\nabla u)\|_{0,\beta_1}^2 + \|\nabla\cdot\boldsymbol{\sigma} + Xu - f\|_{0,\beta_2}^2 \tag{2.2}$$

for functions $(u, \boldsymbol{\sigma}) \in \mathcal{V}\times\mathcal{W}$, where

$$\mathcal{V} = \{v \in H_{\beta_1}^{1}(\Omega) : v = 0 \text{ on } \partial\Omega\}$$

and

$$\mathcal{W} = \{\boldsymbol{\tau} : r^{\beta_1}\boldsymbol{\tau} \in L^2(\Omega),\ r^{\beta_2}\nabla\cdot\boldsymbol{\tau} \in L^2(\Omega)\}.$$

The nonweighted functional (i.e., $\beta_1 = \beta_2 = 0$) yields less-than-optimal global discretization rates for problems with $\alpha < 1$. We thus investigate appropriate values for β_1 and β_2 to recover better rates of convergence and eliminate the pollution effect.

The least-squares minimization problem is to find $(u, \boldsymbol{\sigma}) \in \mathcal{V} \times \mathcal{W}$ such that

$$G(u, \boldsymbol{\sigma}; f) = \inf_{(v,\boldsymbol{\tau}) \in \mathcal{V} \times \mathcal{W}} G(v, \boldsymbol{\tau}; f).$$

It may be noted that the minimization above is equivalent to the symmetric variational problem of finding $(u, \boldsymbol{\sigma}) \in \mathcal{V} \times \mathcal{W}$ such that

$$b(u, \boldsymbol{\sigma}; v, \boldsymbol{\tau}) = f(v, \boldsymbol{\tau}) \qquad \forall\ (v, \boldsymbol{\tau}) \in \mathcal{V} \times \mathcal{W},$$

where the bilinear and linear forms are given by

$$b(u, \boldsymbol{\sigma}; v, \boldsymbol{\tau}) := \langle r^{2\beta_1} A^{-1}(\boldsymbol{\sigma} + A\nabla u), \boldsymbol{\tau} + A\nabla v \rangle + \langle r^{2\beta_2}(\nabla \cdot \boldsymbol{\sigma} + Xu), \nabla \cdot \boldsymbol{\tau} + Xv \rangle$$

and

$$f(v, \boldsymbol{\tau}) := \langle r^{2\beta_2} f, \nabla \cdot \boldsymbol{\tau} + Xv \rangle.$$

Let $\mathcal{T}_h$ be a regular affine triangulation of Ω with a meshsize of $O(h)$, where on element K we denote $h_K = diam(K)$. Denote by $P_k(K)$ the standard space of polynomials of degree $\leq k$ on element K, and consider the Raviart–Thomas space of degree k on element K defined by

$$RT_k(K) = P_k(K)^d + \mathbf{x} P_k(K),$$

where $\mathbf{x} = (x_1, \ldots, x_d)^t$. For this paper we define the spaces

$$\mathcal{V}^h = \{v \in \mathcal{V} \cap C^0(\Omega) : v|_K \in P_1(K)\ \forall\ K \in \mathcal{T}_h\},$$

$$\mathcal{W}^h = \{\boldsymbol{\tau} \in \mathcal{W} : \boldsymbol{\tau}|_K \in RT_0(K)\ \forall\ K \in \mathcal{T}_h\}.$$

The discrete minimization problem is to find $(u^h, \boldsymbol{\sigma}^h) \in \mathcal{V}^h \times \mathcal{W}^h$ such that

$$G(u^h, \boldsymbol{\sigma}^h; f) = \min_{(v^h,\boldsymbol{\tau}^h) \in \mathcal{V}^h \times \mathcal{W}^h} G(v^h, \boldsymbol{\tau}^h; f).$$

This leads to a symmetric, positive definite linear system of equations that can be efficiently solved by standard multigrid methods.

3. Theory. In the following analysis, we assume that A is a $d \times d$ matrix of functions in $L^2(\Omega)$ and that $\mathbf{b}$ and c are smooth, bounded functions. We further assume that A is uniformly symmetric positive definite; that is, there exist positive constants λ_1 and λ_2 such that

$$\lambda_1 \boldsymbol{\xi}^T \boldsymbol{\xi} \leq \boldsymbol{\xi}^T A \boldsymbol{\xi} \leq \lambda_2 \boldsymbol{\xi}^T \boldsymbol{\xi} \tag{3.1}$$

for all $\boldsymbol{\xi} \in \mathbb{R}^d$ and almost all $\boldsymbol{\xi} \in \overline{\Omega}$. Under an ellipticity assumption on the operator for problem (1.1), a general regularity result for second-order problems is established in [1]. We state here an H^1_β estimate sufficient for our purposes.

THEOREM 3.1. *Assume that the operator $L := -\nabla \cdot A\nabla + X$ satisfies the assumptions of Theorem* 5.2 *in* [1] *and that $f \in H^{-1}_\beta(\Omega)$. Then there exists some $\alpha > 0$ such that problem* (1.1) *admits the estimate*

$$\|u\|_{1,\beta} \leq C \|f\|_{-1,\beta}$$

for any $|\beta| < \alpha$. In general, α depends on the operator and domain. If $L = -\Delta$, then $\alpha = \pi/\omega$, where ω is the interior angle at $\mathcal{E}$.

Proof. See Theorem 5.4 in [1]. ∎

We note that values for α are considered in the classic works [7, 8]. In [12], values for α are explicitly shown to satisfy $\alpha \geq 1/2$ for the Stokes and elasticity systems. In general, the value of α for Poisson's equation serves as an upper bound for second-order problems on similar geometries. While the method described in this paper is based on the value of α for the operator of the problem, we note that the exact value of α need not be known for the method to be effective.

Let $\mathcal{I}^h$ be the standard interpolation operator onto $P_k(K)$, where, for all $v \in H^m(K)$,

$$\|v - \mathcal{I}^h v\|_{s;K} \leq C h_K^{m-s} |v|_{m;K} \tag{3.2}$$

holds for $0 \leq s \leq m$ and $1 < m \leq k+1$. A proof of this classical result can be found in [2].

Define the modified interpolation operator, $\mathcal{I}_0^h$, by

$$\mathcal{I}_0^h v|_K = \begin{cases} \mathcal{I}^h v|_K & \text{if } \overline{K} \cap \mathcal{E} = \emptyset, \\ \displaystyle\sum_{a_i \notin \mathcal{E}} v(a_i)\phi_i & \text{if } \overline{K} \cap \mathcal{E} \neq \emptyset, \end{cases}$$

where a_i are the nodal points corresponding to the basis functions, ϕ_i. Thus, the modified interpolation has a value of zero at $\mathcal{E}$ and is identical to $\mathcal{I}^h$ away from $\mathcal{E}$. From [10, 11], if $\beta > 0$, then $\mathcal{I}_0^h$ satisfies

$$\sum_{K \in \mathcal{T}_h} \|v - \mathcal{I}_0^h v\|_{1,\beta;K}^2 \leq C h^2 \|v\|_{2,\beta}^2 \tag{3.3}$$

for $v \in H_\beta^2(\Omega)$.

Let I^h be the standard interpolation operator from $H(\mathrm{div}) \cap L^p(\Omega)$ onto RT_k with $p > 2$. Then the following approximation properties of RT_k hold (see, e.g., [3]) for any $\boldsymbol{\tau} \in (H^m(K))^d$ with $1 \leq m \leq k+1$:

$$\|\boldsymbol{\tau} - I^h \boldsymbol{\tau}\|_{s;K} \leq C h_K^{m-s} |\boldsymbol{\tau}|_{m;K} \quad \text{with} \quad s = 0,\ 1, \tag{3.4}$$

and

$$\|\nabla \cdot (\boldsymbol{\tau} - I^h \boldsymbol{\tau})\|_{0;K} \leq C h_K^m |\nabla \cdot \boldsymbol{\tau}|_{m;K} \tag{3.5}$$

if $\nabla \cdot \boldsymbol{\tau} \in H^m(K)$. For any $\boldsymbol{\tau} \in (H^\alpha(K))^d$ and $\nabla \cdot \boldsymbol{\tau} \in H^\alpha(K)$ with $\alpha < 1$, we assume the following approximation properties:

$$\|\boldsymbol{\tau} - I^h \boldsymbol{\tau}\| \leq C h^{\tilde{\alpha}} |\boldsymbol{\tau}|_\alpha \tag{3.6}$$

and

$$\|\nabla \cdot (\boldsymbol{\tau} - I^h \boldsymbol{\tau})\| \leq C h^{\tilde{\alpha}} |\nabla \cdot \boldsymbol{\tau}|_\alpha \tag{3.7}$$

hold for some constant $\tilde{\alpha} \in (0, \alpha]$.

THEOREM 3.2. *Assume that $\nabla \cdot \boldsymbol{\tau}$ and each $\boldsymbol{\tau}_j$ are in $H_\beta^1(\Omega) \cap H^\alpha(\Omega)$. Then for any $\beta \geq 0$, there exists a positive constant C such that*

$$\|\boldsymbol{\tau} - I^h \boldsymbol{\tau}\|_{0,\beta} \leq C h^{m_1} \left(|\boldsymbol{\tau}|_{1,\beta} + |\boldsymbol{\tau}|_\alpha\right) \tag{3.8}$$

and that

$$\|\nabla\cdot(\boldsymbol{\tau}-I^h\boldsymbol{\tau})\|_{0,\beta}\le Ch^{m_2}\left(|\nabla\cdot\boldsymbol{\tau}|_{1,\beta}+|\nabla\cdot\boldsymbol{\tau}|_{\alpha}\right), \tag{3.9}$$

where $m_1=\min\{1,\beta+\tilde{\alpha}\}$, $m_2=\min\{1,\beta+\alpha\}$, *and* C *depends only on* k, β, *and the shape of* Ω.

Proof. We first prove the bound in (3.8). Note that (3.4) and (3.5) apply since $\boldsymbol{\tau}$ is smooth enough to be in the domain of I^h. Denote the set of elements adjacent to $\mathcal{E}$ by

$$\mathcal{K}_0=\{K\in\mathcal{T}_h:\overline{K}\cap\mathcal{E}\neq\emptyset\}.$$

For any element $K\in\mathcal{T}_h\backslash\mathcal{K}_0$, we note that $\boldsymbol{\tau}\in H^1(K)$ and that

$$\frac{1}{c_K}h_K\le r_{\min}\le r\le r_{\max}\le r_{\min}+c_Kh_K, \tag{3.10}$$

where $r_{\min}$ and $r_{\max}$ are the minimum and maximum distances from K to $\mathcal{E}$, respectively, and c_K depends only on d and the shape of K (e.g., $c_K=\sqrt{d}$ for a uniform mesh). Inequality (3.10) implies

$$\frac{r_{\max}}{r_{\min}}\le 1+c_K^2.$$

Now, it follows from (3.10) and (3.4) with $m=1$ and $s=0$ that

$$\begin{aligned}\|\boldsymbol{\tau}-I^h\boldsymbol{\tau}\|_{0,\beta;K}&=\|r^{\beta}(\boldsymbol{\tau}-I^h\boldsymbol{\tau})\|_{0;K}\le r_{\max}^{\beta}\|\boldsymbol{\tau}-I^h\boldsymbol{\tau}\|_{0;K}\\&\le Cr_{\max}^{\beta}h_K|\boldsymbol{\tau}|_{1;K}\le Cr_{\max}^{\beta}h_K\|(r/r_{\min})^{\beta}\nabla\boldsymbol{\tau}\|_{0;K}\\&=C\left(\frac{r_{\max}}{r_{\min}}\right)^{\beta}h_K|\boldsymbol{\tau}|_{1,\beta;K}=Ch_K|\boldsymbol{\tau}|_{1,\beta;K}.\end{aligned} \tag{3.11}$$

For any element $K\in\mathcal{K}_0$, it follows from the fact that $r\le h_K$ and (3.4) that

$$\begin{aligned}\|\boldsymbol{\tau}-I^h\boldsymbol{\tau}\|_{0,\beta;K}&=\|r^{\beta}(\boldsymbol{\tau}-I^h\boldsymbol{\tau})\|_{0;K}\le h_K^{\beta}\|\boldsymbol{\tau}-I^h\boldsymbol{\tau}\|_{0;K}\\&\le Ch_K^{\beta+\tilde{\alpha}}|\boldsymbol{\tau}|_{\alpha;K},\end{aligned}$$

which, together with (3.11), yields the bound in (3.8).

To prove (3.9), we first consider $K\in\mathcal{T}_h\backslash\mathcal{K}_0$. A calculation identical to (3.11) using (3.5) with $m=1$ gives

$$\|\nabla\cdot(\boldsymbol{\tau}-I^h\boldsymbol{\tau})\|_{0,\beta;K}\le Ch_K|\nabla\cdot\boldsymbol{\tau}|_{1,\beta;K}. \tag{3.12}$$

Now consider $K\in\mathcal{K}_0$. Let Π^h be the L^2-projection onto discontinuous piecewise polynomial of degree k. Working with affine elements allows the use of the well-known commutativity property,

$$\nabla\cdot(I^h\boldsymbol{\tau})=\Pi^h\nabla\cdot\boldsymbol{\tau}.$$

This, (3.5), and the approximation property of Π^h give

$$\begin{aligned}\|\nabla\cdot(\boldsymbol{\tau}-I^h\boldsymbol{\tau})\|_{0,\beta;K}&=\|(I-\Pi^h)\nabla\cdot\boldsymbol{\tau}\|_{0,\beta;K}\\&\le h_K^{\beta}\|(I-\Pi^h)\nabla\cdot\boldsymbol{\tau}\|_{0;K}\\&\le Ch_K^{\beta+\alpha}|\nabla\cdot\boldsymbol{\tau}|_{\alpha;K}.\end{aligned}$$

This and (3.12) lead to the bound in (3.9). The proof of the theorem is now complete. ∎

LEMMA 3.3. *If $\beta \le \gamma$, then*

$$\|\phi\|_{0,\gamma} \le C\,\|\phi\|_{0,\beta} \quad \forall\ \phi \in H^0_\beta(\Omega). \tag{3.13}$$

Proof. Let $\Omega_1 = \Omega \cap \{r < 1\}$. For $r < 1$, since $r^\gamma \le r^\beta$, we then have

$$\|\phi\|_{0,\gamma;\Omega_1} = \|r^\gamma \phi\|_{0;\Omega_1} \le \|r^\beta \phi\|_{0;\Omega_1} = \|\phi\|_{0,\beta;\Omega_1}.$$

Now, (3.13) is a direct consequence of the fact that $\|\phi\|_{0,\gamma;\Omega_2} \le C\,\|\phi\|_{0,\beta;\Omega_2}$, where $\Omega_2 = \Omega \setminus \overline{\Omega}_1$. ∎

The following theorem establishes equivalence of $G(u,\boldsymbol{\sigma};0)^{1/2}$ to the norm

$$|||(u,\boldsymbol{\sigma})|||_{\beta_1,\beta_2} := \left(\|u\|^2_{1,\beta_1} + \|\boldsymbol{\sigma}\|^2_{0,\beta_1} + \|\nabla\cdot\boldsymbol{\sigma}\|^2_{0,\beta_2}\right)^{1/2}.$$

Since $G(u^h,\boldsymbol{\sigma}^h;f) = G(u-u^h,\boldsymbol{\sigma}-\boldsymbol{\sigma}^h;0)$, this equivalence implies that minimizing the functional is equivalent to minimizing the error in this natural norm.

THEOREM 3.4. *For $\beta_1 \le \beta_2 \le \beta_1 + 1$ and $|\beta_1| < \alpha$, there exist positive constants c_0 and c_1 such that*

$$c_0|||(u,\boldsymbol{\sigma})|||^2_{\beta_1,\beta_2} \le G(u,\boldsymbol{\sigma};0) \le c_1|||(u,\boldsymbol{\sigma})|||^2_{\beta_1,\beta_2} \tag{3.14}$$

for all $(u,\boldsymbol{\sigma}) \in \mathcal{V}\times\mathcal{W}$.

Proof. We first recall the definition of the homogeneous functional

$$G(u,\boldsymbol{\sigma};0) = \|A^{-1/2}(\boldsymbol{\sigma} + A\nabla u)\|^2_{0,\beta_1} + \|\nabla\cdot\boldsymbol{\sigma} + Xu\|^2_{0,\beta_2}.$$

Since $\beta_1 - 1 < \beta_1 \le \beta_2$, (3.13) with the respective $\phi = \nabla u$, $\beta = \beta_1$, $\gamma = \beta_2$ and $\phi = u$, $\beta = \beta_1 - 1$, $\gamma = \beta_2$ implies

$$\|\nabla u\|_{0,\beta_2} + \|u\|_{0,\beta_2} \le C\ (\|\nabla u\|_{0,\beta_1} + \|u\|_{0,\beta_1-1}) = C\,\|u\|_{1,\beta_1}. \tag{3.15}$$

These, together with the triangle inequality and the assumptions on A and X, prove the validity of the upper bound in (3.14).

For the lower bound in (3.14), since $|\beta_1| < \alpha$, Theorem 3.1 with $\beta = \beta_1$ and the triangle inequality gives

$$\begin{aligned} \|u\|_{1,\beta_1} &\le C\|-\nabla\cdot A\nabla u + Xu\|_{-1,\beta_1} \\ &\le C(\|\nabla\cdot\boldsymbol{\sigma} + \nabla\cdot A\nabla u\|_{-1,\beta_1} + \|\nabla\cdot\boldsymbol{\sigma} + Xu\|_{-1,\beta_1}). \end{aligned} \tag{3.16}$$

We consider each term on the right side above separately. For the first term, integration by parts and the Cauchy–Schwarz inequality give

$$\begin{aligned} \|\nabla\cdot\boldsymbol{\sigma} + \nabla\cdot A\nabla u\|_{-1,\beta_1} &= \sup_{\phi\in D^1_{-\beta_1}} \frac{\langle\nabla\cdot(\boldsymbol{\sigma} + A\nabla u),\phi\rangle}{\|\phi\|_{1,-\beta_1}} \\ &= \sup_{\phi\in D^1_{-\beta_1}} \frac{\langle -r^{\beta_1}(\boldsymbol{\sigma} + A\nabla u), r^{-\beta_1}\nabla\phi\rangle}{\|\phi\|_{1,-\beta_1}} \\ &\le \|r^{\beta_1}(\boldsymbol{\sigma} + A\nabla u)\| \sup_{\phi\in D^1_{-\beta_1}} \frac{\|\nabla\phi\|_{0,-\beta_1}}{\|\phi\|_{1,-\beta_1}} \\ &\le \|\boldsymbol{\sigma} + A\nabla u\|_{0,\beta_1}. \end{aligned} \tag{3.17}$$

Since $\beta_2 \le \beta_1 + 1$, (3.13) with $\beta = -(\beta_1+1)$ and $\gamma = -\beta_2$ gives

$$\|\phi\|_{0,-\beta_2} \le C\,\|\phi\|_{0,-(\beta_1+1)} = C\,\|r^{-\beta_1-1}\phi\| \le C\,\|\phi\|_{1,-\beta_1}. \tag{3.18}$$

Now, the second term may be bounded similarly using the Cauchy–Schwarz inequality and (3.18),

$$\begin{aligned}
\|\nabla\cdot\boldsymbol{\sigma} + Xu\|_{-1,\beta_1} &= \sup_{\phi\in D^1_{-\beta_1}} \frac{\langle \nabla\cdot\boldsymbol{\sigma} + Xu, \phi\rangle}{\|\phi\|_{1,-\beta_1}} \\
&= \sup_{\phi\in D^1_{-\beta_1}} \frac{\langle r^{\beta_2}(\nabla\cdot\boldsymbol{\sigma} + Xu), r^{-\beta_2}\phi\rangle}{\|\phi\|_{1,-\beta_1}} \\
&\le \|\nabla\cdot\boldsymbol{\sigma} + Xu\|_{0,\beta_2} \sup_{\phi\in D^1_{-\beta_1}} \frac{\|\phi\|_{0,-\beta_2}}{\|\phi\|_{1,-\beta_1}} \\
&\le C\,\|\nabla\cdot\boldsymbol{\sigma} + Xu\|_{0,\beta_2}.
\end{aligned} \tag{3.19}$$

Combining equations (3.16), (3.17), and (3.19) gives

$$\|u\|_{1,\beta_1} \le CG(u,\boldsymbol{\sigma};0)^{1/2}. \tag{3.20}$$

It follows from the triangle inequality, (3.1), (3.20), and (3.15) that

$$\begin{aligned}
&\|\boldsymbol{\sigma}\|_{0,\beta_1} + \|\nabla\cdot\boldsymbol{\sigma}\|_{0,\beta_2} \\
&\le \|\boldsymbol{\sigma} + A\nabla u\|_{0,\beta_1} + \|A\nabla u\|_{0,\beta_1} + \|\nabla\cdot\boldsymbol{\sigma} + Xu\|_{0,\beta_2} + \|Xu\|_{0,\beta_2} \\
&\le CG(u,\boldsymbol{\sigma};0)^{1/2} + C\|u\|_{1,\beta_1} + C(\|\nabla u\|_{0,\beta_2} + \|u\|_{0,\beta_2}) \\
&\le CG(u,\boldsymbol{\sigma};0)^{1/2},
\end{aligned}$$

which, together with (3.20), implies the lower bound in (3.14) and, hence, completes the proof of the theorem. ∎

We may now prove error estimates.

THEOREM 3.5. *Let $(u,\boldsymbol{\sigma}) \in \mathcal{V}\times\mathcal{W}$ be the solution of system (2.1) and let $(u^h,\boldsymbol{\sigma}^h)$ be the minimizer of the weighted functional (2.2) over $\mathcal{V}^h\times\mathcal{W}^h$. Assume that $f\in H^\alpha(\Omega)$. Then for $\beta_1\in[1-\alpha,\alpha)$ and $\beta_2\in[\beta_1,\beta_1+1]$, the following error estimate holds:*

$$|||(u-u^h,\boldsymbol{\sigma}-\boldsymbol{\sigma}^h)|||_{\beta_1,\beta_2} \le Ch^{m_1}\left(\|u\|_{2,\beta_1} + \|u\|_{1+\alpha} + \|f\|_\alpha\right)$$

with $m_1 = \min\{1, \beta_1+\tilde{\alpha}\}$.

Proof. For any $(v,\boldsymbol{\tau})\in\mathcal{V}^h\times\mathcal{W}^h$, it follows from Theorem 3.4 and the orthogonality property that

$$\begin{aligned}
c_0|||(u-u^h,\boldsymbol{\sigma}-\boldsymbol{\sigma}^h)|||^2_{\beta_1,\beta_2} &\le G(u-u^h,\boldsymbol{\sigma}-\boldsymbol{\sigma}^h;0) \\
&= G(u-v^h,\boldsymbol{\sigma}-\boldsymbol{\tau}^h;0) \le c_1|||(u-v^h,\boldsymbol{\sigma}-\boldsymbol{\tau}^h)|||^2_{\beta_1,\beta_2},
\end{aligned}$$

which implies

$$\begin{aligned}
|||(u-u^h,\boldsymbol{\sigma}-\boldsymbol{\sigma}^h)|||^2_{\beta_1,\beta_2} &\le \frac{c_1}{c_0}\inf_{(v,\boldsymbol{\tau})\in\mathcal{V}^h\times\mathcal{W}^h} |||(u-v,\boldsymbol{\sigma}-\boldsymbol{\tau})|||^2_{\beta_1,\beta_2} \\
&\le C\left(\inf_{v\in\mathcal{V}}\|u-v\|^2_{1,\beta_1} + \inf_{\boldsymbol{\tau}\in\mathcal{W}^h}\|\boldsymbol{\sigma}-\boldsymbol{\tau}\|^2_{0,\beta_1} + \inf_{\boldsymbol{\tau}\in\mathcal{W}^h}\|\nabla\cdot(\boldsymbol{\sigma}-\boldsymbol{\tau})\|^2_{0,\beta_2}\right).
\end{aligned}$$

This, with the approximation properties in (3.3), (3.8), and (3.9), results in

$$|||(u-u^h,\boldsymbol{\sigma}-\boldsymbol{\sigma}^h)|||^2_{\beta_1,\beta_2} \le Ch^{2m_1}\left(\|u\|^2_{2,\beta_1}+|\boldsymbol{\sigma}|^2_{1,\beta_1}+|\boldsymbol{\sigma}|^2_{\alpha}+|\nabla\cdot\boldsymbol{\sigma}|^2_{\alpha}\right).$$

Choosing $\beta_1 \ge 1-\alpha$ ensures that $u \in H^2_{\beta_1}(\Omega)$. The terms above involving $\boldsymbol{\sigma}$ each may be bounded as follows:

$$|\boldsymbol{\sigma}|_{1,\beta_1} = |A\nabla u|_{1,\beta_1} \le C\|u\|_{2,\beta_1},$$
$$|\boldsymbol{\sigma}|_{\alpha} = |A\nabla u|_{\alpha} \le C\|u\|_{1+\alpha},$$
$$|\nabla\cdot\boldsymbol{\sigma}|_{\alpha} = |f-Xu|_{\alpha} \le |f|_{\alpha}+|Xu|_{\alpha} \le |f|_{\alpha}+C\|u\|_{1+\alpha},$$

and the theorem follows by combining these bounds above. ∎

Theorem 3.5 shows improvement in the order of approximation by choosing $\beta_1 \in (0,\alpha]$. We note that Theorem 3.5 assumes $\beta_1 \ge 1-\alpha$ to be consistent with interpolation bound (3.3). In practice this restriction is unnecessary. Choosing $\beta_1 = 1-\alpha$ restores exactly $O(h)$ convergence. Since for many elliptic problems (Poisson, Stokes, linear elasticity, etc.) we have $\alpha \in [0.5, 1)$, we may choose $\beta_1 \in [1-\alpha, \alpha)$ or simply $\beta_1 = 0.5$. This flexibility in parameter choice shows that almost no a priori information of the singular solutions is required for this approach to apply. Rough lower bounds on α suffice.

We also may choose to apply the techniques here locally to each singular point in the domain. The loss of smoothness of the solution at irregular boundary points is a local phenomenon, but it causes global difficulties. Our motivation is to provide a local treatment that eliminates the global effect of the irregular boundary point. To this end, let $\Omega_R = \{\mathbf{x} \in \Omega : r > R\}$ be a subdomain of Ω of fixed size that does not include $\mathcal{E}$. Since on Ω_R the norms $|||(u,\boldsymbol{\sigma})|||_{0,0}$ and $|||(u,\boldsymbol{\sigma})|||_{\beta_1,\beta_2}$ are equivalent, Theorem 3.5 indicates that the error is reduced at optimal asymptotic rates away from $\mathcal{E}$—thus eliminating the pollution effect.

We conclude this section with a brief justification of the reduction of the error with respect to a nonweighted norm.

THEOREM 3.6. *Let $\beta_1 \ge 1-\tilde{\alpha}$. Then under the same assumptions as Theorem* 3.5, *the error in approximating $\boldsymbol{\sigma}$ satisfies*

$$\|\boldsymbol{\sigma}-\boldsymbol{\sigma}^h\| \le C\,h^{\tilde{\alpha}}, \tag{3.21}$$

where C is independent of h.

Proof. By the triangle inequality we take

$$\|\boldsymbol{\sigma}-\boldsymbol{\sigma}^h\| \le \|\boldsymbol{\sigma}-I^h\boldsymbol{\sigma}\| + \|\boldsymbol{\sigma}^h-I^h\boldsymbol{\sigma}\| \tag{3.22}$$

and consider the terms above separately. The first term may be bounded by (3.6),

$$\|\boldsymbol{\sigma}-I^h\boldsymbol{\sigma}\| \le Ch^{\tilde{\alpha}}|\boldsymbol{\sigma}|_{\alpha}. \tag{3.23}$$

For the second term, we require an inequality proved in [10]:

$$\|\boldsymbol{\tau}^h\| \le Ch^{-\delta}\|\boldsymbol{\tau}^h\|_{0,\delta} \tag{3.24}$$

for $\delta > 0$ and all $\boldsymbol{\tau}^h \in \mathcal{W}^h$. Now, using (3.24), the triangle inequality, Theorem 3.5, and (3.8) results in

$$\begin{aligned}\|\boldsymbol{\sigma}^h-I^h\boldsymbol{\sigma}\| &\le Ch^{-\beta_1}\|\boldsymbol{\sigma}^h-I^h\boldsymbol{\sigma}\|_{0,\beta_1}\\ &\le Ch^{-\beta_1}\left(\|\boldsymbol{\sigma}-\boldsymbol{\sigma}^h\|_{0,\beta_1}+\|\boldsymbol{\sigma}-I^h\boldsymbol{\sigma}\|_{0,\beta_1}\right)\\ &\le C\,h^{-\beta_1}h^{\beta_1+\tilde{\alpha}} \le C\,h^{\tilde{\alpha}}.\end{aligned} \tag{3.25}$$

Combining (3.22), (3.23), and (3.25) yields the bound in (3.21) and thus completes the theorem. ∎

4. Numerical results. In this section we validate the weighted-norm method by presenting two numerical test problems. Define the domain $\Omega = \{(r, \theta) : r < 1, \theta \in (0, 7\pi/4)\}$ as pictured in Figure 4.1, which results in singularities with $\alpha = 4/7$.

While the analysis in this paper assumes one singular corner with norms weighted by r^{β_1} and r^{β_2}, generalization to more than one singular corner is straightforward. We illustrate that here by defining the local weight functions

$$w_i = \begin{cases} (r/R)^{\beta_i} & \text{for } r < R, \\ 1 & \text{for } r \geq R \end{cases}$$

for $i = 1, 2$ and $R = 0.25$. We minimize the weighted functional

$$G(u, \boldsymbol{\sigma}; f) = \|w_1 A^{-1/2}(\boldsymbol{\sigma} + A\nabla u)\|^2 + \|w_2(\nabla \cdot \boldsymbol{\sigma} + \mathbf{b} \cdot \nabla u + cu - f)\|^2$$

for u^h using continuous P_1 finite elements and $\boldsymbol{\sigma}^h$ using RT_0 elements. We take $\beta_1 = \beta_2 = 0.57 \approx \alpha$ as the power of the weight functions. A regular triangulation of the domain is used, where the meshsize parameter is $h = \min_{K \in \mathcal{T}^h} diam(K)$.

In Table 4.1 we describe a sequence of nested discretizations used to test convergence of the method. Numerical results are summarized in Table 4.2, where we report the computed convergence rate from the two finest meshes (conv. rate), which may be compared with the asymptotic optimal convergence rate that the interpolant admits (asy. opt.).

We denote the exact solution by u, strongly impose Dirichlet boundary conditions on $\partial\Omega$, and compute the right-hand function f by analytically computing

$$f = -\nabla \cdot A\nabla u + \mathbf{b} \cdot \nabla u + cu.$$

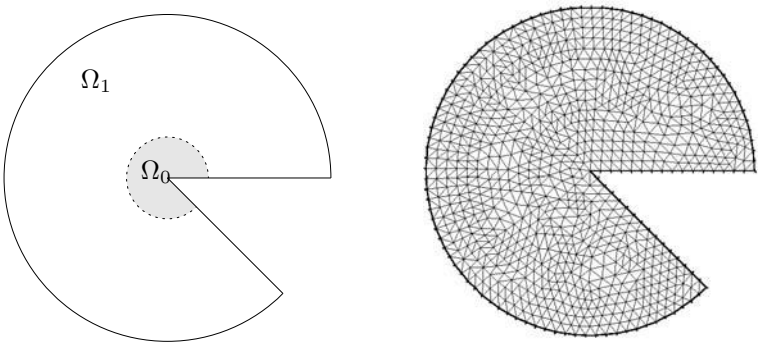


FIG. 4.1. *Test domain Ω, partitioned into subdomains Ω_0 and Ω_1, and the initial discretization, M_1.*

TABLE 4.1
Computational meshes used for both example problems.

mesh	# elements	h
M_1	1576	5.72e-2
M_2	6304	2.86e-2
M_3	25216	1.43e-2
M_4	100864	7.15e-3

TABLE 4.2
Numerical results for Problems 1 *and* 2.

	Problem 1			Problem 2		
$G^{1/2}$	Ω	Ω_0	Ω_1	Ω	Ω_0	Ω_1
mesh M_1	8.57e-2	7.65e-2	3.84e-2	8.94e-1	7.45e-1	4.94e-1
M_2	4.30e-2	3.85e-2	1.91e-2	4.67e-1	3.94e-1	2.49e-1
M_3	2.15e-2	1.93e-2	9.52e-3	2.40e-1	2.05e-1	1.25e-1
M_4	1.08e-2	9.66e-3	4.75e-3	1.23e-1	1.06e-2	6.23e-2
conv. rate	1.00	1.00	1.00	0.98	0.96	1.00
asy. opt.	1	1	1	1	1	1
$\|\boldsymbol{\sigma} - \boldsymbol{\sigma}^h\|$	Ω	Ω_0	Ω_1	Ω	Ω_0	Ω_1
mesh M_1	1.19e-1	1.16e-1	2.84e-2	8.60e-2	8.12e-2	2.84e-2
M_2	8.15e-2	8.03e-2	1.39e-2	5.21e-2	5.06e-2	1.26e-2
M_3	5.52e-2	5.48e-2	6.93e-3	3.21e-2	3.16e-2	5.84e-3
M_4	3.73e-2	3.72e-2	3.46e-3	2.03e-2	2.02e-2	2.79e-3
conv. rate	0.57	0.56	1.00	0.61	0.60	1.05
asy. opt.	4/7	4/7	1	4/7	4/7	1
$\|u - u^h\|_1$	Ω	Ω_0	Ω_1	Ω	Ω_0	Ω_1
mesh M_1	1.10e-1	1.07e-1	2.58e-2	2.51e-1	2.41e-1	7.17e-2
M_2	7.42e-2	7.31e-2	1.30e-2	1.66e-1	1.62e-1	3.56e-2
M_3	4.99e-2	4.95e-2	6.51e-3	1.09e-1	1.07e-1	1.75e-2
M_4	3.36e-2	3.34e-2	3.26e-3	7.12e-2	7.07e-2	8.66e-3
conv. rate	0.57	0.57	1.00	0.61	0.60	1.02
asy. opt.	4/7	4/7	1	4/7	4/7	1

We also analytically define $\boldsymbol{\sigma} = -A\nabla u$ for comparison to the numerical approximation, $\boldsymbol{\sigma}^h$.

Problem 1 (*Poisson's equation*). As a baseline case we begin with the Laplace operator,

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \text{and} \quad c = 0.$$

The exact solution is chosen to be $u = r^\alpha \sin(\alpha\theta)$, and in this case $f = 0$.

Problem 2 (*convection dominated anisotropic diffusion*). We now consider an operator consisting of anisotropic diffusion dominated by a strong convection with a small reaction term,

$$A = \begin{pmatrix} 0.1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 10 \\ 5 \end{pmatrix}, \quad \text{and} \quad c = 1.$$

The exact solution in this case is chosen to be $u = \tilde{r}^\alpha \sin(\alpha\tilde{\theta})$, where $(\tilde{r}, \tilde{\theta})$ are polar coordinates in the rescaled domain defined by $(\tilde{x}_1, \tilde{x}_2)^t = A^{-1/2}(x_1, x_2)^t$. The resulting data function satisfies $f \in H^k(\Omega)$ for any $k < \alpha$ and takes the form

$$f = 10\alpha\tilde{r}^{\alpha-1} \sin\left((\alpha-1)\tilde{\theta}\right) + 5\alpha\tilde{r}^{\alpha-1} \cos\left((\alpha-1)\tilde{\theta}\right) + \tilde{r}^\alpha \sin\alpha\tilde{\theta}.$$

To distinguish the rates of convergence close to and far from the singular point we define the following partitioning of the domain: $\Omega_0 = \{r < 1/4\} \cap \Omega$ and $\Omega_1 = \Omega \backslash \Omega_0$. Thus, Ω_0 contains the singular point and Ω_1 does not. We solve the problem in Ω but monitor the error in these subdomains as measured in the functional norm, $G^{1/2} = G(u^h, \boldsymbol{\sigma}^h; f)^{1/2}$, the L^2 errors in $\boldsymbol{\sigma}$, and the H^1 seminorm error in u.

Table 4.2 summarizes the computational results of the two problems. In both problems, the weighted functional is reduced at (or very nearly at) optimal $O(h)$, as

predicted by Theorem 3.5. This rate is seen on both partitions of the domain. For both $\|\boldsymbol{\sigma} - \boldsymbol{\sigma}^h\|$ and $|u - u^h|_1$, the optimal rates are $O(h^{4/7})$ and $O(h)$ in Ω_0 and Ω_1, respectively. Both example problems have similar behavior, though Problem 1 approaches the asymptotic rate faster than the more challenging Problem 2. The results also show no pollution effect as evidenced by a clear distinction in convergence rates between partitions Ω_0 and Ω_1.

We also note that in (3.6) and (3.7) we assume $\tilde{\alpha} \in (0, \alpha]$. The rates we achieve in Table 4.2 indicate that $\tilde{\alpha} = \alpha$. We also observe this same result in many numerical tests not shown here.

REFERENCES

[1] C. Bacuta, V. Nistor, and L. Zikatanov, *Improving the rate of convergence of high-order finite elements on polyhedra* I: *A priori estimates*, Numer. Funct. Anal. Optim., 26 (2005), pp. 613–639.

[2] D. Braess, *Finite Elements: Theory, Fast Solvers and Applications in Solid Mechanics*, Cambridge University Press, Cambridge, UK, 2001.

[3] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer-Verlag, New York, 1991.

[4] Z. Cai and J. Ku, *The L^2 norm error estimates for the div least-squares method*, SIAM J. Numer. Anal., 44 (2006), pp. 1721–1734.

[5] Z. Cai, R. Lazarov, T. A. Manteuffel, and S. F. McCormick, *First-order system least squares for second-order partial differential equations: Part* I, SIAM J. Numer. Anal., 31 (1994), pp. 1785–1799.

[6] C. Cox and G. Fix, *On the accuracy of least squares methods in the presence of corner singularities*, Comput. Math. Appl., 10 (1984), pp. 463–475.

[7] P. Grisvard, *Singularities in Boundary Value Problems*, Springer-Verlag, Berlin, 1992.

[8] V. Kondratiev, *Boundary problems for elliptic equations in domains with conical or angular points*, Trans. Moscow Math. Soc., 16 (1967), pp. 227–313.

[9] J. Ku, *A remark on the coercivity for a first-order least-squares method*, Numer. Methods Partial Differential Equations, 23 (2007), pp. 1577–1581.

[10] E. Lee, T. A. Manteuffel, and C. R. Westphal, *Weighted-norm first-order system least squares (FOSLS) for problems with corner singularities*, SIAM J. Numer. Anal., 44 (2006), pp. 1974–1996.

[11] E. Lee, T. Manteuffel, and C. Westphal, *Weighted-norm first-order system least squares (FOSLS) for div/curl systems with three dimensional edge singularities*, SIAM J. Numer. Anal., 46 (2008), pp. 1619–1639.

[12] S. Nicaise, *Regularity of the solutions of elliptic systems in polyhedral domains*, Bull. Belg. Math. Soc. Simon Stevin, 4 (1997), pp. 411–429.

SIAM J. NUMER. ANAL.
Vol. 46, No. 4, pp. 1653–1681
© 2008 Society for Industrial and Applied Mathematics

DISCRETE FOURIER ANALYSIS, CUBATURE, AND INTERPOLATION ON A HEXAGON AND A TRIANGLE*

HUIYUAN LI†, JIACHANG SUN†, AND YUAN XU‡

Abstract. Several problems of trigonometric approximation on a hexagon and a triangle are studied using the discrete Fourier transform and orthogonal polynomials of two variables. A discrete Fourier analysis on the regular hexagon is developed in detail, from which the analysis on the triangle is deduced. The results include cubature formulas and interpolation on these domains. In particular, a trigonometric Lagrange interpolation on a triangle is shown to satisfy an explicit compact formula, which is equivalent to the polynomial interpolation on a planar region bounded by Steiner's hypocycloid. The Lebesgue constant of the interpolation is shown to be in the order of $(\log n)^2$. Furthermore, a Gauss cubature is established on the hypocycloid.

Key words. discrete Fourier series, trigonometric, Lagrange interpolation, hexagon, triangle, orthogonal polynomials

AMS subject classifications. 41A05, 41A10

DOI. 10.1137/060671851

1. Introduction. Approximation by trigonometric polynomials, or equivalently complex exponentials, is at the root of approximation theory. Many problems central to approximation theory were first studied for trigonometric polynomials on the real line or on the unit circle, as illustrated in the classical treatise of Zygmund [26]. Much of the advancement in the theory of trigonometric approximation is due to the periodicity of the function. In the case of one variable, one usually works with 2π periodic functions. A straightforward extension to several variables is the tensor product type, where one works with functions that are 2π periodic in each of their variables. There are, however, other definitions of periodicity in several variables, most notably the periodicity defined by a lattice, which is a discrete subgroup defined by $A\mathbb{Z}^d$, where A is a nonsingular matrix, and the periodic function satisfies $f(x + Ak) = f(x)$ for all $k \in \mathbb{Z}^d$. With such a periodicity, one works with exponentials or trigonometric functions of the form $\mathrm{e}^{i2\pi\alpha\cdot x}$, where α and x are in proper sets of $\mathbb{R}^d$, not necessarily the usual trigonometric polynomials.

If Ω is a bounded open set that tiles $\mathbb{R}^d$ with the lattice $L = A\mathbb{Z}^d$, meaning that $\Omega + A\mathbb{Z}^d = \mathbb{R}^d$, then, according to a theorem of Fuglede [6], the family of exponentials $\{\mathrm{e}^{2\pi i\alpha\cdot x} : \alpha \in L^\perp\}$, where $L^\perp = A^{-\mathrm{tr}}\mathbb{Z}^d$ is the dual lattice of L, forms an orthonormal basis in $L^2(\Omega)$. In particular, this shows that it is appropriate to use such exponentials for approximation. Naturally one can study Fourier series on Ω, defined by these exponentials, which is more or less the usual multivariate Fourier series with a change of variables from $\mathbb{Z}^d$ to $A\mathbb{Z}^d$. Lattices appear prominently in various problems in combinatorics and analysis (see, for example, [2, 5]); the Fourier series associated with them are studied in the context of information science and

*Received by the editors October 9, 2006; accepted for publication (in revised form) December 26, 2007; published electronically April 11, 2008.
http://www.siam.org/journals/sinum/46-4/67185.html

†Institute of Software, Chinese Academy of Sciences, Beijing 100080, China (hynli@mail.rdcps.ac.cn, sun@mail.rdcps.ac.cn). The first and the second authors were supported by NSFC grants 10431050 and 60573023. The second author was supported by National Basic Research Program grant 2005CB321702.

‡Department of Mathematics, University of Oregon, Eugene, OR 97403-1222 (yuan@math.uoregon.edu). This author was supported by NSF grant DMS-0604056.

physics, such as sampling theory [7, 11] and multivariate signal processing [3]. As far as we are aware, they have not been studied much in approximation theory until recently. Part of the reason may lie in the complexity of the set Ω. In the simplest nontensor product case, it is a regular hexagon region on the plane, not one of those regions that people are used to in classical analysis. It turns out, however, that the results on the hexagon are instrumental to the study of trigonometric functions on the triangle [21, 22]. The Fourier series on the hexagon were studied from scratch in [21], and two families of functions analogous to those of sine and cosine but defined using three directions were identified and studied. In [22] the two generalized trigonometric functions were derived as solutions of the eigenfunctions of the Laplace equation on a triangle, and the Fourier series in terms of them were studied.

The purpose of the present paper is to study discrete Fourier analysis on the hexagon and on the triangle; in particular, we will derive results on cubature and Lagrange interpolation on these domains. We shall follow two approaches. The first one relies on the discrete Fourier transforms associated with the lattices, which leads naturally to an interpolation formula, and the result applies to a general lattice. We then apply the result to the hexagonal tiling. The symmetry group of the hexagon lattice is the reflection group $\mathcal{A}_2$. The generalized cosine and sine are invariant and anti-invariant projections, respectively, of the basic exponential function under $\mathcal{A}_2$, and they can be studied as functions defined on a fundamental domain, which is one of the equilateral triangles inside the hexagon. This allows us to pass from the hexagon to the triangle. The discrete Fourier transform can be regarded as a cubature formula. In the case of a triangle, the trigonometric functions for which such a cubature holds is a graded algebra due to a connection to polynomials of two variables, which shows that the cubature is of Gaussian type.

The second approach uses orthogonal polynomials in two variables. It turns out that the generalized cosine and sine functions were studied earlier by Koornwinder in [9], who also started with eigenfunctions of the Laplace operator and identified them as orthogonal polynomials on the region bounded by Steiner's hypocycloid, which he called Chebyshev polynomials of the first and the second kind, respectively. It is well known that orthogonal polynomials are closely related to cubature formulas and polynomial interpolation (see, for example, [4, 14, 20, 24, 25]). Roughly speaking, a cubature formula of high precision exists if the set of its nodes is an algebraic variety of a polynomial ideal generated by (quasi-)orthogonal polynomials, and the size of the variety is equal to the codimension of the ideal. Furthermore, such a cubature formula arises from an interpolation polynomial. The rich structure of the generalized Chebyshev polynomials allow us to study their common zeros and establish the cubature and the polynomial interpolation. One particular result shows that Gaussian cubature formula exists for one integral on the region bounded by the hypocycloid. It should be mentioned that, unlike the case of one variable, Gaussian cubature formulas rarely exist in several variables.

The most concrete result, in our view, is the trigonometric interpolation on the triangle. For a standard triangle $\Delta := \{(x, y) : x \ge 0, y \ge 0, x + y \le 1\}$, the set of interpolation points are $X_n := \{(\frac{i}{n}, \frac{j}{n}) : 0 \le i \le j \le n\}$ in Δ. We found a compact formula for the fundamental interpolation functions in terms of the elementary trigonometric functions, which ensures that the interpolation can be computed efficiently. Moreover, the uniform operator norm (the Lebesgue constant) of the interpolation is shown to grow in the order of $(\log n)^2$ as n goes to infinity. This is in sharp contrast to the algebraic polynomial interpolation on the set X_n, which also exists and

can be given by simple formulas but has an undesirable convergence behavior (see, for example, [17]). In fact, it is well known that equally spaced points are not good for algebraic polynomial interpolation. For example, a classical result shows that the polynomial interpolation of $f(x) = |x|$ on equally spaced points of $[-1, 1]$ diverges at all points other than $-1, 0, 1$.

The paper is organized as follows. In section 2 we state the basic results on lattices and discrete Fourier transform, and establish the connection to the interpolation formula. Section 3 contains results on the regular hexagon, obtained from applying the general result of the previous section. Section 4 studies the generalized sine and cosine functions, and the discrete Fourier analysis on the triangle. In particular, it contains the results on the trigonometric interpolation. The generalized Chebyshev polynomials and their connection to the orthogonal polynomials of two variables are discussed in section 5 and used to establish the existence of Gaussian cubature and an alternative way to derive interpolation polynomial.

2. Continuous and discrete Fourier analysis with lattice.

2.1. Lattice and Fourier series. A lattice in $\mathbb{R}^d$ is a discrete subgroup that contains d linearly independent vectors,

$$L := \{k_1a_1 + k_2a_2 + \cdots + k_da_d : \quad k_i \in \mathbb{Z},\, i = 1, 2, \ldots, d\}, \tag{2.1}$$

where $a_1, \ldots, a_d$ are linearly independent column vectors in $\mathbb{R}^d$. Let A be the $d \times d$ matrix whose columns are $a_1, \ldots, a_d$. Then A is called a generator matrix of the lattice L. We can write L as L_A and a short notation for L_A is $A\mathbb{Z}^d$; that is,

$$L_A = A\mathbb{Z}^d = \left\{Ak : k \in \mathbb{Z}^d\right\}.$$

Any d-dimensional lattice has a dual lattice $L^\perp$ given by

$$L^\perp = \{x \in \mathbb{R}^d : x \cdot y \in \mathbb{Z} \text{ for all } y \in L\},$$

where $x \cdot y$ denotes the usual Euclidean inner product of x and y. The generator matrix of the dual lattice $L_A^\perp$ is $A^{-\mathsf{tr}} := (A^{\mathsf{tr}})^{-1}$, where A^{tr} denote the matrix transpose of A [2]. Throughout this paper whenever we need to treat $x \in \mathbb{R}^d$ as a vector, we treat it as a column vector, so that $x \cdot y = x^{\mathsf{tr}}y$.

A bounded set Ω of $\mathbb{R}^d$ is said to tile $\mathbb{R}^d$ with the lattice L if

$$\sum_{\alpha \in L} \chi_\Omega(x + \alpha) = 1 \qquad \text{for almost all } x \in \mathbb{R}^d,$$

where χ_Ω denotes the characteristic function of Ω. We write this as $\Omega + L = \mathbb{R}^d$. Tiling and Fourier analysis are closely related; see, for example, [6, 8]. We will need the following result due to Fuglede [6]. Let $\langle \cdot, \cdot \rangle_\Omega$ denote the inner product in the space $L^2(\Omega)$ defined by

$$\langle f, g\rangle_\Omega = \frac{1}{\operatorname{mes}(\Omega)} \int_\Omega f(x)\overline{g(x)}dx. \tag{2.2}$$

THEOREM 2.1 (see [6]). *Let Ω in $\mathbb{R}^d$ be a bounded open domain and let L be a lattice of $\mathbb{R}^d$. Then $\Omega + L = \mathbb{R}^d$ if and only if $\{e^{2\pi i\alpha\cdot x} : \alpha \in L^\perp\}$ is an orthonormal basis with respect to the inner product* (2.2).

The orthonormal property is defined with respect to normalized Lebesgue measure on Ω. If $L = L_A$, then the measure of Ω is equal to $|\det A|$. Furthermore, since $L_A^\perp = A^{-\mathsf{tr}}\mathbb{Z}^d$, we can write $\alpha = A^{-\mathsf{tr}}k$ for $\alpha \in L_A^\perp$ and $k \in \mathbb{Z}^d$, so that $\alpha \cdot x = (A^{-\mathsf{tr}}k)\cdot x = k^{\mathsf{tr}}A^{-1}x$. Hence the orthogonality in the theorem is

$$\frac{1}{|\det(A)|}\int_\Omega \mathrm{e}^{2\pi i k^{\mathsf{tr}}A^{-1}\xi}d\xi = \delta_{k,0}, \qquad k \in \mathbb{Z}^d. \tag{2.3}$$

The set Ω is called a spectral (fundamental) set for the lattice L. If $L = L_A$, we also write $\Omega = \Omega_A$.

A function $f \in L^1(\Omega_A)$ can be expanded into a Fourier series

$$f(x) \sim \sum_{k\in\mathbb{Z}^d} c_k \mathrm{e}^{2\pi i k^{\mathsf{tr}}A^{-1}x}, \qquad c_k = \frac{1}{|\det(A)|}\int_\Omega f(x)\mathrm{e}^{-2\pi i k^{\mathsf{tr}}A^{-1}x}dx,$$

which is just a change of variables away from the usual multiple Fourier series. The Fourier transform $\widehat{f}$ of a function $f \in L^1(\mathbb{R}^d)$ and its inverse are defined by

$$\widehat{f}(\xi) = \int_{\mathbb{R}^d} f(x)\mathrm{e}^{-2\pi i\,\xi\cdot x}dx \quad \text{and} \quad f(x) = \int_{\mathbb{R}^d}\widehat{f}(\xi)\mathrm{e}^{2\pi i\,\xi\cdot x}d\xi.$$

One important tool for us is the Poisson summation formula given by

$$\sum_{k\in\mathbb{Z}^d} f(x+Ak) = \frac{1}{|\det(A)|}\sum_{k\in\mathbb{Z}^d}\widehat{f}(A^{-\mathsf{tr}}k)\mathrm{e}^{2\pi i k^{\mathsf{tr}}A^{-1}x}, \tag{2.4}$$

which holds for $f \in L^1(\Omega_A)$ under the usual assumption on the convergence of the series in both sides. This formula has been used by various authors (see, for example, [1, 5, 7, 9, 19]). An immediate consequence of the Poisson summation formula is the following sampling theorem (see, for example, [7, 11]).

PROPOSITION 2.2. *Let Ω be a spectral set of the lattice $A\mathbb{Z}^d$. Assume that $\widehat{f}$ is supported on Ω and $\widehat{f} \in L^2(\Omega)$. Then*

$$f(x) = \sum_{k\in\mathbb{Z}^d} f(A^{-\mathsf{tr}}k)\Psi_\Omega(x - A^{-\mathsf{tr}}k) \tag{2.5}$$

in $L^2(\Omega)$, where

$$\Psi_\Omega(x) = \frac{1}{|\det(A)|}\int_\Omega \mathrm{e}^{2\pi i x\cdot\xi}d\xi. \tag{2.6}$$

In fact, since $\widehat{f}$ is supported on Ω_A, $\widehat{f}(\xi) = \sum_{k\in\mathbb{Z}^d}\widehat{f}(\xi + Ak)$. Integrating over Ω, the sampling equation follows from (2.4) and Fourier inversion. We note that

$$\Psi_\Omega(A^{-\mathsf{tr}}j) = \delta_{0,j} \qquad \text{for all } j \in \mathbb{Z}^d$$

by (2.3), so that Ψ_Ω can be considered as a cardinal interpolation function.

2.2. Discrete Fourier analysis and interpolation. A function f defined on $\mathbb{R}^d$ is called a periodic function with respect to the lattice $A\mathbb{Z}^d$ if

$$f(x + Ak) = f(x) \qquad \text{for all } k \in \mathbb{Z}^d.$$

For a given lattice L, its spectral set is evidently not unique. The orthogonality (2.3) is independent of the choice of Ω. This can be seen directly from the fact that the function $x \mapsto e^{2\pi i k^{\mathrm{tr}} A^{-1} x}$ is periodic with respect to the lattice $A\mathbb{Z}^d$.

In order to carry out a discrete Fourier analysis with respect to a lattice, we shall fix Ω such that Ω contains 0 in its interior, and we further require in this subsection that the tiling $\Omega + A\mathbb{Z}^d = \mathbb{R}^d$ holds pointwise and without overlapping. In other words, we require

$$\sum_{k\in\mathbb{Z}^d} \chi_\Omega(x + Ak) = 1 \text{ for all } x \in \mathbb{R}^d \quad \text{and} \quad (\Omega + Ak) \cap (\Omega + Aj) = \emptyset \text{ if } k \neq j. \tag{2.7}$$

For example, we can take $\Omega = [-\frac{1}{2}, \frac{1}{2})^d$ for the standard lattice $\mathbb{Z}^d$.

DEFINITION 2.3. *Let A and B be two nonsingular matrices such that Ω_A and Ω_B satisfy* (2.7). *Assume that all entries of $N := B^{\mathrm{tr}}A$ are integers. Denote*

$$\Lambda_N := \{k \in \mathbb{Z}^d : B^{-\mathrm{tr}} k \in \Omega_A\} \quad \textit{and} \quad \Lambda_N^\dagger := \{k \in \mathbb{Z}^d : A^{-\mathrm{tr}} k \in \Omega_B\}.$$

Two points $x, y \in \mathbb{R}^d$ are said to be congruent with respect to the lattice $A\mathbb{Z}^d$ if $x - y \in A\mathbb{Z}^d$, and we write $x \equiv y \mod A$.

THEOREM 2.4. *Let A, B, and N be as in Definition* 2.3. *Then*

$$\frac{1}{|\det(N)|} \sum_{j\in\Lambda_N} \mathrm{e}^{2\pi i k^{\mathrm{tr}} N^{-1} j} = \begin{cases} 1 & \textit{if } k \equiv 0 \mod N^{\mathrm{tr}}, \\ 0 & \textit{otherwise.} \end{cases} \tag{2.8}$$

Proof. Let $f \in L^1(\mathbb{R}^d)$ be such that Poisson summation formula holds. Denote

$$\Phi^A f(x) := \sum_{k\in\mathbb{Z}^d} f(x + Ak).$$

Then the Poisson summation formula (2.4) shows that

$$\Phi^A f(B^{-\mathrm{tr}} j) = \frac{1}{|\det(A)|} \sum_{k\in\mathbb{Z}^d} \widehat{f}(A^{-\mathrm{tr}} k) e^{2\pi i k^{\mathrm{tr}} N^{-1} j}, \qquad j \in \mathbb{Z}^d.$$

Since $\mathbb{R}^d = \Omega_B + B\mathbb{Z}^d$, we can write $A^{-\mathrm{tr}} k = x_B + Bm$ for every $k \in \mathbb{Z}^d$ with $x_B \in \Omega_B$ and $m \in \mathbb{Z}^d$. The fact that N has integer entries shows that $A^{\mathrm{tr}} x_B = l \in \mathbb{Z}^d$. That is, every $k \in \mathbb{Z}^d$ can be written as $k = l + N^{\mathrm{tr}} m$ with $l \in \Lambda_N^\dagger$ and $m \in \mathbb{Z}^d$. It is easy to see that $\mathrm{e}^{2\pi i k^{\mathrm{tr}} N^{-1} j} = \mathrm{e}^{2\pi i l^{\mathrm{tr}} N^{-1} j}$ under this decomposition. Consequently,

$$\begin{aligned} \Phi^A f(B^{-\mathrm{tr}} j) &= \frac{1}{|\det(A)|} \sum_{l\in\Lambda_N^\dagger} \sum_{m\in\mathbb{Z}^d} \widehat{f}(A^{-\mathrm{tr}} l + Bm) \mathrm{e}^{2\pi i l^{\mathrm{tr}} N^{-1} j} \\ &= \frac{1}{|\det(A)|} \sum_{l\in\Lambda_N^\dagger} \Phi^B \widehat{f}(A^{-\mathrm{tr}} l) \mathrm{e}^{2\pi i l^{\mathrm{tr}} N^{-1} j}. \end{aligned} \tag{2.9}$$

Similarly, we have

$$\Phi^B \widehat{f}(A^{-\mathrm{tr}} l) = \frac{1}{|\det(B)|} \sum_{j\in\Lambda_N} \Phi^A f(B^{-\mathrm{tr}} j) \mathrm{e}^{-2\pi i l^{\mathrm{tr}} N^{-1} j}. \tag{2.10}$$

Substituting (2.9) into (2.10) leads to the identity

$$\Phi^B \widehat{f}(A^{-\mathsf{tr}} l) = \frac{1}{|\det(AB)|} \sum_{k \in \Lambda_N^\dagger} \Phi^B \widehat{f}(A^{-\mathsf{tr}} k) \sum_{j \in \Lambda_N} \mathrm{e}^{2\pi i (k-l)^{\mathsf{tr}} N^{-1} j},$$

which holds for f in, say, the Schwartz class of functions, from which (2.8) follows immediately. □

THEOREM 2.5. *Let A, B, and N be as in Definition* 2.3. *Define*

$$\langle f, g \rangle_N = \frac{1}{|\det(N)|} \sum_{j \in \Lambda_N} f(B^{-\mathsf{tr}} j) \overline{g(B^{-\mathsf{tr}} j)}$$

for f, g in $C(\Omega_A)$, the space of continuous functions on Ω_A. Then

$$\langle f, g \rangle_{\Omega_A} = \langle f, g \rangle_N \tag{2.11}$$

for all f, g in the finite-dimensional subspace

$$\mathcal{H}_N := \operatorname{span} \left\{ \phi_k : \phi_k(x) = \mathrm{e}^{2\pi i k^{\mathsf{tr}} A^{-1} x},\ k \in \Lambda_N^\dagger \right\}.$$

Proof. By the definition of $\langle \cdot, \cdot \rangle_N$, (2.8) implies that $\langle \phi_l, \phi_k \rangle_N = \delta_{l,k}$, $l, k \in N^{\mathsf{tr}}$, which shows, by (2.3), that (2.11) holds for $f = \phi_l$ and $g = \phi_k$. Since both inner products are sesquilinear, the stated result follows. □

If Λ is a subset of $\mathbb{Z}^d$, we denote by $\#\Lambda$ the number of points in Λ. Evidently the dimension of $\mathcal{H}_N$ is $\#\Lambda_N^\dagger$.

Let $\mathcal{I}_N f$ denote the Fourier expansion of $f \in C(\Omega_A)$ in $\mathcal{H}_N$ with respect to the inner product $\langle \cdot, \cdot \rangle$. Then

$$\begin{aligned} \mathcal{I}_N f(x) &:= \sum_{j \in \Lambda_N^\dagger} \langle f, \phi_j \rangle_N \phi_j(x) \\ &= \frac{1}{|\det(N)|} \sum_{k \in \Lambda_N} f(B^{-\mathsf{tr}} k) \sum_{j \in \Lambda_N^\dagger} \phi_j(x) \overline{\phi_j(B^{-\mathsf{tr}} k)}. \end{aligned}$$

Hence, analogous to the sampling theorem in Proposition 2.2, we have

$$\mathcal{I}_N f(x) = \sum_{k \in \Lambda_N} f(B^{-\mathsf{tr}} k) \Psi_{\Omega_B}^A (x - B^{-\mathsf{tr}} k), \qquad f \in C(\Omega_A), \tag{2.12}$$

where

$$\Psi_{\Omega_B}^A(x) = \frac{1}{|\det(N)|} \sum_{j \in \Lambda_N^\dagger} \mathrm{e}^{2\pi i j^{\mathsf{tr}} A^{-1} x}. \tag{2.13}$$

THEOREM 2.6. *Let A, B, and N be as in Definition* 2.3. *If $\Lambda_{N^{\mathsf{tr}}} = \Lambda_N^\dagger$, then $\mathcal{I}_N f$ is the unique interpolation operator on N in $\mathcal{H}_N$; that is,*

$$\mathcal{I}_N f(B^{-\mathsf{tr}} j) = f(B^{-\mathsf{tr}} j) \qquad \text{for all } j \in \Lambda_N.$$

In particular, $\#\Lambda_N = \#\Lambda_{N^{\mathsf{tr}}} = |\det(N)|$. Furthermore, the fundamental interpolation function $\Psi_{\Omega_A}^B$ satisfies

$$\Psi_{\Omega_B}^A(x) = \sum_{j \in \mathbb{Z}^d} \Psi_{\Omega_B}(x + Aj). \tag{2.14}$$

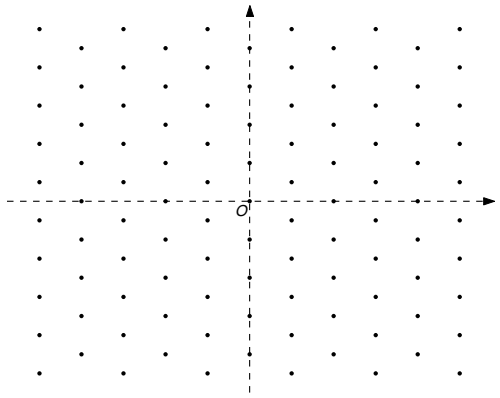


FIG. 1. *The lattice* L_H.

Proof. Using (2.8) with N in place of N^{tr} gives immediately that $\Psi^A_{\Omega_B}(B^{-\mathsf{tr}}k) = \delta_{k,0}$ for $k \in \Lambda_N$, so that the interpolation holds. This also shows that $\{\Psi^A_{\Omega_B}(x - B^{-\mathsf{tr}}k) : k \in \Lambda_N\}$ is linearly independent. Equation (2.8) with $k = 0$ shows immediately that $\#\Lambda_N = |\det(N)|$. Similarly, with N replaced by N^{tr}, we have $\#\Lambda_N = |\det(N^{\mathsf{tr}})|$. This shows that $\#\Lambda_N = |\det(N)| = \#\Lambda_{N^{\mathsf{tr}}}$. Furthermore, let $M = (\phi_k(B^{-\mathsf{tr}}j))_{j,k\in\Lambda_N}$ be the interpolation matrix; then (2.8) shows that M is a unitary matrix, which shows in particular that M is invertible. Consequently the interpolation by $\mathcal{H}_n$ on points in Λ_N is unique.

Finally, since $\widehat{\Psi}_{\Omega_B} = |\det(B)|^{-1}\chi_{\Omega_B}$, we have

$$\begin{aligned}\Psi^A_{\Omega_B}(x) &= \frac{1}{|\det(N)|}\sum_{j\in\mathbb{Z}^d}\phi_j(x)\chi_{\Omega_B}(A^{-\mathsf{tr}}j)\\ &= \frac{1}{|\det(A)|}\sum_{j\in\mathbb{Z}^d}\phi_j(x)\widehat{\Psi}_{\Omega_B}(A^{-\mathsf{tr}}j),\end{aligned}$$

from which (2.14) follows from the Poisson summation formula (2.4). ∎

The result in this section applies to fairly general lattices. In the following section we apply it to the hexagon lattice in two dimensions. We hope to report results for higher dimensional lattices in a future work.

3. Discrete Fourier analysis on the hexagon.

3.1. Hexagon lattice and Fourier analysis. The hexagon lattice is depicted in Figure 1. The generator matrix and the spectral set of the hexagon lattice is given by

$$H = \begin{pmatrix}\sqrt{3} & 0\\ -1 & 2\end{pmatrix}, \qquad \Omega_H = \left\{(x_1,x_2):\ -1 \le x_2, \tfrac{\sqrt{3}}{2}x_1 \pm \tfrac{1}{2}x_2 < 1\right\}.$$

The strict inequality in the definition of Ω_H comes from our assumption in (2.7).

As shown in [21], it is more convenient to use homogeneous coordinates (t_1, t_2, t_3) that satisfy $t_1 + t_2 + t_3 = 0$. We define

$$t_1 = -\frac{x_2}{2} + \frac{\sqrt{3}x_1}{2}, \quad t_2 = x_2, \quad t_3 = -\frac{x_2}{2} - \frac{\sqrt{3}x_1}{2}. \tag{3.1}$$

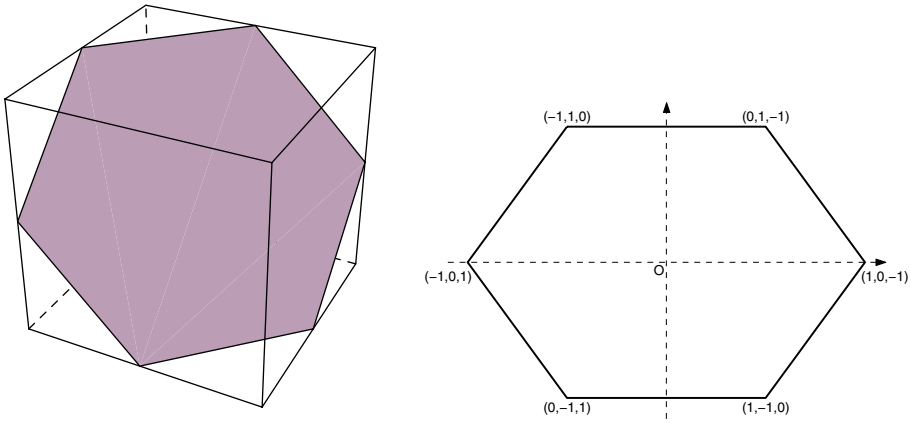


FIG. 2. *The hexagon in homogeneous coordinates.*

Under these homogeneous coordinates, the hexagon Ω_H becomes

$$\Omega = \left\{(t_1, t_2, t_3) \in \mathbb{R}^3 : \ -1 \le t_1, t_2, -t_3 < 1; \, t_1 + t_2 + t_3 = 0\right\},$$

which is the intersection of the plane $t_1 + t_2 + t_3 = 0$ with the cube $[-1, 1]^3$, as shown in Figure 2. Later in the paper we shall depict the hexagon as a two-dimensional figure, but still refer to its points by their homogeneous coordinates, as shown by the right-hand side of Figure 2.

For convenience, we adopt the convention of using bold letters, such as $\mathbf{t}$, to denote points in the space

$$\mathbb{R}^3_H := \{\mathbf{t} = (t_1, t_2, t_3) \in \mathbb{R}^3 : t_1 + t_2 + t_3 = 0\}.$$

In other words, bold letters such as $\mathbf{t}$ stand for homogeneous coordinates. Furthermore, we introduce the notation

$$\mathbb{H} := \mathbb{Z}^3 \cap \mathbb{R}^3_H = \left\{\mathbf{k} = (k_1, k_2, k_3) \in \mathbb{Z}^3 : \ k_1 + k_2 + k_3 = 0\right\}.$$

The inner product on the hexagon under homogeneous coordinates becomes

$$\langle f, g\rangle_H = \frac{1}{|\Omega_H|} \int_{\Omega_H} f(x_1, x_2)\overline{g(x_1, x_2)} dx_1 dx_2 = \frac{1}{|\Omega|} \int_{\Omega} f(\mathbf{t})\overline{g(\mathbf{t})} d\mathbf{t},$$

where $|\Omega|$ denotes the area of Ω.

To apply the general results in the previous section, we choose $A = H$ and $B = \frac{n}{2}H$, where n is a positive integer. Then

$$N = B^{\mathsf{tr}}A = \frac{n}{2}H^T H = \begin{bmatrix} 2n & -n \\ -n & 2n \end{bmatrix}$$

has integer entries. Note that N is a symmetric matrix and $\Lambda_N = \Lambda_N^\dagger$. Using the fact that $B^{-\mathsf{tr}}j = \frac{2}{n}H^{-\mathsf{tr}}j = \frac{1}{n}\left(\frac{2}{\sqrt{3}}j_1 + \frac{1}{\sqrt{3}}j_2, j_2\right)$, it is not hard to see that $B^{-\mathsf{tr}}j \in \Omega_A$, or equivalently $j \in \Lambda_N$, becomes $\mathbf{j} = (j_1, j_2, -j_1 - j_2) \in \mathbb{H}_n$, where

$$\mathbb{H}_n := \{\mathbf{j} \in \mathbb{H} : -n \le j_1, j_2, -j_3 < n\} \qquad \text{and} \qquad \#\mathbb{H}_n = 3n^2.$$

In other words, Λ_N in the previous section becomes $\mathbb{H}_n$ in homogeneous coordinates. The fact that $\#\mathbb{H}_n = 3n^2$ can be easily verified. Moreover, under the change of variables $x = (x_1, x_2) \mapsto (t_1, t_2, t_3)$ in (3.1), it is easy to see that

$$(k_1, k_2)H^{-1}x = \tfrac{1}{3}(k_1, k_2, -k_1 - k_2)\mathbf{t} = \tfrac{1}{3}\mathbf{k}\cdot\mathbf{t}, \tag{3.2}$$

so that $\mathrm{e}^{2\pi i k^{\mathrm{tr}} H^{-1} x} = \mathrm{e}^{\frac{2\pi i}{3}\mathbf{k}^{\mathrm{tr}}\mathbf{t}}$. Therefore, introducing the notation

$$\phi_{\mathbf{j}}(\mathbf{t}) := \mathrm{e}^{\frac{2\pi i}{3}\mathbf{j}^{\mathrm{tr}}\mathbf{t}}, \qquad \mathbf{j} \in \mathbb{H},$$

the orthogonality relation (2.3) becomes

$$\langle \phi_{\mathbf{k}}, \phi_{\mathbf{j}} \rangle_H = \delta_{\mathbf{k},\mathbf{j}}, \qquad \mathbf{k}, \mathbf{j} \in \mathbb{H}. \tag{3.3}$$

The finite-dimensional space $\mathcal{H}_N$ in the previous section becomes

$$\mathcal{H}_n := \operatorname{span}\{\phi_{\mathbf{j}} : \mathbf{j} \in \mathbb{H}_n\} \quad \text{and} \quad \dim \mathcal{H}_n = 3n^2.$$

Under the homogeneous coordinates (3.1), $x \equiv y \pmod{H}$ becomes $\mathbf{t} \equiv \mathbf{s} \mod 3$, where we define

$$\mathbf{t} \equiv \mathbf{s} \mod 3 \quad \Longleftrightarrow \quad t_1 - s_1 \equiv t_2 - s_2 \equiv t_3 - s_3 \mod 3.$$

We call a function f H-periodic if $f(\mathbf{t}) = f(\mathbf{t} + \mathbf{j})$ whenever $\mathbf{j} \equiv 0 \mod 3$. Since $\mathbf{j}, \mathbf{k} \in \mathbb{H}$ implies that $2\mathbf{j}\cdot\mathbf{k} = (j_1 - j_2)(k_1 - k_2) + 3j_3k_3$, we see that $\phi_{\mathbf{j}}$ is H-periodic. The following lemma is convenient in practice.

LEMMA 3.1. *Let $\varepsilon_1 = (2, -1, -1)$, $\varepsilon_2 = (-1, 2, -1)$, $\varepsilon_3 = (-1, -1, 2)$. Then a function $f(\mathbf{t})$ is H-periodic if and only if*

$$f(\mathbf{t} + j\varepsilon_i) = f(\mathbf{t}), \quad j \in \mathbb{Z}, \quad i = 1, 2, 3. \tag{3.4}$$

3.2. Discrete Fourier analysis on the regular hexagon. Using the setup in the previous subsection, Theorem 2.5 becomes, in homogeneous coordinates, the following.

PROPOSITION 3.2. *For $n \ge 0$, define*

$$\langle f, g \rangle_n := \frac{1}{3n^2} \sum_{\mathbf{j} \in \mathbb{H}_n} f(\tfrac{\mathbf{j}}{n})\overline{g(\tfrac{\mathbf{j}}{n})}, \qquad f, g \in C(\Omega). \tag{3.5}$$

Then

$$\langle f, g \rangle_H = \langle f, g \rangle_n, \qquad f, g \in \mathcal{H}_n.$$

The point set $\mathbb{H}_n$, and hence the inner product $\langle \cdot, \cdot \rangle_n$, is not symmetric on Ω in the sense that it contains only part of the points on the boundary, as shown by the left-hand side of Figure 3. We denote by $\mathbb{H}_n^*$ the symmetric point set on Ω,

$$\mathbb{H}_n^* := \{\mathbf{j} \in \mathbb{H} : -n \le j_1, j_2, j_3 \le n\} \quad \text{and} \quad \#\mathbb{H}_n^* = 3n^2 + 3n + 1.$$

The set $\mathbb{H}_n^*$ is shown by the right-hand side of Figure 3.

Using periodicity, we can show that the inner product $\langle \cdot, \cdot \rangle$ is equivalent to a symmetric discrete inner product based on $\mathbb{H}_n^*$. Let

$$\mathbb{H}_n^\circ := \{\mathbf{j} \in \mathbb{H} : -n < j_1, j_2, j_3 < n\}$$

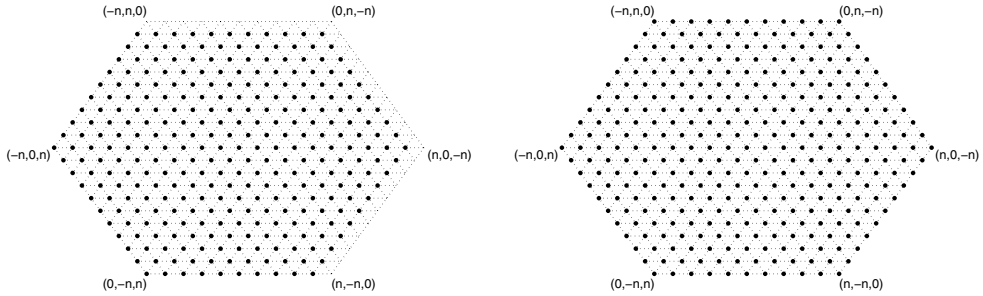


FIG. 3. *The set $\mathbb{H}_n$ (left) and the set $\mathbb{H}_n^*$ (right).*

denote the set of interior points in $\mathbb{H}_n^*$ and $\mathbb{H}_n$. Define

$$\langle f, g\rangle_n^* := \frac{1}{3n^2} \sum_{\mathbf{j}\in\mathbb{H}_n^*} c_{\mathbf{j}}^{(n)} f(\tfrac{\mathbf{j}}{n})\overline{g(\tfrac{\mathbf{j}}{n})}, \qquad f, g \in C(\Omega), \tag{3.6}$$

where

$$c_{\mathbf{j}}^{(n)} = \begin{cases} 1, & \mathbf{j} \in \mathbb{H}_n^\circ \quad \text{(inner points)}, \\ \frac{1}{2}, & \mathbf{j} \in \mathbb{H}_n^* \setminus \mathbb{H}_n^\circ,\ j_1 j_2 j_3 \neq 0 \quad \text{(edge points)}, \\ \frac{1}{3}, & \mathbf{j} \in \mathbb{H}_n^* \setminus \mathbb{H}_n^\circ,\ j_1 j_2 j_3 = 0 \quad \text{(vertices)}, \end{cases}$$

in which we also include the positions of the points; the only one that needs an explanation is the "edge points," which are the points located on the edges but not on the vertices of the hexagon. In analogy to Proposition 3.2, we have the following theorem.

THEOREM 3.3. *For $n \geq 0$,*

$$\langle f, g\rangle_H = \langle f, g\rangle_n = \langle f, g\rangle_n^*, \qquad f, g \in \mathcal{H}_n.$$

Proof. If f is an H-periodic function, then it follows from (3.4) that

$$\begin{aligned} 6 \sum_{\mathbf{j}\in\mathbb{H}_n^*\setminus\mathbb{H}_n} c_{\mathbf{j}}^{(n)} f(\tfrac{\mathbf{j}}{n}) &= 2\left[f(1,-1,0) + f(1,0,-1) + f(0,1,-1) + f(-1,1,0)\right] \\ &\quad + 3 \sum_{1\leq j\leq n-1} \left[f(1, -\tfrac{j}{n}, \tfrac{j}{n} - 1) + f(\tfrac{j}{n} - 1, 1, -\tfrac{j}{n}) + f(1 - \tfrac{j}{n}, \tfrac{j}{n}, -1)\right] \\ &= 3 \sum_{1\leq j\leq n-1} \left[f(-1, 1 - \tfrac{j}{n}, \tfrac{j}{n}) + f(\tfrac{j}{n}, -1, 1 - \tfrac{j}{n}) + f(-\tfrac{j}{n}, \tfrac{j}{n} - 1, 1)\right] \\ &\quad + 4\left[f(0,-1,1) + f(-1,0,1)\right] \\ &= 6 \sum_{\mathbf{j}\in\mathbb{H}_n\setminus\mathbb{H}_{n-1}^*} \left(1 - c_{\mathbf{j}}^{(n)}\right) f(\tfrac{\mathbf{j}}{n}) = 6 \sum_{\mathbf{j}\in\mathbb{H}_n} \left(1 - c_{\mathbf{j}}^{(n)}\right) f(\tfrac{\mathbf{j}}{n}). \end{aligned}$$

Consequently, we have

$$\begin{aligned} \sum_{\mathbf{j}\in\mathbb{H}_n^*} c_{\mathbf{j}}^{(n)} f(\tfrac{\mathbf{j}}{n}) &= \sum_{\mathbf{j}\in\mathbb{H}_n} c_{\mathbf{j}}^{(n)} f(\tfrac{\mathbf{j}}{n}) + \sum_{\mathbf{j}\in\mathbb{H}_n^*\setminus\mathbb{H}_n} c_{\mathbf{j}}^{(n)} f(\tfrac{\mathbf{j}}{n}) \\ &= \sum_{\mathbf{j}\in\mathbb{H}_n} c_{\mathbf{j}}^{(n)} f(\tfrac{\mathbf{j}}{n}) + \sum_{\mathbf{j}\in\mathbb{H}_n} \left(1 - c_{\mathbf{j}}^{(n)}\right) f(\tfrac{\mathbf{j}}{n}) = \sum_{\mathbf{j}\in\mathbb{H}_n} f(\tfrac{\mathbf{j}}{n}). \end{aligned}$$

Replacing f by $f\bar{g}$, we have proved that $\langle f, g\rangle_n = \langle f, g\rangle_n^*$ for H-periodic functions, which clearly applies to functions in $\mathcal{H}_n$. ∎

3.3. Interpolation on the hexagon. First we note that by (3.2), Theorem 2.6 when restricted to the hexagon domain becomes the following.

PROPOSITION 3.4. *For $n \ge 0$, define*

$$\mathcal{I}_n f(\mathbf{t}) := \sum_{\mathbf{j}\in\mathbb{H}_n} f(\tfrac{\mathbf{j}}{n})\Phi_n(\mathbf{t} - \tfrac{\mathbf{j}}{n}), \quad \text{where} \quad \Phi_n(\mathbf{t}) = \frac{1}{3n^2}\sum_{\mathbf{j}\in\mathbb{H}_n} \phi_{\mathbf{j}}(\mathbf{t}),$$

for $f \in C(\Omega)$. Then $\mathcal{I}_n f \in \mathcal{H}_n$ and

$$\mathcal{I}_n f(\tfrac{\mathbf{j}}{n}) = f(\tfrac{\mathbf{j}}{n}) \qquad \text{for all } \mathbf{j} \in \mathbb{H}_n.$$

Furthermore, and perhaps much more interesting, there exists another interpolation operator that works for almost all points in $\mathbb{H}_n^*$. First we need a few more definitions. We denote by $\partial\mathbb{H}_n^* := \mathbb{H}_n^* \setminus \mathbb{H}_n^\circ$ the set of points in $\mathbb{H}_n^*$ that are on the boundary of the hexagon $\{\mathbf{t} : -n \le t_1, t_2, t_3 \le n\}$. We further divide $\partial\mathbb{H}_n^*$ as $\mathbb{H}_n^{\mathrm{v}} \cup \mathbb{H}_n^{\mathrm{e}}$, where $\mathbb{H}_n^{\mathrm{v}}$ consists of the six vertices of $\partial\mathbb{H}_n$ and $\mathbb{H}_n^{\mathrm{e}}$ consists of the other points in $\partial\mathbb{H}_n$, i.e., the edge points. For $\mathbf{j} \in \partial\mathbb{H}_n^*$, we define

$$\mathcal{S}_{\mathbf{j}} := \left\{\mathbf{k} \in \mathbb{H}_n^* : \tfrac{\mathbf{k}}{n} \equiv \tfrac{\mathbf{j}}{n} \mod 3\right\}.$$

Because of the tiling property of the hexagon, if $\mathbf{j} \in \mathbb{H}_n^{\mathrm{e}}$, then $\mathcal{S}_{\mathbf{j}}$ contains two points, $\mathbf{j}$ and $\mathbf{j}^* \in \mathbb{H}_n^{\mathrm{e}}$, where j^* is on the opposite edge, relative to j, of the hexagon; while if $\mathbf{j} \in \mathbb{H}_n^{\mathrm{v}}$, then $\mathcal{S}_{\mathbf{j}}$ contains three vertices, $\mathbf{j}$ and its rotations by the angles $2\pi/3$ and $4\pi/3$.

THEOREM 3.5. *For $n \ge 0$ and $f \in C(\Omega)$, define*

$$\mathcal{I}_n^* f(\mathbf{t}) := \sum_{\mathbf{j}\in\mathbb{H}_n^*} f(\tfrac{\mathbf{j}}{n})\ell_{\mathbf{j},n}(\mathbf{t}),$$

where

$$\ell_{\mathbf{j},n}(\mathbf{t}) = \Phi_n^*(\mathbf{t} - \tfrac{\mathbf{j}}{n}) \quad \text{and} \quad \Phi_n^*(\mathbf{t}) = \frac{1}{3n^2}\sum_{\mathbf{j}\in\mathbb{H}_n^*} c_{\mathbf{j}}^{(n)}\phi_{\mathbf{j}}(\mathbf{t}). \tag{3.7}$$

Then $\mathcal{I}_n^* f \in \mathcal{H}_n^* := \{\phi_{\mathbf{j}} : \mathbf{j} \in \mathbb{H}_n^*\}$ and

$$\mathcal{I}_n^* f(\tfrac{\mathbf{j}}{n}) = \begin{cases} f(\tfrac{\mathbf{j}}{n}), & \mathbf{j} \in \mathbb{H}_n^\circ, \\ \sum_{\mathbf{k}\in\mathcal{S}_{\mathbf{j}}} f(\tfrac{\mathbf{k}}{n}), & \mathbf{j} \in \partial\mathbb{H}_n^*. \end{cases} \tag{3.8}$$

Furthermore, $\Phi_n^*(\mathbf{t})$ is a real function and has a compact formula

$$\begin{aligned}\Phi_n^*(\mathbf{t}) = \frac{1}{3n^2}\Bigg[&\frac{1}{2}\big(\Theta_n(\mathbf{t}) - \Theta_{n-2}(\mathbf{t})\big) \\ &-\frac{1}{3}\big(\cos\tfrac{2n\pi}{3}(t_1 - t_2) + \cos\tfrac{2n\pi}{3}(t_2 - t_3) + \cos\tfrac{2n\pi}{3}(t_3 - t_1)\big)\Bigg],\end{aligned} \tag{3.9}$$

where

$$\Theta_n(\mathbf{t}) := \frac{\sin\frac{(n+1)\pi(t_1-t_2)}{3}\sin\frac{(n+1)\pi(t_2-t_3)}{3}\sin\frac{(n+1)\pi(t_3-t_1)}{3}}{\sin\frac{\pi(t_1-t_2)}{3}\sin\frac{\pi(t_2-t_3)}{3}\sin\frac{\pi(t_3-t_1)}{3}}. \tag{3.10}$$

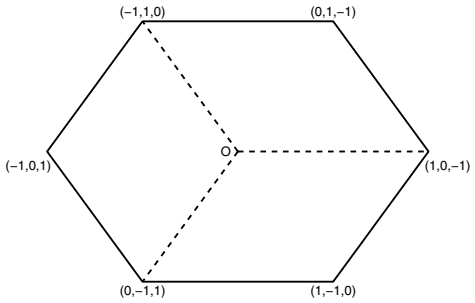


FIG. 4. *The hexagon partitioned into three parallelograms.*

Proof. First let $\mathbf{k} \in \mathbb{H}_n$. From the definition of the inner product $\langle \cdot, \cdot \rangle_n^*$, it follows that if $\mathbf{j} \in \mathbb{H}_n^\circ$, then $\ell_{\mathbf{j},n}(\frac{\mathbf{k}}{n}) = \langle \phi_{\mathbf{k}}, \phi_{\mathbf{j}} \rangle_n^* = \delta_{\mathbf{k},\mathbf{j}}$. If $\mathbf{j} \in \partial\mathbb{H}_n^*$, then one of the points in $\mathcal{S}_{\mathbf{j}}$ will be in $\mathbb{H}_n$, call it $\mathbf{j}^*$, and then $\ell_{\mathbf{j},n}(\frac{\mathbf{k}}{n}) = \delta_{\mathbf{k},\mathbf{j}^*}$. As the roles of $\mathbf{j}$ and $\mathbf{k}$ are symmetric in $\ell_{\mathbf{j},n}(\frac{\mathbf{k}}{n})$, this covers all cases. Putting these cases together, we have proved that

$$\ell_{\mathbf{j},n}(\tfrac{\mathbf{k}}{n}) = \Phi_n^*(\tfrac{\mathbf{k}}{n} - \tfrac{\mathbf{j}}{n}) = \begin{cases} 1 & \text{if } \frac{\mathbf{k}}{n} \equiv \frac{\mathbf{j}}{n} \mod 3, \\ 0 & \text{otherwise,} \end{cases} \qquad \mathbf{k}, \mathbf{j} \in \mathbb{H}_n^*, \tag{3.11}$$

from which the interpolation (3.8) follows immediately.

By the definition of $c_{\mathbf{j}}^{(m)}$, it is easy to see that

$$\Phi_n^*(\mathbf{t}) = \frac{1}{3n^2} \left[\frac{1}{2}\big(D_n(\mathbf{t}) + D_{n-1}(\mathbf{t})\big) - \frac{1}{6} \sum_{\mathbf{j} \in \mathbb{H}_n^{\mathrm{v}}} \phi_{\mathbf{j}}(\mathbf{t}) \right],$$

where D_n is the analogue of the Dirichlet kernel defined by

$$D_n(\mathbf{t}) := \sum_{\mathbf{j} \in \mathbb{H}_n^*} \phi_{\mathbf{j}}(\mathbf{t}). \tag{3.12}$$

By the definitions of $\phi_{\mathbf{k}}$ and $\mathbb{H}_n^{\mathrm{v}}$, we have

$$\sum_{\mathbf{j} \in \mathbb{H}_n^{\mathrm{v}}} \phi_{\mathbf{j}}(\mathbf{t}) = 2\left[\cos \tfrac{2n\pi}{3}(t_1 - t_2) + \cos \tfrac{2n\pi}{3}(t_2 - t_3) + \cos \tfrac{2n\pi}{3}(t_3 - t_1)\right],$$

so that we have to show only that $D_n = \Theta_n - \Theta_{n-1}$, which in fact has already appeared in [21]. Since an intermediate step is needed later, we outline the proof of the identity below. The hexagon domain can be partitioned into three parallelograms, as shown in Figure 4. Accordingly, we split the sum over $\mathbb{H}_n^*$ as

$$D_n(\mathbf{t}) = 1 + \sum_{j_2=-n}^{-1} \sum_{j_1=0}^{n} \phi_{\mathbf{j}}(\mathbf{t}) + \sum_{j_3=-n}^{-1} \sum_{j_2=0}^{n} \phi_{\mathbf{j}}(\mathbf{t}) + \sum_{j_1=-n}^{-1} \sum_{j_3=0}^{n} \phi_{\mathbf{j}}(\mathbf{t}). \tag{3.13}$$

Using the fact that $t_1 + t_2 + t_3 = 0$ and $j_1 + j_2 + j_3 = 0$, each of the three sums can be summed up explicitly. For example,

$$\sum_{j_2=-n}^{-1} \sum_{j_1=0}^{n} \phi_{\mathbf{j}}(\mathbf{t}) \tag{3.14}$$

$$= \frac{\mathrm{e}^{-\frac{\pi i}{3}(n+1)(t_3-t_2)} \sin \frac{n\pi}{3}(t_2 - t_3)}{\sin \frac{\pi}{3}(t_2 - t_3)} \frac{\mathrm{e}^{-\frac{\pi i}{3} n(t_1-t_3)} \sin \frac{(n+1)\pi}{3}(t_1 - t_3)}{\sin \frac{\pi}{3}(t_1 - t_3)}.$$

The formula $D_n = \Theta_n - \Theta_{n-1}$ comes from summing up these terms and carrying out simplification using the fact that if $t_1 + t_2 + t_3 = 0$, then

$$\begin{aligned} \sin 2t_1 + \sin 2t_2 + \sin 2t_3 &= -4 \sin t_1 \sin t_2 \sin t_3, \\ \cos 2t_1 + \cos 2t_2 + \cos 2t_3 &= 4 \cos t_1 \cos t_2 \cos t_3 - 1, \end{aligned} \tag{3.15}$$

which can be easily verified. □

The compact formula of the interpolation function allows us to estimate the operator norm of $\mathcal{I}_n^* : C(\overline{\Omega}) \mapsto C(\overline{\Omega})$ in the uniform norm, which is often referred to as the Lebesgue constant.

THEOREM 3.6. *Let $\|\mathcal{I}_n^*\|_\infty$ denote the operator norm of $\mathcal{I}_n^* : C(\overline{\Omega}) \mapsto C(\overline{\Omega})$. Then there is a constant c, independent of n, such that*

$$\|\mathcal{I}_n^*\|_\infty \le c(\log n)^2.$$

Proof. A standard procedure shows that

$$\|\mathcal{I}_n^*\|_\infty = \max_{\mathbf{t}\in\Omega} \sum_{\mathbf{j}\in\mathbb{H}_n^*} |\Phi_n^*(\mathbf{t} - \tfrac{\mathbf{j}}{n})|.$$

By (3.9) and the fact that $D_n = \Theta_n - \Theta_{n-1}$, it is sufficient to show that

$$\frac{1}{3n^2} \max_{\mathbf{t}\in\Omega} \sum_{\mathbf{j}\in\mathbb{H}_n^*} |D_n(\mathbf{t} - \tfrac{\mathbf{j}}{n})| \le c(\log n)^2, \qquad n \ge 0.$$

By (3.13) and (3.14), the estimate can be further reduced to show that

$$\begin{aligned} J_n(\mathbf{t}) &:= \frac{1}{3n^2} \sum_{\mathbf{j}\in\mathbb{H}_n^*} \left| \frac{\sin \frac{n\pi}{3}(t_2 - t_3 - \frac{j_2 - j_3}{n})}{\sin \frac{\pi}{3}(t_2 - t_3 - \frac{j_2 - j_3}{n})} \frac{\sin \frac{(n+1)\pi}{3}(t_1 - t_3 - \frac{j_1 - j_3}{n})}{\sin \frac{\pi}{3}(t_1 - t_3 - \frac{j_1 - j_3}{n})} \right| \\ &:= \sum_{\mathbf{j}\in\mathbb{H}_n^*} F_n(\mathbf{t} - \tfrac{\mathbf{j}}{n}) \end{aligned} \tag{3.16}$$

and two other similar sums obtained by permuting t_1, t_2, t_3 are bounded by $c(\log n)^2$ for all $\mathbf{t} \in \Omega$. The three sums are similar, so we will work only with (3.16).

Fix a $\mathbf{t} \in \Omega$. If $s_1 + s_2 + s_3 = 0$, then the equations $s_2 - s_3 = t_2 - t_3$ and $s_1 - s_3 = t_1 - t_3$ have the unique solutions $s_2 = t_2$ and $s_3 = t_3$. In particular, there can be at most one $\mathbf{j}$ such that the denominator of $F_n(\mathbf{t} - \frac{\mathbf{j}}{n})$ is zero. Using the well-known inequality $|\sin n\theta / \sin\theta| \le n$ shows that $F_n(\mathbf{t} - \frac{\mathbf{j}}{n}) \le n(n+1)/(3n^2) = \frac{1}{3}(1 + n^{-1})$ for all $\mathbf{j} \in \mathbb{H}_n^*$. We break the summation over $\mathbb{H}_n^*$ into several sums. The first sum is over those $\mathbf{j}$ such that

$$\rho_1(\mathbf{j}) := |t_1 - t_3 - \tfrac{j_1 - j_3}{n}| \quad \text{and} \quad \rho_2(\mathbf{j}) := |t_2 - t_3 - \tfrac{j_2 - j_3}{n}|$$

are both less than n^{-1}. There can be at most 6 such points, so that the sum is bounded by $2(1 + n^{-1})$. The second sum is over those $\mathbf{j}$ such that $\rho_1(\mathbf{j}) < n^{-1}$ but $\rho_2(\mathbf{j}) \ge n^{-1}$. Such $\mathbf{j}$'s are along a line $\rho_1(\mathbf{j}) = 0$, and there are at most $2n$ of them so that this sum is bounded by

$$\begin{aligned} \frac{n+1}{3n^2} \sum_{\rho_2(\mathbf{j})\ge n^{-1}} \left| \frac{\sin \frac{n\pi}{3}(t_2 - t_3 - \frac{j_2 - j_3}{n})}{\sin \frac{\pi}{3}(t_2 - t_3 - \frac{j_2 - j_3}{n})} \right| &\le \frac{c}{n} \sum_{\rho_2(\mathbf{j})\ge n^{-1}} \frac{1}{\rho_2(\mathbf{j})} \\ &\le c \sum_{k=1}^{2n} \frac{1}{k} \le c \log n, \end{aligned}$$

using the fact that $\sin\theta \ge (2/\pi)\theta$ for $0 \le \theta \le \pi/2$. The third sum is over those $\mathbf{j}$ such that $\rho_1(\mathbf{j}) \ge n^{-1}$ but $\rho_2(\mathbf{j}) < n^{-1}$, which can be estimated similarly. The last sum is over those $\mathbf{j}$ such that $\rho_1(\mathbf{j}) \ge n^{-1}$ and $\rho_2(\mathbf{j}) \ge n^{-1}$, which is, using once again the inequalities $\sin\theta \ge (2/\pi)\theta$, bounded by

$$\begin{aligned}\frac{c}{n^2} \sum_{\rho_1(\mathbf{j}),\rho_2(\mathbf{j})\ge n^{-1}} \frac{1}{\rho_1(\mathbf{j})\rho_2(\mathbf{j})} &\le \frac{c}{n^2} \sum_{\rho_1(\mathbf{j})\ge n^{-1}} \frac{1}{\rho_1(\mathbf{j})} \sum_{\rho_2(\mathbf{j})\ge n^{-1}} \frac{1}{\rho_2(\mathbf{j})} \\ &\le c \sum_{k=1}^{2n} \frac{1}{k} \sum_{k=1}^{2n} \frac{1}{k} \le c(\log n)^2.\end{aligned}$$

Putting these estimates together completes the proof. ∎

4. Discrete Fourier analysis on the triangle. Working with functions invariant under the isometries of the hexagon lattice, we can carry out a Fourier analysis on the equilateral triangle based on the analysis on the hexagon.

4.1. Generalized sine and cosine functions. The group $\mathcal{G}$ of isometries of the hexagon lattice is generated by the reflections in the edges of the equilateral triangles inside it, which is the reflection group $\mathcal{A}_2$. In homogeneous coordinates, the three reflections $\sigma_1, \sigma_2, \sigma_3$ are defined by

$$\mathbf{t}\sigma_1 := -(t_1, t_3, t_2), \quad \mathbf{t}\sigma_2 := -(t_2, t_1, t_3), \quad \mathbf{t}\sigma_3 := -(t_3, t_2, t_1).$$

Because of the relations $\sigma_3 = \sigma_1\sigma_2\sigma_1 = \sigma_2\sigma_1\sigma_2$, the group $\mathcal{G}$ is given by

$$\mathcal{G} = \{1, \sigma_1, \sigma_2, \sigma_3, \sigma_1\sigma_2, \sigma_2\sigma_1\}.$$

For a function f in homogeneous coordinates, the action of the group $\mathcal{G}$ on f is defined by $\sigma f(\mathbf{t}) = f(\mathbf{t}\sigma)$, $\sigma \in \mathcal{G}$. Following [9], we call a function f *invariant* under $\mathcal{G}$ if $\sigma f = f$ for all $\sigma \in \mathcal{G}$, and we call it *anti-invariant* under $\mathcal{G}$ if $\sigma f = \rho(\sigma) f$ for all $\sigma \in \mathcal{G}$, where $\rho(\sigma) = -1$ if $\sigma = \sigma_1, \sigma_2, \sigma_3$, and $\rho(\sigma) = 1$ for other elements in $\mathcal{G}$. The following proposition is easy to verify [9].

PROPOSITION 4.1. *Define two operators $\mathcal{P}^+$ and $\mathcal{P}^-$ acting on $f(\mathbf{t})$ by*

$$\mathcal{P}^{\pm} f(\mathbf{t}) = \frac{1}{6}\left[f(\mathbf{t}) + f(\mathbf{t}\sigma_1\sigma_2) + f(\mathbf{t}\sigma_2\sigma_1) \pm f(\mathbf{t}\sigma_1) \pm f(\mathbf{t}\sigma_2) \pm f(\mathbf{t}\sigma_3)\right]. \tag{4.1}$$

Then the operators $\mathcal{P}^+$ and $\mathcal{P}^-$ are projections from the class of H-periodic functions onto the class of invariant, respectively, anti-invariant, functions.

Recall that $\phi_{\mathbf{k}}(\mathbf{t}) = \mathrm{e}^{\frac{2\pi i}{3}\mathbf{k}\cdot\mathbf{t}}$. Following [21], we shall call the functions

$$\begin{aligned}\mathsf{TC}_{\mathbf{k}}(\mathbf{t}) := \mathcal{P}^+\phi_{\mathbf{k}}(\mathbf{t}) =& \frac{1}{6}\left[\phi_{k_1,k_2,k_3}(\mathbf{t}) + \phi_{k_2,k_3,k_1}(\mathbf{t}) + \phi_{k_3,k_1,k_2}(\mathbf{t})\right. \\ &\left. + \phi_{-k_1,-k_3,-k_2}(\mathbf{t}) + \phi_{-k_2,-k_1,-k_3}(\mathbf{t}) + \phi_{-k_3,-k_2,-k_1}(\mathbf{t})\right], \\ \mathsf{TS}_{\mathbf{k}}(\mathbf{t}) := \frac{1}{i}\mathcal{P}^-\phi_{\mathbf{k}}(\mathbf{t}) =& \frac{1}{6i}\left[\phi_{k_1,k_2,k_3}(\mathbf{t}) + \phi_{k_2,k_3,k_1}(\mathbf{t}) + \phi_{k_3,k_1,k_2}(\mathbf{t})\right. \\ &\left. - \phi_{-k_1,-k_3,-k_2}(\mathbf{t}) - \phi_{-k_2,-k_1,-k_3}(\mathbf{t}) - \phi_{-k_3,-k_2,-k_1}(\mathbf{t})\right]\end{aligned}$$

a *generalized cosine* and a *generalized sine*, respectively, where the second equal sign follows from the fact that $\phi_{\mathbf{k}}(\mathbf{t}\sigma) = \phi_{\mathbf{k}\sigma}(\mathbf{t})$ for every $\sigma \in \mathcal{G}$. It follows that $\mathsf{TC}_{\mathbf{k}}$ and $\mathsf{TS}_{\mathbf{k}}$ are invariant and anti-invariant, respectively.

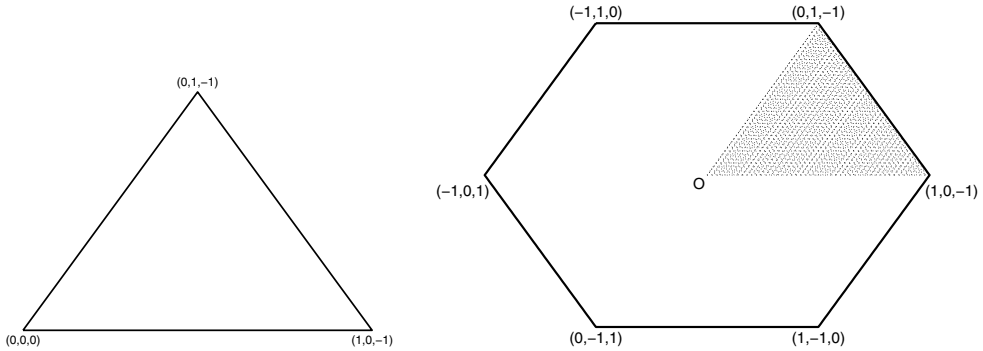


FIG. 5. *The equilateral triangle* Δ *in the hexagon.*

These two functions appear as eigenfunctions of the Laplacian on the triangle and have been studied in [9, 12, 13, 15, 16, 21, 22] under various coordinate systems. In particular, they are shown to share many properties of the classical sine and cosine functions in homogeneous coordinates in [21, 22]. Some of those properties that we shall need will be recalled below. The advantage of homogeneous coordinates lies in the symmetry of the formulas. For example, the equations $t_1 + t_2 + t_3 = 0$ and $k_1 + k_2 + k_3 = 0$ can be used to write $\mathsf{TC}_{\mathbf{k}}$ and $\mathsf{TS}_{\mathbf{k}}$ in several useful forms, such as

$$\begin{aligned}\mathsf{TC}_{\mathbf{k}}(\mathbf{t}) =& \frac{1}{3}\Big[\mathrm{e}^{\frac{i\pi}{3}(k_2-k_3)(t_2-t_3)}\cos k_1\pi t_1 \\ &+ \mathrm{e}^{\frac{i\pi}{3}(k_2-k_3)(t_3-t_1)}\cos k_1\pi t_2 + \mathrm{e}^{\frac{i\pi}{3}(k_2-k_3)(t_1-t_2)}\cos k_1\pi t_3\Big],\end{aligned} \tag{4.2}$$

$$\begin{aligned}\mathsf{TS}_{\mathbf{k}}(\mathbf{t}) =& \frac{1}{3}\Big[\mathrm{e}^{\frac{i\pi}{3}(k_2-k_3)(t_2-t_3)}\sin k_1\pi t_1 \\ &+ \mathrm{e}^{\frac{i\pi}{3}(k_2-k_3)(t_3-t_1)}\sin k_1\pi t_2 + \mathrm{e}^{\frac{i\pi}{3}(k_2-k_3)(t_1-t_2)}\sin k_1\pi t_3\Big],\end{aligned} \tag{4.3}$$

and similar formulas derived from the permutations of t_1, t_2, t_3. In particular, it follows from (4.3) that $\mathsf{TS}_{\mathbf{k}}(\mathbf{t}) \equiv 0$ whenever $\mathbf{k}$ contains one zero component.

For invariant functions, we can use symmetry to translate the results on the hexagon to one of its six equilateral triangles. We shall choose the triangle as

$$\begin{aligned}\Delta :=& \{(t_1,t_2,t_3) : t_1+t_2+t_3 = 0, 0 \le t_1, t_2, -t_3 \le 1\} \\ =& \{(t_1,t_2) : t_1, t_2 \ge 0,\, t_1+t_2 \le 1\}.\end{aligned} \tag{4.4}$$

The region Δ and its relative position in the hexagon is depicted in Figure 5. The inner product on Δ is defined by

$$\langle f, g\rangle_\Delta := \frac{1}{|\Delta|}\int_\Delta f(\mathbf{t})\overline{g(\mathbf{t})}d\mathbf{t} = 2\int_\Delta f(t_1,t_2)\overline{g(t_1,t_2)}dt_1dt_2.$$

If $f\bar{g}$ is invariant under $\mathcal{G}$, then it is easy to see that $\langle f, g\rangle_H = \langle f, g\rangle_\Delta$.

When $\mathsf{TC}_{\mathbf{k}}$ are restricted to the triangle Δ, we only need to consider a subset of $\mathbf{k} \in \mathbb{H}$. In fact, since $\mathsf{TC}_{\mathbf{k}\sigma}(\mathbf{t}) = \mathsf{TC}_{\mathbf{k}}(\mathbf{t}\sigma) = \mathsf{TC}_{\mathbf{k}}(\mathbf{t})$ for $\mathbf{t} \in \Delta$ and $\sigma \in \mathcal{G}$, we can restrict $\mathbf{k}$ to the index set

$$\Lambda := \{\mathbf{k} \in \mathbb{H} : k_1 \ge 0, k_2 \ge 0, k_3 \le 0\}. \tag{4.5}$$

We define the space $\mathcal{T}$ as the collection of finite linear combinations of elements in $\{\mathsf{TC}_{\mathbf{k}}, \mathsf{TS}_{\mathbf{k}} : \mathbf{k} \in \Lambda\}$; that is,

$$\mathcal{T} = \left\{ \sum_{\mathbf{k}\in\Xi} (a_{\mathbf{k}} \mathsf{TC}_{\mathbf{k}} + b_{\mathbf{k}} \mathsf{TS}_{\mathbf{k}}) : a_{\mathbf{k}}, b_{\mathbf{k}} \in \mathbb{C}, \Xi \subset \Lambda, \#\Xi < \infty \right\}. \tag{4.6}$$

Since $\mathsf{TS}_{\mathbf{k}}(\mathbf{t}) = 0$ whenever $\mathbf{k}$ has a zero component, $\mathsf{TS}_{\mathbf{k}}(\mathbf{t})$ are defined only for $\mathbf{k} \in \Lambda^\circ$, where

$$\Lambda^\circ := \{\mathbf{k} \in \mathbb{H} : k_1 > 0, k_2 > 0, k_3 < 0\},$$

which is the set of the interior points of Λ. These functions are orthogonal on Δ.

PROPOSITION 4.2. *For* $\mathbf{k}, \mathbf{j} \in \Lambda$,

$$\langle \mathsf{TC}_{\mathbf{k}}, \mathsf{TC}_{\mathbf{j}} \rangle_\Delta = \delta_{\mathbf{k},\mathbf{j}} \begin{cases} 1, & \mathbf{k} = 0, \\ \frac{1}{3}, & \mathbf{k} \in \Lambda \setminus \Lambda^\circ,\ \mathbf{k} \neq 0, \\ \frac{1}{6}, & \mathbf{k} \in \Lambda^\circ, \end{cases} \tag{4.7}$$

and for $\mathbf{k}, \mathbf{j} \in \Lambda^\circ$,

$$\langle \mathsf{TS}_{\mathbf{k}}, \mathsf{TS}_{\mathbf{j}} \rangle_\Delta = \frac{1}{6}\, \delta_{\mathbf{k},\mathbf{j}}. \tag{4.8}$$

This follows easily from $\langle f, g\rangle_H = \langle f, g\rangle_\Delta$, which holds if $f\bar{g}$ is invariant, the fact that $\mathsf{TC}_{\mathbf{k}}$ is invariant and $\mathsf{TS}_{\mathbf{k}}$ is anti-invariant, and the orthogonality of $\phi_{\mathbf{k}}$ in (3.3). The norms of $\mathsf{TC}_{\mathbf{k}}$ are divided into three cases, since $\mathsf{TC}_0(\mathbf{t}) = 1$ and

$$\mathsf{TC}_{\mathbf{k}}(\mathbf{t}) = \frac{1}{3}\left[\phi_{k_1,k_2,k_3}(\mathbf{t}) + \phi_{k_2,k_3,k_1}(\mathbf{t}) + \phi_{k_3,k_1,k_2}(\mathbf{t})\right], \qquad k_1k_2k_3 = 0.$$

4.2. Discrete inner product on the triangle. Using the fact that $\mathsf{TC}_{\mathbf{k}}$ and $\mathsf{TS}_{\mathbf{k}}$ are invariant and anti-invariant under $\mathcal{G}$ and the orthogonality of $\phi_{\mathbf{k}}$ with respect to the symmetric inner product $\langle\cdot,\cdot\rangle_n^*$, we can deduce a discrete orthogonality for the generalized cosine and sine functions. For this purpose, we define

$$\Lambda_n := \{\mathbf{k} \in \Lambda : -k_3 \le n\} = \{(k_1, k_2) \in \mathbb{Z}^2 : k_1 \ge 0, k_2 \ge 0, k_1 + k_2 \le n\},$$

and we denote by $\Lambda^\circ, \Lambda_n^{\mathrm{e}}, \Lambda_n^{\mathrm{v}}$ the set of points in the interior of Λ_n, on the boundary of Λ_n but not vertices, and at vertices, respectively; that is,

$$\begin{aligned} \Lambda_n^\circ &:= \{\mathbf{k} \in \Lambda_n : k_1 > 0, k_2 > 0, -k_3 < n\}, \\ \Lambda_n^{\mathrm{v}} &:= \{(n, 0, -n), (0, n, -n), (0, 0, 0)\}, \\ \Lambda_n^{\mathrm{e}} &:= \{\mathbf{k} \in \Lambda_n : \mathbf{k} \notin \Lambda_n^\circ \cup \Lambda^{\mathrm{v}}\}. \end{aligned}$$

We then define by $\mathcal{T}C_n$ and $\mathcal{T}S_n$ the subspace of $\mathcal{T}$ defined by

$$\mathcal{T}C_n = \operatorname{span}\{\mathsf{TC}_{\mathbf{k}} : \mathbf{k} \in \Lambda_n\} \quad \text{and} \quad \mathcal{T}S_n = \operatorname{span}\{\mathsf{TS}_{\mathbf{k}} : \mathbf{k} \in \Lambda_n^\circ\},$$

respectively.

THEOREM 4.3. *Let the discrete inner product* $\langle\cdot,\cdot\rangle_{\Delta,n}$ *be defined by*

$$\langle f, g\rangle_{\Delta,n} = \frac{1}{3n^2} \sum_{\mathbf{j}\in\Lambda_n} \lambda_{\mathbf{j}}^{(n)} f(\tfrac{\mathbf{j}}{n}) \overline{g(\tfrac{\mathbf{j}}{n})},$$

where

$$\lambda_{\mathbf{j}}^{(n)} := \begin{cases} 6, & \mathbf{j} \in \Lambda^\circ, \\ 3, & \mathbf{j} \in \Lambda^{\mathrm{e}}, \\ 1, & \mathbf{j} \in \Lambda^{\mathrm{v}}. \end{cases} \tag{4.9}$$

Then

$$\langle f, g\rangle_\Delta = \langle f, g\rangle_{\Delta,n}, \qquad f, g \in \mathcal{T}C_n. \tag{4.10}$$

Proof. We shall deduce the result from the discrete inner product $\langle \cdot, \cdot \rangle_n^*$ defined in (3.6). Let $\mathbf{j}\mathcal{G}$ denote the orbit of $\mathbf{j}$ under $\mathcal{G}$, which is the set $\{\mathbf{j}\sigma : \sigma \in \mathcal{G}\}$. It is easy to see that for $\mathbf{j} \in \mathbb{H}$,

$$|\mathbf{j}\mathcal{G}| = \begin{cases} 6 & \text{if } j_1 j_2 j_3 \neq 0, \\ 3 & \text{if } j_1 j_2 j_3 = 0 \text{ but } \mathbf{j} \neq (0,0,0), \\ 1 & \text{if } \mathbf{j} = (0,0,0). \end{cases} \tag{4.11}$$

Assume that f is invariant. Using (4.11), we see that

$$\begin{aligned} \sum_{\mathbf{j}\in\mathbb{H}_n^*, j_1 j_2 j_3 \neq 0} c_{\mathbf{j}}^{(n)} f(\tfrac{\mathbf{j}}{n}) &= 6 \sum_{\mathbf{j}\in\Lambda_n^\circ} c_{\mathbf{j}}^{(n)} f(\tfrac{\mathbf{j}}{n}) + 6 \sum_{\mathbf{j}\in\Lambda_n^{\mathrm{e}}, j_1 j_2 j_3 \neq 0} c_{\mathbf{j}}^{(n)} f(\tfrac{\mathbf{j}}{n}) \\ &= \sum_{\mathbf{j}\in\Lambda_n, j_1 j_2 j_3 \neq 0} \lambda_{\mathbf{j}}^{(n)} f(\tfrac{\mathbf{j}}{n}) \end{aligned}$$

and

$$\begin{aligned} \sum_{\mathbf{j}\in\mathbb{H}_n^*, j_1 j_2 j_3 = 0} c_{\mathbf{j}}^{(n)} f(\tfrac{\mathbf{j}}{n}) &= f(0,0,0) + 3 \sum_{j\in\Lambda^{\mathrm{e}}, j_1 j_2 j_3 = 0} c_{\mathbf{j}}^{(n)} f(\tfrac{\mathbf{j}}{n}) + 3 \sum_{j\in\Lambda^{\mathrm{v}}, \mathbf{j}\neq 0} c_{\mathbf{j}}^{(n)} f(\tfrac{\mathbf{j}}{n}) \\ &= \sum_{\mathbf{j}\in\Lambda_n, j_1 j_2 j_3 = 0} \lambda_{\mathbf{j}}^{(n)} f(\tfrac{\mathbf{j}}{n}). \end{aligned}$$

Replacing f by $f\bar{g}$ and adding these two sums, we have proved that $\langle f, g\rangle_n^* = \langle f, g\rangle_{\Delta,n}$ for invariant $f\bar{g}$, which applies to $f, g \in \mathcal{T}C_{n-1}$. Thus, the stated result follows immediately from Theorem 3.3 and the fact that $\langle f, g\rangle_H = \langle f, g\rangle_\Delta$. ∎

The proof of the above proposition also applies to $f, g \in \mathcal{T}S_n$ as $f\bar{g}$ is invariant if both f and g are anti-invariant. Notice also that $\mathcal{T}S_{\mathbf{j}}(\frac{\mathbf{j}}{n}) = 0$ when $\mathbf{j} \in \Lambda^{\mathrm{e}}$, and hence we conclude the following result.

PROPOSITION 4.4. *Let the discrete inner product $\langle \cdot, \cdot \rangle_{\Delta^\circ,n}$ be defined by*

$$\langle f, g\rangle_{\Delta^\circ,n} = \frac{2}{n^2} \sum_{\mathbf{j}\in\Lambda_n^\circ} f(\tfrac{\mathbf{j}}{n}) \overline{g(\tfrac{\mathbf{j}}{n})}.$$

Then

$$\langle f, g\rangle_\Delta = \langle f, g\rangle_{\Delta^\circ,n}, \qquad f, g \in \mathcal{T}S_n. \tag{4.12}$$

The relation (4.10) can be regarded as a cubature formula

$$\frac{1}{|\Delta|} \int_\Delta f(\mathbf{t}) d\mathbf{t} = \frac{1}{3n^2} \sum_{\mathbf{j}\in\Lambda_n} \lambda_{\mathbf{j}}^{(n)} f(\tfrac{\mathbf{j}}{n}),$$

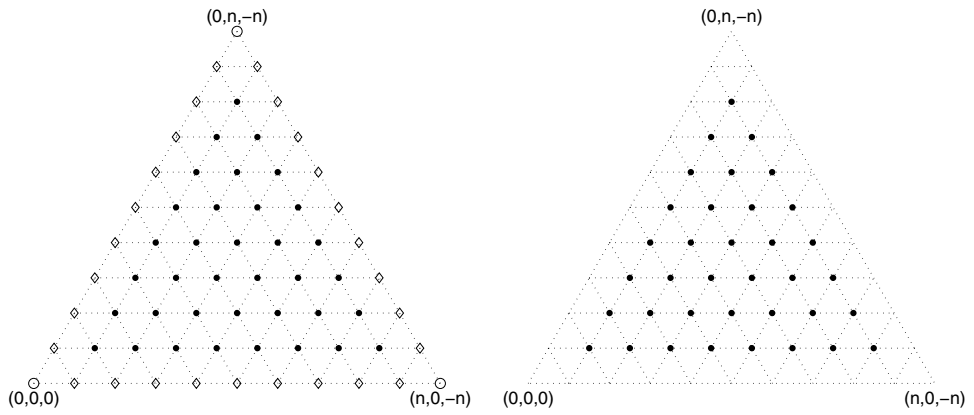


FIG. 6. *Left:* Λ_n *in which* $\bullet$, $\diamond$, *and* $\circ$ *correspond to points of different coefficients in* $\langle\cdot,\cdot\rangle_n^*$. *Right:* $\bullet$ *indicates points in* Λ°.

which is exact for $f = g\bar{h}$, $g, h \in \mathcal{TC}_n$. It turns out, however, that much more is true. We state the following result in terms of the triangle in (t_1, t_2).

THEOREM 4.5. *For* $n \ge 0$, *the cubature formula*

$$2\int_\Delta f(t_1,t_2)dt_1dt_2 = \frac{1}{3n^2}\sum_{j_1=0}^{n}\sum_{j_2=0}^{j_1}\lambda_{\mathbf{j}}^{(n)} f\left(\tfrac{j_1}{n}, \tfrac{j_2}{n}\right) \tag{4.13}$$

is exact for all $f \in \mathcal{TC}_{2n-1}$.

In order to establish such a result, we will need to study the structure of $\mathcal{TC}_n$, which, along with the proof of the theorem, will be given in section 5. One may note that the formulation of the result resembles a Gaussian quadrature. There is indeed a connection, which will also be explored in section 5.

4.3. Interpolation on the triangle. One way to deduce the result on interpolation is by making use of the orthogonality of generalized trigonometric functions with respect to the discrete inner product, as shown in the first theorem below. Recall the operator $\mathcal{P}^\pm$ defined in (4.1).

THEOREM 4.6. *For* $n \ge 0$ *and* $f \in C(\Delta)$ *define*

$$\mathcal{L}_n f(\mathbf{t}) := \sum_{\mathbf{j}\in\Lambda_n^\circ} f(\tfrac{\mathbf{j}}{n})\ell_{\mathbf{j},n}^\circ(\mathbf{t}), \qquad \ell_{\mathbf{j},n}^\circ(\mathbf{t}) = \frac{2}{n^2}\sum_{\mathbf{k}\in\Lambda_n^\circ} \mathsf{TS}_{\mathbf{k}}(\mathbf{t})\overline{\mathsf{TS}_{\mathbf{k}}(\tfrac{\mathbf{j}}{n})}.$$

Then $\mathcal{L}_n$ *is the unique function in* $\mathcal{TS}_n$ *that satisfies*

$$\mathcal{L}_n f(\tfrac{\mathbf{j}}{n}) = f(\tfrac{\mathbf{j}}{n}), \qquad \mathbf{j} \in \Lambda_n^\circ.$$

Furthermore, the fundamental interpolation function $\ell_{\mathbf{j},n}^\circ$ *is real and satisfies*

$$\ell_{\mathbf{j},n}^\circ(\mathbf{t}) = \frac{1}{3n^2}\mathcal{P}_{\mathbf{t}}^-\left[\Theta_{n-1}(\mathbf{t}-\tfrac{\mathbf{j}}{n}) - \Theta_{n-2}(\mathbf{t}-\tfrac{\mathbf{j}}{n})\right],$$

where $\mathcal{P}_{\mathbf{t}}^-$ *means that the operator* $\mathcal{P}^-$ *is acting on the the variable* $\mathbf{t}$ *and* Θ_n *is defined in* (3.10).

Proof. By (4.8) and (4.12), $\langle \mathsf{TS}_{\mathbf{j}}, \mathsf{TS}_{\mathbf{k}}\rangle_{\Delta^\circ,n} = \frac{1}{6}\delta_{\mathbf{k},\mathbf{j}}$, which shows that $\ell_{\mathbf{j},n}^\circ(\frac{\mathbf{k}}{n}) = \delta_{\mathbf{k},\mathbf{j}}$ and verifies the interpolation condition. Furthermore, since $\mathsf{TS}_{\mathbf{k}}(\frac{\mathbf{j}}{n}) = \mathsf{TS}_{\mathbf{j}}(\frac{\mathbf{k}}{n}) = 0$

whenever $\mathbf{k} \in \Lambda_n \setminus \Lambda_n^\circ$, and $\mathsf{TS}_{\mathbf{j}}(\mathbf{t}) = 0$ whenever $j_1 j_2 j_3 = 0$, it follows from the definition of $\mathsf{TS}_{\mathbf{j}}$ that

$$\ell^\circ_{\mathbf{j},n}(\mathbf{t}) = \frac{1}{3n^2}\mathcal{P}^-_{\mathbf{t}}\mathcal{P}^-_{\mathbf{j}} \sum_{\mathbf{k}\in\mathbb{H}^\circ_n} \phi_{\mathbf{k}}(\mathbf{t})\overline{\phi_{\mathbf{k}}(\tfrac{\mathbf{j}}{n})} = \frac{1}{3n^2}\mathcal{P}^-_{\mathbf{t}}\mathcal{P}^-_{\mathbf{j}} D_{n-1}(\mathbf{t} - \tfrac{\mathbf{j}}{n}),$$

where D_n is defined in (3.12). To complete the proof we use the fact that if F is an invariant function, then

$$\mathcal{P}^-_{\mathbf{t}}\mathcal{P}^-_{\mathbf{s}} F(\mathbf{t} - \mathbf{s}) = \mathcal{P}^-_{\mathbf{t}} F(\mathbf{t} - s),$$

which can be easily verified by using the fact that the elements of the group $\mathcal{G}$ satisfy $\sigma_i^2 = 1$, $i = 1, 2, 3$, and $\sigma_3 = \sigma_1\sigma_2\sigma_1 = \sigma_2\sigma_1\sigma_2$. □

We can derive the result of interpolation on the Λ_n using the same approach, but it is more illustrative to derive it from the interpolation on the hexagon given in Theorem 3.5.

THEOREM 4.7. *For $n \ge 0$ and $f \in C(\Delta)$ define*

$$\mathcal{L}^*_n f(\mathbf{t}) := \sum_{\mathbf{j}\in\Lambda_n} f(\tfrac{\mathbf{j}}{n})\ell^\Delta_{\mathbf{j},n}(\mathbf{t}), \qquad \ell^\Delta_{\mathbf{j},n}(\mathbf{t}) = \lambda^{(n)}_{\mathbf{j}}\mathcal{P}^+\ell_{\mathbf{j},n}(\mathbf{t}), \tag{4.14}$$

where $\lambda^{(n)}_{\mathbf{j}}$ and $\ell_{\mathbf{j},n}$ are defined in (4.9) *and* (3.7)*, respectively. Then $\mathcal{L}^*_n$ is the unique function in $\mathcal{T}C_n$ that satisfies*

$$\mathcal{L}^*_n f(\tfrac{\mathbf{j}}{n}) = f(\tfrac{\mathbf{j}}{n}), \qquad \mathbf{j} \in \Lambda_n.$$

Furthermore, the fundamental interpolation function $\ell_{\mathbf{j},n}$ is given by

$$\ell^\Delta_{\mathbf{j},n}(\mathbf{t}) = \frac{\lambda^{(n)}_{\mathbf{j}}}{3n^2}\left[\frac{1}{2}\mathcal{P}^+\big(\Theta_n(\mathbf{t} - \tfrac{\mathbf{j}}{n}) - \Theta_{n-2}(\mathbf{t} - \tfrac{\mathbf{j}}{n})\big) - \Re\left\{\mathsf{TC}_{n,0,-n}(\mathbf{t})\overline{\mathsf{TC}_{n,0,-n}(\tfrac{\mathbf{j}}{n})}\right\}\right].$$

Proof. Throughout this proof, write $\ell_{\mathbf{j}} = \ell_{\mathbf{j},n}$. Recall (3.11), which states that $\ell_{\mathbf{j}}(\frac{\mathbf{k}}{n}) = 1$ if $\mathbf{k} \equiv \mathbf{j} \mod 3$ and $\ell_{\mathbf{j}}(\frac{\mathbf{k}}{n}) = 0$ otherwise. Let $\mathbf{k} \in \Lambda$. If $\mathbf{j} \in \Lambda^\circ$, then we have $\mathcal{P}^+\ell_{\mathbf{j}}(\frac{\mathbf{k}}{n}) = \frac{1}{6}\ell_{\mathbf{j}}(\frac{\mathbf{k}}{n}) = [\lambda^{(n)}_{\mathbf{j}}]^{-1}\delta_{\mathbf{k},\mathbf{j}}$. If $\mathbf{j} \in \Lambda^{\mathrm{e}}$ and $j_1 j_2 j_3 \ne 0$, then $\mathcal{P}^+\ell_{\mathbf{j}}(\frac{\mathbf{k}}{n}) = \frac{1}{6}[\ell_{\mathbf{j}}(\frac{\mathbf{k}}{n}) + \ell_{\mathbf{j}}(\frac{\mathbf{k}^*}{n})] = \frac{1}{3}\delta_{\mathbf{k},\mathbf{j}} = [\lambda^{(n)}_{\mathbf{j}}]^{-1}\delta_{\mathbf{k},\mathbf{j}}$, where k^* is defined as in the proof of Theorem 3.5. If $\mathbf{j} \in \Lambda^{\mathrm{e}}$ and $j_1 j_2 j_3 = 0$, then $\mathcal{P}^+\ell_{\mathbf{j}}(\frac{\mathbf{k}}{n}) = \frac{1}{3}\ell_{\mathbf{j}}(\frac{\mathbf{k}}{n}) = [\lambda^{(n)}_{\mathbf{j}}]^{-1}\delta_{\mathbf{k},\mathbf{j}}$ since $\mathbf{j} \in \mathbb{H}^\circ_n$. If $\mathbf{j} \in \Lambda^{\mathrm{v}}$ and $\mathbf{j} \ne (0,0,0)$, then $\mathcal{P}^+\ell_{\mathbf{j}}(\frac{\mathbf{k}}{n}) = \frac{1}{3}\sum_{\mathbf{l}\in\mathcal{S}_{\mathbf{j}}}\ell_{\mathbf{l}}(\frac{\mathbf{k}}{n}) = \delta_{\mathbf{j},\mathbf{k}} = [\lambda^{(n)}_{\mathbf{j}}]^{-1}\delta_{\mathbf{k},\mathbf{j}}$, where $\mathcal{S}_{\mathbf{j}}$ is defined as in the proof of Theorem 3.5. Finally, if $\mathbf{j} = (0,0,0)$, $\mathcal{P}^+\ell_{\mathbf{j}}(\frac{\mathbf{k}}{n}) = \delta_{\mathbf{k},\mathbf{j}} = [\lambda^{(n)}_{\mathbf{j}}]^{-1}\delta_{\mathbf{k},\mathbf{j}}$. Putting these together, we have proved that $\ell^\Delta_{\mathbf{j}}(\frac{\mathbf{k}}{n}) = \delta_{\mathbf{k},\mathbf{j}}$, which verifies the interpolation condition.

The explicit formula of $\ell^\Delta_{\mathbf{j}}$ follows from (3.9) and the fact that

$$\begin{aligned}\mathcal{P}^+ \sum_{\mathbf{k}\in\mathbb{H}^{\mathrm{v}}_n} \phi_{\mathbf{k}}(\mathbf{t} - \tfrac{\mathbf{j}}{n}) &= \mathcal{P}^+ \sum_{\mathbf{k}\in\mathbb{H}^{\mathrm{v}}_n} \phi_{\mathbf{k}}(\mathbf{t})\overline{\phi_{\mathbf{k}}(\tfrac{\mathbf{j}}{n})} = \sum_{\mathbf{k}\in\mathbb{H}^{\mathrm{v}}_n} \mathsf{TC}_{\mathbf{k}}(\mathbf{t})\overline{\phi_{\mathbf{k}}(\tfrac{\mathbf{j}}{n})} \\ &= 3\mathsf{TC}_{n,0,-n}(\mathbf{t})\overline{\mathsf{TC}_{n,0,-n}(\tfrac{\mathbf{j}}{n})} + 3\mathsf{TC}_{0,n,-n}(\mathbf{t})\overline{\mathsf{TC}_{0,n,-n}(\tfrac{\mathbf{j}}{n})},\end{aligned}$$

as well as the fact that $\overline{\mathsf{TC}_{0,n,-n}(\mathbf{t})} = \mathsf{TC}_{n,0,-n}(\mathbf{t})$. □

As a consequence of Theorem 3.6, we immediately have the following.

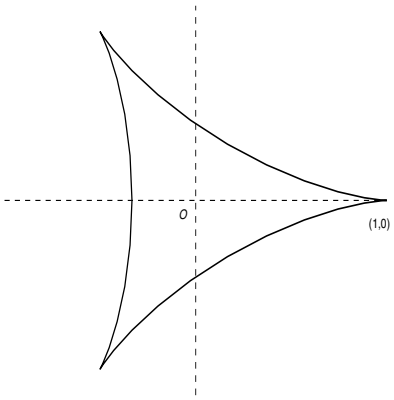


FIG. 7. *The region $\triangle^*$ bounded by Steiner's hypocycloid.*

THEOREM 4.8. *Let $\|\mathcal{L}_n\|$ and $\|\mathcal{L}_n^*\|$ denote the operator norm of $\mathcal{L}_n : C(\Delta) \mapsto C(\Delta)$ and $\mathcal{L}_n^* : C(\Delta) \mapsto C(\Delta)$, respectively. Then there is a constant c, independent of n, such that*

$$\|\mathcal{L}_n\| \le c(\log n)^2 \quad \text{and} \quad \|\mathcal{L}_n^*\| \le c(\log n)^2.$$

In particular, if $f \in C^1(\Omega)$, the class of functions having continuous derivatives on Ω, then both $\mathcal{L}_n f$ and $\mathcal{L}_n^* f$ converge uniformly to f on Ω.

This theorem shows a sharp contrast between interpolation by trigonometric functions on Λ_n or Λ_n° and the interpolation by algebraic polynomials on Λ_n. From the result in one variable [17], the Lebesgue constant for algebraic polynomial interpolation grows exponentially as $n \to \infty$ and, as a consequence, the convergence does not hold in $C^1(\Omega)$.

5. Generalized Chebyshev polynomials and their zeros. Just as in the classical case of one variable, we can use the generalized sine and cosine functions to define analogues of Chebyshev polynomials of the first and the second kind, respectively, which are orthogonal polynomials of the two variables.

5.1. Generalized Chebyshev polynomials. To see the polynomial structure among the generalized trigonometric functions, we make a change of variables. Denote

$$z := \mathsf{TC}_{0,1,-1}(\mathbf{t}) = \tfrac{1}{3}[\phi_{0,1,-1}(\mathbf{t}) + \phi_{1,-1,0}(\mathbf{t}) + \phi_{-1,0,1}(\mathbf{t})], \tag{5.1}$$

whose real part and imaginary part, after simplification using (3.15), are

$$\begin{aligned} x &= \Re z = \tfrac{4}{3}\cos\tfrac{\pi}{3}(t_2 - t_1)\cos\tfrac{\pi}{3}(t_3 - t_2)\cos\tfrac{\pi}{3}(t_1 - t_3) - \tfrac{1}{3}, \\ y &= \Im z = \tfrac{4}{3}\sin\tfrac{\pi}{3}(t_2 - t_1)\sin\tfrac{\pi}{3}(t_3 - t_2)\sin\tfrac{\pi}{3}(t_1 - t_3), \end{aligned} \tag{5.2}$$

respectively. If we change variables $(t_1, t_2) \mapsto (x, y)$, then the region Δ is mapped onto the region Δ^* bounded by Steiner's hypocycloid (see Figure 7),

$$\Delta^* = \{(x, y) : -3(x^2 + y^2 + 1)^2 + 8(x^3 - 3xy^2) + 4 \ge 0\}. \tag{5.3}$$

DEFINITION 5.1. *Under the change of variables* (5.1)*, define*

$$\begin{aligned} T_k^m(z, \bar z) &:= \mathsf{TC}_{k,m-k,-m}(\mathbf{t}), \qquad 0 \le k \le m, \\ U_k^m(z, \bar z) &:= \frac{\mathsf{TS}_{k+1,m-k+1,-m-2}(\mathbf{t})}{\mathsf{TS}_{1,1,-2}(\mathbf{t})}, \qquad 0 \le k \le m. \end{aligned}$$

We will also use the notation $T_k^m(\mathbf{t})$ and $U_k^m(\mathbf{t})$ when it is more convenient to work with $\mathbf{t}$ variable.

These functions are in fact polynomials of degree m in the z and $\bar{z}$ variables since they both satisfy a simple recursive relation.

PROPOSITION 5.2. *Let P_k^m denote either T_k^m or U_k^m. Then*

$$P_{m-k}^m(z,\bar{z}) = \overline{P_k^m(z,\bar{z})}, \qquad 0 \le k \le m, \tag{5.4}$$

and they satisfy the recursion relation

$$P_k^{m+1}(z,\bar{z}) = 3zP_k^m(z,\bar{z}) - P_{k+1}^m(z,\bar{z}) - P_{k-1}^{m-1}(z,\bar{z}) \tag{5.5}$$

for $0 \le k \le m$ and $m \ge 1$, where we use

$$T_{-1}^m(z,\bar{z}) = T_1^{m+1}(z,\bar{z}), \quad T_{m+1}^m(z,\bar{z}) = T_m^{m+1}(z,\bar{z}),$$
$$U_{-1}^m(z,\bar{z}) = 0, \quad U_m^{m-1}(z,\bar{z}) = 0.$$

In particular, we have

$$T_0^0(z,\bar{z}) = 1, \quad T_0^1(z,\bar{z}) = z, \quad T_1^1(z,\bar{z}) = \bar{z},$$
$$U_0^0(z,\bar{z}) = 1, \quad U_0^1(z,\bar{z}) = 3z, \quad U_1^1(z,\bar{z}) = 3\bar{z}.$$

The recursion relation (5.5) follows from a straightforward computation. Together, (5.5) and (5.4) determine all P_k^m recursively, which show that both T_k^m and U_k^m are polynomials of degree m in z and $\bar{z}$. Furthermore, each family inherits an orthogonal relation from the generalized trigonometric functions, so that they are orthogonal polynomials of two variables. They were studied by Koornwinder [9, 10], who called them the generalized Chebyshev polynomials of the first and the second kind, respectively. Moreover, they are special cases of orthogonal polynomials with respect to the weighted inner product

$$\langle f, g\rangle_{w_\alpha} = c_\alpha \int_{\Delta^*} f(x,y)\overline{g(x,y)}w_\alpha(x,y)dxdy,$$

where w_α is the weight function defined in terms of the Jacobian of the change of variables (5.1),

$$w_\alpha(x,y) = \left|\frac{\partial(x,y)}{\partial(t_1,t_2)}\right|^{2\alpha} = \left[\frac{16}{27}\pi^2 \sin\pi t_1 \sin\pi t_2 \sin\pi(t_1+t_2)\right]^{2\alpha}$$
$$= \frac{4^\alpha}{27^\alpha}\pi^{2\alpha}\left[-3(x^2+y^2+1)^2 + 8(x^3-3xy^2) + 4\right]^\alpha,$$

and c_α is a normalization constant, $c_\alpha := 1/\int_\Delta w_\alpha(x,y)dxdy$. We have, in particular, that $c_{-\frac{1}{2}} = 2$ and $c_{\frac{1}{2}} = 81/(8\pi^4)$. Since the change of variables (5.1) shows immediately that

$$\frac{1}{|\Delta|}\int_\Delta f(\mathbf{t})d\mathbf{t} = c_{-\frac{1}{2}}\int_{\Delta^*} f(x,y)w_{-\frac{1}{2}}(x,y)dxdy,$$

it follows from the orthogonality of $\mathsf{TC_j}$ and $\mathsf{TS_j}$ that T_k^m and U_k^m are orthogonal polynomials with respect to $w_{-\frac{1}{2}}$ and $w_{\frac{1}{2}}$, respectively. Furthermore, from Proposition 4.2

we have

$$\langle T_k^m, T_j^m\rangle_{w_{-\frac{1}{2}}} = d_k^m \delta_{k,j}, \quad d_k^m := \begin{cases} 1, & m = k = 0, \\ \frac{1}{3}, & k = 0 \text{ or } k = m,\ m > 0, \\ \frac{1}{6}, & 1 \le k \le m-1,\ m > 0, \end{cases} \tag{5.6}$$

and

$$\langle U_k^m, U_j^m\rangle_{w_{\frac{1}{2}}} = \frac{1}{6}\delta_{k,j}, \qquad 0 \le j,k \le m. \tag{5.7}$$

Let $\widetilde{T}_k^m = \sqrt{d_k^m} T_k^m$ and $\widetilde{U}_k^m = \sqrt{6} U_k^m$. These are orthonormal polynomials. We introduce the notation

$$\mathbb{P}_m = \{P_0^m, P_1^m, \ldots, P_m^m\} \tag{5.8}$$

for either $P_k^m = \widetilde{T}_k^m$ or $\widetilde{U}_k^m$. The polynomials in $\mathbb{P}_n$ consist of an orthonormal basis for the space $\mathcal{V}_n$ of orthogonal polynomials of degree m. Evidently both T_k^m and U_k^m are complex polynomials. As a consequence of (5.4), however, it is easy to derive real orthogonal polynomials with respect to $w_{\pm\frac{1}{2}}(x,y)$ on Δ^*. In fact, let $\mathbb{P}_m$ be as in (5.8). Then the polynomials in the set $\mathbb{P}_m$ defined by

$$\begin{aligned} \mathbb{P}_{2m-1}^R(x,y) &:= \{\sqrt{2}\Re\{P_k^{2m-1}(z,\bar z)\}, \sqrt{2}\Im\{P_k^{2m-1}(z,\bar z)\}, 0 \le k \le \lfloor m-1 \rfloor\}, \\ \mathbb{P}_{2m}^R(x,y) &:= \{P_m^{2m}, \sqrt{2}\Re\{P_k^{2m}(z,\bar z)\}, \sqrt{2}\Im\{P_k^{2m}(z,\bar z)\}, 0 \le k \le \lfloor m \rfloor\} \end{aligned} \tag{5.9}$$

are real orthonormal polynomials and form a basis for $\mathcal{V}_n$ with $n = 2m-1$ or $2m$, respectively.

5.2. Zeros of Chebyshev polynomials and Gaussian cubature formula. Let w be a nonnegative weight function defined on a compact set Ω in $\mathbb{R}^2$. A cubature formula of degree $2n-1$ is a sum of point evaluations that satisfies

$$\int_\Omega f(x,y) w(x,y)\,dxdy = \sum_{j=1}^N \lambda_j f(x_j, y_j), \qquad \lambda_j \in \mathbb{R},$$

for every $f \in \Pi_{2n-1}^2$, where Π_n^2 denote the space of polynomials of degree at most n in two variables. The points (x_j, y_j) are called nodes of the cubature and the numbers λ_i are called weights. It is well known that a cubature formula of degree $2n-1$ exists only if $N \ge \dim \Pi_{n-1}^2$. A cubature that attains such a lower bound is called a Gaussian cubature.

Let $\mathbb{P}_n := \{P_k^n : 0 \le k \le n\}$ denote a basis of orthonormal polynomials of degree n with respect to $w(x,y)dxdy$. It is known that a Gaussian cubature exists if and only if $\mathbb{P}_n$ (every element in $\mathbb{P}_n$) has $\dim \Pi_{n-1}^2$ many common zeros, which are necessarily real and distinct. Furthermore, the weight λ_j in a Gaussian cubature is given by

$$\lambda_j = [K_{n-1}(x_j, y_j)]^{-1}, \qquad K_n(x,y) := \sum_{k=0}^n \sum_{j=0}^k P_j^k(x) P_j^k(y). \tag{5.10}$$

The definition of $K_n(x,y)$ shows that it is the reproducing kernel of Π_n^2 in $L^2(\mu)$, so that it is independent of the choice of the basis.

Unlike the case of one variable, however, Gaussian cubatures do not exist in general. We refer to [4, 14, 20] for results and further discussions. It is sufficient to say that Gaussian cubatures are rare. In fact, the family of weight functions identified in [18] for which the Gaussian cubature exist for all n remains the only known example up to now.

We now test the above theory on the generalized Chebyshev polynomials. Because of the relation (5.9), we can work with the zeros of the complex polynomials in $\mathbb{P}_n$. Furthermore, it follows from (5.4) that the reproducing kernel

$$K_n(z,w) := \sum_{k=0}^{n} \mathbb{P}_k^{\mathsf{tr}}(z,\bar{z})\mathbb{P}_k(w,\bar{w}) = \sum_{k=0}^{n}\sum_{j=0}^{k} P_j^k(z,\bar{z})P_j^k(w,\bar{w})$$

is real and, by (5.9), is indeed the reproducing kernel of Π_n^2.

For the Chebyshev polynomials of the first kind, the polynomials in $\mathbb{P}_n$ do not have common zeros in general. For example, in the case of $m = 2$, we have $T_0^2(z,\bar{z}) = 3z^2 - 2\bar{z}$ and $T_1^2(z,\bar{z}) = (3z\bar{z} - 1)/2$, so that the three real orthogonal polynomials of degree 2 can be given by

$$\left\{3(x^2 - y^2) - 2x, \quad 6xy + 2y, \quad 3(x^2 + y^2) - 1\right\},$$

which has no common zero at all. Therefore, there is no Gaussian cubature for the weigh function $w_{-\frac{1}{2}}$ on Δ^*.

The Chebyshev polynomials of the second kind, on the other hand, do have the maximum number of common zeros, so that a Gaussian cubature exists for $w_{\frac{1}{2}}(\mathbf{t})$. Let

$$X_n := \Lambda_{n+2}^\circ = \{(\tfrac{k_1}{n+2}, \tfrac{k_2}{n+2}) | k_1 \ge 1, k_2 \ge 1, k_1 + k_2 \le n+1\},$$

which is in the interior of Δ, and let X_n^* denote the image of X_n in Δ^* under the mapping (5.2).

THEOREM 5.3. *For the weight function $w_{\frac{1}{2}}$ on Δ^* a Gaussian cubature formula exists; that is,*

(5.11)
$$c_{\frac{1}{2}} \int_{\Delta^*} f(x,y) w_{\frac{1}{2}}(x,y)dxdy = \sum_{j_1=1}^{n+1} \sum_{j_2=1}^{n+1-j_1} \mu_j^{(n)} f(\tfrac{j_1}{n+2}, \tfrac{j_2}{n+2}) \quad \textit{for all } f \in \Pi_{2n-1},$$

where $j = (j_1, j_2)$,

$$\mu_j^{(n)} = \tfrac{32}{9(n+2)^2} \sin^2 \pi \tfrac{j_1}{n+2} \sin^2 \pi \tfrac{j_2}{n+2} \sin^2 \pi \tfrac{j_1+j_2}{n+2}.$$

Proof. The numerator of $U_k^n(\mathbf{t})$ is $\mathsf{TS}_{k+1,n-k+1,-n-2}(\mathbf{t})$, which is zero at all $\frac{j}{n+2}$ according to (4.3). The first equation of (3.15) shows that the denominator of $U_k^n(\mathbf{t})$ is

$$\mathsf{TS}_{1,1,-2}(\mathbf{t}) = \frac{4}{3} \sin \pi t_1 \sin \pi t_2 \sin \pi t_3, \tag{5.12}$$

which vanishes at the integer points on the boundary point of Δ. Consequently, $U_k^n(\mathbf{t})$ does not vanish on the boundary of Δ but vanishes on X_n. It is easy to see that the cardinality of X_n is $\dim \Pi_{n-1}^2$. Consequently, the corresponding $\mathbb{P}_n$ has $\dim \Pi_{n-1}^2$ common zeros in X_n^*, so that the Gaussian cubature with respect to $w_{\frac{1}{2}}$ exists, which

takes the form (5.11). To obtain the formula for $\mu_j^{(n)}$, we note that the change of variables (5.2) shows that (5.11) is the same as

$$\frac{1}{|\Delta|}\int_\Delta f(\mathbf{t})\left[\mathsf{TS}_{1,1,-2}(\mathbf{t})\right]^2 d\mathbf{t} = \frac{2}{(n+2)^2}\sum_{j\in\Lambda^\circ_{n+2}}\left[\mathsf{TS}_{1,1,-2}(\tfrac{\mathbf{j}}{n+2})\right]^2 f(\tfrac{\mathbf{j}}{n+2}),$$

where the last step follows from the fact that $\mathsf{TS}_{1,1,-2}(\mathbf{t})$ vanishes on the boundary of Δ. This is exactly the cubature (4.13) applied to the function $f(\mathbf{t})\left[\mathsf{TS}_{1,1,-2}(\mathbf{t})\right]^2$. ∎

We note that $w_{\frac12}$ is only the second example of a weight function for which a Gaussian cubature exists for all n. Furthermore, the example of $w_{-\frac12}$ shows that the existence of Gaussian cubature does not hold for all weight functions in the family of w_α. In [18], a family of weight functions depending on three parameters that are on a region bounded by a parabola and two lines is shown to ensure Gaussian cubature for all permissible parameters.

The set of nodes of a Gaussian cubature is poised for polynomial interpolation, which shows that there is a unique polynomial P in Π^2_{n-1} such that $P(\xi) = f(\xi)$ for all $\xi \in X_n^*$, where f is an arbitrary function on X_n^*. In fact the general theory (cf. [4]) immediately gives the following result.

PROPOSITION 5.4. *The unique interpolation polynomial of degree n on X_n^* is given by*

$$\mathcal{L}_n f(x) = \sum_{j_1=1}^{n+1}\sum_{j_2=1}^{n+1-j_1} f(\tfrac{j}{n})\mu_j^{(n)} K_n(\mathbf{t}, \tfrac{j}{n}), \tag{5.13}$$

where $x = (x_1, x_2)$ and $z(\mathbf{t}) = x_1 + ix_2$, and $\mathcal{L}_n f$ is of total degree $n-1$ in x.

Under the mapping (5.2), the above interpolation corresponds to a trigonometric interpolation on Λ°_{n+2}. Since $\mu_j^{(n)} = \frac{2}{(n+2)^2}|\mathsf{TS}_{1,1,-2}(\frac{\mathbf{j}}{n})|^2$, it follows readily that

$$\mu_j^{(n)} K_n(\mathbf{t}, \tfrac{j}{n}) = \frac{2}{(n+2)^2}\mathsf{TS}_{1,1,-2}(\tfrac{\mathbf{j}}{n})\sum_{\mathbf{k}\in\Lambda^\circ_{n+2}}\frac{\mathsf{TS}_{\mathbf{k}}(\mathbf{t})\mathsf{TS}_{\mathbf{k}}(\frac{\mathbf{j}}{n})}{\mathsf{TS}_{1,1,-2}(\mathbf{t})},$$

where $\mathbf{j} = (j_1, j_2, -j_1 - j_2)$. Consequently, the interpolation (5.13) is closely related to the trigonometric interpolation on the triangle given in Theorem 4.6.

5.3. Cubature formula and Chebyshev polynomials of the first kind. As discussed in the previous subsection, Gaussian cubature does not exist for $w_{-\frac12}$ on Δ^*. It turns out, however, that there is a Gauss–Lobatto-type cubature for this weight function whose nodes are the points in the set

$$Y_n := \Lambda_n = \left\{\left(\tfrac{k_1}{n}, \tfrac{k_2}{n}\right) | k_1 \ge 0, k_2 \ge 0, k_1 + k_2 \le n\right\},$$

which is located inside Δ. Again we let Y_n^* denote the image of Y_n in Δ^* under the mapping (5.2).

The cardinality of Y_n is $\dim \Pi^2_n$. According to a characterization of cubature formula in the language of ideals and varieties (see [25]), there will be a cubature formula of degree $2n-1$ based on Y_n^* if there are $n+2$ linearly independent polynomials of degree $n+1$ that vanish on Y_n^*. In other words, Y_n^* is the variety of an ideal generated by $n+2$ algebraic polynomials of degree $n+1$, which are necessarily quasi-orthogonal in the sense that they are orthogonal to all polynomials of degree $n-2$. Such an ideal and a basis are described in the following proposition.

PROPOSITION 5.5. *The set Y_n^* is the variety of the polynomial ideal*

$$\langle T_0^{n+1} - T_1^n, \quad T_k^{n+1} - T_{k-1}^{n-1} \quad (1 \le k \le n), \quad T_{n+1}^{n+1} - T_{n-1}^n \rangle. \tag{5.14}$$

Proof. The mapping (5.2) allows us to work with $T_k^n(\mathbf{t})$ on Y_n. Using the explicit formula for $\mathsf{TC}_{k,n-k,-n}$ in (4.2), we have

$$\begin{aligned} T_k^n(\mathbf{t}) = \frac{1}{3}\Big[& e^{\frac{i\pi}{3}(n-2k)(t_2-t_3)} \cos n\pi t_1 \\ & + e^{\frac{i\pi}{3}(n-2k)(t_3-t_1)} \cos n\pi t_2 + e^{\frac{i\pi}{3}(n-2k)(t_1-t_2)} \cos n\pi t_3 \Big]. \end{aligned}$$

Since $n+1-2k = n-1-2(k-1)$, the above formula implies that

$$\begin{aligned} T_k^{n+1}(\mathbf{t}) - T_{k-1}^{n-1}(\mathbf{t}) = \frac{-1}{3}\Big[& e^{\frac{i\pi}{3}(n-2k)(t_2-t_3)} \sin \pi t_1 \sin m\pi t_1 \\ & + e^{\frac{i\pi}{3}(n-2k)(t_3-t_1)} \sin \pi t_2 \sin n\pi t_2 \\ & + e^{\frac{i\pi}{3}(n-2k)(t_1-t_2)} \sin \pi t_3 \sin n\pi t_3 \Big], \end{aligned}$$

from which it follows immediately that $T_k^{n+1} - T_{k-1}^{n+1}$ vanishes on Y_n for $1 \le k \le n$. Furthermore, a tedious computation shows that

$$\begin{aligned} & T_k^n\left(\tfrac{j_1}{n}, \tfrac{j_2}{n}, \tfrac{-j_1-j_2}{n}\right) = \frac{-1}{6} e^{\frac{-2i\pi}{3n}(n+1)j_1 + j_2} \Big[(-2 + e^{2i\pi j_2} + e^{2i\pi(j_1+j_2)}) \\ & + e^{2i\pi \frac{j_2}{n}} (-2e^{2i\pi j_2} + e^{2i\pi(j_1+j_2)} + 1) + e^{2i\pi \frac{j_1+j_2}{n}} (-2e^{2i\pi j_1} + e^{2i\pi j_2} + 1) \Big], \end{aligned}$$

which is evidently zero whenever j_1 and j_2 are integers. Finally, we note that $T_{n+1}^{n+1} - T_{n-1}^n$ is the conjugate of $T_0^{n+1} - T_1^n$. This completes the proof. ∎

As mentioned before, this proposition already implies that a cubature formula of degree $2n-1$ based on the nodes of Y_n^* exists.

THEOREM 5.6. *For the weight function $w_{-\frac{1}{2}}$ on Δ^* the cubature formula*

$$c_{-\frac{1}{2}} \int_{\Delta^*} f(x,y) w_{-\frac{1}{2}}(x,y) dxdy = \frac{1}{3n^2} \sum_{k_1=0}^{n} \sum_{k_2=0}^{k_1} \lambda_k^{(n)} f\left(\tfrac{k_1}{n}, \tfrac{k_2}{n}\right) \quad \text{for all } f \in \Pi_{2n-1} \tag{5.15}$$

holds, where $\lambda_k^{(n)} = \lambda_{k,n-k,-n}^{(n)}$ with $\lambda_{\mathbf{j}}^{(n)}$ given by (4.9).

In fact, the change of variables back to Δ shows that the cubature (5.15) is exactly of the form (4.13), which gives the values of $\lambda_k^{(n)}$. In particular, this gives a proof of Theorem 4.5.

Another way of determining the cubature weight $\lambda_k^{(n)}$ is to work with the reproducing kernel and the polynomial ideal. This approach also gives another way of determining the formula for polynomial interpolation. Furthermore, it fits into a general framework (see [24] and the references therein), and the intermediate results are of independent interest. In the rest of this section, we follow through with this approach.

The recursive relation (5.5) for T_k^m can be conveniently written in matrix form. Recall that for T_k^m, the orthonormal polynomials are

$$P_0^m = \sqrt{3} T_0^m, \quad P_k^m = \sqrt{6} T_k^m \quad (1 \le k \le m-1), \quad P_m^m = \sqrt{3} T_m^m$$

for $m > 0$ and $P_0^0 = 1$. In the following we treat $\mathbb{P}_m$ defined in (5.8) as a column vector. The relation (5.4) shows an important relation

$$\overline{\mathbb{P}_m} = J_{m+1}\mathbb{P}_m, \qquad J_m = \begin{bmatrix} \bigcirc & & 1 \\ & \iddots & \\ 1 & & \bigcirc \end{bmatrix}. \tag{5.16}$$

Using this notation, we have the following three-term relations.

PROPOSITION 5.7. *Define $\mathbb{P}_{-1} = 0$. For $m \ge 0$,*

$$\begin{aligned} z\mathbb{P}_m &= A_m\mathbb{P}_{m+1} + B_m\mathbb{P}_m + C_m\mathbb{P}_{m-1}, \\ \bar{z}\mathbb{P}_m &= C_{m+1}^{\mathsf{tr}}\mathbb{P}_{m+1} + B_m^{\mathsf{tr}}\mathbb{P}_m + A_{m-1}^{\mathsf{tr}}\mathbb{P}_{m-1}, \end{aligned} \tag{5.17}$$

where A_m, B_m, and C_m are given by

$$A_m = \frac{1}{3}\begin{bmatrix} 1 & & & \bigcirc & 0 \\ & \ddots & & & \vdots \\ & & 1 & & \\ \bigcirc & & & \sqrt{2} & 0 \end{bmatrix}, \quad B_m = \frac{1}{3}\begin{bmatrix} 0 & \sqrt{2} & 0 & \dots & \bigcirc \\ 0 & & 1 & & \\ \vdots & & & \ddots & \\ 0 & \bigcirc & & & 1 \quad 0 \\ 0 & 0 & \dots & & 0 \quad \sqrt{2} \end{bmatrix},$$

and $C_m = J_m A_{m-1}^{\mathsf{tr}} J_{m-1}$.

In fact, the first relation is exactly (5.5). To verify the second relation, take the complex conjugate of the first equation and then use (5.16).

In the following we slightly abuse the notation by writing $\mathbb{P}_n(\mathbf{t})$ in place of $\mathbb{P}_n(z)$, when $z = z(\mathbf{t})$ is given by (5.1). Similarly, we write $K_n(\mathbf{t}, \mathbf{s})$ in place of $K_n(z, w)$ when $z = z(\mathbf{t})$ and $w = w(\mathbf{s})$. Just like the case of real orthogonal polynomials, the three-term relation implies a Christoph–Darboux-type formula (see [4]).

PROPOSITION 5.8. *For $m \ge 0$, we have*

$$(z - w)K_n(z, w) := \mathbb{P}_{n+1}^{\mathsf{tr}}(z)A_n^{\mathsf{tr}}\,\overline{\mathbb{P}_n(w)} - \mathbb{P}_n^{\mathsf{tr}}(z)C_{n+1}^{\mathsf{tr}}\overline{\mathbb{P}_{n+1}(w)}. \tag{5.18}$$

Next we incorporate the information on the ideal in (5.14) into the kernel function. The $n + 2$ polynomials in the ideal can be written as a set

$$\mathbb{Q}_{n+1} = \mathbb{P}_{n+1} - \Gamma_0\mathbb{P}_n - \Gamma_1\mathbb{P}_{n-1}, \tag{5.19}$$

which we also treat as a column vector below, where

$$\Gamma_0 = \begin{bmatrix} 0 & \frac{1}{\sqrt{2}} & 0 & \dots & 0 \\ \bigcirc & \cdots & \bigcirc & \cdots & \bigcirc \\ 0 & \dots & 0 & \frac{1}{\sqrt{2}} & 0 \end{bmatrix} \quad \text{and} \quad \Gamma_1 = \begin{bmatrix} 0 & \dots & 0 & \dots & 0 \\ \sqrt{2} & & & & \bigcirc \\ & 1 & & & \\ & & \ddots & & \\ & & & 1 & \\ \bigcirc & & & & \sqrt{2} \\ 0 & \dots & 0 & \dots & 0 \end{bmatrix}.$$

PROPOSITION 5.9. *For $n \ge 0$, define*

$$\Phi_n(\mathbf{t}, \mathbf{s}) := \frac{1}{2}\left[K_n(\mathbf{t}, \mathbf{s}) + K_{n-1}(\mathbf{t}, \mathbf{s})\right] - \frac{1}{2}\left[T_0^n(\mathbf{t})\overline{T_0^n(\mathbf{s})} + \overline{T_0^n(\mathbf{t})}T_0^n(\mathbf{s})\right],$$

and we again write $\Phi_n(z,w)$ *in place of* $\Phi_n(\mathbf{t},\mathbf{s})$ *for* $z = z(\mathbf{t})$ *and* $w = w(\mathbf{s})$. *Then*

$$(z-w)\Phi_n(z,w) = \mathbb{Q}^{\mathsf{tr}}_{n+1}(z)A^{\mathsf{tr}}_n M_n \overline{\mathbb{P}_n(w)} - \mathbb{P}^{\mathsf{tr}}_n(z)M_n C^{\mathsf{tr}}_{n+1}\overline{\mathbb{Q}_{n+1}(w)}, \tag{5.20}$$

where $M_n = \operatorname{diag}\{-1/3, 1/2, \ldots, 1/2, -1/3\}$ *is a diagonal matrix.*

Proof. Inserting $\mathbb{P}_{n+1} = \mathbb{Q}_{n+1} + \Gamma_0\mathbb{P}_n + \Gamma_1\mathbb{P}_{n-1}$ into (5.18) gives

$$\begin{aligned}(z-w)K_n(z,w) = {}& \mathbb{Q}^{\mathsf{tr}}_{n+1}(z)A^{\mathsf{tr}}_n\,\overline{\mathbb{P}_n(w)} - \mathbb{P}^{\mathsf{tr}}_n(z)C^{\mathsf{tr}}_{n+1}\overline{\mathbb{Q}_{n+1}(w)}\\ &+ \mathbb{P}^{\mathsf{tr}}_n(z)\Gamma^{\mathsf{tr}}_0 A^{\mathsf{tr}}_n\,\overline{\mathbb{P}_n(w)} - \mathbb{P}^{\mathsf{tr}}_n(z)C^{\mathsf{tr}}_{n+1}\Gamma_0\overline{\mathbb{P}_n(w)}\\ &+ \mathbb{P}^{\mathsf{tr}}_{n-1}(z)\Gamma^{\mathsf{tr}}_1 A^{\mathsf{tr}}_n\,\overline{\mathbb{P}_n(w)} - \mathbb{P}^{\mathsf{tr}}_n(z)C^{\mathsf{tr}}_{n+1}\Gamma_1\overline{\mathbb{P}_{n-1}(w)}.\end{aligned}$$

Let D_n be an $n\times n$ matrix. Then similarly we have

$$\begin{aligned}(z-w)\mathbb{P}^{\mathsf{tr}}_n(z)D_n\overline{\mathbb{P}_n(w)} = {}& \mathbb{Q}^{\mathsf{tr}}_{n+1}(z)A^{\mathsf{tr}}_n D_n\overline{\mathbb{P}_n(w)} - \mathbb{P}^{\mathsf{tr}}_n(z)D_nC^{\mathsf{tr}}_{n+1}\overline{\mathbb{Q}_{n+1}(w)}\\ &+ \mathbb{P}^{\mathsf{tr}}_n(z)(A_n\Gamma_0 + B_n)^{\mathsf{tr}}D_n\overline{\mathbb{P}_n(w)} - \mathbb{P}^{\mathsf{tr}}_n(z)D_n(C^{\mathsf{tr}}_{n+1}\Gamma_0 + B^{\mathsf{tr}}_n)\overline{\mathbb{P}_n(w)}\\ &+ \mathbb{P}^{\mathsf{tr}}_{n-1}(z)(A_n\Gamma_1 + C_n)^{\mathsf{tr}}D_n\overline{\mathbb{P}_n(w)} - \mathbb{P}^{\mathsf{tr}}_n(z)D_n(C^{\mathsf{tr}}_{n+1}\Gamma_1 + A^{\mathsf{tr}}_{n-1})\overline{\mathbb{P}_{n-1}(w)}.\end{aligned}$$

In order to prove the stated result, we take the difference of the two equations and choose D_n such that

$$\begin{aligned}&(A_n\Gamma_0 + B_n)^{\mathsf{tr}}D_n - D_n(C^{\mathsf{tr}}_{n+1}\Gamma_0 + B^{\mathsf{tr}}_n) = \Gamma^{\mathsf{tr}}_0 A^{\mathsf{tr}}_n - C^{\mathsf{tr}}_{n+1}\Gamma_1,\\ &(A_n\Gamma_1 + C_n)^{\mathsf{tr}}D_n = \Gamma^{\mathsf{tr}}_1 A^{\mathsf{tr}}_n, \qquad D_n(C^{\mathsf{tr}}_{n+1}\Gamma_1 + A^{\mathsf{tr}}_{n-1}) = C^{\mathsf{tr}}_{n+1}\Gamma_1.\end{aligned} \tag{5.21}$$

It turns out that the unique solution of the above three equations is given by $D_n = \operatorname{diag}\{2/3, 1/2, 1/2, \ldots, 1/2, 2/3\}$, which is completely determined by the last two equations and then verified to satisfy the first equation. As a consequence of this choice of D_n, we obtain

$$\begin{aligned}(z-w)\left[K_n(z,w) - \mathbb{P}^{\mathsf{tr}}_n(z)D_n\overline{\mathbb{P}(w)}\right] = {}& \mathbb{Q}^{\mathsf{tr}}_{n+1}(z)A^{\mathsf{tr}}_n(I-D_n)\overline{\mathbb{P}_n(w)}\\ &- \mathbb{P}^{\mathsf{tr}}_n(z)(I-D_n)C^{\mathsf{tr}}_{n+1}\overline{\mathbb{Q}_{n+1}(w)},\end{aligned}$$

where I denotes the identity matrix. Setting $M_n = I - D_n$ we see that the right-hand side agrees with the right-hand side of (5.20). Finally we note that

$$\begin{aligned}\mathbb{P}^{\mathsf{tr}}_n(z)D_n\overline{\mathbb{P}_n(w)} &= \frac{1}{2}\mathbb{P}^{\mathsf{tr}}_n(z)\overline{\mathbb{P}_n(w)} + \left[P^n_0(z)\overline{P^n_0(w)} + P^n_n(z)\overline{P^n_n(w)}\right]\\ &= \frac{1}{2}\left[K_n(z,w) - K_{n-1}(z,w)\right] + \frac{1}{2}\left[T^n_0(z)\overline{T^n_0(w)} + \overline{T^n_0(z)}T^n_0(z)\right],\end{aligned}$$

which completes the proof. ∎

As a consequence of the above proposition, it is readily seen that $\Phi_n(\frac{j}{n},\frac{k}{n}) = 0$ for $j \neq k$. Furthermore, we note that $\Phi_n(\mathbf{t},\mathbf{s})$ is in fact a real-valued function. Consequently, we have proved the following.

PROPOSITION 5.10. *The unique interpolation polynomial on* Y^*_n *is given by*

$$\mathcal{L}^*_n f(x,y) = \sum_{j_1=0}^{n}\sum_{j_2=0}^{n-j_1} f(\tfrac{j}{n})\frac{\Phi_n(\mathbf{t},\frac{j}{n})}{\Phi_n(\frac{j}{n},\frac{j}{n})},$$

where $z(\mathbf{t}) = x + iy$, and $\mathcal{L}_n^* f$ is of total degree n in x and y.

In fact, upon changing variables (5.1), this interpolation polynomial is exactly the trigonometric interpolation function $\mathcal{L}_n^* f(\mathbf{t})$ at (4.14). Here the uniqueness of the interpolation follows from the general theory. Moreover, integrating $\mathcal{L}_n^* f$ with respect to $w_{-\frac{1}{2}}$ over Δ^* yields the cubature formula (5.15) with

$$\lambda_j^{(n)} = \left[\Phi_n\left(\tfrac{j}{n}, \tfrac{j}{n}\right)\right]^{-1}, \tag{5.22}$$

which can be used to determine $\lambda_j^{(n)}$ if needed. The function $K_n(\mathbf{t}, \mathbf{s})$, and thus $\Phi_n(\mathbf{t}, \mathbf{s})$, enjoys a compact formula, which facilitates the computation of $\Phi_n(\frac{j}{n}, \frac{j}{n})$. The compact formula follows from the following identity.

PROPOSITION 5.11. *Let D_n be defined in* (3.12). *Then*

$$K_n(\mathbf{s}, \mathbf{t}) = \sum_{k=0}^{n} \sum_{j=0}^{k} \widetilde{T}_j^k(\mathbf{t}) \overline{\widetilde{T}_j^k(\mathbf{s})} = \mathcal{P}_t^+ D_n(\mathbf{t} - \mathbf{s}).$$

The identity can be established from the definition of T_j^k and D_n as in the proof of Theorem 4.6 upon using the fact that $\mathcal{P}_{\mathbf{t}}^+ \mathcal{P}_{\mathbf{s}}^+ F(\mathbf{t} - \mathbf{s}) = \mathcal{P}^+ F(\mathbf{t} - \mathbf{s})$. We omit the details. From this identity and (5.22), we immediately have a compact formula for the interpolation polynomial, whose trigonometric counterpart under the change of variables (5.2) is exactly the function $\ell_{\mathbf{j},n}^{\Delta}$ in Theorem 4.7.

Acknowledgment. The authors thank the referees for their meticulous and prudent comments.

REFERENCES

[1] V. V. ARESTOV AND E. E. BERDYSHEVA, *Turán's problem for positive definite functions with supports in a hexagon*, Proc. Steklov Inst. Math. 2001, Approximation Theory. Asymptotical Expansions, Suppl. 1, pp. S20–S29.

[2] J. H. CONWAY AND N. J. A. SLOANE, *Sphere Packings, Lattices and Groups*, 3rd ed., Springer, New York, 1999.

[3] D. E. DUDGEON AND R. M. MERSEREAU, *Multidimensional Digital Signal Processing*, Prentice–Hall, Englewood Cliffs, NJ, 1984.

[4] C. F. DUNKL AND Y. XU, *Orthogonal Polynomials of Several Variables*, Encyclopedia Math. Appl. 81, Cambridge University Press, Cambridge, UK, 2001.

[5] W. EBELING, *Lattices and Codes*, Vieweg, Braunschweig/Wiesbaden, 1994.

[6] B. FUGLEDE, *Commuting self-adjoint partial differential operators and a group theoretic problem*, J. Funct. Anal., 16 (1974), pp. 101–121.

[7] J. R. HIGGINS, *Sampling Theory in Fourier and Signal Analysis, Foundations*, Oxford Science Publications, New York, 1996.

[8] M. KOLOUNTZAKIS, *The study of translation tiling with Fourier analysis*, in Fourier Analysis and Convexity, Appl. Numer. Harmon. Anal., Birkhäuser Boston, Boston, MA, 2004, pp. 131–187.

[9] T. KOORNWINDER, *Orthogonal polynomials in two variables which are eigenfunctions of two algebraically independent partial differential operators*, Nederl. Acad. Wetensch. Proc. Ser. A77 = Indag. Math., 36 (1974), pp. 357–381.

[10] T. KOORNWINDER, *Two-variable analogues of the classical orthogonal polynomials*, in Theory and Applications of Special Functions, R. A. Askey, ed., Academic Press, New York, 1975, pp. 435–495.

[11] R. J. MARKS II, *Introduction to Shannon Sampling and Interpolation Theory*, Springer, New York, 1991.

[12] B. J. MCCARTIN, *Eigenstructure of the equilateral triangle, Part* I: *The Dirichlet problem*, SIAM Rev., 45 (2003), pp. 267–287.

[13] B. J. MCCARTIN, *Eigenstructure of the equilateral triangle, Part* II: *The Neumann problem*, Math. Prob. Eng., 8 (2002), pp. 517–539.

[14] I. P. Mysoviskikh, *Interpolatory Cubature Formulas*, Nauka, Moscow, 1981.
[15] M. A. Pinsky, *The eigenvalues of an equilateral triangle*, SIAM J. Math. Anal., 11 (1980), pp. 819–827.
[16] M. A. Pinsky, *Completeness of the eigenfunctions of the equilateral triangle*, SIAM J. Math. Anal., 16 (1985), pp. 848–851.
[17] T. J. Rivlin, *An Introduction to the Approximation of Functions*, Dover, New York, 1981.
[18] H. J. Schmid and Y. Xu, *On bivariable Gaussian cubature formulae*, Proc. Amer. Math. Soc., 122 (1994), pp. 833–842.
[19] I. H. Sloan and T. R. Osborn, *Multiple integration over bounded and unbounded regions*, J. Comput. Appl. Math., 17 (1987), pp. 181–196.
[20] A. Stroud, *Approximate Calculation of Multiple Integrals*, Prentice–Hall, Englewood Cliffs, NJ, 1971.
[21] J. Sun, *Multivariate Fourier series over a class of non tensor-product partition domains*, J. Comput. Math., 21 (2003), pp. 53–62.
[22] J. Sun and H. Li, *Generalized Fourier transform on an arbitrary triangular domain*, Adv. Comput. Math., 22 (2005), pp. 223–248.
[23] G. Szegő, *Orthogonal Polynomials*, 4th ed., Amer. Math. Soc. Colloq. Publ. 23, AMS, Providence, RI, 1975.
[24] Y. Xu, *On orthogonal polynomials in several variables*, in Special Functions, q-Series and Related Topics, Fields Inst. Commun. 14, AMS, Providence, RI, 1997, pp. 247–270.
[25] Y. Xu, *Polynomial interpolation in several variables, cubature formulae, and ideals*, Adv. Comput. Math., 12 (2000), pp. 363–376.
[26] A. Zygmund, *Trigonometric Series*, Cambridge University Press, Cambridge, UK, 1959.

SIAM J. NUMER. ANAL.
Vol. 46, No. 4, pp. 1682–1703
© 2008 Society for Industrial and Applied Mathematics

CONVERGENCE ANALYSIS OF HIGH ORDER ALGEBRAIC FRACTIONAL STEP SCHEMES FOR TIME-DEPENDENT STOKES EQUATIONS*

PAOLA GERVASIO†

Dedicated to Fausto Saleri

Abstract. In this paper we analyze the family of Yosida algebraic fractional step schemes proposed in [A. Quarteroni, F. Saleri, and A. Veneziani, *Comput. Methods Appl. Mech. Engrg.*, 188 (2000), pp. 505–526], [F. Saleri and A. Veneziani, *SIAM J. Numer. Anal.*, 43 (2005), pp. 174–194], and [P. Gervasio, F. Saleri, and A. Veneziani, *J. Comput. Phys.*, 214 (2006), pp. 347–365] when applied to time-dependent Stokes equations. Under suitable regularity assumptions on the data, splitting error estimates both for velocity and pressure are established. In particular we analyze the first three methods of this family, providing, respectively, convergence (of the fractional step solution towards the numerical solution achieved without any operator splitting) of orders 3/2, 5/2, 7/2 for the velocity and 1, 2, 3 for the pressure. Moreover a general way to set up higher-order schemes is proposed. The present analysis is carried out when spectral element methods are employed for space discretization.

Key words. algebraic fractional step methods, time-dependent Stokes equations, spectral element methods

AMS subject classifications. 65M12, 76D07, 76M22

DOI. 10.1137/070682800

1. Introduction.

The aim of this paper is to provide convergence estimates for the class of *Yosida* methods proposed in [12, 13, 15, 8] for the numerical solution of time-dependent Stokes equations in the primitive variables velocity and pressure.

Yosida schemes are algebraic fractional step schemes in which the incompressibility constraint is relaxed in order to earn computational efficiency, and they are based on an inexact LU factorization of the matrix $\mathcal{A}$ arising from the full discretization (in space and time) of Stokes equations. The basic Yosida scheme was introduced in [13]. The accuracy analysis, carried out in [12], shows that when the time discretization is based on a backward Euler method, Yosida splitting still maintains the first order accuracy in time both for velocity and pressure (with respect to appropriate norms).

Successively, in [15, 8] two improved versions of the basic Yosida method (named *Yosida*-3 and *Yosida*-4 in the present paper) have been proposed. A pressure correction step is introduced in these schemes, with the aim of increasing the accuracy in time for both velocity and pressure. A preliminary analysis of the consistency error has been carried out for *Yosida*-3 (resp., *Yosida*-4) in [15] (resp., [8]) proving that the perturbation on the discretization matrix $\mathcal{A}$ is of third (resp., fourth) order in time with respect to the time-step. However, a complete splitting error analysis, i.e., the convergence analysis of the fractional step solution towards the numerical solution achieved without any operator splitting, is still missing. The aim of the present paper is essentially to perform such analysis when the space discretization is based on spectral element methods of $\mathbb{Q}_N - \mathbb{Q}_{N-2}$ type [11, 1, 3, 4], but we also emphasize the derivation of Yosida methods from a more general point of view than that given in

*Received by the editors February 16, 2007; accepted for publication (in revised form) December 29, 2007; published electronically April 11, 2008.
http://www.siam.org/journals/sinum/46-4/68280.html
†Department of Mathematics, University of Brescia, 25133 Brescia, Italy (gervasio@ing.unibs.it).

[15, 8], suggesting the formulation of higher-order schemes. The main result of the paper is Theorem 3.7, in which we prove that the splitting error induced by Yosida schemes (when they are coupled to BDFq (with $q = 1,\ 2$) for time discretization of the time-dependent Stokes equations) behaves like $\Delta t^{p+3/2}$ (resp., Δt^{p+1}) for the velocity (resp., for the pressure), where $p = 0$ for basic *Yosida*, $p = 1$ for *Yosida*-3, and $p = 2$ for *Yosida*-4. The thesis of Theorem 3.7 when $q = 3, 4$ has not been proved yet. Nevertheless in this paper we consider the more general contest with $q = 1, \ldots, 4$ since numerical results relative to the choice $q = p + 2$ are encouraging and provide global approximation errors on the velocity of order q with respect to Δt. In the present paper no evidence is given to computational aspects concerned with Yosida methods, since the topic has been extensively discussed in [8].

This paper is organized as follows. In section 2 we recall notation and settings about Stokes equations, spectral elements discretization, and some useful properties on symmetric positive definite matrices. In section 3 we present the derivation of Yosida methods for the unsteady Stokes equations, perform the splitting error analysis when spectral element methods are used for space discretization, and give some perspective for extending the analysis to unsteady Navier–Stokes equations. In section 4 we present some numerical results corroborating the convergence analysis developed in section 3. Even if the theory presented in this paper holds in $\mathbb{R}^d$, with $d = 2, 3$, only numerical results for the case $d = 2$ are reported here.

2. Problem statement and settings. We consider time-dependent Stokes equations for Newtonian incompressible fluids in the velocity-pressure formulation. For any open bounded domain $\Omega \subset \mathbb{R}^d$ ($d = 2, 3$) with a Lipschitz boundary $\partial\Omega$, and a positive T fixed, given a solenoidal datum $\mathbf{u}_0 \in [H^1(\Omega)]^d$, an external field $\mathbf{f} \in [L^2(0, T; H^{-1}(\Omega))]^d$, and a boundary datum $\mathbf{g} \in [L^2(0, T; H^{1/2}(\partial\Omega))]^d$, we look for the velocity field $\mathbf{u} \in [L^2(0, T; H^1(\Omega))]^d$ and the pressure field $p \in L^2(0, T; L^2_0(\Omega))$ solutions of

$$\begin{cases} \dfrac{\partial \mathbf{u}}{\partial t} - \nu\Delta\mathbf{u} + \nabla p = \mathbf{f} & \text{in } \Omega \times (0, T), \\ \nabla \cdot \mathbf{u} = 0 & \text{in } \Omega \times (0, T), \\ \mathbf{u} = \mathbf{g} & \text{on } \partial\Omega \times (0, T), \\ \mathbf{u} = \mathbf{u}_0 & \text{in } \Omega \times \{0\}, \end{cases} \tag{2.1}$$

where $\nu > 0$ is the kinematic viscosity. It is well known that problem (2.1) admits a unique solution (see, e.g., [16]).

We approximate the time derivative by a *backward differentiation formula* (BDF) of order q. Given $\Delta t \in (0, T)$, we set $t^0 = 0$, $t^n = t^0 + n\Delta t$ (for any $n \geq 1$) and $N_T = \left[\frac{T}{\Delta t}\right]$; therefore, for any integer $n = n_0 (= q - 1), \ldots, N_T - 1$, we look for the solution $(\mathbf{u}^{n+1}, p^{n+1})$ of the system

$$\begin{cases} \dfrac{\beta_{-1}}{\Delta t}\mathbf{u}^{n+1} - \nu\Delta\mathbf{u}^{n+1} + \nabla p^{n+1} = \mathbf{f}^{n+1} + \displaystyle\sum_{j=0}^{q-1} \frac{\beta_j}{\Delta t}\mathbf{u}^{n-j} & \text{in } \Omega, \\ \nabla \cdot \mathbf{u}^{n+1} = 0 & \text{in } \Omega, \\ \mathbf{u}^{n+1} = \mathbf{g}^{n+1} & \text{on } \partial\Omega, \end{cases} \tag{2.2}$$

where $\mathbf{u}^0 = \mathbf{u}_0$ and β_j (for $j = -1, \ldots, q$) are the coefficients of BDF of order q. When BDF with order greater than one are used, initial data could be provided by suitable explicit schemes (e.g., Runge–Kutta) of the same order as the BDF used.

About the space discretization, we choose quadrilateral conforming spectral elements [11]. In order to overcome instabilities due to the mixed formulation of Stokes equations, we make use of the $\mathbb{Q}_N - \mathbb{Q}_{N-2}$ scheme with staggered grids [1], according to which local polynomials of degree N in each variable are used to approximate every component of the velocity field, and local polynomials of degree $N-2$ in each variable are used to approximate the pressure ($\mathbb{Q}_N$ denotes the space of polynomials of degree less than or equal to N in each variable). By this choice, the inf-sup condition is satisfied with a constant β which is proportional to $N^{(1-d)/2}$ [1].

We introduce a conformal, regular, and quasi-uniform (see, e.g., [14]) partition $\mathcal{T}_h$ of Ω in N_e quadrilaterals T_k such that

$$\overline{\Omega} = \bigcup_{k=1}^{N_e} \overline{T}_k, \quad \text{with} \quad h = \max_{T_k \in \mathcal{T}_h} h_k, \qquad h_k = \operatorname{diam}(T_k), \quad k = 1, \ldots, N_e.$$

Let us assume that every $T_k \in \mathcal{T}_h$ is the image of the reference square $\hat{T} = (-1,1)^2$ through a smooth invertible mapping $\mathbf{F}_k : \hat{T} \mapsto T_k$ with Jacobian J_{F_k} satisfying $\det J_{F_k}(\hat{\mathbf{x}}) > 0 \ \forall \hat{\mathbf{x}} \in \hat{T}$. We set

$$\mathbb{Q}_{\mathcal{H}}(\Omega) = \{v_{\mathcal{H}} \in \mathcal{C}^0(\overline{\Omega}) : \ v_{\mathcal{H}|T_k} \circ \mathbf{F}_k \in \mathbb{Q}_N \quad \forall T_k \in \mathcal{T}_h\} \tag{2.3}$$

and $v_{N,k} := v_{\mathcal{H}|T_k} \ \forall T_k \in \mathcal{T}_h$. The subscript $\mathcal{H}$ represents the discretization level and it stands for the couple (h, N). Given $u_{\mathcal{H}},\ v_{\mathcal{H}} \in \mathbb{Q}_{\mathcal{H}}(\Omega)$, we set

$$(u_{\mathcal{H}}, v_{\mathcal{H}})_{\mathcal{H},\Omega} = \sum_{k=1}^{N_e} (u_{N,k}, v_{N,k})_{N,T_k}, \tag{2.4}$$

where $(\cdot,\cdot)_{N,T_k}$ denotes the discrete inner product in $L^2(T_k)$, based on the Gauss–Lobatto–Legendre (GLL) quadrature formulas [3]. In each element T_k of the partition we define a local GLL grid of $(N+1)^2$ points and a local Gauss–Legendre (GL) grid of $(N-1)^2$ points. The last grid is staggered with respect to the former one and is internal to T_k. The unknowns of the discrete problem will be both the set of the velocity values on the GLL grid and the set of the pressure values on the GL grid [1]. By setting the finite-dimensional spaces

$$\begin{aligned} &\mathbf{V}_{\mathcal{H}} := [\mathbb{Q}_{\mathcal{H}}(\Omega)]^2, \qquad \mathbf{V}^0_{\mathcal{H}} := \left[\mathbb{Q}_{\mathcal{H}}(\Omega) \cap H^1_0(\Omega)\right]^2, \\ &Q_{\mathcal{H}} := \{q_{\mathcal{H}} \in L^2_0(\Omega) : \ q_{\mathcal{H}|T_k} \circ \mathbf{F}_k \in \mathbb{Q}_{N-2} \quad \forall T_k \in \mathcal{T}_h\}, \end{aligned}$$

we reformulate problem (2.1) following the Galerkin approach and replace all exact integrals in $L^2(\Omega)$ with GLL quadrature formulas; the resulting approach is named the *spectral element method with numerical integration* (SEM-NI) [4]. For any $t_n \in (0,T)$, we denote by $\tilde{\mathbf{g}}(t_n) \in [H^1(\Omega)]^2$ the extension of $\mathbf{g}(t_n)$ to Ω by any possible continuous operator from $[H^{1/2}(\partial\Omega)]^2$ to $[H^1(\Omega)]^2$; moreover, we denote its discrete counterpart by $\tilde{\mathbf{g}}^n_{\mathcal{H}} \in \mathbf{V}_{\mathcal{H}}$. At each time-step t_{n+1} (for $n = n_0, \ldots, N_T - 1$) we look for the numerical solution $(\mathbf{u}^{n+1}_{\mathcal{H}}, p^{n+1}_{\mathcal{H}}) \in \mathbf{V}_{\mathcal{H}} \times Q_{\mathcal{H}}$, satisfying $(\mathbf{u}^{n+1}_{\mathcal{H}} - \tilde{\mathbf{g}}^{n+1}_{\mathcal{H}}) \in \mathbf{V}^0_{\mathcal{H}}$ and

$$\begin{cases} \dfrac{\beta_{-1}}{\Delta t}(\mathbf{u}^{n+1}_{\mathcal{H}}, \mathbf{v}_{\mathcal{H}})_{\mathcal{H},\Omega} + \nu(\nabla \mathbf{u}^{n+1}_{\mathcal{H}}, \nabla \mathbf{v}_{\mathcal{H}})_{\mathcal{H},\Omega} \\ \qquad - (p^{n+1}_{\mathcal{H}}, \nabla \cdot \mathbf{v}_{\mathcal{H}})_{\mathcal{H},\Omega} = \left(\mathbf{f}^{n+1} + \displaystyle\sum_{j=0}^{q-1} \frac{\beta_j}{\Delta t} \mathbf{u}^{n-j}_{\mathcal{H}}, \mathbf{v}_{\mathcal{H}} \right)_{\mathcal{H},\Omega} & \forall \mathbf{v}_{\mathcal{H}} \in \mathbf{V}^0_{\mathcal{H}}, \\ (\nabla \cdot \mathbf{u}^{n+1}_{\mathcal{H}}, q_{\mathcal{H}})_{\mathcal{H},\Omega} = 0 & \forall q_{\mathcal{H}} \in Q_{\mathcal{H}}. \end{cases} \tag{2.5}$$

Therefore, we set N_v (resp., N_p) the total number of GLL (resp., GL) grid points in Ω. For any $\mathbf{u}_\mathcal{H} \in \mathbf{V}_\mathcal{H}$, we denote by $\mathbf{U} \in \mathbb{R}^{2N_v}$ the vector of the expansion coefficients of $\mathbf{u}_\mathcal{H}$ with respect to the Lagrange basis $\{\boldsymbol{\varphi}_j\}_{j=1}^{2N_v}$ (defined on the GLL grid) in $\mathbf{V}_\mathcal{H}$ and by $\mathsf{V}_G \subset \mathbb{R}^{2N_v}$ (resp., $\mathsf{V}_G^0 \subset \mathbb{R}^{2N_v}$) the set of the vectors $\mathbf{U}$ corresponding to functions $\mathbf{u}_\mathcal{H} \in \mathbf{V}_\mathcal{H}$ (resp., $\mathbf{V}_\mathcal{H}^0$). In a similar way, for any $p_\mathcal{H} \in Q_\mathcal{H}$ we denote by $\mathbf{P} \in \mathbb{R}^{N_p}$ the vector of the expansion coefficients of $p_\mathcal{H}$ with respect to the Lagrange basis $\{\eta_l\}_{l=1}^{N_p}$ (defined on the GL grid) in $Q_\mathcal{H}$ and by $\mathsf{Q}_G \subset \mathbb{R}^{N_p}$ the set of the arrays $\mathbf{P}$ corresponding to functions $p_\mathcal{H} \in Q_\mathcal{H}$.

We denote by $M \in \mathbb{R}^{2N_v \times 2N_v}$ the mass matrix $M_{ij} = (\boldsymbol{\varphi}_j, \boldsymbol{\varphi}_i)_{\mathcal{H},\Omega}$, by $K \in \mathbb{R}^{2N_v \times 2N_v}$ the stiffness matrix $K_{ij} = (\nabla\boldsymbol{\varphi}_j, \nabla\boldsymbol{\varphi}_i)_{\mathcal{H},\Omega}$, and by $B \in \mathbb{R}^{2N_v \times N_p}$ the matrix related to the discretization of $-\nabla\cdot$: $B_{lj} = -(\nabla \cdot \boldsymbol{\varphi}_j, \eta_l)_{\mathcal{H},\Omega}$, and we set

$$C = \frac{\beta_{-1}}{\Delta t} M + \nu K. \tag{2.6}$$

Remark 1. Matrices M, K, and C are symmetric and positive definite. Moreover, thanks to the fact that the space discretization chosen satisfies the inf-sup condition, the matrix B is a full rank matrix and $\ker(B^T) = \{\mathbf{0}\}$. Finally, we remark that when SEM-NI is used, the mass matrix M is diagonal.

The system is reduced to the unknowns internal to Ω, and the right-hand side is modified accordingly, taking into account the contributions that Lagrange basis functions associated to Dirichlet boundary nodes give to those associated to internal nodes. This step produces a right-hand side $[\mathbf{F}_1^{n+1}, \mathbf{F}_2^{n+1}]^T$ that is nonzero also in the continuity equation.

The algebraic form of (2.5) reads as follows: for $n = n_0, \ldots, N_T - 1$ solve

$$\begin{cases} C\mathbf{U}^{n+1} + B^T\mathbf{P}^{n+1} = \mathbf{F}_1^{n+1} + \dfrac{1}{\Delta t}\displaystyle\sum_{j=0}^{q-1} \beta_j M\mathbf{U}^{n-j}, \\ B\mathbf{U}^{n+1} = \mathbf{F}_2^{n+1}, \end{cases} \tag{2.7}$$

or equivalently, for $n = n_0, \ldots, N_T - 1$ solve

$$\mathcal{A}\mathbf{W}^{n+1} = \mathbf{G}^{n+1}, \qquad \text{with} \qquad \mathcal{A} = \begin{bmatrix} C & B^T \\ B & 0 \end{bmatrix} \tag{2.8}$$

and

$$\mathbf{W}^{n+1} = \begin{bmatrix} \mathbf{U}^{n+1} \\ \mathbf{P}^{n+1} \end{bmatrix}, \quad \mathbf{G}^{n+1} = \begin{bmatrix} \mathbf{G}_1^{n+1} \\ \mathbf{G}_2^{n+1} \end{bmatrix} := \begin{bmatrix} \mathbf{F}_1^{n+1} + \dfrac{1}{\Delta t}\displaystyle\sum_{j=0}^{q-1} \beta_j M\mathbf{U}^{n-j} \\ \mathbf{F}_2^{n+1} \end{bmatrix}.$$

System (2.8) could be solved by a global approach such as a preconditioned Krylov method with either algebraic or differential preconditioners. Alternatively, a block LU factorization of $\mathcal{A}$ can be performed with

$$L = \begin{bmatrix} C & 0 \\ B & -BC^{-1}B^T \end{bmatrix}, \qquad U = \begin{bmatrix} I & C^{-1}B^T \\ 0 & I \end{bmatrix}, \tag{2.9}$$

so that, for any $n = n_0, \ldots, N_T - 1$, system (2.7) reads also

$$
\begin{aligned}
&L\text{-step}: \text{ find } \mathbf{U}^{n+1/2},\ \mathbf{P}^{n+1/2}: && \begin{cases} C\mathbf{U}^{n+1/2} = \mathbf{G}_1^{n+1}, \\ \Sigma\mathbf{P}^{n+1/2} = \mathbf{G}_2^{n+1} - B\mathbf{U}^{n+1/2}, \end{cases} \\
&U\text{-step}: \text{ find } \mathbf{U}^{n+1},\ \mathbf{P}^{n+1}: && \begin{cases} \mathbf{P}^{n+1} = \mathbf{P}^{n+1/2}, \\ C(\mathbf{U}^{n+1/2} - \mathbf{U}^{n+1}) = B^T\mathbf{P}^{n+1}. \end{cases}
\end{aligned} \tag{2.10}
$$

The matrix

$$\Sigma := -BC^{-1}B^T \in \mathbb{R}^{N_p \times N_p} \tag{2.11}$$

is the so-called *pressure Schur complement* matrix and $-\Sigma > 0$. Solving system (2.8) by a block LU factorization offers the advantage of splitting the original problem into subproblems of smaller size, even if the assembling of Σ is, however, quite expensive. The idea of Yosida schemes consists in replacing the block LU factorization (2.9) with an inexact one or, equivalently, in replacing algorithm (2.10) with a more efficient one from the computational point of view. The next section will be devoted to present the derivation of Yosida schemes, starting from (2.8).

We are now going to introduce some notation and report some properties we will use in the next sections, referring to [17] and [9] for an exhaustive treatment of these subjects.

D1. Given a real square matrix $\mathcal{B}$, we write $\mathcal{B} > 0$ if it is symmetric positive definite (s.p.d.), while we write $\mathcal{B} \geq 0$ if it is symmetric semipositive definite. Moreover, if $\mathcal{B}$, $\mathcal{C} \in \mathbb{R}^{n\times n}$ are symmetric, we write $\mathcal{B} > \mathcal{C}$ if $\mathcal{B} - \mathcal{C} > 0$ and $\mathcal{B} \geq \mathcal{C}$ if $\mathcal{B} - \mathcal{C} \geq 0$.

D2. The eigenvalues of a matrix $\mathcal{B} \in \mathbb{R}^{n\times n}$ are denoted by $\lambda_i(\mathcal{B})$ and the *spectral radius* of $\mathcal{B}$ is defined as $\rho(\mathcal{B}) = \max\{|\lambda_i(\mathcal{B})|,\ i = 1, \ldots, n\}$.

D3. For any square matrix $\mathcal{M}$ of the form

$$\mathcal{M} = \begin{bmatrix} \mathcal{C} & \mathcal{B}^T \\ \mathcal{B} & \mathcal{D} \end{bmatrix}$$

the Schur complement of the nonsingular block $\mathcal{C}$ in $\mathcal{M}$ is defined as $\mathcal{M}/\mathcal{C} := \mathcal{D} - \mathcal{B}\mathcal{C}^{-1}\mathcal{B}^T$. Moreover, $\det(\mathcal{M}) = \det(\mathcal{C})\det(\mathcal{M}/\mathcal{C})$ (*Schur's formula* [17, Thm. 2.2]).

D4. The *inertia* $In(\mathcal{B})$ of a square matrix $\mathcal{B}$ is defined to be the ordered triplet (i_+, i_-, i_0), where i_+, i_-, and i_0 are the numbers of positive, negative, and zero real parts of the eigenvalues of $\mathcal{B}$.

D5. For any matrix $\mathcal{B} \in \mathbb{R}^{n\times n}$, whose coefficients depend on Δt, we say that

$$\mathcal{B} = \mathcal{O}(\Delta t^k) \qquad \text{if} \quad |||\mathcal{B}||| = \mathcal{O}(\Delta t^k) \qquad \text{for } \Delta t \to 0,$$

where $|||\cdot|||$ denotes the 2-norm for matrices.

P1. For any invertible matrices $\mathcal{B}$, $\mathcal{C} \in \mathbb{R}^{n\times n}$ it holds that

$$(\mathcal{B} + \mathcal{C})^{-1} = \mathcal{B}^{-1}[\mathcal{B} - (\mathcal{C}^{-1} + \mathcal{B}^{-1})^{-1}]\mathcal{B}^{-1}, \tag{2.12}$$

$$\mathcal{B} - \mathcal{C} = -\mathcal{B}(\mathcal{B}^{-1} - \mathcal{C}^{-1})\mathcal{B}, \tag{2.13}$$

$$\mathcal{B}(\mathcal{B} + \mathcal{C})^{-1}\mathcal{B} = \mathcal{B} - \mathcal{C} + \mathcal{C}(\mathcal{B} + \mathcal{C})^{-1}\mathcal{C}. \tag{2.14}$$

P2. As a consequence of the Courant–Fisher minimax theorem [17], for any real symmetric matrices $\mathcal{B}$, $\mathcal{C} \in \mathbb{R}^{n\times n}$ it holds that $\lambda_i(\mathcal{B}) + \lambda_{min}(\mathcal{C}) \leq \lambda_i(\mathcal{B} + \mathcal{C}) \leq \lambda_i(\mathcal{B}) + \lambda_{max}(\mathcal{C})$, $i = 1, \ldots, n$.

P3. If $\mathcal{B}$, $\mathcal{C} \in \mathbb{R}^{n\times n}$ are s.p.d. and $\mathcal{B}-\mathcal{C} > 0$, then $\mathcal{C}^{-1} - \mathcal{B}^{-1} > 0$.

P4. If $\mathcal{B}$, $\mathcal{C} \in \mathbb{R}^{n\times n}$ are symmetric and $\mathcal{C} > 0$, then the eigenvalues of $\mathcal{BC}$ are all real and $In(\mathcal{BC}) = In(\mathcal{B})$ (see, e.g., [5]).

P5. If a matrix is the product of two s.p.d. matrices, then it is similar to an s.p.d. matrix (this is a consequence of P4).

P6. For any s.p.d. matrix $\mathcal{C} \in \mathbb{R}^{n\times n}$ and $\mathcal{X} \in \mathbb{R}^{m\times n}$ with $\ker(\mathcal{X}^T) = \{\mathbf{0}\}$, it holds [18] that

$$\mathcal{C}^{-1} - \mathcal{X}^T(\mathcal{X}\mathcal{C}\mathcal{X}^T)^{-1}\mathcal{X} \geq 0. \tag{2.15}$$

3. Yosida schemes. Yosida schemes come basically from an inexact block LU factorization of matrix $\mathcal{A}$. The first method of this family was introduced in [12, 13] and consists in replacing the Schur complement Σ by a suitable matrix S.

In [15] and [8] two improved versions of the *Yosida* method (here named *Yosida*-3 and *Yosida*-4) were proposed. Such improved Yosida schemes introduce a correction step on the pressure with the aim of increasing the accuracy in time for both velocity and pressure. In particular, these improved Yosida schemes are defined by a new matrix Q, inside the U-step of (2.10) and acting on the pressure, which corrects the approximation of the Schur complement Σ. All is done at the algebraic level, without a compulsory differential interpretation. This feature allows us to neglect the setting up of boundary conditions for the substeps. All methods and the analysis developed hereafter can be applied to different kinds of boundary conditions as well.

The idea of Yosida schemes is as follows: at each time-step, instead of solving system (2.8) by the block LU factorization set in (2.9), we are interested in solving a new system

$$\hat{\mathcal{A}}\widehat{\mathbf{W}}^{n+1} = \widehat{\mathbf{G}}^{n+1}, \tag{3.1}$$

where $\hat{\mathcal{A}} := \hat{L}\hat{U}$ is generated by a suitable $\hat{L}\hat{U}$ inexact factorization of $\mathcal{A}$, $\widehat{\mathbf{W}}^{n+1}$ is an approximation of $\mathbf{W}^{n+1}$ defined as

$$\widehat{\mathbf{W}}^{n+1} = \begin{bmatrix} \hat{\mathbf{U}}^{n+1} \\ \hat{\mathbf{P}}^{n+1} \end{bmatrix}, \quad \text{while} \qquad \widehat{\mathbf{G}}^{n+1} = \begin{bmatrix} \mathbf{F}_1^{n+1} + \dfrac{1}{\Delta t}\displaystyle\sum_{j=0}^{q-1} \beta_j M \widehat{\mathbf{U}}^{n-j} \\ \mathbf{F}_2^{n+1} \end{bmatrix}. \tag{3.2}$$

Let $H = (\Delta t/\beta_{-1})M^{-1}$; the inexact factors $\hat{L}$ and $\hat{U}$ of $\mathcal{A}$ are chosen as follows:

$$\hat{L} = \begin{bmatrix} C & 0 \\ B & -BHB^T \end{bmatrix}, \quad \hat{U} = \begin{bmatrix} I & C^{-1}B^T \\ 0 & Q \end{bmatrix}, \tag{3.3}$$

where the nonsingular matrix Q may be set up by following different strategies (see section 3.1), and its choice leads the accuracy of the whole scheme. By setting

$$S := -BHB^T, \tag{3.4}$$

matrix $\hat{\mathcal{A}}$ reads

$$\hat{\mathcal{A}} = \begin{bmatrix} C & B^T \\ B & SQ - \Sigma \end{bmatrix}. \tag{3.5}$$

Matrix S is an approximation of the pressure Schur complement Σ of $\mathcal{A}$ and $-S > 0$.

For any $n = n_0, \ldots, N_T - 1$, system (3.1) reads also

$$\text{(3.6)} \qquad \begin{cases} C\widehat{\mathbf{U}}^{n+1} + B^T\widehat{\mathbf{P}}^{n+1} = \mathbf{F}_1^{n+1} + \dfrac{1}{\Delta t}\displaystyle\sum_{j=0}^{q-1}\beta_j M\widehat{\mathbf{U}}^{n-j}, \\ B\widehat{\mathbf{U}}^{n+1} - (\Sigma - SQ)\widehat{\mathbf{P}}^{n+1} = \mathbf{F}_2^{n+1} \end{cases}$$

(this is the Yosida counterpart of (2.7)) or again

$$\text{(3.7)} \qquad \begin{array}{ll} \hat{L}\text{-step}: \text{ find } \widehat{\mathbf{U}}^{n+1/2},\ \widehat{\mathbf{P}}^{n+1/2}: & \begin{cases} C\widehat{\mathbf{U}}^{n+1/2} = \widehat{\mathbf{G}}_1^{n+1}, \\ S\widehat{\mathbf{P}}^{n+1/2} = \widehat{\mathbf{G}}_2^{n+1} - B\widehat{\mathbf{U}}^{n+1/2}, \end{cases} \\ \hat{U}\text{-step}: \text{ find } \widehat{\mathbf{U}}^{n+1},\ \widehat{\mathbf{P}}^{n+1}: & \begin{cases} Q\widehat{\mathbf{P}}^{n+1} = \widehat{\mathbf{P}}^{n+1/2}, \\ C(\widehat{\mathbf{U}}^{n+1/2} - \widehat{\mathbf{U}}^{n+1}) = B^T\widehat{\mathbf{P}}^{n+1} \end{cases} \end{array}$$

(this is the Yosida counterpart of (2.10)).

For any t^n, system (3.1) is well posed iff $\hat{\mathcal{A}}$ is invertible, i.e., iff Q is invertible. As a matter of fact, recalling the definition of $\hat{\mathcal{A}}$ and thanks to Schur's formula, we can write $\det(\hat{\mathcal{A}}) = \det(C)\det(SQ)$.

In the following section we will present some possible choices for nonsingular matrices Q and provide algebraic fractional step schemes supplying different accuracy properties.

3.1. How to set up matrix Q. System (2.7) is equivalent to (3.6) if $\hat{\mathcal{A}} = \mathcal{A}$, i.e., if either $Q = S^{-1}\Sigma$ or $Q^{-1} = \Sigma^{-1}S$, but the exact computation of Q is infeasible, since it would imply computing Σ directly, which is what we want to avoid. Our aim is to set suitable sequences of matrices $\{\tilde{Q}_p\}_{p\geq 0}$ approximating Q such that

$$\text{(3.8)} \qquad \exists C > 0 \qquad |||\Sigma - S\tilde{Q}_p||| \leq C\Delta t^{p+2} \qquad \text{as} \qquad \Delta t \to 0,$$

where $|||\cdot|||$ denotes the 2-norm for matrices. The quantity $|||\Sigma - S\tilde{Q}_p|||$ is named the *consistency error* and is due to the approximation of $\mathcal{A}$ by $\hat{\mathcal{A}}$.

The direct definition of $Q = S^{-1}\Sigma$ and its definition by the inverse $Q^{-1} = \Sigma^{-1}S$ suggest two different strategies for deriving matrices $\tilde{Q}_p$ invoked in (3.8). The first strategy exploits the expansion in series of Σ (through the expansion of C^{-1}) and the matrices so derived will be denoted by $\{Q_p\}_{p\geq 0}$. The second strategy takes advantage of the expansion of Σ^{-1} as well and the resulting matrices will be denoted by $\{\hat{Q}_p\}_{p\geq 0}$.

Let us begin to set up matrices Q_p. If $\rho(\nu HK) < 1$, we can expand Σ as follows:

$$\text{(3.9)} \qquad \Sigma = -BC^{-1}B^T = -B(I + \nu HK)^{-1}HB^T = -\sum_{k\geq 0} B(-\nu HK)^k HB^T.$$

By setting

$$\text{(3.10)} \qquad D_k := B(-\nu HK)^k HB^T = \mathcal{O}(\Delta t^{k+1}) \quad \text{for } k = 0, 1, 2, \ldots$$

we have $\Sigma = -\sum_{k\geq 0} D_k$ and $Q = S^{-1}\Sigma = -S^{-1}\sum_{k\geq 0} D_k$. Therefore, we define

$$\text{(3.11)} \qquad Q_p := -S^{-1}\sum_{k=0}^{p} D_k \quad \text{for } \ p = 0,\ 1,\ 2, \ldots .$$

The explicit forms of the first three items of $\{Q_p\}$ are

$$Q_0 = I, \qquad Q_1 = I - S^{-1}D_1, \qquad Q_2 = I - S^{-1}D_1 - S^{-1}D_2. \tag{3.12}$$

Remark 2. In the case of SEM-NI discretization, the assumption $\rho(\nu HK) < 1$ turns into

$$\nu\frac{\Delta t}{\beta_{-1}}\rho(M^{-1}K) = c\nu\frac{\Delta t}{\beta_{-1}}N^4h^{-2} < 1,$$

where c denotes a generic positive constant independent of Δt, N, and h [3, 4]. It follows that the convergence of the series $\sum_{k\geq 0}(-\nu HK)^k$ to $(I+\nu HK)^{-1}$ is guaranteed if

$$\Delta t < c\frac{\beta_{-1}h^2}{\nu\ N^4}. \tag{3.13}$$

It is worth remarking that condition (3.13) implies the nonsingularity of matrices Q_p as well.

To derive the second sequence of matrices $\{\hat{Q}_p\}$ we note that $S = -D_0$ and we write

$$Q^{-1} = \Sigma^{-1}S = \left(-\sum_{k\geq 0} D_k\right)^{-1} \cdot S = \left(S - \sum_{k\geq 1} D_k\right)^{-1} \cdot S = \left(I - S^{-1}\sum_{k\geq 1} D_k\right)^{-1}.$$

By putting $R := S^{-1}(\sum_{k\geq 1} D_k)$ it holds that $Q^{-1} = (I-R)^{-1}$ and, by assuming that $\rho(R) < 1$, it holds that $Q^{-1} = \sum_{k\geq 0} R^k$ as well. Then we rewrite Q^{-1} as a sum of matrices $\tilde{R}_k$ s.t. $\tilde{R}_k = \mathcal{O}(\Delta t^k)$ in order to highlight the dependence on powers of Δt. We have $Q^{-1} = \sum_{k\geq 0}\tilde{R}_k$ and then we set

$$\hat{Q}_p := \left(\sum_{k=0}^{p}\tilde{R}_k\right)^{-1} \qquad \text{for } p = 0,\ 1,\ 2\ldots. \tag{3.14}$$

To give the explicit form of $\hat{Q}_p$ (for $p = 0, 1, 2$) we compute the powers of R up to degree 3:

$$\begin{aligned} R &= S^{-1}D_1 + S^{-1}D_2 + S^{-1}D_3 + \mathcal{O}(\Delta t^4),\\ R^2 &= (S^{-1}D_1)^2 + S^{-1}D_1S^{-1}D_2 + S^{-1}D_2S^{-1}D_1 + \mathcal{O}(\Delta t^4),\\ R^3 &= (S^{-1}D_1)^3 + \mathcal{O}(\Delta t^4), \end{aligned}$$

and recalling that $D_k = \mathcal{O}(\Delta t^{k+1})$ we have

$$\begin{aligned} &\tilde{R}_0 = I,\\ &\tilde{R}_1 = S^{-1}D_1,\\ &\tilde{R}_2 = S^{-1}D_2 + (S^{-1}D_1)^2,\\ &\tilde{R}_3 = (S^{-1}D_1)^3 + S^{-1}D_1S^{-1}D_2 + S^{-1}D_2S^{-1}D_1 + S^{-1}D_3. \end{aligned} \tag{3.15}$$

From definition (3.14) it follows that

$$\begin{aligned} &\hat{Q}_0 = I,\\ &\hat{Q}_1 = (I + S^{-1}D_1)^{-1} = -D^{-1}S,\\ &\hat{Q}_2 = (I + S^{-1}D_1 + S^{-1}D_2 + (S^{-1}D_1)^2)^{-1}\\ &\qquad = (-D + DS^{-1}D + B(HC)^2HB^T)^{-1}S, \end{aligned} \tag{3.16}$$

where we have set $D = BHCHB^T$ and employed the fact that $C = H^{-1} + \nu K$.

Remark 3. We analyze the bound $\rho(R) < 1$ when SEM-NI are used. By construction, the leading part of R, for $\Delta t \to 0$, is $S^{-1}D_1$, and we can infer that, for sufficiently small Δt,

$$\rho(R) \simeq \rho(S^{-1}D_1) = \frac{\nu\Delta t}{\beta_{-1}}\rho((BM^{-1}B^T)^{-1}BM^{-1}KM^{-1}B^T).$$

It is possible to prove that the condition $\rho(R) < 1$ is satisfied under assumption (3.13). As a matter of fact we set $E_0 = BM^{-1}B^T$, $E_1 = BM^{-1}KM^{-1}B^T$, $E = E_0^{-1}E_1$, and we denote by $(\lambda_i, \mathbf{p}_i)$ an eigenvalue-eigenvector pair of E or, equivalently, an eigenvalue-eigenvector pair of the generalized eigenvalue problem $E_1\mathbf{p} = \lambda E_0\mathbf{p}$. Since $E_0 > 0$, the eigenvectors $\mathbf{p}_i$ can be chosen E_0-orthogonal to one other, i.e., $\mathbf{p}_i^T E_0 \mathbf{p}_j = 0$ if $i \neq j$. Then, for any eigenvector $\mathbf{p}_i$ $(i = 1, \ldots, N_p)$, we set $\mathbf{v}_i := M^{-1/2}B^T\mathbf{p}_i$ and denote by V_E the subspace of $\mathbb{R}^{2N_v}$ spanned by $\{\mathbf{v}_i\}_{i=1}^{N_p}$. It is easy to see that $M^{-1/2}KM^{-1/2}\mathbf{v}_i = \lambda_i\mathbf{v}_i$ holds for any $i = 1, \ldots, N_p$, that is, the eigenvalues λ_i of E are bounded by the eigenvalues of $M^{-1}K$.

PROPOSITION 3.1. *Matrices $\hat{Q}_p$ (for $p = 0, 1, 2$) are nonsingular.*

Proof. $\hat{Q}_0$ is the identity matrix, and $\hat{Q}_1 = -D^{-1}S$ is nonsingular since both D and S are nonsingular matrices. Proving $\hat{Q}_2$ is nonsingular is equivalent to proving $\hat{Q}_2^{-1} = \tilde{R}_0 + \tilde{R}_1 + \tilde{R}_2$ is nonsingular. By definition

$$\begin{aligned}\hat{Q}_2^{-1} &= -S^{-1}D + (S^{-1}D)^2 + S^{-1}B(HC)^2HB^T \\ &= S^{-1}BHC(-C^{-1} - HB^T(BHB^T)^{-1}BH + H)CHB^T\end{aligned}$$

and, since B is a full rank matrix, $\hat{Q}_2^{-1}$ is nonsingular iff the matrix

$$W = H - C^{-1} - HB^T(BHB^T)^{-1}BH \tag{3.17}$$

is nonsingular. By recalling that $H = \frac{\beta_{-1}}{\Delta t}M^{-1} > 0$, $C = H^{-1} + \nu K > 0$, and $K > 0$, we have $C - H^{-1} > 0$, and then $H - C^{-1} > 0$ (see P3 in section 2), which implies that $(H - C^{-1})$ is nonsingular. Since $\ker(B^T) \neq \{\mathbf{0}\}$, also BHB^T is nonsingular. Therefore, we set

$$\mathcal{M} = \begin{bmatrix} BHB^T & BH \\ HB^T & H - C^{-1} \end{bmatrix}, \tag{3.18}$$

and we note that matrix W is nothing more than the Schur complement of BHB^T in $\mathcal{M}$, i.e., $\mathcal{M}/(BHB^T)$.

By applying Schur's formula (D3, section 2), $\det(\mathcal{M}) = \det(BHB^T)\det(W)$ holds. On the other hand we can write $\det(\mathcal{M}) = \det(H - C^{-1})\det(\mathcal{M}/(H - C^{-1}))$ as well. Therefore

$$\det(W) = \det(\mathcal{M}/(BHB^T)) = \frac{\det(H - C^{-1})}{\det(BHB^T)}\det(\mathcal{M}/(H - C^{-1}))$$

and $\det(W) \neq 0$ iff $\det(\mathcal{M}/(H - C^{-1})) \neq 0$.

Recalling (2.12) we have $(H - C^{-1})^{-1} = H^{-1}[H - (-C + H^{-1})^{-1}]H^{-1}$, and since $-C + H^{-1} = -\nu K$, it holds that

$$\begin{aligned}\mathcal{M}/(H - C^{-1}) &= BHB^T - BH(H - C^{-1})^{-1}HB^T \\ &= BHB^T - BHH^{-1}[H + (\nu K)^{-1}]H^{-1}HB^T = -B(\nu K)^{-1}B^T.\end{aligned}$$

The thesis follows since K is nonsingular. ∎

As stated by the following lemma, both sequences of matrices Q_p and $\hat{Q}_p$ yield a consistency error satisfying estimate (3.8). This lemma generalizes the results given in [15, 8].

LEMMA 3.2. *There exist* $\overline{\Delta t} = \overline{\Delta t}(\nu, N, h) > 0$ *and positive constants* $c_p = c_p(\nu, N, h)$ *and* $\hat{c}_p = \hat{c}_p(\nu, N, h)$ *such that for any* $\Delta t \in (0, \overline{\Delta t})$ *and* $p = 0, 1, 2$ *the following estimates hold:*

$$|||\Sigma - SQ_p||| = c_p \Delta t^{p+2} + o(\Delta t^{p+2}), \tag{3.19}$$

$$|||\Sigma - S\hat{Q}_p||| = \hat{c}_p \Delta t^{p+2} + o(\Delta t^{p+2}). \tag{3.20}$$

Proof. By definition (3.12) we have $\Sigma - SQ_p = -\sum_{k \geq p+1} D_k$.

Estimate (3.19), with $c_p = \frac{\nu^{p+1}}{(\beta_{-1})^{p+2}} \rho(B(M^{-1}K)^{p+1} M^{-1} B^T)$, immediately follows by recalling that $D_k = \mathcal{O}(\Delta t^{k+1})$ (for any integer $k \geq 0$).

In order to prove (3.20) we begin to consider the matrix $\Sigma^{-1} - \hat{Q}_p^{-1} S^{-1}$. We set $Z = \sum_{k \geq p+1} \tilde{R}_k$, so that

$$\Sigma^{-1} - \hat{Q}_p^{-1} S^{-1} = \left(\sum_{k \geq 0} \tilde{R}_k - \sum_{k=0}^{p} \tilde{R}_k \right) S^{-1} = Z S^{-1}.$$

By using (2.13) we have

$$\Sigma - S\hat{Q}_p = -\Sigma(\Sigma^{-1} - \hat{Q}_p^{-1} S^{-1}) S\hat{Q}_p = -\Sigma Z \hat{Q}_p. \tag{3.21}$$

Estimate (3.20) proceeds by recalling that $\Sigma = -\sum_{k \geq 0} D_k = \mathcal{O}(\Delta t)$, $Z = \sum_{k \geq p+1} \tilde{R}_k = \mathcal{O}(\Delta t^{p+1})$ and by noting that $\hat{Q}_p = \left(\sum_{k=0}^{p} \tilde{R}_k \right)^{-1} = \mathcal{O}(1)$. ∎

Even though both matrices Q_p and $\hat{Q}_p$ produce consistency errors of the same order with respect to Δt, we are inclined to choose matrices $\hat{Q}_p$ instead of Q_p to set up Yosida schemes, when $p > 0$ (recall that $Q_0 = \hat{Q}_0 = I$). As a matter of fact we can easily prove that

$$\hat{c}_1 < c_1, \tag{3.22}$$

which implies $|||\Sigma - S\hat{Q}_1||| < |||\Sigma - SQ_1|||$ for sufficiently small Δt.

By exploiting the definition of $\hat{Q}_1$ and by using (2.14), with $\mathcal{B} = S = -D_0$ and $\mathcal{C} = D_1$, we write the difference $\Sigma - S\hat{Q}_1$ as follows:

$$\begin{aligned} \Sigma - S\hat{Q}_1 &= -\sum_{k \geq 0} D_k - S(S + D_1)^{-1} S = -\sum_{k \geq 0} D_k + D_0 + D_1 - D_1(D_1 - D_0)^{-1} D_1 \\ &= -D_2 - D_1(D_1 - D_0)^{-1} D_1 + o(\Delta t^3). \end{aligned}$$

On the other hand, $\Sigma - SQ_1 = -D_2 + o(\Delta t^3)$. It is immediate to verify that both D_2 and $D_1(D_1 - D_0)^{-1} D_1$ are symmetric matrices, and then the 2-norm $|||\cdot|||$ coincides with the spectral radius $\rho(\cdot)$; moreover it holds that $D_2 > 0$ and $D_1(D_1 - D_0)^{-1} D_1 < 0$. If we apply P2 of section 2, we can conclude that

$$\lambda_{max}(D_2 + D_1(D_1 - D_0)^{-1} D_1) \leq \lambda_{max}(D_2) + \lambda_{max}(D_1(D_1 - D_0)^{-1} D_1) < \lambda_{max}(D_2),$$

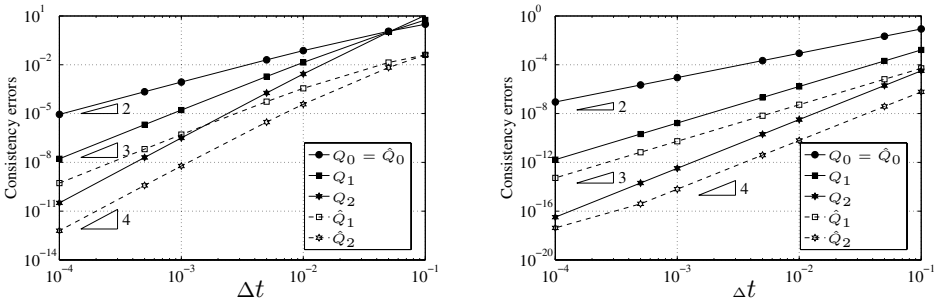


FIG. 3.1. *The consistency errors* $|||\Sigma - SQ_p|||$ *and* $|||\Sigma - S\hat{Q}_p|||$ *for* $p = 0, 1, 2$. *The viscosity is* $\nu = 10^{-2}$ *at left and* $\nu = 10^{-4}$ *at right.* 4×4 *equal square spectral elements with polynomial degree* $N = 5$ *are used.*

that is,

$$\hat{c}_1 \Delta t^3 = \rho(-D_2 - D_1(D_1 - D_0)^{-1} D_1) < \rho(-D_2) = c_1 \Delta t^3.$$

Results shown in Figure 3.1 confirm (3.22). The proof that $\hat{c}_2 < c_2$ (which implies $|||\Sigma - S\hat{Q}_2||| \leq |||\Sigma - SQ_2|||$ for small Δt) is quite long and technical, so we only report numerical results (again in Figure 3.1) confirming such assertion.

In order to compare the two approaches from a computational point of view, we note that both matrices Q_p and $\hat{Q}_p$ are involved in the solution of a linear system like $Q\hat{\mathbf{P}}^{n+1} = \hat{\mathbf{P}}^{n+1/2}$ at each time-step. When $p = 0$ it holds that $Q_0 = \hat{Q}_0 = I$. When $p \geq 1$, solving systems $Q_p\hat{\mathbf{P}}^{n+1} = \hat{\mathbf{P}}^{n+1/2}$ means performing 2 $(N_p \times 2N_v)$-matrix-vector products and solving one linear system whose matrix is $\sum_{k=0}^{p} D_k$. If we choose a direct algebraic solver, the matrix $\sum_{k=0}^{p} D_k$ may be assembled and factorized once, and the computational cost related to the solution of all systems of type $Q\hat{\mathbf{P}}^{n+1} = \hat{\mathbf{P}}^{n+1/2}$ (for any $n = n_0, \ldots, N_T - 1$) is proportional to $N_p^3 + N_p \cdot N_v \cdot N_T$. Otherwise, solving systems $\hat{Q}_p\hat{\mathbf{P}}^{n+1} = \hat{\mathbf{P}}^{n+1/2}$ means performing $2 \cdot p$ $(N_p \times 2N_v)$-matrix-vector products, $2 \cdot p$ $(N_v \times N_v)$-matrix-vector products, and solving p linear systems whose matrix is S. By recalling that matrix S is already assembled and factorized inside the algorithm, since it is used in the $\hat{L}$-substep of (3.7), the computational cost related to the solution of all $\hat{Q}\hat{\mathbf{P}}^{n+1} = \hat{\mathbf{P}}^{n+1/2}$ (for any $n = n_0, \ldots, N_T - 1$) is proportional to $(N_v + N_p) \cdot N_v \cdot N_T$. In conclusion, when we choose a direct algebraic solver to solve $Q\hat{\mathbf{P}}^{n+1} = \hat{\mathbf{P}}^{n+1/2}$, if N_T is large (i.e., $N_T \gtrsim N_v$), the choice $Q = Q_p$ will be more convenient than $Q = \hat{Q}_p$. Nevertheless, matrices $\hat{Q}_p$ (3.14) are preferable to matrices Q_p (3.11) for what concerns accuracy with respect to Δt. In this paper we have preferred accuracy properties and have chosen to approximate Q by $\hat{Q}_p$ inside the definition of $\hat{\mathcal{A}}$ (or equivalently in either (3.6) or (3.7)).

The choice $\hat{Q}_0$ yields the *Yosida* (or *Yosida*-2) method [13, 12], while the choice $\hat{Q}_1$ (resp., $\hat{Q}_2$) provides the *Yosida*-3 (resp., *Yosida*-4) scheme introduced in [15] (resp., [8]). From now on the term *Yosida*-$(p + 2)$ (for $p = 0, 1, 2$) will denote the Yosida method obtained by replacing Q with $\hat{Q}_p$.

Finally, for any $n \geq n_0$ we will denote by $(\hat{\mathbf{u}}_{\mathcal{H},p}^n, \hat{p}_{\mathcal{H},p}^n) \in \mathbf{V}_{\mathcal{H}} \times Q_{\mathcal{H}}$ the numerical solution of (3.6) obtained by replacing Q with $\hat{Q}_p$, while we will denote by $\hat{\mathbf{U}}_p^n$ (resp., $\hat{\mathbf{P}}_p^n$) the vector of the expansion coefficients of $\hat{\mathbf{u}}_{\mathcal{H},p}$ (resp., $\hat{p}_{\mathcal{H},p}$) with respect to the Lagrange basis in $V_{\mathcal{H}}$ (resp., $Q_{\mathcal{H}}$).

3.2. Convergence analysis. We define the *global errors* as the errors between the exact solution of problem (2.1) and the numerical solution $(\hat{\mathbf{u}}_{\mathcal{H},p}, \hat{p}_{\mathcal{H},p})$ obtained with any *Yosida*-$(p+2)$ method, i.e.,

$$\begin{aligned} \|\mathbf{u} - \hat{\mathbf{u}}_{\mathcal{H},p}\|_{\ell^2(H^1)} &:= \left(\Delta t \sum_{n=n_0}^{N_T-1} \|\mathbf{u}(t_n) - \hat{\mathbf{u}}^n_{\mathcal{H},p}\|^2_{H^1(\Omega)} \right)^{1/2}, \\ \|p - \hat{p}_{\mathcal{H},p}\|_{\ell^2(L^2)} &:= \left(\Delta t \sum_{n=n_0}^{N_T-1} \|p(t_n) - \hat{p}^n_{\mathcal{H},p}\|^2_{L^2(\Omega)} \right)^{1/2}. \end{aligned} \tag{3.23}$$

We can upper-bound global errors (3.23) as follows:

$$\begin{aligned} \|\mathbf{u} - \hat{\mathbf{u}}_{\mathcal{H},p}\|_{\ell^2(H^1)} &\le \|\mathbf{u} - \mathbf{u}_{\mathcal{H}}\|_{\ell^2(H^1)} + \|\mathbf{u}_{\mathcal{H}} - \hat{\mathbf{u}}_{\mathcal{H},p}\|_{\ell^2(H^1)}, \\ \|p - \hat{p}_{\mathcal{H},p}\|_{\ell^2(L^2)} &\le \|p - p_{\mathcal{H}}\|_{\ell^2(L^2)} + \|p_{\mathcal{H}} - \hat{p}_{\mathcal{H},p}\|_{\ell^2(L^2)}, \end{aligned}$$

where the first terms on the right-hand sides are the errors due to BDF, while the second ones, induced by the *Yosida*-$(p+2)$ scheme, are named *splitting errors* and are the errors between the solution $(\mathbf{u}_{\mathcal{H}}, p_{\mathcal{H}})$ of (2.10) and the solution $(\hat{\mathbf{u}}_{\mathcal{H},p}, \hat{p}_{\mathcal{H},p})$ obtained by solving (3.6) with $Q \simeq \hat{Q}_p$. Therefore, for $p = 0$, 1, 2, we are interested in analyzing the *splitting errors*:

$$\begin{aligned} \|\mathbf{u}_{\mathcal{H}} - \hat{\mathbf{u}}_{\mathcal{H},p}\|_{\ell^2(H^1)} &:= \left(\Delta t \sum_{n=n_0}^{N_T-1} \|\mathbf{u}^n_{\mathcal{H}} - \hat{\mathbf{u}}^n_{\mathcal{H},p}\|^2_{H^1(\Omega)} \right)^{1/2}, \\ \|p_{\mathcal{H}} - \hat{p}_{\mathcal{H},p}\|_{\ell^2(L^2)} &:= \left(\Delta t \sum_{n=n_0}^{N_T-1} \|p^n_{\mathcal{H}} - \hat{p}^n_{\mathcal{H},p}\|^2_{L^2(\Omega)} \right)^{1/2}. \end{aligned} \tag{3.24}$$

Remark 4. From the analysis carried out in [10] it is known that the solution of the Navier–Stokes problem can be regular for $t \to 0$ under unrealistic nonlocal compatibility conditions. On the other hand, accuracy in the usual norms $\ell^2(H^1)$ and $\ell^2(L^2)$ can be attained only if the solution is regular enough, even when using high order discretization schemes. Otherwise, high order in time can be recovered only in suitable weighted norms. In what follows, we will assume that high regularity hypotheses hold. This ensures that when using BDFq time discretization of order $q(>1)$, without the fractional step approach, optimal accuracy is q for both velocity and pressure. Without such high regularity assumptions, optimal convergence estimates given in the present paper will hold only far from $t = 0$. We denote by $||\cdot||_{\ell^2_w(V)}$ a suitable time-weighted norm; the scheme shown in Table 3.1 summarizes results known in literature.

From now on, by $(\cdot,\cdot)$ and $\|\cdot\| = (\cdot,\cdot)^{1/2}$ we will denote the classic Euclidean inner product and the Euclidean norm, respectively, in $\mathbb{R}^n$.

Let $\mathbf{U},\ \mathbf{V} \in \mathsf{V}_G$ be the vectors associated to $\mathbf{u}_{\mathcal{H}},\ \mathbf{v}_{\mathcal{H}} \in \mathbf{V}_{\mathcal{H}}$. By recalling that M and K are the mass and stiffness matrices, respectively, it holds that $(M\mathbf{U}, \mathbf{V}) = \mathbf{V}^T M\mathbf{U} = (\mathbf{u}_{\mathcal{H}}, \mathbf{v}_{\mathcal{H}})_{\mathcal{H},\Omega}$ and $(K\mathbf{U}, \mathbf{U}) = (\nabla\mathbf{u}_{\mathcal{H}}, \nabla\mathbf{v}_{\mathcal{H}})_{\mathcal{H},\Omega}$. Then we set

$$\|\mathbf{U}\|_0 := \sqrt{(M\mathbf{U}, \mathbf{U})}, \qquad \|\mathbf{U}\|_1 := \sqrt{(K\mathbf{U}, \mathbf{U})}.$$

It is worth pointing out that the norm $||\cdot||_1$ is the discrete counterpart to the L^2-norm of $\nabla\mathbf{u}$. Thanks to the Poincaré inequality, which states that, for any function

TABLE 3.1

	Low regularity assumptions on the exact solution (no compatibility conditions imposed)	High regularity assumptions on the exact solution
BDF1 in time	$\|\mathbf{u}-\mathbf{u}_{\mathcal{H}}\|_{\ell^2(H^1)}=\mathcal{O}(\Delta t)$ $\|p-p_{\mathcal{H}}\|_{\ell^2(L^2)}=\mathcal{O}(\Delta t^{1/2})$	$\|\mathbf{u}-\mathbf{u}_{\mathcal{H}}\|_{\ell^2(H^1)}=\mathcal{O}(\Delta t)$ $\|p-p_{\mathcal{H}}\|_{\ell^2(L^2)}=\mathcal{O}(\Delta t)$
Proof:	see [2]	guidelines of [2]
BDFq in time	$\|\mathbf{u}-\mathbf{u}_{\mathcal{H}}\|_{\ell^2(H^1)}=\mathcal{O}(\Delta t)$ $\|p-p_{\mathcal{H}}\|_{\ell^2(L^2)}=\mathcal{O}(\Delta t^{1/2})$ $\|\mathbf{u}-\mathbf{u}_{\mathcal{H}}\|_{\ell^2_w(H^1)}=\mathcal{O}(\Delta t^q)$ $\|p-p_{\mathcal{H}}\|_{\ell^2_w(L^2)}=\mathcal{O}(\Delta t^{q-1})$	$\|\mathbf{u}-\mathbf{u}_{\mathcal{H}}\|_{\ell^2(H^1)}=\mathcal{O}(\Delta t^q)$ $\|p-p_{\mathcal{H}}\|_{\ell^2(L^2)}=\mathcal{O}(\Delta t^q)$
Proof:	see [6] for $q=2$	guidelines of [6, 2]

$u\in H^1_0(\Omega)$, there exists a positive constant C_Ω such that $\|u\|_{L^2(\Omega)}\le C_\Omega\|\nabla u\|_{L^2(\Omega)}$, and thanks to the equivalence between the continuous L^2-norm and its discrete counterpart induced by the discrete inner product (2.4) [14, 7], $\|u_{\mathcal{H}}\|_{\mathcal{H},\Omega}\le\|u_{\mathcal{H}}\|_{L^2(\Omega)}\le 3\|u_{\mathcal{H}}\|_{\mathcal{H},\Omega}$, the following (discrete) Poincaré inequality holds:

$$\|\mathbf{V}\|_0\le 3C_\Omega\|\mathbf{V}\|_1 \qquad \text{for any } \mathbf{V}\in\mathsf{V}^0_G. \tag{3.25}$$

Moreover, since K is s.p.d., for any $\mathbf{U},\mathbf{V}\in\mathsf{V}_G$ it holds that

$$(K\mathbf{U},\mathbf{V})=(K^{1/2}\mathbf{U},K^{1/2}\mathbf{V})\le\|K^{1/2}\mathbf{U}\|\|K^{1/2}\mathbf{V}\|=\|\mathbf{U}\|_1\|\mathbf{V}\|_1. \tag{3.26}$$

Finally, from Lemma 3.2, we immediately deduce that, for $p=0,1,2$,

$$((\Sigma-S\hat{Q}_p)\mathbf{P},\mathbf{Q})\le\hat{c}_p\Delta t^{p+2}\|\mathbf{P}\|\|\mathbf{Q}\| \qquad \text{for any } \mathbf{P},\mathbf{Q}\in\mathsf{Q}_G. \tag{3.27}$$

Remark 5. In order to prove convergence estimates for the Yosida schemes, we need to prove that matrix $\Sigma-S\hat{Q}_p$ is definite. With Lemmas 3.3, 3.4, and 3.5 we prove that $\Sigma-S\hat{Q}_0>0$, $-(\Sigma-S\hat{Q}_1)\ge 0$ and, under suitable conditions, that $\Sigma-S\hat{Q}_2>0$. Lemma 3.3 has been proved in [13] by using the analogy with the Yosida regularization operator; we present here a shorter proof, based on s.p.d. matrices properties. Lemma 3.4 has been proved in [15] with different arguments. Here we follow an approach exploiting (2.15), yielding a shorter proof.

LEMMA 3.3. $\Sigma-S\hat{Q}_0>0$.

Proof. Inside the proof of Lemma 3.1 we have proved that $(H-C^{-1})>0$. Since $\ker(B^T)=\{\mathbf{0}\}$, it follows that $\Sigma-S\hat{Q}_0=\Sigma-S=B(H-C^{-1})B^T>0$. ∎

LEMMA 3.4. $-(\Sigma-S\hat{Q}_1)\ge 0$.

Proof. By definitions of Σ, S, and $\hat{Q}_1$ we have

$$-(\Sigma-S\hat{Q}_1)=B[C^{-1}-(BH)^T(BHCHB^T)^{-1}BH]B^T$$

and since $\ker(B^T)=\{\mathbf{0}\}$, $-(\Sigma-S\hat{Q}_1)\ge 0$ iff $C^{-1}-(BH)^T(BHCHB^T)^{-1}BH\ge 0$. The thesis follows by putting $\mathcal{X}=BH$ and $\mathcal{C}=C$ in (2.15). ∎

Inside the proof of the following lemma, we will use the matrix

$$\tilde{\tilde{B}}:=-D_3-D_1S^{-1}D_2-D_2S^{-1}D_1-D_1(S^{-1}D_1)^2. \tag{3.28}$$

Matrix $\tilde{\tilde{B}}$ is symmetric, and numerical computations have shown that $\tilde{\tilde{B}}$ has real nonnegative eigenvalues for any space discretization we have considered. Nevertheless

we cannot prove that it is positive definite, even if it is worth noting that nonnegativeness of its eigenvalues does not depend on the time-step (recalling definitions of matrices D_k, all the addend in (3.28) are $\mathcal{O}(\Delta t^3)$).

LEMMA 3.5. *If* $\tilde{\tilde{B}} > 0$ *and if* Δt *is sufficiently small, then* $\Sigma - S\hat{Q}_2 > 0$.

Proof. As a particular case of (3.21) we have

$$\Sigma - S\hat{Q}_2 = -\Sigma(\Sigma^{-1} - \hat{Q}_2^{-1}S^{-1})S\hat{Q}_2, \tag{3.29}$$

and we analyze the sign of the three factors of the right-hand side of (3.29). Let us start by proving that if Δt is small enough, then $-S\hat{Q}_2 > 0$. We set

$$\tilde{B} := D - DS^{-1}D - B(HC)^2HB^T, \tag{3.30}$$

so that $-S\hat{Q}_2 = S\tilde{B}^{-1}S$. Since $\tilde{B}$ is symmetric, then $-S\hat{Q}_2$ is symmetric as well (and $\Sigma - S\hat{Q}_2$ is also symmetric); moreover we can apply Sylvester's law of inertia [9] and conclude that $-S\hat{Q}_2 > 0$ iff $\tilde{B} > 0$. In order to prove that $\tilde{B} > 0$ we apply P2 of section 2. By construction, both D and $-DS^{-1}D - B(HC)^2HB^T$ are symmetric matrices, and we can write that $\lambda_{min}(\tilde{B}) \geq \lambda_{min}(D) + \lambda_{min}(-DS^{-1}D - B(HC)^2HB^T)$. Since $D = -S - D_1 = \mathcal{O}(\Delta t)$ and $-DS^{-1}D - B(HC)^2HB^T = -D_2 - D_1S^{-1}D_1 = \mathcal{O}(\Delta t^3)$, there exist two real constants c_1 and c_2 such that $\lambda_{min}(D) = c_1\Delta t$ and $\lambda_{min}(-DS^{-1}D - B(HC)^2HB^T) = c_2\Delta t^3$; therefore $c_1 + c_2\Delta t^2 > 0$ is a sufficient condition to have $\lambda_{min}(\tilde{B}) > 0$. Since $D = BHCHB^T > 0$, then $c_1 > 0$, while we cannot give any estimate on c_2. If $c_2 > 0$, no restrictions on Δt are required; otherwise if $c_2 < 0$, we have to assume that $\Delta t < \overline{\Delta t} := (-c1/c2)^{1/2}$.

Let us consider now the matrix $-(\Sigma^{-1} - \hat{Q}_2^{-1}S^{-1}) = -\tilde{R}_3S^{-1} - \sum_{k\geq 4}\tilde{R}_kS^{-1}$, where $\tilde{R}_3S^{-1} = \mathcal{O}(\Delta t^2)$ and $\sum_{k\geq 4}\tilde{R}_kS^{-1} = \mathcal{O}(\Delta t^3)$, and we analyze only the sign of $-\tilde{R}_3S^{-1}$. By (3.15) we have $-\tilde{R}_3S^{-1} = S^{-1}\tilde{\tilde{B}}S^{-1}$, so that $-\tilde{R}_3S^{-1} > 0$ since $\tilde{\tilde{B}} > 0$. Whatever the sign of $-\sum_{k\geq 4}\tilde{R}_kS^{-1}$ may be, for sufficiently small Δt, the matrix $-(\Sigma^{-1} - \hat{Q}_2^{-1}S^{-1})$ will be s.p.d.

Finally, starting from (3.29) we can write $(\Sigma - S\hat{Q}_2)(-S\hat{Q}_2)^{-1} = \Sigma(\Sigma^{-1} - \hat{Q}_2^{-1}S^{-1})$ and the thesis follows by noting that the matrix on the right is similar to a s.p.d. matrix (P5 of section 2) and by applying P4 of section 2, with $\mathcal{B} = \Sigma - S\hat{Q}_2$ and $\mathcal{C} = (-S\hat{Q}_2)^{-1}$. ∎

It is worth noting that the stability of BDFq+*Yosida*-$(p+2)$ depends on stability of both BDFq scheme and Yosida splitting. The following lemma extends stability results proved in [12, 15] and highlights the stability of Yosida splitting when the associated BDFq method is absolutely stable, i.e., $q = 1, 2$. On the contrary, when either BDF3 or BDF4 is considered, it is natural to expect that a stability condition due to BDF has to be ensured as well.

LEMMA 3.6 (stability). *BDF*q+*Yosida*-$(p+2)$ *scheme* (3.7) *is unconditionally stable if* $q = 1, 2$ *and* $p = 0$, *while it is conditionally stable if* $q = 1, 2$ *and* $p > 0$.

Proof. We consider system (3.6) with $Q = \hat{Q}_p$, $\mathbf{F}_1 = \mathbf{0}$, $\mathbf{F}_2 = \mathbf{0}$. By using either the identity $2(a-b,a) = a^2 - b^2 + (a-b)^2$ (when $q = 1$) or $2(3a - 4b + c, a) = a^2 - b^2 + (2a-b)^2 - (2b-c)^2 + (a - 2b + c)^2$ (when $q = 2$), by summing from $n = n_0$ up to $N_T - 1$, by Lemma 3.3, and by neglecting some positive terms, it holds that $\|\hat{\mathbf{U}}_p^{N_T}\|_0^2 \leq C$, where C is a positive constant depending only on initial data.

When $q = 1, 2$ and $p = 1$, we proceed as before. In particular, when $p = 1$, in view of Lemma 3.4, a sufficient condition to ensure stability is furnished by $\nu\|\hat{\mathbf{U}}^{n+1}\|_1^2 + ((\Sigma - S\hat{Q}_p)\hat{\mathbf{P}}^{n+1}, \hat{\mathbf{P}}^{n+1}) \geq 0$ for any $n \geq n_0$. By recalling (3.20), the previous estimate

infers a bound on Δt. Finally, when $p = 2$, Lemma 3.5 ensures positiveness of $\Sigma - S\hat{Q}_p$ under suitable restrictions on Δt. ∎

We set $\mathbf{E}^n_{U,p} = \mathbf{U}^n - \hat{\mathbf{U}}^n_p$, $\mathbf{E}^n_{P,p} = \mathbf{P}^n - \hat{\mathbf{P}}^n_p$, and $\mathbf{E}^n_p = [\mathbf{E}^n_{U,p}, \mathbf{E}^n_{P,p}]^T$, recalling that p is the subindex of $\hat{Q}_p$.

THEOREM 3.7 (convergence). *Let us consider systems* (2.7) *and* (3.6) *with either* $q = 1$ *or* $q = 2$, *and* (3.6) *with* $Q = \hat{Q}_p$ *for* $p = 0, 1, 2$. *If* Δt *is sufficiently small, then there exist two positive constants* C_p *and* $\tilde{C}_p$, *depending on space discretization, but independent of* Δt, *such that the following convergence estimates hold:*

$$\|\mathbf{E}^{N_T}_{U,p}\|^2_0 + \nu\Delta t \sum_{n=n_0}^{N_T-1} \|\mathbf{E}^{n+1}_{U,p}\|^2_1 \le C_p \Delta t^{2p+3}, \tag{3.31}$$

$$\Delta t \sum_{n=n_0}^{N_T-1} \|\mathbf{E}^{n+1}_{P,p}\|^2 \le \tilde{C}_p \Delta t^{2p+2}, \tag{3.32}$$

i.e., $\|\mathbf{u}_{\mathcal{H}} - \hat{\mathbf{u}}_{\mathcal{H},p}\|_{l^\infty(L^2)} \le C_p \Delta t^{p+3/2}$, $\|\mathbf{u}_{\mathcal{H}} - \hat{\mathbf{u}}_{\mathcal{H},p}\|_{l^2(H^1)} \le C_p \Delta t^{p+3/2}$, *and* $\|p_{\mathcal{H}} - \hat{p}_{\mathcal{H},p}\|_{l^2(L^2)} \le \tilde{C}_p \Delta t^{p+1}$.

*Proof. BDF*1. We start by considering BDF1 for time approximation of both systems (2.7) and (3.6). They read

$$\begin{cases} \dfrac{M}{\Delta t}(\mathbf{U}^{n+1} - \mathbf{U}^n) + \nu K \mathbf{U}^{n+1} + B^T \mathbf{P}^{n+1} = \mathbf{F}^{n+1}_1, \\ B\mathbf{U}^{n+1} = \mathbf{F}^{n+1}_2 \end{cases} \tag{3.33}$$

and

$$\begin{cases} \dfrac{M}{\Delta t}(\hat{\mathbf{U}}^{n+1}_p - \hat{\mathbf{U}}^n_p) + \nu K \hat{\mathbf{U}}^{n+1}_p + B^T \hat{\mathbf{P}}^{n+1}_p = \mathbf{F}^{n+1}_1, \\ B\hat{\mathbf{U}}^{n+1}_p - (\Sigma - S\hat{Q}_p)\hat{\mathbf{P}}^{n+1}_p = \mathbf{F}^{n+1}_2. \end{cases} \tag{3.34}$$

By subtracting (3.34) from (3.33), we obtain

$$\begin{cases} \dfrac{M}{\Delta t}(\mathbf{E}^{n+1}_{U,p} - \mathbf{E}^n_{U,p}) + \nu K \mathbf{E}^{n+1}_{U,p} + B^T \mathbf{E}^{n+1}_{P,p} = \mathbf{0}, \\ B\mathbf{E}^{n+1}_{U,p} - (\Sigma - S\hat{Q}_p)\mathbf{E}^{n+1}_{P,p} = -(\Sigma - S\hat{Q}_p)\mathbf{P}^{n+1}, \end{cases} \tag{3.35}$$

then multiply the first and the second equations of (3.35) by $\mathbf{V}^T = 2\Delta t(\mathbf{E}^{n+1}_{U,p})^T$ and $\mathbf{Q}^T = 2\Delta t(\mathbf{E}^{n+1}_{P,p})^T$, respectively, and we subtract the second equation from the first one. We obtain

$$\begin{aligned} &(M(\mathbf{E}^{n+1}_{U,p} - \mathbf{E}^n_{U,p}), \mathbf{E}^{n+1}_{U,p}) + 2\nu\Delta t(K\mathbf{E}^{n+1}_{U,p}, \mathbf{E}^{n+1}_{U,p}) \\ &\quad + 2\Delta t((\Sigma - S\hat{Q}_p)\mathbf{E}^{n+1}_{P,p}, \mathbf{E}^{n+1}_{P,p}) = 2\Delta t((\Sigma - S\hat{Q}_p)\mathbf{P}^{n+1}, \mathbf{E}^{n+1}_{P,p}). \end{aligned} \tag{3.36}$$

Moreover, by using the identity $2(a-b,a) = |a|^2 - |b|^2 + |a-b|^2$ and the definite positivity/negativity of $(\Sigma - S\hat{Q}_p)$ (Lemmas 3.3, 3.4, and 3.5), we have

$$\begin{aligned} &\|\mathbf{E}^{n+1}_{U,p}\|^2_0 - \|\mathbf{E}^n_{U,p}\|^2_0 + \|\mathbf{E}^{n+1}_{U,p} - \mathbf{E}^n_{U,p}\|^2_0 + 2\nu\Delta t\|\mathbf{E}^{n+1}_{U,p}\|^2_1 \\ &\quad \le 2\Delta t((\Sigma - S\hat{Q}_p)\mathbf{P}^{n+1}, \mathbf{E}^{n+1}_{P,p}) - 2\gamma\Delta t((\Sigma - S\hat{Q}_p)\mathbf{E}^{n+1}_{P,p}, \mathbf{E}^{n+1}_{P,p}), \end{aligned}$$

where $\gamma = 0$ when $p = 0$ or when $p = 2$ and $\tilde{\tilde{B}} > 0$; otherwise $\gamma = 1$ when $p = 1$ or $p = 2$ and $\tilde{\tilde{B}}$ is not positive definite.

By applying the Cauchy–Schwarz inequality, estimate (3.27), and the Young inequality it follows that

$$\begin{aligned}\|\mathbf{E}_{U,p}^{n+1}\|_0^2 - \|\mathbf{E}_{U,p}^{n}\|_0^2 + \|\mathbf{E}_{U,p}^{n+1} - \mathbf{E}_{U,p}^{n}\|_0^2 + 2\nu\Delta t\|\mathbf{E}_{U,p}^{n+1}\|_1^2 \\ \leq \frac{1}{2\varepsilon}\|(\Sigma - S\hat{Q}_p)\mathbf{P}^{n+1}\|^2 + 2\varepsilon\Delta t^2\|\mathbf{E}_{P,p}^{n+1}\|^2 + 2\gamma\hat{c}_p^2\Delta t^{p+3}\|\mathbf{E}_{P,p}^{n+1}\|^2 \\ \leq \frac{\hat{c}_p^2}{2\varepsilon}\Delta t^{2p+4}\|\mathbf{P}^{n+1}\|^2 + \left(2\varepsilon\Delta t^2 + 2\gamma\hat{c}_p^2\Delta t^{p+3}\right)\|\mathbf{E}_{P,p}^{n+1}\|^2,\end{aligned}$$

and by summing on $n = n_0, \ldots, N_T - 1$ we obtain

$$\begin{aligned}\|\mathbf{E}_{U,p}^{N_T}\|_0^2 + \sum_{n=n_0}^{N_T-1}\|\mathbf{E}_{U,p}^{n+1} - \mathbf{E}_{U,p}^{n}\|_0^2 + 2\nu\Delta t\sum_{n=n_0}^{N_T-1}\|\mathbf{E}_{U,p}^{n+1}\|_1^2 \\ \leq \frac{\hat{c}_p^2}{2\varepsilon}\Delta t^{2p+3}\left(\Delta t\sum_{n=n_0}^{N_T-1}\|\mathbf{P}^{n+1}\|^2\right) \\ + \left(2\varepsilon\Delta t^2 + 2\gamma\hat{c}_p^2\Delta t^{p+3}\right)\sum_{n=n_0}^{N_T-1}\|\mathbf{E}_{P,p}^{n+1}\|^2.\end{aligned} \tag{3.37}$$

We denote by β the *inf-sup* constant which depends on the space discretization but that is independent of Δt. Thanks to the *inf-sup* condition, to the first equation in (3.35), to (3.25) and (3.26), we have

$$\begin{aligned}\|\mathbf{E}_{P,p}^{n+1}\| &\leq \frac{1}{\beta}\sup_{\substack{\mathbf{V}\in\mathbf{V}_G^0\\ \mathbf{V}\neq\mathbf{0}}}\frac{|(B^T\mathbf{E}_{P,p}^{n+1},\mathbf{V})|}{\|\mathbf{V}\|_1} = \frac{1}{\beta}\sup_{\substack{\mathbf{V}\in\mathbf{V}_G^0\\ \mathbf{V}\neq\mathbf{0}}}\frac{|(\frac{M}{\Delta t}(\mathbf{E}_{U,p}^{n+1} - \mathbf{E}_{U,p}^{n}) + \nu K\mathbf{E}_{U,p}^{n+1},\mathbf{V})|}{\|\mathbf{V}\|_1} \\ &\leq \frac{1}{\beta}\left[\frac{3C_\Omega}{\Delta t}\|\mathbf{E}_{U,p}^{n+1} - \mathbf{E}_{U,p}^{n}\|_0 + \nu\|\mathbf{E}_{U,p}^{n+1}\|_1\right]\end{aligned}$$

and then

$$\Delta t\sum_{n=n_0}^{N_T-1}\|\mathbf{E}_{P,p}^{n+1}\|^2 \leq \frac{\Delta t}{\beta^2}\left[\frac{18C_\Omega^2}{\Delta t^2}\sum_{n=n_0}^{N_T-1}\|\mathbf{E}_{U,p}^{n+1} - \mathbf{E}_{U,p}^{n}\|_0^2 + 2\nu^2\sum_{n=n_0}^{N_T-1}\|\mathbf{E}_{U,p}^{n+1}\|_1^2\right]. \tag{3.38}$$

Now we note that there exists a positive constant C_P such that $\Delta t(\sum_{n=n_0}^{N_T-1}\|\mathbf{P}^{n+1}\|^2) \leq C_P$. As a matter of fact, $p \in L^2(0,T;L_0^2(\Omega))$, while the discretization error on the pressure associated to the pure BDF1 scheme (3.33) satisfies the inequality $\Delta t\sum_{n=n_0}^{N_T-1}\|p(t_{n+1}) - p_{\mathcal{H}}(t_{n+1})\|^2 \leq c\Delta t$ [2].

Therefore, we replace (3.38) in (3.37) and obtain

$$\begin{aligned}\|\mathbf{E}_{U,p}^{N_T}\|_0^2 + \sum_{n=n_0}^{N_T-1}\|\mathbf{E}_{U,p}^{n+1} - \mathbf{E}_{U,p}^{n}\|_0^2 + 2\nu\Delta t\sum_{n=n_0}^{N_T-1}\|\mathbf{E}_{U,p}^{n+1}\|_1^2 \\ \leq \frac{\hat{c}_p^2C_P}{2\varepsilon}\Delta t^{2p+3} + \left(\frac{\varepsilon\Delta t}{2} + \gamma\hat{c}_p^2\Delta t^{p+2}\right)\frac{\Delta t}{\beta^2} \\ \times\left[\frac{18C_\Omega^2}{\Delta t^2}\sum_{n=n_0}^{N_T-1}\|\mathbf{E}_{U,p}^{n+1} - \mathbf{E}_{U,p}^{n}\|_0^2 + 2\nu^2\sum_{n=n_0}^{N_T-1}\|\mathbf{E}_{U,p}^{n+1}\|_1^2\right].\end{aligned}$$

If $\gamma = 0$, we choose $\varepsilon = \beta^2/(18C_\Omega^2)$, we move on the left the terms depending on $\mathbf{E}_U$, and, under the assumption $\Delta t \le 18C_\Omega^2/\nu$, we obtain

$$\|\mathbf{E}_{U,p}^{N_T}\|_0^2 + \frac{1}{2}\sum_{n=n_0}^{N_T-1}\|\mathbf{E}_{U,p}^{n+1} - \mathbf{E}_{U,p}^{n}\|_0^2 + \nu\Delta t\sum_{n=n_0}^{N_T-1}\|\mathbf{E}_{U,p}^{n+1}\|_1^2 \le \frac{9C_\Omega^2\hat{c}_p^2C_P}{\beta^2}\Delta t^{2p+3}. \tag{3.39}$$

Otherwise, if $\gamma = 1$, we choose $\varepsilon = \beta^2/(27C_\Omega^2)$, again we move on the left the terms depending on $\mathbf{E}_{U,p}$, and we have

$$\begin{aligned}\|\mathbf{E}_{U,p}^{N_T}\|_0^2 &+ \left(\frac{2}{3} - \Delta t^{p+1}\frac{18\hat{c}_p^2C_\Omega^2}{\beta^2}\right)\sum_{n=n_0}^{N_T-1}\|\mathbf{E}_{U,p}^{n+1} - \mathbf{E}_{U,p}^{n}\|_0^2\\ &+ \nu\Delta t\left(2 - \frac{\nu\Delta t}{27C_\Omega^2} - \frac{2\nu\hat{c}_p^2\Delta t^{2p+3}}{\beta^2}\right)\sum_{n=n_0}^{N_T-1}\|\mathbf{E}_{U,p}^{n+1}\|_1^2 \le \frac{27C_\Omega^2\hat{c}_p^2C_P}{2\beta^2}\Delta t^{2p+3}.\end{aligned}$$

If we require that $\Delta t \le \min\left\{\left(\frac{\beta}{6\sqrt{3}\hat{c}_pC_\Omega}\right)^{(p+1)/2}, \frac{27C_\Omega^2}{2\nu}, \left(\frac{\beta}{2\hat{c}_p\sqrt{\nu}}\right)^{1/(2p+3)}\right\}$, an estimate like (3.39), but with a different numeric constant, is true as well.

In conclusion, by neglecting some positive terms the estimate (3.31) holds, that is, the *Yosida*-$(p+2)$ method is $(p+3/2)$th order accurate on the velocity with respect to both $\ell^\infty(L^2)$- and $\ell^2(H^1)$-norms, with $C_p = C_p(C_\Omega, \hat{c}_p, C_P, \beta)$.

Finally, from (3.38) and (3.31), the estimate (3.32) on the pressure holds as well, that is, the *Yosida*-$(p+2)$ is $(p+1)$th order accurate on the pressure with respect to the $\ell^2(L^2)$-norm, with $\tilde{C}_p = \tilde{C}_p(C_\Omega, \hat{c}_p, C_P, \beta)$.

*BDF*2. We take into account both systems (2.7) and (3.6) with $q = 2$ and subtract the second system from the first one, obtaining

$$\begin{cases}\dfrac{M}{2\Delta t}(3\mathbf{E}_{U,p}^{n+1} - 4\mathbf{E}_{U,p}^{n} + \mathbf{E}_{U,p}^{n-1}) + \nu K\mathbf{E}_{U,p}^{n+1} + B^T\mathbf{E}_{P,p}^{n+1} = \mathbf{0},\\ B\mathbf{E}_{U,p}^{n+1} - (\Sigma - S\hat{Q}_p)\mathbf{E}_{P,p}^{n+1} = -(\Sigma - S\hat{Q}_p)\mathbf{P}^{n+1}.\end{cases} \tag{3.40}$$

Then we multiply the first and the second equations of (3.40) by $\mathbf{V}^T = 4\Delta t(\mathbf{E}_{U,p}^{n+1})^T$ and $\mathbf{Q}^T = 4\Delta t(\mathbf{E}_{P,p}^{n+1})^T$, respectively, and subtract the second equation from the first one. We obtain

$$\begin{aligned}2(M(3\mathbf{E}_{U,p}^{n+1} - 4\mathbf{E}_{U,p}^{n} + \mathbf{E}_{U,p}^{n-1}), \mathbf{E}_{U,p}^{n+1}) &+ 4\nu\Delta t(K\mathbf{E}_{U,p}^{n+1}, \mathbf{E}_{U,p}^{n+1})\\ &+ 4\Delta t((\Sigma - S\hat{Q}_p)\mathbf{E}_{P,p}^{n+1}, \mathbf{E}_{P,p}^{n+1}) = 4\Delta t((\Sigma - S\hat{Q}_p)\mathbf{P}^{n+1}, \mathbf{E}_{P,p}^{n+1}).\end{aligned} \tag{3.41}$$

Moreover, by using the identity $2(3a - 4b + c, a) = |a|^2 - |b|^2 + |2a - b|^2 - |2b - c|^2 + |a - 2b + c|^2$, working as in the proof for BDF1, and by neglecting some positive terms, we obtain

$$\begin{aligned}\|\mathbf{E}_{U,p}^{N_T}\|_0^2 &+ \sum_{n=n_0}^{N_T-1}\|\mathbf{E}_{U,p}^{n+1} - 2\mathbf{E}_{U,p}^{n} + \mathbf{E}_{U,p}^{n-1}\|_0^2 + 4\nu\Delta t\sum_{n=n_0}^{N_T-1}\|\mathbf{E}_{U,p}^{n+1}\|_1^2\\ &\le \frac{\hat{c}_p^2}{2\varepsilon}C_P\Delta t^{2p+3} + 4\left(2\varepsilon\Delta t^2 + \gamma\hat{c}_p^2\Delta t^{p+3}\right)\sum_{n=n_0}^{N_T-1}\|\mathbf{E}_{P,p}^{n+1}\|^2.\end{aligned} \tag{3.42}$$

By applying the *inf-sup* condition, we have

$$\begin{aligned}\|\mathbf{E}_{P,p}^{n+1}\| &\le \frac{1}{\beta}\sup_{\substack{\mathbf{v}\in\mathbf{v}_G^0\\ \mathbf{v}\neq\mathbf{0}}}\frac{|(B^T\mathbf{E}_{P,p}^{n+1},\mathbf{V})|}{\|\mathbf{V}\|_1}\\ &\le \frac{1}{\beta}\left[\frac{3C_\Omega}{2\Delta t}\|3\mathbf{E}_{U,p}^{n+1}-4\mathbf{E}_{U,p}^{n}+\mathbf{E}_{U,p}^{n-1}\|_0+\nu\|\mathbf{E}_{U,p}^{n+1}\|_1\right]\\ &\le \frac{1}{\beta}\left[\frac{3C_\Omega}{2\Delta t}\|\mathbf{E}_{U,p}^{n+1}-2\mathbf{E}_{U,p}^{n}+\mathbf{E}_{U,p}^{n-1}\|_0+2\|\mathbf{E}_{U,p}^{n+1}-\mathbf{E}_{U,p}^{n}\|_0+\nu\|\mathbf{E}_{U,p}^{n+1}\|_1\right]\end{aligned}$$

and then

$$\begin{aligned}\Delta t\sum_{n=n_0}^{N_T-1}\|\mathbf{E}_{P,p}^{n+1}\|^2 \le \frac{3\Delta t}{\beta^2}\Bigg[&\nu^2\sum_{n=n_0}^{N_T-1}\|\mathbf{E}_{U,p}^{n+1}\|_1^2\\ &+\frac{9C_\Omega^2}{4\Delta t^2}\sum_{n=n_0}^{N_T-1}\left(\|\mathbf{E}_{U,p}^{n+1}-2\mathbf{E}_{U,p}^{n}+\mathbf{E}_{U,p}^{n-1}\|_0^2+4\|\mathbf{E}_{U,p}^{n+1}-\mathbf{E}_{U,p}^{n}\|_0^2\right)\Bigg].\end{aligned}\tag{3.43}$$

Now we have to estimate the term $\|\mathbf{E}_{U,p}^{n+1}-\mathbf{E}_{U,p}^{n}\|_0^2$ which is not present on the left-hand side of (3.42). To this aim we multiply the first equation of (3.40) by $4\Delta t(\mathbf{E}_{U,p}^{n+1}-\mathbf{E}_{U,p}^{n})^T$, the second equation of (3.42) at both time t_{n+1} and t_n by $4\Delta t(\mathbf{E}_{p,p}^{n+1})^T$, and we linearly combine them in order to cut the term $(B\mathbf{E}_{P,p}^{n+1},\mathbf{E}_{U,p}^{n+1}-\mathbf{E}_{U,p}^{n})$. By rewriting $3\mathbf{E}_{U,p}^{n+1}-4\mathbf{E}_{U,p}^{n}+\mathbf{E}_{U,p}^{n-1}=2(\mathbf{E}_{U,p}^{n+1}-\mathbf{E}_{U,p}^{n})+(\mathbf{E}_{U,p}^{n+1}-\mathbf{E}_{U,p}^{n})-(\mathbf{E}_{U,p}^{n}-\mathbf{E}_{U,p}^{n-1})$ by applying $(2a,a-b)=|a|^2-|b|^2+|a-b|^2$, we have

$$\begin{aligned}&5\|\mathbf{E}_{U,p}^{n+1}-\mathbf{E}_{U,p}^{n}\|_0^2-\|\mathbf{E}_{U,p}^{n}-\mathbf{E}_{U,p}^{n-1}\|_0^2+\|\mathbf{E}_{U,p}^{n+1}-2\mathbf{E}_{U,p}^{n}+\mathbf{E}_{U,p}^{n-1}\|_0^2\\ &\quad+2\nu\Delta t\left(\|\mathbf{E}_{U,p}^{n+1}\|_1^2-\|\mathbf{E}_{U,p}^{n}\|_1^2+\|\mathbf{E}_{U,p}^{n+1}-\mathbf{E}_{U,p}^{n}\|_1^2\right)\\ &\quad=-4\Delta t((\Sigma-S\hat{Q}_p)(\mathbf{P}^{n+1}-\mathbf{P}^{n}),\mathbf{E}_{P,p}^{n+1})+4\Delta t((\Sigma-S\hat{Q}_p)(\mathbf{E}_{P,p}^{n+1}-\mathbf{E}_{P,p}^{n}),\mathbf{E}_{P,p}^{n+1}).\end{aligned}$$

Therefore, by applying the same techniques as before, we have

$$\begin{aligned}&4\sum_{n=n_0}^{N_T-1}\|\mathbf{E}_{U,p}^{n+1}-\mathbf{E}_{U,p}^{n}\|_0^2+\sum_{n=n_0}^{N_T-1}\|\mathbf{E}_{U,p}^{n+1}-2\mathbf{E}_{U,p}^{n}+\mathbf{E}_{U,p}^{n-1}\|_0^2\\ &\qquad+2\nu\Delta t\sum_{n=n_0}^{N_T-1}\|\mathbf{E}_{U,p}^{n+1}-\mathbf{E}_{U,p}^{n}\|_1^2\\ &\qquad\le\frac{4}{\varepsilon}\hat{c}_p^2C_P\Delta t^{2p+3}+\left(2\varepsilon\Delta t+3\hat{c}_p\Delta t^{p+2}\right)\left(\Delta t\sum_{n=n_0}^{N_T-1}\|\mathbf{E}_{P,p}^{n+1}\|^2\right),\end{aligned}\tag{3.44}$$

where we have used $\sum_n\|\mathbf{P}^{n+1}-\mathbf{P}^{n}\|^2\le 2\sum_n\|\mathbf{P}^{n+1}\|^2$ and $\sum_n\|\mathbf{E}_{P,p}^{n+1}-\mathbf{E}_{P,p}^{n}\|^2\le 2\sum_n\|\mathbf{E}_{P,p}^{n+1}\|^2$. By summing inequalities (3.42) and (3.44) and by neglecting some

positive terms, it holds that

$$\begin{aligned}
&\|\mathbf{E}_{U,p}^{N_T}\|_0^2 + 2\sum_{n=n_0}^{N_T-1}\|\mathbf{E}_{U,p}^{n+1} - 2\mathbf{E}_{U,p}^{n} + \mathbf{E}_{U,p}^{n-1}\|_0^2 + 4\sum_{n=n_0}^{N_T-1}\|\mathbf{E}_{U,p}^{n+1} - \mathbf{E}_{U,p}^{n}\|_0^2 \\
&\quad + 2\nu\Delta t\sum_{n=n_0}^{N_T-1}\|\mathbf{E}_{U,p}^{n+1}\|_1^2 \\
&\quad \le \frac{9}{2\varepsilon}\hat{c}_p^2 C_P \Delta t^{2p+3} + (10\varepsilon\Delta t + 7\hat{c}_p\Delta t^{p+2})\left(\Delta t\sum_{n=n_0}^{N_T-1}\|\mathbf{E}_{P,p}^{n+1}\|^2\right).
\end{aligned} \tag{3.45}$$

By using (3.43) in (3.45) and by moving the terms with $\mathbf{E}_{U,p}$ on the left, we have

$$\begin{aligned}
&\|\mathbf{E}_{U,p}^{N_T}\|_0^2 + \left(4 - \frac{3}{\beta_2}(10\varepsilon\Delta t + 7\hat{c}_p\Delta t^{p+2})\frac{9c_\Omega^2}{4\Delta t}\right)\sum_{n=n_0}^{N_T-1}\|\mathbf{E}_{U,p}^{n+1} - \mathbf{E}_{U,p}^{n}\|_0^2 \\
&\quad + \left(2 - \frac{3}{\beta_2}(10\varepsilon\Delta t + 7\hat{c}_p\Delta t^{p+2})\frac{9c_\Omega^2}{4\Delta t}\right)\sum_{n=n_0}^{N_T-1}\|\mathbf{E}_{U,p}^{n+1} - 2\mathbf{E}_{U,p}^{n} + \mathbf{E}_{U,p}^{n-1}\|_0^2 \\
&\quad + \left(4\nu\Delta t - \frac{3}{\beta_2}(10\varepsilon\Delta t + 7\hat{c}_p\Delta t^{p+2})\nu^2\Delta t\right)\sum_{n=n_0}^{N_T-1}\|\mathbf{E}_{U,p}^{n+1}\|_1^2 \le \frac{9}{2\varepsilon}\hat{c}_p^2 C_P\Delta t^{2p+3}.
\end{aligned}$$

We choose $\varepsilon = \frac{1}{10}(\beta^2/(27C_\Omega^2) - 7\hat{c}_p\Delta t^{p+1})$ and, under the assumption $\Delta t \le \min\{27\, C_\Omega^2/\nu, (4\beta^2/(189\hat{c}_p C_\Omega^2))^{1/(p+1)}\}$, the estimate (3.31) follows. The estimate (3.32) follows by both (3.31) and (3.43). ∎

Remark 6. It is worth noting that the assumptions required on the time-step by the previous theorem are not very restrictive; note that for $\Omega \subset \mathbb{R}^d$ it holds that $C_\Omega = \frac{2meas(\Omega)}{d}$ and $\beta = \mathcal{O}(N^{(1-d)/2})$. The most restrictive condition on the time-step is that given in (3.13) ensuring the series $-\sum_{k\ge 0} D_k$ converges to Σ.

Remark 7. When either BDF3 or BDF4 are considered, under suitable stability conditions on the time-step, we expect that convergence estimates like (3.31)–(3.32) hold for the couple BDFq-*Yosida*-$(p+2)$, as well as for $q = 3, 4$, as numerical results of the next section show.

Remark 8 (extensions to the Navier–Stokes case). We consider now the semi-discretization of unsteady Navier–Stokes equations by the following semi-implicit scheme: for any $n = n_0, \ldots, N_T - 1$, look for the solution $(\mathbf{u}^{n+1}, p^{n+1})$ of the system

$$\begin{cases}
\dfrac{\beta_{-1}}{\Delta t}\mathbf{u}^{n+1} - \nu\Delta\mathbf{u}^{n+1} + \mathcal{N}(\mathbf{u}^{n-q+1}, \ldots, \mathbf{u}^n; \mathbf{u}^{n+1}) \\
\qquad\qquad + \nabla p^{n+1} = \mathbf{f}^{n+1} + \displaystyle\sum_{j=0}^{q-1}\frac{\beta_j}{\Delta t}\mathbf{u}^{n-j} & \text{in } \Omega, \\
\nabla\cdot\mathbf{u}^{n+1} = 0 & \text{in } \Omega, \\
\mathbf{u}^{n+1} = \mathbf{g}^{n+1} & \text{on } \partial\Omega,
\end{cases} \tag{3.46}$$

where $\mathcal{N}$ represents either a full *explicit* discretization of the convective term if

$$\mathcal{N}(\mathbf{u}^{n-q+1}, \ldots, \mathbf{u}^n; \mathbf{u}^{n+1}) := \sum_{j=0}^{q-1}\alpha_j(\mathbf{u}^{n-j}\cdot\nabla)\mathbf{u}^{n-j} \tag{3.47}$$

or a *semi-implicit* discretization of the convective term if

$$\mathcal{N}(\mathbf{u}^{n-q+1},\dots,\mathbf{u}^n;\mathbf{u}^{n+1}) = (\mathbf{u}^*\cdot\nabla)\mathbf{u}^{n+1}, \qquad \mathbf{u}^* = \sum_{j=0}^{q-1}\alpha_j\mathbf{u}^{n-j}, \tag{3.48}$$

and where $\alpha_j \in \mathbb{R}$ (for $j = 0,\dots,q-1$) are the coefficients of an extrapolation formula of order q. When the convective term is handled by the explicit form (3.47) the Yosida method can be used to solve (3.46), (3.47) and the analysis can be easily extended provided that $\mathcal{N}(\mathbf{u}_{\mathcal{H}}^{n-q+1},\dots,\mathbf{u}_{\mathcal{H}}^{n};\mathbf{u}_{\mathcal{H}}^{n+1}) - \mathcal{N}(\hat{\mathbf{u}}_{\mathcal{H}}^{n-q+1},\dots,\hat{\mathbf{u}}_{\mathcal{H}}^{n};\hat{\mathbf{u}}_{\mathcal{H}}^{n+1})$ is bounded. Otherwise, when the convective term is treated by the semi-implicit form (3.48) the counterpart of (3.6) reads as follows: for $n = n_0,\dots,N_T-1$ solve

$$\begin{cases} \left(\dfrac{\beta_{-1}}{\Delta t}M + C_n(\{\widehat{\mathbf{U}}^{n-j}\}_{j=0}^{q-1})\right)\widehat{\mathbf{U}}^{n+1} + B^T\widehat{\mathbf{P}}^{n+1} = \mathbf{F}_1^{n+1} + \dfrac{1}{\Delta t}\displaystyle\sum_{j=0}^{q-1}\beta_j M\widehat{\mathbf{U}}^{n-j}, \\ B\widehat{\mathbf{U}}^{n+1} - (-B(C_n(\{\widehat{\mathbf{U}}^{n-j}\}_{j=0}^{q-1}))^{-1}B^T - SQ)\widehat{\mathbf{P}}^{n+1} = \mathbf{F}_2^{n+1}, \end{cases} \tag{3.49}$$

where $C_n(\{\mathbf{U}^{n-j}\}_{j=0}^{q-1}) := \frac{\beta_{-1}}{\Delta t}M + \nu K + N(\mathbf{U}^{n-q+1},\dots,\mathbf{U}^n)$, and $N(\mathbf{U}^{n-q+1},\dots,\mathbf{U}^n)$ is the matrix related to the discretization of the convective term. It is immediate to see that matrix Σ is now time-dependent and both the derivation and analysis of Yosida schemes cannot be easily extended starting from the Stokes case.

Nevertheless, many numerical results presented in [8, 7] show that the convergence orders proved in Theorem 3.7 for the Stokes case still hold for Navier–Stokes equations with semi-implicit treatment of the convective term.

4. Numerical results on time-dependent Stokes equations. We consider the computational domain $\Omega = (-1,1)^2$ and $t \in (0,1)$, while the right-hand side, the boundary conditions, and the initial conditions are set so that the exact solution is

$$\begin{aligned} &\mathbf{u}(x,y,t) = [(t+1)\sin(x)\sin((t+1)y), \cos(x)\cos((t+1)y)]^T, \\ &p(x,y,t) = \cos(x)\sin((t+1)y). \end{aligned} \tag{4.1}$$

In Figure 4.1 we report the splitting errors (3.24) obtained by using either BDF1 or BDF2. In Figure 4.2 we show the global errors (3.23) for $p = 0,1,2$ by using

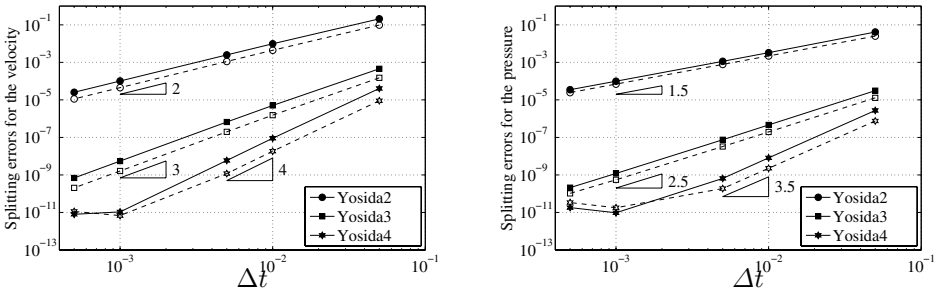


FIG. 4.1. *Splitting errors* (3.24) *for the exact solution* (4.1). *Black symbols refer to runs with BDF*1*, while empty symbols refer to runs with BDF*2. $\nu = 10^{-3}$, $N = 16$, *and* $h = 2$ *(one spectral element). The convergence history of Yosida-*4 *is stopped by the spectral element method discretization error.*

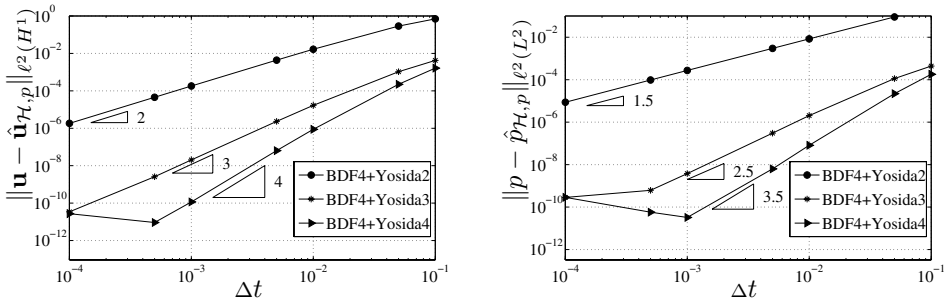


FIG. 4.2. *Global errors* (3.23) *for the exact solution* (4.1). $\nu = 10^{-3}$, $N = 16$, *and* $h = 1$ *(*2×2 *spectral elements). The convergence history of BDF*4 + *Yosida-*4 *is stopped by both spectral element method discretization and rounding errors. The splitting errors prevail over the errors due to BDF*4.

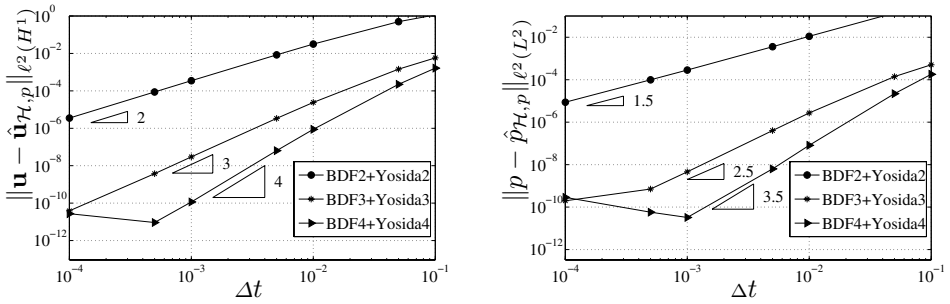


FIG. 4.3. *Global errors* (3.23) *for the exact solution* (4.1). $\nu = 10^{-3}$, $N = 16$, *and* $H = 1$ *(*2×2 *spectral elements). The convergence history of BDF*4 + *Yosida-*4 *is stopped by both spectral element method discretization and rounding errors. The splitting errors due to Yosida-*$(p+2)$ *schemes are of the same order as those produced by BDF*p.

BDF4 time approximation; finally in Figure 4.3 we show the global errors (3.23) for $p = 0, 1, 2$ by using BDFq time approximation, with $q = p + 2$.

Even if Theorem 3.7 ensures that the splitting errors (3.24) behave like $\Delta t^{p+3/2}$ for the velocity and Δt^{p+1} for the pressure, numerical results on the solution (4.1) provide that the splitting errors of *Yosida-*$(p+2)$ schemes behave like Δt^{p+2} for the velocity and like $\Delta t^{p+3/2}$ for the pressure for both BDF1 and BDF2. Global errors depend on convergence order of both BDF and *Yosida-*$(p+2)$ schemes; in particular for the results of Figure 4.2 we have used BDF4, so that, for $p = 0, 1, 2$, splitting errors prevail over the error of BDF. For the results of Figure 4.3 we have coupled *Yosida-*2 with BDF2, *Yosida-*3 with BDF3, and *Yosida-*4 with BDF4 and have obtained a global approximation error of order $p + 2 = q$ for the velocity and of order $p + 3/2$ for the pressure. We refer to [8, 7, 15] for numerical results about the approximation of Navier–Stokes equations.

Acknowledgment. The author is pleased to acknowledge the discussions and helpful comments on the manuscript that have been provided by Lucia Gastaldi, Fausto Saleri, and Alessandro Veneziani.

REFERENCES

[1] C. BERNARDI AND Y. MADAY, *Approximations Spectrales de Problèmes aux Limites Elliptiques*, Springer, Paris, 1992.

[2] J. Blasco and R. Codina, *Space and time error estimates for a first order, pressure stabilized finite element method for the incompressible Navier-Stokes equations*, Appl. Numer. Math., 38 (2001), pp. 475–497.
[3] C. Canuto, M. Y. Hussaini, A. Quarteroni, and T. A. Zang, *Spectral Methods. Fundamentals in Single Domains*, Springer, Heidelberg, 2006.
[4] C. Canuto, M. Y. Hussaini, A. Quarteroni, and T. A. Zang, *Spectral Methods. Evolution to Complex Geometries and Applications to Fluid Dynamics*, Springer, Heidelberg, 2007.
[5] D. H. Carlson, *On real eigenvalues of complex matrices*, Pacific J. Math., 15 (1965), pp. 1119–1129.
[6] E. Emmrich, *Error of the two-step BDF for the incompressible Navier-Stokes problem*, M2AN Math. Model. Numer. Anal., 38 (2004), pp. 757–764.
[7] P. Gervasio and F. Saleri, *Algebraic fractional-step schemes for time-dependent incompressible Navier-Stokes equations*, J. Sci. Comput., 27 (2006), pp. 257–269.
[8] P. Gervasio, F. Saleri, and A. Veneziani, *Algebraic fractional step schemes with spectral methods for the incompressible Navier-Stokes equations*, J. Comput. Phys., 214 (2006), pp. 347–365.
[9] G. Golub and C. V. Loan, *Matrix Computation*, The Johns Hopkins University Press, Baltimore, MD, 1989.
[10] J. G. Heywood and R. Rannacher, *Finite element approximation of the nonstationary Navier–Stokes problem.* I. *Regularity of solutions and second-order error estimates for spatial discretization*, SIAM J. Numer. Anal., 19 (1982), pp. 275–311.
[11] Y. Maday and A. Patera, *Spectral element methods for the incompressible Navier-Stokes equations*, in State-of-the-Art Surveys on Computational Mechanics, A. K. Noor and J. T. Oden, ASME, New York, 1989, pp. 71–143.
[12] A. Quarteroni, F. Saleri, and A. Veneziani, *Analysis of the Yosida method for the incompressible Navier-Stokes equations*, J. Math. Pures Appl. (9), 78 (1999), pp. 473–503.
[13] A. Quarteroni, F. Saleri, and A. Veneziani, *Factorization methods for the numerical approximation of Navier-Stokes equations*, Comput. Methods Appl. Mech. Engrg., 188 (2000), pp. 505–526.
[14] A. Quarteroni and A. Valli, *Numerical Approximation of Partial Differential Equations*, Springer, Heidelberg, 1994.
[15] F. Saleri and A. Veneziani, *Pressure correction algebraic splitting methods for the incompressible Navier–Stokes equations*, SIAM J. Numer. Anal., 43 (2005), pp. 174–194.
[16] R. Temam, *Navier–Stokes Equations and Nonlinear Functional Analysis*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 41, SIAM, Philadelphia, PA, 1983.
[17] F. Zhang, *Matrix Theory*, Springer, New York, 1999.
[18] F. Zhang, *Schur complements and matrix inequalities in the Löwner ordering*, Linear Algebra Appl., 321 (2000), pp. 399–410.

SIAM J. NUMER. ANAL.
Vol. 46, No. 4, pp. 1704–1731
© 2008 Society for Industrial and Applied Mathematics

FINITE ELEMENT APPROXIMATIONS OF THE ERICKSEN–LESLIE MODEL FOR NEMATIC LIQUID CRYSTAL FLOW*

ROLAND BECKER†, XIAOBING FENG‡, AND ANDREAS PROHL§

Abstract. The Ericksen–Leslie model describes dynamics of low molar-mass nematic liquid crystals, where the spatiotemporal distribution of defects defining texture is represented by the director unit vector field $\mathbf{d} : \Omega_T \to \mathbb{S}^2$. It consists of the Navier–Stokes equations with an extra viscous stress tensor, and a convective harmonic map heat flow equation to govern the dynamics of the director field. Two fully discrete finite element methods, for a regularized system using the Ginzburg–Landau regularization and for the limiting Ericksen–Leslie model, are proposed, and well-posedness and related discrete energy laws are established. For the regularized model, unconditional convergence of finite element solutions towards weak solutions of the continuum model as well as convergence towards measure-valued solutions of the limiting Ericksen–Leslie model are verified when the mesh parameters and the regularization parameter successively tend to zero. Computational experiments are also presented to show the importance of balancing numerical and regularization parameters, to compare regularized and direct approaches, and to show numerically the finite-time formation, annihilation, and evolution of point defects.

Key words. nematic liquid crystals, dissipative parabolic system, Ericksen–Leslie model, harmonic map heat flow, finite element method, fully discrete scheme, convergence analysis

AMS subject classifications. 65M12, 65M60, 35K55, 35Q35

DOI. 10.1137/07068254X

1. Introduction.

The Ericksen–Leslie continuum theory gives a successful model for describing dynamics of low molar-mass nematic liquid crystals [14, 23, 22, 34]; this intermediate phase in certain materials is characterized by orientational order but positional disorder of anisotropic molecules, with rigid segments in the structure represented by a unit vector field $\mathbf{d} : \Omega_T \to \mathbb{S}^{N-1}$. The presence of order suggests the possibility of defects, and their spatial distribution defines the texture of a liquid crystal, which characterizes the liquid-crystalline phase and significantly influences flow behavior.

In this paper we propose and analyze some fully discrete finite element approximations of the equations [17, 16, 23]

$$\mathbf{u}_t - \nu\Delta\mathbf{u} + (\mathbf{u}\cdot\nabla)\mathbf{u} + \nabla p + \lambda\,\mathrm{div}\,(\nabla\mathbf{d}\odot\nabla\mathbf{d}) = \mathbf{g} \quad \text{in } \Omega_T := (0,T)\times\Omega\,, \tag{1.1}$$

$$\mathbf{d}_t + (\mathbf{u}\cdot\nabla)\mathbf{d} - \gamma\Delta\mathbf{d} = \gamma|\nabla\mathbf{d}|^2\mathbf{d} \quad \text{in } \Omega_T, \tag{1.2}$$

$$|\mathbf{d}| = 1, \qquad \mathrm{div}\,\mathbf{u} = 0 \quad \text{in } \Omega_T, \tag{1.3}$$

with the following initial and boundary conditions:

*Received by the editors February 13, 2007; accepted for publication (in revised form) January 8, 2008; published electronically April 11, 2008.
http://www.siam.org/journals/sinum/46-4/68254.html

†Laboratoire de Mathematiques Appliquees, Université de Pau et des Pays de l'Adour, BP 1155, 64013 Pau cedex, France (becker@univ-pau.fr).

‡Department of Mathematics, University of Tennessee, Knoxville, TN 37996 (xfeng@math.utk.edu). The work of this author was partially supported by NSF grant DMS-0410266.

§Mathematisches Institut, Universität Tübingen, Auf der Morgenstelle 10, D-72076 Tübingen, Germany (prohl@na.uni-tuebingen.de). The work of this author was supported by the University of Pau.

$$\mathbf{u}(0,\mathbf{x}) = \mathbf{u}_0(\mathbf{x}), \qquad \mathbf{d}(0,\mathbf{x}) = \mathbf{d}_0(\mathbf{x}) \qquad \forall\, \mathbf{x} \in \Omega\,, \tag{1.4}$$

$$\mathbf{u}(t,\mathbf{x}) = 0, \qquad \partial_{\mathbf{n}}\mathbf{d}(t,\mathbf{x}) = 0 \qquad \forall\, (t,\mathbf{x}) \in \partial\Omega_T := (0,T)\times\partial\Omega\,. \tag{1.5}$$

Here and below, let $\Omega \subset \mathbb{R}^N\,(N = 2,3)$ be a bounded domain with a Lipschitz boundary $\partial\Omega$ and outward pointing normal $\mathbf{n} \in \mathbb{S}^{N-1}$. The vector $\mathbf{u}(t,\mathbf{x}) \in \mathbb{R}^N$ and the scalar $p(t,\mathbf{x}) \in \mathbb{R}$ denote the velocity and the pressure of the fluid at the space-time point $(t,\mathbf{x})$, respectively. The vector $\mathbf{d}(t,\mathbf{x}) \in \mathbb{S}^{N-1}$ is known as the *local director* and indicates the directions of the molecules of the liquid crystal at $(t,\mathbf{x}) \in \Omega_T$. Let ν, λ, and γ be positive constants. The vector $\mathbf{g}(t,\mathbf{x}) \in \mathbb{R}^N$ represents a body force on the fluid. The (i,j)th entry of the $N\times N$ matrix $\nabla\mathbf{d}$ is $\frac{\partial d_i}{\partial x_j}$, and $\nabla\mathbf{d}\odot\nabla\mathbf{d}$ stands for the $N\times N$ matrix whose (i,j)th entry is $\mathbf{d}_{x_i}\cdot\mathbf{d}_{x_j}$. Equation (1.1) without the extra stress term $\lambda\,\mathrm{div}\,(\nabla\mathbf{d}\odot\nabla\mathbf{d})$ is the Navier–Stokes equation [36], and (1.2) without the convection term $(\mathbf{u}\cdot\nabla)\mathbf{d}$ is the harmonic map heat flow equation [11, 35].

The system (1.1)–(1.3) was derived by Lin [25] from the original equations of the continuum theory of nematic liquid crystals proposed by Ericksen [17, 16] and Leslie [23]. A nematic liquid crystal behaves like a regular liquid with molecules of similar size; however, it displays anisotropic properties due to the molecule alignment, which is usually described by the local director $\mathbf{d}$ as seen in (1.2). The system (1.1)–(1.3) is known as the simplest approximation of the full Ericksen–Leslie model which retains the basic nonlinear structure and the essential features of the full hydrodynamic equations for nematic liquid crystals. However, it turns out that the mathematical techniques developed for studying the system are applicable to treating the general and more complicated Ericksen–Leslie model. We refer to [17, 23, 25, 26, 28] and the references therein for the derivation of the above system and further physical background on the continuum theory of liquid crystals.

In addition to the well-known barriers for analyzing the Navier–Stokes equations, there are two more features which make the system (1.1)–(1.3) more difficult to study. First, the presence of the extra stress term leads to a strong coupling of the equations. Second, the nonconvex side constraint $|\mathbf{d}| = 1$ causes further difficulties (cf. [9, 7]), both analytically and numerically, and is a main source for possible formation of defects of the system at finite times. Hence, a widely used approach is to approximate the constraint $|\mathbf{d}| = 1$ by a penalty function such as the Ginzburg–Landau approximation $\frac{1}{\varepsilon^2}\mathbf{f}(\mathbf{d})$ for $0 < \varepsilon \ll 1$, where $[\mathbf{f}(\mathbf{z})]_i = (|\mathbf{z}|^2-1)z_i = \frac{\partial}{\partial z_i}F(\mathbf{z})$ for $F(\mathbf{z}) := \frac{1}{4}(|\mathbf{z}|^2-1)^2$ (cf. [25, 26, 9, 7]). The regularized model then reads as

$$\mathbf{u}^\varepsilon_t - \nu\Delta\mathbf{u}^\varepsilon + (\mathbf{u}^\varepsilon\cdot\nabla)\mathbf{u}^\varepsilon + \nabla p^\varepsilon + \lambda\,\mathrm{div}\,(\nabla\mathbf{d}^\varepsilon\odot\nabla\mathbf{d}^\varepsilon) = \mathbf{g} \quad \text{in } \Omega_T\,, \tag{1.6}$$

$$\mathbf{d}^\varepsilon_t + (\mathbf{u}^\varepsilon\cdot\nabla)\mathbf{d}^\varepsilon - \gamma\big(\Delta\mathbf{d}^\varepsilon - \frac{1}{\varepsilon^2}\mathbf{f}(\mathbf{d}^\varepsilon)\big) = 0 \quad \text{in } \Omega_T\,, \tag{1.7}$$

$$\mathrm{div}\,\mathbf{u}^\varepsilon = 0 \quad \text{in } \Omega_T\,, \tag{1.8}$$

subject to initial and boundary conditions (1.4) and (1.5). It is known [28] that as $\varepsilon \searrow 0$, the regularized system (1.6)–(1.8) converges to system (1.1)–(1.3); see section 2 below for more details.

It was observed in [26] that by using the differential identity $\mathrm{div}(\nabla\mathbf{z}\odot\nabla\mathbf{z}) = (\nabla\mathbf{z})^T\Delta\mathbf{z} + \frac{1}{2}\nabla(|\nabla\mathbf{z}|^2)$, (1.6) can be rewritten as

$$\mathbf{u}^\varepsilon_t - \nu\Delta\mathbf{u}^\varepsilon + (\mathbf{u}^\varepsilon\cdot\nabla)\mathbf{u}^\varepsilon + \nabla\widehat{p}^\varepsilon + \lambda\,(\nabla\mathbf{d}^\varepsilon)^T\Delta\mathbf{d}^\varepsilon = \mathbf{g}\,, \quad \widehat{p}^\varepsilon := p^\varepsilon + \frac{\lambda}{2}|\nabla\mathbf{d}^\varepsilon|^2\,. \tag{1.9}$$

This then leads to the key discovery that the system (1.6)–(1.8) satisfies the following basic energy law:

$$\frac{d}{dt}\mathcal{J}_{\varepsilon,\lambda}(\mathbf{u}^\varepsilon,\mathbf{d}^\varepsilon)+\int_\Omega\Big(\nu|\nabla\mathbf{u}^\varepsilon|^2+\lambda\gamma\big|\Delta\mathbf{d}^\varepsilon-\frac{1}{\varepsilon^2}\mathbf{f}(\mathbf{d}^\varepsilon)\big|^2\Big)\,\mathrm{d}\mathbf{x}=\int_\Omega\langle\mathbf{g},\mathbf{u}^\varepsilon\rangle\,\mathrm{d}\mathbf{x}\,, \tag{1.10}$$

where

$$\mathcal{J}_{\varepsilon,\lambda}(\mathbf{u}^\varepsilon,\mathbf{d}^\varepsilon):=\int_\Omega\Big[\,\frac{1}{2}|\mathbf{u}^\varepsilon|^2+\frac{\lambda}{2}|\nabla\mathbf{d}^\varepsilon|^2+\frac{\lambda}{\varepsilon^2}F(\mathbf{d}^\varepsilon)\,\Big]\,\mathrm{d}\mathbf{x}\,. \tag{1.11}$$

Based on the above energy law, Lin and Liu [26] established, for every fixed $\varepsilon>0$ and under appropriate assumptions on the initial data and Ω, global existence of weak solutions and local existence of classical solutions for the regularized system (1.6)–(1.8). Later, they also proved that the singular set of suitable weak solutions of the system has zero one-dimensional Hausdorff measure [27].

The first numerical studies for the regularized problem (1.6)–(1.8) appeared in [31], where Hermite bicubic finite elements were used to approximate the director field $\mathbf{d}^\varepsilon:\Omega_T\to\mathbb{R}^N$. To avoid using Hermite elements, a mixed finite element method was subsequently developed in [32]. Later, a Fourier spectral method was studied in [15]. However, a priori bounds for the numerical solutions in all these works for quantities on the left-hand side of (1.10) are of order $\mathcal{O}\big(\exp(\frac{t}{\varepsilon^2})\big)$, and no uniform discrete energy law was obtained. Consequently, (i) solvability of these schemes by Brouwer fixed point argument requires very small time steps $k=\mathcal{O}\big(\exp(-\frac{1}{\varepsilon^2})\big)$, (ii) no information about the long-time dynamics for the energy is available for the numerical solutions, and (iii) limiting behavior (as $\varepsilon\to0$) of discretizations is not clear.

The first goal of this paper is to construct a practical, fully discrete finite element discretization for the regularized system (1.6)–(1.8). The main features of the proposed scheme include the following: (i) it is unconditionally solvable (with respect to mesh parameters); (ii) the numerical solution satisfies a discrete energy law; (iii) the numerical solution unconditionally converges to a weak solution of (1.6)–(1.8) (as $k,h\to0$) and to a measure-valued solution of (1.1)–(1.3) (as $\varepsilon\to0$). To achieve the above goal, our main idea is to introduce a new "pressure"

$$\widetilde{p}^\varepsilon:=\widehat{p}^\varepsilon+\frac{\lambda}{\varepsilon^2}F(\mathbf{d}^\varepsilon)\equiv p^\varepsilon+\frac{\lambda}{2}|\nabla\mathbf{d}^\varepsilon|^2+\frac{\lambda}{\varepsilon^2}F(\mathbf{d}^\varepsilon) \tag{1.12}$$

and use the fact that $\nabla F(\mathbf{z})=(\nabla\mathbf{z})^T\mathbf{f}(\mathbf{z})$ to rewrite (1.9), and hence (1.6), as

$$\mathbf{u}_t^\varepsilon-\nu\Delta\mathbf{u}^\varepsilon+(\mathbf{u}^\varepsilon\cdot\nabla)\mathbf{u}^\varepsilon+\nabla\widetilde{p}^\varepsilon-\lambda(\nabla\mathbf{d}^\varepsilon)^T\mathbf{w}^\varepsilon=\mathbf{g}\,,\qquad\mathbf{w}^\varepsilon:=-\Delta\mathbf{d}^\varepsilon+\frac{1}{\varepsilon^2}\mathbf{f}(\mathbf{d}^\varepsilon)\,. \tag{1.13}$$

The fully discrete numerical method is then constructed based on this $(\mathbf{u}^\varepsilon,\mathbf{d}^\varepsilon,\widetilde{p}^\varepsilon,\mathbf{w}^\varepsilon)$-formulation (see section 3 for the details). The new formulation allows us to establish a discrete energy law which exactly mimics the differential energy law (1.10). With the help of the discrete energy law, we are able to establish both unconditional well-posedness and convergence of the numerical solution to the solution of system (1.6)–(1.8) for every fixed $\varepsilon>0$ as the mesh parameters independently tend to zero, and to a measure-valued solution of the Ericksen–Leslie model (1.1)–(1.3) as the mesh parameters and successively the parameter ε all approach zero.

It is interesting to note that the sum of the last two terms in the definition of $\widetilde{p}^\varepsilon$ is *exactly* the energy density function for the Ginzburg–Landau equation (cf. [7]). It turns out that this new formulation allows us to use simple Lagrange finite elements to

approximate both $\mathbf{u}^\varepsilon$ and $\mathbf{d}^\varepsilon$ and to avoid using Hermite finite elements or higher order spectral finite elements to approximate $\mathbf{d}^\varepsilon$. From a more practical viewpoint, unless very fine meshes are used, penalization of the sphere constraint $(1.3)_1$ causes some smearing of textures at finite scales of regularization and discretization parameters, which provides motivation to discretize (1.1)–(1.3) directly.

The second goal of this paper concerns the construction of a fully discrete scheme for (1.1)–(1.5) which uses low order finite elements and enjoys a discrete energy law. However, to construct a stable discretization for (1.1)–(1.3) turns out to be even harder than for (1.6)–(1.8), which is now explained. For the differential problem, in order to verify the energy law (for smooth solutions) it is crucial to use the sphere constraint $|\mathbf{d}| = 1$, which is difficult to achieve at the discrete level due to the facts that finite element functions (as well as time differences) have restricted flexibility and lower regularity. In [5], a convergent discretization of the (related nonconvective) harmonic map flow for data $\mathbf{w}_0 : \Omega \to \mathbb{S}^2$ is considered, where $\mathbf{w} : \Omega_T \to \mathbb{S}^2$ satisfies $\mathbf{w}(0,\cdot) = \mathbf{w}_0$ in $\Omega \subset \mathbb{R}^3$, and

$$\mathbf{w}_t - \Delta \mathbf{w} = |\nabla \mathbf{w}|^2 \mathbf{w} \quad \text{in } \Omega_T, \qquad \partial_{\mathbf{n}} \mathbf{w} = 0 \quad \text{on } \partial\Omega_T. \tag{1.14}$$

The construction of a discrete scheme is based on the following reformulation of (1.14):

$$\mathbf{w}_t + \mathbf{w} \times (\mathbf{w} \times \Delta \mathbf{w}) = 0 \quad \text{in } \Omega_T, \qquad \partial_{\mathbf{n}} \mathbf{w} = 0 \quad \text{on } \partial\Omega_T. \tag{1.15}$$

Here, we used the vector identity $\mathbf{a} \times (\mathbf{b} \times \mathbf{c}) = \langle \mathbf{a}, \mathbf{c} \rangle \mathbf{b} - \langle \mathbf{a}, \mathbf{b} \rangle \mathbf{c}$ for all $\mathbf{a}, \mathbf{b}, \mathbf{c} \in \mathbb{R}^3$, together with $|\mathbf{w}| = 1$, to obtain the identity $\mathbf{w} \times (\mathbf{w} \times \Delta \mathbf{w}) = -|\nabla \mathbf{w}|^2 \mathbf{w} - \Delta \mathbf{w}$.

The proposed discretization of (1.15) satisfies a nodal sphere constraint and a discrete energy law (cf. [5]), and hence (subsequences of discrete) solutions unconditionally converge to a weak solution of (1.14) (as $k, h \to 0$). However, the Ericksen–Leslie model is more complicated, mainly due to the convective term in (1.2) which makes it harder to tackle the nodal sphere constraint in a fully discrete context, and due to the strongly nonlinear term $\operatorname{div}\big(\nabla \mathbf{d} \odot \nabla \mathbf{d}\big)$ in (1.1).

Combining reformulations (1.9) and (1.15), we arrive at the following reformulation of (1.1)–(1.3):

$$\mathbf{u}_t - \nu \Delta \mathbf{u} + (\mathbf{u} \cdot \nabla)\mathbf{u} + \nabla \hat{p} + \lambda (\nabla \mathbf{d})^T \Delta \mathbf{d} = \mathbf{g} \quad \text{in } \Omega_T, \tag{1.16}$$

$$\mathbf{d}_t + (\mathbf{u} \cdot \nabla)\mathbf{d} + \gamma\, \mathbf{d} \times (\mathbf{d} \times \Delta \mathbf{d}) = 0 \quad \text{in } \Omega_T, \tag{1.17}$$

$$|\mathbf{d}| = 1, \qquad \operatorname{div} \mathbf{u} = 0 \quad \text{in } \Omega_T. \tag{1.18}$$

Multiplying (1.16) and (1.17) with $\mathbf{u}$ and $\Delta \mathbf{d}$, respectively, and using the formula $\langle \mathbf{a} \times \mathbf{b}, \mathbf{c} \rangle = \langle \mathbf{a} \times \mathbf{c}, \mathbf{b} \rangle$ for all $\mathbf{a}, \mathbf{b}, \mathbf{c} \in \mathbb{R}^3$, we obtain the energy law

$$\frac{d}{dt} \mathcal{J}(\mathbf{u}, \mathbf{d}) + \int_\Omega \Big(\nu\, |\nabla \mathbf{u}|^2 + \lambda\gamma\, |\mathbf{d} \times \Delta \mathbf{d}|^2 \Big)\, \mathrm{d}\mathbf{x} = \int_\Omega \langle \mathbf{g}, \mathbf{u} \rangle\, \mathrm{d}\mathbf{x}\,, \tag{1.19}$$

where

$$\mathcal{J}(\mathbf{u}, \mathbf{d}) = \frac{1}{2} \int_\Omega \big\{ |\mathbf{u}|^2 + \lambda |\nabla \mathbf{d}|^2 \big\}\, \mathrm{d}\mathbf{x}.$$

In section 4, we propose a fully discrete scheme for (1.16)–(1.18) using low-order Lagrangian finite elements (see Algorithm 4.1), and prove its conditional solvability, as well as a discrete energy law. The construction uses the midpoint rule for temporal

discretization and a discrete Laplacian. Appropriate limits (as $k, h \to 0$) are shown to satisfy properties in (1.18) and an inequality version of (1.19). Unfortunately, we are not able to identify limits in the PDEs for the discrete velocity and director fields to recover existing measure-valued solutions to (1.1)–(1.3) in the limit $k, h \to 0$ via Algorithm 4.1; the reason for this is mainly due to the additional stress which leads to the last term on the left-hand side of (1.16).

The paper is organized as follows. In section 2, we introduce notation and cite some useful known results. In section 3, we present an unconditionally solvable scheme for (1.6)–(1.8) which satisfies a discrete energy law. Convergence to weak solutions of (1.6)–(1.8) as $k, h \to 0$ and to measure-valued solutions of (1.1)–(1.3) are shown. Section 4 studies a direct discretization of problem (1.1)–(1.3), where conditional solvability and a discrete energy law are proved. Computational experiments are presented in section 5 to compare the two numerical approaches, to study proper scalings between h, k, and ε for optimal convergence of Algorithm 3.1, and to simulate formation, annihilation, and dynamics of point defects admitted by the model (1.1)–(1.3).

2. Preliminaries. Standard Sobolev space notation is used in this paper; we refer to [1, 36, 30] for their precise definitions. In particular, $(\cdot,\cdot)$ denotes the standard $L^2(\Omega)$-inner product, and $\langle\cdot,\cdot\rangle$ denotes the standard inner product in $\mathbb{R}^N$. We also introduce the function spaces

$$L^2_0(\Omega) := \{q \in L^2(\Omega);\, (q,1) = 0\}, \qquad \boldsymbol{\mathcal{V}} := \{\mathbf{v} \in C^\infty_0(\Omega,\mathbb{R}^3);\, \operatorname{div}\mathbf{v} = 0 \text{ in } \Omega\}\,,$$

$$\mathbf{H} := \{\mathbf{v} \in L^2(\Omega,\mathbb{R}^3);\, \operatorname{div}\mathbf{v} = 0 \text{ weakly in } \Omega \text{ and } \langle\mathbf{v},\mathbf{n}\rangle|_{\partial\Omega} = 0\},$$

$$\mathbf{V} := \{\mathbf{v} \in H^1_0(\Omega,\mathbb{R}^3);\, \operatorname{div}\mathbf{v} = 0 \text{ weakly in } \Omega\}.$$

Let $\mathbf{V}'$ denote the dual of $\mathbf{V}$. In the following, we assume $\Omega \subset \mathbb{R}^N$ to be a bounded Lipschitz domain such that $\Delta : W^{1,2}_0(\Omega) \cap W^{2,2}(\Omega) \to L^2(\Omega)$ is an isomorphism; see [19, 20]. In addition, C is used to denote a generic positive constant which is independent of $\mathbf{u}$, p, $\mathbf{d}$, and the regularization and mesh parameters ε, k, and h.

2.1. Solvability of systems (1.4)–(1.8) and (1.1)–(1.5). Letting $N = 2, 3$, we recall the notion of weak solutions for (1.4)–(1.8).

DEFINITION 2.1. *Let* $(\mathbf{u}_0, \mathbf{d}_0) \in \mathbf{H} \times W^{1,2}(\Omega,\mathbb{R}^N)$, $\mathbf{g} \in L^2\big(0,T;\mathbf{V}')\big)$, *and* $\varepsilon > 0$. *Given* $T > 0$, *a tuple* $(\mathbf{u}^\varepsilon, \mathbf{d}^\varepsilon)$ *with*

$$\mathbf{u}^\varepsilon \in L^2\big(0,T;\mathbf{V}\big) \cap W^{1,\frac{4}{3}}\big(0,T;\mathbf{V}'\big),$$

$$\mathbf{d}^\varepsilon \in L^2\big(0,T;W^{2,2}(\Omega,\mathbb{R}^N)\big) \cap W^{1,\frac{4}{3}}\big(0,T;L^2(\Omega,\mathbb{R}^N)\big)$$

is called a weak solution to (1.4)–(1.8) *if* (1.6)–(1.8) *hold in the distributional sense,* (1.4)–(1.5) *are valid in the sense of traces, and for almost every* $t \in [0,T]$ *there holds*

$$\mathcal{J}_{\varepsilon,\lambda}\big(\mathbf{u}^\varepsilon(t,\cdot),\mathbf{d}^\varepsilon(t,\cdot)\big) + \int_{\Omega_t}\Big(\nu\,|\nabla\mathbf{u}|^2 + \lambda\gamma\,|\Delta\mathbf{d}^\varepsilon - \frac{1}{\varepsilon^2}\mathbf{f}(\mathbf{d}^\varepsilon)|^2\Big)\,\mathrm{d}\mathbf{x}\mathrm{d}s$$

$$\leq \mathcal{J}_{\varepsilon,\lambda}\big(\mathbf{u}_0,\mathbf{d}_0\big) + \int_{\Omega_t}\langle\mathbf{g},\mathbf{u}^\varepsilon\rangle\,\mathrm{d}\mathbf{x}\mathrm{d}s\,. \tag{2.1}$$

Weak solutions to (1.4)–(1.8) for every $\varepsilon > 0$ were constructed in [26] by the Faedo–Galerkin method. They are locally strong for more regular data, and even global if the fluid viscosity $\nu > 0$ is sufficiently large.

A solution concept for the limiting system (1.1)–(1.3) that uses a Radon measure $\nu = \{\nu_{t,\mathbf{x}}\} \in L^\infty_w\big(\Omega_T; \mathcal{M}(\mathbb{R}^{N\times N})\big) \equiv \big(L^1\big(\Omega_T; C_0(\mathbb{R}^{N\times N})\big)\big)'$ to interpret the limiting behavior ($\varepsilon \to 0$) of the extra stress in (1.6) is defined next.

DEFINITION 2.2. *Let* $(\mathbf{u}_0, \mathbf{d}_0) \in \mathbf{H} \times W^{1,2}(\Omega, \mathbb{R}^N)$, *such that* $|\mathbf{d}_0| = 1$ *a.e. in* Ω, *and* $\mathbf{g} \in L^2\big(0,T; \mathbf{V}'\big)$. *Given* $T > 0$, *a triple* $(\mathbf{u}, \mathbf{d}, \nu)$ *with*

$$\mathbf{u} \in L^2\big(0,T;\mathbf{V}\big) \cap W^{1,\frac{4}{3}}\big(0,T;\mathbf{V}'\big),$$

$$(\mathbf{d}, \nu) \in L^\infty\big(0,T; W^{1,2}(\Omega,\mathbb{R}^N)\big) \times L^\infty_w\big(\Omega_T; \mathcal{M}(\mathbb{R}^{N\times N})\big)$$

is called a generalized solution to (1.1)–(1.5) *if*

(i) $|\mathbf{d}| = 1$ *a.e. in* Ω_T,

(ii) *equation* (1.2) *holds in the distributional sense, and for all* $(\boldsymbol{\varphi}, \boldsymbol{\psi}) \in C_0^\infty\big([0,T); \boldsymbol{\mathcal{V}}\big) \times C_0^\infty\big([0,T); C_0^\infty(\Omega,\mathbb{R}^N)\big)$ *there hold*

$$-\int_{\Omega_T} \Big[\langle \mathbf{u}, \boldsymbol{\varphi}_t\rangle - \nu\langle \nabla\mathbf{u}, \nabla\boldsymbol{\varphi}\rangle + \big\langle (\mathbf{u}\cdot\nabla)\boldsymbol{\varphi}, \mathbf{u}\big\rangle + \lambda\left\langle \nabla\boldsymbol{\varphi}, \int_{\mathbb{R}^{N\times N}} \boldsymbol{\lambda}\odot\boldsymbol{\lambda}\,\mathrm{d}\nu(\boldsymbol{\lambda})\right\rangle\Big]\,\mathrm{d}\mathbf{x}\mathrm{d}t = \int_{\Omega_T}\langle \mathbf{g}, \boldsymbol{\varphi}\rangle\,\mathrm{d}\mathbf{x}\mathrm{d}t + \big(\mathbf{u}_0, \boldsymbol{\varphi}(0,\cdot)\big),$$

$$-\int_{\Omega_T} \langle \mathbf{d}, \boldsymbol{\psi}_t\rangle - \langle \nabla\mathbf{d}, \nabla\boldsymbol{\psi}\rangle + \big\langle (\mathbf{U}\cdot\nabla)\boldsymbol{\psi}, \mathbf{d}\big\rangle + \big\langle |\nabla\mathbf{d}|^2\mathbf{d}, \boldsymbol{\psi}\big\rangle\,\mathrm{d}\mathbf{x}\mathrm{d}t = \big(\mathbf{d}_0, \boldsymbol{\psi}(0,\cdot)\big),$$

and

$$\nabla\mathbf{d}(t,\mathbf{x}) = \int_{\mathbb{R}^{N\times N}} \boldsymbol{\lambda}\,\mathrm{d}\nu_{t,\mathbf{x}}(\boldsymbol{\lambda}) \quad \text{a.e. in } \Omega_T\,.$$

The Radon measure ν can be generated from convergent subsequences $\{\nabla\mathbf{d}^\varepsilon\}_{\varepsilon>0}$ which solve (1.6)–(1.8) in the limit, i.e.,

$$\delta_{\nabla\mathbf{d}^\varepsilon} \overset{*}{\rightharpoonup} \nu \qquad \text{in } L^\infty_w\big(\Omega_T; \mathcal{M}(\mathbb{R}^{N\times N})\big) \qquad (\varepsilon \to 0)\,.$$

Local existence of strong solutions ($N = 2$) to (1.1)–(1.5) for more regular data is verified in [33, Chapter 6]. The next lemma states conservation of the initial sphere constraint for classical solutions to (1.1)–(1.5).

LEMMA 2.3. *Suppose* $(\mathbf{u}_0, \mathbf{d}_0) \in \boldsymbol{\mathcal{V}} \times C^2(\Omega, \mathbb{S}^{N-1})$ *and that* $(\mathbf{u}, \mathbf{d})$ *is a classical solution to* (1.1), (1.2), $(1.3)_2$, (1.4), (1.5). *Then* $|\,\mathbf{d}\,| = 1$ *in* Ω_T.

Proof. Multiply (1.2) by $\mathbf{d}$, and set $z = |\,\mathbf{d}\,|^2$. Because of

$$z_t = 2\langle \mathbf{d}, \mathbf{d}_t\rangle\,, \quad \nabla z = 2(\nabla\mathbf{d})\mathbf{d}\,, \quad \Delta z = 2\langle \mathbf{d}, \Delta\mathbf{d}\rangle + 2|\,\nabla\mathbf{d}\,|^2\,,$$

we find

$$z_t - \Delta z + 2\langle (\mathbf{u}\cdot\nabla)\mathbf{d}, \mathbf{d}\rangle = 2|\,\nabla\mathbf{d}\,|^2\big(|\,\mathbf{d}\,|^2 - 1\big) \quad \text{in } \Omega_T\,.$$

Put $w = z - 1$, and note $w(\cdot, 0) = 0$ everywhere in Ω, and

$$w_t - \Delta w + 2\langle (\mathbf{u}\cdot\nabla)\mathbf{d}, \mathbf{d}\rangle = 2|\,\nabla\mathbf{d}\,|^2 w \qquad \text{in } \Omega_T\,.$$

Testing the above equation with w, integrating over Ω_T, and using $\big((\mathbf{u}\cdot\nabla)w, w\big) = 0$ lead to ($T > 0$)

$$\frac{1}{2}\|\,w(T,\cdot)\,\|^2_{L^2} + \int_{\Omega_T} |\,\nabla w\,|^2\,\mathrm{d}\mathbf{x}\mathrm{d}s \le \frac{1}{2}\|\,w(0,\cdot)\,\|^2_{L^2}\exp\left(\int_{\Omega_T}|\,\nabla\mathbf{d}\,|^2\,\mathrm{d}\mathbf{x}\mathrm{d}s\right) = 0$$

by Gronwall's lemma. Hence, $w \equiv 0$ or $|\mathbf{d}| \equiv 1$. ∎

2.2. Finite element spaces. Throughout this section we assume that $\mathcal{T}_h$ is a quasi-uniform triangulation of a bounded polygonal or polyhedral domain $\Omega \subset \mathbf{R}^N$ into triangles or tetrahedrons of maximal diameter $h > 0$ for $N = 2, 3$, i.e., $\overline{\Omega} = \bigcup_{K \in \mathcal{T}_h} \overline{K}$. Let $\mathcal{N}_h = \{\mathbf{x}_\ell\}_{\ell \in L}$ denote the set of all nodes of $\mathcal{T}_h$. Choose

$$\mathbf{Y}_h = \{\mathbf{a}_h \in C(\overline{\Omega}, \mathbb{R}^N) : \mathbf{a}_h|_K \in P_1(K, \mathbb{R}^N)\},$$

and the nodal interpolation operator $\boldsymbol{\mathcal{I}}_{\mathbf{Y}_h} : C(\overline{\Omega}, \mathbb{R}^N) \to \mathbf{Y}_h$, such that $\boldsymbol{\mathcal{I}}_{\mathbf{Y}_h}\boldsymbol{\psi} := \sum_{\mathbf{z} \in \mathcal{N}_h} \boldsymbol{\psi}(\mathbf{z})\boldsymbol{\varphi}_\mathbf{z}$, where $\{\boldsymbol{\varphi}_\mathbf{z} : \mathbf{z} \in \mathcal{N}_h\} \subset \mathbf{Y}_h$ denotes the nodal basis for $\mathbf{Y}_h$ and $\boldsymbol{\psi} \in C(\overline{\Omega}, \mathbb{R}^N)$.

For functions $\boldsymbol{\phi}, \boldsymbol{\psi} \in C(\overline{\Omega}, \mathbb{R}^N)$, we define

$$(\boldsymbol{\phi}, \boldsymbol{\psi})_h := \int_\Omega \mathcal{I}_h(\langle \boldsymbol{\phi}, \boldsymbol{\psi} \rangle)\, \mathrm{d}\mathbf{x} = \sum_{\mathbf{z} \in \mathcal{N}_h} \beta_\mathbf{z} \big\langle \boldsymbol{\phi}(\mathbf{z}), \boldsymbol{\psi}(\mathbf{z}) \big\rangle,$$

$$\|\boldsymbol{\phi}_h\|_h^2 := (\boldsymbol{\phi}_h, \boldsymbol{\phi}_h)_h,$$

where $\beta_\mathbf{z} = \int_\Omega \varphi_\mathbf{z}\, \mathrm{d}\mathbf{x}$ for $\mathbf{z} \in \mathcal{N}_h$. It is easy to check that for all $\boldsymbol{\phi}_h, \boldsymbol{\psi}_h \in \mathbf{Y}_h$,

$$\|\boldsymbol{\phi}_h\|_{L^2} \le \|\boldsymbol{\phi}_h\|_h \le (N+2)^{1/2} \|\boldsymbol{\phi}_h\|_{L^2}, \tag{2.2}$$

$$\big|(\boldsymbol{\phi}_h, \boldsymbol{\psi}_h)_h - (\boldsymbol{\phi}_h, \boldsymbol{\psi}_h)\big| \le Ch \|\boldsymbol{\phi}_h\|_{L^2} \|\nabla \boldsymbol{\psi}_h\|_{L^2}. \tag{2.3}$$

We also define a discrete Laplace operator $\Delta_h : \mathbf{W}^{1,2}(\Omega, \mathbb{R}^N) \to \mathbf{Y}_h$ by

$$(\Delta_h \boldsymbol{\phi}, \boldsymbol{\chi}_h) = -(\nabla \boldsymbol{\phi}, \nabla \boldsymbol{\chi}_h) \qquad \forall\, \boldsymbol{\chi}_h \in \mathbf{Y}_h. \tag{2.4}$$

Note that (see [5]) there exists a constant $C > 0$ such that for all $\boldsymbol{\phi}_h \in \mathbf{Y}_h$ there holds

$$\|\Delta_h \boldsymbol{\phi}_h\|_{L^2} \le Ch^{-2} \|\boldsymbol{\phi}_h\|_{L^2} \quad \text{and} \quad \|\Delta_h \boldsymbol{\phi}_h\|_{L^\infty} \le Ch^{-2} \|\boldsymbol{\phi}_h\|_{L^\infty} \quad \forall\, \boldsymbol{\phi}_h \in \mathbf{Y}_h.$$

The following discrete Sobolev interpolation inequality immediately follows from [21, Lemma 4.4]:

$$\|\nabla \boldsymbol{\phi}_h\|_{L^4} \le C \|\nabla \boldsymbol{\phi}_h\|_{L^2}^{\frac{4-N}{4}} \big[\|\nabla \boldsymbol{\phi}_h\| + \|\Delta_h \boldsymbol{\phi}_h\|_{L^2}\big]^{\frac{N}{4}} \qquad \forall\, \boldsymbol{\phi}_h \in \mathbf{Y}_h. \tag{2.5}$$

The next result is a discrete version of the Aubin–Lions compactness lemma.

LEMMA 2.4. *Let $N = 2, 3$. Suppose $\frac{2N}{N+2} < \ell < \infty$, and let $\mathcal{T}_h$ be a quasi-uniform triangulation of $\Omega \subset \mathbb{R}^N$. Let $\{\boldsymbol{\Phi}_h\}_{h \ge 0} \subset L^2(\Omega_T, \mathbb{R}^N)$ be a sequence of functions, such that $\boldsymbol{\Phi}_h(t, \cdot) \in \mathbf{Y}_h$ for all $t \in [0, T]$, and there holds*

$$\|\boldsymbol{\Phi}_h\|_{W^{1,2}(0,T;L^\ell(\Omega))} + \|\boldsymbol{\Phi}_h\|_{L^\infty(0,T;L^2(\Omega)) \cap L^2(0,T;W^{1,2}(\Omega))} + \|\Delta_h \boldsymbol{\Phi}_h\|_{L^2(\Omega_T)} \le C.$$

Then there exist a subsequence $\{\boldsymbol{\Phi}_h\}_{h>0}$ (not relabeled) and $\boldsymbol{\Phi}$, such that

$$\boldsymbol{\Phi}_h \to \boldsymbol{\Phi} \quad \text{in } L^2\big(0, T; W^{1,2}(\Omega, \mathbb{R}^N)\big) \qquad (h \to 0).$$

Proof. For every $T > 0$ and $h > 0$, consider a weak solution $\mathbf{w}_h : \Omega_T \to \mathbb{R}^3$ of

$$\frac{\partial \mathbf{w}_h}{\partial t} - \Delta \mathbf{w}_h = \mathbf{g}_h \equiv \boldsymbol{\Phi}_{h,t} - \Delta_h \boldsymbol{\Phi}_h \quad \text{in } \Omega_T, \quad \partial_\mathbf{n}[\mathbf{w}_h - \boldsymbol{\Phi}_h] = 0 \quad \text{on } \partial\Omega_T, \tag{2.6}$$

such that $\mathbf{w}_h(0,\cdot) = \mathbf{\Phi}_h(0,\cdot)$. The following maximal L^p-L^ℓ-estimates $(1 < p, \ell < \infty)$ for (2.6) are well known [24]:

$$\begin{aligned}\|(\mathbf{w}_h)_t\|_{L^p(L^\ell)} + \|\mathbf{w}_h\|_{L^p(W^{2,\ell})} &\le C\|\mathbf{g}_h\|_{L^p(L^\ell)} \\ &\le C\left[\|\mathbf{\Phi}_{h,t}\|_{L^p(L^\ell)} + \|\Delta_h\mathbf{\Phi}_h\|_{L^p(L^\ell)}\right].\end{aligned}$$

By Sobolev embedding $W^{2,\ell}(\Omega) \overset{c}{\hookrightarrow} W^{1,2}(\Omega)$ for $\ell > \frac{2N}{2+N}$ and the Aubin–Lions compactness lemma we conclude that there exists a subsequence (not relabeled) of $\{\mathbf{w}_h\}$ such that $\mathbf{w}_h \to \mathbf{\Phi}$ in $L^2\big(0,T;W^{1,2}(\Omega,\mathbb{R}^N)\big)$.

Next, testing (2.6) with $[\boldsymbol{\mathcal{I}}_{\mathbf{Y}_h}\mathbf{w}_h - \mathbf{\Phi}_h](t,\cdot) \in \mathbf{Y}_h$ we obtain

$$\begin{aligned}&\frac{1}{2}\frac{d}{dt}\|\mathbf{w}_h - \mathbf{\Phi}_h\|_{L^2}^2 + \|\nabla\big[\mathbf{w}_h - \mathbf{\Phi}_h\big]\|_{L^2}^2 \\ &\quad= \Big([\mathbf{w}_h - \mathbf{\Phi}_h]_t, \mathbf{w}_h - \boldsymbol{\mathcal{I}}_{\mathbf{Y}_h}\mathbf{w}_h\Big) + \Big(\nabla[\mathbf{w}_h - \mathbf{\Phi}_h], \nabla[\mathbf{w}_h - \boldsymbol{\mathcal{I}}_{\mathbf{Y}_h}\mathbf{w}_h]\Big) \\ &\quad\le \big(\|\mathbf{w}_{h,t}\|_{L^\ell} + \|\mathbf{\Phi}_{h,t}\|_{L^\ell}\big)\|\mathbf{w}_h - \boldsymbol{\mathcal{I}}_{\mathbf{Y}_h}\mathbf{w}_h\|_{L^{\ell'}} \\ &\qquad\qquad + \|\nabla[\mathbf{w}_h - \mathbf{\Phi}_h]\|_{L^2}\|\mathbf{w}_h - \boldsymbol{\mathcal{I}}_{\mathbf{Y}_h}\mathbf{w}_h\|_{W^{1,2}}.\end{aligned}$$

Put $p = 2$ in the above estimate. Integrating in time, applying Young's inequality, and using standard interpolation results yield

$$\begin{aligned}&\|\mathbf{w}_h - \mathbf{\Phi}_h\|_{L^\infty(L^2)}^2 + \frac{1}{2}\|\mathbf{w}_h - \mathbf{\Phi}_h\|_{L^2(W^{1,2})}^2 \\ &\qquad\le C\left[\|\mathbf{w}_{h,t}\|_{L^2(L^\ell)} + \|\mathbf{\Phi}_{h,t}\|_{L^2(L^\ell)}\right]\|\mathbf{w}_h - \boldsymbol{\mathcal{I}}_{\mathbf{Y}_h}\mathbf{w}_h\|_{L^2(L^{\ell'})} \\ &\qquad\quad + 2\|\mathbf{w}_h - \boldsymbol{\mathcal{I}}_{\mathbf{Y}_h}\mathbf{w}\|_{L^2(W^{1,2})}^2 \\ &\qquad\le C\Big[h^{2+N\min\{1-\frac{2}{\ell},0\}} + h^{2[1+N\min\{\frac{\ell-2}{2\ell},0\}]}\Big]\|\mathbf{w}_h\|_{L^2(W^{2,\ell})}.\end{aligned}$$

Hence, for $\ell > \frac{2N}{N+2}$, we have $\mathbf{w}_h - \mathbf{\Phi}_h \to 0$ in $L^2\big(0,T;W^{1,2}(\Omega,\mathbb{R}^N)\big)$ as $h \to 0$. Consequently, $\mathbf{\Phi}_h \to \mathbf{\Phi}$ in $L^2\big(0,T;W^{1,2}(\Omega,\mathbb{R}^N)\big)$. ∎

For a nonnegative integer r, let $P_r(K)$ denote the space of polynomials of degree less than or equal to r on K. Let $(\mathbf{X}_h, M_h)$ be an LBB-stable mixed finite element pair; i.e., $\mathbf{X}_h \subset W_0^{1,2}(\Omega,\mathbb{R}^N)$ and $M_h \subset L_0^2(\Omega)$ satisfy the inf-sup condition

$$\sup_{\mathbf{v}_h \in \mathbf{X}_h} \frac{(\operatorname{div}\mathbf{v}_h, q_h)}{\|\nabla\mathbf{v}_h\|_{L^2}} \ge C\|q_h\|_{L^2} \qquad \forall q_h \in M_h. \tag{2.7}$$

A well-known example that satisfies (2.7) is (see [20, 10])

$$\begin{aligned}\mathbf{X}_h &= \{\mathbf{v}_h \in C^0(\overline{\Omega},\mathbb{R}^N) \cap W_0^{1,2}(\Omega,\mathbb{R}^N);\ \mathbf{v}_h|_K \in P_2(K,\mathbb{R}^N)\},\\ M_h &= \{q_h \in L_0^2(\Omega);\ q_h|_K \in P_0(K)\}.\end{aligned}$$

Let

$$\mathbf{V}_h := \{\mathbf{v}_h \in \mathbf{X}_h :\ (\operatorname{div}\mathbf{v}_h, q_h) = 0\ \forall q_h \in M_h\}.$$

In the following, we use L^2-orthogonal projections $Q_{\mathbf{V}_h} : L^2(\Omega,\mathbb{R}^N)\to\mathbf{V}_h$ and $Q_{\mathbf{Y}_h} : L^2(\Omega,\mathbb{R}^N)\to\mathbf{Y}_h$, i.e., for all $\mathbf{z}\in L^2(\Omega,\mathbb{R}^N)$

$$(\mathbf{z}-Q_{\mathbf{V}_h}\mathbf{z},\mathbf{v}_h)=0\qquad\forall\mathbf{v}_h\in\mathbf{V}_h\,,\qquad(\mathbf{z}-Q_{\mathbf{Y}_h}\mathbf{z},\mathbf{a}_h)=0\qquad\forall\mathbf{a}_h\in\mathbf{Y}_h\,,$$

and the $W^{1,2}$-orthogonal projection $R_{\mathbf{Y}_h} : W^{1,2}(\Omega,\mathbb{R}^N)\to\mathbf{Y}_h$ such that for all $\mathbf{z}\in W^{1,2}(\Omega,\mathbb{R}^N)$

$$\big(\nabla[\mathbf{a}-R_{\mathbf{Y}_h}\mathbf{z}],\nabla\mathbf{b}_h\big)+\big(\mathbf{a}-R_{\mathbf{Y}_h}\mathbf{z},\mathbf{b}_h\big)=0\qquad\forall\,\mathbf{b}_h\in\mathbf{Y}_h\,.$$

The following estimates can be found in [21]:

$$\|\,\mathbf{z}-Q_{\mathbf{V}_h}\mathbf{z}\,\|_2+h\|\,\nabla(\mathbf{z}-Q_{\mathbf{V}_h}\mathbf{z})\,\|_{L^2}\le Ch^2\|\,D^2\mathbf{z}\,\|_{L^2}\quad\forall\mathbf{z}\in\mathbf{V}\cap W^{2,2}(\Omega,\mathbb{R}^N),\tag{2.8}$$

$$\|\,\mathbf{z}-Q_{\mathbf{V}_h}\mathbf{z}\,\|_2\le Ch\,\|\,\nabla\mathbf{z}\,\|_{L^2}\quad\forall\,\mathbf{z}\in\mathbf{V}.\tag{2.9}$$

By the inverse inequality, a standard duality argument which gives $\|\,\mathbf{a}-R_{\mathbf{Y}_h}\mathbf{a}\,\|_{L^2}\le Ch\,\|\,\nabla\mathbf{a}\,\|_{L^2}$ for all $\mathbf{a}\in W^{1,2}(\Omega,\mathbb{R}^N)$, together with (2.9) and $W^{1,2}$-stability of $R_{\mathbf{Y}_h}$: $W^{1,2}(\Omega;\,\mathbb{R}^N)\to\mathbf{Y}_h$, we have

$$\begin{aligned}\|\,Q_{\mathbf{V}_h}\mathbf{z}\,\|_{W^{1,2}}&\le\|\,R_{\mathbf{Y}_h}\mathbf{z}-Q_{\mathbf{V}_h}\mathbf{z}\,\|_{W^{1,2}}+\|\,R_{\mathbf{Y}_h}\mathbf{z}\,\|_{W^{1,2}}\\&\le Ch^{-1}\|\,R_{\mathbf{Y}_h}\mathbf{z}\pm\mathbf{z}-Q_{\mathbf{V}_h}\mathbf{z}\,\|_{L^2}+\|\,\mathbf{z}\,\|_{W^{1,2}}\\&\le C\|\,\mathbf{z}\,\|_{W^{1,2}}\qquad\forall\,\mathbf{z}\in\mathbf{V}\,.\end{aligned}\tag{2.10}$$

Hence, $Q_{\mathbf{V}_h}$ is stable in $W^{1,2}(\Omega)$.

2.3. Discrete time-derivatives and interpolations. Given a time-step size $k>0$ and a sequence $\{\boldsymbol{\phi}_j\}$ in some Banach space $\mathbf{X}$, we set $d_t\boldsymbol{\phi}^j:=k^{-1}\big(\boldsymbol{\phi}^j-\boldsymbol{\phi}^{j-1}\big)$ and $\boldsymbol{\phi}^{j+1/2}:=\frac12\big(\boldsymbol{\phi}^j+\boldsymbol{\phi}^{j+1}\big)$ for $j\ge0$. Note that $\langle d_t\boldsymbol{\phi}^{j+1},\boldsymbol{\phi}^{j+1/2}\rangle=\frac12 d_t\|\,\boldsymbol{\phi}^{j+1}\,\|_X^2$ if $\mathbf{X}$ is a Hilbert space. Piecewise constant interpolations of $\big\{\boldsymbol{\phi}^j\big\}$ are defined for $t\in[t_j,t_{j+1})$ and $j\in\{0,1,\ldots,J-1\}$ by

$$\overline{\boldsymbol{\Phi}}(t):=\boldsymbol{\phi}^{j+1/2},\quad\boldsymbol{\Phi}^-(t):=\boldsymbol{\phi}^j\,,\quad\boldsymbol{\Phi}^+(t):=\boldsymbol{\phi}^{j+1}\,,$$

and a piecewise linear interpolation on $[t_j,t_{j+1})$ for $j\in\{0,1,\ldots,J-1\}$ is defined by

$$\boldsymbol{\Phi}(t):=\frac{t-t_j}{k}\boldsymbol{\phi}^{j+1}+\frac{t_{j+1}-t}{k}\boldsymbol{\phi}^j.$$

Note that there holds

$$\|\,\boldsymbol{\Phi}^\pm-\boldsymbol{\Phi}\,\|_X+\|\,\overline{\boldsymbol{\Phi}}-\boldsymbol{\Phi}\,\|_X\le2k\,\|\,d_t\boldsymbol{\Phi}\,\|_X.$$

3. Fully discrete finite element methods for the regularized Ericksen–Leslie model (1.6)–(1.8). The plan of this section is as follows. A priori bounds for solutions to the regularized Ericksen–Leslie model are collected in section 3.1. A fully discrete finite element approximation is studied in section 3.2, and unconditional solvability and a discrete energy law are shown in Lemma 3.4 and Theorem 3.5; convergence of numerical solutions to weak solutions of (1.6)–(1.8) and to generalized solutions of (1.1)–(1.3) is proved in Theorem 3.7.

3.1. A priori bounds for regular solutions to (1.6)–(1.8). We recall the basic energy law and derive some additional a priori estimates for solutions of the system (1.6)–(1.8); most of the a priori estimates to be given below were derived in [26] for the case $\nu = \lambda = \gamma = \varepsilon = 1$.

We first state some properties of the differential of the vector-valued function $\mathbf{f}(\mathbf{a}) = \big(|\mathbf{a}|^2 - 1\big)\mathbf{a}$. Since the proofs are elementary, we omit them.

LEMMA 3.1. *Let $\mathbf{a} \in \mathbf{R}^N$; then the following properties hold for the differential $D\mathbf{f}$ of $\mathbf{f}(\mathbf{a}) = (|\mathbf{a}|^2 - 1)\mathbf{a}$:*

(i) $D\mathbf{f}(\mathbf{a}) = (|\mathbf{a}|^2 - 1)I + 2\mathbf{a}\mathbf{a}^T$.

(ii) $\mu_1(|\mathbf{a}|) = 3|\mathbf{a}|^2 - 1$ *and* $\mu_2(|\mathbf{a}|) = |\mathbf{a}|^2 - 1$ *are the only two distinct eigenvalues of $D\mathbf{f}(\mathbf{a})$ with multiplicities 1 and $N-1$, respectively. Moreover, $\mathcal{V}_{\mu_1} = span\{\mathbf{a}\}$ and $\mathcal{V}_{\mu_2} = \mathcal{V}_{\mu_1}^{\perp}$ are the eigenspaces corresponding to μ_1 and μ_2.*

(iii) $D\mathbf{f}(\mathbf{a})$ *is a singular $N \times N$ matrix for $|\mathbf{a}| = 1$.*

(iv) $\langle D\mathbf{f}(\mathbf{a})\mathbf{b}, \mathbf{b}\rangle = (|\mathbf{a}|^2 - 1)|\mathbf{b}|^2 + 2|\langle \mathbf{a}, \mathbf{b}\rangle|^2$ *for all $\mathbf{b} \in \mathbf{R}^N$.*

(v) $\langle D\mathbf{f}(\mathbf{a})A, A\rangle = (|\mathbf{a}|^2 - 1)|A|^2 + 2|A^T\mathbf{a}|^2$ *for all $A \in \mathbf{R}^{N\times N}$.*

The next lemma derives the basic energy law associated with (1.6)–(1.8) and its induced a priori estimates.

LEMMA 3.2. *Let $\Omega \subset \mathbb{R}^N$ $(N = 2, 3)$ be a convex polyhedron and let $T > 0$. Let $\mathbf{g} \in L^2\big((0,T); \mathbf{V}'\big)$ and the initial values $\mathbf{u}_0^\varepsilon \in \mathbf{H}$ and $\mathbf{d}_0^\varepsilon \in W^{1,2}(\Omega, \mathbb{R}^N)$ such that $|\mathbf{d}_0^\varepsilon| \le 1$ a.e. in Ω. Then every regular smooth solution $(\mathbf{u}^\varepsilon, \mathbf{d}^\varepsilon)$ of the system (1.6)–(1.8), (1.4)–(1.5) satisfies*

(i) $\operatorname*{ess\,sup}_{t\in[0,T]} \| \mathbf{d}^\varepsilon(t,\cdot) \|_{L^\infty} \le 1\,,$

(ii) $\operatorname*{ess\,sup}_{t\in[0,T]} \Big\{ \frac{1}{2} \| \mathbf{u}^\varepsilon(t,\cdot) \|_{L^2}^2 + \frac{\lambda}{2} \| \nabla \mathbf{d}^\varepsilon(t,\cdot) \|_{L^2}^2 + \lambda\varepsilon^{-2}\big(F(\mathbf{d}^\varepsilon(t,\cdot)), 1\big) \Big\} \le C\,,$

(iii) $\dfrac{\nu}{2}\displaystyle\int_0^T \| \nabla \mathbf{u}^\varepsilon(t,\cdot) \|_{L^2}^2 \mathrm{d}t + \left\{ \begin{array}{l} \lambda\gamma \displaystyle\int_0^T \| \Delta \mathbf{d}^\varepsilon(t,\cdot) - \varepsilon^{-2}\mathbf{f}(\mathbf{d}^\varepsilon(t,\cdot)) \|_{L^2}^2 \mathrm{d}t \\ \dfrac{\lambda}{\gamma} \displaystyle\int_0^T \| \mathbf{d}_t^\varepsilon(t,\cdot) + (\mathbf{u}^\varepsilon(t,\cdot)\cdot\nabla)\mathbf{d}^\varepsilon(t,\cdot) \|_{L^2}^2 \mathrm{d}t \end{array} \right\} \le C,$

(iv) $\displaystyle\int_0^T \Big\{ \| \mathbf{u}_t^\varepsilon(t,\cdot) \|_{(\mathbf{V}\cap L^\infty)'}^2 + \| \mathbf{d}_t^\varepsilon(t,\cdot) \|_{L^{\frac{3}{2}}}^2 \Big\} \mathrm{d}t \le C,$

(v) $\displaystyle\int_0^T \Big\{ \| \Delta \mathbf{d}^\varepsilon(t,\cdot) \|_{L^2}^2 + \| \mathbf{d}_t^\varepsilon(t,\cdot) \|_{L^2}^{\frac{4}{3}} + \| \mathbf{u}_t^\varepsilon(t,\cdot) \|_{\mathbf{V}'}^{\frac{4}{3}} \Big\} \mathrm{d}t \le C\,\varepsilon^{-2}\,.$

Proof. Assertion (i) follows from a weak maximum principle on testing (1.7) with $(|\mathbf{d}^\varepsilon| - 1)^+\mathbf{d}^\varepsilon/|\mathbf{d}^\varepsilon|$, where

$$\frac{(z-1)^+}{z} := \left\{ \begin{array}{ll} 0 & \text{if } z \in [0,1], \\ \frac{z-1}{z} & \text{if } z \ge 1. \end{array} \right.$$

Next, multiplying (1.13) with $\mathbf{u}^\varepsilon$, and (1.7) with $\lambda\big(-\Delta\mathbf{d}^\varepsilon + \varepsilon^{-2}\mathbf{f}(\mathbf{d}^\varepsilon)\big)$, and adding the resulting equations lead to the basic energy equation (1.10).

The bounds (ii), (iii) follow easily from integrating (1.10) with respect to t from 0 to T on noting the inequality

$$|\langle \mathbf{g}, \mathbf{u}^\varepsilon\rangle| \le \frac{\nu}{2} \| \nabla \mathbf{u}^\varepsilon \|_{L^2}^2 + \frac{C_1}{\nu} \| \mathbf{g} \|_{\mathbf{V}'}^2\,.$$

To show $(\mathrm{iv})_1$, testing (1.13) with $\mathbf{v} \in \mathbf{V} \cap L^\infty(\Omega, \mathbb{R}^3)$ yields

$$\begin{aligned}\big(\mathbf{u}_t^\varepsilon, \mathbf{v}\big) &= -\nu\big(\nabla \mathbf{u}^\varepsilon, \nabla \mathbf{v}\big) + \lambda\big(\mathbf{w}^\varepsilon \nabla \mathbf{d}^\varepsilon, \mathbf{v}\big) - \big((\mathbf{u}^\varepsilon \cdot \nabla)\mathbf{u}^\varepsilon, \mathbf{v}\big) + \langle \mathbf{g}, \mathbf{v}\rangle \\ &\le \nu \| \nabla \mathbf{u}^\varepsilon \|_{L^2} \| \nabla \mathbf{v} \|_{L^2} + \lambda \| \mathbf{w}^\varepsilon \|_{L^2} \| \nabla \mathbf{d}^\varepsilon \|_{L^2} \| \mathbf{v} \|_{L^\infty} \\ &\qquad + \| \nabla \mathbf{u}^\varepsilon \|_{L^2} \| \mathbf{u}^\varepsilon \|_{L^2} \| \mathbf{v} \|_{L^\infty} + \| \mathbf{g} \|_{\mathbf{V}'} \| \mathbf{v} \|_{W^{1,2}},\end{aligned}$$

which together with (ii), (iii) implies the assertion. To verify $(\mathrm{iv})_2$, note that (1.7) can be rewritten as $\mathbf{d}_t^\varepsilon = -\gamma \mathbf{w}^\varepsilon - (\mathbf{u}^\varepsilon \cdot \nabla)\mathbf{d}^\varepsilon$. The second assertion in (iv) then uses (ii), (iii), and the inequality

$$\| (\mathbf{u}^\varepsilon \cdot \nabla)\mathbf{d}^\varepsilon \|_{L^{\frac{3}{2}}} \le \| \mathbf{u}^\varepsilon \|_{L^6} \| \nabla \mathbf{d}^\varepsilon \|_{L^2}.$$

To show $(\mathrm{v})_1$, multiply the equation $\mathbf{w}^\varepsilon = -\Delta \mathbf{d}^\varepsilon + \varepsilon^{-2} \mathbf{f}(\mathbf{d}^\varepsilon)$ with $-\Delta \mathbf{d}^\varepsilon$:

$$\begin{aligned}\| \Delta \mathbf{d}^\varepsilon \|_{L^2}^2 &= -(\mathbf{w}^\varepsilon, \Delta \mathbf{d}^\varepsilon) + \frac{1}{\varepsilon^2}\big(\mathbf{f}(\mathbf{d}^\varepsilon), \Delta \mathbf{d}^\varepsilon\big) \\ &\le \| \mathbf{w}^\varepsilon \|_{L^2} \| \Delta \mathbf{d}^\varepsilon \|_{L^2} - \frac{1}{\varepsilon^2}\big(D\mathbf{f}(\mathbf{d}^\varepsilon)\nabla \mathbf{d}^\varepsilon, \nabla \mathbf{d}^\varepsilon\big) \\ &\le \frac{1}{2} \| \Delta \mathbf{d}^\varepsilon \|_{L^2}^2 + \frac{1}{2} \| \mathbf{w}^\varepsilon \|_{L^2}^2 + \frac{1}{\varepsilon^2} \| \nabla \mathbf{d}^\varepsilon \|_{L^2}^2.\end{aligned}$$

Here we have used (v) of Lemma 3.1 in combination with (i). The result then is a consequence of (iii). Estimate $(\mathrm{v})_2$ is an immediate consequence of (iii), $(\mathrm{v})_1$, and the bounds

$$\| \mathbf{d}_t^\varepsilon \|_{L^2}^{\frac{4}{3}} \le C \left[\| \mathbf{w}_\varepsilon \|_{L^2}^{\frac{4}{3}} + \| (\mathbf{u}^\varepsilon \cdot \nabla)\mathbf{d}^\varepsilon \|_{L^2}^{\frac{4}{3}} \right],$$

$$\| (\mathbf{u}^\varepsilon \cdot \nabla)\mathbf{d}^\varepsilon \|_{L^2}^{\frac{4}{3}} \le C \| \mathbf{u}^\varepsilon \|_{L^4}^{\frac{4}{3}} \| \nabla \mathbf{d}^\varepsilon \|_{L^4}^{\frac{4}{3}} \le C \left[\| \mathbf{u}^\varepsilon \|_{L^2}^{\frac{2}{3}} \| \nabla \mathbf{u}^\varepsilon \|_{L^2}^2 + \| \nabla \mathbf{d}^\varepsilon \|_{L^2}^{\frac{2}{3}} \| \mathbf{d}^\varepsilon \|_{W^{2,2}}^2 \right].$$

Finally, to show $(\mathrm{v})_3$, testing (1.13) with $\mathbf{v} \in \mathbf{V}$ yields

$$\begin{aligned}\big(\mathbf{u}_t^\varepsilon, \mathbf{v}\big) &= -\nu\big(\nabla \mathbf{u}^\varepsilon, \nabla \mathbf{v}\big) + \lambda\big(\mathbf{w}^\varepsilon \nabla \mathbf{d}^\varepsilon, \mathbf{v}\big) - \big((\mathbf{u}^\varepsilon \cdot \nabla)\mathbf{u}^\varepsilon, \mathbf{v}\big) + \langle \mathbf{g}, \mathbf{v}\rangle \\ &\le \nu \| \nabla \mathbf{u}^\varepsilon \|_{L^2} \| \nabla \mathbf{v} \|_{L^2} + \lambda \| \mathbf{w}^\varepsilon \|_{L^2} \| \nabla \mathbf{d}^\varepsilon \|_{L^3} \| \mathbf{v} \|_{L^6} \\ &\qquad + \| \nabla \mathbf{u}^\varepsilon \|_{L^2} \| \mathbf{u}^\varepsilon \|_{L^3} \| \mathbf{v} \|_{L^6} + \| \mathbf{g} \|_{W^{-1,2}} \| \mathbf{v} \|_{W^{1,2}} \\ &\le \nu \| \nabla \mathbf{u}^\varepsilon \|_{L^2} \| \nabla \mathbf{v} \|_{L^2} + \lambda \| \mathbf{w}^\varepsilon \|_{L^2} \| \mathbf{d}^\varepsilon \|_{W^{2,2}}^{\frac{1}{2}} \| \nabla \mathbf{d}^\varepsilon \|_{L^2}^{\frac{1}{2}} \| \mathbf{v} \|_{L^6} \\ &\qquad + \| \nabla \mathbf{u}^\varepsilon \|_{L^2}^{\frac{3}{2}} \| \mathbf{u}^\varepsilon \|_{L^2}^{\frac{1}{2}} \| \mathbf{v} \|_{L^6} + \| \mathbf{g} \|_{W^{-1,2}} \| \mathbf{v} \|_{W^{1,2}},\end{aligned}$$

which together with (ii), (iii), and $(\mathrm{v})_1$ leads to the assertion. ∎

A priori estimates for $(\mathbf{u}^\varepsilon, \mathbf{d}^\varepsilon)$ in higher norms show exponential dependence on $\varepsilon > 0$, in general. Without proof, we collect some bounds for strong solutions to (1.6)–(1.8) for $\Omega \subset \mathbb{R}^2$. For the three-dimensional case, corresponding estimates hold locally in time, unless a smallness assumption for initial data is made.

LEMMA 3.3. *In addition to the assumptions of Lemma* 3.2, *let* $\Omega = (0,1)^2$, *and suppose periodic boundary conditions. Let* $\mathbf{g} \in L^2\big(0, T; L^2(\Omega, \mathbb{R}^2)\big)$ *and* $(\mathbf{u}_0^\varepsilon, \mathbf{d}_0^\varepsilon) \in \mathbf{V} \times W^{2,2}(\Omega, \mathbb{R}^2)$. *Then every regular solution* $(\mathbf{u}^\varepsilon, \mathbf{d}^\varepsilon)$ *of problem* (1.6)–(1.8) *satisfies*

$$\text{(i)}\quad \operatorname*{ess\,sup}_{[0,T]}\left\{\|\nabla\mathbf{u}^\varepsilon\|_{L^2}^2+\frac{\lambda^2}{\gamma}\|\Delta\mathbf{d}^\varepsilon\|_{L^2}^2\right\}$$
$$+\int_0^T\left(\nu\,\|\Delta\mathbf{u}^\varepsilon(s,\cdot)\|_{L^2}^2+\lambda^2\|\nabla\Delta\mathbf{d}^\varepsilon(s,\cdot)\|_{L^2}^2\right)\mathrm{d}s\le C\exp\left(\frac{T}{\varepsilon^2}\right),$$

$$\text{(ii)}\quad \int_0^T\left(\|\mathbf{u}_t^\varepsilon(s,\cdot)\|_{L^2}^2+\|\nabla p^\varepsilon(s,\cdot)\|_{L^2}^2\right)\mathrm{d}s\le C\exp\left(\frac{T}{\varepsilon^2}\right).$$

3.2. A fully discrete scheme for (1.6)–(1.8): Unconditional solvability, discrete energy law, and convergence analysis. As mentioned in section 1, our fully discrete finite element approximation for (1.6)–(1.8) is based on the formulation (1.9) using the new "pressure" $\tilde{p}^\varepsilon$.

ALGORITHM 3.1.

(1) *Set* $\mathbf{u}_h^0=Q_{\mathbf{V}_h}\mathbf{u}_0^\varepsilon$ *and* $\mathbf{d}_h^0:=R_{\mathbf{Y}_h}\mathbf{d}_0^\varepsilon$.

(2) *For* $m=1,\dots,M$, *let* $\mathbf{f}_h^m:=|\mathbf{d}_h^m|^2\mathbf{d}_h^m-\mathbf{d}_h^{m-1}$. *Find* $(\mathbf{u}_h^m,\mathbf{d}_h^m,\tilde{p}_h^m,\mathbf{w}_h^m)\in\mathbf{X}_h\times\mathbf{Y}_h\times M_h\times\mathbf{Y}_h$ *such that for all* $(\mathbf{v}_h,\mathbf{a}_h,\mathbf{b}_h,q_h)\in\mathbf{X}_h\times\big[\mathbf{Y}_h\big]^2\times M_h$

$$\big(d_t\mathbf{u}_h^m,\mathbf{v}_h\big)+\nu\big(\nabla\mathbf{u}_h^m,\nabla\mathbf{v}_h\big)+\big((\mathbf{u}_h^{m-1}\cdot\nabla)\mathbf{u}_h^m,\mathbf{v}_h\big)+\frac{1}{2}\big((\operatorname{div}\mathbf{u}_h^{m-1})\,\mathbf{u}_h^m,\mathbf{v}_h\big)$$
$$-\big(\tilde{p}_h^m,\operatorname{div}\mathbf{v}_h\big)-\lambda\big((\nabla\mathbf{d}_h^{m-1})^T\mathbf{w}_h^m,\mathbf{v}_h\big)=\big\langle\mathbf{g}(t_m,\cdot),\mathbf{v}_h\big\rangle,\tag{3.1}$$

$$\big(d_t\mathbf{d}_h^m,\mathbf{a}_h\big)+\big((\mathbf{u}_h^m\cdot\nabla)\mathbf{d}_h^{m-1},\mathbf{a}_h\big)+\gamma\big(\mathbf{w}_h^m,\mathbf{a}_h\big)=0,\tag{3.2}$$

$$\big(\operatorname{div}\mathbf{u}_h^m,q_h\big)=0,\tag{3.3}$$

$$\big(\nabla\mathbf{d}_h^m,\nabla\mathbf{b}_h\big)+\frac{1}{\varepsilon^2}\big(\mathbf{f}_h^m,\mathbf{b}_h\big)_h-\big(\mathbf{w}_h^m,\mathbf{b}_h\big)=0.\tag{3.4}$$

The crux for proving unconditional solvability and convergence of Algorithm 3.1 is to establish a discrete energy law, which mimics the basic energy law (1.10) for (1.6)–(1.8). Note that not every numerical method possesses such a property. Define $\mathcal{J}_{\varepsilon,\lambda,h}:\mathbf{X}_h\times\mathbf{Y}_h\to\mathbb{R}$ by

$$\mathcal{J}_{\varepsilon,\lambda,h}\big(\boldsymbol{\varphi}_h,\boldsymbol{\psi}_h\big):=\int_\Omega\Big[\frac{1}{2}|\boldsymbol{\varphi}_h|^2+\frac{\lambda}{2}|\nabla\boldsymbol{\psi}_h|^2\Big]\mathrm{d}\mathbf{x}+\frac{\lambda}{\varepsilon^2}\big(F(\boldsymbol{\varphi}_h),1\big)_h.\tag{3.5}$$

LEMMA 3.4. *Let* $\big\{(\mathbf{u}_h^m,\mathbf{d}_h^m,\widetilde{p}_h^m,\mathbf{w}_h^m)\big\}_{m=1}^M$ *be generated by Algorithm* 3.1*; then for all* $1\le\ell\le M$,

$$\mathcal{J}_{\varepsilon,\lambda,h}(\mathbf{u}_h^\ell,\mathbf{d}_h^\ell)+k\sum_{m=1}^{\ell}\Big(\frac{k}{2}\|\,d_t\mathbf{u}_h^m\,\|_{L^2}^2+\frac{k\lambda}{2}\|\,\nabla d_t\mathbf{d}_h^m\,\|_{L^2}^2+\nu\|\,\nabla\mathbf{u}_h^m\,\|_{L^2}^2+\lambda\gamma\|\,\mathbf{w}_h^m\,\|_{L^2}^2\Big)$$
$$\le k\sum_{m=1}^{\ell}\langle\mathbf{g}(t_m,\cdot),\mathbf{u}_h^m\rangle+\mathcal{J}_{\varepsilon,\lambda,h}(\mathbf{u}_h^0,\mathbf{d}_h^0).$$

Proof. Setting $\mathbf{v}_h=\mathbf{u}_h^m$ in (3.1), $\mathbf{a}_h=\lambda\mathbf{w}_h^m$ in (3.2), $q_h=\widetilde{p}_h^m$ in (3.3), $\mathbf{b}_h=\lambda d_t\mathbf{d}_h^m$ in (3.4) and adding the resulting equations lead to

$$\big(d_t\mathbf{u}_h^m,\mathbf{u}_h^m\big)+\nu\|\,\nabla\mathbf{u}_h^m\,\|_{L^2}^2+\big((\mathbf{u}_h^{m-1}\cdot\nabla)\mathbf{u}_h^m,\mathbf{u}_h^m\big)+\frac{1}{2}\big((\operatorname{div}\mathbf{u}_h^{m-1}),|\mathbf{u}_h^m|^2\big)$$
$$+\lambda\gamma\|\,\mathbf{w}_h^m\,\|_{L^2}^2+\lambda\big(\nabla\mathbf{d}_h^m,\nabla(d_t\mathbf{d}_h^m)\big)+\frac{\lambda}{\varepsilon^2}\big(\mathbf{f}_h^m,d_t\mathbf{d}_h^m\big)_h=\langle\mathbf{g}(t_m,\cdot),\mathbf{u}_h^m\rangle.$$

The result of the lemma then follows after applying the summation operator $k\sum_{m=1}^{\ell}$ and using the relationships

$$\big(d_t\mathbf{u}_h^m,\mathbf{u}_h^m\big)=\frac{1}{2}\left\{d_t\|\,\mathbf{u}_h^m\,\|_{L^2}^2+k\,\|\,d_t\mathbf{u}_h^m\,\|_{L^2}^2\right\},$$

$$\big((\mathbf{u}_h^{m-1}\cdot\nabla)\mathbf{u}_h^m,\mathbf{u}_h^m\big)+\frac{1}{2}\big(\operatorname{div}\mathbf{u}_h^{m-1},|\mathbf{u}_h^m|^2\big)=0\,,$$

$$\big(\nabla d_t\mathbf{u}_h^m,\nabla\mathbf{u}_h^m\big)=\frac{1}{2}\left\{d_t\|\,\nabla\mathbf{u}_h^m\,\|_{L^2}^2+k\,\|\,\nabla d_t\mathbf{u}_h^m\,\|_{L^2}^2\right\},$$

$$\begin{aligned}\big(\mathbf{f}_h^m,d_t\mathbf{d}_h^m\big)_h&=\big((|\mathbf{d}_h^m|^2-1)\mathbf{d}_h^m+kd_t\mathbf{d}_h^m,d_t\mathbf{d}_h^m\big)_h\\&\geq\frac{1}{4}\big(d_t\big||\mathbf{d}_h^m|^2-1\big|^2,1\big)_h+\frac{k}{2}\big(|d_t\mathbf{d}_h^m|^2,1\big)_h\\&\geq\frac{1}{k}\big(F(\mathbf{d}_h^m)-F(\mathbf{d}_h^{m-1}),1\big)_h\,.\qquad\square\end{aligned}$$

Solvability of Algorithm 3.1 for every $m\geq 1$ now follows from the discrete energy law and an application of the Brouwer fixed point theorem (cf. Corollary 1.1 on p. 279 of [20]).

THEOREM 3.5. *For all $k,h,\varepsilon>0$, and $m\geq 1$, there exists a solution $(\mathbf{u}_h^m,\mathbf{d}_h^m,\tilde{p}_h^m,\mathbf{w}_h^m)$ to (3.1)–(3.4).*

Remark 3.1. Due to the lack of a discrete energy law in [31, 32, 15], existence for those schemes requires $k=\mathcal{O}(e^{-\frac{1}{\varepsilon^2}})$. In [31, 32, 15], the authors follow the strategy given in [26] to construct weak solutions to (1.6)–(1.8) as proper limits of existing Faedo–Galerkin solutions, where existence of local weak solutions was verified by the Schauder fixed point argument, which was then globally extended by a continuation argument using the basic energy law. However, in order to repeat the fixed point argument as used in section 2 of [26], it requires certain smallness of the time interval, which manifests to the above restrictive mesh constraint in the fully discrete case.

Lemma 3.4 now implies the following uniform (in ε) a priori estimates for $(\mathbf{u}_h^m,\mathbf{d}_h^m,\tilde{p}_h^m,\mathbf{w}_h^m)$.

LEMMA 3.6. *Let $\mathcal{T}_h$ be a quasi-uniform triangulation of $\Omega\subset\mathbb{R}^N$ for $N=2,3$, and let $T>0$. Suppose that $\big\{(\mathbf{u}_h^m,\mathbf{d}_h^m,\widetilde{p}_h^m,\mathbf{w}_h^m)\big\}_{m=1}^N$ solves (3.1)–(3.4), where $\mathbf{g}\in L^2\big(0,T;\mathbf{V}'\big)$ and $\mathcal{J}_{\varepsilon,\lambda,h}(\mathbf{u}_h^0,\mathbf{d}_h^0)\leq C$. Then the following estimate holds:*

$$k\sum_{m=1}^{M}\Big\{\|\,d_t\mathbf{u}_h^m\,\|_{(\mathbf{V}\cap W^{2,2})'}^{\frac{4}{3}}+\|\,d_t\mathbf{d}_h^m\,\|_{L^{\frac{3}{2}}}^2+\varepsilon^4\lambda\gamma\|\,\Delta_h\mathbf{d}_h^m\,\|_{L^2}^2\Big\}\leq C\,.$$

Proof. To bound the first term, choose $\mathbf{v}_h=Q_{\mathbf{V}_h}\mathbf{v}$ with any $\mathbf{v}\in\mathbf{V}\cap W^{2,2}(\Omega,\mathbb{R}^3)$ in (3.1) and use the approximation and stability properties of $Q_{\mathbf{V}_h}$ (cf. (2.8)–(2.10)) to conclude

$$
\begin{aligned}
(3.6)\quad (d_t\mathbf{u}_h^m,\mathbf{v}) &= -\nu\big(\nabla\mathbf{u}_h^m,\nabla Q_{\mathbf{V}_h}\mathbf{v}\big)-\big((\mathbf{u}_h^{m-1}\cdot\nabla)\mathbf{u}_h^m,Q_{\mathbf{V}_h}\mathbf{v}\big)+\langle\mathbf{g}(t_m,\cdot),Q_{\mathbf{V}_h}\mathbf{v}\rangle\\
&\quad-\frac{1}{2}\big((\operatorname{div}\mathbf{u}_h^{m-1})\,\mathbf{u}_h^m,Q_{\mathbf{V}_h}\mathbf{v}\big)+\lambda\big((\nabla\mathbf{d}_h^{m-1})^T\mathbf{w}_h^m,Q_{\mathbf{V}_h}\mathbf{v}\big)\\
&\le C\nu\|\nabla\mathbf{u}_h^m\|_{L^2}\|\nabla\mathbf{v}\|_{L^2}+C\|\mathbf{u}_h^{m-1}\|_{L^3}\|\nabla\mathbf{u}_h^m\|_{L^2}\|\nabla Q_{\mathbf{V}_h}\mathbf{v}\|_{L^2}\\
&\quad+C\|\nabla\mathbf{u}_h^{m-1}\|_{L^2}\|\mathbf{u}_h^m\|_{L^3}\|\nabla\mathbf{v}\|_{L^2}+\|\mathbf{g}(t_m,\cdot)\|_{\mathbf{V}'}\|\mathbf{v}\|_{W^{1,2}}\\
&\quad+\lambda\|\mathbf{w}_h^m\|_{L^2}\|\nabla\mathbf{d}_h^{m-1}\|_{L^2}\Big[\|Q_{\mathbf{V}_h}\mathbf{v}\|_{L^\infty}+\|\mathbf{v}-Q_{\mathbf{V}_h}\mathbf{v}\|_{L^\infty}\Big].
\end{aligned}
$$

For the last term above, we use standard interpolation estimates, an inverse estimate, and error bound (2.8).

$$
\begin{aligned}
\|\mathbf{v}-Q_{\mathbf{V}_h}\mathbf{v}\|_{L^\infty}&\le\|\mathbf{v}-\boldsymbol{\mathcal{I}}_{\mathbf{X}_h}\mathbf{v}\|_{L^\infty}+\|Q_{\mathbf{V}_h}\mathbf{v}-\boldsymbol{\mathcal{I}}_{\mathbf{X}_h}\mathbf{v}\|_{L^\infty}\\
&\le Ch^{2-\frac{N}{2}}\|\mathbf{v}\|_{W^{2,2}}+Ch^{-\frac{N}{2}}\|\boldsymbol{\mathcal{I}}_{\mathbf{X}_h}\mathbf{v}-Q_{\mathbf{V}_h}\mathbf{v}\|_{L^2}\\
&\le Ch^{2-\frac{N}{2}}\|\mathbf{v}\|_{W^{2,2}}+Ch^{-\frac{N}{2}}\|\mathbf{v}-Q_{\mathbf{V}_h}\mathbf{v}\|_{L^2}.
\end{aligned}
$$

Together with Sobolev's inequality, the right-hand side of (3.6) can be controlled by

$$
\begin{aligned}
&C\nu\|\nabla\mathbf{u}_h^m\|_{L^2}\|\nabla\mathbf{v}\|_{L^2}+C\|\mathbf{u}_h^m\|_{L^3}\|\mathbf{u}_h^{m-1}\|_{L^6}\|\mathbf{v}\|_{W^{1,2}}+\|\mathbf{g}(t_m,\cdot)\|_{\mathbf{V}'}\|\mathbf{v}\|_{W^{1,2}}\\
&\quad+C\lambda\|\mathbf{w}_h^m\|_{L^2}\|\nabla\mathbf{d}_h^{m-1}\|_{L^2}\big(\|\Delta\mathbf{v}\|_{L^2}+\|\nabla\mathbf{v}\|_{L^2}\big),
\end{aligned}
$$

which yields the first assertion, thanks to interpolation of L^3 between L^2 and $W^{1,2}$.

To bound the second term, notice that (3.2) can be rewritten as

$$
d_t\mathbf{d}_h^m=-\gamma\mathbf{w}_h^m-Q_{\mathbf{Y}_h}((\mathbf{u}_h^m\cdot\nabla)\mathbf{d}_h^{m-1})\quad\text{in }\mathbf{Y}_h.
$$

The L^2-stability of $Q_{\mathbf{Y}_h}$, Lemma 3.4, and the inequality

$$
\|(\mathbf{u}_h^m\cdot\nabla)\mathbf{d}_h^{m-1}\|_{L^{\frac{3}{2}}}\le\|\mathbf{u}_h^m\|_{L^6}\|\nabla\mathbf{d}_h^{m-1}\|_{L^2}
$$

then lead to the desired estimate.

The estimate for the last term follows from (3.4) with $\mathbf{a}_h=-\Delta_h\mathbf{d}_h^m$. We use (2.2) and L^p-stability of Lagrange interpolation, together with Sobolev's inequality, to conclude

$$
\begin{aligned}
\|\Delta_h\mathbf{d}_h^m\|_{L^2}^2&\le C\|\mathbf{w}_h^m\|_{L^2}^2+\frac{C}{\varepsilon^2}\|\mathbf{f}_h^m\|_h^2\\
&\le C\|\mathbf{w}_h^m\|_{L^2}^2+\frac{C}{\varepsilon^4}\|\nabla\mathbf{d}_h^m\|_{L^2}^6+\frac{C}{\varepsilon^4}\|\mathbf{d}_h^{m-1}\|_{L^2}^2. \qquad \square
\end{aligned}
$$

Let $\mathbf{U}_\varepsilon,\mathbf{D}_\varepsilon:\Omega_T\to\mathbb{R}^N$ and $P_\varepsilon:\Omega_T\to\mathbb{R}$ denote continuous, piecewise linear interpolations in time of the fully discrete solution $\{(\mathbf{u}_h^m,\mathbf{d}_h^m,p_h^m)\}_{m=1}^M$. See section 2.2 for the definition.

We are ready to state the main result of this section.

THEOREM 3.7. *Suppose that the assumptions of Lemma* 3.6 *hold, and* $\varepsilon>0$. *Then there exists a convergent (not relabeled) subsequence of* $\{(\mathbf{U}_\varepsilon,\mathbf{D}_\varepsilon,\mathbf{W}_\varepsilon^+)\}$ *and a triple* $(\mathbf{u}^\varepsilon,\mathbf{d}^\varepsilon,\mathbf{w}^\varepsilon)$ *such that*

(i) $\displaystyle\lim_{k,h\to0}\Big(\|\mathbf{U}_\varepsilon-\mathbf{u}^\varepsilon\|_{L^2(\Omega_T)}+\|\nabla(\mathbf{D}_\varepsilon-\mathbf{d}^\varepsilon)\|_{L^2(\Omega_T)}\Big)=0,$

(ii) $\mathbf{W}_\varepsilon\rightharpoonup\mathbf{w}^\varepsilon$ *in* $L^2(\Omega_T,\mathbb{R}^N)\qquad(k,h\to0).$

Moreover, $(\mathbf{u}^\varepsilon, \mathbf{d}^\varepsilon)$ *is a weak solution of problem* (1.6)–(1.8)*, and* $\mathbf{w}^\varepsilon = -\Delta \mathbf{d}^\varepsilon + \varepsilon^{-2}\mathbf{f}(\mathbf{d}^\varepsilon)$ *holds in distributional sense.*

Proof. Step 1*: Extracting convergent subsequences.* Lemmas 3.4 and 3.6 imply the following estimates (uniform in k, h, and ε):

$$\| \mathbf{U}_\varepsilon^+ \|_{L^\infty(L^2)} + \sqrt{\lambda}\| \nabla \mathbf{D}_\varepsilon^+ \|_{L^\infty(L^2)} + \varepsilon^{-1}\sqrt{\lambda}\| \, | \mathbf{D}_\varepsilon^+ |^2 - 1 \|_{L^\infty(L^2)} \leq C \,, \tag{3.7}$$

$$\sqrt{\nu}\| \nabla \mathbf{U}_\varepsilon^+ \|_{L^2(\Omega_T)} + \sqrt{\lambda\gamma}\| \mathbf{W}_\varepsilon^+ \|_{L^2(\Omega_T)} \leq C \,, \tag{3.8}$$

$$\| \mathbf{U}_{\varepsilon,t} \|_{L^2((\mathbf{V}\cap W^{2,2})')} + \| \mathbf{D}_{\varepsilon,t} \|_{L^2(L^{\frac{3}{2}})} \leq C \,, \tag{3.9}$$

$$\varepsilon^2\sqrt{\lambda\gamma}\| \Delta_h \mathbf{D}_\varepsilon^+ \|_{L^2(L^2)} \leq C \,. \tag{3.10}$$

Then, on noting (3.3), there exist a convergent subsequence of $\{(\mathbf{U}_\varepsilon, \mathbf{D}_\varepsilon, \mathbf{W}_\varepsilon^+)\}$ (still denoted by the same notation) and functions $(\mathbf{u}^\varepsilon, \mathbf{d}^\varepsilon, \mathbf{w}^\varepsilon)$ such that

$$\mathbf{u}^\varepsilon \in L^\infty\big(0,T;\mathbf{H}\big) \cap L^2\big(0,T;\mathbf{V}\big) \cap W^{1,2}\big(0,T;[\mathbf{V}\cap W^{2,2}(\Omega,\mathbb{R}^N)]'\big) \,,$$

$$\mathbf{d}^\varepsilon \in L^\infty(0,T;W^{1,2}(\Omega,\mathbb{R}^N)\big) \cap L^4\big(0,T;W^{1,4}(\Omega,\mathbb{R}^N)\big) \cap W^{1,2}\big(0,T;L^{\frac{3}{2}}(\Omega,\mathbb{R}^N)\big) \,,$$

$$\mathbf{w}^\varepsilon \in L^2\big(\Omega_T,\mathbb{R}^N\big) \,,$$

and as $k, h \to 0$

$$\begin{aligned}
\mathbf{U}_\varepsilon, \mathbf{U}_\varepsilon^+ &\overset{*}{\rightharpoonup} \mathbf{u}^\varepsilon \quad \text{in } L^\infty\big(0,T;L^2(\Omega,\mathbb{R}^N)\big) \,,\\
\mathbf{U}_\varepsilon, \mathbf{U}_\varepsilon^+ &\rightharpoonup \mathbf{u}^\varepsilon \quad \text{in } L^2\big(0,T;W^{1,2}(\Omega,\mathbb{R}^N)\big)\\
\mathbf{U}_\varepsilon &\rightharpoonup \mathbf{U}_\varepsilon \quad \text{in } W^{1,\frac{4}{3}}\big(0,T;[\mathbf{V}\cap W^{2,2}(\Omega,\mathbb{R}^N)]'\big) \,,\\
\mathbf{U}_\varepsilon &\to \mathbf{u}^\varepsilon \quad \text{in } L^2\big(0,T;L^2(\Omega,\mathbb{R}^N)\big) \,,\\
\mathbf{D}_\varepsilon, \mathbf{D}_\varepsilon^+ &\overset{*}{\rightharpoonup} \mathbf{d}^\varepsilon \quad \text{in } L^\infty\big(0,T;W^{1,2}(\Omega,\mathbb{R}^N)\big) \,,\\
\mathbf{D}_\varepsilon, \mathbf{D}_\varepsilon^+ &\rightharpoonup \mathbf{d}^\varepsilon \quad \text{in } L^4\big(0,T;W^{1,4}(\Omega,\mathbb{R}^N)\big)\\
\mathbf{D}_\varepsilon &\rightharpoonup \mathbf{d}^\varepsilon \quad \text{in } W^{1,2}\big(0,T;L^{\frac{3}{2}}(\Omega,\mathbb{R}^N)\big) \,,\\
\mathbf{D}_\varepsilon &\to \mathbf{d}^\varepsilon \quad \text{in } L^2\big(0,T;W^{1,2}(\Omega,\mathbb{R}^N)\big) \,,\\
\mathbf{W}_\varepsilon, \mathbf{W}_\varepsilon^+ &\rightharpoonup \mathbf{w}^\varepsilon \quad \text{in } L^2\big(0,T;L^2(\Omega,\mathbb{R}^N)\big) \,.
\end{aligned} \tag{3.11}$$

Here, we have used the Aubin–Lions lemma to conclude $(3.11)_3$ from (3.3), and $(3.11)_8$ from (2.5) and Lemma 2.4.

From Lemma 3.4 we also have, as $k \to 0$,

$$\begin{aligned}
\| \mathbf{U}_\varepsilon - \mathbf{U}_\varepsilon^+ \|_{L^2(L^2)}^2 &= \sum_{m=1}^M \| \mathbf{u}_h^m - \mathbf{u}_h^{m-1} \|_{L^2}^2 \int_{t_{m-1}}^{t_m} \Big(\frac{s-t_m}{k}\Big)^2 \,\mathrm{d}s\\
&= \frac{k}{3}\sum_{m=1}^M \| \mathbf{u}_h^m - \mathbf{u}_h^{m-1} \|_{L^2}^2 \to 0 \,,
\end{aligned}$$

$$\| \nabla(\mathbf{D}_\varepsilon - \mathbf{D}_\varepsilon^+) \|_{L^2(L^2)}^2 = \frac{k}{3}\sum_{m=1}^M \| \nabla(\mathbf{d}_h^m - \mathbf{d}_h^{m-1}) \|_{L^2}^2 \to 0 \,.$$

Hence, the sequences $\{\mathbf{U}_\varepsilon\}$ and $\{\mathbf{U}_\varepsilon^+\}$ converge to the same limit as $k, h \to 0$, and so do the sequences $\{\mathbf{D}_\varepsilon\}$ and $\{\mathbf{D}_\varepsilon^+\}$. Similarly, we can show that $\{\mathbf{U}_\varepsilon\}$ and $\{\mathbf{U}_\varepsilon^-\}$ have the same limit, as do $\{\mathbf{D}_\varepsilon\}$ and $\{\mathbf{D}_\varepsilon^-\}$.

Step 2*: Passing to the limit.* We now pass to the limit in (3.1)–(3.4) and show that $(\mathbf{u}^\varepsilon, \mathbf{d}^\varepsilon)$ is a weak solution to problem (1.6)–(1.8) with the initial values $\mathbf{u}^\varepsilon(0,\cdot) = \mathbf{u}_0^\varepsilon$ and $\mathbf{d}^\varepsilon(0,\cdot) = \mathbf{d}_0^\varepsilon$. To the end, we rewrite (3.1)–(3.4) as follows: For every $t > 0$, find $\big(\mathbf{U}_\varepsilon(t,\cdot), \mathbf{D}_\varepsilon(t,\cdot)\big) \in \mathbf{V}_h \times \mathbf{Y}_h$, such that for every $(\mathbf{v}_h, \mathbf{a}_h, q_h) \in \mathbf{V}_h \times \mathbf{Y}_h \times M_h$

$$
\begin{aligned}
&(\mathbf{U}_{\varepsilon,t}, \mathbf{v}_h) + \nu\big(\nabla \mathbf{U}_\varepsilon^+, \nabla \mathbf{v}_h\big) + \big((\mathbf{U}_\varepsilon^- \cdot \nabla)\mathbf{U}_\varepsilon^+, \mathbf{v}_h\big) \\
&\qquad + \frac{1}{2}\big((\operatorname{div} \mathbf{U}_\varepsilon^-)\,\mathbf{U}_\varepsilon^+, \mathbf{v}_h\big) + \lambda\big((\nabla \mathbf{D}_\varepsilon^-)^T \mathbf{W}_\varepsilon^+, \mathbf{v}_h\big) = \langle \mathbf{g}^+, \mathbf{v}_h\rangle\,, \\
&(\mathbf{D}_{\varepsilon,t}, \mathbf{a}_h) + \big((\mathbf{U}_\varepsilon^+ \cdot \nabla)\mathbf{D}_\varepsilon^-, \mathbf{a}_h\big) + \gamma\,\big(\mathbf{W}_\varepsilon^+, \mathbf{a}_h\big) = 0\,, \\
&(\nabla \mathbf{D}_\varepsilon^+, \nabla \mathbf{a}_h) + \frac{1}{\varepsilon^2}\big(\,\overline{\mathbf{f}}_\varepsilon, \mathbf{a}_h\,\big)_h + \big(\mathbf{W}_\varepsilon^+, \mathbf{a}_h\big) = 0\,, \\
&(\operatorname{div} \mathbf{U}_\varepsilon^+, q_h) = 0\,.
\end{aligned}
$$

For any $(\mathbf{v}, \mathbf{a}) \in \boldsymbol{\mathcal{V}} \times C^\infty(\Omega, \mathbb{R}^N)$, let $(\mathbf{v}_h, \mathbf{a}_h) \equiv (Q_{\mathbf{V}_h}\mathbf{v}, R_{\mathbf{Y}_h}\mathbf{a}) \in \mathbf{V}_h \times \mathbf{Y}_h$, which satisfy $(\mathbf{v}_h, \mathbf{a}_h) \to (\mathbf{v}, \mathbf{a})$ in $\big[W^{1,p}(\Omega, \mathbb{R}^N)\big]^2$ $(h \to 0)$ for all $1 \le p \le \frac{2N}{N-2}$ (cf. (2.8)). For $k, h \to 0$, we use (3.11) to obtain

$$
\langle \mathbf{u}_t^\varepsilon, \mathbf{v}\rangle + \nu\big(\nabla \mathbf{u}^\varepsilon, \nabla \mathbf{v}\big) + \big((\mathbf{u}^\varepsilon \cdot \nabla)\mathbf{u}^\varepsilon, \mathbf{v}\big) + \lambda\big((\nabla \mathbf{d}^\varepsilon)^T \mathbf{w}^\varepsilon, \mathbf{v}\big) = \langle \mathbf{g}, \mathbf{v}\rangle\,, \tag{3.12}
$$

$$
\big(\mathbf{d}_t^\varepsilon, \mathbf{a}\big) + \big((\mathbf{u}^\varepsilon \cdot \nabla)\mathbf{d}^\varepsilon, \mathbf{a}\big) - \gamma\,\big(\mathbf{w}^\varepsilon, \mathbf{a}\big) = 0\,, \tag{3.13}
$$

$$
\big(\operatorname{div} \mathbf{u}^\varepsilon, q\big) = 0, \tag{3.14}
$$

$$
\big(\nabla \mathbf{d}^\varepsilon, \nabla \mathbf{a}\big) + \frac{1}{\varepsilon^2}\big(\mathbf{f}(\mathbf{d}^\varepsilon), \mathbf{a}\big) + \big(\mathbf{w}^\varepsilon, \mathbf{a}\big) = 0\,, \tag{3.15}
$$

and $\mathbf{u}^\varepsilon(0,\cdot) = \mathbf{u}_0^\varepsilon$, $\mathbf{d}^\varepsilon(0,\cdot) = \mathbf{d}_0^\varepsilon$. Here, we have used $\big(\operatorname{div} \mathbf{u}^\varepsilon, \langle \mathbf{u}^\varepsilon, \mathbf{v}\rangle\big) = 0$, which is a consequence of $\mathbf{u}^\varepsilon(t,\cdot) \in \mathbf{V}$ for almost every $t > 0$.

Equation (3.15) immediately implies that $\mathbf{w}^\varepsilon = -\Delta \mathbf{d}^\varepsilon + \frac{1}{\varepsilon^2}\mathbf{f}(\mathbf{d}^\varepsilon)$ in Ω_T, and $\partial_{\mathbf{n}} \mathbf{d}^\varepsilon = 0$ on $\partial\Omega_T$ in the sense of distribution. Hence, $(\mathbf{u}^\varepsilon, \mathbf{d}^\varepsilon)$ is a weak solution of (1.4)–(1.8). □

Remark 3.2. (a) In the case of unique weak solutions (valid for $N = 2$), the proof also verifies that the limit of *every* convergent subsequence of $\{(\mathbf{U}_\varepsilon, \mathbf{D}_\varepsilon, \mathbf{W}_\varepsilon)\}$ must be a weak solution of (1.6)–(1.8). Hence, the whole sequence $\{(\mathbf{U}_\varepsilon, \mathbf{D}_\varepsilon)\}$ converges to the unique weak solution $(\mathbf{u}_*^\varepsilon, \mathbf{d}_*^\varepsilon)$.

(b) The proof does not provide information about convergence of the pressure $\widetilde{P}_\varepsilon$ as $(k, h) \mapsto 0$, which is in accordance with the general theory for incompressible Navier–Stokes equations (cf. p. 83 of [30]). However, its convergence in L^2 could be proved under additional regularity assumptions on the weak solution $(\mathbf{u}_*^\varepsilon, \mathbf{d}_*^\varepsilon)$.

An immediate consequence of Theorem 3.7 and the convergence result of [28] is the following convergence theorem.

THEOREM 3.8. *Under the assumptions of Theorem* 3.7, *there exist a subsequence of* $\{(\mathbf{U}_\varepsilon, \mathbf{D}_\varepsilon)\}$ *and a triple* $(\mathbf{u}_*, \mathbf{d}_*, \nu_*) \in \big[L^\infty\big(0,T;L^2(\Omega,\mathbb{R}^N)\big) \cap L^2(0,T;\mathbf{V})\big] \times \big[L^\infty\big(0,T;W^{1,2}(\Omega,\mathbb{R}^N)\big) \cap L^2\big(0,T;W^{2,2}(\Omega,\mathbb{R}^N)\big)\big]$ *such that*

$$\lim_{\varepsilon\to 0} \lim_{k,h\to 0} \Big(\| \mathbf{U}_\varepsilon - \mathbf{u}_* \|_{L^2(L^2)} + \| \mathbf{D}_\varepsilon - \mathbf{d}_* \|_{L^2(L^2)} \Big) = 0\,, \tag{3.16}$$

$$\delta_{\nabla \mathbf{D}_\varepsilon} \overset{*}{\rightharpoonup} \nu_* \quad \text{in } L^\infty_w\big(\Omega_T, \mathcal{M}(\mathbb{R}^{N\times N})\big) \quad (\text{first } k,h\to 0, \text{ then } \varepsilon \to 0)\,, \tag{3.17}$$

where $(\mathbf{u}_*, \mathbf{d}_*, \nu_*)$ *is a generalized solution to* (1.1)–(1.5).

The polynomial bound with respect to ε^{-1} in (3.10) suggests that a polynomial scaling law $F(\varepsilon,k,h) \geq 0$ would be sufficient to ensure convergence of the numerical solutions to weak solutions of (1.4)–(1.8) and to generalized solutions of (1.1)–(1.5). This observation is in contrast to convergence results for the schemes proposed in [31, 32, 15], where optimal rates of convergence involve constants that grow exponentially with respect to ε^{-1} (cf. Lemma 3.3 for a priori estimates of strong solutions). We shall investigate this issue further in section 5 and now turn to studying a fully discrete scheme which approximates the original problem (1.1)–(1.5) directly.

4. A fully discrete scheme for (1.1)–(1.5): Conditional solvability and discrete energy law. In this section, we propose and analyze a fully discrete scheme for (1.1)–(1.5) which uses a stable tuple $(\mathbf{X}_h, M_h)$ of finite element spaces for velocity and pressure fields. It also crucially utilizes the discrete Laplacian $\Delta_h : W^{1,2}(\Omega,\mathbb{R}^N) \to \mathbf{Y}_h$ and the midpoint rule for temporal discretization.

ALGORITHM 4.1.

(1) *Let* $\mathbf{u}_h^0 = Q_{\mathbf{V}_h}\mathbf{u}^0$ *and* $\mathbf{d}_h^0 = \boldsymbol{\mathcal{I}}_{\mathbf{Y}_h}\mathbf{d}^0$.

(2) *Let* $m = 1,\ldots,M$. *Find* $(\mathbf{u}_h^m, \mathbf{d}_h^m) \in \mathbf{X}_h \times \mathbf{Y}_h$ *such that for all* $(\mathbf{v}_h, \mathbf{a}_h, q_h) \in \mathbf{X}_h \times \mathbf{Y}_h \times M_h$ *there holds*

$$\begin{aligned}(d_t\mathbf{u}_h^m, \mathbf{v}_h) + \nu(\nabla\mathbf{u}_h^m, \nabla\mathbf{v}_h) + \big((\mathbf{u}_h^{m-1}\cdot\nabla)\mathbf{u}_h^m, \mathbf{v}_h\big) + \frac{1}{2}\big((\operatorname{div}\mathbf{u}_h^{m-1})\mathbf{u}_h^m, \mathbf{v}_h\big) \\ - (\hat{p}_h^m, \operatorname{div}\mathbf{v}_h) - \lambda\big((\nabla\mathbf{d}_h^{m-1})^T\Delta_h\mathbf{d}^{m-1/2}, \mathbf{v}_h\big) = \big\langle \mathbf{g}(t_m,\cdot), \mathbf{v}_h\big\rangle\,,\end{aligned} \tag{4.1}$$

$$\begin{aligned}(d_t\mathbf{d}_h^m, \mathbf{a}_h) + \big((\mathbf{u}_h^m\cdot\nabla)\mathbf{d}_h^{m-1}, \mathbf{a}_h\big) \\ + \gamma\big(\mathbf{d}_h^{m-1/2}\times(\mathbf{d}_h^{m-1/2}\times\Delta_h\mathbf{d}_h^{m-1/2}), \mathbf{a}_h\big) = 0\,,\end{aligned} \tag{4.2}$$

$$(\operatorname{div}\mathbf{u}_h^m, q_h) = 0\,. \tag{4.3}$$

Note that an implicit treatment of coupling terms is used in (4.1)–(4.2), in contrast to the semi-implicit discretization in Algorithm 3.1. The following lemma verifies a discrete energy law for solutions of Algorithm 4.1.

LEMMA 4.1. *Let* $(\mathbf{u}_h^0, \mathbf{d}_h^0) \in \mathbf{X}_h \times \mathbf{Y}_h$, *such that* $\mathcal{J}(\mathbf{u}_h^0, \mathbf{d}_h^0) \leq C$. *Then for all* $1 \leq \ell \leq M$,

$$\begin{aligned}\mathcal{J}(\mathbf{u}_h^\ell, \mathbf{d}_h^\ell) + k\sum_{m=1}^{\ell}\Big[\frac{k}{2}\|\,d_t\mathbf{u}_h^m\,\|_{L^2}^2 + \nu\,\|\,\nabla\mathbf{u}_h^m\,\|_{L^2}^2 + \|\,\mathbf{d}_h^{m-1/2}\times\Delta_h\mathbf{d}_h^{m-1/2}\,\|_{L^2}^2\Big] \\ = k\sum_{m=1}^{\ell}\langle\mathbf{g}^m, \mathbf{u}_h^m\rangle + \mathcal{J}(\mathbf{u}_h^0, \mathbf{d}_h^0)\,.\end{aligned}$$

Proof. Setting $(\mathbf{v}_h, \mathbf{a}_h, \chi_h) = (\mathbf{u}_h^m, -\lambda\Delta_h \mathbf{d}_h^{m-1/2}, \hat{p}_h^m)$ in (4.1)–(4.3) and adding the resulting equations yield

$$\frac{1}{2} d_t \big[\| \mathbf{u}_h^m \|_{L^2}^2 + \lambda \| \nabla \mathbf{d}_h^m \|_{L^2}^2 \big] + \frac{k}{2} \| d_t \mathbf{u}_h^m \|_{L^2}^2 + \nu \| \nabla \mathbf{u}_h^m \|_{L^2}^2$$

$$+ \lambda\gamma \| \mathbf{d}_h^{m-1/2} \times \Delta_h \mathbf{d}^{m-1/2} \|_{L^2}^2 + \Big[\big((\mathbf{u}_h^{m-1} \cdot \nabla) \mathbf{u}_h^m, \mathbf{u}_h^m \big) + \frac{1}{2} \big((\operatorname{div} \mathbf{u}_h^{m-1}, | \mathbf{u}_h^m |^2 \big) \Big]$$

$$+ \lambda \Big[\big((\nabla \mathbf{d}_h^{m-1})^T \Delta_h \mathbf{d}_h^{m-1/2}, \mathbf{u}_h^m \big) + \big((\mathbf{u}_h^m \cdot \nabla) \mathbf{d}_h^{m-1}, -\Delta_h \mathbf{d}^{m-1/2} \big) \big) \Big] = \langle \mathbf{g}^m, \mathbf{u}_h^m \rangle .$$

Since the two pairs of terms in the brackets on the left-hand side of the equation vanish, hence, the assertion follows by applying the summation operator $k \sum_{m=1}^{\ell}$. ∎

We now verify existence of sequences $\big\{ (\mathbf{u}_h^m, \mathbf{d}_h^m) \big\}_{m=1}^M \subset \mathbf{X}_h \times \mathbf{Y}_h$ in the case of restricted mesh parameters. For each $m \geq 0$, consider the continuous mapping $\boldsymbol{\mathcal{F}} : \mathbf{V}_h \times \mathbf{Y}_h \to \mathbf{V}_h \times \mathbf{Y}_h$ defined as follows: Given $(\mathbf{u}_h^{m-1}, \mathbf{d}_h^{m-1}) \in \mathbf{V}_h \times \mathbf{Y}_h$, find $(\mathbf{U}, \mathbf{D}) \in \mathbf{V}_h \times \mathbf{Y}_h$ such that

$$\big(\boldsymbol{\mathcal{F}}[\mathbf{U}, \mathbf{D}], [\mathbf{v}_h, \mathbf{a}_h] \big) := \big((\boldsymbol{\mathcal{F}}_1[\mathbf{U}, \mathbf{D}], \mathbf{v}_h), (\boldsymbol{\mathcal{F}}_2[\mathbf{U}, \mathbf{D}], \mathbf{a}_h) \big) = \mathbf{0} \tag{4.4}$$

for any $(\mathbf{v}_h, \mathbf{a}_h) \in \mathbf{V}_h \times \mathbf{Y}_h$. Where

$$\big(\boldsymbol{\mathcal{F}}_1[\mathbf{U}, \mathbf{D}], \mathbf{v}_h \big) := \frac{1}{k} \big(\mathbf{U} - \mathbf{u}_h^{m-1}, \mathbf{v}_h \big) + \nu (\nabla \mathbf{U}, \nabla \mathbf{v}_h) + \big((\mathbf{u}_h^{m-1} \cdot \nabla) \mathbf{U}, \mathbf{v}_h \big)$$

$$+ \frac{1}{2} \big((\operatorname{div} \mathbf{u}_h^{m-1}) \mathbf{U}, \mathbf{v}_h \big) - \lambda \big((\nabla \mathbf{d}_h^{m-1})^T \Delta_h \mathbf{D}, \mathbf{v}_h \big) - \langle \mathbf{g}(t_m, \cdot), \mathbf{v}_h \rangle ,$$

$$\big(\boldsymbol{\mathcal{F}}_2[\mathbf{U}, \mathbf{D}], \mathbf{a}_h \big) := \frac{2}{k} \big(\mathbf{D} - \mathbf{d}_h^{m-1}, \mathbf{a}_h \big) + \big((\mathbf{U} \cdot \nabla) \mathbf{d}_h^{m-1}, \mathbf{a}_h \big) + \gamma \big(\mathbf{D} \times (\mathbf{D} \times \Delta_h \mathbf{D}), \mathbf{a}_h \big) .$$

Suppose $\mathcal{T}_h$ is a quasi-uniform triangulation of a bounded polyhedral domain $\Omega \subset \mathbb{R}^N$ for $N = 2, 3$. In what follows, consider tuple $\big(\mathbf{U}, \mathbf{D} \big) \in \mathbf{V}_h \times \mathbf{Y}_h$ which satisfies for existing nonnegative constants $\overline{C} \equiv \overline{C} \big(\| \mathbf{u}_h^{m-1} \|_{L^2}, \| \mathbf{d}_h^{m-1} \|_{W^{1,2}}, \| \mathbf{g}(t_m, \cdot) \|_{\mathbf{V}'} \big)$ and $\hat{C}$

$$\min \big\{ \| \mathbf{U} \|_{L^2}, \| \mathbf{D} \|_{L^2} \big\} \geq \overline{C} \quad \text{and} \quad k \leq \hat{C} h^{2 + \frac{N}{2}} .$$

First, on noting $\langle \mathbf{g}(t_m, \cdot), \mathbf{U} \rangle \leq \| \mathbf{U} \|_{W^{1,2}} \| \mathbf{g}(t_m, \cdot) \|_{W^{-1,2}}$, and by the inverse inequality, we observe that for some $\tilde{C} = \tilde{C}(\Omega) > 0$

$$\begin{aligned} \big| \big((\nabla \mathbf{d}_h^{m-1})^T \Delta_h \mathbf{D}, \mathbf{U} \big) \big| &\leq \| \nabla \mathbf{d}^{m-1} \|_{L^2} \| \Delta_h \mathbf{D} \|_{L^2} \| \mathbf{U} \|_{L^\infty} \\ &\leq \tilde{C} h^{-(2 + \frac{N}{2})} \| \nabla \mathbf{d}^{m-1} \|_{L^2} \| \mathbf{D} \|_{L^2} \| \mathbf{U} \|_{L^2} , \\ \big| \big((\mathbf{U} \cdot \nabla) \mathbf{d}_h^{m-1}, \mathbf{D} \big) \big| &\leq \| \mathbf{U} \|_{L^\infty} \| \nabla \mathbf{d}_h^{m-1} \|_{L^2} \| \mathbf{D} \|_{L^2} \\ &\leq \tilde{C} h^{-\frac{N}{2}} \| \mathbf{U} \|_{L^2} \| \nabla \mathbf{d}_h^{m-1} \|_{L^2} \| \mathbf{D} \|_{L^2} . \end{aligned}$$

By Lemma 4.1 we conclude that for $k = \mathcal{O}(h^{2+\frac{N}{2}})$ the following holds:

$$\begin{aligned}
\big(\boldsymbol{\mathcal{F}}[\mathbf{U},\mathbf{D}],[\mathbf{U},\mathbf{D}]\big) &= \frac{1}{k}\Big[\|\mathbf{U}\|_{L^2}^2 - (\mathbf{u}_h^{m-1},\mathbf{U}) + 2\|\mathbf{D}\|_{L^2}^2 - 2(\mathbf{d}_h^{m-1},\mathbf{D})\Big] \\
&\quad + \nu\|\nabla\mathbf{U}\|_{L^2}^2 - \lambda\big((\nabla\mathbf{d}_h^{m-1})^T\Delta_h\mathbf{D},\mathbf{U}\big) - \langle\mathbf{g}(t_m,\cdot),\mathbf{U}\rangle \\
&\quad + \big((\mathbf{U}\cdot\nabla)\mathbf{d}_h^{m-1},\mathbf{D}\big) \\
&\geq \frac{1}{k}\Big[\|\mathbf{U}\|_{L^2}\Big(\Big[1-\frac{\tilde{C}k}{h^{2+\frac{N}{2}}}\|\nabla\mathbf{d}_h^{m-1}\|_{L^2}\Big]\|\mathbf{U}\|_{L^2} - \|\mathbf{u}_h^{m-1}\|_{L^2}\Big) \\
&\quad + 2\|\mathbf{D}\|_{L^2}\Big(\Big[1-\frac{\tilde{C}k}{h^{2+\frac{N}{2}}}\|\nabla\mathbf{d}_h^{m-1}\|_{L^2}\Big]\|\mathbf{D}\|_{L^2} - \|\mathbf{d}_h^{m-1}\|_{L^2}\Big)\Big] \\
&\quad + \frac{\nu}{2}\|\nabla\mathbf{U}\|_{L^2}^2 - \frac{2}{\nu}\|\mathbf{g}(t_m,\cdot)\|_{\mathbf{V}'}^2 \geq 0\,.
\end{aligned}$$

Brouwer's fixed point theorem then implies the existence of $(\mathbf{U}^*,\mathbf{D}^*) \in \mathbf{V}_h \times \mathbf{Y}_h$, which solves (4.4) (cf. Corollary 1.1 on p. 279 of [20]), and $(\mathbf{u}_h^m,\mathbf{d}_h^m) := (\mathbf{U}^*, 2\mathbf{D}^* - \mathbf{d}_h^{m-1}) \in \mathbf{V}_h \times \mathbf{Y}_h$ is a solution to Algorithm 4.1. Hence, we have shown the following result.

THEOREM 4.2. *Let $\mathcal{T}_h$ be a quasi-uniform triangulation of a bounded polyhedral domain $\Omega \subset \mathbb{R}^N$ for $N = 2, 3$. There exists $\hat{C} > 0$ such that for all $k \leq \hat{C}h^{2+\frac{N}{2}}$ there exists a solution $(\mathbf{u}_h^m, \mathbf{d}_h^m, \hat{p}_h^m)$ to (4.1)–(4.3).*

The following result provides an estimate for the first order discrete time derivative of $\{\mathbf{d}_h^m\}$.

LEMMA 4.3. *Suppose $(\mathbf{u}_h^0,\mathbf{d}_h^0) \in \mathbf{V}_h \times \mathbf{Y}_h$ and $\mathcal{J}(\mathbf{u}_h^0,\mathbf{d}_h^0) \leq C$. There holds*

$$k\sum_{m=1}^{M}\|d_t\mathbf{d}_h^m\|_{L^p}^2 \leq C \quad \Big(1 \leq p < \frac{N}{N-1}\Big)\,.$$

In the following, we use the L^2-projection $Q_{\mathbf{Y}_h} : L^2(\Omega,\mathbb{R}^N) \to \mathbf{Y}_h$, where for any $\mathbf{b} \in L^2(\Omega,\mathbb{R}^N)$ there holds $(Q_{\mathbf{Y}_h}\mathbf{b} - \mathbf{b}, \mathbf{a}_h) = 0$ for every $\mathbf{a}_h \in \mathbf{Y}_h$.

Proof. We use the L^2-projection $Q_{\mathbf{Y}_h}$ to write $\|d_t\mathbf{d}_h^m\|_{L^p} = \sup_{\boldsymbol{\xi}\in L^{p'}} \frac{(d_t\mathbf{d}_h^m, Q_{\mathbf{Y}_h}\boldsymbol{\xi})}{\|\boldsymbol{\xi}\|_{L^{p'}}}$. On using (4.2), with $\mathbf{a}_h := Q_{\mathbf{Y}_h}\boldsymbol{\xi}$, we find upper bounds ($r > 2$)

$$\begin{aligned}
&\big|\big(\mathbf{d}_h^{m-1/2}\times(\mathbf{d}_h^{m-1/2}\times\Delta_h\mathbf{d}_h^{m-1/2}),\mathbf{a}_h\big)\big| \\
&\qquad\qquad \leq \|\mathbf{d}_h^{m-1/2}\|_{L^r}\|\mathbf{d}_h^{m-1/2}\times\Delta_h\mathbf{d}_h^{m-1/2}\|_{L^2}\|\mathbf{a}_h\|_{L^{\frac{2r}{r-2}}}\,, \\
&\big|\big((\mathbf{u}_h^m\cdot\nabla)\mathbf{d}_h^{m-1},\mathbf{a}_h\big)\big| \leq \|\mathbf{u}_h^m\|_{L^r}\|\nabla\mathbf{d}_h^{m-1}\|_{L^2}\|\mathbf{a}_h\|_{L^{\frac{2r}{r-2}}}\,.
\end{aligned}$$

Then we employ L^q-stability of $Q_{\mathbf{Y}_h}$ (cf. [13]) and Lemma 4.1, which verifies that sequences $\{\mathbf{d}_h^{m-1/2}\} \in \ell^\infty\big(\{t_\ell\}_{\ell=1}^M; L^r(\Omega,\mathbb{R}^N)\big)$ and $\{\mathbf{u}_h^m\} \in \ell^2\big(\{t_\ell\}_{\ell=1}^M; L^r(\Omega,\mathbb{R}^N)\big)$ are bounded uniformly in k,h for all $r < \frac{2N}{N-2}$, which gives the assertion. ∎

Using the notation of section 2.3, by Lemma 4.1, Theorem 4.2, and Lemma 4.3 there exist a convergent subsequence of $\{(\mathbf{U},\mathbf{D})\}$ and a tuple $(\mathbf{u},\mathbf{d})$ such that

$$\begin{aligned}
&\mathbf{u} \in L^\infty(0,T;\mathbf{H}) \cap L^2(0,T;\mathbf{V}), \\
&\mathbf{d} \in L^\infty\big(0,T;W^{1,2}(\Omega,\mathbb{R}^N)\big) \cap W^{1,2}\big(0,T;L^p(\Omega,\mathbb{R}^N)\big)\,,
\end{aligned}$$

and for $k,h \to 0$ and $1 \leq p < \frac{N}{N-1}$

$$\mathbf{U}, \mathbf{U}^+ \overset{*}{\rightharpoonup} \mathbf{u} \quad \text{in } L^\infty\big(0,T;L^2(\Omega,\mathbb{R}^N)\big)\,, \tag{4.5}$$

$$\rightharpoonup \mathbf{u} \quad \text{in } L^2\big(0,T;W^{1,2}(\Omega,\mathbb{R}^N)\big)\,,$$

$$\mathbf{D}, \mathbf{D}^+ \overset{*}{\rightharpoonup} \mathbf{d} \quad \text{in } L^\infty\big(0,T;W^{1,2}(\Omega,\mathbb{R}^N)\big)\,, \tag{4.6}$$

$$\mathbf{D} \rightharpoonup \mathbf{d} \in W^{1,2}\big(0,T;L^p(\Omega,\mathbb{R}^N)\big)\,,$$

$$\to \mathbf{d} \in L^p\big(0,T;L^2(\Omega,\mathbb{R}^N)\big) \qquad (1 \le p < \infty)\,.$$

Note that due to the viscous stress term no estimate for $\mathbf{u}_t$ is obtained. Moreover, by the weak lower semicontinuity property of norms there holds for almost all $t \in (0,T)$

$$\mathcal{J}\big(\mathbf{u}(t,\cdot),\mathbf{d}(t,\cdot)\big) + \nu \int_0^t \|\nabla \mathbf{u}(s,\cdot)\|_{L^2}^2 \,\mathrm{d}s \le \mathcal{J}\big(\mathbf{u}_0,\mathbf{d}_0\big) + \int_0^t \langle \mathbf{g}(s,\cdot), \mathbf{u}(s,\cdot)\rangle \,\mathrm{d}s\,. \tag{4.7}$$

The next result establishes $|\mathbf{d}| = 1$ a.e. in Ω_T for $N = 2$.

THEOREM 4.4. *Let $T > 0$, $\Omega \subset \mathbb{R}^2$ be a bounded polygonal domain, and let $\mathcal{T}_h$ be a quasi-uniform triangulation of Ω. Suppose $(\mathbf{u}_h^0, \mathbf{d}_h^0) \in \mathbf{V}_h \times \mathbf{Y}_h$ satisfies $\mathcal{J}(\mathbf{u}_h^0, \mathbf{d}_h^0) \le C$ and $|\mathbf{d}_h^0(\mathbf{z})| = 1$ for all $\mathbf{z} \in \mathcal{N}_h$. Assume $k \le \hat{C}h^3$. Then the above limiting director field $\mathbf{d}$ satisfies $|\mathbf{d}| = 1$ a.e. in Ω_T.*

Proof. Choose $\mathbf{a}_h = \boldsymbol{\mathcal{I}}_{\mathbf{Y}_h}\big[\frac{1}{2}\big(|\,\mathbf{d}_h^m\,|^2 + |\,\mathbf{d}_h^{m-1}\,|^2 - 2\big)\mathbf{d}_h^{m-1/2}\big]$ in (4.2). Rewrite

$$\frac{1}{2}\Big(d_t\mathbf{d}_h^m, \big[\boldsymbol{\mathcal{I}}_{\mathbf{Y}_h} \pm \mathbf{Id}\big]\big[\big(|\,\mathbf{d}_h^m\,|^2 + |\,\mathbf{d}_h^{m-1}\,|^2 - 2\big)\mathbf{d}_h^{m-1/2}\big]\Big) = A_1 + A_2\,, \tag{4.8}$$

where

$$A_1 = \frac{1}{4}\Big(d_t\big[|\,\mathbf{d}_h^m\,|^2 - 1\big], |\,\mathbf{d}_h^m\,|^2 + |\,\mathbf{d}_h^{m-1}\,|^2 - 2\Big) = \frac{1}{4}d_t\|\,|\,\mathbf{d}_h^m\,|^2 - 1\,\|_{L^2}^2\,,$$

$$\begin{aligned}
|\,A_2\,| &\le \|\,d_t\mathbf{d}_h^m\,\|_{L^p}\|\,\big[\mathbf{Id} - \boldsymbol{\mathcal{I}}_{\mathbf{Y}_h}\big]\big[\big(|\,\mathbf{d}_h^m\,|^2 + |\,\mathbf{d}_h^{m-1}\,|^2 - 2\big)\mathbf{d}_h^{m-1/2}\big]\,\|_{L^{p'}} \\
&\le Ch\,\|\,d_t\mathbf{d}_h^m\,\|_{L^p} \max_{\eta_i \in \{0,\frac{1}{2},1\}} \Big[\int_\Omega |\,\mathbf{d}_h^{m-\eta_1}\,|^{2p'}\,|\,\nabla \mathbf{d}_h^{m-\eta_2}\,|^{p'}\,\mathrm{d}\mathbf{x}\Big]^{1/p'} \\
&\le Ch\,\|\,d_t\mathbf{d}_h^m\,\|_{L^p} \max_{\eta_i \in \{0,\frac{1}{2},1\}} \big\{\|\,\mathbf{d}_h^{m-\eta_1}\,\|_{L^\infty}^2\,\|\,\nabla \mathbf{d}_h^{m-\eta_2}\,\|_{L^{p'}}\big\}
\end{aligned}$$

for $1 \le p < \infty$. When $N = 2$, it follows from Sobolev's inequality, the inverse inequality, and Lemma 4.1 that

$$k\sum_{m=1}^{M} |\,A_2\,| = o(1) \qquad \text{as } h \to 0.$$

The second term in (4.2) can be bounded as follows:

$$\frac{1}{2}\Big((\mathbf{u}_h^m \cdot \nabla)\mathbf{d}_h^{m-1}, \big[\boldsymbol{\mathcal{I}}_{\mathbf{Y}_h} \pm \mathbf{Id}\big]\big[\big(|\,\mathbf{d}_h^m\,|^2 + |\,\mathbf{d}_h^{m-1}\,|^2 - 2\big)\mathbf{d}_h^{m-1/2}\big]\Big) = B_1 + B_2\,, \tag{4.9}$$

where

(4.10)

$$
\begin{aligned}
&k\sum_{m=1}^{M} |\, B_2 \,| \\
&\quad \le k\sum_{m\ge 1}^{M} \| \mathbf{u}_h^m \|_{L^\infty} \| \nabla \mathbf{d}_h^{m-1} \|_{L^2} \| \big[\mathbf{Id} - \boldsymbol{\mathcal{I}}_{\mathbf{Y}_h}\big]\big[\big(|\, \mathbf{d}_h^m \,|^2 + |\, \mathbf{d}_h^{m-1} \,|^2 - 2\big)\mathbf{d}_h^{m-1/2}\big] \|_{L^2} \\
&\quad \le Ch \max_{\eta_i\in\{0,\frac12,1\}} \left[k\sum_{m=1}^{M} \| \mathbf{u}_h^m \|_{L^\infty} \| \mathbf{d}_h^{m-\eta_1} \|_{L^\infty}^2 \| \nabla \mathbf{d}_h^{m-\eta_2} \|_{L^2}^2 \right] \to 0
\end{aligned}
$$

as $h \to 0$. In order to bound $B_1 = \overline{B}_1 + \tilde{B}_1$, we substitute $\mathbf{d}_h^{m-1}$ by $\mathbf{d}_h^{m-1/2} - \frac{k}{2} d_t \mathbf{d}_h^m$ in the first argument in (4.9) and use the inverse inequality and Lemmas 4.1 and 4.3 to get

$$
\overline{B}_1 := k \Big| \Big((\mathbf{u}_h^m \cdot \nabla) d_t \mathbf{d}_h^m, \big(|\, \mathbf{d}_h^m \,|^2 + |\, \mathbf{d}_h^{m-1} \,|^2 - 2\big)\mathbf{d}_h^{m-1/2} \Big) \Big| \to 0
$$

for admissible $k \le \hat{C} h^3$ as $k, h \to 0$. To estimate $\tilde{B}_1$, we first employ the parallelogram identity to rewrite

$$
\begin{aligned}
\tilde{B}_1 &\equiv \frac12 \Big((\mathbf{u}_h^m \cdot \nabla)\mathbf{d}_h^{m-1/2}, \big(|\, \mathbf{d}_h^m \,|^2 + |\, \mathbf{d}_h^{m-1} \,|^2 - 2\big)\mathbf{d}_h^{m-1/2} \Big) \\
&= \frac12 \Big((\mathbf{u}_h^m \cdot \nabla)\big[|\, \mathbf{d}_h^{m-1/2} \,|^2 - 1\big], 2\big[|\, \mathbf{d}_h^{m-1/2} \,|^2 - 1\big] + \frac{k^2}{2} |\, d_t \mathbf{d}_h^m \,|^2 \Big) \\
&= -\Big(\operatorname{div} \mathbf{u}_h^m, \big[|\, \mathbf{d}_h^{m-1/2} \,|^2 - 1\big]^2 \Big) + \frac{k^2}{4} \Big((\mathbf{u}_h^m \cdot \nabla)\big[|\, \mathbf{d}_h^{m-1/2} \,|^2 - 1\big], |\, d_t \mathbf{d}_h^m \,|^2 \Big) \\
&\equiv B_{11} + B_{12} \,.
\end{aligned}
$$

We now use (4.3) to bound B_{11} by

$$
\begin{aligned}
&\Big| \Big(\operatorname{div} \mathbf{u}_h^m, \big[|\mathbf{d}^{m-1/2}|^2 - 1\big]^2 \Big) - P_{M_h}\big[|\mathbf{d}^{m-1/2}|^2 - 1\big]^2 \Big) \Big| \\
&\qquad\qquad \le Ch \, \|\operatorname{div} \mathbf{u}_h^m\|_{L^2} \|\mathbf{d}^{m-1/2}\|_{L^\infty}^2 \|\nabla \mathbf{d}_h^{m-1/2}\|_{L^2} \,.
\end{aligned}
$$

Hence, by Lemma 4.3 we conclude

$$
k\sum_{m=1}^{M} |\, B_{11} \,| = o(1)
$$

for admissible $k, h \to 0$.

To show

$$
k\sum_{m\ge 1} |\, B_{12} \,| \to 0
$$

for admissible $k, h \to 0$, let $\eta_1, \eta_2 \in \{0, \frac12, 1\}$. By the inverse inequality,

$$
\begin{aligned}
|\, B_{12} \,| &\le Ck^2 \| \mathbf{u}_h^m \|_{L^\infty} \| \mathbf{d}_h^{m-\eta_1} \nabla \mathbf{d}^{m-\eta_2} \|_{L^2} \| d_t \mathbf{d}_h^m \|_{L^4}^2 \\
&\le Ck \| \mathbf{u}_h^m \|_{L^\infty} \| \mathbf{d}_h^{m-\eta_1} \|_{L^\infty} \| \nabla \mathbf{d}_h^{m-\eta_2} \|_{L^2} \| d_t \mathbf{d}_h^m \|_{L^2} \sum_{\eta=0}^{1} \| \mathbf{d}_h^{m-\eta} \|_{W^{1,2}} .
\end{aligned}
$$

Convergence to zero for admissible $k, h \to 0$ now follows easily.

For the third term in (4.2), we observe that testing the equation with $\mathbf{a}_h = \big[\boldsymbol{\mathcal{I}}_{\mathbf{Y}_h} \pm \mathbf{Id}\big]\big[\big(|\mathbf{d}_h^m|^2 + |\mathbf{d}_h^{m-1}|^2 - 2\big)\mathbf{d}_h^{m-1/2}\big]$ gives only one nonzero term, which can be bounded as follows, on using Lemma 4.1 and Sobolev's inequality ($N = 2$):

$$\begin{aligned}
&\left| k \sum_{m=1}^{M} \Big(\mathbf{d}_h^{m-1/2} \times (\mathbf{d}_h^{m-1/2} \times \Delta_h \mathbf{d}_h^{m-1/2}), \right. \\
&\qquad \left. \big[\boldsymbol{\mathcal{I}}_{\mathbf{Y}_h} - \mathbf{Id}\big]\big[\big(|\mathbf{d}_h^m|^2 + |\mathbf{d}_h^{m-1}|^2 - 2\big)\mathbf{d}_h^{m-1/2}\big] \Big) \right| \\
&\quad \le Chk \sum_{m=1}^{M} \|\mathbf{d}_h^{m-1/2}\|_{L^\infty} \|\mathbf{d}_h^{m-1/2} \times \Delta_h \mathbf{d}_h^{m-1/2}\|_{L^2} \\
&\qquad \Big\| \nabla \Big[|\mathbf{d}_h^m|^2 + |\mathbf{d}_h^{m-1}|^2 - 2\big)\mathbf{d}_h^{m-1/2} \Big] \Big\|_{L^2} \\
&\quad \to 0 \qquad \text{as } k, h \to 0.
\end{aligned}$$

Finally, putting all the pieces together yields for all $M > 0$

$$\| \, |\, \mathbf{d}_h^m \,|^2 - 1 \, \|_{L^2}^2 \le \| \, |\mathbf{d}_h^0|^2 - 1 \, \|_{L^2}^2 + o(1) \, .$$

Hence, $|\, \mathbf{D} \,| \to 1$ a.e. in Ω_T as $k, h \to 0$. ∎

5. Computational experiments. Two new algorithms have been proposed and analyzed in the previous sections for problem (1.1)–(1.5):

(1) Solutions to Algorithm 3.1 satisfy a discrete energy law (Lemma 3.4) and unconditionally converge to weak solutions (Theorem 3.7) of the regularized problem ($\varepsilon > 0$) and generalized solutions (Theorem 3.8) of the limiting problem ($\varepsilon = 0$); we refer again to section 2.1 for a definition of the solution concepts.

(2) Algorithm 4.1 is based on a direct discretization of reformulation (1.16)–(1.18) of problem (1.1)–(1.5) to avoid penalization: Existence of solutions has been proven under a restrictive mesh constraint given in Theorem 4.2; in addition, the solutions satisfy a (modified) discrete energy law (Lemma 4.1). Convergence to generalized solutions of (1.1)–(1.5) is not known so far.

In this section, we present numerical results for two test problems using both algorithms: The first one has a smooth solution and is used to investigate the convergence behavior of the schemes with respect to the discretization parameters. The second example shows annihilation of (regularized) initial singularities.

We use Newton's method for the solution of the nonlinear system at each time step. The stopping criterion used is to require the L^2-norm of the residual to be less than 10^{-10}. For this, approximately three Newton iterations per time step are sufficient in our computations.

5.1. Smooth solution. We approximate smooth solutions of (1.1)–(1.3). Although we did not prove strong convergence with rates in the case of existing smooth solutions, we conjecture this error behavior for both Algorithms 3.1 and 4.1. In particular, Algorithm 4.1 does not employ an additional penalty parameter, which is the motivation for comparing computed solutions for varying $\varepsilon > 0$ of Algorithm 3.1 with those of Algorithm 4.1.

Example 5.1. We consider (1.1)–(1.5) for $\Omega = (-1, 1)^2$, $\mathbf{g} \equiv 0$, and

$$\mathbf{u}_0 \equiv 0 \, , \qquad \mathbf{d}_0 = (\sin(a), \cos(a))^\top, \quad a := 2.0\pi(\cos(x) - \sin(y)) \, .$$

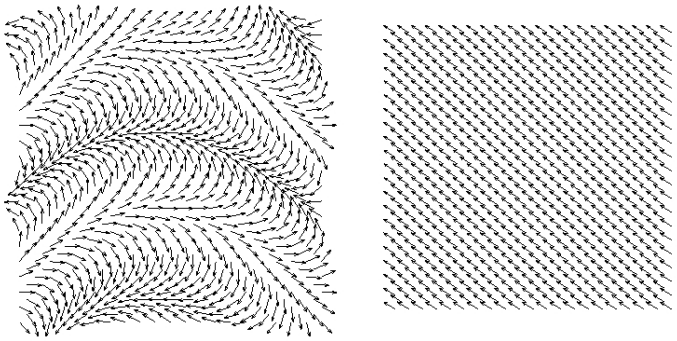


FIG. 5.1. *Example* 5.1*: Initial and final director fields (smooth solution).*

Parameters are set as $\lambda = \gamma = 1$ and $\nu = 0.1$.

The initial condition $\mathbf{d}_0$ and the final state towards which the system evolves are shown in Figure 5.1. We use a uniform crisscross triangulation of Ω with uniform mesh sizes $h = 1/20, 1/40, 1/80$.

First, we investigate the convergence behavior of Algorithm 4.1 with respect to the time-step size k and the mesh size h based on the decay of the elastic energy $\mathcal{J}_{\rm ela}(\mathbf{d}_h^m) = \frac{\lambda}{2}\int_\Omega |\nabla \mathbf{d}_h^m|^2 \,\mathrm{d}\mathbf{x}$. To this end, we compute on the finest mesh with varying $k > 0$, and with varying $h > 0$ for the finest time-step size. The corresponding approximate errors, defined as the difference to the most accurate solution, are shown in Figure 5.2. In addition, the violation of the sphere condition, measured in terms of the penalized energy $\mathcal{J}_{\rm pen}(\mathbf{d}_h^m) = \frac{\lambda}{\varepsilon^2}\big(F(\mathbf{d}_h^m), 1\big)_h$ (here with $\varepsilon = 1$), is shown with respect to varying h. As can be seen from the graphs, the algorithm shows a second order convergence with respect to h; the rate of convergence with respect to k is between first and second order.

Next, we present test results for Algorithm 3.1. For fixed $\varepsilon > 0$, the convergence behavior is similar to that of Algorithm 4.1 depicted in Figure 5.2. It is clear that ε has to be chosen in dependence of h. Figure 5.3 shows that it is quite delicate to balance ε and h in practice. Here, the computation results with a fixed mesh size h and time step k are presented. The discretization parameters are chosen sufficiently small in order to ensure that the difference between the next finer mesh and smaller time step are negligible for Algorithm 4.1. We then compare the different energies of interest with the results of Algorithm 4.1 on the same mesh with the same time step and with varying ε.

As can be seen from Figure 5.3, the correct decay in time of $\mathcal{J}_{\rm ela}$ requires a relatively precise choice of ε. A large choice of ε leads to a significant artificial energy $\mathcal{J}_{\rm pen}(\mathbf{d}_h^m)$; however, a small choice of ε leads to a large elastic energy and a too-small kinetic energy $\mathcal{J}_{\rm kin}(\mathbf{u}_h^m) = \frac{1}{2}\int_\Omega |\mathbf{u}_h^m|^2 \,\mathrm{d}\mathbf{x}$ (for a fixed mesh). Despite the more favorable theoretical results for Algorithm 3.1, it seems nontrivial to correctly capture the dynamics using the algorithm in this example.

5.2. Annihilation. The following example is taken from [31] to study annihilation of (regularized initial) singularities; cf. also [33, p. 284]. Here, the phenomenon critically depends on initial data, which in the limit ($\eta \to 0$) has infinite initial energy.

Example 5.2. Consider (1.1)–(1.5) for $\Omega = (-1,1)^2$, $\mathbf{g} \equiv 0$, and

$$\mathbf{u}_0 \equiv 0\,, \qquad \mathbf{d}_0 = \hat{\mathbf{d}}/\sqrt{|\hat{\mathbf{d}}|^2 + \eta^2} \qquad \text{for} \quad \hat{\mathbf{d}}(x_1, x_2) = (x_1^2 + x_2^2 - 0.25, x_2)^\top .$$

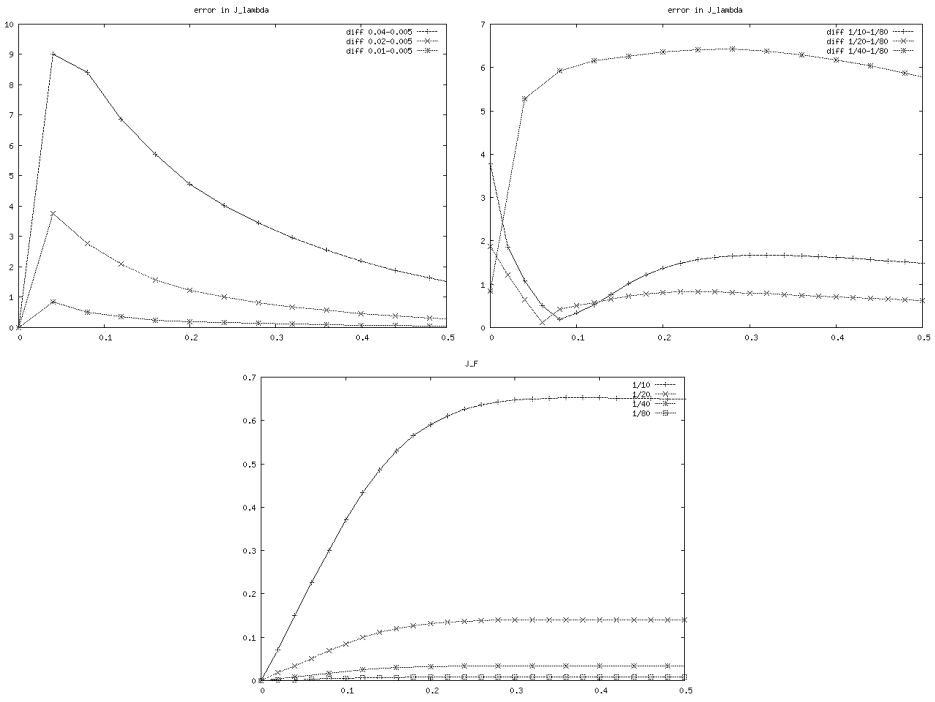


FIG. 5.2. *Example* 5.1*: Approximate errors (as a function of time) in energy* $\mathcal{J}_{\rm ela}$ *for* $k = 0.04, 0.02, 0.01$ *(top left) and* $h = 1/10, 1/20, 1/40$ *(top right);* $\mathcal{J}_{\rm pen}(\mathbf{d}_h^m)$ *for* $h = 1/10, 1/20, 1/40$ *(bottom).*

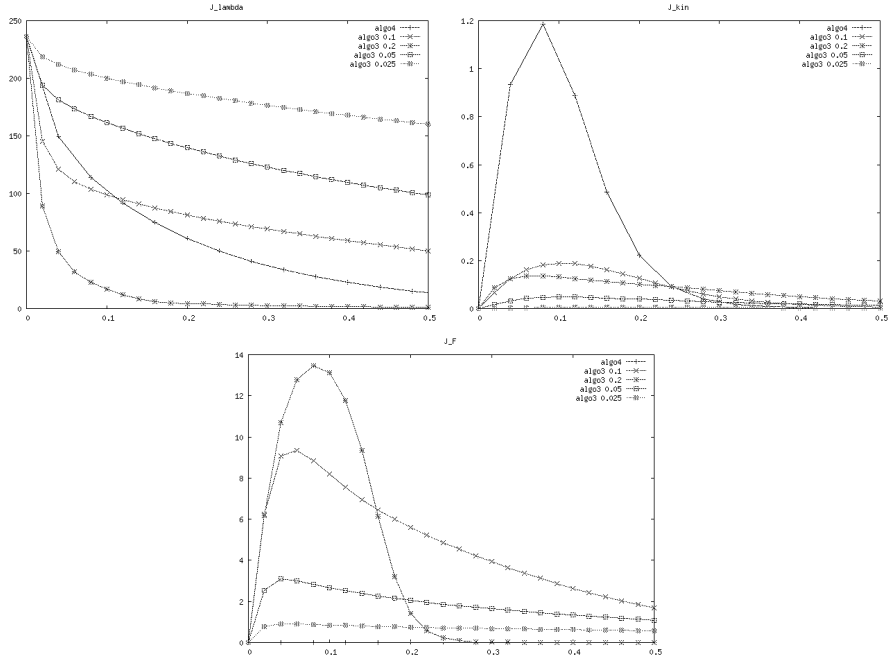


FIG. 5.3. *Example* 5.1*: Comparison of Algorithms* 4.1 *and* 3.1 *for different* $\varepsilon = 0.2, 0.1, 0.05, 0.025$*:* $\mathcal{J}_{\rm ela}$ *(top left),* $\mathcal{J}_{kin}$ *(top right), and* $\mathcal{J}_{\rm pen}(\mathbf{d}_h^m)$ *(bottom).*

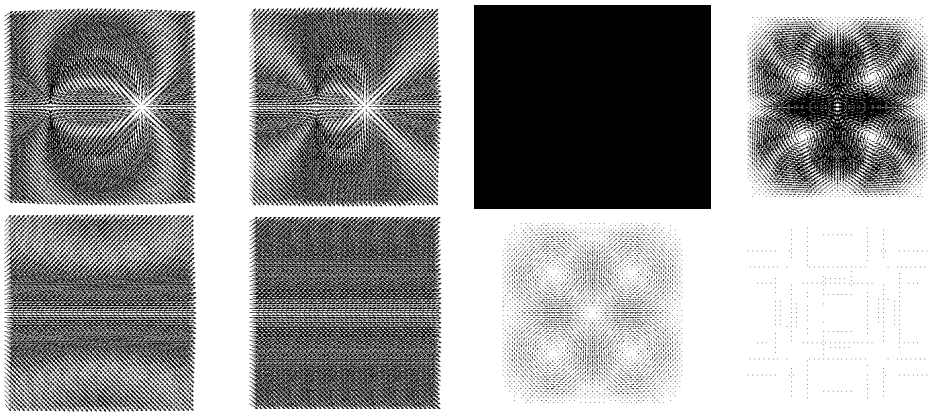


FIG. 5.4. *Example* 5.2, *using Algorithm* 3.1*: Snapshots at times* $t = 0, 0.3, 0.6, 0.9$ *of* $\{\mathbf{d}_h^m\}$ *(left) and* $\{\mathbf{u}_h^m\}$*(right). (The director and velocity fields are scaled by a factor of* 0.1 *and* 2.0*, respectively.)*

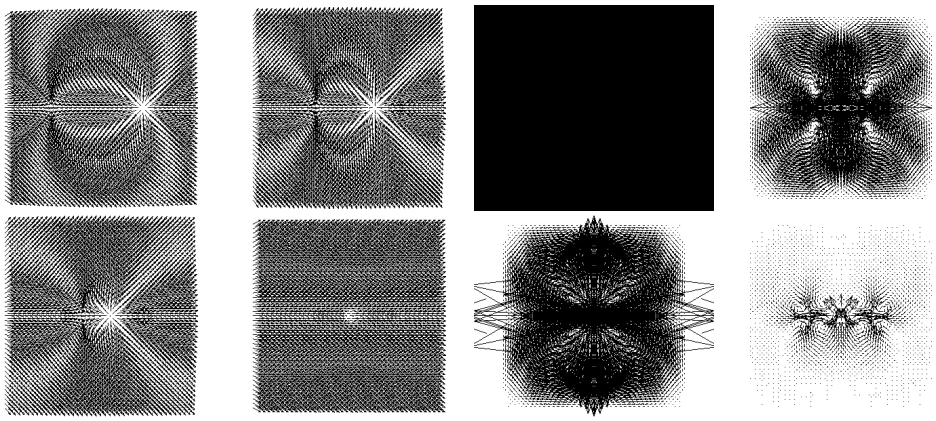


FIG. 5.5. *Example* 5.2, *using Algorithm* 4.1*: Snapshots at times* $t = 0, 0.3, 0.6, 0.9$ *of* $\{\mathbf{d}_h^m\}$ *(left) and* $\{\mathbf{u}_h^m\}$ *(right). (The director and velocity fields are scaled by a factor of* 0.1 *and* 2.0*, respectively.)*

Parameters are set as $\lambda = \gamma = \nu = 1$ and $\eta = 0.05$. We use a uniform crisscross triangulation of Ω, and $(\varepsilon, k, h) = (0.05, 0.01, 0.1)$ if not stated otherwise.

Snapshots of the evolution are displayed in Figure 5.4 for iterates $\{\mathbf{d}_h^m, \mathbf{u}_h^m\}$, which agree well with the computations presented in [31].

Figure 5.6 (first column) displays evolution of kinetic, elastic $\mathcal{J}_{\rm ela}(\mathbf{d}_h^m)$, and penalization $\mathcal{J}_{\rm pen}(\mathbf{d}_h^m)$ energies, as well as the total energy

$$\mathcal{J}_h(\mathbf{u}_h^m, \mathbf{d}_h^m) \equiv \mathcal{J}_{\varepsilon,\lambda,h}(\mathbf{u}_h^m, \mathbf{d}_h^m) := \mathcal{J}_{\rm kin}(\mathbf{u}_h^m) + \mathcal{J}_{\rm ela}(\mathbf{d}_h^m) + \mathcal{J}_{\rm pen}(\mathbf{d}_h^m)\,.$$

In the beginning of the simulation, we observe an oscillatory behavior of nonzero kinetic energy up to an annihilation time $t_{m^*} \approx 0.5$, where $\| \mathbf{u}^{m^*} \|_{L^\infty}$ is large. Both $\| \mathbf{u}^m \|_{L^\infty}$ and $\| \nabla \mathbf{d}^m \|_{L^\infty}$ rapidly approach zero for $m > m^*$; cf. Figure 5.6, first line. The dynamics of $\| \nabla \mathbf{d}_h^m \|_{L^\infty}$ for $m \approx m_*$ and annihilation time t_{m^*} are both sensitive to choices of $\varepsilon = \varepsilon(h)$, as is shown in Figure 5.7.

A drawback of the regularization approach (Algorithm 3.1) is the need to balance

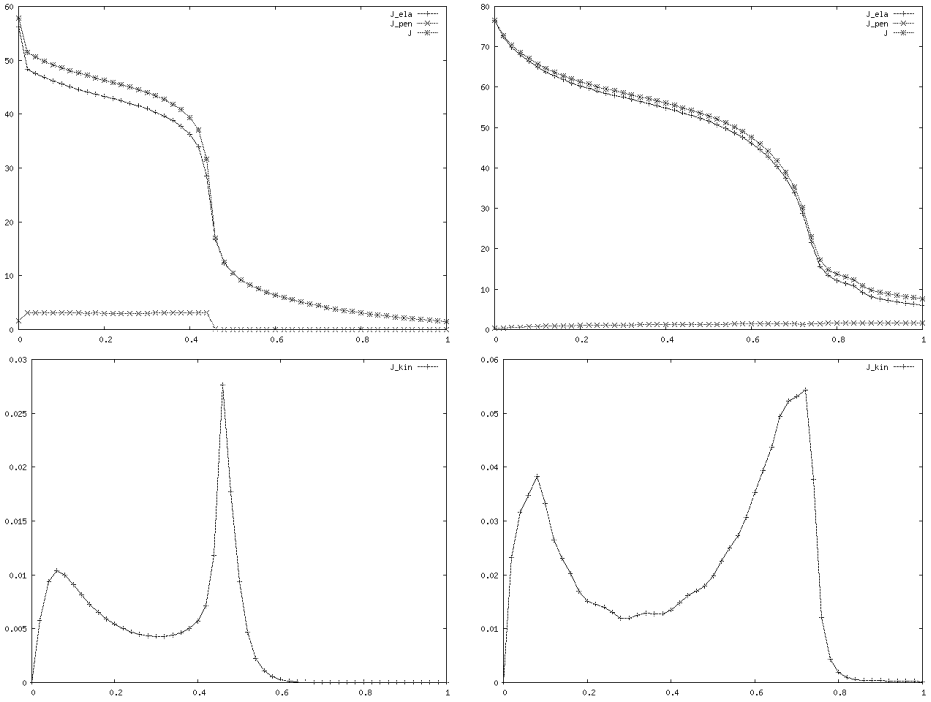


FIG. 5.6. *Example* 5.2*:* (a) *first line: Plot of elastic, penalty and total energies of tuple* $\{\mathbf{d}_h^m, \mathbf{u}_h^m\}$ *for Algorithm* 3.1 *(left) and Algorithm* 4.1 *(right), displayed in Figures* 5.4 *and* 5.5. (b) *second line: (Scaled) Plot of kinetic energies for Algorithm* 3.1 *(left) and Algorithm* 4.1 *(right).*

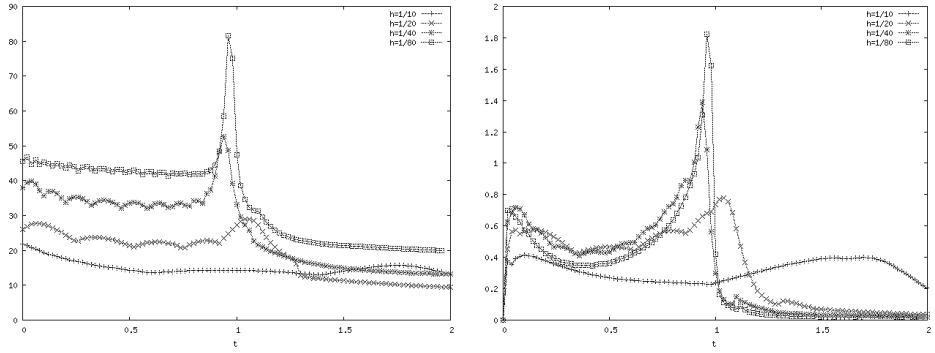


FIG. 5.7. *Example* 5.2*: Plot of* $L^\infty(\Omega)$ *for* $\nabla \mathbf{d}_h^m$ *(left) and* $\mathbf{u}_h^m$ *(right) for varying* $h = 1/10, 1/20, 1/40, 1/80$*, using Algorithm* 4.1.

penalization and discretization parameters $\varepsilon, k, h > 0$, which turned out to be quite subtle. In contrast, computation results of Algorithm 4.1 are much cleaner.

Next we use Algorithm 4.1 to solve the problem stated in Example 5.2. Snapshots of the simulation are collected in Figure 5.5, and the corresponding evolution of energies can be found in Figure 5.6 (second column). Finally, a possible blow-up behavior (simultaneous for $\| \mathbf{u}_h^m \|_{L^\infty}$, $\| \nabla \mathbf{d}_h^m \|_{L^\infty}$, as $m \to m^*$) is evident, with clear blow-up time t_{m^*}.

Finally, we investigate the behavior with respect to the parameter η used to

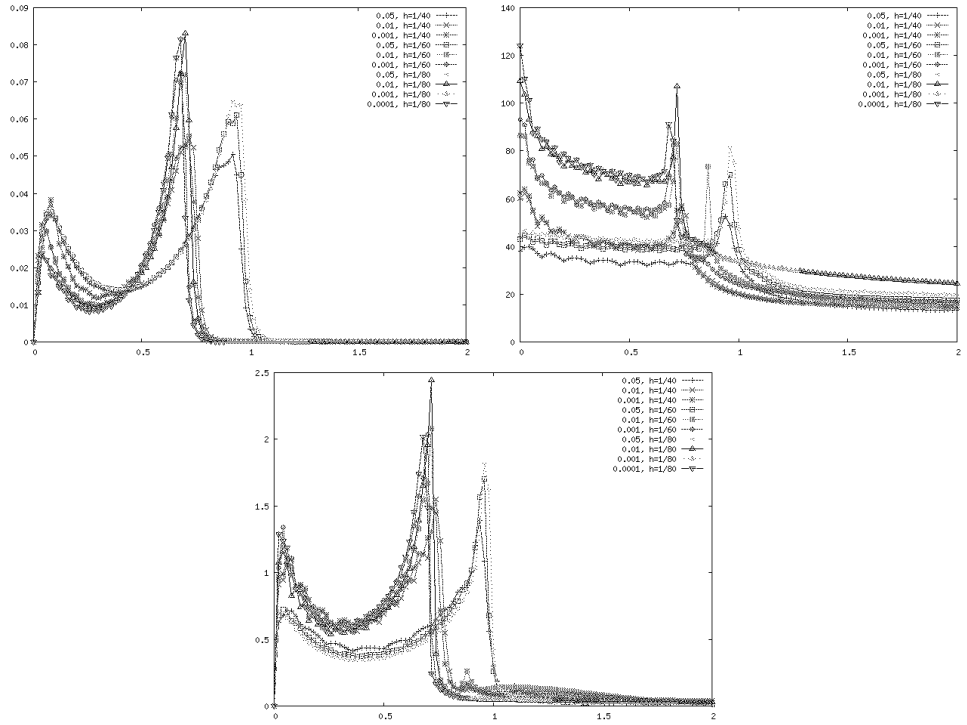


FIG. 5.8. *Example* 5.2*: Plot of* J *(top left),* $L^\infty(\Omega)$ *for* $\nabla \mathbf{d}_h^m$ *(top right), and* $\mathbf{u}_h^m$ *(bottom) for* $h = 1/40, 1/60, 1/80$ *and* $\eta = 0.05, 0.01, 0.001$*, using Algorithm* 4.1.

mollify the initial condition. Computational results based on Algorithm 4.1 are shown in Figure 5.8, where the total energy, velocity, and gradient of director in $L^\infty(\Omega)$ are plotted for varying $\eta = 0.1, 0.01, 0.001$ and mesh sizes $h = 1/20, 1/40, 1/80$.

Acknowledgment. This project started while X. F. and A. P. participated in the 2003 program *"Computational Challenges in Partial Differential Equations"* at the Isaac Newton Institute, Cambridge, UK. The authors are grateful to J. W. Barrett of Imperial College, London for many interesting suggestions and discussions on this topic. A. P. acknowledges support by the University of Pau during his stay as a "professeur invité."

REFERENCES

[1] R. A. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1975.
[2] F. ALOUGES, *A new algorithm for computing liquid crystal stable configurations: The harmonic mapping case*, SIAM J. Numer. Anal., 34 (1997), pp. 1708–1726.
[3] J. W. BARRETT, X. FENG, AND A. PROHL, *Convergence of a fully discrete finite element method for a degenerated parabolic system modeling nematic liquid crystals with variable degree of orientation*, M2AN Math. Model. Numer. Anal., 40 (2006), pp. 175–199.
[4] S. BARTELS, *Robust a priori error analysis for the approximation of degree-one Ginzburg–Landau vortices*, M2AN Math. Model. Numer. Anal., 39 (2005), pp. 863–882.
[5] S. BARTELS AND A. PROHL, *Constraint preserving implicit finite element discretization of harmonic map heat flow into spheres*, Math. Comp., 76 (2007), pp. 1847–1859.
[6] C. BERNARDI AND G. RAUGEL, *A conforming finite element method for the time-dependent the Navier–Stokes equations*, SIAM J. Numer. Anal., 22 (1985), pp. 455–473.
[7] F. BETHUEL, H. BREZIS, AND F. HÉLEIN, *Ginzburg-Landau Vertices*, Birkhäuser, New York, 1994.

[8] J. H. Bramble and J. Xu, *Some estimates for a weighted L^2 projection*, Math. Comp., 56 (1991), pp. 463–576.
[9] H. Brezis, *The interplay between analysis and topology in some nonlinear PDE problems*, Bull. Amer. Math. Soc. (N.S.), 40 (2003), pp. 179–201.
[10] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer, New York, 1992.
[11] Y. M. Chen and M. Struwe, *Regularity for the heat flow for harmonic maps*, Math. Z., 201 (1989), pp. 83–103.
[12] R. Cohen, S. Lin, and M. Luskin, *Relaxation and gradient methods for molecular orientation in liquid crystals*, J. Comput. Phys., 53 (1989), pp. 455–468.
[13] M. Crouzeix and V. Thomee, *The stability in L^p and W_p^1 of the L^2-projection onto finite element function spaces*, Math. Comp., 48 (1987), pp. 521–532.
[14] P. G. de Gennes and J. Prost, *The Physics of Liquid Crystals*, Oxford University Press, Oxford, 1993.
[15] Q. Du, B. Guo, and J. Shen, *Fourier spectral approximation to a dissipative system modeling the flow of liquid crystals*, SIAM J. Numer. Anal., 39 (2001), pp. 735–762; corrigendum: SIAM J. Numer. Anal., 41 (2003), pp. 796–798.
[16] J. Ericksen, *Conservation laws for liquid crystals*, Trans. Soc. Rheology, 5 (1961), pp. 22–34.
[17] J. Ericksen, *Continuum theory of nematic liquid crystals*, Res. Mechanica, 21 (1987), pp. 381–392.
[18] X. Feng and A. Prohl, *Numerical analysis of the Allen-Cahn equation and approximation for mean curvature flows*, Numer. Math., 94 (2003), pp. 33–65.
[19] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, 2nd ed., Springer, New York, 2000.
[20] V. Girault and P. A. Raviart, *Finite Element Method for Navier-Stokes Equations: Theory and Algorithms*, Springer, Berlin, Heidelberg, New York, 1981.
[21] J. G. Heywood and R. Rannacher, *Finite element approximation of the nonstationary Navier–Stokes problem.* I. *Regularity of solutions and second-order error estimates for spatial discretization*, SIAM J. Numer. Anal., 19 (1982), pp. 275–311.
[22] F. Leslie, *Some constitutive equations for liquid crystals*, Arch. Ration. Mech. Anal., 28 (1968), pp. 265–283.
[23] F. Leslie, *Theory of flow phenomena in liquid crystals*, in Advances in Liquid Crystals, Vol. 4, G. Brown, ed., Academic Press, New York, 1997, pp. 1–81.
[24] G. M. Lieberman, *Second Order Parabolic Differential Equations*, World Scientific, River Edge, NJ, 1996.
[25] F. H. Lin, *Nonlinear theory of defects in nematic liquid crystals: Phase transition and flow phenomena*, Comm. Pure Appl. Math., 42 (1989), pp. 789–814.
[26] F. H. Lin and C. Liu, *Nonparabolic dissipative systems modeling the flow of liquid crystals*, Comm. Pure Appl. Math., 48 (1995), pp. 501–537.
[27] F. H. Lin and C. Liu, *Partial regularities of the nonlinear dissipative systems modeling the flow of liquid crystals*, Discrete Contin. Dynam. Systems, 2 (1996), pp. 1–22.
[28] F. H. Lin and C. Liu, *Existence of solutions for Ericksen-Leslie system*, Arch. Ration. Mech. Anal., 154 (2000), pp. 135–156.
[29] S.-Y. Lin and M. Luskin, *Relaxation methods for liquid crystal problems*, SIAM J. Numer. Anal., 26 (1989), pp. 1310–1324.
[30] P. L. Lions, *Mathematical Topics in Fluid Mechanics, Vol.* 1: *Incompressible Models*, Oxford Science Publications, New York, 1996.
[31] C. Liu and N. J. Walkington, *Approximation of liquid crystal flows*, SIAM J. Numer. Anal., 37 (2000), pp. 725–741.
[32] C. Liu and N. J. Walkington, *Mixed method for the approximation of liquid crystal flows*, M2AN Math. Model. Numer. Anal., 36 (2002), pp. 205–222.
[33] A. Prohl, *Computational Micromagnetism*, Teubner, Stuttgart, 2001.
[34] A. D. Rey and M. M. Denn, *Dynamical phenomena in liquid-crystalline materials*, Annu. Rev. Fluid Mech., 34 (2002), pp. 233–266.
[35] M. Struwe, *Geometric Evolution Problems*, IAS/Park City Math. Ser. 2, AMS, Providence, RI, 1996, pp. 259–339.
[36] R. Temam, *Navier-Stokes Equations*, Stud. Math. Appl. 2, North-Holland, Amsterdam, 1977.
[37] E. G. Virga, *Variational Theories for Liquid Crystals*, Chapman & Hall, London, 1994.

SIAM J. NUMER. ANAL.
Vol. 46, No. 4, pp. 1732–1768
© 2008 Society for Industrial and Applied Mathematics

ADAPTIVE WEAK APPROXIMATION OF DIFFUSIONS WITH JUMPS*

E. MORDECKI†, A. SZEPESSY‡, R. TEMPONE§, AND G. E. ZOURARIS¶

Abstract. This work develops adaptive time stepping algorithms for the approximation of a functional of a diffusion with jumps based on a jump augmented Monte Carlo Euler–Maruyama method, which achieve a prescribed precision. The main result is the derivation of new expansions for the time discretization error, with computable leading order term in a posteriori form, which are based on stochastic flows and discrete dual backward functions. Combined with proper estimation of the statistical error, they lead to efficient and accurate computation of global error estimates, extending the results by A. Szepessy, R. Tempone, and G. E. Zouraris [*Comm. Pure Appl. Math.*, 54 (2001), pp. 1169–1214]. Adaptive algorithms for either deterministic or trajectory-dependent time stepping are proposed. Numerical examples show the performance of the proposed error approximations and the adaptive schemes.

Key words. diffusions with jumps, weak approximation, error control, Euler–Maruyama method, a posteriori error estimates, backward dual functions

AMS subject classifications. 65C30, 65Y20, 65L50, 65H35, 60H35, 60H10

DOI. 10.1137/060669632

1. Introduction.

Since the introduction by Kiyoshi Itô of the theory of stochastic integration [26], diffusion processes have become a benchmark model in the mathematical sciences. Nevertheless, the study of more complex phenomena, including abrupt changes, pathwise features different from the ones of diffusions, and heavy tails, leads to the introduction of models with jumps.

In our view, the introduction of jumps in stochastic modeling is realized at least in the following three major directions. First, diffusions with jumps, the case we consider in this work, are obtained by adding jumps to a diffusion-type dynamics. Jumps have exponentially distributed interarrival times and amplitudes generated by time-dependent mark distributions. This jump structure ensures the unpredictability of the epoch and the magnitude of the jumps, a fact that is observed in financial markets, where it leads to the incompleteness of the market. This first approach was initiated with the work by Merton [37]. For other models we refer to Cont and Tankov [11] and the references therein. In a second approach, the jump intensity depends on the state of the process, implying a certain possibility to predict the jump

*Received by the editors September 12, 2006; accepted for publication (in revised form) January 11, 2008; published electronically April 11, 2008. This work was supported by Swedish Research Council for Engineering Science (TFR) grant 222-148, the Swedish National Network in Applied Mathematics (NTM), start up funds from the School of Computational Science, Florida State University, the Dahlquist research fellowship at KTH, Project 10.101 FCE - DINACYT, Uruguay, CSIC - Udelar, Human Resources Program, the Facultad de Ciencias, Udelar, Montevideo, Uruguay, and the Institute of Applied and Computational Mathematics, FO.R.T.H., Greece.

http://www.siam.org/journals/sinum/46-4/66963.html

†Centro de Matemática, Facultad de Ciencias, Universidad de la República, Iguá 4225 C.P. 11400, Montevideo, Uruguay (mordecki@cmat.edu.uy).

‡Department of Mathematics, KTH, S-100 44 Stockholm, Sweden (szepessy@nada.kth.se).

§Department of Mathematics and School of Computational Science, Florida State University, 400 Dirac Science Library, Tallahassee, FL 32306-4120, and School of Computer Sciences and Communication, KTH, S-100 44 Stockholm, Sweden (rtempone@scs.fsu.edu).

¶Department of Mathematics, University of Crete, GR-714 09 Heraklion, Crete, Greece, and Institute of Applied and Computational Mathematics, FO.R.T.H., GR-711 10 Heraklion, Crete, Greece (zouraris@math.uoc.gr).

epoch. State-dependent jump diffusion processes are intensively used in default risk theory, for instance, in the form of affine jump diffusion processes [13]. The third approach assumes that not only the large movements but also the small movements observed in the dynamics are generated by jumps. In some works even the Brownian driving process is replaced by a stable-type process (i.e., a Lévy process with jump measure with infinite mass around zero). This gives a process with an infinite number of jumps on every bounded time interval, a situation known as *infinite activity*. Initial references to infinite activity are, for instance, [34] and [17]. More recent works include [10].

The computation of functionals of the above-mentioned stochastic processes, when closed solutions are not available, led to the development of time discretization schemes, following the approach by Maruyama [35]. In this wide field of applied mathematics, standard references are the books by Kloeden and Platen [31] and by Milstein and Tretyakov [39]. The main emphasis in these books is on the development of different numerical schemes and the determination of their *order* of convergence. Consider a given functional θ (for instance, an expectation, related to an option price in finance) of a stochastic process defined on a time interval $[0, T]$. Assuming the necessary regularity, we say that a scheme has (weak) order r if it produces a sample average approximate value, $\widehat{\theta}$, such that $|\widehat{\theta}-\theta| = \mathcal{O}\big((\Delta t_{max})^r + M^{-1/2}\big)$, using a time discretization $\{0 = t_0 < t_1 < \cdots < t_N = T\}$ and M independent samples.

Motivated by practical reasons, we formulate in this work a weak approximation problem in the following way: given the functional θ, the numerical scheme for time discretization, and a certain desired accuracy $\mathrm{TOL} > 0$, produce a *computable* sample average estimator $\widehat{\theta}$ such that $|\widehat{\theta} - \theta| \leq \mathrm{TOL}$ with a probability close to one. We also seek to construct $\widehat{\theta}$ using minimal computational effort.

To solve this problem we use the so-called jump augmented Euler–Maruyama scheme (also known as jump adapted Euler–Maruyama scheme; cf. [47]) and develop novel time discretization error estimation formulas. Based on these computable error estimates we propose *a posteriori adaptive* methods that compute the approximation $\widehat{\theta}$ and estimate its numerical error due to time discretization and statistical sampling. If the prescribed accuracy TOL is not yet achieved, the adaptive algorithms decide whether to subdivide a set of time intervals with largest error contribution or to increase the number of realizations, up to the point of achieving the accuracy goal. To the best of our knowledge, there are no references that follow this approach for any class of jump processes.

Regarding the jump intensity λ, we consider a moderate number of jumps, motivated by reported results from the calibration of jump diffusion models in finance. For instance, the value $\lambda = 0.089$ was reported for the calibration of the Merton model with option prices on the S&P 500 index, and $\lambda \approx 1$ for the Kou model applied to the German DAX index; see [1] and [12]. These empirical facts allow us to rely on the augmentation of the grid with the jump times, reported to be more accurate for strong approximation of diffusion with jumps [6]. In Remark 4.1 we discuss the corresponding additional work due to the inclusion of the jump points in the time discretization.

The main contribution in this work is the generalization of the approach proposed by Szepessy, Tempone, and Zouraris for diffusions in [51] to the context of diffusion with large jumps. This generalization is based on the probabilistic properties of diffusions with jumps, performing two steps of simulation. The first step relates to the jump structure, and the second relates to the diffusion part. We provide

corresponding analysis for the approximation errors in each step. In section 2 we present an example with a singularity and many jumps. There, our new algorithms save a factor of one hundred in the number of time steps required to produce the same accuracy as the uniform time stepping algorithm. Furthermore, for one of those algorithms the additional computational work to generate the adaptive time steps is negligible.

The other two cases not considered here, namely, diffusions with state-dependent intensity jumps and high intensity of small jumps, seem to require other techniques and would therefore be interesting for future work; see the end of section 1.2.1.

Efficient adaptive time stepping methods, with theoretical basis, use a posteriori error information, since in general a priori knowledge is not as precise as the a posteriori one. Adaptive time stepping methods can be based on Theorem 2.2 or Theorem 3.1, where error estimates with leading order terms in computable a posteriori form are obtained. These estimations require the use of stochastic flows and dual functions in the presence of jumps. Dual functions are standard in optimal control theory, in particular for adaptive mesh control for ordinary and partial differential equations (see [4], [5], [16], [28], [29], [41], [43], and [44]) and were successfully applied in the probabilistic context in [15], [40], [45], and [51].

Alternative deterministic methods involve the numerical approximation of the Kolmogorov backward equation (2.5), which is a nonlocal partial integrodifferential equation. Direct discretization of such time-dependent integrodifferential equations requires the approximation of integrals at each time step and may suffer from the curse of dimensionality. On the other hand, the Euler–Maruyama method avoids these expensive computations. Examples of this deterministic approach include the work by Andersen and Andreasen [1], which uses operator splitting to approximate the integral term explicitly in Fourier space while approximating the other terms in the equation implicitly, and the work by Matache, von Petersdorff, and Schwab [36], which discretizes the partial integrodifferential equation by the θ-scheme in time and a wavelet Galerkin method in space.

For work which proposes and analyzes numerical methods for diffusions with jumps, we refer the reader to Liu and Li [33], where weak convergence schemes are analyzed, and Higham and Kloeden [22], [23], [24], where implicit methods for diffusion with jumps are presented, analyzing strong convergence and nonlinear stability. We refer also to Bruti-Liberati and Platen [6], [7], [8], [9], Kubilius and Platen [32], and Mikulevičius and Platen [38].

1.1. Problem's setting. Let $d \in \mathbb{N}$, $T > 0$ and let $X : [0,T] \to \mathbb{R}^d$ be a stochastic process solving the following stochastic differential equation:

$$\begin{aligned} X(t) = X(0) &+ \int_0^t a(\, s, X(s^-)\,)\, ds + \sum_{\ell=1}^{\ell_0} \int_0^t b^\ell(\, s, X(s^-)\,)\, dW^\ell(s) \\ &+ \int_0^t \int_{\mathbf{z}} c(\, s, X(s^-), z\,)\, p(ds, dz) \quad \forall\, t \in [0,T] \end{aligned} \tag{1.1}$$

(see [19, Chap. II.2], [27, Chap. III.2c], or [25]), with initial condition $X(0) = x_0$ that we assume deterministic for simplicity (see Remark 2.1 for a random initial condition). The randomness in this equation is generated by the following:

(i) A standard ℓ_0-dimensional Wiener process

$$W = \{W(t) = (W^1(t), \ldots, W^{\ell_0}(t)) \colon t \in [0,T]\},$$

defined on a filtered probability space $(\Omega^W, \mathcal{F}^W, \mathbf{F}^W, P^W)$, where

$$\mathbf{F}^W = \{\mathcal{F}_t^W : t \in [0,T]\}$$

is a filtration and $\mathcal{F}_T^W = \mathcal{F}^W$.

(ii) A Poisson random measure $p(\omega, dt, dz)$ on $[0,T]\times\mathbf{Z}$, where $\mathbf{Z} \equiv \mathbb{R}^{\ell_1}\backslash\{0\}$, with $\ell_1 \in \mathbb{N}$, and deterministic compensator $q(dt,dz) = \lambda(t)\mu(t,dz)dt$. Here $\lambda\colon [0,T] \to [0,\lambda_{\max}]$ is the time intensity of jumps with $\lambda_{\max} \in (0,+\infty)$ and, for each fixed $t \geq 0$, the measure $\mu(t,dz)$ is a probability on $\mathbf{Z}$. The jump measure is defined in a filtered probability space $(\Omega^J, \mathcal{F}^J, \mathbf{F}^J, P^J)$, where $\mathbf{F}^J = \{\mathcal{F}_t^J : t \in [0,T]\}$ is a filtration and $\mathcal{F}_T^J = \mathcal{F}^J$.

As a consequence, the process X is defined on a stochastic basis $(\Omega, \mathcal{F}, \mathbf{F}, P)$, where $\Omega = \Omega^W \times \Omega^J$, $\mathbf{F} = \{\mathcal{F}_t^W \otimes \mathcal{F}_t^J : t \in [0,T]\}$, and $P = P^W \times P^J$; as usual in this framework, the Wiener process and the Poisson random measure are mutually independent. The coefficients $a\colon [0,T]\times\mathbb{R}^d \to \mathbb{R}^d$, $b\colon [0,T]\times\mathbb{R}^d \to \mathbb{R}^{d\times\ell_0}$, and $c\colon [0,T]\times\mathbb{R}^d\times\mathbf{Z} \to \mathbb{R}^d$ are assumed to be Borel functions, satisfying regularity conditions stated in Lemma 2.1.

Given a scalar function $g\colon \mathbb{R}^d \to \mathbb{R}$, a functional of the stochastic process X of the form $\theta = E\big[\, g\big(X(T)\big)\big]$, and a tolerance $\mathrm{TOL} > 0$, the goal of our work is to derive an algorithm that produces a computable approximation $\widehat{\theta}$ to θ such that

$$\big|\, \widehat{\theta} - \theta \,\big| \leq \mathrm{TOL}$$

with a probability close to one and minimal computational effort.

1.2. The jump augmented Euler–Maruyama simulation scheme. The scheme has two steps. After choosing an initial deterministic partition of the time interval $[0,T]$ we sample the jump epochs and magnitudes, to later implement the numerical method in the *augmented* grid, including the original points and the jump points. We first describe the jump structure construction.

1.2.1. Construction of the integral with respect to the Poisson random measure. Consider a sequence $(\mathbf{e}_j)_{j=1}^{\infty}$ of random variables defined on a probability space $(\Omega^J, \mathcal{F}^J, P^J)$, which are independent and have common exponential distribution with parameter one. Define an auxiliary function

$$\Lambda(t) = \int_0^t \lambda(s)\, ds, \quad t \in [0,T]. \tag{1.2}$$

Then the number of jumps of the random Poisson measure $p(dt,dz)$ in an interval $[0,t]$ is

$$\nu(t) = \max\left\{ k\colon \sum_{j=1}^{k} \mathbf{e}_j \leq \Lambda(t) \right\}$$

and the total number of jumps in $[0,T]$ is denoted by $\nu \equiv \nu(T)$. The jump times of the Poisson measure can be defined by $\tau_0 = 0$, and $\tau_k = \Lambda^{-1}(\mathbf{e}_1 + \cdots + \mathbf{e}_k)$ for $k = 1,\ldots,\nu$, and can be computed recursively by solving for τ_k in the equations

$$\mathbf{e}_k = \int_{\tau_{k-1}}^{\tau_k} \lambda(s)\, dt = \Lambda(\tau_k) - \Lambda(\tau_{k-1}) \quad \text{for} \quad k = 1,\ldots,\nu. \tag{1.3}$$

Once the jump times are computed, we proceed to sample the marks $\{Z_k\}$, which, conditionally on the values of the jumps times, are independent random variables distributed respectively according to $\{\mu(\tau_k, dz)\}$. Observe that the random measure $p(dt, dz)$, characterized by its intensity $q(dt, dz) = \lambda(t)\mu(t, dz)dt$, can then be constructed as

$$p(ds, dz) = \sum_{k=1}^{\nu} \delta_{(\tau_k, Z_k)}(ds, dz).$$

Here $\delta_{(t,Z)}(ds, dz)$ is a Dirac distribution at the point (t, Z). As a consequence of this construction, the stochastic integral with respect to the Poisson random measure, i.e., the last term in (1.1), can be computed as a finite sum, i.e.,

$$\int_0^t \int_{\mathbf{z}} c(\, s, X(s^-), z\,)\, p(ds, dz) = \sum_{k=1}^{\nu(t)} c\big(\tau_k, X(\tau_k^-), Z_k\big), \quad t \in [0, T].$$

Some relevant remarks concerning the jump structure and its simulation are in order. First, the number of jumps in the interval $[0, T]$ is almost surely finite, its expectation being $\Lambda(T) = \int_0^T \lambda(s)ds \le \lambda_{\max} T$. Second, since the time intensity $\lambda(t)$ is deterministic, the jump times $\{\tau_k\}$ can be directly obtained before solving for the process X; it is enough to find the function Λ by accurately integrating (1.2) and successively finding $\{\tau_k\}$ by solving (1.3). Furthermore, if λ is a constant, then we simply sample $\tau_k = (\mathbf{e}_1 + \cdots + \mathbf{e}_k)/\lambda$. The jump augmented Euler–Maruyama scheme admits a direct generalization in the case of stochastic intensity, as long as this intensity is independent of the Wiener process and it admits exact simulation. This is the case, for instance, of a regime-switching intensity process (see, for instance, [2]).

In the second approach described before, the jump intensity λ depends on the current value $X(t)$ of the process. Then it is not possible to sample the jump times separately from the process and it is necessary to implement a different algorithm. For instance, see Glasserman and Merener [20], [21], where a rejection algorithm is proposed, reducing the case of uniformly bounded state-dependent intensity $\lambda = \lambda(t, x) \le \lambda_o(t) < \lambda_{\max}$ to the case of a time-dependent intensity $\lambda = \lambda_o(t) < \lambda_{\max}$, which is our case.

Regarding the third approach described above and the possibility of adapting our methodology to the case of infinite activity, a possible approach is to separate small and large jumps, simulating the large jumps and approximating the small jumps by an additional independent Brownian motion, as suggested by Asmussen and Rosiński [3]. Nevertheless, one should be cautious, as the efficient use of the jump augmented Euler–Maruyama method assumes a moderate number of jumps, while the cut-off approximation of the small jump part by a Brownian motion may produce a high intensity compound Poisson process. This delicate tradeoff between "small" and "large" jumps is characteristic in infinite activity Lévy processes problems, should be separately examined, and is out of the scope of the present work (see [50]).

1.2.2. Jump augmented Euler–Maruyama time stepping algorithm. After constructing the jumps and its marks as in section 1.2.1, we simulate the trajectory $\overline{X}$ using the following scheme.

Jump augmented Euler–Maruyama simulation scheme.

Input: Given N, time nodes $0 = \tilde{t}_0 < \tilde{t}_1 < \cdots < \tilde{t}_N = T$, and sample jump times $0 < \tau_1(\omega) < \cdots < \tau_{\nu(\omega)}(\omega) < T$ with corresponding marks $Z_1(\omega), \ldots, Z_{\nu(\omega)}(\omega)$,

set the jump counter $k = 1$.

Time stepping: Consider an augmented partition given by the union

$$\{t_n(\omega)\}_{n=0}^{\kappa(\omega)} = \{\tilde{t}_n\}_{n=0}^{N} \cup \{\tau_n(\omega)\}_{n=0}^{\nu(\omega)},$$

where $\kappa(\omega) = N + \nu(\omega)$ (a.s.) is the number of time steps.

Set the initial condition $\overline{X}(t_0, \omega) = x_0$.

For time steps $n = 0, \ldots, \kappa(\omega) - 1$

Compute the remaining approximate grid values, $\overline{X}(t_{n+1}, \omega)$, by first constructing the left limit value of the approximated process,

$$\begin{aligned}\overline{X}(t_{n+1}^-, \omega) &= \overline{X}(t_n, \omega) + a(t_n, \overline{X}(t_n, \omega))(t_{n+1} - t_n) \\ &\quad + \sum_{\ell=1}^{\ell_0} b^\ell(t_n, \overline{X}(t_n, \omega))(W^\ell(t_{n+1}, \omega) - W^\ell(t_n, \omega)).\end{aligned} \tag{1.4}$$

When needed, introduce the correction due to jump discontinuities:

if $(t_{n+1} = \tau_k(\omega))$ **then**

$$\overline{X}(t_{n+1}, \omega) = \overline{X}(t_{n+1}^-, \omega) + c\big(t_{n+1}, \overline{X}(t_{n+1}^-, \omega), Z_k(\omega)\big), \tag{1.5}$$

increase k to $k + 1$

else

$$\overline{X}(t_{n+1}, \omega) = \overline{X}(t_{n+1}^-, \omega), \tag{1.6}$$

end-if

end-for

Several important particular cases deserve comment. First, taking $c = 0$ we obtain the previous algorithm in [51]. For the other particular case when $a = b = 0$ the approximate process $\overline{X}$ has the same law as X, due to the form of the grid proposed in the numerical method (see [20]). A similar situation occurs when only $b = 0$; there, conditioned on the jumps information, a higher order method for ordinary differential equation integration should be used to approximate $\overline{X}$ between jump times, giving a time discretization error which is essentially negligible with respect to the statistical error.

1.3. Error control. The aim, for a given TOL > 0, is to choose time steps $\{0 = \tilde{t}_0 < \tilde{t}_1 < \cdots < \tilde{t}_N = T\}$ and the number M of independent identically distributed (i.i.d.) samples $\{\overline{X}(\cdot\,, \omega_j)\}_{j=1}^{M}$ such that the computational work, defined as M times the average number of time steps, i.e., $M \times E[\kappa]$ is minimal, constrained by the condition that the computational error, $\mathcal{E}$, defined by

$$\mathcal{E} = E[g(X(T))] - \tfrac{1}{M} \sum_{j=1}^{M} g(\overline{X}(T, \omega_j)), \tag{1.7}$$

is such that the event

$$\big|\mathcal{E}\big| \leq \text{TOL} \tag{1.8}$$

has a probability close to one. The computational error $\mathcal{E}$ naturally splits as the sum of the (deterministic) time discretization error $\mathcal{E}_T$ and the statistical error $\mathcal{E}_S$ given,

respectively, by

$$\mathcal{E}_T = E\big[g(X(T))\big] - E\big[g(\overline{X}(T))\big], \tag{1.9}$$

$$\mathcal{E}_S = E[g(\overline{X}(T))] - \tfrac{1}{M}\sum_{j=1}^{M} g(\overline{X}(T,\omega_j)). \tag{1.10}$$

The number M of independent realizations of $\overline{X}$ is computed from the statistical properties of $\mathcal{E}_S$. Theorems 2.2 and 3.1 state approximations for the time discretization error in terms of error contributions from each time subinterval, that is,

$$\mathcal{E}_T \approx E\left(\sum_m \rho(t_m,\omega)(\Delta t_m)^2\right)$$

(see (2.7) and (3.4)). These approximations are used to estimate the time discretization error in the adaptive algorithms that we propose. The error density $\rho(t_m,\omega_j)$ in (2.8) and (3.5), respectively, are computed through the discrete dual functions that we describe in section 1.4.

1.4. Discrete dual equations. This subsection is dedicated to the determination of the discrete dual functions $\varphi(t) \in \mathbb{R}^d$, $\varphi'(t) \in \mathbb{R}^{d\times d}$, and $\varphi''(t) \in \mathbb{R}^{d\times d\times d}$, where t is a node of the (stochastic) partition of the time interval $[0,T]$. First, we introduce some notation.

DEFINITION 1.1 (local discrete solution operator). *Write the Euler time stepping for the time nodes $t = t_0, t_1^-, t_1, t_2^-, t_2, \dots$ as*

$$\overline{X}(t_{next},\omega) = \Phi(t,\overline{X}(t)), \tag{1.11}$$

where the next computational time is given by

$$t_{next} = \begin{cases} t_n & \text{if} \quad t = t_n^-, \\ t_{n+1}^- & \text{if} \quad t = t_n, \end{cases}$$

and, correspondingly, the previous computational time is

$$t_{prev} = \begin{cases} t_{n-1} & \text{if} \quad t = t_n^-, \\ t_n^- & \text{if} \quad t = t_n. \end{cases}$$

Observe that, depending on ω and t, we may have

$$\Phi(t,x,\omega) = \begin{cases} x + a(t,x)\Delta t + b^\ell(t,x)\Delta W^\ell & \text{if } t = t_n, \\ x + c(t,x,Z_k(\omega)) & \text{if } t = t_n^- \text{ and } t_n \text{ is the } k\text{th jump time}, \\ x & \text{if } t = t_n^- \text{ and } t_n \text{ is not a jump time}. \end{cases}$$

Then, for each realization, φ, φ', and φ'' are constructed by the following algorithm.

DUAL BACKWARD TIME STEPPING ALGORITHM.

Set the initial backward values

$$t = T, \tag{1.12}$$
$$\varphi_i(T) = \partial_i g\big(\overline{X}(T)\big),$$
$$\varphi'_{ik}(T) = \partial_{ik} g\big(\overline{X}(T)\big),$$
$$\varphi''_{ikm}(T) = \partial_{ikm} g\big(\overline{X}(T)\big).$$

while $t > 0$ set $t = t_{prev}$

$$\varphi_i(t) = \partial_i \Phi_j(t, \overline{X}(t))\, \varphi_j(t_{next}), \tag{1.13}$$
$$\varphi'_{ik}(t) = \partial_i \Phi_j(t, \overline{X}(t))\, \partial_k \Phi_p(t, \overline{X}(t))\, \varphi'_{jp}(t_{next}) \tag{1.14}$$
$$+ \partial_{ik} \Phi_j(t, \overline{X}(t))\, \varphi_j(t_{next}), \tag{1.15}$$
$$\begin{aligned}\varphi''_{ikm}(t) = {} & \partial_i \Phi_j(t, \overline{X}(t)) \partial_k \Phi_p(t, \overline{X}(t)) \partial_m \Phi_r(t, \overline{X}(t)) \varphi''_{jpr}(t_{next}) \\ & + \partial_{im} \Phi_j(t, \overline{X}(t)) \partial_k \Phi_p(t, \overline{X}(t)) \varphi'_{jp}(t_{next}) \\ & + \partial_i \Phi_j(t, \overline{X}(t)) \partial_{km} \Phi_p(t, \overline{X}(t)) \varphi'_{jp}(t_{next}) \\ & + \partial_{ik} \Phi_j(t, \overline{X}(t)) \partial_m \Phi_p(t, \overline{X}(t)) \varphi'_{jp}(t_{next}) \\ & + \partial_{ikm} \Phi_j(t, \overline{X}(t)) \varphi_j(t_{next}).\end{aligned} \tag{1.16}$$

end-while

1.5. Adaptivity. After choosing the initial jump augmented grid, we propose two different ways of a posteriori refinement or time stepping in order to achieve the prescribed TOL: a first algorithm that refines simultaneously all the simulated trajectories, which we call *deterministic*, and a second one that refines the simulated trajectories one by one, which we call *stochastic*.

In the deterministic algorithm, the refinement is taken in order to achieve

$$(\Delta \tilde{t}_n)^2 |E[\rho(\tilde{t}_n, \omega)]| = \text{constant}.$$

Observe that we can only control the deterministic points $\tilde{t}_n$ and not the whole augmented grid. This is the basis of our adaptive algorithm in section 4.3.

In the stochastic algorithm the time steps are determined by the condition

$$(\Delta \tilde{t}_n)^2 |\rho(\tilde{t}_n, \omega)| = \text{constant},$$

where, in contrast to the previous algorithm, there is no expectation. This is achieved through a test performed at each interval of each realization, to decide whether to refine the given interval (see section 4.4). In this case, when a node is added, a Brownian bridge is used to obtain the value of both the Wiener process and the approximated process, $\overline{X}$, at this new intermediate time node.

Stochastic time steps are advantageous for problems with singularities at random times, but since their use entails more work per realization than the deterministic time steps they should be judiciously used. A natural application of stochastic time steps appears in the weak approximation of killed diffusions; see [15]. The optimal stochastic steps depend on the whole solution $\overline{X}(t)$, $0 < t < T$, and in particular the step $\Delta t(t)$ at time t depends also on $W(u)$ for $t < u$. In stochastic analysis the concept of *adaptedness* means, intuitively, that the values of the process at time t depend only on the events generated by the sources of randomness up to time t,

i.e., they do not depend on future events. In numerical analysis a method is said to be *adaptive* when the approximate solution is used to control the error, e.g., to determine the time steps. Our stochastic time stepping algorithm is in this sense adaptive and nonadapted, since the time steps $\Delta t(t)$ depend on values of $W(u)$ for $t < u$, i.e., on future values of the Wiener process. The backward evolutions (1.13)–(1.14) of the weight functions φ and φ' avoid solving for the two variables t, s present in $\frac{\partial \overline{X}(t)}{\partial x(s)}$, which appears in the forward t-evolution equation for $\frac{\partial \overline{X}(t)}{\partial x(s)}$ in the identity $\varphi_i(t_n) = \partial_j g(\overline{X}(T)) \frac{\partial \overline{X}_j(T)}{\partial x_i(t_n)}$; see (6.25). A solution with two variables s and t would require work of the order N^2 for each realization, instead of the corresponding work of the order N in Theorem 2.2.

The rest of the paper is organized as follows. Section 2 contains a first error expansion based on the dual functions and their first variations. Section 3 contains a second error expansion based on the duals and their first and second variations. Section 4 presents implementations of adaptive algorithms. Section 5 includes results from numerical experiments, and section 6 presents the proofs.

2. A first estimate of the time discretization error. In this section we present in Theorem 2.2 an error expansion in a posteriori computable form for the time discretization error $\mathcal{E}_T$ in (1.9) based on the first and second dual functions φ and φ'. All the proofs are given in section 6.

Our analysis uses the fact that the simulated process can be extended to the interval $[0, T]$. To do this we first define the piecewise constant coefficients

$$\begin{aligned} &\overline{a}(s; \overline{X}) \equiv a(t_n, \overline{X}(t_n)) \quad \forall\, s \in [t_n, t_{n+1}), \\ &\overline{b}^{\ell}(s; \overline{X}) \equiv b^{\ell}(t_n, \overline{X}(t_n)) \quad \forall\, s \in [t_n, t_{n+1}), \quad \ell = 1, \ldots, \ell_0, \end{aligned} \tag{2.1}$$

for $n = 0, \ldots, \nu - 1$, and then consider the stochastic differential equation

$$\begin{aligned} \overline{X}(t) = \overline{X}(0) &+ \int_0^t \overline{a}(s; \overline{X}) ds + \sum_{\ell=1}^{\ell_0} \int_0^t \overline{b}^{\ell}(s; \overline{X})\, dW^{\ell}(s) \\ &+ \int_0^t \int_{\mathbf{z}} c\big(s, \overline{X}(s^-), z\big)\, p(ds, dz) \quad \forall\, t \in [0, T]. \end{aligned} \tag{2.2}$$

A key point of our approach is the possibility of perfect simulation of the jump structure, avoiding the necessity of the consideration of an approximated jump coefficient $\bar{c}$. For simplicity we introduce the notation $\partial_k \equiv \frac{\partial}{\partial x_k}$, $\partial_{ki} \equiv \frac{\partial^2}{\partial x_k \partial x_i}, \ldots$, $\partial_t \equiv \frac{\partial}{\partial t}$ and use the summation convention, i.e., if the same subscript appears twice in a term, the term denotes the sum over the range of this subscript, e.g., $c_{ik}\partial_k b_j \equiv \sum_{k=1}^{d} c_{ik}\partial_k b_j$, and consequently

$$d_{ij} \equiv \frac{1}{2} \sum_{\ell=1}^{\ell_0} b_i^{\ell} b_j^{\ell} = \frac{1}{2} b_i^{\ell} b_j^{\ell}, \quad \overline{d}_{ij} \equiv \frac{1}{2} \sum_{\ell=1}^{\ell_0} \overline{b}_i^{\ell} \overline{b}_j^{\ell} = \frac{1}{2} \overline{b}_i^{\ell} \overline{b}_j^{\ell}.$$

For a derivative ∂_α the notation $|\alpha|$ is its order.

LEMMA 2.1. (a) *Assume that the following regularity conditions hold:*

(i) *The functions $a(t, x)$, $b(t, x)$, and $c(t, x, z)$ are continuous in (t, x) and are twice continuously differentiable with respect to x.*

(ii) *The partial derivatives of first and second order with respect to x of the functions a and b are uniformly bounded.*

(iii) *The function* $c(t,x,z)$ *satisfies*

(2.3)
$$\int_0^T \int_{\mathbf{z}} \left(|c(t,x,z)| + |\nabla_x c(t,x,z)|^4 + |\nabla_x^2 c(t,x,z)|^2 \right) \lambda(t) \mu(t,dz) dt \leq C,$$

where ∇_x *denotes the gradient with respect to* x*, and the constant* C *is independent of* x*.*

(iv) *The function* g *is twice continuously differentiable, and together with its partial derivatives of first and second order it is uniformly bounded.*

Then the cost to go *function, defined by*

$$u(t,x) = E\big[g(X(T)) \mid X(t) = x\big], \tag{2.4}$$

satisfies the Kolmogorov equation

(2.5)
$$\partial_t u(t,x) + a_k \partial_k u(t,x) + d_{kn} \partial_{kn} u(t,x) + \lambda(t) \int_{\mathbf{z}} [u(t,x+c(t,x,z)) - u(t,x)] \mu(t,dz) = 0,$$

with the boundary condition

$$u(T,\cdot) = g.$$

(b) *Furthermore, if the following regularity conditions are satisfied:*

(i) *the functions* $\partial_\beta a(t,\cdot)$*,* $\partial_\beta b(t,\cdot)$*, and* $\partial_\beta c(t,\cdot,z)$ *are bounded uniformly in* t *and* z *for* $1 \leq |\beta| \leq 8$*;*

(ii) *the jump intensity* $\lambda(t)$ *and the measure* $\mu(t,dz)$ *have continuous time derivatives; in addition, the functions* $a(\cdot,x)$*,* $b(\cdot,x)$*,* $c(\cdot,x,z)$ *have continuous and uniformly bounded first order time derivatives;*

(iii) $\int_0^T \int_{\mathbf{z}} |c(t,0,z)|^{2k} \mu(t,dz) dt < \infty$ *for all* $k \in \mathbb{N}$*;*

(iv) *the function* g *has spatial derivatives* $\partial_\beta g$*, with polynomial growth for* $|\beta| \leq 8$*;*

then the function u *has continuous partial derivatives with respect to* x *up to the order* 8*, satisfying the following polynomial growth condition: for all* $i \in \{0,1,2\}$ *and* $\alpha \in \mathbb{N}^d$ *with* $i + |\alpha| \leq 8$ *there exists* $p_{\alpha,i} \in \mathbb{N}$ *and* $C_{\alpha,i} > 0$ *such that*

$$\max_{0 \leq t \leq T} \left| \partial_t^i \partial_\alpha u(t,x) \right| \leq C_{\alpha,i} \left(1 + |x|^{p_{\alpha,i}} \right) \quad \forall\, x \in \mathbb{R}^d.$$

LEMMA 2.2. *Assume that the regularity conditions described in paragraph* (b) *of Lemma* 2.1 *hold. Then the time discretization error in* (1.9) *satisfies*

$$\begin{aligned} \mathcal{E}_T = & \int_0^T E\Big[\left(a_k(t, \overline{X}(t^-)) - \overline{a}_k(t; \overline{X}) \right) \partial_k u(t, \overline{X}(t^-)) \Big] \, dt \\ & + \int_0^T E\Big[\left(d_{ij}(t, \overline{X}(t^-)) - \overline{d}_{ij}(t; \overline{X}) \right) \partial_{ij} u(t, \overline{X}(t^-)) \Big] \, dt, \end{aligned} \tag{2.6}$$

where $u = u(t,x)$ *is the cost to go function defined in* (2.4)*.*

Remark 2.1. It is possible to relax the assumption of the process departing from a fixed value $x_0 \in \mathbb{R}$. We can assume that it departs from a random variable $X(0)$, independent of the jump structure and the Wiener process, and we can assume that either this initial random variable admits perfect simulation by $\overline{X}(0)$ or it does not. Both situations easily adapt to our approach, giving (in the latter case) an additional term of the form $E[u(0,X(0)) - u(0,\overline{X}(0))]$ in the error expansion in (2.6).

Lemma 2.2 is combined with stochastic flows to derive the a posteriori error expansion in Theorem 2.2 below. This error expansion is based on variations of the processes X and $\overline{X}$. The proof of Theorem 2.2 uses mainly the fact that the error in replacing $g(X(T))$ in Lemma 2.2 by $g(\overline{X}(T))$, in the representation (2.4) of $\partial_\alpha u$, yields the small deterministic remainder term $\int_0^T \mathcal{O}\big((\Delta t)^2\big)dt$ in (2.7) of Theorem 2.2, which is analogous to the $\mathcal{O}(N^{-2})$ term in Talay and Tubaro's expansion, [53], and needs some a priori estimate to be controlled. Lemma 2.2 can be applied to estimate this error. The second important ingredient in the proof is the Markov property of $\overline{X}$ satisfied at the discrete times t_n. Based on the fact that $\overline{X}(t_n)$ is $\mathcal{F}_{t_n}$ measurable, the nested expected values

$$E\big[a_j(t_n, \overline{X}(t_n))\partial_{x_j(t_n)} E[g(\overline{X}(T)) \mid \mathcal{F}_{t_n}]\big]$$

in (2.6) can, by the definition of φ and its implication (6.25), be decoupled to

$$E\big[a_j(t_n, \overline{X}(t_n))\varphi_j(t_n)\big],$$

which reduces the computational complexity substantially; see Lemma 6.3.

THEOREM 2.2 (first error expansion). *Suppose that a, b, g, X, and $\overline{X}$ satisfy the assumptions in Lemma* 2.2. *Then the time discretization error in* (1.9) *has the expansion*

(2.7)
$$\mathcal{E}_T = E\left[\sum_{m=0}^{N-1} \rho(\tilde{t}_m, \omega)(\Delta\tilde{t}_m)^2\right] + E\left[\sum_{m=0}^{N-1} (\Delta\tilde{t}_m)^2 \left\{\mathcal{O}(\Delta\tilde{t}_m) + \sum_{k=m}^{N-1} \mathcal{O}((\Delta\tilde{t}_k)^2)\right\}\right],$$

where the leading order error term is in computable a posteriori form:

(2.8)
$$\begin{aligned}\rho(\tilde{t}_m, \omega) \equiv{}& \frac{1}{2}\sum_{n\in\mathcal{J}_m} \big[\big(a_i(t_{n+1}, \overline{X}(t_{n+1}^-, \omega)) - a_i(t_n, \overline{X}(t_n, \omega))\big)\varphi_i(t_{n+1}^-, \omega)\big]\frac{\Delta t_n}{(\Delta\tilde{t}_m)^2}\\ &+ \frac{1}{2}\sum_{n\in\mathcal{J}_m} \big[\big(d_{ik}(t_{n+1}, \overline{X}(t_{n+1}^-, \omega)) - d_{ik}(t_n, \overline{X}(t_n, \omega))\big)\varphi'_{ik}(t_{n+1}^-, \omega)\big]\frac{\Delta t_n}{(\Delta\tilde{t}_m)^2}\end{aligned}$$

with $\mathcal{J}_m \equiv \{n : \tilde{t}_m \le t_n < \tilde{t}_{m+1}\}$, $m = 0, \ldots, N-1$, and the discrete dual functions $\varphi(t_n) \in \mathbb{R}^d$ and $\varphi'(t_n) \in \mathbb{R}^{d\times d}$ defined in (1.13) *and* (1.14), *respectively.*

Remark 2.3. When implementing a deterministic time stepping algorithm based on this result the expectation of the sum of errors in (2.7) is approximated by the mean of the errors along the M simulated trajectories, i.e.,

$$E\left[\sum_{m=0}^{N-1} \rho(\tilde{t}_m, \omega)(\Delta\tilde{t}_m)^2\right] \approx \frac{1}{M}\sum_{j=1}^{M}\sum_{m=0}^{N-1} \rho(\tilde{t}_m, \omega_j)(\Delta\tilde{t}_m)^2.$$

The statistical error of this approximation can be expressed as

$$E\left[\sum_{m=0}^{N-1} \rho(\tilde{t}_m, \omega)(\Delta\tilde{t}_m)^2\right] - \frac{1}{M}\sum_{j=1}^{M}\sum_{m=0}^{N-1} \rho(\tilde{t}_m, \omega_j)(\Delta\tilde{t}_m)^2 = \int_0^T (I_M + J_M)dt,$$

where the distributions of the statistical errors $\sqrt{M}I_M$ and $\sqrt{M}J_M$ weakly converge to normal distributions with mean zero and time interval–dependent variances given

by

$$\mathrm{Var}\left[\sum_{n\in\mathcal{J}_m}\left(a_i(t_{n+1},\overline{X}(t_{n+1}^-))-a_i(t_n,\overline{X}(t_n))\right)\varphi_i(t_{n+1}^-)\right]=\mathcal{O}(\Delta\tilde{t}_m)$$

and

$$\mathrm{Var}\left[\sum_{n\in\mathcal{J}_m}\left(d_{ik}(t_{n+1},\overline{X}(t_{n+1}^-))-d_{ik}(t_n,\overline{X}(t_n))\right)\varphi'_{ik}(t_{n+1}^-)\right]=\mathcal{O}(\Delta\tilde{t}_m),$$

respectively. Observe that the number of realizations to determine a reliable error estimate is in general TOL^{-1}, much smaller than the—proportional to TOL^{-2}—number of realizations to approximate $E[g(X(T))]$. For more details on this and the statistical approximation of the error density ρ see Remark 2.7 in [51].

3. A second estimate of the time discretization error. In this section we present in Theorem 3.1 an error expansion in a posteriori computable form for the time discretization error $\mathcal{E}_T$ in (1.9) based on the dual functions and its first and second derivatives φ, φ', and φ''. The analysis uses the Malliavin derivative, $\partial_{W(t)}Y$, which is the first variation of a process Y with respect to a perturbation $dW(t)$, at time t of the Wiener process; cf. [46]. As we will see, although computationally more expensive than the expansion in Theorem 2.2, it gives a more precise estimation of the error, measured through its variance.

We shall restrict the analysis to time steps which are constructed by first sampling the jump times to augment an a priori given time discretization $\Delta\tilde{t}$, obtaining $\Delta t[0]$ (see (3.3)). Using the refinement criterion

$$|\rho(t,\omega)|\big(\Delta t(t)\big)^2<\text{constant},\tag{3.1}$$

we construct the partition $\Delta t(t)$ that satisfies

$$\Delta t(t)=\Delta t[0](t)/2^n\ \text{ for some natural number } n=n(t,\omega),$$

with an approximate error density function ρ, satisfying for $s,t\in[0,T]$ and all outcomes ω the uniform upper and lower bounds

$$\begin{aligned}\rho_{low}(\mathrm{TOL})\le|\rho(s,\omega)|\le\rho_{up}(\mathrm{TOL}),\\ |\partial_{W(t)}\rho(s,\omega)|\le\rho_{up}(\mathrm{TOL}).\end{aligned}\tag{3.2}$$

Here ρ_{low} and ρ_{up} are positive functions with $\mathrm{TOL}/\rho_{low}(\mathrm{TOL})\to 0$ as $\mathrm{TOL}\to 0$. For each realization successive subdivisions of the steps yield the largest time steps satisfying (3.1). The corresponding stochastic increments ΔW will have the correct distribution, with the necessary independence, if the increments ΔW related to the new steps are generated by Brownian bridges [30], i.e., the time steps are generated by conditional expected values of the Wiener process.

Let δ be a constant approximating $\frac{\mathrm{TOL}}{E[N]}$, where $E[N]$ is the expected number of steps. The analysis in this section with adaptive nonadapted time steps, satisfying (3.1)–(3.2), is based on the *stochastic time stepping algorithm* described below.

In what follows, Lemma 3.1 and Theorem 3.1 show that although the steps generated by (3.1)–(3.2) are not adapted, the method indeed converges to the correct limit, which is the same as the limit of the forward Euler method with adapted time steps.

LEMMA 3.1 (strong convergence). *Suppose that a, b, c, g, X satisfy the assumptions in Lemma 2.2, that $\overline{X}$ is constructed by the forward Euler method, based on the stochastic time stepping algorithm above, with step sizes Δt_n satisfying (3.1)–(3.2), and that their corresponding ΔW_n are generated by Brownian bridges. Then*

$$\sup_{0\le t\le T}\sqrt{E[|X(t)-\overline{X}(t)|^2]} = \mathcal{O}\left(\sqrt{\Delta t_{\sup}}\right) = \mathcal{O}\left(\sqrt{\frac{\mathrm{TOL}}{\rho_{low}(\mathrm{TOL})}}\right) \to 0$$

as TOL $\to 0$, *where* $\Delta t_{\sup} \equiv \sup_{n,\omega} \Delta t_n(\omega)$.

STOCHASTIC TIME STEPPING ALGORITHM.

Do for M realizations $\omega_j,\ j = 1, \dots, M$:

Sample jump times: $0 < \tau_1(\omega_j) < \cdots < \tau_{\nu(\omega_j)}(\omega_j) < T$.

Consider the initial augmented partition given by the union

$$\Delta t[0] = \{t_n(\omega_j)\}_{n=0}^{\kappa(\omega_j)} = \{\tilde{t}_n\}_{n=0}^{N} \cup \{\tau_n(\omega_j)\}_{n=0}^{\nu(\omega_j)} \tag{3.3}$$

that has $\kappa(\omega_j) = N + \nu(\omega_j)$ (a.s.) time steps.

`STEP 1`: Set $k = 0$. Start with the initial coarse mesh $\Delta t[0]$ and compute $\Delta W[0]$.

`STEP 2`: For the piecewise constant mesh function $\Delta t[k]$ with corresponding noise $\Delta W[k]$, compute $\overline{X}[k]$ and the weight function $\rho[k]$ defined in (3.4)–(3.5).

`STEP 3`: Define $r(t) \equiv |\rho[k](t)|(\Delta t[k](t))^2$ and let for all t

$$\Delta t[k+1](t) = \begin{cases} \Delta t[k](t) & \text{if } r(t) < \delta, \quad (\dagger) \\ \Delta t[k](t)/2 & \text{if } r(t) \ge \delta, \quad (\star) \end{cases}$$

and in the refinement case $(\star)$ construct $\Delta W[k+1]$ by Brownian bridges based on the already known $\Delta W[k]$.

`STEP 4`: **If** at least one step of $\Delta t[k]$ is refined by $(\star)$, increment k by 1 and **goto** Step 2. **Else** all steps of $\Delta t[k]$ satisfy $(\dagger)$; thus accept the approximation $g(\overline{X}(T, \omega_j))$ and **goto** the next realization of p and W.

Enddo

If the statistical error, $E[g(\overline{X}(T))] - \frac{1}{M}\sum_{j=1}^{M} g(\overline{X}(T, \omega_j))$, is sufficiently small **stop**, **else** restart with a larger M.

Endif

For a detailed description of the stochastic time stepping algorithm see section 4.4.

Proof of Lemma 3.1. The proof relies on the two-step simulation scheme that we construct and on the fact that the jump nodes do not contribute to the global error. Observe that $\mathcal{F}^J$ is the σ-algebra generated by the jumps and marks constructed in section 1.2.1. Consider the conditional expectation $E[|X(t) - \overline{X}(t)|^2 \mid \mathcal{F}^J]$ and apply Lemma 3.1 from [51], using also that there is no time discretization error at the jump nodes. ∎

In addition to the dual functions φ and φ' in Theorem 2.2, the new error expansion for stochastic time steps in Theorem 3.1 below also uses, for $t = t_{n+1}^-$, the discrete dual variation $\varphi''(t)$ defined in (1.16). It derives an error estimate applicable both to adaptive deterministic time steps and to the stochastic time stepping algorithm; the assumptions and the proof of the theorem focus on stochastic steps; however, a modification to deterministic time steps is straightforward. The computable error density $\tilde{\rho}$ of this error estimate can then be cut off for small and large values to satisfy (3.2). This truncation procedure is explained in formulas (4.6) and (4.15).

THEOREM 3.1 (second error expansion). *Suppose that the functions a, b, c, g, and the process X satisfy the assumptions in Lemma 2.2 with $X(0)$ deterministic. Let $\overline{X}$ be constructed by the forward Euler method with step sizes Δt_n satisfying (3.1)–(3.2) and let the corresponding ΔW_n be generated by Brownian bridges, following the stochastic time stepping algorithm in Lemma 3.1. Then the time discretization error defined in (1.9) has the following expansion, based on both the drift and diffusion fluxes and the discrete dual functions φ, φ', and φ'' given in (1.13)–(1.16), with computable leading order terms:*

(3.4)
$$\mathcal{E}_T = E\left[\sum_{n=0}^{\kappa-1} \tilde{\rho}(t_n,\omega)(\Delta t_n)^2\right] + \mathcal{O}\left(\Big(\frac{\mathrm{TOL}}{\rho_{low}(\mathrm{TOL})}\Big)^{1/2}\Big(\frac{\rho_{up}(\mathrm{TOL})}{\rho_{low}(\mathrm{TOL})}\Big)^{\epsilon}\right) E\left[\sum_{n=0}^{\kappa-1}(\Delta t_n)^2\right],$$

for any $\epsilon > 0$ and where

$$\begin{aligned}\tilde{\rho}(t_n,\omega) \equiv \frac{1}{2}\Big(&\big(\partial_t a_k + \partial_j a_k a_j + \partial_{ij} a_k d_{ij}\big)\varphi_k(t_{n+1}^-)\\ &+ \big(\partial_t d_{km} + \partial_j d_{km} a_j + \partial_{ij} d_{km} d_{ij} + 2\partial_j a_k d_{jm}\big)\varphi'_{km}(t_{n+1}^-)\\ &+ \big(2\partial_j d_{km} d_{jr}\big)\varphi''_{kmr}(t_{n+1}^-)\Big)\end{aligned} \tag{3.5}$$

and the terms in the sum of (3.5) are evaluated at the a posteriori known points $(t_n, \overline{X}(t_n))$, i.e.,

$$\partial_\alpha a \equiv \partial_\alpha a(t_n, \overline{X}(t_n)), \quad \partial_\alpha b \equiv \partial_\alpha b(t_n, \overline{X}(t_n)), \quad \partial_\alpha d \equiv \partial_\alpha d(t_n, \overline{X}(t_n)).$$

Observe that the constant in $\mathcal{O}$ that appears in (3.4) may not be uniform with respect to the value ϵ. Thus, in practice one chooses $\epsilon = \epsilon(\mathrm{TOL})$ to minimize the contribution of the remainder term to the error expansion (3.4).

4. Adaptive time stepping algorithms. Here, we describe two adaptive time stepping algorithms for the weak approximation problem based on the approximation error formula (3.4). They are very similar to those introduced in [45]. Algorithm D augments a mesh that is the same for all realizations with jump times that are sampled differently for each realization; its adaptive strategy is based on averaged information from the a posteriori error formula. On the other hand, Algorithm S can adapt the mesh differently for each realization. Both adaptive algorithms choose adaptively the number of realizations and the size of time steps to efficiently bound the approximation error by a prescribed error tolerance.

4.1. Computational error splitting. As we described in section 1.3 the weak approximation computational error $\mathcal{E}$ of the jump augmented Euler–Maruyama scheme naturally splits into time discretization and statistical errors, namely, $\mathcal{E} = \mathcal{E}_T + \mathcal{E}_S$. Thus, the control of the computational error is related to the combined control of the time discretization error via the choice of the time steps Δt, and of the statistical error via the choice of the number M of realizations. Therefore, we split a given computational error tolerance, $\mathrm{TOL} > 0$, into a statistical error tolerance TOL_S and a time discretization error tolerance TOL_T (see [51], [45]) by

$$\mathrm{TOL}_T = \frac{1}{3}\,\mathrm{TOL} \quad \text{and} \quad \mathrm{TOL}_S = \frac{2}{3}\mathrm{TOL}. \tag{4.1}$$

4.2. Control of the statistical error. For M independent samples $\{Y(\omega_j)\}_{j=1}^M$ of a random variable Y, with $E\left[\,|Y|^6\,\right] < \infty$, define the sample average by

$$\mathcal{A}(Y;M) \equiv \frac{1}{M}\sum_{j=1}^{M} Y(\omega_j)$$

and the sample standard deviation by

$$\mathcal{S}(Y;M) \equiv \left\{\mathcal{A}(Y^2;M) - (\mathcal{A}(Y;M))^2\right\}^{\frac{1}{2}}.$$

Let $\sigma_Y \equiv \left\{E\left[\left|Y - E[Y]\right|^2\right]\right\}^{\frac{1}{2}}$ and $Z_M \equiv \frac{\sqrt{M}}{\sigma_Y}\left(\mathcal{A}(Y;M) - E[Y]\right)$ with cumulative distribution function $F_{Z_M}(x) \equiv P(Z_M \le x),\ x \in \mathbb{R}$. Let $\gamma \equiv (E[|Y - E[Y]|^3])^{1/3}/\sigma_Y < \infty$. The Berry–Essen theorem [14] gives the following estimation in the central limit theorem:

$$\sup_{x\in\mathbb{R}} |F_{Z_M}(x) - \mathcal{N}(x)| \le \frac{3}{\sqrt{M}}\gamma^3$$

for the rate of convergence of F_{Z_M} to the distribution function $\mathcal{N}$ of a normal random variable with mean zero and variance one, i.e.,

$$\mathcal{N}(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}s^2\right) ds. \tag{4.2}$$

Provided M is sufficiently large, i.e., $M \gg 36\gamma^6$, the statistical error

$$\mathcal{E}_S(Y;M) \equiv E[Y] - \mathcal{A}(Y;M)$$

satisfies, by the Berry–Essen theorem, the probability approximation

$$P\left(\left[|\mathcal{E}_S(Y;M)| \le c_0\frac{\sigma_Y}{\sqrt{M}}\right]\right) \approx 2\mathcal{N}(c_0) - 1.$$

In practice choose some constant $c_0 \ge 1.65$, so the normal distribution satisfies $1 > 2\mathcal{N}(c_0) - 1 \ge 0.901$ and the event

$$|\mathcal{E}_S(Y;M)| \le \mathtt{E}_S(Y;M) \equiv c_0\ \frac{\mathcal{S}(Y;M)}{\sqrt{M}} \tag{4.3}$$

has probability close to one, which involves the additional step to approximate σ_Y by $\mathcal{S}(Y;M)$; cf. [18]. Thus, in the computations $\mathtt{E}_S(Y;M)$ is a good approximation of the statistical error $\mathcal{E}_S(Y;M)$.

For a given $\mathrm{TOL}_S > 0$, the goal is to find M such that $\mathtt{E}_S(Y;M) \le \mathrm{TOL}_S$. The following algorithm adaptively finds the number of realizations M to compute the sample average $\mathcal{A}(Y;M)$ as an approximation to $E[Y]$. With probability close to one, depending on c_0, the statistical error in the approximation is then bounded by TOL_S.

`routine Monte-Carlo`(TOL_S, M_0; EY)
Set the batch counter $m = 1$, $M[m] = M_0$ and $\mathtt{E}_S[m] = +\infty$.
Do while ($\mathtt{E}_S[m] > \mathrm{TOL}_S$)

(4.4)
Compute $M[m]$ new samples of Y, along with the sample average $EY \equiv \mathcal{A}(Y;M[m])$, the sample standard deviation $\mathcal{S}[m] \equiv \mathcal{S}(Y;M[m])$ and the statistical error estimation $\mathtt{E}_S[m+1] \equiv \mathtt{E}_S(Y,M[m])$. Compute $M[m+1]$ by `change_`M ($M[m]$, $\mathcal{S}[m]$, TOL_S; $M[m+1]$).
Increase m by 1.

end-do
`end of Monte-Carlo`

`routine change_`M $(M_{in}, \mathcal{S}_{in}, \mathrm{TOL}_S; M_{out})$

$$\begin{aligned} M^* &= \min\Big\{\text{integer part } \Big(\tfrac{c_0\,\mathcal{S}_{in}}{\mathrm{TOL}_S}\Big)^2,\ \mathrm{MCH}\times M_{in}\Big\},\\ n &= \text{integer part } (\log_2 M^*) + 1,\\ M_{out} &= 2^n. \end{aligned} \tag{4.5}$$

`end of change_`M

Here, M_0 is a given initial value for M, and $\mathrm{MCH} > 1$ is a positive integer parameter introduced to avoid a large new number of realizations in the next batch due to a possibly inaccurate sample standard deviation $\mathcal{S}[m]$. Indeed, $M[m+1]$ cannot be greater than $\mathrm{MCH}\times M[m]$.

4.3. Deterministic time stepping algorithm. Following closely [40] and [45], we present an adaptive algorithm based on the cut-off of the error density $\tilde{\rho}$ in (3.5) of Theorem 3.1 which is defined as

$$\rho_D^n \equiv \min\left(\max\left(\left|\frac{1}{(\Delta\tilde{t}_n)^2}\sum_{\ell\in\mathcal{J}_n}(\Delta t_\ell)^2\tilde{\rho}(t_\ell,\overline{X})\right|, \mathrm{TOL}^{\frac{1}{9}}\right), \mathrm{TOL}^{-1}\right), \quad n=1,\dots,N. \tag{4.6}$$

The error expansion in Theorem 3.1 motivates us to approximate the time discretization error by

$$|\mathcal{E}_T| \lesssim E\left[\sum_{n=1}^{N} r_n\right], \tag{4.7}$$

where the error indicator, r_n, is defined by

$$r_n \equiv \rho_D^n\,(\Delta\tilde{t}_n)^2, \quad n=1,\dots,N. \tag{4.8}$$

The main advantage of the deterministic time stepping algorithm over the stochastic time stepping algorithm is that the number M_T of realizations necessary to determine the optimal deterministic stepping scheme is considerably smaller than M, whereas, in the stochastic time stepping algorithm, a refinement of the partition is carried out in each one of the M trajectories, leading to a considerably larger amount of computational work.

As pointed out before, the error expansion derived in Theorem 3.1 is also valid with deterministic time steps. Therefore, we have some flexibility in the choice of error densities for the adaptive algorithm with deterministic time steps because we can also use the results from Theorem 2.2. There are some practical differences to mention. On the one hand, the variance of the averaged error density from (2.8) is $\mathcal{O}(\frac{1}{\Delta\tilde{t}_n M_T})$. This feature has been observed in [42] and [51], where a local filtering procedure was proposed to reduce the variance of the error density estimator. A positive feature of this error density is that it does not require the computation of the second variation, φ'', which may be computationally expensive for large d. On the other hand, the averaged error density from (3.5) has a much smaller variance $\mathcal{O}(\frac{1}{M_T})$

which does not need filtering but requires the computation of φ''. In this work, based on the comments above and on previous experience (mainly [51]), we do not discuss further this choice and only show numerical results with adaptive deterministic time steps based on the error density (3.5).

The approximation of the time discretization error in the right-hand side of (4.7) can be separated into two parts,

$$E\left[\sum_{n=1}^{N} r_n\right] \leq \mathcal{A}\left(\sum_{n=1}^{N} r_n; M_T\right) + \left|E\left[\sum_{n=1}^{N} r_n\right] - \mathcal{A}\left(\sum_{n=1}^{N} r_n; M_T\right)\right|, \tag{4.9}$$

where the second error term in the right-hand side of (4.9) is with probability close to one asymptotically bounded by

$$\left|E\left[\sum_{n=1}^{N} r_n\right] - \mathcal{A}\left(\sum_{n=1}^{N} r_n; M_T\right)\right| \lesssim \ \mathtt{E}_{TS} \ \equiv c_0 \frac{\mathcal{S}\left(\sum_{n=1}^{N} r_n; M_T\right)}{\sqrt{M_T}} \tag{4.10}$$

and the first term defines $\mathtt{E}_{TT} \equiv \mathcal{A}(\sum_{n=1}^{N} r_n; M_T)$. Then for a given $\mathrm{TOL}_T > 0$, the goal is to construct a partition $\Delta\tilde{t}$ of $[0,T]$ with as few time steps and realizations M_T as possible, such that $\mathtt{E}_{TT} + \mathtt{E}_{TS} \leq \mathrm{TOL}_T$. To this end, first split the time discretization tolerance TOL_T into two positive parts, TOL_{TT} and TOL_{TS} for $\mathtt{E}_{TT}$ and $\mathtt{E}_{TS}$, respectively. The statistical error of the time discretization using the density (4.6) is $\mathcal{O}(\frac{\Delta t_{\sup}}{\sqrt{M_T}})$. Therefore the percentage of the tolerance, TOL, devoted to the control of the statistical time discretization error can be arbitrary small as $\Delta t_{\sup} \to 0$. In practice we choose

$$\mathrm{TOL}_{TT} = \frac{2}{3}\mathrm{TOL}_T = \frac{2}{9}\mathrm{TOL}, \qquad \mathrm{TOL}_{TS} = \frac{1}{3}\mathrm{TOL}_T = \frac{1}{9}\mathrm{TOL}. \tag{4.11}$$

The control of the statistical time discretization error determines the number of realizations M_T necessary to ensure a reliable choice of the time discretization in the deterministic time stepping algorithm. Similarly as in [45], here it is optimal to equidistribute the error contributions from different time intervals. Thus, the goal of the adaptive algorithm described below is to construct a deterministic time partition $\Delta\tilde{t}$ of $[0,T]$ such that

$$\bar{r}_n \equiv \mathcal{A}(r_n; M_T) \leq d_1 \frac{\mathrm{TOL}_{TT}}{N}, \quad n = 1, \ldots, N, \tag{4.12}$$

where $d_1 = 2$; see [45, Remark 3.9].

To achieve (4.12), start with an initial partition $\Delta\tilde{t}[1]$ and then specify iteratively a new partition $\Delta\tilde{t}[k+1]$, from $\Delta\tilde{t}[k]$, using the refinement strategy

for $n = 1, 2, \ldots, N[k]$
if $\bar{r}_n[k] \geq d_1 \frac{\mathrm{TOL}_{TT}}{N[k]}$, **then** divide $\Delta\tilde{t}_n[k]$ into 2 uniform substeps. (4.13)
else let the new step be the same as the old
endif
endfor

until the following stopping criteria is satisfied:

$$\textbf{if } \left(\max_{1 \leq n \leq N[k]} \bar{r}_n[k] < D_1 \frac{\mathrm{TOL}_{TT}}{N[k]}\right) \textbf{ then stop.} \tag{4.14}$$

Here D_1 is a given constant satisfying $D_1 > \frac{2}{\hat{c}} d_1$, where $\hat{c} \approx 1/2$; see [45]. The combination of (4.9) and (4.14) asymptotically guarantees a given level of accuracy, $\mathtt{E}_{TT} < D_1 \mathrm{TOL}_{TT}$. The positive number D_1 is introduced to avoid slow convergence in case almost all $\bar{r}_n$ satisfy (4.14), as in section 4.4.

Now we are ready for the detailed definition of the adaptive algorithm with deterministic steps.

ALGORITHM D.

Initialization Choose:

1. an error tolerance, $\mathrm{TOL} \equiv \mathrm{TOL}_S + \mathrm{TOL}_{TT} + \mathrm{TOL}_{TS}$,
2. a number $N[1]$ of initial uniform steps $\Delta\tilde{t}[1]$ for $[0, T]$,
3. a number $M[1]$ of initial realizations and set $M_T[1] = M[1]$,
4. a number $d_1 = 2$ in (4.13) and $\hat{c} = 1/2 + 1/20$ to compute D_1 using $D_1 > \frac{2}{\hat{c}} d_1$,
5. a constant $c_0 \geq 1.65$ and an integer $\mathrm{MCH} \geq 2$ to determine the number of realizations in (4.5), and
6. a constant $\epsilon = 1/2$ used to control the statistical error in the time discretization error estimate.

Set the iteration counter, k, for time refinement levels, to 1 and set the statistical error, $\mathtt{E}_{TS} = +\infty$ and $\bar{r}[k] = +\infty$.

Do while ($\bar{r}[k]$ violates the stopping criterion (4.14) or $\mathtt{E}_{TS} > \mathrm{TOL}_{TS}$)

Compute the sample averages and the error estimates on $\Delta\tilde{t}[k]$
by calling `Euler`. Set $M_T[k+1] = M_T[k]$ and $\Delta\tilde{t}[k+1] = \Delta\tilde{t}[k]$.

Set $\widetilde{\mathrm{TOL}}_{TS,\epsilon} = \max(\mathrm{TOL}_{TS}, \epsilon\, \mathtt{E}_{TT})$.

If $(\mathtt{E}_{TS} > \widetilde{\mathrm{TOL}}_{TS,\epsilon})$

Update $M_T[k+1]$ by `change_M` $(M_T[k], \mathcal{S}_{TS}[k], \widetilde{\mathrm{TOL}}_{TS,\epsilon}; M_T[k+1])$.

elseif ($\bar{r}[k]$ violates the stopping criterion (4.14))

For all time steps $i = 1, \ldots, N[k]$, do the refinement process (4.13)
to update $\Delta\tilde{t}[k+1]$ from $\Delta\tilde{t}[k]$.

end-if

Increase k by 1.

end-do

Compute an approximation, Eg, for $E[g(\overline{X}(T))]$ with fixed time mesh $\Delta\tilde{t} = \Delta\tilde{t}[k]$ by `Monte-Carlo`$(\mathrm{TOL}_S,\ M_T[k];\ Eg)$ in (4.4).

Accept Eg as an approximation of $E[g(X(T))]$, since the estimate of the computational error is bounded by TOL.

`routine Euler`

For each $M_{\tau}[k]$ new realizations, sample jump times with their corresponding marks.

As described in section 1.2, use $\Delta\tilde{t}[k]$ to compute corresponding realizations of the Euler method.

Update the approximations of the time discretization error indicators $\bar{r}[k]$ and the statistical time discretization error $\mathtt{E}_{TS}[k]$ and compute the sample standard deviation $\mathcal{S}_{TS}[k] \equiv \mathcal{S}(g(\overline{X}(T)); M_T[k])$.

`end of Euler`

4.4. The stochastic time stepping algorithm. Now we describe an adaptive algorithm with stochastic time steps based on a cut-off of the error density $\tilde{\rho}$ introduced in (3.5) of Theorem 3.1, ρ_s, defined as

$$\rho_s^n \equiv \min\left(\max\left(|\tilde{\rho}(t_n, \overline{X})|, \mathrm{TOL}^{\frac{1}{9}}\right), \mathrm{TOL}^{-1}\right), \quad n = 1, \ldots, \kappa. \tag{4.15}$$

Following the error expansion in Theorem 3.1, the time discretization error is approximated by

$$|\mathcal{E}_T| \lesssim E\left[\sum_{n=1}^{\kappa} r_n\right], \tag{4.16}$$

where the error indicator, r_n, is defined by

$$r_n \equiv \rho_S^n \left(\Delta t_n\right)^2, \quad n = 1, \ldots, \kappa. \tag{4.17}$$

In this case it is optimal (cf. [45]) to equidistribute the error contributions among all time steps and all realizations. In other words, the goal of the adaptive algorithm is to construct a time partition Δt of $[0, T]$ for each realization such that

$$r_n \leq s_1 \frac{\mathrm{TOL}_T}{E[\kappa]}, \quad n = 1, \ldots, \kappa, \tag{4.18}$$

where $s_1 = 2$; see Remark 3.1 in [45]. Note that in practice the quantity $E[\kappa]$ is not known and we can only estimate it by a sample average $\mathcal{A}(\kappa; M)$ from the previous batch of realizations. The statistical error $|E[\kappa] - \mathcal{A}(\kappa; M)|$ is then bounded by $\mathtt{E}_S(\kappa; M)$, with probability close to one, by the same argument as in (4.3).

Let $\bar{\kappa}[j] \equiv \mathcal{A}(\kappa; M[j])$ be the sample average of the final number of time steps in the jth batch of $M[j]$ realizations. To achieve (4.18) for each realization, start with an initial partition $\Delta t[1]$ and then specify iteratively a new partition $\Delta t[k+1]$, from $\Delta t[k]$, using the following refinement strategy:

for each realization in the jth batch
for each time step $n = 1, \ldots, \kappa[k]$
if $r_n[k] \geq s_1 \frac{\mathrm{TOL}_T}{\bar{\kappa}[j-1]}$, **then** divide $\Delta t_n[k]$ into 2 uniform substeps.
else let the new step be the same as the old
endif
endfor
endfor (4.19)

The refinement strategy (4.19) motivates the following stopping criteria: for each realization of the jth batch

$$\textbf{if } \left(\max_{1 \leq n \leq \kappa[k]} r_n[k] < S_1 \frac{\mathrm{TOL}_T}{\bar{\kappa}[j-1]}\right) \quad \textbf{then} \quad \text{stop}, \tag{4.20}$$

where $S_1 > \frac{2}{\hat{c}}\, s_1$ with $\hat{c} \approx \frac{1}{2}$; see [45].

Now we are ready for the detailed definition of the adaptive algorithm with stochastic steps.

ALGORITHM S.

Initialization Choose:

1. an error tolerance, $\mathrm{TOL} \equiv \mathrm{TOL}_S + \mathrm{TOL}_T$,
2. a number $N[1]$ of initial uniform steps $\Delta t[1]$ for $[0, T]$ and set $\bar{\kappa} = N[1]$,
3. a number $M[1]$ of initial realizations,
4. a number $s_1 = 2$ in (4.19) and $\hat{c} = \frac{1}{2} + \frac{1}{20}$ to compute S_1 using $S_1 > \frac{2}{\hat{c}}\, s_1$, and

5. a constant $c_0 \geq 1.65$ and an integer MCH≥ 2 to determine the number of realizations in (4.5).

Set the iteration counter for batches $m = 1$ and the stochastic error $\mathtt{E}_S[m] = +\infty$.
Do while ($\mathtt{E}_S[m] > \mathrm{TOL}_S$)
  For realizations $j = 1, \ldots, M[m]$
    Set $k = 1$ and $r[k] = +\infty$.
    Generate the jump times and their marks $(\tau, Z) = \{(\tau_\ell, Z_\ell)\}_{\ell=1}^{\nu}$.
    Start with the initial partition $\Delta t[k]$ and generate $\Delta W[k]$.
    Compute $g(\overline{X}(T))[J]$ and $N[J]$ by `routine Control-Time-Error`.
  end-for
  Compute the sample average $Eg \equiv \mathcal{A}\left(g(\overline{X}(T)); M[m]\right)$, the sample standard deviation $\mathcal{S}[m] \equiv \mathcal{S}(g(\overline{X}(T)); M[m])$ and the a posteriori bound for the statistical error $\mathtt{E}_S[m] \equiv \mathtt{E}_S(g(\overline{X}(T)), M[m])$ in (4.3).
  if ($\mathtt{E}_S[m] > \mathrm{TOL}_S$)
    Compute $M[m+1]$ by `change_`$M(M[m], \mathcal{S}[m], \mathrm{TOL}_S; M[m+1])$, (see (4.5)) and update $\bar{\kappa} = \mathcal{A}(\kappa[J]; M[m])$, where the random variable $\kappa[J]$ is the final number of time steps on each realization.
  end-if
  Increase m by 1.
end-do
Accept Eg as an approximation of $E[g(X(T))]$, since the estimate of the computational error is bounded by TOL.

`routine Control-Time-Error`$(\Delta t[k], \Delta W[k], r[k], (\tau, Z); g(\overline{X}(T))[J], N[J])$
  Do while ($r[k]$ violates the stopping criterion (4.20))
    Compute the Euler approximation $\overline{X}[k]$ in section 1.2 and the error indicator $r[k]$ in (4.8) using the error density (4.15) on $\Delta t[k]$ with the known Wiener increments $\Delta W[k]$.
    If ($r[k]$ violates the stopping criterion (4.20))
      For time steps $i = 1, \ldots, N[k]$
        Do the refinement process (4.19) to compute $\Delta t[k+1]$ from $\Delta t[k]$ and compute $\Delta W[k+1]$ from $\Delta W[k]$ using Brownian bridges.
      end-for
    end-if
    Increase k by 1.
  end-do
  Set the number of the final level $J = k - 1$.
`end of Control-Time-Error`

Remark 4.1 (computational limitations of the jump augmented Euler scheme). Since the Euler method has weak order one, the number of discretization points needed to control the time discretization error is $\mathcal{O}(1/\mathrm{TOL})$. Therefore, as long as the product $\Lambda(T)\mathrm{TOL} \ll 1$, we can still use the jump augmented Euler method to solve the weak approximation problem with an amount of computational work that is similar to that required to solve the same problem without jumps. We still find it important to emphasize that our error approximation theory does not rely on the assumption $\Lambda(T)\mathrm{TOL} \ll 1$; the key point is that it may become too costly to apply it if this condition does not hold. Indeed, if $\Lambda(T)\mathrm{TOL} \approx 1$ (or larger), one can suggest (for overall computational efficiency reasons) using another time discretization which does not include jump times in the computational grid, for instance, one related to those discussed in [22] and [33].

5. Numerical experiments. This section shows numerical results from the implementation of the a posteriori error approximation formula presented in section 2 and of the adaptive algorithms described in section 4. Our implementation uses double-precision FORTRAN 77 and is an extension of the code written for the numerical experiments in [51]. For the numerical simulation of the uniform distribution $\mathcal{U}(0,1)$ and the normal distribution $\mathcal{N}(0,1)$, we implemented a double-precision modification of the functions `ran1` and `gasdev` proposed in [48], with a user-provided initial seed which must be a negative integer. In particular, we use three different seeds for the independent sources of randomness in our equations, namely, `iseed` for the simulation of the Wiener process increments, `zseed` for the simulation of the jump marks, and `tseed` for the simulation of the jump times.

Example 1. Consider the stochastic differential equation (1.1) with $d = 2$, $\ell_0 = \ell_1 = 1$,

$$a(t,x) = \left(-x_2,\, x_1 + \frac{1}{2}\,\lambda(t)\,x_2 \right), \quad b^1(t,x) = \left(\sqrt{\frac{\lambda(t)}{1+t}}\,\sin(x_1),\, 0 \right),$$
$$c(t,x,z) = \left(0,\, z\,\frac{\cos(x_1)}{\sqrt{t+1}} - x_2 \right)$$

and a time-dependent intensity $\lambda(t) = (1+t)^{-1}$. The distribution for the jump marks is time dependent and such that $E[Z_k^2] = 1$. In particular, we use $Z_k = \cos(2\pi\tau_k) + \sin(2\pi\tau_k)\,2\sqrt{3}\,(U_k - \frac{1}{2})$, where $\{\tau_k\}$ are the jumps constructed in section 1.2.1 and $\{U_k\}$ is a sequence of $\mathcal{U}(0,1)$ i.i.d. random variables. In this case, the inverse function Λ^{-1} is given explicitly by $\Lambda^{-1}(s) = \exp(s) - 1$. This example is a generalization of Example 5.1 in [33]. Here, we have both time-dependent jump intensity and mark distribution for the underlying Poisson process. Taking $g(x) = |x|^2$, $T = 1$, and $X(0) = (0,0)$, the exact solution of the corresponding weak approximation problem is given by the formula

$$E\big[g(X(T))\big] = |X(0)|^2 + \int_0^T \frac{\lambda(s)}{1+s}\,ds = 1 - \frac{1}{1+T} = \frac{1}{2}.$$

The value of the parameters needed in the simulations are MCH $= 10$, $c_0 = 1.65$, `iseed` $= -7$, `zseed` $= -101$, and `tseed` $= -20$. We note that the expected value of the number of jump points equals $\Lambda(T) = \Lambda(1) = \log(2) \approx 0.693$. Therefore, the computational cost of including the jump times of the process X in our discretization is fairly low; see the statistic for the sampled number of jumps, ν. In fact, the extra work to simulate the exact jump times of the process is asymptotically negligible as the required accuracy, TOL, tends to zero.

Deterministic time stepping algorithm. First, we perform overkilling runs in order to test how realistic the a posteriori error approximation of the time discretization error described in Theorem 2.2 is. The results, shown in Table 5.1, confirm that the ratio of the computational error and its computable approximation tends to 1 as the number of uniform time steps N increases. For each value of N, we choose the number of realizations M large enough in order to keep the total statistical error at the level of 1% of the size of the obtained approximation of the time discretization error. Then Table 5.2 contains the results of the Algorithm D with an adaptive choice of both the deterministic time steps and the number of realizations. The program starts with $M = 100$ and $N = 5$ subintervals as an initial uniform partition of the time interval $[0,T] = [0,1]$. The tolerance TOL $= 0.02$ is divided into $\text{TOL}_S = 0.01333$,

TABLE 5.1
Computing the efficiency index of the Algorithm D in Example 1.

	$c_0 = 1.65$	$\texttt{iseed} = -7$	$A = \frac{\mathtt{E}_{TT} - \mathtt{E}_S - \mathtt{E}_{TS}}{\mathcal{E}}$	$B = \frac{\mathtt{E}_{TT} + \mathtt{E}_S + \mathtt{E}_{TS}}{\mathcal{E}}$
N	M	$\mathtt{E}_{TT}$	$\mathtt{E}_S + \mathtt{E}_{TS}$	$[\min\{A,B\}, \quad \max\{A,B\}]$
5	10×10^6	-6.02×10^{-2}	5.87×10^{-4}	[1.026, 1.046]
10	50×10^6	-3.14×10^{-2}	2.33×10^{-4}	[1.019, 1.035]
20	100×10^6	-1.59×10^{-2}	1.54×10^{-4}	[1.008, 1.028]

TABLE 5.2
Adaptive choice of M and Δt with Algorithm D when $\mathrm{TOL} = 0.02$ *in Example* 1.

Iter.	N	M	$\mathcal{E}$	$\mathtt{E}_{TT}$	$\mathtt{E}_{TS}$	$\mathtt{E}_S$
1	5	100	-2.66×10^{-2}	-5.22×10^{-2}	1.82×10^{-2}	1.44×10^{-1}
2	10	100	-9.19×10^{-3}	-2.48×10^{-2}	8.74×10^{-3}	1.35×10^{-1}
3	20	100	9.16×10^{-2}	-1.16×10^{-2}	4.25×10^{-3}	1.30×10^{-1}
4	20	1000	-1.78×10^{-2}	–	–	4.54×10^{-2}
5	20	10000	-1.77×10^{-2}	–	–	1.50×10^{-2}
6	20	14088	-1.29×10^{-2}	–	–	1.23×10^{-2}

$\mathrm{TOL}_T = 0.00444$, and $\mathrm{TOL}_{TS} = 0.00222$. When the algorithm stops, the size of the total approximation error is less than 2TOL according to the stopping criterion (4.14), and it agrees with the size of the computational error.

Stochastic time stepping algorithm. To observe the performance of the stochastic time steps Algorithm S, we apply it for different values of TOL, starting with a number of realizations $M = 100$ and a number of uniform time steps $N = 5$. Table 5.3 contains the obtained results which show that Algorithm S is also effective in giving us an approximation of the quantity of interest within the margin of 2TOL due to the criterion (4.14).

Example 2. The second example motivates the use of adaptivity by applying Algorithm D to a linear one-dimensional ($d = 1$) stochastic differential equation with time-dependent coefficients. It is interesting to observe that although here the evolution of X has many jumps, computing with Algorithm D based on jump augmented grids is still substantially more economical than doing the time stepping on a uniform mesh. The goal of the computation is given by the function $g(x) = x$, the initial condition $x_0 = 1$, and the final time $T = 1$. The coefficients of the equation have the form

$$a(t,x) = a(t)x, \quad b(t,x) = \frac{x}{\sqrt{2}}, \quad c(t,x,z) = c_* xz,$$

where $c_* > 0$ is a given constant defined below. The compensator of the jump measure is given by $q(dt,dz) = \lambda(t)\mu(dz)dt$, where $\mu(dz)$ is the uniform distribution in the interval $[-\sqrt{3},\sqrt{3}]$, and the jump intensity exhibits two different regimes,

$$\lambda(t) = \begin{cases} 0 & \text{if } 0 \le t \le 2/3, \\ 3N_J & \text{if } 2/3 < t \le 1, \end{cases}$$

with $N_J = 32^2 = 1024$. Thus, we expect in the mean $N_J = 1024$ jumps per realization. The time varying drift function is

$$a(t) = \begin{cases} 0 & \text{if } t < 1/3, \\ \frac{1}{2\sqrt{t-1/3+\mathrm{TOL}^4}} & \text{if } 1/3 \le t \le 1. \end{cases}$$

TABLE 5.3
Adaptive runs with Algorithm S corresponding to different required accuracies in Example 1.

TOL	M	$\mathcal{A}(\kappa;M)$	$[\kappa_{\min},\kappa_{\max}]$	$\mathcal{S}(\kappa;M)$	$\nu_{\max}$	E_S	$\mathcal{E}$
0.040	3.2×10^3	8.3	[5, 41]	5.2	6	0.0250	−0.002
0.020	12×10^3	10.9	[5, 73]	9.2	6	0.0130	−0.006
0.010	48×10^3	20.5	[10, 146]	16.9	6	0.0066	−0.010
0.005	186×10^3	30.1	[10, 193]	32.2	7	0.0033	−0.003

Finally, the constant that defines the jump amplitude is $c_* = 1/\sqrt{N_J}$. The exact solution of the corresponding weak approximation problem is

$$E[g(X(T))] = \exp\left(\int_0^T a(t)dt\right) = \exp\left(\sqrt{T - 1/3 + \mathrm{TOL}^4} - \mathrm{TOL}^2\right).$$

Due to the time discontinuity of the drift function and to ensure optimal convergence of the adaptive algorithms, we modify the diffusion step in the Euler method by

$$\overline{X}(t_{n+1}^-) - \overline{X}(t_n) = a(\hat{t}, \overline{X}_n)\Delta t_n + \overline{X}(t_n)\Delta W_n, \quad n = 0, \ldots, \kappa - 1,$$

where we choose the stochastic evaluation time $\hat{t} \in \{t_n, t_{n+1}\}$ by

$$|a(\hat{t}, \overline{X}_n)| = \max(|a(t_n, \overline{X}_n)|, |a(t_{n+1}^-, \overline{X}_n)|).$$

Observe that the use of $\hat{t}$ does not change the adapted nature of the Euler method. To illustrate the problem dynamics, Figure 5.1 shows three realizations of the approximate process, $\overline{X}$, and the backward dual function, $\varphi(t_n)$. The latter represents the pathwise sensitivity of $g(\overline{X}(T))$ with respect to perturbations in the initial data of X at time t_n. It coincides in this example with the first variation of $\overline{X}$, namely, $\overline{X}'(T;t_n)$. Figure 5.2 and Table 5.4 show information related to the adaptive runs for two values of the accuracy tolerance, starting in both cases with $M = 10^2$ realizations and $N = 10$ uniform time steps. In particular, Figure 5.2 shows the adaptive time steps needed to control the numerical error. Observe that these deterministic time steps are prior to the augmentation with the sampled jump times and are the ones we can choose to control the time discretization error. Figure 5.2 shows that although we may have more than a thousand time steps added in the interval $[2/3, 1]$ due to jump times, we still need to significantly refine our deterministic time steps near the singularity at $t = 1/3$. Table 5.4 shows information related to the computational effort necessary to solve the problem. There we can see that the amount of realizations needed to find our deterministic time steps, M_T, is much smaller than the amount needed to solve the problem, M. For instance, when $\mathrm{TOL} = 1.0 \times 10^{-2}$ we need only $M_T \approx M \times 10^{-3}$ realizations to control the time discretization error, making the adaptivity procedure completely inexpensive. Using the approximate error density we have also estimated the number of uniform time steps, N_U, needed to achieve the prescribed accuracy. Table 5.4 shows that N_U is approximately 100 times larger than the number of deterministic time steps, N. This fact motivates further the use of Algorithm D in this problem. Finally, we emphasize that the computational effort due to the augmentation of the time discretization with the jump times decreases as TOL decreases. Even in this case with $N_J = 1024$ jump times in the mean we

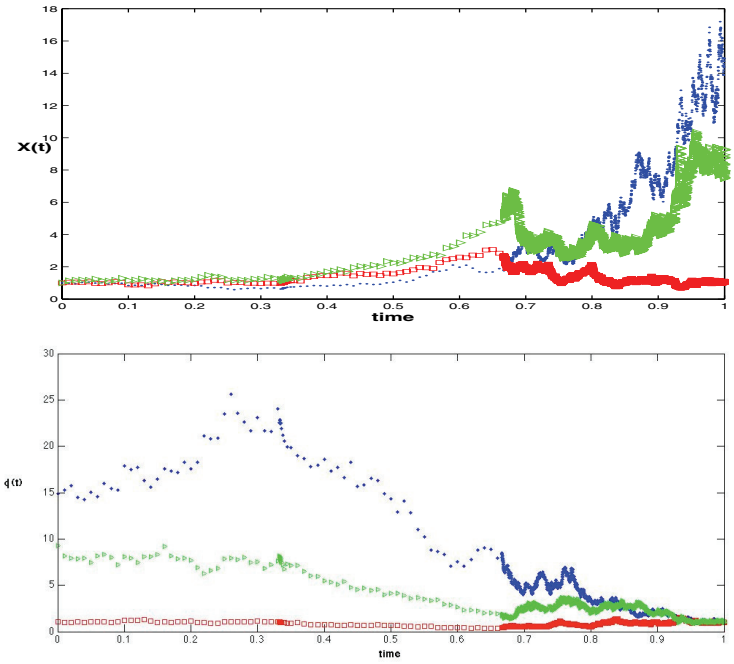


FIG. 5.1. *Three sample realizations of the processes* $\overline{X}$ *(above) and* φ *(below) obtained with modified Euler in Example* 2.

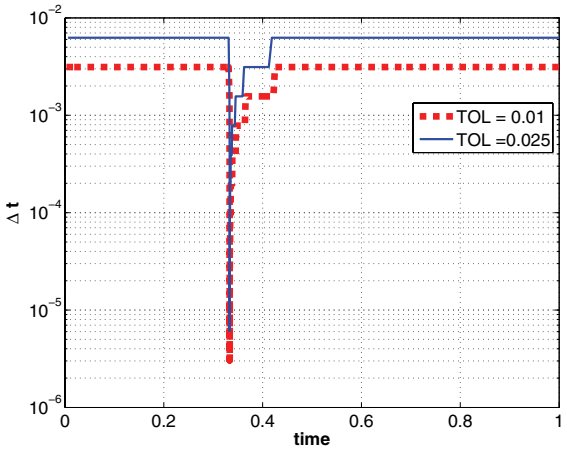


FIG. 5.2. *Final time discretizations (jumps not included) for the adaptive runs with Algorithm D and* $\mathrm{TOL} = 2.5 \times 10^{-2}$, $\mathrm{TOL} = 1 \times 10^{-2}$ *in Example* 2.

TABLE 5.4
Adaptive choice of M *and* Δt *with Algorithm D in Example* 2.

TOL	M	N	M_T	N_U	$\mathcal{E}$
2.5×10^{-2}	2.0×10^{5}	220	1.7×10^{3}	2×10^{4}	-2.0×10^{-2}
1.0×10^{-2}	1.45×10^{6}	450	1.7×10^{3}	6×10^{4}	-1.4×10^{-2}

anticipate that for a smaller value of TOL $\approx 10^{-3}$ the number of deterministic time steps needed to control the discretization error will be approximately 4 times larger than the mean number of jumps.

6. Proofs. This section contains the proofs postponed in the previous sections.

6.1. Proof of Lemma 2.1. The proof follows closely the proof of Corollary II.2.2 to Theorem II.2.6 (pp. 237–238 in [19]). The main two differences with this result are (i) our process has finite Lévy measure, while there a general Lévy measure is considered; (ii) our jump distribution depends on time, while there it does not. Theorem II.2.6 in [19] covers our case, and our regularity conditions on a and b are the same as in the mentioned corollary. Concerning the jumps, it is direct to see that (2.3) replaces the corresponding condition on the jump coefficient in Corollary II.2.2 in [19], following now the result from Theorem II.2.6 [19, p. 237]. The other minor difference consists of the compensation of the jump part of the differential equation (1.1). Defining

$$\alpha(t,x) = a(t,x) + \lambda(t)\int_{\mathbf{z}} c(t,x,z)\mu(t,dz),$$

we see that our equation can be written as

$$\begin{aligned} X(t) = X(0) &+ \int_0^t \alpha(\,s, X(s^-)\,)\,ds + \sum_{\ell=1}^{\ell_0}\int_0^t b^\ell(\,s, X(s^-)\,)\,dW^\ell(s) \\ &+ \int_0^t\int_{\mathbf{z}} c(\,s, X(s^-), z\,)\,\left(p(ds,dz) - q(ds,dz)\right), \end{aligned}$$

which is the form of the equation in [19]. This concludes the proof of part (a).

Part (b) is a consequence of the representation of the derivatives in terms of the fluxes, by application of the chain rule, as developed in formulas (6.9)–(6.12). We apply the Cauchy–Schwarz inequality to each summand in each representation to obtain the estimates based on the respective polynomial growth of the function g and its derivatives, and the boundedness of the moments of the process X and its variations. Since the full proof is somewhat lengthy, we do not include it here. □

6.2. Proof of Lemma 2.2. Based on Lemma 2.1 we know that the cost to go function $u(t,x)$ is the solution of the Kolmogorov backward equation (2.5), satisfying the regularity conditions described in (iv) of part (b) of Lemma 2.2. The Itô formula for processes with jumps (cf. [25, p. 66]) applied to (2.2) gives

$$\begin{aligned} u(T,\overline{X}(T)) = u(0,X(0)) &+ \int_0^T \left[\,\partial_t u(t,\overline{X}(t^-)) + \overline{a}_i(t;\overline{X})\,\partial_i u(t,\overline{X}(t^-))\,\right]\,dt \\ &+ \int_0^T \overline{d}_{ij}(t;\overline{X})\,\partial_{ij}u(t,\overline{X}(t^-))\,dt \\ &+ \int_0^T \overline{b}_i^\ell(t;\overline{X})\,\partial_i u(t,\overline{X}(t^-))\,dW^\ell(t) \\ &+ \int_0^T\int_{\mathbf{z}} \left[\,u\big(t,\overline{X}(t^-) + c(t,\overline{X}(t^-),z)\big) - u(t,\overline{X}(t^-))\,\right]\,p(dt,dz), \end{aligned}$$

which is written equivalently as

$$\begin{aligned}
u(T,\overline{X}(T)) = u(0,X(0)) &+ \int_0^T \big[\,\partial_t u(t,\overline{X}(t^-)) + \overline{a}_i(t;\overline{X})\,\partial_i u(t,\overline{X}(t^-))\,\big]\,dt \\
&+ \int_0^T \overline{d}_{ij}(t;\overline{X})\,\partial_{ij}u(t,\overline{X}(t^-))\,dt \\
&+ \int_0^T \overline{b}_i^{\ell}(t;\overline{X})\,\partial_i u(t,\overline{X}(t^-))\,dW^{\ell}(t) \\
&+ \int_0^T \lambda(t)\int_{\mathbf{z}}\big[\,u\big(t,\overline{X}(t^-)+c(t,\overline{X}(t^-),z)\big)-u(t,\overline{X}(t^-))\,\big]\,\mu(t,dz)dt \\
&+ \int_0^T\int_{\mathbf{z}}[u(t,\overline{X}(t^-)+c(t,\overline{X}(t^-),z))-u(t,\overline{X}(t^-))]\,\big(p(dt,dz)-q(dt,dz)\big),
\end{aligned}$$

where the last equality is obtained by the compensation of the integral with respect to p (remember that $q(dt,dz)=\lambda(t)\mu(t,dz)dt$). Now, use (2.5) to substitute $\partial_t u(t,\overline{X}(t^-))$ in the latter formula to obtain

$$\begin{aligned}
u(0,X(0)) = g(\overline{X}(T)) &+ \int_0^T \big[\,a_i(t,\overline{X}(t^-)) - \overline{a}_i(t;\overline{X})\,\big]\,\partial_i u(t,\overline{X}(t^-))\,dt \\
&+ \int_0^T \big[\,d_{ij}(t,\overline{X}(t^-)) - \overline{d}_{ij}(t;\overline{X})\big]\,\partial_{ij}u(t,\overline{X}(t^-))\,dt \\
&- \int_0^T \overline{b}_i^{\ell}(t;\overline{X})\,\partial_i u(t,\overline{X}(t^-))\,dW^{\ell}(t) \\
&- \int_0^T\int_{\mathbf{z}}\big[\,u(t,\overline{X}(t^-)+c(t,\overline{X}(t^-),z)) - u(t,\overline{X}(t^-))\big]\,\big(p(dt,dz)-q(dt,dz)\big).
\end{aligned} \tag{6.1}$$

The expected value of the last two integrals is zero, the first one by the martingale property of Itô integrals and the second one due to the fact that $p(dt,dz)-q(dt,dz)$ is a compensated random measure (i.e., a martingale measure). Using the representation of u in (2.4) we obtain $E\big[u(0,X(0))\big] = E\big[g(X(T))\big]$. Therefore, taking expected values in both sides of (6.1) we arrive at the error representation (2.6). □

6.3. Proof of Theorem 2.2. The main content of Theorem 2.2 is the replacement of the (noncomputable) error estimate in Lemma 2.2 by an error expansion with computable leading order term. For this purpose the derivatives of the expected value

$$\partial_\alpha u(t,x) = \partial_\alpha E\big[g(X(T))|X(t)=x\big]$$

appearing in the integral in Lemma 2.2 are approximated by the corresponding derivatives of

$$\bar{u}(t,x) \equiv E[g(\overline{X}(T)) \mid \overline{X}(t)=x],$$

which depends on the simulated solution $\overline{X}$. The proof is divided into three steps:

(i) in Lemma 6.1 we estimate the quadrature error;

(ii) in Lemma 6.2 we bound the error in replacing $\partial_\alpha u$ by $\partial_\alpha \bar{u}$ with the use of stochastic flows and its variations;

(iii) in Lemma 6.3 we use the discrete dual functions φ and φ' (which solve the backward evolution problems; see (1.13)–(1.14) in section 1.4) to derive a computable representation of $\partial_\alpha \bar{u}$.

Step 1. We begin with the estimation of the error of the quadrature approximation.

LEMMA 6.1. *Let* $\mathcal{J}_m \equiv \{n : \tilde{t}_m \leq t_n < \tilde{t}_{m+1}\}$ *for* $m = 0, \ldots, N-1$. *If the assumptions in Lemma* 2.2 *hold, then the following quadrature error estimates hold:*

$$\int_{\tilde{t}_m}^{\tilde{t}_{m+1}} E\big[\big(a_i(t, \overline{X}(t^-)) - \overline{a}_i(t; \overline{X})\big)\, \partial_i u(t, \overline{X}(t^-))\big] dt = Q_m(a_i) + \mathcal{O}\big((\Delta \tilde{t}_m)^3\big)$$

and

$$\int_{\tilde{t}_m}^{\tilde{t}_{m+1}} E\big[\big(d_{ij}(t, \overline{X}(t^-)) - \overline{d}_{ij}(t; \overline{X})\big)\partial_{ij} u(t, \overline{X}(t^-))\big] dt = \widetilde{Q}_m(d_{ij}) + \mathcal{O}\big((\Delta \tilde{t}_m)^3\big)$$

for $m = 0, \ldots, N-1$, *where*

$$Q_m(a_i) \equiv E\left[\sum_{n \in \mathcal{J}_m} \frac{\Delta t_n}{2} \left[a_i(t_{n+1}, \overline{X}(t_{n+1}^-)) - a_i(t_n, \overline{X}(t_n))\right] \partial_i u(t_{n+1}, \overline{X}(t_{n+1}^-))\right]$$

and

$$\widetilde{Q}_m(d_{ij}) \equiv E\left[\sum_{n \in \mathcal{J}_m} \frac{\Delta t_n}{2} \left[d_{ij}(t_{n+1}, \overline{X}(t_{n+1}^-)) - d_{ij}(t_n, \overline{X}(t_n))\right] \partial_{ij} u(t_{n+1}, \overline{X}(t_{n+1}^-))\right].$$

Proof. Denote by $\mathcal{F}^J$ the σ-algebra generated by the jumps and marks in $[0, T]$ constructed in section 1.2.1. Observe that in $[t_n, t_{n+1})$ the conditioned process $\overline{X}$ has no jump discontinuities, and introduce the auxiliary functions

$$\begin{aligned} \mathcal{H}(t, \overline{X}(t^-)) &\equiv \left[\, d_{ij}(t, \overline{X}(t^-)) - \overline{d}_{ij}(t; \overline{X}) \,\right] \partial_{ij} u(t, \overline{X}(t^-)), \\ \overline{\mathcal{H}}_n(t) &\equiv \frac{t - t_n}{\Delta t_n}\, \mathcal{H}(t_{n+1}, \overline{X}(t_{n+1}^-)), \quad n = 0, \ldots, \kappa - 1. \end{aligned}$$

Then we have

$$\begin{aligned} \int_{\tilde{t}_m}^{\tilde{t}_{m+1}} E[\mathcal{H}(t, \overline{X}(t^-))]\, dt &= E\left[\sum_{n \in \mathcal{J}_m} \int_{t_n}^{t_{n+1}} E\big[\mathcal{H}(t, \overline{X}(t^-))|\mathcal{F}^J\big]\, dt\right] \\ &= \widetilde{Q}_m(d_{ij}) + E\left[\sum_{n \in \mathcal{J}_m} \int_{t_n}^{t_{n+1}} E\big[\mathcal{H}(t, \overline{X}(t^-)) - \overline{\mathcal{H}}_n(t)|\mathcal{F}^J\big]\, dt\right]. \end{aligned} \tag{6.2}$$

Observing that the function $E\big[\,\overline{\mathcal{H}}_n(t)|\mathcal{F}^J\,\big]$ is the linear interpolant of the smooth function $E\big[\,\mathcal{H}(t, \overline{X}(t^-))|\mathcal{F}^J\big]$ at the nodes $\{t_n, t_{n+1}\}$, a standard interpolation error

estimate yields

$$\begin{aligned}\left|\int_{t_n}^{t_{n+1}} E[\mathcal{H}(t,\overline{X}(t^-)) - \overline{\mathcal{H}}_h(t)|\mathcal{F}^J]\,dt\right| \\ \leq \frac{(\Delta t_n)^2}{8}\int_{t_n}^{t_{n+1}}\left|\frac{d^2}{dt^2}E[\mathcal{H}(t,\overline{X}(t^-))|\mathcal{F}^J]\right|dt.\end{aligned} \tag{6.3}$$

Denoting $\mathcal{L}\mathcal{H} \equiv \partial_t\mathcal{H}+\overline{a}_i\partial_i\mathcal{H}+\overline{d}_{ij}\partial_{ij}\mathcal{H}$, Itô's formula and the conditions in Lemma 2.1 show that there exist constants K_1, K_2 and $p \in \mathbb{N}$ such that

$$\begin{aligned}\frac{d}{dt}E[\mathcal{H}(t,\overline{X}(t^-))|\mathcal{F}^J] &= E[\mathcal{L}\mathcal{H}(t,\overline{X}(t^-))|\mathcal{F}^J] \\ &\leq K_1 + K_2\,E[\max_{t\in[0,T]}|\overline{X}_t|^{2p}|\mathcal{F}^J], \\ \frac{d^2}{dt^2}E[\mathcal{H}(t,\overline{X}(t^-))|\mathcal{F}^J] &= E[\mathcal{L}^2\mathcal{H}(t,\overline{X}(t^-))|\mathcal{F}^J] \\ &\leq K_1 + K_2\,E[\max_{t\in[0,T]}|\overline{X}_t|^{2p}|\mathcal{F}^J],\end{aligned} \tag{6.4}$$

as all the functions involved in the derivatives of $\mathcal{H}(t,x)$ in the computation of $\mathcal{L}\mathcal{H}$ and $\mathcal{L}^2\mathcal{H}$ have polynomial growth, and $\overline{X}$ has bounded moments of all order, uniformly with respect to Δt_{max} (from this fact follows the finiteness of the moments of the supremum of the process). The combination of (6.2), (6.3), and (6.4) proves the estimate of the diffusion term in the lemma. The estimate of the drift term follows analogously. □

Step 2. In the second step of the proof the derivative $\partial_\alpha u$ and its approximation $\partial_\alpha \bar{u}$ are evaluated, respectively, through expected values of *stochastic flows* of X and $\overline{X}$ and its *variations*, which we now introduce. Let

$$\delta_{ik} \equiv \begin{cases} 0, & i \neq k, \\ 1, & i = k. \end{cases}$$

The following equation for the first variation of the process X at time $s > t$ holds:

$$\begin{aligned} dX'_{ij}(s) &= \partial_k a_i(s,X(s^-))\,X'_{kj}(s^-)\,ds + \partial_k b_i^\ell(s,X(s^-))\,X'_{kj}(s^-)\,dW^\ell(s) \\ &\quad + \int_{\mathbf{z}} \partial_k c_i(s,X(s^-),z)\,X'_{kj}(s^-)\,p(ds,dz), \\ X'_{ij}(t) &= \delta_{ij}. \end{aligned} \tag{6.5}$$

Similarly, for the second variation of the process X at time $s > t$ we have

$$\begin{aligned} dX''_{ijn}(s) &= \left[\partial_k a_i(s,X(s^-))\,X''_{kjn}(s^-) + \partial_{kr}a_i(s,X(s^-))\,X'_{kj}(s^-)\,X'_{rn}(s^-)\right]ds \\ &\quad + \left[\partial_k b_i^\ell(s,X(s^-))\,X''_{kjn}(s^-)\right. \\ &\qquad \left. + \partial_{kr}b_i^\ell(s,X(s^-))\,X'_{kj}(s^-)\,X'_{rn}(s^-)\right]dW^\ell(s) \\ &\quad + \int_{\mathbf{z}}\left[\partial_k c_i(s,X(s^-),z)\,X''_{kjn}(s^-)\right. \\ &\qquad \left. + \partial_{kr}c_i(s,X(s^-),z)\,X'_{kj}(s^-)\,X'_{rn}(s^-)\right]p(ds,dz), \\ X''_{ijn}(t) &= 0. \end{aligned} \tag{6.6}$$

For the third variation of the process X at time $s > t$ we have

$$
\begin{aligned}
dX'''_{ijnm}(s) = \Big[& \partial_k a_i(s, X(s^-))\, X'''_{kjnm}(s^-) \\
& + \partial_{kr} a_i(s, X(s^-))\, X'_{kj}(s^-)\, X''_{rnm}(s^-) \\
& + \partial_{kr} a_i(s, X(s^-))\, X'_{kn}(s^-)\, X''_{rjm}(s^-) \\
& + \partial_{kr} a_i(s, X(s^-))\, X'_{km}(s^-)\, X''_{rjn}(s^-) \\
& + \partial_{krv} a_i(s, X(s^-))\, X'_{kj}(s^-)\, X'_{rn}(s^-)\, X'_{vm}(s^-) \Big]\, ds \\
+ \Big[& \partial_k b_i^\ell(s, X(s^-))\, X'''_{kjnm}(s^-) \\
& + \partial_{kr} b_i^\ell(s, X(s^-))\, X'_{kj}(s^-)\, X''_{rnm}(s^-) \\
& + \partial_{kr} b_i^\ell(s, X(s^-))\, X'_{kn}(s^-)\, X''_{rjm}(s^-) \\
& + \partial_{kr} b_i^\ell(s, X(s^-))\, X'_{km}(s^-)\, X''_{rjn}(s^-) \\
& + \partial_{krv} b_i^\ell(s, X(s^-))\, X'_{kj}(s^-)\, X'_{rn}(s^-)\, X'_{vm}(s^-) \Big]\, dW^\ell(s) \\
+ \int_{\mathbf{z}} \Big[& \partial_k c_i(s, X(s^-))\, X'''_{kjnm}(s^-) \\
& + \partial_{kr} c_i(s, X(s^-), z)\, X'_{kj}(s^-) X''_{rnm}(s^-) \\
& + \partial_{kr} c_i(s, X(s^-), z)\, X'_{kn}(s^-)\, X''_{rjm}(s^-) \\
& + \partial_{kr} c_i(s, X(s^-), z)\, X'_{km}(s^-)\, X''_{rjn}(s^-) \\
& + \partial_{krv} c_i(s, X(s^-), z)\, X'_{kj}(s^-)\, X'_{rn}(s^-)\, X'_{vm}(s^-) \Big]\, p(ds, dz), \\
X'''_{ijnm}(t) = & \, 0,
\end{aligned}
\tag{6.7}
$$

and similarly for the fourth variation of the process X at time $s > t$:

$$
dX''''_{ijnmp} = \ldots, \qquad X''''_{ijnmp}(t) = 0. \tag{6.8}
$$

Let $I^{d\times d}$ be the $d \times d$ identity matrix. The equations above imply the representation of the derivatives of expectations with stochastic flows (cf. [19], [49], and [52]) as follows: for the first order derivatives

$$
\partial_k u(t,x) = E\big[\, \partial_i g(X(T))\, X'_{ik}(T) \mid X^{I}(t) = (x, I^{d\times d}) \,\big] \tag{6.9}
$$

with $X^{I} \equiv (X, X')$; for the second order derivatives

$$
\begin{aligned}
\partial_{kn} u(t,x) = E\big[& \partial_i g(X(T))\, X''_{ikn}(T) \\
& + \partial_{ir} g(X(T))\, X'_{ik}(T)\, X'_{rn}(T) \mid X^{II}(t) = (x, I^{d\times d}, 0) \,\big]
\end{aligned}
\tag{6.10}
$$

with $X^{II} \equiv (X, X', X'')$; for the third order derivatives

$$
\begin{aligned}
\partial_{knm} u(t,x) = E\big[& \partial_i g(X(T))\, X'''_{iknm}(T) + \partial_{ir} g(X(T))\, X'_{ik}(T)\, X''_{rnm}(T) \\
& + \partial_{ir} g(X(T))\, X'_{in}(T)\, X''_{rkm}(T) \\
& + \partial_{irv} g(X(T))\, X'_{ik}(T)\, X'_{rn}(T)\, X'_{vm}(T) \\
& + \partial_{ir} g(X(T))\, X'_{im}(T)\, X''_{rkn}(T) \mid X^{III}(t) = (x, I^{d\times d}, 0, 0) \big]
\end{aligned}
\tag{6.11}
$$

with $X^{III} \equiv (X, X', X'', X''')$; and, analogously, for the fourth order derivatives

$$
\partial_{knmp} u(t,x) = \ldots. \tag{6.12}
$$

Let $Y = (X, X', X'', X''', X'''')^T$. Then the system consisting of (1.1) and (6.5)–(6.8) can be written as

$$\begin{aligned} dY(t) &= A(t, Y(t^-))\, dt + B^\ell(t, Y(t^-))\, dW^\ell(t) \\ &\quad + \int_{\mathbf{z}} C(t, Y(t^-), z)\, p(dt, dz), \quad t > t_0, \\ Y(t_0) &= (x, I^{d\times d}, 0, 0, 0)^T. \end{aligned} \tag{6.13}$$

Furthermore, in order to simplify the writing of the formulas (6.9)–(6.12), we introduce below additional notation:

$$\begin{aligned} f_i(Y(T)) &\equiv \partial_k g(X(T))\, X'_{ki}(T), \\ f_{ij}(Y(T)) &\equiv \partial_k g(X(T))\, X''_{kij}(T) + \partial_{kn} g(X(T))\, X'_{ki}(T)\, X'_{nj}(T), \\ f_{ijm}(Y(T)) &\equiv \partial_k g(X(T))\, X'''_{kijm}(T) + \partial_{kn} g(X(T))\, X'_{ki}(T)\, X''_{njm}(T) \\ &\quad + \partial_{kn} g(X(T))\, X'_{kj}(T)\, X''_{nim}(T) \\ &\quad + \partial_{kn} g(X(T))\, X'_{km}(T)\, X''_{nji}(T) \\ &\quad + \partial_{knv} g(X(T))\, X'_{ki}(T)\, X'_{nj}(T)\, X'_{vm}(T), \\ f_{ijmn}(Y(T)) &\equiv \dots. \end{aligned} \tag{6.14}$$

The Euler approximation of Y, the solution of (6.13), is denoted by

$$\overline{Y} = \big(\overline{Y}^0, \overline{Y}^1, \overline{Y}^2, \overline{Y}^3, \overline{Y}^4\big)^\tau$$

and can be extended as the solution of the stochastic differential equation with piecewise constant drift and diffusion fluxes

$$d\overline{Y} = \overline{A}(t; \overline{Y})\, dt + \overline{B}^\ell(t; \overline{Y})\, dW^\ell(t) + \int_{\mathbf{z}} C(t, \overline{Y}, z)\, p(dt, dz), \tag{6.15}$$

defined as in (1.4), (1.6), and (2.2), with coefficients defined as in (2.1).

An important consequence of the Euler method is that the variation and the Euler discretization commute. This yields for each α the representation

$$\begin{aligned} &\partial_\alpha u(t,x) - \partial_\alpha \bar{u}(t,x) \\ &\qquad = E\big[f_\alpha(Y(T)) - f_\alpha(\overline{Y}(T)) \mid Y(t) = \overline{Y}(t) = (x, I^{d\times d}, 0, 0, 0)^T\big]. \end{aligned} \tag{6.16}$$

Now we rely on Lemma 2.2 applied to the process Y to obtain the representation

$$\begin{aligned} &E\big[f_\alpha(Y(T)) - f_\alpha(\overline{Y}(T)) \mid Y(t) = \overline{Y}(t) = (x, I^{d\times d}, 0, 0, 0)^T\big] \\ &\quad = \int_t^T E\Big[(A - \overline{A})_k\, \partial_k \nu^\alpha(s, \overline{Y}(s^-)) \,\Big|\, \overline{Y}(t) = (x, I^{d\times d}, 0, 0, 0)^T\Big]\, ds \\ &\qquad + \int_t^T E\Big[(D - \overline{D})_{kn}\, \partial_{kn} \nu^\alpha(s, \overline{Y}(s^-)) \,\Big|\, \overline{Y}(t) = (x, I^{d\times d}, 0, 0, 0)^T\Big]\, ds \end{aligned} \tag{6.17}$$

to be used in Lemmas 6.2 and 6.3 below, where the cost to go function ν^α is, for each f_α in (6.14), defined as

$$\nu^\alpha(t, y) = E\big[f_\alpha(Y(T)) \mid Y(t) = y\big].$$

We use the notation $D_{kn} = \frac{1}{2}B_k^\ell B_n^\ell$ and $\overline{D}_{kn} = \frac{1}{2}\overline{B}_k^\ell \overline{B}_n^\ell$, and the corresponding Kolmogorov backward equation for $t > T$ is

$$\begin{aligned}\mathcal{L}^Y \nu^\alpha &\equiv \partial_t \nu^\alpha + A_k \partial_k \nu^\alpha + D_{kn}\partial_{kn}\nu^\alpha \\ &\quad + \lambda(t)\int_{\mathbf{z}} [\nu^\alpha(t, Y + C(t,Y,z)) - \nu^\alpha(t,Y)]\,\mu(t,dz) = 0,\\ \nu^\alpha(T,\cdot) &= f_\alpha.\end{aligned}$$

We are ready for the second step.

LEMMA 6.2 (approximation of $\partial_\alpha u$). *Let the piecewise constant mesh function Δt be defined by $\Delta t(s) \equiv \Delta t_n$ for $s \in [t_n, t_{n+1})$ and $n = 0, \ldots, \kappa - 1$. Suppose that the assumptions in Lemma 2.2 hold. Then the discretization errors of the stochastic flows, for $|\alpha| \le 4$, satisfy*

$$\partial_\alpha(u - \bar{u})(t_n, \overline{X}(t_n)) = \int_{t_n}^T E[\mathcal{O}(\Delta t(s)) \mid \mathcal{F}_{t_n}]\, ds = \mathcal{O}(\Delta \tilde{t}_{max}). \tag{6.18}$$

Furthermore

$$\begin{aligned}E\Big[\big[a_i(t_{n+1}, \overline{X}(t_{n+1}^-)) - a_i(t_n, \overline{X}(t_n))\big]\, \partial_i(u-\bar{u})(t_{n+1}, \overline{X}(t_{n+1}^-)) \mid \mathcal{F}^J\Big] \\ = \Delta t_n \int_{t_{n+1}}^T \mathcal{O}(\Delta t(s))\, ds\end{aligned} \tag{6.19}$$

and

$$\begin{aligned}E\Big[\big[\, d_{ij}(t_{n+1}, \overline{X}(t_{n+1}^-)) - d_{ij}(t_n, \overline{X}(t_n))\,\big]\, \partial_{ij}(u-\bar{u})(t_{n+1}, \overline{X}(t_{n+1}^-)) \mid \mathcal{F}^J\Big] \\ = \Delta t_n \int_{t_{n+1}}^T \mathcal{O}(\Delta t(s))\, ds.\end{aligned} \tag{6.20}$$

Proof. The combination of (6.16) and (6.17) gives

$$\begin{aligned}\partial_\alpha(u-\bar{u})(t_n, \overline{X}(t_n)) &= \int_{t_n}^T E\big[(A_i - \overline{A}_i)\,\partial_i \nu^\alpha(s, \overline{Y}(s^-)) \mid \mathcal{F}_{t_n}\big]\, ds \\ &\quad + \int_{t_n}^T E\big[(D_{ij} - \overline{D}_{ij})\,\partial_{ij}\nu^\alpha(s, \overline{Y}(s^-)) \mid \mathcal{F}_{t_n}\big]\, ds.\end{aligned} \tag{6.21}$$

Now, for $t_m \le t < t_{m+1}$, introduce the notation

$$h(t, \overline{Y}(t)) \equiv (A_i - \overline{A}_i)\partial_i \nu^\alpha(t, \overline{Y}(t)) + (D_{ij} - \overline{D}_{ij})\partial_{ij}\nu^\alpha(t, \overline{Y}(t))$$

and

$$\begin{aligned}\mathcal{L}_0 w(t,y) &\equiv \big(\partial_t w + \overline{A}_n \partial_n w + \overline{D}_{kn}\partial_{kn} w\big)(t,y),\\ \mathcal{L}^{\overline{Y}} w(t,y) &\equiv \mathcal{L}_0 w(t,y) + \lambda(t)\int_{\mathbf{z}} [w(t, y + C(t,y,z)) - w(t,y)]\mu(t,dz).\end{aligned}$$

Observe that $h(t_m, \overline{Y}(t_m)) = 0$, and apply Itô's formula in the interval $[t_m, t]$ to obtain

$$E[h(t, \overline{Y}(t^-)) \mid \mathcal{F}_{t_m}] = \int_{t_m}^t E[\mathcal{L}^{\overline{Y}} h(s, \overline{Y}(s^-)) \mid \mathcal{F}_{t_m}] ds = \mathcal{O}(\Delta t_m), \quad t_m \le t < t_{m+1}, \tag{6.22}$$

where the bound follows as in (6.4). Combining (6.22) with (6.21) proves (6.18).

The estimate (6.20) follows similarly by defining now

$$\tilde{h}(t,\overline{X}(t)) \equiv (d_{ij}-\overline{d}_{ij})\partial_{ij}(u-\bar{u})(t,\overline{X}(t)).$$

Then, as in (6.3)–(6.4), the Itô's formula shows

$$E[\tilde{h}(t,\overline{X}(t)) \mid \mathcal{F}^J] = \int_{t_n}^{t} E[\mathcal{L}_0\tilde{h}(s,\overline{X}(s^-)) \mid \mathcal{F}^J]ds, \quad t_n \le t < t_{n+1}.$$

The final step to proving (6.20) is to establish

$$E[\mathcal{L}_0\,\tilde{h}(s,\overline{X}(s^-)) \mid \mathcal{F}^J] = \int_{s}^{T} \mathcal{O}(\Delta t(\tau))d\tau.$$

The function $\mathcal{L}_0\,\tilde{h}(s,\overline{X}(s^-))$ splits into the two types of terms $h_1 \equiv (d-\overline{d})v$ and $h_2 \equiv v\partial_\alpha(u-\bar{u})$, with smooth functions v of $(s,\overline{X}(s))$. The Itô formula again shows that

$$E[h_1(s,\overline{X}(s^-)) \mid \mathcal{F}^J] = \int_{t_n}^{s} E[\mathcal{L}_0 h_1(\tau,\overline{X}(\tau^-)) \mid \mathcal{F}^J]d\tau = \mathcal{O}(\Delta t_n), \quad t_n \le s < t_{n+1}.$$

Moreover (6.18) implies

$$E[h_2(s,\overline{X}(s^-)) \mid \mathcal{F}^J] = \int_{s}^{T} \mathcal{O}(\Delta t(\tau))d\tau,$$

and consequently (6.20) holds. The estimate (6.19) of the drift terms follows analogously. □

Step 3. The last step is to show a result on the representation of the derivative of $\bar{u}$ using discrete duals.

LEMMA 6.3. *Suppose that the assumptions in Lemma* 2.2 *hold. Then the dual functions* φ *and* φ', *defined in* (1.13) *and* (1.14), *satisfy for* $t = t_n$ *and* $t = t_{n+1}^-$

$$\partial_i\bar{u}(t,\overline{X}(t)) = E[\varphi_i(t) \mid \mathcal{F}_t], \tag{6.23}$$

$$\partial_{ij}\bar{u}(t,\overline{X}(t)) = E[\varphi'_{ij}(t) \mid \mathcal{F}_t]. \tag{6.24}$$

Proof. Equations (2.2), (6.5), and (6.9) show that the first variation of the Euler approximation $\overline{X}$ is in fact equal to the Euler approximation of the first variation X', and consequently

$$\partial_i\bar{u}(t,\overline{X}(t)) = E[\partial_j g(\overline{X}(T))\overline{X}'_{ji}(T;t) \mid \mathcal{F}_t],$$

where $\overline{X}'_{ji}(s;t)$ $(s > t)$, is the Euler approximation (6.15) of X' with initial data $\overline{X}'_{ji}(t;t) = \delta_{ji}$.

For $t = \ldots, t_n, t_{n+1}^-, \ldots$, let

$$\varphi_i(t) \equiv \partial_{x_i(t)} g(\overline{X}(T)),$$

i.e.,

$$\varphi_i(t) = \partial_j g(\overline{X}(T))\overline{X}'_{ji}(T;t). \tag{6.25}$$

We prove inductively that $\varphi_i(t)$ is the solution of the corresponding problem in (1.12)–(1.13). Since (1.12) is trivially true it remains to prove the inductive step. By the chain rule we have

$$\partial_k g(\overline{X}(T))\overline{X}'_{ki}(T;t) = \partial_k g(\overline{X}(T))\overline{X}'_{kj}(T;t_{next})\overline{X}'_{ji}(t_{next};t)$$

or, in other words,

$$\varphi_i(t) = \varphi_j(t_{next})\partial_i \Phi_j(t, \overline{X}(t)), \tag{6.26}$$

which is equivalent to (1.12)–(1.13), which is what we wanted to prove.

The equality (6.25) implies that

$$\partial_{ij}\bar{u}(t, \overline{X}(t)) = E\big[\partial_{x_j(t)}\varphi_i(t) \mid \mathcal{F}_t\big].$$

The next step is to verify that the first variation of φ,

$$\varphi'_{ij}(t) \equiv \partial_{x_j(t)}\varphi_i(t) \equiv \frac{\partial \varphi_i(t; \overline{X}(t) = x)}{\partial x_j}, \tag{6.27}$$

satisfies the backward recursive equation (1.14). First, differentiate the equation (6.26) to obtain

$$\begin{aligned} \varphi'_{ik}(t) &= \partial_i \Phi_j \partial_{x_k(t)}\varphi_j(t_{next}) + \partial_k \partial_i \Phi_j \varphi_j(t_{next}), \quad t < T, \\ \varphi'_{ik}(T) &= \partial_{ik} g(\overline{X}(T)). \end{aligned} \tag{6.28}$$

Observe that problem (1.12)–(1.16) shows that $\varphi(t_{next})$ depends only on the point values

$$\{\overline{X}(s) : \ t_{next} \le s \le T\},$$

so that

$$\partial_{x_k(t)}\varphi_j(t_{next}) = \partial_{x_p(t_{next})}\varphi_j(t_{next})\partial_{x_k(t)}\overline{X}_p(t_{next}). \tag{6.29}$$

Finally, the definitions of $\overline{X}$ and Φ in (2.2) and (1.11) imply

$$\partial_k \Phi_p(t, \overline{X}(t)) = \partial_{x_k(t)}\overline{X}_p(t_{next}), \tag{6.30}$$

which together with (6.28)–(6.29) proves that φ' satisfies the recursive equation

$$\begin{aligned} \varphi'_{ik}(t) &= \partial_i \Phi_j \ \ \varphi'_{jp}(t_{next})\partial_k \Phi_p(t, \overline{X}(t)) + \partial_k \partial_i \Phi_j \ \ \varphi_j(t_{next}), \quad t < T, \\ \varphi'_{ik}(T) &= \partial_{ik} g(\overline{X}(T)), \end{aligned} \tag{6.31}$$

which is equivalent to (1.12)–(1.14). ∎

Remark 6.1. The measurability of $\big(a_i(t^-_{n+1}, \overline{X}(t^-_{n+1})) - \overline{a}_i(t_n, \overline{X}(t_n))\big) \in \mathcal{F}_{t^-_{n+1}}$ proves that for $t = t^-_{n+1}$ and any random variable β we have

$$\begin{aligned} E\big[\big(a_i(t, \overline{X}(t)) - \overline{a}_i(t_n, \overline{X}(t_n))\big)E[\beta \mid \mathcal{F}_t]\big] &= E\big[E[\big(a_i(t, \overline{X}(t)) - \overline{a}_i(t_n; \overline{X})\big)\beta \mid \mathcal{F}_t]\big] \\ &= E\big[\big(a_i(t, \overline{X}(t)) - \overline{a}_i(t_n; \overline{X})\big)\beta\big], \end{aligned}$$

and in a completely similar way we obtain the same result for the diffusion terms, i.e.,

$$E\big[\big(d_{ij}(t, \overline{X}(t)) - \overline{d}_{ij}(t_n; \overline{X})\big)E[\beta \mid \mathcal{F}_t]\big] = E\big[\big(d_{ij}(t, \overline{X}(t)) - \overline{d}_{ij}(t_n; \overline{X})\big)\beta\big].$$

The proof of Theorem 2.2 is now concluded by combining Lemmas 6.1, 6.2, and 6.3, Remark 6.1, and the central limit theorem to estimate I_M and J_M. We have

$$\begin{aligned}
&\int_{\tilde{t}_m}^{\tilde{t}_{m+1}} E\big[\big(a_i(t,\overline{X}(t^-)) - \overline{a}_i(t;\overline{X})\big)\,\partial_i u(t,\overline{X}(t^-))\big]dt \\
&\quad = E\left[\sum_{n\in\mathcal{J}_m} \frac{\Delta t_n}{2}\left[a_i(t_{n+1},\overline{X}(t_{n+1}^-)) - a_i(t_n,\overline{X}(t_n))\right]\partial_i u(t_{n+1},\overline{X}(t_{n+1}^-))\right] \\
&\qquad + \sum_{m=0}^{N-1}\mathcal{O}\big((\Delta\tilde{t}_m)^3\big) \\
&\quad = E\left[\sum_{n\in\mathcal{J}_m} \frac{\Delta t_n}{2}\left[\,a_i(t_{n+1},\overline{X}(t_{n+1}^-)) - a_i(t_n,\overline{X}(t_n))\,\right]\partial_i \bar{u}(t_{n+1},\overline{X}(t_{n+1}^-))\right] \\
&\qquad + \sum_{m=0}^{N-1} E\left[\sum_{n\in\mathcal{J}_m}(\Delta t_n)^2\int_{t_{n+1}}^{T}\mathcal{O}(\Delta t(s))\,ds\right] + \sum_{m=0}^{N-1}\mathcal{O}\big((\Delta\tilde{t}_m)^3\big) \\
&\quad = E\left[\sum_{n\in\mathcal{J}_m} \frac{\Delta t_n}{2}\left[\,a_i(t_{n+1},\overline{X}(t_{n+1}^-)) - a_i(t_n,\overline{X}(t_n))\,\right]E\big[\varphi_i(t_{n+1-})\,\big|\,\mathcal{F}_{t_{n+1}^-}\big]\right] \\
&\qquad + \sum_{m=0}^{N-1} E\left[\sum_{n\in\mathcal{J}_m}(\Delta t_n)^2\int_{t_{n+1}}^{T}\mathcal{O}(\Delta t(s))\,ds\right] + \sum_{m=0}^{N-1}\mathcal{O}\big((\Delta\tilde{t}_m)^3\big) \\
&\quad = E\left[\sum_{n\in\mathcal{J}_m} \frac{\Delta t_n}{2}\left[\,a_i(t_{n+1},\overline{X}(t_{n+1}^-)) - a_i(t_n,\overline{X}(t_n))\,\right]\varphi_i(t_{n+1-})\right] \\
&\qquad + \sum_{m=0}^{N-1} E\left[\sum_{n\in\mathcal{J}_m}(\Delta t_n)^2\int_{t_{n+1}}^{T}\mathcal{O}(\Delta t(s))\,ds\right] + \sum_{m=0}^{N-1}\mathcal{O}\big((\Delta\tilde{t}_m)^3\big).
\end{aligned}$$

The expansion of the diffusion term appearing in (2.8) follows analogously. ∎

6.4. Proof of Theorem 3.1. We consider the difference

$$g(X(t)) - g(\overline{X}(t))$$

and apply Theorem 3.3 from [51], using also the fact that there is no time discretization error at the jump nodes. To this end, denote the set of stochastic time nodes by $\mathcal{J} \equiv \{0 = t_0, t_1^-, t_1, t_2^-, t_2, \ldots, t_N = T\}$ and recall that the notation

$$\overline{X}(t_{next}) = \Phi(\overline{X}(t)), \quad t \in \mathcal{J}, \tag{6.32}$$

introduced in Definition 1.1 denotes one step with the Euler method. Write similarly one step with the exact solution

$$X(t_{next}) = \hat{\Phi}(X(t)), \quad t \in \mathcal{J}. \tag{6.33}$$

Introduce the notation $X^t \equiv X(t)$ and $\overline{X}^t \equiv \overline{X}(t)$. Now verify the representation

$$g(X(T)) - g(\overline{X}(T)) = \sum_{n=0}^{\kappa-1}(\hat{\Phi}(\overline{X}(t_n)) - \Phi(\overline{X}(t_n)))_i\tilde{\varphi}_i(t_{n+1}^-), \tag{6.34}$$

where the weight functions are defined recursively by linear backward recursion:

$$\begin{aligned}\tilde{\varphi}_i(T) &= \int_0^1 \partial_i g(sX(T) + (1-s)\overline{X}(T))ds,\\ \tilde{\varphi}_i(t) &= \left(\int_0^1 \partial_i \hat{\Phi}_j(sX(t) + (1-s)\overline{X}(t))ds\right)\tilde{\varphi}_j(t_{next}),\ t \in \mathcal{J}.\end{aligned} \tag{6.35}$$

To verify (6.34), first observe that by construction of the Euler method at every jump point $t = t_n^-$ there is no local error in computing the next $\overline{X}$ value at time t_n, i.e.,

$$\hat{\Phi}(\overline{X}^t) - \Phi(\overline{X}^t) = 0,$$

so (6.34) is equivalent to

$$g(X(T)) - g(\overline{X}(T)) = \sum_{t\in\mathcal{J}} (\hat{\Phi}(\overline{X}^t) - \Phi(\overline{X}^t))_i \tilde{\varphi}_i(t_{next}). \tag{6.36}$$

Then telescoping cancellation gives

$$g(X(T)) - g(\overline{X}(T)) = \sum_{t\in\mathcal{J}} \big((X^{t_{next}} - \overline{X}^{t_{next}})_i \tilde{\varphi}_i(t_{next}) - (X^t - \overline{X}^t)_i \tilde{\varphi}_i(t)\big). \tag{6.37}$$

Use the definitions (6.32) and (6.33) and split the first term in the sum of (6.37) into

$$(\hat{\Phi}(X^t) - \hat{\Phi}(\overline{X}^t) + \hat{\Phi}(\overline{X}^t) - \Phi(\overline{X}^t))_i \tilde{\varphi}_i(t_{next}).$$

The two first terms above and the last term in the sum of (6.37) combine to zero by (6.35):

$$(\hat{\Phi}(X^t) - \hat{\Phi}(\overline{X}^t))_i \tilde{\varphi}_i(t_{next})(X^t - \overline{X}^t)_i \tilde{\varphi}_i(t),$$

which proves (6.34).

The next step is to use the Malliavin derivative to analyze the expectation of the representation (6.34) by studying the dependence of $\overline{X}$ and $\tilde{\varphi}$ on a small increment dW. This follows exactly the lines of the proof of Theorem 3.3 from [51], and it is not reproduced here. □

Remark 6.2. The assumptions in Theorems 2.2 and 3.1 can be relaxed. For instance, the results of Theorem 2.2 also hold when λ and μ are piecewise continuously differentiable, provided that (i) they have a finite number of discontinuities; (ii) those discontinuity times are included in the time discretization; and (iii) for each time subinterval between discontinuities, the first derivatives of λ and μ are bounded and have lateral limits.

Acknowledgment. The authors are indebted to the anonymous referees for their useful comments regarding the manuscript.

REFERENCES

[1] L. Andersen and J. Andreasen, *Jump-diffusion processes: Volatility smile fitting and numerical methods for options pricing*, Rev. Deriv. Res., 4 (2000), pp. 231–262.

[2] S. Asmussen, F. Avram, and M.R. Pistorius, *Russian and American put options under exponential phase-type Lévy models*, Stochastic Process. Appl., 109 (2004), pp. 79–111.

[3] S. Asmussen and J. Rosiński, *Approximations of small jumps of Lévy processes with a view towards simulation*, J. Appl. Probab., 38 (2001), pp. 482–493.
[4] I. Babuška, A. Miller, and M. Vogelius, *Adaptive methods and error estimation for elliptic problems of structural mechanics*, in Adaptive Computational Methods for Partial Differential Equations (College Park, MD, 1983), SIAM, Philadelphia, PA, 1983, pp. 57–73.
[5] R. Becker and R. Rannacher, *A feed-back approach to error control in finite element methods: Basic analysis and examples*, East-West J. Numer. Math., 4 (1996), pp. 237–264.
[6] N. Bruti-Liberati and E. Platen, *On the Strong Approximation of Jump-Diffusion Processes*, Research paper 157, Quantitative Finance Research Center, University of Technology Sydney, Sydney, Australia, 2005.
[7] N. Bruti-Liberati and E. Platen, *On Weak Predictor-Corrector Schemes for Jump-Diffusion Processes in Finance*, Research paper 179, Quantitative Finance Research Center, University of Technology Sydney, Sydney, Australia, 2006.
[8] N. Bruti-Liberati and E. Platen, *Strong approximations of stochastic differential equations with jumps*, J. Comput. Appl. Math., 205 (2007), pp. 982–1001.
[9] N. Bruti-Liberati and E. Platen, *Approximation of jump diffusions in finance and economics*, Comput. Econom., 29 (2007), pp. 283–312.
[10] P. Carr, H. Geman, D.B. Madan, and M. Yor, *The fine structure of asset returns: An empirical investigation*, J. Business, 75 (2002), pp. 305–332.
[11] R. Cont and P. Tankov, *Financial Modelling with Jump Processes*, Chapman Hall/CRC Financ. Math. Ser., Chapman & Hall/CRC, Boca Raton, FL, 2004.
[12] R. Cont and P. Tankov, *Non-parametric calibration of jump-diffusion option pricing models*, J. Comput. Finance, 7 (2002), pp. 1–49.
[13] D. Duffie, D. Filipović, and W. Schachermayer, *Affine processes and applications in finance*, Ann. Appl. Probab., 13 (2003), pp. 984–1053.
[14] R. Durett, *Probability: Theory and Examples*, 3rd ed., Duxbury Advanced Series, Brooks/Cole-Thomson Learning, Belmont, CA, 2005.
[15] A. Dzougoutov, K.-S. Moon, E. von Schwerin, A. Szepessy, and R. Tempone, *Adaptive Monte Carlo algorithms for stopped diffusion*, in Multiscale Methods in Science and Engineering, B. Enguist, P. Lötstedt, and O. Runborg, eds., Lect. Notes Comput. Sci. Eng. 44, Springer-Verlag, Berlin, 2005, pp. 59–88.
[16] K. Eriksson, D. Estep, P. Hansbo, and C. Johnson, *Introduction to adaptive methods for differential equations*, Acta Numer., 4 (1995), pp. 105–158.
[17] E.F. Fama, *Portfolio analysis in a stable Paretian market*, Management Sci., 11 (1965), pp. 404–419.
[18] G.S. Fishman, *Monte Carlo. Concepts, Algorithms and Applications*, Springer Ser. Oper. Res., Springer-Verlag, New York, 1996.
[19] I.I. Gikhman and A.V. Skorokhod, *The Theory of Stochastic Processes* III, Monographs in Probability Theory and Mathematical Statistics, Nauka, Moscow, 1975 (in Russian).
[20] P. Glasserman and N. Merener, *Numerical solution of jump-diffusion LIBOR market models*, Finance Stoch., 7 (2003), pp. 1–27.
[21] P. Glasserman and N. Merener, *Convergence of a discretization scheme for jump-diffusion processes with state-dependent intensities*, Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 460 (2004), pp. 111–127.
[22] D.J. Higham and P.E. Kloeden, *Numerical methods for nonlinear stochastic differential equations with jumps*, Numer. Math., 101 (2005), pp. 101–119.
[23] D.J. Higham and P.E. Kloeden, *Convergence and stability of implicit methods for jump-diffusion systems*, Int. J. Numer. Anal. Model., 3 (2006), pp. 125–140.
[24] D.J. Higham and P.E. Kloeden, *Strong convergence rates for backward Euler on a class of nonlinear jump-diffusion problems*, J. Comput. Appl. Math., 205 (2007), pp. 949–956.
[25] N. Ikeda and S. Watanabe, *Stochastic Differential Equations and Diffusion Processes*, 2nd ed., North-Holland Math. Library 24, North-Holland, Amsterdam, 1989.
[26] K. Itô, *Essentials of Stochastic Processes*, Transl. Math. Monogr. 231, translated from the 1957 Japanese original by Y. Itô, AMS, Providence, RI, 2006.
[27] J. Jacod and A.N. Shiryaev, *Limit Theorems for Stochastic Processes*, Grundlehren Math. Wiss. 288, Springer-Verlag, Berlin, 1987.
[28] C. Johnson, *Error estimates and adaptive time-step control for a class of one-step methods for stiff ordinary differential equations*, SIAM J. Numer. Anal., 25 (1988), pp. 908–926.
[29] C. Johnson and A. Szepessy, *Adaptive finite element methods for conservation laws based on a posteriori error estimates*, Comm. Pure Appl. Math., 48 (1995), pp. 199–234.
[30] I. Karatzas and S.E. Shreve, *Brownian Motion and Stochastic Calculus*, Grad. Texts in Math. 113, Springer-Verlag, New York, 1988.

[31] P.E. Kloeden and E. Platen, *Numerical Solution of Stochastic Differential Equations*, Appl. Math. 23, Springer-Verlag, New York, 1992.
[32] K. Kubilius and E. Platen, *Rate of weak convergence of the Euler approximation for diffusion processes with jumps*, Monte Carlo Methods Appl., 8 (2002), pp. 83–96.
[33] X.Q. Liu and C.W. Li, *Weak approximations and extrapolations of stochastic differential equations with jumps*, SIAM J. Numer. Anal., 37 (2000), pp. 1747–1767.
[34] B. Mandelbrot, *The variation of certain speculative prices*, J. Business, 36 (1963), pp. 394–419.
[35] G. Maruyama, *Continuous Markov processes and stochastic equations*, Rend. Circ. Mat. Palermo, 4 (1955), pp. 48–90.
[36] A.-M. Matache, T. von Petersdorff, and C. Schwab, *Fast deterministic pricing of options on Lévy driven assets*, Math. Model. Numer. Anal., 38 (2004), pp. 37–71.
[37] R.C. Merton, *Option pricing when the underlying stock returns are discontinuous*, J. Financial Econom., 3 (1976), pp. 125–144.
[38] R. Mikulevičius and E. Platen, *Time discrete Taylor approximations for Itô processes with jump component*, Math. Nachr., 138 (1988), pp. 99–104.
[39] G.N. Milstein and M.V. Tretyakov, *Stochastic Numerics for Mathematical Physics*, Sci. Comput., Springer-Verlag, Berlin, Heidelberg, 2004.
[40] K.-S. Moon, E. von Schwerin, A. Szepessy, and R. Tempone, *An adaptive algorithm for ordinary, stochastic and partial differential equations*, in Recent Advances in Adaptive Computation, Z.-C. Shi, Z. Chen, T. Tang, and D. Yu, eds., Contemp. Math. 383, AMS, Providence, RI, 2005, pp. 325–343.
[41] K.-S. Moon, E. von Schwerin, A. Szepessy, and R. Tempone, *Convergence rates for an adaptive dual weighted residual finite element algorithm*, BIT, 46 (2006), pp. 367–407.
[42] K.-S. Moon, A. Szepessy, R. Tempone, and G.E. Zouraris, *Hyperbolic differential equations and adaptive numerics*, in Theory and Numerics of Differential Equations (Durham, 2000), J.F. Blowey, J.P Coleman, and A.W. Craig, eds., Universitext, Springer-Verlag, Berlin, 2001, pp. 231–280.
[43] K.-S. Moon, A. Szepessy, R. Tempone, and G.E. Zouraris, *A variational principle for adaptive approximation of ordinary differential equations*, Numer. Math., 96 (2003), pp. 131–152.
[44] K.-S. Moon, A. Szepessy, R. Tempone, and G.E. Zouraris, *Convergence rates for adaptive approximation of ordinary differential equations*, Numer. Math., 96 (2003), pp. 99–129.
[45] K.-S. Moon, A. Szepessy, R. Tempone, and G.E. Zouraris, *Convergence rates for adaptive weak approximation of stochastic differential equations*, Stochastic Anal. Appl., 23 (2005), pp. 511–558.
[46] D. Nualart, *The Malliavin Calculus and Related Topics*, Probab. Appl., Springer-Verlag, New York, 1995.
[47] E. Platen, *An approximation method for a class of Itô processes with jump component*, Litovsk. Mat. Sb., 22 (1982), pp. 124–136.
[48] W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery, *Numerical Recipes in FORTRAN. The Art of Scientific Computing*, 2nd ed., Cambridge University Press, Cambridge, UK, 1992.
[49] P. Protter, *Stochastic Integration and Differential Equations. A New Approach*, Appl. Math. 21, Springer-Verlag, Berlin, 1990.
[50] P. Protter and D. Talay, *The Euler scheme for Lévy driven stochastic differential equations*, Ann. Probab., 25 (1997), pp. 393–423.
[51] A. Szepessy, R. Tempone, and G.E. Zouraris, *Adaptive weak approximation of stochastic differential equations*, Comm. Pure Appl. Math., 54 (2001), pp. 1169–1214.
[52] D.W. Strook, *Lectures on Topics in Stochastic Differential Equations*, TATA Institute of Fundamental Research, Bombay, India, 1982.
[53] D. Talay and L. Tubaro, *Expansion of the global error for numerical schemes solving stochastic differential equations*, Stochastic Anal. Appl., 8 (1990), pp. 483–509.

SIAM J. NUMER. ANAL.
Vol. 46, No. 4, pp. 1769–1779
© 2008 Society for Industrial and Applied Mathematics

FINITE ELEMENT RUNGE–KUTTA DISCRETIZATIONS OF POROUS MEDIUM–TYPE EQUATIONS*

ESKIL HANSEN† AND ALEXANDER OSTERMANN†

Abstract. In this paper we analyze the convergence properties of full discretizations of a class of generalized porous medium equations. For the spatial and time discretizations, we use continuous piecewise linear finite elements and algebraically stable Runge–Kutta methods, respectively. We prove our convergence result without any assumption on the *spatial regularity*. It is shown that, under a certain stability assumption, the *temporal order* of convergence is given by the stage order of the method, whereas the spatial order is essentially one. Numerical experiments illustrate our stability assumption and the convergence result.

Key words. porous medium equation, degenerate parabolic problems, Runge–Kutta time discretization, high order convergence in time

AMS subject classifications. 65M12, 65L06, 35K65

DOI. 10.1137/070680953

1. Introduction.

Consider the equation

$$\begin{cases} \dot{u} = \Delta\alpha(u) & \text{in } \Omega \times (0,T], \\ \alpha(u) = 0 & \text{on } \partial\Omega \quad \text{and } u(0) = \eta \text{ in } \Omega, \end{cases} \tag{1.1}$$

where Ω is an open and convex polyhedral domain in $\mathbb{R}^d$, with $d = 1$, 2, or 3. We assume that $\alpha : \mathbb{R} \to \mathbb{R}$ is a continuous and strictly increasing map with range $R(\alpha) = \mathbb{R}$, and $\alpha(0) = 0$. The classical example of (1.1) is the porous medium equation (PME),

$$\dot{u} = \Delta|u|^{m-1}u, \quad m > 1,$$

which is found when modeling the flow of an ideal gas through a homogeneous porous medium, where the unknown u represents the (normalized) gas density. Other applications can be found, for example, in nonlinear heat transfers and population dynamics. For further details regarding the applications and the mathematical theory of porous medium–type equations, we refer to the recent book by Vázquez [14].

In order to characterize the nature of the problem class, we temporarily restrict our attention to the PME. Rewriting the equation in divergence form,

$$\dot{u} = \nabla \cdot (m|u|^{m-1}\nabla u),$$

reveals a nonlinear diffusion process where the diffusion coefficient $m|u|^{m-1}$ may vanish, i.e., the PME is a degenerate parabolic equation. Another feature of the PME is that an initial value η with compact support in Ω evolves with a finite speed of propagation. Quite a few families of explicit solutions have been derived for the PME

*Received by the editors January 25, 2007; accepted for publication (in revised form) January 14, 2008; published electronically April 11, 2008. This work was supported by the Austrian Science Fund under grant M961-N13.

http://www.siam.org/journals/sinum/46-4/68095.html

†Institut für Mathematik, Universität Innsbruck, Technikerstraße 13, A-6020 Innsbruck, Austria (eskil.hansen@uibk.ac.at, alexander.ostermann@uibk.ac.at).

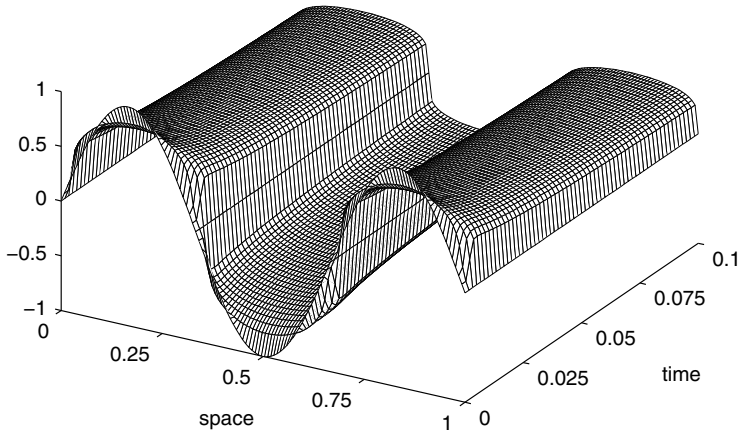


FIG. 1.1. *The evolution of the initial value* $\eta(x) = \sin(3\pi x)$ *in time under the PME with* $m = 5$.

in the context of initial data with compact support, and these solutions usually lack high regularity in space and time.

Due to the lack of regularity, the numerical studies concerning porous medium–type equations (or even more general problems) have predominantly considered discretizations based on continuous piecewise linear finite elements and the implicit Euler method [3, 7, 11, 12, 13]. For these discretizations, Rulla and Walkington [13] have derived an (essentially) optimal convergence of the form $\mathcal{O}(h + k)$ in the space $L^\infty(0, T : H^{-1}(\Omega))$, where h is the mesh width of the finite elements and k is the time step size. For studies of other spatial discretizations, see Eymard, Hilhorst, and Vohralík [8] and the references therein.

Numerical experiments, however, clearly indicate that the *time regularity* is improved for a large class of initial values which are sufficiently regular and do not vanish on open subsets of Ω. For example, consider the evolution presented in Figure 1.1. Here, the gradient of the evolution ∇u is most likely unbounded, whereas the evolution seems to be smooth in time. The same behavior can also be found in higher dimensional problems; see Figure 6.1. In the presence of higher time regularity, one may apply high order time-stepping schemes, e.g., Runge–Kutta methods, which enables the usage of large time steps k without compromising the accuracy of the discretization.

The aim of the present study is to derive an optimal convergence analysis for discretizations based on linear finite elements and high order implicit Runge–Kutta methods. An outline of the paper is as follows: In section 2 we give an appropriate H^{-1} framework for analyzing (1.1). The analogous framework for the spatial discretization is introduced in section 3. In section 4 we investigate the convergence properties of the spatial discretization. The main ingredient of the analysis is to consider auxiliary implicit Euler discretizations in combination with the results of [6, 13]. The convergence of the full discretization is then analyzed in section 5. The analysis will be carried out in a discrete H^{-1} norm in space. Under a certain stability assumption on the spatial discretization, we derive an (essentially) optimal convergence result of the form $\mathcal{O}(h + k^q)$, with q denoting the stage order of the Runge–Kutta method. The time convergence analysis employs results from the first author's paper

[10]. We conclude by proposing an implementation in section 6 which validates our assumptions and the derived convergence orders.

Henceforth, C denotes a generic constant that assumes different values at different occurrences. The inner product of a Hilbert space H is denoted by $(\cdot,\cdot)_H$ and the duality pairing between H and H' is referred to as $\langle\cdot,\cdot\rangle$.

2. Problem setting. We will interpret the porous medium–type equation (1.1) in an H^{-1} framework. For the convenience of the reader, we summarize some basic properties of this space. Recall that $H^{-1}(\Omega)$ is the dual space of $H_0^1(\Omega)$ and the inclusions $H_0^1(\Omega) \subset L^2(\Omega) = L^2(\Omega)' \subset H^{-1}(\Omega)$ are continuous. The related Riesz map $\mathcal{R} : H_0^1(\Omega) \to H^{-1}(\Omega)$ is the isomorphism defined by

$$\langle \mathcal{R}u, v\rangle := (u,v)_{H_0^1(\Omega)} = \int_\Omega \nabla u \cdot \nabla v \, \mathrm{d}x;$$

hence, $\mathcal{R}$ is the Dirichlet Laplacian $-\Delta$ on $H_0^1(\Omega)$. The inner product on $H^{-1}(\Omega)$ can be defined as

$$(u,v)_{H^{-1}(\Omega)} := (\mathcal{R}^{-1}u, \mathcal{R}^{-1}v)_{H_0^1(\Omega)} = \langle u, \mathcal{R}^{-1}v\rangle = \langle \mathcal{R}^{-1}u, v\rangle.$$

In the present H^{-1} context the map $\Delta\alpha$, with domain

$$D(\Delta\alpha) := \{u \in H^{-1}(\Omega) \cap L^1(\Omega) : \alpha(u) \in H_0^1(\Omega)\},$$

has the following properties

(i) $(\Delta\alpha(u) - \Delta\alpha(v), u - v)_{H^{-1}(\Omega)} \le 0$ for all u and v in $D(\Delta\alpha)$;
(ii) $R(I - \lambda\Delta\alpha) = H^{-1}(\Omega)$ for all $\lambda > 0$.

In other words, the map $-\Delta\alpha$ is maximal monotone, and the following existence result is obtained from nonlinear semigroup theory.

LEMMA 2.1. *For any initial value $\eta \in D(\Delta\alpha)$ there exists a unique function $u \in W^{1,\infty}(0,T : H^{-1}(\Omega))$ satisfying*

1. *$u(t) \in D(\Delta\alpha)$ for all $t \in [0,T]$;*
2. *$\dot{u} = \Delta\alpha(u)$ a.e. on $(0,T)$;*
3. *$u(0) = \eta$.*

These results are given in [5]; for a contemporary reference see [2, Proposition 2.2.12 and Theorem 4.1.4]. For the forthcoming error analysis, we assume the following.

Assumption 1. The map $\alpha^{-1} : \mathbb{R} \to \mathbb{R}$ satisfies the bound

$$|\alpha^{-1}(x)| \le C(1 + |x|^r),$$

where C is a positive constant and the parameter r is in the interval $(0,\infty)$ when $\Omega \subset \mathbb{R}^d$, with $d \le 2$, and in the interval $(0,1]$ when $\Omega \subset \mathbb{R}^3$.

This assumption is fulfilled for the PME, where $r = 1/m$, and it implies the inclusions $D(\Delta\alpha) \subset L^p(\Omega)$ for all $p \in [1,\infty)$ when $\Omega \subset \mathbb{R}^d$, with $d \le 2$, and $D(\Delta\alpha) \subset L^6(\Omega)$ when $\Omega \subset \mathbb{R}^3$. The stated inclusions follow by the observations

$$\|u\|_{L^{p/r}(\Omega)} \le C\|1 + |\alpha(u)|^r\|_{L^{p/r}(\Omega)} \le C(1 + \|\alpha(u)\|^r_{L^p(\Omega)})$$

and $\alpha(u) \in H_0^1(\Omega) \subset L^p(\Omega)$, with p given by the Sobolev embedding theorem [1]. Note that the upper bound of r in the three-dimensional case is not optimal. However, we omit these considerations in order to keep the technicalities to a minimum.

3. Spatial discretization. In preparation for the spatial discretization of (1.1), we introduce the standard finite element space $S_h \subset H_0^1(\Omega)$ of continuous piecewise linear functions on Ω, with h denoting the maximum diameter of the related tetrahedral partition; see [4] for further details. As a direct analogue to the continuous case, we identify the Riesz map $\mathcal{R}_h : S_h \to S_h'$ with the discrete Laplacian $-\Delta_h$, where

$$\langle -\Delta_h u, v\rangle = (u, v)_{S_h} = \int_\Omega \nabla u \cdot \nabla v \, \mathrm{d}x.$$

Furthermore, we define the inner product of S_h' as

$$(u, v)_{S_h'} := (\mathcal{R}_h^{-1} u, \mathcal{R}_h^{-1} v)_{S_h} = \langle u, \mathcal{R}_h^{-1} v\rangle = \langle \mathcal{R}_h^{-1} u, v\rangle.$$

Note that $u \in H^{-1}(\Omega)$ can be interpreted as an element in S_h', by mere restriction, and that $u_h \in S_h$ can be seen as an element in both S_h' and $H^{-1}(\Omega)$, via the construction $(u_h, \cdot)_{L^2(\Omega)}$. Next, consider the map $\Delta_h \alpha$, with domain

$$D(\Delta_h \alpha) := \{u \in L^2(\Omega) : \alpha(u) \in S_h\},$$

and the spatial discretization of (1.1),

$$\dot{u}_h = \Delta_h \alpha(u_h), \quad u_h(0) = \eta_h \in D(\Delta_h \alpha). \tag{3.1}$$

Lemma 2.1 is in fact valid for an arbitrary maximal monotone map on a Hilbert space (see [2, Theorem 4.1.4]). Therefore, the very same existence result, as obtained for the original equation (1.1), can be derived for (3.1) by the lemma below.

LEMMA 3.1. *The map $-\Delta_h \alpha$ is maximal monotone.*

Proof. (i) Monotonicity: As $\alpha : \mathbb{R} \to \mathbb{R}$ is an increasing function, it follows that

$$\begin{aligned}(\Delta_h \alpha(u) - \Delta_h \alpha(v), u - v)_{S_h'} &= -\langle \alpha(u) - \alpha(v), u - v\rangle \\ &= -(\alpha(u) - \alpha(v), u - v)_{L^2(\Omega)} \le 0\end{aligned}$$

for all u and v in $D(\Delta_h \alpha)$.

(ii) Range condition: In order to prove that $R(I - \lambda \Delta_h \alpha) = S_h'$ for all $\lambda > 0$, consider the map $\alpha^{-1} - \lambda \Delta_h : S_h \to S_h'$ which is *hemicontinuous*, i.e., $t \mapsto \langle (\alpha^{-1} - \lambda \Delta_h)(u + tv), w\rangle$ is continuous on $[0, 1]$ for all u, v, and w in S_h, and *strongly monotone*, i.e.,

$$\big\langle (\alpha^{-1} - \lambda \Delta_h)(u) - (\alpha^{-1} - \lambda \Delta_h)(v), u - v\big\rangle \ge \lambda \|u - v\|_{S_h}^2$$

for all u and v in S_h. Hence, it follows by the Browder and Minty theorem (see [15, Theorem 26.A]) that $R(\alpha^{-1} - \lambda \Delta_h) = S_h'$, and the range condition is then implied by the identity $I - \lambda \Delta_h \alpha = (\alpha^{-1} - \lambda \Delta_h)\alpha$. ∎

4. Spatial error analysis. The aim of this section is to bound the spatial error, $u(t) - u_h(t)$, generated when passing from (1.1) to (3.1). The key idea of the analysis is to introduce the auxiliary implicit Euler discretizations

$$u^n := (I - \tau \Delta \alpha)^{-n}(\eta) \quad \text{and} \quad u_h^n := (I - \tau \Delta_h \alpha)^{-n}(\eta_h)$$

of (1.1) and (3.1), respectively, where $\tau := t/n$, $u^n \approx u(t)$, and $u_h^n \approx u_h(t)$. The influence of the spatial discretization can then be isolated to a time-independent term, $u^n - u_h^n$, by the decomposition

$$u(t) - u_h(t) = (u(t) - u^n) + (u^n - u_h^n) + (u_h^n - u_h(t)).$$

Before we proceed with the analysis, we again recapitulate a well-known result from nonlinear semigroup theory.

LEMMA 4.1. *Consider the evolution equation*

$$\dot{u} = f(u), \quad u(0) = \eta \in D(f),$$

on the (real) Hilbert space H, and its implicit Euler discretization $u^n = (I - \tau f)^{-n}(\eta)$. If the map $-f$ is maximal monotone, then

$$\|u(t) - u^n\|_H \le C\,\tau^{1/2}\,\|f(\eta)\|_H,$$

with a uniform constant C on the time interval $[0, T]$.

This result is due to [6]. Bounds of the term $u^n - u_h^n$ can also be found in the literature [13].

LEMMA 4.2. *If the initial values η and η_h satisfy $\eta \in L^\infty(\Omega)$, $\alpha(\eta) \in W^{2,1}(\Omega)$, $\|\alpha(\eta_h)\|_{S_h} \le C$, $\|\eta - \eta_h\|_{S_h'} \le Ch$, and Assumption 1 is valid, then*

$$\|u^n - u_h^n\|_{S_h'} \le C\varphi_d(\tau, h)h,$$

where $\varphi_d(\tau, h) = 1 + |\log h|^{1+r/2}$ for $d \le 2$ and $\varphi_3(\tau, h) = \tau^{-r/12}$.

Note that Lemma 4.2 is stated in our problem setting, but the results follow along the lines of the proofs presented in [13], where the one- and two-dimensional results are given in Lemma 2.6, and the three-dimensional result is stated in the remark following Theorem 2.7.

Example 4.3. Assume that the initial value η is Lipschitz continuous. The above hypotheses on η_h are then valid if $\eta_h := \alpha^{-1}(I_h\alpha(\eta))$, where I_h is the linear interpolation operator. Note that in this case the bound of $\eta - \eta_h$ holds even in $L^\infty(\Omega)$. To this end, consider a tetrahedron K of the partition with nodes $\{x_i\}$. As $I_h\eta_h = I_h\eta$ and $\eta_h|_K$ is the image of a linear function under a monotone map, we have the bound $|\eta(x_i) - \eta_h(x)| \le \max_j |\eta(x_i) - \eta(x_j)|$ for all x in K. Hence,

$$|\eta(x) - \eta_h(x)| \le |\eta(x) - \eta(x_i)| + |\eta(x_i) - \eta_h(x)| \le 2\,\mathrm{Lip}(\eta)h$$

for all x in K, and consequently $\|\eta - \eta_h\|_{L^\infty(\Omega)} \le C(\eta)h$.

Combining these results yields the following spatial error bound.

THEOREM 4.4. *Consider the solution u of the porous medium–type equation (1.1) and its spatial approximation u_h, given by (3.1). If $\eta \in L^\infty(\Omega)$, $\alpha(\eta) \in W^{2,1}(\Omega)$, $\|\alpha(\eta_h)\|_{S_h} \le C$, $\|\eta - \eta_h\|_{S_h'} \le Ch$, and Assumption 1 is valid, then*

$$\|u(t) - u_h(t)\|_{S_h'} \le C\phi_d(h)h,$$

where $\phi_d(h) = 1 + |\log h|^{1+r/2}$ for $d \le 2$ and $\phi_3(h) = h^{-r/(6+r)}$. The constant C is uniform in the time interval $[0, T]$.

Proof. By Lemmas 4.1 and 4.2, one obtains the inequality

$$\begin{aligned}\|u(t) - u_h(t)\|_{S_h'} &\le \|u(t) - u^n\|_{H^{-1}(\Omega)} + \|u^n - u_h^n\|_{S_h'} + \|u_h^n - u_h(t)\|_{S_h'} \\ &\le C\left(\tau^{1/2} + \varphi_d(\tau, h)h + \tau^{1/2}\right).\end{aligned}$$

The sought after bound now follows by setting $\tau = h^2$ when $\Omega \subset \mathbb{R}^d$, with $d \le 2$, or $\tau = h^{12/(6+r)}$ when $\Omega \subset \mathbb{R}^3$. ∎

The bound in Theorem 4.4 is close to the optimal form $\mathcal{O}(h)$, as the term $\phi_d(h)$ will not be observable in numerical experiments, e.g., for the three-dimensional PME the term becomes $\phi_3(h) = h^{-1/(6m+1)}$.

5. Full spatial time discretization. In order to obtain a full spatial time approximation of (1.1), we discretize (3.1) in time by an implicit Runge–Kutta method. The main issue when bounding the related error is that the bound of the time discretization must be independent of the stiffness of the problem. This prevents the usage of explicit integrators and the classic ODE theory. We will instead use implicit methods and employ the uniform results of [10].

For fixed h and a sufficiently regular α, (3.1) is a smooth ODE on a finite-dimensional space. Consequently, its solution u_h is smooth with bounded derivatives. For the convergence analysis we have in mind, however, it is essential that we can control the size of certain time derivatives of u_h.

The uniform boundedness of these derivatives cannot be guaranteed, in general. A typical example is given by the Barenblatt solution of the PME (see [14, section 4.4]) for which the time derivatives of u_h grow as $h \to 0$ and become unbounded in the limit. This implies that the error constant of the time discretization grows as $h \to 0$; see the Barenblatt evolution (6.7) in Table 6.2 below.

Extensive numerical experiments, however, indicate that the situation is often more favorable. For example, the evolutions (6.5), (6.6), and (6.8), considered in our simulations in section 6, clearly show that the error constant of the time discretization remains bounded. Even though the current analytic theory does not provide tools to characterize an explicit class of initial values guaranteeing the sought after regularity, the convincing numerical results motivate the following sufficient stability assumption.

Assumption 2. The function $u_h \in C^{q+1}([0,T] : S_h')$ satisfies (3.1) on the interval $(0,T]$ and $\|u_h^{(q+1)}\|_{L^\infty(0,T:S_h')}$ is uniformly bounded in h.

Note that this assumption is of a theoretical nature as some growth of the time derivatives in the error constant can always be compensated for by relating the time step to the spatial grid size.

The s-stage Runge–Kutta approximation u_h^{n+1} of $u_h(t_{n+1})$ is defined by the S_h' difference equation

$$\begin{cases} U^{ni} = u_h^n + k\sum_{j=1}^s a_{ij}\Delta_h\alpha(U^{nj}), \quad i = 1,\dots,s, \\ u_h^{n+1} = u_h^n + k\sum_{i=1}^s b_i\Delta_h\alpha(U^{ni}), \end{cases} \tag{5.1}$$

where $k = T/N$, $t_n = kn$ for $n = 0,\dots,N$, and $u_h^0 = \eta_h$. The stage values U^{ni} approximate the solution at times $t_n + c_i k$ with $c_i = \sum_{j=1}^s a_{ij}$. For further details on Runge–Kutta methods, we refer to [9].

One can characterize a set of matrices $A = \{a_{ij}\}$ for which (5.1) has a unique solution, as we prove in [10, Theorem 5.2]. We recall this result for the convenience of the reader.

LEMMA 5.1. *If the matrix A is invertible and there exists a positive diagonal matrix D such that DA^{-1} is positive definite, then there exists a unique solution $\{U^{ni}\}_{i=1}^s$ of* (5.1).

Implicit Runge–Kutta methods are referred to as *algebraically stable*, and with *stage order* $q \geq 1$, whenever $b_i \geq 0$, the matrix $S = \{b_i a_{ij} + b_j a_{ji} - b_i b_j\}$ is positive semidefinite, and the coefficients satisfy the order conditions

$$\sum_{i=1}^s b_i c_i^{k-1} = 1/k \quad \text{and} \quad \sum_{j=1}^s a_{ij} c_j^{k-1} = c_i^k/k \quad \text{for} \quad k = 1,\dots,q.$$

For these methods one has the following analogue to the global error bounds obtained for stiff ODEs (compare, e.g., with the results of [9, Theorem 15.3]).

LEMMA 5.2. *If there exists a positive diagonal matrix D such that DA^{-1} is positive definite, and Assumption 2 is valid, then*

$$\|u_h(t_n) - u_h^n\|_{S_h'} \le Ck^q,$$

where the constant C is uniform in the time interval $[0,T]$.

The proof is given in [10, Corollary 6.3].

Example 5.3. Some examples of algebraically stable Runge–Kutta methods which satisfy the assumptions imposed on the matrix A in Lemma 5.2 are the Gauss and Radau IIA methods. The explicit expressions of the optimal D matrices are summarized in [9, section IV.14].

We are now able to derive the following bound for the error of the full discretization by combining Theorem 4.4 and Lemma 5.2.

THEOREM 5.4. *Consider the solution u of the porous medium–type equation (1.1) and its full finite element Runge–Kutta approximation u_h^n, given by (5.1). If $\eta \in L^\infty(\Omega)$, $\alpha(\eta) \in W^{2,1}(\Omega)$, $\|\alpha(\eta_h)\|_{S_h} \le C$, $\|\eta - \eta_h\|_{S_h'} \le Ch$, the Runge–Kutta method is algebraically stable with stage order q, there exists a positive diagonal matrix D such that DA^{-1} is positive definite, and Assumptions 1 and 2 are valid, then*

$$\|u(t_n) - u_h^n\|_{S_h'} \le C(\phi_d(h)h + k^q),$$

where $\phi_d(h) = 1 + |\log h|^{1+r/2}$ for $d \le 2$ and $\phi_3(h) = h^{-r/(6+r)}$. The constant C is uniform in the time interval $[0,T]$. ∎

6. Numerical experiments. We finally propose an implementation for the full spatial time discretization and present a series of numerical experiments that validate the regularity assumptions as well as the derived convergence orders.

To ease the notation, we start off by considering the full discretization based on the implicit Euler method,

$$u_h^{n+1} = u_h^n + k\Delta_h \alpha(u_h^{n+1}).$$

The only computable quantity is $\alpha(u_h^{n+1}) \in S_h$, and we therefore rewrite the discretization as

$$\alpha^{-1}(v_h^{n+1}) = \alpha^{-1}(v_h^n) + k\Delta_h v_h^{n+1},$$

where $v_h^n := \alpha(u_h^n) \in S_h$. Introduce a basis $\{\psi_i\}_{i=1}^M$ of S_h and let $v_h^n = \sum_{j=1}^M d_j^n \psi_j$. The discretization is then equivalent to

$$g(d^{n+1}) = g(d^n) - kKd^{n+1}, \tag{6.1}$$

where $d^n = (d_1^n, \dots, d_M^n)^{\mathrm{T}}$, $K = \{(\nabla\psi_j, \nabla\psi_i)_{L^2(\Omega)}\}$ is the stiffness matrix, and

$$g(d^n)_i = \int_\Omega \alpha^{-1}\Big(\textstyle\sum_{j=1}^M d_j^n \psi_j\Big)\,\psi_i \,\mathrm{d}x, \quad i = 1, \dots, M.$$

Next, choose $\{\psi_i\}$ as the standard basis related to linear Lagrange elements [4]. Approximating the integrals in (6.1) by the nodal quadrature rule, i.e., the multidimensional counterpart of the trapezoidal rule, and changing variables, $y^n := \alpha^{-1}(d^n)$, gives the scheme

$$Dy^{n+1} = Dy^n - kK\alpha(y^{n+1}),$$

where the diagonal matrix D is the so-called lumped mass matrix. (The application of α and α^{-1} to a vector is understood componentwise.) Hence, the full discretization u_h^n is approximated by $\tilde{u}_h^n := \alpha^{-1}(\sum_{i=1}^M \alpha(y_i^n)\psi_i)$.

The very same considerations can be made for full discretizations based on stiffly accurate Runge–Kutta methods, i.e., $a_{sj} = b_j$ and $c_s = 1$. A typical example is given by the class of Radau IIA methods. For stiffly accurate methods, the scheme becomes

$$\begin{cases} DY^{ni} = Dy^n - k\sum_{j=1}^s a_{ij} K\alpha(Y^{nj}), \quad i = 1, \dots, s, \\ y^{n+1} = Y^{ns}. \end{cases} \tag{6.2}$$

In the numerical experiments below, where the PME is chosen as the model problem, the map α is differentiable, and we then propose using Newton's method when approximating the solution to (6.2).

Throughout the experiments we will employ the three-stage Radau IIA method and with the discrete initial value η_h chosen as in Example 4.3. In order to approximate the discretization error in S_h', we consider the linear interpolation of the numerical solution $I_h\tilde{u}_h^n = \sum_{i=1}^M y_i^n \psi_i$. For the one- and two-dimensional settings we have the estimate

$$\begin{aligned} \|\tilde{u}_h^n - I_h\tilde{u}_h^n\|_{S_h'} &\le C(1 + |\log h|)\, \|\tilde{u}_h^n - I_h\tilde{u}_h^n\|_{L^1(\Omega)} \\ &\le C(1 + |\log h|)h^d \sum_{\ell=1}^M \max_{K_\ell} |\tilde{u}_h^n(x_i) - \tilde{u}_h^n(x_j)|, \end{aligned}$$

where d is the spatial dimension and K_ℓ denotes the ℓth tetrahedron of the partition with nodes $\{x_i\}$. In the context of the PME, one observes a spatially smooth solution except in regions where the equation degenerates. Hence, the differences $\tilde{u}_h^n(x_i) - \tilde{u}_h^n(x_j)$ are (at least) of the form $\mathcal{O}(h)$ in partitions where the solution is smooth and are of the form $\mathcal{O}(1)$ in partitions where the solution jumps. Furthermore, the solution will degenerate in a fixed number of points or curves and the number of degenerate partitions is therefore proportional to h^{-d+1}. This implies the relation

$$\sum_{\ell=1}^M \max_{K_\ell} |\tilde{u}_h^n(x_i) - \tilde{u}_h^n(x_j)| = \mathcal{O}(h^{-d+1}), \tag{6.3}$$

and we then obtain that $\|\tilde{u}_h^n - I_h\tilde{u}_h^n\|_{S_h'} = \mathcal{O}(|\log h|\, h)$. Note that the sum in (6.3) is dependent only on the nodal values of the numerical solution, and it is easily checked in the simulation runs that the equality (6.3) is valid for our experiments. If we finally compute a reference solution $\tilde{u}_{h_r}^{N_r}$ by solving (6.2) with sufficiently small parameters (h_r, k_r), then the global error can be approximated as

$$\|u(T) - u_h^N\|_{S_h'} \approx \|\tilde{u}_{h_r}^{N_r} - \tilde{u}_h^N\|_{S_h'} \le \|I_{h_r}\tilde{u}_{h_r}^{N_r} - I_h\tilde{u}_h^N\|_{S_h'} + \mathcal{O}(|\log h|\, h),$$

where the second term from the right is explicitly computable:

$$\|I_{h_r}\tilde{u}_{h_r}^{N_r} - I_h\tilde{u}_h^N\|_{S_h'} = (I_{h_r}\tilde{u}_{h_r}^{N_r} - I_h\tilde{u}_h^N, -\Delta_h^{-1}(I_{h_r}\tilde{u}_{h_r}^{N_r} - I_h\tilde{u}_h^N))_{L^2(\Omega)}^{1/2}. \tag{6.4}$$

To validate the convergence result of Theorem 5.4, we consider the one-dimensional PME with

$$\alpha(u) = u^5, \quad \eta(x) = \sin(3\pi x), \quad \Omega = (0, L), \quad L = 1, \quad \text{and} \quad T = 0.1. \tag{6.5}$$

TABLE 6.1
Numerically observed errors and orders for the full (three-stage Radau IIA*) discretization.*

	$h/L = 2^{-13}$		$h/L = k/T$		$h/L = 32(k/T)^2$	
h/L	error	order	error	order	error	order
1/32	1.07 E-7		6.23 E-4		6.23 E-4	
1/64	2.48 E-8	2.11	2.50 E-4	1.32	9.72 E-5	2.68
1/128	3.71 E-9	2.74	9.69 E-5	1.37	1.62 E-5	2.58
1/256	4.59 E-10	3.02	4.04 E-5	1.26	2.77 E-6	2.55
1/512	5.26 E-11	3.13	1.58 E-5	1.35	4.20 E-7	2.72
Thm. 5.4		3		1		2

The results of the simulations are presented in Table 6.1, where the scaled spatial parameter h/L was kept at a constant value of 2^{-13}, varied as k/T, or varied proportionally to k^2. As can be seen from the results, the term (6.4) displays a convergence of roughly the form $\mathcal{O}(h^{1.3} + k^3)$. The presence of the third order convergence in time indicates that there is sufficient time regularity present, and the results are indeed in agreement with the convergence of form $\mathcal{O}(h + k^3)$, predicted by Theorem 5.4.

We conclude by investigating which type of initial values yields solutions satisfying Assumption 2. One way to conduct such an investigation is to consider the time discretization error

$$\|u_h(T) - u_h^N\|_{S_h'} \approx \|I_h(\tilde{u}_h^{N_r} - \tilde{u}_h^N)\|_{S_h'} + \mathcal{O}(|\log h|\, h).$$

If the solution of the spatial discretization u_h fulfills Assumption 2, one would expect the error to remain constant as the spatial parameter h tends to zero and, similarly, one expects that $\lim_{h\to 0} \|u_h(T) - u_h^N\|_{S_h'} = \infty$ when $u_h^{(4)}$ is not uniformly bounded with respect to h. For our experiment we consider the example given in (6.5) and the evolutions

$$\alpha(u) = |u|^{2.2}u,\ \eta(x) = \sin(4.5x)(1-x),\ \Omega = (0, L),\ L = 1,\ \text{and } T = 3, \tag{6.6}$$

$$\alpha(u) = u^5,\ \eta(x) = [1 - x^2/15]_+^{1/4},\ \Omega = (-L, L),\ L = 10,\ \text{and } T = 5, \tag{6.7}$$

$$\alpha(u) = u^5,\ \eta(x, y) = \sin(3\pi x)\sin(3\pi y),\ \Omega = (0, L)^2,\ L = 1,\ \text{and } T = 0.1. \tag{6.8}$$

The first problem is chosen in order to illustrate that Assumption 2 is valid for (regular) unsymmetric initial values and noninteger values of the exponent m in the PME. The second evolution (6.7) is equipped with the initial data of the classical Barenblatt solution [14], for which the solution does not possess high time regularity, and the last evolution is the two-dimensional version of (6.5); see Figure 6.1.

The observed values of the term $\|I_h(\tilde{u}_h^{N_r} - \tilde{u}_h^N)\|_{S_h'}$ are presented in Table 6.2, where the spatial parameter was varied. The scaled time steps k/L were chosen as 2^{-6} for (6.5) and (6.6), and as 2^{-8} for (6.7) and (6.8), respectively. It is clear from the simulations that the time discretization error $\|u_h(T) - u_h^N\|_{S_h'}$ remains bounded for the evolutions (6.5), (6.6), (6.8), and grows for the Barenblatt solution (6.7); i.e., Assumption 2 is realistic for the cases with regular initial values.

Acknowledgments. We thank Juan Luis Vázquez for providing us with a preliminary version of his book [14] and Noel Walkington for a proof of the remark following Theorem 2.7 in [13].

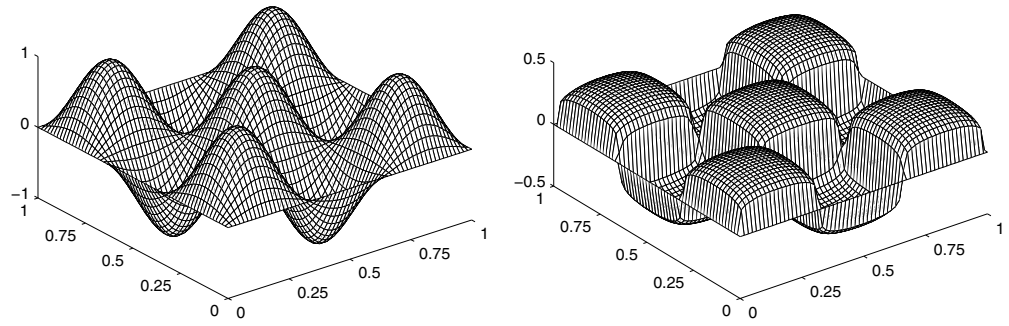


FIG. 6.1. *The evolution of the initial value* $\eta(x,y) = \sin(3\pi x)\sin(3\pi y)$ *(left) by the PME with* $m = 5$ *at time* $T = 0.1$ *(right).*

TABLE 6.2
The time discretization errors obtained when varying only the parameter h.

Evolution	(6.5)	(6.6)	(6.7)
h/L	error	error	error
1/512	2.38 E-8	3.78 E-7	8.90 E-11
1/1024	2.43 E-8	3.79 E-7	5.87 E-10
1/2048	2.46 E-8	3.74 E-7	6.15 E-9
1/4096	2.47 E-8	3.68 E-7	1.74 E-8
1/8192	2.48 E-8	3.67 E-7	6.22 E-8

Evolution	(6.8)
h/L	error
1/16	2.84 E-10
1/32	3.49 E-10
1/64	2.84 E-10
1/128	2.78 E-10

REFERENCES

[1] R. A. ADAMS, *Sobolev Spaces*, Academic Press, New York, London, 1975.
[2] V. BARBU, *Analysis and Control of Nonlinear Infinite-Dimensional Systems*, Academic Press, Boston, MA, 1993.
[3] J. W. BARRETT AND P. KNABNER, *Finite element approximation of the transport of reactive solutes in porous media*. II: *Error estimates for equilibrium adsorption processes*, SIAM J. Numer. Anal., 34 (1997), pp. 455–479.
[4] S. C. BRENNER AND L. R. SCOTT, *The Mathematical Theory of Finite Element Methods*, Springer-Verlag, New York, 1994.
[5] H. BRÉZIS, *Monotonicity methods in Hilbert spaces and some applications to nonlinear partial differential equations*, in Contributions to Nonlinear Functional Analysis (Proceedings of the Symposium, Math. Res. Center, University of Wisconsin, Madison, WI, 1971), Academic Press, New York, 1971, pp. 101–156.
[6] M. G. CRANDALL AND T. M. LIGGETT, *Generation of semi-groups of nonlinear transformations on general Banach spaces*, Amer. J. Math., 93 (1971), pp. 265–298.
[7] C. EBMEYER, *Error estimates for a class of degenerate parabolic equations*, SIAM J. Numer. Anal., 35 (1998), pp. 1095–1112.
[8] R. EYMARD, D. HILHORST, AND M. VOHRALÍK, *A combined finite volume–nonconforming/ mixed-hybrid finite element scheme for degenerate parabolic problems*, Numer. Math., 105 (2006), pp. 73–131.
[9] E. HAIRER AND G. WANNER, *Solving Ordinary Differential Equations* II. *Stiff and Differential-Algebraic Problems*, 2nd ed., Springer-Verlag, Berlin, 1996.
[10] E. HANSEN, *Runge–Kutta time discretizations of nonlinear dissipative evolution equations*, Math. Comp., 75 (2006), pp. 631–640.
[11] R. H. NOCHETTO AND C. VERDI, *Approximation of degenerate parabolic problems using numerical integration*, SIAM J. Numer. Anal., 25 (1988), pp. 784–814.
[12] I. S. POP AND W. YONG, *A numerical approach to degenerate parabolic equations*, Numer. Math., 92 (2002), pp. 357–381.

[13] J. Rulla and N. J. Walkington, *Optimal rates of convergence for degenerate parabolic problems in two dimensions*, SIAM J. Numer. Anal., 33 (1996), pp. 56–67.
[14] J. L. Vázquez, *The Porous Medium Equation. Mathematical Theory*, Oxford University Press, Oxford, 2007.
[15] E. Zeidler, *Nonlinear Functional Analysis and Its Applications* II/B. *Nonlinear Monotone Operators*, Springer-Verlag, New York, 1990.

SIAM J. NUMER. ANAL.
Vol. 46, No. 4, pp. 1780–1798
© 2008 Society for Industrial and Applied Mathematics

CONVERGENCE AND NUMERICS OF A MULTISECTION METHOD FOR SCATTERING BY THREE-DIMENSIONAL ROUGH SURFACES*

ERIC HEINEMEYER†, MARKO LINDNER‡, AND ROLAND POTTHAST‡

Abstract. We introduce a novel *multisection method* for the solution of integral equations on unbounded domains. The method is applied to the rough surface scattering problem in three dimensions, in particular to a Brakhage–Werner-type integral equation for acoustic scattering by an unbounded rough surface with Dirichlet boundary condition, where the fundamental solution is replaced by some appropriate half-space Green's function. The basic idea of the multisection method is to solve an integral equation $A\varphi = f$ by *approximately* solving the equation $P_\varrho A P_\tau \varphi = P_\varrho f$ for some positive constants ϱ, τ. Here P_ϱ is a projection operator that truncates a function to a ball with radius $\varrho > 0$. For a very general class of operators A, for which the Brakhage–Werner equation from acoustic scattering is a particular example, we will show existence of approximate solutions to the multisection equation and show that approximate solutions to the multisection equation approximate the true solution φ_0 of the operator equation $A\varphi = f$. Finally, we describe a numerical implementation of the multisection algorithm and provide numerical examples for the case of rough surface scattering in three dimensions.

Key words. rough surfaces, scattering, multisection method

AMS subject classifications. 35J05, 45P05, 31B10, 65R20

DOI. 10.1137/060673321

1. Introduction. The goal of this work is the introduction and analysis of a new numerical scheme for a class of linear operator equations

$$A\varphi = f, \tag{1}$$

where A has particular decay properties.

Model problem. The basic model problem for our work is the solution of integral equations over unbounded domains which appear in what is usually called the *rough surface scattering problem* in the engineering literature. It denotes scattering of acoustic or electromagnetic waves by a surface which is a nonlocal perturbation of an infinite plane surface. We will restrict our attention to the case where the scattering surface

$$\Gamma = \{x = (x_1, x_2, x_3) \in \mathbb{R}^3 : x_3 = f(x_1, x_2)\} \tag{2}$$

is the graph of some bounded continuous function $f : \mathbb{R}^2 \to \mathbb{R}$. The domain of wave propagation is the upper perturbed half-space

$$D := \{x = (x_1, x_2, x_3) : x_3 > f(x_1, x_2)\}. \tag{3}$$

For this work we will assume that $f \in C^{1,\alpha}$, $\alpha \in (0, 1)$, is a Lyapunov function; i.e., it is continuously differentiable with Hölder continuous first derivative.

*Received by the editors October 25, 2006; accepted for publication (in revised form) April 19, 2007; published electronically April 16, 2008.
http://www.siam.org/journals/sinum/46-4/67332.html
†Institute for Numerical Applied Mathematics, University of Göttingen, Lotzestr. 16-18, 37083 Göttingen, Germany (heinemey@math.uni-goettingen.de).
‡Department of Mathematics, University of Reading, Whiteknights, P.O. Box 220, Reading, Berkshire, RG6 6AX, UK (m.lindner@reading.ac.uk, r.w.e.potthast@reading.ac.uk).

Rough surface scattering problems arise in important applications for acoustic and electromagnetic waves. Well-known areas are outdoor sound propagation or optical scattering in nanotechnology. We refer the reader to the extensive literature reviewed in [12], [16], [15], [17], and [5]. Recently, Chandler-Wilde, Heinemeyer, and Potthast [1], [2] provided some rigorous existence theory for the integral equation approach in three dimensions. But to our knowledge there is no rigorous numerical analysis for the solution of these integral equations in three dimensions. This is the starting point of our work.

However, the approach presented here is based only on very general properties of the operator equations under consideration. We expect that it can be used for a whole range of different problems, and it will be the starting point for further research.

The finite section method. If we regard (1) as an equation on the space $Y = L^2(\mathbb{R}^2)$, for example, then the well-known *finite section method* consists in replacing (1) by

$$P_\varrho A P_\varrho \varphi = P_\varrho f, \tag{4}$$

where $\varrho > 0$ and the projection operator $P_\varrho : Y \to Y$ is given by

$$(P_\varrho \psi)(x) := \begin{cases} \psi(x), & |x| < \varrho \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

The operator P_ϱ truncates a function on $\mathbb{R}^2$ to its values inside the disk

$$B_\varrho := \{x \in \mathbb{R}^2 : |x| < \varrho\}. \tag{6}$$

In these notations $|\cdot|$ denotes an arbitrary norm in $\mathbb{R}^2$. In particular, in our numerical part in section 4 we will use the maximum norm $|x| := \max(|x_1|, |x_2|)$.

So the finite section method (4) simply restricts the right-hand side f, the solution φ, and the resulting function $A\varphi$ all to the same disk B_ϱ. The idea behind this method is, provided (1) is uniquely solvable for every right-hand side f, to hope that also (4) is uniquely solvable (now in $L^2(B_\varrho)$, of course) and that its solution φ approximates the exact solution φ_0 of (1) if only one chooses ϱ large enough. If this is the case, then this method is called applicable. There are, however, simple examples of operators A that show that the finite section method is not necessarily applicable in general.

There are recent results on the applicability of the finite section method for the fairly large class of all so-called band-dominated operators in terms of their limit operators (see [10], [13], and [14]). The operators originating from two-dimensional (2D) rough surface scattering are band-dominated, and we refer the reader to [11], [3], and further literature cited therein for the study and application of the finite section method to these equations. Although, also for three-dimensional (3D) rough surface scattering, the corresponding operators belong to this class, we will leave the finite section method aside and study what we call the multisection method. We underline that the framework for this method not only covers the set of band-dominated operators, it even applies to the much larger class, called $L(Y, \mathcal{P})$ in [10] and [13], that consists of all bounded linear operators A on Y for which

$$\|(I - P_\varrho)AP_\tau\| \to 0 \quad \text{and} \quad \|P_\tau A(I - P_\varrho)\| \to 0 \quad \text{as} \quad \varrho \to \infty \tag{7}$$

for every fixed $\tau > 0$. In fact, the multisection method to be presented here only requires the first condition in (7) to be true.

The multisection method. Already very simple examples like the shift operator

$$(A\varphi)(x) \;=\; \varphi(x-a), \qquad x \in \mathbb{R}^2,$$

on $Y = L^2(\mathbb{R}^2)$ with a fixed $a \in \mathbb{R}^2 \setminus \{0\}$ show that the finite section method fails to apply in general. Indeed, $P_\varrho A P_\varrho$ is neither surjective nor injective on $L^2(B_\varrho)$, however large ϱ is chosen. The new method proposed here will overcome some of these difficulties and will, in fact, apply to all bounded linear operators on Y that are subject to the first condition in (7).

In contrast to exactly solving the truncated equation (4), that is,

$$P_\varrho A P_\varrho \varphi \;=\; P_\varrho f,$$

for large ϱ, we propose to look for a function $\varphi \in Y$ with

$$P_\varrho A P_\tau \varphi \;\approx\; P_\varrho f \tag{8}$$

for large ϱ, τ and a given discrepancy allowance δ in the "$\approx$" sign. So we have two main differences from the finite section method:

(a) We allow two different cutoff parameters ϱ and τ instead of just one.

(b) We work with approximate instead of exact solutions.

Point (a) is the reason why we call this method the *multisection method*. From the matrix perspective it means that we cut rectangular rather than quadratic finite matrices out of the original infinite matrix that represents our operator A.

We will show that for every precision $\delta > 0$ and every sufficiently large parameter $M > 0$, there are positive constants ϱ, τ such that there exists a function $\varphi \in Y$ which satisfies the conditions

$$P_\tau \varphi = \varphi, \tag{9}$$

$$\|P_\varrho A P_\tau \varphi - P_\varrho f\|_Y \le \delta; \tag{10}$$

i.e., φ is supported on B_τ and it approximately satisfies the multisection equation (8).

In our main convergence result we show that, given any $\varepsilon > 0$, we can choose the parameters δ, ϱ, τ such that every solution φ of the multisection method (9)–(10) approximates the true solution φ_0 of the original equation (1) with

$$\|\varphi - \varphi_0\|_Y \le \varepsilon.$$

We would like to clarify the relation between the multisection method and the theory of regularization of ill-posed problems. Clearly, in general, the truncated equation $P_\varrho A P_\tau \varphi = P_\varrho f$ does not have a solution and, thus, describes an ill-posed equation. In the theory of inverse problems such equations are solved approximately by using a family R_α of bounded linear inversion operators which (for injective operators) tend pointwise to the inverse of the equation under consideration; see [6]. However, here we are not interested in the case where $\delta \to 0$ while ρ and τ are fixed, which would be the case interesting in the framework of inverse problems. Here we keep δ fixed and then choose τ, ρ sufficiently large; i.e., we change the equation which we solve.

Contents of the paper. Section 2 introduces the 3D rough surface scattering problems. Here we study the operators arising from this application and prove that they satisfy the conditions on which the multisection method is built. Section 3 contains the main analysis for the multisection method. We provide this analysis in a very general setting which reflects the potential of this method. Section 4 provides details about the algorithmical realization of the multisection method and numerical examples which demonstrate the applicability of the ideas.

2. Scattering by rough surfaces. We restrict our attention to time-harmonic acoustic waves, which are modelled by the *Helmholtz equation*

$$\triangle u + \kappa^2 u = 0. \tag{11}$$

Here, κ denotes the wave number, which for the real-valued case is linked to the speed of sound c and the frequency ω via $\kappa = \omega/c > 0$. Often, κ is admitted to be a complex number $\kappa = \kappa_0 + i\sigma$, where the imaginary part σ models the properties of some lossy medium.

For our scattering surface Γ we assume that $f \in BC^{1,\alpha}(\mathbb{R}^2)$ for some $\alpha \in (0, 1]$. Further, f is assumed to satisfy the bounds

$$0 < f^- \le f(x) \le f^+, \quad x \in \mathbb{R}^2. \tag{12}$$

We consider the scattering of an *incident acoustic wave* u^i by the surface Γ. The *total field* $u := u^i + u^s$ is the sum of the incident field and the *scattered field* u^s. The scattered field is a solution to the Helmholtz equation (11) in D. Further, we assume that u satisfies the *Dirichlet boundary condition*

$$u(x) = 0, \quad x \in \Gamma. \tag{13}$$

The scattered field is required to be bounded in D; i.e.,

$$|u^s(x)| \le c, \quad x \in D, \tag{14}$$

for some constant c. Further, we follow [2] and require that u satisfies the *limiting absorption principle* as follows. If u is considered in its dependence on the complex wave number κ, we write $u = u^{(\kappa)}$. We assume that, for all sufficiently small $\sigma > 0$, a solution of (11), (13), (14) exists and satisfies

$$u^{(\kappa_0+i\sigma)} \to u^{(\kappa_0)}, \quad \sigma \to 0, \tag{15}$$

for all $x \in D$. The limiting absorption principle is a kind of radiation condition which is needed to obtain the physically correct solution to the scattering problem.

The free-space fundamental solution in three dimensions is given by the function

$$\Phi(x, y) = \frac{1}{4\pi} \frac{e^{i\kappa|x-y|}}{|x-y|}, \quad x \ne y \in \mathbb{R}^3. \tag{16}$$

The decay of this function for $|y| \to \infty$ is very weak, and it has not been possible to use it for the solution of scattering problems over unbounded domains in three dimensions. However, in [1] Chandler-Wilde, Heinemeyer, and Potthast employed the Green's function for the half-space $\{x \in \mathbb{R}^3 : \ x_3 > 0\}$,

$$G(x, y) := \Phi(x, y) - \Phi(x, y') \tag{17}$$

with $y' := (y_1, y_2, -y_3)$, to derive an integral equation of the second kind for the solution of the rough surface scattering problem. The decay of the function G is given in [1, equation (3.8)] as

$$\Big|G(x, y)\Big| \le \frac{C}{|x-y|^2} \tag{18}$$

for all $x, y \in \mathbb{R}^3$ with $f^- < x_3, y_3 < f^+$ with some constant C. For the normal derivative of G the estimate

$$\Big|\frac{\partial G(x,y)}{\partial \nu(y)}\Big| \leq \frac{C}{|x-y|^2} \tag{19}$$

holds for all $x, y \in \mathbb{R}^3$ with $f^- < x_3, y_3 < f^+$ with some constant C; compare equation (3.11) in [1].

We are now prepared to formulate the scattering problem under consideration.

DEFINITION 2.1 (point source rough surface scattering problem). *For the incident field $u^i(x) := \Phi(x,z)$ with $z \in D$ we seek a scattered field $u^s \in C^2(D) \cap C(\overline{D})$ which satisfies the Helmholtz equation* (11) *in D, the Dirichlet boundary condition* (13) *on Γ, the bound* (14)*, and the limiting absorption principle* (15).

The rough surface scattering problem is transformed into a boundary value problem via the ansatz

$$u^s(x) = v(x) - \Phi(x, z'), \tag{20}$$

where v satisfies the Helmholtz equation (11), the bound (14), the limiting absorption principle (15), and the boundary condition

$$v(x) = g(x), \quad x \in \Gamma, \tag{21}$$

with

$$g(x) := -(\Phi(x,z) - \Phi(x,z')) = -G(x,z). \tag{22}$$

A solution to the boundary value problem can be found via the single- and double-layer potential approach. We define the *single-layer potential*

$$u_1(x) = \int_\Gamma G(x,y)\varphi(y)\, ds(y), \quad x \in \mathbb{R}^3, \tag{23}$$

and the *double-layer potential*

$$u_2(x) = \int_\Gamma \frac{\partial G(x,y)}{\partial \nu(y)}\varphi(y)\, ds(y), \quad x \in \mathbb{R}^3. \tag{24}$$

The boundary values of these potentials can be calculated using the boundary integral operators S and K defined by

$$(S\varphi)(x) = 2\int_\Gamma G(x,y)\varphi(y)\, ds(y), \quad x \in \Gamma, \tag{25}$$

and the *double-layer potential*

$$(K\varphi)(x) = 2\int_\Gamma \frac{\partial G(x,y)}{\partial \nu(y)}\varphi(y)\, ds(y), \quad x \in \Gamma. \tag{26}$$

It is shown in [1] that the combined single- and double-layer potential

$$v(x) := u_2(x) - i\eta u_1(x), \quad x \in D, \tag{27}$$

with parameter $\eta > 0$ satisfies the boundary value problem (21) if and only if the density $\varphi \in L^2(\Gamma)$ satisfies the integral equation

$$(I + K - i\eta S)\varphi = 2g, \tag{28}$$

which is of the form $(I - W)\varphi = f$ with $W = -K + i\eta S$ and $f = 2g$. The basic uniqueness and existence result is given by Theorem 3.4 of [2] as follows.

THEOREM 2.2. *The operator $I + K - i\eta S$ is boundedly invertible on $L^2(\Gamma)$, and for the norm of its inverse, one has*

$$\|(I + K - i\eta S)^{-1}\| \leq B, \tag{29}$$

where the constant B, as given in (3.4) *in* [2], *depends only on the quotient κ/η and the Lipschitz constant of f.*

As a result of Theorem 2.2 we obtain the existence of the solution for the rough surface scattering problem. In principle, the solution to the scattering problem is given by the combined potential (27) with a density φ which satisfies (28). Our main topic here is the numerical solution of such integral equations in three dimensions. We need to treat the numerical integration of functions over unbounded surfaces.

3. The multisection method (MSM). We will formulate our multisection method (from this point abbreviated by MSM) for the following abstract setting which includes the rough surface scattering problems discussed above.

Let Y be a Banach space, and let $\{P_\varrho\}_{\varrho>0}$ be a family of linear operators on Y with the following three properties:

(P1) $P_\varrho P_\tau = P_\tau = P_\tau P_\varrho$ for all $\varrho \geq \tau > 0$,

(P2) $\|P_\varrho\| = 1$ for all $\varrho > 0$,

(P3) $P_\varrho \to I$ pointwise, which means $P_\varrho\varphi \to \varphi$ for all $\varphi \in Y$, as $\varrho \to \infty$.

From (P1) with $\varrho = \tau$ we get that every P_ϱ is a projection operator. We will also have to deal with the complementary projectors $I - P_\varrho$ which, for brevity, shall be denoted by Q_ϱ for every $\varrho > 0$.

Now suppose A is a bounded linear operator on Y with

(A1) A invertible (and therefore boundedly invertible) on Y,

(A2) $\|Q_\varrho A P_\tau\| \to 0$ as $\varrho \to \infty$ for every fixed $\tau > 0$.

For illustration we give an example of Y, $\{P_\varrho\}_{\varrho>0}$, and A that includes the setting from (28) in section 2.

Example 3.1. Let $Y = L^p(\mathbb{R}^n)$ with $1 \leq p < \infty$ and $n \in \mathbb{N}$, and define P_ϱ as in (5) for every $\varrho > 0$. Then Y and the family $\{P_\varrho\}_{\varrho>0}$ are clearly subject to our assumptions (P1)–(P3).

Now let $A = I - W$, where W is a well-defined and bounded integral operator

$$(W\varphi)(x) = \int_{\mathbb{R}^n} k(x,y)\,\varphi(y)\,dy, \qquad x \in \mathbb{R}^n, \tag{30}$$

on Y. Hence the kernel function $k(\cdot,\cdot)$ must satisfy a decay condition

$$|k(x,y)| \leq \frac{C}{|x-y|^\gamma} \quad \text{for} \quad |x-y| > 1 \tag{31}$$

with constants $\gamma > 0$ and some $C > 0$.

Note that this example, with $p = n = \gamma = 2$, covers the operators in the boundary integral formulation (28) arising from 3D rough surface scattering problems as discussed in section 2.

The following lemma shows that our assumption (A2) automatically holds for the operator A from Example 3.1 if $\gamma p > n$.

LEMMA 3.2. *Let $A = I - W$ act boundedly on $Y = L^p(\mathbb{R}^n)$ with $1 \le p < \infty$, $n \in \mathbb{N}$ and W as in (30) and (31) with $C > 0$ and $\gamma > 0$. Then, for every $\tau > 0$, we have*

$$\|Q_\varrho A P_\tau\| \le \frac{c}{\varrho^{\gamma - n/p}}, \qquad \varrho > 2\tau, \tag{32}$$

with some constant $c > 0$ depending on τ. In particular, if $\gamma p > n$, then assumption (A2) *holds.*

Proof. Let $B_\varrho = \{x \in \mathbb{R}^n : |x| \le \varrho\}$ and $\partial B_\varrho = \{x \in \mathbb{R}^n : |x| = \varrho\}$ be the ball and sphere of radius $\varrho > 0$ in $\mathbb{R}^n$, and denote their n- and $(n-1)$-dimensional measure by $|B_\varrho|$ and $|\partial B_\varrho|$, respectively. Now take some $\varrho > 2\tau > 0$ and first suppose $1 < p < \infty$. Using Hölder's inequality with $1/p + 1/q = 1$ we get the following:

$$\begin{aligned}
\|Q_\varrho A P_\tau \varphi\|^p_{L^p(\mathbb{R}^n)} &= \int_{|x|\ge\varrho} \Big|(AP_\tau\varphi)(x)\Big|^p dx \\
&= \int_{|x|\ge\varrho} \Big|(WP_\tau\varphi)(x)\Big|^p dx \\
&= \int_{|x|\ge\varrho} \Big| \int_{|y|<\tau} k(x,y)\,\varphi(y)\; dy\Big|^p dx \\
&\le \int_{|x|\ge\varrho} \left(\left(\int_{|y|<\tau} |k(x,y)|^q\; dy \right)^{1/q} \cdot \|P_\tau\varphi\|_{L^p(\mathbb{R}^n)} \right)^p dx \\
&\le \int_{|x|\ge\varrho} \left(\int_{|y|||<\tau} |k(x,y)|^q\; dy \right)^{p/q} dx \;\cdot\; \|\varphi\|^p_{L^p(\mathbb{R}^n)}.
\end{aligned}$$

Consequently, using the bound (31) and the inequality

$$|x| \ge \varrho > 2\tau > 2|y|, \quad \text{which implies} \quad |x-y| \ge |x| - |y| > |x|/2,$$

we get

$$\begin{aligned}
\|Q_\varrho A P_\tau\|^p &\le \int_{|x|\ge\varrho} \left(\int_{|y|<\tau} \frac{C^q}{|x-y|^{\gamma q}}\; dy \right)^{p/q} dx \qquad (33) \\
&\le \int_{|x|\ge\varrho} \left(\int_{|y|<\tau} \frac{C^q}{(|x|/2)^{\gamma q}}\; dy \right)^{p/q} dx \\
&= \left(\int_{|x|\ge\varrho} \frac{1}{|x|^{\gamma p}}\; dx \right) \left(\int_{|y|<\tau} dy \right)^{p/q} 2^{\gamma p}\, C^p \\
&= \left(\int_{r=\varrho}^{+\infty} \frac{1}{r^{\gamma p}}\; |\partial B_r|\; dr \right) |B_\tau|^{p/q}\, 2^{\gamma p}\, C^p \\
&= \left(\int_{r=\varrho}^{+\infty} \frac{r^{n-1}}{r^{\gamma p}}\; dr \right) |\partial B_1|\, |B_\tau|^{p/q}\, 2^{\gamma p}\, C^p \\
&= \varrho^{\,n-\gamma p}\; \frac{1}{\gamma p - n}\, |\partial B_1|\, |B_\tau|^{p/q}\, 2^{\gamma p}\, C^p.
\end{aligned}$$

Finally, taking pth roots proves (32). The proof for $p = 1$ is similar. But instead of using Hölder's inequality one immediately arrives at (33), with p/q replaced by 0. ∎

We illustrate the generality of our approach with three more examples of Banach spaces Y with projections P_ϱ and operators A acting on them.

Example 3.3. Let $Y = \ell^p(\mathbb{Z}^n)$ with $1 \le p < \infty$ and $n \in \mathbb{N}$. Define P_ϱ, with $\varrho > 0$, literally as in (5) but with $x \in \mathbb{Z}^n$, of course. Also here, (P1)–(P3) clearly hold.

In this case, in fact every bounded linear operator A on Y is subject to (A2)! This can be seen as follows. Since all P_τ are compact operators on Y, also AP_τ is compact for every $\tau > 0$. But since, by (P3), $Q_\varrho \to 0$ pointwise as $\varrho \to \infty$, and since pointwise convergence on compact sets is uniformly, we get that even (A2) holds.

Example 3.4. Let $Y = C[0, 1]$. For $m \in \mathbb{N}$, let $P_m\varphi$ denote the piecewise linear function which interpolates $\varphi \in Y$ at the points $j/2^m$ with $j = 0, \dots, 2^m$, and for arbitrary $\varrho > 0$, put $P_\varrho = P_{[\varrho]}$ with $[\varrho]$ denoting the integer part of ϱ. Then it is easy to see that (P1)–(P3) are fulfilled.

Example 3.5. Let $Y = L^2(\mathbb{T})$, where $\mathbb{T} := \{z \in \mathbb{C} : |z| = 1\}$ is the complex unit circle. For $m \in \mathbb{N}$, let

$$(P_m\varphi)(t) = \sum_{k=-m}^{m} c_k\, t^k, \quad t \in \mathbb{T},$$

be the truncated Fourier series of φ with c_k denoting its kth Fourier coefficient. And again put $P_\varrho = P_{[\varrho]}$ for arbitrary $\varrho > 0$. Then (P1)–(P3) are fulfilled.

Note that this example is isometrically isomorphic to Example 3.3 with $n = 1$ and $p = 2$, the isomorphism being the Fourier transform. As a result we get that also here every bounded linear operator A on Y is subject to (A2).

Given a Banach space Y and a family of projections $\{P_\varrho\}_{\varrho>0}$ on Y with properties (P1)–(P3), an operator A on Y with properties (A1) and (A2), and an arbitrary element $f \in Y$, we are looking for the (unique) solution $\varphi =: \varphi_0$ of (1); that is,

$$A\varphi = f.$$

For the approximate solution of (1) we propose the following method.

DEFINITION 3.6 (MSM). *For a given precision $\delta > 0$ and sufficiently large cutoff parameters ϱ and τ, calculate a solution $\varphi \in Y$ of the system*

$$\left.\begin{aligned} P_\tau\varphi &= \varphi, \\ \|P_\varrho A P_\tau \varphi - P_\varrho f\| &\le \delta. \end{aligned}\right\} \quad \text{(MSM)}$$

DEFINITION 3.7. *We say that $\tau_0 > 0$ is an* admissible τ-bound *for a given precision $\delta > 0$ if the system (MSM) is solvable in Y for all $\varrho > 0$ and $\tau > \tau_0$.*

THEOREM 3.8 (existence of solutions to (MSM)). *For every $\delta > 0$, there is an admissible τ-bound $\tau_0 = \tau_0(\delta) > 0$.*

Proof of Theorem 3.8. Let $\varrho > 0$ be arbitrary. We demonstrate how to choose τ_0 so that

$$\varphi_1 := P_\tau\varphi_0 = P_\tau A^{-1} f \tag{34}$$

solves the system (MSM) for every $\tau > \tau_0$, with φ_0 being the exact solution of (1).

Since P_τ is a projector, we have

$$P_\tau\varphi_1 = P_\tau^2\varphi_0 = P_\tau\varphi_0 = \varphi_1$$

for every $\tau > 0$.

Furthermore, for all $\varrho > 0$ and $\tau > 0$, we have

$$\begin{aligned}\|P_\varrho A P_\tau \varphi_1 - P_\varrho f\| &= \|P_\varrho A P_\tau^2 A^{-1} f - P_\varrho f\| \\ &\le \|P_\varrho A A^{-1} f - P_\varrho f\| \ + \ \|P_\varrho A Q_\tau A^{-1} f\| \\ &\le 0 \ + \ \|A\| \cdot \|Q_\tau A^{-1} f\|.\end{aligned}$$

But, by assumption (P3), there is a $\tau_0 > 0$ such that

$$\|Q_\tau A^{-1} f\| \le \frac{\delta}{\|A\|} \tag{35}$$

for all $\tau > \tau_0$, so that finally

$$\|P_\varrho A P_\tau \varphi_1 - P_\varrho f\| \le \delta$$

holds, and hence φ_1 is a solution of the system (MSM) for all $\tau > \tau_0$ and $\varrho > 0$. ∎

LEMMA 3.9. *Let $\tau_0 > 0$ be an admissible τ-bound for a given precision $\delta > 0$. If $\tau > \tau_0$ and $\varrho > 0$ are such that $\|Q_\varrho A P_\tau\| < 1/\|A^{-1}\|$, then the set of all solutions of the system (MSM) is a bounded subset of Y. Precisely, every solution $\varphi \in Y$ of the system (MSM) is subject to $\|\varphi\|_Y \le M$ with M given by* (36).

Proof. Suppose $\varphi \in Y$ solves the system (MSM) for given parameters $\delta, \varrho, \tau > 0$. Then

$$\begin{aligned}\|A\varphi\| - \|P_\varrho f\| \le \|A\varphi - P_\varrho f\| \ &= \ \|AP_\tau\varphi - P_\varrho f\| \\ &\le \|AP_\tau\varphi - P_\varrho A P_\tau \varphi\| + \|P_\varrho A P_\tau \varphi - P_\varrho f\| \\ &\le \|Q_\varrho A P_\tau\| \cdot \|\varphi\| + \delta,\end{aligned}$$

together with $\|\varphi\| \le \|A^{-1}\| \cdot \|A\varphi\|$, implies that

$$\begin{aligned}\frac{\|\varphi\|}{\|A^{-1}\|} \ \le \ \|A\varphi\| &\le \|P_\varrho f\| + \|Q_\varrho A P_\tau\| \cdot \|\varphi\| + \delta \\ &\le \|f\| + \|Q_\varrho A P_\tau\| \cdot \|\varphi\| + \delta,\end{aligned}$$

and hence

$$\|\varphi\| \ \le \ M \ := \ \frac{\|f\| \ + \ \delta}{1/\|A^{-1}\| \ - \ \|Q_\varrho A P_\tau\|}. \tag{36}$$ ∎

THEOREM 3.10 (convergence of the MSM). *For every $\varepsilon > 0$, there are parameters δ, ϱ, τ such that every solution $\varphi \in Y$ of the system (MSM) is an approximation*

$$\|\varphi - \varphi_0\|_Y < \varepsilon \tag{37}$$

of the exact solution φ_0 of (1). *Precisely, there are functions $\delta_0, \tau_0 : \mathbb{R}_+ \to \mathbb{R}_+$ and $\varrho_0 : \mathbb{R}_+^3 \to \mathbb{R}_+$ such that if $\delta < \delta_0(\varepsilon)$, $\tau > \tau_0(\delta)$, and $\varrho > \varrho_0(\varepsilon, \delta, \tau)$, then every solution $\varphi \in Y$ of (MSM) is subject to* (37).

Proof. Let $\varepsilon > 0$ be given. We start the proof with three preliminary steps.

(a) Choose $\delta < \delta_0 := \frac{\varepsilon}{3\|A^{-1}\|}$.

(b) Choose $\tau_0 > 0$ such that $\big(\|Q_\tau\varphi_0\| = \big)\|Q_\tau A^{-1}f\| \le \frac{\delta}{\|A\|}$ for all $\tau > \tau_0$, so that τ_0 is an admissible τ-bound for δ (see inequality (35)). Now let $\tau > \tau_0$.

(c) Choose $\varrho_0 > 0$ such that $\|Q_\varrho f\| < \frac{\varepsilon}{3\|A^{-1}\|}$ and

$$\|Q_\varrho A P_\tau\| \;<\; \frac{1}{\|A^{-1}\|}\left(1 - \frac{1}{1 + \frac{\varepsilon}{3(\|f\|+\delta)\cdot\|A^{-1}\|}}\right) \tag{38}$$

for all $\varrho > \varrho_0$, and fix some $\varrho > \varrho_0$.

Now let $\varphi \in Y$ be a solution of (MSM) with parameters δ, τ, and ϱ as chosen above. From (38) we get $\|Q_\varrho A P_\tau\| < 1/\|A^{-1}\|$, and hence, by Lemma 3.9,

$$\|\varphi\| \;\le\; M \tag{39}$$

with M as defined in (36). Moreover, inequality (38) is equivalent to

$$\|Q_\varrho A P_\tau\| \;<\; \frac{1}{\|A^{-1}\|} \cdot \frac{\frac{\varepsilon}{3(\|f\|+\delta)\cdot\|A^{-1}\|}}{1 + \frac{\varepsilon}{3(\|f\|+\delta)\cdot\|A^{-1}\|}},$$

and hence to

$$\left(1 + \frac{\varepsilon}{3(\|f\|+\delta)\cdot\|A^{-1}\|}\right) \cdot \|Q_\varrho A P_\tau\| \;<\; \frac{1}{\|A^{-1}\|} \cdot \frac{\varepsilon}{3(\|f\|+\delta)\cdot\|A^{-1}\|}.$$

This, moreover, is equivalent to

$$\begin{aligned}\|Q_\varrho A P_\tau\| &< \frac{1}{\|A^{-1}\|} \cdot \frac{\varepsilon}{3(\|f\|+\delta)\cdot\|A^{-1}\|} - \frac{\varepsilon}{3(\|f\|+\delta)\cdot\|A^{-1}\|} \cdot \|Q_\varrho A P_\tau\| \\ &= \frac{\varepsilon\,\big(1/\|A^{-1}\| - \|Q_\varrho A P_\tau\|\big)}{3(\|f\|+\delta)\cdot\|A^{-1}\|} \;=\; \frac{\varepsilon}{3M\|A^{-1}\|} \end{aligned} \tag{40}$$

with M as defined in (36). Then we have

$$\begin{aligned}\|\varphi - \varphi_0\| = \|P_\tau\varphi - \varphi_0\| \;&=\; \|A^{-1}AP_\tau\varphi - A^{-1}f\| \;\le\; \|A^{-1}\| \cdot \|AP_\tau\varphi - f\| \\ &\le \|A^{-1}\| \cdot \big(\|AP_\tau\varphi - P_\varrho AP_\tau\varphi\| \;+\; \|P_\varrho AP_\tau\varphi - P_\varrho f\| \;+\; \|P_\varrho f - f\|\big) \\ &\le \|A^{-1}\| \cdot \big(\|Q_\varrho AP_\tau\| \cdot \|\varphi\| \;+\; \delta \;+\; \|Q_\varrho f\|\big) \\ &\le \frac{\varepsilon}{3} \;+\; \frac{\varepsilon}{3} \;+\; \frac{\varepsilon}{3} \;=\; \varepsilon,\end{aligned}$$

using (40) and (39) and the bounds on δ and $\|Q_\varrho f\|$ in the last step. ∎

Remark 3.11. One way to effectively solve the system (MSM) for given parameters ϱ, τ, and δ is to compute a $\varphi \in Y$ that minimizes the discrepancy in (8), for example, using a gradient method or, if possible, by directly applying the Moore Penrose pseudoinverse $B^\dagger$ of $B := P_\varrho A P_\tau$ to the right-hand side $P_\varrho f$.

If Y is a Hilbert space, then it is well known that $\varphi \in Y$ minimizes the residual $\|B\varphi - P_\varrho f\|$ if and only if $B^*(B\varphi - P_\varrho f) = 0$ with B^* denoting the adjoint operator of B. If, in addition, P_ϱ is self-adjoint for all $\varrho > 0$, then, after resubstituting B, the latter is equivalent to

$$P_\tau A^* P_\varrho A P_\tau \varphi \;=\; P_\tau A^* P_\varrho f. \tag{41}$$

However, if ϱ is sufficiently large, then, by (A2) and (P3), (41) is just a small perturbation of

$$P_\tau A^* A P_\tau \varphi \;=\; P_\tau A^* f, \tag{42}$$

which is nothing but the finite section method for the equation

$$A^* A \varphi \;=\; A^* f. \tag{43}$$

Note that the finite section method (42) is applicable since A^*A is positive definite (see, e.g., Theorem 1.10b in [7]). Clearly, if A is invertible, as we require in (A1), then also its adjoint A^* is invertible, and (43) is equivalent to our original equation (1).

Summarizing, if Y is a Hilbert space and all P_ϱ are self-adjoint, then minimizing $\|P_\varrho A P_\tau \varphi - P_\varrho f\|$ is equivalent to solving a slight perturbation (41) of the finite section method (42) for (43).

The MSM can be applied to the rough surface scattering problem of section 2, as we know from Example 3.1, Lemma 3.2, and inequalities (18) and (19). In particular, note that the bound (29) on $\|A^{-1}\|$ enables us to actually compute the corresponding terms in steps (a) and (c) in the proof of Theorem 3.10.

We summarize the results in the following theorem.

THEOREM 3.12 (MSM for rough surface scattering). *The MSM, as defined in Definition* 3.6, *applied to rough surface scattering* (11), (13), (14), *and* (15), *is convergent in the sense of Theorem* 3.10.

4. Numerical realization and examples. The goal of this final section is to provide a numerical algorithm for the solution of the integral equation (28) using the MSM. For simplicity, we employ a low-order scheme for the treatment of the singularity of the operator, leaving more sophisticated high-order approaches to [8].

Our numerical approach to the solution of the system (MSM) for the concrete operator equation (28) is to choose some large parameters ϱ and τ and to choose the discrepancy δ as small as it possibly can be by looking for a function $\varphi \in L^2(B_\tau)$ that minimizes the MSM residual $\|P_\varrho A P_\tau \varphi - P_\varrho f\|$.

We have implemented both a *direct solver* for the minimization of the corresponding MSM residual and an *iterative scheme* based on a gradient method. In both cases we first transform the truncated approximate equation (8) by numerical quadrature and projection as described in (52)–(56) below into an approximate matrix equation which, for simplicity, we denote by

$$\mathbf{A}\boldsymbol{\varphi} \approx \mathbf{f}. \tag{44}$$

The direct solver then approximately solves this equation via the Moore–Penrose pseudoinverse $\mathbf{A}^\dagger$ of $\mathbf{A}$. Precisely,

$$\boldsymbol{\varphi} = \mathbf{A}^\dagger \circ \mathbf{f} \tag{45}$$

has the smallest norm among all $\boldsymbol{\varphi}$ that minimize the residual $\|\mathbf{A}\boldsymbol{\varphi} - \mathbf{f}\|$.

This scheme is rather quick, but for standard desktop PCs[1] limits the number of unknowns N to several thousands, since we need to store fully occupied matrices of size N^2. This leads to a bound of approximately 10 on the number of wavelengths which can be resolved, since, for example, $N = 2500$ unknowns correspond to a grid

[1]For our computations we used workstations with 2 GB main memory.

size of 50×50 grid points on our surface, and we need at least 5 points per wavelength for its resolution.

For problems with more unknowns we used an iterative solver, where the functional

$$\mu(\boldsymbol{\varphi}) := \|\mathbf{A}\boldsymbol{\varphi} - \mathbf{f}\|^{\mathbf{2}} \tag{46}$$

is minimized. There are well-known schemes available for this task, for example, the CG method, GMRES, or gradient schemes. Here, for the minimization we employed the *gradient method*

$$\boldsymbol{\varphi}_{n+1} \;=\; \boldsymbol{\varphi}_n \;-\; h_n \,\mathrm{grad}_{\boldsymbol{\varphi}}\, \mu(\boldsymbol{\varphi}_n), \quad n = 0, 1, 2, \ldots, \tag{47}$$

with starting value $\boldsymbol{\varphi} = 0$ and adaptive step size $h_n > 0$. Simple calculations show that the gradient of μ with respect to the complex density $\boldsymbol{\varphi}$ is given by

$$\mathrm{grad}_{\boldsymbol{\varphi}}\, \mu(\varphi) \;=\; 2\mathbf{A}^* \circ (\mathbf{A}\boldsymbol{\varphi} - \mathbf{f}). \tag{48}$$

To see this one simply writes $\boldsymbol{\varphi} \in \mathbb{C}^N$ as a function $\tilde{\boldsymbol{\varphi}} \in \mathbb{R}^{2N}$, calculates the gradient in $\mathbb{R}^{2N}$, and converts it back into $\mathbb{C}^N$. The calculation of

$$\mathbf{s_1} := \mathbf{A}\boldsymbol{\varphi}, \tag{49}$$

$$\mathbf{s_2} \;=\; \mathbf{s_1} - \mathbf{f}, \tag{50}$$

$$\mathrm{grad}_{\boldsymbol{\varphi}}\, \mu(\varphi) \;=\; 2\mathbf{A}^* \circ \mathbf{s_2} \tag{51}$$

can be carried out without storing the matrices $\mathbf{A}$ and $\mathbf{A}^*$; i.e., for the gradient method we only need to store vectors of length N. Here, in principle, we can work with up to several million unknowns N if we can invest the time to calculate the updates. In this case we can resolve up to 400 wavelengths with a grid size of 2000×2000.

For the iteration scheme the computation time per iteration scales quadratically in the number of unknowns. If we want to keep the time for one gradient step in the time frame of hours,[2] this limits the size of the problem to approximately $N = 100.000$, which corresponds to a grid of 350×350 or a resolution of 70 wavelengths. It is clear that at this point matrix compression schemes like the fast multipole method need to be employed. Since our main point here is the presentation of the MSM and the proof of its applicability, we leave this to future research.

The main ingredient of (28) is the integral operator $I + K - i\eta S$, which, for the MSM, is applied to some density $P_\tau \varphi$ supported on B_τ and is evaluated on the set B_ϱ. We denote the kernel of $K - i\eta S$ by $k(x, y)$, $x \in B_\varrho$, $y \in B_\tau$.

We employ the basic ideas of Nyström's method, however, applied to the approximate equation (8). We use standard numerical quadrature for the calculation of

$$I(x) := \int_{B_\tau} k(x, y)\varphi(y)\, dy, \quad x \in B_\varrho; \tag{52}$$

i.e., we replace the integral by a finite sum

$$I_N(x) := \sum_{\xi=1}^{N} \alpha_\xi \tilde{k}(x, y_\xi)\varphi(y_\xi), \quad x \in B_\varrho, \tag{53}$$

[2]Calculations have been carried out on a PC with processor speed of 3.6 GHz.

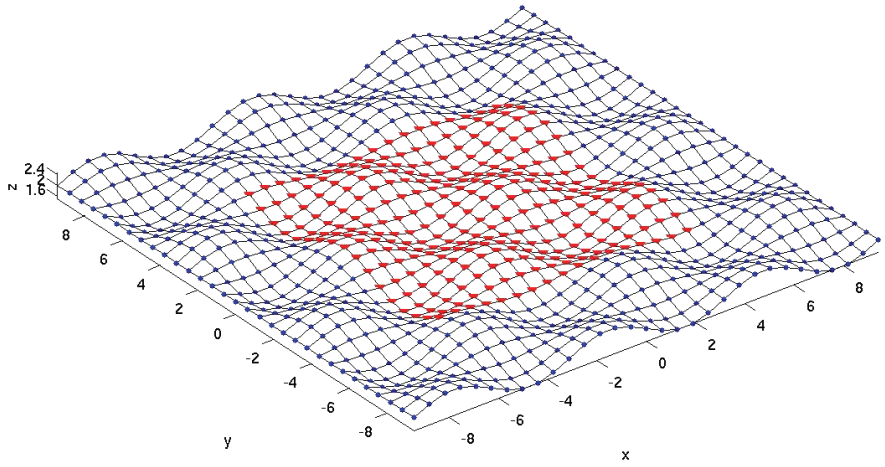


FIG. 1. *Setting for the MSM. The set B_ϱ is shown with dots, the set B_τ with triangles.*

with *quadrature points* $y_\xi \in B_\varrho$ and *weights* α_ξ. For details about a variety of such quadratures we refer the reader to [4] and [9]. Here, we employed a trapezoidal rule, i.e., the exact integration of linear functions, combined with a cutoff of the singular parts of the kernels. For more sophisticated schemes for the treatment of the singularity we refer the reader to [8]. For a convergent quadrature formula the values (53) approximate the integral (52) and we have

$$I_N(x) \to I(x), \quad N \to \infty, \quad x \in B_\varrho. \tag{54}$$

Next, we choose points $x_\eta \in B_\varrho$ for $\eta = 1, \ldots, M$ and define

$$V := \{x_\eta : \ \eta = 1, \ldots, M\}, \qquad L := \{y_\xi : \ \xi = 1, \ldots, N\}. \tag{55}$$

For our numerics of the direct method we chose the points such that $L \subset V$; compare Figure 1 for the case of an equidistant grid. Then we evaluate $I_N(x_\eta)$ at the points x_η, $\eta = 1, \ldots, M$. This leads to the discrete system

$$\|\mathbf{A}\boldsymbol{\varphi} - \mathbf{f}\|_{\mathbf{2}} \le \delta \tag{56}$$

with

$$\mathbf{A} = (\mathbf{I} - \mathbf{W}) \circ \mathbf{P}_\tau \qquad \text{(see Figure 2)}, \tag{57}$$

$$\boldsymbol{\varphi} = \begin{pmatrix} \varphi(x_1) \\ \vdots \\ \varphi(x_M) \end{pmatrix}, \tag{58}$$

$$\mathbf{f} = \begin{pmatrix} f(x_1) \\ \vdots \\ f(x_M) \end{pmatrix}, \tag{59}$$

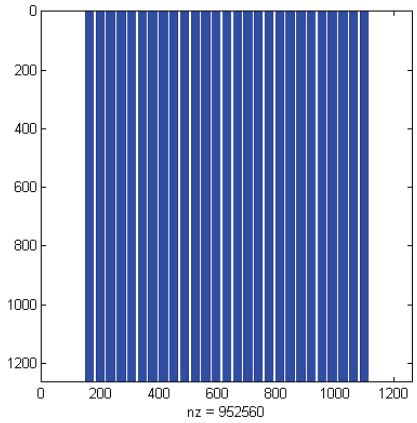


FIG. 2. *Structure of the matrix* $\mathbf{A} = (\mathbf{I} - \mathbf{W}) \circ \mathbf{P}_\tau$.

$$\mathbf{W} = \begin{pmatrix} \alpha_1 k(x_1, x_1) & \dots & \alpha_M k(x_1, x_M) \\ \vdots & & \vdots \\ \alpha_1 k(x_M, x_1) & \dots & \alpha_M k(x_M, x_M) \end{pmatrix}, \tag{60}$$

and the projector $\mathbf{P}_\tau$ is defined by

$$\mathbf{P}_\tau = \operatorname{diag}(\mathbf{v}), \quad \mathbf{v_j} = \begin{cases} 1 & \text{if } x_j \in B_\tau, \\ 0 & \text{otherwise} \end{cases} \tag{61}$$

for $j = 1, \dots, M$.

The solution of (56) is calculated using the Moore–Penrose pseudoinverse $\mathbf{A}^\dagger$ (in MATLAB realized via the function pinv($\mathbf{A}$)) applied to $\mathbf{f}$, i.e., via

$$\boldsymbol{\varphi}_{\text{sol}} := \Big((\mathbf{I} - \mathbf{W}) \circ \mathbf{P}_\tau \Big)^\dagger \mathbf{f}. \tag{62}$$

As a test example for the numerical realization of the MSM we use the surface Γ defined by

$$f(x_1, x_2) := c_0 \cdot \sin(x_1) \cdot \cos(x_2) + 2$$

with $c_0 = 1/2$. Here, we place the source point z of the incident point source $\Phi(\cdot, z)$ below the surface Γ, but above the plane $x_3 = 0$. In this case the scattered field is given by $-\Phi(\cdot, z)$. Thus, we can test the precision of the combined single- and double-layer potential with density φ calculated via the MSM.

Figure 3(top) shows the original fields for an incident point source with source point $(-2, 0, 0.2)$. The reconstructed field is shown in Figure 3(bottom). Here, we worked with the direct method and employed the MSM with $\varrho = 3\pi$ and $\tau = 0.8\,\varrho$. The figures show a grid of 40×40 points.

Figure 4 demonstrates the real part of the scattered field for an incident point source with source point $(-2, 0, 4)$. We employed the direct method and used the same choice of parameters as in Figure 3.

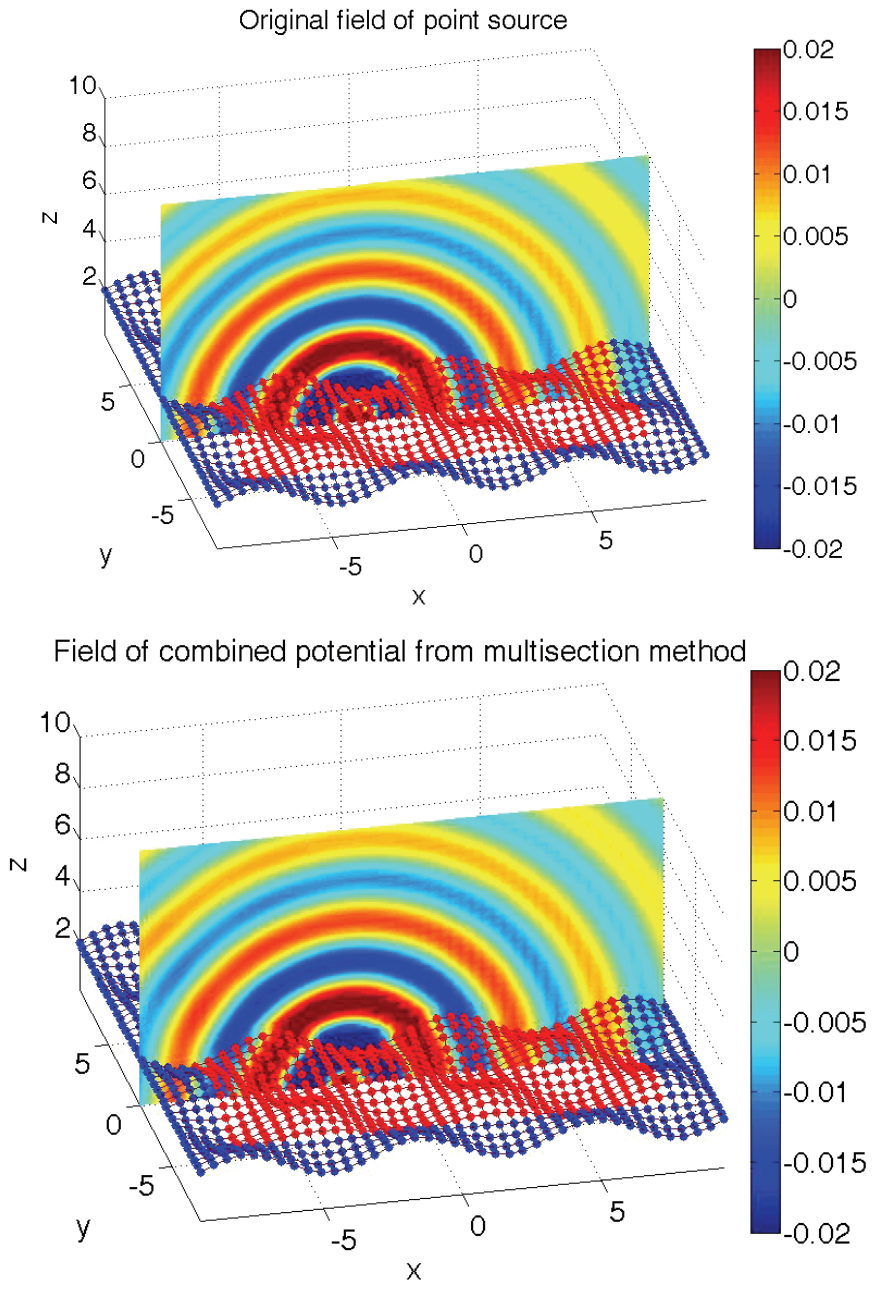


FIG. 3. *Top: Test field of a point source located between the surface and the plane* $x_3 = 0$. *Bottom: Constructed field via the combined layer potential with a density given by the MSM. Here, the relative error for the reconstruction in the points with* $x_3 \geq 3$ *above the set* B_ϱ *is below* 5%.

The behavior of the solution density φ of the MSM is shown in Figure 5. The different sections are indicated in the third row. The right-hand side shows the "true" density, which is calculated using larger patches.

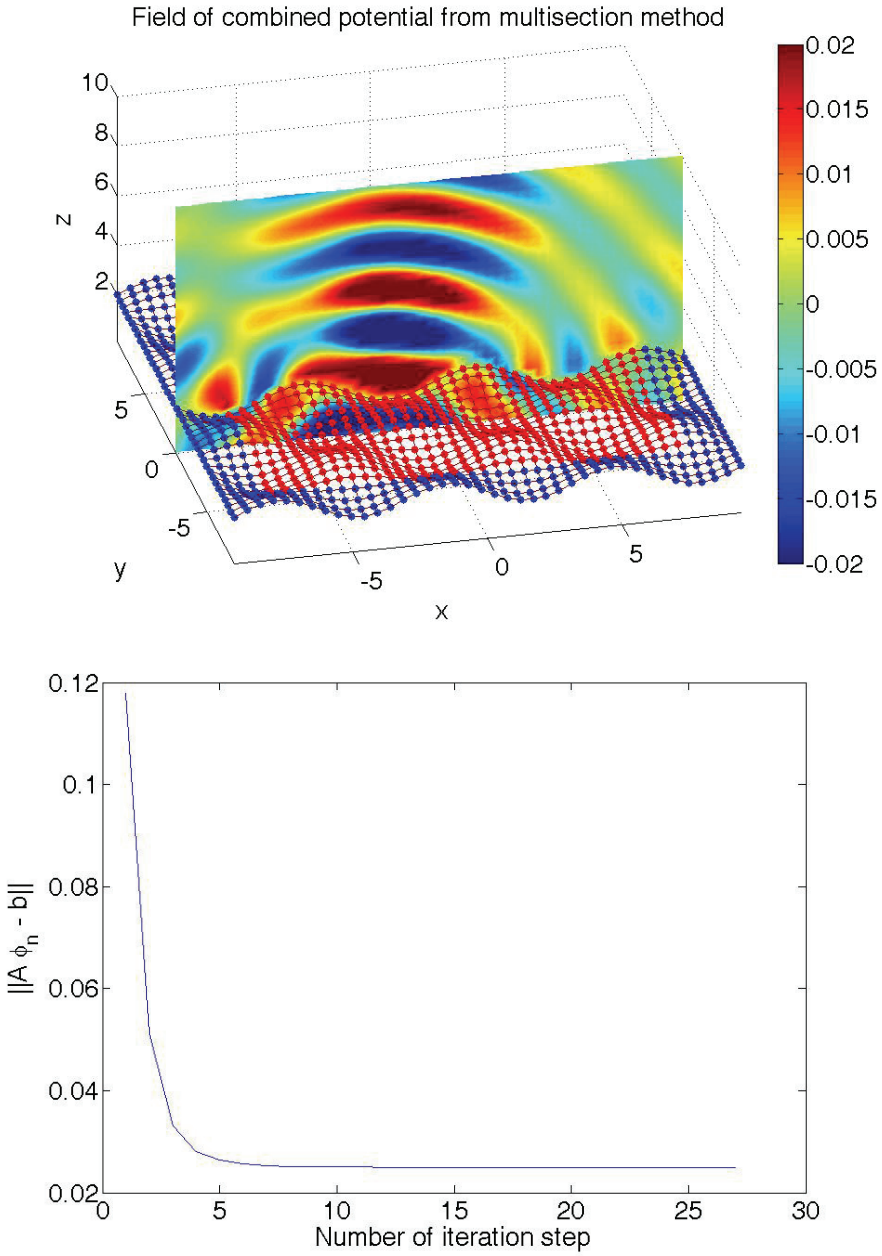


FIG. 4. *Top: Scattered field for an incident point source with source point* $z = (-2, 0, 4)$. *Bottom: Error behavior for the iterates of the gradient method.*

The behavior of the solutions for the *iterative* scheme is demonstrated in Figures 6 and 4. Figure 6(top) is a schematic presentation of the relation between the area ϱ^2 of the calculation domain, the total number of unknowns N, and the one-dimensional step size h. The blue lines are lines of constant step size, where both N and ϱ are increased. To achieve convergence of the numerical method, we need to increase both ϱ and N such that $h = h(\varrho, N)$ tends to zero.

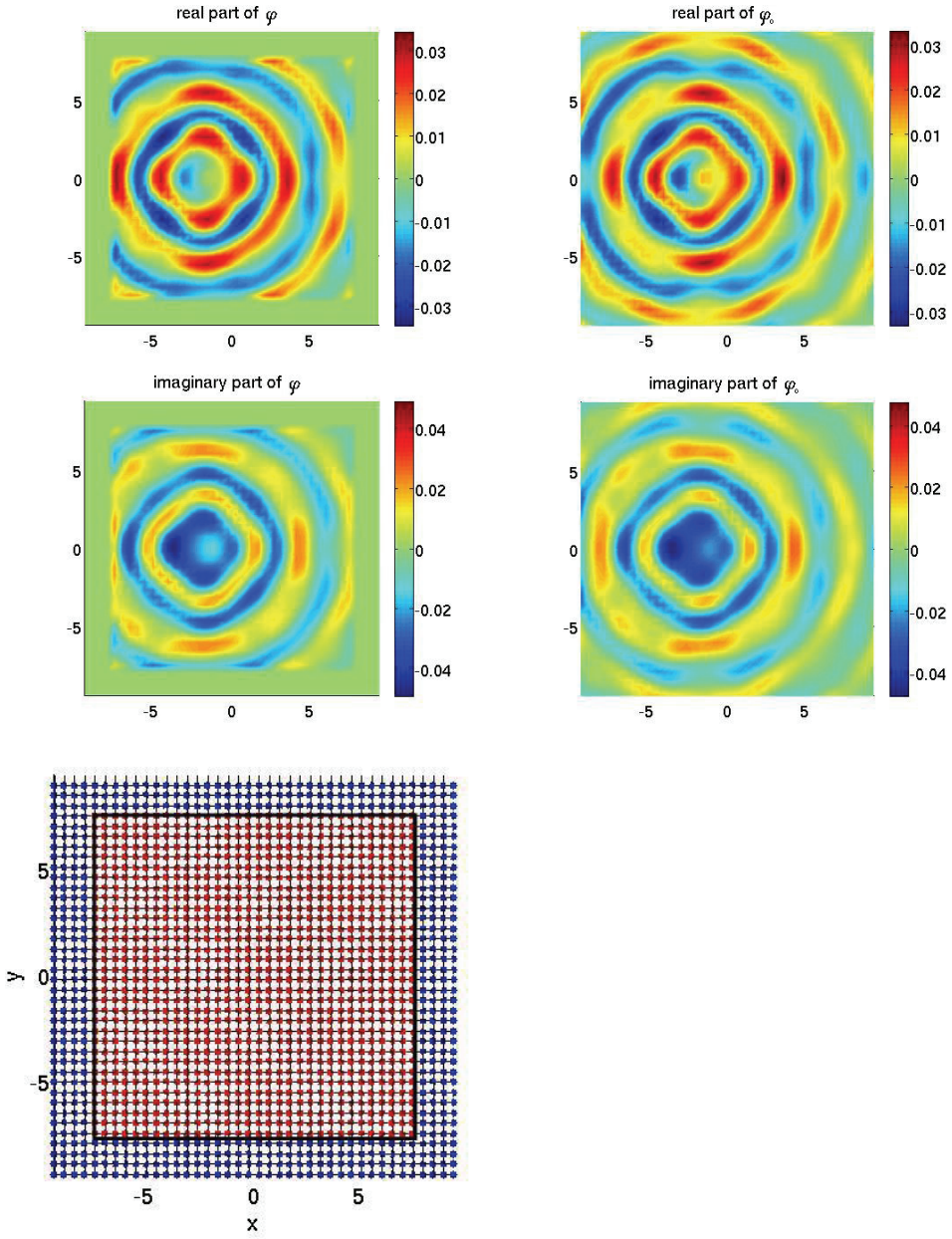


FIG. 5. *The images show the real and imaginary parts of the density φ solving the multisection inequality in the left column and the true density φ_0 in the right column. The location of the two different sections B_τ and B_ϱ is indicated in the image of the third row. On the set B_τ the approximation is good and on $B_\varrho \setminus B_\tau$ the density φ is zero by construction.*

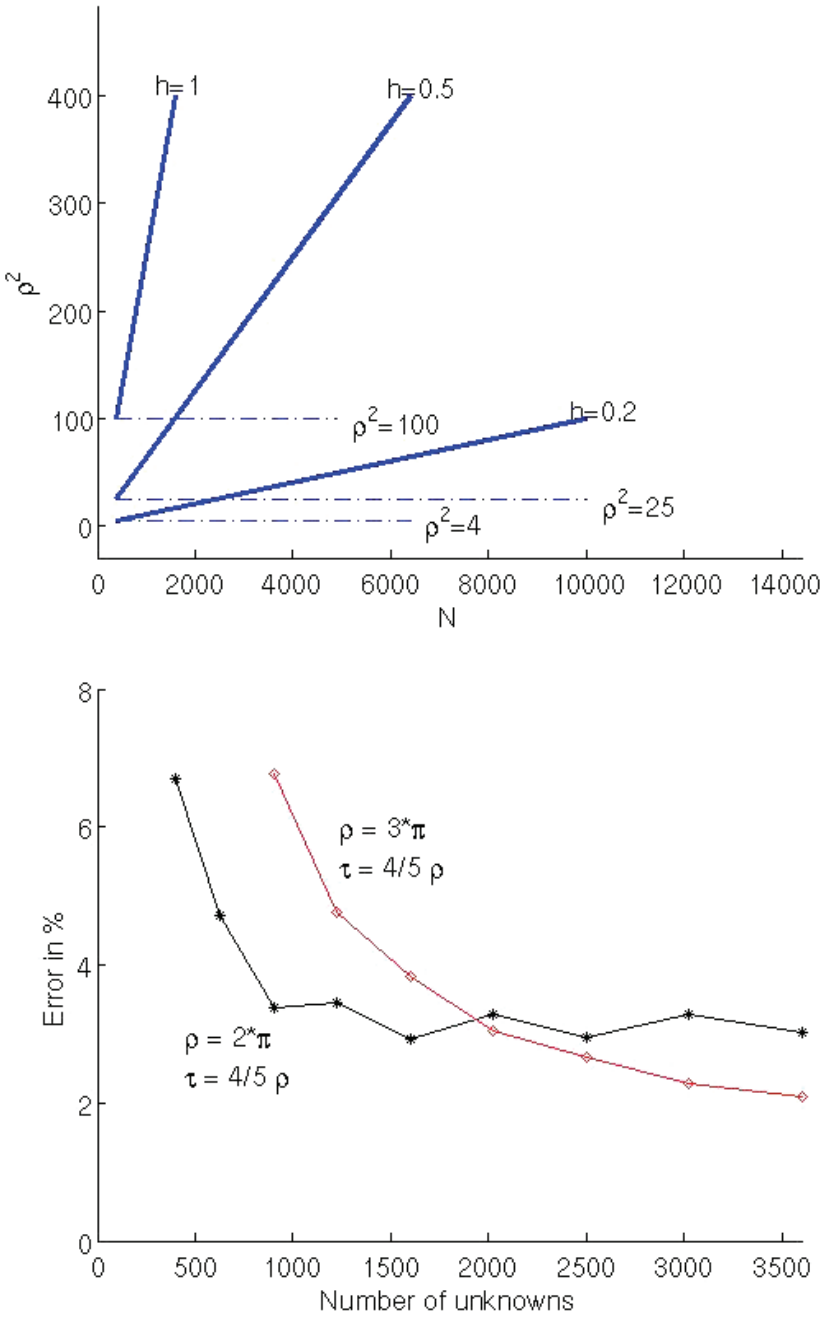


FIG. 6. *Top: Scheme to discuss the discretization and cutoff error and convergence properties of the MSM. Bottom: Illustration of the relative error measured in the calculated field for different choices of cutoff parameters* ϱ *and* τ *dependent on the number of unknowns* N*. In both cases we have chosen a fixed ratio of* $\rho/\tau = 4/5$*. The curve with stars shows the error for* $\rho = 2\pi$*, and the curve with diamonds shows the error for* $\rho = 3\pi$*.*

The relative error for the calculation of the field of Figure 3 for different choices of ϱ and N is demonstrated in Figure 6(bottom). For a fixed small number of unknowns the error for a large calculation domain is larger than for a small calculation domain. This can be explained by the two types of error which add up to the total error:

1. the discretization error from approximating the integral by a quadrature formula,
2. the cutoff error from the limited domain of calculation.

If for a fixed number of unknowns we enlarge the calculation domain, then the cutoff error decreases, but the discretization error increases. Here with $N = 3500$ unknowns and a sufficiently large calculation area we reach an error of 2%.

The behavior of the residual for the gradient method is illustrated in Figure 4(bottom). The error usually quickly drops and then tends to some fixed residual which does not change significantly when further iterating. The residual is caused by the discretization and cutoff error. We usually used 15–20 iterations in our calculations.

REFERENCES

[1] S. N. Chandler-Wilde, E. Heinemeyer, and R. Potthast, *Acoustic scattering by mildly rough unbounded surfaces in three dimensions*, SIAM J. Appl. Math., 66 (2006), pp. 1002–1026.

[2] S. N. Chandler-Wilde, E. Heinemeyer, and R. Potthast, *A well-posed integral equation formulation for three-dimensional rough surface scattering*, Proc. Roy. Soc. London Ser. A. Math. Phys. Eng. Sci., 462 (2006), pp. 3683–3705.

[3] S. N. Chandler-Wilde and M. Lindner, *Boundary integral equations on unbounded rough surfaces: Fredholmness and the finite section method*, J. Integral Equations Appl., to appear.

[4] D. Colton and R. Kress, *Inverse Acoustic and Electromagnetic Scattering Theory*, 2nd ed., Appl. Math. Sci. 93, Springer-Verlag, Berlin, 1998.

[5] DeSanto, *Scattering by rough surfaces*, in Scattering: Scattering and Inverse Scattering in Pure and Applied Science, R. Pike and P. Sabatier, eds., Academic Press, New York, 2002, pp. 15–36.

[6] H. W. Engl, M. Hanke, and A. Neubauer, *Regularization of Inverse Problems*, Math. Appl. 375, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1996.

[7] R. Hagen, S. Roch, and B. Silbermann, *C^*-Algebras and Numerical Analysis*, Marcel Dekker, New York, 2001.

[8] E. Heinemeyer, *Integral Equation Methods for Rough Surface Scattering Problems in Three Dimensions*, Ph.D. dissertation, University of Göttingen, Göttingen, Germany, 2008.

[9] R. Kress, *Linear Integral Equations*, Appl. Math. Sci. 82, Springer-Verlag, Berlin, 1989.

[10] M. Lindner, *Infinite Matrices and Their Finite Sections: An Introduction to the Limit Operator Method*, Front. Math., Birkhäuser-Verlag, Basel, Boston, Berlin, 2006.

[11] A. Meier and S. N. Chandler-Wilde, *On the stability and convergence of the finite section method for integral equation formulations of rough surface scattering*, Math. Methods Appl. Sci., 24 (2001), pp. 209–232.

[12] J. A. Ogilvy, *Theory of Wave Scattering from Random Rough Surfaces*, Adam Hilger, Bristol, UK, 1991.

[13] V. S. Rabinovich, S. Roch, and B. Silbermann, *Limit Operators and Their Applications in Operator Theory*, Oper. Theory. Adv. Appl. 150, Birkhäuser-Verlag, Basel, Boston, Berlin, 2004.

[14] S. Roch, *Finite sections of band-dominated operators*, Mem. Amer. Math. Soc., to appear.

[15] M. Saillard and A. Sentenac, *Rigorous solutions for electromagnetic scattering from rough surfaces*, Waves Random Media, 11 (2001), pp. R103–R137.

[16] A. G. Voronovich, *Wave Scattering from Rough Surfaces*, 2nd ed., Springer-Verlag, Berlin, 1998.

[17] K. F. Warnick and W. C. Chew, *Numerical simulation methods for rough surface scattering*, Waves Random Media, 11 (2001), pp. R1–R30.

SIAM J. NUMER. ANAL.
Vol. 46, No. 4, pp. 1799–1820
© 2008 Society for Industrial and Applied Mathematics

A SPLINE PRODUCT QUASI-INTERPOLATION METHOD FOR WEAKLY SINGULAR FREDHOLM INTEGRAL EQUATIONS*

EERO VAINIKKO† AND GENNADI VAINIKKO†

Abstract. A discrete method of accuracy $O(h^m)$ is constructed and justified for a class of Fredholm integral equations of the second kind with kernels that may have weak diagonal and boundary singularities. The method is based on (i) improving the boundary behavior of the kernel with the help of a change of variables, and (ii) the product integration using quasi-interpolation by smooth splines of order m.

Key words. weakly singular integral equations, boundary singularities, spline quasi-interpolation, spline interpolation, product integration methods, Nyström-type methods

AMS subject classifications. 65R20, 65D07, 65D30

DOI. 10.1137/070693308

1. Introduction. In the present paper we call attention to some high order fully discrete methods for a class of Fredholm integral equations of the second kind with different weak singularities. The idea of the method is close to those in [2, 3, 9]. Consider the weakly singular integral equation

$$u(x) = \int_0^1 \big(a(x,y)|x-y|^{-\nu} + b(x,y)\big)\, u(y)dy + f(x), \qquad 0 \le x \le 1, \tag{1.1}$$

depending on the parameter ν, $0 < \nu < 1$. A traditional product integration method (see [2, 3] and the literature quoted there) is based on the approximation of (1.1) by the equation

$$u(x) = \int_0^1 \big(|x-y|^{-\nu} P_{n,m}(a(x,y)u(y)) + P_{n,m}(b(x,y)u(y))\big)\, dy + f(x), \qquad 0 \le x \le 1,$$

where $P_{n,m}$ is an interpolation projector onto the space of piecewise polynomial functions of degree $m-1$ associated with a subdivision of $[0,1]$ into n subintervals by some knots $0 = x_{0,n} < x_{1,n} < \cdots < x_{n,n} = 1$; $P_{n,m}$ is applied to the products $a(x,y)u(y)$ and $b(x,y)u(y)$ as functions of y treating x as a parameter. This results in some Nyström-type methods of accuracy $O(n^{-m})$, provided that the grid $\{x_{0,n}, x_{1,n}, \ldots, x_{n,n}\}$ is properly graded to compensate the generic boundary singularities of the derivatives of the exact solution to (1.1); even the accuracy $O(n^{-m-1+\nu})$ can be achieved for a skilled choice of the interpolation points; see [13]. The approach of [9] is different: the authors first perform a change of variables which improves the boundary behavior of the solution and of the coefficient functions $a(x,y), b(x,y)$. After that, using a polynomial approximation of the solution, they apply a Gauss-type quadrature formula for the integral in the transformed equation.

In our approach we follow the methods of [9, 10], performing a smoothing change of variables and then, instead of discontinuous piecewise polynomial interpolation

*Received by the editors May 30, 2007; accepted for publication (in revised form) January 10, 2008; published electronically April 16, 2008. This work was supported by the Estonian Information Technology Foundation, grant 07-03-01-01, and by the Estonian Science Foundation, grant 7353.
http://www.siam.org/journals/sinum/46-4/69330.html
†Faculty of Mathematics and Computer Science, University of Tartu, Liivi 2, Tartu 50409, Estonia (eero.vainikko@ut.ee, gennadi.vainikko@ut.ee).

by $P_{n,m}$, we use a product quasi-interpolation (or even the real interpolation) by smooth splines of degree $m-1$ on the uniform grid of the step size $h = 1/n$. This leads to some reduction of degrees of freedom compared with the use of piecewise polynomials. Our algorithm still has features of the Nyström method. In Nyström-type methods for weakly singular integral equations, the computation of the quadrature coefficients is a laborious and delicate part of the algorithms. The use of quasi-interpolation (or real interpolation) by smooth splines on uniform grids enables us to obtain simple but still somewhat delicate formulae for the quadrature coefficients, and so we obtain a fully discrete method of accuracy $O(h^m)$. This is the main message of the present paper.

In our considerations, the coefficient functions $a(x,y)$ and $b(x,y)$ may have boundary singularities with respect to y. Denote by T the integral operator of (1.1),

$$(Tu)(x) = \int_0^1 \big(a(x,y)|x-y|^{-\nu} + b(x,y)\big)\, u(y)dy.$$

LEMMA 1.1. *Assume that $a, b \in C([0,1]\times(0,1))$ satisfy the inequalities*

$$|a(x,y)| \le cy^{-\lambda_0}(1-y)^{-\lambda_1}, \quad |b(x,y)| \le cy^{-\mu_0}(1-y)^{-\mu_1}, \quad (x,y)\in[0,1]\times(0,1),$$

where $0<\nu<1$, $\lambda_0, \lambda_1, \mu_0, \mu_1 \in \mathbb{R}$, $\lambda_0+\nu<1$, $\lambda_1+\nu<1$, $\mu_0<1$, $\mu_1<1$. Then T maps $C[0,1]$ into $C[0,1]$, and $T: C[0,1]\to C[0,1]$ is compact.

The proof is standard; cf. [5, 7]; a detailed argument can be found in [11].

For $m\in\mathbb{N}$, θ_0, $\theta_1\in\mathbb{R}$, θ_0, $\theta_1<1$, denote by $C^m_\star(0,1)$ and $C^{m,\theta_0,\theta_1}(0,1)$ the weighted spaces of functions $u\in C[0,1]\cap C^m(0,1)$ such that, respectively,

$$\|u\|_{C^m_\star(0,1)} := \sum_{k=0}^m \sup_{0<x<1} x^k(1-x)^k|u^{(k)}(x)| < \infty,$$

$$\|u\|_{C^{m,\theta_0,\theta_1}(0,1)} := \sum_{k=0}^m \sup_{0<x<1} w_{k-1+\theta_0}(x)w_{k-1+\theta_1}(1-x)|u^{(k)}(x)| < \infty,$$

where

$$w_\rho(r) = \begin{cases} 1, & \rho<0, \\ 1/(1+|\log r|), & \rho=0, \\ r^\rho, & \rho>0, \end{cases} \qquad r,\rho\in\mathbb{R}, r>0.$$

Clearly, $C^m[0,1]\subset C^{m,\theta_0,\theta_1}(0,1)\subset C^m_\star(0,1)$.

LEMMA 1.2 (see [11]). *Let $a, b\in C^m([0,1]\times(0,1))$ and*

$$|\partial_x^k\partial_y^l a(x,y)| \le cy^{-\lambda_0-l}(1-y)^{-\lambda_1-l}, \quad |\partial_x^k\partial_y^l b(x,y)| \le cy^{-\mu_0-l}(1-y)^{-\mu_1-l}, \tag{1.2}$$
$$(x,y)\in[0,1]\times(0,1), k+l\le m,$$

where $m\in\mathbb{N}$, $0<\nu<1$, $\lambda_0+\nu<1$, $\lambda_1+\nu<1$, $\mu_0<1$, $\mu_1<1$. Then T maps $C^m_\star(0,1)$ into $C^m_\star(0,1)$, $T: C^m_\star(0,1)\to C^m_\star(0,1)$ is bounded, and $T^2: C^m_\star(0,1)\to C^m_\star(0,1)$ is compact. Further, T maps $C^{m,\theta_0,\theta_1}(0,1)$ with $\theta_0 = \max\{\lambda_0+\nu, \mu_0\}$, $\theta_1 = \max\{\lambda_1+\nu, \mu_1\}$ into $C^{m,\theta_0,\theta_1}(0,1)$, and $T: C^{m,\theta_0,\theta_1}(0,1)\to C^{m,\theta_0,\theta_1}(0,1)$ is compact.

Here $\partial_x^k\partial_y^l = (\partial/\partial x)^k(\partial/\partial y)^l$. An example of a satisfying (1.2) is given by $a(x,y) = y^{-\lambda_0}(1-y)^{-\lambda_1}\overline{a}(x,y)$, where $\overline{a}\in C^m([0,1]\times[0,1])$.

Denote $\mathcal{N}(I-T)=\{u\in C[0,1]: u=Tu\}$. The following theorem is a consequence of Lemmas 1.1 and 1.2.

THEOREM 1.3. *Assume the conditions of Lemma* 1.2 *and* $\mathcal{N}(I-T)=\{0\}$. *Then for* $f\in C^m_\star(0,1)$, (1.1) *has a solution* $u\in C^m_\star(0,1)$ *which is unique in* $C[0,1]$, *and* $\|u\|_{C^m_\star(0,1)}\le c\|f\|_{C^m_\star(0,1)}$, *where the constant* c *is independent of* f. *Further, if* $f\in C^{m,\theta_0,\theta_1}(0,1)$, $\theta_0=\max\{\lambda_0+\nu,\mu_0\}$, $\theta_1=\max\{\lambda_1+\nu,\mu_1\}$, *then* $u\in C^{m,\theta_0,\theta_1}(0,1)$, *and* $\|u\|_{C^{m,\theta_0,\theta_1}(0,1)}\le c_{m,\theta_0,\theta_1}\|f\|_{C^{m,\theta_0,\theta_1}(0,1)}$, *where the constant* c_{m,θ_0,θ_1} *is independent of* f.

The main results of the paper are established under assumptions of Theorem 1.3. Using the approach of [12], our treatment can be extended to the case where f has a singularity at a point $x=\xi$, $\xi\in(0,1)$, and/or a and b have certain singularities on the lines $x=\xi$ and $y=\xi$ or on a system of such inner points/lines—we can build the approximate solution on $[0,\xi]$ and $[\xi,1]$ in a manner similar to that done on $[0,1]$ in section 4.

The rest of the paper is organized as follows. In section 2 we recall some results about the interpolation and quasi-interpolation of functions by splines with a uniform knot set and, in particular, about the optimality of such approximations. In section 3 we transform (1.1) into a new form by the change of variables. Section 4 is central in the paper—we introduce the product quasi-interpolation method for the transformed equation, examine its accuracy, and present the matrix form of the method. In section 5 we discuss different computational details of the method, and in section 6 we test the algorithms on a numerical example.

Actually the method can be applied to a somewhat more general class of equations rather than (1.1). The case of an integral equation with logarithmic diagonal singularity of the kernel will be treated in a separate work with more attention to the computational environment. In [17] the method is modified for the Volterra integral equation corresponding to (1.1).

2. Approximation tools.

2.1. The father B-spline. The *father* B*-spline* B_m *of order* m in the terminology of [4, 15], or of degree $m-1$ in the terminology of [6, 16], can be defined by the formula

$$B_m(x)=\frac{1}{(m-1)!}\sum_{i=0}^{m}(-1)^i\mathcal{C}_i^{(m)}(x-i)_+^{m-1},\quad x\in\mathbb{R},\quad m\in\mathbb{N}. \tag{2.1}$$

Here, as usual, $0!=1$, $0^0:=\lim_{x\downarrow 0}x^x=1$,

$$\mathcal{C}_i^{(m)}=\left(\begin{array}{c} m\\ i\end{array}\right)=\frac{m!}{i!(m-i)!},\quad (x-i)_+^{m-1}=\begin{cases}(x-i)^{m-1}, & x-i\ge 0,\\ 0, & x-i<0.\end{cases}$$

Let us recall some properties of B_m: $B_m\in C^{(m-2)}(\mathbb{R})$,

$$\mathrm{supp}B_m=[0,m],\quad B_m(x)=B_m(m-x)>0 \text{ for } 0<x<m, \tag{2.2}$$

$$B_m^{(m-1)}(x)=(-1)^l\mathcal{C}_l^{(m-1)} \text{ for } l<x<l+1,\quad l=0,\dots,m-1, \tag{2.3}$$

$$\sum_{j\in\mathbb{Z}}B_m(x-j)=1 \text{ for } x\in\mathbb{R},\quad \int_{\mathbb{R}}B_m(x)dx=1. \tag{2.4}$$

2.2. Spline interpolation on the uniform grid in $\mathbb{R}$. Introduce in $\mathbb{R}$ the uniform grid $h\mathbb{Z} = \{ih : i \in \mathbb{Z}\}$ of the step size $h > 0$. Denote by $S_{h,m}$, $m \in \mathbb{N}$, the space of splines of order m (of degree $m-1$) and defect 1 with the knot set $h\mathbb{Z}$. The family B-spline $B_m(h^{-1}x - j)$, $j \in \mathbb{Z}$, belongs to $S_{h,m}$, and the same is true for $\sum_{j\in\mathbb{Z}} d_j B_m(h^{-1}x - j)$ with arbitrary coefficients d_j; there are no problems with the convergence of the series since it is locally finite: it follows from (2.2) that

$$\sum_{j\in\mathbb{Z}} d_j B_m(h^{-1}x - j) = \sum_{j=i-m+1}^{i} d_j B_m(h^{-1}x - j) \text{ for } x \in [ih, (i+1)h), \quad i \in \mathbb{Z}.$$

Given a function $f \in C(\mathbb{R})$, bounded or of at most polynomial growth as $|x| \to \infty$, we determine the interpolant $Q_{h,m}f \in S_{h,m}$ by the conditions

$$(Q_{h,m}f)(x) = \sum_{j\in\mathbb{Z}} d_j B_m(h^{-1}x - j), \quad x \in \mathbb{R}, \tag{2.5}$$

$$(Q_{h,m}f)\left(\left(k + \frac{m}{2}\right)h\right) = f\left(\left(k + \frac{m}{2}\right)h\right), \qquad k \in \mathbb{Z}. \tag{2.6}$$

For $m = 1$ and $m = 2$, $Q_{h,m}f$ is the usual piecewise constant, respectively, piecewise linear interpolant, which can be determined on every subinterval $[ih, (i+1)h)$, $i \in \mathbb{Z}$, independently of other subintervals. For $m \geq 3$, the value of $Q_{h,m}f$ at a given point $x \in \mathbb{R}$ depends on the values of f at all interpolation knots $(k + \frac{m}{2})h$, $k \in \mathbb{Z}$. It occurs (see [14, 16]) that for $m \geq 3$ conditions (2.5)–(2.6) really determine d_j, $j \in \mathbb{Z}$, uniquely in the space of bounded or polynomially growing bisequences (d_j), namely,

$$d_j = \sum_{k\in\mathbb{Z}} \alpha_{j-k,m} f\left(\left(k + \frac{m}{2}\right)h\right) = \sum_{k\in\mathbb{Z}} \alpha_{k,m} f\left(\left(j - k + \frac{m}{2}\right)h\right), \quad j \in \mathbb{Z}, \tag{2.7}$$

where

$$\alpha_{k,m} = \sum_{l=1}^{m_0} \frac{z_{l,m}^{m_0-1}}{P_m'(z_{l,m})} z_{l,m}^{|k|}, \quad k \in \mathbb{Z}, \quad m_0 = \begin{cases} (m-2)/2 & \text{if } m \text{ is even,} \\ (m-1)/2 & \text{if } m \text{ is odd,} \end{cases} \tag{2.8}$$

and $z_{l,m} \in (-1, 0)$, $l = 1, \ldots, m_0$, are the roots of the characteristic polynomial

$$P_m(z) = \sum_{|k|\leq m_0} B_m\left(k + \frac{m}{2}\right) z^{k+m_0}$$

(it is a polynomial of degree $2m_0$); it occurs that P_m has exactly m_0 simple roots $z_{l,m}$, $l = 1, \ldots, m_0$, in the interval $(-1, 0)$, and the remaining m_0 roots are of the form $z_{l+m_0,m} = 1/z_{l,m} \in (-\infty, -1)$, $l = 1, \ldots, m_0$.

Denote by $BC(\mathbb{R})$ the space of bounded continuous functions on $\mathbb{R}$ equipped with the norm $\|f\|_\infty = \sup_{x\in\mathbb{R}} |f(x)|$, by $V^{m,\infty}(\mathbb{R})$ the space of functions having bounded mth (distributional) derivative in $\mathbb{R}$, by $W^{m,\infty}(\mathbb{R})$ the standard Sobolev space of functions on $\mathbb{R}$ having bounded derivatives of order $\leq m$, and by $W^{m,\infty}_{(0,1)}(\mathbb{R})$ the space of functions $f \in W^{m,\infty}(\mathbb{R})$ with supports in $(0, 1)$.

LEMMA 2.1 (see [6, 21, 22]). *For $f \in V^{m,\infty}(\mathbb{R})$, it holds that $f - Q_{h,m}f \in BC(\mathbb{R})$ and*

$$\|f - Q_{h,m}f\|_\infty \leq \Phi_{m+1}\pi^{-m}h^m\|f^{(m)}\|_\infty, \tag{2.9}$$

where $\Phi_m = \frac{4}{\pi}\sum_{k=0}^{\infty} \frac{(-1)^{km}}{(2k+1)^m}$, $m \in \mathbb{N}$, is the Favard constant,

$$\Phi_1 < \Phi_3 < \Phi_5 < \cdots < 4/\pi < \cdots < \Phi_6 < \Phi_4 < \Phi_2, \quad \lim_{m\to\infty} \Phi_m = 4/\pi.$$

The following lemma tells that, in some sense, the spline interpolation yields the best approximation of the function classes $W^{m,\infty}(\mathbb{R})$ and $V^{m,\infty}(\mathbb{R})$, asymptotically also of $W^{m,\infty}_{(0,1)}(\mathbb{R})$, compared with other methods that use the same information as $Q_{h,m}f$—the values $f|_{\mathbb{Z}_{h,m}}$ of f on the grid $\mathbb{Z}_{h,m} = \{(j + \frac{m}{2})h : j \in \mathbb{Z}\}$. Denote by $C(\mathbb{Z}_{h,m})$ the vector space of all (grid) functions defined on $\mathbb{Z}_{h,m}$.

LEMMA 2.2 (see [21, 22]). *For given $\gamma > 0$, we have by Lemma* 2.1

$$\sup_{f\in V^{m,\infty}(\mathbb{R}),\|f^{(m)}\|_\infty\le\gamma} \|f - Q_{h,m}f\|_\infty \le \Phi_{m+1}\pi^{-m}h^m\gamma,$$

whereas for any mapping $M_h : C(\mathbb{Z}_{h,m}) \to C(\mathbb{R})$ (linear or nonlinear, continuous or discontinuous), it holds that

$$\sup_{f\in W^{m,\infty}(\mathbb{R}),\|f^{(m)}\|_\infty\le\gamma} \|f - M_h(f|_{\mathbb{Z}_{h,m}})\|_\infty \ge \Phi_{m+1}\pi^{-m}h^m\gamma,$$

$$\liminf_{h\to 0} \sup_{f\in W^{m,\infty}_{(0,1)}(\mathbb{R}),\|f^{(m)}\|_\infty\le\gamma} \|f - M_h(f|_{\mathbb{Z}_{h,m}})\|_\infty/(\Phi_{m+1}\pi^{-m}h^m\gamma) \ge 1.$$

2.3. Quasi-interpolation by splines. Let $m \ge 3$. In a quasi-interpolant $Q^{(p)}_{h,m}f$, the infinite sum (2.7) defining the coefficients d_j of the spline interpolant (2.5)–(2.6) is replaced by a finite sum:

$$(Q^{(p)}_{h,m}f)(x) = \sum_{j\in\mathbb{Z}} d^{(p)}_j B_m(h^{-1}x - j), \quad x \in \mathbb{R}, \tag{2.10}$$

$$d^{(p)}_j = \sum_{|k|\le p-1} \alpha^{(p)}_{k,m} f\left(\left(j - k + \frac{m}{2}\right)h\right), \quad p \in \mathbb{N}. \tag{2.11}$$

A simple truncation of the series in (2.7) does not give acceptable results. Using a special difference calculus for fast decaying bisequences, the following formulae for $\alpha^{(p)}_{k,m}$ have been proposed in [8]:

$$\alpha^{(p)}_{k,m} = \sum_{q=|k|}^{p-1} (-1)^{k+q} \begin{pmatrix} 2q \\ k+q \end{pmatrix} \gamma_{q,m}, \quad |k| \le p-1, \tag{2.12}$$

$$\gamma_{0,m} = 1, \quad \gamma_{q,m} = \sum_{l=1}^{m_0} \frac{(1 + z_{l,m})z^{m_0+q-1}_{l,m}}{(1 - z_{l,m})^{2q+1}P'_m(z_{l,m})}, \quad q \ge 1. \tag{2.13}$$

Then it occurs that

$$\|Q_{h,m}f - Q^{(p)}_{h,m}f\|_\infty \le c_{m,p}h^{2p}\|f^{(2p)}\|_\infty \text{ for } f \in V^{2p,\infty}(\mathbb{R}) \tag{2.14}$$

with a constant $c_{m,p}$ that can be described. A consequence of (2.14) is that for $f \in V^{m,\infty}(\mathbb{R})$ with uniformly continuous $f^{(m)}$ and $2p > m$, it holds that $\|Q_{h,m}f - Q^{(p)}_{h,m}f\|_\infty h^{-m} \to 0$ as $h \to 0$, i.e., the quasi-interpolant $Q^{(p)}_{h,m}f$ is asymptotically of the same accuracy as the interpolant $Q_{h,m}f$. It is reasonable to take the smallest $p \in \mathbb{N}$ for which $2p > m$; denote it by m_1,

$$m_1 = \left\{ \begin{array}{ll} \frac{m}{2} + 1, & m \text{ even} \\ \frac{m+1}{2}, & m \text{ odd} \end{array} \right\} = m - m_0. \tag{2.15}$$

Denote also $Q'_{h,m} := Q^{(m_1)}_{h,m}$, $\alpha'_{k,m} := \alpha^{(m_1)}_{k,m}$, $|k| < m_1$. Note that $(Q'_{h,m}f)(x)$ is well defined for $x \in [ih, (i+1)h]$ with an $i \in \mathbb{Z}$ if f is given on $[(i-m+1)h, (i+m)h] \cap \mathbb{Z}_{h,m}$.

LEMMA 2.3 (see [8]). *For $i \in \mathbb{Z}$, $f \in C^m[(i-m+1)h, (i+m)h]$, it holds that*

$$\max_{ih \le x \le (i+1)h} |f(x) - (Q'_{h,m}f)(x)| \le (\Phi_{m+1}\pi^{-m} + q_m c'_m)h^m \sup_{(i-m+1)h \le x \le (i+m)h} |f^{(m)}(x)|. \tag{2.16}$$

For a relatively compact set $\mathcal{M}$ in $C[-\delta, 1+\delta]$, $\delta > 0$, it holds that

$$\sup_{f \in \mathcal{M}} \max_{0 \le x \le 1} |f(x) - (Q'_{h,m}f)(x)| \to 0 \text{ as } h \to 0. \tag{2.17}$$

Formulae for $q_m := \|Q_{h,m}\|_{BC(\mathbb{R}) \to BC(\mathbb{R})}$, $q'_m := \|Q'_{h,m}\|_{BC(\mathbb{R}) \to BC(\mathbb{R})}$, and c'_m from (2.16) are presented in [8] with the following numerical values:

m	3	4	5	6	7	8	9	10	20
c'_m	0.016	0.019	0.015	0.0085	0.0060	0.0030	0.0022	0.0010	6.5e-6
q_m	1.414	1.549	1.706	1.816	1.916	2.000	2.075	2.142	2.583
$q_m c'_m$	0.023	0.029	0.026	0.015	0.012	0.006	0.005	0.002	1.7e-5
q'_m	1.250	1.354	1.329	1.403	1.356	1.413	1.378	1.419	1.514

Relatively small values of the norms q_m and q'_m tell us about good stability properties of interpolation and quasi-interpolation processes.

We assumed that $m \ge 3$. For $m = 1$ and $m = 2$, we may put $Q'_{h,m} = Q_{h,m}$.

3. A smoothing change of variables. In the integral equation (1.1), we perform the change of variables

$$x = \varphi(t), \quad y = \varphi(s), \quad 0 \le t \le 1, \quad 0 \le s \le 1, \tag{3.1}$$

where $\varphi : [0,1] \to [0,1]$ is defined by the formula

$$\varphi(t) = \frac{1}{c_\star} \int_0^t \sigma^{r_0 - 1}(1-\sigma)^{r_1 - 1} d\sigma, \tag{3.2}$$

$$c_\star = \int_0^1 \sigma^{r_0 - 1}(1-\sigma)^{r_1 - 1} d\sigma = \frac{\Gamma(r_0)\Gamma(r_1)}{\Gamma(r_0 + r_1)},$$

Γ is the Euler gamma function. We assume that r_0, $r_1 \in \mathbb{R}$, $r_0 \ge 1$, $r_1 \ge 1$, but we keep in mind that practicable algorithms correspond to the case where at least one of the parameters r_0 and r_1 is natural. If $r_1 \in \mathbb{N}$, (3.2) takes the form

$$\varphi(t) = \frac{1}{c_\star} t^{r_0} \sum_{k=0}^{r_1 - 1} \frac{(-1)^k}{r_0 + k} \mathcal{C}_k^{(r_1 - 1)} t^k, \quad 0 \le t \le 1, \quad c_\star = \frac{(r_1 - 1)!}{r_0(r_0+1)\dots(r_0 + r_1 - 1)},$$

but in our computations, the following expansion of φ with positive terms occurred to be preferable (essentially more stable numerically):

$$\varphi(t) = t^{r_0} \left[1 + \sum_{k=1}^{r_1 - 1} \frac{r_0(r_0+1)\dots(r_0+k-1)}{k!} (1-t)^k \right], \quad 0 \le t \le 1. \tag{3.3}$$

If $r_0, r_1 \in \mathbb{N}$, the integral in (3.2) can be alternatively evaluated [9] in a stable way by an exact Gauss rule, since the integrand is a polynomial of degree $r_0 + r_1 - 2$.

Clearly, $\varphi(0) = 0$, $\varphi(1) = 1$, and $\varphi(t)$ is strictly increasing in $[0,1]$. Hence $\frac{\varphi(t)-\varphi(s)}{t-s} > 0$, $|\varphi(t) - \varphi(s)| = \frac{\varphi(t)-\varphi(s)}{t-s}|t-s|$ for $s \neq t$, and (1.1) takes with respect to $v(t) = u(\varphi(t))$ the form

$$v(t) = \int_0^1 \left(\mathcal{A}(t,s)|t-s|^{-\nu} + \mathcal{B}(t,s)\right) v(s)ds + g(t), \quad 0 \le t \le 1, \tag{3.4}$$

where

$$\mathcal{A}(t,s) = a(\varphi(t),\varphi(s))\Phi(t,s)^{-\nu}\varphi'(s), \qquad \Phi(t,s) = \begin{cases} \frac{\varphi(t)-\varphi(s)}{t-s}, & t \neq s, \\ \varphi'(s), & t = s, \end{cases} \tag{3.5}$$

$$\mathcal{B}(t,s) = b(\varphi(t),\varphi(s))\varphi'(s), \qquad g(t) = f(\varphi(t)). \tag{3.6}$$

Let us characterize the boundary behavior of functions in (3.4). Clearly,

$$\begin{gathered} 0 \le \varphi(t) \le ct^{r_0}, \quad 0 \le 1 - \varphi(t) \le c(1-t)^{r_1}, \\ |\varphi^{(k)}(t)| \le ct^{r_0-k}(1-t)^{r_1-k}, \quad 0 < t < 1, \quad k = 1, \dots, m. \end{gathered} \tag{3.7}$$

LEMMA 3.1. *For $u \in C^m_\star(0,1)$, $v(t) = u(\varphi(t))$, it holds that $v \in C^m_\star(0,1)$ and*

$$\|v\|_{C^m_\star(0,1)} \le c\|u\|_{C^m_\star(0,1)}, \tag{3.8}$$

where the constant c is independent of u. Further, if $u \in C^{m,\theta_0,\theta_1}(0,1)$, $\theta_0 < 1$, $\theta_1 < 1$, then for $j = 1, \dots, m$, $0 < t < 1$,

$$\begin{aligned} |v^{(j)}(t)| &\le c\|u\|_{C^{m,\theta_0,\theta_1}(0,1)} \\ &\times \left\{ \begin{array}{r} t^{r_0-j}, \quad \theta_0 < 0 \\ t^{r_0-j}(1+|\log t|), \quad \theta_0 = 0 \\ t^{(1-\theta_0)r_0-j}, \quad \theta_0 > 0 \end{array} \right\} \left\{ \begin{array}{r} (1-t)^{r_1-j}, \quad \theta_1 < 0 \\ (1-t)^{r_1-j}(1+|\log(1-t)|), \quad \theta_1 = 0 \\ (1-t)^{(1-\theta_1)r_1-j}, \quad \theta_1 > 0 \end{array} \right\}. \end{aligned} \tag{3.9}$$

Proof. Clearly $\|v\|_\infty = \|u\|_\infty$. The proof of (3.8) and (3.9) is based on (3.7) and the formula of Faà di Bruno,

$$\left(\frac{d}{dt}\right)^j u(\varphi(t)) = \sum_{k_1+\cdots+jk_j=j} \frac{j!}{k_1!..k_j!} u^{(k_1+\cdots+k_j)}(\varphi(t)) \left(\frac{\varphi'(t)}{1!}\right)^{k_1} \cdots \left(\frac{\varphi^{(j)}(t)}{j!}\right)^{k_j},$$

where the sum is taken over all $k_1 \ge 0, \dots, k_j \ge 0$ such that $k_1 + 2k_2 + \cdots + jk_j = j$. Here are the details for (3.9). For $0 < t \le \frac{1}{2}$, $1 \le j \le m$, we have

$$\begin{aligned} |v^{(j)}(t)| &\le c \sum_{k_1+2k_2\dots+jk_j=j} \left\{ \begin{array}{cc} 1, & k_1+\cdots+k_j < 1-\theta_0 \\ 1+|\log t|, & k_1+\cdots+k_j = 1-\theta_0 \\ t^{(1-\theta_0-k_1-\cdots-k_j)r_0}, & k_1+\cdots+k_j > 1-\theta_0 \end{array} \right\} \\ &\times \|u\|_{C^{m,\theta_0,\theta_1}} t^{k_1(r_0-1)} t^{k_2(r_0-2)} \dots t^{k_j(r_0-j)} \\ &= c \sum_{k_1+\cdots+jk_j=j} \left\{ \begin{array}{cc} t^{(k_1+\cdots+k_j)r_0-j}, & k_1+\cdots+k_j < 1-\theta_0 \\ t^{(1-\theta_0)r_0-j}(1+|\log t|), & k_1+\cdots+k_j = 1-\theta_0 \\ t^{(1-\theta_0)r_0-j}, & k_1+\cdots+k_j > 1-\theta_0 \end{array} \right\} \|u\|_{C^{m,\theta_0,\theta_1}}. \end{aligned}$$

The condition $k_1+2k_2\cdots+jk_j=j$ implies the following: (i) The relation $k_1+\cdots+k_j<1-\theta_0$ is possible only if $\theta_0<0$, and then t^{r_0-j} is the leading term as $t\to 0$ in the last sum (it corresponds to $k_1=\cdots=k_{j-1}=0$, $k_j=1$). (ii) The relation $k_1+\cdots+k_j=1-\theta_0$ is possible only if $\theta_0\le 0$, $\theta_0\in\mathbb{Z}$; in the case $\theta_0=0$, the leading term is $t^{r_0-j}(1+|\log t|)$, whereas in case $\theta_0\le -1$, similarly as in (i), the leading term is t^{r_0-j}. (iii) If $0<\theta_0<1$, then all terms in the last sum are $t^{(1-\theta_0)r_0-j}$. These considerations prove (3.9) for $0<t\le\frac{1}{2}$. By an analogous argument we obtain (3.9) also for $\frac{1}{2}\le t<1$. □

LEMMA 3.2. *Let a and b satisfy the conditions of Lemma* 1.2. *Then for $j=0,\dots,m$, $0\le t\le 1$, $0<s<1$, it holds that*

$$|\partial_s^j a(\varphi(t),\varphi(s))|\le cs^{-r_0\lambda_0-j}(1-s)^{-r_1\lambda_1-j}, \tag{3.10}$$

$$|\partial_s^j b(\varphi(t),\varphi(s))|\le cs^{-r_0\mu_0-j}(1-s)^{-r_1\mu_1-j}. \tag{3.11}$$

These inequalities elementarily follow by the formula of Faà di Bruno.

The function $\Phi(t,s)^{-\nu}$ has singularities at $(0,0)$ and $(1,1)$, the only zeros of $\Phi(t,s)$ in $[0,1]\times[0,1]$. It is easy to see that

$$\partial_s^k\Phi(t,s)\asymp(t+s)^{r_0-k-1}((1-t)+(1-s))^{r_1-k-1} \text{ as } t,s\to 0 \text{ or as } t,s\to 1,$$

which together with the formula of Faà di Bruno implies the following result.

LEMMA 3.3. *For $j=0,\dots,m$, $0\le t\le 1$, $0<s<1$, it holds that*

$$\left|\partial_s^j\left((\Phi(t,s))^{-\nu}\right)\right|\le c(t+s)^{-\nu(r_0-1)-j}((1-t)+(1-s))^{-\nu(r_1-1)-j}. \tag{3.12}$$

Next we present estimates for functions $\mathcal{A}(t,s)$, $\mathcal{B}(t,s)$, and $\partial_s^m[\mathcal{A}(t,s)v(s)]$, $\partial_s^m[\mathcal{B}(t,s)v(s)]$ in somewhat specific form for the needs of section 4.

COROLLARY 3.4. *Let a and b satisfy the conditions of Lemma* 1.1. *Then the following holds true:* (i) *if*

$$r_0,r_1\ge 1,\quad r_0>(1-\nu)/(1-\nu-\lambda_0),\quad r_1>(1-\nu)/(1-\nu-\lambda_1), \tag{3.13}$$

then with $\delta_0:=(1-\nu-\lambda_0)r_0-(1-\nu)>0$, $\delta_1:=(1-\nu-\lambda_1)r_1-(1-\nu)>0$, it holds that

$$|\mathcal{A}(t,s)|\le cs^{\delta_0}(1-s)^{\delta_1},\quad (t,s)\in[0,1]\times(0,1); \tag{3.14}$$

(ii) *if*

$$r_0,r_1\ge 1,\quad r_0>1/(1-\mu_0),\quad r_1>1/(1-\mu_1), \tag{3.15}$$

then with $\delta_0:=(1-\mu_0)r_0-1>0$, $\delta_1:=(1-\mu_1)r_1-1>0$, it holds that

$$|\mathcal{B}(t,s)|\le cs^{\delta_0}(1-s)^{\delta_1},\quad (t,s)\in[0,1]\times(0,1). \tag{3.16}$$

Proof. These estimates are direct consequences of Lemmas 3.2 and 3.3. By (3.7), (3.10), and (3.12) we have

$$|\mathcal{A}(t,s)|\le cs^{-r_0\lambda_0+r_0-1}(t+s)^{-\nu(r_0-1)}(1-s)^{-r_1\lambda_1+r_1-1}(2-t-s)^{-\nu(r_1-1)},$$

which for r_0, r_1 satisfying (3.13) yields (3.14). Similarly, by (3.7), (3.11), and (3.12),

$$|\mathcal{B}(t,s)|\le cs^{-r_0\mu_0+r_0-1}(1-s)^{-r_1\mu_1+r_1-1},$$

which for r_0, r_1 satisfying (3.15) yields (3.16). □

COROLLARY 3.5. *Let a and b satisfy the conditions of Lemma* 1.2, *and let $u \in C^m_\star(0,1)$, $v(t) = u(\varphi(t))$. Then the following estimates hold true for $(t,s) \in [0,1] \times (0,1)$:* (i) *if*

$$r_0, r_1 \geq 1, \quad r_0 > m/(1-\nu-\lambda_0), \quad r_1 > m/(1-\nu-\lambda_1), \tag{3.17}$$

then with $\delta_0 := (1-\nu-\lambda_0)r_0 - m > 0$, $\delta_1 := (1-\nu-\lambda_1)r_1 - m > 0$,

$$|\mathcal{A}(t,s)| \leq cs^{m-(1-\nu)+\delta_0}(1-s)^{m-(1-\nu)+\delta_1}, \tag{3.18}$$

$$|\partial_s^m[\mathcal{A}(t,s)v(s)]| \leq cs^{-(1-\nu)+\delta_0}(1-s)^{-(1-\nu)+\delta_1}\|u\|_{C^m_\star(0,1)}; \tag{3.19}$$

(ii) *if*

$$r_0, r_1 \geq 1, \quad r_0 > m/(1-\mu_0), \quad r_1 > m/(1-\mu_1), \tag{3.20}$$

then with $\delta_0 := (1-\mu_0)r_0 - m > 0$, $\delta_1 := (1-\mu_1)r_1 - m > 0$,

$$|\mathcal{B}(t,s)| \leq cs^{m-1+\delta_0}(1-s)^{m-1+\delta_1}, \tag{3.21}$$

$$|\partial_s^m[\mathcal{B}(t,s)v(s)]| \leq cs^{-1+\delta_0}(1-s)^{-1+\delta_1}\|u\|_{C^m_\star(0,1)}; \tag{3.22}$$

(iii) *if*

$$r_0, r_1 \geq 1, \quad r_0 > (m+\nu)/(1-\lambda_0), \quad r_1 > (m+\nu)/(1-\lambda_1), \tag{3.23}$$

then with $\delta_0 := (1-\lambda_0)r_0 - m - \nu > 0$, $\delta_1 := (1-\lambda_1)r_1 - m - \nu > 0$,

$$\varphi'(t)^\nu|\mathcal{A}(t,s)| \leq cs^{m-(1-\nu)+\delta_0}(1-s)^{m-(1-\nu)+\delta_1}, \tag{3.24}$$

$$\varphi'(t)^\nu|\partial_s^m[\mathcal{A}(t,s)v(s)]| \leq cs^{-(1-\nu)+\delta_0}(1-s)^{-(1-\nu)+\delta_1}\|u\|_{C^m_\star(0,1)}. \tag{3.25}$$

Proof. These estimates are direct consequences of Lemmas 3.1–3.3. Observe that $|\partial_s^k[\varphi'(t)^\nu\Phi(t,s)^{-\nu}]| \leq c(t+s)^{-k}((1-t)+(1-s))^{-k}$. ∎

Let us extend $\mathcal{A}(t,s)$ and $\mathcal{B}(t,s)$ with respect to s outside $(0,1)$ by the zero value. Under conditions (3.13) and (3.15) we obtain continuous functions on $[0,1] \times \mathbb{R}$; see (3.14) and (3.16).

4. The product quasi-interpolation method.

4.1. The operator form of the method. Let $h = 1/n$, $n \in \mathbb{N}$, $n \geq m$. For $0 \leq s \leq 1$, the quasi-interpolant $Q'_{h,m}w$ of a function $w \in C[-\delta, 1+\delta]$ has the expansion

$$(Q'_{h,m}w)(s) = \sum_{j=-m+1}^{n-1} \sum_{|p|<m_1} \alpha'_{p,m} w\left(\left(j-p+\frac{m}{2}\right)h\right) B_m(ns-j). \tag{4.1}$$

We approximate (3.4) by its discretized version: for $0 \leq t \leq 1$,

$$v_n(t) = \int_0^1 [|t-s|^{-\nu}Q'_{h,m}(\mathcal{A}(t,s)v_n(s)) + Q'_{h,m}(\mathcal{B}(t,s)v_n(s))]ds + g(t), \tag{4.2}$$

where $Q'_{h,m}$ is applied to the products $\mathcal{A}(t,s)v(s)$ and $\mathcal{B}(t,s)v(s)$ as functions of s. This can be done for v given on $[0,1]$ since $\mathcal{A}(t,s) = \mathcal{B}(t,s) = 0$ for $s \leq 0$ and for $s \geq 1$; recall that under conditions (3.13) and (3.15), $\mathcal{A}, \mathcal{B} \in C([0,1] \times \mathbb{R})$.

THEOREM 4.1. (i) *Let a and b satisfy the conditions of Lemma* 1.1, $\mathcal{N}(I-T)=\{0\}$, $f\in C[0,1]$, *and let r_0 and r_1 satisfy conditions* (3.13) *and* (3.15). *Then for sufficiently large n,* (4.2) *has a unique solution v_n, and*

$$\|v-v_n\|_\infty := \max_{0\le t\le 1}|v(t)-v_n(t)|\to 0 \text{ as } n\to\infty, \tag{4.3}$$

where v is the solution of (3.4).

(ii) *Under the assumptions of Theorem* 1.3, $f\in C^m_\star(0,1)$, *and conditions* (3.17) *and* (3.20) *on r_0 and r_1, it holds that*

$$\|v-v_n\|_\infty \le ch^m\|f\|_{C^m_\star(0,1)}. \tag{4.4}$$

(iii) *Under the assumptions of Theorem* 1.3, $f\in C^m_\star(0,1)$, *and conditions* (3.20) *and* (3.23) *on r_0 and r_1, it holds that*

$$\max_{0\le t\le 1}\varphi'(t)^\nu|v(t)-v_n(t)|\le ch^m\|f\|_{C^m_\star(0,1)}. \tag{4.5}$$

The constant c in (4.4) *and* (4.5) *is independent of n and f.*

Proof. Accept the assumptions formulated in (i). Denote by $\mathcal{T}$ and $\mathcal{T}_n$ the integral operators of (3.4) and (4.2),

$$(\mathcal{T}v)(t)=\int_0^1[|t-s|^{-\nu}\mathcal{A}(t,s)+\mathcal{B}(t,s)]v(s)ds,$$

$$(\mathcal{T}_nv)(t)=\int_0^1[|t-s|^{-\nu}Q'_{h,m}(\mathcal{A}(t,s)v(s))+Q'_{h,m}(\mathcal{B}(t,s)v(s))]ds.$$

We claim that $\mathcal{T}_n\to\mathcal{T}$ compactly in $C[0,1]$ as $n\to\infty$, i.e.,

$$\|\mathcal{T}_nv-\mathcal{T}v\|_\infty\to 0 \text{ for every } v\in C[0,1], \tag{4.6}$$

$$(v_n)\subset C[0,1], \|v_n\|_\infty\le 1 \Rightarrow (\mathcal{T}_nv_n) \text{ is relatively compact in } C[0,1]. \tag{4.7}$$

Indeed, the sets $\{\mathcal{A}(t,\cdot):0\le t\le 1\}$ and $\{\mathcal{B}(t,\cdot):0\le t\le 1\}$ are relatively compact in $C[-\delta,1+\delta]$, and by Lemma 2.3, for a fixed $v\in C[0,1]$ extended by $v(t)=v(0)$ for $-\delta\le s\le 0$, $v(t)=v(1)$ for $1\le s\le 1+\delta$, it holds that

$$\sup_{0\le t\le 1}\max_{0\le s\le 1}|\mathcal{A}(t,s)v(s)-Q'_{h,m}(\mathcal{A}(t,s)v(s))|\to 0 \text{ for } n\to\infty,$$

$$\sup_{0\le t\le 1}\max_{0\le s\le 1}|\mathcal{B}(t,s)v(s)-Q'_{h,m}(\mathcal{B}(t,s)v(s))|\to 0 \text{ for } n\to\infty.$$

This together with the equality $\|Q'_{h,m}\|=q'_m$ implies (4.6). The proof of (4.7) can be built using the Arzelà theorem.

Due to the condition $\mathcal{N}(I-T)=\{0\}$, also $\mathcal{N}(I-\mathcal{T})=\{0\}$. As is well known (see [1, 2, 7] or [18]), relations (4.6), (4.7) and $\mathcal{N}(I-\mathcal{T})=\{0\}$ imply that, for sufficiently large n, the operators $I-\mathcal{T}_n$ are invertible and the inverses are uniformly bounded:

$$\|(I-\mathcal{T}_n)^{-1}\|_{C[0,1]\to C[0,1]}\le c, \qquad n\ge n_0. \tag{4.8}$$

Let v and v_n be the solutions of (3.4) and (4.2), respectively. Then

$$v-v_n=(I-\mathcal{T}_n)^{-1}(\mathcal{T}v-\mathcal{T}_nv), \tag{4.9}$$

and due to (4.8) and (4.6), the convergence (4.3) follows.

Next we establish the uniform estimate (4.4). Assume the conditions of Theorem 1.3, $f \in C^m_\star(0,1)$, and (3.17), (3.20) hold. For the solutions u and v of (1.1) and (3.4) we have $v(t) = u(\varphi(t))$ and $u \in C^m_\star(0,1)$ by Theorem 1.3. On the basis of (2.16) and (3.19),

$$\begin{aligned}\int_{2mh}^{1-2mh} &|t-s|^{-\nu}|\mathcal{A}(t,s)v(s) - Q'_{h,m}(\mathcal{A}(t,s)v(s))|ds \\ &\le ch^m \int_{mh}^{1-mh} |t-s|^{-\nu} s^{-(1-\nu)+\delta_0}(1-s)^{-(1-\nu)+\delta_1} ds \|u\|_{C^m_\star(0,1)} \\ &\le c'h^m \|u\|_{C^m_\star(0,1)},\end{aligned}$$

where we have taken into account that for $2mh \le s \le 1-2mh$, it holds that

$$\max_{|\sigma-s|\le mh} \sigma^{-(1-\nu)+\delta_0}(1-\sigma)^{-(1-\nu)+\delta_1} \le cs^{-(1-\nu)+\delta_0}(1-s)^{-(1-\nu)+\delta_1}.$$

Further, due to (3.18)

$$\max_{0\le t\le 1, s\in[0,2mh]\cup[1-2mh,1]} |\mathcal{A}(t,s)| \le ch^{m-(1-\nu)+\delta'}, \quad \delta' = \min\{\delta_0,\delta_1\} > 0,$$

which implies

$$\begin{aligned}\left(\int_0^{2mh} + \int_{1-2mh}^1\right) &|t-s|^{-\nu}|\mathcal{A}(t,s)v(s) - Q'_{h,m}(\mathcal{A}(t,s)v(s))|ds \\ &\le (1+q'_m)ch^{m-(1-\nu)+\delta'} \left(\int_0^{2mh} + \int_{1-2mh}^1\right) |t-s|^{-\nu} ds \|v\|_\infty \\ &\le c'h^{m+\delta'}\|u\|_\infty\end{aligned}$$

and together with the previous estimate,

$$\int_0^1 |t-s|^{-\nu}|\mathcal{A}(t,s)v(s) - Q'_{h,m}(\mathcal{A}(t,s)v(s))|ds \le ch^m\|u\|_{C^m_\star(0,1)}, \quad 0 \le t \le 1.$$

With the help of (3.21) and (3.22) we obtain in a similar way that

$$\int_0^1 |\mathcal{B}(t,s)v(s) - Q'_{h,m}(\mathcal{B}(t,s)v(s))|ds \le ch^m\|u\|_{C^m_\star(0,1)}, \quad 0 \le t \le 1.$$

Exploiting also Theorem 1.3 we get

$$\|\mathcal{T}v - \mathcal{T}_n v\|_\infty \le ch^m\|u\|_{C^m_\star(0,1)} \le c'h^m\|f\|_{C^m_\star(0,1)}.$$

Together with (4.8) and (4.9) this proves (4.4).

Finally, let us prove the weighted error estimate (4.5) under conditions (3.20), (3.23) on r_0 and r_1. Let v and v_n be the solutions of (3.4) and (4.2), respectively; then $w(t) = \varphi'(t)^\nu v(t)$ and $w_n(t) = \varphi'(t)^\nu v_n(t)$ are the solutions of equations

$$w(t) = \int_0^1 [|t-s|^{-\nu}\overline{\mathcal{A}}(t,s)w(s) + \overline{\mathcal{B}}(t,s)w(s)]ds + \varphi'(t)^\nu g(t), \tag{4.10}$$

$$\begin{aligned}w_n(t) = \int_0^1 &[|t-s|^{-\nu}Q'_{h,m}(\overline{\mathcal{A}}(t,s)w_n(s)) + Q'_{h,m}(\overline{\mathcal{B}}(t,s))w_n(s)]ds \\ &+ \varphi'(t)^\nu g(t), \quad 0 \le t \le 1,\end{aligned} \tag{4.11}$$

respectively, where

$$\overline{\mathcal{A}}(t,s) = \varphi'(t)^\nu \mathcal{A}(t,s)\varphi'(s)^{-\nu} = \varphi'(t)^\nu a(\varphi(t),\varphi(s))\Phi(t,s)^{-\nu}\varphi'(s)^{1-\nu},$$
$$\overline{\mathcal{B}}(t,s) = \varphi'(t)^\nu \mathcal{B}(t,s)\varphi'(s)^{-\nu} = \varphi'(t)^\nu b(\varphi(t),\varphi(s))\varphi'(s)^{1-\nu}.$$

Denote by $\overline{\mathcal{T}}$ and $\overline{\mathcal{T}}_n$ the integral operators of (4.10) and (4.11). Arguing similarly as above it is easily seen that $\overline{\mathcal{T}}_n \to \overline{\mathcal{T}}$ compactly in $C[0,1]$; by the way, this holds even under the conditions of part (i) of the theorem, in particular, (3.13) and (3.15) guarantee the continuity of $\overline{\mathcal{A}}(t,s)$ and $\overline{\mathcal{B}}(t,s)$ on $[0,1]\times\mathbb{R}$. Hence

$$\|(I-\overline{\mathcal{T}}_n)^{-1}\|_{C[0,1]\to C[0,1]} \le c, \quad n \ge n_0, \tag{4.12}$$
$$\|w - w_n\|_\infty \le c\|\overline{\mathcal{T}}w - \overline{\mathcal{T}}_n w\|_\infty = c \max_{0\le t\le 1} \varphi'(t)^\nu |(\mathcal{T}v)(t) - (\mathcal{T}_n v)(t)|.$$

Using (3.24), (3.25), (3.21), (3.22), and Theorem 1.3, we obtain

$$\max_{0\le t\le 1} \varphi'(t)^\nu |(\mathcal{T}v)(t) - (\mathcal{T}_n v)(t)| \le ch^m \|u\|_{C^m_\star(0,1)} \le c'h^m \|f\|_{C^m_\star(0,1)},$$

which completes the proof of estimate (4.5). □

Remark 4.1. Under the conditions of Theorem 4.1(i), we have

$$\|\mathcal{T}_n\|_{C[0,1]\to C[0,1]} \le c, \|\overline{\mathcal{T}}_n\|_{C[0,1]\to C[0,1]} \le c, \qquad n \ge m. \tag{4.13}$$

Remark 4.2. Under the conditions of Theorem 4.1(ii) but using slack inequalities instead of strict ones in (3.17), (3.20), a slight modification in the proof yields

$$\|v - v_n\|_\infty \le ch^m |\log h| \|f\|_{C^m_\star(0,1)}. \tag{4.14}$$

A similar remark concerns Theorem 4.1(iii).

The numerical example in section 6 does not confirm the presence of the factor $|\log h|$ in (4.14). If it really can be dropped, the proof needs new ideas.

Remark 4.3. Assume that m is odd, $f \in C^{m+1}_\star(0,1)$, $a, b \in C^{m+1}([0,1]\times(0,1))$ satisfy (1.2) for $k+l \le m+1$, and $\mathcal{N}(I-T) = \{0\}$. Then under the conditions

$$r_0, r_1 \ge 1, \quad r_0 \ge (m+1-\nu)/(1-\nu-\lambda_0), r_1 \ge (m+1-\nu)/(1-\nu-\lambda_1),$$
$$r_0 \ge (m+1-\nu)/(1-\mu_0), r_1 \ge (m+1-\nu)/(1-\mu_1)$$

the convergence order (4.4) of method (4.2) can be improved (cf. [13]):

$$\|v - v_n\|_\infty \le ch^{m+1-\nu} \|f\|_{C^{m+1}_\star(0,1)}. \tag{4.15}$$

The outlines of the proof are as follows. Due to (2.12), $\alpha^{(p)}_{k,m} = \alpha^{(p)}_{-k,m}$. Together with the symmetry property (2.2) of B_m this implies that for a function e which if odd with respect to a point $(j+\frac{1}{2})h$ also $Q'_{h,m}e$ has the same parity property, a consequence is that $\int_{jh}^{(j+1)h}(e_m - Q'_{h,m}e_m)ds = 0$ for odd m and polynomials e_m of degree $\le m$ (note that, due to (2.9) and (2.14), $e_{m-1} - Q'_{h,m}e_{m-1} = 0$ for polynomials of degree $\le m-1$). Hence for odd m,

$$\left|\int_{jh}^{(j+1)h} (w - Q'_{h,m}w)ds\right| \le ch^{m+2} \max_{(j-m+\frac{1}{2})h\le s\le(j+m-\frac{1}{2})h} |w^{(m+1)}(s)|. \tag{4.16}$$

For $w(s) = \mathcal{A}(t,s)v(s)$, $0 < s < 1$, $0 \le t \le 1$, the conditions on r_0, r_1 yield

$$|w^{(m)}(s)| \le c\|v\|_{C^m_\star(0,1)}, \quad |w^{(m+1)}(s)| \le c[s^{-1} + (1-s)^{-1}]\|v\|_{C^{m+1}_\star(0,1)},$$

and with the help of (2.16) and (4.16) we obtain

$$\varepsilon_j := \max_{jh\le s\le (j+1)h} |w(s)-(Q'_{h,m}w)(s)| \le ch^m \|v\|_{C^m_\star(0,1)}, \quad 0\le j\le n-1,$$

$$\varepsilon'_j := \left| \int_{jh}^{(j+1)h} (w-Q'_{h,m}w)ds \right| \le ch^{m+2}[(jh)^{-1}+((n-j)h)^{-1}]\|v\|_{C^{m+1}_\star(0,1)},$$
$$j=m+1,\dots,n-m-2.$$

With $i\in\mathbb{Z}$, $0\le i\le n-1$, such that $|t-(i+\frac12)h|\le\frac{h}{2}$, we get

$$\left|\int_0^1 |t-s|^{-\nu}[w(s)-(Q'_{h,m}w)(s)]ds\right| \le \left(\sum_{j=0}^{m}+\sum_{j=n-m-1}^{n-1}\right)\varepsilon_j\int_{jh}^{(j+1)h}|t-s|^{-\nu}ds$$
$$+\sum_{j=m+1}^{n-m-2}\left\{\varepsilon_j\int_{jh}^{(j+1)h}\left||t-s|^{-\nu}-\left|\left(i+\frac12-j\right)h\right|^{-\nu}\right|ds+\varepsilon'_j\left|\left(i+\frac12-j\right)h\right|^{-\nu}\right\}$$
$$\le ch^{m+1-\nu}\|v\|_{C^{m+1}_\star(0,1)}, \quad 0\le t\le 1.$$

For $w(s)=\mathcal{B}(t,s)v(s)$ we get in a straightforward manner, as in the proof of Theorem 4.1, that $|\int_0^1(w-Q'_{h,m}w)ds|\le ch^{m+1-\nu}\|v\|_{C^{m+1}_\star(0,1)}$, and now (4.15) follows.

For even m, an inequality like (4.16) is violated, and estimate (4.4) for the method (4.2) cannot be improved; for $m=4$, a numerical confirmation can be seen from Table 6.1, column $r=9$. It is possible to modify the basic method (4.2) so that (4.15) will hold for even m; we do not go into the details here.

4.2. Matrix form of the method. With the help of (4.1) we rewrite (4.2) in the form

$$v_n(t)=\sum_{j=-m+1}^{n-1}\sum_{p\in\mathbb{Z}:|p|<m_1,0<j-p+\frac{m}{2}<n}\alpha'_{p,m}\left[\mathcal{A}\left(t,\left(j-p+\frac{m}{2}\right)h\right)\beta_j(t)\right.$$
$$\left.+\mathcal{B}\left(t,\left(j-p+\frac{m}{2}\right)h\right)\beta_j^0\right]v_n\left(\left(j-p+\frac{m}{2}\right)h\right)+g(t), \quad 0\le t\le 1, \tag{4.17}$$

where

$$\beta_j(t)=\int_0^1|t-s|^{-\nu}B_m(ns-j)ds, \quad \beta_j^0=\int_0^1 B_m(ns-j)ds;$$

we took into account that $\mathcal{A}(t,s)=0$, $\mathcal{B}(t,s)=0$ for $s\le 0$ and for $s\ge 1$. The solution v_n of (4.17) is determined by its values at the points $(k+\frac{m}{2})h$ for k satisfying $0<(k+\frac{m}{2})h<1$, that is, for $k=-m_0,\dots,n-m_1$ with m_0 and m_1 defined in (2.8), (2.15). Collocating (4.17) at $(i+\frac{m}{2})h$ and denoting

$$v_{i,n}=v_n\left(\left(i+\frac{m}{2}\right)h\right), \quad g_i=g\left(\left(i+\frac{m}{2}\right)h\right),$$
$$i=-m_0,\dots,n-m_1,$$
$$a_{i,k}=\mathcal{A}\left(\left(i+\frac{m}{2}\right)h,\left(k+\frac{m}{2}\right)h\right), \quad b_{i,k}=\mathcal{B}\left(\left(i+\frac{m}{2}\right)h, \quad \left(k+\frac{m}{2}\right)h\right),$$
$$i,k=-m_0,\dots,n-m_1,$$

we arrive at the system of linear algebraic equations

$$v_{i,n} = \sum_{j=-m+1}^{n-1} \sum_{p\in\mathbb{Z}:|p|<m_1, 0<j-p+\frac{m}{2}<n} \alpha'_{p,m}(\beta_{i,j}a_{i,j-p} + \beta_j^0 b_{i,j-p})v_{j-p,n} + g_i, \tag{4.18}$$
$$i = -m_0, \dots, n-m_1,$$

where for $i = -m_0, \dots, n-m_1$, $j = -m+1, \dots, n-1$,

$$\beta_{i,j} = \int_0^1 \left|\left(i+\frac{m}{2}\right)h - s\right|^{-\nu} B_m(ns-j)ds, \quad \beta_j^0 = \int_0^1 B_m(ns-j)ds. \tag{4.19}$$

To present (4.18) in a standard form, we change the summation ordering:

$$\sum_{j=-m+1}^{n-1} \sum_{p\in\mathbb{Z}:|p|<m_1, 0<j-p+\frac{m}{2}<n} \alpha'_{p,m}(\beta_{i,j}a_{i,j-p} + \beta_j^0 b_{i,j-p})v_{j-p,n}$$
$$= \sum_{j=-m+1}^{n-1} \sum_{k=\max\{-m_0, j-m_1+1\}}^{\min\{n-m_1, j+m_1-1\}} \alpha'_{j-k,m}(\beta_{i,j}a_{i,k} + \beta_j^0 b_{i,k})v_{k,n}$$
$$= \sum_{k=-m_0}^{n-m_1} \sum_{j=k-m_1+1}^{k+m_1-1} \alpha'_{j-k,m}(\beta_{i,j}a_{i,k} + \beta_j^0 b_{i,k})v_{k,n}$$

(the summation goes over $(j,k) \in D\cap\mathbb{Z}^2$, where $D \subset \mathbb{R}^2$ is the parallelogram restricted by the lines $k = j - m_1 + 1$, $k = j + m_1 - 1$, $k = -m_0$, $k = n - m_1$; note that $m_0 + m_1 = m$). We arrive at the following form of system (4.18):

$$v_{i,n} = \sum_{k=-m_0}^{n-m_1} \tau_{i,k}v_{k,n} + g_i, \quad i = -m_0, \dots, n-m_1, \tag{4.20}$$

where

$$\tau_{i,k} = \sum_{j=k-m_1+1}^{k+m_1-1} \alpha'_{j-k,m}(\beta_{i,j}a_{i,k} + \beta_j^0 b_{i,k}), \quad i,k = -m_0, \dots, n-m_1. \tag{4.21}$$

The dimension of system (4.20) is, respectively, $n-1$ and n for even and odd m. The computation of matrix elements (4.21) from $\alpha'_{j,m}$, $\beta_{i,j}$, β_j^0, $a_{i,k}$, $b_{i,k}$ costs approximately $3mn^2$ flops.

4.3. Quasi-interpolation extension of the discrete solution. Having found the solution $v_{i,n}$, $i = -m_0, \dots, n-m_1$, of system (4.20) we can use (4.17) to compute $v_n(t)$ for any $t \in [0,1]$. A cheaper way is to construct on every $[ih, (i+1)h]$, $i = 0, \dots, n-1$, the quasi-interpolant $\widetilde{v}_n(t)$ of degree $m-1$ using the solution of (4.20) completed by values $v_{i,n} = v_{-m_0,n}$ for $i < -m_0$ and $v_{i,n} = v_{n-m_1,n}$ for $i > n-m_1$; a slightly more accurate way is to compute $v_n(0)$ and $v_n(1)$ from (4.17) and set $v_{i,n} = v_n(0)$ for $i < -m_0$ and $v_{i,n} = v_n(1)$ for $i > n-m_1$. To maintain the convergence orders (4.4) and (4.5) for $\widetilde{v}_n$, we have to strengthen conditions on f and parameters r_0 and r_1.

THEOREM 4.2. (i) *Under the assumptions of Theorem* 4.1(i), $\|v - \widetilde{v}_n\|_\infty \to 0$ *as* $n \to \infty$.

(ii) *Under the assumptions of Theorem* 1.3, $f \in C^{m,\theta_0,\theta_1}(0,1)$ *with* $\theta_0 = \max\{\lambda_0+\nu, \mu_0\}$, $\theta_1 = \max\{\lambda_1+\nu, \mu_1\}$, (3.17), (3.20), *and* $r_0 > m$, $r_1 > m$, *it holds that*

$$\|v - \widetilde{v}_n\|_\infty \le ch^m \|f\|_{C^{m,\theta_0,\theta_1}(0,1)}. \tag{4.22}$$

(iii) *Under the assumptions of Theorem* 1.3, $f \in C^{m,\theta_0,\theta_1}(0,1)$ *as in* (ii), (3.20), (3.23), *and* $r_0 > (m+\nu)/(1+\nu)$, $r_1 > (m+\nu)/(1+\nu)$, *it holds that*

$$\max_{0\le t\le 1} \varphi'(t)^\nu |v(t) - \widetilde{v}_n(t)| \le ch^m \|f\|_{C^{m,\theta_0,\theta_1}(0,1)}. \tag{4.23}$$

Proof. Accept the assumptions of (ii). Due to (3.9) and the conditions on r_0, r_1, it holds that $v^{(j)}(0) = v^{(j)}(1) = 0$, $j = 1, \dots, m$. Extending $v(t) = v(0)$ for $t < 0$, $v(t) = v(1)$ for $t > 1$, we have $v \in C^m(\mathbb{R})$. The equality $v - \widetilde{v}_n = v - Q'_{h,m}v + Q'_{h,m}(v - v_n)$ together with (2.16), (3.9), and (4.4) implies

$$\|v - \widetilde{v}_n\|_\infty \le \|v - Q'_{h,m}v\|_\infty + q'_m \|v - v_n\|_\infty$$

$$\le ch^m \|v^{(m)}\|_\infty + ch^m \|f\|_{C^m_\star(0,1)} \le ch^m \|u\|_{C^{m,\theta_0,\theta_1}(0,1)} + ch^m \|f\|_{C^m_\star(0,1)},$$

where u is the solution of (1.1), $\|u\|_{C^{m,\theta_0,\theta_1}(0,1)} \le c\|f\|_{C^{m,\theta_0,\theta_1}(0,1)}$ by Theorem 1.3. This completes the proof of (4.22). The proof of (4.23) is similar. ∎

4.4. Modification: The product interpolation method. The considerations of sections 4.1–4.3 can be easily modified for the case of the product interpolation method defined by

$$v_n(t) = \int_0^1 [|t-s|^{-\nu} Q_{h,m}(\mathcal{A}(t,s)v_n(s)) + Q_{h,m}(\mathcal{B}(t,s)v_n(s))]ds + g(t) \tag{4.24}$$

(the spline interpolation projector $Q_{h,m}$ is used instead of the quasi-interpolation operator $Q'_{h,m}$; cf. (4.2); $\mathcal{A}(t,\cdot)v$ and $\mathcal{B}(t,\cdot)v$ are still extended from $[0,1]$ to $\mathbb{R}$ with the zero value). For $0 \le s \le 1$ and a function w defined on $\mathbb{R}$ we have (see (2.5), (2.7))

$$(Q_{h,m}w)(s) = \sum_{j=-m+1}^{n-1} \sum_{k\in\mathbb{Z}} \alpha_{j-k,m} w\left(\left(k + \frac{m}{2}\right)h\right) B_m(ns - j).$$

Since in the functions $w(s) = \mathcal{A}(t,s)v(s)$ and $w(s) = \mathcal{B}(t,s)v(s)$ we have $w(s) = 0$ for $s \le 0$, and for $s \ge 1$, the infinite sum over k reduces to the sum over $k = -m_0, \dots, n-1$, and the matrix form of method (4.24) takes the form (4.20) with (cf. (4.21))

$$\tau_{i,k} = \sum_{j=-m+1}^{n-1} \alpha_{j-k,m}(\beta_{i,j}a_{i,k} + \beta^0_j b_{i,k})$$

$$= a_{i,k} \sum_{j=-m+1}^{n-1} \alpha_{j-k,m}\beta_{i,j} + b_{i,k} \sum_{j=-m+1}^{n-1} \alpha_{j-k,m}\beta^0_j, \quad i,k = -m_0, \dots, n-m_1.$$

Now the computation of the elements of the matrix costs $O(n^3)$ flops, but using the convolution structure of the sums in the last representation form of $\tau_{i,k}$, the cost can be reduced to $O(n^2 \log n)$ flops with the help of FFT. Nevertheless, this is more complicated and slightly more expensive than in the case of the product quasi-interpolation method.

Theorem 4.1 remains valid also for method (4.24); in the proof we simply exploit Lemma 2.1 instead of Lemma 2.3.

5. Some computational details.

5.1. A numerically stable evaluation of $\Phi(t,s)$. The computation of the divided difference $\Phi(t,s) = (\varphi(t)-\varphi(s))/(t-s)$ (see (3.5) and (3.2)) for small $|t-s| \neq 0$ may cause a loss of accuracy. For $r_0, r_1 \in \mathbb{N}$, the division by $t-s$ can be performed analytically. For $0 \le s < t \le 1$ we have

$$\Phi(t,s) = \frac{1}{t-s}\int_s^t \varphi'(\sigma)d\sigma = \frac{1}{c_\star(t-s)}\int_s^t \sigma^{r_0-1}(1-\sigma)^{r_1-1}d\sigma$$

$$= \frac{1}{c_\star}\int_0^1 [(t-s)\xi+s]^{r_0-1}[(1-t)+(t-s)(1-\xi)]^{r_1-1}d\xi, \tag{5.1}$$

where we undertook the change of variables $\sigma = (t-s)\xi + s$. The last integral can be computed in a fast and stable way by an exact Gauss rule since the integrand is a polynomial of degree r_0+r_1-2 in ξ. Alternatively, applying the Newton binomial formula, we obtain the expansion

$$\Phi(t,s) = (r_0+r_1-1)!\sum_{p=0}^{r_0-1}\sum_{q=0}^{r_1-1}\frac{s^{r_0-1-p}(1-t)^{r_1-1-q}(t-s)^{p+q}}{(r_0-1-p)!(r_1-1-q)!(p+q+1)!}$$

with nonnegative terms for $0 \le s \le t \le 1$. Nevertheless, the use of this stable formula is more expensive than the use of an exact Gauss rule in (5.1). For $0 \le t < s \le 1$, the symmetry $\Phi(t,s) = \Phi(s,t)$ can be used.

5.2. Computation of the quadrature coefficients. The integrals $\beta_j(t)$ in (4.17), in particular the quadrature coefficients $\beta_{i,j}$ and β_j^0 defined in (4.19), can be evaluated on the basis of the following two lemmas.

LEMMA 5.1. *For a locally integrable function $w \in L^1_{\mathrm{loc}}(\mathbb{R})$, it holds that*

$$\int_{\mathbb{R}} w(s)B_m(ns-j)ds = h(D_h^m w^{(-m)})(jh), \quad h = 1/n, \quad j \in \mathbb{Z}, \tag{5.2}$$

where $w^{(-m)}$ is an integral function of w of order m, i.e., in the sense of distributions, $(d/ds)^m w^{(-m)}(s) = w(s)$, and $D_h^m = (D_h)^m$ is defined by

$$(D_h u)(x) = h^{-1}(u(x+h)-u(x)), \quad (D_h^m u)(x) = h^{-m}\sum_{k=0}^{m}(-1)^{m-k}\binom{m}{k}u(x+kh).$$

Proof. Integrating by parts $m-1$ times and using (2.2)–(2.4) we find that

$$\int_{\mathbb{R}} w(s)B_m(ns-j)ds = (-1)^{m-1}n^{m-1}\int_{\mathbb{R}} w^{-(m-1)}(s)B_m^{(m-1)}(ns-j)ds$$

$$= (-1)^{m-1}n^{m-1}\sum_{l=0}^{m-1}(-1)^l\binom{m-1}{l}\int_{(j+l)h}^{(j+l+1)h} w^{-(m-1)}(s)ds$$

$$= (-1)^{m-1}h^{-m+1}\sum_{l=0}^{m-1}(-1)^l\binom{m-1}{l}h(D_h w^{(-m)})((j+l)h).$$

We get (5.2) since $(-1)^{m-1}h^{-m+1}\sum_{l=0}^{m-1}(-1)^l\binom{m-1}{l}u(x+lh) = (D_h^{m-1}u)(x)$. ∎

Of course, the right-hand side of (5.2) is independent of the choice of a particular integral function $w^{(-m)}$ since D_h^m nullifies the difference of any two integral functions—those differ by a polynomial of degree $\le m-1$.

LEMMA 5.2. *For $w \in L^1(0,1)$, let $w^{(-m)}(s)$ be its integral function of order m on $[0,1]$; put $w^{(k-m)}(s) = (d/ds)^k w^{(-m)}(s)$, $0 \le s \le 1$, $k = 0, \dots, m-1$. Then*

$$\int_0^1 w(s) B_m(ns-j)ds = h(D_h^m \overline{w}^{(-m)})(jh), \quad h = 1/n, \quad j \in \mathbb{Z}, \tag{5.3}$$

where

$$\overline{w}^{(-m)}(s) = \begin{cases} \sum_{k=0}^{m-1} \frac{w^{(k-m)}(0)}{k!} s^k \text{ for } s < 0, \\ w^{(-m)}(s) \text{ for } 0 \le s \le 1, \\ \sum_{k=0}^{m-1} \frac{w^{(k-m)}(1)}{k!} (s-1)^k \text{ for } s > 1. \end{cases}$$

Proof. This immediately follows from (5.2) representing

$$\int_0^1 w(s) B_m(ns-j)ds = \int_{\mathbb{R}} \overline{w}(s) B_m(ns-j)ds,$$

where $\overline{w}$ is the extension of w by the zero value outside $(0,1)$. The integral functions of $\overline{w}$ and w can be taken equal on $[0,1]$; for $s < 0$ the integral function of $\overline{w}$ must be a polynomial of degree $m-1$, and hence the Taylor polynomial of $w^{(-m)}$ with the expansion center 0; similarly for $s > 1$. ∎

To compute $\beta_j^0 = \int_0^1 B_m(ns-j)ds$, Lemma 5.2 can be applied with $\overline{w}^{(-m)}(s) = \frac{1}{m!} s_+^m$ for $-\infty < s \le 1$. According to (5.3),

$$\beta_j^0 = h \frac{1}{m!} \triangle^m \gamma_j^0, \quad j = -m+1, \dots, n-m+1, \quad \gamma_j^0 = \begin{cases} 0, & j < 0, \\ j^m, & 0 \le j \le n, \end{cases} \tag{5.4}$$

where $\triangle^m$ is the forward difference operator of order m defined for $(\gamma_j)_{j \in \mathbb{Z}}$ by

$$\Delta \gamma_j = \gamma_{j+1} - \gamma_j, \quad \triangle^m \gamma_j = \sum_{k=0}^{m} (-1)^{m-k} \binom{m}{k} \gamma_{j+k}, \quad j \in \mathbb{Z}.$$

Actually it is sufficient to compute $\beta_j^0 = h \frac{1}{m!} \triangle^m \gamma_j^0$ by (5.4) only for $j = -m+1, \dots, -1$, since $\beta_j^0 = h$ for $0 \le j \le n-m$ and by a symmetry argument $\beta_{n-m+k}^0 = \beta_{-k}^0$, $k = 1, 2, \dots$; see (2.2).

For $w_t(s) = |t-s|^{-\nu}$, $t \in (0,1)$, an integral function $w_t^{(-m)}(s)$ on $[0,1]$ is given by

$$\overline{w}^{(-m)}(s) = w_t^{(-m)}(s) = \frac{1}{(1-\nu)\dots(m-\nu)} \begin{cases} (-1)^m (t-s)^{m-\nu}, & 0 \le s \le t, \\ (s-t)^{m-\nu}, & t \le s \le 1. \end{cases}$$

Respectively,

$$\text{for } s \le 0, \quad \overline{w}^{(-m)}(s) = \sum_{k=0}^{m-1} \frac{1}{k!} \frac{(-1)^{m-k} t^{m-k-\nu}}{(1-\nu)\dots(m-k-\nu)} s^k,$$

$$\text{for } s \ge 1, \quad \overline{w}^{(-m)}(s) = \sum_{k=0}^{m-1} \frac{1}{k!} \frac{(1-t)^{m-k-\nu}}{(1-\nu)\dots(m-k-\nu)} (s-1)^k.$$

For $\beta_{i,j} = \int_0^1 |(i+\frac{m}{2})h - s|^{-\nu} B_m(ns-j)ds$, Lemma 5.2 yields the formula

$$\beta_{i,j} = h^{1-\nu} \triangle_j^m \gamma_{i,j}, \quad i = -m_0, \dots, n-m_1, \quad j = -m+1, \dots, n-1, \tag{5.5}$$

where $\triangle^m$ is still applied to $(\gamma_{i,j})$ with respect to j and

$$\gamma_{i,j} = \begin{cases} \sum_{k=0}^{m-1} \frac{(-1)^{m-k}}{k!(1-\nu)\dots(m-k-\nu)} (i+\frac{m}{2})^{m-k-\nu} j^k, & j < 0, \\ \frac{(-1)^m}{(1-\nu)\dots(m-\nu)} (i-j+\frac{m}{2})^{m-\nu}, & 0 \le j < i+\frac{m}{2}, \\ \frac{1}{(1-\nu)\dots(m-\nu)} (j-i-\frac{m}{2})^{m-\nu}, & i+\frac{m}{2} \le j \le n, \\ \sum_{k=0}^{m-1} \frac{1}{k!(1-\nu)\dots(m-k-\nu)} (n-i-\frac{m}{2})^{m-k-\nu} (j-n)^k, & j > n. \end{cases} \tag{5.6}$$

There are some symmetries which enable us to reduce the computation cost of (5.5)–(5.6): due to (2.2), for $\widetilde{\beta}_{i,j} := \int_{\mathbb{R}} |(i+\frac{m}{2})h - s|^{-\nu} B_m(ns-j)ds$ we have $\widetilde{\beta}_{i,j} = \widetilde{\beta}_{0,|i-j|}$, $i, j \in \mathbb{Z}$, and $\beta_{i,j} = \widetilde{\beta}_{i,j} = \widetilde{\beta}_{0,|i-j|}$ for $i = -m_0, \dots, n-m_1$, $j = 0, \dots, n-m$.

One must be careful using exact formulae like (5.5)–(5.6): the differences $\triangle^m \gamma_{i,j}$ may cause a loss of accuracy in a standard floating point arithmetic. For instance, for $j - i \ge m$ we have

$$\triangle^m \left(j - i - \frac{m}{2}\right)^{m-\nu} = \left(\frac{d}{dz}\right)^m z^{m-\nu}\Big|_{z=\xi} = (1-\nu)\dots(m-\nu)\xi^{-\nu},$$

where ξ is a point from the interval $(j-i-\frac{m}{2}, j-i+\frac{m}{2})$; thus $\triangle^m (j-i-\frac{m}{2})^{m-\nu}$ is approximately $(j-i)^m/[(1-\nu)\dots(m-\nu)]$ times smaller than $(j-i-\frac{m}{2})^{m-\nu}$ itself. To avoid the loss of accuracy, $\gamma_{i,j}$ and their differences should be presented with sufficiently long mantissas. A simple rule is as follows: in cases of single-precision computations, the 7 decimal digits of $\triangle^m \gamma_{i,j}$ are correct if $\gamma_{i,j}$ and their differences are computed with double precision and $n^m \le 10^7 \cdot [(1-\nu)\dots(m-\nu)]$; in cases of double-precision computations, the 15 decimal digits of $\triangle^m \gamma_{i,j}$ are correct if $\gamma_{i,j}$ and their differences are computed with quadruple precision and

$$n^m \le 10^{15} \cdot [(1-\nu)\dots(m-\nu)]. \tag{5.7}$$

Actually, condition (5.7) can be somewhat relaxed, since usually we do not need maximally high precision of $\beta_{i,j}$.

Forgetting exact formulae, $\beta_{i,j}$ can be computed in the framework of a standard arithmetic approximately using suitable quadratures, with a given accuracy; this is more laborious than the use of (5.5)–(5.6). Exact formulae for the quadrature coefficient seem to be a delicate problem also in other methods of Nyström type for weakly singular integral equations. For instance, in [9] an unstable recurrence formula is proposed and successfully used for the computations of the quadrature coefficients in a Gauss-type quadrature.

5.3. Solving the system. Due to (4.8) and (4.13), the condition number of the matrix $(\delta_{i,k} - \tau_{i,k})_{i,k=-m_0}^{n-m_1}$ of system (4.20) is bounded uniformly in n; therefore standard methods are expected to work well when solving (4.20). Gauss elimination with its $O(n^3)$ operations can be recommended for moderate n. For large n, iteration methods with stopping when the discrepancy achieves a value δn^{-r}, $\delta > 0$, $r \ge m$, are preferable. The two-grid iteration method with a fixed coarse level [2] is of complexity $O(n^2 \log n)$ flops; with a suitable enlarging strategy for the dimension of the coarse level, the complexity can be reduced to $O(n^2)$ flops (cf. [20]); here n^2 is the cost of one

matrix to vector multiplication. The complexity of GMRES is $o(n^2 \log n)$ flops due to (4.6), (4.7) (see [19]). In our numerical computations treated in section 6, rarely more than 20 GMRES iterations were needed to achieve the discrepancy level 10^{-14}–10^{-16}.

5.4. Approximate solution of (1.1). The solutions of (1.1) and (3.4) are in the relation $v(t) = u(\varphi(t))$, or $u(x) = v(\varphi^{-1}(x))$, where $\varphi^{-1} : [0,1] \to [0,1]$ is the inverse function of φ defined in (3.2)/(3.3). Having solved the system (4.20), we find that $v_{i,n} = v_n((i + \frac{m}{2})h)$ are approximations to $u(x_{i,n})$ at $x_{i,n} = \varphi((i + \frac{m}{2})h)$, $0 < (i + \frac{m}{2})h < 1$, and a local polynomial interpolation of degree $m-1$ can be used to obtain an approximation for $u(x)$ between the knots. A more precise way is to determine $t = \varphi^{-1}(x)$ and put $\tilde{u}_n(x) = \widetilde{v}_n(\varphi^{-1}(x))$; see section 4.3. Under the conditions of Theorem 4.2(ii), we have the uniform error bound

$$\max_{0\le x\le 1} |u(x) - \tilde{u}_n(x)| = \max_{0\le t\le 1} |v(t) - \widetilde{v}_n(t)| \le ch^m \|f\|_{C^{m,\theta_0,\theta_1}(0,1)};$$

under the conditions of Theorem 4.2(iii) we have the weighted error bound

$$\begin{aligned}&\max_{0\le x\le 1} x^{\nu(r_0-1)/r_0}(1-x)^{\nu(r_1-1)/r_1}|u(x) - \tilde{u}_n(x)|\\ &\le c' \max_{0\le t\le 1} \varphi'(t)^{\nu}|v(t) - \widetilde{v}_n(t)| \le ch^m \|f\|_{C^{m,\theta_0,\theta_1}(0,1)}.\end{aligned}$$

In general φ is too complicated to present a closed formula for $t = \varphi^{-1}(x)$. For given $x \in (0,1)$, $t = \varphi^{-1}(x)$ can be approximated by the Newton method,

$$t_k = t_{k-1} - (\varphi(t_{k-1}) - x)/\varphi'(t_{k-1}), \quad k = 1, 2, \dots;$$

note that according to (3.2) $\varphi'(t) = (1/c_\star)t^{r_0-1}(1-t)^{r_1-1}$. For any $x \in (0,1)$, starting from the initial guess $t_0 = (r_0 - 1)/(r_0 + r_1 - 2)$, which is the only flex point of φ for $r_0, r_1 > 1$, the iterations converge monotonically to $t = \varphi^{-1}(x)$. For x close to 0 or 1, the initial guess $t_0 = (c_\star r_0 x)^{1/r_0}$, respectively, $t_0 = 1 - (c_\star r_1(1-x))^{1/r_1}$, may be preferable.

6. Numerical example. For the testing of the algorithms we took a simple equation (1.1) with $\nu = 1/2$:

$$u(x) = \int_0^1 |x-y|^{-1/2} u(y)dy + f(x), \quad 0 \le x \le 1.$$

We put $u(x) = 1 + x^{1/2} + (1-x)^{1/2}$ to be the exact solution; it corresponds to $f(x) = 1 - \frac{\pi}{2} - 2x^{1/2} - 2(1-x)^{1/2} - x\log(1 + (1-x)^{1/2}) - (1-x)\log(1 + x^{1/2}) + \frac{1}{2}x\log x + \frac{1}{2}(1-x)\log(1-x)$. We composed system (4.20), (4.21) for $m = 3, 4, \dots, 10$, $n = 2^k$ with $k = 4, 5, \dots, 12$, and the smoothing parameters $r_0 = r_1 =: r \in \mathbb{N}$ with $2 \le r \le 2m+1$; we solved the system by GMRES. Double precision was used everywhere except in the computation of $\beta_{i,j}$ by formulae (5.5)–(5.6), where quadruple precision was involved (see section 5.2). The errors

$$\varepsilon_{m,n,r} := \max_{-m_0\le i\le n-m_1} \left| v\left(\left(i + \frac{m}{2}\right)h\right) - v_{i,n}\right|,$$

$$\overline{\varepsilon}_{m,n,r} := \max_{-m_0\le i\le n-m_1} \left[\varphi'\left(\left(i + \frac{m}{2}\right)h\right)\right]^{1/2} \left| v\left(\left(i + \frac{m}{2}\right)h\right) - v_{i,n}\right|$$

are presented in Tables 6.1 and 6.2 for $m = 4$, different n and r. The error $\varepsilon_{m,n,r}$ for $m = 6$, different r, is presented graphically on Figure 6.1 as a function of $\log_2 n$. In

TABLE 6.1
Uniform errors $\varepsilon_{4,n,r}$.

$n\backslash r$	2	4	6	7	8	9
16	9.145e-02	1.036e-02	4.228e-03	5.434e-03	8.920e-03	1.322e-02
32	3.269e-02	2.318e-03	2.994e-04	2.786e-04	3.003e-04	4.377e-04
64	1.335e-02	5.648e-04	3.382e-05	1.843e-05	1.466e-05	1.930e-05
128	5.950e-03	1.407e-04	4.005e-06	1.353e-06	8.458e-07	1.113e-06
256	2.794e-03	3.521e-05	4.842e-07	1.085e-07	5.206e-08	6.994e-08
512	1.350e-03	8.812e-06	5.936e-08	8.876e-09	3.321e-09	4.460e-09
1024	6.632e-04	2.204e-06	7.339e-09	7.372e-10	2.114e-10	2.839e-10
2048	3.285e-04	5.513e-07	9.121e-10	6.218e-11	1.341e-11	1.809e-11
4096	1.634e-04	1.379e-07	1.137e-10	5.709e-12	8.606e-13	1.163e-12

TABLE 6.2
Weighted errors $\bar{\varepsilon}_{4,n,r}$.

$n\backslash r$	2	3	4	5	6	7
16	5.422e-02	9.333e-03	2.967e-03	1.093e-03	3.582e-03	8.021e-03
32	1.393e-02	1.591e-03	1.872e-04	1.868e-05	1.356e-04	3.039e-04
64	4.055e-03	2.203e-04	1.004e-05	1.746e-06	8.275e-06	1.540e-05
128	1.283e-03	2.865e-05	5.313e-07	1.788e-07	5.478e-07	9.365e-07
256	4.268e-04	3.640e-06	2.924e-08	1.437e-08	3.566e-08	5.952e-08
512	1.460e-04	4.581e-07	1.684e-09	1.019e-09	2.287e-09	3.799e-09
1024	5.074e-05	5.744e-08	1.004e-10	6.794e-11	1.454e-10	2.414e-10
2048	1.777e-05	7.196e-09	6.101e-12	4.372e-12	9.160e-12	1.519e-11
4096	6.254e-06	9.164e-10	4.716e-13	3.281e-13	6.533e-13	9.770e-13

Figure 6.2, the error $\varepsilon_{m,n,2m}$ is presented graphically for different m. The numerical results are in good accordance with Theorem 4.1, which in the present case ($\nu = 1/2$, $\lambda_0 = \lambda_1 = 0$, $b(x,y) \equiv 0$) claims that (i) $\varepsilon_{m,n,r} \to 0$ as $n \to \infty$ for $r > 1$; (ii) $\varepsilon_{m,n,r} = O(n^{-m})$ for $r > 2m$; (iii) $\bar{\varepsilon}_{m,n,r} = O(n^{-m})$ for $r > m + \frac{1}{2}$. We can see from Tables 6.1 and 6.2 that in reality, the best results for $m = 4$ in the uniform norm are obtained for $r = 2m = 8$, whereas in the weighted norm, the best results are obtained for $r = m + 1 = 5$; for the uniform norm, the empirical convergence order is $\varepsilon_{4,n,r} = O(n^{-4})$ for $r \geq 8$ and lower for $r \leq 7$; for the weighted norm, the empirical convergence order is $\bar{\varepsilon}_{4,n,r} = O(n^{-4})$ for $r \geq 5$ and lower for $m \leq 4$; on the other hand, the logarithmic factor of estimate (4.14) for $r = 2m = 8$ cannot be observed in the present example.

For $r > 2m$ recommended by Theorem 4.1(ii), and also for $r = 2m$, the computations appeared to be rather stable in a wide scale of n; see Figure 6.2. A side remark is that the numerical results presented on Figure 6.2 encourage the use of high order splines in computations: in the present example, already beginning from $n = 32$, the use of tenth order splines gave more precise numerical results than the use of splines of lower order.

For $m = 4$ and fixed $r \geq 2$, we can see from Table 6.1 a monotone decrease of errors $\varepsilon_{4,n,r}$ in n. For large m but r significantly smaller than $2m$, the error $\varepsilon_{m,n,r}$ starts to grow beginning from some (relatively large) $n = n_{m,r}$; see Figure 6.1 ($m = 6$). We run into difficulties trying to explain this kind of instability for small r ($r = 2$ on Figure 6.1 for $m = 6$). Condition (5.7) guaranteeing a safe use of the quadruple precision in formulae (5.5)–(5.6) is violated for $m = 6$, $n \geq 739$. Nevertheless, when the precision of the computations in (5.5)–(5.6) was increased, no essential change in the behavior of $\varepsilon_{6,n,r}$ was observed compared with Figure 6.1. Hence the instability for large m, n and small r cannot be explained by the violation of (5.7) only.

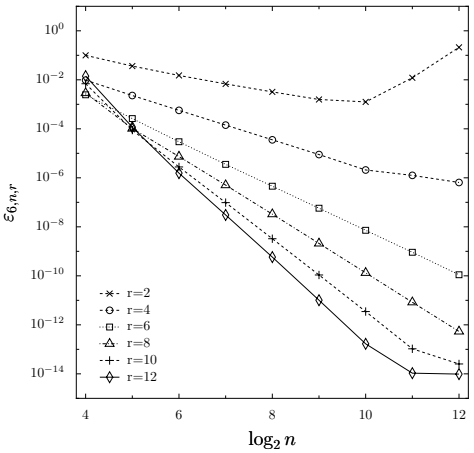


FIG. 6.1. *Errors* $\varepsilon_{6,n,r}$.

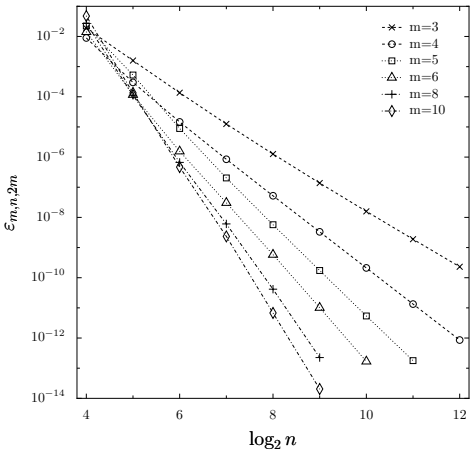


FIG. 6.2. *Errors* $\varepsilon_{m,n,2m}$.

All programs for given computations were written using the SciPy package (Scientific Tools for Python, http://www.scipy.org). For carrying out higher precision floating point arithmetic, a Python interface to a Fortran95 subroutine was used, which uses the `quadruple precision` datatype for internal calculations. As an alternative to Fortran95, we used also Python Decimal datatype for higher precision calculations. With the Decimal type there is a performance penalty compared to Fortran95 calculations. But the flexibility of specifying an arbitrary length of the mantissa in floating point calculations proved to be very helpful in finding out the sensitivity of our algorithms to precision changes.

The source code and numerical results in more detail of provided tests can be found at http://www.ut.ee/∼eero/WSIE/Fredholm/.

Further prospects. We solved the equation $\lambda u(x) = \int_0^1 |x-y|^{-\nu} u(y)du + f(x)$ numerically also for some other values of $\nu \in (0,1)$. For small and moderate ν, the method behaves as expected by Theorem 4.1. For ν close to 1, the uniform accuracy becomes lower due to great values of the smoothing parameter $r \sim m/(1-\nu)$; see Theorem 4.1(ii).

Returning to the general problem (1.1), it is possible to modify our approach towards smaller values of the smoothing parameters r_0 and r_1. In the arguments of section 4 we exploited the fact that the smoothing change of variables suppresses the coefficient function by $|t-s|^{-\nu}$ but since $v(0) \neq 0$, $v(1) \neq 0$ in general, we were not able to use the fact that also the derivatives of the solution v of (3.4) are suppressed if $f \in C^{m,\theta_0,\theta_1}(0,1)$ with $\theta_0 < 1$, $\theta_1 < 1$. The situation changes if $u(0) = u(1) = 0$ for the solution of (1.1), then also $v(0) = v(1) = 0$. The idea of the modification is to rewrite (1.1) with respect to unknowns $u(0)$, $u(1)$, and $U(x) = u(x) - u(0) - (u(1) - u(0))\psi(x)$, where ψ is a function of type (3.2), achieving $V^{(j)}(0) = V^{(j)}(1) = 0$, $j = 0, \ldots, m$, for $V(t) = U(\varphi(t))$. In the algorithm and its justification, some new problems arrive. We cannot go into details here because of the length of the paper but we will return to the question elsewhere.

Acknowledgment. The authors are grateful to the referees for the useful discussion and valuable comments that enabled them to improve some details of the paper.

REFERENCES

[1] P. M. Anselone, *Collectively Compact Operator Approximation Theory*, Prentice-Hall, Englewood Cliffs, NJ, 1971.

[2] K. E. Atkinson, *The Numerical Solution of Integral Equations of the Second Kind*, Cambridge University Press, Cambridge, UK, 1997.

[3] H. Brunner, *Collocation Methods for Volterra Integral and Related Functional Equations*, Cambridge University Press, Cambridge, UK, 2004.

[4] C. de Boor, *A Practical Guide to Splines*, Springer, New York, 2001.

[5] W. Hackbusch, *Integral Equations*, Birkhäuser, Basel, 1995.

[6] N. P. Korneychuk, *Splines in the Approximation Theory*, Nauka, Moscow, 1984 (in Russian).

[7] R. Kress, *Linear Integral Equations*, Springer, Berlin, 1989.

[8] E. Leetma and G. Vainikko, *Quasi-interpolation by splines on the uniform knot sets*, Math. Model. Anal., 12 (2007), pp. 107–120.

[9] G. Monegato and L. Scuderi, *High order methods for weakly singular integral equations with nonsmooth input functions*, Math. Comp., 67 (1998), pp. 1493–1515.

[10] A. Pedas and G. Vainikko, *Smoothing transformation and piecewise polynomial collocation for weakly singular integral equations*, Computing, 73 (2004), pp. 271–293.

[11] A. Pedas and G. Vainikko, *Integral equations with diagonal and boundary singularities of the kernel*, Z. Anal. Anwend., 25 (2006), pp. 487–516.

[12] A. Pedas and G. Vainikko, *On the regularity of solutions to integral equations with nonsmooth kernels*, J. Comput. Appl. Math., to appear.

[13] C. Schneider, *Product integration for weakly singular integral equations*, Math. Comp., 36 (1981), pp. 207–213.

[14] I. J. Schoenberg, *Cardinal Spline Interpolation*, SIAM, Philadelphia, 1973.

[15] L. L. Schumaker, *Spline Functions: Basic Theory*, Krieger, Malabar, FL, 1993.

[16] S. B. Stechkin and Yu. N. Subbotin, *Splines in Numerical Mathematics*, Nauka, Moscow, 1976 (in Russian).

[17] E. Vainikko and G. Vainikko, *A product quasi-interpolation method for weakly singular Volterra integral equations*, in Numerical Analysis and Applied Mathematics, Vol. 936, T. E. Simos, ed., AIP Conference Proceedings, American Institute of Physics, Melville, NY, 2007, pp. 570–573.

[18] G. Vainikko, *Funktionalanalysis der Diskretisierungsmethoden*, Teubner, Leipzig, 1976.

[19] G. Vainikko, *GMRES and discrete approximation of operators*, Proc. Estonian Acad Sci. Phys. Math., 53 (2004), pp. 124–131.

[20] G. Vainikko, *Fast solvers of integral equations of the second kind: Quadrature methods*, J. Integral Equations Appl., 17 (2005), pp. 91–120.

[21] G. Vainikko, *On the best approximation of function classes from values on a uniform grid in the real line*, WSEAS Trans. Math., 4/6 (2006), pp. 523–528.

[22] G. Vainikko, *Error estimates for the cardinal spline interpolation*, Z. Anal. Anwend., to appear.

SIAM J. NUMER. ANAL.
Vol. 46, No. 4, pp. 1821–1841
© 2008 Society for Industrial and Applied Mathematics

NUMERICAL SOLUTIONS OF NEUTRAL STOCHASTIC FUNCTIONAL DIFFERENTIAL EQUATIONS*

FUKE WU† AND XUERONG MAO‡

Abstract. This paper examines the numerical solutions of neutral stochastic functional differential equations (NSFDEs) $d[x(t) - u(x_t)] = f(x_t)dt + g(x_t)dw(t)$, $t \geq 0$. The key contribution is to establish the strong mean square convergence theory of the Euler–Maruyama approximate solution under the local Lipschitz condition, the linear growth condition, and contractive mapping. These conditions are generally imposed to guarantee the existence and uniqueness of the true solution, so the numerical results given here are obtained under quite general conditions. Although the way of analysis borrows from [X. Mao, *LMS J. Comput. Math.*, 6 (2003), pp. 141–161], to cope with $u(x_t)$, several new techniques have been developed.

Key words. neutral stochastic functional differential equations, strong convergence, Euler–Maruyama method, local Lipschitz condition

AMS subject classifications. 65C20, 65C30

DOI. 10.1137/070697021

1. Introduction.

The theory of stochastic functional differential equations (SFDEs) has received a great deal of attention in recent decades. More recently researchers have given special interest to the study of equations in which the variable delay argument occurs in the derivative of the state variable, or so-called neutral stochastic functional differential equations (NSFDEs). Many well-known theorems in SFDEs are successfully extended to NSFDEs; see, for example, [10, 13, 14, 15, 16, 18].

As is the case with deterministic neutral functional differential equations and SFDEs, most NSFDEs cannot be solved explicitly, so numerical methods become one of the most powerful techniques. A number of papers study the numerical analysis of deterministic neutral functional differential equations; see, for example, [1, 3, 5, 7, 11] and the references therein. Numerical solutions of SFDEs have also been studied extensively by many authors. Here we mention some of them, for example, [2, 4, 6, 9, 12, 16, 17].

However, little is as yet known about numerical solutions for NSFDEs although it may be seen as a combination of deterministic neutral functional differential equations and SFDEs. This paper will fill that gap.

We study the Euler–Maruyama numerical solutions of the NSFDE

$$d[x(t) - u(x_t)] = f(x_t)dt + g(x_t)dw(t), \quad t \geq 0, \tag{1.1}$$

with initial data $x_0 = \xi \in L^p_{\mathcal{F}_0}([-\tau, 0], \mathbb{R}^n)$. Here,

$$f : C([-\tau, 0]; \mathbb{R}^n) \to \mathbb{R}^n, \quad g : C([-\tau, 0]; \mathbb{R}^n) \to \mathbb{R}^{n\times m}, \quad u : C([-\tau, 0]; \mathbb{R}^n) \to \mathbb{R}^n,$$

*Received by the editors July 12, 2007; accepted for publication (in revised form) January 11, 2008; published electronically April 16, 2008. This work was supported by the Chinese Scholarship Council.

http://www.siam.org/journals/sinum/46-4/69702.html

†Department of Mathematics, Huazhong University of Science and Technology, Wuhan, Hubei 430 074, People's Republic of China (wufuke@mail.hust.edu.cn).

‡Department of Statistics and Modelling Science, University of Strathclyde, Glasgow G1 1XH, UK (xuerong@stams.strath.ac.uk).

$x(t) \in \mathbb{R}^n$ for each t,

$$x_t = \{x(t+\theta) : -\tau \le \theta \le 0\} \in C([-\tau, 0]; \mathbb{R}^n),$$

and $w(t)$ is an m-dimensional Brownian motion. The initial data ξ is an $\mathcal{F}_0$-measurable $C([-\tau, 0]; \mathbb{R}^n)$-valued random variable such that $\mathbb{E}\|\xi\|^p < \infty$ for some $p \ge 2$. Our main aim is to extend the method developed by [16] and [17] to NSFDEs and study strong convergence for the Euler–Maruyama numerical approximations in the case where f and g satisfy both the local Lipschitz condition and the linear growth condition and where u is a contractive mapping. These three conditions are standard for the existence and uniqueness of the true solutions.

Although the method of analysis borrows from [16], the existence of the neutral term $u(x_t)$ essentially changes the problem, and several new techniques have been developed to cope with the difficulties which have risen from the neutral term.

In section 2, we introduce some necessary assumptions and auxiliary results, define the Euler–Maruyama method for NSFDEs, and state our main result that the approximate solutions strongly converge to the exact solution. The proof of the result is rather technical so we present several lemmas in section 3 and then complete the proof in section 4. In the final section, under the global Lipschitz condition, we reveal the order of convergence of approximate solutions.

2. The Euler–Maruyama method for NSFDEs. Throughout this paper, unless otherwise specified, we use the following notation. Let $|\cdot|$ be the Euclidean norm in $\mathbb{R}^n$. If A is a vector or matrix, its transpose is denoted by A^T. If A is a matrix, its trace norm is denoted by $|A| = \sqrt{\text{trace}(A^T A)}$. Let $\mathbb{R}_+ = [0, \infty)$, and let $\tau > 0$. Denote by $C([-\tau, 0], \mathbb{R}^n)$ the family of continuous functions from $[-\tau, 0]$ to $\mathbb{R}^n$ with the norm $\|\varphi\| = \sup_{-\tau \le \theta \le 0} |\varphi(\theta)|$.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a complete probability space with a filtration $\{\mathcal{F}_t\}_{t \ge 0}$ satisfying the usual conditions; that is, it is right continuous and increasing while $\mathcal{F}_0$ contains all P-null sets. Let $w(t) = (w_1(t), \dots, w_m(t))^T$ be an m-dimensional Brownian motion defined on the probability space. Let $p > 0$ and let $L^p_{\mathcal{F}_0}([-\tau, 0]; \mathbb{R}^n)$ be the family of $\mathcal{F}_0$-measurable $C([-\tau, 0]; \mathbb{R}^n)$-valued random variables ξ such that $\mathbb{E}\|\xi\|^p < \infty$. If $x(t)$ is an $\mathbb{R}^n$-valued stochastic process on $t \in [-\tau, \infty)$, we let $x_t = \{x(t+\theta) : -\tau \le \theta \le 0\}$ for $t \ge 0$.

Let $f : C([-\tau, 0]; \mathbb{R}^n) \to \mathbb{R}^n$, $g : C([-\tau, 0]; \mathbb{R}^n) \to \mathbb{R}^{n \times m}$, and $u : C([-\tau, 0]; \mathbb{R}^n) \to \mathbb{R}^n$. In this paper we impose the following hypotheses.

Assumption 2.1 (local Lipschitz condition). For each integer $j \ge 1$, there exists a positive constant C_j such that

$$|f(\varphi) - f(\psi)|^2 \vee |g(\varphi) - g(\psi)|^2 \le C_j \|\varphi - \psi\|^2 \tag{2.1}$$

for $\varphi, \psi \in C([-\tau, 0]; \mathbb{R}^n)$ with $\|\varphi\| \vee \|\psi\| \le j$.

Assumption 2.2 (linear growth condition). There is a constant $K > 0$ such that

$$|f(\varphi)|^2 \vee |g(\varphi)|^2 \le K(1 + \|\varphi\|^2) \tag{2.2}$$

for $\varphi \in C([-\tau, 0]; \mathbb{R}^n)$.

Assumption 2.3 (contractive mapping). There exists a constant $\kappa \in (0, 1)$ such that for all $\varphi, \psi \in C([-\tau, 0]; \mathbb{R}^n)$,

$$|u(\varphi) - u(\psi)| \le \kappa \|\varphi - \psi\| \tag{2.3}$$

for $\varphi, \psi \in C([-\tau, 0]; \mathbb{R}^n)$ and $u(0) = 0$.

Consider the n-dimensional NSFDE

$$d[x(t) - u(x_t)] = f(x_t)dt + g(x_t)dw(t), \quad t \geq 0, \tag{2.4}$$

with initial data $x_0 = \xi \in L^p_{\mathcal{F}_0}([-\tau, 0]; \mathbb{R}^n)$.

We know that Assumptions 2.1, 2.2, and 2.3 are standard conditions for the existence and uniqueness of the solution of (2.4) except that we impose $u(0) = 0$. Actually, $u(0) = 0$ is also standard for the boundedness of the solution's moments (for example, see [14]).

We impose the following condition on the initial data.

Assumption 2.4. $\xi \in L^p_{\mathcal{F}_0}([-\tau, 0]; \mathbb{R}^n)$ for some $p \geq 2$, and there exists a nondecreasing function $\alpha(\cdot)$ such that

$$\mathbb{E}\Big(\sup_{-\tau \leq s \leq t \leq 0} |\xi(t) - \xi(s)|^2 \Big) \leq \alpha(t - s) \tag{2.5}$$

with the property $\alpha(s) \to 0$ as $s \to 0$.

From Mao [14, p. 211], we may therefore state the following theorem.

THEOREM 2.1. *Let $p \geq 2$. If Assumptions* 2.2–2.4 *are satisfied, then*

$$\mathbb{E}\Big(\sup_{-\tau \leq t \leq T} |x(t)|^p \Big) \leq (1 + \bar{C}\mathbb{E}\|\xi\|^p) e^{\widetilde{C}T} \tag{2.6}$$

for any $T > 0$, where

$$\widetilde{C} = \frac{2p(1+\kappa)^{p-2}}{(1-p)^p}[\sqrt{2K}(1+\kappa) + K(33p-1)] \quad \textit{and} \quad \bar{C} = \frac{1}{1-\kappa} + \frac{2(1+\kappa)^p}{(1-\kappa)^p}.$$

From now on, we will consider (2.4) on the finite time interval $[0, T]$, since our aim is to discuss the finite time convergence of the Euler–Maruyama method. Without loss of generality, we may assume that T/τ is a rational number; otherwise we may replace T by a larger number. Let the step size $\triangle \in (0, 1)$ be a fraction of τ and T, namely, $\triangle = \tau/N = T/M$ for some integers $N > \tau$ and $M > T$. The explicit discrete Euler–Maruyama approximate solution $\bar{y}(k\triangle)$, $k \geq -N$, is defined as follows:

$$\begin{cases} \bar{y}(k\triangle) = \xi(k\triangle), & -N \leq k \leq 0, \\ \bar{y}((k+1)\triangle) = \bar{y}(k\triangle) + u(\bar{y}_{k\triangle}) - u(\bar{y}_{(k-1)\triangle}) + f(\bar{y}_{k\triangle})\triangle + g(\bar{y}_{k\triangle})\triangle w_k, & \\ \qquad\qquad 0 \leq k \leq M-1, & \end{cases} \tag{2.7}$$

where $\triangle w_k = w((k+1)\triangle) - w(k\triangle)$ and $\bar{y}_{k\triangle} = \{\bar{y}_{k\triangle}(\theta) : -\tau \leq \theta \leq 0\}$ is a $C([-\tau, 0]; \mathbb{R}^n)$-valued random variable defined by

$$\begin{aligned} \bar{y}_{k\triangle}(\theta) &= \bar{y}((k+i)\triangle) + \frac{\theta - i\triangle}{\triangle}[\bar{y}((k+i+1)\triangle) - \bar{y}((k+i)\triangle)] \\ &\qquad \text{for } i\triangle \leq \theta \leq (i+1)\triangle,\ i = -N, -(N-1), \ldots, -1, \end{aligned} \tag{2.8}$$

where in order for $\bar{y}_{-\triangle}$ to be well defined, we set $\bar{y}(-(N+1)\triangle) = \xi(-N\triangle)$.

It is easy to see from (2.8) that $\bar{y}_{k\triangle}(\cdot)$ is the linear interpolation of $\bar{y}((k-N)\triangle), \ldots, \bar{y}(k\triangle)$ and (2.8) can be rewritten as

$$\bar{y}_{k\triangle}(\theta) = \frac{(i+1)\triangle - \theta}{\triangle}\bar{y}((k+i)\triangle) + \frac{\theta - i\triangle}{\triangle}\bar{y}((k+i+1)\triangle), \tag{2.9}$$

which yields

$$|\bar{y}_{k\triangle}(\theta)| \le \frac{(i+1)\triangle-\theta}{\triangle}|\bar{y}((k+i)\triangle)| + \frac{\theta-i\triangle}{\triangle}|\bar{y}((k+i+1)\triangle)|$$
$$\le |\bar{y}((k+i)\triangle)| \vee |\bar{y}((k+i+1)\triangle)|$$

because $(i+1)\triangle-\theta$ and $\theta-i\triangle \in [0,\triangle]$. We therefore have

$$\|\bar{y}_{k\triangle}\| = \max_{-N\le i\le 0}|\bar{y}((k+i)\triangle)| \quad \forall k=-1,0,1,2,\ldots,M-1. \tag{2.10}$$

It is obvious that $\|\bar{y}_{-\triangle}\| \le \|\bar{y}_0\|$.

In our analysis it will be more convenient to use continuous-time approximations. We hence introduce the $C([-\tau,0];\mathbb{R}^n)$-valued step process

$$\bar{y}_t = \sum_{k=0}^{M-2}\bar{y}_{k\triangle}1_{[k\triangle,(k+1)\triangle)}(t) + \bar{y}_{(M-1)\triangle}1_{[(M-1)\triangle,M\triangle]}(t), \tag{2.11}$$

and we define the continuous Euler–Maruyama approximate solution as follows: let $y(t)=\xi(t)$ for $-\tau\le t\le 0$, while for $t\in[k\triangle,(k+1)\triangle]$, $k=0,1,\ldots,M-1$,

$$y(t) = \xi(0) + u\Big(\bar{y}_{(k-1)\triangle} + \frac{t-k\triangle}{\triangle}(\bar{y}_{k\triangle}-\bar{y}_{(k-1)\triangle})\Big) - u(\bar{y}_{-\triangle})$$
$$+\int_0^t f(\bar{y}_s)ds + \int_0^t g(\bar{y}_s)dw(s). \tag{2.12}$$

Clearly, (2.12) can also be written as

$$y(t) = \bar{y}(k\triangle) + u\Big(\bar{y}_{(k-1)\triangle} + \frac{t-k\triangle}{\triangle}(\bar{y}_{k\triangle}-\bar{y}_{(k-1)\triangle})\Big) - u(\bar{y}_{(k-1)\triangle})$$
$$+\int_{k\triangle}^t f(\bar{y}_s)ds + \int_{k\triangle}^t g(\bar{y}_s)dw(s). \tag{2.13}$$

In particular, this shows that $y(k\triangle)=\bar{y}(k\triangle)$, that is, the discrete and continuous Euler–Maruyama approximate solutions coincide at the gridpoints. We know that $y(t)$ is not computable because it requires knowledge of the entire Brownian path, not just its $\triangle$-increments. However, $y(k\triangle)=\bar{y}(k\triangle)$, so the error bound for $y(t)$ will automatically imply the error bound for $\bar{y}(k\triangle)$. It is then obvious that

$$\|\bar{y}_{k\triangle}\| \le \|y_{k\triangle}\| \quad \forall k=0,1,2,\ldots,M-1. \tag{2.14}$$

Moreover, for any $t\in[0,T]$,

$$\begin{aligned}\sup_{0\le t\le T}\|\bar{y}_t\| &= \sup_{0\le k\le M-1}\|\bar{y}_{k\triangle}\|\\ &\le \sup_{0\le k\le M-1}\|y_{k\triangle}\|\\ &= \sup_{0\le k\le M-1}\sup_{-\tau\le\theta\le 0}|y(k\triangle+\theta)|\\ &\le \sup_{0\le t\le T}\sup_{-\tau\le\theta\le 0}|y(t+\theta)|\\ &\le \sup_{-\tau\le s\le T}|y(s)|,\end{aligned} \tag{2.15}$$

and letting $[t/\triangle]$ be the integer part of $t/\triangle$, then

$$\|\bar{y}_t\| = \|\bar{y}_{[t/\triangle]\triangle}\| \le \|y_{[t/\triangle]\triangle}\| \le \sup_{-\tau\le s\le t}|y(s)|. \tag{2.16}$$

These properties will be used frequently in what follows, without further explanation.

The primary aim of this paper is to establish the following strong mean square convergence theorem for the Euler–Maruyama approximations.

THEOREM 2.2. *Under Assumptions* 2.1–2.4,

$$\lim_{\triangle\to 0}\mathbb{E}\Big(\sup_{0\le t\le T}|x(t)-y(t)|^2\Big) = 0. \tag{2.17}$$

The proof of this theorem is very technical, so we present some lemmas in the next section, and then complete the proof in section 4.

3. Lemmas.

LEMMA 3.1. *Under Assumptions* 2.2–2.4, *for any* $p \ge 2$, *there exists a constant* $H(p)$ *such that*

$$\mathbb{E}\Big(\sup_{-\tau\le t\le T}|y(t)|^p\Big) \le H(p), \tag{3.1}$$

where $H(p)$ *is independent of* $\triangle$.

Proof. For $t \in [k\triangle,(k+1)\triangle]$, $k = 0,1,2,\dots,M-1$, set $\tilde{y}(t) := y(t) - u(\bar{y}_{(k-1)\triangle} + (t-k\triangle)(\bar{y}_{k\triangle} - \bar{y}_{(k-1)\triangle})/\triangle)$ and

$$h(t) := \mathbb{E}\left(\sup_{-\tau\le s\le t}|y(s)|^p\right);$$

$$\tilde{h}(t) := \mathbb{E}\left(\sup_{0\le s\le t}|\tilde{y}(s)|^p\right).$$

Recall the inequality that for $p \ge 1$ and any $\varepsilon > 0$,

$$|x+y|^p \le (1+\varepsilon)^{p-1}(|x|^p + \varepsilon^{1-p}|y|^p).$$

This, together with Assumption 2.3, yields

$$\begin{aligned}|y(t)|^p &\le (1+\varepsilon)^{p-1}(|\tilde{y}(t)|^p + \varepsilon^{1-p}|u(\bar{y}_{(k-1)\triangle} + (t-k\triangle)(\bar{y}_{k\triangle} - \bar{y}_{(k-1)\triangle})/\triangle)|^p)\\ &\le (1+\varepsilon)^{p-1}\Big(|\tilde{y}(t)|^p + \varepsilon^{1-p}\kappa^p\Big\|\bar{y}_{(k-1)\triangle} + \frac{t-k\triangle}{\triangle}(\bar{y}_{k\triangle} - \bar{y}_{(k-1)\triangle})\Big\|^p\Big).\end{aligned}$$

By (2.9) and $\|\bar{y}_{-\triangle}\| \le \|\bar{y}_0\|$, noting $k = 0,1,2,\dots,M-1$,

$$\begin{aligned}&\Big\|\bar{y}_{(k-1)\triangle} + \frac{t-k\triangle}{\triangle}(\bar{y}_{k\triangle} - \bar{y}_{(k-1)\triangle})\Big\|^p\\ &\le \Big|\frac{(k+1)\triangle - t}{\triangle}\|\bar{y}_{(k-1)\triangle}\| + \frac{t-k\triangle}{\triangle}\|\bar{y}_{k\triangle}\|\Big|^p\\ &\le \Big[\frac{(k+1)\triangle - t}{\triangle}\Big(\sup_{-\tau\le s\le t}|y(s)|\Big) + \frac{t-k\triangle}{\triangle}\Big(\sup_{-\tau\le s\le t}|y(s)|\Big)\Big]^p\\ &\le \sup_{-\tau\le s\le t}|y(s)|^p.\end{aligned}$$

Consequently,

$$|y(t)|^p \le (1+\varepsilon)^{p-1}\Big[|\tilde{y}(t)|^p + \varepsilon^{1-p}\kappa^p\Big(\sup_{-\tau\le s\le t}|y(s)|^p\Big)\Big].$$

Choose $\varepsilon = \kappa/(1-\kappa)$. Then

$$|y(t)|^p \le (1-\kappa)^{1-p}|\tilde{y}(t)|^p + \kappa\Big(\sup_{-\tau\le s\le t}|y(s)|^p\Big).$$

Hence,

$$\begin{aligned} h(t) &\le \mathbb{E}\|\xi\|^p + \mathbb{E}\Big(\sup_{0\le s\le t}|y(s)|^p\Big) \\ &\le \mathbb{E}\|\xi\|^p + \kappa h(t) + (1-\kappa)^{1-p}\tilde{h}(t), \end{aligned}$$

which implies

$$h(t) \le \frac{\mathbb{E}\|\xi\|^p}{1-\kappa} + \frac{\tilde{h}(t)}{(1-\kappa)^p}. \tag{3.2}$$

Since

$$\tilde{y}(t) = \tilde{y}(0) + \int_0^t f(\bar{y}_s)ds + \int_0^t g(\bar{y}_s)dw(s)$$

with $\tilde{y}(0) = \bar{y}(0) - u(\bar{y}_{-\triangle})$, by the Hölder inequality, we have

$$|\tilde{y}(t)|^p \le 3^{p-1}\left[|\tilde{y}(0)|^p + t^{p-1}\int_0^t |f(\bar{y}_s)|^p ds + \left|\int_0^t g(\bar{y}_s)dw(s)\right|^p\right].$$

Hence, for any $t_1 \in [0,T]$

$$\begin{aligned} \tilde{h}(t_1) \le 3^{p-1}\Bigg[&\mathbb{E}|\tilde{y}(0)|^p + T^{p-1}\mathbb{E}\int_0^{t_1}|f(\bar{y}_s)|^p ds \\ &+ \mathbb{E}\left(\sup_{0\le t\le t_1}\left|\int_0^t g(\bar{y}_s)dw(s)\right|^p\right)\Bigg]. \end{aligned} \tag{3.3}$$

By Assumption 2.3 and the fact $\|\bar{y}_{-\triangle}\| \le \|\bar{y}_0\|$, we compute that

$$\begin{aligned} \mathbb{E}|\tilde{y}(0)|^p &= \mathbb{E}|\bar{y}(0) - u(\bar{y}_{-\triangle})|^p \\ &\le \mathbb{E}(|\bar{y}(0)| + \kappa\|\bar{y}_{-\triangle}\|)^p \\ &\le \mathbb{E}(|\bar{y}(0)| + \kappa\|\bar{y}_0\|)^p \\ &\le \mathbb{E}(|\xi(0)| + \kappa\|\xi\|)^p \\ &\le (1+\kappa)^p\mathbb{E}\|\xi\|^p. \end{aligned} \tag{3.4}$$

Assumption 2.2 and the Hölder inequality give

$$\begin{aligned} \mathbb{E}\int_0^{t_1}|f(\bar{y}_s)|^p ds &\le \mathbb{E}\int_0^{t_1} K^{\frac{p}{2}}(1+\|\bar{y}_s\|^2)^{\frac{p}{2}}ds \\ &\le K^{\frac{p}{2}}2^{\frac{p-2}{2}}\mathbb{E}\int_0^{t_1}(1+\|\bar{y}_s\|^p)ds \\ &\le K^{\frac{p}{2}}2^{\frac{p-2}{2}}\left[T + \int_0^{t_1}\mathbb{E}\left(\sup_{-\tau\le t\le s}|y(t)|^p ds\right)\right]. \end{aligned} \tag{3.5}$$

Applying the Burkholder–Davis–Gundy inequality, the Hölder inequality, and Assumption 2.2 yields

$$\begin{aligned}\mathbb{E}\left(\sup_{0\le t\le t_1}\left|\int_0^t g(\bar{y}_s)dw(s)\right|^p\right) &\le C_p\mathbb{E}\left(\int_0^{t_1}|g(\bar{y}_s)|^2ds\right)^{\frac{p}{2}}\\ &\le C_pT^{\frac{p-2}{2}}\mathbb{E}\int_0^{t_1}|g(\bar{y}_s)|^pds\\ &\le C_pT^{\frac{p-2}{2}}\mathbb{E}\int_0^{t_1}K^{\frac{p}{2}}\left(1+\|\bar{y}_s\|^2\right)^{\frac{p}{2}}ds\\ &\le C_pT^{\frac{p-2}{2}}K^{\frac{p}{2}}2^{\frac{p-2}{2}}\mathbb{E}\int_0^{t_1}\left(1+\|\bar{y}_s\|^p\right)ds\\ &\le C_pT^{\frac{p-2}{2}}K^{\frac{p}{2}}2^{\frac{p-2}{2}}\left[T+\int_0^{t_1}\mathbb{E}\Big(\sup_{-\tau\le t\le s}|y(t)|^p\Big)ds\right], \qquad (3.6)\end{aligned}$$

where C_p is a constant dependent only on p. Substituting (3.4), (3.5), and (3.6) into (3.3) gives

$$\begin{aligned}\tilde{h}(t_1) \le\ & 3^{p-1}\Big[(1+\kappa)^p\mathbb{E}\|\xi\|^p+K^{\frac{p}{2}}2^{\frac{p-2}{2}}T^p+C_p(2T)^{\frac{p-2}{2}}K^{\frac{p}{2}}T\Big]\\ &+3^{p-1}\Big[K^{\frac{p}{2}}2^{\frac{p-2}{2}}T^{p-1}+C_p(2T)^{\frac{p-2}{2}}K^{\frac{p}{2}}\Big]\int_0^{t_1}\mathbb{E}\Big(\sup_{-\tau\le t\le s}|y(t)|^p\Big)ds\\ =:\ & C_1+C_2\int_0^{t_1}h(s)ds.\end{aligned}$$

Hence from (3.2), we have

$$\begin{aligned}h(t_1)&\le\frac{\mathbb{E}\|\xi\|^p}{1-\kappa}+\frac{1}{(1-\kappa)^p}\left[C_1+C_2\int_0^{t_1}h(s)ds\right]\\ &\le\frac{\mathbb{E}\|\xi\|^p}{1-\kappa}+\frac{C_1}{(1-\kappa)^p}+\frac{C_2}{(1-\kappa)^p}\int_0^{t_1}h(s)ds.\end{aligned}$$

By the Gronwall inequality we find that

$$h(T)\le\left[\frac{\mathbb{E}\|\xi\|^p}{1-\kappa}+\frac{C_1}{(1-\kappa)^p}\right]e^{\frac{C_2T}{(1-\kappa)^p}}.$$

From the expressions of C_1 and C_2, we know that they are positive constants dependent only on ξ, κ, K, p, and T but independent of $\triangle$. The required assertion must hold. □

LEMMA 3.2. *If Assumptions* 2.2–2.4 *hold, then for any integer* $l>1$,

$$\mathbb{E}\Big(\sup_{0\le k\le M-1}\|\bar{y}_{k\triangle}-\bar{y}_{(k-1)\triangle}\|^2\Big)\le c_1\alpha(\triangle)+\bar{c}_1(l)\triangle^{\frac{l-1}{l}}=:\gamma(\triangle), \qquad (3.7)$$

where $c_1=1/(1-\kappa)$, *and* $\bar{c}_1(l)$ *is a constant dependent on* l *but independent of* $\triangle$.

Proof. For $\theta \in [i\triangle, (i+1)\triangle]$, where $i = -N, -(N+1), \ldots, -1$, from (2.9),

$$\begin{aligned}|\bar{y}_{k\triangle} - \bar{y}_{(k-1)\triangle}| &\le \frac{(i+1)\triangle - \theta}{\triangle}|\bar{y}((k+i)\triangle) - \bar{y}((k-1+i)\triangle)| \\ &\quad + \frac{\theta - i\triangle}{\triangle}|\bar{y}((k+i+1)\triangle) - \bar{y}((k+i)\triangle)| \\ &\le |\bar{y}((k+i)\triangle) - \bar{y}((k-1+i)\triangle)| \vee |\bar{y}((k+i+1)\triangle) - \bar{y}((k+i)\triangle)|,\end{aligned}$$

so

$$\|\bar{y}_{k\triangle} - \bar{y}_{(k-1)\triangle}\| \le \sup_{-N\le i\le 0}|\bar{y}((k+i)\triangle) - \bar{y}((k-1+i)\triangle)|.$$

We therefore have

$$\begin{aligned}&\mathbb{E}\Big(\sup_{0\le k\le M-1}\|\bar{y}_{k\triangle} - \bar{y}_{(k-1)\triangle}\|^2\Big) \\ &\le \mathbb{E}\Big[\sup_{0\le k\le M-1}\Big(\sup_{-N\le i\le 0}|\bar{y}((k+i)\triangle) - \bar{y}((k-1+i)\triangle)|^2\Big)\Big] \\ &\le \mathbb{E}\Big(\sup_{-N\le k\le M-1}|\bar{y}(k\triangle) - \bar{y}((k-1)\triangle)|^2\Big). \qquad (3.8)\end{aligned}$$

When $-N \le k \le 0$, by Assumption 2.4 and $\bar{y}(-(N+1)\triangle) = \xi(-N\triangle)$,

$$\begin{aligned}\mathbb{E}\Big(\sup_{-N\le k\le 0}|\bar{y}(k\triangle) - \bar{y}((k-1)\triangle)|^2\Big) &\le \mathbb{E}\Big(\sup_{-N\le k\le 0}|\xi(k\triangle) - \xi((k-1)\triangle)|^2\Big) \\ &\le \alpha(\triangle). \qquad (3.9)\end{aligned}$$

When $1 \le k \le M-1$, from (2.7), we have

$$\bar{y}(k\triangle) - \bar{y}((k-1)\triangle) = u(\bar{y}_{(k-1)\triangle}) - u(\bar{y}_{(k-2)\triangle}) + f(\bar{y}_{(k-1)\triangle})\triangle + g(\bar{y}_{(k-1)\triangle})\triangle w_k.$$

Recall the elementary inequality, for any $x, y > 0$ and $\varepsilon \in (0,1)$,

$$(x+y)^2 \le \frac{x^2}{\varepsilon} + \frac{y^2}{1-\varepsilon}.$$

Then we have

$$\begin{aligned}&|\bar{y}(k\triangle) - \bar{y}((k-1)\triangle)|^2 \\ &\le \frac{1}{\varepsilon}|u(\bar{y}_{(k-1)\triangle}) - u(\bar{y}_{(k-2)\triangle})|^2 + \frac{1}{1-\varepsilon}|f(\bar{y}_{(k-1)\triangle})\triangle + g(\bar{y}_{(k-1)\triangle})\triangle w_k|^2 \\ &\le \frac{\kappa^2}{\varepsilon}\|\bar{y}_{(k-1)\triangle} - \bar{y}_{(k-2)\triangle}\|^2 + \frac{2}{1-\varepsilon}|f(\bar{y}_{(k-1)\triangle})|^2\triangle^2 + \frac{2}{1-\varepsilon}|g(\bar{y}_{(k-1)\triangle})\triangle w_k|^2,\end{aligned}$$

and consequently,

$$\begin{aligned}&\mathbb{E}\Big(\sup_{1\le k\le M-1}|\bar{y}(k\triangle) - \bar{y}((k-1)\triangle)|^2\Big) \\ &\le \frac{\kappa^2}{\varepsilon}\mathbb{E}\Big(\sup_{1\le k\le M-1}\|\bar{y}_{(k-1)\triangle} - \bar{y}_{(k-2)\triangle}\|^2\Big) \\ &\quad + \frac{2\triangle^2}{1-\varepsilon}\mathbb{E}\Big(\sup_{1\le k\le M-1}|f(\bar{y}_{(k-1)\triangle})|^2\Big) \\ &\quad + \frac{2}{1-\varepsilon}\mathbb{E}\Big(\sup_{1\le k\le M-1}|g(\bar{y}_{(k-1)\triangle})\triangle w_k|^2\Big). \qquad (3.10)\end{aligned}$$

We deal with these three terms separately. By (3.9),

$$\begin{aligned}
&\mathbb{E}\Big(\sup_{1\le k\le M-1}\|\bar{y}_{(k-1)\triangle}-\bar{y}_{(k-2)\triangle}\|^2\Big)\\
&\le \mathbb{E}\Big(\sup_{1\le k\le M-1}\sup_{-N\le i\le 0}|\bar{y}((k+i-1)\triangle)-\bar{y}((k-2+i)\triangle)|^2\Big)\\
&\le \mathbb{E}\Big(\sup_{-N\le k\le M-1}|\bar{y}(k\triangle)-\bar{y}((k-1)\triangle)|^2\Big)\\
&\le \mathbb{E}\Big(\sup_{-N\le k\le 0}|\bar{y}(k\triangle)-\bar{y}((k-1)\triangle)|^2\Big)\\
&\quad+\mathbb{E}\Big(\sup_{1\le k\le M-1}|\bar{y}(k\triangle)-\bar{y}((k-1)\triangle)|^2\Big)\\
&\le \alpha(\triangle)+\mathbb{E}\Big(\sup_{1\le k\le M-1}|\bar{y}(k\triangle)-\bar{y}((k-1)\triangle)|^2\Big). \qquad (3.11)
\end{aligned}$$

Noting that $\mathbb{E}[\sup_{-\tau\le t\le T}|\bar{y}(t)|^2]\le\mathbb{E}[\sup_{-\tau\le t\le T}|y(t)|^2]\le H(2)$ (where $H(p)$ has been defined in Lemma 3.1), by Assumption 2.2 and (2.10),

$$\begin{aligned}
\mathbb{E}\Big(\sup_{1\le k\le M-1}|f(\bar{y}_{(k-1)\triangle})|^2\Big)&\le\mathbb{E}\Big(\sup_{1\le k\le M-1}K(1+\|\bar{y}_{(k-1)\triangle}\|^2)\Big)\\
&\le K+K\mathbb{E}\Big(\sup_{1\le k\le M-1}\sup_{-N\le i\le 0}|\bar{y}((k-1+i)\triangle)|^2\Big)\\
&\le K+K\mathbb{E}\Big(\sup_{-N\le k\le M-1}|\bar{y}(k\triangle)|^2\Big)\\
&\le K+K\mathbb{E}\Big(\sup_{-\tau\le t\le T}|\bar{y}(t)|^2\Big)\\
&\le K(1+H(2)). \qquad (3.12)
\end{aligned}$$

By the Hölder inequality, for any integer $l>1$,

$$\begin{aligned}
&\mathbb{E}\Big(\sup_{1\le k\le M-1}|g(\bar{y}_{k\triangle})\triangle w_k|^2\Big)\\
&\le\mathbb{E}\Big(\sup_{1\le k\le M-1}|g(\bar{y}_{(k-1)\triangle})|^2\sup_{1\le k\le M-1}|\triangle w_k|^2\Big)\\
&\le\Big[\mathbb{E}\Big(\sup_{1\le k\le M-1}|g(\bar{y}_{(k-1)\triangle})|^{\frac{2l}{l-1}}\Big)\Big]^{\frac{l-1}{l}}\Big[\mathbb{E}\Big(\sup_{0\le k\le M-1}|\triangle w_k|^{2l}\Big)\Big]^{\frac{1}{l}}\\
&\le\Big[\mathbb{E}\Big(\sup_{0\le k\le M-1}(K(1+\|\bar{y}_{k\triangle}\|^2))^{\frac{l}{l-1}}\Big)\Big]^{\frac{l-1}{l}}\Big[\mathbb{E}\Big(\sum_{k=0}^{M-1}|\triangle w_k|^{2l}\Big)\Big]^{\frac{1}{l}}\\
&\le\Big[K^{\frac{l}{1-l}}\mathbb{E}\Big(1+\Big(\sup_{0\le k\le M-1}\|\bar{y}_{k\triangle}\|^2\Big)\Big)^{\frac{l}{l-1}}\Big]^{\frac{l-1}{l}}\Big[\Big(\sum_{k=0}^{M-1}\mathbb{E}|\triangle w_k|^{2l}\Big)\Big]^{\frac{1}{l}}\\
&\le\Big[2^{\frac{1}{l-1}}K^{\frac{l}{1-l}}\Big(1+\mathbb{E}\Big(\sup_{0\le k\le M-1}\|\bar{y}_{k\triangle}\|^{\frac{2l}{l-1}}\Big)\Big)\Big]^{\frac{l-1}{l}}\Big[\Big(\sum_{k=0}^{M-1}(2l-1)!!\triangle^l\Big)\Big]^{\frac{1}{l}}\\
&\le\Big[2^{\frac{1}{l-1}}K^{\frac{l}{1-l}}\Big(1+H\Big(\frac{2l}{l-1}\Big)\Big)\Big]^{\frac{l-1}{l}}[(2l-1)!!T\triangle^{l-1}]^{\frac{1}{l}}\\
&\le D(l)\triangle^{\frac{l-1}{l}}, \qquad (3.13)
\end{aligned}$$

where $(2l-1)!! = 1 \cdot 3 \cdots (2l-1)$, $D(l)$ is a constant dependent on l.

Substituting (3.11), (3.12), and (3.13) into (3.10), choosing $\varepsilon = \kappa$, and noting that $\triangle \in (0,1)$, we have

$$\begin{aligned} &\mathbb{E}\Big(\sup_{1\le k\le M-1}|\bar{y}(k\triangle)-\bar{y}((k-1)\triangle)|^2\Big) \\ &\le \frac{\kappa}{1-\kappa}\alpha(\triangle)+\frac{2K(1+H(2))+2D(l)}{(1-\kappa)^2}\triangle^{\frac{l-1}{l}}. \end{aligned} \tag{3.14}$$

Combining (3.9) with (3.14), from (3.8) we have

$$\begin{aligned} &\mathbb{E}\Big(\sup_{0\le k\le M-1}\|\bar{y}_{k\triangle}-\bar{y}_{(k-1)\triangle}\|^2\Big) \\ &\le \mathbb{E}\Big(\sup_{-N\le k\le M-1}|\bar{y}(k\triangle)-\bar{y}((k-1)\triangle)|^2\Big) \\ &\le \mathbb{E}\Big(\sup_{-N\le k\le 0}|\bar{y}(k\triangle)-\bar{y}((k-1)\triangle)|^2\Big) \\ &\quad + \mathbb{E}\Big(\sup_{1\le k\le M-1}|\bar{y}(k\triangle)-\bar{y}((k-1)\triangle)|^2\Big) \\ &\le \frac{1}{1-\kappa}\alpha(\triangle)+\frac{2K(1+H(2))+2D(l)}{(1-\kappa)^2}\triangle^{\frac{l-1}{l}}, \end{aligned} \tag{3.15}$$

as required. □

LEMMA 3.3. *If Assumptions* 2.2–2.4 *hold,*

$$\mathbb{E}\Big(\sup_{0\le s\le T}\|y_s-\bar{y}_s\|^2\Big)\le c_2\alpha(2\triangle)+\bar{c}_2(l)\triangle^{\frac{l-1}{l}} =: \beta(\triangle), \tag{3.16}$$

where c_2 is a constant independent of l and $\triangle$, and $\bar{c}_2(l)$ is a constant dependent on l but independent of $\triangle$.

Proof. Fix any $s \in [0,T]$ and $\theta \in [-\tau, 0]$. Let $k_s \in \{0,1,2,\ldots,M-1\}$, $k_\theta \in \{-N, -N+1, \ldots, -1\}$, and $k_{s\theta} \in \{-N, -N+1, \ldots, M-1\}$ be the integers for which $s \in [k_s\triangle, (k_s+1)\triangle]$, $\theta \in [k_\theta\triangle, (k_\theta+1)\triangle]$, and $s+\theta \in [k_{s\theta}\triangle, (k_{s\theta}+1)\triangle]$, respectively. Clearly,

$$0 \le s+\theta-(k_s+k_\theta) \le 2\triangle, \tag{3.17}$$

$$k_{s\theta}-(k_s+k_\theta) \in \{0,1,2\}. \tag{3.18}$$

From (2.8),

$$\begin{aligned} \bar{y}_s &= \bar{y}_{k_s\triangle}(\theta) \\ &= \bar{y}((k_s+k_\theta)\triangle)+\frac{\theta-k_\theta\triangle}{\triangle}(\bar{y}((k_s+k_\theta+1)\triangle)-\bar{y}(k_s+k_\theta)\triangle), \end{aligned}$$

which yields

$$\begin{aligned} |y_s-\bar{y}_s| &= |y(s+\theta)-\bar{y}_{k_s\triangle}(\theta)| \\ &\le |y(s+\theta)-\bar{y}((k_s+k_\theta)\triangle)|+\frac{\theta-k_\theta\triangle}{\triangle}|\bar{y}((k_s+k_\theta+1)\triangle)-\bar{y}((k_s+k_\theta)\triangle)| \\ &\le |y(s+\theta)-\bar{y}((k_s+k_\theta)\triangle)|+|\bar{y}((k_s+k_\theta+1)\triangle)-\bar{y}((k_s+k_\theta)\triangle)|, \end{aligned}$$

so by (3.8) and Lemma 3.2, noting that $\bar{y}(M\triangle) = \bar{y}((M-1)\triangle)$ from (2.11),

$$\begin{aligned}
&\mathbb{E}\Big(\sup_{0\le s\le T}\|y_s-\bar{y}_s\|^2\Big)\\
&\le 2\mathbb{E}\Big[\sup_{0\le s\le T}\Big(\sup_{-\tau\le\theta\le 0}|y(s+\theta)-\bar{y}((k_s+k_\theta)\triangle)|^2\Big)\Big]\\
&\quad+2\mathbb{E}\Big[\sup_{0\le k_s\le M-1}\Big(\sup_{-N\le k_\theta\le 0}|\bar{y}((k_s+k_\theta+1)\triangle)-\bar{y}((k_s+k_\theta)\triangle)|^2\Big)\Big]\\
&\le 2\mathbb{E}\Big(\sup_{-\tau\le\theta\le 0,0\le s\le T}|y(s+\theta)-\bar{y}((k_s+k_\theta)\triangle)|^2\Big)+2\gamma(\triangle). \qquad (3.19)
\end{aligned}$$

Therefore, it is key to compute $\mathbb{E}\big(\sup_{-\tau\le\theta\le 0,0\le s\le T}|y(s+\theta)-\bar{y}((k_s+k_\theta)\triangle)|^2\big)$. We discuss the following four possible cases.

Case 1: $k_s+k_\theta\ge 0$. We again divide this case into three possible subcases according to $k_{s\theta}-(k_s+k_\theta)\in\{0,1,2\}$.

Subcase 1: $k_{s\theta}-(k_s+k_\theta)=0$. From (2.13)

$$\begin{aligned}
&y(s+\theta)-\bar{y}((k_s+k_\theta)\triangle)\\
&=u\Big(\bar{y}_{(k_{s\theta}-1)\triangle}+\frac{s+\theta-k_{s\theta}\triangle}{\triangle}(\bar{y}_{k_{s\theta}\triangle}-\bar{y}_{(k_{s\theta}-1)\triangle})\Big)-u(\bar{y}_{(k_{s\theta}-1)\triangle})\\
&\quad+\int_{k_{s\theta}\triangle}^{s+\theta}f(\bar{y}_r)dr+\int_{k_{s\theta}\triangle}^{s+\theta}g(\bar{y}_r)dw(r),
\end{aligned}$$

which yields

$$\begin{aligned}
&\mathbb{E}\Big(\sup_{-\tau\le\theta\le 0,0\le s\le T,k_s+k_\theta\ge 0}|y(s+\theta)-\bar{y}((k_s+k_\theta)\triangle)|^2\Big)\\
&\le 3\mathbb{E}\Big(\sup_{-\tau\le\theta\le 0,0\le s\le T,k_{s\theta}\ge 0}\Big|u\Big(\bar{y}_{(k_{s\theta}-1)\triangle}+\frac{s+\theta-k_{s\theta}\triangle}{\triangle}(\bar{y}_{k_{s\theta}\triangle}-\bar{y}_{(k_{s\theta}-1)\triangle})\Big)\\
&\quad-u(\bar{y}_{(k_{s\theta}-1)\triangle})\Big)\Big|^2+3\mathbb{E}\Big(\sup_{-\tau\le\theta\le 0,0\le s\le T,k_{s\theta}\ge 0}\Big|\int_{k_{s\theta}\triangle}^{s+\theta}f(\bar{y}_r)dr\Big|^2\Big)\\
&\quad+3\mathbb{E}\Big(\sup_{-\tau\le\theta\le 0,0\le s\le T,k_{s\theta}\ge 0}\Big|\int_{k_{s\theta}\triangle}^{s+\theta}g(\bar{y}_r)dw(r)\Big|^2\Big). \qquad (3.20)
\end{aligned}$$

From Assumption 2.3 and Lemma 3.2, we have

$$\begin{aligned}
&\mathbb{E}\Big(\sup_{-\tau\le\theta\le 0,0\le s\le T,k_{s\theta}\ge 0}\Big|u\Big(\bar{y}_{(k_{s\theta}-1)\triangle}\\
&\quad+\frac{s+\theta-k_{s\theta}\triangle}{\triangle}(\bar{y}_{k_{s\theta}\triangle}-\bar{y}_{(k_{s\theta}-1)\triangle})\Big)-u(\bar{y}_{(k_{s\theta}-1)\triangle})\Big)\Big|^2\\
&\le\kappa^2\mathbb{E}\Big(\sup_{-\tau\le\theta\le 0,0\le s\le T,k_{s\theta}\ge 0}\Big\|\frac{s+\theta-k_{s\theta}\triangle}{\triangle}(\bar{y}_{k_{s\theta}\triangle}-\bar{y}_{(k_{s\theta}-1)\triangle})\Big\|^2\Big)\\
&\le\kappa^2\mathbb{E}\Big(\sup_{-\tau\le\theta\le 0,0\le s\le T,k_{s\theta}\ge 0}\|(\bar{y}_{k_{s\theta}\triangle}-\bar{y}_{(k_{s\theta}-1)\triangle})\|^2\Big)\\
&\le\kappa^2\mathbb{E}\Big(\sup_{0\le k_{s\theta}\le M-1}\|(\bar{y}_{k_{s\theta}\triangle}-\bar{y}_{(k_{s\theta}-1)\triangle})\|^2\Big)\\
&\le\kappa^2\gamma(\triangle). \qquad (3.21)
\end{aligned}$$

By the Hölder inequality, Assumption 2.2, and Lemma 3.1,

$$\begin{aligned}
&\mathbb{E}\Big(\sup_{-\tau\le\theta\le 0,0\le s\le T,k_{s\theta}\ge 0}\Big|\int_{k_{s\theta}\triangle}^{s+\theta} f(\bar{y}_r)dr\Big|^2\Big)\\
&\le \triangle\mathbb{E}\Big(\sup_{-\tau\le\theta\le 0,0\le s\le T,k_{s\theta}\ge 0}\int_{k_{s\theta}\triangle}^{s+\theta}|f(\bar{y}_r)|^2dr\Big)\\
&\le K\triangle\mathbb{E}\Big(\sup_{-\tau\le\theta\le 0,0\le s\le T,k_{s\theta}\ge 0}\int_{k_{s\theta}\triangle}^{s+\theta}(1+\|\bar{y}_r\|^2)dr\Big)\\
&\le K\triangle\mathbb{E}\Big[\int_0^T\Big(1+\sup_{0\le r\le T}\|\bar{y}_r\|^2\Big)dr\Big]\\
&\le K\triangle\int_0^T\Big[1+\mathbb{E}\Big(\sup_{-\tau\le t\le T}|y(t)|^2\Big)\Big]dr\\
&\le K\triangle\int_0^T[1+H(2)]dr\\
&\le KT[1+H(2)]\triangle. \qquad (3.22)
\end{aligned}$$

Setting $v=s+\theta$ and $k_v=k_{s\theta}$ and applying the Hölder inequality yield

$$\begin{aligned}
&\mathbb{E}\Big(\sup_{-\tau\le\theta\le 0,0\le s\le T,k_{s\theta}\ge 0}\Big|\int_{k_{s\theta}\triangle}^{s+\theta} g(\bar{y}_r)dw(r)\Big|^2\Big)\\
&\quad=\mathbb{E}\Big(\sup_{0\le v\le T,0\le k_v\le M-1}|g(\bar{y}_{k_v\triangle})(w(v)-w(k_v\triangle))|^2\Big)\\
&\quad\le\Big[\mathbb{E}\Big(\sup_{0\le v\le T,0\le k_v\le M-1}|g(\bar{y}_{k_v\triangle})|^{\frac{2l}{l-1}}\Big)\Big]^{\frac{l-1}{l}}\\
&\qquad\Big[\mathbb{E}\Big(\sup_{0\le v\le T,0\le k_v\le M-1}|w(v)-w(k_v\triangle)|^{2l}\Big)\Big]^{\frac{1}{l}}.
\end{aligned}$$

The Doob martingale inequality gives

$$\begin{aligned}
&\mathbb{E}\Big(\sup_{0\le v\le T,0\le k_v\le M-1}|w(v)-w(k_v\triangle)|^{2l}\Big)\\
&=\mathbb{E}\Big(\sup_{0\le k_v\le M-1}\Big(\sup_{k_v\triangle\le v\le(k_v+1)\triangle}|w(v)-w(k_v\triangle)|^{2l}\Big)\Big)\\
&\le\mathbb{E}\Big(\sum_{k_v=0}^{M-1}\Big(\sup_{k_v\triangle\le v\le(k_v+1)\triangle}|w(v)-w(k_v\triangle)|^{2l}\Big)\Big)\\
&=\sum_{k_v=0}^{M-1}\mathbb{E}\Big(\sup_{k_v\triangle\le v\le(k_v+1)\triangle}|w(v)-w(k_v\triangle)|^{2l}\Big)\\
&\le\Big(\frac{2l}{2l-1}\Big)^{2l}\sum_{k_v=0}^{M-1}\mathbb{E}|w((k_v+1)\triangle)-w(k_v\triangle)|^{2l}.
\end{aligned}$$

By (3.13), we therefore have

$$\begin{aligned}
&\mathbb{E}\Big(\sup_{-\tau\le\theta\le 0,0\le s\le T,k_{s\theta}\ge 0}\Big|\int_{k_{s\theta}\triangle}^{s+\theta} g(\bar{y}_r)dw(r)\Big|^2\Big)\\
&\le \Big(\frac{2l}{2l-1}\Big)^2\Big[\mathbb{E}\Big(\sup_{0\le k_v\le M-1}|g(\bar{y}_{k_v\triangle})|^{\frac{2l}{l-1}}\Big)\Big]^{\frac{l-1}{l}}\\
&\times\Big[\sum_{k_v=0}^{M-1}\mathbb{E}|w((k_v+1)\triangle)-w(k_v\triangle)|^{2l}\Big]^{\frac{1}{l}}\\
&\le \Big(\frac{2l}{2l-1}\Big)^2 D(l)\triangle^{\frac{l-1}{l}}. \qquad (3.23)
\end{aligned}$$

Substituting (3.21), (3.22), and (3.23) into (3.20) and noting $\triangle\in(0,1)$ give

$$\begin{aligned}
&\mathbb{E}\Big(\sup_{-\tau\le\theta\le 0,0\le s\le T,k_s+k_\theta\ge 0}|y(s+\theta)-\bar{y}((k_s+k_\theta)\triangle)|^2\Big)\\
&\le 3\kappa^2\gamma(\triangle)+3\Big(KT(1+H(2))+\Big(\frac{2l}{2l-1}\Big)^2 D(l)\Big)\triangle^{\frac{l-1}{l}}\\
&=: 3\kappa^2\gamma(\triangle)+c_1(l)\triangle^{\frac{l-1}{l}}. \qquad (3.24)
\end{aligned}$$

Subcase 2: $k_{s\theta}-(k_s+k_\theta)=1$. From (2.13) and (2.7),

$$\begin{aligned}
&y(s+\theta)-\bar{y}((k_s+k_\theta)\triangle)\\
&=y(k_{s\theta}\triangle)-\bar{y}((k_s+k_\theta)\triangle)+y(s+\theta)-\bar{y}(k_{s\theta}\triangle)\\
&\le u(\bar{y}_{(k_s+k_\theta)\triangle})-u(\bar{y}_{(k_s+k_\theta-1)\triangle})+f(\bar{y}_{(k_s+k_\theta)\triangle})\triangle+g(\bar{y}_{(k_s+k_\theta)\triangle})\triangle w_{k_s+k_\theta}\\
&\quad+y(s+\theta)-\bar{y}(k_{s\theta}\triangle),
\end{aligned}$$

so we have

$$\begin{aligned}
&\mathbb{E}\Big(\sup_{-\tau\le\theta\le 0,0\le s\le T,k_s+k_\theta\ge 0}|y(s+\theta)-\bar{y}((k_s+k_\theta)\triangle)|^2\Big)\\
&\le 4\Big\{\mathbb{E}\Big(\sup_{-\tau\le\theta\le 0,0\le s\le T,k_s+k_\theta\ge 0}|u(\bar{y}_{(k_s+k_\theta)\triangle})-u(\bar{y}_{(k_s+k_\theta-1)\triangle})|^2\Big)\\
&\quad+\mathbb{E}\Big(\sup_{-\tau\le\theta\le 0,0\le s\le T,k_s+k_\theta\ge 0}|f(\bar{y}_{(k_s+k_\theta)\triangle})\triangle|^2\Big)\\
&\quad+\mathbb{E}\Big(\sup_{-\tau\le\theta\le 0,0\le s\le T,k_s+k_\theta\ge 0}|g(\bar{y}_{(k_s+k_\theta)\triangle})\triangle w_{k_s+k_\theta}|^2\Big)\\
&\quad+\mathbb{E}\Big(\sup_{-\tau\le\theta\le 0,0\le s\le T,k_s+k_\theta\ge 0}|y(s+\theta)-\bar{y}(k_{s\theta}\triangle)|^2\Big)\Big\}. \qquad (3.25)
\end{aligned}$$

Since

$$\begin{aligned}
&\mathbb{E}\Big(\sup_{-\tau\le\theta\le 0,0\le s\le T,k_s+k_\theta\ge 0}|u(\bar{y}_{(k_s+k_\theta)\triangle})-u(\bar{y}_{(k_s+k_\theta-1)\triangle})|^2\Big)\\
&\le\kappa^2\mathbb{E}\Big(\sup_{-\tau\le\theta\le 0,0\le s\le T,k_s+k_\theta\ge 0}\|\bar{y}_{(k_s+k_\theta)\triangle}-\bar{y}_{(k_s+k_\theta-1)\triangle}\|^2\Big)\\
&\le\kappa^2\gamma(\triangle), \qquad (3.26)
\end{aligned}$$

from (3.12), (3.13), and Subcase 1, noting that $\triangle \in (0,1)$, we have

$$\begin{aligned}&\mathbb{E}\Big(\sup_{-\tau\le\theta\le 0,0\le s\le T,k_s+k_\theta\ge 0}|y(s+\theta)-\bar{y}((k_s+k_\theta)\triangle)|^2\Big)\\ &\le 4\Big\{\kappa^2\gamma(\triangle)+K[1+H(2)]\triangle^2+D(l)\triangle^{\frac{l-1}{l}}+3\kappa^2\gamma(\triangle)+c_1(l)\triangle^{\frac{l-1}{l}}\Big\}\\ &=: 16\kappa^2\gamma(\triangle)+c_2(l)\triangle^{\frac{l-1}{l}}. \qquad (3.27)\end{aligned}$$

Subcase 3: $k_{s\theta}-(k_s+k_\theta)=2$. From (2.13) and (2.7), we have

$$\begin{aligned}&y(s+\theta)-\bar{y}((k_s+k_\theta)\triangle)\\ &=y(s+\theta)-y((k_{s\theta}-1)\triangle)+y((k_{s\theta}-1)\triangle)-\bar{y}((k_s+k_\theta)\triangle),\end{aligned}$$

so from Subcase 2, we have

$$\begin{aligned}&\mathbb{E}\Big(\sup_{-\tau\le\theta\le 0,0\le s\le T,k_s+k_\theta\ge 0}|y(s+\theta)-\bar{y}((k_s+k_\theta)\triangle)|^2\Big)\\ &\le 2\mathbb{E}\Big(\sup_{-\tau\le\theta\le 0,0\le s\le T,k_s+k_\theta\ge 0}|y(s+\theta)-\bar{y}((k_{s\theta}-1)\triangle)|^2\Big)\\ &\quad+2\mathbb{E}\Big(\sup_{-\tau\le\theta\le 0,0\le s\le T,k_s+k_\theta\ge 0}|y((k_{s\theta}-1)\triangle)-\bar{y}((k_s+k_\theta)\triangle)|^2\Big)\\ &\le 2[16\kappa^2\gamma(\triangle)+c_2(l)\triangle^{\frac{l-1}{l}}]+2[\kappa^2\gamma(\triangle)+K[1+H(2)]\triangle^2+D(l)\triangle^{\frac{l-1}{l}}]\\ &=: 34\kappa^2\gamma(\triangle)+c_3(l)\triangle^{\frac{l-1}{l}}. \qquad (3.28)\end{aligned}$$

From these three subcases, we have

$$\begin{aligned}&\mathbb{E}\Big(\sup_{-\tau\le\theta\le 0,0\le s\le T,k_s+k_\theta\ge 0}|y(s+\theta)-\bar{y}((k_s+k_\theta)\triangle)|^2\Big)\\ &\le 53\kappa^2\gamma(\triangle)+[c_1(l)+c_2(l)+c_3(l)]\triangle^{\frac{l-1}{l}}. \qquad (3.29)\end{aligned}$$

Case 2: $k_s+k_\theta=-1$ *and* $0\le s+\theta\le\triangle$. In this case, applying Assumption 2.4 and Case 1, we have

$$\begin{aligned}&\mathbb{E}\Big(\sup_{-\tau\le\theta\le 0,0\le s\le T}|y(s+\theta)-\bar{y}((k_s+k_\theta)\triangle)|^2\Big)\\ &\le 2\mathbb{E}\Big(\sup_{-\tau\le\theta\le 0,0\le s\le T}|y(s+\theta)-\bar{y}(0)|^2\Big)+2E|y(0)-y(-\triangle)|^2\\ &\le 106\kappa^2\gamma(\triangle)+2[c_1(l)+c_2(l)+c_3(l)]\triangle^{\frac{l-1}{l}}+2\alpha(\triangle). \qquad (3.30)\end{aligned}$$

Case 3: $k_s+k_\theta=-1$ *and* $-\triangle\le s+\theta\le 0$. In this case, using Assumption 2.4

$$\mathbb{E}\Big(\sup_{-\tau\le\theta\le 0,0\le s\le T}|y(s+\theta)-\bar{y}((k_s+k_\theta)\triangle)|^2\Big)\le\alpha(\triangle). \qquad (3.31)$$

Case 4: $k_s+k_\theta\le -2$. In this case, $s+\theta\le 0$, so using Assumption 2.4

$$\mathbb{E}\Big(\sup_{-\tau\le\theta\le 0,0\le s\le T}|y(s+\theta)-\bar{y}((k_s+k_\theta)\triangle)|^2\Big)\le\alpha(2\triangle). \qquad (3.32)$$

Substituting these four cases into (3.19) and noting the expression of $\gamma(\triangle)$, there exist c_2 and $\bar{c}_2(l)$ such that

$$\mathbb{E}\Big(\sup_{0\le s\le T}\|y_s-\bar{y}_s\|^2\Big)\le c_2\alpha(2\triangle)+\bar{c}_2(l)\triangle^{\frac{l-1}{l}}, \tag{3.33}$$

namely, the required assertion holds. □

4. Proof of Theorem 2.2. From Lemma 3.1 and Theorem 2.1, there exists a positive constant $\widetilde{H}$ such that

$$\mathbb{E}\Big(\sup_{-\tau\le t\le T}|x(t)|^p\Big)\vee\mathbb{E}\Big(\sup_{-\tau\le t\le T}|y(t)|^p\Big)\le\widetilde{H}. \tag{4.1}$$

Let j be a sufficient large integer. Define the stopping times

$$u_j:=\inf\{t\ge 0:\|x_t\|\ge j\},\quad v_j:=\inf\{t\ge 0:\|y_t\|\ge j\},\quad \rho_j:=u_j\wedge v_j,$$

where we set $\inf\emptyset=\infty$ as usual. Let

$$e(t):=x(t)-y(t).$$

Obviously,

$$\begin{aligned}\mathbb{E}\Big(\sup_{0\le t\le T}|e(t)|^2\Big)&=\mathbb{E}\Big(\sup_{0\le t\le T}|e(t)|^2 1_{\{u_j>T,\ v_j>T\}}\Big)\\&\quad+\mathbb{E}\Big(\sup_{0\le t\le T}|e(t)|^2 1_{\{u_j\le T\text{ or }v_j\le T\}}\Big).\end{aligned}$$

Recall the following elementary inequality:

$$a^\gamma b^{1-\gamma}\le\gamma a+(1-\gamma)b\quad\forall a,b>0,\ \gamma\in[0,1]. \tag{4.2}$$

We thus have, for any $\delta>0$,

$$\begin{aligned}&\mathbb{E}\Big(\sup_{0\le t\le T}|e(t)|^2 1_{\{u_j\le T\text{ or }v_j\le T\}}\Big)\\&\quad\le\mathbb{E}\Big[\Big(\delta\sup_{\{0\le t\le T\}}|e(t)|^p\Big)^{\frac{2}{p}}\Big(\delta^{-\frac{2}{p-2}}1_{\{u_j\le T\text{ or }v_j\le T\}}\Big)^{\frac{p-2}{p}}\Big]\\&\quad\le\frac{2\delta}{p}\mathbb{E}\Big(\sup_{0\le t\le T}|e(t)|^p\Big)+\frac{p-2}{p\delta^{2/(p-2)}}\mathbb{P}(u_j\le T\text{ or }v_j\le T).\end{aligned}$$

Hence

$$\begin{aligned}\mathbb{E}\Big(\sup_{0\le t\le T}|e(t)|^2\Big)&\le\mathbb{E}\Big(\sup_{0\le t\le T}|e(t)|^2 1_{\{\rho_j>T\}}\Big)+\frac{2\delta}{p}\mathbb{E}\Big(\sup_{0\le t\le T}|e(t)|^p\Big)\\&\quad+\frac{p-2}{p\delta^{2/(p-2)}}\mathbb{P}(u_j\le T\text{ or }v_j\le T).\end{aligned} \tag{4.3}$$

Now,

$$\begin{aligned}\mathbb{P}(u_j\le T)&\le\mathbb{E}\Big(1_{\{u_j\le T\}}\frac{\|x_t\|^p}{j^p}\Big)\\&\le\frac{1}{j^p}\mathbb{E}\Big(\sup_{-\tau\le t\le T}|x(t)|^p\Big)\\&\le\frac{\widetilde{H}}{j^p}.\end{aligned}$$

Similarly,

$$\mathbb{P}(v_j \le T) \le \frac{\widetilde{H}}{j^p}.$$

Thus

$$\begin{aligned}\mathbb{P}(v_j \le T \text{ or } u_j \le T) &\le \mathbb{P}(v_j \le T) + \mathbb{P}(u_j \le T)\\ &\le \frac{2\widetilde{H}}{j^p}. \end{aligned} \tag{4.4}$$

We also have

$$\begin{aligned}\mathbb{E}\Big(\sup_{0\le t\le T}|e(t)|^p\Big) &\le 2^{p-1}\mathbb{E}\Big(\sup_{0\le t\le T}(|x_t|^p + |y_t|^p)\Big)\\ &\le 2^p\widetilde{H}. \end{aligned} \tag{4.5}$$

Moreover,

$$\begin{aligned}\mathbb{E}\Big(\sup_{0\le t\le T}|e(t)|^2 1_{\{\rho_j>T\}}\Big) &= \mathbb{E}\Big(\sup_{0\le t\le T}|e(t\wedge\rho_j)|^2 1_{\{\rho_j>T\}}\Big)\\ &\le \mathbb{E}\Big(\sup_{0\le t\le T}|e(t\wedge\rho_j)|^2\Big).\end{aligned}$$

Using these bounds in (4.3) yields

$$\mathbb{E}\Big(\sup_{0\le t\le T}|e(t)|^2\Big) \le \mathbb{E}\Big(\sup_{0\le t\le T}|e(t\wedge\rho_j)|^2\Big) + \frac{2^{p+1}\delta\widetilde{H}}{p} + \frac{2(p-2)\widetilde{H}}{p\delta^{2/(p-2)}j^p}. \tag{4.6}$$

Setting $r := t\wedge\rho_j$ and for any $\varepsilon\in(0,1)$, by the Hölder inequality, when $r\in[k\triangle,(k+1)\triangle]$, for $k=0,1,2,\dots,M-1$,

$$\begin{aligned}|e(r)|^2 &= |x(r)-y(r)|^2\\ &\le \Big|u(x_r) - u(\bar{y}_{(k-1)\triangle} + (r-k\triangle)(\bar{y}_{k\triangle}-\bar{y}_{(k-1)\triangle})/\triangle)\\ &\quad + \int_0^r [f(x_s)-f(\bar{y}_s)]ds + \int_0^r [g(x_s)-g(\bar{y}_s)]dw(s)\Big|^2\\ &\le \frac{1}{\varepsilon}|u(x_r) - u(\bar{y}_{(k-1)\triangle} + (r-k\triangle)(\bar{y}_{k\triangle}-\bar{y}_{(k-1)\triangle})/\triangle)|^2\\ &\quad + \frac{2}{1-\varepsilon}\Big[T\int_0^r [f(x_s)-f(\bar{y}_s)]^2ds + \Big|\int_0^r [g(x_s)-g(\bar{y}_s)]dw(s)\Big|^2\Big].\end{aligned}$$

Since

$$\begin{aligned}&\Big|u(x_r) - u\Big(\bar{y}_{(k-1)\triangle} + \frac{r-k\triangle}{\triangle}(\bar{y}_{k\triangle}-\bar{y}_{(k-1)\triangle})\Big)\Big|^2\\ &\le \kappa^2\Big\|x_r - \bar{y}_{(k-1)\triangle} - \frac{r-k\triangle}{\triangle}(\bar{y}_{k\triangle}-\bar{y}_{(k-1)\triangle})\Big\|^2\\ &\le \kappa^2\Big\||x_r-y_r| + |y_r-\bar{y}_r| + |\bar{y}_{k\triangle}-\bar{y}_{(k-1)\triangle}| - \frac{r-k\triangle}{\triangle}|\bar{y}_{k\triangle}-\bar{y}_{(k-1)\triangle}|\Big\|^2\\ &\le \frac{\kappa^2}{\varepsilon}\|x_r-y_r\|^2 + \frac{2\kappa^2}{1-\varepsilon}\Big(\|y_r-\bar{y}_r\|^2 + \|\bar{y}_{k\triangle}-\bar{y}_{(k-1)\triangle}\|^2\Big),\end{aligned}$$

we have

$$|e(r)|^2 \le \frac{\kappa^2}{\varepsilon^2}\|x_r - y_r\|^2 + \frac{2\kappa^2}{\varepsilon(1-\varepsilon)}(\|y_r - \bar{y}_r\|^2 + \|\bar{y}_{k\triangle} - \bar{y}_{(k-1)\triangle}\|^2)$$
$$+\frac{2}{1-\varepsilon}\Big[T\int_0^r [f(x_s) - f(\bar{y}_s)]^2 ds + \Big|\int_0^r [g(x_s) - g(\bar{y}_s)]dw(s)\Big|^2\Big].$$

Hence, for any $t_1 \in [0, T]$, by Lemmas 3.2 and 3.3,

$$\begin{aligned}
&\mathbb{E}\Big[\sup_{0\le t\le t_1}|e(t\wedge\rho_j)|^2\Big]\\
&\le \frac{\kappa^2}{\varepsilon^2}\mathbb{E}\Big(\sup_{0\le t\le t_1}\|x_{t\wedge\rho_j} - y_{t\wedge\rho_j}\|^2\Big) + \frac{2\kappa^2}{\varepsilon(1-\varepsilon)}\Big[\mathbb{E}\Big(\sup_{0\le t\le t_1}\|y_{t\wedge\rho_j} - \bar{y}_{t\wedge\rho_j}\|^2\Big)\\
&\quad + \mathbb{E}\Big(\sup_{0\le k\le M-1}\|\bar{y}_{k\triangle\wedge\rho_j} - \bar{y}_{(k-1)\triangle\wedge\rho_j}\|^2\Big)\Big]\\
&\quad + \frac{2T}{1-\varepsilon}\mathbb{E}\int_0^{t_1\wedge\rho_j}[f(x_s) - f(\bar{y}_s)]^2 ds\\
&\quad + \frac{2}{1-\varepsilon}\mathbb{E}\Big[\sup_{0\le t\le t_1}\Big|\int_0^{t\wedge\rho_j}[g(x_s) - g(\bar{y}_s)]dw(s)\Big|^2\Big]\\
&\le \frac{\kappa^2}{\varepsilon^2}\mathbb{E}\Big(\sup_{0\le t\le t_1}\|x_{t\wedge\rho_j} - y_{t\wedge\rho_j}\|^2\Big) + \frac{2\kappa^2}{\varepsilon(1-\varepsilon)}(\gamma(\triangle) + \beta(\triangle))\\
&\quad + \frac{2T}{1-\varepsilon}\mathbb{E}\int_0^{t_1\wedge\rho_j}[f(x_s) - f(\bar{y}_s)]^2 ds\\
&\quad + \frac{2}{1-\varepsilon}\mathbb{E}\Big[\sup_{0\le t\le t_1}\Big|\int_0^{t\wedge\rho_j}[g(x_s) - g(\bar{y}_s)]dw(s)\Big|^2\Big]. \qquad (4.7)
\end{aligned}$$

Since $x(t) = y(t) = \xi(t)$ when $t \in [-\tau, 0]$, we have

$$\begin{aligned}
&\mathbb{E}\Big(\sup_{0\le t\le t_1}\|x_{t\wedge\rho_j} - y_{t\wedge\rho_j}\|^2\Big)\\
&\le \mathbb{E}\Big(\sup_{-\tau\le\theta\le 0}\sup_{0\le t\le t_1}|x(t\wedge\rho_j + \theta) - y(t\wedge\rho_j + \theta)|^2\Big)\\
&\le \mathbb{E}\Big(\sup_{-\tau\le t\le t_1}|x(t\wedge\rho_j) - y(t\wedge\rho_j)|^2\Big)\\
&= \mathbb{E}\Big(\sup_{0\le t\le t_1}|x(t\wedge\rho_j) - y(t\wedge\rho_j)|^2\Big). \qquad (4.8)
\end{aligned}$$

By Assumption 2.1 and Lemma 3.3, we may compute

$$\begin{aligned}
&\mathbb{E}\int_0^{t_1\wedge\rho_j}[f(x_s) - f(\bar{y}_s)]^2 ds\\
&\le C_j\mathbb{E}\int_0^{t_1\wedge\rho_j}\|x_s - \bar{y}_s\|^2 ds\\
&\le 2C_j\mathbb{E}\int_0^{t_1\wedge\rho_j}\|x_s - y_s\|^2 ds + 2C_j\mathbb{E}\int_0^{t_1\wedge\rho_j}\|y_s - \bar{y}_s\|^2 ds
\end{aligned}$$

$$\le 2C_j\mathbb{E}\int_0^{t_1\wedge\rho_j}\sup_{-\tau\le\theta\le 0}|x(s+\theta)-y(s+\theta)|^2ds+2C_j\int_0^{t_1}\mathbb{E}\|y_s-\bar{y}_s\|^2ds$$

$$\le 2C_j\mathbb{E}\int_0^{t_1}\sup_{-\tau\le\theta\le 0}|x(s\wedge\rho_j+\theta)-y(s\wedge\rho_j+\theta)|^2ds+2C_jT\beta(\triangle)$$

$$\le 2C_j\mathbb{E}\int_0^{t_1}\sup_{-\tau\le r\le s}|x(r\wedge\rho_j)-y(r\wedge\rho_j)|^2ds+2C_jT\beta(\triangle)$$

$$= 2C_j\mathbb{E}\int_0^{t_1}\sup_{0\le r\le s}|x(r\wedge\rho_j)-y(r\wedge\rho_j)|^2ds+2C_jT\beta(\triangle). \tag{4.9}$$

By the Doob martingale inequality and Assumption 2.1,

$$\mathbb{E}\Big[\sup_{0\le t\le t_1}\Big|\int_0^{t\wedge\rho_j}[g(x_s)-g(\bar{y}_s)]dw(s)\Big|^2\Big]$$

$$\le 4\mathbb{E}\int_0^{t_1\wedge\rho_j}|g(x_s)-g(\bar{y}_s)|^2ds$$

$$\le 4C_j\mathbb{E}\int_0^{t_1\wedge\rho_j}\|x_s-\bar{y}_s\|^2ds$$

$$\le 8C_j\int_0^{t_1}\mathbb{E}\Big(\sup_{0\le r\le s}|x(r\wedge\rho_j)-y(r\wedge\rho_j)|^2\Big)ds+8C_jT\beta(\triangle). \tag{4.10}$$

Substituting (4.8), (4.9), (4.10) into (4.7) yields

$$\Big(1-\frac{\kappa^2}{\varepsilon^2}\Big)\mathbb{E}\Big[\sup_{0\le t\le t_1}|e(t\wedge\rho_j)|^2\Big]\le\frac{2\kappa^2}{\varepsilon(1-\varepsilon)}[\beta(\triangle)+\gamma(\triangle)]+\frac{4C_jT(T+4)}{1-\varepsilon}\beta(\triangle)$$

$$+\frac{4C_j(T+4)}{1-\varepsilon}\int_0^{t_1}\mathbb{E}\Big(\sup_{0\le r\le s}|e(r\wedge\rho_j)|^2\Big)ds. \tag{4.11}$$

Choosing $\varepsilon=(1+\kappa)/2$ and noting that $\kappa\in(0,1)$, we have

$$\mathbb{E}\Big(\sup_{0\le t\le t_1}|e(t\wedge\rho_j)|^2\Big)$$

$$\le\frac{8\kappa^2(1+\kappa)}{(1-\kappa)^2(1+3\kappa)}[\beta(\triangle)+\gamma(\triangle)]+\frac{8C_jT(1+\kappa)^2(T+4)}{(1-\kappa)^2(1+3\kappa)}\beta(\triangle)$$

$$+\frac{8C_j(1+\kappa)^2(T+4)}{(1-\kappa)^2(1+3\kappa)}\int_0^{t_1}\mathbb{E}\Big(\sup_{0\le s\le r}|e(s\wedge\rho_j)|^2\Big)dr$$

$$\le\frac{8}{(1-\kappa)^2}[\beta(\triangle)+\gamma(\triangle)]+\frac{16}{(1-\kappa)^2}C_jT(T+4)\beta(\triangle)$$

$$+\frac{16}{(1-\kappa)^2}C_j(T+4)\int_0^{t_1}\mathbb{E}\Big(\sup_{0\le s\le r}|e(s\wedge\rho_j)|^2\Big)dr.$$

The Gronwall inequality yields

$$\mathbb{E}\Big[\sup_{0\le t\le t_1}|e(t\wedge\rho_j)|^2\Big]\le\frac{8}{(1-\kappa)^2}[\beta(\triangle)+\gamma(\triangle)+2C_jT(T+4)\beta(\triangle)]$$

$$\times e^{\frac{16}{(1-\kappa)^2}C_jT(T+4)}. \tag{4.12}$$

By (4.6),

$$\mathbb{E}\Big(\sup_{0\le t\le T}|e(t)|^2\Big) \le \frac{8}{(1-\kappa)^2}[\beta(\triangle)+\gamma(\triangle)+2C_jT(T+4)\beta(\triangle)]e^{\frac{16}{(1-\kappa)^2}C_jT(T+4)}$$

$$+\frac{2^{p+1}\delta\widetilde{H}}{p}+\frac{2(p-2)\widetilde{H}}{p\delta^{2/(p-2)}j^p}. \tag{4.13}$$

Given any $\epsilon>0$, we can now choose δ sufficiently small such that $2^{p+1}\delta\widetilde{H}/p\le\epsilon/3$, then choose j sufficiently large such that

$$\frac{2(p-2)\widetilde{H}}{p\delta^{2/(p-2)}j^p}<\frac{\epsilon}{3},$$

and finally choose $\triangle$ so small that

$$\frac{8}{(1-\kappa)^2}[\beta(\triangle)+\gamma(\triangle)+2C_jT(T+4)\beta(\triangle)]e^{\frac{16}{(1-\kappa)^2}C_jT(T+4)}<\frac{\epsilon}{3}$$

and thus $\mathbb{E}(\sup_{0\le t\le T}|e(t)|^2)\le\epsilon$, as required. □

Remark. It should be pointed out that much simpler proofs of Lemmas 3.2 and 3.3 can be obtained by choosing $l=2$ if we only want to prove Theorem 2.2. The reason why we want to control the stochastic term by $c_1(l)\Delta^{\frac{l-1}{l}}$ for any $l>1$ in section 3 is for the use in the next section, where we will show the order of the strong convergence.

5. Order of convergence under the global Lipschitz condition. Theorem 2.2 shows that under Assumptions 2.1–2.4 the Euler–Maruyama approximate solutions strongly converge to the true solution. However, this theorem does not give the order of the convergence. In this section we reveal the order of the convergence, but here we need to replace the local Lipschitz condition by the global Lipschitz condition. To be more precise, we state the assumption as follows.

Assumption 5.1 (global Lipschitz condition). There exists a constant C such that for all $\varphi,\psi\in C([-\tau,0];\mathbb{R}^n)$ and $t\in[0,T]$

$$|f(\varphi)-f(\psi)|^2\vee|g(\varphi)-g(\psi)|^2\le C\|\varphi-\psi\|^2. \tag{5.1}$$

It is easy to see from the global Lipschitz condition that for any $\varphi\in C([-\tau,0];\mathbb{R}^n)$,

$$|f(\varphi)|^2\vee|f(\psi)|^2\le 2(|f(0)|^2\vee|f(0)|^2)+C\|\varphi\|^2. \tag{5.2}$$

In other words, the global Lipschitz condition implies a linear growth condition with the growth coefficient

$$K=2[|f(0)|^2\vee|g(0)|^2\vee C].$$

In addition, we need to slightly strengthen Assumption 2.4 on the initial data.

Assumption 5.2. $\xi\in L^p_{\mathcal{F}_0}([-\tau,0];\mathbb{R}^n)$ for some $p\ge2$, and there exists a positive constant λ such that

$$\mathbb{E}\Big(\sup_{-\tau\le s\le t\le0}|\xi(t)-\xi(s)|^2\Big)\le\lambda(t-s). \tag{5.3}$$

We can state another theorem which reveals the order of the convergence.

THEOREM 5.1. *Under Assumptions* 5.1, 5.2, *and* 2.3, *for any positive constant* $\varepsilon \in (0,1)$,

$$\mathbb{E}\Big(\sup_{0\le t\le T}|x(t)-y(t)|^2\Big) \le O(\triangle^{1-\varepsilon}). \tag{5.4}$$

Proof. Since $\alpha(\triangle)$ may be replaced by $\lambda\triangle$, from Lemmas 3.2 and 3.3, there exists constants $\tilde{c}_1(l)$ and $\tilde{c}_2(l)$ such that $\beta(\triangle) \le \tilde{c}_1(l)\triangle^{\frac{l-1}{l}}$ and $\gamma(\triangle) \le \tilde{c}_2(l)\triangle^{\frac{l-1}{l}}$. Here we do not need to define the stopping times u_j and v_j; we may repeat the proof in section 4 and directly compute

$$\begin{aligned}\mathbb{E}\Big(\sup_{0\le t\le T}|e(t)|^2\Big) \le{}& \frac{8}{(1-\kappa)^2}[\tilde{c}_1(l)+\tilde{c}_2(l)+2\tilde{c}_1(l)CT(T+4)]\\ &\times e^{\frac{16CT(T+4)}{(1-\kappa)^2}}\triangle^{\frac{l-1}{l}}.\end{aligned} \tag{5.5}$$

Choosing $l \ge 1/\varepsilon$ yields the required assertion. ∎

Remark. Let us give some closing comments. Theorem 5.1 reveals that the order of convergence in mean square is close to 1, although the order of the strong convergence in mean square for the Euler–Maruyama scheme applied to both SDEs and SFDEs is 1 (see [8, 16]). Noting that $\tilde{c}_i(l)(i=1,2)$ appeared in (5.5) will tend to infinity as $l \to \infty$, we cannot simply let $l \to \infty$ to show that the order of the convergence is 1. It remains open whether the order of the convergence is 1 in the case of NSFDEs.

We would also like to point out that although we concentrate on the convergence in mean square (namely, in L^2), our methods can easily be extended to establish the strong convergence in L^p for any $p \ge 2$. In particular, under the global Lipschitz condition, we can show that the order of the strong convergence in L^p is $p/2-\varepsilon$ for any $\varepsilon > 0$.

Acknowledgments. The authors would like to thank the Chinese Scholarship Council for its financial support. They also wish to thank the referees for their detailed comments and helpful suggestions.

REFERENCES

[1] D. BAHUGUNA AND S. AGARWAL, *Approximations of solutions to neutral functional differential equations with nonlocal history conditions*, J. Math. Anal. Appl., 317 (2006), pp. 583–602.
[2] C. T. H. BAKER AND E. BUCKWAR, *Numerical analysis of explicit one-step methods for stochastic delay differential equations*, LMS J. Comput. Math., 3 (2000), pp. 315–335.
[3] A. BELLEN, Z. JACKIEWICZ, AND M. ZENNARO, *Stability analysis of Runge-Kutta methods for neutral delay-differential equations*, Numer. Math., 52 (1988), pp. 605–619.
[4] E. BUCKWAR, *Introduction to the numerical analysis of stochastic delay differential equations*, J. Comput. Appl. Math., 125 (2000), pp. 297–307.
[5] F. HARTUNG, T. L. HERDMAN, AND J. TURI, *On existence, uniqueness and numerical approximation for neutral equations with state-dependent delays*, Appl. Numer. Math., 24 (1997), pp. 393–409.
[6] Y. HU, S.-E. A. MOHAMMED, AND F. YAN, *Discrete-time approximations of stochastic delay equations: The Milstein scheme*, Ann. Probab., 32 (2004), pp. 265–314.
[7] Z. JACKIEWICZ, *One-step methods of any order for neutral functional differential equations*, SIAM J. Numer. Anal., 21 (1984), pp. 486–511.
[8] P. E. KLOEDEN AND E. PLATEN, *Numerical Solutions of Stochastic Differential Equations*, Springer, Berlin, 1992.
[9] U. KÜCHLER AND E. PLETEN, *Strong discrete time approximation of stochastic differential equations with time delay*, Math. Comput. Simulation, 54 (2000), pp. 189–205.
[10] K. LIU AND X. XIA, *On the exponential stability in mean square of neutral stochastic functional differential equations*, Systems Control Lett., 37 (1999), pp. 207–215.

[11] Y. Liu, *Numerical solution of implicit neutral functional differential equations*, SIAM J. Numer. Anal., 36 (1999), pp. 516–528.
[12] X. Mao, *Approximate solutions for a class of stochastic evolution equations with variable delays—part* II, Numer. Funct. Anal. Optim., 15 (1994), pp. 65–76.
[13] X. Mao, *Exponential stability in mean square of neutral stochastic differential functional equations*, Systems Control Lett., 26 (1995), pp. 245–251.
[14] X. Mao, *Stochastic Differential Equations and Applications*, Horwood, Chichester, 1997.
[15] X. Mao, *Razumikhin-type theorems on exponential stability of neutral stochastic functional differential equations*, SIAM J. Math. Anal., 28 (1997), pp. 389–401.
[16] X. Mao, *Numerical solutions of stochastic functional differential equations*, LMS J. Comput. Math., 6 (2003), pp. 141–161.
[17] X. Mao and S. Sabanis, *Numerical solutions of stochastic differential delay equations under local Lipschitz condition*, J. Comput. Appl. Math., 151 (2003), pp. 215–227.
[18] J. Randielovic and S. Jankovic, *On the pth moment exponential stability criteria of neutral stochastic functional differential equations*, J. Math. Anal. Appl., 326 (2007), pp. 266–280.

SIAM J. NUMER. ANAL.
Vol. 46, No. 4, pp. 1842–1866
© 2008 Society for Industrial and Applied Mathematics

ENERGY-CONSERVING ALGORITHMS FOR A COROTATIONAL FORMULATION*

J. SALOMON†, A. A. WEISS†, AND B. I. WOHLMUTH†

Abstract. Standard nonlinear schemes for the the simulation of elastodynamic problems have several shortcomings when considering high-speed rotations. To tackle these problems, we use a corotational framework and a corresponding specific linearization to design new energy-conserving numerical schemes. In the two-dimensional case, an algorithm preserving also angular momentum is presented. The existence of a solution for the fully discrete setting of this algorithm is established. Numerical results illustrate the flexibility and efficiency of the proposed algorithms.

Key words. elastodynamic, corotational formulation, energy-conservative schemes, large displacements, small strains

AMS subject classifications. 35L70, 74B05, 74B20, 74H15

DOI. 10.1137/060669863

1. Introduction.

The time discretization of elastodynamic problems involving large rotations and translations is a well-known challenge in the simulation of continuum mechanics. Usually, geometrically nonlinear strain measures such as Green–Lagrange strains are applied to tackle this kind of problem. The resulting nonlinear equation requires iterative solvers which can be very costly, for example when considering high-speed rotation in combination with large time steps. Moreover, such methods show unsatisfying numerical results in these cases as the global motion is not reproduced correctly and spurious oscillations can be observed.

In order to accelerate the computations, one can use the framework of linear elasticity, i.e., linearize the evolution equation with respect to parameters associated with the displacements of the body. However, this kind of linearization does not show satisfying results for rotating bodies. In particular, rigid body rotations wrongly contribute to the elastic energy as the small deformation assumption is not fulfilled anymore.

As a remedy, the motion can be split into a rigid body motion and a purely elastic effective deformation. This idea was initially introduced in [10] to trace the global motion of elastic bodies in the context of aircraft and spacecraft dynamics. Then, the approach was put in an abstract framework by Fraejis de Veubeke [20], where different criteria were presented to guarantee the uniqueness of the decomposition. Nowadays, this idea is often used by the engineering community where it is referred to as "corotational formulation." In this paper, we use this splitting to carry out the linearization with respect to the effective displacements. To this end, we make use of the Fraejis de Veubeke's criterion and choose the corotational decomposition that minimizes the L^2-norm of the effective displacements. The well-posedness of the associated system has been proved recently in [23, 24].

*Received by the editors September 14, 2006; accepted for publication (in revised form) April 4, 2007; published electronically April 18, 2008. This work was supported in part by the Deutsche Forschungsgemeinschaft, SFB 404, B8.

http://www.siam.org/journals/sinum/46-4/66986.html

†Institut für Angewandte Analysis und Numerische Simulation (IANS), Universität Stuttgart, Pfaffenwaldring 57, D-70569 Stuttgart, Germany (salomon@ians.uni-stuttgart.de, weiss@ians.uni-stuttgart.de, wohlmuth@ians.uni-stuttgart.de).

The problem of handling large rotations in the Lagrangian kinematic description of continuum mechanics has spawned several publications, e.g., [1, 2, 3, 25]. The application of the splitting of the deformation in combination with finite elements was introduced by Wempner [37] and Belytschko and Hsieh [7]. In the engineering literature, the term "corotational finite elements" was established for this method starting with [5, 6] where the splitting was used for each element, resulting in additional degrees of freedom per element. Some developments of this approach were presented in [17, 18, 34, 36]. The method was improved by defining a corotated frame per element [8, 9]; see also the overview [27]. Crisfield [12, 13] developed the concept of consistent linearization in the corotational formulation. Nour-Omid and Rankin [26], Rankin and Brognan [32], and Rankin and Nour-Omid [33] finally invented the element-independent corotational formulation, where projectors were used to separate the total deformation. Recently, the corotational formulation has been used for control and regulation [4, 28], sensitivity analysis [29, 30, 31], and modern solution techniques for large-scale simulations [15, 16]. For a more exhaustive overview of the development of the corotational formulation, we refer the reader to [19]. However, the design of suitable time integration methods for the corotational framework is still an open question; see topic E of section 7 in [19].

An important property of a suitable time discretization is the conservation of energy and angular momentum. For standard schemes, one can proceed using the discrete energy momentum method by Simo and Tarnow; see [35] or similar variants, e.g., [22]. This method has also been successfully adapted to the corotational framework for special finite element models; see [14, 21]. However, conservation of the angular momentum is not always guaranteed and the effect of linearizations with respect to effective deformations has not been considered so far.

In this paper, we show how the linearization induced by a corotational formulation enables us to design a new energy-conserving time discretization of elastodynamic problems in two and three dimensions. In two dimensions, one of our algorithms also preserves the angular momentum. The existence of a solution for the fully discrete setting is established in this case. In the other cases, the angular momentum is asymptotically preserved and a priori estimates are obtained. These new algorithms are compared to standard algorithms used in the simulation of elastodynamics involving large rigid body motion.

Our new schemes are quite attractive from the computational point of view and show high accuracy in the case of large rigid body motions. Moreover, the approach can be easily extended to nonlinear strain measures and nonlinear material laws. Numerical simulations indicate that the resulting algorithms are almost as efficient as the linearized version.

Though most of the concepts presented in this paper hold for both two and three dimensions, we focus on the two-dimensional (2D) case for simplicity. However, the necessary adaptations to the three-dimensional (3D) case are presented together with a 3D version of one of our algorithms.

The rest of the paper is organized as follows. In section 2, we derive the 2D model and present the continuous linearized system of equations describing the evolution of the system. We state two time-discretized energy conservative algorithms in section 3. In section 4, the space discretization on a constrained space and the existence of a solution is discussed. In section 5, we extend one of our algorithms to the 3D case. Finally, in section 6, we show the efficiency and flexibility of our approach by means of various numerical tests.

2. Problem setting. Throughout this paper, we use the notation

$$\dot{a}(x,t) := \frac{\partial}{\partial t} a(x,t), \quad \ddot{a}(x,t) := \frac{\partial^2}{\partial t^2} a(x,t).$$

Given a matrix A, we denote by $A^\top$ its transposed and by $\operatorname{tr}(A)$ its trace. The identity matrix of $\mathbb{R}^2$ is denoted by I and R_θ represents the rotation of θ, $R_\theta := \left(\begin{smallmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{smallmatrix}\right)$. The matrix Π is defined by $\Pi := R_{\frac{\pi}{2}} = \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$ and $A : B := \operatorname{tr}(A^\top B)$ is the usual matrix scalar product. We note that using this notation the derivative of the rotation matrix reads $\frac{d}{d\theta} R_\theta = \Pi R_\theta$.

2.1. Nonlinear evolution equation. Let $\Omega \subset \mathbb{R}^2$ be a bounded domain with a piecewise sufficiently smooth boundary $\Gamma := \partial\Omega$ describing an elastic body with mass density $\rho(x)$. The coordinate system is chosen such that the center of gravity has the coordinates $(0,0)$, i.e., $\int_\Omega \rho x = 0$. The body is subject to some volume forces $f(x,t)$ and boundary tractions $g(x,t)$. Let $\varphi : \Omega \times [0,T] \to \mathbb{R}^2$ be the deformation mapping. In the fully nonlinear setting, the evolution of φ is described by the weak form of the equilibrium condition [11],

$$\int_\Omega \rho \ddot{\varphi} \cdot \eta + \int_\Omega \frac{\partial}{\partial E} W\big(E(d)\big) : F^\top \nabla \eta = \int_\Omega f \cdot \eta + \int_\Gamma g \cdot \eta, \tag{2.1}$$

where η is a function in $V := [H^1(\Omega)]^2$, $d := \varphi - x$, and $F := \nabla\varphi = I + \nabla d$ is the deformation gradient. The so-called Green–Lagrange strain is given by

$$E(d) := \frac{1}{2}(\nabla d + \nabla d^\top + \nabla d^\top \nabla d). \tag{2.2}$$

We assume a Saint Venant–Kirchhoff material model where the stored energy $W\big(E(d)\big)$ is calculated from

$$W\big(E(d)\big) := \frac{\lambda}{2} \left(\operatorname{tr}\big(E(d)\big)\right)^2 + \mu \operatorname{tr}\big(E(d)^2\big),$$

with Lamé parameters λ and μ. For this linear material model, the second Piola–Kirchhoff stress tensor $\Sigma(d) := \frac{\partial}{\partial E} W\big(E(d)\big)$ reads

$$\Sigma(d) = 2\mu E(d) + \lambda \operatorname{tr} E(d) I.$$

Energy- and momentum-conserving time integration schemes for this setting are well-known; see, e.g., [35]. However, when considering high-speed rotation in combination with large timesteps, these standard time integration schemes show poor accuracy: In this case, the rotation speed is not reproduced correctly which results in a wrong solution; see section 6.1.

2.2. Corotational decomposition of the motion. As a remedy, the corotational formulation proposes to work with effective displacements in smaller spaces by removing the large rigid body component of the displacements involved in the elastic term of (2.1). The deformation is decomposed as follows [24]:

$$\varphi(x,t) = \tau(t) + R_{\theta(t)}\big(x + u(x,t)\big). \tag{2.3}$$

In this equation, $\tau(t) \in \mathbb{R}^2$ describes the rigid body translation of the body and R_θ its rigid body rotation. After this change of variable, the effective displacements are described by u. An important property of the decomposition (2.3) is that

$$E\big(d\big) = E\big(u\big), \tag{2.4}$$

which shows that the nonlinear model correctly cancels the rigid body motion appearing in φ. In order to define a unique triple (τ, θ, u) and to work with small effective displacements, we consider the decomposition that minimizes the weighted L^2-norm of u, $\int_\Omega \rho u^2$; see [20]. A straightforward computation shows that if $\int_\Omega \rho(x+u)\cdot x \neq 0$, the minimization condition for u is equivalent to (see [23, 24])

$$\int_\Omega \rho u = 0, \quad \int_\Omega \rho u \cdot \Pi x = 0. \tag{2.5}$$

We note that the first condition results from the translation and the second condition from the rotation term. Thus, we are looking for the effective displacements u in the linear space $X := X^{\mathrm{trans}} \cap X^{\mathrm{rot}}$ defined by

$$X^{\mathrm{trans}} := \left\{ u \in V;\ \int_\Omega \rho u = 0 \right\}, \quad X^{\mathrm{rot}} := \left\{ u \in V;\ \int_\Omega \rho u \cdot \Pi x = 0 \right\}.$$

Let us now consider the effects of this change of variable on (2.1).
We start with the elastic term of (2.1). Defining $\hat{F} := I + \nabla u$, we have, thanks to (2.4),

$$\int_\Omega \frac{\partial}{\partial E} W\big(E(d)\big) : F^\top \nabla \eta = \int_\Omega \frac{\partial}{\partial E} W\big(E(u)\big) : \hat{F}^\top R_\theta^\top \nabla \eta = \int_\Omega \Sigma(u) : \hat{F}^\top R_\theta^\top \nabla \eta. \tag{2.6}$$

The inertial term of (2.1) can also be rewritten in this new setting. To this end, we define the relative velocity by

$$s(x,t) := R_{\theta(t)}^\top \big(\dot{\varphi}(x,t) - \dot{\tau}(t)\big) = \dot{u}(x,t) + \dot{\theta}(t)\Pi\big(x + u(x,t)\big). \tag{2.7}$$

The acceleration term can then be expressed by

$$\ddot{\varphi}(x,t) = \ddot{\tau}(t) + R_{\theta(t)} \left(\dot{s}(x,t) + \dot{\theta}(t)\Pi s(x,t) \right). \tag{2.8}$$

Using this representation of $\ddot{\varphi}$ and reducing the test space to X, (2.1) now reads

$$\int_\Omega \rho \left(\dot{s} + \dot{\theta}\Pi s \right) \cdot v + \int_\Omega \Sigma(u) : \hat{F}^\top \nabla v = B_\theta(v) \quad \forall v \in X, \tag{2.9}$$

where we have set $v = R_\theta^\top \eta$ and

$$B_\theta(v) := \int_\Omega f \cdot R_\theta v + \int_\Gamma g \cdot R_\theta v. \tag{2.10}$$

We note that the term $\int_\Omega \ddot{\tau} \cdot R_\theta v$ cancels because $R_\theta v \in X^{\mathrm{trans}}$. To compensate this reduction of the test space, additional constraints have to be imposed which are then used to determine the correct values of θ and τ. Thus, in addition to (2.9), we require that the linear and angular momentum are preserved. To obtain the corresponding conditions, we formally test (2.9) with $v^1_{\mathrm{trans}} = R_\theta^\top (0,1)^\top$, $v^2_{\mathrm{trans}} = R_\theta^\top (1,0)^\top$, and $v_{\mathrm{rot}} = \Pi(x+u)$, respectively, to get the linear and angular momentum equations. Using $u \in X^{\mathrm{trans}}$, this gives the conservation of the linear momentum,

$$\mathcal{M}\ddot{\tau} = \int_\Omega f + \int_\Gamma g, \tag{2.11}$$

and the conservation of the angular momentum,

$$\dot{\mathcal{J}} = B_\theta(\Pi(x+u)), \tag{2.12}$$

where $\mathcal{M}$ and $\mathcal{J}$ are defined by

$$\mathcal{M} := \int_\Omega \rho, \quad \mathcal{J} := \int_\Omega \rho s \cdot \Pi(x+u).$$

2.3. Linearization of the elastic term. In order to accelerate the computation or when considering numerical methods to solve (2.9), some additional linearization can be introduced. For small effective displacements u, the strain relationship can be linearized. Using the following approximation of (2.2):

$$E(u) \approx \varepsilon(u) := \frac{1}{2}(\nabla u + \nabla u^\top),$$

and $\hat{F} \approx I$, the equilibrium equation (2.9) becomes linear with respect to u,

$$\int_\Omega \rho \left(\dot{s} + \dot{\theta}\Pi s\right) \cdot v + \int_\Omega \sigma(u) : \nabla v = B_\theta(v), \tag{2.13}$$

where $\sigma(u) := 2\mu\varepsilon(u) + \lambda \operatorname{tr} \varepsilon(u) I$.

We will consider the nonlinear models (2.1) and (2.9) as a reference to measure the efficiency of our algorithms.

Let us now introduce the space Y defined by

$$Y := \left\{ u \in X;\ \int_\Omega \rho(x+u) \cdot x > 0 \right\}.$$

In a recent paper [23] (see also [24]), Grandmont, Maday, and Metier have obtained the following existence result concerning the system (2.11)–(2.13).

THEOREM 2.1. *Let τ_0, $\dot{\tau}_0$ be given in $\mathbb{R}^2$, θ_0, $\dot{\theta}_0$ be real numbers, $u_0 \in Y, \dot{u}_0 \in L^2(\Omega)$, $f \in L^2_{loc}(0,+\infty; L^2(\Omega))$, and $g \in H^1_{loc}(0,+\infty; H^{-\frac{1}{2}}(\Gamma)) \cap L^2_{loc}(0,+\infty; L^1(\Gamma))$. There exists $T^\star > 0$ such that for all $T < T^\star$ there exists a unique triplet (τ, θ, u) with $\tau \in H^2(0,T)$, $\theta \in H^2(0,T)$, $u \in H^2(0,T; (H^1(\Omega))') \cap W^{1,\infty}(0,T; L^2(\Omega)) \cap L^\infty(0,T; H^1(\Omega))$, and $u \in \mathcal{C}^0([0,T]; Y)$ satisfying*

$$\begin{cases} \mathcal{M}\ddot{\tau} = \displaystyle\int_\Omega f + \int_\Gamma g \text{ in } L^2(0,T), \\ \dot{\mathcal{J}} = B_\theta(\Pi(x+u)) \text{ in } \mathcal{D}'(0,T), \\ \forall v \in X, \displaystyle\int_\Omega \rho \left(\dot{s} + \dot{\theta}\Pi s\right) \cdot v + \sigma(u) : \varepsilon(v) = B_\theta(v) \text{ in } \mathcal{D}'(0,T), \end{cases} \tag{2.14}$$

where s is defined by (2.7) together with $\tau(0) = \tau_0$, $\dot{\tau}(0) = \dot{\tau}_0$, $\theta(0) = \theta_0$, $\dot{\theta}(0) = \dot{\theta}_0$, $u(0) = u_0$, $\dot{u}(0) = \dot{u}_0$.

Moreover, the following alternative holds:

- $T^\star = +\infty$ *or*
- $\lim_{t \to T^\star} \int_\Omega \rho(x+u) \cdot x = 0$.

3. Energy conservative algorithms. We now present our energy conservative algorithms which are based on two different time discretizations of (2.14). The first one guarantees the conservation of the energy, but not of the angular momentum. Next, an alternative scheme is presented that preserves both quantities. We briefly outline the strategies we have followed to design our algorithms. Denoting by $\frac{d}{dt}$ the time derivation operator, we find in terms of (2.7) and (2.8),

$$R_\theta^\top(\ddot{\varphi}-\ddot{\tau}) = \left(\frac{d}{dt}+\dot{\theta}\Pi\right)s \tag{3.1}$$

$$= \left(\frac{d}{dt}+\dot{\theta}\Pi\right)^2(x+u)$$

$$= \ddot{u}+\Pi\left(\frac{d}{dt}(\dot{\theta}u)+\dot{\theta}\dot{u}\right)-(\dot{\theta})^2(x+u). \tag{3.2}$$

The first algorithm is now obtained by discretizing each term of (3.2). In the second algorithm, we first discretize $s = (\frac{d}{dt}+\dot{\theta}\Pi)(x+u)$ (see (3.16)) and then discretize $R_\theta^\top(\ddot{\varphi}-\ddot{\tau})$ via (3.1) (see the first term of (3.15)).

3.1. Midpoint time discretization and discrete setting. Let us denote the timestep by Δt and the time index by n with $t_n := n\Delta t$. We use a midpoint time discretization which has the property

$$\dot{a}_{n+1/2}b_{n+1/2}+a_{n+1/2}\dot{b}_{n+1/2} = [\dot{ab}]_{n+1/2}, \tag{3.3}$$

where we have used the following notations for a generic time-discretized quantity $\star$:

$$\star_{n+1/2} := \frac{\star_{n+1}+\star_n}{2}, \quad \dot{\star}_{n+1/2} := \frac{\star_{n+1}-\star_n}{\Delta t}. \tag{3.4}$$

Let us also define the total energy of the linearized system at time t_n by

$$\mathcal{E}_n := \frac{1}{2}\mathcal{M}\dot{\tau}_n^2+\frac{1}{2}\int_\Omega \rho s_n^2+\frac{1}{2}\int_\Omega \sigma(u_n):\varepsilon(u_n). \tag{3.5}$$

The angular momentum is given by

$$\mathcal{J}_n := \int_\Omega \rho s_n\cdot\Pi(x+u_n). \tag{3.6}$$

Here s_n is a time discretization of (2.7), depending on the algorithm under consideration.

Note that without linearization of the strain term, the energy corresponding to (2.9) is defined by

$$\mathcal{E}_n^{nl} := \frac{1}{2}\mathcal{M}\dot{\tau}_n^2+\frac{1}{2}\int_\Omega \rho s_n^2+\frac{1}{2}\int_\Omega \Sigma(u_n):E(u_n), \tag{3.7}$$

which coincides with (3.5) when neglecting terms of second order with respect to ∇u_n. Finally, we discretize B_θ at time $t_{n+1/2}$ by

$$B_{n+1/2}(v) := \int_\Omega f_{n+1/2}\cdot R_{\theta_{n+1/2}}v+\int_\Gamma g_{n+1/2}\cdot R_{\theta_{n+1/2}}v.$$

3.2. Algorithm I. A time discretization of (2.14) can be done in terms of (3.2). Special care has to be taken about the energy conservation property. To this end, we choose suitable approximations, for example $(\dot\theta(t_{n+1/2}))^2 \approx \dot\theta_n\dot\theta_{n+1}$.

Algorithm I then reads: Let τ_0, $\dot\tau_0$, θ_0, $\dot\theta_0$, u_0, as well as the external forces $(f_{n+1/2})_{n\in\mathbb{N}}$ and $(g_{n+1/2})_{n\in\mathbb{N}}$ be given, and suppose that $\tau_n, \dot\tau_n, \theta_n, \dot\theta_n$ and $u_n, \dot u_n \in X$ have already been computed. The derivation of $\tau_{n+1}, \dot\tau_{n+1}$, $\theta_{n+1}, \dot\theta_{n+1}$ and u_{n+1}, $\dot u_{n+1} \in X$ is carried out by solving

$$\mathcal{M}\ddot\tau_{n+1/2} = \int_\Omega f_{n+1/2} + \int_\Gamma g_{n+1/2}, \tag{3.8}$$

$$\begin{aligned}\frac{1}{\Delta t}\left(\int_\Omega \rho\dot\theta_{n+1}(x+u_{n+1})^2 - \int_\Omega \rho\dot\theta_n(x+u_n)^2\right) + \int_\Omega \rho\Pi\frac{\dot\theta_{n+1}u_{n+1}+\dot\theta_n u_n}{2\dot\theta_{n+1/2}}\cdot\ddot u_{n+1/2}\\ = B_{n+1/2}(\Pi(x+u_{n+1/2})),\end{aligned} \tag{3.9}$$

$$\begin{aligned}\int_\Omega \rho\left(\ddot u_{n+1/2} + \Pi\frac{\dot\theta_{n+1}u_{n+1}-\dot\theta_n u_n}{\Delta t} + \Pi\dot\theta_{n+1/2}\dot u_{n+1/2} - \dot\theta_n\dot\theta_{n+1}(x+u_{n+1/2})\right)\cdot v\\ + \int_\Omega \sigma(u_{n+1/2}):\varepsilon(v) = B_{n+1/2}(v) \quad \forall v\in X.\end{aligned} \tag{3.10}$$

Equations (3.8)–(3.10) correspond to the three equations of (2.14). We refer the reader to section 4 for details about how to guarantee that $u_{n+1} \in X$.

We note that using the discretized relative velocity

$$s_n := \dot u_n + \dot\theta_n\Pi(x+u_n), \tag{3.11}$$

and the fact that $u_n, \dot u_n \in X$, the angular momentum $\mathcal{J}_n$ can be written as

$$\mathcal{J}_n = \int_\Omega \rho s_n\cdot\Pi(x+u_n) = \dot\theta_n\int_\Omega \rho(x+u_n)^2 + \int_\Omega \rho\dot u_n\cdot\Pi u_n.$$

This quantity is not preserved exactly by Algorithm I because of the third term of the left-hand side of (3.9). It is, however, possible to specify the local order with respect to Δt. Indeed, we have

$$\mathcal{J}_{n+1} - \mathcal{J}_n - \Delta t B_{n+1/2}\big(\Pi(x+u_{n+1/2})\big) = \frac{\Delta t^2}{4}\frac{\ddot\theta_{n+1/2}}{\dot\theta_{n+1/2}}\Pi\dot u_{n+1/2}\cdot\ddot u_{n+1/2}. \tag{3.12}$$

Although the momentum is asymptotically preserved if Δt tends to zero, we observe that local oscillations occur; see Figure 6.11. This motivates the introduction of Algorithm II.

3.3. Algorithm II. We now describe a time discretization of (2.14) that preserves both the energy and the angular momentum.

Let τ_0, $\dot\tau_0$, θ_0, u_0, s_0, the external forces $(f_{n+1/2})_{n\in\mathbb{N}}$, and $(g_{n+1/2})_{n\in\mathbb{N}}$ be given, and suppose that τ_n, $\dot\tau_n$, θ_n, $u_n \in X$ and s_n have already been computed. The derivation of $\tau_{n+1}, \dot\tau_{n+1}$, θ_{n+1}, $u_{n+1} \in X$ and $s_{n+1} \in V$ is carried out by solving

$$\mathcal{M}\ddot\tau_{n+1/2} = \int_\Omega f_{n+1/2} + \int_\Gamma g_{n+1/2}, \tag{3.13}$$

$$\dot{\mathcal{J}}_{n+1/2} = B_{n+1/2}\big(\Pi(x+u_{n+1/2})\big), \tag{3.14}$$

$$G(\dot s_{n+1/2}, s_{n+1/2}, \dot\theta_{n+1/2}, u_{n+1/2}; v) = B_{n+1/2}(v) \quad \forall v\in X, \tag{3.15}$$

with

$$G(\dot{s}, s, \dot{\theta}, u; v) := \int_\Omega \rho(\dot{s} + \dot{\theta}\Pi s) \cdot v + \int_\Omega \sigma(u) : \varepsilon(v)$$

and

$$s_{n+1/2} := \dot{u}_{n+1/2} + \dot{\theta}_{n+1/2}\Pi\big(x + u_{n+1/2}\big). \tag{3.16}$$

Equations (3.13)–(3.15) correspond to the three equations of (2.14). Equation (3.14) guarantees that the angular momentum is preserved by this algorithm.

Remark. We note that the time discretization of the auxiliary variable s is done at the timestep n (see (3.11)) in Algorithm I and at the staggered timestep $(n+1/2)$ (see (3.16)) in Algorithm II. Moreover, given a parameter $\star$ and denoting by $\star^I$ and $\star^{II}$ its time discretizations in Algorithms I and II, respectively, we have that

$$s^I_{n+1/2} = s^{II}_{n+1/2} + \frac{\Delta t^2}{4}\ddot{\theta}^{II}_{n+1/2}\Pi\dot{u}^{II}_{n+1/2}.$$

3.4. Energy conservation. Let us state the main property of our algorithms.

THEOREM 3.1. *Algorithms* I *and* II *defined by* (3.8)–(3.10) *and* (3.13)–(3.15), *respectively, guarantee the energy conservation in the sense that*

$$\mathcal{E}_{n+1} - \mathcal{E}_n = \Delta t\Big(\int_\Omega f_{n+1/2} \cdot \dot{\varphi}_{n+1/2} + \int_\Gamma g_{n+1/2} \cdot \dot{\varphi}_{n+1/2}\Big).$$

Here, $\mathcal{E}_n$ is the total energy defined by (3.5) *where s is defined by* (3.11) *when considering Algorithm* I *and by* (3.16) *when considering Algorithm* II. *The time discretization of $\dot{\varphi}$ is done by*

$$\dot{\varphi}_{n+1/2} := \dot{\tau}_{n+1/2} + R_{\theta_{n+1/2}} s_{n+1/2}.$$

Proof. We give only the proof for the case of Algorithm II, since the one corresponding to Algorithm I can be obtained in a similar way.

Testing (3.15) with $v = \dot{u}_{n+1/2}$, we obtain

$$\int_\Omega \rho(\dot{s}_{n+1/2} + \dot{\theta}_{n+1/2}\Pi s_{n+1/2}) \cdot \dot{u}_{n+1/2} + \int_\Omega \sigma(u_{n+1/2}) : \varepsilon(\dot{u}_{n+1/2}) = B_{n+1/2}(\dot{u}_{n+1/2}). \tag{3.17}$$

On the other hand, using (3.3), (3.14) reads

$$\int_\Omega \rho\dot{s}_{n+1/2} \cdot \Pi(x + u_{n+1/2}) + \int_\Omega \rho s_{n+1/2} \cdot \Pi\dot{u}_{n+1/2} = B_{n+1/2}\big(\Pi(x + u_{n+1/2})\big). \tag{3.18}$$

Multiplying (3.18) by $\Delta t\dot{\theta}_{n+1/2}$ and (3.17) by Δt and adding the two resulting equations, we get by means of (3.16),

$$\begin{aligned}\Delta t \int_\Omega \rho\dot{s}_{n+1/2} \cdot \big(\dot{u}_{n+1/2} + \dot{\theta}_{n+1/2}\Pi(x + u_{n+1/2})\big) + \Delta t \int_\Omega \sigma(u_{n+1/2}) : \varepsilon(\dot{u}_{n+1/2}) \\ = \frac{1}{2}\left(\int_\Omega \rho s^2_{n+1} + \sigma(u_{n+1}) : \varepsilon(u_{n+1}) - \int_\Omega \rho s^2_n + \sigma(u_n) : \varepsilon(u_n)\right),\end{aligned}$$

and thus

(3.19)

$$\frac{1}{2}\left(\int_\Omega \rho s_{n+1}^2 + \sigma(u_{n+1}) : \varepsilon(u_{n+1}) - \int_\Omega \rho s_n^2 + \sigma(u_n) : \varepsilon(u_n)\right) = \Delta t B_{n+1/2}(s_{n+1/2}).$$

Using (3.13), we end up with

$$\frac{1}{2}\mathcal{M}\dot{\tau}_{n+1}^2 - \frac{1}{2}\mathcal{M}\dot{\tau}_n^2 = \Delta t\left(\int_\Omega f_{n+1/2}\cdot\dot{\tau}_{n+1/2} + \int_\Gamma g_{n+1/2}\cdot\dot{\tau}_{n+1/2}\right),$$

which yields the energy conservation in combination with (3.19). ∎

In contrast to Algorithm I(see (3.12)), Algorithm II preserves the angular momentum; thanks to (3.14), we have

$$\mathcal{J}_{n+1} - \mathcal{J}_n - \Delta t B_{n+1/2}\big(\Pi(x+u_{n+1/2})\big) = 0. \tag{3.20}$$

4. Space discretization. In this section, we present a suitable space discretization for the time-discretized systems. We focus on (3.15), although the same method can be used for the discretization of (3.10). The space X motivates the use of a Galerkin method based on the spectral decomposition of the operator $-\mathrm{div}(\sigma(\cdot))$. Indeed, we note that X is spanned by the eigenvectors with nonzero eigenvalue and is orthogonal (in the sense of the $L^2(\Omega)$ scalar product) on the kernel of this operator.

Unfortunately, finite elements do not fit into this setting. We observe that the conditions (2.5) are not satisfied by nodal finite element basis functions. The orthogonal projection of these basis functions onto the space X leads to a nonlocal support of the finite element basis and additional computational costs. We therefore use a Lagrange multiplier approach to guarantee (2.5) and state the problem in a saddle point framework. For this fully discretized setting, we show the existence of a solution of (3.13)–(3.15).

4.1. Finite element approximation of the space X. We assume that the domain Ω is resolved by a mesh consisting of quadrilaterals and/or triangles. We use a conforming finite element space of lowest order, denoting the nodal basis functions by ϕ_i and the finite element space by V_h. The mass matrix is referred to by M, $M_{ij} := \int_\Omega \rho\phi_i\cdot\phi_j$, and the stiffness matrix for linear elasticity by S, $S_{ij} := \int_\Omega \sigma(\phi_i) : \varepsilon(\phi_j)$. The subspace X_h of X is defined by $X_h := X \cap V_h$.

To guarantee that $u \in V_h$ is also in X_h, i.e., satisfies (2.5), we introduce two Lagrange multipliers $\alpha \in \mathbb{R}$ and $\beta \in \mathbb{R}^2$ and rewrite (3.15) as

$$\begin{aligned} G(\dot{s}_{n+1/2}, s_{n+1/2}, \dot{\theta}_{n+1/2}, u_{n+1/2}; v) + \alpha\int_\Omega \rho v\cdot\Pi x + \int_\Omega \rho\beta\cdot v &= B_{n+1/2}(v)\ \forall v\in V_h,\\ \int_\Omega \rho u_{n+1/2}\cdot\Pi x &= 0,\\ \int_\Omega \rho u_{n+1/2} &= 0. \end{aligned} \tag{4.1}$$

The Lagrange multiplier β can be eliminated by static condensation as follows. First, we recall that $s_n, u_n, u_{n+1/2} \in X^{\mathrm{trans}}$, implying $\int_\Omega \rho\Pi u_{n+1/2} = \Pi\int_\Omega \rho u_{n+1/2} = 0$. Therefore $\Pi u_{n+1/2} \in X^{\mathrm{trans}}$ and thus $s_{n+1/2}, \dot{s}_{n+1/2} \in X^{\mathrm{trans}}$. Using a constant

$v \in V_h$, we have $\varepsilon(v) = 0$ and therefore $G(\dot{s}_{n+1/2}, s_{n+1/2}, \dot{\theta}_{n+1/2}, u_{n+1/2}; v) = 0$. We use $v^1 = (1,0)^\top$ and $v^2 = (0,1)^\top$, respectively, in (4.1) and finally obtain

$$\beta = \frac{1}{\mathcal{M}} \left(\int_\Omega R^\top_{\theta_{n+1/2}} f_{n+1/2} + \int_\Gamma R^\top_{\theta_{n+1/2}} g_{n+1/2} \right) = R^\top_{\theta_{n+1/2}} \ddot{\tau}_{n+1/2}. \tag{4.2}$$

Inserting (4.2) in (4.1) gives the following system:

$$\begin{aligned} G(\dot{s}_{n+1/2}, s_{n+1/2}, \dot{\theta}_{n+1/2}, u_{n+1/2}; v) + \alpha \int_\Omega \rho v \cdot \Pi x = \tilde{B}_{n+1/2}(v) \quad \forall v \in V_h, \\ \int_\Omega \rho u_{n+1/2} \cdot \Pi x = 0, \end{aligned} \tag{4.3}$$

where $\tilde{B}_{n+1/2}$ is defined by

$$\tilde{B}_{n+1/2}(v) := B_{n+1/2}(v) - \int_\Omega \rho R^\top_{\theta_{n+1/2}} \ddot{\tau}_{n+1/2} \cdot v.$$

We note that the first line of (4.3) ensures $\int_\Omega \rho u_{n+1/2} = 0$, i.e., $u_{n+1/2} \in X^{\text{trans}}$.

We rewrite (4.3) in terms of $u_{n+1/2}$ and obtain the following linear system for the vector of nodal values:

$$\left(\begin{array}{c|c} \tilde{G}(\theta_{n+1} - \theta_n, \Delta t) & M\Pi x \\ \hline (M\Pi x)^\top & 0 \end{array} \right) \cdot \left(\begin{array}{c} u_{n+1/2} \\ \hline \alpha_{n+1/2} \end{array} \right) = \left(\begin{array}{c} \tilde{c}_{n+1/2}(\theta_{n+1} - \theta_n, \Delta t) \\ \hline 0 \end{array} \right), \tag{4.4}$$

where

- $\tilde{G}(\delta, \Delta t) := M\big(2I + \delta\Pi\big)^2 + \Delta t^2 S$,
- $\tilde{c}_{n+1/2}(\delta, \Delta t) := M\big(2I + \delta\Pi\big)(2u_n - \delta\Pi x) + 2M\Delta t s_n + \Delta t^2 \tilde{B}_{n+1/2}$.

We note that we follow the widely applied abuse of notation; i.e., the symbols u_n, $u_{n+1/2}$, s_n, and $\tilde{B}_{n+1/2}$ denote the function or the vector of coefficients of the finite element representation, depending on the situation. We also keep the notation "$\cdot$" for the scalar product in what follows.

Equation (3.14) now reads

$$h(\theta_{n+1} - \theta_n, \Delta t) = 0, \tag{4.5}$$

with

$$\begin{aligned} h(\delta, \Delta t) :&= 4\delta(x + u_{n+1/2}) \cdot M(x + u_{n+1/2}) \\ &\quad + \left(-\delta M(x + u_n) + 2\Delta t M \Pi s_n + \Delta t^2 \Pi R^\top_{\theta_n + \delta/2} \tilde{B}_{n+1/2} \right) \cdot (x + u_{n+1/2}), \end{aligned}$$

where $u_{n+1/2}$ is the solution of (4.4) corresponding to $\theta_{n+1} - \theta_n = \delta$.

4.2. Existence of a solution for the fully discretized algorithm. This section is devoted to the existence of a solution of the fully discretized system associated with Algorithm II.

THEOREM 4.1. *Given $\mathcal{E}_{max} > 0$, $B_{max} > 0$, and $\nu > 0$, suppose that at time t_n we have $\mathcal{E}_n \le \mathcal{E}_{max}$, $\|\tilde{B}_{n+1/2}\| \le B_{max}$, where $\|\cdot\|$ is a given norm on V_h and*

$$\nu \le x \cdot M(x + u_n). \tag{4.6}$$

Then there exists $\Delta t^* > 0$ depending on $\mathcal{E}_{max}$, B_{max}, and ν such that, given $\Delta t \leq \Delta t^*$, τ_n, $\dot{\tau}_n$, θ_n, u_n, and s_n, as well as the external forces $f_{n+1/2}$ and $g_{n+1/2}$, there exists $\theta_{n+1} \in \mathbb{R}$ satisfying (4.5) (with $u_{n+1/2}$ the corresponding solution of (4.4)).
Moreover, there exist $\delta^+(\Delta t)$, $\delta^-(\Delta t)$ such that

- $\lim_{\Delta t \to 0} \delta^+(\Delta t) = 2$,
- $\lim_{\Delta t \to 0} \delta^-(\Delta t) = -2$,
- $\theta_{n+1} \in [\theta_n + \delta^-(\Delta t), \theta_n + \delta^+(\Delta t)]$.

Proof. Given $\Delta t > 0$, $\eta \in]0,1[$, and $\delta \in L := [-2-\eta, 2+\eta]$, we denote by $u(\delta, \Delta t) := u_{n+1/2}$ the solution of (4.4) corresponding to $\theta_{n+1} - \theta_n = \delta$. Since $\tilde{G}(\delta, 0)$ is invertible for $\delta \in L$, there exists Δt_0 (depending only on S) such that $\tilde{G}(\delta, \Delta t)$ is invertible on $L \times [0, \Delta t_0]$. Then, let us define $q(\delta, \Delta t)$ by

$$q(\delta, \Delta t) := \Pi x \cdot \tilde{G}(\delta, \Delta t)^{-1} \Pi x.$$

The inversion of (4.4) provides

$$u(\delta, \Delta t) = \tilde{G}(\delta, \Delta t)^{-1} \left(\tilde{c}_{n+1/2}(\delta, \Delta t) - \frac{1}{q(\delta, \Delta t)} \Big((\Pi x) \cdot \tilde{G}(\delta, \Delta t)^{-1} \tilde{c}_{n+1/2}(\delta, \Delta t) \Big) \Pi x \right).$$

By means of (4.6), we get for $\delta \neq 0$,

$$(\Pi x) \cdot \tilde{G}(\delta, 0)^{-1} \tilde{c}_{n+1/2}(\delta, 0) = -\frac{2\delta}{(4+\delta^2)^2} x \cdot M(x + u_n) \neq 0.$$

There exists $\Delta t_0'$ depending on $\mathcal{E}_{max}$, B_{max}, and ν such that

$$\forall (\delta, \Delta t) \in \{[-2-\eta, -2+\eta] \cup [2-\eta, 2+\eta]\} \times [0, \Delta t_0'],$$
$$(\Pi x) \cdot \tilde{G}(\delta, \Delta t)^{-1} \tilde{c}_{n+1/2}(\delta, \Delta t) \neq 0. \tag{4.7}$$

Let us focus now on the behavior of $q(\cdot, \cdot)$ on $\{[-2-\eta, -2+\eta] \cup [2-\eta, 2+\eta]\} \times [0, \Delta t_0']$. We have

$$q(\delta, 0) = \frac{4 - \delta^2}{(4+\delta^2)^2} x \cdot Mx,$$

such that $q(-2, 0) = q(2, 0) = 0$ and $q(\delta, 0) \neq 0$ for $\delta \neq \pm 2$. Moreover, since

$$\frac{\partial q}{\partial \delta}(\pm 2, 0) = \mp \frac{1}{16} x \cdot Mx \neq 0,$$

the implicit function theorem proves the existence of $\Delta t_0''$ (depending only on S) and two functions δ^+ and δ^- defined on $[0, \Delta t_0'']$ such that

$$\forall \Delta t \in [0, \Delta t_0''], \qquad q\big(\delta^+(\Delta t), \Delta t\big) = 0, \ q\big(\delta^-(\Delta t), \Delta t\big) = 0,$$
$$\lim_{\Delta t \to 0} \delta^-(\Delta t) = -2, \ \lim_{\Delta t \to 0} \delta^+(\Delta t) = 2.$$

We have then proved the existence of two roots of $\delta \mapsto q(\delta, \Delta t)$ in L for all $\Delta t \leq \Delta t_0''$. By decreasing $\Delta t_0''$, these two roots can be guaranteed to be unique: indeed, suppose that there exists a sequence $(\Delta t_n)_{n \in \mathbb{N}}$ of $[0, \Delta t_0'']$ converging towards 0 and $(r_n)_{n \in \mathbb{N}}$ a sequence of roots of $h(\cdot, \Delta t_n)$ such that

$$\forall n \in \mathbb{N}, \quad r_n \neq \delta^+(\Delta t_n), \ r_n \neq \delta^-(\Delta t_n).$$

Since L is compact, we can assume that $(r_n)_{n\in\mathbb{N}}$ is convergent. By continuity of $q(\cdot,\cdot)$, the possible limits are -2 and 2. Suppose, for example, that

$$\lim_{n\to+\infty} r_n = 2;$$

then

$$\lim_{n\to\infty} \frac{q\big(\delta^+(\Delta t_n),\Delta t_n\big) - q(r_n,\Delta t_n)}{\delta^+(\Delta t_n) - r_n} = 0 = \frac{\partial q}{\partial \delta}(2,0) \neq 0,$$

which gives a contradiction. The case $\lim_{n\to+\infty} r_n = -2$ is treated similarly.

Thanks to (4.7), we obtain that for $\Delta t \leq \Delta t^* := \min(\Delta t_0, \Delta t_0', \Delta t_0'')$, the function $\phi : \delta \mapsto u(\delta,\Delta t)\cdot Mu(\delta,\Delta t)$ satisfies

- ϕ is continuous on $L - \{\delta^-(\Delta t), \delta^+(\Delta t)\}$,
- $\lim_{\delta\to\delta^-(\Delta t)} \phi(\delta) = \lim_{\delta\to\delta^+(\Delta t)} \phi(\delta) = +\infty$.

Since $h(\delta,\Delta t) = 4\delta\phi(\delta) + a_{n+1/2}(\delta).u(\delta,\Delta t) + b_{n+1/2}(\delta)$, where $a_{n+1/2}$ and $b_{n+1/2}$ are bounded on L, the existence of θ_{n+1} follows. □

Remark. Note that the hypothesis (4.6) is the discretized version of the one of Theorem 2.1 and relies on the fact that the model is valid only for motions close to the rigid body motion, whereas the hypothesis on the parameters B_{max} and $\mathcal{E}_{max}$ stem from the discretization.

5. Extension to the 3D case. This section is devoted to the extension of Algorithm II to the 3D case. Here the rotation is parameterized by a vector. As a consequence, Algorithm I cannot be extended to three dimensions since it strongly relies on the scalar nature of the angular parameter; see, e.g., (3.9).

The subset Ω is now a domain in $\mathbb{R}^3$. We denote by $\times$ the cross product defined by

$$\begin{pmatrix} a \\ b \\ c \end{pmatrix} \times \begin{pmatrix} a' \\ b' \\ c' \end{pmatrix} := \begin{pmatrix} bc' - b'c \\ a'c - ac' \\ ab' - a'b \end{pmatrix}.$$

5.1. Parameterization of 3D rotations. The major problem met when extending Algorithm II to three dimensions is to parameterize the rotation corresponding to the previous R_θ. We choose to describe it by using the Euler angles and refer the reader to Appendix A of [19] for a review on other parameterizations. We remark that the representation of a rotation with respect to Euler angles is not unique. Given $\theta \in \mathbb{R}$, let us introduce the rotation matrices R^1_θ, R^2_θ, and R^3_θ given by

$$R^1_\theta := \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{pmatrix},\ R^2_\theta := \begin{pmatrix} \cos\theta & 0 & -\sin\theta \\ 0 & 1 & 0 \\ \sin\theta & 0 & \cos\theta \end{pmatrix},$$

$$R^3_\theta := \begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

We also define for $i = 1,2,3$ the matrices $\Pi^i := \frac{1}{2}\big(R^i_{\frac{\pi}{2}} - R^i_{-\frac{\pi}{2}}\big)$.

Consider now $\boldsymbol{\theta} = (\theta_1,\theta_2,\theta_3) \in \mathbb{R}^3$. We can express a general 3D rotation matrix $R_{\boldsymbol{\theta}}$ as follows:

$$R_{\boldsymbol{\theta}} := R^1_{\theta_1} R^2_{\theta_2} R^3_{\theta_3}.$$

For $\boldsymbol{\theta}$ depending on the time variable t, the derivative of $R_{\boldsymbol{\theta}}$ reads

$$\dot{R}_{\boldsymbol{\theta}} = R_{\boldsymbol{\theta}} \Theta^{\boldsymbol{\theta}}(t),$$

where $\Theta^{\boldsymbol{\theta}}(t)$ is the skew-symmetric matrix defined by

(5.1)
$$\Theta^{\boldsymbol{\theta}}(t) := \dot{\theta}_1(t)(R^3_{\theta_3(t)})^\top (R^2_{\theta_2(t)})^\top \Pi^1 R^3_{\theta_3(t)} R^2_{\theta_2(t)} + \dot{\theta}_2(t)(R^3_{\theta_3(t)})^\top \Pi^2 R^3_{\theta_3(t)} + \dot{\theta}_3(t)\Pi^3.$$

Note that $\Theta^{\boldsymbol{\theta}}(t)$ depends on both $\boldsymbol{\theta}(t)$ and $\dot{\boldsymbol{\theta}}(t)$.

5.2. Corotational model. Before considering the system of evolution equations, we introduce the space $\widetilde{X} = \widetilde{X}^{\mathrm{trans}} \cap \widetilde{X}^{\mathrm{rot}}$, defined by

$$\widetilde{X}^{\mathrm{trans}} := \left\{ u \in \widetilde{V};\ \int_\Omega \rho u = 0 \right\}, \quad \widetilde{X}^{\mathrm{rot}} := \left\{ u \in \widetilde{V};\ \int_\Omega \rho u \times x = 0 \right\},$$

where $\widetilde{V} := [H^1(\Omega)]^3$.

Let us now look at the time derivatives of the deformation φ. As for the 2D case, we decompose φ into

$$\varphi(x,t) = \tau(t) + R_{\boldsymbol{\theta}}(t)(x + u(x,t)),$$

with $\tau(t) \in \mathbb{R}^3$ and $u \in \widetilde{X}$.

Introducing the relative velocity s defined by

$$s(x,t) := \dot{u}(x,t) + \Theta^{\boldsymbol{\theta}}(t)\big(x + u(x,t)\big),$$

we obtain

$$\dot{\varphi}(x,t) = \dot{\tau}(t) + R_{\boldsymbol{\theta}(t)} s(x,t)$$

and

$$\ddot{\varphi}(x,t) = \ddot{\tau}(t) + R_{\boldsymbol{\theta}(t)} \Big(\dot{s}(x,t) + \Theta^{\boldsymbol{\theta}}(t) s(x,t) \Big).$$

The angular momentum is given by

$$\mathcal{J} := R_{\boldsymbol{\theta}} \int_\Omega \rho s \times (x + u).$$

Note that, in contrast to the 2D case, we have to keep the rotation $R_{\boldsymbol{\theta}}$ in this definition, since $\int_\Omega \rho s \times (x+u)$ cannot be collinear to the axis of the rotation. This fact motivates the introduction of the relative angular momentum m defined by

$$m := \int_\Omega \rho s \times (x + u).$$

In terms of m, the time derivative of $\mathcal{J}$ reads

$$\dot{\mathcal{J}} = R_{\boldsymbol{\theta}} \big(\dot{m} + \Theta^{\boldsymbol{\theta}} m \big).$$

The dynamics of our system can be described by the following three equations:

$$\mathcal{M}\ddot{\tau} = \int_\Omega f + \int_\Gamma g,$$

$$\dot{m} + \Theta^{\boldsymbol{\theta}} m = \int_\Omega R_{\boldsymbol{\theta}}^\top f \times (x + u) + \int_\Gamma R_{\boldsymbol{\theta}}^\top g \times (x + u),$$

and

$$\int_\Omega \rho(\dot s + \Theta^{\boldsymbol\theta} s)\cdot v + \int_\Omega \sigma(u):\varepsilon(v) = B_{\boldsymbol\theta}(v)\ \forall v \in \widetilde X,$$

where

$$B_{\boldsymbol\theta}(v) := \int_\Omega f\cdot R_{\boldsymbol\theta} v + \int_\Gamma g\cdot R_{\boldsymbol\theta} v.$$

5.3. 3D Algorithm. We are now able to extend Algorithm II to the 3D case. We keep the notation corresponding to the midpoint time discretization; see (3.4).

Let τ_0, $\dot\tau_0$, $\boldsymbol\theta_0$, u_0, s_0, as well as the external forces $(f_{n+1/2})_{n\in\mathbb N}$ and $(g_{n+1/2})_{n\in\mathbb N}$ be given, and suppose that τ_n, $\dot\tau_n$, $\boldsymbol\theta_n$, $u_n \in \widetilde X$, and s_n have already been computed. The derivation of $\tau_{n+1}, \dot\tau_{n+1}$, $\boldsymbol\theta_{n+1}$, $u_{n+1} \in \widetilde X$ and $s_{n+1} \in \widetilde V$ is carried out by the resolution of

$$\mathcal M\ddot\tau_{n+1/2} = \int_\Omega f_{n+1/2} + \int_\Gamma g_{n+1/2}, \tag{5.2}$$

$$\begin{aligned} K(\dot m_{n+1/2}, m_{n+1/2}, \dot{\boldsymbol\theta}_{n+1/2}, \boldsymbol\theta_{n+1/2}) = {}& \int_\Omega R^\top_{\boldsymbol\theta_{n+1/2}} f_{n+1/2}\times(x+u_{n+1/2}) \\ & + \int_\Gamma R^\top_{\boldsymbol\theta_{n+1/2}} g_{n+1/2}\times(x+u_{n+1/2}), \end{aligned} \tag{5.3}$$

(5.4)

$$G(\dot s_{n+1/2}, s_{n+1/2}, \dot{\boldsymbol\theta}_{n+1/2}, \boldsymbol\theta_{n+1/2}, u_{n+1/2}; v) = B_{n+1/2}(v) \qquad \forall v \in \widetilde X,$$

with

$$\begin{aligned} K(\dot m, m, \dot{\boldsymbol\theta}, \boldsymbol\theta) &:= \dot m + \Theta^{\boldsymbol\theta} m, \\ G(\dot s, s, \dot{\boldsymbol\theta}, \boldsymbol\theta, u; v) &:= \int_\Omega \rho(\dot s + \Theta^{\boldsymbol\theta} s)\cdot v + \int_\Omega \sigma(u):\varepsilon(v), \\ B_{n+1/2}(v) &:= \int_\Omega f_{n+1/2}\cdot R_{\boldsymbol\theta_{n+1/2}} v + \int_\Gamma g_{n+1/2}\cdot R_{\boldsymbol\theta_{n+1/2}} v, \end{aligned}$$

and

(5.5)

$$m_{n+1/2} := \int_\Omega \rho s_{n+1/2}\times\big(x+u_{n+1/2}\big), \qquad s_{n+1/2} := \dot u_{n+1/2} + \Theta^{\boldsymbol\theta}_{n+1/2}\big(x+u_{n+1/2}\big).$$

Here, $\Theta^{\boldsymbol\theta}_{n+1/2}$ is the time discretization of $\Theta^{\boldsymbol\theta}(t_{n+1/2})$ computed with $\boldsymbol\theta_{n+1/2}$ and $\dot{\boldsymbol\theta}_{n+1/2}$ by (5.1).

5.4. Angular momentum and energy conservation. The energy conservation of the 3D Algorithm can be established as in the 2D case. However, regarding the angular momentum, we only get a weaker conservation property, namely that the norm of the angular momentum vector is preserved.

THEOREM 5.1. *The 3D Algorithm defined by* (5.2)–(5.4) *guarantees the energy conservation in the sense that*

$$\mathcal E_{n+1} - \mathcal E_n = \Delta t\Big(\int_\Omega f_{n+1/2}\cdot\dot\varphi_{n+1/2} + \int_\Gamma g_{n+1/2}\cdot\dot\varphi_{n+1/2}\Big),$$

where $\mathcal{E}_n$ is the total energy defined by (3.5). *The time discretization of $\dot{\varphi}$ is done by*

$$\dot{\varphi}_{n+1/2} := \dot{\tau}_{n+1/2} + R_{\boldsymbol{\theta}_{n+1/2}} s_{n+1/2},$$

with $s_{n+1/2}$ as given in (5.5).

The proof of Theorem 5.1 can be obtained by following the one of Theorem 3.1. Note that we do not have exact angular momentum conservation since $\dot{\mathcal{J}}$ cannot be written as an explicit variation of a time discretization of $\mathcal{J}$ between two timesteps. As for Algorithm I in two dimensions, it is however possible to specify the local order with respect to Δt. Defining the time-discretized angular momentum $\mathcal{J}$ by

$$\mathcal{J}_n := R_{\boldsymbol{\theta}_n} m_n,$$

where $\boldsymbol{\theta}_n, u_n$, and s_n are computed by the 3D Algorithm, it can be shown that there exists $C_{n+1/2}$, depending only on $\dot{\boldsymbol{\theta}}_{n+1/2}$, such that

$$\Big\| \mathcal{J}_{n+1} - \mathcal{J}_n - \Delta t \Big(\int_\Omega f_{n+1/2} \times R_{\boldsymbol{\theta}_{n+1/2}}(x + u_{n+1/2}) + \int_\Gamma g_{n+1/2} \times R_{\boldsymbol{\theta}_{n+1/2}}(x + u_{n+1/2}) \Big) \Big\| \le C_{n+1/2} \Delta t^2,$$

where $\|\cdot\|$ is a norm in $\mathbb{R}^3$. Moreover, the norm of the angular momentum is preserved in the following sense:

$$\mathcal{J}_{n+1}^2 - \mathcal{J}_n^2 = 2\Delta t \Big(\int_\Omega f_{n+1/2} \times R_{\boldsymbol{\theta}_{n+1/2}}(x + u_{n+1/2}) \cdot R_{\boldsymbol{\theta}_{n+1/2}} m_{n+1/2} + \int_\Gamma g_{n+1/2} \times R_{\boldsymbol{\theta}_{n+1/2}}(x + u_{n+1/2}) \cdot R_{\boldsymbol{\theta}_{n+1/2}} m_{n+1/2} \Big).$$

6. Numerical examples. In this section, we present some numerical examples which show the flexibility and efficiency of our method. We start with several 2D examples based on the linear Saint Venant–Kirchhoff material model, which are numerically studied using the energy- and momentum-conserving algorithm (Algorithm II). Then, we apply our corotational scheme to a nonlinear material law. At the end, we present a 3D example.

In our 2D examples, we compare the two corotational formulations (with and without the linearization of strains introduced in section 2.3) with the standard nonlinear formulation, where the linearization is not possible due to the rotation. In what follows, we call these three formulations "corotational linear," "corotational nonlinear," and "(standard) nonlinear," where "nonlinear" is always referred to using the nonlinearized Green–Lagrange strains (2.2). We remark that in the corotational nonlinear setting, two nonlinearities occur: the nonlinearity to determine the correct rotation angle and the nonlinearity in the stiffness term. For the nonlinear formulations, we use an energy-conserving time discretization as introduced, e.g., in [35].

To compare the results, we compute from the solution φ_n of the standard nonlinear setting in a postprocessing step the translation τ_n and the rotation angle θ_n such that the minimization condition (2.5) is fulfilled, i.e.,

$$\tau_n := \int_\Omega \rho \varphi_n, \quad \int_\Omega \rho R_{\theta_n}^\top (\varphi_n - \tau_n) \cdot \Pi x = 0.$$

We then set $\dot{\tau}_{n+1/2} = \frac{\tau_{n+1}-\tau_n}{\Delta t}$ and $\dot{\theta}_{n+1/2} = \frac{\theta_{n+1}-\theta_n}{\Delta t}$. The energy for the standard nonlinear formulation is computed by

$$\frac{1}{2}\int_\Omega \rho\dot{\varphi}_n^2 + \frac{1}{2}\int_\Omega \Sigma(d_n) : E(d_n),$$

with $d_n := \varphi_n - x$, whereas the angular momentum is given by

$$\int_\Omega \rho\dot{\varphi}_n \cdot \Pi\varphi_n.$$

For the corotational formulations, the energies are computed by (3.5) and (3.7), respectively, and the angular momentum by (3.6).

As initial deformations, we use either $u_0 = 0$, i.e., $\varphi_0 = x$, or $u_0 = u_{\rm rot}(\dot{\theta}_0)$ and $\varphi_0 = \varphi_{\rm rot}(\dot{\theta}_0)$, respectively, where $u_{\rm rot}$, $\varphi_{\rm rot}$ are the effective displacements and deformations of a body rotating with a given constant angular speed $\dot{\theta}_0$.

We obtain $u_{\rm rot}(\dot{\theta}_0)$ for the linearized setting by solving the system

$$-(\dot{\theta}_0)^2 M(x + u_{\rm rot}) + S u_{\rm rot} = 0. \tag{6.1}$$

We note that multiplying (6.1) by Πx gives $M u_{\rm rot} \cdot \Pi x = 0$ and thus $u_{\rm rot} \in X_h$.

For the nonlinearized strains, $\varphi_{\rm rot}(\dot{\theta}_0)$ is given by the solution of

$$-(\dot{\theta}_0)^2 M\varphi_{\rm rot} + S(\varphi_{\rm rot} - x) = 0. \tag{6.2}$$

Here, $S(\cdot)$ denotes the discretized nonlinear elasticity operator.

The nonlinear equation (4.5) is solved for our examples using a Newton scheme. We note that the derivative $\frac{d}{d\theta}u_{n+1/2}$ can be computed by solving the system (4.4) with a modified right-hand side. For the corotational nonlinear setting, we do not use a full Newton scheme to resolve both nonlinearities but a nested iteration scheme leading here to an inner iteration where the tangential matrix can be computed in the same way as in the standard nonlinear setting. Restricting the number of inner iterations (also called inexact solving), we get an efficient solver that is easy to implement.

6.1. Examples using the standard nonlinear setting. We start with two examples demonstrating the effect of large timesteps when using the standard nonlinear setting to simulate high-speed rotating bodies; see Figure 6.1. First, we simulate a rigid body motion with a constant rotation speed of 20; for details see Example 1 later. Here we see that the larger the stepwidth is, the worse the rotational speed is approximated. Second (see Example 3), a rigid body motion with high-speed elastic oscillations is depicted. These high-speed oscillations cannot be resolved for large timesteps. Yet, it turns out that the standard nonlinear setting with energy- and momentum-conserving time integration scheme not only reproduces a wrong mean value of the rotational speed but also artificially increases the oscillation.

6.2. Example 1: Rotating beam. We consider a beam that is rotating with a constant speed. The domain Ω is given by $[-0.5, 0.5] \times [-0.05, 0.05]$ (see Figure 6.2), and the material parameters are given by $E = 1.62 \cdot 10^5$, $\nu = 0.2$, and $\rho = 1.0$.

We assume no boundary and no volume forces, i.e., $g = 0$ and $f = 0$. As initial rotation speed, we use $\dot{\theta}_0 = 20$. For the corotational setting, we use the initial effective displacement $u_0 = u_{\rm rot}(\dot{\theta}_0)$ and $u_0 = \varphi_{\rm rot}(\dot{\theta}_0) - x$ for the linear and nonlinear strain

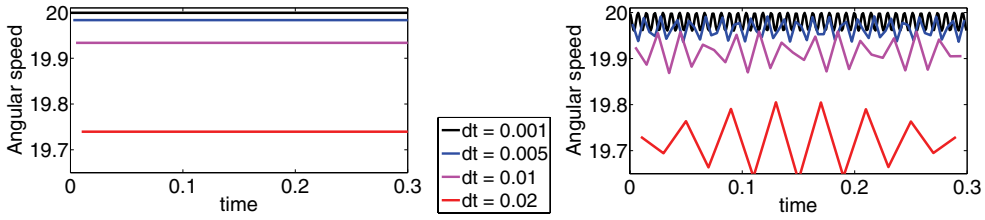


FIG. 6.1. *Calculated rotation speed using standard time discretization. Left: rigid body motion. Right: rotation plus oscillation.*

FIG. 6.2. *Example* 1. *Left: grid. Middle: initial condition for corotational formulation. Right: initial condition for nonlinear formulation (displacement with factor* 200, *effective stress).*

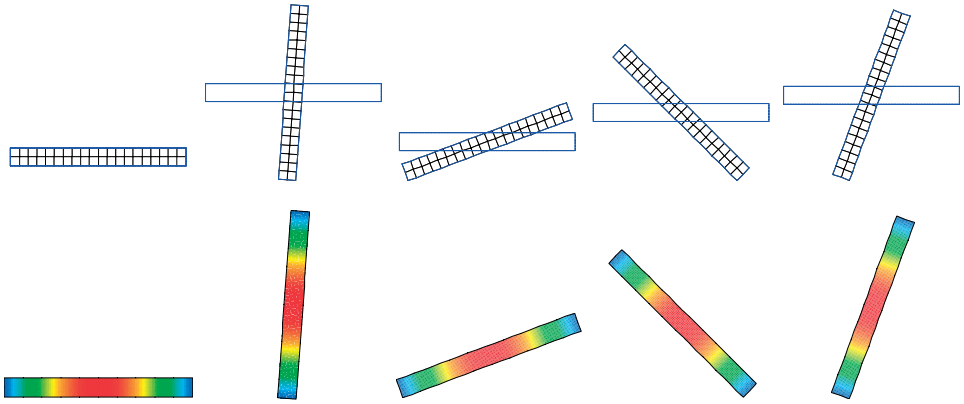


FIG. 6.3. *Example* 1. *Deformed mesh and effective stress (corotational linear) at time* $t = 0, 0.1, 0.2, 0.3, 0.4$.

setting, respectively, and set $s_0 = \dot{\theta}_0 \Pi(x + u_0)$, $\theta_0 = 0$. $\dot{\tau}_0 = (1,1)^\top$, $\tau_0 = (0,0)^\top$. With these initial conditions, it is easy to verify that the solution of (3.13)–(3.15) is given by $\theta_n = n\Delta t \dot{\theta}_0$, $\dot{\theta}_n = \dot{\theta}_0$, $\tau_n = n\Delta t \dot{\tau}_0$, $\dot{\tau}_n = \dot{\tau}_0$, $u_n = u_0$, $s_n = s_0$. The solution at various timesteps is given in Figure 6.3.

For the standard nonlinear setting, we set $\varphi_0 = \varphi_{\mathrm{rot}}(\dot{\theta}_0)$ and $\dot{\varphi}_0 = \dot{\theta}_0 \Pi \varphi_0$. As timesteps we use $\Delta t = 0.05, 0.01, 0.001$. We note that in our corotational formulations the solution is independent of the choice of Δt. Moreover, only one Newton iteration is needed if the iteration is started with θ_{n+1} such that $\dot{\theta}_{n+1/2} = \dot{\theta}_n$ which is the correct solution. This holds for both corotational formulations with and without linearized strain. For the standard nonlinear setting, we observe an increasing number of Newton steps with increasing Δt; see Figure 6.4. Moreover, it turns out that for large Δt the nonlinear approach is not able to handle the rotation correctly resulting in a wrong rotational speed.

6.3. Example 2: Rotation of a soft disc with no initial deformation. In this example, we consider a soft rotating disc, given by a circle with radius $r = 1.0$

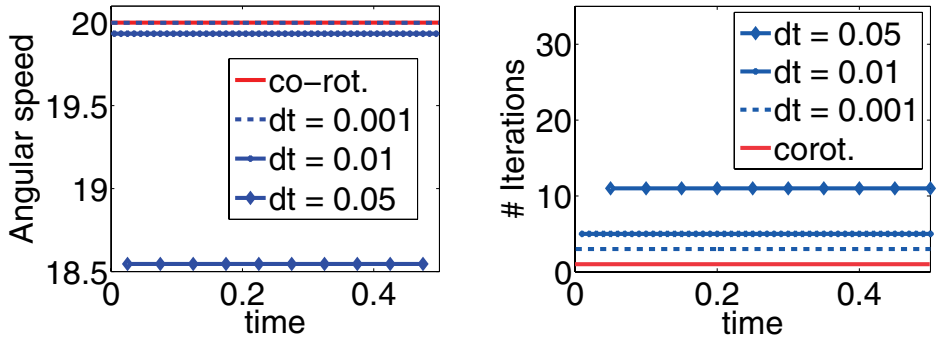


FIG. 6.4. *Example* 1. *Computed angular speed and number of iterations, both corotational formulations vs. nonlinear approach with different timesteps.*

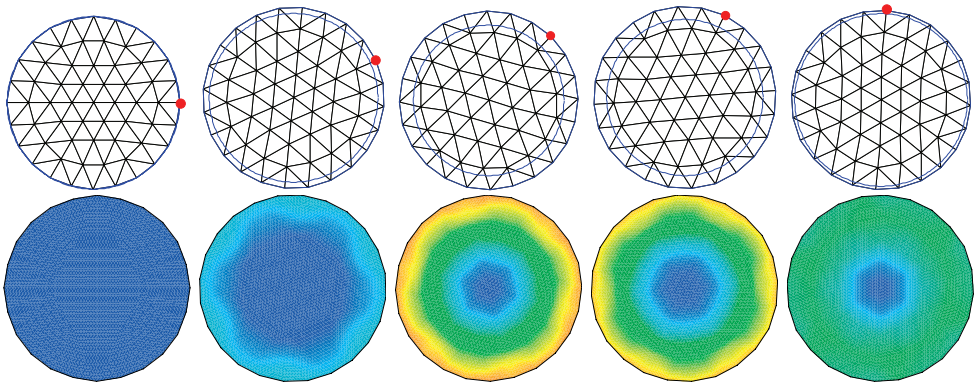


FIG. 6.5. *Example* 2. *Deformed mesh and effective stress (corotational linear) at time* t_n, $n = 0, 5, 10, 15, 20$.

and the material parameters $E = 500$, $\nu = 0.2$, and $\rho = 1.0$. We assume no boundary and volume forces and use the timestep $\Delta t = 0.001$. As initial conditions for the corotational settings, we use $\dot{\theta}_0 = 20$ and assume no initial displacements, therefore $u_0 = 0$. The initial speed is given by $s_0 = \dot{\theta}_0 \Pi x$ and the initial rotation angle by $\theta_0 = 0$. For the standard nonlinear setting, we have $\varphi_0 = x$, $\dot{\varphi}_0 = \dot{\theta}_0 \Pi x$. We assume zero translation, i.e., $\tau = (0,0)^\top$, $\dot{\tau} = (0,0)^\top$.

In Figure 6.5, the solution of the corotational linear setting at various times is shown. The calculated angular speed for the corotational and standard nonlinear approaches is depicted in Figure 6.6. We observe that the corotational linear approach fails to resolve the frequency of the solution. This is due to the linearized strain which leads to a wrong approximation of the elastic term for large deformations. However, the corotational nonlinear setting perfectly reproduces the results obtained by the standard nonlinear setting. When comparing the energies and angular momenta for the three formulations (Figure 6.7), we see that all settings yield conservation of these quantities.

In Figure 6.8, the necessary Newton steps per timestep for the standard nonlinear and the two corotational settings are shown. For $\Delta t = 0.01$ we see that the standard nonlinear setting even performs somewhat better than the corotational formulations. In a second test, we increase the stepwidth to $\Delta t = 0.1$. Here, a comparable

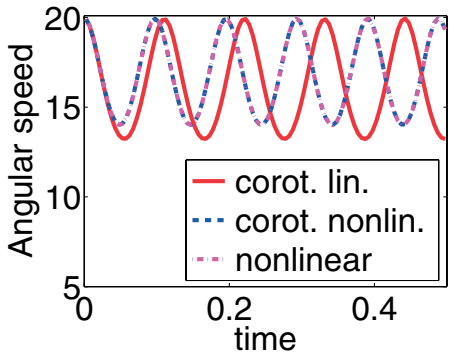


FIG. 6.6. *Example 2. Angular speed using nonlinear, corotational nonlinear, and corotational linear formulation.*

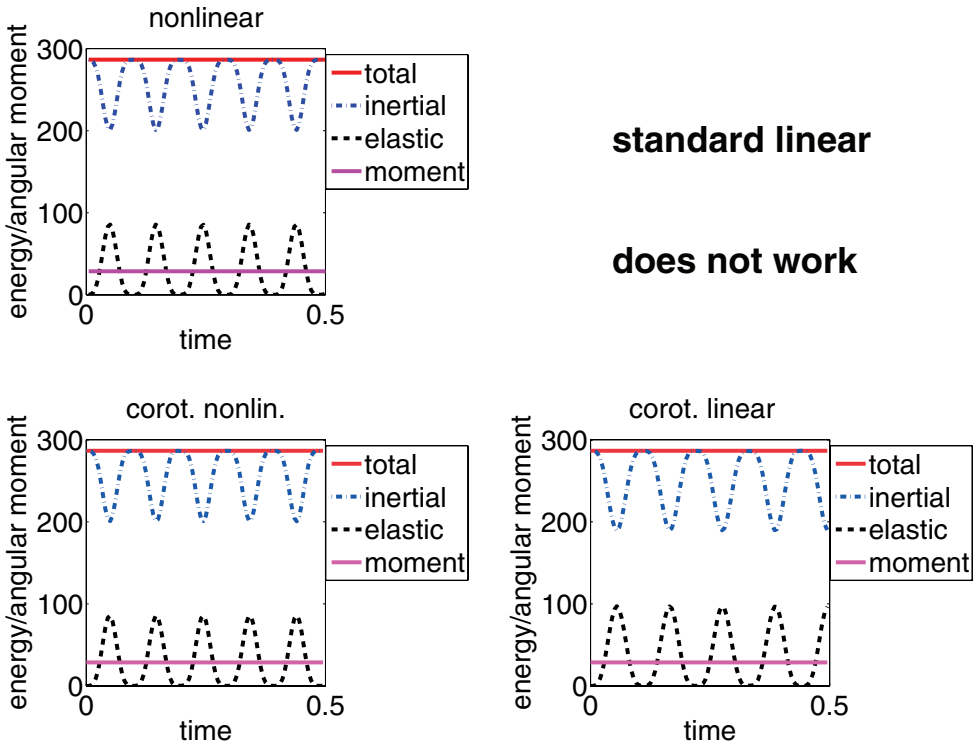


FIG. 6.7. *Example 2. Energies for nonlinear, corotational nonlinear, and corotational linear formulation.*

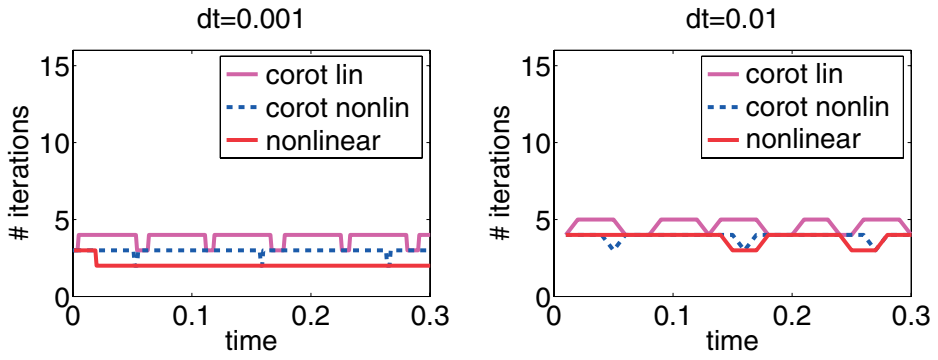


FIG. 6.8. *Example* 2. *Number of Newton iterations for nonlinear, corotational nonlinear, and corotational linear formulation at different timesteps.*

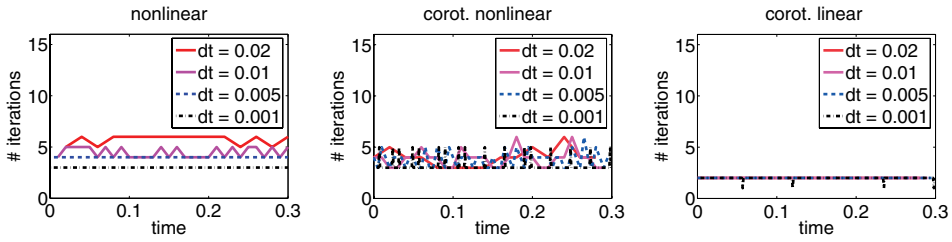


FIG. 6.9. *Example* 3. *Number of Newton iterations comparing standard nonlinear with corotational settings.*

performance is obtained for all three schemes.

6.4. Example 3: Rotation of a hard disc with no initial deformation. Now, we consider a rotating hard disc given by a circle with radius $r = 1.0$ and the material parameters $E = 1.62 \cdot 10^5$, $\nu = 0.2$, and $\rho = 1.0$. We use the same initial and boundary conditions as in the previous example and the timesteps $\Delta t = 0.001, 0.005, 0.01, 0.02$.

In Figure 6.9, the necessary Newton steps per timestep for the three formulations are shown. One can see that for small timesteps all settings perform quite well. For larger timesteps, however, the number of iterations using the standard nonlinear approach increases while the corotational formulations stay at low iteration numbers. When comparing the corotational linear with the corotational nonlinear setting, we can clearly see that in this example the linearization enhances the performance, which is expectable because the very stiff body is close to a rigid body motion.

The angular speed $\dot{\theta}$ obtained by the three settings is depicted in Figure 6.10. We see that for large timesteps the time integration scheme is no longer able to handle the oscillations correctly. Moreover, the standard nonlinear approach gives a wrong average rotation speed. This points out the fact that the standard time integration scheme with a large timestep is not able to handle the rotation correctly even when using the nonlinear strain relationship. Here, the corotational approach yields a good approximation of the angular speed for both strain-displacement relationships.

In Figure 6.11, we compare the angular momentum for the two different algorithms presented in this paper. While Algorithm II preserves the angular momentum, we see a small oscillation when using Algorithm I.

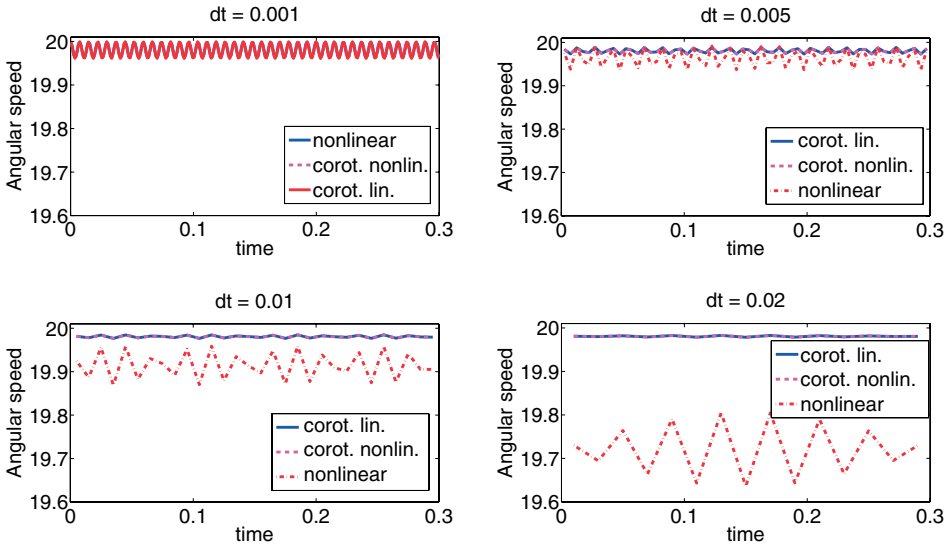


FIG. 6.10. *Example* 3. *Angular speed for nonlinear and corotational settings with different* Δt.

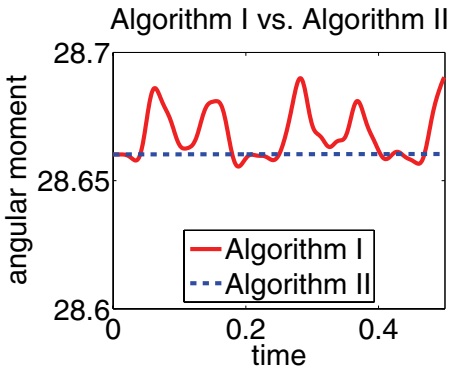


FIG. 6.11. *Example* 3. *Angular momentum using Algorithms* I *and* II.

6.5. Example 4: Rotation of a beam under surface traction. We consider a rotating beam under surface traction. For such rotational problems, boundary forces normal or perpendicular to the rotated configuration are quite common. This leads to a boundary force $g = R_\theta g_{\text{rel}}$, where g_{rel} is the boundary force relative to the rotated configuration and θ is the actual rotation angle that can be defined, for example, by (2.5). This setting causes extra difficulties when solving with standard time integration schemes as the right-hand side depends on the actual solution. However, in the corotational framework, this type of boundary condition can be modeled in a natural way and moreover gives a right-hand side that is independent of the actual rotation.

We use the same setting and initial conditions as in Example 1 and a timestep of $\Delta t = 0.01$. Here, the surface traction is given by $g_{\text{rel}} = (0, -20)^\top$ on the right boundary and $g_{\text{rel}} = (0, 20)^\top$ on the left boundary. On the top and bottom, we assume no surface traction. The rotation angle and angular speed using the corotational linear framework are depicted on the left of Figure 6.12, and the energy and angular

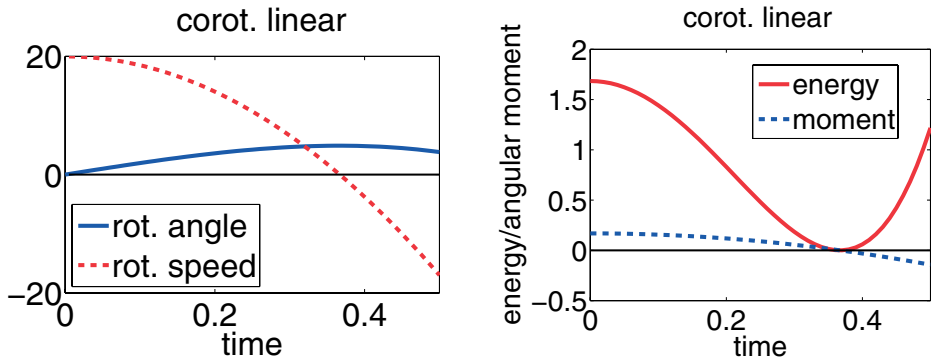


FIG. 6.12. *Example* 4. *Left: rotation angle and rotation speed. Right: energy and angular momentum.*

FIG. 6.13. *Example* 4. *Effective stress and effective displacement* u *(factor* 200*) at time* $t = 0, 0.2, 0.4, 0.5$.

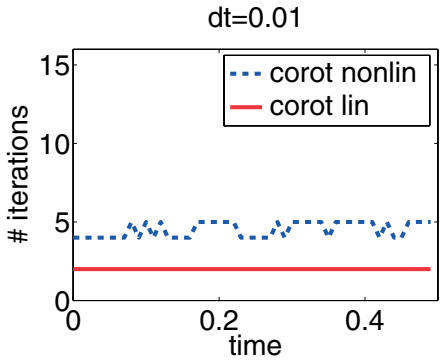


FIG. 6.14. *Example* 4. *Number of iterations.*

momentum are shown on the right. In Figure 6.13, the effective displacement u is plotted at various timesteps. We note that for this example the number of necessary Newton iterations is two for all timesteps; see Figure 6.14. Computing the same setting with the corotational nonlinear scheme leads to iteration numbers between 4 and 5 without improving the results.

6.6. Example 5: Nonlinear material law. In this example, we apply our corotational formulation to a nonlinear material law. We use the Mooney–Rivlin material model with $E = 500$, $\nu = 0.2$, and $c_m = \frac{1}{11}$. The computational domain and initial conditions are the same as in Examples 2 and 3, i.e., a circle with initial speed 20 and no initial displacements. Again, we use a stepwidth of $\Delta t = 0.001$. To achieve energy conservation, we use the method of Gonzalez [22]. The results for the rotation angle and angular speed comparing the standard nonlinear scheme with the corotational nonlinear schemes are given in Figure 6.15. Comparing with the linear

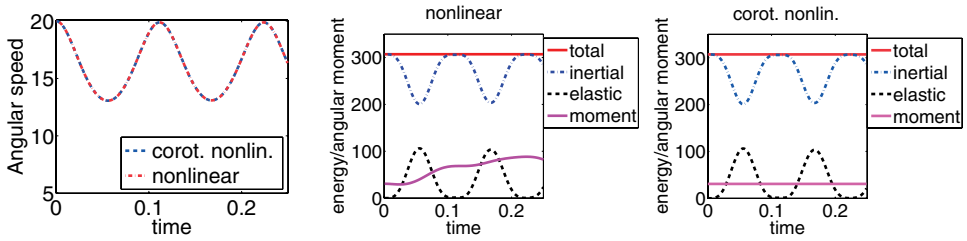


FIG. 6.15. *Example* 5. *Left*: *angular speed. Middle and right*: *energies and angular momenta.*

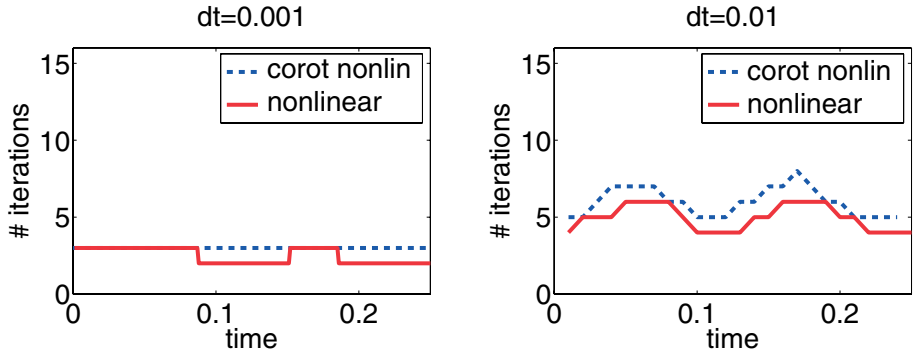


FIG. 6.16. *Example* 5. *Iteration numbers.*

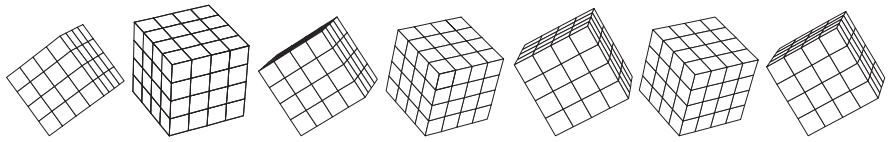


FIG. 6.17. *Example* 6. *Deformed mesh at time* $t = 0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3$.

Saint Venant–Kirchhoff material model used in Example 2, we see that the body is artificially hardened by the linearization in the material law. Good agreement of the angular speed is achieved for the two schemes. This can also be observed in the energy plots. However, in contrast to our new corotational scheme, the standard nonlinear scheme does not preserve the angular momentum; see Figure 6.15. The number of iterations per timestep is compared in Figure 6.16 for stepwidths $\Delta t = 0.001$ and $\Delta t = 0.01$. Here, both schemes yield similar iteration numbers with the standard nonlinear scheme performing somewhat better.

6.7. Example 6: Rotation of a 3D cube. In our last example, we apply our algorithm to a 3D problem. We model a rotating cube with constant rotation axis $(1, 1, 1)^\top$. We note that for this kind of rotation, R can be represented in terms of just one rotation angle. Moreover, the angular momentum is exactly preserved in this example as the algorithm preserves the norm of the angular momentum. The domain is given by $[0, 1] \times [0, 1] \times [0, 1]$ and the material parameters by $E = 1.62 \cdot 10^4$, $\nu = 0.2$, and $\rho = 1.0$. We use $\Delta t = 0.001$, an initial angular speed of 20, and zero initial effective displacement. The deformed mesh at various timesteps is presented in Figure 6.17. As the cube is not rotationally symmetric with respect to the rotation

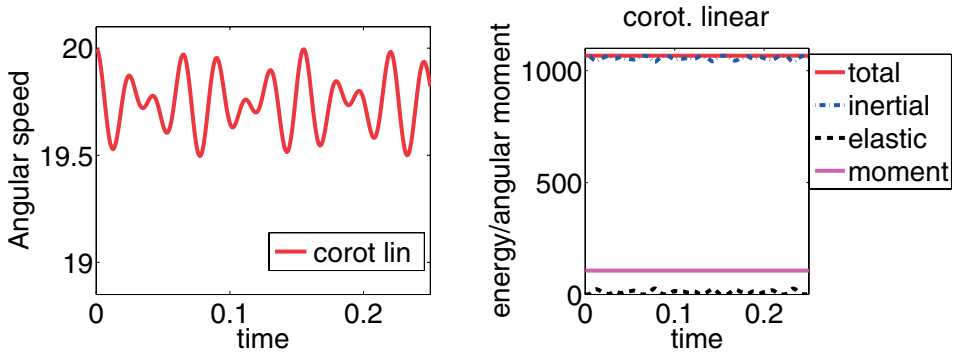


FIG. 6.18. *Example* 6. *Left*: *angular speed. Right*: *energy and angular momentum.*

axis, we do not get the periodic behavior as in the 2D examples; see Figure 6.18. However, the energy and angular momentum are still preserved.

REFERENCES

[1] J. H. ARGYRIS, *An excursion into large rotations*, Comput. Methods Appl. Mech. Engrg., 32 (1982), pp. 85–155.

[2] K.-J. BATHE, *Finite Element Procedures in Engineering Analysis*, Prentice–Hall, Englewood Cliffs, NJ, 1982.

[3] K.-J. BATHE, E. RAMM, AND E. L. WILSON, *Finite element formulations for large deformation dynamic analysis*, Internat. J. Numer. Methods Engrg., 7 (1973), pp. 255–271.

[4] J.-M. BATTINI AND C. PACOSTE, *Corotational beam elements with warping effects in instability problems*, Comput. Methods Appl. Mech. Engrg., 191 (2002), pp. 1755–1789.

[5] T. BELYTSCHKO, *An overview of semidiscretization and time integration operators*, in Computational Methods for Transient Analysis, T. Belytschko and T. J. R. Hughes, eds., North–Holland, Amsterdam, 1983, pp. 1–66.

[6] T. BELYTSCHKO AND L.W. GLAUM, *Application of higher order corotational stretch theories to nonlinear finite element analysis*, Comput. & Structures, 11 (1979), pp. 175–182.

[7] T. BELYTSCHKO AND B. J. HSIEH, *Nonlinear transient finite element analysis with convected coordinates*, Internat. J. Numer. Methods Engrg., 7 (1973), pp. 255–271.

[8] P. G. BERGAN AND G. HORRIGMOE, *Incremental variational principles and finite element models for nonlinear problems*, Comput. Methods Appl. Mech. Engrg., 7 (1976), pp. 201–217.

[9] P. G. BERGAN AND M. K. NYGARD, *Finite elements with increased freedom in choosing shape functions*, Internat. J. Numer. Methods Engrg., 20 (1984), pp. 643–664.

[10] M. A. BIOT, *The Mechanics of Incremental Deformations*, McGraw–Hill, New York, 1962.

[11] P. G. CIARLET, *Mathematical Elasticity, Volume* I, North–Holland, Amsterdam, 1988.

[12] M. A. CRISFIELD, *A consistent corotational formulation for nonlinear three-dimensional beam elements*, Comput. Methods Appl. Mech. Engrg., 81 (1990), pp. 131–150.

[13] M. A. CRISFIELD, *Nonlinear Finite Element Analysis of Solids and Structures, Vol.* 2, *Advanced Topics*, John Wiley, New York, 1997.

[14] M. A. CRISFIELD AND J. SHI, *A corotational element/time-integration strategy for non-linear dynamics*, Internat. J. Numer. Methods Engrg., 37 (1994), pp. 1897–1913.

[15] C. FARHAT, P. GEUZAINE, AND G. BROWN, *Application of a three-field nonlinear fluid-structure formulation to the prediction of the aeroelastic parameters of an F-16 fighter*, Comput. & Fluids, 32 (2003), pp. 3–29.

[16] C. FARHAT, K. PIERSON, AND M. LESOINNE, *The second generation of FETI methods and their application to the parallel solution of large-scale linear and geometrically nonlinear structural analysis problems*, Comput. Methods Appl. Mech. Engrg., 184 (2000), pp. 333–374.

[17] C. FELIPPA, *Recent advances in finite element templates*, in Computational Mechanics for the Twenty-First Century, B. H. V. Tropping, ed., Saxe-Coburn Publications, Edinburgh, 2000, pp. 71–98.

[18] C. FELIPPA, *A study of optimal membrane triangles with drilling freedoms*, Comput. Methods Appl. Mech. Engrg., 192 (2003), pp. 2125–2168.

[19] C. A. Felippa and B. Haugen, *A unified formulation of small-strain corotational finite elements:* I. *Theory*, Comput. Methods Appl. Mech. Engrg., 194 (2005), pp. 2285–2335.
[20] B. Fraejis de Veubeke, *The dynamics of flexible bodies*, Internat. J. Engrg. Sci., 14 (1976), pp. 895–913.
[21] U. Galvanetto and M. A. Crisfield, *An energy-conserving corotational procedure for the dynamics of planar beam structures*, Internat. J. Numer. Methods Engrg., 39 (1996), pp. 2265–2282.
[22] O. Gonzalez, *Time integration and discrete Hamiltonian systems*, J. Nonlinear Sci., 6 (1996), pp. 449–468.
[23] C. Grandmont, Y. Maday, and P. Métier, *Modeling and analysis of an elastic problem in large displacements and small strains*, J. Elasticity, 87 (2007), pp. 29–72.
[24] C. Grandmont, Y. Maday, and P. Metier, *Existence of a solution for an unsteady elasticity problem in large displacement and small perturbation*, C. R. Acad. Sci. Paris Sér. I Math., 334 (2002), pp. 521–526.
[25] T. J. R. Hughes and J. Winget, *Finite rotation effects in numerical integration of rate constitutive equations arising in large deformation analysis*, Internat. J. Numer. Methods Engrg., 15 (1980), pp. 1862–1867.
[26] B. Nour-Omid and C. C. Rankin, *Finite rotation analysis and consistent linearizations using projectors*, Comput. Methods Appl. Mech. Engrg., 93 (1991), pp. 353–384.
[27] M. K. Nygard and P. G. Bergan, *Advances in treating large rotations for nonlinear problems*, in State of the Art Surveys on Computational Mechanics, A. K. Noor and J. T. Oden, eds., ASME, New York, 1989, pp. 305–332.
[28] C. Pacoste, *Corotational flat facet triangular elements for shell instability analysis*, Comput. Methods Appl. Mech. Engrg., 156 (1998), pp. 75–110.
[29] J. M. Pajot and K. Maute, *Analytical sensitivity analysis of geometrically nonlinear structures based on the corotational finite element method*, Finite Elem. Anal. Des., 42 (2006), pp. 900–913.
[30] C. Pedersen, *Topology optimization of* 2*D-frame structures with path dependencies*, Internat. J. Numer. Methods Engrg., 57 (2003), pp. 1471–1501.
[31] E. Perente and L. E. Vaz, *On evaluation of shape sensitivities for non-linear critical loads*, Internat. J. Numer. Methods Engrg., 56 (2003), pp. 809–846.
[32] C. C. Rankin and F. A. Brognan, *An element-independent corotational procedure for the treatment of large rotations*, ASME J. Pressure Vessel Technology, 108 (1986), pp. 165–174.
[33] C. C. Rankin and B. Nour-Omid, *The use of projectors to improve finite element performance*, Comput. & Structures, 30 (1988), pp. 257–267.
[34] J. C. Simo, *A finite strain beam formulation. Part* I: *The three-dimensional dynamic problem*, Comput. Methods Appl. Mech. Engrg., 49 (1985), pp. 55–70.
[35] J. C. Simo and N. Tarnow, *The discrete energy-momentum method: Conserving algorithms for nonlinear elastodynamics*, Z. Angew. Math. Phys., 43 (1992), pp. 757–792.
[36] M. L. Szwabowicz, *Variational formulation in the geometrically nonlinear thin elastic shell theory*, Internat. J. Engrg. Sci., 22 (1986), pp. 1161–1175.
[37] G. A. Wempner, *Finite elements, finite rotations and small strains of flexible shells*, Internat. J. Solids Structures, 5 (1969), pp. 117–153.

SIAM J. NUMER. ANAL.
Vol. 46, No. 4, pp. 1867–1891
© 2008 Society for Industrial and Applied Mathematics

NONCONFORMING, ANISOTROPIC, RECTANGULAR FINITE ELEMENTS OF ARBITRARY ORDER FOR THE STOKES PROBLEM*

THOMAS APEL† AND GUNAR MATTHIES‡

Abstract. The numerical solution of the Stokes problem is analyzed for four families of rectangular, nonconforming finite elements of arbitrary order. While two of them become unstable as the maximum aspect ratio of the finite elements increases, the other two remain stable. A complete convergence analysis is given for these two families on two-level triangulations with isotropic rectangular macroelements and certain anisotropic refinement strategies within the macros.

Key words. nonconforming finite elements, anisotropic meshes, consistency error, inf-sup condition

AMS subject classifications. 65N30, 65N15

DOI. 10.1137/060666652

1. Introduction.

We consider the stationary Stokes problem

$$\begin{aligned} -\triangle \boldsymbol{u} + \nabla p &= \boldsymbol{f} && \text{in } \Omega, \\ \operatorname{div} \boldsymbol{u} &= 0 && \text{in } \Omega, \\ \boldsymbol{u} &= \boldsymbol{0} && \text{on } \partial\Omega, \end{aligned} \tag{1.1}$$

in a two-dimensional domain Ω. For the discretization of (1.1), one has to make the fundamental decision of choosing either a continuous or a discontinuous pressure approximation. We advocate the latter due to a better mass conservation on the element level. Furthermore, coupled multigrid solvers work in general better for discontinuous pressure spaces [16, 17]. Moreover, one has to distinguish between conforming and nonconforming velocity spaces. Since the degrees of freedom in two-dimensional nonconforming methods are edge-oriented, these methods are advantageous for parallel implementation due to missing cross-point communication.

Well-known examples for low order nonconforming discretizations of (1.1) can be found in [8, 11, 13]. Families of nonconforming triangular elements of arbitrarily high polynomial order were given in [22, 23] where only in [22] the application to the Stokes problem was considered. Three families of quadrilateral elements and a family of hexahedral elements were derived recently in [21]. They generalize the well-known rotated Q_1 element [24].

The spaces for approximating the velocity and the pressure in Stokes and Navier–Stokes problems cannot be chosen independently but have to satisfy the discrete inf-sup stability condition. The solution of the Navier–Stokes problem includes in general layer phenomena. To capture this anisotropic behavior, one usually employs adequately refined anisotropic meshes. The small element size in one direction is used to compensate for large derivatives in this direction. It is essential in such

*Received by the editors August 2, 2006; accepted for publication (in revised form) September 7, 2007; published electronically April 18, 2008.
http://www.siam.org/journals/sinum/46-4/66665.html
†Institut für Mathematik und Bauinformatik, Universität der Bundeswehr München, 85577 Neubiberg, Germany (thomas.apel@unibw.de).
‡Fakultät für Mathematik, Ruhr-Universität Bochum, 44780 Bochum, Germany (gunar.matthies@ruhr-uni-bochum.de).

applications that the constant in the discrete inf-sup condition does not degenerate for an increasing aspect ratio of the elements. We refer the reader to [5] for the state of the art in about 2002. However, the above-mentioned families were until now not investigated concerning their behavior on meshes which contain elements with a large aspect ratio. In this paper, we focus on anisotropic meshes and analyze the properties of the three quadrilateral families of [21] and a further new family which generalizes the first order element developed in [4]. All considered elements are parametric. The first mentioned three families generalize the parametric rotated Q_1 element, which itself is not inf-sup stable on anisotropic meshes. Although we investigate families of arbitrary polynomial degree, we focus on an h-version of the finite element method, which means that the dependence of the constants on the polynomial degree is not elaborated.

The paper is organized as follows. We start with basic notation in section 2. All four considered families will be introduced in section 3. The numerical results which are obtained by using these families on isotropic meshes differ only slightly. However, a numerical study on selected anisotropic meshes shows that for two families the constant in the inf-sup condition degenerates to zero as the aspect ratio increases. We succeeded in analyzing the remaining two families and present interpolation error estimates, the proof of the inf-sup condition, and an estimate of the consistency error in sections 5, 6, and 7, respectively. Since the proofs are technical, we summarize our results in section 4.

2. Notation. Let Ω be a bounded domain in $\mathbb{R}^2$. For a one- or two-dimensional domain $G \subset \Omega$ we will use the Sobolev spaces $W^{m,p}(G)$ with norm $\|\cdot\|_{m,p,G}$ and semi-norm $|\cdot|_{m,p,G}$, where the last index will be omitted if $G = \Omega$. We write in the case $p = 2$, as usual, $H^m(G) := W^{m,2}(G)$, and we omit the index p in the norms. Note further that $L^p(G) := W^{0,p}(G)$. The vector-valued version of all spaces will be denoted by the corresponding bold-face letter. The inner product in $L^2(G)$ and its vector-valued and tensor-valued versions will be denoted by $(\cdot,\cdot)_G$. For a one-dimensional manifold $F \subset \partial G$, the inner product in $L^2(F)$ and its vector-valued versions are written as $\langle\cdot,\cdot\rangle_F$.

Let $\boldsymbol{X} := \boldsymbol{H}^1_0(\Omega)$ and $M := L^2_0(\Omega) = \{v \in L^2(\Omega) : (v,1) = 0\}$. A weak formulation of (1.1) reads as follows: Find $(\boldsymbol{u}, p) \in \boldsymbol{X} \times M$ such that

$$\begin{aligned} a(\boldsymbol{u},\boldsymbol{v}) - b(\boldsymbol{v},p) &= (\boldsymbol{f},\boldsymbol{v}) \qquad \forall \boldsymbol{v} \in \boldsymbol{X}, \\ b(\boldsymbol{u},q) &= 0 \qquad \forall q \in M, \end{aligned} \tag{2.1}$$

where

$$a(\boldsymbol{u},\boldsymbol{v}) := (\nabla\boldsymbol{u}, \nabla\boldsymbol{v}), \qquad b(\boldsymbol{v},q) := (\operatorname{div}\boldsymbol{v}, q).$$

Let $\{\mathcal{T}_h\}$ be a family of admissible triangulations of Ω into rectangular elements. For a domain $G \subset \mathbb{R}^d$, $d = 1, 2$, let $P_k(G)$ denote the space of polynomials of degree less than or equal to k. For discretizing the pressure, we will use the space

$$M_h := \{p \in L^2_0(\Omega) : p|_K \in P_{r-1}(K)\, \forall K \in \mathcal{T}_h\}$$

of generally discontinuous, piecewise polynomials of degree less than or equal to $r-1$. The velocity will be approximated by a nonconforming finite element space $\boldsymbol{X}_h := X_h^2$. The detailed definition for several choices of X_h will be given in section 3.

Since the bilinear forms a and b are not defined for general functions from $\boldsymbol{X}_h$, we introduce the bilinear forms a_h and b_h by

$$a_h(\boldsymbol{u}_h, \boldsymbol{v}_h) := \sum_{K \in \mathcal{T}_h} (\nabla \boldsymbol{u}_h, \nabla \boldsymbol{v}_h)_K, \qquad b_h(\boldsymbol{v}_h, q_h) := \sum_{K \in \mathcal{T}_h} (\operatorname{div} \boldsymbol{v}_h, q_h)_K,$$

which coincide for regular arguments with a and b, i.e., $a_h(\boldsymbol{u}, \boldsymbol{v}) = a(\boldsymbol{u}, \boldsymbol{v})$ for $\boldsymbol{u}, \boldsymbol{v} \in \boldsymbol{X}$ and $b_h(\boldsymbol{v}, q) = b(\boldsymbol{v}, q)$ for $\boldsymbol{v} \in \boldsymbol{X}$, $q \in M$. Furthermore, we define

$$|v_h|_{1,h} := \left(\sum_{K \in \mathcal{T}_h} |v_h|_{1,K}^2 \right)^{1/2},$$

which will be a norm on X_h; see (3.2) for the definition of X_h. The discretized Stokes problem reads as follows: Find $(\boldsymbol{u}_h, p_h) \in \boldsymbol{X}_h \times M_h$ such that

$$\begin{aligned} a_h(\boldsymbol{u}_h, \boldsymbol{v}_h) - b_h(\boldsymbol{v}_h, p_h) &= (\boldsymbol{f}, \boldsymbol{v}_h) \qquad && \forall \boldsymbol{v}_h \in \boldsymbol{X}_h, \\ b_h(\boldsymbol{u}_h, q_h) &= 0 && \forall q_h \in M_h. \end{aligned} \tag{2.2}$$

We define for the spaces $\boldsymbol{X}_h$ and M_h the quantity

$$\beta_h := \inf_{q_h \in M_h} \sup_{\boldsymbol{v}_h \in \boldsymbol{X}_h} \frac{b_h(\boldsymbol{v}_h, q_h)}{|\boldsymbol{v}|_{1,h} \, \|q_h\|_0}. \tag{2.3}$$

The positivity of β_h ensures that (2.2) is uniquely solvable. Moreover, the inverse of β_h enters the estimate of the discretization error; see [8, sect. II.2.2]. Since we are interested in anisotropic discretizations, the quantity β_h should be bounded away from zero not only for h tending to zero but also for an increasing aspect ratio. We call this the inf-sup condition, and we shall see that it is not generally satisfied.

Within this paper, we will frequently use the notation $x \lesssim y$ for the case that $x \leq C\, y$ holds true where C is a positive constant which is independent of the mesh parameters, i.e., independent of the mesh size h and the aspect ratio.

3. Families of finite elements. For a nonnegative integer k, we denote by L_k the kth one-dimensional Legendre polynomial normalized such that $L_k(1) = 1$. Let

$$\delta_k := \int_{-1}^{1} L_k^2(t)\, dt = \frac{2}{2k+1}, \qquad k \geq 0.$$

The Legendre polynomials are orthogonal, i.e.,

$$\int_{-1}^{1} L_i(t) L_j(t)\, dt = \delta_{ij} \delta_i,$$

where δ_{ij} denotes the Kronecker delta. The integrated Legendre polynomials $\widehat{L}_k$ which are defined by

$$\widehat{L}_k(t) := \int_{-1}^{t} L_k(s)\, ds$$

are related to the Legendre polynomials by

$$\widehat{L}_k(t) = \begin{cases} \frac{\delta_k}{2}\big(L_{k+1}(t) - L_{k-1}(t)\big), & k \geq 1, \\ L_1(t) + L_0(t), & k = 0. \end{cases} \tag{3.1}$$

For defining the nonconforming finite element space, we will use the triple notation in the sense of Ciarlet [9]. Let $(\widehat{K}, \widehat{V}, \widehat{\mathcal{N}})$ be a finite element on the reference cell $\widehat{K}$. All quantities which are connected to the reference element will be indicated by a hat. Note that as an exception both the Legendre polynomials L_k and the integrated Legendre polynomials $\widehat{L}_k$ are always defined on reference interval $[-1, 1]$. Let $F_K := \widehat{K} \to K$ be the affine transformation which maps $\widehat{K}$ onto K. Then, the function space $V(K)$ on K is defined by

$$V(K) := \big\{ v = \hat{v} \circ F_K^{-1} \,:\, \hat{v} \in \widehat{V} \big\}.$$

Using the notation $\mathcal{S}_h^i$ and $\mathcal{S}_h^b$ for the set of interior and boundary edges of $\mathcal{T}_h$, the global scalar finite element space X_h is given by

$$\begin{aligned} X_h := \Big\{ & v \in L^2(\Omega) \,:\, v|_K \in V(K)\, \forall K \in \mathcal{T}_h,\ \langle q, [v]_E \rangle_E = 0\, \forall q \in P_{r-1}(E), \forall E \in \mathcal{S}_h^i, \\ & \langle q, v \rangle_E = 0\, \forall q \in P_{r-1}(E), \forall E \in \mathcal{S}_h^b \Big\}. \end{aligned} \tag{3.2}$$

As usual, $[v]_E$ denotes the jump of the function v across the edge E. The vector-valued finite element space $\boldsymbol{X}_h$ consists of X_h in each component.

Now, we will define several finite elements on the reference cell $\widehat{K} = (-1, 1)^2$. We introduce the nodal functionals

$$\widehat{N}_{ij}(\hat{v}) := \frac{1}{\delta_i \delta_j} \int_{-1}^{1} \int_{-1}^{1} \hat{v}(\hat{x}_1, \hat{x}_2)\, L_i(\hat{x}_1) L_j(\hat{x}_2)\, d\hat{x}_2\, d\hat{x}_1, \qquad i, j \geq 0,$$

and, for nonnegative integers k,

$$\widehat{N}_{k,-}(\hat{v}) := \frac{1}{\delta_k} \int_{-1}^{1} \hat{v}(\hat{x}_1, -1)\, L_k(\hat{x}_1)\, d\hat{x}_1, \qquad \widehat{N}_{k,+}(\hat{v}) := \frac{1}{\delta_k} \int_{-1}^{1} \hat{v}(\hat{x}_1, +1)\, L_k(\hat{x}_1)\, d\hat{x}_1,$$

$$\widehat{N}_{-,k}(\hat{v}) := \frac{1}{\delta_k} \int_{-1}^{1} \hat{v}(-1, \hat{x}_2)\, L_k(\hat{x}_2)\, d\hat{x}_2, \qquad \widehat{N}_{+,k}(\hat{v}) := \frac{1}{\delta_k} \int_{-1}^{1} \hat{v}(+1, \hat{x}_2)\, L_k(\hat{x}_2)\, d\hat{x}_2.$$

For a fixed integer $r \geq 1$, the set $\widehat{\mathcal{N}}_r$ of nodal functionals is given by

$$\widehat{\mathcal{N}}_r := \big\{ \widehat{N}_{ij} \,:\, 0 \leq i + j \leq r - 2 \big\} \cup \big\{ \widehat{N}_{k,-}, \widehat{N}_{k,+}, \widehat{N}_{-,k}, \widehat{N}_{+,k} \,:\, k = 0, \ldots, r - 1 \big\}. \tag{3.3}$$

Now, we look for suitable function spaces $\widehat{V}_r \supset P_r(\widehat{K})$ with the property that the set $\widehat{\mathcal{N}}_r$ of nodal functionals is unisolvent with respect to the space $\widehat{V}_r$. The following four choices for $\widehat{V}_r$ fulfill these conditions. An illustration of these families for $r = 11$ is given in Figure 3.1 where the grey boxes indicate the space $P_r(\widehat{K})$, the dark boxes stand for further functions of form $L_i(\hat{x}_1) L_j(\hat{x}_2)$, and if a pair of bright boxes is connected by an arc, then only the difference of the involved functions belongs to the space $\widehat{V}_r$.

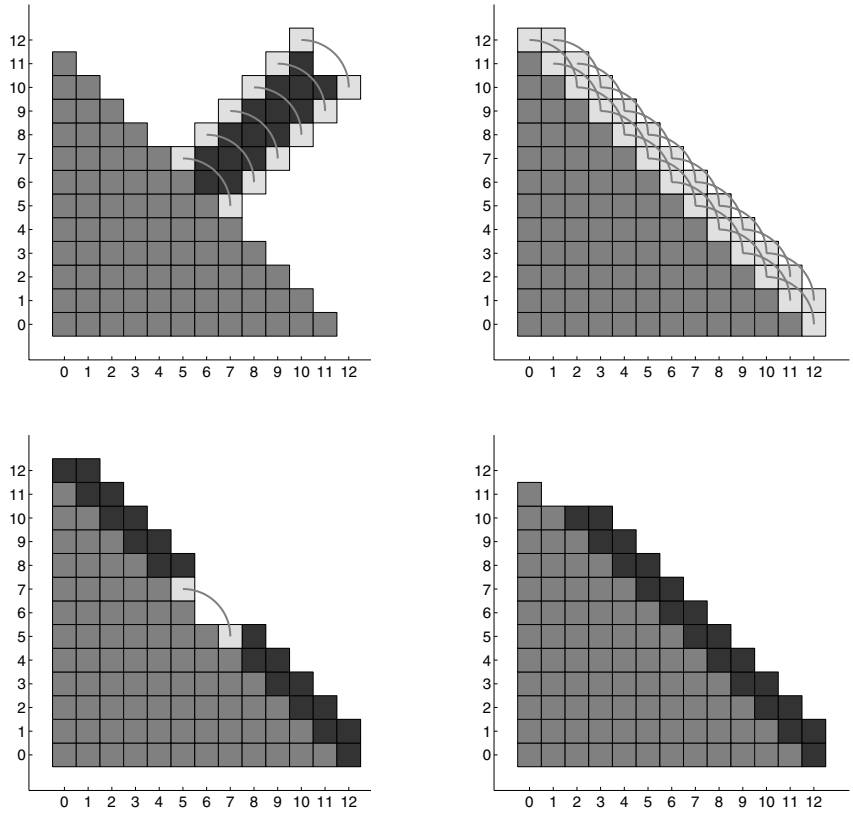


FIG. 3.1. *Illustration of the function spaces* $\widehat{V}_{11}^{\nu}$ *for* $\nu = 1, \ldots, 4$, *from top left to bottom right.*

A first choice was given in Example 5 in [14]. We set

$$\widehat{V}_r^1 := P_r(\widehat{K}) \oplus \widehat{R}_r \oplus \widehat{S}_r, \tag{3.4}$$

where

$$\widehat{R}_r := \begin{cases} \operatorname{span}\Big\{ L_{i+1}(\hat{x}_1)L_i(\hat{x}_2),\ L_i(\hat{x}_1)L_{i+1}(\hat{x}_2) \\ \qquad L_{i+2}(\hat{x}_1)L_i(\hat{x}_2) - L_i(\hat{x}_1)L_{i+2}(\hat{x}_2),\ :\ i = r/2 \Big\}, & r \text{ even}, \\ \operatorname{span}\Big\{ L_{i+2}(\hat{x}_1)L_i(\hat{x}_2) - L_i(\hat{x}_1)L_{i+2}(\hat{x}_2)\ :\ i = (r-1)/2 \Big\}, & r \text{ odd}, \end{cases}$$

and

$$\widehat{S}_r := \operatorname{span}\Big\{ L_i(\hat{x}_1)L_i(\hat{x}_2),\ L_{i+1}(\hat{x}_1)L_i(\hat{x}_2),\ L_i(\hat{x}_1)L_{i+1}(\hat{x}_2), \\ L_{i+2}(\hat{x}_1)L_i(\hat{x}_2) - L_i(\hat{x}_1)L_{i+2}(\hat{x}_2)\ :\ r/2 < i \le r-1 \Big\}.$$

A second variant which is given by

(3.5)
$$\widehat{V}_r^2 := P_r(\widehat{K}) \oplus \operatorname{span}\Big\{ L_{i+2}(\hat{x}_1)L_{r-1-i}(\hat{x}_2) - L_i(\hat{x}_1)L_{r+1-i}(\hat{x}_2)\ ,\ i=0,\dots,r-1 \Big\}$$
$$\oplus \operatorname{span}\Big\{ L_{i+2}(\hat{x}_1)L_{r-i}(\hat{x}_2) - L_i(\hat{x}_1)L_{r+2-i}(\hat{x}_2)\ ,\ i=1,\dots,r-1 \Big\}$$

was introduced in [21]. The following third version can also be found there:

(3.6)
$$\begin{aligned}\widehat{V}_r^3 :=& P_r(\widehat{K}) \\ &\oplus \operatorname{span}\Big\{ L_i(\hat{x}_1)L_{r+1-i}(\hat{x}_2),\ L_{r+1-i}(\hat{x}_1)L_i(\hat{x}_2)\ ,\ i=0,\dots,\lfloor r/2\rfloor-1 \Big\} \\ &\oplus \operatorname{span}\Big\{ L_i(\hat{x}_1)L_{r+2-i}(\hat{x}_2),\ L_{r+2-i}(\hat{x}_1)L_i(\hat{x}_2)\ ,\ i=1,\dots,\lfloor (r+1)/2\rfloor-1 \Big\} \\ &\oplus \operatorname{span}\Big\{ L_{\lfloor r/2\rfloor+2}(\hat{x}_1)L_{\lfloor r/2\rfloor}(\hat{x}_2) - L_{\lfloor r/2\rfloor}(\hat{x}_1)L_{\lfloor r/2\rfloor+2}(\hat{x}_2) \Big\},\end{aligned}$$

where $\lfloor s\rfloor$ is the largest integer which is less than or equal to s. Note that all these spaces generalize the parametric rotated Q_1 element, i.e., $\widehat{V}_1^1 = \widehat{V}_1^2 = \widehat{V}_1^3 = Q_1^{\mathrm{rot}}$. Finally, we introduce a new family,

(3.7)
$$\begin{aligned}\widehat{V}_r^4 := P_r(\widehat{K}) &\oplus \operatorname{span}\big\{ L_{r+1-i}(\hat{x}_1)L_i(\hat{x}_2),\ i=0,\dots,r-1 \big\} \\ &\oplus \operatorname{span}\big\{ L_{r+2-i}(\hat{x}_1)L_i(\hat{x}_2),\ i=1,\dots,r-1 \big\},\end{aligned}$$

which recovers for $r=1$ the space introduced in [4]. Note that this space is not symmetric with respect to $\hat{x}_1$ and $\hat{x}_2$. It is adapted to anisotropic elements which are elongated in the $\hat{x}_1$-direction. If the anisotropic element is stretched in the $\hat{x}_2$-direction, the roles of $\hat{x}_1$ and $\hat{x}_2$ must be exchanged in the definition of $\widehat{V}_r^4$. For isotropic elements one can use both variants or one of the families $\widehat{V}_r^\nu$, $\nu=1,2,3$.

All four families are unisolvent with respect to the set $\widehat{\mathcal{N}}_r$ of nodal functionals defined in (3.3). The proof for $\widehat{V}_r^1$ can be found in [14]. The unisolvence results for $\widehat{V}_r^2$ and $\widehat{V}_r^3$ are given in [21]. The proof for the family $\widehat{V}_r^4$ will be given in the appendix; see Lemma A.1 on page 1888.

Note that $\widehat{V}_r^\nu \subset \widehat{V}_{r+1}^\nu$, $\nu=1,2,3,4$, holds true, but the resulting nonconforming finite element spaces X_h are nonnested since a higher value of r also requires a higher order of consistency; see the definition (3.2) of X_h.

In order to assess these families of elements, we present the results of numerical tests performed with the finite element package MooNMD [18]. Table 3.1 shows that the results for different elements of the same order are comparable on isotropic meshes. The presented results were obtained on a uniform 64×64-decomposition of the unit square. The right-hand side and the boundary conditions were chosen such that

$$u(x,y) = \begin{pmatrix} \sin(x)\sin(y) \\ \cos(x)\cos(y) \end{pmatrix}, \qquad p(x,y) = 2\cos(x)\sin(y) - p_0$$

TABLE 3.1
Results for second and third order elements on isotropic meshes.

	$\lvert \mathbf{u} - \mathbf{u}_h \rvert_{1,h}$		$\lVert \mathbf{u} - \mathbf{u}_h \rVert_0$		$\lVert p - p_h \rVert_0$	
Pair	Error	Order	Error	Order	Error	Order
V_2^1/P_1^{disc}	2.475-05	1.98	4.732-08	2.99	2.147-05	2.00
V_2^2/P_1^{disc}	2.679-05	1.99	5.637-08	2.99	2.145-05	2.00
V_2^3/P_1^{disc}	2.321-05	2.01	4.643-08	3.01	2.146-05	2.00
V_2^4/P_1^{disc}	2.542-05	2.00	5.955-08	2.99	2.341-05	2.00
V_3^1/P_2^{disc}	6.990-08	2.99	1.117-10	3.99	6.655-08	3.00
V_3^2/P_2^{disc}	6.794-08	3.00	1.158-10	4.00	6.650-08	3.00
V_3^3/P_2^{disc}	6.869-08	3.02	1.151-10	4.02	6.807-08	3.03
V_3^4/P_2^{disc}	6.642-08	2.99	1.095-10	3.99	6.681-08	3.00

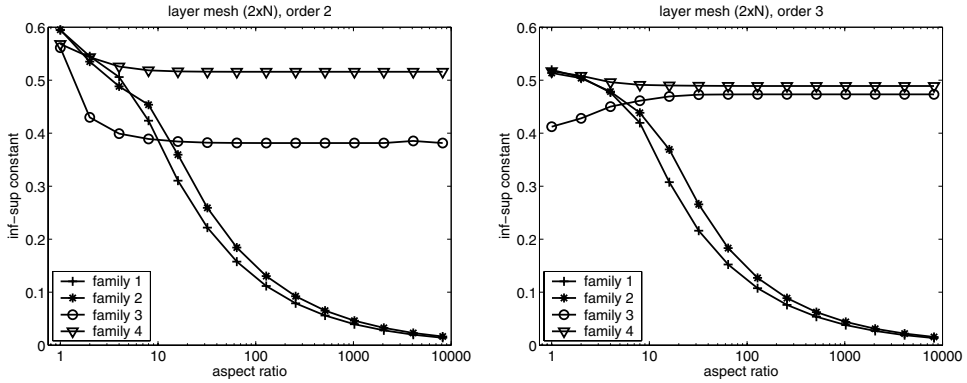


FIG. 3.2. *Plot of β_h against the aspect ratio for a family of layer meshes. Left: $r = 2$. Right: $r = 3$.*

is the given solution of (2.1), where the constant p_0 is determined by $p \in L_0^2(\Omega)$. The given orders were calculated from the data on the 64×64 mesh and the next coarser 32×32 mesh. All convergence rates are optimal.

We display in Figure 3.2 the inf-sup constants β_h for the finite element spaces $(\boldsymbol{X}_h, M_h)$ defined via $\widehat{V}_r^\nu$, $r = 2, 3$, $\nu = 1, 2, 3, 4$, on a family of meshes with increasing aspect ratio. These rectangular $2 \times N$ meshes are defined by their nodes (x_i, y_j), $x_i = i/2$, $i = 0, 1, 2$, $y_j = j/N$, $j = 0, \ldots, N$, yielding the aspect ratio $N/2$ for all elements. We see that the inf-sup constant β_h tends to zero for the families with $\nu = 1, 2$. This means that these families are not suited when anisotropic meshes are employed. For $\nu = 3, 4$, however, the inf-sup constant β_h remains bounded away from zero on a quite large level. Therefore, we will investigate these two families further.

Since we will not focus on a p-version of the finite element method, the dependence of all constants on the parameter r is not elaborated within this paper. For $\nu = 3, 4$, we define by

$$\widehat{N}(\widehat{I}^\nu \hat{v}) = \widehat{N}(\hat{v}) \qquad \forall \widehat{N} \in \widehat{\mathcal{N}}$$

the canonical interpolation operators $\widehat{I}^\nu$ which map onto $\widehat{V}^\nu$. Note that these interpolation operators are well defined for all functions which belong to $W^{1,1}(\widehat{K})$. The global interpolant I_h^ν is defined elementwise as usual via $\big(I_h^\nu u\big)|_K = \big(\widehat{I}^\nu (u|_K \circ F_K)\big) \circ F_K^{-1}$. Since the edge functionals from $\widehat{\mathcal{N}}$ are transformed into functionals which are

associated with the edges of $\mathcal{T}_h$, the elementwise defined interpolation operators I_h^ν, $\nu = 3, 4$, map into the corresponding nonconforming space X_h. The vector-valued interpolation operators $\boldsymbol{I}_h^\nu : \boldsymbol{X} \to \boldsymbol{X}_h$, where $\boldsymbol{X}_h$ is based on $\widehat{V}^\nu$, $\nu = 3, 4$, are defined componentwise.

4. Main results. We will investigate in the remaining part of this paper the families 3 and 4, where we succeed to prove, on special meshes, the fulfillment of the inf-sup condition and estimates for the interpolation error and the consistency error. These three properties result in the final discretization error estimate of order r. Since the proofs are quite lengthy, we summarize the results in this section. The proofs of Theorems 4.1–4.4 are given in sections 5–7.

The first ingredient is an anisotropic local interpolation error estimate on an element K which is a rectangle with edge sizes $h_{K,1}$ and $h_{K,2}$. We assume that K is elongated in the x_1-direction; i.e., $h_{K,1} \geq h_{K,2}$. We abbreviate $\partial_1 := \partial/\partial x_1$, $\partial_2 := \partial/\partial x_2$, $\hat{\partial}_1 := \partial/\partial \hat{x}_1$, $\hat{\partial}_2 := \partial/\partial \hat{x}_2$, and we use the multi-index notation with $\alpha = (\alpha_1, \alpha_2)$, $\alpha_1, \alpha_2 \geq 0$, $|\alpha| = \alpha_1 + \alpha_2$, $D^\alpha = \partial_1^{\alpha_1} \partial_2^{\alpha_2}$, and $h_K^\alpha = h_{K,1}^{\alpha_1} h_{K,2}^{\alpha_2}$.

THEOREM 4.1 (interpolation of smooth functions). *Let $u \in W^{\ell,p}(K)$, where $\ell \in \mathbb{N}$, $2 \leq \ell \leq r+1$, $p \in [1, \infty]$. Fix $m \in \{0, 1\}$ and $q \in [1, \infty]$ such that $W^{\ell-m,p}(K) \hookrightarrow L^q(K)$. Then, the anisotropic interpolation error estimate*

$$|u - I_h^\nu u|_{m,q,K} \lesssim |K|^{1/q-1/p} \sum_{|\alpha|=\ell-m} h_K^\alpha |D^\alpha u|_{m,p,K}, \quad \nu = 3, 4,$$

holds true.

Although the interpolation operators I_h^ν, $\nu = 3, 4$, are well defined for functions from $W^{1,1}(K)$, the anisotropic interpolation error estimate holds only for functions from $W^{2,1}(K)$ if $r \geq 2$. The reason is that we need traces of derivatives in the proof; see also Example 5.5. Note that for $r = 1$ and $\ell = 1$, the anisotropic interpolation error estimate holds for I_h^4 (see [3, Corollary 1]) but not for I_h^3 (see [6]), since the parametric version is considered here.

Interpolation error estimates for regular functions are needed for the estimation of the best approximation error. However, for the proof of the fulfillment of the inf-sup condition, one needs an interpolation estimate for H^1-functions. Fortunately, the regularity assumption $u \in W^{2,1}(K)$ in Theorem 4.1 can be weakened if functions which vanish on the short sides of K are considered. To this end, we introduce the function space

$$W_E^{1,p}(K) := \big\{ v \in W^{1,p}(K) \,:\, v|_{E_{-,*}} = v|_{E_{+,*}} = 0 \big\}, \tag{4.1}$$

where we denote by $E_{-,*}$ and $E_{+,*}$ the edges on the left-hand and right-hand sides of the element K, respectively; see Figure 4.1. The anisotropic interpolation error estimate on $W_E^{1,p}(K)$ is stated in the following theorem.

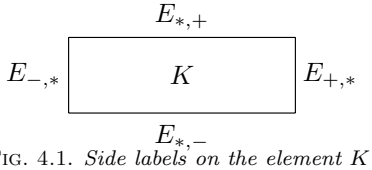


FIG. 4.1. *Side labels on the element K.*

THEOREM 4.2 (interpolation of nonsmooth functions). *Let $u \in W_E^{1,p}(K)$, $p \in [1, \infty]$. Fix $m \in \{0, 1\}$ and $q \in [1, \infty]$ such that $W^{1-m,p}(K) \hookrightarrow L^q(K)$. Then, the*

anisotropic interpolation error estimate

$$|u - I_h^\nu u|_{m,q,K} \lesssim |K|^{1/q-1/p} \sum_{|\alpha|=1-m} h^\alpha |D^\alpha u|_{m,p,K}, \quad \nu = 3, 4,$$

holds true.

To be able to use this result for checking the inf-sup condition, we employ the macroelement technique from [7]; see also [13, sect. II.1.4]. To this end, let $\mathcal{T}_H$ be an admissible macrotriangulation of Ω into isotropic rectangular macroelements Q:

$$\overline{\Omega} = \bigcup_{Q \in \mathcal{T}_H} \overline{Q}.$$

The finally used triangulation $\mathcal{T}_h$ is obtained from $\mathcal{T}_H$ by applying local refinement strategies to the macroelements Q in such a way that $\mathcal{T}_h$ is admissible. In particular, here we investigate boundary layer and corner patches which are defined in what follows. Other patches can be included provided a local inf-sup condition, as given by Lemmas 6.2 and 6.3, holds.

Let $Q = (x_-, x_+) \times (y_-, y_+)$ be a patch, and let $\tau_x = \{(x_{i-1}, x_i), i = 1, \dots, n\}$ and $\tau_y = \{(y_{j-1}, y_j), j = 1, \dots, m\}$ be decompositions of (x_-, x_+) and (y_-, y_+) into subintervals, respectively. In this way, the tensor-product triangulation $\tau_x \times \tau_y$ of Q is induced. If τ_x is the trivial decomposition of (x_-, x_+) into one interval, then the obtained patch is called a boundary layer patch. If $m = n$, $x_0 = x_-$, $y_0 = y_-$, $x_i = x_- + (x_+ - x_-)\sigma^{n-i}$, $i = 1, \dots, n$, and $y_j = y_- + (y_+ - y_-)\sigma^{n-j}$, $j = 1, \dots, n$, with a parameter $\sigma \in (0, 1)$, then the resulting patch is called a corner patch. Typical boundary layer and corner patches are shown in Figure 4.2. Note that the trivial decomposition of Q into one cell is allowed. The constants σ and n may vary from patch to patch. It is required only that the global mesh $\mathcal{T}_h$ is admissible.

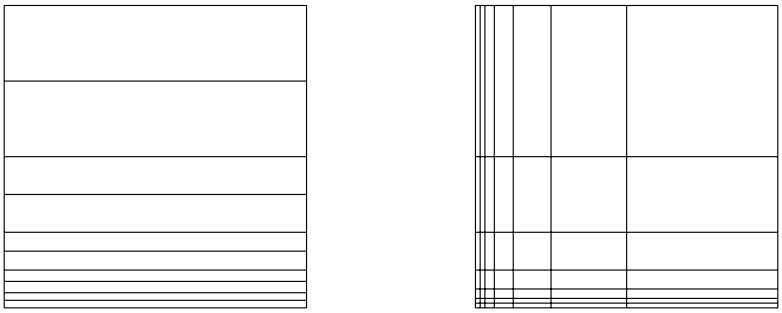


FIG. 4.2. *Typical boundary layer (left) and corner patches (right).*

THEOREM 4.3 (inf-sup condition). *Let $\mathcal{T}_h$ be a triangulation of Ω as described above and consider $\mathbf{X}_h$ based on $\widehat{V}_r^3$ or $\widehat{V}_r^4$ with $r \geq 3$. Then, there exists a constant $\beta > 0$ such that $\beta_h \geq \beta$. Like all constants in this paper, β is independent of the mesh parameter h and the maximal aspect ratio.*

Our method of proof works for all polynomial degrees $r \geq 3$ but not for $r = 2$. Since some numerical tests with $r = 2$ showed stability, the validity of a uniform inf-sup condition must be considered open for this case.

The consistency error is estimated in the following theorem.

THEOREM 4.4 (consistency error). *Let $\boldsymbol{u} \in \boldsymbol{X} \cap \boldsymbol{H}^{r+1}(\Omega)$ and $p \in M \cap H^r(\Omega)$ be the solution of (2.1). Then, the consistency error estimate*

$$\sup_{\boldsymbol{v}_h \in \boldsymbol{X}_h} \frac{|a_h(\boldsymbol{u},\boldsymbol{v}_h) - b_h(\boldsymbol{v}_h,p) - (\boldsymbol{f},\boldsymbol{v}_h)|}{|\boldsymbol{v}_h|_{1,h}}$$

$$\lesssim \left(\sum_{K\in\mathcal{T}_h} \sum_{|\alpha|=r} h_K^{2\alpha} \left(\|D^\alpha(\partial_1 u_1 - p)\|_{0,K}^2 + \|D^\alpha \partial_1 u_2\|_{0,K}^2 \right) \right)^{1/2}$$

$$+ \left(\sum_{K\in\mathcal{T}_h} h_{K,1}^2 \sum_{|\alpha|=r-1} h_K^{2\alpha} \|D^\alpha \boldsymbol{f}\|_{0,K}^2 \right)^{1/2}$$

holds true. As is the case above, the space $\mathbf{X}_h$ is based on $\widehat{V}_r^3$ or $\widehat{V}_r^4$.

Now, we can formulate our main result.

THEOREM 4.5 (discretization error estimate). *Let $(\boldsymbol{u},p) \in \boldsymbol{X} \times M$ and $(\boldsymbol{u}_h,p_h) \in \boldsymbol{X}_h \times M_h$ be the solutions of (2.1) and (2.2), respectively. The space $\mathbf{X}_h$ is based on $\widehat{V}_r^3$ or $\widehat{V}_r^4$ with $r \geq 3$. Then, the discretization error is estimated by*

$$|\boldsymbol{u}-\boldsymbol{u}_h|_{1,h} + \|p-p_h\|_0 \lesssim \left(\sum_{K\in\mathcal{T}_h} \sum_{|\alpha|=r} h_K^{2\alpha} \left(|D^\alpha \boldsymbol{u}|_{1,K}^2 + \|D^\alpha p\|_{0,K}^2 \right) \right)^{1/2}$$

$$+ \left(\sum_{K\in\mathcal{T}_h} h_{K,1}^2 \sum_{|\alpha|=r-1} h_K^{2\alpha} \|D^\alpha \boldsymbol{f}\|_{0,K}^2 \right)^{1/2}$$

provided that $\boldsymbol{u} \in \boldsymbol{H}^{r+1}(\Omega)$, $p \in H^r(\Omega)$, and $\boldsymbol{f} \in \boldsymbol{H}^{r-1}(\Omega)$.

Proof. Under the assumption that the uniform inf-sup condition $\beta_h \geq \beta > 0$ is valid, Proposition II.2.16 of [8] gives

$$|\boldsymbol{u}-\boldsymbol{u}_h|_{1,h} + \|p-p_h\|_0 \lesssim \inf_{\boldsymbol{v}_h \in \boldsymbol{X}_h} |\boldsymbol{u}-\boldsymbol{v}_h|_{1,h} + \inf_{q_h \in M_h} \|p-q_h\|_0$$

$$+ \sup_{\boldsymbol{v}_h \in \boldsymbol{X}_h} \frac{|a_h(\boldsymbol{u},\boldsymbol{v}_h) - b_h(\boldsymbol{v}_h,p) - (\boldsymbol{f},\boldsymbol{v}_h)|}{|\boldsymbol{v}_h|_{1,h}}.$$

The estimate for the error of the best approximation is obtained by combining the local estimates from Theorem 4.1. The consistency term is estimated in Theorem 4.4. The assertion is concluded by observing that parts of the right-hand side of the consistency error estimate can be bounded by best approximation terms. ∎

Note that the estimate is standard in the sense that we have a method of order r for $\boldsymbol{u} \in \boldsymbol{H}^{r+1}(\Omega)$, $p \in H^r(\Omega)$, and $\boldsymbol{f} \in \boldsymbol{H}^{r-1}$. The advantage is that we can compensate large derivatives of $\boldsymbol{u}$ and p in the x_2-direction by choosing small values for $h_{K,2}$.

5. Proof of the interpolation error estimates. Estimates for $\|u - I_h u\|_{0,q,K}$ and $\|\partial_1(u - I_h u)\|_{0,q,K}$ can be proved in the standard way since the element K is assumed to be stretched in the x_1-direction. Even nonsmooth functions $u \in W^{1,p}(K)$ can be treated by this approach.

LEMMA 5.1. *Let $u \in W^{\ell,p}(K)$, where $\ell \in \mathbb{N}$, $1 \le \ell \le r+1$, $p \in [1,\infty]$. Fix $m \in \{0,1\}$ and $q \in [1,\infty]$ such that $W^{\ell-m,p}(K) \hookrightarrow L^q(K)$. Then, the anisotropic interpolation error estimate*

$$\|\partial_1^m(u - I_h^\nu u)\|_{0,q,K} \lesssim |K|^{1/q-1/p} \sum_{|\alpha|=\ell-m} h_K^\alpha |D^\alpha u|_{m,p,K}, \quad \nu = 3,4,$$

holds true.

Proof. We prove the estimate first on the reference element $\widehat{K}$. Since the interpolation operator $\widehat{I}^\nu$ preserves polynomials of degree r and since $\widehat{I}^\nu : W^{\ell,p}(\widehat{K}) \to W^{m,q}(\widehat{K})$ is a bounded operator, we have for all $\hat{w} \in P_{\ell-1}(\widehat{K}) \subset P_r(\widehat{K}) \subset \widehat{V}_r^\nu$

$$\|\hat{\partial}_1^m(\hat{u} - \widehat{I}^\nu \hat{u})\|_{0,q,\widehat{K}} = \|\hat{\partial}_1^m(\hat{u} - \hat{w}) - \hat{\partial}_1^m \widehat{I}^\nu(\hat{u} - \hat{w})\|_{0,q,\widehat{K}} \lesssim \|\hat{u} - \hat{w}\|_{\ell,p,\widehat{K}}.$$

By using the Deny–Lions lemma, we obtain

$$\|\hat{\partial}_1^m(\hat{u} - \widehat{I}^\nu \hat{u})\|_{0,q,\widehat{K}} \lesssim |\hat{u}|_{\ell,p,\widehat{K}}.$$

The assertion is proved by transforming this estimate from $\widehat{K}$ to K by using $h_{K,2} \le h_{K,1}$. ∎

The challenge is the proof for $\|\partial_2(u - I_h^\nu u)\|_{0,q,K}$; see [2, Ex. 2.1]. Due to [2, Lemma 2.2], we succeed if we find a set $\widehat{\mathcal{F}}_r$ of $\dim \hat{\partial}_2 \widehat{V}_r^\nu$ linear functionals such that

$$\widehat{F} \in \big(W^{\ell-1,p}(\widehat{K})\big)' \quad \forall \widehat{F} \in \widehat{\mathcal{F}}_r, \tag{5.1}$$

$$\widehat{F}\big(\hat{\partial}_2(\hat{u} - \widehat{I}^\nu \hat{u})\big) = 0 \quad \forall \widehat{F} \in \widehat{\mathcal{F}}_r, \quad \forall \hat{u} \in W^{\ell,p}(\widehat{K}), \tag{5.2}$$

$$\hat{v} \in \hat{\partial}_2 \widehat{V}_r^\nu \quad \text{and} \quad \widehat{F}(\hat{v}) = 0 \quad \forall \widehat{F} \in \widehat{\mathcal{F}}_r \quad \Rightarrow \quad \hat{v} \equiv 0. \tag{5.3}$$

To this end, we define the set

$$\begin{aligned} \widehat{\mathcal{F}}_r := \big\{\widehat{F}_{i,*} \,:\, i = 0,\dots,r-1\big\} \cup \big\{\widehat{F}_{ij} \,:\, 0 \le i+j \le r-2\big\} \\ \cup \big\{\widehat{F}_{+,i}, \widehat{F}_{-,i} \,:\, i = 1,\dots,r-1\big\} \end{aligned} \tag{5.4}$$

with

$$\widehat{F}_{i,*}(\hat{v}) := \frac{1}{\delta_i \delta_0} \int_{-1}^1 \int_{-1}^1 L_i(\hat{x}_1) L_0(\hat{x}_2) \hat{v}(\hat{x}_1, \hat{x}_2)\, d\hat{x}_1\, d\hat{x}_2,$$

$$\widehat{F}_{ij}(\hat{v}) := \frac{2}{\delta_i \delta_j} \int_{-1}^1 \int_{-1}^1 L_i(\hat{x}_1) \widehat{L}_j(\hat{x}_2) \hat{v}(\hat{x}_1, \hat{x}_2)\, d\hat{x}_1\, d\hat{x}_2,$$

$$\widehat{F}_{+,i}(\hat{v}) := \frac{2}{\delta_i} \int_{-1}^1 \widehat{L}_i(\hat{x}_2) \hat{v}(+1, \hat{x}_2)\, d\hat{x}_2,$$

$$\widehat{F}_{-,i}(\hat{v}) := \frac{2}{\delta_i} \int_{-1}^1 \widehat{L}_i(\hat{x}_2) \hat{v}(-1, \hat{x}_2)\, d\hat{x}_2.$$

One easily checks that both the dimension of $\hat{\partial}_2 \widehat{V}_r^\nu$, $\nu = 3,4$, and the number of functionals in $\widehat{\mathcal{F}}_r$ are equal to $r^2/2 + 5r/2 - 2$. It remains to prove that the set $\widehat{\mathcal{F}}_r$

satisfies the properties (5.1)–(5.3). We start with the unisolvence property (5.3) which needs the most complex proof, done for $\widehat{V}_r^4$ and $\widehat{V}_r^3$ separately.

LEMMA 5.2. *The set $\widehat{\mathcal{F}}_r$ of functionals is unisolvent on $\hat{\partial}_2\widehat{V}_r^4$.*

Proof. The definition of $\widehat{V}_r^4$ gives immediately

$$\begin{aligned}\hat{\partial}_2\widehat{V}_r^4 = P_{r-1}(\widehat{K}) &\oplus \operatorname{span}\big\{L_{r-i}(\hat{x}_1)L_i(\hat{x}_2),\ i=0,\dots,r-2\big\}\\ &\oplus \operatorname{span}\big\{L_{r+1-i}(\hat{x}_1)L_i(\hat{x}_2),\ i=0,\dots,r-2\big\}.\end{aligned} \tag{5.5}$$

We will show that $\hat{v}\in\hat{\partial}_2\widehat{V}_r$ and $\widehat{F}(\hat{v})=0$ for all $\widehat{F}\in\widehat{\mathcal{F}}_r$ give $\hat{v}\equiv 0$. To this end, a general element $\hat{v}\in\hat{\partial}_2\widehat{V}_r$ is written as

$$\hat{v} = \sum_{0\le i+j\le r-1}\alpha_{ij}L_i(\hat{x}_1)L_j(\hat{x}_2) + \sum_{i=0}^{r-2}\beta_i L_{r-i}(\hat{x}_1)L_i(\hat{x}_2) + \sum_{i=0}^{r-2}\gamma_i L_{r+1-i}(\hat{x}_1)L_i(\hat{x}_2).$$

Applying $\widehat{F}_{n,*}$ to $\hat{v}$ and using the orthogonality of the Legendre polynomials it follows that

$$0=\widehat{F}_{n,*}(\hat{v})=\alpha_{n0},\qquad n=0,\dots,r-1. \tag{5.6}$$

Hence, the function $\hat{v}$ can be written as

$$\hat{v} = \sum_{i=0}^{r-2}\sum_{j=1}^{r-1-i}\alpha_{ij}L_i(\hat{x}_1)L_j(\hat{x}_2) + \sum_{i=0}^{r-2}\beta_i L_{r-i}(\hat{x}_1)L_i(\hat{x}_2) + \sum_{i=0}^{r-2}\gamma_i L_{r+1-i}(\hat{x}_1)L_i(\hat{x}_2).$$

Applying $\widehat{F}_{m,0}$, $m=0,\dots,r-2$, and using the orthogonality of the Legendre polynomials together with $2/\delta_0=1$ result in

$$\begin{aligned}0=\widehat{F}_{m,0}(\hat{v}) = &\sum_{i=0}^{r-2}\sum_{j=1}^{r-1-i}\alpha_{ij}\delta_{im}(\delta_{j1}\delta_1+\delta_{j0}\delta_0) + \sum_{i=0}^{r-2}\beta_i\delta_{m,r-i}(\delta_{i1}\delta_1+\delta_{i0}\delta_0)\\ &+\sum_{i=0}^{r-2}\gamma_i\delta_{m,r+1-i}(\delta_{i1}\delta_1+\delta_{i0}\delta_0) = \alpha_{m1}\delta_1\end{aligned} \tag{5.7}$$

since all other terms vanish due to the range of i and j in the various sums. Hence, $\alpha_{m1}=0$ for $m=0,\dots,r-2$. Now, we apply $\widehat{F}_{mn}$ for $n\ge 1$, i.e., for $n=1,\dots,r-2$, $m=0,\dots,r-2-n$. Applying relation (3.1) between Legendre polynomials and integrated Legendre polynomials, we obtain

$$\begin{aligned}0=\widehat{F}_{mn}(\hat{v}) = &\sum_{i=0}^{r-2}\sum_{j=2}^{r-1-i}\alpha_{ij}\delta_{im}(\delta_{j,n+1}\delta_{n+1}+\delta_{j,n-1}\delta_{n-1})\\ &+\sum_{i=0}^{r-2}\beta_i\delta_{m,r-i}(\delta_{i,n+1}\delta_{n+1}+\delta_{i,n-1}\delta_{n-1})\\ &+\sum_{i=0}^{r-2}\gamma_i\delta_{m,r+1-i}(\delta_{i,n+1}\delta_{n+1}+\delta_{i,n-1}\delta_{n-1}),\end{aligned}$$

where again the orthogonality of the Legendre polynomials was used. The first double sum can be simplified, and the sums with β_i and γ_i are again zero due to $m+n \le r-2$. Thus, the equations become

$$0 = \alpha_{m,n+1}\delta_{n+1} - \alpha_{m,n-1}\delta_{n-1}.$$

For $n = 1$ one gets $\alpha_{m2} = 0$, $m = 0, \dots, r-3$, since $\alpha_{m0} = 0$. By induction we obtain with (5.6) and (5.7)

$$\alpha_{ij} = 0, \qquad 0 \le i + j \le r - 1.$$

Hence, we have

$$\hat{v} = \sum_{i=0}^{r-2} \Big(\beta_i L_{r-i}(\hat{x}_1) + \gamma_i L_{r+1-i}(\hat{x}_1)\Big) L_i(\hat{x}_2).$$

Using $\widehat{F}_{+,n}(\hat{v}) = 0$ and $\widehat{F}_{-,n}(\hat{v}) = 0$, $n = 1, \dots, r-1$, we get

$$0 = \sum_{i=0}^{r-2} \Big(\beta_i (+1)^{r-i} + \gamma_i (+1)^{r+1-i}\Big)\big(\delta_{n+1,i}\delta_{n+1} - \delta_{n-1,i}\delta_{n-1}\big),$$

$$0 = \sum_{i=0}^{r-2} \Big(\beta_i (-1)^{r-i} + \gamma_i (-1)^{r+1-i}\Big)\big(\delta_{n+1,i}\delta_{n+1} - \delta_{n-1,i}\delta_{n-1}\big).$$

These equations reduce in the cases $n = r-1$ and $n = r-2$ to

$$\beta_{r-2} + \gamma_{r-2} = 0, \qquad \beta_{r-2} - \gamma_{r-2} = 0,$$

$$\beta_{r-3} + \gamma_{r-3} = 0, \qquad \beta_{r-3} - \gamma_{r-3} = 0;$$

i.e., we get

$$\beta_{r-2} = \gamma_{r-2} = \beta_{r-3} = \gamma_{r-3} = 0.$$

Using this result, we get by induction on $n = r-3, \dots, 1$ the final result

$$\beta_i = \gamma_i = 0, \qquad i = 0, \dots, r-2.$$

This means that $\hat{v} \equiv 0$. ∎

LEMMA 5.3. *The set $\widehat{\mathcal{F}}_r$ of functionals is unisolvent on $\hat{\partial}_2 \widehat{V}_r^3$.*

Proof. In order to get an expression for $\hat{\partial}_2 \widehat{V}_r^3$, we have to consider the difference term carefully. Setting

$$k := \lfloor r/2 \rfloor \tag{5.8}$$

and using the differentiated version of (3.1), we obtain

$$\begin{aligned}
&\frac{\partial}{\partial \hat{x}_2}\big(L_{k+2}(\hat{x}_1)L_k(\hat{x}_2) - L_k(\hat{x}_1)L_{k+2}(\hat{x}_2)\big) \\
&\quad = L_{k+2}(\hat{x}_1)\left[L'_{k-2}(\hat{x}_2) + \frac{2}{\delta_{k-1}} L_{k-1}(\hat{x}_2)\right] - L_k(\hat{x}_1)\left[L'_k(\hat{x}_2) + \frac{2}{\delta_{k+1}} L_{k+1}(\hat{x}_2)\right] \\
&\quad = \frac{2}{\delta_{k+1}}\left[\frac{\delta_{k+1}}{\delta_{k-1}} L_{k+2}(\hat{x}_1)L_{k-1}(\hat{x}_2) - L_k(\hat{x}_1)L_{k+1}(\hat{x}_2)\right] + q_{r-1}
\end{aligned}$$

with $q_{r-1} \in P_{r-1}(\widehat{K})$. Hence, the space $\hat{\partial}_2 \widehat{V}_r^3$ can be written as

(5.9)

$$\begin{aligned}\hat{\partial}_2 \widehat{V}_r^3 = {} & P_{r-1}(\widehat{K}) \\ & \oplus \operatorname{span}\Big\{ L_{r-i}(\hat{x}_1) L_i(\hat{x}_2), L_{r+1-i}(\hat{x}_1) L_i \,:\, i = 0, \dots, k-2 \text{ or } i = k+2, \dots, r \Big\} \\ & \oplus \operatorname{span}\left\{ \frac{\delta_{k+1}}{\delta_{k-1}} L_{k+2}(\hat{x}_1) L_{k-1}(\hat{x}_2) - L_k(\hat{x}_1) L_{k+1}(\hat{x}_2) \right\} \\ & \oplus \operatorname{span}\Big\{ L_m(\hat{x}_1) L_n(\hat{x}_2) \Big\},\end{aligned}$$

where k is defined in (5.8) and

$$(m,n) := \begin{cases} (k+3, k-1) & \text{for } r \text{ odd}, \\ (k-1, k+1) & \text{for } r \text{ even}. \end{cases} \tag{5.10}$$

As usual, we will show that $\hat{v} \in \hat{\partial}_2 \widehat{V}_r^3$ with $\widehat{F}(\hat{v}) = 0$ for all $\widehat{F} \in \widehat{\mathcal{F}}_r$ results in $\hat{v} \equiv 0$. In order to do so, we write $\hat{v}$ as a linear combination of the basis functions given in (5.9) and the Legendre basis of $P_{r-1}(\widehat{K})$. Proceeding as in the beginning of the proof of Lemma 5.2, we get that all coefficients for functions from $P_{r-1}(\widehat{K})$ vanish. Hence, the function $\hat{v}$ can be represented as

$$\begin{aligned}\hat{v} = {} & \sum_{i=0}^{k-2} \Big(\beta_i L_{r-i}(\hat{x}_1) L_i(\hat{x}_2) + \gamma_i L_{r+1-i}(\hat{x}_1) L_i(\hat{x}_2) \Big) \\ & + \sum_{i=k+2}^{r} \Big(\beta_i L_{r-i}(\hat{x}_1) L_i(\hat{x}_2) + \gamma_i L_{r+1-i}(\hat{x}_1) L_i(\hat{x}_2) \Big) \\ & + \mu \left(\frac{\delta_{k+1}}{\delta_{k-1}} L_{k+2}(\hat{x}_1) L_{k-1}(\hat{x}_2) - L_k(\hat{x}_1) L_{k+1}(\hat{x}_2) \right) \\ & + \vartheta L_m(\hat{x}_1) L_n(\hat{x}_2)\end{aligned}$$

with k from (5.8) and m, n from (5.10). Now, the cases of odd and even r have to be distinguished. Let r be even. Applying the nodal functionals $\widehat{F}_{\pm,k}$, we obtain

$$\begin{aligned} 0 = \widehat{F}_{\pm,k}(\hat{v}) = {} & \mu \frac{\delta_{k+1}}{\delta_{k-1}} \int_{-1}^{1} (\pm 1)^{k+2} L_{k-1}(\hat{x}_2) \big(L_{k+1}(\hat{x}_2) - L_{k-1}(\hat{x}_2) \big) \, d\hat{x}_2 \\ & - \mu \int_{-1}^{1} (\pm 1)^{k} L_{k+1}(\hat{x}_2) \big(L_{k+1}(\hat{x}_2) - L_{k-1}(\hat{x}_2) \big) \, d\hat{x}_2 \\ & + \vartheta \int_{-1}^{1} (\pm 1)^{k-1} L_{k+1}(\hat{x}_2) \big(L_{k+1}(\hat{x}_2) - L_{k-1}(\hat{x}_2) \big) \, d\hat{x}_2 \\ = {} & (\pm 1)^k \left[\frac{\delta_{k+1}}{\delta_{k-1}} (-\delta_{k-1}) \mu - \delta_{k+1} \mu + (\pm 1) \delta_{k+1} \vartheta \right] \end{aligned}$$

by again using the orthogonality of the Legendre polynomials. Hence, we get

$$0 = -2\mu + \vartheta, \qquad 0 = -2\mu - \vartheta,$$

which immediately give $\mu = \vartheta = 0$. We obtain in a similar way that $\mu = \vartheta = 0$ also for odd r. For showing that the coefficients β_i and γ_i will vanish, we proceed as in the end of the proof of Lemma 5.2. Here, we apply $\widehat{F}_{\pm,\ell}$ for $\ell = k+1, \dots, r-1$ and $\ell = k-1, \dots, 0$. Finally, we end up with $\hat{v} \equiv 0$ since all coefficients vanish. ∎

We are now prepared to prove the anisotropic interpolation error estimate for $\partial_2(u - I_h^\nu u)$, $\nu = 3, 4$. We start with the case of smooth functions $u \in W^{\ell,p}(K)$, $\ell \geq 2$.

LEMMA 5.4. *Let $u \in W^{\ell,p}(K)$, where $\ell \in \mathbb{N}$, $2 \leq \ell \leq r+1$, $p \in [1, \infty]$. Fix $q \in [1, \infty]$ such that $W^{\ell-1,p}(K) \hookrightarrow L^q(K)$. Then, the anisotropic interpolation error estimate*

$$\|\partial_2(u - I_h^\nu u)\|_{0,q,K} \lesssim |K|^{1/q-1/p} \sum_{|\alpha|=\ell-1} h_K^\alpha |D^\alpha u|_{1,p,K}, \quad \nu = 3, 4,$$

holds true.

Proof. We prove first an estimate on the reference element $\widehat{K}$. In order to be able to apply [2, Lemma 2.2], it remains to show that the functionals $\widehat{F} \in \widehat{\mathcal{F}}_r$ satisfy the properties (5.1) and (5.2) since the property (5.3) was already proved in Lemmas 5.2 and 5.3.

Property (5.1) is satisfied since $\ell - 1 \geq 1$ and all functionals are well defined for $\hat{u} \in W^{1,1}(\widehat{K})$. We note here that weaker functions cannot be used since their trace on an edge of $\widehat{K}$ must be integrable. This is the reason for the assumption $\ell \geq 2$.

Property (5.2) can also be proved simultaneously for both $\widehat{V}_r^3$ and $\widehat{V}_r^4$ since the set of nodal functionals $\widehat{\mathcal{N}}$ is the same. We have

$$\begin{aligned}
\widehat{F}_{i,*}\big(\hat{\partial}_2(\hat{u} - \widehat{I}^\nu \hat{u})\big) &= \frac{1}{\delta_i \delta_0} \int_{-1}^{1} \int_{-1}^{1} L_i(\hat{x}_1) L_0(\hat{x}_2) \hat{\partial}_2\big(\hat{u} - \widehat{I}^\nu \hat{u}\big)(\hat{x}_1, \hat{x}_2)\, d\hat{x}_1\, d\hat{x}_2 \\
&= \frac{1}{\delta_i \delta_0} \int_{-1}^{1} L_i(\hat{x}_1)\big(\hat{u} - \widehat{I}^\nu \hat{u}\big)(\hat{x}_1, +1)\, d\hat{x}_1 \\
&\qquad - \frac{1}{\delta_i \delta_0} \int_{-1}^{1} L_i(\hat{x}_1)\big(\hat{u} - \widehat{I}^\nu \hat{u}\big)(\hat{x}_1, -1)\, d\hat{x}_1 \\
&= \frac{1}{\delta_0}\big(\widehat{N}_{i,+}(\hat{u} - \widehat{I}^\nu \hat{u}) - \widehat{N}_{i,-}(\hat{u} - \widehat{I}^\nu \hat{u})\big) = 0, \qquad i = 0, \dots, r-1, \\
\widehat{F}_{\pm,i}\big(\hat{\partial}_2(\hat{u} - \widehat{I}^\nu \hat{u})\big) &= \frac{2}{\delta_i} \int_{-1}^{1} \widehat{L}_i(\hat{x}_2) \hat{\partial}_2\big(\hat{u} - \widehat{I}^\nu \hat{u}\big)(\pm 1, \hat{x}_2)\, d\hat{x}_2 \\
&= -\frac{2}{\delta_i} \int_{-1}^{1} L_i(\hat{x}_2)\big(\hat{u} - \widehat{I}^\nu \hat{u}\big)(\pm 1, \hat{x}_2)\, d\hat{x}_2 \\
&= \widehat{N}_{\pm,i}\big(\hat{u} - \widehat{I}^\nu \hat{u}\big) = 0, \qquad i = 1, \dots, r-1,
\end{aligned}$$

$$\begin{aligned}\widehat{F}_{ij}\big(\hat\partial_2(\hat u-\widehat I^\nu\hat u)\big) &= \frac{2}{\delta_i\delta_j}\int_{-1}^1 L_i(\hat x_1)\int_{-1}^1 \widehat L_j(\hat x_2)\hat\partial_2\big(\hat u-\widehat I^\nu\hat u\big)(\hat x_1,\hat x_2)\,d\hat x_2\,d\hat x_1\\ &= \delta_{j0}\frac{2}{\delta_i\delta_0}\int_{-1}^1 L_i(\hat x_1)\widehat L_0(+1)\big(\hat u-\widehat I^\nu\hat u\big)(\hat x_1,1)\,d\hat x_1\\ &\quad -\frac{2}{\delta_i\delta_j}\int_{-1}^1 L_i(\hat x_1)\int_{-1}^1 L_j(\hat x_2)\big(\hat u-\widehat I^\nu\hat u\big)(\hat x_1,\hat x_2)\,d\hat x_2\,d\hat x_1\\ &= 2\delta_{j0}\widehat N_{i,+}\big(\hat u-\widehat I^\nu\hat u\big)-2\widehat N_{ij}\big(\hat u-\widehat I^\nu\hat u\big)=0,\qquad 0\le i+j\le r-2.\end{aligned}$$

Consequently, Lemma 2.2 of [2] delivers

$$\|\hat\partial_2(\hat u-\widehat I^\nu\hat u)\|_{0,q,\widehat K}\lesssim|\hat\partial_2\hat u|_{\ell-1,p,\widehat K}. \tag{5.11}$$

Note that the condition $\hat u\in C(\widehat K)$ in [2, Lemma 2.2] is not needed here but only for Lagrangian finite elements. The transformation of (5.11) from $\widehat K$ to K gives

$$\begin{aligned}\|\partial_2(u-I_h^\nu u)\|_{0,q,K}&\lesssim|K|^{1/q-1/p}\sum_{|\alpha|=\ell-1}h_K^\alpha\|D^\alpha\partial_2 u\|_{0,p,K}\\ &\lesssim|K|^{1/q-1/p}\sum_{|\alpha|=\ell-1}h_K^\alpha|D^\alpha u|_{1,p,K}.\end{aligned}$$

This is the desired estimate. ∎

Example 5.5. We show that the assumption $\ell\ge 2$ in Lemma 5.4 is essential. The validity of

$$\|\partial_2(u-I_h^\nu u)\|_{0,q,K}\lesssim|K|^{1/q-1/p}|u|_{1,p,K}$$

is equivalent to

$$\|\hat\partial_2(\hat u-\widehat I^\nu\hat u)\|_{0,q,\widehat K}\lesssim\|\hat\partial_2\hat u\|_{0,p,\widehat K}; \tag{5.12}$$

compare this with [2, Example 2.1]. In order to show that estimate (5.12) is not valid, we adapt Example 2.3 of [2] and consider the sequence of functions

$$\hat u_\varepsilon(\hat x_1,\hat x_2):=\hat x_2\cdot\min\left\{1,\varepsilon\left|\ln\frac{\hat x_1+1}{2}\right|\right\},\quad \varepsilon\to+0.$$

The limit function is defined pointwise by

$$\hat u_0(\hat x_1,\hat x_2):=\lim_{\varepsilon\to0}\hat u_\varepsilon(\hat x_1,\hat x_2)=\begin{cases}\hat x_2 & \text{for } \hat x_1=-1,\\ 0 & \text{for } \hat x_1>-1.\end{cases}$$

Since, for $p<\infty$,

$$\lim_{\varepsilon\to0}\|\hat\partial_2\hat u_\varepsilon\|_{0,p,\widehat K}=0\quad\text{and}\quad\lim_{\varepsilon\to0}\|\hat\partial_2(\hat u_\varepsilon-\widehat I^\nu\hat u_\varepsilon)\|_{0,q,\widehat K}=\|\hat\partial_2\widehat I^\nu\hat u_0\|_{0,q,\widehat K}=C>0,$$

the estimate (5.12) cannot hold.

The last equality is proved by contradiction: If $\|\hat\partial_2\widehat I^\nu\hat u_0\|_{0,q,\widehat K}=0$, then $I^\nu\hat u_0=:\hat v(\hat x_1)$ is polynomial in $\hat x_1$, and $(\hat u_0-\widehat I^\nu\hat u_0)(-1,\hat x_2)=\hat x_2-\hat v(-1)$ is a nonvanishing first

order polynomial in $\hat{x}_2$. On the other hand, we obtain from the definition of $\widehat{I}^\nu$ that this polynomial is orthogonal to all polynomials of order $r-1$. This is not possible for a nonvanishing first order polynomial if $r \geq 2$.

In order to prove an anisotropic interpolation error estimate for nonsmooth functions $u \in W^{1,p}(K)$, we have to deal with the difficulty that $\widehat{F}_{-,i}$ and $\widehat{F}_{+,i}$, $i = 1, \ldots, r-1$, are in general not defined for functions $\hat{v} \in L^p(\widehat{K})$. The alternative is to consider functions from $W^{1,p}_E(K)$ which vanish on $E_{-,*}$ and $E_{+,*}$; see (4.1).

LEMMA 5.6. *Let $u \in W^{1,p}_E(K)$ defined in (4.1), $p, q \in [1, \infty]$, $q \leq p$. Then, the estimate*

$$\|\partial_2(u - I_h^\nu u)\|_{0,q,K} \lesssim |K|^{1/q-1/p} |\partial_2 u|_{0,p,K}, \quad \nu = 3, 4,$$

holds true.

Proof. We start with an estimate on the reference element $\widehat{K}$. We have with respect to (5.3) and (5.4) the following equivalence of norms on the space $\hat{\partial}_2 \widehat{V}_r^\nu$, $\nu = 3, 4$:

$$\|\hat{\partial}_2 \widehat{I}^\nu \hat{u}\|_{0,q,\widehat{K}} \lesssim \sum_{0 \leq i+j \leq r-2} |\widehat{F}_{ij}(\hat{\partial}_2 \widehat{I}^\nu \hat{u})| + \sum_{i=0}^{r-1} |\widehat{F}_{i,*}(\hat{\partial}_2 \widehat{I}^\nu \hat{u})|$$
$$+ \sum_{i=1}^{r-1} |\widehat{F}_{+,i}(\hat{\partial}_2 \widehat{I}^\nu \hat{u})| + \sum_{i=1}^{r-1} |\widehat{F}_{+,i}(\hat{\partial}_2 \widehat{I}^\nu \hat{u})|.$$

For the functionals $\widehat{F}_{ij}$ and $\widehat{F}_{i,*}$, we use their property (5.2) and their boundedness for functions from $L^1(\widehat{K})$:

$$|\widehat{F}_{ij}(\hat{\partial}_2 \widehat{I}^\nu \hat{u})| = |\widehat{F}_{ij}(\hat{\partial}_2 \hat{u})| \lesssim \|\hat{\partial}_2 \hat{u}\|_{0,p,\widehat{K}}, \quad 0 \leq i+j \leq r-2,$$

$$|\widehat{F}_{i,*}(\hat{\partial}_2 \widehat{I}^\nu \hat{u})| = |\widehat{F}_{i,*}(\hat{\partial}_2 \hat{u})| \lesssim \|\hat{\partial}_2 \hat{u}\|_{0,p,\widehat{K}}, \quad i = 0, \ldots, r-1.$$

Moreover, we integrate by parts for the functionals $\widehat{F}_{\pm,i}$ and employ $\widehat{L}_i(\pm 1) = 0$ for $i \geq 1$ to obtain

$$\left|\widehat{F}_{\pm,i}(\hat{\partial}_2 \widehat{I}^\nu \hat{u})\right| = \frac{2}{\delta_i} \left| \int_{-1}^{1} \widehat{L}_i(\hat{x}_2)(\hat{\partial}_2 \widehat{I}^\nu \hat{u})(\pm 1, \hat{x}_2)\, d\hat{x}_2 \right| = \frac{2}{\delta_i} \left| \int_{-1}^{1} L_i(\hat{x}_2) \widehat{I}^\nu \hat{u}(\pm 1, \hat{x}_2)\, d\hat{x}_2 \right|$$
$$= \left|\widehat{N}_{\pm,i}(\widehat{I}^\nu \hat{u})\right| = \left|\widehat{N}_{\pm,i}(\hat{u})\right| = 0.$$

Therefore, we can conclude

$$\|\hat{\partial}_2 \widehat{I}^\nu \hat{u}\|_{0,q,\widehat{K}} \lesssim \|\hat{\partial}_2 \hat{u}\|_{0,p,\widehat{K}},$$

$$\|\hat{\partial}_2 (\hat{u} - \widehat{I}^\nu \hat{u})\|_{0,q,\widehat{K}} \lesssim \|\hat{\partial}_2 \hat{u}\|_{0,p,\widehat{K}},$$

where we have used $\|\cdot\|_{0,q,\widehat{K}} \lesssim \|\cdot\|_{0,p,\widehat{K}}$ for $q \leq p$. The assertion of the lemma is proved by transforming this estimate from $\widehat{K}$ to K. ∎

Lemmas 5.1 and 5.4 yield Theorem 4.1, whereas Lemmas 5.1 and 5.6 give Theorem 4.2.

6. Checking the inf-sup condition. For showing that the inf-sup condition holds true, we will use the macroelement technique introduced in [7]; see also [13, sect. II.1.4]. As already described in section 4, the domain Ω is decomposed into isotropic rectangular macroelements Q such that

$$\overline{\Omega} = \bigcup_{Q\in\mathcal{T}_H} \overline{Q}.$$

On the macroelements Q from this macrotriangulation $\mathcal{T}_H$, we define subspaces of $\boldsymbol{X}_h$ and Q_h in the following way:

$$\boldsymbol{X}_h(Q) := \bigl\{\boldsymbol{v} \,:\, \boldsymbol{v}\in\boldsymbol{X}_h,\, \boldsymbol{v}\equiv\boldsymbol{0} \text{ in } \Omega\setminus\overline{Q}\bigr\},$$

$$M_h(Q) := \bigl\{q|_Q : q\in M_h\bigr\} \cap L^2_0(Q).$$

Furthermore, let

$$\overline{M}_h := \bigl\{q\in L^2_0(\Omega) \,:\, q|_Q = \text{const}\ \forall Q\in\mathcal{T}_H\bigr\}$$

be the space of piecewise constant functions with respect to the macrotriangulation $\mathcal{T}_H$. Now, we can give the uniform local inf-sup condition

$$\exists\beta_\ell>0\ \forall Q\in\mathcal{T}_H \,:\, \inf_{q_h\in M_h(Q)} \sup_{\boldsymbol{v}_h\in\boldsymbol{X}_h(Q)} \frac{b_h(\boldsymbol{v}_h,q_h)}{|\boldsymbol{v}_h|_{1,h}\,\|q_h\|_{0,Q}} \geq \beta_\ell.$$

Adapting the results from [7] to the nonconforming case, we can state the following lemma.

LEMMA 6.1. *Let the local inf-sup condition be fulfilled with a constant β_ℓ which is independent of the number of macroelements and the mesh size h. Furthermore, we assume that there exists a subspace $\overline{\boldsymbol{X}}_h$ of $\boldsymbol{X}_h$ such that $(\overline{\boldsymbol{X}}_h,\overline{M}_h)$ satisfies an inf-sup condition with a constant $\overline{\beta}>0$ which does not depend on h. Then, there exists a constant $\beta>0$ independent of h such that $(\boldsymbol{X}_h,M_h)$ is inf-sup stable with the constant β.*

Proof. For conforming finite elements, the proof can be found in [7] and [13, sect. II.1.4]. For the adaptation of the idea to nonconforming finite elements, we refer the reader to [10, 19, 20]. ∎

A careful inspection of the macroelement techniques shows that the global inf-sup constant is independent of all mesh parameters if the constants β_ℓ and $\overline{\beta}$ are independent of the mesh parameters.

For our purposes, we will use for $\overline{\boldsymbol{X}}_h$ the space of continuous functions v_H with $v_H|_Q \in Q'_2(Q)$ (serendipity elements) for all $Q\in\mathcal{T}_H$. Note that this space is a subspace of $\boldsymbol{X}_h$ for $r\geq 3$ and forms together with $\overline{M}_h$ an inf-sup stable pair. Hence, it remains to show that the local inf-sup condition will be satisfied on boundary layer patches and corner patches. Let us start with boundary layer patches.

LEMMA 6.2. *Let Q be a boundary layer patch. Then, there exists a constant $\beta>0$ independent of the aspect ratio of the elements such that*

$$\beta_h := \inf_{q_h\in M_h(Q)} \sup_{\boldsymbol{v}_h\in\boldsymbol{X}_h(Q)} \frac{b_h(\boldsymbol{v}_h,q_h)}{|\boldsymbol{v}|_{1,Q}\,\|q_h\|_{0,Q}} > \beta \tag{6.1}$$

holds true provided $r\geq 2$.

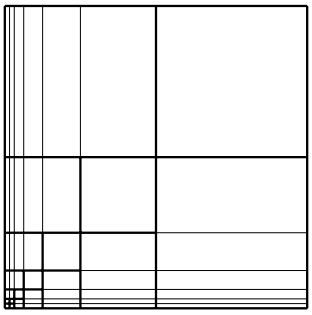


FIG. 6.1. *Corner patch with macrotriangulation.*

Proof. Due to Fortin [12], the fulfillment of the inf-sup condition on a macroelement Q is equivalent to the existence of an interpolation operator $\boldsymbol{i}_Q : \boldsymbol{H}_0^1(Q) \to \boldsymbol{X}_h$ which has the following properties:

$$b_h(\boldsymbol{i}_Q \boldsymbol{v}, q_h) = b_h(\boldsymbol{v}, q_h) \qquad \forall q_h \in M_h(Q),\ \boldsymbol{v} \in \boldsymbol{H}_0^1(Q), \tag{6.2}$$

$$|\boldsymbol{i}_Q \boldsymbol{v}|_{1,h} \lesssim |\boldsymbol{v}|_1 \qquad \forall \boldsymbol{v} \in \boldsymbol{H}_0^1(Q). \tag{6.3}$$

We can choose in the considered situation the natural interpolation operator $\boldsymbol{I}^\nu$, $\nu = 3, 4$, for $\boldsymbol{i}_Q$. The condition (6.3) follows directly from the interpolation error estimate given in Theorem 4.2 for the case $m = 1$, $p = 2$. The condition (6.2) is a consequence of the natural interpolation operator which is defined via the nodal functionals from $\widehat{\mathcal{N}}_r$. Indeed, we have

$$\begin{aligned} b_h(\boldsymbol{I}^\nu \boldsymbol{v}, q_h) &= \sum_{K \subset Q} (\operatorname{div} \boldsymbol{I}^\nu \boldsymbol{v}, q_h)_K \\ &= \sum_{K \subset Q} \big[-(\boldsymbol{I}^\nu \boldsymbol{v}, \nabla q_h)_K + \langle \boldsymbol{n}^K \cdot \boldsymbol{I}^\nu \boldsymbol{v}, q_h \rangle_{\partial K} \big] \\ &= \sum_{K \subset Q} \big[-(\boldsymbol{v}, \nabla q_h)_K + \langle \boldsymbol{n}^K \cdot \boldsymbol{v}, q_h \rangle_{\partial K} \big] \\ &= \sum_{K \subset Q} (\operatorname{div} \boldsymbol{v}, q_h)_K = b_h(\boldsymbol{v}, q_h), \end{aligned}$$

where we have used that $\nabla q_h|_K \in \boldsymbol{P}_{r-2}(K)$ and $q_h|_E \in P_{r-1}(E)$ for all edges $E \in \partial K$. Since both conditions (6.2) and (6.3) are fulfilled and the constant in (6.3) is independent of all mesh parameters, we conclude that the local inf-sup condition is satisfied with a constant which does not depend on the mesh parameters. ∎

LEMMA 6.3. *Let Q be a corner patch. Then there exists a constant $\beta > 0$ independent of the aspect ratio of the elements such that* (6.1) *holds true provided $r \geq 3$.*

Proof. The corner patch Q can be considered as a macrotriangulation $\Delta = \{K\}$ with hanging nodes where the macros K are boundary layer patches. This is illustrated in Figure 6.1 by using lines with different widths. Furthermore, we denote by $\widehat{\Delta} = \{\widehat{K}\}$ the affine image of Δ on $\widehat{Q} = (0,1)^2$. Schötzau, Schwab, and Stenberg [25, sect. 4.2] considered on each $\widehat{K} \in \widehat{\Delta}$ the velocity space

$$L^1(\widehat{K}) := Q_1(\widehat{K})^2 \oplus \operatorname{span}\big\{\hat{b}_1 \hat{\boldsymbol{n}}_1, \ldots, \hat{b}_4 \hat{\boldsymbol{n}}_4\big\},$$

where $\hat{\boldsymbol{n}}_i$ is the outward normal on the ith edge of $\widehat{K}$ and $\hat{b}_i \in Q_2'(\widehat{K})$ is the corresponding edge bubble function, $i = 1, \ldots, 4$. Moreover, they defined the spaces

$$L_0^{1,1}(\widehat{\Delta}) := \bigl\{\hat{\boldsymbol{u}} \in H_0^1(\widehat{Q})^2 : \hat{\boldsymbol{u}}|_{\widehat{K}} \in L^1(\widehat{K})\ \forall \widehat{K} \in \widehat{\Delta}\bigr\},$$

$$S_0^{0,0}(\widehat{\Delta}) := \bigl\{\hat{p} \in L_0^2(\widehat{Q}) : \hat{p}|_{\widehat{K}} \in P_0(\widehat{K})\ \forall \widehat{K} \in \widehat{\Delta}\bigr\}$$

and proved inf-sup stability for this pair [25, Theorem 4.9]. The inf-sup constant depends just on the grading factor σ. The stability for the pair $\bigl(L_0^{1,1}(\Delta), S_0^{0,0}(\Delta)\bigr)$ is obtained by scaling. Since $L^1(\widehat{K}) \subset Q_2'(\widehat{K})^2$, we also have stability for the pair of spaces which is defined on the basis of (Q_2', P_0). This result can also be derived by a closer look at the proofs of Heuveline and Schieweck [15], who investigated the (Q_2, P_0) pair on meshes with hanging nodes.

Since $Q_2'(\widehat{K}) \subset \widehat{V}_r^3, \widehat{V}_r^4$ for $r \geq 3$ and due to the stability on boundary layer patches as proved in Lemma 6.2, we can conclude the assertion of Lemma 6.3 by the macroelement technique. ∎

Numerical tests for the case $r = 2$ have shown that all four families are stable on corner patches with $\sigma = \frac{1}{2}$. Nevertheless, our method of proof does not work for $r = 2$. Therefore, the validity of a uniform inf-sup condition must be considered open for this case.

Now, we can finish this section with the following proof.

Proof of Theorem 4.3. Since we have shown all necessary ingredients for the macroelement technique, the stated inf-sup condition holds true with a constant which is independent of the mesh parameter, aspect ratio, and mesh size. ∎

7. Proof of the consistency error estimate. For proving the estimate of the consistency error, we start with an estimate for an auxiliary problem.

LEMMA 7.1. *For arbitrary $\boldsymbol{\eta} \in \boldsymbol{H}^r(\Omega)$ and $g = -\operatorname{div}\boldsymbol{\eta}$, the estimate*

$$\sup_{v_h \in X_h} \frac{(\boldsymbol{\eta}, \nabla v_h)_h - (g, v_h)}{|v_h|_{1,h}} \lesssim \left(\sum_{K \in \mathcal{T}_h} \sum_{|\alpha|=r} h_K^{2\alpha} \|D^\alpha \boldsymbol{\eta}\|_{0,K}^2 \right)^{1/2} + \left(\sum_{K \in \mathcal{T}_h} h_{K,1}^2 \sum_{|\alpha|=r-1} h_K^{2\alpha} \|D^\alpha g\|_{0,K}^2 \right)^{1/2}$$

holds true where

$$(\boldsymbol{\eta}, \nabla v_h)_h := \sum_{K \in \mathcal{T}_h} (\boldsymbol{\eta}, \nabla v_h)_K.$$

Proof. By using $g = -\operatorname{div}\boldsymbol{\eta}$ and integration by parts, we can rewrite the numerator as

$$(\boldsymbol{\eta}, \nabla v_h)_h - (g, v_h) = \sum_{K \in \mathcal{T}_h} \langle \boldsymbol{\eta} \cdot \boldsymbol{n}_K, v_h \rangle_{\partial K} = T_1 + T_2$$

with

$$T_1 = \sum_{K \in \mathcal{T}_h} \sum_{E \in \mathcal{E}_S(K)} \int_E \boldsymbol{\eta} \cdot \boldsymbol{n}\, v_h \, d\gamma, \qquad T_2 = \sum_{K \in \mathcal{T}_h} \sum_{E \in \mathcal{E}_L(K)} \int_E \boldsymbol{\eta} \cdot \boldsymbol{n}\, v_h \, d\gamma,$$

where $\mathcal{E}_L(K)$ and $\mathcal{E}_S(K)$ denote the sets of long and short edges of K, respectively. Let us further assume that the long edges are parallel to the x_1-direction while the short edges are parallel to the x_2-direction. This gives that $n_2 = 0$ on short edges and $n_1 = 0$ on long edges. Hence, T_1 and T_2 simplify to

$$T_1 = \sum_{E\in\mathcal{E}_S} \int_E \eta_1 n_1^E [v_h]_E \, d\gamma, \qquad T_2 = \sum_{E\in\mathcal{E}_L} \int_E \eta_2 n_2^E [v_h]_E \, d\gamma.$$

Here, $\mathcal{E}_L$ and $\mathcal{E}_S$ are the global sets of long and short edges, respectively. Let Π denote the special interpolation operator into the space of continuous Q_{r-1} elements defined in [13, p. 108]. Due to the definition of X_h, the jump $[v_h]_E$ is orthogonal to all polynomials of degree $r-1$. Hence, we obtain

$$T_1 = \sum_{E\in\mathcal{E}_S} \int_E (\eta_1 - \Pi\eta_1) n_1^E [v_h]_E \, d\gamma, \qquad T_2 = \sum_{E\in\mathcal{E}_L} \int_E (\eta_2 - \Pi\eta_2) n_2^E [v_h]_E \, d\gamma$$

since the normal vectors are constant on the edges. Using the Gaussian theorem and the product rule, we get

$$\begin{aligned} T_1 &= \sum_{K\in\mathcal{T}_h} \int_K \partial_1 \big[(\eta_1 - \Pi\eta_1) v_h\big] \, dx \\ &= \sum_{K\in\mathcal{T}_h} \int_K \partial_1(\eta_1 - \Pi\eta_1) v_h \, dx + \sum_{K\in\mathcal{T}_h} \int_K (\eta_1 - \Pi\eta_1) \partial_1 v_h \, dx, \\ T_2 &= \sum_{K\in\mathcal{T}_h} \int_K \partial_2 \big[(\eta_2 - \Pi\eta_2) v_h\big] \, dx \\ &= \sum_{K\in\mathcal{T}_h} \int_K \partial_2(\eta_2 - \Pi\eta_2) v_h \, dx + \sum_{K\in\mathcal{T}_h} \int_K (\eta_2 - \Pi\eta_2) \partial_2 v_h \, dx. \end{aligned}$$

Due to the choice of the special interpolation operator Π, we have

$$\int_K \partial_i(\eta_i - \Pi\eta_i) \, dx = 0.$$

Hence, we get

$$\begin{aligned} T_1 &= \sum_{K\in\mathcal{T}_h} \int_K \partial_1(\eta_1 - \Pi\eta_1)(v_h - v_h^K) \, dx + \sum_{K\in\mathcal{T}_h} \int_K (\eta_1 - \Pi\eta_1) \partial_1 v_h \, dx, \\ T_2 &= \sum_{K\in\mathcal{T}_h} \int_K \partial_2(\eta_2 - \Pi\eta_2)(v_h - v_h^K) \, dx + \sum_{K\in\mathcal{T}_h} \int_K (\eta_2 - \Pi\eta_2) \partial_2 v_h \, dx, \end{aligned}$$

where v_h^K is the integral mean of v_h on the cell K. The interpolation operator Π fulfills

$$\|\partial_i(w - \Pi w)\|_{0,K} \lesssim \sum_{|\alpha|=r-1} h_K^\alpha \|D^\alpha \partial_i w\|_{0,K}, \qquad \|w - \Pi w\|_{0,K} \lesssim \sum_{|\alpha|=r} h_K^\alpha \|D^\alpha w\|_{0,K}$$

provided that $w \in H^r(K)$; see [1, 26]. We end up with

$$|T_1| \lesssim \left[\left(\sum_{K\in\mathcal{T}_h} h_{K,1}^2 \sum_{|\alpha|=r-1} h_K^{2\alpha} \|D^\alpha \partial_1 \eta_1\|_{0,K}^2 \right)^{1/2} + \left(\sum_{K\in\mathcal{T}_h} \sum_{|\alpha|=r} h_K^{2\alpha} \|D^\alpha \eta_1\|_{0,K}^2 \right)^{1/2} \right] |v_h|_{1,h},$$

$$|T_2| \lesssim \left[\left(\sum_{K\in\mathcal{T}_h} h_{K,1}^2 \sum_{|\alpha|=r-1} h_K^{2\alpha} \|D^\alpha \partial_2 \eta_2\|_{0,K}^2 \right)^{1/2} + \left(\sum_{K\in\mathcal{T}_h} \sum_{|\alpha|=r} h_K^{2\alpha} \|D^\alpha \eta_2\|_{0,K}^2 \right)^{1/2} \right] |v_h|_{1,h},$$

where the Cauchy–Schwarz inequality was used. The desired estimate is obtained by using $\partial_2\eta_2 = -g - \partial_1\eta_1$. ∎

We finish this section with the following proof.

Proof of Theorem 4.4. The proof is accomplished by applying Lemma 7.1 to

$$\boldsymbol{\eta}^{(1)} = \begin{pmatrix} \partial_1 u_1 \\ \partial_2 u_1 \end{pmatrix}, \quad g^{(1)} = f_1 - \partial_1 p, \quad \text{and} \quad \boldsymbol{\eta}^{(2)} = \begin{pmatrix} \partial_1 u_2 \\ \partial_2 u_2 - p \end{pmatrix}, \quad g^{(2)} = f_2.$$

This choice leads to

$$\begin{aligned} &|a_h(\boldsymbol{u},\boldsymbol{v}_h) - b_h(\boldsymbol{v}_h,p) - (\boldsymbol{f},\boldsymbol{v}_h)| \\ &\quad = \left|(\boldsymbol{\eta}^{(1)},\nabla v_{h,1})_h - (g^{(1)},v_{h,1}) + (\boldsymbol{\eta}^{(2)},\nabla v_{h,2})_h - (g^{(2)},v_{h,2})\right| \\ &\quad \lesssim \left(\sum_{K\in\mathcal{T}_h} \sum_{|\alpha|=r} h_K^{2\alpha} \left(|D^\alpha \boldsymbol{u}|_{1,K}^2 + \|D^\alpha p\|_{0,K}^2 \right) \right)^{1/2} |\boldsymbol{v}_h|_{1,h} \\ &\qquad + \left(\sum_{K\in\mathcal{T}_h} h_{K,1}^2 \sum_{|\alpha|=r-1} h_K^{2\alpha} \left(\|D^\alpha \boldsymbol{f}\|_{0,K}^2 + \|D^\alpha \partial_1 p\|_{0,K}^2 \right) \right)^{1/2} |\boldsymbol{v}_h|_{1,h}. \end{aligned}$$

The pressure terms in the last row are part of those in the second last row. This finishes the proof. ∎

Appendix. Unisolvence. This appendix contains the proof of the unisolvence results for $\widehat{V}_r^4$.

LEMMA A.1. *The set $\widehat{\mathcal{N}}_r$ is unisolvent on $\widehat{V}_r^4$.*

Proof. We have to show that $\hat{v} \in \widehat{V}_r^4$ and $\widehat{N}(\hat{v}) = 0$ for all $\widehat{N} \in \widehat{\mathcal{N}}_r$ results in $\hat{v} \equiv 0$. To this end, we represent an arbitrary function $\hat{v} \in \widehat{V}_r^4$ in the form

(A.1)
$$\hat{v} = \sum_{0\le i+j\le r} \alpha_{ij} L_i(\hat{x}_1) L_j(\hat{x}_2) + \sum_{i=0}^{r-1} \beta_i L_{r+1-i}(\hat{x}_1) L_i(\hat{x}_2) + \sum_{i=1}^{r-1} \gamma_i L_{r+2-i}(\hat{x}_1) L_i(\hat{x}_2).$$

The orthogonality of the Legendre polynomials gives

$$0 = \widehat{N}_{mn}(\hat{v}) = \alpha_{mn}, \qquad 0 \le m + n \le r - 2.$$

Hence, we have

$$\hat{v} = \sum_{i=0}^{r-1} \alpha_{r-1-i,i} L_{r-1-i}(\hat{x}_1) L_i(\hat{x}_2) + \sum_{i=0}^{r} \alpha_{r-i,i} L_{r-i}(\hat{x}_1) L_i(\hat{x}_2)$$
$$+ \sum_{i=0}^{r-1} \beta_i L_{r+1-i}(\hat{x}_1) L_i(\hat{x}_2) + \sum_{i=1}^{r-1} \gamma_i L_{r+2-i}(\hat{x}_1) L_i(\hat{x}_2).$$

The application of $\widehat{N}_{-,n}$ and $\widehat{N}_{+,n}$, $0 \le n \le r-1$, results in

$$0 = \alpha_{r-1-n,n} + \alpha_{r-n,n} + \beta_n + \gamma_n, \qquad n = 0, \dots, r-1, \tag{A.2}$$

$$0 = \alpha_{r-1-n,n} - \alpha_{r-n,n} + \beta_n - \gamma_n, \qquad n = 0, \dots, r-1, \tag{A.3}$$

where again the orthogonality of the Legendre polynomials is used. The parameter $\gamma_0 = 0$ is introduced for the convenience of notation. The sum and the difference of (A.2) and (A.3) give

$$0 = \alpha_{r-1-n,n} + \beta_n, \qquad n = 0, \dots, r-1, \tag{A.4}$$

$$0 = \alpha_{r-n,n} + \gamma_n, \qquad n = 0, \dots, r-1, \tag{A.5}$$

respectively. If we apply $\widehat{N}_{m,-}$ and $\widehat{N}_{m,+}$, $m = 0, \dots, r-1$, then we get

$$0 = \alpha_{m,r-1-m} - \alpha_{m,r-m} + \beta_{r+1-m} - \gamma_{r+2-m}, \qquad m = 0, \dots, r-1, \tag{A.6}$$

$$0 = \alpha_{m,r-1-m} + \alpha_{m,r-m} + \beta_{r+1-m} + \gamma_{r+2-m}, \qquad m = 0, \dots, r-1, \tag{A.7}$$

where we set $\beta_r = \beta_{r+1} = 0$ and $\gamma_r = \gamma_{r+1} = \gamma_{r+2} = 0$. Taking the sum and the difference of (A.6) and (A.7), we obtain

$$0 = \alpha_{m,r-1-m} + \beta_{r+1-m}, \qquad m = 0, \dots, r-1, \tag{A.8}$$

$$0 = \alpha_{m,r-m} + \gamma_{r+2-m}, \qquad m = 0, \dots, r-1, \tag{A.9}$$

respectively. Using (A.4) for n and (A.8) for $m = r - 1 - n$, we obtain

$$\alpha_{r-1-n,n} + \beta_n = 0 = \alpha_{r-1-n,n} + \beta_{n+2}, \qquad n = 0, \dots, r-1.$$

Hence, we get $\beta_n = \beta_{n+2}, n = 0, \dots, r-1$. We conclude with $\beta_r = \beta_{r+1} = 0$ that

$$\beta_n = 0, \qquad n = 0, \dots, r-1,$$

and with (A.4)

$$\alpha_{r-1-n,n} = 0, \qquad n = 0, \dots, r-1.$$

For $n = 1, \dots, r-1$, we obtain from (A.5) for n and from (A.9) for $m = r - n$

$$\alpha_{r-n,n} + \gamma_n = 0 = \alpha_{r-n,n} + \gamma_{n+2}, \qquad n = 1, \dots, r-1.$$

Hence, we have $\gamma_n = \gamma_{n+2}, n = 1, \ldots, r-1$. Together with $\gamma_r = \gamma_{r+1} = 0$, we end up with

$$\gamma_n = 0, \qquad n = 1, \ldots, r-1.$$

Putting this into (A.5), we get

$$\alpha_{r-n,n} = 0, \qquad n = 1, \ldots, r-1.$$

It remains to consider $\alpha_{r,0}$ and $\alpha_{0,r}$. From (A.5) for $n = 0$ we get $\alpha_{r,0} = 0$ since $\gamma_0 = 0$. Setting $m = 0$ in (A.9) results in $\alpha_{0,r} = 0$ due to $\gamma_{r+2} = 0$. Finally, we have seen that all coefficients in the representation (A.1) vanish. Hence, $\hat{v} \equiv 0$, and the set of nodal functionals $\widehat{\mathcal{N}}_r$ is unisolvent on $\widehat{V}_r^4$. □

REFERENCES

[1] G. Acosta, T. Apel, R. G. Durán, and A. L. Lombardi, *Anisotropic error estimates for an interpolant defined via moments*, Computing, submitted.

[2] T. Apel, *Anisotropic Finite Elements: Local Estimates and Applications*, Adv. Numer. Math., Teubner, Stuttgart, 1999.

[3] T. Apel, S. Nicaise, and J. Schöberl, *Crouzeix-Raviart type finite elements on anisotropic meshes*, Numer. Math., 89 (2001), pp. 193–223.

[4] T. Apel, S. Nicaise, and J. Schöberl, *A non-conforming finite element method with anisotropic mesh grading for the Stokes problem in domains with edges*, IMA J. Numer. Anal., 21 (2001), pp. 843–856.

[5] T. Apel and H. M. Randrianarivony, *Stability of discretizations of the Stokes problem on anisotropic meshes*, Math. Comput. Simulation, 61 (2003), pp. 437–447.

[6] R. Becker and R. Rannacher, *Finite element solution of the incompressible Navier-Stokes equations on anisotropically refined meshes*, in Fast Solvers for Flow Problems (Kiel, 1994), Notes Numer. Fluid Mech. 49, Vieweg, Braunschweig, 1995, pp. 52–62.

[7] J. M. Boland and R. A. Nicolaides, *Stability of finite elements under divergence constraints*, SIAM J. Numer. Anal., 20 (1983), pp. 722–731.

[8] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer Ser. Comput. Math. 15, Springer-Verlag, New York, 1991.

[9] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North–Holland, Amsterdam, 1978.

[10] M. Crouzeix and R. S. Falk, *Nonconforming finite elements for the Stokes problem*, Math. Comp., 52 (1989), pp. 437–456.

[11] M. Crouzeix and P. A. Raviart, *Conforming and nonconforming finite elements for solving the stationary Stokes equations* I, Rev. Française Automat. Informat. Recherche Opérationnelle Sér. Rouge, 7 (1973), pp. 33–75.

[12] M. Fortin, *An analysis of the convergence of mixed finite element methods*, RAIRO Anal. Numér., 11 (1977), pp. 341–354.

[13] V. Girault and P.-A. Raviart, *Finite Element Methods for Navier–Stokes Equations*, Springer-Verlag, Berlin, 1986.

[14] J. P. Hennart, J. Jaffré, and J. E. Roberts, *A constructive method for deriving finite elements of nodal type*, Numer. Math., 53 (1988), pp. 701–738.

[15] V. Heuveline and F. Schieweck, *On the inf-sup condition for higher order mixed fem on meshes with hanging nodes*, Math. Model. Numer. Anal., 41 (2007), pp. 1–20.

[16] V. John, *Higher order finite element methods and multigrid solvers in a benchmark problem for the 3d Navier-Stokes equations*, Internat. J. Numer. Methods Fluids, 40 (2002), pp. 775–798.

[17] V. John and G. Matthies, *Higher order finite element discretizations in a benchmark problem for incompressible flows*, Internat. J. Numer. Methods Fluids, 37 (2001), pp. 885–903.

[18] V. John and G. Matthies, *MooNMD—a program package based on mapped finite element methods*, Comput. Vis. Sci., 6 (2004), pp. 163–170.

[19] P. Knobloch, *On the application of the P_1^{mod} element to incompressible flow problems*, Comput. Vis. Sci., 6 (2004), pp. 185–195.

[20] P. Knobloch, *On the inf-sup condition for the $P_3^{\mathrm{mod}}/P_2^{\mathrm{disc}}$ element*, Computing, 76 (2006), pp. 41–54.

[21] G. Matthies, *Inf-sup stable nonconforming finite elements of higher order on quadrilaterals and hexahedra*, M2AN Math. Model. Numer. Anal., 41 (2007), pp. 855–874.

[22] G. Matthies and L. Tobiska, *Inf-sup stable non-conforming finite elements of arbitrary order on triangles*, Numer. Math., 102 (2005), pp. 293–309.

[23] J. Maubach and P. Rabier, *Nonconforming Finite Elements of Arbitrary Degree over Triangles*, RANA report 0328, Technical University of Eindhoven, Eindhoven, The Netherlands, 2003.

[24] R. Rannacher and S. Turek, *Simple nonconforming quadrilateral Stokes element*, Numer. Methods Partial Differential Equations, 8 (1992), pp. 97–111.

[25] D. Schötzau, C. Schwab, and R. Stenberg, *Mixed hp-FEM on anisotropic meshes.* II. *Hanging nodes and tensor products of boundary layer meshes*, Numer. Math., 83 (1999), pp. 667–697.

[26] M. Stynes and L. Tobiska, *Using rectangular Q_p elements in the SDFEM for a convection-diffusion problem with a boundary layer*, Preprint 06-08, Fakultät für Mathematik, Otto-von-Guericke-Universität Magdeburg, Magdeburg, Germany, 2006. Appl. Numer. Math., to appear.

SIAM J. NUMER. ANAL.
Vol. 46, No. 4, pp. 1892–1911
© 2008 Society for Industrial and Applied Mathematics

ON USING APPROXIMATE FINITE DIFFERENCES IN MATRIX-FREE NEWTON–KRYLOV METHODS*

PETER N. BROWN†, HOMER F. WALKER‡, REBECCA WASYK‡, AND CAROL S. WOODWARD†

Abstract. A Newton–Krylov method is an implementation of Newton's method in which a Krylov subspace method is used to solve approximately the linear systems that characterize steps of Newton's method. Newton–Krylov methods are often implemented in "matrix-free" form, in which the Jacobian-vector products required by the Krylov solver are approximated by finite differences. Here we consider using approximate function values in these finite differences. We first formulate a finite-difference Arnoldi process that uses approximate function values. We then outline a Newton–Krylov method that uses an implementation of the GMRES or Arnoldi method based on this process, and we develop a local convergence analysis for it, giving sufficient conditions on the approximate function values for desirable local convergence properties to hold. We conclude with numerical experiments involving particular function-value approximations suitable for nonlinear diffusion problems. For this case, conditions are given for meeting the convergence assumptions for both lagging and linearizing the nonlinearity in the function evaluation.

Key words. matrix-free Newton–Krylov methods, Newton–GMRES methods, Newton–Arnoldi methods, Newton's method, Krylov subspace methods, GMRES method, Arnoldi method

AMS subject classifications. 65H10, 65F10

DOI. 10.1137/060652749

1. Introduction.

The problem of interest is to determine $u_* \in \mathbb{R}^n$ satisfying a system of nonlinear equations

$$F(u_*) = 0, \tag{1.1}$$

where $F : \mathbb{R}^n \to \mathbb{R}^n$ is continuously differentiable. A classical algorithm for approximately solving (1.1) is *Newton's method*, which, at a current approximate solution u, generates a next approximate solution $u + s$ through a step s that satisfies the linear *Newton equation*

$$F'(u)s = -F(u). \tag{1.2}$$

Our interest here is in *Newton–Krylov methods* (cf. [2], [11], [12], [15]), in which *Krylov subspace methods* are used to solve (1.2) approximately. An extensive introduction to Krylov subspace methods is beyond the scope of this paper; we touch on only the most relevant aspects here and refer the reader to [8], [9], [18], [19] for full treatments.

*Received by the editors February 23, 2006; accepted for publication (in revised form) September 25, 2007; published electronically April 18, 2008.
http://www.siam.org/journals/sinum/46-4/65274.html

†Center for Applied Scientific Computing, Lawrence Livermore National Laboratory, Box 808, L-561, Livermore, CA 94551 (pnbrown@llnl.gov, cswoodward@llnl.gov). The work of these authors was performed under the auspices of the U.S. Department of Energy by University of California, Lawrence Livermore National Laboratory under contract W-7405-ENG-48.

‡Department of Mathematical Sciences, Worcester Polytechnic Institute, Worcester, MA 01609-2280 (walker@wpi.edu, wasykrd@wpi.edu). The work of these authors was supported in part by the Center for Simulation of Accidental Fires and Explosions funded at the University of Utah by the U.S. Department of Energy under contracts LLNL B341493 and B524196.

A Krylov subspace method applied to a general linear system $Ax = b$, $A \in \mathbb{R}^{n\times n}$ invertible, begins with an initial approximate solution x_0 and, at the jth iteration, determines $x_j = x_0 + z_j$ through a correction z_j in the jth *Krylov subspace*

$$\mathcal{K}_j \equiv \text{span}\,\{r_0, Ar_0, \ldots, A^{j-1}r_0\}, \tag{1.3}$$

where $r_0 = b - Ax_0$. Specific methods are determined by the choice of $z_j \in \mathcal{K}_j$.

We focus here on two Krylov subspace methods: the *generalized minimal residual method* (GMRES) [16], in which each z_j is chosen to minimize the residual norm over all corrections in $\mathcal{K}_j$, and the *Arnoldi method* (also known as the *full orthogonalization method*) [17], in which each z_j is chosen to make the residual orthogonal to $\mathcal{K}_j$, if that is possible. (We consider these methods in only their basic forms and do not consider "restarted" or "truncated" variants.) These have in common that their iterates are generated directly from the *Arnoldi process* [1], which we review in section 2.1.

In the Newton–Krylov context, a particular advantage of the most widely used Krylov subspace methods, including the GMRES and Arnoldi methods, is that they require only products of F' with vectors for their implementation. Thus they allow "matrix-free" Newton–Krylov formulations (cf. [12]), in which these products are evaluated or approximated without creating or storing F'. Perhaps the most popular way of approximating these products is with a first-order forward difference

$$F'(u)v \approx \frac{F(u+\sigma v) - F(u)}{\sigma}, \tag{1.4}$$

where σ is an appropriately chosen difference step.

Each approximation (1.4) requires a new F-evaluation. If F-evaluations are expensive or if many iterations of the Krylov solver are required, then the aggregate cost of these evaluations over all Krylov iterations may be undesirably high. In this case, it is natural to consider using an inexpensive approximation of $F(u+\sigma v)$ in (1.4), if one is available.

In the following, we explore this possibility. Our goal is to lay out a framework that will provide guidelines for using such approximations, specifically to outline conditions on such approximations that lead to desirable convergence properties of the resulting Newton–Krylov methods. Although we consider only the GMRES and Arnoldi methods here, it is our expectation that the results will provide useful guidance when other Krylov subspace methods are used as well.

In section 2, we first describe the approximations of F that we allow in difference quotients and formulate a finite-difference Arnoldi process that uses them. We then outline a Newton–Krylov method that uses GMRES or Arnoldi implementations based on this process, and we develop a local convergence analysis for it. In section 3, we discuss numerical experiments involving two illustrative approximations of F that are appropriate for a broad class of nonlinear diffusion problems. We also show how these approximations can be chosen to satisfy the convergence assumptions in section 2. We offer a summary discussion and conclusions in section 4.

This work builds on and extends work of [3]. We assume throughout that F is continuously differentiable on domains of interest and, for convenience, that $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$ are the Euclidean norm and inner-product on $\mathbb{R}^n$, respectively. (All results can be easily extended to the case of an arbitrary inner-product norm on $\mathbb{R}^n$.) For $u \in \mathbb{R}^n$ and $\delta > 0$, we denote $B(u,\delta) \equiv \{v \in \mathbb{R}^n : \|v-u\| < \delta\}$. We implicitly assume that arithmetic is exact and do not consider errors due to rounding in finite-precision arithmetic.

2. The approximate finite-difference framework. We assume that there is a function $\tilde{F} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ such that $\tilde{F}(u,u) = F(u)$ for all $u \in \mathbb{R}^n$. The presumption is that, in an application of interest, one can determine $\tilde{F}$ so that, for each u and w near u, evaluating $\tilde{F}(u,w)$ is preferred over evaluating $F(w)$. Thus, in lieu of (1.4), we consider approximations of the form

$$F'(u)v \approx \frac{\tilde{F}(u, u+\sigma v) - F(u)}{\sigma}. \tag{2.1}$$

We make further assumptions on $\tilde{F}$ after outlining in section 2.1 a version of the Arnoldi process that uses these approximations.

2.1. The approximate finite-difference Arnoldi process. We begin by recalling the usual Arnoldi process for a general linear system $Ax = b$, $A \in \mathbb{R}^{n\times n}$ invertible, and initial $x_0 \in \mathbb{R}^n$.

ALGORITHM 2.1: ARNOLDI PROCESS

GIVEN x_0, SET $r_0 = b - Ax_0$.

IF $r_0 = 0$, STOP; OTHERWISE SET $v_1 = r_0/\|r_0\|$.

FOR $j = 1, 2, \ldots,$ DO:

SET $w_j = Av_j - \sum_{i=1}^{j} h_{ij} v_i$, WHERE $h_{ij} = \langle Av_j, v_i \rangle$.

IF $h_{j+1,j} \equiv \|w_j\| = 0$, STOP; OTHERWISE, SET

$v_{j+1} = w_j/h_{j+1,j}$ AND CONTINUE.

If $r_0 \neq 0$, then the process terminates at step $j > 0$ if and only if $h_{j+1,j} = 0$, which occurs if and only if $\mathcal{K}_j$ is an invariant subspace of A. Thus the process must terminate for some $j \leq n$. For each j up to termination, $\{v_i\}_{i=1,\ldots,j}$ constitutes an orthonormal basis of $\mathcal{K}_j$. After j steps, particular products of the process are the matrices $V_j \equiv (v_1, \ldots, v_j)$ and upper-Hessenberg

$$H_j \equiv \begin{pmatrix} h_{11} & \cdots & h_{1j} \\ h_{21} & \cdots & h_{2j} \\ & \ddots & \vdots \\ & & h_{j+1,j} \end{pmatrix}.$$

The relationship between V_j and H_j is characterized by

$$AV_j = \begin{cases} V_{j+1}H_j & \text{if } h_{j+1,j} \neq 0, \\ V_j \bar{H}_j & \text{if } h_{j+1,j} = 0, \end{cases}$$

where $\bar{H}_j$ is obtained from H_j by deleting the $(j+1)$st row.

The iterates of the GMRES and Arnoldi methods applied to $Ax = b$ are uniquely determined by the matrices V_j and H_j. In the case of the GMRES method, an approximate solution is defined at each iteration; in the case of the Arnoldi method, an approximate solution may not be defined at every iteration, depending on A and b. However, for both methods, with A invertible, the approximate solution is defined and equal to the exact solution at some step if and only if Algorithm 2.1 terminates at that step. In both methods, the approximate solution is computed only upon termination. During the intermediate iterations, the value of the residual norm is maintained recursively without computing the approximate solution, and termination is based on this value, which we refer to below as the *recursive residual norm*.

To formulate our approximate finite-difference Arnoldi process, we consider a Jacobian system

$$F'(u)x = c \tag{2.2}$$

and outline a process analogous to Algorithm 2.1 that uses approximations of the form (2.1). In this, we use "~" to denote approximate counterparts of quantities in Algorithm 2.1 that are derived from $\tilde{F}$.

ALGORITHM 2.2: APPROXIMATE FINITE-DIFFERENCE ARNOLDI PROCESS

GIVEN x_0, SET $q_0 = \frac{\tilde{F}(u,u+\sigma_0 x_0)-F(u)}{\sigma_0}$ AND $\tilde{r}_0 = c - q_0$.
IF $\tilde{r}_0 = 0$, STOP; OTHERWISE, SET $\tilde{v}_1 = \tilde{r}_0/\|\tilde{r}_0\|$.
FOR $j = 1, 2, \ldots,$ DO:
SET $q_j = \frac{\tilde{F}(u,u+\sigma_j \tilde{v}_j)-F(u)}{\sigma_j}$ AND $\tilde{w}_j = q_j - \sum_{i=1}^{j} \tilde{h}_{ij}\tilde{v}_i$,
WHERE $\tilde{h}_{ij} = \langle q_j, \tilde{v}_i \rangle$.
IF $\tilde{h}_{j+1,j} \equiv \|\tilde{w}_j\| = 0$, STOP; OTHERWISE, SET
$\tilde{v}_{j+1} = \tilde{w}_j/\tilde{h}_{j+1,j}$ AND CONTINUE.

If $\tilde{r}_0 \neq 0$, then the process terminates at step $j > 0$ if and only if $\tilde{h}_{j+1,j} = 0$. For each j up to termination, $\{\tilde{v}_i\}_{i=1,\ldots,j}$ is an orthonormal set; hence, termination must occur for some $j \leq n$.

The GMRES and Arnoldi methods applied to (2.2) can be based on Algorithm 2.2 in an obvious way. In particular, for each j, we define $\tilde{V}_j \equiv (\tilde{v}_1, \ldots, \tilde{v}_j)$ and upper-Hessenberg

$$\tilde{H}_j \equiv \begin{pmatrix} \tilde{h}_{11} & \cdots & \tilde{h}_{1j} \\ \tilde{h}_{21} & \cdots & \tilde{h}_{2j} \\ & \ddots & \vdots \\ & & \tilde{h}_{j+1,j} \end{pmatrix}.$$

Then, like their counterparts V_j and H_j produced by Algorithm 2.1, the matrices $\tilde{V}_j$ and $\tilde{H}_j$ uniquely determine the iterates of the GMRES and Arnoldi methods based on Algorithm 2.2. As before, the GMRES approximate solution is defined at every iteration, but the Arnoldi approximate solution may not be.

We further define

$$\begin{aligned} e_0 &\equiv q_0 - F'(u)x_0, \\ e_i &\equiv q_i - F'(u)\tilde{v}_i, \quad i = 1, \ldots, j. \end{aligned} \tag{2.3}$$

With the orthonormality of $\{\tilde{v}_i\}_{i=1,\ldots,j}$, one easily verifies that

$$\{F'(u) + e_i \tilde{v}_i^T\}\tilde{v}_i = q_i, \quad i = 1, \ldots, j,$$

and, therefore,

$$\{F'(u) + E_j\}\tilde{V}_j = (q_1, \ldots, q_j) = \begin{cases} \tilde{V}_{j+1}\tilde{H}_j & \text{if } \tilde{h}_{j+1,j} \neq 0, \\ \tilde{V}_j \bar{\tilde{H}}_j & \text{if } \tilde{h}_{j+1,j} = 0, \end{cases} \tag{2.4}$$

where

$$E_j \equiv (e_1, \ldots, e_j)\tilde{V}_j^T \tag{2.5}$$

and $\bar{\tilde{H}}_j \in \mathbb{R}^{j \times j}$ is obtained from $\tilde{H}_j$ by deleting the $(j+1)$st row.

In view of (2.3)–(2.5), the following theorem is immediate.

THEOREM 2.3. *Suppose that, for some $j \in \{1, \ldots, n\}$, $\tilde{V}_j$ and $\tilde{H}_j$ have been produced by j steps of Algorithm 2.2 with an initial x_0. Then $\tilde{V}_j$ and $\tilde{H}_j$ are the same*

as, respectively, V_j and H_j produced by j steps of Algorithm 2.1 *with the same initial x_0 and with $A = F'(u) + E_j$ and $b = c - e_0 + E_j x_0$, where E_j and e_0 are defined by* (2.3) *and* (2.5).

Proof. In Algorithm 2.2, one has $\tilde{v}_1 = \tilde{r}_0/\|\tilde{r}_0\|$ and, with (2.3),

$$\tilde{r}_0 = c - q_0 = c - e_0 + E_j x_0 - (F'(u) + E_j)\, x_0.$$

The theorem follows from this and (2.4). ∎

COROLLARY 2.4. *Suppose that, for some $j \in \{1, \ldots, n\}$, the GMRES or Arnoldi method based on Algorithm* 2.2 *has been applied to* (2.2) *for j steps. If $F'(u) + E_j$ is invertible, then the approximate solution determined at step j by the method (if defined) is the same as the approximate solution of the perturbed system*

$$\{F'(u) + E_j\}x = c - e_0 + E_j x_0 \tag{2.6}$$

determined at step j by the same method based on Algorithm 2.1. *The approximate solution at step j is defined and is the exact solution of* (2.6) *if and only if Algorithm* 2.2 *terminates at step j.*

2.2. A local convergence analysis. Newton–Krylov methods are usually implemented as *inexact Newton methods* [5], in which the Newton equation (1.2) is replaced with an *inexact Newton condition*

$$\|F(u) + F'(u)s\| \le \eta\|F(u)\|, \tag{2.7}$$

where $\eta \in [0, 1)$. In a Newton–Krylov implementation, (2.7) normally provides a criterion for terminating the Krylov iterations: one first chooses $\eta \in [0, 1)$ and then applies the Krylov solver to (1.2) until (2.7) holds. Used in this way, η is often called a *forcing term* (cf. [7]). In the seminal paper [5], it is shown that the local convergence of inexact Newton methods is determined by the forcing terms.

In the context of interest here, the Krylov solver is the GMRES or Arnoldi method based on Algorithm 2.2, and the approximate solution it produces is actually that of the perturbed linear system (2.6) with $c = -F(u)$. Thus an accurate evaluation of the linear residual norm $\|F(u) + F'(u)s\|$ is very likely to be undesirably expensive or unavailable. If this is the case, then (2.7) cannot be used directly for terminating the Krylov iterations. A likely course is to replace $\|F(u) + F'(u)s\|$ in (2.7) by the recursive residual norm maintained by the Krylov solver. This leads to the following algorithm.

ALGORITHM 2.5: APPROXIMATE FINITE-DIFFERENCE GMRES/ARNOLDI NEWTON–KRYLOV METHOD

SUPPOSE THAT u_0 IS GIVEN.

FOR $k = 0, 1, \ldots,$ DO:

CHOOSE $\eta_k \in [0, 1)$ AND APPLY THE GMRES OR ARNOLDI METHOD BASED ON ALGORITHM 2.2 TO $F'(u_k)s = -F(u_k)$ TO OBTAIN AN APPROXIMATE SOLUTION s_k FOR WHICH THE RECURSIVE RESIDUAL NORM IS LESS THAN OR EQUAL TO $\eta_k\|F(u_k)\|$.

SET $u_{k+1} = u_k + s_k$.

In the remainder of this subsection, we develop a local convergence analysis for Algorithm 2.5. This culminates in our main result (Theorem 2.12) to the effect that, near a solution of (1.2) at which the Jacobian is invertible, Algorithm 2.5 does not break down and enjoys desirable convergence properties if the difference steps are chosen sufficiently small and the forcing terms are chosen appropriately.

We begin with Lemma 2.6 below, which shows that, if the recursive residual norm satisfies a specified inexact Newton condition, then the actual residual norm satisfies a related inexact Newton condition. This lemma is a perturbation result based on Corollary 2.4. With $c = -F(u)$, the perturbed linear system (2.6) can be viewed as a specific instance of a more general perturbed linear system considered by Catinas in [4]. Perturbation results are given in [4] that are somewhat similar in spirit to Lemma 2.6 but are not as directly suited to our purposes.

For convenience in Lemma 2.6 and that which follows, we regard the initial residual norm as the recursive residual norm at step zero. Throughout the discussion, e_0 and E_j for $j > 0$ are defined by (2.3) and (2.5), with x_0 the initial approximate solution in the Krylov solver and with $c = -F(u)$ in Algorithm 2.2.

LEMMA 2.6. *Suppose that $\eta \in [0, 1)$ is given and that the GMRES or Arnoldi method based on Algorithm* 2.2 *is applied to* (1.2) *until, for some $j \in \{0, \ldots, n\}$, the recursive residual norm is less than or equal to $\eta\|F(u)\|$ at step j. Let s denote the approximate solution of* (1.2) *obtained by the method at step j. If $j = 0$, or if $j > 0$ and $F'(u) + E_j$ is invertible, then*

$$\|F(u) + F'(u)s\| \le \tilde{\eta}\|F(u)\|, \tag{2.8}$$

where

$$\tilde{\eta} \equiv \begin{cases} \eta + \dfrac{\|e_0\|}{\|F(u)\|} & \text{if } j = 0, \\ \eta + (1+\eta)\|E_j(F'(u)+E_j)^{-1}\| & \\ \qquad + \left[1 + \|E_j(F'(u)+E_j)^{-1}\|\right] \dfrac{\|e_0 - E_j\, x_0\|}{\|F(u)\|} & \text{if } j > 0. \end{cases} \tag{2.9}$$

Remark. If the initial approximate solution x_0 in the GMRES or Arnoldi method is zero, as is usually the case in Newton–Krylov methods, then $e_0 = 0$ and (2.9) reduces to

$$\tilde{\eta} \equiv \begin{cases} \eta & \text{if } j = 0, \\ \eta + (1+\eta)\|E_j(F'(u)+E_j)^{-1}\| & \text{if } j > 0. \end{cases} \tag{2.10}$$

Proof. If $j = 0$, then $s = x_0$ in Algorithm 2.2, and, with q_0 as in Algorithm 2.2, we have

$$\begin{aligned} \|F(u) + F'(u)s\| &\le \|F(u) + q_0\| + \|F'(u)s - q_0\| \\ &\le \eta\|F(u)\| + \|e_0\| \\ &= \left(\eta + \frac{\|e_0\|}{\|F(u)\|}\right)\|F(u)\|. \end{aligned}$$

If $j > 0$, then, by using Corollary 2.4 with $c = -F(u)$, we have

$$\begin{aligned} \|F(u) + F'(u)s\| &\le \|F(u) + e_0 - E_j\, x_0 + (F'(u) + E_j)\, s\| \\ &\qquad + \|E_j\, s\| + \|e_0 - E_j\, x_0\| \\ &\le \eta\|F(u)\| + \|E_j(F'(u)+E_j)^{-1}\| \cdot \|(F'(u) + E_j)\, s\| \\ &\qquad + \|e_0 - E_j\, x_0\|. \end{aligned} \tag{2.11}$$

Writing

$$(F'(u)+E_j)s = -F(u) - e_0 + E_j x_0 + r, \quad \|r\| \le \eta\|F(u)\|,$$

one verifies that

$$\begin{aligned}\|(F'(u)+E_j)s\| &\le (1+\eta)\|F(u)\| + \|e_0 - E_j\, x_0\| \\ &= \left(1+\eta+\frac{\|e_0 - E_j\, x_0\|}{\|F(u)\|}\right)\|F(u)\|.\end{aligned}$$

Substituting this in (2.11) and rearranging gives

$$\begin{aligned}\|F(u)+F'(u)s\| &\le \left\{\eta + \|E_j(F'(u)+E_j)^{-1}\|\cdot\left(1+\eta+\frac{\|e_0 - E_j\, x_0\|}{\|F(u)\|}\right)\right. \\ &\qquad \left. + \frac{\|e_0 - E_j\, x_0\|}{\|F(u)\|}\right\}\|F(u)\| \\ &= \left\{\eta + (1+\eta)\|E_j(F'(u)+E_j)^{-1}\|\right. \\ &\qquad \left. + \left[1+\|E_j(F'(u)+E_j)^{-1}\|\right]\frac{\|e_0 - E_j\, x_0\|}{\|F(u)\|}\right\}\|F(u)\|. \qquad \square\end{aligned}$$

We now formulate the first of two additional assumptions on $\tilde{F}$ and establish Lemma 2.8 and Corollary 2.9 below, which will be helpful in developing our local convergence results and also in justifying the assumptions in Lemma 2.6 near a solution of (1.2) at which the Jacobian is invertible.

Assumption 2.7. There exists $\Omega \subseteq \mathbb{R}^n$ such that

$$\omega(\sigma) \equiv \sup_{u\in\Omega,\ \|v\|\le 1}\left\|\frac{\tilde{F}(u,u+\sigma v)-F(u)}{\sigma} - F'(u)v\right\| \tag{2.12}$$

satisfies $\omega_* \equiv \limsup_{\sigma\to 0}\omega(\sigma) < \infty$.

The role of Assumption 2.7 in that which follows is to ensure that approximations of the form (2.1) are uniformly accurate when σ is appropriately small. We note that, if this assumption holds with $\omega_* = 0$, then the Frechet derivative $\tilde{F}_w(u,w) \equiv \frac{\partial \tilde{F}(u,w)}{\partial w}$ exists for $u \in \Omega$ and $w = u$ and, moreover, $\tilde{F}_w(u,u) = F'(u)$ for $u \in \Omega$.

LEMMA 2.8. *Suppose that Assumption* 2.7 *holds and that* $u_* \in \Omega$ *is such that* $F'(u_*)$ *is invertible and*

$$n\|F'(u_*)^{-1}\|\omega_* < 1/2. \tag{2.13}$$

Then, for any ϵ *such that*

$$\frac{n\|F'(u_*)^{-1}\|\omega_*}{1-n\|F'(u_*)^{-1}\|\omega_*} < \epsilon < 1, \tag{2.14}$$

there exist $\delta > 0$ *and* $\sigma_* > 0$ *such that if* $u \in B(u_*,\delta)$ *and, for some* $j \in \{1,\dots,n\}$, *Algorithm* 2.2 *has been applied with* $0 < |\sigma_i| \le \sigma_*$ *for* $i = 1, \dots, j$, *then* $F'(u)+E_j$ *is invertible, and*

$$\|E_j(F'(u)+E_j)^{-1}\| \le \epsilon. \tag{2.15}$$

Moreover, there is a λ independent of j and u for which

$$\|E_j(F'(u)+E_j)^{-1}\| \le \lambda \max_{1\le i\le j}\omega(\sigma_i)\,. \tag{2.16}$$

Proof. Since F is continuously differentiable and $F'(u_*)$ is invertible, there exists $\delta>0$ such that $B(u_*,\delta)\subset\Omega$, $F'(u)$ is invertible whenever $u\in B(u_*,\delta)$, and $\sup_{u\in B(u_*,\delta)}\|F'(u)^{-1}\|<\infty$. Moreover, in view of (2.13) and (2.14), we can choose δ and $\sigma_*>0$ sufficiently small to have $n\|F'(u)^{-1}\|\omega(\sigma)\le 1/2$ and

$$\frac{n\|F'(u)^{-1}\|\omega(\sigma)}{1-n\|F'(u)^{-1}\|\omega(\sigma)}\le\epsilon \tag{2.17}$$

whenever $u\in B(u_*,\delta)$ and $0<|\sigma|\le\sigma_*$. Note that, for $0<|\sigma|\le\sigma_*$, (2.17) implies that

$$\omega(\sigma)\le\frac{\epsilon}{(1+\epsilon)n\|F'(u)^{-1}\|}. \tag{2.18}$$

Suppose that $u\in B(u_*,\delta)$ and that Algorithm 2.2 has been applied with $0<|\sigma_i|\le\sigma_*$ for $i=1,\ldots,j$. Then

$$\begin{aligned}\|E_j\| &= \max_{\|v\|=1}\|E_jv\| = \max_{\|v\|=1}\|(e_1,\ldots,e_j)\tilde V_j^Tv\|\\ &= \max_{\|v\|=1}\left\|\sum_{i=1}^j\langle\tilde v_i,v\rangle\,e_i\right\| \le \max_{\|v\|=1}\sum_{i=1}^j|\,\langle\tilde v_i,v\rangle\,|\,\|e_i\|\\ &\le \textstyle\sum_{i=1}^j\|e_i\|\end{aligned} \tag{2.19}$$

since $|\,\langle\tilde v_i,v\rangle\,|\le\|\tilde v_i\|\,\|v\|=1$. Since

$$\|e_i\| = \|q_i-F'(u)\tilde v_i\| = \left\|\frac{\tilde F(u,u+\sigma_i\tilde v_i)-F(u)}{\sigma_i}-F'(u)\tilde v_i\right\|\le\omega(\sigma_i)$$

for each i, it follows from (2.19) and (2.18) that

$$\|E_j\|\le j\max_{1\le i\le j}\omega(\sigma_i)\le n\max_{1\le i\le j}\omega(\sigma_i)\le\frac{\epsilon}{(1+\epsilon)\|F'(u)^{-1}\|}. \tag{2.20}$$

From (2.20), we have

$$\|E_jF'(u)^{-1}\|\le\|E_j\|\,\|F'(u)^{-1}\|\le\frac{\epsilon}{1+\epsilon}<1, \tag{2.21}$$

and it follows by using well-known arguments (see, e.g., [6, Thm. 3.1.4]) that $I+E_jF'(u)^{-1}$ is invertible and

$$\left\|\left(I+E_jF'(u)^{-1}\right)^{-1}\right\|\le\frac{1}{1-\epsilon/(1+\epsilon)}=1+\epsilon. \tag{2.22}$$

Consequently, $F'(u)+E_j=(I+E_jF'(u)^{-1})F'(u)$ is invertible, and

$$\begin{aligned}\left\|E_j(F'(u)+E_j)^{-1}\right\| &= \left\|E_jF'(u)^{-1}\left(I+E_jF'(u)^{-1}\right)^{-1}\right\|\\ &\le \|E_jF'(u)^{-1}\|\,\|\left(I+E_jF'(u)^{-1}\right)^{-1}\|\\ &\le \frac{\epsilon}{1+\epsilon}\cdot(1+\epsilon)=\epsilon\,.\end{aligned} \tag{2.23}$$

To complete the proof, one easily verifies from the inequalities in (2.18)–(2.23) that

$$\|E_j(F'(u)+E_j)^{-1}\| \le n \max_{1\le i\le j} \omega(\sigma_i)\, \|F'(u)^{-1}\|\, (1+\epsilon)\,.$$

It follows that (2.16) holds with $\lambda = n \sup_{u\in B(u_*,\delta)} \|F'(u)^{-1}\|\,(1+\epsilon)$. ∎

COROLLARY 2.9. *Suppose that Assumption* 2.7 *holds and that* $u_* \in \Omega$ *is such that* $F'(u_*)$ *is invertible and* $F'(u_*)^{-1}$ *satisfies* (2.13). *Then there exist* $\delta > 0$ *and* $\sigma_* > 0$ *such that if* $u \in B(u_*, \delta)$ *and the GMRES or Arnoldi method based on Algorithm* 2.2 *is applied to* (1.2) *with* $0 < |\sigma_j| \le \sigma_*$ *for each* j, *then for some* $j \in \{0, \dots, n\}$ *the approximate solution obtained by the method at step* j *is defined and the recursive residual norm is zero. In particular, given any* $\eta \in [0, 1)$, *the method determines at some step an approximate solution for which the recursive residual norm is less than or equal to* $\eta\|F(u)\|$.

Proof. Suppose that the GMRES or Arnoldi method based on Algorithm 2.2 is applied to (1.2). If $\tilde{r}_0 = 0$, then there is nothing to prove. If $\tilde{r}_0 \neq 0$, then Algorithm 2.2 terminates for some $j \in \{1, \dots, n\}$. By Lemma 2.8, there exist $\delta > 0$ and $\sigma_* > 0$ such that if $u \in B(u_*, \delta)$ and $0 < |\sigma_i| \le \sigma_*$ for $i = 1, \dots, j$, then $F'(u) + E_j$ is invertible. Consequently, assuming $u \in B(u_*, \delta)$ and $0 < |\sigma_i| \le \sigma_*$ for $i = 1, \dots, j$, one has from Corollary 2.4 that the approximate solution determined by the method at step j is defined and is the exact solution of the perturbed system

$$\{F'(u) + E_j\}x = -F(u) - e_0 + E_j x_0. \tag{2.24}$$

Since the recursive residual norm is the norm of the residual of (2.24), it follows that the recursive residual norm is zero. ∎

We use the following somewhat specialized assumption of Hölder continuity of $\tilde{F}_w$. For a general definition of Hölder continuity, see, e.g., [14, section 3.1.9].

Assumption 2.10. With Ω as in Assumption 2.7, there is a $\delta_\Omega > 0$ such that $\tilde{F}_w(u, u+v)$ exists for all $u \in \Omega$ and all v with $\|v\| \le \delta_\Omega$. Moreover, there exist γ and $p \in (0, 1]$ such that

$$\|\tilde{F}_w(u, u+v) - \tilde{F}_w(u, u)\| \le \gamma \|v\|^p$$

for all $u \in \Omega$ and all v with $\|v\| \le \delta_\Omega$.

LEMMA 2.11. *Suppose that Assumption* 2.7 *holds with* $\omega_* = 0$ *and that Assumption* 2.10 *also holds. Then* $\omega(\sigma)$ *given by* (2.12) *satisfies*

$$\omega(\sigma) \le \mu\, |\sigma|^p \tag{2.25}$$

for sufficiently small $|\sigma|$ *and* $\mu = \gamma(\delta_\Omega)^{1+p}/(1+p)$.

Proof. Assume for convenience that $\sigma > 0$. (The proof requires only trivial changes when $\sigma < 0$.) By recalling that $\omega_* = 0$ implies $F'(u) = \tilde{F}_w(u, u)$ and adapting familiar reasoning (see, e.g., [14, section 3.2.12]), we have for $u \in \Omega$, $\|v\| \le \delta_\Omega$, and

sufficiently small $\sigma > 0$ that

$$\begin{aligned}\|\tilde{F}(u, u+\sigma v) - F(u) - \sigma F'(u)v\| &= \left\| \left\{ \int_0^\sigma [\tilde{F}_w(u, u+\tau v) - \tilde{F}_w(u,u)]\, d\tau \right\} v \right\| \\ &\le \left\{ \int_0^\sigma \|\tilde{F}_w(u, u+\tau v) - \tilde{F}_w(u,u)\|\, d\tau \right\} \|v\| \\ &\le \left\{ \int_0^\sigma \gamma \tau^p\, d\tau \right\} \|v\|^{1+p} \\ &= \frac{\gamma}{1+p}\sigma^{1+p}\|v\|^{1+p} \le \frac{\gamma(\delta_\Omega)^{1+p}}{1+p}\sigma^{1+p},\end{aligned}$$

which immediately yields (2.25). ∎

The theorem below is our local convergence result for Algorithm 2.5. It can be viewed as a counterpart of Theorems 2.3 and 3.3 of [5], which are used in its proof. For definitions of the types of convergence referred to in the theorem, see, e.g., [6]. For convenience, we denote by σ_{jk} the jth difference step used by the GMRES or Arnoldi method at the kth iteration of Algorithm 2.5. We also assume that, at each iteration of Algorithm 2.5, the GMRES or Arnoldi method begins with the zero initial approximate solution, and, therefore, the applicable perturbation result from Lemma 2.6 is (2.8) followed by (2.10) rather than (2.9).

THEOREM 2.12. *Suppose that Assumption* 2.7 *holds and that $u_* \in \Omega$ is such that $F(u_*) = 0$, $F'(u_*)$ is invertible, and $n\|F'(u_*)^{-1}\|\omega_* < 1/2$. Then for any t_* and η_* such that*

$$\mu_* \equiv \frac{n\|F'(u_*)^{-1}\|\omega_*}{1 - n\|F'(u_*)^{-1}\|\omega_*} < t_* < 1 \quad \text{and} \quad 0 \le \eta_* < \frac{t_* - \mu_*}{1+\mu_*}, \tag{2.26}$$

there exist $\delta_* > 0$ and $\sigma_* > 0$ such that if $u_0 \in B(u_*, \delta_*)$, $0 \le \eta_k \le \eta_*$ for each k, and $0 < |\sigma_{jk}| \le \sigma_*$ for each j and k, then Algorithm 2.5 *produces $\{u_k\}_{k=0,1,\ldots}$ that converges to u_* with*

$$\|u_{k+1} - u_*\|_* \le t_*\|u_k - u_*\|_* \tag{2.27}$$

for each k, where $\|v\|_* \equiv \|F'(u_*)v\|$ for $v \in \mathbb{R}^n$. Additionally, if $\omega_* = 0$ and if $\eta_k \to 0$ and $\max_j |\sigma_{jk}| \to 0$ as $k \to \infty$, then the convergence is q-superlinear. If Assumption 2.10 *also holds, F' is Hölder continuous with exponent p at u_*, and $\eta_k = O(\|F(u_k)\|^p)$ and $\max_j |\sigma_{jk}| = O(\|F(u_k)\|)$ as $k \to \infty$, then the convergence is of q-order $1+p$.*

Proof. Suppose that t_* and η_* satisfy (2.26). Let ϵ be such that $\mu_* < \epsilon < t_*$ and $\eta_* < (t_* - \epsilon)/(1+\epsilon)$. Note that $\eta_{\max} \equiv \eta_* + (1+\eta_*)\epsilon < t_*$. By [5, Thm. 2.3], there is a $\delta_* > 0$ such that any sequence $\{u_k\}_{k=0,1,\ldots}$ for which $u_0 \in B(u_*, \delta_*)$ and

$$\left.\begin{aligned} u_{k+1} &= u_k + s_k \\ \|F(u_k) + F'(u_k)s_k\| &\le \eta_{\max}\|F(u_k)\| \end{aligned}\right\} \quad k = 0, 1, \ldots$$

satisfies (2.27) for each k and converges to u_*.

Let $\delta > 0$ and $\sigma_* > 0$ be chosen so that the conclusions of Lemma 2.8 and Corollary 2.9 hold. Let M be such that

$$\frac{1}{M}\|v\| \le \|v\|_* \le M\|v\|$$

for all $v \in \mathbb{R}^n$. By taking δ_* smaller if necessary, assume that $M^2\delta_* < \delta$. Suppose that Algorithm 2.5 is applied with $u_0 \in B(u_*, \delta_*)$, $0 \le \eta_k \le \eta_*$ for all k, and $0 < |\sigma_{jk}| \le \sigma_*$ for all j and k. If, for some $k \ge 0$, the algorithm has determined $\{u_j\}_{j=0,\dots,k} \subseteq B(u_*, \delta)$, then, at step k of the algorithm, it follows from Corollary 2.9 that the GMRES or Arnoldi method based on Algorithm 2.2 determines at some iteration an approximate solution s_k for which the recursive residual norm is less than or equal to $\eta_k \|F(u_k)\|$. Then u_{k+1} is defined. Moreover, with Lemmas 2.8 and 2.6, we have, by using (2.8)–(2.10) and (2.15), that

$$\|F(u_k) + F'(u_k)s_k\| \le [\eta_k + (1+\eta_k)\epsilon]\, \|F(u_k)\| \le \eta_{\max}\|F(u_k)\|,$$

and we can apply (2.27) to obtain

$$\begin{aligned} \|u_{k+1} - u_*\| &\le M\|u_{k+1} - u_*\|_* \le M t_*^{k+1}\|u_0 - u_*\|_* \\ &\le M^2 t_*^{k+1}\|u_0 - u_*\| < M^2\delta_* < \delta. \end{aligned}$$

It follows inductively that Algorithm 2.2 produces $\{u_k\}_{k=0,1,\dots} \subseteq B(u_*, \delta)$ that converges to u_*, with (2.27) holding for each k.

Assuming $u_0 \in B(u_*, \delta_*)$, we also have from (2.8)–(2.10) and (2.16) that

$$\|F(u_k) + F'(u_k)s_k\| \le [\eta_k + (1+\eta_k)\,\lambda \max_j \omega(\sigma_{jk})]\, \|F(u_k)\| \tag{2.28}$$

for each k, where λ is independent of k. Consequently, if $\omega_* \equiv \lim_{\sigma\to 0}\omega(\sigma) = 0$ and if $\eta_k \to 0$ and $\max_j |\sigma_{jk}| \to 0$ as $k \to \infty$, then it follows from [5, Thm. 3.3] that $\{u_k\}_{k=0,1,\dots}$ converges to u_* q-superlinearly. If Assumption 2.10 also holds, then (2.28) and (2.25) yield

$$\|F(u_k) + F'(u_k)s_k\| \le [\eta_k + (1+\eta_k)\,\lambda\,\mu\,(\max_j |\sigma_{jk}|)^p]\, \|F(u_k)\|\,,$$

where μ is independent of k. It follows from [5, Thm. 3.3] that, if $\eta_k = O(\|F(u_k)\|^p)$ and $\max_j |\sigma_{jk}| = O(\|F(u_k)\|)$ as $k \to \infty$, then the convergence is of q-order $1+p$. ∎

Remark. The framework and analysis given above cover the case in which $F = F_1 + F_2$, where F_1 and F_2 are differentiable and such that, for each u and v, $F_1'(u)v$ can be readily evaluated but $F_2'(u)v$ must be approximated with a finite difference.[1] Indeed, setting $\tilde{F}(u, w) \equiv F_1(u) + F_1'(u)(w-u) + F_2(w)$, we obtain

$$\frac{\tilde{F}(u, u+\sigma v) - F(u)}{\sigma} = F_1'(u)v + \frac{F_2(u+\sigma v) - F_2(u)}{\sigma}.$$

Then Assumption 2.7 becomes

$$\omega_* \equiv \sup_{u\in\Omega,\ \|v\|\le 1} \left\| \frac{F_2(u+\sigma v) - F_2(u)}{\sigma} - F_2'(u)v \right\| < \infty,$$

and Assumption 2.10 becomes an assumption of Hölder continuity of $F_2'(u)$. With this understanding, it is straightforward to apply Corollary 2.9 and Theorem 2.12.

3. Numerical experiments. In this section we first describe the test problems considered. We then focus on two possible choices for $\tilde{F}$ and analyze to what degree they satisfy Assumptions 2.7 and 2.10. Finally, we present numerical results obtained with these choices on the test problems.

[1] This case can also be addressed with a modest extension of the results in [3].

TABLE 1
Test cases.

Test case	$D(u)$
1	$\sqrt{u^2+u+1}$
2	$\frac{1}{\sqrt{u^2+u+1}}$
3	$u^{3/5}e^u$
4	$K\sqrt{S(u)}[1-(1-S(u)^{\frac{1}{\mu}})^{\mu}]^2$, where $S(u)=(1+\lvert\alpha u\rvert^{\nu})^{-\mu}$ and K,α,μ,ν are constants

3.1. Test problems. The test problems are of the form

$$\mathcal{F}(u) \equiv \nabla \cdot D(u)\nabla u - g(u) + f = 0. \tag{3.1}$$

This class of nonlinear diffusion problems arises in many applications, such as flow through porous media, radiation transport, phase transition, biochemistry, and dynamics of biological groups. In our experiments, we considered these problems on square domains in $\mathbb{R}^2$ with homogeneous Dirichlet boundary conditions.

We are motivated by circumstances in which g-evaluations are inexpensive, but D-evaluations are relatively costly, and less expensive approximations may be desirable. In our experiments, we considered four choices for D, labeled test cases 1–4 in Table 1. For cases 1–3, we chose $g(u) \equiv u^2$. In case 4, we used $g(u) \equiv 0$. Except in test case 4, we took the domain to be $[0,1] \times [0,1]$ and chose f so that the solution of (3.1) was $u_c(x,y) \equiv cx(1-x)y(1-y)$, with $c = 1$, 2, 5, and 10. Choosing f in this way allowed us to determine success or failure with certainty and also to exercise some control over the magnitude of the solution and its derivatives. Test case 4 was inspired by a formulation of Richards's equation describing fluid flow in unsaturated porous media that appeared in [13]. In this case, the domain was $[0, \frac{1}{2}] \times [0, \frac{1}{2}]$ and f was chosen so that the solution was $u_c(x,y) \equiv \frac{c}{4}x(1-2x)y(1-2y) + \beta$, where $\beta > 0$ is a small constant chosen to avoid singularities ($\beta = \frac{1}{16}$ in the experiments reported here) and again $c = 1$, 2, 5, and 10.

In our tests, (3.1) was discretized by using central differences on an $m \times m$ grid of regularly spaced interior points with grid spacing $h = 1/(m+1)$ in test cases 1–3 and $h = 1/2(m+1)$ in test case 4. The resulting system of equations is

$$\begin{aligned} F_{i,j}(u) \equiv \frac{1}{h^2}\Bigg[& D\left(\frac{u_{i+1,j}+u_{i,j}}{2}\right)(u_{i+1,j}-u_{i,j}) - D\left(\frac{u_{i,j}+u_{i-1,j}}{2}\right)(u_{i,j}-u_{i-1,j}) \\ & + D\left(\frac{u_{i,j+1}+u_{i,j}}{2}\right)(u_{i,j+1}-u_{i,j}) - D\left(\frac{u_{i,j}+u_{i,j-1}}{2}\right)(u_{i,j}-u_{i,j-1})\Bigg] \\ & - g(u_{i,j}) + f(x_i,y_j) = 0 \end{aligned} \tag{3.2}$$

for $1 \le i \le m$ and $1 \le j \le m$, where (x_i, y_j) denotes the ijth grid point and $u_{i,j}$ denotes the approximate solution there. (The boundary conditions give $u_{0,j} = u_{m+1,j} = u_{i,0} = u_{i,m+1} = 0$.) We denote this system by $F(u) = 0$, where, for convenience, u denotes the vector of $u_{i,j}$'s as well as the solution of (3.1).

3.2. Choices of $\tilde{F}$. In our experiments, we considered two illustrative choices of $\tilde{F}$ based on our motivating assumption that D-evaluations are significantly more expensive than g-evaluations. (Other choices of $\tilde{F}$ are possible; see the remark at

the end of this subsection.) We refer to these choices as the "lagged" and "linear" approximations of F and denote them by $\tilde{F}^{\mathrm{Lag}}(u,w)$ and $\tilde{F}^{\mathrm{Lin}}(u,w)$, respectively. These correspond to the respective approximations

$$\tilde{\mathcal{F}}^{\mathrm{Lag}}(u,w) \equiv \nabla \cdot D(u)\nabla w - g(w) + f$$

and

$$\tilde{\mathcal{F}}^{\mathrm{Lin}}(u,w) \equiv \nabla \cdot [D(u) + D'(u)(w-u)]\nabla w - g(w) + f$$

of $\mathcal{F}$ in the continuous problem (3.1). They are defined by respective ijth components

$$\begin{aligned}
&\tilde{F}^{\mathrm{Lag}}_{i,j}(u,w) \\
&\quad \equiv \frac{1}{h^2}\left[D\left(\frac{u_{i+1,j}+u_{i,j}}{2}\right)(w_{i+1,j}-w_{i,j}) - D\left(\frac{u_{i,j}+u_{i-1,j}}{2}\right)(w_{i,j}-w_{i-1,j})\right. \\
&\qquad \left. + D\left(\frac{u_{i,j+1}+u_{i,j}}{2}\right)(w_{i,j+1}-w_{i,j}) - D\left(\frac{u_{i,j}+u_{i,j-1}}{2}\right)(w_{i,j}-w_{i,j-1})\right] \\
&\qquad - g(w_{i,j}) + f(x_i,y_j),
\end{aligned} \tag{3.3}$$

$$\begin{aligned}
&\tilde{F}^{\mathrm{Lin}}_{i,j}(u,w) \\
&\quad \equiv \frac{1}{h^2}\left\{\left[D\left(\frac{u_{i+1,j}+u_{i,j}}{2}\right) + \frac{1}{2}D'\left(\frac{u_{i+1,j}+u_{i,j}}{2}\right)(w_{i+1,j}+w_{i,j}-u_{i+1,j}-u_{i,j})\right]\right. \\
&\qquad \cdot (w_{i+1,j}-w_{i,j}) \\
&\qquad - \left[D\left(\frac{u_{i,j}+u_{i-1,j}}{2}\right) + \frac{1}{2}D'\left(\frac{u_{i,j}+u_{i-1,j}}{2}\right)(w_{i,j}+w_{i-1,j}-u_{i,j}-u_{i-1,j})\right] \\
&\qquad \cdot (w_{i,j}-w_{i-1,j}) \\
&\qquad + \left[D\left(\frac{u_{i,j+1}+u_{i,j}}{2}\right) + \frac{1}{2}D'\left(\frac{u_{i,j+1}+u_{i,j}}{2}\right)(w_{i,j+1}+w_{i,j}-u_{i,j+1}-u_{i,j})\right] \\
&\qquad \cdot (w_{i,j+1}-w_{i,j}) \\
&\qquad - \left[D\left(\frac{u_{i,j}+u_{i,j-1}}{2}\right) + \frac{1}{2}D'\left(\frac{u_{i,j}+u_{i,j-1}}{2}\right)(w_{i,j}+w_{i,j-1}-u_{i,j}-u_{i,j-1})\right] \\
&\qquad \left. \cdot (w_{i,j}-w_{i,j-1})\right\} - g(w_{i,j}) + f(x_i,y_j).
\end{aligned} \tag{3.4}$$

It is clear that both $\tilde{F}^{\mathrm{Lag}}(u,w)$ and $\tilde{F}^{\mathrm{Lin}}(u,w)$ satisfy $\tilde{F}(u,u) = F(u)$. To apply the theoretical results obtained in the previous section, we must determine whether they also satisfy Assumptions 2.7 and 2.10. To facilitate verifying these assumptions, we note that the discrete problem (3.2) is a particular case of the model problem

$$F(u) \equiv A(u)u - b(u) + c = 0, \tag{3.5}$$

where $u \in \mathbb{R}^n$, $A(u) \in \mathbb{R}^{n\times n}$ has continuously differentiable entries $a_{ij}(u)$, and $b(u) \in \mathbb{R}^n$ has continuously differentiable components $b_i(u)$. For this model problem, we have that

$$F'(u)v = A(u)v + [A'(u)v]u - b'(u)v,$$

where $[A'(u)v] \in \mathbb{R}^{n\times n}$ has entries $[A'(u)v]_{ij} = \nabla a_{ij}(u)^T v$ and $\nabla a_{ij}(u)$ denotes the gradient of $a_{ij}(u)$ with respect to u. We define

$$\tilde{F}^{\text{Lag}}(u,w) \equiv A(u)w - b(w) + c, \tag{3.6}$$

$$\tilde{F}^{\text{Lin}}(u,w) \equiv (A(u) + [A'(u)(w-u)])\, w - b(w) + c. \tag{3.7}$$

We show the following results for problem (3.5) and the approximations (3.6)–(3.7).

THEOREM 3.1. *Let F, $\tilde{F}^{\text{Lag}}$, and $\tilde{F}^{\text{Lin}}$ be defined by* (3.5)–(3.7). *Then*

$$\frac{\tilde{F}^{\text{Lag}}(u,u+\sigma v) - F(u)}{\sigma} - F'(u)v = -[A'(u)v]u - \left[\frac{b(u+\sigma v) - b(u)}{\sigma} - b'(u)v\right] \tag{3.8}$$

and

$$\frac{\tilde{F}^{\text{Lin}}(u,u+\sigma v) - F(u)}{\sigma} - F'(u)v = \sigma[A'(u)v]v - \left[\frac{b(u+\sigma v) - b(u)}{\sigma} - b'(u)v\right]. \tag{3.9}$$

If, in addition, there exists an $\Omega \subseteq \mathbb{R}^n$ such that, for all $u \in \Omega$, $|\partial a_{ij}(u)/\partial u_k| \le C_1 < \infty$ and $|\partial b_i(u)/\partial u_j| \le C_2 < \infty$ for all $i,j,k \in \{1,\dots,n\}$ and some constants C_1 and C_2, then both $\tilde{F}^{\text{Lag}}$ and $\tilde{F}^{\text{Lin}}$ satisfy Assumption 2.7 *with $\omega_*^{\text{Lag}} < \infty$ for $\tilde{F}^{\text{Lag}}$ and $\omega_*^{\text{Lin}} = 0$ for $\tilde{F}^{\text{Lin}}$.*

Proof. The proof follows by substituting the definitions of the lagged and linear approximations into the above expressions and then deleting appropriate terms. ∎

THEOREM 3.2. *Let the assumptions regarding $\tilde{F}^{\text{Lin}}$ in Theorem* 3.1 *hold. In addition, assume that there exist γ and $p \in (0,1]$ such that*

$$\|b'(u+v) - b'(u)\| \le \gamma \|v\|^p$$

for all $u \in \Omega$ and all v, with $\|v\| \le \delta_\Omega$. Then Assumption 2.10 *holds for $\tilde{F}^{\text{Lin}}$ with the same value of p.*

Proof. We first show that, for any v,

$$\tilde{F}^{\text{Lin}}_w(u,w)v = A(u)v + [A'(u)(w-u)]v + [A'(u)v]w - b'(w)v.$$

We have

$$\begin{aligned}
&\sigma^{-1}\left[\tilde{F}^{\text{Lin}}(u,w+\sigma v) - \tilde{F}^{\text{Lin}}(u,w)\right] \\
&\quad = \sigma^{-1}\left[(A(u) + [A'(u)(w+\sigma v - u)])(w+\sigma v) - (A(u) + [A'(u)(w-u)])w\right] \\
&\qquad - \sigma^{-1}\left[b(w+\sigma v) - b(w)\right] \\
&\quad = A(u)v + [A'(u)(w-u)]v + [A'(u)v](w+\sigma v) - \sigma^{-1}\left[b(w+\sigma v) - b(w)\right].
\end{aligned}$$

Letting $\sigma \to 0$ gives the result. Next, for any vector $\hat{v}$

$$\begin{aligned}
&\tilde{F}^{\text{Lin}}_w(u,u+v)\hat{v} - F'(u)\hat{v} \\
&\qquad = A(u)\hat{v} + [A'(u)v]\hat{v} + [A'(u)\hat{v}](u+v) - b'(u+v)\hat{v} \\
&\qquad\quad - A(u)\hat{v} - [A'(u)\hat{v}]u + b'(u)\hat{v} \\
&\qquad = [A'(u)v]\hat{v} + [A'(u)\hat{v}]v - (b'(u+v) - b'(u))\hat{v}.
\end{aligned}$$

Letting $\|\cdot\|_F$ denote the Frobenius norm on $\mathbb{R}^{n\times n}$, we have

$$\|[A'(u)v]\|_F = \left(\sum_{i,j=1}^n (\nabla a_{ij}(u)^T v)^2\right)^{1/2} \le \|v\| \left(\sum_{i,j=1}^n \|\nabla a_{ij}(u)\|^2\right)^{1/2} \le K \cdot \|v\|$$

for some constant K since $|\partial a_{ij}(u)/\partial u_k| \le C_1 < \infty$ for all $i, j, k = 1, \ldots, n$. Thus,

$$\begin{aligned}\|\tilde{F}_w^{\mathrm{Lin}}(u, u+v)\hat{v} - F'(u)\hat{v}\| &\le 2K \cdot \|v\| \cdot \|\hat{v}\| + \|b'(u+v) - b'(u)\| \cdot \|\hat{v}\| \\ &\le 2K \cdot \|v\| \cdot \|\hat{v}\| + \gamma\|v\|^p \cdot \|\hat{v}\| \\ &= \left(2K\delta_\Omega^{1-p} + \gamma\right)\|v\|^p \cdot \|\hat{v}\|\end{aligned}$$

for any vector $\hat{v}$. Therefore,

$$\|\tilde{F}_w^{\mathrm{Lin}}(u, u+v) - F'(u)\| \le \left(2K\delta_\Omega^{1-p} + \gamma\right)\|v\|^p$$

for all $u \in \Omega$ and all v with $\|v\| \le \delta_\Omega$, which completes the proof. □

For the discrete problem (3.2), a short computation reveals that $\tilde{F}^{\mathrm{Lag}}$ and $\tilde{F}^{\mathrm{Lin}}$ given by (3.3) and (3.5) both satisfy the assumptions of Theorem 3.1 if there exists a convex set Ω such that $|D'(u)|$ and $|g'(u)|$ are uniformly bounded for $u \in \Omega$. Additionally, $\tilde{F}^{\mathrm{Lin}}$ satisfies the assumptions of Theorem 3.2 if $g'(u)$ is Hölder continuous in Ω. Thus, under these mild assumptions, the local convergence results in section 2.2 can be applied to these two cases.

We note that using $\tilde{F} = \tilde{F}^{\mathrm{Lag}}$ given by (3.3) requires minimal new calculations at each step of the linear solve in Algorithm 2.5. Indeed, each step requires only one g-evaluation at each grid point in addition to a very modest amount of arithmetic, since the necessary D- and f-values are already available. Thus, the computational effort required to implement $\tilde{F}^{\mathrm{Lag}}$ is so small that the time savings may be considerable for problems for which convergence can be achieved with this approximation.

Using $\tilde{F} = \tilde{F}^{\mathrm{Lin}}$ given by (3.5) in Algorithm 2.5 requires evaluating D' at points intermediate to the grid points, which may be problematic if D'-evaluations are expensive or unavailable. However, this needs to be done only once at the outset of each linear solve and, in many applications, may be no more difficult than evaluating D at those points. Once the necessary D'-values have been computed, each step of the linear solve requires the same g-evaluations as $\tilde{F}^{\mathrm{Lag}}$ together with a somewhat greater but still very modest amount of arithmetic.

Remark. One of the referees suggested an alternative to (3.7) in defining $\tilde{F}^{\mathrm{Lin}}$ for the model problem (3.5), viz.,

$$\tilde{F}^{\mathrm{Lin}}(u, w) = A(u)w + [A'(u)(w-u)]u - b(w) + c. \tag{3.10}$$

This has the error term

$$\frac{\tilde{F}^{\mathrm{Lin}}(u, u+\sigma v) - F(u)}{\sigma} - F'(u)v = -\left[\frac{b(u+\sigma v) - b(u)}{\sigma} - b'(u)v\right],$$

which is slightly simpler than (3.9). Implementing this $\tilde{F}^{\mathrm{Lin}}$ on the discrete problem (3.2) has essentially the same cost as implementing $\tilde{F}^{\mathrm{Lin}}$ given by (3.7). With this $\tilde{F}^{\mathrm{Lin}}$, the analysis in section 2.2 can be applied as in the remark at the end of section 2.2 by writing F in (3.5) as $F = F_1 + F_2$, where $F_1(u) \equiv A(u)u$ and $F_2(u) \equiv -b(u) + c$. Then (3.10) becomes

$$\tilde{F}^{\mathrm{Lin}}(u, w) = F_1(u) + F_1'(u)(w-u) + F_2(w),$$

and Assumptions 2.7 and 2.10 become assumptions on F_2 and, hence, on b as in the remark at the end of section 2.2.

TABLE 2
Key to abbreviations.

Abbreviation	Meaning
NEQ	Number of equations (unknowns)
NNI	Number of nonlinear iterations
NLI	Number of linear iterations

3.3. Numerical results. In our numerical experiments, the test problems described above were treated by applying the KINSOL Newton–Krylov code from the SUNDIALS suite [10]; see [10] for details of the algorithm not specified here. KINSOL was applied in matrix-free mode with modifications to allow the approximation (2.1) with $\tilde{F} = \tilde{F}^{\mathrm{Lag}}$ and $\tilde{F} = \tilde{F}^{\mathrm{Lin}}$ to be used in place of (1.4). The linear solver used was GMRES with a banded block-diagonal preconditioner provided by SUNDIALS using finite differences of F-values; see [10]. The linear and nonlinear solver specifications were as follows: the maximum number of GMRES iterations allowed was 100, with no restarts; each forcing term η_k was the constant 10^{-3}; the nonlinear iterations terminated successfully if $\|F(u_k)\| \leq 10^{-8}$. These tight tolerances were chosen to bring out performance differences resulting from the different choices of $\tilde{F}$. Additionally, failure was declared if either the number of nonlinear iterations exceeded 200 or $\|s_k\| \leq 10^{-13}$; however, these failure modes were not observed in our tests.

All runs were done on the ASC Frost machine at Lawrence Livermore National Laboratory, an IBM SP parallel platform running the AIX operating system with 1088 375 MHz processors grouped into 16-processor nodes. In each of our runs, the spatial domain was subdivided into a $p \times p$ array of square subdomains having an equal number of grid points, and the grid points in each subdomain were mapped to one processor.

In test cases 1–3, the domain was divided into 400×400 zones yielding $160,000$ unknowns. In test case 4, a smaller problem was also considered, in which the domain was divided into 200×200 zones yielding 40,000 unknowns. In all cases, the grid was adjusted so that the number of grid points per subdomain (and processor) was always 10,000. Thus, for the 160,000-unknown problems, 16 processors were used, and four were used for the smaller test case 4 problem.

The results of our runs are given in the tables below. A key to the abbreviations used in these tables is provided in Table 2. In Tables 3–6, "None" refers to using no approximation, i.e., using (1.4); "Linear" and "Lagged" refer to using $\tilde{F} = \tilde{F}^{\mathrm{Lag}}$ and $\tilde{F} = \tilde{F}^{\mathrm{Lin}}$, respectively, in (2.1). Run times are given in seconds, and each normalized run time is calculated by dividing the run time by that obtained by using no approximation.

3.3.1. Case 1: $D(u) = \sqrt{u^2 + u + 1}$. This choice for D is the least expensive to evaluate of all of those we experimented with, and so we expected the time savings using the approximations $\tilde{F}^{\mathrm{Lag}}$ and $\tilde{F}^{\mathrm{Lin}}$ to be less significant than in the other test problems. Recall that the problem was posed so that the exact solution was $u_c(x, y) = cx(1-x)y(1-y)$. Here we took $c = 10$ with the initial approximate solution $u_0 \equiv c$. (No significant additional information was gained in this case by considering the other values of c.) The results are given in Table 3. There were no unsuccessful runs in this case.

As expected, since D is not very expensive to evaluate, using the approximations $\tilde{F}^{\mathrm{Lag}}$ and $\tilde{F}^{\mathrm{Lin}}$ did not significantly reduce computational time. In fact, the lagged approximation required more GMRES iterations and, consequently, more run time than

TABLE 3
Case 1: $D(u) = \sqrt{u^2 + u + 1}$, $c = 10$.

$\tilde{F}$ approx.	NNI	NLI	Run time	Norm. run time
None	9	416	38.2	1.00
Linear	9	417	35.3	0.92
Lagged	9	530	43.8	1.15

TABLE 4
Case 2: $D(u) = \frac{1}{\sqrt{u^2+u+1}}$.

c	$\tilde{F}$ approx.	NNI	NLI	Run time	Norm. run time
1	None	7	329	32.7	1.00
	Linear	7	331	30.3	0.93
	Lagged	6	323	29.1	0.89
2	None	12	387	41.1	1.00
	Linear	12	387	38.2	0.93
	Lagged	8	442	37.7	0.92
5	None	7	329	32.1	1.00
	Linear	7	329	29.5	0.92
	Lagged	8	436	37.1	1.16
10	None	7	345	33.7	1.00
	Linear	7	343	30.8	0.91
	Lagged	10	566	47.2	1.40

the two alternatives, apparently because of less compatibility between the approximate Jacobian-vector products obtained with $\tilde{F}^{\mathrm{Lag}}$ and the preconditioner provided by SUNDIALS using exact F-values. Consistent with the theory, there is very close agreement between the numbers of nonlinear and linear iterations obtained by using $\tilde{F}^{\mathrm{Lin}}$ and those obtained by using no approximation.

3.3.2. Case 2: $D(u) = \frac{1}{\sqrt{u^2+u+1}}$. This choice for D is only slightly more expensive to evaluate than the choice in Case 1; however, there are some notable differences in the computational results. For $c = 1$ and $c = 2$, we took $u_0 \equiv c$ as in Case 1. However, for $c = 5$ and $c = 10$, we took $u_0 \equiv 1$; this was necessary in order to obtain convergence to the solution, even when no approximation was used. With these choices of u_0, all runs were successful. The results are shown in Table 4.

This choice for D is interesting because the results with the lagged approximation $\tilde{F}^{\mathrm{Lag}}$ show sensitivity to the value of c. As c gets larger, the lagged approximation becomes less accurate, and, as a result, its performance degrades. The results obtained by using the linear approximation $\tilde{F}^{\mathrm{Lin}}$ are similar to those in Case 1. Again we note very close agreement between the numbers of linear and nonlinear iterations obtained by using $\tilde{F}^{\mathrm{Lin}}$ and those obtained by using no approximation.

3.3.3. Case 3: $D(u) = u^{\frac{3}{5}}e^u$. This choice for D entails the largest values of $D'(u)$ at the solutions $u_c(x, y) = cx(1 - x)y(1 - y)$, and so we expected differences in the effectiveness of $\tilde{F}^{\mathrm{Lag}}$ and $\tilde{F}^{\mathrm{Lin}}$ as approximations in (2.1) to be significant and to become increasingly pronounced as c increases. We took the initial approximate solution to be $u_0 \equiv c$ for each value of c. With this u_0, there were two failures when $\tilde{F}^{\mathrm{Lag}}$ was used. The results are given in Table 5.

TABLE 5
Case 3: $D(u) = u^{\frac{3}{5}} e^u$.

c	$\tilde{F}$ Approx.	NNI	NLI	Run time	Norm. run time
1.0	None	16	600	103.4	1.00
	Linear	16	598	81.5	0.79
	Lagged	33	2374	273.6	2.65
2.0	None	17	645	110.0	1.00
	Linear	17	632	85.1	0.77
	Lagged	35	2503	287.0	2.61
5.0	None	19	771	124.8	1.00
	Linear	19	782	98.0	0.79
	Lagged	-	-	-	-
10.0	None	23	1201	199.2	1.00
	Linear	23	1199	152.4	0.77
	Lagged	-	-	-	-

The results in Table 5 are consistent with our expectations. The lagged approximation resulted in failures of the method and greatly increased run times when it did not fail. (The failures resulted from GMRES failing to produce the requested residual norm reduction within the allowable 100 iterations.) In contrast, the linear approximation led to success in every case and to significant reductions in run times. For all values of c, as seen in the previous cases, there is very close agreement between the numbers of linear and nonlinear iterations obtained by using $\tilde{F}^{\mathrm{Lin}}$ and those obtained by using no approximation.

3.3.4. Case 4: $D(u) = K\sqrt{S(u)}[1 - (1 - S(u)^{\frac{1}{\mu}})^{\mu}]^2$. This problem is the most realistic of all of those considered here, and, of all of the choices for D, this is the most expensive to evaluate. The problem is related to a formulation of Richards's equation, which is often used for modeling fluid flows in unsaturated porous media. Here we chose values of the parameters appropriate for groundwater flow in an unsaturated dune sand (see [13] and the references therein), as follows: $K = 5.040$, $\alpha = 5.470$, $\nu = 4.264$, and $\mu = \frac{\nu-1}{\nu}$. In each test case, we began with $u_0 \equiv \beta = \frac{1}{16}$, for which all runs were successful. Results are given in Table 6.

As in the previous cases, there is very close agreement between the numbers of linear and nonlinear iterations obtained by using the linear approximation and those obtained by using no approximation. For the linear approximation, the greater runtime reductions on the larger problems are notable and likely reflect the benefit of amortizing the cost of evaluating D' over larger numbers of GMRES iterations per inexact Newton step. While not as effective as the linear approximation, the lagged approximation, for all values of c except 10, also resulted in significantly reduced run times, with greater reductions on the larger problems, which required many more GMRES iterations per inexact Newton step than the smaller problems.

4. Summary and conclusions. We have considered Newton–Krylov methods for solving (1.1) in which the matrix-vector products required by the Krylov solver are approximated by finite differences of the form (2.1) involving an approximating function $\tilde{F} \approx F$. The particular Krylov subspace methods considered are the GMRES and Arnoldi methods, which are based on the Arnoldi process (Algorithm 2.1); however, the developments here should also be helpful when other Krylov subspace methods are of interest.

TABLE 6
Case 4: $D(u) = K\sqrt{S(u)}[1 - (1 - S(u)^{\frac{1}{\mu}})^{\mu}]^2$.

NEQ	c	$\tilde{F}$ approx.	NNI	NLI	Run time	Norm. run time
40,000	1.0	None	4	107	98.5	1.00
		Linear	4	105	79.9	0.81
		Lagged	4	104	79.2	0.80
160,000	1.0	None	4	220	141.2	1.00
		Linear	4	220	102.3	0.72
		Lagged	4	223	101.6	0.72
40,000	2.0	None	4	109	99.2	1.00
		Linear	4	104	79.9	0.81
		Lagged	5	135	84.8	0.85
160,000	2.0	None	4	215	139.0	1.00
		Linear	4	215	101.5	0.73
		Lagged	5	280	113.1	0.81
40,000	5.0	None	5	138	110.4	1.00
		Linear	5	139	87.1	0.79
		Lagged	6	163	90.3	0.82
160,000	5.0	None	5	278	163.6	1.00
		Linear	5	278	114.2	0.70
		Lagged	6	330	122.6	0.75
40,000	10.0	None	5	144	112.9	1.00
		Linear	5	147	88.8	0.79
		Lagged	8	232	103.6	0.92
160,000	10.0	None	5	281	162.7	1.00
		Linear	5	289	115.7	0.71
		Lagged	9	514	156.9	0.96

In Algorithm 2.2, we formulate an Arnoldi process that uses approximate finite differences of the form (2.1) on a Jacobian system (2.2), and we consider the GMRES and Arnoldi methods applied to (2.2) that are based on this process. In Theorem 2.3 and Corollary 2.4, it is shown that an approximate solution of (2.2) determined at some step of the GMRES or Arnoldi method based on Algorithm 2.2 is the same as the approximate solution of the perturbed system (2.6) obtained at the same step of the method based on Algorithm 2.1.

In Algorithm 2.5, we outline a Newton–Krylov method that uses the GMRES or Arnoldi method based on Algorithm 2.2. Our main result for Algorithm 2.5 is Theorem 2.12, which can be viewed as a counterpart of results in [5, Thms. 2.3 and 3.3]. This theorem asserts that, under Assumption 2.7 on the approximating function $\tilde{F}$ as well as mild assumptions on F, the iterates produced by Algorithm 2.5 converge to a solution locally and q-linearly in a certain norm, provided the forcing terms are uniformly bounded below one and the difference steps are uniformly sufficiently small. If the forcing terms and difference steps approach zero, then the convergence is q-superlinear. If Assumption 2.10 on $\tilde{F}$ also holds, then the convergence is of q-order $1 + p$, where $p \in (0, 1]$ is the exponent of Hölder continuity in Assumption 2.10.

In section 3, we report on numerical experiments with two illustrative choices of $\tilde{F}$ that are suitable for a broad class of nonlinear diffusion problems. One of these choices, the "linear" approximation $\tilde{F}^{\mathrm{Lin}}$ satisfies the assumptions of Theorem 2.12 under mild conditions and, therefore, has a sound theoretical basis. In our tests, this choice yielded run times that were always less than those obtained by using exact values

of F, with greater run-time reductions observed for problems with more expensive F-evaluations. Additionally, in all test cases, this choice resulted in numbers of linear and nonlinear iterations that were very close to those obtained by using exact values of F. Thus this choice promises to be both robust and helpful in improving efficiency in many nonlinear diffusion problems, especially those in which F-evaluations are expensive. However, its implementation may be problematic in some applications (see section 3.2). Our other choice, the "lagged" approximation $\tilde{F}^{\mathrm{Lag}}$, does not normally satisfy the assumptions of Theorem 2.12; however, it is relatively easy and inexpensive to apply. It improved efficiency in some of our tests, especially in Case 2 (see Table 4); however, it was inefficient in other tests and suffered failures in Case 3 (see Table 5). Thus this choice should be kept in mind for ease of application and potential for improving efficiency but should not be counted on for improvement in all problems.

REFERENCES

[1] W. E. Arnoldi, *The principle of minimized iterations in the solution of the matrix eigenvalue problem*, Quart. Appl. Math., 9 (1951), pp. 17–29.

[2] P. N. Brown and Y. Saad, *Hybrid Krylov methods for nonlinear systems of equations*, SIAM J. Sci. Stat. Comput., 11 (1990), pp. 450–481.

[3] P. N. Brown, *A local convergence theory for combined inexact-Newton/finite-difference projection methods*, SIAM J. Numer. Anal., 24 (1987), pp. 407–434.

[4] E. Catinas, *Inexact perturbed Newton methods and applications to a class of Krylov solvers*, J. Optim. Theory Appl., 108 (2001), pp. 543–571.

[5] R. S. Dembo, S. C. Eisenstat, and T. Steihaug, *Inexact Newton methods*, SIAM J. Numer. Anal., 19 (1982), pp. 400–408.

[6] J. E. Dennis, Jr., and R. B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Series in Automatic Computation, Prentice-Hall, Englewood Cliffs, NJ, 1983.

[7] S. C. Eisenstat and H. F. Walker, *Choosing the forcing terms in an inexact Newton method*, SIAM J. Sci. Comput., 17 (1996), pp. 16–32.

[8] R. W. Freund, G. H. Golub, and N. M. Nachtigal, *Iterative solution of linear systems*, Acta Numer. 1992, 1 (1992), pp. 57–100.

[9] G. H. Golub and C. Van Loan, *Matrix Computations*, 3rd ed., Johns Hopkins University Press, Baltimore, 1996.

[10] A. C. Hindmarsh, P. N. Brown, K. E. Grant, S. L. Lee, R. Serban, D. E. Shumaker, and C. S. Woodward, *SUNDIALS: Suite of nonlinear and differential/algebraic equation solvers*, ACM Trans. Math. Software, 31 (2005), pp. 363–396.

[11] C. T. Kelley, *Solving Nonlinear Equations with Newton's Method*, Fundam. Algorithms 1, SIAM, Philadelphia, PA, 2003.

[12] D. A. Knoll and D. E. Keyes, *Jacobian-free Newton–Krylov methods: A survey of approaches and applications*, J. Comput. Phys., 193 (2004), pp. 357–397.

[13] C. T. Miller, G. A. Williams, C. T. Kelley, and M. D. Tocci, *Robust solution of Richards' equation for non-uniform porous media*, Water Resources Research, 34 (1998), pp. 2599–2610.

[14] J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Comput. Sci. Appl. Math., Academic Press, New York, 1970.

[15] R. P. Pawlowski, J. N. Shadid, J. P. Simonis, and H. F. Walker, *Globalization techniques for Newton–Krylov methods and applications to the fully coupled solution of the Navier–Stokes equations*, SIAM Rev., 48 (2006), pp. 700–721.

[16] Y. Saad and M. H. Schultz, GMRES: *A generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Stat. Comput., 7 (1986), pp. 856–869.

[17] Y. Saad, *Krylov subspace methods for solving large unsymmetric linear systems*, Math. Comp., 37 (1981), pp. 105–126.

[18] Y. Saad, *Iterative Methods for Sparse Linear Systems*, PWS Publishing, Boston, 1996.

[19] H. A. van der Vorst, *Iterative Krylov Methods for Large Linear Systems*, Cambridge University Press, Cambridge, 2003.

SIAM J. NUMER. ANAL.
Vol. 46, No. 4, pp. 1912–1934
© 2008 Society for Industrial and Applied Mathematics

RENORMALIZED MESHFREE SCHEMES I: CONSISTENCY, STABILITY, AND HYBRID METHODS FOR CONSERVATION LAWS*

NATHALIE LANSON† AND JEAN-PAUL VILA‡

Abstract. This paper is devoted to the study of a new kind of meshfree scheme based on a new class of meshfree derivatives: the renormalized meshfree derivatives, which improve the consistency of the original weighted particle methods. The weak renormalized meshfree scheme, built from the weak formulation of general conservation laws, turns out to be L^2 stable under some geometrical conditions on the distribution of particles and some regularity conditions of the transport field. A time discretization is then performed by analogy with finite volume methods, and the L^1, L^∞, and BV stabilities of the obtained time discretized scheme are studied. From the same analogy with finite volume methods, a hybrid particle scheme is built using the Godunov method and is numerically compared to the weak renormalized scheme.

Key words. meshfree methods, SPH, renormalization, weak scheme, nonlinear conservation law, finite volume, hybrid particle schemes

AMS subject classifications. 65M12, 65M99

DOI. 10.1137/S0036142903427718

1. Introduction.

We focus on conservation laws of type

$$\begin{cases} L_a(u) + \operatorname{div}(F(u,x,t)) = S(u,x,t), \quad t \in \mathbb{R}_+,\ x \in \mathbb{R}^d, \\ u(x,0) = u_0(x), \end{cases} \tag{1}$$

where L_a is the transport operator defined as $L_a(u) = \frac{\partial u}{\partial t} + \operatorname{div}(a(x,t) \otimes u)$, a the transport field $a = (a^\alpha)_{\alpha=1}^d$, $a^\alpha : \mathbb{R}^d \times \mathbb{R}_+ \to \mathbb{R}$, and $u \in \mathbb{R}^p$ is the exact solution of the problem. We denote $u_0 \in \mathbb{R}^p$ to be the initial condition. The flux $F = (F^\alpha)_{\alpha=1}^d$, composed of d vectors $F^\alpha \in \mathbb{R}^p$, and the source term $S = (S^l)_{l=1}^p$ are as follows: $F^\alpha : \mathbb{R}^p \times \mathbb{R}^d \times \mathbb{R}_+ \to \mathbb{R}^p\ \forall \alpha = 1, \dots, d$ and $S : \mathbb{R}^p \times \mathbb{R}^d \times \mathbb{R}_+ \to \mathbb{R}^p$.

Many numerical schemes aimed at solving this problem already exist and can be classified as mesh-based or meshfree methods. The main advantage of meshfree methods lies in their ability to handle complex situations such as extremely large deformation of bodies, impact problems, treatment of discontinuities as in multiphases flows, or free surface flows. Mathematical analysis of weighted particle methods [14] pointed out the lack of consistency of these methods. We focus on the renormalization, a recent tool introduced in [13, 7], which seems to give accurate results and increases the consistency of weighted particle methods. According to the literature [4] this tool, although very attractive for applications, seems to make the global conservativity of the original method disappear. In this paper, this drawback is alleviated by introducing a new class of meshfree approximation of derivatives, the renormalized meshfree derivatives. We propose to use this new class of derivatives for the

*Received by the editors May 6, 2003; accepted for publication (in revised form) December 6, 2007; published electronically April 18, 2008.
http://www.siam.org/journals/sinum/46-4/42771.html
†Applied Mathematics Department, University of Waterloo, Waterloo, ON N2L 3G1, Canada (nlanson@math.uwaterloo.ca). This author was supported by the Association Française des Femmes Diplomées des Universités.
‡Département Génie Mathématique et Modélisation, INSA, 31077 Toulouse, France (vila@gmm.insa-tlse.fr).

approximation of the weak formulation of conservation laws (see [17, 2]) to build and study the conservative weak renormalized meshfree scheme.

We will start by recalling some basic principles of particle approximation, and we will build the new class of meshfree approximate derivatives and study their consistency. The end of the second section is devoted to the construction and the computational study of a time continuous scheme based on the weak formulation of (1): the weak renormalized meshfree scheme. In section 3, we study the L^2 stability property of the time continuous weak renormalized meshfree scheme and give some geometrical conditions necessary for stability to be achieved. We verify these geometrical conditions in the case of a regular transport field and a regular initial distribution of particles. Section 4 is devoted to the construction and stability study of a time discretized weak renormalized scheme. We will apply the obtained scheme to general scalar conservation laws and give L^∞, L^1, and BV stability results. Finally, in section 5, we apply the analogy made with finite volumes in section 4 to develop another type of renormalized meshfree method: *the renormalized hybrid particle method*. This new method is applied to Euler equations, and comparisons with the weak renormalized meshfree scheme are given in the case of free surface flows.

2. Renormalized particle approximation.

2.1. Basic principles of particle approximation. We consider a regular vector field a and introduce a system of Lagrangian coordinates (ξ, t). Let us take a set of moving particles P, with coordinates $x_j(t)$ $\forall j \in P$, volumes $\Omega_j(t)$ which represent a volume surrounding the particle j at time t (see the example in Figure 1(a)), and weights $w_j(t)$, such that $w_j(t) = \text{measure}(\Omega_j(t))$. The *characteristic distance* between particles is denoted $h(t)$ and represents the order of magnitude of the distances between particles at time t (see Figure 1(a)). h depends on the time t since the particles are moving. Indeed, consider $h_j(t)$ the minimum of the distances between a particle j and its immediate neighbors. The characteristic distance $h(t)$ is then defined as the maximum of the $h_j(t)$. The particles are moved along the characteristic curves of the field a and the weights are modified in order to take the deformations of the system into account:

$$\begin{aligned} &\frac{dx_j}{dt} = a(x_j, t), && \text{with } x_j(0) = \xi_j, \\ &\frac{dw_j(t)}{dt} = w_j(t)\text{div}(a(x_j, t)), && \text{with } w_j(0) = w_j^0. \end{aligned} \tag{2}$$

We consider that the vector field is regular enough, such that at any time t: $\mathbb{R}^d = \cup_{j\in P}\Omega_j(t)$. In the following, in order to simplify the notation, the time dependency will not be explicitly written.

The particle approximation of a function f is defined as

$$\Pi^h f(x) = \sum_{j\in P} w_j\, f(x_j)\, \delta(x - x_j). \tag{3}$$

In order to deal with derivatives, the particle weighted approximation $\Pi^h_\varepsilon f$ of a function f is defined as in (3), where the delta functions are approximated by the regularizing kernels $\zeta_\varepsilon(x - x_j)$, with ε the length of its compact support, also called the *smoothing length*. Such kernel functions ζ_ε are defined by

$$\zeta_\varepsilon(x - x_j) = \frac{1}{\varepsilon^d}\zeta\left(\frac{\|x - x_j\|}{\varepsilon}\right), \tag{4}$$

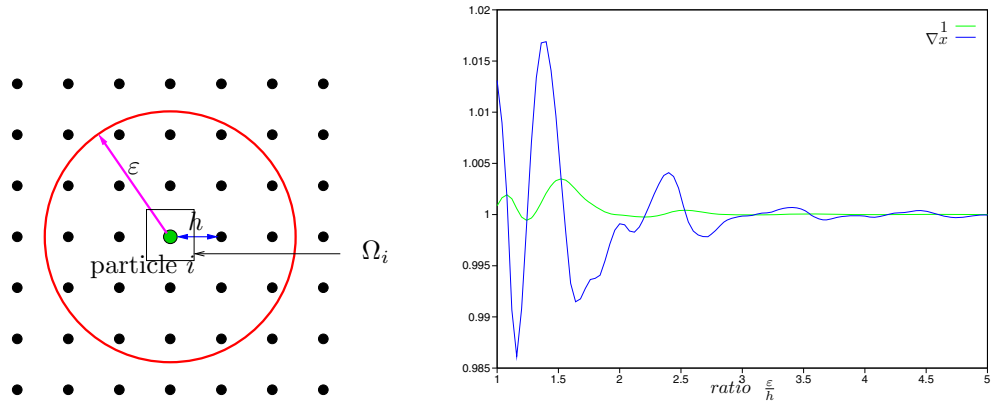


FIG. 1. *Geometry of the particle set and smoothed particle approximation.*

where the function ζ is a positive function with compact support. For instance, in the case of a cubic B-spline (see [11]), support$(\zeta) \subset [0,2]$ and ζ is given by

$$\zeta(r) = C \times \begin{cases} 1 - \frac{3}{2}r^2 + \frac{3}{4}r^3 & \text{if } 0 \le r \le 1, \\ \frac{1}{4}(2-r)^3 & \text{if } 1 \le r \le 2, \end{cases} \tag{5}$$

where $C = \frac{2}{3}$ for $d = 1$, $\frac{10}{7\pi}$ for $d = 2$, $\frac{1}{\pi}$ for $d = 3$.

The smoothing length varies in time and space according to

$$\frac{d\varepsilon}{dt} = \frac{\varepsilon}{d}\mathrm{div}a, \tag{6}$$

where d is again the dimension of the domain. The initial value of $\varepsilon(x,0)$ is computed using the initial characteristic length $h(0)$. If the particle are initially equidistributed, we take $\varepsilon(x,0) = Ch(0)\ \forall x$, where the coefficient C is chosen so that each particle has enough interaction with its surrounding neighboring particles.

We will suppose that the smoothing length stays bounded (this is true if a is regular enough). In section 3.2 we will take $a \in (\mathcal{C}^2(\mathbb{R}^d \times \mathbb{R}_+))^d)$:

$$\exists C, \varepsilon_0 \text{ constant such that } \forall x, t \quad \varepsilon_0 \le \varepsilon \le C\varepsilon_0. \tag{7}$$

When ε is computed from (6), the C depends on the transport field a.

The weighted particle approximation of derivatives is then given by the operator

$$D^{\varepsilon}_{\alpha,h} f(x) = \sum_{j \in P} w_j\, f(x_j)\, D_\alpha \zeta_\varepsilon(x - x_j). \tag{8}$$

Consider now that the approximation (8) is taken at $x = x_i$; then ε defines the set of particles that have a contribution to the calculation of $D^{\varepsilon}_{\alpha,h} f(x_i)$ as shown in Figure 1(a). This set of particles, referred to as the *neighborhood* of the particle i, will be denoted V_i in the following:

$$V_i = \{ j \in P \text{ s.t. } ||x_i - x_j|| \le \varepsilon \}. \tag{9}$$

It is shown in [14] that the approximation of derivatives depends on $\frac{h}{\varepsilon}$. This ratio $\frac{h}{\varepsilon}$ has to tend to zero in order to get convergence, which is not the case for numerical applications. Figure 1(b) represents the accuracy of the approximation of two functions, 1 and ∇x, on a regular grid in 2 dimensions of space, using the smoothed particle approximation (8), for different values of the ratio $\frac{\varepsilon}{h}$. One can see that, in two dimensions, a good value is $\varepsilon = 1.2h$. This value is different depending on the dimension of the space we work in. Within computations, however, the distribution of particles is no longer regular and it becomes difficult to ensure the consistency of the method.

2.2. A new class of meshfree approximation of derivatives. In order to alleviate the smooth particle methods' lack of consistency, we introduce the class of renormalized meshfree derivatives. We will consider a general kernel μ instead of $D_\alpha \zeta_\varepsilon$, as the renormalization applies to any kernel with compact support. Suppose that the approximation of the derivatives is given by

$$D^{\varepsilon}_{\alpha,h} f(x) = \sum_{j \in P} w_j \, f(x_j) \mu^\alpha(x, x_j) \quad \forall \alpha = 1, \dots, d, \tag{10}$$

where $\mu = \mu(x, y)$ is a vector-valued function with spherical compact support, of radius $\varepsilon = \varepsilon(x, y)$:

$$\mu(x, y) = 0 \quad \text{for } ||x - y|| \geq \varepsilon(x, y), \tag{11}$$

and ε satisfies (7). The following lemma can be proved.

LEMMA 2.1. *Under the hypotheses* (7), (11) *and the scaling hypothesis*

$$\exists C \text{ constant such that } ||\mu(x, y)|| \leq \frac{C}{\varepsilon_0^{d+1}} \quad \forall x, y \tag{12}$$

there exists $C > 0$ *(depending on the constants in* (7) *and* (12)*) such that* $\forall x, t$

$$\sum_{j \in P} w_j ||\mu(x, x_j)|| \leq \frac{C}{\varepsilon_0}. \tag{13}$$

The general renormalized meshfree derivatives of a function with respect to x^α, $\alpha \in \{1 \cdots d\}$, is given by

$$D_{\alpha,h} f(x) = \sum_{j \in P} w_j \left(f(x_j) - f(x)\right) \, A^\alpha(x, x_j), \tag{14}$$

where $A^\alpha(x, x_j)$ is an element of the vector $A(x, x_j)$ that we choose equal to

$$A(x, x_j) = B(x).\mu(x, x_j). \tag{15}$$

The *renormalization matrix* $B(x)$ allows us to achieve the consistency we are seeking.

The discrete operator of (14) and its adjoint are given by the following $\forall \alpha = 1, \dots, d$:

$$D_{\alpha,h} f(x_i) = \sum_{j \in P} w_j (f(x_j) - f(x_i)) A^\alpha_{ij}, D^*_{\alpha,h} f(x_i) = \sum_{j \in P} w_j (f(x_i) A^\alpha_{ij} - f(x_j) A^\alpha_{ji}). \tag{16}$$

Remark 2.2. In (16), when we consider $\mu(x, y) = \nabla \zeta(x, y)$, the vectors A_{ij} are defined by $A_{ij} := B_i . \nabla \zeta_{ij}$, which clearly reduces to the standard particle weighted approximation of derivatives if we take $B_i = B(x_i) = \mathbb{1}$.

The renormalization matrices $B_i = B(x_i)\ \forall i \in P$ are computed in order to get

$$D_h f = Df \quad \forall f \in \mathbb{P}^1(\mathbb{R}^d), \tag{17}$$

where $\mathbb{P}^1(\mathbb{R}^d)$ is the space of polynomials of $\mathbb{R}^d$ of order 1. The definition of $B_i\ \forall i \in P$ that satisfies the condition (17) is given in the following proposition.

PROPOSITION 2.3. *The condition* (17) *is satisfied if and only if* $B_i = (E_i)^{-1}$ *with*

$$E_i^{\alpha\beta} = \sum_{j\in P} w_j (x_j^\beta - x_i^\beta)\, \mu_{ij}^\alpha \quad \forall (\alpha,\beta) \in \{1,\dots,d\}^2, \tag{18}$$

where $\mu_{ij} = \mu(x_i, x_j)$*, and* μ *is the general kernel function used in* (10).

We can now prove the following error estimate.

PROPOSITION 2.4. *Suppose that the matrix* $E(x)$ *defined in* (18) *is invertible* $\forall x$*. Then, under the hypotheses* (7), (11), (12)*, there exists* $C > 0$ *such that* $\forall \varphi \in W^{2,\infty}(\mathbb{R}^d)$*, the renormalized meshfree derivatives* (14) *computed with* (18) *satisfy*

$$\|D_h\varphi(x) - D\varphi(x)\| \le C\varepsilon_0 \|B(x)\| \|D^2\varphi\|_\infty,$$

where D^2 *denotes the matrix of the second order derivatives of* φ*.*

Proof. Let us start from a Taylor series expansion:

$$\varphi_j - \varphi(x) = D\varphi(x).(x_j - x) + \overbrace{\frac{1}{2}\int_0^1 (1-s)(x_j - x).D^2\varphi(x - s(x_j - x)).(x_j - x)ds}^{R_j(\varphi)}.$$

We deduce that

$$D\varphi(x) - D_h\varphi(x) = \overbrace{D\varphi(x) - \sum_{j\in P} D\varphi(x).(x_j - x)B(x)\mu(x,x_j)}^{I} + \sum_{j\in P} w_j R_j(\varphi) B(x)\mu(x,x_j).$$

Using Proposition 2.3, we verify that

$$\begin{aligned} I^\alpha &= D_\alpha\varphi(x) - \sum_{\gamma,\beta=1}^d D_\gamma\varphi(x) B(x)^{\alpha\beta} \sum_{j\in P} (x_j^\gamma - x^\gamma)\mu^\beta(x,x_j) \\ &= D_\alpha\varphi(x) - \sum_{\gamma,\beta=1}^d \delta^{\alpha,\gamma} D_\gamma\varphi(x) = 0. \end{aligned}$$

It remains to estimate the term $\sum_{j\in P} w_j R_j(\varphi) B(x)\mu(x,x_j)$. Knowing that $\mu(x,x_j) = 0$ when $||x_j - x|| \ge C\varepsilon_0$, and using Lemma 2.1, we get

$$||D\varphi(x) - D_h\varphi(x)|| \le \frac{C\varepsilon_0^2}{2} ||D^2\varphi||_\infty |||B(x)||| \frac{C}{2\varepsilon_0} \le C\varepsilon_0 |||B(x)|||\ ||D^2\varphi||_\infty. \qquad \square$$

It turns out that the approximation depends now solely on the parameter ε_0, while the ratio h/ε_0 must tend to zero for classical weighted particles approximation.

The kernel is taken spherical in order to have $\nabla\zeta_{ij} = -\nabla\zeta_{ji}$. To simplify the

scheme and for local conservation, we choose $B_{ij} = \frac{B_i+B_j}{2}$ instead of B_i, so that $A_{ij} = -A_{ji}$. We can also prove that Proposition 2.4 also holds in the following case.

PROPOSITION 2.5. *Under the same hypothesis as in Proposition* 2.4 *and if the condition*

$$\forall (i,j) \text{ such that } ||x_j - x_i|| \leq C\varepsilon \Rightarrow B_j = B_i + O(\varepsilon) \tag{19}$$

is satisfied, then the L^∞ *error estimate of the operator* D_h *with* $A_{ij} = \frac{B_i+B_j}{2}\mu_{ij} = -A_{ji}$ *holds:*

$$||D_h\varphi(x) - D\varphi(x)|| \leq C\varepsilon_0 ||B(x)|| \ (||D^2\varphi||_\infty + ||D\varphi||_\infty). \tag{20}$$

The principle of the proof is the same as for Proposition 2.4, and it is left to the reader.

Hypotheses (7), (11), (12) are related to the choice of the shape function $\mu(x)$ and the variations of $\varepsilon(x)$. The important assumption in Proposition 2.4 is that the matrices $E(x)$ must be invertible $\forall x$. We will prove in section 3.2 that this is satisfied when the ratio $\frac{\varepsilon}{h}$ is of order 1, the transport field is $a \in (\mathcal{C}^2(\mathbb{R}^d \times \mathbb{R}_+))^d$, and the particles are initially equidistributed.

2.3. The time continuous weak renormalized meshfree scheme. We consider the general conservation law (1). In order to deal with nonlinear hyperbolic conservation laws and the computation of discontinuous solutions, it is necessary to introduce the notion of a weak solution of (1): $\forall \varphi \in (\mathcal{C}_c^2(\mathbb{R}^d \times \mathbb{R}_+^*))^p$

$$\int_{\mathbb{R}^d \times \mathbb{R}^+} (u(x,t)L_a^*(\varphi) + F(u,x,t).\nabla(\varphi) + S(u,x,t)\varphi)\, dxdt = 0, \tag{21}$$

where $-L_a^*$ is the adjoint operator of L_a defined as $L_a^*(\varphi) = \frac{\partial \varphi}{\partial t} + \sum_{\alpha=1}^d a^\alpha \frac{\partial \varphi}{\partial x^\alpha}$.

We replace the integration over space in (21) by the discrete scalar product $(.,.)_h$ defined as $(\varphi, \Phi)_h = \sum_i w_i \varphi_i \Phi_i$ and the derivative $\nabla(\varphi)$ by its approximation $D_h\varphi$. We then get the following system, whose solution is denoted by u^h:

$$\int_{\mathbb{R}_+} \left[\left(u^h(.,t), L_a^*(\varphi)\right)_h + \sum_{\alpha=1}^d \left(F^\alpha(u^h,.,t), D_{\alpha,h}\varphi\right)_h + \left(S(u^h,.,t),\varphi\right)_h \right] dt = 0. \tag{22}$$

Let us choose $u^h(x,t) = u_i^h(t)\ \forall x \in \Omega_i$. Making an integration by parts with respect to t, we find that (22) is true if and only if

$$\begin{cases} \text{(i)} \quad \dfrac{d}{dt}x_i = a(x_i,t), & x_i(0) = \xi_i, \\ \text{(ii)} \quad \dfrac{d}{dt}w_i = \text{div } (a(x_i,t))w_i, & w_i(0) = w_i^0, \\ \text{(iii)} \quad \dfrac{d}{dt}(w_i u_i^h) + w_i \displaystyle\sum_{\alpha=1}^d D_{\alpha,h}^*(F^\alpha)(u_i^h, x_i, t) = w_i S(u_i^h, x_i, t), & u_i^h(0) = u^0(\xi_i), \end{cases} \tag{23}$$

where $x_i = x_i(t)$, $u_i^h = u_i^h(t)$, and $w_i = w_i(t)$.

Remark 2.6. It is possible to build a strong scheme by simply replacing the derivative operator in (1) by the approximation operator $D_{\alpha,h}$ (14). We will not discuss this scheme in this paper, but it is studied in [8].

2.4. Application to Euler equations. Euler equations read as follows:

$$\begin{cases} \dfrac{\partial \rho}{\partial t} + \operatorname{div}(\rho \vec{v}) = 0, \\ \dfrac{\partial \rho \vec{v}}{\partial t} + \operatorname{div}(\rho \vec{v} \otimes \vec{v}) + \vec{\nabla} p = 0, \\ \dfrac{\partial E}{\partial t} + \operatorname{div}(\vec{v}(E+p)) = 0, \end{cases}$$

where ρ is the density, $\vec{v}$ the velocity (the transport field), E the total energy, and $p = p(\rho, u)$ the pressure. The total energy E is given by $E = \rho(u + \frac{1}{2}\|\vec{v}\|^2)$ with u the specific internal energy.

Euler equations can be rewritten as a conservation law. Applying the weak renormalized scheme (23), we get

$$\begin{cases} \dfrac{d}{dt} m_i = \dfrac{d}{dt}(w_i \rho_i) = 0, \\ \dfrac{d}{dt} v_i = -\displaystyle\sum_{j \in P} \frac{m_j}{\rho_i \rho_j}(p_i + p_j) A_{ij} - \sum_{j \in P} m_j \pi_{ij} A_{ij}, \\ \dfrac{d}{dt} u_i = \displaystyle\sum_{j \in P} \frac{m_j}{\rho_i \rho_j} p_j (v_i - v_j) A_{ij} + \frac{1}{2} \sum_{j \in P} \pi_{ij} (v_i - v_j) A_{ij}, \end{cases} \tag{24}$$

where we have chosen $A_{ij} = -A_{ji}$ as explained in Proposition 2.5. As we will show in section 4.1, when using an Euler explicit type of time discretization, one must add a numerical viscosity, which we have denoted π_{ij} in the above equations. For now, we will use the classical Von Neumann–Richtmeyer numerical viscosity (see [11] for instance). It remains to compute the density ρ and the smoothing length ε, as we want to move ε according to the distortion of the domain (see (6)) in order to keep a good number of neighbors per particle. We use

$$\frac{d}{dt}\rho_i = -\left(\rho \sum_{\alpha=1,\dots,d} D_h^\alpha v \right)_i = -\rho_i \sum_{j \in P} \frac{m_j}{\rho_j}(v_j - v_i) A_{ij}, \tag{25}$$

$$\frac{d}{dt}\varepsilon_i = \frac{1}{d}\left(\varepsilon \sum_{\alpha=1,\dots,d} D_h^\alpha v \right)_i = \frac{1}{d}\varepsilon_i \sum_{j \in P} \frac{m_j}{\rho_j}(v_j - v_i) A_{ij}. \tag{26}$$

Let us notice that the time-continuous scheme (24) satisfies the following properties: conservation of momentum (impulse and angular), conservation of total energy, second law of thermodynamics: $T_i \frac{ds_i}{dt} \geq 0$ (T = temperature, s = entropy).

Let us now present some numerical results that will demonstrate the efficiency of the renormalized method compared to the classical weighted particle methods. Consider the case of a one-dimensional implosion problem. The initial conditions are $\rho = 1$, $v = 1$, $p = 1$ in the domain $-1 \leq x \leq 0$ and $\rho = 1$, $v = -1$, $p = 1$ in the domain $0 \leq x \leq 1$. We used ghost particles to compute the wall boundary conditions at each end of the line of particles. The exact solution is composed of two shock waves, with a pressure plateau of $p_e = 2.92665$. We compare the results obtained from the classical SPH (smooth particle hydrodynamics) weighted particle scheme (based on the weak formulation of the conservation law as well) and the weak renormalized scheme. Both schemes have been implemented using the cubic B-spline (5) and with

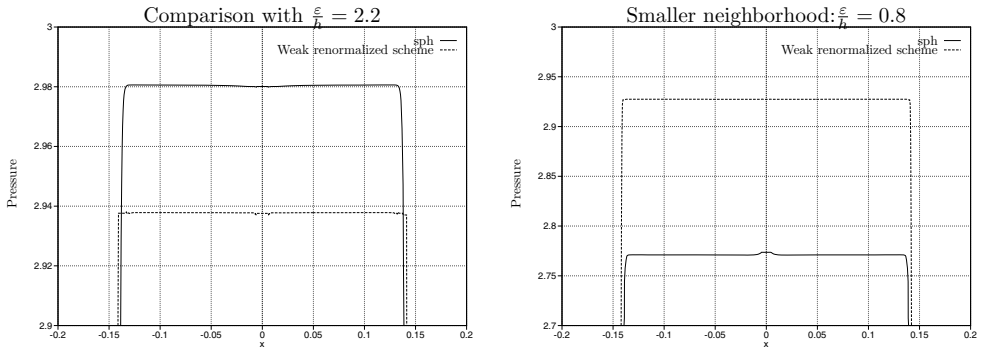


FIG. 2. *Comparison between SPH and weak renormalized schemes.*

two different values of the initial ratio $\frac{\varepsilon}{h}$: $\frac{\varepsilon}{h} = 2.2$, which is the optimal value of this ratio for a one-dimensional problem, and $\frac{\varepsilon}{h} = 0.8$. As explained earlier, ε will be computed according to (26), so that the number of neighbors for each particle will remain roughly the same within the computation. Our aim is not only to prove the efficiency of the renormalized method, but also to investigate the possibility of reducing the neighborhood of each particle to a minimum number of particles (for a one-dimensional problem, 2 particles) in order to reduce the computation time (as the computation time is directly linked to the number of neighbors per particles in a meshfree code). The results are given in Figure 2. The pressure plateaus obtained with the SPH scheme are $p = 2.984$ for $\frac{\varepsilon}{h} = 2.2$ and 2.780 for $\frac{\varepsilon}{h} = 0.8$. As for the weak renormalized scheme, we get $p = 2.934$ for $\frac{\varepsilon}{h} = 2.2$ and $p = 2.928$ for $\frac{\varepsilon}{h} = 0.8$. It is clear the weak renormalized scheme gives better results than the classical SPH weighted particle scheme, even in the case $\frac{\varepsilon}{h} = 2.2$, where the behavior of the classical SPH weighted particle scheme is supposed to be optimum. We refer to [2] for more numerical results and comparisons.

3. L^2 stability of the time continuous weak renormalized meshfree scheme. In this section, we will study the L^2 stability of the time continuous weak renormalized meshfree scheme (23) through the special case of Friedrichs systems. Let us consider the case where the physical problem admits a regular transport field $a = (a^\alpha)_{\alpha=1}^d$. Friedrichs systems are then written as

$$\begin{aligned} &\frac{\partial u}{\partial t} + \sum_{\alpha=1}^{d} D_\alpha(a^\alpha u) + \sum_{\alpha=1}^{d} D_\alpha(\tilde{\mathcal{F}}^\alpha u) + B_0 u = S \quad \forall (t,x) \in [0,T] \times \mathbb{R}^d, \\ &u(0,x) = u_0(x) \quad \forall x \in \mathbb{R}^d, \end{aligned} \tag{27}$$

where $u : [0,T] \times \mathbb{R}^d \to \mathbb{R}^p$ is the solution of the system, u_0 the initial condition $u_0 : \mathbb{R}^d \to \mathbb{R}^p$, S and B_0 the parameters of the source term such that $S : [0,T] \times \mathbb{R}^d \to \mathbb{R}^p$, $B_0 : [0,T] \times \mathbb{R}^d \to \mathcal{L}(\mathbb{R}^p)$, and $\tilde{\mathcal{F}}^\alpha$, $\alpha = 1, \ldots, d$, the flux matrices $\tilde{\mathcal{F}}^\alpha : [0,T] \times \mathbb{R}^d \to \mathcal{L}(\mathbb{R}^p)$ $\forall \alpha \in \{1, \ldots, d\}$, which are supposed symmetric. Applying (23) to the set of equations (27), one gets

$$\begin{cases} \dfrac{d}{dt} x_i = a_i, \quad x_i(0) = \xi_i, & \dfrac{d}{dt} w_i = w_i \mathrm{div}\,(a_i), \quad w_i(0) = w_i, \\ \dfrac{d}{dt}(w_i u_i^h) + w_i \displaystyle\sum_{\alpha=1}^{d} D^*_{\alpha,h}(\tilde{\mathcal{F}}^\alpha u^h)_i + w_i B_{0,i} u_i^h = w_i S_i, \quad u_i^h(0) = u^0(\xi_i), \end{cases} \tag{28}$$

where $x_i = x_i(t)$, $w_i = w_i(t)$, $a_i = a(x_i,t)$, $u_i^h = u^h(x_i,t)$, $S_i = S(x_i,t)$, $B_{0,i} = B_0(x_i,t)$, and $(\tilde{\mathcal{F}}^\alpha u^h)_i = (\tilde{\mathcal{F}}^\alpha u^h)(x_i,t) = \tilde{\mathcal{F}}^\alpha(x_i,t)u_i^h$.

3.1. L^2 stability analysis. We prove the following result, which is a key point for the stability analysis.

PROPOSITION 3.1. *Suppose that the shape vector A_{ij} satisfies the following: $\exists C$ s.t.*

$$\text{(i)} \quad \sup_{i\in P}\sum_{j\in P} w_j||A_{ij}|| \le \frac{C}{\varepsilon_0}, \tag{29a}$$

$$\text{(ii)} \quad \sup_{i\in P}||\sum_{j\in P} w_j A_{ij}|| \le C, \tag{29b}$$

$$\text{(iii)} \quad \sup_{i\in P}||A_{ij}+A_{ji}|| \le \frac{C}{\varepsilon_0^d}. \tag{29c}$$

Then, under the hypotheses (7), (11), *$\exists C$ such that $\forall\varphi\in(L^2(\mathbb{R}^d))^m$, $\forall\alpha\in\{1,\dots,d\}$, and for all regular symmetric matrices M^1, M^2 satisfying $M^1M^2 = M^2M^1$, we have*

$$\left|\left(M^2\varphi, D_{\alpha,h}(M^1\varphi)\right)_h\right| = \left|\left(M^1\varphi, D^*_{\alpha,h}(M^2\varphi)\right)_h\right| \le C||\varphi||_h^2\ , \tag{30}$$

where C depends on M^1, M^2 and their first order derivatives.

Proof. Let us first define the quantities S_1 and S_2:

$$\left(M^1\varphi, D_h^{\alpha,*}(M^2\varphi)\right)_h = \sum_{i,j} w_iw_j\left((M_i^1\varphi_i, M_i^2\varphi_i)A_{ij}^\alpha - (M_i^1\varphi_i, M_j^2\varphi_j)A_{ji}^\alpha\right) = S_1+S_2.$$

We will estimate S_1 and S_2 separately.

Term S_1.

$$||S_1|| = \sum_i w_i(M_i^1\varphi_i, M_i^2\varphi_i)\left(\sum_j w_jA_{ij}^\alpha\right). \tag{31}$$

Using the conditions (29b), one deduces (with $||M||_\infty = \sup_x|||M(x)||||$)

$$||S_1|| \le C||M^1M^2||_\infty\sum_i w_i||\varphi_i||^2 \le C||M^1M^2||_\infty||\varphi||_h^2. \tag{32}$$

Term S_2. Since M^1 and M^2 are symmetric, we have $S_2 = \frac{1}{2}(S_{21}+S_{22})$:

$$S_{21} = \sum_{i,j} w_iw_j(\varphi_i, (M_i^2M_j^1 - M_i^1M_j^2).\varphi_j)A_{ji}^\alpha, \tag{33}$$

$$S_{22} = \sum_{i,j} w_iw_j(\varphi_i, (M_i^2M_j^1).\varphi_j)(A_{ij}^\alpha + A_{ji}^\alpha). \tag{34}$$

From $M_i^2M_j^1 - M_i^1M_j^2 = (M_i^2-M_j^2)M_j^1 + (M_j^1-M_i^1)M_j^2$, and the hypotheses (7), (11), we obtain the following from the regularity of the matrices M^1 and M^2:

$$|||M_i^2M_j^1 - M_i^1M_j^2||| \le C_{\varepsilon_0}(||M^2||_\infty||DM^1||_\infty + ||M^1||_\infty||DM^2||_\infty)\varepsilon_0 \le C_2\varepsilon_0,$$

where DM represents the derivatives of the matrix M. We deduce

(35)
$$||S_{21}|| \leq C_2\varepsilon_0 \sum_{i,j} w_i w_j ||\varphi_j||\ ||\varphi_i||\ ||A_{ij}|| \leq C_2 \frac{\varepsilon_0}{2} \sum_{i,j} w_i w_j (||\varphi_j||^2 + ||\varphi_i||^2)||A_{ij}||$$
$$\leq C_2 \frac{\varepsilon_0}{2} \sum_i w_i ||\varphi_i||^2 \left(\sum_j w_j (||A_{ij}|| + ||A_{ji}||) \right).$$

Using (29a), it is found that $\varepsilon_0 \sum_{j\in P} w_j ||A_{ij}||$ is uniformly bounded. Arguing that $\sum_{j\in P} = O(\frac{\varepsilon_0^d}{h^d})$ and using (29c), we get

$$\sum_{j\in P} w_j ||A_{ij} + A_{ji}|| \leq C. \tag{36}$$

As a consequence, we also get that $\varepsilon_0 \sum_{j\in P} w_j ||A_{ji}|| \leq C$. It follows that

$$||S_{21}|| \leq (||M^2||_\infty ||DM^1||_\infty + ||M^1||_\infty ||DM^2||_\infty) C ||\varphi||_h^2. \tag{37}$$

We then find an upper bound of the term S_{22} from the condition (29c):

(38)
$$||S_{22}|| \leq \frac{||M^2||_\infty ||M^1||_\infty}{2} \sum_{i,j} w_i w_j (||\varphi_i||^2 + ||\varphi_j||^2)||A_{ij} + A_{ji}||$$
$$\leq ||M^2||_\infty ||M^1||_\infty C ||\varphi||_h^2.$$

The proof is achieved by combining (32), (37), and (38). ∎

Remark 3.2. The conditions (i) and (ii) are consequences of hypotheses (7), (11), (12), Lemma 2.1, and the invertibility of the matrices $E(x)$. The last condition (iii) is important for ensuring the L^2 stability. The validity of these three hypotheses will be established in section 3.2.

We can now prove the main stability result for the time continuous scheme.

THEOREM 3.3. *Consider the same hypothesis as Proposition* 3.1*, and also suppose that* $\tilde{\mathcal{F}}^\alpha$, $\alpha = 1, \ldots, d$, *and* a *are in* $L^\infty(\mathbb{R}_+; W^{1,\infty}(\mathbb{R}^d))$, S *and* B_0 *are in* $L^\infty(\mathbb{R}_+ \times \mathbb{R}^d)$, *and* $u_0 \in L^\infty(\mathbb{R}^d)$. *Then the solution of the time continuous weak renormalized scheme applied to the Friedrichs system* (28) *is bounded in* $C(0, T; (L^2(\mathbb{R}^d))^p)$.

Proof. Let us differentiate $||u^h||_h^2 = \sum_{i\in P} w_i u_i^2$ with respect to time t:

$$\frac{d}{dt}(||u^h||_h^2) = 2 \sum_{i\in P} u_i^h . \frac{d}{dt}(w_i u_i^h) - \sum_{i\in P} w_i ||u_i^h||^2 \mathrm{div} a_i.$$

Using the ODE of the scheme (28), we get

$$\frac{d}{dt}(||u^h||_h^2) = -2 \sum_{\alpha=1}^d (u^h, D_{\alpha,h}(\tilde{\mathcal{F}}^\alpha . u^h))_h - (u^h, u^h \mathrm{div} a)_h + 2(u^h, B_0 u^h + S)_h.$$

We thus obtain (using Proposition 3.1)

$$\frac{d}{dt}(||u^h||_h^2) \leq (2C + ||\mathrm{div} a||_\infty + 2||B_0||_\infty)||u^h||_h^2 + ||u^h||_h^2 + ||S||_h^2 \leq C_1 ||u^h||_h^2 + C_2.$$

A direct integration gives

$$||u^h(., t)||_h^2 \leq ||u^h(., 0)||_h^2 \exp(C_1 t) + \frac{C_2}{C_1}(\exp(C_1 t) - 1).$$

The L^2 stability follows. ∎

From this stability result, we can prove the following convergence theorem.

THEOREM 3.4. *Let u^h be an approximate solution to* (27) *computed with the weak renormalized scheme* (28)*. Under the same assumptions as Theorem* 3.3*, u^h weakly converges towards the exact solution u of* (27) *in $L^2(0,T;(L^2(R^d))^p)$.*

Proof. We proved stability in Theorem 3.3; it remains to prove consistency. If consistency can be proved for a general conservation law, it would then also be true for Friedrichs systems. We want to prove that if u^h computed from the scheme (23) converges towards a function $\bar{u}$ as ε_0 and $h \to 0$, then $\bar{u}$ is a weak solution of (1), thus satisfying (21). From the construction of the scheme (23), it is clear that $\bar{u}$ satisfies the following: $\forall \varphi \in C_c^2(\mathbb{R}^d \times \mathbb{R}^{+,*})$

$$\int_{\mathbb{R}^+} \left[(\bar{u}(.,t), L_a^*(\varphi))_h + \sum_{\alpha=1}^{d} (F^\alpha(\bar{u},.,t), D_{\alpha,h}\varphi)_h + (S(\bar{u},.,t),\varphi)_h \right] dt = 0. \tag{39}$$

The discrete scalar product $(\varphi,\Phi)_h = \sum_i w_i \varphi_i \Phi_i$ used to develop the weak renormalized scheme (23) tends towards the scalar product in $L^2(\mathbb{R}^d)$ when h tends to 0. From Propositions 2.4 and 2.5, we know that $D_{\alpha,h}\varphi \to_{\varepsilon_0 \to 0} D_\alpha \varphi$. Applying the Lebesgue theorem, we establish that $\bar{u}$ satisfies (21), and thus is a weak solution of (1). ∎

Remark 3.5. The convergence of the weak renormalized scheme applied to Friedrichs systems, and discretized in time, towards a physically relevant solution is proved in [9]. A convergence rate is also given.

3.2. Geometrical interpretation. In the previous subsection, we established a set of conditions that the kernels A_{ij} must satisfy in order to obtain the L^2 stability for linear systems of PDEs. We will now discuss the validity of these conditions for a chosen regularizing kernel μ. Let us consider the following regularizing function:

$$\mu_{ij} = \mu(x_i,x_j) = \frac{1}{\varepsilon_{ij}^{d+2}} r_{ij}\theta_{ij}, \tag{40}$$

where $r_{ij} = x_j - x_i$ and $\theta_{ij} = \theta(x_i,x_j)$, with θ a function with compact support. In this paper, we will consider the function θ that classically appears in the smoothed particle approximation, with variable smoothing length:

$$\theta_{ij} = \frac{\varepsilon_{ij}}{||r_{ij}||} \zeta' \left(\frac{||r_{ij}||}{\varepsilon_{ij}} \right), \tag{41}$$

where ζ is, for instance, the cubic B-spline (5).

Remark 3.6. We chose a symmetric variable smoothing length $\varepsilon_{ij} = \frac{\varepsilon_i + \varepsilon_j}{2}$. One could also choose ε_i or ε_j instead of ε_{ij} in (41). The results proved in the following still hold for these two other choices.

The inverses E_i of the renormalization matrices B_i are rewritten as

$$E_i^{\alpha\beta}(t) = \sum_{j \in P} w_j \frac{r_{ij}^\alpha r_{ji}^\beta}{\varepsilon_{ij}^{d+2}} \theta_{ij}, \quad (\alpha,\beta) \in 1,\dots,d^2. \tag{42}$$

We will also use the following notation:

$$\Delta_i^h(t) = \sum_{j \in P} w_j \frac{r_{ji}}{\varepsilon_{ij}^{d+2}} \theta_{ij}. \tag{43}$$

We will in the following give three results that will identify some geometrical conditions under which (29) are satisfied. The following proposition gives a first set of sufficient conditions that ensure that (29) and (19) are satisfied.

PROPOSITION 3.7. *Suppose that* (7) *is satisfied, and that*

$$\text{(i)}\quad E_i \text{ is uniformly invertible}, \tag{44a}$$

$$\text{(ii)}\quad \sup_{i\in P} ||\Delta_i^h(t)|| = O(1), \tag{44b}$$

$$\text{(iii)}\quad \sup_{i\in P, j\in V_i} ||E_i - E_j|| = O(\varepsilon_0); \tag{44c}$$

then the inequalities (29) *and the condition* (19) *are verified:*

(A_1) (i) $\Rightarrow ||B(.)||_\infty \le C \Rightarrow$ (29a),
(A_2) (i) *and* (ii) $\Rightarrow$ (29b),
(A_3) (i) *and* (iii) $\Rightarrow$ (29c),
(A_4) (i) *and* (iii) $\Rightarrow$ (19).

Proof.

(A_1) (44a) $\Rightarrow ||B(.)||_\infty \le C$ is obvious.
Hypothesis (7) on ε being satisfied, and given the definition of μ (40), (41), one can use Lemma 2.1 and obtain (29a):

$$\sup_{i\in P}\sum_{j\in P} w_j||A_{ij}|| \le \sup_{i\in P}\sum_{j\in V_i} w_j||B_i||\ ||\mu_{ij}|| \le \frac{C}{\varepsilon_0}.$$

Using $\frac{B_i+B_j}{2}$ instead of B_i gives the same result.

(A_2) $\sum_{j\in P} w_j A_{ij} = \sum_{j\in P} w_j B_i.\mu_{ij} = B_i.(\sum_{j\in P} w_j\mu_{ij})$. Hypotheses (44a) and (44b) allow us to conclude the following: $||\sum_{j\in P} w_j A_{ij}|| \le ||B_i||\ ||\sum_{j\in P} w_j\mu_{ij}|| \le C$. If one uses $\frac{B_i+B_j}{2}$ instead of B_i in the definition of the kernel A_{ij}, one has to use hypothesis (19), which is proved as assertion A_4.

(A_3) From the definition of μ_{ij}, we have $\mu_{ij} + \mu_{ji} = \frac{1}{\varepsilon_{ij}^{d+2}} r_{ij}(\theta_{ij} - \theta_{ji}) = 0$. Considering $A_{ij} + A_{ji} = B_j.(E_j - E_i).A_{ij} + B_j.(\mu_{ji} + \mu_{ij})$, we obtain

$$||A_{ij} + A_{ji}|| \le ||B_j||\ ||E_j - E_i||\ ||A_{ij}|| + ||B_j||\ ||\mu_{ji} + \mu_{ij}||,$$

which can be upper bounded from (44a), (44c):

$$||A_{ij} + A_{ji}|| \le C\varepsilon_0||A_{ij}|| \le C\varepsilon_0||B_i||\ ||\mu_{ij}|| = O\left(\frac{1}{\varepsilon_0^d}\right).$$

If we use $\frac{B_i+B_j}{2}$ instead of B_i, (29c) is automatically satisfied.

(A_4) (44c) implies that $E_j = E_i + O(\varepsilon_0)$. Considering (44a), we deduce (19). ∎

Let us now check that the conditions (44) are satisfied for the chosen function θ. Since ε varies, we introduce a bounded function $\kappa(x,t)$ such that $\varepsilon_{ij} = \kappa_{ij}\varepsilon_0 = \frac{\kappa_i+\kappa_j}{2}\varepsilon_0$, where $\kappa_i = \kappa(x_i,t)$. Hypothesis (7) becomes

$$0 \le \kappa_{ij} \le C \quad \forall i,j \in P. \tag{45}$$

Let us now introduce the matrix of the Lagrangian coordinates DX and the Jacobian of the transformation J:

$$DX(\xi,t) = \left(\frac{\partial x^\alpha}{\partial \xi^\beta}(\xi,t)\right)_{\alpha,\beta=1,d}, \quad J(\xi,t) = \det(DX(\xi,t)). \tag{46}$$

Using these elements, we have the following Taylor series expansion:

$$\begin{aligned} &r_{ij} = DX_i.\xi_{ij} + O(\varepsilon_0^2), \qquad w_j = w_j^0(J_i + O(\varepsilon_0)), \\ &J_j = J_i + O(\varepsilon_0), \quad \kappa_{ij} = \kappa_i + \frac{\nabla\kappa_i}{2}.DX_i.\xi_{ij} + O(\varepsilon_0^2), \end{aligned} \tag{47}$$

where $\xi_{ij} = \xi_j - \xi_i$, $DX_i = DX(\xi_i, t)$, and $J_i = J(\xi_i, t)$. Let us denote

$$\nu = \frac{\xi}{\varepsilon_0}. \tag{48}$$

Introducing the Taylor series (47) into the definitions (42) and (43), we get

$$\Delta_i^h(t) = J_i DX_i.\tilde{\Delta}_i^h(t) + O(1), \quad E_i(t) = J_i DX_i.\tilde{E}_i(t).DX_i^T + O(\varepsilon_0), \tag{49}$$

with

$$\tilde{\Delta}_i^h(t) = \frac{1}{\varepsilon_0}\sum_{j\in P} b_{ij}(t)\nu_{ji}, \quad \tilde{E}_i^{\alpha\beta}(t) = \sum_{j\in P} b_{ij}(t)\nu_{ij}^\alpha\nu_{ji}^\beta, \quad b_{ij}(t) = w_j^0\frac{\theta_{ij}(t)}{(\kappa_{ij})^{d+2}\varepsilon_0^d}. \tag{50}$$

From these expansions, we can prove the following result.

PROPOSITION 3.8. *Suppose that the transport field $a(x,t)$ is regular, $a \in (\mathcal{C}^2(\mathbb{R}^d \times \mathbb{R}_+))^d$, and that $b_{ij}(t)$ and ν_{ij} defined by* (50) *and* (48) *satisfy the following:*

(51a) (i) *for $d \geq 2$, no particle initially has its neighbors on the same line;*

(51b) (ii) $\displaystyle\sum_{j\in P} b_{ij}(t)\nu_{ij} = O(\varepsilon_0)$;

(51c) (iii) $\displaystyle\sum_{j\in P} b_{ij}(t)\nu_{ij}^\alpha\nu_{ij}^\beta - C_0(t,\xi_i)\delta^{\alpha\beta} = O(\varepsilon_0)$,

where $C_0(t,\xi)$ is bounded with respect to t and Lipschitz with respect to ξ. Then the conditions (44) *of Proposition* 3.7 *are satisfied.*

Proof.

Condition (44a). From its definition (41), θ is positive, which implies that the elements $b_{ij}(t)$ are positive. Consider Λ in the following quadratic form:

$$\Lambda_i(x) = \frac{1}{2}\sum_{\alpha,\beta\in\{1,\dots,d\}}\sum_{j\in V_i} x^\alpha x^\beta b_{ij}(t)\nu_{ij}^\alpha\nu_{ij}^\beta,$$

where V_i is the set defined in (9).

We clearly have $\tilde{E}_i(t) = -D_2\Lambda_i(x)$. $\tilde{E}_i(t)$ is thus the Hessian matrix of the quadratic form Λ_i. If Λ_i is uniformly strictly convex, then $\tilde{E}_i(t)$ is symmetric negative definite and consequently uniformly invertible. Due to the definition of θ $A_{ij} \geq 0$ $\forall i,j \in P$, the quadratic form can be rewritten as

$$\Lambda_i(x) = \frac{1}{2}\sum_{j\in V_i} b_{ij}(t)\left(\sum_{\alpha\in\{1,\dots,d\}} x^\alpha\nu_{ij}^\alpha\right)^2 = \frac{1}{2}\sum_{j\in V_i} b_{ij}(t)|x.\nu_{ij}|^2 \geq b_{*,i}||x||^2,$$

where $b_{*,i} = \inf_{y;||y||=1}\frac{1}{2}\sum_{j\in V_i} b_{ij}(t)|y.\nu_{ij}|^2$. The quantities $b_{*,i}$ must be different from 0 in order to have Λ strictly convex. When working in a space of dimension

$d \geq 2$, if all the particle neighbors of i are initially on the same straight line, it suffices to take y orthogonal to this straight line to get $b_{*,i} = 0$, as we would then have $|y.\nu_{ij}| = 0\ \forall j \in V_i$. Consequently, we must impose a condition on the initial distribution of particles, so that no particle will have all its neighbors on the same straight line (condition (51a)).

Condition (44b). We have $\Delta_i^h(t) = J_i DX_i.\tilde{\Delta}_i^h(t) + O(1)$. From (51b), and considering the regularity of the transport field, we have

$$||J_i DX_i.\tilde{\Delta}_i^h(t)|| \leq C||J||_{L^\infty(P)}||DX||_{L^\infty(P)}$$

and we deduce (44b).

Condition (44c). Let us write

$$\begin{aligned} E_i - E_j &= J_i DX_i.\tilde{E}_i.DX_i^T - J_j DX_j.\tilde{E}_j.DX_j^T + O(\varepsilon_0) \\ &= (J_i DX_i - J_j DX_j).\tilde{E}_i.DX_i^T + J_j DX_j.\tilde{E}_i.(DX_i - DX_j)^T \\ &\quad + J_j DX_j.(\tilde{E}_i - \tilde{E}_j).DX_j^T + O(\varepsilon_0). \end{aligned}$$

Using the hypothesis of regularity of the transport field a, we prove

$$||J_i - J_j||_{L^\infty(P)} \leq C||\xi_j - \xi_i||,$$

$$||DX_i - DX_j||_{L^\infty(P)} \leq C||\xi_j - \xi_i||.$$

From (51c), we deduce $||\tilde{E}_i||_{L^\infty(P)} \leq C$ and $||\tilde{E}_i - \tilde{E}_j||_{L^\infty(P)} \leq C\varepsilon_0$. Knowing that $j \in V_i$, we deduce $||\xi_j - \xi_i|| \leq C\varepsilon_0$ and consequently (44c). ∎

We now identify a special initial geometry of the distribution of particles sufficient to satisfy the conditions (51).

PROPOSITION 3.9. *Suppose that the particles are initially distributed on a regular net of characteristic length h:*

$$\text{(52)} \qquad \begin{cases} P = \mathbb{Z}^d = \{j = (j_1, \ldots, j_d), \quad j_l \in \mathbb{Z}, \quad l \in \{1, \ldots, d\}\}, \\ x_j = (x_{j_1}^1, \ldots, x_{j_d}^d), \qquad x_{j_l}^l = j_l h. \end{cases}$$

Suppose also that the transport field $a \in (\mathcal{C}^2(\mathbb{R}^d \times \mathbb{R}_+))^d$ and the smoothing length coefficient $\kappa \in \mathcal{C}^2(\mathbb{R}^d \times \mathbb{R}_+)$. Then the conditions (51) *are satisfied.*

Proof. Since the particles are equidistributed according to (52), we will consider in the following $\Omega_i(0) = \Pi_{\alpha=1}^d [(i_\alpha - \frac{1}{2})h, (i_\alpha + \frac{1}{2})h]$ such that $w_i^0 = h^d$.

Condition (51b). Since we have considered a regular initial distribution of particles (52), for every particle j, there exists a particle $s(j)$ symmetric of j with respect to i. This particle $s(j)$ satisfies $\nu_{ij} + \nu_{is(j)} = 0$, and we have

$$\sum_{j \in P} b_{ij}(t)\nu_{ij} = \frac{1}{2} \sum_{j \in P} (b_{ij}(t) - b_{is(j)}(t))\nu_{ij}.$$

Using the Taylor series expansions (47), together with the definition (41) of θ, and considering only the terms of order ε_0, we get

$$\begin{aligned} b_{ij}(t) &= \left(\frac{h}{\varepsilon_0}\right)^d \frac{1}{\kappa_i^{d+2}} \left(||DX_i.\nu_{ij}||^{-1} \zeta' \left(\left\| \frac{DX_i.\nu_{ij}}{\kappa_i} \right\| \right) + O(\varepsilon_0) \right) \\ &= \left(\frac{h}{\varepsilon_0}\right)^d \frac{1}{\kappa_i^{d+2}} \left(||DX_i.\nu_{is(j)}||^{-1} \zeta' \left(\frac{||DX_i.\nu_{is(j)}||}{\kappa_i} \right) + O(\varepsilon_0) \right) = b_{is(j)}(t) + O(\varepsilon_0), \end{aligned}$$

and condition (51b) follows.

Condition (51c). Let us study the cases $\alpha = \beta$ and $\alpha \neq \beta$ separately.

For $d \geq 2$ and $\alpha \neq \beta$, from the initial distribution of particles (52), for all particles j there exists a symmetric particle $s^{\alpha\beta}(j)$ with respect to i that satisfies $\nu_{ij}^{\alpha} + \nu_{is^{\alpha\beta}(j)}^{\alpha} = 0$ and $\nu_{ij}^{\beta} = \nu_{is^{\alpha\beta}(j)}^{\beta}$, such that

$$\sum_{j\in P} b_{ij}(t)\nu_{ij}^{\alpha}\nu_{ij}^{\beta} = \frac{1}{2}\sum_{j\in P}(b_{ij}(t) - b_{is^{\alpha\beta}(j)}(t))\nu_{ij}^{\alpha}\nu_{ij}^{\beta}.$$

Using the Taylor series expansion (47), we get

$$\sum_{j\in P} b_{ij}(t)\nu_{ij}^{\alpha}\nu_{ij}^{\beta} = \sum_{j\in P}\left(\frac{h}{\varepsilon_0}\right)^{d} O(\varepsilon_0) = O(\varepsilon_0). \tag{53}$$

Thus, in the case $\alpha \neq \beta$, the condition (51c) is satisfied.

For case $\alpha = \beta$, $d \geq 1$, let us take

$$\sum_{j\in P} b_{ij}(t)\nu_{ij}^{\alpha}\nu_{ij}^{\alpha} = \left(\frac{h}{\varepsilon_0}\right)^{d}\sum_{j\in P}\frac{1}{\kappa_{ij}^{d+2}}\theta_{ij}(t)(\nu_{ij}^{\alpha})^2 = C_0(t,\xi_i).$$

Combining the two cases, we obtain

$$\sum_{j\in P} b_{ij}(t)\nu_{ij}^{\alpha}\nu_{ij}^{\beta} = C_0(t,\xi_i)\delta^{\alpha\beta} + O(\varepsilon_0).$$

It remains to make sure that $C_0(t,\xi)$ is Lipschitz with respect to ξ. Using the Taylor expansions (47), we prove that

$$C_0(\xi_i,t) = \left(\frac{h}{\varepsilon_0}\right)^{d}\frac{1}{\kappa_i^{d+1}}\sum_{j\in P}\frac{1}{||DX_i.\nu_{ij}||}\zeta'\left(\frac{||DX_i.\nu_{ij}||}{\kappa_i}\right)(\nu_{ij}^{\alpha})^2 + O(\varepsilon_0).$$

From the regularity of the transport field a and of the function κ, we can prove $C_0(\xi_i,t) = C_0(\xi_k,t) + O(\xi_i - \xi_k) + O(\varepsilon_0)$. □

In this section, we proved that if the transport field and the smoothing length are regular enough, and if the initial distribution of particles is regular, the time continuous scheme will be L^2 stable. The smoothing length regularity is clearly linked to the transport fields' regularity when using (6) to compute ε. We also established that the existence of the renormalization matrix at each particle $i \in P$ depends on whether the neighborhood of i is on one line. The case where the neighbors are on one line (or close to a line) corresponds to the case of a very flat element in finite element methods. However, the possibility of varying the smoothing length ε allows us to increase the neighborhood in order to avoid these situations.

Although the regularity $a \in (\mathcal{C}^2(\mathbb{R}^d \times \mathbb{R}_+))^d$ is required in our proofs, the numerical study in section 2.4 clearly shows that even when the transport field is initially not regular, the weak renormalized code gives accurate results. In practice, we do not require $a \in (\mathcal{C}^2(\mathbb{R}^d \times \mathbb{R}_+))^d$.

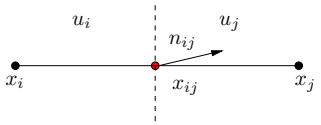


FIG. 3. *Interaction between two particles.*

4. On the L^∞ and BV stabilities of the time discretized weak renormalized meshfree scheme.

4.1. Time discretization: Analogy with finite volume. Our aim is to discretize the numerical scheme (23). Using an explicit Euler-like scheme leads to an unstable scheme. One means of stabilizing it is to introduce some numerical viscosity by analogy with finite volume schemes. Thus, let n_{ij} be the normal vector $n_{ij} = \frac{A_{ij}}{||A_{ij}||}$, and let us consider the interaction between two particles i and j as shown in Figure 3.

Consider the Riemann problem at the interface of position $x_{ij} = \frac{x_i+x_j}{2}$ in the unidimensional direction n_{ij}:

$$\frac{\partial v}{\partial t} + \frac{\partial}{\partial x}(F(v, x_{ij}, t).n_{ij}) = 0, \quad v(x, t^n) = \begin{cases} u_j^{h,n} \text{ if } x > x_{ij}, \\ u_i^{h,n} \text{ if } x < x_{ij}. \end{cases} \tag{54}$$

Let us introduce a numerical flux $g(n_{ij}, u_i^h, u_j^h)$ in order to define a numerical scheme representing the interaction between i and j. The numerical flux $g(n, u, v)$, analogous to a finite volume flux, includes the numerical viscosity. We suppose that this flux belongs to the E-scheme flux class (see [12, 15]) and satisfies

$$\text{(i)} \quad g(n, u, u) = F(u, x, t).n, \quad \text{(ii)} \quad g(n, u, v) = -g(-n, v, u). \tag{55}$$

The numerical viscosity is then defined as

$$Q(n_{ij}, u_i^h, u_j^h) = \frac{F(u_i^h, x_{ij}, t).n_{ij} - 2g(n_{ij}, u_i^h, u_j^h) + F(u_j^h, x_{ij}, t).n_{ij}}{u_j^h - u_i^h}. \tag{56}$$

Let us now introduce the contribution of particle j on particle i, which we denote by $u_i^{h,n+1,j}$ (with $F_{i,ij}^n = F(u_i^{h,n}, x_{ij}, t^n)$ and $Q_{ij}^n = Q(n_{ij}, u_i^{h,n}, u_j^{h,n})$):

$$\begin{aligned} u_i^{h,n+1,j} &= u_i^{h,n} - \frac{\Delta t}{\overline{\Delta x}_i^n} \overbrace{\left(\frac{F_{j,ij}^n - F_{i,ij}^n}{2}.n_{ij} + Q_{ij}^n (u_i^{h,n} - u_j^{h,n}) \right)}^{C_{ij}^n} \\ &= u_i^{h,n} - \frac{\Delta t}{\overline{\Delta x}_i^n} \left(g(n_{ij}, u_i^{h,n}, u_j^{h,n}) - F_{i,ij}^n \right), \end{aligned} \tag{57}$$

where C_{ij}^n is the incremental coefficient and

$$\overline{\Delta x}_i^n = \frac{1}{2\sum_{l\in P} w_l^n ||A_{il}^n||}, \tag{58}$$

and is clearly equivalent to a space step as $\sum_{l\in P} w_l^n ||A_{il}^n|| = O(\frac{1}{\varepsilon_0})$. Indeed, it can be viewed as a mean distance separating i from its neighbors. In (58), we use the notation $A_{ij}^n = A_{ij}(t^n)$, since it is clear from the definition of A_{ij} that it depends on

the time variable, since x_i and x_j depend on the time. When referring to the γth component of the vector A_{ij}^n, we will write $A_{ij}^{n,\gamma}$.

Let us denote

$$\alpha_{ij}^n = \frac{w_j^n ||A_{ij}^n||}{\sum_{l\in P} w_l^n ||A_{il}^n||}. \tag{59}$$

From a convex combination of these single contributions, we get $\tilde{u^h}_i^{n+1}$:

$$\tilde{u^h}_i^{n+1} = \sum_{j\in P} \alpha_{ij}^n u_i^{h,n+1,j} - 2\Delta t \sum_{j\in P} w_j^n F_{i,ij}^n A_{ij}^n + \Delta t S_i^n, \tag{60}$$

where $S_i^n = S(u_i^{h,n}, x_i, t^n)$. The convex combination $\tilde{u^h}_i^{n+1}$ corresponds to an explicit Euler time discretized scheme, where the convection part is not taken into account. Taking the convection into account, we get the time discretized scheme, with constant time step Δt:

$$\left\{ \begin{array}{l} \tilde{u^h}_i^{n+1} = \displaystyle\sum_{j\in P} \alpha_{ij}^n u_i^{h,n+1,j} - 2\Delta t \sum_{j\in P} w_j^n F_{i,ij}^n A_{ij}^n + \Delta t S_i^n \\ \qquad\qquad = u_i^{h,n} - \Delta t \left(\mathrm{div}_h^* F_{i,ij}^n + \displaystyle\sum_{j\in P} w_j^n Q_{ij}^n ||A_{ij}^n|| - S_i^n \right), \\ \dfrac{w_i^{n+1}}{w_i^n} u_i^{h,n+1} = \tilde{u^h}_i^{n+1}. \end{array} \right. \tag{61}$$

Remark 4.1. We choose a constant time step in order to simplify the notation. The stability results still hold for a variable time step.

Finally, the expression of $u_i^{h,n+1,j}$ (57) allows us to give a CFL condition:

$$\sup_{\substack{|u|,\,|v| \le C(T,\,||u_0||_\infty) \\ i,\,j\,\in\,P,\ n\Delta t \le T}} \frac{\Delta t}{\overline{\Delta x}_i} |Q(n_{ij}, u, v)| \le 1 - \beta, \quad 0 < \beta < 1. \tag{62}$$

4.2. L^∞ and BV stability results. We will consider the following regularities on F, S, and a:

$$\begin{array}{l} a \in (\mathcal{C}^2(\mathbb{R}^d \times \mathbb{R}_+))^d, \ \ F \in (\mathcal{C}^2(\mathbb{R} \times \mathbb{R}^d \times \mathbb{R}_+))^d, \\ F(0,.,.), D_\beta F(0,.,.) \in (L^1(\mathbb{R}^d \times \mathbb{R}_+))^d, \ \ \beta = 1, \ldots, d, \\ S \in C^1(\mathbb{R} \times \mathbb{R}^d \times \mathbb{R}_+), \ \ S(0,.,.) \in L^1(\mathbb{R}^d \times \mathbb{R}_+). \end{array} \tag{63}$$

Since we will introduce here some stability results with functional norms, let us introduce the function $u^h(x,t)$ associated with the scheme (61):

$$\left\{ \begin{array}{ll} u^h(x,t) = u^h(x_i^{n+1}, t^{n+1}) & \text{if } \ t \in]t^n, t^{n+1}], \ x \in \Omega_i^{n+1} \text{ with } \ n \ge 1, \\ u^h(x,t) = u^h(x_i^0, 0) & \text{if } \ t \in [t^0, t^1], \ x \in \Omega_i^0. \end{array} \right. \tag{64}$$

We first give the L^∞ and L^1 properties.

PROPOSITION 4.2. *Let u^h be the resulting function of the scheme* (61), (64). *Let us assume that a, S, and F satisfy* (63) *and the initial condition $u_0 \in L^\infty(\mathbb{R}^d) \cap L^1(\mathbb{R}^d)$. We also suppose that the coefficients A_{ij}^n satisfy* (29) *$\forall i \in P$, $\forall j \in P$, and*

for all time t^n. Then $\forall T > 0$, there are two constants $C_\infty(T, F, S, a, ||u_0||_\infty)$ and $C_1(T, F, S, a, ||u_0||_\infty, ||u_0||_1)$ such that if the time step Δt satisfies the CFL condition (62)*, we have the following $\forall t \leq T$:*

$$||u^h(.,t)||_{L^\infty(\mathbb{R}^d)} \leq C_\infty(T, F, S, a, ||u_0||_{L^\infty(\mathbb{R}^d)}), \tag{65}$$

$$||u^h(.,t)||_{L^1(\mathbb{R}^d)} \leq C_1(T, F, S, a, ||u_0||_{L^\infty(\mathbb{R}^d)}, ||u_0||_{L^1(\mathbb{R}^d)}). \tag{66}$$

Proof. The proof is based on the study of SPH [3]. However, some differences occur in the estimates, and the proof is given in [8, pp. 202–203]. ∎

We now focus on the a priori BV estimate.

PROPOSITION 4.3. *Let us suppose that $u_0 \in L^\infty(\mathbb{R}^d) \cap L^1(\mathbb{R}^d)$. Then, under the same assumptions as in Proposition* 4.2*, $\forall T > 0$, there exists a positive constant $C_{est}(T, F, S, a, u_0)$ such that the following inequality holds:*

$$\sum_{\substack{(i,j) \in P^2 \\ n;\, n\Delta t \leq T}} \Delta t w_i^n w_j^n \frac{\Delta t}{\overline{\Delta x_i}} |C_{ij}^n|^2 |u_i^{h,n} - u_j^{h,n}|^2 |A_{ij}^n| \leq C_{est}, \tag{67}$$

where $\overline{\Delta x_i}$ is defined by (58) *and $C_{ij}^n = C(n_{ij}, u_i^h, u_j^h)$ is the incremental coefficient* (57)*. Furthermore, if the particles belong to a bounded domain $\Omega \in \mathbb{R}^d$, then there exists a constant $C_{BV}(u_0, F, S, a, T, \Omega)$ such that*

$$\sum_{\substack{i \in P;\, x_i \in \Omega \\ n;\, t^n \leq T}} \Delta t w_i^n |u_i^{h,n+1} - u_i^{h,n}| \leq C_{BV}(\sqrt{h} + h). \tag{68}$$

Proof. In order to prove (67), we use the local energy estimate in [1], and (68) is deduced from inequality (67), together with the regularity of the transport field a. The proof is standard and is given in [8, p. 203]. ∎

Results from this section are used to prove the convergence of the weak renormalized meshfree scheme applied to the scalar nonlinear conservation law in [10].

5. Comparison with hybrid finite volume schemes.

5.1. Basic principle. In subsection 4.1, we explained how the time discretized weak renormalized scheme can be built from a finite volume interpretation. The same process allows us to define another type of renormalized scheme, directly based on finite volume Godunov methods: *the hybrid renormalized finite volume schemes*. The basic principle is to use the definition (55) of the numerical flux $g(n, u, v)$ in order to redefine the time continuous scheme. It remains to define the fluxes $g(n_{ij}, u_i^h, u_j^h)$. As explained in the following subsection, we use a Godunov-type method, where the fluxes at each cell interface are computed from solving a Riemann problem of type (54) at each interface. For the computation, we will not be able to use the actual computed transport field, but instead we will have to introduce a smoother transport field to move the particles. This approach is connected with the notion of arbitrary Lagrange Euler approximation and is also in some sense a generalization of ideas developed in [6].

Higher order version of the method. In the field of particle methods a classical tool for increasing accuracy is for increasing the smoothness of the kernel and making its momentum equal to zero. This approach has been used in SPH calculations, for example, by using a "super Gaussian" kernel. The increase of accuracy by modifying

the kernel introduces some instability due to the nonpositiveness of the kernel. Thus, we propose a different method based upon the well-known MUSCL schemes developed by Van Leer [16] for finite difference schemes, and extended more recently to finite volume methods (see, e.g., [1] for a mathematical analysis). The idea is to account for information given by $\nabla^h(u)$ (where $\nabla^h(u)$ is the approximate gradient operator computed with (14)) to compute a numerical flux which increases the accuracy of the method: we replace the flux $g(n_{ij}, u_i, u_j)$ which approximates the flux of the conservation law (54) located at x_{ij} by $g(\boldsymbol{n}_{ij}, u_{ij}, u_{ji})$. u_{ij} is an approximation of u at x_{ij} given by a first order Taylor expansion from the point x_i: $u_{ij} = u_i + \nabla^h(u)_i.(x_{ij} - x_i)$. It is well known in the field of MUSCL finite difference schemes that such an approximation leads to unstable schemes unless we introduce a limitation of the derivative $\nabla^h(u)_i$. We propose to start from the value $\nabla^h(u^\alpha)_i = \sum_{j\in P} w_j \left(u_j^\alpha - u_i^\alpha\right) \boldsymbol{A}_{ij}$, and then make a loop over all the neighboring particles such that each component of the gradient is reduced in a way that for all neighboring particles j

$$\begin{cases} u_{ij}^\alpha - u_i^\alpha = \lambda_{ij}^\alpha (u_j^\alpha - u_i^\alpha), \\ \text{with } 0 \le \lambda_{ij}^\alpha \le 1, \end{cases}$$

where the interface values at x_{ij} have been computed according to $u_{ij}^\alpha = u_i^\alpha + \nabla^h(u^\alpha)_i.(x_{ij} - x_i)$. In the scalar case it can be proven that the method is L^∞ stable and convergent with suitable assumptions on the numerical flux.

5.2. Application to Euler equations: Use of Riemann solvers. Consider a regular transport field $a(x,t)$ and the following conservative form of Euler equations in two dimensions (for simplicity):

$$L_a(u) + \sum_{\alpha=1,d} \frac{\partial}{\partial x^\alpha}(F_E^\alpha(u) - a^\alpha u) = 0,$$

where the fluxes F_E^α are given by

$$\text{(i)} \quad F_E^1(u) = \begin{pmatrix} \rho v^1 \\ p + \rho (v^1)^2 \\ \rho v^1 v^2 \\ v^1 (p+E) \end{pmatrix}, \qquad \text{(ii)} \quad F_E^2(u) = \begin{pmatrix} \rho v^2 \\ \rho v^1 v^2 \\ p + \rho (v^2)^2 \\ v^2 (p+E) \end{pmatrix}.$$

We thus have to solve between each particle i and j the Riemann problem

$$\text{(69)} \qquad \begin{cases} \dfrac{\partial}{\partial t}(u) + \dfrac{\partial}{\partial x}((F_E(u).\boldsymbol{n}_{ij} - a(x_{ij},t).\boldsymbol{n}_{ij} u)) = 0, \\ u(x,0) = \begin{cases} u_i & \text{if } x < x_{ij}, \\ u_j & \text{if } x > x_{ij}. \end{cases} \end{cases}$$

A reasonable choice for the flux $g_E(n_{ij}, u_i, u_j)$ of the Godunov scheme associated with our smooth particle approximation is

$$\begin{cases} \lambda_{ij}^a = a(x_{ij},t).\boldsymbol{n}_{ij}, \\ u_{ij}(\lambda_{ij}^a) = u_E(\lambda_{ij}^a; u_i, u_j), \\ G_E(u_i, u_j) = F_E(u_{ij}(\lambda_{ij}^a)) - a(x_{ij},t) \otimes u_{ij}(\lambda_{ij}^a), \\ g_E(\boldsymbol{n}_{ij}, u_i, u_j) = G_E(u_i, u_j).\boldsymbol{n}_{ij}. \end{cases}$$

The resulting particle approximation is given by

$$\frac{d}{dt}(x_i) = a(x_i,t), \qquad \frac{d}{dt}(w_i) = w_i \mathrm{div}(a(x_i,t)),$$
$$\frac{d}{dt}(w_i u_i) + w_i \sum_{j\in P} w_j 2G_E(u_i,u_j)\boldsymbol{A}_{ij} = 0, \qquad u_i(0) = u^0(\xi_i).$$

The detailed equations for mass, momentum, and total energy conservation are

$$\left\{ \begin{array}{l} \dfrac{d}{dt}(w_i\rho_i) + w_i \displaystyle\sum_{j\in P} w_j 2\rho^a_{E,ij}(v^a_{E,ij} - a(x_{ij},t)).\boldsymbol{A}_{ij} = 0, \\ \dfrac{d}{dt}(w_i\rho_i v_i) + w_i \displaystyle\sum_{j\in P} w_j 2\left[\rho^a_{E,ij} v^a_{E,ij} \otimes (v^a_{E,ij} - a(x_{ij},t)) + p^a_{E,ij}\right].\boldsymbol{A}_{ij} = 0, \\ \dfrac{d}{dt}(w_i\rho_i E_i) + w_i \displaystyle\sum_{j\in P} w_j 2 \ \left[E^a_{E,ij}(v^a_{E,ij} - a(x_{ij},t)) + p^a_{E,ij} v^a_{E,ij}\right].\boldsymbol{A}_{ij} = 0, \end{array} \right. \tag{70}$$

where $(\rho^a_{E,ij}, \rho^a_{E,ij} v^a_{E,ij}, E^a_{E,ij})^T = u_{ij}(\lambda^a_{ij})$.

Remark 5.1. Even if we take the transport field equal to the velocity of the particles, $m_i = w_i\rho_i$ does not stay constant as in the renormalized method developed in the previous sections. Although we lose this nice property of the method, we keep global conservation of mass, momentum, and energy. We also expect more robustness since we have better convergence results in the scalar case (see [1, 10]).

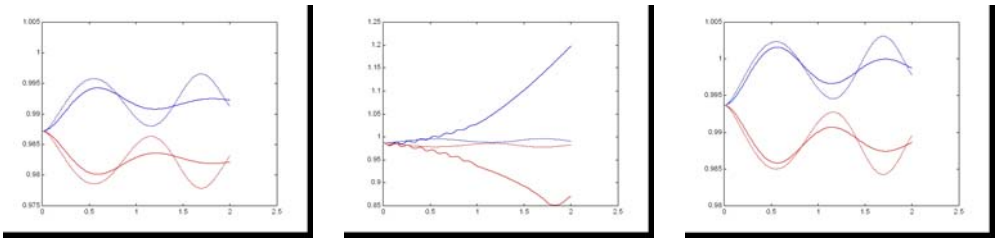
5.3. Applications to free surface flow. We present here some results concerning the simulation of incompressible free surface flows such as those that can arise in sloshing of satellite tanks or launcher tanks. All the computations are performed with an artificial compressibility of the water. Practically we use a linear pressure law, with a sound velocity chosen such that the effective Mach number of the flow is less than 0.1.

5.3.1. A linear regime test case. We first present a test case in two-dimensional plane geometry in the linear regime. We have a two-dimensional tank filled with water, with length $L = 1.0$ m and height $H = 1.0$ m. At time $t = 0$ s the water is at rest with a hydrostatic pressure. We suddenly impose a horizontal acceleration $g = 0.0981$. The characteristic Froude number of the surface waves is $F = 0.001$. In Figure 4 we present successively the computed (full line) and exact position (dotted line) of the water-free surface on the two sides of the tank. In Case 1 we use a 40×40 initial distribution of particles and the hybrid renormalized Godunov–MUSCL (70) scheme, in Case 2 the meshfree weak renormalized scheme (numerical viscosity $\alpha = 0.5, \beta = 1.0$) with 40×40 particles, and in Case 3 the hybrid renormalized Godunov–MUSCL with 80×80 particles.

The weak renormalized scheme with artificial viscosity is not able to reasonably compute this solution, due to the sensitivity of the result to small perturbations (in practice the displacement of any particle is less than 1/100 of the average distance between any particles). It is almost ill conditioned for the SPH Lagrangian-type scheme. Our new method offers the possibility to perform such a computation.

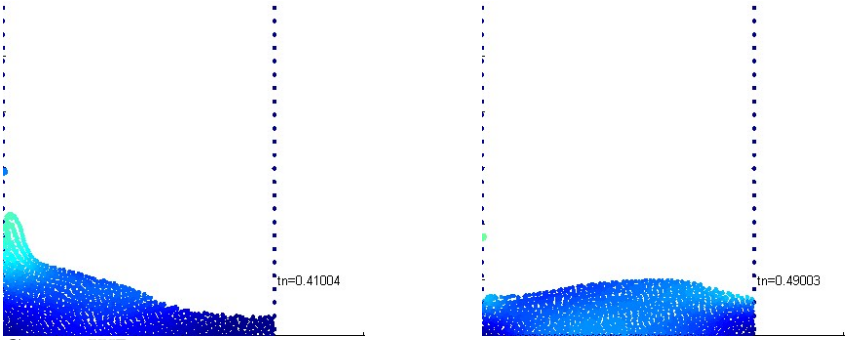
5.3.2. Effect of dissipation in a dam break type problem. We now study the evolution of water in a two-dimensional tank of width $L = 0.15$ m. The water is initially located in a "dam" of height 0.15 m and width 0.05 m. The inner wall is removed at $t = 0$ s.

In Cases 1 and 2 of Figure 5 we use an initial set of 20×60 particles. In Case 3, we use 40×120 particles. In Cases 2 and 3 we use our hybrid renormalized

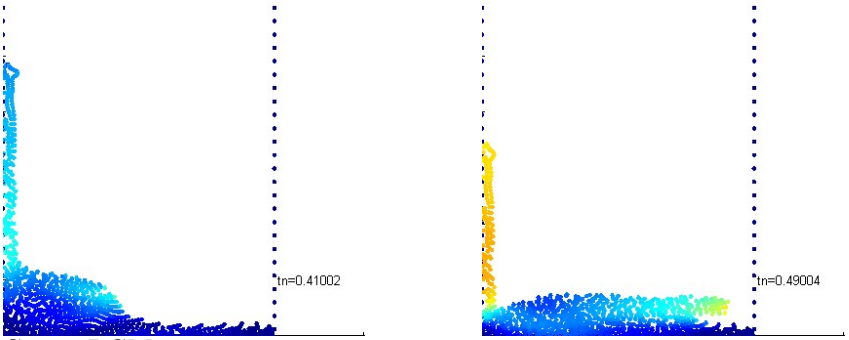


Case 1 Case 2 Case 3

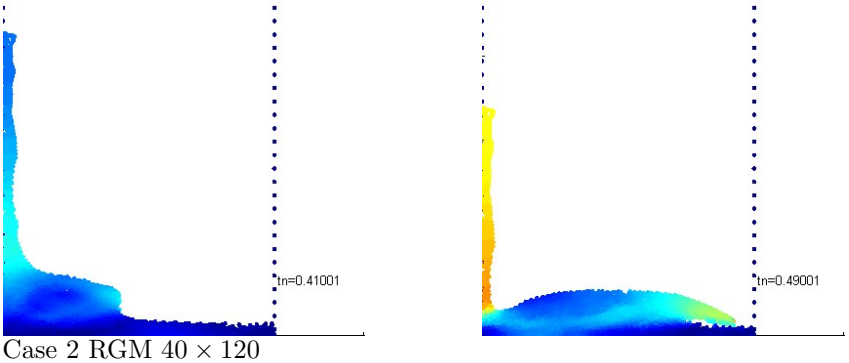
FIG. 4. *Two-dimensional tank of water in linear regime.*



Case 1 WR 20×60



Case 2 RGM 20×60



Case 2 RGM 40×120

FIG. 5. *Dam breaking problem.*

Godunov–MUSCL (RGM) scheme (70), while in Case 1 we use the weak renormalized (WR) scheme (numerical viscosity $\alpha = 0.5, \beta = 1.0$). Clearly the weak renormalized scheme (Case 1) suffers from a great dissipation effect. Comparison of Cases 2 and 3 proves that even with the thicker mesh the renormalized Godunov–MUSCL scheme gives accurate results.

6. Conclusions. We introduced a new class of meshfree derivatives (16) that is more consistent, and we derived the weak renormalized scheme (23) to approach the solution of a general conservation law (1). We established that the time continuous version of this scheme is L^2 stable under the regularity condition on the transport field $a \in (\mathcal{C}^2(\mathbb{R}^d \times \mathbb{R}_+))^d$ and for a regular initial distribution of particles. We also gave L^∞ and L^1 stability results as well as a BV estimate for the solution of the scheme. These results of stability will be essential in the study of convergence (see [10]). The time discretization was performed by analogy with finite volume methods and the concept of E-schemes. The same analogy allowed us to build another type of scheme, based on the renormalization as well, but including more of the Godunov finite volume approach. The hybrid renormalized particle scheme shows great potential, as the numerical results established that it behaves better than the weak renormalized scheme (subsection 5.3), which already gives much better results than classical weighted particle methods, as shown in subsection 2.4.

REFERENCES

[1] S. Benharbit, A. Chalabi, and J. P. Vila, *Numerical viscosity and convergence of finite volume methods for conservation laws with boundary conditions*, SIAM J. Numer. Anal., 32 (1995), pp. 775–796.

[2] B. Ben Moussa, N. Lanson, and J. P. Vila, *Convergence of meshless methods for conservation laws: Applications to Euler equations*, in Hyperbolic Problems: Theory, Numerics, Applications, Vol. I (Zürich, 1998), Internat. Ser. Numer. Math. 129, Birkhäuser, Basel, 1999, pp. 31–40.

[3] B. Ben Moussa and J. P. Vila, *Convergence of SPH method for scalar nonlinear conservation laws*, SIAM J. Numer. Anal., 37 (2000), pp. 863–887.

[4] G. V. Bicknell, *The equations of motion of particles in smoothed particle hydrodynamics*, SIAM J. Sci. Statist. Comput., 12 (1991), pp. 1198–1206.

[5] R. A. Gingold and J. J. Monaghan, *Shock simulation by the particle method SPH*, J. Comput. Phys., 52 (1983), pp. 374–389.

[6] A. Harten and J. M. Hyman, *Self adjusting grid methods for one-dimensional hyperbolic conservation laws*, J. Comput. Phys., 50 (1983), pp. 235–269.

[7] G. R. Johnson and S. R. Beissel, *Normalized smoothing functions for SPH +impact computations* Internat. J. Numer. Methods Engrg., 39 (1996), pp. 2725–2741.

[8] N. Lanson, *Etude des méthodes particulaires renormalisées. Applications aux probèmes de dynamique rapide.*, Ph.D. thesis, INSA-Université Paul Sabatier, Toulouse, France, 2001, www.math.uwaterloo.ca/˜nlanson/thesis1.pdf.

[9] N. Lanson and J. P. Vila, *Convergence of renormalized particle methods for Friedrichs systems*, Les Comptes Rendus de l'Académie des Sciences, 340 (2005), pp. 465–470.

[10] N. Lanson and J. P. Vila, *Renormalized meshfree schemes* II: *Convergence for scalar conservation laws*, SIAM J. Numer. Anal., 46 (2008), pp. 1935–1964.

[11] J. J. Monaghan, *Smooth particle hydrodynamics*, Ann. Rev. Astron. Astrop., 30 (1992), pp. 543–574.

[12] S. Osher, *Riemann solvers, the entropy condition, and difference approximations*, SIAM J. Numer. Anal., 21 (1984), pp. 217–235.

[13] R. W. Randles and L. D. Libertsky, *Smoothed particle hydrodynamics: Some recent improvements and applications*, Comput. Methods Appl. Mech. Engrg., 139 (1996) pp. 375–408.

[14] P. A. Raviart, *An analysis of particle methods*, in Numerical Methods in Fluid Dynamics (Como, 1983), F. Brezzi, ed., Lecture Notes in Math. 1127, Springer-Verlag, Berlin, 1985, pp. 243–324.

[15] E. Tadmor, *Numerical viscosity and the entropy condition for conservative difference schemes*, Math. Comp., 43 (1984), pp. 369–381.
[16] B. Van Leer, *Towards the ultimate conservative difference scheme.* V, J. Comput. Phys., 32 (1979), pp. 101–136.
[17] J. P. Vila, *Weighted particle methods and smooth particle hydrodynamics*, Math. Models Methods Appl. Sci., 9 (1999), pp. 161–209.
[18] J. P. Vila, *Weighted particle-finite volume hybrid schemes*, in Finite Volumes for Complex Applications, F. Benkhaldoun and R. Vilsmeier, eds., Hermes, Paris, 1996.

SIAM J. NUMER. ANAL.
Vol. 46, No. 4, pp. 1935–1964
© 2008 Society for Industrial and Applied Mathematics

RENORMALIZED MESHFREE SCHEMES II: CONVERGENCE FOR SCALAR CONSERVATION LAWS*

NATHALIE LANSON† AND JEAN-PAUL VILA‡

Abstract. We establish a convergence rate of the renormalized meshfree numerical scheme applied to scalar conservation laws. Renormalization is a tool introduced in order to eliminate the smoothed particle hydrodynamics (SPH) lack of consistency. A conservative scheme, the weak renormalized scheme, is derived from the general conservation laws weak formulation and is time discretized by using finite volume techniques. Because of the new form of the derivative approximation, the convergence proof of the classical SPH method cannot be applied. Thus, we use a Kruskov technique to obtain an L^1 norm comparison between the approximate solution and the solution of a regularized conservation law, the pseudoviscous problem.

Key words. meshfree methods, smoothed particle hydrodynamics, renormalization, weak scheme, pseudoviscosity, Kruskov–Kuznetsov entropy method, convergence rate

AMS subject classifications. 65M12, 65M15, 65M99, 35B45

DOI. 10.1137/S003614290444739X

1. Introduction.

We focus on scalar conservation laws:

$$\begin{cases} L_a(u) + \operatorname{div}(F(u,x,t)) = S(u,x,t), \quad t \in \mathbb{R}_+,\ x \in \mathbb{R}^d\ , \\ u(x,0) = u_0(x)\ , \end{cases} \tag{1}$$

where L_a is the transport operator defined as $L_a(u) = \frac{\partial u}{\partial t} + \operatorname{div}(a(x,t)u)$, where a is the transport field $a = (a^\alpha)_{\alpha=1}^d$, $a^\alpha : \mathbb{R}^d \times \mathbb{R}_+ \to \mathbb{R}$, and $u \in \mathbb{R}$ is the exact solution of the problem. $u_0 \in \mathbb{R}$ is the initial condition. The flux $F = (F^\alpha)_{\alpha=1}^d$ and the source term S are $\mathbb{R}$-valued functions: $F : \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}_+ \to \mathbb{R}$ and $S : \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}_+ \to \mathbb{R}$.

The aim of the present paper is to analyze the convergence of the weak renormalized meshfree scheme developed in [10] for the scalar conservation law (1). We consider the following hypothesis on the transport vector field a, the flux F^α, $\alpha = 1, \dots, d$, the source term S, and the initial condition u_0:

$$\begin{aligned} &u_0 \in L^\infty(\mathbb{R}^d) \cap BV_{loc}(\mathbb{R}^d), \qquad a \in (\mathcal{C}^2(\mathbb{R}^d \times \mathbb{R}_+))^d, \\ &F^\alpha \in \mathcal{C}^2(\mathbb{R} \times \mathbb{R}^d \times \mathbb{R}_+), F^\alpha(0,.,.), D_\beta F^\alpha(0,.,.) \in L^1(\mathbb{R}^d \times \mathbb{R}_+), \\ &S \in C^1(\mathbb{R} \times \mathbb{R}^d \times \mathbb{R}_+), S(0,.,.) \in L^1(\mathbb{R}^d \times \mathbb{R}_+). \end{aligned} \tag{2}$$

These regularity conditions are the same as the one required for the stability analysis of the weak renormalized scheme developed in [10].

The existence and uniqueness of the weak entropy solution $u \in L^\infty(\mathbb{R}^d \times \mathbb{R}_+)$ of the problem (1) was proved by Kruszkov in [7]. These solutions satisfy the following

*Received by the editors August 24, 2004; accepted for publication (in revised form) December 6, 2007; published electronically April 18, 2008.
http://www.siam.org/journals/sinum/46-4/44739.html
†Applied Mathematics Department, University of Waterloo, Waterloo, ON N2L 3G1, Canada (nlanson@math.uwaterloo.ca).
‡Département Génie Mathematique et Modélisation, INSA, 31077 Toulouse, France (vila@gmm.insa-tlse.fr).

entropy inequality: $\forall \varphi \geq 0 \in C_c^1(\mathbb{R}^d \times \mathbb{R}_+)$, $\forall c \in \mathbb{R}$

$$
\begin{aligned}
&\int_{\mathbb{R}^d \times \mathbb{R}_+} |u-c| L_a^*(\varphi) + H_c(u,x,t).\nabla_x \varphi dx dt \\
&- \int_{\mathbb{R}^d \times \mathbb{R}_+} sign(u-c)(c\mathrm{div}(a) + \mathrm{div}(F(c,x,t)) - S(u,x,t))\varphi dx dt \\
&+ \int_{\mathbb{R}^d} |u-c|\varphi(x,0) dx \geq 0,
\end{aligned} \tag{3}
$$

where $H_c(u,x,t) = sign(u-c)(F(u,x,t)-F(c,x,t))$ and L_a^* is the transposed operator of L_a and the only physically relevant weak solution. Our aim is to prove that the weak renormalized meshfree scheme developed in [10] converges towards such a solution.

The paper is organized as follows. In section 2, the renormalization principles are briefly recalled, and the time discretized weak renormalized meshfree scheme for conservation laws is given. Some stability results proved in [10] are restated in the end of section 2. In section 3, the classical smoothed particle hydrodynamics (SPH) convergence proof is briefly reviewed, and its inapplicability to the weak renormalized meshfree scheme is explained. In order to prove the convergence, we then introduce a regularized Kruskov entropy function and the pseudoviscous problem associated to the conservation law needed to handle the adjoint operator of the renormalized meshfree derivatives. The organization of the convergence proof for the weak renormalized meshfree numerical scheme is then given. The pseudoviscous problem and its properties are studied in section 4, with an emphasis on building an L^∞ estimate of the second order derivative of the pseudoviscous solution. In section 5, entropy inequalities, using the specific regularized Kruskov entropy, are built for both the solution of the numerical scheme and the solution of the pseudoviscous problem. These two entropy inequalities are compared in section 6. Finally, the main convergence result is given section 7. In order to make the proof clearer, the definitions of the remainders and their estimates are given in Appendices A and B, as well as some quadrature results. The reader can also find a table of symbols at the end of Appendix C.

2. Renormalized particle approximation.

2.1. Basic principles. Consider a set of moving particles P, with coordinates $x_j(t)$, $\forall j \in P$, volumes $\Omega_j(t)$, and weights $w_j(t)$, as explained in [10]. The particle position and weights are moved and are calculated according to:

$$
\begin{aligned}
&\frac{dx_j}{dt} = a(x_j,t), && \text{with } x_j(0) = \xi_j, \\
&\frac{dw_j(t)}{dt} = \mathrm{div}(a(x_j,t))w_j(t), && \text{with } w_j(0) = w_j^0.
\end{aligned} \tag{4}
$$

We consider that the vector field is regular enough such that at any time t: $\mathbb{R}^d = \cup_{j\in P}\Omega_j(t)$, where $\Omega_j(t)$ are distinct volumes of measure $w_j(t)$. In the following the time dependence of x_i and w_i, $\forall i \in P$, will not be explicitly written.

As explained in [10], the *renormalized meshfree derivatives* is introduced to alleviate the lack of consistency of the original particle method. The renormalized meshfree derivatives $D_{\alpha,h}$, $\alpha = 1,\ldots,d$, are given by:

$$
D_{\alpha,h} f(x) = \sum_{j\in P} w_j \left(f(x_j) - f(x)\right)\ A^\alpha(x,x_j), \tag{5}
$$

where the vector is chosen equal to:

$$A(x,x_j)=\frac{B(x)+B(x_j)}{2}.\mu(x,x_j). \tag{6}$$

The regularizing kernel $\mu(x,y)$ has compact support of radius $\varepsilon(x,y)=\frac{\varepsilon(x)+\varepsilon(y)}{2}$, where the smoothing length ε varies in both space and time following $\frac{d\varepsilon}{dt}=\frac{\varepsilon}{d}\mathrm{div}a$ and satisfies $\exists C$ and ε_0 such that, $\forall x,t,\varepsilon_0\leq\varepsilon\leq C\varepsilon_0$. In [10], we saw that, by computing $B_i=B(x_i)$, $\forall i\in P$, as follows:

$$E_i^{\alpha\beta}=\sum_{j\in P}w_j(x_j^\beta-x_i^\beta)\,\mu^\alpha(x_i,x_j),\forall(\alpha,\beta)\in\{1,\dots,d\}^2,\qquad B_i=(E_i)^{-1}, \tag{7}$$

we obtain that the error estimate of the renormalized meshfree derivatives (5) is

$$\forall\varphi\in W^{2,\infty}(\mathbb{R}^d),\quad \|D_h\varphi(x)-D\varphi(x)\|\leq C\varepsilon_0\|B(x)\|\|D^2\varphi\|_\infty, \tag{8}$$

as soon as $B(x)$ is uniformly bounded. We proved in [10] that $B(x)$ is uniformly bounded under the condition that $a\in\mathcal{C}^2(\mathbb{R}_+\times\mathbb{R}^d)$ for a regular initial distribution of particles.

2.2. The weak renormalized meshfree scheme. The construction of the weak renormalized meshfree scheme is given in [10]. It is derived from the weak formulation of (1) together with the use of the approximate derivatives (5). A time discretization is performed in analogy with finite volume methods. The time discretized weak renormalized meshfree scheme:

$$\begin{cases}\tilde{u^h}_i^{n+1}=u_i^{h,n}-\Delta t\left(\mathrm{div}_h^*F_{i,ij}^n+\displaystyle\sum_{j\in P}w_j^nQ_{ij}^n||A_{ij}^n||-\Delta tS_i^n\right),\\[2ex] \dfrac{w_i^{n+1}}{w_i^n}u_i^{h,n+1}=\tilde{u^h}_i^{n+1},\end{cases} \tag{9}$$

where $S_i^n=S(u_i^{h,n},x_i,t^n)$, $Q_{ij}^n=Q(n_{ij},u_i^{h,n},u_j^{h,n})$, is a numerical viscosity, that is assumed to be the numerical viscosity of an E-scheme (see [12, 14]). The numerical viscosity is added in analogy with finite volume methods to stabilize the numerical scheme. The operator $\mathrm{div}_h^*F_{i,ij}^n$ is the transposed operator of the renormalized approximation of the divergence and is defined by:

$$\mathrm{div}_h^*F(u_i^h,x_{ij},t)=\sum_{j\in P}w_j(F(u_i^h,x_{ij},t))+F(u_j^h,x_{ij},t)\dot{A}_{ij}), \tag{10}$$

where $x_{ij}=\frac{x_i+x_j}{2}$. This scheme has a CFL type of stability condition:

$$\sup_{\substack{|u|,\,|v|\,\leq\,C(T,\,||u_0||_\infty)\\ i,\,j\,\in\,P,\ n\Delta t\,\leq\,T}}\frac{\Delta t}{\overline{\Delta x_i}}|Q(n_{ij},u,v)|\leq 1-\beta,\quad 0<\beta<1, \tag{11}$$

where

$$\overline{\Delta x_i}^n=\frac{1}{2\sum_{l\in P}w_l^n||A_{il}^n||} \tag{12}$$

can be viewed as a mean distance separating i from its neighbors. Finally, let us introduce the function $u^h(x,t)$ associated with the scheme (9):

$$\left\{\begin{array}{ll} u^h(x,t) = u^h(x_i^{n+1}, t^{n+1}) & \text{if } t \in]t^n, t^{n+1}],\ x \in \Omega_i^{n+1},\ \text{with } n \geq 1, \\ u^h(x,t) = u^h(x_i^0, 0) & \text{if } t \in [t^0, t^1],\ x \in \Omega_i^0. \end{array}\right. \tag{13}$$

We proved in [10] that this scheme is L_∞, L_1, and BV stable (see Propositions 4.5 and 4.6 of [10]), under the condition that

$$\begin{array}{l} \text{(i) } \sup\limits_{i\in P} \sum\limits_{j\in P} w_j^n ||A_{ij}^n|| \leq \dfrac{C}{\varepsilon_0}, \quad \text{(ii) } \sup\limits_{i\in P} \left\| \sum\limits_{j\in P} w_j^n A_{ij}^n \right\| \leq C, \\ \text{(iii) } \sup\limits_{i\in P} ||A_{ij}^n + A_{ji}^n|| \leq \dfrac{C}{\varepsilon_0^d}. \end{array} \tag{14}$$

It was shown in [10] that the vectors A_{ij}^n satisfy those conditions if $a \in \mathcal{C}^2(\mathbb{R}_+ \times \mathbb{R}^d)$, for a regular initial distribution of particles.

3. First elements on the convergence analysis.

3.1. Methodology and its justification. In the case of classical SPH for nonlinear scalar conservation laws, the convergence proof (see [2]) uses the notion of measure-valued solutions. In the frame of SPH, it is sufficient to prove that, for all convex functions η, the approximate solution of the SPH scheme u^h satisfies (with the definition of (16)):

$$\begin{array}{l} \displaystyle\int_{\mathbb{R}^d\times\mathbb{R}_+} \eta(u^h) L_a^*(\varphi) + H(u^h,x,t)\nabla_x\varphi dxdt \\ \displaystyle + \int_{\mathbb{R}^d\times\mathbb{R}_+} \sum_{\alpha=1}^d \left(D_\alpha H^\alpha(u^h,x,t) - \eta'(u^h) D_\alpha F^\alpha(u^h,x,t) \right) \varphi dxdt \\ \displaystyle + \int_{\mathbb{R}^d\times\mathbb{R}_+} \left(\eta'(u^h) S(u^h,x,t) - (\eta(u^h) - \eta'(u^h)u^h)\mathrm{div}(a) \right) \varphi dxdt \\ \displaystyle + \int_{\mathbb{R}^d} \eta(u_0)\varphi(x,0)dx \geq 0\ . \end{array} \tag{15}$$

Having the adjoint operator of approximation of the divergence tending strongly towards the exact divergence $\mathrm{div}_x F$ is crucial within the convergence proof. Unfortunately, our transposed operator (10) does not satisfy this requirement, and we cannot apply the proof developed in [2] to our scheme. We will use a Kruskov–Kuznetsov-like method, similar to the work in [15, 3, 4, 5]. However, another difficulty arises: The term $sign(u^h - c)\mathrm{div}_{h,x}^* F^\alpha(c,x,t)$, when using Kruskov entropy, appears in the entropy inequality of the approximate solution. Since our renormalized meshfree derivative adjoint operator $\mathrm{div}_{h,x}^*$ does not tend strongly toward div_x, the term $sign(u^h - c)\mathrm{div}_{h,x}^* F^\alpha(c,x,t)$ can only weakly tend to $sign(u^h - c)\mathrm{div}_x F^\alpha(c,x,t)$. Consequently, one needs some regularity on $sign(u^h - c)$. The first step is to regularize the sign function (see Definition 3.1). Second, the approximate solution u^h is not regular enough, and neither is the exact solution u ($W^{2,\infty}$ is required). We introduce the pseudoviscous problem (section 4), and we will show that the solution of the pseudoviscous problem u^ϵ has enough regularity to make possible the handling of such a term as $sign(u^\epsilon - c)\mathrm{div}_{h,x}^* F^\alpha(c,x,t)$. Thus, we build an entropy inequality for the pseudoviscous solution u^ϵ (26), that we can compare to the entropy inequality of the approximate solution u^h (24).

3.2. Entropy and regularized Kruskov entropy functions.

(16) Let H be the flux associated to the entropy η, such that $\partial_u H = \eta'(u)\partial_u F$,

where F is the flux of the conservation law (1) with $p = 1$.

The numerical entropy flux h_e satisfying $h_e(n, u, u) = H(u, x, t).n$ is defined by:

$$h_e(n_{ij}, u_i^h, u_j^h) = \frac{(H(u_i^h, x_{ij}, t) + H(u_j^h, x_{ij}, t)).n_{ij}}{2} - Q^\eta(n_{ij}, u_i^h, u_j^h). \tag{17}$$

As explained earlier, we take a $\mathcal{C}^2$ regularization of the Kruskov entropy.

DEFINITION 3.1. *The regularized function* $(\eta_\delta)'(u) = sign_\delta(u - c)$ *is defined by:*

$$sign_\delta(u-c) = \begin{cases} -1 & \text{if } u - c \le -\delta, \\ \frac{15}{8\delta}(u-c) - \frac{5}{4\delta^3}(u-c)^3 + \frac{3}{8\delta^5}(u-c)^5 & \text{if } u - c \in]-\delta;\delta[, \\ 1 & \text{if } u - c \ge \delta, \end{cases} \tag{18}$$

and the associated flux entropy H *is given by:*

$$\begin{aligned} H_\delta(u,x,t) &= \int_c^u sign_\delta(v-c)\partial_v F(v,x,t)dv = H_\delta^c(u,x,t) + \Psi_\delta^c(u,x,t), \\ H_\delta^c(u,x,t) &= sign_\delta(u-c)(F(u,x,t)-F(c,x,t)), \\ \Psi_\delta^c(u,x,t) &= \int_c^u (sign_\delta(u-c)-sign_\delta(v-c))\partial_v F(v,x,t)dv. \end{aligned} \tag{19}$$

These definitions satisfy the following results.

LEMMA 3.2. *Definitions* (18) *and* (19) *imply that*

$$\big||u|^\delta - |u|\big| \le C\delta, \tag{20}$$

$$||\Psi_\delta^c(u,x,t)|| \le C_\Psi(u)\delta, \quad |\mathrm{div}_x \Psi_\delta^c(u,x,t)| \le C'_\Psi(u)\delta, \tag{21}$$

where $C_\Psi(u) = \sup_{v\in[c,u]} |\partial_u F(v,x,t)|$ *and* $C'_\Psi(u) = \sup_{v\in[c,u]} |\mathrm{div}_x \partial_u F(v,x,t)|$,

$$\begin{aligned} &sign_\delta(u-c)F(u,x,t) - H_\delta^c(u,x,t) = sign_\delta(u-c)F(c,x,t), \\ &sign_\delta(u-c)\mathrm{div}_x F(u,x,t) - \mathrm{div}_x H_\delta^c(u,x,t) = sign_\delta(u-c)\mathrm{div}_x F(c,x,t), \end{aligned} \tag{22}$$

and the following inequalities: $\forall$*compact* $Q \in \mathbb{R}^d \times \mathbb{R}_+$

$$\begin{aligned} &||H_\delta(u,.,.)||_{L^\infty(Q)} \le ||u-c||_\infty ||sign_\delta||_\infty ||\partial_u F(u,.,.)||_{L^\infty(Q)}, \\ &||\mathrm{div}_x H_\delta(u,.,.)||_{L^\infty(Q)} \le ||u-c||_\infty ||sign_\delta||_\infty ||\mathrm{div}_x \partial_u F(u,.,.)||_{L^\infty(Q)}, \\ &||H_\delta^c(u,.,.)||_{L^\infty(Q)} \le 2||sign_\delta||_\infty ||F(u,.,.)||_{L^\infty(Q)}, \\ &||\mathrm{div}_x H_\delta^c(u,.,.)||_{L^\infty(Q)} \le 2||sign_\delta||_\infty ||\mathrm{div}_x F(u,.,.)||_{L^\infty(Q)}, \\ &||\partial_t H_\delta^c(u,.,.)||_{L^\infty(Q)} \le 2||sign_\delta||_\infty ||\partial_t F(u,.,.)||_{L^\infty(Q)}. \end{aligned} \tag{23}$$

Proof. The only nontrivial part is to prove the estimates (21). The proof is given in [9, page 161]. ∎

3.3. Organization of the convergence proof.

1. We prove in section 5.1 that the solution u^h of the scheme (9), (13) satisfies:

(24)
$$\underbrace{\int_{\mathbb{R}_+\times\mathbb{R}^d} |u^h(x,t)-c|^\delta L_a^*(\varphi)dxdt}_{\text{(a)}} + \underbrace{\int_{\mathbb{R}_+\times\mathbb{R}^d} H_\delta^c(u^h(x,t),x,t).\nabla\varphi dxdt}_{\text{(b)}}$$
$$\underbrace{-\int_{\mathbb{R}_+\times\mathbb{R}^d} (u^h(x,t)sign_\delta(u^h(x,t)-c)-|u^h(x,t)-c|^\delta)\mathrm{div}a(x,t)\varphi dxdt}_{\text{(c)}}$$
$$\underbrace{+\int_{\mathbb{R}_+\times\mathbb{R}^d} sign_\delta(u^h(x,t)-c)S(u^h(x,t),x,t)\varphi dxdt}_{\text{(d)}} + \underbrace{\int_{\mathbb{R}^d} |u^h(x,0)-c|^\delta\varphi dx}_{\text{(e)}}$$
$$\underbrace{-\int_{\mathbb{R}_+\times\mathbb{R}^d} sign_\delta(u^h(x,t)-c)\overline{\mathrm{div}_h^*}F(c,x,t)\varphi dxdt}_{\text{(f)}} \geq \langle\mu_h,\varphi\rangle,$$

(25) where $\mu_h \to 0$ and
$$\overline{\mathrm{div}_h^*}F(c,x,t) = \mathrm{div}_h^*F(c,x_i,t^n) \text{ for } x\in\Omega_i^n, t\in]t^n,t^{n+1}].$$

2. The term $\mathrm{div}_{h,x}^*(F(c,x,t))$ tends only weakly towards $\mathrm{div}_x(F(c,x,t))$. Thus, we have to transform the entropy equation satisfied by the exact solution of problem (1) in order to get a similar term $\mathrm{div}_{h,x}^*F$. Then another problem arises: Considering the assumptions (2) on F, a, S, and u_0, our exact solution is not regular enough to handle this transformation. Indeed, the error estimate on the renormalized meshfree derivatives implies that u should be $W^{2,\infty}$. Thus, we introduce the *pseudoviscous solution* u^ϵ in section 4, give its properties, and deduce the L^∞ estimate of its second order derivatives.
3. We prove in section 5.2 that the pseudoviscous solution u^ϵ satisfies:

(26)
$$\underbrace{\int_{\mathbb{R}_+\times\mathbb{R}^d} |u^\epsilon(y,s)-c|^\delta L_a^*(\varphi)dyds}_{\text{(a')}} + \underbrace{\int_{\mathbb{R}_+\times\mathbb{R}^d} H_\delta^c(u^\epsilon(y,s),y,s).\nabla\varphi dyds}_{\text{(b')}}$$
$$\underbrace{-\int_{\mathbb{R}_+\times\mathbb{R}^d} (u^\epsilon(y,s)sign_\delta(u^\epsilon(y,s)-c)-|u^\epsilon(y,s)-c|^\delta)\mathrm{div}a(y,s)\varphi dyds}_{\text{(c')}}$$
$$\underbrace{+\int_{\mathbb{R}_+\times\mathbb{R}^d} sign_\delta(u^\epsilon(y,s)-c)S(u^\epsilon(y,s),y,s)\varphi dyds}_{\text{(d')}} + \underbrace{\int_{\mathbb{R}^d} \eta_\delta(u_0^\epsilon-c)\varphi dy}_{\text{(e')}}$$
$$\underbrace{-\int_{\mathbb{R}_+\times\mathbb{R}^d} \varphi sign_\delta(u^\epsilon(y,s)-c)\overline{\mathrm{div}_h^*}F(c,y,s)dyds}_{\text{(f')}} \geq \langle\nu_\epsilon,\varphi\rangle,$$

where ν_ϵ is a measure term tending to zero.
4. In section 6 we sum the inequality (24), integrated over (s,y) with $c=u(y,s)$, and the inequality (26), integrated over (t,x) with $c=u^h(x,t)$. The chosen

test function $\varphi = \varphi(x, y, s, t)$ (Definition 6.1) together with the error estimate given in Appendix A leads to the L^1 error estimate between u^ϵ and u^h. By using the convergence properties of u^ϵ, we get the L^1 convergence of the approximate solution u^h towards the exact solution u (Theorem 7.1).

4. Pseudoviscous solution and its properties. The pseudoviscous problem is given by (with $\epsilon > 0$):

$$\left\{ \begin{array}{l} \dfrac{\partial u^\epsilon}{\partial s} + \mathrm{div} F(u^\epsilon, y, s) = \epsilon \Delta u^\epsilon + S(u^\epsilon, y, s), \\ u^\epsilon(0, .) = u^{\epsilon,0}, \end{array} \right. \tag{27}$$

where $u^{\epsilon,0} = u_0 * \zeta_\epsilon$; ζ_ϵ is a regularizing kernel. Note that the flux F includes the transport term a. We use this convention to simplify the notations in this section.

4.1. Estimates on u^ϵ. *Note: In the following $TV(v)$ represents the total variation of a function v.*

This problem has been studied in [6] for a space- and time-dependent flux and in [11] for a flux and a source term depending on u, x, and t. Lemmas 4.13 and 2.9 and Theorem 2.29 of [13] are summed up below.

LEMMA 4.1. *If $u_0 \in L^1(\mathbb{R}^d) \cap L^\infty(\mathbb{R}^d) \cap BV(\mathbb{R}^d)$, then the solution of (27) satisfies: $\exists C$ and $C(Q)$ (Q compact subset of $\mathbb{R}^d \times \mathbb{R}_+$) independent of ϵ such that*

$$||u^\epsilon(., s)||_{L^\infty(\mathbb{R}^d)} \leq ||u^{0,\epsilon}||_{L^\infty(\mathbb{R}^d)},\ ||u^\epsilon(., s)||_{L^1(\mathbb{R}^d)} \leq ||u^{0,\epsilon}||_{L^1(\mathbb{R}^d)} + C\ TV(u_0),$$
$$||\nabla u^\epsilon(., s)||_{L^1(\mathbb{R}^d)} \leq TV(u_0),\ ||\partial_t u^\epsilon(., s)||_{L^1(\mathbb{R}^d)} \leq C\ TV(u_0),\ ||\nabla u^\epsilon||_{L^2(Q)} \leq \frac{C(Q)}{(\epsilon)^{\frac{1}{2}}}.$$

By using the ideas developed in [13, 11], one can also prove the following.

LEMMA 4.2. *Under the assumption that $u_0 \in L^1(\mathbb{R}^d) \cup L^\infty(\mathbb{R}^d) \cup BV(\mathbb{R}^d)$, there exists a constant $C'(Q)$ (independ of ϵ) such that u^ϵ satisfies:*

$$||D^2 u^\epsilon||_{L^\infty(Q)} \leq \frac{C'(Q)}{\epsilon^{\frac{5}{2}}}. \tag{28}$$

Proof. In the following, $u^\varepsilon_\alpha = \frac{\partial u^\varepsilon}{\partial x^\alpha}$. First differentiate the parabolic equation (27) with respect to x^α, and then integrate the resulting equation by parts over space, with the test function $\varphi = u^\epsilon_\alpha \chi^2$, where χ is a function with compact support $Q \subset \mathbb{R}^+ \times \mathbb{R}^d$ and such that $\chi = 1\ \forall (x, t) \in Q$ and $\chi = 0$ if distance$((t, x), Q) > 1$. Since $(u^\epsilon_\alpha \chi)^2(T)$ is positive, we have:

$$\int_Q |\epsilon^{\frac{1}{2}} \nabla (u^\epsilon_\alpha \chi)|^2 dx dt \leq \sum_{k=1}^{8} I_k, \text{ with}$$

$$I_1 = \left| \int_Q \frac{\partial}{\partial u^2} F^\beta(u^\epsilon, x, t) u^\epsilon_\beta (u^\epsilon_\alpha)^2 \chi^2 dx dt \right|, \qquad I_4 = \left| \int_Q (u^\epsilon_\alpha)^2 \chi \frac{\partial \chi}{\partial t} dx dt \right|,$$

$$I_2 = \left| \int_Q \frac{\partial}{\partial u} F^\beta(u^\epsilon, x, t) u^\epsilon_{\beta\alpha} u^\epsilon_\alpha \chi^2 dx dt \right|, \qquad I_5 = \epsilon \left| \int_Q (u^\epsilon_\alpha)^2 |\nabla \chi|^2 dx dt \right|,$$

$$I_3 = \left| \int_Q \frac{\partial^2}{\partial x^\beta \partial x^\alpha} F^\beta(u^\epsilon, x, t) u^\epsilon_\alpha \chi^2 dx dt \right|, \qquad I_6 = \left| \int_Q \frac{\partial}{\partial x^\alpha} S(u^\epsilon, x, t) u^\epsilon_\alpha \chi^2 dx dt \right|,$$

$$I_7 = \left| \int_Q \frac{\partial}{\partial u} S(u^\epsilon, x, t) (u^\epsilon_\alpha)^2 \chi^2 dx dt \right|, \qquad I_8 = \int_{\mathbb{R}^d} (u^\epsilon_\alpha \chi)^2(0) dx.$$

Integrals I_k, $k=3,\dots,7$, are bounded from the hypothesis (2) on the flux F and the source term S and from the estimates of Lemma 4.1:

$$
\begin{aligned}
&I_3 \leq \sup_{v;|v|\leq C(u_0)} \left\{ \frac{\partial^2}{\partial x^\beta \partial x^\alpha} F^\beta(u^\epsilon,.,.)||_{L^\infty(Q)} \right\} TV(u_0),\\
&I_4 \leq \frac{C(Q)}{\sqrt{\epsilon}}, \quad I_5 \leq \frac{C(Q)}{\sqrt{\epsilon}},\\
&I_6 \leq \sup_{v;|v|\leq C(u_0)} \left\{ \left\| \frac{\partial}{\partial x^\alpha} S(v,.,.) \right\|_{L^\infty(Q)} \right\} TV(u_0),\\
&I_7 \leq \sup_{v;|v|\leq C(u_0)} \{||\partial_v S(v,.,.)||_{L^\infty(Q)}\} \frac{C(Q)}{\sqrt{\epsilon}}.
\end{aligned}
\tag{29}
$$

The second integral is bounded by using the Cauchy–Schwartz inequality:

$$
I_2 \leq \sup_{v;|v|\leq C(u_0)} \{||\partial_v F(v,.,.)||_{L^\infty(Q)}\} ||u^\epsilon_{\alpha\beta}||_{L^2(Q)} ||u^\epsilon_\alpha||_{L^2(Q)}. \tag{30}
$$

The first integral is estimated by using an integration by parts and estimates (4.1):

(31)

$$
\begin{aligned}
I_1 &\leq \sup_{v;|v|\leq C(u_0)} \{||\partial_u^2 F(v,.,.)||_{L^\infty(Q)}\} \int_Q |u^\epsilon_\beta| (u^\epsilon_\alpha)^2 \chi^2 dxdt\\
&\leq C(F,u_0) \sup_t ||u^\epsilon(.,t)||_{L^\infty(\mathbb{R}^d)} \left(||u^\epsilon_{\beta\alpha}||_{L^2(Q)} ||u^\epsilon_\alpha||_{L^2(Q)} + ||u^\epsilon_\beta||_{L^2(Q)} ||u^\epsilon_{\alpha\alpha}||_{L^2(Q)} \right.\\
&\qquad \left. +2||u^\epsilon_\beta||_{L^2(Q)} ||u^\epsilon_\alpha||_{L^2(Q)} \right),\\
&\leq C(F,S,u_0,Q) \left(\frac{||u^\epsilon_{x_j x_i}||_{L^2(Q)}}{\sqrt{\epsilon}} + \frac{||u^\epsilon_{x_i x_i}||_{L^2(Q)}}{\sqrt{\epsilon}} + \frac{1}{\epsilon} \right).
\end{aligned}
$$

Finally, integral I_8 is bounded by using Lemma 4.6 of [13], which proves that $||u^\epsilon(0,.)||_{L^\infty(\mathbb{R}^d)} \leq ||u_0||_{L^\infty(\mathbb{R}^d)}$ and $||\Delta u^\epsilon(0,.)||_{L^1(\mathbb{R}^d)} \leq \frac{C}{\varepsilon} TV(u_0)$. Consequently, we have:

$$
\begin{aligned}
I_8 &= \int_{\mathbb{R}^d} (u^\epsilon_\alpha(0,x))^2 \chi(0,x)^2 dx \leq \int_{\mathbb{R}^d} (u^\epsilon_\alpha(0,x))^2 dx\\
&\leq \left| \int_{\mathbb{R}^d} u^\epsilon(0,x) u^\epsilon_{\alpha\alpha}(0,x) dx \right| \leq \frac{C}{\epsilon} TV(u_0) ||u_0||_{L^\infty(\mathbb{R}^d)} \leq \frac{C(u_0)}{\epsilon}.
\end{aligned}
$$

The upper bounds of the integrals I_k, $k=1,\dots,8$, lead to:

$$
\begin{aligned}
&\epsilon ||D^2 u^\epsilon||^2_{L^2(Q)} \leq C(F,S,u_0,Q) \left(1 + \frac{1}{\epsilon} + \frac{1}{\sqrt{\epsilon}} + \frac{1}{\epsilon} + \frac{1}{\sqrt{\epsilon}} ||D^2 u^\epsilon||_{L^2(Q)} \right)\\
&\Rightarrow ||D^2 u^\epsilon||_{L^2(Q)} \leq \frac{C(F,S,u_0,Q)}{\epsilon^{\frac{3}{2}}}.
\end{aligned}
\tag{32}
$$

$$
\text{Let us use the Sobolev injection } ||D^2 u^\epsilon||_{L^\infty(\Omega)} \leq C ||D^2 u^\epsilon||_{H^1(\Omega)} \ \forall \Omega. \tag{33}
$$

We now have to estimate the L^2 norm of $D^3 u^\epsilon$. This is achieved by rederiving the proof above, starting by differentiating (27) a second time with respect to x_γ, and using the test function $\varphi = u_{x_\gamma x_\alpha} \chi^2$. We then prove that

$$
||D^3 u^\epsilon||_{L^2(Q)} \leq \frac{C(F,S,u_0,Q)}{\epsilon^{\frac{5}{2}}}. \tag{34}
$$

One deduces (28) from the above inequality, together with the Sobolev injection (33) (we keep only the dependence in Q in the constant C' for simplicity). ∎

4.2. Convergence of the pseudoviscous solution. We recall here a classical result, established by Kuznetsov [8].

THEOREM 4.3 (Kuznetsov). *$\exists C > 0$ such that, if $u_0 \in BV(\mathbb{R}^d)$ and if $u(0) = u^\epsilon(0) = u_0$, then*

$$||u^\epsilon(.,s) - u(.,s)||_1 \leq C\sqrt{\epsilon s}\ TV(u_0). \tag{35}$$

5. Kruzkov entropy inequalities for approximate and pseudoviscous solutions.

5.1. Entropy inequality for the approximate solution.

PROPOSITION 5.1. *Let us suppose that assumptions* (2) *are satisfied, and let u^h be the solution of* (9), (13). *Then u^h verifies the continuous entropy inequality* (24).

Proof.

- *Construction of a first inequality for a general entropy η*: The following result is proved in [14] and [1]: If g is the flux of an E-scheme consistent with (1) (with $p = 1$), and if the CFL is satisfied: $\lambda Q(n,u,v) \leq 1$, then there exists a convex entropy flux η such that:

$$\begin{aligned}\eta(u - \lambda(g(n,u,v) - F(u,x,t).n)) - \eta(u) + \lambda[h_e(n,u,v)\\ -H(u,x,t).n] \leq 0.\end{aligned} \tag{36}$$

η is a convex function, and thus $\eta(u+v) \leq \eta(u) + v\eta'(u+v)$; consequently:

$$\begin{aligned}\eta(\tilde{u^h}_i^{n+1}) \leq \eta\left(\sum_{j\in P} \alpha_{ij}^n u^{h,n+1,j}\right) - \Delta t\eta'(\tilde{u^h}_i^{n+1}) \sum_{j\in P} w_j^n (F_{i,ij}^n - F_i^n).A_{ij}^n\\ + \Delta t\eta'(\tilde{u^h}_i^{n+1})S_i^n,\end{aligned}$$

where $u_i^{h,n+1,j}$ is defined section 4.1 of [10], $\alpha_{ij}^n = \frac{w_j^n||A_{ij}^n||}{\sum_{l\in P} w_l^n||A_{il}^n||}$, $F_{i,ij}^n = F(u_i^{h,n}, x_{ij}, t^n)$, $F_i^n = F(u_i^{h,n}, x_i, t^n)$, and $S_i^n = S(u_i^{h,n}, x_i, t^n)$. By using Jensen's inequality (since $\sum_{j\in P} \alpha_{ij}^n = 1$), and applying (36) to $u_i^{h,n+1,j}$, with $\lambda = \frac{\Delta t}{\overline{\Delta x_i^n}}$ ($\overline{\Delta x_i^n}$ defined in (12)), we get:

$$\begin{aligned}\eta(\tilde{u^h}_i^{n+1}) + 2\Delta t \sum_{j\in P} w_j^n (h_{e,ij}^n - H_{i,ij}^n.n_{ij})||A_{ij}^n|| \leq \eta(u_i^{h,n})\\ - \eta'(\tilde{u^h}_i^{n+1})\Delta t \sum_{j\in P} w_j^n (F_{i,ij}^n - F_i^n).A_{ij}^n + \Delta t\eta'(\tilde{u^h}_i^{n+1})S_i^n,\end{aligned}$$

where $H_{i,ij}^n = H(u_i^{h,n}, x_{ij}, t^n)$ and $h_{e,ij}^n = h_e(n_{ij}, u_i^{h,n}, u_j^{h,n})$. After using the definition of the numerical flux h_e (17), we multiply the above inequality by a test function $(-\varphi)$, with $\varphi \geq 0$, and sum over all particles and time steps. By transposing the adjoint derivatives approximate operator on H to φ, one finally gets, by using the notation $(\Phi, \varphi)_h^n$ for the l^2 scalar product

$\sum_{i\in P} w_i^n \Phi_i^n \varphi_i^n$:

$$
\begin{aligned}
&- \sum_{n\geq 0} \Delta t \left\{ \overbrace{\left(\frac{\eta(\tilde{u^h}^{n+1}) - \eta(u^{h,n})}{\Delta t}, \varphi^n \right)_h^n}^{I_1} - (H^n, D\varphi^n)_h^n \right\} \\
&\geq -2\Delta t \sum_i w_i^n \varphi_i^n \sum_{j\in P} w_j^n (H_{i,ij}^n - \eta'(\tilde{u^h}_i^{n+1}) F_{i,ij}^n).A_{ij}^n \\
&- \sum_{n\geq 0} \Delta t \left(\eta'(\tilde{u^h}^{n+1}) S^n, \varphi^n \right)_h^n + R_{11}^h(\varphi) + R_{12}^h(\varphi) + R_{13}^h(\varphi),
\end{aligned} \tag{37}
$$

where $R_{11}^h(\varphi)$, $R_{12}^h(\varphi)$, and $R_{13}^h(\varphi)$ are defined in (72a).
In the following, we will use the results below, deduced from (4) and (9):

$$
\begin{aligned}
&w_i^{n+1} = w_i^n + \Delta t \mathrm{div} a_i^n w_i^n, \\
&\eta(\tilde{u^h}_i^{n+1}) \leq \eta(u_i^{h,n+1}) + u_i^{h,n+1} \Delta t \mathrm{div} a_i^n \eta'(\tilde{u^h}_i^{n+1}), \\
&\eta(u_i^{h,n+1}) \leq \eta(\tilde{u^h}_i^{n+1}) + \eta'(u_i^{h,n+1}) \left(u_i^{h,n+1} - \tilde{u^h}_i^{n+1} \right).
\end{aligned} \tag{38}
$$

Suppose that φ has a compact support in $\mathbb{R}_+ \times \mathbb{R}^d$ and $\varphi(T,.) = 0$. By dividing the sum and making a discrete integration by parts, one obtains:

$$
\begin{aligned}
I_1 = &\overbrace{- \sum_{n\geq 0} \sum_{i\in P} (w_i^n \eta(\tilde{u^h}_i^{n+1}) - w_i^{n+1} \eta(u_i^{h,n+1})) \varphi_i^n}^{I_{11}} \\
&+ \sum_{n\geq 0} \sum_{i\in P} w_i^{n+1} \eta(u_i^{h,n+1}) (\varphi_i^{n+1} - \varphi_i^n) + \sum_{i\in P} w_i \eta(u_i^{h,0}) \varphi_i^0.
\end{aligned} \tag{39}
$$

From (38), we have:

$$
\begin{aligned}
I_{11} &\leq \sum_{n\geq 0} \sum_{i\in P} \left((w_i^{n+1} - w_i^n) \eta(u_i^{h,n+1}) - \eta'(u_i^{h,n+1}) u_i^{h,n+1} (w_i^{n+1} - w_i^n) \right) \varphi_i^n \\
&\leq \sum_{n\geq 0} \Delta t \sum_{i\in P} w_i^n \left(\eta(u_i^{h,n+1}) - \eta'(u_i^{h,n+1}) u_i^{h,n+1} \right) \mathrm{div} a_i^n \varphi_i^n.
\end{aligned} \tag{40}
$$

By using (40) and (39) in (37), one gets $\forall \eta, H$, and $\forall \varphi$ regular with compact support:

$$
\begin{aligned}
&\sum_{n\geq 0} \Delta t \left(\eta(u^{h,n+1}), \frac{(\varphi^{n+1} - \varphi^n)}{\Delta t} \right)_h^{n+1} + \sum_{n\geq 0} \Delta t (H(u^{h,n}, ., t^n), \nabla \varphi^n)_h^n \\
&+ \sum_{n\geq 0} \Delta t ((\eta(u^{h,n+1}) - \eta'(u^{h,n+1}) u^{h,n+1}) \mathrm{div} a^n, \varphi^n)_h^n + \left(\eta(u^{h,0}), \varphi^0 \right)_h^0 \\
&- 2\Delta t \sum_i \varphi_i^n \sum_{j\in P} w_j^n (H_{i,ij}^n - \eta'(\tilde{u^h}_i^{n+1}) F_{i,ij}^n).A_{ij}^n \\
&+ \sum_{n\geq 0} \Delta t (\eta'(\tilde{u^h}^{n+1}) S(u^{h,n}, ., t^n), \varphi^n)_h^n \geq R_1^h(\varphi),
\end{aligned} \tag{41}
$$

where R_1^h is defined in (72a), with $H = H_\delta$ and $\eta = \eta_\delta$.

- *Application of Kruszkov entropy*: Let us replace the entropy η by the regularized Kruszkov entropy $|.|^\delta$ and the entropy flux H by $H^c_\delta + \Psi^c_\delta$ (see Definition 3.1) in (41). We then obtain R^h_2 (72b) (linked to the introduction of Ψ^c_δ). By using equality (22), one gets:

$$\begin{aligned} &- 2\Delta t \sum_i \varphi^n_i \sum_{j\in P} w^n_j (H^{c,n}_{\delta,i,ij} - sign_\delta(u^{h,n+1}_i) F^n_{i,ij}).A^n_{ij} \\ &= 2\sum_n \Delta t \sum_i w^n_i sign_\delta(u^{h,n+1}_i)\varphi^n_i \sum_j w^n_j F(c, x_{ij}, t^n) A^n_{ij}, \end{aligned}$$

that we transform into $\mathrm{div}^{h,*}_x F(c, x_i, t^n)$ using $R^h_3(\varphi)$ defined in (72c).
In order to fit with the definition of u_h (13), we change all dependencies in $u^{h,n}$ or $\tilde{u^h}^{n+1}$ into $u^{h,n+1}$, and, with the definition (72d) of $R^h_4(\varphi)$, we get:
(42)

$$\begin{aligned} &\sum_{n\geq 0} \Delta t \left(|u^{h,n+1} - c|^\delta, \frac{(\varphi^{n+1} - \varphi^n)}{\Delta t} \right)^{n+1}_h + \sum_{n\geq 0} \Delta t (H^c_\delta(u^{h,n+1}, ., t^n), \nabla\varphi^n)^n_h \\ &+ \sum_{n\geq 0} \Delta t ((|u^{h,n+1} - c|^\delta - sign_\delta(u^{h,n+1} - c) u^{h,n+1}) \mathrm{div} a^n, \varphi^n)^n_h \\ &- \sum_{n\geq 0} \Delta t (sign_\delta(u^{h,n+1} - c) \mathrm{div}^{h,*}_x F(c, ., t^n), \varphi^n)^n_h + \left(|u^{h,0} - c|^\delta, \varphi^0\right)^0_h \\ &+ \sum_{n\geq 0} \Delta t \left(sign_\delta(u^{h,n+1} - c) S(u^{h,n+1}, ., t^n), \varphi^n\right)^n_h \geq \sum_{k=1}^4 R^h_k(\varphi). \end{aligned}$$

- *Continuous entropy inequality for the scheme*: We finally obtain (24) by introducing $R^h_5(\varphi)$ (see (72e)), representing the transition between discrete sums and continuous integrals with $\langle \mu_h, \varphi\rangle = \sum_{k=1}^5 R^h_k(\varphi)$. ∎

5.2. Entropy inequality for pseudoviscous solution.

PROPOSITION 5.2. *Let us suppose that assumptions* (2) *are satisfied, and let* u^ϵ *be the solution of* (27). *Then* u^ϵ *verifies the continuous entropy inequality* (26).

Proof. Let η be a convex function and H the associated entropy flux. We multiply (27) by $\eta'(u^\epsilon)$ and integrate over t and x. The term resulting from the perturbation is then transformed as follows:

$$\begin{aligned} -\epsilon \int_{\mathbb{R}_+\times\mathbb{R}^d} \eta'(u^\epsilon)\Delta u^\epsilon \varphi dyds &= -\epsilon \int_{\mathbb{R}_+\times\mathbb{R}^d} \nabla(\eta'(u^\epsilon)\nabla u^\epsilon)\varphi dxdt \\ &\quad + \epsilon \int_{\mathbb{R}_+\times\mathbb{R}^d} \eta''(u^\epsilon)(\nabla u^\epsilon)^2 \varphi dyds. \end{aligned}$$

Since η is a convex function, $\eta''(u^\epsilon) \geq 0$, the second term on the right-hand side of the above equality is positive, and we deduce the entropy inequality for u^ϵ:

$$\begin{aligned} &\int_{\mathbb{R}_+\times\mathbb{R}^d} \eta(u^\epsilon(y,s)) \frac{d\varphi}{dt} - (u^\epsilon(y,s)\eta'(u^\epsilon(y,s)) - \eta(u^\epsilon(y,s))) \mathrm{div} a(y,s)\varphi \\ &\qquad + H(u^\epsilon(y,s), y, s).\nabla\varphi - \eta'(u^\epsilon(y,s)) \mathrm{div}_y F(u^\epsilon(y,s), y, s)\varphi \\ &\qquad + \eta'(u^\epsilon(y,s)) S(u^\epsilon(y,s), y, s)\varphi - \epsilon\eta'(u^\epsilon(y,s))\nabla u^\epsilon(y,s).\nabla\varphi dyds \\ &+ \int_{\mathbb{R}^d} \eta(u^\epsilon_0)\varphi dy \geq 0. \end{aligned} \tag{43}$$

We apply $\eta(u) = |u-c|^\delta$ (see Definition 3.1) in (43) to get $R_1^\epsilon(\varphi)$ defined in (94a). By considering $H = H_\delta^c + \Psi_\delta^c$, we introduce the remainder $R_2^\epsilon(\varphi)$ defined in (94b). It remains to transform the term including the adjoint approximate operator of div$F$ into a term including the derivative of $\varphi sign_\delta(u^\epsilon)$. This is achieved by considering $R_3^\epsilon(\varphi)$ defined in (94c). We finally get (26) with $\langle \nu_\epsilon, \varphi \rangle = \sum_{k=1}^3 R_k^\epsilon(\varphi)$. ∎

6. Comparison of the two entropy inequalities. We present the test function that allows us to compare inequalities (24) and (26) and establish the comparison result that will lead to the proof of the main result.

6.1. Definition and properties of the test function. The test function depends on x, y, s, and t and must be with compact support with respect to x and t. Our aim is to have (s,y) tending to (t,x). Thus we use some approximations of the delta function with smoothing length γ.

DEFINITION 6.1. *The test function φ is equal to:*

$$\varphi(x,y,t,s) = \psi(t,x)\zeta^\gamma(t-s)\bar{\zeta}^\gamma(x-y), \tag{44}$$

where

$$\zeta^\gamma(x-y) = \frac{1}{\gamma^d}\zeta\left(\frac{||x-y||}{\gamma}\right) \quad \bar{\zeta}^\gamma(t-s) = \frac{1}{\gamma}\bar{\zeta}\left(\frac{|t-s|}{\gamma}\right), \tag{45}$$

$$\psi(t,x) = \begin{cases} e^{C_\psi t}\rho(|x|^\delta + \omega t)\frac{T-t}{T} & \text{if } x \in \mathbb{R}^d, t \in [0,T], \\ 0 & \text{if } x \in \mathbb{R}^d, t \geq T. \end{cases} \tag{46}$$

The positive functions $\zeta \in C_c^\infty(\mathbb{R}^d)$ and $\bar{\zeta} \in C_c^\infty(\mathbb{R})$ are with compact support such that Supp$(\bar{\zeta}) \subset [-1,0]$ and Supp$(\zeta) \subset [-1,1]$, and

$$\int_{\mathbb{R}^d} \zeta(||x||)dx = 1, \quad \int_{\mathbb{R}_+} \bar{\zeta}(t)dt = 1. \tag{47}$$

The function ρ in (46) *has the following properties:*

$$\begin{cases} \rho(r) = 1 \text{ if } r \in [0,R], \\ \rho(r) = 0 \text{ if } r \in [R+1, +\infty[, \\ \rho'(r) \leq 0 \ \forall r \in \mathbb{R}_+. \end{cases} \tag{48}$$

We first remark that $\varphi(x,y,t,0) = 0$, since $\zeta^\gamma(t-0) = 0$ for all $t \geq 0$, and that ρ goes to zero smoothly. Also, the derivatives of functions ζ^γ and $\bar{\zeta}^\gamma$ satisfy $\nabla_x \zeta^\gamma(x-y) = -\nabla_y \zeta^\gamma(x-y)$ and $\partial_t \bar{\zeta}^\gamma(t-s) = -\partial_s \bar{\zeta}^\gamma(t-s)$. This definition of φ ensures that φ has a compact support: $(x,t) \in Q$ and $(y,s) \in \tilde{Q}$, where Q and $\tilde{Q}$ are defined by

$$Q = \Omega \times [0,T] \text{ and } \tilde{Q} = \tilde{\Omega} \times [0, T+\gamma], \ \tilde{\Omega} = \text{extension of } \Omega \text{ by a width } \gamma. \tag{49}$$

The upper bounds of the L^∞ norm of some derivatives of ψ and φ will be needed later to estimate the remainders. They can be found in [9, page 190, equation (6.134) for φ and page 172, equation (6.89) for ψ].

6.2. Comparison of the entropy inequalities. We take $c = u(y,s)$ (resp., $c = u^h(x,t)$) in the entropy inequality for u^h (24) (resp., u^ϵ (26)) and integrate over

y and s (resp., x and t). By summing the two resulting inequalities and considering the definition of H_c^δ (19), (24)+(26) becomes:

$$\begin{aligned}
(\mathrm{a})+(\mathrm{a}') &= \int_{Q\times\tilde{Q}} |u^h(x,t)-u^\epsilon(y,s)|^\delta \psi_t(t,x)\zeta^\gamma(x-y)\zeta^\gamma(t-s)dxdydtds \\
&\quad + \int_{Q\times\tilde{Q}} |u^h(x,t)-u^\epsilon(y,s)|^\delta (a(x,t)-a(y,s)) \\
&\qquad \cdot \psi(t,x)\nabla_x\zeta^\gamma(x-y)\bar{\zeta}^\gamma(t-s)dxdydtds \qquad (50)\\
&\quad + \int_{Q\times\tilde{Q}} |u^h(x,t)-u^\epsilon(y,s)|^\delta a(x,t) \\
&\qquad \cdot \nabla_x\psi(t,x)\zeta^\gamma(x-y)\bar{\zeta}^\gamma(t-s)dxdydtds \\
&= T_1+T_2+T_3, \\
(\mathrm{b})+(\mathrm{b}') &= \int_{Q\times\tilde{Q}} sign_\delta(u^h(x,t)-u^\epsilon(y,s))\left(F(u^h(x,t),x,t)-F(u^\epsilon(y,s),x,t)\right) \\
&\quad \cdot \nabla_x\psi(t,x)\zeta^\gamma(x-y)\zeta^\gamma(t-s)dxdydtds + \int_{Q\times\tilde{Q}} sign_\delta(u^h(x,t) \\
&\quad - u^\epsilon(y,s))\left((F(u^h(x,t),x,t)-F(u^h(x,t),y,s))+(F(u^\epsilon(y,s),y,s)\right. \qquad (51)\\
&\quad \left. - F(u^\epsilon(y,s),x,t))\right).\psi(t,x)\nabla_x\zeta^\gamma(x-y)\bar{\zeta}^\gamma(t-s)dxdydsdt \\
&= T_4+T_5, \\
(\mathrm{c})+(\mathrm{c}') &= \int_{Q\times\tilde{Q}} |u^h(x,t)-u^\epsilon(y,s)|^\delta(\mathrm{div}a(y,s)+\mathrm{div}a(x,t)) \\
&\qquad \cdot \psi(t,x)\zeta^\gamma(x-y)\zeta^\gamma(t-s)dxdydtds \\
&\quad - \int_{Q\times\tilde{Q}} sign_\delta(u^h(x,t)-u^\epsilon(y,s))\left(u^h(x,t)\mathrm{div}a(x,t)\right. \qquad (52)\\
&\qquad \left. -u^\epsilon(y,s)\mathrm{div}a(y,s)\right)\psi(t,x)\zeta^\gamma(x-y)\zeta^\gamma(t-s)dxdydtds \\
&= T_6-T_7, \\
(\mathrm{f})+(\mathrm{f}') &= -\int_{Q\times\tilde{Q}} sign_\delta(u^h(x,t)-u^\epsilon(y,s))(\overline{\mathrm{div}_x^{h,*}}F(u^\epsilon(y,s),x,t) \qquad (53)\\
&\qquad -\overline{\mathrm{div}_y^{h,*}}F(u^h(x,t),y,s))\varphi(x,y,t,s)dxdydtds = -T_8, \\
(\mathrm{d})+(\mathrm{d}') &= -\int_{Q\times\tilde{Q}} sign_\delta(u^h(x,t)-u^\epsilon(y,s))(S(u^h(x,t),x,t) \qquad (54)\\
&\qquad -S(u^\epsilon(y,s),y,s))\varphi(x,y,t,s)dxdydtds = T_9,
\end{aligned}$$

and

$$(\mathrm{e})+(\mathrm{e}') = \int_{\Omega\times\tilde{Q}} |u^h(x,0)-u^\epsilon(y,s)|^\delta\psi(x,0)\zeta^\gamma(x-y)\zeta^\gamma(s)dxdyds = T_{10}. \qquad (55)$$

By considering that $\varphi(x,y,t,0)=0$, we get:

$$\sum_{i=1}^{10} T_i \geq \int_{(\mathbb{R}^+\times\mathbb{R}^d)} \langle\mu_h,\varphi(.,y,.,s)\rangle dyds + \int_{(\mathbb{R}^+\times\mathbb{R}^d)} \langle\nu_\epsilon,\varphi(x,.,t,.)\rangle dxdt. \qquad (56)$$

Inequality (56) is the basis to prove the following lemma.

LEMMA 6.2. *Let u^h be the solution of the scheme* (9), (13) *and u^ϵ the solution of the pseudoviscous equation* (27)*, and then the following inequality holds:*

(57)
$$\int_{(\mathbb{R}_+\times\mathbb{R}^d)^2} \big(|u^h(x,t)-u^\epsilon(y,s)|\partial_t\psi(t,x) + sign_\delta(u^h(x,t)-u^\epsilon(y,s))\left(F(u^h(x,t),x,t)-F(u^\epsilon(y,s),x,t)\right).\nabla_x\psi(t,x)\big) \cdot\zeta^\gamma(x-y)\zeta^\gamma(t-s)dxdydtds + \int_{\Omega\times\tilde{Q}} |u^h(x,0)-u^\epsilon(y,s)|\psi(0,x)\zeta^\gamma(x-y)\zeta^\gamma(s)dxdyds \geq T_{L^1}(\varphi)+\langle\lambda_1,\varphi\rangle + \int_{\tilde{Q}}\langle\mu_h,\varphi(.,y,.,s)\rangle dyds + \int_Q\langle\nu_\epsilon,\varphi(x,.,t,.)\rangle dxdt,$$

where μ_h and ν_ϵ are, respectively, defined in (72) *and* (94)*, and $\langle\lambda_1,\varphi\rangle=\sum_{k=1}^8 R_k^c(\varphi)$, where $R_k^c(\varphi), k=1,\dots,8$, are given in* (104)*. The term $T_{L^1}(\varphi)$ represents all elements which can be bounded by $||u^h-u^\epsilon||_L^1(\Omega\times[0,T])$ and is defined in* (118)*.*

Proof. Each term T_k, $k=1,\dots,10$, from inequality (56) not including derivatives of ψ is transformed in order to get terms either tending to 0 or upper bounded by $||u^h-u^\epsilon||_{L^1}$. The fourth term T_4 is kept unchanged. We use the characteristic function χ such that χ_V is equal to 1 on V and 0 elsewhere, where V is a subspace of $\mathbb{R}$ or $\mathbb{R}^d$. The subspaces Q and $\tilde{Q}$ were defined in (49).

- *Term T_1*: The transformation of T_1 concerns only the regularization of $|.|$

$$T_1 = \int_{Q\times\tilde{Q}} |u^h(x,t)-u^\epsilon(y,s)|\psi_t(t,x)\zeta^\gamma(x-y)\bar{\zeta}^\gamma(t-s)dxdydsdt + R_1^c(\varphi), \tag{58}$$

  where $R_1^c(\varphi)$ is defined in (104a).
- *Term T_2*: By using the regularity (2) on a, the inequality $|v|^\delta\leq|v|$ (from Definition (3.1)), and the compact supports of ζ and $\bar{\zeta}$ (see (45)), one gets:

$$|T_2|\leq(||Da||_{L^\infty(\tilde{Q})}+||\partial_t a||_{L^\infty(\tilde{Q})})||\nabla\zeta^\gamma||_\infty||\bar{\zeta}^\gamma||_\infty\gamma \int_{Q\times\tilde{Q}} |u^h(x,t)-u^\epsilon(y,s)|\ |\psi(t,x)|\chi_{[-\gamma,0]}(t-s)\chi_{B(0,\gamma)}(x-y)dydsdxdt.$$

  By considering

$$\int_{\tilde{Q}}\chi_{[-\gamma,0]}(t-s)\chi_{B(0,\gamma)}(x-y)dyds=\gamma^{d+1}, \tag{59}$$

  let us transform the integral in the following way:

$$\int_{Q\times\tilde{Q}} |u^h(x,t)-u^\epsilon(y,s)|\ |\psi(t,x)|\chi_{[-\gamma,0]}(t-s)\chi_{B(0,\gamma)}(x-y)dydsdxdt = \gamma^{d+1}\int_Q|u^h(x,t)-u^\epsilon(x,t)|\ |\psi(t,x)|dxdt + \int_{Q\times\tilde{Q}}|u^\epsilon(x,t)-u^\epsilon(y,s)|\ |\psi(t,x)|\chi_{[-\gamma,0]}(t-s)\chi_{B(0,\gamma)}(x-y)dydsdxdt.$$

By using the definition of $R^{c,L^1}(\psi)$ given in (104k), we deduce that

$$|T_2| \leq C(u_0,T,\Omega,F,S)\left(\int_Q |u^h(x,t)-u^\epsilon(x,t)|\ |\psi(t,x)|dxdt + R^{c,L^1}(\psi)\right). \tag{60}$$

- *Term* T_3: We proceed as for the term T_1: With $R_2^c(\psi)$ defined in (104b),

$$T_3 = \int_{Q\times\tilde{Q}} |u^h(x,t)-u^\epsilon(y,s)|a(x,t).\nabla\psi(t,x)\zeta^\gamma(x-y)\bar{\zeta}^\gamma(t-s)dxdydsdt + R_2^c(\varphi).$$

- *Term* T_5: By using hypothesis (2) on F, one can use Taylor's formula on $F(u^h(x,t),a,b) - F(u^\epsilon(y,s),a,b)$ for $(a,b)=(x,t)$ or (y,s). We get:

$$\begin{aligned} T_5 = \int_{Q\times\tilde{Q}}\int_0^1 sign_\delta(u^h(x,t)-u^\epsilon(y,s))\left(\partial_1 F(u^\epsilon(y,s)+\theta(u^h(x,t)\right. \\ - u^\epsilon(y,s)),x,t) - \partial_1 F(u^\epsilon(y,s)+\theta(u^h(x,t) \\ \left. - u^\epsilon(y,s)),y,s)\right)(u^\epsilon(y,s)-u^h(x,t)) \\ \cdot \psi(t,x)\nabla_x\zeta^\gamma(x-y)\bar{\zeta}^\gamma(t-s)dxdydtdsd\theta. \end{aligned}$$

Let us bound T_5 as follows:

$$\begin{aligned} |T_5| \leq ||sign_\delta||_{L^\infty(\mathbb{R})}\int_{Q\times\tilde{Q}} \{\sup_{c\in\mathbb{R}} |\partial_1 F(c,x,t)-\partial_1 F(c,y,s)|\} \\ \cdot |u^\epsilon(y,s)-u^h(x,t)|\ |\psi(t,x)|\ |\nabla_x\zeta^\gamma(x-y)|\bar{\zeta}^\gamma(t-s)dxdydtds. \end{aligned}$$

By using the regularity hypothesis on F and proceeding as for T_2, there exists $C(u_0,F,S,a,T)$ such that:

$$|T_5| \leq C(u_0,F,S,a,T)\left(\int_Q |u^\epsilon(x,t)-u^h(x,t)|\ |\psi(t,x)|dxdt + R^{c,L^1}(\psi)\right), \tag{61}$$

with $R^{c,L^1}(\psi)$ defined in (104k).

- *Term* T_6: We write $T_6 = T_6' + R_3^c(\varphi) + R_4^c(\varphi)$, where $R_3^c(\varphi)$, $R_4^c(\varphi)$, and T_6' are defined in (104c), (104d), and (118c), respectively, and T_6' satisfies (as for the term T_2):

$$|T_6'| \leq C(u_0,F,S,a,T)\left(\int_Q |u^\epsilon(x,t)-u^h(x,t)|\ |\psi(t,x)|dxdt + R^{c,L^1}(\psi)\right). \tag{62}$$

- *Term* T_7: We write $T_7 = T_7' + R_5^c(\Psi)$, where $R_5^c(\varphi)$ and T_7' are defined in (104e) and (118d), respectively, and T_7' satisfies (as for the term T_2):

$$|T_7'| \leq C(u_0,F,S,a,T)\left(\int_Q |u^\epsilon(x,t)-u^h(x,t)|\ |\psi(t,x)|dxdt + R^{c,L^1}(\psi)\right). \tag{63}$$

- *Term T_8*: We write $T_8 = T_8' + R_6^c$, where $R_6^c(\varphi)$ and T_8' are defined in (104f) and (118e), respectively.
  Concerning T_8', let us use the regularity of F with its first variable:

$$\underline{T_8'} \leq \left(\sup_{v,|v|\leq C(u_0)} ||\partial_1 F(v,.,.)||_{L^\infty(\tilde{Q})} + \sup_{v,|v|\leq C(u_0)} ||\nabla \partial_1 F(v,.,.)||_{L^\infty(\tilde{Q})} \right)$$
$$\cdot ||sign_\delta'||_\infty \int_{Q\times\tilde{Q}} |\varphi(x,y,t,s)|\ |u^\epsilon(y,s) - u^h(x,t)| dxdyds dt \tag{64}$$
$$\leq C(u_0,F,S,a,T,\Omega) \left(\int_Q |u^\epsilon(x,t) - u^h(x,t)| \psi(t,x) dxdt + R^{c,L^1}(\psi) \right).$$

- *Term T_9*: We set $T_9 = T_9' + R_7^c(\varphi)$, where $R_7^c(\varphi)$ and T_9' are defined in (104g) and (118f), respectively. From the hypothesis (2) on the source term S, one prove that

$$|T_9'| \leq ||sign_\delta||_\infty \sup_{c;||c||\leq C} \{ ||\partial_u S(c,.,.)||_{L^\infty(\tilde{(}Q))} \} \tag{65}$$
$$\cdot \int_{Q\times\tilde{Q}} |u^\epsilon(y,s) - u^h(x,t)| \psi(t,x) \zeta^\gamma(x-y) \bar{\zeta}^\gamma(t-s) dxdydtds$$
$$\leq C(u_0,F,S,a,T) \left(\textstyle\int_Q |u^\epsilon(x,t) - u^h(x,t)|\ |\psi(t,x)| dxdt + R^{c,L^1}(\psi) \right).$$

- *Term T_{10}*: We set $T_{10} = T_{10}' + R_8^c(\varphi)$, where $R_8^c(\varphi)$ is defined in (104h), and $T_{10}' = \int_{\tilde{Q}\times\Omega} |u^h(x,0) - u^\epsilon(x,0)| \zeta^\gamma(x-y) \bar{\zeta}^\gamma(s) dyds \psi(x,0) dx$ and satisfies:

$$|T_{10}'| \leq C(u_0,T,\Omega,F,a,S) \int_\Omega |u^h(x,0) - u^\epsilon(x,0)| \psi(x,0) dx. \tag{66}$$

- *Conclusion*: From the above transformations, we obtain inequality (57), with $\langle \lambda_1, \varphi \rangle = \sum_{k=1}^8 R_k^c(\varphi)$, and T_{L^1} defined in (118), bounded by $||u^h - u^\epsilon||_{L^1}$: $\exists C_{L^1}(u_0,F,S,a,T,\Omega)$ such that

$$\boxed{|T_{L^1}(\varphi)| \leq C_{L^1} \left(\int_Q |u^h(x,t) - u^\epsilon(x,t)|\ |\psi(x,t)| dxdt + R^{c,L^1}(\psi) \right).} \tag{67}$$
□

6.3. Gronwall formula. From Lemma 6.2, we deduce the following proposition.

PROPOSITION 6.3. *Let u^h be the solution of the scheme* (9), (13) *and u^ϵ the solution of problem* (27) *with S, F, a, and u_0 satisfying* (2). *Then, with the definition* (44) *and* (49) *for $\varphi(x,y,t,s)$, Q, and $\tilde{Q}$, we have for all compact E in $\mathbb{R}^d \times \mathbb{R}_+$:*

$$\int_E |u^h(x,t) - u^\epsilon(x,t)| dxdt \leq \int_\Omega |u^h(x,0) - u^\epsilon(x,0)| dx + T \int_{\tilde{Q}} \langle \mu_h, \varphi(.,y,.,s) \rangle dyds$$
$$+ T \int_Q \langle \nu_\epsilon, \varphi(x,.,t,.) \rangle dxdt + T \langle \lambda, \varphi \rangle, \tag{68}$$

where μ_h, ν_ϵ, and λ are defined in (72), (94), *and* (104), *respectively.*

Proof. By using the definition of ψ (46), the left-hand side of (57) (called I) reads:

$$\begin{aligned}
I = &\int_{Q\times\tilde{Q}} |u^h(x,t)-u^\epsilon(y,s)| C_\psi e^{C_\psi t}\rho(|x|^\delta+\omega t)\frac{T-t}{T}\zeta^\gamma(x-y)\bar{\zeta}^\gamma(t-s)dxdydsdt\\
&-\int_{Q\times\tilde{Q}} |u^h(x,t)-u^\epsilon(y,s)| e^{C_\psi t}\rho(|x|^\delta+\omega t)\frac{1}{T}\zeta^\gamma(x-y)\bar{\zeta}^\gamma(t-s)dxdydsdt\\
&+\int_{Q\times\tilde{Q}} |u^h(x,t)-u^\epsilon(y,s)| e^{C_\psi t}\omega\rho'(|x|^\delta+\omega t)\frac{T-t}{T}\zeta^\gamma(x-y)\bar{\zeta}^\gamma(t-s)dxdydsdt\\
&+\int_{Q\times\tilde{Q}} sign_\delta(u^h(x,t)-u^\epsilon(y,s))\left(F(u^h(x,t),x,t)-F(u^\epsilon(y,s),x,t)\right)\\
&\cdot e^{C_\psi t}\vec{sign}_\delta(x)\rho'(|x|^\delta+\omega t)\frac{T-t}{T}\zeta^\gamma(x-y)\bar{\zeta}^\gamma(t-s)dxdydsdt.
\end{aligned}$$

From hypothesis (2), we have $|F(\tilde{u},x,t)-F(u,x,t)|\le D_F^1|\tilde{u}-u|$. By taking $\omega=D_F^1$ and using $\bar{\zeta}\ge 0$, $\zeta\ge 0$, $\rho'\le 0$, and $||sign_\delta||_\infty\le 1$, we get:

$$\begin{aligned}
&\int_{Q\times\tilde{Q}} e^{C_\psi t}\rho'(|x|^\delta+D_F^1 t)\frac{T-t}{T}\zeta^\gamma(x-y)\zeta^\gamma(t-s)\left(D_F^1|u^h(x,t)-u^\epsilon(y,s)|\right.\\
&\left.+\,sign(u^h(x,t)-u^\epsilon(y,s))\left(F(u^h(x,t),x,t)-F(u^\epsilon(y,s),x,t)\right).\vec{sign}_\delta(x)\right)dxdydsdt\le 0.
\end{aligned}$$

From the upper bound of $T_{L^1}(\varphi)$ (67), one gets:

$$\begin{aligned}
&\int_Q |u^h(x,t)-u^\epsilon(x,t)| e^{C_\psi t}\rho(|x|^\delta+D_F^1 t)\frac{1}{T}dxdt\\
&\le \int_Q |u^h(x,t)-u^\epsilon(x,t)| C_\psi e^{C_\psi t}\rho(|x|^\delta+\omega t)\frac{T-t}{T}dxdt\\
&+\int_\Omega |u^h(x,0)-u^\epsilon(x,0)|\rho(|x|^\delta)dx + C_{L^1}\int_Q |u^h(x,t)-u^\epsilon(x,t)| e^{C_\psi t}\rho(|x|^\delta+D_F^1 t)\\
&\quad\cdot\frac{T-t}{T}\Big|dxdt+\langle\lambda,\varphi\rangle+\int_{\tilde{Q}}\langle\mu_h,\varphi(.,y,.,s)\rangle dyds\\
&\quad+\int_Q\langle\nu_\epsilon,\varphi(x,.,t,.)\rangle dxdt,
\end{aligned}$$

where we take $\langle\lambda,\varphi\rangle=\langle\lambda_1,\varphi\rangle+R_9^c(\varphi)+R_{10}^c(\varphi)+R^{c,L^1}(\psi)$, R_9^c and R_{10}^c being defined in (104i) and (104j), respectively.

Take $C_\psi=-C_{L^1}(u_0,F,S,a,T,\Omega)$. It remains that

(69)
$$\begin{aligned}
\int_Q |u^h(x,t)-u^\epsilon(x,t)| e^{C_\psi t}\rho(|x|^\delta+\omega t)\frac{1}{T}dxdt &\le \int_\Omega |u^h(x,0)-u^\epsilon(x,0)|\ \rho(|x|^\delta)\ dx\\
&+\langle\lambda,\varphi\rangle+\int_{\tilde{Q}}\langle\mu_h,\varphi(.,y,.,s)\rangle dyds+\int_Q\langle\nu,\varphi(x,.,t,.)\rangle dxdt.
\end{aligned}$$

The integration domain with respect to t being $0\le t\le T$ and the constant C_ψ being negative, we have $e^{C_\psi T}\ge e^{C_\psi t}$. Moreover, if we consider any subspace $E\in\cup_{0\le t\le T}(B(0,R-\omega t)\times\{t\})$, the function ρ is equal to 1 when $(x,t)\in E$, and we get:

$$\int_E |u^h(x,t)-u^\epsilon(x,t)|dxdt\le\frac{1}{e^{(C_\psi T)}}\int_E |u^h(x,t)-u^\epsilon(x,t)| e^{C_\psi t}\rho(|x|^\delta+\omega t)dxdt. \tag{70}$$

By combining this equation with (69), one gets the result of Proposition 6.3. ∎

7. Main result.

THEOREM 7.1. *Let us assume that a, F, S, and u_0 satisfy the hypothesis* (2)*, that the time step satisfies the CFL condition* (11)*, and that the ratio $\frac{\varepsilon}{h} = \mathcal{O}(1)$. Suppose that the shape vectors A_{ij}^n verify the conditions* (14) *and the Lipschitz condition:*

$$\left| \sum_{j \in P} w_j^n A_{ij}^n - \sum_{j \in P} w_j^m A_{kj}^m \right| \leq C(|t^m - t^n| + |\xi_i - \xi_k| + O(\varepsilon_0)).$$

Let us denote by u the entropic solution of (1) *satisfying* (3) *and u^h the solution of the numerical scheme* (9), (13)*. Then u^h converges towards u in $L^1_{loc}(\mathbb{R}^d \times \mathbb{R}_+)$, and we have the following estimate: For all compact E of $\mathbb{R}^d \times \mathbb{R}_+$, there exists a constant $\mathcal{C}$ depending only on E, S, F, u_0, and a such that:*

$$\int_E |u^h(x,t) - u(x,t)| dx dt \leq \mathcal{C} h^{\frac{1}{20}}. \tag{71}$$

Proof. From the result of Proposition 6.3 and the estimates (93), (103), (117) given in Appendix A of μ_h, ν_ϵ, and λ, respectively, one gets:

$$\begin{aligned}
&\int_E |u^h(x,t) - u^\epsilon(x,t)| dx dt \leq \int_{\mathbb{R}^d} |u^h(x,0) - u^\epsilon(x,0)| \rho(|x|^\delta) \frac{1}{T} dx \\
&+ \mathcal{C} \left(h + \sqrt{h} + \delta + \frac{\sqrt{h}}{\gamma} + \frac{h}{\delta} + \frac{h}{\gamma} + \frac{h}{\gamma^2} + \frac{\delta}{\gamma} + \gamma + \frac{h}{\delta \gamma \epsilon^{\frac{1}{2}}} + \frac{h}{\delta \epsilon^{\frac{5}{2}}} + \frac{h}{\epsilon^3 \delta^2} + \frac{\epsilon}{\gamma} \right),
\end{aligned}$$

where $\mathcal{C} = \mathcal{C}(u_0, F, S, a, T, \Omega)$. The integral at $t = 0$ is bounded by:

$$\begin{aligned}
\int_{\mathbb{R}^d} |u^h(x,0) - u^\epsilon(x,0)| \rho(|x|^\delta) \frac{1}{T} dx &\leq C(T) \int_{\mathbb{R}^d} |u^h(x,0) - u^\epsilon(x,0)| dx \\
&\leq C(T) \left(\sum_{i \in P} \omega_i |u^h(x_i,0) - u^\epsilon(x_i,0)| + \sum_{i \in P} \int_{\Omega_i} |u^\epsilon(x_i,0) - u^\epsilon(x,0)| dx \right) \\
&\leq C(T, u_0)(h + \epsilon),
\end{aligned}$$

so that we have:

$$\begin{aligned}
\int_E |u^h(x,t) - u^\epsilon(x,t)| dx dt \leq \mathcal{C} \Bigg(& h + \sqrt{h} + \epsilon + \gamma + \delta + \frac{\sqrt{h}}{\gamma} + \frac{h}{\delta} + \frac{h}{\gamma} + \frac{h}{\gamma^2} + \frac{\delta}{\gamma} \\
& + \frac{h}{\delta \gamma \epsilon^{\frac{1}{2}}} + \frac{h}{\delta \epsilon^{\frac{5}{2}}} + \frac{h}{\epsilon^3 \delta^2} + \frac{\epsilon}{\gamma} \Bigg).
\end{aligned}$$

The optimal values of ϵ, γ, and δ are $\epsilon = h^{\frac{1}{10}}$, $\gamma = h^{\frac{1}{20}}$, and $\delta = h^{\frac{1}{4}}$, respectively. By combining this with the convergence result (Theorem 4.3) of u^ϵ towards the exact solution of (1) u, one gets the final result (71) and the proof of convergence. ∎

Because of the problems related to the definition of the renormalized derivative operator, as explained in section 5, we used the Kruskov–Kuznetsov entropy method in order to prove the convergence of the weak renormalized meshfree scheme. The advantage of this method is that an error estimate is obtained, and the proof can be applied to problems with irregular initial conditions, as the initial condition can always be regularized in order to be BV. However, the rate of convergence of $\frac{1}{20}$ is not the right convergence rate as it is much less than what we observe in real computations.

Appendix A. Remainders and their estimates. In this appendix section, we give the definition of all of the estimates presented within the article and an upper bound of their absolute value. We recall that $\frac{\Delta t}{h} = O(1)$ and $\frac{\varepsilon}{h} = O(1)$ (see the hypothesis of Theorem 7.1).

A.1. Measure μ_h. We note that $||D\varphi|| = ||D\varphi(.,y,.,s)||$ for any operator D and any norm.

$$\langle \mu_h, \varphi \rangle = R_1^h(\varphi) + R_2^h(\varphi) + R_3^h(\varphi) + R_4^h(\varphi) + R_5^h(\varphi), \tag{72}$$

$$R_1^h(\varphi) = R_{11}^h(\varphi) + R_{12}^h(\varphi) + R_{13}^h(\varphi), \tag{72a}$$

$$R_{11}^h(\varphi) = \sum_{n\geq 0} \Delta t \sum_{i,j\in P} w_i^n \varphi_i^n w_j^n Q_{ij}^{\eta_\delta,n} ||A_{ij}^n||, \tag{72a(i))}$$

$$R_{12}^h(\varphi) = \sum_{n\geq 0} \Delta t \sum_{i,j\in P} w_i^n \varphi_i^n w_j^n (H_{\delta,j}^n - H_{\delta,i,ij}^n + H_{\delta,i}^n - H_{\delta,j,ij}^n).A_{ij}^n, \tag{72a(ii))}$$

$$R_{13}^h(\varphi) = \sum_{n\geq 0} \Delta t \, (H_\delta, \nabla\varphi - \nabla_h\varphi)_h^n \,, \tag{72a(iii))}$$

$$R_2^h(\varphi) = -\sum_{n;i,j\in P} \Delta t w_i^n \Psi_{\delta,i}^{c,n}.w_j^n(\varphi_j^n - \varphi_i^n)A_{ij}^n + 2\sum_{n;i,j\in P} \Delta t w_i^n \varphi_i^n w_j^n \Psi_{\delta,i,ij}^{c,n}.A_{ij}^n, \tag{72b}$$

$$\begin{aligned} R_3^h(\varphi) = \sum_{n;i,j\in P} \Delta t w_i^n sign_\delta(u_i^{h,n+1})\varphi_i^n w_j^n (F(c,x_i,t^n) - F(c,x_{ij},t^n) \\ + F(c,x_j,t^n) - F(c,x_{ij},t^n)).A_{ij}^n, \end{aligned} \tag{72c}$$

$$R_4^h(\varphi) = R_{41}^h(\varphi) + R_{42}^h(\varphi), \tag{72d}$$

(72d(i))

$$\begin{aligned} R_{41}^h(\varphi) &= \sum_{n\geq 0} \Delta t \left((H_\delta(u^{h,n+1},.,t^n) - H_\delta(u^{h,n},.,t^n)), \nabla_h(\varphi^n) \right)_h^n \\ &+ \sum_{n\geq 0} \Delta t \left(sign_\delta(\tilde{u^h}^{n+1})(\mathrm{div}_{h,x}^* F(u^{h,n+1},.,t^n) - \mathrm{div}_{h,x}^* F(u^{h,n},.,t^n)), \varphi^n \right)_h^n \\ &- \sum_{n\geq 0} \Delta t \left(sign_\delta(\tilde{u^h}^{n+1})(S(u^{h,n+1},.,t^n) - S(u^{h,n},.,t^n)), \varphi^n \right)_h^n \\ &- \sum_{n\geq 0} \Delta t \left((\mathrm{div}_{h,x}^* H_\delta(u^{h,n+1},.,t^n) - \mathrm{div}_{h,x}^* H_\delta(u^{h,n},.,t^n)), \varphi^n \right)_h^n, \end{aligned}$$

$$\begin{aligned} R_{42}^h(\varphi) = \sum_{n\geq 0} \Delta t \Big((sign_\delta(u^{h,n+1}) - sign_\delta(\tilde{u^h}^{n+1})) \\ (\mathrm{div}_{h,x}^* F(u^{h,n+1},.,t^n) - S(u^{h,n+1},.,t^n)), \varphi^n \Big)_h^n, \end{aligned} \tag{72d(ii))}$$

$$R_5^h(\varphi) = R_{51}^h(\varphi) + R_{52}^h(\varphi) + R_{53}^h(\varphi) + R_{54}^h(\varphi) + R_{55}^h(\varphi) + R_{56}^h(\varphi), \tag{72e}$$

$$R_{51}^h(\varphi) = \sum_{n;i\in P} w_i^{n+1} |u_i^{h,n+1} - c|^\delta (\varphi_i^{n+1} - \varphi_i^n) - \int_Q |u^h - c|^\delta \frac{d\varphi}{dt} dxdt, \tag{72e(i))}$$

$$\begin{aligned} R_{52}^h(\varphi) = \sum_{n;i\in P} \int_{t^n}^{t^{n+1}} \int_{\Omega_i^n} H_\delta^c(u_i^{h,n+1},x,t).\nabla\varphi(x,t) \\ - H_\delta^c(u_i^{h,n+1},x_i,t^n).\nabla\varphi(x_i,t^n) dxdt, \end{aligned} \tag{72e(ii))}$$

(72e(iii)) $$R_{53}^h(\varphi) = \sum_{n,i\in P} \int_{t^n}^{t^{n+1}} \int_{\Omega_i^n} (u_i^{h,n+1} sign_\delta(u_i^{h,n+1}-c) - |u_i^{h,n+1}-c|^\delta) \\ (\mathrm{div}a(x,t)\varphi(x,t) - \mathrm{div}a(x_i,t^n)\varphi(x_i,t^n))dxdt,$$

(72e(iv)) $$R_{54}^h(\varphi) = \sum_{n\geq 0}\sum_i \int_{t^n}^{t^{n+1}} \int_{\Omega_i^n} (\varphi(x_i,t^n) sign_\delta(u_i^{h,n+1}-c) \\ - \varphi(x,t) sign_\delta(u^h(x,t)-c))\overline{\mathrm{div}}_h^* F(c,x,t)dxdt,$$

(72e(v)) $$R_{55}^h(\varphi) = \sum_{n;i\in P} \int_{t^n}^{t^{n+1}} \int_{\Omega_i^n} sign_\delta(u_i^{h,n+1}) S(u_i^{h,n+1},x_i,t^n)\varphi(x_i,t^n) \\ - sign_\delta(u_i^{h,n+1}) S(u_i^{h,n+1},x,t)\varphi(x,t)dxdt,$$

(72e(vi)) $$R_{56}^h(\varphi) = \sum_{i\in P} w_i^0 \varphi_i^0 |u_i^{h,0}-c|^\delta - \int_\Omega |u^h(x,0)-c|^\delta \varphi(x,0)dx.$$

The remainders R_i^h, $i=1,\dots,5$, are estimated as follows.

Remainder R_1^h:

- R_{11}^h: Since we chose $Q_{ij}^{\eta_\delta,n}$ to be antisymmetric, we rewrite $R_{11}^h(\varphi)$ as follows:

$$R_{11}^h(\varphi) = \frac{1}{2}\sum_{n\geq 0}\Delta t \sum_{i,j\in P} w_i^n w_j^n Q_{ij}^{\eta_\delta,n}(\varphi_i^n - \varphi_j^n)||A_{ij}^n||.$$

Let A and B be two real variables not depending on i and j. We have:

$$R_{11}^h(\varphi) = \frac{1}{2}\sum_{n\geq 0}\Delta t \sum_{i,j\in P} w_i^n w_j^n Q_{ij}^{\eta_\delta,n}(\varphi_i^n - \varphi_j^n + A - B)||A_{ij}^n||.$$

We know that $2h_{\delta,ij}^n = H_{\delta,i,ij}^n.n_{ij} + H_{\delta,j,ij}^n.n_{ij} - Q_{ij}^{\eta_\delta,n}$, and thus:

$$R_{11}^h(\varphi) = -\sum_{n\geq 0}\Delta t \sum_{i,j\in P} w_i^n w_j^n (h_{\delta,ij}^n - H_{\delta,i,ij}^n.n_{ij})(\varphi_i^n - \varphi_j^n + A - B)||A_{ij}^n|| \\ + \frac{1}{2}\sum_{n\geq 0}\Delta t \sum_{i,j\in P} w_i^n w_j^n (H_{\delta,i,ij}^n - H_{\delta,j,ij}^n)(\varphi_i^n - \varphi_j^n + A - B).A_{ij}^n \\ = R_{11}^{1,h}(\varphi) + R_{11}^{2,h}(\varphi).$$

We set $A = \inf_i \inf_{x\in\Omega_i}(\varphi(t,x)-\varphi(t,x_i))$ and $B = \sup_i \sup_{x\in\Omega_i}(\varphi(t,x) - \varphi(t,x_i))$.

- *Term* $R_{11}^{1,h}(\varphi)$: Let us recall the following inequality proved in [1]:

(73) $$h_e(n,u,v) - H(u,x,t).n \leq \eta'(u)(g(n,u,v) - F(u,x,t).n).$$

Since we have $-(\varphi_i^n - \varphi_j^n + A - B) \geq 0$, by applying (73), we get:

$$R_{11}^{1,h}(\varphi) \leq \frac{1}{2}\sum_{n\geq 0}\Delta t \sum_{i,j\in P} w_i^n w_j^n (\varphi_i^n - \varphi_j^n + A - B) \\ sign_\delta(u_i^{h,n}) \underbrace{(g_{ij}^n - F_{i,ij}^n.n_{ij})}_{=C_{ij}^n(u_i^{h,n}-u_j^{h,n})} ||A_{ij}^n||.$$

By using the Cauchy–Schwartz inequality and Proposition 4.2 of [10], one gets:

$$|R_{11}^{1,h}(\varphi)| \leq C(T, |\Omega|, u_0, F, S, a)||\nabla\varphi||_\infty h^{\frac{1}{2}}. \tag{74}$$

– *Term* $R_{11}^{2,h}(\varphi)$: Since $A_{ij}^n = -A_{ji}^n$, and knowing that $|B-A| \leq ||\nabla\varphi||_\infty h$:

$$\begin{aligned}|R_{11}^{2,h}(\varphi)| \leq\ & ||\nabla\varphi||_{L^\infty(Q)} h \sum_{n;i\in P} \Delta t w_i^n \\ & \left(\sup_{v;||v||_\infty \leq C_\infty} ||H_\delta(v,.,.)||_\infty \left| \sum_{j\in P} w_j^n A_{ij}^n \right| \right. \\ & \left. + \sup_{v;||v||_\infty \leq C_\infty} ||\mathrm{div}_x H_\delta(v,.,.)||_\infty \sum_{j\in P} w_j^n |x_i - x_{ij}|\ |A_{ij}^n| \right).\end{aligned}$$

From (14), (23), and the regularity of F (2), we get:

$$|R_{11}^{2,h}(\varphi)| \leq C(F, u_0, a, S, \Omega, T)||\nabla\varphi||_\infty h. \tag{75}$$

- R_{12}^h: We have

$$R_{12}^h(\varphi) = \sum_{n\geq 0} \Delta t \sum_i w_i^n \sum_{j\in P} w_j^n (H_{\delta,i}^n - H_{\delta,i,ij}^n).(\varphi_i^n - \varphi_j^n) A_{ij}^n.$$

From (14), (23), and the regularity of F (2), one gets:

$$\begin{aligned}|R_{12}^h(\varphi)| &\leq C \sup_{v;|v|\leq C(u_0)} ||\mathrm{div}_x H_\delta(v,.,.)||_\infty ||\nabla\varphi||_\infty h \\ &\leq C(u_0, F, S, a, T, \Omega)||\nabla\varphi||_\infty h.\end{aligned} \tag{76}$$

- R_{13}^h: By using (8), (23), and the regularity of φ and F (2), one gets:

$$|R_{13}^h(\varphi)| \leq C(u_0, F, S, a, T, \Omega)||D^2\varphi||_{L^\infty(Q)} h. \tag{77}$$

- *Conclusion for* $R_1^h(\varphi)$: From (74), (75), (76), and (77), one gets:

$$\underline{|R_1^h(\varphi)| \leq C(u_0, F, S, a, T, \Omega) \left(||\varphi||_{L^\infty([0,T], C_c^1(\Omega))} \sqrt{h} + ||\varphi||_{L^\infty([0,T], C_c^2(\Omega))} h \right).} \tag{78}$$

Remainder R_2^h: This remainder is bounded by using inequalities (21):

$$|R_2^h(\varphi)| \leq \left(\sup_{u;||u||\leq C} ||C_{\Psi(u)}||_\infty(Q) + \sup_{u;||u||\leq C} ||C'_{\Psi(u)}||_\infty \right) C|\Omega|T(2||\varphi||_\infty + ||\nabla\varphi||_\infty)\delta. \tag{79}$$

Since u^h is L^∞ stable and $c = u^\epsilon$ is also bounded in $\mathbb{R}$ by a constant depending on u_0, $\sup_{u;||u||\leq C} ||C_{\Psi(u)}||_\infty$ and $\sup_{u;||u||\leq C} ||C'_{\Psi(u)}||_\infty$ are both bounded by $C(u_0, F, S, a, \Omega, T)$. Consequently:

$$\underline{|R_2^h(\varphi)| \leq C(u_0, F, S, a, T, \Omega)(||\varphi||_\infty + ||\nabla\varphi||_\infty)\delta.} \tag{80}$$

Remainder R_3^h: By using (14) and the regularity of F (2), one gets:

$$\underline{|R_3^h(\varphi)| \leq C(u_0, F, S, a, T, \Omega)||\nabla\varphi||_\infty h.} \tag{81}$$

Remainder R_4^h:

- *Term* $R_{41}^h(\varphi)$: From the regularity (2) on F, S, and thus H_δ, we have:

$$|R_{41}^h(\varphi)| \leq \left\{ \left(\sup_{u;||u||\leq C} ||\partial_u H_\delta(u,.,.)||_\infty + \sup_{u;||u||\leq C} ||\partial_u \mathrm{div} H_\delta(u,.,.)||_\infty \right) ||\nabla\varphi||_\infty \right.$$
$$+ \left(||sign_\delta||_\infty (\sup_{u;||u||\leq C} ||\partial_u F(u,.,.)||_\infty + \sup_{u;||u||\leq C} ||\partial_u \mathrm{div} F(u,.,.)||_{L^\infty(Q)} \right.$$
$$\left. \left. + \sup_{u;||u||\leq C} ||\partial_u S(u,.,.)||_{L^\infty(Q)}) \right) \ ||\varphi||_{L^\infty(Q)} \right\} \sum_{n;i\in P} \Delta t w_i^n |u_i^{h,n+1} - u_i^{h,n}| \ .$$

From Proposition 4.3 of [10] and (23), we get $|R_{41}^h(\varphi)| \leq C(F,S,a,u_0,T,\Omega)$ $(||\varphi||_\infty + ||\nabla\varphi||_\infty)(h+\sqrt{h})$.

- *Term* $R_{42}^h(\varphi)$: By using the finite increases theorem, we have:

$$|R_{42}^h(\varphi)| \leq 2C \left(\sup_{u;||u||\leq C} ||F(u,.,.)||_\infty + \sup_{u;||u||\leq C} ||\nabla_x F(u,.,.)||_\infty \right.$$
$$\left. + \sup_{u;||u||\leq C} ||S(u,.,.)||_\infty \right) ||\varphi||_\infty ||(sign_\delta)'||_{L^\infty(\mathbb{R})} \sum_n \Delta t \sum_{i\in P} w_i^n |\tilde{u^h}_i^{n+1} - u_i^{h,n+1}|.$$

From Propositions 4.2 and 4.3 in [10], we get:

$$|R_{42}^h(\varphi)| \leq C(u_0,F,S,a,T,\Omega)||\varphi||_\infty \frac{h}{\delta}. \tag{82}$$

- *Conclusion for* $R_4^h(\varphi)$: Estimates on $R_{41}^h(\varphi)$ and $R_{42}^h(\varphi)$ finally lead to:

$$|R_4^h(\varphi)| \leq C(F,S,a,u_0,T,\Omega) \left(||\varphi||_\infty \left(h + \sqrt{h} + \frac{h}{\delta} \right) + ||\nabla\varphi||_\infty (h+\sqrt{h}) \right). \tag{83}$$

Remainder R_5^h: Each term of R_5^h is estimated by using Lemma A.1 given in the appendix of [9, page 293].

- $R_{51}^h(\varphi)$, $R_{52}^h(\varphi)$, $R_{53}^h(\varphi)$, $R_{55}^h(\varphi)$:

$$R_{51}^h(\varphi) = \sum_{n;i\in P} \Delta t \omega_i J_i^{n+1} |u_i^{h,n+1} - c|^\delta \frac{(\varphi_i^{n+1} - \varphi_i^n)}{\Delta t} - \int_Q J(\xi,t)|u^h(x,t) - c|^\delta \frac{d\varphi}{dt} d\xi dt$$
$$= \sum_{n;i\in P} \int_{t^n}^{t^{n+1}} \int_{\Omega_i} J_i^{n+1} |u^h(x,t) - c|^\delta \frac{(\varphi_i^{n+1} - \varphi_i^n)}{\Delta t} - J(\xi,t)|u^h(x,t) - c|^\delta \frac{d\varphi}{dt} d\xi dt$$
$$\leq ||u^h||_\infty \sum_{n;i\in P} \int_{t^n}^{t^{n+1}} \int_{\Omega_i} \left| J_i^{n+1} \frac{(\varphi_i^{n+1} - \varphi_i^n)}{\Delta t} - J(\xi,t) \frac{d\varphi}{dt} \right| d\xi dt,$$

$J(\xi,t)$ being the Jacobian of the transformation into Lagrangian coordinates. We can now use Lemma A.1 of [9] with $b \leftarrow J$ and $\varphi \leftarrow \varphi$:

$$|R_{51}^h(\varphi)| \leq C_2 ||J||_{C^1(Q)} ||\varphi||_{C^2(Q)} (\Delta t + h) \leq C(u_0,F,S,a,T,|\Omega|)||\varphi||_{C^2(Q)} h. \tag{84}$$

By using a similar method, we get:

$$|R_{52}^h(\varphi)| \leq C_3||J||_{C^1}||H_\delta^c \nabla\varphi||_{C^1} \leq C(u_0, F, S, a, T, |\Omega|)||\varphi||_{C^1(Q)}h, \tag{85}$$

$$|R_{53}^h(\varphi)| \leq C(u_0, F, S, a, T, |\Omega|)||\varphi||_{C^1(Q)}h, \tag{86}$$

$$|R_{55}^h(\varphi)| \leq C(u_0, F, S, a, T, \Omega)||\varphi||_{C^1(Q)}h. \tag{87}$$

- $R_{54}^h(\varphi)$: We split this term into two parts:

$$\begin{aligned} R_{54}^h(\varphi) = &\sum_{n;i\in P}\int_{t^n}^{t^{n+1}}\int_{\Omega_i^n}\varphi(x_i,t^n)(sign_\delta(u_i^{h,n+1}-c)-sign_\delta(u^h(x,t)-c)) \\ &\overline{\mathrm{div}_h^*}F(c,x,t)dxdt + \sum_{n;i\in P}\int_{t^n}^{t^{n+1}}\int_{\Omega_i^n}(\varphi(x_i^n,t^n) \\ &-\varphi(x,t))sign_\delta(u^h(x,t)-c)\overline{\mathrm{div}_h^*}F(c,x,t)dxdt. \end{aligned} \tag{88}$$

Since $\overline{\mathrm{div}_h^*}F(c,x,t)$ is bounded, one can prove, by using the definition of u_h (13), that:

$$\begin{aligned} &\left|\sum_{n;i\in P}\int_{t^n}^{t^{n+1}}\int_{\Omega_i^n}\varphi(x_i,t^n)(sign_\delta(u_i^{h,n+1}-c)\right. \\ &\qquad\left.-sign_\delta(u^h(x,t)-c))\overline{\mathrm{div}_h^*}F(c,x,t)dxdt\right| \\ &\leq C(F,u_0)||\varphi||_\infty\sum_{n;i\in P}\int_{t^n}^{t^{n+1}}\int_{\Omega_i}J_i^n|sign_\delta(u_i^{h,n+1}-c)-sign_\delta(u^h(x,t)-c)|dxdt \\ &\leq C(F,u_0)||\varphi||_\infty||sign_\delta||_{L^\infty(\mathbb{R})}\sum_{n;i\in P}\int_{t^n}^{t^{n+1}}\int_{\Omega_i}\|J_i^n-J_i^{n+1}|dxdt \\ &\leq C(u_0,F,a,T,\Omega)||\varphi||_\infty h. \end{aligned}$$

By using Lemma A.1 of [9] with $b \leftarrow J$, $\varphi \leftarrow \varphi$ for the second integral of (89), we get:

$$|R_{54}^h(\varphi)| \leq C(u_0, F, S, a, T, \Omega)||\varphi||_{C_c^1(Q)}h. \tag{89}$$

- $R_{56}^h(\varphi)$:

$$|R_{56}^h(\varphi)| \leq \sum_{i\in P}|u_i^{h,0}-c|\left|\int_{\Omega_i}\varphi(\xi_i,0)-\varphi(\xi,0)d\xi\right| \leq C(u_0)||\nabla_x\varphi(0,.)||_{L^\infty(\Omega)}h. \tag{90}$$

- *Conclusion for* $R_5^h(\varphi)$: By using (84), (85), (86), (89), (87), and (90), we finally get

$$\underline{|R_5^h(\varphi)| \leq C(u_0, F, S, a, T, |\Omega|)||\varphi||_{C_c^2(Q)}h.} \tag{91}$$

Conclusion for μ_h: There exist $C(u_0, F, S, a, T, \Omega)$ such that:

$$\begin{aligned}\langle \mu_h, \varphi(.,y,.,s)\rangle \leq C \Big(& ||\varphi(.,y,.,s)||_{C^2(Q)} h + ||\varphi(.,y,.,s)||_{L^\infty([0,T];C^1(\Omega)} \sqrt{h} \\ & + ||\varphi(.,y,.,s)||_{L^\infty([0,T];C^1(\Omega))} \delta + ||\varphi(.,y,.,s)||_\infty \frac{h}{\delta} \Big).\end{aligned} \tag{92}$$

By integrating the previous inequality over y and s, and using the properties 6.134 in [9] and the upper bounds 6.89 in [9] of the chosen test function φ, one gets:

$$\boxed{\int_{\tilde{Q}} \langle \mu_h, \varphi(.,y,.,s)\rangle dyds \leq \mathcal{C} \left(h + \sqrt{h} + \delta + \gamma + \frac{\sqrt{h}}{\gamma} + \frac{h}{\gamma} + \frac{h}{\gamma^2} + \frac{h}{\delta} + \frac{\delta}{\gamma} \right).} \tag{93}$$

A.2. Measure ν_ϵ.

$$\langle \nu_\epsilon, \varphi\rangle = R_1^\epsilon(\varphi) + R_2^\epsilon(\varphi) + R_3^\epsilon(\varphi), \tag{94}$$

$$R_1^\epsilon(\varphi) = \epsilon \int_{\tilde{Q}} \nabla(sign_\delta(u^\epsilon)\nabla u^\epsilon)\varphi dyds, \tag{94a}$$

$$R_2^\epsilon(\varphi) = \int_{\tilde{Q}} \Psi_\delta^c.\nabla\varphi + \mathrm{div}\Psi_\delta^c \varphi dyds, \tag{94b}$$

$$\begin{aligned} R_3^\epsilon = & \int_{\tilde{Q}} \varphi(y,s) sign_\delta(u^\epsilon(y,s)) \overline{\mathrm{div}_h^*} F(c,y,s) dyds \\ & + \int_{\tilde{Q}} F(c,y,s).\nabla(\varphi sign_\delta(u^\epsilon)) dyds. \end{aligned} \tag{94c}$$

We now denote that $||D\varphi|| = ||D\varphi(x,.,t,.)||$ for any operator D and any norm.

Remainder R_1^ϵ: From the estimates (Lemma 4.1) on u^ϵ, we get:

$$\begin{aligned} \underline{|R_1^\epsilon(\varphi)|} & \leq \epsilon ||sign_\delta||_\infty ||\nabla\varphi||_\infty \sup_{t\in[0,T]} ||\nabla u^\epsilon(.,t)||_{L^1(\Omega)} \\ & \leq \underline{C(u_0,F,S,a,T,\Omega)||\nabla\varphi||_\infty \epsilon}. \end{aligned} \tag{95}$$

Remainder R_2^ϵ: We use similar arguments to estimate R_2^ϵ and get:

$$\underline{|R_2^\epsilon(\varphi)| \leq C(u_0,F,S,a,T,\Omega)(||\varphi||_\infty + ||\nabla\varphi||_\infty)\delta}. \tag{96}$$

Remainder R_3^ϵ: We split this remainder into $R_3^\epsilon = R_{31}^\epsilon + R_{32}^\epsilon + R_{33}^\epsilon + R_{34}^\epsilon$ with

$$\begin{aligned} R_{31}^\epsilon = & \sum_{n;i\in P} \int_{t^n}^{t^{n+1}} \int_{\Omega_i^n} (\varphi(y,s) \\ & sign_\delta(u^\epsilon(y,s)) - \varphi(x_i^n,t^n) sign_\delta(u^\epsilon(x_i^n,t^n))) \overline{\mathrm{div}_h^*} F(c,y,s) dyds, \\ R_{32}^\epsilon = & -\sum_{n;i\in P} \int_{t^n}^{t^{n+1}} \int_{\Omega_i^n} (F(c,x_i^n,t^n) - F(c,y,s)).\nabla_h(\varphi sign_\delta(u^\epsilon))(x_i^n,t^n) dyds, \\ R_{33}^\epsilon = & -\sum_{n;i\in P} \int_{t^n}^{t^{n+1}} \int_{\Omega_i^n} F(c,y,s).(\nabla_h(\varphi \\ & sign_\delta(u^\epsilon))(x_i^n,t^n) - \nabla(\varphi sign_\delta(u^\epsilon))(x_i^n,t^n)) dyds, \\ R_{34}^\epsilon = & -\sum_{n;i\in P} \int_{t^n}^{t^{n+1}} \int_{\Omega_i^n} F(c,y,s).(\nabla(\varphi sign_\delta(u^\epsilon))(y,s) - \nabla(\varphi sign_\delta(u^\epsilon))(x_i^n,t^n)) dyds. \end{aligned}$$

- *Estimate of* R_{31}^{ϵ}: The remainder R_{31}^{ϵ} is bounded by:

$$|R_{31}^{\epsilon}| \leq ||F||_{\infty} Ch \Big(||sign_{\delta}||_{\infty} (||\nabla \varphi||_{\infty} + ||\partial_s \varphi||_{\infty}) + ||\varphi||_{\infty} ||sign'_{\delta}||_{\infty} (||\partial_t u^{\epsilon}||_{L^1(\tilde{Q})} + ||\nabla u^{\epsilon}||_{L^1(\tilde{Q})}) \Big),$$

which leads to, by considering the estimates on u^{ϵ} (4.1) and (4.2):

$$|R_{31}^{\epsilon}| \leq C(u_0, F, S, a, T, \Omega)(||\nabla \varphi||_{\infty} + ||\partial_s \varphi||_{\infty} + ||\varphi||_{\infty})h. \tag{97}$$

- *Estimate of* R_{32}^{ϵ}: One can prove, by using estimates (4.1) and (4.2) that

$$|R_{32}^{\epsilon}| \leq \left(\sup_{v, |v| \leq C(u_0)} ||\nabla F(v, ., .)||_{\infty} + \sup_{v, |v| \leq C(u_0)} ||\partial_t F(v, ., .)||_{\infty} \right) Ch \left| \int_0^T \int_{\Omega} \nabla_h (\varphi sign_{\delta}(u^{\epsilon}))(y, s) dy ds \right| \leq C(u_0, F, S, a, \Omega, T) h ||\nabla \varphi||_{L^{\infty}([0,T]; C_c^1(\Omega))}. \tag{98}$$

- *Estimate of* R_{33}^{ϵ}: Using Proposition 4.2, one can deduce that

$$|R_{33}^{\epsilon}| \leq C||F^c||_{\infty} ||D^2(\varphi sign_{\delta}(u^{\epsilon}))||_{\infty} h. \tag{99}$$

By using (4.1) and (4.2), $\exists C = C(u_0, F, S, a, T, \Omega)$ such that (s.t.):

$$|R_{33}^{\epsilon}| \leq Ch \left(||D^2 \varphi||_{\infty} + \frac{1}{\delta \epsilon^{\frac{1}{2}}} ||\nabla \varphi||_{\infty} + ||\varphi||_2 \frac{1}{\epsilon^{\frac{5}{2}} \delta} + ||\varphi||_{\infty} \frac{1}{\epsilon^3 \delta^2} \right). \tag{100}$$

- *Estimate of* R_{34}^{ϵ}: By taking the $W^{1,\infty}$ of $\nabla(\varphi sign_{\delta}(u^{\epsilon}))$, one gets:

$$|R_{34}^{\epsilon}| \leq h||F^c||_{\infty} ||D^2(\varphi sign_{\delta}(u^{\epsilon}))||_{\infty}. \tag{101}$$

We thus obtain the same upper bound as for the remainder R_{33}^{ϵ}.

Conclusion for $\boldsymbol{\nu_{\epsilon}}$: There exist $C = C(u_0, F, S, a, T, \Omega)$ such that:

$$\langle \nu_{\epsilon}, \varphi(x, ., t, .) \rangle \leq C \Bigg((||\varphi(x, ., t, .)||_{C_c^1(\tilde{Q})} + ||D_y^2 \varphi(x, ., t, .)||_{\infty})h + ||\varphi(x, ., t, .)||_{L^{\infty}([0, T+\gamma]; C_c^1(\tilde{\Omega}))} \left(\epsilon + \delta + \frac{h}{\delta \epsilon^{\frac{1}{2}}} \right) + ||\varphi(x, ., t, .)||_{\infty} \left(\frac{h}{\delta \epsilon^{\frac{5}{2}}} + \frac{h}{\delta^2 \epsilon^3} \right) \Bigg). \tag{102}$$

By integrating the previous inequality over x and t, and using the properties 6.134 in [9] and the upper bounds 6.89 in [9] of the chosen test function φ, one gets:

$$\boxed{\int_Q \langle \nu_{\epsilon}, \varphi(x, ., t, .) \rangle dx dt \leq \mathcal{C} \left(\delta + h + \frac{\delta}{\gamma} + \frac{\epsilon}{\gamma} + \frac{h}{\gamma} + \frac{h}{\gamma^2} + \frac{h}{\delta \gamma \epsilon^{\frac{1}{2}}} + \frac{h}{\delta \epsilon^{\frac{5}{2}}} + \frac{h}{\epsilon^3 \delta^2} \right).} \tag{103}$$

A.3. Measure λ. The remainders of λ are defined by:

$$\langle \lambda, \varphi \rangle = \sum_{i=1}^{10} R_i^c(\varphi) + R^{c,L^1}(\psi), \tag{104}$$

$$R_1^c(\varphi) = \int_{Q\times\tilde{Q}} \left(|u^h(x,t) - u^\epsilon(y,s)| - |u^h(x,t) - u^\epsilon(y,s)|^\delta \right) \psi_t(t,x)\zeta^\gamma(x-y)\bar{\zeta}^\gamma(t-s)dxdydsdt, \tag{104a}$$

$$R_2^c(\varphi) = \int_{Q\times\tilde{Q}} \left(|u^h(x,t) - u^\epsilon(y,s)| - |u^h(x,t) - u^\epsilon(y,s)|^\delta \right) a(x,t) .\nabla\psi(t,x)\zeta^\gamma(x-y)\bar{\zeta}^\gamma(t-s)dxdydsdt, \tag{104b}$$

$$R_3^c(\varphi) = \int_{Q\times\tilde{Q}} |u^h(x,t) - u^\epsilon(y,s)|^\delta (\mathrm{div}a(y,s) - \mathrm{div}a(x,t)) \psi(t,x)\zeta^\gamma(x-y)\bar{\zeta}^\gamma(t-s)dxdydtds, \tag{104c}$$

$$R_4^c(\varphi) = 2\int_{Q\times\tilde{Q}} \left(|u^h(x,t) - u^\epsilon(y,s)|^\delta - |u^h(x,t) - u^\epsilon(y,s)| \right) \mathrm{div}a(x,t) \psi(t,x)\zeta^\gamma(x-y)\bar{\zeta}^\gamma(t-s)dxdydtds, \tag{104d}$$

$$R_5^c(\varphi) = \int_{Q\times\tilde{Q}} sign_\delta(u^h(x,t) - u^\epsilon(y,s))u^\epsilon(y,s)(\mathrm{div}(x,t) - \mathrm{div}(y,s)) \psi(t,x)\zeta^\gamma(x-y)\bar{\zeta}^\gamma(t-s)dxdydtds, \tag{104e}$$

$$R_6^c(\varphi) = \int_{Q\times\tilde{Q}} sign_\delta(u^h(x,t) - u^\epsilon(y,s)) \left(\overline{\mathrm{div}_x^{h,*}} F(u^h(x,t),x,t) - \overline{\mathrm{div}_x^{h,*}} F(u^h(x,t),y,s) \right) \psi(t,x)\zeta^\gamma(x-y)\bar{\zeta}^\gamma(t-s)dxdydtds, \tag{104f}$$

$$R_7^c(\varphi) = \int_{Q\times\tilde{Q}} sign_\delta(u^h(x,t) - u^\epsilon(y,s)) \left(S(u^\epsilon(y,s),y,s) - S(u^\epsilon(y,s),x,t) \right) \psi(t,x)\zeta^\gamma(x-y)\bar{\zeta}^\gamma(t-s)dxdydtds, \tag{104g}$$

$$R_8^c(\varphi) = \int_{\Omega\times\tilde{Q}} \left(|u^h(x,0) - u^\epsilon(y,s)|^\delta - |u^h(x,0) - u^\epsilon(y,s)| \right) \psi(x,0)\zeta^\gamma(x-y)\bar{\zeta}^\gamma(s)dxdyds, \tag{104h}$$

$$R_9^c(\varphi) = \int_{Q\times\tilde{Q}} (1 + C_\psi(T-t)) \left(|u^h(x,t) - u^\epsilon(y,s)| - |u^h(x,t) - u^\epsilon(x,t)| \right) \psi(t,x)\zeta^\gamma(x-y)\bar{\zeta}^\gamma(t-s)dxdydsdt, \tag{104i}$$

$$R_{10}^c(\varphi) = \int_{\Omega\times\tilde{Q}} (|u^h(x,0) - u^\epsilon(y,s)| - |u^h(x,0) - u^\epsilon(x,0)|) \psi(0,x)\zeta^\gamma(x-y)\bar{\zeta}^\gamma(s)dxdyds, \tag{104j}$$

$$R^{c,L^1}(\psi) = \int_{Q\times\tilde{Q}} |u^\epsilon(x,t) - u^\epsilon(y,s)| \, |\psi(t,x)| \chi_{[-\gamma,0]}(t-s)\chi_{B(0,\gamma)}(x-y)dydsdxdt. \tag{104k}$$

Remainder $R_1^c(\varphi)$, $R_2^c(\varphi)$, $R_4^c(\varphi)$, $R_8^c(\varphi)$: From (20), $|\,|u|^\delta - |u|\,| \le C\delta$. Thus:

$$|R_1^c(\varphi)| \le C(u_0,F,S,a,T,\Omega)\delta||\partial_t\psi||_\infty, \tag{105}$$

$$\underline{|R_2^c(\varphi)| \le C(u_0,F,S,a,T,\Omega)\delta||\nabla\psi||_\infty}, \tag{106}$$

$$|R_4^c(\varphi)| \le C(u_0,F,S,a,T,\Omega)\delta||\psi||_\infty, \tag{107}$$

$$\underline{|R_8^c(\varphi)| \le C(u_0)\delta||\psi(.,0)||_{L^1(\Omega)} \le C(u_0)|\Omega|\delta||\psi(.,0)||_{L^\infty(\Omega)}}. \tag{108}$$

Remainder $R_3^c(\varphi)$: From hypothesis (2), Proposition 4.2 of [10], and Lemma 4.1, we have:

$$\begin{aligned}|R_3^c(\varphi)| &\leq (||\partial_t \mathrm{div} a||_\infty + ||\nabla_x \mathrm{div} a||_\infty)\,(||u^\epsilon||_\infty + ||u^h||_\infty)\gamma||\psi||_{L^1(Q)} \\ &\leq C(u_0, F, S, a, T, \Omega)\gamma||\psi||_\infty.\end{aligned} \tag{109}$$

Remainder $R_5^c(\varphi)$: From hypothesis (2) for a and Lemma 4.1 for u^ϵ, we get:

$$|R_5^c(\varphi)| \leq C(u_0, F, S, a, T, \Omega)\gamma||\psi||_\infty. \tag{110}$$

Remainder $R_6^c(\varphi)$: We prove in [9, page 218] that $\overline{\mathrm{div}^*_{h,x}}F$ is Lipschitz with respect to x and t:

$$|\overline{\mathrm{div}^*_{h,x}}F(u^h(x,t),x,t) - \overline{\mathrm{div}^*_{h,x}}F(u^h(x,t),y,s)| \leq C(|t^m - t^n| + |x_i - x_k| + O(h))$$

$$\begin{aligned}\Rightarrow |R_6^c(\varphi)| &\leq C\sum_{n,m}\int_{t^n}^{t^{n+1}}\int_{s^m}^{s^{m+1}}\sum_{i,k}\int_{\Omega_i}\int_{\Omega_k}|\psi(x,t)|\zeta^\gamma(x-y)\zeta^\gamma(t-s) \\ &\qquad (|t^m - t^n| + |x_i - x_k| + O(h))dxdydtds \\ &\leq C(u_0, F, S, a, T, \Omega)(h+\gamma)||\psi||_\infty.\end{aligned} \tag{111}$$

The last upper bound comes from: If $x \in \Omega_i$ and $y \in \Omega_k$, then $|x_i - x_k| \leq |x_i - x| + |x - y| + |y - x_k|$.

Remainder $R_7^c(\varphi)$: From hypothesis (2) for S, we have:

$$\begin{aligned}&|R_7^c(\varphi)| \leq ||sign_\delta||_\infty||\psi||_1(\sup_{c;||c||\leq C}\{||\mathrm{div}_x S(c,.,.)||_\infty\} \\ &+\sup_{c;||c||\leq C}\{||\partial_t S(c,.,.)||_\infty\})\gamma \leq C(u_0, F, S, a, T, \Omega)\gamma||\psi||_\infty.\end{aligned} \tag{112}$$

Remainders $R_9^c(\varphi)$ and $R^{c,L^1}(\psi)$: From the definition of the functions ζ and $\bar{\zeta}$ and the fact that $C_\psi = C(u_0, T, \Omega, S, F, a)$, $R_9^c(\varphi)$ and $R^{c,L^1}(\psi)$ are linked by:

$$|R_9^c(\varphi)| \leq C(u_0, T, \Omega, S, F, a)R^{c,L^1}(\psi). \tag{113}$$

By using the regularity of u^ϵ in Lemma 4.1, one gets:

$$\begin{aligned}&\int_{Q\times\bar{Q}}|\varphi(x,y,t,s)|\;|u^\epsilon(y,s) - u^\epsilon(x,t)|dxdydsdt \\ &\leq C||\psi||_\infty\int_{Q\times\bar{Q}}\int_0^1\sup_{|\tau_x|\leq\gamma}(|\nabla_x u^\epsilon(x+\tau_x, t+\theta(s-t))|\;|x-y| \\ &\qquad + |\partial_t u^\epsilon(x+\tau_x, t+\theta(s-t))|\;|t-s|)d\theta \\ &\qquad \chi_{[-\gamma,0]}(t-s)\chi_{B(0,\gamma)}(x-y)dxdydtds.\end{aligned}$$

Let τ_x^0 be the value for which the sup value is reached, and we then can bound by L^1 norms of ∇u^ϵ and $\partial_t u^\epsilon$:

$$R^{c,L^1}(\psi) \leq C||\psi||_\infty\gamma\int_0^T(||\nabla_x u^\epsilon(.,t+\tau_t)||_{L^1(\tilde{\Omega})} + ||\partial_t u^\epsilon(.,t+\tau_t)||_{L^1(\tilde{\Omega})})dt,$$

where $t + \tau_t$ is between 0 and $T + \gamma$, so that we can use estimates (4.1) and get:

$$\underline{R^{c,L^1}(\psi) \leq C(u_0, F, S, a, T, \Omega)\gamma||\psi||_\infty.} \tag{114}$$

Remainder $R_{10}^c(\varphi)$: We first have

$$|R_{10}^c(\varphi)| \leq C \int_{Q\times\tilde{Q}} |u^\epsilon(y,s) - u^\epsilon(x,0)|\ |\psi(x,0)| \chi_{[-\gamma,0]}(-s)\chi_{B(0,\gamma)}(x-y)dxdyds. \tag{115}$$

We then proceed exactly as for the remainder R^{c,L^1} with $t=0$, and we get:

$$\underline{|R_{10}^c(\varphi)| \leq C(u_0,F,S,a,\Omega)\gamma||\psi||_\infty.} \tag{116}$$

A.3.1. Conclusion for λ. From (106), (107), (109), (108), (110), (111), (112), (113), (116), (114), and the upper bounds on the norms of ψ 6.89 in [9], we get:

$$\boxed{\langle \lambda, \varphi\rangle \leq C(u_0,F,S,a,T,\Omega)\left(\delta+\gamma+h+\frac{h}{\gamma}+\frac{h}{\gamma^2}\right).} \tag{117}$$

Appendix B. L^1 terms.

$$T_{L^1} = T_2 + T_5 + T_6' + T_7' + T_8' + T_9', \tag{118}$$

$$\begin{aligned} T_2 = \int_{Q\times\tilde{Q}} & |u^h(x,t) - u^\epsilon(y,s)|^\delta (a(x,t)-a(y,s)).\psi(t,x) \\ & \nabla\zeta^\gamma(x-y)\bar{\zeta}^\gamma(t-s)dxdydsdt, \end{aligned} \tag{118a}$$

$$\begin{aligned} T_5 = \int_{Q\times\tilde{Q}} & sign_\delta(u^h(x,t)-u^\epsilon(y,s))\left(F(u^h(x,t),x,t)\right. \\ & \left. - F(u^h(x,t),y,s)\right).\psi(t,x)\nabla_x\zeta^\gamma(x-y)\bar{\zeta}^\gamma(t-s) \\ & + sign_\delta(u^h(x,t)-u^\epsilon(y,s))\left(F(u^\epsilon(y,s),y,s)\right. \\ & \left. - F(u^\epsilon(y,s),x,t)\right)\psi(t,x).\nabla_x\zeta^\gamma(x-y)\bar{\zeta}^\gamma(t-s)dxdydtds, \end{aligned} \tag{118b}$$

$$T_6' = 2\int_{Q\times\tilde{Q}} |u^h(x,t)-u^\epsilon(y,s)|\mathrm{div}a(x,t)\psi(t,x)\zeta^\gamma(x-y)\bar{\zeta}^\gamma(t-s)dxdydtds, \tag{118c}$$

$$\begin{aligned} T_7' = \int_{Q\times\tilde{Q}} & sign_\delta(u^h(x,t)-u^\epsilon(y,s))(u^h(x,t)-u^\epsilon(y,s))\mathrm{div}a(x,t) \\ & \psi(t,x)\zeta^\gamma(x-y)\bar{\zeta}^\gamma(t-s)dxdydtds, \end{aligned} \tag{118d}$$

$$\begin{aligned} T_8' = \int_{Q\times\tilde{Q}} & sign_\delta(u^h(x,t)-u^\epsilon(y,s))\left(\overline{\mathrm{div}^*_{h,x}}F(u^\epsilon(y,s),x,t)\right. \\ & \left. - \overline{\mathrm{div}^*_{h,x}}F(u^h(x,t),x,t)\right)\psi(t,x)\zeta^\gamma(x-y)\bar{\zeta}^\gamma(t-s)dxdydtds, \end{aligned} \tag{118e}$$

$$\begin{aligned} T_9' = \int_{Q\times\tilde{Q}} & sign_\delta(u^h(x,t)-u^\epsilon(y,s))\left(S(u^\epsilon(y,s),x,t)\right. \\ & \left. - S(u^h(x,t),x,t)\right)\psi(t,x)\zeta^\gamma(x-y)\bar{\zeta}^\gamma(t-s)dxdydtds. \end{aligned} \tag{118f}$$

Appendix C. Table of symbols.

Notations	Definitions
P, d	Set of particles, space dimension (section 2)
$\Omega_i(t),\ w_i(t),\ \ i \in P$	Particle volumes and weights at time t (section 2)
$h,\ \varepsilon$	Characteristic distance between particles—smoothing length (section 2)
$\left(A_{ij}^n\right), B_i^n$	Shape vector (6), renormalization matrices (7)
D_h, D_h^*	Derivative approximation operator (5) and its adjoint
$\eta(u), \eta_\delta(u), \delta$	Entropy function, regularized Kruskov entropy (18), and reg. parameter
u	Exact solution of the conservation law (1)
u^ϵ,ϵ	Solution of the pseudoviscous problem (27) and reg. parameter
u^h	Solution of the numerical scheme (9), (13)
$a, F(v,x,t), S(v,x,t), u_0(x)$	Transport, flux, source, and initial condition of the conservation law (1)
L_a, L_a^*	Transport operator and its adjoint
$n(x),\ n_{ij}$	Normal vector, normal vector from i to j
$g(n(x),u,v), Q(n(x),u,v)$	Numerical flux and viscosity of the scheme
$\Delta t,\ \overline{\Delta x_i^n}$	Time step—mean distance between i and its neighbors (12)
$h_e(n(x),u,v), h_\delta$	Entropic numerical flux (17) and Kruskov entropic numerical flux
$H(u,x,t),\ H_\delta,\ H_\delta^c,\ \Psi_\delta^c$	General entropy flux and regularized Kruskov entropy flux (19)
$Q^\eta(n(x),u,v), Q^{\eta_\delta}$	Entropic numerical viscosity and Kruskov entropic numerical viscosity (see (17))
$\mu_h,\ R^h$	Measure and remainders related to u^h entropy inequality
$\nu_\epsilon,\ R^\epsilon$	Measure and remainders related to u^ϵ entropy inequality
$\lambda,\ \lambda_1,\ R^c$	Measures and remainders due to the comparison of u^h and u^ϵ entropy inequalities
$\varphi,\ \psi,\ \zeta,\ \bar{\zeta},\ \Omega,\ T,\ \gamma$	Test function and its parameters (Definition 6.1)

REFERENCES

[1] S. Benharbit, A. Chalabi, and J. P. Vila, *Numerical viscosity and convergence of finite volume methods for conservation laws with boundary conditions*, SIAM J. Numer. Anal., 32 (1995), pp. 775–796.

[2] B. Ben Moussa and J. P. Vila, *Convergence of SPH method for scalar nonlinear conservation laws*, SIAM J. Numer. Anal., 37 (2000), pp. 863–887.

[3] F. Bouchut and B. Perthame, *Kruzkov's estimate for conservation laws revisited*, Trans. Amer. Math. Soc., 350 (1998), pp. 2847–2870.

[4] B. Cockburn, F. Coquel, and P. LeFloch, *Convergence of the finite volume method for multidimensional conservation laws*, SIAM J. Numer. Anal., 32 (1995), pp. 687–705.

[5] R. Eymard, T. Gallouet, and R. Herbin, *The finite volume method*, in Handbook of Numerical Analysis, Ph. Ciarlet and J. L. Lions eds., North-Holland, Amsterdam, 2000.

[6] E. Godlewski and P. A. Raviart, *Hyperbolic systems of conservation laws*, SMAI Math. Appl., 3/4 (1991).

[7] N. Kruskov, *First order quasilinear equations with several space variables*, Mat. Sb., 10 (1970), pp. 217–243.

[8] N. N. Kuznetsov, *Accuracy of some approximate methods for computing the weak solutions of a first order quasi-linear equation*, Comput. Math. Math. Phys., 16 (1976), pp. 105–119.

[9] N. Lanson, *Etude des Méthodes Particulaires Renormalisées. Applications aux Probèmes de Dynamique Rapide*, Ph.D. thesis, INSA-Université Paul Sabatier, Toulouse, France, 2001; also available online from www.math.uwaterloo.ca/˜ nlanson/thesis1.pdf.

[10] N. Lanson and J. P. Vila, *Renormalized meshfree schemes: Consistency, stability and hybrid methods for conservation laws*, SIAM J. Numer. Anal., submitted.

[11] A. Leroux, *Approximation de Quelques Problèmes Hyperboliques non Linéaires*, Ph.D. thesis, Université de Rennes, France, 1979.

[12] S. Osher, *Riemann solvers, the entropy condition, and difference approximations*, SIAM J. Numer. Anal., 21 (1984), pp. 217–235.
[13] J. Málek, J. Necas, M. Rokyta, and M. Ruzicka, *Weak and Measure Valued Solutions to Evolutionary PDE's*, Applied Mathematics and Mathematical Computation 13, Chapman and Hall, London, 1996.
[14] E. Tadmor, *Numerical viscosity and the entropy condition for conservative difference schemes*, Math. Comput., 43 (1984), pp. 369–381.
[15] J. P. Vila, *Weighted particle-finite volume hybrid schemes*, in Finite Volume for Complex Applications, Benkhaldoun-Vilsmeir, eds., HERMES, London, 1996.

SIAM J. NUMER. ANAL.
Vol. 46, No. 4, pp. 1965–1984
© 2008 Society for Industrial and Applied Mathematics

A FAST FOURIER–GALERKIN METHOD FOR SOLVING SINGULAR BOUNDARY INTEGRAL EQUATIONS*

HAOTAO CAI† AND YUESHENG XU‡

Abstract. We propose in this paper a convenient way to compress the dense matrix representation of a compact integral operator with a smooth kernel under the Fourier basis. The compression leads to a sparse matrix with only $\mathcal{O}(n \log n)$ nonzero entries, where $2n$ or $2n+1$ denotes the order of the matrix. Based on this compression strategy, we develop a fast Fourier–Galerkin method for solving a class of singular boundary integral equations. We prove that the fast Fourier–Galerkin method gives the optimal convergence order $\mathcal{O}(n^{-t})$, where t denotes the degree of regularity of the exact solution. Moreover, we design a fast scheme for solving the corresponding truncated linear system. We show that solving this system requires only an $\mathcal{O}(n \log^2 n)$ number of multiplications. We present numerical examples to confirm the theoretical estimates and to demonstrate the efficiency and accuracy of the proposed algorithm.

Key words. singular boundary integral equations, Fourier–Galerkin methods, fast solutions

AMS subject classifications. 65R20, 45E05

DOI. 10.1137/070703478

1. Introduction.

The Galerkin method using the trigonometric polynomial basis (which we call the Fourier basis) among other numerical methods (cf. [1, 2, 3, 4, 10, 12, 13, 14, 18, 20, 21]) is a standard numerical treatment for solving singular boundary integral equations. In this paper, we call such a method the *Fourier–Galerkin method*. This method enjoys nice convergence properties but at the same time suffers from having large computational costs. The coefficient matrix (which we call the Galerkin matrix) of the corresponding discrete linear system is a dense matrix. As a result, it has $\mathcal{O}(n^2)$ nonzero entries, where $2n$ or $2n+1$ denotes the order of the system. When n is large, the computational cost needed to set up the Galerkin matrix and then to solve the system is large. Therefore, approximating the dense Galerkin matrix by a sparse matrix is practically important.

The main purpose of this paper is to develop a fast Fourier–Galerkin method for solving the equation by compressing the dense Galerkin matrix to a sparse matrix. A large class of boundary integral operators can be written in the form $\mathcal{A}+\mathcal{B}$, where $\mathcal{A}$ is a singular but simple integral operator and $\mathcal{B}$ is a compact integral operator with a smooth kernel. Operator $\mathcal{A}$ carries the main singularity characteristic of the boundary integral operator, and it has the Fourier basis functions as its eigenfunctions. Hence, the matrix representation of the operator $\mathcal{A}$ under the Fourier basis is a diagonal matrix, and we observe that the matrix representation of the compact operator $\mathcal{B}$

*Received by the editors September 24, 2007; accepted for publication January 22, 2008; published electronically April 18, 2008.

http://www.siam.org/journals/sinum/46-4/70347.html

†Department of Mathematics and Statistics, Shandong University of Finance, Jinan, 250002, People's Republic of China, and Academy of Mathematics and System Science, Chinese Academy of Sciences, Beijing, 100080, People's Republic of China (caihaotao@126.com).

‡Corresponding author. Department of Mathematics, Syracuse University, Syracuse, NY 13244, and Academy of Mathematics and System Science, Chinese Academy of Sciences, Beijing, 100080, People's Republic of China (yxu06@syr.edu). This author's research was supported in part by the U.S. National Science Foundation under grants CCR-0407476 and DMS-0712827, by the National Aeronautics and Space Administration under cooperative agreement NNX07AC37A, by the Natural Science Foundation of China under grants 10371122 and 10631080, and by the Education Ministry of the People's Republic of China under the Changjiang Scholar Chair Professorship Program through Sun Yat-sen University.

under the same basis can be compressed to a sparse matrix without loss of critical information encoded in the matrix. This serves as the base for the development of the fast Fourier–Galerkin method for solving the boundary integral equations. We propose a truncation strategy for the Galerkin matrix. We prove that the truncated sparse matrix has only $\mathcal{O}(n \log n)$ nonzero entries, when the trigonometric polynomials of degree n are used in the Galerkin method and that the approximate solution obtained from the truncated Fourier–Galerkin method has the same convergence order as that obtained from the conventional Fourier–Galerkin method. In other words, the proposed compression strategy preserves the convergence property of the conventional Fourier–Galerkin method but reduces the number of nonzero entries of the Galerkin matrix from $\mathcal{O}(n^2)$ to $\mathcal{O}(n \log n)$. We also develop a fast solver for the corresponding compressed linear system and prove that solving the linear system in the proposed way requires only an $\mathcal{O}(n \log^2 n)$ number of multiplications.

We organize this paper in seven sections. In section 2, we review the Fourier–Galerkin method for solving boundary integral equations. A truncation strategy for the matrix representation of a compact integral operator with a smooth kernel under the Fourier basis is proposed in section 3. This truncation strategy leads to a *truncated* Fourier–Galerkin method for solving boundary integral equations. We show that the truncated Galerkin matrix has only $\mathcal{O}(n \log n)$ nonzero entries. We prove in section 4 the stability of the truncated Fourier–Galerkin method, which ensures the unique solvability of the truncated method. In section 5, we establish the optimal convergence order of the truncated method, which is the same as that of the conventional Fourier–Galerkin method. Section 6 is devoted to the development of a fast solver for the linear system resulting from the truncated Fourier–Galerkin method. The fast algorithm is based on a multilevel augmentation method proposed in [6]. We prove that this fast solution of the linear system will not ruin the convergence order of the truncated Fourier–Galerkin method and show that the total number of multiplications required for solving the linear system is $\mathcal{O}(n \log^2 n)$. Finally in section 7, we present numerical examples to demonstrate the efficiency and accuracy of the proposed algorithm.

2. The Fourier–Galerkin method for boundary integral equations. We describe in this section the Fourier–Galerkin method for solving singular boundary integral equations.

Let us first recall the formulation of boundary integral equations. As examples, we present three specific boundary integral equations which are reformulations of the two-dimensional Laplace equation with either the Dirichlet or the Neumann boundary condition. We begin with an introduction of appropriate Sobolev spaces. For $k \in \mathbb{Z} := \{\dots, -1, 0, 1, \dots\}$, we set

$$e_k(x) := \frac{1}{\sqrt{2\pi}} e^{ikx}, \quad x \in I := [0, 2\pi],$$

where i is the imaginary unit such that $i^2 := -1$. For $s \geq 0$, we denote by $H^s(I)$ the Sobolev space of functions $\phi \in L^2(I)$ with property $\sum_{k\in\mathbb{Z}}(1+k^2)^s|\phi_k|^2 < +\infty$, where ϕ_k are the Fourier coefficients of ϕ defined by $\phi_k := \int_I \phi(x) e_k(x) dx, \ k \in \mathbb{Z}$. The inner product of space $H^s(I)$ is defined by

$$(\phi, \psi)_s := \sum_{k\in\mathbb{Z}} (1+k^2)^s \phi_k \overline{\psi}_k, \quad \phi, \psi \in H^s(I),$$

and its norm is defined by $\|\phi\|_s := (\phi, \phi)^{\frac{1}{2}}_s$.

For two nonnegative numbers s_1 and s_2, we suppose that $X \subseteq H^{s_1}(I)$ and $Y \subseteq H^{s_2}(I)$ are two subspaces and that $\mathcal{A} : X \to Y$ is a bounded linear operator defined by

$$(\mathcal{A}w)(x) = \int_I a(x,y)w(y)dy, \quad x \in I,$$

which has a bounded inversion $\mathcal{A}^{-1} : Y \to X$. The kernel a of the operator $\mathcal{A}$ is either weakly singular, strongly singular, or hypersingular. We assume that $\mathcal{B} : X \to Y$ is a compact operator defined by

$$(\mathcal{B}w)(x) = \int_I b(x,y)w(y)dy, \quad x \in I,$$

with a smooth kernel b. We consider solving the operator equation

$$(\mathcal{A} + \mathcal{B})u = g, \tag{2.1}$$

where $g \in Y$ is a given function and $u \in X$ is the unknown to be determined.

Many boundary integral equations have the form (2.1). We now follow [1, 11] to describe three boundary integral equations in the form (2.1). To this end, we let D be a bounded open simply connected region in $\mathbb{R}^2$ and ∂D be its boundary. We suppose that the boundary ∂D has a parametrization γ defined by $\gamma(x) = (\gamma_1(x), \gamma_2(x))$, $x \in I$. We denote by $\mathbf{n}_P$ the unit normal vector at the point P on the boundary ∂D.

In the first case, we consider the interior Dirichlet problem in the plane

$$\begin{cases} \triangle U(P) = 0, & P \in D, \\ U(P) = f(P), & P \in \partial D. \end{cases} \tag{2.2}$$

The solution U of (2.2) can be expressed as a single-layer potential in the form

$$U(P) = \int_{\partial D} \rho(Q) \log |P - Q| ds_Q, \quad P \in D,$$

where the unknown density function ρ satisfies the boundary integral equation of the first kind

$$\int_{\partial D} \rho(Q) \log |P - Q| ds_Q = f(P), \quad P \in \partial D. \tag{2.3}$$

We next rewrite (2.3) in the form (2.1) by introducing two kernels

$$a(x,y) := -\frac{1}{\pi} \log \left| 2e^{-1} \sin \frac{y-x}{2} \right|, \quad x, y \in I, \tag{2.4}$$

and

$$b(x,y) := \begin{cases} -\frac{1}{\pi} \log \left| \frac{e^{\frac{1}{2}} |\gamma(x) - \gamma(y)|}{|2 \sin(\frac{x-y}{2})|} \right|, & x - y \neq 2k\pi, \quad k \in \mathbb{Z}, \\ -\frac{1}{\pi} \log |e^{\frac{1}{2}} \gamma'(y)|, & \text{otherwise}, \end{cases} \qquad x, y \in I.$$

We also let $g(x) := -\frac{1}{\pi} f(\gamma(x))$ and $u(x) := \rho(\gamma(x))|\gamma'(x)|$, for $x \in I$. It is known (cf. [1, 3, 20, 22]) that u satisfies (2.1) with the kernel a being weakly singular when the boundary is smooth. In this case, the domain space and the range space for the corresponding integral operators are $X = H^0(I)$ and $Y = H^1(I)$, respectively.

The second boundary integral equation is for the interior Neumann problem

$$\begin{cases} \triangle U(P) = 0, & P \in D, \\ \frac{\partial U(P)}{\partial \mathbf{n}_P} = f(P), & P \in \partial D. \end{cases} \tag{2.5}$$

The solution U of (2.5) can be expressed by a double-layer potential

$$U(P) := \int_{\partial D} \rho(Q) \frac{\partial}{\partial \mathbf{n}_Q} \log |P - Q| ds_Q, \quad P \in D. \tag{2.6}$$

Likewise, we define two kernels by

$$a(x, y) := \frac{\sin(y - x)}{2(1 - \cos(y - x))}, \quad x, y \in I, \tag{2.7}$$

and

$$b(x, y) := \frac{\gamma'(x) \cdot (\gamma(y) - \gamma(x))}{|\gamma(x) - \gamma(y)|^2} - \frac{\sin(y - x)}{2(1 - \cos(y - x))}, \quad x, y \in I,$$

and introduce functions $g(x) := f(\gamma(x))|\gamma'(x)|$ and $u(x) := \frac{d\rho(\gamma(x))}{dx}$ for $x \in I$. According to [1, 15] we find that the unknown u satisfies (2.1) with the kernels defined above. Note that the kernel a has a strong singularity along the diagonal. The domain space and the range space of the corresponding integral operators are $X = Y = H_0^0(I)$, where

$$H_0^0(I) := \left\{ \phi \in H^0(I) : \int_I \phi(x) dx = 0 \right\}.$$

The third case is the hypersingular integral equation reformulation of the interior Neumann problem (2.5). Again, we express the solution U of (2.5) as a double-layer potential (2.6). This time, we choose the kernels differently. Specifically, we let

$$a(x, y) := -\frac{1}{4\pi} \csc^2 \frac{y - x}{2}, \quad x, y \in I, \tag{2.8}$$

$$b(x, y) := \frac{1}{\pi} \frac{\partial}{\partial y} \left[\frac{\gamma(x) \cdot (\gamma(y) - \gamma(x))}{|\gamma(x) - \gamma(y)|^2} - \frac{\sin(y - x)}{2(1 - \cos(y - x))} \right], \quad x, y \in I,$$

$g(x) := \frac{1}{\pi} f(\gamma(x))|\gamma'(x)|$, and $u(x) := \rho(\gamma(x))$ for $x \in I$. Then it is known (cf. [15]) that u satisfies the hypersingular boundary integral equation having the form (2.1). The domain space and the range space for the corresponding integral operators are $X = H_0^0(I) \cap H^1(I)$ and $Y = H_0^0(I)$, respectively.

We now turn to describing the Fourier–Galerkin method for solving the boundary integral equations (2.1). To this end, we introduce several index sets: $\mathbb{N} := \{1, 2, \dots\}$, $\mathbb{N}_0 := \{0, 1, 2, \dots\}$, $\mathbb{Z}_n^+ := \{1, 2, \dots, n\}$, and $\mathbb{Z}_n := \mathbb{Z}_n^+ \cup \{0\}$. For each $n \in \mathbb{N}$, we define a finite-dimensional subspace X_n by $X_n := \operatorname{span}\{e_k(x) : |k| \in \mathbb{Z}_n^*\}$, where $\mathbb{Z}_n^*$ denotes either $\mathbb{Z}_n$ or $\mathbb{Z}_n^+$, according to the type of integral operators $\mathcal{A}$ in (2.1). For example, if the operator is weakly singular, then we choose $\mathbb{Z}_n^* := \mathbb{Z}_n$, and if it is strongly singular or hypersingular, then we choose $\mathbb{Z}_n^* := \mathbb{Z}_n^+$.

The Fourier–Galerkin method for solving (2.1) is to project the equation onto the space X_n generated by the Fourier basis. Let $\mathcal{P}_n$ be the orthogonal projection from

X to X_n. It is well known (cf. [1, 11]) that there exists a positive constant c such that, for any $v \in H^\mu(I)$ and for any $\nu \in [0, \mu]$,

$$\|v - \mathcal{P}_n v\|_\nu \le c n^{\nu-\mu} \|v\|_\mu, \tag{2.9}$$

and $\lim_{n\to\infty} \|v - \mathcal{P}_n v\|_\mu = 0$. Thus, there exists a positive constant ς such that, for all n, $\|\mathcal{P}_n\|_\mu \le \varsigma$. The Fourier–Galerkin method for solving (2.1) is to seek $u_n \in X_n$ such that

$$(\mathcal{P}_n\mathcal{A} + \mathcal{P}_n\mathcal{B})u_n = \mathcal{P}_n g. \tag{2.10}$$

Throughout this paper, we always assume that the operator $\mathcal{A}$ has the Fourier basis functions e_k as its eigenfunctions. Note that the singular integral operators $\mathcal{A}$ in the three boundary integral equation examples all have this property. Thus, the operator $\mathcal{A}$ commutes with the projection $\mathcal{P}_n$; that is,

$$\mathcal{A}\mathcal{P}_n = \mathcal{P}_n\mathcal{A}. \tag{2.11}$$

Equation (2.11) is equivalent to

$$\mathcal{A}^{-1}\mathcal{P}_n = \mathcal{P}_n\mathcal{A}^{-1}. \tag{2.12}$$

Let $\mathcal{B}_n := \mathcal{P}_n\mathcal{B}|_{X_n}$ and $g_n := \mathcal{P}_n g$. By condition (2.11), (2.10) is rewritten as:

$$(\mathcal{A} + \mathcal{B}_n)u_n = g_n. \tag{2.13}$$

Once a solution of this equation is solved, we can obtain an approximation of $U(P)$ via the potential. Efficient numerical methods for computation of the potential may be found in [19].

The following stability and convergence results regarding the Fourier–Galerkin method can be found in [1, 11].

THEOREM 2.1. *Suppose that* (2.1) *has the unique solution* u. *Then there exist a positive integer* n_0 *and a positive constant* ρ *such that, for* $n \ge n_0$ *and any* $w \in X_n$,

$$\|(\mathcal{A} + \mathcal{B}_n)w\|_{s_2} \ge \rho \|w\|_{s_1}, \tag{2.14}$$

and there exists a positive constant c *such that, for all* $n \in \mathbb{N}$,

$$\|u - u_n\|_{s_1} \le c\|u - \mathcal{P}_n u\|_{s_1}.$$

Moreover, if $u \in X \cap H^{s_1+t}(I)$, *then there exists a positive constant* c *such that*

$$\|u - u_n\|_{s_1} \le c n^{-t} \|u\|_{s_1+t}.$$

To close this section, we make remarks for the regularity of the solution of the boundary integral equations (2.1) and the corresponding convergence order of the Fourier–Galerkin method for the equations if operator $\mathcal{A}$ is one of the three singular integral operators described earlier.

In the case when operator $\mathcal{A}$ is the weakly singular integral operator defined by the kernel (2.4), we have $s_1 = 0$ and $s_2 = 1$. By following [1, 11], if $b(\cdot, y) \in H^{t+1}(I)$ for any $y \in I$ and $g \in H^{t+1}(I)$, then $u \in H^t(I)$, and, thus, there exists a positive constant c such that

$$\|u - u_n\|_0 \le c n^{-t} \|u\|_t. \tag{2.15}$$

In the case when operator $\mathcal{A}$ is the strongly singular integral operator defined by the kernel (2.7), we have $s_1 = 0$ and $s_2 = 0$. By following [1, 15], we conclude that if $b(\cdot, y) \in H^t(I)$ for any $y \in I$ and $g \in H_0^0(I) \cap H^t(I)$, then $u \in H_0^0(I) \cap H^t(I)$, and, thus, there exists a positive constant c such that

$$\|u - u_n\|_0 \le cn^{-t}\|u\|_t. \tag{2.16}$$

Finally, in the case when operator $\mathcal{A}$ is the hypersingular integral operator defined by the kernel (2.8), we have $s_1 = 1$ and $s_2 = 0$. By following [1, 4], if $b(\cdot, y) \in H^t(I)$ for any $y \in I$ and $g \in H_0^0(I) \cap H^t(I)$, then $u \in H_0^0(I) \cap H^{t+1}(I)$. Thus, in this case, there exists a positive constant c such that

$$\|u - u_n\|_1 \le cn^{-t}\|u\|_{t+1}. \tag{2.17}$$

3. A matrix truncation strategy. In this section, we introduce a truncation strategy for the Galerkin matrix of operator $\mathcal{A} + \mathcal{B}$.

We first present the Galerkin matrix. The Fourier–Galerkin method (2.13) is to seek a function $u_n \in X_n$ in the form

$$u_n(x) := \sum_{|k| \in \mathbb{Z}_n^*} a_k e_k(x), \quad x \in I,$$

that satisfies the equation

$$((\mathcal{A} + \mathcal{B})u_n, e_k)_{s_2} = (g, e_k)_{s_2}, \quad |k| \in \mathbb{Z}_n^*. \tag{3.1}$$

We next write a matrix form of (3.1). For $|k|, |l| \in \mathbb{Z}_n$, we define

$$a_{k,l} := \int_I \int_I a(x, y) e_k(x) e_l(y) dx dy$$

and introduce 2×2 matrices

$$A_{k,l} := \begin{bmatrix} a_{-k,l} & a_{-k,-l} \\ a_{k,l} & a_{k,-l} \end{bmatrix}.$$

Throughout this paper, we write $\mathbf{v} = [v_k, v_{-k} : k \in \mathbb{Z}_n^+]$ to denote the vector $\mathbf{v} = [v_1, v_{-1}, \ldots, v_n, v_{-n}]$ and write $\mathbf{v} = [v_k, v_{-k} : k \in \mathbb{Z}_n]$ to denote the vector $\mathbf{v} = [v_0, v_1, v_{-1}, \ldots, v_n, v_{-n}]$. We also write $\mathbf{v}^- = [v_{-k}, v_k : k \in \mathbb{Z}_n^*]$. With this notation, we then define matrices

$$\begin{aligned} &A_1 := [A_{k,l} : k, l \in \mathbb{Z}_n^+], \quad A' := [a_{0,0}], \\ &A'' := [a_{0,k}, a_{0,-k} : k \in \mathbb{Z}_n^+], \quad A''' := [a_{-k,0}, a_{k,0} : k \in \mathbb{Z}_n^+]^T. \end{aligned}$$

We let

$$\mathbf{A}_n := A_1 \ \text{ if } \ \mathbb{Z}_n^* := \mathbb{Z}_n^+ \ \text{ and } \ \mathbf{A}_n := \begin{bmatrix} A' & A'' \\ A''' & A_1 \end{bmatrix} \ \text{ if } \ \mathbb{Z}_n^* := \mathbb{Z}_n.$$

By replacing the kernel a by the kernel b, we likewise define the matrix $\mathbf{B}_n$ and its corresponding blocks. For $|k| \in \mathbb{Z}_n$, we let

$$\widetilde{g}_k := \int_I g(x) e_k(x) dx, \quad \mathbf{g}^- := [\widetilde{g}_{-k}, \widetilde{g}_k : k \in \mathbb{Z}_n^*]^T.$$

Associated with u_n, the solution of (2.13), we introduce the vector $\mathbf{u}_n := [a_k, a_{-k} : k \in \mathbb{Z}_n^*]^T$. By using these matrix and vector notations, we have the following matrix form of (3.1):

$$(\mathbf{A}_n + \mathbf{B}_n)\mathbf{u}_n = \mathbf{g}^-. \tag{3.2}$$

In general, the Galerkin matrix $\mathbf{A}_n + \mathbf{B}_n$ is a dense matrix. When n is large, it requires large computational effort to set up the matrix and to solve the corresponding linear system subsequently. We propose a truncation strategy which allows us to approximate the dense coefficient matrix by a sparse matrix and preserves the optimal convergence order of the conventional Fourier–Galerkin method for the truncated method.

We first consider the matrix $\mathbf{A}_n$. Under the hypothesis that, for each $k \in \mathbb{Z}$, e_k is an eigenfunction of the operator $\mathcal{A}$, matrix $\mathbf{A}_n$ is diagonal. This is the case when the operator $\mathcal{A}$ is one of the three singular boundary integral operators described in the last section. In fact, the matrices $\mathbf{A}_n$ of the singular boundary integral operator $\mathcal{A}$ corresponding to the kernel a having the form (2.4), (2.7), and (2.8) are given, respectively, by

$$\mathbf{A}_n := \operatorname{diag}\left(1, 1, 1, \frac{1}{2}, \frac{1}{2}, \ldots, \frac{1}{n}, \frac{1}{n}\right), \quad \mathbf{A}_n := \operatorname{diag}\left(\pi i, -\pi i, \pi i, -\pi i, \ldots, \pi i, -\pi i\right), \quad \text{and}$$

$$\mathbf{A}_n := \operatorname{diag}\left(1, -1, 2, -2, \ldots, n, -n\right).$$

The matrix $\mathbf{B}_n$, which is usually dense, requires a truncation. We now describe the truncation strategy. For each $n \in \mathbb{N}$, the truncation strategy uses the *hyperbolic* curve $kl = n$. We keep only the blocks $B_{k,l}$ for the pairs (k, l) satisfying $kl \le n$ and replace all of the blocks $B_{k,l}$ by the 2×2 zero matrix $0_{2\times 2}$ for the pairs (k, l) satisfying $kl > n$. Specifically, for $k, l \ge 1$, we define a block truncation matrix from $\mathbf{B}_n$ by setting

$$\widetilde{B}_{k,l} := \begin{cases} B_{k,l}, & kl \le n, \\ 0_{2\times 2}, & \text{otherwise}, \end{cases} \quad \text{and} \quad \widetilde{B}_1 := [\widetilde{B}_{k,l} : k, l \in \mathbb{Z}_n^+] \tag{3.3}$$

and letting

$$\widetilde{\mathbf{B}}_n := \widetilde{B}_1 \ \text{ if } \ \mathbb{Z}_n^* := \mathbb{Z}_n^+ \ \text{ and } \ \widetilde{\mathbf{B}}_n := \begin{bmatrix} B' & B'' \\ B''' & \widetilde{B}_1 \end{bmatrix} \ \text{ if } \ \mathbb{Z}_n^* := \mathbb{Z}_n.$$

By replacing the dense matrix $\mathbf{B}_n$ in (3.2) by the sparse matrix $\widetilde{\mathbf{B}}_n$, we obtain the truncated linear system

$$(\mathbf{A}_n + \widetilde{\mathbf{B}}_n)\widetilde{\mathbf{u}}_n = \mathbf{g}^-, \tag{3.4}$$

where $\widetilde{\mathbf{u}}_n := [\widetilde{a}_k, \widetilde{a}_{-k} : k \in \mathbb{Z}_n^*]^T$. We call this numerical method the truncated Fourier–Galerkin method. Now that the coefficient matrix of the truncated linear system is sparse, it is a fast method. Issues regarding the truncated Fourier–Galerkin method include: (1) How sparse is the matrix? (2) Does the truncated linear system have a unique solution? (3) How does the solution of (3.4) approximate the exact solution of the boundary integral equation? We answer the first question at the end of this section and postpone the other two questions to the next two sections.

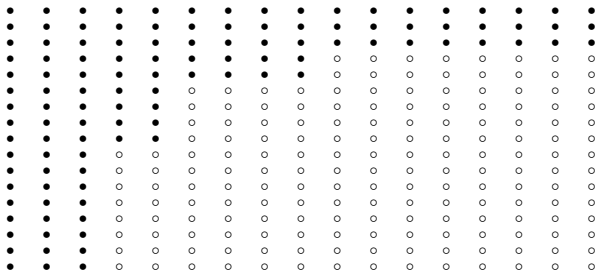


FIG. 1. *Sparse matrix* $\widetilde{\mathbf{B}}_8$ *when* $\mathbb{Z}_8^* = \mathbb{Z}_8$.

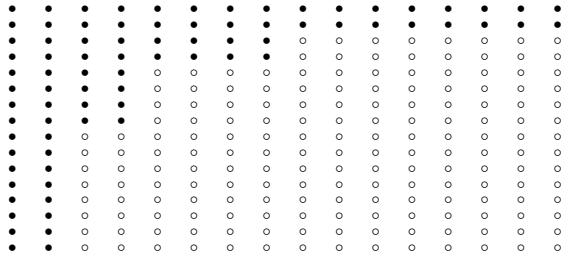


FIG. 2. *Sparse matrix* $\widetilde{\mathbf{B}}_8$ *when* $\mathbb{Z}_8^* = \mathbb{Z}_8^+$.

In the next theorem we estimate precisely the number of nonzero entries in matrix $\widetilde{\mathbf{B}}_n$. To this end, we use $\mathcal{N}(\mathbf{G})$ to denote the number of nonzero entries in matrix $\mathbf{G}$ and $\lfloor x \rfloor$ to denote the largest integer not greater than x.

THEOREM 3.1. *There exists a positive constant* c *such that, for all* $n \in \mathbb{N}$,

$$\mathcal{N}(\widetilde{\mathbf{B}}_n) \leq cn \log n.$$

Proof. For $k, l \geq 1$, by using truncation strategy (3.3) and the fact that $\mathcal{N}(\widetilde{B}_{kl}) \leq 4$, we have

$$\mathcal{N}(\widetilde{B}_1) = \sum_{k \in \mathbb{Z}_n^+} \sum_{l \in \mathbb{Z}_n^+} \mathcal{N}(\widetilde{B}_{k,l}) \leq 4 \sum_{k \in \mathbb{Z}_n^+} \left\lfloor \frac{n}{k} \right\rfloor \leq 4n \sum_{k \in \mathbb{Z}_n^+} \frac{1}{k}.$$

Employing the equation

$$\sum_{k \in \mathbb{Z}_n} \frac{1}{k} = \mathcal{O}(\log n)$$

in the last inequality, we conclude that $\mathcal{N}(\widetilde{B}_1) \leq 4n \log n$. By noting that

$$\mathcal{N}(B') \leq 1, \quad \mathcal{N}(B'') \leq 2n, \quad \mathcal{N}(B''') \leq 2n,$$

we have the desired estimate. ∎

We call this truncation strategy a Γ-type truncation strategy. This truncation strategy admits a fast algorithm to generate the coefficient matrix of linear system (3.4) and to solve it subsequently.

The Γ-type truncation strategy is illustrated in Figures 1 and 2, where the dots indicate the entries kept and the circles indicate the entry replaced by zero. Figure 1 illustrates the truncation strategy for the weakly singular boundary integral equation and Figure 2 for the strongly singular or hypersingular boundary integral equation.

4. Stability analysis. We analyze in this section the stability of the fast algorithm presented in the last section.

We first present four technical lemmas necessary for the stability analysis. For a vector $\mathbf{d} := [d_k : k \in \mathbb{Z}_n^+]^T$, we let $\|\mathbf{d}\|$ denote its spectral norm, and, for a nonnegative number μ, we define a weighted vector $\mathbf{d}_\mu := [(1+k^2)^{\frac{\mu}{2}} d_k : k \in \mathbb{Z}_n^+]^T$. We set

$$\mathbb{J}_n := \left\{(k,l) : l = \left\lceil \frac{n}{k} \right\rceil, \left\lceil \frac{n}{k} \right\rceil + 1, \ldots, n, k \in \mathbb{Z}_n^+\right\},$$

where $\lceil x \rceil$ denotes the smallest integer not less than x.

LEMMA 4.1. *If a sequence $f_{k,l} \in \mathbb{C}$, $k,l \in \mathbb{Z}_n^+$, satisfies the condition that there exists a positive constant θ such that, for all $n \in \mathbb{N}$,*

$$\sum_{k \in \mathbb{Z}_n^+} \sum_{l \in \mathbb{Z}_n^+} (1+k^2)^{\mu+\nu}(1+l^2)^{\mu+\nu}|f_{k,l}|^2 \le \theta^2, \tag{4.1}$$

for some nonnegative constants μ and ν, then, for two arbitrary vectors $\mathbf{d} := [d_k : k \in \mathbb{Z}_n^+]^T$ and $\mathbf{h} := [h_k : k \in \mathbb{Z}_n^+]^T$,

$$\sum_{(k,l) \in \mathbb{J}_n} |(1+k^2)^{\mu} d_k h_l f_{k,l}| \le \theta \|\mathbf{d}_\mu\| \|\mathbf{h}\| n^{-\nu}. \tag{4.2}$$

Proof. Associated with the sequence $f_{k,l}$, for $k,l \in \mathbb{Z}_n^+$, we define

$$g_{k,l} := (1+k^2)^{\frac{\mu+\nu}{2}}(1+l^2)^{\frac{\mu+\nu}{2}}|f_{k,l}|,$$

and, associated with the vectors $\mathbf{d}$ and $\mathbf{h}$, we define

$$h_{k,l} := (1+k^2)^{\frac{\mu-\nu}{2}}(1+l^2)^{-\frac{\mu+\nu}{2}}|d_k||h_l|.$$

We then let S denote the inner product of these two sequences on the index set $\mathbb{J}_n$; that is,

$$S := \sum_{(k,l) \in \mathbb{J}_n} h_{k,l} g_{k,l}.$$

Applying the Cauchy–Schwartz inequality to the inner product S yields

$$S \le \left(\sum_{(k,l) \in \mathbb{J}_n} h_{k,l}^2 \right)^{\frac{1}{2}} \left(\sum_{(k,l) \in \mathbb{J}_n} g_{k,l}^2 \right)^{\frac{1}{2}}. \tag{4.3}$$

By using hypothesis (4.1), we obtain

$$\sum_{(k,l) \in \mathbb{J}_n} g_{k,l}^2 \le \theta^2. \tag{4.4}$$

Substituting (4.4) and the inequality

$$\sum_{(k,l) \in \mathbb{J}_n} h_{k,l}^2 \le \|\mathbf{d}_\mu\|^2 \|\mathbf{h}\|^2 n^{-2\nu}$$

into the right-hand side of (4.3) yields the estimate (4.2). ∎

In the next lemma, we estimate the difference between two matrices $\mathbf{B}_n$ and $\widetilde{\mathbf{B}}_n$. To this end, for $\mathbf{v}^- := [v_{-k}, v_k : k \in \mathbb{Z}_n^+]^T \in \mathbb{C}^{2n}$ and a nonnegative number μ, we define $\mathbf{v}_\mu^- := [(1+k^2)^{\frac{\mu}{2}} v_{-k}, (1+k^2)^{\frac{\mu}{2}} v_k : k \in \mathbb{Z}_n^+]^T$.

LEMMA 4.2. *If there exists a positive constant β such that, for all n,*

$$\sum_{k\in\mathbb{Z}_n^+}\sum_{l\in\mathbb{Z}_n^+}(1+k^2)^{\mu+\nu}(1+l^2)^{\mu+\nu}|b_{k,l}|^2 \le \beta^2, \tag{4.5}$$

for some nonnegative constants μ and ν, then, for two arbitrary vectors $\mathbf{v}^-, \mathbf{w} \in \mathbb{C}^{2n}$,

$$|{\mathbf{v}_{2\mu}^-}^T(\mathbf{B}_n - \widetilde{\mathbf{B}}_n)\mathbf{w}| \le 4\beta\|\mathbf{v}_\mu^-\|\|\mathbf{w}\|n^{-\nu}. \tag{4.6}$$

Proof. According to the truncation strategy (3.3), we write ${\mathbf{v}_{2\mu}^-}^T(\mathbf{B}_n - \widetilde{\mathbf{B}}_n)\mathbf{w}$ in the form

(4.7)

$${\mathbf{v}_{2\mu}^-}^T(\mathbf{B}_n-\widetilde{\mathbf{B}}_n)\mathbf{w} = \sum_{(k,l)\in\mathbb{J}_n}(1+k^2)^{\mu}(v_k w_l b_{-k,l}+v_{-k}w_l b_{-k,-l}+v_k w_{-l}b_{k,l}+v_{-k}w_{-l}b_{k,-l}).$$

Upon introducing four summations

$$I_1 := \sum_{(k,l)\in\mathbb{J}_n}(1+k^2)^{\mu}v_k w_l b_{-k,l}, \quad I_2 := \sum_{(k,l)\in\mathbb{J}_n}(1+k^2)^{\mu}v_{-k}w_l b_{-k,-l},$$
$$I_3 := \sum_{(k,l)\in\mathbb{J}_n}(1+k^2)^{\mu}v_k w_{-l} b_{k,l}, \quad I_4 := \sum_{(k,l)\in\mathbb{J}_n}(1+k^2)^{\mu}v_{-k}w_{-l}b_{k,-l},$$

we rewrite (4.7) as

$${\mathbf{v}_{2\mu}^-}^T(\mathbf{B}_n - \widetilde{\mathbf{B}}_n)\mathbf{w} = I_1 + I_2 + I_3 + I_4. \tag{4.8}$$

It suffices to estimate I_k for $k = 1, 2, 3, 4$. We first estimate I_1. In Lemma 4.1, we choose $f_{k,l} := b_{-k,l}$, $\mathbf{d}_\mu := [(1+k^2)^{\frac{\mu}{2}} v_k : k \in \mathbb{Z}_n^+]^T$, and $\mathbf{h} := [w_k : k \in \mathbb{Z}_n^+]^T$. Clearly, the sequence $f_{k,l}$ satisfies condition (4.1), and vectors $\mathbf{d}_\mu$ and $\mathbf{h}$ have the bounds

$$\|\mathbf{d}_\mu\| \le \|\mathbf{v}_\mu^-\| \quad \text{and} \quad \|\mathbf{h}\| \le \|\mathbf{w}\|,$$

respectively. By Lemma 4.1 with the above bounds, we obtain the bound for I_1 given by

$$|I_1| \le \beta\|\mathbf{v}_\mu^-\|\|\mathbf{w}\|n^{-\nu}. \tag{4.9}$$

In a similar manner, we obtain the following estimates for I_k:

$$|I_k| \le \beta\|\mathbf{v}_\mu^-\|\|\mathbf{w}\|n^{-\nu}, \quad k = 2, 3, 4. \tag{4.10}$$

Substituting estimates (4.9) and (4.10) into (4.8) yields the desired estimate. ∎

Let $\widetilde{\mathcal{B}}_n$ be the linear operator such that its matrix representation under the basis $\{e_k : |k| \in \mathbb{Z}_n^*\}$ is $\widetilde{\mathbf{B}}_n$. The next lemma translates the estimate in the last lemma to an estimate of the difference $\mathcal{B} - \mathcal{B}_n$.

LEMMA 4.3. *If there exists a positive constant β such that, for all n, (4.5) with $\mu = s_2$ and $\nu = t > 0$ holds, then there exists a positive integer n_0 such that, for all $n \ge n_0$ and for each $w \in X$,*

$$\|(\mathcal{B}_n - \widetilde{\mathcal{B}}_n)\mathcal{P}_n w\|_{s_2} \le 4\beta\|w\|_0 n^{-t}. \tag{4.11}$$

Proof. By the definition of norm,

$$\begin{aligned}\|(\mathcal{B}_n-\widetilde{\mathcal{B}}_n)\mathcal{P}_n w\|_{s_2} &= \sup_{v\in Y,\|v\|_{s_2}=1}\left|\Big((\mathcal{B}_n-\widetilde{\mathcal{B}}_n)\mathcal{P}_n w, v\Big)_{s_2}\right| \\ &= \sup_{v\in Y,\|v\|_{s_2}=1}\left|\Big((\mathcal{B}_n-\widetilde{\mathcal{B}}_n)\mathcal{P}_n w, \mathcal{P}_n v\Big)_{s_2}\right|.\end{aligned} \tag{4.12}$$

For $v\in Y$ and $w\in X$, we express their projection onto X_n as

$$(\mathcal{P}_n v)(x)=\sum_{|k|\in\mathbb{Z}_n^*} v_k e_{-k}(x),\quad (\mathcal{P}_n w)(x)=\sum_{|k|\in\mathbb{Z}_n^*} w_k e_k(x),\quad x\in I. \tag{4.13}$$

Two coefficient vectors in (4.13) are denoted by $\mathbf{v}^- := [v_{-k}, v_k : k\in\mathbb{Z}_n^+]^T$ and $\mathbf{w} := [w_k, w_{-k} : k\in\mathbb{Z}_n^+]^T$, respectively. It can be verified that

$$\|\mathbf{v}^-_{s_2}\|\le\|v\|_{s_2},\quad \|\mathbf{w}\|\le\|w\|_0. \tag{4.14}$$

From (4.13) and the definition of the operators $\mathcal{B}_n$ and $\widetilde{\mathcal{B}}_n$, we obtain

$$((\mathcal{B}_n-\widetilde{\mathcal{B}}_n)\mathcal{P}_n w,\mathcal{P}_n v)_{s_2} = {\mathbf{v}^-_{2s_2}}^T(\mathbf{B}_n-\widetilde{\mathbf{B}}_n)\mathbf{w}. \tag{4.15}$$

By applying Lemma 4.2 to the right-hand side of (4.15) with the estimate (4.14), we have

$$\left|((\mathcal{B}_n-\widetilde{\mathcal{B}}_n)\mathcal{P}_n w,\mathcal{P}_n v)_{s_2}\right|\le 4\beta\|v\|_{s_2}\|w\|_0 n^{-t}. \tag{4.16}$$

Substituting the above estimate into (4.12) yields the desired estimate of this lemma. ∎

To continue the stability analysis, we need a regularity hypothesis on kernel b. Let $L^2(I^2)$ denote the Hilbert space with inner product defined by

$$(\phi,\varphi) := \int_I\int_I \phi(x,y)\overline{\varphi(x,y)}dxdy.$$

For $\mu,\nu\ge 0$, let $H^{\mu,\nu}(I^2)$ denote the space of functions $\phi\in L^2(I^2)$ with the property

$$\sum_{k\in\mathbb{Z}}\sum_{l\in\mathbb{Z}}(1+k^2)^\mu(1+l^2)^\nu|\phi_{k,l}|^2<+\infty,$$

where $\phi_{k,l}$ is the Fourier coefficient of ϕ defined by

$$\phi_{k,l} := \int_I\int_I \phi(x,y)e_k(x)e_l(y)dxdy,\quad k,l\in\mathbb{Z}.$$

The space $H^{\mu,\nu}(I^2)$ is a Hilbert space with norm

$$\|\phi\|_{\mu,\nu} := \left\{\sum_{k,l\in\mathbb{Z}}(1+k^2)^\mu(1+l^2)^\nu|\phi_{k,l}|^2\right\}^{1/2};$$

see, for example, [16, 17]. By using integration by parts, we can verify that if the kernel $b(x,y)$ has 2π-periodic and continuous $\lceil s_2+t\rceil$th partial derivatives with respect to both variables x and y, then $b\in H^{s_2+t,s_2+t}(I^2)$.

LEMMA 4.4. *If $b \in H^{s_2+t,s_2+t}(I^2)$ for $t > 0$, then there exists a positive integer n_0 such that, for all $n \geq n_0$ and for each $w \in X$,*

$$\|(\mathcal{B}_n - \widetilde{\mathcal{B}}_n)\mathcal{P}_n w\|_{s_2} \leq 4\|b\|_{s_2+t,s_2+t}\|w\|_0 n^{-t}. \tag{4.17}$$

Proof. Because $b \in H^{s_2+t,s_2+t}(I^2)$, by the definition of the norm,

$$\sum_{k\in\mathbb{Z}_n^+}\sum_{l\in\mathbb{Z}_n^+}(1+k^2)^{s_2+t}(1+l^2)^{s_2+t}|b_{k,l}|^2 = \|b\|^2_{s_2+t,s_2+t} < +\infty. \tag{4.18}$$

Hence, estimate (4.17) follows directly from Lemma 4.3 with (4.18). ∎

We are now ready to present the stability result.

THEOREM 4.5. *If $b \in H^{s_2+t,s_2+t}(I^2)$ for $t > 0$, then there exists a positive integer n_0 such that, for all $n \geq n_0$ and for all $w \in X_n$,*

$$\|(\mathcal{A} + \widetilde{\mathcal{B}}_n)w\|_{s_2} \geq \frac{\rho}{2}\|w\|_{s_1}, \tag{4.19}$$

where $\rho > 0$ is the constant that appears in Theorem 2.1.

Proof. The proof is done by employing the stability of the Fourier–Galerkin method (Theorem 2.1) and Lemma 4.4.

Since $b \in H^{s_2+t,s_2+t}(I^2)$, we have $\|b\|^2_{s_2+t,s_2+t} < +\infty$. By Lemma 4.4 with the fact that $n^{-t} \to 0$ as $n \to \infty$ and the inequality $\|w\|_0 \leq \|w\|_{s_1}$, we conclude that there exists a positive integer n_0 such that, for all $n \geq n_0$ and for all $w \in X_n$,

$$\|(\mathcal{B}_n - \widetilde{\mathcal{B}}_n)w\|_{s_2} \leq \frac{\rho}{2}\|w\|_{s_1}. \tag{4.20}$$

It follows from (2.14) and (4.20) that

$$\|(\mathcal{A} + \widetilde{\mathcal{B}}_n)w\|_{s_2} \geq \|(\mathcal{A} + \mathcal{B}_n)w\|_{s_2} - \|(\mathcal{B}_n - \widetilde{\mathcal{B}}_n)w\|_{s_2} \geq \frac{\rho}{2}\|w\|_{s_1},$$

proving the stability estimate (4.19). ∎

Theorem 4.5 ensures that for $n \geq n_0$ the equation

$$(\mathcal{A} + \widetilde{\mathcal{B}}_n)\widetilde{u}_n = g_n \tag{4.21}$$

has a unique solution

$$\widetilde{u}_n(x) := \sum_{|k|\in\mathbb{Z}_n^*} \widetilde{a}_k e_k(x), \quad x \in I. \tag{4.22}$$

In other words, the truncated linear system (3.4), which is equivalent to (4.21), has a unique solution $\widetilde{\mathbf{u}}_n$ for $n \geq n_0$.

5. Convergence analysis. Now we turn to considering the convergence order of the solution $\widetilde{u}_n$ of (4.21). We will show that the truncation strategy will not ruin the convergence order of the conventional Fourier–Galerkin method.

THEOREM 5.1. *If $b \in H^{s_2+t,s_2+t}(I^2)$ and $g \in H^{s_2+t}(I)$ for $t > 0$, then there exist a positive constant c and a positive integer n_0 such that, for all $n \geq n_0$,*

$$\|u - \widetilde{u}_n\|_{s_1} \leq cn^{-t}\|u\|_{s_1+t}.$$

Proof. By our general hypothesis, if $b \in H^{s_2+t,s_2+t}(I^2)$, then $u \in X \cap H^{s_1+t}(I)$. By applying the triangle inequality, we obtain

$$\|u - \widetilde{u}_n\|_{s_1} \le \|u - \mathcal{P}_n u\|_{s_1} + \|\mathcal{P}_n u - \widetilde{u}_n\|_{s_1}. \tag{5.1}$$

We first study the second term in the right-hand side of (5.1). Since $b \in H^{s_2+t,s_2+t}(I^2)$, we can employ the stability theorem established in the last section. By Theorem 4.5, there exist a positive constant c and a positive integer n_0 such that, for all $n \ge n_0$,

$$\|\mathcal{P}_n u - \widetilde{u}_n\|_{s_1} \le c\|(\mathcal{I} + \mathcal{A}^{-1}\widetilde{\mathcal{B}}_n)(\mathcal{P}_n u - \widetilde{u}_n)\|_{s_1}. \tag{5.2}$$

On the other hand, by applying the operator $\mathcal{A}^{-1}\mathcal{P}_n$ to both sides of (2.1) with a help of (2.12), we observe that

$$\mathcal{P}_n(\mathcal{I} + \mathcal{A}^{-1}\mathcal{B})u = \mathcal{A}^{-1} g_n. \tag{5.3}$$

Substituting (4.21) into the right-hand side of (5.3) yields

$$\mathcal{P}_n(\mathcal{I} + \mathcal{A}^{-1}\mathcal{B})u = (\mathcal{I} + \mathcal{A}^{-1}\widetilde{\mathcal{B}}_n)\widetilde{u}_n. \tag{5.4}$$

A direct computation using (2.12) and (5.4) confirms that there holds the equation

$$(\mathcal{I} + \mathcal{A}^{-1}\widetilde{\mathcal{B}}_n)(\mathcal{P}_n u - \widetilde{u}_n) = \mathcal{P}_n(\mathcal{I} + \mathcal{A}^{-1}\mathcal{B})(\mathcal{P}_n u - u) + \mathcal{A}^{-1}(\widetilde{\mathcal{B}}_n - \mathcal{B}_n)\mathcal{P}_n u. \tag{5.5}$$

Consequently, we have

$$\|(\mathcal{I} + \mathcal{A}^{-1}\widetilde{\mathcal{B}}_n)(\mathcal{P}_n u - \widetilde{u}_n)\|_{s_1} \le c\|\mathcal{P}_n u - u\|_{s_1} + c\|(\widetilde{\mathcal{B}}_n - \mathcal{B}_n)\mathcal{P}_n u\|_{s_2}. \tag{5.6}$$

By substituting (5.6) into the right-hand side of (5.2) and then substituting the resulting inequality into (5.1), we obtain the inequality

$$\|u - \widetilde{u}_n\|_{s_1} \le c\|\mathcal{P}_n u - u\|_{s_1} + c\|(\widetilde{\mathcal{B}}_n - \mathcal{B}_n)\mathcal{P}_n u\|_{s_2}. \tag{5.7}$$

Since $u \in X \cap H^{s_1+t}(I)$, by estimate (2.9) with $\nu = s_1$ and $\mu = s_1 + t$, there exist a positive constant c and a positive integer n_0 such that, for all $n \ge n_0$,

$$\|u - \mathcal{P}_n u\|_{s_1} \le cn^{-t}\|u\|_{s_1+t}. \tag{5.8}$$

By using Lemma 4.4 and the inequality $\|u\|_0 \le \|u\|_{s_1+t}$, the second term of (5.7) is also bounded by the same quantity. This proves the desired estimate of this theorem. ∎

Theorem 5.1 confirms that the truncated Fourier–Galerkin method gives the optimal order of convergence.

When applying the fast Fourier–Galerkin method to solve the three types of singular boundary integral equations described in section 2, by using Theorem 5.1 and the regularity of the solutions of these equations, we have the following three corollaries regarding their convergence order.

COROLLARY 5.2. *If operator $\mathcal{A}$ is the weakly singular integral operator defined by the kernel* (2.4) *and if $b \in H^{t+1,t+1}(I^2)$ and $g \in H^{t+1}(I)$, then there exist a positive constant c and a positive integer n_0 such that, for all $n \ge n_0$,*

$$\|u - \widetilde{u}_n\|_0 \le cn^{-t}\|u\|_t.$$

COROLLARY 5.3. *If operator $\mathcal{A}$ is the strongly singular integral operator defined by the kernel* (2.7) *and if $b \in H^{t,t}(I^2)$ and $g \in H_0^0(I) \cap H^t(I)$, then there exist a positive constant c and a positive integer n_0 such that, for all $n \geq n_0$,*

$$\|u - \widetilde{u}_n\|_0 \leq cn^{-t}\|u\|_t.$$

COROLLARY 5.4. *If operator $\mathcal{A}$ is the hypersingular integral operator defined by the kernel* (2.8) *and if $b \in H^{t,t}(I^2)$ and $g \in H_0^0(I) \cap H^t(I)$, then there exist a positive constant c and a positive integer n_0 such that, for all $n \geq n_0$,*

$$\|u - \widetilde{u}_n\|_1 \leq cn^{-t}\|u\|_{t+1}.$$

6. A fast solver for the truncated linear system. In this section we develop a fast solver for the truncated linear system by taking advantage of the structure of the truncated matrix $\widetilde{\mathbf{B}}_n$. The idea of this algorithm comes from the multilevel augmentation method presented in [6, 7] (see also [5, 8, 9]). We consider only the case $\mathbb{Z}_n^* = \mathbb{Z}_n^+$ since the other case can be similarly handled.

To present the fast algorithm, we assume that $n := 2^{m+q}$, where both m and q are positive integers with m *fixed*. The integer m here will be used to denote the initial level of approximation. We write $\widetilde{\mathbf{B}}_n$ in a block form

$$\widetilde{\mathbf{B}}_n := \begin{bmatrix} \overline{\mathbf{B}}_{0,0} & \overline{\mathbf{B}}_{0,1} & \cdots & \overline{\mathbf{B}}_{0,q} \\ \overline{\mathbf{B}}_{1,0} & \overline{\mathbf{B}}_{1,1} & \cdots & \overline{\mathbf{B}}_{1,q} \\ \vdots & \vdots & & \vdots \\ \overline{\mathbf{B}}_{q,0} & \overline{\mathbf{B}}_{q,1} & \cdots & \overline{\mathbf{B}}_{q,q} \end{bmatrix}, \tag{6.1}$$

where $\overline{\mathbf{B}}_{0,0}$ has order $2^{m+1} \times 2^{m+1}$, $\overline{\mathbf{B}}_{0,k}$ has order $2^{m+1} \times 2^{m+k}$, $\overline{\mathbf{B}}_{k,0}$ has order $2^{m+k} \times 2^{m+1}$ for $k = 1, 2, \ldots, q$, and $\overline{\mathbf{B}}_{j,l}$ has order $2^{m+j} \times 2^{m+l}$ for $j, l = 1, 2, \ldots, q$. For each $k = 0, 1, 2, \ldots, q$, we define a submatrix $\mathbf{G}_{m,k}$ of $\widetilde{\mathbf{B}}_n$ by replacing the subscript q in the right-hand side matrix of (6.1) with k. Clearly, $\mathbf{G}_{m,0} := \overline{\mathbf{B}}_{0,0}$, $\mathbf{G}_{m,q} := \widetilde{\mathbf{B}}_n$, and $\mathbf{G}_{m,k}$ is obtained from $\mathbf{G}_{m,k-1}$ by adding new rows and columns. We split $\mathbf{G}_{m,k}$ into two matrices

$$\mathbf{C}_{m,k} := \begin{bmatrix} \overline{\mathbf{B}}_{0,0} & \overline{\mathbf{B}}_{0,1} & \cdots & \overline{\mathbf{B}}_{0,k} \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix} \quad \text{and} \quad \mathbf{D}_{m,k} := \begin{bmatrix} 0 & 0 & \cdots & 0 \\ \overline{\mathbf{B}}_{1,0} & \overline{\mathbf{B}}_{1,1} & \cdots & \overline{\mathbf{B}}_{1,k} \\ \vdots & \vdots & & \vdots \\ \overline{\mathbf{B}}_{k,0} & \overline{\mathbf{B}}_{k,1} & \cdots & \overline{\mathbf{B}}_{k,k} \end{bmatrix}.$$

Matrix $\mathbf{C}_{m,k}$ consists of the "low frequency" component of the matrix $\mathbf{G}_{m,k}$, and matrix $\mathbf{D}_{m,k}$ consists of its "high frequency" component. By using this splitting, (2.13) with $n = 2^{m+k}$ becomes

$$(\mathbf{A}_{2^{m+k}} + \mathbf{C}_{m,k})\mathbf{u}_{2^{m+k}} = g^-_{2^{m+k}} - \mathbf{D}_{m,k}\mathbf{u}_{2^{m+k}} \tag{6.2}$$

for $k = 0, 1, \ldots, q$. Note that the matrices $\mathbf{A}_{2^{m+k}} + \mathbf{C}_{m,k}$ have a very simple structure: $\mathbf{A}_{2^{m+k}}$ are diagonal, and all blocks of $\mathbf{C}_{m,k}$ except those in the first row are zero. The computational cost for inverting these matrices is comparable to that for inverting the matrix $\mathbf{A}_{2^m} + \mathbf{C}_{m,0}$. Equation (6.2) leads to the following fast algorithm.

ALGORITHM 6.1. *Let $m > 0$ be a fixed positive integer.*

Step 1*: Solve $\mathbf{u}^*_{2^m}$ from an equation of the initial step*

$$(\mathbf{A}_{2^m} + \mathbf{C}_{m,0})\mathbf{u}^*_{2^m} = \mathbf{g}^-_{2^m}. \tag{6.3}$$

Step 2: *For $k = 1, 2, \ldots, q$, suppose that $\mathbf{u}^*_{2^{m+k-1}}$ has been obtained, and do the following:*

- *Augment the matrices $\mathbf{C}_{m,k-1}$ and $\mathbf{D}_{m,k-1}$ to form $\mathbf{C}_{m,k}$ and $\mathbf{D}_{m,k}$, respectively.*
- *Augment the vector $\mathbf{u}^*_{2^{m+k-1}}$ by setting $\mathbf{v}^*_{2^{m+k}} = \begin{bmatrix} \mathbf{u}^*_{2^{m+k-1}} \\ \mathbf{0}_{2\times 2} \end{bmatrix}$.*
- *Solve $\mathbf{u}^*_{2^{m+k}}$ from the equation*

$$(\mathbf{A}_{2^{m+k}} + \mathbf{C}_{m,k})\mathbf{u}^*_{2^{m+k}} = \mathbf{g}^-_{2^{m+k}} - \mathbf{D}_{m,k}\mathbf{v}^*_{2^{m+k}}. \tag{6.4}$$

At each step, we need only to solve (6.4). Due to the structure of the matrix $\mathbf{A}_{2^{m+k}} + \mathbf{C}_{m,k}$, the computational effort to invert (6.4) is comparable to that to invert (6.3), which is fixed, independent of n. Hence, the main computational effort for solving the entire problem is concentrated on matrix-vector-multiplication $\mathbf{D}_{m,k}\mathbf{v}^*_{2^{m+k}}$. The computational effort of this algorithm is analyzed in the next theorem.

THEOREM 6.2. *The total number of multiplications required for obtaining $\mathbf{u}^*_{2^{m+q}}$ is $\mathcal{O}(n \log^2 n)$.*

Proof. We assume that the number of multiplications required for computing $\mathbf{u}^*_{2^m}$ is $\mathcal{M}(m)$. From Algorithm 6.1, the total number of multiplications required for obtaining $\mathbf{u}^*_{2^{m+q}}$ is

$$(q+1)\mathcal{M}(m) + \sum_{k=1}^{q} \mathcal{N}(\mathbf{A}_{2^{m+k}} + \mathbf{C}_{m,k} + \mathbf{D}_{m,k}). \tag{6.5}$$

Because the matrix $\mathbf{A}_{2^{m+k}}$ is diagonal, we have

$$\mathcal{N}(\mathbf{A}_{2^{m+k}}) = 2^{m+k+1}. \tag{6.6}$$

Based on the truncation strategy, we conclude that

$$\mathcal{N}(\mathbf{C}_{m,k} + \mathbf{D}_{m,k}) \le 4 \sum_{j=1}^{2^{m+k}} \frac{2^{m+q}}{j} \le 5(2^{m+q})(m+k). \tag{6.7}$$

By using (6.6) and (6.7), we obtain

$$\sum_{k=1}^{q} \mathcal{N}(\mathbf{A}_{2^{m+k}+p_0} + \mathbf{C}_{m,k} + \mathbf{D}_{m,k}) \le c(2^{m+q} + p_0)(m+q)^2.$$

This with (6.5) gives the desired result. ∎

Suppose that the output of Algorithm 6.1 is the vector $\mathbf{u}^*_n := [a^*_k, a^*_{-k} : k \in \mathbb{Z}^*_n]^T$. As in (4.22), we define

$$u^*_n(x) := \sum_{|k| \in \mathbb{Z}^*_n} a^*_k e_k(x), \quad x \in I. \tag{6.8}$$

The function u^*_n can be viewed as a new approximation of u. The next theorem shows that this new approximate solution has the same optimal order of convergence as $\widetilde{u}_n$.

THEOREM 6.3. *If $b \in H^{s_2+t, s_2+t}(I^2)$, then there exist a positive integer m_0 and a positive constant c such that, for all $m \ge m_0$ and all $q \in \mathbb{N}$,*

$$\|u - u^*_n\|_{s_1} \le cn^{-t}\|u\|_{s_1+t},$$

where $n = 2^{m+q}$.

Proof. The convergence order of Algorithm 6.1 is a direct consequence of Theorem 2.2 in [6]. ∎

TABLE 1

q	$N(n)$	C.R.	C.T.	$\|u-u_n^*\|_0$	C.O.
0	129	0.0982	< 0.01	$2.49e-1$	
1	257	0.0546	< 0.01	$1.76e-1$	0.4972
2	513	0.0300	0.015	$1.25e-1$	0.4986
3	1025	0.0164	0.031	$8.83e-2$	0.4993
4	2049	0.0089	0.060	$6.25e-2$	0.4996
5	4097	0.0048	0.123	$4.42e-2$	0.4997

7. Numerical examples. In this section we present numerical examples to illustrate the efficiency and accuracy of the proposed methods. All computer programs are run on a personal computer with a 1.70 GHz celeron CPU and 256 M memory.

In the numerical results reported in this section, the remaining nonzero entries in the truncated Galerkin matrix are computed by using a numerical quadrature scheme based on the fast Fourier transform. Moreover, we use m to denote the initial level in Algorithm 6.1. We denote by $N(n)$ the order of the matrix $\widetilde{\mathbf{B}}_n$. The notations "C.R." and "C.O." stand for the compression rate and convergence order defined, respectively, by

$$\text{C.R.} = \frac{\mathcal{N}(\widetilde{\mathbf{B}}_n)}{\mathcal{N}(\mathbf{B}_n)} \quad \text{and} \quad \text{C.O.} = \log_2 \frac{\|u-u_{2n}^*\|_{s_1}}{\|u-u_n^*\|_{s_1}}.$$

The symbol "C.T." denotes the computing time measured in seconds, used in solving the truncated linear system by Algorithm 6.1.

Example 1. In this example, we consider solving the boundary integral equation (2.1) with the weakly singular kernel a as in (2.4) and a smooth kernel

$$b(x,y) := e^{\sin x \cos 2y} \log|\cos y + 5|, \quad x, y \in I,$$

by using Algorithm 6.1. We choose the right-hand side function g so that $u(x) := x$ is the exact solution of the equation. The integral equation is solved by Algorithm 6.1 with the initial level $m = 6$. Since u is not a periodic function, $u \in H^{0.5-\epsilon}(I)$, with $\epsilon > 0$ being an arbitrary number, and, thus, the theoretical convergence order is $\mathcal{O}(n^{-0.5+\epsilon})$. The numerical results are presented in Table 1.

Example 2. Consider solving the interior Dirichlet problem (2.2) with the domain D bounded by the curve

$$\gamma(x) := (\cos x, 0.4 \sin x), \quad x \in I.$$

We choose the boundary function f so that the exact solution is given by $U(x,y) := e^x \sin y$, $(x,y) \in D$. The associated boundary integral equation is given by (2.1) with the kernel a as in (2.4), a kernel

$$b(x,y) := -\frac{1}{2\pi}\left\{1 + \log\left[\sin^2\frac{x+y}{2} + 0.16\cos^2\frac{x+y}{2}\right]\right\}, \quad x, y \in I,$$

and the right-hand side function

$$g(x) := -\frac{1}{\pi} e^{\cos x} \cos(0.4 \sin x), \quad x \in I.$$

An approximate solution $\widetilde{u}_n$ of the boundary integral equation is solved by Algorithm 6.1 with the initial level $m = 5$. The corresponding approximate solution U_n of the

TABLE 2

q	$N(n)$	δ_1	δ_2	δ_3	δ_4	C.T.
0	65	$3.1115e-04$	$1.3084e-04$	$9.4113e-05$	$3.4259e-04$	< 0.01
1	129	$3.6387e-05$	$2.3389e-06$	$4.8140e-06$	$1.9217e-06$	< 0.01
2	257	$1.0728e-07$	$9.4553e-08$	$8.8520e-08$	$8.6981e-08$	0.014
3	513	$5.4137e-09$	$1.5789e-10$	$6.2348e-09$	$3.2646e-09$	0.030
4	1025	$2.2982e-13$	$4.8406e-14$	$1.7808e-13$	$1.2723e-13$	0.061

TABLE 3

q	$N(n)$	C.R.	C.T.	$\|u-u_n^*\|_0$	C.O.
0	128	0.0840	< 0.01	$7.88e-4$	
1	256	0.0472	< 0.01	$2.80e-4$	1.4915
2	512	0.0263	0.015	$9.94e-5$	1.4958
3	1024	0.0145	0.031	$3.52e-5$	1.4979
4	2048	0.0079	0.063	$1.25e-5$	1.4989
5	4096	0.0043	0.125	$4.40e-6$	1.4992

Dirichlet problem is given by

$$U_n(x,y) := \frac{1}{2}\int_I \widetilde{u}_n(z)\log[(x-\cos z)^2+(y-0.4\sin z)^2]dz, \quad (x,y)\in D.$$

In Table 2, we present the errors of the approximate solution U_n at the points $P_k := \gamma_k(\cos\frac{\pi}{4}, \sin\frac{\pi}{4})$, $k=1,2,3,4$, with $\gamma_1 = 0.2$, $\gamma_2 = 0.4$, $\gamma_3 = 0.6$, and $\gamma_4 = 0.8$, where $\delta_k := |U(P_k) - U_n(P_k)|$, $k=1,2,3,4$.

Example 3. We consider solving the boundary integral equation (2.1) with the strongly singular kernel a as in (2.7) and the smooth kernel

$$b(x,y) := e^{\cos x \sin 2y}\sin x \sin 2y \log|\sin y + 5|, \quad x,y\in I.$$

The right-hand side function g is chosen so that the exact solution $u(x) := \frac{1}{12}(3x^2 - 6\pi x + 2\pi^2)$. It can be seen that $u\in H^{1.5-\epsilon}(I)$, with $\epsilon > 0$ being an arbitrary number. We solve the boundary integral equation by Algorithm 6.1 with the initial level $m = 6$.

In Table 3, we list the numerical results for this example. Note that, since the exact solution $u\in H^{1.5-\epsilon}$, the theoretical convergence order is $\mathcal{O}(n^{-1.5+\epsilon})$, which is confirmed by the computed order of convergence.

Example 4. Consider solving the interior Neumann problem (2.5) with the domain D bounded by the curve by

$$\gamma(x) := (\cos x, 2\sin x), \quad x\in I.$$

We choose the boundary function f so that the exact solution is given by $U(x,y) = e^x \sin y$ for $(x,y)\in D$. The associated boundary integral equation is given by (2.1) with a strongly singular kernel a as in (2.7), a smooth kernel

$$b(x,y) := \frac{3\sin(x+y)}{10+6\cos(x+y)}, \quad x,y\in I,$$

and the right-hand side function

$$g(x) := 2e^{\cos x}\sin(2\sin x)\cos x + e^{\cos x}\cos(2\sin x)\sin x, \quad x\in I.$$

TABLE 4

q	$N(n)$	δ_1	δ_2	δ_3	δ_4	C.T.
0	64	$3.6761e-04$	$4.7679e-04$	$5.1798e-04$	$6.6145e-04$	< 0.01
1	128	$8.7346e-06$	$8.2797e-05$	$8.5745e-05$	$1.7674e-05$	< 0.01
2	256	$8.3246e-08$	$1.1360e-07$	$1.0144e-07$	$9.7936e-08$	0.015
3	512	$5.4328e-12$	$2.5468e-11$	$2.0348e-11$	$5.7835e-11$	0.031
4	1024	$6.2728e-15$	$1.0769e-14$	$2.0650e-14$	$2.0650e-14$	0.062

TABLE 5

q	$N(n)$	C.R.	C.T.	$\|u-u_n^*\|_1$	C.O.
0	128	0.0840	< 0.01	$8.80e-2$	
1	256	0.0472	< 0.01	$6.24e-2$	0.4979
2	512	0.0263	0.014	$4.42e-2$	0.4986
3	1024	0.0145	0.030	$3.12e-2$	0.4993
4	2048	0.0079	0.061	$2.21e-2$	0.4996
5	4096	0.0043	0.123	$1.56e-2$	0.4998

We solve the boundary integral equation by Algorithm 6.1 with the initial level $m = 5$ and obtain an approximate solution $\widetilde{u}_n$ having the form $\widetilde{u}_n = \sum_{|k|\in\mathbb{Z}_n^+} \widetilde{a}_k e^{ikx}$. The corresponding approximate solution of the interior Neumann problem is then given by

$$U_n(x,y) := \int_I \widetilde{\rho}_n(z) K(x,y,z)dz, \quad (x,y) \in D,$$

where

$$\widetilde{\rho}_n(x) := \sum_{|k|\in\mathbb{Z}_n^+} \frac{\operatorname{sign}(k)}{|k|} \widetilde{a}_k e^{ikx}, \quad x \in I,$$

and

$$K(x,y,z) = \frac{-2\cos z(x-\cos z) + \sin z(y-2\sin z)}{(x-\cos z)^2 + (y-2\sin z)^2}, \quad (x,y) \in D, \quad z \in I.$$

In Table 4, we present the error of the approximate solution $\delta_k := |U(P_k) - U_n(P_k)|$, where P_k, $k = 1, 2, 3, 4$, are the same points as those in Example 2.

Example 5. We consider solving the boundary integral equation (2.1) with the hypersingular kernel a as in (2.8) and a smooth kernel

$$b(x,y) := (x-\pi)(y-\pi), \quad x, y \in I.$$

We choose the right-hand side function g so that $u(x) := \frac{1}{12}(3x^2 - 6\pi x + 2\pi^2)$ is the exact solution of the equation. The integral equation is solved by Algorithm 6.1 with the initial level $m = 6$. Since $u \in H^{1.5-\epsilon}(I)$, with $\epsilon > 0$ being an arbitrary number, the theoretical convergence order is $\mathcal{O}(n^{-0.5+\epsilon})$. This is confirmed by the numerical result presented in Table 5.

Example 6. Consider the same interior Neumann problem as in Example 4. But this time we reformulate the Neumann problem as a boundary integral equation (2.1) with the hypersingular kernel a as in (2.8), the smooth kernel

$$b(x,y) := \frac{1}{\pi} \frac{30\cos(x+y) + 18}{(10 + 6\cos(x+y))^2}, \quad x, y \in I,$$

TABLE 6

q	$N(n)$	δ_1	δ_2	δ_3	δ_4	C.T.
0	64	$3.1598e-03$	$5.8998e-03$	$6.8311e-03$	$3.7228e-03$	< 0.01
1	128	$9.4629e-05$	$1.9808e-05$	$9.0284e-06$	$3.9735e-05$	< 0.01
2	256	$7.9961e-07$	$9.9455e-07$	$1.7150e-08$	$2.1955e-08$	0.015
3	512	$3.4572e-11$	$8.4628e-11$	$3.0479e-11$	$4.6393e-11$	0.031
4	1024	$2.6701e-14$	$6.1506e-14$	$4.2855e-14$	$3.8636e-14$	0.062

and the right-hand side function

$$g(x) := -\frac{2}{\pi}e^{\cos x}\sin(2\sin x)\cos x - \frac{1}{\pi}e^{\cos x}\cos(2\sin x)\sin x, \quad x \in I.$$

We solve the boundary integral equation by using Algorithm 6.1 with the initial level $m = 5$ and obtain an approximate solution. A corresponding approximate solution of the interior Neumann problem is then obtained by

$$U_n(x,y) := \int_I \widetilde{u}_n(z)K(x,y,z)dz, \quad (x,y) \in D,$$

where $K(x,y,z)$ is the same kernel as in Example 4. In Table 6, we report the error of the approximate solution at points P_k for $k = 1, 2, 3, 4$.

REFERENCES

[1] K. E. Atkinson, *The Numerical Solution of Integral Equations of Second Kind*, Cambridge University Press, Cambridge, 1997.

[2] K. E. Atkinson, *A discrete Galerkin method for the first kind integral equations with a logarithmic kernel*, J. Integral Equations Appl., 1 (1988), pp. 343–363.

[3] K. E. Atkinson and I. H. Sloan, *The numerical solution of the first kind logarithmic kernel integral equations on smooth open curves*, Math. Comp., 56 (1991), pp. 119–139.

[4] D. Chien and K. E. Atkinson, *A discrete Galerkin method for a hypersingular boundary integral equations*, IMA. J. Numer. Anal., 17 (1997), pp. 463–478.

[5] Z. Chen, C. A. Micchelli, and Y. Xu, *A multilevel method for solving operator equations*, J. Math. Anal. Appl., 262 (2001), pp. 688–699.

[6] Z. Chen, B. Wu, and Y. Xu, *Multilevel augmentation methods for solving operator equations*, Numer. Math. J. Chinese Univ., 14 (2005), pp. 244–251.

[7] Z. Chen, B. Wu, and Y. Xu, *Multilevel augmentation methods for differential equations*, Adv. Comput. Math., 24 (2006), pp. 213–238.

[8] Z. Chen, H. Yang, and Y. Xu, *A multilevel augmentation method for solving ill-posed operator equations*, Inverse Problems, 22 (2006), pp. 155–174.

[9] W. Fang, F. Ma, and Y. Xu, *Multilevel iteration methods for solving integral equations of the second kind*, J. Integral Equations Appl., 14 (2002), pp. 355–376.

[10] G. C. Hsiao and W. L. Wendland, *A finite element method for some integral equations of the first kind*, J. Math. Anal. Appl., 58 (1977), pp. 449–481.

[11] R. Kress, *Linear Integral Equations*, Springer-Verlag, New York, 1989.

[12] R. Kress, *On the numerical solution of a hypersingular integral equation in scattering theory*, J. Comput. Appl. Math., 51 (1995), pp. 345–350.

[13] W. Mclean, *A spectral Galerkin method for a boundary integral equation*, Math. Comp., 47 (1986), pp. 597–607.

[14] W. Mclean, S. Prossdorf, and W. L. Wenland, *A fully-discrete trigonometric collocation method*, J. Integral Equations Appl., 5 (1993), pp. 103–129.

[15] N. I. Muskhelishvili, *Singular Integral Equations*, Noordhoff, Groningen, 1953.

[16] J. Saranen and G. Vainikko, *Two-grid solutions of Symm's integral equation*, Math. Nachr., 177 (1996), pp. 265–279.

[17] J. Saranen and G. Vainikko, *Fast solution of integral and pseudodifferential equations on closed curves*, Math. Comp., 67 (1998), pp. 1473–1491.

[18] I. Sloan, *Error analysis of boundary integral methods*, Acta Numer., 1 (1992), pp. 287–339.
[19] Y. Xu, H.-L. Chen, and Q. Zou, *Limit values of derivatives of the Cauchy integrals and computation of the logarithmic potentials*, Computing, 73 (2004), pp. 295–327.
[20] Y. Xu and Y. Zhao, *Quadratures for boundary integral equations of the first kind with logarithmic kernels*, J. Integral Equations Appl., 8 (1996), pp. 239–268.
[21] Y. Xu and Y. Zhao, *A extrapolation method for a class of boundary integral equations*, Math. Comp., 65 (1996), pp. 587–610.
[22] Y. Yan and I. H. Sloan, *On integral equations of the first kind with logarithmic kernels*, J. Integral Equations Appl., 1 (1988), pp. 549–579.

SIAM J. NUMER. ANAL.
Vol. 46, No. 4, pp. 1985–1997
© 2008 Society for Industrial and Applied Mathematics

CONSTRAINED SPLINE SMOOTHING*

K. KOPOTUN†, D. LEVIATAN‡, AND A. V. PRYMAK§

Abstract. Several results on constrained spline smoothing are obtained. In particular, we establish a general result, showing how one can constructively smooth any monotone or convex piecewise polynomial function (ppf) (or any q-monotone ppf, $q \geq 3$, with one additional degree of smoothness) to be of minimal defect while keeping it close to the original function in the $\mathbb{L}_p$-(quasi)norm. It is well known that approximating a function by ppfs of minimal defect (splines) avoids introduction of artifacts which may be unrelated to the original function; thus it is always preferable. On the other hand, it is usually easier to construct constrained ppfs with as few requirements on smoothness as possible. Our results allow us to obtain shape-preserving splines of minimal defect with equidistant or Chebyshev knots. The validity of the corresponding Jackson-type estimates for shape-preserving spline approximation is summarized; in particular, we show that the $\mathbb{L}_p$-estimates, $p \geq 1$, can be immediately derived from the $\mathbb{L}_\infty$-estimates.

Key words. splines, smoothing, minimal defect, moduli of smoothness, degree of approximation, Jackson-type estimates

AMS subject classifications. 65D07, 65D10, 41A15, 41A29, 41A25, 26A15

DOI. 10.1137/060658746

1. Introduction and the main results.

1.1. Notation. Let $\mathcal{S}_r(\mathbf{z}_n)$ be the space of all piecewise polynomial functions (ppf) of degree r (order $r+1$) with the knots $\mathbf{z}_n := (z_i)_{i=0}^n$, $-1 =: z_0 < z_1 < \cdots < z_{n-1} < z_n := 1$. In other words, we say that $s \in \mathcal{S}_r(\mathbf{z}_n)$ if, on each interval (z_i, z_{i+1}), $0 \leq i \leq n-1$, $s \in \Pi_r$, where Π_r denotes the space of algebraic polynomials of degree $\leq r$. Also, let $\widetilde{\mathcal{S}}_r(\mathbf{z}_n) := \mathcal{S}_r(\mathbf{z}_n) \cap \mathbb{C}^{r-1}$ be the corresponding space of splines of minimal defect (highest smoothness).

As usual, $\mathbb{L}_p(J)$, $0 < p \leq \infty$, denotes the space of all measurable functions f on an interval J such that $\|f\|_{\mathbb{L}_p(J)} < \infty$, where $\|f\|_{\mathbb{L}_p(J)} := (\int_J |f(x)|^p\,dx)^{1/p}$ if $p < \infty$, and $\|f\|_{\mathbb{L}_\infty(J)} := \operatorname{ess\,sup}_{x\in J} |f(x)|$. For $\mu \in \mathbb{N}$, the space of all μ-times continuously differentiable functions on J is denoted by $\mathbb{C}^\mu(J)$. Also, $\mathbb{C}(J)$ and $AC(J)$ denote the spaces of all continuous and locally absolutely continuous functions, respectively. (Note that if $f \in \mathbb{C}(J)$, then $\|f\|_{\mathbb{C}(J)} = \|f\|_{\mathbb{L}_\infty(J)}$.) The Sobolev space is defined by $\mathbb{W}_p^r(J) := \big\{f \in \mathbb{L}_p(J) \mid f^{(r-1)} \in AC(J) \text{ and } f^{(r)} \in \mathbb{L}_p(J)\big\}$.

For $k \in \mathbb{N}_0$, define

$$\Delta_h^k(f,x,J) := \begin{cases} \displaystyle\sum_{i=0}^{k} \binom{k}{i} (-1)^{k-i} f(x - kh/2 + ih) & \text{if } x \pm kh/2 \in J, \\ 0 & \text{otherwise.} \end{cases}$$

*Received by the editors May 2, 2006; accepted for publication (in revised form) April 8, 2007; published electronically April 23, 2008.

http://www.siam.org/journals/sinum/46-4/65874.html

†Department of Mathematics, University of Manitoba, Winnipeg R3T 2N2, MB, Canada (kopotunk@cc.umanitoba.ca). This author's research was supported in part by NSERC of Canada.

‡School of Mathematical Sciences, Raymond and Beverley Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv 69978, Israel (leviatan@post.tau.ac.il).

§Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton T6G 2G1, AB, Canada (prymak@gmail.com).

The kth modulus of smoothness of $f \in \mathbb{L}_p(J)$ is defined by

$$\omega_k(f,t,J)_p := \sup_{0<h\le t} \left\|\Delta_h^k(f,\cdot,J)\right\|_{\mathbb{L}_p(J)},$$

and the Ditzian–Totik moduli of smoothness is

$$\omega_k^\varphi(f,t)_p := \sup_{0<h\le t} \left\|\Delta_{h\varphi(\cdot)}^k(f,\cdot)\right\|_{\mathbb{L}_p[-1,1]},$$

where $\varphi(x) := \sqrt{1-x^2}$.

The set of q-monotone functions on J is denoted by $\Delta^q(J)$. Recall that $f \in \Delta^q(J)$ if the divided differences $[f;t_0,\dots,t_q]$ of order q of f are nonnegative for all choices of $(q+1)$ distinct points $t_0,\dots,t_q$ in J. If f is continuous, then $f \in \Delta^q(J)$ if $\Delta_h^q(f,x,J) \ge 0$ for all $x \in J$ and $h > 0$. It is well known that, for $q \ge 2$ and an open interval J, $f \in \Delta^q(J)$ if and only if $f^{(q-2)}$ exists and is convex on J.

The error of unconstrained approximation of f from a set U is denoted by

$$E(f,U)_p := \inf_{u\in U} \|f-u\|_{\mathbb{L}_p(J)},$$

and the error of q-monotone approximation (i.e., approximation of f by q-monotone elements of U) is

$$E^{(q)}(f,U)_p := E(f, U\cap\Delta^q(J))_p.$$

Throughout this paper, we also use the notation $\|f\|_p := \|f\|_{\mathbb{L}_p[-1,1]}$, $\Delta^q := \Delta^q[-1,1]$, $|J| := \operatorname{meas}(J)$, and

$$\omega_k(f,J)_p := \omega_k(f,|J|,J)_p \quad \text{and} \quad \omega_k(f,t)_p := \omega_k(f,t,[-1,1])_p.$$

Given a partition $\mathbf{z}_n := (z_i)_{i=0}^n$, $-1 =: z_0 < z_1 < \cdots < z_{n-1} < z_n := 1$, we say that the partition $\tilde{\mathbf{z}}_m = (\tilde z_i)_{i=0}^m$ is a *δ-remesh* of $\mathbf{z}_n$ if, for each $0 \le j \le n-1$,

$$\max\{\tilde z_{i+1} - \tilde z_i \mid [\tilde z_i, \tilde z_{i+1}] \cap (z_j, z_{j+1}) \ne \emptyset\} \le \delta \min_{\nu=j-1,j,j+1} |z_{\nu+1} - z_\nu|$$

with z_{-1} and z_{n+1} defined to be (in this definition only!) $-\infty$ and $+\infty$, respectively. In other words, the largest interval $[\tilde z_i, \tilde z_{i+1}]$ intersecting (z_j, z_{j+1}) should have length at most δ times the length of $[z_j, z_{j+1}]$ or the lengths of (one or two) intervals adjacent to $[z_j, z_{j+1}]$, whichever is smaller. The class of all δ-remeshes of $\mathbf{z}_n$ is denoted by $\mathcal{R}_\delta(\mathbf{z}_n)$.

Clearly, m, n, and δ are not independent. The smallest m such that there is $\tilde{\mathbf{z}}_m$ in $\mathcal{R}_\delta(\mathbf{z}_n)$ is determined not only by n and δ but also by the scale of the partition $\mathbf{z}_n$ (see (1.1)).

It is well known that it is easier and less costly to obtain a good piecewise approximation to a function f than to assure, at the same time, the maximum smoothness of the approximating ppfs. On the other hand, replacing f by a ppf with minimal defect (spline) avoids introduction of artifacts which are unrelated to the original function, and may affect and complicate the problem one deals with. Thus, given two partitions $\mathbf{z}_n$ and $\tilde{\mathbf{z}}_m$ with the only requirement that $\tilde{\mathbf{z}}_m$ is somewhat denser than $\mathbf{z}_n$ (more precisely, $\tilde{\mathbf{z}}_m$ is an *arbitrary* δ-remesh of $\mathbf{z}_n$), we show in this paper how to smooth a general ppf on $\mathbf{z}_n$ to a spline of maximum smoothness on $\tilde{\mathbf{z}}_m$, while staying close to the original ppf and preserving its shape characteristics. We prove that

such a $\delta > 0$ exists that splines on an arbitrary δ-remesh of $\mathbf{z}_n$ may be constructed to satisfy the above requirements. In particular, as an illustration of these general results, we consider two special kinds of partitions of $[-1,1]$ which are important in applications: the *uniform* partition $\mathbf{u}_n := (-1+2j/n)_{j=0}^n$ and the *Chebyshev* partition $\mathbf{t}_n := (-\cos(j\pi/n))_{j=0}^n$.

The following are some properties of classes $\mathcal{R}_\delta(\mathbf{z}_n)$.

- For any partition $\mathbf{z}_n$ and $0 < \delta_1 \le \delta_2$, $\mathcal{R}_{\delta_1}(\mathbf{z}_n) \subset \mathcal{R}_{\delta_2}(\mathbf{z}_n)$.
- If a partition $\mathbf{z}^*_{kn}$ is a refinement of the partition $\mathbf{z}_n$ obtained by subdividing each interval in $\mathbf{z}_n$ into k equal subintervals, and $\tilde{\mathbf{z}}_m \in \mathcal{R}_\delta(\mathbf{z}_n)$, then $\tilde{\mathbf{z}}_m \in \mathcal{R}_{k\delta}(\mathbf{z}^*_{kn})$.
- If $0 < \delta < 1$ and $\tilde{\mathbf{z}}_m \in \mathcal{R}_\delta(\mathbf{z}_n)$, then any interval in $\tilde{\mathbf{z}}_m$ is contained in the union of at most two intervals in $\mathbf{z}_n$.
- For any $m \ge n/\delta$, $\mathbf{u}_m \in \mathcal{R}_\delta(\mathbf{u}_n)$.
- For any $m \ge \max\{25/\delta, 1\}n$, $\mathbf{t}_m \in \mathcal{R}_\delta(\mathbf{t}_n)$.

1.2. Constrained smoothing. Let $\mathbf{z}_n := \{z_0, \dots, z_n | -1 =: z_0 < z_1 < \cdots < z_n := 1\}$ be a partition of $[-1,1]$, and extend the notation by setting $z_j := z_0$, $j < 0$, and $z_j := z_n$, $j > n$. We denote the scale of the partition $\mathbf{z}_n$ by

$$\vartheta(\mathbf{z}_n) := \max_{0\le j\le n-1} \frac{|J_{j\pm1}|}{|J_j|}, \tag{1.1}$$

where $J_j := [z_j, z_{j+1}]$. We also denote $\mathfrak{I}_j := [(z_{j-1}+z_j)/2, (z_j+z_{j+1})/2]$.

The following theorem is our main result on constrained spline smoothing.

THEOREM 1.1. *Let $q \in \mathbb{N}$, $r \in \mathbb{N}$, and let $\mathbf{z}_n = (z_i)_{i=0}^n$ be a partition of $[-1,1]$. There is a constant $\delta = \delta(q,r)$ such that for each $s \in \mathcal{S}_{q+r}(\mathbf{z}_n) \cap \Delta^q$ such that*

$$s \in \mathbb{C}^{q-1}[-1,1], \tag{1.2}$$

and any $\tilde{\mathbf{z}}_m \in \mathcal{R}_\delta(\mathbf{z}_n)$ (i.e., $\tilde{\mathbf{z}}_m$ is a δ-remesh of $\mathbf{z}_n$), there exists a spline $\tilde{s} \in \widetilde{\mathcal{S}}_{q+r}(\tilde{\mathbf{z}}_m) \cap \Delta^q$ satisfying

$$\|s - \tilde{s}\|_{\mathbb{L}_p(\mathfrak{I}_j)} \le c(p,q,r)\omega_{q+r+1}(s, \mathfrak{I}_j)_p, \quad 0 \le j \le n,$$

for all $0 < p \le \infty$. Moreover, the construction of $\tilde{s}$ does not depend on p.

The proof of Theorem 1.1 (as well as the proof of Theorem 1.2 below) is postponed until section 3.

Recall that, for any $q \ge 2$, $\Delta^q \subset \mathbb{C}^{q-2}(-1,1)$; i.e., any q-monotone function is in $\mathbb{C}^{q-2}$. Hence, the smoothness provided by shape itself does not guarantee the applicability of the above result; one needs to assume/gain one additional smoothness degree.

It turns out that for $q = 1$ and $q = 2$ the gain of this additional degree of smoothness is not difficult (see section 3.2), and so we get the following stronger result in the case for $q \le 2$.

THEOREM 1.2. *Let $q = 1$ or 2, $r \in \mathbb{N}$, and let $\mathbf{z}_n = (z_i)_{i=0}^n$ be a partition of $[-1,1]$. There is a constant $\delta = \delta(r)$ such that for each $s \in \mathcal{S}_{q+r}(\mathbf{z}_n) \cap \Delta^q$ and any $\tilde{\mathbf{z}}_m \in \mathcal{R}_\delta(\mathbf{z}_n)$ there exists a spline $\tilde{s} \in \widetilde{\mathcal{S}}_{q+r}(\tilde{\mathbf{z}}_m) \cap \Delta^q$ satisfying*

$$\|s - \tilde{s}\|_{\mathbb{L}_p(\mathfrak{I}_j)} \le c(p,r)\omega_{q+r+1}(s, \mathfrak{I}_j)_p, \quad 0 \le j \le n,$$

for all $0 < p \le \infty$.

Note that an analog of Theorem 1.2 is also valid (and is actually simpler) in the case $r = 0$ (see Lemma 3.8 and Corollary 3.9).

In case the partitions $\mathbf{z}_n$ and $\tilde{\mathbf{z}}_m$ are either both uniform or both Chebyshev, Theorem 1.1 can be restated as follows.

COROLLARY 1.3. *Let $q \in \mathbb{N}$, $r \in \mathbb{N}$, and let $\mathbf{z}_n = (z_i)_{i=0}^n$ denote either $\mathbf{u}_n$ or $\mathbf{t}_n$. There is a constant $m_0 = m_0(q,r)$ such that for each $s \in \mathcal{S}_{q+r}(\mathbf{z}_n) \cap \Delta^q \cap \mathbb{C}^{q-1}[-1,1]$, and any $m \ge m_0 n$, there exists a spline $\tilde{s} \in \widetilde{\mathcal{S}}_{q+r}(\mathbf{z}_m) \cap \Delta^q$, where $\mathbf{z}_m$ is either $\mathbf{u}_m$ or $\mathbf{t}_m$, respectively, satisfying*

$$\|s - \tilde{s}\|_{\mathbb{L}_p(\mathcal{I}_j)} \le c(p,q,r)\omega_{q+r+1}(s, \mathcal{I}_j)_p, \quad 0 \le j \le n,$$

for all $0 < p \le \infty$.

We remark that in view of Theorem 1.2, in the case $q = 1$ and $q = 2$, the condition that s is in $\mathbb{C}^{q-1}[-1,1]$ in Corollary 1.3 can be removed.

Finally, we remark that it is still an open question whether the condition (1.2) in the statement of Theorem 1.1 can be removed if $q \ge 3$.

2. Applications: Jackson-type estimates.

2.1. Monotone and convex spline approximation: $p = \infty$. The following theorem is rather well known. Its positive part follows from Whitney's inequality ($q = 1$ and $1 \le k + \nu \le 2$), [11, Lemma 2] ($q = 1$, $\nu \ge 1$), [13, Corollary 2.4] ($q = 2$, $\nu \ge 2$), and [6] ($q = 2$, $2 \le k + \nu \le 3$). The negative part follows from [17] and [15, p. 141].

THEOREM 2.1 ($p = \infty$). *Let $q = 1$ or 2, and $k, \nu \in \mathbb{N}_0$ be such that either $\nu \ge q$ or $q \le k + \nu \le q + 1$ (see Figures 1 and 2). Then, for every $f \in \Delta^q \cap \mathbb{C}^\nu[-1,1]$, $n \in \mathbb{N}$, and any partition $\mathbf{z}_n = (z_i)_{i=0}^n$ of $[-1,1]$, there exists $s \in \mathcal{S}_{k+\nu-1}(\mathbf{z}_n) \cap \Delta^q$ such that*

$$\|f - s\|_{\mathbb{L}_\infty(J_j)} \le c(k,\nu,\vartheta(\mathbf{z}_n))|J_j|^\nu \omega_k(f^{(\nu)}, J_j)_\infty, \qquad 0 \le j \le n-1.$$

Moreover, this estimate is no longer true in general for k and ν which do not satisfy the above conditions. (This means that for each partition $\mathbf{z}_n$ and any constant $c(k,\nu,\vartheta(\mathbf{z}_n))$, there exists a function $f \in \Delta^q \cap \mathbb{C}^\nu[-1,1]$, such that the above estimate is invalid for any $s \in \mathcal{S}_{k+\nu-1}(\mathbf{z}_n) \cap \Delta^q$.)

We note that the case $q = 2$, $k + \nu = 1$ is excluded from the statement of Theorem 2.1 (as well as the statements of Theorem 2.4 and Corollaries 2.2, 2.3, 2.5, and 2.6). Indeed, the only convex piecewise constant ppfs are constant functions on $[-1,1]$.

Using Theorem 1.2 we can now obtain the following consequence of this result for monotone and convex approximation by splines of any smoothness.

COROLLARY 2.2 ($p = \infty$). *Let $q = 1$ or 2, and $k, \nu \in \mathbb{N}_0$ be such that either $\nu \ge q$ or $q \le k + \nu \le q + 1$. Then, for any $r \ge k + \nu - 1$ and partition $\mathbf{z}_n = (z_i)_{i=0}^n$ of $[-1,1]$, there exists a constant $\delta = \delta(r)$ such that for any $\tilde{\mathbf{z}}_m \in \mathcal{R}_\delta(\mathbf{z}_n)$ and every $f \in \Delta^q \cap \mathbb{C}^\nu[-1,1]$, there exists a spline $\tilde{s} \in \widetilde{\mathcal{S}}_r(\tilde{\mathbf{z}}_m) \cap \Delta^q$ satisfying*

$$\|f - \tilde{s}\|_{\mathbb{L}_\infty(\mathcal{I}_j)} \le c(k,\nu,r,\vartheta(\mathbf{z}_n))|J_j|^\nu \omega_k(f^{(\nu)}, [z_{j-1}, z_{j+1}])_\infty$$

for all $0 \le j \le n$. Moreover, this estimate is no longer true in general for k and ν which do not satisfy the above conditions.

Taking into account Corollary 1.3 and the fact that $|J_j| = n^{-1}$ (if $\mathbf{z}_n = \mathbf{u}_n$) and $|J_j| \sim \varphi(x)n^{-1} + n^{-2}$, $x \in J_j$ (if $\mathbf{z}_n = \mathbf{t}_n$), this, in turn, immediately implies the following result.

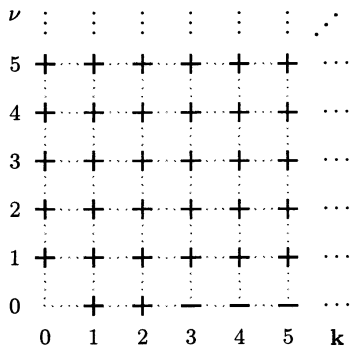


FIG. 1. $q = 1, p = \infty$.

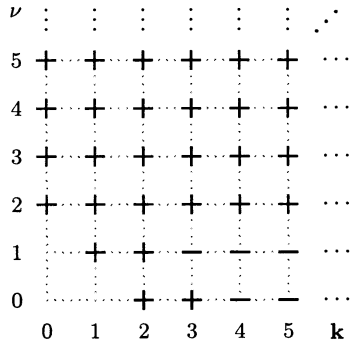


FIG. 2. $q = 2, p = \infty$.

COROLLARY 2.3. *Let $q = 1$ or 2, and $k, \nu \in \mathbb{N}_0$ be such that either $\nu \geq q$ or $q \leq k + \nu \leq q + 1$. Then, for every $f \in \Delta^q \cap \mathbb{C}^\nu[-1, 1]$, $n \in \mathbb{N}$, and $r \geq k + \nu - 1$, we have*

$$E^{(q)}(f, \widetilde{\mathcal{S}}_r(\mathbf{u}_n))_\infty \leq c(k, \nu, r) n^{-\nu} \omega_k(f^{(\nu)}, n^{-1})_\infty \tag{2.1}$$

and

$$E^{(q)}(f, \widetilde{\mathcal{S}}_r(\mathbf{t}_n))_\infty \leq c(k, \nu, r) n^{-\nu} \omega_k^\varphi(f^{(\nu)}, n^{-1})_\infty. \tag{2.2}$$

Moreover, these estimates are no longer true in general for k and ν which do not satisfy the above conditions.

For the reader's convenience we describe the above results using arrays in Figures 1 and 2. In these figures as well as Figures 3 and 4 in the case $1 \leq p < \infty$ (with obvious modifications), the symbols "−" and "+" have the following meaning.

- The symbol "+" in the position (k, ν) means that inequalities (2.1) and (2.2) are valid for all $f \in \Delta^q \cap \mathbb{C}^\nu[-1, 1]$.
- The symbol "−" in the position (k, ν) means that inequalities (2.1) and (2.2) are not true in general, i.e., there are functions $f \in \Delta^q \cap \mathbb{C}^\nu[-1, 1]$ for which these inequalities fail.

We remark that the estimates in Corollary 2.3 (and the arrays in Figures 1 and 2) are well known for algebraic polynomials as well as splines of low smoothness. For splines of minimal defect, only some special cases when $\mathbf{z}_n = \mathbf{u}_n$ were known (see, e.g., [3, Theorem 1], [10, Theorem 2.1], [2, Theorem 1], [8, Corollaries 5.3 and 5.4]).

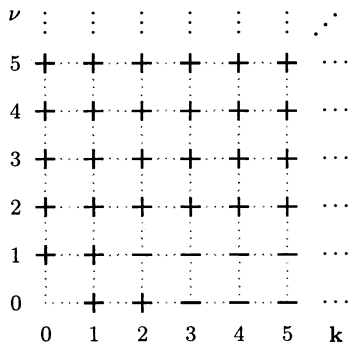


FIG. 3. $q = 1, 1 \le p < \infty$.

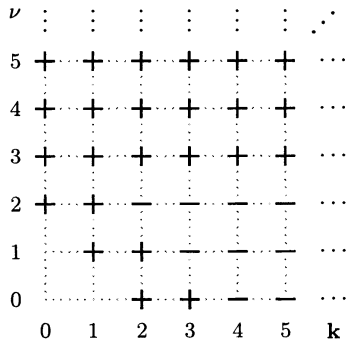


FIG. 4. $q = 2, 1 \le p < \infty$.

2.2. Monotone and convex spline approximation: $1 \le p < \infty$.

THEOREM 2.4 ($1 \le p < \infty$). *Let $q = 1$ or 2, $1 \le p < \infty$, and $k, \nu \in \mathbb{N}_0$ be such that either $\nu \ge q+1$ or $q \le k+\nu \le q+1$. Then, for every $f \in \Delta^q \cap \mathbb{W}_p^\nu[-1,1]$, $n \in \mathbb{N}$, and any partition $\mathbf{z}_n = (z_i)_{i=0}^n$ of $[-1,1]$, there exists $s \in \mathcal{S}_{k+\nu-1}(\mathbf{z}_n) \cap \Delta^q$ such that*

$$\|f - s\|_{\mathbb{L}_p(J_j)} \le c(k, \nu, \vartheta(\mathbf{z}_n))|J_j|^\nu \omega_k(f^{(\nu)}, J_j)_p, \quad 0 \le j \le n-1.$$

Moreover, this estimate is no longer true in general for k and ν which do not satisfy the above conditions.

The positive part of Theorem 2.4 for $\nu \ge 1$ follows from Theorem 2.1 and the well-known inequality

$$\omega_{r+1}(f, J)_\infty \le c|J|^{1-1/p}\, \omega_r(f', J)_p,$$

where J is a closed interval, $f \in \mathbb{W}_p^1(J)$, $1 \le p < \infty$, and $r \in \mathbb{N}_0$. For $\nu = 0$, the case $(q,k) = (1,1)$ is straightforward, $(q,k) = (2,3)$ is shown in [4, Theorem 1.2], and the cases $(q,k) = (1,2)$ and $(2,2)$ are established following the proof of [4, Theorem 1.2] using piecewise linear ppfs which interpolate f at the knots instead of piecewise quadratic functions in [4, Lemma 2.2]. The negative part of Theorem 2.4 is a consequence of [7, Theorem 1].

As in the case $p = \infty$, Theorem 1.2 yields the following stronger result for monotone and convex approximation by splines of any smoothness.

COROLLARY 2.5 ($1 \le p < \infty$). *Let $q = 1$ or 2, $1 \le p < \infty$, and $k, \nu \in \mathbb{N}_0$ be such that either $\nu \ge q+1$ or $q \le k+\nu \le q+1$. Then, for any $r \ge k+\nu-1$ and*

any partition $\mathbf{z}_n = (z_i)_{i=0}^n$ *of* $[-1,1]$, *there exists a constant* $\delta = \delta(r)$ *such that for any* $\tilde{\mathbf{z}}_m \in \mathcal{R}_\delta(\mathbf{z}_n)$ *and every* $f \in \Delta^q \cap \mathbb{W}_p^\nu[-1,1]$ *there exists a spline* $\tilde{s} \in \widetilde{\mathcal{S}}_r(\tilde{\mathbf{z}}_m) \cap \Delta^q$ *satisfying*

$$\|f - \tilde{s}\|_{\mathbb{L}_p(\mathcal{I}_j)} \le c(k,\nu,r,\vartheta(\mathbf{z}_n))|J_j|^\nu \omega_k(f^{(\nu)},[z_{j-1},z_{j+1}])_p$$

for all $0 \le j \le n$. *Moreover, this estimate is no longer true in general for* k *and* ν *which do not satisfy the above conditions.*

Recalling that for $f \in \mathbb{L}_p[-1,1]$, $1 \le p < \infty$, and $k, \mu \in \mathbb{N}$ the following estimates are true (see, e.g., [14], [5]):

$$\sum_{j=0}^{n-\mu-1} \omega_k\left(f, \cup_{i=j}^{j+\mu} J_i\right)_p^p \le \begin{cases} c(k,\mu)\,\omega_k(f,n^{-1})_p^p & \text{if } \mathbf{z}_n = \mathbf{u}_n, \\ c(k,\mu)\,\omega_k^\varphi(f,n^{-1})_p^p & \text{if } \mathbf{z}_n = \mathbf{t}_n, \end{cases}$$

we get the following consequence of Corollary 2.5.

COROLLARY 2.6. *Let* $q = 1$ *or* 2, $1 \le p < \infty$, *and* $k, \nu \in \mathbb{N}_0$ *be such that either* $\nu \ge q+1$ *or* $q \le k+\nu \le q+1$. *Then, for any* $f \in \Delta^q \cap \mathbb{W}_p^\nu[-1,1]$, $n \in \mathbb{N}$, *and* $r \ge k+\nu-1$, *we have*

$$E^{(q)}(f, \widetilde{\mathcal{S}}_r(\mathbf{u}_n))_p \le c(k,\nu,r) n^{-\nu} \omega_k(f^{(\nu)}, n^{-1})_p \tag{2.3}$$

and

$$E^{(q)}(f, \widetilde{\mathcal{S}}_r(\mathbf{t}_n))_p \le c(k,\nu,r) n^{-\nu} \omega_k^\varphi(f^{(\nu)}, n^{-1})_p. \tag{2.4}$$

Moreover, these estimates are no longer true in general for k *and* ν *which do not satisfy the above conditions.*

3. Further results and proofs. Everywhere in this section, if p is a polynomial piece of a spline s on an interval J, i.e., $p := s|_J$, then $p(x)$ for $x \notin J$ is the polynomial extension of $s|_J$.

3.1. Constrained spline smoothing: Proof of Theorem 1.1.

LEMMA 3.1 (see [9, Lemma 2.4 and Corollary 2.5]). *Let* $s \in \mathcal{S}_r(\mathbf{z}_n)$, *and suppose that the polynomials* $p_j := s|_{[z_j,z_{j+1}]}$, $0 \le j \le n-1$. *Then, for every* $1 \le j \le n-1$ *and all* $0 < p \le \infty$,

$$|J_j|^{k+1/p}|p_j^{(k)}(z_j) - p_{j-1}^{(k)}(z_j)| \le c(p,r,\vartheta(\mathbf{z}_n))\omega_{r+1}(s,[z_{j-1},z_{j+1}])_p$$

for $0 \le k \le r$, *and*

$$\|p_j - p_{j-1}\|_{\mathbb{L}_p[z_{j-1},z_{j+1}]} \le c(p,r,\vartheta(\mathbf{z}_n))\omega_{r+1}(s,[z_{j-1},z_{j+1}])_p.$$

LEMMA 3.2 (Markov's inequality). *For any polynomial* $\mathfrak{p} \in \Pi_r$,

$$\|\mathfrak{p}'\|_{\mathbb{C}[a,b]} \le \frac{2r^2}{b-a}\,\|\mathfrak{p}\|_{\mathbb{C}[a,b]}.$$

LEMMA 3.3. *Let* $r \in \mathbb{N}$, $S \in \mathcal{S}_r(\{-1,0,1\})$ *(i.e., S is a ppf of degree* $\le r$ *on* $[-1,1]$ *with the only breakpoint at* 0*), and let* p_1 *and* p_2 *be the polynomial pieces of* S: $p_1 := S|_{[-1,0)}$, $p_2 := S|_{(0,1]}$. *If* S *is nonnegative on* $[-1,1]$, *then there exists an interval* I, $I \subset [-1,0]$ *or* $I \subset [0,1]$ *such that* $|I| \ge \frac{1}{4r^2}$ *and*

$$S(x) \ge c_1(r)\,\|p_1 - p_2\|_\infty \quad \textit{for all } x \in I.$$

Proof. Let $p \in \Pi_r$ be nonnegative on $[0,1]$ (the same argument applies for $[-1,0]$). Then, there exists an interval $I \subset [0,1]$, $|I| \geq \frac{1}{4r^2}$ such that

$$p(x) \geq \frac{1}{2} \|p\|_{\mathbb{C}[0,1]} \quad \text{for all } x \in I.$$

Indeed, let $x^* \in [0,1]$ be a point satisfying $p(x^*) = \|p\|_{\mathbb{C}[0,1]}$. By Lemma 3.2, $\|p'\|_{\mathbb{C}[0,1]} \leq 2r^2 \|p\|_{\mathbb{C}[0,1]}$, and so

$$p(x) \geq p(x^*) - 2r^2 \|p\|_{\mathbb{C}[0,1]} |x - x^*|, \quad x \in [0,1].$$

Hence, if $x \in [0,1]$ and $|x - x^*| \leq \frac{1}{4r^2}$, then $p(x) \geq \frac{1}{2} \|p\|_{\mathbb{C}[0,1]}$, as required.

Now suppose, without loss of generality, that $\|p_2\|_\infty \geq \|p_1\|_\infty$. Then

$$\|p_1 - p_2\|_\infty \leq \|p_1\|_\infty + \|p_2\|_\infty \leq 2 \|p_2\|_\infty \leq c(r) \|p_2\|_{\mathbb{C}[0,1]},$$

and the above observation can be used with $p := p_2$. ∎

The following lemma follows from Beatson [1, Lemma 3.2].

LEMMA 3.4. *Let $r \in \mathbb{N}$, $d := 2r^2$, and $\mathfrak{p}_1, \mathfrak{p}_2 \in \Pi_r$. For any knot sequence $\mathbf{x}_d := (x_i)_{i=0}^d$, $a = x_0 < x_1 < \cdots < x_d = b$, there exists a spline $s \in \widetilde{\mathbb{S}}_r(\mathbf{x}_d)$ such that*

(i) *$s(x)$ is a number between $\mathfrak{p}_1(x)$ and $\mathfrak{p}_2(x)$ for $x \in [a,b]$,*

(ii) *$s \equiv \mathfrak{p}_1$ on $(-\infty, a]$, and $s \equiv \mathfrak{p}_2$ on $[b, \infty)$.*

LEMMA 3.5. *Let $r \in \mathbb{N}$, $S \in \mathbb{S}_r(\{-1,0,1\})$ (i.e., S is a ppf of degree $\leq r$ on $[-1,1]$ with the only breakpoint at 0), and assume that S is nonnegative on $[-1,1]$. Suppose that $d := 2r^2$, and a knot sequence $\mathbf{z}_{2d+1} := (z_i)_{i=0}^{2d+1}$, $-1 < z_0 < z_1 < \cdots < z_{2d+1} < 1$ is such that $z_d \leq 0 \leq z_{d+1}$ (in other words, there are $(d+1)$ knots to the left and to the right of zero). Then, there exists $\widetilde{S} \in \widetilde{\mathbb{S}}_r(\mathbf{z}_{2d+1})$ satisfying*

(i) *$\widetilde{S} \geq 0$ on $[-1,1]$,*

(ii) *$\widetilde{S}(x) = S(x)$ for all $x \notin [z_0, z_{2d+1}]$,*

(iii) *$\|S - \widetilde{S}\|_\infty \leq 2 \|p_1 - p_2\|_\infty$, where $p_1 := S|_{[-1,0)}$, $p_2 := S|_{(0,1]}$ are the polynomial pieces of S.*

Proof. Let $\tilde{p}(x) := p_1(x) + \|p_1 - p_2\|_\infty$. Note that $\tilde{p}(x) \geq p_1(x) \geq 0$ for $x \in [-1,0]$, and $\tilde{p}(x) \geq p_2(x) \geq 0$ for $x \in [0,1]$.

We now use Lemma 3.4 twice. First, we "glue" p_1 with $\tilde{p}$ using Lemma 3.4 with $\mathfrak{p}_1 = p_1$, $\mathfrak{p}_2 = \tilde{p}$, $\mathbf{x}_d = (z_i)_{i=0}^d$, and $[a,b] = [z_0, z_d]$. Second, we "glue" $\tilde{p}$ with p_2 using Lemma 3.4 with $\mathfrak{p}_1 = \tilde{p}$, $\mathfrak{p}_2 = p_2$, $\mathbf{x}_d = (z_i)_{i=d+1}^{2d+1}$, and $[a,b] = [z_{d+1}, z_{2d+1}]$. As a result, we get a spline $\widetilde{S} \in \widetilde{\mathbb{S}}_r(\mathbf{z}_{2d+1})$ such that $\tilde{p} \geq \widetilde{S} \geq p_1 \geq 0$ on $[z_0, z_d]$, $\tilde{p} \geq \widetilde{S} \geq p_2 \geq 0$ on $[z_{d+1}, z_{2d+1}]$, $\widetilde{S}(x) = S(x) \geq 0$ for $x \notin [z_0, z_{2d+1}]$, and $\widetilde{S}(x) = \tilde{p}(x) \geq 0$ for $x \in [z_d, z_{d+1}]$. Clearly,

$$\left\| S - \widetilde{S} \right\|_\infty \leq \max_{i=1,2} \|\tilde{p} - p_i\|_\infty \leq 2 \|p_1 - p_2\|_\infty,$$

and the proof is complete. ∎

The following lemma is the main tool for constrained smoothing.

LEMMA 3.6. *Let $q \in \mathbb{N}$, $r \in \mathbb{N}$. There exists a constant $\delta(q,r) > 0$ such that, for each ppf $s \in \Delta^q \cap \mathbb{C}^{q-1}[-1,1]$ of degree $\leq q + r$ with the only knot at 0 (i.e., $s \in \mathbb{S}_{q+r}(\{-1,0,1\})$), and each partition $\mathbf{z}_l = (z_i)_{i=0}^l$, $-1 = z_0 < z_1 < \cdots < z_{l-1} < z_l = 1$ with*

$$\max_{0 \leq i \leq l-1} (z_{i+1} - z_i) =: \delta < \delta(q,r),$$

there exists a spline $\tilde{s} \in \Delta^q \cap \widetilde{\mathbb{S}}_{q+r}(\mathbf{z}_l)$ *(i.e.,* $\tilde{s}$ *is a spline of minimal defect) such that* $\tilde{s} \equiv s$ *in neighborhoods of* -1 *and* 1*, and*

$$\|s - \tilde{s}\|_p \le c(p,q,r)\omega_{q+r+1}(s,1)_p \tag{3.1}$$

for all $0 < p \le \infty$.

Proof. Since $s^{(q-1)}$ is absolutely continuous and differentiable everywhere except for 0 (because it is a continuous ppf of degree $\le r+1$ with the only breakpoint at 0), we conclude that $S := s^{(q)}$ is a nonnegative ppf of degree $\le r$ with the only knot at 0. Let $p_1 := S|_{[-1,0)}$, $p_2 := S|_{(0,1]}$ be the polynomial pieces of S, and denote $\eta := \|p_1 - p_2\|_\infty$ and $d := 2r^2$. Lemma 3.3 implies that there exists an interval I (without loss of generality, we may assume that it is a subset of $[0,1]$) such that $|I| \ge (2d)^{-1}$ and

$$S(x) \ge c_1\eta \quad \text{for all } x \in I, \tag{3.2}$$

where c_1 depends on r only.

If $P_1 := s|_{[-1,0)}$, $P_2 := s|_{(0,1]}$ are the original polynomial pieces of s, then $p_1 = P_1^{(q)}$, $p_2 = P_2^{(q)}$, and using Lemma 3.2 as well as Lemma 3.1 we have

$$\begin{aligned}\eta &= \left\|P_1^{(q)} - P_2^{(q)}\right\|_\infty \le (q+r)^{2q}\|P_1 - P_2\|_\infty \\ &\le c(p,q,r)\|P_1 - P_2\|_p \le c(p,q,r)\omega_{q+r+1}(s,1)_p.\end{aligned}$$

Now, let I' be the subinterval of I with the same right endpoint as I and length $|I'| = (5d)^{-1}$, so that

$$I' \cap [-(5d)^{-1}, (5d)^{-1}] = \emptyset.$$

Suppose now that the index $i^\star$ is such that $z_{i^\star} \le 0 < z_{i^\star+1}$, and that $\delta < (5d(d+2))^{-1}$. Then each of the intervals $[-(5d)^{-1}, 0]$ and $[0, (5d)^{-1}]$ contains at least $(d+1)$ points from $\mathbf{z}_l$:

$$(z_i)_{i=i^\star-d}^{i^\star} \subset [-(5d)^{-1}, 0] \quad \text{and} \quad (z_i)_{i=i^\star+1}^{i^\star+d+1} \subset [0, (5d)^{-1}].$$

We can apply Lemma 3.5 to construct a nonnegative spline $\widetilde{S}$ of degree r and highest smoothness ($\widetilde{S} \in \mathbb{C}^{r-1}$) having knots $(z_i)_{i=i^\star-d}^{i^\star+d+1}$ only, coinciding with S outside $J_\star := [z_{i^\star-d}, z_{i^\star+d+1}]$, and such that

$$\left\|S - \widetilde{S}\right\|_\infty \le 2\eta.$$

We define s_* to be such that $s_*^{(q)} \equiv \widetilde{S}$ and $s_*^{(\nu)}(-1) = s^{(\nu)}(-1)$ for all $0 \le \nu \le q-1$. Therefore,

$$s_*(x) := \sum_{\nu=0}^{q-1} \frac{s^{(\nu)}(-1)}{\nu!}(x+1)^\nu + \frac{1}{(q-1)!}\int_{-1}^{x}(x-t)^{q-1}\widetilde{S}(t)\,dt.$$

The spline s_* is almost what we need. Namely, it is in $\Delta^q \cap \widetilde{\mathbb{S}}_{q+r}(\mathbf{z}_l)$, it coincides with s in a neighborhood of -1, and since S and $\widetilde{S}$ may differ only on an interval $J_\star$ of length $|J_\star| \le (2d+1)\delta$, we have

$$\begin{aligned}\|s - s_*\|_\infty &\le \left\|\frac{1}{(q-1)!}\int_{-1}^{x}(x-t)^{q-1}\left(S(t) - \widetilde{S}(t)\right)\,dt\right\|_\infty \\ &\le \frac{2^q\eta}{(q-1)!}|J_\star| \le 2^q(2d+1)\eta\delta =: c_2\eta\delta.\end{aligned}$$

The only property that s_* does not have is that it may not be coinciding with s in a neighborhood of 1. To remedy this, we use Lemma 3.4 and "glue" s_* to s on the interval I' preserving its properties.

Let $[a,b] := I'$ and recall that $b-a=(5d)^{-1}$. Now, let $d_1 := 2(q+r)^2$ and take $\delta < h := \frac{b-a}{2(d_1+1)} = \frac{1}{10d(d_1+1)}$. Then, each interval $[a+jh, a+(j+1)h]$, $0 \le j \le 2d_1+1$, contains at least one point z_{i_j} from $\mathbf{z}_l$, and we set $\mathbf{x}_{d_1} = (x_j)_{j=0}^{d_1}$, where $x_j := z_{i_{(2j)}}$, $0 \le j \le d_1$. Note that

$$\max_{0\le j\le d_1}(x_{j+1}-x_j) \le 3h \quad \text{and} \quad \min_{0\le j\le d_1}(x_{j+1}-x_j) \ge h.$$

Taking into account that both s and s_* are polynomials of degree $\le q+r$ on I', Lemma 3.4 implies that there exists a spline $\tilde{s} \in \widetilde{\mathcal{S}}_{q+r}(\mathbf{z}_l)$ such that the only knots of $\tilde{s}$ inside I' are x_j, $0 \le j \le d_1$, $\tilde{s}(x)$ is a number between $s(x)$ and $s_*(x)$ for $x \in [x_0, x_{d_1}] \subset I'$, and $\tilde{s}$ coincides with s_* on $[-1,x_0]$ and with s on $[x_{d_1},1]$. Hence,

$$\|s-\tilde{s}\|_\infty \le \|s-s_*\|_\infty \le c_2\eta\delta$$

(so that (3.1) immediately follows), and it only remains to verify that $\tilde{s} \in \Delta^q[x_0, x_{d_1}]$. To this end, it suffices to show that $\tilde{s}^{(q)}(x) \ge 0$ for all $x \in J_j := [x_j, x_{j+1}]$, $0 \le j \le d_1-1$. Let such an interval J_j be fixed. Using Lemma 3.2 and the fact that $\tilde{s}-s$ is a polynomial of degree $\le q+r$ on J_j, we have

$$\begin{aligned}\left\|\tilde{s}^{(q)} - s^{(q)}\right\|_{\mathbb{C}(J_j)} &\le \left(\frac{2(q+r)^2}{|J_j|}\right)^q \|s-\tilde{s}\|_{\mathbb{C}(J_j)} \le \left(\frac{2(q+r)^2}{h}\right)^q c_2\eta\delta \\ &= (10dd_1(d_1+1))^q\, c_2\eta\delta =: c_3\eta\delta,\end{aligned}$$

where c_3 depends only on r and q. Therefore, for every $x \in J_j$, using (3.2) we have

$$\tilde{s}^{(q)}(x) \ge s^{(q)}(x) - c_3\eta\delta \ge c_1\eta - c_3\eta\delta \ge 0,$$

provided $\delta \le c_1/c_3$. Combining the above restrictions on δ we see that it is possible to take

$$\delta(q,r) := \min\left\{(10d(d_1+1))^{-1}, c_1/c_3\right\},$$

and the construction of $\tilde{s}$ is complete. □

COROLLARY 3.7. *Let $q \in \mathbb{N}$, $r \in \mathbb{N}$. Then there exists a constant $\delta(q,r) > 0$ such that, for each ppf $s \in \Delta^q \cap \mathbb{C}^{q-1}[a,b]$ of degree $\le q+r$ with the only knot at $c \in (a,b)$ (i.e., $s \in \mathcal{S}_{q+r}(\{a,c,b\})$), and each partition $\mathbf{z}_l = (z_i)_{i=0}^l$, $a = z_0 < z_1 < \cdots < z_{l-1} < z_l = b$ with*

$$\max_{0\le i\le l-1}(z_{i+1}-z_i) < \delta(q,r)\min\{b-c, c-a\},$$

there exists a spline $\tilde{s} \in \Delta^q \cap \widetilde{\mathcal{S}}_{q+r}(\mathbf{z}_l)$ (i.e., $\tilde{s}$ is a spline of minimal defect) such that $\tilde{s} \equiv s$ in neighborhoods of a and b, and

$$\|s-\tilde{s}\|_{\mathbb{L}_p[a,b]} \le c(p,q,r)\omega_{q+r+1}(s,[a,b])_p$$

for all $0 < p \le \infty$.

Proof. Without loss of generality, assume $c \ge (b+a)/2$. We apply Lemma 3.6 to $s^*(x) := s(c+(b-c)x)$ and those knots $(z_i-c)/(b-c)$ which belong to $[-1,1]$. Now, extending the resulting ppf $\tilde{s}^*$ polynomially from $[-1,1]$ to all $\mathbb{R}$, we can go back to the original interval $[a,b]$ using $\tilde{s}(t) := \tilde{s}^*((t-c)/(b-c))$. □

Proof of Theorem 1.1. It is sufficient to apply Corollary 3.7 with $[a,b] = \mathfrak{I}_j$ and $c = z_j$, for each $1 \le j \le n-1$. Construction of $\tilde{s}$ is now obvious. □

3.2. Convex and monotone spline smoothing: Auxiliary results for the proof of Theorem 1.2. Recall that for a partition $\mathbf{z}_n := (z_i)_{i=0}^n$, $a =: z_0 < z_1 < \cdots < z_{n-1} < z_n := b$ of an interval $[a,b]$, $z_j := a$, $j < 0$, and $z_j := b$, $j > n$, and that $J_j = [z_j, z_{j+1}]$. The following lemma shows how a piecewise polynomial convex function can be smoothed to be continuously differentiable without adding any extra knots and keeping the error small.

LEMMA 3.8 (convex smoothing). *Let $r \in \mathbb{N}_0$, $\mathbf{z}_n := (z_i)_{i=0}^n$, $a =: z_0 < z_1 < \cdots < z_{n-1} < z_n := b$ be a partition of $[a,b]$, and let $s \in \Delta^2 \cap \mathcal{S}_{r+2}(\mathbf{z}_n)$. Then, there exists $\tilde{s} \in \Delta^2 \cap \mathcal{S}_{r+2}(\mathbf{z}_n) \cap \mathbb{C}^1[a,b]$ such that, for any $1 \le j \le n-1$ and all $0 < p \le \infty$,*

$$\|s - \tilde{s}\|_{\mathbb{L}_p[z_{j-1}, z_{j+1}]} \le c(p, r, \vartheta(\mathbf{z}_n))\omega_{r+3}(s, [z_{j-2}, z_{j+2}])_p \tag{3.3}$$

and

$$\|s' - \tilde{s}'\|_{\mathbb{L}_p[z_{j-1}, z_{j+1}]} \le c(p, r, \vartheta(\mathbf{z}_n))\omega_{r+2}(s', [z_{j-2}, z_{j+2}])_p. \tag{3.4}$$

Moreover,

$$\tilde{s}^{(\nu)}(a) = s^{(\nu)}(a) \quad \text{and} \quad \tilde{s}^{(\nu)}(b) = s^{(\nu)}(b), \quad \nu = 0, 1. \tag{3.5}$$

Proof. This lemma is actually a simpler version of [12, Lemma 1] (see also [16]). It was not proved in [12] for all $p > 0$, and the construction of $\tilde{s}$ was more involved there (because s was allowed to change convexity). For completeness, we recall this construction from [12], adopting it to our case and showing how the estimates can be obtained for all $p > 0$.

With $p_i := s|_{J_i}$, $0 \le i \le n-1$, denote

$$a_i(x) := \frac{(z_{i+2} - z_{i+1})\left(p_{i+1}'(z_{i+1}) - p_i'(z_{i+1})\right)}{2(z_{i+1} - z_i)(z_{i+2} - z_i)}(x - z_i)^2, \quad 0 \le i \le n-2,$$

and

$$b_i(x) := \frac{(z_i - z_{i-1})\left(p_i'(z_i) - p_{i-1}'(z_i)\right)}{2(z_{i+1} - z_i)(z_{i+1} - z_{i-1})}(x - z_{i+1})^2, \quad 1 \le i \le n-1,$$

and also set $a_{n-1}(x) \equiv 0$ and $b_0(x) \equiv 0$. Then, for $x \in [z_i, z_{i+1}]$, $0 \le i \le n-1$, we define

$$\tilde{s}(x) := p_i(x) + a_i(x) + b_i(x).$$

It is now straightforward to verify that s is a continuously differentiable convex function satisfying (3.5), and it remains to prove (3.3) and (3.4). For $1 \le j \le n-1$, we have

$$\begin{aligned}
\|s - \tilde{s}\|_{\mathbb{L}_p[z_{j-1}, z_{j+1}]} &\le 2^{1/p} \max_{i=j-1,j} \|a_i + b_i\|_{\mathbb{L}_p[z_i, z_{i+1}]} \\
&\le c(p, r, \vartheta(\mathbf{z}_n))|J_i|^{1+1/p} \max_{i=j-1,j,j+1\,;\,1\le i\le n-1} \left|p_i'(z_i) - p_{i-1}'(z_i)\right| \\
&\le c(p, r, \vartheta(\mathbf{z}_n))\omega_{r+3}(s, [z_{j-2}, z_{j+2}])_p,
\end{aligned}$$

where the last inequality follows from Lemma 3.1. Similarly, Lemma 3.1 implies

$$\begin{aligned}
\|s' - \tilde{s}'\|_{\mathbb{L}_p[z_{j-1}, z_{j+1}]} &\le 2^{1/p} \max_{i=j-1,j} \|a_i' + b_i'\|_{\mathbb{L}_p[z_i, z_{i+1}]} \\
&\le c(p, r, \vartheta(\mathbf{z}_n))|J_i|^{1/p} \max_{i=j-1,j,j+1\,;\,1\le i\le n-1} \left|p_i'(z_i) - p_{i-1}'(z_i)\right| \\
&\le c(p, r, \vartheta(\mathbf{z}_n))\omega_{r+2}(s', [z_{j-2}, z_{j+2}])_p. \qquad \square
\end{aligned}$$

Lemma 3.8 immediately implies the following result for monotone spline smoothing.

COROLLARY 3.9 (monotone smoothing). *Let $r \in \mathbb{N}_0$, $\mathbf{z}_n := (z_i)_{i=0}^n$, $a =: z_0 < z_1 < \cdots < z_{n-1} < z_n := b$ be a partition of $[a,b]$, and let $s \in \Delta^1 \cap \mathcal{S}_{r+1}(\mathbf{z}_n)$. Then, there exists $\tilde{s} \in \Delta^1 \cap \mathcal{S}_{r+1}(\mathbf{z}_n) \cap \mathbb{C}[a,b]$ such that, for any $1 \le j \le n-1$ and $0 < p \le \infty$,*

$$\|s - \tilde{s}\|_{\mathbb{L}_p[z_{j-1}, z_{j+1}]} \le c(p, r, \vartheta(\mathbf{z}_n))\omega_{r+2}(s, [z_{j-2}, z_{j+2}])_p.$$

Moreover,

$$\tilde{s}(a) = s(a) \quad \text{and} \quad \tilde{s}(b) = s(b).$$

In the case $n = 2$, i.e., when a ppf has only one breakpoint inside an interval $[a,b]$, we get the following corollaries for convex and monotone spline smoothing.

COROLLARY 3.10. *Let $r \in \mathbb{N}_0$, $\mathfrak{z} := (z_i)_{i=0}^2$, $a =: z_0 < z_1 < z_2 := b$ be a partition of $[a,b]$, and let $s \in \Delta^2 \cap \mathcal{S}_{r+2}(\mathfrak{z})$. Then, there exists $\tilde{s} \in \Delta^2 \cap \mathcal{S}_{r+2}(\mathfrak{z}) \cap \mathbb{C}^1[a,b]$ such that*

$$\|s - \tilde{s}\|_{\mathbb{L}_p[a,b]} \le c(p, r, \vartheta(\mathfrak{z}))\omega_{r+3}(s, [a,b])_p$$

for any $0 < p \le \infty$. Moreover,

$$\tilde{s}^{(\nu)}(a) = s^{(\nu)}(a) \quad \text{and} \quad \tilde{s}^{(\nu)}(b) = s^{(\nu)}(b), \quad \nu = 0, 1.$$

COROLLARY 3.11. *Let $r \in \mathbb{N}_0$, $\mathfrak{z} := (z_i)_{i=0}^2$, $a =: z_0 < z_1 < z_2 := b$ be a partition of $[a,b]$, and let $s \in \Delta^1 \cap \mathcal{S}_{r+1}(\mathfrak{z})$. Then, there exists $\tilde{s} \in \Delta^1 \cap \mathcal{S}_{r+1}(\mathfrak{z}) \cap \mathbb{C}[a,b]$ such that*

$$\|s - \tilde{s}\|_{\mathbb{L}_p[a,b]} \le c(p, r, \vartheta(\mathfrak{z}))\omega_{r+2}(s, [a,b])_p$$

for any $0 < p \le \infty$. Moreover, $\tilde{s}(a) = s(a)$ and $\tilde{s}(b) = s(b)$.

Proof of Theorem 1.2. Let $\mathbf{z}_{3n}^\star$ be a refinement of $\mathbf{z}_n$ obtained by adding two extra knots l_i and r_i in each interval $J_i = [z_i, z_{i+1}]$, $0 \le i \le n-1$, where $l_i := z_i + |J_i|/4$ and $r_i := z_{i+1} - |J_i|/4$. We now apply Corollaries 3.10 and 3.11 for every $1 \le i \le n-1$, with $\mathfrak{z} = (r_{i-1}, z_i, l_i)$ to obtain $s^\star \in \mathcal{S}_{q+r}(\mathbf{z}_{3n}^\star) \cap \Delta^q \cap \mathbb{C}^{q-1}[-1,1]$ satisfying

$$\|s - s^\star\|_{\mathbb{L}_p[r_{i-1}, l_i]} \le c(p,r)\omega_{q+r+1}(s, [r_{i-1}, l_i])_p, \qquad 1 \le j \le n-1.$$

Moreover, $s^\star \equiv s$ on $[l_i, r_i]$, $1 \le i \le n-2$, and on $[-1, r_0]$ and $[l_{n-1}, 1]$, and therefore

$$\|s - s^\star\|_{\mathbb{L}_p(\mathfrak{I}_j)} \le c(p,r)\omega_{q+r+1}(s, \mathfrak{I}_j)_p, \qquad 0 \le j \le n. \tag{3.6}$$

Now, using Theorem 1.1 with $s^\star$ and $\mathbf{z}_{3n}^\star$ instead of s and $\mathbf{z}_n$, respectively, and taking into account that, if $\tilde{\mathbf{z}}_m \in \mathcal{R}_\delta(\mathbf{z}_n)$, then $\tilde{\mathbf{z}}_m \in \mathcal{R}_{4\delta}(\mathbf{z}_{3n}^\star)$, we conclude that

$$\|s^\star - \tilde{s}\|_{\mathbb{L}_p(\tilde{\mathfrak{I}}_j)} \le c(p,q,r)\omega_{q+r+1}(s^\star, \tilde{\mathfrak{I}}_j)_p, \qquad 0 \le j \le n,$$

where $\tilde{\mathfrak{I}}_j$ stands for $[(l_{i-1}+r_{i-1})/2, (r_{i-1}+z_i)/2]$, $[(r_{i-1}+z_i)/2, (z_i+l_i)/2]$, or $[(z_i+l_i)/2, (l_i+r_i)/2]$, and $l_i = r_i := -1$ if $i < 0$, and $l_i = r_i := 1$ if $i \ge n$. Finally, we notice that $(l_i + r_i)/2 = (z_i + z_{i+1})/2$ and, hence,

$$\|s^\star - \tilde{s}\|_{\mathbb{L}_p(\mathfrak{I}_j)} \le c(p,q,r)\omega_{q+r+1}(s^\star, \mathfrak{I}_j)_p, \qquad 0 \le j \le n,$$

which, together with (3.6), completes the proof of the theorem. ∎

REFERENCES

[1] R. K. Beatson, *Restricted range approximation by splines and variational inequalities*, SIAM J. Numer. Anal., 19 (1982), pp. 372–380.

[2] R. K. Beatson, *Convex approximation by splines*, SIAM J. Math. Anal., 12 (1981), pp. 549–559.

[3] R. A. DeVore, *Monotone approximation by splines*, SIAM J. Math. Anal., 8 (1977), pp. 891–905.

[4] R. A. DeVore, Y. K. Hu, and D. Leviatan, *Convex polynomial and spline approximation in L_p $(0 < p < \infty)$*, Constr. Approx., 12 (1996), pp. 409–422.

[5] R. A. DeVore, D. Leviatan, and X. M. Yu, *Polynomial approximation in L_p $(0 < p < 1)$*, Constr. Approx., 8 (1992), pp. 187–201.

[6] K. A. Kopotun, *Pointwise and uniform estimates for convex approximation of functions by algebraic polynomials*, Constr. Approx., 10 (1994), pp. 153–178.

[7] K. A. Kopotun, *On K-monotone polynomial and spline approximation in L_p, $0 < p < \infty$ (quasi)norm*, in Approximation Theory VIII, Vol. 1 (College Station, TX, 1995), Ser. Approx. Decompos. 6, World Scientific, River Edge, NJ, 1995, pp. 295–302.

[8] K. A Kopotun, *Univariate splines: Equivalence of moduli of smoothness and applications*, Math. Comp., 76 (2007), pp. 931–945.

[9] K. A. Kopotun, D. Leviatan, and A. V. Prymak, *Nearly monotone spline approximation in $\mathbb{L}_p$*, Proc. Amer. Math. Soc., 134 (2006), pp. 2037–2047.

[10] D. Leviatan and H. N. Mhaskar, *The rate of monotone spline approximation in the L_p-norm*, SIAM J. Math. Anal., 13 (1982), pp. 866–874.

[11] D. Leviatan and I. A. Shevchuk, *Nearly comonotone approximation*, J. Approx. Theory, 95 (1998), pp. 53–81.

[12] D. Leviatan and I. A. Shevchuk, *Coconvex approximation*, J. Approx. Theory, 118 (2002), pp. 20–65.

[13] D. Leviatan and I. A. Shevchuk, *Coconvex polynomial approximation*, J. Approx. Theory, 121 (2003), pp. 100–118.

[14] P. P. Petrushev and V. A. Popov, *Rational Approximation of Real Functions*, Encyclopedia Math. Appl. 28, Cambridge University Press, Cambridge, UK, 1987.

[15] I. A. Shevchuk, *Approximation by Polynomials and Traces of the Functions Continuous on an Interval*, Naukova Dumka, Kyiv, 1992.

[16] I. A. Shevchuk, *One construction of cubic convex spline*, in Approximation and Optimization, Vol. I (Cluj-Napoca, 1996), Transylvania, Cluj-Napoca, 1997, pp. 357–368.

[17] A. S. Švedov, *Orders of coapproximation of functions by algebraic polynomials*, Mat. Zametki, 29 (1981), pp. 117–130 (in Russian).

SIAM J. NUMER. ANAL.
Vol. 46, No. 4, pp. 1998–2021
© 2008 Society for Industrial and Applied Mathematics

A LOCAL DISCONTINUOUS GALERKIN METHOD FOR THE CAMASSA–HOLM EQUATION*

YAN XU† AND CHI-WANG SHU‡

Abstract. In this paper, we develop, analyze, and test a local discontinuous Galerkin (LDG) method for solving the Camassa–Holm equation which contains nonlinear high-order derivatives. The LDG method has the flexibility for arbitrary h and p adaptivity. We prove the L^2 stability for general solutions and give a detailed error estimate for smooth solutions, and provide numerical simulation results for different types of solutions of the nonlinear Camassa–Holm equation to illustrate the accuracy and capability of the LDG method.

Key words. local discontinuous Galerkin method, Camassa–Holm equation, stability, error estimate

AMS subject classifications. 65M60, 35Q53

DOI. 10.1137/070679764

1. Introduction.

In this paper, we consider numerical approximations to the Camassa–Holm (CH) equation [4, 5]

$$u_t - u_{xxt} + 2\kappa u_x + 3uu_x = 2u_x u_{xx} + uu_{xxx}, \tag{1.1}$$

where κ is a constant. Such nonlinearly dispersive partial differential equations (PDEs) support peakon solutions. The lack of smoothness at the peak of the peakon introduces high-frequency dispersive errors into the calculation. It is a challenge to design stable and accurate numerical schemes for solving this equation.

We develop a class of local discontinuous Galerkin (LDG) methods for this nonlinear CH equation. Our proposed scheme is high-order accurate, nonlinear stable, and flexible for arbitrary h and p adaptivity. The proof of the L^2 stability of the scheme is given for general solutions. Error estimates are given for smooth solutions. To our best knowledge, this is the first provably stable finite element method for the CH equation.

The main motivation for the algorithm discussed in this paper originates from the LDG techniques which have been developed for convection-diffusion equations (containing second derivatives) [9] and nonlinear wave equations with high-order derivatives [32, 33, 24, 27, 28, 29, 30]. In these papers, stable LDG methods for quite general nonlinear wave equations including multidimensional and system cases have been developed. The proofs of the nonlinear L^2 stability of these methods are usually given and successful numerical experiments demonstrate their capability. These results indicate that the LDG method is a good tool for solving nonlinear equations in mathematical physics.

*Received by the editors January 10, 2007; accepted for publication (in revised form) April 18, 2007; published electronically April 23, 2008.
http://www.siam.org/journals/sinum/46-4/67976.html

†Department of Applied Mathematics, University of Twente, P. O. Box 217, 7500 AE, Enschede, The Netherlands. Current address: Department of Mathematics, University of Science and Technology of China, Hefei, Anhui 230026, People's Republic of China (yxu@ustc.edu.cn). The research of this author was supproted by NSFC grant 10601055.

‡Division of Applied Mathematics, Brown University, Providence, RI 02912 (shu@dam.brown.edu). The research of this author was supported by ARO grant W911NF-04-1-0291 and NSF grant DMS-0510345.

There are only a few works in the literature for error estimates of the LDG method for nonlinear wave equations with high-order derivatives. In [31], a priori error estimates are given for the LDG method for nonlinear convection-diffusion equations and KdV equations. There are technical difficulties to derive the L^2 a priori error estimates from the cell entropy inequality and approximation results, because of the possible lack of control on some of the jump terms at cell boundaries, which appear because of the discontinuous nature of the finite element space for the discontinuous Galerkin (DG) method. The remedy in [31] to handle such jump terms is via a special projection, which eliminates such troublesome jump terms in the error equation. It is more challenging to perform L^2 a priori error estimates for nonlinear PDEs with high-order derivatives than for first-order hyperbolic PDEs in [34].

The CH equation (1.1) was derived as a model for the propagation of the unidirectional gravitational waves in a shallow water approximation, with u representing the free surface of water over a flat bed [4, 5, 18]. The CH equation has a very intriguing structure. For instance, it is completely integrable and models wave breaking for a large class of initial data. This equation has attracted a lot of attention in the literature. Lenells gave a detailed discussion of its conservation law in [23] and also classified all the weak traveling wave solutions in [22]. Johnson [19] implemented the inverse-scattering transform method to the solution of the CH equation. Li and Zhang [25] used the same method to solve a multiple-soliton solution for the CH equation. Bressan and Constantin [2, 3] developed a new approach for its analysis, particularly for the investigation of the wave breaking.

There are only a few numerical works in the literature to solve the CH equation. Holden and Raynaud [17] proved the convergence of a finite difference method for the CH equation and they also developed a convergent numerical scheme based on multipeakon in [16]. Several different aspects of periodic traveling wave solutions of the CH equation were numerically explored in [20] and the error analysis of a spectral projection of the periodic CH equation was given in [21]. A finite-volume method was developed in [1] to simulate the dynamics of peakons.

The DG method we discuss in this paper is a class of finite element methods using completely discontinuous piecewise polynomial space for the numerical solution and the test functions in the spatial variables. The DG discretization results in an extremely local, element based discretization, which is beneficial for parallel computing and maintaining high-order accuracy on unstructured meshes. In particular, DG methods are well suited for *hp*-adaptation, which consists of local mesh refinement and/or the adjustment of the polynomial order in individual elements. They also have excellent provable nonlinear stability. The LDG method for the CH equation (1.1) that we design in this paper shares all these nice properties. More general information about DG methods can be found in [7, 10, 11, 12]. Recently, Eskilsson and Sherwin [13, 14, 15] also presented discontinuous spectral element methods for simulating one-dimensional linear Boussinesq-type equations, dispersive shallow water systems, and two-dimensional Boussinesq equations.

The paper is organized as follows. In section 2, we present and analyze our LDG method for the CH equation (1.1). In section 2.2, we present the LDG method. Details related to the implementation of the method are described in section 2.3. We give a proof of the L^2 stability in section 3, and present an a priori error estimate in section 4. Section 5 contains numerical results to demonstrate the accuracy and capability of the methods. Concluding remarks are given in section 6. Some of the more technical proofs of several lemmas are collected in the appendix.

2. The LDG method for the CH equation.

2.1. Notation. We denote the mesh by $I_j = [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}]$ for $j = 1, \ldots, N$. The center of the cell is $x_j = \frac{1}{2}(x_{j-\frac{1}{2}} + x_{j+\frac{1}{2}})$ and the mesh size is denoted by $h_j = x_{j+\frac{1}{2}} - x_{j-\frac{1}{2}}$, with $h = \max_{1\le j\le N} h_j$ being the maximum mesh size. We assume that the mesh is regular, namely the ratio between the maximum and the minimum mesh sizes stays bounded during mesh refinements. We define the piecewise polynomial space V_h as the space of polynomials of the degree up to k in each cell I_j, i.e.,

$$V_h = \{v : v \in P^k(I_j) \text{ for } x \in I_j,\ j = 1, \ldots, N\}.$$

Note that functions in V_h are allowed to have discontinuities across element interfaces.

The solution of the numerical scheme is denoted by u_h, which belongs to the finite element space V_h. We denote by $(u_h)^+_{j+\frac{1}{2}}$ and $(u_h)^-_{j+\frac{1}{2}}$ the values of u_h at $x_{j+\frac{1}{2}}$, from the right cell I_{j+1} and from the left cell I_j, respectively. We use the usual notations $[u_h] = u_h^+ - u_h^-$ and $\bar{u}_h = \frac{1}{2}(u_h^+ + u_h^-)$ to denote the jump and the mean of the function u_h, respectively, at each element boundary point.

2.2. The LDG method. In this section, we define our LDG method for the CH equation (1.1), written in the following form:

$$u - u_{xx} = q, \tag{2.1}$$

$$q_t + f(u)_x = \frac{1}{2}(u^2)_{xxx} - \frac{1}{2}((u_x)^2)_x \tag{2.2}$$

with an initial condition

$$u(x, 0) = u_0(x) \tag{2.3}$$

and periodic boundary conditions

$$u(x, t) = u(x + L, t), \tag{2.4}$$

where L is the period in the x direction and $f(u) = 2\kappa u + \frac{3}{2}u^2$. Notice that the assumption of periodic boundary conditions is for simplicity only and is not essential: the method can be easily designed for nonperiodic boundary conditions.

To define the LDG method, we further rewrite (2.1) as a first-order system:

$$\begin{aligned} u - r_x &= q, \\ r - u_x &= 0. \end{aligned} \tag{2.5}$$

The LDG method for (2.5), where q is assumed known and we would want to solve for u, is formulated as follows: Find u_h, $r_h \in V_h$ such that, for all test functions ρ, $\phi \in V_h$,

$$\int_{I_j} u_h \rho dx + \int_{I_j} r_h \rho_x dx - (\widehat{r}_h \rho^-)_{j+\frac{1}{2}} + (\widehat{r}_h \rho^+)_{j-\frac{1}{2}} = \int_{I_j} q_h \rho dx, \tag{2.6}$$

$$\int_{I_j} r_h \phi dx + \int_{I_j} u_h \phi_x dx - (\widehat{u}_h \phi^-)_{j+\frac{1}{2}} + (\widehat{u}_h \phi^+)_{j-\frac{1}{2}} = 0. \tag{2.7}$$

The "hat" terms in (2.6)–(2.7) in the cell boundary terms from integration by parts are the so-called *numerical fluxes*, which are single-valued functions defined on the

edges and should be designed based on different guiding principles for different PDEs to ensure stability. For the standard elliptic equation (2.5), we can take the simple choices such that

$$\widehat{r}_h = r_h^-, \quad \widehat{u}_h = u_h^+, \tag{2.8}$$

where we have omitted the half-integer indices $j + \frac{1}{2}$ as all quantities in (2.8) are computed at the same points (i.e., the interfaces between the cells). We remark that the choice of the fluxes (2.8) is not unique. We can, for example, also choose the following numerical flux

$$\widehat{r}_h = r_h^+, \quad \widehat{u}_h = u_h^-. \tag{2.9}$$

For (2.2), we can also rewrite it into a first-order system:

$$\begin{aligned} q_t + f(u)_x - p_x + B(r)_x &= 0, \\ p - (b(r)u)_x &= 0, \\ r - u_x &= 0, \end{aligned} \tag{2.10}$$

where $B(r) = \frac{1}{2}r^2$ and $b(r) = B'(r) = r$. Now we can define an LDG method to (2.10), resulting in the following scheme: Find q_h, p_h, $r_h \in V_h$ such that, for all test functions φ, ψ, $\eta \in V_h$,

$$\begin{aligned} &\int_{I_j} (q_h)_t \varphi dx - \int_{I_j} (f(u_h) - p_h + B(r_h))\varphi_x dx \\ &\quad + ((\widehat{f} - \widehat{p}_h + \widehat{B(r_h)})\varphi^-)_{j+\frac{1}{2}} - ((\widehat{f} - \widehat{p}_h + \widehat{B(r_h)})\varphi^+)_{j-\frac{1}{2}} = 0, \end{aligned} \tag{2.11}$$

$$\int_{I_j} p_h \psi dx + \int_{I_j} b(r_h) u_h \psi_x dx - (\widehat{b(r_h)}\widetilde{u}_h \psi^-)_{j+\frac{1}{2}} + (\widehat{b(r_h)}\widetilde{u}_h \psi^+)_{j-\frac{1}{2}} = 0, \tag{2.12}$$

$$\int_{I_j} r_h \eta dx + \int_{I_j} u_h \eta_x dx - (\widehat{u}_h \eta^-)_{j+\frac{1}{2}} + (\widehat{u}_h \eta^+)_{j-\frac{1}{2}} = 0. \tag{2.13}$$

The numerical fluxes in (2.11)–(2.13) are chosen as

$$\widehat{p}_h = p_h^-, \ \widehat{u}_h = u_h^+, \ \widehat{B(r_h)} = B(r_h^-), \ \widehat{b(r_h)} = \frac{B(r_h^+) - B(r_h^-)}{r_h^+ - r_h^-}, \ \widetilde{u}_h = u_h^+. \tag{2.14}$$

Here $\widehat{f}(u_h^-, u_h^+)$ is a monotone flux for solving conservation laws; i.e., it is Lipschitz continuous in both arguments, consistent ($\widehat{f}(u_h, u_h) = f(u_h)$), nondecreasing in the first argument, and nonincreasing in the second argument. Examples of monotone fluxes which are suitable for DG methods can be found in, e.g., [8]. We could, for example, use the simple Lax–Friedrichs flux

$$\widehat{f}(u_h^-, u_h^+) = \frac{1}{2}(f(u_h^-) + f(u_h^+) - \alpha(u_h^+ - u_h^-)), \quad \alpha = \max |f'(u_h)|,$$

where the maximum is taken over a relevant range of u_h. This Lax–Friedrichs flux is used in the numerical experiments in next section. The definition of the algorithm is now complete.

We remark that the choice of the fluxes (2.14) is not unique. In fact the crucial part is taking $\widehat{p}_h$ and $\widehat{u}_h$ from opposite sides and $\widehat{B(r_h)}$ and $\widetilde{u}_h$ from opposite sides.

2.3. Algorithm flowchart. In this section, we give details related to the implementation of the method.

- First, from (2.6)–(2.8), we obtain q_h in the following matrix form:

$$\boldsymbol{q}_h = \mathbf{A}\boldsymbol{u}_h, \tag{2.15}$$

  where $\boldsymbol{q}_h$ and $\boldsymbol{u}_h$ are the vectors containing the degrees of freedom for q_h and u_h, respectively.
- From (2.11)–(2.14), we obtain the LDG discretization of the residual $-f(u)_x + \frac{1}{2}(u^2)_{xxx} - \frac{1}{2}((u_x)^2)_x$ in the following vector form:

$$(\boldsymbol{q}_h)_t = \mathbf{res}(\boldsymbol{u}_h). \tag{2.16}$$

- We then combine (2.15) and (2.16) to obtain

$$\mathbf{A}(\boldsymbol{u}_h)_t = \mathbf{res}(\boldsymbol{u}_h). \tag{2.17}$$

- We use a time discretization method to solve

$$(\boldsymbol{u}_h)_t = \mathbf{A}^{-1}\mathbf{res}(\boldsymbol{u}_h). \tag{2.18}$$

  This step involves a linear solver with the matrix $\mathbf{A}$. We perform an LU decomposition for $\mathbf{A}$ at the beginning and use it for all time steps. Any standard ODE solvers can be used here, for example the Runge–Kutta methods.

The LDG matrix $\mathbf{A}$ is a sparse block matrix, hence its multiplication with vectors and a linear solver, involving it as the coefficient matrix, can be implemented efficiently.

3. L^2 stability of the LDG method. In this section, we prove the L^2 stability of the LDG method for the CH equation defined in the previous section.

PROPOSITION 3.1 (L^2 stability). *The solution to the schemes* (2.6)–(2.8) *and* (2.11)–(2.14) *satisfies the L^2 stability*

$$\frac{d}{dt}\int_0^L (u_h^2 + r_h^2)dx \le 0. \tag{3.1}$$

Proof. For (2.6), we first take the time derivative and get

$$\int_{I_j} (u_h)_t \rho dx + \int_{I_j} (r_h)_t \rho_x dx - (\widehat{(r_h)_t}\rho^-)_{j+\frac{1}{2}} + (\widehat{(r_h)_t}\rho^+)_{j-\frac{1}{2}} = \int_{I_j} (q_h)_t \rho dx. \tag{3.2}$$

Since (3.2), (2.7), and (2.11)–(2.13) hold for any test functions in V_h, we can choose

$$\rho = -u_h, \quad \phi = (r_h)_t, \quad \varphi = u_h, \quad \psi = -r_h, \quad \eta = p_h.$$

With these choices of test functions and summing up the five equations in (3.2), (2.7), and (2.11)–(2.13), we obtain

$$\begin{aligned}
&\int_{I_j} ((u_h)_t u_h + (r_h)_t r_h)dx - \int_{I_j} f(u_h)(u_h)_x dx + (\widehat{f(u_h)}u_h^-)_{j+\frac{1}{2}} - (\widehat{f(u_h)}u_h^+)_{j-\frac{1}{2}} \\
&\quad + \int_{I_j} (p_h u_h)_x dx - (\widehat{p}_h u_h^- + \widehat{u}_h p_h^-)_{j+\frac{1}{2}} + (\widehat{p}_h u_h^+ + \widehat{u}_h p_h^+)_{j-\frac{1}{2}} \\
&\qquad - \int_{I_j} (B(r_h)u_h)_x dx + (\widehat{B(r_h)}u_h^- + \widehat{b(r_h)}\widetilde{u}_h r_h^-)_{j+\frac{1}{2}} - (\widehat{B(r_h)}u_h^+ + \widehat{b(r_h)}\widetilde{u}_h r_h^+)_{j-\frac{1}{2}}
\end{aligned}$$

$$+\int_{I_j}((r_h)_t u_h)_x dx - (\widehat{(r_h)}_t u_h^- + \widehat{u}_h(r_h^-)_t)_{j+\frac{1}{2}} + (\widehat{(r_h)}_t u_h^+ + \widehat{u}_h(r_h^+)_t)_{j-\frac{1}{2}} = 0.$$

Taking $F(u_h) = \int^{u_h} f(\tau)d\tau$, we have

$$\int_{I_j}((u_h)_t u_h + (r_h)_t r_h)dx + \Psi_{j+\frac{1}{2}} - \Psi_{j-\frac{1}{2}} + \Theta_{j-\frac{1}{2}} = 0, \tag{3.3}$$

where the numerical entropy fluxes are given by

$$\begin{aligned}\Psi_{j+\frac{1}{2}} = \Big(&-F(u_h^-) + \widehat{f}u_h^- + p_h^- u_h^- - (\widehat{p}_h u_h^- + \widehat{u}_h p_h^-)\\ &- B(r_h^-)u_h^- + \widehat{B(r_h)}u_h^- + \widehat{b(r_h)}\widetilde{u}_h r_h^- + (r_h^-)_t u_h^- - (\widehat{(r_h)}_t u_h^- + \widehat{u}_h(r_h^-)_t)\Big)_{j+\frac{1}{2}},\end{aligned}$$

and the extra term Θ is given by

$$\begin{aligned}\Theta_{j-\frac{1}{2}} = \Big(&[F(u_h)] - \widehat{f}[u_h] - [p_h u_h] + \widehat{p}_h[u_h] + \widehat{u}_h[p_h]\\ &+ [B(r_h)u_h] - \widehat{b(r_h)}\widetilde{u}_h[r_h] - \widehat{B(r_h)}[u_h] - [(r_h)_t u_h] + \widehat{(r_h)}_t[u_h] + \widehat{u}_h[(r_h)_t]\Big)_{j-\frac{1}{2}}.\end{aligned}$$

With the definitions (2.8) and (2.14) of the numerical fluxes and after some algebraic manipulation, we easily obtain

$$\begin{aligned}-[p_h u_h] + \widehat{p}_h[u_h] + \widehat{u}_h[p_h] &= 0,\\ [B(r_h)u_h] - \widehat{b(r_h)}\widetilde{u}_h[r_h] - \widehat{B(r_h)}[u_h] &= 0,\\ -[(r_h)_t u_h] + \widehat{(r_h)}_t[u_h] + \widehat{u}_h[(r_h)_t] &= 0,\end{aligned}$$

and hence

$$\Theta_{j-\frac{1}{2}} = ([F(u_h)] - \widehat{f}[u_h])_{j-\frac{1}{2}} \geq 0,$$

where the last inequality follows from the monotonicity of the flux

$$[F(u_h)] - \widehat{f}[u_h] = \int_{u_h^-}^{u_h^+}(f(s) - \widehat{f}(u_h^-, u_h^+))ds \geq 0.$$

By summing up the cell entropy inequalities, we obtain the desired L^2 stability (3.1). ∎

4. Error estimates of the LDG method. In this section, we present the procedure to obtain a priori error estimates of the LDG method for the CH equation.

4.1. Notations, definitions, and auxiliary results. In this section we introduce notations and definitions to be used later in the paper and also present some auxiliary results. We first review an important quantity measuring the relationship between the numerical flux and physical flux introduced in [34, 31]. We then define some projections and present certain interpolation and inverse properties for the finite element spaces that will be used in the error analysis.

4.1.1. Notations for different constants. We will adopt the following convention for different constants. These constants may have a different value in each occurrence.

We will denote by C a positive constant independent of h, which may depend on the solution of the problem considered in this paper. Especially, to emphasize the nonlinearity of the flux $f(u)$ (or other nonlinear fluxes), we will denote by $C_\star$ a positive constant depending on the maximum of $|f''|$ or/and $|f'''|$. We remark that $C_\star = 0$ for a linear flux $f = cu$ with a constant c. For problems considered in this section, the exact solution is assumed to be smooth with periodic or compactly supported boundary condition. Also, $0 \le t \le T$ for a fixed T. Therefore, the exact solution is always bounded. We follow the convention [34] to redefine the nonlinear functions $f(u)$, $B(u)$, etc. outside their ranges such that the derivatives of these nonlinear functions, $f'(u)$, $f''(u)$, etc., become globally bounded functions.

4.1.2. A quantity related to the numerical flux. For notational convenience we would like to introduce the following numerical flux related to the DG spatial discretization. $\widehat{f}(\omega^-, \omega^+)$ is a given monotone numerical flux that depends on the two values of the function ω at the discontinuity point $x_{j+\frac{1}{2}}$, namely $\omega^{\pm}_{j+\frac{1}{2}} = \omega(x^{\pm}_{j+\frac{1}{2}})$. The numerical flux $\widehat{f}(\omega^-, \omega^+)$ satisfies the following conditions:

(a) it is locally Lipschitz continuous, so it is bounded when $\omega^{\pm}$ are in a bounded interval;
(b) it is consistent with the flux $f(\omega)$, i.e., $\widehat{f}(\omega, \omega) = f(\omega)$;
(c) it is a nondecreasing function of its first argument, and a nonincreasing function of its second argument.

In [34], Zhang and Shu introduced an important quantity to measure the difference between the numerical flux and the physical flux. In [31], Xu and Shu introduced the idea of "uniform dissipative flux." For completeness, we give their definition in the following lemma.

LEMMA 4.1 (see [34]). *For any piecewise smooth function $\omega \in L^2(0,1)$, on each cell boundary point we define*

$$\alpha(\widehat{f};\omega) \equiv \alpha(\widehat{f};\omega^-,\omega^+) \triangleq \begin{cases} [\omega]^{-1}(f(\bar{\omega}) - \widehat{f}(\omega)), & \text{if} [\omega] \neq 0; \\ \frac{1}{2}|f'(\bar{\omega})|, & \text{if} [\omega] = 0, \end{cases} \tag{4.1}$$

where $\widehat{f}(\omega) \equiv \widehat{f}(\omega^-, \omega^+)$ is a monotone numerical flux consistent with the given flux f. Then $\alpha(\widehat{f};\omega)$ is nonnegative and bounded for any $(\omega^-, \omega^+) \in \mathbb{R}^2$. Moreover, we have

$$\frac{1}{2}|f'(\bar{\omega})| \le \alpha(\widehat{f};\omega) + C_\star|[\omega]|, \tag{4.2}$$

$$-\frac{1}{8}f''(\bar{\omega})[\omega] \le \alpha(\widehat{f};\omega) + C_\star|[\omega]|^2. \tag{4.3}$$

Remark 4.1. Examples of monotone fluxes which are suitable for the DG methods can be found in, e.g., [8]. For our error estimates, we rewrite the numerical flux in a viscosity form

$$\widehat{f}(\omega^-, \omega^+) = \frac{1}{2}(f(\omega^-) + f(\omega^+) - \lambda(\omega^-, \omega^+)(\omega^+ - \omega^-)), \tag{4.4}$$

and assume that the viscosity coefficient $\lambda(\omega^-, \omega^+)$ satisfies

$$\lambda(\omega^-, \omega^+) \ge \lambda_0 > 0, \quad \lambda_0 \text{ is constant.} \tag{4.5}$$

We call such flux a "uniform dissipative flux" [31]. The well-known Lax–Friedrichs flux is a uniform dissipative flux with a proper choice of λ. This property is necessary in our proof because of a lack of control for certain jump terms at cell boundaries due to the nonlinear terms and the high-order derivative terms.

We would also like to use the following simplified notation. For any functions ω and ϕ, we denote

$$\alpha(\widehat{f};\omega)[\phi]^2 = \sum_{1\le j\le N} \alpha(\widehat{f};\omega)_{j+\frac{1}{2}}[\phi]^2_{j+\frac{1}{2}}.$$

4.1.3. Projection and interpolation properties. In what follows, we will consider the standard L^2 projection of a function ω with $k+1$ continuous derivatives into space V_h, denoted by $\mathcal{P}$; i.e., for each j,

$$\int_{I_j} (\mathcal{P}\omega(x) - \omega(x))v(x)dx = 0 \quad \forall v \in P^k(I_j), \tag{4.6}$$

and the special projection $\mathcal{P}^{\pm}$ into V_h, which satisfy, for each j,

$$\begin{aligned}&\int_{I_j} (\mathcal{P}^+\omega(x) - \omega(x))v(x)dx = 0 \quad \forall v \in P^{k-1}(I_j), \quad \text{and} \quad \mathcal{P}^+\omega(x^+_{j-\frac{1}{2}}) = \omega(x_{j-\frac{1}{2}}),\\ &\int_{I_j} (\mathcal{P}^-\omega(x) - \omega(x))v(x)dx = 0 \quad \forall v \in P^{k-1}(I_j), \quad \text{and} \quad \mathcal{P}^-\omega(x^-_{j+\frac{1}{2}}) = \omega(x_{j+\frac{1}{2}}).\end{aligned} \tag{4.7}$$

For the projections mentioned above, it is easy to show that (cf. [6])

$$\|\omega^e\| + h\|\omega^e\|_\infty + h^{\frac{1}{2}}\|\omega^e\|_{\Gamma_h} \le Ch^{k+1}, \tag{4.8}$$

where $\omega^e = \mathcal{P}\omega - \omega$ or $\omega^e = \mathcal{P}^{\pm}\omega - \omega$. The positive constant C, solely depending on ω, is independent of h. Γ_h denotes the set of boundary points of all elements I_j.

4.1.4. Inverse properties. Finally, we list some inverse properties of the finite element space V_h that will be used in our error analysis. For any $\omega_h \in V_h$, there exists a positive constant C independent of ω_h and h, such that

(4.9)

$$\text{(i) } \|\partial_x\omega_h\| \le Ch^{-1}\|\omega_h\|, \quad \text{(ii) } \|\omega_h\|_{\Gamma_h} \le Ch^{-\frac{1}{2}}\|\omega_h\|, \quad \text{(iii) } \|\omega_h\|_\infty \le Ch^{-\frac{1}{2}}\|\omega_h\|.$$

For more details of these inverse properties, we refer the readers to [6].

4.2. The main error estimate result. We state the main error estimates of the semidiscrete LDG scheme for the CH equation.

THEOREM 4.2. *Let u be the exact solution of the problem* (2.1)–(2.2)*, which is sufficiently smooth with bounded derivatives, and assume that $f \in C^3$. Let u_h be the numerical solution of the semidiscrete LDG scheme,* (2.6)–(2.8) *and* (2.11)–(2.14)*, and denote the corresponding numerical error by $e_u = u - u_h$ and $e_r = r - r_h$, where $r = u_x$ is defined by* (2.5)*. For regular triangulations of $I = (0,1)$, if the finite element space V_h is the piecewise polynomials of degree $k \ge 2$, then for small enough h the following error estimates hold:*

$$\|u - u_h\|^2 + \|r - r_h\|^2 \le Ch^{2k}, \tag{4.10}$$

where the constant C depends on the final time T, k, $\|u\|_{k+1}$, $\|r\|_{k+1}$, and the bounds on the derivatives $|f^{(m)}|$, $m = 1,2,3$. Here $\|u\|_{k+1}$ and $\|r\|_{k+1}$ are the maximum over $0 \le t \le T$ of the standard Sobolev $k+1$ norm in space.

Remark 4.2. Although we could not obtain the optimal error estimates $O(h^{k+1})$ for u due to some extra boundary terms arising from high-order derivatives, numerical examples in section 5 verify the optimal order $O(h^{k+1})$ for u. For the solution r_h, our numerical results indicate that kth-order accuracy is sharp.

4.2.1. The error equation. In order to obtain the error estimate to smooth solutions for the considered semidiscrete LDG scheme (2.6)–(2.8) and (2.11)–(2.14), we need to first obtain the error equation.

Notice that the scheme (2.6) and (2.11)–(2.13) is also satisfied when the numerical solutions are replaced by the exact solutions. We then obtain the error equation

$$\begin{aligned}
&\int_{I_j}(u-u_h)_t\rho dx-\int_{I_j}\Big((q-q_h)_t(\rho+\varphi)+(r-r_h)\phi+(p-p_h)\psi+(r-r_h)\eta\Big)dx\\
&\quad+\int_{I_j}(r-r_h)_t\rho_x dx-((r_t-\widehat{(r_h)}_t)\rho^-)_{j+\frac12}+((r_t-\widehat{(r_h)}_t)\rho^+)_{j-\frac12}\\
&\quad+\int_{I_j}(u-u_h)\phi_x dx-((u-\widehat{u}_h)\phi^-)_{j+\frac12}+((u-\widehat{u}_h)\phi^+)_{j-\frac12}\\
&\quad+\int_{I_j}(p-p_h)\varphi_x dx-((p-\widehat{p}_h)\varphi^-)_{j+\frac12}+((p-\widehat{p}_h)\varphi^+)_{j-\frac12}\\
&\quad+\int_{I_j}(u-u_h)\eta_x dx-((u-\widehat{u}_h)\eta^-)_{j+\frac12}+((u-\widehat{u}_h)\eta^+)_{j-\frac12}\\
&\qquad-\int_{I_j}(B(r)-B(r_h))\varphi_x dx+((B(r)-\widehat{B(r_h)})\varphi^-)_{j+\frac12}-((B(r)-\widehat{B(r_h)})\varphi^+)_{j-\frac12}\\
&\quad+\int_{I_j}(b(r)u-b(r_h)u_h)\psi_x dx-((b(r)u-\widehat{b(r_h)}\widetilde{u}_h)\psi^-)_{j+\frac12}+((b(r)u-\widehat{b(r_h)}\widetilde{u}_h)\psi^+)_{j-\frac12}\\
&\qquad-\int_{I_j}(f(u)-f(u_h))\varphi_x dx+((f(u)-\widehat{f})\varphi^-)_{j+\frac12}-((f(u)-\widehat{f})\varphi^+)_{j-\frac12}=0
\end{aligned}$$

for all ρ, ϕ, φ, ψ, $\eta\in V_h$.

Define

(4.11)

$$\begin{aligned}
&\mathcal{B}_j(u-u_h,q-q_h,p-p_h,r-r_h;\rho,\phi,\varphi,\psi,\eta)\\
&\quad=\int_{I_j}(u-u_h)_t\rho dx-\int_{I_j}\Big((q-q_h)_t(\rho+\varphi)+(r-r_h)\phi+(p-p_h)\psi+(r-r_h)\eta\Big)dx\\
&\qquad+\int_{I_j}(r-r_h)_t\rho_x dx-((r_t-\widehat{(r_h)}_t)\rho^-)_{j+\frac12}+((r_t-\widehat{(r_h)}_t)\rho^+)_{j-\frac12}\\
&\qquad+\int_{I_j}(u-u_h)\phi_x dx-((u-\widehat{u}_h)\phi^-)_{j+\frac12}+((u-\widehat{u}_h)\phi^+)_{j-\frac12}\\
&\qquad+\int_{I_j}(p-p_h)\varphi_x dx-((p-\widehat{p}_h)\varphi^-)_{j+\frac12}+((p-\widehat{p}_h)\varphi^+)_{j-\frac12}\\
&\qquad+\int_{I_j}(u-u_h)\eta_x dx-((u-\widehat{u}_h)\eta^-)_{j+\frac12}+((u-\widehat{u}_h)\eta^+)_{j-\frac12},
\end{aligned}$$

(4.12)

$$\mathcal{H}_j(f;u,u_h;\varphi)=\int_{I_j}(f(u)-f(u_h))\varphi_x dx-((f(u)-\widehat{f})\varphi^-)_{j+\frac12}+((f(u)-\widehat{f})\varphi^+)_{j-\frac12},$$

and

(4.13)
$$\begin{aligned}&\mathbf{R}_j(b,B;r,u,r_h,u_h;\varphi,\psi)\\&=\int_{I_j}(B(r)-B(r_h))\varphi_x dx-((B(r)-\widehat{B(r_h)})\varphi^-)_{j+\frac12}+((B(r)-\widehat{B(r_h)})\varphi^+)_{j-\frac12}\\&-\int_{I_j}(b(r)u-b(r_h)u_h)\psi_x dx+((b(r)u-\widehat{b(r_h)}\widetilde{u}_h)\psi^-)_{j+\frac12}-((b(r)u-\widehat{b(r_h)}\widetilde{u}_h)\psi^+)_{j-\frac12}.\end{aligned}$$

Summing over j, the error equation becomes

$$\begin{aligned}&\sum_{j=1}^{N}\mathcal{B}_j(u-u_h,q-q_h,p-p_h,r-r_h;\rho,\phi,\varphi,\psi,\eta)\\&\qquad=\sum_{j=1}^{N}\Big(\mathcal{H}_j(f;u,u_h;\varphi)+\mathbf{R}_j(b,B;r,u,r_h,u_h;\varphi,\psi)\Big)\end{aligned}\tag{4.14}$$

for all ρ, ϕ, φ, ψ, $\eta\in V_h$.

Denoting

$$\begin{aligned}&s=\mathcal{P}^+u-u_h,\qquad &&s^e=\mathcal{P}^+u-u,\\&\xi=\mathcal{P}q-q_h,\qquad &&\xi^e=\mathcal{P}q-q,\\&v=\mathcal{P}p-p_h,\qquad &&v^e=\mathcal{P}p-p,\\&\sigma=\mathcal{P}r-r_h,\qquad &&\sigma^e=\mathcal{P}r-r\end{aligned}\tag{4.15}$$

and taking the test functions

$$\rho=-s,\quad \phi=\sigma_t,\quad \varphi=s,\quad \psi=-\sigma,\quad \eta=v,$$

we obtain the important *energy equality*

$$\begin{aligned}&\sum_{j=1}^{N}\mathcal{B}_j(s-s^e,\xi-\xi^e,v-v^e,\sigma-\sigma^e;-s,\sigma_t,s,-\sigma,v)\\&\qquad=\sum_{j=1}^{N}\Big(\mathcal{H}_j(f;u,u_h;s)+\mathbf{R}_j(b,B;r,u,r_h,u_h;s,-\sigma)\Big).\end{aligned}\tag{4.16}$$

4.2.2. Proof of the main result. In this subsection, we will follow the idea of [31] to present the main proof of Theorem 4.2. We shall prove the theorem by analyzing each term of the energy equation (4.16).

First, we consider the left-hand side of the energy equation (4.16). The estimates for the left-hand side of the energy equation are given in the following lemma. The proof of this lemma will be given in Appendix A.1.

LEMMA 4.3. *The following equation holds*:

$$\begin{aligned}&\sum_{j=1}^{N}\mathcal{B}_j(s-s^e,\xi-\xi^e,v-v^e,\sigma-\sigma^e;-s,\sigma_t,s,-\sigma,v)\\&\qquad=\int_0^1(ss_t+\sigma_t\sigma)dx-\int_0^1 s_t^e s dx-\sum_{j=1}^{N}((\widehat{v^e}+\widehat{\sigma_t^e})[s])_{j+\frac12}.\end{aligned}\tag{4.17}$$

To deal with the nonlinearity of the flux $f(u)$ we would like to make an a priori assumption that, for small enough h, there holds

$$\|u - u_h\| \le h. \tag{4.18}$$

We will justify the validity of this a priori assumption later. For the linear flux $f(u) = cu$, this a priori assumption is unnecessary.

COROLLARY 4.4. *Suppose that the interpolation property* (4.8) *is satisfied; then the a priori assumption* (4.18) *implies that*

$$\|e_u\|_\infty \le Ch^{\frac{1}{2}} \quad \textit{and} \quad \|\mathbb{Q}u - u_h\|_\infty \le Ch^{\frac{1}{2}}, \tag{4.19}$$

where $\mathbb{Q} = \mathcal{P}$ *or* $\mathbb{Q} = \mathcal{P}^{\pm}$ *is the projection operator.*

Next, we consider the right-hand side of the energy equation (4.16). We can rewrite it into the following form:

$$\begin{aligned}
&\sum_{j=1}^{N} \mathcal{H}_j(f; u, u_h; s) = \sum_{j=1}^{N} \int_{I_j} (f(u) - f(u_h)) s_x dx \\
&\quad + \sum_{j=1}^{N} ((f(u) - f(\bar{u}_h))[s])_{j+\frac{1}{2}} + \sum_{j=1}^{N} ((f(\bar{u}_h) - \widehat{f})[s])_{j+\frac{1}{2}}, \qquad (4.20)\\
&\sum_{j=1}^{N} \mathbf{R}_j(b, B; r, u, r_h, u_h; s, -\sigma) \\
&\quad = \sum_{j=1}^{N} \int_{I_j} (B(r) - B(r_h)) s_x dx + \sum_{j=1}^{N} ((B(r) - B(r_h^-))[s])_{j+\frac{1}{2}} \qquad (4.21)\\
&\qquad + \sum_{j=1}^{N} \int_{I_j} (b(r)u - b(r_h)u_h)\sigma_x dx + \sum_{j=1}^{N} ((b(r)u - \widehat{b}(r_h)u_h^+)[\sigma])_{j+\frac{1}{2}},
\end{aligned}$$

where we take into account the periodic boundary condition and recall that the average $\bar{u}_h$ is defined by $\bar{u}_h = \frac{1}{2}(u_h^+ + u_h^-)$.

The estimate for (4.20) is given in the following lemma.

LEMMA 4.5. *Suppose that the interpolation property* (4.8) *is satisfied; then we have the following estimate for* (4.20)*:*

$$\begin{aligned}
&\sum_{j=1}^{N} \mathcal{H}_j(f; u, u_h; s) \qquad (4.22)\\
&\quad \le -\frac{1}{4}\alpha(\widehat{f}; u_h)[s]^2 + (C + C_\star(\|s\|_\infty + h^{-1}\|e_u\|_\infty^2))\|s\|^2 + (C + C_\star h^{-1}\|e_u\|_\infty^2)h^{2k+1}.
\end{aligned}$$

Remark 4.3. For the proof of this lemma, we refer readers to Lemmas 3.4 and 3.5 in [31]. For $f(u) = 2\kappa u + \frac{3}{2}u^2$ in the CH equation (1.1), we have $f'''(u) = 0$ and we do not need the estimate for $\mathcal{T}_6$ in [31].

The estimate for (4.21) is given in the following lemma. The proof of this lemma will be given in Appendix A.2.

LEMMA 4.6. *Suppose that the interpolation property* (4.8) *is satisfied, then we have the following estimate for* (4.21)*:*

$$\left|\sum_{j=1}^{N}\mathbf{R}_j(b,B;r,u,r_h,u_h;s,-\sigma)\right| \le \sum_{j=1}^{N}(|b'(r)u\bar{\sigma^e}[\sigma]|+|b(r)(\sigma^e)^-[s]|)_{j+\frac12}+\frac12\|s\|^2+C\|\sigma\|^2+Ch^{2k+2}. \tag{4.23}$$

Now we are ready to get the final error estimates (4.10). Combining (4.16), (4.17), (4.22), and (4.23), we obtain

$$\begin{aligned}&\int_0^1(s_ts+\sigma_t\sigma)dx+\frac14\alpha(\widehat{f};u_h)[s]^2\\ &\quad\le\int_0^1 s_t^e s\,dx+\sum_{j=1}^{N}((\widehat{v^e}+\widehat{\sigma_t^e})[s])_{j+\frac12}+\sum_{j=1}^{N}(|b'(r)u\bar{\sigma^e}[\sigma]|+|b(r)(\sigma^e)^-[s]|)_{j+\frac12}\\ &\qquad+(C+C_\star(\|s\|_\infty+h^{-1}\|e_u\|_\infty^2))\|s\|^2+(C+C_\star h^{-1}\|e_u\|_\infty^2)h^{2k+1}+C\|\sigma\|^2.\end{aligned}$$

Again by Young's inequality and the interpolation property (4.8), the inequality becomes

$$\begin{aligned}&\int_0^1(s_ts+\sigma_t\sigma)dx+\frac18\alpha(\widehat{f};u_h)[s]^2\\ &\quad\le(C+C_\star(\|s\|_\infty+h^{-1}\|e_u\|_\infty^2))\|s\|^2+(C+C_\star h^{-1}\|e_u\|_\infty^2)h^{2k+1}+Ch^{2k}+C\|\sigma\|^2,\end{aligned}$$

where we use the uniform dissipation property of the numerical flux $\widehat{f}$. Using the results (4.19) implied by the a priori assumption (4.18) and the positive property of $\alpha(\widehat{f};u_h)$, we can get the following error estimate:

$$\frac12\frac{d}{dt}\int_0^1(s^2+\sigma^2)dx\le C(\|s\|^2+\|\sigma\|^2)+Ch^{2k}.$$

Thus Theorem 4.2 follows by the triangle inequality and the interpolating property (4.8).

To complete the proof, let us verify the a priori assumption (4.18). For $k\ge2$, we can consider h small enough so that $Ch^k<\frac12 h$, where C is the constant in (4.10) determined by the final time T. Then, if $t^*=\sup\{t:\|u(t)-u_h(t)\|\le h\}$, we would have $\|u(t^*)-u_h(t^*)\|=h$ by continuity if t^* is finite. On the other hand, our proof implies that (4.10) holds for $t\le t^*$, in particular $\|u(t^*)-u_h(t^*)\|\le Ch^k<\frac12 h$. This is a contradiction if $t^*<T$. Hence $t^*\ge T$ and our a priori assumption (4.18) is justified.

5. Numerical results. In this section we provide numerical examples to illustrate the accuracy and capability of the method. Time discretization is by the third-order explicit TVD Runge–Kutta method in [26]. This is not the most efficient method for the time discretization to our LDG scheme. However, we will not address the issue of time discretization efficiency in this paper. We have verified with the aid of successive mesh refinements that, in all cases, the results shown are numerically convergent. We will give the numerical test results for the CH equation

$$u_t-u_{xxt}+3uu_x=2u_xu_{xx}+uu_{xxx} \tag{5.1}$$

TABLE 5.1
Accuracy test for the CH equation (5.1) *with the exact solution* (5.2). *Periodic boundary condition. Uniform meshes with* N *cells at time* $t = 0.5$.

		$u - u_h$				$r - r_h$			
	N	L^2 error	Order	L^∞ error	Order	L^2 error	Order	L^∞ error	Order
P^0	10	1.42E-01	—	3.08E-01	—	1.42E-01	—	3.08E-01	—
	20	7.95E-02	0.84	1.77E-01	0.80	7.95E-02	0.83	1.77E-01	0.57
	40	4.23E-02	0.91	9.41E-01	0.91	4.23E-02	0.94	9.41E-02	0.87
	80	2.18E-02	0.95	4.83E-02	0.96	2.18E-02	0.98	4.83E-02	0.97
P^1	10	1.16E-02	—	6.63E-02	—	1.16E-02	—	6.63E-02	—
	20	3.12E-03	1.90	1.86E-02	1.84	3.12E-03	0.68	1.86E-02	0.24
	40	8.05E-04	1.95	4.76E-03	1.96	8.05E-04	0.85	4.76E-03	0.63
	80	2.04E-04	1.98	1.19E-02	2.00	2.04E-04	0.93	1.19E-03	0.87
P^2	10	1.41E-03	—	6.75E-03	—	1.41E-03	—	6.75E-03	—
	20	1.49E-04	3.24	9.06E-04	2.90	1.49E-04	2.64	9.06E-04	2.64
	40	1.70E-05	3.13	9.85E-05	3.20	1.70E-05	2.06	9.85E-05	1.45
	50	8.95E-06	2.88	4.96E-05	3.07	8.95E-06	1.95	4.96E-05	1.77

with different initial conditions.

Example 5.1. Smooth solution. In this example, we test the scheme with smooth traveling waves. Smooth traveling waves are solutions of the form

$$u(x,t) = \phi(x - ct), \tag{5.2}$$

where ϕ is a solution of the second-order ODE

$$\phi_{xx} = \phi - \frac{\alpha}{(\phi - c)^2}. \tag{5.3}$$

In order to get a smooth traveling wave, we choose the constants c and α as in [16], i.e., $\alpha = c = 3$. The initial conditions for ϕ is

$$\phi(0) = 1, \quad \frac{d\phi}{dx}(0) = 0. \tag{5.4}$$

It gives rise to a smooth traveling wave with period $a \simeq 6.46954603635$. We use fourth-order explicit Runge–Kutta method with 100000 points to approximate the solution of (5.3). This high-precision solution is used as a reference solution for the smooth traveling wave. The L^2 and L^∞ errors and the numerical orders of accuracy for u at time $t = 0.5$ with uniform meshes are contained in Table 5.1. Periodic boundary conditions are used. We can see that the method with P^k elements gives a uniform $(k+1)$th order of accuracy for u in both norms. For the solution r_h, the accuracy is of kth order in both norms for $k \geq 1$ and this indicates that our error estimates result for r is sharp.

Example 5.2. Accuracy test. The peakon solutions of the CH equation (5.1) are well known and are the only traveling waves for which there is a simple explicit formula. The peaked traveling wave solution is

$$u(x,t) = ce^{-|x-ct|}, \tag{5.5}$$

where c is the wave speed. We give the accuracy test with $c = 0.25$. The accuracy is measured in smooth parts of the solution, 1/5 of the computational domain away from the peak. The L^2 and L^∞ errors and the numerical orders of accuracy for u at time $t = 1$ with uniform meshes in $[-25, 25]$ are contained in Table 5.2. We can see

TABLE 5.2
Accuracy test for the CH equation (5.1) *with the exact solution* (5.5). *Periodic boundary condition.* $c = 0.25$. *Uniform meshes with* N *cells at time* $t = 1$.

	N	L^2 error	Order	L^∞ error	Order
P^0	10	8.71E-03	—	1.90E-02	—
	20	2.18E-03	2.00	4.95E-03	1.94
	40	9.58E-04	1.18	2.36E-03	1.07
	80	4.08E-04	1.23	1.19E-03	0.98
P^1	10	1.45E-02	—	3.17E-02	—
	20	8.33E-04	4.12	2.06E-03	3.94
	40	1.14E-04	2.87	3.74E-04	2.46
	80	1.80E-05	2.67	8.82E-05	2.08
P^2	10	1.60E-02	—	3.50E-02	—
	20	4.05E-04	5.31	1.23E-03	4.83
	40	2.81E-05	3.85	9.65E-05	3.68
	80	3.54E-06	2.99	1.29E-05	2.90

that the method with P^k elements gives a uniform $(k+1)$th order of accuracy for u in both norms.

Example 5.3. Peakon solution. In this example, we present the wave propagation of the periodized version of the solution (5.5). In the single peak case, the initial condition is

$$u_0(x) = \begin{cases} \dfrac{c}{\cosh(a/2)} \cosh(x - x_0), & |x - x_0| \le a/2, \\ \dfrac{c}{\cosh(a/2)} \cosh(a - (x - x_0)), & |x - x_0| > a/2, \end{cases} \tag{5.6}$$

where x_0 is the position of the trough and a is the period. We present the wave propagation for the CH equation with parameters $c = 1$, $a = 30$, and $x_0 = -5$. The computational domain is $[0, a]$. In Figure 1, the peak profile at $t = 0$, 5, 10 and the space time graph of the solutions up to $t = 10$ are shown. The lack of smoothness at the peak of peakon introduces high-frequency dispersive errors into the calculation and will cause numerical oscillation near the peak. In our computation of the LDG method, we use the P^5 element with $N = 320$ cells to resolve the peak. We can see clearly that the moving peak profile is resolved very well.

Example 5.4. Two-peakon interaction. In this example, we consider the two-peakon interaction of the CH equation with the initial condition

$$u_0(x) = \phi_1(x) + \phi_2(x), \tag{5.7}$$

where

$$\phi_i(x) = \begin{cases} \dfrac{c_i}{\cosh(a/2)} \cosh(x - x_i), & |x - x_i| \le a/2, \\ \dfrac{c_i}{\cosh(a/2)} \cosh(a - (x - x_i)), & |x - x_i| > a/2, \end{cases} \quad i = 1, 2. \tag{5.8}$$

The parameters are $c_1 = 2$, $c_2 = 1$, $x_1 = -5$, $x_2 = 5$, $a = 30$. The computational domain is $[0, a]$. We use the P^5 element with $N = 320$ cells in our computation of the LDG method. In Figure 2, the two-peakon interaction at $t = 0$, 5, 12, and 18 is shown. We can see clearly that the moving peak interaction is resolved very well.

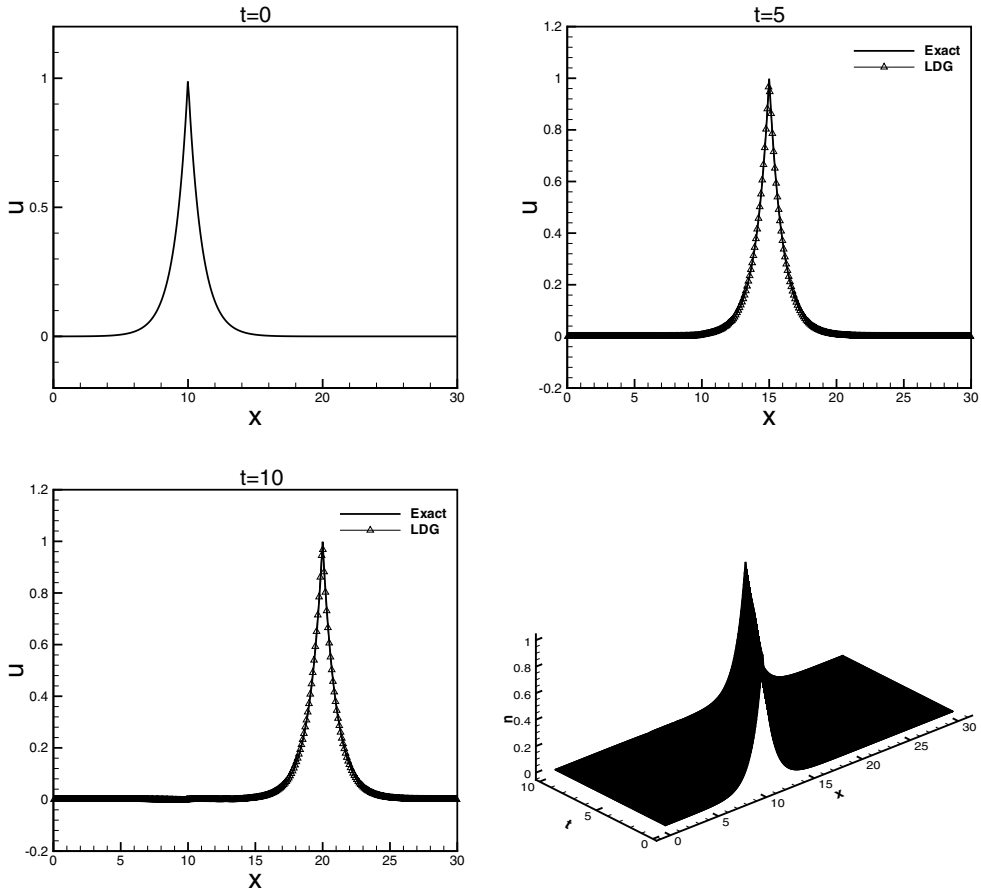


FIG. 1. *The peak profile of the CH equation* (5.1) *with the initial condition* (5.6). *Periodic boundary condition in* $[0, 30]$. P^5 *elements and a uniform mesh with* $N = 320$ *cells.*

Example 5.5. Three-peakon interaction. In this example, we consider the three-peakon interaction of the CH equation with the initial condition

$$u_0(x) = \phi_1(x) + \phi_2(x) + \phi_3(x), \tag{5.9}$$

where ϕ_i, $i = 1, 2, 3$, are defined as in (5.8). The parameters are $c_1 = 2$, $c_2 = 1$, $c_3 = 0.8$, $x_1 = -5$, $x_2 = -3$, $x_3 = -1$, $a = 30$. The computational domain is $[0, a]$. We use the P^5 element with $N = 320$ cells in our computation of the LDG method. In Figure 3, the three-peakon interaction at $t = 0$, 1, 2, 3, 4, and 6 is shown. We can see clearly that the moving peak interaction is resolved very well.

Example 5.6. Solution with a discontinuous derivative. In this example, we consider an initial data function u_0 which has a discontinuous derivative as in [16]. The initial condition is

$$u_0(x) = \frac{10}{(3 + |x|)^2}. \tag{5.10}$$

The computational domain is $[-30, 30]$. We use the P^2 element with $N = 640$ cells in our computation of the LDG method. The solutions at time $t = 5$, 10, 15, and 20 are

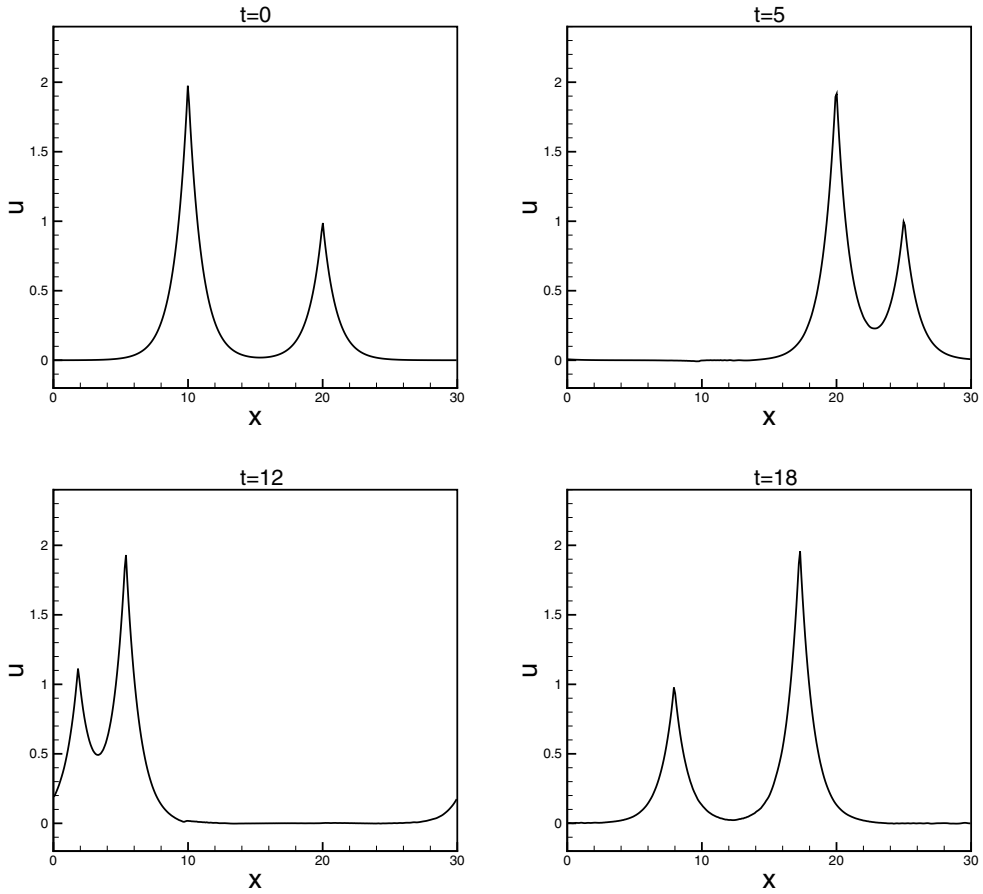


FIG. 2. *The two-peakon interaction of the CH equation* (5.1) *with the initial condition* (5.7). *Periodic boundary condition in* $[0, 30]$. P^5 *elements and a uniform mesh with* $N = 320$ *cells.*

shown in Figure 4. Even if we do not use higher-order polynomials and more cells, we still obtain a good resolution of the solution comparable with that in [16].

Example 5.7. Breakup of the plateau traveling wave. In this example, we consider a cutoff peakon in [1], i.e., a plateau function $u(x,t) = \phi(x - ct)$ with

$$\phi(x) = \begin{cases} ce^{x+k}, & x \le -k, \\ c, & |x| \le k, \\ ce^{-x+k}, & x \ge k. \end{cases} \tag{5.11}$$

We put $c = 0.6$ and $k = 5$. The computational domain is $[-40, 40]$. We use the P^2 element with $N = 800$ cells in our computation of the LDG method. The breakup of the plateau traveling wave is shown in Figure 5 at different time. The solution is resolved very well compared with the result in [1].

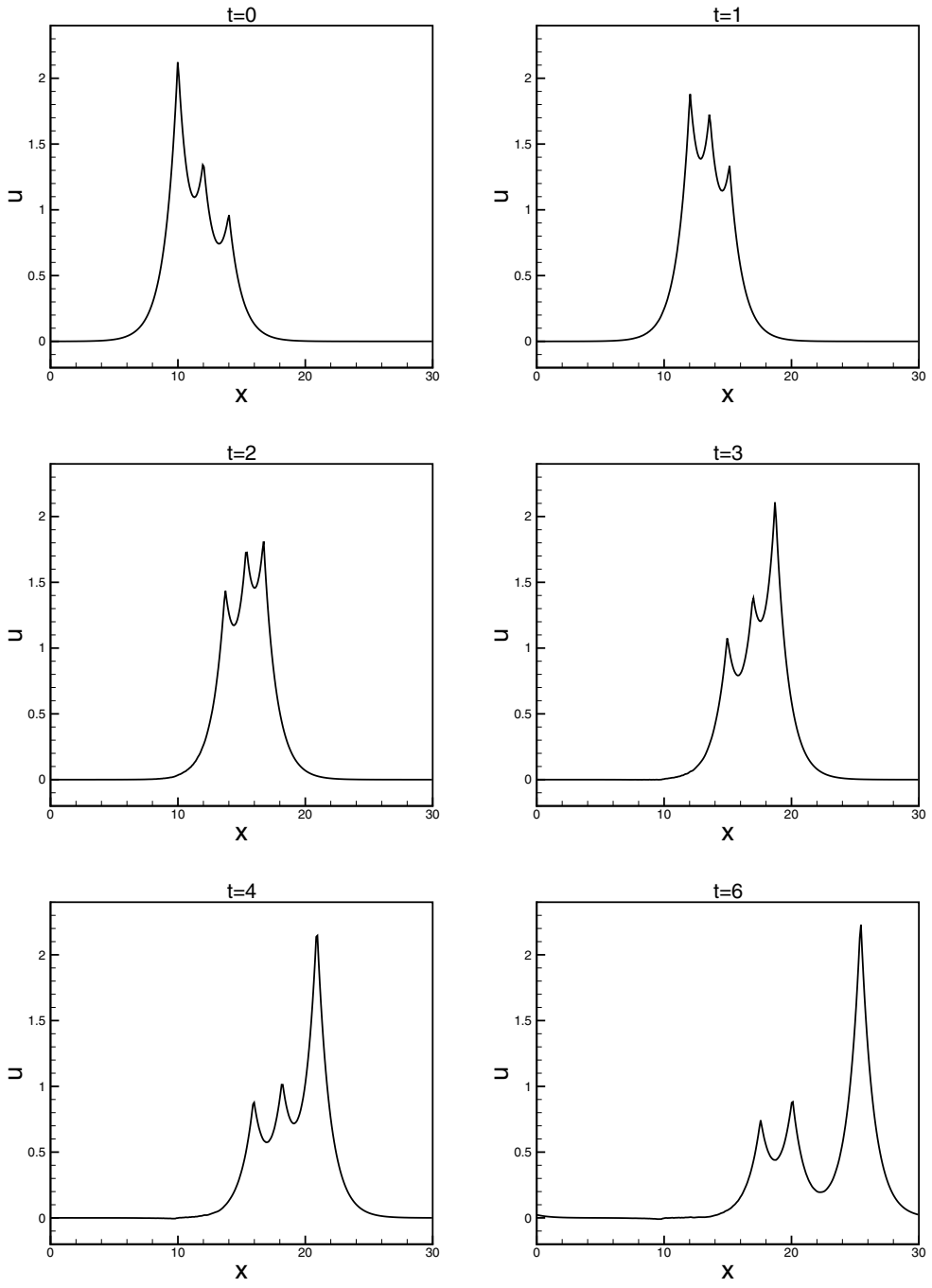


FIG. 3. *The three-peakon interaction of the CH equation* (5.1) *with the initial condition* (5.9). *Periodic boundary condition in* $[0, 30]$. P^5 *elements and a uniform mesh with* $N = 320$ *cells.*

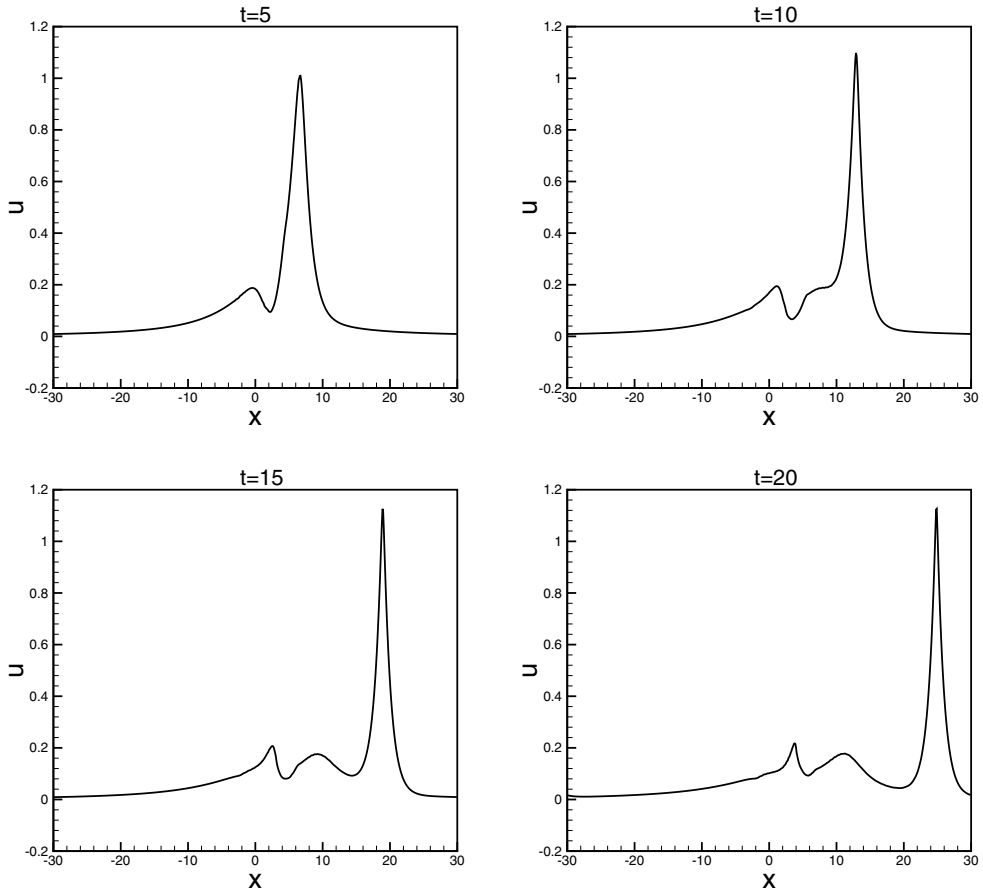


FIG. 4. *The solution with discontinuous derivative of the CH equation* (5.1) *with the initial condition* (5.10). *Periodic boundary condition in* $[-30, 30]$. P^2 *elements and a uniform mesh with* $N = 640$ *cells.*

6. Conclusion. We have developed an LDG method to solve the CH equation. L^2 stability is proven for general solutions, and an a priori error estimate is obtained for smooth solutions. Numerical examples are given to illustrate the accuracy and capability of the methods. Although not addressed in this paper, the LDG methods are flexible for general geometry, unstructured meshes and *hp*-adaptivity, and have excellent parallel efficiency. The LDG method has a good potential in solving the CH equation and similar nonlinear equations in mathematical physics.

Appendix A. Proof of several lemmas.

A.1. Proof of Lemma 4.3.

$$\begin{aligned}(\text{A.1})\quad &\mathcal{B}_j(s - s^e, \xi - \xi^e, v - v^e, \sigma - \sigma^e; -s, \sigma_t, s, -\sigma, v)\\ &= \mathcal{B}_j(s, \xi, v, \sigma; -s, \sigma_t, s, -\sigma, v) - \mathcal{B}_j(s^e, \xi^e, v^e, \sigma^e; -s, \sigma_t, s, -\sigma, v).\end{aligned}$$

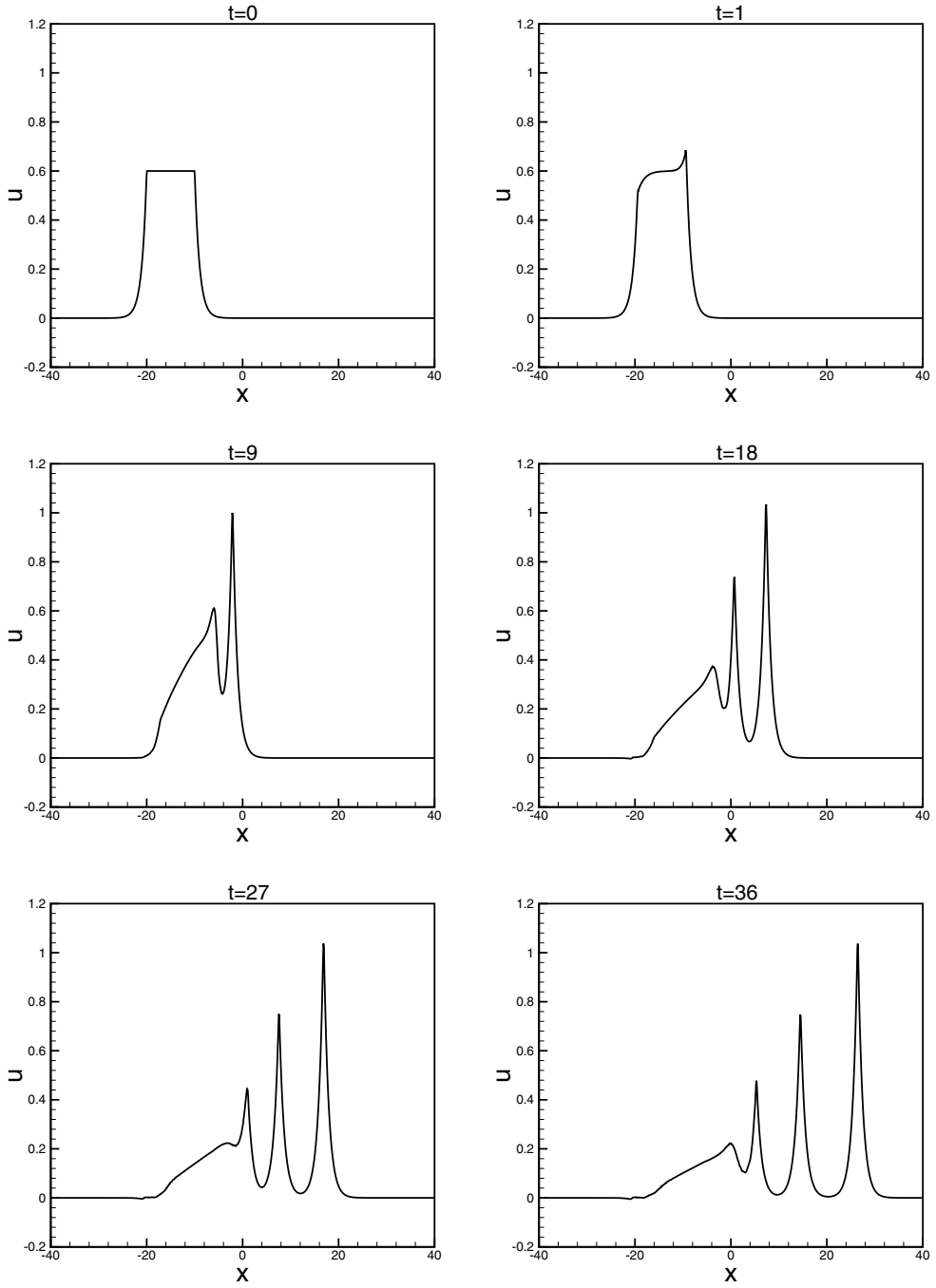


FIG. 5. *The break up of plateau traveling wave of the CH equation* (5.1) *with the initial condition* (5.11). *Periodic boundary condition in* $[-40, 40]$. P^2 *elements and a uniform mesh with* $N = 800$ *cells.*

By the same argument as that used for the L^2 stability in Proposition 3.1, the first term of the right-hand side in (A.1) becomes

$$\mathcal{B}_j(s,\xi,v,\sigma;-s,\sigma_t,s,-\sigma,v)=\int_{I_j}(s_ts+\sigma_t\sigma)dx+\Psi_{j+\frac{1}{2}}-\Psi_{j-\frac{1}{2}}, \tag{A.2}$$

where $\Psi=v^-s^--\widehat{v}s^--\widetilde{s}v^-+\sigma_t^-s^--\widehat{\sigma}_ts^--\widehat{s}\sigma_t^-$.

As to the second term of the right-hand side in (A.1), we have

(A.3)

$$\begin{aligned}&\mathcal{B}_j(s^e,\xi^e,v^e,\sigma^e;-s,\sigma_t,s,-\sigma,v)\\&\quad=\int_{I_j}(s_t^es+\sigma^e\sigma_t)dx+\int_{I_j}(\sigma^ev-v^e\sigma)dx+\int_{I_j}(\sigma_t^es_x+s^e(\sigma_t)_x+v^es_x+s^ev_x)dx\\&\qquad+((\widehat{v^e}+\widehat{\sigma_t^e})[s])_{j-\frac{1}{2}}+(\widehat{s^e}[\sigma_t])_{j-\frac{1}{2}}+(\widetilde{s^e}[v])_{j-\frac{1}{2}}+\Phi_{j+\frac{1}{2}}-\Phi_{j-\frac{1}{2}},\end{aligned}$$

where $\Phi=-\widehat{v^e}s^--\widetilde{s^e}v^--\widehat{\sigma_t^e}s^--\widehat{s^e}\sigma_t^-$. Because $\mathcal{P}$ is a local L^2 projection, and $\mathcal{P}^+$, even though not a local L^2 projection, does have the property that $s-\mathcal{P}^+s$ is locally orthogonal to all polynomials of degree up to $k-1$, we have

$$\int_{I_j}\sigma^e\sigma_tdx+\int_{I_j}(\sigma^ev-v^e\sigma)dx+\int_{I_j}(\sigma_t^es_x+s^e(\sigma_t)_x+v^es_x+s^ev_x)dx=0.$$

Noticing the special interpolating property of the projection $\mathcal{P}^-$, we also have

$$(\widehat{s^e}[\sigma_t])_{j-\frac{1}{2}}+(\widetilde{s^e}[v])_{j-\frac{1}{2}}=0.$$

Equation (A.3) then becomes

$$\mathcal{B}_j(s^e,\xi^e,v^e,\sigma^e;-s,\sigma_t,s,-\sigma,v)=\int_{I_j}s_t^esdx+((\widehat{v^e}+\widehat{\sigma_t^e})[s])_{j-\frac{1}{2}}+\Phi_{j+\frac{1}{2}}-\Phi_{j-\frac{1}{2}}.$$

Combining the above equation with (A.2), summing over j, and taking into account the periodic boundary condition, we obtain the desired equality (4.17).

A.2. Proof of Lemma 4.6. The proof of Lemma 4.6 is similar to that of Lemma 3.5 in [31]. The main difference is that we take $\bar{r}_h=\frac{1}{2}(r_h^++r_h^-)$, and u_h^+ and r_h^- as the reference values of the functions r_h and u_h at each boundary point. For the nonlinear terms $b(r)u$ and $B(r)$, we use the following Taylor expansions:

$$\begin{aligned}&B(r)-B(r_h)=b(r)\sigma-\frac{1}{2}b'(r)\sigma^2-b(r)\sigma^e+b'(r)\sigma\sigma^e-\frac{1}{2}b'(r)(\sigma^e)^2,\\&B(r)-B(r_h^-)=b(r)\sigma^--\frac{1}{2}b'(r)(\sigma^-)^2-b(r)(\sigma^e)^-+b'(r)\sigma^-(\sigma^e)^--\frac{1}{2}b'(r)((\sigma^e)^-)^2,\end{aligned}$$

where we use the properties $B'(r)=b(r)$ and $b^{(m)}(r)=0, m\geq 2$,

$$\begin{aligned}&b(r)u-b(r_h)u_h\\&\quad=b'(r)u\sigma+b(r)s-b'(r)s\sigma-b'(r)u\sigma^e-b(r)s^e+b'(r)s\sigma^e+b'(r)\sigma s^e-b'(r)\sigma^es^e,\\&b(r)u-\widehat{b}(r_h)u_h^+\\&\quad=b'(r)u\bar{\sigma}+b(r)s^+-b'(r)\bar{\sigma}s^+-b'(r)u\bar{\sigma}^e-b(r)(s^e)^+\end{aligned}$$

$$+ b'(r)s^+\bar{\sigma}^e + b'(r)\bar{\sigma}(s^e)^+ - b'(r)\bar{\sigma}^e(s^e)^+.$$

These imply the following representation:

$$\sum_{j=1}^{N} \mathbf{R}_j(b, B; r, u, r_h, u_h; s, -\sigma) = \mathcal{S}_1 + \mathcal{S}_2 + \mathcal{S}_3 + \mathcal{S}_4 + \mathcal{S}_5, \tag{A.4}$$

where

$$\begin{aligned}
\mathcal{S}_1 &= \sum_{j=1}^{N} \int_{I_j} b'(r)u\sigma\sigma_x dx + \sum_{j=1}^{N} (b'(r)u\bar{\sigma}[\sigma])_{j+\frac{1}{2}} \\
&\quad + \sum_{j=1}^{N} \int_{I_j} b(r)(s\sigma)_x dx + \sum_{j=1}^{N} (b(r)(s^+[\sigma] + \sigma^-[s]))_{j+\frac{1}{2}}, \\
\mathcal{S}_2 &= -\frac{1}{2}\left(\sum_{j=1}^{N} \int_{I_j} b'(r)(s\sigma^2)_x dx + \sum_{j=1}^{N} (b'(r)(2s^+\bar{\sigma}[\sigma] + (\sigma^-)^2[s]))_{j+\frac{1}{2}} \right), \\
\mathcal{S}_3 &= -\sum_{j=1}^{N} \int_{I_j} b'(r)u\sigma^e\sigma_x dx - \sum_{j=1}^{N} (b'(r)u\bar{\sigma}^e[\sigma])_{j+\frac{1}{2}} \\
&\quad - \sum_{j=1}^{N} \int_{I_j} b(r)(\sigma^e s_x + s^e\sigma_x) dx - \sum_{j=1}^{N} (b(r)((s^e)^+[\sigma] + (\sigma^e)^-[s]))_{j+\frac{1}{2}}, \\
\mathcal{S}_4 &= \sum_{j=1}^{N} \int_{I_j} b'(r)\sigma^e(s\sigma)_x dx + \sum_{j=1}^{N} (b'(r)(s^+\bar{\sigma}^e[\sigma] + (\sigma^e)^-\sigma^-[s]))_{j+\frac{1}{2}} \\
&\quad + \sum_{j=1}^{N} \int_{I_j} b'(r)s^e\sigma\sigma_x dx + \sum_{j=1}^{N} (b'(r)\bar{\sigma}(s^e)^+[\sigma])_{j+\frac{1}{2}}, \\
\mathcal{S}_5 &= -\frac{1}{2}\left(\sum_{j=1}^{N} \int_{I_j} b'(r)((\sigma^e)^2 s_x + 2\sigma^e s^e\sigma_x) dx \right. \\
&\quad \left. + \sum_{j=1}^{N} (b'(r)(2(s^e)^+\bar{\sigma}^e[\sigma] + ((\sigma^e)^-)^2[s]))_{j+\frac{1}{2}} \right)
\end{aligned}$$

will be estimated separately.

- The $\mathcal{S}_1$ term.
  After a simple integration by parts, it is easy to obtain

$$\mathcal{S}_1 = -\frac{1}{2}\sum_{j=1}^{N} \int_{I_j} (b'(r)u)_x\sigma^2 dx - \sum_{j=1}^{N} \int_{I_j} (b(r))_x s\sigma dx \le C\|\sigma\|^2 + \frac{1}{8}\|s\|^2. \tag{A.5}$$

- The $\mathcal{S}_2$ term.
  After a simple integration by parts, it is easy to obtain

$$\mathcal{S}_2 = \frac{1}{2}\sum_{j=1}^{N} \int_{I_j} (b'(r))_x s\sigma^2 dx = 0. \tag{A.6}$$

- The $\mathcal{S}_3$ term.
  We can rewrite $\mathcal{S}_3$ into the following form

$$\begin{aligned}\mathcal{S}_3 = &-\sum_{j=1}^{N}\int_{I_j}(b'(r)u - b'(r_j)u_j)\sigma^e\sigma_x dx - \sum_{j=1}^{N}\int_{I_j} b'(r_j)u_j\sigma^e\sigma_x dx\\ &-\sum_{j=1}^{N}\int_{I_j}(b(r)-b(r_j))(\sigma^e s_x + s^e\sigma_x)dx - \sum_{j=1}^{N}\int_{I_j} b(r_j)(\sigma^e s_x + s^e\sigma_x)dx\\ &-\sum_{j=1}^{N}(b'(r)u\bar{\sigma}^e[\sigma])_{j+\frac{1}{2}} - \sum_{j=1}^{N}(b(r)(\sigma^e)^-[s])_{j+\frac{1}{2}} - \sum_{j=1}^{N}(b(r)(s^e)^+[\sigma])_{j+\frac{1}{2}}.\end{aligned}$$

  The second, the fourth, and the last terms in the above equation are zero by the definition of the special projection. Since $|b'(r)u - b'(r_j)u_j| = O(h)$ on each element I_j, then by the inverse property (i) in (4.9), together with the interpolation property (4.8), the first term in the above equation is estimated by

$$\left|\sum_{j=1}^{N}\int_{I_j}(b'(r)u - b'(r_j)u_j)\sigma^e\sigma_x dx\right| \le C\|\sigma^e\|\|\sigma\| \le C\|\sigma\|^2 + Ch^{2k+2}.$$

  By the same argument we can also get the estimate for the third term

$$\left|\sum_{j=1}^{N}\int_{I_j}(b(r)-b(r_j))(\sigma^e s_x + s^e\sigma_x)dx\right| \le C\|\sigma\|^2 + \frac{1}{8}\|s\|^2 + Ch^{2k+2}.$$

  Now we can get the estimate of $\mathcal{S}_3$,

$$\mathcal{S}_3 \le \sum_{j=1}^{N}(|b'(r)u\bar{\sigma^e}[\sigma]| + |b(r)(\sigma^e)^-[s]|)_{j+\frac{1}{2}} + C\|\sigma\|^2 + \frac{1}{8}\|s\|^2 + Ch^{2k+2}. \tag{A.7}$$

- $\mathcal{S}_4$ and $\mathcal{S}_5$ terms.
  Because $\mathcal{S}_4$ and $\mathcal{S}_5$ are high-order terms in the Taylor expansion, it is easy to show by Young's inequality and the inverse properties (i) and (ii) in (4.9) that

$$\mathcal{S}_4 \le Ch^{-1}(\|\sigma^e\|_\infty + \|s^e\|_\infty)\|\sigma\|\|s\| \le C\|\sigma\|^2 + \frac{1}{8}\|s\|^2, \tag{A.8}$$

$$\mathcal{S}_5 \le Ch^{-1}\|\sigma^e\|_\infty(\|\sigma^e\|\|s\| + \|s^e\|\|\sigma\|) \le C\|\sigma\|^2 + \frac{1}{8}\|s\|^2 + Ch^{2k+2}. \tag{A.9}$$

Therefore, summing up the above estimates from (A.5) to (A.9), we complete the proof of Lemma 4.6.

REFERENCES

[1] R. Artebrant and H. J. Schroll, *Numerical simulation of Camassa–Holm peakons by adaptive upwinding*, Appl. Numer. Math., 56 (2006), pp. 695–711.

[2] A. Bressan and A. Constantin, *Global conservative solutions of the Camassa–Holm equation*, Arch. Ration Mech. Anal., 183 (2007), pp. 215–239.

[3] A. Bressan and A. Constantin, *Global dissipative solutions of the Camassa–Holm equation*, Anal. Appl., 5 (2007), pp. 1–27.
[4] R. Camassa and D. Holm, *An integrable shallow water equation with peaked solitons*, Phys. Rev. Lett., 71 (1993), pp. 1661–1664.
[5] R. Camassa, D. Holm, and J. Hyman, *A new integrable shallow water equation*, Adv. Appl. Mech., 31 (1994), pp. 1–33.
[6] P. Ciarlet, *The Finite Element Method for Elliptic Problem*, North–Holland, Amsterdam, 1975.
[7] B. Cockburn, *Discontinuous Galerkin methods for methods for convection-dominated problems*, in High-Order Methods for Computational Physics, T. J. Barth and H. Deconinck, eds., Lecture Notes in Computational Science and Engineering 9, Springer, 1999, pp. 69–224.
[8] B. Cockburn and C.-W. Shu, *TVB Runge–Kutta local projection discontinuous Galerkin finite element method for conservation laws* II: *General framework*, Math. Comp., 52 (1989), pp. 411–435.
[9] B. Cockburn and C.-W. Shu, *The local discontinuous Galerkin method for time-dependent convection-diffusion systems*, SIAM J. Numer. Anal., 35 (1998), pp. 2440–2463.
[10] B. Cockburn and C.-W. Shu, *Runge–Kutta discontinuous Galerkin methods for convection-dominated problems*, J. Sci. Comput., 16 (2001), pp. 173–261.
[11] B. Cockburn and C.-W. Shu, *Foreword for the special issue on discontinuous Galerkin method*, J. Sci. Comput., 22/23 (2005), pp. 1–3.
[12] C. Dawson, *Foreword for the special issue on discontinuous Galerkin method*, Comput. Methods Appl. Mech. Engrg., 195 (2006), p. 3183.
[13] C. Eskilsson and S. J. Sherwin, *A discontinuous spectral element model for Boussinesq-type equations*, J. Sci. Comput., 17 (2002), pp. 143–152.
[14] C. Eskilsson and S. J. Sherwin, *Discontinuous Galerkin spectral/hp element modelling of dispersive shallow water systems*, J. Sci. Comput., 22/23 (2005), pp. 269–288.
[15] C. Eskilsson and S. J. Sherwin, *Spectral/hp discontinuous Galerkin methods for modelling 2D Boussinesq equations*, J. Comput. Phys., 212 (2006), pp. 566–589.
[16] H. Holden and X. Raynaud, *A convergent numerical scheme for the Camassa–Holm equation based on multipeakons*, Discrete Contin. Dyn. Syst., 14 (2006), pp. 505–523.
[17] H. Holden and X. Raynaud, *Convergence of a finite difference scheme for the Camassa–Holm equation*, SIAM J. Numer. Anal., 44 (2006), pp. 1655–1680.
[18] R. S. Johnson, *Camassa–Holm, Korteweg-de Vries and related models for water waves*, J. Fluid Mech., 455 (2002), pp. 63–82.
[19] R. S. Johnson, *On solutions of the Camassa–Holm equation*, R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci., 459 (2003), pp. 1687–1708.
[20] H. Kalisch and J. Lenells, *Numerical study of traveling-wave solutions for the Camassa–Holm equation*, Chaos Solitons Fractals, 25 (2005), pp. 287–298.
[21] H. Kalisch and X. Raynaud, *Convergence of a spectral projection of the Camassa–Holm equation*, Numer. Methods Partial Differential Equations, 22 (2006), pp. 1197–1215.
[22] J. Lenells, *Traveling wave solutions of the Camassa–Holm equation*, J. Differential Equations, 217 (2005), pp. 393–430.
[23] J. Lenells, *Conservation laws of the Camassa–Holm equation*, J. Phys. A, 38 (2005), pp. 869–880.
[24] D. Levy, C.-W. Shu, and J. Yan, *Local discontinuous Galerkin methods for nonlinear dispersive equations*, J. Comput. Phys., 196 (2004), pp. 751–772.
[25] Y. S. Li and J. E. Zhang, *The multiple-soliton solution of the Camassa–Holm equation*, Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 460 (2004), pp. 2617–2627.
[26] C.-W. Shu and S. Osher, *Efficient implementation of essentially nonoscillatory shock-capturing schemes*, J. Comput. Phys., 77 (1988), pp. 439–471.
[27] Y. Xu and C.-W. Shu, *Local discontinuous Galerkin methods for three classes of nonlinear wave equations*, J. Comput. Math., 22 (2004), pp. 250–274.
[28] Y. Xu and C.-W. Shu, *Local discontinuous Galerkin methods for nonlinear Schrödinger equations*, J. Comput. Phys., 205 (2005), pp. 72–97.
[29] Y. Xu and C.-W. Shu, *Local discontinuous Galerkin methods for two classes of two dimensional nonlinear wave equations*, Phys. D, 208 (2005), pp. 21–58.
[30] Y. Xu and C.-W. Shu, *Local discontinuous Galerkin methods for the Kuramoto–Sivashinsky equations and the Ito-type coupled KdV equations*, Comput. Methods Appl. Mech. Engrg., 195 (2006), pp. 3430–3447.
[31] Y. Xu and C.-W. Shu, *Error estimates of the semi-discrete local discontinuous Galerkin method for nonlinear convection-diffusion and KdV equations*, Comput. Methods Appl.

Mech. Engrg., 196 (2007), pp. 3805–3822.

[32] J. Yan and C.-W. Shu, *A local discontinuous Galerkin method for KdV type equations*, SIAM J. Numer. Anal., 40 (2002), pp. 769–791.

[33] J. Yan and C.-W. Shu, *Local discontinuous Galerkin methods for partial differential equations with higher order derivatives*, J. Sci. Comput., 17 (2002), pp. 27–47.

[34] Q. Zhang and C.-W. Shu, *Error estimates to smooth solutions of Runge–Kutta discontinuous Galerkin methods for scalar conservation laws*, SIAM J. Numer. Anal., 42 (2004), pp. 641–666.

SIAM J. NUMER. ANAL.
Vol. 46, No. 4, pp. 2022–2038
© 2008 Society for Industrial and Applied Mathematics

HIGH-ORDER EXPONENTIAL OPERATOR SPLITTING METHODS FOR TIME-DEPENDENT SCHRÖDINGER EQUATIONS*

MECHTHILD THALHAMMER†

Abstract. In this paper, we deduce high-order error bounds for exponential operator splitting methods. The employed techniques are specific to linear differential equations of the form $u'(t) = A\,u(t) + B\,u(t)$, $t \geq 0$, involving an unbounded operator A. In particular, evolutionary Schrödinger equations with sufficiently regular initial values are included in the analysis.

Key words. exponential operator splitting, Schrödinger equations, high-order methods, convergence, stability

AMS subject classifications. 65L05, 65M12, 65J10

DOI. 10.1137/060674636

1. Introduction.

In this paper, we study exponential operator splitting methods for initial value problems of the form

$$u'(t) = A\,u(t) + B\,u(t), \quad t \geq 0, \qquad u(0) \text{ given.} \tag{1.1}$$

We are primarily interested in evolution equations that are related to time-dependent linear Schrödinger equations or spatial discretizations thereof. That is, we allow the operator norm of A to be of arbitrary size and suppose B to be a bounded linear operator.

For the time integration of (1.1), we consider exponential operator splitting methods composed by several exponentials

$$u_n = \prod_{j=1}^{s} \mathrm{e}^{\,b_j h B}\, \mathrm{e}^{a_j h A}\, u_{n-1}, \quad n \geq 1, \qquad u_0 \text{ given,} \tag{1.2}$$

with coefficients $a_j, b_j \in \mathbb{R}$ for $1 \leq j \leq s$. In particular, the symmetric second-order splitting scheme

$$u_n = \mathrm{e}^{\frac{1}{2} h B}\, \mathrm{e}^{h A}\, \mathrm{e}^{\frac{1}{2} h B}\, u_{n-1}, \quad n \geq 1, \qquad u_0 \text{ given,} \tag{1.3}$$

referred to as *Strang* [20] or *symmetric Trotter* [21] splitting, is contained in the method class (1.2).

In Jahnke and Lubich [11] error bounds for (1.3) when applied to pseudospectral discretizations of time-dependent linear Schrödinger equations are given. In the present paper, we extend this approach to splitting methods of the general form (1.2).

So far, despite the fact that exponential operator splitting methods are widely used in the time integration of partial differential equations, it remains open to provide a convergence analysis for the numerical method class (1.2) when applied to stiff problems. In this work, we deduce a theoretical result on the convergence and stability behavior of exponential operator splitting methods that contributes to filling the blank

*Received by the editors November 10, 2006; accepted for publication (in revised form) June 27, 2007; published electronically April 23, 2008.
http://www.siam.org/journals/sinum/46-4/67463.html
†Institut für Mathematik, Leopold-Franzens-Universität Innsbruck, Technikerstrasse 13/VII, A–6020 Innsbruck, Austria (Mechthild.Thalhammer@uibk.ac.at).

for evolutionary linear Schrödinger equations and to justifying the practical use of example methods proposed in the literature for this type of problem; for instance, our error bound is applicable to symmetric symplectic fourth- and sixth-order methods, given in Blanes and Moan [3], and it shows that the schemes retain their convergence orders for time-dependent linear Schrödinger equations with sufficiently regular data.

This paper is organized as follows. In section 2, we introduce the precise assumptions on the linear evolution equation (1.1). Our abstract framework is based on the theory of $\mathscr{C}_0$-(semi)groups on Banach spaces and includes evolutionary linear Schrödinger equations; a thorough treatment of one-parameter (semi)groups and applications to partial differential equations is found in [7, 10, 17]. In section 3, we derive an expansion for the local error of the exponential operator splitting method (1.2) that remains valid within the analytic framework of section 2; for that purpose, extending Jahnke and Lubich [11, Proof of Thm. 2.1], we associate the order conditions with quadrature order conditions for multiple integrals. In section 4, we then prove a global error bound under reasonable regularity requirements on the initial value. Essential tools for our convergence analysis are estimates for repeated commutators; for instance, such bounds are valid for time-dependent linear Schrödinger equations subject to periodic boundary conditions, provided that the potential is sufficiently regular. The error estimate is finally illustrated by a numerical example.

A variety of works is concerned with exponential splitting methods for differential equations; as a small excerpt from the literature we mention [1, 2, 12, 18, 19] and the recent contributions [4, 5, 6, 15]; see also the references therein. In particular, we refer the reader to Lubich [14], where a rigorous convergence analysis of the Strang splitting for the cubic Schrödinger equation is given; Magnus integrators for linear Schrödinger equations involving a time-dependent potential are analyzed in Hochbruck and Lubich [9]. Basic information on splitting methods is also found in the survey article of McLachlan and Quispel [16] and the monograph of Hairer, Lubich, and Wanner [8].

2. Splitting methods for evolutionary Schrödinger equations. In this section, we give the basic hypotheses on the abstract differential equation (1.1). As illustration, we consider time-dependent linear Schrödinger equations and formulate them as evolution equations; for notational simplicity, we restrict ourselves to one space dimension. Moreover, we restate the exponential operator splitting method (1.2) and introduce several auxiliary abbreviations used throughout.

2.1. Evolution equations. We employ the following assumptions on the abstract initial value problem (1.1); see also Engel and Nagel [7, Thm. II.3.8]. To simplify matters, following Jahnke and Lubich [11], we require e^{tA} to be bounded by one.

HYPOTHESIS 1. *Let $(X, \|\cdot\|_X)$ denote the underlying Banach space and $\|\cdot\|_{X\leftarrow X}$ the induced operator norm. We assume that the densely defined and closed linear operator $A : D \subset X \to X$ is the infinitesimal generator of a $\mathscr{C}_0$-semigroup $\left(\mathrm{e}^{tA}\right)_{t\geq 0}$ on X satisfying the bound*

$$\left\|\mathrm{e}^{tA}\right\|_{X\leftarrow X} \leq 1, \qquad 0 \leq t \leq T. \tag{2.1}$$

Further, we suppose $B : X \to X$ to be a bounded linear operator.

The above assumptions ensure that for $u(0) \in D$ the uniquely determined (classical) solution $u \in \mathscr{C}^1([0,\infty), X)$ of the initial value problem (1.1) is given by

$$u(t) = \mathrm{e}^{t(A+B)}\, u(0), \qquad t \geq 0. \tag{2.2a}$$

Moreover, the following representation by the variation-of-constants formula

$$u(t) = \mathrm{e}^{tA}\, u(0) + \int_0^t \mathrm{e}^{(t-\tau)A}\, B\, u(\tau)\, \mathrm{d}\tau, \qquad t \geq 0, \tag{2.2b}$$

is valid; see [7, Chap. II.6/III.1].

For some $\omega \in \mathbb{R}$ and $\vartheta > 0$ the fractional powers A_ω^ϑ of the shifted operator $A_\omega = A - \omega I$ are well defined on a subspace of X (see [7, Chap. II.5c]); we denote by $\big(X_\vartheta, \|\cdot\|_{X_\vartheta}\big)$ the domain of A_ω^ϑ, endowed with the graph norm

$$\|v\|_{X_\vartheta} = \|v\|_X + \big\|A_\omega^\vartheta\, v\big\|_X, \qquad v \in X_\vartheta;$$

in particular, we set $A^0 = I$ and $X_0 = X$.

Henceforth, we assume that the linear operator B remains bounded on X_ϑ for some $\vartheta \geq 0$, that is, there exists a constant $C > 0$ such that

$$\big\|B\big\|_{X_\vartheta \leftarrow X_\vartheta} \leq C; \tag{2.3a}$$

consequently, the bound

$$\big\|\mathrm{e}^{tB}\big\|_{X_\vartheta \leftarrow X_\vartheta} \leq C, \qquad 0 \leq t \leq T, \tag{2.3b}$$

holds with some constant $C > 0$. Hypothesis 1 implies that $\mathrm{e}^{tA} : X \to X$ is uniformly bounded on finite time intervals; furthermore, for any $\vartheta \geq 0$ it follows that

$$\big\|\mathrm{e}^{tA}\big\|_{X_\vartheta \leftarrow X_\vartheta} \leq 1, \quad \big\|\mathrm{e}^{t(A+B)}\big\|_{X_\vartheta \leftarrow X_\vartheta} \leq C, \qquad 0 \leq t \leq T, \tag{2.3c}$$

with some constant $C > 0$.

For integers $j \geq 0$, the linear operators $\varphi_j(tB)$ are defined through

$$\varphi_0(z) = \mathrm{e}^z, \qquad \varphi_j(z) = \frac{1}{(j-1)!} \int_0^1 \tau^{j-1}\, \mathrm{e}^{(1-\tau)z}\, \mathrm{d}\tau, \quad j \geq 1, \qquad z \in \mathbb{C}. \tag{2.4a}$$

We conclude from the above relation (2.3b) that the estimate

$$\big\|\varphi_j(tB)\big\|_{X_\vartheta \leftarrow X_\vartheta} \leq \frac{1}{j!}\, C, \qquad 0 \leq t \leq T, \tag{2.4b}$$

is valid for $j \geq 0$.

In the situation of Hypothesis 1, the exponential operator splitting method (1.2) is well defined provided that $a_j \geq 0$ for $1 \leq j \leq s$. For schemes with negative coefficients, we employ a stronger framework; in particular, relation (2.3c) then remains valid for $0 \leq |t| \leq T$.

Hypothesis 2. *The densely defined closed linear operator $A : D \subset X \to X$ generates a $\mathscr{C}_0$-group $\big(\mathrm{e}^{tA}\big)_{t\in\mathbb{R}}$ on X such that the bound*

$$\big\|\mathrm{e}^{tA}\big\|_{X \leftarrow X} \leq 1, \qquad 0 \leq |t| \leq T, \tag{2.5}$$

is satisfied.

In order to obtain high-order error bounds for (1.2), we require the solution of the initial value problem (1.1) to fulfill certain commutator bounds on fractional power spaces of A. We henceforth employ the standard notation

$$\mathrm{ad}_A^0(B) = B, \qquad \mathrm{ad}_A^j(B) = \big[A, \mathrm{ad}_A^{j-1}(B)\big], \quad j \geq 1, \tag{2.6}$$

where $[A, L] = AL - LA$ for some linear operator L; e.g., see [8, Chap. III.4.1].

HYPOTHESIS 3. *We suppose that for certain $\vartheta, \widetilde{\vartheta} \geq 0$ the estimate*

$$\big\|\mathrm{ad}_A^j(B)\big\|_{X_\vartheta \leftarrow X_{\tilde{\vartheta}}} \leq C$$

is valid with constant $C > 0$.

As indicated in section 2.2, the above hypothesis is reasonable in connection with evolutionary linear Schrödinger equations that are subject to periodic boundary conditions and involve a smooth potential.

2.2. Time-dependent Schrödinger equations. Let $V : \Omega = [-\pi, \pi] \to \mathbb{R}$ be a periodic map that fulfills suitable regularity requirements. We consider the time-dependent linear Schrödinger equation

$$\mathrm{i}\,\partial_t U(x,t) = -\,\Delta\, U(x,t) + V(x)\, U(x,t), \qquad x \in \Omega, \quad t \geq 0, \tag{2.7}$$

subject to periodic boundary conditions on Ω and an initial condition $U(x,0) = U_0(x)$, $x \in \Omega$.

The above initial-boundary value problem is interpreted as an initial value problem of the form (1.1) by setting $u(t) = U(\cdot, t)$ and

$$\big(A\, v\big)(x) = \mathrm{i}\,\Delta\, v(x), \qquad \big(B\, v\big)(x) = -\,\mathrm{i}\, V(x)\, v(x)$$

for $v : \Omega \to \mathbb{C}$ a sufficiently regular function. Similarly as in Pazy [17, Chap. 7.5], it is shown that the linear operators $A : D \to X$ and $B : X \to X$ satisfy the assumptions of Hypothesis 1 with $X = L^2(\Omega, \mathbb{C})$ and $D = \big\{v \in H^2(\Omega, \mathbb{C}) : v \text{ periodic on } \Omega\big\}$. Furthermore, it holds that $X_{k/2} = \big\{v \in H^k(\Omega, \mathbb{C}) : v \text{ periodic on } \Omega\big\}$ for integers $k \geq 1$.

We note that Hypothesis 3 is fulfilled with $\vartheta = k/2$ and $\widetilde{\vartheta} = \vartheta + j/2$ for integers $j, k \geq 0$, provided that V is sufficiently often differentiable. In fact, the commutator of the one-dimensional Laplace operator $\mathscr{A} = \partial_x^2$ and a jth-order differential operator

$$\mathscr{B} = \sum_{\ell=0}^{j} \beta_\ell(x)\, \partial_x^\ell$$

with smooth space-dependent coefficients β_ℓ, $0 \leq \ell \leq j$, yields a differential operator of order $j + 1$. More precisely, it follows that

$$\big[\mathscr{A}, \mathscr{B}\big] = \mathscr{A}\mathscr{B} - \mathscr{B}\mathscr{A} = \sum_{\ell=0}^{j} \big(2\, \beta_\ell'\, \partial_x^{\ell+1} + \beta_\ell''\, \partial_x^\ell\big).$$

Proceeding by induction, we conclude that the iterated commutator $\mathrm{ad}_{\mathscr{A}}^j(V)$ is a differential operator of order j and therefore obtain

$$\big\|\mathrm{ad}_{\mathscr{A}}^j(V)\, v\big\|_{H^k(\Omega)} \leq C \|v\|_{H^{k+j}(\Omega)}, \qquad v \in H^{k+j}(\Omega),$$

for $k, j \geq 0$.

It is straightforward to generalize the above considerations to higher space dimensions.

2.3. Exponential operator splitting methods. Throughout, the gridpoints associated with a constant stepsize $h > 0$ are denoted by $t_n = nh$ for $n \geq 0$. For an exponential operator splitting method with real coefficients $a_j, b_j \in \mathbb{R}$, $1 \leq j \leq s$, numerical approximation values $u_n \approx u(t_n)$ to the true solution of (1.1) are given by the recurrence formula

$$u_n = \mathrm{e}^{b_s hB}\, \mathrm{e}^{a_s hA} \cdots \mathrm{e}^{b_2 hB}\, \mathrm{e}^{a_2 hA}\, \mathrm{e}^{b_1 hB}\, \mathrm{e}^{a_1 hA}\, u_{n-1}, \quad n \geq 1, \qquad u_0 \text{ given;} \tag{2.8a}$$

see Hypotheses 1 and 2. We note that the linear operators arising in the above product do not commute in general.

In order to write (2.8a) in short notation, it is useful to employ the following abbreviations. For linear operators $L_j : X \to X$, $\ell \leq j \leq k$, we define

$$\prod_{j=\ell}^{k} L_j = L_k \cdots L_{\ell+1}\, L_\ell, \quad k \geq \ell, \qquad \prod_{j=\ell}^{k} L_j = I, \quad k < \ell.$$

Here, $I : X \to X$ denotes the identity operator on X. Furthermore, for $k \leq s$ we set

$$P_k = \prod_{j=1}^{k} \mathrm{e}^{b_j hB}\, \mathrm{e}^{a_j hA}, \tag{2.8b}$$

and thus the exponential operator splitting method (2.8a) takes the compact form

$$u_n = P_s\, u_{n-1}, \quad n \geq 1, \qquad u_0 \text{ given;} \tag{2.8c}$$

see (1.2).

Inserting the exact solution values into the numerical scheme (2.8c) defines the defect at time t_n

$$u(t_n) = P_s\, u(t_{n-1}) + d_n, \qquad n \geq 1. \tag{2.9}$$

As a consequence, a recurrence relation for the error of the splitting method

$$e_n = u_n - u(t_n) = P_s\, e_{n-1} - d_n, \qquad n \geq 1,$$

follows. Resolving this recursion, we finally obtain

$$e_n = P_s^n\, e_0 - \sum_{j=0}^{n-1} P_s^{n-j-1}\, d_{j+1}, \qquad n \geq 0. \tag{2.10}$$

Our main result in section 4 is a convergence estimate for (2.8). Its proof is based on the above representation; that is, suitable estimations of the splitting operator and the defects are required; in this context, several auxiliary results provided in section 3.1 are of use.

3. Local error. In the present section, our concern is to deduce a suitable expansion of the local error

$$d_n = u(t_n) - P_s\, u(t_{n-1}), \qquad n \geq 1, \tag{3.1}$$

that remains well defined within the analytical framework of section 2; see (2.9). Furthermore, from this representation, the order conditions for the exponential operator splitting method (2.8) follow.

We start with expanding the exact solution and the splitting operator.

3.1. Auxiliary expansions. (i) *Exact solution.* An expansion of the exact solution value $u(t_n)$ that is specific to evolutionary Schrödinger equations is obtained by means of a reapplication of the variation-of-constants formula (2.2b). Namely, replacing in

$$u(t_n) = u(t_{n-1}+h) = \mathrm{e}^{hA}\, u(t_{n-1}) + \int_0^h \mathrm{e}^{(h-\tau_1)A}\, B\, u(t_{n-1}+\tau_1)\,\mathrm{d}\tau_1$$

the solution value $u(t_{n-1}+\tau_1)$ by

$$u(t_{n-1}+\tau_j) = \mathrm{e}^{\tau_j A}\, u(t_{n-1}) + \int_0^{\tau_j} \mathrm{e}^{(\tau_j-\tau_{j+1})A}\, B\, u(t_{n-1}+\tau_{j+1})\,\mathrm{d}\tau_{j+1}, \tag{3.2}$$

where $j=1$, yields the following relation:

$$\begin{aligned} u(t_n) &= \mathrm{e}^{hA}\, u(t_{n-1}) + \int_0^h \mathrm{e}^{(h-\tau_1)A}\, B\, \mathrm{e}^{\tau_1 A}\, u(t_{n-1})\,\mathrm{d}\tau_1 \\ &\quad + \int_0^h \int_0^{\tau_1} \mathrm{e}^{(h-\tau_1)A}\, B\, \mathrm{e}^{(\tau_1-\tau_2)A}\, B\, u(t_{n-1}+\tau_2)\,\mathrm{d}\tau_2\,\mathrm{d}\tau_1. \end{aligned} \tag{3.3}$$

Henceforth, we employ the compact vector notation $\tau = (\tau_1, \tau_2, \ldots, \tau_k) \in \mathbb{R}^k$ and set

$$f_k(\tau) = \prod_{\ell=1}^{k} \Big(\mathrm{e}^{(\tau_{k-\ell} - \tau_{k-\ell+1})A}\, B \Big), \qquad g_k(\tau) = f_k(\tau)\, \mathrm{e}^{\tau_k A}, \tag{3.4a}$$

with $\tau_0 = h$; as before, the product is defined downwards. Applying repeatedly the substitution (3.2) to (3.3), we therefore obtain

$$\begin{aligned} u(t_n) &= \mathrm{e}^{hA}\, u(t_{n-1}) + \sum_{k=1}^{p} I_k\, u(t_{n-1}) + R_{p+1}^{(1)}, \\ I_k &= \int_{\Delta_k} g_k(\tau)\,\mathrm{d}\tau, \qquad R_{p+1}^{(1)} = \int_{\Delta_{p+1}} f_{p+1}(\tau)\, u(t_{n-1}+\tau_{p+1})\,\mathrm{d}\tau, \end{aligned} \tag{3.4b}$$

where $\Delta_k = \big\{ \tau \in \mathbb{R}^k : 0 \le \tau_\ell \le \tau_{\ell-1}, 1 \le \ell \le k \big\}$. We note that the above expansion remains well defined in the situation of Hypothesis 1; see also section 4.

(ii) *Splitting operator.* We next specify a representation for the splitting operator P_s that allows, in an easy manner, a comparison with (3.4b). Our main tool for deducing such an expansion is a recurrence formula for the φ-functions

$$\varphi_j(z) = \frac{1}{j!} + z\, \varphi_{j+1}(z), \qquad j \ge 0, \quad z \in \mathbb{C}, \tag{3.5a}$$

obtained from (2.4a) by a partial integration. Moreover, we make use of the identity

$$\prod_{j=1}^{k} \big(K_j + L_j\big) = \prod_{\ell=1}^{k} K_\ell + \sum_{j=1}^{k} \prod_{\ell=j+1}^{k} K_\ell\; L_j \prod_{\ell=1}^{j-1} \big(K_\ell + L_\ell\big) \tag{3.5b}$$

that is valid for (noncommuting) linear operators K_j, L_j, $1 \le j \le k$. By applying the relations (3.5) to P_s, we obtain

$$P_s = \prod_{j=1}^{s} \mathrm{e}^{\,b_j h B}\, \mathrm{e}^{a_j h A} = \prod_{j=1}^{s} \Big(\mathrm{e}^{a_j h A} + h\, b_j\, B\, \varphi_1\big(b_j h B\big)\, \mathrm{e}^{a_j h A} \Big), \tag{3.6}$$

and further the representation

$$P_s = \mathrm{e}^{c_s hA} + h \sum_{j=1}^{s} b_j \, \mathrm{e}^{(c_s - c_j)\, hA} \, B \, \varphi_1(b_j hB) \, \mathrm{e}^{a_j hA} \, P_{j-1}. \tag{3.7}$$

Here, we employ the abbreviation

$$c_k = \sum_{j=1}^{k} a_j, \qquad 1 \le k \le s. \tag{3.8a}$$

Regarding the h-expansions (3.4b) and (3.7), it is seen that in (3.1) the $\mathscr{O}(1)$ term vanishes, provided that

$$c_s = 1; \tag{3.8b}$$

thus, we henceforth suppose the above order condition to be fulfilled. For the following considerations, it is useful to introduce the abbreviations

$$\begin{gathered} \Phi_j(\lambda) = F(\lambda) \, \varphi_j(b_{\lambda_k} hB) \, \mathrm{e}^{a_{\lambda_k} hA} \, P_{\lambda_k - 1}, \\ F(\lambda) = \prod_{\ell=1}^{k} b_{\lambda_\ell} \, f_k(c_\lambda h), \qquad G(\lambda) = F(\lambda) \, \mathrm{e}^{c_{\lambda_k} hA} = \prod_{\ell=1}^{k} b_{\lambda_\ell} \, g_k(c_\lambda h), \end{gathered} \tag{3.9}$$

where we set $c_\lambda = (c_{\lambda_1}, c_{\lambda_2}, \dots, c_{\lambda_k})$ for multi-indices $\lambda = (\lambda_1, \lambda_2, \dots, \lambda_k) \in \mathbb{N}^k$ and $\lambda_0 = s$; see (3.4a). With the help of this notation, relation (3.7) can be written as

$$P_s = \mathrm{e}^{hA} + h \sum_{\lambda_1 = 1}^{s} \Phi_1(\lambda_1). \tag{3.10}$$

In order to expand the splitting operator P_s further, we proceed by induction; that is, we repeatedly apply a recurrence relation for Φ_j to (3.10) which we derive next. Formula (3.5a) implies

$$\Phi_j(\lambda) = \frac{1}{j!} \, F(\lambda) \, \mathrm{e}^{a_{\lambda_k} hA} \, P_{\lambda_k - 1} + h \, \Phi_{j+1}(\lambda, \lambda_k);$$

moreover, from the analogue of (3.7),

$$\begin{aligned} \mathrm{e}^{a_{\lambda_k} hA} \, P_{\lambda_k - 1} = \mathrm{e}^{c_{\lambda_k} hA} \\ + h \sum_{\lambda_{k+1} = 1}^{\lambda_k - 1} b_{\lambda_{k+1}} \, \mathrm{e}^{(c_{\lambda_k} - c_{\lambda_{k+1}})\, hA} \, B \, \varphi_1(b_{\lambda_{k+1}} hB) \, \mathrm{e}^{a_{\lambda_{k+1}} hA} \, P_{\lambda_{k+1} - 1}, \end{aligned}$$

we obtain the recurrence formula

$$\Phi_j(\lambda) = \frac{1}{j!} \, G(\lambda) + h \left(\Phi_{j+1}(\lambda, \lambda_k) + \sum_{\lambda_{k+1} = 1}^{\lambda_k - 1} \frac{1}{j!} \, \Phi_1(\lambda, \lambda_{k+1}) \right). \tag{3.11}$$

As an illustration, we apply the above relation twice to (3.10). In a first step, we get

$$\begin{gathered} P_s = \mathrm{e}^{hA} + Q_1 + r_2^{(2)}, \\ Q_1 = h \sum_{\lambda_1 = 1}^{s} G(\lambda_1), \qquad r_2^{(2)} = h^2 \Bigg(\sum_{\lambda_1 = 1}^{s} \Phi_2(\lambda_1, \lambda_1) + \sum_{\lambda_1 = 1}^{s} \sum_{\lambda_2 = 1}^{\lambda_1 - 1} \Phi_1(\lambda_1, \lambda_2) \Bigg). \end{gathered}$$

A further expansion of $r_2^{(2)}$ by means of (3.11) yields

$$P_s = \mathrm{e}^{hA} + \sum_{k=1}^{2} Q_k + r_3^{(2)},$$

$$Q_1 = h \sum_{\lambda_1=1}^{s} G(\lambda_1), \qquad Q_2 = h^2 \Bigg(\sum_{\lambda_1=1}^{s} \tfrac{1}{2}\, G(\lambda_1,\lambda_1) + \sum_{\lambda_1=1}^{s} \sum_{\lambda_2=1}^{\lambda_1-1} G(\lambda_1,\lambda_2) \Bigg), \tag{3.12a}$$

with remainder $r_3^{(2)}$ given by

$$\begin{aligned} r_3^{(2)} = h^3 \Bigg(& \sum_{\lambda_1=1}^{s} \Phi_3(\lambda_1,\lambda_1,\lambda_1) + \sum_{\lambda_1=1}^{s} \sum_{\lambda_2=1}^{\lambda_1-1} \tfrac{1}{2}\, \Phi_1(\lambda_1,\lambda_1,\lambda_2) \\ & + \sum_{\lambda_1=1}^{s} \sum_{\lambda_2=1}^{\lambda_1-1} \Phi_2(\lambda_1,\lambda_2,\lambda_2) + \sum_{\lambda_1=1}^{s} \sum_{\lambda_2=1}^{\lambda_1-1} \sum_{\lambda_3=1}^{\lambda_2-1} \Phi_1(\lambda_1,\lambda_2,\lambda_3) \Bigg). \end{aligned} \tag{3.12b}$$

By reason of brevity, we write Q_2 as

$$Q_2 = h^2 \sum_{\lambda_1=1}^{s} \sum_{\lambda_2=1}^{\lambda_1} \alpha_{\lambda_1\lambda_2}\, G(\lambda_1,\lambda_2), \qquad \alpha_{\lambda_1\lambda_2} = \begin{cases} \frac{1}{2} & \text{if } \lambda_2 = \lambda_1, \\ 1 & \text{if } \lambda_2 \le \lambda_1 - 1. \end{cases} \tag{3.12c}$$

Applying (3.11) several times to (3.12), we finally end up with the expansion

$$P_s = \mathrm{e}^{hA} + \sum_{k=1}^{p} Q_k + r_{p+1}^{(2)}. \tag{3.13a}$$

Here, the multiple sum Q_k comprises terms of the form $h^k\, \alpha_\lambda\, G(\lambda)$ for certain $\alpha_\lambda \in \mathbb{R}$; more precisely, we have

$$Q_k = h^k \sum_{\lambda \in \Lambda_k} \alpha_\lambda\, G(\lambda) = h^k \sum_{\lambda_1=1}^{s} \sum_{\lambda_2=1}^{\lambda_1} \cdots \sum_{\lambda_k=1}^{\lambda_{k-1}} \alpha_{\lambda_1\lambda_2\cdots\lambda_k}\, G(\lambda_1,\lambda_2,\ldots,\lambda_k), \tag{3.13b}$$

where we denote $\Lambda_k = \big\{\lambda \in \mathbb{N}^k : 1 \le \lambda_\ell \le \lambda_{\ell-1}, 1 \le \ell \le k\big\}$ with $\lambda_0 = s$, according to (3.4). The coefficients α_λ are displayed in Table A.1 for $1 \le k \le 4$. The remainder $r_{p+1}^{(2)}$ is given by

$$r_{p+1}^{(2)} = h^{p+1} \sum_{j=1}^{p+1} \sum_{\lambda \in \Lambda_{p+1}} \widetilde{\alpha}_{j\lambda}\, \Phi_j(\lambda) \tag{3.13c}$$

for certain $\widetilde{\alpha}_{j\lambda} \in \mathbb{R}$; we omit further details and refer the reader to the MATLAB code `recurrence`, available at http://techmath.uibk.ac.at/mecht/research/research.html, that determines the coefficients α_λ and $\widetilde{\alpha}_{j\lambda}$.

(iii) *Quadrature formulas.* From the above considerations, we obtain the following expansion of the defect:

$$d_n = \sum_{k=1}^{p} \big(I_k - Q_k\big)\, u(t_{n-1}) + R_{p+1}^{(1)} - R_{p+1}^{(2)}, \qquad R_{p+1}^{(2)} = r_{p+1}^{(2)}\, u(t_{n-1}); \tag{3.14}$$

see also (3.1), (3.4), (3.8), and (3.13). We next relate the difference $I_k - Q_k$ to the error of quadrature formulas for multiple integrals. Regarding (3.9) and (3.13b), we write Q_k as

$$Q_k = h^k \sum_{\lambda \in \Lambda_k} \beta_\lambda \, g_k(c_\lambda h), \qquad \beta_\lambda = \beta_{\lambda_1 \lambda_2 \cdots \lambda_k} = \alpha_{\lambda_1 \lambda_2 \cdots \lambda_k} \prod_{\ell=1}^{k} b_{\lambda_\ell}; \tag{3.15}$$

see (3.4a) for the definition of g_k. As usual, in order to determine the defects of the quadrature formula (3.15), we employ a Taylor series expansion of order $M = p - k$:

$$\begin{aligned} g_k(\tau) &= \sum_{m=0}^{p-k} \frac{1}{m!} \, g_k^{(m)}(0) \, \tau^m + \varrho_{k,p-k+1}(\tau), \\ \varrho_{k,M+1}(\tau) &= \frac{1}{M!} \int_0^1 (1-z)^M \, g_k^{(M+1)}(z\,\tau) \, \tau^{M+1} \, \mathrm{d}z. \end{aligned} \tag{3.16a}$$

In view of section 4, we next specify the mth derivative of g_k. It holds that

$$g_k^{(m)}(\sigma) \, \tau^m = \sum_{\substack{\mu \in \mathbb{N}^k \\ |\mu| = m}} \frac{m!}{\mu!} \, \partial_\tau^\mu g_k(\sigma) \, \tau^\mu; \tag{3.16b}$$

here, for $\mu = (\mu_1, \mu_2, \ldots, \mu_k) \in \mathbb{N}^k$ and $\tau = (\tau_1, \tau_2, \ldots, \tau_k) \in \mathbb{R}^k$, we let

$$|\mu| = \sum_{\ell=1}^{k} \mu_\ell, \qquad \mu! = \prod_{\ell=1}^{k} \mu_\ell!, \qquad \tau^\mu = \prod_{\ell=1}^{k} \tau_\ell^{\mu_\ell}, \qquad \partial_\tau^\mu = \prod_{\ell=1}^{k} \partial_{\tau_\ell}^{\mu_\ell}. \tag{3.16c}$$

Proceeding by induction, it follows that

$$\partial_\tau^\mu \, g_k(\tau) = (-1)^{|\mu|} \prod_{\ell=1}^{k} \bigl(\mathrm{e}^{(\tau_{k-\ell} - \tau_{k-\ell+1})A} \, \mathrm{ad}_A^{\mu_{k-\ell+1}}(B) \bigr) \, \mathrm{e}^{\tau_k A}; \tag{3.16d}$$

we recall definition (2.6) of the repeated commutators. In particular, we have

$$\partial_\tau^\mu \, g_k(0) = (-1)^{|\mu|} \, \mathrm{e}^{hA} \prod_{\ell=1}^{k} \mathrm{ad}_A^{\mu_{k-\ell+1}}(B). \tag{3.16e}$$

Inserting the Taylor series expansion (3.16) into

$$I_k - Q_k = \int_{\Delta_k} g_k(\tau) \, \mathrm{d}\tau - h^k \sum_{\lambda \in \Lambda_k} \beta_\lambda \, g_k(c_\lambda h)$$

(see (3.4b) and (3.15)) and making use of the fact that

$$\int_{\Delta_k} \tau^\mu \, \mathrm{d}\tau = h^{k+|\mu|} \prod_{\ell=1}^{k} \frac{1}{\mu_\ell + \cdots + \mu_k + k - \ell + 1}, \tag{3.17}$$

we obtain the representation

$$\begin{aligned} I_k - Q_k = \sum_{m=0}^{p-k} h^{k+m} \sum_{\substack{\mu \in \mathbb{N}^k \\ |\mu| = m}} \frac{1}{\mu!} \biggl(\prod_{\ell=1}^{k} \frac{1}{\mu_\ell + \cdots + \mu_k + k - \ell + 1} - \sum_{\lambda \in \Lambda_k} \beta_\lambda \, c_\lambda^\mu \biggr) \partial_\tau^\mu g_k(0) \\ + \int_{\Delta_k} \varrho_{k,p-k+1}(\tau) \, \mathrm{d}\tau - h^k \sum_{\lambda \in \Lambda_k} \beta_\lambda \, \varrho_{k,p-k+1}(c_\lambda h). \end{aligned}$$

Altogether, inserting the above relation into (3.14) implies

$$d_n = \sum_{k=1}^{p} \sum_{m=0}^{p-k} h^{k+m} \sum_{\substack{\mu\in\mathbb{N}^k \\ |\mu|=m}} \frac{1}{\mu!} \Big(\prod_{\ell=1}^{k} \frac{1}{\mu_\ell + \cdots + \mu_k + k - \ell + 1} - \sum_{\lambda\in\Lambda_k} \beta_\lambda \, c_\lambda^\mu \Big)$$
$$\times \, \partial_\tau^\mu g_k(0)\, u(t_{n-1}) + R_{p+1}^{(1)} - R_{p+1}^{(2)} + R_{p+1}^{(3)},$$
$$R_{p+1}^{(3)} = \sum_{k=1}^{p} \Big(\int_{\Delta_k} \varrho_{k,p-k+1}(\tau)\, \mathrm{d}\tau - h^k \sum_{\lambda\in\Lambda_k} \beta_\lambda \, \varrho_{k,p-k+1}(c_\lambda h) \Big)\, u(t_{n-1});$$

see also (3.4b) and (3.13c) for the definition of $R_{p+1}^{(1)}$ and $R_{p+1}^{(2)}$.

3.2. Local error expansion and order conditions. In Lemma 1, we restate the expansion for the defect (3.1) deduced in section 3.1. In view of section 4, we resume the employed abbreviations; see also (2.2a), (2.4a), (2.6), and (3.16c).

LEMMA 1. *We set* $\Lambda_k = \big\{\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_k) \in \mathbb{N}^k : 1 \le \lambda_\ell \le \lambda_{\ell-1}, 1 \le \ell \le k\big\}$ *and* $\Delta_k = \big\{\tau = (\tau_1, \tau_2, \ldots, \tau_k) \in \mathbb{R}^k : 0 \le \tau_\ell \le \tau_{\ell-1}, 1 \le \ell \le k\big\}$, *where* $\lambda_0 = s$ *and* $\tau_0 = h$. *Further, we denote* $c_k = a_1 + a_2 + \cdots + a_k$ *for* $1 \le k \le s$. *Provided that the condition* $c_s = 1$ *is fulfilled, the defect of the exponential operator splitting method* (2.8) *equals*

(3.18)
$$d_n = \sum_{k=1}^{p} \sum_{\substack{\mu\in\mathbb{N}^k \\ |\mu|\le p-k}} \frac{(-1)^{|\mu|}}{\mu!}\, h^{k+|\mu|} \Big(\prod_{\ell=1}^{k} \frac{1}{\mu_\ell + \cdots + \mu_k + k - \ell + 1} - \sum_{\lambda\in\Lambda_k} \alpha_\lambda \prod_{\ell=1}^{k} b_{\lambda_\ell} c_{\lambda_\ell}^{\mu_\ell} \Big)$$
$$\times \, \mathrm{e}^{hA} \prod_{\ell=1}^{k} \mathrm{ad}_A^{\mu_{k-\ell+1}}(B)\, \mathrm{e}^{t_{n-1}(A+B)}\, u(0) + R_{p+1}.$$

The remainder $R_{p+1} = R_{p+1}^{(1)} - R_{p+1}^{(2)} + R_{p+1}^{(3)}$ *is given by*

$$R_{p+1}^{(1)} = \int_{\Delta_{p+1}} \prod_{j=1}^{p+1} \Big(\mathrm{e}^{(\tau_{p+1-j} - \tau_{p+1-j+1})A}\, B \Big)\, \mathrm{e}^{(t_{n-1}+\tau_{p+1})(A+B)}\, u(0)\, \mathrm{d}\tau,$$
$$R_{p+1}^{(2)} = h^{p+1} \sum_{j=1}^{p+1} \sum_{\lambda\in\Lambda_{p+1}} \widetilde{\alpha}_{j\lambda} \prod_{\ell=1}^{p+1} \Big(b_{\lambda_\ell}\, \mathrm{e}^{(c_{\lambda_{p-\ell+1}} - c_{\lambda_{p-\ell+2}})\, hA}\, B \Big)$$
$$\times \, \varphi_j(b_{\lambda_{p+1}} hB)\, \mathrm{e}^{a_{\lambda_{p+1}} hA}\, P_{\lambda_{p+1}-1}\, \mathrm{e}^{t_{n-1}(A+B)}\, u(0),$$

$$\varrho_{k,p-k+1}(\tau) = (p-k+1) \int_0^1 (1-z)^{p-k} \sum_{\substack{\mu\in\mathbb{N}^k \\ |\mu|=p-k+1}} \frac{(-1)^{|\mu|}}{\mu!}$$
$$\times \prod_{\ell=1}^{k} \Big(\mathrm{e}^{(\tau_{k-\ell} - \tau_{k-\ell+1})\, zA}\, \mathrm{ad}_A^{\mu_{k-\ell+1}}(B) \Big)\, \mathrm{e}^{\tau_k z A}\, \tau^\mu\, \mathrm{d}z,$$

$$R_{p+1}^{(3)} = \sum_{k=1}^{p} \Big(\int_{\Delta_k} \varrho_{k,p-k+1}(\tau)\, \mathrm{d}\tau - h^k \sum_{\lambda\in\Lambda_k} \alpha_\lambda \prod_{\ell=1}^{k} b_{\lambda_\ell}\, \varrho_{k,p-k+1}(c_\lambda h) \Big)\, \mathrm{e}^{t_{n-1}(A+B)}\, u(0).$$

The coefficients α_λ and $\widetilde{\alpha}_{m\lambda}$ are obtained by recurrence.[1]

Obviously, the first term in the local error expansion (3.18) vanishes whenever (combinations of) the coefficients

$$C_\mu = \prod_{\ell=1}^{k} \frac{1}{\mu_\ell + \cdots + \mu_k + k - \ell + 1} - \sum_{\lambda \in \Lambda_k} \alpha_\lambda \prod_{\ell=1}^{k} b_{\lambda_\ell} c_{\lambda_\ell}^{\mu_\ell} \tag{3.19}$$

vanish for all $1 \le k \le p$ and $|\mu| \le p - k$. In order to show that the classical order conditions for a general exponential operator splitting method (2.8) coincide with the stiff order conditions, we meanwhile assume (1.1) to be a *nonstiff* problem; that is, we require the linear operator A to be bounded on X. In this situation, the expansion

$$\mathrm{e}^{hA} = \sum_{j=0}^{\infty} \frac{1}{j!}\, h^j A^j$$

is well defined on X and further yields

$$d_\mu = \sum_{\substack{k \ge 1,\, j \ge 0,\, \mu \in \mathbb{N}^k \\ k + j + |\mu| \le p}} \frac{(-1)^{|\mu|}}{j!\,\mu!}\, h^{k+j+|\mu|}\, C_\mu\, E_{j\mu}(A,B)\, \mathrm{e}^{t_{\mu-1}(A+B)}\, \mu(0) + \mathscr{O}(\mu^{p+1}),$$

$$E_{j\mu}(A,B) = A^j \prod_{\ell=1}^{k} \mathrm{ad}_A^{\mu_{k-\ell+1}}(B).$$

The splitting method is consistent of order p iff for any $1 \le q \le p$ the term involving h^q vanishes. Thus, taking into account all combinations of $k \ge 1$, $j \ge 0$, and $\mu \in \mathbb{N}^k$ such that $k + |j| + |\mu| = q$ for some $1 \le q \le p$ fixed, we conclude that

$$\sum_{\substack{k \ge 1,\, j \ge 0,\, \mu \in \mathbb{N}^k \\ k + j + |\mu| = q}} \frac{(-1)^{|\mu|}}{j!\,\mu!}\, C_\mu\, E_{j\mu}(A,B) = 0.$$

For $j > 0$, due to lower order conditions, the corresponding terms vanish; if $j = 0$, the same conditions as in the stiff case arise.

Alternatively, in order to show that for evolutionary Schrödinger equations the stiff and nonstiff order conditions coincide, one could also derive the classical order conditions by means of the Campbell–Baker–Hausdorff formula (see [8]); however, as further tedious calculations are involved, we did not follow this approach here.

The proof of our convergence result for general exponential operator splitting methods (2.8) relies on a suitable estimation of the local error expansion given in Lemma 1. Provided that the conditions

$$\sum_{\ell=1}^{s} a_\ell = 1,$$

$$\sum_{\lambda \in \Lambda_k} \alpha_\lambda \prod_{\ell=1}^{k} b_{\lambda_\ell} c_{\lambda_\ell}^{\mu_\ell} = \prod_{\ell=1}^{k} \frac{1}{\mu_\ell + \cdots + \mu_k + k - \ell + 1}, \qquad 1 \le k \le p, \quad |\mu| \le p - k, \tag{3.20}$$

[1] A MATLAB code `recurrence` is available at http://techmath.uibk.ac.at/mecht/research/research.html.

are fulfilled (see also (3.19)), the first term in the local error expansion vanishes and it remains to estimate R_{p+1}. Clearly, for *nonstiff* problems (1.1), the iterated commutators are bounded; however, whenever the differential equation involves an unbounded linear operator A, this is not true in general. For evolutionary Schrödinger equations, reasonable regularity assumptions on the initial value allow us to estimate the decisive term $R^{(3)}_{p+1}$ by means of Hypothesis 3.

3.3. Example methods. As an illustration, we next specify various exponential operator splitting methods and verify that the conditions (3.20) are satisfied; see also Appendix A. Our main result in section 4 ensures that the example methods retain their convergence orders for evolutionary Schrödinger equations.

Order 1. The conditions for $p = 1$ are given in (A.1a). In particular, for $s = 1$ it follows that $a_1 = 1 = b_1$, and we retain the *Lie–Trotter* splitting

$$u_n = \mathrm{e}^{hB}\,\mathrm{e}^{hA}\,u_{n-1}, \quad n \geq 1, \qquad u_0 \text{ given.}$$

Order 2. For $p = 2$ the additional conditions are given in (A.1b). Choosing $s = 2$ and simplifying (A.1a) and (A.1b) yields $a_1 + a_2 = 1$, $b_1 + b_2 = 1$, and $b_1 a_1 + b_2 = \frac{1}{2}$. For the choice $a_1 = 0$ we thus retain the *Strang* splitting (1.3); alternatively, by setting $a_1 = \frac{1}{2}$ it follows that

$$u_n = \mathrm{e}^{\frac{1}{2}hA}\,\mathrm{e}^{hB}\,\mathrm{e}^{\frac{1}{2}hA}\,u_{n-1}, \quad n \geq 1, \qquad u_0 \text{ given.}$$

Order 4/6. Symmetric symplectic methods that can be cast in the form (2.8) are proposed in Blanes and Moan [3, Tables 2–3]. The coefficients of a fourth- and sixth-order scheme are collected in Table 3.1; the conditions (3.20) are fulfilled (up to machine precision) for $p = 4$ and $p = 6$, respectively (see also Appendix A).

TABLE 3.1
Coefficients of a fourth-order method given in Blanes and Moan [3].

j	a_j
1,7	0.0792036964311957
2,6	0.3531729060497740
3,5	-0.0420650803577195
4	$1 - 2\,(a_1 + a_2 + a_3)$

j	b_j
1,6	0.209515106613362
2,5	-0.143851773179818
3,4	$1/2 - (b_1 + b_2)$
7	0

4. Global error. In this section, we derive a convergence result for the exponential operator splitting method (2.8) when applied to the abstract initial value problem (1.1) and further illustrate the error bound by a numerical example.

4.1. Global error bound in terms of the initial value. In the formulation of Theorem 1, we focus on evolutionary linear Schrödinger equations; that is, the unbounded linear operator A is related to the Laplacian. Provided that the method coefficients a_j are nonnegative for $1 \leq j \leq s$, it suffices to require that A generates a $\mathscr{C}_0$-semigroup; see Hypothesis 1.

THEOREM 1. *Assume that the coefficients of the exponential operator splitting method* (2.8) *fulfill the classical order conditions for some integer* $p \geq 1$. *Suppose further that the linear operators* A *and* B *satisfy the requirements of Hypotheses* 2 *and* 3 *with* $\vartheta = k/2$ *and* $\widetilde{\vartheta} = (j+k)/2$ *for integers* $j, k \geq 0$ *such that* $j + k \leq p$. *Then, provided that* $u(0) \in X_{p/2}$, *the error estimate*

$$\big\|u_n - u(t_n)\big\|_X \leq C\,\big\|u(0) - u_0\big\|_X + C\,h^p\,\big\|u(0)\big\|_{X_{p/2}}, \qquad 0 \leq nh \leq T,$$

is valid with some constant $C > 0$ depending in particular on T but not on n and h.

Proof. Our main tools for deriving the above convergence result are a stability bound for the splitting operator and an estimate for the defect of the exponential operator splitting method (2.8). For notational simplicity, we do not distinguish the arising constants.

We first verify the boundedness of P_s^n. To that purpose, we employ the relation

$$P_s = \prod_{j=1}^{s} \big(I + h\, b_j\, B\, \varphi_1(b_j h B)\big)\, \mathrm{e}^{a_j h A}$$

obtained by means of (3.5a); see also (3.6). Applying (2.3) and (2.4) and using that $1 + z \le \mathrm{e}^z$ for $z > 0$, it follows that

$$\begin{aligned} \big\|P_s\big\|_{X\leftarrow X} &\le \prod_{j=1}^{s} \big\|I + h\, b_j\, B\, \varphi_1\big(b_j h B\big)\big\|_{X\leftarrow X}\, \big\|\mathrm{e}^{a_j h A}\big\|_{X\leftarrow X} \\ &\le \prod_{j=1}^{s} \big(1 + C\, h\, |b_j|\big) \le \mathrm{e}^{Ch}. \end{aligned}$$

Furthermore, we conclude that for any $h > 0$ the estimate

$$\big\|P_s^n\big\|_{X\leftarrow X} \le C, \qquad 0 \le nh \le T, \tag{4.1}$$

holds true on finite time intervals with constant $C > 0$ depending on T but not on n and h.

We next deduce a suitable bound for the defects. Regarding the expansion given in Lemma 1 and using that the order conditions (3.20) are fulfilled, it remains to estimate the remainder. By means of (2.3), we obtain

$$\begin{aligned} \big\|R_{p+1}^{(1)}\big\|_X &\le \int_{\Delta_{p+1}} \prod_{j=1}^{p+1} \Big(\big\|\mathrm{e}^{(\tau_{k-j} - \tau_{k-j+1})A}\big\|_{X\leftarrow X}\, \|B\|_{X\leftarrow X}\Big) \\ &\quad \times \big\|\mathrm{e}^{(t_{n-1} + \tau_{p+1})(A+B)}\big\|_{X\leftarrow X}\, \big\|u(0)\big\|_X\, \mathrm{d}\tau \\ &\le C\, h^{p+1}\, \big\|u(0)\big\|_X; \end{aligned}$$

set $|\mu| = 0$ in (3.17). In a similar manner, it follows that

$$\big\|R_{p+1}^{(2)}\big\|_X \le C\, h^{p+1} \big\|u(0)\big\|_X.$$

Hypothesis 3, together with (2.3), implies

$$\big\|\varrho_{k,p-k+1}(\tau)\big\|_{X\leftarrow X_{\vartheta_k}} \le C \sum_{|\mu| = p-k+1} \frac{\tau^\mu}{\mu!}, \qquad \vartheta_k = \frac{p-k+1}{2}\,. \tag{4.2}$$

Thus, by means of (3.17), we obtain

$$\begin{aligned} \Big(\int_{\Delta_k} \big\|\varrho_{k,p-k+1}(\tau)\big\|_{X\leftarrow X_{\vartheta_k}}\, \mathrm{d}\tau + h^k \sum_{\lambda\in\Lambda_k} \alpha_\lambda \prod_{\ell=1}^{k} |b_{\lambda_\ell}|\, \big\|\varrho_{k,p-k+1}(c_\lambda h)\big\|_{X\leftarrow X_{\vartheta_k}}\Big) \\ \times \big\|\mathrm{e}^{t_{n-1}(A+B)}\big\|_{X_{\vartheta_k}\leftarrow X_{\vartheta_k}}\, \big\|u(0)\big\|_{X_{\vartheta_k}} \\ \le C\, h^{p+1}\, \big\|u(0)\big\|_{X_{\vartheta_k}}. \end{aligned}$$

Employing the reasonable assumption

$$\|v\|_{X_\vartheta} \le \|v\|_{X_{\widetilde{\vartheta}}}, \qquad \vartheta \le \widetilde{\vartheta}, \quad v \in X_{\widetilde{\vartheta}}, \tag{4.3}$$

this finally gives the bound

$$\big\|R_{p+1}^{(3)}\big\|_X \le C\, h^{p+1}\, \big\|u(0)\big\|_{X_{p/2}}.$$

From the above considerations, we obtain the following estimate for the defect:

$$\big\|d_n\big\|_X = \big\|R_{p+1}\big\|_X \le C\, h^{p+1} \big\|u(0)\big\|_{X_{p/2}};$$

see also Lemma 1, (3.20), and (4.3). With the help of (4.1), it is now straightforward to estimate the global error; see (2.10). Altogether, we obtain

$$\begin{aligned} \big\|e_n\big\|_X &\le \big\|P_s^n\big\|_{X\leftarrow X}\, \big\|e_0\big\|_X + \sum_{j=0}^{n-1} \big\|P_s^{n-j-1}\big\|_{X\leftarrow X}\, \big\|d_{j+1}\big\|_X \\ &\le C\, \big\|e_0\big\|_X + C\, h^p\, \big\|u(0)\big\|_{X_{p/2}}, \qquad 0 \le nh \le T, \end{aligned}$$

with constant $C > 0$ depending in particular on T. □

In the present situation, as seen in the proof of Theorem 1, general exponential splitting methods remain stable for any choice of the time stepsize $0 < h \le T$; however, for evolutionary Schrödinger equations (1.1) involving a time-dependent or solution-dependent operator B, in general, a stepsize restriction is expected; see [9, 14].

The above error bound implies that the example methods given in section 3.2 retain their convergence orders when applied to time-dependent Schrödinger equations subject to a periodic boundary condition, provided that the data are sufficiently differentiable. For less regular initial values, the following convergence result holds true; order reduction phenomena for the Strang splitting are also studied in [11, 13].

COROLLARY 1. *Under the assumptions of Theorem* 1*, whenever* $u(0) \in X_{k/2}$ *but* $u(0) \notin X_{(k+1)/2}$ *for some* $1 \le k \le p-1$*, the error bound*

$$\big\|u_n - u(t_n)\big\|_X \le C\, \big\|u(0) - u_0\big\|_X + C\, h^k\, \big\|u(0)\big\|_{X_{k/2}}, \qquad 0 \le nh \le T,$$

is valid with constant $C > 0$ *depending on* T *but not on* n *and* h*.*

We note that the above convergence analysis also applies to equations involving a differential operator of higher order; in this case, the relation between the quantities $\vartheta, \widetilde{\vartheta}$ arising in Hypothesis 3 has to be adapted.

4.2. Numerical example. We next illustrate the error bound of Theorem 1 by a numerical example for a time-dependent linear Schrödinger equation. In the present paper, we do not recapitulate how the obtained results for abstract differential equations are applicable to pseudospectral discretizations of time-dependent linear Schrödinger equations; in this respect and for a more detailed description of the realization of the numerical example, we refer the reader to Jahnke and Lubich [11, sect. 3].

As test problem, we consider (2.7) with $V(x) = 1 - \cos x$; in order to study numerically the influence of the initial value on the temporal convergence order of an exponential operator splitting method, we choose various initial data that correspond to $u(0) \in X_{k/2}$ for $2 \le k \le 6$. The initial-boundary value problem is discretized in space by the Fourier pseudospectral method with $M = 2^{13}$ gridpoints.

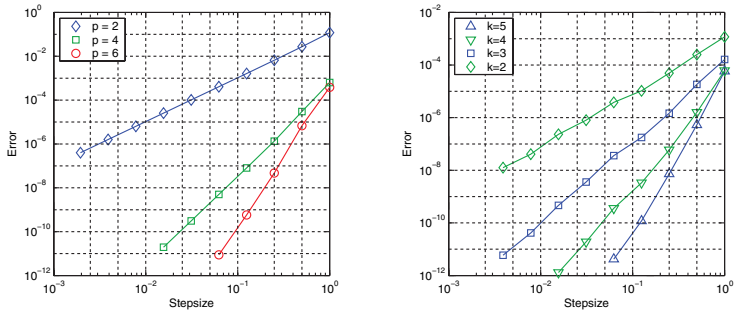


FIG. 4.1. *Numerically observed convergence orders of exponential operator splitting methods. Schemes of various orders (left picture). Initial values of various regularity (right picture).*

For a sufficiently regular initial value, i.e., $k = 6$, we apply the second-order Strang splitting and the fourth- and sixth-order schemes proposed by Blanes and Moan [3] for various time stepsizes (see section 3.3); a reference solution at final time $T = 1$ is computed by the sixth-order splitting method with stepsize $h = 2^{-11}$. The obtained temporal errors, displayed in Figure 4.1 (left picture), confirm the assertion of Theorem 1; the splitting methods retain their convergence orders for time-dependent Schrödinger equations.

For less regular initial values, the order reduction predicted by Corollary 1 is also observed numerically; for the sixth-order scheme the numerical convergence orders are displayed in Figure 4.1 (right picture).

Appendix A. In this appendix, we state the coefficients α_λ, $\lambda \in \Lambda_k$ (see Table A.1), and further specify the conditions (3.20) for $1 \le k \le p = 4$:

$$\sum_{\ell=1}^{s} a_\ell = 1, \qquad \sum_{\lambda_1=1}^{s} b_{\lambda_1} = 1, \tag{A.1a}$$

$$\sum_{\lambda_1=1}^{s} b_{\lambda_1} c_{\lambda_1} = \frac{1}{2}, \qquad \sum_{\lambda_1=1}^{s} \frac{1}{2}\, b_{\lambda_1}^2 + \sum_{\lambda_1=1}^{s} \sum_{\lambda_2=1}^{\lambda_1-1} b_{\lambda_1} b_{\lambda_2} = \frac{1}{2}, \tag{A.1b}$$

TABLE A.1
Coefficients $\alpha_{\lambda_\kappa} = \alpha_{\lambda_{\kappa_1}\lambda_{\kappa_2}\cdots\lambda_{\kappa_k}}$ *for* $\kappa = (\kappa_1, \kappa_2, \ldots, \kappa_k)$.

k	κ	α_{λ_κ}
1	1	1
2	11	1/2
	12	1
3	111	1/6
	112,122	1/2
	123	1
4	1111	1/24
	1112,1222	1/6
	1122	1/4
	1123,1223,1233	1/2
	1234	1

$$\sum_{\lambda_1=1}^{s} b_{\lambda_1} c_{\lambda_1}^2 = \frac{1}{3},$$

$$\sum_{\lambda_1=1}^{s} \frac{1}{2}\, b_{\lambda_1}^2 c_{\lambda_1} + \sum_{\lambda_1=1}^{s} \sum_{\lambda_2=1}^{\lambda_1-1} b_{\lambda_1} b_{\lambda_2} c_{\lambda_1} = \frac{1}{3},$$

$$\sum_{\lambda_1=1}^{s} \frac{1}{2}\, b_{\lambda_1}^2 c_{\lambda_1} + \sum_{\lambda_1=1}^{s} \sum_{\lambda_2=1}^{\lambda_1-1} b_{\lambda_1} b_{\lambda_2} c_{\lambda_2} = \frac{1}{6},$$

$$\sum_{\lambda_1=1}^{s} \frac{1}{6}\, b_{\lambda_1}^3 + \sum_{\lambda_1=1}^{s} \sum_{\lambda_2=1}^{\lambda_1-1} \frac{1}{2}\, b_{\lambda_1} b_{\lambda_2} \big(b_{\lambda_1} + b_{\lambda_2}\big) + \sum_{\lambda_1=1}^{s} \sum_{\lambda_2=1}^{\lambda_1-1} \sum_{\lambda_3=1}^{\lambda_2-1} b_{\lambda_1} b_{\lambda_2} b_{\lambda_3} = \frac{1}{6},$$

$$\sum_{\lambda_1=1}^{s} b_{\lambda_1} c_{\lambda_1}^3 = \frac{1}{4},$$

$$\sum_{\lambda_1=1}^{s} \frac{1}{2}\, b_{\lambda_1}^2 c_{\lambda_1}^2 + \sum_{\lambda_1=1}^{s} \sum_{\lambda_2=1}^{\lambda_1-1} b_{\lambda_1} b_{\lambda_2} c_{\lambda_1}^2 = \frac{1}{4},$$

$$\sum_{\lambda_1=1}^{s} \frac{1}{2}\, b_{\lambda_1}^2 c_{\lambda_1}^2 + \sum_{\lambda_1=1}^{s} \sum_{\lambda_2=1}^{\lambda_1-1} b_{\lambda_1} b_{\lambda_2} c_{\lambda_1} c_{\lambda_2} = \frac{1}{8},$$

$$\sum_{\lambda_1=1}^{s} \frac{1}{2}\, b_{\lambda_1}^2 c_{\lambda_1}^2 + \sum_{\lambda_1=1}^{s} \sum_{\lambda_2=1}^{\lambda_1-1} b_{\lambda_1} b_{\lambda_2} c_{\lambda_2}^2 = \frac{1}{12},$$

$$\sum_{\lambda_1=1}^{s} \frac{1}{6}\, b_{\lambda_1}^3 c_{\lambda_1} + \sum_{\lambda_1=1}^{s} \sum_{\lambda_2=1}^{\lambda_1-1} \frac{1}{2}\, b_{\lambda_1} b_{\lambda_2} c_{\lambda_1} \big(b_{\lambda_1} + b_{\lambda_2}\big) + \sum_{\lambda_1=1}^{s} \sum_{\lambda_2=1}^{\lambda_1-1} \sum_{\lambda_3=1}^{\lambda_2-1} b_{\lambda_1} b_{\lambda_2} b_{\lambda_3} c_{\lambda_1} = \frac{1}{8},$$

$$\sum_{\lambda_1=1}^{s} \frac{1}{6}\, b_{\lambda_1}^3 c_{\lambda_1} + \sum_{\lambda_1=1}^{s} \sum_{\lambda_2=1}^{\lambda_1-1} \frac{1}{2}\, b_{\lambda_1} b_{\lambda_2} \big(b_{\lambda_1} c_{\lambda_1} + b_{\lambda_2} c_{\lambda_2}\big) + \sum_{\lambda_1=1}^{s} \sum_{\lambda_2=1}^{\lambda_1-1} \sum_{\lambda_3=1}^{\lambda_2-1} b_{\lambda_1} b_{\lambda_2} b_{\lambda_3} c_{\lambda_2} = \frac{1}{12},$$

$$\sum_{\lambda_1=1}^{s} \frac{1}{6}\, b_{\lambda_1}^3 c_{\lambda_1} + \sum_{\lambda_1=1}^{s} \sum_{\lambda_2=1}^{\lambda_1-1} \frac{1}{2}\, b_{\lambda_1} b_{\lambda_2} c_{\lambda_2} \big(b_{\lambda_1} + b_{\lambda_2}\big) + \sum_{\lambda_1=1}^{s} \sum_{\lambda_2=1}^{\lambda_1-1} \sum_{\lambda_3=1}^{\lambda_2-1} b_{\lambda_1} b_{\lambda_2} b_{\lambda_3} c_{\lambda_3} = \frac{1}{24},$$

$$\sum_{\lambda_1=1}^{s} \frac{1}{24}\, b_{\lambda_1}^4 + \sum_{\lambda_1=1}^{s} \sum_{\lambda_2=1}^{\lambda_1-1} b_{\lambda_1} b_{\lambda_2} \left(\frac{1}{6}\, b_{\lambda_1}^2 + \frac{1}{4}\, b_{\lambda_1} b_{\lambda_2} + \frac{1}{6}\, b_{\lambda_2}^2 \right) + \sum_{\lambda_1=1}^{s} \sum_{\lambda_2=1}^{\lambda_1-1} \sum_{\lambda_3=1}^{\lambda_2-1} \frac{1}{2}\, b_{\lambda_1} b_{\lambda_2} b_{\lambda_3} \big(b_{\lambda_1} + b_{\lambda_2} + b_{\lambda_3}\big) + \sum_{\lambda_1=1}^{s} \sum_{\lambda_2=1}^{\lambda_1-1} \sum_{\lambda_3=1}^{\lambda_2-1} \sum_{\lambda_4=1}^{\lambda_3-1} b_{\lambda_1} b_{\lambda_2} b_{\lambda_3} b_{\lambda_4} = \frac{1}{24}.$$

REFERENCES

[1] W. Bao, D. Jaksch, and P. Markowich, *Numerical solution of the Gross–Pitaevskii equation for Bose–Einstein condensation*, J. Comput. Phys., 187 (2003), pp. 318–342.
[2] C. Besse, B. Bidégaray, and S. Descombes, *Order estimates in time of splitting methods for the nonlinear Schrödinger equation*, SIAM J. Numer. Anal., 40 (2002), pp. 26–40.
[3] S. Blanes and P.C. Moan, *Practical symplectic partitioned Runge–Kutta and Runge–Kutta–Nyström methods*, J. Comput. Appl. Math., 142 (2002), pp. 313–330.
[4] S. Blanes, F. Casas, and A. Murua, *On the Linear Stability of Splitting Methods*, Numerical Analysis Reports 8/2006, University of Cambridge, Cambridge, UK, 2006.
[5] E. Celledoni and N. Saefstroem, *A symmetric splitting method for rigid body dynamics*, Model. Identif. Control, 27 (2006), pp. 95–108.
[6] P. Chartier and A. Murua, *Preserving First Integrals and Volume Forms of Additively Split Systems*, INRIA Report RR-6016, Rennes, France, 2006.
[7] K.J. Engel and R. Nagel, *One-Parameter Semigroups for Linear Evolution Equations*, Springer, New York, 2000.
[8] E. Hairer, Ch. Lubich, and G. Wanner, *Geometric Numerical Integration. Structure-Preserving Algorithms for Ordinary Differential Equations*, Springer, Berlin, 2002.
[9] M. Hochbruck and Ch. Lubich, *On Magnus integrators for time-dependent Schrödinger equations*, SIAM J. Numer. Anal., 41 (2003), pp. 945–963.
[10] E. Hille and R.S. Phillips, *Functional Analysis and Semi-Groups*, AMS, Providence, RI, 1957.
[11] T. Jahnke and Ch. Lubich, *Error bounds for exponential operator splittings*, BIT, 40 (2000), pp. 735–744.
[12] R. Kozlov, A. Kværnø, and B. Owren, *The Behaviour of the Local Error in Splitting Methods Applied to Stiff Problems*, Preprint Numerics 1/2003, University of Trondheim, Trondheim, Norway, 2003; also available online from http://www.math.ntnu.no/~bryn/trliste.imf (6th report).
[13] R. Kozlov and B. Owren, *Order Reduction in Operator Splitting Methods*, Preprint Numerics 6/1999, University of Trondheim, Trondheim, Norway, 1999.
[14] Ch. Lubich, *On Splitting Methods for Schrödinger-Poisson and Cubic Nonlinear Schrödinger Equations*, preprint, Universität Tübingen, Tübingen, Germany, 2007.
[15] R. McLachlan and R. Quispel, *Geometric integration of ODEs*, J. Phys., 39 (2006), pp. 5251–5286.
[16] R. McLachlan and R. Quispel, *Splitting methods*, Acta Numer., 11 (2002), pp. 341–434.
[17] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer, New York, 1983.
[18] Q. Sheng, *Global error estimates for exponential splitting*, IMA J. Numer. Anal., 14 (1994), pp. 27–56.
[19] B. Sportisse and J.G. Verwer, *A Note on Operator Splitting in the Stiff Linear Case*, Technical report MAS-R9830, CWI Amsterdam, Amsterdam, The Netherlands, 1998.
[20] G. Strang, *On the construction and comparison of difference schemes*, SIAM J. Numer. Anal., 5 (1968), pp. 506–517.
[21] H.F. Trotter, *On the product of semi-groups of operators*, Proc. Amer. Math. Soc., 10 (1959), pp. 545–551.

SIAM J. NUMER. ANAL.
Vol. 46, No. 4, pp. 2039–2067
© 2008 Society for Industrial and Applied Mathematics

REDUCED BASIS ERROR BOUND COMPUTATION OF PARAMETER-DEPENDENT NAVIER–STOKES EQUATIONS BY THE NATURAL NORM APPROACH*

SIMONE DEPARIS†

Abstract. This work focuses on the a posteriori error estimation for the reduced basis method applied to partial differential equations with quadratic nonlinearity and affine parameter dependence. We rely on natural norms—*local* parameter-dependent norms—to provide a sharp and computable lower bound of the inf-sup constant. We prove a formulation of the Brezzi–Rappaz–Raviart existence and uniqueness theorem in the presence of two distinct norms. This allows us to relax the existence condition and to sharpen the field variable error bound. We also provide a robust algorithm to compute the Sobolev embedding constants involved in the error bound and in the inf-sup lower bound computation. We apply our method to a steady natural convection problem in a closed cavity, with a Grashof number varying from 10 to 10^7.

Key words. reduced basis methods, a posteriori error estimation, Brezzi–Rappaz–Raviart theory, steady incompressible Navier–Stokes equations, natural convection

AMS subject classifications. 35Q30, 46E35, 65N30, 76D03, 76D05

DOI. 10.1137/060674181

1. Introduction.

In this work we are interested in the numerical approximation of parameter- ($\boldsymbol{\mu}$-) dependent quadratic nonlinear partial differential equations—in particular the natural convection problem for Navier–Stokes equations—and the prediction of an "output of interest" which is a functional of the field variable $\boldsymbol{u}^{\mathrm{e}}(\boldsymbol{\mu})$,

$$s^{\mathrm{e}}(\boldsymbol{\mu}) = \ell(\boldsymbol{u}^{\mathrm{e}}(\boldsymbol{\mu})) \in \mathbb{R}, \tag{1.1}$$

where ℓ is a continuous linear form. The solution $\boldsymbol{u}^{\mathrm{e}}(\boldsymbol{\mu})$ may not be unique; nevertheless, we consider only one solution branch; hence we presume local uniqueness.

The outputs are related to energies or forces, stresses or strains, flow rates or pressure drops, temperatures or fluxes and are functions of an "input" parameter P-vector $\boldsymbol{\mu} \in \mathcal{D} \subset \mathbb{R}^P$, which is related to geometry, physical properties, boundary conditions, or loads.

We directly adopt the description and notation of Sen et al. [32], where the natural norms are introduced and applied to the error bound computation of coercive and noncoercive linear elliptic partial differential equations.

The field variable $\boldsymbol{u}^{\mathrm{e}}(\boldsymbol{\mu}) \in X^{\mathrm{e}}$—say, velocity, pressure, or temperature—satisfies in weak form the $\boldsymbol{\mu}$-parametrized nonlinear partial differential equation

$$a(\boldsymbol{u}^{\mathrm{e}}(\boldsymbol{\mu}), \boldsymbol{v}; \boldsymbol{\mu}) = f(\boldsymbol{v}) \quad \forall\, \boldsymbol{v} \in X^{\mathrm{e}}. \tag{1.2}$$

Here X^{e} is the appropriate function space with norm $\|\cdot\|_X$, $a(\cdot, \cdot; \boldsymbol{\mu})$ is continuous, quadratic in the first variable, and linear in the second one, and f is a continuous linear form.

*Received by the editors November 7, 2006; accepted for publication (in revised form) December 12, 2007; published electronically April 23, 2008. This work was supported by DARPA and AFOSR under grant FA9550-05-1-0114, by the Singapore-MIT Alliance, and by PSC grant ASC040010P.
http://www.siam.org/journals/sinum/46-4/67418.html
†Mechanical Engineering Department, Massachusetts Institute of Technology, Cambridge, MA 02139 (simone.deparis@epfl.ch).

Unfortunately, to achieve the desired accuracy, the evaluation $\boldsymbol{\mu} \to s^{\mathrm{e}}(\boldsymbol{\mu})$ by discrete projection methods such as finite element or spectral methods is simply too costly in the many-query and real-time contexts often of interest in engineering. Low-order models—we consider here reduced basis approximations—are increasingly popular in the engineering analysis, parameter estimation, design optimization, and control contexts.

In the reduced basis approach [1, 18, 6, 22, 20, 13, 23, 11, 12, 25], we approximate $s^{\mathrm{e}}(\boldsymbol{\mu})$, $\boldsymbol{u}^{\mathrm{e}}(\boldsymbol{\mu})$ with $s_N(\boldsymbol{\mu}), \boldsymbol{u}_N(\boldsymbol{\mu})$: Given $\boldsymbol{\mu} \in \mathcal{D}$,

$$s_N(\boldsymbol{\mu}) = \ell(\boldsymbol{u}_N(\boldsymbol{\mu})),$$

where $\boldsymbol{u}_N(\boldsymbol{\mu}) \in W_N$ satisfies[1]

$$a(\boldsymbol{u}_N(\boldsymbol{\mu}), \boldsymbol{v}; \boldsymbol{\mu}) = f(\boldsymbol{v}) \quad \forall\, \boldsymbol{v} \in W_N. \tag{1.3}$$

Here W_N is a problem-specific space of dimension $N \le N_{\max}$ that focuses on the (typically very smooth) parametric manifold of interest—$\{\boldsymbol{u}^{\mathrm{e}}(\boldsymbol{\mu}) \,|\, \boldsymbol{\mu} \in \mathcal{D}\}$—and thus the reduced basis method enjoys very rapid convergence as N increases [6, 15]; $N_{\max}$ is the maximal size of our reduced basis.

Our own effort is dedicated to the development of a posteriori error estimators for reduced basis approximations in the nonlinear case [13, 23, 35, 17, 11, 12, 24]. Since, in general, we cannot find the exact (our superscript "e" above) solution, we replace $s^{\mathrm{e}}(\boldsymbol{\mu}), \boldsymbol{u}^{\mathrm{e}}(\boldsymbol{\mu})$ with a Galerkin finite element approximation[2] $s^{\mathcal{N}}(\boldsymbol{\mu})$, $\boldsymbol{u}^{\mathcal{N}}(\boldsymbol{\mu})$. $\mathcal{N}$ represents the dimension of our finite element space.

For some fixed sufficiently large $\mathcal{N} = \mathcal{N}_{\mathrm{t}}$, we presume that the finite element and the exact solution are very close. We shall call $\boldsymbol{u}^{\mathcal{N}_t}(\boldsymbol{\mu})$ a "truth" solution, upon which we provide inexpensive—complexity *independent* of $\mathcal{N}_{\mathrm{t}}$—and sharp error bounds $\Delta_N(\boldsymbol{\mu})$ and $\Delta_N^s(\boldsymbol{\mu})$ such that

$$\|\boldsymbol{u}^{\mathcal{N}_{\mathrm{t}}}(\boldsymbol{\mu}) - \boldsymbol{u}_N(\boldsymbol{\mu})\|_X \le \Delta_N(\boldsymbol{\mu}) \quad \text{and}$$
$$|s^{\mathcal{N}_{\mathrm{t}}}(\boldsymbol{\mu}) - s_N(\boldsymbol{\mu})| \le \Delta_N^s(\boldsymbol{\mu}) \quad \forall\, \boldsymbol{\mu} \in \mathcal{D}.$$

In the absence of such rigorous error bounds we cannot efficiently determine if N is too small—and our reduced basis approximation unacceptably inaccurate—or if N is too large—and our reduced basis approximation unnecessarily expensive—or we cannot establish the very *existence* of a "truth" solution $\boldsymbol{u}^{\mathcal{N}_{\mathrm{t}}}(\boldsymbol{\mu})$ [8, 34, 33]. We cannot determine in "real time" if critical design conditions and constraints are satisfied. In fact, we cannot even construct an efficient and well-conditioned reduced basis approximation space W_N [17, 16].

The dramatic *dimension reduction*—from $\mathcal{N}_t$ to N—in conjunction with *offline/online computational procedures* [13, 23, 2, 9], yields very large savings in the many-query and real-time contexts: The online complexity depends only on the size of the reduced basis space N, which is typically orders of magnitude smaller than the dimension of the finite element space $\mathcal{N}_{\mathrm{t}}$. Hence we may consider a highly accurate "truth" approximation, at the detriment of the offline complexity only.

Relative to earlier work [35, 17, 34, 33], we adapt the *natural norms* technique proposed in [32] to the quadratic case in order to improve the construction and evaluation of a lower bound of the inf-sup constant (which we shall subsequently call the

[1] For simplicity, we consider a purely primal approach; we shall subsequently describe also a primal-dual formulation.

[2] Other projection methods, such as spectral or spectral elements methods, may be used as well; cf. [11, 12].

"inf-sup lower bound"). We also extend the Brezzi–Rappaz–Raviart theory for the case of mixed norms (in our case, X and natural norms); this enables one to better control the effectivity of our error bound and to significantly reduce the constraint on the dual norm of the residual for the existence and uniqueness of a (finite element) solution in the neighborhood of our reduced basis approximation.

The effectivity relative to the natural norm of the field variable error is proven to be very accurate: locally of order one when the dual norm of the residual is small. However, when considering only a primal formulation—as in the example that we provide—we have no control of the effectivity on the output. To partially address this issue, we also propose a primal-dual formulation, whose apparent advantages would require further investigation, in particular in conjunction with deflation.

Our improvements enable a feasible and fast inf-sup lower bound construction based on a relatively small number of local norms. The inf-sup lower bound is of order one in the *local* natural norm and hence is intrinsically much simpler to compute than in existing works [34, 17, 33]. The lower bound components are also simplified: They rely on the computation of Sobolev embedding constants and eigenvalues with respect to the natural norms. We develop robust tools to compute them based on multiple initial guesses modes and, regarding the Sobolev embedding constants, the restriction to a small discrete subspace to provide the initial guess and on a fixed point algorithm.

As a result, we are now able to cover much larger parameter sets than in earlier works. For example, in [17, 33] the authors considered natural convection with a Grashof number (ratio buoyancy over viscosity) from 10 to 10^4. Instead, we have applied the reduced basis method for a Grashof number up to 10^7 without increasing—actually by simplifying—the complexity of the error bound computation. To enable the finite element resolution of our problem at these values of the Grashof number and to speed up the offline work, we implemented a parallel code and a domain decomposition preconditioner based on a sequential one proposed in [5].

We first introduce the basic assumptions that we are going to use throughout the paper and define our natural norms for nonlinear partial differential equations (section 2). We then prove an existence and unicity result based on the Brezzi–Rappaz–Raviart theory and provide an output error bound based on a pure primal approach (section 3). In section 4 we develop our inf-sup lower bound construction that enables a feasible offline/online computational strategy (section 5). We then apply our approach to a steady natural convection problem (section 6), provide the numerical results (section 7), and present our conclusions (section 8). The appendix is devoted to the output error bound by a primal-dual formulation.

2. Problem statement.

2.1. Reduced basis formulation. For any $\boldsymbol{\mu} \equiv (\mu_1, \ldots, \mu_P)$ in the closed input domain $\mathcal{D} \subset \mathbb{R}^P$, $a(\cdot, \cdot; \boldsymbol{\mu})\colon X^{\mathrm{e}} \times X^{\mathrm{e}} \to \mathbb{R}$ is a parameter-dependent form, which is quadratic with respect to its first argument and linear with respect to the second one. The forms ℓ and $f\colon X^{\mathrm{e}} \to \mathbb{R}$ are parameter-independent and linear (extension to parameter-dependent linear forms is straightforward; we provide also an extension to quadratic outputs).

Our "truth" or "reference" finite element approximation to the exact output and field variable $s(\boldsymbol{\mu}) \equiv s^{\mathcal{N}_{\mathrm{t}}}(\boldsymbol{\mu})$ and $\boldsymbol{u}(\boldsymbol{\mu}) \equiv \boldsymbol{u}^{\mathcal{N}_{\mathrm{t}}}(\boldsymbol{\mu}) \in X^{\mathcal{N}_{\mathrm{t}}} \equiv X$, for a given $\boldsymbol{\mu} \in \mathcal{D}$,

$$\begin{aligned} s(\boldsymbol{\mu}) &= \ell(\boldsymbol{u}(\boldsymbol{\mu})) \quad \text{and} \\ a(\boldsymbol{u}(\boldsymbol{\mu}), \boldsymbol{v}; \boldsymbol{\mu}) &= f(\boldsymbol{v}) \quad \forall\, \boldsymbol{v} \in X. \end{aligned} \tag{2.1}$$

We assume that $\mathcal{N}_\mathrm{t}$ is chosen sufficiently large so that $s(\boldsymbol{\mu})$ and $\boldsymbol{u}(\boldsymbol{\mu})$ are essentially indistinguishable from $s^\mathrm{e}(\boldsymbol{\mu})$ and $\boldsymbol{u}^\mathrm{e}(\boldsymbol{\mu})$, respectively. We shall build our reduced basis approximation upon this "truth" approximation, and we shall evaluate the error in our reduced basis approximation with respect to this "truth" approximation. The online complexity (and stability) of our reduced basis approach is *independent* of $\mathcal{N}_\mathrm{t}$ (cf. [17, 33, 32]); hence, we may choose $\mathcal{N}_\mathrm{t}$ to be "arbitrarily" large at *no detriment to (online) performance*.

We shall suppose that our form a is "affine" in the parameter: For some fixed integers Q_0 and Q_1—typically larger than P, sometimes by a considerable factor—we require that

$$a(\boldsymbol{u},\boldsymbol{v};\boldsymbol{\mu}) = \sum_{q=1}^{Q_0} \Theta_0^q(\boldsymbol{\mu}) a_0^q(\boldsymbol{u},\boldsymbol{v}) + \frac{1}{2}\sum_{q=1}^{Q_1} \Theta_1^q(\boldsymbol{\mu}) a_1^q(\boldsymbol{u},\boldsymbol{u},\boldsymbol{v}) \quad \forall\, \boldsymbol{u},\boldsymbol{v}\in X,\ \forall\, \boldsymbol{\mu}\in\mathcal{D}, \tag{2.2}$$

where $\Theta_0^q, \Theta_1^q\colon\ \mathcal{D}\to\mathbb{R}$, $a_0^q\colon X\times X\to\mathbb{R}$, $1\le q\le Q_0$, and $a_1^q\colon X\times X\times X\to\mathbb{R}$, $1\le q\le Q_1$, are parameter-dependent functions and parameter-independent continuous bilinear and trilinear forms, respectively. (In particular, this assumption presumes a quadratic nonlinearity in our partial differential equation.) Without loss of generality, we assume that, for $1\le q\le Q_1$ and $\forall \boldsymbol{u},\boldsymbol{v},\boldsymbol{w}\in X$, $a_1^q(\boldsymbol{u},\boldsymbol{w},\boldsymbol{v}) = a_1^q(\boldsymbol{w},\boldsymbol{u},\boldsymbol{v})$. We shall further assume that $\Theta_i^q\in\mathcal{C}^1(\mathcal{D})$, $1\le q\le Q_i$, $i=0,1$.

We denote the inner product and norm associated with our Hilbert space $X(\equiv X^{\mathcal{N}_\mathrm{t}}$, finite-dimensional) as $(\boldsymbol{w},\boldsymbol{v})_X$ and $\|\boldsymbol{v}\|_X = \sqrt{(\boldsymbol{v},\boldsymbol{v})_X}$, respectively. We further define the dual norm for any bounded linear functional f as

$$\|f\|_{X'} \equiv \sup_{\boldsymbol{v}\in X} \frac{f(\boldsymbol{v})}{\|\boldsymbol{v}\|_X}.$$

For example, when considering two-dimensional Navier–Stokes equations, the exact space X^e shall satisfy $H_0^1(\Omega)\times H_0^1(\Omega)\times L_0^2(\Omega)\subset X^\mathrm{e}\subset H^1(\Omega)\times H^1(\Omega)\times L^2(\Omega)$: Here $\Omega\subset\mathbb{R}^2$ is a spatial domain with suitably regular boundary $\partial\Omega$; $L^2(\Omega)$ is the Hilbert space of integrable functions; $H^1(\Omega)$ is the usual Hilbert space of derivative-square-integrable functions; $H_0^1(\Omega)$ indicates the functions with null trace on $\partial\Omega$ and $L_0^2(\Omega)$ the zero mean ones. The typical choice for our inner product $(\cdot,\cdot)_X$ is the standard $H^1(\Omega)\times H^1(\Omega)\times L^2(\Omega)$ inner product.

If in this example the Reynolds number is the parameter, then the affine decomposition (2.2) becomes, for $\boldsymbol{\mu}=(\mathrm{Re})$, $\boldsymbol{u}=(u_1,u_2,p)$, $\boldsymbol{v}=(v_1,v_2,q)$, $Q_0=2$, and $Q_1=1$,

$$\begin{aligned}
a_0^1(\boldsymbol{u},\boldsymbol{v}) &= -\int_\Omega p\partial_i v_i - \int_\Omega q\partial_i u_i + \lambda\int_\Omega p + \gamma\int_\Omega q, && \Theta_0^1(\boldsymbol{\mu}) = 1,\\
a_0^2(\boldsymbol{u},\boldsymbol{v}) &= \int_\Omega \partial_i u_j \partial_i v_j, && \Theta_0^2(\boldsymbol{\mu}) = \mathrm{Re}^{-1},\\
a_1^1(\boldsymbol{u},\boldsymbol{z},\boldsymbol{v}) &= \int_\Omega u_j\partial_i z_i v_j + \int_\Omega z_j\partial_i u_i v_j, && \Theta_1^1(\boldsymbol{\mu}) = 1
\end{aligned}$$

(with summation over repeated indices and partial derivatives with respect to x_1, x_2).

By recalling (2.2) and the symmetry of the a_1^q's in the first two variables, the Fréchet derivative of $a(\cdot,\cdot;\boldsymbol{\mu})$ with respect to the first variable at a point $\boldsymbol{w}\in X$ can be expressed as

$$da(\boldsymbol{w};\boldsymbol{\mu})(\boldsymbol{u},\boldsymbol{v}) = \sum_{q=1}^{Q_0} \Theta_0^q(\boldsymbol{\mu}) a_0^q(\boldsymbol{u},\boldsymbol{v}) + \sum_{q=1}^{Q_1} \Theta_1^q(\boldsymbol{\mu}) a_1^q(\boldsymbol{u},\boldsymbol{w},\boldsymbol{v}) \quad \forall\, \boldsymbol{u},\boldsymbol{v}\in X,\ \forall\, \boldsymbol{\mu}\in\mathcal{D}.$$

For any $\boldsymbol{\mu}$ in $\mathcal{D}$ and any solution $\boldsymbol{u}(\boldsymbol{\mu})$ of (2.1) in the region of our interest, we assume that $da(\boldsymbol{u}(\boldsymbol{\mu});\boldsymbol{\mu})$ is "stable" and continuous in the sense that there exist $\beta_0 > 0$ and $\gamma_0 \in \mathbb{R}$ such that, $\forall \boldsymbol{\mu} \in \mathcal{D}$,

$$0 < \beta_0 < \beta(\boldsymbol{\mu}) \equiv \inf_{\boldsymbol{w}\in X} \sup_{\boldsymbol{v}\in X} \frac{da(\boldsymbol{u}(\boldsymbol{\mu});\boldsymbol{\mu})(\boldsymbol{w},\boldsymbol{v})}{\|\boldsymbol{w}\|_X \|\boldsymbol{v}\|_X}, \tag{2.3}$$

$$\infty > \gamma_0 > \gamma(\boldsymbol{\mu}) \equiv \sup_{\boldsymbol{w}\in X} \sup_{\boldsymbol{v}\in X} \frac{da(\boldsymbol{u}(\boldsymbol{\mu});\boldsymbol{\mu})(\boldsymbol{w},\boldsymbol{v})}{\|\boldsymbol{w}\|_X \|\boldsymbol{v}\|_X}.$$

It then follows that in the neighborhood of $\boldsymbol{u}(\boldsymbol{\mu})$ the solution to (2.1) is unique. We further assume that ℓ and f are in X'—bounded linear functionals.

2.2. Natural norms. Let $\boldsymbol{u}_N(\boldsymbol{\mu})$ be the reduced basis approximation to $\boldsymbol{u}(\boldsymbol{\mu})$ given by (1.3). We next introduce [35, 17, 7, 14, 32] the parametrized linear operator $T_N^{\boldsymbol{\mu}}\colon X \to X$ such that

$$(T_N^{\boldsymbol{\mu}}\boldsymbol{w},\boldsymbol{v})_X = da(\boldsymbol{u}_N(\boldsymbol{\mu});\boldsymbol{\mu})(\boldsymbol{w},\boldsymbol{v}) \quad \forall\, \boldsymbol{v},\boldsymbol{w} \in X. \tag{2.4}$$

Our method—in particular our inf-sup lower bound construction—requires a discrete set of K parameter values $\mathcal{V}^K \equiv \{\overline{\boldsymbol{\mu}}_1,\dots,\overline{\boldsymbol{\mu}}_K\} \subset \mathcal{D}$—upon which to construct local norms—a fixed integer $\overline{N} \le N_{\max}$,[3] and an indicator function $\mathcal{I}^K\colon\ \mathcal{D} \to \mathcal{V}^K$ which associates to any $\boldsymbol{\mu}$ in $\mathcal{D}$ a member of $\mathcal{V}^K$.[4] Typically, $\mathcal{I}^K$ defines the "closest"—in a sense that we have to define—element of $\mathcal{D}$.

We assume that, $\forall \overline{\boldsymbol{\mu}} \in \mathcal{V}^K$,

$$0 < \beta_{\overline{N}}(\overline{\boldsymbol{\mu}}) \equiv \inf_{\boldsymbol{w}\in X} \sup_{\boldsymbol{v}\in X} \frac{da(\boldsymbol{u}_{\overline{N}}(\overline{\boldsymbol{\mu}});\overline{\boldsymbol{\mu}})(\boldsymbol{w},\boldsymbol{v})}{\|\boldsymbol{w}\|_X \|\boldsymbol{v}\|_X} \equiv \inf_{\boldsymbol{w}\in X} \frac{\|T_{\overline{N}}^{\overline{\boldsymbol{\mu}}}\boldsymbol{w}\|_X}{\|\boldsymbol{w}\|_X}, \tag{2.5}$$

$$\infty > \gamma_{\overline{N}}(\overline{\boldsymbol{\mu}}) \equiv \sup_{\boldsymbol{w}\in X} \sup_{\boldsymbol{v}\in X} \frac{da(\boldsymbol{u}_{\overline{N}}(\overline{\boldsymbol{\mu}});\overline{\boldsymbol{\mu}})(\boldsymbol{w},\boldsymbol{v})}{\|\boldsymbol{w}\|_X \|\boldsymbol{v}\|_X} \equiv \sup_{\boldsymbol{w}\in X} \frac{\|T_{\overline{N}}^{\overline{\boldsymbol{\mu}}}\boldsymbol{w}\|_X}{\|\boldsymbol{w}\|_X}. \tag{2.6}$$

Note that this is true if $\boldsymbol{u}_{\overline{N}}(\overline{\boldsymbol{\mu}})$ is close enough to $\boldsymbol{u}(\overline{\boldsymbol{\mu}})$ (even though this is not required here); from the Cauchy–Schwarz inequality, $\boldsymbol{v} = T_{\overline{N}}^{\overline{\boldsymbol{\mu}}}\boldsymbol{w}$ is the inner supremizer in (2.5) and (2.6).

We then introduce, for any given $\overline{\boldsymbol{\mu}} \in \mathcal{V}^K$ and $\overline{N}$, our "natural norm"

$$|||\boldsymbol{v}|||_{\overline{\boldsymbol{\mu}}} \equiv \|T_{\overline{N}}^{\overline{\boldsymbol{\mu}}}\boldsymbol{v}\|_X \quad \forall\, \boldsymbol{v} \in X. \tag{2.7}$$

(To simplify the notation we have dropped the index $\overline{N}$ from the norm symbol.) This norm is the extension of the natural norm introduced in [32] for the linear case. Note that, thanks to our assumptions on $\beta_{\overline{N}}(\overline{\boldsymbol{\mu}})$ and $\gamma_{\overline{N}}(\overline{\boldsymbol{\mu}})$, (2.7) does indeed define a norm, which is equivalent to $\|\cdot\|_X$.

2.3. Trilinear forms continuity constants. In the development of our error bound, we assume that there exists a positive $\rho_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})$ such that

$$\left|da(\boldsymbol{z}^2;\boldsymbol{\mu})(\boldsymbol{v},\boldsymbol{w}) - da(\boldsymbol{z}^1;\boldsymbol{\mu})(\boldsymbol{v},\boldsymbol{w})\right| \le \rho_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})\, |||\boldsymbol{z}^2 - \boldsymbol{z}^1|||_{\overline{\boldsymbol{\mu}}}\, |||\boldsymbol{v}|||_{\overline{\boldsymbol{\mu}}}\, \|\boldsymbol{w}\|_X \tag{2.8}$$

for all $\boldsymbol{z}^1$, $\boldsymbol{z}^2$, $\boldsymbol{v}$, and $\boldsymbol{w}$ in X.

[3] In practice, $\overline{N}$ may depend on $\overline{\boldsymbol{\mu}}$, and the results of this work still hold.

[4] The process by which we select "good" $\mathcal{V}^K$ and $\mathcal{I}^K$ is described in section 5.

Similarly, in the construction of an inf-sup lower bound, we assume that there exists a positive $\rho_{X,\overline{\boldsymbol{\mu}}}$ such that

$$\left|\sum_{q=1}^{Q_1}\Theta_1^q(\overline{\boldsymbol{\mu}})a_1^q(\boldsymbol{v},\boldsymbol{z},\boldsymbol{w})\right| \le \rho_{X,\overline{\boldsymbol{\mu}}}\,\|\boldsymbol{z}\|_X\,|||\boldsymbol{v}|||_{\overline{\boldsymbol{\mu}}}\,\|\boldsymbol{w}\|_X, \tag{2.9}$$

$$\left|\sum_{q=1}^{Q_1}\left(\Theta_1^q(\boldsymbol{\mu})-\Theta_1^q(\overline{\boldsymbol{\mu}})\right)a_1^q(\boldsymbol{v},\boldsymbol{z},\boldsymbol{w})\right| \le \max_{q=1,\dots,Q_1}\frac{|\Theta_1^q(\boldsymbol{\mu})-\Theta_1^q(\overline{\boldsymbol{\mu}})|}{|\Theta_1^q(\overline{\boldsymbol{\mu}})|}\rho_{X,\overline{\boldsymbol{\mu}}}\,\|\boldsymbol{z}\|_X\,|||\boldsymbol{v}|||_{\overline{\boldsymbol{\mu}}}\,\|\boldsymbol{w}\|_X \tag{2.10}$$

for all $\boldsymbol{z}$, $\boldsymbol{v}$, and $\boldsymbol{w}$ in X.

To a given problem, we have to provide these constants; in sections 6.3 and 6.5 we show how to compute these constants, which are related to the Sobolev embedding theorem, for the model problem at hand. The algorithm proposed in section 6.5 can be easily extended to other examples.

3. Existence and uniqueness. Let $\boldsymbol{\mu}\in\mathcal{D}$, $\overline{\boldsymbol{\mu}}\in\mathcal{V}^K$, $N\le N_{\max}$, and $\overline{N}\le N_{\max}$; we define (implicitly dropping the subscripts $_{\overline{N}}$ and $_N$)

$$\beta_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu}) \equiv \inf_{\boldsymbol{w}\in X}\sup_{\boldsymbol{v}\in X}\frac{(T_N^{\boldsymbol{\mu}}\boldsymbol{w},\boldsymbol{v})_X}{|||\boldsymbol{w}|||_{\overline{\boldsymbol{\mu}}}||\boldsymbol{v}||_X} = \inf_{\boldsymbol{w}\in X}\frac{\|T_N^{\boldsymbol{\mu}}\boldsymbol{w}\|_X}{|||\boldsymbol{w}|||_{\overline{\boldsymbol{\mu}}}} = \inf_{\boldsymbol{w}\in X}\frac{\|T_N^{\boldsymbol{\mu}}\boldsymbol{w}\|_X}{\|T_{\overline{N}}^{\overline{\boldsymbol{\mu}}}\boldsymbol{w}\|_X}, \tag{3.1}$$

$$\gamma_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu}) \equiv \sup_{\boldsymbol{w}\in X}\sup_{\boldsymbol{v}\in X}\frac{(T_N^{\boldsymbol{\mu}}\boldsymbol{w},\boldsymbol{v})_X}{|||\boldsymbol{w}|||_{\overline{\boldsymbol{\mu}}}||\boldsymbol{v}||_X} = \sup_{\boldsymbol{w}\in X}\frac{\|T_N^{\boldsymbol{\mu}}\boldsymbol{w}\|_X}{|||\boldsymbol{w}|||_{\overline{\boldsymbol{\mu}}}} = \sup_{\boldsymbol{w}\in X}\frac{\|T_N^{\boldsymbol{\mu}}\boldsymbol{w}\|_X}{\|T_{\overline{N}}^{\overline{\boldsymbol{\mu}}}\boldsymbol{w}\|_X}. \tag{3.2}$$

We assume that $\beta_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})>0$ (this is the case when $\beta(\boldsymbol{\mu})>0$ (cf. (2.3)) and $\boldsymbol{u}_N(\boldsymbol{\mu})$ is "nearby" to $\boldsymbol{u}(\boldsymbol{\mu})$) and also that we explicitly know either $\beta_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})$ or a rigorous positive lower bound to it. The results in this section extend straightforwardly in the latter case by replacing $\beta_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})$ by its lower bound.

3.1. Uniqueness. We are now ready to prove a uniqueness result in the neighborhood—with respect to the natural norms—of a given reduced basis approximation. Let $B_{\overline{\boldsymbol{\mu}}}(\boldsymbol{u},\epsilon)$ be an open ball around $\boldsymbol{u}\in X$, of radius $\epsilon>0$, and metric induced by $|||\cdot|||_{\overline{\boldsymbol{\mu}}}$.

LEMMA 3.1. *Assume that $\beta_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})>0$. There can be at most one solution $\boldsymbol{u}(\boldsymbol{\mu})$ to* (2.1) *such that*

$$\boldsymbol{u}(\boldsymbol{\mu})\in B_{\overline{\boldsymbol{\mu}}}\left(\boldsymbol{u}_N(\boldsymbol{\mu}),\frac{\beta_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})}{\rho_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})}\right).$$

Proof. The proofs of this lemma and the following one are based on the original proofs in [4] and are extended to the case of two distinct norms. For these proofs we need to introduce the operators $A(\cdot;\boldsymbol{\mu}):X\mapsto X'$ and $DA(\boldsymbol{z},\boldsymbol{\mu}):X\mapsto X'$, for $\boldsymbol{z}$ in X, associated to a (and f) and da, and the operator $H(\cdot;\boldsymbol{\mu}):X\to X$,

$$\langle A(\boldsymbol{u};\boldsymbol{\mu}),\boldsymbol{v}\rangle = a(\boldsymbol{u},\boldsymbol{v};\boldsymbol{\mu})-f(\boldsymbol{v}) \quad \forall \boldsymbol{u},\boldsymbol{v}\in X,$$

$$\langle DA(\boldsymbol{z},\boldsymbol{\mu})\boldsymbol{u},\boldsymbol{v}\rangle = da(\boldsymbol{z};\boldsymbol{\mu})(\boldsymbol{u},\boldsymbol{v}) \quad \forall \boldsymbol{u},\boldsymbol{v}\in X, \tag{3.3}$$

$$H(\boldsymbol{w};\boldsymbol{\mu}) = \boldsymbol{w}-DA(\boldsymbol{u}_N(\boldsymbol{\mu});\boldsymbol{\mu})^{-1}A(\boldsymbol{w};\boldsymbol{\mu}). \tag{3.4}$$

$DA(\boldsymbol{u}_N(\boldsymbol{\mu});\boldsymbol{\mu})$ is invertible thanks to our assumption that $\beta_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})>0$. A fixed point of H must be a solution to (2.1) and vice versa; we are going to show that there exists at most one.

We consider $H(\boldsymbol{z}^2,\boldsymbol{\mu})-H(\boldsymbol{z}^1,\boldsymbol{\mu})$, which from the definition (3.4) can be expressed as

$$H(\boldsymbol{z}^2,\boldsymbol{\mu})-H(\boldsymbol{z}^1,\boldsymbol{\mu})=(\boldsymbol{z}^2-\boldsymbol{z}^1)-DA(\boldsymbol{u}_N(\boldsymbol{\mu});\boldsymbol{\mu})^{-1}\left(A(\boldsymbol{z}^2;\boldsymbol{\mu})-A(\boldsymbol{z}^1;\boldsymbol{\mu})\right). \tag{3.5}$$

We observe that, since our nonlinearity is quadratic,

$$A(\boldsymbol{z}^2;\boldsymbol{\mu})-A(\boldsymbol{z}^1;\boldsymbol{\mu})=DA\left(\frac{\boldsymbol{z}^1+\boldsymbol{z}^2}{2};\boldsymbol{\mu}\right)(\boldsymbol{z}^2-\boldsymbol{z}^1). \tag{3.6}$$

Applying (3.5) and (3.6) gives

$$\begin{aligned}&DA(\boldsymbol{u}_N(\boldsymbol{\mu});\boldsymbol{\mu})\left(H(\boldsymbol{z}^2;\boldsymbol{\mu})-H(\boldsymbol{z}^1;\boldsymbol{\mu})\right)\\&\quad=DA(\boldsymbol{u}_N(\boldsymbol{\mu});\boldsymbol{\mu})\left(\boldsymbol{z}^2-\boldsymbol{z}^1\right)-\left(A(\boldsymbol{z}^2;\boldsymbol{\mu})-A(\boldsymbol{z}^1;\boldsymbol{\mu})\right)\\&\quad=\left[DA(\boldsymbol{u}_N(\boldsymbol{\mu});\boldsymbol{\mu})-DA\left(\frac{\boldsymbol{z}^1+\boldsymbol{z}^2}{2};\boldsymbol{\mu}\right)\right](\boldsymbol{z}^2-\boldsymbol{z}^1).\end{aligned}$$

We then have, by recalling (2.8),

$$\begin{aligned}&\left\langle DA(\boldsymbol{u}_N(\boldsymbol{\mu});\boldsymbol{\mu})(H(\boldsymbol{z}^2;\boldsymbol{\mu})-H(\boldsymbol{z}^1;\boldsymbol{\mu})),\boldsymbol{v}\right\rangle\\&\quad\le\rho_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})\left|\!\left|\!\left|\boldsymbol{u}_N(\boldsymbol{\mu})-\frac{1}{2}(\boldsymbol{z}^2+\boldsymbol{z}^1)\right|\!\right|\!\right|_{\overline{\boldsymbol{\mu}}}|||\boldsymbol{z}^2-\boldsymbol{z}^1|||_{\overline{\boldsymbol{\mu}}}\|\boldsymbol{v}\|_X.\end{aligned}$$

Furthermore, from the definition of $\beta_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})$, $T_N^{\boldsymbol{\mu}}$, and $DA(\boldsymbol{u}_N(\boldsymbol{\mu});\boldsymbol{\mu})$, ((3.1), (2.4), and (3.3), respectively), we have

$$\begin{aligned}&\beta_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})|||H(\boldsymbol{z}^2;\boldsymbol{\mu})-H(\boldsymbol{z}^1;\boldsymbol{\mu})|||_{\overline{\boldsymbol{\mu}}}\|T_N^{\boldsymbol{\mu}}(H(\boldsymbol{z}^2;\boldsymbol{\mu})-H(\boldsymbol{z}^1;\boldsymbol{\mu}))\|_X\\&\quad\le\left\langle DA(\boldsymbol{u}_N(\boldsymbol{\mu});\boldsymbol{\mu})(H(\boldsymbol{z}^2;\boldsymbol{\mu})-H(\boldsymbol{z}^1;\boldsymbol{\mu})),T_N^{\boldsymbol{\mu}}(H(\boldsymbol{z}^2;\boldsymbol{\mu})-H(\boldsymbol{z}^1;\boldsymbol{\mu}))\right\rangle.\end{aligned} \tag{3.7}$$

Let $\alpha>0$; if $\boldsymbol{z}^1$ and $\boldsymbol{z}^2$ are in $B_{\overline{\boldsymbol{\mu}}}(\boldsymbol{u}_N(\boldsymbol{\mu}),\alpha)$, then $|||\boldsymbol{u}_N(\boldsymbol{\mu})-\frac{1}{2}(\boldsymbol{z}^2+\boldsymbol{z}^1)|||_{\overline{\boldsymbol{\mu}}}<\alpha$, and the last two inequalities give

$$|||H(\boldsymbol{z}^2;\boldsymbol{\mu})-H(\boldsymbol{z}^1;\boldsymbol{\mu})|||_{\overline{\boldsymbol{\mu}}}<\frac{\rho_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})\alpha}{\beta_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})}|||\boldsymbol{z}^2-\boldsymbol{z}^1|||_{\overline{\boldsymbol{\mu}}},$$

which implies that $H(\cdot;\boldsymbol{\mu})$ is a contraction mapping for

$$\alpha<\frac{\beta_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})}{\rho_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})}.$$

It follows that there can be only (at most) one fixed point of $H(\cdot;\mu)$ inside $B_{\overline{\boldsymbol{\mu}}}(\boldsymbol{u}_N(\boldsymbol{\mu}),\frac{\beta_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})}{\rho_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})})$ and hence at most one solution $\boldsymbol{u}(\boldsymbol{\mu})$. ∎

3.2. Existence. We define the following quantities:

$$\epsilon_N(\boldsymbol{\mu})\equiv\|A(\boldsymbol{u}_N(\boldsymbol{\mu});\boldsymbol{\mu})\|_{X'}, \tag{3.8}$$

$$\tau_{N,\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})\equiv\frac{2\rho_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})\epsilon_N(\boldsymbol{\mu})}{\beta_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})^2}, \tag{3.9}$$

$$\Delta_{N,\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})\equiv\frac{\beta_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})}{\rho_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})}\left[1-\sqrt{1-\tau_{N,\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})}\right]. \tag{3.10}$$

LEMMA 3.2. *If $\beta_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu}) > 0$ and $\tau_{N,\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu}) \le 1$, there exists a unique solution $\boldsymbol{u}(\boldsymbol{\mu})$ to (2.1) such that*

$$\boldsymbol{u}(\boldsymbol{\mu}) \in B_{\overline{\boldsymbol{\mu}}}\Big(\boldsymbol{u}_N(\boldsymbol{\mu}), \Delta_{N,\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})\Big).$$

Proof. Let $\alpha > 0$ and $\boldsymbol{z}$ in $B_{\overline{\boldsymbol{\mu}}}(\boldsymbol{u}_N(\boldsymbol{\mu}), \alpha)$; consider

$$\begin{aligned} H(\boldsymbol{z};\boldsymbol{\mu}) - \boldsymbol{u}_N(\boldsymbol{\mu}) &= \boldsymbol{z} - \boldsymbol{u}_N(\boldsymbol{\mu}) \\ &- DA(\boldsymbol{u}_N(\boldsymbol{\mu});\boldsymbol{\mu})^{-1}\left(A(\boldsymbol{z};\boldsymbol{\mu}) - A(\boldsymbol{u}_N(\boldsymbol{\mu});\boldsymbol{\mu})\right) - DA(\boldsymbol{u}_N(\boldsymbol{\mu});\boldsymbol{\mu})^{-1}A(\boldsymbol{u}_N(\boldsymbol{\mu});\boldsymbol{\mu}). \end{aligned}$$

By using arguments similar to those invoked in the previous lemma (among others, that $\boldsymbol{u}(\boldsymbol{\mu})$ is a fixed point of $H(\cdot;\boldsymbol{\mu})$), we have

$$\begin{aligned} &\langle DA(\boldsymbol{u}_N(\boldsymbol{\mu});\boldsymbol{\mu})(H(\boldsymbol{z};\boldsymbol{\mu}) - \boldsymbol{u}_N(\boldsymbol{\mu})), \boldsymbol{v}\rangle \\ &= \langle DA(\boldsymbol{u}_N(\boldsymbol{\mu});\boldsymbol{\mu})(\boldsymbol{z} - \boldsymbol{u}_N(\boldsymbol{\mu})), \boldsymbol{v}\rangle - \langle A(\boldsymbol{z};\boldsymbol{\mu}) - A(\boldsymbol{u}_N(\boldsymbol{\mu});\boldsymbol{\mu}), \boldsymbol{v}\rangle - \langle A(\boldsymbol{u}_N(\boldsymbol{\mu});\boldsymbol{\mu}), \boldsymbol{v}\rangle \\ &= \left\langle \left[DA(\boldsymbol{u}_N(\boldsymbol{\mu});\boldsymbol{\mu}) - DA\left(\frac{\boldsymbol{z} + \boldsymbol{u}_N(\boldsymbol{\mu})}{2};\boldsymbol{\mu}\right)\right](\boldsymbol{z} - \boldsymbol{u}_N(\boldsymbol{\mu})), \boldsymbol{v}\right\rangle - \langle A(\boldsymbol{u}_N(\boldsymbol{\mu});\boldsymbol{\mu}), \boldsymbol{v}\rangle \\ &\le \frac{1}{2}\rho_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})|||\boldsymbol{z} - \boldsymbol{u}_N(\boldsymbol{\mu})|||^2_{\overline{\boldsymbol{\mu}}}\,\|\boldsymbol{v}\|_X + \epsilon_{N,\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})\|\boldsymbol{v}\|_X \end{aligned} \tag{3.11}$$

for all $\boldsymbol{v}$ in X. Similarly to (3.7),

$$\begin{aligned} \beta_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})|||H(\boldsymbol{z};\boldsymbol{\mu}) - \boldsymbol{u}_N(\boldsymbol{\mu})|||_{\overline{\boldsymbol{\mu}}}\|T_N^{\boldsymbol{\mu}}(H(\boldsymbol{z};\boldsymbol{\mu}) - \boldsymbol{u}_N(\boldsymbol{\mu}))\|_X \\ \le \langle DA(\boldsymbol{u}_N(\boldsymbol{\mu});\boldsymbol{\mu})(H(\boldsymbol{z};\boldsymbol{\mu}) - \boldsymbol{u}_N(\boldsymbol{\mu})), T_N^{\boldsymbol{\mu}}(H(\boldsymbol{z};\boldsymbol{\mu}) - \boldsymbol{u}_N(\boldsymbol{\mu}))\rangle\,. \end{aligned} \tag{3.12}$$

Since $\boldsymbol{z}$ is in the open ball $B_{\overline{\boldsymbol{\mu}}}(\boldsymbol{u}_N(\boldsymbol{\mu}), \alpha)$, choosing $\boldsymbol{v} = T_N^{\boldsymbol{\mu}}(H(\boldsymbol{z};\boldsymbol{\mu}) - \boldsymbol{u}_N(\boldsymbol{\mu}))$, (3.11), and (3.12) gives

$$|||H(\boldsymbol{z};\boldsymbol{\mu}) - \boldsymbol{u}_N(\boldsymbol{\mu})|||_{\overline{\boldsymbol{\mu}}} \le \frac{\rho_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})}{2\beta_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})}|||\boldsymbol{z} - \boldsymbol{u}_N(\boldsymbol{\mu})|||^2_{\overline{\boldsymbol{\mu}}} + \frac{\epsilon_{N,\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})}{\beta_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})} < \frac{\rho_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})\alpha^2}{2\beta_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})} + \frac{\epsilon_{N,\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})}{\beta_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})}. \tag{3.13}$$

We now look for values of α (in particular, the smallest will give us a sharper error bound) such that

$$\frac{\rho_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})\alpha^2}{2\beta_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})} + \frac{\epsilon_{N,\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})}{\beta_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})} \le \alpha.$$

Provided that $\tau_{N,\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu}) \le 1$, we have that the smallest value is given by $\alpha = \Delta_{N,\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})$. Then for this choice and from (3.13) we have

$$|||H(\boldsymbol{z};\boldsymbol{\mu}) - \boldsymbol{u}_N(\boldsymbol{\mu})|||_{\overline{\boldsymbol{\mu}}} < \alpha;$$

i.e., H maps $\in B_{\overline{\boldsymbol{\mu}}}(\boldsymbol{u}_N(\boldsymbol{\mu}), \Delta_{N,\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu}))$ into itself. The proof follows from $\Delta_{N,\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu}) \le \frac{\beta_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})}{\rho_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})}$, the fact that H is a contraction (cf. the proof of the previous lemma), and the contraction mapping theorem. ∎

3.3. A posteriori error estimation.

THEOREM 3.3. *Let $\boldsymbol{\mu} \in \mathcal{D}$, and assume that, for a $\overline{\boldsymbol{\mu}} \in \mathcal{V}^K$, $\beta_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu}) > 0$ and $\tau_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu}) \le 1$. Then there exists a unique solution $\boldsymbol{u}(\boldsymbol{\mu})$ such that*

$$|||\boldsymbol{u}(\boldsymbol{\mu}) - \boldsymbol{u}_N(\boldsymbol{\mu})|||_{\overline{\boldsymbol{\mu}}} < \frac{\beta_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})}{\rho_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})};$$

the error with respect to the natural norm is

$$|||\boldsymbol{u}(\boldsymbol{\mu}) - \boldsymbol{u}_N(\boldsymbol{\mu})|||_{\overline{\boldsymbol{\mu}}} \leq \Delta_{N,\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu}) \tag{3.14}$$

with effectivity

$$\Delta_{N,\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu}) \leq \left[\frac{2\gamma_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})}{\beta_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})} + \tau_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})\right] |||\boldsymbol{u}(\boldsymbol{\mu}) - \boldsymbol{u}_N(\boldsymbol{\mu})|||_{\overline{\boldsymbol{\mu}}}. \tag{3.15}$$

This theorem provides an error bound with respect to the natural norms that, compared to previous results (cf. [17, 33]), first provides better effectivity in the field variable, since we expect $\gamma_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})$ and $\beta_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})$ to be of order one and second indirectly sharpens the existence's condition since in (3.9) again $\beta_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})$ is of order one.

Note that we expect that $\gamma_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})$ and $\beta_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})$ are of order one, since $\overline{\boldsymbol{\mu}} = \mathcal{I}^K(\boldsymbol{\mu})$ is "close" to $\boldsymbol{\mu}$ and hence $T_N^{\boldsymbol{\mu}}$ and $T_N^{\overline{\boldsymbol{\mu}}}$ have similar spectra. In (3.9) the term that is the worse controlled is $\rho_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})$; this is related to a Sobolev embedding constant (cf. [33] and section 6). However, the equivalent condition in [33] reads

$$\tau_N(\boldsymbol{\mu}) = \frac{2\rho\,\epsilon_N(\boldsymbol{\mu})}{\beta_N(\boldsymbol{\mu})^2},$$

where ρ and $\beta_N(\boldsymbol{\mu})$ are the Sobolev embedding constants and equivalent to (2.8) and (3.1), respectively, with respect to the X-norm only. Since $\beta_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu}) \geq \frac{\beta_N(\boldsymbol{\mu})}{\gamma(\overline{\boldsymbol{\mu}})}$ and $\rho_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu}) \leq \frac{\rho}{\beta(\overline{\boldsymbol{\mu}})^2}$, we have

$$\tau_{N,\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu}) \leq \tau_N(\boldsymbol{\mu}) \frac{\gamma(\overline{\boldsymbol{\mu}})^2}{\beta(\overline{\boldsymbol{\mu}})^2};$$

i.e., if $\gamma_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})$ and $\beta_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})$ are of order one, the condition $\tau_{N,\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu}) \leq 1$ is less stringent than $\tau_N(\boldsymbol{\mu}) < 1$, which is required in [33].

In our simulations, for a large Grashof number, we actually get that $\rho_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})$ is one order smaller than $\frac{\rho}{\beta_N(\boldsymbol{\mu})^2}$, and hence $\tau_{N,\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})$ is also one order smaller than $\tau_N(\boldsymbol{\mu})$.

Proof. We just need to prove (3.15). Let $\boldsymbol{e}(\boldsymbol{\mu}) = \boldsymbol{u}(\boldsymbol{\mu}) - \boldsymbol{u}_N(\boldsymbol{\mu})$ and $\hat{\boldsymbol{e}}(\boldsymbol{\mu})$ such that

$$\begin{aligned}(\hat{\boldsymbol{e}}(\boldsymbol{\mu}), \boldsymbol{v})_X &= -\langle A(\boldsymbol{u}_N(\boldsymbol{\mu}); \boldsymbol{\mu}), \boldsymbol{v}\rangle \\ &= f(\boldsymbol{v}) - a(\boldsymbol{u}_N(\boldsymbol{\mu}), \boldsymbol{v}; \boldsymbol{\mu}) = a(\boldsymbol{u}(\boldsymbol{\mu}), \boldsymbol{v}; \boldsymbol{\mu}) - a(\boldsymbol{u}_N(\boldsymbol{\mu}), \boldsymbol{v}; \boldsymbol{\mu}).\end{aligned} \tag{3.16}$$

Note that, from (3.8), $\|\hat{\boldsymbol{e}}(\boldsymbol{\mu})\|_X = \epsilon_{N,\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})$ and that, since our nonlinearity is quadratic,

$$\begin{aligned}&a(\boldsymbol{u}(\boldsymbol{\mu}), \boldsymbol{v}; \boldsymbol{\mu}) - a(\boldsymbol{u}_N(\boldsymbol{\mu}), \boldsymbol{v}; \boldsymbol{\mu}) \\ &\qquad = da(\boldsymbol{u}_N(\boldsymbol{\mu}); \boldsymbol{\mu})(\boldsymbol{e}(\boldsymbol{\mu}), \boldsymbol{v}) + \frac{1}{2}\Big(da(\boldsymbol{u}(\boldsymbol{\mu}); \boldsymbol{\mu}) - da(\boldsymbol{u}_N(\boldsymbol{\mu}); \boldsymbol{\mu})\Big)(\boldsymbol{e}(\boldsymbol{\mu}), \boldsymbol{v})\end{aligned}$$

for all $\boldsymbol{v}$ in X. This, (3.16), (3.2), and (2.8) yield to

$$\begin{aligned}\|\hat{\boldsymbol{e}}(\boldsymbol{\mu})\|_X^2 &= da(\boldsymbol{u}_N(\boldsymbol{\mu}); \boldsymbol{\mu})(\boldsymbol{e}(\boldsymbol{\mu}), \hat{\boldsymbol{e}}(\boldsymbol{\mu})) + \frac{1}{2}\Big(da(\boldsymbol{u}(\boldsymbol{\mu}); \boldsymbol{\mu}) - da(\boldsymbol{u}_N(\boldsymbol{\mu}); \boldsymbol{\mu})\Big)(\boldsymbol{e}(\boldsymbol{\mu}), \hat{\boldsymbol{e}}(\boldsymbol{\mu})) \\ &\leq \gamma_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu}) |||\boldsymbol{e}(\boldsymbol{\mu})|||_{\overline{\boldsymbol{\mu}}} \|\hat{\boldsymbol{e}}(\boldsymbol{\mu})\|_X + \frac{1}{2}\rho_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})\, |||\boldsymbol{e}(\boldsymbol{\mu})|||_{\overline{\boldsymbol{\mu}}}^2 \|\hat{\boldsymbol{e}}(\boldsymbol{\mu})\|_X\end{aligned}$$

and therefore

$$\epsilon_{N,\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu}) \le \gamma_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})|||\boldsymbol{e}(\boldsymbol{\mu})|||_{\overline{\boldsymbol{\mu}}} + \frac{1}{2}\rho_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})|||\boldsymbol{e}(\boldsymbol{\mu})|||^2_{\overline{\boldsymbol{\mu}}}. \tag{3.17}$$

Since $0 \le \tau_{N,\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu}) \le 1$ and $1 - \sqrt{1 - \tau_{N,\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})} \le \tau_{N,\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})$, from (3.9)–(3.10) we get

$$\Delta_{N,\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu}) \le 2\frac{\epsilon_{N,\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})}{\beta_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})} \quad \text{and} \tag{3.18}$$

$$\frac{\rho_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})}{\beta_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})}\Delta_{N,\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu}) \le \tau_{N,\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu}). \tag{3.19}$$

Then, from (3.17) and (3.18)

$$\Delta_{N,\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu}) \le \frac{2\gamma_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})}{\beta_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})}|||\boldsymbol{e}(\boldsymbol{\mu})|||_{\overline{\boldsymbol{\mu}}} + \frac{\rho_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})}{\beta_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})}|||\boldsymbol{e}(\boldsymbol{\mu})|||^2_{\overline{\boldsymbol{\mu}}};$$

we split the square of $|||\boldsymbol{e}(\boldsymbol{\mu})|||_{\overline{\boldsymbol{\mu}}}$ and get

$$\begin{aligned}\Delta_{N,\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu}) &\le \frac{2\gamma_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})}{\beta_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})}|||\boldsymbol{e}(\boldsymbol{\mu})|||_{\overline{\boldsymbol{\mu}}} + \frac{\rho_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})}{\beta_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})}|||e(\boldsymbol{\mu})|||_{\overline{\boldsymbol{\mu}}}\Delta_{N,\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu}) \\ &\le \frac{2\gamma_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})}{\beta_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})}|||\boldsymbol{e}(\boldsymbol{\mu})|||_{\overline{\boldsymbol{\mu}}} + \tau_{N,\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})|||\boldsymbol{e}(\boldsymbol{\mu})|||_{\overline{\boldsymbol{\mu}}},\end{aligned}$$

thanks to (3.14) in the first inequality and (3.19) in the second one. ∎

3.4. Output error bound. Often the precision required in the prediction of the output is obtained by the primal approximation alone. This is the case in the model problems that we propose; in fact, the requirements to ensure existence and unicity already imply a very small error in the output.

For a given $\boldsymbol{\mu} \in \mathcal{D}$, we are interested in bounding the error on the output $s(\boldsymbol{\mu}) = \ell(\boldsymbol{u}(\boldsymbol{\mu}))$. For $\overline{\boldsymbol{\mu}} \in \mathcal{V}^K$, the *natural* dual norm of the output linear functional is defined as

$$|||\ell|||_{\overline{\boldsymbol{\mu}}'} = \sup_{\boldsymbol{v}\in Y} \frac{|\ell(\boldsymbol{v})|}{|||\boldsymbol{v}|||_{\overline{\boldsymbol{\mu}}}}.$$

(Note that, in the case of an affine output, $|||\ell|||_{\overline{\boldsymbol{\mu}}'} \le \sum_{q=1}^{Q_\ell} |\theta^q_\ell(\boldsymbol{\mu})|\ |||\ell^q|||_{\overline{\boldsymbol{\mu}}'}$.) If $\beta_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu}) > 0$ and $\tau_{N,\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu}) \le 1$, we can then state the output error bound as

$$|s(\boldsymbol{\mu}) - s_N(\boldsymbol{\mu})| \le |||\ell|||_{\overline{\boldsymbol{\mu}}'}\Delta_{N,\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu}) \equiv \Delta^s_{N,\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu}).$$

However, we do not have any effectivity result about these bounds. In our example, the effectivities are at high values of the Grashof number reasonable but not as accurate at low values. In the appendix we propose a dual approach that partially addresses this issue; further improvements shall require deflation (cf. [32]).

4. Lower bound of the inf-sup constant. We are going to build a lower bound for the inf-sup parameter $\beta_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})$. The ingredients, similarly to [32], are the selection of a candidate supremizer, a Taylor expansion with the remainder, and a lower bound for the terms in the Taylor expansion.

It is clear that, for $\boldsymbol{\mu} = \overline{\boldsymbol{\mu}}$ in (3.1) and (3.2), $\beta_{\overline{\boldsymbol{\mu}}}(\overline{\boldsymbol{\mu}}) = \gamma_{\overline{\boldsymbol{\mu}}}(\overline{\boldsymbol{\mu}}) = 1$; the natural norm can thus be viewed as a generalization of the usual energy norm (for symmetric, coercive operators) to the nonlinear case. It can be demonstrated that

$$\beta(\overline{\boldsymbol{\mu}})\beta_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu}) \le \beta(\boldsymbol{\mu}) \le \beta_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})\gamma(\overline{\boldsymbol{\mu}}).$$

In what follows we shall require an "intermediate" inf-sup parameter—an approximation to $\beta_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})$—which we denote as $\overline{\beta}_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})$: For given $\boldsymbol{\mu} \in \mathcal{D}$ and $\overline{\boldsymbol{\mu}} \in \mathcal{V}^K$,

$$\overline{\beta}_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu}) \equiv \inf_{\boldsymbol{w}\in X} \frac{(T_N^{\boldsymbol{\mu}}\boldsymbol{w}, T_N^{\overline{\boldsymbol{\mu}}}\boldsymbol{w})_X}{|||\boldsymbol{w}|||^2_{\overline{\boldsymbol{\mu}}}}.$$

It follows directly from the Cauchy–Schwarz inequality—or, equivalently, we may observe that $T_N^{\overline{\boldsymbol{\mu}}}\boldsymbol{w}$ is a candidate supremizer $\boldsymbol{v}$ in (3.1)—that

$$\overline{\beta}_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu}) \le \beta_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu}) \quad \forall\, \boldsymbol{\mu} \in \mathcal{D}.$$

(Note that $\overline{\beta}_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})$ is not necessarily positive.)

We can also show that $\overline{\beta}_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})$ is a "good" lower bound for $\beta_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})$.

LEMMA 4.1. *Let*

$$\delta_{\overline{\boldsymbol{\mu}}}^{\boldsymbol{\mu}} = \sup_{\boldsymbol{w}\in X} \frac{\left(T_N^{\boldsymbol{\mu}}\boldsymbol{w} - T_N^{\overline{\boldsymbol{\mu}}}\boldsymbol{w}, T_N^{\boldsymbol{\mu}}\boldsymbol{w} + T_N^{\overline{\boldsymbol{\mu}}}\boldsymbol{w}\right)_X}{\|T_N^{\overline{\boldsymbol{\mu}}}\boldsymbol{w}\|_X^2},$$

and assume that

$$\sup_{\boldsymbol{w}\in X} \frac{\|T_N^{\boldsymbol{\mu}}\boldsymbol{w} - T_N^{\overline{\boldsymbol{\mu}}}\boldsymbol{w}\|_X}{\|T_N^{\overline{\boldsymbol{\mu}}}\boldsymbol{w}\|_X} = o(\delta_{\overline{\boldsymbol{\mu}}}^{\boldsymbol{\mu}})$$

as $\boldsymbol{\mu} \to \overline{\boldsymbol{\mu}}$. *Then*

$$1 - \beta_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu}) = -\frac{1}{2}\delta_{\overline{\boldsymbol{\mu}}}^{\boldsymbol{\mu}} + o(|\delta_{\overline{\boldsymbol{\mu}}}^{\boldsymbol{\mu}}|^2), \tag{4.1}$$

and there exists $c(\overline{\boldsymbol{\mu}}, \boldsymbol{\mu})$ *in* $[0,1]$ *such that*

$$c(\overline{\boldsymbol{\mu}}, \boldsymbol{\mu})\,[1 - \beta_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})] + 2\left[\,\overline{\beta}_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu}) - \beta_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})\right] = o(|\delta_{\overline{\boldsymbol{\mu}}}^{\boldsymbol{\mu}}|^2). \tag{4.2}$$

In general, we can assume that $\delta_{\overline{\boldsymbol{\mu}}}^{\boldsymbol{\mu}}$ is of order $o(|\boldsymbol{\mu} - \overline{\boldsymbol{\mu}}|)$; hence, as for the linear case [32], $1 - \beta_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu}) = o(|\boldsymbol{\mu} - \overline{\boldsymbol{\mu}}|)$ as $\boldsymbol{\mu} \to \overline{\boldsymbol{\mu}}$ and

$$0 \le \beta_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu}) - \overline{\beta}_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu}) \le \frac{1}{2}\,[1 - \beta_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})] + o(|\overline{\boldsymbol{\mu}} - \boldsymbol{\mu}|^2).$$

Proof. From the definition of $\beta_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})$ and $\overline{\beta}_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})$, $1 - \beta_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})^2 = -\delta_{\overline{\boldsymbol{\mu}}}^{\boldsymbol{\mu}}$ and $\overline{\beta}_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})^2 \le \beta_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})^2 \le \overline{\beta}_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})^2 + \frac{1}{2}\delta_{\overline{\boldsymbol{\mu}}}^{\boldsymbol{\mu}} + o(|\delta_{\overline{\boldsymbol{\mu}}}^{\boldsymbol{\mu}}|^2)$; hence

$$\overline{\beta}_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})^2 - \beta_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})^2 = -\frac{c(\overline{\boldsymbol{\mu}}, \boldsymbol{\mu})}{2}\delta_{\overline{\boldsymbol{\mu}}}^{\boldsymbol{\mu}} + o(|\delta_{\overline{\boldsymbol{\mu}}}^{\boldsymbol{\mu}}|^2),$$

where $c(\overline{\boldsymbol{\mu}}, \boldsymbol{\mu}) \in [0,1]$. From the Taylor series of the square root follows (4.1) and

$$\overline{\beta}_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu}) - \beta_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu}) = \frac{c(\overline{\boldsymbol{\mu}}, \boldsymbol{\mu})}{4}\frac{\delta_{\overline{\boldsymbol{\mu}}}^{\boldsymbol{\mu}}}{\beta_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})} + o(|\delta_{\overline{\boldsymbol{\mu}}}^{\boldsymbol{\mu}}|^2).$$

Since $\frac{1}{1-x} = 1 + x + o(x)$, and from (4.1), $\overline{\beta}_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu}) - \beta_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu}) = \frac{c(\overline{\boldsymbol{\mu}}, \boldsymbol{\mu})}{4}\delta_{\overline{\boldsymbol{\mu}}}^{\boldsymbol{\mu}} + o(|\delta_{\overline{\boldsymbol{\mu}}}^{\boldsymbol{\mu}}|^2)$, which proves (4.2). ∎

4.1. Local lower bound. For a fixed $\overline{\boldsymbol{\mu}}$ in $\mathcal{V}^K$, we expand $\overline{\beta}_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})$ as follows:

(4.3)
$$\overline{\beta}_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu}) = \inf_{\boldsymbol{w}\in X} \frac{\left(T_N^{\boldsymbol{\mu}}\boldsymbol{w}, T_{\overline{N}}^{\overline{\boldsymbol{\mu}}}\boldsymbol{w}\right)_X}{|||\boldsymbol{w}|||^2_{\overline{\boldsymbol{\mu}}}}$$

$$= \inf_{\boldsymbol{w}\in X} \left\{ 1 + \frac{1}{|||\boldsymbol{w}|||^2_{\overline{\boldsymbol{\mu}}}} \sum_{q=1}^{Q_0} \left(\Theta_0^q(\boldsymbol{\mu}) - \Theta_0^q(\overline{\boldsymbol{\mu}})\right) a_0^q(\boldsymbol{w}, T_{\overline{N}}^{\overline{\boldsymbol{\mu}}}\boldsymbol{w}) \right. \tag{4.3a}$$

$$+ \frac{1}{|||\boldsymbol{w}|||^2_{\overline{\boldsymbol{\mu}}}} \sum_{q=1}^{Q_1} \left(\Theta_1^q(\boldsymbol{\mu}) - \Theta_1^q(\overline{\boldsymbol{\mu}})\right) a_1^q(\boldsymbol{w}, \boldsymbol{u}_{\overline{N}}(\overline{\boldsymbol{\mu}}), T_{\overline{N}}^{\overline{\boldsymbol{\mu}}}\boldsymbol{w}) \tag{4.3b}$$

$$+ \frac{1}{|||\boldsymbol{w}|||^2_{\overline{\boldsymbol{\mu}}}} \sum_{q=1}^{Q_1} \Theta_1^q(\overline{\boldsymbol{\mu}})\, a_1^q(\boldsymbol{w},\, \boldsymbol{u}_N(\boldsymbol{\mu}) - \boldsymbol{u}_{\overline{N}}(\overline{\boldsymbol{\mu}})\,, T_{\overline{N}}^{\overline{\boldsymbol{\mu}}}\boldsymbol{w}) \tag{4.3c}$$

$$\left. + \frac{1}{|||\boldsymbol{w}|||^2_{\overline{\boldsymbol{\mu}}}} \sum_{q=1}^{Q_1} \left(\Theta_1^q(\boldsymbol{\mu}) - \Theta_1^q(\overline{\boldsymbol{\mu}})\right) a_1^q(\boldsymbol{w},\, \boldsymbol{u}_N(\mu) - \boldsymbol{u}_{\overline{N}}(\overline{\boldsymbol{\mu}})\,, T_{\overline{N}}^{\overline{\boldsymbol{\mu}}}\boldsymbol{w}) \right\}. \tag{4.3d}$$

We approximate the partial derivative of $\frac{\partial \boldsymbol{u}_{\overline{N}}}{\partial \mu_p}$ at $\overline{\boldsymbol{\mu}}$ with respect to the pth parameter in our reduced space $W_{\overline{N}}$:

$$da(\boldsymbol{u}_{\overline{N}}(\overline{\boldsymbol{\mu}}); \overline{\boldsymbol{\mu}}) \left(\frac{\partial \boldsymbol{u}_{\overline{N}}}{\partial \mu_p}(\overline{\boldsymbol{\mu}}), \boldsymbol{v} \right) \tag{4.4}$$
$$= -\left[\sum_{q=1}^{Q_0} \frac{\partial \Theta_0^q(\overline{\boldsymbol{\mu}})}{\partial \mu_p} a_0^q(\boldsymbol{u}_{\overline{N}}(\overline{\boldsymbol{\mu}}), \boldsymbol{v}) + \frac{1}{2} \sum_{q=1}^{Q_1} \frac{\partial \Theta_1^q(\overline{\boldsymbol{\mu}})}{\partial \mu_p} a_1^q(\boldsymbol{u}_{\overline{N}}(\overline{\boldsymbol{\mu}}), \boldsymbol{u}_{\overline{N}}(\overline{\boldsymbol{\mu}}), \boldsymbol{v}) \right]$$

for all $\boldsymbol{v}$ in $W_{\overline{N}}$. (If f affinely depends on the parameter, we have to add $\sum_{q=1}^{Q_f} \frac{\partial \Theta^f(\overline{\boldsymbol{\mu}})}{\partial \mu_p} f^q(\boldsymbol{v})$ to the right-hand side of (4.4).) We then rewrite

$$a_1^q(\boldsymbol{w}, \boldsymbol{u}_N(\boldsymbol{\mu}) - \boldsymbol{u}_{\overline{N}}(\overline{\boldsymbol{\mu}}), T_{\overline{N}}^{\overline{\boldsymbol{\mu}}}\boldsymbol{w})$$
$$= \sum_{p=1}^{P} \kappa_p \left(\mu_p - \overline{\mu}_p\right) a_1^q \left(\boldsymbol{w}, \frac{\partial \boldsymbol{u}_{\overline{N}}}{\partial \mu_p}(\overline{\mu}), T_{\overline{N}}^{\overline{\boldsymbol{\mu}}} w \right) + a_1^q(\boldsymbol{w}, \boldsymbol{\alpha}(\boldsymbol{\mu}, \overline{\boldsymbol{\mu}}, \boldsymbol{\kappa}), T_{\overline{N}}^{\overline{\boldsymbol{\mu}}}\boldsymbol{w}),$$

where $\boldsymbol{\kappa} \in \mathbb{R}^P$ and $\boldsymbol{\alpha}(\boldsymbol{\mu}, \overline{\boldsymbol{\mu}}, \boldsymbol{\kappa}) \in W_{N_{\max}}$ is defined as

$$\boldsymbol{\alpha}(\boldsymbol{\mu}, \overline{\boldsymbol{\mu}}, \boldsymbol{\kappa}) = \boldsymbol{u}_N(\boldsymbol{\mu}) - \boldsymbol{u}_{\overline{N}}(\overline{\boldsymbol{\mu}}) - \sum_{p=1}^{P} \kappa_p \left(\mu_p - \overline{\mu}_p\right) \frac{\partial \boldsymbol{u}_{\overline{N}}}{\partial \mu_p}(\overline{\boldsymbol{\mu}}).$$

As for the linear case [32], we perform a Taylor expansion with respect to $\boldsymbol{\mu}$ at $\overline{\boldsymbol{\mu}}$. The first-order terms in (4.3a), (4.3b), and (4.3c) are collected: For $p = 1, \ldots, P$,

(4.5)
$$a_p^{\overline{\boldsymbol{\mu}}}(\cdot,\cdot) \equiv \sum_{q=1}^{Q_0} \frac{\partial \Theta_0^q}{\partial \mu_p}(\overline{\boldsymbol{\mu}}) a_0^q(\cdot,\cdot) + \sum_{q=1}^{Q_1} \frac{\partial \Theta_1^q}{\partial \mu_p}(\overline{\boldsymbol{\mu}}) a_1^q(\cdot, \boldsymbol{u}_{\overline{N}}(\overline{\boldsymbol{\mu}}), \cdot) + \sum_{q=1}^{Q_1} \Theta_1^q(\overline{\boldsymbol{\mu}}) a_1^q \left(\cdot, \frac{\partial \boldsymbol{u}_{\overline{N}}}{\partial \mu_p}(\overline{\boldsymbol{\mu}}), \cdot\right).$$

We denote the extreme eigenvalues with respect to $|||\cdot|||_{\overline{\boldsymbol{\mu}}}$ of $a_p^{\overline{\boldsymbol{\mu}}}$, a_0^q, and $a_1^q(\cdot, \boldsymbol{u}_{\overline{N}}(\overline{\boldsymbol{\mu}}), \cdot)$ as

$$\lambda_{p,\inf}^{\overline{\boldsymbol{\mu}}} = \inf_{\boldsymbol{w}\in X} \frac{a_p^{\overline{\boldsymbol{\mu}}}(\boldsymbol{w}, T_{\overline{N}}^{\overline{\boldsymbol{\mu}}}\boldsymbol{w})}{|||\boldsymbol{w}|||_{\overline{\boldsymbol{\mu}}}^2}, \quad \lambda_{p,\sup}^{\overline{\boldsymbol{\mu}}} = \sup_{\boldsymbol{w}\in X} \frac{a_p^{\overline{\boldsymbol{\mu}}}(\boldsymbol{w}, T_{\overline{N}}^{\overline{\boldsymbol{\mu}}}\boldsymbol{w})}{|||\boldsymbol{w}|||_{\overline{\boldsymbol{\mu}}}^2}, \quad p = 1, \ldots, P,$$

$$\gamma_{0,q,\inf}^{\overline{\boldsymbol{\mu}}} = \inf_{\boldsymbol{w}\in X} \frac{a_0^q(\boldsymbol{w}, T_{\overline{N}}^{\overline{\boldsymbol{\mu}}}\boldsymbol{w})}{|||\boldsymbol{w}|||_{\overline{\boldsymbol{\mu}}}^2}, \quad \gamma_{0,q,\sup}^{\overline{\boldsymbol{\mu}}} = \sup_{\boldsymbol{w}\in X} \frac{a_0^q(\boldsymbol{w}, T_{\overline{N}}^{\overline{\boldsymbol{\mu}}}\boldsymbol{w})}{|||\boldsymbol{w}|||_{\overline{\boldsymbol{\mu}}}^2}, \quad q = 1, \ldots, Q_0,$$

$$\gamma_{1,q,\inf}^{\overline{\boldsymbol{\mu}}} = \inf_{\boldsymbol{w}\in X} \frac{a_1^q(\boldsymbol{w}, \boldsymbol{u}_{\overline{N}}(\overline{\boldsymbol{\mu}}), T_{\overline{N}}^{\overline{\boldsymbol{\mu}}}\boldsymbol{w})}{|||\boldsymbol{w}|||_{\overline{\boldsymbol{\mu}}}^2},$$

$$\gamma_{1,q,\sup}^{\overline{\boldsymbol{\mu}}} = \sup_{\boldsymbol{w}\in X} \frac{a_1^q(\boldsymbol{w}, \boldsymbol{u}_{\overline{N}}(\overline{\boldsymbol{\mu}}), T_{\overline{N}}^{\overline{\boldsymbol{\mu}}}\boldsymbol{w})}{|||\boldsymbol{w}|||_{\overline{\boldsymbol{\mu}}}^2}, \quad q = 1, \ldots, Q_1.$$

From our assumption (2.9) we have

$$\frac{1}{|||\boldsymbol{w}|||_{\overline{\boldsymbol{\mu}}}^2} \left| \sum_{q=1}^{Q_1} \Theta_1^q(\overline{\boldsymbol{\mu}}) a_1^q(\boldsymbol{w}, \boldsymbol{\alpha}(\boldsymbol{\mu}, \overline{\boldsymbol{\mu}}, \boldsymbol{\kappa}), T_{\overline{N}}^{\overline{\boldsymbol{\mu}}}\boldsymbol{w}) \right| \le \rho_{X,\overline{\boldsymbol{\mu}}} \|\boldsymbol{\alpha}(\boldsymbol{\mu}, \overline{\boldsymbol{\mu}}, \boldsymbol{\kappa})\|_X. \tag{4.6}$$

Similarly, from (2.10),

$$\frac{1}{|||\boldsymbol{w}|||_{\overline{\boldsymbol{\mu}}}^2} \left| \sum_{q=1}^{Q_1} \left(\Theta_1^q(\boldsymbol{\mu}) - \Theta_1^q(\overline{\boldsymbol{\mu}})\right) a_1^q(\boldsymbol{w}, \boldsymbol{u}_N(\boldsymbol{\mu}) - \boldsymbol{u}_{\overline{N}}(\overline{\boldsymbol{\mu}}), T_{\overline{N}}^{\overline{\boldsymbol{\mu}}}\boldsymbol{w}) \right|$$
$$\le \max_{q=1,\ldots,Q_1} \frac{|\Theta_1^q(\boldsymbol{\mu}) - \Theta_1^q(\overline{\boldsymbol{\mu}})|}{|\Theta_1^q(\overline{\boldsymbol{\mu}})|} \rho_{X,\overline{\boldsymbol{\mu}}} \|\boldsymbol{u}_N(\boldsymbol{\mu}) - \boldsymbol{u}_{\overline{N}}(\overline{\boldsymbol{\mu}})\|_X.$$

As a result, our bound to (4.3) is defined, for any $\boldsymbol{\kappa}$ in $\mathbb{R}^P$, as

$$\begin{aligned}
\overline{\beta}_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu}) \ge & \overline{B}_{\overline{\boldsymbol{\mu}}}^{LB}(\boldsymbol{\mu}, \boldsymbol{\kappa}) \equiv 1 + \sum_{p=1}^{P} \min_{\lambda_p = \lambda_{p,\,\inf\,|\,\sup}^{\overline{\boldsymbol{\mu}}}} \kappa_p(\mu_p - \overline{\mu}_p)\lambda_p \\
& + \sum_{q=1}^{Q_0} \min_{\gamma_q = \gamma_{0,q,\,\inf\,|\,\sup}^{\overline{\boldsymbol{\mu}}}} \left(\Theta_0^q(\boldsymbol{\mu}) - \Theta_0^q(\overline{\boldsymbol{\mu}}) - \sum_{p=1}^{P} \kappa_p(\mu_p - \overline{\mu}_p) \frac{\partial \Theta_0^q}{\partial \mu_p}(\overline{\boldsymbol{\mu}}) \right) \gamma_q \\
& + \sum_{q=1}^{Q_1} \min_{\gamma_q = \gamma_{1,q,\,\inf\,|\,\sup}^{\overline{\boldsymbol{\mu}}}} \left(\Theta_1^q(\boldsymbol{\mu}) - \Theta_1^q(\overline{\boldsymbol{\mu}}) - \sum_{p=1}^{P} \kappa_p(\mu_p - \overline{\mu}_p) \frac{\partial \Theta_1^q}{\partial \mu_p}(\overline{\boldsymbol{\mu}}) \right) \gamma_q \\
& - \rho_{X,\overline{\boldsymbol{\mu}}} \left(\|\boldsymbol{\alpha}(\boldsymbol{\mu}, \overline{\boldsymbol{\mu}}, \boldsymbol{\kappa})\|_X + \max_{q=1,\ldots,Q_1} \frac{|\Theta_1^q(\boldsymbol{\mu}) - \Theta_1^q(\overline{\boldsymbol{\mu}})|}{|\Theta_1^q(\overline{\boldsymbol{\mu}})|} \|\boldsymbol{u}_N(\boldsymbol{\mu}) - \boldsymbol{u}_{\overline{N}}(\overline{\boldsymbol{\mu}})\|_X \right),
\end{aligned} \tag{4.7}$$

where we denote the minimum of a function g over two values as

$$\min_{\lambda=\lambda_{\inf\,|\,\sup}} g(\lambda) \equiv \min\left\{g(\lambda_{\inf}), g(\lambda_{\sup})\right\}.$$

Since (4.7) is valid for any $\boldsymbol{\kappa} \in \mathbb{R}^P$, we define our inf-sup lower bound as

$$\overline{B}_{\overline{\boldsymbol{\mu}}}^{LB}(\boldsymbol{\mu}) \equiv \max_{\boldsymbol{\kappa}\in\mathbb{R}^P} \overline{B}_{\overline{\boldsymbol{\mu}}}^{LB}(\boldsymbol{\mu}, \boldsymbol{\kappa}). \tag{4.8}$$

4.2. $\boldsymbol{\kappa}$ optimization. Ideally, for a given $\boldsymbol{\mu}$, we have to find the value of $\boldsymbol{\kappa}$ that maximizes $\overline{B}^{LB}_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu},\boldsymbol{\kappa})$. The optimization of $\boldsymbol{\kappa}$ relies on quadratic programming. We propose a simpler heuristic method: The bound involving the eigenvalues $\lambda^{\overline{\boldsymbol{\mu}}}_{p,\inf|\sup}$ is sharper than (4.6), and therefore we compute $\boldsymbol{\kappa}$ such that $\|\boldsymbol{\alpha}(\boldsymbol{\mu},\overline{\boldsymbol{\mu}},\boldsymbol{\kappa})\|_X$ is the smallest. Moreover, in our algorithms we will require an X-orthonormal basis for $W_{N_{\max}}$; hence this minimization is achieved by the least-square solution $\boldsymbol{\kappa}^*$ that satisfies

$$\begin{aligned}(4.9)\qquad &\left[\begin{array}{c}\vdots\\ (\mu_p-\overline{\mu}_p)\partial_p U_{\overline{N},j}(\overline{\boldsymbol{\mu}})\\ \vdots\end{array}\right]^T_{j,p}\left[\begin{array}{c}\vdots\\ (\mu_p-\overline{\mu}_p)\partial_p U_{\overline{N},j}(\overline{\boldsymbol{\mu}})\\ \vdots\end{array}\right]_{j,p}\boldsymbol{\kappa}^*\\ &\qquad=\left[\begin{array}{c}\vdots\\ (\mu_p-\overline{\mu}_p)\partial_p U_{\overline{N},j}(\overline{\boldsymbol{\mu}})\\ \vdots\end{array}\right]^T_{j,p}\left[\begin{array}{c}\vdots\\ U_{N,j}(\boldsymbol{\mu})-U_{\overline{N},j}(\overline{\boldsymbol{\mu}})\\ \vdots\end{array}\right]_{j},\end{aligned}$$

where $U_{N,j}(\boldsymbol{\mu})$, $U_{\overline{N},j}(\overline{\boldsymbol{\mu}})$, and $\partial_p U_{\overline{N},j}(\overline{\boldsymbol{\mu}})$ are the jth components of $\boldsymbol{u}_N(\boldsymbol{\mu})$, $\boldsymbol{u}_{\overline{N}}(\overline{\boldsymbol{\mu}})$, and $\frac{\partial \boldsymbol{u}_{\overline{N}}(\overline{\boldsymbol{\mu}})}{\partial \mu_p}$ with respect to the basis of $W_{N_{\max}}$.

5. Computational strategy.

5.1. Offline algorithm. Our offline algorithm includes the computation of the reduced basis ingredients—optimal selection of the basis and computation of matrices and vector in the reduced space (cf., e.g., [32])—and of the a posteriori error estimation ingredients—selection of $\mathcal{V}^K=\{\overline{\boldsymbol{\mu}}_1,\dots,\overline{\boldsymbol{\mu}}_K\}$, solution of the eigenvalue problems, and computation of the Sobolev embedding constants. Because of the presence of the reduced basis approximations $\boldsymbol{u}_{\overline{N}}(\overline{\boldsymbol{\mu}})$ in the definition of the natural norms, these two stages are dependent on each other. (In the linear case (cf. [32]) these steps are ordered: first to compute the a posteriori error estimation ingredients, then to select the reduced basis.)

In the first step, we provide surrogates $\overline{B}^{LB}_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})$ and $\rho_{\overline{\boldsymbol{\mu}}}(\overline{\boldsymbol{\mu}})$ in order to compute an error bound extrapolation. We expect that $\overline{B}^{LB}_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})$ is of order one, so we set its surrogate to, say, 0.2; to compute a surrogate $\rho_{\overline{\boldsymbol{\mu}}}(\overline{\boldsymbol{\mu}})$, we replace $\boldsymbol{u}_{\overline{N}}(\overline{\boldsymbol{\mu}})$ by $\boldsymbol{u}(\overline{\boldsymbol{\mu}})$ in (2.4) (hence in (2.7))—in other words we do not need a reduced basis. As a result we have an efficient—online fast—tool to approximately compute the error for a given $\boldsymbol{\mu}$ in $\mathcal{D}$.

In the second step, we compute the eigenvalues and the Sobolev embedding constants that we need in our inf-sup lower bound construction.

We hence propose the following algorithm:

1. Manually set K, select a representative set $\mathcal{V}^K=\{\overline{\boldsymbol{\mu}}_1,\dots,\overline{\boldsymbol{\mu}}_K\}$, and compute surrogates for each $\rho_{\overline{\boldsymbol{\mu}}}$;
2. start an optimal search algorithm: For a large random[5] set of $\boldsymbol{\mu}$'s
   (a) solve the reduced basis problem with the existing basis;
   (b) select the "nearest" $\overline{\boldsymbol{\mu}}$ in $\mathcal{V}^K$, and compute $\epsilon_N(\mu)$, $\tau_{N,\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})$, and $\Delta_{N,\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})$ by replacing $\rho_{\overline{\boldsymbol{\mu}}}$ and $\beta_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})$ by their surrogates;
   (c) select the optimal $\boldsymbol{\mu}$ as follows:

[5]This can be replaced by a deterministic search over a large set of sample points.

   i. If for some $\boldsymbol{\mu}$'s, (2a) did not converge, select the "furthest" from the previously selected ones;
   ii. if for some $\boldsymbol{\mu}$'s, $\tau_{N,\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})$ is bigger than 1, select the one with the largest $\tau_{N,\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})$;
   iii. otherwise, select the one with the largest $\Delta_{N,\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})$.
3. Solve the finite element problem (2.1) for the selected $\boldsymbol{\mu}$;
4. increase N and enrich the reduced basis space[6] W_N with $\boldsymbol{u}(\boldsymbol{\mu})$;[7]
5. if the maximum number of bases wished or the tolerance requested for $\Delta_{N,\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})$ are achieved, go to 6; otherwise, go to 2;
6. set $\overline{N}$ to the reached number of basis functions, and compute all of the ingredients for the computation of $\overline{B}^{LB}_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})$;
7. start a random process: For a large set of $\boldsymbol{\mu}$'s check that the reduced basis problem can be solved, that $\overline{B}^{LB}_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})$ is positive for at least one $\overline{\boldsymbol{\mu}}$, and that $\tau_{N,\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu}) \leq 1$ for this $\overline{\boldsymbol{\mu}}$.
   If this is ok, then all of the online components are ready, and we set $N_{\max} = N$; otherwise, either add some $\overline{\boldsymbol{\mu}}$'s to $\mathcal{V}^K$ and go to 6 or go to 2 and replace $\beta_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})$ by $\overline{B}^{LB}_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})$ instead of by 0.2.

Steps 2(a) and 2(b) are online fast. Moreover, in our parallel offline code, each processor performs the optimal search on a different random set. As a result, we can explore our parameter space very rapidly.

To ensure that we are on a selected solution branch, in the finite element resolution we perform a homotopy procedure in $\boldsymbol{\mu}$ starting from an already computed solution at a nearby parameter; in the reduced basis we simply start from a solution at a nearby parameter used to generate the W_N. (In some applications, it may be necessary to perform homotopy also in the reduced basis context.)

Note that we can admit small errors in the resolution of (2.1); indeed, we need approximations to $\boldsymbol{u}(\boldsymbol{\mu})$ at selected parameter points to build our reduced basis. However, the error bound is given with respect to the exact "truth" solution $\boldsymbol{u}(\boldsymbol{\mu})$.

The reader is referred to [33] and [32] for the exact offline procedure to compute the reduced basis and a posteriori error estimation components.

5.2. Online algorithm. The online procedure, for given $\boldsymbol{\mu} \in \mathcal{D}$ and $N \leq N_{\max}$, reads:

I. Solve the reduced basis problem (1.3)[8] and compute $s_N(\boldsymbol{\mu}) = \ell(\boldsymbol{u}(\boldsymbol{\mu}))$;
II. compute $\epsilon_N(\boldsymbol{\mu})$;
III. order $\mathcal{V}^K$ in increasing distance[9] from $\boldsymbol{\mu}$. Compute $\overline{B}^{LB}_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})$ and $\tau_{N,\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})$ until we find $\overline{\boldsymbol{\mu}}$ in $\mathcal{V}$ such that $\overline{B}^{LB}_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu}) > 0$ and $\tau_{N,\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu}) \leq 1$. We set $\mathcal{I}^K(\boldsymbol{\mu}) = \overline{\boldsymbol{\mu}}$;
IV. compute $\Delta_{N,\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})$ and $\Delta^s_{N,\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})$.

If in step III we do not find any $\overline{\boldsymbol{\mu}}$, our method fails: We cannot provide an error bound nor even existence; if possible, N must be increased; otherwise, our basis is not rich enough.

[6] This step involves Gramm–Schmidt orthogonalization and the computation of the matrix- and scalar-products necessary for the reduced basis matrices and the online computation of the Y-dual norm of the residual (cf. [23, 32]).

[7] In our model problems, $\boldsymbol{u}(\boldsymbol{\mu})$ is split into its physical components and a stabilizing supremizer is computed; the reduced space is hence enriched by four functions (cf. section 6.2).

[8] We solve this by a Newton algorithm with the initial guess a solution at a nearby parameter used to generate the reduced basis.

[9] The distance function in $\mathcal{D}$ is arbitrary.

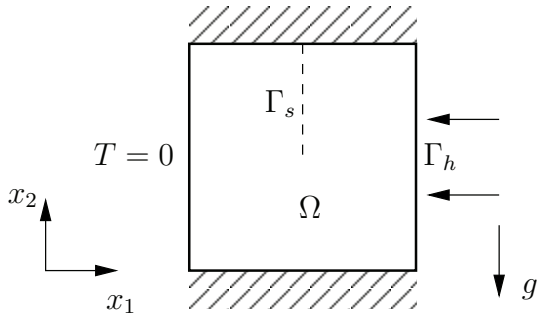


FIG. 1. *Closed cavity. On the left the temperature is constant, on the right the heat flux is constant, and the top and bottom are insulated.*

The computational effort is independent of $\mathcal{N}$: Steps I and II have dominant complexities $N^3 + Q_1 N^2$ times the number of Newton iterations required and $Q_1^2 N^4 + Q_0 Q_1 N^3$, respectively. The complexity of steps III and IV is dominated by that of step II.

6. Application: Steady natural convection. We are interested in solving a natural convection problem in a two-dimensional square cavity under vertical gravity (Figure 1), with constant temperature on one side, heated on the opposite side, insulated at the top and bottom, and Prandtl number $\mathrm{Pr} = 1$ (dissipation to conduction ratio). We consider one parameter $\boldsymbol{\mu} = (\frac{1}{\sqrt{\mathrm{Gr}}})$, where Gr is the Grashof number, which is defined as $(\beta \Delta T g L^3)/\nu^2$ (g is the gravity constant, β is the thermal expansion coefficient, ΔT is the temperature difference, L is the length scale, and ν is the kinematic viscosity).

We impose homogeneous Dirichlet boundary conditions for the velocity (u_1, u_2), while for the temperature θ we impose homogeneous Dirichlet boundary conditions on the left side, homogeneous Neumann boundary conditions on the top and the bottom, and constant unitary Neumann boundary conditions on the heated side Γ_h ($\partial_n \theta = 1$ on Γ_h). We scale the pressure such that $\int_\Omega p = 0$.

Our domain $\Omega \subset \mathbb{R}^2$ is a closed square cavity with unitary sides. We want to solve (with summation over repeated indices and partial derivatives with respect to x_1 and x_2)

$$
\begin{aligned}
u_j \partial_j u_1 &= -\partial_1 p + \frac{1}{\sqrt{\mathrm{Gr}}} \partial^2_{jj} u_1, \\
u_j \partial_j u_2 &= -\partial_2 p + \theta + \frac{1}{\sqrt{\mathrm{Gr}}} \partial^2_{jj} u_2, \\
\partial_i u_i &= 0, \\
u_j \partial_j \theta &= \frac{1}{\sqrt{\mathrm{Gr}}\,\mathrm{Pr}} \partial^2_{jj} \theta
\end{aligned}
$$

in Ω.

Our outputs of interest are the inverse of the Nusselt number and the flux through the middle half section Γ_s parallel to Γ_h:

$$
s_{(1)}(\boldsymbol{\mu}) = \frac{1}{|\Gamma_h|} \int_{\Gamma_h} \theta(\boldsymbol{\mu}), \qquad s_{(2)}(\boldsymbol{\mu}) = \frac{1}{|\Gamma_s|} \int_{\Gamma_s} u_1(\boldsymbol{\mu}).
$$

6.1. Affine decomposition. Our reference finite element space X is a subset of $H_0^1(\Omega)\times H_0^1(\Omega)\times H^1(\Omega)\times L^2(\Omega)\times\mathbb{R}$. We denote elements in X as $\boldsymbol{u}=(u_1,u_2,\theta,p,\lambda)$, $\boldsymbol{v}=(v_1,v_2,\chi,q,\gamma)$, and $\boldsymbol{z}=(z_1,z_2,\zeta,\dots)$.

Recalling (2.2), a decomposes into $Q_0=2$ bilinear and $Q_1=1$ trilinear components,

$$a_0^1(\boldsymbol{u},\boldsymbol{v})=-\int_\Omega p\partial_i v_i-\int_\Omega q\partial_i u_i+\lambda\int_\Omega p+\gamma\int_\Omega q-\int_\Omega \theta v_2, \qquad \Theta_0^1(\boldsymbol{\mu})=1,$$

$$a_0^2(\boldsymbol{u},\boldsymbol{v})=\int_\Omega \partial_i u_j\partial_i v_j+\int_\Omega \partial_i\theta\partial_i\chi, \qquad \Theta_0^2(\boldsymbol{\mu})=\mu_1,$$

$$a_1^1(\boldsymbol{u},\boldsymbol{z},\boldsymbol{v})=\int_\Omega u_j\partial_i z_i v_j+\int_\Omega z_j\partial_i u_i v_j+\int_\Omega z_i\partial_i\theta\,\chi+\int_\Omega u_i\partial_i\zeta\,\chi, \qquad \Theta_1^1(\boldsymbol{\mu})=1.$$

The right-hand side, which linearly depends on μ_1, and the (linear) outputs take the form

$$f^1(\boldsymbol{v})=\int_{\Gamma_h}\chi, \qquad \Theta_f^1(\boldsymbol{\mu})=\mu_1,$$

$$\ell_{(1),0}^1(\boldsymbol{v})=\int_{\Gamma_h}\chi, \qquad \Theta_{\ell,(1)}^1(\boldsymbol{\mu})=1,$$

$$\ell_{(2),0}^1(\boldsymbol{v})=\int_{\Gamma_s}v_1, \qquad \Theta_{\ell,(2)}^1(\boldsymbol{\mu})=1.$$

6.2. Reduced basis and supremizer. To achieve faster convergence (cf. [17, 33]), in step 4 of the algorithm in section 5.1, the solution $\boldsymbol{u}(\boldsymbol{\mu})=(u_1,u_2,\chi,p,\lambda)$ is split into its physical components. To ensure solvability of the reduced system, we preliminarily add the constant pressure $(0,0,0,1,0)$ and Lagrange multiplier $(0,0,0,0,1)$ to our reduced space. To ensure stability, to each new added pressure mode, we add a supremizer (cf. [26, 27]) $\sigma(\boldsymbol{u}(\boldsymbol{\mu}))\in X$, defined by a null temperature, pressure, and Lagrange multiplier, and

$$(\boldsymbol{\sigma}(\boldsymbol{u}(\boldsymbol{\mu})),\boldsymbol{v})_X=\int_\Omega p\partial_i v_i \quad \forall\boldsymbol{v}\in X.$$

(Note that there are alternative definitions of the supremizer; cf. [26, 27].)

Therefore step 4 is replaced by

4bis. Add the three basis functions $(u_1,u_2,0,0,0)$, $(0,0,\chi,0,0)$, and $(0,0,0,p,0)$ as well as the supremizer $\boldsymbol{\sigma}(\boldsymbol{u}(\boldsymbol{\mu}))$ to our reduced basis space.

6.3. Sobolev embedding constant. We define a seminorm $|\cdot|_4$ as the restriction of the L^4-norm to the velocity and temperature fields:

$$|\boldsymbol{v}|_4^4=\int_\Omega\left(v_1^2+v_2^2+\chi^2\right)^2.$$

PROPOSITION 6.1. *For all $\boldsymbol{u}$, $\boldsymbol{v}$, and $\boldsymbol{z}$ in X,*

$$\left|a_1^1(\boldsymbol{u},\boldsymbol{z},\boldsymbol{v})\right|\le 2|\boldsymbol{u}|_4\,\|\boldsymbol{z}\|_X\,|\boldsymbol{v}|_4+|\boldsymbol{u}|_4\,|\boldsymbol{z}|_4\,\|\boldsymbol{v}\|_X.$$

Proof. We denote that $u_3 = \theta$, $v_3 = \chi$, and $z_3 = \zeta$; we then have

$$\int_\Omega \sum_{i=1}^{2}\sum_{j=1}^{3} u_i \partial_i z_j v_j \le \int_\Omega \left(\sum_{j=1}^{3} v_j^2\right)^{\frac12} \left(\sum_{j=1}^{3}\left(\sum_{i=1}^{2} u_i \partial_i z_j\right)^2\right)^{\frac12} \tag{6.1}$$

$$\le \left[\int_\Omega \left(\sum_{j=1}^{3} v_j^2\right)^2\right]^{\frac14} \left[\int_\Omega \left(\sum_{j=1}^{3}\left(\sum_{i=1}^{2} u_i \partial_i z_j\right)^2\right)^{\frac23}\right]^{\frac34}$$

$$\le |\boldsymbol{v}|_4 \left[\int_\Omega \left(\sum_{i=1}^{2} u_i^2\right)^{\frac23} \left(\sum_{j=1}^{3}\sum_{i=1}^{2} (\partial_i z_j)^2\right)^{\frac23}\right]^{\frac34}$$

$$\le |\boldsymbol{v}|_4 \left[\int_\Omega \left(\sum_{i=1}^{2} u_i^2\right)^2\right]^{\frac14} \left[\int_\Omega \sum_{j=1}^{3}\sum_{i=1}^{2} (\partial_i z_j)^2\right]^{\frac12}$$

$$\le |\boldsymbol{v}|_4 \, \|(u_1,u_2)\|_{L^4(\Omega)^2} \, |(z_1,z_2,z_3)|_{H^1(\Omega)^3}.$$

From

$$\sum_{i=1}^{2}\sum_{j=1}^{3} \partial_i (u_j z_i) = \sum_{i=1}^{2}\sum_{j=1}^{3} u_j \partial_i z_i + \sum_{i=1}^{2}\sum_{j=1}^{3} z_i \partial_i u_j,$$

integration by parts, and $z_1 = z_2 = 0$ on $\partial\Omega$, we have

$$\int_\Omega \sum_{i=1}^{2}\sum_{j=1}^{3} z_i \partial_i u_j v_j = -\int_\Omega \sum_{i=1}^{2}\sum_{j=1}^{3} u_j z_i \partial_i v_j - \int_\Omega \sum_{i=1}^{2}\sum_{j=1}^{3} u_j \partial_i z_i v_j.$$

We bound the first term on the right-hand side by applying (6.1) and the second one by

$$\int_\Omega \sum_{i=1}^{2}\sum_{j=1}^{3} u_j \partial_i z_i v_j \le |\boldsymbol{v}|_4 \left[\int_\Omega \left(\sum_{j=1}^{3} u_j^2 \sum_{i=1}^{2} (\partial_i z_i)^2\right)^{\frac23}\right]^{\frac34}$$

$$\le |\boldsymbol{v}|_4 \left[\int_\Omega \left(\sum_{j=1}^{3} u_j^2\right)^2\right]^{\frac14} \left[\int_\Omega \sum_{i=1}^{2} (\partial_i z_i)^2\right]^{\frac12} = |\boldsymbol{v}|_4 \, |\boldsymbol{u}|_4 \, |(z_1,z_2)|_{H^1_{\mathrm{div}}(\Omega)^2}.$$

$\|(u_1,u_2)\|_{L^4(\Omega)^2} \le |\boldsymbol{u}|_4$, $|(z_1,z_2,z_3)|_{H^1(\Omega)^3} < \|\boldsymbol{z}\|_X$, $|(z_1,z_2)|_{H^1_{\mathrm{div}}(\Omega)^2} < \|\boldsymbol{z}\|_X$, and the definition of a_1^1 complete the proof. ∎

From the Sobolev embedding theorem we know that there exists a constant ρ_X such that

$$|\boldsymbol{v}|_4 \le \rho_X \|\boldsymbol{v}\|_X \;\forall\; \boldsymbol{v} \in X \tag{6.2}$$

and from the equivalence of the norms $\|\cdot\|_X$ and $|||\cdot|||_{\overline{\boldsymbol{\mu}}}$ that there exists a constant $\rho_{\overline{\boldsymbol{\mu}}} \le \frac{\rho_X}{\beta(\overline{\boldsymbol{\mu}})}$ such that

$$|\boldsymbol{v}|_4 \le \rho_{\overline{\boldsymbol{\mu}}} |||\boldsymbol{v}|||_{\overline{\boldsymbol{\mu}}} \;\forall\; \boldsymbol{v} \in X. \tag{6.3}$$

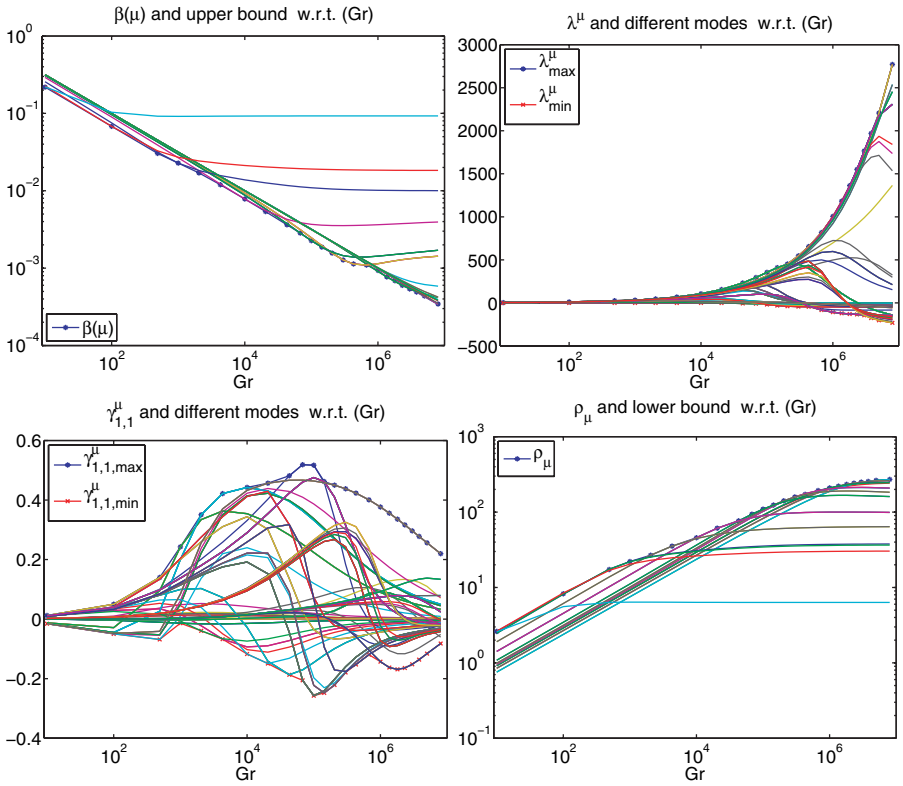


FIG. 2. *Upper left:* $\beta(\overline{\boldsymbol{\mu}})$ *and upper bounds given by the Rayleigh quotients of the minimizing modes. Upper right and lower left:* $\lambda_1^{\overline{\boldsymbol{\mu}}}$ *and* $\gamma_{1,1}^{\overline{\boldsymbol{\mu}}}$*, respectively, and Rayleigh quotients of several maximizing modes (those relative to* $\lambda^{\overline{\boldsymbol{\mu}}}$*,* $\gamma_{0,1}^{\overline{\boldsymbol{\mu}}}$*,* $\gamma_{0,2}^{\overline{\boldsymbol{\mu}}}$*, and* $\gamma_{1,1}^{\overline{\boldsymbol{\mu}}}$*). Lower right:* $\rho_{\overline{\boldsymbol{\mu}}}$ *and lower bounds given by the quotients of our* L^4 *seminorm and natural norm of the maximizing modes. The* x*-axis represents the Grashof number relative to* $\overline{\boldsymbol{\mu}}$*.*

With

$$\rho_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu}) \equiv \frac{\rho_{\overline{\boldsymbol{\mu}}}\rho_X}{\beta(\overline{\boldsymbol{\mu}})} + 2\rho_{\overline{\boldsymbol{\mu}}}^2 \quad \text{and} \quad \rho_{X,\overline{\boldsymbol{\mu}}} \equiv 3\rho_{\overline{\boldsymbol{\mu}}}\rho_X,$$

the assumptions (2.8), (2.9), and (2.10) are satisfied.

6.4. Computation of the eigenvalues. The computation of the inf-sup constants $\beta(\overline{\boldsymbol{\mu}})$ and of the extreme eigenvalues $\lambda_{p,\,\inf|\sup}^{\overline{\boldsymbol{\mu}}}$, $p = 1, \ldots, P$, $\gamma_{i,q,\,\inf|\sup}^{\overline{\boldsymbol{\mu}}}$, $q = 1, \ldots, Q_i$, $i = 0, 1$, heavily depends on the initial guesses. We start by computing $\beta(\overline{\boldsymbol{\mu}})$ for all $\overline{\boldsymbol{\mu}}$ in $\mathcal{V}^K$ from low to high Grashof numbers by inverse iterations. We store all of the modes, and at each $\overline{\boldsymbol{\mu}}$ we select the initial guess as the stored mode with the smallest Rayleigh quotient.

Then we compute the remaining eigenvalues, also by selecting among the stored modes and by storing the new modes found. In this case, we use a locally optimized block preconditioned gradient method [10]. Figure 2 shows $\beta(\overline{\boldsymbol{\mu}})$, $\lambda_{1,\,\inf|\sup}^{\overline{\boldsymbol{\mu}}}$, $\gamma_{1,1,\,\inf|\sup}^{\overline{\boldsymbol{\mu}}}$, and the comparison with the stored modes; some maximizing modes are almost identical for different values of $\overline{\boldsymbol{\mu}}$.

6.5. Computation of the Sobolev embedding constants. We have developed an algorithm [19] for the computation of the Sobolev embedding constants ρ_X and $\rho_{\overline{\boldsymbol{\mu}}}$ in (6.2) and (6.3).

Let $\boldsymbol{u}^* \in X$, $\|\boldsymbol{u}^*\| = 1$, be the (possibly nonunique) supremizer of

$$\rho = \sup_{\boldsymbol{v} \in X} \frac{|\boldsymbol{v}|_4}{\|\boldsymbol{v}\|}, \tag{6.4}$$

where $\|\cdot\|$ denotes the X or the natural norms, and let $z : X \to L^2(\Omega)$,

$$z(\boldsymbol{v}) = \frac{v_1^2 + v_2^2 + \chi^2}{|\boldsymbol{v}|_4^2}.$$

Note that $\|z(\boldsymbol{v})\|_{L^2(\Omega)} = 1$ for all $\boldsymbol{v}$ in X.

For a nonnegative function $z \in L^2(\Omega)$, we introduce the eigenproblem: Find $\boldsymbol{u} = (u_1, u_2, \theta, \dots) \in X$ and $\lambda \in \mathbb{R}$ such that $\|\boldsymbol{u}\| = 1$ and

$$\int_\Omega z\,(u_1 v_1 + u_2 v_2 + \theta\chi) = \lambda\,(\boldsymbol{u}, \boldsymbol{v}) \quad \forall \boldsymbol{v} = (v_1, v_2, \chi, \dots) \in X,$$

where $(\cdot,\cdot)$ is the X or natural scalar products. We denote by $\lambda_{\max}(z)$ and $\boldsymbol{u}_{\max}(z)$ the largest eigenvalue and associated eigenfunction, respectively. Note that we choose the eigenfunctions always in the same half-space.

We observe that $\lambda_{\max}(z(\boldsymbol{u}^*)) = \rho^2$ and $\boldsymbol{u}_{\max}(z(\boldsymbol{u}^*)) = \boldsymbol{u}^*$, which suggests the following fixed point algorithm:

Let $\boldsymbol{u}^0 \in X$ be our initial guess; for $k \geq 0$, define

$$\boldsymbol{u}^{k+1} = \boldsymbol{u}_{\max}\left(z(\boldsymbol{u}^k)\right), \qquad \lambda^{k+1} = \lambda_{\max}\left(z(\boldsymbol{u}^k)\right).$$

A fixed point of this algorithm is not necessarily the supremizer of (6.4), but it is at least a local supremizer. In fact

$$\lambda^{k+1} - \lambda^k = |\boldsymbol{u}^{k-1}|_4^2 \int_\Omega (z(\boldsymbol{u}^{k-1}) - z(\boldsymbol{u}^{k-2}))z(\boldsymbol{u}^{k-1}) + o(\|z(\boldsymbol{u}^{k-1}) - z(\boldsymbol{u}^{k-2})\|^2_{L^2(\Omega)}),$$

and, since $\|z(\boldsymbol{v})\|_{L^2(\Omega)} = 1$,

$$\int_\Omega (z(\boldsymbol{u}^{k-1}) - z(\boldsymbol{u}^{k-2}))z(\boldsymbol{u}^{k-1}) \geq 0;$$

if $z(\boldsymbol{u}^{k-1})$ and $z(\boldsymbol{u}^{k-2})$ are not parallel (i.e, not equal), this integral is strictly positive.

The most critical point of this algorithm is hence the initial guess. To provide a robust initial guess, we first look for the maximizing mode in the low-dimensional subspace generated by the eigenmodes of $T_N^{\overline{\boldsymbol{\mu}}}$ and the already computed supremizers of (6.4). We apply our algorithm in this space, and we select our initial guesses among the basis elements (i.e., among the eigenmodes of $T_N^{\overline{\boldsymbol{\mu}}}$ and the supremizers of (6.4)). This process depends on the order that we choose in $\mathcal{V}^K$; to relax this dependency, when the computations are completed for all $\overline{\boldsymbol{\mu}}$ in $\mathcal{V}^K$, we store the modes and restart the computation of all of the embedding constants. The last picture in Figure 2 shows the embedding constants and the comparison with the quotient (6.4) of the stored modes. In Figure 3 note how, for a high Grashof number, $\rho_{\overline{\boldsymbol{\mu}}}$ deflates significantly from its upper bound $\rho_X/\beta(\boldsymbol{\mu})$.

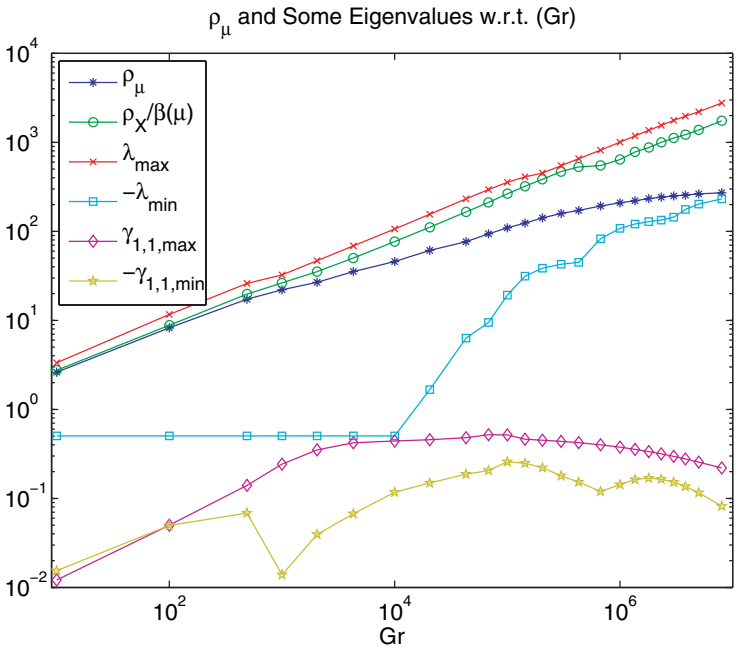


FIG. 3. *Plot of* $\rho_{\overline{\boldsymbol{\mu}}}$, $\rho_X/\beta(\overline{\boldsymbol{\mu}})$, $\lambda^{\overline{\boldsymbol{\mu}}}_{1,\inf|\sup}$, *and* $\gamma^{\overline{\boldsymbol{\mu}}}_{1,1,\inf|\sup}$. *Note that* $\rho_{\overline{\boldsymbol{\mu}}} \leq \rho_X/\beta(\overline{\boldsymbol{\mu}})$ *and that they are similar for a low Grashof number, but that at a higher Grashof number the difference is almost of one order of magnitude. This encourages one to use the mixed formulation of the Brezzi–Rappaz–Raviart theory.*

7. Numerical results.

7.1. Finite element solver. We implemented a parallel finite element solver that uses the Trilinos [36] library, in particular Epetra as MPI interface, Amesos [31, 30] as dense direct solver, Aztec00 as linear iterative solver with domain decomposition preconditioners from IFPACK [29], and Anasazi as eigenvalue solver, in particular LOBPCG [10].

We exploited the PSC resources [37].

We used Taylor–Hood P1/P2 finite elements for a total of 38000 degrees of freedom. We modified a preconditioner proposed by Elman, David, and Wathen [5]: At a point $\boldsymbol{u} \in X$, we define P^0 as $da(\boldsymbol{u}, \boldsymbol{\mu})$ where the divergence operator is replaced by the pressure mass matrix on the pressure block diagonal. We then construct a one-level Schwarz preconditioner P [29, 28] to P^0 and perform a local LU factorization. We solve the Jacobian system with restarted PGMRES(500) with preconditioner P. The resulting operator P, which depends on $\boldsymbol{u}$ and $\boldsymbol{\mu}$, is fast, distributed, and an effective preconditioner for our problem up to $\mathrm{Gr} = 10^7$.

We are interested in one solution branch; therefore, we use homotopy with respect to the Grashof number when solving the finite element problem. In contrast, in our example, the reduced basis problem does not need homotopy: Our initial guess is a known solution for a nearby parameter (a solution that has generated our reduced space).

In Figures 4 and 5 we present the streamlines and the profiles for an increasing Grashof number. We note the development of two vortices and a boundary layer against the wall. We also show the behavior of the Reynolds number (given by

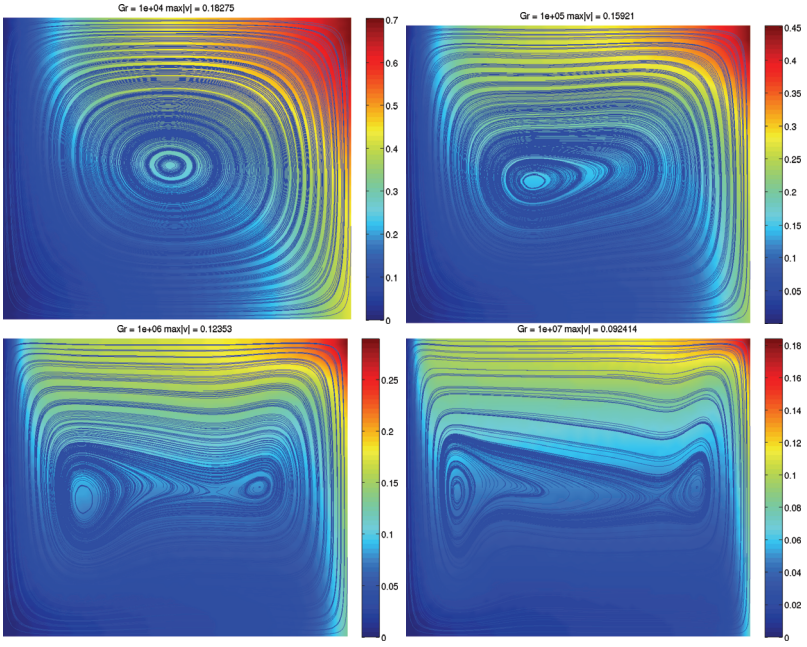


FIG. 4. *Streamlines at Grashof number* 10^4, 10^5, 10^6, *and* 10^7.

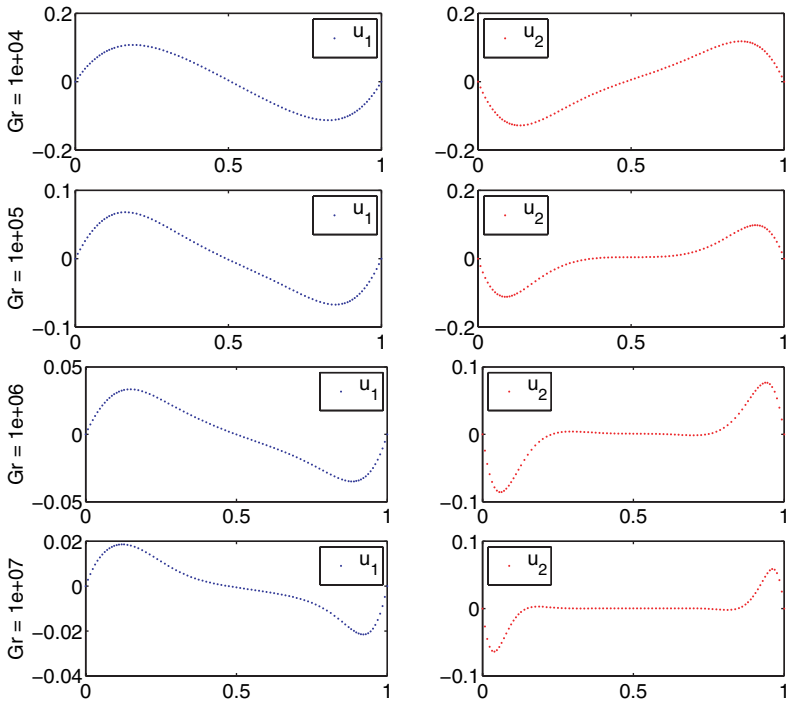


FIG. 5. *Profiles at Grashof number* 10^4, 10^5, 10^6, *and* 10^7. *On the left is the vertical central section, on the right the horizontal one.*

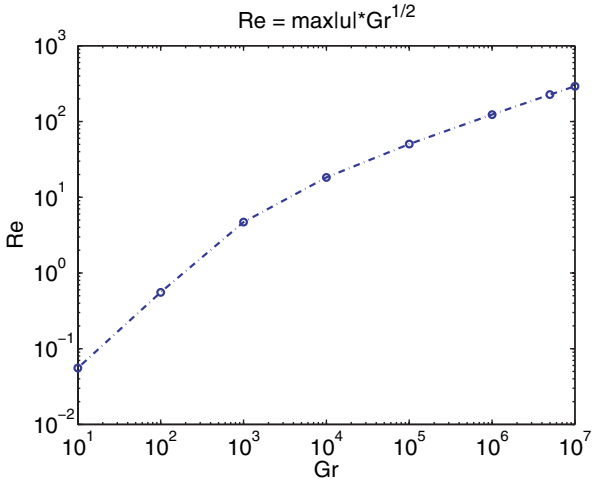


FIG. 6. *Approximative relation between Grashof and Reynolds numbers.*

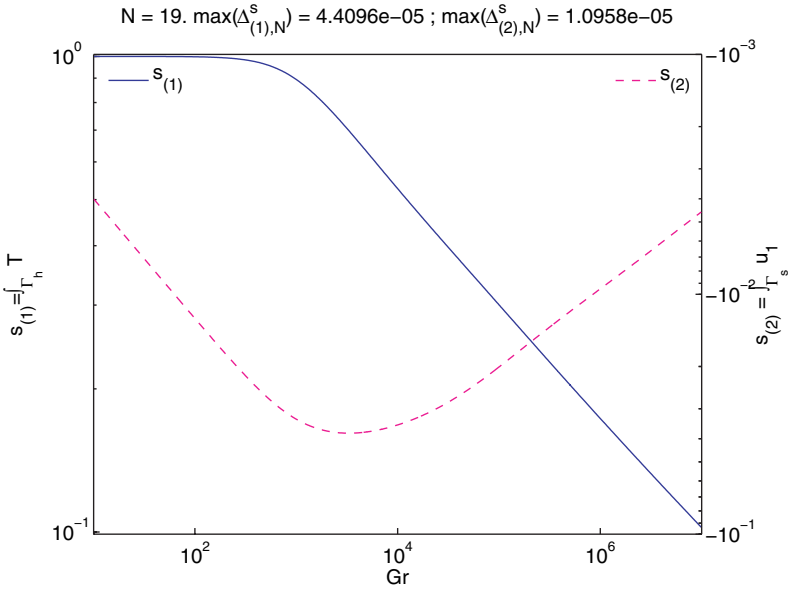


FIG. 7. *Outputs* $s_{(1)}(\boldsymbol{\mu}) = \frac{1}{|\Gamma_h|}\int_{\Gamma_h}\theta(\boldsymbol{\mu})$ *and* $s_{(2)}(\boldsymbol{\mu}) = \frac{1}{|\Gamma_s|}\int_{\Gamma_s}u_1(\boldsymbol{\mu})$ *with* $N = N_{\max} = 2{+}4{\cdot}19$.

$\sqrt{\mathrm{Gr}}\cdot\max|(u_1,u_1)|$) with respect to the Grashof number (Figure 6) and our outputs (Figure 7, computed by the reduced basis method).

The offline computation, i.e., the computation of the reduced basis and the inf-sup lower bound ingredients, required about 36 wall-time hours on a 16 CPU cluster.

7.2. Reduced basis resolution and error bounds. Table 1 compares the CPU time needed by an online reduced basis resolution and a finite element resolution. In the latter, we optimize the initial guess by taking the reduced basis solution such that the X-dual norm of the residual is already below 10^{-6} and two Newton iterations

TABLE 1

Convergence rate and CPU time comparison over 1000 *random samples (different for different* N*).* $\tau_{N,\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu}) = \infty$ *means that for some parameters our inf-sup lower bound is negative. The maximal error bounds are taken over the samples for which we can provide existence and unicity. The RBcpu counts the mean CPU time for solving the reduced basis plus the time to compute the error bound (dual norm of the residual and inf-sup lower bounds); the FEcpu counts the mean CPU time for computing the finite element solution with a dual norm of the residual smaller than* 10^{-10} *and the reduce basis solution as the initial guess.*

N	$\max \tau_N$	$\max \Delta_N$	$\max \Delta^s_{(1),N}$	$\max \Delta^s_{(2),N}$	RBcpu	FEcpu
2+4·10	∞	$4.05\cdot10^{-4}$	$5.60\cdot10^{-3}$	$4.57\cdot10^{-4}$	15+19 ms	
2+4·12	∞	$3.60\cdot10^{-4}$	$4.81\cdot10^{-3}$	$3.93\cdot10^{-4}$	18+35 ms	
2+4·14	2.26	$2.94\cdot10^{-5}$	$6.92\cdot10^{-4}$	$9.24\cdot10^{-5}$	22+66 ms	
2+4·16	0.128	$6.32\cdot10^{-6}$	$3.08\cdot10^{-4}$	$2.98\cdot10^{-5}$	25+108 ms	
2+4·18	0.113	$3.16\cdot10^{-6}$	$1.154\cdot10^{-4}$	$1.49\cdot10^{-5}$	27+157 ms	
2+4·19	0.122	$5.12\cdot10^{-7}$	$3.34\cdot10^{-5}$	$8.20\cdot10^{-6}$	29+188 ms	16×19s

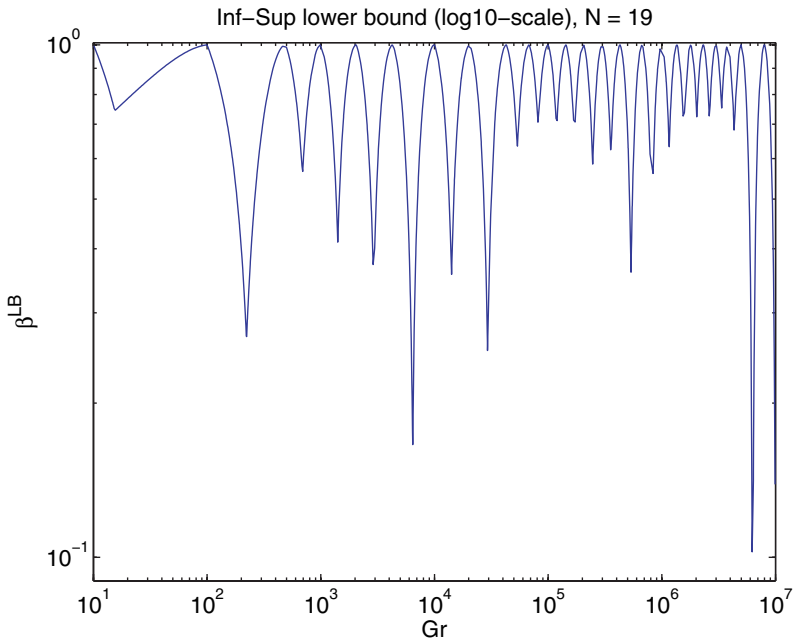


FIG. 8. $\overline{B}^{LB}_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})$*, with* $N = N_{\max} = 2{+}4{\cdot}19$ *and* $\boldsymbol{\kappa}$ *satisfying* (4.9).

are usually enough to reduce it to 10^{-10}. Although this optimization is very generous with the finite element code, the wall time is about two orders of magnitude smaller when using the reduced basis method. Moreover, the online reduced basis problem can be solved on a notebook (one CPU and about 1 GB of memory), while the finite element problem has been solved on a cluster using 16 CPUs.

Once $\tau_{N,\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu}) \leq 1$, our greedy algorithm seeks for a reduction in the error, not in $\tau_{N,\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})$ itself; our random set changes at every new search, which explains why in the last two lines of the table $\tau_{N,\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})$ increases.

Figure 8 shows that our inf-sup lower bound is everywhere between 0.1 and 1. We also recognize the elements of $\mathcal{V}^K = \{\overline{\boldsymbol{\mu}}_1, \ldots, \overline{\boldsymbol{\mu}}_{24}\}$ where $\overline{B}^{LB}_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu}) = 1$.

The error bounds for $N = 2{+}4{\cdot}19$ are small, and, also important, the effectivities are small thanks to our natural norms approach (cf. (3.15) and Figure 9). Note how the error bounds are closely related to the parameters that generate our reduced basis,

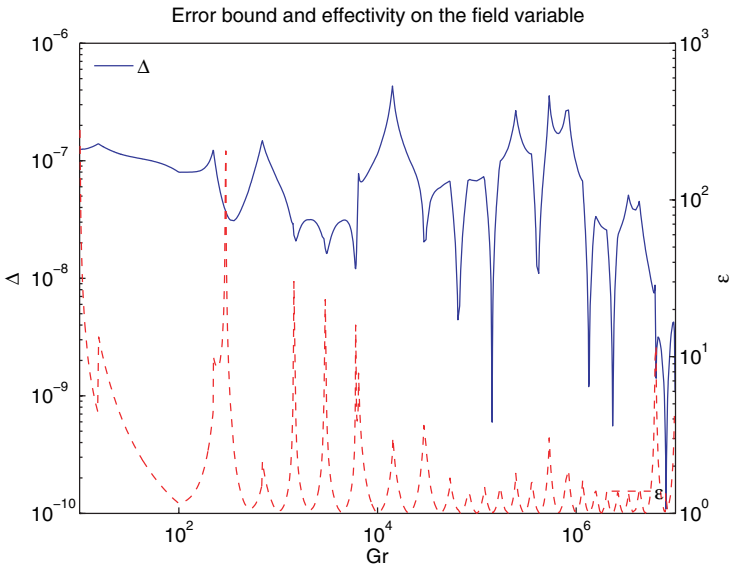


FIG. 9. *Error bound and effectivities* $\Delta_N/\|u-u_N\|_{\overline{\boldsymbol{\mu}}}$ *over* 1000 *random samples,* $N = N_{\max} = 2{+}4{\cdot}19$.

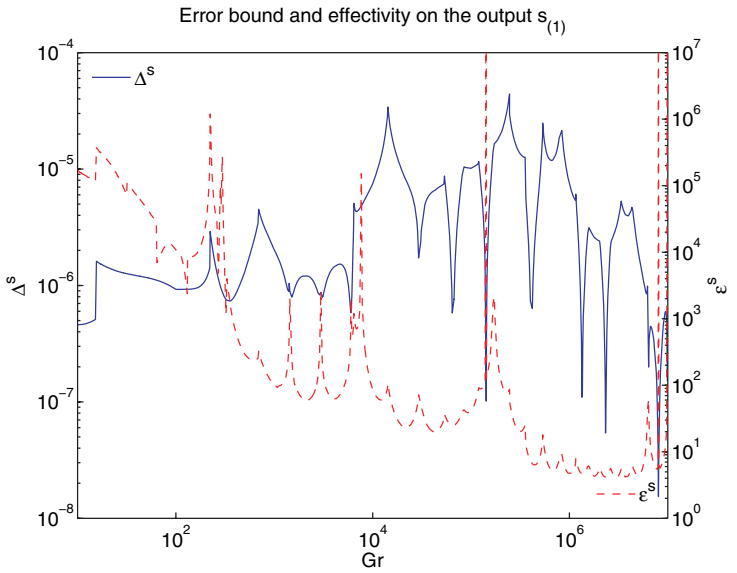


FIG. 10. *Error bound and effectivities* $\Delta_N^s/|s-s_N|$ *over* 1000 *random samples for output* $s_{(1)}$, $N = N_{\max} = 2{+}4{\cdot}19$.

while the effectivities are closely related to the inf-sup lower bound (Figure 8). Since our inf-sup lower bound is better "nearby a $\overline{\boldsymbol{\mu}}$" (cf. (4.1) and (4.2)) the effectivity is better "nearby a $\overline{\boldsymbol{\mu}}$."

In Figure 7 we show our outputs $s_{(1)}$ and $s_{(2)}$ with respect to the Grashof number, computed with our reduced basis on 1000 random points; the error bound is always smaller than $5{\cdot}10^{-5}$. The effectivities for the first output (Figure 10) are less sharp than for the field variable; anyway, when the error bound is small, a "bad" effectivity is acceptable, and when the error bound is the largest, the effectivity is good.

8. Conclusions. We have modified the Brezzi–Rappaz–Raviart theory and defined the natural norm in the quadratic case; the latter enables a feasible and locally sharp inf-sup lower bound construction. These three components provide a fast and feasible error bound computation, which is essential in both the online and the offline stages of the reduced basis method. Our approach has proven good effectivities in the field variable error bounds and empirical good effectivities in the output error bounds.

One essential point is that the error bounds—with respect to the *truth* finite element solution—are exact; hence, the computation of the inf-sup lower bound components have to be exact, too. We have developed an algorithm to remedy a robust Sobolev embedding constant computation.

We are hence able, first, to prepare all of the inf-sup lower bound components offline with a clear methodology and in a reasonable CPU wall time and, second, to solve online quadratic partial differential equations—natural convection with a Grashof number up to 10^7—and provide fast and reliable error bounds, with gains in term of online resources—a notebook against a 16 processor cluster—and in terms of CPU wall time—two orders of magnitude.

Future work will be to show that even in the multiparameter case our approach is feasible; i.e., the offline CPU time is acceptable and the online performance is not affected (preliminary tests with Grashof and Prandl numbers as parameters are promising), including in the presence of geometrical changes (where, e.g., the aspect ratio of the cavity is a further parameter).

We have presumed that the (parameter-dependent) solution lays on an isolated branch. The natural norm approach should also allow us to consider bifurcations since with respect to the "local" norm the inf-sup constant is of order one. Nevertheless, the natural norm degenerates near the bifurcation because of a singular mode. This mode probably has to be treated by deflation, and the eigenvalue solver should take into account the norm degeneration.

Appendix. Dual approximation. In addition to our "truth" primal problem, we may also consider a "truth" dual (or adjoint) problem [23, 14, 3, 21, 33, 32] associated with our particular output functional to improve the convergence of the output.

In this section we consider the case of the quadratic, parameter-independent output (the extension to the linear or the affine-dependent cases is straightforward)

$$\ell(\boldsymbol{v}) = \ell_0(\boldsymbol{v}) - \ell_1(\boldsymbol{v}, \boldsymbol{v}),$$

where ℓ_0 is continuous and linear and ℓ_1 continuous, symmetric, and bilinear.

For a given approximation $\boldsymbol{u}_N(\boldsymbol{\mu})$ in W_N to $\boldsymbol{u}(\boldsymbol{\mu})$, the dual problem reads: Find $\boldsymbol{\psi}^N$ in X such that

$$da(\boldsymbol{u}_N(\boldsymbol{\mu}); \boldsymbol{\mu})(\boldsymbol{v}, \boldsymbol{\psi}^N(\boldsymbol{\mu})) = -\ell_0(\boldsymbol{v}) - 2\ell_1(\boldsymbol{u}_N(\boldsymbol{\mu}), \boldsymbol{v}) \quad \forall \boldsymbol{v} \in X. \tag{A.1}$$

We apply our reduced basis method to solve the dual problem (A.1) and solve for the reduced basis solution $\boldsymbol{\psi}_N^N(\boldsymbol{\mu})$ in W_N^{du} such that

$$da(\boldsymbol{u}_N(\boldsymbol{\mu}); \boldsymbol{\mu})(\boldsymbol{v}, \boldsymbol{\psi}_N^N(\boldsymbol{\mu})) = -\ell_0(\boldsymbol{v}) - 2\ell_1(\boldsymbol{u}_N(\boldsymbol{\mu}), \boldsymbol{v}) \quad \forall \boldsymbol{v} \in W_N^{du},$$

where W_N^{du} (in general different from W_N) is a reduced basis space of dimension N (for simplicity in this presentation, we consider the same dimension for the reduced

primal and dual spaces). We denote the primal and dual residuals as

$$\begin{aligned} r(\boldsymbol{v};\boldsymbol{\mu}) &= a(\boldsymbol{u}(\boldsymbol{\mu}),\boldsymbol{v};\mu) - a(\boldsymbol{u}_N(\boldsymbol{\mu}),\boldsymbol{v};\boldsymbol{\mu}) = f(\boldsymbol{v}) - a(\boldsymbol{u}_N(\boldsymbol{\mu}),\boldsymbol{v};\boldsymbol{\mu}) \quad \forall \boldsymbol{v}\in X,\\ r^{N,du}(\boldsymbol{v};\boldsymbol{\mu}) &= da(\boldsymbol{u}_N(\boldsymbol{\mu});\boldsymbol{\mu})(\boldsymbol{v},\boldsymbol{\psi}^N(\boldsymbol{\mu})) - da(\boldsymbol{u}_N(\boldsymbol{\mu});\boldsymbol{\mu})(\boldsymbol{v},\boldsymbol{\psi}_N^N(\boldsymbol{\mu}))\\ &= -\ell_0(\boldsymbol{v}) - 2\ell_1(\boldsymbol{u}_N(\boldsymbol{\mu}),\boldsymbol{v}) - da(\boldsymbol{u}_N(\boldsymbol{\mu});\boldsymbol{\mu})(\boldsymbol{v},\boldsymbol{\psi}_N^N(\boldsymbol{\mu})) \quad \forall \boldsymbol{v}\in X, \end{aligned}$$

respectively, and the primal error as $\boldsymbol{e}(\boldsymbol{\mu}) = \boldsymbol{u}(\boldsymbol{\mu}) - \boldsymbol{u}_N(\boldsymbol{\mu})$. We define the dual corrected output as

$$s_N^{N,du}(\boldsymbol{\mu}) = \ell(\boldsymbol{u}_N(\boldsymbol{\mu})) - r(\boldsymbol{\psi}_N^N(\boldsymbol{\mu});\boldsymbol{\mu}).$$

The following lemma holds.

LEMMA A.1.

$$\begin{aligned} &s(\boldsymbol{\mu}) - s_N^{N,du}(\boldsymbol{\mu})\\ &= \ell_1(\boldsymbol{e}(\boldsymbol{\mu}),\boldsymbol{e}(\boldsymbol{\mu})) - r^{N,du}(\boldsymbol{e}(\boldsymbol{\mu});\boldsymbol{\mu}) + \frac{1}{2}\Big[da(\boldsymbol{u}(\boldsymbol{\mu});\boldsymbol{\mu}) - da(\boldsymbol{u}_N(\boldsymbol{\mu});\boldsymbol{\mu})\Big](\boldsymbol{e}(\boldsymbol{\mu}),\boldsymbol{\psi}_N^N(\boldsymbol{\mu})). \end{aligned} \tag{A.2}$$

Proof.

$$\begin{aligned} s(\mu) - s_N^{N,du}(\mu) &= \ell_0(\boldsymbol{u}(\boldsymbol{\mu})) - \ell_0(\boldsymbol{u}_N(\boldsymbol{\mu})) + \ell_1\left(\boldsymbol{u}(\boldsymbol{\mu}) - \boldsymbol{u}_N(\boldsymbol{\mu}), \boldsymbol{u}(\boldsymbol{\mu}) - \boldsymbol{u}_N(\boldsymbol{\mu})\right)\\ &\quad + 2\,\ell_1\left(\boldsymbol{u}(\boldsymbol{\mu}) - \boldsymbol{u}_N(\boldsymbol{\mu}), \boldsymbol{u}_N(\boldsymbol{\mu})\right) + r(\boldsymbol{\psi}_N^N(\boldsymbol{\mu});\boldsymbol{\mu})\\ &= \ell_0(\boldsymbol{e}(\boldsymbol{\mu})) + \ell_1\left(\boldsymbol{e}(\boldsymbol{\mu}),\boldsymbol{e}(\boldsymbol{\mu})\right) + 2\,\ell_1\left(\boldsymbol{u}_N(\boldsymbol{\mu}),\boldsymbol{e}(\boldsymbol{\mu})\right)\\ &\quad + a(\boldsymbol{u}(\boldsymbol{\mu}),\boldsymbol{\psi}_N^N(\boldsymbol{\mu});\boldsymbol{\mu}) - a(\boldsymbol{u}_N(\boldsymbol{\mu}),\boldsymbol{\psi}_N^N(\boldsymbol{\mu});\boldsymbol{\mu})\\ &= \ell_1(\boldsymbol{e}(\boldsymbol{\mu}),\boldsymbol{e}(\boldsymbol{\mu})) - r^{N,du}(\boldsymbol{e}(\boldsymbol{\mu});\boldsymbol{\mu}) - da(\boldsymbol{u}_N(\boldsymbol{\mu});\boldsymbol{\mu})\left(\boldsymbol{e}(\boldsymbol{\mu}),\boldsymbol{\psi}_N^N(\boldsymbol{\mu})\right)\\ &\quad + \frac{1}{2}\Big[da(\boldsymbol{u}(\boldsymbol{\mu});\boldsymbol{\mu}) + da(\boldsymbol{u}_N(\boldsymbol{\mu});\boldsymbol{\mu})\Big](\boldsymbol{e}(\boldsymbol{\mu}),\boldsymbol{\psi}_N^N(\boldsymbol{\mu})). \qquad \square \end{aligned}$$

We are now able to compute an upper bound for the dual corrected output. Let

$$\begin{aligned} |||\ell_1|||_{\overline{\boldsymbol{\mu}}\overline{\boldsymbol{\mu}}} &\equiv \sup_{\boldsymbol{v}\in X} \frac{|\ell_1(\boldsymbol{v},\boldsymbol{v})|}{|||\boldsymbol{v}|||_{\overline{\boldsymbol{\mu}}}^2},\\ \|r^{N,du}(\cdot;\boldsymbol{\mu})\|_{\overline{\boldsymbol{\mu}}'} &\equiv \sup_{\boldsymbol{v}\in X} \frac{|r^{N,du}(\boldsymbol{v};\boldsymbol{\mu})|}{|||\boldsymbol{v}|||_{\overline{\boldsymbol{\mu}}}}, \end{aligned} \tag{A.3}$$

and

$$\Delta_{N,\overline{\mu}}^s(\mu) \equiv \|r^{N,du}(\cdot;\boldsymbol{\mu})\|_{\overline{\boldsymbol{\mu}}'}\Delta_{N,\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu}) + \left(\frac{1}{2}\rho_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})\|\boldsymbol{\psi}_N^N(\boldsymbol{\mu})\|_X + |||\ell_1|||_{\overline{\boldsymbol{\mu}}\overline{\boldsymbol{\mu}}}\right)\Delta_{N,\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})^2. \tag{A.4}$$

Then, from (2.8), (3.14), (A.2), and (A.3), the following output error bound holds.

THEOREM A.2.

$$|s(\boldsymbol{\mu}) - s_N^{N,du}(\boldsymbol{\mu})| \le \Delta_{N,\overline{\boldsymbol{\mu}}}^s(\boldsymbol{\mu}).$$

As we expect from the dual approach, the error bound is quadratic with respect to the dual norm of the residuals r and $r^{N,du}$. The main advantage of (A.4) is that there is no dependence on $\beta(\overline{\boldsymbol{\mu}})$, which is the square root of the worse mode of $T_N^{\overline{\boldsymbol{\mu}}}$;

however, this mode may be hidden behind other terms like the natural dual norm of the dual residual or $\rho_{\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})$ (both explicitly) or in $\Delta_{N,\overline{\boldsymbol{\mu}}}(\boldsymbol{\mu})$.

Note that, to simplify the offline work, the term $\|r^{N,du}(\cdot;\boldsymbol{\mu})\|_{\overline{\boldsymbol{\mu}}'}$ may also be replaced by using

$$\|r^{N,du}(\cdot;\boldsymbol{\mu})\|_{\overline{\boldsymbol{\mu}}'} \leq \frac{1}{\beta(\overline{\boldsymbol{\mu}})}\|r^{N,du}(\cdot;\boldsymbol{\mu})\|_{X'}$$

at the detriment of the independency on $\beta(\overline{\boldsymbol{\mu}})$.

Acknowledgments. I thank Professor Anthony T. Patera for his valuable comments and many contributions. I am grateful to Dr. Karen Veroy–Grepl and Dr. Gianluigi Rozza for many discussions on the subject.

REFERENCES

[1] B. O. Almroth, P. Stern, and F. A. Brogan, *Automatic choice of global shape functions in structural analysis*, AIAA J., 16 (1978), pp. 525–528.

[2] E. Balmes, *Parametric families of reduced finite element models: Theory and applications*, Mech. Syst. Signal Processing, 10 (1996), pp. 381–394.

[3] R. Becker and R. Rannacher, *An optimal control approach to a posteriori error estimation in finite element methods*, Acta Numer., 10 (2001), pp. 1–102.

[4] F. Brezzi, J. Rappaz, and P. A. Raviart, *Finite dimensional approximation of nonlinear problems. Part* I: *Branches of nonsingular solutions*, Numer. Math., 36 (1980), pp. 1–25.

[5] H. C. Elman, J. S. David, and A. J. Wathen, *Finite Elements and Fast Iterative Solvers: With Applications in Incompressible Fluid Dynamics*, Numer. Math. Sci. Comput., Oxford University Press, New York, 2005.

[6] J. P. Fink and W. C. Rheinboldt, *On the error behavior of the reduced basis technique for nonlinear finite element approximations*, Z. Angew. Math. Mech., 63 (1983), pp. 21–28.

[7] M. A. Grepl, N. C. Nguyen, K. Veroy, A. T. Patera, and G. R. Liu, *Certified rapid solution of partial differential equations for real-time parameter estimation and optimization*, in Proceedings of the 2nd Sandia Workshop of PDE-Constrained Optimization: Towards Real-Time and On-Line PDE-Constrained Optimization, SIAM Computational Science and Engineering Book Series, SIAM, Philadelphia, 2007.

[8] K. Ito and S. S. Ravindran, *A reduced basis method for control problems governed by PDEs*, in Control and Estimation of Distributed Parameter Systems, W. Desch, F. Kappel, and K. Kunisch, eds., Birkhäuser, Cambridge, MA, 1998, pp. 153–168.

[9] K. Ito and S. S. Ravindran, *A reduced-order method for simulation and control of fluid flows*, J. Comput. Phys., 143 (1998), pp. 403–425.

[10] A. V. Knyazev, *Toward the optimal preconditioned eigensolver: Locally optimal block preconditioned conjugate gradient method*, SIAM J. Sci. Comput., 23 (2001), pp. 517–541.

[11] A. E. Løvgren, Y. Maday, and E. M. Rønquist, *A reduced basis element method for the steady Stokes problem*, M2AN Math. Model. Numer. Anal., 40 (2006), pp. 529–552.

[12] A. E. Løvgren, Y. Maday, and E. M. Rønquist, *The reduced basis element method for fluid flows*, in Analysis and Simulation of Fluid Dynamics, C. Calgaro, J.-F. Coulombel, and T. Goudon, eds., Adv. Math. Fluid Mech. Vol. VII, Birkhäuser, Cambridge, MA, 2007, pp. 129–154.

[13] L. Machiels, Y. Maday, I. B. Oliveira, A. T. Patera, and D. V. Rovas, *Output bounds for reduced-basis approximations of symmetric positive definite eigenvalue problems*, C. R. Acad. Sci. Paris, Série I, 331 (2000), pp. 153–158.

[14] Y. Maday, A. T. Patera, and D. V. Rovas, *A blackbox reduced-basis output bound method for noncoercive linear problems*, in Nonlinear Partial Differential Equations and Their Applications, Collége de France Seminar Volume XIV, D. Cioranescu and J.-L. Lions, eds., Elsevier Science B.V., Amsterdam, 2002, pp. 533–569.

[15] Y. Maday, A. T. Patera, and G. Turinici, *Global a priori convergence theory for reduced-basis approximation of single-parameter symmetric coercive elliptic partial differential equations*, C. R. Acad. Sci. Paris, Série I, 335 (2002), pp. 289–294.

[16] N. C. Nguyen, *Reduced-Basis Approximation and A Posteriori Error Bounds for Nonaffine and Nonlinear Partial Differential Equations: Application to Inverse Analysis*, Ph.D. thesis, Singapore-MIT Alliance, National University of Singapore, 2005.

[17] N. C. Nguyen, K. Veroy, and A. T. Patera, *Certified real-time solution of parametrized partial differential equations*, in Handbook of Materials Modeling, S. Yip, ed., Springer, New York, 2005, pp. 1523–1558.
[18] A. K. Noor and J. M. Peters, *Reduced basis technique for nonlinear analysis of structures*, AIAA J., 18 (1980), pp. 455–462.
[19] A. T. Patera, *private communication.*
[20] J. S. Peterson, *The reduced basis method for incompressible viscous flow calculations*, SIAM J. Sci. Comput., 10 (1989), pp. 777–786.
[21] N. A. Pierce and M. B. Giles, *Adjoint recovery of superconvergent functionals from PDE approximations*, SIAM Rev., 42 (2000), pp. 247–264.
[22] T. A. Porsching, *Estimation of the error in the reduced basis method solution of nonlinear equations*, Math. Comp., 45 (1985), pp. 487–496.
[23] C. Prud'homme, D. Rovas, K. Veroy, Y. Maday, A. T. Patera, and G. Turinici, *Reliable real-time solution of parametrized partial differential equations: Reduced-basis output bound methods*, J. Fluids Engrg., 124 (2002), pp. 70–80.
[24] A. Quarteroni and G. Rozza, *Numerical solution of parametrized Navier-Stokes equations by reduced basis methods*, Numer. Methods Partial Differential Equations, 23 (2007), pp. 923–948.
[25] A. Quarteroni, G. Rozza, and A. Quaini, *Reduced basis methods for optimal control of advection-diffusion problem*, in Adv. Numer. Math., W. Fitzgibbon, R. Hoppe, J. Periaux, O. Pironneau, and Y. Vassilevski, eds., Institute of Numerical Mathematics, Russian Academy of Sciences and Department of Mathematics, University of Houston, 2006, pp. 193–216.
[26] G. Rozza, *Optimization, control and shape design for an arterial bypass*, Internat. J. Numer. Methods Fluids, 47 (2005), pp. 1411–1419.
[27] G. Rozza and K. Veroy, *On the stability of the reduced basis method for Stokes equations in parametrized domains*, Comput. Methods Appl. Mech. Engrg., 196 (2007), pp. 1244–1260.
[28] M. Sala, *Domain Decomposition Preconditioners: Theoretical Properties, Application to the Compressible Euler Equations, Parallel Aspects*, Ph.D. thesis, EPFL, Lausanne, 2003.
[29] M. Sala and M. Heroux, *Robust Algebraic Preconditioners with IFPACK* 3.0, Technical report SAND-0662, Sandia National Laboratories, Albuquerque, NM, 2005.
[30] M. Sala, K. Stanley, and M. Heroux, *Amesos: A set of general interfaces to sparse direct solver libraries*, in Proceedings of PARA'06 Conference, Umea, Sweden, Springer, New York, 2006.
[31] M. Sala, K. Stanley, and M. Heroux, *On the Design of Interfaces to Sparse Direct Solvers*, submitted.
[32] S. Sen, K. Veroy, D.B.P. Huynh, S. Deparis, N. C. Nguyen, and A. T. Patera, *"Natural norm" a posteriori error estimators for reduced basis approximations*, J. Comput. Phys., 217 (2006), pp. 37–62.
[33] K. Veroy and A. T. Patera, *Certified real-time solution of the parametrized steady incompressible Navier-Stokes equations; rigorous reduced-basis a posteriori error bounds*, Internat. J. Numer. Methods Fluids, 47 (2005), pp. 773–788.
[34] K. Veroy, C. Prud'homme, and A. T. Patera, *Reduced-basis approximation of the viscous Burgers equation: Rigorous a posteriori error bounds*, C. R. Acad. Sci. Paris, Série I, 337 (2003), pp. 619–624.
[35] K. Veroy, C. Prud'homme, D. V. Rovas, and A. T. Patera, *A posteriori error bounds for reduced-basis approximation of parametrized noncoercive and nonlinear elliptic partial differential equations*, in Proceedings of the 16th AIAA Computational Fluid Dynamics Conference, 2003.
[36] http://software.sandia.gov/trilinos/.
[37] http://www.psc.edu.

SIAM J. NUMER. ANAL.
Vol. 46, No. 4, pp. 2068–2089
© 2008 Society for Industrial and Applied Mathematics

A POSTERIORI ANALYSIS AND IMPROVED ACCURACY FOR AN OPERATOR DECOMPOSITION SOLUTION OF A CONJUGATE HEAT TRANSFER PROBLEM*

D. ESTEP†, S. TAVENER‡, AND T. WILDEY‡

Abstract. We consider the accuracy of an operator decomposition finite element method for a conjugate heat transfer problem consisting of two materials coupled through a common boundary. We derive accurate a posteriori error estimates that account for the transfer of error between components of the operator decomposition method as well as the differences between the adjoints of the full problem and the discrete iterative system. We use these estimates to guide adaptive mesh refinement. In addition, we address a loss of order of convergence that results from the decomposition and show that the order of convergence is limited by the accuracy of the transferred gradient information. We employ a boundary flux recovery method to regain the expected order of accuracy in an efficient manner.

Key words. a posteriori error analysis, adaptive mesh refinement, adjoint problem, boundary flux method, conjugate heat transfer, domain decomposition, finite element method, generalized Green's function, goal-oriented error estimates, operator decomposition, residual, transfer error

AMS subject classifications. 65N15, 65N30, 65N50

DOI. 10.1137/060678737

1. Introduction.

In this paper, we consider the solution of a conjugate heat transfer problem by an operator decomposition approach. The goal is to compute a functional of the temperature of a body composed of two distinct components that share a common boundary or interface. The two components may have different conductivities and be subject to different heat sources and boundary conditions. Our problem therefore consists of a pair of elliptic boundary value problems that are coupled through boundary conditions posed at the common interface. More generally, our interest is in situations where there is a small number of interfaces between the components.

One approach to solve a conjugate heat transfer problem is to directly discretize the system of coupled elliptic problems and to solve the resulting large global system. However, there are situations in which this approach is infeasible, e.g., because of the size of the full global system or because different numerical methods and codes are to be used for the component problems. Operator decomposition provides an alternative strategy to solving the fully coupled system. We compute the solution in each component individually subject to the conditions on the common interface that

*Received by the editors December 28, 2006; accepted for publication (in revised form) January 14, 2008; published electronically April 23, 2008.
http://www.siam.org/journals/sinum/46-4/67873.html

†Department of Mathematics and Department of Statistics, Colorado State University, Fort Collins, CO 80523 (estep@math.colostate.edu). This author's work was supported in part by the Department of Energy (DE-FG02-04ER25620, DE-FG02-05ER25699, and DE-FC02-07ER54909), the National Aeronautics and Space Administration (NNG04GH63G), the National Science Foundation (DMS-0107832, DMS-0715135, DGE-0221595003, MSPA-CSE-0434354, and ECCS-0700559), Idaho National Laboratory (00069249), and the Sandia Corporation (PO299784).

‡Department of Mathematics, Colorado State University, Fort Collins, CO 80523 (tavener@math.colostate.edu, wildey@math.colostate.edu). The second author's work was supported in part by the Department of Energy (DE-FG02-04ER25620). The third author's work was supported in part by the Department of Energy (DE-FG02-04ER25620), the National Science Foundation (DMS-0107832), and the Sandia Corporation (PO299784).

are obtained from a numerical solution of the other component. Obtaining a solution of the full system by operator decomposition therefore requires a (nominally) infinite iteration in which updated boundary conditions are passed between components.

Operator decomposition is a widely used technique for solving multiphysics, multiscale problems. The general approach is to decompose the problem into components involving simpler physics over a relatively limited range of scales and then to seek the solution of the entire system through an iterative procedure involving solutions of the individual components. This approach is appealing because there is generally a good understanding of how to solve a broad spectrum of single physics problems accurately and efficiently and because it provides an alternative to accommodating multiple scales in one discretization. In some situations, operator decomposition can be viewed as domain decomposition that allows for very different discretizations in each component. However, operator decomposition presents an entirely new set of accuracy and stability issues, some of which are obvious and some subtle and all of which are difficult to correct. In the case of the conjugate heat transfer, the operator decomposition causes a loss in the order of the numerical approximation.

In this paper, we perform an a posteriori error analysis of a finite element implementation of the operator decomposition technique and obtain accurate error estimates that are used to guide an adaptive discretization strategy. Our approach is based on the standard techniques using variational analysis, residuals, and the generalized Green's function solution to an adjoint problem [3, 6, 15, 17, 18, 22], which we modify to account for several new features arising from the operator decomposition. These include the following:

- Numerical errors in the solution of each component are propagated to the other component through the boundary conditions that are applied.
- Numerical errors in the solution of each of the components at one step of the iterative procedure are propagated to the next step.
- The adjoint of the full problem and the adjoint of the iterative procedure are distinct operators whose differences must be recognized when seeking to obtain accurate error estimates.

These kinds of effects are characteristic of operator decomposition discretizations, e.g., [9, 10, 14], and generally require extensions of the usual a posteriori analysis techniques.

In addition to obtaining accurate estimates, we seek to improve the accuracy of the operator decomposition method in an efficient way. In particular, we adapt the "boundary flux recovery" technique developed by Wheeler [26] and Carey [7, 8] to compute normal derivatives on a boundary and show that this can be used to improve accuracy and, in particular, restore the order of convergence of the numerical approximation that is lost due to the operator decomposition.

The operator decomposition method we consider is closely related to the nonoverlapping domain decomposition methods as described in, for example, [24, 29, 25, 30]. The major concern of these authors is, however, the convergence of different iterative methods and the design of effective preconditioners or relaxation parameters, while the accuracy of the final finite element approximation is usually not addressed, despite the fact that some of the numerical results indicate a loss of accuracy. To our knowledge, this paper is the first to address the loss of accuracy. Our results would extend to this particular type of nonoverlapping domain decomposition.

In section 2, we introduce the conjugate heat transfer problem and provide some notation. We describe the iterative operator decomposition finite element method

and some modifications in section 3, as well as the boundary flux recovery method used to compute gradients on the common interface. We perform two a posteriori error analyses in section 4, by using first the adjoint to the fully coupled problem and then the adjoint to the iterative scheme, and present numerical results to highlight the differences between the two. We end the section with a brief discussion of adaptive mesh refinement. In section 5, we carry out an analysis which identifies the transferred gradient information as being responsible for the loss of order of the numerical approximation and show that using the recovered boundary flux restores the order of convergence. Our conclusions are presented in section 6.

2. The model for conjugate heat transfer. Let Ω_1 and Ω_2 be polygonal domains in $\mathbb{R}^2$ or $\mathbb{R}^3$ with boundaries $\partial\Omega_1$ and $\partial\Omega_2$, respectively, intersecting along an interface $\Gamma = \partial\Omega_1 \cap \partial\Omega_2$. We consider a system of second order linear elliptic problems, where the components are coupled through boundary conditions imposed on Γ,

$$\begin{cases} L_1 u_1 = f_1, & \mathbf{x} \in \Omega_1, \\ u_1 = 0, & \mathbf{x} \in \partial\Omega_1 \backslash \Gamma, \\ \begin{cases} u_1 = u_2, \\ A_1 \partial_n u_1 = A_2 \partial_n u_2, \end{cases} & \mathbf{x} \in \Gamma, \\ L_2 u_2 = f_2, & \mathbf{x} \in \Omega_2, \\ u_2 = 0, & \mathbf{x} \in \partial\Omega_2 \backslash \Gamma, \end{cases} \tag{2.1}$$

where, for $i = 1, 2$, $L_i u_i = -\nabla \cdot (A_i \nabla u_i) + c_i u_i$, $A_i \geq A_{i,0} > 0, c_i, f_i$ are sufficiently smooth functions and ∂_n is the partial derivative in the direction of the unit normal vector that is directed outwards from $\partial\Omega_1$. The results of this paper extend easily to general elliptic operators and general Dirichlet, Neumann, and Robin boundary conditions on the boundaries in the complement of the interface.

We let $L^2(\Omega_i)$ denote the space of square integrable functions on Ω_i with inner product $(\cdot,\cdot)_{\Omega_i}$ and norm $\|\cdot\|_{\Omega_i}$, but we use $(\cdot,\cdot) = (\cdot,\cdot)_{\Omega_i}$ when the domain is clear. We use $H^s(\Omega_i)$ to denote the Sobolev space with real index s associated with the norm $\|\cdot\|_{\Omega_i,s}$ and seminorm $|\cdot|_{\Omega_i,s}$. We also use the subspaces $H_0^1(\Omega_i) = \left\{v \in H^1(\Omega_i), v = 0 \text{ on } \partial\Omega_i \backslash \Gamma\right\}$.

The weak formulation of (2.1) seeks $u_i \in H_0^1(\Omega_i)$ such that $u_1 = u_2$ on Γ and

$$\sum_{i=1}^{2} a_i(u_i, v_i) = \sum_{i=1}^{2} (f_i, v_i) \tag{2.2}$$

for all $v_i \in H_0^1(\Omega_i)$, with $a_i(u_i, v) = \int_{\Omega_i} (A_i \nabla u_i \cdot \nabla v + c_i u_i v)\, dx$ for $i = 1, 2$. By assuming that each $a_i(\cdot,\cdot)$ is coercive, (2.2) admits a unique weak solution in $H_0^1(\Omega_i)$ [2, 5, 11].

3. An iterative operator decomposition method. We consider iterative operator decomposition methods to compute the numerical solution of (2.1). Let $u_2^{(0)}$ be an initial guess for the Dirichlet data along the interface. We solve

$$\begin{cases} L_1 u_1^{\{k\}} = f_1, & \mathbf{x} \in \Omega_1, \\ u_1^{\{k\}} = 0, & \mathbf{x} \in \partial\Omega_1 \backslash \Gamma, \\ u_1^{\{k\}} = u_2^{\{k-1\}}, & \mathbf{x} \in \Gamma, \end{cases} \tag{3.1}$$

followed by

$$\begin{cases} L_2 u_2^{\{k\}} = f_2, & \mathbf{x} \in \Omega_2, \\ u_2^{\{k\}} = 0, & \mathbf{x} \in \partial\Omega_2 \backslash \Gamma, \\ A_1 \partial_n u_1^{\{k\}} = A_2 \partial_n u_2^{\{k\}}, & \mathbf{x} \in \Gamma, \end{cases} \tag{3.2}$$

for $k = 1, 2, \ldots, N$, and iterate until (hopefully) a convergence criteria is satisfied. Throughout this paper, we iterate until $\|u_1^{\{k\}} - u_2^{\{k\}}\|_\Gamma$ is less than a prescribed tolerance, i.e., until the continuity condition in the fully coupled problem is satisfied to within a given tolerance.

3.1. Finite element discretization. We let $T_{i,h}$ be a triangulation of Ω_i into elements K where the length of the longest edge is h_K and $h = \max_{K \in T_{i,h}} h_K$. We assume that each triangulation is locally quasi-uniform and $\overline{\Omega}_i = \cup_{K \in T_{i,h}} K$. However, the triangulations on either side of Γ are not assumed to be aligned.

We use the piecewise polynomial spaces

$$S_1 = \left\{ v \text{ continuous on } \Omega_1,\ v \in P^1(K) \text{ for all } K \in T_{1,h} \right\},$$
$$S_2 = \left\{ v \text{ continuous on } \Omega_2,\ v \in P^1(K) \text{ for all } K \in T_{2,h} \right\}$$

and the associated spaces

$$S_{1,0} = \left\{ v \in S_1 \mid v = 0, \mathbf{x} \in \partial\Omega_1 \right\},$$
$$S_{2,0} = \left\{ v \in S_2 \mid v = 0, \mathbf{x} \in \partial\Omega_2 \backslash \Gamma \right\},$$

where $P^1(K)$ denotes the space of linear polynomials on an element K. We let π_i be a projection into S_i as well as the projection into S_i along the interface Γ.

For the finite element approximation, given $U_2^{\{0\}}$ along Γ, compute $U_1^{\{k\}} \in S_1$ satisfying

$$\begin{cases} a_1(U_1^{\{k\}}, v) = (f_1, v)_{\Omega_1} & \text{for all } v \in S_{1,0}, \\ U_1^{\{k\}} = 0, & \mathbf{x} \in \partial\Omega_1 \backslash \Gamma, \\ U_1^{\{k\}} = \pi_1 U_2^{\{k-1\}}, & \mathbf{x} \in \Gamma, \end{cases} \tag{3.3}$$

followed by $U_2^{\{k\}} \in S_2$ such that

$$\begin{cases} a_2(U_2^{\{k\}}, v) = (f_2, v)_{\Omega_2} - (A_1 \partial_n U_1^{\{k\}}, v)_\Gamma & \text{for all } v \in S_{2,0}, \\ U_2^{\{k\}} = 0, & \mathbf{x} \in \partial\Omega_2 \backslash \Gamma, \end{cases} \tag{3.4}$$

for $k = 1, 2, \ldots, N$, and iterate until $\|U_1^{\{k\}} - \pi_1 U_2^{\{k\}}\|_\Gamma$ is smaller than a given tolerance.

3.2. Relaxed iterations. Unfortunately, the simple iterative given by (3.3)–(3.4) may not converge. In particular, the convergence depends on the values of A_1 and A_2 along the interface and the geometry of each region [21, 24, 29]. As a result, we also consider the following relaxation scheme.

We choose $\alpha \in [0, 1)$ and update the Dirichlet interface values with

$$U_1^{\{k\}} = \alpha U_1^{\{k-1\}} + (1 - \alpha) \pi_1 U_2^{\{k-1\}}. \tag{3.5}$$

Optimal values of α can be found in [24, 29] for $A_1 = A_2$, but optimal relaxation parameters are often difficult to find for the general problem. Recognition of this difficulty has prompted the development of, for example, Newton–Krylov methods [31, 16]. We do not pursue this issue further but rather choose to focus on analyzing the loss accuracy in the case that the iterations can be forced to converge.

3.3. Flux correction. In section 5, we show that operator decomposition results in a loss of order of convergence of the finite element solution with mesh size h. This is not completely unexpected due to the lower order of accuracy in the derivative, specifically $\mathbf{O}(h)$ rather than $\mathbf{O}(h^2)$ for the approximation spaces used here. This reduces the order of the overall approximation to first order. To mitigate this effect, we use a postprocessing technique developed by Wheeler [26] and Carey [7, 8] to compute a more accurate boundary flux.

We define the set of elements in $T_{1,h}$ that intersect the boundary by

$$T_{1,h}^{\Gamma} = \{K \in T_{1,h} \mid \overline{K} \cap \partial\Omega \neq \emptyset\}$$

and the corresponding space

$$W_h = \{v \in P^1(K) \text{ with } K \in T_{1,h}^{\Gamma},\ v(\eta_i) = 0 \text{ if } \eta_i \notin \partial\Omega\}, \tag{3.6}$$

where $\{\eta_i\}$ denotes the nodes of element K, so the degrees of freedom correspond to the nodes on the boundary. This space will arise later as the difference between a projection onto $T_{1,h}$ and a second projection onto $T_{1,h}$ that is also required to be zero on the boundary. We seek $\sigma^{\{k\}} \in W_h$ satisfying

$$-(\sigma^{\{k\}}, v)_{\Gamma} = (f_1, v)_{\Omega_1} - a_1(U_1^{\{k\}}, v) \quad \text{for all } v \in W_h, \tag{3.7}$$

where $U_1^{\{k\}}$ solves (3.3). Green's identity implies that $\sigma^{\{k\}}$ gives an approximation to the normal flux on the boundary which is relatively inexpensive to compute.

In general, the accuracy of the recovered boundary flux approximation depends on the regularity of an associated Green's function [20, 27]. In some cases, though perhaps not all, the postprocessed flux is $\mathbf{O}(h^2)$ rather than the standard $\mathbf{O}(h)$ for the normal flux of a piecewise linear finite element approximation. However, we show that the recovered boundary flux leads to a cancellation of the "transfer error" term in the error representation formula, which is the source of the loss of order. This fortunate cancellation of errors means that the accuracy of this recovered boundary flux is only of peripheral interest for our purposes.

We stress that $\sigma^{\{k\}}$ is not assumed to be continuous. In fact, if the domain has a corner on the boundary, the normal derivative is, in general, discontinuous due to the jump in the normal vector. When Dirichlet conditions are given on each boundary segment, we implement the method described in [8] and allow two degrees of freedom to account for the discontinuity. For the application to more general boundary conditions, such as a Neumann or Robin condition on one segment, or an interface condition on both, we refer the reader to [28].

Now, as an alternative to (3.4), we may solve

$$\begin{cases} a_2(U_2^{\{k\}}, v) = (f_2, v)_{\Omega_2} - (\sigma^{\{k\}}, v)_{\Gamma} & \text{for all } v \in S_{2,0}, \\ U_2^{\{k\}} = 0, & \mathbf{x} \in \partial\Omega_2 \backslash \Gamma. \end{cases} \tag{3.8}$$

Other possible approaches to mitigating the loss of order include approximating the boundary flux by using a gradient recovery technique such as the Zienkiewicz–Zhu patch recovery technique [33] or the polynomial-preserving recovery method [32] or by using higher order polynomials near the interface to improve the accuracy of the finite element flux [1, 22]. We had mixed computational success using these other approaches.

4. A posteriori error analysis. To estimate the error of the operator decomposition finite element approximation, we apply a posteriori techniques based on variational analysis and the adjoint problem. In this case, however, the adjoint for the fully coupled original problem differs significantly from the adjoint problem associated with the discretization of the decomposed system. The motivation to use operator decomposition suggests that the adjoint of the full problem is unavailable in practice, and therefore the decomposed adjoint needs to be used to compute error estimates. The difference between the adjoint problems can be reduced by increasing the computational work used to solve the decomposed adjoint; however, in practice, we must consider the issue that the two adjoints lead to significantly different error representations.

In order to understand the differences, we derive a posteriori error estimates by using both adjoints. Both analyses begin in the same way. We wish to estimate the difference between the exact solution to the full problem and the numerical approximation to the solution of the iterative procedure, i.e., $E^{\{k\}} = u - U^{\{k\}}$. We decompose the error:

$$E^{\{k\}} = u - U^{\{k\}} = \left(u - u^{\{k\}}\right) + \left(u^{\{k\}} - U^{\{k\}}\right) = c^{\{k\}} + e^{\{k\}}.$$

The first component $c^{\{k\}}$ measures the difference between the exact solution to the full problem and the exact solution to the iterative problem. If the iterative method converges, then $c^{\{k\}} \to 0$. The second component $e^{\{k\}}$ measures the numerical error in solving the iterative problem. The adjoint of the fully coupled system can be used to estimate the total error $E^{\{k\}}$, while the decomposed adjoint estimates $e^{\{k\}}$. The two estimates are nearly identical if $c^{\{k\}} \approx 0$.

It is helpful to remark that, in the conventional approach, the adjoint problem is defined in such a way as to make the formal bilinear identity hold [23, 19], i.e., to make the boundary terms arising from integration by parts equal to zero. In the analysis below, we find it convenient to consider the values passed between components as quantities of interest defined on the common boundary, and we have to alter the definition of the adjoint problem accordingly.

4.1. The adjoint to the fully coupled problem. The adjoint boundary value problem for the quantity of interest $(\boldsymbol{\psi}, \boldsymbol{u}) = (\psi_1, u_1) + (\psi_2, u_2)$ for the coupled problem (2.1) is

$$\begin{cases} L_1^* \phi_1 = \psi_1, & \mathbf{x} \in \Omega_1, \\ \phi_1 = 0, & \mathbf{x} \in \partial\Omega_1 \backslash \Gamma, \\ \begin{cases} \phi_1 = \phi_2, \\ A_1 \partial_n \phi_1 = A_2 \partial_n \phi_2, \end{cases} & \mathbf{x} \in \Gamma, \\ L_2^* \phi_2 = \psi_2, & \mathbf{x} \in \Omega_2, \\ \phi_2 = 0, & \mathbf{x} \in \partial\Omega_2 \backslash \Gamma, \end{cases} \tag{4.1}$$

where $L_i^* \phi_i = -\nabla \cdot (A_i \nabla \phi_i) + c_i \phi_i$. We solve (4.1) numerically by using an iterative operator decomposition approach as for the forward problem. The iterations are completely independent of the forward iterations.

We can derive an error representation formula for the basic scheme (3.1)–(3.2) and a weighted relaxation technique (3.5) or when using the postprocessed flux as in (3.8). In the discussion below, we use $\theta_h^{\{k\}}$ to denote the numerical flux passed at the kth iteration from Ω_1 to Ω_2.

To begin, we multiply by $(\psi_1, \psi_2)^T$, integrate, and apply the divergence theorem, noting that $u_1 = u_2$ and $A_1\partial_n\phi_1 = A_2\partial_n\phi_2$ along Γ, to obtain

$$(\psi_1, e_1) + (\psi_2, e_2) = a_1(e_1, \phi_1) + a_2(e_2, \phi_2) + (U_1^{\{k\}}, A_1\partial_n\phi_1)_\Gamma - (U_2^{\{k\}}, A_2\partial_n\phi_2)_\Gamma.$$

Observe that the test space $S_{1,0}$ consists of functions that are zero along the interface, while, in general, ϕ_1 is not zero along Γ. This means that the projection of ϕ_1 into $S_{1,0}$ cannot be the interpolant. We define a new projection $\pi_1^0 : H^2 \to S_{1,0}$ such that for any node x_i

$$\pi_1^0\phi(x_i) = \begin{cases} \pi_1\phi(x_i), & x_i \notin \Gamma, \\ 0, & x_i \in \Gamma. \end{cases} \tag{4.2}$$

We also observe that

$$a_2(e_2, \pi_2\phi_2) = -(A_1\partial_n u_1, \pi_2\phi_2)_\Gamma + (\theta_h^{\{k\}}, \pi_2\phi_2)_\Gamma,$$

and

$$a_1(u_1, \phi_1 - \pi_1^0\phi_1) = (f_1, \phi_1 - \pi_1^0\phi_1) + (A_1\partial_n u_1, \phi_1)_\Gamma,$$
$$a_2(u_2, \phi_2 - \pi_2\phi_2) = (f_2, \phi_2 - \pi_2\phi_2) - (A_1\partial_n u_1, \phi_2 - \pi_2\phi_2)_\Gamma,$$

since the adjoint solutions are not zero along Γ.

By using the projection (4.2) when invoking Galerkin orthogonality and the equalities above, we have

$$\begin{aligned}(\psi_1, e_1) + (\psi_2, e_2) = {} & (f_1, \phi_1 - \pi_1^0\phi_1) - a_1(U_1^{\{k\}}, \phi_1 - \pi_1^0\phi_1) \\ & + (f_2, \phi_2 - \pi_2\phi_2) - a_2(U_2^{\{k\}}, \phi_2 - \pi_2\phi_2) \\ & + (U_1^{\{k\}} - U_2^{\{k\}}, A_1\partial_n\phi_1)_\Gamma + (\theta_h^{\{k\}}, \pi_2\phi_2)_\Gamma.\end{aligned}$$

Next, we define $\pi_\partial\phi_1 = \pi_1\phi_1 - \pi_1^0\phi_1$, which is nonzero only near the interface due to the definition of $\pi_1^0\phi_1$ and is an element of W_h defined in (3.6). Substituting $\pi_1^0\phi_1 = \pi_1\phi_1 - \pi_\partial\phi_1$ gives

$$\begin{aligned}(\psi_1, e_1) + (\psi_2, e_2) = {} & (f_1, \phi_1 - \pi_1\phi_1) - a_1(U_1^{\{k\}}, \phi_1 - \pi_1\phi_1) \\ & + (f_2, \phi_2 - \pi_2\phi_2) - a_2(U_2^{\{k\}}, \phi_2 - \pi_2\phi_2) \\ & + (f_1, \pi_\partial\phi_1) - a_1(U_1^{\{k\}}, \pi_\partial\phi_1) \\ & + (U_1^{\{k\}} - U_2^{\{k\}}, A_1\partial_n\phi_1)_\Gamma + (\theta_h^{\{k\}}, \pi_2\phi_2)_\Gamma.\end{aligned}$$

Finally (3.7) implies that the recovered boundary flux $\sigma^{\{k\}}$ satisfies

$$-(\sigma^{\{k\}}, \pi_\partial\phi_1)_\Gamma = (f_1, \pi_\partial\phi_1)_{\Omega_1} - a_1(U_1^{\{k\}}, \pi_\partial\phi_1).$$

Since $\pi_\partial\phi_1 = \pi_1\phi_1$ along Γ, we conclude the following result.

THEOREM 4.1. *The errors $e_1 = u_1 - U_1^{\{k\}}$ and $e_2 = u_2 - U_2^{\{k\}}$ satisfy*

$$\begin{aligned}(\psi_1, e_1) + (\psi_2, e_2) = {} & (f_1, \phi_1 - \pi_1\phi_1) - a_1(U_1^{\{k\}}, \phi_1 - \pi_1\phi_1) \\ & + (f_2, \phi_2 - \pi_2\phi_2) - a_2(U_2^{\{k\}}, \phi_2 - \pi_2\phi_2) \\ & + (U_1^{\{k\}} - U_2^{\{k\}}, A_1\partial_n\phi_1)_\Gamma \\ & + (\theta_h^{\{k\}}, \pi_2\phi_2)_\Gamma - (\sigma^{\{k\}}, \pi_1\phi_1)_\Gamma.\end{aligned} \tag{4.3}$$

The error has been decomposed into two typical discretization components, an iterative component, and a component reflecting the error arising from the transfer of derivative information. The iterative component can be decomposed:

$$(U_1^{\{k\}} - U_2^{\{k\}}, A_1\partial_n\phi_1)_\Gamma = (U_1^{\{k\}} - \pi_1 U_2^{\{k\}}, A_1\partial_n\phi_1)_\Gamma + (\pi_1 U_2^{\{k\}} - U_2^{\{k\}}, A_1\partial_n\phi_1)_\Gamma,$$

which represents an iteration error and a projection error. The choice of derivative information transferred from Ω_1 to Ω_2 has a significant impact on the transfer component.

- Suppose that we set $\theta_h^{\{k\}} = A_1\partial_n U_1^{\{k\}}$, and then

$$(\theta_h^{\{k\}}, \pi_2\phi_2)_\Gamma - (\sigma^{\{k\}}, \pi_1\phi_1)_\Gamma \\ = (A_1\partial_n U_1^{\{k\}} - \sigma^{\{k\}}, \pi_2\phi_2)_\Gamma + (\sigma^{\{k\}}, \pi_2\phi_2 - \pi_1\phi_1)_\Gamma,$$

  which represents a transfer error and a projection error.
- Suppose that we set $\theta_h^{\{k\}} = \sigma^{\{k\}}$, and then

$$(\theta_h^{\{k\}}, \pi_2\phi_2)_\Gamma - (\sigma^{\{k\}}, \pi_1\phi_1)_\Gamma = (\sigma^{\{k\}}, \pi_2\phi_2 - \pi_1\phi_1)_\Gamma,$$

  which represents only a projection error with *no transfer error*.

4.2. The adjoint to the operator decomposition iterative scheme. Next, we derive an error representation by using the natural adjoint for the iterative system (3.1)–(3.2), where we set $U_1^{\{k\}} = \pi_1 U_2^{\{k-1\}}$ and $A_2\partial_n U_2^{\{k\}} = A_1\partial_n U_1^{\{k\}}$. This adjoint avoids the formulation of a globally coupled adjoint and reads

$$\begin{cases} L_1^*\phi_1^{\{k\}} = \psi_1^{\{k\}}, & \mathbf{x} \in \Omega_1, \\ \phi_1^{\{k\}} = 0, & \mathbf{x} \in \partial\Omega_1\backslash\Gamma, \\ \phi_1^{\{k\}} = \phi_2^{\{k\}}, & \mathbf{x} \in \Gamma, \end{cases} \tag{4.4}$$

$$\begin{cases} L_2^*\phi_2^{\{k\}} = \psi_2^{\{k\}}, & \mathbf{x} \in \Omega_2, \\ \phi_2^{\{k\}} = 0, & \mathbf{x} \in \partial\Omega_2\backslash\Gamma, \\ A_2\partial_n\phi_2^{\{k\}} = A_1\partial_n\phi_1^{\{k+1\}}, & \mathbf{x} \in \Gamma, \end{cases} \tag{4.5}$$

for $k = N, \ldots, 1$, with $A_1\partial_n\phi_1^{\{N+1\}} = 0$. Note that the adjoint system is defined as "backwards." In the reasonable case in which we seek information from the final iterate, the data for the adjoint problem would be $\psi_1^{\{1\}} = \psi_2^{\{1\}} = \cdots = \psi_1^{\{N-1\}} = \psi_2^{\{N-1\}} = 0$, with $\psi_1^{\{N\}}$ and $\psi_2^{\{N\}}$ chosen appropriately.

We derive the error representation formula for this system by observing that

$$\sum_{k=1}^{N} \left((\psi_1^{\{k\}}, e_1^{\{k\}}) + (\psi_2^{\{k\}}, e_2^{\{k\}})\right) = \sum_{k=1}^{N} \left((L_1^*\phi_1^{\{k\}}, e_1^{\{k\}}) + (L_2^*\phi_2^{\{k\}}, e_2^{\{k\}})\right)$$

and applying the steps used to derive (4.3). This provides the following theorem.

THEOREM 4.2. *The errors* $e_1^{\{1\}}, e_2^{\{1\}}, \ldots, e_1^{\{N\}}, e_2^{\{N\}}$ *satisfy*

$$\begin{aligned}(4.6)\quad &\sum_{k=1}^{N}\left((\psi_1^{\{k\}}, e_1^{\{k\}}) + (\psi_2^{\{k\}}, e_2^{\{k\}})\right)\\ &\quad= \sum_{k=1}^{N}\Bigg((f_1, \phi_1^{\{k\}} - \pi_1\phi_1^{\{k\}}) - a_1(U_1^{\{k\}}, \phi_1^{\{k\}} - \pi_1\phi_1^{\{k\}})\\ &\qquad + (f_2, \phi_2^{\{k\}} - \pi_2\phi_2^{\{k\}}) - a_2(U_2^{\{k\}}, \phi_2^{\{k\}} - \pi_2\phi_2^{\{k\}})\\ &\qquad\quad + (U_1^{\{k\}} - U_2^{\{k-1\}}, A_1\partial_n\phi_1^{\{k\}})_\Gamma\\ &\qquad\qquad + (A_1\partial_n U_1^{\{k\}}, \pi_1\phi_1^{\{k\}})_\Gamma - (\sigma^{\{k\}}, \pi_2\phi_2^{\{k\}})_\Gamma\Bigg).\end{aligned}$$

It is possible to define adjoints for more complicated iterative methods, but it may be difficult. For example, relaxation techniques couple $U_1^{\{k\}}$ to both $U_2^{\{k-1\}}$ and $U_1^{\{k-1\}}$, which affects how the adjoint variables are coupled. In addition, using the postprocessed numerical flux given in section 3.2 requires defining the adjoint of the postprocessing procedure.

4.3. Numerical results. We illustrate the a posteriori estimates in sections 4.1 and 4.2.

Example 4.1. We triangulate the domains $\Omega_1 = [-0.25, 0.25] \times [-0.25, 0.25]$ and $\Omega_2 = ([-1,1]\times[-1,1])\backslash\Omega_1$ independently (see Figure 4.1) and consider the elliptic interface problem

$$(4.7)\qquad \begin{cases} -\nabla\cdot(A_1\nabla u_1) = f_1, & \mathbf{x}\in\Omega_1,\\ \begin{cases} u_1 = u_2,\\ A_1\partial_n u_1 = A_2\partial_n u_2, \end{cases} & \mathbf{x}\in\Gamma,\\ -\nabla\cdot(A_2\nabla u_2) = f_2, & \mathbf{x}\in\Omega_2,\\ u_2 = 0, & \mathbf{x}\in\partial\Omega_2\backslash\Gamma, \end{cases}$$

with very different diffusion coefficients $A_1 = 1$, $A_2 = 10^3$. The data $f_1(x,y)$ and $f_2(x,y)$ are chosen so the true solutions are $u_1 = 10^3\sin(4\pi x)\sin(4\pi y)$ and $u_2 = \sin(4\pi x)\sin(4\pi y)$. We solve this by using (3.3)–(3.4) with $U_1^{\{k\}} = \pi_1 U_2^{\{k-1\}}$ and

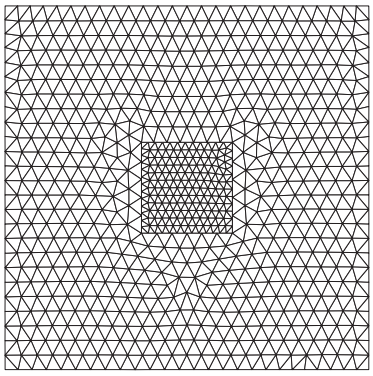


FIG. 4.1. *Triangulations of* Ω_1 *and* Ω_2 *that do not match along the interface.*

TABLE 4.1
Error estimates and effectivity ratios using the adjoint of the fully coupled (F.C.) system.

	ψ_1	ψ_2	True error	F.C. adjoint	Effect. ratio
1.	4	4/15	−0.87344	−0.87309	0.9996
2.	4	0	−0.84291	−0.84258	0.9996
3.	0	4/15	−0.03053	−0.03051	0.9992
4.	$\hat{\delta}_0$	0	−0.39863	−0.39688	0.9956

TABLE 4.2
Error estimates and effectivity ratios using the adjoint of the operator decomposition (O.D.) system.

	ψ_1	ψ_2	True error	O.D. adjoint	Effect. ratio
1.	4	4/15	−0.87344	−0.87309	0.9996
2.	4	0	−0.84291	−0.84259	0.9996
3.	0	4/15	−0.03053	−0.03051	0.9992
4.	$\hat{\delta}_0$	0	−0.39863	−0.38688	0.9956

iterate until $\|U_1 - \pi_1 U_2\|_\Gamma < 1E-6$. We consider several quantities of interest:

1. the global average value,
2. the average value over Ω_1,
3. the average value over Ω_2,
4. the value of u_1 at the point $(0,0)$ computed by using the approximate delta function $\hat{\delta}_0(x,y) = \frac{400}{\pi}\exp(-400x^2 - 400y^2)$.

First, we use an independent iterative scheme to solve the adjoint of the fully coupled problem and iterate until our convergence criteria are met. We display the error estimates for each of the quantities of interest in Table 4.1.

Next, we solve the adjoint to the operator decomposition method and give these results in Table 4.2. Note that the number of iterations as well as the values of the relaxation parameters are determined by the number of iterations used in the solution of the forward operator decomposition method. In this case, four iterations were required with $\alpha = 0$.

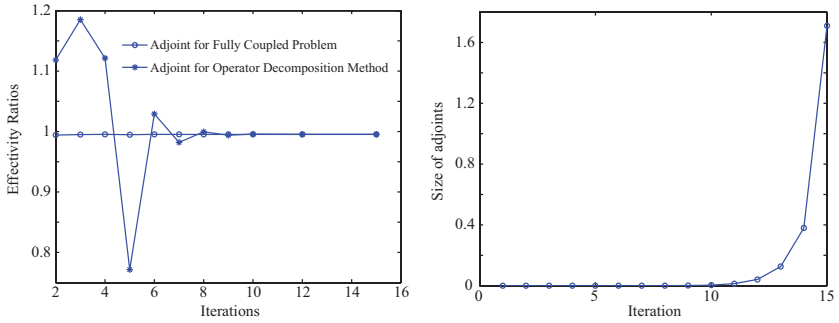
Note that, when the forward iterations have fully converged, the error estimates provided by the fully coupled and operator decomposition adjoints are nearly identical.

Example 4.2. Next we consider neighboring domains $\Omega_1 = [0,1]\times[0,1]$ and $\Omega_2 = ([1,2]\times[0,1])$ with independent triangulations and the elliptic interface problem

$$\begin{cases} -\nabla\cdot(A_1\nabla u_1) = f_1, & \mathbf{x}\in\Omega_1,\\ u_1 = 0, & \mathbf{x}\in\partial\Omega_1\backslash\Gamma,\\ \begin{cases} u_1 = u_2,\\ A_1\partial_n u_1 = A_2\partial_n u_2, \end{cases} & \mathbf{x}\in\Gamma,\\ -\nabla\cdot(A_2\nabla u_2) = f_2, & \mathbf{x}\in\Omega_2,\\ u_2 = 0, & \mathbf{x}\in\partial\Omega_2\backslash\Gamma, \end{cases} \tag{4.8}$$

where $A_1 = 1$, $A_2 = 3$, and $f_1(x,y)$ and $f_2(x,y)$ are chosen so the true solutions are $u_2 = \sin(2\pi x)\sin(2\pi y)$ and $u_1 = 3u_2$. The quantity of interest is the average value over $\Omega_1\cup\Omega_2$, so $\psi_1 = \psi_2 = 1$.

This experiment demonstrates that the error representation formulas (4.3) and (4.6) may give different estimates if the operator decomposition method used to solve the forward problem has failed to converge completely. We fix the number of iterations



(a) Effectivity ratios (b) Decay of influence

FIG. 4.2. *On the left, we show a comparison of the effectivity ratios using the two adjoints for a given number of iterations used for the operator decomposition method for the forward problem. On the right, we plot* $\|\phi_1^{\{k\}}\| + \|\phi_2^{\{k\}}\|$ *to show the decay of influence of errors that occur in previous iterations.*

in the operator decomposition method for (4.8) and use (3.3) and (3.4) to approximate the solution. We solve the fully coupled adjoint iteratively by using a sufficiently large number of iterations for convergence. However, the number of iterations for the adjoint of the iterative system (4.4)–(4.5) is determined by the number of iterations for the forward problem.

In Figure 4.2(a), we plot the effectivity ratios for the two estimates against the number of iterations. We observe that the adjoint for the operator decomposition method does not produce accurate estimates until a sufficient number of iterations for the forward problem have been carried out. This implies that, during the early part of the iterative process, the iterative component of the error $c^{\{k\}} = u - u^{\{k\}}$ dominates the estimate.

If the iterative method converges, we expect that the effect of the initial guess and the first few iterations would diminish. This is reflected in Figure 4.2(b), where we fix the number of iterations $N = 15$ and plot the L^2 norm of the iterative adjoint $\phi_1^{\{k\}}$ and $\phi_2^{\{k\}}$ over Ω_1 and Ω_2, respectively, at each iteration for $1 \le k \le 15$. We observe that the norm decays rapidly as k decreases from N. Recall that the adjoint to the iterative system is solved "backwards" from $N = 15$. The rapid decay indicates that the norm of the fixed-point operator, and hence the norm of the adjoint fixed-point operator, is smaller than one. Thus, the influence of errors in previous iterations is rapidly damped and has little influence on the current error. This implies that we can compute accurate estimates by using the truncated series

$$
\begin{aligned}
(4.9)\quad & (\psi_1^{\{N\}}, e_1^{\{N\}}) + (\psi_2^{\{N\}}, e_2^{\{N\}}) \\
&= \sum_{k=M}^{N} \Big((f_1, \phi_1^{\{k\}} - \pi_1\phi_1^{\{k\}}) - a_1(U_1^{\{k\}}, \phi_1^{\{k\}} - \pi_1\phi_1^{\{k\}}) \\
&\qquad + (f_2, \phi_2^{\{k\}} - \pi_2\phi_2^{\{k\}}) - a_2(U_2^{\{k\}}, \phi_2^{\{k\}} - \pi_2\phi_2^{\{k\}}) \\
&\qquad\quad + (U_1^{\{k\}} - U_2^{\{k-1\}}, A_1\partial_n\phi_1^{\{k\}})_\Gamma \\
&\qquad\qquad + (A_1\partial_n U_1^{\{k\}}, \pi_1\phi_1^{\{k\}})_\Gamma - (\sigma^{\{k\}}, \pi_2\phi_2^{\{k\}})_\Gamma \Big),
\end{aligned}
$$

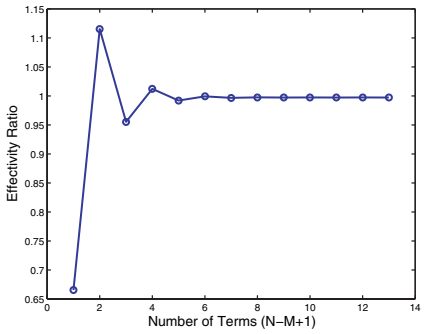


FIG. 4.3. *A comparison of the effectivity ratios computed by using a truncated error representation.*

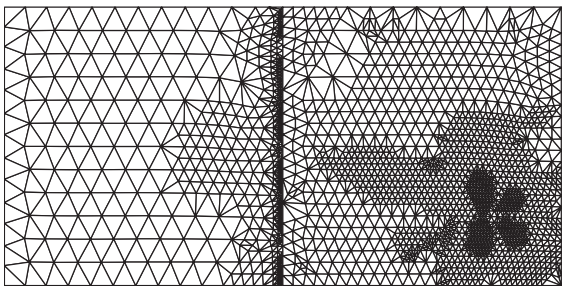


FIG. 4.4. *Adaptive mesh for the quantity of interest equal to the value of* u_2 *at the point* $(1.75, 0.25)$.

where $1 \leq M \leq N$, in place of the full series (4.6). This significantly reduces the computational cost associated with computing a series of adjoint problems. In Figure 4.3 we plot the effectivity ratio for the global average value where we truncate the error representation and compute only the last $N - M + 1$ terms.

4.4. Adaptive refinement. We use the a posteriori error estimate as the basis for adaptivity by employing the standard "optimization framework" after writing the estimate as a sum of element contributions and introducing norms [12, 13, 4, 3]. An element K is marked for refinement when the local error indicator is larger than a local tolerance, usually the global tolerance divided by the current number of elements.

Example 4.3. We demonstrate the adaptive procedure on the model problem (4.8) with the quantity of interest equal to the value of u_2 at the point $(1.75, 0.25)$, which we approximate by choosing $\psi_1 = 0$ and $\psi_2 = 400/\pi \exp(-400(x-1.75)^2 - 400(y-0.25)^2)$. Figure 4.4 shows the results after 3 refinement steps. We solve (3.3)–(3.4) and observe that refinement occurs within Ω_1 along Γ, which reflects the impact of numerical errors in the normal derivative on the quantity of interest.

5. An analysis of the loss of order. In practice, the iterative operator decomposition technique (3.3) and (3.4) is occasionally observed to result in $\mathbf{O}(h)$ convergence rather than the $\mathbf{O}(h^2)$ convergence that is obtained when solving the fully coupled problem. This loss of order results from passing the normal derivative of the finite element approximation, which is only $\mathbf{O}(h)$. Passing the recovered boundary flux restores the full order of convergence. We use the adjoint for the fully coupled problem to derive a posteriori error bounds for the iterative approximations. Numerical examples are provided at the end of the section.

5.1. L^2 error bounds. Let $u_1 \in H^{1+\alpha_1}(\Omega_1)$ and $u_2 \in H^{1+\alpha_2}(\Omega_2)$ be the solutions to (2.1), with $0 \leq \alpha_1, \alpha_2 \leq 1$, and $U_1^{\{k\}}$ and $U_2^{\{k\}}$ be the solutions of the operator decomposition scheme at the kth iteration. Let $\theta_h^{\{k\}}$ denote the flux passed at the kth iteration. Let $\phi_1 \in H^{1+\alpha_1}(\Omega_1)$ and $\phi_2 \in H^{1+\alpha_2}(\Omega_2)$, and pose the adjoint problem (4.1) with $\psi_1 = e_1/\|e_1\|_{\Omega_1}$ and $\psi_2 = e_2/\|e_2\|_{\Omega_2}$. Starting with (4.3), integration by parts over each element K gives

$$\|e_1\|_{\Omega_1} + \|e_2\|_{\Omega_2} = I_1 + I_2 + I_3 + I_4,$$

where

$$\begin{aligned}
I_1 &= \sum_{K \in T_{1,h}} \left(f_1 - L_1 U_1^{\{k\}}, \phi_1 - \pi_1\phi_1\right)_K + \frac{1}{2}\left([A_1\partial_n U_1^{\{k\}}], \phi_1 - \pi_1\phi_1\right)_{\partial K} \\
&\quad + \sum_{K \in T_{2,h}} \left(f_2 - L_2 U_2^{\{k\}}, \phi_2 - \pi_2\phi_2\right)_K + \frac{1}{2}\left([A_2\partial_n U_2^{\{k\}}], \phi_2 - \pi_2\phi_2\right)_{\partial K}, \\
I_2 &= \left(A_2\partial_n U_2^{\{k\}}, \phi_2 - \pi_2\phi_2\right)_\Gamma - \left(A_1\partial_n U_1^{\{k\}}, \phi_1 - \pi_1\phi_1\right)_\Gamma, \\
I_3 &= \left(A_1\partial_n\phi_1, U_1^{\{k\}} - U_2^{\{k\}}\right)_\Gamma, \\
I_4 &= \left(\theta_h^{\{k\}}, \pi_2\phi_2\right)_\Gamma - \left(\sigma^{\{k\}}, \pi_1\phi_1\right)_\Gamma,
\end{aligned}$$

where $[A_i\partial_n U_i^{\{k\}}]$ denotes the jump in the normal derivative across an element edge.

The first term I_1 is a standard a posteriori error bound for elliptic problems and is not affected by nonmatching triangulations along the interface or by transfer error. The second term I_2 is similar to the jump terms along element edges in I_1. The third term I_3 represents the jump in the Dirichlet values across the interface. Finally, the fourth term I_4 represents the difference between the flux passed from Ω_1 to Ω_2 and the flux obtained via the boundary flux recovery technique.

We construct Lemmas 5.1–5.4 below to bound I_1–I_4 individually. In each of these lemmas we first provide the general bound when the triangulations do not match across the boundary and then show the simplification that arises for matching triangulations. We then combine these four lemmas into Theorems 5.5 and 5.7, which give error bounds for the basic iteration (3.3) and (3.4) and when using the recovered boundary flux (3.8), respectively. These two theorems describe the general result when the triangulations do not match across the boundary, while the simplification given matching triangulations is provided as a corollary.

In the following analysis, it is useful to define a smooth function to separate the discretization error arising within an iteration from the transfer and iteration errors. Let $\tilde{u}_1^{\{k\}} \in H^{1+\alpha_1}(\Omega_1)$ solve

$$\begin{cases} a_1(\tilde{u}_1^{\{k\}}, v) = (f_1, v)_{\Omega_1} & \text{for all } v \in H_0^1(\Omega_1), \\ \tilde{u}_1^{\{k\}} = 0, & \mathbf{x} \in \partial\Omega_1 \backslash \Gamma, \\ \tilde{u}_1^{\{k\}} = g_1^{\{k\}}, & \mathbf{x} \in \Gamma, \end{cases} \tag{5.1}$$

where $g_1^{\{k\}}$ is a smooth function, with $\pi_1 g_1^{\{k\}}$ the Dirichlet data provided on Γ at the kth iteration.

Similarly, let $\tilde{u}_2^{\{k\}} \in H^{1+\alpha_2}(\Omega_2)$ solve

$$\begin{cases} a_2(\tilde{u}_2^{\{k\}}, v) = (f_2, v)_{\Omega_2} + (g_2^{\{k\}}, v)_\Gamma & \text{for all } v \in H_0^1(\Omega_2), \\ \tilde{u}_2^{\{k\}} = 0, & \mathbf{x} \in \partial\Omega_2 \backslash \Gamma, \end{cases} \tag{5.2}$$

where $g_2^{\{k\}}$ is a smooth function such that the L^2 projection of $g_2^{\{k\}}$ is the Neumann data on Γ at the kth iteration, namely, $\theta_h^{\{k\}}$.

Standard finite element theory can be applied to bound

$$\|\tilde{u}_1^{\{k\}} - U_1^{\{k\}}\|_{\Omega_1,1} \le Ch_1^{\alpha_1}\left(\|f_1\|_{\Omega_1} + \|g_1^{\{k\}}\|_\Gamma\right),$$
$$\|\tilde{u}_2^{\{k\}} - U_2^{\{k\}}\|_{\Omega_2,1} \le Ch_2^{\alpha_2}\left(\|f_2\|_{\Omega_2} + \|g_2^{\{k\}}\|_\Gamma\right).$$

LEMMA 5.1 (bound on I_1).

$$I_1 \le \sum_{K\in T_{1,h}} \begin{pmatrix} \|f_1 - L_1U_1^{\{k\}}\|_K \\ \frac{1}{2}\|h_K^{-1/2}[A_1\partial_n U_1^{\{k\}}]\|_{\partial K}\end{pmatrix} \cdot \begin{pmatrix} Ch_K^{1+\alpha_1}|\phi_1|_{K,1+\alpha_1} \\ Ch_K^{1+\alpha_1}|\phi_1|_{K,1+\alpha_1}\end{pmatrix}$$
$$+ \sum_{K\in T_{2,h}} \begin{pmatrix} \|f_2 - L_2U_2^{\{k\}}\|_K \\ \frac{1}{2}\|h_K^{-1/2}[A_2\partial_n U_2^{\{k\}}]\|_{\partial K}\end{pmatrix} \cdot \begin{pmatrix} Ch_K^{1+\alpha_2}|\phi_2|_{K,1+\alpha_2} \\ Ch_K^{1+\alpha_2}|\phi_2|_{K,1+\alpha_2}\end{pmatrix}.$$

Proof. Apply standard arguments by using the Cauchy–Schwarz inequality and a trace inequality. ∎

Remark 5.1. If $u_1^{\{k\}} \in H^2(\Omega_1)$, we add and subtract $A_1\partial_n\tilde{u}_1^{\{k\}}$ and use the Cauchy–Schwarz and trace inequalities to bound the jump terms

$$\|[A_1\partial_n U_1^{\{k\}}]\|_{\partial K} \le Ch_K^{1/2}\|\tilde{u}_1^{\{k\}}\|_{2,K}.$$

A similar analysis shows that

$$\|[A_2\partial_n U_2^{\{k\}}]\|_{\partial K} \le Ch_K^{1/2}\|\tilde{u}_2^{\{k\}}\|_{2,K}$$

for $u_2^{\{k\}} \in H^2(\Omega_2)$. Substituting these bounds above shows that I_1 converges $\mathbf{O}(h_1^2 + h_2^2)$ as expected.

LEMMA 5.2 (bound on I_2). *If the triangulations $T_{1,h}$ and $T_{2,h}$ do not match along the interface, then*

$$I_2 \le \left(\|h_1^{-1/2}(\theta_h^{\{k\}} - A_1\partial_n U_1^{\{k\}})\|_\Gamma\right)\cdot\left(Ch_1^{1+\alpha_1}|\phi_1|_{1+\alpha_1}\right)$$
$$+ \left(\|h_2^{-1/2}(A_2\partial_n U_2^{\{k\}} - \theta_h^{\{k\}})\|_\Gamma\right)\cdot\left(Ch_2^{1+\alpha_2}|\phi_2|_{1+\alpha_2}\right)$$
$$+ \left(\|\theta_h^{\{k\}}\|_\Gamma\right)\cdot\left(\|\pi_1\phi_1 - \pi_2\phi_2\|_\Gamma\right),$$

where

$$\|\pi_1\phi_1 - \pi_2\phi_2\|_\Gamma \le \|\pi_1\phi_1 - \phi_1\|_\Gamma + \|\phi_2 - \pi_2\phi_2\|_\Gamma$$
$$\le C_1h_1^{1/2+\alpha_1}|\phi_1|_{\Gamma,1/2+\alpha_1} + C_2h_2^{1/2+\alpha_2}|\phi_2|_{\Gamma,1/2+\alpha_2}.$$

If the triangulations $T_{1,h}$ and $T_{2,h}$ match along the interface, then

$$I_2 \le \left(\|h_1^{-1/2}(\theta_h^{\{k\}} - A_1\partial_n U_1^{\{k\}})\|_\Gamma\right)\cdot\left(Ch_1^{1+\alpha_1}|\phi_1|_{1+\alpha_1}\right)$$
$$+ \left(\|h_2^{-1/2}(A_2\partial_n U_2^{\{k\}} - \theta_h^{\{k\}})\|_\Gamma\right)\cdot\left(Ch_2^{1+\alpha_2}|\phi_2|_{1+\alpha_2}\right).$$

Proof. We add and subtract $(\theta_h^{\{k\}}, \phi_1 - \pi_1\phi_1)_\Gamma$ and $(\theta_h^{\{k\}}, \phi_2 - \pi_2\phi_2)_\Gamma$ and use the fact that $\phi_1 = \phi_2$ on the interface to give

$$I_2 = (\theta_h^{\{k\}} - A_1\partial_n U_1^{\{k\}}, \phi_1 - \pi_1\phi_1)_\Gamma + (A_2\partial_n U_2^{\{k\}} - \theta_h^{\{k\}}, \phi_2 - \pi_2\phi_2)_\Gamma + (\theta_h^{\{k\}}, \pi_1\phi_1 - \pi_2\phi_2)_\Gamma.$$

Observe that $\pi_1\phi_1 = \pi_2\phi_2$ if the triangulations match along the interface. We apply the Cauchy–Schwarz and trace inequalities and an interpolation result to complete the proof. ∎

Remark 5.2. If we assume that $\tilde{u}_1^{\{k\}} \in H^2(\Omega_1)$ and $\tilde{u}_2^{\{k\}} \in H^2(\Omega_2)$, then we can bound

$$\|\theta_h^{\{k\}} - A_1\partial_n U_1^{\{k\}}\|_\Gamma \le C h_1^{1/2} \|\tilde{u}_1^{\{k\}}\|_{2,\Omega_1},$$
$$\|A_2\partial_n U_2^{\{k\}} - \theta_h^{\{k\}}\|_\Gamma \le C h_2^{1/2} \|\tilde{u}_2^{\{k\}}\|_{2,\Omega_2},$$

by adding and subtracting $A_1\partial_n\tilde{u}_1^{\{k\}}$ and $g_2^{\{k\}}$ to the first and second equations, respectively, and applying standard finite element results and the fact that the recovered boundary flux is at least as accurate as the finite element flux [26, 7, 8, 27]. Note that the first equation is zero if $\theta_h^{\{k\}} = A_1\partial_n U_1^{\{k\}}$. Substituting these bounds above shows that I_2 converges $\mathbf{O}(h_1^2 + h_2^2)$.

LEMMA 5.3 (bound on I_3). *If the triangulations $T_{1,h}$ and $T_{2,h}$ do not match along the interface, then*

$$I_3 \le (\|A_1\partial_n\phi_1\|_\Gamma) \cdot \left(\|U_1^{\{k\}} - \pi_1 U_2^{\{k\}}\|_\Gamma\right) + (\|A_1\partial_n\phi_1\|_\Gamma) \cdot \left(\|\pi_1 U_2^{\{k\}} - U_2^{\{k\}}\|_\Gamma\right),$$

where

$$\|\pi_1 U_2^{\{k\}} - U_2^{\{k\}}\|_\Gamma \le C_1 h_1^{1/2+\alpha_1} |\hat{u}^{\{k\}}|_{\Gamma,1/2+\alpha_1} + C_2 h_2^{1/2+\alpha_2} |\hat{u}^{\{k\}}|_{\Gamma,1/2+\alpha_2},$$

with $\hat{u}^{\{k\}}$ a smooth function defined on the interface such that $\pi_1\hat{u}^{\{k\}} = \pi_1 U_2^{\{k\}}$ and $\pi_2\hat{u}^{\{k\}} = U_2^{\{k\}}$. If the triangulations $T_{1,h}$ and $T_{2,h}$ match along the interface, then

$$I_3 \le (\|A_1\partial_n\phi_1\|_\Gamma) \cdot \left(\|U_1^{\{k\}} - \pi_1 U_2^{\{k\}}\|_\Gamma\right).$$

Proof. Rewrite I_3 by adding and subtracting $(A_1\partial_n\phi_1, \pi_1 U_2^{\{k\}})_\Gamma$ to get

$$I_3 = (A_1\partial_n\phi_1, \pi_1 U_1^{\{k\}} - \pi_1 U_2^{\{k\}})_\Gamma + (A_1\partial_n\phi_1, \pi_1 U_2^{\{k\}} - U_2^{\{k\}})_\Gamma.$$

To bound $\|\pi_1 U_2^{\{k\}} - U_2^{\{k\}}\|_\Gamma$ we add and subtract $\hat{u}^{\{k\}}$ and apply Cauchy–Schwarz and trace inequalities giving

$$\|\pi_1 U_2^{\{k\}} - U_2^{\{k\}}\|_\Gamma \le C_1 h_1^{1/2+\alpha_1} |\hat{u}^{\{k\}}|_{\Gamma,1/2+\alpha_1} + C_2 h_2^{1/2+\alpha_2} |\hat{u}^{\{k\}}|_{\Gamma,1/2+\alpha_2}.$$

Observe that $\pi_1 U_2^{\{k\}} = U_2^{\{k\}}$ if the triangulations match along the interface. We apply the Cauchy–Schwarz inequality to complete the proof. ∎

LEMMA 5.4 (bound on I_4).

(i) *If the triangulations $T_{1,h}$ and $T_{2,h}$ do not match along the interface and $\theta_h^{\{k\}} = A_1\partial_n U_1^{\{k\}}$, then*

$$I_4 \le \left(\|A_1\partial_n U_1^{\{k\}} - A_1\partial_n \tilde{u}_1^{\{k\}}\|_\Gamma + \|A_1\partial_n \tilde{u}_1^{\{k\}} - \sigma^{\{k\}}\|_\Gamma\right)\cdot\left(\|\pi_2\phi_2\|_\Gamma\right) + \left(\|\sigma^{\{k\}}\|_\Gamma\right)\cdot\left(\|\pi_2\phi_2 - \pi_1\phi_1\|_\Gamma\right),$$

where

$$\|\pi_2\phi_2 - \pi_1\phi_1\|_\Gamma \le C_1 h_1^{1/2+\alpha_1}|\phi_1|_{\Gamma,1/2+\alpha_1} + C_2 h_2^{1/2+\alpha_2}|\phi_2|_{\Gamma,1/2+\alpha_2}.$$

(ii) *If the triangulations $T_{1,h}$ and $T_{2,h}$ match along the interface and $\theta_h^{\{k\}} = A_1\partial_n U_1^{\{k\}}$, then*

$$I_4 \le \left(\|A_1\partial_n U_1^{\{k\}} - A_1\partial_n \tilde{u}_1^{\{k\}}\|_\Gamma + \|A_1\partial_n \tilde{u}_1^{\{k\}} - \sigma^{\{k\}}\|_\Gamma\right)\cdot\left(\|\pi_2\phi_2\|_\Gamma\right).$$

(iii) *If the triangulations $T_{1,h}$ and $T_{2,h}$ do not match along the interface and $\theta_h^{\{k\}} = \sigma^{\{k\}}$, then*

$$I_4 \le \left(\|\sigma^{\{k\}}\|_\Gamma\right)\cdot\left(\|\pi_2\phi_2 - \pi_1\phi_1\|_\Gamma\right),$$

where

$$\|\pi_2\phi_2 - \pi_1\phi_1\|_\Gamma \le C_1 h_1^{1/2+\alpha_1}|\phi_1|_{\Gamma,1/2+\alpha_1} + C_2 h_2^{1/2+\alpha_2}|\phi_2|_{\Gamma,1/2+\alpha_2}.$$

(iv) *If the triangulations $T_{1,h}$ and $T_{2,h}$ match along the interface and $\theta_h^{\{k\}} = \sigma^{\{k\}}$, then*

$$I_4 = 0.$$

Proof. We add and subtract $(\sigma^{\{k\}}, \pi_2\phi_2)_\Gamma$ and use $\phi_1 = \phi_2$ along Γ to get

$$I_4 = (\theta_h^{\{k\}} - \sigma^{\{k\}}, \pi_2\phi_2)_\Gamma + (\sigma^{\{k\}}, \pi_2\phi_2 - \pi_1\phi_1)_\Gamma.$$

Clearly, if $\theta_h^{\{k\}} = \sigma^{\{k\}}$, then the first term drops out. If the triangulations match along the interface, then $\pi_1\phi_1 = \pi_2\phi_2$, and the second term drops out. We add and subtract $(A_1\partial_n\tilde{u}_1^{\{k\}}, \pi_2\phi_2)_\Gamma$ to the first term and apply the Cauchy–Schwarz inequality to complete the proof. ∎

Remark 5.3. In practice, the error in the normal derivative is typically the same accuracy as the H^1 error, namely, $\mathbf{O}(h_1^{\alpha_1})$. However, an application of the trace theorem proves only $\mathbf{O}(h_1^{\alpha_1/2})$ accuracy. This is not an important issue, however, since we intend to use the fact that this term is less accurate than the others and therefore pollutes the L^2 error. We assume the error in the normal derivative can be bounded

$$\|A_1\partial_n U_1^{\{k\}} - \tilde{u}_1^{\{k\}}\|_\Gamma \le C h_1^{\beta_1}\|f_1\|_{\Omega_1}$$

for $\alpha_1/2 \le \beta_1 \le \alpha_1$. The error in the recovered boundary flux can be bounded

$$\|A_1\partial_n \tilde{u}_1^{\{k\}} - \sigma^{\{k\}}\|_\Gamma \le C S_1 h_1^{\beta_2}\|f_1\|_{\Omega_1}$$

for $\beta_1 \leq \beta_2 \leq \alpha_1 + 1$, where S_1 is a stability factor defined by an associated Green's function [20, 27]. In certain situations, the recovered boundary flux is the same order as the finite element approximation [26, 7, 8], but this result is not known in general. We emphasize that the accuracy of the recovered boundary flux is immaterial due to the cancellation in the error representation.

The bounds in Lemmas 5.1–5.4 are combined in the following theorems and corollaries. First we consider the case $\theta_h^{\{k\}} = A_1 \partial_n U_1^{\{k\}}$.

THEOREM 5.5. *Assume that the triangulations $T_{1,h}$ and $T_{2,h}$* do not *match along the interface Γ. If $U_1^{\{k\}}$ and $U_2^{\{k\}}$ solve* (3.3) *and* (3.4)*, respectively, then the errors $e_1 = u_1 - U_1^{\{k\}}$ and $e_2 = u_2 - U_2^{\{k\}}$ satisfy*

$$\begin{aligned}
\|e_1\|_{\Omega_1} + \|e_2\|_{\Omega_2} \leq & \sum_{K \in T_{1,h}} \begin{pmatrix} \|f_1 - L_1 U_1^{\{k\}}\|_K \\ \frac{1}{2}\|h_K^{-1/2}[A_1 \partial_n U_1^{\{k\}}]\|_{\partial K} \end{pmatrix} \cdot \begin{pmatrix} C h_K^{1+\alpha_1} |\phi_1|_{K,1+\alpha_1} \\ C h_K^{1+\alpha_1} |\phi_1|_{K,1+\alpha_1} \end{pmatrix} \\
& + \sum_{K \in T_{2,h}} \begin{pmatrix} \|f_2 - L_2 U_2^{\{k\}}\|_K \\ \frac{1}{2}\|h_K^{-1/2}[A_2 \partial_n U_2^{\{k\}}]\|_{\partial K} \end{pmatrix} \cdot \begin{pmatrix} C h_K^{1+\alpha_2} |\phi_2|_{K,1+\alpha_2} \\ C h_K^{1+\alpha_2} |\phi_2|_{K,1+\alpha_2} \end{pmatrix} \\
& + \left(\|h_2^{-1/2}(A_2 \partial_n U_2^{\{k\}} - \theta_h^{\{k\}})\|_\Gamma \right) \cdot \left(C h_2^{1+\alpha_2} |\phi_2|_{1+\alpha_2} \right) \\
& + \left(\|A_1 \partial_n \phi_1\|_\Gamma \right) \cdot \left(\|U_1^{\{k\}} - \pi_1 U_2^{\{k\}}\|_\Gamma \right) \\
& + \left(C h_1^{\beta_1} \|f_1\|_{\Omega_1} + C S_1 h_1^{\beta_2} \|f_1\|_{\Omega_1} \right) \cdot \left(\|\pi_2 \phi_2\|_\Gamma \right) \\
& + C_1 h_1^{1/2+\alpha_1} \left((S_2 + S_3) |\phi_1|_{\Gamma,1/2+\alpha_1} + S_4 |\hat{u}|_{\Gamma,1/2+\alpha_1} \right) \\
& + C_2 h_2^{1/2+\alpha_2} \left((S_2 + S_3) |\phi_2|_{\Gamma,1/2+\alpha_2} + S_4 |\hat{u}|_{\Gamma,1/2+\alpha_2} \right),
\end{aligned}$$

with $\alpha_1/2 \leq \beta_1 \leq \alpha_1$, $\beta_1 \leq \beta_2 \leq \alpha_1 + 1$, S_1 is a stability factor independent of h, $S_2 = \|A_1 \partial_n U_1^{\{k\}}\|_\Gamma$, $S_3 = \|\sigma^{\{k\}}\|_\Gamma$, and $S_4 = \|A_1 \partial_n \phi_1\|_\Gamma$.

COROLLARY 5.6. *Assume that the triangulations $T_{1,h}$ and $T_{2,h}$ match along the interface Γ. If $U_1^{\{k\}}$ and $U_2^{\{k\}}$ solve* (3.3) *and* (3.4)*, respectively, then the errors $e_1 = u_1 - U_1^{\{k\}}$ and $e_2 = u_2 - U_2^{\{k\}}$ satisfy*

$$\begin{aligned}
\|e_1\|_{\Omega_1} + \|e_2\|_{\Omega_2} \leq & \sum_{K \in T_{1,h}} \begin{pmatrix} \|f_1 - L_1 U_1^{\{k\}}\|_K \\ \frac{1}{2}\|h_K^{-1/2}[A_1 \partial_n U_1^{\{k\}}]\|_{\partial K} \end{pmatrix} \cdot \begin{pmatrix} C h_K^{1+\alpha_1} |\phi_1|_{K,1+\alpha_1} \\ C h_K^{1+\alpha_1} |\phi_1|_{K,1+\alpha_1} \end{pmatrix} \\
& + \sum_{K \in T_{2,h}} \begin{pmatrix} \|f_2 - L_2 U_2^{\{k\}}\|_K \\ \frac{1}{2}\|h_K^{-1/2}[A_2 \partial_n U_2^{\{k\}}]\|_{\partial K} \end{pmatrix} \cdot \begin{pmatrix} C h_K^{1+\alpha_2} |\phi_2|_{K,1+\alpha_2} \\ C h_K^{1+\alpha_2} |\phi_2|_{K,1+\alpha_2} \end{pmatrix} \\
& + \left(\|h_2^{-1/2}(A_2 \partial_n U_2^{\{k\}} - \theta_h^{\{k\}})\|_\Gamma \right) \cdot \left(C h_2^{1+\alpha_2} |\phi_2|_{1+\alpha_2} \right) \\
& + \left(\|A_1 \partial_n \phi_1\|_\Gamma \right) \cdot \left(\|U_1^{\{k\}} - \pi_1 U_2^{\{k\}}\|_\Gamma \right) \\
& + \left(C h_1^{\beta_1} \|f_1\|_{\Omega_1} + C S_1 h_1^{\beta_2} \|f_1\|_{\Omega_1} \right) \cdot \left(\|\pi_2 \phi_2\|_\Gamma \right),
\end{aligned}$$

with $\alpha_1/2 \leq \beta_1 \leq \alpha_1$, $\beta_1 \leq \beta_2 \leq \alpha_1 + 1$, and S_1 is a stability factor independent of h_1.

The theorem is valid independently of the size of $\|U_1^{\{k\}} - \pi_1 U_2^{\{k\}}\|_\Gamma$, but, when the iteration converges such that this term is negligible, it is clear that the term

containing h_1^β decreases at a slower rate than the other terms. In fact, if we assume that $u_1 \in H^2(\Omega_1)$ and $u_2 \in H^2(\Omega_2)$, then the other terms are $\mathbf{O}(h_1^2)$ or $\mathbf{O}(h_2^2)$, while the finite element flux is $\mathbf{O}(h_1)$ at best. Suppose that instead we set $\theta_h^{\{k\}} = \sigma^{\{k\}}$ instead of the finite element flux. As shown in Lemma 5.4, this reduces I_4 to

$$I_4 = (\sigma^{\{k\}}, \pi_2\phi_2 - \pi_1\phi_1)_\Gamma,$$

resulting in the following theorem.

THEOREM 5.7. *Assume that the triangulations $T_{1,h}$ and $T_{2,h}$* do not *match along the interface Γ. If $U_1^{\{k\}}$ and $U_2^{\{k\}}$ solve* (3.3) *and* (3.8)*, respectively, then the errors $e_1 = u_1 - U_1^{\{k\}}$ and $e_2 = u_2 - U_2^{\{k\}}$ satisfy*

$$\begin{aligned}
\|e_1\|_{\Omega_1} + \|e_2\|_{\Omega_2} \le & \sum_{K\in T_{1,h}} \begin{pmatrix} \|f_1 - L_1U_1^{\{k\}}\|_K \\ \frac{1}{2}\|h_K^{-1/2}[A_1\partial_n U_1^{\{k\}}]\|_{\partial K}\end{pmatrix} \cdot \begin{pmatrix} Ch_K^{1+\alpha_1}|\phi_1|_{K,1+\alpha_1} \\ Ch_K^{1+\alpha_1}|\phi_1|_{K,1+\alpha_1}\end{pmatrix} \\
& + \sum_{K\in T_{2,h}} \begin{pmatrix} \|f_2 - L_2U_2^{\{k\}}\|_K \\ \frac{1}{2}\|h_K^{-1/2}[A_2\partial_n U_2^{\{k\}}]\|_{\partial K}\end{pmatrix} \cdot \begin{pmatrix} Ch_K^{1+\alpha_2}|\phi_2|_{K,1+\alpha_2} \\ Ch_K^{1+\alpha_2}|\phi_2|_{K,1+\alpha_2}\end{pmatrix} \\
& + \left(\|h_1^{-1/2}(\sigma^{\{k\}} - A_1\partial_n U_1^{\{k\}})\|_\Gamma\right) \cdot \left(Ch_1^{1+\alpha_1}|\phi_1|_{1+\alpha_1}\right) \\
& + \left(\|h_2^{-1/2}(A_2\partial_n U_2^{\{k\}} - \sigma^{\{k\}})\|_\Gamma\right) \cdot \left(Ch_2^{1+\alpha_2}|\phi_2|_{1+\alpha_2}\right) \\
& + \left(\|A_1\partial_n\phi_1\|_\Gamma\right) \cdot \left(\|U_1^{\{k\}} - \pi_1 U_2^{\{k\}}\|_\Gamma\right) \\
& + C_1h_1^{1/2+\alpha_1}\left((S_2+S_3)|\phi_1|_{\Gamma,1/2+\alpha_1} + S_4|\hat{u}|_{\Gamma,1/2+\alpha_1}\right) \\
& + C_1h_2^{1/2+\alpha_2}\left((S_2+S_3)|\phi_2|_{\Gamma,1/2+\alpha_2} + S_4|\hat{u}|_{\Gamma,1/2+\alpha_2}\right),
\end{aligned}$$

where $S_2 = \|\sigma^{\{k\}}\|_\Gamma$, $S_3 = \|\sigma^{\{k\}}\|_\Gamma$, and $S_4 = \|A_1\partial_n\phi_1\|_\Gamma$.

COROLLARY 5.8. *Assume that the triangulations $T_{1,h}$ and $T_{2,h}$ match along the interface Γ. If $U_1^{\{k\}}$ and $U_2^{\{k\}}$ solve* (3.3) *and* (3.8)*, respectively, then the errors $e_1 = u_1 - U_1^{\{k\}}$ and $e_2 = u_2 - U_2^{\{k\}}$ satisfy*

$$\begin{aligned}
\|e_1\|_{\Omega_1} + \|e_2\|_{\Omega_2} \le & \sum_{K\in T_{1,h}} \begin{pmatrix} \|f_1 - L_1U_1^{\{k\}}\|_K \\ \frac{1}{2}\|h_K^{-1/2}[A_1\partial_n U_1^{\{k\}}]\|_{\partial K}\end{pmatrix} \cdot \begin{pmatrix} Ch_K^{1+\alpha_1}|\phi_1|_{K,1+\alpha_1} \\ Ch_K^{1+\alpha_1}|\phi_1|_{K,1+\alpha_1}\end{pmatrix} \\
& + \sum_{K\in T_{2,h}} \begin{pmatrix} \|f_2 - L_2U_2^{\{k\}}\|_K \\ \frac{1}{2}\|h_K^{-1/2}[A_2\partial_n U_2^{\{k\}}]\|_{\partial K}\end{pmatrix} \cdot \begin{pmatrix} Ch_K^{1+\alpha_2}|\phi_2|_{K,1+\alpha_2} \\ Ch_K^{1+\alpha_2}|\phi_2|_{K,1+\alpha_2}\end{pmatrix} \\
& + \left(\|h_1^{-1/2}(\sigma^{\{k\}} - A_1\partial_n U_1^{\{k\}})\|_\Gamma\right) \cdot \left(Ch_1^{1+\alpha_1}|\phi_1|_{1+\alpha_1}\right) \\
& + \left(\|h_2^{-1/2}(A_2\partial_n U_2^{\{k\}} - \sigma^{\{k\}})\|_\Gamma\right) \cdot \left(Ch_2^{1+\alpha_2}|\phi_2|_{1+\alpha_2}\right) \\
& + \left(\|A_1\partial_n\phi_1\|_\Gamma\right) \cdot \left(\|U_1^{\{k\}} - \pi_1 U_2^{\{k\}}\|_\Gamma\right).
\end{aligned}$$

Comparing Theorem 5.7 with Theorem 5.5 and Corollary 5.8 with Corollary 5.6, we see that the terms involving h^{β_1} and h^{β_2} have dropped out and the optimal order of convergence of the numerical approximation has been restored.

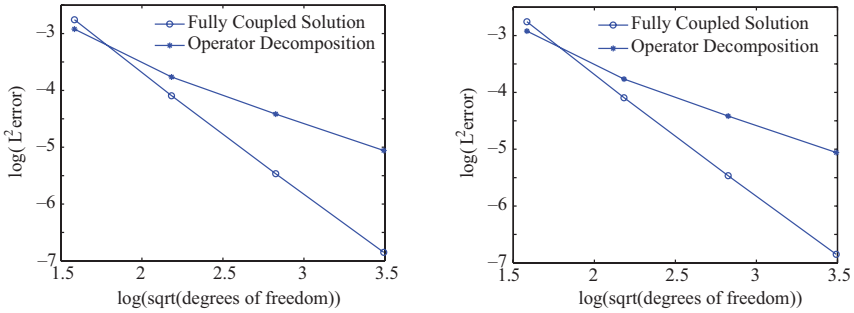


FIG. 5.1. *Comparison of the L^2 error in the fully coupled approximation and the operator decomposition approximation over Ω_1 (left) and Ω_2 (right) with mixed boundary conditions on $\partial\Omega_i\backslash\Gamma$ for $i = 1, 2$.*

5.2. Numerical results.

Example 5.1. Let $\Omega_1 = [0,1] \times [0,1]$ and $\Omega_2 = [1,2] \times [0,1]$, and assume that the triangulations match along $\Gamma = \{(x,y) |\ x = 1,\ 0 \le y \le 1\}$. Consider the elliptic interface problem

$$\begin{cases} -\nabla \cdot (\nabla u_1) = f_1, & x \in \Omega_1, \\ \begin{cases} u_1 = u_2, \\ \partial_n u_1 = \partial_n u_2, \end{cases} & x \in \Gamma, \\ -\nabla \cdot (\nabla u_2) = f_2, & x \in \Omega_2, \end{cases} \tag{5.3}$$

where the data $f_1(x,y)$ and $f_2(x,y)$ are chosen so the true solutions are $u_1 = u_2 = \sin(\pi x/2)\sin(2\pi y)$. The purpose of this experiment is to show how the asymptotic $\mathbf{O}(h_1^\beta)$ convergence rate may be masked by apparently minor changes to the problem and therefore may be difficult to observe in all situations. We first solve (5.3) with the boundary conditions

$$\begin{cases} u_1 = 0 & \text{along } x = 0 \text{ and } y = 0, \\ \partial_n u_1 = 2\pi \sin(\pi x/2) & \text{along } y = 1, \\ u_2 = 0 & \text{along } x = 2 \text{ and } y = 0, \\ \partial_n u_2 = 2\pi \sin(\pi x/2) & \text{along } y = 1. \end{cases}$$

We set $\alpha = 1/2$ and iterate until the convergence criteria is met. For comparison, we also solve (5.3) by using a fully coupled method on the same mesh in *COMSOL*. We see, in Figure 5.1, that the fully coupled solution converges quadratically, while the iterative solution converges linearly. Next, we solve (5.3) with Dirichlet boundary conditions

$$\begin{cases} u_1 = 0, & x \in \partial\Omega_1\backslash\Gamma, \\ u_2 = 0, & x \in \partial\Omega_2\backslash\Gamma, \end{cases}$$

and again we also solve the fully coupled problem for comparison. In Figure 5.2, we observe that the operator decomposition solution appears to converge quadratically, but eventually the $\mathbf{O}(h_1^\beta)$ term dominates and the convergence is ultimately linear.

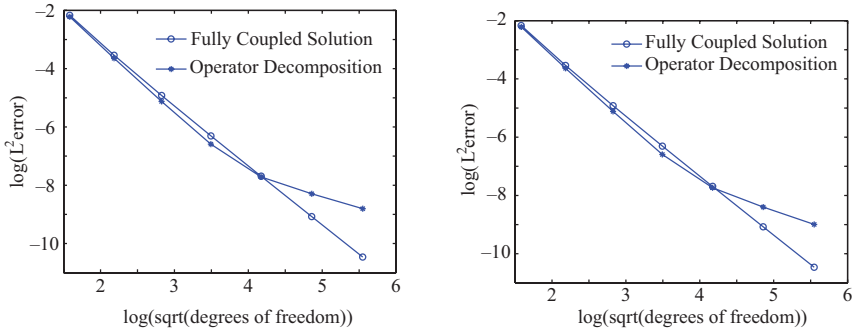


FIG. 5.2. *Comparison of the L^2 error in the fully coupled approximation and the iterative approximation over Ω_1 (left) and Ω_2 (right) with Dirichlet boundary conditions on $\partial\Omega_i\backslash\Gamma$ for $i = 1, 2$.*

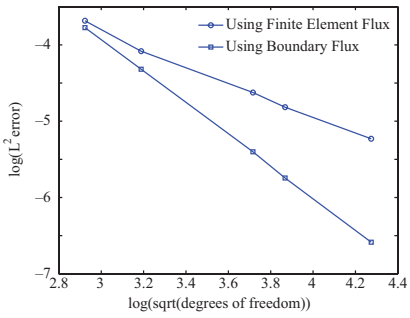


FIG. 5.3. *Comparison of the L^2 error in the iterative approximation using the finite element flux and the recovered boundary flux over $\Omega_1 \cup \Omega_2$ with mixed boundary conditions on $\partial\Omega_i\backslash\Gamma$ for $i = 1, 2$.*

Example 5.2. In the final example we assume that the triangulations do not match along the interface and consider (5.3) with the mixed boundary conditions

$$\begin{cases} u_1 = 0 & \text{along } x = 0, \\ A_1\partial_n u_1 = 2\pi \sin(\pi x/2) & \text{along } y = 1, \\ A_1\partial_n u_1 = -2\pi \sin(\pi x/2) & \text{along } y = 0, \\ u_2 = 0 & \text{along } x = 2, \\ A_2\partial_n u_2 = 2\pi \sin(\pi x/2) & \text{along } y = 1, \\ A_2\partial_n u_2 = -2\pi \sin(\pi x/2) & \text{along } y = 0. \end{cases}$$

First, we solve the problem iteratively by using operator decomposition by passing the finite element flux $A_1\partial_n U_1^{\{k\}}$. Next, we use the recovered boundary flux method to compute and pass $\sigma^{\{k\}}$. In Figure 5.3, we compare the L^2 errors over $\Omega_1 \cup \Omega_2$ on a series of meshes. We see that the approximations using the finite element flux converge linearly, while the approximations using the recovered boundary flux converge quadratically.

6. Conclusion. We have conducted an a posteriori error analysis of a finite element implementation of a number of operator decomposition strategies for a canonical conjugate heat transfer problem. By modifying the standard approach based on variational analysis, residuals, and the generalized Green's function, we derive accurate

error estimates and use these estimates to guide adaptive mesh refinement. Modifications to the standard error analysis account for the transfer of error between components of the decomposed operator and interpolation error. Our analysis also accounts for the differences between the adjoints of the full problem and the operator decomposition system. We show that the loss of order typically observed in the operator decomposition method is due to inaccuracy of the transferred gradient information, and we show how the boundary flux method can be used to efficiently regain the expected optimal order of convergence.

REFERENCES

[1] I. Babuska and M. Suri, *The p and h-p versions of the finite element method, basic principles and properties*, SIAM Rev., 36 (1994), pp. 578–632.

[2] I. Babuska, *The finite element method for elliptic equations with discontinuous coefficients*, Computing, 5 (1970), pp. 207–213.

[3] W. Bangerth and R. Rannacher, *Adaptive Finite Element Methods for Differential Equations*, Birkhauser-Basel, Verlag, Switzerland, 2003.

[4] R. Becker and R. Rannacher, *An optimal control approach to a posteriori error estimation in finite element methods*, Acta Numer., (2001), pp. 1–102.

[5] J. H. Bramble and J. T. King, *A finite element method for interface problems in domains with smooth boundaries and interfaces*, Adv. Comput. Math., 67 (1998), pp. 1–19.

[6] T. Cao, D. Kelly, and M. Ainsworth, *Some useful techniques for pointwise and local error estimates of the quantities of interest in the finite element approximation*, ANZIAM J., 42 (2000), pp. 317–339.

[7] G. F. Carey, S. S. Chow, and M. K. Seager, *Approximate boundary-flux calculations*, Comput. Methods Appl. Mech. Engrg., 50 (1985), pp. 107–120.

[8] G. F. Carey, *Derivative calculation from finite element solutions*, Comput. Methods Appl. Mech. Engrg., 35 (1982), pp. 1–14.

[9] V. Carey, D. Estep, and S. Tavener, *A posteriori analysis and adaptive error control for multiscale operator decomposition of elliptic systems* I: *Triangular systems*, SIAM J. Numer. Anal., to appear.

[10] V. Carey, D. Estep, and S. Tavener, *A Posteriori Analysis and Adaptive Error Control for Multiscale Operator Decomposition of Elliptic Systems* II: *Fully Coupled Systems*, manuscript, 2007.

[11] Z. Chen and J. Zou, *Finite element methods and their convergence for elliptic and parabolic interface problems*, Numer. Math, 79 (1998), pp. 175–202.

[12] K. Eriksson, D. Estep, P. Hansbo, and C. Johnson, *Introduction to adaptive methods for differential equations*, Acta Numer., (1995), pp. 105–158.

[13] K. Eriksson, D. Estep, P. Hansbo, and C. Johnson, *Computational Differential Equations*, Cambridge University Press, London, 1996.

[14] D. Estep, V. Ginting, D. Ropp, J. Shadid, and S. Tavener, *An a posteriori - A priori analysis of multiscale operator splitting*, SIAM J. Numer. Anal., 46 (2008), pp. 1116–1146.

[15] D. Estep, M. Holst, and M. Larson, *Generalized Green's functions and the effective domain of influence*, SIAM J. Sci. Comput., 26 (2005), pp. 1314–1339.

[16] D. Estep, S. Tavener, and T. Wildey, *Iterative Techniques for the Solution of Coupled Multiphysics Problems*, preprint, 2008.

[17] D. Estep, R. Williams, and M. Larson, *Estimating the error of numerical solutions of systems of reaction-diffusion equations*, Mem. Amer. Math. Soc., 146 (2000).

[18] D. Estep, *A posteriori error bounds and global error control for approximation of ordinary differential equations*, SIAM J. Numer. Anal., 32 (1995), pp. 1–48.

[19] D. Estep, *A Short Course on Duality, Adjoint Operators, Green's Functions, and A-Posteriori Error Analysis*, Sandia National Laboratories, Albuquerque, NM, 2004.

[20] M. Giles, M. G. Larson, J. M. Levenstam, and E. Suli, *Adaptive Error Control for Finite Element Approximations of the Lift and Drag Coefficients in Viscous Flow*, Technical report NA-97-06, Oxford University Computing Laboratory, 1997.

[21] M. Giles, *Stability analysis of numerical interface boundary conditions in fluid-structure thermal analysis*, Internat. J. Numer. Methods Fluids, 25 (1997), pp. 421–436.

[22] V. Heuveline and R. Rannacher, *Duality-based adaptivity in the hp-finite element method*, J. Numer. Math, 11 (2003), pp. 95–113.

[23] C. Lanczos, *Linear Differential Operators*, Dover, New York, 1997.
[24] J. Rice, P. Tsompanopoulus, and E. Vavalis, *Interface relaxation methods for elliptic differential equations*, Appl. Numer. Math., 32 (2000), pp. 219–245.
[25] B. Smith, P. Bjorstad, and W. Gropp, *Domain Decomposition*, Cambridge University Press, London, 1994.
[26] M. F. Wheeler, *A Galerkin procedure for estimating the flux for two-point boundary-value problems using continuous piecewise-polynomial spaces*, Numer. Math, 2 (1974), pp. 99–109.
[27] T. Wildey, D. Estep, and S. Tavener, *A posteriori error estimation of approximate boundary fluxes*, Comm. Numer. Methods Engrg., to appear.
[28] T. Wildey, *A Posteriori Analysis of Operator Decomposition in Interface Problems*, Ph.D. thesis, Colorado State University, 2007.
[29] D. Yang, *A parallel nonoverlapping schwarz domain decomposition method for elliptic interface problems*, IMA J. Numer. Anal., 16 (1996), pp. 75–91.
[30] I. Yotov, *A multilevel Newton-Krylov interface solver for multiphysics coupling of flow in porous media*, Numer. Linear Algebra Appl., 8 (2001), pp. 551–570.
[31] I. Yotov, *Interface solvers and preconditioners of domain decomposition type for multiphase flow in multiblock porous media*, in Adv. Comput. Theory Pract. 7, Nova Science, Hauppauge, NY, 2001, pp. 157–167.
[32] Z. Zhang and A. Naga, *A new finite element gradient recovery method: Superconvergence property*, SIAM J. Sci. Comput., 26 (2005), pp. 1192–1213.
[33] O. Zienkiewicz and J. Zhu, *The superconvergent patch recovery and a posteriori error estimates*, Internat. J. Numer. Methods Engrg., 33 (1992), pp. 1331–1364 and 1365–1382.

SIAM J. NUMER. ANAL.
Vol. 46, No. 4, pp. 2090–2111
© 2008 Society for Industrial and Applied Mathematics

A BDDC ALGORITHM FOR MORTAR DISCRETIZATION OF ELASTICITY PROBLEMS*

HYEA HYUN KIM†

Abstract. A balancing domain decomposition by constraints (BDDC) algorithm is developed for compressible elasticity problems in three dimensions with mortar discretization on geometrically nonconforming subdomain partitions. Material parameters of the elasticity problems may have jump across the subdomain interface. Coarse basis functions in the BDDC algorithm are constructed from primal constraints on faces, which are similar to the average matching condition and the moment matching condition considered in [A. Klawonn and O. B. Widlund, *Comm. Pure Appl. Math.*, 59 (2006), pp. 1523–1572] and [H. H. Kim, *A FETI-DP Formulation of Three Dimensional Elasticity Problems with Mortar Discretization*, Technical report, New York University, 2005]. A condition number bound is proved to be $C(1+\log(H/h))^3$ for geometrically nonconforming partitions and to be $C(1+\log(H/h))^2$ for geometrically conforming partitions. The bound is not affected by the jump of the material parameters across the subdomain interface. Numerical results are included.

Key words. BDDC algorithm, mortar discretization, preconditioner, elasticity problems

AMS subject classifications. 65N30, 65N55

DOI. 10.1137/060649677

1. Introduction.

This study is intended for development of an efficient balancing domain decomposition by constraints (BDDC) algorithm for solving mortar discretization of elasticity problems. The mortar discretization is applied to geometrically nonconforming subdomain partitions. In the geometrically nonconforming partitions, a pair of subdomains can intersect over only part of a subdomain face or a subdomain edge. Allowing geometrically nonconforming partitions makes mortar discretizations much more practical for problems given on three-dimensional complex structures with composite materials.

BDDC algorithms have been studied extensively in [16, 18, 14, 13, 15, 7] since their introduction by Dohrmann [3]. A coarse component of BDDC preconditioners is built from the weighted sum of functions that minimize discrete local energies subject to certain primal constraints across the subdomain interface. The BDDC algorithms solve linear systems of primal unknowns, in contrast to FETI-DP algorithms, which have been applied successfully to solving algebraic systems of dual unknowns during the last decade [4, 11, 12, 10, 6, 9]. This feature makes the BDDC algorithms more robust and more flexible than the FETI-DP algorithms; see [18, 15]. On the other hand, the BDDC and FETI-DP algorithms have much in common. They share the same spectra except the eigenvalues 0 and 1 when the same set of primal constraints is chosen; see [16, 14]

Both of these methods have been applied to solving problems arising from mortar discretizations. In a recent study by the author jointly with Dryja and Widlund [7], a BDDC algorithm is developed for elliptic problems. This algorithm was shown to share the same spectra with the FETI-DP algorithm considered in [8] by the author

*Received by the editors January 12, 2006; accepted for publication (in revised form) April 10, 2007; published electronically May 2, 2008. The work of this author was supported in part by the BK21 project and in part by the U.S. Department of Energy under contract DE-FC02-01ER25482.
http://www.siam.org/journals/sinum/46-4/64967.html

†National Institute for Mathematical Sciences, 385-16, Doryong-dong, Yuseong-gu, Daejeon 305–340, South Korea. Current address: Department of Mathematics, Chonnam National University, Gwangju, South Korea (hhk@nims.re.kr).

and Lee. In addition, it is generalized to the geometrically nonconforming partitions with a slightly weaker condition number bound, $C(1+\log(H_i/h_i))^3$. The FETI-DP algorithm in [8] has been extended to solving elasticity problems by the author [6]. However, the work is limited to geometrically conforming partitions. This study aims at developing a BDDC algorithm for the elasticity problems that allows the geometrically nonconforming partitions.

FETI-DP algorithms for compressible elasticity problems employing the exact matching condition have been developed by Klawonn and Widlund [10]. Even for the compressible case, extending FETI-DP algorithms to the three-dimensional elasticity problems needed much more effort than the standard second-order problems, i.e., elliptic problems. A special difficulty arises in building certain functionals dual to six rigid body motions. These functionals are important in providing a condition number bound independent of the number of subdomains. The most important ingredient of this study is a new set of primal constraints that makes it possible to construct such functionals in geometrically nonconforming partitions employing the mortar matching condition. Using the new set of primal constraints, we build a scalable BDDC algorithm with its condition number bound, $C(1+\log(H_i/h_i))^3$.

This paper is organized as follows. In section 2, we introduce a model compressible elasticity problem. In section 3, a BDDC algorithm is developed for mortar discretization of the elasticity problem employing coarse basis functions related to the primal constraints across the subdomain interface. Section 4 is devoted to analyzing the condition number bound of the BDDC algorithm. Numerical results are provided in section 5.

Throughout the paper, c and C denote generic positive constants independent of mesh parameters, number of subdomains, and coefficients of the elasticity problems. We will use h_i and H_i to denote the mesh size and the subdomain size of Ω_i, respectively.

2. A model problem. Let Ω be a polyhedral domain in $\mathbf{R}^3$. The Sobolev space $H^1(\Omega)$ is a set of functions in $L^2(\Omega)$ that are square integrable up to first weak derivatives, and it is equipped with the Sobolev norm

$$\|v\|^2_{1,\Omega} := |v|^2_{1,\Omega} + \frac{1}{H^2}\|v\|^2_{0,\Omega},$$

where $|v|^2_{1,\Omega} = \int_\Omega \nabla v \cdot \nabla v \, dx$, $\|v\|^2_{0,\Omega} = \int_\Omega v^2 \, dx$, and H is the diameter of Ω.

We assume that $\partial\Omega$ is divided into two parts, $\partial\Omega_D$ and $\partial\Omega_N$, on which a Dirichlet boundary condition and a natural boundary condition are given, respectively. The subspace $H^1_D(\Omega) \subset H^1(\Omega)$ is a set of functions having zero trace on $\partial\Omega_D$. We introduce the vector-valued Sobolev spaces, $[H^1_D(\Omega)]^3$ and $[H^1(\Omega)]^3$, equipped with the usual product norm.

We then consider the elasticity problem: Find $\mathbf{u} \in [H^1_D(\Omega)]^3$ such that

$$\int_\Omega G(\mathbf{x})\varepsilon(\mathbf{u}) : \varepsilon(\mathbf{v}) \, d\mathbf{x} + \int_\Omega G(\mathbf{x})\beta(\mathbf{x})\nabla\cdot\mathbf{u}\nabla\cdot\mathbf{v} \, d\mathbf{x} = \langle \mathbf{F}, \mathbf{v}\rangle \quad \forall \mathbf{v} \in [H^1_D(\Omega)]^3, \tag{2.1}$$

where $G = E/(1+\nu)$ and $\beta = \nu/(1-2\nu)$ are material parameters depending on the Young's modulus $E > 0$ and the Poisson ratio $\nu \in (0, 1/2]$. We assume that ν is bounded away from $1/2$ so that we exclude the case of incompressible elasticity problems. We note that $G(\mathbf{x})$ and $\beta(\mathbf{x})$ can be discontinuous; i.e., they can have

different values at different parts of Ω. The linearized strain tensor is defined by

$$\varepsilon_{ij}(\mathbf{u}) := \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right), \quad i,j = 1,2,3,$$

and the tensor product and the force term are given by

$$\varepsilon(\mathbf{u}) : \varepsilon(\mathbf{v}) = \sum_{i,j=1}^{3} \varepsilon_{ij}(\mathbf{u})\varepsilon_{ij}(\mathbf{v}), \quad \langle \mathbf{F}, \mathbf{v}\rangle = \int_{\Omega} \mathbf{f}\cdot\mathbf{v}\, d\mathbf{x} + \int_{\partial\Omega_N} \mathbf{g}\cdot\mathbf{v} d\sigma.$$

Here $\mathbf{f}$ is the body force and $\mathbf{g}$ is the surface force on the natural boundary part $\partial\Omega_N$.

The space $\mathbf{ker}(\varepsilon)$ has the following six rigid body motions as its basis, which are three translations

$$\mathbf{r}_1 = \begin{pmatrix}1\\0\\0\end{pmatrix}, \quad \mathbf{r}_2 = \begin{pmatrix}0\\1\\0\end{pmatrix}, \quad \mathbf{r}_3 = \begin{pmatrix}0\\0\\1\end{pmatrix}, \tag{2.2}$$

and three rotations

$$\mathbf{r}_4 = \frac{1}{H}\begin{pmatrix}x_2 - \widehat{x}_2\\ -x_1 + \widehat{x}_1\\ 0\end{pmatrix}, \quad \mathbf{r}_5 = \frac{1}{H}\begin{pmatrix}-x_3 + \widehat{x}_3\\ 0\\ x_1 - \widehat{x}_1\end{pmatrix}, \quad \mathbf{r}_6 = \frac{1}{H}\begin{pmatrix}0\\ x_3 - \widehat{x}_3\\ -x_2 + \widehat{x}_2\end{pmatrix}. \tag{2.3}$$

Here $\widehat{\mathbf{x}} = (\widehat{x}_1, \widehat{x}_2, \widehat{x}_3) \in \Omega$ and H is the diameter of Ω. This shift and the scaling make the L_2-norm of the six vectors scale in the same way with H.

When Ω is partitioned into a set of subdomains, the elasticity problem given on a floating subdomain has purely natural boundary condition. The Korn inequalities provided in section 6 of [10] concern this case. In the following, Σ is any open subset of $\partial\Omega$ with positive measure and its diameter is comparable to that of Ω. Let $(\mathbf{u},\mathbf{r})_\Sigma$ be an L_2-inner product given by

$$(\mathbf{u},\mathbf{r})_\Sigma = \int_\Sigma \mathbf{u}\cdot\mathbf{r}\, ds.$$

We introduce two seminorms provided for the space $[H^{1/2}(\Sigma)]^3$, which is a trace space of $[H^1(\Omega)]^3$,

$$|\mathbf{u}|_{1/2,\Sigma} := \inf_{\substack{\mathbf{v}\in[H^1(\Omega)]^3\\ \mathbf{v}|_\Sigma = \mathbf{u}}} |\mathbf{v}|_{1,\Omega}, \quad |\mathbf{u}|_{E(\Sigma)} := \inf_{\substack{\mathbf{v}\in[H^1(\Omega)]^3\\ \mathbf{v}|_\Sigma = \mathbf{u}}} \|\varepsilon(\mathbf{v})\|_{0,\Omega}.$$

The following Korn inequality is provided in [10, Lemma 6.4].

LEMMA 2.1. *Let Ω be a* Lipschitz *domain and let Σ be a subset of $\partial\Omega$ with positive measure. Then there exists a constant $c > 0$ such that*

$$c|\mathbf{u}|_{1/2,\Sigma} \le |\mathbf{u}|_{E(\Sigma)} \le |\mathbf{u}|_{1/2,\Sigma}$$

for $\mathbf{u} \in [H^{1/2}(\Sigma)]^3$ satisfying $(\mathbf{u},\mathbf{r})_\Sigma = 0$ for all $\mathbf{r} \in \mathbf{ker}(\varepsilon)$.

Another important inequality can be found in [10, Lemma 6.6].

LEMMA 2.2. *Let Ω be a* Lipschitz *domain of the diameter H and let $\Sigma \subset \partial\Omega$ be an open subset with positive measure. Then there exists a constant $C > 0$ such that*

$$\inf_{\mathbf{r}\in\mathbf{ker}(\varepsilon)} \|\mathbf{u} - \mathbf{r}\|_{0,\Sigma}^2 \le CH|\mathbf{u}|_{E(\Sigma)}^2 \quad \forall \mathbf{u} \in [H^{1/2}(\Sigma)]^3.$$

We note that the infimum above occurs when $\mathbf{u} - \mathbf{r}$ satisfies

$$(\mathbf{u} - \mathbf{r}, \mathbf{s})_\Sigma = 0 \quad \forall \mathbf{s} \in \mathbf{ker}(\varepsilon).$$

3. BDDC formulation for elasticity.

3.1. Mortar discretization. We divide the domain Ω into a geometrically nonconforming partition $\{\Omega_i\}_{i=1}^N$, which is shape regular. We consider a compressible elasticity problem with coefficients $G(\mathbf{x})$ and $\beta(\mathbf{x})$ positive constants in each subdomain

$$G(\mathbf{x})|_{\Omega_i} = G_i, \quad \beta(\mathbf{x})|_{\Omega_i} = \beta_i.$$

The conforming P_1-finite element space $\mathbf{X}_i$ is associated with a quasi-uniform triangulation T_i of each subdomain Ω_i. In addition, functions in the space $\mathbf{X}_i$ satisfy the Dirichlet boundary condition on $\partial\Omega_i \cap \partial\Omega_D$. The triangulations $\{T_i\}_{i=1}^N$ may not match across the subdomain interface. We note that it is more economical to use larger mesh sizes for larger values G_i, since there are only small changes in the solution when $\mathbf{F}$ in (2.1) does not reflect the jump in the coefficient $G(\mathbf{x})$, so that allowing nonmatching triangulations is useful for implementing such adaptivity. We denote by $\mathbf{W}_i$ the trace space of $\mathbf{X}_i$ on $\partial\Omega_i$ and define the spaces

$$\mathbf{X} = \prod_{i=1}^N \mathbf{X}_i, \quad \mathbf{W} = \prod_{i=1}^N \mathbf{W}_i.$$

Functions in these spaces can be discontinuous across the subdomain interface.

In three dimensions, a pair of subdomains can have a face, an edge, or a vertex in common. We will regard only the common faces as the subdomain interface. In a geometrically nonconforming partition, a common face can be only a part of a subdomain face. The union of all common faces is denoted by

$$\Gamma = \left(\bigcup_{ij} \partial\Omega_i \cap \partial\Omega_j \right) \setminus \partial\Omega.$$

Among the subdomain faces, we can select a collection of faces F_l (see [17, section 4.1]) such that

$$\bigcup_l \overline{F}_l = \overline{\Gamma}, \quad F_l \cap F_k = \emptyset,\ l \neq k.$$

We call F_l nonmortar faces. They are a full subdomain face.

Since the subdomain partition can be geometrically nonconforming, a single nonmortar face $F_l \subset \partial\Omega_i$ may intersect with several subdomain boundaries $\partial\Omega_j$. This provides F_l with a partition

$$\overline{F}_l = \bigcup_j \overline{F}_{ij}, \quad F_{ij} = \partial\Omega_i \cap \partial\Omega_j.$$

We call the opposite part F_{ij} mortar neighbors of F_l.

A dual or a standard Lagrange multiplier space $\mathbf{M}_l$ is given for each nonmortar face F_l. The space $\mathbf{M}_l$ is required to have the same dimension as the space

$$\overset{\circ}{\mathbf{W}}(F_l) := \mathbf{W}_i|_{F_l} \cap [H_0^1(F_l)]^3 \tag{3.1}$$

and to contain constant functions. Constructions of such Lagrange multiplier spaces have been studied in [1, 2, 19, 5]. On a nonmortar face $F_l \subset \partial\Omega_i$, a function $\boldsymbol{\phi}$, that is provided from its mortar neighbors of F_l, is given by

$$\boldsymbol{\phi} = \mathbf{v}_j \text{ on } F_{ij} \quad \forall j,$$

where $\mathbf{v} = (\mathbf{v}_1, \ldots, \mathbf{v}_N) \in \mathbf{X}$. The mortar matching condition on $\mathbf{v} \in \mathbf{X}$ is

$$\int_{F_l} (\mathbf{v}_i - \boldsymbol{\phi}) \cdot \boldsymbol{\lambda}\, ds = 0 \quad \forall \boldsymbol{\lambda} \in \mathbf{M}_l,\ \forall F_l. \tag{3.2}$$

The mortar discretization of the problem (2.1) is to approximate the solution by Galerkin's method in the mortar finite element space,

$$\widehat{\mathbf{X}} := \{\mathbf{v} \in \mathbf{X} \,:\, \mathbf{v} \text{ satisfies the mortar matching condition (3.2)}\}.$$

3.2. Primal constraints and change of variables. We will first select certain primal constraints on functions in $\mathbf{X}$ (or $\mathbf{W}$) from the mortar matching condition (3.2) to construct coarse basis functions. These basis functions will be used to build a preconditioner in our BDDC algorithm. In addition, we introduce a change of variables (basis) for the unknowns (functions) in the space $\mathbf{W}$ to make the primal constraints explicit.

On a nonmortar face F_l, we consider the rigid body motions $\{\mathbf{r}_k\}_{k=1}^6$ as in (2.2) and (2.3), where H is the diameter of the face F_l and $\widehat{\mathbf{x}}$ is a point in F_l. We define a nodal interpolant $I_{F_l} : L^2(F_l) \to M_l$ such that

$$I_{F_l}(w)(x) = \sum_m w(x_m)\psi_m(x), \tag{3.3}$$

where ψ_m is the basis element of M_l to the node x_m in the finite element space $\overset{\circ}{\mathbf{W}}(F_l)$; see (3.1). We extend the nodal interpolant to the product spaces $[L^2(F_l)]^3$ and $\mathbf{M}_l (= [M_l]^3)$, and we denote it by $\mathbf{I}_{F_l}$. By using the interpolated rigid body motions $\{\mathbf{I}_{F_l}(\mathbf{r}_k)\}_{k=1}^6$, we select the following six primal constraints from (3.2) on each nonmortar face $F_l \subset \partial\Omega_i$:

$$\int_{F_l} (\mathbf{w}_i - \boldsymbol{\phi}) \cdot \mathbf{I}_{F_l}(\mathbf{r}_k)\, ds = 0 \quad \forall k = 1, \ldots, 6, \tag{3.4}$$

where $\boldsymbol{\phi} = \mathbf{w}_j$ on $F_{ij}\, (= \partial\Omega_i \cap \partial\Omega_j) \subset F_l$.

In geometrically conforming partitions, i.e., F_l is a full face of two subdomains, the above constraints with $\{\mathbf{I}_{F_l}(\mathbf{r}_k)\}_{k=1}^3$ are face average matching conditions because $\mathbf{I}_{F_l}(\mathbf{r}_k) = \mathbf{r}_k$, $k = 1, 2, 3$, and $\boldsymbol{\phi} = \mathbf{w}_j$ on F_l. The remaining constraints with $\{\mathbf{I}_{F_l}(\mathbf{r}_k)\}_{k=4}^6$ are similar to the moment matching constraints which were introduced for fully primal edges in [10]. We call the constraints with $\{\mathbf{I}_{F_l}(\mathbf{r}_k)\}_{k=1}^3$ average constraints and call the other constraints with $\{\mathbf{I}_{F_l}(\mathbf{r}_k)\}_{k=4}^6$ moment constraints.

We now introduce a change of variables to make the primal constraints explicit; see Li and Widlund [14] or Kim, Dryja, and Widlund [7]. The change of variables leads to much simpler analysis of the BDDC algorithms.

On a nonmortar face $F_l \subset \partial\Omega_i$, unknowns $\mathbf{w}_i$ are represented by new unknowns $(\widehat{\mathbf{w}}_{\Delta,i}, \widehat{\mathbf{w}}_{\Pi,i})$ using a transform T_l such that

$$\mathbf{w}_i|_{F_l} = T_l \begin{pmatrix} \widehat{\mathbf{w}}_{\Delta,i} \\ \widehat{\mathbf{w}}_{\Pi,i} \end{pmatrix}, \tag{3.5}$$

with the unknowns $\widehat{\mathbf{w}}_{\Pi,i}$ of six components,

$$(\widehat{\mathbf{w}}_{\Pi,i})_k = \frac{\int_{F_l} \mathbf{w}_i \cdot \mathbf{I}_{F_l}(\mathbf{r}_k)\, ds}{\int_{F_l} \mathbf{I}_{F_l}(\mathbf{r}_k) \cdot \mathbf{I}_{F_l}(\mathbf{r}_k)\, ds}, \qquad k = 1, \ldots, 6,$$

and with $\widehat{\mathbf{w}}_{\Delta,i}$ giving the following six values zero, i.e.,

$$\int_{F_l} \mathbf{w}_{\Delta,i} \cdot \mathbf{I}_{F_l}(\mathbf{r}_k)\, ds = 0, \quad k = 1, \ldots, 6, \qquad \text{with } \mathbf{w}_{\Delta,i} = T_l \begin{pmatrix} \widehat{\mathbf{w}}_{\Delta,i} \\ 0 \end{pmatrix}.$$

We call $\widehat{\mathbf{w}}_{\Pi,i}$ primal unknowns. In addition, the transform can be built in such a way that it changes unknowns of which nodal basis are supported in the face F_l and it retains the other unknowns. Therefore, the transform can be applied to each nonmortar face F_l independently; see [7, section 2.2] for a detailed construction.

We now consider a change of variables on mortar neighbors of F_l. The nonmortar face F_l is partitioned into $\{F_{ij}\}_j$, where $F_{ij} = \partial\Omega_i \cap \partial\Omega_j$. On each mortar interface F_{ij} we introduce primal variables $\widehat{\mathbf{w}}_{\Pi,ij}$ of six components,

$$(\widehat{\mathbf{w}}_{\Pi,ij})_k = \frac{\int_{F_{ij}} \mathbf{w}_j \cdot \mathbf{I}_{F_l}(\mathbf{r}_k)\, ds}{\int_{F_{ij}} \mathbf{I}_{F_l}(\mathbf{r}_k) \cdot \mathbf{I}_{F_l}(\mathbf{r}_k)\, ds}, \qquad k = 1, \ldots, 6.$$

Similarly as before, we take a transform T_{ij} such that

$$\mathbf{w}_j|_{F_{ij}} = T_{ij} \begin{pmatrix} \widehat{\mathbf{w}}_{\Delta,ij} \\ \widehat{\mathbf{w}}_{\Pi,ij} \end{pmatrix},$$

with $\widehat{\mathbf{w}}_{\Delta,ij}$ giving the following six values zero:

$$\int_{F_{ij}} \mathbf{w}_{\Delta,ij} \cdot \mathbf{I}_{F_l}(\mathbf{r}_k)\, ds = 0, \quad k = 1, \ldots, 6, \qquad \text{with } \mathbf{w}_{\Delta,ij} = T_{ij} \begin{pmatrix} \widehat{\mathbf{w}}_{\Delta,ij} \\ 0 \end{pmatrix}.$$

The primal constraints (3.4) on F_l are then written as

$$(\widehat{\mathbf{w}}_{\Pi,i})_k = \sum_j A_{jk} (\widehat{\mathbf{w}}_{\Pi,ij})_k, \quad k = 1, \ldots, 6, \tag{3.6}$$

where

$$A_{jk} = \frac{\int_{F_{ij}} \mathbf{I}_{F_l}(\mathbf{r}_k) \cdot \mathbf{I}_{F_l}(\mathbf{r}_k)\, ds}{\int_{F_l} \mathbf{I}_{F_l}(\mathbf{r}_k) \cdot \mathbf{I}_{F_l}(\mathbf{r}_k)\, ds},$$

and $(\mathbf{v})_k$ denotes the kth component of the vector $\mathbf{v}$.

Remark 3.1. We note that the primal constraints (3.6) are new. In the previous work on elliptic problems [7], primal constraints are given on every piece F_{ij} of F_l

$$\int_{F_{ij}} (w_i - w_j)\, \psi_{ij}\, ds,$$

where ψ_{ij} is the sum of basis of M_l supported in $\overline{F}_{ij}$. Note that $\psi_{ij} = 1$ for the geometrically conforming partitions.

Remark 3.2. Instead of the interpolant $\mathbf{I}_{F_l}$, we might be able to introduce the interpolant $\mathbf{I}_{F_{ij}} \colon [L^2(F_{ij})]^3 \to \mathbf{M}_{ij}$, where $\mathbf{M}_{ij}$ is the space with its basis $\boldsymbol{\psi} \in \mathbf{M}_l$ supported in $\overline{F}_{ij}$. Correspondingly, we might select primal constraints similar to those in [7] such that

$$\int_{F_{ij}} (\mathbf{w}_i - \mathbf{w}_j) \cdot \mathbf{I}_{F_{ij}}(\mathbf{r}_k)\, ds = 0.$$

However, the above constraints need more primal unknowns on the nonmortar face F_l than those in (3.6).

These transforms can be applied independently to each nonmortar face F_l and each mortar interface F_{ij}. After applying the change of variables, we write the unknowns in each subdomain boundary as

$$\widehat{\mathbf{w}}_i = \begin{pmatrix} \widehat{\mathbf{w}}_{\Delta}^{(i)} \\ \widehat{\mathbf{w}}_{\Pi}^{(i)} \end{pmatrix},$$

where the subscript Π denotes the primal unknowns. To simplify notation, we will omit the hat notation for the new unknowns. Accordingly, the space $\mathbf{W}_i$ can be decomposed into

$$\mathbf{W}_i = \mathbf{W}_{\Delta}^{(i)} \times \mathbf{W}_{\Pi}^{(i)}.$$

Here the space $\mathbf{W}_{\Pi}^{(i)}$ consists of primal unknowns in Ω_i and $\mathbf{W}_{\Delta}^{(i)}$ is the space of the other unknowns. Similarly, after the change of unknowns (basis), the Schur complement matrix S_i and the Schur complement forcing vector $\mathbf{g}_i$ of the local elasticity problem can be written as

$$S_i = \begin{pmatrix} S_{\Delta\Delta}^{(i)} & S_{\Delta\Pi}^{(i)} \\ S_{\Pi\Delta}^{(i)} & S_{\Pi\Pi}^{(i)} \end{pmatrix}, \quad \mathbf{g}_i = \begin{pmatrix} \mathbf{g}_{\Delta}^{(i)} \\ \mathbf{g}_{\Pi}^{(i)} \end{pmatrix}.$$

We note that S_i is obtained from the stiffness matrix K_i of the local elasticity problem after the elimination of the interior unknowns:

$$K_i = \begin{pmatrix} K_{II}^{(i)} & K_{I\Delta}^{(i)} & K_{I\Pi}^{(i)} \\ K_{\Delta I}^{(i)} & K_{\Delta\Delta}^{(i)} & K_{\Delta\Pi}^{(i)} \\ K_{\Pi I}^{(i)} & K_{\Pi\Delta}^{(i)} & K_{\Pi\Pi}^{(i)} \end{pmatrix},$$

$$S_i = \begin{pmatrix} K_{\Delta\Delta}^{(i)} & K_{\Delta\Pi}^{(i)} \\ K_{\Pi\Delta}^{(i)} & K_{\Pi\Pi}^{(i)} \end{pmatrix} - \begin{pmatrix} K_{\Delta I}^{(i)} \\ K_{\Pi I}^{(i)} \end{pmatrix} {K_{II}^{(i)}}^{-1} \begin{pmatrix} K_{I\Delta}^{(i)} & K_{I\Pi}^{(i)} \end{pmatrix}.$$

We introduce the space $\widetilde{\mathbf{W}}$ of functions satisfying the primal constraints:

$$\widetilde{\mathbf{W}} = \{\mathbf{w} \in \mathbf{W} \,:\, \mathbf{w} \text{ satisfies the primal constraints (3.4)}\}.$$

In the following, we derive a matrix representation of the mortar matching condition (3.2) on functions in this space $\widetilde{\mathbf{W}}$.

We denote by $\mathbf{W}_{\Pi}$ the space of primal unknowns of all mortar faces and by $\mathbf{W}_{\Delta}$ the space of unknowns other than the primal unknowns, i.e.,

$$\mathbf{W}_{\Delta} = \prod_{i=1}^{N} \mathbf{W}_{\Delta}^{(i)}.$$

We call $\mathbf{W}_{\Pi}$ the space of global primal unknowns. From the relation (3.6), we can define the mapping from the global primal unknowns to the local primal unknowns:

$$R_{\Pi}^{(i)} \,:\, \mathbf{W}_{\Pi} \to \mathbf{W}_{\Pi}^{(i)}. \tag{3.7}$$

The matrix $R_{\Pi}^{(i)}$ is not a Boolean matrix for the geometrically nonconforming case, while it is a Boolean matrix for the geometrically conforming case. The mortar matching condition (3.2) on $\mathbf{w} \in \widetilde{\mathbf{W}}$ can be written as

$$B_{\Delta}\mathbf{w}_{\Delta} + B_{\Pi}\mathbf{w}_{\Pi} = 0. \tag{3.8}$$

Since $\mathbf{w} \in \widetilde{\mathbf{W}}$ satisfies the primal constraints (3.4) that are selected from (3.2), the above constraints (3.8) are redundant on $\mathbf{w} \in \widetilde{\mathbf{W}}$. We can make them nonredundant after deleting six equations from (3.8) corresponding to each nonmortar face F_l. We will use the same notation (3.8) for the nonredundant mortar matching constraints.

We further decompose

$$\mathbf{W}_{\Delta} = \mathbf{W}_{\Delta,n} \times \mathbf{W}_{\Delta,m},$$

where n and m denote unknowns at nonmortar faces (open) and the remaining unknowns, respectively. The condition (3.8) is then written as

$$B_n\mathbf{w}_n + B_m\mathbf{w}_m + B_{\Pi}\mathbf{w}_{\Pi} = 0. \tag{3.9}$$

We note that B_n is square and invertible, since the above equations are nonredundant and the number of unknowns $\mathbf{w}_n$ is equal to the number of the equations in (3.9).

3.3. A BDDC algorithm. We now formulate a BDDC operator for the mortar discretization of the elasticity problem (2.1). Since the matrix B_n in (3.9) is invertible,

$$\mathbf{w}_n = -B_n^{-1}(B_m\mathbf{w}_m + B_{\Pi}\mathbf{w}_{\Pi}). \tag{3.10}$$

We then define the matrix

$$R_{\Gamma} = \begin{pmatrix} -B_n^{-1}B_m & -B_n^{-1}B_{\Pi} \\ I & 0 \\ 0 & I \end{pmatrix}, \tag{3.11}$$

which maps the unknowns $(\mathbf{w}_m^t, \mathbf{w}_{\Pi}^t)^t$ to the unknowns $(\mathbf{w}_n^t, \mathbf{w}_m^t, \mathbf{w}_{\Pi}^t)^t$ satisfying the mortar matching condition (3.9). In practical implementation, it is beneficial to use the original mortar matching matrices without the change of unknowns. When $\mathbf{w}_m$ and $\mathbf{w}_{\Pi}$ are given, we perform the computation R_{Γ} as follows. From $\mathbf{w}_m$ and $\mathbf{w}_{\Pi}$, we can recover the original unknowns $\mathbf{w}_j$ on each nonmortar interface $F_{ij} \subset F_l$ using the transform T_{ij}. We denote by $\overline{\mathbf{w}}_m$ the unknowns $\mathbf{w}_j$ on the mortar interfaces F_{ij} and the unknowns in the boundary of the nonmortar face F_l. Let $\overline{\mathbf{w}}_n$ be the unknowns in the nonmortar face F_l except those on the boundary of F_l. Then we obtain $\overline{\mathbf{w}}_n$ by solving

$$\overline{\mathbf{w}}_n = \overline{B}_n^{-1}\overline{B}_m\overline{\mathbf{w}}_m.$$

Here $\overline{B}_n$ and $\overline{B}_m$ are the mortar matching matrix before the change of unknowns. After obtaining $\overline{\mathbf{w}}_n$, we perform a change of unknowns to $\overline{\mathbf{w}}_n$ and then take only the dual part $\mathbf{w}_n$ of $\overline{\mathbf{w}}_n$, so that $\mathbf{w}_n$ satisfies (3.10) for the given $\mathbf{w}_m$ and $\mathbf{w}_{\Pi}$. Lagrange multiplier spaces with dual basis give $\overline{B}_n$ a diagonal matrix, while the one with standard basis gives $\overline{B}_n$ a sparse matrix. The computation $\overline{B}_n^{-1}$ is much simpler and more efficient for the case with dual basis than for the standard case.

The mortar finite element space $\widehat{\mathbf{W}}$ can be described as

$$\widehat{\mathbf{W}} = \{\mathbf{w} = (\mathbf{w}_n, \mathbf{w}_m, \mathbf{w}_{\Pi}) \ : \ \mathbf{w} \text{ satisfies (3.9)}\}.$$

Here $\mathbf{w}_n$, $\mathbf{w}_m$, and $\mathbf{w}_\Pi$ are unknowns in the spaces $\mathbf{W}_{\Delta,n}$, $\mathbf{W}_{\Delta,m}$, and $\mathbf{W}_\Pi$, respectively. In addition, we express unknowns $(\mathbf{w}_n, \mathbf{w}_m, \mathbf{w}_\Pi)$ in the space $\widehat{\mathbf{W}}$ by using the matrix R_Γ and the unknowns $(\mathbf{w}_m, \mathbf{w}_\Pi)$:

$$\begin{pmatrix} \mathbf{w}_n \\ \mathbf{w}_m \\ \mathbf{w}_\Pi \end{pmatrix} = R_\Gamma \begin{pmatrix} \mathbf{w}_m \\ \mathbf{w}_\Pi \end{pmatrix}.$$

The mortar discretization of the elasticity problem is then given by

$$R_\Gamma^t \widetilde{S} R_\Gamma \begin{pmatrix} \mathbf{w}_m \\ \mathbf{w}_\Pi \end{pmatrix} = R_\Gamma^t \begin{pmatrix} \mathbf{g}_m \\ \mathbf{g}_\Pi \end{pmatrix}, \tag{3.12}$$

where

$$\widetilde{S} = \begin{pmatrix} S_{\Delta\Delta} & S_{\Delta\Pi} \\ S_{\Pi\Delta} & S_{\Pi\Pi} \end{pmatrix}, \quad \mathbf{g}_m = \begin{pmatrix} \mathbf{g}_m^{(1)} \\ \vdots \\ \mathbf{g}_m^{(N)} \end{pmatrix}, \quad \mathbf{g}_\Pi = \sum_i (R_\Pi^{(i)})^t \mathbf{g}_\Pi^{(i)}, \tag{3.13}$$

with $\mathbf{g}_m^{(i)}$ the part of $\mathbf{g}_\Delta^{(i)}$ to the unknowns except those interior to the nonmortar face, and

$$\begin{aligned} &S_{\Delta\Delta} = \mathrm{diag}_i \left(S_{\Delta\Delta}^{(i)} \right), \\ &S_{\Pi\Delta} = \textstyle\sum_i (R_\Pi^{(i)})^t S_{\Pi\Delta}^{(i)}, \quad S_{\Delta\Pi} = S_{\Pi\Delta}^t, \\ &S_{\Pi\Pi} = \textstyle\sum_i (R_\Pi^{(i)})^t S_{\Pi\Pi}^{(i)} R_\Pi^{(i)}. \end{aligned} \tag{3.14}$$

We note that $R_\Pi^{(i)}$ is the mapping defined in (3.7).

Our BDDC algorithm aims at solving (3.12) efficiently by the PCGM (preconditioned conjugate gradient method) with an appropriate preconditioner. The preconditioner will have local problems in Ω_i and a coarse problem in Ω as its building blocks.

The coarse problem is obtained from the coarse finite element space that is related to the primal constraints in (3.4) or (3.6). The coarse finite element space is spanned by the columns of the matrix Ψ,

$$\Psi = \begin{pmatrix} \Psi^{(1)} R_\Pi^{(1)} \\ \vdots \\ \Psi^{(N)} R_\Pi^{(N)} \end{pmatrix}, \quad \Psi^{(i)} = \begin{pmatrix} \Psi_\Delta^{(i)} \\ I_\Pi^{(i)} \end{pmatrix} = \begin{pmatrix} -(S_{\Delta\Delta}^{(i)})^{-1} S_{\Delta\Pi}^{(i)} \\ I_\Pi^{(i)} \end{pmatrix}, \tag{3.15}$$

where the matrix $I_\Pi^{(i)}$ is the identity matrix of size equal to the number of primal unknowns in Ω_i. We note that $\Psi_\Delta^{(i)}$ satisfies

$$\begin{pmatrix} S_{\Delta\Delta}^{(i)} & S_{\Delta\Pi}^{(i)} \\ S_{\Pi\Delta}^{(i)} & S_{\Pi\Pi}^{(i)} \end{pmatrix} \begin{pmatrix} \Psi_\Delta^{(i)} \\ I_\Pi^{(i)} \end{pmatrix} = \begin{pmatrix} 0 \\ F_{\Pi\Pi}^{(i)} \end{pmatrix}, \tag{3.16}$$

with

$$F_{\Pi\Pi}^{(i)} = S_{\Pi\Pi}^{(i)} - S_{\Pi\Delta}^{(i)} {S_{\Delta\Delta}^{(i)}}^{-1} S_{\Delta\Pi}^{(i)}. \tag{3.17}$$

Each column $\boldsymbol{\psi}$ of the matrix Ψ is related to one primal unknown in $\mathbf{W}_\Pi$. Since $\boldsymbol{\psi} \in \mathbf{W}$ satisfies the primal constraints (3.6), we take $\overline{\boldsymbol{\psi}} = (\boldsymbol{\psi}_\Delta^t, \boldsymbol{\psi}_\Pi^t)^t$ from the vector $\boldsymbol{\psi}$ so that $\boldsymbol{\psi}_\Delta \in \mathbf{W}_\Delta$ and $\boldsymbol{\psi}_\Pi \in \mathbf{W}_\Pi$. The matrix $\overline{\Psi}$ with the columns $\overline{\boldsymbol{\psi}}$ is then expressed by

$$\overline{\Psi} = R_\Pi^t - \sum_{i=1}^{N} (R_\Delta^{(i)})^t (S_{\Delta\Delta}^{(i)})^{-1} S_{\Delta\Pi}^{(i)} R_\Pi^{(i)}, \tag{3.18}$$

where $R_\Pi^t : \mathbf{W}_\Pi \to \mathbf{W}_\Delta \times \mathbf{W}_\Pi$ and $(R_\Delta^{(i)})^t : \mathbf{W}_\Delta^{(i)} \to \mathbf{W}_\Delta \times \mathbf{W}_\Pi$ are zero extensions.

We introduce

$$R_{D,\Gamma} = \begin{pmatrix} D_{nn} & & \\ & D_{mm} & \\ & & D_{\Pi\Pi} \end{pmatrix} R_\Gamma, \tag{3.19}$$

where the matrices D_{nn}, D_{mm}, and $D_{\Pi\Pi}$ will be specified later. We then propose the following preconditioner M^{-1} for the problem (3.12) with the local and coarse problems as its building blocks:

$$M^{-1} = R_{D,\Gamma}^t \left\{ \begin{pmatrix} S_{\Delta\Delta}^{-1} & 0 \\ 0 & 0 \end{pmatrix} + \overline{\Psi} (\Psi^t S \Psi)^{-1} \overline{\Psi}^t \right\} R_{D,\Gamma}, \tag{3.20}$$

where

$$S = \operatorname{diag}_i (S_i), \quad S_i = \begin{pmatrix} S_{\Delta\Delta}^{(i)} & S_{\Delta\Pi}^{(i)} \\ S_{\Pi\Delta}^{(i)} & S_{\Pi\Pi}^{(i)} \end{pmatrix},$$

and $S_{\Delta\Delta}$ is given in (3.14).

We will now express (3.20) in a simpler form. From the definition of Ψ in (3.15), we have

$$\Psi^t S \Psi = \sum_{i=1}^{N} (R_\Pi^{(i)})^t (\Psi^{(i)})^t S_i \Psi^{(i)} R_\Pi^{(i)}.$$

From (3.15)–(3.17), we obtain

$$\Psi^t S \Psi = \sum_{i=1}^{N} (R_\Pi^{(i)})^t F_{\Pi\Pi}^{(i)} R_\Pi^{(i)} = S_{\Pi\Pi} - S_{\Pi\Delta} S_{\Delta\Delta}^{-1} S_{\Delta\Pi}. \tag{3.21}$$

We define

$$F_{\Pi\Pi} = S_{\Pi\Pi} - S_{\Pi\Delta} S_{\Delta\Delta}^{-1} S_{\Delta\Pi}.$$

The definitions of the above matrices are given in (3.14). Using the block Cholesky factorization of $\widetilde{S}$ as in Li and Widlund [14] and above (see also (3.13)), we have

$$\begin{aligned} \widetilde{S}^{-1} = & \begin{pmatrix} S_{\Delta\Delta}^{-1} & 0 \\ 0 & 0 \end{pmatrix} \\ & + \left(R_\Pi^t - \sum_{i=1}^{N} (R_\Delta^{(i)})^t (S_{\Delta\Delta}^{(i)})^{-1} S_{\Delta\Pi}^{(i)} R_\Pi^{(i)} \right) F_{\Pi\Pi}^{-1} \\ & \times \left(R_\Pi^t - \sum_{i=1}^{N} (R_\Delta^{(i)})^t (S_{\Delta\Delta}^{(i)})^{-1} S_{\Delta\Pi}^{(i)} R_\Pi^{(i)} \right)^t. \end{aligned}$$

By combining the above equation with (3.18) and (3.21), we obtain

$$\widetilde{S}^{-1} = \begin{pmatrix} S_{\Delta\Delta}^{-1} & 0 \\ 0 & 0 \end{pmatrix} + \overline{\Psi}(\Psi^t S \Psi)^{-1} \overline{\Psi}^t.$$

Therefore, the BDDC operator (see (3.12)) with the preconditioner M^{-1} in (3.20) can be simply written as

$$B_{DDC} = R_{D,\Gamma}^t \widetilde{S}^{-1} R_{D,\Gamma} R_\Gamma^t \widetilde{S} R_\Gamma. \tag{3.22}$$

To help understand the BDDC algorithm we compare its implementation with the standard BDDC algorithm employing the direct matching condition, i.e., $u_i(x) = u_j(x)$ for the nodal points x on the subdomain interface.

Step 1. Set local stiffness matrices K_i and mortar matching matrices B_i.

Step 2. Perform change of basis to K_i depending on the primal constraints.

Step 3. Order K_i into interior, dual, and primal unknowns and B_i into unknowns in nonmortar faces (open) and the other unknowns:

$$K_i = \begin{pmatrix} K_{II}^{(i)} & K_{I\Delta}^{(i)} & K_{I\Pi}^{(i)} \\ K_{\Delta I}^{(i)} & K_{\Delta\Delta}^{(i)} & K_{\Delta\Pi}^{(i)} \\ K_{\Pi I}^{(i)} & K_{\Pi\Delta}^{(i)} & K_{\Pi\Pi}^{(i)} \end{pmatrix}, \quad B_i = \begin{pmatrix} B_n^{(i)} & B_m^{(i)} \end{pmatrix}.$$

Step 4. Compute LU factorization for $K_{II}^{(i)}$, $K_{rr}^{(i)}$, and $B_n^{(i)}$ (note that $K_{rr}^{(i)}$ stands for the part of K_i except the rows and columns to the primal unknowns).

Step 5. Compose the coarse problem matrix $F_{\Pi\Pi}$ and perform LU factorization for it:

$$F_{\Pi\Pi} = \sum_i (R_\Pi^{(i)})^t (K_{\Pi\Pi}^{(i)} - K_{\Pi r}^{(i)} (K_{rr}^{(i)})^{-1} K_{r\Pi}^{(i)}) R_\Pi^{(i)}.$$

Step 5. Perform CG iterations with the preconditioner. In each iteration, we perform the matrix-vector multiplications for the given residual r_Γ,

$$y_\Gamma = R_\Gamma^t \widetilde{S} R_\Gamma r_\Gamma, \quad z_\Gamma = R_{D,\Gamma}^t \widetilde{S}^{-1} R_{D,\Gamma} y_\Gamma.$$

We note that in Step 5 the LU factorizations of $K_{II}^{(i)}$ are used in the matrix-vector multiplication with $\widetilde{S}$, and the LU factorizations of $K_{rr}^{(i)}$ and $F_{\Pi\Pi}$ are used in the matrix-vector multiplication with $\widetilde{S}^{-1}$. In addition, LU factorizations of $B_n^{(i)}$ are used in $R_{\Gamma,D}$ and R_Γ. We note that B_i denotes the original mortar matching matrix without the change of unknowns. The computation R_Γ can be done using B_i as explained in the beginning of section 3.3. Additional complication compared to the standard BDDC algorithms is that we need to build the mortar matching matrices B_i and to perform the change of unknowns based on the primal constraints selected from the mortar matching condition. We also need to perform LU factorization to $B_n^{(i)}$. The other part of the BDDC algorithm is the same as the standard one.

4. Condition number estimate. In this section, we provide a condition number bound of the BDDC algorithm. We will first construct functionals $\{f_m\}_{m=1}^6$, dual to the space $\mathbf{ker}(\varepsilon)$, which satisfy the following properties:

$$f_m(\mathbf{r}_k) = \delta_{mk}, \quad m, k = 1, \ldots, 6,$$

$$|f_m(\mathbf{v})|^2 \le C \frac{\|\mathbf{v}\|_{0,\partial\Omega_i}^2}{H_i^2} \quad \text{for } \mathbf{v} \in [L^2(\partial\Omega_i)]^3. \tag{4.1}$$

Here $\mathbf{r}_k$ are basis of $\mathbf{ker}(\varepsilon)$ scaled with respect to a face $F \subset \partial\Omega_i$; this means that we take $\widehat{\mathbf{x}} \in F$ and $H = \mathrm{diam}(F)$ in (2.3). Such dual functionals were first introduced by Klawonn and Widlund [10]. An arbitrary rigid body motion $\mathbf{r}$ can then be represented by a linear combination of the elements of the basis $\{\mathbf{r}_k\}_{k=1}^6$,

$$\mathbf{r} = \sum_{k=1}^{6} f_k(\mathbf{r})\mathbf{r}_k.$$

We now introduce six linearly independent functionals $\left\{g_k^l\right\}_{k=1}^6$ related to the primal constraints across a nonmortar face $F_l \subset \partial\Omega_i$,

$$g_k^l(\mathbf{v}) = \frac{\int_{F_l} \mathbf{v} \cdot \mathbf{I}_{F_l}(\mathbf{r}_k^l)\, ds}{\int_{F_l} \mathbf{I}_{F_l}(\mathbf{r}_k^l) \cdot \mathbf{I}_{F_l}(\mathbf{r}_k^l)\, ds} \quad \text{for } \mathbf{v} \in [L^2(F_l)]^3,\ k = 1, \ldots, 6, \tag{4.2}$$

where $\mathbf{I}_{F_l}(\mathbf{r}_k^l)$ is the nodal interpolant defined in (3.3) and $\left\{\mathbf{r}_k^l\right\}_{k=1}^6$ are six rigid body motions scaled with respect to the face F_l. Since these functionals are linearly independent, they provide a basis of the dual space $(\mathbf{ker}(\varepsilon))'$. Thus there exists $\{\beta_{mk}\}_{m,k=1}^6$ such that

$$f_m^l = \sum_{k=1}^{6} \beta_{mk} g_k^l, \quad f_m^l(\mathbf{r}_k^l) = \delta_{mk}, \quad m = 1, \ldots, 6. \tag{4.3}$$

We will now show that $\{f_m^l\}_m$ satisfy (4.1). Since $\mathbf{r}_k^l$ are linear functions, one can show that

$$\|\mathbf{I}_{F_l}(\mathbf{r}_k^l)\|_{0,F_l}^2 \geq C H_i^2. \tag{4.4}$$

Here the constant C does not depend on any mesh parameters for sufficiently small h_i. From (4.2), (4.4), and Hölder's inequality, we obtain

$$|g_k^l(\mathbf{v})|^2 \leq C \frac{\|\mathbf{v}\|_{0,\partial\Omega_i}^2}{H_i^2} \quad \text{for } \mathbf{v} \in [L^2(\partial\Omega_i)]^3.$$

From this bound and (4.3), $\left\{f_m^l\right\}_{m=1}^6$ satisfy the bound in (4.1). We note that for $\mathbf{w} = (\mathbf{w}_1, \ldots, \mathbf{w}_N) \in \widetilde{\mathbf{W}}$, that satisfies the primal constraints, we have $g_k^l(\mathbf{w}_i) = g_k^l(\boldsymbol{\phi})$ (see (4.2) and (3.4)), so that

$$f_m^l(\mathbf{w}_i) = f_m^l(\boldsymbol{\phi}), \quad m = 1, \ldots, 6,$$

where $\boldsymbol{\phi} = \mathbf{w}_j$ on $F_{ij} \subset F_l$.

In the following, we will provide several lemmas which will be used to analyze the condition number bound of the BDDC operator. For a face $F \subset \partial\Omega_i$, the space $H_{00}^{1/2}(F)$ consists of the functions whose zero extension to the whole boundary $\partial\Omega_i$ belongs to the space $H^{1/2}(\partial\Omega_i)$ and is equipped with the norm

$$\|v\|_{H_{00}^{1/2}(F)} := \left(|v|_{H^{1/2}(F)}^2 + \int_F \frac{v(x)^2}{\mathrm{dist}(x, \partial F)}\, ds(x) \right)^{1/2},$$

where

$$|v|_{H^{1/2}(F)}^2 = \int_F \int_F \frac{|v(x) - v(y)|^2}{|x - y|^3}\, ds(x)\, ds(y).$$

The norm can be extended to the product space $[H_{00}^{1/2}(F)]^3$ by using the usual product norm.

We now provide several inequalities for the mortar projection of functions. We recall the space $\mathring{\mathbf{W}}(F_l)$ (see (3.1)) and the Lagrange multiplier space $\mathbf{M}_l$ given on the nonmortar face $F_l \subset \partial\Omega_i$.

DEFINITION 4.1. *The mortar projection* $\pi_l : [L^2(F_l)]^3 \to \mathring{\mathbf{W}}(F_l)$ *is defined by*

$$\int_{F_l} (\pi_l(\mathbf{v}) - \mathbf{v}) \cdot \boldsymbol{\psi}\, ds = 0 \quad \forall \boldsymbol{\psi} \in \mathbf{M}_l.$$

The mortar projection is continuous in both the L^2- and the $H_{00}^{1/2}$-norms when appropriate Lagrange multiplier spaces $\mathbf{M}_l$ are chosen; see [5, 19].

Let $F_l \subset \partial\Omega_i$ be a nonmortar face that is partitioned into $\{F_{ij}\}_j$ with $F_{ij} = \partial\Omega_i \cap \partial\Omega_j$, and let $\boldsymbol{\phi} = \mathbf{w}_j$ on F_{ij}. We then have the following result; see [7, Lemma 4.4 and Remark 4.5].

LEMMA 4.2.

$$\|\pi_l(\mathbf{w}_i)\|^2_{H_{00}^{1/2}(F_l)} \le C \left(1 + \log \frac{H_i}{h_i}\right)^2 \|\mathbf{w}_i\|^2_{1/2,\partial\Omega_i},$$

$$\|\pi_l(\boldsymbol{\phi})\|^2_{H_{00}^{1/2}(F_l)} \le C \left(1 + \log \frac{H_i}{h_i}\right)^3 \sum_j \|\mathbf{w}_j\|^2_{1/2,\partial\Omega_j}.$$

Remark 4.3. The additional factor $(1 + \log(H_i/h_i))$ in the second estimate came from the geometric nonconformity that causes $\boldsymbol{\phi} \in H^{1/2-\epsilon}(F_l)$, with $\epsilon > 0$. When F_l has only one mortar part, i.e., $F_l = F_{ij}$ and $\boldsymbol{\phi} = \mathbf{w}_j$, we obtain the usual estimate

$$\|\pi_l(\mathbf{w}_j)\|^2_{H_{00}^{1/2}(F_l)} \le C \left(1 + \log \frac{H_i}{h_i}\right)^2 \|\mathbf{w}_j\|^2_{1/2,\partial\Omega_j}. \tag{4.5}$$

The proof of the following lemma is provided in [6, Lemma 4.6].

LEMMA 4.4. *For the basis* $\{\mathbf{r}_m^l\}_{m=1}^6$ *of* $\mathbf{ker}(\varepsilon)$, *scaled with respect to the face* $F_l \subset \partial\Omega_i$, *we have*

$$\|\pi_l(\mathbf{r}_m^l)\|^2_{H_{00}^{1/2}(F_l)} \le C \left(1 + \log \frac{H_i}{h_i}\right) H_i \|\mathbf{r}_m^l\|^2_{\infty,F_l}.$$

DEFINITION 4.5. *A nonmortar face* $F_l \subset \partial\Omega_i$ *is called a simple face if it has only one mortar subdomain* Ω_j *such that* $F_l = \partial\Omega_i \cap \partial\Omega_j$.

Assumption 4.6. On each simple face F_l $(= \partial\Omega_i \cap \partial\Omega_j)$, its nonmortar subdomain Ω_i and its mortar subdomain Ω_j have been chosen so as to satisfy $G_i \le G_j$.

LEMMA 4.7. *Let* $F_l \subset \Omega_i$ *be a simple face, i.e.,* $F_l = \partial\Omega_i \cap \partial\Omega_j$, *and satisfy Assumption* 4.6. *Then, for* $\mathbf{w} = (\mathbf{w}_1, \ldots, \mathbf{w}_N) \in \widetilde{\mathbf{W}}$, *we have*

$$G_i \|\pi_l(\mathbf{w}_i - \mathbf{w}_j)\|^2_{H_{00}^{1/2}(F_l)} \le C \left(1 + \log \frac{H_i}{h_i}\right)^2 \left(|\mathbf{w}_i|^2_{S_i} + |\mathbf{w}_j|^2_{S_j}\right),$$

where $|\mathbf{w}_l|^2_{S_l} = \langle S_l \mathbf{w}_l, \mathbf{w}_l \rangle$ *for* $l = i, j$.

Proof. Let $\{\mathbf{r}_m^l\}_{m=1}^6$ be basis of $\mathbf{ker}(\varepsilon)$ scaled with respect to the face F_l, and let $\{f_m^l\}_{m=1}^6$ be the corresponding dual functionals to the basis. Since $\mathbf{w} =$

$(\mathbf{w}_1,\ldots,\mathbf{w}_N)\in\widetilde{\mathbf{W}}$ satisfies the primal constraints across the face F_l, the following identity holds:

$$\sum_{m=1}^{6} f_m^l(\mathbf{w}_i)\mathbf{r}_m^l = \sum_{m=1}^{6} f_m^l(\mathbf{w}_j)\mathbf{r}_m^l.$$

We then have

$$\begin{aligned}\|\pi_l(\mathbf{w}_i-\mathbf{w}_j)\|^2_{H_{00}^{1/2}(F_l)} \le\ & 2\left\|\pi_l\left(\mathbf{w}_i-\sum_{m=1}^{6} f_m^l(\mathbf{w}_i)\mathbf{r}_m^l\right)\right\|^2_{H_{00}^{1/2}(F_l)}\\ & +2\left\|\pi_l\left(\mathbf{w}_j-\sum_{m=1}^{6} f_m^l(\mathbf{w}_j)\mathbf{r}_m^l\right)\right\|^2_{H_{00}^{1/2}(F_l)}.\end{aligned}$$

We now estimate

$$\begin{aligned}&\left\|\pi_l\left(\mathbf{w}_i-\sum_{m=1}^{6} f_m^l(\mathbf{w}_i)\mathbf{r}_m^l\right)\right\|^2_{H_{00}^{1/2}(F_l)}\\ &\quad=\left\|\pi_l\left(\mathbf{w}_i-\mathbf{r}_i-\sum_{m=1}^{6} f_m^l(\mathbf{w}_i-\mathbf{r}_i)\mathbf{r}_m^l\right)\right\|^2_{H_{00}^{1/2}(F_l)}\\ &\quad\le C\left(\|\pi_l(\mathbf{w}_i-\mathbf{r}_i)\|^2_{H_{00}^{1/2}(F_l)}+\sum_{m=1}^{6}|f_m^l(\mathbf{w}_i-\mathbf{r}_i)|^2\|\pi_l(\mathbf{r}_m^l)\|^2_{H_{00}^{1/2}(F_l)}\right),\end{aligned}$$

where $\mathbf{r}_i\in\ \mathbf{ker}\ (\varepsilon)$ is an arbitrary rigid body motion. We note that for any $\mathbf{r}\in\ \mathbf{ker}\ (\varepsilon)$, $\mathbf{r}=\sum_{m=1}^{6} f_m^l(\mathbf{r})\mathbf{r}_m^l$.

From Lemmas 4.2, 2.1, and 2.2, the first term of the above expression is bounded by

(4.6)
$$\|\pi_l(\mathbf{w}_i-\mathbf{r}_i)\|^2_{H_{00}^{1/2}(F_l)} \le C\left(1+\log\frac{H_i}{h_i}\right)^2|\mathbf{w}_i|^2_{E(\partial\Omega_i)} \le C\frac{1}{G_i}\left(1+\log\frac{H_i}{h_i}\right)^2|\mathbf{w}_i|^2_{S_i},$$

and from (4.1) and Lemmas 4.4 and 2.2, the second term is bounded by

$$\begin{aligned}&|f_m^l(\mathbf{w}_i-\mathbf{r}_i)|^2\ \|\pi_l(\mathbf{r}_m^l)\|^2_{H_{00}^{1/2}(F_l)}\\ &\quad\le C\frac{\|\mathbf{w}_i-\mathbf{r}_i\|^2_{0,\partial\Omega_i}}{H_i^2}\left(1+\log\frac{H_i}{h_i}\right)H_i\|\mathbf{r}_m^l\|^2_{\infty,F_l}\\ &\quad\le C\left(1+\log\frac{H_i}{h_i}\right)|\mathbf{w}_i|^2_{E(\partial\Omega_i)}\\ &\quad\le C\frac{1}{G_i}\left(1+\log\frac{H_i}{h_i}\right)|\mathbf{w}_i|^2_{S_i}.\end{aligned}\tag{4.7}$$

Similarly, we obtain

$$\|\pi_l(\mathbf{w}_j-\mathbf{r}_j)\|^2_{H_{00}^{1/2}(F_l)} \le C\frac{1}{G_j}\left(1+\log\frac{H_i}{h_i}\right)^2|\mathbf{w}_j|^2_{S_j}$$

and

$$|f_m^l(\mathbf{w}_j - \mathbf{r}_j)|^2 \, \|\pi_l(\mathbf{r}_m^l)\|^2_{H_{00}^{1/2}(F_l)} \leq C \frac{1}{G_j} \left(1 + \log \frac{H_i}{h_i}\right) |\mathbf{w}_j|^2_{S_j}.$$

Using Assumption 4.6, i.e., $G_i \leq G_j$, the required bound has been shown. ∎

DEFINITION 4.8. *A pair of subdomains Ω_j and Ω_k has a* simple path *if there exists a path $(\Omega_j, \Omega_{k_1}, \ldots, \Omega_{k_p}, \Omega_k)$ connecting Ω_j and Ω_k through simple nonmortar faces, i.e., all common faces appearing in the path are simple faces.*

DEFINITION 4.9. *We say that a nonmortar face $F_l \subset \partial\Omega_i$, that is not a simple face, has* simple mortar neighbors *if any pair of subdomains (Ω_j, Ω_k) in its mortar neighbors has a simple path whose length is less than TOL and the subdomains Ω_m in the path satisfy*

$$\frac{G_i}{G_m} \leq TOL,$$

where G_i and G_m are the coefficients given in the subdomains Ω_i and Ω_m, respectively.

Assumption 4.10. We assume that any nonmortar face, that is not a simple face, has simple mortar neighbors.

Remark 4.11. When most of the nonmortar faces are simple faces and when nonmortar subdomains are chosen to have smaller G_i, Assumption 4.10 can be satisfied for such partitions with a relatively small TOL.

LEMMA 4.12. *Let a nonmortar face $F_l \subset \Omega_i$, that is not a simple face, satisfy Assumption* 4.10. *For $\mathbf{w} = (\mathbf{w}_1, \ldots, \mathbf{w}_N) \in \widetilde{\mathbf{W}}$, we then have*

$$G_i \|\pi_l(\mathbf{w}_i - \boldsymbol{\phi})\|^2_{H_{00}^{1/2}(F_l)} \leq C(TOL) \left(1 + \log \frac{H_i}{h_i}\right)^3 \left(|\mathbf{w}_i|^2_{S_i} + \sum_{m \in P(F_l)} |\mathbf{w}_m|^2_{S_m}\right),$$

where $P(F_l)$ is the set of subdomain indices appearing in the simple paths connecting any two mortar neighbors in $\{\Omega_j\}_j$ and the constant $C(TOL)$ depends only on TOL.

Proof. Let

$$\mathcal{F}_l(\mathbf{v}) = \sum_{m=1}^{6} f_m^l(\mathbf{v}) \mathbf{r}_m^l,$$

where $\{f_m^l\}_m$ are dual functionals provided for the basis $\{\mathbf{r}_m^l\}_m$, that are scaled with respect to the face F_l. We note that $\mathcal{F}_l(\mathbf{r}) = \mathbf{r}$ for any rigid body motion $\mathbf{r}$. On F_l, we define $\boldsymbol{\phi} = \mathbf{w}_j$ on $F_{ij}\,(:= \partial\Omega_i \cap \partial\Omega_j) \subset F_l$. Since $\mathbf{w}$ satisfies the primal constraints, we have $\mathcal{F}_l(\mathbf{w}_i) = \mathcal{F}_l(\boldsymbol{\phi})$ and obtain

$$\begin{aligned} &\|\pi_l(\mathbf{w}_i - \boldsymbol{\phi})\|^2_{H_{00}^{1/2}(F_l)} \\ &\qquad \leq 2\|\pi_l\left(\mathbf{w}_i - \mathcal{F}_l(\mathbf{w}_i)\right)\|^2_{H_{00}^{1/2}(F_l)} + \|\pi_l\left(\boldsymbol{\phi} - \mathcal{F}_l(\boldsymbol{\phi})\right)\|^2_{H_{00}^{1/2}(F_l)}. \end{aligned} \tag{4.8}$$

We can estimate the first term of (4.8) as in (4.6) and (4.7),

$$\|\pi_l\left(\mathbf{w}_i - \mathcal{F}_l(\mathbf{w}_i)\right)\|^2_{H_{00}^{1/2}(F_l)} \leq C \frac{1}{G_i} \left(1 + \log \frac{H_i}{h_i}\right)^2 |\mathbf{w}_i|^2_{S_i}. \tag{4.9}$$

Applying Lemma 4.2 to the second term of (4.8) gives

$$\|\boldsymbol{\phi}-\mathcal{F}_l(\boldsymbol{\phi})\|^2_{H^{1/2}_{00}(F_l)} \leq \left(1+\log\frac{H_i}{h_i}\right)^3 \sum_j \|\mathbf{w}_j-\mathcal{F}_l(\boldsymbol{\phi})\|^2_{1/2,\partial\Omega_j}.$$

It suffices to prove that

$$\sum_j \|\mathbf{w}_j-\mathcal{F}_l(\boldsymbol{\phi})\|^2_{1/2,\partial\Omega_j} \leq C \sum_{m\in P(F_l)} \frac{1}{G_m}|\mathbf{w}_m|^2_{S_m}, \tag{4.10}$$

where $P(F_l)$ is the set of subdomain indices appearing in the simple paths connecting any two subdomains in $\{\Omega_j\}_j$ that are mortar neighbors of Ω_i across F_l.

We now consider

$$\begin{aligned}
&\|\mathbf{w}_j-\mathcal{F}_l(\boldsymbol{\phi})\|^2_{1/2,\partial\Omega_j}\\
&\quad\leq 2\|\mathbf{w}_j-\mathbf{r}_j\|^2_{1/2,\partial\Omega_j}+2\|\mathbf{r}_j-\mathcal{F}_l(\boldsymbol{\phi})\|^2_{1/2,\partial\Omega_j}\\
&\quad\leq C\frac{1}{G_j}|\mathbf{w}_j|^2_{S_j}+2\|\mathcal{F}_l(\mathbf{r}_j-\boldsymbol{\phi})\|^2_{1/2,\partial\Omega_j}. \qquad (4.11)
\end{aligned}$$

The second term in (4.11) is estimated by

$$\begin{aligned}
&\|\mathcal{F}_l(\mathbf{r}_j-\boldsymbol{\phi})\|^2_{1/2,\partial\Omega_j}\\
&\quad\leq C\frac{\|\mathbf{r}_j-\boldsymbol{\phi}\|^2_{0,F_l}}{H_i^2}\max_m\left\{\|\mathbf{r}^l_m\|^2_{1/2,\partial\Omega_j}\right\}\\
&\quad\leq C\frac{1}{H_i^2}\frac{H_j^3}{H_i^2}\sum_{k,\,F_{ik}\subset F_l}\|\mathbf{r}_j-\mathbf{w}_k\|^2_{0,F_{ik}}\\
&\quad\leq C\frac{1}{H_i}\sum_{k,\,F_{ik}\subset F_l}\|\mathbf{r}_j-\mathbf{w}_k\|^2_{0,F_{ik}}. \qquad (4.12)
\end{aligned}$$

Here we have used the regularity of the partition, i.e., $H_i \simeq H_j$ for neighboring subdomains Ω_i and Ω_j, and

$$\|\mathbf{r}^l_m\|^2_{1/2,\partial\Omega_j} \leq \|\mathbf{r}^l_m\|^2_{1,\Omega_j} \leq CH_j^3/H_i^2.$$

We now consider the term $\|\mathbf{r}_j-\mathbf{w}_k\|^2_{0,F_{ik}}$. When $k=j$, from Lemmas 2.2 and 2.1 we obtain

$$\frac{1}{H_i}\|\mathbf{r}_j-\mathbf{w}_j\|^2_{0,F_{ij}} \leq C\frac{1}{G_j}|\mathbf{w}_j|^2_{S_j}.$$

When $k\neq j$, we note that there exists a path $(\Omega_j,\Omega_{k_1},\ldots,\Omega_{k_p},\Omega_k)$, where each pair of neighboring subdomains Ω_q and Ω_r in the path intersect on a simple face F_{qr}. Let

$$\mathcal{F}_{qr}(\mathbf{v})=\sum_m f^{qr}_m(\mathbf{v})\mathbf{r}^{qr}_m,$$

where $\{f^{qr}_m\}_m$ are dual functionals to the basis $\{\mathbf{r}^{qr}_m\}_m$ of $\mathbf{ker}(\varepsilon)$ scaled with respect to the face F_{qr}. We note that $\mathcal{F}_{qr}(\mathbf{w}_q)=\mathcal{F}_{qr}(\mathbf{w}_r)$.

We then obtain

$$
\begin{aligned}
&\|\mathbf{r}_j - \mathbf{w}_k\|^2_{0,F_{ik}} \\
&\quad = \|\mathbf{r}_j - \mathcal{F}_{jk_1}(\mathbf{w}_j) + \mathcal{F}_{jk_1}(\mathbf{w}_{k_1}) - \mathcal{F}_{k_1k_2}(\mathbf{w}_{k_1}) \\
&\qquad + \cdots + \mathcal{F}_{k_{p-1}k_p}(\mathbf{w}_{k_p}) - \mathcal{F}_{k_pk}(\mathbf{w}_{k_p}) + \mathcal{F}_{k_pk}(\mathbf{w}_k) - \mathbf{w}_k\|^2_{0,F_{ik}} \\
&\quad \le C\Big(\|\mathcal{F}_{jk_1}(\mathbf{r}_j - \mathbf{w}_j)\|^2_{0,F_{il}} \\
&\qquad + \|\mathcal{F}_{jk_1}(\mathbf{w}_{k_1} - \mathbf{r}_{k_1})\|^2_{0,F_{ik}} + \|\mathcal{F}_{k_1k_2}(\mathbf{w}_{k_1} - \mathbf{r}_{k_1})\|^2_{0,F_{ik}} \\
&\qquad + \cdots + \|\mathcal{F}_{k_{p-1}k_p}(\mathbf{w}_{k_p} - \mathbf{r}_{k_p})\|^2_{0,F_{ik}} + \|\mathcal{F}_{k_pk}(\mathbf{w}_{k_p} - \mathbf{r}_{k_p})\|^2_{0,F_{ik}} \\
&\qquad + \|\mathcal{F}_{k_pk}(\mathbf{w}_k - \mathbf{r}_k)\|^2_{0,F_{ik}} + \|\mathbf{r}_k - \mathbf{w}_k\|^2_{0,F_{ik}} \Big). \qquad (4.13)
\end{aligned}
$$

We note that the rigid body motions $\{\mathbf{r}_m^{qr}\}_{m=1}^6$ scaled with respect to the face F_{qr} in the path satisfy

$$\|\mathbf{r}_m^{qr}\|_{\infty,F_{ik}} \le C(TOL),$$

since the length of the path is less than TOL and the subdomain partition is regular. Using the above bound and Lemma 2.2, each term in (4.13) can be bounded in a similar way to

$$
\begin{aligned}
\|\mathcal{F}_{k_{p-1}k_p}(\mathbf{w}_{k_p} - \mathbf{r}_{k_p})\|^2_{0,F_{ik}} &\le C\frac{1}{H_{k_p}^2}\|\mathbf{w}_{k_p} - \mathbf{r}_{k_p}\|^2_{F_{k_{p-1}k_p}} \max_m \left\{\|\mathbf{r}_m^{k_{p-1}k_p}\|^2_{\infty,F_{ik}}\right\} H_i^2 \\
&\le C(TOL)\frac{H_i^2}{H_{k_p}}\frac{1}{H_{k_p}}\|\mathbf{w}_{k_p} - \mathbf{r}_{k_p}\|^2_{0,\partial\Omega_{k_p}} \\
&\le C(TOL)\frac{H_i^2}{H_{k_p}}|\mathbf{w}_{k_p}|^2_{E(\partial\Omega_{k_p})}.
\end{aligned}
$$

Therefore we obtain

$$
\begin{aligned}
\frac{1}{H_i}\|\mathbf{r}_j - \mathbf{w}_k\|^2_{0,F_{ik}} &\le C(TOL)\sum_{m\in P(j,k)}\frac{H_i}{H_m}|\mathbf{w}_m|^2_{E(\partial\Omega_m)} \\
&\le C(TOL)\sum_{m\in P(j,k)}\frac{1}{G_m}|\mathbf{w}_m|^2_{S_m}, \qquad (4.14)
\end{aligned}
$$

where $P(j,k)$ is the set of subdomain indices appearing in the path from Ω_j to Ω_k. Since the subdomain partition is regular, the factors H_i and H_m cancel each other in the final estimate. Combining (4.14) with (4.12) and (4.11), we obtain the estimate (4.10). Using $G_i/G_m \le TOL$ for $m \in P(j,k)$, the desired bound follows. ∎

A condition number analysis of BDDC algorithms has been carried out using an appropriate average operator by Li and Widlund [13]. The same technique was used for a condition number estimate of the BDDC algorithm for mortar discretization by

Kim, Dryja, and Widlund [7]. A bound for an average operator E_D in a certain norm is central in the analysis. We recall the definitions of R_Γ and $R_{D,\Gamma}$ in (3.11) and (3.19), respectively. We note that

$$R_\Gamma : \mathbf{W}_{\Delta,m} \times \mathbf{W}_\Pi \to \mathbf{W}_{\Delta,n} \times \mathbf{W}_{\Delta,m} \times \mathbf{W}_\Pi,$$

where $\mathbf{W}_\Pi$ is the space of the global primal unknowns, $\mathbf{W}_{\Delta,n}$ is the space of unknowns on nonmortar faces except the primal unknowns, and $\mathbf{W}_{\Delta,m}$ is the space of the remaining unknowns. We also note that

$$\mathbf{W}_\Delta = \mathbf{W}_{\Delta,n} \times \mathbf{W}_{\Delta,m}, \quad \widetilde{\mathbf{W}} = \mathbf{W}_\Delta \times \mathbf{W}_\Pi.$$

We now define the operator E_D by

$$E_D = R_\Gamma R_{D,\Gamma}^t. \tag{4.15}$$

Here the weight matrix D in (3.19) will be chosen so that

(P1) $R_\Gamma^t R_{D,\Gamma} = R_{D,\Gamma}^t R_\Gamma = I$,

(P2) $|E_D \mathbf{w}|_{\widetilde{S}}^2 \le C \max_i \left\{ \left(1 + \log \frac{H_i}{h_i}\right)^3 \right\} |\mathbf{w}|_{\widetilde{S}}^2$,

where $|\mathbf{w}|_{\widetilde{S}}^2 = \langle \widetilde{S}\mathbf{w}, \mathbf{w} \rangle$; see (3.13). We then have

$$R_\Gamma^t R_{D,\Gamma} \begin{pmatrix} \mathbf{w}_m \\ \mathbf{w}_\Pi \end{pmatrix} = \begin{pmatrix} -B_m^t (B_n^t)^{-1} D_{nn} \mathbf{z}_n + D_{mm} \mathbf{w}_m \\ -B_\Pi^t (B_n^t)^{-1} D_{nn} \mathbf{z}_n + D_{\Pi\Pi} \mathbf{w}_\Pi \end{pmatrix},$$

where

$$\mathbf{z}_n = -B_n^{-1}(B_m \mathbf{w}_m + B_\Pi \mathbf{w}_\Pi).$$

In order to satisfy property (P1), the weight matrix D is chosen to be

$$D_{nn} = 0, \quad D_{mm} = I, \quad D_{\Pi\Pi} = I. \tag{4.16}$$

Using Lemma 4.12, we can establish property (P2) for the operator E_D with the weight matrix D just given.

LEMMA 4.13. *With Assumptions* 4.6 *and* 4.10, *the operator* E_D *satisfies*

$$|E_D \mathbf{w}|_{\widetilde{S}}^2 \le C(TOL) \max_i \left\{ \left(1 + \log \frac{H_i}{h_i}\right)^3 \right\} |\mathbf{w}|_{\widetilde{S}}^2.$$

Proof. With the weight matrix D in (4.16), the operator E_D in (4.15) is given by

$$E_D \begin{pmatrix} \mathbf{w}_n \\ \mathbf{w}_m \\ \mathbf{w}_\Pi \end{pmatrix} = \begin{pmatrix} \mathbf{w}_n - B_n^{-1}(B_n \mathbf{w}_n + B_m \mathbf{w}_m + B_\Pi \mathbf{w}_\Pi) \\ \mathbf{w}_m \\ \mathbf{w}_\Pi \end{pmatrix}.$$

Let

$$\mathbf{z}_n = \mathbf{w}_n - B_n^{-1}(B_n \mathbf{w}_n + B_m \mathbf{w}_m + B_\Pi \mathbf{w}_\Pi).$$

We take $\mathbf{z}_i$ by restricting the unknowns $(\mathbf{z}_n, \mathbf{w}_m, \mathbf{w}_\Pi)$ to the subdomain Ω_i. Similarly, we take $\mathbf{w}_i$ from $(\mathbf{w}_n, \mathbf{w}_m, \mathbf{w}_\Pi)$. Let

$$\mathbf{z} = (\mathbf{z}_1, \dots, \mathbf{z}_n), \quad \mathbf{w} = (\mathbf{w}_1, \dots, \mathbf{w}_n).$$

We note that $\mathbf{w}$ satisfies the primal constraints and $\mathbf{z}$ satisfies the mortar matching condition. Each $\mathbf{z}_i$ is of the form

$$\mathbf{z}_i = \mathbf{w}_i - \sum_{F \subset \partial\Omega_i} E_F^{(i)} \pi_F(\mathbf{w}_i - \boldsymbol{\phi}),$$

where F is a nonmortar face in $\partial\Omega_i$, π_F is the mortar projection given on the face F, $E_F^{(i)}$ is the zero extension of functions defined on F to $\partial\Omega_i$, and $\boldsymbol{\phi} = \mathbf{w}_j$ on $F_{ij} (:= \partial\Omega_j \cap \partial\Omega_i) \subset F$. We then obtain

$$\begin{aligned}
|E_D\mathbf{w}|_{\widetilde{S}}^2 &= \sum_{i=1}^N \langle S_i \mathbf{z}_i, \mathbf{z}_i \rangle \\
&\le C \sum_{i=1}^N \left(\langle S_i \mathbf{w}_i, \mathbf{w}_i \rangle + \sum_{F \subset \partial\Omega_i} \langle S_i E_F^{(i)} \pi_F(\mathbf{w}_i - \boldsymbol{\phi}), E_F^{(i)} \pi_F(\mathbf{w}_i - \boldsymbol{\phi}) \rangle \right) \\
&\le C \sum_{i=1}^N \langle S_i \mathbf{w}_i, \mathbf{w}_i \rangle + \sum_{i=1}^N \sum_{F \subset \partial\Omega_i} G_i \|\pi_F(\mathbf{w}_i - \boldsymbol{\phi})\|_{H_{00}^{1/2}(F)}^2 \\
&\le C(TOL) \max_i \left\{ \left(1 + \log \frac{H_i}{h_i} \right)^3 \right\} \sum_{i=1}^N \langle S_i \mathbf{w}_i, \mathbf{w}_i \rangle \\
&\le C(TOL) \max_i \left\{ \left(1 + \log \frac{H_i}{h_i} \right)^3 \right\} \langle \widetilde{S}\mathbf{w}, \mathbf{w} \rangle.
\end{aligned}$$

Here we have used $\langle S_i \mathbf{w}_i, \mathbf{w}_i \rangle \le C G_i |\mathbf{w}_i|_{H^{1/2}(\partial\Omega_i)}^2$, $|E_F^{(i)} \pi_F(\mathbf{w}_i - \boldsymbol{\phi})|_{H^{1/2}(\partial\Omega_i)} \simeq \|\pi_F(\mathbf{w}_i - \boldsymbol{\phi})\|_{H_{00}^{1/2}(F)}$, and Lemma 4.12. ∎

By using the properties (P1) and (P2), we can show the following condition number bound of the BDDC operator in (3.22); see [7, Theorem 4.7] for the proof.

THEOREM 4.14. *If E_D satisfies the properties* (P1) *and* (P2), *then the BDDC operator has the following condition number bound*:

$$\kappa(B_{DDC}) \le C(TOL) \max_i \left\{ \left(1 + \log \frac{H_i}{h_i} \right)^3 \right\}.$$

Remark 4.15. The analysis above can be modified for the partitions where all nonmortar faces are simple, i.e., they have only one mortar interface. For such a case, geometrically conforming partitions as an example, we have a better condition number bound (see Lemma 4.7),

$$\kappa(B_{DDC}) \le C \max_i \left\{ \left(1 + \log \frac{H_i}{h_i} \right)^2 \right\}.$$

5. Numerical results. We present numerical results on a simple model problem to show the scalability of the suggested BDDC algorithm. We consider a model problem of (2.1) in a parallelepiped domain $\Omega = [0,1] \times [0,1] \times [0,1/2]$ with a zero Dirichlet boundary condition on $\partial\Omega$. We perform the BDDC algorithm for the case when the exact solution is zero, i.e., $\mathbf{F}$ is zero in (2.1). In the CG iteration, we start with a nonzero initial and perform the iteration up to the relative residual norm reduced by a factor of 10^6.

We consider a geometrically nonconforming partition of Ω for a given N as follows. We first partition Ω into N^3 uniform parallelepiped subregions. Each subregion has a geometrically nonconforming partition that consists of three subdomains as in Figure 1. One is a parallelepiped and the other two are cubes. As a result, we obtain a geometrically nonconforming subdomain partition of Ω. In the partition, we have cubical subdomains and parallelepiped subdomains. For a given n, the cubical subdomains are equipped with $(n-1)^3$ uniform cubic elements, and the parallelepiped subdomains are equipped with uniform parallelepiped elements by dividing their x, y, and z directional edges into $n-1$, $n+2$, and $n-1$ intervals, respectively. We test the BDDC algorithm in both the cases when the coefficient $G(x)$ is continuous and when $G(x)$ is discontinuous across the subdomain interface. Lagrange multipliers with dual basis developed in [5] are used in our computation.

We first present the result when $G(x) = 1$, i.e., continuous across the subdomain interface. We perform the BDDC algorithm as the number of subdomains increases and as the subdomain problem size increases. Table 1 presents the number of iterations and the condition numbers. The number of iterations and the condition numbers get stable with the increase of the number of the subdomains. Moreover, the con-

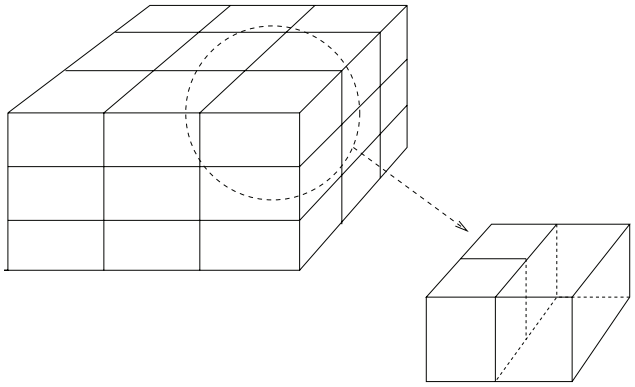


FIG. 1. *A geometrically nonconforming subdomain partition when $N = 3$; the domain Ω is partitioned into N^3 uniform parallelepiped subregions, and each subregion is then partitioned into a geometrically nonconforming subdomain partition.*

TABLE 1

Performance of the BDDC algorithm when $G(x) = 1$ with the increase of N, i.e., the number of subdomains, with a fixed local problem size ($n = 6$) and with the increase of the local problem size n with a fixed subdomain partition ($N = 3$); Cond (condition number), Iter (the number of iterations).

N	Cond	Iter	n	Cond	Iter
3	21.45	26	6	21.45	20
4	22.40	26	8	27.49	22
5	22.87	27	10	32.20	23
6	23.20	27	12	35.99	26

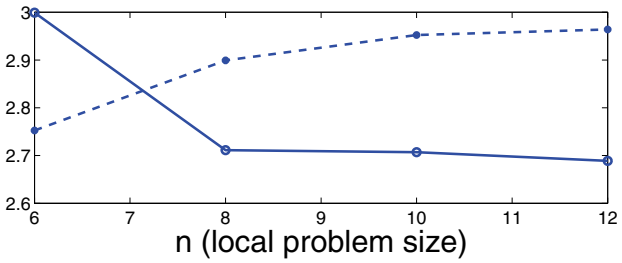


FIG. 2. *Plot of* $C = Cond/(1 + log(n))^2$ *as the local problem size* n *increases; dotted line for* $G(x) = 1$ *and solid line for* $G(x) = G_i$ *in* Ω_i *with Assumption* 4.6.

TABLE 2

Performance of the BDDC algorithm when $G(x) = G_i$ *for* $x \in \Omega_i$ *with the increase of* N *with a fixed local problem size* ($n = 6$) *and with the increase of the local problem size* n *in a geometrically nonconforming partition with* $N = 3$; *Cond (condition number), Iter (the number of iterations).*

With Assumption 4.6						Without Assumption 4.6					
N	Cond	Iter	n	Cond	Iter	N	Cond	Iter	n	Cond	Iter
2	18.61	21	6	23.37	26	2	626.6	108	6	913.3	133
3	23.37	26	8	25.71	27	3	913.3	133	8	1099	178
4	29.93	28	10	29.52	28	4	933.9	141	10	1226	188
5	34.15	30	12	32.65	30	5	941.3	145	12	1324	196

dition numbers show $\log^2$-behavior with the increase of the local problem size. We note that in our computation the nonmortar faces were selected to be simple, i.e., to have only one mortar face, so that we have the $\log^2$-bound for this case even in the geometrically nonconforming partitions; see Figure 2.

We perform the BDDC algorithm when $G(x)$ has jump across the subdomain interface. We select $G_i = 1$ for cubical subdomains and $G_i = 100$ for parallelepiped subdomains. Table 2 presents the performance of the algorithm with the increase of the number of subdomains and with the increase of the local problem size. To justify Assumption 4.6, we test both the cases when the selection of mortar and nonmortar faces satisfies and violates Assumption 4.6. The results confirm that the performance is nice when Assumption 4.6 is satisfied, while the performance is very poor when Assumption 4.6 is not satisfied. We can also observe that the condition number bound $C(1+\log(H/h))^2$ is in fact optimal for both cases when $G(x) = 1$ and $G(x) = G_i$, once the selection of mortar and nonmortar faces satisfies Assumption 4.6 and all nonmortar faces are simple; see Figure 2.

Acknowledgment. The author is grateful to Professor Olof B. Widlund at the Courant Institute for suggesting this problem.

REFERENCES

[1] F. B. BELGACEM AND Y. MADAY, *The mortar element method for three dimensional finite elements*, M2AN Math. Model. Numer. Anal., 31 (1997), pp. 289–302.

[2] C. BERNARDI, Y. MADAY, AND A. T. PATERA, *A new nonconforming approach to domain decomposition: The mortar element method*, in Nonlinear Partial Differential Equations and Their Applications, Collège de France Seminar, Vol. XI (Paris, 1989–1991), Pitman Res. Notes Math. Ser. 299, Longman Scientific and Technical, Harlow, UK, 1994, pp. 13–51.

[3] C. R. DOHRMANN, *A preconditioner for substructuring based on constrained energy minimization*, SIAM J. Sci. Comput., 25 (2003), pp. 246–258.

[4] C. Farhat, M. Lesoinne, P. LeTallec, K. Pierson, and D. Rixen, *FETI-DP: A dual-primal unified FETI method*. I. *A faster alternative to the two-level FETI method*, Internat. J. Numer. Methods Engrg., 50 (2001), pp. 1523–1544.
[5] C. Kim, R. D. Lazarov, J. E. Pasciak, and P. S. Vassilevski, *Multiplier spaces for the mortar finite element method in three dimensions*, SIAM J. Numer. Anal., 39 (2001), pp. 519–538.
[6] H. H. Kim, *A FETI-DP Formulation of Three Dimensional Elasticity Problems with Mortar Discretization*, Technical report, New York University, 2005.
[7] H. H. Kim, M. Dryja, and O. B. Widlund, *A BDDC Algorithm for Problems with Mortar Discretization*, Technical report, New York University, 2005.
[8] H. H. Kim and C.-O. Lee, *A preconditioner for the FETI-DP formulation with mortar methods in two dimensions*, SIAM J. Numer. Anal., 42 (2005), pp. 2159–2175.
[9] H. H. Kim and C.-O. Lee, *A Neumann–Dirichlet preconditioner for a FETI-DP formulation of the two-dimensional Stokes problem with mortar methods*, SIAM J. Sci. Comput., 28 (2006), pp. 1133–1152.
[10] A. Klawonn and O. B. Widlund, *Dual-primal FETI methods for linear elasticity*, Comm. Pure Appl. Math., 59 (2006), pp. 1523–1572.
[11] A. Klawonn, O. B. Widlund, and M. Dryja, *Dual-primal FETI methods for three-dimensional elliptic problems with heterogeneous coefficients*, SIAM J. Numer. Anal., 40 (2002), pp. 159–179.
[12] J. Li, *Dual-Primal FETI Methods for Incompressible Stokes and Linearized Navier-Stokes Equations*, Technical report, New York University, 2002.
[13] J. Li and O. Widlund, *BDDC Algorithms for Incompressible Stokes Equations*, Technical report, New York University, 2005.
[14] J. Li and O. B. Widlund, *FETI-DP, BDDC, and block Cholesky methods*, Internat. J. Numer. Methods Engrg., 66 (2006), pp. 250–271.
[15] J. Li and O. B. Widlund, *On the use of inexact subdomain solvers for BDDC algorithms*, Comput. Methods Appl. Mech. Engrg., 196 (2007), pp. 1415–1428.
[16] J. Mandel, C. R. Dohrmann, and R. Tezaur, *An algebraic theory for primal and dual substructuring methods by constraints*, Appl. Numer. Math., 54 (2005), pp. 167–193.
[17] D. Stefanica, *Domain Decomposition Methods for Mortar Finite Elements*, Ph.D. thesis, Department of Computer Science, Courant Institute, New York University, 2000.
[18] X. Tu, *Three-Level BDDC in Two Dimensions*, Technical report, New York University, 2004.
[19] B. I. Wohlmuth, *Discretization Methods and Iterative Solvers Based on Domain Decomposition*, Lecture Notes in Comput. Sci. Engrg. 17, Springer-Verlag, Berlin, 2001.

SIAM J. NUMER. ANAL.
Vol. 46, No. 4, pp. 2112–2132
© 2008 Society for Industrial and Applied Mathematics

INEXACT NEWTON DOGLEG METHODS*

ROGER P. PAWLOWSKI†, JOSEPH P. SIMONIS‡, HOMER F. WALKER‡, AND JOHN N. SHADID†

Abstract. The dogleg method is a classical trust-region technique for globalizing Newton's method. While it is widely used in optimization, including large-scale optimization via truncated-Newton approaches, its implementation in general inexact Newton methods for systems of nonlinear equations can be problematic. In this paper, we first outline a very general dogleg method suitable for the general inexact Newton context and provide a global convergence analysis for it. We then discuss certain issues that may arise with the standard dogleg implementational strategy and propose modified strategies that address them. Newton–Krylov methods have provided important motivation for this work, and we conclude with a report on numerical experiments involving a Newton–GMRES dogleg method applied to benchmark CFD problems.

Key words. dogleg methods, trust-region methods, inexact Newton methods, Newton's method, Newton iterative methods, Newton–Krylov methods, Newton–GMRES methods, truncated-Newton methods, globalized Newton methods, fully-coupled solution methods, Navier–Stokes equations

AMS subject classifications. 65H10, 65F10

DOI. 10.1137/050632166

1. Introduction.

The problem of interest is a system of nonlinear equations

$$F(x) = 0, \tag{1.1}$$

where $F : \mathbb{R}^n \to \mathbb{R}^n$ is continuously differentiable. A classical algorithm for solving (1.1) is *Newton's method*, which generates a sequence of iterates $\{x_k\}$ through steps $s_k = x_{k+1} - x_k$ that satisfy the linear *Newton equation*

$$F'(x_k)s_k = -F(x_k). \tag{1.2}$$

An *inexact Newton method*, defined in [9], is an extension of Newton's method having the following basic form:

Algorithm IN. INEXACT NEWTON METHOD [9]

Let x_0 be given.
For $k = 0, 1, \ldots$ (until convergence) do:
  Choose $\eta_k \in [0, 1)$ and s_k^{IN} such that

$$\|F(x_k) + F'(x_k)\, s_k^{\mathrm{IN}}\| \le \eta_k \|F(x_k)\|. \tag{1.3}$$

  Set $x_{k+1} = x_k + s_k^{\mathrm{IN}}$.

*Received by the editors May 24, 2005; accepted for publication (in revised form) August 2, 2007; published electronically May 2, 2008.
http://www.siam.org/journals/sinum/46-4/63216.html

†Department of Computational Sciences, Sandia National Laboratories, MS 0316, P.O. Box 5800, Albuquerque, NM 87185-0316 (rppawlo@sandia.gov, jnshadi@sandia.gov). The work of these authors was partially supported by the US DOE ASC program and the US DOE Office of Science MICS program at Sandia National Laboratories under contract DE-AC04-94AL85000.

‡Department of Mathematical Sciences, Worcester Polytechnic Institute, Worcester, MA 01609-2280 (jpsimoni@wpi.edu, walker@wpi.edu). The work of these authors was supported in part by Sandia National Laboratories under the ASC program and in part by the Sandia National Laboratories Computer Science Research Institute (contract 16099 with WPI). The third author's work was also supported in part by NSF grant DMS-0540684 with Worcester Polytechnic Institute and by the Center for Simulation of Accidental Fires and Explosions funded at the University of Utah by the U.S. Department of Energy under contracts LLNL B341493 and B524196.

The *inexact Newton condition* (1.3) is a relaxation of (1.2) that expresses the defining property of an inexact Newton method. The usual point of view is that, for each k, the *inexact Newton step* s_k^{IN} is chosen to reduce the norm of $F(x_k) + F'(x_k)s$, the *local linear model* of F at x_k, to an extent specified by η_k. Viewed in this way, η_k is often called a *forcing term* (cf. [13]).

While there is no restriction in Algorithm IN on how s_k^{IN} satisfying (1.3) is determined, inexact Newton methods are often implemented as *Newton iterative methods*, in which s_k^{IN} is found by applying an iterative linear algebra method to (1.2) until (1.3) holds for a specified η_k. For most large-scale applications, the most robust and efficient such methods are *Newton–Krylov methods*, in which the iterative linear algebra methods are *Krylov subspace methods*.[1] Newton–Krylov methods, which have provided considerable motivation for this paper, can enable the efficient solution of (1.2) on large-scale distributed architectures and may also allow "matrix-free" implementations (see [25]) that do not require the evaluation or storage of F'. The term "Newton–Krylov" appears to have originated in [3], but these methods date back at least to the *truncated Newton method* for optimization introduced in [10] (see also [28] and [46], which anticipated some algorithmic features), in which the Krylov subspace method is the preconditioned conjugate-gradient (PCG) method or its Lanczos variant. A Newton–Krylov method that uses a specific Krylov subspace method is often designated by appending the name of the Krylov solver to "Newton," as in the *Newton–GMRES method*, which uses the generalized minimal residual (GMRES) method [38]. For recent general references on Newton–Krylov methods, see [24], [25], and [30].

Inexact Newton methods, like all Newton-like methods, must usually be *globalized*, i.e., augmented with certain auxiliary procedures (*globalizations*) that increase the likelihood of convergence to a solution when good initial approximate solutions are not available. Globalizations are typically constructed to test whether a trial step gives satisfactory progress toward a solution and, if necessary, to modify it in some way to obtain a step that does give satisfactory progress. There are two major categories of globalizations: *backtracking* (*linesearch*, *damping*) methods, in which step lengths are adjusted (usually shortened) to obtain satisfactory steps; and *trust-region* methods, in which a step is ideally chosen to minimize the norm of the local linear model of F within a specified "trust region."[2] (More specifically, the trust-region step from an approximate solution x_k is ideally $\arg\min_{\|s\|\le\delta}\|F(x_k) + F'(x_k)s\|$, where $\delta > 0$ is a given *trust-region radius* and $\|\cdot\|$ is a norm of interest.) Both backtracking and trust-region methods have strong theoretical support; see, e.g., [11], [7], and [12] and the references therein. Backtracking methods are relatively easy to implement; however, each step direction is restricted to be that of the initial trial step. While this step is normally constructed to be a descent direction for $\|F\|$, and is always a descent direction for $\|F\|$ if the step satisfies (1.3) [4, Prop. 3.3], [12, Lem. 7.1], it may be only a weak descent direction, especially if the Jacobian is ill-conditioned [42]. Since trust-region steps are increasingly nearly in the steepest-descent direction as the trust-region radius decreases, trust-region methods have the potential advantage of producing modified steps that may be stronger descent directions than the initial trial step; however, their implementation in practical methods may be problematic.

[1] An introduction to Krylov subspace methods is beyond the scope of this paper; we refer the reader to the surveys [16] and [20] and the books [19], [37], and [48].

[2] See [11, Ch. 6] for a general discussion of classical globalizations and [7] for a more recent and very extensive treatment of trust-region methods.

A fundamental issue is that it is often not feasible to compute the ideal trust-region step with much accuracy (see, e.g., [11, Ch. 6] and [7, Ch. 7]), and various ways of approximating it have been devised.

The *dogleg method* [33], [32] provides perhaps the most popular way of approximating the trust-region step. In the traditional dogleg method for solving (1.1), given an approximate solution x_k, one defines the *dogleg curve* Γ_k^{DL} to be the piecewise-linear curve joining 0, the *Cauchy point* s_k^{CP} (defined to be the minimizer of $\|F(x_k) + F'(x_k)s\|$ in the steepest-descent direction, provided $\|\cdot\|$ is an inner-product norm),[3] and the Newton step $s_k^{\mathrm{N}} \equiv -F'(x_k)^{-1}F(x_k)$. Then, given a trust-region radius $\delta > 0$, the dogleg step s_k is defined as follows: If $\|s_k^{\mathrm{N}}\| \leq \delta$, then $s_k = s_k^{\mathrm{N}}$; otherwise, s_k is chosen on Γ_k^{DL} such that $\|s_k\| = \delta$. It can be shown (see, e.g., [11, Ch. 6]) that, as a point s traverses Γ_k^{DL} from 0 to s_k^{N}, $\|s\|$ is monotone increasing and $\|F(x_k) + F'(x_k)s\|$ is monotone decreasing. Thus, the dogleg step s_k is uniquely defined, and $s_k = \arg\min_{s \in \Gamma_k^{\mathrm{DL}},\ \|s\| \leq \delta} \|F(x_k) + F'(x_k)s\|$. Moreover, if δ is sufficiently small (specifically, if $\delta \leq \|s_k^{\mathrm{CP}}\|$), then s_k is a short step in the steepest-descent direction.

One can extend the dogleg method to the inexact Newton context in a straightforward way by substituting s_k^{IN} for s_k^{N} in the definition of Γ_k^{DL}. The resulting dogleg curve retains the appealing property of offering a continuum of steps ranging from s_k^{IN} to short steps in the steepest-descent direction. However, some desirable properties are lost. First, for any $\eta_k > 0$, no matter how small, the local linear model norm may not decrease monotonically along Γ_k^{DL} between s_k^{CP} and s_k^{IN}. For example, this will be the case if s_k^{N} is nearly on and interior to the line segment joining s_k^{CP} and s_k^{IN}, as illustrated on the left in Figure 1.1. Note that, in this illustration, the point on Γ_k^{DL} having norm δ does not minimize the local linear model norm along Γ_k^{DL} within the trust region. Second, unless η_k is sufficiently small, $\|s\|$ may not increase monotonically as s traverses Γ_k^{DL} from s_k^{CP} to s_k^{IN}; as a consequence, Γ_k^{DL} may have more than one point of intersection with the trust-region boundary, as illustrated on the right in Figure 1.1. (If, in this illustration, s_k^{IN} were moved to the right along the same contour to a point slightly outside the trust-region boundary, then Γ_k^{DL} would have *three* points of intersection with the trust-region boundary.) Thus, with this straightforward extension, the dogleg step cannot be reliably characterized as the unique point on Γ_k^{DL} having norm δ; neither can we always expect a point on Γ_k^{DL} having norm δ to minimize the local linear model norm along Γ_k^{DL} within the trust region.

Remark. These issues do not arise in the case of a truncated Newton method for minimizing $f : \mathbb{R}^n \to \mathbb{R}^1$. In this case, PCG is applied to (1.2) with $F(x_k) = \nabla f(x_k)$ and $F'(x_k) = \nabla^2 f(x_k)$ until either a specified inexact-Newton condition holds, a PCG iterate falls outside of the trust region, or a nonpositive eigenvalue of $\nabla^2 f(x_k)$ is detected. In this context, the model of interest is the local quadratic model of f, rather than the local linear model of $F = \nabla f$. One can show [44, Th. 2.1], [7, section 7.5] that, in an appropriate norm $\|\cdot\|$, the first PCG iterate is the Cauchy point for the local quadratic model and, with the minimizing properties of the PCG iterates, that the model is monotone decreasing along the dogleg curve joining 0, the Cauchy point, and any subsequent PCG iterate. Moreover, the PCG iterate norms are monotone

[3] Our definition of the Cauchy point follows that in [11, Ch. 6]. In [7], the Cauchy point is defined to be the minimizer in the steepest-descent direction *within the trust region*. In section 2 below, we recall the definition of the steepest-descent direction with respect to an inner-product norm of interest and discuss potential difficulties in evaluating it.

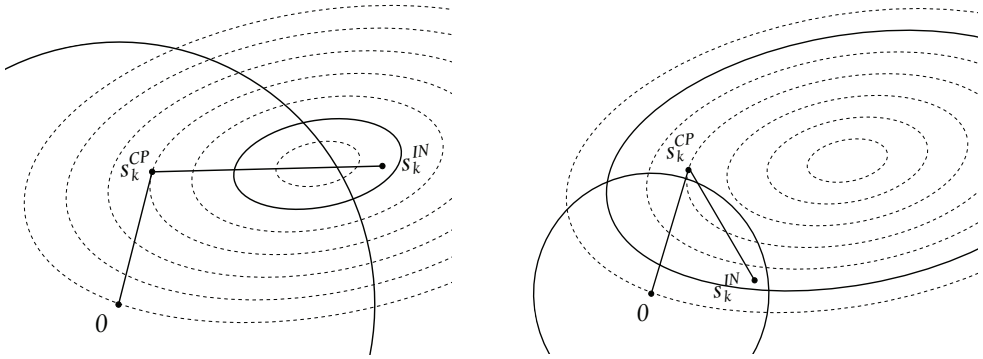


FIG. 1.1. *Illustrative inexact Newton dogleg curves in* $\mathbf{R}^2$. *The ellipses represent level curves of the local linear model norm. The solid ellipses represent* $\{s \in \mathbf{R}^n : \|F(x_k) + F'(x_k)s\| = \eta_k\|F(x_k)\|\}$. *The circular arcs represent trust-region boundaries.*

increasing [44, Th. 2.1], [7, Th. 7.5.1], and it follows that $\|s\|$ is monotone increasing as s traverses the dogleg curve from 0 to the final PCG iterate.

With these issues in mind, our goal in this paper is to develop an understanding of dogleg methods in the general inexact Newton context that will place them on a firm footing, both theoretically and practically. In section 2, we outline a very general inexact Newton dogleg method and provide a convergence analysis for it. Since computing exact Cauchy points may be undesirably expensive or infeasible, as discussed further in section 2, the method allows approximations of these points in addition to inexact Newton steps in defining the dogleg curves.[4] The method also allows a great deal of flexibility in determining steps along the dogleg curves in order to accommodate a variety of strategies for selecting steps. In section 3, we address specific step-selection strategies, noting possible shortcomings of the traditional strategy and suggesting certain alternatives and enhancements. In section 4, we discuss implementational details and report on numerical experiments with Newton–GMRES dogleg methods applied to benchmark problems involving the steady-state Navier–Stokes equations.[5] We conclude with summary observations in section 5.

2. The general method. We assume throughout that $\langle \cdot, \cdot \rangle$ is an inner product of interest and that $\| \cdot \|$ is the associated norm. We use $\langle \cdot, \cdot \rangle_2$ and $\| \cdot \|_2$ to denote the Euclidean inner product and norm, and recall that there is a unique symmetric positive-definite $D \in \mathbb{R}^{n \times n}$ such that $\langle u, v \rangle = \langle u, Dv \rangle_2$ for all $u, v \in \mathbb{R}^n$. (In practice, D is likely to be explicitly available, but this is not assumed here.) We define the merit function

$$f(x) \equiv \frac{1}{2}\|F(x)\|^2$$

[4]The use of approximate Cauchy points is particularly relevant in the context of interest here. In order to keep the focus on the issues of major interest, we do not consider inaccuracy from other sources. See [6] for a general discussion of inaccuracy in inexact Newton methods and [2] and [5] for a treatment of the effects of particular sources of inaccuracy in Jacobian-vector products.

[5]Globalization is usually less critical for time-dependent problems, since time-steps are usually sufficiently small to maintain convergence of inexact Newton iterates in implicit time-stepping methods. However, globalization has been used to advantage in some cases; for example, it has been observed in groundwater flow simulations that globalization appears necessary (and a linesearch globalization is sufficient) to achieve desirably long time steps [49].

and note that f is continuously differentiable since F is, with $\nabla f(x) = F'(x)^T DF(x)$. One has that $\nabla f(x)^T s = \langle \nabla f(x), s\rangle_2 = \langle D^{-1}\nabla f(x), s\rangle$; hence, the steepest-descent direction for f with respect to $\langle\cdot,\cdot\rangle$ is

$$d(x) \equiv -D^{-1}\nabla f(x) = -D^{-1}F'(x)^T DF(x). \tag{2.1}$$

Accurately evaluating $d(x)$ may be problematic in practice for at least two reasons: first, solving systems with D may be undesirable, even if D is available; second, multiplication by $F'(x)^T$ may be infeasible, e.g., in applications of "matrix-free" Newton–Krylov methods in which $F'(x)$ is unavailable. However, useful approximations of $d(x)$ may be available in many circumstances. If $F'(x)^T$-products can be evaluated, then $-\nabla f(x)$ may be an acceptable substitute for $d(x)$; see Remark 2 after Theorem 2.1. If accurate $F'(x)^T$-products are not available, then one may be able to approximate them using approximations of $F'(x)$ that can be readily computed, e.g., by omitting terms in $F'(x)$ that cannot be easily evaluated or that lie outside a desirable sparsity pattern. Additionally, in the Newton–Krylov context, certain approximations of $d(x)$ may be available at little cost, as explained further at the end of this section. Thus, in outlining our general inexact Newton dogleg method, we allow at each iterate x_k an approximate steepest-descent direction $\hat{d}_k \approx d(x_k)$ and prescribe in Theorem 2.1 how good this approximation must be in order to ensure desirable convergence properties of the method.

With each $\hat{d}_k$, we have an approximate Cauchy point

$$\hat{s}_k^{\mathrm{CP}} \equiv \arg\min\{\|F(x_k) + F'(x_k)s\| :\ s = \lambda\hat{d}_k, -\infty < \lambda < \infty\}.$$

(Here, and in the following, we use "ˆ" to designate approximations derived from $\hat{d}_k$.) Then, given an inexact Newton step s_k^{IN}, we define an approximate dogleg curve $\hat{\Gamma}_k^{\mathrm{DL}}$ to be the piecewise-linear curve connecting 0, $\hat{s}_k^{\mathrm{CP}}$, and s_k^{IN}. Our general method allows great latitude in selecting trial steps on this approximate dogleg curve within the trust region, imposing only very mild minimum-length requirements.

Our test for accepting a step is based on a comparison of the actual reduction of $\|F\|$ and the reduction "predicted" by the local linear model, defined at the kth iteration by, respectively,

$$\begin{aligned} ared_k &\equiv \|F(x_k)\| - \|F(x_k + s_k)\|, \\ pred_k &\equiv \|F(x_k)\| - \|F(x_k) + F'(x_k)\, s_k\|. \end{aligned} \tag{2.2}$$

Specifically, our test requires $ared_k \geq t \cdot pred_k > 0$ for each k, where $t \in (0,1)$ is independent of k. Tests of this type have been used by a number of others; see, e.g., [14], [15], [34], and [12] for such tests applied to systems of nonlinear equations and [27], [43], and [7] for analogous tests applied to unconstrained minimization.

The following is our general inexact Newton dogleg method. See section 4 for the specific parameter values and ancillary procedures used in our test implementation.

Algorithm INDL. Inexact newton dogleg method

Let x_0, $\eta_{\max} \in [0,1)$, $t \in (0,1)$, $\theta_{\max} \in (0,1)$, $\delta_{\min} > 0$ and $\delta \geq \delta_{\min}$ be given.
For $k = 0, 1, \ldots$ (until convergence) do:
Choose $\eta_k \in [0, \eta_{\max}]$ and s_k^{IN} such that
$$\|F(x_k) + F'(x_k)\, s_k^{\mathrm{IN}}\| \leq \eta_k \|F(x_k)\|.$$

Determine $\hat{d}_k$ and evaluate $\hat{s}_k^{\mathrm{CP}}$.
Determine $s_k \in \hat{\Gamma}_k^{\mathrm{DL}}$ with $\min\{\delta_{\min}, \|s_k^{\mathrm{IN}}\|\} \le \|s_k\| \le \delta$.
While $ared_k < t \cdot pred_k$ do:
    If $\delta = \delta_{\min}$, stop; else choose $\theta \in (0, \theta_{\max}]$.
    Update $\delta \leftarrow \max\{\theta\delta, \delta_{\min}\}$.
    Redetermine $s_k \in \hat{\Gamma}_k^{\mathrm{DL}}$ with $\min\{\delta_{\min}, \|s_k^{\mathrm{IN}}\|\} \le \|s_k\| \le \delta$.
Set $x_{k+1} = x_k + s_k$ and update δ in $[\delta_{\min}, \infty)$.

As it is stated, Algorithm INDL requires determining both s_k^{IN} and $\hat{s}_k^{\mathrm{CP}}$ before determining the initial $s_k \in \hat{\Gamma}_k^{\mathrm{DL}}$. In a practical implementation, only one of these may be needed; see section 3.1.

The algorithm employs $\delta_{\min} > 0$ that serves as a lower bound on allowable trust-region radii and is also used in determining a lower bound on admissible dogleg step lengths. This $\delta_{\min}$ is used in the proof of Theorem 2.1. Such a $\delta_{\min}$ is likely to be a part of any practical algorithm and, in practice, can be related to a step-length stopping tolerance or taken so small that it does not undesirably restrict choices of the trust-region radii or the dogleg steps.

It is easily verified that the while-loop cannot continue indefinitely; that is, once an initial s_k has been obtained, either an acceptable s_k is determined or the algorithm terminates with $\delta = \delta_{\min}$ after at most a finite number of iterations of the while-loop. For the purposes of Theorem 2.1, there is complete freedom in choosing $\theta \in (0, \theta_{\max}]$. In particular, because of the constrained update $\delta \leftarrow \max\{\theta\delta, \delta_{\min}\}$, it is not necessary to impose a positive lower bound on θ, although one may well do this in practice. (There are a number of practical possibilities for choosing θ; see, in particular, [11] for choices based on minimizing interpolating polynomials over a fixed subinterval $[\theta_{\min}, \theta_{\max}] \subset (0, 1)$.) Similarly, for the purposes of Theorem 2.1, there is complete freedom in updating δ in $[\delta_{\min}, \infty)$ following the while-loop. Of course, one would want to do this judiciously in practice; see, e.g., [11] and [7] for practically effective procedures.

Our global convergence result for Algorithm INDL is given in Theorem 2.1 below. For this, we recall that $x \in \mathbb{R}^n$ is a *stationary point* of $\|F\|$ if $\|F(x)\| \le \|F(x) + F'(x)\, s\|$ for every $s \in \mathbb{R}^n$. In the present context, x is a stationary point if and only if $\nabla f(x) = 0$.

THEOREM 2.1. *Assume that F is continuously differentiable. Suppose that $\{x_k\}$ is produced by Algorithm* INDL *and that, for d defined in* (2.1) *and some $\epsilon > 0$,*

$$\frac{\langle \hat{d}_k, d(x_k)\rangle}{\|\hat{d}_k\|\|d(x_k)\|} \ge \epsilon \tag{2.3}$$

for every k. If x_* is a limit point of $\{x_k\}$, then x_* is a stationary point of $\|F\|$. If, additionally, $F'(x_*)$ is nonsingular, then $F(x_*) = 0$ and $x_k \to x_*$. Moreover, for all sufficiently large k, the initial s_k is accepted without modification in the while-loop, and $s_k = s_k^{\mathrm{IN}}$ is an admissible step.

Remark 1. In Theorem 2.1, we implicitly assume that $F(x_k)$, $\hat{d}_k$, and $d(x_k)$ are nonzero for every k.

Remark 2. Inequality (2.3) states that the vectors $\hat{d}_k$ are uniformly bounded away from orthogonality with the vectors $d(x_k)$. Note that

$$\frac{\langle -\nabla f(x_k), d(x_k)\rangle}{\|-\nabla f(x_k)\|\|d(x_k)\|} = \frac{\|\nabla f(x_k)\|_2^2}{\langle \nabla f(x_k), D\nabla f(x_k)\rangle_2^{1/2} \langle \nabla f(x_k), D^{-1}\nabla f(x_k)\rangle_2^{1/2}} \ge \epsilon$$

for $0 < \epsilon \le \kappa_2(D)^{-1/2}$, where $\kappa_2(D) = \|D\|_2\|D^{-1}\|_2$. Thus (2.3) holds for some $\epsilon > 0$ if $\hat{d}_k$ is $-\nabla f(x_k)$ or a sufficiently accurate approximation of it.

Remark 3. If $x_k \to x_*$ with $F(x_*) = 0$ and $F'(x_*)$ nonsingular and if $s_k = s_k^{\mathrm{IN}}$ for all sufficiently large k, then the convergence is ultimately governed by the choices of the forcing terms η_k as in the local convergence theory of [9] and [13].

Proof. Let $\{x_k\}$ be produced by Algorithm INDL and suppose that (2.3) holds for every k. Since $\langle F'(x_k)\hat{d}_k, F(x_k)\rangle = -\langle \hat{d}_k, d(x_k)\rangle \neq 0$ by (2.3), we have that $F'(x_k)\hat{d}_k \neq 0$ for every k. Then a straightforward calculation yields

$$\|\hat{s}_k^{\mathrm{CP}}\| = \frac{|\langle F(x_k), F'(x_k)\hat{d}_k\rangle|}{\|F'(x_k)\hat{d}_k\|^2}\|\hat{d}_k\|, \tag{2.4}$$

$$\hat{\eta}_k^{\mathrm{CP}} \equiv \frac{\|F(x_k) + F'(x_k)\hat{s}_k^{\mathrm{CP}}\|}{\|F(x_k)\|} = \sqrt{1 - \frac{\langle F(x_k), F'(x_k)\hat{d}_k\rangle^2}{\|F(x_k)\|^2\|F'(x_k)\hat{d}_k\|^2}}. \tag{2.5}$$

Our plan is to build on (2.4)–(2.5) to obtain upper bounds on $\|\hat{s}_k^{\mathrm{CP}}\|$ and $\hat{\eta}_k^{\mathrm{CP}}$ at points of interest that will allow us to use results from [12], which are stated as Theorem 2.2 following the proof. Equation (2.4) immediately gives

$$\|\hat{s}_k^{\mathrm{CP}}\| \le \frac{\|F(x_k)\|\,\|\hat{d}_k\|}{\|F'(x_k)\hat{d}_k\|}\,. \tag{2.6}$$

Also, from (2.3), we obtain

$$\begin{aligned}\epsilon \le \frac{\langle \hat{d}_k, d(x_k)\rangle}{\|\hat{d}_k\|\,\|d(x_k)\|} &= -\frac{\langle \hat{d}_k, F'(x_k)^T DF(x_k)\rangle_2}{\|\hat{d}_k\|\,\|d(x_k)\|} \\ &= -\frac{\langle F'(x_k)\hat{d}_k, F(x_k)\rangle}{\|\hat{d}_k\|\,\|d(x_k)\|} \le \frac{\|F'(x_k)\hat{d}_k\|}{\|\hat{d}_k\|}\cdot\frac{\|F(x_k)\|}{\|d(x_k)\|}\,,\end{aligned}$$

whence

$$\frac{\|\hat{d}_k\|}{\|F'(x_k)\hat{d}_k\|} \le \frac{\|F(x_k)\|}{\epsilon\|d(x_k)\|}\,.$$

Then (2.6) yields

$$\|\hat{s}_k^{\mathrm{CP}}\| \le \frac{\|F(x_k)\|^2}{\epsilon\|d(x_k)\|}\,. \tag{2.7}$$

From (2.3), we also have

$$\begin{aligned}-\frac{\langle F(x_k), F'(x_k)\hat{d}_k\rangle}{\|F(x_k)\|\,\|F'(x_k)\hat{d}_k\|} &= \frac{\langle \hat{d}_k, d(x_k)\rangle}{\|\hat{d}_k\|\,\|d(x_k)\|}\cdot\frac{\|\hat{d}_k\|\,\|d(x_k)\|}{\|F(x_k)\|\,\|F'(x_k)\hat{d}_k\|} \\ &\ge \epsilon\cdot\frac{\|\hat{d}_k\|}{\|F'(x_k)\hat{d}_k\|}\cdot\frac{\|d(x_k)\|}{\|F(x_k)\|} \ge \frac{\epsilon}{\|F'(x_k)\|}\cdot\frac{\|d(x_k)\|}{\|F(x_k)\|}\,,\end{aligned}$$

and it follows from (2.5) that

$$\hat{\eta}_k^{\mathrm{CP}} \le \sqrt{1 - \frac{\epsilon^2}{\|F'(x_k)\|^2}\cdot\frac{\|d(x_k)\|^2}{\|F(x_k)\|^2}}\,. \tag{2.8}$$

Let x_* be a limit point of $\{x_k\}$, and suppose that x_* is not a stationary point of $\|F\|$. Then $F(x_*) \neq 0$. Additionally, $d(x_*) = -D^{-1}\nabla f(x_*) \neq 0$ since x_* is not a stationary point of $\|F\|$. Then, by continuity, there is a neighborhood N_* of x_* such that $0 < \inf_{x\in N_*} \|F(x)\| \leq \sup_{x\in N_*} \|F(x)\| < \infty$, $0 < \inf_{x\in N_*} \|d(x)\|$, and $0 < \inf_{x\in N_*} \|F'(x)\| \leq \sup_{x\in N_*} \|F'(x)\| < \infty$. Then one sees from (2.7)–(2.8) that there are bounds M and $\eta_{\max}^{\mathrm{CP}} < 1$ such that $\|\hat{s}_k^{\mathrm{CP}}\| \leq M$ and $\hat{\eta}_k^{\mathrm{CP}} \leq \eta_{\max}^{\mathrm{CP}}$ whenever $x_k \in N_*$.

Suppose that $x_k \in N_*$ and s_k is determined by Algorithm INDL. If s_k lies on $\hat{\Gamma}_k^{\mathrm{DL}}$ between $\hat{s}_k^{\mathrm{CP}}$ and s_k^{IN}, then it follows from norm convexity that

$$\|F(x_k) + F'(x_k)\, s_k\| \leq \max\{\hat{\eta}_k^{\mathrm{CP}}, \eta_k\}\|F(x_k)\| \leq \max\{\eta_{\max}^{\mathrm{CP}}, \eta_{\max}\}\|F(x_k)\|. \tag{2.9}$$

If s_k lies on $\hat{\Gamma}_k^{\mathrm{DL}}$ between 0 and $\hat{s}_k^{\mathrm{CP}}$, then $\min\{\delta_{\min}, \|s_k^{\mathrm{IN}}\|\} \leq \|s_k\| \leq \|\hat{s}_k^{\mathrm{CP}}\|$. We have

$$\|F(x_k)\| - \|F'(x_k)s_k^{\mathrm{IN}}\| \leq \|F(x_k) + F'(x_k)s_k^{\mathrm{IN}}\| \leq \eta_k\|F(x_k)\| \leq \eta_{\max}\|F(x_k)\|,$$

which implies

$$\|s_k^{\mathrm{IN}}\| \geq \frac{(1-\eta_{\max})}{\|F'(x_k)\|}\|F(x_k)\|. \tag{2.10}$$

The right-hand side of (2.10) is bounded away from zero for $x_k \in N_*$; thus, there is a $\bar{\delta} > 0$ such that $\bar{\delta} \leq \min\{\delta_{\min}, \|s_k^{\mathrm{IN}}\|\}$ whenever $x_k \in N_*$. Since the local linear model norm is monotone decreasing along the segment of $\hat{\Gamma}_k^{\mathrm{DL}}$ between 0 and $\hat{s}_k^{\mathrm{CP}}$, we have, again using norm convexity, that

$$\begin{aligned}
\|F(x_k) + F'(x_k)\, s_k\| &\leq \left\|F(x_k) + F'(x_k)\, \left(\frac{\bar{\delta}}{\|\hat{s}_k^{\mathrm{CP}}\|}\hat{s}_k^{\mathrm{CP}}\right)\right\| \\
&\leq \left(1 - \frac{\bar{\delta}}{\|\hat{s}_k^{\mathrm{CP}}\|}\right)\|F(x_k)\| + \frac{\bar{\delta}}{\|\hat{s}_k^{\mathrm{CP}}\|}\|F(x_k) + F'(x_k)\,\hat{s}_k^{\mathrm{CP}}\| \\
&\leq \left[1 - \frac{\bar{\delta}}{\|\hat{s}_k^{\mathrm{CP}}\|}(1-\hat{\eta}_k^{\mathrm{CP}})\right]\|F(x_k)\| \\
&\leq \left[1 - \frac{\bar{\delta}}{M}(1-\eta_{\max}^{\mathrm{CP}})\right]\|F(x_k)\|.
\end{aligned} \tag{2.11}$$

From (2.9) and (2.11), one concludes that whenever $x_k \in N_*$, we have $\|F(x_k) + F'(x_k)\, s_k\| \leq \bar{\eta}\|F(x_k)\|$, where

$$\bar{\eta} \equiv \max\left\{\eta_{\max}^{\mathrm{CP}}, \eta_{\max}, 1 - \frac{\bar{\delta}}{M}(1-\eta_{\max}^{\mathrm{CP}})\right\} < 1;$$

therefore, with $pred_k$ defined as in (2.2),

$$\frac{pred_k}{\|F(x_k)\|} = \frac{\|F(x_k)\| - \|F(x_k) + F'(x_k)\, s_k\|}{\|F(x_k)\|} \geq (1-\bar{\eta}) > 0. \tag{2.12}$$

Since $x_k \in N_*$ for infinitely many values of k, (2.12) implies that $\sum_{k=0}^{\infty} pred_k/\|F(x_k)\|$ diverges, and it follows from Theorem 2.2 that $F(x_*) = 0$. This is a contradiction; therefore, x_* must be a stationary point of $\|F\|$.

Suppose now that x_* is a limit point of $\{x_k\}$ and that $F'(x_*)$ is nonsingular. Then x_* must be a stationary point of $\|F\|$, and, since $F'(x_*)$ is nonsingular, we must have

$F(x_*) = 0$. Additionally, we can let N_* be a neighborhood of x_* such that $F'(x)$ is invertible for $x \in N_*$ and $\sup_{x\in N_*}\{\|F(x)\|, \|F'(x)\|, \|F'(x)^{-1}\|\} < \infty$. Noting that

$$\frac{\|F(x_k)\|}{\|d(x_k)\|} \leq \| \left[D^{-1}F'(x_k)^T D\right]^{-1} \|, \tag{2.13}$$

we have from (2.7) that, for $x_k \in N_*$,

$$\|\hat{s}_k^{\mathrm{CP}}\| \leq \frac{1}{\epsilon} \| \left[D^{-1}F'(x_k)^T D\right]^{-1} \| \, \|F(x_k)\|. \tag{2.14}$$

One concludes that there is again a bound M such that $\|\hat{s}_k^{\mathrm{CP}}\| \leq M$ whenever $x_k \in N_*$. Moreover, it follows from (2.13) and (2.8) that, for $x_k \in N_*$,

$$\hat{\eta}_k^{\mathrm{CP}} \leq \sqrt{1 - \frac{\epsilon^2}{\|F'(x_k)\|^2 \, \| \left[D^{-1}F'(x_k)^T D\right]^{-1} \|^2}} \, .$$

Consequently, there is again a bound $\eta_{\max}^{\mathrm{CP}} < 1$ such that $\hat{\eta}_k^{\mathrm{CP}} \leq \eta_{\max}^{\mathrm{CP}}$ whenever $x_k \in N_*$.

Suppose that $x_k \in N_*$ and s_k is determined by Algorithm INDL. If s_k lies on $\hat{\Gamma}_k^{\mathrm{DL}}$ between $\hat{s}_k^{\mathrm{CP}}$ and s_k^{IN}, then (2.9) holds as before. If s_k lies on $\hat{\Gamma}_k^{\mathrm{DL}}$ between 0 and $\hat{s}_k^{\mathrm{CP}}$, then we again have $\min\{\delta_{\min}, \|s_k^{\mathrm{IN}}\|\} \leq \|s_k\| \leq \|\hat{s}_k^{\mathrm{CP}}\|$. Reasoning as before, we have that

$$\begin{aligned}
\|F(x_k) + F'(x_k)\, s_k\| \;&\leq\; \left\| F(x_k) + F'(x_k) \left(\frac{\min\{\delta_{\min}, \|s_k^{\mathrm{IN}}\|\}}{\|\hat{s}_k^{\mathrm{CP}}\|} \hat{s}_k^{\mathrm{CP}} \right) \right\| \\
&\leq\; \left(1 - \frac{\min\{\delta_{\min}, \|s_k^{\mathrm{IN}}\|\}}{\|\hat{s}_k^{\mathrm{CP}}\|} \right) \|F(x_k)\| \\
&\qquad + \frac{\min\{\delta_{\min}, \|s_k^{\mathrm{IN}}\|\}}{\|\hat{s}_k^{\mathrm{CP}}\|} \|F(x_k) + F'(x_k)\, \hat{s}_k^{\mathrm{CP}}\| \\
&\leq\; \left[1 - \frac{\min\{\delta_{\min}, \|s_k^{\mathrm{IN}}\|\}}{\|\hat{s}_k^{\mathrm{CP}}\|} (1 - \hat{\eta}_k^{\mathrm{CP}}) \right] \|F(x_k)\| \\
&\leq\; \left[1 - \min\left\{ \frac{\delta_{\min}}{M}, \frac{\|s_k^{\mathrm{IN}}\|}{\|\hat{s}_k^{\mathrm{CP}}\|} \right\} (1 - \eta_{\max}^{\mathrm{CP}}) \right] \|F(x_k)\|.
\end{aligned} \tag{2.15}$$

Inequality (2.10) again holds, and (2.10) and (2.14) yield

$$\frac{\|s_k^{\mathrm{IN}}\|}{\|\hat{s}_k^{\mathrm{CP}}\|} \geq \frac{\epsilon(1-\eta_{\max})}{\|F'(x_k)\| \, \| \left[D^{-1}F'(x_k)^T D\right]^{-1} \|} \geq \inf_{x\in N_*} \frac{\epsilon(1-\eta_{\max})}{\|F'(x)\| \, \| \left[D^{-1}F'(x)^T D\right]^{-1} \|} > 0.$$

It follows that there is a $\bar{\delta} > 0$ independent of k such that

$$\bar{\delta} \leq \min\left\{ \frac{\delta_{\min}}{M}, \frac{\|s_k^{\mathrm{IN}}\|}{\|\hat{s}_k^{\mathrm{CP}}\|} \right\},$$

and we have from (2.15) that

$$\|F(x_k) + F'(x_k)\, s_k\| \leq \left[1 - \bar{\delta}(1 - \eta_{\max}^{\mathrm{CP}})\right] \|F(x_k)\|.$$

As before, one concludes that (2.12) holds whenever $x_k \in N_*$, with $\bar{\eta}$ now given by

$$\bar{\eta} \equiv \max\{\eta_{\max}^{\mathrm{CP}}, \eta_{\max}, 1 - \bar{\delta}(1 - \eta_{\max}^{\mathrm{CP}})\} < 1,$$

and we again have that $\sum_{k=0}^{\infty} pred_k/\|F(x_k)\|$ diverges. Since $F'(x_*)$ is nonsingular, it follows from Theorem 2.2 below that $x_k \to x_*$.

To complete the proof, we note that, since $x_k \to x_*$ as $k \to \infty$ with $F(x_*) = 0$ and $F'(x_*)$ nonsingular, it follows from (2.14) that $\hat{s}_k^{\mathrm{CP}} \to 0$ as $k \to \infty$. Additionally, writing $F'(x_k)\, s_k = -F(x_k) + r_k$ with $\|r_k\| \le \eta_k \|F(x_k)\|$, we have that

$$\begin{aligned} \|s_k^{\mathrm{IN}}\| &= \|F'(x_k)^{-1}[-F(x_k) + r_k]\| \le \|F'(x_k)^{-1}\|\,(1 + \eta_k)\,\|F(x_k)\| \\ &\le 2\,\|F'(x_k)^{-1}\|\,\|F(x_k)\|. \end{aligned}$$

Consequently, $s_k^{\mathrm{IN}} \to 0$ as $k \to \infty$. One easily verifies that, for all sufficiently large k, every admissible $s_k \in \hat{\Gamma}_k^{\mathrm{DL}}$ must be sufficiently small that $ared_k \ge t \cdot pred_k$ holds; thus, for all sufficiently large k, the initial s_k is accepted without modification. Finally, for all sufficiently large k, one has $\|s_k^{\mathrm{IN}}\| \le \delta_{\min} \le \delta$, and $s_k = s_k^{\mathrm{IN}}$ is admissible. ∎

The following is the result from [12] used in the proof of Theorem 2.1.

THEOREM 2.2 (see [12, Cor. 3.6]). *Let $F : \mathbb{R}^n \to \mathbb{R}^n$ be continuously differentiable and assume that $\{x_k\} \subset \mathbb{R}^n$ is such that $pred_k \ge 0$ and $ared_k \ge t \cdot pred_k$ for each k, where $t \in (0, 1)$ is independent of k and $ared_k$ and $pred_k$ are given by (2.2) with $s_k \equiv x_{k+1} - x_k$. If $\sum_{k \ge 0} pred_k/\|F(x_k)\|$ is divergent, then $F(x_k) \to 0$. If, in addition, x_* is a limit point of $\{x_k\}$ such that $F'(x_*)$ is invertible, then $F(x_*) = 0$ and $x_k \to x_*$.*

Applications to Newton–GMRES and Newton–Arnoldi methods. It follows from observations in [3, section 4] that if either GMRES or the Arnoldi method [35] is used (without restarting) to solve (1.2), then the orthogonal projection of $d(x_k)$ determined by (2.1) onto each Krylov subspace generated by the method is available at very little cost. Using results from [4], we show how Theorem 2.1 can be applied when these projections are used in Algorithm INDL to approximate the vectors $d(x_k)$, provided that the condition numbers $\kappa(F'(x_k)) \equiv \|F'(x_k)\|\,\|F'(x_k)^{-1}\|$ are bounded uniformly in k. In the spirit of [4], we show this for orthogonal projections onto subspaces that are not necessarily Krylov subspaces.

Suppose that $x \in \mathbb{R}^n$, $\mathcal{K}$ is a subspace of $\mathbb{R}^n$, and $s \in \mathcal{K}$ is such that $\|F(x) + F'(x)s\| \le \eta \|F(x)\|$ for some $\eta \in [0, 1)$. Let $\hat{d}$ denote the orthogonal projection onto $\mathcal{K}$ of $d(x)$ given by (2.1). A straightforward extension of [4, Cor. 3.5] to the context of a general inner-product norm gives

$$\|\hat{d}\| \ge \frac{1 - \eta}{(1 + \eta)\kappa(F'(x))} \|d(x)\|.$$

Since $\langle \hat{d}, d(x) \rangle = \|\hat{d}\|^2$, it follows that

$$\frac{|\langle \hat{d}, d(x) \rangle|}{\|\hat{d}\|\,\|d(x)\|} = \frac{\|\hat{d}\|}{\|d\|} \ge \frac{1 - \eta}{(1 + \eta)\kappa(F'(x))}. \tag{2.16}$$

This immediately gives the following corollary of Theorem 2.1.

COROLLARY 2.3. *Assume that F is continuously differentiable and that $\{x_k\}$ is produced by Algorithm* INDL. *For each k, suppose that there is a subspace $\mathcal{K}_k \subseteq \mathbb{R}^n$*

such that $s_k^{\mathrm{IN}} \in \mathcal{K}_k$*, and let* $\hat{d}_k$ *denote the orthogonal projection onto* $\mathcal{K}_k$ *of* $d(x_k)$ *determined by* (2.1)*. If* $\{\kappa(F'(x_k))\}$ *is bounded and if* x_* *is a limit point of* $\{x_k\}$*, then* $F(x_*) = 0$ *and* $x_k \to x_*$*. Moreover, for all sufficiently large* k*, the initial* s_k *is accepted without modification in the while-loop, and* $s_k = s_k^{\mathrm{IN}}$ *is an admissible step.*

Proof. Let M be such that $\kappa(F'(x_k)) \le M$ for all k. Then, since $\eta_k \in [0, \eta_{\max}]$, (2.16) implies that

$$\frac{|\langle \hat{d}_k, d(x_k)\rangle|}{\|\hat{d}_k\|\,\|d(x_k)\|} \ge \frac{1-\eta_{\max}}{(1+\eta_{\max})M}$$

for each k, and (2.3) holds with $\epsilon = \frac{1-\eta_{\max}}{(1+\eta_{\max})M}$. The result follows from Theorem 2.1. ∎

3. Step selection strategies. Algorithm INDL allows great flexibility in determining each step $s_k \in \hat{\Gamma}_k^{\mathrm{DL}}$, and particular choices of the steps may strongly affect the behavior of the algorithm. In this section, we discuss specific strategies for selecting admissible steps, first recalling the traditional strategy and then suggesting an alternative that may have advantages. We conclude by outlining certain refinements with which these strategies can be augmented.

3.1. Two strategies. The traditional strategy outlined in the following is a straightforward adaptation of the usual procedure for determining dogleg steps in the exact-Newton context (see, e.g., [11]).

Procedure 3.1. Traditional strategy

If $\|s_k^{\mathrm{IN}}\| \le \delta$, set $s_k = s_k^{\mathrm{IN}}$.
Else,
  If $\|\hat{s}_k^{\mathrm{CP}}\| \ge \delta$, set $s_k = (\delta/\|\hat{s}_k^{\mathrm{CP}}\|)\hat{s}_k^{\mathrm{CP}}$.
  Else, set $s_k = (1-\gamma)\hat{s}_k^{\mathrm{CP}} + \gamma s_k^{\mathrm{IN}}$ for $\gamma \in (0,1)$
    such that $\|s_k\| = \delta$.

This procedure always determines s_k uniquely, notwithstanding that, as noted in section 1, the usual exact-Newton characterization of the dogleg step may fail to do so. However, there are still issues that may be of concern. As seen in the left illustration in Figure 1.1, for any $\eta_k > 0$, no matter how small, s_k determined by Procedure 3.1 may not minimize the local linear model norm along $\hat{\Gamma}_k^{\mathrm{DL}}$ within the trust region. Perhaps of greater concern is that, if η_k is not small, then, as seen in the right illustration in Figure 1.1, the procedure may specify $s_k = s_k^{\mathrm{IN}}$ without taking $\hat{s}_k^{\mathrm{CP}}$ into account, even though there are steps in the $\hat{s}_k^{\mathrm{CP}}$-direction, or biased toward it, that give significantly greater reduction of the local linear model norm along $\hat{\Gamma}_k^{\mathrm{DL}}$ within the trust region.

As a step toward addressing these concerns, we offer the following.

Procedure 3.2. Alternative strategy

If $\|\hat{s}_k^{\mathrm{CP}}\| \ge \delta$, set $s_k = (\delta/\|\hat{s}_k^{\mathrm{CP}}\|)\hat{s}_k^{\mathrm{CP}}$.
Else,
  If $\|F(x_k) + F'(x_k)\,\hat{s}_k^{\mathrm{CP}}\| \le \eta_k\|F(x_k)\|$, set $s_k = \hat{s}_k^{\mathrm{CP}}$.
  Else,
    If $\|s_k^{\mathrm{IN}}\| \le \delta$, set $s_k = s_k^{\mathrm{IN}}$.
    Else, set $s_k = (1-\gamma)\hat{s}_k^{\mathrm{CP}} + \gamma s_k^{\mathrm{IN}}$ for $\gamma \in (0,1)$
      such that $\|s_k\| = \delta$.

In Procedure 3.2, $\hat{s}_k^{\mathrm{CP}}$ is examined first. If it lies outside the trust region, then, as in the traditional strategy in the exact-Newton context, s_k is a step in the $\hat{s}_k^{\mathrm{CP}}$-direction and scaled to lie on the trust-region boundary. If $\hat{s}_k^{\mathrm{CP}}$ lies within the trust region and also satisfies the prescribed inexact Newton condition, then $\hat{s}_k^{\mathrm{CP}}$ achieves the desired linear model norm reduction, and we take $s_k = \hat{s}_k^{\mathrm{CP}}$. If $\hat{s}_k^{\mathrm{CP}}$ does not satisfy the inexact Newton condition, then s_k^{IN} is examined, and s_k is taken to be either s_k^{IN}, if it lies within the trust region, or, otherwise, the unique point between $\hat{s}_k^{\mathrm{CP}}$ and s_k^{IN} at which $\hat{\Gamma}_k^{\mathrm{DL}}$ intersects the trust-region boundary, as in the traditional strategy in the exact-Newton context.

Note that, in a practical implementation within Algorithm INDL, Procedure 3.2 would require computing $\hat{s}_k^{\mathrm{CP}}$ for each k but would require computing s_k^{IN} only when $\hat{s}_k^{\mathrm{CP}}$ lies within the trust region and does not satisfy the inexact Newton condition. In contrast, Procedure 3.1 would require computing s_k^{IN} for each k but would require computing $\hat{s}_k^{\mathrm{CP}}$ only when s_k^{IN} lies outside the trust region. Also, in the Newton-iterative context, Procedure 3.2 offers the possibility of using $\hat{s}_k^{\mathrm{CP}}$ as the initial approximate solution in the iterative solver when it is necessary to compute s_k^{IN}, thereby allowing advantage to be taken of the linear model norm reduction already achieved by $\hat{s}_k^{\mathrm{CP}}$.

3.2. Further refinements. As noted in section 1 and seen in the left illustration in Figure 1.1, the local linear model norm may not be monotone decreasing along the dogleg curve and, indeed, may not be minimized at the intersection of the dogleg curve and the trust-region boundary. Note that, in this illustration, this intersection would be the dogleg step determined by both Procedure 3.1 and Procedure 3.2. We reproduce this illustration on the left in Figure 3.1 and include the point s_{min} at which the local linear model norm is minimized along the dogleg curve within the trust region. As we show in the following, minimizers such as s_{min} can be readily computed; thus s_{min} is likely to be the preferred dogleg step.

To show further possibilities, we include the illustration on the right in Figure 3.1, in which the minimizer of the local linear model norm within the trust region along the line joining s_k^{CP} and s_k^{IN} (again denoted by s_{min}) occurs *beyond* the end of the dogleg curve. In this illustration, both Procedure 3.1 and Procedure 3.2 would determine the dogleg step to be s_k^{IN}. (In practice, Procedure 3.1 would do so without computing s_k^{CP}, while Procedure 3.2 would compute both s_k^{CP} and s_k^{IN}.) Since s_{min} satisfies the inexact Newton condition satisfied by s_k^{IN}, we can regard it as an admissible dogleg step simply by replacing s_k^{IN} by s_{min} in determining the dogleg curve. Since this s_{min} can also be readily computed, it is likely to be the preferred dogleg step in this case.

In the following, we formulate refinements of Procedures 3.1 and 3.2 that determine dogleg steps to be local linear model norm minimizers such as those illustrated in Figure 3.1. In these refinements, we modify the procedures to minimize explicitly only along the line joining $\hat{s}_k^{\mathrm{CP}}$ and s_k^{IN} in the final steps, when both $\hat{s}_k^{\mathrm{CP}}$ and s_k^{IN} would have been computed in a practical implementation. Minimizing along the line joining zero and $\hat{s}_k^{\mathrm{CP}}$ is already ensured by the definition of $\hat{s}_k^{\mathrm{CP}}$ and the convexity of the local linear model norm along that line. Additionally, it seems unlikely to be worthwhile to compute either $\hat{s}_k^{\mathrm{CP}}$ or s_k^{IN} only for the purposes of minimization when the procedures do not otherwise require them. For example, in the illustration on the right in Figure 1.1, computing both $\hat{s}_k^{\mathrm{CP}}$ (which is just s_k^{CP} in this case) and s_k^{IN} would, in fact, allow one to determine the minimizer along the entire dogleg curve within the trust region; however, not much would be gained over the scaled step in the s_k^{CP}-direction determined by Procedure 3.2. In any event, it would be easy to further modify the procedures to perform this additional minimization, if desired.

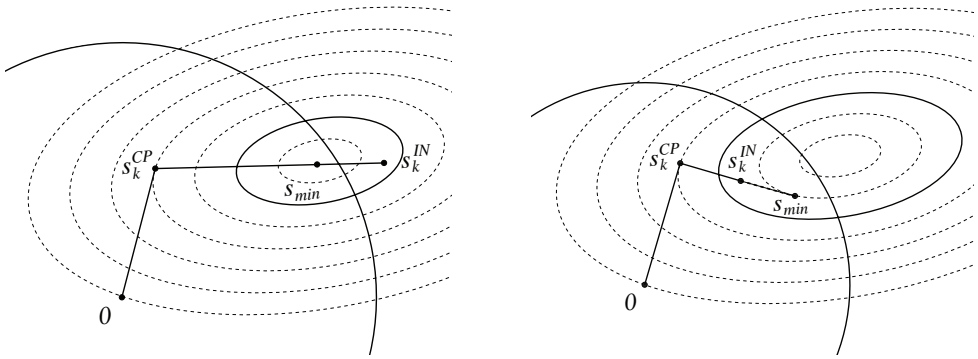


FIG. 3.1. *Illustrative local linear model norm minimizers in* $\boldsymbol{R}^2$. *The ellipses represent level curves of the local linear model norm. The solid ellipses represent* $\{s \in \boldsymbol{R}^n : \|F(x_k) + F'(x_k)s\| = \eta_k\|F(x_k)\|\}$. *The circular arcs represent trust-region boundaries. Each point* s_{min} *is the minimizer of the local linear model norm along the line joining* $s_k^{\rm CP}$ *and* $s_k^{\rm IN}$.

We begin with two elementary propositions, the proofs of which are left to the reader. These are included primarily to show what might be involved in computing minimizing steps. In these, we define

$$s(\gamma) \equiv (1-\gamma)\hat{s}_k^{\rm CP} + \gamma s_k^{\rm IN}.$$

PROPOSITION 3.3. *If* $F'(x_k)(\hat{s}_k^{\rm CP} - s_k^{\rm IN}) \neq 0$, *then* $\|F(x_k) + F'(x_k)s(\gamma)\|$ *is minimized at* $\gamma = \gamma_{\min}$, *given by*

$$\gamma_{\min} \equiv \frac{\langle F(x_k) + F'(x_k)\hat{s}_k^{\rm CP},\ F'(x_k)(\hat{s}_k^{\rm CP} - s_k^{\rm IN})\rangle}{\|F'(x_k)(\hat{s}_k^{\rm CP} - s_k^{\rm IN})\|^2} = \frac{\langle \hat{r}_k^{\rm CP},\ \hat{r}_k^{\rm CP} - r_k^{\rm IN}\rangle}{\|\hat{r}_k^{\rm CP} - r_k^{\rm IN}\|^2}, \tag{3.1}$$

where $\hat{r}_k^{\rm CP} \equiv F(x_k) + F'(x_k)\hat{s}_k^{\rm CP}$ *and* $r_k^{\rm IN} \equiv F(x_k) + F'(x_k)s_k^{\rm IN}$.

One sees from (3.1) that, if $\hat{r}_k^{\rm CP}$ and $r_k^{\rm IN}$ are already available, then $\gamma_{\min}$ can be determined with only a modest amount of arithmetic. If these are not available, then one or, at most, two Jacobian-vector products may be required in addition to some arithmetic.

PROPOSITION 3.4. *If* $\|\hat{s}_k^{\rm CP}\| < \delta$ *and* $s_k^{\rm IN} \neq \hat{s}_k^{\rm CP}$, *then* $\|s(\gamma)\| = \delta$ *for exactly two values* $\gamma = \gamma_\delta^+$ *and* $\gamma = \gamma_\delta^-$, *given by*

$$\gamma_\delta^\pm = \frac{\langle \hat{s}_k^{\rm CP},\ \hat{s}_k^{\rm CP} - s_k^{\rm IN}\rangle \pm \sqrt{\langle \hat{s}_k^{\rm CP},\ \hat{s}_k^{\rm CP} - s_k^{\rm IN}\rangle^2 + (\delta^2 - \|\hat{s}_k^{\rm CP}\|^2)\|\hat{s}_k^{\rm CP} - s_k^{\rm IN}\|^2}}{\|\hat{s}_k^{\rm CP} - s_k^{\rm IN}\|^2}. \tag{3.2}$$

One sees from (3.2) that $\gamma_\delta^\pm$ can be computed with a modest amount of arithmetic. Also, we have that $\gamma_\delta^- < 0 < \gamma_\delta^+$. Additionally, in the context of interest, viz., the final steps of Procedures 3.1 and 3.2, the assumption that $\|\hat{s}_k^{\rm CP}\| < \delta$ and $s_k^{\rm IN} \neq \hat{s}_k^{\rm CP}$ always holds. Indeed, in this context, we have $\|\hat{s}_k^{\rm CP}\| < \delta < \|s_k^{\rm IN}\|$ in Procedure 3.1 and both $\|\hat{s}_k^{\rm CP}\| < \delta$ and $\|F(x_k) + F'(x_k)s_k^{\rm IN}\| \leq \eta_k < \|F(x_k) + F'(x_k)\hat{s}_k^{\rm CP}\|$ in Procedure 3.2. Note that the first pair of inequalities implies that $\gamma_\delta^+ < 1$, and it follows from the final pair that $\gamma_{\min} > 0$.

The following are our refinements of Procedures 3.1 and 3.2. In these, $\gamma_{\min}$ and $\gamma_\delta^\pm$ are defined by (3.1) and (3.2), respectively. In Procedure 3.6, some simplification is allowed in the last step by the fact that $\gamma_{\min} > 0$ there, as noted above.

Procedure 3.5. Traditional strategy with model norm minimization

If $\|s_k^{\mathrm{IN}}\| \le \delta$, set $s_k = s_k^{\mathrm{IN}}$.
Else,
  If $\|\hat{s}_k^{\mathrm{CP}}\| \ge \delta$, set $s_k = (\delta/\|\hat{s}_k^{\mathrm{CP}}\|)\hat{s}_k^{\mathrm{CP}}$.
  Else, set $s_k = (1-\gamma)\hat{s}_k^{\mathrm{CP}} + \gamma s_k^{\mathrm{IN}}$ for $\gamma = \max\{\gamma_\delta^-, \min\{\gamma_{\min}, \gamma_\delta^+\}\}$.

Procedure 3.6. Alternative strategy with model norm minimization

If $\|\hat{s}_k^{\mathrm{CP}}\| \ge \delta$, set $s_k = (\delta/\|\hat{s}_k^{\mathrm{CP}}\|)\hat{s}_k^{\mathrm{CP}}$.
Else,
  If $\|F(x_k) + F'(x_k)\,\hat{s}_k^{\mathrm{CP}}\| \le \eta_k\|F(x_k)\|$, set $s_k = \hat{s}_k^{\mathrm{CP}}$.
  Else, set $s_k = (1-\gamma)\hat{s}_k^{\mathrm{CP}} + \gamma s_k^{\mathrm{IN}}$ for $\gamma = \min\{\gamma_{\min}, \gamma_\delta^+\}$.

Note that, in the circumstance illustrated on the left in Figure 3.1, both Procedure 3.5 and Procedure 3.6 would yield $s_k = s_{min}$. However, in the circumstance illustrated on the right, Procedure 3.6 would produce $s_k = s_{min}$, but Procedure 3.5 would result in $s_k = s_k^{\mathrm{IN}}$ because s_k^{CP} would not be computed in this case.

It is observed in section 3.1 that, in a Newton-iterative implementation of Procedure 3.2, it may be advantageous to use $\hat{s}_k^{\mathrm{CP}}$ as the initial approximate solution in the iterative solver when it is necessary to compute s_k^{IN}. If this is done and if the iterative method is unrestarted GMRES or a mathematically equivalent method, then the minimization of Procedure 3.6 would be unneccesary. Indeed, Procedure 3.2 would produce the same step as Procedure 3.6, since the GMRES step from $\hat{s}_k^{\mathrm{CP}}$ to s_k^{IN} already minimizes the local linear model norm over the Krylov subspace in which it lies and, in particular, minimizes it along the line joining $\hat{s}_k^{\mathrm{CP}}$ and s_k^{IN}.

We also note an additional possibility with Procedures 3.2 and 3.6: If $\|\hat{s}_k^{\mathrm{CP}}\| < \delta$ and $\|F(x_k) + F'(x_k)\,\hat{s}_k^{\mathrm{CP}}\| \le \eta_k\|F(x_k)\|$, then, rather than take $s_k = \hat{s}_k^{\mathrm{CP}}$, one can reduce η_k so that $\eta_k < \|F(x_k) + F'(x_k)\,\hat{s}_k^{\mathrm{CP}}\|$ and compute s_k^{IN} for the reduced η_k. With this s_k^{IN}, the final steps of Procedures 3.2 and 3.6 should yield steps that further reduce the local linear model norm within the trust region.

4. Numerical experiments. In this section, we report on numerical experiments in which a Newton–GMRES implementation of Algorithm INDL employing step-selection strategies from section 3 and certain other options was applied to several well-known benchmark problems involving the steady-state Navier–Stokes equations. In the following, we outline details of the implementation and the overall computational environment, describe the test problems, and give the results of the experiments. These experiments involve challenging large-scale problems and advanced computing platforms but are somewhat limited in scope. Our primary goal is to demonstrate the basic effectiveness of Algorithm INDL with the selected strategies and options on these test problems, and not to provide a definitive comparison of all algorithmic possibilities on a broad variety of applications.

4.1. The implementation and the computational environment. In our implementation of Algorithm INDL, each inexact Newton step s_k^{IN} was determined using restarted GMRES with a restart value of 200. Up to three restarts (600 total GMRES iterations) were allowed for each k. The preconditioner was an additive Schwarz preconditioner with an ILUT [36] factorization and solve on each subdomain. With this preconditioner, GMRES almost never restarted in our experiments. For each k, the initial approximate solution in GMRES was either zero or, optionally, the approximate Cauchy point $\hat{s}_k^{\mathrm{CP}}$ determined by $\hat{d}_k = -\nabla f(x_k)$, which could be

evaluated analytically by our codes and for which condition (2.3) of Theorem 2.1 holds for $0 < \epsilon \le \kappa_2(D)^{-1/2}$ (see Remark 2 following Theorem 2.1). Products of the Jacobian with vectors needed by GMRES were also evaluated analytically.

We used two choices for each forcing term η_k in our experiments: a small constant choice $\eta_k = 10^{-4}$, which should result in fast linear convergence in a neighborhood of a solution; and an adaptive choice from [13], called "Choice 1" there and given by

$$\eta_k = \min\left\{\eta_{\max}, \frac{\left|\|F(x_k)\| - \|F(x_{k-1}) + F'(x_{k-1})\, s_{k-1}\|\right|}{\|F(x_{k-1})\|}\right\},\ k = 1, 2, \ldots, \tag{4.1}$$

which should yield a certain superlinear convergence near a solution [13, Th. 2.2]. As in [13], (4.1) was followed with the safeguard

$$\eta_k \leftarrow \max\{\eta_k, \eta_{k-1}^{(1+\sqrt{5})/2}\} \quad \text{whenever } \eta_{k-1}^{(1+\sqrt{5})/2} > 0.1$$

to prevent the forcing terms from becoming too small too quickly away from a solution. In our implementation, we used $\eta_0 = .01$ and $\eta_{\max} = .9$.

Dogleg steps were selected using the standard and alternative strategies described in Procedures 3.1 and 3.2, respectively. Since, as noted previously, GMRES rarely restarted, Procedure 3.6 almost always produced the same steps as Procedure 3.2 when GMRES was started from the approximate Cauchy point. Thus there seemed to be no significant advantage in using Procedure 3.6 rather than Procedure 3.2 in this case. For consistency, then, we used Procedures 3.1 and 3.2 rather than Procedures 3.5 and 3.6 throughout.

In the test $ared_k < t \cdot pred_k$ at the top of the while-loop, we took $t = 10^{-4}$, so that even very modest agreement would result in accepting the step. Within the while-loop, we used the fixed choice $\theta = .25$ for each update $\delta \leftarrow \max\{\theta\delta, \delta_{\min}\}$. In updating δ in $[\delta_{\min}, \infty)$ following the while-loop, we used a procedure similar to that outlined in [11], in which the trust region is shrunk (subject to the constraint $\delta \ge \delta_{\min}$) if $\|F(x_k + s_k)\|$ and $\|F(x_k) + F'(x_k)\, s_k\|$ do not agree well, expanded if they agree especially well, and left unchanged otherwise. The specific procedure, in which $0 < \rho_s < \rho_e < 1$, $0 < \beta_s < 1 < \beta_e$, and $\delta_{\max} > \delta_{\min}$, is as follows:

Procedure 4.1.

If $ared_k/pred_k < \rho_s$,
  If $\|s_k^{IN}\| < \delta$, update $\delta \leftarrow \max\{\|s_k^{IN}\|, \delta_{\min}\}$.
  Else, update $\delta \leftarrow \max\{\beta_s\delta, \delta_{\min}\}$.
Else,
  If $ared_k/pred_k > \rho_e$ and $\|s_k\| = \delta$, update $\delta \leftarrow \min\{\beta_e\delta, \delta_{\max}\}$.

In our implementation, we took $\rho_s = 0.1$, $\rho_e = 0.75$, $\beta_s = .25$, $\beta_e = 4.0$, $\delta_{\min} = 10^{-6}$, and $\delta_{\max} = 10^{10}$. The initial δ was determined after the computation of s_0^{IN} as follows: If $\|s_0^{IN}\| < \delta_{\min}$, then $\delta = 2\delta_{\min}$; otherwise, $\delta = \|s_0^{IN}\|$.

Our implementation of Algorithm INDL was done within the NOX nonlinear solver package [26], a C++ object-oriented library developed at Sandia National Laboratories for the efficient solution of systems of nonlinear equations. The GMRES implementation and preconditioners were provided by the AztecOO package [22], an extension of the Aztec library [47], which provides an extensive suite of Krylov solvers and preconditioners for the parallel solution of linear systems. The parallel finite-element reacting-flow code MPSalsa [40] was used to set up the finite-element discretization of the test problems described in section 4.2 and also to invoke

the solvers. MPSalsa uses Chaco [21], a general graph partitioning tool, to partition the FE mesh into subdomains and assign subdomains to processors. Successful termination of the nonlinear solution algorithm was declared if both of the following held: (1) $\frac{1}{n}\|Ws_k\| < 1$, where n is the total number of unknowns and W is a diagonal weighting matrix with entries $W_{ii} = 1/(\varepsilon_r|x_k^{(i)}| + \varepsilon_a)$, in which $x_k^{(i)}$ is the ith component of x_k (see [1]); (2) $\|F(x_k)\| \le \epsilon_F\|F(x_0)\|$. In our tests, $\varepsilon_r = 10^{-3}$, $\varepsilon_a = 10^{-8}$, and $\epsilon_F = 10^{-2}$. The first criterion is employed by MPSalsa and, in our experiments, was typically more stringent than the second and necessary to ensure that finer physical details of solutions are adequately resolved.

Our experiments were performed on an IBM cluster at Sandia National Laboratories having 16 nodes, with each node containing two one-GHz Pentium III processors with one GB of RAM each. Tests on the 2D problems were done using four nodes (eight processors); tests on the 3D problems were done using 15 nodes (30 processors).

4.2. The test problems. The three benchmark problems considered here have been widely studied. We describe these in only qualitative terms and refer the reader to [30] and [41] for more specific descriptions of the governing partial differential equations, boundary conditions, and discretizations used in these experiments.

The problems are flow problems involving the steady-state Navier–Stokes equations for low-speed, incompressible flow, specifically the equations for momentum transport, total mass conservation, and, in one case, thermal energy transport. The unknowns are the fluid velocity vector, the hydrodynamic pressure, and, in the one case, the temperature. See [30] for precise formulations of the equations and the boundary conditions in each case. In each of our experiments, an algebraic system of the form (1.1) was obtained from the PDEs using a stabilized finite-element formulation following [23] and [45]. The Jacobian of the system is nonsymmetric and, when needed, was evaluated analytically in our tests.

Our first test problem is the thermal-convection problem [8]. This involves modelling the flow of fluid in a differentially heated square box in the presence of gravity. The PDEs are the equations for momentum and thermal energy transport and for mass conservation. The unknowns include the temperature as well as the fluid velocity and pressure. The temperature is held fixed at different values on two opposite faces of the box, with zero heat flux imposed on the remaining sides. The fluid velocity is zero on all sides. When suitably nondimensionalized, the equations involve two parameters: the Rayleigh number Ra and the Prandtl number Pr. As Ra increases for fixed Pr, the nonlinear effects of the convection terms increase and the solution becomes increasingly difficult to obtain. In our experiments, we took Pr = 1 and varied Ra. We considered 2D and 3D forms of this problem. The 2D problem was discretized on a 100×100 equally spaced mesh, which resulted in 40,804 unknowns for the discretized problem. In 3D, the discretization was on a $32 \times 32 \times 32$ equally spaced mesh, resulting in 179,685 unknowns.

Our second test problem is the lid-driven cavity problem [18], [39]. This requires simulating flow in the unit square in $\mathbb{R}^2$ or the unit cube in $\mathbb{R}^3$ driven by a moving upper boundary. The PDEs are the momentum transport and mass conservation equations; the unknowns are the fluid velocity and pressure. The fluid velocity on the top side is fixed at a nonzero value in the x-axis direction and held at zero on all other sides. An appropriately nondimensionalized form of the PDEs leads to the Reynolds number Re, a nondimensional parameter expressing the ratio of convective to diffusive momentum transport. As Re increases, the nonlinear components of the equations

become more dominant and the problem becomes more difficult. We considered this problem in both 2D and 3D. The 2D computations were done on a 100×100 equally spaced grid, which led to 30,603 unknowns for the discretized problem. In 3D, the discretization was a $32 \times 32 \times 32$ equally spaced grid, yielding 143,748 unknowns.

Our third test problem is the backward-facing step problem [17]. This involves the simulation of flow through a rectangular channel that is initially constricted and subsequently expands over a reentrant backward-facing step. As above, the PDEs are those for momentum transport and mass conservation, and the unknowns are the fluid velocity and pressure. The nondimensionalized form of the equations is again parametrized by the Reynolds number Re, with problem difficulty increasing as Re increases. We considered this problem only in 2D. The flow was computed only in the expanded portion of the channel; flow entering from the constricted portion was simulated by imposing a parabolic velocity profile in the upper half of the inlet boundary and zero velocity on the lower half. See [30] for further details of the domain and boundary conditions. The problem was discretized on a 20×400 unequally spaced mesh (with a finer mesh near the step), which resulted in 25,263 unknowns.

4.3. The test results. In our experiments, we first conducted a robustness study. In this, we applied our implementation of Algorithm INDL with the algorithmic options described in section 4.1 to the test problems in section 4.2 and tabulated the numbers of failures with different options as problem parameters varied. The specific parameter values considered are as follows:

2D and 3D Thermal Convection	$\mathrm{Ra} = 10^3, 10^4, 10^5, 10^6,$
2D Lid-Driven Cavity	$\mathrm{Re} = 1,000, 2,000, \ldots, 10,000,$
3D Lid-Driven Cavity	$\mathrm{Re} = 100, 200, \ldots, 1,000,$
2D Backward-Facing Step	$\mathrm{Re} = 100, 200, \ldots, 700, 750, 800.$

The results of the robustness study are given in Table 4.1. For comparison, the table also includes results for NOX Newton–GMRES implementations with no globalization and with a backtracking globalization from [12] (see also [41], [31], [30]); these results were obtained in an earlier study involving identical test conditions (see [29, Tables 2–3]). Table 4.1 shows that both the dogleg and backtracking globalizations significantly improved robustness in these experiments, reducing the total numbers of failures with both forcing terms by 69% to 77%, compared to the method without globalization. Improvement was especially pronounced when adaptive forcing terms were used, a result consistent with other studies of globalized Newton–GMRES methods [31], [41], [30]. As also observed elsewhere [3], [41, section 5], [30], the dogleg and backtracking globalizations performed fairly comparably overall on these test problems, although there were some differences in the numbers of failures and where failures occurred. In view of the limited scope of these tests, we feel that further testing is needed to assess more fully the comparative robustness of these methods.

We next conducted an efficiency study of the dogleg methods, in which we compiled run times and other statistics for Algorithm INDL with different options on a selected set of test problem cases. This set included all cases considered in the robustness study in which all of the dogleg methods succeeded. Additionally, since all dogleg methods using the small constant forcing terms failed on the 2D lid-driven cavity problem for each value of Re above $1,000$, we included cases of this problem with $100 \leq \mathrm{Re} \leq 1,000$. The specific cases considered are as follows:

TABLE 4.1

*Robustness study: For each method, the upper and lower rows to the right show the numbers of failures with, respectively, the adaptive and small constant forcing terms. "TC," "LDC," and "BFS" denote, respectively, the thermal-convection, lid-driven cavity, and backward-facing step problems. * GMRES solves starting from zero. ** GMRES solves starting from the approximate Cauchy point.*

Method	2D problems			3D problems		Failure totals	
	TC	LDC	BFS	TC	LDC		
Algorithm INDL, Procedure 3.1	0	0	0	0	0	0	10
	1	9	0	0	0	10	
Algorithm INDL, Procedure 3.2*	0	2	0	0	0	2	12
	0	9	0	0	1	10	
Algorithm INDL, Procedure 3.2**	0	0	0	0	0	0	9
	0	9	0	0	0	9	
Backtracking globalization	0	0	0	0	0	0	10
	0	9	1	0	0	10	
No globalization	1	9	5	1	0	16	39
	1	10	7	1	4	23	

2D Thermal Convection	$\mathrm{Ra} = 10^3, 10^4, 10^5,$
3D Thermal Convection	$\mathrm{Ra} = 10^3, 10^4, 10^5, 10^6,$
2D Lid-Driven Cavity	$\mathrm{Re} = 100, 200, \ldots, 1,000,$
3D Lid-Driven Cavity	$\mathrm{Re} = 100, 200, \ldots, 900,$
2D Backward-Facing Step	$\mathrm{Re} = 100, 200, \ldots, 700, 750, 800.$

The results of the efficiency study are given in Table 4.2.[6] One sees that, as before, the forcing terms significantly influenced the performance of the methods. The adaptive forcing terms resulted in more inexact Newton steps on average than the more aggressive small constant forcing terms; however, the adaptive forcing terms led to significantly fewer GMRES iterations, both overall and per inexact Newton step. In the balance, the adaptive forcing terms yielded smaller run times. For a particular choice of forcing terms (adaptive or small constant), the methods performed fairly similarly, although when small constant forcing terms were used, the methods with Procedure 3.2 were somewhat less efficient in all statistics.

5. Concluding summary. In the preceding, we provide a theoretical framework and practical strategies for implementing dogleg globalizations of general inexact Newton methods for solving a nonlinear system (1.1).

In section 2, we outline a very general inexact Newton dogleg method. In this, the dogleg curves are defined using general inexact Newton steps together with approximate steepest-descent directions and Cauchy points. To be admissible, trial steps along these curves must satisfy only mild minimum-length requirements in addition to being within trust regions. A convergence theorem for this method is given; this asserts that if the approximate steepest-descent directions are bounded away from

[6] See [30, Table 5.2] for a similar study comparing Algorithm INDL using Procedure 3.1 with backtracking and linesearch globalizations.

TABLE 4.2

Efficiency study: For each method, the upper and lower rows to the right show results with the adaptive and small constant forcing terms, respectively. All statistics are geometric means. Times are relative to that of Algorithm INDL *with Procedure* 3.1 *and small constant forcing terms. "INS" stands for "Inexact Newton Steps."*

** GMRES solves starting from zero. ** GMRES solves starting from the approximate Cauchy point.*

Method	Inexact Newton steps	Function evaluations	GMRES iterations	GMRES iterations per INS	Normalized time
Algorithm INDL, Procedure 3.1	18.82	21.11	1160	61.65	0.87
	10.30	12.45	1330	129.2	1.00
Algorithm INDL, Procedure 3.2*	16.62	18.87	1200	72.21	0.92
	11.27	13.71	1650	146.4	1.21
Algorithm INDL, Procedure 3.2**	18.40	20.57	1122	60.99	0.88
	11.15	13.55	1634	146.6	1.23

orthogonality with the exact steepest-descent directions, then every limit point of a sequence of iterates generated by the method is a stationary point of $\|F\|$. Moreover, if there is a limit point at which the Jacobian of F is nonsingular, then that point must be a solution of (1.1) and the iterates must converge to it. In this case, initial trial steps are ultimately accepted without modification by the method, and the inexact Newton steps are ultimately admissible trial steps.

In section 3, we discuss strategies for selecting trial steps along the dogleg curves. It is noted that the step-selection strategy traditionally used in the exact-Newton case may exhibit certain shortcomings, principally that, when the forcing term is not small, this strategy may select the inexact Newton step even though a step biased toward the approximate Cauchy point may give significantly greater linear model norm reduction within the trust region. An alternative strategy is proposed that avoids this shortcoming. As a refinement of these strategies, simple modifications are introduced that further determine each dogleg step to minimize the local linear model norm along the dogleg curve within the trust region, provided both the approximate Cauchy point and the inexact Newton step have been computed.

In section 4, we report on numerical experiments in which a Newton–GMRES implementation of the general inexact Newton dogleg method in section 2 using the step-selection strategies in section 3 was applied to several two- and three-dimensional benchmark problems involving the steady-state Navier–Stokes equations. Although somewhat limited in scope, these experiments demonstrate the robustness of the method with these step-selection strategies and other algorithmic options on these challenging problems. They also provide an indication of the relative efficiency of these strategies and options. The use of adaptive forcing terms resulted in significantly greater robustness and some improvement in efficiency compared to the constant forcing term strategy; these results are consistent with previous studies of globalized inexact Newton methods. Otherwise, these experiments do not show dramatic differences in performance among the strategies and options that we tested. We feel that this may reflect the particular nature of the limited test set and also the rather modest problem sizes; there may be more pronounced differences in performance on other, broader test sets and higher-resolution problems. A more comprehensive numerical study will be the subject of future work.

Acknowledgment. The authors thank the anonymous referees for their thoughtful reviews, which led to significant improvements in this paper.

REFERENCES

[1] K. E. Brenan, S. L. Campbell, and L. R. Petzold, *Numerical Solution of Initial-value problems in differential-algebraic equations*, Classics Appl. Math. 14, SIAM, Philadelphia, 1996.

[2] P. N. Brown, *A local convergence theory for combined inexact-Newton/finite-difference projection methods*, SIAM J. Numer. Anal., 24 (1987), pp. 407–434.

[3] P. N. Brown and Y. Saad, *Hybrid Krylov methods for nonlinear systems of equations*, SIAM J. Sci. Statist. Comput., 11 (1990), pp. 450–481.

[4] P. N. Brown and Y. Saad, *Convergence theory of nonlinear Newton–Krylov algorithms*, SIAM J. Optim., 4 (1994), pp. 297–330.

[5] P. N. Brown, H. F. Walker, R. Wasyk, and C. S. Woodward, *On using approximate finite differences in matrix-free Newton–Krylov methods*, SIAM J. Numer. Anal., 46 (2008), pp. 1892–1911.

[6] E. Catinas, *Inexact perturbed Newton methods and applications to a class of Krylov solvers*, J. Optim. Theory Appl., 108 (2001), pp. 543–570.

[7] A. R. Conn, N. I. M. Gould, and Ph. L. Toint, *Trust-Region Methods*, MPS-SIAM Ser. Optim., SIAM, Philadelphia, 2000.

[8] G. De Vahl Davis and C. P. Jones, *Natural convection in a square cavity: A comparison exercise*, Int. J. Numer. Methods Fluids, 3 (1983), pp. 227–248.

[9] R. S. Dembo, S. C. Eisenstat, and T. Steihaug, *Inexact Newton methods*, SIAM J. Numer. Anal., 19 (1982), pp. 400–408.

[10] R. S. Dembo and T. Steihaug, *Truncated Newton algorithms for large-scale optimization*, Math. Programming, 26 (1983), pp. 190–212.

[11] J. E. Dennis, Jr. and R. B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice Hall Ser. Comput. Math., Prentice-Hall, Englewood Cliffs, NJ, 1983.

[12] S. C. Eisenstat and H. F. Walker, *Globally convergent inexact Newton methods*, SIAM J. Optim., 4 (1994), pp. 393–422.

[13] S. C. Eisenstat and H. F. Walker, *Choosing the forcing terms in an inexact Newton method*, SIAM J. Sci. Comput., 17 (1996), pp. 16–32.

[14] M. El Hallabi, *A Global Convergence Theory for Arbitrary Norm Trust-Region Methods for Nonlinear Equations*, Ph.D. thesis, Department of Mathematical Sciences, Rice University, Houston, TX, 1987.

[15] M. El Hallabi and R. A. Tapia, *A Global Convergence Theory for Arbitrary Norm Trust-Region Methods for Nonlinear Equations*, Technical report TR87-25, Department of Mathematical Sciences, Rice University, Houston, TX, 1987, revised 1989.

[16] R. W. Freund, G. H. Golub, and N. M. Nachtigal, *Iterative solution of linear systems*, Acta Numer., 1 (1992), pp. 57–100.

[17] D. K. Gartling, *A test problem for outflow boundary conditions–flow over a backward facing step*, Internat. J. Numer. Methods Fluids, 11 (1990), pp. 953–967.

[18] U. Ghia, K. N. Ghia, and C. T. Shin, *High-Re solutions for incompressible flow using the Navier-Stokes equations and a multigrid method*, J. Comput. Phys., 48 (1982), pp. 387–411.

[19] G. H. Golub and C. Van Loan, *Matrix Computations*, 3rd ed., Johns Hopkins University Press, Baltimore, 1996.

[20] M. H. Gutknecht, *Lanczos-type solvers for nonsymmetric linear systems of equations*, Acta Numer., 6 (1997), pp. 271–397.

[21] B. Hendrickson and R. Leland, *The Chaco User's Guide–Version* 1.0, Technical report Sand93-2339, Sandia National Laboratories, Albuquerque NM, 1993.

[22] M. Heroux, *AztecOO: Object-Oriented Aztec Linear Solver Package*, available online at http://software.sandia.gov/trilinos/packages/aztecoo/index.html.

[23] T. J. R. Hughes, L. P. Franca, and G. M. Hulbert, *A new finite element formulation for computational fluid dynamics:* VII. *The Galerkin/least-squares method for advective-diffusive equations*, Comput. Methods Appl. Mech. Engrg., 73 (1989), pp. 173–189.

[24] C. T. Kelley, *Solving Nonlinear Equations with Newton's Method*, Fundam. Algorithms, SIAM, Philadelphia, 2003.

[25] D. A. Knoll and D. E. Keyes, *Jacobian-free Newton–Krylov methods: A survey of approaches*

and applications, J. Comput. Phys., 193 (2004), pp. 357–397.
[26] T. G. Kolda and R. P. Pawlowski, *NOX Nonlinear Solver Project*, available online at http://software.sandia.gov/nox.
[27] J. J. Moré and D. C. Sorensen, *Computing a trust region step*, SIAM J. Sci. Statist. Comput., 4 (1983), pp. 553–572.
[28] S. G. Nash, *Truncated Newton Methods*, Ph.D. thesis, Computer Science Department, Stanford University, Palo Alto, CA, 1982.
[29] R. P. Pawlowski, J. N. Shadid, J. P. Simonis, and H. F. Walker, *Globalization Techniques for Newton–Krylov Methods and Applications to the Fully-Coupled Solution of the Navier–Stokes Equations*, Technical report Sand2004-1777, Sandia National Laboratories, Albuquerque NM, 2004.
[30] R. P. Pawlowski, J. N. Shadid, J. P. Simonis, and H. F. Walker, *Globalization techniques for Newton–Krylov methods and applications to the fully-coupled solution of the Navier–Stokes equations*, SIAM Rev., 48 (2006), pp. 700–721.
[31] M. Pernice and H. F. Walker, *NITSOL: A Newton iterative solver for nonlinear systems*, SIAM J. Sci. Comput., 19 (1998), pp. 302–318.
[32] M. J. D. Powell, *A FORTRAN Subroutine for Unconstrained Minimization, Requiring First Derivatives of the Objective Function*, Technical report AERE-R. 6469, Mathematics Branch, A.E.R.E. Harwell, Berkshire, England, 1970.
[33] M. J. D. Powell, *A hybrid method for nonlinear equations*, in Numerical Methods for Nonlinear Algebraic Equations, P. Rabinowitz, ed., Gordon and Breach, London, 1970, pp. 87–114.
[34] M. J. D. Powell, *General algorithms for discrete nonlinear approximation calculations*, in Approximation Theory IV, C. K. Chui, L. L. Schumaker, and J. D. Ward, eds., Academic Press, New York, 1983, pp. 187–218.
[35] Y. Saad, *Krylov subspace methods for solving large unsymmetric linear systems*, Math. Comp., 37 (1981), pp. 105–126.
[36] Y. Saad, *ILUT: A dual threshold incomplete ILU factorization*, Numer. Linear Algebra Appl., 1 (1994), pp. 387–402.
[37] Y. Saad, *Iterative Methods for Sparse Linear Systems*, PWS Publishing Company, Boston, 1996.
[38] Y. Saad and M. H. Schultz, *GMRES: A generalized minimal residual method for solving nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 856–869.
[39] R. Schreiber and H. B. Keller, *Driven cavity flows by efficient numerical techniques*, J. Comput. Phys., 49 (1983), pp. 310–333.
[40] J. N. Shadid, H. K. Moffat, S. A. Hutchinson, G. L. Hennigan, K. D. Devine, and A. G. Salinger, *MPSalsa: A Finite Element Computer Program for Reacting Flow Problems Part* 1: *Theoretical Development*, Technical report Sand95-2752, Sandia National Laboratories, Albuquerque NM, 1996.
[41] J. N. Shadid, R. S. Tuminaro, and H. F. Walker, *An inexact Newton method for fully coupled solution of the Navier–Stokes equations with heat and mass transport*, J. Comput. Phys., 137 (1997), pp. 155–185.
[42] J. N. Shadid, R. S. Tuminaro, and H. F. Walker, *On backtracking failure in Newton–GMRES methods with a demonstration for the Navier–Stokes equations*, J. Comput. Phys., 180 (2002), pp. 549–558.
[43] G. A. Shultz, R. B. Schnabel, and R. H. Byrd, *A family of trust-region-based algorithms for unconstrained minimization with strong global convergence properties*, SIAM J. Numer. Anal., 22 (1985), pp. 47–67.
[44] T. Steihaug, *The conjugate gradient method and trust regions in large scale optimization*, SIAM J. Numer. Anal., 20 (1983), pp. 626–637.
[45] T. E. Tezduyar, *Stabilized finite element formulations for incompressible flow computations*, Adv. App. Mech., 28 (1992), pp. 1–44.
[46] Ph. L. Toint, *Towards an efficient sparsity exploiting Newton method for minimization*, in Sparse Matrices and Their Uses, I. S. Duff, ed., Academic Press, New York, 1981, pp. 57–87.
[47] R. S. Tuminaro, M. Heroux, S. A. Hutchinson, and J. N. Shadid, *Aztec User's Guide–Version* 2.1, Technical report Sand99-8801J, Sandia National Laboratories, Albuquerque NM, 1999.
[48] H. A. van der Vorst, *Iterative Krylov Methods for Large Linear Systems*, Cambridge Monogr. Appl. Comput. Math. 13, Cambridge University Press, Cambridge, 2003.
[49] C. S. Woodward, *private communication*, 2006.

SIAM J. NUMER. ANAL.
Vol. 46, No. 4, pp. 2133–2152
© 2008 Society for Industrial and Applied Mathematics

AN OPTIMAL-ORDER ERROR ESTIMATE FOR A FAMILY OF ELLAM-MFEM APPROXIMATIONS TO POROUS MEDIUM FLOW*

HONG WANG†

Abstract. Mathematical models used to describe porous medium flow lead to coupled systems of time-dependent nonlinear partial differential equations, which present serious mathematical and numerical difficulties. Standard methods tend to generate numerical solutions with nonphysical oscillations or numerical dispersion along with spurious grid-orientation effect. The ELLAM-MFEM time-stepping procedure, in which an Eulerian–Lagrangian localized adjoint method (ELLAM) is used to solve the transport equation and a mixed finite element method (MFEM) is used for the pressure equation, simulates porous medium flow accurately even if large spatial grids and time steps are used. In this paper we prove an optimal-order error estimate for a family of ELLAM-MFEM approximations.

Key words. advection-diffusion equations, characteristic methods, convergence analysis, error estimates, Eulerian–Lagrangian methods, numerical analysis, porous medium flow, subsurface flow

AMS subject classifications. 65M12, 65M15, 65M25, 65M60, 65Z05, 76M10, 76S05

DOI. 10.1137/S0036142903428281

1. Introduction.

Mathematical models used to describe porous medium flow processes in petroleum reservoir simulation, groundwater contaminant transport, and other applications lead to coupled systems of time-dependent nonlinear partial differential equations (PDEs) [2, 11]. Due to the advection-diffusion feature of the transport PDEs, the nonlinearity and couplings of the PDEs in the systems, and the strong effect of sources and sinks, these systems admit solutions with moving steep fronts and complicated structures. Consequently, these problems present serious mathematical and numerical difficulties. Standard finite difference or finite element methods (FDMs and FEMs, respectively) tend to generate solutions with nonphysical oscillations. Upwind methods are often used to stabilize the numerical approximations, but they often produce numerical dispersion and the grid-orientation effect [11]. Extensive research has been carried out on the development of improved methods for advection-diffusion PDEs [16, 18, 21].

An FEM-MFEM time-stepping procedure was presented and analyzed in [8, 9]. In the procedure, a mixed finite element method (MFEM) scheme [4, 5, 20] was used to solve the pressure PDE, and an FEM was used to solve the transport PDE. MFEM schemes generate an accurate approximation to the Darcy velocity, which is required for accurate approximation to the concentration because advection and diffusion dispersion in the transport PDE are governed by Darcy velocity. MFEMs minimize the numerical difficulties occurring in FDMs or FEMs caused by differentiation of the pressure and then multiplication by rough coefficients [22]. Subsequently, an MMOC-MFEM time-stepping procedure was proposed and analyzed in [12, 13], in which the modified method of characteristics (MMOC) [10] was used to replace the FEM scheme in [9]. The MMOC symmetrizes and stabilizes the transport PDE, greatly reduces temporal errors, and so allows for large time steps in a simulation without loss of

*Received by the editors May 15, 2003; accepted for publication (in revised form) October 25, 2007; published electronically May 2, 2008.
http://www.siam.org/journals/sinum/46-4/42828.html
†Department of Mathematics, University of South Carolina, Columbia, SC 29208 (hwang@math.sc.edu).

accuracy. However, the MMOC fails to conserve mass and has difficulties in handling general boundary conditions.

The Eulerian–Lagrangian localized adjoint method (ELLAM) framework [6, 21] provides a systematic approach to solve transient advection-diffusion PDEs with general boundary conditions in a mass-conservative manner, while maintaining accuracy and efficiency of Eulerian–Lagrangian methods. Numerical experiments show that the ELLAM is very competitive in the context of model transport PDEs [26, 28] and miscible porous medium flow [3, 17, 30, 32]. Nevertheless, corresponding theoretical analysis falls far behind. To date, error estimates for the ELLAM and related schemes have been proved in the context of linear advection-dominated transport PDEs [1, 14, 15, 24, 25, 27, 29, 31]. This is partly due to the fact that the analyses are more complex than those for the MMOC and related methods [10, 29, 21]. In this paper we derive an optimal-order error estimate for an ELLAM-MFEM approximation to the coupled system of miscible porous medium flow.

The rest of the paper is organized as follows: In section 2 we review the mathematical model. In section 3 we describe the ELLAM-MFEM time-stepping procedure. In section 4 we present known theoretical results used in the analysis. In section 5 we prove an optimal-order error estimate for the ELLAM-MFEM time-stepping procedure. In section 6 we prove several lemmas used in section 5.

2. Mathematical model and notation. We present a mathematical model for porous medium flow and introduce the functional spaces used in this paper.

2.1. Mathematical model. Let $c(\mathbf{x},t)$ be the concentration of an invading fluid or a concerned solute/solvent, and let $p(\mathbf{x},t)$ and $\mathbf{u}(\mathbf{x},t)$ be the pressure and Darcy velocity of the fluid mixture, respectively. The mass conservation for the fluid mixture incorporated with the incompressibility condition, Darcy's law, and the mass conservation for the invading fluid lead to the following system of PDEs [2, 11]:

$$\nabla\cdot\mathbf{u}=q,\qquad \mathbf{u}=-\frac{\mathbf{K}}{\mu(c)}(\nabla p-\rho g\nabla d),\qquad \mathbf{x}\in\Omega,\quad t\in[0,T], \tag{2.1}$$

$$\phi\frac{\partial c}{\partial t}+\nabla\cdot(\mathbf{u}c-\mathbf{D}(\mathbf{x},\mathbf{u})\nabla c)=\bar{c}q,\qquad \mathbf{x}\in\Omega,\quad t\in[0,T], \tag{2.2}$$

$$\begin{aligned}\mathbf{u}\cdot\mathbf{n}&=(\mathbf{D}\nabla c)\cdot\mathbf{n}=0, && \mathbf{x}\in\partial\Omega,\quad t\in[0,T],\\ c(\mathbf{x},0)&=c_0(\mathbf{x}), && \mathbf{x}\in\Omega.\end{aligned} \tag{2.3}$$

We assume that the medium is homogeneous vertically and take $\Omega\subset\mathbb{R}^2$ with a nonuniform local elevation. $\phi(\mathbf{x})$ and $\mathbf{K}(\mathbf{x})$ are the porosity and the permeability tensor of the medium, respectively, $\mu(c)$ and ρ are the viscosity and the density of the fluid mixture, respectively, g is the gravitational acceleration, $d(\mathbf{x})$ is the reservoir depth, and $q(\mathbf{x},t)$ is the source and sink term. $\mathbf{D}(\mathbf{x},\mathbf{u})=\phi(\mathbf{x})d_m\,\mathbf{I}+d_t|\mathbf{u}|+(d_l-d_t)(u_iu_j)_{i,j=1}^2/|\mathbf{u}|$ is the diffusion-dispersion tensor, with d_m, d_t, and d_l being the molecular diffusion and the transverse and longitudinal dispersivities, respectively, and $\mathbf{I}$ is the identity tensor. $\bar{c}(\mathbf{x},t)$ is specified at sources and $\bar{c}(\mathbf{x},t)=c(\mathbf{x},t)$ at sinks. $c_0(\mathbf{x})$ is the initial concentration.

The combination of the first equations in (2.1) and (2.3) implies that q has mean 0. In addition, (2.1) with the noflow boundary condition (2.3) can determine the pressure $p(\mathbf{x},t)$ only up to an additive constant for all the time $t\in[0,T]$. But this indeterminacy is of no consequence since $\mathbf{u}$ is uniquely determined by Darcy's law, and only $\mathbf{u}$ (not p) is needed in (2.2).

2.2. Notation. Let $W_q^m(\Omega)$ be the Sobolev spaces consisting of functions whose derivatives up to order-m are qth integrable on Ω, and $H^m(\Omega) := W_2^m(\Omega)$. Let $L_0^2(\Omega)$ be the subspace of $L^2(\Omega)$ with mean 0. We introduce vector-valued Sobolev spaces

$$H^m(\mathrm{div};\Omega) := \Big\{\mathbf{f}(\mathbf{x}) = (f_1, f_2) \ : \ f_1, f_2, \nabla\cdot\mathbf{f} \in H^m(\Omega)\Big\},$$
$$\|\mathbf{f}\|_{H^m(\mathrm{div};\Omega)} := \Big(\|f_1\|^2_{H^m(\Omega)} + \|f_2\|^2_{H^m(\Omega)} + \|\nabla\cdot\mathbf{f}\|^2_{H^m(\Omega)}\Big)^{1/2},$$
$$H_0(\mathrm{div};\Omega) := \Big\{\mathbf{f}(\mathbf{x}) \in H^0(\mathrm{div};\Omega) \ : \ \mathbf{f}(\mathbf{x})\cdot\mathbf{n}(\mathbf{x}) = 0,\ \mathbf{x}\in\partial\Omega\Big\}.$$

We drop Ω in these notations when it is clear from the context.

For any Banach space X, we introduce Sobolev spaces involving time

$$W_q^m(t_1,t_2;X) := \left\{ f(\mathbf{x},t) \ : \ \Big\|\frac{\partial^\alpha f}{\partial t^\alpha}(\cdot,t)\Big\|_X \in L^q(t_1,t_2),\ 0\le\alpha\le m,\ 1\le q\le\infty \right\},$$
$$\|f\|_{W_q^m(t_1,t_2;X)} := \begin{cases} \displaystyle\left(\sum_{\alpha=0}^m \int_{t_1}^{t_2} \Big\|\frac{\partial^\alpha f}{\partial t^\alpha}(\cdot,t)\Big\|_X^q dt\right)^{1/q}, & 1\le q<\infty, \\ \displaystyle\max_{0\le\alpha\le m} \mathrm{esssup}_{t\in(t_1,t_2)} \Big\|\frac{\partial^\alpha f}{\partial t^\alpha}(\cdot,t)\Big\|_X, & q=\infty. \end{cases}$$

We also define the discrete norms

$$\|f\|_{\hat{L}_c^\infty(0,T;X)} := \max_{0\le n\le N}\|f(\mathbf{x},t_n^c)\|_X, \qquad \|f\|_{\hat{L}_p^\infty(0,T;X)} := \max_{0\le m\le M}\|f(\mathbf{x},t_m^p)\|_X,$$
$$\|f\|_{\hat{L}_c^2(0,T;X)} := \left(\sum_{n=0}^N \|f(\mathbf{x},t_n^c)\|_X^2 \Delta t_n^c\right)^{1/2},$$

with t_n^c and t_m^p being defined in (3.1) and (3.11), respectively. If $(t_1,t_2) = (0,T)$, we drop it from these notations.

In this paper we use ε to denote an arbitrarily small positive number, A_i, K_i, and Q_i to denote fixed positive constants, and Q to denote a generic positive constant that could assume different values at different occurrences. All of these constants are independent of spatial and temporal grid parameters.

3. An ELLAM-MFEM time-stepping procedure. In this procedure an ELLAM scheme is used to solve the transport PDE (2.2), and an MFEM scheme is used for the pressure system (2.1).

3.1. An ELLAM formulation for the transport equation. We define a temporal partition on the interval $[0,T]$ by

$$0 =: t_0^c < t_1^c < \cdots < t_n^c < \cdots < t_{N-1}^c < t_N^c := T, \tag{3.1}$$

with $\Delta t_n^c := t_n^c - t_{n-1}^c$ and $\Delta t_c := \max_{1\le n\le N}\Delta t_n^c$.

By multiplying (2.2) by any space-time test functions z that are continuous and piecewise smooth, vanish outside the space-time strip $\Omega\times(t_{n-1}^c, t_n^c]$, and are discon-

tinuous in time at time t_{n-1}^c, we obtain a weak formulation for (2.2)

$$\begin{aligned}
&\int_\Omega \phi(\mathbf{x}) c(\mathbf{x}, t_n^c) z(\mathbf{x}, t_n^c) d\mathbf{x} + \int_{t_{n-1}^c}^{t_n^c} \int_\Omega \nabla z(\mathbf{y}, \theta) \cdot \mathbf{D}(\mathbf{y}, \mathbf{u}(\mathbf{y}, \theta)) \nabla c(\mathbf{y}, \theta) d\mathbf{y} d\theta \\
&\quad - \int_{t_{n-1}^c}^{t_n^c} \int_\Omega c(\mathbf{y}, \theta) \left[\phi(\mathbf{y}) \frac{\partial z(\mathbf{y}, \theta)}{\partial \theta} + \mathbf{u}(\mathbf{y}, \theta) \cdot \nabla z(\mathbf{y}, \theta) \right] d\mathbf{y} d\theta \\
&\quad = \int_\Omega \phi(\mathbf{x}) c(\mathbf{x}, t_{n-1}^c) z(\mathbf{x}, t_{n-1}^{c,+}) d\mathbf{x} + \int_{t_{n-1}^c}^{t_n^c} \int_\Omega \bar{c}(\mathbf{y}, \theta) q(\mathbf{y}, \theta) z(\mathbf{y}, \theta) \, d\mathbf{y} d\theta,
\end{aligned} \tag{3.2}$$

where $z(\mathbf{x}, t_{n-1}^{c,+}) := \lim_{t \to t_{n-1}^c,\ t > t_{n-1}^c} z(\mathbf{x}, t)$ accounts for the discontinuity of $z(\mathbf{x}, t)$ in time at time t_{n-1}^c. We replace the dummy variables $\mathbf{x}$ and t in the space-time integrals in (3.2) by $\mathbf{y}$ and θ and reserve $\mathbf{x}$ for the point in Ω at time t_n^c or t_{n-1}^c.

In the ELLAM framework [6, 21, 30], the test functions in (3.2) are defined to be the solutions of the adjoint equation of the hyperbolic part of (2.2)

$$\phi(\mathbf{y}) \, \frac{\partial z(\mathbf{y}, \theta)}{\partial \theta} + \mathbf{u}(\mathbf{y}, \theta) \cdot \nabla z(\mathbf{y}, \theta) = 0, \quad \mathbf{y} \in \overline{\Omega}, \quad \theta \in [t_{n-1}^c, t_n^c]. \tag{3.3}$$

Thus, the last term on the left-hand side of (3.2) vanishes. Once the test functions $z(\mathbf{x}, t_n^c)$ are specified in Ω, they are determined in the space-time strip $\Omega \times (t_{n-1}^c, t_n^c]$ by constant extension along the characteristics $\mathbf{y} = \mathbf{r}(\theta; \mathbf{x}, t_n^c)$ defined by

$$\frac{d\mathbf{r}}{d\theta} = \frac{\mathbf{u}(\mathbf{r}, \theta)}{\phi(\mathbf{r})} \quad \text{and} \quad \mathbf{r}(\theta; \bar{\mathbf{x}}, \bar{t})\Big|_{\theta = \bar{t}} = \bar{\mathbf{x}}. \tag{3.4}$$

Evaluating the source term in (3.2) by the Euler formula at time t_n^c yields

$$\begin{aligned}
&\int_{t_{n-1}^c}^{t_n^c} \int_\Omega \bar{c}(\mathbf{y}, \theta) q(\mathbf{y}, \theta) z(\mathbf{y}, \theta) \, d\mathbf{y} d\theta \\
&\quad = \int_\Omega \int_{t_{n-1}^c}^{t_n^c} \bar{c}(\mathbf{r}(\theta; \mathbf{x}, t_n^c), \theta) q(\mathbf{r}(\theta; \mathbf{x}, t_n^c), \theta) z(\mathbf{x}, t_n^c) \det \left(\frac{\partial \mathbf{r}(\theta; \mathbf{x}, t_n^c)}{\partial \mathbf{x}} \right) d\theta d\mathbf{x} \\
&\quad = \Delta t_n^c \int_\Omega \bar{c}(\mathbf{x}, t_n^c) q(\mathbf{x}, t_n^c) z(\mathbf{x}, t_n^c) d\mathbf{x} + E_q(\bar{c}, z),
\end{aligned} \tag{3.5}$$

where $E_q(\bar{c}, z)$ is the local truncation error term given by

$$\begin{aligned}
E_q(\bar{c}, z) := - \int_\Omega z(\mathbf{x}, t_n^c) \Bigg\{ \int_{t_{n-1}^c}^{t_n^c} \Bigg[&\bar{c}(\mathbf{x}, t_n^c) q(\mathbf{x}, t_n^c) \\
&- \bar{c}(\mathbf{r}(\theta; \mathbf{x}, t_n^c), \theta) q(\mathbf{r}(\theta; \mathbf{x}, t_n^c), \theta) \det \left(\frac{\partial \mathbf{r}(\theta; \mathbf{x}, t_n^c)}{\partial \mathbf{x}} \right) \Bigg] d\theta \Bigg\} d\mathbf{x}.
\end{aligned} \tag{3.6}$$

In (3.5) we have implicitly used the fact that the traceback operator $\mathbf{r}$ is a diffeomorphism and the determinant is positive [10].

Likewise, we evaluate the diffusion-dispersion term in (3.2) to obtain

$$
\begin{aligned}
&\int_{t_{n-1}^c}^{t_n^c}\int_\Omega \nabla z(\mathbf{y},\theta)\cdot\mathbf{D}(\mathbf{y},\mathbf{u}(\mathbf{y},\theta))\nabla c(\mathbf{y},\theta)d\mathbf{y}d\theta\\
&\quad=\int_\Omega\int_{t_{n-1}^c}^{t_n^c}\nabla_{\mathbf{r}} z(\mathbf{x},t_n^c)\cdot(\mathbf{D}(\mathbf{r}(\theta;\mathbf{x},t_n^c),\mathbf{u}(\mathbf{r}(\theta;\mathbf{x},t_n^c),\theta))\\
&\qquad\qquad\nabla c_{\mathbf{r}}(\mathbf{r}(\theta;\mathbf{x},t_n^c),\theta))\det\left(\frac{\partial\mathbf{r}(\theta;\mathbf{x},t_n^c)}{\partial\mathbf{x}}\right)d\theta d\mathbf{x}\\
&\quad=\int_\Omega\int_{t_{n-1}^c}^{t_n^c}\det\left(\frac{\partial\mathbf{r}(\theta;\mathbf{x},t_n^c)}{\partial\mathbf{x}}\right)\left(\frac{\partial\mathbf{r}(\theta;\mathbf{x},t_n^c)}{\partial\mathbf{x}}\right)^{-1}\nabla z(\mathbf{x},t_n^c)\\
&\qquad\cdot(\mathbf{D}(\mathbf{r}(\theta;\mathbf{x},t_n^c),\mathbf{u}(\mathbf{r}(\theta;\mathbf{x},t_n^c),\theta))\nabla c_{\mathbf{r}}(\mathbf{r}(\theta;\mathbf{x},t_n^c),\theta))d\theta d\mathbf{x}\\
&\quad=\Delta t_n^c\int_\Omega\nabla z(\mathbf{x},t_n^c)\cdot\mathbf{D}(\mathbf{x},\mathbf{u}(\mathbf{x},t_n^c))\nabla c(\mathbf{x},t_n^c)\;d\mathbf{x}+E_{\mathbf{D}}(c,z).
\end{aligned}
\tag{3.7}
$$

The subscript $\mathbf{r}$ emphasizes that the gradient is taken with respect to $\mathbf{r}$. The local truncation error $E_{\mathbf{D}}(c,z)$ is given by

$$
\begin{aligned}
E_{\mathbf{D}}(c,z):=&\int_\Omega\int_{t_{n-1}^c}^{t_n^c}\left(\det\left(\frac{\partial\mathbf{r}(\theta;\mathbf{x},t_n^c)}{\partial\mathbf{x}}\right)\left(\frac{\partial\mathbf{r}(\theta;\mathbf{x},t_n^c)}{\partial\mathbf{x}}\right)^{-1}-\mathbf{I}\right)\nabla z(\mathbf{x},t_n^c)\\
&\qquad\cdot(\mathbf{D}(\mathbf{r}(\theta;\mathbf{x},t_n^c),\mathbf{u}(\mathbf{r}(\theta;\mathbf{x},t_n^c),\theta))\nabla c(\mathbf{r}(\theta;\mathbf{x},t_n^c),\theta))d\theta d\mathbf{x}\\
&-\int_\Omega\int_{t_{n-1}^c}^{t_n^c}\nabla z(\mathbf{x},t_n)\cdot(\mathbf{D}(\mathbf{x},\mathbf{u}(\mathbf{x},t_n^c))\nabla c(\mathbf{x},t_n^c)\\
&\qquad\qquad-\mathbf{D}(\mathbf{r}(\theta;\mathbf{x},t_n^c),\mathbf{u}(\mathbf{r}(\theta;\mathbf{x},t_n^c),\theta))\nabla c(\mathbf{r}(\theta;\mathbf{x},t_n^c),\theta))d\theta d\mathbf{x}.
\end{aligned}
\tag{3.8}
$$

We substitute (3.5) and (3.7) into (3.2) to obtain an ELLAM reference equation for the transport PDE (2.2)

$$
\begin{aligned}
&\int_\Omega\phi(\mathbf{x})c(\mathbf{x},t_n^c)z(\mathbf{x},t_n^c)d\mathbf{x}+\Delta t_n^c\int_\Omega\nabla z(\mathbf{x},t_n^c)\cdot\mathbf{D}(\mathbf{x},\mathbf{u}(\mathbf{x},t_n^c))\nabla c(\mathbf{x},t_n^c)d\mathbf{x}\\
&\quad=\int_\Omega\phi(\mathbf{x})c(\mathbf{x},t_{n-1}^c)z(\mathbf{x},t_{n-1}^{c,+})d\mathbf{x}+\Delta t_n^c\int_\Omega\bar{c}(\mathbf{x},t_n^c)q(\mathbf{x},t_n^c)z(\mathbf{x},t_n^c)d\mathbf{x}\\
&\qquad-E_{\mathbf{D}}(c,z)+E_q(\bar{c},z).
\end{aligned}
\tag{3.9}
$$

3.2. An MFEM formulation for the pressure and Darcy velocity. We multiply the second equation in (2.1) by $\mu(c)\mathbf{K}^{-1}(\mathbf{x})$ and any test functions $\mathbf{v}\in H_0(\mathrm{div};\Omega)$ and apply the divergence theorem to the ∇p term. We then multiply the first equation in (2.1) by any test functions $w(\mathbf{x})\in L_0^2(\Omega)$ and integrate over Ω. The system (2.1) is expressed as a time-parameterized saddle-point problem of finding a map $(\mathbf{u}(\mathbf{x},t),p(\mathbf{x},t))\in H_0(\mathrm{div};\Omega)\times L_0^2(\Omega)$ such that

$$
\begin{aligned}
\int_\Omega\mu(c)\mathbf{K}^{-1}\mathbf{u}\cdot\mathbf{v}d\mathbf{x}-\int_\Omega p\nabla\cdot\mathbf{v}d\mathbf{x}&=\int_\Omega\rho g\nabla d\cdot\mathbf{v}d\mathbf{x},\\
\int_\Omega w\nabla\cdot\mathbf{u}d\mathbf{x}&=\int_\Omega q(\mathbf{x},t)wd\mathbf{x}\\
\forall(\mathbf{v}(\mathbf{x}),w(\mathbf{x}))\in H_0(\mathrm{div};\Omega)\times L_0^2(\Omega),t\in[0,T].&
\end{aligned}
\tag{3.10}
$$

It was proved [4, 8] that problem (3.10) has a unique solution $(\mathbf{u}(\mathbf{x},t),p(\mathbf{x},t)) \in H_0(\mathrm{div};\Omega)\times L_0^2(\Omega)$ for all $t\in[0,T]$.

Let $V^k \subset H_0(\mathrm{div};\Omega)$ and $S_p^k \subset L_0^2(\Omega)$ be MFEM spaces of index $k\geq 0$ on a quasi-uniform partition of $\Omega = \cup\Omega_e^p$ with the diameter h_p [5, 20]. We define a temporal partition on the time interval $[0,T]$ for the pressure grid by

$$0 =: t_0^p < t_1^p < \cdots < t_m^p < \cdots < t_{M-1}^p < t_M^p := T, \tag{3.11}$$

with $\Delta t_m^p := t_m^p - t_{m-1}^p$ and $\Delta t_p := \max_{1\leq m\leq M}\Delta t_m^p$.

Given a concentration approximation $c_h(\mathbf{x},t_m^p)$ at time t_m^p, the MFEM scheme determines the velocity $\mathbf{u}_h(\mathbf{x},t_m^p)\in V^k$ and the pressure $p_h(\mathbf{x},t_m^p)\in S_p^k$ such that

$$\begin{aligned}
&\int_\Omega \mu(c_h(\mathbf{x},t_m^p))\mathbf{K}^{-1}(\mathbf{x})\mathbf{u}_h(\mathbf{x},t_m^p)\cdot\mathbf{v}_h(\mathbf{x})d\mathbf{x} - \int_\Omega p_h(\mathbf{x},t_m^p)\nabla\cdot\mathbf{v}_h(\mathbf{x})d\mathbf{x}\\
&\quad = \int_\Omega \rho g\nabla d(\mathbf{x})\cdot\mathbf{v}_h(\mathbf{x})d\mathbf{x} \qquad\qquad \forall \mathbf{v}_h(\mathbf{x})\in V^k,\\
&\int_\Omega w_h(\mathbf{x})\nabla\cdot\mathbf{u}_h(\mathbf{x},t_m^p)d\mathbf{x} = \int_\Omega q(\mathbf{x},t_m^p)w_h(\mathbf{x})d\mathbf{x} \qquad \forall w_h(\mathbf{x})\in S_p^k.
\end{aligned} \tag{3.12}$$

3.3. An ELLAM-MFEM time-stepping procedure. The velocity field usually changes less rapidly than the concentration. Moreover, at each time step the MFEM system (3.12) is more expensive to solve than the ELLAM scheme for the transport PDE (2.2). Therefore, we allow a larger time step for the pressure than that for the concentration [9]. It is often computationally convenient to define the time partition (3.1) for the concentration by subdividing the time partition (3.11) for the pressure. Namely, there exist $0 =: N_0 < N_1 < \cdots < N_m < \cdots < N_{M-1} < N_M := N$ such that $t_{N_m}^c = t_m^p$ for $m = 1,2,\ldots,M$. For $n = N_{m-1}+1, N_{m-1}+2,\ldots,N_m$, the concentration time step t_n^c relates to the pressure time steps by $t_{m-1}^p < t_n^c \leq t_m^p$. In the ELLAM scheme we define a velocity approximation $\mathbf{u}_h^{\mathrm{E}}(\mathbf{x},t_n^c)$ by an extrapolation of $\mathbf{u}_h(\mathbf{x},t_{m-1}^p)$ and earlier values [9]

$$\mathbf{u}_h^{\mathrm{E}}(\mathbf{x},t_n^c) := \begin{cases} \left(1+\dfrac{t_n^c - t_{m-1}^p}{\Delta t_{m-1}^p}\right)\mathbf{u}_h(\mathbf{x},t_{m-1}^p) - \dfrac{t_n^c - t_{m-1}^p}{\Delta t_{m-1}^p}\mathbf{u}_h(\mathbf{x},t_{m-2}^p), \\ \qquad N_{m-1}+1\leq n\leq N_m, \quad 2\leq m\leq M, \\ \mathbf{u}_h(\mathbf{x},0), \quad 1\leq n\leq N_1, \quad m=1. \end{cases} \tag{3.13}$$

Let $S_c^l \subset W_\infty^1(\Omega)$ be an FEM space, which contains the space of continuous piecewise polynomials of degree at most l on a quasi-uniform partition of diameter h_c. For example, the FEM space S_c^l could be the space of piecewise polynomials of degree at most l on triangular elements or the space of piecewise polynomials of degree at most l in each coordinate direction on rectangular elements. We often utilize the fact that velocity is smoother than the concentration to use a much larger grid size h_p than h_c and to further reduce computational cost since (3.12) is more expensive to solve than (3.14).

Let $c_h(\mathbf{x},0)$ be an approximation to $c_0(\mathbf{x})$ (e.g., its L^2 or Ritz projection, or interpolation). An ELLAM-MFEM time-stepping procedure is formulated as follows: For $m = 1,\ldots,M$, solve the MFEM scheme (3.12) at the pressure time step t_{m-1}^p. For $n = N_{m-1}+1, N_{m-1}+2,\ldots,N_m$, solve the following ELLAM scheme at each

concentration time step t_n^c: Find $c_h(\mathbf{x}, t_n^c) \in S_c^l$ such that for all $z_h(\mathbf{x}, t_n^c) \in S_c^l$

$$\begin{aligned}&\int_\Omega \phi(\mathbf{x}) c_h(\mathbf{x}, t_n^c) z_h(\mathbf{x}, t_n^c) d\mathbf{x} \\ &+ \Delta t_n^c \int_\Omega \nabla z_h(\mathbf{x}, t_n^c) \cdot \mathbf{D}(\mathbf{x}, \mathbf{u}_h^E(\mathbf{x}, t_n^c)) \nabla c_h(\mathbf{x}, t_n^c) d\mathbf{x} \\ &= \int_\Omega \phi(\mathbf{x}) c_h(\mathbf{x}, t_{n-1}^c) z_h(\mathbf{x}, t_{n-1}^{c,+}) d\mathbf{x} + \Delta t_n^c \int_\Omega \bar{c}_h(\mathbf{x}, t_n^c) q(\mathbf{x}, t_n^c) z_h(\mathbf{x}, t_n^c) d\mathbf{x}.\end{aligned} \tag{3.14}$$

All of the integrals in (3.14), except for the first on the right-hand side, are in Eulerian coordinates and can be evaluated as in FEMs. The first term on the right-hand side of (3.14) is in Lagrangian coordinates and requires extra attention. By definition, the test functions z in the ELLAM reference equation (3.9) are constant along the characteristics $\mathbf{r}$ defined by the initial-value problem (3.4). Set

$$\tilde{\mathbf{x}} := \mathbf{r}(t_n^c; \mathbf{x}, t_{n-1}^c) = \mathbf{x} + \int_{t_{n-1}^c}^{t_n^c} \frac{\mathbf{u}(\mathbf{r}(\theta; \mathbf{x}, t_{n-1}^c), \theta)}{\phi(\mathbf{r}(\theta; \mathbf{x}, t_{n-1}^c))} d\theta; \tag{3.15}$$

the test function $z(\mathbf{x}, t_{n-1}^c)$ in (3.9) is determined by

$$z(\mathbf{x}, t_{n-1}^{c,+}) = z(\tilde{\mathbf{x}}, t_n^c). \tag{3.16}$$

Unlike in linear advection-diffusion PDEs [6, 26], the velocity $\mathbf{u}$ is unknown. In (3.14) we compute an approximate characteristic $\mathbf{r}_h(t_n^c; \mathbf{x}, t_{n-1}^c)$ by solving a numerical analogue of problem (3.4)

$$\frac{d\mathbf{r}_h}{d\theta} = \frac{\mathbf{u}_h^{\mathrm{E}}(\mathbf{r}_h, \theta)}{\phi(\mathbf{r}_h)} \quad \text{and} \quad \mathbf{r}_h(\theta; \bar{\mathbf{x}}, \bar{t})\Big|_{\theta=\bar{t}} = \bar{\mathbf{x}}. \tag{3.17}$$

For clarity of exposition, we assume an Euler tracking and set

$$\begin{aligned}\tilde{\mathbf{x}}_h := \mathbf{r}_h(t_n^c; \mathbf{x}, t_{n-1}^c) &= \mathbf{x} + \int_{t_{n-1}^c}^{t_n^c} \frac{\mathbf{u}_h^{\mathrm{E}}(\mathbf{r}_h(\theta; \mathbf{x}, t_{n-1}^c), \theta)}{\phi(\mathbf{r}_h(\theta; \mathbf{x}, t_{n-1}^c))} d\theta \\ &= \mathbf{x} + \frac{\mathbf{u}_h^{\mathrm{E}}(\mathbf{x}, t_{n-1}^c)}{\phi(\mathbf{x})} \Delta t_n^c.\end{aligned} \tag{3.18}$$

The test functions $z_h(\mathbf{x}, t_{n-1}^c)$ in the ELLAM scheme (3.14) are evaluated by

$$z_h(\mathbf{x}, t_{n-1}^{c,+}) = z_h(\tilde{\mathbf{x}}_h, t_n^c). \tag{3.19}$$

This further complicates the analysis for the ELLAM-MFEM procedure.

4. Preliminaries. This section cites well-established estimates to be used in the proof of the main theorem in section 5.

The finite element space S_c^l has the approximation and inverse properties [7] for $1 \le m \le l+1, 1 \le p, q \le \infty$

$$\inf_{z_h \in S_c^l} (\|z - z_h\|_{L^q} + h_c \|z - z_h\|_{W_q^1}) \le A_1 h_c^{m+1+(\frac{2}{q} - \frac{2}{p})} \|z\|_{W_p^m} \ \forall z \in W_p^m(\Omega) \tag{4.1}$$

and

$$\begin{aligned}&\|z_h\|_{H^1} \le K_1 h_c^{-1} \|z_h\|_{L^2}, \qquad \|z_h\|_{L^\infty} \le K_1 \, |\ln h_c|^{1/2} \|z_h\|_{H^1}, \\ &\|z_h\|_{W_q^m} \le K_1 h_c^{-(1-\frac{2}{q})} \|z_h\|_{H^m} \quad \forall z_h \in S_c^l, \ m = 0, 1.\end{aligned} \tag{4.2}$$

The MFEM spaces (V^k, S_p^k) have the approximation and inverse properties [4, 5, 7, 19] for $2 \le q \le +\infty$ and $1 \le m \le k+1$

$$
\begin{aligned}
&\inf_{\mathbf{v}_h \in V^k} \|\mathbf{v} - \mathbf{v}_h\|_{L^q} \le A_2 h_p^{m-(1-\frac{2}{q})} \|\mathbf{v}\|_{H^m} && \forall \mathbf{v} \in H^m, \\
&\inf_{\mathbf{v}_h \in V^k} \|\mathbf{v} - \mathbf{v}_h\|_{H(\mathrm{div})} \le A_2 h_p^m \|\mathbf{v}\|_{H^m(\mathrm{div})} && \forall \mathbf{v} \in H^m(\mathrm{div}), \\
&\inf_{g_h \in S_p^k} \|g - g_h\|_{L^2} \le A_2 h_p^m \|g\|_{H^m} && \forall g \in H^m
\end{aligned} \tag{4.3}
$$

and

$$
\begin{aligned}
&\|\mathbf{v}_h\|_{L^q} \le K_2 h_p^{-(1-\frac{2}{q})} \|\mathbf{v}_h\|_{L^2} && \forall \mathbf{v}_h \in V^k, \\
&\|\mathbf{v}_h\|_{W_q^1} \le K_2 h_p^{-1} \|\mathbf{v}_h\|_{L^q} && \forall \mathbf{v}_h \in V^k.
\end{aligned} \tag{4.4}
$$

It is understood in (4.4) that

$$
\|\mathbf{v}_h\|_{W_q^1} := \begin{cases} \displaystyle\left(\sum_{\Omega_e^p \subset \Omega} \|\mathbf{v}_h\|_{W_q^1(\Omega_e^p)}^q \right)^{1/q} & \text{for } \ 2 \le q < +\infty, \\ \displaystyle\max_{\forall \Omega_e^p \subset \Omega} \|\mathbf{v}_h\|_{W_q^1(\Omega_e^p)} & \text{for } \ q = +\infty, \end{cases}
$$

where $\Omega_e^p \subset \Omega$ are all of the elements of the pressure mesh.

Let $\Pi c(\mathbf{x}, t) \in S_c^l,\ t \in [0, T]$, be the Ritz projection of $c(\mathbf{x}, t)$ defined by [23, 33]

$$
\begin{aligned}
&\int_\Omega \chi(\mathbf{x})(\Pi c(\mathbf{x}, t) - c(\mathbf{x}, t)) d\mathbf{x} \\
&\quad + \int_\Omega \nabla \chi(\mathbf{x}) \cdot \mathbf{D}(\mathbf{x}, \mathbf{u}(\mathbf{x}, t)) \nabla (\Pi c(\mathbf{x}, t) - c(\mathbf{x}, t)) d\mathbf{x} = 0 \quad \forall \chi \in S_c^l.
\end{aligned} \tag{4.5}
$$

The following estimates hold [7, 19, 23, 33] for $2 \le q \le +\infty,\ 1 \le m \le l+1$:

$$
\begin{aligned}
\|\Pi c - c\|_{L^\infty(L^q)} + h_c \|\Pi c - c\|_{L^\infty(W_q^1)} &\le A_1 h_c^m \|c\|_{L^\infty(W_q^m)}, \\
\|\Pi c - c\|_{H^1(L^q)} &\le A_1 h_c^m \|c\|_{H^1(W_q^m)}.
\end{aligned} \tag{4.6}
$$

Here the constant A_1 is independent of c and h_c.

Let $\mathrm{I}c(\mathbf{x}, t) \in S_c^l,\ t \in [0, T]$, be the interpolation of $c(\mathbf{x}, t)$. We use the estimates (4.1) with $p = 2$ and $q = +\infty$, (4.2) with $q = +\infty$, and (4.6) with $q = 2$ to conclude that for $c \in L^\infty(W_\infty^1 \cap H^2)$

$$
\begin{aligned}
\|\Pi c\|_{L^\infty(W_\infty^1)} &\le \|\Pi c - \mathrm{I}c\|_{L^\infty(W_\infty^1)} + \|\mathrm{I}c - c\|_{L^\infty(W_\infty^1)} + \|c\|_{L^\infty(W_\infty^1)} \\
&\le K_1 h_c^{-1} \|\Pi c - \mathrm{I}c\|_{L^\infty(H^1)} + (A_1 + 1) \|c\|_{L^\infty(W_\infty^1)} \\
&\le K_1 h_c^{-1} \big(\|\Pi c - c\|_{L^\infty(H^1)} + \|\mathrm{I}c - c\|_{L^\infty(H^1)} \big) \\
&\quad + (A_1 + 1) \|c\|_{L^\infty(W_\infty^1)} \\
&\le 2 A_1 K_1 \|c\|_{L^\infty(H^2)} + (A_1 + 1) \|c\|_{L^\infty(W_\infty^1)} =: K_3.
\end{aligned} \tag{4.7}
$$

We similarly define a mapping $(\Pi\mathbf{u}, \Pi p)$ from $H(\mathrm{div}) \times L_0^2$ to $V^k \times S_p^k$ by

$$\begin{aligned}
&\int_\Omega \mu(c(\mathbf{x},t))\mathbf{K}^{-1}(\mathbf{x})(\Pi\mathbf{u}(\mathbf{x},t) - \mathbf{u}(\mathbf{x},t)) \cdot \mathbf{v}_h(\mathbf{x})d\mathbf{x} \\
&\qquad - \int_\Omega (\Pi p(\mathbf{x},t) - p(\mathbf{x},t))\nabla \cdot \mathbf{v}_h(\mathbf{x})d\mathbf{x} = 0 \quad \forall \mathbf{v}_h \in V^k, \\
&\int_\Omega w_h(\mathbf{x})\nabla \cdot (\Pi\mathbf{u}(\mathbf{x},t) - \mathbf{u}(\mathbf{x},t))d\mathbf{x} = 0 \qquad \forall w_h \in S_p^k.
\end{aligned} \tag{4.8}$$

The following estimates hold, e.g., for Raviart–Thomas spaces [4, 8, 9]:

$$\begin{aligned}
&\|\Pi\mathbf{u} - \mathbf{u}\|_{L^\infty(H(\mathrm{div}))} + \|\Pi p - p\|_{L^\infty(L^2)} \\
&\quad \le A\left(\inf_{\mathbf{v}_h \in V^k} \|\mathbf{u} - \mathbf{v}_h\|_{L^\infty(H(\mathrm{div}))} + \inf_{g_h \in S_p^k} \|p - g_h\|_{L^\infty(L^2)}\right) \\
&\quad \le A_2 h_p^{k+1}(\|\mathbf{u}\|_{L^\infty(H^{k+1}(\mathrm{div}))} + \|p\|_{L^\infty(H^{k+1})}).
\end{aligned} \tag{4.9}$$

Here A_2 is independent of h_p, u, p, and c.

We let I be an interpolation operator from $H(\mathrm{div}) \times L_0^2$ to $V^k \times S_p^k$ for $t \in [0,T]$. We use the estimates (4.4) and (4.9) to conclude that

$$\begin{aligned}
\|\Pi\mathbf{u}\|_{L^\infty(L^\infty)} &\le \|\Pi\mathbf{u} - \mathrm{I}\mathbf{u}\|_{L^\infty(L^\infty)} + \|\mathrm{I}\mathbf{u} - \mathbf{u}\|_{L^\infty(L^\infty)} + \|\mathbf{u}\|_{L^\infty(L^\infty)} \\
&\le K_2 h_p^{-1}\|\Pi\mathbf{u} - \mathrm{I}\mathbf{u}\|_{L^\infty(L^2)} + (A_2+1)\|\mathbf{u}\|_{L^\infty(L^\infty)} \\
&\le K_2 h_p^{-1}\big(\|\Pi\mathbf{u} - \mathbf{u}\|_{L^\infty(L^2)} + \|\mathrm{I}\mathbf{u} - \mathbf{u}\|_{L^\infty(L^2)}\big) \\
&\quad + (A_2+1)\|\mathbf{u}\|_{L^\infty(L^\infty)} \\
&\le 2A_2K_2\big(\|\mathbf{u}\|_{L^\infty(H^1(\mathrm{div}))} + \|p\|_{L^\infty(H^1)}\big) \\
&\quad + (A_2+1)\|\mathbf{u}\|_{L^\infty(L^\infty)} =: K_4.
\end{aligned} \tag{4.10}$$

For the analysis in section 5 we introduce an extrapolation of the exact velocity $\mathbf{u}$

$$\mathbf{u}^{\mathrm{E}}(\mathbf{x}, t_n^c) := \begin{cases} \left(1 + \dfrac{t_n^c - t_{m-1}^p}{\Delta t_{m-1}^p}\right)\mathbf{u}(\mathbf{x}, t_{m-1}^p) - \dfrac{t_n^c - t_{m-1}^p}{\Delta t_{m-1}^p}\mathbf{u}(\mathbf{x}, t_{m-2}^p), \\ \qquad N_{m-1}+1 \le n \le N_m, \quad 2 \le m \le M, \\ \mathbf{u}(\mathbf{x},0), \quad 1 \le n \le N_1, \quad m = 1. \end{cases} \tag{4.11}$$

Then we routinely see that for $2 \le q \le +\infty$

$$\begin{aligned}
&\|\mathbf{u}^E(\mathbf{x},t) - \mathbf{u}(\mathbf{x},t)\|_{L^q} \\
&\quad \le \begin{cases} A_3(\Delta t_p)^{\frac{3}{2}}\|\mathbf{u}\|_{H^2(t_{m-2}^p, t_m^p; L^q)} & \forall t \in [t_{m-1}^p, t_m^p],\ m \ge 2, \\ A_3\Delta t_p^1\|\mathbf{u}\|_{W_1^\infty(t_0^p, t_1^p; L^q)} & \forall t \in [t_0^p, t_1^p], \quad m = 1. \end{cases}
\end{aligned} \tag{4.12}$$

5. An optimal-order error estimate. In this section we prove an optimal-order error estimate for the ELLAM-MFEM time-stepping procedure with any order of approximating polynomials ($k \ge 0$, $l \ge 1$).

5.1. On the main theorem and its proof. Note that the test functions in the ELLAM formulation are defined by (3.3), which can be evaluated by using (3.19). However, the velocity $\mathbf{u}$ is unknown. This is in contrast to the ELLAM schemes for

linear transport PDEs in which the true velocity is known and is used to evaluate characteristics [6, 21, 26]. In the ELLAM-MFEM time-stepping procedure in section 3, we have to use the numerical velocity $\mathbf{u}_h^{\mathrm{E}}$ to evaluate the approximate characteristics $\mathbf{r}_h$ by solving problem (3.17). These issues further complicate the analysis [21, 25].

THEOREM 5.1. *Suppose that the solution* $(c, p, \mathbf{u})$ *of problem* (2.1)–(2.3) *satisfies* $c \in L^\infty(W^{l+1}_{2+\gamma}) \cap L^\infty(W^1_\infty) \cap H^1(H^{l+1})$ *for some* $\gamma > 0$, $p \in L^\infty(H^{k+1})$, *and* $\mathbf{u} \in L^\infty(H^{k+1}(\mathrm{div}) \cap W^1_\infty) \cap W^1_\infty(L^\infty) \cap H^2(L^2)$. *Let* $(c_h(\mathbf{x}, t_n^c), p_h(\mathbf{x}, t_m^p), \mathbf{u}_h(\mathbf{x}, t_m^p))$ *be the solution of the ELLAM-MFEM time-stepping procedure* (3.12) *and* (3.14) *with* $l \geq 1$ *and* $k \geq 0$. *Assume that the discretization parameters obey the relations*

$$\Delta t_c = O(h_p), \quad h_c^{l+1} = o(h_p), \quad \Delta t_p^1 = O(h_p^{2/3}), \quad \Delta t_p = O(h_p^{1/2}). \tag{5.1}$$

There exist positive constants h_c^*, h_p^*, Δt_c^*, Δt_p^*, *and* Q^* *such that the following optimal-order error estimate holds for* $0 < h_c \leq h_c^*$, $0 < h_p \leq h_p^*$, $0 < \Delta t_c \leq \Delta t_c^*$, *and* $0 < \Delta t_p \leq \Delta t_p^*$:

$$\begin{aligned}
&\|c_h - c\|_{\hat{L}_c^\infty(L^2)} + |\mathbf{D}|_{min}^{1/2} \|c_h - c\|_{\hat{L}_c^2(H^1)} \\
&\quad + \|\mathbf{u}_h - \mathbf{u}\|_{\hat{L}_p^\infty(H(\mathrm{div}))} + \|p_h - p\|_{\hat{L}_p^\infty(L^2)} \\
&\leq Q^* \Delta t_c \left(\|\bar{c}\|_{L^2(L^2)} + \|c\|_{L^2(H^1)} + \left\| \frac{d\bar{c}}{dt} \right\|_{L^2(L^2)} + \left\| \frac{dc}{dt} \right\|_{L^2(H^1)} \right) \\
&\quad + Q^* ((\Delta t_p^1)^{3/2} + (\Delta t_p)^2) \|\mathbf{u}\|_{H^2(L^2)} + Q^* h_c^{l+1} (\|c\|_{L^\infty(W^{l+1}_{2+\gamma})} \\
&\quad + \|c\|_{H^1(H^{l+1})}) + Q^* h_p^{k+1} (\|\mathbf{u}\|_{L^\infty(H^{k+1}(\mathrm{div}))} + \|p\|_{L^\infty(H^{k+1})}).
\end{aligned} \tag{5.2}$$

Here $\frac{d}{dt}$ *refers to the material derivative along the characteristics defined in* (3.4). *The constant* $Q^* = Q^*(h_c^*, h_p^*, \Delta t_c^*, \Delta t_p^*, T)$, *but* Q^* *is independent of the discretization parameters* h_c, h_p, Δt_c, *or* Δt_p.

To prove the theorem, we use (3.10), (3.12), and (4.8) to derive a relation

$$\begin{aligned}
&\int_\Omega \mu(c_h(\mathbf{x}, t_m^p)) \mathbf{K}^{-1}(\mathbf{x}) (\mathbf{u}_h(\mathbf{x}, t_m^p) - \Pi \mathbf{u}(\mathbf{x}, t_m^p)) \cdot \mathbf{v}_h(\mathbf{x}) d\mathbf{x} \\
&\qquad - \int_\Omega (p_h(\mathbf{x}, t_m^p) - \Pi p(\mathbf{x}, t_m^p)) \nabla \cdot \mathbf{v}_h(\mathbf{x}) d\mathbf{x} \\
&\qquad = \int_\Omega (\mu(c(\mathbf{x}, t_m^p)) - \mu(c_h(\mathbf{x}, t_m^p))) \mathbf{K}^{-1}(\mathbf{x}) \Pi \mathbf{u}(\mathbf{x}, t_m^p) \cdot \mathbf{v}_h(\mathbf{x}) d\mathbf{x}, \\
&\int_\Omega w_h(\mathbf{x}) \nabla \cdot (\mathbf{u}_h(\mathbf{x}, t_m^p) - \Pi \mathbf{u}(\mathbf{x}, t_m^p)) d\mathbf{x} = 0 \quad \forall (\mathbf{v}_h, w_h) \in V^k \times S_p^k.
\end{aligned}$$

Combining this equation with (4.10) yields an estimate [4, 5, 8]

$$\begin{aligned}
&\|\mathbf{u}_h(\mathbf{x}, t_m^p) - \Pi \mathbf{u}(\mathbf{x}, t_m^p)\|_{H(\mathrm{div})} + \|p_h(\mathbf{x}, t_m^p) - \Pi p(\mathbf{x}, t_m^p)\| \\
&\quad \leq Q(1 + \|\Pi \mathbf{u}(\mathbf{x}, t_m^p)\|_{L^\infty}) \, \|c_h(\mathbf{x}, t_m^p) - c(\mathbf{x}, t_m^p)\| \\
&\quad \leq Q_1 \|c_h(\mathbf{x}, t_m^p) - c(\mathbf{x}, t_m^p)\|, \qquad 0 \leq m \leq M.
\end{aligned} \tag{5.3}$$

For convenience, we have dropped the subscript L^2.

The estimates (4.9) and (5.3) show that the bound on $\|\mathbf{u}_h-\mathbf{u}\|_{\hat{L}_p^\infty(H(\mathrm{div}))}+\|p_h-p\|_{\hat{L}_p^\infty(L^2)}$ in (5.2) is a consequence of the bound on $\|c_h-c\|_{\hat{L}_c^\infty(L^2)}$.

5.2. Proof of the key estimate. To bound $\|c_h-c\|_{\hat{L}_c^\infty(L^2)}$, we set $\xi(\mathbf{x},t_n^c):=c_h(\mathbf{x},t_n^c)-\Pi c(\mathbf{x},t_n^c)$ and $\eta(\mathbf{x},t_n^c):=\Pi c(\mathbf{x},t_n^c)-c(\mathbf{x},t_n^c)$. Note that $c_h-c=\xi+\eta$ and that the estimate for η is known from (4.6). The key to prove the theorem is to derive an estimate of the form (5.2) for ξ.

Proof. We subtract (3.14) from (3.9) for $n=N_{m-1}+1,\dots,N_m$ to generate an error equation for any $z_h(\mathbf{x},t_n^c)\in S_c^l$

$$\begin{aligned}
&\int_\Omega \phi(\mathbf{x})(c_h(\mathbf{x},t_n^c)-c(\mathbf{x},t_n^c))z_h(\mathbf{x},t_n^c)d\mathbf{x}\\
&\quad+\Delta t_n^c\int_\Omega \nabla z_h(\mathbf{x},t_n^c)\cdot(\mathbf{D}(\mathbf{x},\mathbf{u}_h^E(\mathbf{x},t_n^c))\nabla c_h(\mathbf{x},t_n^c)-\mathbf{D}(\mathbf{x},\mathbf{u}(\mathbf{x},t_n^c))\nabla c(\mathbf{x},t_n^c))d\mathbf{x}\\
&=\int_\Omega \phi(\mathbf{x})(c_h(\mathbf{x},t_{n-1}^c)-c(\mathbf{x},t_{n-1}^c))z_h(\mathbf{x},t_{n-1}^{c,+})d\mathbf{x}\\
&\quad+\Delta t_n^c\int_\Omega(\bar{c}_h(\mathbf{x},t_n^c)-\bar{c}(\mathbf{x},t_n^c))q(\mathbf{x},t_n^c)z_h(\mathbf{x},t_n^c)d\mathbf{x}\\
&\quad+E_{\mathbf{D}}(c,z_h)-E_q(\bar{c},z_h).
\end{aligned}$$

In the second term on the right-hand side, $\bar{c}_h:=\bar{c}$ is specified at sources so the difference $\bar{c}_h-\bar{c}$ vanishes. At sinks $q=q^-$ is specified but $\bar{c}_h=c_h$ and $\bar{c}=c$. We rewrite the preceding equation with $z_h=\xi$ as

$$\begin{aligned}
&\int_\Omega \phi(\mathbf{x})\xi^2(\mathbf{x},t_n^c)d\mathbf{x}+\Delta t_n^c\int_\Omega \nabla\xi(\mathbf{x},t_n^c)\cdot\mathbf{D}(\mathbf{x},\mathbf{u}_h^E(\mathbf{x},t_n^c))\nabla\xi(\mathbf{x},t_n^c)d\mathbf{x}\\
&\quad=\int_\Omega \phi(\mathbf{x})\xi(\mathbf{x},t_{n-1}^c)\xi(\mathbf{x},t_{n-1}^{c,+})d\mathbf{x}+\int_\Omega \phi(\mathbf{x})\eta(\mathbf{x},t_{n-1}^c)\xi(\mathbf{x},t_{n-1}^{c,+})d\mathbf{x}\\
&\qquad+\Delta t_n^c\int_\Omega q^-(\mathbf{x},t_n^c)\xi^2(\mathbf{x},t_n^c)d\mathbf{x}+\Delta t_n^c\int_\Omega q^-(\mathbf{x},t_n^c)\eta(\mathbf{x},t_n^c)\xi(\mathbf{x},t_n^c)d\mathbf{x}\\
&\qquad-\Delta t_n^c\int_\Omega \nabla\xi(\mathbf{x},t_n^c)\cdot(\mathbf{D}(\mathbf{x},\mathbf{u}_h^E(\mathbf{x},t_n^c))-\mathbf{D}(\mathbf{x},\mathbf{u}(\mathbf{x},t_n^c)))\nabla\Pi c(\mathbf{x},t_n^c)d\mathbf{x}\\
&\qquad-\Delta t_n^c\int_\Omega \nabla\xi(\mathbf{x},t_n^c)\cdot\mathbf{D}(\mathbf{x},\mathbf{u}(\mathbf{x},t_n^c))\nabla\eta(\mathbf{x},t_n^c)d\mathbf{x}\\
&\qquad-\int_\Omega \phi(\mathbf{x})\eta(\mathbf{x},t_n^c)\xi(\mathbf{x},t_n^c)d\mathbf{x}+E_{\mathbf{D}}(c,\xi)-E_q(\bar{c},\xi).
\end{aligned}\tag{5.4}$$

We bound the third and fourth terms on the right-hand side of (5.4) by

$$\begin{aligned}
&\left|\Delta t_n^c\int_\Omega q^-(\mathbf{x},t_n^c)\xi^2(\mathbf{x},t_n^c)d\mathbf{x}+\Delta t_n^c\int_\Omega q^-(\mathbf{x},t_n^c)\eta(\mathbf{x},t_n^c)\xi(\mathbf{x},t_n^c)d\mathbf{x}\right|\\
&\quad\le Q\Delta t_n^c\|\xi(\mathbf{x},t_n^c)\|^2+\Delta t_n^c\|\eta(\mathbf{x},t_n^c)\|^2\\
&\quad\le Q\Delta t_n^c\|\xi(\mathbf{x},t_n^c)\|^2+A_1^2\Delta t_n^c h_c^{2l+2}\|c\|^2_{L^\infty(H^{l+1})}.
\end{aligned}$$

We use (4.5) and (4.6) to bound the sixth term on the right-hand side of (5.4)

$$\begin{aligned}
&\left|\Delta t_n^c \int_\Omega \nabla\xi(\mathbf{x},t_n^c)\cdot\mathbf{D}(\mathbf{x},\mathbf{u}(\mathbf{x},t_n^c))\nabla\eta(\mathbf{x},t_n^c)d\mathbf{x}\right| \\
&\quad= \left|\Delta t_n^c \int_\Omega \xi(\mathbf{x},t_n^c)\eta(\mathbf{x},t_n^c)d\mathbf{x}\right| \\
&\quad\le Q\Delta t_n^c\|\xi(\mathbf{x},t_n^c)\|^2 + \Delta t_n^c\|\eta(\mathbf{x},t_n^c)\|^2 \\
&\quad\le Q\Delta t_n^c\|\xi(\mathbf{x},t_n^c)\|^2 + A_1^2\Delta t_n^c h_c^{2l+2}\|c\|^2_{L^\infty(H^{l+1})}.
\end{aligned}$$

The fifth term on the right-hand side of (5.4) is bounded in Lemma 6.3

$$\begin{aligned}
&\left|\Delta t_n^c \int_\Omega \nabla\xi(\mathbf{x},t_n^c)\cdot(\mathbf{D}(\mathbf{x},\mathbf{u}_h^E(\mathbf{x},t_n^c)) - \mathbf{D}(\mathbf{x},\mathbf{u}(\mathbf{x},t_n^c)))\nabla\Pi c(\mathbf{x},t_n^c)d\mathbf{x}\right| \\
&\quad\le \varepsilon\Delta t_n^c\|\nabla\xi(\mathbf{x},t_n^c)\|^2 + Q\Delta t_n^c(\|\xi(\mathbf{x},t_{m-1}^p)\|^2 + \|\xi(\mathbf{x},t_{m-2}^p)\|^2) \\
&\qquad + Q\Delta t_n^c\Big(h_c^{2l+2}\|c\|^2_{L^\infty(H^{l+1})} + h_p^{2k+2}(\|\mathbf{u}\|^2_{L^\infty(H^{k+1}(\mathrm{div}))} + \|p\|^2_{L^\infty(H^{k+1})}) \\
&\qquad + \delta_{m,1}(\Delta t_p^1)^2\|\mathbf{u}\|^2_{W_1^\infty(0,t_1^p;L^2)} + (1-\delta_{m,1})(\Delta t_p)^3\|\mathbf{u}\|^2_{H^2(t_{m-2}^p,t_m^p;L^2)}\Big),
\end{aligned} \tag{5.5}$$

where $\delta_{i,j}=1$ if $i=j$ or 0 otherwise.

Note that the evaluation of source and sink terms and the diffusion-dispersion term in the ELLAM scheme (3.14) does not involve any characteristic tracking. Hence, (3.6) and (3.8) still hold, leading to the estimate

$$\begin{aligned}
|E_q(\bar c,\xi)| &= \left|\int_\Omega \xi(\mathbf{x},t_n^c)\int_{t_{n-1}^c}^{t_n^c}\left\{\int_\theta^{t_n^c}\frac{d}{dt}(\bar c(\mathbf{r}(t;\mathbf{x},t_n^c),t)q(\mathbf{r}(t;\mathbf{x},t_n^c),t))dt\right.\right. \\
&\qquad \left.\left. + \bar c(\mathbf{r}(\theta;\mathbf{x},t_n^c),\theta)q(\mathbf{r}(\theta;\mathbf{x},t_n^c),\theta)\left[\det\left(\frac{\partial\mathbf{r}(\theta;\mathbf{x},t_n^c)}{\partial\mathbf{x}}\right)-1\right]\right\}d\theta d\mathbf{x}\right| \\
&\le Q\Delta t_n^c\|\xi(\mathbf{x},t_n^c)\|^2 + Q(\Delta t_n^c)^2\left(\|\bar c\|^2_{L^2(t_{n-1}^c,t_n^c;L^2)} + \left\|\frac{d\bar c}{dt}\right\|^2_{L^2(t_{n-1}^c,t_n^c;L^2)}\right).
\end{aligned}$$

We use the estimates (6.1) and (6.2) in Lemma 6.1 to bound $|E_{\mathbf{D}}(c,\xi)|$ by

$$\begin{aligned}
|E_{\mathbf{D}}(c,\xi)| &\le \left|\int_\Omega\int_{t_{n-1}^c}^{t_n^c}\left(\det\left(\frac{\partial\mathbf{r}(\theta;\mathbf{x},t_n^c)}{\partial\mathbf{x}}\right)\left(\frac{\partial\mathbf{r}(\theta;\mathbf{x},t_n^c)}{\partial\mathbf{x}}\right)^{-1} - \mathbf{I}\right)\nabla\xi(\mathbf{x},t_n^c)\right. \\
&\qquad \left.\cdot(\mathbf{D}(\mathbf{r}(\theta;\mathbf{x},t_n^c),\mathbf{u}(\mathbf{r}(\theta;\mathbf{x},t_n^c),\theta))\nabla c(\mathbf{r}(\theta;\mathbf{x},t_n^c),\theta))d\theta d\mathbf{x}\right| \\
&\quad + \left|\int_\Omega\int_{t_{n-1}^c}^{t_n^c}\nabla\xi(\mathbf{x},t_n)\cdot(\mathbf{D}(\mathbf{x},\mathbf{u}(\mathbf{x},t_n^c))\nabla c(\mathbf{x},t_n^c)\right. \\
&\qquad \left. - \mathbf{D}(\mathbf{r}(\theta;\mathbf{x},t_n^c),\mathbf{u}(\mathbf{r}(\theta;\mathbf{x},t_n^c),\theta))\nabla c(\mathbf{r}(\theta;\mathbf{x},t_n^c),\theta))d\theta d\mathbf{x}\right| \\
&\le \varepsilon\Delta t_n^c\|\nabla\xi(\mathbf{x},t_n^c)\|^2 + Q(\Delta t_n^c)^2\left(\|c\|^2_{L^2(t_{n-1}^c,t_n^c;H^1)} + \left\|\frac{dc}{dt}\right\|^2_{L^2(t_{n-1}^c,t_n^c;H^1)}\right).
\end{aligned}$$

The first term on the right-hand side of (5.4) is estimated in Lemma 6.4

$$\begin{aligned}
&\left|\int_\Omega \phi(\mathbf{x})\xi(\mathbf{x},t_{n-1}^c)\xi(\mathbf{x},t_{n-1}^{c,+})d\mathbf{x}\right| \\
&\quad\le \frac{1}{2}\int_\Omega \phi(\mathbf{x})\xi^2(\mathbf{x},t_{n-1}^c)d\mathbf{x} + \frac{1}{2}\int_\Omega \phi(\mathbf{x})\xi^2(\mathbf{x},t_n^c)d\mathbf{x} + \varepsilon\Delta t_n^c(\|\nabla\xi(\mathbf{x},t_n^c)\|^2 \\
&\qquad + \|\nabla\xi(\mathbf{x},t_{n-1}^c)\|^2) + Q\Delta t_n^c(\|\xi(\mathbf{x},t_n^c)\|^2 + \|\xi(\mathbf{x},t_{n-1}^c)\|^2).
\end{aligned} \tag{5.6}$$

The second and seventh terms on the right-hand side of (5.4) are estimated in Lemma 6.5

$$\begin{aligned}
&\left|\int_\Omega \phi(\mathbf{x})\eta(\mathbf{x},t_n^c)\xi(\mathbf{x},t_n^c)d\mathbf{x} - \int_\Omega \phi(\mathbf{x})\eta(\mathbf{x},t_{n-1}^c)\xi(\mathbf{x},t_{n-1}^{c,+})d\mathbf{x}\right| \\
&\quad\le \varepsilon\Delta t_n^c\ \|\nabla\xi(\mathbf{x},t_n^c)\|^2 + Q\Delta t_n^c\ \|\xi(\mathbf{x},t_n^c)\|^2 \\
&\qquad + Qh_c^{2l+2}\|c\|^2_{H^1(t_{n-1}^c,t_n^c;H^{l+1})} + Q\Delta t_n^c h_c^{2l+2}\|c\|^2_{L^\infty(W^{l+1}_{2+\gamma})}.
\end{aligned} \tag{5.7}$$

We incorporate the preceding estimates into (5.4) to get

$$\begin{aligned}
&\int_\Omega \phi(\mathbf{x})\xi^2(\mathbf{x},t_n^c)d\mathbf{x} + |\mathbf{D}|_{min}\ \Delta t_n^c\|\nabla\xi(\mathbf{x},t_n^c)\|^2 \\
&\quad\le \frac{1}{2}\int_\Omega \phi(\mathbf{x})\xi^2(\mathbf{x},t_n^c)d\mathbf{x} + \frac{1}{2}\int_\Omega \phi(\mathbf{x})\xi^2(\mathbf{x},t_{n-1}^c)d\mathbf{x} + 3\varepsilon\Delta t_n^c\|\nabla\xi(\mathbf{x},t_n^c)\|^2 \\
&\qquad + \varepsilon\Delta t_n^c\|\nabla\xi(\mathbf{x},t_{n-1}^c)\|^2 + Q\Delta t_n^c\Big(\|\xi(\mathbf{x},t_n^c)\|^2 + \|\xi(\mathbf{x},t_{n-1}^c)\|^2 \\
&\qquad + \|\xi(\mathbf{x},t_{m-1}^p)\|^2 + \|\xi(\mathbf{x},t_{m-2}^p)\|^2\Big) + Q(\Delta t_n^c)^2\Big(\|\bar c\|^2_{L^2(t_{n-1}^c,t_n^c;L^2)} \\
&\qquad + \|c\|^2_{L^2(t_{n-1}^c,t_n^c;H^1)} + \Big\|\frac{d\bar c}{dt}\Big\|^2_{L^2(t_{n-1}^c,t_n^c;L^2)} + \Big\|\frac{dc}{dt}\Big\|^2_{L^2(t_{n-1}^c,t_n^c;H^1)}\Big) \\
&\qquad + Q\Delta t_n^c\Big(\delta_{m,1}(\Delta t_p^1)^2\|\mathbf{u}\|^2_{W^1_\infty(0,t_1^p;L^2)} + (1-\delta_{m,1})(\Delta t_p)^3\|\mathbf{u}\|^2_{H^2(t_{m-2}^p,t_m^p;L^2)}\Big) \\
&\qquad + Qh_c^{2l+2}\Big(\Delta t_n^c\|c\|^2_{L^\infty(W^{l+1}_{2+\gamma})} + \|c\|^2_{H^1(t_{n-1}^c,t_n^c;H^{l+1})}\Big) \\
&\qquad + Q\Delta t_n^c h_p^{2k+2}\Big(\|\mathbf{u}\|^2_{L^\infty(H^{k+1}(\mathrm{div}))} + \|p\|^2_{L^\infty(H^{k+1})}\Big).
\end{aligned} \tag{5.8}$$

We choose $\varepsilon = |\mathbf{D}|_{min}/8$, sum this estimate for $n = 1, 2, \ldots, n^*$, with $n^* \le N_m$, and cancel like terms to obtain

$$\begin{aligned}
&\int_\Omega \phi(\mathbf{x})\xi^2(\mathbf{x},t_{n^*}^c)d\mathbf{x} + |\mathbf{D}|_{min}\sum_{n=1}^{n^*}\Delta t_n^c\|\nabla\xi(\mathbf{x},t_n^c)\|^2 \\
&\quad\le Q\sum_{n=1}^{n^*}\Delta t_n^c\|\xi(\mathbf{x},t_n^c)\|^2 + Q(\Delta t_n^c)^2\Big(\|\bar c\|^2_{L^2(L^2)} + \|c\|^2_{L^2(H^1)} \\
&\qquad + \Big\|\frac{d\bar c}{dt}\Big\|^2_{L^2(L^2)} + \Big\|\frac{dc}{dt}\Big\|^2_{L^2(H^1)}\Big) + Q\big((\Delta t_p)^4 + (\Delta t_p^1)^3\big)\|\mathbf{u}\|^2_{H^2(L^2)} \\
&\qquad + Qh_c^{2l+2}\big(\|c\|^2_{L^\infty(W^{l+1}_{2+\gamma})} + \|c\|^2_{H^1(H^{l+1})}\big) \\
&\qquad + Qh_p^{2k+2}\big(\|\mathbf{u}\|^2_{L^\infty(H^{k+1}(\mathrm{div}))} + \|p\|^2_{L^\infty(H^{k+1})}\big).
\end{aligned} \tag{5.9}$$

Here we have used the estimate (6.8) to bound the initial value $\|\xi(\mathbf{x},0)\|$ that has appeared on the right-hand side.

We choose Δt_n^c small enough such that $Q\Delta t_n^c < \phi_{min}/2$ and apply Gronwall inequality to (5.9) to get

$$
\begin{aligned}
&\|\xi\|^2_{\hat{L}_c^\infty(0,t^c_{n^*};L^2)} + |\mathbf{D}|_{min}\,\|\nabla\xi\|^2_{\hat{L}_c^2(0,t^c_{n^*};L^2)} \\
&\quad \le Q(\Delta t_n^c)^2\Big(\|\bar{c}\|^2_{L^2(L^2)} + \|c\|^2_{L^2(H^1)} + \Big\|\frac{d\bar{c}}{dt}\Big\|^2_{L^2(L^2)} + \Big\|\frac{dc}{dt}\Big\|^2_{L^2(H^1)}\Big) \\
&\quad + Q\big((\Delta t_p)^4 + (\Delta t_p^1)^3\big)\|\mathbf{u}\|^2_{H^2(L^2)} + Qh_c^{2l+2}\big(\|c\|^2_{L^\infty(W^{l+1}_{2+\gamma})} + \|c\|^2_{H^1(H^{l+1})}\big) \\
&\quad + Q\,h_p^{2k+2}\big(\|\mathbf{u}\|^2_{L^\infty(H^{k+1}(\mathrm{div}))} + \|p\|^2_{L^\infty(H^{k+1})}\big).
\end{aligned}
\tag{5.10}
$$

Combining this estimate with (4.6) finishes the proof. ∎

5.3. Discussion. The error estimate (5.2) contains only the H^1 norm of c and its material derivative in the Δt_c term. These norms are much smaller than those of the temporal derivatives c due to the Lagrangian nature of the transport PDEs. Hence, larger time steps and spatial grids can be used in the ELLAM-MFEM time-stepping procedure. In other words, Theorem 5.1 theoretically justifies the mathematical and numerical strength of the ELLAM-MFEM time-stepping procedure, which was previously observed numerically [3, 17, 21, 30, 32].

In the ELLAM-MFEM time-stepping procedure (3.12) and (3.14), a linear extrapolation in time is used to define the numerical velocity field $\mathbf{u}_h^{\mathrm{E}}(\mathbf{x},t_n^c)$ in (3.13). A constant-in-time interpolation can be used instead:

$$
\mathbf{u}_h^{\mathrm{E}}(\mathbf{x},t_n^c) := \mathbf{u}_h(\mathbf{x},t_{m-1}^p) \qquad \text{for} \quad N_{m-1}+1 \le n \le N_m,\ 1 \le m \le M.
$$

The proof in Theorem 5.1 still goes through, leading to an estimate of the form

$$
\begin{aligned}
&\|c_h - c\|_{\hat{L}_c^\infty(L^2)} + \|\mathbf{u}_h - \mathbf{u}\|_{\hat{L}_p^\infty(H(\mathrm{div}))} + \|p_h - p\|_{\hat{L}_p^\infty(L^2)} \\
&\quad \le Q(h_c^{l+1} + h_p^{k+1} + \Delta t_c + \Delta t_p).
\end{aligned}
$$

An error estimate similar to Theorem 5.1 was proved in [9] for a Galerkin FEM-MFEM time-stepping procedure for problem (2.1)–(2.3) and in [13] for an MMOC-MFEM time-stepping procedure. These estimates require a restrictive condition that

$$
\Delta t_c = o(h_p). \tag{5.11}
$$

In other words, these procedures are guaranteed to converge only if the Courant number tends to zero asymptotically, which is more restrictive than the CFL condition for an explicit scheme. This is especially important for the MMOC-MFEM time-stepping procedure, since the strength of the MMOC scheme is really reflected in the large time steps the MMOC scheme allows.

In Theorem 5.1 the restriction (5.11) is relaxed to be

$$
\Delta t_c = O(h_p). \tag{5.12}
$$

This implies that the ELLAM-MFEM time-stepping procedure converges for any size of Courant numbers. Thus, Theorem 5.1 provides a solid theoretical foundation to show the competitiveness of the ELLAM-MFEM procedure, which was previously illustrated and observed numerically [21, 30, 32].

6. Auxiliary lemmas. In this section we prove several lemmas that were used in the proof of Theorem 5.1.

6.1. Estimates on tracking. In this subsection, we derive various bounds for tracking-related transformations.

LEMMA 6.1. *Assume that* $\phi \in W^1_\infty$ *and* $\mathbf{u} \in L^\infty(W^1_\infty)$. *Then the Jacobian matrix of the characteristic* $\mathbf{r}(t; x, t^c_{n-1})$ *is bounded in the Frobenius norm* $|\cdot|_2$ *by*

$$\left| \frac{\partial \mathbf{r}(t; \mathbf{x}, t^c_{n-1})}{\partial \mathbf{x}} - \mathbf{I} \right|_2 \le Q(t - t^c_{n-1}), \qquad t \in [t^c_{n-1}, t^c_n]. \tag{6.1}$$

Here Q *depends on* $\|\mathbf{u}\|_{L^\infty(W^1_\infty)}$ *but not on discretization parameters. In addition,*

$$\det\left(\frac{\partial \mathbf{r}(t; \mathbf{x}, t^c_{n-1})}{\partial \mathbf{x}} \right) = 1 + O(t - t^c_{n-1}), \qquad t \in [t^c_{n-1}, t^c_n]. \tag{6.2}$$

Proof. Differentiating (3.15) with respect to $\mathbf{x}$ yields, for $t \in [t^c_{n-1}, t^c_n]$,

$$\frac{\partial \mathbf{r}(t; \mathbf{x}, t^c_{n-1})}{\partial \mathbf{x}} - \mathbf{I} = \int_{t^c_{n-1}}^{t} \nabla \left(\frac{\mathbf{u}(\mathbf{r}(\theta; \mathbf{x}, t^c_{n-1}), \theta)}{\phi(\mathbf{r}(\theta; \mathbf{x}, t^c_{n-1}), \theta)} \right) \frac{\partial \mathbf{r}(\theta; \mathbf{x}, t^c_{n-1})}{\partial \mathbf{x}} d\theta. \tag{6.3}$$

We apply Cauchy inequality to bound the Frobenius norm of $\frac{\partial \mathbf{r}}{\partial \mathbf{x}} - \mathbf{I}$ by

$$\begin{aligned}
\left| \frac{\partial \mathbf{r}(t; \mathbf{x}, t^c_{n-1})}{\partial \mathbf{x}} - \mathbf{I} \right|^2_2 &\le (t - t^c_{n-1}) \int_{t^c_{n-1}}^{t} \left| \nabla \left(\frac{\mathbf{u}(\mathbf{r}(\theta; \mathbf{x}, t^c_{n-1}), \theta)}{\phi(\mathbf{r}(\theta; \mathbf{x}, t^c_{n-1}), \theta)} \right) \frac{\partial \mathbf{r}(\theta; \mathbf{x}, t^c_{n-1})}{\partial \mathbf{x}} \right|^2_2 d\theta \\
&\le (t - t^c_{n-1}) \left\| \frac{\mathbf{u}}{\phi} \right\|^2_{L^\infty(W^1_\infty)} \int_{t^c_{n-1}}^{t} \left| \frac{\partial \mathbf{r}(\theta; \mathbf{x}, t^c_{n-1})}{\partial \mathbf{x}} \right|^2_2 d\theta \\
&\le Q(t - t^c_{n-1})^2 + Q(t - t^c_{n-1}) \int_{t^c_{n-1}}^{t} \left| \frac{\partial \mathbf{r}(\theta; \mathbf{x}, t^c_{n-1})}{\partial \mathbf{x}} - \mathbf{I} \right|^2_2 d\theta.
\end{aligned}$$

Applying Gronwall inequality finishes the proof of (6.1), which directly gives (6.2). ∎

LEMMA 6.2. *Under the assumptions of Theorem* 5.1,

$$h_p^{-1} h_c^{l+1} \|c_0\|_{H^{l+1}} \le 1, \tag{6.4}$$

and

$$\frac{\Delta t_c}{h_p} \le Q_2, \qquad \frac{\Delta t_p}{h_p^{1/2}} \le Q_2, \qquad \frac{\Delta t_p^1}{h_p^{2/3}} \le Q_2, \tag{6.5}$$

we have the following estimate for the difference $\tilde{\mathbf{x}}_h - \tilde{\mathbf{x}}$*:*

$$\|\tilde{\mathbf{x}}_h - \tilde{\mathbf{x}}\|_{L^q} \le Q_3 \Delta t^c_n h_p^{\frac{2}{q}}. \tag{6.6}$$

Proof. We use (3.15) and (3.18) and the estimates (4.3) and (5.3) to bound $\|\tilde{\mathbf{x}}_h - \tilde{\mathbf{x}}\|_{L^q}$ for $2 \le q \le +\infty$ by

$$
\begin{aligned}
\|\tilde{\mathbf{x}}_h - \tilde{\mathbf{x}}\|_{L^q} &= \left\| \frac{\mathbf{u}_h^{\mathrm{E}}(\mathbf{x}, t_{n-1}^c)}{\phi(\mathbf{x})} \Delta t_n^c - \int_{t_{n-1}^c}^{t_n^c} \frac{\mathbf{u}(\mathbf{r}(\theta; \mathbf{x}, t_{n-1}^c), \theta)}{\phi(\mathbf{r}(\theta; \mathbf{x}, t_{n-1}^c))} d\theta \right\|_{L^q} \\
&= \left\| \Delta t_n^c \left(\frac{\mathbf{u}_h^{\mathrm{E}}(\mathbf{x}, t_{n-1}^c) - \mathbf{u}^E(\mathbf{x}, t_{n-1}^c)}{\phi(\mathbf{x})} + \frac{\mathbf{u}^E(\mathbf{x}, t_{n-1}^c) - \mathbf{u}(\mathbf{x}, t_{n-1}^c)}{\phi(\mathbf{x})} \right) \right. \\
&\qquad \left. - \int_{t_{n-1}^c}^{t_n^c} \left(\frac{\mathbf{u}(\mathbf{r}(\theta; \mathbf{x}, t_{n-1}^c), \theta)}{\phi(\mathbf{r}(\theta; \mathbf{x}, t_{n-1}^c))} - \frac{\mathbf{u}(\mathbf{x}, t_{n-1}^c)}{\phi(\mathbf{x})} \right) d\theta \right\|_{L^q} \qquad (6.7) \\
&\le 3\phi_{min}^{-1} \Delta t_n^c \Big(\|\mathbf{u}_h - \mathbf{u}\|_{\hat{L}_p^\infty(t_{m-2}^p, t_{m-1}^p; L^q)} + \delta_{m,1} \Delta t_p^1 \|\mathbf{u}\|_{W_1^\infty(t_0^p, t_1^p; L^q)} \\
&\quad + (1 - \delta_{m,1})(\Delta t_p)^{3/2} \|\mathbf{u}\|_{H^2(t_{m-2}^p, t_m^p; L^q)} + Q \Delta t_n^c \Big).
\end{aligned}
$$

Here Q depends on $\|\frac{d\mathbf{u}}{dt}\|_{L^\infty(L^\infty)}$ and $\|\mathbf{u}\|_{L^\infty(L^\infty)}$ but not on discretization parameters.

We combine (4.1), (4.6), and (5.3) to conclude that

$$
\begin{aligned}
\|\xi(\mathbf{x}, 0)\| &\le \|c_h(\mathbf{x}, 0) - \Pi c_0(\mathbf{x})\| \qquad (6.8) \\
&\le \|c_h(\mathbf{x}, 0) - c_0(\mathbf{x})\| + \|\Pi c_0(\mathbf{x}) - c_0(\mathbf{x})\| \le 2A_1 h_c^{l+1} \|c_0\|_{H^{l+1}}
\end{aligned}
$$

and

$$
\begin{aligned}
&\|(\mathbf{u}_h - \Pi \mathbf{u})(\mathbf{x}, 0)\|_{L^q} \\
&\le K_2 h_p^{-(1-\frac{2}{q})} \|(\mathbf{u}_h - \Pi \mathbf{u})(\mathbf{x}, 0)\| \le K_2 Q_1 h_p^{-(1-\frac{2}{q})} \|c_h(\mathbf{x}, 0) - c_o(\mathbf{x})\| \qquad (6.9) \\
&\le A_1 K_2 Q_1 h_p^{-(1-\frac{2}{q})} h_c^{l+1} \|c_o\|_{H^{l+1}} \le A_1 K_2 Q_1 h_p^{\frac{2}{q}},
\end{aligned}
$$

provided that condition (6.4) is true.

By (4.4) and (5.1), the estimate (6.7) reduces to the following for $n = 1, 2, \ldots, N_1$:

$$
\begin{aligned}
\|\tilde{\mathbf{x}}_h - \tilde{\mathbf{x}}\|_{L^q} &\le 3\phi_{min}^{-1} \Delta t_n^c \Big(\|(\mathbf{u}_h - \mathbf{u})(\mathbf{x}, 0)\|_{L^q} + \Delta t_p^1 \|\mathbf{u}\|_{W_1^\infty(t_0^p, t_1^p; L^q)} + Q \Delta t_n^c \Big) \\
&\le 3\phi_{min}^{-1} \Delta t_n^c \Big(A_2 K_2 h_p^{k+\frac{2}{q}} (\|\mathbf{u}\|_{L^\infty(H^{k+1}(\mathrm{div}))} + \|p\|_{L^\infty(H^{k+1})}) \qquad (6.10) \\
&\qquad + A_1 K_2 Q_1 h_p^{\frac{2}{q}} + \Delta t_p^1 \|\mathbf{u}\|_{W_1^\infty(t_0^p, t_1^p; L^q)} + Q \Delta t_n^c \Big) \\
&\le Q_4 \Delta t_n^c h_p^{\frac{2}{q}}.
\end{aligned}
$$

For $n = N_{m-1} + 1, 2, \ldots, N_m$, we use (4.4), (5.1), (5.2), and (6.5) to bound (6.7) as follows:

$$
\begin{aligned}
&\|\tilde{\mathbf{x}}_h - \tilde{\mathbf{x}}\|_{L^q} \\
&\le 3\phi_{min}^{-1} \Delta t_n^c \Big[\|\mathbf{u}_h - \mathbf{u}\|_{\hat{L}_p^\infty(t_{m-2}^p, t_{m-1}^p; L^q)} + \delta_{m,1} \Delta t_p^1 \|\mathbf{u}\|_{W_1^\infty(t_0^p, t_1^p; L^q)} \\
&\quad + (1 - \delta_{m,1})(\Delta t_p)^{3/2} \|\mathbf{u}\|_{H^2(t_{m-2}^p, t_m^p; L^q)} + Q \Delta t_n^c \Big] \\
&\le 3\phi_{min}^{-1} \Delta t_n^c h_p^{\frac{2}{q}} \Big[(A_2 + Q^*) K_2 h_p^k (\|\mathbf{u}\|_{L^\infty(H^{k+1}(\mathrm{div}))} + \|p\|_{L^\infty(H^{k+1})}) \qquad (6.11) \\
&\quad + Q^* Q_2 \Big(\|\bar{c}\|_{L^2(L^2)} + \|c\|_{L^2(H^1)} + \Big\| \frac{d\bar{c}}{d\tau} \Big\|_{L^2(L^2)} + \Big\| \frac{dc}{d\tau} \Big\|_{L^2(H^1)} \Big) \\
&\quad + Q^* Q_2 \|\mathbf{u}\|_{H^2(L^2)} + Q^* \Big] \\
&\le Q_3 \Delta t_n^c h_p^{\frac{2}{q}}. \qquad \square
\end{aligned}
$$

6.2. Miscellaneous estimates. In this subsection we prove the estimates (5.5), (5.6), and (5.7) for the fifth, the first, and the second and seventh terms on the right-hand side of (5.4), respectively.

LEMMA 6.3. *Under the conditions of Theorem* 5.1*, estimate* (5.5) *holds for* $n = N_{m-1}+1,\dots,N_m$.

Proof. It is straightforward to see that [8]

$$|\mathbf{D}(\mathbf{x},\mathbf{u})-\mathbf{D}(\mathbf{x},\mathbf{v})| \le Q_4\ |\mathbf{u}-\mathbf{v}|.$$

With this we bound the left-hand side of (5.5) by

$$\begin{aligned}
&\left|\Delta t_n^c \int_\Omega \nabla\xi(\mathbf{x},t_n^c)\cdot(\mathbf{D}(\mathbf{x},\mathbf{u}_h^E(\mathbf{x},t_n^c))-\mathbf{D}(\mathbf{x},\mathbf{u}(\mathbf{x},t_n^c)))\nabla\Pi c(\mathbf{x},t_n^c)d\mathbf{x}\right| \\
&\quad\le Q_4\Delta t_n^c\|\nabla\xi(\mathbf{x},t_n^c)\|\ \|\mathbf{u}_h^E(\mathbf{x},t_n^c)-\mathbf{u}(\mathbf{x},t_n^c)\|\ \|\Pi c(\mathbf{x},t_n^c)\|_{W_\infty^1} \\
&\quad\le K_3Q_4\Delta t_n^c\|\nabla\xi(\mathbf{x},t_n^c)\|\ (\|\mathbf{u}_h^E(\mathbf{x},t_n^c)-(\Pi\mathbf{u})^E(\mathbf{x},t_n^c)\| \\
&\qquad+\|(\Pi\mathbf{u})^E(\mathbf{x},t_n^c)-\mathbf{u}^E(\mathbf{x},t_n^c)\|+\|\mathbf{u}^E(\mathbf{x},t_n^c)-\mathbf{u}(\mathbf{x},t_n^c)\|),
\end{aligned} \tag{6.12}$$

where at the last "$\le$" sign we have used (4.7).

The last term in the bracket is bounded in (4.12). We use the estimate (4.9) to bound the second term in the bracket to get

$$\begin{aligned}
&\|(\Pi\mathbf{u})^E(\mathbf{x},t_n^c)-\mathbf{u}^E(\mathbf{x},t_n^c)\| \\
&\quad\le Q(\|\Pi\mathbf{u}(\mathbf{x},t_{m-2}^p)-\mathbf{u}(\mathbf{x},t_{m-2}^p)\|+\|\Pi\mathbf{u}(\mathbf{x},t_{m-1}^p)-\mathbf{u}(\mathbf{x},t_{m-1}^p)\|) \\
&\quad\le Qh_p^{k+1}(\|\mathbf{u}\|_{L^\infty(H^{k+1}(\mathrm{div}))}+\|p\|_{L^\infty(H^{k+1})}).
\end{aligned}$$

We use (4.6) and (5.3) to bound the first term in the bracket of (6.12) by

$$\begin{aligned}
&\|\mathbf{u}_h^E(\mathbf{x},t_n^c)-(\Pi\mathbf{u})^E(\mathbf{x},t_n^c)\| \\
&\quad\le Q\ (\|\mathbf{u}_h(\mathbf{x},t_{m-2}^p)-\Pi\mathbf{u}(\mathbf{x},t_{m-2}^p)\|+\|\mathbf{u}_h(\mathbf{x},t_{m-1}^p)-\Pi\mathbf{u}(\mathbf{x},t_{m-1}^p)\|) \\
&\quad\le Q\ (\|c_h(\mathbf{x},t_{m-2}^p)-c(\mathbf{x},t_{m-2}^p)\|+\|c_h(\mathbf{x},t_{m-1}^p)-c(\mathbf{x},t_{m-1}^p)\|) \\
&\quad\le Q\ (\|\xi(\mathbf{x},t_{m-2}^p)\|+\|\xi(\mathbf{x},t_{m-1}^p)\|)+Qh_c^{l+1}\|c\|_{L^\infty(H^{l+1})}.
\end{aligned}$$

We combine these two estimates with (4.12) to complete the proof. ∎

LEMMA 6.4. *Under the conditions of Theorem* 5.1*, estimate* (5.6) *holds for* $n = N_{m-1}+1,\dots,N_m$.

Proof. The first term on the right-hand side of (5.4) can be written as

$$\begin{aligned}
\int_\Omega \phi(\mathbf{x})\xi(\mathbf{x},t_{n-1}^c)\xi(\mathbf{x},t_{n-1}^{c,+})d\mathbf{x} &= \int_\Omega \phi(\mathbf{x})\xi(\mathbf{x},t_{n-1}^c)\xi(\tilde{\mathbf{x}}_h,t_n^c)d\mathbf{x} \\
&= \int_\Omega \phi(\mathbf{x})\xi(\mathbf{x},t_{n-1}^c)\xi(\tilde{\mathbf{x}},t_n^c)d\mathbf{x} \\
&\quad+\int_\Omega \phi(\mathbf{x})\xi(\mathbf{x},t_{n-1}^c)(\xi(\tilde{\mathbf{x}}_h,t_n^c)-\xi(\tilde{\mathbf{x}},t_n^c))d\mathbf{x}.
\end{aligned} \tag{6.13}$$

We use estimate (6.2) to bound the first term on the right-hand side of (6.13)

$$
\begin{aligned}
&\left|\int_\Omega \phi(\mathbf{x})\xi(\mathbf{x},t_{n-1}^c)\xi(\tilde{\mathbf{x}},t_n^c)d\mathbf{x}\right| \\
&\quad\le \frac{1}{2}\int_\Omega \phi(\mathbf{x})\xi^2(\mathbf{x},t_{n-1}^c)d\mathbf{x}+\frac{1}{2}\int_\Omega \phi(\mathbf{x})\xi^2(\tilde{\mathbf{x}},t_n^c)d\mathbf{x} \\
&\quad= \frac{1}{2}\int_\Omega \phi(\mathbf{x})\xi^2(\mathbf{x},t_{n-1}^c)d\mathbf{x}+\frac{1}{2}\int_\Omega \phi(\tilde{\mathbf{x}})\xi^2(\tilde{\mathbf{x}},t_n^c)\det\Big(\frac{\partial \mathbf{x}}{\partial \tilde{\mathbf{x}}}\Big)d\tilde{\mathbf{x}} \\
&\quad\quad+\frac{1}{2}\int_\Omega (\phi(\mathbf{x})-\phi(\tilde{\mathbf{x}}))\xi^2(\tilde{\mathbf{x}},t_n^c)\det\Big(\frac{\partial \mathbf{x}}{\partial \tilde{\mathbf{x}}}\Big)d\tilde{\mathbf{x}} \\
&\quad\le \frac{1}{2}\int_\Omega \phi(\mathbf{x})\xi^2(\mathbf{x},t_{n-1}^c)d\mathbf{x}+\frac{1}{2}\int_\Omega \phi(\mathbf{x})\xi^2(\mathbf{x},t_n^c)d\mathbf{x}+Q\Delta t_n^c\|\xi(\mathbf{x},t_n^c)\|^2.
\end{aligned}
\tag{6.14}
$$

Here we have rewritten $\tilde{\mathbf{x}}$ as $\mathbf{x}$ in the second term.

We use (4.2), (6.6), and the relationship

$$
\xi(\tilde{\mathbf{x}}_h,t_n^c)-\xi(\tilde{\mathbf{x}},t_n^c)=(\tilde{\mathbf{x}}_h-\tilde{\mathbf{x}})\cdot\int_0^1 \nabla\xi(\tilde{\mathbf{x}}+s(\tilde{\mathbf{x}}_h-\tilde{\mathbf{x}}),t_n^c)ds,
\tag{6.15}
$$

to bound the second term on the right-hand side of (6.13) by

$$
\begin{aligned}
&\left|\int_\Omega \phi(\mathbf{x})\xi(\mathbf{x},t_{n-1}^c)(\xi(\tilde{\mathbf{x}}_h,t_n^c)-\xi(\tilde{\mathbf{x}},t_n^c))d\mathbf{x}\right| \\
&\quad\le Q\|\xi(\mathbf{x},t_{n-1}^c)\|_{L^\infty}\|\tilde{\mathbf{x}}_h-\tilde{\mathbf{x}}\|\left(\int_0^1\int_\Omega |\nabla\xi(\tilde{\mathbf{x}}+s(\tilde{\mathbf{x}}_h-\tilde{\mathbf{x}}),t_n^c)|^2d\mathbf{x}ds\right)^{1/2} \\
&\quad\le K_1Q|\ln h_c|^{1/2}\|\xi(\mathbf{x},t_{n-1}^c)\|_{H^1}\|\tilde{\mathbf{x}}_h-\tilde{\mathbf{x}}\|\ \|\nabla\xi(\mathbf{x},t_{n-1}^c)\| \\
&\quad\le \varepsilon\Delta t_n^c(\|\nabla\xi(\mathbf{x},t_n^c)\|^2+\|\nabla\xi(\mathbf{x},t_{n-1}^c)\|^2+\|\xi(\mathbf{x},t_{n-1}^c)\|^2),
\end{aligned}
\tag{6.16}
$$

provided that

$$
2K_1QQ_3h_p^{2q}|\ln h_c|^{1/2}\le\varepsilon,\qquad 2K_1QQ_4h_p^{2q}|\ln h_c|^{1/2}\le\varepsilon. \qquad \square
\tag{6.17}
$$

LEMMA 6.5. *Under the conditions of Theorem* 5.1, *estimate* (5.7) *holds for* $n=N_{m-1}+1,\dots,N_m$.

Proof. We decompose the second and seventh terms on the right-hand side of (5.4) as follows:

$$
\begin{aligned}
&\int_\Omega \phi(\mathbf{x})\eta(\mathbf{x},t_n^c)\xi(\mathbf{x},t_n^c)d\mathbf{x}-\int_\Omega \phi(\mathbf{x})\eta(\mathbf{x},t_{n-1}^c)\xi(\mathbf{x},t_{n-1}^{c,+})d\mathbf{x} \\
&\quad=\int_\Omega \phi(\mathbf{x})\left[\int_{t_{n-1}^c}^{t_n^c}\frac{\partial\eta(\mathbf{x},t)}{\partial t}dt\right]\xi(\mathbf{x},t_n^c)d\mathbf{x} \\
&\quad\quad-\int_\Omega \phi(\mathbf{x})\eta(\mathbf{x},t_{n-1}^c)(\xi(\tilde{\mathbf{x}}_h,t_n^c)-\xi(\tilde{\mathbf{x}},t_n^c))d\mathbf{x} \\
&\quad\quad-\int_\Omega \phi(\mathbf{x})\eta(\mathbf{x},t_{n-1}^c)(\xi(\tilde{\mathbf{x}},t_n^c)-\xi(\mathbf{x},t_n^c))d\mathbf{x}.
\end{aligned}
\tag{6.18}
$$

The first term on the right-hand side of (6.18) is bounded by

$$
\begin{aligned}
&\left|\int_\Omega \phi(\mathbf{x})\left[\int_{t_{n-1}^c}^{t_n^c}\frac{\partial\eta(\mathbf{x},t)}{\partial t}dt\right]\xi(\mathbf{x},t_n^c)d\mathbf{x}\right| \\
&\quad\le Q(\Delta t_n^c)^{1/2}\ \|\xi(\mathbf{x},t_n^c)\|\ \|\eta\|_{H^1(t_{n-1}^c,t_n^c;L^2)} \\
&\quad\le Q\Delta t_n^c\ \|\xi(\mathbf{x},t_n^c)\|^2+A_1^2h_c^{2l+2}\|c\|^2_{H^1(t_{n-1}^c,t_n^c;H^{l+1})}.
\end{aligned}
\tag{6.19}
$$

We use (3.15) to bound the third term on the right-hand side of (6.18)

$$\begin{aligned}
&\left|\int_\Omega \phi(\mathbf{x})\eta(\mathbf{x},t_{n-1}^c)(\xi(\tilde{\mathbf{x}},t_n^c)-\xi(\mathbf{x},t_n^c))d\mathbf{x}\right| \\
&\quad=\left|\int_\Omega \phi(\mathbf{x})\eta(\mathbf{x},t_{n-1}^c)(\tilde{\mathbf{x}}-\mathbf{x})\cdot\int_0^1 \nabla\xi(\mathbf{x}+s(\tilde{\mathbf{x}}-\mathbf{x}),t_n^c)d\mathbf{x}\right| \\
&\quad\le Q\Delta t_n^c\|\nabla\xi(\mathbf{x},t_n)\|\ \|\eta(\mathbf{x},t_{n-1}^c)\| \\
&\quad\le \varepsilon\Delta t_n^c\ \|\nabla\xi(\mathbf{x},t_n^c)\|^2+Q\Delta t_n^c h_c^{2l+2}\|c\|^2_{L^\infty(H^{l+1})}.
\end{aligned} \tag{6.20}$$

We use the estimate (6.6) with $q=2(2+\gamma)/\gamma$ and the generalized Hölder inequality to bound the second term on the right-hand side of (6.18) in a similar manner to the estimate of (6.16) as follows:

$$\begin{aligned}
&\left|\int_\Omega \phi(\mathbf{x})\eta(\mathbf{x},t_{n-1}^c)(\xi(\tilde{\mathbf{x}}_h,t_n^c)-\xi(\tilde{\mathbf{x}},t_n^c))d\mathbf{x}\right| \\
&\quad\le Q\left\|\tilde{\mathbf{x}}_h-\tilde{\mathbf{x}}\right\|_{L^{\frac{2(2+\gamma)}{\gamma}}}\|\eta(\mathbf{x},t_{n-1}^c)\|_{L^{2+\gamma}}\|\nabla\xi(\mathbf{x},t_n^c)\| \\
&\quad\le \varepsilon\Delta t_n^c\|\nabla\xi(\mathbf{x},t_n^c)\|^2+Q\Delta t_n^c h_c^{2l+2}\|c\|^2_{L^\infty(W^{l+1}_{2+\gamma})},
\end{aligned} \tag{6.21}$$

provided that

$$QQ_3h_p^{\frac{\gamma}{2+\gamma}}\le\varepsilon,\qquad QQ_4h_p^{\frac{\gamma}{2+\gamma}}\le\varepsilon. \qquad \square \tag{6.22}$$

Acknowledgments. The author expresses his sincere thanks to the referees for their very helpful comments and suggestions, which greatly improved the quality of this paper.

REFERENCES

[1] T. Arbogast and M.F. Wheeler, *A characteristics-mixed finite element method for advection-dominated transport problems*, SIAM J. Numer. Anal., 32 (1995), pp. 404–424.
[2] J. Bear, *Dynamics of Fluids in Porous Materials*, Elsevier, New York, 1972.
[3] P.J. Binning and M.A. Celia, *A finite volume Eulerian-Lagrangian localized adjoint method for solution of the contaminant transport equations in two-dimensional multi-phase flow systems*, Water Resour. Res., 32 (1996), pp. 103–114.
[4] F. Brezzi, *On the existence, uniqueness and approximation of saddle-point problems arising from Lagrangian multipliers*, RAIRO Anal. Numer., 8 (1974), pp. 129–151.
[5] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer-Verlag, New York, 1991
[6] M.A. Celia, T.F. Russell, I. Herrera, and R.E. Ewing, *An Eulerian-Lagrangian localized adjoint method for the advection-diffusion equation*, Adv. Wat. Resour., 13 (1990), pp. 187–206.
[7] P.G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1978.
[8] J. Douglas, Jr., R.E. Ewing, and M.F. Wheeler, *The approximation of the pressure by a mixed method in the simulation of miscible displacement*, RAIRO Anal. Numer., 17 (1983), pp. 17–33.
[9] J. Douglas, Jr., R.E. Ewing, and M.F. Wheeler, *A time-discretization procedure for a mixed finite element approximation of miscible displacement in porous media*, RAIRO Anal. Numer., 17 (1983), pp. 249–265.
[10] J. Douglas, Jr., and T.F. Russell, *Numerical methods for convection-dominated diffusion problems based on combining the method of characteristics with finite element or finite difference procedures*, SIAM J. Numer. Anal., 19 (1982), pp. 871–885.
[11] R.E. Ewing, ed., *The Mathematics of Reservoir Simulation*, SIAM, Philadelphia, 1984.

[12] R.E. Ewing, T.F. Russell, and M.F. Wheeler, *Simulation of miscible displacement using mixed methods and a modified method of characteristics*, SPE 12241, (1983), pp. 71–81.
[13] R.E. Ewing, T.F. Russell, and M.F. Wheeler, *Convergence analysis of an approximation of miscible displacement in porous media by mixed finite elements and a modified method of characteristics*, Comput. Methods Appl. Mech. Engrg., 47 (1984), pp. 73–92.
[14] R.E. Ewing and H. Wang, *Eulerian-Lagrangian localized adjoint methods for linear advection or advection-reaction equations and their convergence analysis*, Comput. Mech., 12 (1993), pp. 97–121.
[15] R.E. Ewing and H. Wang, *An optimal-order error estimate to Eulerian-Lagrangian localized adjoint methods for variable-coefficient advection-reaction problems*, SIAM J. Numer. Anal., 33 (1996), pp. 318–348.
[16] R.E. Ewing and H. Wang, *A summary of numerical methods for time-dependent advection-dominated partial differential equations*, J. Comput. Appl. Math., 128 (2001), pp. 423–445.
[17] C.I. Heberton, T.F. Russell, L.F. Konikow, and G.Z. Hornberger, *Three-dimensional finite-volume ELLAM implementation*, in Computational Methods in Water Resources XIII, L. Bently et al. eds., Balkema, Rotterdam, pp. 603–610.
[18] K.W. Morton, *Numerical Solution of Convection-Diffusion Problems*, Chapman & Hall, London, 1996.
[19] R. Rannacher and R. Scott, *Some optimal error estimate for piecewise linear finite element approximations*, Math. Comp., 38 (1982), pp. 437–445.
[20] P.-A. Raviart and J.-M. Thomas, *A Mixed Finite Element Method for Second Order Elliptic Problems*, Lecture Notes in Math. 606, Galligani and Magenes, eds., Springer-Verlag, Berlin, 1977, pp. 292–315.
[21] T.F. Russell and M.A. Celia, *An overview of research on Eulerian-Lagrangian localized adjoint methods (ELLAM)*, Adv. Wat. Resour., 25 (2002), pp. 1215–1231.
[22] T.F. Russell and M.F. Wheeler, *Finite element and finite difference methods for continuous flows in porous media*, in The Mathematics of Reservoir Simulation, R. E. Ewing, ed., SIAM, Philadelphia, 1984, pp. 35–106.
[23] V. Thomée, *Galerkin Finite Element Methods for Parabolic Problems*, Lecture Notes in Math. 1054, Springer-Verlag, New York, 1984.
[24] H. Wang, *A family of ELLAM Schemes for advection-diffusion-reaction equations and their convergence analyses*, Numer. Methods Partial Differential Equations, 14 (1998), pp. 739–780.
[25] H. Wang, *An optimal-order error estimate for an ELLAM scheme for two-dimensional linear advection-diffusion equations*, SIAM J. Numer. Anal., 37 (2000), pp. 1338–1368.
[26] H. Wang, H.K. Dahle, R.E. Ewing, M.S. Espedal, R.C. Sharpley, and S. Man, *An ELLAM scheme for advection-diffusion equations in two dimensions*, SIAM J. Sci. Comput., 20 (1999), pp. 2160–2194.
[27] H. Wang and R.E. Ewing, *Optimal-order convergence rates for Eulerian-Lagrangian localized adjoint methods for reactive transport and contamination in groundwater*, Numer. Methods Partial Differential Equations, 11 (1995), pp. 1–31.
[28] H. Wang, R.E. Ewing, G. Qin, S.L. Lyons, M. Al-Lawatia, and S. Man, *A family of Eulerian-Lagrangian localized adjoint methods for multi-dimensional advection-reaction equations*, J. Comput. Phys., 152 (1999), pp. 120–163.
[29] H. Wang, R.E. Ewing, and T.F. Russell, *Eulerian-Lagrangian localized methods for convection-diffusion equations and their convergence analysis*, IMA J. Numer. Anal., 15 (1995), pp. 405–459.
[30] H. Wang, D. Liang, R.E. Ewing, S.L. Lyons, and G. Qin, *An approximation to miscible fluid flows in porous media with point sources and sinks by an Eulerian–Lagrangian localized adjoint method and mixed finite element methods*, SIAM J. Sci. Comput., 22 (2000), pp. 561–581.
[31] K. Wang, *A uniformly optimal-order error estimate of an ELLAM scheme for unsteady-state advection-diffusion equations*, Internat. J. Numer. Anal. and Modeling, 5 (2008), pp. 286–302.
[32] K. Wang, H. Wang, and M. Al-Lawatia, *An Eulerian-Lagrangian discontinuous Galerkin method for transient advection-diffusion equations*, Numer. Methods Partial Differential Equations, 23 (2007), pp. 1343–1367.
[33] M.F. Wheeler, *A priori L_2 error estimates for Galerkin approximations to parabolic partial differential equations*, SIAM J. Numer. Anal., 10 (1973), pp. 723–759.

SIAM J. NUMER. ANAL.
Vol. 46, No. 4, pp. 2153–2168
© 2008 Society for Industrial and Applied Mathematics

DOMAIN DECOMPOSITION FOR LESS REGULAR SUBDOMAINS: OVERLAPPING SCHWARZ IN TWO DIMENSIONS*

CLARK R. DOHRMANN†, AXEL KLAWONN‡, AND OLOF B. WIDLUND§

Abstract. In the theory of domain decomposition methods, it is often assumed that each subdomain is the union of a small set of coarse triangles or tetrahedra. In this study, extensions to the existing theory which accommodate subdomains with much less regular shapes are presented; the subdomains are required only to be John domains. Attention is focused on overlapping Schwarz preconditioners for problems in two dimensions with a coarse space component of the preconditioner, which allows for good results even for coefficients which vary considerably. It is shown that the condition number of the domain decomposition method is bounded by $C(1 + H/\delta)(1 + \log(H/h))^2$, where the constant C is independent of the number of subdomains and possible jumps in coefficients between subdomains. Numerical examples are provided which confirm the theory and demonstrate very good performance of the method for a variety of subregions including those obtained when a mesh partitioner is used for the domain decomposition.

Key words. domain decomposition, overlapping Schwarz, preconditioners, iterative methods, irregular subdomains, John domains

AMS subject classifications. 65F10, 65N30, 65N55

DOI. 10.1137/070685841

1. Introduction.

In the theory for overlapping Schwarz domain decomposition methods, we typically assume that each subdomain is quite regular, e.g., the union of a small set of coarse triangles or tetrahedra; see, e.g., [28, Assumption 4.3]. However, this is often unrealistic, especially if the subdomains result from using a mesh partitioner. The subdomain boundaries might then not even be uniformly Lipschitz continuous; i.e., the number of patches required to cover the boundary of the region, in each of which the boundary is the graph of a Lipschitz continuous function, might not be uniformly bounded independently of the finite element mesh size. We note that the existing theory for iterative substructuring establishes bounds on the convergence rates of the algorithms which are insensitive to even large jumps in the material properties across subdomain boundaries as reflected in the coefficients of the problem. The theory for overlapping Schwarz methods is less restrictive as far as the subdomain shapes are concerned (see, e.g., [28, Chapter 3]) but relatively little has been known about the effect of large changes in the coefficients; see, however, [25] and recent work reported in [13, 26].

The purpose of this paper is to begin the development of a theory under much weaker assumptions on the partitioning; a companion study for iterative substruc-

*Received by the editors March 21, 2007; accepted for publication (in revised form) December 17, 2007; published electronically May 2, 2008.

http://www.siam.org/journals/sinum/46-4/68584.html

†Structural Dynamics Research Department, Sandia National Laboratories, Albuquerque, NM 87185 (crdohrm@sandia.gov). Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

‡Fachbereich Mathematik, Universität Duisburg-Essen, Universitätsstr. 3, 45117 Essen, Germany (axel.klawonn@uni-due.de).

§Department of Mathematics, Courant Institute, New York, NY 10012 (widlund@cims.nyu.edu).
This author's work was supported in part by the U.S. Department of Energy under contracts DE-FG02-06ER25718 and DE-FC02-01ER25482 and in part by National Science Foundation grant DMS-0513251.

turing methods, by Rheinbach and the second and third authors of this paper, has also been completed [16]. Our results in this paper concern an overlapping Schwarz algorithm which combines a coarse space adopted from an iterative substructuring method (see [28, Algorithm 5.16] and [11]) with local preconditioner components selected as in classical overlapping Schwarz methods, i.e., based on solving problems on overlapping subdomains. This choice of the coarse component will allow us to prove results which are independent of coefficient jumps. We note that there are earlier studies on multilevel methods [10, 24] in which the coarsest components are similarly borrowed from iterative substructuring algorithms.

We will use nonoverlapping subdomains and denote them by $\Omega_i, i = 1, \ldots, N$, as well as overlapping subdomains $\Omega'_j, j = 1, \ldots, N'$. The interface between the Ω_i will be denoted by Γ.

So far, complete results have been obtained only for problems in the plane. To simplify our presentation, we will also confine ourselves to scalar elliptic problems of the following form:

$$-\nabla \cdot (\rho(x)\nabla u(x)) \;=\; f(x), \quad x \in \Omega \subset \mathbb{R}^2, \tag{1.1}$$

with a Dirichlet boundary condition on a measurable subset $\partial\Omega_D$ of $\partial\Omega$, the boundary of Ω, and a Neumann condition on $\partial\Omega_N = \partial\Omega \setminus \partial\Omega_D$. The coefficient $\rho(x)$ is strictly positive and assumed to be a constant ρ_i for $x \in \Omega_i$. We use piecewise linear, continuous finite elements and triangulations with shape-regular elements and assume that each subdomain is the union of a set of elements. The weak formulation of the elliptic problem is written in terms of the bilinear form

$$a(u,v) \;:=\; \sum_{i=1}^{N} a_i(u,v) \;:=\; \sum_{i=1}^{N} \rho_i \int_{\Omega_i} \nabla u \cdot \nabla v dx.$$

Our study requires the generalization of some technical tools used in the proof of a bound of the convergence rate of this type of algorithm. Some of the standard tools, such as those collected in [28, section 4.6], are no longer available, and we therefore also have to modify the reasoning in the proof of our main result. Three auxiliary results, namely, a Poincaré inequality, a Sobolev-type inequality for finite element functions, and a bound for certain edge terms, will be required in our proof; see Lemmas 2.2, 3.2, and 3.4. We will work with John domains, introduced in section 2, and will be able to express our bounds on the convergence of our algorithm in terms of a few geometric parameters.

2. A Poincaré inequality and John domains. In any analysis of any domain decomposition method with more than one level, we need a Poincaré inequality. This inequality is closely related to an isoperimetric inequality; see [12, 20, 18, 21, 22].

LEMMA 2.1 (isoperimetric inequality). *Let $\Omega \subset \mathbb{R}^n$ be a domain—an open, bounded, and connected set—and let u be sufficiently smooth. Then*

$$\inf_{c \in R} \left(\int_\Omega |u - c|^{n/(n-1)}\, dx \right)^{(n-1)/n} \leq \gamma(\Omega, n) \int_\Omega |\nabla u|\, dx$$

if and only if

$$(\min(|A|, |B|))^{1-1/n} \leq \gamma(\Omega, n) |\partial A \cap \partial B|. \tag{2.1}$$

Here $A \subset \Omega$ is an arbitrary open set, and $B = \Omega \setminus \bar{A}$; $\gamma(\Omega, n)$ is the best possible constant, and $|A|$ is the measure of the set A, etc.

We note that the domain does not need to be star-shaped or Lipschitz. A small value of $\gamma(\Omega, n)$ is desirable for our purposes. Thus, the domain cannot be very thin, and its boundary should not have very pointed parts.

For two dimensions, we immediately obtain a standard Poincaré inequality by using the Cauchy–Schwarz inequality. The best choice of c is $\bar{u}_\Omega$, the average of u over the domain.

LEMMA 2.2 (Poincaré's inequality). *Consider a domain $\Omega \subset \mathbb{R}^2$. Then*

$$\|u - \bar{u}_\Omega\|^2_{L_2(\Omega)} \leq (\gamma(\Omega, 2))^2 |\Omega| \|\nabla u\|^2_{L_2(\Omega)} \quad \forall u \in H^1(\Omega).$$

For $n = 3$, a similar bound is obtained by using Hölder's inequality several times. In Lemma 2.2, we then should replace $|\Omega|$ by $|\Omega|^{2/3}$.

Throughout, we will use weighted $H^1(\Omega_i)$-norms defined by

$$\|u\|^2_{H^1(\Omega_i)} := |u|^2_{H^1(\Omega_i)} + 1/H_i^2 \|u\|^2_{L_2(\Omega_i)} := \int_{\Omega_i} \nabla u \cdot \nabla u dx + 1/H_i^2 \int_{\Omega_i} |u|^2 dx.$$

Here H_i is the diameter of Ω_i. The weight originates from a dilation of a domain with diameter 1. We will use Lemma 2.2 to remove L_2-terms from full H^1-norms.

We will also use another variant of the Poincaré inequality; in fact we will use it only in the simple case when the region is a circular disk.

LEMMA 2.3 (Poincaré, L_1-variant). *Consider a domain $\Omega \subset \mathbb{R}^2$. Then*

$$\|u - \bar{u}_\Omega\|_{L_1(\Omega)} \leq \gamma(\Omega, 2) |\Omega|^{1/2} \|\nabla u\|_{L_1(\Omega)} \quad \forall u \in W^{1,1}(\Omega).$$

Proof. By the Cauchy–Schwarz inequality,

$$\|u - \bar{u}_\Omega\|_{L_1(\Omega)} \leq |\Omega|^{1/2} \|u - \bar{u}_\Omega\|_{L_2(\Omega)}.$$

The proof is then completed by using Lemma 2.1. ∎

We next give a definition of a John domain; see [14] and the references therein.

DEFINITION 2.4 (John domain). *A domain $\Omega \subset \mathbb{R}^n$ is a John domain if there exists a constant $C_J \geq 1$ and a distinguished central point $x_0 \in \Omega$ such that each $x \in \Omega$ can be joined to it by a rectifiable curve $\gamma : [0, 1] \to \Omega$ such that $\gamma(0) = x_0$, $\gamma(1) = x$, and* distance$(\gamma(t), \partial\Omega) \geq C_J^{-1} |x - \gamma(t)|$ *for all $t \in [0, 1]$.*

This condition can be viewed as a twisted cone condition. We note that certain snowflake curves with fractal boundaries are John domains and that the length of the boundary of a John domain can be arbitrarily much larger than its diameter; see Figure 5.1. We also note that, by restricting the refinement to the neighborhood of a point, the contribution to the length of the boundary from a small neighborhood of the point can be made to dominate. A John domain can have cusps facing inwards but not outwards; this means, however, that there can be no cusps on the interface since any cusp would face outwards for one of two neighboring subdomains; see [21, 22] for further information.

We note that, for any choice of the point x_0, there is a point $x \in \Omega$ at a distance of at least diameter$(\Omega)/2$. We find that diameter$(\Omega) \leq 2C_J r_\Omega$, where r_Ω is the radius of the largest ball inscribed in Ω and centered at x_0.

It is known that any simply connected plane domain with a finite Poincaré parameter $\gamma(\Omega, 2)$ is a John domain; see [7]. It is also known (see [2]) that any John domain has a bounded Poincaré parameter $\gamma(\Omega, n)$. Since we have been unable to find a simple estimate of $\gamma(\Omega, n)$ in terms of C_J in the literature, we have chosen to express our results in terms of both of these geometric parameters.

3. The algorithm, technical tools, and the main result. The domain $\Omega \subset \mathbb{R}^2$ of the elliptic problem is decomposed into nonoverlapping subdomains Ω_i, each of which is the union of elements, and with the finite element nodes on the boundaries of neighboring subdomains matching across the interface Γ, which is the union of the parts of the subdomain boundaries which are common to at least two subdomains. The interface Γ is composed of edges and vertices. An edge $\mathcal{E}^{ik}$ is an open subset of Γ, which contains the nodes which are shared only by the boundaries of a particular pair of subdomains Ω_i and Ω_k. The subdomain vertices $\mathcal{V}^\ell$ are end points of edges and are typically shared by more than two; see [17, Definition 3.1] for more details on how these sets can be defined for quite general situations. We denote the standard finite element space of continuous, piecewise linear functions on Ω_i by $V^h(\Omega_i)$ and assume that these functions vanish at the nodes on $\partial\Omega_i \cap \partial\Omega_D$.

We will view our algorithm as an additive Schwarz method, as in [28, Chapters 2 and 3], defined in terms of a set of subspaces. To simplify the discussion, we will use exact solvers for both the coarse problem and the local ones. All that is then required for the analysis of our algorithm is an estimate of a parameter in a stable decomposition of any function in the finite element space; see [28, Assumption 2.2 and Lemma 2.5]. Thus, we need to estimate C_0^2 in

$$a(u_0, u_0) + \sum_{j=1}^{N'} a_{\Omega'_j}(u_j, u_j) \leq C_0^2 a(u, u) \quad \forall u \in V^h,$$

for some decomposition $\{u_j\}$, such that

$$u = \sum_{j=0}^{N'} R_j^T u_j, \quad u_j \in V_j.$$

Here $a_{\Omega'_j}(u, v)$ is the bilinear form obtained by integrating only over the subdomain Ω'_j, and $R_j^T : V_j \longrightarrow V^h$ is an interpolation operator from the space of the jth subproblem V_j into the space V^h. The local spaces $V_j, j = 1, \ldots N'$, are defined as

$$V_j = V^h(\Omega'_j) \cap H_0^1(\Omega'_j), \tag{3.1}$$

except for subdomains with a boundary which intersects $\partial\Omega_N$. In the latter case, the functions of the local space are constrained to vanish only at the nodes of $\partial\Omega'_j \setminus \partial\Omega_N$. This is the same standard choice as in [28, Chapter 3]. We will assume that each Ω'_j has a diameter comparable to those of the subdomains Ω_i which it intersects.

Associated with each local space V_j is a projection P_j defined by

$$a_{\Omega'_j}(\tilde{P}_j u, v) = a(u, R_j^T v) \quad \forall v \in V_j, \text{ and } P_j = R_j^T \tilde{P}_j.$$

The coarse space V_0 is spanned by vertex and edge functions extended as discrete harmonic functions into the interiors of the subdomains Ω_i. The discrete harmonic extensions minimize the energy; see [28, section 4.4]. There will be one basis function $\theta_{\mathcal{V}^\ell}(x)$ for each subdomain vertex; it is the discrete harmonic extension of the interface values of the standard nodal basis function. For each edge $\mathcal{E}^{ik}$ there is also a basis function $\theta_{\mathcal{E}^{ik}}(x) \in V^h$, which equals 1 at all nodes on the edge and vanishes at all other interface nodes. The vertex and edge functions provide a partition of unity on the interface Γ. We will use a projection P_0 onto V_0, which is defined by

$$a(P_0 u, v) = a(u, v) \quad \forall v \in V_0. \tag{3.2}$$

The additive Schwarz operator, the preconditioned operator used in our iteration, is given by

$$P_{ad} = \sum_{j=0}^{N'} P_j.$$

By using [28, Lemmas 2.5 and 2.10], we find that the condition number $\kappa(P_{ad})$ of the additive Schwarz operator can be bounded by $(N^C + 1)C_0^2$, where N^C is the minimal number of colors required to color the subdomains Ω_j' such that no pair of intersecting subdomains have the same color.

The overlap between the subdomains is characterized by parameters δ_j, as in [28, Assumption 3.1]; δ_j is the minimum width of the subset Ω_{j,δ_j} of Ω_j' which is also covered by at least one neighboring overlapping subdomain.

We can now formulate our main result, which is also valid for compressible elasticity with piecewise constant Lamé parameters, with jumps only across the interface, provided that the coarse space is enriched to include all rigid body functions; see [8]. The extension to the case of linear elasticity requires a Korn inequality; see [1, 17], and [28, Chapter 8]. We note that there are also recent results on almost incompressible elasticity discretized by mixed finite elements with discontinuous pressure spaces; see [9].

THEOREM 3.1 (condition number estimate). *Let $\Omega \subset \mathbb{R}^2$ be partitioned into nonoverlapping subdomains Ω_i, which are John domains, each with a shape-regular triangulation. The condition number of our domain decomposition method then satisfies*

$$\kappa(P_{ad}) \leq C\,(1 + H/\delta)(1 + \log(H/h))^2,$$

where $C > 0$ is a constant which depends only on the John and Poincaré parameters of the subdomains, the number of colors required for the overlapping subdomains, and the shape regularity of the elements.

As in many domain decomposition results, H/h is shorthand for $\max_i(H_i/h_i)$, where h_i is the diameter of the smallest element of Ω_i. Similarly, H/δ is the largest ratio of H_i and the smallest of the δ_j of the overlapping subregions that intersect Ω_i.

The logarithmic bound of our main result can be improved to a first power if all of the subregions have boundaries which are uniformly Lipschitz. This is true, in particular, if they satisfy [28, Assumption 4.3]. If, in addition, the coefficients do not have any large jumps across the interface and the coarse space is enriched to contain all piecewise linear functions on a coarse mesh as in [28, Assumption 4.3], we can also eliminate the logarithmic factors altogether by using ideas from the proof of [28, Theorem 3.13].

To prove Theorem 3.1, we need two auxiliary results, in addition to Poincaré's inequality. The first is a discrete Sobolev inequality.

LEMMA 3.2 (discrete Sobolev inequality). *Let Ω_i be a John domain in the plane with diameter H. Then*

$$\|u - \bar{u}_{\Omega_i}\|^2_{L_\infty(\Omega_i)} \leq C(1 + \log(H/h))|u|^2_{H^1(\Omega_i)} \quad \forall u \in V^h(\Omega_i). \tag{3.3}$$

The constant $C > 0$ depends only on the John parameter $C_J(\Omega_i)$, the Poincaré parameter $\gamma(\Omega_i, 2)$, and the shape regularity of the elements.

We note that this lemma is well-known in the theory of iterative substructuring methods. Proofs for domains satisfying an interior cone condition are given in [4] and

in [6, section 4.9]; see also [3]. For a different proof, see [28, Lemma 4.15]. For a proof that this inequality is sharp, see [5].

In preparation for our proof of this result for John domains, we will formulate and prove a lemma which is directly inspired by a paper by Hajłasz [14]. We first introduce a set of open circular disks. We consider a John domain Ω_i with diameter H and with a triangulation using a mesh size h. Let $\gamma(t)$ be a John curve from a central point $x_0 \in \Omega_i$ to an arbitrary point x. A family of circular disks $B_i,\ i \geq 0$, is introduced as follows: the first disk is $B_0 := B(x_0, \text{dist}(x_0, \partial\Omega_i)/4)$, the open disk centered at x_0 and with radius $r_0 = \text{dist}(x_0, \partial\Omega_i)/4$. We will assume that x lies outside $2B_0$, the concentric disk with twice the radius of B_0; the other case is trivial. The other disks are defined similarly as $B_i := B(x_i, r_i)$, centered at x_i and with radius $r_i := |x - x_i|/4C_J, i \geq 1$. All of the x_i's lie on the John curve, and all of the B_i's are clearly subsets of Ω_i. Given B_i, we select $x_{i+1} \in \partial B_i$ as the last point of exit of $\gamma(t)$ from B_i.

Several properties of the B_i can now be established.

LEMMA 3.3. *There exists a constant $M = M(C_J)$ such that*

1. *no point in Ω_i belongs to more than M of the B_i (finite covering);*
2. *$|B_i \cup B_{i+1}| \leq M|B_i \cap B_{i+1}|$;*
3. *$r_i \to 0$ and $x_i \to x$ as $i \to \infty$;*
4. *the number of x_i that are at a distance larger than r from x can be estimated by $M \log(H/r)$.*

Proof. Our proof will rely heavily on arguments developed in [14], in particular, on the proof of Theorem 10 of that paper, which we will reproduce in part; see also references to earlier contributions in that paper.

Let a point y belong to $B_k, B_{k+1}, \ldots B_{k+p}$. Then

$$|x - y| \leq |x - x_{k+m}| + |x_{k+m} - y| \leq (4C_J + 1)r_{k+m}, \quad m = 0, \ldots, p.$$

Similarly,

$$4C_J r_{k+n} = |x - x_{k+n}| \leq |x - y| + |y - x_{k+n}| \leq |x - y| + r_{k+n}, \quad n = 0, \ldots, p.$$

Therefore, we have

$$(4C_J - 1)r_{k+n} \leq (4C_J + 1)r_{k+m}, \quad m, n = 0, \ldots, p.$$

Since no center of a disk belongs to any other disk, it is then easy to show that, for $m \neq n$,

$$|x_{k+n} - x_{k+m}| \geq (4C_J - 1)/(4C_J + 1)r_k, \tag{3.4}$$

which provides a uniform bound on p since only a finite number of such x_i can satisfy (3.4) while being within r_i of the common point y.

The second statement now follows easily by using the bound on the ratio of the radii of the two circles and elementary geometry.

We next consider a circle, centered at x, which covers all of Ω_i. We decompose this circle into annuli with a fixed ratio between the outer and inner circles of any annulus. It is easy to see that the number of x_i contained in each annulus will be uniformly bounded since the radii of the disks associated with an annulus are determined by the distance to x. This shows that $r_i \to 0$ and also that we can, by a simple covering argument, bound the number of x_i's that are at a distance larger than r from x by $M \log(H/r)$. ∎

Proof of Lemma 3.2. We will now confine x to be a centroid of an element K; by following [6, Formula (4.9.6)] we can then easily obtain our estimate for any $x \in K$. We denote by $N = N(h)$ the index of the first B_i which lies entirely inside K.

We will work with averages $\bar{u}_{B_i}$ of sufficiently smooth functions over the circular disks. We note that the average of a linear function over a circular disk B_i coincides with the value at its center if $B_i \subset K$.

Then, by following [14] and using the triangle inequality, we find that

$$|u(x_N) - \bar{u}_{B_0}| \le \sum_{i=0}^{N-1} |\bar{u}_{B_i} - \bar{u}_{B_{i+1}}|$$

$$\le \sum_{i=0}^{N-1} (|\bar{u}_{B_i} - \bar{u}_{B_i \cap B_{i+1}}| + |\bar{u}_{B_i \cap B_{i+1}} - \bar{u}_{B_{i+1}}|). \tag{3.5}$$

The expression $\bar{u}_{B_i} - \bar{u}_{B_i \cap B_{i+1}}$ is invariant under a shift of u by a constant. By replacing u by $u - \bar{u}_{\mathcal{B}_i}$, we therefore find that

$$|\bar{u}_{B_i} - \bar{u}_{B_i \cap B_{i+1}}| \le \frac{1}{|B_i \cap B_{i+1}|} \int_{B_i \cap B_{i+1}} |u - \bar{u}_{\mathcal{B}_i}|$$

$$\le \frac{1}{|B_i \cap B_{i+1}|} \int_{B_i} |u - \bar{u}_{\mathcal{B}_i}| = \frac{|B_i|}{|B_i \cap B_{i+1}|} \fint_{B_i} |u - \bar{u}_{B_i}|.$$

Here $\fint_{B_i} v$ denotes the average of v over the disk B_i. The second term of (3.5) can be estimated similarly, and we can now use the second statement of Lemma 3.3 to obtain

$$|u(x_N) - \bar{u}_{B_0}| \le (2M) \sum_{i=0}^{N-1} \fint_{B_i} |u - \bar{u}_{B_i}|.$$

By using Lemma 2.3 for the disks B_i, we obtain

$$\fint_{B_i} |u - \bar{u}_{B_i}| \le C \int_{B_i} \frac{|\nabla u(z)|}{r_i} dz.$$

We now use the Cauchy–Schwarz inequality and obtain

$$|u(x_N) - \bar{u}_{B_0}|^2 \le CN \sum_{i=0}^{N-1} \int_{B_i} |\nabla u(z)|^2 dz.$$

By using the finite covering property—the first statement of Lemma 3.3—this sum can be estimated by $\int_{\Omega_i} |\nabla u(z)|^2 dz$.

Given that B_N is the first disk which is a subset of K and that x_N lies on the boundary of B_{N-1}, we find that $\operatorname{dist}(x_N, \partial K) \le 2r_{N-1}$. We also know that

$$r_{N-1} \le \frac{4C_J + 1}{4C_J - 1} r_N \quad \text{and that} \quad r_N = \frac{|x - x_N|}{4C_J}.$$

Since

$$\operatorname{dist}(x, \partial K) \le |x_N - x| + \operatorname{dist}(x_N, \partial K),$$

we see that $|x - x_N|$ is bounded from below by $c(C_J)h$, where $c(C_J) > 0$ and h is the diameter of the element K. We can now use the logarithmic estimate of N—the fourth statement of Lemma 3.3—and find that

$$|u(x_N) - \bar{u}_{B_0}|^2 \leq C(1 + \log(H/h)) \int_{\Omega_i} |\nabla u(z)|^2 dz.$$

It is also elementary to estimate $|\bar{u}_{B_0} - \bar{u}_{\Omega_i}|$. We find that

$$|\bar{u}_{B_0} - \bar{u}_{\Omega_i}| \leq |B_0|^{-1} \|u - \bar{u}_{\Omega_i}\|_{L_1(B_0)} \leq |B_0|^{-1} \|u - \bar{u}_{\Omega_i}\|_{L_1(\Omega_i)}.$$

We complete the argument by using the Cauchy–Schwarz inequality, Lemma 2.2, and the fact that there is a bound on $|\Omega_i|$ in terms of $|B_0|$. We note that this term does not require any logarithmic factor.

We then have a bound for one point in the element K, and, as already indicated, we can easily estimate the difference between the values at any pair of points in the element K as in [6, Formula (4.9.6)]. ∎

A second important tool provides estimates of the edge functions.

LEMMA 3.4. *The edge function $\theta_{\mathcal{E}^{ik}}$ can be bounded as follows:*

$$\|\theta_{\mathcal{E}^{ik}}\|^2_{H^1(\Omega_i)} \leq C(1 + \log(H_i/h_i)), \tag{3.6}$$

and

$$\|\theta_{\mathcal{E}^{ik}}\|^2_{L_2(\Omega_i)} \leq CH_i^2(1 + \log(H_i/h_i)). \tag{3.7}$$

Here the constant C depends only on the John parameter $C_J(\Omega_i)$. We can remove the logarithmic factor from (3.7) *if all angles in the triangulation are acute, i.e.,*

$$\|\theta_{\mathcal{E}^{ik}}\|^2_{L_2(\Omega_i)} \leq CH_i^2. \tag{3.8}$$

A proof of this lemma is given in [16].

4. Proof of Theorem 3.1. As in many other proofs of results on domain decomposition algorithms, we will work with one subdomain Ω_i at a time. With local bounds, there are now no difficulties in handling variations of the coefficients across the interface.

We recall that the coarse space is spanned by the $\theta_{\mathcal{V}^\ell}$, the discrete harmonic extensions of the restrictions of the standard nodal basis functions to Γ, and the edge functions $\theta_{\mathcal{E}^{ik}}$. The coarse space component $u_0 \in V_0$ in the decomposition of an arbitrary finite element function $u(x)$ can be chosen as

$$u_0(x) \;=\; \sum_\ell u(\mathcal{V}^\ell)\theta_{\mathcal{V}^\ell}(x) + \sum_{ik} \bar{u}_{\mathcal{E}^{ik}} \theta_{\mathcal{E}^{ik}}(x). \tag{4.1}$$

Here $\bar{u}_{\mathcal{E}^{ik}}$ is the average of u over the edge. This interpolation formula is a two-dimensional analog of [28, Formula (5.13)].

Since the operator defined by (4.1) reproduces constants, we can replace u by $u - \bar{u}_{\Omega_i}$ and u_0 by $u_0 - \bar{u}_{\Omega_i}$. We now estimate the terms of the right-hand side of the resulting formula term by term. The energy of the vertex basis functions is uniformly bounded, and the values of $u(\mathcal{V}^\ell) - \bar{u}_{\Omega_i}$ are bounded by using Lemma 3.2. Turning to the edge terms, we use Lemma 3.4 to estimate the energy of the edge basis functions. The coefficient $\bar{u}_{\mathcal{E}^{ik}} - \bar{u}_{\Omega_i}$ is bounded by $\|u - \bar{u}_{\Omega_i}\|_{L_\infty(\Omega_i)}$. Thus,

$$|u_0|^2_{H^1(\Omega_i)} = |u_0 - \bar{u}_{\Omega_i}|^2_{H^1(\Omega_i)} \leq C(1 + \log(H/h))^2 |u|^2_{H^1(\Omega_i)}, \tag{4.2}$$

and we find that

$$a(u_0, u_0) \leq C(1 + \log(H/h))^2 a(u, u).$$

Remark 1. If the subdomains Ω_i have Lipschitz boundaries, we can estimate the edge averages by using the Cauchy–Schwarz inequality and an elementary trace theorem; see Nečas [23, Theorem 1.2]. This eliminates one of the two logarithmic factors.

Similarly, we can prove that

$$\|u - u_0\|^2_{L_2(\Omega_i)} \leq C(1 + \log(H/h))^2 H_i^2 |u|^2_{H^1(\Omega_i)}. \tag{4.3}$$

If all angles of the triangulation of Ω_i are acute, then by using (3.8) of Lemma 3.4 we can remove one of the logarithmic factors, and if the subdomain boundaries are Lipschitz, no logarithmic factors are required on the right-hand side of (4.3); see [28, Lemma 4.25] for a three-dimensional result for regular subdomains.

Remark 2. In the case of linear elasticity, we must include the restriction of all rigid body modes to the edges in the coarse space; see [19] or [27] for a discussion of the necessity of this null space property. In this case, we should extend the values given on the interface into the subdomains by solving homogeneous discrete elasticity problems.

We now turn to the estimate related to the local spaces. Again, we will carry out the work on one subdomain Ω_i at a time. Let $w := u - u_0$, and define a local term in the decomposition by $u_j = I^h(\theta_j w)$. We can borrow extensively from [28, sections 3.2 and 3.6]. Thus, I^h interpolates into V^h, and the $\theta_j \in R_j^T V_j$ provide a partition of unity. These functions vary between 0 and 1, and their gradients are bounded by $|\nabla \theta_j| \leq C/\delta_j$ and vanish outside the regions of overlap Ω_{j,δ_j} .

We note that the number of Ω_j' that intersect Ω_i is uniformly bounded; we will consider only the contribution from one of them, Ω_j'. As in our earlier work [28, section 3.5], the only term that requires a careful consideration is $\nabla \theta_j w$. We cover the set $\Omega_{j,\delta_j} \cap \Omega_i$ by square patches with sides on the order of δ_j and note that on the order of H_i/δ_j of them will suffice. Just as in the proof of [28, Lemma 3.10], we have

$$\int_{\Omega_i} |\nabla \theta_j w|^2 \leq C/\delta_j^2 (\delta_j^2 |w|^2_{H^1(\Omega_i)} + (H_i/\delta_j)\delta_j^2 \|w\|^2_{H^1(\Omega_i)}).$$

The proof is completed by combining this inequality with the bounds in (4.2) and (4.3).

Remark 3. Should the overlap between the subdomains be more generous in places, we can modify the relevant θ_j by making them vary from 0 to 1 over a distance of δ_j, effectively making the set Ω_{j,δ_j} more evenly wide.

5. Numerical examples. Numerical examples are presented in this section for the three different types of subdomains shown in Figure 5.1. Type 1 subdomains have a square geometry and are partitioned into square bilinear elements. Type 2 subdomains also consist of square bilinear elements, but their boundaries have a ragged shape. For type 3 subdomains, we employ equilateral linear triangular elements, and the edges of the subdomains have both straight-line and fractal segments. Numerical examples are also presented for subdomains obtained from a mesh partitioner; see Figure 5.5.

The main purpose of this first set of numerical examples is to verify certain estimates used in the proof of Theorem 3.1. For type 1 and 2 subdomains, the ratio

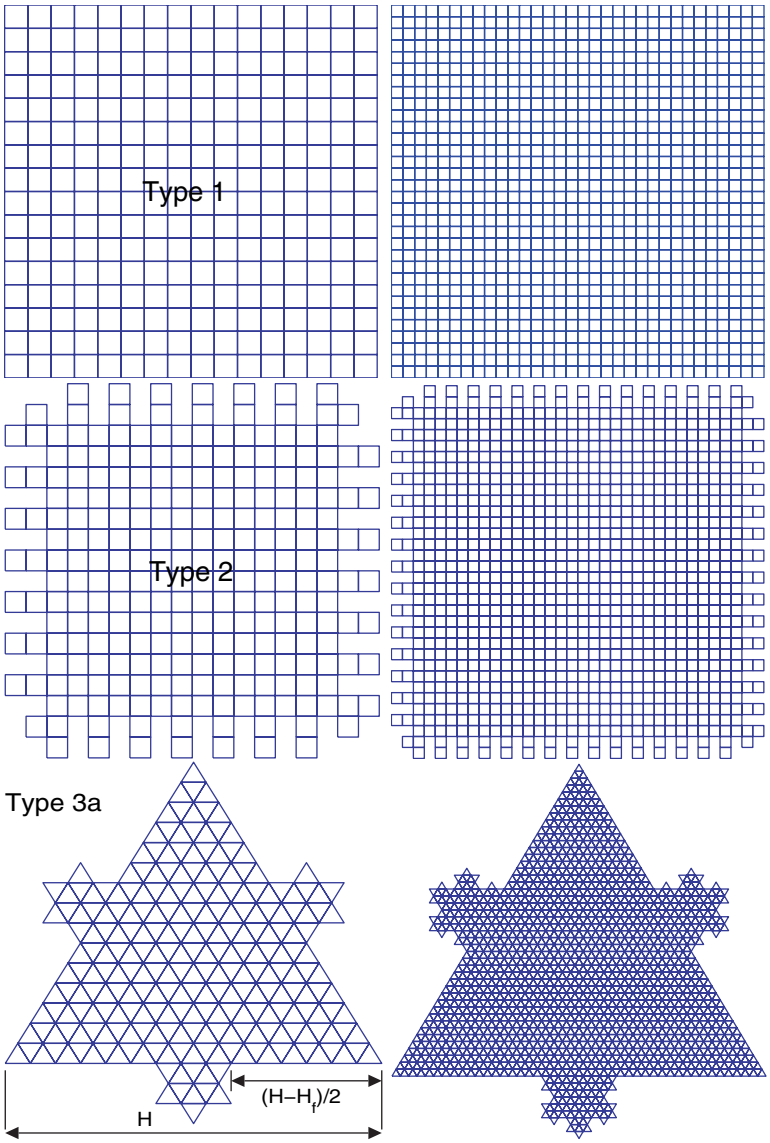


FIG. 5.1. *Three types of subdomains used in the numerical examples. Two levels of mesh refinement are shown, and the type* 3 *subdomains have a constant value of* $H/H_f = 5$. *For each refinement of type* 3 *subdomains, every element edge on the fractal part of the boundary is first divided into three shorter edges of* $1/3$ *the length. The middle one of these edges is then replaced by two other edges with which it forms an equilateral triangle. Thus, for fixed* H/H_f, *the length of the fractal part of the boundary increases by a factor of* $4/3$ *with each additional level of refinement.*

H/h is increased by a factor of 2 with each additional level of mesh refinement. At the ith ($i \geq 0$) level of refinement for type 3 subdomains,

$$H/h = (H/H_f)3^{i+1},$$

where $H/H_f = 5$ is fixed for type 3a subdomains and $H/H_f = 5 + 2i$ for type 3b. We note that in the asymptotic limit the fractal segment lengths grow by a factor of $4/3$ with each mesh refinement for both type 3 subdomains, whereas the straight-line

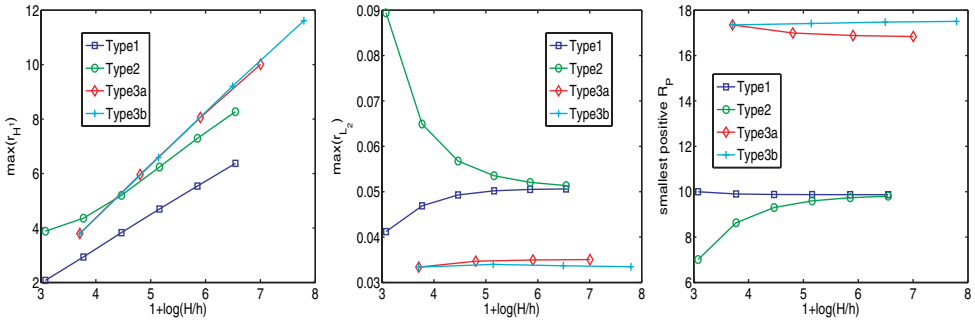


FIG. 5.2. *Rayleigh quotients for different subdomain types.*

segment lengths remain bounded. The primary difference between type 3a and 3b subdomains is that H/H_f is fixed in the former while H/H_f increases without bound in the latter. Thus, for type 3b subdomains, the protruding regions bounded by fractal segments decrease in area, while fractal segment lengths continue to increase with mesh refinement.

We recall the estimates (4.2) and (4.3) from the proof of Theorem 3.1 and consider the associated Rayleigh quotients

$$r_{H^1} = \frac{|u_0|^2_{H^1(\Omega_i)}}{|u|^2_{H^1(\Omega_i)}} \quad \text{and} \quad r_{L_2} = \frac{\|u - u_0\|^2_{L_2(\Omega_i)}}{|u|^2_{H^1(\Omega_i)}}$$

and also the Rayleigh quotient

$$r_P = \frac{|u|^2_{H^1(\Omega_i)}}{\|u\|^2_{L_2(\Omega_i)}}.$$

Stationary values of these Rayleigh quotients correspond to eigenvalues of generalized eigenproblems. For example, a stationary value of r_P is an eigenvalue λ of

$$K\phi = \lambda M\phi,$$

where K and M are the stiffness and mass matrices, respectively, of the finite element discretization of Ω_i with Neumann boundary conditions. We note in this case that the reciprocal of the smallest positive eigenvalue provides an estimate of the constant in Lemma 2.2. Maximum values of r_{H^1} and r_{L_2}, along with the smallest positive stationary value of r_P, are plotted versus $1 + \log(H/h)$ in Figure 5.2. We observe essentially the same dependence on $\log(H/h)$ for all four subdomain types. The maximum value of r_{H^1} appears to vary linearly with $\log(H/h)$ in the asymptotic limit, whereas the maximum value of r_{L_2} appears to be converging to a constant. This behavior is consistent with (4.2) and (4.3), but the estimates do not appear to be sharp for the different subdomain types and ranges of H/h considered. Consistent with the isoperimetric inequality, the rightmost plot in Figure 5.2 suggests that the smallest positive stationary value of r_P is uniformly bounded from below. We have yet to identify a subdomain shape with numerical behavior essentially different from the regular-shaped case.

The next group of numerical examples are for the domain decompositions shown in Figure 5.3. For type 3 subdomains, we fix $H/H_f = 5$. Homogeneous essential

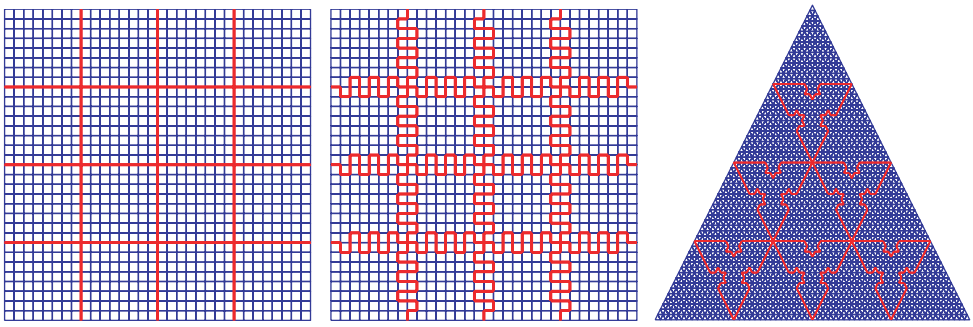


FIG. 5.3. *Sample domain decompositions, corresponding to subdomains of type* 1, 2, *and* 3a, *used in numerical examples. Subdomain boundaries are shown in red.*

TABLE 5.1

Results for domain decompositions shown in Figure 5.3. *Table headings iter, cond, and ndof denote the number of iterations, condition number estimates, and the number of unknowns in the problem, respectively. Material properties are the same for all subdomains.*

Type	H/h	d_j	iter	cond	ndof
1	8	1	25	9.11	1056
	16	3	29	11.07	4160
	32	7	31	12.99	16,512
	64	15	32	14.87	65,792
	128	31	34	16.71	262,656
2	8	1	28	13.53	1056
	16	3	30	15.70	4160
	32	7	32	17.81	16,512
	64	15	33	19.76	65,792
	128	31	34	21.60	262,656
3	15	3	30	18.56	1830
	45	11	34	22.79	16,290
	135	35	36	26.66	146,070
	405	107	37	30.14	1,313,010

boundary conditions are applied to the bottom edge of each domain, and $\rho_i = 1$ for each subdomain. Preconditioned conjugate gradients are used to solve the associated linear systems to a relative residual tolerance of 10^{-8} for random right-hand sides. The amount of overlap is specified by the integer d_j. If $\Omega_{i,h}$ contains all of the nodes in the closure of Ω_i, then the nodes in overlapping subdomain Ω_i' are the union of $\Omega_{i,h}$ with all nodes a graph distance d_j or less from one or more nodes in $\Omega_{i,h}$. We note that nodes on the boundary of subdomain Ω_i are also contained in $\Omega_{i,h}$; the designation nonoverlapping refers to an element rather than to a nodal decomposition.

The numbers of iterations and condition number estimates from conjugate gradient iterations are reported in Table 5.1. The condition numbers from Table 5.1 are also plotted in Figure 5.4. Provided the overlap $d_j + 1$ remains in a fixed proportion to H/h, i.e., H/δ is constant, the growth in condition number appears to be bounded by $C(1 + \log(H/h))$ for all three subdomain types considered. These results are consistent with Theorem 3.1 but suggest that the estimate in the theorem may not be sharp.

The next group of examples is identical to the previous one, but we now set $\rho_i = 1$ for the bottom two rows of subdomains and $\rho_i = \sigma$ for the top two rows

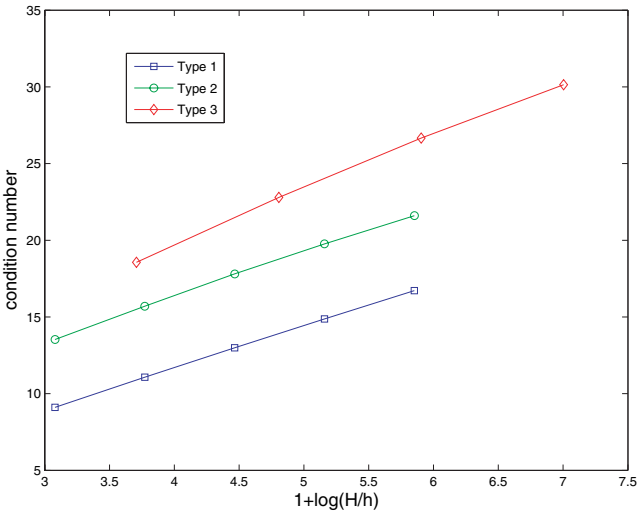


FIG. 5.4. *Condition numbers versus* $1 + \log(H/h)$ *for three different subdomain types.*

TABLE 5.2

Results for domain decompositions shown in Figure 5.3. *The bottom two rows of subdomains have* $\rho_i = 1$, *and the top two rows have* $\rho_i = \sigma$. *Material property jumps are aligned with subdomain boundaries.*

Type	H/h	d_j	σ	iter	cond	ndof
1	16	3	10^{-4}	32	10.10	4160
			10^{-2}	29	10.16	
			1	29	11.07	
			10^{2}	30	11.45	
			10^{4}	32	11.45	
2	16	3	10^{-4}	33	11.86	4160
			10^{-2}	32	12.19	
			1	30	15.70	
			10^{2}	31	16.75	
			10^{4}	32	16.76	
3	45	11	10^{-4}	34	18.84	16,290
			10^{-2}	33	18.92	
			1	34	22.79	
			10^{2}	33	27.72	
			10^{4}	35	27.86	

of subdomains in Figure 5.3. Consistent with the theory, the results in Table 5.2 suggest that condition numbers are bounded independently of material property jumps between subdomains.

The final group of examples deals with subdomains generated by a mesh partitioner. Consider a graph in which each vertex of the graph corresponds to a finite element. There is an edge between two vertices in the graph if the corresponding finite elements share an edge, i.e., they have two nodes in common. The graph partitioning software Metis [15] is then used to decompose such graphs into N subdomains. Example domain decompositions obtained in this manner are shown in Figure 5.5 for the same square geometry as in the previous examples. Results in Table 5.3 suggest that using a general-purpose mesh partitioner to generate subdomains does not degrade performance significantly. Results are also shown in the table for problems in which

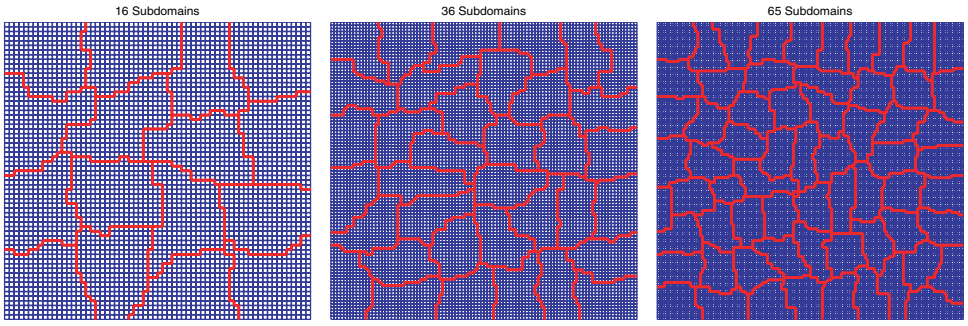


FIG. 5.5. *Sample domain decompositions obtained from a mesh partitioner.*

TABLE 5.3

Results for square subdomains and subdomains obtained from a mesh partitioner, where N designates the number of subdomains. In the upper right quadrant of the square domain $\rho(x) = \sigma$, and elsewhere $\rho(x) = 1$. For $N = 64$ and $H/h = 16$, the mesh partitioner generated one subdomain consisting of two disconnected components. These two components were treated as distinct subdomains resulting in $N = 65$. Results in the bottom part of the table are for a checkerboard arrangement of material properties.

				Square subdomains		Mesh partitioner		
N	H/h	σ	d_j	iter	cond	iter	cond	ndof
16	16	1	3	29	11.07	31	10.31	4160
	32		7	31	12.99	33	11.96	16,512
36	16		3	31	11.87	36	14.05	9312
	32		7	33	14.11	39	14.78	37,056
64/65	16		3	33	12.25	38	14.49	16,512
64	32		7	36	14.63	41	15.42	65,792
25	16	10^{-4}	3	31	11.55	38	17.18	6480
		10^{-2}		31	11.55	38	17.18	
		1		30	11.56	37	17.15	
		10^{2}		34	14.34	38	17.14	
		10^{4}		34	14.58	38	17.14	
5x5 checkerboard with $(16 \cdot 5)^2$ total elements								
20		10^{-4}	3			34	28.39	6480
		1				36	14.39	
		10^{4}				41	32.53	
25		10^{-4}	3	26	9.40	46	46.87	6480
		1		30	11.56	37	17.15	
		10^{4}		26	9.55	37	18.86	
30		10^{-4}	3			36	16.69	6480
		1				34	13.09	
		10^{4}				43	24.44	

$\rho(x) = \sigma$ in the upper right quadrant of the square domain and $\rho(x) = 1$ elsewhere. The bottom part of Table 5.3 shows results for a 5×5 checkerboard where $\rho(x) = 1$ in red squares and $\rho(x) = \sigma$ in black squares. Except for the checkerboard problem decomposed into 25 square subdomains, the subdomain boundaries are not aligned with jumps in material properties. Although our theory does not apply to such cases, good performance is evident for both regular and irregular domain decompositions.

Acknowledgments. The authors are grateful to Professor Fanghua Lin of the Courant Institute for introducing the third author to John domains and Poincaré's

inequality for very general domains. They are also grateful to Professors Piotr Hajłasz and Marcus Sarkis for their interest and insight.

REFERENCES

[1] G. Acosta, R. G. Durán, and M. A. Muschietti, *Solutions of the divergence operator on John domains*, Adv. Math., 206 (2006), pp. 373–401.

[2] B. Bojarski, *Remarks on Sobolev imbedding inequalities*, in Complex Analysis, Joensuu 1987, Lecture Notes in Math. 1351, Springer-Verlag, Berlin, 1988, pp. 52–68.

[3] J. Bramble, *A second order finite difference analog of the first biharmonic boundary value problem*, Numer. Math., 9 (1966), pp. 236–249.

[4] J. H. Bramble, J. E. Pasciak, and A. H. Schatz, *The construction of preconditioners for elliptic problems by substructuring.* I, Math. Comp., 47 (1986), pp. 103–134.

[5] S. Brenner and L.-Y. Sung, *Discrete Sobolev and Poincaré inequalities via Fourier series*, East-West J. Numer. Math., 8 (2000), pp. 83–92.

[6] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, Texts Appl. Math. 15, 2nd ed., Springer-Verlag, New York, 2002.

[7] S. M. Buckley and P. Koskela, *Sobolev-Poincaré implies John*, Math. Res. Lett., 2 (1995), pp. 577–593.

[8] C. R. Dohrmann, A. Klawonn, and O. B. Widlund, *A family of energy minimizing coarse spaces for overlapping Schwarz preconditioners*, in Proceedings of the 17th International Conference on Domain Decomposition Methods in Science and Engineering, Strobl, Austria, 2006, U. Langer, M. Discacciati, D. Keyes, O. Widlund, and W. Zulehner, eds., Lect. Notes Comput. Sci. Eng. 60, Springer-Verlag, Berlin, 2007.

[9] C. R. Dohrmann and O. B. Widlund, *An Overlapping Schwarz Algorithm for Almost Incompressible Elasticity*, manuscript, 2007.

[10] M. Dryja, M. V. Sarkis, and O. B. Widlund, *Multilevel Schwarz methods for elliptic problems with discontinuous coefficients in three dimensions*, Numer. Math., 72 (1996), pp. 313–348.

[11] M. Dryja, B. F. Smith, and O. B. Widlund, *Schwarz analysis of iterative substructuring algorithms for elliptic problems in three dimensions*, SIAM J. Numer. Anal., 31 (1994), pp. 1662–1694.

[12] H. Federer and W. H. Fleming, *Normal and integral currents*, Ann. of Math. 2, 72 (1960), pp. 458–520.

[13] I. G. Graham, P. Lechner, and R. Scheichl, *Domain decomposition for multiscale PDEs*, Numer. Math., (2007), pp. 589–626.

[14] P. Hajłasz, *Sobolev inequalities, truncation method, and John domains*, in Papers on Analysis, Rep. Univ. Jyväskylä Dep. Math. Stat. 83, University of Jyväskylä, Jyväskylä, 2001, pp. 109–126.

[15] G. Karypis and V. Kumar, *METIS Version* 4.0, University of Minnesota, Department of Computer Science, Minneapolis, MN, 1998.

[16] A. Klawonn, O. Rheinbach, and O. B. Widlund, *An analysis of a FETI–DP algorithm on irregular subdomains in the plane*, SIAM J. Numer Anal., to appear.

[17] A. Klawonn and O. B. Widlund, *Dual-primal FETI methods for linear elasticity*, Comm. Pure Appl. Math., 59 (2006), pp. 1523–1572.

[18] F. Lin and X. Yang, *Geometric Measure Theory—An Introduction*, Adv. Math. (Beijing/Boston) 1, Science Press, Beijing, 2002.

[19] J. Mandel, *Iterative solvers by substructuring for the p-version finite element method*, Comput. Methods Appl. Mech. Engrg., 80 (1990), pp. 117–128.

[20] V. G. Maz'ja, *Classes of domains and imbedding theorems for function spaces*, Soviet Math. Dokl., 1 (1960), pp. 882–885.

[21] V. G. Maz'ja, *Sobolev Spaces*, Springer Ser. Soviet Math., Springer-Verlag, Berlin, 1985, translated from the Russian by T. O. Shaposhnikova.

[22] V. G. Maz'ya and S. V. Poborchi, *Differentiable Functions on Bad Domains*, World Scientific, River Edge, NJ, 1997.

[23] J. Nečas, *Les méthodes Directes en Théorie des Équations Elliptiques*, Academia, Prague, 1967.

[24] M. Sarkis, *Nonstandard coarse spaces and Schwarz methods for elliptic problems with discontinuous coefficients using non-conforming elements*, Numer. Math., 77 (1997), pp. 383–406.

[25] M. V. Sarkis, *Partition of unity coarse spaces: Enhanced versions, discontinuous coefficients and applications to elasticity*, in Domain Decomposition Methods in Science and Engineer-

ing, I. Herrera, D. E. Keyes, O. B. Widlund, and R. Yates, eds., Natl. Auton. Univ. Mex., México, 2003, pp. 149–158.

[26] R. Scheichl and E. Vainikko, *Additive Schwarz and aggregation-based coarsening for elliptic problems with highly variable coefficients*, Computing, 80 (2007), pp. 319–343.

[27] B. F. Smith, P. Bjørstad, and W. Gropp, *Domain Decomposition: Parallel Multilevel Methods for Elliptic Partial Differential Equations*, Cambridge University Press, New York, 1996.

[28] A. Toselli and O. B. Widlund, *Domain Decomposition Methods - Algorithms and Theory*, Springer Ser. Comput. Math. 34, Springer-Verlag, Berlin, 2005.

SIAM J. NUMER. ANAL.
Vol. 46, No. 4, pp. 2169–2182
© 2008 Society for Industrial and Applied Mathematics

SMOOTHNESS ANALYSIS OF SUBDIVISION SCHEMES ON REGULAR GRIDS BY PROXIMITY*

PHILIPP GROHS†

Abstract. Subdivision is a powerful way of approximating a continuous object $f(x, y)$ by a sequence $((S^l p_{i,j})_{i,j\in\mathbb{Z}})_{l\in\mathbb{N}}$ of discrete data on finer and finer grids. The rule S that maps an approximation on a coarse grid $S^l p$ to the approximation on the next finer grid $S^{l+1}p$ is called a subdivision scheme. If for a given scheme S every continuous object $f(x, y)$ constructed by S is of C^k smoothness, then S is said to have smoothness order k. Subdivision schemes are well understood if they are linear. However, for various applications the data have values in a manifold which is not a vector space (for example, when our data are positions of a moving rigid body). Under these circumstances, subdivision schemes become nonlinear and much harder to analyze. One way of analyzing such schemes is to relate them to a given linear scheme and to establish a so-called proximity condition between the two schemes, which helps in proving that the two schemes share the same smoothness. The present paper uses this method to show the C^1-smoothness of a wide class of nonlinear multivariate schemes.

Key words. linear subdivision schemes, nonlinear subdivision schemes, smoothness, multivariate subdivision, lazy wavelets

AMS subject classifications. 41A05, 41A25, 26B05, 22E05, 53B20, 68U05

DOI. 10.1137/060669759

1. Introduction.

Subdivision is a powerful way of approximating a continuous object $f(x, y)$ by a sequence $((S^l p_{i,j})_{i,j\in\mathbb{Z}})_{l\in\mathbb{N}}$ of discrete data on finer and finer grids. The rule S that maps an approximation on a coarse grid $S^l p$ to the approximation on the next finer grid $S^{l+1}p$ is called a subdivision scheme. Generally this rule depends on l but we will assume that it is always the same rule. The subdivision scheme is then called stationary; this is the class of schemes in which we are interested. If for a given scheme S every continuous object $f(x, y)$ constructed by S is of C^k-smoothness, we say that the smoothness order of S is k. The fact that in this introduction we use a bivariate function $f(x, y)$ is immaterial; this paper is concerned with the general multivariate case.

To store a continuous object which is constructed via subdivision, one needs only to store a coarse approximation and the subdivision scheme constructing the object. This fact makes subdivision a valuable tool in computer-aided geometric design. Another application of subdivision schemes is data compression and noise removal via wavelet-type transforms associated with the subdivision scheme [2, 3]. All of these concepts can be applied to manifold-valued data using nonlinear subdivision schemes if the underlying nonlinear scheme is of sufficient smoothness [10, 4].

Subdivision schemes are especially well understood if they are linear. However, for various applications the data have values in a manifold which is not a vector space (for example, when our data are positions of a moving rigid body). Under these circumstances, subdivision schemes become nonlinear and much harder to analyze. One way of analyzing such schemes is to relate them to a given linear scheme and to establish

*Received by the editors September 14, 2006; accepted for publication (in revised form) January 24, 2008; published electronically May 2, 2008. This research was funded by the Austrian Science Fund under grant P18575.

http://www.siam.org/journals/sinum/46-4/66975.html

†Institut für Diskrete Mathematik und Geometrie, Technische Universität Wien, Wiedner Hauptstr. 8-10/104, A-1040 Wien, Austria (grohs@geometrie.tuwien.ac.at).

a so-called proximity condition between the two schemes, which helps in proving that the two schemes share the same smoothness. In [12, 14] this method has been used to prove smoothness properties for certain nonlinear curve schemes. The problem in proving higher-order smoothness is to verify sufficient proximity conditions. In [15] it has been shown that on the sphere and related manifolds any univariate nonlinear scheme that is constructed from an interpolatory linear scheme via closest-point projection shares the smoothness of the linear scheme.

All of these results apply only to univariate schemes. The present paper treats the general multivariate case and shows the C^1-smoothness of a wide class of nonlinear multivariate schemes. We follow the lines of [12] and also remove a condition on the decay rates of the linear scheme needed there.

1.1. An example: The Kobbelt scheme and its log-exp analogue. First we describe, as an example, the construction of a nonlinear subdivision scheme for matrix groups which is analogous to a well-known interpolatory linear scheme (we construct the log-exp version of Kobbelt's scheme [8]).

The Kobbelt scheme for regular input data is the bivariate tensor product rule of the interpolatory four-point scheme [5] with itself. It is defined as

$$\begin{aligned}
Sp_{2i,2j} &= p_{i,j}, \quad Sp_{2i+1,2j} = p_{i,j} - \frac{1}{16}(v_{-1,0} + 9v_{1,0} - v_{2,0}),\\
Sp_{2i,2j+1} &= p_{i,j} - \frac{1}{16}(v_{0,-1} + 9v_{0,1} - v_{0,2}),\\
Sp_{2i+1,2j+1} &= p_{i,j} + \frac{1}{256}(v_{2,2} - 9v_{2,1} - 9v_{2,0} + v_{2,-1} - 9v_{1,2} + 81v_{1,1} + 81v_{1,0} -\\
&\qquad 9v_{1,-1} - 9v_{0,2} + 81v_{0,1} - 9v_{0,-1} + v_{-1,2} - 9v_{-1,1} - 9v_{-1,0} + v_{-1,-1}),
\end{aligned}$$

where $v_{k,l} := p_{i+k,j+l} - p_{i,j}$ is the vector pointing from $p_{i,j}$ to $p_{i+k,j+l}$ and $(Sp_{i,j})_{i,j\in\mathbb{Z}}$ means the grid function produced from $(p_{i,j})_{i,j\in\mathbb{Z}}$ via subdivision.

The rules that define the usual Kobbelt scheme are all of the form

$$Sp_{2i+k,2j+l} = p_{i,j} + \sum_{r,s} \lambda^{k,l}_{r,s} v_{r,s}, \tag{1.1}$$

where the vector $\sum_{r,s} \lambda^{k,l}_{r,s} v_{r,s}$ is added to the point $p_{i,j}$. In a matrix group, this point-vector addition of a point p and a vector v is realized with the matrix-exponential function exp, as $p \cdot \exp(v)$. The vector pointing from a group element p to the group element q is given by

$$\overrightarrow{pq} = \log(p^{-1}q). \tag{1.2}$$

We can thus define an analogue of the Kobbelt scheme which works in matrix groups as follows: for the rule given by (1.1), we define its log-exp analogue as

$$Tp_{2i+k,2j+l} = p_{i,j} \cdot \exp\left(\sum_{r,s} \lambda^{k,l}_{r,s} v_{r,s}\right), \quad v_{r,s} = \log(p_{i,j}^{-1} p_{i+u,j+v}).$$

To what extent does this nonlinear scheme share the properties of the Kobbelt scheme? It is certainly interpolatory, but does it converge? If so, does it produce smooth limit functions? As we will see, the answer to both questions is yes.

1.2. Other ways of constructing nonlinear analogues of linear schemes. Aside from the log-exp approach there are two other nonlinear variants that we want to study. They both rely on the observation in [14] that every affinely invariant scheme can be expressed in terms of the affine averaging operator

$$\mathrm{av}_\lambda(x,y) = \lambda x + (1-\lambda)y, \quad \lambda \in \mathbb{R}.$$

In Euclidean space $\mathbb{R}^n$, the shortest path from a point x to another point y is described by the curve $c(t) := tx + (1-t)y$. We see that $\mathrm{av}_\lambda(x,y)$ is given by $c(\lambda)$. This can be generalized to surfaces in $\mathbb{R}^n$, and Riemannian manifolds as well: in such spaces geodesics play the role of straight lines, and if we define the *geodesic average* $\text{g-av}_\lambda(x,y) := c(\lambda)$, where c is the shortest geodesic, with $c(0) = x$ and $c(1) = y$, we can replace every ocurrence of the av_λ-operator by the g-av_λ-operator. This gives us an analogue of a linear scheme that assumes its values on a given manifold. A unique shortest geodesic connecting two points need not always exist. However, it always exists locally, and, if the distances between consecutive points in the grid function we want to subdivide are sufficiently small, we will not run into any problems (cf. the discussion in [12]).

Another approach is to apply a projection operator to the averages. A projection operator for a surface $\mathcal{M} \subseteq \mathbb{R}^n$ in our setting is a smooth function $P : \mathbb{R}^n \to \mathcal{M}$ with the property that $P(x) = x$ for all $x \in \mathcal{M}$. For the nonlinear scheme we apply the operator $P \circ \mathrm{av}_\lambda(x,y)$ instead of $\mathrm{av}_\lambda(x,y)$. This defines the *projection analogue* of a linear scheme.

2. Outline of this paper. In this work we analyze nonlinear subdivision schemes that satisfy a proximity condition with a certain linear scheme and show that if the linear scheme is convergent (resp., smooth), then the nonlinear scheme is convergent (resp., smooth). In section 3 we review some basic concepts of subdivision. Section 4 derives results of the linear theory needed for our analysis. Section 5 shows that if a scheme is in proximity with a linear convergent scheme, it is convergent. Section 6 does the same with smoothness. In section 7 we verify that a proximity condition holds for various nonlinear analogues of linear schemes. Section 8 summarizes the results, until then valid only for surfaces and matrix groups, and extends them to abstract Lie groups and Riemannian manifolds.

3. Preliminaries. Let $(p_\gamma)_{\gamma\in\mathbb{Z}^s}$ be a bounded grid function of points $p_\gamma \in \mathbb{R}^n$ for $n \in \mathbb{N}$. In that case $(p_\gamma)_{\gamma\in\mathbb{Z}^s}$ can be viewed as an element of $(l^\infty(\mathbb{Z}^s))^n$. A subdivision scheme on a regular grid is a mapping that maps such a grid function to a new grid function $(Sp_\gamma)_{\gamma\in\mathbb{Z}^s}$ under the condition that there exists an integer $N > 1$ with the property that for all grid functions p, q with $q_\gamma = p_{\gamma+\delta_i}$, where δ_i is the ith canonical basis vector in $\mathbb{Z}^s$, the relation $(Sq)_\gamma = (Sp)_{\gamma+N\delta_i}$ holds. N is called the dilation factor of S (typically $N = 2$). The case $s = 1$ corresponds to curve subdivison and $s = 2$ to surface subdivision as in our example in section 1. A subdivision scheme S is called linear if S is a linear mapping. It is easy to see that for a linear scheme there exists a sequence $(S_\gamma)_{\gamma\in\mathbb{Z}^s}$ with the property

$$(Sp)_\gamma = \sum_{\beta\in\mathbb{Z}^s} S_{\gamma-N\beta} p_\beta.$$

This sequence is called the mask of S. We consider only subdivision schemes with a finite mask, meaning that $S_\gamma \neq 0$ only for a finite number of γ's. Thus the assumption

that the grid function $(p_\gamma)_{\gamma\in\mathbb{Z}^s}$ is bounded is not actually a restriction, because a finite mask means that the subdivison operator is local.

If the elements S_γ of the mask are $n\times n$ matrices, one speaks of matrix subdivision schemes, and if they are scalars, i.e., elements of $\mathbb{R}$, we speak of scalar subdivision schemes. In this paper we are concerned only with the analysis of scalar subdivision schemes.

An interpretation of the definition of a subdivision scheme would be to view $(p_\gamma)_{\gamma\in\mathbb{Z}^s}$ as a discrete approximation of some continuous data defined in some domain in $\mathbb{R}^s$. The subdivision operator S maps this approximation to a finer approximation on the grid $\frac{1}{N}\mathbb{Z}^s$, S^2 maps the approximation on the even finer grid $\frac{1}{N^2}\mathbb{Z}^s$, and so on. If this sequence of approximations converges to a function defined on $\mathbb{R}^s$, the subdivision scheme is called convergent or, more precisely (see [1, Definition 2.1]), the following.

DEFINITION 3.1. *Let $S : (l^\infty(\mathbb{Z}^s))^n \to (l^\infty(\mathbb{Z}^s))^n$ be a subdivision scheme with dilation factor N. S is called convergent if for any $p \in (l^\infty(\mathbb{Z}^s))^n$ there exists a continuous function $f : \mathbb{R}^s \to \mathbb{R}^n$ with*

$$\lim_{k\to\infty} \max_{\gamma\in\mathbb{Z}^s} |(S^k p)_\gamma - f(N^{-k}\gamma)| = 0,$$

where $|\cdot|$ is a fixed norm on $\mathbb{R}^n$. Moreover, if always $f \in C^k(\mathbb{R}^s, \mathbb{R}^n)$, S is called C^k, $k \in \mathbb{N}$.

Clearly this definition is equivalent to saying that the functions $\mathcal{F}_l(S^l p)$, given by the piecewise s-multilinear interpolation of $S^l p$ on $\frac{1}{N^l}\mathbb{Z}^s$, converge uniformly to f, as $l \to \infty$.

4. Linear theory. We review some results from the theory of linear schemes. For further information concerning linear schemes, the reader may consult [1, 6, 17, 7]. We let $n = 1$; i.e., the points p_γ are elements of $\mathbb{R}$. This is no restriction, since convergence and smoothness properties can be treated componentwise in the case of scalar linear schemes. In the univariate setting, the (unique) difference operator plays an important role. On grids we look at all of the difference operators in the different directions on the grid simultaneously. This allows us to define derived schemes as discrete analogues of the left derivatives.

Let Δ_i be the operator that maps grid function $(p_\gamma)_{\gamma\in\mathbb{Z}^s}$ to the grid function $(p_\gamma - p_{\gamma-\delta_i})_{\gamma\in\mathbb{Z}^s}$, where δ_i denotes the ith canonical basis vector in $\mathbb{Z}^s$.

DEFINITION 4.1. *For $p \in l^\infty(\mathbb{Z}^s)$ we define $\mathbf{\Delta} p \in (l^\infty(\mathbb{Z}^s))^s$ as*

$$\mathbf{\Delta} p := \big(\Delta_1 p\ ,\ \Delta_2 p\ ,\ \dots\ ,\ \Delta_s p\big)^T.$$

For a subdivision scheme to converge, we need the distance between consecutive points to become small.

DEFINITION 4.2. *$D(p) := \sup_{\gamma\in\mathbb{Z}^s} \|\mathbf{\Delta} p_\gamma\|$, where $\|\mathbf{\Delta} p_\gamma\| := \max_{i=1,\dots,s} |\Delta_i p_\gamma|$.*

While a linear scheme is always defined for all input data, this is no longer true for a nonlinear subdivision scheme T, which is, in general, defined only on certain grid functions with values in some subset $\mathcal{D}(T) \subseteq \mathbb{R}^n$ and maps such grid functions to grid functions with values in $\mathcal{D}(T)$. The definition of convergence and smoothness is the same as in the linear case; it has to hold only for all grid functions where T is defined. Typically the set $\mathcal{D}(T)$ will be a manifold.

The mth derived scheme of a linear multivariate subdivision scheme S is defined by the relation

$$S^{[m]}\mathbf{\Delta}^m = N^m \mathbf{\Delta}^m S \quad \text{for } m \geq 1, \quad \text{and } S^{[0]} = S.$$

Note that for $s > 1$ the derived schemes are always matrix subdivision schemes. It is possible to characterize the smoothness of a linear subdivision scheme by the decay of the derived schemes.

THEOREM 4.3. *Let S be a linear subdivision scheme. If there exists a constant $\mu_0 < 1$ such that for some $l \in \mathbb{Z}^+$*

$$D(S^l p) \le \mu_0 D(p) \text{ for all } p, \tag{4.1}$$

then S is convergent. Convergence of S implies the existence of $S^{[1]}$.

If in addition there exists a constant $\mu_1 < 1$ with

$$D((S^{[1]})^l \mathbf{\Delta} p) \le \mu_1 D(\mathbf{\Delta} p) \text{ for all } p, \tag{4.2}$$

then S is C^1.

The sufficient conditions (4.1) and (4.2) are almost necessary for convergence (resp., smoothness).

THEOREM 4.4 (see [6]). *Let S be a convergent linear subdivison scheme. Then $S^{[1]}$ exists and there is $l \in \mathbb{Z}^+$ such that S^l satisfies* (4.1) *with $\mu_{0l} = \frac{1}{N^l}\|(S^{[1]})^l|_{\mathbf{\Delta}}\| < 1$. If in addition S is C^1 and $S^{[2]}$ exists, then there exists $l \in \mathbb{Z}^+$ such that $(S^{[1]})^l$ satisfies* (4.2) *with $\mu_{1l} = \frac{1}{N^l}\|(S^{[2]})^l|_{\mathbf{\Delta}^2}\| < 1$.*

Remark 4.5. With the notation of Theorem 4.4 we may assume that there exists a constant A such that $\|(S^{[1]})^k|_{\mathbf{\Delta}}\| < A$ for all $k \in \mathbb{Z}^+$. This follows from the fact that the derived scheme $S^{[1]}$ converges to the first derivative of the limit function of S. This immediately implies that the spectral radius

$$\rho_\infty\big((S^{[1]})|_{\mathbf{\Delta}}\big) = \liminf_{k\to\infty} \|(S^{[1]})^k|_{\mathbf{\Delta}}\|^{1/k} = 1. \tag{4.3}$$

Remark 4.6. The existence of the first derived scheme $S^{[1]}$ is equivalent to a purely algebraic condition on the mask of S, namely,

$$\sum_{\beta\in\mathbb{Z}^s} S_{\gamma - N\beta} = 1 \quad \text{for all } \gamma \in \mathbb{Z}^s.$$

This property of S is in the literature usually referred to as *reproduction of constants*.

For a linear convergent scheme S, the rate of convergence is known (compare [6, Proposition 6.7]):

LEMMA 4.7. *Let S be a linear subdivision scheme, and assume that $S^{[1]}$ exists. Then there exists a constant $C > 0$ with*

$$\|\mathcal{F}_{j+1}(Sp) - \mathcal{F}_j(p)\|_\infty \le C D(p) \tag{4.4}$$

and

$$\sup_{\gamma\in\mathbb{Z}^s} \inf_{\beta\in\mathbb{Z}^s} |(Sp)_\gamma - p_\beta| \le C D(p). \tag{4.5}$$

If in addition $S^{[2]}$ exists, then there exists another constant $C > 0$ with

$$\|\mathcal{F}_{j+1}(S^{[1]}\mathbf{\Delta} p) - \mathcal{F}_j(\mathbf{\Delta} p)\|_\infty \le C D(\mathbf{\Delta} p). \tag{4.6}$$

Here, $\mathcal{F}_j$ is applied componentwise.

5. Convergence and well definedness. In this section we show that a nonlinear scheme is convergent if it satisfies a proximity condition with a linear and convergent scheme. Since most of the nonlinear schemes we are interested in are not defined for all input data, we have to restrict our analysis to initial data defined in certain subsets of $\mathbb{R}^n$. The following type of subsets is useful.

DEFINITION 5.1. *$\mathcal{P}_{M,\varepsilon}$ denotes the set of grid functions p, whose values p_γ are contained in a subset M of the domain of definition $\mathcal{D}(T)$ of T and which fulfill the condition $D(p) \leq \varepsilon$.*

We want to obtain convergence and smoothness properties for nonlinear schemes derived from linear ones, and we would like these properties to be directly related to the respective properties of the original linear schemes. For this purpose we need a condition that ensures that the nonlinear scheme is not too far apart from the linear scheme. As we will see, the following proximity condition serves us well. The next definition is the multivariate analogue of Definition 7 in [12].

DEFINITION 5.2. *Let S and T be two subdivision schemes. We say that S and T satisfy a proximity condition for the class $\mathcal{P}_{M,\varepsilon}$ of order $\alpha > 1$ if there is a constant C such that, for all $p \in \mathcal{P}_{M,\varepsilon}$,*

$$\sup_{\gamma \in \mathbb{Z}^s} |Sp_\gamma - Tp_\gamma| \leq C \cdot D(p)^\alpha. \tag{5.1}$$

From Theorem 4.4 we know that for a convergent linear scheme S there exists a contractivity constant $\mu_0 < 1$, with

$$D(S^l p) \leq \mu_0 D(p) \tag{5.2}$$

for some $l \in \mathbb{Z}^+$. The idea about how to prove that a nonlinear scheme T in proximity with a linear convergent scheme S is convergent is as follows: from the proximity condition we show that if S satisfies (5.2), i.e., is convergent, then T satisfies a similar condition. After that we show that if such a condition is satisfied for a nonlinear scheme T in proximity with a linear scheme S, the scheme T is itself convergent. We make the following definition.

DEFINITION 5.3. *A subdivision scheme T is said to satisfy a convergence condition with factor $\mu_0 < 1$ and the class $\mathcal{P}_{M,\varepsilon}$ if there exists $l \in \mathbb{Z}^+$ such that*

$$D(T^l p) \leq \mu_0 D(p) \tag{5.3}$$

for all initial data $p \in \mathcal{P}_{M,\varepsilon}$.

The following central lemma shows that a nonlinear scheme T satisfies a convergence condition if it is in proximity with a convergent linear scheme. Unfortunately, the proof of (5.3) requires us to deal with the additional difficulty that choosing initial data in a set $\mathcal{P}_{M,\varepsilon}$ will result in data with values no longer contained in M. We are therefore forced to employ various different sets of type $\mathcal{P}_{M,\varepsilon}$. Essentially we must show that it is possible to restrict the domain of definition of T a little bit such that all iterates of T are well defined and additionally (5.3) is valid.

In what follows we denote the constant from (4.5) by C'. Recall that the subdivision scheme T maps input data with values in its domain of definition $\mathcal{D}(T)$ to data with values in $\mathcal{D}(T)$. We use the notation $\rho \cdot B$ for the ball of radius ρ, and $M + \rho \cdot B$ is a Minkowski sum; i.e., the set of points whose distance from M does not exceed ρ (w.r.t. the norm $|\cdot|$ on $\mathbb{R}^n$).

LEMMA 5.4. *Let S and T satisfy the proximity condition as in Definition 5.2 for input data in $\mathcal{P}_{M,\varepsilon}$. Assume that S is linear and satisfies a convergence condition*

with factor $\mu_0 < 1$. Then for any $M' \subseteq M$ which stays away from the boundary of M, i.e., $\exists\, \rho > 0$ with

$$(M' + \rho \cdot B) \cap \mathcal{D}(T) \subseteq M, \tag{5.4}$$

there is $\delta \in (0, \varepsilon)$ and $\overline{\mu_0} < 1$ such that data $T^l p$ are contained in $\mathcal{P}_{M,\delta}$ for all l and T satisfies a convergence condition with factor $\overline{\mu_0}$ for input data p in $\mathcal{P}_{M',\delta}$. By choosing δ small enough, we can achieve that $\overline{\mu_0} - \mu_0$ is arbitrarily small.

Proof. To make the proof easier to follow, we assume without loss of generality that the convergence condition for S is satisfied with $l = 1$. We choose $\delta \in (0, \varepsilon)$ small enough to satisfy the following two conditions:

$$\overline{\mu_0} := \mu_0 + 2C\delta^{\alpha-1} < 1 \tag{5.5}$$

and

$$\delta \left(\frac{C'}{1 - \overline{\mu_0}} + \frac{C\delta^{\alpha-1}}{1 - \overline{\mu_0}^{\alpha}} \right) < \rho. \tag{5.6}$$

Condition (5.6) will ensure that the sequence $T^l p$ stays within M. We show that for $p \in \mathcal{P}_{M',\delta}$ the norm of the difference sequence $D(Tp)$ can be estimated by $\overline{\mu_0} D(p)$:

$$D(Tp) \le D(Sp - Tp) + D(Sp) \le 2 \sup_{\gamma \in \mathbb{Z}^s} |Sp_\gamma - Tp_\gamma| + D(Sp).$$

Since S is convergent with contractivity constant μ_0 and T is in proximity with S, we can deduce that

$$D(Tp) \le 2CD(p)^\alpha + \mu_0 D(p) \le D(p)(\mu_0 + 2C\delta^{\alpha-1}) = \overline{\mu_0} D(p). \tag{5.7}$$

Now we show that Tp is still contained in $\mathcal{P}_{M,\delta}$:

$$\begin{aligned} \sup_{\gamma \in \mathbb{Z}^s} \inf_{x \in M'} |Tp_\gamma - x| &\le \sup_{\gamma \in \mathbb{Z}^s} \inf_{\beta \in \mathbb{Z}^s} |Tp_\gamma - p_\beta| \\ &\le \sup_{\gamma \in \mathbb{Z}^s} \inf_{\beta \in \mathbb{Z}^s} (|Tp_\gamma - Sp_\gamma| + |Sp_\gamma - p_\beta|) \\ &\le \sup_{\gamma \in \mathbb{Z}^s} |Sp_\gamma - Tp_\gamma| + C'D(p) \le CD(p)^\alpha + C'D(p) < \rho. \end{aligned}$$

By (5.4) and the fact that $Tp_\gamma \in \mathcal{D}(T)$, the grid function Tp takes values in M. Together with (5.7) we have $Tp \in \mathcal{P}_{M,\delta}$.

Now we use induction on l to prove that $D(T^l p) \le \overline{\mu_0}^l D(p)$ and that $T^l p \in \mathcal{P}_{M,\delta}$ under the assumption that $T^{l-1} p \in \mathcal{P}_{M,\delta}$ and $D(T^{l-1} p) \le \overline{\mu_0}^{l-1} D(p)$. First,

$$D(T^l p) \le D(T^l p - ST^{l-1} p) + D(ST^{l-1} p) \le 2 \sup_{\gamma \in \mathbb{Z}^s} |ST^{l-1} p_\gamma - TT^{l-1} p_\gamma| + D(ST^{l-1} p).$$

By the proximity condition and the convergence condition satisfied by S, we get

$$D(T^l p) \le 2CD(T^{l-1} p)^\alpha + \mu_0 D(T^{l-1} p) \le \overline{\mu_0} D(T^{l-1} p) \le \overline{\mu_0}^l D(p).$$

It remains to show that $T^l p \in \mathcal{P}_{M,\delta}$ (for initial data $p \in \mathcal{P}_{M',\delta}$). To this end we

estimate

$$\begin{aligned}\sup_{\gamma\in\mathbb{Z}^s}\inf_{x\in M'}|T^l p_\gamma - x| &\le \sup_{\gamma\in\mathbb{Z}^s}\inf_{\beta\in\mathbb{Z}^s}|T^l p_\gamma - p_\beta| \le \sum_{j=1}^{l}\sup_{\gamma\in\mathbb{Z}^s}\inf_{\beta\in\mathbb{Z}^s}|T^j p_\gamma - T^{j-1}p_\beta| \\ &\le \sum_{j=1}^{l}\sup_{\gamma\in\mathbb{Z}^s}\inf_{\beta\in\mathbb{Z}^s}(|T^j p_\gamma - ST^{j-1}p_\gamma| + |ST^{j-1}p_\gamma - T^{j-1}p_\beta|) \\ &\le \sum_{j=1}^{l} CD(T^{j-1}p)^\alpha + C'D(T^{j-1}p) \\ &\le \sum_{j=0}^{l-1} C\overline{\mu_0}^{j\alpha}D(p)^\alpha + C'\overline{\mu_0}^{j}D(p) \\ &< \rho.\end{aligned}$$

The last estimate follows from summing the geometric series and using condition (5.6) on δ.

The same reasoning as for the case $l = 1$ above and the assumption (5.4) establishes the fact that $T^l p \in \mathcal{P}_{M,\delta}$, and the induction is complete.

The statement regarding the difference $\overline{\mu_0} - \mu_0$ is immediate. ∎

PROPOSITION 5.5. *Under the assumptions of Lemma* 5.4*, the subdivision scheme* T *is convergent for all initial data* $p \in \mathcal{P}_{M',\delta}$.

Proof. We have

$$\begin{aligned}\|\mathcal{F}_{j+1}(T^{j+1}p) - \mathcal{F}_j(T^j p)\|_\infty &\le \|\mathcal{F}_{j+1}(T^{j+1}p) - \mathcal{F}_{j+1}(ST^j p)\|_\infty \\ &\quad + \|\mathcal{F}_{j+1}(ST^j p) - \mathcal{F}_j(T^j p)\|_\infty \\ &\le C\cdot D(T^j p)^\alpha + C'\cdot D(T^j p)\end{aligned}$$

by the proximity condition and by (4.4). By Proposition 5.4, this expression is bounded by a factor times μ^j, with $\mu < 1$. It follows that $T^j p$ is a Cauchy sequence with respect to the sup norm. ∎

6. Smoothness. In this section we show that a nonlinear scheme that is in proximity with a linear smooth scheme is itself smooth. To facilitate our analysis, we assume that the second derived scheme exists, which is not really a restriction (see [1]). We begin with a sufficient condition for a possibly nonlinear subdivision scheme to converge to a smooth limit (see [7] for a proof).

LEMMA 6.1. *Let* T *be a subdivision scheme and* $p^l := T^l p$*. Assume that* $\mathcal{F}_l(p^l)$ *converges to a continuous function* f *in the sup norm. If* $g_l := \mathcal{F}_l(N^l\mathbf{\Delta}p^l)$ *is a Cauchy sequence and* $\lim_{l\to\infty} D(N^l\mathbf{\Delta}p^l) = 0$*, then* f *is* C^1.

DEFINITION 6.2. *A subdivision scheme* T *is said to satisfy a mixed smoothness condition if* T *satisfies a convergence condition and there is* $\mu_1 < 1$ *such that*

$$D(N\mathbf{\Delta}T^l p) \le \mu_1^l P_1(l)D(p) \text{ for all } l, p, \tag{6.1}$$

where P_1 *is a linear function with positive coefficients.*

LEMMA 6.3. *Suppose that* S *and* T *satisfy a proximity condition for* $p \in \mathcal{P}_{M,\varepsilon}$ *with* $\alpha > 1$ *and that* S *satisfies a smoothness condition with factors* μ_0, μ_1 *such that*

$$\mu_0 < \mu_0^* = N^{-1/\alpha}, \quad \mu_1 < 1. \tag{6.2}$$

Then there is a $\delta > 0$ such that T satisfies a mixed smoothness condition with factors $\overline{\mu_0}, \overline{\mu_1}$ which also satisfy (6.2)*, for all $p \in \mathcal{P}_{M',\delta}$, where M' is a subset of M satisfying* (5.4).

Proof. By Proposition 5.4 we know that there exists a $\delta > 0$ and a $\overline{\mu_0} < 1$ so that

$$D(T^l p) \le \overline{\mu_0}^l D(p) \text{ for } D(p) < \delta.$$

We want to show that there is a $\overline{\mu_1} < 1$ and a linear polynomial with positive coefficients $\overline{P_1}$ such that $D(N^l \mathbf{\Delta} T^l p) \le \overline{\mu_1}^l \overline{P_1}(l) D(p)$, for all p with $D(p) < \delta$. We write $d_l := D(N^l \mathbf{\Delta} T^l p)$ and $q := T^{l-1} p$. Next we proceed just as in the proof of Theorem 5.4 and make use of the fact that $\sup_{\gamma \in \mathbb{Z}^s} \|\mathbf{\Delta} S q_\gamma - \mathbf{\Delta} T q_\gamma\| \le 2 \sup_{\gamma \in \mathbb{Z}^s} |Sq - Tq|$.

$$\begin{aligned} d_l = D(N^l \mathbf{\Delta} T q) &\le D(N^l \mathbf{\Delta} S q) + 2N^l \sup_{\gamma \in \mathbb{Z}^s} \|\mathbf{\Delta} S q_\gamma - \mathbf{\Delta} T q_\gamma\| \\ &\le D(N^l \mathbf{\Delta} S q) + 4N^l \sup_{\gamma \in \mathbb{Z}^s} |S q_\gamma - T q_\gamma| \\ &\le \mu_1 D(N^{l-1} \mathbf{\Delta} q) + 4N^l C D(q)^\alpha, \end{aligned} \tag{6.3}$$

because of the smoothness condition satisfied by S and the proximity condition. We can estimate the last summand as follows:

$$4N^l C D(q)^\alpha \le 4N^l C (\overline{\mu_0}^{l-1} D(p))^\alpha = 4NC (N \overline{\mu_0}^\alpha)^{l-1} D(p)^\alpha.$$

Because of Proposition 5.4, (6.2), and the fact that we can make $D(p)$ arbitrarily small, we can choose $\overline{\mu_0} := \overline{\mu_0} \le N^{-1/\alpha} < 1$, and thus

$$4CN^l D(q)^\alpha \le \tilde{\mu_1}^{l-1} D(p)^\alpha P_0,$$

with a positive constant P_0. Substituting into (6.3) yields

$$d_l \le d_{l-1} \mu_1 + P_0 D(p)^\alpha \sum_{j=0}^{l-1} \mu_1^{l-j-1} \tilde{\mu_1}^j.$$

Now we define $\overline{\mu_1} := \max\{\mu_1, \tilde{\mu_1}\}$ and get

$$d_l \le \overline{\mu_1}^l d_0 + D(p)^\alpha \overline{\mu_1}^{l-1} l P_0.$$

There is a constant $C' > 0$, with $\frac{1}{\delta^{\alpha-1} C'} < \overline{\mu_1}$, so that we have

$$D(p)^\alpha \overline{\mu_1}^{l-1} \le D(p) C' \overline{\mu_1}^l.$$

This shows that $d_l \le \overline{\mu_1}^l \left(D(\mathbf{\Delta} p) + C' D(p) l P_0\right) \le \overline{\mu_1}^l (2 + C' l P_0) D(p)$, and the proof is complete. ∎

Condition (6.2) is automatically fulfilled if the second derived scheme of S exists.

LEMMA 6.4. *Assume that S produces C^1 limit functions. If the derived schemes $S^{[1]}$ and $S^{[2]}$ exist, then* (6.2) *holds for an iterate of S.*

Proof. For the iterated schemes S^k condition (6.2) reads $\mu_{0k} \le N^{-k/\alpha}$, where μ_{0k} is the first contractivity constant of S^k given by $\frac{\|(S^{[1]})^k|_{\mathbf{\Delta}}\|}{N^k}$. Thus the statement of the lemma is equivalent to

$$\rho_\infty(\|S^{[1]}|_{\mathbf{\Delta}}\|) < N^{1-1/\alpha}. \tag{6.4}$$

By Remark 4.5 the spectral radius of the first derived scheme is equal to 1, and thus (6.4) is satisfied for every $\alpha > 1$. ∎

Now we can show that a nonlinear scheme T in proximity with a linear C^1 scheme that satisfies (6.2) is itself C^1.

PROPOSITION 6.5. *Let S be linear and of finite mask with C^1 limits such that $S^{[1]}$ and $S^{[2]}$ exist. Then the limit function $\lim_{l\to\infty} \mathcal{F}_l(T^l p)(x_1,\dots,x_s)$ is C^1.*

Proof. First of all we have $D(N^l \mathbf{\Delta} T^l p) \to 0$ by Proposition 6.3. Since we assume the affine invariance of the derived scheme $S^{[1]}$, (4.6) implies that

$$\|\mathcal{F}_{j+1}(S^{[1]}\mathbf{\Delta} p) - \mathcal{F}_j(\mathbf{\Delta} p)\|_\infty \le C \cdot D(\mathbf{\Delta} p),$$

with a constant $C > 0$. Note that the smoothness condition is nothing but a convergence condition for the derived scheme $S^{[1]}$. Then

$$\begin{aligned}
&\|\mathcal{F}_{l+1}(N^{l+1}\mathbf{\Delta} T^{l+1} p) - \mathcal{F}_l(N^l \mathbf{\Delta} T^l p)\|_\infty \\
&\le \|\mathcal{F}_{l+1}(N\mathbf{\Delta} T - S^{[1]}\mathbf{\Delta}) N^l T^l p\|_\infty + \|(\mathcal{F}_{l+1} S^{[1]} - \mathcal{F}_l)\mathbf{\Delta} N^l T^l p\|_\infty \\
&\le (2N) \sup_{\gamma\in\mathbb{Z}^s} |(T-S)(N^l T^l p)_\gamma| + C D(\mathbf{\Delta} N^l T^l p) \\
&\le (2N) C' N^l (\mu_0 D(p))^{\alpha l} + C \mu_1^l P_1(l) D(p) =: \beta_l.
\end{aligned}$$

By Lemma 6.4 we may assume that $\mu_0^\alpha N < 1$ and $\sum_l \beta_l < \infty$. Therefore the sequence $\mathcal{F}_l(N^l \mathbf{\Delta} T^l p)$ is a Cauchy sequence, so we can apply Lemma 6.1 to complete the proof. ∎

7. Verification of the proximity conditions. In order to show that a proximity condition actually holds between an affinely invariant subdivision scheme and the perturbations that we want to study, we use a generalization of Lemma 5 in [12].

LEMMA 7.1. *Let S be a linear subdivision scheme on $\mathbb{Z}^s$ that reproduces constants. Then S is expressible by the repeated application of the affine averaging operator given by*

$$\mathrm{av}_\lambda(x,y) = \lambda x + (1-\lambda) y, \quad \lambda \in [0,1].$$

Assume that $(f_\lambda(x,y))_{\lambda\in[0,1]}$ is a family of functions from $\mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ satisfying

$$|\mathrm{av}_\lambda(x,y) - f_\lambda(x,y)| \le C \cdot |x-y|^2 \tag{7.1}$$

for a constant $C > 0$. If T is defined by exchanging the affine averaging operators av_λ by f_λ, then S and T satisfy a proximity condition of order 2 in the sense of Definition 5.2.

Proof. The fact that S is expressible via the affine averaging operator is just Theorem 1 in [12]. For the second assertion we perform induction on the number of iterations of averaging operators. For a subdivision scheme with dilation factor N we have N^s different rules, and for each of them we have to verify

$$|S p_{\alpha+N\beta} - T p_{\alpha+N\beta}| \le C \cdot D(p)^2. \tag{7.2}$$

Let m be the number of iterations of the averaging operator for the rule associated with $\beta \in \mathbb{Z}^s$. For $m = 1$ we get

$$S p_{\alpha+N\beta} = \mathrm{av}_{\lambda_\beta}(p_{\alpha+r_\beta}, p_{\alpha+s_\beta}).$$

Then by (7.1) there exists a constant C such that (7.2) holds true. Now we perform an induction step. Assume that our assertion is correct for m. Let y and z be

points obtained from at most m iterations of affine averaging operators and y', z' the corresponding points obtained by replacing the averaging operators by f. By the induction hypothesis,

$$|y - y'|,\ |z - z'| \le C' \cdot D(p)^2, \text{ with a constant } C' > 0. \tag{7.3}$$

Clearly we also have

$$|y - z| \le C \cdot D(p), \tag{7.4}$$

with a constant $C > 0$. Let $x := \mathrm{av}_\lambda(y, z)$, $x' := f_\lambda(y', z')$, and $x'' := \mathrm{av}_\lambda(y', z')$. We need to show that there exists a constant $\tilde{C} > 0$, with

$$|x - x'| \le \tilde{C} \cdot D(p)^2. \tag{7.5}$$

We estimate

$$\begin{aligned} |x - x'| &\le |x' - x''| + |x - x''| \le C'' \cdot |y' - z'|^2 + |\mathrm{av}_\lambda(y - y', z - z')| \\ &\le C'' \cdot (|y - y'| + |y - z| + |z - z'|)^2 + C''' \cdot \max(|y - y'|, |z - z'|). \end{aligned}$$

Because of (7.3) and (7.4) this implies (7.5). ∎

7.1. Geodesic averaging. For a surface embedded in the Euclidean space $\mathbb{R}^n$ we construct the geodesic analogue of a linear scheme S by replacing the affine averaging operators av_λ by the geodesic averaging operators g-av_λ, where $\text{g-av}_\lambda(x, y) := c(\lambda)$, with c the shortest geodesic with $c(0) = x$ and $c(1) = y$. Lemma 4 in [12] states that (7.1) is fulfilled with $f_\lambda = \text{g-av}_\lambda$, and therefore, by Lemma 7.1, a proximity condition holds between a linear scheme and its geodesic analogue.

7.2. Projection analogue. Another possibility of perturbing a linear scheme that is studied in this paper is obtained by applying a projection operator P to the subdivided points; i.e., we set $f_\lambda := P \circ \mathrm{av}_\lambda$. P is required only to be smooth and leave the points of the manifold fixed. In [12] it has been shown that (7.1) holds for this class of schemes, and therefore a proximity condition holds between S and the perturbed scheme.

7.3. Log-exp scheme. We go on to prove that a proximity condition holds between a linear scheme S and its log-exp analogue T. In [13] a different log-exp analogue is analyzed, and a proximity condition that ensures C^2-smoothness is proven. However, this different log-exp analogue is not capable of being generalized to the multivariate setting.

PROPOSITION 7.2. *Let S be a linear subdivision scheme that reproduces constants and T its G-valued* log-exp *analogue for a matrix group G. Then for every bounded subset $M \subseteq G$ such that* $\exp|_M$ *is a diffeomorphism, S and T satisfy a proximity condition for initial data p with $D(p)$ small enough.*

Proof. The matrix exponential function is given by

$$\exp(v) = I + v + \frac{v^2}{2} + \frac{v^3}{3!} \dots,$$

where $v \in \mathbb{R}^{m \times m}$. Since exp is a diffeomorphism in a neighborhood of the identity matrix I, we can choose $\delta_1 > 0$ such that $\|\exp(v) - I\| < \delta_1$ implies that

$$\alpha_1 \|\exp(v) - I\| \le \|v\| \le \alpha_2 \|\exp(v) - I\| \quad \text{for some } \alpha_1,\ \alpha_2 > 0, \tag{7.6}$$

where $\|v\| := \sqrt{tr(v^T v)}$, the Frobenius norm.

Since we are operating in a bounded subset of G, we can assume that $\|p_\alpha\| \leq \gamma_1$, $\|p_\alpha^{-1}\| \leq \gamma_2$ for some $\gamma_1, \gamma_2 > 0$.

As explained in section 1 the construction of new points in a linear subdivision scheme S and its log-exp analogue T has the form

$$\begin{aligned} s &= p + \sum_{\beta \in F} \lambda_\beta (q_\beta - p) \qquad \text{(linear)}, \\ t &= p \cdot \exp\Big(\sum_{\beta \in F} \lambda_\beta v_\beta \Big) \quad \text{(nonlinear)}, \end{aligned}$$

where $q_\beta = p \cdot \exp(v_\beta)$ and F is a finite subset of $\mathbb{Z}^s$. We show that

$$\|s - t\| \leq C \cdot \max_{\beta \in F} \|q_\beta - p\|^2 \quad \text{for a } C > 0, \tag{7.7}$$

which is clearly equivalent to the required proximity condition

$$\begin{aligned} \|s - t\| &= \left\| p \cdot \left(I + \sum_{\beta \in F} \lambda_\beta (\exp(v_\beta) - I) - \exp\left(\sum_{\beta \in F} \lambda_\beta v_\beta \right) \right) \right\| \\ &\leq \gamma_1 \left\| \sum_{\beta \in F} \lambda_\beta \left(\frac{v_\beta^2}{2} + \frac{v_\beta^3}{3!} + \cdots \right) - \frac{1}{2} \left(\sum_{\beta \in F} \lambda_\beta v_\beta \right)^2 - \frac{1}{3!} \left(\sum_{\beta \in F} \lambda_\beta v_\beta \right)^3 - \cdots \right\| \\ &\leq C_1 \max_{\beta \in F} \|v_\beta\|^2 \end{aligned} \tag{7.8}$$

for $\|v_\beta\|$ small enough for the series to converge. By (7.6) we can estimate

$$\begin{aligned} \|v_\beta\| &\leq \alpha_2 \| \exp(v_\beta) - I\| = \alpha_2 \|p^{-1} q_\beta - I\| \\ &\leq \alpha_2 \|p^{-1}\| \|q_\beta - p\| \leq \alpha_2 \gamma_2 \|q_\beta - p\|. \end{aligned} \tag{7.9}$$

If we combine (7.9) with (7.8), we get the proximity condition that we want to establish. ∎

8. Results. In the previous section we verified that a proximity condition holds between a linear subdivision scheme and the nonlinear analogues we want to study. We combine these results with the inequalities obtained in sections 5 and 6 and are able to state some theorems.

THEOREM 8.1. *Let S be a linear subdivision scheme with a finite mask defined on $\mathbb{Z}^s$, $s \geq 1$. If S is convergent and produces continuous limit functions, then, for initial data p with $D(p)$ sufficiently small, the following hold:*

(i) *If T is the geodesic analogue of S in a smooth surface $\mathcal{M}$, then T is convergent.*

(ii) *If T is the projection analogue of S obtained from a smooth projection mapping $P : \mathbb{R}^n \to \mathcal{M}$, with $P(x) = x$ for all $x \in \mathcal{M}$, then T is convergent.*

(iii) *If $(G, \cdot)$ is a matrix group and T is the G-valued* log-exp *version of S, then T is convergent.*

Proof. By the results of section 7, a proximity condition is satisfied between T and S, and therefore, by Proposition 5.5, the scheme T produces continuous limit functions. ∎

The assumption that $D(p)$ is sufficiently small does not mean that there exists a global constant $\delta > 0$ such that $T^l p$ converges for all initial data with $D(p) < \delta$. Such

a constant in general just exists locally. If the surface (resp., Lie group) is compact, there exists such a global constant. A global δ also exists in the following situations (compare [12] where the univariate situation is discussed):

- for the geodesic analogue if the normal curvatures of the surface are bounded,
- for the projection analogue if the gradient and second derivative of P are bounded,
- in general if p is contained in a compact set.

For certain special cases, for example, the geodesic analogue of the Chaikin algorithm, it is possible to show that the nonlinear analogue is convergent for all initial data [16].

THEOREM 8.2. *Let S be a linear subdivision scheme with a finite mask defined on $\mathbb{Z}^s$, $s \geq 1$, that produces C^1 limit functions. Assume that the first and second derived schemes $S^{[1]}$ and $S^{[2]}$ exist. Then the limit curves of T enjoy C^1-smoothness in the following situations:*

(i) *T is the geodesic analogue of S,*

(ii) *T is the projection analogue of S obtained from a smooth projection mapping $P : \mathbb{R}^n \to \mathcal{M}$, with $P(x) = x$ for all $x \in \mathcal{M}$,*

(iii) *if $(G, \cdot)$ is a matrix group and T is the G-valued* log-exp *version of S.*

Proof. This is a direct consequence of the proximity condition and Proposition 6.5. ∎

Up to this point we have considered only surfaces embedded in the Euclidean space for the geodesic approach. This assumption can be dropped since, by Nash's global embedding theorem [9], any abstract Riemannian manifold can be smoothly embedded in some Euclidean space $\mathbb{R}^n$. Therefore the following theorem is true.

THEOREM 8.3. *Both Theorems* 8.1(i) *and* 8.2(i) *also hold if $\mathcal{M}$ is an abstract Riemannian manifold.*

In a similar manner one can define the log-exp analogue for a general finite dimensional Lie group $(G, \cdot)$ with Lie algebra $\mathfrak{g}$: if a subdivision rule is given by $p_{i,j} + \sum_{r,s} \lambda_{r,s} v_{r,s}$, as in (1.1), we define the associated log-exp rule by

$$p_{i,j} \cdot \exp\left(\sum_{r,s} \lambda_{r,s} v_{r,s}\right), \text{ where } v_{r,s} := \log(p_{i,j}^{-1} \cdot p_{i+r,j+s}) \in \mathfrak{g}.$$

The mapping "exp" is the exponential mapping of G. The log function which is the inverse of the exponential mapping is not defined globally but only for a certain neighborhood of e in G. Since we are dealing only with schemes with a finite mask, we can restrict ourselves to a small neighborhood of the identity element e of G. By Ado's theorem [11] every finite-dimensional Lie group is locally isomorphic to a local matrix group, and therefore, for $D(p)$ small enough, the previous results also apply to general finite-dimensional Lie groups.

THEOREM 8.4. *Theorems* 8.1(iii) *and* 8.2(iii) *also hold if $(G, \cdot)$ is an abstract finite-dimensional Lie group.*

Acknowledgments. The author wants to thank Johannes Wallner for his continuing support and the anonymous referee for valuable suggestions.

REFERENCES

[1] A. S. CAVARETTA, C. A. MICCHELLI, AND W. DAHMEN, *Stationary Subdivision*, American Mathematical Society, Boston, MA, 1991.

[2] D. L. DONOHO AND T. P. YU, *Deslauriers-Dubuc: Ten years after*, in Spline Functions and the Theory of Wavelets, S. Dubuc and G. Deslauriers, eds., CRM Proc. Lecture Notes 18, American Mathematical Society, Providence, RI, 1999, pp. 355–370.

[3] D. L. Donoho, *Interpolating Wavelet Transforms*, Technical report, Stanford University, Stanford, CA, 1992; also available from http://www-stat.stanford.edu/∼donoho/Reports/1992/interpol.pdf.

[4] D. L. Donoho, *Wavelet-type representation of Lie-valued data*, in Proceedings of the IMI Approximation and Computation Meeting, Charleston, SC, Industrial Mathematics Institute, Columbia, SC, 2001.

[5] N. Dyn, D. Levin, and J. A. Gregory, *A 4-point interpolatory subdivision scheme for curve design*, Comput. Aided Geom. Design, 4 (1987), pp. 257–268.

[6] N. Dyn, *Subdivision schemes in CAGD*, in Advances in Numerical Analysis Vol. II, W. A. Light, ed., Oxford University Press, London, 1992, pp. 36–104.

[7] P. Grohs, *Smoothness Analysis of Nonlinear Subdivision Schemes on Regular Grids*, Ph.D. thesis, Vienna University of Technology, Wien, Austria, 2007.

[8] L. Kobbelt, *Interpolatory subdivision on open quadrilateral nets with arbitrary topology*, Comput. Graph. Forum, 15 (1996), pp. 409–420.

[9] J. Nash, *The imbedding problem for Riemannian manifolds*, Ann. of Math., 63 (1956), pp. 20–63.

[10] I. U. Rahman, I. Drori, V. C. Stodden, D. L. Donoho, and P. Schröder, *Multiscale representations for manifold-valued data*, Multiscale Model. Simul., 4 (2005), pp. 1201–1232.

[11] M. Spivak, *A Comprehensive Introduction to Differential Geometry*, Vol. I–V, Publish or Perish, Boston, MA, 1979.

[12] J. Wallner and N. Dyn, *Convergence and C^1 analysis of subdivision schemes on manifolds by proximity*, Comput. Aided Geom. Design, 22 (2005), pp. 593–622.

[13] J. Wallner, E. Nava Yazdani, and P. Grohs, *Smoothness properties of Lie group subdivision schemes*, Multiscale Model. Simul., 6 (2007), pp. 493–505.

[14] J. Wallner, *Smoothness analysis of subdivision schemes by proximity*, Constr. Approx., 24 (2006), pp. 289–318.

[15] G. Xie and T. P.-Y. Yu, *Smoothness equivalence properties of manifold-valued data subdivision schemes based on the projection approach*, SIAM J. Numer. Anal., 45 (2007), pp. 1220–1225.

[16] T. P.-Y. Yu, *Cutting corners on the sphere*, in Wavelets and Splines: Athens 2005, G. Chen and M.-J. Lai, eds., Nashboro Press, Brentwood, TN, 2006, pp. 496–506.

[17] D. X. Zhou, *Stability of refinable functions, multiresolution analysis, and haar bases*, SIAM J. Math. Anal., 27 (1996), pp. 891–904.

SIAM J. NUMER. ANAL.
Vol. 46, No. 4, pp. 2183–2209
© 2008 Society for Industrial and Applied Mathematics

AN ARTIFICIAL BOUNDARY METHOD FOR AMERICAN OPTION PRICING UNDER THE CEV MODEL*

HOI YING WONG† AND JING ZHAO‡

Abstract. The Black–Scholes asset price dynamics is well known to be inadequate for capturing the volatility smile in the financial market. Therefore, the constant elasticity of variance (CEV) model has become a popular alternative to valuing options fitting to the smile. American option pricing, however, is computationally intensive, as there are no analytical formulas available. This paper proposes an artificial boundary method for partial differential equations (PDEs) to compute American option prices and Greeks under the CEV model. The idea is to reduce the infinite computational domain to a finite one by introducing an artificial boundary on which an exact boundary condition is derived. We then develop a Crank–Nicolson scheme to solve the PDE with the artificial boundary condition implemented with a numerical Laplace inversion. With a finite computational domain, the optimal exercise boundary can be determined efficiently. Our numerical examples show that the proposed scheme is accurate, robust with respect to the truncation size, and more efficient than alternative methods for accurate option prices.

Key words. American option, artificial boundary, constant elasticity of variance model, finite difference method

AMS subject classifications. 35A35, 35A40, 65N99

DOI. 10.1137/060671541

1. Introduction.

The American feature in an option contract represents the right for the option holders to exercise their option prematurely. American options are actually the most popular derivatives in the financial market in the sense that their trading volume is even larger than that of their European counterparts. The corresponding valuation has been a fundamental challenge in financial engineering, as it involves a determination of the early exercise strategy within the pricing mechanism.

Under the Black–Scholes asset price dynamics, the valuation of American options has been studied thoroughly. Merton [17] formulates it as a dynamic programming problem. Kim [14] finds that the American option price can be expressed as a European option plus an early exercise premium. Carr, Jarrow, and Myneni [3] formulate the problem as a free boundary value problem for a partial differential equation (PDE). Although they manage to write the solution in a closed form that depends on the early exercise boundary, the option value should be determined numerically, as the early exercise boundary is not known in advance. An alternative way to obtain the option value applies numerical PDE methods, such as the lattice approach, finite difference methods (FDMs), and finite element methods. We refer to [19] for more details on this topic.

In fact, the lattice approach, which can be viewed as an explicit FDM, is usually derived from the probabilistic approach. It is well known that an explicit FDM is conditionally stable; i.e., a condition of the type $\Delta t \sim \Delta x^2$ is required for stability. In contrast, an implicit FDM can achieve unconditional stability. The Crank–Nicolson (CN)

*Received by the editors October 5, 2006; accepted for publication (in revised form) January 15, 2008; published electronically May 2, 2008.
http://www.siam.org/journals/sinum/46-4/67154.html
†Corresponding author. Department of Statistics, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong (hywong@cuhk.edu.hk). The research of this author was supported by HKSAR-RGC grants 403608.
‡Department of Statistics, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong (jzhao@sta.cuhk.edu.hk).

method, which is an average between forward-difference and backward-difference, is unconditionally stable and is a second higher order scheme in time and state [20]. Unfortunately, implicit schemes, including both the implicit FDM and CN method are originally constructed for a PDE with a bounded domain. Therefore, their implementation requires the truncation of the infinite domain into a finite one. Kwok, Wong, and Lau [15] demonstrate that, when the CN scheme is applied to value path-dependent options, the boundary condition can be tricky and the order of accuracy deteriorates near the prescribed boundaries. Kangro and Nicolaides [13] study the far-field boundary conditions for Black–Scholes equations. Han and Wu [12] propose an artificial Neumann boundary condition for PDEs for pricing American options. This method works very well under Black–Scholes dynamics.

In the financial market, the volatility implied by option prices shows a smile shape across strike prices. This means that the constant volatility assumption of the Black–Scholes model is incorrect. To capture the volatility smile, one popular alternative is the constant elasticity of variance (CEV) model suggested by Cox [5]. The implications of the CEV model are studied by Becker [1]. It has been shown that this model can fit the volatility skew of equity options. Emanuel and MacBeth [9] derive a closed-form solution for European options under positive elasticity. Davydov and Linetsky [6] study the pricing and hedging of barrier and lookback options under CEV. Detemple and Tian [7] derive a recursive integral equation for the exercise boundary under CEV, which can be used as a starting point for numerical methods of American option valuation. Although its computational accuracy and efficiency are not reported in the paper, the recursive integral equation is computationally intensive. In practice, the FDM is the most common numerical tool for pricing options because of its efficiency and robustness. Our interest is in developing an FDM for option pricing under CEV. Lu and Hsu [16] obtain a binomial tree approximation for European options under CEV and speculate that it is also useful for American option pricing. However, its efficiency and stability are not reported.

This study generalizes the artificial boundary method of Han and Wu [12] to the CEV model. This extension is nontrivial. The Black–Scholes equation can be transformed into the standard heat equation in which the artificial boundary is classic [11], [21]. For the CEV model, the PDE cannot be transformed into a linear equation with constant coefficients, and the corresponding artificial boundary condition has not been considered before. We derive the artificial boundary condition (ABC) by using the Laplace transform and give several demonstrations of the application. An efficient CN scheme is then constructed by approximating the artificial boundary condition with a numerical Laplace inversion. Numerical examples show that the proposed artificial boundary approach is efficient, accurate, and robust with respect to the truncation size. In addition, it outperforms the binomial method and the CN scheme using the far-field boundary condition. We also prove that the artificial boundary condition is exact and the proposed CN scheme is unconditionally stable. This lays the theoretical grounds for option pricing under CEV with an FDM.

The remainder of the paper is organized as follows. Section 2 presents some basic results in option pricing under CEV. Section 3 formulates the proposed artificial boundary approach. Some theoretical results are obtained. Numerical examples are demonstrated in section 4. Section 5 concludes.

2. Option pricing under CEV. The CEV model assumes that, under the risk-neutral probability measure, the underlying asset price evolves according to the stochastic differential equation:

$$\mathrm{d}S = (r-q)S\mathrm{d}t + \sigma S^{\alpha}\,\mathrm{d}Z, \tag{2.1}$$

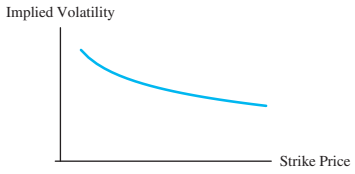


FIG. 2.1. *Volatility smile for equities.*

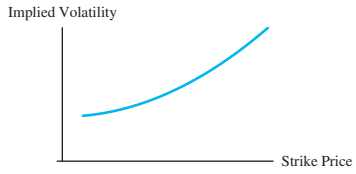


FIG. 2.2. *Volatility smile for options on futures.*

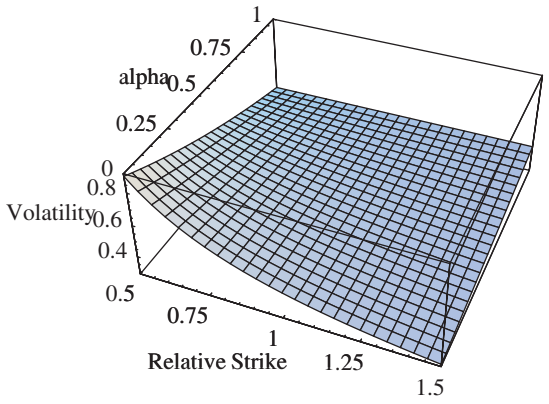


FIG. 2.3. *The CEV skew for varying values of* α.

where r is the risk-free rate, q is the dividend yield, Z is a Wiener process, σ is a volatility parameter, and α is a constant known as the elasticity factor.

The CEV model nests several asset price processes as special cases. For instance, it incorporates a local volatility given by $\sigma(S) = \sigma S^{\alpha-1}$. When $\alpha = 1$, the process (2.1) becomes the geometric Brownian motion and is reduced to the Black–Scholes model. We have the Ornstein–Uhlenbeck process by setting α to zero. If $\alpha = \frac{1}{2}$, then the square-root model is obtained.

There are interesting interpretations of the CEV model. When $\alpha < 1$, the local volatility increases as the stock price decreases. This creates a probability distribution similar to the empirical distribution of equities that has a heavy left tail and a less heavy right tail. Since the financial crisis of 1987, the volatility smile used by traders to price equity options (both on individual stocks and on stock indices) has had the general form shown in Figure 2.1. When $\alpha > 1$, the local volatility increases as the stock price increases. This creates a probability distribution with a heavy right tail and a less heavy left tail, which corresponds to a volatility smile in which the implied volatility is an increasing function of the strike price. This volatility smile often appears for options on futures, as shown in Figure 2.2.

The parameters of the CEV model can be estimated by calibration, that is, to fit the model prices of vanilla options as closely as possible to market prices by minimizing the sum of the squared differences between them.

Figure 2.3 shows the volatility skew under the CEV model by varying the value of α. The plot uses a current stock price of 100, a time to maturity of one year, an interest rate of 5%, and an at-the-money volatility of 45%. When α gets close to zero, the CEV model generates a steeper skew for the implied volatility smile. Conversely, the skew becomes flatter as α tends to one and ultimately becomes as flat as the Black–Scholes model.

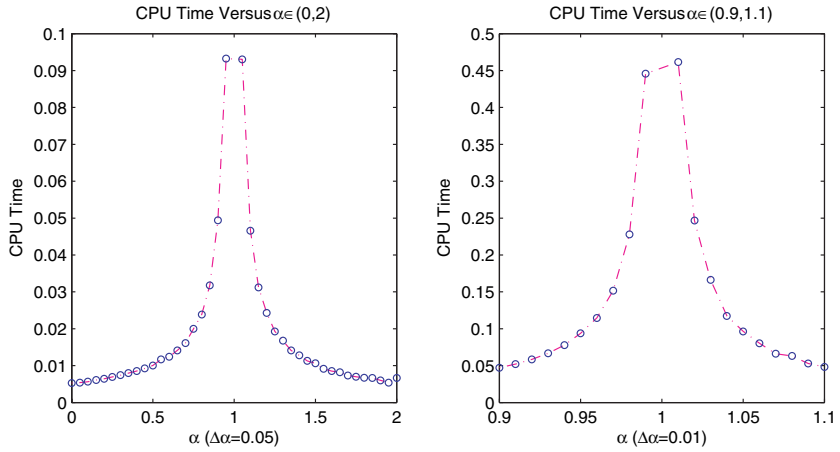


FIG. 2.4. *Computational time for different* α.

2.1. The valuation of European options. Closed-form pricing formulas of European call and put options are available for the CEV model. Specifically, Cox [5] shows that

$$C_E = S_0 e^{-qT}[1-\chi^2(b, c+2, d)] - Ke^{-rT}\chi^2(d, c, b), \tag{2.2}$$
$$P_E = Ke^{-rT}[1-\chi^2(d, c, b)] - S_0 e^{-qT}\chi^2(b, c+2, d),$$

when $\alpha < 1$, and Emanuel and MacBeth [9] derive

$$C_E = S_0 e^{-qT}[1-\chi^2(d, -c, b)] - Ke^{-rT}\chi^2(b, 2-c, d), \tag{2.3}$$
$$P_E = Ke^{-rT}[1-\chi^2(b, 2-c, d)] - S_0 e^{-qT}\chi^2(d, -c, b),$$

when $\alpha > 1$, where

$$b = \frac{[Ke^{-(r-q)T}]^{2(1-\alpha)}}{(1-\alpha)^2 w}, \quad c = \frac{1}{1-\alpha}, \tag{2.4}$$
$$d = \frac{S^{2(1-\alpha)}}{(1-\alpha)^2 w}, \quad w = \frac{\sigma^2}{2(r-q)(\alpha-1)}[e^{2(r-q)(\alpha-1)T}-1],$$

and $\chi^2(z, k, \ell)$ is the cumulative distribution function (cdf) of a noncentral chi-squared random variable with noncentrality parameter ℓ and k degrees of freedom.

These formulas feature the noncentral chi-squared cdf, which is available in some computational software. For instance, Matlab provides the NCX2CDF function to compute it as a series expansion in terms of chi-squared cdfs with Poisson weights. Schroder [18], however, points out that the evaluation of the infinite sum can be computationally expensive when both z and ℓ are large (i.e., α is close to 1, volatility is low, or T is small in the CEV model).

Figure 2.4 shows the computational time of the closed-form European call price against α, where the current stock price is 100, strike price 110, time to expiry 6 months, $r = 5\%$, $q = 0$, and at-the-money volatility 20%. It can be seen that the computational time increases significantly when α gets close to 1.

The numerical PDE method offers an efficient alternative approach to evaluating option prices under CEV. We compare the computational accuracy of the closed-form

TABLE 2.1
Values of European put options for different α.

α	0.999	0.9999	0.99999	BS price	1.00001	1.0001	1.001
Closed-form	10.199	10.201	10.261	10.191	10.275	10.200	10.199
CPU(s)	3.9586	35.9467	367.35		397.62	35.3435	3.745
FDM	10.190	10.191	10.191	10.191	10.191	10.191	10.191
CPU(s)	1.1444	1.1421	1.1453		1.1425	1.1430	1.149

solution with that of the FDM when α gets close to 1. We consider α near 1 because the Black–Scholes pricing formula provides a base to compare the accuracy. Table 2.1 indicates that the closed-form formula is unstable and inefficient for α close to 1. Consequently, an efficient and accurate numerical method for option pricing under CEV is invaluable. Table 2.1 also shows that the traditional CN FDM with 500 grid points converges to the Black–Scholes (BS) price more efficiently. The computational times are similar for all values of α. Moreover, the FDM generates option values on the entire grid of S value. Thus, interpolation can be employed to obtain option prices for all initial stock prices.

The aforementioned CN scheme is based on the PDE for CEV option pricing with a far-field boundary condition. Following the arguments of Carr, Jarrow, and Myneni [3], the price of a European option is the unique solution of the following problem under general diffusion processes:

$$\begin{cases} \frac{\partial V}{\partial t} - rV + \frac{1}{2}\sigma^2(S)S^2\frac{\partial^2 V}{\partial S^2} + (r-q)S\frac{\partial V}{\partial S} = 0, \quad 0 \le S, \quad t \le T, \\ V(S,T) = V_T(S), \quad 0 \le S. \end{cases} \tag{2.5}$$

The payoff function $V_T(S)$ is determined by the contractual agreement of the option. For example, the payoffs of vanilla calls and puts are given by

$$V(S,T) = \max(\theta(S-K),0), \quad 0 \le S, \tag{2.6}$$

where $\theta = 1$ for a call option and -1 for a put option.

To make the option payoff an initial condition, we reverse the time by setting $t^* = T - t$ and then substituting it into (2.5):

$$\begin{cases} \frac{\partial V}{\partial t^*} + rV - \frac{1}{2}\sigma^2(S)S^2\frac{\partial^2 V}{\partial S^2} - (r-q)S\frac{\partial V}{\partial S} = 0, \quad 0 \le S, \quad t^* \le T, \\ V(S,0) = V_T(S), \quad 0 \le S. \end{cases} \tag{2.7}$$

FDMs are simple and robust approaches that have been widely used to solve PDEs numerically. As (2.7) is defined for $S \in [0,\infty)$, the option should be valued over an unbounded domain, which is computationally infeasible for implicit schemes. A traditional approach works on a truncated domain $[0, S_{max}]$ for some suitably large S_{max}. The left and right boundary conditions can be approximated through the option's properties as follows:

$$P_E(0,t^*) = Ke^{-rt^*}, \quad P_E(S_{max},t^*) \approx 0, \tag{2.8}$$

where $P_E(S,t^*)$ represents a European put option price. By the put-call parity, equivalent boundary conditions for a European call $C_E(S,t^*)$ can be obtained as follows:

$$C_E(0,t^*) = 0, \quad C_E(S_{max},t^*) \approx S_{max}e^{-qt^*} - Ke^{-rt^*}. \tag{2.9}$$

Note that the origin is an absorbing boundary for S to guarantee the absence of arbitrage. Implicit schemes can then be implemented over the bounded domain with the imposed approximate boundary conditions.

2.2. The valuation of American options. As there is no closed-form solution for finite maturity American options under CEV, the valuation should rely on numerical methods. Let $P_A(S,t)$ be the American put option price and $S_p^*(t)$ be the optimal exercise price. Then the entire domain is divided into two regions separated by a free boundary. In the exercise region $[0, S_p^*(t)] \times [0,T]$, we have

$$\begin{cases} \frac{\partial P_A}{\partial t} - rP_A + \frac{1}{2}\sigma^2 S^{2\alpha} \frac{\partial^2 P_A}{\partial S^2} + (r-q)S\frac{\partial P_A}{\partial S} < 0, \\ P_A(S,t) = K - S = \max(K-S,0). \end{cases} \tag{2.10}$$

In the continuation region $(S_p^*(t), \infty) \times [0,T]$, we have

$$\begin{cases} \frac{\partial P_A}{\partial t} - rP_A + \frac{1}{2}\sigma^2 S^{2\alpha} \frac{\partial^2 P_A}{\partial S^2} + (r-q)S\frac{\partial P_A}{\partial S} = 0, \\ P_A(S,t) > \max(K-S,0). \end{cases} \tag{2.11}$$

On the free boundary, we should have

$$P_A(S_p^*(t),t) = K - S_p^*(t)$$

and

$$\frac{\partial P_A(S_p^*(t),t)}{\partial S}(S_p^*(t),t) = -1. \tag{2.12}$$

Equation (2.12) is known as the high-contact condition for American put options. This condition must hold for all continuous asset price processes, including the CEV model [3]. The inequality (2.10) describes a situation in which the value of the option if exercised exceeds its value if held. Thus, it is optimal to exercise the option prematurely. If $P_A(S,t) > \max(K-S,0)$, then we should hold the option for an obvious reason, and hence the option value satisfies the governing equation. When $P_A(S,t) > \max(K-S,0)$, $P_A(S,t-\Delta t) > P_A(S,t)$ for any positive Δt.

From the facts above, it is easy to see that the following inequality holds: $S_p^*(t) > S_p^*(t-\Delta t), \Delta t > 0$, which implies that $S_p^*(t)$ is an increasing function of t. It is known that $S_p^\infty \le S_p^*(t) \le S_p^0, 0 \le t \le T$, where S_p^∞ is the optimal exercise price of the perpetual American put option and $S_p^0 = S_p^*(T) = \min(\frac{rK}{q}, K)$. Moreover,

$$\lim_{S\to\infty} P_A(S,t) = 0. \tag{2.13}$$

This PDE problem is defined on a semi-infinite domain. To implement an implicit FDM, the domain could be truncated by utilizing the far-field boundary conditions:

$$P_A(0,t^*) = K, \quad P_A(S_{max},t^*) \approx 0. \tag{2.14}$$

The numerical valuation of American options is similar that of European options, except that early exercise opportunities are required to be searched for at each time point. If the optimal exercise strategy depends only on the current value of the underlying process and time, then the value of the option can be determined by using the Bellman principle of dynamic programming:

$$P_A(S) = \max\{f(S), PV_t[P_A(S+\mathrm{d}S, t+\mathrm{d}t)]\}, \tag{2.15}$$

where $PV_t[\cdot]$ is the present value at time t and $f(S)$ is the intrinsic value of the option.

This principle is easily implemented with the lattice approach or an FDM because it simply means looking for early exercise opportunities at each node at each time point. If it is optimal to exercise early at a particular node, then the option value is replaced with the intrinsic value at that node. However, this can be time-consuming for a large domain.

According to the put-call symmetry for American options, if the local volatility $\sigma(S,t)$ depends solely on S, then the following relations hold [22]:

$$C_A(S,t;r,q,\sigma(S)) = P_A(E^2/S,t;q,r,\sigma(S))S/E, \tag{2.16}$$
$$P_A(S,t;r,q,\sigma(S)) = C_A(E^2/S,t;q,r,\sigma(S))S/E.$$

Thus, it suffices to consider the valuation of either the American put or call option. In this paper, we focus on the put option.

2.3. How "far" is enough? To deal with an unbounded domain, we usually impose boundary conditions at a place that is far away from the region of interest so that, even if the imposed boundary conditions are imperfect, the solution will not be affected materially. This approach is referred to as the far-field boundary condition [13].

How far is sufficient? To produce an accurate result, a large domain is preferred, but we have to deal with many points for an evenly spaced grid. Consequently, the computational time is increased. Although a small domain results in fast computation, the accuracy can be greatly affected by the error associated with the imposed boundary conditions. In addition, the effect of the far-field boundary condition for the CEV model is not known, and it lacks theoretical support.

There is always a trade-off between accuracy and efficiency with the far-field approach. The more accurately the boundary conditions conform to the characteristics of the desired solution, the closer we can place the boundary to the domain of interest. Thus, an artificial boundary method is a remedy for this implementation problem. Although not demonstrated in this paper, an unevenly spaced grid (fine spacing near the strike price and coarse spacing near the boundaries) can offer considerable computational advantages on top of the proposed artificial boundary method.

3. The proposed approach. In this section, we derive a class of exact artificial boundary conditions for the CEV model on an unbounded domain. An exact artificial boundary condition represents a boundary condition by which we obtain an initial boundary value (IBV) problem on a bounded domain that shares the same solution with the original initial value problem. Therefore, we perform FDM on the IBV problem on the introduced bounded domain.

3.1. Standardized form of the CEV model. Under the Black–Scholes model, the governing equation can be transformed into a heat equation so that classical numerical methods can be applied. Here we transform the PDE under CEV into a standard form. We consider the standard time-reversing transformation $t^* = T - t$ and (2.5). By setting $V = e^{-rt^*}V_1$, the equation is reduced to

$$\frac{\sigma^2(S)}{2}S^2\frac{\partial^2 V_1}{\partial S^2} + (r-q)S\frac{\partial V_1}{\partial S} = \frac{\partial V_1}{\partial t^*}. \tag{3.1}$$

To eliminate the drift term, we introduce a new variable $S' = Se^{(r-q)t^*}$ and transform the PDE into

$$\frac{\sigma^2(S)}{2}S'^2\frac{\partial^2 V_1}{\partial S'^2} = \frac{\partial V_1}{\partial t^*}. \tag{3.2}$$

After recognizing $\sigma(S) = \sigma S^{\alpha-1}$ in the CEV model, the governing equation (3.2) is given by

$$\frac{\sigma^2 S'^{2\alpha}}{2}\frac{\partial^2 V_1}{\partial S'^2} = \frac{\partial V_1}{\partial t^*}e^{(r-q)t^*(2\alpha-2)}. \tag{3.3}$$

We reparameterize the elasticity by letting $2\alpha = 2 - \frac{1}{\nu}$. The PDE now becomes

$$\frac{\sigma^2 S'^{(2-\frac{1}{\nu})}}{2}\frac{\partial^2 V_1}{\partial S'^2} = \frac{\partial V_1}{\partial t^*}e^{(r-q)t^*(-\frac{1}{\nu})}. \tag{3.4}$$

To simplify the left-hand side of (3.4), we let $X = S'^{\frac{1}{\nu}}$ and obtain

$$\sigma^2 X\frac{\partial^2 V_1}{\partial X^2} + \sigma^2(1-\nu)\frac{\partial V_1}{\partial X} = 2\nu^2 e^{(r-q)t^*(-\frac{1}{\nu})}\frac{\partial V_1}{\partial t^*}. \tag{3.5}$$

The right-hand side of (3.5) can be further simplified by introducing a new time variable τ for

$$d\tau = \frac{1}{2\nu^2}e^{(r-q)t^*(\frac{1}{\nu})}dt^* \Rightarrow \tau = \frac{1}{2\nu(r-q)}(e^{\frac{(r-q)(T-t)}{\nu}} - 1).$$

Consequently, we arrive at the standardized CEV equation

$$\frac{\partial V_1}{\partial \tau} = \sigma^2 X\frac{\partial^2 V_1}{\partial X^2} + \sigma^2(1-\nu)\frac{\partial V_1}{\partial X}. \tag{3.6}$$

This governing equation can be linked to a simple parabolic equation with a nonconstant drift by letting $Y = \frac{2}{\sigma}X^{\frac{1}{2}}$:

$$\frac{\partial V_1}{\partial \tau} = \frac{\partial^2 V_1}{\partial Y^2} + \frac{1-2\nu}{Y}\frac{\partial V_1}{\partial Y}. \tag{3.7}$$

The overall changes are summarized as follows:

$$\begin{cases} \tau = \frac{1}{2\nu(r-q)}(e^{\frac{(r-q)(T-t)}{\nu}} - 1), \\ \quad \text{when } t = 0, \ \ \tau = \frac{1}{2\nu(r-q)}(e^{\frac{(r-q)T}{\nu}} - 1) = T^*, \\ \quad \text{when } t = T, \ \ \tau = 0, \\ 2\alpha = 2 - \frac{1}{\nu}, \\ Y = \frac{2}{\sigma}(Se^{(r-q)(T-t)})^{\frac{1}{2\nu}}, \\ \quad \text{when } t = 0, \ \ Y = \frac{2}{\sigma}(Se^{(r-q)T})^{\frac{1}{2\nu}}, \\ \quad \text{when } t = T, \ \ Y = \frac{2}{\sigma}S^{\frac{1}{2\nu}}, \\ V_1 = e^{r(T-t)}V, \\ \quad \text{when } t = 0, \ \ V_1 = e^{rT}V, \\ \quad \text{when } t = T, \ \ V_1 = V. \end{cases} \tag{3.8}$$

The American option pricing problem under the CEV model can be formulated as

$$\begin{cases} \frac{\partial V_1}{\partial \tau} = \frac{\partial^2 V_1}{\partial Y^2} + \frac{1-2\nu}{Y}\frac{\partial V_1}{\partial Y}, \ \ Y > 0, \ 0 \le \tau, \\ V_1|_{\tau=0} = \max(K - (\frac{\sigma Y}{2})^{2\nu}, 0), \ \ Y_p^*(\tau) \le Y, \\ V_1\,|_{Y_p^*(\tau)} = Ke^{r(T-t)} - \left(\frac{\sigma Y_p^*(\tau)}{2}\right)^{2\nu}e^{q(T-t)}, \\ \frac{\partial V_1}{\partial Y}\big|_{Y=Y_p^*(\tau)} = -\nu\sigma^{2\nu}\left(\frac{Y_p^*(\tau)}{2}\right)^{2\nu-1}e^{q(T-t)}, \end{cases} \tag{3.9}$$

where $Y_p^*(\tau)$ is the optimal exercise value of Y at time τ and satisfies

$$Y_p^*(\tau) = \frac{2}{\sigma}(S_p^*(t)e^{(r-q)(T-t)})^{\frac{1}{2\nu}}. \tag{3.10}$$

The last equation of (3.9) comes from the high-contact condition (2.12) of the original problem. Its derivation is the direct consequence of the transformation (3.8) and the fact that $\frac{\partial}{\partial S} = \frac{\partial Y}{\partial S}\frac{\partial}{\partial Y}$. In the continuation region, we have

$$V_1(Y,\tau) > \max\left(Ke^{r(T-t)} - \left(\frac{\sigma Y}{2}\right)^{2\nu} e^{q(T-t)}, 0\right). \tag{3.11}$$

3.2. Exact artificial boundary conditions. Armed with the standardized CEV equation, this is the right moment to derive the exact artificial boundary condition for American put options. We consider the artificial boundary on the asset value to be larger than the strike price K as the optimal exercise price is less than K. When $\alpha = 1$, the CEV model reduces to the Black–Scholes model, and the corresponding artificial boundary has been studied by Han and Wu [12]. Thus, we focus on the remaining two cases: $\alpha > 1$ and $\alpha < 1$.

3.2.1. Case 1: $\frac{1}{\nu} < 0$ ($\alpha > 1$). This case is simple. From (3.8), we recognize that $Y = \frac{2}{\sigma}(Se^{(r-q)(T-t)})^{\frac{1}{2\nu}}$ is a decreasing function of S. When $S \to \infty$, $Y \to 0$. By the exact boundary condition of (2.13), we deduce that

$$\lim_{Y\to 0} V_1(Y,\tau) = 0. \tag{3.12}$$

Therefore, the original PDE defined on the unbounded domain $(S_p^*(t), \infty) \times [0, T]$ is converted into a new PDE defined on the bounded domain $(0, Y_p^*(\tau)) \times [0, T^*]$ with the exact boundary condition (3.12) and the initial condition

$$V_1(Y,0) = \max\left(K - \left(\frac{\sigma Y}{2}\right)^{2\nu}, 0\right). \tag{3.13}$$

Moreover, the value of an American put option should respect the constraint:

$$V_1(Y,\tau) \geq \max\left(Ke^{r(T-t)} - \left(\frac{\sigma Y}{2}\right)^{2\nu} e^{q(T-t)}, 0\right). \tag{3.14}$$

Thus, the CN scheme can be applied to the PDE (3.9) defined on this bounded domain.

3.2.2. Case 2: $\frac{1}{\nu} > 0$ ($\alpha < 1$). The challenge goes to the case in which $\alpha < 1$. From (3.8), $Y = \frac{2}{\sigma}(Se^{(r-q)(T-t)})^{\frac{1}{2\nu}}$ is an increasing function of S under the present situation. We recall that the governing equation

$$\frac{\partial V_1}{\partial \tau} = \frac{\partial^2 V_1}{\partial Y^2} + \frac{1-2\nu}{Y}\frac{\partial V_1}{\partial Y} \tag{3.15}$$

is defined on an unbounded domain $\Omega(\tau)$ as

$$\Omega(\tau) = \{(Y,\tau) \in R^2 \mid Y_p^*(\tau) \leq Y < +\infty, 0 \leq \tau \leq T^*\}. \tag{3.16}$$

We now introduce an artificial boundary Γ_a:

$$\Gamma_a = \{(Y,\tau) \mid Y = a, 0 \le \tau \le T^*\}, \tag{3.17}$$

where a is a constant prespecified by the user. Exact boundary conditions are derived to ensure that the solution on the computational (truncated) domain coincides with the exact solution of the original problem.

The unbounded domain $\Omega(\tau)$ is divided by Γ_a into the bounded time-dependent interior domain

$$\Omega_{\text{int}}(\tau) = \{(Y,\tau) \in R^2 \mid Y_p^*(\tau) \le Y < a, 0 \le \tau \le T^*\} \tag{3.18}$$

and the unbounded time-independent exterior domain

$$\Omega_{\text{ext}} = \{(Y,\tau) \in R^2 \mid Y > a, 0 \le \tau \le T^*\}. \tag{3.19}$$

We postulate that the initial condition is compactly supported in the interior (bounded) domain $\Omega_{\text{int}}(\tau)$ so that the initial condition is zero in the exterior (unbounded) domain. Hence, the artificial boundary should be above the free boundary.

For the derivation of the exact artificial boundary condition at Γ_a, we consider the exterior problem:

$$\begin{cases} \frac{\partial V_1}{\partial \tau} = \frac{\partial^2 V_1}{\partial Y^2} + \frac{1-2\nu}{Y}\frac{\partial V_1}{\partial Y}, \quad (Y,\tau) \in \Omega_{\text{ext}}, \\ V_1 \mid_{\tau=0} = 0, \quad Y > a, \\ V_1 \mid_{Y=a} = V_1(a,\tau), \quad 0 \le \tau \le T^*, V_1(a,0) = 0, \\ V_1 \to 0 \ \text{ when } \ Y \to +\infty, 0 \le \tau \le T^*. \end{cases} \tag{3.20}$$

As $V_1(a,\tau)$ is an unknown function, the problem of (3.20) is not a complete-posed problem and cannot be solved independently. Certainly, it becomes a well-posed problem once we determine the value of $V_1(Y,\tau)$ on the artificial boundary Γ_a. For the moment, we consider $V_1(a,\tau) = V_0$ and solve the exterior problem (3.20) by using the Laplace transform, as follows:

$$\hat{V_1}(Y,s) = \int_0^\infty V_1(Y,\tau)e^{-s\tau}d\tau := L_s(V_1(Y,\tau)), \tag{3.21}$$

where $s > 0$ is the parameter of the Laplace transform. Then the subsidiary equation for (3.20) with a zero initial condition is

$$\frac{d^2\hat{V_1}}{dY^2} + \frac{1-2\nu}{Y}\frac{d\hat{V_1}}{dY} = s\hat{V_1}. \tag{3.22}$$

The corresponding boundary conditions are

$$\lim_{Y\to\infty} \hat{V_1} = 0, \qquad \hat{V_1}(a) = \frac{V_0}{s}. \tag{3.23}$$

Let $\hat{V_1} = Y^\nu \overline{V}_1$ and $z = \sqrt{s}Y$. Then the ordinary differential equation (3.22) is transformed into the modified Bessel differential equation:

$$\frac{d^2\overline{V}_1}{dz^2} + \frac{1}{z}\frac{d\overline{V}_1}{dz} - \left(1 + \frac{\nu^2}{z^2}\right)\overline{V}_1 = 0. \tag{3.24}$$

The solution of (3.24) is classic:

$$\overline{V}_1 = AI_\nu(z) + BK_\nu(z), \tag{3.25}$$

where $I_\nu(z)$ and $K_\nu(z)$ are, respectively, the modified Bessel functions of the first and the second kind. The constants A and B are chosen such that $\overline{V}_1$ satisfies the boundary conditions (3.23). A simple calculation shows that

$$A = 0, \qquad B = \frac{V_0}{sa^\nu K_\nu(\sqrt{s}a)}. \tag{3.26}$$

Hence, we obtain

$$\hat{V}_1 = \left(\frac{Y}{a}\right)^\nu \frac{V_0 K_\nu(\sqrt{s}Y)}{sK_\nu(\sqrt{s}a)}. \tag{3.27}$$

By inverting the Laplace transform, the solution of (3.20) defined on the exterior domain with a zero initial condition and a constant boundary condition over Γ_a is found:

$$V_1 = L_s^{-1}\left(\left(\frac{Y}{a}\right)^\nu \frac{V_0 K_\nu(\sqrt{s}Y)}{sK_\nu(\sqrt{s}a)}\right). \tag{3.28}$$

Thus, we identify the function as follows:

$$G(Y,\tau) = L_s^{-1}\left(\left(\frac{Y}{a}\right)^\nu \frac{K_\nu(\sqrt{s}Y)}{sK_\nu(\sqrt{s}a)}\right)\Bigg|_{t=\tau}. \tag{3.29}$$

By using Duhamel's theorem [4], we conclude that

$$\begin{aligned} V_1(Y,\tau) &= \int_0^\tau V_1(a,\lambda)\frac{\partial}{\partial\tau}G(Y,\tau-\lambda)d\lambda = -\int_0^\tau V_1(a,\lambda)\frac{\partial}{\partial\lambda}G(Y,\tau-\lambda)d\lambda \\ &= -V_1(a,\lambda)G(Y,\tau-\lambda)\,|_{\lambda=0}^{\lambda=\tau} + \int_0^\tau \frac{\partial V_1(a,\lambda)}{\partial\lambda}G(Y,\tau-\lambda)d\lambda. \end{aligned} \tag{3.30}$$

As $V_1(a,0) = 0$ and $G(Y,0) = 0$, (3.30) reads

$$V_1(Y,\tau) = \int_0^\tau \frac{\partial V_1(a,\lambda)}{\partial\lambda}G(Y,\tau-\lambda)d\lambda. \tag{3.31}$$

An exact Neumann boundary condition can be obtained from (3.31). Consider

$$\frac{\partial V_1}{\partial Y} = \int_0^\tau \frac{\partial V_1(a,\lambda)}{\partial\lambda}\frac{\partial G(Y,\tau-\lambda)}{\partial Y}d\lambda, \tag{3.32}$$

where

$$\frac{\partial G(Y,\tau-\lambda)}{\partial Y} = L_s^{-1}\left(\frac{\partial}{\partial Y}\left[\left(\frac{Y}{a}\right)^\nu \frac{K_\nu(\sqrt{s}Y)}{sK_\nu(\sqrt{s}a)}\right]\right)\Bigg|_{t=\tau-\lambda}.$$

Due to the fact that

$$zK_\nu'(z) + \nu K_\nu(z) = -zK_{\nu-1}(z), \tag{3.33}$$

we define the integral kernel as

$$\frac{\partial G(Y,\tau-\lambda)}{\partial Y}\bigg|_{Y=a} = L_s^{-1}\left(\frac{-K_{\nu-1}(\sqrt{s}a)}{\sqrt{s}K_\nu(\sqrt{s}a)}\right)\bigg|_{t=\tau-\lambda} = H_\nu(\tau-\lambda). \tag{3.34}$$

Then we are able to write

$$\frac{\partial V_1(Y,\tau)}{\partial Y}\bigg|_{Y=a} = \int_0^\tau \frac{\partial V_1(a,\lambda)}{\partial \lambda} H_\nu(\tau-\lambda)d\lambda. \tag{3.35}$$

The above computation shows the following proposition.

PROPOSITION 3.1. *For $\alpha < 1$, the solution of the original problem* (3.9) *over an unbounded domain satisfies the following PDE over a bounded domain:*

$$\begin{cases} \frac{\partial V_1}{\partial \tau} = \frac{\partial^2 V_1}{\partial Y^2} + \frac{1-2\nu}{Y}\frac{\partial V_1}{\partial Y}, \quad Y_p^*(\tau) \le Y \le a, \quad 0 \le \tau \le T^*, \\ V_1\,|_{\tau=0} = \max(K - (\frac{\sigma Y}{2})^{2\nu}, 0), \quad Y_p^*(\tau) \le Y \le a, \\ V_1\,|_{Y_p^*(\tau)} = Ke^{r(T-t)} - \left(\frac{\sigma Y_p^*(\tau)}{2}\right)^{2\nu} e^{q(T-t)}, \\ \frac{\partial V_1}{\partial Y}|_{Y=a} = \int_0^\tau \frac{\partial V_1(a,\lambda)}{\partial \lambda} H_\nu(\tau-\lambda)d\lambda. \end{cases} \tag{3.36}$$

It remains for us to show that the solution to (3.36) is unique so that we are allowed to use the numerical solution of (3.36) to approximate that of (3.9).

PROPOSITION 3.2. *The solution of* (3.36) *is unique.*

Proof. Suppose that $V_{11}(Y,\tau)$ and $V_{12}(Y,\tau)$ are two solutions to problem (3.36). We define their difference to be $\varepsilon(Y,\tau) = V_{11}(Y,\tau) - V_{12}(Y,\tau)$. In Ω_{int}, $\varepsilon(Y,\tau)$ satisfies

$$\begin{cases} \frac{\partial \varepsilon}{\partial \tau} = \frac{\partial^2 \varepsilon}{\partial Y^2} + \frac{1-2\nu}{Y}\frac{\partial \varepsilon}{\partial Y}, \quad Y_p^*(\tau) \le Y \le a, \quad 0 \le \tau \le T^*, \\ \varepsilon\,|_{\tau=0} = 0, \quad Y_p^*(\tau) \le Y \le a, \\ \varepsilon\,|_{Y_p^*(\tau)} = 0, \quad 0 \le \tau \le T^*, \\ \frac{\partial \varepsilon}{\partial Y}\,|_{Y=a} = \int_0^\tau \frac{\partial \varepsilon(a,\lambda)}{\partial \lambda} H_\nu(\tau-\lambda)d\lambda. \end{cases} \tag{3.37}$$

By multiplying $Y\varepsilon$ by both sides of the PDE in (3.37) and performing integrations over Ω_{int}, we obtain

$$\begin{aligned} &\int_0^{T^*}\int_{Y_p^*(\tau)}^a Y\left(\frac{\partial \varepsilon}{\partial Y}\right)^2 dY d\tau + \frac{1}{2}\int_{Y_p^*(\tau)}^a Y\varepsilon^2\bigg|_{t=T^*} dY \\ &- \int_0^{T^*} \varepsilon Y \frac{\partial \varepsilon}{\partial Y}\bigg|_{Y=a} d\tau + 2\nu\int_0^{T^*} \frac{\varepsilon^2}{2}\bigg|_{Y=a} d\tau = 0. \end{aligned} \tag{3.38}$$

We then consider the following problem on the unbounded domain Ω_{ext}:

$$\begin{cases} \frac{\partial \varepsilon^*}{\partial \tau} = \frac{\partial^2 \varepsilon^*}{\partial Y^2} + \frac{1-2\nu}{Y}\frac{\partial \varepsilon^*}{\partial Y}, \quad a \le Y \le \infty, \quad 0 \le \tau \le T^*, \\ \varepsilon^*\,|_{\tau=0} = 0, \quad a \le Y \le \infty, \\ \varepsilon^*\,|_{Y=a} = \varepsilon\,|_{Y=a}, \quad 0 \le \tau \le T^*, \\ \varepsilon^* \to 0 \ \text{ when } \ Y \to \infty. \end{cases} \tag{3.39}$$

Given $\varepsilon(Y,\tau)$, problem (3.39) has a unique solution $\varepsilon^*(Y,\tau)$. Moreover,

$$\frac{\partial \varepsilon^*}{\partial Y}\bigg|_{Y=a} = \int_0^\tau \frac{\partial \varepsilon(a,\lambda)}{\partial \lambda} H_\nu(\tau-\lambda)d\lambda = \frac{\partial \varepsilon}{\partial Y}\bigg|_{Y=a}. \tag{3.40}$$

By multiplying $Y\varepsilon^*$ by both sides of (3.39) and integrating over Ω_{ext}, we obtain

$$\begin{aligned}(3.41)\qquad &\int_0^{T^*}\int_a^{\infty} Y\left(\frac{\partial \varepsilon^*}{\partial Y}\right)^2 dY d\tau + \frac{1}{2}\int_a^{\infty} Y\varepsilon^{*2}\Big|_{t=T^*} dY \\ &= -\int_0^{T^*} \varepsilon^* Y \frac{\partial \varepsilon^*}{\partial Y}\Big|_{Y=a} d\tau + 2\nu \int_0^{T^*} \frac{\varepsilon^{*2}}{2}\Big|_{Y=a} d\tau \geq 0.\end{aligned}$$

According to (3.38),

$$\begin{aligned}(3.42)\qquad &-\int_0^{T^*} \varepsilon^* Y \frac{\partial \varepsilon^*}{\partial Y}\Big|_{Y=a} d\tau + 2\nu \int_0^{T^*} \frac{\varepsilon^{*2}}{2}\Big|_{Y=a} d\tau \\ &= -\int_0^{T^*} \varepsilon Y \frac{\partial \varepsilon}{\partial Y}\Big|_{Y=a} d\tau + 2\nu \int_0^{T^*} \frac{\varepsilon^{2}}{2}\Big|_{Y=a} d\tau \leq 0.\end{aligned}$$

By combining (3.41)–(3.42) with (3.39), we end up with $\varepsilon(Y,\tau)\equiv 0$. The proof is then completed. □

PROPOSITION 3.3. *For $\alpha < 1$, according to the transformations in* (3.8)*, the American put price $P_A(S,t)$ under the CEV model can be expressed as follows:*

$$P_A(S,t) = \begin{cases} e^{-r(T-t)}\int_0^{\tau} \frac{\partial V_1(a,\lambda)}{\partial \lambda} G(Y,\tau-\lambda)d\lambda \quad when \quad S > (\frac{\sigma a}{2})^{2\nu} e^{-(r-q)(T-t)}, \\ V_1(Y,\tau)e^{-r(T-t)} \quad when \quad S_p^*(t) < S \leq (\frac{\sigma a}{2})^{2\nu} e^{-(r-q)(T-t)}, \\ K - S \quad when \quad 0 \leq S \leq S_p^*(t), \end{cases} \tag{3.43}$$

where $V_1(Y,\tau)$ is the solution of (3.36)*,*

$$\begin{aligned} G(Y,\tau) &= L_s^{-1}\left[\left(\frac{Y}{a}\right)^{\nu} \frac{K_\nu(\sqrt{s}Y)}{sK_\nu(\sqrt{s}a)}\right]\Big|_{t=\tau}, \\ &= 1 + \frac{2}{\pi}\left(\frac{Y}{a}\right)^{\nu} \int_0^{\infty} e^{-u^2\tau} \frac{J_\nu(uY)Z_\nu(ua) - Z_\nu(uY)J_\nu(ua)}{J_\nu^2(ua) + Z_\nu^2(ua)} \frac{du}{u}, \end{aligned}$$

and $J_\nu(z)$ and $Z_\nu(z)$ are, respectively, the Bessel functions of the first and the second kind. In addition,

$$S_p^*(t) = \left(\frac{\sigma Y_p^*(\tau)}{2}\right)^{2\nu} e^{-(r-q)(T-t)}.$$

Proof. The proof is a direct consequence of Proposition 3.1 and (3.31). □

3.3. Integral kernels and numerical Laplace inversion. The function

$$H_\nu(\tau-\lambda) = L_s^{-1}\left[\frac{-K_{\nu-1}(\sqrt{s}a)}{\sqrt{s}K_\nu(\sqrt{s}a)}\right]\Big|_{t=\tau-\lambda},$$

which appears in the exact artificial boundary condition (3.36), requires a Laplace inversion. By the inversion theorem, we have

$$(3.44)\quad H_\nu(t) = -\frac{1}{2\pi i}\int_{\gamma-i\infty}^{\gamma+i\infty} \frac{K_{\nu-1}(\sqrt{s}a)}{\sqrt{s}K_\nu(\sqrt{s}a)} ds = -\frac{4}{\pi^2}\int_0^{\infty} \frac{e^{-u^2 t}}{au\{J_\nu^2(au) + Z_\nu^2(au)\}} du.$$

Therefore, the numerical method involves the computation of this integral over a semi-infinite domain, which, if not handled carefully, may deteriorate the efficiency once the analytic solution of the Laplace inversion is not available.

In this paper, we use a mixture of the Euler summation and the fast Fourier transform (FFT) to compute the Laplace inversion numerically. These methods are summarized in the appendix. FFT can significantly reduce the computational time, especially for inverting $H_\nu(t)$ for many t-values. As is discussed in [8], FFT methods are effective only when a large number of terms is involved. Specifically, FFT deals with the finite Fourier transform pairs $(X(j), A(k))$:

$$X(j) = \sum_{k=0}^{N-1} A(k) \exp\left(\frac{2\pi i}{N} jk\right), \qquad j = 0, 1, \ldots, N-1. \tag{3.45}$$

A direct calculation of (3.45) requires N^2 operations. FFT is a well-constructed algorithm that reduces the computation to an order of $N \log_2 N$, where $\log_2 N$ is an integer.

However, it is easy to see that $H_\nu(t)$ has a complicated infinity at $t = 0$. Because the Laplace inversion calculated by the FFT has a large error for values near singularity, i.e., when the option is close to maturity in the finance term, we employ the Euler summation for the first several points. This not only accelerates the convergence but also avoids divergent series generated from the singularity. The advantage of the Euler algorithm is obvious for a few t-values. This mixture of the FFT and the Euler summation reduces the computational error remarkably and retains the efficiency.

Let $\mathbf{e} = (e_0, e_1, \ldots, e_{N-1})^T$ be the error vector of N components such that

$$e_j = \widehat{H}_\nu(t_j) - H_\nu(t_j), \qquad j = 0, 1, \ldots, N-1, \tag{3.46}$$

where t_j are the elements of the time vector $\mathbf{t} = (t_0, t_1, \ldots, t_{N-1})$, $\widehat{H}_\nu(t_j)$ is obtained by the numerical Laplace inversion, and $H_\nu(t_j)$ is the exact value of the integral kernel. The L_1, L_2, and L_∞ norms of the error vector $\mathbf{e}$ are defined as follows:

$$\|\mathbf{e}\|_{L_1} = \sum_{i=0}^{N-1} |e_i|, \quad \|\mathbf{e}\|_{L_2} = \left(\frac{1}{N} \sum_{i=0}^{N-1} |e_i^2|\right)^{1/2}, \quad \|\mathbf{e}\|_{L_\infty} = \max_{0 \le i \le N-1} |e_i|. \tag{3.47}$$

The efficiency and accuracy of our method is investigated. For comparison, we look at several special cases in which simplification can be applied to compute the integral kernel.

When $\nu = \frac{1}{2}$, the modified Bessel functions in the integrand of $H_\nu(t)$ can be expressed in terms of trigonometric functions:

$$J_{\frac{1}{2}}(z) = Z_{-\frac{1}{2}}(z) = \left(\frac{1}{2}\pi z\right)^{-\frac{1}{2}} \sin z, \quad Z_{\frac{1}{2}}(z) = -J_{-\frac{1}{2}}(z) = -\left(\frac{1}{2}\pi z\right)^{-\frac{1}{2}} \cos z. \tag{3.48}$$

In fact, the Bessel functions with the order $\nu = n + \frac{1}{2}$ $(n = 0, \pm 1, \pm 2, \ldots)$ are elementary functions. In addition, for $n = 0, 1, 2, \ldots$,

$$J_{-n-\frac{1}{2}}(z) = (-1)^{n+1} Z_{n+\frac{1}{2}}(z), \quad Z_{-n-\frac{1}{2}}(z) = (-1)^n J_{n+\frac{1}{2}}(z), \tag{3.49}$$

$$J_{n+\frac{1}{2}}(z) = (-1)^n \left(\frac{1}{2}\pi z\right)^{-\frac{1}{2}} z^{n+1} \left(\frac{d}{z dz}\right)^n \frac{\sin z}{z},$$

TABLE 3.1
Errors in different norms.

α	N	Euler			FFT			CPU(s)
		L_1	L_2	L_∞	L_1	L_2	L_∞	
0	64	4.1045e-7	4.1045e-7	4.0145e-7	3.0803e-5	2.5662e-5	2.5383e-5	0.0126
	128	9.7115e-7	6.9675e-7	5.6891e-7	6.5080e-6	4.8217e-6	4.6712e-6	0.0245
	256	1.8399e-6	1.0905e-6	8.0544e-7	2.3845e-6	1.5816e-6	1.4899e-6	0.0296
	512	3.1733e-6	1.6449e-6	1.1397e-6	1.2262e-6	7.0317e-7	6.3379e-7	0.0326
2/3	64	2.9565e-7	2.9565e-7	2.9565e-7	2.5962e-5	2.1808e-5	2.1596e-5	0.0140
	128	7.5813e-7	5.4865e-7	4.6164e-7	5.6266e-6	4.1961e-6	4.0707e-6	0.0239
	256	1.5178e-6	9.0999e-7	6.9711e-7	2.1337e-6	1.4202e-6	1.3387e-6	0.0287
	512	2.7406e-6	1.4370e-6	1.0306e-6	1.1525e-6	6.6987e-7	6.0817e-7	0.0332

and

$$Z_{n+\frac{1}{2}}(z) = -(-1)^n \left(\frac{1}{2}\pi z\right)^{-\frac{1}{2}} z^{n+1} \left(\frac{d}{zdz}\right)^n \frac{\cos z}{z}. \tag{3.50}$$

Two integral kernels are used in our numerical examples:

1. When $\nu = \frac{1}{2}$ $(\alpha = 0)$,

$$H_{\frac{1}{2}}(t) = -\frac{4}{\pi^2}\int_0^\infty \frac{\pi e^{-u^2 t}}{2} du = -\frac{1}{\sqrt{\pi t}}\mathrm{Erf}(+\infty) = -\frac{1}{\sqrt{\pi t}}. \tag{3.51}$$

2. When $\nu = \frac{3}{2}$ $(\alpha = \frac{2}{3})$,

$$H_{\frac{3}{2}}(t) = -\int_0^\infty \frac{2a^2u^2e^{-u^2t}}{\pi(1+a^2u^2)} du = -\frac{1}{\sqrt{\pi t}} + \frac{1}{a}\exp\left(\frac{t}{a^2}\right)\mathrm{Erfc}\left(\frac{\sqrt{t}}{a}\right). \tag{3.52}$$

Consider the values of $H_\nu(t)$ for $t \in (0, 3)$. The errors under different norms for the numerical Laplace inversion are shown in Table 3.1. The points near singularity are computed by the Euler summation, whereas the others are computed by the FFT. If N increases, then more points will be placed in the neighborhood of the singularity and computed by the Euler summation. The CPU times are similar for different integral kernels and are less than 0.04 seconds for 512 points. It is clear that the mixture of the FFT method and the Euler summation achieves a high level of accuracy and efficiency.

4. Numerical experiments. As the closed-form pricing formulas of European options under the CEV model are available as shown in (2.4), we first investigate the sensitivity on the truncation parameter a to European option prices by using the proposed numerical method. Then we present the results for a European put option with different grid points. This attempts to show the convergence of the proposed scheme and demonstrate that our method outperforms the traditional CN scheme. Finally, we implement our method with American put options and make comparisons with alternative approaches.

As there is no exact solution for American options, Han and Wu [12] regard the binomial method with many time steps as true values to examine their numerical scheme. We subscribe to the same notion and use the lattice methods for the CEV model, as proposed by Lu and Hsu [16], with many time steps as the benchmark.

After imposing an artificial boundary Γ_a to bound the domain $\Omega_{\text{int}} = \{(Y, \tau) \in R^2 \mid Y_p^*(\tau) \le Y < a, 0 \le \tau \le T^*\}$, we need only to solve problem (3.36) with

the derived exact artificial boundary condition (3.35). Although all FDMs for the boundary value problem are possible choices, we consider the CN scheme with an evenly spaced grid, as it is the most widely used method for option pricing.

4.1. The numerical scheme. The interval $[0, T^*]$ is divided into N equally time-spaced subintervals, namely,

$$0 = \tau_0 < \tau_1 < \cdots < \tau_n < \cdots < \tau_N = T^*. \tag{4.1}$$

The interval $[0, a]$ for the state variable Y is divided into M subintervals with equal length, namely,

$$0 = Y_0 < Y_1 < \cdots < Y_m < \cdots < Y_M = a. \tag{4.2}$$

The central-time and central-space finite difference approximations are used as follows:

$$\frac{\partial V_1}{\partial \tau} \approx \frac{u_m^{n+1} - u_m^n}{\Delta \tau}, \tag{4.3}$$

$$\frac{\partial V_1}{\partial Y} \approx \frac{1}{2}\left(\frac{u_{m+1}^{n+1} - u_{m-1}^{n+1}}{2\Delta Y} + \frac{u_{m+1}^{n} - u_{m-1}^{n}}{2\Delta Y}\right), \tag{4.4}$$

$$\frac{\partial^2 V_1}{\partial Y^2} \approx \frac{1}{2}\left(\frac{u_{m+1}^{n+1} - 2u_m^{n+1} + u_{m-1}^{n+1}}{\Delta Y^2} + \frac{u_{m+1}^{n} - 2u_m^{n} + u_{m-1}^{n}}{\Delta Y^2}\right), \tag{4.5}$$

where $u_m^n = V_1(Y_m, \tau_n)$. The truncation error of the scheme is $O(\Delta Y^2, \Delta \tau^2)$.

By substituting the approximations into (3.36), we obtain

$$\begin{aligned}\frac{u_m^{n+1} - u_m^n}{\Delta \tau} = {} & \frac{1}{2}\left(\frac{u_{m+1}^{n+1} - 2u_m^{n+1} + u_{m-1}^{n+1}}{\Delta Y^2} + \frac{u_{m+1}^{n} - 2u_m^{n} + u_{m-1}^{n}}{\Delta Y^2}\right) \\ & + \frac{1 - 2\nu}{m \times \Delta Y} \times \frac{1}{2}\left(\frac{u_{m+1}^{n+1} - u_{m-1}^{n+1}}{2\Delta Y} + \frac{u_{m+1}^{n} - u_{m-1}^{n}}{2\Delta Y}\right).\end{aligned} \tag{4.6}$$

Let

$$g(Y, \tau) = \max\left(Ke^{r(T-t)} - \left(\frac{\sigma Y}{2}\right)^{2\nu} e^{q(T-t)}, 0\right), \tag{4.7}$$

$$g_m^n = g(Y_m, \tau_n). \tag{4.8}$$

A linear system of equations is established to compute the numerical values across each time interval, as follows:

$$\begin{aligned}\left(-\frac{\rho}{2} + \frac{1-2\nu}{4m}\rho\right) u_{m-1}^{n+1} + (\rho + 1) u_m^{n+1} + \left(-\frac{\rho}{2} - \frac{1-2v}{4m}\rho\right) u_{m+1}^{n+1} = \\ \left(\frac{\rho}{2} - \frac{1-2\nu}{4m}\rho\right) u_{m-1}^{n} + (-\rho + 1) u_m^{n} + \left(\frac{\rho}{2} + \frac{1-2v}{4m}\rho\right) u_{m+1}^{n},\end{aligned} \tag{4.9}$$

$$u_m^n \geq g_m^n, \tag{4.10}$$

$$u_m^0 = g_m^0, \tag{4.11}$$

where $n = 0, 1, 2, \ldots, N-1$, $m = 1, 2, \ldots, M$, and $\rho = \frac{\Delta \tau}{\Delta Y^2}$ is the mesh ratio.

For the exact artificial boundary condition on Γ_a, we use the approximation

$$\left.\frac{\partial V_1(Y,\tau)}{\partial Y}\right|_{Y=a} = \int_0^\tau \frac{\partial V_1(a,\lambda)}{\partial \lambda} H_\nu(\tau-\lambda)d\lambda \approx \frac{1}{\Delta\tau}\sum_{\ell=0}^{n-1}(u_M^{\ell+1}-u_M^\ell)\int_{\tau_\ell}^{\tau_{\ell+1}} H_\nu(\tau_n-\lambda)d\lambda, \tag{4.12}$$

where the truncation error is $O(\Delta\tau^\beta)$, with β depending on the integral kernel. For instance, $\beta = 3/2$ when $\alpha = 0$ [21].

We approximate the exact artificial boundary condition as follows:

$$\left.\frac{\partial V_1(Y,\tau)}{\partial Y}\right|_{Y=a} \approx \sum_{\ell=0}^{n-1}(u_M^{\ell+1}-u_M^\ell)\times\frac{H_\nu(\tau_n-\tau_{\ell+1})+H_\nu(\tau_n-\tau_\ell)}{2}, \tag{4.13}$$

where the function $H_\nu(t)$ can be evaluated by the numerical Laplace inversion shown in the appendix. The approximation in (4.13) is in fact the trapezoid rule, which approximates the integral $\int_a^b f(x)dx$ by $(b-a)\frac{f(a)+f(b)}{2} + O((b-a)^3)$. Thus, the aggregated error associated with the approximated boundary condition should be of $O(\Delta t^\beta + \Delta t^2)$. As β depends on the integral kernel, which is generally obtained from the Laplace inversion, determining the aggregated error for the approximated boundary involves tedious calculation. However, the most important thing should be the error for the numerical option price itself. Numerical examples are used to check the accuracy for European and American option prices. As a byproduct, we also check the order of convergence.

In the traditional CN scheme, the domains for the PDEs (2.7) and (2.11) are truncated to $[0, S_{max}]$ for a large S_{max}. In our example, we set S_{max} to twice the strike price for a put option. The far-field boundary condition is then given by

$$P_E(S_{max},\tau) \approx 0. \tag{4.14}$$

By expressing the problem in matrix-vector form, the values for u_m^n can be solved efficiently in the matrix-friendly Matlab program. All of the computations are implemented on a PC with an Intel Core 2 Duo E6300 1.86G Hz CPU and 1.0 GB of RAM. For the above discrete problem (4.11)–(4.14), we perform the following stability analysis.

PROPOSITION 4.1. *The discrete problem* (4.7)–(4.10) *is unconditionally stable, in the sense that* $\|u^n\|_\infty \le \|u^0\|_\infty, n = 1, 2, \ldots, N$, *independent of the time step and mesh spacing.*

Proof. Let $\{u^n\}$ be the solution of (4.7)–(4.10). Denote that

$$u_m^n = \|u^n\|_\infty = \max_{1\le i\le M}\{u_i^n\}. \tag{4.15}$$

Moreover, note that $u_m^n \ge g_m^n$. If $u_m^n = g_m^n$, then we have

$$u_m^n = \|u^n\|_\infty = g_m^n \le g_m^{n-1} \le u_m^{n-1} \le u_m^n = \|u^{n-1}\|_\infty,$$

where $n = 1, 2, \ldots, N$.

Otherwise, by (4.9), we have

$$\left(-\frac{\rho}{2}+\frac{1-2\nu}{4m}\rho\right)u_{m-1}^n + (\rho+1)u_m^n + \left(-\frac{\rho}{2}-\frac{1-2v}{4m}\rho\right)u_{m+1}^n = \left(\frac{\rho}{2}-\frac{1-2\nu}{4m}\rho\right)u_{m-1}^{n-1} + (-\rho+1)u_m^{n-1} + \left(\frac{\rho}{2}+\frac{1-2v}{4m}\rho\right)u_{m+1}^{n-1},$$

and hence we can obtain

$$(\rho+1)\|u^n\|_\infty \le \left(\frac{\rho}{2}+\frac{1-2\nu}{4m}\rho+\frac{\rho}{2}-\frac{1-2\nu}{4m}\rho\right)\|u^n\|_\infty + \left(\frac{\rho}{2}-\frac{1-2\nu}{4m}\rho-\rho+1+\frac{\rho}{2}+\frac{1-2\nu}{4m}\rho\right)\|u^{n-1}\|_\infty,$$

which can be simplified as

$$\|u^n\|_\infty \le \|u^{n-1}\|_\infty, \quad n=1,2,\ldots,N.$$

The proof is then completed by recurrence on n. □

Remark 1. Note that an evenly spaced grid is employed not solely for computational convenience but also for comparison to the widely used CN scheme. The proposed scheme can be further enhanced by considering variably sized time steps to obtain quadratic convergence. We refer to [10] for details.

Remark 2. For $\alpha > 1$, the artificial boundary condition is not required after transforming the original PDE into the standardized CEV equation. When $\alpha > 1$, the standardization procedure in section 3 leads to the PDE (3.9) defined on the bounded domain $(0, Y_p^*(\tau))\times[0,T^*]$ in which the exact boundary condition $\lim_{Y\to 0} V_1(Y,\tau) = 0$ is observed. For American put options, the early exercise boundary $S_p^\infty \le S_p^*(t) \le S_p^0$ for $0 \le t \le T$ corresponds to that using a new set of independent variables: $Y_p^0 \le Y_p^*(\tau) \le Y_p^\infty$ for $0 \le \tau \le T^*$. Thus, the traditional CN scheme can be implemented over the bounded domain $(0, Y_p^\infty)$ for $\alpha > 1$.

4.2. Sensitivity on the truncation size. The sensitivity on the truncation parameter a plays a key role in the proposed numerical method. According to the transformation $Y = \frac{2}{\sigma}(Se^{(r-q)(T-t)})^{\frac{1}{2\nu}}$, we choose

$$a = \max\left(\frac{2}{\sigma}(S_{max}e^{(r-q)T})^{\frac{1}{2\nu}}, \frac{2}{\sigma}(S_{max})^{\frac{1}{2\nu}}\right). \tag{4.16}$$

Figure 4.1 plots the errors of the proposed ABC scheme and the traditional CN scheme (FDM) against the value of S_{max}. The computation is based on 256×256 grids, a current stock price of 100, $r = 5\%$, and $q = 0$. The value of σ is set such that the local volatility is σ_0 given the value of α. It can be seen from Figures 4.1(c) and 4.1(d) that the ABC method is more robust with respect to the truncation size and the model parameters than is the FDM. In particular, the sufficiently large S_{max} is of levity in the FDM, and a large domain is often required to secure the accuracy, especially when the volatility is high and/or the maturity is long. Therefore, the literature suggests setting S_{max} to twice the strike price.

Figures 4.1(a) and 4.1(b) show that the errors of the proposed method are less than 0.001 once $S_{max} \ge 120$. Thus, imposing an artificial boundary condition around 1.2 times the strike price gives a reasonably good numerical values for the option price.

As the size of the computational domain and the grids are the same for the ABC and the FDM, Figure 4.1 indicates that the enhanced performance of the ABC approach is not because of a finer grid spacing. To further evaluate the costs and benefits of the FDM and the ABC, it is more desirable to make a comparison based on computational times and errors.

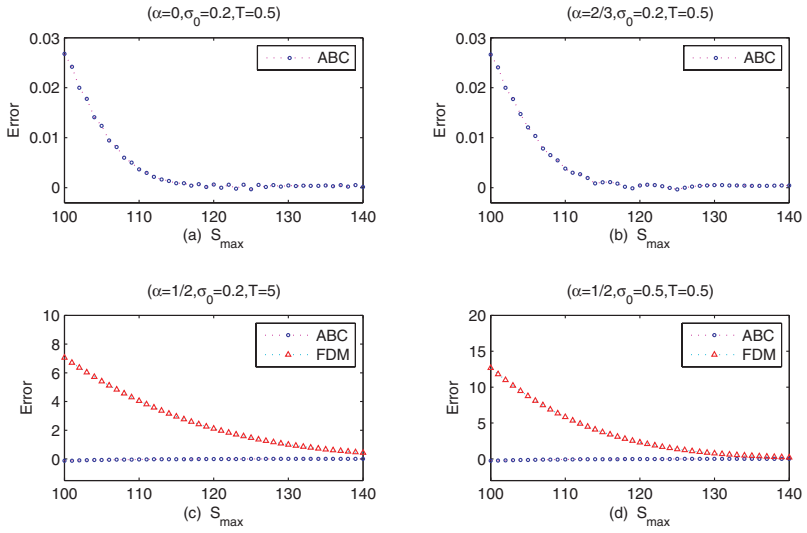


FIG. 4.1. *Sensitivity of the errors on the truncation parameter.*

TABLE 4.1
European and American puts with $\alpha = 0$.

	European				American			
$\alpha = 0$	K=90	K=100	K=110	CPU(s)	K=90	K=100	K=110	CPU(s)
Closed-form	1.4680	4.4244	9.9552	0.0093	/	/	/	/
FDM:64*64	1.4443	4.3896	9.9202	0.0051	1.4998	4.5970	10.7273	0.0059
(error)	(0.0237)	(0.0349)	(0.0350)		(0.0124)	(0.0420)	(0.0242)	
FDM:128*128	1.4645	4.4157	9.9496	0.0229	1.5102	4.6274	10.7441	0.0242
(error)	(0.0035)	(0.0088)	(0.0056)		(0.0020)	(0.0116)	(0.0074)	
FDM:256*256	1.4668	4.4222	9.9535	0.1492	1.5112	4.6357	10.7483	0.1534
(error)	(0.0012)	(0.0023)	(0.0017)		(0.0010)	(0.0033)	(0.0032)	
FDM:512*512	1.4679	4.4238	9.9550	1.0204	1.5119	4.6380	10.7504	1.0424
(error)	(0.0001)	(0.0007)	(0.0002)		(0.0003)	(0.0010)	(0.0011)	
ABC:64*64	1.4645	4.4209	9.9456	0.0057	1.5085	4.6335	10.7455	0.0095
(error)	(0.0035)	(0.0036)	(0.0096)		(0.0037)	(0.0055)	(0.0060)	
ABC:128*128	1.4676	4.4233	9.9523	0.0244	1.5117	4.6374	10.7507	0.0320
(error)	(0.0004)	(0.0012)	(0.0029)		(0.0005)	(0.0016)	(0.0008)	
ABC:256*256	1.4677	4.4242	9.9543	0.1498	1.5119	4.6383	10.7509	0.1660
(error)	(0.0003)	(0.0003)	(0.0009)		(0.0003)	(0.0007)	(0.0006)	
ABC:512*512	1.4680	4.4245	9.9552	1.0344	1.5122	4.6388	10.7514	1.1085
(error)	(0.0000)	(0.0000)	(0.0000)		(0.0000)	(0.0002)	(0.0001)	
Binomial:5000	1.4680	4.4245	9.9552	982.99	1.5122	4.6390	10.7515	1041.7

4.3. Examples. We now give several examples to show the use of the proposed artificial boundary method for CEV option pricing.

Example 4.1. Consider the pricing of European and American put options with $\nu = \frac{1}{2}$, which corresponds to $\alpha = 0$. We have discussed the properties of the integral kernel under this situation in section 3.3. It is clear that the standardized CEV equation is reduced to a heat equation when $\nu = \frac{1}{2}$.

Comparisons are made by using the closed-form put pricing formula in (2.2)–(2.4). The current stock price is 100, the time to expiry is six months, $r = 5\%$, $q = 0$ (nondividend), and the at-the-money volatility is 20%. We choose $S_{max} = 120$ for the ABC and $S_{max} = 200$ for the traditional CN scheme (FDM), respectively. It can be seen from Table 4.1 that the ABC method is efficient and accurate. Answers accurate

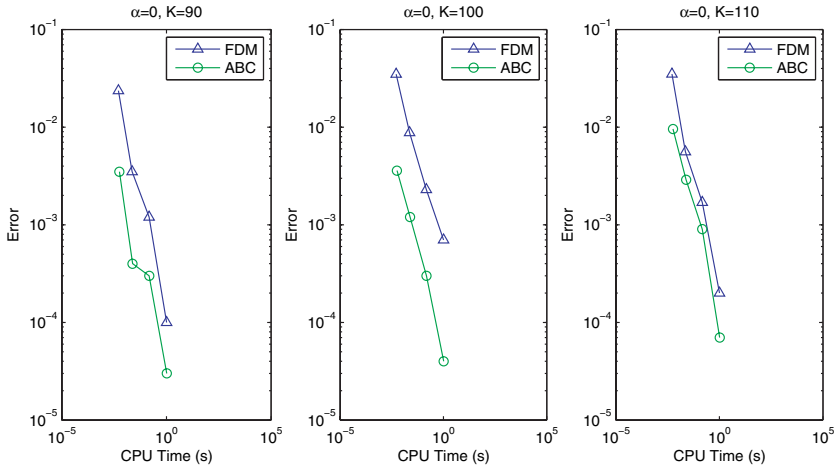


FIG. 4.2. *Error for the FDM and the ABC with* $\alpha = 0$.

to within 0.01 are given by using 64×64 grids for all three cases. The CPU time is even less than in the closed-form method. Compared to the traditional CN method with 128×128 grids, the proposed method is more than four times faster, with the same level of accuracy. We stress that the finite difference scheme computes not one, but many, option prices at the same time. For the case of 64×64 grids, the computational time refers to the time to obtain 64 option prices. By using interpolation, we are able to obtain the option value on the entire domain. These many option prices are useful for computing option Greeks, such as Delta and Gamma. The binomial method with 5000 time steps also generates accurate results but is much less efficient.

The performance of the proposed method for valuing American options is investigated next. As no closed-form solution is available, we compare the results of the ABC to those obtained by the FDM with many more grid points and the binomial method with 5000 time steps. Once again, the ABC method outperforms the traditional CN scheme with its high convergent speed. Moreover, the ABC method exceeds the binomial method with its greater efficiency. From Table 4.1, when the strike price is set to 110, the result of the traditional CN scheme with 512×512 grids is 10.7504 with an evaluation time of 1.0424 seconds. By using the ABC method, a result of 10.7507 is obtained for 128×128 grids, and the computational time is only 0.0320 seconds. The binomial method is much more time-consuming for the same level of accuracy.

We plot the computational error against the execution time for different K in Figure 4.2. The lower the curve, the better the performance of the scheme. It is obvious that, for all three cases, the ABC method outperforms the traditional CN scheme, especially for out-of-the-money and at-the-money options.

Example 4.2. We now take a look at a situation in which $0 < \alpha < 1$. In particular, we consider $\alpha = 2/3$, which is equivalent to $\nu = 3/2$. We use the properties in section 3.3 to implement the ABC method. The other parameters are the same as those in Example 4.1.

Table 4.2 summarizes the computational results. First we concentrate on the European put option. Unlike in Example 4.1, the integral kernel involves the computation of the error function in the present case. The evaluation time will be a little bit longer. However, the ABC method still outperforms the traditional CN scheme and

TABLE 4.2
European and American puts with $\alpha = \frac{2}{3}$.

	European				American			
$\alpha = 2/3$	K=90	K=100	K=110	CPU(s)	K=90	K=100	K=110	CPU(s)
Closed-form	1.3380	4.4202	10.1099	0.0244	/	/	/	/
FDM:64*64	1.3138	4.3856	10.0759	0.0057	1.3731	4.6066	10.8693	0.0061
(error)	(0.0222)	(0.0347)	(0.0340)		(0.0113)	(0.0425)	(0.0249)	
FDM:128*128	1.3347	4.4146	10.1043	0.0238	1.3828	4.6374	10.8860	0.0264
(error)	(0.0033)	(0.0087)	(0.0056)		(0.0016)	(0.0117)	(0.0082)	
FDM:256*256	1.3369	4.4181	10.1075	0.1538	1.3834	4.6457	10.8910	0.1562
(error)	(0.0011)	(0.0022)	(0.0024)		(0.0010)	(0.0034)	(0.0032)	
FDM:512*512	1.3376	4.4197	10.1094	1.0982	1.3840	4.6481	10.8931	1.1038
(error)	(0.0004)	(0.0006)	(0.0005)		(0.0004)	(0.0010)	(0.0011)	
ABC:64*64	1.3326	4.4156	10.1021	0.0100	1.3782	4.6399	10.8808	0.0176
(error)	(0.0054)	(0.0047)	(0.0078)		(0.0062)	(0.0092)	(0.0134)	
ABC:128*128	1.3369	4.4184	10.1063	0.0309	1.3832	4.6457	10.8878	0.0567
(error)	(0.0011)	(0.0019)	(0.0036)		(0.0012)	(0.0034)	(0.0064)	
ABC:256*256	1.3376	4.4198	10.1088	0.1570	1.3840	4.6476	10.8934	0.2082
(error)	(0.0004)	(0.0005)	(0.0011)		(0.0004)	(0.0015)	(0.0008)	
ABC:512*512	1.3378	4.4204	10.1097	1.5662	1.3841	4.6489	10.8939	1.7849
(error)	(0.0002)	(0.0001)	(0.0002)		(0.0003)	(0.0002)	(0.0003)	
Binomial:5000	1.3380	4.4203	10.1099	1699.2	1.3844	4.6491	10.8942	1913.3

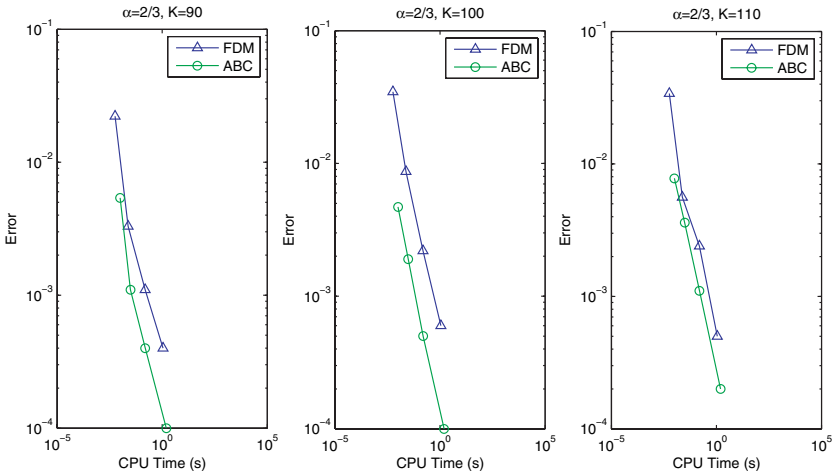


FIG. 4.3. *Error for the FDM and the ABC with* $\alpha = 2/3$.

the binomial method with 5000 time steps in terms of both efficiency and accuracy. When the error is plotted against the evaluation time in Figure 4.3, the superiority of the ABC method becomes obvious.

The ABC method is still very efficient when we use it to price American options, as reported in Table 4.2. For example, the ABC method with 256×256 grids generates the same accuracy level as the traditional CN scheme with 512×512 grids. The evaluation time is much less than the binomial methods with 5000 time steps.

Figure 4.3 gives a more complete picture to compare the performance. It can be seen that the curve corresponding to the ABC is always beneath that of the FDM approach. We are comfortable with concluding that the ABC approach outperforms the FDM ones for $\alpha = 2/3$.

TABLE 4.3
Values and Greeks of European put and American put with $\alpha = -3$.

	European				American			
$\alpha = -3$	K=90	K=100	K=110	CPU(s)	K=90	K=100	K=110	CPU(s)
Closed-form	2.2210	4.5055	9.3486	0.0045	/	/	/	/
FDM:64*64	2.1968	4.4684	9.3103	0.0056	2.2254	4.6010	10.2207	0.0067
(error)	(0.0242)	(0.0371)	(0.0383)		(0.0161)	(0.0401)	(0.0216)	
FDM:128*128	2.2171	4.4973	9.3426	0.0234	2.2342	4.6304	10.2344	0.0259
(error)	(0.0039)	(0.0082)	(0.0060)		(0.0023)	(0.0107)	(0.0079)	
FDM:256*256	2.2195	4.5032	9.3467	0.1476	2.2399	4.6382	10.2393	0.1532
(error)	(0.0015)	(0.0023)	(0.0019)		(0.0016)	(0.0029)	(0.0030)	
FDM:512*512	2.2205	4.5046	9.3480	1.2362	2.2412	4.6403	10.2412	1.5490
(error)	(0.0005)	(0.0009)	(0.0006)		(0.0003)	(0.0008)	(0.0011)	
ABC:64*64	2.2148	4.4998	9.3415	0.0241	2.2350	4.6328	10.2333	0.0335
(error)	(0.0062)	(0.0057)	(0.0071)		(0.0065)	(0.0083)	(0.0090)	
ABC:128*128	2.2198	4.5036	9.3452	0.0676	2.2402	4.6386	10.2386	0.0764
(error)	(0.0012)	(0.0019)	(0.0034)		(0.0013)	(0.0025)	(0.0037)	
ABC:256*256	2.2207	4.5051	9.3478	0.2654	2.2411	4.6398	10.2409	0.3123
(error)	(0.0003)	(0.0004)	(0.0008)		(0.0004)	(0.0013)	(0.0014)	
ABC:512*512	2.2209	4.5053	9.3485	1.5823	2.2413	4.6409	10.2419	2.2215
(error)	(0.0001)	(0.0002)	(0.0001)		(0.0002)	(0.0002)	(0.0004)	
Binomial:5000	2.2209	4.5053	9.3487	1724.3	2.2415	4.6411	10.2423	2012.9
delta	-0.1079	-0.2898	-0.6181	0.3578	-0.1106	-0.3142	-0.7934	0.4821
gamma	0.0106	0.0262	0.0399	0.3924	0.0217	0.0320	0.0324	0.5335
rho	-11.6295	-22.3032	-39.5805	0.5186	-11.0972	-13.7731	-18.3743	0.7130
vega	26.5072	28.7688	20.7496	0.6201	26.5556	27.9550	19.5724	0.8689
theta	4.1384	3.5241	2.1918	0.3870	4.1926	3.7212	2.9739	0.4902

Example 4.3. Another example is used to verify the performance of the ABC method. As the empirical results cited by Davydov and Linetsky [6] suggest that $\alpha = -3$ may be appropriate for the S&P 500 index, this example demonstrates the effect of this choice on the European and American option prices and their corresponding Greeks. Except for the value of α, the other parameters are the same as those in Example 4.1.

The results of the European and American put options are displayed in Table 4.3. Because the analytical solution of the integral kernel is unavailable, we resort to the numerical Laplace inversion. Hence, the evaluation time will be a little longer. The ABC method clearly outperforms the binomial method with 5000 time steps in both the European and the American options.

To compare the ABC method with the FDM, Figure 4.4 is constructed to examine the trade-off between accuracy and efficiency in both approaches. It can be seen that both approaches exhibit similar performance if one accepts that the error can be larger than 5×10^{-3}. In practice, option prices correct to three decimal places are quoted because deeply out-of-money option prices are typically low. Therefore, it is practical to consider the error to be less than or equal to 10^{-3}. For this range of errors, the ABC approach outperforms the FDM because all of the curves corresponding to the ABC are beneath those of the FDM, as shown in Figure 4.4. The ABC method has a clear practical advantage for the CEV model.

As important as option prices are the hedging parameters: *delta*, *gamma*, *vega*, *theta*, and *rho*. In the financial market, the aim of the trader is to manage the Greeks so that all risks are acceptable. Because finite difference calculations generate option values on an entire grid of S values, the calculation of *delta*, *gamma*, and *theta* can be determined with a negligible additional computational cost. The evaluation of

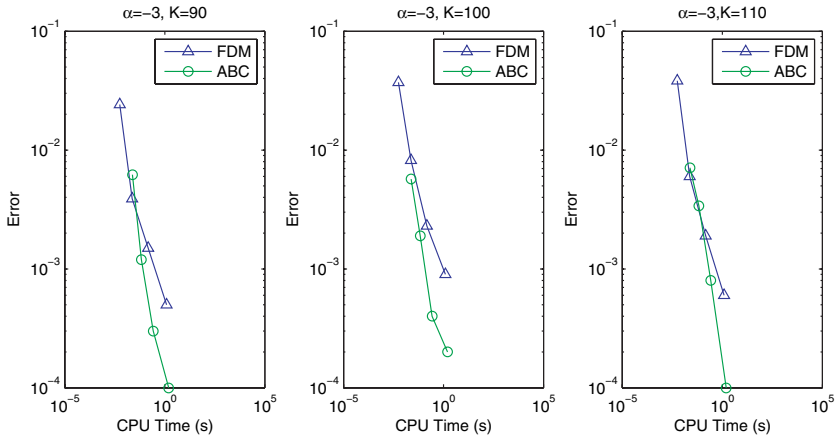


FIG. 4.4. *Error for the FDM and the ABC with* $\alpha = -3$.

the other hedging parameters in this manner requires the repeated solution of the equation for the option value. In Table 4.3, the results of the Greeks for European and American put options are obtained by using the ABC method with 256×256 grids. Our method works well.

4.4. Order of convergence. To show the order of convergence of the proposed scheme numerically, we consider Examples 4.1–4.3 again. As these examples share the same set of parameters, except for α, we can examine the order of convergence for different values of α, given other things being fixed.

Let $\mathbf{V_1}$ be a vector of option prices calculated by using the proposed scheme and $\mathbf{u}$ be that by using the exact solution. The exact solution for European put option is clear. We regard the numerical value obtained from the binomial method with very fine time steps as the exact solution of the American put option. We speculate that the error in the L_2 norm takes approximately the following form:

$$||\mathbf{V_1} - \mathbf{u}||_{L_2} \approx C\Delta\tau^p, \tag{4.17}$$

where C and p are constant values. To investigate the order of convergence with respect to the time step size, we estimate the order parameter p by using a linear regression:

$$\log ||\mathbf{V_1} - \mathbf{u}||_{L_2} \approx p \log \Delta\tau + \log C. \tag{4.18}$$

The estimated value of p is denoted as $\hat{p}$. To get a better estimate for p, we expand our calculation to cover grids of 32×32, 64×64, 128×128, 256×256, 512×512, 1024×1024, and 2048×2048.

Table 4.4 shows the results for European and American options with different combinations of K and α. It can be seen that all of the values of $\hat{p}$ are consistently larger than 1.5, and some of them even come close to 2. As the option payoff, or equivalently the initial condition of the PDE, is not a smooth function at the at-the-money point, quadratic convergence is impossible for an evenly grid CN scheme. Forsyth and Vetzal [10] propose a penalty approach that automatically generates variably sized time steps to obtain the quadratic convergence. Their method can be applied to our case as it does not conflict with the exact artificial boundary condition

TABLE 4.4
Estimated order of convergence $\hat{p}$.

	European			American		
	K=90	K=100	K=110	K=90	K=100	K=110
$\alpha = 0$	1.8734	1.9163	1.8932	1.8742	1.6492	1.7689
$\alpha = 2/3$	1.6376	1.8737	1.7932	1.6064	1.8347	1.8624
$\alpha = -3$	1.8721	1.7143	1.7689	1.7149	1.7239	1.6889

approach. As the present paper concentrates on the ABC approach, we would like to highlight its effect and thus do not implement the Forsyth and Vetzal [10] approach here.

5. Conclusion. We study the artificial boundary finite difference method for pricing American options under the CEV model. The exact artificial boundary condition is derived and implemented in several practical examples. Our numerical experiments show that the proposed approach outperforms the classical FDM and the binomial method in terms of both efficiency and robustness for highly accurate option prices.

Appendix A. Numerical Laplace inversion. The numerical Laplace inversion computes the original function $f(t)$ for a given single value of t. The FFT algorithm can often be exploited. The method used in this paper belongs to the Dubber–Abate family, which starts with the complex inversion formula

$$f(t) = \frac{1}{2\pi i}\int_{a-i\infty}^{a+i\infty} e^{st}F(s)ds, \tag{A.1}$$

where it is assumed that $F(s)$ is holomorphic in a right-half plane such that a can be any positive number. By assuming that $F(s)$ is real in the sense that $F(\bar{s}) = \overline{F(s)}$, the formula is transformed into

$$f(t) = \frac{2e^{at}}{\pi}\int_0^\infty \Re[F(a+iu)]\cos(ut)du = -\frac{2e^{at}}{\pi}\int_0^\infty \Im[F(a+iu)]\sin(ut)du,$$

where $\Re$ stands for the real part and $\Im$ stands for the imaginary part. The cosine-transform integral is numerically evaluated by means of the trapezoidal rule, which is a numerical approximation of the integral defined as follows:

$$\int_{x_0}^{x_1} g(x)\mathrm{d}x = \frac{x_1 - x_0}{2}(g(x_0) + g(x_1)) - \frac{1}{12}g''(c)\left(\frac{x_1 - x_0}{2}\right)^3, \tag{A.2}$$

where c lies somewhere between x_0 and x_1.

With step size h, we obtain

$$f(t) \approx f_h(t) := h\left\{\frac{e^{at}}{\pi}\Re[F(a)] + \frac{2e^{at}}{\pi}\sum_{k=1}^{\infty}\Re[F(a+ikh)]\cos(kht)\right\}. \tag{A.3}$$

For the given t and a positive number A, let us choose the step $h = \frac{\pi}{T}$, which yields the formula

$$f_h(t) := \frac{2e^{at}}{T}\left\{\frac{1}{2}\Re[F(a)] + \sum_{k=1}^{\infty}\Re\left[F\left(a + \frac{k\pi i}{T}\right)\right]\cos\left(\frac{k\pi}{T}t\right)\right\}. \tag{A.4}$$

We now cast the method into a form compatible with the finite Fourier transform (3.45) as follows:

$$f_h(t) := \frac{e^{at}}{T}\left\{\sum_{n=-\infty}^{+\infty} \Re\left[F\left(a+i\frac{\pi n}{T}\right)\right]\exp\left(i\frac{n\pi}{T}t\right)\right\}. \tag{A.5}$$

If $|f(t)|$ is bounded by a constant K, we obtain the error estimator (see [8]):

$$|error| = |f_h(t) - f(t)| \le \exp(-aT + at)\cosh(at)/\sinh(aT). \tag{A.6}$$

Since the error becomes prohibitively large as t approaches T, our representation is restricted to the interval $0 \le t \le T/2$. Now let the situation be such that we require $f(t)$ for the equidistant points $t = j\Delta t$, where $j = 0, 1, 2, \ldots, N = N_1/2$. Then Δt is the desired sampling interval, and the maximum t-value is $t_{max} = N\Delta t$. We have $T = 2t_{max}$ or $T = 2N\Delta t = N_1\Delta t$. By using these definitions, we are able to manipulate the series so that (A.5) can be put into the following form:

$$\frac{f(j\Delta t)}{b(j)} = \sum_{k=0}^{N_1-1} A(k)\exp\left(\frac{2\pi i}{N_1}jk\right), \tag{A.7}$$

where

$$A(k) = \frac{1}{N_1\Delta t}\sum_{n=-\infty}^{+\infty} \Re\left[F\left(a+\frac{2\pi i}{N_1}(k+nN_1)\right)\right], \quad b(j) = 2\exp(aj\Delta t). \tag{A.8}$$

The right-hand side of (A.7) is of the same form as (3.45), thereby permitting the use of FFT on the problem. The procedure is as follows:

1. Compute $A(k)$ from $F(s)$ by use of (A.8).
2. Let $A(k)$ be the input to the FFT subroutine. The output will be $X(j)$ for $j = 0, 1, 2, \ldots, N_1 - 1$.
3. The inverse function is given by $f(j\Delta t) = b(j)X(j)$, where $b(j)$ is defined by (A.8). However, due to our error analysis we accept only $b(j)X(j)$ as a representation of the function f for $j = 0, 1, 2, \ldots, N_1/2 = N$.

For the t-values near singularity, we employ another version of the Fourier series method with Euler summation. The cosine factor in (A.3) may be eliminated by choosing the step $h = \frac{\pi}{2t}$, i.e., $T = 2t$, and the integration abscissa $a = \frac{A}{2t}$ such that a nearly alternating series results:

$$f_h(t) := \frac{e^{A/2}}{2t}F\left(\frac{A}{2t}\right) + \frac{e^{A/2}}{t}\sum_{k=1}^{\infty}(-1)^k\Re\left[F\left(\frac{A+2k\pi i}{2t}\right)\right]. \tag{A.9}$$

However, the coefficients then become t-dependent and must be evaluated for each new value of t. The remaining problem is to evaluate the above infinite sum. Because we have a nearly alternating series, an acceleration method for alternating series, i.e., the Euler summation, is proposed. With

$$s_n(t) := \frac{e^{A/2}}{2t}F\left(\frac{A}{2t}\right) + \frac{e^{A/2}}{t}\sum_{k=1}^{n}(-1)^k\Re\left[F\left(\frac{A+2k\pi i}{2t}\right)\right], \tag{A.10}$$

the Euler summation is applied to m terms after an initial n (typically $m = 11$ and $n = 15$):

$$f(t) \approx f_h(t) \approx E(m,n,t) := \sum_{k=0}^{m} \left(\begin{array}{c} m \\ k \end{array} \right) 2^{-m} s_{n+k}(t). \tag{A.11}$$

It is possible to gain some insight into the discretization error that results from passing from $f(t)$ to $f_h(t)$. To establish this result, we use the formula derived by the discrete Laplace transform:

$$\sum_{k=0}^{\infty} f(kT+\tau)e^{-kTs} = pv\frac{1}{T}\sum_{k=-\infty}^{\infty} F\left(s + i\frac{2\pi}{T}k\right)\exp\left[\left(s + i\frac{2\pi}{T}k\right)\tau\right], \tag{A.12}$$

where pv means principle value. The formula holds for $0 < \tau < T$. By substituting t for τ, b for s, and $T = 2\pi/h$, we obtain

$$\sum_{k=0}^{\infty} f\left(t + k\frac{2\pi}{h}\right)e^{-k2\pi b/h} = pv\frac{h}{2\pi}\sum_{k=-\infty}^{\infty} F(b+ihk)\exp[(b+ihk)t] \tag{A.13}$$

and, with $h = \pi/t$ and $b = A/(2t)$,

$$\sum_{k=0}^{\infty} f((2k+1)t)e^{-kA} = pv\frac{e^{A/2}}{2t}\sum_{k=-\infty}^{\infty} F\left(\frac{A+2\pi ik}{2t}\right)(-1)^k. \tag{A.14}$$

This can be written as

$$\begin{aligned} &f(t) + \sum_{k=1}^{\infty} f((2k+1)t)e^{-kA} \\ &= \frac{e^{A/2}}{2t}F\left(\frac{A}{2t}\right) + pv\frac{e^{A/2}}{2t}\sum_{k=1}^{\infty}(-1)^k \Re\left[F\left(\frac{A+2\pi ik}{2t}\right)\right]. \end{aligned} \tag{A.15}$$

Therefore,

$$f(t) = f_h(t) - \sum_{k=1}^{\infty} f((2k+1)t)e^{-kA}. \tag{A.16}$$

This shows that the sampling error associated with the trapezoidal rule is given by

$$err_d = \sum_{k=1}^{\infty} f[(2k+1)t]e^{-kA}. \tag{A.17}$$

If we know, for example, that $|f(t)|$ is bounded by a constant K, then we obtain the estimate

$$|err_d| \le K\frac{e^{-A}}{1-e^{-A}} \approx Ke^{-A} \tag{A.18}$$

when e^{-A} is small.

REFERENCES

[1] S. Becker, *The constant elasticity of variance model and its implications for option pricing*, J. Finance, 35 (1980), pp. 661–673.

[2] M. Brennan and E. Schwartz, *The valuation of American put options*, J. Finance, 32 (1977), pp. 449–462.

[3] P. Carr, R. Jarrow, and R. Myneni, *Alternative characterizations of American put options*, Math. Finance, 2 (1992), pp. 87–106.

[4] H. S. Carslay and J. C. Jaeger, *Conduction of Heat in Solids*, Clarendon Press, Oxford, 1959.

[5] J. Cox, *Notes on Option Pricing* I: *Constant Elasticity of Variance Diffusions*, working paper, Stanford University, 1975 (reprinted in J. Portfolio Management, 22 (1996), pp. 15–17).

[6] D. Davydov and V. Linetsky, *Pricing and hedging path dependent options under the CEV process*, Management Sci., 47 (2001), pp. 949–965.

[7] J. Detemple and W. T. Tian, *The valuation of American options for a class of diffusion processes*, Management Sci., 48 (2002), pp. 917–937.

[8] H. Dubner and J. Abate, *Numerical inversion of Laplace transforms by relating them to the finite Fourier cosine transform*, J. ACM, 15 (1968), pp. 115–123.

[9] D. Emanuel and J. MacBeth, *Further results on the constant elasticity of variance call option pricing model*, J. Financial and Quantitative Anal., 17 (1982), pp. 533–554.

[10] P. A. Forsyth and K. R. Vetzal, *Quadratic convergence for valuing American options using a penalty method*, SIAM J. Sci. Comput., 23 (2002), pp. 2095–2122.

[11] H. Han and Z. Huang, *A class of artificial boundary conditions for heat equation in unbounded domains*, Comput. Math. Appl., 43 (2000), pp. 889–900.

[12] H. Han and X. Wu, *A fast numerical method for the Black–Scholes equation of American options*, SIAM J. Numer. Anal., 41 (2003), pp. 2081–2095.

[13] R. Kangro and R. Nicolaides, *Far field boundary conditions for Black–Scholes equations*, SIAM J. Numer. Anal., 38 (2000), pp. 1357–1368.

[14] I. J. Kim, *The analytic valuation of American puts*, Rev. Financial Studies, 3 (1990), pp. 547–572.

[15] Y. K. Kwok, H. Y. Wong, and K. W. Lau, *Pricing algorithms of multivariate path dependent options*, J. Complexity, 17 (2001), pp. 773–794.

[16] R. Lu and Y. H. Hsu, *Valuation of standard options under the constant elasticity of variance model*, Internat. J. Bus. Econ., 4 (2005), pp. 157–165.

[17] R. C. Merton, *The theory of rational option pricing*, Bell J. Econ. Management Sci., 4 (1973), pp. 141–183.

[18] M. Schroder, *Computing the constant elasticity of variance option pricing formula*, J. Finance, 44 (1989), pp. 211–219.

[19] D. Tavella and C. Randall, *Pricing Financial Instruments: The Finite Difference Method*, John Wiley, New York, 2000.

[20] H. Windcliff, P. A. Forsyth, and K. R. Vetzal, *Analysis of the stability of the linear boundary condition for the Black-Scholes equation*, J. Comput. Finance, 8 (2004), pp. 65–92.

[21] X. Wu and Z. Sun, *Convergence of difference scheme for heat equation in unbounded domains using artificial boundary conditions*, Appl. Numer. Math., 50 (2004), pp. 261–277.

[22] Y. Zhu, X. Wu, and I. Chern, *Derivative Securities and Difference Methods*, Springer, New York, 2004.

SIAM J. NUMER. ANAL.
Vol. 46, No. 5, pp. 2211–2237
© 2008 Society for Industrial and Applied Mathematics

CONVERGENCE OF GODUNOV-TYPE SCHEMES FOR SCALAR CONSERVATION LAWS UNDER LARGE TIME STEPS*

JING-MEI QIU† AND CHI-WANG SHU†

Abstract. In this paper, we consider convergence of classical high-order Godunov-type schemes towards entropy solutions for scalar conservation laws. It is well known that sufficient conditions for such convergence include total variation boundedness of the reconstruction and cell or wavewise entropy inequalities. We prove that under large time steps, we only need total variation boundedness of the reconstruction to guarantee such convergence. We discuss high-order total variation bounded reconstructions to fulfill this sufficient condition and provide numerical examples on one-dimensional convex conservation laws to assess the performance of such large time step Godunov-type methods. To demonstrate the generality of this approach, we also prove convergence and give numerical examples for a large time step Godunov-like scheme involving Sanders' third-order total variation diminishing reconstruction using both cell averages and point values at cell boundaries.

Key words. hyperbolic conservation laws, Godunov-type scheme, high-order accuracy, total variation diminishing, total variation bounded, convergence, entropy solution

AMS subject classifications. 65M06, 65M12

DOI. 10.1137/060657911

1. Introduction.

In this paper, we consider the Cauchy problem of the scalar conservation law,

$$\begin{cases} u_t + \nabla \cdot f(u) = 0 & \text{in } \mathbb{R}^m \times [0,T], \\ u(x,0) = u_0(x) & \text{in } \mathbb{R}^m, \end{cases} \tag{1.1}$$

where $u_0(x) \in \mathrm{BV}$, the space of functions with bounded variation. We do not consider boundary conditions and assume that the initial condition is either compactly supported or periodic. In most of our presentation we will concentrate on the one-dimensional case $m = 1$, and will point out similarities and differences for higher dimensions $m > 1$. The scheme under consideration is the classical Godunov-type scheme, consisting of the following three stages to evolve the cell averages from time level n (denoted by $\bar{u}^n$) to time level $n+1$ (denoted by $\bar{u}^{n+1}$).

1. *Reconstruction:* obtain a high-order piecewise polynomial reconstruction $u^n(x)$ whose cell averages agree with the given cell averages $\bar{u}^n$. We denote this reconstruction operator as $Re(\bar{u}^n)$.
2. *Evolution:* evolve $u^n(x)$ by the conservation law (1.1) *exactly* for a time step Δt to obtain a solution $\tilde{u}^{n+1}(x)$, which is in general *not* piecewise polynomial anymore. We denote this evolution operator as $S_{\Delta t}(u^n)$.
3. *Averaging:* average the function $\tilde{u}^{n+1}(x)$ to obtain the cell averages $\bar{u}^{n+1}$ at time level $n+1$. We denote this averaging operator as $A(\tilde{u}^{n+1})$.

*Received by the editors April 21, 2006; accepted for publication (in revised form) April 24, 2007; published electronically May 9, 2008. This research supported by ARO grant W911NF-04-1-0291, NSF grant DMS-0510345, and AFOSR grant FA9550-05-1-0123.
http://www.siam.org/journals/sinum/46-5/65791.html
†Division of Applied Mathematics, Brown University, Providence, RI 02912 (jqiu@dam.brown.edu, shu@dam.brown.edu).

The Godunov-type scheme, using the notations introduced above, can be described abstractly as

$$\begin{cases} \bar{u}^0 = A(u_0), \\ \bar{u}^{n+1} = A \circ S_{\Delta t} \circ Re\,(\bar{u}^n), \qquad n = 0, 1, \ldots, N-1. \end{cases} \tag{1.2}$$

The purpose of the reconstruction step is to increase the formal accuracy of the scheme. It is usually achieved by using cell averages of several neighboring cells to obtain a high-order reconstructed polynomial. In order to be more specific, let us first introduce some notations. For example, in the one-dimensional case, the mesh is given by $I_j = [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}]$. The center of the cell I_j is given by $x_j = \frac{1}{2}(x_{j-\frac{1}{2}} + x_{j+\frac{1}{2}})$, and the size of the cell I_j is denoted by $\Delta x = x_{j+\frac{1}{2}} - x_{j-\frac{1}{2}}$, which is considered to be a constant (uniform mesh) in this paper for simplicity of notations. The cell average of a function $u(x)$ in the cell I_j is defined as

$$\bar{u}_j = A(u)_j = \frac{1}{\Delta x} \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} u(x)\,dx.$$

A second-order reconstruction would be a piecewise linear function, and the linear reconstruction $p_j(x)$ in cell I_j could be the unique linear polynomial satisfying

$$\frac{1}{\Delta x} \int_{x_{j-\frac{3}{2}}}^{x_{j-\frac{1}{2}}} p_j(x)\,dx = \bar{u}_{j-1}, \qquad \frac{1}{\Delta x} \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} p_j(x)\,dx = \bar{u}_j,$$

namely, $p_j(x)$ agrees with the given cell averages of $u(x)$ over the two cells $\{I_{j-1}, I_j\}$, which is called the *stencil* of the reconstruction $p_j(x)$. Usually, we require that the stencil for the reconstruction in cell I_j contain at least I_j itself. Additional cells in the stencil can be chosen according to symmetry or stability (nonoscillation) considerations. For example, the stencil for the linear polynomial $p_j(x)$ above could also be $\{I_j, I_{j+1}\}$, which would of course produce a different reconstructed linear polynomial.

The reconstruction is called (formally) kth-order accurate, namely,

$$|Re(\bar{u})(x) - u(x)| \le C\Delta x^k, \qquad x \in I_j, \tag{1.3}$$

if the reconstructed polynomial in I_j is of $(k-1)$th degree with a stencil of k cells over which the function $u(x)$ is smooth. Here and below, C (or $\tilde{C}$) will denote a generic constant independent of the mesh sizes Δx and Δt, which may have different values in different locations. For more details of second- and higher-order reconstructions, we refer the reader to [14, 21].

A reconstruction is called total variation bounded (TVB) if

$$TV(Re(\bar{u})) \le (1 + C\Delta x) TV(\bar{u}), \tag{1.4}$$

and it is called total variation diminishing (TVD) if (1.4) holds with $C = 0$. Here the total variation of a piecewise smooth function $g(x)$ is defined as usual,

$$TV(g) = \int |g'(x)|\,dx,$$

and the total variation of $\bar{u}$ in (1.4) is defined with $\bar{u}$ considered as a piecewise constant function.

Since the exact evolution operator $S_{\Delta t}$ and the averaging operator A do not increase the total variation of a function, the total variation of the final solution $\bar{u}^N$ of the Godunov-type scheme at a fixed time $T = N\Delta t$ is bounded by a constant C independent of the mesh sizes Δx and Δt, when both of them go to zero, if the reconstruction is TVB. This constant C is the total variation of the initial data if the reconstruction is TVD.

By compactness, the sequence of numerical solutions of a Godunov-type scheme with a TVB reconstruction, when the mesh sizes Δx and Δt both go to zero, has a convergent subsequence in L^1. Since the scheme is conservative, the limit of such a convergent subsequence is a weak solution to (1.1). However, because of the nonuniqueness of weak solutions, we cannot conclude the convergence of the original sequence of numerical solutions, nor can we claim that the limit of a convergent subsequence is the physically relevant entropy solution. Typically, we would need an additional cell entropy inequality, e.g., [18], or a wavewise entropy inequality, e.g., [24], to imply such convergence. Such cell or wavewise entropy inequalities are very difficult to obtain for high-order schemes, especially for nonconvex conservation laws for which such entropy inequalities are needed for all entropies rather than just for one of them. As far as we know, within self-similar finite difference or finite volume framework, such cell or wavewise entropy inequalities have been obtained only for second-order MUSCL-type schemes for scalar one-dimensional convex conservation laws [18, 24], and for a special second-order scheme for general conservation laws which evolves cell averages and cell slopes independently [2]. These schemes are all explicit schemes with the usual CFL restriction $\Delta t = O(\Delta x)$. It seems that for such situations a cell or wavewise entropy inequality is necessary for convergence towards entropy solutions; see, for example, [18] in which it is proven that finite volume schemes (those which evolve only cell averages and rely on reconstruction to achieve higher-order accuracy) which satisfy all entropy conditions are necessarily only first-order accurate, [1] in which an entropy condition can be obtained only when a first-order antidiffusive scheme is suitably modified and enhanced, and [23] in which a counterexample is shown to demonstrate that a second-order TVD Godunov-type scheme can converge to an entropy violating solution for nonconvex conservation laws. We remark that cell entropy inequality for the square entropy can be obtained in more general situations (arbitrary order of accuracy, general triangulation in multispace dimensions, arbitrary scalar conservation laws) for the semidiscrete discontinuous Galerkin method [9], but this is a totally different class of schemes in which no reconstruction is involved and the complete polynomial within a cell is evolved in time to obtain high-order accuracy.

We now mention a few papers in the literature which directly motivated our work in this paper. Lions and Souganidis [16] proved the convergence of an implicit MUSCL scheme to the entropy solution of strictly convex scalar one-dimensional Hamilton–Jacobi equation and conservation law under a large time step assumption, namely, $\Delta x = o(\Delta t)$. For one-dimensional convex Hamilton–Jacobi equations, which are very closely related to conservation laws (1.1), Ferretti [5] proved the convergence of a semi-Lagrangian scheme to the viscosity solution (the solution corresponding to the entropy solution for conservation laws), again under a large time step assumption $\Delta x = o(\Delta t)$. This convergence holds for quite general high-order reconstructions, e.g., the ENO reconstruction [8] up to fifth order and WENO reconstruction [10] up to ninth order; see [5, 3]. The message in [16, 5, 3] seems to be that it is easier to prove convergence for large time step ($\Delta x = o(\Delta t)$) schemes. Of course, such a choice of the time step violates the traditional CFL condition for local explicit schemes. One would

need to take either implicit time discretization, as in [16], or semi-Lagrangian time discretization, namely following the characteristics, as in [5, 3]. In related work, we mention the large time step generalization of the first-order accurate Godunov method in [13], which involves an approximate but not exact entropy solution evolution. We also mention the large time step semi-Lagrangian method in [4], for which an entropy condition can be proven for convex scalar one-dimensional conservation laws.

In this paper we prove that, indeed, the large time step assumption $\Delta x = o(\Delta t)$ allows us to bypass the verification of cell or wavewise entropy conditions and prove the convergence towards the entropy solution of Godunov-type schemes solely under the assumption of a TVB reconstruction, for general scalar conservation laws (1.1). Since self-similar TVB reconstruction is available in one dimension to third-order accuracy [15], we then have third-order accurate Godunov-type schemes in one dimension which can be proven convergent towards the entropy solution. Of course, to actually implement the large time step Godunov-type schemes is not trivial. We discuss an effective implementation, similar to the semi-Lagrangian scheme in [5, 3] for Hamilton–Jacobi equations, for one-dimensional scalar convex conservations.

The paper is organized as follows. Section 2 contains our main convergence result and its proof. In section 3, the TVB property of the numerical solution is considered. Examples of high-order TVB reconstruction are given. Section 4 discusses an effective implementation of the large time step Godunov-type scheme for one-dimensional scalar convex conservation laws and provides numerical results verifying the performance of such schemes. In order to demonstrate the generality of this approach, we consider in section 5 a Godunov-like scheme using a third-order TVD reconstruction due to Sanders [19], which uses both cell averages and point values at cell boundaries for the reconstruction and does not exactly fit our large time step Godunov-type scheme framework, but the convergence can be obtained along similar lines. Concluding remarks are given in section 6.

2. Convergence of large time step Godunov-type schemes. We consider the Godunov-type scheme (1.2) for solving the scalar conservation law (1.1) under the following three assumptions:

1. the reconstruction operator Re satisfies

$$\|\bar{u} - Re(\bar{u})\|_{L^1} \le C\,\Delta x TV(\bar{u}) \tag{2.1}$$

   for any piecewise constant function $\bar{u}$;
2. the total variation of the numerical solution is uniformly bounded,

$$TV(\bar{u}^n) \le C \tag{2.2}$$

   for all mesh sizes Δx and Δt satisfying $n\Delta t \le T$;
3. large time step condition is

$$\lim_{\Delta x \to 0} \frac{\Delta x}{\Delta t} = 0. \tag{2.3}$$

We have the following convergence result under these assumptions.

THEOREM 2.1. *The numerical solution of the Godunov-type scheme* (1.2) *for solving the scalar conservation law* (1.1) *converges in* L^1 *to the entropy solution of* (1.1) *when the mesh size* Δx *goes to zero if the three assumptions* (2.1), (2.2), *and* (2.3) *are satisfied.*

Proof. We denote the exact entropy solution of (1.1) at time level t^n by $u(\cdot, t^n)$ and the L^1 difference between this exact solution and the reconstructed numerical solution $u^n = Re(\bar{u}^n)$ at time level t^n by e^n,

$$e^n = \|u(\cdot, t^n) - Re(\bar{u}^n)\|_{L^1}. \tag{2.4}$$

First, we notice that

$$\begin{aligned}
\|A(u(\cdot, t^n)) - \bar{u}^n\|_{L^1} &= \|A(u(\cdot, t^n)) - A \circ S_{\Delta t} \circ Re\,(\bar{u}^{n-1})\|_{L^1} \\
&= \|A\,\left(u(\cdot, t^n) - S_{\Delta t} \circ Re\,(\bar{u}^{n-1})\right)\|_{L^1} \\
&\le \|u(\cdot, t^n) - S_{\Delta t} \circ Re\,(\bar{u}^{n-1})\|_{L^1} \\
&= \|S_{\Delta t}\,(u(\cdot, t^{n-1})) - S_{\Delta t} \circ Re\,(\bar{u}^{n-1})\|_{L^1} \\
&\le \|u(\cdot, t^{n-1}) - Re\,(\bar{u}^{n-1})\|_{L^1} = e^{n-1}, \qquad (2.5)
\end{aligned}$$

where the third inequality is due to the fact that the averaging operator A does not increase the L^1 norm of a function, and the fifth inequality is due to the L^1 contraction property of the exact entropy solution operator $S_{\Delta t}$ of (1.1); see [12]. Also for a bounded variation function w, it is easy to verify that

$$\|w - A(w)\|_{L^1} \le C\Delta x TV(w). \tag{2.6}$$

In fact, for the one-dimensional case

$$\begin{aligned}
\|w - A(w)\|_{L^1} &= \sum_j \int_{I_j} |w(x) - \bar{w}_j| dx \\
&\le \sum_j \Delta x \left(\sup_{x \in I_j} w(x) - \inf_{x \in I_j} w(x) \right) \le \Delta x TV(w).
\end{aligned}$$

The proof for the multidimensional case is similar.

We can now proceed to prove the desired convergence,

$$\begin{aligned}
e^n &= \|u(\cdot, t^n) - Re(\bar{u}^n)\|_{L^1} \\
&\le \|u(\cdot, t^n) - A(u(\cdot, t^n))\|_{L^1} + \|A(u(\cdot, t^n)) - \bar{u}^n\|_{L^1} + \|\bar{u}^n - Re(\bar{u}^n)\|_{L^1} \\
&\le C\Delta x TV(u(\cdot, t^n)) + e^{n-1} + C\Delta x TV(\bar{u}^n) \\
&\le \tilde{C}\Delta x + e^{n-1} \\
&\le \cdots \\
&\le \tilde{C} n \Delta x + e^0 \\
&= \tilde{C} T \frac{\Delta x}{\Delta t} + e^0 \to 0 \qquad \text{as} \quad \Delta x \to 0, \qquad (2.7)
\end{aligned}$$

where for the third inequality we have used (2.6), (2.5), and the assumption (2.1), for the fourth inequality we have used the fact that the exact entropy solution $u(\cdot, t^n)$ is of bounded variation and the assumption (2.2), and for the last limit we have used the large time step assumption (2.3). ∎

Next, we prove that the assumption (2.1) is satisfied by all commonly used reconstructions, including the high-order ENO and WENO reconstructions [8, 10]. As before, for simplicity we divide the computational domains into uniform cells (in one dimension these are nonoverlapping intervals of the same length Δx covering the whole

computational domain). We require the reconstruction to be a piecewise polynomial function; namely, it is a polynomial over each cell I_j. The reconstructed polynomial in I_j is uniquely determined by requiring that it have the same cell average as the given one over each cell in a stencil, which is a set of finitely many cells (the number of cells in this set is fixed and does not depend on the mesh size Δx) and contains the given cell I_j. For such reconstructions we have the following result.

LEMMA 2.2. *The assumption* (2.1) *is satisfied by any piecewise polynomial reconstruction with a local stencil containing finitely many cells.*

Proof. We prove this lemma for the one-dimensional case. The proof for the multidimensional case is similar. Assume the reconstruction stencil for cell I_j is $I_{j-l(j)}, \ldots, I_{j+r(j)}$. Since the stencil contains I_j and has only finitely many cells, we have $0 \le l(j), r(j) \le M$ for a fixed integer M. The reconstructed polynomial in cell I_j can be written as

$$p(x) = \sum_{i=-l(j)}^{r(j)} p_i\left(\frac{x - x_j}{\Delta x}\right) \bar{u}_{j+i}, \tag{2.8}$$

where $p_i(\xi)$ are polynomials independent of Δx and the location I_j; see, for example, [21] for the explicit expression of $p_i(\xi)$. We therefore have $|p_i(\xi)| \le C$ for some constant C when ξ is in the relevant range $|\xi| \le \frac{1}{2}$. Also, by consistency (i.e., the reconstruction is exact if $\bar{u}_j \equiv 1$), we have

$$\sum_{i=-l(j)}^{r(j)} p_i\left(\frac{x - x_j}{\Delta x}\right) = 1.$$

Hence

$$p(x) - \bar{u}_j = \sum_{i=-l(j)}^{r(j)} p_i\left(\frac{x - x_j}{\Delta x}\right) (\bar{u}_{j+i} - \bar{u}_j).$$

We now have

$$\begin{aligned}
\|\bar{u} - Re(\bar{u})\|_{L^1} &= \sum_j \|\bar{u} - Re(\bar{u})\|_{L^1(I_j)} \\
&= \sum_j \int_{I_j} \left| \sum_{i=-l(j)}^{r(j)} p_i\left(\frac{x - x_j}{\Delta x}\right) (\bar{u}_{j+i} - \bar{u}_j) \right| dx \\
&\le CM(2M+1)\Delta x \sum_j |\bar{u}_{j+1} - \bar{u}_j| \\
&= \tilde{C}\Delta x TV(\bar{u}). \qquad \square
\end{aligned}$$

Clearly, Lemma 2.2 covers both the linear reconstructions, for which the reconstruction stencil relative to I_j is fixed, i.e., $l(j)$ and $r(j)$ do not depend on j, as well as the ENO reconstruction [8], for which the reconstruction stencil depends on the local smoothness of the data $\bar{u}$. Since the WENO reconstructions [10] are linear combinations of fixed stencil reconstructions with bounded weights (linear combination coefficients, which depend on the local smoothness of the data $\bar{u}$), Lemma 2.2 also covers WENO reconstructions. Notice that Lemma 2.2 is valid for reconstructions of arbitrarily high-order accuracy.

Remark 2.3. By combining the results in Theorem 2.1 and Lemma 2.2, we conclude that, if the reconstruction operator Re is based on piecewise polynomial reconstruction with a local stencil containing finitely many cells, then the numerical solution of the Godunov-type scheme (1.2) for solving the scalar conservation law (1.1) converges in L^1 to the entropy solution of (1.1) when the mesh size Δx goes to zero, provided that the two assumptions (2.2) and (2.3) are satisfied.

Remark 2.4. The proof (2.7) for Theorem 2.1 gives a pessimistic error bound of $O(\Delta x/\Delta t)$ for the scheme. While there is reason to believe that this pessimistic error bound could actually be sharp for general bounded variation solutions, e.g., for the example given in [23] where the Godunov-type scheme converges to an entropy violating solution for $\Delta t = O(\Delta x)$, for smooth solutions the convergence rate is expected to be much better. To look at the local truncation error, assume we start from the cell averages $\bar{u}_j^n = Au(\cdot, t^n)|_{I_j}$ of the exact solution $u(\cdot, t)$ of the PDE and use our scheme to advance for one time step to obtain the numerical solution $\bar{u}^{n+1}$. The error at the end of this one time step is estimated by

$$\begin{aligned}\|Au(\cdot, t^{n+1}) - \bar{u}^{n+1}\|_{L^1} &= \|A \circ S_{\Delta t}(u(\cdot, t^n)) - A \circ S_{\Delta t} \circ Re(\bar{u}^n)\|_{L^1} \\ &\le \|S_{\Delta t}(u(\cdot, t^n)) - S_{\Delta t} \circ Re(\bar{u}^n)\|_{L^1} \\ &\le \|u(\cdot, t^n) - Re(\bar{u}^n)\|_{L^1}.\end{aligned}$$

That is, the L^1 error at next time step is no larger than the reconstruction error in L^1. For general bounded variation solutions this reconstruction error is expected to be $O(\Delta x)$, by combining Lemma 2.2 and (2.6). Therefore, the local truncation error, which is this error divided by Δt, is at most $O(\Delta x/\Delta t)$, which would go to zero by our choice of a large time step Δt. For a smooth solution u, a more detailed analysis indicates that the local truncation error is of the order $O(\Delta x^k)$ for a kth-order reconstruction using piecewise polynomials of degree $k-1$. We include this analysis in the appendix, since it is quite technical but does reveal the somewhat surprising result that the local truncation error is independent of the time step Δt.

3. TVB property of the numerical solution and examples of the TVB reconstruction operator. The result of the previous section indicates that, apart from the large time step assumption (2.3), which causes difficulties in the actual implementation of the scheme and will be discussed in more detail in next section, we would only need to have a TVB numerical solution to imply convergence. If the reconstruction operator Re is TVB, i.e., it satisfies (1.4), then we easily have

$$\begin{aligned}TV(\bar{u}^n) &= TV(A \circ S_{\Delta t} \circ Re\,(\bar{u}^{n-1})) \\ &\le TV(Re(\bar{u}^{n-1})) \\ &\le (1 + C\Delta x) TV(\bar{u}^{n-1}) \\ &\le \cdots \\ &\le (1 + C\Delta x)^n\, TV(\bar{u}^0) \\ &\le (1 + C\Delta t)^n\, TV(u_0) \\ &\le e^{CT} TV(u_0)\end{aligned}$$

for any n and Δt such that $n\Delta t \le T$. Here we have used the fact that neither the averaging operator A nor the exact entropy solution evolution operator $S_{\Delta t}$ increases the total variation of a function for the second inequality, and the assumption (2.3) for the sixth inequality (in fact, here we do not need the large time step assumption; we

just need the assumption that $\Delta x \leq C\Delta t$). We therefore only need to verify the total variation boundedness of the reconstruction (1.4) in order to guarantee convergence towards entropy solutions for a large time step Godunov-type scheme.

Next we discuss two examples of TVB reconstructions (1.4). We concentrate our attention on self-similar reconstructions, namely, reconstructions which do not depend explicitly on the mesh size Δx (see the example of the reconstruction (2.8) in which the polynomials $p_i(\xi)$ do not depend on Δx [21]). In fact, almost all known self-similar TVB reconstructions are actually TVD. It is very difficult to obtain self-similar TVB reconstructions in two or higher dimensions [6]; hence we restrict our attention to one dimension.

1. *Second-order MUSCL reconstruction.* This reconstruction [22] is piecewise linear and is second-order accurate except at extrema, where it degenerates to first-order accuracy. The reconstructed polynomial in cell I_j is given by

$$p(x) = \bar{u}_j + s_j\,(x - x_j)\,, \tag{3.1}$$

where the slope s_j is given by

$$s_j = m\left(\frac{\bar{u}_{j+1} - \bar{u}_j}{\Delta x}, \frac{\bar{u}_j - \bar{u}_{j-1}}{\Delta x}\right)$$

with the usual definition of the *minmod* function m (see [7])

$$m(a,b) = \frac{\operatorname{sign}(a) + \operatorname{sign}(b)}{2}\min(|a|,|b|).$$

In words, the slope is taken as the smaller one (in magnitude) between those of the two reconstruction linear polynomials based on the stencils $\{I_{j-1}, I_j\}$ and $\{I_j, I_{j+1}\}$, unless these two slopes are of opposite sign (i.e., we are at an extremum of the cell average $\bar{u}$), in which case the slope is taken as zero. This reconstruction is clearly TVD.

2. *Third-order nonoscillatory central reconstruction.* This reconstruction [15] is piecewise quadratic and is third-order accurate except at extrema, where it degenerates to second-order accuracy. The reconstructed polynomial in cell I_j is based on $q_j(x)$, the reconstructed quadratic polynomial over the stencil $\{I_{j-1}, I_j, I_{j+1}\}$, with modification to satisfy a nonoscillatory property (no new extremum) and a maximum principle. The first step is to obtain the nonoscillatory property and is given by

$$p_j^1(x) = \bar{u}_j + \theta_j\,(q_j(x) - \bar{u}_j)\,, \tag{3.2}$$

where the limiter θ_j is defined as

$$\theta_j = \begin{cases} \min\left(\dfrac{M_{j+\frac{1}{2}} - \bar{u}_j}{M_j - \bar{u}_j}, \dfrac{m_{j-\frac{1}{2}} - \bar{u}_j}{m_j - \bar{u}_j}, 1\right) & \text{if } \bar{u}_{j-1} < \bar{u}_j < \bar{u}_{j+1}, \\ \min\left(\dfrac{M_{j-\frac{1}{2}} - \bar{u}_j}{M_j - \bar{u}_j}, \dfrac{m_{j+\frac{1}{2}} - \bar{u}_j}{m_j - \bar{u}_j}, 1\right) & \text{if } \bar{u}_{j-1} > \bar{u}_j > \bar{u}_{j+1}, \\ 1 & \text{if } (\bar{u}_{j+1} - \bar{u}_j)\,(\bar{u}_j - \bar{u}_{j-1}) \leq 0, \end{cases}$$

where

$$M_j = \max_{x \in I_j} q_j(x), \qquad m_j = \min_{x \in I_j} q_j(x),$$

$$M_{j\pm\frac{1}{2}} = \max\left(\frac{1}{2}(\bar{u}_j + \bar{u}_{j\pm1}), q_{j\pm1}(x_{j\pm\frac{1}{2}})\right),$$
$$m_{j\pm\frac{1}{2}} = \min\left(\frac{1}{2}(\bar{u}_j + \bar{u}_{j\pm1}), q_{j\pm1}(x_{j\pm\frac{1}{2}})\right).$$

The second step is to further modify $p_j^1(\mathrm{x})$ to satisfy a maximum principle. If we denote

$$M = \max_{x\in R} u_0(x), \qquad m = \min_{x\in R} u_0(x)$$

and

$$M_j^1 = \max_{x\in I_j} p_j^1(x), \qquad m_j^1 = \min_{x\in I_j} p_j^1(x),$$

then the final reconstructed polynomial is given by

(3.3)
$$p_j(x) = \begin{cases} \varepsilon^{(1)}\left(p_j^1(x) - M_j^1\right) + M & \text{when} \quad M_j^1 > M,\ m_j^1 \geq m, \\ \varepsilon^{(2)}\left(p_j^1(x) - m_j^1\right) + m & \text{when} \quad M_j^1 \leq M,\ m_j^1 < m, \\ \varepsilon^{(1)}\left(p_j^1(x) - M_j^1\right) + M & \text{when} \quad \varepsilon^{(1)} < \varepsilon^{(2)},\ M_j^1 > M, m_j^1 < m, \\ \varepsilon^{(2)}\left(p_j^1(x) - m_j^1\right) + m & \text{when} \quad \varepsilon^{(1)} \geq \varepsilon^{(2)},\ M_j^1 > M, m_j^1 < m, \\ p_j^{(1)}(x) & \text{when} \quad M_j^1 \leq M,\ m_j^1 \geq m, \end{cases}$$

with

$$\varepsilon^{(1)} = \frac{M - \bar{u}_j}{M_j^1 - \bar{u}_j}, \qquad \varepsilon^{(2)} = \frac{m - \bar{u}_j}{m_j^1 - \bar{u}_j}.$$

It is proven in [15] that the reconstruction (3.3) is TVB. Moreover, we have

$$\|p_j^1(x) - \bar{u}_j\|_{L^1} = \theta_j \|q_j(x) - \bar{u}_j\|_{L^1}$$

and

$$\|p_j(x) - \bar{u}_j\|_{L^1} = \varepsilon^{(i)}\|p_j^1(x) - \bar{u}_j\|_{L^1}, \quad i = 1, 2, \quad \text{or}$$
$$\|p_j(x) - \bar{u}_j\|_{L^1} = \|p_j^1(x) - \bar{u}_j\|_{L^1}$$

with $0 \leq \theta_j \leq 1$ and $0 \leq \varepsilon^{(i)} \leq 1$. Therefore,

$$\|p_j(x) - \bar{u}_j\|_{L^1} \leq \|p_j^1(x) - \bar{u}_j\|_{L^1} \leq \|q_j(x) - \bar{u}_j\|_{L^1}.$$

Hence clearly assumption (2.1) is also satisfied by this reconstruction, using Lemma 2.2.

Other possible TVB reconstructions in one dimension include the very high-order TVD reconstructions in [17] (however, these reconstructions degenerate to first-order accuracy at smooth extrema), or the high-order TVB reconstructions in [20] (however, the reconstructions are no longer self-similar).

4. Implementation of the scheme and numerical examples for one-dimensional scalar convex conservation laws. In this section, we discuss a practical implementation of the large time step Godunov-type scheme discussed in the previous section, for one-dimensional scalar convex conservation laws, using a

semi-Lagrangian-type scheme similar to that used in [5, 3] for Hamilton–Jacobi equations. However, we emphasize that our implementation in this section is exactly the Godunov-type scheme discussed in the convergence analysis in previous sections. No extra error (other than computer roundoff errors) has been introduced. In particular, the scheme is conservative, unlike many semi-Lagrangian-type schemes.

Recall that there are three stages to evolve the cell averages $\bar{u}^n$ at time level n to $\bar{u}^{n+1}$ at time level $n+1$, that is, reconstruction (Re), exact evolution ($S_{\Delta t}$), and averaging (A).

1. Reconstruction: reconstruct a piecewise polynomial function $u^n(x)$ from the cell averages $\bar{u}^n$. Specifically, for second-order TVD or third-order TVB reconstruction, the implementation follows the procedure described in (3.1) and (3.2).
2. Evolution: evolve $u^n(x)$ by (1.1) exactly for a time step Δt to obtain the solution $\tilde{u}^{n+1}$. In general, it is very difficult to numerically obtain $\tilde{u}^{n+1}$.
3. Averaging: average $\tilde{u}^{n+1}$ over each cell I_j to obtain $\bar{u}^{n+1}$.

Fortunately, we do not need the detailed information about $\tilde{u}^{n+1}$ in the second stage above (which would be very difficult to obtain exactly), but only its cell averages $\bar{u}^{n+1}$ over each cell. Therefore, we will combine the second stage ($S_{\Delta t}$) and third stage (A) together in our implementation.

We make the following observations on the x-t plane (see Figure 1).

Observation 1. At each of the cell boundary points at time level $n+1$, say $(x_{j+\frac{1}{2}}, t^{n+1})$, there exists a backward characteristic line, denoted as $\Gamma_{j+\frac{1}{2}}$, with the foot located on time level n at $y_{j+\frac{1}{2}}$. These backward characteristic lines, $\{\Gamma_{j+\frac{1}{2}}\}$, do not cross each other in the x-t plane, which is a property of the entropy solution. Along each of these characteristic lines, the entropy solution remains a constant. For example, the solution along $\Gamma_{j+\frac{1}{2}}$ is $u^n(y_{j+\frac{1}{2}})$ if the function $u^n(x)$ is continuous at $x = y_{j+\frac{1}{2}}$. The slope of the characteristic line is also known. For example, the slope of $\Gamma_{j+\frac{1}{2}}$ is $f'(u^n(y_{j+\frac{1}{2}}))$.

Observation 2. For each cell I_j at time level $n+1$, there is a trapezoid region Ω_j (see Figure 1) with four vertices $(x_{j-\frac{1}{2}}, t^{n+1})$, $(x_{j+\frac{1}{2}}, t^{n+1})$, $(y_{j-\frac{1}{2}}, t^n)$, and $(y_{j+\frac{1}{2}}, t^n)$. These trapezoid regions are mutually disjoint and they together cover the whole computational domain times the interval $[t^n, t^{n+1}]$.

Based on the two observations listed above, for each cell I_j, we apply the integral form of conservation laws over the region Ω_j,

$$\int_{\Omega_j} (u_t + f(u)_x) dx dt = 0. \tag{4.1}$$

By the divergence theorem and Observation 1, the left-hand side of (4.1) can be written in the following explicit form:

$$\begin{aligned} \int_{\Omega_j} (u_t + f(u)_x) dx dt &= \int_{\partial\Omega_j} (u n_t + f(u) n_x) ds \\ &= \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} u dx + \int_{\Gamma_{j+\frac{1}{2}}} (u n_t + f(u) n_x) ds \end{aligned} \tag{4.2}$$

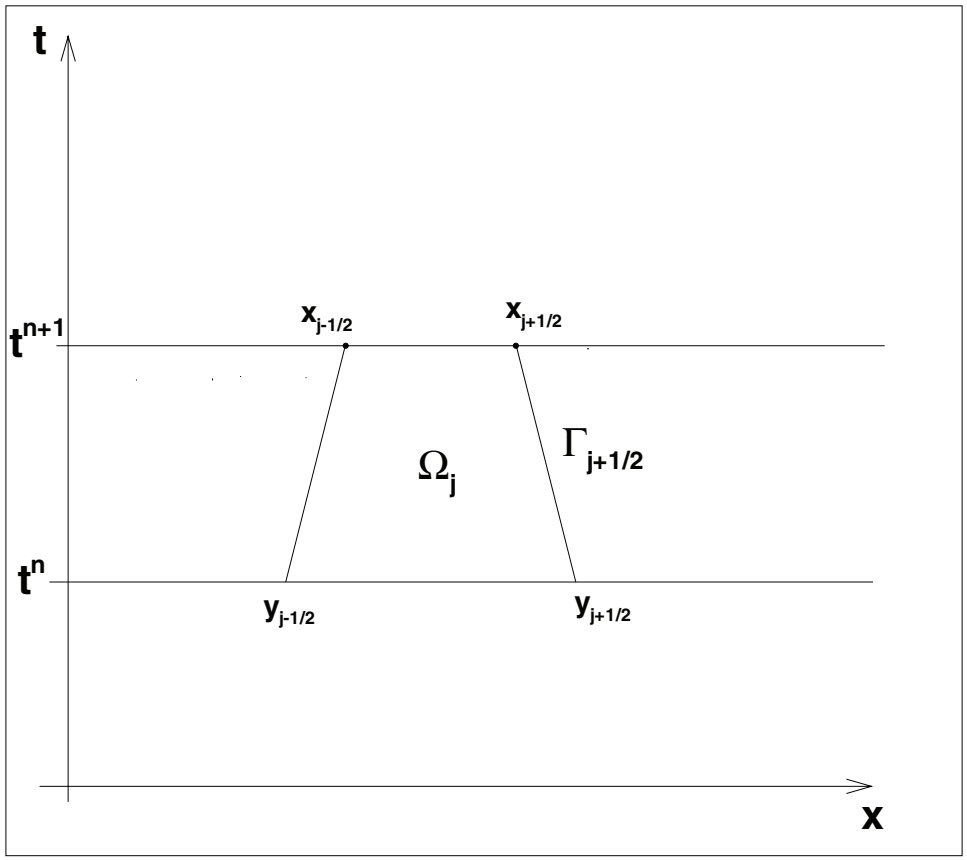


FIG. 1. *The trapezoid region Ω_j in the x-t plane.*

$$-\Delta x\bar{u}_j^{n+1} + \int_{\Gamma_{j-\frac{1}{2}}} (un_t + f(u)n_x)ds \tag{4.3}$$

$$= \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} u^n(x)dx - \left(\Delta t f(u_{j+\frac{1}{2}}) + (y_{j+\frac{1}{2}} - x_{j+\frac{1}{2}})u_{j+\frac{1}{2}}\right)$$

$$-\Delta x\bar{u}_j^{n+1} + \left(\Delta t f(u_{j-\frac{1}{2}}) + (y_{j-\frac{1}{2}} - x_{j-\frac{1}{2}})u_{j-\frac{1}{2}}\right), \tag{4.4}$$

where (n_x, n_t) is the unit outward normal vector for the boundaries of the region Ω_j. Hence, we can obtain the following formula for computing $\bar{u}_j^{n+1}$:

$$\Delta x\bar{u}_j^{n+1} = \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} u^n(x)dx - \left(\Delta t f\left(u_{j+\frac{1}{2}}\right) + \left(y_{j+\frac{1}{2}} - x_{j+\frac{1}{2}}\right)u_{j+\frac{1}{2}}\right)$$

$$+ \left(\Delta t f\left(u_{j-\frac{1}{2}}\right) + \left(y_{j-\frac{1}{2}} - x_{j-\frac{1}{2}}\right)u_{j-\frac{1}{2}}\right). \tag{4.5}$$

This is the formula we use in coding up the algorithm. Clearly, if we can find the backward characteristic line $\Gamma_{j+\frac{1}{2}}$, hence also the (constant) value of the entropy solution $u_{j+\frac{1}{2}}$ along it, for each point $(x_{j+\frac{1}{2}}, t^{n+1})$, then the Godunov-type scheme discussed in the previous sections in the convergence analysis is implemented exactly, only subject to roundoff errors.

Now, the only difficulty left for the implementation of the Godunov-type scheme is to find the backward characteristic line $\Gamma_{j+\frac{1}{2}}$ for each target point $(x_{j+\frac{1}{2}}, t^{n+1})$. Notice that there might be multiple forward characteristic lines from $u^n(x)$ in the previous time level which cross at the target point. Even though only one of those characteristic lines (or at most two if the solution happens to be discontinuous at $(x_{j+\frac{1}{2}}, t^{n+1})$ should be correct and the remaining forward characteristic lines will have crossed other characteristic lines before the time t^{n+1} and should be ignored as they are no longer valid after such crossing, it seems difficult to numerically implement the finding of this correct characteristic line with low computational cost.

We propose the following method to obtain the backward characteristic line $\Gamma_{j+\frac{1}{2}}$ for $(x_{j+\frac{1}{2}}, t^{n+1})$.

(a) At time level n, for each cell I_i, we determine its domain of influence J_i, and for each cell boundary point $x_{i+\frac{1}{2}}$, we determine its domain of influence $\tilde{J}_{i+\frac{1}{2}}$.
- If the function $u^n(x)$ is monotone in $[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$, then J_i is determined by the two forward characteristics from the cell boundaries $(x^{\pm}_{i\mp\frac{1}{2}}, t^n)$. Specifically, from $(x^{\pm}_{i\mp\frac{1}{2}}, t^n)$, the forward characteristic lines cross $t = t^{n+1}$ at $\xi^{\pm}_{i\mp\frac{1}{2}}$, respectively. If $\xi^{+}_{i-\frac{1}{2}} \le \xi^{-}_{i+\frac{1}{2}}$, then J_i is simply the interval $[\xi^{+}_{i-\frac{1}{2}}, \xi^{-}_{i+\frac{1}{2}}]$. Otherwise $J_i = \emptyset$, since the forward characteristics have crossed each other before they reach t^{n+1}. If the function $u^n(x)$ is not monotone in $[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$, then we divide the interval $[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$ into a union of several nonoverlapping subintervals; in each of them the function $u^n(x)$ is monotone. Since $u^n(x)$ is a reconstructed polynomial in $[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$, we have at most k such subintervals, where k is the polynomial degree. For each such subinterval we use the previous technique to determine its domain of influence J_i^{ℓ}. J_i is then the union of all these J_i^{ℓ}.
- $\tilde{J}_{i+\frac{1}{2}}$ is determined by the two forward characteristics from the cell boundary $(x^{\mp}_{i+\frac{1}{2}}, t^n)$. Specifically, from $(x^{\mp}_{i+\frac{1}{2}}, t^n)$, the forward characteristic lines cross $t = t^{n+1}$ at $\xi^{\mp}_{i+\frac{1}{2}}$, respectively. If $\xi^{-}_{i+\frac{1}{2}} \le \xi^{+}_{i+\frac{1}{2}}$, then $\tilde{J}_{i+\frac{1}{2}}$ is simply the interval $[\xi^{-}_{i+\frac{1}{2}}, \xi^{+}_{i+\frac{1}{2}}]$. Otherwise $\tilde{J}_{i+\frac{1}{2}} = \emptyset$, since the boundary point $(x_{i+\frac{1}{2}}, t^n)$ generates a shock.

(b) For each of the target points $(x_{j+\frac{1}{2}}, t^{n+1})$, find the set of candidate intervals, denoted as $C_{j+\frac{1}{2}}$, and the set of candidate cell boundary points, denoted as $\tilde{C}_{j+\frac{1}{2}}$, such that

$$C_{j+\frac{1}{2}} = \left\{ i : \ x_{j+\frac{1}{2}} \in J_i \right\}, \qquad \tilde{C}_{j+\frac{1}{2}} = \left\{ i : \ x_{j+\frac{1}{2}} \in \tilde{J}_{i+\frac{1}{2}} \right\}. \tag{4.6}$$

In words, a candidate interval or a candidate cell boundary point should be the one with the target point contained in its domain of influence. Because of the convexity of $f(u)$ and hence the monotonicity of $f'(u)$, the sets $C_{j+\frac{1}{2}}$ and $\tilde{C}_{j+\frac{1}{2}}$ include all candidate intervals and cell boundary points at time level t^n which contain forward characteristic lines crossing the target point $(x_{j+\frac{1}{2}}, t^{n+1})$. Furthermore, we can use the minimum and maximum values of $f'(u^n)$ to estimate the largest possible region of dependency for the point $(x_{j+\frac{1}{2}}, t^{n+1})$ at the time level t^n, and search for candidates of the sets $C_{j+\frac{1}{2}}$

and $\tilde{C}_{j+\frac{1}{2}}$ only within this region of dependency to save computational cost.

(c) Find the candidate trace-back points within the sets $C_{j+\frac{1}{2}}$ and $\tilde{C}_{j+\frac{1}{2}}$. For each J_i (or J_i^ℓ if J_i is a union of at least two of them) with $i \in C_{j+\frac{1}{2}}$, we can find a unique forward characteristic line from (x_i^*, t^n), for some point $x_i^* \in I_i$, such that it crosses the target point $(x_{j+\frac{1}{2}}, t^{n+1})$. The solution $u^* = u^n(x_i^*)$ is a constant along this line, which can be found through the implicit characteristic equation

$$u^* = u^n\left(x_{j+\frac{1}{2}} - f'(u^*)\,\Delta t\right). \tag{4.7}$$

For the Burgers equation with piecewise polynomial reconstruction of degree no larger than 4, u^* can be obtained analytically. For the general situation a Newton iteration might be needed to find u^*. Likewise, for each $\tilde{J}_{i+\frac{1}{2}}$ with $i \in \tilde{C}_{j+\frac{1}{2}}$, we can find a unique forward characteristic line from $(x_{i+\frac{1}{2}}, t^n)$ with a solution value u^* between $u^n(x^-_{i+\frac{1}{2}})$ and $u^n(x^+_{i+\frac{1}{2}})$, such that it crosses the target point $(x_{j+\frac{1}{2}}, t^{n+1})$. The solution u^* can be found by the characteristic equation which is simpler than (4.7),

$$f'(u^*) = \frac{x_{j+\frac{1}{2}} - x_{i+\frac{1}{2}}}{\Delta t}. \tag{4.8}$$

(d) Use the Lax–Hopf formula to pick the correct foot of the backward characteristic line of the entropy solution among all the $\{x_i^*, u^*\}$ and $\{x_{i+\frac{1}{2}}, u^*\}$ that we have found in the previous step. The classical Lax formula [11] is as follows:

$$g\left(y_{j+\frac{1}{2}}\right) = \min_y g(y), \quad \text{where} \quad g(y) = \int_{-\infty}^{y} u^n(\xi)d\xi + \Delta t L\left(\frac{x_{j+\frac{1}{2}} - y}{\Delta t}\right), \tag{4.9}$$

where the lower limit $-\infty$ in the integral is irrelevant and can be replaced by any fixed point, and

$$L(v) = \int^{v} (f')^{-1}(u)du$$

is the Legendre transform of $f(u)$. For convex conservation laws, the Legendre transform of $f(u)$ can be obtained in closed form as

$$L(v) = v(f')^{-1}(v) - f((f')^{-1}(v)).$$

Especially, in (4.9),

$$L\left(\frac{x_{j+\frac{1}{2}} - y}{\Delta t}\right) = \frac{x_{j+\frac{1}{2}} - y}{\Delta t}u^* - f(u^*) = f'(u^*)u^* - f(u^*) \tag{4.10}$$

with u^* solved from (4.7) or (4.8).

If there are multiple minimizers to $g(y)$, corresponding to the situation that the solution is discontinuous at $(x_{j+\frac{1}{2}}, t^{n+1})$, then we will just take the smallest minimizer as $y_{j+\frac{1}{2}}$, corresponding to the left limit value of the solution at the discontinuity $(x_{j+\frac{1}{2}}, t^{n+1})$. It is here that we use the property that

the conservation law has a convex flux $f(u)$, as the Lax formula is only valid for convex conservation laws. Notice that the minimum is taken over finitely many candidates $\{x_i^*, u^*\}$ and $\{x_{i+\frac{1}{2}}, u^*\}$ that we have found in the previous step; hence there is no error incurred in this step either.

We now summarize our implementation of the large time step Godunov-type scheme for a scalar one-dimensional convex conservation law (1.1), from $\bar{u}^n$ to $\bar{u}^{n+1}$.

1. Perform a reconstruction to obtain a piecewise polynomial $u^n(x)$. For example, for second-order TVD or third-order TVB reconstruction, the implementation follows the procedure described in (3.1) and (3.2).
2. Find the domain of influence J_i for each cell I_i and the domain of influence $\tilde{J}_{i+\frac{1}{2}}$ for each cell boundary point $x_{i+\frac{1}{2}}$.
3. Find the set $C_{j+\frac{1}{2}}$ of all candidate intervals and the set $\tilde{C}_{j+\frac{1}{2}}$ of all candidate cell boundary points at time level t^n which will contain forward characteristic lines crossing the target point $(x_{j+\frac{1}{2}}, t^{n+1})$, via (4.6).
4. For each J_i with $i \in C_{j+\frac{1}{2}}$, or for each J_i^ℓ if J_i is a union of at least two of them, find a forward characteristic line from (x_i^*, t^n) with the solution value u^*, via (4.7). Likewise, for each $\tilde{J}_{i+\frac{1}{2}}$ with $i \in \tilde{C}_{j+\frac{1}{2}}$, find a forward characteristic line from $(x_{i+\frac{1}{2}}, t^n)$ with a solution value u^*, via (4.8).
5. Use the Lax formula (4.9) to determine the correct backward characteristic line $\{y_{j+\frac{1}{2}}, u_{j+\frac{1}{2}}\}$ among all the (finitely many) $\{x_i^*, u^*\}$ and $\{x_{i+\frac{1}{2}}, u^*\}$ that we have found above.
6. Use (4.5) to obtain $\bar{u}_j^{n+1}$ for all j.

Remark 4.1. In general, there might be errors introduced in the exact evolution stage of our scheme; for example, the exact characteristics may not be easy to obtain and an approximation is used instead. Let $\tilde{S}_{\Delta t}$ be the approximate evolution operator and let

$$\|\tilde{S}_{\Delta t}(u) - S_{\Delta t}(u)\|_{L^1} \le \delta(\Delta x, \Delta t) \tag{4.11}$$

for our piecewise polynomial, bounded variation numerical solution u, or weaker still

$$\|A \circ \tilde{S}_{\Delta t}(u) - A \circ S_{\Delta t}(u)\|_{L^1} \le \delta(\Delta x, \Delta t). \tag{4.12}$$

A reasonable assumption is that $\delta(\Delta x, \Delta t) \le C\Delta x$. Under this assumption, Theorem 2.1 still holds with a very minor modification in the proof. Specifically,

$$\begin{aligned}
\|A(u(\cdot, t^n)) - \bar{u}^n\|_{L^1} &= \|A(u(\cdot, t^n)) - A \circ \tilde{S}_{\Delta t} \circ Re\,(\bar{u}^{n-1})\|_{L^1} \\
&\le \|A(u(\cdot, t^n)) - A \circ S_{\Delta t} \circ Re\,(\bar{u}^{n-1})\|_{L^1} \qquad (4.13) \\
&\quad + \|A \circ S_{\Delta t} \circ Re\,(\bar{u}^{n-1}) - A \circ \tilde{S}_{\Delta t} \circ Re\,(\bar{u}^{n-1})\|_{L^1} \\
&\le \|u(\cdot, t^n) - S_{\Delta t} \circ Re\,(\bar{u}^{n-1})\|_{L^1} + C\Delta x \\
&= \|S_{\Delta t}\,(u(\cdot, t^{n-1})) - S_{\Delta t} \circ Re\,(\bar{u}^{n-1})\|_{L^1} + C\Delta x \\
&\le \|u(\cdot, t^{n-1}) - Re\,(\bar{u}^{n-1})\|_{L^1} + C\Delta x \qquad (4.14) \\
&= e^{n-1} + C\Delta x, \qquad (4.15)
\end{aligned}$$

and according to (2.7), $e^n \to 0$ as $\Delta x \to 0$. Again, with the assumption (4.11) or

(4.12), the local truncation error of our scheme would be

$$\begin{aligned}\|Au(\cdot,t^{n+1})-\bar{u}^{n+1}\|_{L^1} &= \|A\circ S_{\Delta t}(u(\cdot,t^n))-A\circ\tilde{S}_{\Delta t}\circ Re(\bar{u}^n)\|_{L^1}\\ &\le \|A\circ S_{\Delta t}(u(\cdot,t^n))-A\circ S_{\Delta t}\circ Re(\bar{u}^n)\|_{L^1}\\ &\quad +\|A\circ S_{\Delta t}\circ Re(\bar{u}^n)-A\circ\tilde{S}_{\Delta t}\circ Re(\bar{u}^n)\|_{L^1}\\ &\le \|S_{\Delta t}(u(\cdot,t^n))-S_{\Delta t}\circ Re(\bar{u}^n)\|_{L^1}+\delta(\Delta x,\Delta t)\\ &\le \|u(\cdot,t^n)-Re(\bar{u}^n)\|_{L^1}+\delta(\Delta x,\Delta t).\end{aligned}$$

That is, the local truncation error is no larger than the reconstruction error plus the error in the evolution operator.

In the following we provide numerical results of the large time step Godunov-type scheme applied to the Burgers equation, namely (1.1) with $f(u)=\frac{u^2}{2}$, for several different reconstructions and for both smooth and discontinuous solutions. We take $\Delta t=\Delta x^{0.9}$ in all runs.

Example 4.2. Burgers equation with the initial condition

$$u_0(x)=\sin(\pi x)+0.5,\qquad -1\le x<1, \tag{4.16}$$

with periodic boundary condition.

The solution at $t=0.2$ is still smooth, and we record the L^1 errors and numerical orders of accuracy in Table 4.1 for the first-order Godunov method (no reconstruction), second-order MUSCL reconstruction, and third-order nonoscillatory reconstruction in [15]. All these cases correspond to TVB reconstructions and hence provable convergence for the large time step Godunov-type scheme according to the discussions in the previous two sections. Moreover, since the solution is smooth, we expect the accuracy should be that of the reconstruction. We can see clearly from Table 4.1 that all schemes achieve their designed order of accuracy.

Even though the third-order linear central reconstruction (namely, the stencil for reconstructing the polynomial in I_j is fixed as $\{I_{j-1},I_j,I_{j+1}\}$) and the third-order ENO reconstruction [8] are not proven to be TVB, they perform well in this smooth test case, yielding the expected third-order accuracy for the large time step Godunov-type scheme; see Table 4.2.

Next, we continue the numerical simulation to $t=0.97$ when a shock has already appeared. We record the L^1 errors and numerical orders of accuracy in Table 4.3 for the first-order Godunov method (no reconstruction), the second-order MUSCL reconstruction, the third-order nonoscillatory reconstruction in [15], and the third-order ENO reconstruction [8]. Our theory predicts convergence for such discontinuous solutions for the first three reconstructions, which are TVB, but only guarantees a

TABLE 4.1

L^1 errors and numerical orders of accuracy for different TVD reconstructions. Burgers equation with initial condition (4.16)*, $t=0.2$.*

Number of cells	Constant		MUSCL		Liu and Osher [15]	
	L^1 error	Order	L^1 error	Order	L^1 error	Order
40	9.11E-3	—	3.79E-4	—	7.18E-5	—
80	5.29E-3	0.78	1.51E-4	1.32	2.87E-5	1.32
160	3.01E-3	0.81	2.93E-5	2.36	6.60E-6	2.12
320	1.70E-3	0.82	8.25E-6	1.82	9.59E-7	2.78
640	8.89E-4	0.93	2.21E-6	1.89	1.16E-7	3.04
1280	4.52E-4	0.97	7.24E-7	1.61	1.80E-8	2.68

TABLE 4.2

L^1 errors and numerical orders of accuracy for third-order linear central and ENO reconstructions. Burgers equation with initial condition (4.16), $t = 0.2$.

Number of cells	Central		ENO	
	L^1 error	Order	L^1 error	Order
40	7.54E-5	—	7.03E-5	—
80	2.06E-5	1.86	1.66E-5	2.08
160	2.92E-6	2.82	2.31E-6	2.84
320	4.22E-7	2.78	3.18E-7	2.86
640	5.76E-8	2.87	4.07E-8	2.96
1280	7.01E-9	3.03	5.41E-8	2.91

pessimistic error bound $\Delta x/\Delta t$; see (2.7). In the actual numerical test, we observe much better convergence rates, often close to the order of the reconstruction, even for this solution with a shock, although the convergence rate is not uniform, indicating a fluctuation in the accuracy of the solution in the shocked cell.

In Figure 2, we plot the numerical solution (cell averages) versus the exact cell averages of the entropy solution, for the third-order nonoscillatory reconstruction in [15] with various mesh sizes. We see clearly that the scheme can even resolve the transition cell average inside the shock very well. It would be interesting to study in more detail this good convergence behavior; however, the tools in previous sections are not enough for this detailed analysis. Finally, even though the ENO reconstruction is not TVB and hence is not covered by our theory, the convergence for this shocked problem is also very good; see again Table 4.3.

Example 4.3. Burgers equation with a discontinuous initial condition,

$$u_0(x) = \begin{cases} 1 & \text{if } -1 \le x < 0, \\ 3 & \text{if } 0 \le x < 1, \end{cases} \tag{4.17}$$

with periodic boundary condition. The solutions of the large time step Godunov-type scheme at $t = 0.2$ for different reconstructions are shown in Figure 3. We can see in general good agreement between the numerical and the exact entropy solutions. The first-order Godunov scheme is showing some stair-like features in the rarefaction region, which is well known, and the schemes based on linear reconstructions, namely the second-order upwind reconstruction (for this case with $f'(u) > 0$ the stencil for reconstructing the polynomial in I_j is fixed as $\{I_{j-1}, I_j\}$) and the third-order central reconstruction are both slightly oscillatory. The schemes based on MUSCL, Liu–Osher nonoscillatory, and ENO reconstructions are all nonoscillatory as expected.

TABLE 4.3

L^1 errors and numerical orders of accuracy for different reconstructions. Burgers equation with initial condition (4.16), $t = 0.97$.

Number of cells	Constant		MUSCL		Liu and Osher [15]		3rd-order ENO	
	L^1 error	Order	L^1 error	Order	L^1 error	Order	L^1 error	Order
38	1.16E-2	—	4.67E-4	—	1.02E-3	—	2.30E-5	—
76	6.51E-3	0.78	8.59E-5	2.44	2.68E-5	5.25	2.18E-5	0.08
152	3.59E-3	0.81	3.08E-5	1.47	2.78E-4	−3.37	5.43E-7	5.33
304	1.87E-3	0.82	1.04E-5	1.56	9.20E-5	1.60	5.44E-8	3.32
608	8.84E-4	0.93	1.70E-6	2.61	4.16E-6	4.46	6.21E-9	3.12
1216	4.15E-4	0.97	4.43E-7	1.94	1.69E-5	−2.02	7.92E-10	2.97

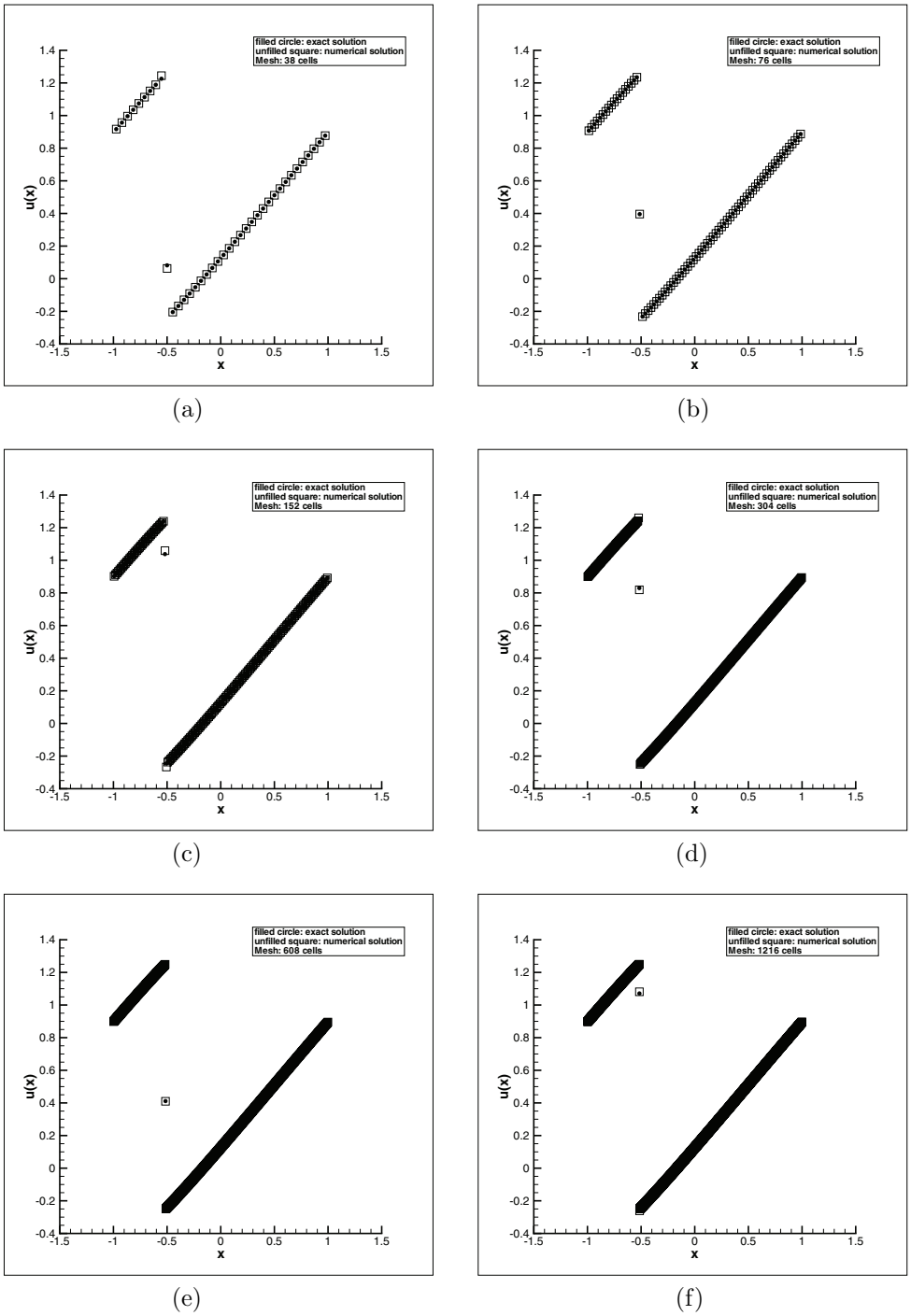


FIG. 2. *Burgers equation with initial condition* (4.16), $t = 0.97$. *Liu and Osher's third-order nonoscillatory reconstruction:* (a) 38 *uniform cells;* (b) 76 *uniform cells;* (c) 152 *uniform cells;* (d) 304 *uniform cells;* (e) 608 *uniform cells;* (f) 1216 *uniform cells. Solid symbol: cell averages of the exact solution; empty squares: numerical cell averages of the scheme.*

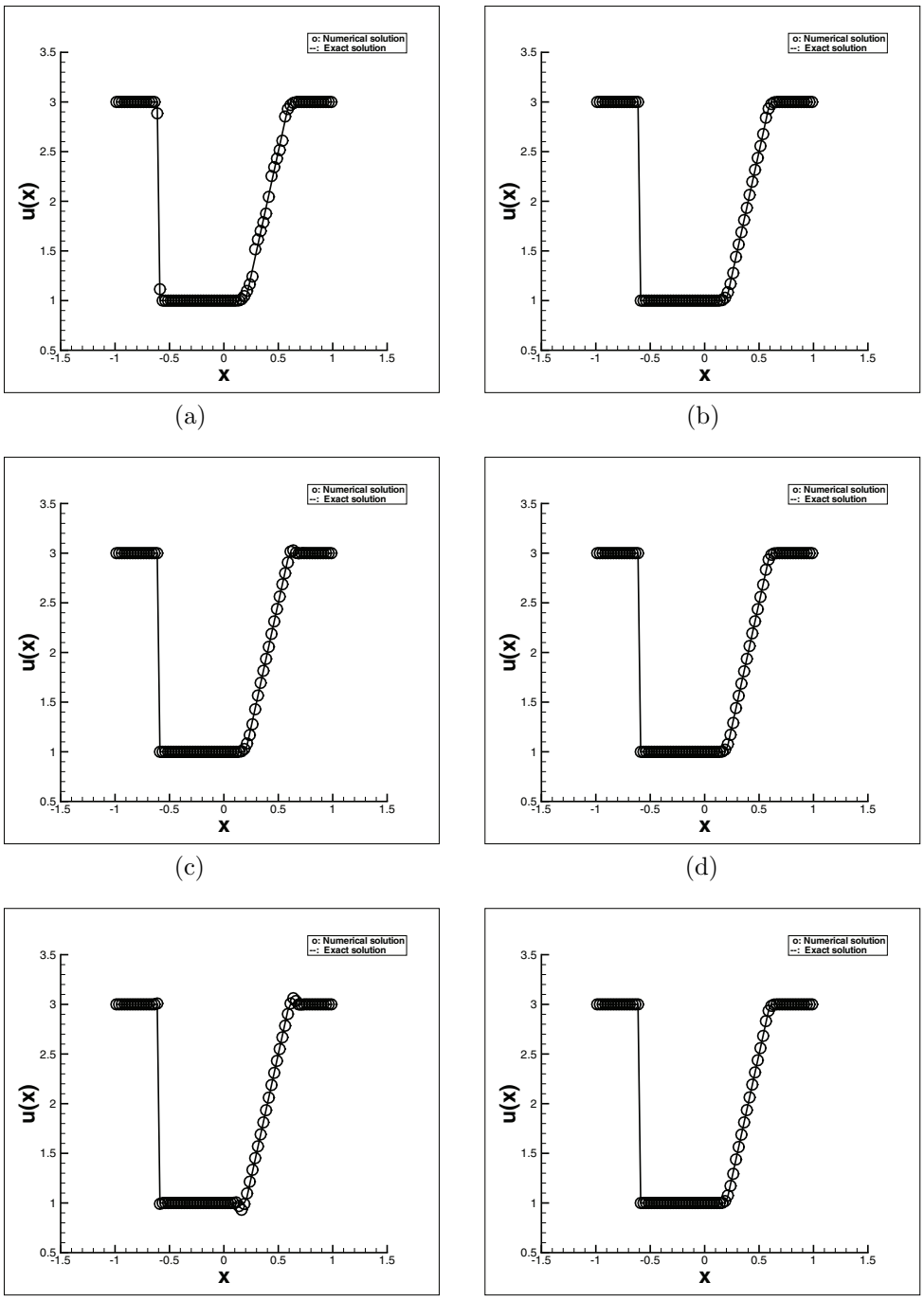


FIG. 3. *Burgers equation with the discontinuous initial condition* (4.17), $t = 0.2$. 80 *uniform cells.* (a) *First-order scheme with no reconstruction;* (b) *second-order MUSCL reconstruction;* (c) *second-order upwind reconstruction;* (d) *third-order ENO reconstruction;* (e) *third-order linear central reconstruction;* (f) *third-order nonoscillatory reconstruction of Liu and Osher. Solid line: exact solution; empty circle: numerical solution.*

5. Convergence of the large time step Godunov-like scheme based on Sanders' TVD reconstruction. Sanders [19] proposed a third-order accurate TVD (he called it total variation nonexpansive) scheme by evolving both the cell average and the values at two end points of a cell. He can then obtain a third-order TVD scheme instead of the usual Godunov-type TVD scheme which degenerates to second order at best. This scheme does not fit the framework of Godunov-type schemes that we have discussed before, since it evolves not just the cell averages but also point values of the solution at cell boundaries. However, the solution procedure is sufficiently similar to Godunov-type schemes, and therefore we will call it a Godunov-like scheme. The scheme is based on a reconstruction operator Rs, which produces a piecewise quadratic function. After a suitable initialization to obtain a piecewise quadratic initial numerical solution $u^0(x)$, the Godunov-type scheme based on Sanders' reconstruction operator Rs can be summarized as follows:

1. *Evolution:* evolve $u^n(x)$ by the conservation law (1.1) *exactly* for a time step Δt, to obtain a solution $\tilde{u}^{n+1}(x)$ which is in general *not* piecewise polynomial anymore. Recall that this evolution operator is denoted as $S_{\Delta t}(u^n)$.
2. *Reconstruction:* obtain a piecewise quadratic polynomial reconstruction $u^n(x)$ by Sanders' reconstruction. We denote this reconstruction operator as $u^{n+1} = Rs(\tilde{u}^{n+1})$.

This Godunov-like scheme, using the notations introduced above, can be described abstractly as

$$\text{(5.1)} \qquad \begin{cases} u^0 = Rs(u_0), \\ u^{n+1} = Rs \circ S_{\Delta t}(u^n), \qquad n = 0, 1, \ldots, N-1. \end{cases}$$

We will first discuss the TVD reconstruction operator Rs, then prove the convergence of the large time step Godunov-like scheme based on this reconstruction. Finally, we will discuss the implementation of this scheme, again for one-dimensional scalar convex conservation laws, and provide numerical examples.

5.1. Sanders' TVD reconstruction. Sanders' reconstruction Rs maps a bounded variation function $u(x)$ to a piecewise quadratic polynomial, satisfying the following four properties:

1. agreement of the cell averages:

$$\text{(5.2)} \qquad \int_{I_j} Rs(u)(x)dx = \int_{I_j} u(x)dx \qquad \forall j;$$

2. TVD:

$$\text{(5.3)} \qquad TV(Rs(u)) \le TV(u);$$

3. maximum (and minimum) principle:

$$\sup Rs(u)(x) \le \sup u(x), \qquad \inf Rs(u)(x) \ge \inf u(x);$$

4. third-order accuracy:

$$Rs(u)(x) = u(x) + O(\Delta x^3)$$

when u is locally smooth.

Let us now describe the details of the reconstruction operator Rs. First, given the cell average $\bar{u}_j$ and the cell boundary values $u^+_{j-\frac{1}{2}}$ and $u^-_{j+\frac{1}{2}}$, we can easily construct a unique quadratic polynomial $q_j(x)$ such that

$$\frac{1}{\Delta x}\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} q_j(x)\,dx = \bar{u}_j, \qquad q_j\left(x_{j-\frac{1}{2}}\right) = u^+_{j-\frac{1}{2}}, \qquad q_j\left(x_{j+\frac{1}{2}}\right) = u^-_{j+\frac{1}{2}}.$$

In fact, if we denote

$$a = u^+_{j-\frac{1}{2}} - \bar{u}_j, \qquad b = u^-_{j+\frac{1}{2}} - \bar{u}_j, \tag{5.4}$$

then the quadratic polynomial $q_j(x)$ is given by

$$q_j(x) = \hat{q}(a,b)(x) + \bar{u}_j = 3(a+b)\left(\frac{x-x_j}{\Delta x}\right)^2 + (b-a)\left(\frac{x-x_j}{\Delta x}\right) - \frac{1}{4}(a+b) + \bar{u}_j. \tag{5.5}$$

We define M to be the larger (in magnitude) of a and b and m the smaller one, $\rho = \frac{m}{M}$,

$$E = \begin{cases} \sup\limits_{I_j}(u - \bar{u}_j) & \text{if} \quad M \le 0, \\ \inf\limits_{I_j}(u - \bar{u}_j) & \text{if} \quad M > 0, \end{cases} \tag{5.6}$$

and $\hat{E} = \frac{E}{M}$. Let

$$\tau_+ = -\hat{E}\cdot\frac{3(1+\rho)}{1+\rho+\rho^2}, \qquad \tau_- = -\frac{1}{2}\left((\rho+3\hat{E}) - 3(\hat{E}-\rho)(3\hat{E}+\rho)^{\frac{1}{2}}\right),$$

and

$$\tau_l = \begin{cases} 1 & \text{if} \quad -1 \le \rho \le -\frac{1}{2}, \\ 1 & \text{if} \quad -\frac{1}{2} < \rho < 0 \quad \text{and} \quad a = M, \\ \min(\tau_-, 1) & \text{if} \quad -\frac{1}{2} < \rho < 0 \quad \text{and} \quad b = M, \\ \min(\tau_+, 1) & \text{if} \quad 0 \le \rho \le 1, \end{cases}$$

$$\tau_r = \begin{cases} 1 & \text{if} \quad -1 \le \rho \le -\frac{1}{2}, \\ 1 & \text{if} \quad -\frac{1}{2} < \rho < 0 \quad \text{and} \quad b = M, \\ \min(\tau_-, 1) & \text{if} \quad -\frac{1}{2} < \rho < 0 \quad \text{and} \quad a = M, \\ \min(\tau_+, 1) & \text{if} \quad 0 \le \rho \le 1; \end{cases}$$

then the TVD reconstruction in I_j is defined by

$$Rs(u)(x) = \hat{q}(\tau_l a, \tau_r b)(x) + \bar{u}_j \tag{5.7}$$

with $\hat{q}$ defined in (5.5). It is proved in [19] that the reconstruction Rs is third-order accurate and TVD.

5.2. Convergence of the Godunov-like scheme based on Sanders' reconstruction. In this subsection we prove the convergence of the Godunov-like scheme (5.1) based on Sanders' reconstruction.

THEOREM 5.1. *The numerical solution of the Godunov-like scheme* (5.1) *with Sanders' reconstruction* (5.7) *for solving the scalar conservation law* (1.1) *converges*

in L^1 to the entropy solution of (1.1) *when the mesh size Δx goes to zero if the large time step assumption* (2.3) *is satisfied.*

Proof. We again denote the exact entropy solution of (1.1) at time level t^n by $u(\cdot, t^n)$ and the L^1 difference between this exact solution and the numerical solution u^n at time level t^n by e^n,

$$e^n = \|u(\cdot, t^n) - u^n\|_{L^1}. \tag{5.8}$$

First, the total variation of the numerical solution is bounded (in fact, diminishing),

$$\begin{aligned} TV(u^n) &= TV(Rs \circ S_{\Delta t}\,(u^{n-1})) \\ &\le TV(u^{n-1}) \\ &\le \cdots \\ &\le TV(u^0) \le TV(u_0), \end{aligned} \tag{5.9}$$

where we have used the fact that neither the reconstruction operator Rs nor the exact entropy solution evolution operator $S_{\Delta t}$ increases the total variation of a function, for the second inequality.

We now proceed to prove the desired convergence

$$\begin{aligned} e^n &= \|u(\cdot, t^n) - u^n\|_{L^1} \\ &\le \|u(\cdot, t^n) - S_{\Delta t}(u^{n-1})\|_{L^1} + \|S_{\Delta t}(u^{n-1}) - A \circ S_{\Delta t}(u^{n-1})\|_{L^1} \\ &\quad + \|A \circ S_{\Delta t}(u^{n-1}) - u^n\|_{L^1} \\ &= \|S_{\Delta t}(u(\cdot, t^{n-1})) - S_{\Delta t}(u^{n-1})\|_{L^1} + \|S_{\Delta t}(u^{n-1}) - A \circ S_{\Delta t}(u^{n-1})\|_{L^1} \\ &\qquad + \|A \circ Rs \circ S_{\Delta t}(u^{n-1}) - Rs \circ S_{\Delta t}(u^{n-1})\|_{L^1} \\ &\le \|u(\cdot, t^{n-1}) - u^{n-1}\|_{L^1} + C\Delta x TV(S_{\Delta t}(u^{n-1})) + C\Delta x TV(Rs \circ S_{\Delta t}(u^{n-1})) \\ &\le e^{n-1} + C\Delta x TV(u^{n-1}) + C\Delta x TV(u^{n-1}) \\ &\le e^{n-1} + \tilde{C}\Delta x \\ &\le \cdots \\ &\le \tilde{C} n \Delta x + e^0 \\ &= \tilde{C} T \frac{\Delta x}{\Delta t} + e^0 \to 0 \qquad \text{as} \quad \Delta x \to 0, \end{aligned}$$

where for the third equality we have used the cell average preserving property (5.2) of Rs; for the fourth inequality we have used the fact that the exact entropy solution operator $S_{\Delta t}$ is an L^1 contraction, and the inequality (2.6) with $w = S_{\Delta t}(u^{n-1})$ and $w = Rs \circ S_{\Delta t}(u^{n-1})$, respectively; for the fifth inequality we have used the fact that the reconstruction operator Rs and the exact entropy solution operator $S_{\Delta t}$ are both TVD; and for the sixth inequality we have used the fact that the total variation of the numerical solution is bounded (5.9). ∎

5.3. Implementation and numerical examples. For one-dimensional scalar convex conservation law (1.1), the implementation of the large time step Godunov-like scheme based on Sanders' reconstruction operator Rs is similar to that for the large time step Godunov-type scheme described in detail in section 4. This is because the procedure described in section 4 provides both the cell average $\bar{u}_j^{n+1}$ and the cell boundary values $u_{j-\frac{1}{2}}$ and $u_{j+\frac{1}{2}}$; hence a and b in (5.4) are available. The only additional information needed is the relevant extremum value E in (5.6), which is in

TABLE 5.1

L^1 errors and numerical orders of accuracy for the unmodulated third-order reconstruction (5.5) *and Sanders' TVD reconstruction Rs. Burgers equation with initial condition* (4.16), $t = 0.2$.

Number of cells	Unmodulated third-order		Sanders	
	L^1 error	Order	L^1 error	Order
38	1.30E-5	—	1.34E-5	—
76	2.47E-6	2.39	2.55E-6	2.38
152	3.46E-7	2.83	3.69E-7	2.78
304	4.57E-8	2.92	4.92E-8	2.90
608	6.06E-9	2.92	6.41E-9	2.93

TABLE 5.2

L^1 errors and numerical orders of accuracy for the unmodulated third-order reconstruction (5.5) *and Sanders' TVD reconstruction Rs. Burgers equation with initial condition* (4.16), $t = 0.97$.

Number of cells	Unmodulated third-order		Sanders	
	L^1 error	Order	L^1 error	Order
38	2.31E-6	—	2.31E-6	—
76	3.13E-7	2.88	3.13E-7	2.88
152	4.32E-8	2.85	4.32E-8	2.85
304	5.93E-9	2.86	5.93E-9	2.86
608	8.40E-10	2.82	8.40E-10	2.82

general difficult to obtain. We follow Sanders [19] and take instead the corresponding extremum value E at the previous time level t^n between the feet of the characteristics $y_{j-\frac{1}{2}}$ and $y_{j+\frac{1}{2}}$. This approximation of E is exact if there is no discontinuity within the trapezoid Ω_j (see Figure 1), and is overestimating (larger for the sup and smaller for the inf) if there is at least one shock within Ω_j. It does not affect the accuracy and TVD property of the reconstruction Rs.

In the numerical experiments we take $\Delta t = \Delta x^{0.9}$ as before. We test the large time step Godunov-like scheme based on both an unmodulated third-order reconstruction (q_j as given by (5.5)) and Sanders' TVD reconstruction Rs.

We again first test the smooth solution as in Example 4.1. The L^1 errors and numerical orders of accuracy at $t = 0.2$ are shown in Table 5.1. We can see clearly third-order accuracy for the scheme based on both reconstructions.

Next, we record the L^1 errors and numerical orders of accuracy at $t = 0.97$, when the shock has already appeared, in Table 5.2. We observe again that the convergence rates are surprisingly good for both reconstructions, close to that of the reconstruction, even though the solution contains a shock. In Figures 4 and 5, we plot the numerical solution (cell averages) versus the exact cell averages of the entropy solution, for the scheme based on Sanders' TVD reconstruction Rs and based on the unmodulated third-order reconstruction (5.5), respectively, with various mesh sizes. We again observe that the scheme can resolve the transition cell average inside the shock very well.

Next, we test the discontinuous solution as in Example 4.2. Figure 6 displays the numerical solution. We can again see good agreement between the numerical and the exact entropy solutions. The scheme based on the unmodulated third-order reconstruction is slightly oscillatory, and the scheme based on Sanders' TVD reconstruction is not oscillatory, as expected.

6. Concluding remarks. The main result of this paper is the proof of convergence towards entropy solutions of large time step Godunov-type schemes with TVB

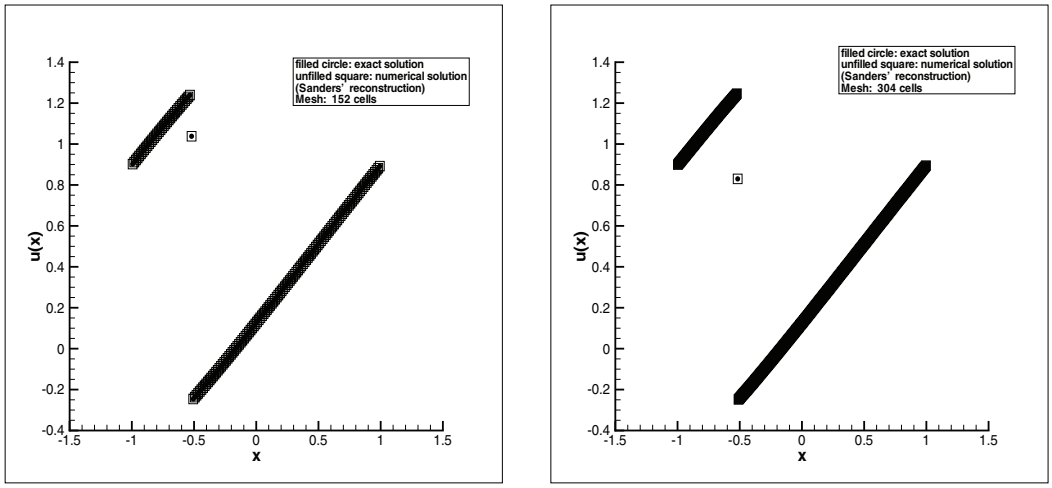


FIG. 4. *Burgers equation with the discontinuous initial condition* (4.16), $t = 0.97$. *Sanders' TVD reconstruction. Left*: 152 *uniform cells*; *right*: 304 *uniform cells.*

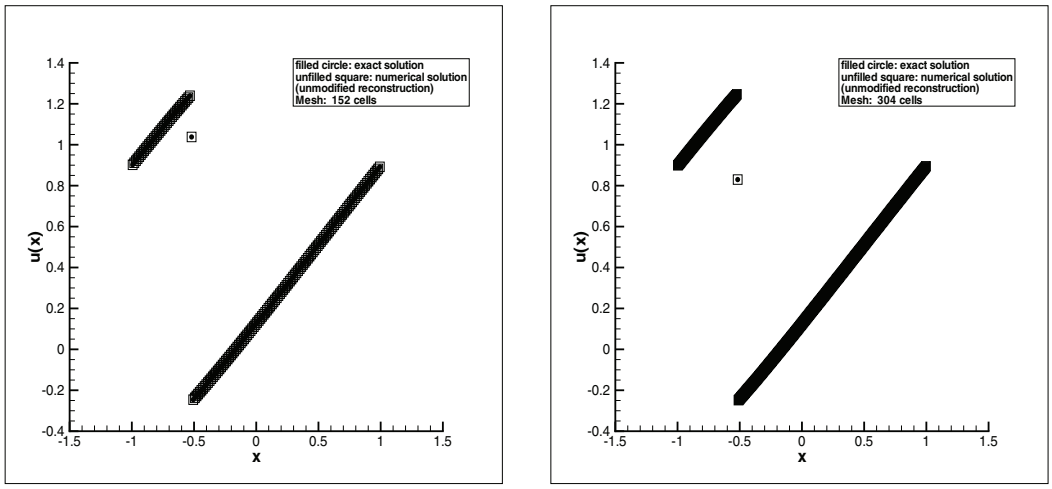


FIG. 5. *Burgers equation with the discontinuous initial condition* (4.16), $t = 0.97$. *Unmodulated third-order reconstruction. Left*: 152 *uniform cells*; *right*: 304 *uniform cells.*

reconstructions for general scalar conservation laws. It is somewhat surprising that no cell or wavewise entropy inequalities are needed in the proof. An intuitive explanation is that larger time step sizes lead to fewer time steps (asymptotically much less than $O(1/\Delta x)$) to reach the final time; hence there are fewer reconstructions performed, which are the main source of possible erroneous entropy production. We have also discussed practical implementation of such large time step Godunov-type schemes for scalar one-dimensional convex conservation laws and have provided numerical examples to demonstrate the performance of these schemes. Future work will include investigations on practical implementation for more general equations, and on possible elimination of the TVB assumption on the reconstruction (or the replacement of this condition by a weaker condition easier to satisfy for high-order reconstructions in multispace dimensions), at least for conservation laws with convex fluxes.

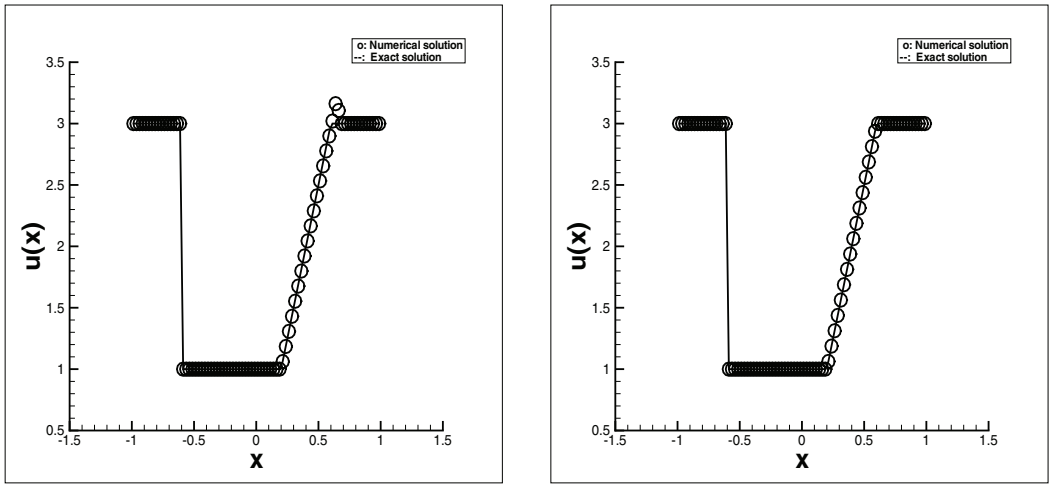


FIG. 6. *Burgers equation with the discontinuous initial condition* (4.17), $t = 0.2$. 80 *uniform cells. Left*: *unmodulated third-order reconstruction*; *right*: *Sanders' TVD reconstruction.*

Appendix A. Truncation error analysis. To look at the local truncation error of the scheme (1.2), we assume that we start from the cell averages $\bar{u}_j^n = Au(\cdot, t^n)|_{I_j}$ of the exact solution $u(\cdot, t)$ of the PDE at time level t^n and use the scheme to advance for one time step to obtain the numerical solution $\bar{u}^{n+1}$.

Let $\bar{u}_j^{e,n} = \bar{u}_j^n = Au(\cdot, t^n)|_{I_j}$ and $\bar{u}_j^{e,n+1} = Au(\cdot, t^{n+1})|_{I_j}$. The local truncation error is therefore defined as

$$LTE_j = \frac{1}{\Delta t}\left(\bar{u}_j^{e,n+1} - \bar{u}_j^{n+1}\right). \tag{A.1}$$

We will prove below that this local truncation error is at the level of $O(\Delta x^k)$ when piecewise polynomial of degree at most $k-1$ is used in the reconstruction, for smooth solutions, regardless of the size of Δt.

Evolving the PDE (1.1) from the two sets of initial data, the exact solution $u(\cdot, t^n)$ and the reconstructed solution $Re(\bar{u}^n)$ from the exact cell averages, respectively, give us $y^e_{j+\frac{1}{2}}$ and $y_{j+\frac{1}{2}}$, the feet of the traced-back characteristic lines from $x_{j+\frac{1}{2}}$ at t^n (see Figure 1), as well as $u^e_{j+\frac{1}{2}}$ and $u_{j+\frac{1}{2}}$, the values of the solution along these traced-back characteristics.

When the solution is smooth, according to the characteristic equations,

$$u_{j+\frac{1}{2}} = u^n\left(x_{j+\frac{1}{2}} - f'\left(u_{j+\frac{1}{2}}\right)\Delta t\right), \tag{A.2}$$

$$u^e_{j+\frac{1}{2}} = u\left(x_{j+\frac{1}{2}} - f'\left(u^e_{j+\frac{1}{2}}\right)\Delta t, t^n\right) = u^n\left(x_{j+\frac{1}{2}} - f'\left(u^e_{j+\frac{1}{2}}\right)\Delta t\right) + \mathcal{O}(\Delta x^k), \tag{A.3}$$

we have

$$|u^e_{j+\frac{1}{2}} - u_{j+\frac{1}{2}}| \sim \mathcal{O}(\Delta x^k), \qquad |y^e_{j+\frac{1}{2}} - y_{j+\frac{1}{2}}| \sim \mathcal{O}(\Delta x^k \Delta t). \tag{A.4}$$

Again if the solution is smooth, we have

$$u_{j+\frac{1}{2}} - u_{j-\frac{1}{2}} \sim \mathcal{O}(\Delta x), \qquad y_{j+\frac{1}{2}} - y_{j-\frac{1}{2}} \sim \mathcal{O}(\Delta x). \tag{A.5}$$

From the scheme, we can obtain the formula (4.5) for computing $\bar{u}_j^{n+1}$. Similarly, evolving (1.1) with the initial data $u(\cdot, t^n)$ gives us

$$\begin{aligned}\Delta x \bar{u}_j^{e,n+1} = &\int_{y^e_{j-\frac{1}{2}}}^{y^e_{j+\frac{1}{2}}} u(x,t^n)dx - \left(\Delta t f(u^e_{j+\frac{1}{2}}) + \left(y^e_{j+\frac{1}{2}} - x_{j+\frac{1}{2}}\right) u^e_{j+\frac{1}{2}}\right) \\ &+ \left(\Delta t f(u^e_{j-\frac{1}{2}}) + \left(y^e_{j-\frac{1}{2}} - x_{j-\frac{1}{2}}\right) u^e_{j-\frac{1}{2}}\right). \end{aligned} \tag{A.6}$$

Subtracting (4.5) from (A.6), we obtain

$$\begin{aligned}\Delta x(\bar{u}_j^{e,n+1} - \bar{u}_j^{n+1}) = &\left(\int_{y^e_{j-\frac{1}{2}}}^{y^e_{j+\frac{1}{2}}} u(x,t^n)dx - \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} u^n(x)dx\right) \\ &- \left(\Delta t\left(f\left(u^e_{j+\frac{1}{2}}\right) - f\left(u_{j+\frac{1}{2}}\right)\right) - \Delta t\left(f\left(u^e_{j-\frac{1}{2}}\right) - f\left(u_{j-\frac{1}{2}}\right)\right)\right) \\ &+ \left(\left(y^e_{j+\frac{1}{2}} - x_{j+\frac{1}{2}}\right) u^e_{j+\frac{1}{2}} - \left(y_{j+\frac{1}{2}} - x_{j+\frac{1}{2}}\right) u_{j+\frac{1}{2}}\right. \\ &\quad \left. - \left(y^e_{j-\frac{1}{2}} - x_{j-\frac{1}{2}}\right) u^e_{j-\frac{1}{2}} + \left(y_{j-\frac{1}{2}} - x_{j-\frac{1}{2}}\right) u_{j-\frac{1}{2}}\right). \end{aligned} \tag{A.7}$$

By (A.4), the first term on the second line of (A.7) is of the size $\mathcal{O}(\Delta x^k \Delta t)$. The second term on that line is identical to the first term with a shift of the spatial grid index from $j+\frac{1}{2}$ to $j-\frac{1}{2}$. By (A.5) and the smoothness of the leading coefficient of the error term $\mathcal{O}(\Delta x^k \Delta t)$, the difference of the two terms on the second line of (A.7) gives an additional $\mathcal{O}(\Delta x)$; that is, this line is of the size $\mathcal{O}(\Delta x^{k+1}\Delta t)$. With the same reason, the difference of the two terms on the third and fourth lines of (A.7) is also of the order $\mathcal{O}(\Delta x^{k+1}\Delta t)$. Now let us look at the two terms on the right-hand side of the first line of (A.7). Without loss of generality, we assume that $y^e_{j-\frac{1}{2}} > y_{j-\frac{1}{2}}$ and $y^e_{j+\frac{1}{2}} > y_{j+\frac{1}{2}}$. We also assume that $y_{j-\frac{1}{2}}$ and $y^e_{j-\frac{1}{2}}$ fall into the same interval, say I_{jl}, and that $y_{j+\frac{1}{2}}$ and $y^e_{j+\frac{1}{2}}$ fall into the same neighboring interval $I_{jr} = I_{jl+1}$. Other situations can be handled similarly. We now have

$$\begin{aligned}&\int_{y^e_{j-\frac{1}{2}}}^{y^e_{j+\frac{1}{2}}} u(x,t^n)dx - \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} u^n(x)dx \\ &\quad = \left(\int_{y^e_{j-\frac{1}{2}}}^{x_{jl+\frac{1}{2}}} + \int_{x_{jr-\frac{1}{2}}}^{y^e_{j+\frac{1}{2}}}\right) u(x,t^n)dx - \left(\int_{y_{j-\frac{1}{2}}}^{x_{jl+\frac{1}{2}}} + \int_{x_{jr-\frac{1}{2}}}^{y_{j+\frac{1}{2}}}\right) u^n(x)dx \\ &\quad = \left(\left(\int_{x_{jr-\frac{1}{2}}}^{y^e_{j+\frac{1}{2}}} + \int_{x_{jl-\frac{1}{2}}}^{x_{jl+\frac{1}{2}}} - \int_{x_{jl-\frac{1}{2}}}^{y^e_{j-\frac{1}{2}}}\right) u(x,t^n)dx\right) \\ &\qquad - \left(\left(\int_{x_{jr-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} + \int_{x_{jl-\frac{1}{2}}}^{x_{jl+\frac{1}{2}}} - \int_{x_{jl-\frac{1}{2}}}^{y_{j-\frac{1}{2}}}\right) u^n(x)dx\right) \\ &\quad = \left(\int_{x_{jr-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} - \int_{x_{jl-\frac{1}{2}}}^{y_{j-\frac{1}{2}}}\right)(u(x,t^n) - u^n(x))dx + \left(\int_{y_{j+\frac{1}{2}}}^{y^e_{j+\frac{1}{2}}} - \int_{y_{j-\frac{1}{2}}}^{y^e_{j-\frac{1}{2}}}\right) u(x,t^n)dx, \end{aligned} \tag{A.8}$$

where we have used the fact that the integrals of $u(x,t^n)$ and of $u^n(x)$ over any cell I_i are the same. The first term in the last line of (A.8) is a difference of two integrals,

TABLE A.1

The local truncation error (A.1). *Godunov-type scheme with one time step evolution. L^1 norm and numerical order of accuracy for different TVD reconstructions. Burgers equation with initial condition $u_0(x) = \sin(\pi x)$. $\Delta t = \Delta x^{0.5}$.*

Number of cells	Constant		MUSCL		Liu and Osher [15]	
	L^1 error	Order	L^1 error	Order	L^1 error	Order
160	1.11E-2	—	2.11E-4	—	3.45E-6	—
320	6.06E-3	0.88	5.28E-5	2.00	4.64E-7	2.89
640	3.23E-3	0.90	1.28E-5	2.04	6.13E-8	2.92

with the integrand of the order $\mathcal{O}(\Delta x^k)$ and the length of the integral interval of the order $\min(\Delta x, \Delta t)$. Therefore, this first term is a difference of two identical $\mathcal{O}(\Delta x^k \min(\Delta x, \Delta t))$ terms with only a shift of the spatial grid index from $j+\frac{1}{2}$ to $j-\frac{1}{2}$. As before, we conclude that this first term is of the order $\mathcal{O}(\Delta x^{k+1} \min(\Delta x, \Delta t))$, or no more than the order $\mathcal{O}(\Delta x^{k+1}\Delta t)$. The second term in the last line of (A.8) is a difference of two integrals, with the integrand of the order $\mathcal{O}(1)$ and the length of the integral interval of the order $\mathcal{O}(\Delta x^k \Delta t)$. Hence, this second term is also of the order $\mathcal{O}(\Delta x^{k+1}\Delta t)$.

We can now conclude that the right-hand side of (A.7), $\Delta x(\bar{u}_j^{e,n+1} - \bar{u}_j^{n+1})$, is, therefore, of order $\mathcal{O}(\Delta x^{k+1}\Delta t)$. This implies that the local truncation error (A.1) is of the order $\mathcal{O}(\Delta x^k)$.

We can numerically verify the analysis above by giving the L^1 norm and order of accuracy of the local truncation error (A.1) for evolving the Burgers equation for one time step with different choices of Δt. In Table A.1 we list the results for the choice of $\Delta t = \Delta x^{0.5}$ with different orders of reconstruction, verifying that the local truncation error (A.1) is of the order $\mathcal{O}(\Delta x^k)$. We have also performed the numerical experiments for other choices of Δt, including $\Delta t = \Delta x^0, \Delta x^1$, and Δx^2, obtaining the same orders of accuracy for the local truncation error. We omit these error tables to save space.

Acknowledgment. We would like to thank the anonymous referees for their helpful and constructive comments, which led to an improvement of the paper.

REFERENCES

[1] F. BOUCHUT, *An antidiffusive entropy scheme for monotone scalar conservation laws*, J. Sci. Comput., 21 (2004), pp. 1–30.

[2] F. BOUCHUT, C. BOURDARIAS, AND B. PERTHAME, *A MUSCL method satisfying all the numerical entropy inequalities*, Math. Comp., 65 (1996), pp. 1439–1461.

[3] E. CARLINI, R. FERRETTI, AND G. RUSSO, *A weighted essentially nonoscillatory, large time-step scheme for Hamilton–Jacobi equations*, SIAM J. Sci. Comput., 27 (2005), pp. 1071–1091.

[4] L. CORRIAS, M. FALCONE, AND R. NATALINI, *Numerical schemes for conservation laws via Hamilton–Jacobi equations*, Math. Comp., 64 (1995), pp. 555–580.

[5] R. FERRETTI, *Convergence of semi-Lagrangian approximation to convex Hamilton–Jacobi equations under (very) large Courant numbers*, SIAM J. Numer. Anal., 40 (2003), pp. 2240–2253.

[6] J. GOODMAN AND R. LEVEQUE, *On the accuracy of stable schemes for 2D scalar conservation laws*, Math. Comp., 45 (1985), pp. 15–21.

[7] A. HARTEN, *High resolution schemes for hyperbolic conservation laws*, J. Comput. Phys., 49 (1983), pp. 357–393.

[8] A. HARTEN, B. ENGQUIST, S. OSHER, AND S. CHAKRAVARTHY, *Uniformly high order accurate essentially nonoscillatory schemes*, III, J. Comput. Phys., 71 (1987), pp. 231–303.

[9] G.-S. JIANG AND C.-W. SHU, *On cell entropy inequality for discontinuous Galerkin methods*, Math. Comp., 62 (1994), pp. 531–538.

[10] G.-S. Jiang and C.-W. Shu, *Efficient implementation of weighted ENO schemes*, J. Comput. Phys., 126 (1996), pp. 202–228.
[11] P. Lax, *Hyperbolic systems of conservation laws* II, Comm. Pure Appl. Math., 10 (1957), pp. 537–566.
[12] P. D. Lax, *Hyperbolic Systems of Conservation Laws and the Mathematical Theory of Shock Waves*, CBMS-NSF Regional Conf. Ser. Appl. Math. 11, SIAM, Philadelphia, 1973.
[13] R. J. LeVeque, *A large time step generalization of Godunov's method for systems of conservation laws*, SIAM J. Numer. Anal., 22 (1985), pp. 1051–1073.
[14] R. J. LeVeque, *Numerical Methods for Conservation Laws*, Birkhäuser-Verlag, Basel, 1990.
[15] X.-D. Liu and S. Osher, *Nonoscillatory high order accurate self-similar maximum principle satisfying shock capturing schemes* I, SIAM J. Numer. Anal., 33 (1996), pp. 760–779.
[16] P.-L. Lions and P. E. Souganidis, *Convergence of MUSCL and filtered schemes for scalar conservation law and Hamilton–Jacobi equations*, Numer. Math., 69 (1995), pp. 441–470.
[17] S. Osher and S. Chakravarthy, *Very high order accurate TVD schemes*, in Oscillation Theory, Computation, and Methods of Compensated Compactness, C. Dafermos, J. L. Ericksen, D. Kinderlehrer, and M. Slemrod, eds., IMA Vol. Math. Appl. 2, Springer-Verlag, 1986, pp. 229–274.
[18] S. Osher and E. Tadmor, *On the convergence of the difference approximations to scalar conservation laws*, Math. Comp., 50 (1988), pp. 19–51.
[19] R. Sanders, *A third-order accurate variation nonexpansive difference scheme for single nonlinear conservation law*, Math. Comp., 51 (1988), pp. 535–558.
[20] C.-W. Shu, *TVB uniformly high order schemes for conservation laws*, Math. Comp., 49 (1987), pp. 105–121.
[21] C.-W. Shu, *Essentially non-oscillatory and weighted essentially non-oscillatory schemes for hyperbolic conservation laws*, in Advanced Numerical Approximation of Nonlinear Hyperbolic Equations, A. Quarteroni, ed., Lecture Notes in Math. 1697, Springer-Verlag, Berlin, 1998, pp. 325–432.
[22] B. van Leer, *Towards the ultimate conservative difference scheme* V. *A second order sequel to Godunov's method*, J. Comput. Phys., 32 (1979), pp. 101–136.
[23] H. Yang, *Convergence of Godunov type schemes*, Appl. Math. Lett., 9 (1996), pp. 63–67.
[24] H. Yang, *On wavewise entropy inequality for high resolution schemes* II: *Fully discrete MUSCL schemes with exact evolution in small time*, SIAM J. Numer. Anal., 36 (1998), pp. 1–31.

SIAM J. NUMER. ANAL.
Vol. 46, No. 5, pp. 2238–2253
© 2008 Society for Industrial and Applied Mathematics

ON NUMERICAL ALGORITHMS FOR THE SOLUTION OF A BELTRAMI EQUATION*

DENIS GAIDASHEV† AND DMITRY KHMELEV

Abstract. The paper concerns numerical algorithms for solving the Beltrami equation $f_{\bar{z}}(z) = \mu(z)f_z(z)$ for a compactly supported μ. First, we study an efficient algorithm that has been proposed in [P. Daripa, *J. Comput. Phys.*, 106 (1993), pp. 355–365] and [P. Daripa and D. Mashat, *Numer. Algorithms*, 18 (1998), pp. 133–157] and present its rigorous justification. We then propose a different scheme for solving the Beltrami equation which has a comparable speed and accuracy, but has the virtue of easier implementation by avoiding the use of the Hilbert transform. The present paper can also be viewed as a prologue to one important application of the Beltrami equation: it provides a detailed description of the algorithm that has been used in [D. Gaidashev, *Nonlinearity*, 20 (1998), pp. 713–741] and [D. Gaidashev and M. Yampolsky, *Experiment. Math.*, 16 (2007), pp. 215–226] to address an important issue in complex dynamics—conjectural universality for Siegel disks.

Key words. Beltrami equation, Hilbert transform, Cauchy transform, measurable Riemann mapping, quasi-conformal map

AMS subject classifications. 65Y20, 65E05, 45P05, 44A20, 30E20

DOI. 10.1137/050640710

1. Introduction.

The problem of existence of a quasi-conformal mapping g with a given complex dilatation μ is one of the central problems of classical complex analysis.

We will first recall a definition of a quasi-conformal map. Given an open set $\Omega \subset \mathbb{C}$, a map $f : \Omega \to f(\Omega) \subset \mathbb{C}$ is said to be absolutely continuous on lines (ACL) in a rectangle $R \subset \Omega$ with sides parallel to the x and y axes if f is absolutely continuous on almost every horizontal and vertical line in R. The map f is ACL on Ω if it is ACL on every rectangle in Ω. Partial derivatives $f_z = (f_x - if_y)/2$ and $f_{\bar{z}} = (f_x + if_y)/2$ of such a map exist a.e. in Ω.

DEFINITION 1.1. *A homeomorphism $f : \Omega \to f(\Omega)$ is K-quasi-conformal if and only if the following hold:*

(i) *f is ACL on Ω.*

(ii) *$|f_{\bar{z}}| \le \frac{K-1}{K+1}|f_z|$ a.e. in Ω.*

The complex dilatation of a quasi-conformal homeomorphism f defined a.e. is

$$\mu_f(z) = f_{\bar{z}}(z)/f_z(z).$$

A mapping f with a prescribed complex dilatation μ solves the following *Beltrami equation.*

$$f_{\bar{z}}(z) = \mu(z)f_z(z). \tag{1.1}$$

The question of existence of a solution of (1.1) is addressed in the famous Ahlfors–Bers–Boyarskii theorem (see [3] and [6]) (below, $\hat{\mathbb{C}}$ signifies the Riemann sphere).

*Received by the editors September 19, 2005; accepted for publication (in revised form) May 30, 2007; published electronically May 9, 2008. This project was completed by the first author after D. Khmelev's sudden passing.

http://www.siam.org/journals/sinum/46-5/64071.html

†Institutionen för Matematik, Kgl Tekniska Högskolan, 100 44 Stockholm, Sweden (gaidash@math.kth.se).

THEOREM 1.2 (Ahlfors and Bers [3], Boyarskii [6]). *Let $\mu \in L_\infty(\hat{\mathbb{C}})$ satisfy $\|\mu\|_\infty \le k < 1$. Then there exists a unique map $f^\mu : \hat{\mathbb{C}} \to \hat{\mathbb{C}}$ such that f^μ satisfies the Beltrami equation $f^\mu_{\bar{z}}(z) = \mu(z) f^\mu_z(z)$, f^μ fixes 0, 1, and ∞, and f^μ is a $(1+k)/(1-k)$-quasi-conformal map.*

The goal of the present paper will be to study two algorithms for finding the normalized solution of the Beltrami equation numerically.

There seem to be three main techniques that have been used over time to solve a Beltrami equation with a compactly supported differential, or to find a quasi-conformal map of a simply or multiply connected domain to a "canonical" domain (disk, rectangle, etc.): finite differences, circle packings, and a numerical version of Ahlfors and Bers' original method, which we will later describe in greater detail.

Several works have proposed numerical constructions of quasi-conformal mappings by solving a system of elliptic equations with finite difference techniques. Reference [21] presents a finite difference construction of a quasi-conformal map of a simply connected domain in $\mathbb{C}$ onto a rectangle. Finite-difference techniques have been also used in [26] and [27] to find quasi-conformal maps from simply, doubly, and triply connected domains. One particular application is finite grid generation [4], [22].

Another method that has been used to solve the Beltrami equation numerically is circle packings. A circle packing—a collection of closed disks with disjoint interiors and prescribed tangencies—was suggested by W. Thurston for a geometric proof of the Riemann mapping theorem. The details of this proof have been worked out by Rodin and Sullivan in [24]. A generalization of this scheme to the Ahlfors–Bers–Boyarskii theorem has been given in [16] and [18]. A numerical implementation of this procedure was studied in [19].

In this paper we will study a third method: an algorithm first proposed in [11] and based on the original iterative scheme of Ahlfors and Bers which uses singular integral transforms.

According to Ahlfors and Bers' classical proof of Theorem 1.2, a solution of the Beltrami equation can be constructed with the help of two operators, the first operator being the Hilbert transform,

$$T[h](z) = \frac{i}{2\pi} \lim_{\epsilon \to 0} \int\int_{\mathbb{C}\setminus B(z,\epsilon)} \frac{h(\xi)}{(\xi - z)^2}\, d\bar{\xi} \wedge d\xi, \tag{1.2}$$

and the second the Cauchy transform,

$$P[h](z) = \frac{i}{2\pi} \int\int_{\mathbb{C}} \frac{h(\xi)}{\xi - z} - \frac{h(\xi)}{\xi}\, d\bar{\xi} \wedge d\xi. \tag{1.3}$$

If the differential μ is compactly supported, then the quasi-conformal homeomorphism f is constructed in two steps. First, one solves the following equation:

$$h^* = T[\mu h^*] + T[\mu]. \tag{1.4}$$

If h^* satisfying (1.4) is found, then f solving (1.1) on the Riemann sphere $\hat{\mathbb{C}}$, normalized as to fix 0, 1, and ∞, is given by

$$f(z) = P[\mu(h^* + 1)](z) + z.$$

The operator P is well-defined and is Hölder continuous for $h \in L^p(\mathbb{C})$, $p > 2$. This implies that a solution to (1.4) must also be constructed in $L_p(\mathbb{C})$, $p > 2$. The celebrated Calderon–Zygmund inequality (cf. [9]) states that

$$\|T[h]\|_p \le C_p \|h\|_p$$

for all $h \in L_p(\mathbb{C})$, where C_p is some constant, depending only on p, which can be chosen so that $C_p \to 1$ as $p \to 2$. While the optimal value of C_p for all p is not known, it is known that $C_2 = 1$, and T is an isometry in $L_2(\mathbb{C})$:

$$\|T[h]\|_2 = \|h\|_2.$$

To solve (1.4) one has to choose a p such that $kC_p < 1$. Then the operator $h \mapsto T[\mu h]$ is a contraction in $L_p(\mathbb{C})$ and iterations

$$h^{n+1} = T[\mu h^n] + T[\mu] \tag{1.5}$$

converge to a solution of (1.4): $h^n \to h^*$ as $n \to \infty$.

We can now state the extended version of the Ahlfors–Bers–Boyarskii theorem for compactly supported differentials.

THEOREM 1.3. *Let $\mu \in L_\infty(\mathbb{C})$ be compactly supported and satisfy $\|\mu\|_\infty \le k < 1$. Then for every $p > 2$ such that $C_p k < 1$, the operator $h \to T[\mu(h+1)]$ is a contraction on $L_p(\mathbb{C})$ with a unique fixed point h^*. Moreover, the solution of the Beltrami equation, f^μ, is given by*

$$f^\mu = P[\mu(h^* + 1)] + \mathrm{id}, \tag{1.6}$$

where id *is the identity function, and is such that $f(0) = 0$, f is continuous and has distributional derivatives, and $f_z - 1 \in L_p(\mathbb{C})$.*

We will concentrate on the numerical implementation of the procedure described in this theorem. This is an important practical problem which has been addressed by various authors.

Efficient algorithms for computing the two required transforms T and P have been proposed in [11], [12], and [13]. We will proceed with a very brief description of these algorithms.

As we have pointed out, the Cauchy transform of an L_p-function, $p > 2$, is well-defined and is known to be Hölder continuous with exponent $1 - 2/p$ (cf. [2], [20]). In [11] an effective algorithm for the numerical computation of the following integral transform was suggested:

$$\tilde{P}[h](z) = \frac{i}{2\pi} \int\int_{\mathbb{C}} \frac{h(\xi)}{(\xi - z)}\, d\bar{\xi} \wedge d\xi. \tag{1.7}$$

A comparison with (1.3) immediately shows that

$$P[h](z) = \tilde{P}[h](z) - \tilde{P}[h](0). \tag{1.8}$$

Represent h and $\tilde{P}[h]$ as a Fourier series:

$$h(re^{i\theta}) = \sum_{k=-\infty}^{\infty} h_k(r) e^{ik\theta}, \tag{1.9}$$

$$\tilde{P}[h](re^{i\theta}) = \sum_{k=-\infty}^{\infty} p_k(r) e^{ik\theta}, \tag{1.10}$$

where the coefficients of the $\tilde{P}$-transform are given by

$$p_k(r) = \frac{1}{2\pi} \int_0^{2\pi} e^{-ik\theta} \tilde{P}[h](re^{i\theta})\, d\theta. \tag{1.11}$$

Then, according to [11],

$$p_k(r) = \begin{cases} 2\displaystyle\int_0^r \left(\frac{r}{\rho}\right)^k h_{k+1}(\rho)d\rho, & k<0, \\ -2\displaystyle\int_r^\infty \left(\frac{r}{\rho}\right)^k h_{k+1}(\rho)d\rho, & k\geq 0. \end{cases} \tag{1.12}$$

Next, to obtain similar formulae for the Hilbert transform, assume that h is a Hölder continuous function compactly supported in an open disk around zero of radius R, $B(0,R)\subset\mathbb{C}$. The Hilbert transform of such a function is well-defined (see the next section). Represent the Hilbert transform as a Fourier series:

$$T[h](re^{i\theta}) = \sum_{k=-\infty}^{\infty} c_k(r)e^{ik\theta}, \quad c_k(r) = \frac{1}{2\pi}\int_0^{2\pi} e^{-ik\theta}T[h](re^{i\theta})\, d\theta. \tag{1.13}$$

In [12] and [13] the authors arrive at the following expressions for these coefficients:

$$c_0(0) = -2\lim_{\epsilon\to 0}\int_\epsilon^R \frac{h_2(\rho)}{\rho}d\rho, \text{ and } c_k(0)=0, \text{ whenever } k\neq 0, \tag{1.14}$$

$$c_k(r) = A_k\int_0^r \frac{r^k}{\rho^{k+1}}h_{k+2}(\rho)d\rho + B_k\int_r^R \frac{r^k}{\rho^{k+1}}h_{k+2}(\rho)d\rho + h_{k+2}(r), \tag{1.15}$$

where

$$A_k = \begin{cases} 0, & k\geq 0, \\ 2(k+1), & k<0, \end{cases} \quad \text{and} \quad B_k = \begin{cases} -2(k+1), & k\geq 0, \\ 0, & k<0. \end{cases} \tag{1.16}$$

Given values of h, for instance, on a circular $N\times M$ grid that contains the compact support of h, one can use a fast Fourier transform (FFT; cf. [23]) to find the values of the coefficients h_k at the radii r_i, $1\leq i\leq M$. Next, one can use these values to construct a piecewise constant, a piecewise linear, or a spline approximation of the functions h_k (the choice of approximation, of course, depends on the known or expected smoothness of h_k). This allows one to compute integrals in (1.12) and in (1.14)–(1.15). Armed with these implementations of the Hilbert and Cauchy transforms, one can try to solve the Beltrami equation (1.1), first by running iterations (1.5) for some time, and, finally, applying (1.6). It is convenient to use the pointwise multiplication of grid values of h and $\mu+1$ inside the Hilbert transform in (1.5), rather than the multiplication of their Fourier series: the order of the computational complexity of the pointwise multiplication is $O(NM)$, as opposed to $O(NM^2)$ for the series. The transition from the representation of h as a Fourier series to point values at each iteration step can be performed with the help of the FFT. This way, the computational complexity of one iteration step becomes $O(NM\log_2 M)$.

The objective of our paper is manifold. First, we will present a rigorous mathematical justification of the formulae (1.14)–(1.15) for the singular operator T different from that of [13]. Second, we will propose an algorithm which avoids programming formulae (1.14)–(1.15) altogether. It should be noted that in our experience writing the code for the Hilbert transform can be rather cumbersome; therefore this second algorithm can be viewed as an improvement.

Furthermore, we would like to reiterate the well-known fact that solutions of the Beltrami equation are often the central ingredient of many uniformization problems

appearing, in particular, in complex dynamics. In this light, the present paper can be also viewed as a detailed account of the numerical algorithms that have been used, in particular, in [14] and [15] to address an important issue in complex dynamics—conjectural universality for Siegel disks. The interested reader is referred to [14] for a description of a parallelized implementation of the algorithms of the present paper, which also includes rigorous computer-assisted bounds on the difference between a numerical approximation of the solution of the Beltrami equation with the true one.

2. An alternative justification of Daripa and Mashat's algorithm for the Hilbert transform. The Hilbert transform is a singular integral transform, and therefore, first of all, one has to specify the conditions under which this integral can be computed. The issue of existence of a rather general class of singular integral transforms (including the Hilbert transform) has been addressed in the classical work [9] of Calderon and Zygmund. The sufficient condition for the existence of the Hilbert transform itself can be stated in a very concise way (cf. [2], [10]).

LEMMA 2.1. *If h is Hölder continuous, then the Hilbert integral transform of h exists as a Cauchy principle value.*

In this light, the main result of [13] can be stated as follows.

THEOREM 2.1. *If $T[h](z)$ exists as a Cauchy principle value, then the Fourier coefficients of this Hilbert transform are given by formulae* (1.14)–(1.15).

The proof of this theorem, presented in [13], starts with an observation that according to (1.2) and (1.13), coefficients of the Hilbert transform are given by the following formula:

$$c_k(r) = -\frac{1}{2\pi^2}\int_0^{2\pi} e^{-ik\theta}\lim_{\epsilon\to 0}\int\int_{\mathbb{C}\setminus B(re^{i\theta},\epsilon)} \frac{h(\rho e^{i\alpha})}{(\rho e^{i\alpha}-re^{i\theta})^2}\ \rho d\rho d\alpha d\theta. \tag{2.1}$$

Daripa and Mashat proceed in their proof by computing the Hilbert transform in the problematic region (ρ close to r) explicitly. We will take a different route here: a crucial step in our proof will be to interchange the order of the first integration and the limit in (2.1).

We will now proceed with our argument. Let h be Hölder continuous with exponent γ and constant A, and let it be compactly supported in an open disk around zero of radius R, $B(0,R)\subset\mathbb{C}$. Notice that the variable z in (1.2) can range over all of $\mathbb{C}$ and not only over $B(0,R)$. We will start by showing that the Hilbert transform of a Hölder continuous function is bounded.

Choose a $\delta < 1/R$. For all $\epsilon < \delta R$ define

$$\begin{aligned} T_\epsilon[h](z) &= \frac{i}{2\pi}\int\int_{B(0,R)\setminus B(z,\epsilon)} \frac{h(\xi)}{(\xi-z)^2}\ d\bar{\xi}\wedge d\xi \\ &= \frac{i}{2\pi}\int\int_{B(0,R)\setminus B(z,\delta R)} \frac{h(\xi)}{(\xi-z)^2}\ d\bar{\xi}\wedge d\xi \\ &\quad + \frac{i}{2\pi}\int\int_{B(z,\delta R)\setminus B(z,\epsilon)} \frac{h(\xi)}{(\xi-z)^2}\ d\bar{\xi}\wedge d\xi \\ &= T^1[h](z) + T^2_\epsilon[h](z). \end{aligned} \tag{2.2}$$

LEMMA 2.2. *There exists a constant C such that $\left|T^2_\epsilon[h](z)\right| < C$ for all $0<\epsilon<\delta R$ and all $z\in\overline{B(0,R)}$.*

Proof.

$$\begin{aligned}
\left|T_\epsilon^2[h](z)\right| &= \left|\frac{i}{2\pi}\int\int_{B(z,\delta R)\setminus B(z,\epsilon)} \frac{h(\xi)}{(\xi-z)^2}\, d\bar{\xi}\wedge d\xi\right| \\
&= \left|-\frac{1}{\pi}\int_\epsilon^{\delta R}\int_0^{2\pi} \frac{h(z+\rho e^{i\theta})}{\rho^2 e^{2i\theta}}\rho\, d\rho d\theta\right| \\
&= \left|-\frac{1}{\pi}\int_\epsilon^{\delta R}\int_0^{\frac{\pi}{2}} \frac{h\left(z+\rho e^{i\theta}\right)-h\left(z+\rho e^{i(\theta+\pi/2)}\right)}{\rho e^{2i\theta}} d\rho d\theta\right. \\
&\quad \left.-\frac{1}{\pi}\int_\epsilon^{\delta R}\int_0^{\frac{\pi}{2}} \frac{h\left(z+\rho e^{i(\theta+\pi)}\right)-h\left(z+\rho e^{i(\theta+3\pi/2)}\right)}{\rho e^{2i\theta}} d\rho d\theta\right| \\
&\le \frac{1}{\pi}\int_\epsilon^{\delta R}\int_0^{\frac{\pi}{2}} \frac{\left|h\left(z+\rho e^{i\theta}\right)-h\left(z+\rho e^{i(\theta+\pi/2)}\right)\right|}{\rho}\, d\rho d\theta \\
&\quad +\frac{1}{\pi}\int_\epsilon^{\delta R}\int_0^{\frac{\pi}{2}} \frac{\left|h\left(z+\rho e^{i(\theta+\pi)}\right)-h\left(z+\rho e^{i(\theta+3\pi/2)}\right)\right|}{\rho}\, d\rho d\theta \\
&\le \frac{1}{2}\int_\epsilon^{\delta R} \frac{A\rho^\gamma\left|1-e^{i\frac{\pi}{2}}\right|^\gamma}{\rho}\, d\rho + \frac{1}{2}\int_\epsilon^{\delta R} \frac{A\rho^\gamma\left|1-e^{i\frac{\pi}{2}}\right|^\gamma}{\rho}\, d\rho \\
&\le A\left|1-i\right|^\gamma \frac{1}{\gamma}(\delta R)^\gamma = C,
\end{aligned}$$

where we have used Hölder continuity. ∎

Next, we will derive the expressions for the Fourier coefficients of the Hilbert transform separately for $z \in \mathrm{int}B(0,R)-\{0\}$, $z \in \partial B(0,R)$, $z=0$, and $z \in \mathbb{C}\setminus\overline{B(0,R)}$.

2.1. Case $z \in \mathbf{int}B(0,R)-\{0\}$.

COROLLARY 2.2.

$$\begin{aligned}
c_k(r) &= -\frac{1}{2\pi^2}\lim_{\epsilon\to 0}\int_0^{r-\epsilon}\int_0^{2\pi} h(\rho e^{i\alpha})\int_0^{2\pi} \frac{e^{-ik\theta}}{(\rho e^{i\alpha}-re^{i\theta})^2}\, d\theta d\alpha\rho d\rho \\
&\quad -\frac{1}{2\pi^2}\lim_{\epsilon\to 0}\int_{r+\epsilon}^{R}\int_0^{2\pi} h(\rho e^{i\alpha})\int_0^{2\pi} \frac{e^{-ik\theta}}{(\rho e^{i\alpha}-re^{i\theta})^2}\, d\theta d\alpha\rho d\rho + h_{k+2}(r)
\end{aligned}$$

for all $0<r<R$.

Proof. Let $z=re^{i\theta}$. Recall that $T_\epsilon[h]=T^1[h]+T_\epsilon^2[h]$, where $T^1[h]$ and $T_\epsilon^2[h]$ are as in (2.2). The first of these operators is clearly bounded, and the second is bounded by the preceding lemma. Therefore, $|T_\epsilon[h](z)|$ is bounded by a constant function, while the function $\psi_\epsilon(\theta)=e^{-i\theta}T_\epsilon[h](re^{i\theta})$ is bounded by $\mathrm{const}\cdot e^{-i\theta}$, which is certainly in $L_1([0,2\pi))$. A direct application of the Lebesgue dominated convergence theorem shows that the limiting procedure and the integration over the circle can be interchanged:

$$\begin{aligned}
c_k(r) &= \frac{i}{4\pi^2}\int_0^{2\pi}\lim_{\epsilon\to 0}\psi_\epsilon(\theta)d\theta = \frac{i}{4\pi^2}\lim_{\epsilon\to 0}\int_0^{2\pi}\psi_\epsilon(\theta)d\theta \\
&= \frac{i}{4\pi^2}\lim_{\epsilon\to 0}\int_0^{2\pi}\int\int_{B(0,R)\setminus B(z,\epsilon)} h(\xi)\frac{e^{-ik\theta}}{(\xi-re^{i\theta})^2}\, d\bar{\xi}\wedge d\xi d\theta = \lim_{\epsilon\to 0} c_k^\epsilon(r).
\end{aligned}$$

We will denote $\Omega = B(0, r+\epsilon) \setminus (B(0, r-\epsilon) \cup B(z, \epsilon))$ for the purpose of notational brevity. Then, for an arbitrary but fixed ϵ, one can interchange the order of integration (according to Fubini's theorem) in the following way:

$$\begin{aligned}
c_k^\epsilon(r) &= \frac{i}{4\pi^2}\int_0^{2\pi}\int\int_{B(0,r-\epsilon)} h(\xi)\frac{e^{-ik\theta}}{(\xi - re^{i\theta})^2}\, d\bar{\xi}\wedge d\xi d\theta \\
&+ \frac{i}{4\pi^2}\int_0^{2\pi}\int\int_{B(0,R)\setminus B(0,r+\epsilon)} h(\xi)\frac{e^{-ik\theta}}{(\xi - re^{i\theta})^2}\, d\bar{\xi}\wedge d\xi d\theta \\
&+ \frac{i}{4\pi^2}\int_0^{2\pi}\int\int_{\Omega} h(\xi)\frac{e^{-ik\theta}}{(\xi - re^{i\theta})^2}\, d\bar{\xi}\wedge d\xi d\theta \\
&= -\frac{1}{2\pi^2}\int_0^{r-\epsilon}\int_0^{2\pi} h(\rho e^{i\alpha})\int_0^{2\pi}\frac{e^{-ik\theta}}{(\rho e^{i\alpha} - re^{i\theta})^2}\, d\theta d\alpha\rho d\rho \\
&- \frac{1}{2\pi^2}\int_{r+\epsilon}^{R}\int_0^{2\pi} h(\rho e^{i\alpha})\int_0^{2\pi}\frac{e^{-ik\theta}}{(\rho e^{i\alpha} - re^{i\theta})^2}\, d\theta d\alpha\rho d\rho \\
&+ \frac{i}{4\pi^2}\int_0^{2\pi}\int\int_{\Omega} h(\xi)\frac{e^{-ik\theta}}{(\xi - re^{i\theta})^2}\, d\bar{\xi}\wedge d\xi d\theta. \qquad (2.3)
\end{aligned}$$

Next, define $\Omega_2 = B(z, \epsilon^\beta) \setminus B(z, \epsilon)$ for some real positive $\beta < 1$ to be specified later ($B(z,\epsilon) \subset B(z,\epsilon^\beta)$ since ϵ is always less than 1 by our choice of δ), and $\Omega_1 = \Omega \setminus \Omega_2$. Using Hölder continuity once again, one can show that the difference

$$I_\epsilon = \left| \frac{i}{4\pi^2}\int_0^{2\pi}\int\int_{\Omega}\frac{h(\xi)e^{-ik\theta}}{(\xi - z)^2}\, d\bar{\xi}\wedge d\xi d\theta - \frac{i}{4\pi^2}\int_0^{2\pi}\int\int_{\Omega}\frac{h(z)e^{-ik\theta}}{(\xi - z)^2}\, d\bar{\xi}\wedge d\xi d\theta \right| \qquad (2.4)$$

can be bounded in the following way:

$$\begin{aligned}
I_\epsilon &\le \frac{1}{4\pi^2}\int_0^{2\pi}\int\int_{\Omega}\frac{|h(\xi) - h(z)|}{|\xi - z|^2}\, d\bar{\xi}\wedge d\xi d\theta \\
&\le \frac{1}{4\pi^2}\int_0^{2\pi}\int\int_{\Omega} A|\xi - z|^{\gamma-2}\, d\bar{\xi}\wedge d\xi d\theta \\
&\le \frac{A}{2\pi}\int\int_{\Omega}\rho^{\gamma-2}\, d\alpha\rho d\rho.
\end{aligned}$$

If $\gamma \ge 1$, then I_ϵ can be clearly bounded by

$$I_\epsilon \le \frac{A}{2\pi}\epsilon^{\gamma-2}\mathrm{Area}(\Omega) \le \frac{A}{2\pi}\epsilon^{\gamma-2}4\pi r\epsilon = 2Ar\epsilon^{\gamma-1}. \qquad (2.5)$$

Otherwise

$$\begin{aligned}
I_\epsilon &\le \frac{A}{2\pi}\int\int_{\Omega_1}\rho^{\gamma-2}\, d\alpha\rho d\rho + \frac{A}{2\pi}\int\int_{\Omega_2}\rho^{\gamma-2}\, d\alpha\rho d\rho \\
&\le \frac{A}{2\pi}\epsilon^{\beta(\gamma-2)}\mathrm{Area}(\Omega_1) + \frac{A}{\gamma}(\epsilon^{\beta\gamma} - \epsilon^\gamma) \le \frac{A}{2\pi}\epsilon^{\beta(\gamma-2)}4\pi r\epsilon + \frac{A}{\gamma}(\epsilon^{\beta\gamma} - \epsilon^\gamma) \\
&\le 2Ar\epsilon^{\beta(\gamma-2)+1} + \frac{A}{\gamma}(\epsilon^{\beta\gamma} - \epsilon^\gamma).
\end{aligned}$$

Notice that in both cases

$$\lim_{\epsilon\to 0} I_\epsilon = 0 \tag{2.6}$$

as long as $\beta(2-\gamma)<1$. Now, it is not hard to show that

$$\frac{i}{4\pi^2}\int\int_\Omega \frac{d\bar{\xi}\wedge d\xi}{(\xi-z)^2} = \frac{1}{2\pi}\frac{(r-\epsilon)^2}{z^2} \tag{2.7}$$

(see [13]) and, therefore, combining (2.6) and (2.7), that

$$\begin{aligned}
&\left|\frac{i}{4\pi^2}\lim_{\epsilon\to 0}\int_0^{2\pi}\int\int_\Omega \frac{h(\xi)e^{-ik\theta}}{(\xi-z)^2}\, d\bar{\xi}\wedge d\xi d\theta - \frac{1}{2\pi}\int_0^{2\pi} h(z)e^{-ik\theta}\frac{r^2}{z^2}\, d\theta\right| \\
&= \left|\frac{i}{4\pi^2}\lim_{\epsilon\to 0}\int_0^{2\pi}\int\int_\Omega \frac{h(\xi)e^{-ik\theta}}{(\xi-z)^2}\, d\bar{\xi}\wedge dd\xi d\theta - \frac{1}{2\pi}\lim_{\epsilon\to 0}\int_0^{2\pi} h(z)e^{-ik\theta}\frac{(r-\epsilon)^2}{z^2}\, d\theta\right| \\
&= \lim_{\epsilon\to 0}\left|\frac{i}{4\pi^2}\int_0^{2\pi}\int\int_\Omega \frac{h(\xi)e^{-ik\theta}}{(\xi-z)^2}\, d\bar{\xi}\wedge d\xi d\theta - \frac{1}{2\pi}\int_0^{2\pi} h(z)e^{-ik\theta}\frac{(r-\epsilon)^2}{z^2}\, d\theta\right| \\
&= \lim_{\epsilon\to 0}\left|\frac{i}{4\pi^2}\int_0^{2\pi}\int\int_\Omega \frac{h(\xi)e^{-ik\theta}}{(\xi-z)^2}\, d\bar{\xi}\wedge d\xi d\theta - \frac{i}{4\pi^2}\int_0^{2\pi}\int\int_\Omega \frac{h(z)e^{-ik\theta}}{(\xi-z)^2}\, d\bar{\xi}\wedge d\xi d\theta\right| \\
&\le \lim_{\epsilon\to 0} I_\epsilon = 0,
\end{aligned}$$

where in the first equality we again used the Lebesgue dominated convergence theorem. We conclude that

$$\frac{i}{4\pi^2}\lim_{\epsilon\to 0}\int_0^{2\pi}\int\int_\Omega \frac{h(\xi)e^{-ik\theta}}{(\xi-z)^2}\, d\bar{\xi}\wedge d\xi d\theta = \frac{1}{2\pi}\int_0^{2\pi} h(z)e^{-ik\theta}\frac{r^2}{z^2}\, d\theta. \tag{2.8}$$

Observe that

$$\frac{1}{2\pi}\int_0^{2\pi} h(z)e^{-ik\theta}\frac{r^2}{z^2}\, d\theta = \frac{1}{2\pi}\int_0^{2\pi}\frac{e^{-ik\theta}\sum_{n=-\infty}^{\infty} h_n(r)e^{in\theta}}{e^{2i\theta}}\, d\theta = h_{k+2}(r). \tag{2.9}$$

The claim of the corollary follows from (2.3), (2.8), and (2.9). □

Now, consider

$$I(\xi,r) = \int_0^{2\pi}\frac{e^{-ik\theta}}{(\xi-re^{i\theta})^2}\, d\theta = \frac{1}{r^2}\int_0^{2\pi}\frac{e^{-ik\theta}}{(\xi/r-e^{i\theta})^2}\, d\theta = \frac{1}{ir^2}\oint_{\mathbb{T}^1}\frac{1}{w^{k+1}(\xi/r-w)^2}\, dw.$$

We will first compute this integral in the case when $|\xi|/r>1$:

$$I(\xi,r) = \frac{2\pi}{r^2}\mathrm{Res}\left[\frac{1}{w^{k+1}(\xi/r-w)^2}\right]_{w=0} = -B_k\pi\frac{r^k}{\xi^{k+2}},$$

where

$$B_k = \begin{cases} -2(k+1), & k\ge 0, \\ 0, & k<0. \end{cases}$$

Similarly, for $|\xi|/r < 1$,

$$I(\xi,r) = \frac{2\pi}{r^2}\mathrm{Res}\left[\frac{1}{w^{k+1}(\xi/r-w)^2}\right]_{w=\frac{\xi}{r}} + \frac{2\pi}{r^2}\mathrm{Res}\left[\frac{1}{w^{k+1}(\xi/r-w)^2}\right]_{w=0} = -A_k\pi\frac{r^k}{\xi^{k+2}},$$

where

$$A_k = \begin{cases} 0, & k \geq 0, \\ 2(k+1), & k < 0. \end{cases}$$

Finally, we obtain

$$\begin{aligned} c_k(r) &= \frac{A_k}{2\pi}\int_0^r\int_0^{2\pi} h(\rho e^{i\alpha})\frac{r^k}{\xi^{k+2}}\ \rho d\rho d\alpha + \frac{B_k}{2\pi}\int_r^R\int_0^{2\pi} h(\rho e^{i\alpha})\frac{r^k}{\xi^{k+2}}\ \rho d\rho d\alpha + h_{k+2}(r) \\ &= \frac{A_k}{2\pi}\int_0^r\int_0^{2\pi} \frac{r^k\sum_{-\infty}^{\infty} h_n(\rho)e^{in\alpha}}{\rho^{k+1}e^{i(k+2)\alpha}}\ d\rho d\alpha + \frac{B_k}{2\pi}\int_r^R\int_0^{2\pi} \frac{r^k\sum_{-\infty}^{\infty} h_n(\rho)e^{in\alpha}}{\rho^{k+1}e^{i(k+2)\alpha}}\ d\rho d\alpha \\ &+ h_{k+2}(r) = A_k\int_0^r \frac{r^k}{\rho^{k+1}}h_{k+2}(\rho)d\rho + B_k\int_r^R \frac{r^k}{\rho^{k+1}}h_{k+2}(\rho)d\rho + h_{k+2}(r). \end{aligned}$$

2.2. Case $z \in \partial B(0,R)$. We will briefly outline a counterpart of Corollary 2.2 for $|z| = R$.

Corollary 2.3.

$$c_k(R) = -\frac{1}{2\pi^2}\lim_{\epsilon\to 0}\int_0^{R-\epsilon}\int_0^{2\pi} h(\rho e^{i\alpha})\int_0^{2\pi}\frac{e^{-ik\theta}}{(\rho e^{i\alpha}-Re^{i\theta})^2}\ d\theta d\alpha\rho d\rho.$$

Proof. According to Lemma 2.2, we can exchange the order of the integration and the limit:

$$\begin{aligned} c_k(R) &= \frac{i}{4\pi^2}\int_0^{2\pi}\lim_{\epsilon\to 0}\int\int_{B(0,R)\setminus B(z,\epsilon)} h(\xi)\frac{e^{-ik\theta}}{(\xi-Re^{i\theta})^2}\ d\bar{\xi}\wedge d\xi d\theta \\ &= \frac{i}{4\pi^2}\lim_{\epsilon\to 0}\int_0^{2\pi}\int\int_{B(0,R)\setminus B(z,\epsilon)} h(\xi)\frac{e^{-ik\theta}}{(\xi-Re^{i\theta})^2}\ d\bar{\xi}\wedge d\xi d\theta = \lim_{\epsilon\to 0} c_k^{\epsilon}(R). \end{aligned}$$

Denote $\Omega = B(0,R)\setminus(B(0,R-\epsilon)\cup B(z,\epsilon))$. Then we obtain for an arbitrary but fixed ϵ:

$$\begin{aligned} c_k^{\epsilon}(R) &= -\frac{1}{2\pi^2}\int_0^{R-\epsilon}\int_0^{2\pi} h(\rho e^{i\alpha})\int_0^{2\pi}\frac{e^{-ik\theta}}{(\rho e^{i\alpha}-Re^{i\theta})^2}\ d\theta d\alpha\rho d\rho \\ &+ \frac{i}{4\pi^2}\int_0^{2\pi}\int\int_{\Omega} h(\xi)\frac{e^{-ik\theta}}{(\xi-Re^{i\theta})^2}\ d\bar{\xi}\wedge d\xi d\theta. \end{aligned} \tag{2.10}$$

Again, define $\Omega_2 = B(z,\epsilon^{\beta})\setminus B(z,\epsilon)$ for some real positive $\beta < 1$, and $\Omega_1 = \Omega\setminus\Omega_2$. As in Corollary 2.2, one shows that the difference (2.4) can be bounded by $Ar\epsilon^{\gamma-1}$ if $\gamma \geq 1$, and by $Ar\epsilon^{\beta(\gamma-2)+1} + (A/2\gamma)(\epsilon^{\beta\gamma}-\epsilon^{\gamma})$ otherwise, and that $\lim_{\epsilon\to 0} I_{\epsilon} = 0$ as long as $\beta(2-\gamma) < 1$. However, $h(z) \equiv 0$ for all $z \in \partial B(0,R)$; hence

$$I_{\epsilon} = \left|\frac{i}{4\pi^2}\int_0^{2\pi}\int\int_{\Omega}\frac{h(\xi)e^{-ik\theta}}{(\xi-z)^2}\ d\bar{\xi}\wedge d\xi d\theta\right|$$

and therefore

$$\frac{i}{4\pi^2}\lim_{\epsilon\to 0}\int_0^{2\pi}\int\!\!\int_{\Omega} h(\xi)\frac{e^{-ik\theta}}{(\xi-Re^{i\theta})^2}\,d\bar{\xi}\wedge d\xi d\theta = 0. \tag{2.11}$$

The claim of the lemma follows from (2.10) and (2.11). □

To obtain an expression for $c_k(R)$, consider the integral $\int_0^{2\pi} e^{-ik\theta}/(\xi - Re^{i\theta})^2\,d\theta$. It is not hard to show that this integral is equal to $-A_k\pi R^k/\xi^{k+2}$, and, therefore,

$$\begin{aligned}
c_k(R) &= -\frac{1}{2\pi^2}\lim_{\epsilon\to 0}\int_0^{R-\epsilon}\int_0^{2\pi} h(\rho e^{i\alpha})\int_0^{2\pi}\frac{e^{-ik\theta}}{(\rho e^{i\alpha}-Re^{i\theta})^2}\,d\theta d\alpha\rho d\rho \\
&= \frac{1}{2\pi}\int_0^R\int_0^{2\pi} h(\rho e^{i\alpha})A_k\frac{R^k}{\xi^{k+2}}\,d\alpha\rho d\rho \\
&= \frac{A_k}{2\pi}\int_0^R\int_0^{2\pi}\frac{R^k\sum_{n=-\infty}^{\infty}h_n(\rho)e^{in\theta}}{\rho^{k+1}e^{i(k+2)\theta}}\,d\rho d\theta \\
&= A_k\int_0^R\frac{R^k}{\rho^{k+1}}h_{k+2}(\rho)\,d\rho.
\end{aligned}$$

2.3. Case $z = 0$. We will prove a counterpart of Corollary 2.2, at the same time obtaining an expression for $c_k(0)$.

COROLLARY 2.4.

$$c_k(0) = \begin{cases} 0, & k\neq 0, \\ -2\int_0^R \frac{h_2(\rho)}{\rho}d\rho, & k = 0. \end{cases}$$

Proof. As before, according to Lemma 2.2, for an arbitrary but fixed ϵ, one can use the Lebesgue dominated convergence theorem to change the order of the integration and the limit:

$$\begin{aligned}
c_k(0) &= \frac{i}{4\pi^2}\int_0^{2\pi}\lim_{\epsilon\to 0}\int\!\!\int_{B(0,R)\setminus B(0,\epsilon)} h(\xi)\frac{e^{-ik\theta}}{\xi^2}\,d\bar{\xi}\wedge d\xi d\theta \\
&= \frac{i}{4\pi^2}\lim_{\epsilon\to 0}\int_0^{2\pi}\int\!\!\int_{B(0,R)\setminus B(0,\epsilon)} h(\xi)\frac{e^{-ik\theta}}{\xi^2}\,d\bar{\xi}\wedge d\xi d\theta \\
&= -\frac{1}{2\pi^2}\lim_{\epsilon\to 0}\int_\epsilon^R\int_0^{2\pi}\frac{h(\rho e^{i\alpha})}{\rho^2 e^{2i\alpha}}\int_0^{2\pi} e^{-ik\theta}\,d\theta d\alpha\rho d\rho \\
&= -\frac{1}{2\pi^2}\int_0^R\int_0^{2\pi}\frac{\sum_{n=-\infty}^{n=\infty}h_n(\rho)e^{in\alpha}}{\rho^2 e^{2i\alpha}}\int_0^{2\pi} e^{-ik\theta}\,d\theta d\alpha\rho d\rho \\
&= \begin{cases} 0, & k\neq 0, \\ -2\int_0^R \frac{h_2(\rho)}{\rho}d\rho, & k = 0. \end{cases}
\end{aligned}$$

The Hölder property of h and the fact that $h_2(0) = 0$ show that the last integral ($k = 0$) is convergent. □

2.4. Case $z \in \mathbb{C} \setminus \overline{B(0,R)}$. In this case the limiting procedure and the angular integration in (2.1) can be interchanged right away.

$$\int_0^{2\pi} \frac{e^{-ik\theta}}{(\xi - re^{i\theta})^2}\, d\theta = \frac{2\pi}{r^2}\mathrm{Res}\left[\frac{1}{w^{k+1}(\xi/r - w)^2}\right]_{w=0} = -A_k \pi \frac{r^k}{\xi^{k+2}}$$

and

$$\begin{aligned} c_k(r) &= \frac{i}{4\pi^2}\int\int_{B(0,R)} h(\xi) A_k \frac{r^k}{\xi^{k+2}}\, d\bar{\xi}\wedge d\xi \\ &= \frac{A_k}{2\pi}\int\int_{B(0,R)} \sum_{n=-\infty}^{\infty} h_n(\rho) e^{in\theta}\frac{r^k}{\xi^{k+2}}\, \rho d\rho d\theta \\ &= A_k \int_0^R \frac{r^k}{\rho^{k+1}} h_{k+2}(\rho)\, d\rho. \end{aligned}$$

3. Recursive formulae. Define

$$d_k(r,r') = 2(k+1)\int_r^{r'} \frac{1}{\rho}\left(\frac{r'}{\rho}\right)^k h_{k+2}(\rho)\, d\rho.$$

Then, using (1.14)–(1.16), one readily gets the following recursive formula for the coefficients of the Hilbert transform,

$$c_k(r) = \left(\frac{r}{r'}\right)^k (c_k(r') - h_{k+2}(r') - d_k(r,r')) + h_{k+2}(r), \tag{3.1}$$

which holds for any pair of nonnegative radii r and r' for which the ratio $(r/r')^k$ is defined.

We will also note that it is convenient to compute $p_k(r)$ in (1.12) using the following recursive relations:

$$p_k(r) = \left(\frac{r}{r'}\right)^k p_k(r') - e_k(r,r'), \qquad e_k(r,r') = 2\int_r^{r'} \left(\frac{r}{\rho}\right)^k h_{k+1}(\rho) d\rho. \tag{3.2}$$

These recursive formulae are extremely useful because they allow us to compute the value of a coefficient at the ith radius using its value at either the $(i+1)$th or the $(i-1)$th radius, thus avoiding the full integration in formulae (1.12) and (1.14)–(1.16).

For reference purposes we will rewrite formulae (3.1) and (3.2) for a piecewise-constant interpolation of data $h_k(r_i)$ in the appendix (section 6). Strictly speaking, our derivation of the expressions for the Fourier coefficients of the Hilbert transform does not hold for a discontinuous h with piecewise-constant coefficients. However, given an $\epsilon > 0$ there always is a Hölder continuous function whose coefficients are ϵ-close to $h_k(r)$ in the L_1 norm, and, therefore, there always is a Hölder continuous smoothing of piecewise-constant data such that the coefficients of its Hilbert transform are arbitrarily close to those given by expressions (6.1)–(6.3) in the sup norm.

4. Another way to compute T. In this section we propose an alternative way for computing $T[h]$. By a theorem presented in [2], the following identity holds for all $h \in C_0^2$ (twice differentiable h with a compact support):

$$T[h](z) = \partial_z P[h](z) = \partial_z \tilde{P}[h](z). \tag{4.1}$$

As we have already mentioned, in [12] Daripa suggested an effective algorithm for a numerical computation of integral transform (1.7). The algorithm in [12] for computing $\tilde{P}$ is defined in terms of a circular $N \times M$ grid, and its computational complexity is $O(NM \log_2 M)$. In order to use this algorithm it is necessary to find an approximation for h_z, given the values of h on a circular grid. A formula for h_z in polar coordinates is supplied in the following lemma, whose easy proof will be omitted.

LEMMA 4.1. *Suppose that f is $C^1(\mathbb{R}^2)$, and in polar coordinates f is represented by g, i.e., $f(x,y) = g(r,\theta)$, where $x + iy = re^{i\theta}$. Then*

$$f_z(x,y) = \frac{e^{-i\theta}}{2}(g_r(r,\theta) + g_\theta(r,\theta)/ir), \tag{4.2}$$

$$f_{\bar{z}}(x,y) = \frac{ie^{i\theta}}{2}(-ig_r(r,\theta) + g_\theta(r,\theta)/r). \tag{4.3}$$

5. Comparison of the two algorithms. We have implemented the algorithm for the Cauchy transform (see formulae (1.7)–(1.12) and (3.2)) and both algorithms for the Hilbert transform ((1.13)–(1.16), (3.1), and (4.1)) for the piecewise linear approximation of Fourier coefficients as a set of routines in the programming language Ada 95 (cf. [25] for the language standard). We have used the public version 3.15p of the GNAT compiler [1]. Our programs can be found at [17].

Of particular interest to us will be a comparison of the two implementations of the iteration scheme $h \mapsto T[\mu(h+1)]$:

(1) *In the first of them, the Hilbert transform has been computed through formula* (3.1). *This technique will be referred to as* Scheme 1.
(2) *In the other, the Hilbert transform is calculated as the z-derivative of transform* (1.7). *This method will be called* Scheme 2.

The prototypical Beltrami differential for algorithm testing in the literature seems to be $\mu(z) = z^p \bar{z}^q$ in some disk and $\mu(z) = 0$ outside of that disk. Both Hilbert and Cauchy transforms of this differential, as well as the solution of the Beltrami equation itself, can be found exactly. Here we will quote the formula for the Hilbert transform:

$$T[z^p \bar{z}^q] = \frac{p\,\eta(R-|z|)}{q+1} z^{p-1} \bar{z}^{q+1} + [\eta(q+1-p) - \eta(R-|z|)]\,\frac{p-q-1}{q+1} R^{2(q+1)} z^{p-q-2}, \tag{5.1}$$

where η is the Heaviside step function. The derivation of this formula is relatively straightforward and can be done either by direct calculation or by differentiating the expression for the Cauchy transform of $z^p \bar{z}^q$ which have appeared in several publications (cf., e.g., [11]).

This formula suggests yet one more algorithm for the Hilbert transform of C_0^k functions representable as finite power series on its compact support:

(3) *Let $h(z,\bar{z}) = \eta(R-|z|) \sum_{i,j=0}^{m,n} a_{i,j} z^i \bar{z}^j$ be such a function. Then one can first use the linearity of the Hilbert transform to represent $T[h]$ as*

$$T[h](z,\bar{z}) = \sum_{i,j=0}^{m,n} a_{i,j} T[z^i \bar{z}^j],$$

compute each $T[z^i \bar{z}^j]$ in this series using (5.1), *and sum up. We will refer to this procedure as* Scheme 3.

TABLE 1
Convergence rate ($\|h^n - h^{n-1}\|_\infty$) for the two iteration schemes.

Number of iterations, n	5	10	20
Scheme 1	1.17×10^{-6}	3.83×10^{-11}	1.80×10^{-14}
Scheme 2	1.20×10^{-6}	3.91×10^{-10}	6.79×10^{-14}

Of course, this scheme has a limited value, as many differentials do not have sufficient smoothness, or are sometimes given in numerical experiments simply as an array of values on a grid.

Here, we will use a quartic test Beltrami differential which readily yields itself to the procedure that we just described:

$$\mu(z) = \begin{cases} a - \frac{a}{s^2} z\bar{z} + \frac{a}{4s^4} z^2 \bar{z}^2, & |z| \le \sqrt{2}s, \\ 0, & |z| > \sqrt{2}s, \end{cases} \tag{5.2}$$

where a and s are some real positive parameters. Our μ is given by a real-valued, nonnegative, radially symmetric C_0^2 function on $\mathbb{C}$ with a single maximum at zero, compactly supported in the disk of radius $\sqrt{2}s$. We would like to emphasize that iterations $h^{n+1} \mapsto T[h^n(\mu+1)]$ can be performed *exactly* using Scheme 3 (that is, the only error is the rounding error of our computing device). In particular, there is no need for any approximation of the functions being transformed, as it was the case with the choice of the spline approximation of the coefficients h_k in formulae (1.14)–(1.15). We will use Scheme 3 as a measuring stick for the accuracy of Schemes 1 and 2.

Let $g(r,\theta) = \sum_k g_k(r) e^{ik\theta}$ be smooth. Then it follows from (4.2) that the Fourier coefficients of the z-derivative of this function, denoted by $\tilde{g}$, are given by the following exact expression:

$$\tilde{g}_{k-1}(r) = \frac{g_k'(r) + k g_k(r)/r}{2}. \tag{5.3}$$

We have used this formula to compute the coefficients of the z-derivative of the iteration $h \mapsto \partial_z P[\mu(h+1)]$ for our choice (5.2) of μ; $g_k'(r)$'s were supplied by right two-point approximations through values of $g_k(r)$ on the radial grid.

We should also mention that we have used the FFT together with the pointwise multiplication of function values on the grid to perform the multiplication of μ and h in both schemes. Indeed, the computational cost of the FFT is of order $O(NM \log_2 M)$, and that of the pointwise multiplication of functional values is $O(NM)$; at the same time, the complexity of the multiplication of two Fourier series representing μ and h would be $O(NM^2)$.

We subjected the two schemes to several tests on a 1.5 GHz Intel Pentium class PC. Our first test was that of convergence of the two iteration schemes. Table 1 compares the convergence rate (measured by the norm $\|h^n - h^{n-1}\|_\infty$ rounded to 3 significant figures) for both iteration schemes and several values of n. We have used zero initial conditions, $h^0 = \mu$ in both cases. We have fixed the values of the parameters in the Beltrami differential to be $a = 0.5$ and $s = 1.0$. The dimensions of the radial grid were $N = 500$ (i.e., the disk of radius $\sqrt{2}$ contained 500 radii of the grid) and $M = 512$.

It is not surprising that the convergence rates of the schemes are comparable: these rates should depend only on the value of the essential supremum of μ and not on the particular realization of the Hilbert transform.

TABLE 2
Execution time for several grids (in seconds).

$N \times M$	500×256	1000×512	5000×1024	7000×2048
Scheme 1	8	31	313	965
Scheme 2	10	38	375	1166

TABLE 3
Comparison of the performance of Schemes 1 *and* 2 *against Scheme* 3.

$N \times M$	500×256	1000×512	5000×1024	7000×2048
$\|h_1^{10} - h_3^{10}\|_\infty$	4.00×10^{-7}	1.00×10^{-7}	7.12×10^{-8}	1.89×10^{-7}
$\|h_2^{10} - h_3^{10}\|_\infty$	8.39×10^{-5}	4.20×10^{-5}	8.39×10^{-6}	5.99×10^{-6}

Next, we have measured the execution time (rounded to the nearest second) of 10 iteration steps for several grids. Table 2 displays the results (the values of a, s, and h^0 are as before).

As should be expected, Scheme 2 takes somewhat longer: the computational complexity of both integral transforms is the same, $O(NM \log_2 M)$; however, the iteration $h \mapsto \partial_z P[\mu(h+1)]$ also involves a differentiation whose computational complexity is $O(NM)$. Nevertheless, for the tested range of grids, the speed benefits of Scheme 1 probably do not justify an extra effort in programming the Hilbert transform: as we have already mentioned, in our experience writing and testing such code might be rather cumbersome and lengthy, and could significantly increase the project development time.

The l_∞ norm of the difference of the outcomes of Schemes 1 and 3 and Schemes 2 and 3 after 10 steps (with the same parameters and initial conditions as before) for several grids is reported in Table 3. As we have already explained, Scheme 3 serves as a measuring stick for our purposes.

With regards to the accuracy of Schemes 1 and 2, Table 3 demonstrates that Scheme 1 is the scheme of choice. A bigger error in Scheme 2 can be attributed to the fact that we are using a numerical approximation for the differentiation in the iteration $h \mapsto \partial_z P[\mu(h+1)]$. This approximation introduces an extra error which is absent in Scheme 1. However, if the degree of differentiability of h is known, one can use a proper spline approximation of the derivative and thus significantly reduce this error. We still feel that in those cases when a high accuracy is not required, Scheme 2 can be more preferable.

6. Appendix. We will now return to formula (3.1). Consider a collection of real numbers $0 = r_1 < r_2 < \cdots < r_{L-1} < r_L = R < r_{L+1} < \cdots < r_N$ that specify a radial grid. We will denote the midpoint of the interval $[r_i, r_{i+1}]$ by ϱ_i; ϱ_0 will be 0 by definition.

If h_k's are compactly supported *piecewise-constant* functions

$$h_k(r) = \begin{cases} h_{k,i}, & r \in [\varrho_{i-1}, \varrho_i), \quad 1 \le i \le L, \\ 0, & r > \varrho_{L+1}, \end{cases}$$

then the coefficients of the Cauchy and the Hilbert transforms are given by the following set of expressions.

Coefficients of the Hilbert transform for piecewise-constant data.

• $r = r_1$:

$$c_k(r_1) = \begin{cases} c_0(r_2) - s_2(r_2) - 2s_2(r_2)\ln\frac{r_2}{\varrho_0}, & k = 0, \\ 0, & k \neq 0. \end{cases} \tag{6.1}$$

• $r \in (\varrho_{i-1}, \varrho_i]$, $1 \le i \le L$:

$$c_k(r) = \begin{cases} \left(\frac{r}{r_{i+1}}\right)^k (c_k(r_{i+1}) - s_{k+2}(r_{i+1})) - d_{k,i}(r) + s_{k+2}(r), & k \ge 1, \\ c_0(r_{i+1}) - s_2(r_{i+1}) - 2s_2(r_{i+1})\ln\frac{r_{i+1}}{\varrho_i} - 2s_2(r_i)\ln\frac{\varrho_i}{r} + s_2(r), & k = 0, \\ \left(\frac{r}{r_{i-1}}\right)^k (c_k(r_{i-1}) - s_{k+2}(r_{i-1})) + b_{k,i}(r) + s_{k+2}(r), & k < 0, \end{cases} \tag{6.2}$$

where

$$d_{k,i}(r) = 2\frac{k+1}{k}\left[s_{k+2}(r_{i+1})\left(\frac{r^k}{\varrho_i^k} - \frac{r^k}{r_{i+1}^k}\right) + s_{k+2}(r_i)\left(1 - \frac{r^k}{\varrho_i^k}\right)\right],$$

$$b_{k,i}(r) = 2\frac{k+1}{k}\left[s_{k+2}(r_{i-1})\left(\frac{r^k}{r_{i-1}^k} - \frac{r^k}{\varrho_{i-1}^k}\right) + s_{k+2}(r_i)\left(\frac{r^k}{\varrho_{i-1}^k} - 1\right)\right].$$

• $r \in (\varrho_{i-1}, \varrho_i]$, $L < i \le N$:

$$c_k(r) = \begin{cases} 0, & k \neq 0, \\ \frac{r^k}{r_{i-1}^k} c_k(r_{i-1}), & k < 0. \end{cases} \tag{6.3}$$

Coefficients of the Cauchy transform for piecewise-constant data.

• $r = r_1$:

$$c_k(r_1) = \begin{cases} c_0(r_2) - 2s_1(r_2)\varrho_0 - 2s_1(r_2)(r_2 - \varrho_1), & k = 0, \\ 0, & k \neq 0. \end{cases}$$

• $r \in (\varrho_{i-1}, \varrho_i]$, $1 \le i \le L$:

$$c_k(r) = \begin{cases} \left(\frac{r}{r_{i+1}}\right)^k c_k(r_{i+1}) - d_{k,i}(r), & k > 1, \\ \frac{r}{r_{i+1}} c_1(r_{i+1}) - 2rs_2(r_i)\ln\frac{\varrho_i}{r} - 2rs_2(r_{i+1})\ln\frac{r_{i+1}}{\varrho_i}, & k = 1, \\ \left(\frac{r}{r_{i-1}}\right)^k c_k(r_{i-1}) + b_{k,i}(r), & k <= 0, \end{cases}$$

where

$$d_{k,i}(r) = \frac{2r}{1-k}\left[s_{k+1}(r_i)\left(\frac{r^{k-1}}{\varrho_i^{k-1}} - 1\right) + s_{k+1}(r_{i+1})\left(\frac{r^{k-1}}{r_{i+1}^{k-1}} - \frac{r^{k-1}}{\varrho_i^{k-1}}\right)\right],$$

$$b_{k,i}(r) = \frac{2r}{1-k}\left[s_{k+1}(r_{i-1})\left(\frac{\varrho_{i-1}^{1-k}}{r^{1-k}} - \frac{r_{i-1}^{1-k}}{r^{1-k}}\right) + s_{k+1}(r_i)\left(1 - \frac{\varrho_{i-1}^{1-k}}{r^{1-k}}\right)\right].$$

• $r \in (\varrho_{i-1}, \varrho_i]$, $L < i \le N$:

$$c_k(r) = \begin{cases} 0, & k \neq 0, \\ \frac{r^k}{r_{i-1}^k} c_k(r_{i-1}), & k < 0. \end{cases}$$

Acknowledgment. The author would like to thank Michael Yampolsky for his numerous helpful comments.

REFERENCES

[1] Ada Core Technologies, New York, NY 10003, USA; see also ftp://cs.nyu.edu/pub/gnat.

[2] L. Ahlfors, *Lectures on Quasiconformal Mappings*, Princeton University Press, Princeton, NJ, 1966.

[3] L. Ahlfors and L. Bers, *Riemann's mapping theorem for variable metrics*, Ann. of Math., 72 (1960), pp. 385–404.

[4] P. P. Belinskii, S. K. Godunov, Yu. B. Ivanov, and I. K. Yanenko, *An application of a class of quasiconformal mappings to the construction of difference grids in domains with curvilinear boundaries*, in Numerical Methods in Plasma Physics, Nauka, Moscow, 1977, pp. 31–52 (in Russian).

[5] L. Bers, *Quasiconformal mappings, with applications to differential equations, function theory and topology*, Bull. Amer. Math. Soc., 83 (1977), pp. 1083–1100.

[6] B. V. Boyarskii, *Generalized solutions of systems of differential equations of first order and elliptic type with discontinuous coefficients*, Mat. Sb. N.S., 43 (1957), pp. 451–503.

[7] B. V. Boyarskii and T. Iwanier, *Quasiconformal mappings and nonlinear elliptic equations in two variables*, I, Bull. Acad. Polon. Sci. Sér. Sci. Math. Astronom. Phys., 12 (1974), pp. 473–478.

[8] B. V. Boyarskii and T. Iwanier, *Quasiconformal mappings and nonlinear elliptic equations in two variables*, II, Bull. Acad. Polon. Sci. Sér. Sci. Math. Astronom. Phys., 12 (1974), pp. 479–484.

[9] A. P. Calderon and A. Zygmund, *On singular integrals*, Amer. J. Math., 78 (1956), pp. 289–309.

[10] L. Carleson and Th. W. Gamelin, *Complex Dynamics*, Springer, New York, 1991.

[11] P. Daripa, *A fast algorithm to solve nonhomogeneous Cauchy–Riemann equations in the complex plane*, SIAM J. Sci. Statist. Comput., 13 (1992), pp. 1418–1432.

[12] P. Daripa, *A fast algorithm to solve the Beltrami equation with applications to quasiconformal mappings*, J. Comput. Phys., 106 (1993), pp. 355–365.

[13] P. Daripa and D. Mashat, *Singular integral transforms and fast numerical algorithms*, Numer. Algorithms, 18 (1998), pp. 133–157.

[14] D. Gaidashev, *Cylinder renormalization for Siegel discs and a constructive measurable Riemann mapping theorem*, Nonlinearity, 20 (1998), pp. 713–741.

[15] D. Gaidashev and M. Yampolsky, *Cylinder renormalization of Siegel disks*, Experiment. Math., 16 (2007), pp. 215–226.

[16] Z.-X. He, *Solving Beltrami equations by circle packing*, Trans. Amer. Math. Soc., 322 (1990), pp. 657–670.

[17] http://www.math.kth.se/~gaidash/Programs/beltrami_solver.tar.bz2.

[18] S.-Y. Lan and D.-Q. Dai, *Riemann mapping theorems for Beltrami equations by circle packings*, Pacific J. Math., 217 (2004), pp. 139–147.

[19] S.-Y. Lan and D.-Q. Dai, *Numerical analysis of quasiconformal mappings by circle packings*, Acta Math. Sci., 26 (2006), pp. 94–98.

[20] V. Markovic, *Quasiconformal Maps*, Lecture Notes, 2003.

[21] C. W. Mastin and J. F. Thompson, *Discrete quasiconformal mappings*, Z. Angew. Math. Phys., 29 (1978), pp. 1–11.

[22] C. W. Mastin and J. F. Thompson, *Quasiconformal mappings and grid generation*, SIAM J. Sci. Statist. Comput., 5 (1984), pp. 305–310.

[23] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes in Fortran. The Art of Scientific Computing*, Cambridge University Press, Cambridge, UK, 1992.

[24] B. Rodin and D. Sullivan, *The convergence of circle packings to the Riemann mapping*, J. Differential Geom., 26 (1987), pp. 349–360.

[25] S. T. Taft and R. A. Duff, eds., *Ada* 95 *Reference Manual: Language and Standard Libraries, International Standard ISO/IEC* 8652 : 1995*(E)*, Lecture Notes in Comput. Sci. 1246, Springer, Berlin, 1995.

[26] A. P. Vlasyuk and V. G. Mikhalchuk, *Automatic Construction of Conformal and Quasiconformal Mappings of Quadrangle Domains by Means of Difference Grids with "Floating" Nodes*, Akad. Nauk Ukrain. SSR Inst. Mat. Preprint 79, 1989 (in Russian).

[27] A. P. Vlasyuk and V. G. Mikhalchuk, *Automatic Construction of Conformal and Quasiconformal Mappings of Doubly Connected and Triply Connected Domains*, Akad. Nauk Ukrainy Inst. Mat. Preprint 57, 1991 (in Russian).

SIAM J. NUMER. ANAL.
Vol. 46, No. 5, pp. 2254–2275
© 2008 Society for Industrial and Applied Mathematics

A LEGENDRE–GALERKIN SPECTRAL METHOD FOR OPTIMAL CONTROL PROBLEMS GOVERNED BY ELLIPTIC EQUATIONS*

YANPING CHEN†, NIANYU YI‡, AND WENBIN LIU§

Abstract. In this paper, we study the Legendre–Galerkin spectral approximation for a constrained optimal control problem. We first derive a priori error estimates for the spectral approximation of optimal control problems. Then a posteriori error estimates are obtained for both the state and the control approximation. A preconditioning projection algorithm is proposed with some numerical tests.

Key words. Legendre–Galerkin, spectral method, convex optimal control, a priori error, a posteriori error, preconditioning projection algorithms

AMS subject classifications. 49J20, 65L10, 65L12

DOI. 10.1137/070679703

1. Introduction.

The spectral method employs global polynomials as the trial functions for the discretization of PDEs. It provides very accurate approximations with a relatively small number of unknowns when the solutions are smooth. Recently the spectral method has been extended to approximate unconstrained optimal control problems; see, for example, [10].

However spectral accuracy generally cannot be achieved when the approximated solutions have lower regularities, and this is typically the case when, for example, there exist control constraints in an optimal control problem (the so-called constrained optimal control problem). Thus the spectral method is not widely used in solving constrained distributed optimal control problems where the solutions often have singularities at the boundary of constraints even though all the initial data are smooth. Thus although there has been much work on the finite element method for numerically solving constrained optimal control problems, it seems that there has not been much work on the spectral method. Furthermore, the optimality conditions, which are normally the starting point of spectral approximation, are just PDE systems for unconstrained control problems, while those for constrained control problems contain variational inequalities, as shown later on. This also raises new issues in analyzing and solving the systems discretized using the spectral method.

In this work we study a constrained optimal control problem. It in fact represents a class of useful optimal control problems. We show that the solution of this particular control problem can be infinitely smooth if the initial data are so. Then we propose to use the Galerkin spectral method to approximate the solution. As there

*Received by the editors January 10, 2007; accepted for publication (in revised form) November 12, 2007; published electronically May 9, 2008. This work was supported by Program for New Century Excellent Talents in University of China State Education Ministry NCET-04-0776, National Science Foundation of China, the National Basic Research Program under grant 2005CB321703, and Scientific Research Fund of Hunan Provincial Education Department.

http://www.siam.org/journals/sinum/46-5/67970.html

†School of Mathematical Sciences, South China Normal University, Guangzhou 510631, People's Republic of China (yanpingchen@scnu.edu.cn).

‡Hunan Key Laboratory for Computation and Simulation in Science and Engineering, Department of Mathematics, Xiangtan University, Xiangtan 411105, Hunan, People's Republic of China (yinianyu365109@126.com).

§CBS and IMS, University of Kent, Canterbury CT2 7NF, UK (W.B.Liu@ukc.ac.uk).

are variational inequalities in optimal conditions of constrained optimal control problems, approaches have to be used that are different from those used in dealing with spectral approximation of unconstrained optimal control, such as the methods used in [10]. The purpose of this work is to derive a priori and a posteriori error estimates for the spectral approximation of the optimal control problem. Also we propose a preconditioning projection algorithm to solve the control problem with some numerical tests.

We are interested in the following distributed convex optimal control problem:

$$\min_{u\in K}\left\{\frac{1}{2}\int_\Omega (y-y_0)^2+\frac{1}{2}\int_\Omega u^2\right\}, \tag{1.1}$$

$$-\Delta y = f+u \ \text{ in } \Omega, \quad y|_{\partial\Omega}=0, \tag{1.2}$$

where K is a set defined by

$$K=\left\{v\in L^2(\Omega): \int_\Omega v\geq 0\right\}.$$

The details will be specified later.

The outline of this paper is as follows: In section 2, we discuss the spectral approximation of the optimal control problem. In section 3, we establish some important properties of this control problem and study the regularity of its solution. In section 4, a priori error estimates are derived for the optimal control problem. In section 5, a posteriori error estimates in the H^1 and L^2 norms are derived for the optimal control problem. And in section 6 we discuss the *hp* spectral element method for the optimal control problem. In section 7, we propose a preconditioning projection algorithm and present some numerical results.

Let $\Omega\subset\mathbb{R}^d$ ($d=1,2,$ or 3) be a convex open domain with a Lipschitz boundary $\partial\Omega$. In this paper we adopt the standard notation $W^{m,q}(\Omega)$ for Sobolev spaces on Ω with the norm $||\cdot||_{W^{m,q}(\Omega)}$ and the seminorm $|\cdot|_{W^{m,q}(\Omega)}$. We set $W_0^{m,q}(\Omega)\equiv\{w\in W^{m,q}(\Omega):\ w|_{\partial\Omega}=0\}$. We denote $W^{m,2}(\Omega)$ ($W_0^{m,2}(\Omega)$) by $H^m(\Omega)$ ($H_0^m(\Omega)$). In addition, C denotes a general positive constant independent of N, the order of the spectral approximation.

2. Optimal control problem and its spectral approximation. In this section we study Galerkin spectral approximation of the optimal control problem (1.1)–(1.2). We first introduce a weak formulation of the problem. Let $V=H_0^1(\Omega)$ and $U=L^2(\Omega)$. Let

$$a(y,w)=\int_\Omega(\nabla y)\cdot(\nabla w)\qquad \forall y,w\in V,$$

$$(u,v)_U=\int_\Omega uv\qquad \forall (u,v)\in U\times U.$$

Then the standard weak formulation for the state equation reads as follows: find $y(u)\in V$ such that

$$a(y(u),w)=(f+u,w)\qquad \forall w\in V,$$

where $(\cdot,\cdot)$ is the inner product in $L^2(\Omega)$. Therefore the optimal control problem (1.1)–(1.2) can be restated as follows: (OCP) find (y,u) such that

$$\min_{u\in K}\left\{\frac{1}{2}\int_\Omega (y-y_0)^2+\frac{1}{2}\int_\Omega u^2\right\}, \tag{2.1}$$

$$a(y(u),w)=(f+u,w)\quad \forall w\in V. \tag{2.2}$$

It is well known (see, e.g., [20]) that the optimal control problem (OCP) (2.1)–(2.2) has a unique solution (y^*, u^*), and (y^*, u^*) is the solution if and only if there is a costate $p^* \in V$ such that the triplet (y^*, p^*, u^*) satisfies the following optimality conditions: (OCP-OPT)

$$a(y^*, w) = (f + u^*, w) \qquad \forall w \in V, \tag{2.3}$$

$$a(q, p^*) = (y^* - y_0, q) \qquad \forall q \in V, \tag{2.4}$$

$$(u^* + p^*, v - u^*)_U \geq 0 \qquad \forall v \in K \subset U. \tag{2.5}$$

Let us investigate the spectral approximation of the optimal control problem (OCP) (2.1)–(2.2). Here we consider only the Legendre–Galerkin spectral method. From now on, we assume that $\Omega = (-1, 1)^d$.

Let us first introduce some basic notation which will be used in what follows. For $x_i,\ i = 1, \ldots, d$, we denote by $L_r(x_i)$ the rth degree Legendre polynomial in the variable x_i, and we set

$$X_N^i = \text{span}\{L_0(x_i),\ L_1(x_i), \ldots, L_N(x_i)\},$$

where $N \geq 0$ is an integer. We define a product space such as

$$X_N = \prod_{i=1}^{d} X_N^i.$$

We introduce the finite dimensional spaces $V_N = X_N \bigcap V$, $U_N = X_N \bigcap U$, and $K_N = X_N \bigcap K$.

Then the Legendre–Galerkin approximation for (OCP) is $(\text{OCP})^N$

$$\min_{u_N \in K_N \subset U_N} \left\{ \frac{1}{2} \int_\Omega (y_N - y_0)^2 + \frac{1}{2} \int_\Omega u_N^2 \right\}, \tag{2.6}$$

$$a(y_N, w_N) = (f + u_N, w_N) \quad \forall w_N \in V_N \subset H_0^1(\Omega). \tag{2.7}$$

This is a finite dimensional optimization problem and may be solved by existing mathematical programming methods such as the steepest decent method, the conjugate gradient method, the trust domain method, and the sequential quadratic programming method.

It follows that the control problem $(\text{OCP})^N$ has a solution (y_N^*, u_N^*), and (y_N^*, u_N^*) is the solution if and only if there is a costate $p_N^* \in V_N$ such that the triplet (y_N^*, p_N^*, u_N^*) satisfies the following optimality conditions: $(\text{OCP-OPT})^N$

$$a(y_N^*, w_N) = (f + u_N^*, w_N) \qquad \forall w_N \in V_N, \tag{2.8}$$

$$a(q_N, p_N^*) = (y_N^* - y_0, q_N) \qquad \forall q_N \in V_N, \tag{2.9}$$

$$(u_N^* + p_N^*, v_N - u_N^*)_U \geq 0 \qquad \forall v_N \in K_N. \tag{2.10}$$

Let $y(u)$ and $y_N(u_N)$ be the solutions of (2.2) and (2.7), respectively. Set

$$J(u) = \left\{ \frac{1}{2} \int_\Omega (y - y_0)^2 + \frac{1}{2} \int_\Omega u^2 \right\},$$

$$J_N(u_N) = \left\{ \frac{1}{2} \int_\Omega (y_N - y_0)^2 + \frac{1}{2} \int_\Omega u_N^2 \right\}.$$

It is clear that $J(\cdot)$ is uniformly convex. Then there is a $c > 0$ independent of N such that

$$(J^{'}(u^*) - J^{'}(u_N^*), u^* - u_N^*)_U \geq c||u^* - u_N^*||^2_{L^2(\Omega)}, \tag{2.11}$$

where u^* and u_N^* are the solutions of (2.1)–(2.2) and (2.6)–(2.7), respectively.

For any $u \in L^2(\Omega)$, we define the orthogonal projection operator $P_N : \ L^2(\Omega) \to U_N$ which satisfies

$$(P_N u - u, w_N) = 0 \qquad \forall w_N \in U_N.$$

It can be shown (see [6]) that

$$P_N u = \sum_{\max\limits_{1\leq i\leq d}\{k_i\}\leq N} \hat{u}_k \phi_k, \qquad \hat{u}_k = \prod_{i=1}^{d} \left(k_i + \frac{1}{2}\right) \int_\Omega u(x)\phi_k(x),$$

where $k = (k_1, \ldots, k_d)$ and $\phi_k(x) = L_{k_1}(x_1) \cdots L_{k_d}(x_d)$.

For any $u \in H_0^1(\Omega), P_{1,N}^0 : \ H_0^1(\Omega) \to V_N$ is defined by

$$\int_\Omega \nabla(u - P_{1,N}^0 u) \cdot \nabla w = 0 \qquad \forall w \in V_N.$$

The following lemma is important for deriving a posteriori error estimates of residual type. It can be found in the reference book [6].

LEMMA 2.1. *For all $u \in H^m(\Omega)$ $(m \geq 0)$, we have that*

$$||u - P_N u||_{H^l(\Omega)} \leq C N^{\sigma(l)-m} ||u||_{H^m(\Omega)}, \qquad 0 \leq l \leq m,$$

where $\sigma(l) = 0$ if $l = 0$, and $\sigma(l) = 2l - \frac{1}{2}$ for $l > 0$. If $u \in H_0^1(\Omega) \cap H^m(\Omega), m \geq 1$, then we have that

$$||u - P_{1,N}^0 u||_{H^\mu(\Omega)} \leq C N^{\mu-m} ||u||_{H^m(\Omega)}, \qquad 0 \leq \mu \leq 1.$$

3. Preliminary properties of the optimal control problem. In general, regularity of the optimal control for a constrained optimal control is quite low, say only in $H^1(\Omega)$; see [19]. When the constraint set is of obstacle type $\{u : u \geq 0\}$, for example, we have the following relationship between the optimal control u and the costate p: $u = \max(0, -p)$; see [19]. Thus the gradient of u has jumps along the boundary of the zero-set of p, whose location is generally unknown. Due to the special structure of our control constraint set K, we are able to show that the optimal control of our optimal control problem can be infinitely smooth if the initial data are so. To this end we first derive an important relationship between the optimal control and the optimal costate of (OCP-OPT). This relationship is key to our analysis and numerical algorithms.

THEOREM 3.1. *Let (y^*, p^*, u^*) be the solution of* (OCP-OPT). *Then we have*

$$u^* = \max\{0, \overline{p^*}\} - p^*,$$

where $\overline{p} = \int_\Omega p \Big/ \int_\Omega 1$ denotes the integral average on Ω of the function p.

Proof. For any function $p \in V$ we show that

$$u = \max\{0, \overline{p}\} - p$$

satisfies the variational inequality

$$(u+p, v-u)_U \geq 0 \qquad \forall v \in K.$$

If $\overline{p} > 0$, then $u = \overline{p} - p$ and

$$\begin{aligned}(u+p, v-u)_U &= \int_\Omega (u+p)(v-u) \\ &= \int_\Omega \overline{p}(v-\overline{p}+p) \\ &= \overline{p}\int_\Omega v \geq 0 \qquad \text{for } \ v \in K.\end{aligned}$$

If $\overline{p} \leq 0$, then $u = -p$ and

$$(u+p, v-u)_U = 0.$$

Now, note that for the costate solution p^* the solution of (2.5) is unique. Thus we have proved the theorem. ∎

THEOREM 3.2. *Let (y_N^*, p_N^*, u_N^*) be the solution of* (OCP-OPT)^N. *Then we have*

$$u_N^* = \max\{0, \overline{p_N^*}\} - p_N^*.$$

Proof. The proof is similar to the proof of Theorem 3.1. ∎

It follows from Theorem 3.1 that we can derive regularity of the optimal control.

THEOREM 3.3. *Let (y^*, p^*, u^*) satisfy the optimality conditions* (OCP-OPT). *Assume that the data functions f, y_0, and the domain Ω are infinitely smooth. Then the optimal control $u^* \in C^\infty(\bar{\Omega})$.*

Proof. Applying the standard regularity argument of elliptic problems, it is clear that $y \in H^2(\Omega)$ so that $p^* \in H^4(\Omega)$. It follows from the relationship between the control and the costate

$$u^* = \max\{0, \overline{p^*}\} - p^*$$

that $u^* \in H^4(\Omega)$. Thus $y^* \in H^6(\Omega)$. By repeating the above process, we can conclude that $u^* \in C^\infty(\bar{\Omega})$. ∎

Remark 3.1. If Ω is a convex open domain with a Lipschitz boundary $\partial\Omega$ and $f, y_0 \in L^2(\Omega)$, then we have $u^* \in H^2(\Omega)$.

4. A priori error estimates. In this section, we show that u_N^* converges to u^* in spectral accuracy provided that the data are sufficiently smooth. Defining $J(\cdot)$ and $J_N(\cdot)$ as before, it can be shown that (see [20])

$$(J'(u^*), v)_U = (u^* + p^*, v)_U, \tag{4.1}$$

$$(J_N'(u_N^*), v)_U = (u_N^* + p_N^*, v)_U, \tag{4.2}$$

$$(J'(u_N^*), v)_U = (u_N^* + p(u_N^*), v)_U, \tag{4.3}$$

where $p(u_N^*)$ is the solution of the auxiliary equation:

$$a(y(u_N^*), w) = (f + u_N^*, w) \qquad \forall w \in V = H_0^1(\Omega), \tag{4.4}$$

$$a(q, p(u_N^*)) = (y(u_N^*) - y_0, q) \quad \forall q \in V = H_0^1(\Omega). \tag{4.5}$$

THEOREM 4.1. *Let (y^*, p^*, u^*) and (y_N^*, p_N^*, u_N^*) be the solutions of (2.3)–(2.5) and (2.8)–(2.10), respectively. Assume that the solution (y^*, p^*, u^*) is sufficiently regular. Then for any integer $k > 0$ there exists $C > 0$, independent of N, such that*

$$||u^* - u_N^*||_{L^2(\Omega)} + ||y^* - y_N^*||_{H^1(\Omega)} + ||p^* - p_N^*||_{H^1(\Omega)} \le CN^{-k} \qquad \forall k \in \mathbb{Z}^+. \tag{4.6}$$

Proof. First, it follows from the convexity of $J(\cdot)$, (4.1), (4.3), the optimal conditions (2.5) and (2.10), and the Schwarz inequality that for any $v_N \in K_N$,

$$\begin{aligned}
c||u^* - u_N^*||_{L^2(\Omega)}^2 &\le (J'(u^*), u^* - u_N^*)_U - (J'(u_N^*), u^* - u_N^*)_U \\
&= (u^* + p^*, u^* - u_N^*)_U - (u_N^* + p(u_N^*), u^* - u_N^*)_U \\
&= (u^* + p^*, u^* - v_N)_U + (u^* + p^*, v_N - u_N^*)_U \\
&\quad + (u_N^* + p_N^*, u_N^* - v_N)_U + (u_N^* + p_N^*, v_N - u^*)_U \\
&\quad + (p_N^* - p(u_N^*), u^* - u_N^*)_U \\
&\le (u^* + p^*, v_N - u^*)_U + (u_N^* - u^*, v_N - u^*)_U \\
&\quad + (u^* + p^*, v_N - u^*)_U + (p_N^* - p(u_N^*), v_N - u^*)_U \\
&\quad + (p(u_N^*) - p^*, v_N - u^*)_U + (p_N^* - p(u_N^*), u^* - u_N^*)_U \\
&\le 2(u^* + p^*, v_N - u^*)_U \\
&\quad + ||u_N^* - u^*||_{L^2(\Omega)}||v_N - u^*||_{L^2(\Omega)} \\
&\quad + ||p_N^* - p(u_N^*)||_{L^2(\Omega)}||v_N - u^*||_{L^2(\Omega)} \\
&\quad + ||p(u_N^*) - p^*||_{L^2(\Omega)}||v_N - u^*||_{L^2(\Omega)} \\
&\quad + ||p_N^* - p(u_N^*)||_{L^2(\Omega)}||u^* - u_N^*||_{L^2(\Omega)} \\
&\le 2(u^* + p^*, v_N - u^*)_U \\
&\quad + C(\delta)||v_N - u^*||_{L^2(\Omega)}^2 + C(\delta)||p_N^* - p(u_N^*)||_{L^2(\Omega)}^2 \\
&\quad + \delta||p(u_N^*) - p^*||_{L^2(\Omega)}^2 + 2\delta||u^* - u_N^*||_{L^2(\Omega)}^2,
\end{aligned} \tag{4.7}$$

where $\delta > 0$ is a sufficiently small constant.

Next, we note that

$$||y_N^* - y^*||_{H^1(\Omega)} \le ||y_N^* - y(u_N^*)||_{H^1(\Omega)} + ||y(u_N^*) - y^*||_{H^1(\Omega)},$$

$$||p_N^* - p^*||_{H^1(\Omega)} \le ||p_N^* - p(u_N^*)||_{H^1(\Omega)} + ||p(u_N^*) - p^*||_{H^1(\Omega)}.$$

From (2.3) and (4.4), we can obtain the following error equation of the intermediate variable:

$$a(y(u_N^*) - y^*, w) = (u_N^* - u^*, w) \qquad \forall w \in V.$$

Setting $w = y(u_N^*) - y^*$ in the above equation, we have that

$$||y(u_N^*) - y^*||_{H^1(\Omega)} \le c_1||u_N^* - u^*||_{L^2(\Omega)}, \tag{4.8}$$

where $c_1 = C(\Omega) > 0$ is the coefficient of the Poincaré inequality

$$||w||_{H_0^1(\Omega)} \le C(\Omega)||\nabla w||_{L^2(\Omega)}.$$

Using (2.8) and (4.4), we have

$$a(y_N^* - y(u_N^*), w_N) = 0 \qquad \forall w_N \in V_N. \tag{4.9}$$

Hence,

$$\begin{aligned}||y_N^*-y(u_N^*)||^2_{H^1(\Omega)} &\le Ca(y_N^*-y(u_N^*),y_N^*-y(u_N^*))\\ &= Ca(y_N^*-y(u_N^*),y_N^*-y(u_N^*)+w_N)\\ &\le C||y_N^*-y(u_N^*)||_{H^1(\Omega)}\cdot\inf_{w_N\in V_N}||y(u_N^*)-w_N||_{H^1(\Omega)}.\end{aligned}$$

Thus,

$$\begin{aligned}||y_N^*-y(u_N^*)||_{H^1(\Omega)} &\le C\inf_{w_N\in V_N}||y(u_N^*)-w_N||_{H^1(\Omega)}\\ &\le C||y(u_N^*)-P^0_{1,N}y(u_N^*)||_{H^1(\Omega)}\le CN^{-k}\qquad\forall k\in\mathbb{Z}^+,\end{aligned}\tag{4.10}$$

where $P^0_{1,N}$ is the orthogonal projection operator defined in section 2.

Similarly, for the costate we also have that

$$||p(u_N^*)-p^*||_{H^1(\Omega)}\le c_1||y(u_N^*)-y^*||_{L^2(\Omega)}\le c_1||u_N^*-u^*||_{L^2(\Omega)}.\tag{4.11}$$

Using (2.9) and (4.5), we have

$$\begin{aligned}||p(u_N^*)-p_N^*||^2_{H^1(\Omega)} &\le C(||p(u_N^*)-p_N^*-P_N(p(u_N^*)-p_N^*)||_{H^1(\Omega)}||p(u_N^*)-p_N^*||_{H^1(\Omega)}\\ &\quad+||y(u_N^*)-y_N^*||_{L^2(\Omega)}||p(u_N^*)-p_N^*-P_N(p(u_N^*)-p_N^*)||_{L^2(\Omega)}\\ &\quad+||y(u_N^*)-y_N^*||_{L^2(\Omega)}||p(u_N^*)-p_N^*||_{L^2(\Omega)})\\ &\le CN^{-2k}\qquad\forall k\in\mathbb{Z}^+.\end{aligned}\tag{4.12}$$

Then, apply (4.11)–(4.12) in (4.7) to obtain that

$$\begin{aligned}c||u^*-u_N^*||^2_{L^2(\Omega)} &\le (J^{'}(u^*),u^*-u_N^*)_U-(J^{'}(u_N^*),u^*-u_N^*)_U\\ &\le 2(u^*+p^*,v_N-u^*)_U+C(\delta)||v_N-u^*||^2_{L^2(\Omega)}\\ &\quad+C(\delta)N^{-2k}+(2+c_1)\delta||u^*-u_N^*||^2_{L^2(\Omega)},\end{aligned}\tag{4.13}$$

where $\delta>0$ is a sufficiently small constant. Thus, we choose $\delta=c/(4+2c_1)$ to derive that

$$||u^*-u_N^*||^2_{L^2(\Omega)}\le 2(u^*+p^*,v_N-u^*)_U+C(\delta)||v_N-u^*||^2_{L^2(\Omega)}+C(\delta)N^{-2k}.\tag{4.14}$$

Now, set $v_N=P_Nu^*\in U_N$ in (4.14), where P_N is the L^2 orthogonal projection operator defined in section 2. Note that

$$(u^*-P_Nu^*,v)=0\qquad\forall v\in P_N,$$

and especially letting $v=1\in P_N$ we have

$$(u^*-P_Nu^*,v)=\int_\Omega(u^*-P_Nu^*)=0;$$

thus

$$\int_\Omega P_Nu^*=\int_\Omega u^*\ge 0.$$

Then $v_N\in K_N\subset U_N$. It follows from Lemma 2.1 that

$$||u^*-v_N||_{L^2(\Omega)}\le CN^{-k}.\tag{4.15}$$

Applying (4.15) in (4.14), we can obtain that

$$||u^*-u_N^*||_{L^2(\Omega)}\le CN^{-k}.\tag{4.16}$$

Finally, combining (4.8)–(4.12) with (4.16) implies the error estimates (4.6). ∎

5. A posteriori error estimates. In this section, we derive a posteriori error estimates for the spectral approximation of (OCP). First, we use the L^2 norm for estimating the control approximation error and the H^1 norm for the state and costate approximation error, namely, the L^2-H^1 posteriori error estimates. Now, we assume that Ω is a convex open domain with a Lipschitz boundary $\partial\Omega$ and $f, y_0 \in L^2(\Omega)$.

In the following lemma, we can derive a posteriori error estimates for the intermediate variable.

LEMMA 5.1. *Let $y(u_N^*)$, $p(u_N^*)$, y_N^*, and p_N^* be the solutions of* (4.4), (4.5), (2.8), *and* (2.9), *respectively. Then*

$$||y(u_N^*) - y_N^*||_{H^1(\Omega)} + ||p(u_N^*) - p_N^*||_{H^1(\Omega)} \le C\eta, \tag{5.1}$$

where the estimator η is defined as

$$\eta = N^{-1}||y_N^* - y_0 + \Delta p_N^*||_{L^2(\Omega)} + N^{-1}||f + u_N^* + \Delta y_N^*||_{L^2(\Omega)}.$$

Proof. Set $e_p = p(u_N^*) - p_N^*$, and let $e_p^N = P_{1,N}^0 e_p \in V_N$ be the orthogonal projection of e_p. Then it follows from (2.9), (4.5), and Lemma 2.1 that

$$\begin{aligned} c_1^{-1}||p(u_N^*) - p_N^*||_{H^1(\Omega)}^2 &\le (\nabla e_p, \nabla e_p) = (\nabla(e_p - e_p^N), \nabla e_p) + (\nabla e_p^N, \nabla e_p) \\ &= (y(u_N^*) - y_0 + \Delta p_N^*, e_p - e_p^N) + (y(u_N^*) - y_N^*, e_p^N) \\ &= (y_N^* - y_0 + \Delta p_N^*, e_p - e_p^N) + (y(u_N^*) - y_N^*, e_p) \\ &\le ||y_N^* - y_0 + \Delta p_N^*||_{L^2(\Omega)} \cdot ||e_p - e_p^N||_{L^2(\Omega)} + (y(u_N^*) - y_N^*, e_p) \\ &\le CN^{-1}||y_N^* - y_0 + \Delta p_N^*||_{L^2(\Omega)}||e_p||_{H^1(\Omega)} \\ &\quad + ||y(u_N^*) - y_N^*||_{L^2(\Omega)}||e_p||_{L^2(\Omega)} \\ &\le CN^{-2}||y_N^* - y_0 + \Delta p_N^*||_{L^2(\Omega)}^2 + C||y(u_N^*) - y_N^*||_{L^2(\Omega)}^2 \\ &\quad + 2\delta||e_p||_{H^1(\Omega)}^2. \end{aligned}$$

Choosing that $\delta = \frac{1}{4c_1}$, we have

$$||p(u_N^*) - p_N^*||_{H^1(\Omega)}^2 \le CN^{-2}||y_N^* - y_0 + \Delta p_N^*||_{L^2(\Omega)}^2 + C||y(u_N^*) - y_N^*||_{L^2(\Omega)}^2. \tag{5.2}$$

Similarly, set $e_y = y(u_N^*) - y_N^*$, and let $e_y^N = P_{1,N}^0 e_y \in V_N$ be the orthogonal projection of e_y. Then, using (4.9), (2.8), (4.4), and Lemma 2.1, we have that

$$\begin{aligned} c_1^{-1}||y(u_N^*) - y_N^*||_{H^1(\Omega)}^2 &\le (\nabla e_y, \nabla e_y) = (\nabla(e_y - e_y^N), \nabla e_y) \\ &= (f + u_N^* + \Delta y_N^*, e_y - e_y^N) \\ &\le CN^{-2}||f + u_N^* + \Delta y_N^*||_{L^2(\Omega)}^2 + \delta||y(u_N^*) - y_N^*||_{H^1(\Omega)}^2. \end{aligned}$$

Choosing that $\delta = \frac{1}{2c_1}$, we have

$$||y(u_N^*) - y_N^*||_{H^1(\Omega)}^2 \le CN^{-2}||f + u_N^* + \Delta y_N^*||_{L^2(\Omega)}^2. \tag{5.3}$$

Then we combine (5.2) and (5.3) to derive our estimates (5.1). ∎

THEOREM 5.1. *Let (y^*, p^*, u^*) and (y_N^*, p_N^*, u_N^*) be the solutions of* (2.3)–(2.5) *and* (2.8)–(2.10), *respectively. Then we have that*

$$||u^* - u_N^*||_{L^2(\Omega)} + ||y^* - y_N^*||_{H^1(\Omega)} + ||p^* - p_N^*||_{H^1(\Omega)} \le C\eta, \tag{5.4}$$

where η is defined in Lemma 5.1.

Proof. Let $y(u_N^*)$ and $p(u_N^*)$ be the intermediate variables defined in (4.4) and (4.5). For $v_N = P_N u^* \in X_N$, it follows from (4.1)–(4.3), (OCP-OPT) and (OCP-OPT)^N, and Theorem 3.2 that

$$\begin{aligned} c||u^* - u_N^*||^2_{L^2(\Omega)} &\le (J^{'}(u^*), u^* - u_N^*)_U - (J^{'}(u_N^*), u^* - u_N^*)_U \\ &\le -(J^{'}(u_N^*), u^* - u_N^*)_U \\ &= (J_N^{'}(u_N^*), u_N^* - u^*)_U + (J_N^{'}(u_N^*) - J^{'}(u_N^*), u^* - u_N^*)_U \\ &\le (J_N^{'}(u_N^*), v_N - u^*)_U + (J_N^{'}(u_N^*) - J^{'}(u_N^*), u^* - u_N^*)_U \\ &= (J_N^{'}(u_N^*) - J^{'}(u_N^*), u^* - u_N^*)_U \\ &= (p_N^* - p(u_N^*), u^* - u_N^*) \\ &\le C||p(u_N^*) - p_N^*||_{L^2(\Omega)}||u^* - u_N^*||_{L^2(\Omega)}. \end{aligned}$$

Therefore, from Lemma 5.1, we have that

$$||u^* - u_N^*||_{L^2(\Omega)} \le C||p_N^* - p(u_N^*)||_{L^2(\Omega)}. \tag{5.5}$$

Note that

$$\begin{aligned} ||y^* - y_N^*||_{H^1(\Omega)} &\le ||y^* - y(u_N^*)||_{H^1(\Omega)} + ||y(u_N^*) - y_N^*||_{H^1(\Omega)}, \\ ||p^* - p_N^*||_{H^1(\Omega)} &\le ||p^* - p(u_N^*)||_{H^1(\Omega)} + ||p(u_N^*) - p_N^*||_{H^1(\Omega)}. \end{aligned} \tag{5.6}$$

By using (5.5)–(5.6), (4.8), (4.11), and Lemma 5.1, we can obtain the result (5.4). ∎

In many control applications, one is also interested in L^2 estimation of the state and costate approximation error.

In order to estimate $||p(u_N^*) - p_N^*||_{L^2(\Omega)}$, we assume that the auxiliary problem

$$-\Delta\xi = f, \qquad x \in \Omega, \qquad \xi|_{\partial\Omega} = 0, \tag{5.7}$$

possesses the following regularity estimates (see [18], for example):

$$||\xi||_{H^2(\Omega)} \le C||f||_{L^2(\Omega)}. \tag{5.8}$$

Then, we can show that the following lemma holds.

LEMMA 5.2. *Let $(y(u_N^*), p(u_N^*))$ and (y_N^*, p_N^*) be the solutions of* (4.4)–(4.5) *and* (2.8)–(2.9)*, respectively. Then*

$$||p(u_N^*) - p_N^*||_{L^2(\Omega)} + ||y(u_N^*) - y_N^*||_{L^2(\Omega)} \le C\eta_1, \tag{5.9}$$

where

$$\eta_1 = N^{-2}||y_N^* - y_0 + \Delta p_N^*||_{L^2(\Omega)} + N^{-2}||f + u_N^* + \Delta y_N^*||_{L^2(\Omega)}.$$

Proof. Let ξ be the solution of (5.7) with $f(x) = p(u_N^*) - p_N^*$, let $\xi^N = P^0_{1,N}\xi \in V_N$ be the orthogonal projection of ξ. Then it follows from (2.9), (4.5), and Lemma 2.1 that

$$\begin{aligned} ||p(u_N^*) - p_N^*||^2_{L^2(\Omega)} &= (\nabla\xi, \nabla(p(u_N^*) - p_N^*)) \\ &= (\nabla(\xi - \xi^N), \nabla(p(u_N^*) - p_N^*)) + (\nabla\xi^N, \nabla(p(u_N^*) - p_N^*)) \\ &= (y(u_N^*) - y_0 + \Delta p_N^*, \xi - \xi^N) + (y(u_N^*) - y_N^*, \xi^N) \\ &= (y_N^* - y_0 + \Delta p_N^*, \xi - \xi^N) + (y(u_N^*) - y_N^*, \xi) \\ &\le C\big(N^{-2}||y_N^* - y_0 + \Delta p_N^*||_{L^2(\Omega)} + ||y(u_N^*) - y_N^*||_{L^2(\Omega)}\big) \cdot ||\xi||_{H^2(\Omega)} \\ &\le C\big(N^{-2}||y_N^* - y_0 + \Delta p_N^*||_{L^2(\Omega)} + ||y(u_N^*) - y_N^*||_{L^2(\Omega)}\big) \cdot ||f||_{L^2(\Omega)}. \end{aligned} \tag{5.10}$$

Thus we have

$$||p(u_N^*)-p_N^*||_{L^2(\Omega)} \le CN^{-2}||y_N^*-y_0+\Delta p_N^*||_{L^2(\Omega)}+C||y(u_N^*)-y_N^*||_{L^2(\Omega)}. \tag{5.11}$$

Similarly, let ξ be the solution of (5.7) with $f(x)=y(u_N^*)-y_N^*$, and let $\xi^N = P_{1,N}^0\xi \in V_N$ be the orthogonal projection of ξ. Then it follows from (2.8), (4.4), and Lemma 2.1 that

$$\begin{aligned}||y(u_N^*)-y_N^*||^2_{L^2(\Omega)} &= (\nabla\xi, \nabla(y(u_N^*)-y_N^*))\\ &= (\nabla(\xi-\xi^N), \nabla(y(u_N^*)-y_N^*)) + (\nabla\xi^N, \nabla(y(u_N^*)-y_N^*))\\ &= (f+u_N^*+\Delta y_N^*, \xi-\xi^N)\\ &\le CN^{-2}||f+u_N^*+\Delta y_N^*||_{L^2(\Omega)}\cdot||\xi||_{H^2(\Omega)}\\ &\le CN^{-2}||f+u_N^*+\Delta y_N^*||_{L^2(\Omega)}\cdot||f||_{L^2(\Omega)}.\end{aligned}$$

Then

$$||y(u_N^*)-y_N^*||_{L^2(\Omega)} \le CN^{-2}||f+u_N^*+\Delta y_N^*||_{L^2(\Omega)}. \tag{5.12}$$

Thus (5.10)–(5.12) implies (5.9). ∎

THEOREM 5.2. *Let (y^*,p^*,u^*) and (y_N^*,p_N^*,u_N^*) be the solutions of* (2.3)–(2.5) *and* (2.8)–(2.10)*, respectively. Then, we have*

$$||u^*-u_N^*||_{L^2(\Omega)}+||y^*-y_N^*||_{L^2(\Omega)}+||p^*-p_N^*||_{L^2(\Omega)} \le C\eta_1, \tag{5.13}$$

where η_1 is defined in Lemma 5.2.

Proof. By using (5.5)–(5.6), (4.8), (4.11), and Lemma 5.2, we can obtain the result (5.13). ∎

6. *hp* spectral element method for optimal control problem. In this section, we extend the above estimates to the spectral element method. In fact we derive a posteriori error estimates for the *hp* spectral element method. Spectral element methods, proposed by Patera [27], combine the advantages of Galerkin spectral methods with those of finite element methods by a simple application of the spectral method per element. As in finite element methods, the domain is divided into N_τ nonoverlapping subdomains (elements) $\tau_i, 1 \le i \le N_\tau$:

$$\overline{\Omega} = \bigcup_{i=1}^{N_\tau} \overline{\tau_i}, \qquad \tau_i \bigcap \tau_j = \emptyset, \quad i \ne j, \quad 1 \le i,j \le N_\tau.$$

We first consider the *hp* spectral element approximation of (OCP). Let $\mathcal{T} = \{\tau\}$ be a local quasi-uniform partitioning of Ω into nonoverlapping regular elements τ. We denote by $\hat{\tau} = (-1,1)^2$ the reference element. We let $\mathcal{E}(\mathcal{T})$ denote all edges, and let $\mathcal{E}_0(\mathcal{T})$ denote all edges which do not lie on the boundary $\partial\Omega$. Each element τ can be the image of the reference element $\hat{\tau}$ under an affine map $F_\tau : \hat{\tau} \to \tau$. We write $h_\tau :=$ diam τ. We assume that the triangulation is γ-shape regular; i.e.,

$$h_\tau^{-1}\|F_\tau'\| + h_\tau\|(F_\tau')^{-1}\| \le \gamma. \tag{6.1}$$

For γ-shape regular meshes $\mathcal{T}$ on the domain Ω, we associate with each element $\tau \in \mathcal{T}$ a polynomial degree $p_\tau \in \mathbb{N}_0$; these polynomial degrees $\{p_\tau\}$ are collected into

the polynomial degree vector $\boldsymbol{p} = \{p_\tau\}$. Then we can define the spaces of spectral element approximation $U^{\boldsymbol{p}}(\mathcal{T},\Omega)$, $S^{\boldsymbol{p}}(\mathcal{T},\Omega)$, $S_0^{\boldsymbol{p}}(\mathcal{T},\Omega)$ as follows:

$$U^{\boldsymbol{p}}(\mathcal{T},\Omega) := \{u \in L^2(\Omega) : u|_e \circ F_\tau \in \mathcal{P}_{p_\tau}(\hat{\tau})\},$$

$$S^{\boldsymbol{p}}(\mathcal{T},\Omega) := \{v \in H^1(\Omega) : v|_e \circ F_\tau \in \mathcal{P}_{p_\tau}(\hat{\tau})\},$$

$$S_0^{\boldsymbol{p}}(\mathcal{T},\Omega) := S^{\boldsymbol{p}}(\mathcal{T},\Omega) \bigcap H_0^1(\Omega),$$

where $\mathcal{P}_{p_\tau}(\hat{\tau})$ denotes the spaces of polynomials in $\hat{\tau}$ of degree $\le p_\tau$ in each variable, respectively. As to polynomial degree distribution $\boldsymbol{p}$, similarly to (6.1), we assume that the polynomial degrees of neighboring elements are comparable; i.e., there is a constant $\gamma > 0$ such that

$$\gamma^{-1}(p_\tau + 1) \le p_{\tau'} + 1 \le \gamma(p_\tau + 1) \quad \forall \tau, \tau' \in \mathcal{T}, \quad \bar{\tau} \bigcap \bar{\tau}' \neq \emptyset. \tag{6.2}$$

Convergence is obtained either by increasing the degree of the polynomials or by increasing the number of elements N_τ.

Let $K^{h,\boldsymbol{p}} := K \bigcap U^{\boldsymbol{p}}(\mathcal{T},\Omega)$ be the space of approximation for the control, and let $S_0^{\boldsymbol{p}}(\mathcal{T},\Omega)$ be the space of approximation for the state and costate. Then the spectral element approximation to (OCP) reads as

$$\min_{u_{hp} \in K^{h,\boldsymbol{p}}} \left\{ \frac{1}{2} \int_\Omega (y_{hp} - y_0)^2 + \frac{1}{2} \int_\Omega u_{hp}^2 \right\}, \tag{6.3}$$

$$a(y_{hp}, w_{hp}) = (f + u_{hp}, w_{hp}) \quad \forall w_{hp} \in S_0^{\boldsymbol{p}}(\mathcal{T},\Omega). \tag{6.4}$$

It follows that (6.3)–(6.4) has a solution $(y_{hp}, u_{hp}) \in S_0^{\boldsymbol{p}}(\mathcal{T},\Omega) \times K^{h,\boldsymbol{p}}$, and (y_{hp}, u_{hp}) is the solution if and only if there is a costate $p_{hp} \in S_0^{\boldsymbol{p}}(\mathcal{T},\Omega)$ such that the triplet (y_{hp}, p_{hp}, u_{hp}) satisfies the following optimality conditions:

$$a(y_{hp}, w_{hp}) = (f + u_{hp}, w_{hp}) \qquad \forall w_{hp} \in S_0^{\boldsymbol{p}}(\mathcal{T},\Omega), \tag{6.5}$$

$$a(q_{hp}, p_{hp}) = (y_{hp} - y_0, q_{hp}) \qquad \forall q_{hp} \in S_0^{\boldsymbol{p}}(\mathcal{T},\Omega), \tag{6.6}$$

$$(u_{hp} + p_{hp}, v_{hp} - u_{hp})_U \ge 0 \qquad \forall v_{hp} \in K^{h,\boldsymbol{p}}. \tag{6.7}$$

Let $(y(u_{hp}), p(u_{hp}))$ be the solution of the auxiliary equations

$$a(y(u_{hp}), w) = (f + u_{hp}, w) \qquad \forall w \in H_0^1(\Omega), \tag{6.8}$$

$$a(q, p(u_{hp})) = (y(u_{hp}) - y_0, q) \qquad \forall q \in H_0^1(\Omega). \tag{6.9}$$

In what follows, we introduce two lemmas which generalize the well-known Clément-type interpolation operators of [9] and [28] to the *hp* context. The reader is referred to [25] for details.

LEMMA 6.1 (Clément-type quasi interpolation). *Let $\mathcal{T}$ be a γ-shape regular triangulation (see* (6.1)*) of a domain $\Omega \in \mathcal{R}^2$, and let $\boldsymbol{p}$ be a polynomial degree distribution which is comparable (see* (6.2)*). Then there exists a bounded linear operator $\Pi : L^1(\Omega) \to S^{\boldsymbol{p}}(\mathcal{T},\Omega)$, and there exists a constant $C > 0$, which depends only on γ, such that for every $u \in H^1(\Omega)$, all elements $\tau \in \mathcal{T}$, and all edges $e \in \mathcal{E}(\tau)$,*

$$\|u-\Pi u\|_{L^2(\tau)}+\frac{h_\tau}{p_\tau}\|\nabla(u-\Pi u)\|_{L^2(\tau)}\le C\frac{h_\tau}{p_\tau}\|\nabla u\|_{L^2(\omega_\tau)}, \tag{6.10}$$

$$\|u-\Pi u\|_{L^2(e)}\le C\sqrt{\frac{h_e}{p_e}}\|\nabla u\|_{L^2(\omega_e)}, \tag{6.11}$$

where h_e is the length of the edge e and $p_e=\max(p_\tau,p_{\tau'})$, where τ,τ' are elements sharing the edge e, and ω_τ, ω_e are patches covering τ and e with a few layers, respectively. See [25] *for more details on ω_τ and ω_e.*

LEMMA 6.2 (Scott–Zhang-type quasi interpolation). *Let $\mathcal{T}$ be a γ-shape regular triangulation (see* (6.1)*) of a domain $\Omega\in\mathcal{R}^2$, and let $\boldsymbol{p}$ be a polynomial degree distribution which is comparable (see* (6.2)*). Then there exists a linear operator $\hat{\Pi}: H_0^1(\Omega)\to S_0^{\boldsymbol{p}}(\mathcal{T},\Omega)$, and there exists a constant $C>0$, which depends only on γ, such that for every $u\in H_0^1(\Omega)$, all elements $\tau\in\mathcal{T}$, and all edges $e\in\mathcal{E}(\tau)$,*

$$\|u-\hat{\Pi}u\|_{L^2(\tau)}+\frac{h_\tau}{p_\tau}\|\nabla(u-\hat{\Pi}u)\|_{L^2(\tau)}\le C\frac{h_\tau}{p_\tau}\|\nabla u\|_{L^2(\omega_\tau)}, \tag{6.12}$$

$$\|u-\hat{\Pi}u\|_{L^2(e)}\le C\sqrt{\frac{h_e}{p_e}}\|\nabla u\|_{L^2(\omega_e)}, \tag{6.13}$$

where h_e is the length of the edge e and $p_e=\max(p_\tau,p_{\tau'})$, where τ,τ' are elements sharing the edge e, and ω_τ, ω_e are patches covering τ and e with a few layers, respectively. See [25] *for more details on ω_τ and ω_e.*

In what follows, we will derive a posteriori estimates for the optimal control problem. First, we define some notation to represent the error bound as follows:

$$\eta_2^2:=\sum_{\tau\in\mathcal{T}}\left(\frac{h_\tau^2}{p_\tau^2}\|f+u_{hp}+\Delta y_{hp}\|_{L^2(\tau)}^2+\sum_{e\subset(\partial\tau\bigcap\Omega)\bigcap\mathcal{E}_0(\mathcal{T})}\frac{h_e}{2p_e}\left\|\left[\frac{\partial y_{hp}}{n_e}\right]\right\|_{L^2(e)}^2\right), \tag{6.14}$$

$$\eta_3^2:=\sum_{\tau\in\mathcal{T}}\left(\frac{h_\tau^2}{p_\tau^2}\|y_{hp}-y_0+\Delta p_{hp}\|_{L^2(\tau)}^2+\sum_{e\subset(\partial\tau\bigcap\Omega)\bigcap\mathcal{E}_0(\mathcal{T})}\frac{h_e}{2p_e}\left\|\left[\frac{\partial p_{hp}}{n_e}\right]\right\|_{L^2(e)}^2\right), \tag{6.15}$$

$$\eta_4^2:=\sum_{\tau\in\mathcal{T}}\frac{h_\tau^2}{p_\tau^2}\|\nabla(u_{hp}+p_{hp})\|_{L^2(\tau)}^2, \tag{6.16}$$

where we denote the jump of v across the edges by $[v]$ and n_e is the unit outer normal on e.

Using the above notation, we now give the following a posteriori error estimates.

LEMMA 6.3. *Let (y,p,u), (y_{hp},p_{hp},u_{hp}), and $(y(u_{hp}),p(u_{hp}))$ be the solutions of* (2.3)–(2.5)*,* (6.5)–(6.7)*, and* (6.8)–(6.9)*, respectively. Then we have*

$$\|u-u_{hp}\|_{L^2(\Omega)}^2\le C(\eta_4^2+\|p_{hp}-p(u_{hp})\|_{L^2(\Omega)}^2), \tag{6.17}$$

where η_4 is defined by (6.16)*.*

Proof. It follows from (2.5) and (6.7) that for any $v_{hp}\in K^{h,\boldsymbol{p}}$

$$\begin{aligned}||u-u_{hp}||_{L^2(\Omega)}^2&=(u,u-u_{hp})_U-(u_{hp},u-u_{hp})_U\\&\le(-p,u-u_{hp})_U-(u_{hp},u-u_{hp})_U+(u_{hp}+p_{hp},v_{hp}-u_{hp})_U\\&=(u_{hp}+p_{hp},v_{hp}-u)_U+(p_{hp}-p,u-u_{hp})_U.\end{aligned} \tag{6.18}$$

It follows from (2.3), (2.4), (6.8), and (6.9) that we have

$$a(y(u_{hp}) - y, w) = (u_{hp} - u, w) \qquad \forall w \in H_0^1(\Omega),$$

$$a(q, p(u_{hp}) - p) = (y(u_{hp}) - y, q) \qquad \forall q \in H_0^1(\Omega).$$

Let $w = p(u_{hp}) - p$ and $q = y(u_{hp}) - y$. Then

$$(u_{hp} - u, p(u_{hp}) - p) = (y(u_{hp}) - y, y(u_{hp}) - y) \geq 0.$$

It follows from (6.18) that we have

$$\begin{aligned} ||u - u_{hp}||^2_{L^2(\Omega)} &\leq (u_{hp} + p_{hp}, v_{hp} - u)_U + (p_{hp} - p(u_{hp}), u - u_{hp})_U \\ &\quad - (p(u_{hp}) - p, u_{hp} - u)_U \\ &\leq \sum_\tau (u_{hp} + p_{hp}, v_{hp} - u)_\tau + C(\delta)\|p_{hp} - p(u_{hp})\|^2_{H^1(\Omega)} \\ &\quad + \delta||u - u_{hp}||^2_{L^2(\Omega)}. \end{aligned}$$

Let $v_{hp} = Pu$ be the L^2-projection of u in $U^{\boldsymbol{p}}(\mathcal{T}, \Omega)$. As in Theorem 4.1, we have $Pu \in K^{h,\boldsymbol{p}}$. Let Π be the Clément-type interpolator defined in Lemma 6.1. Then we have

(6.19)
$$\begin{aligned} ||u - u_{hp}||^2_{L^2(\Omega)} &\leq \sum_\tau (u_{hp} + p_{hp}, v_{hp} - u)_\tau + C(\delta)\|p_{hp} - p(u_{hp})\|^2_{H^1(\Omega)} \\ &\quad + \delta||u - u_{hp}||^2_{L^2(\Omega)} \\ &= \sum_\tau (u_{hp} + p_{hp} - \Pi(u_{hp} + p_{hp}), P(u - u_{hp}) - (u - u_{hp}))_\tau \\ &\quad + C(\delta)\|p_{hp} - p(u_{hp})\|^2_{H^1(\Omega)} + \delta||u - u_{hp}||^2_{L^2(\Omega)} \\ &\leq C(\delta) \sum_\tau \frac{h_\tau^2}{p_\tau^2} \|\nabla(u_{hp} + p_{hp})\|^2_{L^2(\tau)} + C(\delta)\|p_{hp} - p(u_{hp})\|^2_{H^1(\Omega)} \\ &\quad + 2\delta||u - u_{hp}||^2_{L^2(\Omega)}. \end{aligned}$$

Now choosing $\delta = \frac{1}{4}$, we obtain (6.17). ∎

In the following lemma, we derive a posteriori error estimates for the intermediate variables.

LEMMA 6.4. *Let* (y_{hp}, p_{hp}, u_{hp}) *and* $(y(u_{hp}), p(u_{hp}))$ *be the solutions of* (6.5)–(6.7) *and* (6.8)–(6.9)*, respectively. Then we have*

$$\|y(u_{hp}) - y_{hp}\|^2_{H^1(\Omega)} + \|p(u_{hp}) - p_{hp}\|^2_{H^1(\Omega)} \leq C(\eta_2^2 + \eta_3^2), \tag{6.20}$$

where η_2, η_3 *are defined by* (6.14), (6.15).

Proof. Set $e^p = p(u_{hp}) - p_{hp}$, and let $e_I^p = \hat{\Pi} e^p$, where $\hat{\Pi} : H_0^1(\Omega) \to S_0^{\boldsymbol{p}}(\mathcal{T}, \Omega)$ is the Scott–Zhang-type interpolator defined in Lemma 6.2. Then it follows from (6.6), (6.9), and Lemma 6.2 that

(6.21)
$$
\begin{aligned}
c||p(u_{hp})-p_{hp}||^2_{H^1(\Omega)} &\le (\nabla e^p,\nabla e^p)=(\nabla(e^p-e_I^p),\nabla e^p)+(\nabla e_I^p,\nabla e^p)\\
&=(y(u_{hp})-y_0,e^p)+(\nabla(e^p-e_I^p),\nabla p_{hp})+(y_{hp}-y_0,e_I^p)\\
&=\sum_{\tau\in\mathcal{T}}\int_\tau (y_{hp}-y_0+\Delta p_{hp})(e^p-e_I^p)\\
&\quad-\sum_{e\in\mathcal{E}_0(\mathcal{T})}\int_e\left[\frac{\partial p_{hp}}{n_e}\right](e^p-e_I^p)+(y(u_{hp})-y_{hp},e^p)\\
&\le C\sum_{\tau\in\mathcal{T}}\|y_{hp}-y_0+\Delta p_{hp}\|_{L^2(\tau)}\|e^p-e_I^p\|_{L^2(\tau)}\\
&\quad+C\sum_{e\in\mathcal{E}_0(\mathcal{T})}\left\|\left[\frac{\partial p_{hp}}{n_e}\right]\right\|_{L^2(e)}\|e^p-e_I^p\|_{L^2(e)}\\
&\quad+C\|y(u_{hp})-y_{hp}\|_{L^2(\Omega)}\|e^p\|_{L^2(\Omega)}\\
&\le C\sum_{\tau\in\mathcal{T}}\frac{h_\tau}{p_\tau}\|y_{hp}-y_0+\Delta p_{hp}\|_{L^2(\tau)}\|\nabla e^p\|_{L^2(\omega_\tau)}\\
&\quad+C\sum_{e\in\mathcal{E}_0(\mathcal{T})}\sqrt{\frac{h_e}{p_e}}\left\|\left[\frac{\partial p_{hp}}{n_e}\right]\right\|_{L^2(e)}\|\nabla e^p\|_{L^2(\omega_e)}\\
&\quad+C\|y(u_{hp})-y_{hp}\|_{L^2(\Omega)}\|e^p\|_{L^2(\Omega)}\\
&\le C(\delta)\eta_3^2+C(\delta)\|y(u_{hp})-y_{hp}\|^2_{L^2(\Omega)}+\delta\|e^p\|^2_{H^1(\Omega)}.
\end{aligned}
$$

Letting $\delta=\frac{c}{2}$, we have

$$
||p(u_{hp})-p_{hp}||^2_{H^1(\Omega)}\le C\eta_3^2+C\|y(u_{hp})-y_{hp}\|^2_{L^2(\Omega)}. \tag{6.22}
$$

Similarly, set $e^y=y(u_{hp})-y_{hp}$, and let $e_I^y=\hat{\Pi}e^y$ be the Scott–Zhang interpolator of e^y defined in Lemma 6.2. Then, using (6.5), (6.8), and Lemma 6.3, we have that

$$
\begin{aligned}
c||y(u_{hp})-y_{hp}||^2_{H^1(\Omega)} &\le (\nabla e^y,\nabla e^y)=(\nabla(e^y-e_I^y),\nabla e^y)\\
&=\sum_{\tau\in\mathcal{T}}\int_\tau (f+u_{hp}+\Delta y_{hp})(e^y-e_I^y)\\
&\quad-\sum_{e\in\mathcal{E}_0(\mathcal{T})}\int_e\left[\frac{\partial y_{hp}}{n_e}\right](e^p-e_I^p)\\
&\le C\sum_{\tau\in\mathcal{T}}\frac{h_\tau}{p_\tau}\|f+u_{hp}+\Delta y_{hp}\|_{L^2(\tau)}\|\nabla e^y\|_{L^2(\omega_\tau)}\\
&\quad+C\sum_{e\in\mathcal{E}_0(\mathcal{T})}\sqrt{\frac{h_e}{p_e}}\left\|\left[\frac{\partial y_{hp}}{n_e}\right]\right\|_{L^2(e)}\|\nabla e^y\|_{L^2(\omega_e)}\\
&\le C(\delta)\eta_2^2+\delta\|e^y\|^2_{H^1(\Omega)}.
\end{aligned}
$$

Choosing that $\delta=\frac{1}{2c_1}$, we have

$$
||y(u_{hp})-y_{hp}||^2_{H^1(\Omega)}\le C\eta_2^2. \tag{6.23}
$$

Then, we combine (6.22) and (6.23) to derive our estimates (6.20). ∎

THEOREM 6.1. *Let* (y,p,u) *and* (y_{hp},p_{hp},u_{hp}) *be the solutions of* (2.3)–(2.5) *and* (6.5)–(6.7)*, respectively. Then we have that*

$$||u-u_{hp}||^2_{L^2(\Omega)}+||y-y_{hp}||^2_{H^1(\Omega)}+||p-p_{hp}||^2_{H^1(\Omega)}\le C(\eta_2^2+\eta_3^2+\eta_4^2), \tag{6.24}$$

where $\eta_2^2,\eta_3^2,\eta_4^2$ *are defined by* (6.14)–(6.16).

Proof. We can prove (6.24) as in Theorem 5.1 by using Lemmas 6.3 and 6.4. ∎

7. A preconditioning projection algorithm. In this section, we will present four numerical examples. For the minimization problem consider

$$\min_{u\in K} H(u)=\left\{\frac{1}{2}\int_\Omega (y-y_0)^2+\frac{1}{2}\int_\Omega u^2\right\} \tag{7.1}$$

and

$$K=\left\{v\in L^2(\Omega):\int_\Omega v\geqslant 0\right\}\subset U=L^2(\Omega).$$

We define the iterative scheme $(n=0,1,2,\dots)$

$$\begin{cases}(u_{n+\frac{1}{2}},v)=(u_n,v)-\rho_n(H^{'}(u_n),v) & \forall v\in U,\\ u_{n+1}=P_K(u_{n+\frac{1}{2}}),\end{cases} \tag{7.2}$$

where ρ_n will be specified later. Now define the projection operator $P_K: U\to K$ satisfying the following: For given $w\in U$ find $P_Kw\in K$ such that

$$(P_Kw-w,P_Kw-w)=\min_{u\in K}(u-w,u-w),$$

which is equivalent to

$$(P_Kw-w,v-P_Kw)\ge 0 \qquad \forall v\ \in K. \tag{7.3}$$

Let U_n be the coordinates of u_n in R^n, and let $DH(U_n)$ be the gradient of $H(U_n)$. Then the matrix form of the above algorithm reads as

$$U_{n+\frac{1}{2}}=U_n-\rho_nM_N^{-1}DH(u_n),$$

where M_N is the mass matrix of $(\cdot,\cdot)_{U_N}$, the preconditioner. For the above scheme we have the following convergence result (see [19]).

LEMMA 7.1. *Assume that* $H^{'}$ *is Lipschitz and uniformly monotone in the sense that there are positive constants* c, C *such that*

$$|H^{'}(u)-H^{'}(v)|\le C||u-v||_U \qquad \forall u,v\in U,$$
$$(H^{'}(u)-H^{'}(v),u-v)\ge c||u-v||^2_U \qquad \forall u,v\in U.$$

Then there are $0<\delta<1,\ \epsilon>0$ *such that*

$$||u-u_n||\le\delta^n||u-u_0||, \qquad n=0,1,2,\dots,$$

provided that $\rho_n<\epsilon$.

Remark 7.1. To have the convergence result, we should select ρ_n satisfying $0\le 1+\rho_n(C\rho_n-c)\le\delta$.

Now, it is obvious that the objection functional in (1.1)

$$H(u) = J(u) = \left\{ \frac{1}{2} \int_\Omega (y(u) - y_0)^2 + \frac{1}{2} \int_\Omega u^2 \right\}$$

satisfies all the conditions in Lemma 6.1. Thus, we can apply the above algorithm to compute the discrete elliptic optimal control problem $(\mathrm{OCP})^N$.

Let U_N and K_N be defined in section 2, and let P_{K_N} be the discrete projection operator $P_{K_N} : U_N \to K_N$ satisfying the following: For given $w \in U_N$ find $P_{K_N} w \in K_N$ such that

$$(P_{K_N} w - w, P_{K_N} w - w) = \min_{u \in K_N} (u - w, u - w),$$

which is equivalent to

$$(P_{K_N} w - w, v - P_{K_N} w) \geq 0 \qquad \forall v \ \in K_N. \tag{7.4}$$

It follows from Theorem 3.2 that for any $u_N \in U_N$

$$P_{K_N} u_N = -\min\{0, \overline{u_N}\} + u_N.$$

Then it follows from (7.2) that

$$u_{n+1} = P_{K_N}(u_{n+\frac{1}{2}}) = -\min\{0, \overline{u_{n+\frac{1}{2}}}\} + u_{n+\frac{1}{2}}.$$

For the discrete space U_N, we can apply (7.2) to construct the following algorithm for the discrete elliptic control problem $(\mathrm{OCP})^N$.

Algorithm.

$$\begin{array}{ll} (u_{n+\frac{1}{2}}, v) = (u_n, v) - \rho_n (J'(u_n), v), & u_{n+\frac{1}{2}}, u_n \in U_N \quad \forall v \in U_N, \\ a(y_n, w) = (f + u_n, w), & y_n \in V_N \quad \forall w \in V_N, \\ a(q, p_n) = (y_n - y_0, q), & p_n \in V_N \quad \forall q \in V_N, \\ u_{n+1} = P_{K_N}(u_{n+\frac{1}{2}}) = -\min\{0, \overline{u_{n+\frac{1}{2}}}\} + u_{n+\frac{1}{2}}, & \end{array} \tag{7.5}$$

where we have omitted the subscript N for all the functions in the discrete spaces. In practice, we select the step size $\rho_n \in (0, 1/4)$.

7.1. One-dimensional case. Following [30], we choose the $\phi_k = c_k(L_k(x) - L_{k+2}(x))$ as the basis functions in the spectral approximation with

$$c_k = \frac{1}{\sqrt{4k+6}}.$$

Let $V_N = \mathrm{span}\{\phi_0, \phi_1, \ldots, \phi_{N-2}\}$. It can be verified that (see [30])

$$a_{jk} = (\phi_k', \phi_j') = \begin{cases} 1, & k = j, \\ 0, & k \neq j, \end{cases}$$

$$b_{jk} = b_{kj} = (\phi_k, \phi_j) = \begin{cases} c_k c_j \left(\frac{2}{2j+1} + \frac{2}{2j+5}\right), & k = j, \\ -c_k c_j \frac{2}{2k+1}, & k = j+2, \\ 0 & \text{otherwise.} \end{cases}$$

Example 1. In this example, we let

$$\begin{array}{lll} p = \sin(\pi x), & f = (1 + \pi^4)p, & y_0 = 0, \\ y = y_0 + \pi^2 p, & u = \max\{0, \overline{p}\} - p. & \end{array}$$

Let us denote

$$
\begin{aligned}
&f_k = (P_N f, \phi_k), \qquad F = (f_0, f_1, \ldots, f_{N-2})^T,\\
&g_k = (-P_N y_0, \phi_k), \qquad G = (g_0, g_1, \ldots, g_{N-2})^T,\\
&y_N = \sum_{n=0}^{N-2} \varphi_n \phi_n(x), \qquad \varphi = (\varphi_0, \varphi_1, \ldots, \varphi_{N-2})^T,\\
&p_N = \sum_{n=0}^{N-2} \psi_n \phi_n(x), \qquad \psi = (\psi_0, \psi_1, \ldots, \psi_{N-2})^T,\\
&u_N = \sum_{n=0}^{N-2} \tau_n \phi_n(x), \qquad \tau = (\tau_0, \tau_1, \ldots, \tau_{N-2})^T,\\
&A = (a_{jk}), \qquad B = (b_{jk}), \qquad 0 \le j, k \le N-2.
\end{aligned}
$$

Then, in this case, the algorithm (7.5) is equivalent to

$$
\left\{
\begin{array}{l}
Bu_{n+\frac{1}{2}} = Bu_n - \rho_n(u_n + p_n),\\
Ay_n = F + Bu_n,\\
Ap_n = By_n + G,\\
u_{n+1} = -\min\{0, \overline{u_{n+\frac{1}{2}}}\} + u_{n+\frac{1}{2}}.
\end{array}
\right. \tag{7.6}
$$

We now describe the main steps for solving numerical Example 1:

- *Step* 1. Compute Legendre–Gauss–Lobatto points $\{x_j\}$.
- *Step* 2. Compute the matrixes F, G, A, and B.
- *Step* 3. Use the preconditioning projection algorithm to solve the linear equations (7.6) to obtain $\{\varphi_n\}$, $\{\psi_n\}$, and $\{\tau_n\}$.
- *Step* 4. For every $\{x_j\}$, $j = 0, 1, \ldots, N$, evaluate

$$
\begin{aligned}
y_N(x_j) &= \sum_{n=0}^{N-2} \varphi_n \phi_n(x_j),\\
p_N(x_j) &= \sum_{n=0}^{N-2} \psi_n \phi_n(x_j),\\
u_N(x_j) &= \sum_{n=0}^{N-2} \tau_n \phi_n(x_j).
\end{aligned}
$$

- *Step* 5. Compute the maximum pointwise error of $y - y_N$, $p - p_N$, and $u - u_N$.

The approximation errors are presented in Table 7.1.

We plot the exact solutions and the spectral approximated solutions in Figure 7.1 (in the case of $N = 64$).

TABLE 7.1
The maximum pointwise error of $y - y_N$, $p - p_N$, and $u - u_N$.

N	4	8	16	32	64
$\max\|y - y_N\|$	0.514254	0.000117075	4.04958e-009	4.04958e-009	4.04958e-009
$\max\|p - p_N\|$	0.102066	1.32049e-005	5.22898e-010	5.24445e-010	5.25692e-010
$\max\|u - u_N\|$	0.102066	1.32049e-005	5.22898e-010	5.24444e-010	5.25691e-010

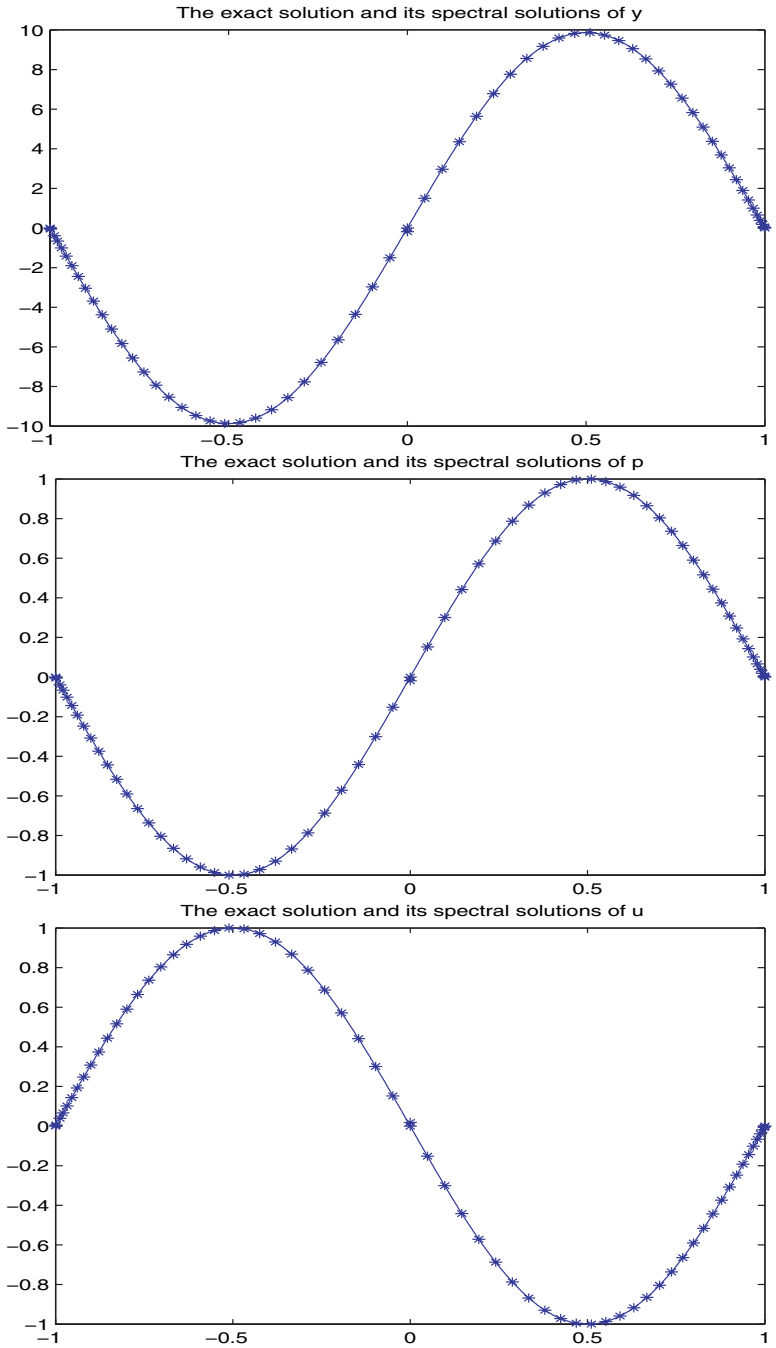


FIG. 7.1. *The exact solutions (solid line) and the approximated solutions (∗).*

Example 2. In this example, we let

$$p = \sin \pi x, \quad y_0 = 1 - |x|^3, \quad y = y_0 + \pi^2 p,$$
$$f = (1 + \pi^4)p, \qquad u = -p.$$

As in Example 1, we can also get the approximation errors (see Table 7.2).

TABLE 7.2
The maximum pointwise error of $y - y_N$, $p - p_N$, and $u - u_N$.

N	4	8	16	32	64
$\max\|y - y_N\|$	0.594161	0.0574371	0.0155496	0.00406608	0.00104092
$\max\|p - p_N\|$	0.142789	0.0201929	0.0052457	0.00134375	0.000340603
$\max\|u - u_N\|$	0.142789	0.0201929	0.0052457	0.00134375	0.000340603

TABLE 7.3
The maximum pointwise error of $y - y_N$, $p - p_N$, and $u - u_N$.

N	4	8	16	32	64
$\max\|y - y_N\|$	0.521042	0.00336548	0.000874284	0.000223959	5.67671e-005
$\max\|p - p_N\|$	0.115384	0.00957285	0.0025916	0.000677681	0.000173486
$\max\|u - u_N\|$	0.115384	0.00957285	0.0025916	0.000677681	0.000173486

In this example, the optimal control is not infinitely smooth due to the absolute function in the initial data. Thus the spectral accuracy is not achieved.

Example 3. In this example, we let

$$p = \sin \pi x + \frac{|x|^3}{6} - \frac{1}{6}, \quad u = -p, \quad y = \pi^2 \sin \pi x,$$
$$y_0 = |x|, \qquad f = (1 + \pi^4) \sin \pi x + \frac{|x|^3}{6} - \frac{1}{6}.$$

We can get the approximation errors presented in Table 7.3.

7.2. Two-dimensional case. In the two-dimensional case, the discrete space for the state and costate is constructed as follows:

$$V_N^2 = \operatorname{span}\{\phi_i(x_1)\phi_j(x_2) : \quad i,\ j = 0, 1, \ldots, N-2\}.$$

Let us denote

$$f_{ij} = (P_N f, \phi_i(x_1)\phi_j(x_2)), \qquad F = (f_{ij}),$$
$$g_{ij} = (-P_N y_0, \phi_i(x_1)\phi_j(x_2)), \quad G = (g_{ij}),$$
$$y_N = \sum_{i,j=0}^{N-2} \varphi_{ij}\phi_i(x_1)\phi_j(x_2), \qquad Y = (\varphi_{ij}),$$
$$p_N = \sum_{i,j=0}^{N-2} \psi_{ij}\phi_i(x_1)\phi_j(x_2), \qquad P = (\psi_{ij}),$$
$$u_N = \sum_{i,j=0}^{N-2} \tau_{ij}\phi_i(x_1)\phi_j(x_2), \qquad U = (\tau_{ij}),$$
$$u_{N+\frac{1}{2}} = \sum_{i,j=0}^{N-2} \sigma_{ij}\phi_i(x_1)\phi_j(x_2), \qquad U_{\frac{1}{2}} = (\sigma_{ij}),$$
$$A = (a_{ij}), \qquad B = (b_{ij}).$$

Taking $w_N,\ q_N = \phi_l(x_1)\phi_m(x_2),\ l,m = 0,1,\ldots,N-2$ in (2.8)–(2.9), we find that the algorithm (7.5) is equivalent to the following matrix equations:

$$\begin{aligned}
&BU_{\frac{1}{2}}B = BUB - \rho_n B(U+P)B,\\
&YB + BY = F + BUB,\\
&PB + BP = G + BYB,\\
&BUB = cD + BU_{\frac{1}{2}}B,
\end{aligned}$$

where $c = -\min\{0, \overline{u_{\frac{1}{2}}}\}$ and $D = (d_{ij}) = (1, \phi_i(x_1)\phi_j(x_2)),\ d_{00} = \frac{2}{3},\ d_{ij} = 0$, when $i \neq 0,\ j \neq 0$.

In particular, these equations can be solved by the matrix decomposition method described in [5], which is well known in the field of spectral methods as the matrix diagonalization method [14]. To this end, let Λ be the diagonal matrix whose diagonal entries $\{\lambda_p\}$ are the eigenvalues of B, and let E be the matrix whose columns are the eigenvectors of B. Then we have

$$E^{-1}BE = \Lambda.$$

Now setting $Y = EV,\ P = EW,\ U = EQ,\ U_{\frac{1}{2}} = EQ_{\frac{1}{2}}$, the equations become

$$\begin{aligned}
&Q^T_{\frac{1}{2}} = Q^T - \rho_n(Q^T + W^T),\\
&BV^T + V^T\Lambda = F^T E^{-T} + BQ^T\Lambda,\\
&BW^T + W^T\Lambda = G^T E^{-T} + BV^T\Lambda,\\
&Q^T = cB^{-1}D^T\Lambda^{-T} + Q^T_{\frac{1}{2}}.
\end{aligned}$$

These equations can be solved in columns (see [30] for the details).

We now describe the main steps for numerically solving the two-dimensional optimal control problem:

- *Step* 1. Compute the eigenvalues $\{\lambda_p\}$ and eigenvectors e_p of B.
- *Step* 2. Compute $F' = F^T E^{-T},\ G' = G^T E^{-T}$.
- *Step* 3. Obtain V, W, Q.
- *Step* 4. Set $Y = EV$, $P = EW$, $U = EQ$.
- *Step* 5. Obtain $y_N,\ p_N,\ u_N$.
- *Step 6.* Compute the maximum pointwise error of $y - y_N,\ p - p_N$, and $u - u_N$.

Now we apply the algorithm to the following example.

Example 4. In this example, the data and solutions are given as follows:

$$\begin{aligned}
&p = \sin(\pi x_1)\sin(\pi x_2), \quad f = (1 + 4\pi^4)p, \quad y_0 = 0,\\
&y = y_0 + 2\pi^2 p, \qquad u = \max\{0, \overline{p}\} - p.
\end{aligned}$$

We can obtain the approximation errors presented in Table 7.4.

TABLE 7.4
The maximum pointwise error of $y - y_N$, $p - p_N$, and $u - u_N$.

N	4	8	16	32	64
$y - y_N$	0.9223248	0.000348741	8.02892e-009	8.04805e-009	8.09281e-009
$p - p_N$	0.0908673	2.23794e-005	4.87379e-010	4.88399e-010	4.89571e-010
$u - u_N$	0.0908673	2.23794e-005	4.87379e-010	4.88399e-010	4.89571e-010

REFERENCES

[1] R.A. ADAMS, *Sololev Spaces*, Academic Press, New York, 1975.

[2] M. AINSWORTH AND J.T. ODEN, *A posteriori error estimators in finite element analysis*, Comput. Methods Appl. Mech. Engrg., 142 (1997), pp. 1–88.

[3] R. BECKER, H. KAPP, AND R. RANNACHER, *Adaptive finite element methods for optimal control of partial differential equations: Basic concept*, SIAM J. Control Optim., 39 (2000), pp. 113–132.

[4] C. BERNARDI, *Indicateurs d'erreur en h-N version des element spectraux*, RAIRO Modél. Math. Anal. Numér., 30 (1996), pp. 1–38.

[5] B.L. BUZBEE, G.H. GOLUB, AND C.W. NIELSON, *On direct methods for solving Poisson's equations*, SIAM J. Numer. Anal., 7 (1970), pp. 627–656.

[6] C. CANUTO, M.Y. HUSSAINI, A. QUARTERONI, AND T.A. ZANG, *Spectral Methods in Fluid Dynamics*, Springer-Verlag, Berlin, 1988.

[7] E. CASAS AND F. TRÖLTZSCH, *Second-order necessary and sufficient optimality conditions for optimization problems and applications to control theory*, SIAM. J. Optim., 13 (2002), pp. 406–431.

[8] P.G. CIARLET, *The Finite Element Method for Elliptic Problems*, North–Holland, Amsterdam, 1978.

[9] PH. CLÉMENT, *Approximation by finite element functions using local regularization*, RAIRO Anal. Numér., 9 (1975), pp. 77–84.

[10] R. GHANEM AND H. SISSAOUI, *A posteriori error estimate by a spectral method of an elliptic optimal control problem*, J. Comput. Math. Optim., 2 (2006), pp. 111–125.

[11] D. GOTTLIEB AND S.A. ORSZAG, *Numerical Analysis of Spectral Methods: Theory and Applications*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 26, SIAM, Philadelphia, 1977.

[12] M.D. GUNZBURGER AND L.S. HOU, *Finite-dimensional approximation of a class of constrained nonlinear optimal control problems*, SIAM J. Control Optim., 34 (1996), pp. 1001–1043.

[13] B. GUO, *Spectral Methods and Their Applications*, World Scientific, River Edge, NJ, 1998.

[14] D.B. HAIDVOGEL AND T.A. ZANG, *The accurate solution of Poisson's equation by expansion in Chebyshev polynomials*, J. Comput. Phys., 30 (1979), pp. 167–180.

[15] M. HINZE AND K. KUNISCH, *Second order methods for optimal control of time-dependent fluid flow*, SIAM J. Control Optim., 40 (2001), pp. 925–946.

[16] Y. HUANG, R. LI, W.B. LIU, AND N.N. YAN, *Adaptive multi-mesh finite element approximation for constrained optimal control problems*, SIAM J. Control Optim., submitted.

[17] A. KUFNER, O. JOHN, AND S. FUCIK, *Function Spaces*, Nordhoff, Leyden, The Netherlands, 1977.

[18] O.A. LADYZHENSKAYA AND N.N. URAĹTSEVA, *Linear and Quasilinear Elliptic Equations*, Academic Press, New York, 1968.

[19] R. LI, W. LIU, H. MA, AND T. TANG, *Adaptive finite element approximation for distributed elliptic optimal control problems*, SIAM J. Control Optim., 41 (2002), pp. 1321–1349.

[20] J.L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, Berlin, 1971.

[21] W.B. LIU AND D. TIBA, *Error estimates for the finite element approximation of a class of nonlinear optimal control problems*, J. Numer. Funct. Optim., 22 (2001), pp. 953–972.

[22] W.B. LIU AND N.N. YAN, *A posteriori error analysis for convex distributed optimal control problems*, Adv. Comput. Math., 15 (2001), pp. 285–309.

[23] W.B. LIU AND N.N. YAN, *A posteriori error estimates for optimal control problems governed by parabolic equations*, Numer. Math., 93 (2003), pp. 497–521.

[24] W. LIU AND N. YAN, *A posteriori error estimates for control problems governed by Stokes equations*, SIAM J. Numer. Anal., 40 (2002), pp. 1850–1869.

[25] J.M. MELENK, *hp-interpolation of nonsmooth functions and an application to hp-a posteriori error estimation*, SIAM. J. Numer. Anal., 43 (2005), pp. 127–155.

[26] J.M. Melenk and B.I. Wohlmuth, *On residual-based a posteriori error estimation in hp-FEM*, Adv. Comput. Math., 15 (2001), pp. 311–331.
[27] A.T. Patera, *A spectral element method for fluid dynamics: Laminar flow in a channel expansion*, J. Comput. Phys., 54 (1984), pp. 468–488.
[28] L.R. Scott and S. Zhang, *Finite element interpolation of nonsmooth functions satisfying boundary conditions*, Math. Comp., 54 (1990), pp. 483–493.
[29] J. Shen, *A spectral-tau approximation for the Stokes and Navier-Stokes equations*, RAIRO Modél. Math. Anal. Numér., 22 (1988), pp. 677–693.
[30] J. Shen, *Efficient spectral-Galerkin method* I. *Direct solvers of second- and fourth-order equations using Legendre polynomials*, SIAM J. Sci. Comput., 15 (1994), pp. 1489–1505.
[31] J. Shen, *Efficient spectral-Galerkin method* II. *Direct solvers of second- and fourth-order equations using Chebyshev polynomials*, SIAM J. Sci. Comput., 16 (1995), pp. 74–87.

SIAM J. NUMER. ANAL.
Vol. 46, No. 5, pp. 2276–2308
© 2008 Society for Industrial and Applied Mathematics

CONDITIONAL STABILITY AND CONVERGENCE OF A FULLY DISCRETE SCHEME FOR THREE-DIMENSIONAL NAVIER–STOKES EQUATIONS WITH MASS DIFFUSION*

F. GUILLÉN-GONZÁLEZ† AND J. V. GUTIÉRREZ-SANTACREU†

Abstract. We construct a fully discrete numerical scheme for three-dimensional incompressible fluids with mass diffusion (in density-velocity-pressure formulation), also called the Kazhikhov–Smagulov model. We will prove conditional stability and convergence, by using at most C^0-finite elements, although the density of the limit problem will have H^2-regularity. The key idea of our argument is first to obtain pointwise estimates for the discrete density by imposing the constraint $\lim_{(h,k)\to 0} h/k = 0$ on the time and space parameters (k, h). Afterwards, under the same constraint on the parameters, strong estimates for the discrete density in $l^\infty(H^1)$ and for the discrete Laplacian of the density in $l^2(L^2)$ are obtained. From here, the compactness and convergence of the scheme can be concluded with similar arguments as we used in [*Math. Comp.*, to appear], where a different scheme is studied for two-dimensional domains which is unconditionally stable and convergent. Moreover, we study the asymptotic behavior of the numerical scheme as the diffusion parameter λ goes to zero, obtaining convergence as $(k, h, \lambda) \to 0$ towards a weak solution of the density-dependent Navier–Stokes system provided that the constraint $\lim_{(\lambda,h,k)\to 0} h/(\lambda^2 k) = 0$ on (h, k, λ) is satisfied.

Key words. three-dimensional Kazhikhov–Smagulov models, density-dependent Navier–Stokes equations, finite elements, stability, convergence

AMS subject classifications. 35Q35, 65M12, 65M60

DOI. 10.1137/07067951X

1. Introduction.

1.1. The model. Let $\Omega \subset \mathbb{R}^3$ be an open bounded set with boundary Γ. We denote by $[0, T]$ $(0 < T < +\infty)$ the time interval of observation. We will use the notation $Q = \Omega \times (0, T)$, $\Sigma = \Gamma \times (0, T)$, and $\boldsymbol{n}(\boldsymbol{x})$ the outwards unit normal vector to Γ at the point $\boldsymbol{x} \in \Gamma$.

We consider the Navier–Stokes system with mass diffusion (the so-called Kazhikhov–Smagulov model) in Q:

$$\left\{ \begin{array}{rcl} \rho \boldsymbol{u}_t + ((\rho \boldsymbol{u} - \lambda \nabla \rho) \cdot \nabla)\boldsymbol{u} - \nabla \cdot (\mu \nabla \boldsymbol{u} - \lambda \rho (\nabla \boldsymbol{u})^t) + \nabla p &=& \rho \boldsymbol{f}, \\ \nabla \cdot \boldsymbol{u} = 0, \qquad \rho_t + \boldsymbol{u} \cdot \nabla \rho - \lambda \Delta \rho &=& 0. \end{array} \right. \tag{1.1}$$

The unknowns for this model are $\rho : Q \to \mathbb{R}^+$, the density of the fluid, $\boldsymbol{u} : Q \to \mathbb{R}^3$, the incompressible (averaged) velocity vector field, and $p : Q \to \mathbb{R}$, a potential function (modified pressure).

Model (1.1) can be derived from the compressible Navier–Stokes system, by imposing that the compressible velocity $\boldsymbol{v}$ can be decomposed as $\boldsymbol{v} = \boldsymbol{u} - \lambda \nabla \log \rho$, with $\nabla \cdot \boldsymbol{u} = 0$ (it decomposes into an incompressible part $\boldsymbol{u}$ and a potential part $-\lambda \nabla \log \rho$), and eliminating λ^2-terms (see [11]).

*Received by the editors January 9, 2007; accepted for publication (in revised form) January 22, 2008; published electronically May 9, 2008. This work was partially supported by the Spanish project BFM2003–06446-C02-01.

http://www.siam.org/journals/sinum/46-5/67951.html

†Departamento E.D.A.N., University of Sevilla, Aptdo. 1160, 41080 Sevilla, Spain (guillen@us.es, juanvi@us.es).

We complete (1.1) with the boundary conditions on Σ:

$$\boldsymbol{u}_{|\Sigma} = 0, \quad \left.\frac{\partial \rho}{\partial \boldsymbol{n}}\right|_{\Sigma} = 0 \tag{1.2}$$

and the initial conditions in Ω:

$$\rho|_{t=0} = \rho_0, \qquad \boldsymbol{u}|_{t=0} = \boldsymbol{u}_0, \tag{1.3}$$

where $\rho_0 : \Omega \to \mathbb{R}^+$ and $\boldsymbol{u}_0 : \Omega \to \mathbb{R}^d$ are given functions.

Throughout this work, we assume the hypothesis on the initial density:

$$0 < m \leq \rho_0(\boldsymbol{x}) \leq M \quad \text{in } \Omega. \tag{1.4}$$

1.2. Known results. Concerning the simplified model (1.1), Kazhikhov and Smagulov [12] proved, via a semi-Galerkin method, the existence of global weak solutions, under the following hypothesis:

$$\lambda < 2\mu/(M-m), \tag{1.5}$$

and the existence of local strong solutions (which is global in the two-dimensional case). Salvi [14] proved the existence of weak solutions for noncylindrical domains. On the other hand, Secchi [16] studied the problem for $\Omega = \mathbb{R}^3$, proving the local existence and uniqueness of strong solutions by using a fixed point argument.

For the complete model (including the λ^2-terms), Beirão da Veiga [2] and Secchi [15] established the local existence of strong solutions by using linearization and a fixed point argument. In [15], global existence and uniqueness are shown for two-dimensional (2D) domains by imposing that λ/μ is small enough as well as the asymptotic behavior as $\lambda \to 0$ towards a weak solution of the density-dependent Navier–Stokes problem:

$$\left\{\begin{array}{rl} & \rho\left[\boldsymbol{u}_t + (\boldsymbol{u}\cdot\nabla)\boldsymbol{u}\right] - \mu\Delta\boldsymbol{u} + \nabla p = \rho\boldsymbol{f} \quad \text{in } Q, \\ & \nabla\cdot\boldsymbol{u} = 0, \quad \rho_t + \boldsymbol{u}\cdot\nabla\rho = 0 \quad \text{in } Q, \\ \boldsymbol{u} = 0 \quad \text{on } \Sigma, & \boldsymbol{u}|_{t=0} = \boldsymbol{u}_0, \quad \rho|_{t=0} = \rho_0 \quad \text{in } \Omega. \end{array}\right. \tag{1.6}$$

In the case of nonnegative initial density and 3D domains, Guillén-González [9] proved the global existence of weak solutions and the behavior, as $\lambda \to 0$, towards the density-dependent Navier–Stokes system (1.6). Recently, the existence and regularity of strong solutions have been proved in [10] by means of an iterative method (jointly with some error estimates).

A time-space numerical scheme has been recently developed by using C^0-finite elements for density and velocity in [11] for model (1.1) in 2D domains, which is unconditionally stable and convergent towards the (unique) weak solution of the continuous problem. This scheme is of the backward Euler type, where in each time step the computation of the density and the velocity pressure are decoupled, by means of linear problems.

Concerning the numerical analysis for the density-dependent Navier–Stokes problem, a stable and convergent scheme is proposed in [13], by using in particular a discontinuous Galerkin finite element method to approximate the density transport equation.

1.3. Main results of the paper. Our main objective is to design a linear scheme by using finite elements to approximate all unknowns (density, velocity, and pressure) of problem (1.1)–(1.3). To this end, we consider for simplicity a uniform partition of $[0,T]$, $(t_n = nk)_{n=0}^{n=N}$, with $k = T/N$ being the time step, and propose a backward Euler time scheme, implicit with respect to the diffusion terms and semi-implicit with respect to the convection terms. The finite element spaces must verify specific properties which we will describe in section 3.1.

In what follows we consider the notation $(\cdot,\cdot)$ and $|\cdot|$ for the $L^2(\Omega)$-inner product and the $L^2(\Omega)$-norm, respectively. Also, we denote that $\|u\| = |\nabla u|$, which is an equivalent norm to the usual one in $H_0^1(\Omega)$.

The scheme is described as follows:

Initialization: Let $(\boldsymbol{u}_h^0, \rho_h^0) \in \boldsymbol{V}_h \times W_h$ be an approximation of $(\boldsymbol{u}_0, \rho_0)$ as $h \to 0$.

Time step $n+1$: Given $(\rho_h^n, \boldsymbol{u}_h^n, p_h^n) \in W_h \times \boldsymbol{V}_h \times M_h$,

1. find $(\boldsymbol{w}_h^n, q_h^n) \in \widetilde{\boldsymbol{V}}_h \times \widetilde{M}_h$ such that, for each $(\bar{\boldsymbol{w}}_h, \bar{q}_h) \in \widetilde{\boldsymbol{V}}_h \times \widetilde{M}_h$,

$$\left(\nabla \boldsymbol{w}_h^n, \nabla \bar{\boldsymbol{w}}_h\right) - \left(q_h^n, \nabla \cdot \bar{\boldsymbol{w}}_h\right) = \left(\nabla \boldsymbol{u}_h^n, \nabla \bar{\boldsymbol{w}}_h\right), \quad \left(\nabla \cdot \boldsymbol{w}_h^n, \bar{q}_h\right) = 0; \tag{1.7}$$

2. find $\rho_h^{n+1} \in W_h$ such that, for each $\bar{\rho}_h \in W_h$,

$$\left(\frac{\rho_h^{n+1} - \rho_h^n}{k}, \bar{\rho}_h\right) + \left(\boldsymbol{w}_h^n \cdot \nabla \rho_h^{n+1}, \bar{\rho}_h\right) + \lambda\left(\nabla \rho_h^{n+1}, \nabla \bar{\rho}_h\right) = 0; \tag{1.8}$$

3. find $(\boldsymbol{u}_h^{n+1}, p_h^{n+1}) \in \boldsymbol{V}_h \times M_h$ such that, for each $(\bar{\boldsymbol{u}}_h, \bar{p}_h) \in \boldsymbol{V}_h \times M_h$,

$$\begin{cases} \left(\rho_h^n \dfrac{\boldsymbol{u}_h^{n+1} - \boldsymbol{u}_h^n}{k}, \bar{\boldsymbol{u}}_h\right) + \dfrac{1}{2}\left(\dfrac{\rho_h^{n+1} - \rho_h^n}{k}\boldsymbol{u}_h^{n+1}, \bar{\boldsymbol{u}}_h\right) + a\left(\rho_h^{n+1}, \boldsymbol{u}_h^{n+1}, \bar{\boldsymbol{u}}_h\right) \\ + c\left(\rho_h^{n+1}\boldsymbol{u}_h^n - \lambda \nabla \rho_h^{n+1}, \boldsymbol{u}_h^{n+1}, \bar{\boldsymbol{u}}_h\right) = \left(\rho_h^{n+1}\boldsymbol{f}^{\,n+1}, \bar{\boldsymbol{u}}_h\right) + \left(p_h^{n+1}, \nabla \cdot \bar{\boldsymbol{u}}_h\right), \end{cases} \tag{1.9}$$

$$\left(\nabla \cdot \boldsymbol{u}_h^{n+1}, \bar{p}_h\right) = 0, \tag{1.10}$$

where

$$\boldsymbol{f}^{\,n+1} = \frac{1}{k}\int_{t_n}^{t_{n+1}} \boldsymbol{f}(t)\,dt,$$

$$a(\rho, \boldsymbol{u}, \boldsymbol{v}) = \mu(\nabla \boldsymbol{u}, \nabla \boldsymbol{v}) + \lambda \int_\Omega \left(\frac{\widetilde{M} + \widetilde{m}}{2} - \rho\right)(\nabla \boldsymbol{u})^t : \nabla \boldsymbol{v}\,d\boldsymbol{x},$$

with $\widetilde{M} > M$, $0 < \widetilde{m} < m$ such that $\lambda\frac{\widetilde{M} - \widetilde{m}}{2} < \mu$ (here (1.5) is imposed), and

$$c(\boldsymbol{w}, \boldsymbol{u}, \boldsymbol{v}) = \frac{1}{2}\left[((\boldsymbol{w} \cdot \nabla)\boldsymbol{u}, \boldsymbol{v}) - \left((\boldsymbol{w} \cdot \nabla)\boldsymbol{v}, \boldsymbol{u}\right)\right].$$

The following properties of continuity and coercivity hold:

$$\begin{aligned} &a(\rho, \boldsymbol{u}, \boldsymbol{u}) \geq \frac{\mu_1}{2}\|\boldsymbol{u}\|^2 \ \text{ if } \ \widetilde{m} \leq \rho \leq \widetilde{M}, \ \text{ with } \ \frac{\mu_1}{2} = \mu - \lambda\frac{\widetilde{M} - \widetilde{m}}{2} (> 0), \\ &a(\rho, \boldsymbol{u}, \boldsymbol{v}) \leq C_\lambda \|\boldsymbol{u}\|\,\|\boldsymbol{v}\| \quad (\text{if } \|\rho\|_{L^\infty(\Omega)} \leq C), \\ &c(\boldsymbol{w}, \boldsymbol{u}, \boldsymbol{u}) = 0, \quad c(\boldsymbol{w}, \boldsymbol{u}, \boldsymbol{v}) \leq C\,\|\boldsymbol{w}\|_{L^3}\|\boldsymbol{u}\|\,\|\boldsymbol{v}\|. \end{aligned} \tag{1.11}$$

Here and in what follows, we denote by C_λ and C different positive constants independent of (h,k) and (h,k,λ), respectively.

From the computational point of view, scheme (1.7)–(1.10) decouples density ρ_h^{n+1} and velocity pressure $(\boldsymbol{u}_h^{n+1}, p_h^{n+1})$, whereas $\boldsymbol{w}_h^n$ is an intermediate velocity obtained as the H^1 orthogonal projection of $\boldsymbol{u}_h^n$ onto a discrete free-divergence space. We will see that scheme (1.7)–(1.10) is conditionally stable and convergent. As in many practical situations, the diffusion parameter is small; we will prove that, when the diffusion parameter λ and the space and time parameters (h,k) goes to zero, scheme (1.7)–(1.10) approximates to a weak solution of the density-dependent Navier–Stokes system (1.6), under a constraint involving the parameters h, k, and λ. In fact, to our knowledge, it is the first convergent scheme to (1.6) based on C^0-finite elements for the discrete density, avoiding to perform directly an algorithm for (1.6) which presents important difficulties by itself, mainly for the approximation of the density transport equation. Recall that in [13] a convergent scheme for (1.6) is given based on a discontinuous Galerkin method for the density.

The corresponding study for the complete model, with λ^2-terms, will be the subject of a forthcoming paper.

By defining in $[0,T]$ piecewise constant functions $\boldsymbol{u}_{h,k}, \rho_{h,k}$ such that $\boldsymbol{u}_{h,k}$, $\rho_{h,k}\big|_{(t_{n-1},t_n]} = \boldsymbol{u}_h^n, \rho_h^n$, respectively, that we will denote by $\boldsymbol{u}_{h,k,\lambda}, \rho_{h,k,\lambda}$ when the case of diffusion parameter $\lambda \to 0$ is considered, we present the following main results of this paper.

THEOREM 1.1. *Assume hypotheses* (H0)–(H5) *given in section* 3.1. *If the constraint on the parameters*

$$h/k \to 0 \quad \text{as } (h,k) \to 0 \tag{S}$$

holds, then there exists a convergent subsequence of $(\boldsymbol{u}_{h,k}, \rho_{h,k})$ *(denoted in the same way) as* $(h,k) \to 0$ *towards a weak solution* $(\boldsymbol{u}, \rho)$ *of problem* (1.1)–(1.3) *(see Definition* 2.1*), in the following sense:* $(\boldsymbol{u}_{h,k}, \rho_{h,k}) \to (\boldsymbol{u}, \rho)$ *in* $L^2(0,T;\boldsymbol{L}^2(\Omega)) \times L^2(0,T;H^1(\Omega))$*-strong, in* $L^\infty(0,T;\boldsymbol{L}^2(\Omega)) \times (H^1(\Omega) \cap L^\infty(\Omega))$*-weak*$\star$*, and in* $L^2(0,T;\boldsymbol{H}_0^1(\Omega)) \times L^4(0,T;W^{1,3}(\Omega))$*-weak. Moreover,* $\widetilde{m} \le \rho_{h,k} \le \widetilde{M}$.

THEOREM 1.2. *Under the hypotheses of Theorem* 1.1 *and by extending* (H2) *by* (H2′) *(given in section* 8.1*) and changing* (S) *by the more restrictive constraint*

$$h/(k\lambda^2) \to 0 \quad \text{as } (\lambda,h,k) \to 0, \tag{S′}$$

then there exists a convergent subsequence of $(\boldsymbol{u}_{h,k,\lambda}, \rho_{h,k,\lambda})$ *as* $(h,k,\lambda) \to 0$ *towards a weak solution* $(\boldsymbol{u}, \rho)$ *of the density-dependent Navier–Stokes problem* (1.6) *(see Definition* 8.1*) in the following sense:* $\boldsymbol{u}_{h,k,\lambda} \to \boldsymbol{u}$ *in* $L^2(0,T;\boldsymbol{L}^2(\Omega))$*-strong, in* $L^\infty(0,T;\boldsymbol{L}^2(\Omega))$*-weak*$\star$*, and in* $L^2(0,T;\boldsymbol{H}_0^1(\Omega))$*-weak, and* $\rho_{h,k,\lambda} \to \rho$ *in* $L^\infty(Q)$*-weak*$\star$.

The main ideas for the derivation of this scheme can be found in [11], where the following scheme has been studied:

Time step $(n+1)$: Given $(\rho_h^n, \boldsymbol{u}_h^n, p_h^n) \in W_h \times \boldsymbol{V}_h \times M_h$,

1. find $\rho_h^{n+1} \in W_h$ such that, for each $\bar{\rho}_h \in W_h$,

$$\left(\frac{\rho_h^{n+1} - \rho_h^n}{k}, \bar{\rho}_h\right) + \lambda\Big(\nabla\rho_h^{n+1}, \nabla\bar{\rho}_h\Big) = -\Big(\boldsymbol{u}_h^n \cdot \nabla\rho_h^n, \bar{\rho}_h\Big); \tag{1.12}$$

2. find $(\boldsymbol{u}_h^{n+1}, p_h^{n+1}) \in \boldsymbol{V}_h \times M_h$ such that, for each $(\bar{\boldsymbol{u}}_h, \bar{p}_h) \in \boldsymbol{V}_h \times M_h$,

$$\left\{\begin{array}{l} \left([\rho_h^n]_T \dfrac{\boldsymbol{u}_h^{n+1}-\boldsymbol{u}_h^n}{k}, \bar{\boldsymbol{u}}_h\right) + \dfrac{1}{2}\left(\dfrac{[\rho_h^{n+1}]_T - [\rho_h^n]_T}{k}, \boldsymbol{u}_h^{n+1}\cdot\bar{\boldsymbol{u}}_h\right) \\ +a\Big([\rho_h^{n+1}]_T, \boldsymbol{u}_h^{n+1}, \bar{\boldsymbol{u}}_h\Big) + c\Big(\rho_h^{n+1}\boldsymbol{u}_h^n - \lambda\nabla\rho_h^{n+1}, \boldsymbol{u}_h^{n+1}, \bar{\boldsymbol{u}}_h\Big) \\ -\Big(p_h^{n+1}, \nabla\cdot\bar{\boldsymbol{u}}_h\Big) = \Big([\rho_h^{n+1}]_T \boldsymbol{f}^{\,n+1}, \bar{\boldsymbol{u}}_h\Big), \end{array}\right. \tag{1.13}$$

$$\Big(\nabla\cdot\boldsymbol{u}_h^{n+1}, \bar{p}_h\Big) = 0, \tag{1.14}$$

where

$$[w_h]_T(\boldsymbol{x}_i) = \left\{\begin{array}{lll} w_h(\boldsymbol{x}_i) & \text{if} & w_h(\boldsymbol{x}_i)\in[m,M], \\ m & \text{if} & w_h(\boldsymbol{x}_i) < m, \\ M & \text{if} & w_h(\boldsymbol{x}_i) > M, \end{array}\right.$$

with $\boldsymbol{x}_i$ the nodes of the mesh $\mathcal{T}_h$ of Ω.

By comparing both schemes we can observe the following differences: The discrete density involved in the mixed variational problem for velocity pressure (1.13)–(1.14), which requires the property of the maximum principle, are truncated, whereas this truncation is not necessary for scheme (1.7)–(1.10). Moreover, in (1.12) the convective term for the density scheme is considered in the explicit form $(\boldsymbol{u}_h^n\cdot\nabla\rho_h^n, \bar{\rho}_h)$, and it is now taken in the semi-implicit form $(\boldsymbol{w}_h^n\cdot\nabla\rho_h^{n+1}, \bar{\rho}_h)$, where $\boldsymbol{w}_h^n$ is a projection of $\boldsymbol{u}_h^n$ onto a discrete zero-divergence space. This space is chosen to hold $(\nabla\cdot\boldsymbol{w}_h^n, \bar{\rho}_h\bar{\rho}_h) = 0$ for all $\bar{\rho}_h \in W_h$.

Concerning the numerical analysis we remark on the following three main differences between both schemes:

1. The argument to obtain pointwise estimates for the discrete density under the constraints (S) done in subsection 3.4 of this paper is completely new. Moreover, the extension of this argument to the scheme studied in [11] is not clear even assuming some constraints on the discrete parameters. This justifies the presence of the truncation operator in the discrete momentum system (1.13). On the other hand, the scheme (1.12)–(1.14) of [11] is unconditionally stable, and now the scheme (1.7)–(1.10) is stable and convergent under the constraint (S).
2. Strong estimates for the discrete density are obtained in two different ways in [11] and in the present paper. In [11], we used the discrete version of the *Gagliardo–Nirenberg* interpolation inequality $\|\nabla\rho\|_{L^4}^2 \le C\,\|\rho\|_{H^1}\|\Delta\rho\|_{L^2}$ which does not need pointwise estimates for the discrete density. Since this interpolation is exclusive for two-dimensional domains, we cannot use it for three-dimensional domains. Accordingly, we change this Gagliardo–Nirenberg interpolation by a discrete version of the interpolation inequality $\|\nabla\rho\|_{L^4}^2 \le C\,\|\rho\|_{L^\infty}\|\Delta\rho\|_{L^2}$ and make a discrete integration by parts (which mimics the argument of the exact problem to obtain strong estimates of the density). Observe that we have to assure a maximum principle or at least pointwise estimates for the discrete density in order for this other interpolation to work.
3. Another difference is the asymptotic behavior with respect to the diffusion parameter λ (jointly with the discretization parameters (k,h)). Due to the fact that the convective term of the discrete density equation is handled in

different ways, we find that the strong estimates of the discrete density furnished in [11] degenerate when $\lambda \to 0$, and we cannot pass to the limit towards a weak solution of the density-dependent Navier–Stokes problem (1.6). However, now the dependence of λ is improved, and the scheme (1.14)–(1.10) gives a numerical approximation for the density-dependent Navier–Stokes problem (1.6) by means of continuous finite elements.

The rest of the paper can be described as follows. The main ideas for the mathematical analysis of problem (1.1)–(1.3) are provided in section 2. In section 3, by using appropriate auxiliary schemes, we establish conditional stability estimates, energy estimates for the velocity, and pointwise estimates for the density. In section 4, strong estimates for the density are obtained, by using the discrete Laplacian of the density. In sections 5, 6, and 7, weak and strong convergences and the passage to the limit are shown, respectively, concluding the proof of Theorem 1.1. In section 8, we study the asymptotic behavior as the diffusion parameter λ goes to zero, proving Theorem 1.2.

2. Analysis of the continuous model. To define the concept of a weak solution of problem (1.1)–(1.3), we introduce the following function spaces:

$$\begin{aligned}
\boldsymbol{H} &= \{\boldsymbol{u} : \boldsymbol{u} \in \boldsymbol{L}^2(\Omega), \nabla \cdot \boldsymbol{u} = 0 \text{ in } \Omega, \boldsymbol{u} \cdot \boldsymbol{n} = 0 \text{ on } \Gamma\},\\
\boldsymbol{V} &= \{\boldsymbol{u} : \boldsymbol{u} \in \boldsymbol{H}_0^1(\Omega), \nabla \cdot \boldsymbol{u} = 0 \text{ in } \Omega\},\\
L_0^2(\Omega) &= \left\{ p : p \in L^2(\Omega), \int_\Omega p(\boldsymbol{x}) d\boldsymbol{x} = 0 \right\},\\
H_N^2(\Omega) &= \left\{ \rho \in H^2(\Omega) : \frac{\partial \rho}{\partial \boldsymbol{n}} = 0 \text{ on } \Gamma, \int_\Omega \rho(\boldsymbol{x}) d\boldsymbol{x} = \int_\Omega \rho_0(\boldsymbol{x}) d\boldsymbol{x} \right\}.
\end{aligned}$$

In $\boldsymbol{V}$ the $\|\boldsymbol{u}\|_{H^1(\Omega)}$-norm is equivalent to $|\nabla \boldsymbol{u}|$ (which will be denoted by $\|\boldsymbol{u}\|$). $H_N^2(\Omega)$ is an affine space: $H_N^2(\Omega) = \frac{1}{|\Omega|}\int_\Omega \rho_0(\boldsymbol{x})d\boldsymbol{x} + H_{N,0}^2(\Omega)$, and in $H_{N,0}^2(\Omega)$ (zero-average function space) the norm $\|\rho\|_{H^1(\Omega)}$ is equivalent to $|\nabla\rho|$ and the norm $\|\rho\|_{H^2(\Omega)}$ is equivalent to $|\Delta\rho|$. In particular, in $H_N^2(\Omega)$ the following norms are equivalents: $\|\rho - \frac{1}{|\Omega|}\int_\Omega \rho_0\|_{H^1(\Omega)} \sim |\nabla\rho|$ and $\|\nabla\rho\|_{H^1(\Omega)} \sim |\Delta\rho|$.

DEFINITION 2.1. *A pair* $(\rho, \boldsymbol{u})$ *is called a weak solution of* (1.1)–(1.3) *in* $(0,T)$ *if it verifies:*

(a) $\boldsymbol{u} \in L^\infty(0,T;\boldsymbol{H}) \cap L^2(0,T;\boldsymbol{V})$, $\rho \in L^\infty(0,T;H^1(\Omega)) \cap L^2(0,T;H_N^2(\Omega))$, *with* $0 < m \le \rho(\boldsymbol{x},t) \le M$, *a.e.* $(\boldsymbol{x},t) \in Q$.

(b) $\forall\, \phi \in C^1([0,T];\boldsymbol{V})$ *such that* $\phi(T) = 0$,

$$\begin{aligned}
&\int_0^T \left\{ -\Big(\boldsymbol{u}, \rho\phi_t + (\rho\boldsymbol{u} - \lambda\nabla\rho)\cdot\nabla\phi\Big) + \Big(\mu\nabla\boldsymbol{u} - \lambda\rho(\nabla\boldsymbol{u})^t, \nabla\phi\Big)\right\} dt\\
&= \int_0^T \Big(\rho\boldsymbol{f}, \phi\Big)\, dt + \Big(\rho_0\boldsymbol{u}_0, \phi(0)\Big).
\end{aligned}$$

(c) *The equation of mass diffusion* $(1.1)_c$ *is verified a.e. in* Q.

Remark 2.2. As usual, the pressure p can be obtained by using (b) and De Rham's lemma [18].

We state the existence of (global in time) weak solutions of problem (1.1)–(1.3) (see [12, 1]).

THEOREM 2.3. *Let* $\boldsymbol{u}_0 \in \boldsymbol{H}$, $\rho_0 \in H^1(\Omega)$ *satisfying* (1.4), *and* $\boldsymbol{f} \in L^2(0,T;\boldsymbol{L}^{6/5}(\Omega))$. *Suppose that the constants* λ, μ, m, *and* M *satisfy* (1.5). *Then there exists at least a weak solution of* (1.1)–(1.3) *in* $(0,T)$.

Proof (outline of proof). The proof is divided into four steps:

(a) Pointwise estimates for the density. From the maximum principle applied to the density equation $(1.1)_c$ and hypothesis (1.4), one gets

$$0 < m \leq \rho(\boldsymbol{x}, t) \leq M \quad \text{in} \quad Q.$$

(b) Weak estimates for the velocity. Adding the momentum system $(1.1)_a$ by $\boldsymbol{u}$ to the density equation $(1.1)_c$ by $\frac{1}{2}\boldsymbol{u}\cdot\boldsymbol{u}$, one arrives at the following energy equality:

$$\frac{1}{2}\frac{d}{dt}\int_\Omega \rho|\boldsymbol{u}|^2 d\boldsymbol{x} + \mu\|\boldsymbol{u}\|^2 = \lambda\int_\Omega \rho(\nabla\boldsymbol{u})^t : \nabla\boldsymbol{u}\, d\boldsymbol{x} + (\rho\boldsymbol{f}, \boldsymbol{u}). \tag{2.1}$$

The first term on the right-hand side of (2.1) can be rewritten as

$$\lambda\int_\Omega \rho(\nabla\boldsymbol{u})^t : \nabla\boldsymbol{u}\, d\boldsymbol{x} = \lambda\int_\Omega \left(\rho - \frac{M+m}{2}\right)(\nabla\boldsymbol{u})^t : \nabla\boldsymbol{u}\, d\boldsymbol{x} \leq \lambda\frac{M-m}{2}\|u\|^2,$$

where we have used the pointwise inequality $|\,\rho - (M+m)/2\,| \leq (M-m)/2$ (obtained from $m \leq \rho \leq M$). By imposing the constraint on the coefficients (1.5), one arrives at the estimate

$$\max_{0\leq t\leq T}|\boldsymbol{u}(t)|^2 + \int_0^T \|\boldsymbol{u}(t)\|^2 dt \leq C.$$

(c) Strong estimates for the density. By multiplying the density equation $(1.1)_c$ by $-\Delta\rho$ and bounding the convective term (previously integrated by parts) thanks to the interpolation inequality

$$\|\nabla\rho\|_{L^4(\Omega)} \leq C\|\rho\|_\infty^{1/2}|\Delta\rho|^{1/2} \leq C|\Delta\rho|^{1/2}, \tag{2.2}$$

as

$$\int_\Omega \boldsymbol{u}\cdot\nabla\rho\,\Delta\rho \leq C\int_\Omega |\nabla\boldsymbol{u}|\,|\nabla\rho|^2 \leq C\,|\nabla\boldsymbol{u}|\,\|\nabla\rho\|_{L^4}^2 \leq C\,|\nabla\boldsymbol{u}|\,|\Delta\rho|,$$

the following estimate holds:

$$\max_{0\leq t\leq T}|\nabla\rho(t)|^2 + \int_0^T |\Delta\rho(t)|^2 dt \leq C_\lambda.$$

(d) Compactness for the velocity. By using a rather technical argument [1], one can get the following estimate of the "time fractional derivative":

$$\int_0^{T-\delta} |\boldsymbol{u}(t+\delta) - \boldsymbol{u}(t)|^2\, dt \leq C_\lambda\,\delta^{1/2} \quad \forall\delta \in (0,T),$$

which implies [17] compactness for the velocity $\boldsymbol{u}$ in $L^2(0,T;\boldsymbol{L}^2(\Omega))$.

From here, it is rather standard to obtain the existence of weak solutions, by using, for instance, the *semi-Galerkin* method [1]. □

3. Weak and pointwise estimates. Since (1.7), (1.8), and (1.9)–(1.10) can be reduced to three independent algebraic linear systems, it suffices to check the uniqueness of the solution to guarantee that these problems are well-posed. In particular, the uniqueness will be a consequence of the weak and pointwise estimates that we will obtain in this section.

3.1. Hypotheses. Throughout this work the following hypotheses will be assumed:

(H0) Hypotheses for the data: Assume (1.5), and let $\widetilde{M} > M$ and $0 < \widetilde{m} < m$ such that $\lambda\frac{\widetilde{M}-\widetilde{m}}{2} < \mu$. Let $\boldsymbol{u}_0 \in \boldsymbol{V}$, $\rho_0 \in H^1(\Omega)$, with $0 < m \le \rho_0 \le M$ in Ω, and $\boldsymbol{f} \in L^2(0,T;\boldsymbol{L}^{6/5}(\Omega))$.

(H1) Assume that Ω is an open, bounded set of $\mathbb{R}^3$ whose boundary is polyhedral and such that the continuous dependence in H^2-norm of the Poisson–Neumann problem and in the $\boldsymbol{H}^2 \times H^1$-norm of the Stokes problem holds (see (4.4) and (3.6), respectively). This is verified, for example, if Ω is convex [8].

(H2) The triangulation of Ω and the discrete spaces verify

- the inverse inequalities:

$$|\nabla\bar\rho_h| \le C\,h^{-1}|\bar\rho_h|, \quad \|\nabla\bar\rho_h\|_{L^3(\Omega)} \le C\,h^{-1/2}|\nabla\bar\rho_h| \quad \forall\,\bar\rho_h \in W_h,$$
$$\|\bar\rho_h\|_{L^\infty(\Omega)} \le C\,h^{-1/2}\|\bar\rho_h\|_{H^1(\Omega)} \quad \forall\,\bar\rho_h \in W_h,$$

- and the interpolation errors:

$$\|\bar{\boldsymbol{u}} - \widetilde{J}_h\bar{\boldsymbol{u}}\|_{H^1(\Omega)} + \|\bar{\boldsymbol{u}} - J_h\bar{\boldsymbol{u}}\|_{H^1(\Omega)} \le C\,h|\bar{\boldsymbol{u}}|_{H^2(\Omega)} \quad \forall\,\bar{\boldsymbol{u}} \in \boldsymbol{H}^2(\Omega)\cap\boldsymbol{H}^1_0(\Omega),$$
$$|\bar p - \widetilde{K}_h\bar p| + |\bar p - K_h\bar p| \le C\,h|\bar p|_{H^1(\Omega)} \quad \forall\,\bar p \in H^1(\Omega)\cap L^2_0(\Omega),$$
$$\|\bar\rho - I_h\bar\rho\|_{L^\infty(\Omega)\cap W^{1,3}(\Omega)} \le C\,h^{1/2}|\bar\rho|_{H^2(\Omega)} \quad \forall\,\bar\rho \in H^2(\Omega),$$
$$|\bar\rho - I_h\bar\rho| + h\,\|\bar\rho - I_h\bar\rho\|_{H^1(\Omega)} \le C\,h^2|\bar\rho|_{H^2(\Omega)} \quad \forall\,\bar\rho \in H^2(\Omega),$$

where J_h, $\widetilde{J}_h$, K_h, $\widetilde{K}_h$, and I_h are interpolation operators from $\boldsymbol{H}^2(\Omega)\cap\boldsymbol{H}^1_0(\Omega)$ into $\boldsymbol{V}_h$, $\boldsymbol{H}^2(\Omega)\cap\boldsymbol{H}^1_0(\Omega)$ into $\widetilde{\boldsymbol{V}}_h$, $H^1(\Omega)\cap L^2_0(\Omega)$ into M_h, $H^1(\Omega)\cap L^2_0(\Omega)$ into $\widetilde{M}_h$, and $H^2(\Omega)$ into W_h, respectively. Here and in what follows, we denote by $|v|_{H^k(\Omega)} = \sum_{|\alpha|=k}|D^\alpha v|$ the standard seminorm of higher order derivatives.

(H3) Inf-sup conditions. There exist $\beta > 0$ and $\widetilde{\beta} > 0$ (independent of h) such that, $\forall\,\bar p_h \in M_h$ and $\forall\,\bar q_h \in \widetilde{M}_h$,

$$\|\bar p_h\|_{L^2_0(\Omega)} \le \beta \sup_{\bar{\boldsymbol{u}}_h \in \boldsymbol{V}_h\setminus\{0\}} \frac{\left(\bar p_h, \nabla\cdot\bar{\boldsymbol{u}}_h\right)}{\|\bar{\boldsymbol{u}}_h\|}, \quad \|\bar q_h\|_{L^2_0(\Omega)} \le \widetilde{\beta} \sup_{\bar{\boldsymbol{w}}_h \in \widetilde{\boldsymbol{V}}_h\setminus\{0\}} \frac{\left(\bar q_h, \nabla\cdot\bar{\boldsymbol{w}}_h\right)}{\|\bar{\boldsymbol{w}}_h\|}.$$

(H4) Compatibility condition between $\widetilde{M}_h$ and W_h: $(W_h\cdot W_h)\cap L^2_0(\Omega) \subset \widetilde{M}_h$, i.e.,

$$\forall\,\bar\rho^1_h,\,\bar\rho^2_h \in W_h, \quad \bar\rho^1_h\,\bar\rho^2_h - \frac{1}{|\Omega|}\int_\Omega \bar\rho^1_h(\boldsymbol{x})\,\bar\rho^2_h(\boldsymbol{x})\,d\boldsymbol{x} \in \widetilde{M}_h.$$

(H5) Compatibility condition between $(M_h, \widetilde{M}_h)$:

$$M_h \subset \widetilde{M}_h.$$

For instance, a way of defining the discrete spaces $(W_h, \boldsymbol{V}_h, M_h, \widetilde{\boldsymbol{V}}_h, \widetilde{M}_h)$ verifying (H2)–(H5) is the following: Let $\{\mathcal{T}_h\}_{h>0}$ be a regular, quasi-uniform family of triangulations of Ω, with $h = \max_{K\in\mathcal{T}_h} h_K$ (h_K = diameter of K), and

$$X^l_h = \{x_h \in C^0(\overline{\Omega}) \text{ such that } x_h|_K \in \mathbb{P}_l(K)\ \forall K \in \mathcal{T}_h,\}.$$

Then we define $W_h = X_h^1$. There are several possibilities to define $(\boldsymbol{V}_h, M_h)$ [8], by using the Taylor–Hood element $(\mathbb{P}_2 \times \mathbb{P}_1)$ or the minielement $(\mathbb{P}_1 + \text{bubble} \times \mathbb{P}_1)$, for instance. For the spaces $(\widetilde{\boldsymbol{V}_h}, \widetilde{M_h})$ we choose $\widetilde{\boldsymbol{V}_h} = \boldsymbol{X}_h^3 \cap \boldsymbol{H}_0^1(\Omega)$ and $\widetilde{M_h} = X_h^2 \cap L_0^2(\Omega)$.

Note that if $\boldsymbol{V}_h = \widetilde{\boldsymbol{V}_h}$ and $M_h = \widetilde{M_h}$ are chosen, we need not consider the projection problem (1.7).

Remark 3.1. Hypothesis (H4) implies that $(\nabla \cdot \boldsymbol{w}_h^n, \bar{\rho}_h^1 \bar{\rho}_h^2) = 0$ for all $\bar{\rho}_h^1, \bar{\rho}_h^2 \in W_h$ (this property will play an important role in our analysis). Indeed, we shall write

$$0 = \left(\nabla \cdot \boldsymbol{w}_h^n, \bar{\rho}_h^1 \bar{\rho}_h^2 - \frac{1}{|\Omega|}\int_\Omega \bar{\rho}_h^1 \bar{\rho}_h^2\right) = \left(\nabla \cdot \boldsymbol{w}_h^n, \bar{\rho}_h^1 \bar{\rho}_h^2\right) - \frac{1}{|\Omega|}\int_\Omega \bar{\rho}_h^1 \bar{\rho}_h^2 \int_\Omega \nabla \cdot \boldsymbol{w}_h^n$$
$$= \left(\nabla \cdot \boldsymbol{w}_h^n, \bar{\rho}_h^1 \bar{\rho}_h^2\right).$$

As a consequence, by taking $\bar{\rho}_h = 1$ in (1.8) we have that $\int_\Omega \rho_h^n = \int_\Omega \rho_h^0$, for each n. This property is the discrete version of the continuous one $\int_\Omega \rho(\boldsymbol{x}, t_1)\, d\boldsymbol{x} = \int_\Omega \rho(\boldsymbol{x}, t_2)\, d\boldsymbol{x}$ for any $t_1, t_2 \in [0, T]$, whose physical meaning is the conservation of mass.

3.2. Auxiliary truncated scheme. To prove a priori estimates for scheme (1.7)–(1.10), we will introduce an auxiliary scheme in which some of the densities appearing in the discrete problem of the momentum system are truncated between $\widetilde{m}$ and $\widetilde{M}$ as follows:

Initialization: Let $\boldsymbol{u}_h^0$ and ρ_h^0 be given as in scheme (1.7)–(1.10).

Time step $n+1$: Given $(\rho_h^n, \boldsymbol{u}_h^n, p_h^n) \in W_h \times \boldsymbol{V}_h \times M_h$,

1. find $(\boldsymbol{w}_h^n, q_h^n) \in \widetilde{\boldsymbol{V}_h} \times \widetilde{M_h}$ such that, for each $(\bar{\boldsymbol{w}}_h, \bar{q}_h) \in \widetilde{\boldsymbol{V}_h} \times \widetilde{M_h}$,

$$\left(\nabla \boldsymbol{w}_h^n, \nabla \bar{\boldsymbol{w}}_h\right) - \left(q_h^n, \nabla \cdot \bar{\boldsymbol{w}}_h\right) = \left(\nabla \boldsymbol{u}_h^n, \nabla \bar{\boldsymbol{w}}_h\right), \quad \left(\nabla \cdot \boldsymbol{w}_h^n, \bar{q}_h\right) = 0; \tag{3.1}$$

2. find $\rho_h^{n+1} \in W_h$ such that, for each $\bar{\rho}_h \in W_h$,

$$\left(\frac{\rho_h^{n+1} - \rho_h^n}{k}, \bar{\rho}_h\right) + \left(\boldsymbol{w}_h^n \cdot \nabla \rho_h^{n+1}, \bar{\rho}_h\right) + \lambda\left(\nabla \rho_h^{n+1}, \nabla \bar{\rho}_h\right) = 0; \tag{3.2}$$

3. find $(\boldsymbol{u}_h^{n+1}, p_h^{n+1}) \in \boldsymbol{V}_h \times M_h$ such that, for each $(\bar{\boldsymbol{u}}_h, \bar{p}_h) \in \boldsymbol{V}_h \times M_h$,

$$\left\{\begin{array}{l} \displaystyle\left([\rho_h^n]_T \frac{\boldsymbol{u}_h^{n+1} - \boldsymbol{u}_h^n}{k}, \bar{\boldsymbol{u}}_h\right) + \frac{1}{2}\left(\frac{[\rho_h^{n+1}]_T - [\rho_h^n]_T}{k}\boldsymbol{u}_h^{n+1}, \bar{\boldsymbol{u}}_h\right) \\ \qquad\qquad + a\left([\rho_h^{n+1}]_T, \boldsymbol{u}_h^{n+1}, \bar{\boldsymbol{u}}_h\right) \\ \displaystyle + c\left(\rho_h^{n+1}\boldsymbol{u}_h^n - \lambda \nabla \rho_h^{n+1}, \boldsymbol{u}_h^{n+1}, \bar{\boldsymbol{u}}_h\right) = \left([\rho_h^{n+1}]_T \boldsymbol{f}^{n+1}, \bar{\boldsymbol{u}}_h\right) + \left(p_h^{n+1}, \nabla \cdot \bar{\boldsymbol{u}}_h\right), \end{array}\right. \tag{3.3}$$

$$\left(\nabla \cdot \boldsymbol{u}_h^{n+1}, q_h\right) = 0. \tag{3.4}$$

Here the truncation $[\cdot]_T$ is defined as follows: Given $\rho_h \in W_h$, then

$$[\rho_h]_T(\boldsymbol{x}) = \left\{\begin{array}{ll} \rho_h(\boldsymbol{x}) & \text{if} \quad \rho_h(\boldsymbol{x}) \in [\widetilde{m}, \widetilde{M}], \\ \widetilde{m} & \text{if} \quad \rho_h(\boldsymbol{x}) < \widetilde{m}, \\ \widetilde{M} & \text{if} \quad \rho_h(\boldsymbol{x}) > \widetilde{M}. \end{array}\right.$$

The idea was to truncate in those density terms which required to hold either L^∞ estimates or positivity in order to obtain weak energy estimates (see the proof of Theorem 2.3).

We may again deduce that (3.1), (3.2), and (3.3)–(3.4) are well-posed problems, obtaining a priori estimates.

3.3. Weak estimates for the truncated scheme.

LEMMA 3.2. *The solution of scheme* (3.1)–(3.4) *verifies the following estimates:*

$$\text{(i)}\ \max_{0\le n\le N}|\boldsymbol{u}_h^n|\le C,\quad \text{(ii)}\ k\sum_{n=0}^{N}\|\boldsymbol{u}_h^n\|^2\le C,\quad \text{(iii)}\ \sum_{n=0}^{N-1}|\boldsymbol{u}_h^{n+1}-\boldsymbol{u}_h^n|^2\le C,$$

$$\text{(iv)}\ \max_{0\le n\le N}|\rho_h^n|\le C,\quad \text{(v)}\ \lambda k\sum_{n=0}^{N-1}|\nabla\rho_h^{n+1}|^2\le C,\quad \text{(vi)}\ \sum_{n=0}^{N-1}|\rho_h^{n+1}-\rho_h^n|^2\le C,$$

where $C>0$ depends on the data $(\rho_0,\boldsymbol{u}_0,\boldsymbol{f})$ but is independent of k,h, and λ.

Proof. To obtain a priori estimates for the velocity $(\boldsymbol{u}_h^n)$, we take $\bar{\boldsymbol{u}}_h=2k\boldsymbol{u}_h^{n+1}$ and $\bar{p}_h=p_h^{n+1}$ as test functions in (3.3)–(3.4), resulting in [11]:

$$\begin{aligned}&|\sqrt{[\rho_h^{n+1}]_T}\boldsymbol{u}_h^{n+1}|^2-|\sqrt{[\rho_h^n]_T}\boldsymbol{u}_h^n|^2+|\sqrt{[\rho_h^n]_T}(\boldsymbol{u}_h^{n+1}-\boldsymbol{u}_h^n)|^2+\mu_1k\|\boldsymbol{u}_h^{n+1}\|^2\\&\le 2k\Big([\rho_h^{n+1}]_T\boldsymbol{f}^{n+1},\boldsymbol{u}_h^{n+1}\Big)\le 2k\|[\rho_h^{n+1}]_T\|_{L^\infty(\Omega)}\|\boldsymbol{f}^{n+1}\|_{L^{6/5}(\Omega)}\|\boldsymbol{u}_h^{n+1}\|_{L^6(\Omega)}\\&\le\frac{\mu_1k}{2}\|\boldsymbol{u}_h^{n+1}\|^2+C\,k\|\boldsymbol{f}^{n+1}\|^2_{L^{6/5}(\Omega)}.\end{aligned}$$

Consequently,

$$\begin{aligned}&|\sqrt{[\rho_h^{n+1}]_T}\boldsymbol{u}_h^{n+1}|^2-|\sqrt{[\rho_h^n]_T}\boldsymbol{u}_h^n|^2+|\sqrt{[\rho_h^n]_T}(\boldsymbol{u}_h^{n+1}-\boldsymbol{u}_h^n)|^2\\&\quad+\frac{\mu_1}{2}k\|\boldsymbol{u}_h^{n+1}\|^2\le C\,k\|\boldsymbol{f}^{n+1}\|^2_{L^{6/5}(\Omega)}.\end{aligned}\tag{3.5}$$

By adding (3.5) for $n=0,\dots,r$ with any $r<N$, estimates (i), (ii), and (iii) hold.

On the other hand, to obtain weak energy estimates for the density (ρ_h^n), we take $\bar{\rho}_h=2\,k\,\rho_h^{n+1}$ in (3.2) and use the fact that $(\nabla\cdot\boldsymbol{w}_h^n,(\rho_h^{n+1})^2)=0$ (see Remark 3.1):

$$|\rho_h^{n+1}|^2-|\rho_h^n|^2+|\rho_h^{n+1}-\rho_h^n|^2+2\,\lambda\,k|\nabla\rho_h^{n+1}|^2=0.$$

By adding over n, one deduces estimates (iv), (v), and (vi). ∎

COROLLARY 3.3. *The following estimates hold:*

$$\text{(vii)}\ \max_{0\le n\le N}|\boldsymbol{w}_h^n|\le C,\quad \text{(viii)}\ k\sum_{n=0}^{N}\|\boldsymbol{w}_h^n\|^2\le C,$$

where $C>0$ is independent of k, h, and λ.

Proof. By taking $\bar{\boldsymbol{w}}_h=\boldsymbol{w}_h^n$ in (1.7) and using (H5), $(\nabla(\boldsymbol{w}_h^n-\boldsymbol{u}_h^n),\nabla\boldsymbol{w}_h^n)=0$, and hence one has $|\nabla\boldsymbol{w}_h^n|\le|\nabla\boldsymbol{u}_h^n|$. So from (ii) we get (viii). Now we are going to get estimate (vii) by using a duality technique and the constraint (S). Indeed, let $(\boldsymbol{z},\xi)\in(\boldsymbol{V}\cap\boldsymbol{H}^2(\Omega))\times(L^2_0(\Omega)\cap H^1(\Omega))$ be the strong solution of the Stokes problem

$$-\Delta\boldsymbol{z}+\nabla\xi=\boldsymbol{w}_h^n-\boldsymbol{u}_h^n,\quad\nabla\cdot\boldsymbol{z}=0\text{ in }\Omega,\quad\boldsymbol{z}=0\text{ on }\Gamma.\tag{3.6}$$

By taking $\boldsymbol{w}_h^n-\boldsymbol{u}_h^n$ as a test function in the variational formulation of (3.6), we get

$$|\boldsymbol{w}_h^n-\boldsymbol{u}_h^n|^2=\Big(\nabla\boldsymbol{z},\nabla(\boldsymbol{w}_h^n-\boldsymbol{u}_h^n)\Big)+\Big(\xi,\nabla\cdot(\boldsymbol{w}_h^n-\boldsymbol{u}_h^n)\Big).\tag{3.7}$$

Let $(\boldsymbol{z}_h, \xi_h) \in \widetilde{V_h} \times \widetilde{M_h}$ be the discrete approximation of (3.6) defined as

$$\begin{cases} \Big(\nabla \boldsymbol{z}_h, \nabla \bar{\boldsymbol{w}}_h\Big) - \Big(\xi_h, \nabla \cdot \bar{\boldsymbol{w}}_h\Big) = \Big(\boldsymbol{w}_h^n - \boldsymbol{u}_h^n, \bar{\boldsymbol{w}}_h\Big) \quad \forall \bar{\boldsymbol{w}}_h \in \widetilde{\boldsymbol{V}_h}, \\ \Big(\nabla \cdot \boldsymbol{z}_h, \bar{q}_h\Big) = 0 \quad \forall \bar{q}_h \in \widetilde{M_h}. \end{cases}$$

In view of hypothesis (H5), $(K_h\xi, \nabla \cdot (\boldsymbol{w}_h^n - \boldsymbol{u}_h^n)) = 0$, and hence we write (3.7) as follows:

$$|\boldsymbol{w}_h^n - \boldsymbol{u}_h^n|^2 = \Big(\nabla \boldsymbol{z} - \nabla \boldsymbol{z}_h, \nabla(\boldsymbol{w}_h^n - \boldsymbol{u}_h^n)\Big) + \Big(\nabla \boldsymbol{z}_h, \nabla(\boldsymbol{w}_h^n - \boldsymbol{u}_h^n)\Big) + \Big(\xi - K_h\xi, \nabla \cdot (\boldsymbol{w}_h^n - \boldsymbol{u}_h^n)\Big),$$

where K_h is the interpolation operator defined in hypothesis (H2). From (1.7), it follows that $(\nabla \boldsymbol{z}_h, \nabla(\boldsymbol{w}_h^n - \boldsymbol{u}_h^n)) = 0$. Thus, we find

$$\begin{aligned} &|\boldsymbol{w}_h^n - \boldsymbol{u}_h^n|^2 \le |\nabla \boldsymbol{z} - \nabla \boldsymbol{z}_h||\nabla(\boldsymbol{w}_h^n - \boldsymbol{u}_h^n)| + |\xi - K_h\xi||\nabla \cdot (\boldsymbol{w}_h^n - \boldsymbol{u}_h^n)| \\ &\le Ch(\|\boldsymbol{z}\|_{H^2(\Omega)} + \|\xi\|_{H^1(\Omega)})|\nabla(\boldsymbol{w}_h^n - \boldsymbol{u}_h^n)| \le Ch|\boldsymbol{w}_h^n - \boldsymbol{u}_h^n||\nabla(\boldsymbol{w}_h^n - \boldsymbol{u}_h^n)|, \end{aligned}$$

where in the second line we have used the approximation property (see [8]) $|\nabla \boldsymbol{z} - \nabla \boldsymbol{z}_h| \le C\, h(\|\boldsymbol{z}\|_{H^2} + \|\xi\|_{H^1})$, the interpolation error $|\xi - K_h\xi| \le C\, h\|\xi\|_{H^1(\Omega)}$ assumed in (H2), and the $H^2 \times H^1$ continuous dependency of the Stokes problem (3.6) $\|\boldsymbol{z}\|_{H^2(\Omega)} + \|\xi\|_{H^1(\Omega)} \le C\, |\boldsymbol{w}_h^n - \boldsymbol{u}_h^n|$ assumed in (H1).

Therefore, we have

$$|\boldsymbol{w}_h^n - \boldsymbol{u}_h^n| \le C\, h|\nabla(\boldsymbol{w}_h^n - \boldsymbol{u}_h^n)|. \tag{3.8}$$

Now, in view of (S), we may get $h \le Ck$ for (h, k) small enough. Then, since $k^{1/2}|\nabla \boldsymbol{w}_h^n| \le k^{1/2}|\nabla \boldsymbol{u}_h^n| \le C$ (thanks to the estimate $|\nabla \boldsymbol{w}_h^n| \le |\nabla \boldsymbol{u}_h^n|$ and estimate (ii) of Lemma 3.2), it is easy to see that $|\boldsymbol{w}_h^n| \le Ck^{1/2} + |\boldsymbol{u}_h^n| \le C$, and we get estimate (vii). ∎

3.4. Discrete maximum principle (of the truncated scheme). In this subsection we prove that the discrete density of scheme (3.2) has pointwise estimates by excess and defect with respect to the upper and lower bounds of the initial density ρ_0, respectively. Namely, we will see that $\widetilde{m} \le \rho_h^n \le \widetilde{M}$ in Ω for all k and h small enough satisfying constraint (S).

3.4.1. Study of an auxiliary time discrete scheme. We define a sequence (ρ^n) associated to $(\boldsymbol{w}_h^n)$ by means of the following time discrete scheme:

Initialization: Let $\rho^0 = \rho_0$.

Time step $n+1$: Given ρ^n, we compute $\rho^{n+1} \in H^2(\Omega)$, verifying

$$\frac{\rho^{n+1} - \rho^n}{k} + \boldsymbol{w}_h^n \cdot \nabla \rho^{n+1} - \lambda \Delta \rho^{n+1} = 0 \text{ in } \Omega, \quad \frac{\partial \rho^{n+1}}{\partial \boldsymbol{n}} = 0 \text{ on } \Gamma. \tag{3.9}$$

LEMMA 3.4. *Let* $\{\boldsymbol{w}_h^n\}_{n=0}^N \subset \boldsymbol{H}_0^1(\Omega)$ *such that* $k\sum_{n=0}^N \|\boldsymbol{w}_h^n\|^2 \le C$. *Then there exists a unique solution* $\rho^{n+1} \in H^2(\Omega)$ *of* (3.9), *which also verifies:*

$$\begin{aligned} &0 < m \le \rho^{n+1}(\boldsymbol{x}) \le M \quad \forall \boldsymbol{x} \in \Omega, \quad \forall\, n = 0, \dots, N-1, \\ &\lambda^2\, k \sum_{n=0}^{N-1} \|\rho^{n+1}\|_{H^2(\Omega)}^2 \le C, \end{aligned} \tag{3.10}$$

where $C > 0$ *is a constant independent of* k, h, *and* λ.

Proof. The proof of this lemma can be found in Appendix A. ∎

3.4.2. Error estimates between $\boldsymbol{\rho^{n+1}}$ and $\boldsymbol{\rho_h^{n+1}}$. Denote by $e_\rho^{n+1} = \rho^{n+1} - \rho_h^{n+1}$ the difference between the solutions of problems (3.9) and (3.2). Our intention now is to state the following error estimate:

$$\|e_\rho^{n+1}\|_{H^1(\Omega)} \leq C\left(\frac{h^2}{\lambda\sqrt{\lambda}k} + \frac{h\sqrt{h}}{\lambda^2 k} + \frac{hh^{1/4}}{\lambda^{3/2}k^{3/4}} + \frac{h}{\lambda\sqrt{k}}\right). \tag{3.11}$$

Indeed, by subtracting (3.9) multiplied by $\bar\rho_h \in W_h$ and (3.2), one has

$$\left(\frac{e_\rho^{n+1}-e_\rho^n}{k}, \bar\rho_h\right) + \Big(\boldsymbol{w}_h^n\cdot\nabla e_\rho^{n+1}, \bar\rho_h\Big) + \lambda\Big(\nabla e_\rho^{n+1}, \nabla\bar\rho_h\Big) = 0$$

for each $\bar\rho_h \in W_h$. By decomposing the convective term as

$$\Big(\boldsymbol{w}_h^n\cdot\nabla e_\rho^{n+1}, \bar\rho_h\Big) = \Big(\boldsymbol{w}_h^n\cdot\nabla(\rho^{n+1} - I_h\rho^{n+1}), \bar\rho_h\Big) + \Big(\boldsymbol{w}_h^n\cdot\nabla(I_h\rho^{n+1} - \rho_h^{n+1}), \bar\rho_h\Big),$$

taking $\bar\rho_h = 2k(e_\rho^{n+1} - \rho^{n+1} + I_h\rho^{n+1}) = 2\,k(I_h\rho^{n+1} - \rho_h^{n+1})$, with $I_h\rho^{n+1} \in W_h$, and using the fact that $(\nabla\cdot\boldsymbol{w}_h^n, \bar\rho_h^2) = 0$ for all $\bar\rho_h \in W_h$ (see Remark 3.1), we get

$$\begin{aligned}
&|e_\rho^{n+1}|^2 - |e_\rho^n|^2 + |e_\rho^{n+1} - e_\rho^n|^2 + 2\lambda k|\nabla e_\rho^{n+1}|^2\\
&\leq 2\Big(e_\rho^{n+1} - e_\rho^n, \rho^{n+1} - I_h\rho^{n+1}\Big) - 2k\Big(\boldsymbol{w}_h^n\cdot\nabla(\rho^{n+1} - I_h\rho^{n+1}), e_\rho^{n+1}\Big)\\
&-k\Big(\nabla\cdot\boldsymbol{w}_h^n, (\rho^{n+1} - I_h\rho^{n+1})^2\Big) + 2\lambda\,k\Big(\nabla e_\rho^{n+1}, \nabla(\rho^{n+1} - I_h\rho^{n+1})\Big).
\end{aligned}$$

Next, by integrating by parts the second term on the right-hand side and bounding adequately, we infer that

$$\begin{aligned}
&|e_\rho^{n+1}|^2 - |e_\rho^n|^2 + |e_\rho^{n+1} - e_\rho^n|^2 + 2\lambda k|\nabla e_\rho^{n+1}|^2\\
&\leq 2|e_\rho^{n+1} - e_\rho^n|\,|\rho^{n+1} - I_h\rho^{n+1}| + Ck\|\boldsymbol{w}_h^n\|\,\|e_\rho^{n+1}\|_{H^1(\Omega)}\,\|\rho^{n+1} - I_h\rho^{n+1}\|_{L^3(\Omega)}\\
&+k\|\boldsymbol{w}_h^n\|\|\rho^{n+1} - I_h\rho^{n+1}\|^2_{L^4(\Omega)} + 2\lambda\,k|\nabla e_\rho^{n+1}|\,|\nabla(\rho^{n+1} - I_h\rho^{n+1})|\\
&\leq \frac{1}{2}|e_\rho^{n+1} - e_\rho^n|^2 + C|\rho^{n+1} - I_h\rho^{n+1}|^2 + C\lambda k|\nabla(\rho^{n+1} - I_h\rho^{n+1})|^2\\
&+\frac{C}{\lambda}k\|\boldsymbol{w}_h^n\|^2\|\rho^{n+1} - I_h\rho^{n+1}\|^2_{L^3} + k\|\boldsymbol{w}_h^n\|\|\rho^{n+1} - I_h\rho^{n+1}\|^2_{L^4} + \lambda k|e_\rho^{n+1}|^2 + \lambda k|\nabla e_\rho^{n+1}|^2,
\end{aligned}$$

and then, by taking into account the interpolation errors

$$|\rho^{n+1} - I_h\rho^{n+1}| + h|\nabla(\rho^{n+1} - I_h\rho^{n+1})| \leq C\,h^2\|\rho^{n+1}\|_{H^2},$$

and

$$\|\rho^{n+1} - I_h\rho^{n+1}\|_{L^3} \leq C\,h^{3/2}\|\rho^{n+1}\|_{H^2}, \qquad \|\rho^{n+1} - I_h\rho^{n+1}\|_{L^4} \leq C\,h^{5/4}\|\rho^{n+1}\|_{H^2}$$

(the last two are a consequence of the previous one and the 3D interpolation inequalities $\|\rho\|_{L^3} \leq C|\rho|^{1/2}\|\rho\|^{1/2}_{H^1}$ and $\|\rho\|_{L^4} \leq C|\rho|^{1/4}\|\rho\|^{3/4}_{H^1}$) and the estimate $k\|\boldsymbol{w}_h^n\|^2 \leq C$ thanks to estimate (vii) of Corollary 3.3, one arrives at

$$\begin{aligned}
&|e_\rho^{n+1}|^2 - |e_\rho^n|^2 + \lambda\,k\,|\nabla e_\rho^{n+1}|^2\\
&\qquad\leq C\left(h^4 + \frac{h^3}{\lambda} + k^{1/2}h^{5/2} + \lambda\,k\,h^2\right)\|\rho^{n+1}\|^2_{H^2} + \lambda\,k|e_\rho^{n+1}|^2.
\end{aligned}$$

By adding up for $n = 0, \dots, l$ and using the fact that $\lambda^2\, k \sum_{n=0}^{l} \|\rho^{n+1}\|_{H^2}^2 \le C$ and by virtue of the generalized discrete Gronwall lemma, we infer that for all (k, λ), with $\lambda\, k < 1$ (for instance, $\lambda\, k \le 1/2$), there exists $C > 0$ independent of λ such that

$$|e_\rho^{l+1}|^2 + \lambda\, k \sum_{n=0}^{l} |\nabla e_\rho^{n+1}|^2 \le C \left(\frac{h^4}{\lambda^2 k} + \frac{h^3}{\lambda^3 k} + \frac{h^{5/2}}{\lambda^2 k^{1/2}} + \frac{h^2}{\lambda} + |e_\rho^0|^2 \right).$$

By taking $|e_\rho^0|^2 = |\rho^0 - \rho_h^0|^2 \le C\, h^2$, we deduce the bound

$$|e_\rho^{n+1}|^2 + \lambda k \sum_{n=0}^{l} |\nabla e_\rho^{n+1}|^2 \le C \left(\frac{h^4}{\lambda^2 k} + \frac{h^3}{\lambda^3 k} + \frac{h^{5/2}}{\lambda^2 k^{1/2}} + \frac{h^2}{\lambda} \right),$$

whence, in particular, (3.11) holds.

3.4.3. Pointwise estimates of the truncated scheme. Here we will prove the following pointwise estimates [6]:

$$\begin{aligned} \rho_h^{n+1} &\ge m - C\, h^{-1/2} \left(\frac{h^2}{\lambda^{3/2} k} + \frac{h\sqrt{h}}{\lambda^2 k} + \frac{h h^{1/4}}{\lambda^2 k^{3/4}} + \frac{h}{\lambda\sqrt{k}} \right), \\ \rho_h^{n+1} &\le M + C\, h^{-1/2} \left(\frac{h^2}{\lambda^{3/2} k} + \frac{h\sqrt{h}}{\lambda^2 k} + \frac{h h^{1/4}}{\lambda^2 k^{3/4}} + \frac{h}{\lambda\sqrt{k}} \right). \end{aligned} \tag{3.12}$$

To prove (3.12) it suffices to prove that

$$\|\rho_h^{n+1} - \rho^{n+1}\|_{L^\infty(\Omega)} \le C\, h^{-1/2} \left(\frac{h^2}{\lambda^{3/2} k} + \frac{h\sqrt{h}}{\lambda^2 k} + \frac{h h^{1/4}}{\lambda^2 k^{3/4}} + \frac{h}{\lambda\sqrt{k}} \right).$$

For this, from the triangle inequality

$$\begin{aligned} \|\rho_h^{n+1} - \rho^{n+1}\|_{L^\infty(\Omega)} &\le \|\rho_h^{n+1} - I_h \rho^{n+1}\|_{L^\infty(\Omega)} + \|I_h \rho^{n+1} - \rho^{n+1}\|_{L^\infty(\Omega)} \\ &\le \|\rho_h^{n+1} - I_h \rho^{n+1}\|_{L^\infty(\Omega)} + C\, \frac{h^{1/2}}{\lambda\, k^{1/2}}, \end{aligned}$$

where in the last line we have used the approximation inequality $\|\rho^{n+1} - I_h \rho^{n+1}\|_{L^\infty} \le C\, h^{1/2} \|\rho^{n+1}\|_{H^2}$ and the estimate $\|\rho^{n+1}\|_{H^2} \le \frac{C}{\lambda\, k^{1/2}}$ (see (3.10)). Hence, it suffices to obtain the inequality

$$\|\rho_h^{n+1} - I_h \rho^{n+1}\|_{H^1(\Omega)} \le C \left(\frac{h^2}{\lambda^{3/2} k} + \frac{h\sqrt{h}}{\lambda^2 k} + \frac{h h^{1/4}}{\lambda^2 k^{3/4}} + \frac{h}{\lambda\sqrt{k}} \right) \tag{3.13}$$

and to use the inverse inequality (see [3]) $\|\bar\rho_h\|_{L^\infty(\Omega)} \le C\, h^{-1/2}\, \|\bar\rho_h\|_{H^1(\Omega)}\ \forall \bar\rho_h \in W_h$. Let us prove (3.13). From the triangle inequality,

$$\|\rho_h^{n+1} - I_h \rho^{n+1}\|_{H^1(\Omega)} \le \|\rho_h^{n+1} - \rho^{n+1}\|_{H^1(\Omega)} + \|\rho^{n+1} - I_h \rho^{n+1}\|_{H^1(\Omega)},$$

and, by using the error estimate (3.11) and the interpolation error $\|w - I_h w\|_{H^1(\Omega)} \le h \|w\|_{H^2(\Omega)}$ for $w = \rho^{n+1}$ jointly with the estimate $h\|\rho^{n+1}\|_{H^2} \le C\, h/\lambda\sqrt{k}$ (thanks to (3.10)), one easily deduces (3.13).

Now, by taking into account hypothesis (S), one has

$$\lim_{(h,k)\to 0}\left(\frac{h^2}{k}+\frac{h\sqrt{h}}{k}+\frac{hh^{1/4}}{k^{3/4}}\right)\Big/\left(\frac{h}{\sqrt{k}}\right)=0,$$

and consequently, for each (h,k) small enough,

$$C\,h^{-1/2}\left(\frac{h^2}{\lambda^{3/2}k}+\frac{h\sqrt{h}}{\lambda^2 k}+\frac{h}{\lambda\sqrt{k}}\right)\le C\,h^{-1/2}\frac{h}{\sqrt{k}}=C\sqrt{\frac{h}{k}}.$$

In particular, thanks to (3.12), by imposing $h\le h_0$, $k\le k_0$ such that

$$C\sqrt{\frac{h}{k}}\le \min\{m-\widetilde{m},\widetilde{M}-M\}, \tag{3.14}$$

one gets $0<\widetilde{m}\le \rho_h^{n+1}\le \widetilde{M}$.

3.4.4. Identification between the truncated and nontruncated schemes. Now it is clear that if ρ_h^{n+1} is the solution of the truncated scheme, then $[\rho_h^{n+1}]_T=\rho_h^{n+1}$, and consequently the truncated scheme and the nontruncated scheme coincide, arriving at the following result.

THEOREM 3.5. *Assume that $h\le h_0$, $k\le k_0$ satisfying* (3.14) *and $\lambda\,k\le 1/2$; then scheme* (1.7)–(1.10) *is well-posed and verifies estimates* (i)–(vi) *of Lemma* 3.2, (vii)–(viii) *of Corollary* 3.3, *and $0<\widetilde{m}\le\rho_h^{n+1}\le\widetilde{M}$ in Ω.*

4. Strong estimates for the density. Let $-\Delta_h : W_h\to W_h$ be the linear operator defined as follows:

$$-\Big(\Delta_h\rho_h,\bar\rho_h\Big)=\Big(\nabla\rho_h,\nabla\bar\rho_h\Big)\quad\forall\,\bar\rho_h\in W_h. \tag{4.1}$$

Then the discrete density equation (1.8) can be rewritten as

$$\left(\frac{\rho_h^{n+1}-\rho_h^n}{k},\bar\rho_h\right)+\Big(\boldsymbol{w}_h^n\cdot\nabla\rho_h^{n+1},\bar\rho_h\Big)-\lambda\Big(\Delta_h\rho_h^{n+1},\bar\rho_h\Big)=0. \tag{4.2}$$

THEOREM 4.1. *Under the hypotheses of Theorem* 3.5, *the solution ρ_h^{n+1} of scheme* (1.8) *verifies the following estimates, for each h and k small enough:*

$$\text{(ix) } \lambda\max_{0\le n\le N}|\nabla\rho_h^n|^2\le C,\qquad \text{(x) } \lambda^2\,k\sum_{n=0}^{N}|\Delta_h\rho_h^{n+1}|^2\le C,$$
$$\text{(xi) } \lambda\sum_{n=0}^{N-1}|\nabla(\rho_h^{n+1}-\rho_h^n)|^2\le C,$$

where $C>0$ is independent of h, k, and λ.

Proof. By taking $\bar\rho_h=-2k\Delta_h\rho_h^{n+1}$ in (4.2), we arrive at:

$$|\nabla\rho_h^{n+1}|^2-|\nabla\rho_h^n|^2+|\nabla(\rho_h^{n+1}-\rho_h^n)|^2+2\lambda k|\Delta_h\rho_h^{n+1}|^2=2k\Big(\boldsymbol{w}_h^n\cdot\nabla\rho_h^{n+1},\Delta_h\rho_h^{n+1}\Big):=I. \tag{4.3}$$

To bound I, we use an idea given in [11], where a regular function associated to the discrete Laplacian function $-\Delta_h\rho_h^{n+1}$ is considered. But here, the type of estimates

used in [11] must be changed, making use of the pointwise estimates of ρ_h^{n+1}. We define $\rho(h) \in H^2(\Omega)$ as the solution of the problem:

$$-\Delta\rho(h) = -\Delta_h \rho_h^{n+1} \ \text{ in } \Omega, \quad \frac{\partial \rho(h)}{\partial \boldsymbol{n}}\Big|_{\partial\Omega} = 0, \quad \int_\Omega \rho(h) = \int_\Omega \rho_h^0. \tag{4.4}$$

From the H^2-regularity of the previous problem $\|\rho(h) - \frac{1}{|\Omega|}\int \rho_h^0\|_{H^2(\Omega)} \le C\,|\Delta_h \rho_h^{n+1}|$, and hence one has in particular

$$|\rho(h)|_{H^2(\Omega)} \le C|\Delta_h \rho_h^{n+1}|. \tag{4.5}$$

We write I as $I = 2k(\boldsymbol{w}_h^n \cdot \nabla\rho(h), \Delta\rho(h)) + 2k(\boldsymbol{w}_h^n \cdot \nabla(\rho_h^{n+1} - \rho(h)), \Delta_h\rho_h^{n+1})$. By integrating by parts the first term on the right-hand side, and using (2.2),

$$\begin{aligned} 2k\Big(\boldsymbol{w}_h^n \cdot \nabla\rho(h), \Delta\rho(h)\Big) &= -2k\Big(\nabla \boldsymbol{w}_h^n, \nabla\rho(h) \otimes \nabla\rho(h)\Big) - 2k\Big((\boldsymbol{w}_h^n \cdot \nabla)\nabla\rho(h), \nabla\rho(h)\Big) \\ &= -2k\Big(\nabla \boldsymbol{w}_h^n, \nabla\rho(h) \otimes \nabla\rho(h)\Big) + k\Big(\nabla \cdot \boldsymbol{w}_h^n, |\nabla\rho(h)|^2\Big) \\ &\le C\,k\|\boldsymbol{w}_h^n\|\|\nabla\rho(h)\|_{L^4(\Omega)}^2 \le C\,k\|\boldsymbol{w}_h^n\|\,\|\rho(h)\|_{L^\infty(\Omega)}|\Delta\rho(h)| \\ &\le C\,k\|\boldsymbol{w}_h^n\|\,\|\rho(h)\|_{L^\infty(\Omega)}|\Delta_h\rho_h^{n+1}|, \end{aligned}$$

where $\boldsymbol{a} \otimes \boldsymbol{b}$ denotes the tensorial product matrix of two vectors $\boldsymbol{a} = (a_i)_{i=1}^2, \boldsymbol{b} = (b_i)_{i=1}^2$, with coefficients $(\boldsymbol{a} \otimes \boldsymbol{b})_{i,j} = a_i b_j$. Accordingly,

$$I \le C\,k\|\boldsymbol{w}_h^n\|\Big(\|\rho(h)\|_{L^\infty(\Omega)} + \|\nabla(\rho_h^{n+1} - \rho(h))\|_{L^3(\Omega)}\Big)|\Delta_h\rho_h^{n+1}|. \tag{4.6}$$

Now we will prove the inequality:

$$\|\nabla(\rho_h^{n+1} - \rho(h))\|_{L^3(\Omega)} \le C\,h^{1/2}|\rho(h)|_{H^2(\Omega)}. \tag{4.7}$$

For this, we write

$$\|\nabla(\rho_h^{n+1} - \rho(h))\|_{L^3} \le \|\nabla(\rho_h^{n+1} - I_h\rho(h))\|_{L^3} + \|\nabla(I_h\rho(h) - \rho(h))\|_{L^3}. \tag{4.8}$$

By multiplying (4.4) by $\bar\rho_h \in W_h$ and subtracting to (4.1), one gets

$$\Big(\nabla\rho_h^{n+1} - \nabla\rho(h), \nabla\bar\rho_h\Big) = 0 \quad \forall\,\bar\rho_h \in W_h.$$

By adding and subtracting $\nabla I_h\rho(h)$, and considering $\bar\rho_h = \rho_h^{n+1} - I_h\rho(h) \in W_h$, we obtain

$$\begin{aligned} |\nabla\rho_h^{n+1} - \nabla I_h\rho(h)|^2 &= -\Big(\nabla I_h\rho(h) - \nabla\rho(h), \nabla\rho_h^{n+1} - \nabla I_h\rho(h)\Big) \\ &\le |\nabla I_h\rho(h) - \nabla\rho(h)|\,|\nabla\rho_h^{n+1} - \nabla I_h\rho(h)|, \end{aligned}$$

whence

$$|\nabla\rho_h^{n+1} - \nabla I_h\rho(h)| \le |\nabla I_h\rho(h) - \nabla\rho(h)| \le C\,h|\rho(h)|_{H^2(\Omega)} \tag{4.9}$$

$$|\nabla\rho_h^{n+1} - \nabla\rho(h)| \le C\,h|\rho(h)|_{H^2(\Omega)}. \tag{4.10}$$

Thus, by using the inverse inequality [7]

$$\|\nabla\rho_h^{n+1} - \nabla I_h\rho(h)\|_{L^3(\Omega)} \le Ch^{-1/2}|\nabla\rho_h^{n+1} - \nabla I_h\rho(h)|$$

and (4.9), we arrive at

$$\|\nabla\rho_h^{n+1} - \nabla I_h\rho(h)\|_{L^3(\Omega)} \le C\,h^{1/2}|\rho(h)|_{H^2(\Omega)}. \tag{4.11}$$

So, from (4.8) and (4.11) one gets (4.7) by taking into account the interpolation error $\|\nabla(\rho(h) - I_h\rho(h))\|_{L^3(\Omega)} \le Ch^{1/2}|\rho(h)|_{H^2(\Omega)}$.

By getting back to (4.6) and using (4.7), we bound

$$I \le C\,k\|\boldsymbol{w}_h^n\|\Big(\|\rho(h) - \rho_h^{n+1}\|_{L^\infty} + \|\rho_h^{n+1}\|_{L^\infty} + h^{1/2}|\rho(h)|_{H^2}\Big)|\Delta_h\rho_h^{n+1}|. \tag{4.12}$$

Now we write

$$\|\rho_h^{n+1} - \rho(h)\|_{L^\infty(\Omega)} \le \|\rho_h^{n+1} - I_h\rho(h)\|_{L^\infty(\Omega)} + \|I_h\rho(h) - \rho(h)\|_{L^\infty(\Omega)}.$$

By using the interpolation error $\|\rho(h)-I_h\rho(h)\|_{L^\infty(\Omega)} \le Ch^{1/2}|\rho(h)|_{H^2(\Omega)}$, the inverse inequality in 3D

$$\begin{aligned}\|\rho_h^{n+1} - I_h\rho(h)\|_{L^\infty(\Omega)} &\le C\,h^{-1/2}\|\rho_h^{n+1} - I_h\rho(h)\|_{H^1(\Omega)}\\ &\le C\,h^{-1/2}(\|\rho_h^{n+1} - \rho(h)\|_{H^1(\Omega)} + \|\rho(h) - I_h\rho(h)\|_{H^1(\Omega)})\\ &\le C\,h^{-1/2}(|\nabla(\rho_h^{n+1} - \rho(h))| + h\,|\rho(h)|_{H^2(\Omega)}),\end{aligned}$$

where the generalized Poincare inequality has been used in the last line, since $\int_\Omega \rho_h^{n+1} = \int_\Omega \rho_h^0$ (see Remark 3.1) and $\int_\Omega \rho_h^0 = \int_\Omega \rho(h)$ (see (4.4)). By using (4.10) and (4.5),

$$\|\rho_h^{n+1} - \rho(h)\|_{L^\infty(\Omega)} \le C\,h^{1/2}|\rho(h)|_{H^2(\Omega)} \le C\,h^{1/2}|\Delta_h\rho_h^{n+1}|.$$

By applying the above estimate in (4.12), we bound

$$\begin{aligned}I &\le C\,k\|\boldsymbol{w}_h^n\|\Big(h^{1/2}|\Delta_h\rho_h^{n+1}| + \widetilde{M}\Big)|\Delta_h\rho_h^{n+1}|\\ &\le C\,k\|\boldsymbol{w}_h^n\|h^{1/2}|\Delta_h\rho_h^{n+1}|^2 + \frac{C}{\lambda}k\|\boldsymbol{w}_h^n\|^2 + \frac{\lambda}{2}k|\Delta_h\rho_h^{n+1}|^2.\end{aligned}$$

By Corollary 3.3 we infer the bound $\|\boldsymbol{w}_h^n\| \le C/k^{1/2}$ (with C independent of λ, h, k), and, by choosing h and k small enough such that

$$C\sqrt{\frac{h}{k}} \le \frac{1}{2}\lambda,$$

we get

$$I \le \frac{C}{\lambda}k\|\boldsymbol{w}_h^n\|^2 + \lambda\,k|\Delta_h\rho_h^{n+1}|^2.$$

Therefore, from (4.3) we get the inequality

$$|\nabla\rho_h^{n+1}|^2 - |\nabla\rho_h^n|^2 + |\nabla(\rho_h^{n+1} - \rho_h^n)|^2 + \lambda k|\Delta_h\rho_h^{n+1}|^2 \le C\,\frac{k}{\lambda}\|\boldsymbol{w}_h^n\|^2. \tag{4.13}$$

By adding (4.13) for $n = 0, \dots, r$, with $r < N$, we arrive at

$$\lambda|\nabla\rho_h^{r+1}|^2 + \lambda\sum_{n=0}^{r}|\nabla(\rho_h^{n+1} - \rho_h^n)|^2 + \lambda^2\,k\sum_{n=0}^{r}|\Delta_h\rho_h^{n+1}|^2 \le C\,k\sum_{n=0}^{r}\|\boldsymbol{w}_h^n\|^2 + \lambda|\nabla\rho_h^0|^2.$$

Finally, from estimate (ii) of Lemma 3.2, one gets the desired estimates (ix)–(xi). ∎

COROLLARY 4.2. *The following inequality holds:*

$$\|\nabla\rho_h^{n+1}\|_{L^3(\Omega)} \le C\,|\nabla\rho_h^{n+1}|^{1/2}|\Delta_h\rho_h^{n+1}|^{1/2}. \tag{4.14}$$

Consequently, under the hypotheses of Theorem 4.1*, one has the estimate*

$$\lambda^3 k\sum_{n=0}^{N-1}\|\rho_h^{n+1}\|_{W^{1,3}(\Omega)}^4 \le C,$$

where $C>0$ is independent of h,k, and λ.

Proof. Thanks to estimates (ix) for (ρ_h^n) and (x) for $(\Delta_h\rho_h^n)$, it suffices to prove (4.14). For this, by considering $\rho(h)$ the solution of problem (4.4), we have

$$\|\nabla\rho_h^{n+1}\|_{L^3(\Omega)} \le \|\nabla(\rho_h^{n+1}-\rho(h))\|_{L^3(\Omega)} + \|\nabla\rho(h)\|_{L^3(\Omega)}.$$

By using inequality (4.7) and the interpolation inequality

$$\|\nabla\rho(h)\|_{L^3(\Omega)} \le C|\nabla\rho(h)|^{1/2}\|\nabla\rho(h)\|_{H^1(\Omega)}^{1/2} \le C|\nabla\rho(h)|^{1/2}|\rho(h)|_{H^2(\Omega)}^{1/2},$$

we arrive at

$$\|\nabla\rho_h^{n+1}\|_{L^3(\Omega)} \le C\,h^{1/2}|\rho(h)|_{H^2(\Omega)} + C\,|\nabla\rho(h)|^{1/2}|\rho(h)|_{H^2(\Omega)}^{1/2}.$$

Next, we bound the term $|\nabla\rho(h)|$ by using (4.10) as follows:

$$|\nabla\rho(h)| \le |\nabla(\rho(h)-\rho_h^{n+1})| + |\nabla\rho_h^{n+1}| \le C\,h\,|\rho(h)|_{H^2(\Omega)} + |\nabla\rho_h^{n+1}|.$$

Therefore, by using (4.5)

$$\begin{aligned}\|\nabla\rho_h^{n+1}\|_{L^3(\Omega)} &\le C\,h^{1/2}|\rho(h)|_{H^2(\Omega)} + C\,|\nabla\rho_h^{n+1}|^{1/2}|\rho(h)|_{H^2(\Omega)}^{1/2}\\ &\le C\,h^{1/2}|\Delta_h\rho_h^{n+1}| + C\,|\nabla\rho_h^{n+1}|^{1/2}|\Delta_h\rho_h^{n+1}|^{1/2}.\end{aligned}$$

On the other hand, by considering $\bar\rho_h = -\Delta_h\rho_h^{n+1}$ in (4.1), we get

$$|\Delta_h\rho_h^{n+1}|^2 \le |\nabla\rho_h^{n+1}||\nabla\Delta_h\rho_h^{n+1}| \le \frac{C}{h}|\nabla\rho_h^{n+1}||\Delta_h\rho_h^{n+1}|,$$

where we have used the inverse inequality between L^2 and H^1. The last two estimates imply (4.14). ∎

5. Weak convergence. To study the convergence of scheme (1.8)–(1.10), we define the following functions.

DEFINITION 5.1. *One defines $\boldsymbol{u}_{h,k}$ (respectively, $\widehat{\boldsymbol{u}}_{h,k}$, $\widehat{\boldsymbol{w}}_{h,k}$, and $p_{h,k}$) as the piecewise constant functions taking values $\boldsymbol{u}_h^{n+1}$ on $(t_n,t_{n+1}]$ (respectively, $\boldsymbol{u}_h^n$, $\boldsymbol{w}_h^n$, and p_h^{n+1}). Analogously, we define $\rho_{h,k}$ and $\widehat\rho_{h,k}$. Moreover, one defines $\widetilde\rho_{h,k}\in C^0([0,T];\boldsymbol{V}_h)$ as the piecewise linear functions such that $\widetilde\rho_{h,k}(t_n)=\rho_h^n$.*

Lemma 3.2, Corollary 3.3, Theorem 3.5, Theorem 4.1, and Corollary 4.2 imply the following estimates (independent of h and k, but some of them depend on λ):

$$\begin{gathered}\{\boldsymbol{u}_{h,k}\}_{h,k},\ \{\widehat{\boldsymbol{u}}_{h,k}\}_{h,k},\ \{\widehat{\boldsymbol{w}}_{h,k}\}_{h,k} \text{ in } L^\infty(0,T;\boldsymbol{L}^2(\Omega))\cap L^2(0,T;\boldsymbol{H}_0^1(\Omega)),\\ \{\widetilde\rho_{h,k}\}_{h,k},\ \{\rho_{h,k}\}_{h,k},\ \{\widehat\rho_{h,k}\}_{h,k} \text{ in } L^\infty(0,T;H^1(\Omega))\cap L^\infty(Q),\\ \{\rho_{h,k}\}_{h,k} \text{ in } L^4(0,T;W^{1,3}(\Omega)).\end{gathered}$$

Now let us pass to the limit in both discrete free-divergence equations (1.10) and (3.1). Consider $q \in C^0([0,T];C^\infty(\Omega))$ such that $\int_\Omega q(\boldsymbol{x})\,d\boldsymbol{x}=0$, $q_h^n = K_h q(t_n) \in M_h$, and $\bar{q}_h^n = \widetilde{K}_h q(t_n) \in \widetilde{M}_h$. Define $q_{h,k}$ and $\bar{q}_{h,k}$ as $p_{h,k}$ in Definition 5.1. On the other hand, we know that there exist two limit functions $\boldsymbol{w}$ and $\boldsymbol{u}$ belonging to $\boldsymbol{H}_0^1(\Omega)$ such that $\boldsymbol{u}_{h,k} \to \boldsymbol{u}$ and $\boldsymbol{w}_{h,k} \to \boldsymbol{w}$ weakly in $L^2(0,T;\boldsymbol{H}_0^1(\Omega))$ as $(h,k) \to 0$. Thus, we write from (1.10)

$$0 = k\sum_{n=1}^{N}(\nabla\cdot\boldsymbol{u}_h^n, q_h^n) = \int_0^T (\nabla\cdot\boldsymbol{u}_{h,k}, q_{h,k})\,dt \to \int_0^T (\nabla\cdot\boldsymbol{u}, q)\,dt = 0$$

for all $q \in C^0([0,T];C^\infty(\Omega))$, with $\int_\Omega q(\boldsymbol{x})\,d\boldsymbol{x}=0$. A density argument says $\boldsymbol{u} \in \boldsymbol{V}$. In an analogous way, we can also prove that $\boldsymbol{w} \in \boldsymbol{V}$.

Next, we wish to derive a test function for (1.7), a discrete free-divergence approximation of a function $\boldsymbol{v} \in \boldsymbol{C}_c^\infty(\Omega)$, with $\nabla\cdot\boldsymbol{v}=0$.

LEMMA 5.2. *Let $\bar{\boldsymbol{w}} \in \boldsymbol{C}_c^\infty(\Omega)$. Then there exists $\bar{\boldsymbol{w}}_h \in \widetilde{\boldsymbol{V}}_h$ such that:*

$$\bar{\boldsymbol{w}}_h \to \bar{\boldsymbol{w}} \ \text{ in } \boldsymbol{H}_0^1(\Omega) \quad \text{and} \quad (\nabla\cdot\bar{\boldsymbol{w}}_h, \bar{q}_h) = (\nabla\cdot\bar{\boldsymbol{w}}, \bar{q}_h)\ \forall\,\bar{q}_h \in \widetilde{M}_h.$$

A proof of this result can be found in [11].

By taking into account previous arguments, we can arrive at the following result.

LEMMA 5.3. *Under the hypotheses of Theorem* 3.5 *there exist subsequences of $\{\boldsymbol{u}_{h,k}\}_{h,k}$, $\{\widehat{\boldsymbol{u}}_{h,k}\}_{h,k}$, $\{\widehat{\boldsymbol{w}}_{h,k}\}_{h,k}$, $\{\rho_{h,k}\}_{h,k,\lambda}$, $\{\widehat{\rho}_{h,k}\}_{h,k}$, and $\{\widetilde{\rho}_{h,k}\}_{h,k}$ (denoted in the same way) and limit functions $\boldsymbol{u}$, ρ verifying the following weak convergences as $(h,k) \to 0$:*

$$\boldsymbol{u}_{h,k} \to \boldsymbol{u}, \quad \widehat{\boldsymbol{u}}_{h,k} \to \boldsymbol{u}, \quad \widehat{\boldsymbol{w}}_{h,k} \to \boldsymbol{u} \quad \text{in} \begin{cases} L^2(0,T;\boldsymbol{H}_0^1(\Omega))\text{-weak}, \\ L^\infty(0,T;\boldsymbol{L}^2(\Omega))\text{-weak}*, \end{cases}$$

$$\rho_{h,k} \to \rho, \quad \widehat{\rho}_{h,k} \to \rho, \quad \widetilde{\rho}_{h,k} \to \rho \quad \text{in} \begin{cases} L^\infty(Q)\text{-weak}*, \\ L^\infty(0,T;H^1(\Omega))\text{-weak}*, \end{cases}$$

$$\widehat{\rho}_{h,k} \to \rho \ \text{in } L^4(0,T;W^{1,3}(\Omega))\text{-weak}.$$

Proof. Let us prove only that $\widehat{\boldsymbol{w}}_{h,k} \to \boldsymbol{u}$ in $L^2(0,T;\boldsymbol{H}_0^1(\Omega))$-weak. Consider $\boldsymbol{v} \in C^0([0,T];\boldsymbol{C}_c^\infty(\Omega))$, with $\nabla\cdot\boldsymbol{v}=0$ and $\boldsymbol{v}_h^n \in \widetilde{\boldsymbol{V}}_h$ an approximation of $\boldsymbol{v}(t_n)$ given by Lemma 5.2. Define $\boldsymbol{v}_{h,k} \in L^\infty(0,T;\widetilde{\boldsymbol{V}}_h)$ as the piecewise constant functions taking values $\boldsymbol{v}_h^{n+1}$ on $(t_n,t_{n+1}]$ which verifies $\boldsymbol{v}_{h,k} \to \boldsymbol{v}$ in $L^\infty(0,T;\boldsymbol{H}_0^1(\Omega))$. Next, by testing (1.7) by the test function $\boldsymbol{v}_h^n$, we have $(\nabla(\boldsymbol{w}_h^n - \boldsymbol{u}_h^n), \nabla\boldsymbol{v}_n^n) = 0$, since the pressure term vanishes. By multiplying by the time step k, summing over n, and passing to limit as (h,k) tend to zero, we infer that

$$\int_0^T (\nabla(\boldsymbol{w}-\boldsymbol{u}), \nabla\boldsymbol{v}) = 0 \quad \forall\boldsymbol{v} \in C^0([0,T];\boldsymbol{C}_c^\infty(\Omega)), \ \text{with } \nabla\cdot\boldsymbol{v}=0.$$

A density argument provides that this equality holds for any $\boldsymbol{v} \in L^2(0,T;\boldsymbol{V})$. Therefore, we can choose $\boldsymbol{v} = \boldsymbol{w}-\boldsymbol{u}$ (since $\nabla\cdot\boldsymbol{u} = 0$ and $\nabla\cdot\boldsymbol{w} = 0$); then $\boldsymbol{w}=\boldsymbol{u}$. ∎

6. Strong convergence. As usual for nonlinear systems, strong convergence in some suitable space is necessary to identify the limit of the nonlinear terms.

6.1. Strong convergence for the density in $L^2(\Omega)$.

LEMMA 6.1. *Under the hypotheses of Theorem* 4.1*, one has:*

$$k\sum_{n=0}^{N}\left|\frac{\rho_h^{n+1}-\rho_h^n}{k}\right|^{4/3}\leq C_\lambda,$$

where $C_\lambda>0$ is independent of h and k (but depends on λ).

Proof. We consider $P_h: L^2(\Omega)\to W_h$ the orthogonal projector, defined as $(P_hw-w,w_h)=0$ for any $w_h\in W_h$. Let $w\in L^2(\Omega)$. By taking in (1.8) as a test function $w_h=P_hw$, we arrive at

$$\left(\frac{\rho_h^{n+1}-\rho_h^n}{k},w\right)+\left(\boldsymbol{w}_h^n\cdot\nabla\rho_h^{n+1},P_hw\right)-\lambda\Big(\Delta_h\rho_h^{n+1},w\Big)=0,$$

where we have used the definition of P_h in the first and last terms. By taking into account the stability of the projector operator $|P_hw|\leq|w|$, we get

$$\left|\frac{\rho_h^{n+1}-\rho_h^n}{k}\right|\leq\|\boldsymbol{w}_h^n\|_{L^6(\Omega)}\|\nabla\rho_h^{n+1}\|_{L^3(\Omega)}+\lambda\,|\Delta_h\rho_h^{n+1}|.$$

By summing up over n and using the estimates of $\{\widehat{\boldsymbol{w}}_{h,k}\}_{h,k}$ in $L^2(0,T;\boldsymbol{L}^6(\Omega))$ (due to the estimates in $L^2(0,T;\boldsymbol{H}^1(\Omega))$), of $\{\lambda\,\Delta_h\rho_{h,k}\}_{h,k}$ in $L^2(0,T;L^2(\Omega))$, and of $\{\nabla\rho_{h,k}\}_{h,k}$ in $L^4(0,T;\boldsymbol{L}^3(\Omega))$ (this last estimate depends on λ), we can conclude the result. ∎

Remark 6.2. As a consequence of the previous lemma, the estimate

$$\left\|\frac{d}{dt}\widetilde{\rho}_{h,k}\right\|_{L^{4/3}(0,T;L^2(\Omega))}\leq C_\lambda$$

holds. On the other hand, from Lemma 3.2, we get $\|\widetilde{\rho}_{h,k}\|_{L^\infty(0,T;H^1(\Omega))}\leq C_\lambda$. Then, thanks to an Aubin–Lions compactness argument, one has $\widetilde{\rho}_{h,k}\to\rho$ in $L^\infty(0,T;L^p(\Omega))$ as $(h,k)\to0$, with $p<6$. From this convergence, we deduce that $\rho_{h,k}$, $\widehat{\rho}_{h,k}\to\rho$ in $L^2(0,T;L^2(\Omega))$ as $(h,k)\to0$, since

$$\|\widetilde{\rho}_{h,k}-\rho_{k,h}\|^2_{L^2(0,T;L^2(\Omega))}\leq\|\widehat{\rho}_{h,k}-\rho_{k,h}\|^2_{L^2(0,T;L^2(\Omega))}=k\sum_{n=0}^{N-1}|\rho_h^{n+1}-\rho_h^n|^2\leq C\,k.$$

6.2. Strong convergence for the velocity.

PROPOSITION 6.3. *Under the hypotheses of Theorem* 4.1*, the following estimate holds:*

$$\int_0^{T-\delta}\left|\sqrt{\rho_{h,k}(t+\delta)}(\boldsymbol{u}_{h,k}(t+\delta)-\boldsymbol{u}_{h,k}(t))\right|^2dt\leq C_\lambda\,\delta^{1/4}\quad\forall\,\delta:\quad0<\delta<T,\tag{6.1}$$

where $C_\lambda>0$ is independent of h, k, and δ (but depends on λ).

Proof. Throughout the proof we will keep in mind Lemmas 3.2 and 4.1 and Remark 3.3. As $\rho_{h,k}$ and $\boldsymbol{u}_{h,k}$ are piecewise constant functions, it suffices to suppose that δ is proportional to the time step k, i.e., $\delta=r\,k$ for any $r=0,\ldots,N$. Then, to obtain (6.1), it suffices to prove that

$$k\sum_{m=0}^{N-r}|\sqrt{\rho_h^{m+r}}(\boldsymbol{u}_h^{m+r}-\boldsymbol{u}_h^m)|^2\leq C_\lambda\,(r\,k)^{1/4}\quad\forall\,r:0\leq r\leq N.\tag{6.2}$$

Let us write the time derivative of the discrete momentum system (1.9) in conservative form. By adding at the right- and left-hand sides of (1.9) the term

$$\frac{1}{2}\left(\frac{\rho_h^{n+1}-\rho_h^n}{k}, \boldsymbol{u}_h^{n+1}\cdot\bar{\boldsymbol{u}}_h\right):$$

$$\begin{cases}\left(\dfrac{\rho_h^{n+1}\boldsymbol{u}_h^{n+1}-\rho_h^n\boldsymbol{u}_h^n}{k},\bar{\boldsymbol{u}}_h\right)+a\Big(\rho_h^{n+1},\boldsymbol{u}_h^{n+1},\bar{\boldsymbol{u}}_h\Big)-\Big(p_h^{n+1},\nabla\cdot\bar{\boldsymbol{u}}_h\Big)\\ +c\Big(\rho_h^{n+1}\boldsymbol{u}_h^n-\lambda\nabla\rho_h^{n+1},\boldsymbol{u}_h^{n+1},\bar{\boldsymbol{u}}_h\Big)=\Big(\rho_h^{n+1}\boldsymbol{f}^{\,n+1},\bar{\boldsymbol{u}}_h\Big)+\dfrac{1}{2}\left(\dfrac{\rho_h^{n+1}-\rho_h^n}{k},\boldsymbol{u}_h^{n+1}\cdot\bar{\boldsymbol{u}}_h\right).\end{cases} \tag{6.3}$$

By multiplying (6.3) by k and summing for $n=m,\dots,m-1+r$, we have

$$\begin{cases}\Big(\rho_h^{m+r}\boldsymbol{u}_h^{m+r}-\rho_h^m\boldsymbol{u}_h^m,\bar{\boldsymbol{u}}_h\Big)+k\displaystyle\sum_{n=m}^{m-1+r}a\Big(\rho_h^{n+1},\boldsymbol{u}_h^{n+1},\bar{\boldsymbol{u}}_h\Big)-\sum_{n=m}^{m-1+r}\Big(p_h^{n+1},\nabla\cdot\bar{\boldsymbol{u}}_h\Big)\\ +k\displaystyle\sum_{n=m}^{m-1+r}c\Big(\rho_h^{n+1}\boldsymbol{u}_h^n-\lambda\nabla\rho_h^{n+1},\boldsymbol{u}_h^{n+1},\bar{\boldsymbol{u}}_h\Big)\\ =k\displaystyle\sum_{n=m}^{m-1+r}\Big(\rho_h^{n+1}\boldsymbol{f}^{\,n+1},\bar{\boldsymbol{u}}_h\Big)+\frac{k}{2}\sum_{n=m}^{m-1+r}\left(\frac{\rho_h^{n+1}-\rho_h^n}{k},\boldsymbol{u}_h^{n+1}\cdot\bar{\boldsymbol{u}}_h\right).\end{cases}$$

By taking $\bar{\boldsymbol{u}}_h=\boldsymbol{u}_h^{m+r}-\boldsymbol{u}_h^m$ and making use of the identity

$$\rho_h^{m+r}\boldsymbol{u}_h^{m+r}-\rho_h^m\boldsymbol{u}_h^m=\rho_h^{m+r}(\boldsymbol{u}_h^{m+r}-\boldsymbol{u}_h^m)+(\rho_h^{m+r}-\rho_h^m)\boldsymbol{u}_h^m, \tag{6.4}$$

we get

$$\begin{cases}|\sqrt{\rho_h^{m+r}}(\boldsymbol{u}_h^{m+r}-\boldsymbol{u}_h^m)|^2=-\Big(\rho_h^{m+r}-\rho_h^m,\boldsymbol{u}_h^m\cdot(\boldsymbol{u}_h^{m+r}-\boldsymbol{u}_h^m)\Big)\\ -k\displaystyle\sum_{n=m}^{m-1+r}\Big\{a\Big(\rho_h^{n+1},\boldsymbol{u}_h^{n+1},\boldsymbol{u}_h^{m+r}-\boldsymbol{u}_h^m\Big)+c\Big(\rho_h^{n+1}\boldsymbol{u}_h^n-\lambda\nabla\rho_h^{n+1},\boldsymbol{u}_h^{n+1},\boldsymbol{u}_h^{m+r}-\boldsymbol{u}_h^m\Big)\Big\}\\ +k\displaystyle\sum_{n=m}^{m-1+r}\left\{\Big(\rho_h^{n+1}\boldsymbol{f}^{\,n+1},\boldsymbol{u}_h^{m+r}-\boldsymbol{u}_h^m\Big)+\frac{1}{2}\left(\frac{\rho_h^{n+1}-\rho_h^n}{k},\boldsymbol{u}_h^{n+1}\cdot(\boldsymbol{u}_h^{m+r}-\boldsymbol{u}_h^m)\right)\right\}.\end{cases} \tag{6.5}$$

On the other hand, by taking $\bar{\rho}_h=\rho_h^{m+r}-\rho_h^m$ as a test function in the density scheme (4.2), multiplying by k, and summing for $n=m,\dots,m-1+r$, we obtain

$$\begin{aligned}|\rho_h^{m+r}-\rho_h^m|^2&=-k\sum_{n=m}^{m-1+r}\Big(\boldsymbol{w}_h^n\cdot\nabla\rho_h^{n+1}-\lambda\,\Delta_h\rho_h^{n+1},\rho_h^m-\rho_h^{m+r}\Big)\\ &\le k\sum_{n=m}^{m-1+r}(\|\boldsymbol{w}_h^n\|_{\boldsymbol{L}^6(\Omega)}\|\nabla\rho_h^n\|_{\boldsymbol{L}^3(\Omega)}+\lambda|\Delta_h\rho_h^{n+1}|)|\rho_h^m-\rho_h^{m+r}|.\end{aligned}$$

Therefore,

$$\begin{aligned}|\rho_h^m-\rho_h^{m+r}|&\le C\left(k\sum_{n=m}^{m-1+r}\|\boldsymbol{w}_h^n\|^2_{\boldsymbol{L}^6(\Omega)}\right)^{1/2}\left(k\sum_{n=m}^{m-1+r}\|\nabla\rho_h^n\|^4_{\boldsymbol{L}^3(\Omega)}\right)^{1/4}(rk)^{1/4}\\ &\quad+C\left(k\sum_{n=m}^{m-1+r}\lambda|\Delta_h\rho_h^{n+1}|^2\right)^{1/2}(rk)^{1/2}\le C_\lambda\,(rk)^{1/4}+C(rk)^{1/2},\end{aligned}$$

where $C_\lambda > 0$ is a constant independent of h and k (and depending on λ). Then we have

$$\max_{1\le m\le N} |\rho_h^m - \rho_h^{m+r}| \le C_\lambda\,(r\,k)^{1/4}. \tag{6.6}$$

By multiplying (6.5) by k and summing for $m = 0,\dots,N-r$, we are going to get the desired bound (6.2) using (6.6). Indeed, from (6.6), one can obtain (with a similar argument as in [11])

$$-k\sum_{m=0}^{N-r}\left(\rho_h^m - \rho_h^{m+r}, \boldsymbol{u}_h^m\cdot(\boldsymbol{u}_h^{m+r}-\boldsymbol{u}_h^m)\right) \le C_\lambda (rk)^{1/4}.$$

We analyze only the two terms whose estimates will be different from the ones done in [11]:

$$J_1 := -k^2\sum_{m=0}^{N-r}\sum_{n=m}^{m-1+r} c\left(\rho_h^{n+1}\boldsymbol{u}_h^n - \lambda\nabla\rho_h^{n+1}, \boldsymbol{u}_h^{n+1}, \boldsymbol{u}_h^{m+r}-\boldsymbol{u}_h^m\right),$$

$$J_2 := \frac{k^2}{2}\sum_{m=0}^{N-r}\sum_{n=m}^{m-1+r}\left(\frac{\rho_h^{n+1}-\rho_h^n}{k}, \boldsymbol{u}_h^{n+1}\cdot(\boldsymbol{u}_h^{m+r}-\boldsymbol{u}_h^m)\right).$$

To estimate J_1, we use (1.11) as follows:

$$J_1 \le C\,k^2\sum_{m=0}^{N-r}\sum_{n=m}^{m-1+r}\|\rho_h^{n+1}\boldsymbol{u}_h^n - \lambda\nabla\rho_h^{n+1}\|_{\boldsymbol{L}^3(\Omega)}\|\boldsymbol{u}_h^{n+1}\|\|\boldsymbol{u}_h^{m+r}-\boldsymbol{u}_h^m\|$$

$$\le C\,k^2\sum_{m=0}^{N-r}\sum_{n=m}^{m-1+r}\left(\|\rho_h^{n+1}\|_{L^\infty(\Omega)}\|\boldsymbol{u}_h^n\|_{\boldsymbol{L}^3(\Omega)} + \lambda\|\nabla\rho_h^{n+1}\|_{\boldsymbol{L}^3(\Omega)}\right)\|\boldsymbol{u}_h^{n+1}\|\|\boldsymbol{u}_h^{m+r}-\boldsymbol{u}_h^m\|.$$

By interchanging the sum order (Fubini's discrete rule) and using the fact that $\|\rho_h^{n+1}\|_{L^\infty(\Omega)} \le C$,

$$J_1 \le\ C\,k^2\sum_{n=0}^{N-1}\left(\|\boldsymbol{u}_h^n\|_{\boldsymbol{L}^3(\Omega)} + \lambda\|\nabla\rho_h^{n+1}\|_{\boldsymbol{L}^3(\Omega)}\right)\|\boldsymbol{u}_h^{n+1}\|\sum_{m=\overline{n-r+1}}^{\overline{n}}\|\boldsymbol{u}_h^{m+r}-\boldsymbol{u}_h^m\|,$$

where

$$\overline{n} = \begin{cases} 0 & \text{if} \quad n<0,\\ n & \text{if} \quad 0\le n\le N-r,\\ N-r & \text{if} \quad n>N-r.\end{cases}$$

Next, by taking into account that $|\overline{n} - \overline{n-r+1}| \le r$ and Corollary 4.2, we get

$$J_1 \le C\,k\sum_{n=0}^{N-1}\left(\|\boldsymbol{u}_h^n\|_{L^3(\Omega)} + \lambda\|\nabla\rho_h^{n+1}\|_{L^3(\Omega)}\right)\|\boldsymbol{u}_h^{n+1}\|$$

$$\left(\sum_{m=\overline{n-r+1}}^{\overline{n}} k\|\boldsymbol{u}_h^{m+r}-\boldsymbol{u}_h^m\|^2\right)^{1/2}\left(\sum_{m=\overline{n-r+1}}^{\overline{n}} k\right)^{1/2}$$

$$\le\ C(r\,k)^{1/2}\left(k\sum_{n=0}^{N-1}\left(\|\boldsymbol{u}_h^n\|_{L^3(\Omega)} + \lambda\|\nabla\rho_h^{n+1}\|_{L^3(\Omega)}\right)^2\right)^{1/2}\left(k\sum_{n=0}^{N-1}\|\boldsymbol{u}_h^{n+1}\|^2\right)^{1/2}$$

$$\le C\,(r\,k)^{1/2}.$$

In the same way, we can bound the term $k^2\sum_{m=0}^{N-r}\sum_{n=m}^{m-1+r} a(\rho_h^{n+1}, \boldsymbol{u}_h^{n+1}, \boldsymbol{u}_h^{m+r}-\boldsymbol{u}_h^m)$.

We bound J_2 as follows:

$$J_2 \le C\,k^2 \sum_{m=0}^{N-r}\sum_{n=m}^{m-1+r}\left|\frac{\rho_h^{n+1}-\rho_h^n}{k}\right| \|\boldsymbol{u}_h^{n+1}\|_{L^3(\Omega)}\|\boldsymbol{u}_h^{m+r}-\boldsymbol{u}_h^m\|$$

$$\text{(Fubini)} \le C\,k^2 \sum_{n=0}^{N-1}\left|\frac{\rho_h^{n+1}-\rho_h^n}{k}\right| \|\boldsymbol{u}_h^{n+1}\|^{1/2}|\boldsymbol{u}_h^{n+1}|^{1/2} \sum_{m=\overline{n-r+1}}^{\overline{n}} \|\boldsymbol{u}_h^{m+r}-\boldsymbol{u}_h^m\|$$

$$\le C\,k \sum_{n=0}^{N-1}\left|\frac{\rho_h^{n+1}-\rho_h^n}{k}\right| \|\boldsymbol{u}_h^{n+1}\|^{1/2} \left(k \sum_{m=\overline{n-r+1}}^{\overline{n}} \|\boldsymbol{u}_h^{m+r}-\boldsymbol{u}_h^m\|^2\right)^{1/2} (rk)^{1/2}$$

$$\le C\,(rk)^{1/2} \left(k\sum_{n=0}^{N-1}\left|\frac{\rho_h^{n+1}-\rho_h^n}{k}\right|^{4/3}\right)^{3/4} \left(k\sum_{n=0}^{N-1}\|\boldsymbol{u}_h^{n+1}\|^2\right)^{1/4}$$

$$\le C_\lambda (rk)^{1/2}. \qquad \square$$

Remark 6.4. From the weak estimates of the discrete velocity $\boldsymbol{u}_{h,k}$ in $L^\infty(0,T;\boldsymbol{L}^2(\Omega))\cap L^2(0,T;\boldsymbol{H}_0^1(\Omega))$ and the fractional in time estimate of $\boldsymbol{u}_{h,k}$ given in (6.1), we can apply a compactness result [17] and obtain $\boldsymbol{u}_{h,k}\to\boldsymbol{u}$ in $L^2(0,T;\boldsymbol{L}^2(\Omega))$-strong. Consequently, thanks to estimate (iii), $\widehat{\boldsymbol{u}}_{h,k}\to\boldsymbol{u}$ in $L^2(0,T;\boldsymbol{L}^2(\Omega))$-strong. Finally, since $\{\boldsymbol{u}_{h,k}\}_{h,k}$ is bounded in $L^\infty(0,T;\boldsymbol{L}^2(\Omega))$ we improve the compactness of $\{\boldsymbol{u}_{h,k}\}_{h,k}$ to the $L^p(0,T;\boldsymbol{L}^2(\Omega))$ space, with $p<\infty$.

6.3. Strong convergence for the density in $\boldsymbol{H^1(\Omega)}$. By using the compactness of the discrete density in $L^2(0,T;\boldsymbol{L}^2(\Omega))$ and comparing the equation for the discrete Laplacian and its limit (see [11]), one can obtain the convergence of the $L^2(0,T;\boldsymbol{L}^2(\Omega))$-norm of $\nabla\rho_{h,k}$ towards the same norm of $\nabla\rho$. Consequently, one has

$$\|\rho_{h,k}-\rho\|_{L^2(0,T;H^1(\Omega))}\to 0 \text{ as } (h,k)\to 0.$$

7. Passing to the limit.

7.1. Convergence for the density scheme. Thanks to the previous convergences, we can prove [11] the convergence of the density scheme as $(h,k)\to 0$, obtaining

$$\rho_t + \boldsymbol{u}\cdot\nabla\rho - \lambda\Delta\rho = 0 \text{ in } Q, \quad \frac{\partial\rho}{\partial\boldsymbol{n}}\Big|_\Sigma = 0, \quad \rho(0)=\rho_0 \text{ in } \Omega. \tag{7.1}$$

7.2. Convergence for the momentum scheme. We use the following convergence result, which is similar to Lemma 5.2.

LEMMA 7.1. *Let $\bar{\boldsymbol{u}}\in\boldsymbol{C}_c^\infty(\Omega)$. Then there exists $\bar{\boldsymbol{u}}_h\in\boldsymbol{V}_h$ such that:*

$$\bar{\boldsymbol{u}}_h\to\bar{\boldsymbol{u}} \ \text{ in } \boldsymbol{H}_0^1(\Omega) \ \text{ and } \ \Big(\nabla\cdot\bar{\boldsymbol{u}}_h, q_h\Big) = \Big(\nabla\cdot\bar{\boldsymbol{u}}, q_h\Big)\ \forall\, q_h\in M_h.$$

To pass to the limit in the discrete momentum system, we consider $\boldsymbol{v}\in C^1([0,T];\mathbf{C}_c^\infty(\Omega))$, with $\nabla\cdot\boldsymbol{v}=0$ and $\boldsymbol{v}(T)=0$. We define $\boldsymbol{v}_h^n$ as the projection of $\boldsymbol{v}(t^n)$ furnished by Lemma 7.1. We define $\boldsymbol{v}_{h,k}\in L^\infty(0,T;\boldsymbol{V}_h)$ as the piecewise constant functions taking values $\boldsymbol{v}_h^{n+1}$ on $(t_n,t_{n+1}]$ and let $\widetilde{\boldsymbol{v}}_{h,k}\in C^0([0,T];\boldsymbol{V}_h)$ be the piecewise linear, globally continuous functions and such that $\widetilde{\boldsymbol{v}}_{h,k}(t_n)=\boldsymbol{v}_h^n$. It is known that, as $(h,k)\to 0$,

$$\boldsymbol{v}_{h,k}\to\boldsymbol{v} \quad \text{in } L^\infty(0,T;\boldsymbol{H}_0^1(\Omega)), \quad \widetilde{\boldsymbol{v}}_{h,k}\to\boldsymbol{v} \quad \text{in } W^{1,\infty}(0,T;\boldsymbol{H}_0^1(\Omega)).$$

By taking $\bar{\boldsymbol{u}}_h = \boldsymbol{v}_h^{n+1}$ as a test function in (6.3), multiplying by k, summing over n, and using the expression (discrete integration by parts in time)

$$\sum_{n=0}^{N-1}\Big(\rho_h^{n+1}\boldsymbol{u}_h^{n+1}-\rho_h^n\boldsymbol{u}_h^n,\boldsymbol{v}_h^{n+1}\Big)=\Big(\rho_h^N\boldsymbol{u}_h^N,\boldsymbol{v}_h^N\Big)-\sum_{n=0}^{N-1}\Big(\rho_h^n\boldsymbol{u}_h^n,\boldsymbol{v}_h^{n+1}-\boldsymbol{v}_h^n\Big)-\Big(\rho_{0h}\boldsymbol{u}_{0h},\boldsymbol{v}_h^0\Big)$$

and the fact that $\boldsymbol{v}_h^N = 0$ (since $\boldsymbol{v}(T) = 0$), the following "conservative" formulation holds:

$$\left\{\begin{array}{l} -\,k\displaystyle\sum_{n=0}^{N-1}\Big(\rho_h^n\boldsymbol{u}_h^n,\frac{\boldsymbol{v}_h^{n+1}-\boldsymbol{v}_h^n}{k}\Big)-\Big(\rho_{0h}\boldsymbol{u}_{0h},\boldsymbol{v}_h^0\Big)\\ +\,k\displaystyle\sum_{n=0}^{N-1}c(\rho_h^{n+1}\boldsymbol{u}_h^n-\lambda\nabla\rho_h^{n+1},\boldsymbol{u}_h^{n+1},\boldsymbol{v}_h^{n+1})+k\sum_{n=0}^{N-1}a(\boldsymbol{u}_h^{n+1},\boldsymbol{v}_h^{n+1})\\ =k\displaystyle\sum_{n=0}^{N-1}\Big(\rho_h^{n+1}\boldsymbol{f}^{\,n+1},\boldsymbol{v}_h^{n+1}\Big)+k\sum_{n=0}^{N-1}\frac{1}{2}\left(\frac{\rho_h^{n+1}-\rho_h^n}{k},\boldsymbol{u}_h^{n+1}\cdot\boldsymbol{v}_h^{n+1}\right).\end{array}\right.$$

Next, by taking into account Definition 5.1,

$$\left\{\begin{array}{l} -\displaystyle\int_0^T\Big(\widehat{\rho}_{h,k}\widehat{\boldsymbol{u}}_{h,k},\frac{\partial}{\partial t}\widetilde{\boldsymbol{v}}_{h,k}\Big)-\Big(\rho_{0h}\boldsymbol{u}_{0h},\boldsymbol{v}_h^0\Big)\\ +\displaystyle\int_0^T c\Big(\rho_{h,k}\widehat{\boldsymbol{u}}_{h,k}-\lambda\nabla\rho_{h,k},\boldsymbol{u}_{h,k},\boldsymbol{v}_{h,k}\Big)+\sum_{n=0}^{N-1}a\Big(\rho_{h,k},\boldsymbol{u}_{h,k},\boldsymbol{v}_{h,k}\Big)\\ =\displaystyle\int_0^T\Big(\rho_{h,k}\boldsymbol{f}_k,\boldsymbol{v}_{h,k}\Big)+\frac{1}{2}\int_0^T\left(\frac{\partial}{\partial t}\widetilde{\rho}_{h,k},\boldsymbol{u}_{h,k}\cdot\boldsymbol{v}_{h,k}\right).\end{array}\right.$$

This variational formulation of the discrete momentum system allows us to pass to the limit in a standard way. We pass to the limit only in the last term on the right-hand side since this term does not appear in the theoretical analysis. We know that $\frac{\partial}{\partial t}\widetilde{\rho}_{h,k}\to\frac{\partial}{\partial t}\rho$ weakly in $L^{4/3}(0,T;L^2(\Omega))$ and $\boldsymbol{u}_{h,k}\to\boldsymbol{u}$ strongly in $L^p(0,T;L^2(\Omega))$, with $p<\infty$, and is bounded in $L^4(0,T;\boldsymbol{L}^3(\Omega))$, and hence $\frac{\partial}{\partial t}\widetilde{\rho}_{h,k}\boldsymbol{u}_{h,k}\to\frac{\partial}{\partial t}\rho\boldsymbol{u}$ weakly in $L^1(0,T;\boldsymbol{L}^{6/5}(\Omega))$. As $\boldsymbol{v}_{h,k}\to\boldsymbol{v}$ in $L^\infty(0,T;\boldsymbol{H}_0^1(\Omega))$, we have

$$\int_0^T\left(\frac{\partial}{\partial t}\widetilde{\rho}_{h,k},\boldsymbol{u}_{h,k}\cdot\boldsymbol{v}_{h,k}\right)\to\int_0^T\left(\frac{\partial}{\partial t}\rho,\boldsymbol{u}\cdot\boldsymbol{v}\right)\ \text{ as } (h,k)\to 0.$$

This concludes the proof of Theorem 1.1. ∎

Remark 7.2. A variant of the Kazhikhoz–Smagulov model is obtained by replacing the linear diffusion term $-\nabla\cdot(\mu\nabla\boldsymbol{u})$ in (1.1) by a nonlinear diffusion term $-\lambda\nabla\cdot(\rho\nabla\boldsymbol{u})$ (i.e., taking $\mu=\lambda\rho$). It is a model of pollution studied by Bresch, Essoufi, and Sy in [4, 5], where they prove the existence of a global in time weak solution, without imposing the restrictive hypothesis (1.5) on the coefficients.

The scheme that we design for this model is obtained by replacing the stabilizing term of the momentum system

$$-\lambda\int_\Omega\frac{M+m}{2}(\nabla\boldsymbol{u}_h^{n+1})^t:\nabla\bar{\boldsymbol{u}}_h\,d\boldsymbol{x}$$

by the term $\frac{\lambda m}{2}(\nabla\cdot\boldsymbol{u}_h^{n+1},\nabla\cdot\bar{\boldsymbol{u}}_h)$ and the remainder with the following scheme:

Given $(\rho_h^n,\boldsymbol{u}_h^n,p_h^n)\in W_h\times\boldsymbol{V}_h\times M_h$,

1. find $(\boldsymbol{w}_h^n, q_h^n) \in \widetilde{\boldsymbol{V}_h} \times \widetilde{M_h}$ such that, for each $(\bar{\boldsymbol{w}}_h, \bar{q}_h) \in \widetilde{\boldsymbol{V}_h} \times \widetilde{M_h}$,

$$\Big(\nabla \boldsymbol{w}_h^n, \nabla \bar{\boldsymbol{w}}_h\Big) - \Big(q_h^n, \nabla \cdot \bar{\boldsymbol{w}}_h\Big) = \Big(\nabla \boldsymbol{u}_h^n, \nabla \bar{\boldsymbol{w}}_h\Big), \quad \Big(\nabla \cdot \boldsymbol{w}_h^n, \bar{q}_h\Big) = 0; \tag{7.2}$$

2. find $\rho_h^{n+1} \in W_h$ such that, for each $\bar{\rho}_h \in W_h$,

$$\left(\frac{\rho_h^{n+1} - \rho_h^n}{k}, \bar{\rho}_h\right) + \Big(\boldsymbol{w}_h^n \cdot \nabla \rho_h^{n+1}, \bar{\rho}_h\Big) + \lambda \Big(\nabla \rho_h^{n+1}, \nabla \bar{\rho}_h\Big) = 0; \tag{7.3}$$

3. find $(\boldsymbol{u}_h^{n+1}, p_h^{n+1}) \in \boldsymbol{V}_h \times M_h$ such that, for each $(\bar{\boldsymbol{u}}_h, \bar{p_h}) \in \boldsymbol{V}_h \times M_h$,

$$\begin{cases} \left(\rho_h^n \dfrac{\boldsymbol{u}_h^{n+1} - \boldsymbol{u}_h^n}{k}, \bar{\boldsymbol{u}}_h\right) + \dfrac{1}{2}\left(\dfrac{\rho_h^{n+1} - \rho_h^n}{k}, \boldsymbol{u}_h^{n+1} \cdot \bar{\boldsymbol{u}}_h\right) \\ + c\Big(\rho_h^{n+1} \boldsymbol{u}_h^n - \lambda \nabla \rho_h^{n+1}, \boldsymbol{u}_h^{n+1}, \bar{\boldsymbol{u}}_h\Big) \\ + \lambda \Big(\rho_h^{n+1} (\nabla \boldsymbol{u}_h^{n+1} - (\nabla \boldsymbol{u}_h^{n+1})^t), \nabla \bar{\boldsymbol{u}}_h\Big) + \dfrac{\lambda\, m}{2}\Big(\nabla \cdot \boldsymbol{u}_h^{n+1}, \nabla \cdot \bar{\boldsymbol{u}}_h\Big) \\ = \Big(\rho_h^{n+1} \boldsymbol{f}^{\,n+1}, \bar{\boldsymbol{u}}_h\Big) + \Big(p_h^{n+1}, \nabla \cdot \bar{\boldsymbol{u}}_h\Big), \end{cases} \tag{7.4}$$

$$\Big(\nabla \cdot \boldsymbol{u}_h^{n+1}, \bar{p}_h\Big) = 0. \tag{7.5}$$

By following the arguments of this paper, one may establish the same conclusions of Theorem 1.1 for this scheme.

8. Asymptotic behavior when $\boldsymbol{\lambda \to 0}$. In this section we are interested in the asymptotic behavior of scheme (1.8)–(1.10) when the diffusion parameter λ goes to zero. More precisely, we will see that, by imposing the stability condition

$$h/(k\lambda^2) \to 0 \quad \text{as } (\lambda, h, k) \to 0 \tag{S$'$}$$

and completing (H2) with the additional approximation property

$$|\bar{\rho} - I_h \bar{\rho}| \le C\, h^{2/3} \|\bar{\rho}\|_{W^{1,3/2}(\Omega)} \quad \forall\, \bar{\rho} \in W^{1,3/2}(\Omega), \tag{H2$'$}$$

then scheme (1.8)–(1.10) approximates, as $(h, k, \lambda) \to 0$, to a weak solution of the density-dependent Navier–Stokes problem (1.6), which is defined as follows.

DEFINITION 8.1. *A pair $(\rho, \boldsymbol{u})$ is said to be a weak solution of* (1.6) *in* $(0, T)$ *if:*

(a) $\boldsymbol{u} \in L^\infty(0, T; \boldsymbol{L}^2(\Omega)) \cap L^2(0, T; \boldsymbol{V})$, $\rho \in L^\infty(Q)$, *with* $0 < m \le \rho(\boldsymbol{x}, t) \le M$ *a.e.* $(\boldsymbol{x}, t) \in Q$;

(b) *for all* $\phi \in C^1([0, T]; \boldsymbol{V})$, *with* $\phi(T) = 0$,

$$\int_0^T \{-(\rho \boldsymbol{u}, \phi_t + (\boldsymbol{u} \cdot \nabla)\phi) + \mu(\nabla \boldsymbol{u}, \nabla \phi)\}\, dt = \int_0^T (\rho \boldsymbol{f}, \phi)\, dt + (\rho_0 \boldsymbol{u}_0, \phi(0));$$

(c) *for all* $\varphi \in C^1([0, T]; \boldsymbol{H}^1(\Omega))$, *with* $\varphi(T) = 0$,

$$-\int_0^T (\rho, \varphi_t)\, dt - \int_0^T (\rho \boldsymbol{u}, \nabla \varphi)\, dt = (\rho_0, \varphi(0)).$$

The rest of this section is devoted to the proof of Theorem 1.2.

8.1. Uniform estimates with respect to (h,k,λ). By following arguments of the previous sections and assuming (S′) and (h,k,λ) small enough, we can obtain the following estimates independent of h, k, and λ (now we denote piecewise functions associated to the scheme also with the parameter λ explicitly):

$$\begin{aligned}
\{\boldsymbol{u}_{h,k,\lambda}\}_{h,k,\lambda},\ \{\widehat{\boldsymbol{u}}_{h,k,\lambda}\}_{h,k,\lambda},\{\widehat{\boldsymbol{w}}_{h,k,\lambda}\}_{h,k,\lambda} &\text{ in } L^\infty(0,T;\boldsymbol{L}^2(\Omega))\cap L^2(0,T;\boldsymbol{H}^1_0(\Omega)),\\
\{\widetilde{\rho}_{h,k,\lambda}\}_{h,k,\lambda},\ \{\rho_{h,k,\lambda}\}_{h,k,\lambda},\ \{\widehat{\rho}_{h,k,\lambda}\}_{h,k,\lambda} &\text{ in } L^\infty(Q),\\
\lambda^{1/2}\{\rho_{h,k,\lambda}\}_{h,k,\lambda},\ \lambda^{1/2}\{\widehat{\rho}_{h,k,\lambda}\}_{h,k,\lambda} &\text{ in } L^\infty(0,T;H^1(\Omega)),\\
\lambda^{3/4}\{\rho_{h,k,\lambda}\}_{h,k,\lambda} &\text{ in } L^4(0,T;W^{1,3}(\Omega)),\\
\lambda^{3/4}\left\{\frac{\partial}{\partial t}\widetilde{\rho}_{h,k,\lambda}\right\}_{h,k,\lambda} &\text{ in } L^{4/3}(0,T;L^2(\Omega)).
\end{aligned}$$

In addition, $0<\widetilde{m}\le\rho_{h,k,\lambda},\widehat{\rho}_{h,k,\lambda},\widetilde{\rho}_{h,k,\lambda}\le\widetilde{M}$ in Q,

$$\begin{aligned}
&\|\widetilde{\rho}_{h,k,\lambda}-\rho_{h,k,\lambda}\|_{L^2(0,T;L^2(\Omega))}\le\|\rho_{h,k,\lambda}-\widehat{\rho}_{h,k,\lambda}\|_{L^2(0,T;L^2(\Omega))}\le C\sqrt{k},\\
&\|\boldsymbol{u}_{h,k,\lambda}-\widehat{\boldsymbol{u}}_{h,k,\lambda}\|_{L^2(0,T;\boldsymbol{H}^1_0(\Omega))}\le C\sqrt{k}.
\end{aligned}$$

In fact, we have the following result.

LEMMA 8.2. *By assuming* (S′) *and* (h,k,λ) *small enough, there exist subsequences of* $\{\boldsymbol{u}_{h,k,\lambda}\}_{h,k}$, $\{\widehat{\boldsymbol{u}}_{h,k,\lambda}\}_{h,k}$, $\{\rho_{h,k,\lambda}\}_{h,k,\lambda}$, $\{\widehat{\rho}_{h,k,\lambda}\}_{h,k}$, *and* $\{\widetilde{\rho}_{h,k,\lambda}\}_{h,k}$ *(denoted in the same way) and limit functions* $\boldsymbol{u}$ *and* ρ, *such that* $\nabla\cdot\boldsymbol{u}=0$ *a.e. in* Ω, *and the following weak convergences hold as* $(h,k,\lambda)\to 0$*:*

$$\begin{aligned}
&\boldsymbol{u}_{h,k,\lambda}\to\boldsymbol{u},\quad \widehat{\boldsymbol{u}}_{h,k,\lambda}\to\boldsymbol{u},\quad \widehat{\boldsymbol{w}}_{h,k,\lambda}\to\boldsymbol{u}\quad \textit{in}\ \begin{cases} L^2(0,T;\boldsymbol{H}^1_0(\Omega))\textit{-weak},\\ L^\infty(0,T;\boldsymbol{L}^2(\Omega))\textit{-weak}*,\end{cases}\\
&\widetilde{\rho}_{h,k,\lambda}\to\rho,\quad \rho_{h,k,\lambda}\to\rho,\quad \widehat{\rho}_{h,k,\lambda}\to\rho\quad \textit{in } L^\infty(Q)\textit{-weak}*\,.
\end{aligned}$$

8.2. Compactness.

PROPOSITION 8.3. *Under the hypotheses of Lemma* 8.2*, the estimate*

$$\int_0^{T-\delta}\left|\sqrt{\rho_{h,k,\lambda}(t+\delta)}(\boldsymbol{u}_{h,k,\lambda}(t+\delta)-\boldsymbol{u}_{h,k,\lambda}(t))\right|^2 dt\le C\,\delta^{1/2}\quad\forall\,\delta:0<\delta<T \tag{8.1}$$

holds, with $C>0$ *independent of* h, k, δ, *and* λ.

In particular, since $\rho_{h,k,\lambda}\ge\widetilde{m}$, then

$$\boldsymbol{u}_{h,k,\lambda}\to\boldsymbol{u},\quad \boldsymbol{w}_{h,k,\lambda}\to\boldsymbol{u}\quad\text{in } L^2(0,T;\boldsymbol{L}^2(\Omega))\ \text{ as } (h,k,\lambda)\to 0.$$

Proof. Again, since $\rho_{h,k,\lambda}$ and $\boldsymbol{u}_{h,k,\lambda}$ are piecewise constant functions, (8.1) is equivalent to

$$k\sum_{m=0}^{N-r}|\sqrt{\rho_h^{m+r}}(\boldsymbol{u}_h^{m+r}-\boldsymbol{u}_h^m)|^2\le C\,(r\,k)^{1/2}\quad\forall\,r:0\le r\le N. \tag{8.2}$$

By following the proof of Proposition 6.3, one observes that the terms which are not necessarily bounded independent of λ are

$$T_1:=k^2\sum_{m=0}^{N-r}\sum_{n=m}^{m-1+r}\left(\frac{\rho_h^{n+1}-\rho_h^n}{k},\boldsymbol{u}_h^{n+1}\cdot(\boldsymbol{u}_h^{m+r}-\boldsymbol{u}_h^m)\right)$$

and

$$T_2 := -k \sum_{m=0}^{N-r} \left(\rho_h^{m+r} - \rho_h^m, \boldsymbol{u}_h^m \cdot (\boldsymbol{u}_h^{m+r} - \boldsymbol{u}_h^m)\right).$$

Here we are going to bound these terms.

We consider the projection operator on W_h with respect to $L^2(\Omega)$-inner product:

$$P_h : L^2(\Omega) \to W_h \quad \text{such that} \quad (P_h v, w) = (v, w) \quad \forall w \in W_h.$$

We take $\bar{\rho}_h = P_h(\boldsymbol{u}_h^{n+1} \cdot (\boldsymbol{u}_h^{m+r} - \boldsymbol{u}_h^m)) \in W_h$ in (4.2), arriving at

$$\begin{aligned} &\left(\frac{\rho_h^{n+1} - \rho_h^n}{k}, \boldsymbol{u}_h^{n+1} \cdot (\boldsymbol{u}_h^{m+r} - \boldsymbol{u}_h^m)\right) + (\boldsymbol{w}_h^n \cdot \nabla \rho_h^{n+1}, P_h(\boldsymbol{u}_h^{n+1} \cdot (\boldsymbol{u}_h^{m+r} - \boldsymbol{u}_h^m))) \\ &\quad - \lambda\Big(\Delta_h \rho_h^{n+1}, \boldsymbol{u}_h^{n+1} \cdot (\boldsymbol{u}_h^{m+r} - \boldsymbol{u}_h^m)\Big) = 0, \end{aligned} \tag{8.3}$$

where we have used the definition of projection operator in $L^2(\Omega)$ in the first and the last term. Thus, we decompose $T_1 = T_{1,1} + T_{1,2}$, where:

$$T_{1,1} = -k^2 \sum_{m=0}^{N-r} \sum_{n=m}^{m-1+r} (\boldsymbol{w}_h^n \cdot \nabla \rho_h^{n+1}, P_h(\boldsymbol{u}_h^{n+1} \cdot (\boldsymbol{u}_h^{m+r} - \boldsymbol{u}_h^m))),$$

$$T_{1,2} = \lambda k^2 \sum_{m=0}^{N-r} \sum_{n=m}^{m-1+r} (\Delta_h \rho_h^{n+1}, \boldsymbol{u}_h^{n+1} \cdot (\boldsymbol{u}_h^{m+r} - \boldsymbol{u}_h^m)).$$

By writing $T_{1,1}$ as follows:

$$\begin{aligned} T_{1,1} = &-k^2 \sum_{m=0}^{N-r} \sum_{n=m}^{m-1+r} (\boldsymbol{w}_h^n \cdot \nabla \rho_h^{n+1}, P_h(\boldsymbol{u}_h^{n+1} \cdot (\boldsymbol{u}_h^{m+r} - \boldsymbol{u}_h^m)) - \boldsymbol{u}_h^{n+1} \cdot (\boldsymbol{u}_h^{m+r} - \boldsymbol{u}_h^m)) \\ &- k^2 \sum_{m=0}^{N-r} \sum_{n=m}^{m-1+r} (\boldsymbol{w}_h^n \cdot \nabla \rho_h^{n+1}, \boldsymbol{u}_h^{n+1} \cdot (\boldsymbol{u}_h^{m+r} - \boldsymbol{u}_h^m)) := T_{1,1}^1 + T_{1,1}^2, \end{aligned}$$

we bound it as

$$T_{1,1}^1 \leq C\, k^2 \sum_{m=0}^{N-r} \sum_{n=m}^{m-1+r} \|\boldsymbol{w}_h^n\|_{L^6(\Omega)} \|\nabla \rho_h^{n+1}\|_{L^3(\Omega)} h^{2/3} \|\boldsymbol{u}_h^{n+1}\| \|\boldsymbol{u}_h^{m+r} - \boldsymbol{u}_h^m\|,$$

where we have used the inequalities $|\bar{\rho} - P_h\bar{\rho}| \leq |\bar{\rho} - I_h\bar{\rho}| \leq C\, h^{2/3} \|\bar{\rho}\|_{W^{1,3/2}(\Omega)}$ and $\|\boldsymbol{u}_h^{n+1} \cdot (\boldsymbol{u}_h^{m+r} - \boldsymbol{u}_h^m)\|_{W^{1,3/2}(\Omega)} \leq \|\boldsymbol{u}_h^{n+1}\| \|\boldsymbol{u}_h^{m+r} - \boldsymbol{u}_h^m\|$. Next, by taking (λ, h, k) small enough and such that

$$h^{2/3} \leq C\, \lambda^{3/4} k^{1/2}, \tag{8.4}$$

using the fact that $\lambda^{3/4} k^{1/2} \|\nabla \rho_h^{n+1}\|_{L^3(\Omega)} \leq C$ (with C independent of λ, h, k), and applying the discrete Fubini rule, we arrive at the estimate $T_{1,1}^1 \leq C\,(rk)^{1/2}$.

The bound $T_{1,1}^2 \le C\,(rk)^{1/2}$ is obtained easily by integrating by parts and using the pointwise bound $\|\rho_h^{n+1}\|_{L^\infty(\Omega)} \le C$ (where C is independent of λ):

$$\begin{aligned}T_{1,1}^2 &\le k^2 \sum_{m=0}^{N-r}\sum_{n=m}^{m-1+r} (|\nabla\cdot \boldsymbol{w}_h^n||\boldsymbol{u}_h^{n+1}\cdot(\boldsymbol{u}_h^{m+r}-\boldsymbol{u}_h^m)| + \|\boldsymbol{w}_h^n\|_{L^6}|\nabla \boldsymbol{u}_h^{n+1}|\|\boldsymbol{u}_h^{m+r}-\boldsymbol{u}_h^m\|_{L^3})\\ &\quad + k^2 \sum_{m=0}^{N-r}\sum_{n=m}^{m-1+r} \|\boldsymbol{w}_h^n\|_{L^6}\|\boldsymbol{u}_h^{n+1}\|_{L^3}|\nabla(\boldsymbol{u}_h^{m+r}-\boldsymbol{u}_h^m)|\\ &\le C\,k^2 \sum_{n=0}^{N-1} \|\boldsymbol{w}_h^n\|\|\boldsymbol{u}_h^{n+1}\| \sum_{m=\overline{n-r+1}}^{\overline{n}} \|\boldsymbol{u}_h^{m+r}-\boldsymbol{u}^m\| \le C(r\,k)^{1/2}.\end{aligned}$$

The bound $T_{1,2} \le C(rk)^{1/2}$ is obtained using the fact that $\lambda^2\,k\sum_{n=0}^{N-1}|\Delta_h\rho_h^{n+1}|^2 \le C$ and Fubini's rule.

By summing up (8.3) multiplied by k^2 for $n = m,\dots,m+r-1$ and then summing up for $m = 0,\dots,N-r$, we can write T_2 as:

$$T_2 = -k^2 \sum_{m=0}^{N-r}\sum_{n=m}^{m+r-1} (\boldsymbol{w}_h^n\cdot\nabla\rho_h^n, P_h(\boldsymbol{u}_h^m\cdot(\boldsymbol{u}_h^{m+r}-\boldsymbol{u}_h^m)) - \lambda(\Delta_h\rho_h^{n+1}, \boldsymbol{u}_h^m\cdot(\boldsymbol{u}_h^{m+r}-\boldsymbol{u}_h^m)).$$

These new terms are bounded in an analogous manner (now, without using Fubinis's rule), resulting in

$$T_2 \le C(rk)^{1/2}.$$

Note that the restriction on the parameters (8.4) imposed to get the bound of $T_{1,1}^1$ is included in the constraint (S′) imposed to obtain a priori estimates. Indeed, (8.4) is equivalent to $\frac{1}{\lambda^{3/4}}\frac{h^{2/3}}{k^{1/4}} \le C$. But as

$$\lim_{(h,k,\lambda)\to 0}\left(\frac{1}{\lambda}\sqrt{\frac{h}{k}}\right)\Big/\left(\frac{1}{\lambda^{3/4}}\frac{h^{2/3}}{k^{1/4}}\right) = \lim_{(h,k,\lambda)\to 0}\frac{1}{\lambda^{1/4}k^{1/2}h^{1/6}} = \infty,$$

then $\frac{1}{\lambda^{3/4}}\frac{h^{2/3}}{k^{1/4}} \le C\frac{1}{\lambda}\sqrt{\frac{h}{k}} \to 0$ (thanks to (S′)). ∎

Remark 8.4. By comparing the fractional in time estimates for the discrete velocity in the case of λ fixed (done in Proposition 6.3) with respect to the case of $\lambda \to 0$ (done now in Proposition 8.3), one can observe that condition (8.4) on the parameters h and k imposed now in the proof of Proposition 8.3 is not necessary in the case of λ fixed; however, in this case only order $(rk)^{1/4}$ is obtained. Now, in the case $\lambda \to 0$, we have to impose constraint (8.4), and the estimate order is improved from $(rk)^{1/4}$ to $(rk)^{1/2}$.

8.3. Passing to the limit.

8.3.1. Density equation. Let $\eta \in C^1([0,T];C_c^\infty(\Omega))$ such that $\eta(T) = 0$. We define η_h^n as the interpolation in W_h of $\eta(t_n)$ and $\eta_{h,k} \in L^\infty(0,T;W_h))$ as the piecewise constant function taking values η_h^{n+1} in $(t_n,t_{n+1}]$ and let $\widetilde{\eta}_{h,k} \in C^0([0,T];W_h))$ be the piecewise linear, globally continuous function such that $\widetilde{\eta}_{h,k}(t_n) = \eta_h^n$. One has, as $(h,k) \to 0$,

$$\eta_{k,h} \to \eta \quad \text{in } L^\infty(0,T;H^1(\Omega)) \quad \text{and} \quad \widetilde{\eta}_{k,h} \to \eta \quad \text{in } W^{1,\infty}(0,T;H^1(\Omega)).$$

By using these discrete test functions in the discrete density formulation (1.8), together with an integration by parts in time, we arrive [11] at the formulation

$$\int_0^T \Big(\widehat{\rho}_{h,k,\lambda}, -\frac{d}{dt}\widetilde{\eta}_{h,k}\Big) + \lambda(\nabla\rho_{h,k,\lambda}, \nabla\eta_{h,k}) + (\widehat{\boldsymbol{w}}_{h,k,\lambda}\cdot\nabla\rho_{h,k,\lambda}, \eta_{h,k}) = (\rho_{0h}, \eta_h^0). \tag{8.5}$$

Before taking a limit in (8.5), we rewrite the convective term in the following form:

$$\int_0^T (\widehat{\boldsymbol{w}}_{h,k,\lambda}\cdot\nabla\rho_{h,k,\lambda}, \eta_{h,k})\,dt = -\int_0^T (\nabla\cdot\widehat{\boldsymbol{w}}_{h,k,\lambda}\rho_{h,k,\lambda}, \eta_{h,k})\,dt,$$
$$-\int_0^T (\rho_{h,k,\lambda}\widehat{\boldsymbol{w}}_{h,k,\lambda}, \nabla\eta_{h,k})\,dt = -\int_0^T (\rho_{h,k,\lambda}\widehat{\boldsymbol{w}}_{h,k,\lambda}, \nabla\eta_{h,k})\,dt,$$

where we have used Remark 3.1. Therefore, (8.5) remains as

$$\int_0^T \Big(\widehat{\rho}_{h,k,\lambda}, -\frac{d}{dt}\widetilde{\eta}_{h,k}\Big) + \lambda(\nabla\rho_{h,k,\lambda}, \nabla\eta_{h,k}) - (\widehat{\boldsymbol{w}}_{h,k,\lambda}\,\rho_{h,k,\lambda}, \nabla\eta_{h,k}) = (\rho_{0h}, \eta_h^0).$$

By taking into account the weak and strong convergences obtained in previous subsections, it is possible to pass to the limit. Notice that, by arguing as in [9], one can prove that

$$\lambda\int_0^T (\nabla\rho_{h,k,\lambda}, \nabla\eta_{h,k})\,dt \to 0 \quad \text{as } (h,k,\lambda)\to 0.$$

8.3.2. Velocity system. Before taking a limit in the discrete momentum system (1.9), we will write (1.9) in a completely conservative form. For this, summing to both sides of (1.9) the terms

$$\frac{1}{2}\left(\frac{\rho_h^{n+1}-\rho_h^n}{k}, \boldsymbol{u}_h^{n+1}\cdot\bar{\boldsymbol{u}}_h\right) - \frac{1}{2}(\rho_h^{n+1}\boldsymbol{u}_h^n - \lambda\nabla\rho_h^{n+1}, \nabla(\boldsymbol{u}_h^{n+1}\cdot\bar{\boldsymbol{u}}_h))$$

provides that, for each $\bar{\boldsymbol{u}}_h \in \mathbf{V}_h$,

$$\begin{cases} \left(\dfrac{\rho_h^{n+1}\boldsymbol{u}_h^{n+1}-\rho_h^n\boldsymbol{u}_h^n}{k}, \bar{\boldsymbol{u}}_h\right) - ((\rho_h^{n+1}\boldsymbol{u}_h^n - \lambda\nabla\rho_h^{n+1})\otimes\boldsymbol{u}_h^{n+1}, \nabla\bar{\boldsymbol{u}}_h) \\ + a\Big(\rho_h^{n+1}, \boldsymbol{u}_h^{n+1}, \bar{\boldsymbol{u}}_h\Big) = (\rho_h^{n+1}\boldsymbol{f}^{\,n+1}, \bar{\boldsymbol{u}}_h) + (p_h^{n+1}, \nabla\cdot\bar{\boldsymbol{u}}_h) \\ + \dfrac{1}{2}\left(\dfrac{\rho_h^{n+1}-\rho_h^n}{k}, \boldsymbol{u}_h^{n+1}\cdot\bar{\boldsymbol{u}}_h\right) - \dfrac{1}{2}(\rho_h^{n+1}\boldsymbol{u}_h^n - \lambda\nabla\rho_h^{n+1}, \nabla(\boldsymbol{u}_h^{n+1}\cdot\bar{\boldsymbol{u}}_h)). \end{cases} \tag{8.6}$$

Now we consider Q_h the projector operator onto W_h with respect to the H^1-norm:

$$Q_h : H^1(\Omega) \to W_h \text{ such that } (Q_h v, w)_{H^1(\Omega)} = (v,w)_{H^1(\Omega)} \quad \forall w \in W_h,$$

where $(\cdot,\cdot)_{H^1(\Omega)}$ denotes the usual H^1-inner product.

By subtracting from the second member of (8.6) the result of taking $\bar\rho_h = \frac12 Q_h(\boldsymbol{u}_h^{n+1}\cdot \bar{\boldsymbol{u}}_h)$ as a test function in (1.8) and integrating by parts the convective term, one has

$$\begin{cases} \left(\dfrac{\rho_h^{n+1}\boldsymbol{u}_h^{n+1}-\rho_h^n\boldsymbol{u}_h^n}{k},\bar{\boldsymbol{u}}_h\right)-\Big((\rho_h^{n+1}\boldsymbol{u}_h^n-\lambda\nabla\rho_h^{n+1})\otimes\boldsymbol{u}_h^{n+1},\nabla\bar{\boldsymbol{u}}_h\Big)\\ +a\Big(\rho_h^{n+1},\boldsymbol{u}_h^{n+1},\bar{\boldsymbol{u}}_h\Big)=\Big(\rho_h^{n+1}\boldsymbol{f}^{\,n+1},\bar{\boldsymbol{u}}_h\Big)+\Big(p_h^{n+1},\nabla\cdot\bar{\boldsymbol{u}}_h\Big)\\ +\dfrac12\left(\dfrac{\rho_h^{n+1}-\rho_h^n}{k},\boldsymbol{u}_h^{n+1}\cdot\bar{\boldsymbol{u}}_h-Q_h(\boldsymbol{u}_h^{n+1}\cdot\bar{\boldsymbol{u}}_h)\right)\\ -\dfrac12\Big(\rho_h^{n+1}\boldsymbol{w}_h^n,\nabla(\boldsymbol{u}_h^{n+1}\cdot\bar{\boldsymbol{u}}_h-Q_h(\boldsymbol{u}_h^{n+1}\cdot\bar{\boldsymbol{u}}_h))\Big)\\ -\dfrac12\Big(\rho_h^{n+1}(\boldsymbol{u}_h^n-\boldsymbol{w}_h^n),\nabla(\boldsymbol{u}_h^{n+1}\cdot\bar{\boldsymbol{u}}_h)\Big)+\dfrac12\Big(\nabla\cdot\boldsymbol{w}_h^n\,\rho_h^{n+1},Q_h(\boldsymbol{u}_h^{n+1}\cdot\bar{\boldsymbol{u}}_h)\Big)\\ -\dfrac{\lambda}{2}\Big(\rho_h^{n+1},\boldsymbol{u}_h^{n+1}\cdot\bar{\boldsymbol{u}}_h-Q_h(\boldsymbol{u}_h^{n+1}\cdot\bar{\boldsymbol{u}}_h)\Big), \end{cases}$$

where we have used the definition of Q_h in the last term on the right-hand side. Note that the term $(\nabla\cdot\boldsymbol{w}_h^n\,\rho_h^{n+1},Q_h(\boldsymbol{u}_h^{n+1}\cdot\bar{\boldsymbol{u}}_h))=0$ thanks to Remark 3.1.

In a similar way to section 7, we arrive at

$$\begin{cases} -\displaystyle\int_0^T\Big(\widehat\rho_{h,k,\lambda}\widehat{\boldsymbol{u}}_{h,k,\lambda},\frac{\partial}{\partial t}\widetilde{\boldsymbol{v}}_{h,k}\Big)-\Big(\rho_{0h}\boldsymbol{u}_{0h},\boldsymbol{v}_h^0\Big)\\ -\displaystyle\int_0^T\Big(\rho_{h,k,\lambda}\widehat{\boldsymbol{u}}_{h,k}-\lambda\nabla\rho_{h,k,\lambda}\otimes\boldsymbol{u}_{h,k,\lambda},\nabla\boldsymbol{v}_{h,k}\Big)+a\Big(\rho_{h,k,\lambda},\boldsymbol{u}_{h,k,\lambda},\boldsymbol{v}_{h,k}\Big)\\ =\displaystyle\int_0^T\Big(\rho_{h,k,\lambda}\boldsymbol{f}_k,\boldsymbol{v}_{h,k}\Big)+\frac12\int_0^T\left(\frac{\partial}{\partial t}\widetilde\rho_{h,k},\boldsymbol{u}_{h,k,\lambda}\cdot\boldsymbol{v}_{h,k}-Q_h(\boldsymbol{u}_{h,k,\lambda}\cdot\boldsymbol{v}_{h,k})\right)\\ -\dfrac12\displaystyle\int_0^T\Big(\rho_{h,k,\lambda}\widehat{\boldsymbol{w}}_{h,k,\lambda},\nabla(\boldsymbol{u}_{h,k,\lambda}\cdot\boldsymbol{v}_{h,k}-Q_h(\boldsymbol{u}_{h,k,\lambda}\cdot\boldsymbol{v}_{h,k}))\Big)\\ -\dfrac12\displaystyle\int_0^T\Big(\rho_{h,k,\lambda}(\widehat{\boldsymbol{u}}_{h,k,\lambda}-\widehat{\boldsymbol{w}}_{h,k,\lambda}),\nabla(\boldsymbol{u}_{h,k,\lambda}\cdot\boldsymbol{v}_{h,k}-Q_h(\boldsymbol{u}_{h,k,\lambda}\cdot\boldsymbol{v}_{h,k}))\Big)\\ -\dfrac{\lambda}{2}k\displaystyle\int_0^T\Big(\rho_{h,k,\lambda},\boldsymbol{u}_{h,k,\lambda}\cdot\boldsymbol{v}_{h,k}-Q_h(\boldsymbol{u}_{h,k,\lambda}\cdot\boldsymbol{v}_{h,k})\Big)\\ :=\displaystyle\int_0^T\Big(\rho_{h,k,\lambda}\boldsymbol{f}_k,\boldsymbol{v}_{h,k}\Big)+R_1+R_2+R_3+R_4, \end{cases}$$

where $\boldsymbol{v}_{h,k}$ and $\widetilde{\boldsymbol{v}}_{h,k}$ are suitable approximations of a test function $\boldsymbol{v}\in C^1([0,T];\boldsymbol{C}_c^\infty(\Omega))$.

Again, by using the estimates independent of λ and arguing as in the continuous case [9], one can prove that

$$\lambda\int_0^T(\nabla\rho_{h,k,\lambda}\otimes\boldsymbol{u}_{h,k,\lambda},\nabla\boldsymbol{v}_{h,k})\to0\quad\text{ as }(h,k,\lambda)\to0,$$

$$\lambda\int_0^T\left(\rho_{h,k,\lambda}-\frac{\widetilde M+\widetilde m}{2}\right)(\nabla\boldsymbol{u}_{h,k,\lambda})^t,\nabla\boldsymbol{v}_{h,k}\to0\quad\text{ as }(h,k,\lambda)\to0.$$

To finish the passage to the limit, we show only that the residual terms R_i vanish as $(h,k,\lambda)\to 0$. For this, we impose that the sequence of test functions $\boldsymbol{v}_{h,k}$ is bounded in $L^\infty(0,T;\boldsymbol{W}^{1,3}(\Omega)\cap\boldsymbol{L}^\infty(\Omega))$.

We bound R_1, thanks to estimates (ii) and (vi) of Lemma 3.2, as follows:

$$R_1 \le C\sum_{n=0}^{N-1}|\rho_h^{n+1}-\rho_h^n|h\|\boldsymbol{u}_h^{n+1}\|\|\boldsymbol{v}_h^n\|_{W^{1,3}\cap L^\infty}\le C\frac{h}{k}\to 0\quad\text{(thanks to (S$'$))}.$$

By integrating by parts R_2,

$$\begin{aligned}R_2 &= -\int_0^T\Big(\widehat{\boldsymbol{u}}_{h,k,\lambda}\cdot\nabla\rho_{h,k,\lambda},\boldsymbol{u}_{h,k,\lambda}\cdot\boldsymbol{v}_{h,k}-Q_h(\boldsymbol{u}_{h,k,\lambda}\cdot\boldsymbol{v}_{h,k})\Big)\\ &\quad-\int_0^T\Big(\nabla\cdot\widehat{\boldsymbol{u}}_{h,k,\lambda}\,\rho_{h,k,\lambda},\boldsymbol{u}_{h,k,\lambda}\cdot\boldsymbol{v}_{h,k}-Q_h(\boldsymbol{u}_{h,k,\lambda}\cdot\boldsymbol{v}_{h,k})\Big):=R_2^1+R_2^2.\end{aligned}$$

By using the (duality) result of the Aubin–Nitsche type $|u-Q_hu|\le C\,h\|u\|_{H^1(\Omega)}$, the first term of R_2 can be estimated as follows:

$$\begin{aligned}R_2^1&\le\int_0^T\|\widehat{\boldsymbol{u}}_{h,k,\lambda}\|\|\nabla\rho_{h,k,\lambda}\|_{L^3(\Omega)}h\|\boldsymbol{u}_{h,k,\lambda}\|\|\boldsymbol{v}_{h,k}\|_{W^{1,\infty}(\Omega)}\\ &\le C\frac{h}{k^{1/4}\lambda^{3/4}}\|\widehat{\boldsymbol{u}}_{h,k,\lambda}\|_{L^2(0,T;H_0^1(\Omega))}\lambda^{3/4}k^{1/4}\|\nabla\rho_{h,k,\lambda}\|_{L^\infty(0,T;L^3(\Omega))}\\ &\qquad\|\boldsymbol{u}_{h,k,\lambda}\|_{L^2(0,T;H_0^1(\Omega))}\\ &\le C\frac{h}{k^{1/4}\lambda^{3/4}}\lambda^{3/4}\|\nabla\rho_{h,k,\lambda}\|_{L^4(0,T;L^3(\Omega))}\le C\frac{h}{k^{1/4}\lambda^{3/4}}\to 0\quad\text{(thanks to (S$'$))}.\end{aligned}$$

The convergence to zero of the other term R_2^2 can be made in a similar way.

The term R_3 is handled as follows:

$$\begin{aligned}R_3&\le\int_0^T\|\rho_{h,k,\lambda}\|_{L^\infty(\Omega)}|\widehat{\boldsymbol{u}}_{h,k,\lambda}-\widehat{\boldsymbol{w}}_{h,k,\lambda}||\nabla(\boldsymbol{u}_{h,k,\lambda}\cdot\boldsymbol{v}_{h,k}-Q_h(\boldsymbol{u}_{h,k,\lambda}\cdot\boldsymbol{v}_{h,k}))|\\ &\le C\,h\int_0^T\|\widehat{\boldsymbol{u}}_{h,k,\lambda}-\widehat{\boldsymbol{w}}_{h,k,\lambda}\|\|\boldsymbol{u}_{h,k,\lambda}\|\|\boldsymbol{v}_{h,k}\|_{W^{1,3}(\Omega)\cap L^\infty(\Omega)}\le C\,h\to 0,\end{aligned}$$

where we have used (3.8) and the stability property of Q_h in the H^1-norm.

Finally, the convergence to zero of R_4 is easy to deduce. This concludes the proof of Theorem 1.2. ∎

Remark 8.5. The asymptotic behavior as λ goes to zero of the scheme (7.2)–(7.5) (see Remark 7.2), associated to a problem with density-dependent diffusion, remains as an open problem. In fact, when $\lambda\to 0$, both diffusion coefficients (viscosity and mass diffusion) vanish. Therefore, we find a viscosity-vanishing problem, which is an open problem even in the continuous case.

Appendix A: Proof of Lemma 3.4. We consider the following auxiliary semi-discrete scheme: Find $\rho^{n+1}\in H^2(\Omega)$ as the solution of the problem:

$$\text{(A.1)}\qquad \frac{\rho^{n+1}-\rho^n}{k}+\boldsymbol{w}_h^n\cdot\nabla T_m^M\rho^{n+1}-\lambda\Delta\rho^{n+1}=0\ \text{ in }\Omega,\quad \frac{\partial\rho^{n+1}}{\partial\boldsymbol{n}}\Big|_{\partial\Omega}=0,$$

where

$$T_m^M\rho^{n+1}(\boldsymbol{x},t)=\begin{cases}\rho^{n+1}(\boldsymbol{x},t) & \text{if}\quad \rho^{n+1}(\boldsymbol{x},t)\in[m,M],\\ m & \text{if}\quad \rho^{n+1}(\boldsymbol{x},t)<m,\\ M & \text{if}\quad \rho^{n+1}(\boldsymbol{x},t)>M.\end{cases}$$

LEMMA A.1. *Problem* (A.1) *has a unique solution.*

Proof. Let $R : H^1(\Omega) \to H^1(\Omega)$ be defined by $Rw = v$, where v is the weak solution of the elliptic problem:

$$\frac{v}{k} - \lambda\Delta v = -\boldsymbol{w}_h^n \cdot \nabla T_m^M w + \frac{1}{k}\rho^n \text{ in } \Omega, \quad \frac{\partial v}{\partial \boldsymbol{n}}\Big|_{\partial\Omega} = 0. \tag{A.2}$$

Notice that $T_m^M w$ is defined a.e. in Q. In addition, as $-\boldsymbol{w}_h^n \cdot \nabla T_m^M w + \frac{1}{k}\rho^n \in H^1(\Omega)'$, R is well-defined, that is, there exists a unique $v \in H^1(\Omega)$ such that $Rw = v$.

By using v as a test function in (A.2) and integrating by parts, we arrive at

$$\frac{1}{k}|v|^2 + \lambda|\nabla v|^2 \le |\nabla \cdot \boldsymbol{w}_h^n|\,\|T_m^M w\|_{L^\infty(\Omega)}|v| + |\boldsymbol{w}_h^n|\|T_m^M w\|_{L^\infty(\Omega)}|\nabla v| + \frac{1}{k}|\rho^n|\,|v|.$$

Since $\|T_m^M w\|_{L^\infty(\Omega)} = M$, then $\|v\|_{H^1(\Omega)} \le C$, with $C > 0$ a constant independent of w. Therefore, by taking any $r \ge C$, one has that if $\|w\|_{H^1(\Omega)} \le r$, then $\|Rw\|_{H^1(\Omega)} \le r$. To apply Schauder's fixed point theorem we have to prove that $R : H^1(\Omega) \to H^1(\Omega)$ is continuous and compact. For this, it suffices to demonstrate that

$$\text{if} \quad w_l \rightharpoonup w \ \text{ in } H^1(\Omega), \quad \text{then} \quad Rw_l \to Rw \text{ in } H^1(\Omega) \text{ as } l \to \infty.$$

Indeed, from $w_l \rightharpoonup w$ in $H^1(\Omega)$, it holds by compactness that $w_l \to w$ in $L^2(\Omega)$. Therefore, there exists a subsequence, that to simplify notation is denoted in the same way, such that $w_l \to w$ a.e. in Ω. By virtue of dominated convergence theorem, we have

$$\boldsymbol{w}_h^n \cdot \nabla T_m^M w_l + \frac{1}{k}\rho^n \longrightarrow \boldsymbol{w}_h^n \cdot \nabla T_m^M w + \frac{1}{k}\rho^n \text{ in } H^1(\Omega)' \text{ as } l \to \infty. \tag{A.3}$$

On the other hand, the mapping $h \in H^1(\Omega)' \to z \in H^1(\Omega)$, where z is the solution

$$\frac{1}{k}\Big(z, \bar{z}\Big) - \lambda\Big(\nabla z, \nabla \bar{z}\Big) = \Big\langle h, \bar{z}\Big\rangle_{H^1(\Omega)' H^1(\Omega)} \quad \forall \bar{z} \in H^1(\Omega),$$

is linear and continuous (by the Lax–Milgram theorem). Then from (A.3) we have $Rw_l \to Rw$ in $H^1(\Omega)$.

In conclusion, $R|_{\overline{B}(0,r)} : \overline{B}(0,r) \to \overline{B}(0,r)$ is continuous and compact. By Schauder's fixed point theorem, we have the desired result after a regularity result for elliptic equations. ∎

THEOREM A.2. *The solution of problem* (A.1) *verifies the maximum principle; i.e., if* $m \le \rho^0 \le M$, *then* $m \le \rho^n \le M$ *for each* $n \ge 1$.

Proof. Let us first see that if $\rho^n \le M$, then $\rho^{n+1} \le M$. By multiplying (A.1) by $(\rho^{n+1} - M)_+$ and integrating over Ω (we denote $f(x)_+ = \begin{cases} f(x) & \text{if } f(x) > 0, \\ 0 & \text{if } f(x) \le 0, \end{cases}$), this gives us

$$\begin{aligned}&\left(\frac{\rho^{n+1} - \rho^n}{k}, (\rho^{n+1} - M)_+\right) + \Big(\boldsymbol{w}_h^n \cdot \nabla T_m^M \rho^{n+1}, (\rho^{n+1} - M)_+\Big) \\ &+ \lambda(\nabla\rho^{n+1}, \nabla(\rho^{n+1} - M)_+) = 0.\end{aligned}$$

By using properties of the positive part function and the fact that $T_m^M \rho^{n+1} = M$ as $(\rho^{n+1} - M)_+ \ne 0$, we rewrite it as

$$\left(\frac{\rho^{n+1} - \rho^n}{k}, (\rho^{n+1} - M)_+\right) + \lambda\Big(\nabla(\rho^{n+1} - M)_+, \nabla(\rho^{n+1} - M)_+\Big) = 0. \tag{A.4}$$

Therefore, we deduce from (A.4) that $(\rho^{n+1}-\rho^n,(\rho^{n+1}-M)_+)\le 0$. By adding and subtracting M, one arrives at $((\rho^{n+1}-M)-(\rho^n-M),(\rho^{n+1}-M)_+)\le 0$. Again, by the properties of the positive part function, we obtain

$$\big((\rho^{n+1}-M),(\rho^{n+1}-M)_+\big)=\big((\rho^{n+1}-M)_+,(\rho^{n+1}-M)_+\big)=|(\rho^{n+1}-M)_+|^2.$$

By the induction hypothesis $\rho^n-M\le 0$, and then $(\rho^n-M,(\rho^{n+1}-M)_+)\le 0$. Consequently, $|(\rho^{n+1}-M)_+|^2\le(\rho^n-M,(\rho^{n+1}-M)_+)\le 0$, and then $\rho^{n+1}\le M$ holds.

The proof of the other bound—if $\rho^n\ge m$, then $\rho^{n+1}\ge m$—can be obtained in the same way. ∎

Corollary A.3. *Problem* (A.1) *is equivalent to problem* (3.9).

Proof. Since ρ^{n+1} the solution of problem (A.1) verifies the maximum principle $m\le\rho^{n+1}\le M$, then, in particular from the definition of truncating operator T_m^M, one has $T_m^M\rho^{n+1}=\rho^{n+1}$, and problem (A.1) is rewritten as problem (3.9). ∎

From the uniqueness of the solution of problem (3.9), we have, in particular, that problem (3.9) verifies the maximum principle

$$0<m\le\rho^n(\boldsymbol{x})\le M\quad\forall\boldsymbol{x}\in\Omega\quad\forall\, n.$$

Therefore, to finish the proof of Lemma 3.4, it remains to prove strong estimates (independent of λ,h,k) for the ρ^{n+1} solution of (3.9). For this, it will be fundamental to use the pointwise estimates $m\le\rho^{n+1}\le M$ in Q.

We define $\eta^{n+1}=\rho^{n+1}-\frac{1}{|\Omega|}\int_\Omega\rho^{n+1}$, and hence $\frac{1}{|\Omega|}\int_\Omega\eta^{n+1}=0$. Then the $\|\eta^{n+1}\|_{H^1}$-norm is equivalent to $|\nabla\eta^{n+1}|$, and the $\|\eta^{n+1}\|_{H^2}$-norm is equivalent to $|\Delta\eta^{n+1}|$. Since $m\le\frac{1}{|\Omega|}\int_\Omega\rho^{n+1}\le M$, the estimates for $\|\eta^{n+1}\|_{H^1}$ and $\|\eta^{n+1}\|_{H^2}$ imply estimates for $\|\rho^{n+1}\|_{H^1}$ and $\|\rho^{n+1}\|_{H^2}$, respectively.

With this definition of η^{n+1}, problem (3.9) can be rewritten as:

$$\frac{\eta^{n+1}-(\rho^n-\frac{1}{|\Omega|}\int_\Omega\rho^{n+1})}{k}+\boldsymbol{w}_h^n\cdot\nabla\eta^{n+1}-\lambda\Delta\eta^{n+1}=0\ \text{ in }\Omega,\quad\frac{\partial\eta^{n+1}}{\partial\boldsymbol{n}}\Big|_{\partial\Omega}=0.$$

By multiplying by $-2\,k\,\Delta\eta^{n+1}$, we arrive at

$$\begin{aligned}&|\nabla\eta^{n+1}|^2-\left|\nabla\left(\rho^n-\frac{1}{|\Omega|}\int_\Omega\rho^{n+1}\right)\right|^2+|\nabla(\rho^{n+1}-\rho^n)|^2+2\lambda\,k|\Delta\eta^{n+1}|^2\\&\le 2k\Big(\boldsymbol{w}_h^n\cdot\nabla\eta^{n+1},\Delta\eta^{n+1}\Big):=K_1.\end{aligned}$$

By integrating by parts in K_1,

$$K_1=-2k(\nabla\boldsymbol{w}_h^n,\nabla\eta^{n+1}\otimes\nabla\eta^{n+1})+k(\nabla\cdot\boldsymbol{w}_h^n,|\nabla\eta^{n+1}|^2).$$

By using $\|\nabla\eta^{n+1}\|_{L^4(\Omega)}\le C\,\|\eta^{n+1}\|_{L^\infty}^{1/2}|\Delta\eta^{n+1}|^{1/2}\le C\,|\Delta\eta^{n+1}|^{1/2}$, we get

$$K_1\le Ck\|\boldsymbol{w}_h^n\|\,\|\nabla\eta^{n+1}\|_{L^4(\Omega)}^2\le Ck\|\boldsymbol{w}_h^n\|\,|\Delta\eta^{n+1}|\le\frac{C_\varepsilon}{\lambda}\,k\,\|\boldsymbol{w}_h^n\|^2+\varepsilon\,\lambda\,k|\Delta\eta^{n+1}|^2.$$

Therefore,

$$|\nabla\eta^{n+1}|^2-\left|\nabla\left(\rho^n-\frac{1}{|\Omega|}\int_\Omega\rho^{n+1}\right)\right|^2+\lambda k|\Delta\eta^{n+1}|^2\le\frac{C}{\lambda}\,k\|\boldsymbol{w}_h^n\|^2.$$

By using the fact that $\nabla(\rho^n - \frac{1}{|\Omega|}\int \rho^{n+1}) = \nabla\rho^n = \nabla\eta^n$, we have

$$\lambda\|\eta^{n+1}\|_{H^1}^2 - \lambda\|\eta^n\|_{H^1}^2 + \lambda\|\rho^{n+1} - \rho^n\|_{H^1}^2 + \lambda^2 k\,\|\eta^{n+1}\|_{H^2}^2 \leq C\,k\|\boldsymbol{w}_h^n\|^2.$$

By summing over n and applying $k\sum_{n=1}^N \|\boldsymbol{w}_h^n\|^2 \leq C$, we get the following bounds:

$$\lambda \max_{0\leq n\leq N}\|\rho^n\|_{H^1(\Omega)}^2 \leq C, \quad \lambda^2\,k\sum_{n=1}^N \|\rho^n\|_{H^2(\Omega)}^2 \leq C,$$

where $C > 0$ is independent of λ, h, k. This concludes the proof of Lemma 3.4. ∎

REFERENCES

[1] S. N. Antontsev, A. V. Kazhikhov, and V. N. Monakhov, *Boundary Value Problems in Mechanics of Nonhomogeneous Fluids*, Stud. Math. Appl. 22, North-Holland, Amsterdam, 1990.

[2] H. Beirão da Veiga, *Diffusion on viscous fluids, existence and asymptotic properties of solutions*, Ann. Sc. Norm. Super. Pisa, 10 (1983), pp. 341–355.

[3] S. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, Texts Appl. Math. 15, Springer-Verlag, Berlin, 1994.

[4] D. Bresch, E. H. Essoufi, and M. Sy, *De nouveaux systèmes de type Kazhikhov-Smagulov: Modèles de propagation de polluants et de combustion à faible nombre de Mach*, C. R. Acad. Sci. Paris, Série I, 335 (2002), pp. 973–978.

[5] D. Bresch, E. H. Essoufi, and M. Sy, *Effects of density dependent viscosities on multiphasic incompressible fluid models*, J. Math. Fluid Mech., 9 (2007), pp. 377–397.

[6] J. Étienne and P. Saramito, *A priori error estimates of the Lagrange-Galerkin method for Kazhikhov-Smagulov type systems*, C. R. Math. Acad. Sci. Paris, 341 (2005), pp. 769–774.

[7] P. G. Ciarlet, *The Tinite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1987.

[8] V. Girault and P. A. Raviart, *Finite Element Methods for Navier-Stokes Equations: Theory and Algorithms*, Springer-Verlag, Berlin, 1986.

[9] F. Guillén-González, *Sobre un modelo asintótico de difusión de masa para fluidos incompresibles, viscoso y no homogéneos*, in Proceedings of the Third Catalan Days on Applied Mathematics, Lleida, Spain, 1996, pp. 103–114.

[10] F. Guillén-González, P. Damázio, and M. A. Rojas-Medar, *Approach of regular solutions for incompressible fluids with mass diffusion by an interative method*, J. Math. Anal. Appl., 326 (2007), pp. 468–487.

[11] F. Guillén-González, and J. V. Gutiérrez-Santacreu, *Unconditional stability and convergence of fully discrete schemes for 2D viscous fluids models with mass diffusion*, Math. Comp., to appear.

[12] A. Kazhikhov and Sh. Smagulov, *The correctness of boundary value problems in a diffusion model of an inhomogeneous fluid*, Sov. Phys. Dokl., 22 (1977), pp. 249–252.

[13] C. Liu and N. J. Walkington, *Convergence of numerical approximations of the incompressible Navier-Stokes equations with variable density and viscosity*, SIAM J. Numer. Anal., 45 (2007), pp. 1287–1304.

[14] R. Salvi, *On the existence of weak solutions of boundary-value problems in a diffusion model of an inhomogeneous liquid in regions with moving boundaries*, Port. Math., 43 (1986), pp. 213–233.

[15] P. Secchi, *On the motion of viscous fluids in the presense of diffusion*, SIAM J. Math. Anal., 19 (1988), pp. 22–31.

[16] P. Secchi, *On the inicial value problem for the equations of motion of viscous incompressible fluids in the presence of diffsion*, Boll. Unione Mat. Ital., 6 (1982), pp. 117–1130.

[17] J. Simon, *Compact sets in the Space $L^p(0,T;B)$*, Ann. Mat. Pura Appl., 146 (1987), pp. 65–97.

[18] R. Temam, *Navier-Stokes Equations. Theory and Numerical Analysis*, North-Holland, Amsterdam, 1977.

SIAM J. NUMER. ANAL.
Vol. 46, No. 5, pp. 2309–2345
© 2008 Society for Industrial and Applied Mathematics

A SPARSE GRID STOCHASTIC COLLOCATION METHOD FOR PARTIAL DIFFERENTIAL EQUATIONS WITH RANDOM INPUT DATA*

F. NOBILE†, R. TEMPONE‡, AND C. G. WEBSTER§

Abstract. This work proposes and analyzes a Smolyak-type sparse grid stochastic collocation method for the approximation of statistical quantities related to the solution of partial differential equations with random coefficients and forcing terms (input data of the model). To compute solution statistics, the sparse grid stochastic collocation method uses approximate solutions, produced here by finite elements, corresponding to a deterministic set of points in the random input space. This naturally requires solving uncoupled deterministic problems as in the Monte Carlo method. If the number of random variables needed to describe the input data is moderately large, full tensor product spaces are computationally expensive to use due to the *curse of dimensionality*. In this case the sparse grid approach is still expected to be competitive with the classical Monte Carlo method. Therefore, it is of major practical relevance to understand in which situations the sparse grid stochastic collocation method is more efficient than Monte Carlo. This work provides error estimates for the fully discrete solution using L^q norms and analyzes the computational efficiency of the proposed method. In particular, it demonstrates algebraic convergence with respect to the total number of collocation points and quantifies the effect of the dimension of the problem (number of input random variables) in the final estimates. The derived estimates are then used to compare the method with Monte Carlo, indicating for which problems the former is more efficient than the latter. Computational evidence complements the present theory and shows the effectiveness of the sparse grid stochastic collocation method compared to full tensor and Monte Carlo approaches.

Key words. collocation techniques, stochastic PDEs, finite elements, uncertainty quantification, sparse grids, Smolyak approximation, multivariate polynomial approximation

AMS subject classifications. 65N30, 65N35, 65N12, 65N15, 65C20

DOI. 10.1137/060663660

Introduction. Mathematical modeling and computer simulations are nowadays widely used tools to predict the behavior of physical and engineering problems. Whenever a particular application is considered, the mathematical models need to be equipped with input data, such as coefficients, forcing terms, boundary conditions, geometry, etc. However, in many applications, such input data may be affected by a relatively large amount of uncertainty. This can be due to an intrinsic variability

*Received by the editors June 26, 2006; accepted for publication (in revised form) February 4, 2008; published electronically May 16, 2008.
http://www.siam.org/journals/sinum/46-5/66366.html

†MOX, Dipartimento di Matematica "F. Brioschi," Politecnico di Milano, 20133 Milano, Italy (fabio.nobile@polimi.it). This author's research was partially supported by M.U.R.S.T. Cofin 2005 "Numerical Modeling for Scientific Computing and Advanced Applications" and SANDIA project 523695.

‡Department of Mathematics and School of Computational Science, Florida State University, 400 Dirac Science Library, Tallahassee, FL 32306-4120, and School of Computer Sciences and Communication, KTH, S–100 44 Stockholm, Sweden (rtempone@scs.fsu.edu). This author's research was partially supported by SANDIA project 523695, the Dahlquist fellowship at the Royal Institute of Technology in Stockholm, Sweden, and the UdelaR in Uruguay.

§Optimization and Uncertainty Quantification Department, Sandia National Laboratories, P.O. Box 5800, Albuquerque, NM 87185-1318 (cgwebst@sandia.gov). This author's research was supported by the School of Computational Science (SCS) at Florida State University and by the John von Neumann Fellowship at the Computer Science Research Institute (CSRI), Sandia National Laboratories. Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE AC04-94-AL85000.

in the physical system, as, for instance, in the mechanical properties of many biomaterials, polymeric fluids, and composite materials or the action of wind or seismic vibrations on civil structures.

In other situations, uncertainty may come from our difficulty in characterizing accurately the physical system under investigation, as in the study of groundwater flows, where the subsurface properties such as porosity and permeability in an aquifer have to be extrapolated from measurements taken only in a few spatial locations.

Such uncertainties can be included in the mathematical model adopting a probabilistic setting, provided enough information is available for a complete statistical characterization of the physical system. In this framework, the input data are modeled as *random variables* or, more generally, as *random fields* with a given spatial (or temporal) correlation structure.

Therefore, the goal of the mathematical and computational analysis becomes the prediction of statistical moments of the solution (mean value, variance, covariance, etc.) or statistics of some given responses of the system (sometimes also called quantities of physical interest which are real valued functionals of the solution), given the probability distribution of the input random data. Examples of quantities of interest could be the solution values in a given region, fluxes across given boundaries, etc.

In order to parametrize the input data for a given partial differential equation (PDE), random fields that are either coefficients or loads can often be expanded as an infinite combination of random variables by, for instance, the so-called Karhunen–Loève [23, 24] or polynomial chaos (PC) expansions [34, 38]. Although such random fields are properly described only by means of an infinite number of random variables, whenever the realizations are slowly varying in space, with a correlation length comparable to the size of the domain, only a few terms in the above-mentioned expansions are typically needed to describe the random field with sufficient accuracy. Therefore, in this case, it is reasonable to limit the analysis to just a few random variables in the expansion (see, e.g., [2, 16]).

In this work we focus on PDEs whose coefficients and forcing terms are described by a finite-dimensional random vector. This is often called *finite-dimensional noise assumption* and may hold either because the problem itself can be described by a finite number of random variables or because the input coefficients are modeled as truncated random fields.

The most popular approach to solving mathematical problems in a probabilistic setting is the Monte Carlo method (see, e.g., [15] and the references therein). The Monte Carlo method is easy to implement and allows one to reuse available deterministic codes. However, the convergence rate is typically very slow, although with a mild dependence on the number on sampled random variables.

In the last few years, other approaches have been proposed, which in certain situations feature a much faster convergence rate. We mention, among others, the spectral Galerkin method [20, 37, 4, 5, 16, 26, 22, 28], stochastic collocation [3, 30, 25, 36], and perturbation methods or Neumann expansions [17, 35, 1, 31].

For certain classes of problems, the solution may have a very regular dependence on the input random variables. For instance, it was shown in [3] and [4] that the solution of a linear elliptic PDE with diffusivity coefficient and/or forcing term described as truncated expansions of random fields is analytic in the input random variables. In such situations, spectral Galerkin or stochastic collocation methods based on orthogonal tensor product polynomials feature a very fast convergence rate.

In particular, our earlier work [3] proposed a stochastic collocation/finite element

method based on standard finite element approximations in space and a collocation on a tensor grid built upon the zeros of orthogonal polynomials with respect to the joint probability density function of the input random variables. It was shown that for an elliptic PDE the error converges exponentially fast with respect to the number of points employed for each random input variable.

The stochastic collocation method can be easily implemented and leads naturally to the solution of uncoupled deterministic problems as in the Monte Carlo method, even in the presence of input data which depend nonlinearly on the driving random variables. It can also treat efficiently the case of nonindependent random variables with the introduction of an auxiliary density and can handle, for instance, cases with log-normal diffusivity coefficient, which is not bounded in $\Omega \times D$ but has bounded realizations. When the number of input random variables is small, stochastic collocation is a very effective numerical tool.

On the other hand, approximations based on tensor product grids suffer from the *curse of dimensionality* since the number of collocation points in a tensor grid grows exponentially fast in the number of input random variables.

If the number of random variables is moderately large, one should rather consider sparse tensor product spaces as first proposed by Smolyak [29] and further investigated by, e.g., [18, 6, 16, 36], which will be the primary focus of this paper. It is natural to expect that the use of sparse grids will reduce dramatically the number of collocation points, while preserving a high level of accuracy and thus being able to successfully compete with Monte Carlo. Our main purpose is to clarify the limitations of the previous statement and to understand in which situations the sparse grid stochastic collocation method is more efficient than Monte Carlo.

Motivated by the above, this work proposes and analyzes a Smolyak-type sparse grid stochastic collocation method for the approximation of statistical quantities related to the solution of PDEs whose input data are described through a finite number of random variables. The sparse tensor product grids are built upon either Clenshaw–Curtis [11] or Gaussian abscissas. After outlining the method, this work provides strong error estimates for the fully discrete solution and analyzes its computational efficiency. In particular, it proves algebraic convergence with respect to the total number of collocation points or, equivalently, the total computational work which is directly proportional to the number of collocation points. The exponent of such algebraic convergence is connected to both the regularity of the solution and the number of input random variables, N, and essentially deteriorates with N by a $1/\log(N)$ factor. Then these error estimates are used to compare the method with the standard Monte Carlo method, indicating for which problems the former is more efficient than the latter.

Moreover, this work addresses the case where the input random variables come from suitably truncated expansions of random fields. There it discusses how to relate the number of points in the sparse grid to the number of random variables retained in the truncated expansion in order to balance discretization error with truncation error in the input random fields. Computational evidence complements the present theory and shows the effectiveness of the sparse grid stochastic collocation method. It also includes a comparison with full tensor and Monte Carlo methods.

The outline of the work is the following: Section 1 introduces the mathematical problem and basic notation and states a regularity assumption to be used later in the error analysis. Section 2 summarizes various collocation techniques and describes the sparse approximation method under study. It also describes two types of abscissas,

Clenshaw–Curtis and Gaussian, that will be employed in the sparse approximation method.

Section 3 is the core of the work. We first develop strong error estimates for the fully discrete solution using L_P^∞ and L_P^2 norms for Clenshaw–Curtis and Gaussian abscissas, respectively (P being the probability measure considered). These norms control the error in the approximation of expected values of smooth functionals of the solution. Then in section 3.2 these error estimates are used to compare the method with the standard Monte Carlo method, explaining cases where the former is more efficient than the latter.

Sections 4 and 5 focus on applications to linear elliptic PDEs with random input data. In section 4 we verify that the assumptions under which our general theory works hold in this particular case. Then we present in section 5 some numerical results showing the effectiveness of the proposed method when compared to the full tensor and Monte Carlo methods.

1. Problem setting. We begin by focusing our attention on a differential operator $\mathcal{L}$, linear or nonlinear, on a domain $D \subset \mathbb{R}^d$, which depends on some coefficients $a(\omega, x)$ with $x \in D$, $\omega \in \Omega$, where $(\Omega, \mathcal{F}, P)$ is a complete probability space. Here Ω is the set of outcomes, $\mathcal{F} \subset 2^\Omega$ is the σ-algebra of events, and $P : \mathcal{F} \to [0,1]$ is a probability measure. Similarly the forcing term $f = f(\omega, x)$ can be assumed random as well.

Consider the stochastic boundary value problem: find a random function, $u : \Omega \times \overline{D} \to \mathbb{R}$, such that P-a.e. in Ω, or in other words almost surely (a.s.), the following equation holds:

$$\mathcal{L}(a)(u) = f \quad \text{in } D \tag{1.1}$$

equipped with suitable boundary conditions. Before introducing some assumptions we denote by $W(D)$ a Banach space of functions $v : D \to \mathbb{R}$ and define, for $q \in [1, \infty]$, the stochastic Banach spaces

$$L_P^q(\Omega; W(D)) = \Big\{ v : \Omega \to W(D) \mid v \text{ is strongly measurable and } \int_\Omega \|v(\omega, \cdot)\|_{W(D)}^q dP(\omega) < +\infty \Big\}$$

and

$$L_P^\infty(\Omega; W(D)) = \Big\{ v : \Omega \to W(D) \mid v \text{ is strongly measurable and } P - \operatorname*{ess\,sup}_{\omega \in \Omega} \|v(\omega, \cdot)\|_{W(D)}^2 < +\infty \Big\}.$$

Of particular interest is the space $L_P^2(\Omega; W(D))$, consisting of Banach valued functions that have finite second moments.

We will now make the following assumptions:

(A_1) The solution to (1.1) has realizations in the Banach space $W(D)$, i.e., $u(\cdot, \omega) \in W(D)$ a.s. and $\forall \omega \in \Omega$

$$\|u(\cdot, \omega)\|_{W(D)} \le C \|f(\cdot, \omega)\|_{W^*(D)},$$

where we denote $W^*(D)$ to be the dual space of $W(D)$, and C is a constant independent of the realization $\omega \in \Omega$.

(A_2) The forcing term $f \in L^2_P(\Omega; W^*(D))$ is such that the solution u is unique and bounded in $L^2_P(\Omega; W(D))$.

Here we give two example problems that are posed in this setting.

Example 1.1. The linear problem

$$\begin{cases} -\nabla \cdot (a(\omega,\cdot)\nabla u(\omega,\cdot)) = f(\omega,\cdot) & \text{in } \Omega \times D, \\ u(\omega,\cdot) = 0 & \text{on } \Omega \times \partial D, \end{cases} \tag{1.2}$$

with $a(\omega,\cdot)$ uniformly bounded and coercive, i.e.,

$$\text{there exists } a_{min}, a_{max} \in (0,+\infty) \text{ such that}$$
$$P(\omega \in \Omega :\ a(\omega,x) \in [a_{min}, a_{max}] \quad \forall x \in \overline{D}) = 1$$

and $f(\omega,\cdot)$ square integrable with respect to P, satisfies assumptions A_1 and A_2 with $W(D) = H^1_0(D)$ (see [3]).

Example 1.2. Similarly, for $k \in \mathbb{N}^+$, the nonlinear problem

$$\begin{cases} -\nabla \cdot (a(\omega,\cdot)\nabla u(\omega,\cdot)) + u(\omega,\cdot)^{2k+1} = f(\omega,\cdot) & \text{in } \Omega \times D, \\ u(\omega,\cdot) = 0 & \text{on } \Omega \times \partial D, \end{cases} \tag{1.3}$$

with $a(\omega,\cdot)$ uniformly bounded and coercive and $f(\omega,\cdot)$ square integrable with respect to P, satisfies assumptions A_1 and A_2 with $W(D) = H^1_0(D) \cap L^{2k+2}(D)$.

Remark 1.3 (goals of the computation). As said in the introduction, the goal of the mathematical and computational analysis is the prediction of statistical moments of the solution u to (1.1) (mean value, variance, covariance, etc.) or statistics of some given quantities of physical interest $\psi(u)$. Examples of quantities of interest could be the average value of the solution in a given region $D_c \subset D$,

$$\psi(u) = \frac{1}{|D_c|} \int_{D_c} u dx,$$

and similarly average fluxes on a given direction $n \in \mathbb{R}^d$. In the case of Examples 1.1 and 1.2 these fluxes can be written as

$$\psi(u) = \frac{1}{|D_c|} \int_{D_c} a \frac{\partial u}{\partial n} dx.$$

1.1. On finite-dimensional noise. In some applications, the coefficient a and the forcing term f appearing in (1.1) can be described by a random vector $[Y_1, \ldots, Y_N] : \Omega \to \mathbb{R}^N$, as in the following examples. In such cases, we will emphasize such dependence by writing a_N and f_N.

Example 1.4 (piecewise constant random fields). Let us consider again problem (1.2), where the physical domain D is the union of nonoverlapping subdomains D_i, $i = 1, \ldots, N$. We consider a diffusion coefficient that is piecewise constant and random on each subdomain, i.e.,

$$a_N(\omega, x) = a_{min} + \sum_{i=1}^{N} \sigma_i\, Y_i(\omega) 1_{D_i}(x).$$

Here 1_{D_i} is the indicator function of the set D_i, σ_i, a_{min} are positive constants, and the random variables Y_i are nonnegative with unit variance.

In other applications the coefficients and forcing terms in (1.1) may have another type of spatial variation that is amenable to being described by an expansion. Depending on the decay of such an expansion and the desired accuracy in our computations we may retain just the first N terms.

Example 1.5 (Karhunen–Loève expansion). We recall that any second order random field $g(\omega, x)$, with continuous covariance function $\mathrm{cov}[g] : \overline{D \times D} \to \mathbb{R}$, can be represented as an infinite sum of random variables by means, for instance, of a Karhunen–Loève expansion [23, 24]. To this end, introduce the compact and self-adjoint operator $T_g : L^2(D) \to L^2(D)$, which is defined by

$$T_g v(\cdot) := \int_D \mathrm{cov}[g](x, \cdot)\, v(x)\, dx \quad \forall v \in L^2(D).$$

Then consider the sequence of nonnegative decreasing eigenvalues of T_g, $\{\lambda_i\}_{i=1}^{\infty}$, and the corresponding sequence of orthonormal eigenfunctions, $\{b_i\}_{i=1}^{\infty}$, satisfying

$$T_g b_i = \lambda_i b_i, \quad (b_i, b_j)_{L^2(D)} = \delta_{ij} \ \text{ for } \ i, j \in \mathbb{N}_+.$$

In addition, define mutually uncorrelated real random variables

$$Y_i(\omega) := \frac{1}{\sqrt{\lambda_i}} \int_D (g(\omega, x) - E[g](x))\, b_i(x) dx, \quad i = 1, \ldots,$$

with zero mean and unit variance, i.e., $E[Y_i] = 0$ and $E[Y_i Y_j] = \delta_{ij}$ for $i, j \in \mathbb{N}_+$. The truncated Karhunen–Loève expansion g_N, of the stochastic function g, is defined by

$$g_N(\omega, x) := E[g](x) + \sum_{i=1}^{N} \sqrt{\lambda_i}\, b_i(x)\, Y_i(\omega) \quad \forall N \in \mathbb{N}_+.$$

Then by Mercer's theorem (cf. [27, p. 245]), it follows that

$$\lim_{N \to \infty} \left\{ \sup_D E\left[(g - g_N)^2\right] \right\} = \lim_{N \to \infty} \left\{ \sup_D \left(\sum_{i=N+1}^{\infty} \lambda_i b_i^2 \right) \right\} = 0.$$

Observe that the N random variables in (1.5), describing the random data, are then weighted differently due to the decay of the eigenpairs of the Karhunen–Loève expansion. The decay of eigenvalues and eigenvectors has been investigated, e.g., in the works [16] and [31].

The above examples motivate us to consider problems whose coefficients are described by finitely many random variables. Thus, we will seek a random field $u_N : \Omega \times \overline{D} \to \mathbb{R}$, such that a.s. the following equation holds:

$$\mathcal{L}(a_N)(u_N) = f_N \quad \text{in } D. \tag{1.4}$$

We assume that (1.4) admits a unique solution $u_N \in L_P^2(\Omega; W(D))$. Therefore, following the same argument as in [3, p. 1010] yields that the solution u_N of the stochastic boundary value problem (1.4) can be described by the $[Y_1, \ldots, Y_N]$ random variables, i.e., $u_N = u_N(\omega, x) = u_N(Y_1(\omega), \ldots, Y_N(\omega), x)$.

We underline that the coefficients a_N and f_N in (1.4) may be an exact representation of the input data as in Example 1.4 or a suitable truncation of the input data

as in Example 1.5. In the latter case, the solution u_N will also be an approximation of the exact solution u in (1.1) and the truncation error $u - u_N$ has to be properly estimated (see section 3.2).

Remark 1.6 (nonlinear coefficients). In certain cases, one may need to ensure qualitative properties on the coefficients a_N and f_N, and it may be worthwhile describing them as nonlinear functions of Y. For instance, in Example 1.1 one is required to enforce positiveness on the coefficient $a_N(\omega, x)$, say $a_N(\omega, x) \geq a_{min}\ \forall x \in D$, a.s. in Ω. Then a better choice is to expand $\log(a_N - a_{min})$. The following standard transformation guarantees that the diffusivity coefficient is bounded away from zero a.s.:

$$\log(a_N - a_{min})(\omega, x) = b_0(x) + \sum_{1 \leq n \leq N} \sqrt{\lambda_n} b_n(x) Y_n(\omega), \tag{1.5}$$

i.e., one performs a Karhunen–Loève expansion for $\log(a_N - a_{min})$, assuming that $a_N > a_{min}$ a.s. On the other hand, the right-hand side of (1.4) can be represented as a truncated Karhunen–Loève expansion $f_N(\omega, x) = c_0(x) + \sum_{1 \leq n \leq N} \sqrt{\mu_n} c_n(x) Y_n(\omega)$.

For this work we denote by $\Gamma_n \equiv Y_n(\Omega)$ the image of Y_n, where we assume $Y_n(\omega)$ to be bounded. Without loss of generality we can assume $\Gamma_n = [-1, 1]$. We also let $\Gamma^N = \prod_{n=1}^N \Gamma_n$ and assume that the random variables $[Y_1, Y_2, \ldots, Y_N]$ have a joint probability density function

$$\rho : \Gamma^N \to \mathbb{R}_+, \quad \text{with} \quad \rho \in L^\infty(\Gamma^N). \tag{1.6}$$

Thus, the plan is to approximate the function $u_N = u_N(y, x)$ for any $y \in \Gamma^N$ and $x \in \overline{D}$ (see [3, 4]).

Remark 1.7 (unbounded random variables). By using a similar approach to the work [3] we can easily deal with unbounded random variables, such as Gaussian or exponential ones. For the sake of simplicity in the presentation we focus our study on bounded random variables only.

The convergence properties of the collocation techniques that will be developed in the next section depend on the regularity that the solution u_N has with respect to y. Denote $\Gamma_n^* = \prod_{\substack{j=1 \\ j \neq n}}^N \Gamma_j$, and let y_n^* be an arbitrary element of Γ_n^*. Here we require the solution to problem (1.1) to satisfy the following assumption.

Assumption 1.8 (regularity). For each $y_n \in \Gamma_n$, there exists $\tau_n > 0$ such that the function $u_N(y_n, y_n^*, x)$ as a function of y_n, $u_N : \Gamma_n \to C^0(\Gamma_n^*; W(D))$ admits an analytic extension $u(z, y_n^*, x)$, $z \in \mathbb{C}$, in the region of the complex plane

$$\Sigma(\Gamma_n; \tau_n) \equiv \{z \in \mathbb{C},\ \ \text{dist}(z, \Gamma_n) \leq \tau_n\}. \tag{1.7}$$

Moreover, $\forall z \in \Sigma(\Gamma_n; \tau_n)$,

$$\|u_N(z)\|_{C^0(\Gamma_n^*; W(D))} \leq \lambda \tag{1.8}$$

with λ a constant independent of n.

This assumption is sound in several problems; in particular, it can be verified for the linear problem that will be analyzed in section 4. In the more general case, this assumption should be verified for each particular application and will have implications on the allowed regularity of the input data, e.g., coefficients, loads, etc., of the stochastic PDE under study. See also Remark 3.14 for related results based on fewer regularity requirements.

2. Collocation techniques. We seek a numerical approximation to the exact solution of (1.4) in a suitable finite-dimensional subspace. To describe such a subspace properly, we introduce some standard approximation subspaces:

- $W_h(D) \subset W(D)$ is a standard finite element space of dimension N_h, which contains continuous piecewise polynomials defined on regular triangulations $\mathcal{T}_h$ that have a maximum mesh-spacing parameter $h > 0$. We suppose that W_h has the following deterministic approximation property: for a given function $\varphi \in W(D)$,

$$\min_{v \in W_h(D)} \|\varphi - v\|_{W(D)} \leq C(s;\varphi)\, h^s, \tag{2.1}$$

where s is a positive integer determined by the smoothness of φ and the degree of the approximating finite element subspace, and $C(s;\varphi)$ is independent of h.

Example 2.1. Let D be a convex polygonal domain and $W(D) = H_0^1(D)$. For piecewise linear finite element subspaces we have

$$\min_{v \in W_h(D)} \|\varphi - v\|_{H_0^1(D)} \leq c\, h\, \|\varphi\|_{H^2(D)}.$$

That is, $s = 1$ and $C(s;\varphi) = c\|\varphi\|_{H^2(D)}$; see, for example, [7].

We will also assume that there exists a finite element operator $\pi_h : W(D) \to W_h(D)$ with the optimality condition

$$\|\varphi - \pi_h\varphi\|_{W(D)} \leq C_\pi \min_{v \in W_h(D)} \|\varphi - v\|_{W(D)} \quad \forall \varphi \in W(D), \tag{2.2}$$

where the constant C_π is independent of the mesh size h. It is worth noticing that in general the operator π_h will depend on the specific problem, as well as on y, i.e., $\pi_h = \pi_h(y)$.

- $\mathcal{P}_{\mathbf{p}}(\Gamma^N) \subset L^2_\rho(\Gamma^N)$ is the span of tensor product polynomials with degree at most $\mathbf{p} = (p_1, \ldots, p_N)$, i.e., $\mathcal{P}_{\mathbf{p}}(\Gamma^N) = \bigotimes_{n=1}^N \mathcal{P}_{p_n}(\Gamma_n)$, with

$$\mathcal{P}_{p_n}(\Gamma_n) = \operatorname{span}(y_n^k,\, k = 0, \ldots, p_n), \quad n = 1, \ldots, N.$$

Hence the dimension of $\mathcal{P}_{\mathbf{p}}(\Gamma^N)$ is $N_p = \prod_{n=1}^N (p_n + 1)$.

Stochastic collocation entails the evaluation of approximate values $\pi_h u_N(y_k) = u_h^N(y_k) \in W_h(D)$ to the solution u_N of (1.4) on a suitable set of points $y_k \in \Gamma^N$. Then the fully discrete solution $u_{h,\mathbf{p}}^N \in C^0(\Gamma^N; W_h(D))$ is a global approximation (sometimes an interpolation) constructed by linear combinations of the point values. That is,

$$u_{h,\mathbf{p}}^N(y,\cdot) = \sum_{k \in \mathcal{K}} u_h^N(y_k,\cdot) l_k^{\mathbf{p}}(y), \tag{2.3}$$

where, for instance, the functions $l_k^{\mathbf{p}}$ can be taken as the Lagrange polynomials (see sections 2.1 and 2.2). This formulation can be used to compute the mean value or variance of u, as described in [3, section 2], or to approximate expected values of functionals $\psi(u)$ (cf. Remark 1.3) by

$$E[\psi(u)] \approx E[\psi(u_{h,\mathbf{p}}^N)] \approx \sum_{k \in \mathcal{K}} \psi(u_h^N(y_k))\, E[l_k^{\mathbf{p}}].$$

In the next sections we consider different choices of the evaluation points y_k and corresponding weights $E[l_k^{\mathbf{p}}]$ in the associated quadrature formula.

2.1. Full tensor product interpolation. In this section we briefly recall interpolation based on Lagrange polynomials. We first introduce an index $i \in \mathbb{N}_+$, $i \geq 1$. Then, for each value of i, let $\{y_1^i, \ldots, y_{m_i}^i\} \subset [-1,1]$ be a sequence of abscissas for Lagrange interpolation on $[-1,1]$.

For $u \in C^0(\Gamma^1; W(D))$ and $N = 1$ we introduce a sequence of one-dimensional Lagrange interpolation operators $\mathscr{U}^i : C^0(\Gamma^1; W(D)) \to V_{m_i}(\Gamma^1; W(D))$,

$$\mathscr{U}^i(u)(y) = \sum_{j=1}^{m_i} u(y_j^i)\, l_j^i(y) \quad \forall u \in C^0(\Gamma^1; W(D)), \tag{2.4}$$

where $l_j^i \in \mathcal{P}_{m_i-1}(\Gamma^1)$ are the Lagrange polynomials of degree $m_i - 1$, i.e., $l_j^i(y) = \prod_{\substack{k=1 \\ k\neq j}}^{m_i} \frac{(y-y_k^i)}{(y_j^i - y_k^i)}$, and

$$V_m(\Gamma^1; W(D)) = \left\{ v \in C^0(\Gamma^1; W(D)) \,:\, v(y,x) = \sum_{k=1}^{m} \widetilde{v}_k(x) l_k(y),\ \{\widetilde{v}_k\}_{k=1}^m \in W(D) \right\}.$$

Formula (2.4) reproduces exactly all polynomials of degree less than m_i. Now, in the multivariate case $N > 1$, for each $u \in C^0(\Gamma^N; W(D))$ and the multi-index $\mathbf{i} = (i_1, \ldots, i_N) \in \mathbb{N}_+^N$ we define the full tensor product interpolation formulas

$$\mathcal{I}_{\mathbf{i}}^N u(y) = \left(\mathscr{U}^{i_1} \otimes \cdots \otimes \mathscr{U}^{i_N}\right)(u)(y) = \sum_{j_1=1}^{m_{i_1}} \cdots \sum_{j_N=1}^{m_{i_N}} u\left(y_{j_1}^{i_1}, \ldots, y_{j_N}^{i_N}\right) \left(l_{j_1}^{i_1} \otimes \cdots \otimes l_{j_N}^{i_N}\right). \tag{2.5}$$

Clearly, the above product needs $\prod_{n=1}^N m_{i_n}$ function evaluations. These formulas will also be used as the building blocks for the Smolyak method, described next.

2.2. Smolyak approximation. Here we follow closely the work [6] and describe the Smolyak *isotropic* formulas $\mathscr{A}(\mathrm{w}, N)$. The Smolyak formulas are just linear combinations of product formulas (2.5) with the following key properties: only products with a relatively small number of points are used. With $\mathscr{U}^0 = 0$ and for $i \in \mathbb{N}_+$ define

$$\Delta^i := \mathscr{U}^i - \mathscr{U}^{i-1}. \tag{2.6}$$

Moreover, given an integer $\mathrm{w} \in \mathbb{N}_+$, hereafter called the *level*, we define the sets

$$X(\mathrm{w}, N) := \left\{ \mathbf{i} \in \mathbb{N}_+^N, \mathbf{i} \geq \mathbf{1} \,:\, \sum_{n=1}^N (i_n - 1) \leq \mathrm{w} \right\}, \tag{2.7a}$$

$$\widetilde{X}(\mathrm{w}, N) := \left\{ \mathbf{i} \in \mathbb{N}_+^N, \mathbf{i} \geq \mathbf{1} \,:\, \sum_{n=1}^N (i_n - 1) = \mathrm{w} \right\}, \tag{2.7b}$$

$$Y(\mathrm{w}, N) := \left\{ \mathbf{i} \in \mathbb{N}_+^N, \mathbf{i} \geq \mathbf{1} \,:\, \mathrm{w} - N + 1 \leq \sum_{n=1}^N (i_n - 1) \leq \mathrm{w} \right\}, \tag{2.7c}$$

and for $\mathbf{i} \in \mathbb{N}_+^N$ we set $|\mathbf{i}| = i_1 + \cdots + i_N$. Then the isotropic Smolyak formula is given by

$$\mathscr{A}(\mathrm{w}, N) = \sum_{\mathbf{i} \in X(\mathrm{w}, N)} \left(\Delta^{i_1} \otimes \cdots \otimes \Delta^{i_N}\right). \tag{2.8}$$

Equivalently, formula (2.8) can be written as (see [33])

$$\mathscr{A}(\mathrm{w},N)=\sum_{\mathbf{i}\in Y(\mathrm{w},N)}(-1)^{\mathrm{w}+N-|\mathbf{i}|}\binom{N-1}{\mathrm{w}+N-|\mathbf{i}|}\cdot\left(\mathscr{U}^{i_1}\otimes\cdots\otimes\mathscr{U}^{i_N}\right). \tag{2.9}$$

To compute $\mathscr{A}(\mathrm{w},N)(u)$, one only needs to know function values on the "sparse grid"

$$\mathscr{H}(\mathrm{w},N)=\bigcup_{\mathbf{i}\in Y(\mathrm{w},N)}\left(\vartheta^{i_1}\times\cdots\times\vartheta^{i_N}\right)\subset[-1,1]^N, \tag{2.10}$$

where $\vartheta^i=\left\{y_1^i,\ldots,y_{m_i}^i\right\}\subset[-1,1]$ denotes the set of abscissas used by $\mathscr{U}^i$. If the sets are nested, i.e., $\vartheta^i\subset\vartheta^{i+1}$, then $\mathscr{H}(\mathrm{w},N)\subset\mathscr{H}(\mathrm{w}+1,N)$ and

$$\mathscr{H}(\mathrm{w},N)=\bigcup_{\mathbf{i}\in\widetilde{X}(\mathrm{w},N)}\left(\vartheta^{i_1}\times\cdots\times\vartheta^{i_N}\right). \tag{2.11}$$

The Smolyak formula is actually interpolatory whenever nested points are used. This result has been proved in [6, Proposition 6, p. 277].

By comparing (2.11) and (2.10), we observe that the Smolyak approximation that employs nested points requires fewer function evaluations than the corresponding formula with nonnested points. In the next section we introduce two particular sets of abscissas, nested and nonnested, respectively. Also, Figure 1 shows, as an example, the sparse grid $\mathscr{H}(5,2)$ obtained in those two cases. Note that the Smolyak approximation formula, as presented in this section, is isotropic, since all directions are treated equally. This can be seen from (2.8) observing that if a multi-index $\mathbf{i}=(i_1,i_2,\ldots,i_N)$ belongs to the set $X(\mathrm{w},N)$, then any permutation of $\mathbf{i}$ also belongs to $X(\mathrm{w},N)$ and contributes to the construction of the Smolyak approximation $\mathscr{A}(\mathrm{w},N)$.

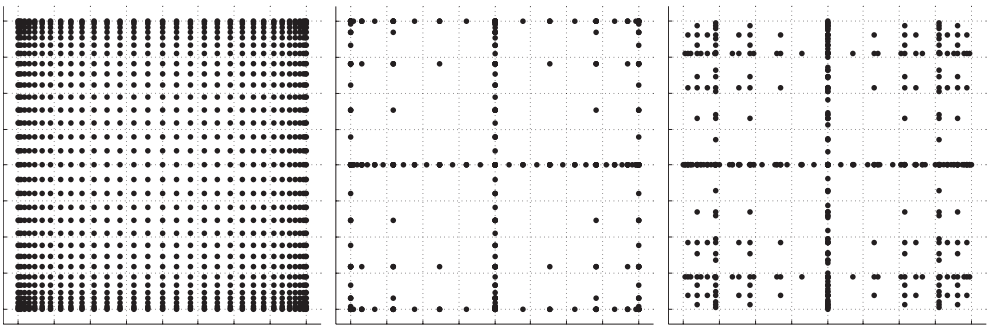


FIG. 1. *For a two-dimensional parameter space ($N=2$) and maximum level* $\mathrm{w}=5$*, we plot the full tensor product grid using the Clenshaw–Curtis abscissas (left) and isotropic Smolyak sparse grids $\mathscr{H}(5,2)$, utilizing the Clenshaw–Curtis abscissas (middle) and the Gaussian abscissas (right).*

2.3. Choice of interpolation abscissas.

Clenshaw–Curtis abscissas. We first suggest using Clenshaw–Curtis abscissas (see [11]) in the construction of the Smolyak formula. These abscissas are the extrema of Chebyshev polynomials and, for any choice of $m_i>1$, are given by

$$y_j^i=-\cos\left(\frac{\pi(j-1)}{m_i-1}\right),\quad j=1,\ldots,m_i. \tag{2.12}$$

In addition, one sets $y_1^i = 0$ if $m_i = 1$ and lets the number of abscissas m_i in each level grow according to the following formula:

$$m_1 = 1 \quad \text{and} \quad m_i = 2^{i-1} + 1 \text{ for } i > 1. \tag{2.13}$$

With this particular choice, one obtains nested sets of abscissas, i.e., $\vartheta^i \subset \vartheta^{i+1}$ and thereby $\mathscr{H}(\mathrm{w}, N) \subset \mathscr{H}(\mathrm{w}+1, N)$. It is important to choose $m_1 = 1$ if we are interested in optimal approximation in relatively large N, because in all other cases the number of points used by $\mathscr{A}(\mathrm{w}, N)$ increases too fast with N.

Gaussian abscissas. We also propose using Gaussian abscissas, i.e., the zeros of the orthogonal polynomials with respect to some positive weight. However, these Gaussian abscissas are in general not nested. Regardless, as in the Clenshaw–Curtis case, we choose the number m_i of abscissas that are used by $\mathscr{U}^i$ as in (2.13). See the work [32] for an insightful comparison of quadrature formulas based on Clenshaw–Curtis and Gaussian abscissas. The natural choice of the weight should be the probability density function ρ of the random variables $Y_i(\omega)$, $i = 1, \ldots, N$. Yet, in the general multivariate case, if the random variables Y_i are not independent, the density ρ does not factorize, i.e., $\rho(y_1, \ldots, y_n) \neq \prod_{n=1}^N \rho_n(y_n)$. To this end, we first introduce an auxiliary probability density function $\hat{\rho} : \Gamma^N \to \mathbb{R}^+$ that can be seen as the joint probability of N independent random variables, i.e., it factorizes as

$$\hat{\rho}(y_1, \ldots, y_n) = \prod_{n=1}^N \hat{\rho}_n(y_n) \ \ \forall y \in \Gamma^N \qquad \text{and is such that} \quad \left\| \frac{\rho}{\hat{\rho}} \right\|_{L^\infty(\Gamma^N)} < \infty. \tag{2.14}$$

For each dimension $n = 1, \ldots, N$, let the m_n Gaussian abscissas be the roots of the m_n degree polynomial that is $\hat{\rho}_n$-orthogonal to all polynomials of degree $m_n - 1$ on the interval $[-1, 1]$. The auxiliary density $\hat{\rho}$ should be chosen as close as possible to the true density ρ, so as to have the quotient $\rho/\hat{\rho}$ not too large. Indeed, such a quotient will appear in the final error estimate (see section 3.1.2).

Examples of isotropic sparse grids, constructed from the nested Clenshaw–Curtis abscissas and the nonnested Gaussian abscissas, are shown in Figure 1. There, we consider a two-dimensional parameter space and a maximum level $\mathrm{w} = 5$ (sparse grid $\mathscr{H}(5, 2)$). To see the reduction in function evaluations with respect to full tensor product grids, we also include a plot of the corresponding Clenshaw–Curtis isotropic full tensor grid having the same maximum number of points in each direction, namely, $2^{\mathrm{w}}+1 = 33$. Observe that if we take m points in each direction, the isotropic full tensor grid will contain m^N points, while the analogous isotropic Smolyak grid $\mathscr{H}(\mathrm{w}, N)$ will contain much fewer points. Figure 2 shows the total number of points contained in the full tensor grid and in the Smolyak sparse grid as a function of the level w (or the corresponding maximum number m of points in each direction) for dimensions $N = 5, 11, 21$.

3. Error analysis. In this section we develop error estimates that will help us compare the efficiency of the isotropic Smolyak approximation with other alternatives, for instance, the Monte Carlo method as explained in section 3.2. Much about this has been claimed in the existing literature based on particular numerical examples. Our main goal is therefore to understand in which situations the sparse grid stochastic collocation method is more efficient than Monte Carlo.

As explained in section 2, it is possible to use collocation methods to approximate the solution $u_N \in C^0(\Gamma^N; W(D))$ using finitely many function values, each of them

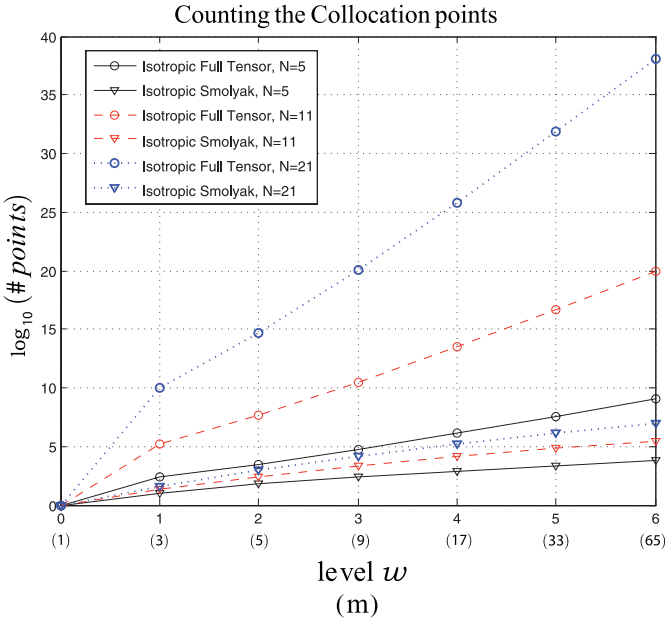


FIG. 2. *For a finite-dimensional* Γ^N *with* $N = 5, 11,$ *and* 21 *we plot the* log *of the number of distinct Clenshaw–Curtis collocation points used by the isotropic Smolyak method and the corresponding isotropic full tensor product method versus the level* w *(or the maximum number of points* m *employed in each direction).*

computed by finite elements. Recall that by Assumption 1.8, u_N admits an analytic extension. Let the fully discrete numerical approximation be $\mathscr{A}(\mathrm{w}, N)\pi_h u_N$. The aim of this section is to give a priori estimates for the total error

$$e = u - \mathscr{A}(\mathrm{w}, N)\pi_h u_N,$$

where the operator $\mathscr{A}(\mathrm{w}, N)$ is described by (2.8) and π_h is the finite element projection operator described by (2.2). We will investigate the error

$$\|u - \mathscr{A}(\mathrm{w}, N)\pi_h u_N\| \le \underbrace{\|u - u_N\|}_{(I)} + \underbrace{\|u_N - \pi_h u_N\|}_{(II)} + \underbrace{\|\pi_h u_N - \mathscr{A}(\mathrm{w}, N)\pi_h u_N\|}_{(III)} \tag{3.1}$$

evaluated in the norm $L^q_P(\Omega; W(D))$ with either $q = 2$ or $q = \infty$. This yields also control of the error in the expected value of u, $\|E[e]\|_{W(D)} \le E\big[\,\|e\|_{W(D)}\,\big] \le \|e\|_{L^q_P(\Omega;W(D))}$, and the error in the approximation of $E[\psi(u)]$, with ψ being a smooth functional of u. In such a case we have

$$|E[\psi(u) - \psi(\mathscr{A}(\mathrm{w}, N)\pi_h u_N)]| \le \left(\int_0^1 \|\delta_e \psi(u + \theta e)\|_{L^{q^*}_P(\Omega;W^*(D))} d\theta\right) \|e\|_{L^q_P(\Omega;W(D))},$$

with $1/q + 1/q^* = 1$ and $\delta_e \psi(u + \theta e)$ denoting the Fréchet derivative of ψ at $u + \theta e$. Although in the previous inequalities it is suboptimal to estimate expectations of Sobolev norms rather than a Sobolev norm of expectations, we will content ourselves in this work with providing estimates for $\|e\|_{L^q_P(\Omega;W(D))}$. Sharper estimates could be obtained generalizing the ideas given in [4, Theorem 5.1].

The quantity (I) controls the truncation error for the case where the input data a_N and f_N are suitable truncations of random fields. This contribution to the total error will be considered in section 3.2. The quantity (I) is otherwise zero if the representation of a_N and f_N is exact, as in Example 1.4. The second term, (II), controls the convergence with respect to h, i.e., the finite element error, which will be dictated by standard approximability properties of the finite element space $W_h(D)$, given by (2.1), and the regularity in space of the solution u (see, e.g., [10, 7]). For example, if we let $q = 2$, we have

$$\|u_N - \pi_h u_N\|_{L^2_\rho(\Gamma^N;W(D))} \le h^s \left(\int_{\Gamma^N} (C_\pi(y) C(s; u(y)))^2 \rho(y)\, dy \right)^{1/2}.$$

The full tensor product convergence results are given by [3, Theorem 1], and therefore we will concern ourselves only with the convergence results when implementing the Smolyak approximation formula described in section 2.2. Namely, our primary concern will be to analyze the Smolyak approximation error

$$(III) = \|\pi_h u_N - \mathscr{A}(\mathrm{w}, N)\pi_h u_N\|_{L^q_\rho(\Gamma^N;W(D))}$$

for both the Clenshaw–Curtis and Gaussian versions of the Smolyak formula.

Under the very reasonable assumption that the semidiscrete finite element solution $\pi_h u_N$ admits an analytic extension as described in Assumption 1.8 with the same analyticity region as for u_N, the behavior of the error (III) will be analogous to $\|u_N - \mathscr{A}(\mathrm{w}, N)u_N\|_{L^q_\rho(\Gamma^N;W(D))}$. For this reason in the next sections we will analyze the latter.

Remark 3.1 (on the randomness of approximate and exact solutions). It is important to emphasize that the exact solution u and its numerical approximation $\bar{u} := \mathscr{A}(\mathrm{w}, N)\pi_h u_N$ do not depend on the same randomness. Therefore, when considering both solutions we should in principle work with a product of probability spaces. For instance, assume now that $\mu_{u,\bar{u}}$ is a joint distribution for u and $\bar{u}$. It is key to recall that we are interested only in the approximation of statistics of u, which depend only on the marginal distribution μ_u, by the corresponding statistics of $\bar{u}$, which depend only on the marginal distribution $\mu_{\bar{u}}$. Let us introduce an auxiliary approximate solution, $\hat{u}$, with the same law as $\bar{u}$ but different joint distribution, $\mu_{u,\hat{u}}$. We have

$$\begin{aligned} |E^{\mu_u}[\psi(u)] - E^{\mu_{\bar{u}}}[\psi(\bar{u})]| &= |E^{\mu_u}[\psi(u)] - E^{\mu_{\hat{u}}}[\psi(\hat{u})]| \\ &= |E^{\mu_{u,\hat{u}}}[\psi(u) - \psi(\hat{u})]| \\ &\le E^{\mu_{u,\hat{u}}}[(\psi(u) - \psi(\hat{u}))^2]^{1/2}. \end{aligned}$$

Observe now that for the purpose of estimating the difference $|E^{\mu_u}[\psi(u)] - E^{\mu_{\bar{u}}}[\psi(\bar{u})]|$ the choice of the auxiliary joint distribution $\mu_{u,\hat{u}}$ is arbitrary. To this end, for the remainder of the work we select the joint distribution $\mu_{u,\hat{u}}$ such that u and $\hat{u}$ are driven by the same random variables Y. We will also avoid making any distinction in the notation between $\hat{u}$ and $\bar{u}$.

3.1. Analysis of the approximation error. In this work the technique to develop error bounds for multidimensional Smolyak approximation is based on one-dimensional results. Therefore, we first address the case $N = 1$. Let us recall the best approximation error for a function $v : \Gamma^1 \to W(D)$ which admits an analytic extension in the region $\Sigma(\Gamma^1; \tau) = \{z \in \mathbb{C},\ \mathrm{dist}(z, \Gamma^1) < \tau\}$ of the complex plane for

some $\tau > 0$. We will still denote the extension by v; in this case, τ is smaller than the distance between $\Gamma^1 \subset \mathbb{R}$ and the nearest singularity of $v(z)$ in the complex plane. Since we are considering only the case of bounded random variables, we recall the following result, whose proof can be found in [3, Lemma 7] and which is an immediate extension of the result given in [12, Chapter 7, section 8].

LEMMA 3.2. *Given a function $v \in C^0(\Gamma^1; W(D))$ which admits an analytic extension in the region of the complex plane $\Sigma(\Gamma^1; \tau) = \{z \in \mathbb{C}, dist(z, \Gamma^1) \leq \tau\}$ for some $\tau > 0$, there holds*

$$E_{m_i} \equiv \min_{w \in V_{m_i}} \|v - w\|_{C^0(\Gamma^1; W(D))} \leq \frac{2}{e^{\hat{\sigma}} - 1} e^{-\hat{\sigma}\, m_i} \max_{z \in \Sigma(\Gamma^1; \tau)} \|v(z)\|_{W(D)}, \tag{3.2}$$

where $0 < \hat{\sigma} = \log\left(\frac{2\tau}{|\Gamma^1|} + \sqrt{1 + \frac{4\tau^2}{|\Gamma^1|^2}}\right)$.

Remark 3.3 (approximation with unbounded random variables). A related result with weighted norms holds for unbounded random variables whose probability density decays as the Gaussian density at infinity (see [3]).

In the multidimensional case, the size of the analyticity region, τ_n (see, e.g., the problem considered in section 4), will depend, in general, on the direction n and on the $(N-1)$ values of the remainder y components, denoted by y_n^* in Assumption 1.8. The same holds for the decay coefficient $\hat{\sigma}_n$. In what follows, we set

$$\hat{\sigma} \equiv \min_{1 \leq n \leq N} \min_{y_n^* \in \Gamma_n^*} \hat{\sigma}_n \tag{3.3}$$

and assume that $\hat{\sigma}$ is strictly positive. Similarly to that, the norm $\max_{z \in \Sigma(\Gamma^1; \tau)} \|v(z)\|_{W(D)}$ appearing in Lemma 3.2 also depends on the direction n and the remainder vector y_n^* and should be replaced, in the multidimensional case, by

$$M(v) = \max_{1 \leq n \leq N} \max_{y_n^* \in \Gamma_n^*} \max_{z \in \Sigma(\Gamma_n; \tau_n)} \|v(z)\|_{W(D)},$$

which will be assumed bounded hereafter. Both assumptions on $\hat{\sigma}$ and $M(v)$ have to be verified for each problem under consideration. In particular, they are satisfied for the application described in section 4.

3.1.1. Interpolation estimates for the Clenshaw–Curtis abscissas. In this section we develop L^∞ error estimates for the Smolyak interpolant based on Clenshaw–Curtis abscissas (cf. (2.12) and (2.13)) applied to analytic functions $u \in C^0(\Gamma^N; W(D))$ that satisfy Assumption 1.8. We remind the reader that even though in the global estimate (3.1) it is enough to bound the approximation error (III) in the $L^2_\rho(\Gamma^N; W(D))$ norm, we will still work with the more stringent $L^\infty(\Gamma^N; W(D))$ norm.

In our notation the norm $\|\cdot\|_{\infty,N}$ is shorthand for $\|\cdot\|_{L^\infty(\Gamma^N; W(D))}$ and will be used henceforth. We also define $I_N : \Gamma^N \to \Gamma^N$ as the identity operator on an N-dimensional space. We begin by letting E_m be the error of the best approximation to functions $u \in C^0(\Gamma^1; W(D))$ by functions $w \in V_m$. Similarly to [6], since the interpolation $\mathscr{U}^i$ is exact on the subspace $V_{m_i - 1}$ we can apply the general formula

$$\left\|u - \mathscr{U}^i(u)\right\|_{\infty,1} \leq E_{m_i - 1}(u) \cdot (1 + \Lambda_{m_i}), \tag{3.4}$$

where Λ_m is the Lebesgue constant for our choice (2.12). It is known that

$$\Lambda_m \leq \frac{2}{\pi} \log(m - 1) + 1 \tag{3.5}$$

for $m \geq 2$; see [13]. Using Lemma 3.2, the best approximation $E_{m_i}(u)$ to functions $u \in C^0(\Gamma^1; W(D))$ that admit an analytic extension as described by Assumption 1.8 is bounded by (3.2). Moreover, for the choice of m_i given in (2.13), it is easy to show that $(1 + \Lambda_{m_i}) \leq 2i \ \forall i > 0$. Hence (3.2)–(3.4) implies

$$\begin{aligned}\left\|(I_1 - \mathscr{U}^i)(u)\right\|_{\infty,1} &\leq C\, i e^{-\sigma 2^i} M(u),\\ \left\|\Delta^i(u)\right\|_{\infty,1} &= \left\|(\mathscr{U}^i - \mathscr{U}^{i-1})(u)\right\|_{\infty,1}\\ &\leq \left\|(I_1 - \mathscr{U}^i)(u)\right\|_{\infty,1} + \left\|(I_1 - \mathscr{U}^{i-1})(u)\right\|_{\infty,1} \leq 2C\, i e^{-\sigma 2^{i-1}} M(u)\end{aligned}$$

$\forall i \in \mathbb{N}_+$, with positive constants

$$C = \frac{4}{e^{2\sigma} - 1} \quad \text{and} \quad \sigma = \frac{\hat{\sigma}}{2} \quad (\text{with } \hat{\sigma} \text{ as in Lemma 3.2}).$$

Since the value $M(u)$ affects the error estimates as a multiplicative constant, from now on we assume it to be one without loss of generality.

The convergence proof will be split into several steps, the main results being given in Theorems 3.7 and 3.10, which state the convergence rates in terms of the level w and the total number of collocation points, respectively. We denote by I_d the identity operator applicable to functions which depend on the first d variables $y_1, \ldots, y_d$. Then the following result holds.

LEMMA 3.4. *For functions $u \in C^0(\Gamma^N; W(D))$ satisfying the assumption of Lemma* 3.2 *the isotropic Smolyak formula* (2.8) *based on Clenshaw–Curtis abscissas satisfies*

$$\left\|\left(I_N - \mathscr{A}(\mathrm{w}, N)\right)(u)\right\|_{\infty,N} \leq \sum_{d=1}^{N} R(\mathrm{w}, d) \tag{3.6}$$

with

$$R(\mathrm{w}, d) := \frac{1}{2} \sum_{\mathbf{i} \in \tilde{X}(\mathrm{w},d)} (2C)^d \left(\prod_{n=1}^{d} i_n\right) e^{-\sigma h(\mathbf{i},d)} \tag{3.7}$$

and

$$h(\mathbf{i}, d) = \sum_{n=1}^{d} 2^{i_n - 1}. \tag{3.8}$$

Proof. We start by providing an equivalent representation of the isotropic Smolyak formula:

$$\begin{aligned}\mathscr{A}(\mathrm{w}, N) &= \sum_{\mathbf{i} \in X(\mathrm{w},N)} \bigotimes_{n=1}^{N} \Delta^{i_n}\\ &= \sum_{\mathbf{i} \in X(\mathrm{w},N-1)} \bigotimes_{n=1}^{N-1} \Delta^{i_n} \otimes \sum_{j=1}^{1+\mathrm{w}-\sum_{n=1}^{N-1}(i_n-1)} \Delta^{j}\\ &= \sum_{\mathbf{i} \in X(\mathrm{w},N-1)} \bigotimes_{n=1}^{N-1} \Delta^{i_n} \otimes \mathscr{U}^{1+\mathrm{w}-\sum_{n=1}^{N-1}(i_n-1)}.\end{aligned}$$

Introducing the one-dimensional identity operator $I_1^{(n)} : \Gamma_n \to \Gamma_n$ for $n = 1, \ldots, N$, we can compute the error estimate recursively using the previous representation, namely,

$$
\begin{aligned}
I_N - \mathscr{A}(\mathrm{w}, N) &= I_N - \sum_{\mathbf{i} \in X(\mathrm{w}, N-1)} \bigotimes_{n=1}^{N-1} \Delta^{i_n} \otimes \left(\mathscr{U}^{1+\mathrm{w}-\sum_{n=1}^{N-1}(i_n-1)} - I_1^{(N)} \right) \\
&\quad - \sum_{\mathbf{i} \in X(\mathrm{w}, N-1)} \bigotimes_{n=1}^{N-1} \Delta^{i_n} \otimes I_1^{(N)} \\
&= \sum_{\mathbf{i} \in X(\mathrm{w}, N-1)} \bigotimes_{n=1}^{N-1} \Delta^{i_n} \otimes \left(I_1^{(N)} - \mathscr{U}^{1+\mathrm{w}-\sum_{n=1}^{N-1}(i_n-1)} \right) \\
&\quad + (I_{N-1} - \mathscr{A}(\mathrm{w}, N-1)) \otimes I_1^{(N)} \\
&= \sum_{d=2}^{N} \left[\tilde{R}(\mathrm{w}, d) \bigotimes_{n=d+1}^{N} I_1^{(n)} \right] + \left(I_1^{(1)} - \mathscr{A}(\mathrm{w}, 1) \right) \bigotimes_{n=2}^{N} I_1^{(n)},
\end{aligned}
\tag{3.9}
$$

where, for a general dimension d, we define

$$
\tilde{R}(\mathrm{w}, d) = \sum_{\mathbf{i} \in X(\mathrm{w}, d-1)} \bigotimes_{n=1}^{d-1} \Delta^{i_n} \otimes \left(I_1^{(d)} - \mathscr{U}^{\hat{i}_d} \right)
$$

and, for any $(i_1, \ldots, i_{d-1}) \in X(\mathrm{w}, d-1)$, we have set $\hat{i}_d = 1 + \mathrm{w} - \sum_{n=1}^{d-1}(i_n - 1)$. Observe that with this definition, the d-dimensional vector $\mathbf{j} = (i_1, \ldots, i_{d-1}, \hat{i}_d)$ belongs to the set $\tilde{X}(\mathrm{w}, d)$, defined in (2.7), and the term $\tilde{R}(\mathrm{w}, d)$ can now be bounded as follows:

$$
\begin{aligned}
\left\| \tilde{R}(\mathrm{w}, d)(u) \right\|_{\infty,d} &\le \sum_{\mathbf{i} \in X(\mathrm{w}, d-1)} \prod_{n=1}^{d-1} \left\| (\Delta^{i_n})(u) \right\|_{\infty,d} \left\| \left(I_1^{(d)} - \mathscr{U}^{\hat{i}_d} \right)(u) \right\|_{\infty,d} \\
&\le \frac{1}{2} \sum_{\mathbf{i} \in X(\mathrm{w}, d-1)} (2C)^d \left(\prod_{n=1}^{d-1} i_n \right) \hat{i}_d \, e^{-\sigma\left(\sum_{n=1}^{d-1} 2^{i_n - 1} + 2^{\hat{i}_d}\right)} \\
&\le \frac{(2C)^d}{2} \sum_{\mathbf{i} \in \tilde{X}(\mathrm{w}, d)} \left(\prod_{n=1}^{d} i_n \right) e^{-\sigma\, h(\mathbf{i}, d)} =: R(\mathrm{w}, d).
\end{aligned}
$$

Hence, the Smolyak approximation error satisfies

$$
\left\| (I_N - \mathscr{A}(\mathrm{w}, N))(u) \right\|_{\infty,N} \le \sum_{d=2}^{N} R(\mathrm{w}, d) + \left\| (I_1^{(1)} - \mathscr{A}(\mathrm{w}, 1))(u) \right\|_{\infty,1}.
$$

Observe that the last term in the previous equation can also be bounded by $R(\mathrm{w}, 1)$

defined in (3.7). Indeed, the set $\tilde{X}(\mathrm{w},1)$ contains only the point $i_1 = 1+\mathrm{w}$ and

$$\begin{aligned}\left\|\left(I_1^{(1)} - \mathscr{A}(\mathrm{w},1)\right)(u)\right\|_{\infty,1} &= \left\|\left(I_1^{(1)} - \mathscr{U}^{1+\mathrm{w}}\right)(u)\right\|_{\infty,1} \\ &\le C\,(1+\mathrm{w})e^{-\sigma\,2^{1+\mathrm{w}}} \\ &\le \sum_{i_1\in\tilde{X}(\mathrm{w},1)} C\,i_1\,e^{-\sigma\,2^{i_1-1}} =: R(\mathrm{w},1),\end{aligned}$$

and this concludes the proof. ∎

LEMMA 3.5. *Let $\delta > 0$. Under the assumptions of Lemma 3.4 the following bound holds for the term $R(\mathrm{w},d)$, $d = 1,\dots,N$:*

$$R(\mathrm{w},d) \le \frac{C_1(\sigma,\delta)^d}{2}\,\exp\left(-\sigma d\left(2^{\mathrm{w}/d} - \delta\tilde{C}_2(\sigma)\mathrm{w}\right)\right), \tag{3.10}$$

where

$$\tilde{C}_2(\sigma) := 1 + \frac{1}{\log(2)}\sqrt{\frac{\pi}{2\sigma}} \tag{3.11}$$

and

$$C_1(\sigma,\delta) := \frac{4C}{e\delta\sigma}\exp\left(\delta\sigma\left\{\frac{1}{\sigma\log^2(2)} + \frac{1}{\log(2)\sqrt{2\sigma}} + 2\left(1+\frac{1}{\log(2)}\sqrt{\frac{\pi}{2\sigma}}\right)\right\}\right). \tag{3.12}$$

Proof. The proof is divided into several steps.

1. Expand the function $h(\mathbf{i},d)$ up to second order around the point $\mathbf{i}^* = (1+\mathrm{w}/d,\dots,1+\mathrm{w}/d)$ on the subspace $\{\mathbf{x}\in\mathbb{R}^d :, |\mathbf{x}-\mathbf{1}| = \mathrm{w}\}$. Observe that $\mathbf{i}^*$ is a constrained minimizer of $h(\mathbf{i},d)$ and
$$h(\mathbf{i},d) \ge d2^{\mathrm{w}/d} + \frac{\log^2(2)}{2}\sum_{n=1}^{d}(i_n - (1+\mathrm{w}/d))^2 \quad \forall\mathbf{i}\in\tilde{X}(\mathrm{w},d). \tag{3.13}$$
2. Combining (3.7) and (3.13), estimate
$$\begin{aligned}R(\mathrm{w},d) &\le \frac{(2C)^d}{2}e^{-\sigma d2^{\mathrm{w}/d}}\sum_{\mathbf{i}\in\tilde{X}(\mathrm{w},d)}\left(\prod_{n=1}^{d} i_n\right)\,e^{-\sigma\,\frac{\log^2(2)}{2}\sum_{n=1}^{d}(i_n-(1+\mathrm{w}/d))^2} \\ &\le \frac{(2C)^d}{2}e^{-\sigma d2^{\mathrm{w}/d}}\left(\sum_{i=1}^{\mathrm{w}+1} i\,e^{-\sigma\,\frac{\log^2(2)}{2}(i-(1+\mathrm{w}/d))^2}\right)^d.\end{aligned} \tag{3.14}$$
3. Next, use (A.2) from Corollary A.4 of the appendix to estimate the term $T_1 := \sum_{i=1}^{\mathrm{w}+1} i\,e^{-\sigma\,\frac{\log^2(2)}{2}(i-(1+\mathrm{w}/d))^2}$. We have
$$T_1 \le 2\left(\frac{1}{\sigma\log^2(2)} + \frac{1}{\log(2)\sqrt{2\sigma}}\right) + 2(\mathrm{int}\,\{\mathrm{w}/d\}+2)\left(1+\frac{1}{\log(2)}\sqrt{\frac{\pi}{2\sigma}}\right), \tag{3.15}$$
where $\mathrm{int}\,\{\cdot\}$ denotes the integer part of a real number. Combine (3.14) and (3.15), arriving at
$$R(\mathrm{w},d) \le C_1(\sigma,d,\mathrm{w})e^{-\sigma d\,2^{\mathrm{w}/d}} \tag{3.16}$$

with

$$C_1(\sigma,d,\mathrm{w}) \le \frac{1}{2}\,(4C)^d \left\{ \left(\frac{1}{\sigma \log^2(2)} + \frac{1}{\log(2)\sqrt{2\sigma}} \right) + (\mathrm{int}\,\{\mathrm{w}/d\}+2)\,\left(1+\frac{1}{\log(2)}\sqrt{\frac{\pi}{2\sigma}}\right)\right\}^d.$$

Now let $\delta > 0$ and use the inequality $x+1 \le e^x$, $x \ge 0$, to bound

$$\begin{aligned} & C_1(\sigma,d,\mathrm{w}) \\ & \le \frac{1}{2}\left(\frac{4C}{\delta\sigma}\right)^d \exp\left(d\delta\sigma \left\{ \left(\frac{1}{\sigma \log^2(2)} + \frac{1}{\log(2)\sqrt{2\sigma}} \right) \right.\right. \\ & \qquad \left.\left. + (\mathrm{int}\,\{\mathrm{w}/d\}+2)\,\left(1+\frac{1}{\log(2)}\sqrt{\frac{\pi}{2\sigma}}\right)\right\} - d \right) \\ & \le \frac{C_1(\sigma,\delta)^d}{2} \exp\left(\delta\sigma \left(1+\frac{1}{\log(2)}\sqrt{\frac{\pi}{2\sigma}}\right)\mathrm{w}\right), \end{aligned} \tag{3.17}$$

with $C_1(\sigma,\delta) := \frac{4C}{e\delta\sigma} \exp\left(\delta\sigma\left\{ \frac{1}{\sigma\log^2(2)} + \frac{1}{\log(2)\sqrt{2\sigma}} + 2\left(1+\frac{1}{\log(2)}\sqrt{\frac{\pi}{2\sigma}}\right)\right\}\right)$ defined as in (3.12).

Estimate (3.10) follows from (3.16) and (3.17). The proof is now complete. ∎

Remark 3.6 (alternative estimate). Observe that an alternative upper bound for T_1 in (3.15) is

$$T_1 \le \exp\left(\sigma\,\log^2(2)\left(1+\frac{\mathrm{w}}{d}\right)\right)\,\frac{(2+\mathrm{w})^2}{2}, \tag{3.18}$$

which remains bounded as $\sigma \to 0$. This does not happen with the bound $C_1(\sigma,d)$ (cf. (3.12)), which blows up as $\sigma \to 0$. As an implication of (3.18), we have

$$R(\mathrm{w},d) \le \frac{(C(2+\mathrm{w})^2)^d}{2}\, e^{-\sigma d(2^{\mathrm{w}/d} - \frac{\mathrm{w}}{d}\log^2(2))},$$

which is an alternative to the estimate (3.10) that has an extra polynomial growth in w but remains bounded for small values of σ.

THEOREM 3.7. *For functions $u \in C^0(\Gamma^N; W(D))$ satisfying the assumption of Lemma* 3.2 *the isotropic Smolyak formula* (2.8) *based on Clenshaw–Curtis abscissas satisfies*

$$\begin{aligned} \|(I_N - \mathscr{A}(\mathrm{w},N))\,(u)\|_{\infty,N} \le & \inf_{\delta\in(0,\frac{\chi}{\sqrt{\pi}})} \hat{C}(\sigma,\delta,N) \\ & \times \begin{cases} e^{-\sigma\mathrm{w}(e\log(2)-\delta\tilde{C}_2(\sigma))} & \text{if } 0 \le \mathrm{w} \le \dfrac{N}{\log(2)}, \\ e^{-\sigma\mathrm{w}\left(\frac{N}{\mathrm{w}}2^{\mathrm{w}/N}-\delta\tilde{C}_2(\sigma)\right)} & \text{otherwise,} \end{cases} \end{aligned} \tag{3.19}$$

where function $\hat{C}(\sigma,\delta,N) = \frac{C_1(\sigma,\delta)}{2}\,\frac{1-C_1(\sigma,\delta)^N}{1-C_1(\sigma,\delta)}$. The values of $\tilde{C}_2(\sigma)$ and $C_1(\sigma,\delta)$ have been defined in (3.11) *and* (3.12)*, respectively.*

Proof. From Lemmas 3.4 and 3.5 we obtain the following bound for the approximation error:

$$\|(I_N - \mathscr{A}(\mathrm{w},N))\,(u)\|_{\infty,N} \le \frac{1}{2}\sum_{d=1}^{N} C_1(\sigma,\delta)^d\, e^{-\sigma d(2^{\mathrm{w}/d}-\frac{\mathrm{w}}{d}\delta\tilde{C}_2(\sigma))},$$

with $C_1(\sigma,\delta)$ defined in (3.12). Then

$$\begin{aligned}\left\|\left(I_N-\mathscr{A}(\mathrm{w},N)\right)(u)\right\|_{\infty,N} &\le \frac{1}{2}\max_{1\le d\le N} e^{-\sigma \mathrm{w}\left(\frac{d}{\mathrm{w}}2^{\mathrm{w}/d}-\delta\tilde{C}_2(\sigma)\right)}\sum_{d=1}^{N} C_1(\sigma,\delta)^d \\ &\le \hat{C}(\sigma,\delta,N)e^{\sigma \mathrm{w}\delta\tilde{C}_2(\sigma)}\max_{1\le d\le N} e^{-\sigma \mathrm{w}\left(\frac{d}{\mathrm{w}}2^{\mathrm{w}/d}\right)},\end{aligned}$$

with

$$\begin{aligned}\hat{C}(\sigma,\delta,N) &:= \frac{1}{2}\sum_{d=1}^{N} C_1(\sigma,\delta)^d \\ &= \frac{C_1(\sigma,\delta)}{2}\,\frac{1-C_1(\sigma,\delta)^N}{1-C_1(\sigma,\delta)}.\end{aligned} \tag{3.20}$$

To finish the proof we further bound

$$\begin{aligned}\left\|\left(I_N-\mathscr{A}(\mathrm{w},N)\right)(u)\right\|_{\infty,N} &\le \hat{C}(\sigma,\delta,N)\, e^{\sigma \mathrm{w}\delta\tilde{C}_2(\sigma)}\, e^{-\sigma \mathrm{w}\left(\min_{1\le d\le N}\frac{d}{\mathrm{w}}2^{\mathrm{w}/d}\right)} \\ &\le \hat{C}(\sigma,\delta,N)\, e^{\sigma \mathrm{w}\delta\tilde{C}_2(\sigma)}\, e^{-\sigma \mathrm{w}\left(\min_{s\in[\mathrm{w}/N,\mathrm{w}]}\frac{1}{s}2^{s}\right)}\end{aligned}$$

and observe that

$$\min_{s\in[w/N,w]}\frac{1}{s}2^s = \begin{cases} e\log(2) & \text{if } 0\le \mathrm{w}\le \dfrac{N}{\log(2)}, \\ \dfrac{N}{\mathrm{w}}2^{\mathrm{w}/N} & \text{otherwise.}\end{cases} \qquad \square$$

Remark 3.8 (alternative estimate). Following Remark 3.6 we have an alternative to (3.19) in the estimate

$$\begin{aligned}\left\|\left(I_N-\mathscr{A}(\mathrm{w},N)\right)(u)\right\|_{\infty,N} &\le \frac{C(2+\mathrm{w})^2}{2}\,\frac{(C(2+\mathrm{w})^2)^N-1}{C(2+\mathrm{w})^2-1} \\ &\quad\times\begin{cases} e^{-\sigma \mathrm{w}\chi} & \text{if } 0\le \mathrm{w}\le \dfrac{N}{\log(2)}, \\ e^{-\sigma \mathrm{w}\left(\frac{N}{\mathrm{w}}2^{\mathrm{w}/N}-\log^2(2)\right)} & \text{otherwise.}\end{cases}\end{aligned} \tag{3.21}$$

Here we used the notation $\chi=\log(2)\,(e-\log(2))\approx 1.4037$. The previous estimate can be used to produce estimates like those in Theorems 3.10 and 3.11. The alternative estimates have constants which do not blow up as $\sigma\to 0$ but have the drawback of exhibiting additional multiplicative powers of $\log(\eta)$, η being the total number of collocation points. A completely identical discussion applies to the estimates based on Gaussian abscissas (see section 3.1.2) and will not be repeated there.

Now we relate the number of collocation points, $\eta=\eta(\mathrm{w},N)=\#\mathscr{H}(\mathrm{w},N)$, to the level w of the isotropic Smolyak formula. We state the result in the following lemma.

LEMMA 3.9. *Using the isotropic Smolyak interpolant described by* (2.8) *with Clenshaw–Curtis abscissas, the total number of points required at level* w *satisfies the following bounds:*

$$N(2^{\mathrm{w}}-1)\le \eta \le (2eN)^{\mathrm{w}}\min\{\mathrm{w}+1,2eN\}. \tag{3.22}$$

Moreover, as a direct consequence of (3.22) *we get that*

$$\frac{\log(\eta)}{1+\log(2)+\log(N)} - 1 \leq \mathrm{w}. \tag{3.23}$$

Proof. By using formula (2.8) and exploiting the nested structure of the Clenshaw–Curtis abscissas, the number of points $\eta = \eta(\mathrm{w}, N) = \#\mathscr{H}(\mathrm{w}, N)$ can be counted in the following way:

$$\eta = \sum_{\mathbf{i}\in X(\mathrm{w},N)} \prod_{n=1}^{N} r(i_n), \text{ where } r(i) := \begin{cases} 1 & \text{if } i = 1, \\ 2 & \text{if } i = 2, \\ 2^{i-2} & \text{if } i > 2. \end{cases} \tag{3.24}$$

Now notice that $\forall n = 1, 2, \ldots, N$ the following bound holds:

$$2^{i_n-2} \leq r(i_n) \leq 2^{i_n-1}. \tag{3.25}$$

We now produce a lower bound and an upper bound for η.

A lower bound on the number η of points can be obtained considering only the contribution from certain tensor grids. Indeed, for a fixed value of $\tilde{\mathrm{w}} = 1, \ldots, \mathrm{w}$, let us consider the N grids with indices $i_n = 1$ for $n \neq m$ and $i_m = \tilde{\mathrm{w}} + 1$, $m = 1, \ldots, N$. Since each of those N grids has $2^{\tilde{\mathrm{w}}-1}$ points, we have

$$\eta \geq \sum_{\tilde{\mathrm{w}}=1}^{\mathrm{w}} N\, 2^{\tilde{\mathrm{w}}-1} = N(2^{\mathrm{w}} - 1).$$

On the other hand, to produce an upper bound for η, we recall that $|\mathbf{i} - \mathbf{1}| = \sum_{n=1}^{N}(i_n - 1) \leq \mathrm{w}$, so the following bounds hold:

$$\begin{aligned}
\eta &= \sum_{\mathbf{i}\in X(\mathrm{w},N)} \prod_{n=1}^{N} r(i_n) \leq \sum_{\mathbf{i}\in X(\mathrm{w},N)} 2^{|\mathbf{i}-\mathbf{1}|} \leq \sum_{j=0}^{\mathrm{w}} \sum_{|\mathbf{i}-\mathbf{1}|=j} 2^j \\
&\leq \sum_{j=0}^{\mathrm{w}} 2^j \binom{N-1+j}{N-1} \leq \sum_{j=0}^{\mathrm{w}} 2^j \prod_{s=1}^{N-1} \left(1 + \frac{j}{s}\right) \\
&\leq \sum_{j=0}^{\mathrm{w}} 2^j \exp\left(\sum_{s=1}^{N-1} \frac{j}{s}\right) \leq \sum_{j=0}^{\mathrm{w}} 2^j \exp\left(j(1+\log(N))\right) \\
&\leq \sum_{j=0}^{\mathrm{w}} (2eN)^j \leq \min\{(\mathrm{w}+1)(2eN)^{\mathrm{w}}, (2eN)^{\mathrm{w}+1}\},
\end{aligned}$$

and this finishes the proof. ∎

The next theorem provides an error bound in terms of the total number η of collocation points. The proof follows directly from the results in Theorem 3.7 (taking $\delta = (e\log(2) - 1)/\tilde{C}_2(\sigma)$) and Lemma 3.9; it is therefore omitted.

THEOREM 3.10 (algebraic convergence). *For functions $u \in C^0(\Gamma^N; W(D))$ satisfying the assumption of Lemma* 3.2 *the isotropic Smolyak formula* (2.8) *based on Clenshaw–Curtis abscissas satisfies*

$$\begin{gathered}
\left\| \left(I_N - \mathscr{A}(\mathrm{w}, N)\right)(u) \right\|_{\infty,N} \leq \frac{C_1(\sigma, \delta^*) e^{\sigma}}{|1 - C_1(\sigma, \delta^*)|} \max\{1, C_1(\sigma, \delta^*)\}^N \eta^{-\mu_1}, \\
\text{with } \mu_1 = \frac{\sigma}{1 + \log(2N)}.
\end{gathered} \tag{3.26}$$

Here $\delta^* = (e\log(2)-1)/\tilde{C}_2(\sigma)$ *and the constants* $\tilde{C}_2(\sigma)$ *and* $C_1(\sigma,\delta^*)$*, defined in* (3.11) *and* (3.12)*, respectively, do not depend on* η*.*

Observe that the previous result indicates at least *algebraic convergence* with respect to the number of collocation points η. Under the same assumptions as those of the previous theorem and with a completely similar derivation, for large values of w we have the following sharper estimate.

THEOREM 3.11 (subexponential convergence). *Under the same assumptions as those of Theorem* 3.10 *and for* $\mathrm{w} > \frac{N}{\log(2)}$ *it holds that*

$$\left\|\left(I_N - \mathscr{A}(\mathrm{w},N)\right)(u)\right\|_{\infty,N} \leq \frac{C_1(\sigma,\delta^*)}{e^{\sigma\delta^*\tilde{C}_2(\sigma)}} \frac{\max\{1,C_1(\sigma,\delta^*)\}^N}{|1-C_1(\sigma,\delta^*)|}\ \eta^{\mu_3}\, e^{-\frac{N\sigma}{2^{1/N}}\eta^{\mu_2}}, \tag{3.27}$$

$$\text{with} \qquad \mu_2 = \frac{\log(2)}{N(1+\log(2N))} \qquad \text{and} \qquad \mu_3 = \frac{\sigma\delta^*\tilde{C}_2(\sigma)}{1+\log(2N)},$$

and with constant $C_1(\sigma,\delta^*)$ *defined in* (3.12) *and independent of* η*.*

Proof. We start from the result stated in Theorem 3.7 and observe that for $\mathrm{w} > N/\log(2)$ the function

$$g(\mathrm{w}) = \sigma(N2^{\mathrm{w}/N} - \mathrm{w}\delta\tilde{C}_2(\sigma))$$

is increasing in w for all values of $\delta < \log(2)e/\tilde{C}_2(\sigma)$. Hence, combining (3.19) with the lower bound (3.23) we obtain the desired result. ∎

The previous theorem indicates at least asymptotic *subexponential convergence* with respect to the number of collocation points η. It should be pointed out, however, that the large values of $w > N/\log(2)$ under which the bound holds are seldom used in practical computations. Therefore, from this point of view estimate (3.27) is less useful than (3.26).

Remark 3.12 (deterioration of the estimates with respect to the dimension N). Depending on the distance to the singularities of the solution, related to the parameter τ introduced in Lemma 3.2, the constant $C_1(\sigma,\delta^*)$ may be less than 1. In such a case the only dependence of the error bounds for $\left\|\left(I_N - \mathscr{A}(\mathrm{w},N)\right)(u)\right\|_{\infty,N}$ is in the exponent, whose denominator slowly grows like $\log(2N)$.

Remark 3.13 (full tensor versus Smolyak). An isotropic full tensor product interpolation converges roughly like $C(\sigma,N)\exp(-\sigma p)$, where p is the order of the polynomial space. Since the number of collocation points relates to p in this case as $\eta = (1+p)^N$ then $\log(\eta) = N\log(1+p) \leq Np$ and with respect to η the convergence rate can be bounded as $C(\sigma,N)\eta^{-\sigma/N}$. The slowdown effect that the dimension N has on the last convergence is known as the *curse of dimensionality*, and it is the reason for not using isotropic full tensor interpolation for large values of N. On the other hand, the isotropic Smolyak approximation seems to be better suited for this case. Indeed, from the estimate (3.26) we see that the Smolyak algebraic convergence has the faster exponent $\mathcal{O}(\frac{\sigma}{\log(2N)})$. This is a clear advantage of the isotropic Smolyak method with respect to the full tensor and justifies our claim that the use of Smolyak approximation greatly reduces the curse of dimensionality. In section 5 numerical results will give computational ground to this claim.

Remark 3.14 (estimates based on bounded mixed derivatives). We can proceed in a similar way to analyze the approximation error for functions that have a bounded mixed derivative of order $(k,\ldots,k)$. In that case, the one-dimensional best approximation error is $\|u-\mathcal{U}^i(u)\| \leq Cm_i^{-k}(1+\Lambda_{m_i})$, with C depending on u and k but not

on m_i, and using again the recursion (3.9) yields

$$\|(I_N - \mathscr{A}(\mathrm{w}, N))(u)\|_{\infty,N} \leq \frac{C}{|C(1+2^k)-1|}(C(1+2^k))^N(\mathrm{w}+1)^{2N}2^{-k\mathrm{w}}. \tag{3.28}$$

Finally, the combination of (3.28) with the counting estimates in Lemma 3.9 yields

$$\|(I_N - \mathscr{A}(\mathrm{w}, N))(u)\|_{\infty,N} \leq \frac{(C(1+2^k))^N}{|1+2^k-1/C|}\left(1+\log_2\left(1+\frac{\eta}{N}\right)\right)^{2N} \min\left\{2^k\eta^{-\frac{k\log(2)}{1+\log(2N)}}, \eta^{-k}\left(1+\log_2\left(1+\frac{\eta}{N}\right)\right)^{Nk}\right\}.$$

This estimate improves the one derived in [6]. Analogous results can be derived for Gaussian abscissas and L^2 norms.

3.1.2. Approximation estimates for Gaussian abscissas. Similarly to the previous section, we now develop error estimates for Smolyak approximation using Gaussian abscissas (cf. section 2.3) of $C^0(\Gamma^N; W(D))$ analytic functions described by Assumption 1.8. As before, we remind the reader that in the global estimate (3.1) we need to bound the approximation error (III) in the norm $L^2_\rho(\Gamma^N; W(D))$. Yet, the Gaussian abscissas defined in section 2.3 are constructed for the auxiliary density $\hat{\rho} = \prod_{n=1}^N \hat{\rho}_n$, still yielding control of the desired norm

$$\|v\|_{L^2_\rho(\Gamma^N;W(D))} \leq \left\|\frac{\rho}{\hat{\rho}}\right\|^{1/2}_{L^\infty(\Gamma^N)} \|v\|_{L^2_{\hat{\rho}}(\Gamma^N;W(D))} \quad \forall v \in C^0(\Gamma^N; W(D)).$$

In what follows we will use the shorthand notation $\|\cdot\|_{\hat{\rho},N}$ for $\|\cdot\|_{L^2_{\hat{\rho}}(\Gamma^N;W(D))}$ and restrict out analysis to the case where $W(D)$ is a Hilbert space. We now quote a useful result from Erdös and Turán [14].

LEMMA 3.15. *For every function $u \in C^0(\Gamma^1; W(D))$ the interpolation error with Lagrange polynomials based on Gaussian abscissas satisfies*

$$\|u - \mathscr{U}^i(u)\|_{\hat{\rho},1} \leq 2\sqrt{\int_{\Gamma^1}\hat{\rho}(y)\,dy}\inf_{v\in V_{m_i}}\|u-v\|_{\infty,1}. \tag{3.29}$$

Similarly to section 3.1.1, the combination of (3.2) with (3.29) yields

$$\left\|(I_1 - \mathscr{U}^i)(u)\right\|_{\hat{\rho},1} \leq \widetilde{C}\,e^{-\sigma 2^i}M(u), \quad \text{with } \tilde{C} = \frac{8}{e^{2\sigma}-1}\sqrt{\int_{\Gamma^1}\hat{\rho}(y)\,dy},$$

$$\begin{aligned}\left\|\Delta^i(u)\right\|_{\hat{\rho},1} &= \left\|(\mathscr{U}^i - \mathscr{U}^{i-1})(u)\right\|_{\hat{\rho},1}\\ &\leq \left\|(I_1-\mathscr{U}^i)(u)\right\|_{\hat{\rho},1} + \left\|(I_1-\mathscr{U}^{i-1})(u)\right\|_{\hat{\rho},1}\\ &\leq 2\widetilde{C}\,e^{-\sigma 2^{i-1}}M(u)\end{aligned}$$

$\forall i \in \mathbb{N}_+$. As in section 3.1.1 we assume the value $M(u)$ to be one without loss of generality. We then present the following lemma and theorem whose proofs follow, with minor changes, those given in Lemma 3.5 and Theorem 3.10, respectively. For instance, we apply (A.1) from Corollary A.4 to bound the corresponding T_1 sum in the estimate of $R(\mathrm{w}, d)$.

LEMMA 3.16. *For functions* $u \in C^0(\Gamma^N; W(D))$ *satisfying the assumption of Lemma* 3.2 *the isotropic Smolyak formula* (2.8) *based on Gaussian abscissas satisfies*

$$\|(I_N - \mathscr{A}(\mathrm{w}, N))(u)\|_{\rho,N} \le \sqrt{\|\rho/\hat{\rho}\|_{L^\infty(\Gamma^N)}}\frac{\tilde{C}_1(\sigma)}{2}\,\frac{1-\tilde{C}_1(\sigma)^N}{1-\tilde{C}_1(\sigma)} \times \begin{cases} e^{-\sigma e \log(2)\,\mathrm{w}} & \text{if } 0 \le \mathrm{w} \le \dfrac{N}{\log(2)}, \\ e^{-\sigma N 2^{\mathrm{w}/N}} & \text{otherwise.} \end{cases} \tag{3.30}$$

Here we have

$$\tilde{C}_1(\sigma) := 4\tilde{C}\left(1 + \frac{1}{\log(2)}\sqrt{\frac{\pi}{2\sigma}}\right). \tag{3.31}$$

Now we relate the number of collocation points $\eta = \eta(\mathrm{w}, N) = \#\mathscr{H}(\mathrm{w}, N)$ to the level w of the Smolyak formula. We state the result in the following lemma.

LEMMA 3.17. *Using the Smolyak formula described by* (2.9) *with Gaussian abscissas, the total number of points required at level* w *satisfies the following bounds:*

$$N(2^{\mathrm{w}} + 1) \le \eta \le (e\, 2^{1+\log_2(1.5)} N)^{\mathrm{w}} \min\{(\mathrm{w}+1), e\, 2^{1+\log_2(1.5)} N\}, \tag{3.32}$$

which implies

$$\frac{\log(\eta)}{\zeta + \log(N)} - 1 \le \mathrm{w},$$

with $\zeta := 1 + (1 + \log_2(1.5))\log(2) \approx 2.1$.

Proof. By using formula (2.9), where we collocate using the Gaussian abscissas, the number of points $\eta = \eta(\mathrm{w}, N) = \#\mathscr{H}(\mathrm{w}, N)$ can be counted in the following way:

$$\eta = \sum_{\mathbf{i} \in Y(\mathrm{w}, N)} \prod_{n=1}^{N} \widetilde{r}(i_n), \text{ where } 2^{i-1} \le \widetilde{r}(i) := \begin{cases} 1 & \text{if } i = 1, \\ 2^{i-1} + 1 & \text{if } i \ge 2. \end{cases} \tag{3.33}$$

Proceeding in a similar way as for the proof of Lemma 3.9, we obtain the following lower bound on the number of points η:

$$\eta \ge N \sum_{\tilde{\mathrm{w}}=\mathrm{w}-N+1}^{\mathrm{w}} \left(2^{\tilde{\mathrm{w}}} + 1\right) \ge N(2^{\mathrm{w}} + 1).$$

On the other hand, an upper bound on η can be obtained following along the same lines as in the proof of Lemma 3.9 and observing that $2^{i-1} \le \widetilde{r}(i) \le 2^{(1+\epsilon)(i-1)}$, with $\epsilon = \log_2(1.5) \approx 0.585$. ∎

Finally, the next theorem relates the error bound (3.30) to the number of collocation points $\eta = \eta(\mathrm{w}, N) = \#\mathscr{H}(\mathrm{w}, N)$, described by Lemma 3.17.

THEOREM 3.18 (algebraic convergence). *For functions* $u \in C^0(\Gamma^N; W(D))$ *satisfying the assumption of Lemma* 3.2 *the isotropic Smolyak formula* (2.8) *based on Gaussian abscissas satisfies*

$$\|(I_N - \mathscr{A}(\mathrm{w}, N))(u)\|_{\rho,N} \le \sqrt{\|\rho/\hat{\rho}\|_{L^\infty(\Gamma^N)}}\, e^{\sigma\, e \log(2)}\, \tilde{C}_1(\sigma)\, \frac{\max\{1, \tilde{C}_1(\sigma)\}^N}{|1 - \tilde{C}_1(\sigma)|}\, \eta^{-\tilde{\mu}_1}, \tag{3.34}$$

$$\tilde{\mu}_1 = \frac{\sigma\, e \log(2)}{\zeta + \log(N)},$$

with $\zeta := 1 + (1 + \log_2(1.5))\log(2) \approx 2.1$. *The constant* $\tilde{C}_1(\sigma)$ *was defined in* (3.31).

Similarly to section 3.1.1 and with the same assumptions as those of the previous theorem, for large values of w we have the following sharper estimate.

THEOREM 3.19 (subexponential convergence). *If* $\mathrm{w} > \frac{N}{\log(2)}$, *then*

(3.35)
$$\left\|\left(I_N - \mathscr{A}(\mathrm{w}, N)\right)(u)\right\|_{\rho,N} \leq \sqrt{\|\rho/\hat{\rho}\|_{L^\infty(\Gamma^N)}}\tilde{C}_1(\sigma)\,\frac{\max\{1, \tilde{C}_1(\sigma)\}^N}{|1 - \tilde{C}_1(\sigma)|}\,e^{-\frac{N\sigma}{2^{1/N}}\eta^{\tilde{\mu}_2}},$$

$$\textit{with } \tilde{\mu}_2 = \frac{\log(2)}{N(\zeta + \log(N))}$$

and $\zeta := 1 + (1 + \log_2(1.5))\log(2) \approx 2.1$.

3.2. Influence of truncation errors. In this section we consider the case where the coefficients a_N and f_N from (1.4) are suitably truncated random fields. Therefore, the truncation error $u - u_N$ is nonzero and contributes to the total error. Such a contribution should be considered as well in the error analysis. In particular, understanding the relationship of this error with the discretization error allows us to compare the efficiency of the isotropic Smolyak method with other computational alternatives, for instance, the Monte Carlo method.

To this end, we make the assumption that the truncation error $u - u_N$ decays as

$$\|u - u_N\|_{L^2_P(\Omega;W(D))} \leq \zeta(N) \tag{3.36}$$

for some monotonic decreasing function $\zeta(N)$ such that $\zeta(N) \to 0$ as $N \to \infty$. For example, if one truncates the input random fields with a Karhunen–Loève expansion (see [16]), the function $\zeta(N)$ is typically related to the decay of the eigenpairs of their covariance operators.

Now, given a desired computational accuracy to achieve, tol > 0, our aim is to choose the dimension $N = N(\mathrm{tol})$ and the level $\mathrm{w} = \mathrm{w}(\mathrm{tol})$ (or equivalently $\eta = \eta(\mathrm{tol})$, the number of collocation points) such that

$$\|u - \mathscr{A}(\mathrm{w}, N)(u_N)\|_{L^2_P(\Omega;W(D))} \leq \zeta(N) + \|u_N - \mathscr{A}(\mathrm{w}, N)(u_N)\|_{L^2_P(\Omega;W(D))} \approx \mathrm{tol}.$$

More precisely, we will impose that both error contributions should be of size tol, i.e.,

$$\zeta(N) \approx \mathrm{tol} \tag{3.37}$$

and

$$\|u_N - \mathscr{A}(\mathrm{w}, N)(u_N)\|_{L^2_P(\Omega;W(D))} \approx \mathrm{tol}. \tag{3.38}$$

Condition (3.37) determines the dimension $N(\mathrm{tol})$, while (3.38) determines the necessary number of collocation points in the isotropic Smolyak approximation. Then this number of collocation points is compared to the number of samples required in the standard Monte Carlo method to approximate a statistical quantity of interest with accuracy tol. The latter is $\mathcal{O}(\mathrm{tol}^{-2})$.

We detail only the procedure for the choice of Clenshaw–Curtis abscissas, since the discussion for Gaussian abscissas is identical. To impose condition (3.38), we apply Theorem 3.7, with the choice $\delta^* = (e\log(2) - 1)/\tilde{C}_2(\sigma)$ and $\tilde{C}_2(\sigma)$ as in (3.11), yielding

$$\|u_N - \mathscr{A}(\mathrm{w}, N)(u_N)\|_{L^2_\rho(\Gamma^N;W(D))} \leq \frac{C_1(\sigma, \delta^*)}{|1 - C_1(\sigma, \delta^*)|}\max\{1, C_1(\sigma, \delta^*)\}^N e^{-\sigma \mathrm{w}},$$

where the constant $C_1(\sigma, \delta^*)$ is defined in (3.12). Now define the constants $C = \frac{C_1(\sigma,\delta^*)}{|1-C_1(\sigma,\delta^*)|}$ and $F = \max\{1, C_1(\sigma, \delta^*)\}$. With this notation and using (3.23), we have an upper bound in terms of the number of collocation points,

$$\begin{aligned}\|u_N - \mathscr{A}(\mathrm{w}, N)(u_N)\|_{L^2_\rho(\Gamma^N;W(D))} &\le C\, F^N e^{-\sigma \mathrm{w}} \\ &\le C\, F^N e^{\sigma}\, \eta^{-\frac{\sigma}{1+\log(2N)}} \approx \mathrm{tol}.\end{aligned}$$

Then, given the value of $N(\mathrm{tol})$, we can find

$$\eta(\mathrm{tol}) \approx \left(\frac{C F^N e^{\sigma}}{\mathrm{tol}}\right)^{\frac{1+\log(2N)}{\sigma}} \tag{3.39}$$

and compare with the number of samples needed to achieve accuracy tol with Monte Carlo, which is $\eta_{MC} \approx \mathrm{tol}^{-2}$.

Exponential truncation error. Here we have $\zeta(N) = \theta e^{-\gamma N}$, with θ and γ positive constants. Therefore the dimension depends on the required accuracy like

$$N(\mathrm{tol}) = \frac{1}{\gamma} \log\left(\frac{\theta}{\mathrm{tol}}\right),$$

and the number of corresponding collocation points, following (3.39), is

$$\eta(\mathrm{tol}) \approx \left(C e^{\sigma} F^{\frac{1}{\gamma}\log(\theta)}\right)^{\frac{\log(2e/\gamma)+\log(\log(\theta/\mathrm{tol}))}{\sigma}} \mathrm{tol}^{-(1+\log(F)/\gamma)\left(\frac{\log(2e/\gamma)+\log(\log(\theta/\mathrm{tol}))}{\sigma}\right)}.$$

From here we can see roughly that for the exponential truncation error case the isotropic Smolyak method would be more efficient than Monte Carlo only if

$$\left(1 + \frac{\log(F)}{\gamma}\right) \frac{\log(2e/\gamma) + \log(\log(\theta/\mathrm{tol}))}{\sigma} < 2.$$

Observe that for sufficiently stringent accuracy requirements, i.e., tol sufficiently small, the Monte Carlo method will have a better convergence rate. On the other hand, due to the very slow growth of the $\log(\log(\theta/\mathrm{tol}))$ term above, these values of tol may be much smaller than the ones we need in practice. Thus, the range of parameters for which the isotropic Smolyak approximation gives a better convergence rate than Monte Carlo can still be relevant in many practical problems with truncated coefficients.

Observe, moreover, that whenever the parameter γ is large, the behavior of the one-dimensional interpolation error varies widely with respect to the different y directions. In such a case, it is likely that the isotropic Smolyak method uses too many points in the directions with fastest decay. For such a case, the isotropic Smolyak method may still be better than Monte Carlo, yet we recommend the use of an anisotropic version of the Smolyak method to obtain faster convergence. For instance, see [19, 8], where anisotropic Smolyak formulas have been proposed.

Algebraic truncation error. Here we have $\zeta(N) = \theta N^{-r}$, with θ and r positive constants. Therefore the dimension is $N(\mathrm{tol}) = (\mathrm{tol}/\theta)^{-\frac{1}{r}}$ and we have

$$\frac{\mathrm{tol}}{F^{(\theta/\mathrm{tol})^{1/r}}} \approx C e^{\sigma} \eta^{\frac{\sigma}{\log(2e)+\frac{1}{r}\log(\theta/\mathrm{tol})}}.$$

After denoting $\widehat{\text{tol}} = \frac{\text{tol}}{F^{(\theta/\text{tol})^{1/r}}} \leq \text{tol}$, the corresponding number of collocation points is

$$\eta(\text{tol}) \approx \left(Ce^{\sigma}\,\widehat{\text{tol}}\right)^{-\frac{\log(2e)+\frac{1}{r}\log(\theta/\text{tol})}{\sigma}}. \tag{3.40}$$

Observe that even for the case where $F = 1$ we now have an asymptotically faster growth of $\eta(\text{tol})$ than in the exponential truncation case. In fact, for such a case we need to have

$$\frac{\log(2e) + \frac{1}{r}\log(\theta/\text{tol})}{\sigma} < 2$$

for the isotropic Smolyak method to be more efficient than Monte Carlo. If $F > 1$, then $\widehat{\text{tol}} < \text{tol}$ and this makes, as tol gets smaller, the comparison even more favorable to Monte Carlo; cf. (3.40).

4. Application to linear elliptic PDEs with random input data. In this section we apply the theory developed so far to the particular linear problem described in Example 1.1. Problem (1.2) can be written in a weak form as follows: Find $u \in L^2_P(\Omega; H^1_0(D))$ such that

$$\int_D E[a\nabla u \cdot \nabla v]\,dx = \int_D E[fv]\,dx \quad \forall\, v \in L^2_P(\Omega; H^1_0(D)). \tag{4.1}$$

A straightforward application of the Lax–Milgram theorem allows one to state the well-posedness of problem (4.1) and yields

$$\|u(\omega)\|_{H^1_0(D)} \leq \frac{C_P}{a_{min}}\|f(\omega,\cdot)\|_{L^2(D)} \quad \text{a.s.},$$

$$\|u\|_{L^2_P(\Omega;H^1_0(D))} \leq \frac{C_P}{a_{min}}\left(\int_D E[f^2]\,dx\right)^{1/2},$$

where C_P denotes the constant appearing in the Poincaré inequality: $\|v\|_{L^2(D)} \leq C_P\|\nabla v\|_{L^2(D)}\ \forall v \in H^1_0(D)$.

Once we have the input random fields described by a finite set of random variables, i.e., $a(\omega, x) = a_N(Y_1(\omega), \ldots, Y_N(\omega), x)$, and similarly for $f(\omega, x)$, the "finite-dimensional" version of the stochastic variational formulation (4.1) has a "deterministic" equivalent, which is the following: Find $u_N \in L^2_\rho(\Gamma^N; H^1_0(D))$ such that

$$\begin{aligned}\int_{\Gamma^N} (a_N\nabla u_N, \nabla v)_{L^2(D)}\,\rho(y)dy \\ = \int_{\Gamma^N} (f_N, v)_{L^2(D)}\,\rho(y)dy \quad \forall\, v \in L^2_\rho(\Gamma^N; H^1_0(D)),\end{aligned} \tag{4.2}$$

where $\rho(y)$ is the joint probability density function defined by (1.6). Observe that in this work the gradient notation, ∇, always means differentiation with respect to $x \in D$ only, unless otherwise stated. The stochastic boundary value problem (4.1) now becomes a deterministic Dirichlet boundary value problem for an elliptic partial differential equation with an N-dimensional parameter. Then it can be shown that problem (4.1) is equivalent to

$$\int_D a_N(y)\nabla u_N(y) \cdot \nabla\phi\,dx = \int_D f_N(y)\phi\,dx \quad \forall \phi \in H^1_0(D),\ \ \rho\text{-a.e. in } \Gamma^N. \tag{4.3}$$

For our convenience, we will suppose that the coefficient a_N and the forcing term f_N admit a smooth extension on the ρ-zero measure sets. Then (4.3) can be extended a.e. in Γ^N with respect to the Lebesgue measure (instead of the measure ρdy).

It has been proved in [3] that problem (4.3) satisfies the analyticity result stated in Assumption 1.8. For instance, if we take the diffusivity coefficient as in Example 1.4 and a deterministic load, the size of the analyticity region is given by

$$\tau_n = \frac{a_{min}}{4\sigma_n}. \tag{4.4}$$

On the other hand, if we take the diffusivity coefficient as a truncated expansion as in Remark 1.6, then the analyticity region $\Sigma(\Gamma_n; \tau_n)$ is given by

$$\tau_n = \frac{1}{4\sqrt{\lambda_n}\|b_n\|_{L^\infty(D)}}. \tag{4.5}$$

Observe that, in the latter case, as $\sqrt{\lambda_n}\|b_n\|_{L^\infty(D)} \to 0$ for a regular enough covariance function (see [16]), the analyticity region increases as n increases. This fact introduces, naturally, an anisotropic behavior with respect to the "direction" n. This effect will not be exploited in the numerical methods proposed in the next sections but is the subject of ongoing research.

The finite element operator π_h can be introduced for this problem by projecting (4.3) onto the subspace $W_h(D)$ for each $y \in \Gamma^N$, i.e., $u_h^N(y) = \pi_h u_N(y)$ satisfies

$$\int_D a_N(y)\nabla u_h^N(y) \cdot \nabla \phi_h \, dx = \int_D f_N(y)\phi_h \, dx \quad \forall \phi_h \in W_h(D) \text{ for a.e. } y \in \Gamma^N. \tag{4.6}$$

Notice that the finite element functions $u_h^N(y)$ satisfy the optimality condition (2.2) $\forall y \in \Gamma^N$. Finally, the Smolyak formula (2.8) can be applied to u_h^N to obtain the fully discrete solution. The error estimates for the Smolyak approximation, stated in Theorems 3.10 and 3.11 for Clenshaw–Curtis abscissas and Theorems 3.18 and 3.19 for Gaussian abscissas, hold in this case, with parameter

$$\sigma = \frac{1}{2} \min_{n=1,\dots,N} \log\left(\frac{2\tau_n}{|\Gamma_n|} + \sqrt{1 + \frac{4{\tau_n}^2}{|\Gamma_n|^2}}\right). \tag{4.7}$$

5. Numerical examples. This section illustrates the convergence of the sparse collocation method for the stochastic linear elliptic problem in two spatial dimensions, as described in section 4. The computational results are in accordance with the convergence rates predicted by the theory. Actually, we observe a faster convergence than stated in Theorems 3.10 and 3.18, which hints that the current estimates may be improved.

We will also use this section to compare the convergence of the isotropic Smolyak approximation, described and analyzed in sections 2.2 and 3.1, respectively, with other ensemble-based methods such as the anisotropic adaptive full tensor product method described in the work [5, section 9] and the Monte Carlo method. The problem is to solve

$$\begin{cases} -\nabla \cdot (a(\omega,\cdot)\nabla u(\omega,\cdot)) = f(\omega,\cdot) & \text{in } \Omega \times D, \\ u(\omega,\cdot) = 0 & \text{on } \Omega \times \partial D, \end{cases} \tag{5.1}$$

with $D = [0,d]^2$ and $d = 1$. We consider a deterministic load $f(\omega,x,z) = \cos(x)\sin(z)$ and construct the random diffusion coefficient $a_N(\omega,x)$ with one-dimensional (layered)

spatial dependence as

$$\log(a_N(\omega,x)-0.5) = 1 + \mathrm{Y}_1(\omega)\left(\frac{\sqrt{\pi}L}{2}\right)^{1/2} + \sum_{n=2}^{N} \zeta_n\,\varphi_n(x)\,\mathrm{Y}_n(\omega), \tag{5.2}$$

where

$$\zeta_n := \left(\sqrt{\pi}L\right)^{1/2} \exp\left(\frac{-\left(\lfloor \frac{n}{2} \rfloor \pi L\right)^2}{8}\right) \quad \text{if } n>1 \tag{5.3}$$

and

$$\varphi_n(x) := \begin{cases} \sin\left(\frac{\lfloor \frac{n}{2} \rfloor \pi x}{L_p}\right) & \text{if } n \text{ even}, \\ \cos\left(\frac{\lfloor \frac{n}{2} \rfloor \pi x}{L_p}\right) & \text{if } n \text{ odd}. \end{cases} \tag{5.4}$$

In this example, the random variables $\{Y_n(\omega)\}_{n=1}^{\infty}$ are independent, have zero mean and unit variance, i.e., $E[Y_n]=0$ and $E[Y_nY_m]=\delta_{nm}$ for $n,m\in\mathbb{N}_+$, and are uniformly distributed in the interval $[-\sqrt{3},\sqrt{3}]$. Consequently, the auxiliary probability density $\hat{\rho}$ defined by (2.14) can be taken equal to the joint probability density function ρ defined by (1.6). Expression (5.2) represents the truncation of a one-dimensional random field with stationary covariance

$$\begin{aligned} &\operatorname{cov}[\log(a_N-0.5)](x_1,x_2) \\ &\quad = E\big[(\log(a)(x_1)-E[\log(a)](x_1))\,(\log(a)(x_2)-E[\log(a)](x_2))\big] \\ &\quad = \exp\left(\frac{-(x_1-x_2)^2}{L_c^2}\right). \end{aligned}$$

For $x\in[0,d]$, let L_c be a desired physical correlation length for the coefficient a, meaning that the random variables $a(x)$ and $a(y)$ become essentially uncorrelated for $|x-y|\gg L_c$. Then the parameter L_p in (5.4) is $L_p=\max\{d,2L_c\}$ and the parameter L in (5.2) and (5.3) is $L=L_c/L_p$.

The rate of convergence of the isotropic Smolyak method is dictated by the decay coefficient σ defined by (4.7), which in this case can be bounded as

$$\sigma \geq \frac{1}{2}\log\left(1+\sqrt{\frac{1}{24\sqrt{\pi}L}}\right). \tag{5.5}$$

From (5.5) we notice that larger correlation lengths will have negative effects on the rate of convergence, i.e., the coefficient σ appearing in the estimates (3.26)–(3.27) and (3.34)–(3.35) is approaching 1 as L_c becomes large. Hence, the effect of increasing L_c is a deterioration of the rate of convergence.

Recall from section 2.3 that the Clenshaw–Curtis abscissas are nested and therefore, in practice, we exploit this fact and construct the isotropic Smolyak interpolant using formula (2.8). Hence, the number of points $\eta=\eta(\mathrm{w},N)=\#\mathscr{H}(\mathrm{w},N)$ can be counted as in formula (3.24). On the other hand, the Gaussian abscissas, which in this case are the roots of the Legendre polynomials, are not nested, and to reduce the number of points necessary to build the isotropic Smolyak formula one utilizes the variant of (2.8), given by (2.9). Consequently, we can count the number of points η used by the Smolyak formula as in (3.33).

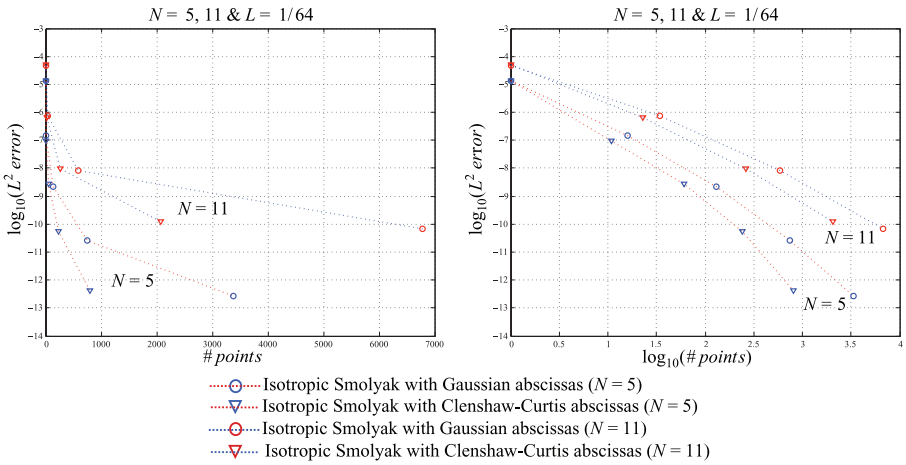


FIG. 3. *The rate of convergence of the isotropic Smolyak approximation for solving problem* (5.1) *with correlation length* $L_c = 1/64$ *using both the Gaussian and Clenshaw–Curtis abscissas. For a finite-dimensional probability space* Γ^N *with* $N = 5$ *and* $N = 11$ *we plot the* $L^2(D)$ *approximation error in the expected value in the* log*-linear scale (left) and* log-log *scale (right).*

The finite element space for the spatial discretization is the span of continuous functions that are piecewise polynomials with degree two over a uniform triangulation of D with 4225 unknowns.

Observe that the collocation method requires only the solution of uncoupled deterministic problems over the set of collocation points, even in the presence of a diffusivity coefficient which depends nonlinearly on the random variables as in (5.2). This is a significant advantage that the collocation method offers compared to the classical stochastic-Galerkin finite element method as considered, for instance, in [4, 16, 26, 37]. To study the convergence of the isotropic Smolyak approximation we consider a problem with a fixed dimension N and investigate the behavior when the level w in the Smolyak formula is increased linearly.

The computational results for the $L^2(D)$ error to the expected value, $E[u]$, using the isotropic Smolyak approximation, are shown in Figure 3. Here we consider the truncated probability space to have dimensions $N = 5$ and $N = 11$. To estimate the computational error in the level w we approximate $\|E[\epsilon]\| \approx \|E[\mathscr{A}(\mathrm{w}, N)\pi_h u_N - \mathscr{A}(\mathrm{w}+1, N)\pi_h u_N]\|$. The results reveal, as expected, that for a small nondegenerate correlation length, i.e., $L_c = 1/64$, the error decreases (sub)exponentially, as the level w increases. We also observe that the convergence rate is dimension dependent and slightly deteriorates as N increases.

To investigate the performance of the isotropic Smolyak approximation by varying the correlation length L_c we also include the cases where $L_c = 1/16$, $L_c = 1/4$, and $L_c = 1/2$ for both $N = 5$ and $N = 11$, as seen in Figure 4. As predicted by (5.5), we observe that the larger correlation lengths do indeed slow down the rate of convergence. Our final interest, then, is to compare our isotropic sparse tensor product method with the Monte Carlo approach, as well as with the anisotropic full tensor product method, proposed in [5].

The anisotropic full tensor product algorithm can be described in the following way: Given a tolerance tol the method computes a multi-index $\mathbf{p} = (p_1, p_2, \ldots, p_N)$, corresponding to the order of the approximating polynomial spaces $\mathcal{P}_{\mathbf{p}}(\Gamma^N)$. This

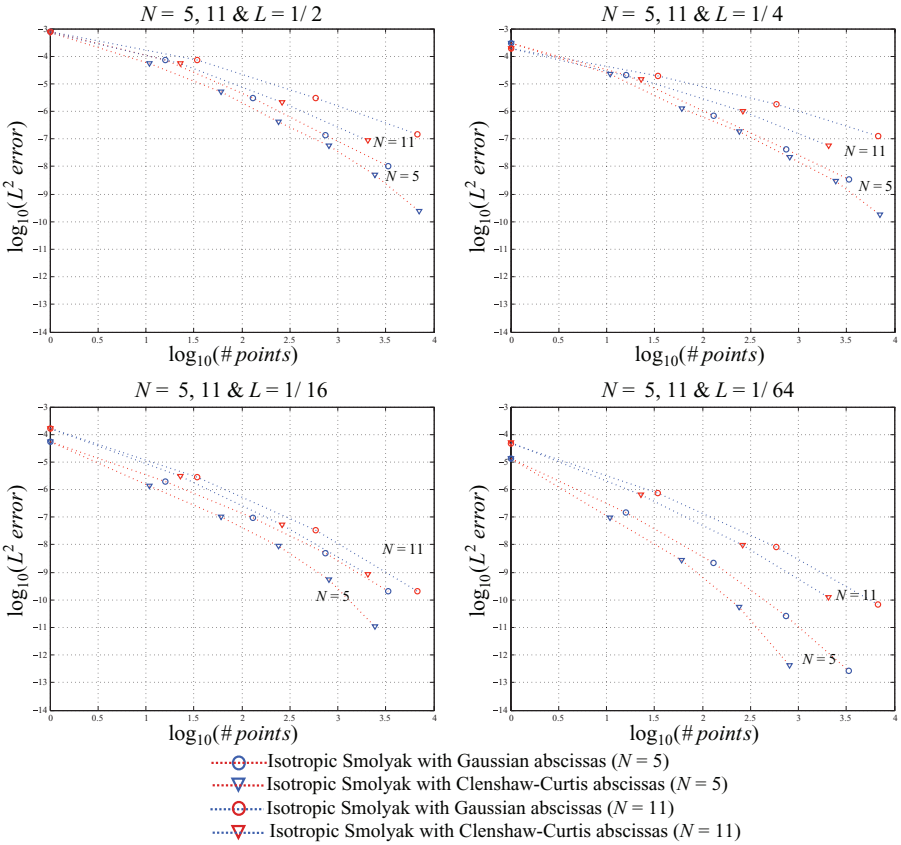


FIG. 4. *The convergence of the isotropic Smolyak approximation for solving problem* (5.1) *with given correlation lengths* $L_c = 1/2, 1/4, 1/16$, *and* $1/64$ *using both the Gaussian and Clenshaw–Curtis abscissas. For a finite-dimensional probability space* Γ^N *with* $N = 5$ *and* $N = 11$ *we plot the* $L^2(D)$ *approximation error in the expected value versus the number of collocation points.*

adaptive algorithm increases the tensor polynomial degree with an anisotropic strategy: it increases the order of approximation in one direction as much as possible before considering the next direction. Tables 1 and 2 show the values of components of the 11-dimensional multi-index $\mathbf{p}$ for different values of tol, corresponding to $L_c = 1/2$ and $L_c = 1/64$, respectively. These tables also give insight into the anisotropic behavior of each particular problem. Notice, in particular, that for the case $L_c = 1/64$ the algorithm predicts a multi-index $\mathbf{p}$ which is equal in all directions, i.e., an isotropic tensor product space. A convergence plot for $L_c = 1/2$ and $L_c = 1/64$ can be constructed by examining each row of Table 1 and Table 2, respectively, and plotting the number of points in the tensor product grid versus the error in expectation. We estimate the error in expectation by $\|E[\epsilon]\| \approx \|E[u^N_{h,\mathbf{p}} - u^N_{h,\widetilde{\mathbf{p}}}]\|$, with $\widetilde{\mathbf{p}} = (p_1 + 1, p_2 + 1, \ldots, p_N + 1)$. This entails an additional computational cost, which is bounded by the factor $\exp\left(\sum_{n=1}^{N} 1/p_n\right)$ times the work to compute $E[u^N_{h,\mathbf{p}}]$.

The standard Monte Carlo finite element method is a popular choice for solving stochastic problems such as (5.1) (see, e.g., [21, 9, 5] and the references therein). If the aim is to compute a functional of the solution such as the expected value, one would approximate $E[u]$ numerically by sample averages of independant and identically dis-

TABLE 1
The $N = 11$ components of the multi-index $\mathbf{p}$ computed by the anisotropic full tensor product algorithm when solving problem (5.1) with a correlation length $L_c = 1/2$.

tol	$N = 1$	$N = 2, 3$	$N = 4, 5$	$N = 6, 7$	$N = 8, 9$	$N = 10, 11$
1.0e-04	1	1	1	1	1	1
1.0e-05	2	1	1	1	1	1
1.0e-06	2	2	1	1	1	1
1.0e-07	3	2	2	1	1	1
1.0e-08	4	3	2	1	1	1
1.0e-09	4	4	3	1	1	1
1.0e-10	5	5	3	2	1	1
1.0e-11	5	5	4	2	1	1
1.0e-12	5	6	4	2	1	1

TABLE 2
The $N = 11$ components of the multi-index $\mathbf{p}$ computed by the anisotropic full tensor product algorithm when solving problem (5.1) with a correlation length $L_c = 1/64$.

tol	$N = 1$	$N = 2, 3$	$N = 4, 5$	$N = 6, 7$	$N = 8, 9$	$N = 10, 11$
1.0e-03	1	1	1	1	1	1
1.0e-06	2	2	2	2	2	2
1.0e-09	3	3	3	3	3	3
1.0e-12	4	4	4	4	4	4

tributed (i.i.d.) realizations of the stochastic input data. Given a number of realizations, $M \in \mathbb{N}_+$, we compute the sample average as follows: For each $k = 1, \ldots, M$, sample i.i.d. realizations of $a(\omega_k, \cdot)$ and $f(\omega_k, \cdot)$, solve problem (5.1), and construct finite element approximations $u_h^N(\omega_k, \cdot)$. We note that once we have fixed $\omega = \omega_k$, the problem is completely deterministic, and may be solved by standard methods as in the collocation approach. Finally, approximate $E[u]$ by the sample average: $\overline{E}[u_{h,k}^N; M](\cdot) := \frac{1}{M} \sum_{k=1}^{M} u_h^N(\omega_k, \cdot)$.

For the cases $L_c = 1/2, 1/4, 1/16$, and $1/64$ we take $M = 2^i$, $i = 0, 1, 2, \ldots, 11$, realizations and compute the approximation to the error in expectation by $\|E[\epsilon]\| \approx \|\overline{E}[u_{h,k}^N; M] - E[\mathscr{A}(\overline{\mathrm{w}}+1, N)\pi_h u_N]\|$, with $\overline{\mathrm{w}} = 4$, so that $\mathscr{A}(5, N)$ is a highly enriched Clenshaw–Curtis isotropic sparse solution.

To study the advantages of utilizing an isotropic sparse tensor product space as opposed to an anisotropic full tensor product space, we show in Figure 5 the convergence of these methods when solving problem (5.1), using correlation lengths $L_c = 1/2, 1/4, 1/16$, and $1/64$ with $N = 11$. We also include five ensembles of the Monte Carlo method described previously. Figure 5 reveals that for the isotropic case with $L_c = 1/64$ the isotropic Smolyak method obtains a faster rate of convergence than the anisotropic full tensor product method. This is due to a slower decay of the eigenvalues expansion (5.2) and hence an almost equal weighing of all $N = 11$ random variables. On the contrary, opposite behavior can be observed for $L_c = 1/2$. Since, in this case, the rate of decay of the expansion is faster, the anisotropic full tensor method weighs heavily the important modes and, therefore, achieves a faster convergence than the isotropic Smolyak method.

In all four cases we observe that the two methods outperform the Monte Carlo method. We know that the amount of work to reach the accuracy ϵ in the Monte Carlo approach can be approximated by $\epsilon \approx O(M^{-1/2})$ times the amount of work per sample, where M is the number of samples. This is affected only by the problem dimension through the eventual increase of the work per sample. Nevertheless, the convergence rate is quite slow and a high level of accuracy is achieved only when a

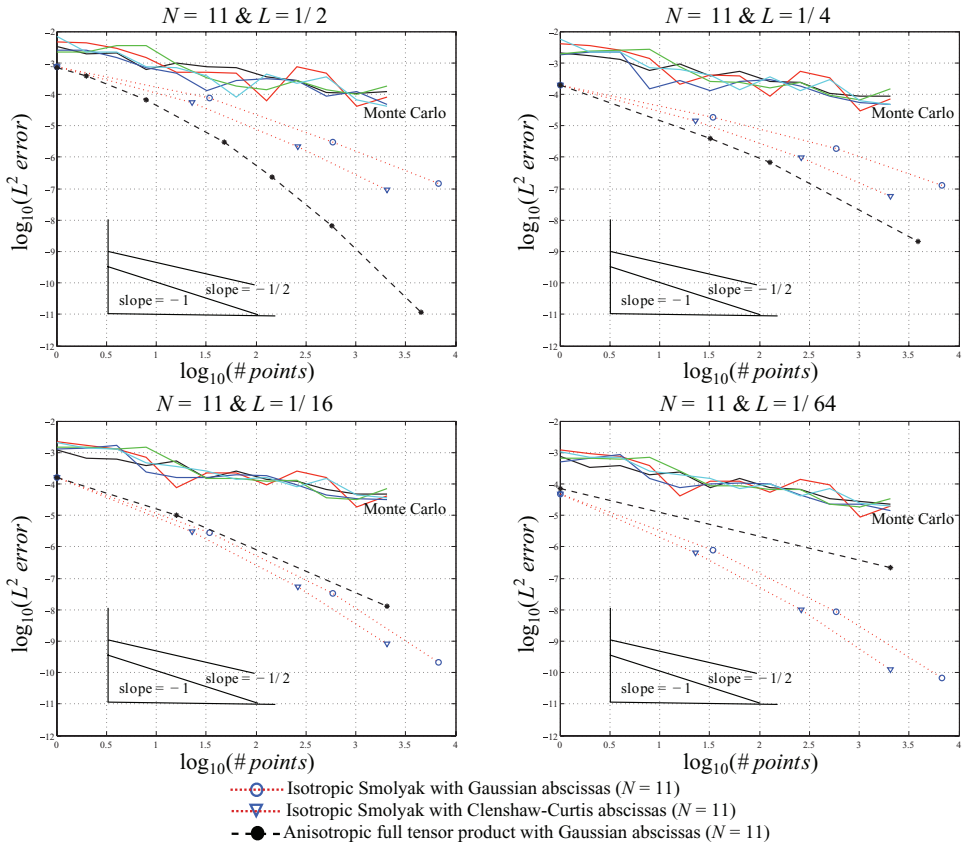


FIG. 5. *An 11-dimensional comparison of the isotropic Smolyak method, the anisotropic full tensor product algorithm, and Monte Carlo approach for solving problem* (5.1) *with correlation lengths* $L_c = 1/2, 1/4, 1/16,$ *and* $1/64$. *We plot the* $L^2(D)$ *approximation error in the expected value versus the number of collocation points (or samples of the Monte Carlo method).*

TABLE 3

For $N = 11$, *we compare the number of function evaluations required by the anisotropic full tensor product method (AF) using Gaussian abscissas, isotropic smolyak (IS) using Clenshaw–Curtis abscissas, and the Monte Carlo (MC) method using random abscissas to reduce the original error of problem* (5.1), *in expectation, by a factor of* 10^4.

L_c	AF	IS	MC
1/2	2.5×10^2	2.5×10^3	5.0×10^9
1/4	1.2×10^3	4.0×10^3	2.0×10^9
1/16	2.0×10^3	5.0×10^2	1.6×10^9
1/64	2.0×10^5	3.6×10^2	1.3×10^9

large amount of function evaluations are required. This can been seen from Figure 5, where we include reference lines with slopes $-1/2$ and -1, respectively, or in Table 3, where, for $N = 11$, we compare the work, proportional to the number of samples, which is the number of collocation points, required by each method to decrease the original error by a factor of 10^4 for all four correlation lengths $L_c = 1/2, 1/4, 1/16,$ and $1/64$.

6. Conclusions. In this work we proposed and analyzed a sparse grid stochastic collocation method for solving PDEs whose coefficients and forcing terms depend on a finite number of random variables. The sparse grids are constructed from the Smolyak formula, utilizing either Clenshaw–Curtis or Gaussian abscissas. The method leads to the solution of uncoupled deterministic problems and, as such, it is simple to implement, allows for the use of legacy codes, and is fully parallelizable like a Monte Carlo method.

This method is an improvement of the stochastic collocation method on tensor product grids proposed in [3]. The use of sparse grids considered in the present work (as opposed to full tensor grids) reduces considerably the curse of dimensionality and allows us to treat effectively problems that depend on a moderately large number of random variables, while keeping a high level of accuracy.

Upon assumption that the solution depends analytically on each random variable (which is a reasonable assumption for a certain class of applications; see [4, 3]), we derived strong error estimates for the fully discrete sparse grid stochastic collocation solution and analyzed its computational efficiency. In particular, the main result is the algebraic convergence with respect to the total number of collocation points; cf. Theorems 3.10 and 3.18. The exponent of such algebraic convergence depends on both the regularity of the solution and the number of input random variables, N. The exponent essentially deteriorates with N by a factor of $1/\log(N)$. The theory is confirmed numerically by the examples presented in section 5. We also utilized the error estimates to compare the method with Monte Carlo in terms of computational work to achieve a given accuracy, indicating for which problems the former is more efficient than the latter. To this effect, in section 3.2 we considered a case where the input random variables come from suitably truncated expansions of random fields and related the number of collocation points in the sparse grid to the number of random variables retained in the truncated expansion. We also developed error estimates with fewer regularity requirements in Remark 3.14.

The sparse grid method is very effective for problems whose input data depend on a moderate number of random variables, which "weigh equally" in the solution. For such an isotropic situation the displayed convergence is faster than standard collocation techniques built upon full tensor product spaces.

On the other hand, the convergence rate deteriorates when we attempt to solve highly anisotropic problems, such as those appearing when the input random variables come, e.g., from Karhunen–Loève truncated expansions of "smooth" random fields. In such cases, a full anisotropic tensor product approximation, as proposed in [5, 3], may still be more effective for a small or moderate number of random variables.

Future directions of this research will include the development and analysis of an anisotropic version of the sparse grid stochastic collocation method, which will combine an optimal treatment of the anisotropy of the problem while reducing the curse of dimensionality via the use of sparse grids.

Appendix. Additional estimates. Here we present auxiliary results that are used in section 3. Let us recall the definition for the integer and fractional parts of a nonnegative real number x that satisfy $x = frac\{x\} + \mathrm{int}\{x\}\ \forall x \in \mathbb{R}^+$, with $\mathrm{int}\{x\}$ being the largest natural number that is smaller than or equal to x.

LEMMA A.1. *Given* $\mathrm{w} \in \mathbb{N}_+$, *for any* $\alpha > 0$ *and* $0 \le \beta < \mathrm{w}$, *we have*

$$\sum_{i=0}^{\mathrm{w}} e^{-\alpha(i-\beta)^2} \le 2e^{-\alpha(\min\{frac\{\beta\},1-frac\{\beta\}\})^2} \left(1 + \frac{1}{2}\sqrt{\frac{\pi}{\alpha}}\right).$$

Proof. Let us write $\beta = \operatorname{int}\{\beta\} + \mathit{frac}\{\beta\}$. Then

$$\begin{aligned}\sum_{i=0}^{\mathrm{w}} e^{-\alpha(i-\beta)^2} &= \sum_{j=-\operatorname{int}\{\beta\}}^{\mathrm{w}-\operatorname{int}\{\beta\}} e^{-\alpha(j-\mathit{frac}\{\beta\})^2} \\ &= \sum_{j=-\operatorname{int}\{\beta\}}^{0} e^{-\alpha(j-\mathit{frac}\{\beta\})^2} + \sum_{j=1}^{\mathrm{w}-\operatorname{int}\{\beta\}} e^{-\alpha(j-\mathit{frac}\{\beta\})^2} \\ &= \sum_{j=0}^{\operatorname{int}\{\beta\}} e^{-\alpha(j+\mathit{frac}\{\beta\})^2} + \sum_{j=0}^{\mathrm{w}-\operatorname{int}\{\beta\}-1} e^{-\alpha(j+(1-\mathit{frac}\{\beta\}))^2}.\end{aligned}$$

Finally, estimate

$$\begin{aligned}\sum_{i=0}^{\mathrm{w}} e^{-\alpha(i-\beta)^2} &\le 2e^{-\alpha(\min\{\mathit{frac}\{\beta\},1-\mathit{frac}\{\beta\}\})^2} \sum_{j=0}^{\infty} e^{-\alpha j^2} \\ &\le 2e^{-\alpha(\min\{\mathit{frac}\{\beta\},1-\mathit{frac}\{\beta\}\})^2} \left(1 + \sum_{j=1}^{\infty} e^{-\alpha j^2}\right) \\ &\le 2e^{-\alpha(\min\{\mathit{frac}\{\beta\},1-\mathit{frac}\{\beta\}\})^2} \left(1 + \int_0^{\infty} e^{-\alpha x^2} dx\right) \\ &\le 2e^{-\alpha(\min\{\mathit{frac}\{\beta\},1-\mathit{frac}\{\beta\}\})^2} \left(1 + \frac{1}{2}\sqrt{\frac{\pi}{\alpha}}\right). \qquad \square\end{aligned}$$

Now we state and prove an auxiliary estimate, to be used later on in the proof of Lemma A.3.

LEMMA A.2. *If $\alpha > 0$, we have*

$$\sum_{k=1}^{\infty} k\, e^{-\alpha k^2} \le \frac{1}{\alpha} + \frac{1}{\sqrt{2\alpha}}.$$

Proof. Observe first that $xe^{-\alpha x^2} \le \frac{1}{\sqrt{2e\alpha}}$ $\forall x \ge 0$ and that the bound is attained at $x^* = \frac{1}{\sqrt{2\alpha}}$. Then, for any integer $k_0 \ge x^*$, we can estimate

$$\sum_{k=1}^{\infty} k\, e^{-\alpha k^2} \le \frac{k_0}{\sqrt{2e\alpha}} + \int_{k_0}^{+\infty} xe^{-\alpha x^2}\, dx \le \frac{k_0}{\sqrt{2e\alpha}} + \frac{e^{-\alpha k_0^2}}{2\alpha}.$$

Finally, choosing $k_0 = \operatorname{int}\{x^*\}+1 = \operatorname{int}\{1/\sqrt{2\alpha}\}+1$, the desired result follows. $\square$

LEMMA A.3. *Given* $\mathrm{w} \in \mathbb{N}_+$, *for any* $\alpha > 0$ *and* $0 \le \beta < \mathrm{w}$, *we have*

$$\begin{aligned}\sum_{i=1}^{\mathrm{w}} i\, e^{-\alpha(i-\beta)^2} &\le 2\, e^{-\alpha(1-\mathit{frac}\{\beta\})^2} \left(\frac{1}{\alpha} + \frac{1}{\sqrt{2\alpha}}\right) \\ &\quad + (\operatorname{int}\{\beta\} + 1)\, 2\, e^{-\alpha(\min\{\mathit{frac}\{\beta\},1-\mathit{frac}\{\beta\}\})^2} \left(1 + \frac{1}{2}\sqrt{\frac{\pi}{\alpha}}\right).\end{aligned}$$

Proof. Write

$$\begin{aligned}\sum_{i=0}^{\mathrm{w}} i\, e^{-\alpha(i-\beta)^2} &= \sum_{j=-\operatorname{int}\{\beta\}}^{\mathrm{w}-\operatorname{int}\{\beta\}} (j+\operatorname{int}\{\beta\}) e^{-\alpha(j-frac\{\beta\})^2} \\ &= \sum_{j=-\operatorname{int}\{\beta\}}^{\mathrm{w}-\operatorname{int}\{\beta\}} (j-1)\, e^{-\alpha(j-frac\{\beta\})^2} \\ &\quad + (\operatorname{int}\{\beta\}+1) \sum_{j=-\operatorname{int}\{\beta\}}^{\mathrm{w}-\operatorname{int}\{\beta\}} e^{-\alpha(j-frac\{\beta\})^2}\end{aligned}$$

and bound

$$\begin{aligned}\sum_{i=0}^{\mathrm{w}} i\, e^{-\alpha(i-\beta)^2} &\le \sum_{j=1}^{\mathrm{w}-\operatorname{int}\{\beta\}} (j-1)\, e^{-\alpha(j-frac\{\beta\})^2} \\ &\quad + (\operatorname{int}\{\beta\}+1)\, 2e^{-\alpha(\min\{frac\{\beta\},1-frac\{\beta\}\})^2} \left(1+\frac{1}{2}\sqrt{\frac{\pi}{\alpha}}\right) \\ &\le e^{-\alpha(1-frac\{\beta\})^2} \sum_{j=1}^{\mathrm{w}-\operatorname{int}\{\beta\}} (j-1)\, e^{-\alpha(j-1)^2} \\ &\quad + (\operatorname{int}\{\beta\}+1)\, 2e^{-\alpha(\min\{frac\{\beta\},1-frac\{\beta\}\})^2} \left(1+\frac{1}{2}\sqrt{\frac{\pi}{\alpha}}\right).\end{aligned}$$

Finally, use the auxiliary Lemma A.2 to estimate

$$\sum_{j=0}^{\mathrm{w}-\operatorname{int}\{\mathrm{w}/d\}} j\, e^{-\alpha j^2} \le \sum_{j=1}^{\infty} j\, e^{-\alpha j^2} \le \left(\frac{1}{\alpha}+\frac{1}{\sqrt{2\alpha}}\right). \qquad \square$$

We have, as a direct consequence of Lemmas A.1 and A.3 the following estimates.

COROLLARY A.4. *There holds*

$$\sum_{i=0}^{\mathrm{w}} e^{-\sigma\frac{\log^2(2)}{2}(i-\mathrm{w}/d)^2} \le 2\left(1+\frac{1}{\log(2)}\sqrt{\frac{\pi}{2\sigma}}\right) \tag{A.1}$$

and

$$\begin{aligned}\sum_{i=0}^{\mathrm{w}} (1+i)\, e^{-\sigma\frac{\log^2(2)}{2}(i-\mathrm{w}/d)^2} &\le 2\left(\frac{1}{\sigma\log^2(2)}+\frac{1}{\log(2)\sqrt{2\sigma}}\right) \\ &\quad + 2(\operatorname{int}\{\mathrm{w}/d\}+2)\left(1+\frac{1}{\log(2)}\sqrt{\frac{\pi}{2\sigma}}\right).\end{aligned} \tag{A.2}$$

Acknowledgments. The third author would like to thank MOX, Dipartimento di Matematica, Politecnico di Milano, Italy, for hosting a visit to complete this research. Also, the third author would like to thank Prof. Max Gunzburger and Dr. John Burkardt for their insight, guidance, and many helpful discussions. The authors also would like to thank the anonymous referee for constructive comments.

REFERENCES

[1] I. M. Babuška and P. Chatzipantelidis, *On solving elliptic stochastic partial differential equations*, Comput. Methods Appl. Mech. Engrg., 191 (2002), pp. 4093–4122.

[2] I. M. Babuška, K. M. Liu, and R. Tempone, *Solving stochastic partial differential equations based on the experimental data*, Math. Models Methods Appl. Sci., 13 (2003), pp. 415–444.

[3] I. M. Babuška, F. Nobile, and R. Tempone, *A stochastic collocation method for elliptic partial differential equations with random input data*, SIAM J. Numer. Anal., 435 (2007), pp. 1005–1034.

[4] I. M. Babuška, R. Tempone, and G. E. Zouraris, *Galerkin finite element approximations of stochastic elliptic partial differential equations*, SIAM J. Numer. Anal., 42 (2004), pp. 800–825.

[5] I. M. Babuška, R. Tempone, and G. E. Zouraris, *Solving elliptic boundary value problems with uncertain coefficients by the finite element method: The stochastic formulation*, Comput. Methods Appl. Mech. Engrg., 194 (2005), pp. 1251–1294.

[6] V. Barthelmann, E. Novak, and K. Ritter, *High dimensional polynomial interpolation on sparse grids*, Adv. Comput. Math., 12 (2000), pp. 273–288.

[7] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, Springer-Verlag, New York, 1994.

[8] H.-J. Bungartz and M. Griebel, *Sparse grids*, Acta Numer., 13 (2004), pp. 147–269.

[9] J. Burkardt, M. Gunzburger, and C. Webster, *Reduced order modeling of some nonlinear stochastic partial differential equations*, Int. J. Numer. Anal. Model., 4 (2007), pp. 368–391.

[10] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North–Holland, New York, 1978.

[11] C. W. Clenshaw and A. R. Curtis, *A method for numerical integration on an automatic computer*, Numer. Math., 2 (1960), pp. 197–205.

[12] R. A. DeVore and G. G. Lorentz, *Constructive Approximation*, Grundlehren Math. Wiss. 303, Springer-Verlag, Berlin, 1993.

[13] V. K. Dzjadyk and V. V. Ivanov, *On asymptotics and estimates for the uniform norms of the Lagrange interpolation polynomials corresponding to the Chebyshev nodal points*, Anal. Math., 9 (1983), pp. 85–97.

[14] P. Erdös and P. Turán, *On interpolation. I. Quadrature- and mean-convergence in the Lagrange-interpolation*, Ann. of Math. (2), 38 (1937), pp. 142–155.

[15] G. S. Fishman, *Monte Carlo: Concepts, Algorithms, and Applications*, Springer Ser. Oper. Res., Springer-Verlag, New York, 1996.

[16] P. Frauenfelder, C. Schwab, and R. A. Todor, *Finite elements for elliptic problems with stochastic coefficients*, Comput. Methods Appl. Mech. Engrg., 194 (2005), pp. 205–228.

[17] A. Gaudagnini and S. Neumann, *Nonlocal and localized analysis of conditional mean steady state flow in bounded, randomly nonuniform domains. 1. Theory and computational approach. 2. Computational examples*, Water Resources Res., 35 (1999), pp. 2999–3039.

[18] T. Gerstner and M. Griebel, *Numerical integration using sparse grids*, Numer. Algorithms, 18 (1998), pp. 209–232.

[19] T. Gerstner and M. Griebel, *Dimension-adaptive tensor-product quadrature*, Computing, 71 (2003), pp. 65–87.

[20] R. G. Ghanem and P. D. Spanos, *Stochastic Finite Elements: A Spectral Approach*, Springer-Verlag, New York, 1991.

[21] M. Grigoriu, *Stochastic Calculus*, Appl. Sci. Engrg., Birkhäuser, Boston, MA, 2002.

[22] O. P. Le Maître, O. M. Knio, H. N. Najm, and R. G. Ghanem, *Uncertainty propagation using Wiener-Haar expansions*, J. Comput. Phys., 197 (2004), pp. 28–57.

[23] M. Loève, *Probability Theory*, I, 4th ed., Grad. Texts in Math. 45, Springer-Verlag, New York, 1977.

[24] M. Loève, *Probability Theory*, II, 4th ed., Grad. Texts in Math. 46, Springer-Verlag, New York, 1978.

[25] L. Mathelin, M. Y. Hussaini, and T. A. Zang, *Stochastic approaches to uncertainty quantification in CFD simulations*, Numer. Algorithms, 38 (2005), pp. 209–236.

[26] H. G. Matthies and A. Keese, *Galerkin methods for linear and nonlinear elliptic stochastic partial differential equations*, Comput. Methods Appl. Mech. Engrg., 194 (2005), pp. 1295–1331.

[27] F. Riesz and B. Sz-Nagy, *Functional Analysis*, Dover, New York, 1990.

[28] L. J. Roman and M. Sarkis, *Stochastic Galerkin method for elliptic SPDEs: A white noise approach*, Discrete Contin. Dyn. Syst. Ser. B, 6 (2006), pp. 941–955.

[29] S. A. Smolyak, *Quadrature and interpolation formulas for tensor products of certain classes of functions*, Dokl. Akad. Nauk SSSR, 4 (1963), pp. 240–243.

[30] M. A. Tatang, *Direct Incorporation of Uncertainty in Chemical and Environmental Engineering Systems*, Ph.D. thesis, MIT, Cambridge, MA, 1995.
[31] R. A. Todor, *Sparse Perturbation Algorithms for Elliptic PDE's with Stochastic Data*, Dissertation 16192, ETH Zurich, Switzerland, 2005.
[32] L. N. Trefethen, *Is Gauss quadrature better than Clenshaw–Curtis?*, SIAM Rev., 50 (2008), pp. 67–87.
[33] G. W. Wasilkowski and H. Woźniakowski, *Explicit cost bounds of algorithms for multivariate tensor product problems*, J. Complexity, 11 (1995), pp. 1–56.
[34] N. Wiener, *The homogeneous chaos*, Amer. J. Math., 60 (1938), pp. 897–936.
[35] C. L. Winter and D. M. Tartakovsky, *Groundwater flow in heterogeneous composite aquifers*, Water Resources Res., 38 (8) (2002), article 1148.
[36] D. Xiu and J. S. Hesthaven, *High-order collocation methods for differential equations with random inputs*, SIAM J. Sci. Comput., 27 (2005), pp. 1118–1139.
[37] D. Xiu and G. E. Karniadakis, *Modeling uncertainty in steady state diffusion problems via generalized polynomial chaos*, Comput. Methods Appl. Mech. Engrg., 191 (2002), pp. 4927–4948.
[38] D. Xiu and G. E. Karniadakis, *The Wiener–Askey polynomial chaos for stochastic differential equations*, SIAM J. Sci. Comput., 24 (2002), pp. 619–644.

SIAM J. NUMER. ANAL.
Vol. 46, No. 5, pp. 2346–2370
© 2008 Society for Industrial and Applied Mathematics

A FETI-DP FORMULATION OF THREE DIMENSIONAL ELASTICITY PROBLEMS WITH MORTAR DISCRETIZATION*

HYEA HYUN KIM†

Abstract. In this paper, a FETI-DP formulation for three dimensional elasticity on nonmatching grids over geometrically nonconforming subdomain partitions is considered. To resolve the nonconformity of the finite elements, a mortar matching condition is imposed on the subdomain interfaces (faces). A FETI-DP algorithm is then built by enforcing the mortar matching condition in dual and primal ways. In order to make the FETI-DP algorithm scalable, a set of primal constraints, which include average and momentum constraints over interfaces, are selected from the mortar matching condition. A condition number bound, $C(1 + \log(H/h))^2$, is then proved for the FETI-DP formulation for the elasticity problems with discontinuous material parameters. Only some faces need to be chosen as primal faces on which the average and momentum constraints are imposed. Numerical results are included.

Key words. FETI-DP, mortar methods, preconditioner, elasticity

AMS subject classifications. 65N30, 65N55

DOI. 10.1137/060663179

1. Introduction.

We will develop an efficient dual-primal finite element tearing and interconnecting (FETI-DP) algorithm for solving linear systems arising from certain nonconforming discretizations of compressible elasticity problems in three dimensions. We consider a nonconforming finite element space with triangulations that are nonmatching across subdomain interfaces. Allowing such triangulations helps make adaptivity for problems with singular points or with jumps in the material parameters easier and more economical and helps save the cost for mesh generation, especially for three dimensional problems. Furthermore, we allow geometrically nonconforming partitions that are useful for biomechanics with joints or with heterogeneous elasticity.

Mortar methods have been developed as nonconforming approximations with the goal of obtaining as accurate an approximate solution as possible for a conforming approximation; see [3, 1, 13, 28]. For this purpose, a mortar matching condition is imposed on the subdomain solutions across the interface. In more detail, the jumps of the solutions across the subdomain interface are orthogonal to a certain Lagrange multiplier space. This condition can be enforced directly on the nonconforming finite element functions to produce elements of the mortar finite element space. Another approach is to impose the condition weakly by introducing Lagrange multipliers, which leads to a saddle-point problem similar to that considered in FETI-type algorithms.

FETI-type algorithms were originally developed for second order elliptic problems with conforming discretizations. These algorithms belong to the iterative substructuring domain decomposition methods of dual type. A separate set of interface unknowns is assigned to each subdomain. Pointwise continuity of solutions across the interface is then enforced using dual Lagrange multipliers, leading to a saddle-point problem. The

*Received by the editors June 19, 2006; accepted for publication (in revised form) February 19, 2008; published electronically May 16, 2008. This work was supported in part by the Applied Mathematical Sciences Program of the U.S. Department of Energy under contract DE-FG02-00ER25053 and DE-FC02-01ER25482 and in part by the Postdoctoral Fellowship Program of Korea Science and Engineering Foundation (KOSEF).

http://www.siam.org/journals/sinum/46-5/66317.html

†Department of Mathematics, Chonnam National University, Youngbong-dong, Buk-gu, Gwangju 500-757, Korea (hyeahyun@gmail.com or hkim@jnu.ac.kr).

subdomain unknowns are then eliminated and the resulting linear system for the dual variables is solved iteratively using a preconditioner. These algorithms have evolved from the one-level FETI methods into two-level FETI and FETI-DP methods; see [12, 11, 9]. In FETI-DP methods, a certain set of continuity constraints is enforced throughout the iteration while the remaining constraints are imposed weakly by dual Lagrange multipliers. FETI-DP algorithms have been further developed for three dimensional elliptic problems with discontinuous coefficients by Klawonn, Widlund, and Dryja [23].

FETI-type algorithms have also been applied to the saddle-point problems resulting from mortar discretizations. A numerical study in [26] showed that such methods applied to these saddle-point problems are as efficient as the FETI methods for conforming discretizations. Lee and the author [16] introduced a FETI-DP algorithm for two dimensional elliptic problems with discontinuous coefficients and showed a condition number bound, $C(1 + \log(H/h))^2$, with a constant C independent of the coefficients and mesh parameters. Numerical results show that it is most efficient for problems with jump coefficients; see [5]. This preconditioner is similar to previously developed FETI-DP preconditioners [7, 8] except that its weights equal zero at the interface unknowns on the mortar sides. We call this preconditioner the Neumann–Dirichlet preconditioner. This algorithm has later been extended to the Stokes problem and to three dimensional elliptic problems; see [17, 14].

The purpose of this study is to extend the FETI-DP algorithm of [16] to three dimensional compressible elasticity problems with mortar discretizations on geometrically nonconforming partitions. FETI-DP methods for three dimensional elasticity problems, with conforming discretization, have been studied extensively both theoretically and numerically; see [10, 20, 22]. In [10], Farhat, Lesoinne, and Pierson introduced face average constraints in addition to vertex constraints as primal constraints and observed that these additional constraints help produce a scalable algorithm. Later Klawonn and Widlund [22] considered various primal constraints for elasticity problems with discontinuous Lamé parameters. In their work, some faces and edges were selected as fully primal faces and fully primal edges. Edge average constraints on fully primal faces and edge average and moment constraints on fully primal edges were then enforced to get a scalable algorithm. In our FETI-DP formulation, we will introduce face average and face moment constraints selected from the mortar matching condition.

This paper is organized as follows. In section 2, we introduce a compressible elasticity problem and Korn inequalities. In section 3, we approximate the solution of the model problem using the mortar discretization. We then build a FETI-DP algorithm considering the mortar matching condition just like the pointwise continuity constraints in conforming finite element approximations. Some primal constraints are selected from the mortar matching condition. In order to make them explicit, we perform a change of unknowns. Section 4 is devoted to our condition number analysis of the FETI-DP algorithm. In the final section, numerical results will be presented to verify the efficiency of the suggested algorithm under heterogeneous elasticity and geometrically nonconforming partitions.

Throughout this paper, c and C denote generic positive constants independent of the mesh size, the number of subdomains, and the problem coefficients. We will use h_i and H_i to denote the mesh size and the subdomain size of Ω_i, respectively.

2. A model problem and Korn's inequality. Let Ω be a polyhedral domain in $\mathbf{R}^3$. The Sobolev space $H^1(\Omega)$ is the set of functions in $L^2(\Omega)$ which are square

integrable up to its first derivatives and equipped with the Sobolev norm;

$$\|v\|_{1,\Omega}^2 := |v|_{1,\Omega}^2 + \frac{1}{H^2}\|v\|_{0,\Omega}^2,$$

where $|v|_{1,\Omega}^2 = \int_\Omega \nabla v \cdot \nabla v \, dx$, $\|v\|_{0,\Omega} = \int_\Omega v^2 \, dx$, and H denotes the diameter of Ω.

We assume that $\partial\Omega$ is divided into two parts $\partial\Omega_D$ and $\partial\Omega_N$ on which a Dirichlet boundary condition and a natural boundary condition, respectively, are specified. The subspace $H_D^1(\Omega) \subset H^1(\Omega)$ is the set of functions having zero traces on $\partial\Omega_D$. We introduce the vector valued Sobolev spaces $[H_D^1(\Omega)]^3$ and $[H^1(\Omega)]^3$, equipped with the usual product norm.

We then consider the elasticity problem:

Find $\mathbf{u} \in [H_D^1(\Omega)]^3$ such that

$$\int_\Omega G(\mathbf{x})\varepsilon(\mathbf{u}) : \varepsilon(\mathbf{v}) \, d\mathbf{x} + \int_\Omega G(\mathbf{x})\beta(\mathbf{x})\nabla\cdot\mathbf{u}\,\nabla\cdot\mathbf{v}\, d\mathbf{x} = \langle \mathbf{F}, \mathbf{v}\rangle \quad \forall \mathbf{v} \in [H_D^1(\Omega)]^3, \tag{2.1}$$

where $G = E/(1+\nu)$ and $\beta = \nu/(1-2\nu)$ are material parameters which depend on the Young's modulus $E > 0$ and the Poisson ratio $\nu \in (0, 1/2)$. We assume that ν is bounded from above away from $1/2$, excluding the case of incompressible elasticity problems. The linearized strain tensor is defined by

$$\varepsilon(\mathbf{u})_{ij} := \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right), \quad i, j = 1, 2, 3,$$

and the tensor product and the force term are given by

$$\varepsilon(\mathbf{u}) : \varepsilon(\mathbf{v}) = \sum_{i,j=1}^{3} \varepsilon_{ij}(\mathbf{u})\varepsilon_{ij}(\mathbf{v}), \quad \langle \mathbf{F}, \mathbf{v}\rangle = \int_\Omega \mathbf{f}\cdot\mathbf{v}\, d\mathbf{x} + \int_{\partial\Omega_N} \mathbf{g}\cdot\mathbf{v} d\sigma.$$

Here $\mathbf{f}$ is the body force and $\mathbf{g}$ is the surface force on the natural boundary part $\partial\Omega_N$.

The space $\mathbf{ker}(\varepsilon)$ has the following six rigid body motions as its basis elements. They are the three translations,

$$\mathbf{r}_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \ \mathbf{r}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \ \mathbf{r}_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \tag{2.2}$$

and the three rotations,

$$\mathbf{r}_4 = \frac{1}{H}\begin{pmatrix} x_2 - \widehat{x}_2 \\ -x_1 + \widehat{x}_1 \\ 0 \end{pmatrix}, \ \mathbf{r}_5 = \frac{1}{H}\begin{pmatrix} -x_3 + \widehat{x}_3 \\ 0 \\ x_1 - \widehat{x}_1 \end{pmatrix}, \ \mathbf{r}_6 = \frac{1}{H}\begin{pmatrix} 0 \\ x_3 - \widehat{x}_3 \\ -x_2 + \widehat{x}_2 \end{pmatrix}. \tag{2.3}$$

Here $\widehat{\mathbf{x}} = (\widehat{x}_1, \widehat{x}_2, \widehat{x}_3) \in \widehat{\Omega}$ and H is the diameter of $\widehat{\Omega}$. This shift and the scaling make the L_2-norm of the six vectors scale in the same way with H. When Ω is partitioned into a set of subdomains, the elasticity problem given on a floating subdomain has a purely natural boundary condition. The Korn inequalities in [22, section 6] address this case. Let $\Sigma \subset \partial\Omega$ be of positive measure. We define an L_2–inner product $(\mathbf{u}, \mathbf{r})_\Sigma$ by

$$(\mathbf{u}, \mathbf{r})_\Sigma = \int_\Sigma \mathbf{u}\cdot\mathbf{r}\, ds.$$

The following Korn inequality is provided in [22, Lemma 6.4].

LEMMA 2.1. *Let Ω be a Lipschitz domain and Σ be a subset of $\partial\Omega$ with positive measure. Then there exist a constant $c > 0$, invariant under dilation, such that*

$$c|\mathbf{u}|_{1,\Omega} \leq \|\varepsilon(\mathbf{u})\|_{0,\Omega} \leq |\mathbf{u}|_{1,\Omega},$$

where $\mathbf{u} \in [H^1(\Omega)]^3$ satisfies $(\mathbf{u}, \mathbf{r})_\Sigma = 0$ for all $\mathbf{r} \in \mathbf{ker}(\varepsilon)$.

Furthermore, we have similar inequalities for seminorms defined in the space $[H^{1/2}(\Sigma)]^3$ which is the trace space of $[H^1(\Omega)]^3$ on $\Sigma \subset \partial\Omega$. For $\mathbf{u} \in [H^{1/2}(\Sigma)]^3$, we define two seminorms by

$$|\mathbf{u}|_{1/2,\Sigma} := \inf_{\substack{\mathbf{v} \in [H^1(\Omega)]^3 \\ \mathbf{v}|_\Sigma = \mathbf{u}}} |\mathbf{v}|_{1,\Omega}, \quad |\mathbf{u}|_{E(\Sigma)} := \inf_{\substack{\mathbf{v} \in [H^1(\Omega)]^3 \\ \mathbf{v}|_\Sigma = \mathbf{u}}} \|\varepsilon(\mathbf{v})\|_{0,\Omega}. \tag{2.4}$$

LEMMA 2.2. *Let Ω be a Lipschitz domain and Σ be a subset of $\partial\Omega$ with positive measure. Then there exists a constant $c > 0$, invariant under dilation, such that*

$$c|\mathbf{u}|_{1/2,\Sigma} \leq |\mathbf{u}|_{E(\Sigma)} \leq |\mathbf{u}|_{1/2,\Sigma},$$

for $\mathbf{u} \in [H^{1/2}(\Sigma)]^3$ satisfying $(\mathbf{u}, \mathbf{r})_\Sigma = 0$ for all $\mathbf{r} \in \mathbf{ker}(\varepsilon)$.

This lemma can be found in [22, Lemma 6.5]. Another important inequality, which follows from this inequality, is given in [22, Lemma 6.6].

LEMMA 2.3. *Let Ω be a Lipschitz domain of diameter H and $\Sigma \subset \partial\Omega$ be an open subset with positive measure. Then there exists a constant $C > 0$ such that*

$$\inf_{\mathbf{r} \in \mathbf{ker}(\varepsilon)} \|\mathbf{u} - \mathbf{r}\|^2_{0,\Sigma} \leq CH|\mathbf{u}|^2_{E(\Sigma)} \quad \forall \mathbf{u} \in [H^{1/2}(\Sigma)]^3.$$

Here the infimum occurs when $\mathbf{r}$ satisfies $(\mathbf{u} - \mathbf{r}, \mathbf{q})_\Sigma = 0$ for all $\mathbf{q} \in \ker(\varepsilon)$.

3. FETI-DP formulation.

3.1. Mortar discretization. We divide the domain Ω into a geometrically nonconforming partition $\{\Omega_i\}_{i=1}^N$, where Ω_i are polyhedral and equipped with quasiuniform triangulations T_i.

Assumption 3.1. The subdomain partition $\{\Omega_i\}_{i=1}^N$ is locally quasiuniform, i.e., neighboring subdomains have comparable diameters and each subdomain is polyhedral.

We consider a model compressible elasticity problem (2.1) with coefficients $G(\mathbf{x})$ and $\beta(\mathbf{x})$ positive constants in each subdomain

$$G(\mathbf{x})|_{\Omega_i} = G_i, \quad \beta(\mathbf{x})|_{\Omega_i} = \beta_i.$$

A conforming P_1–finite element space $\mathbf{X}_i$ is associated with each triangulation T_i. In addition, functions in the space $\mathbf{X}_i$ satisfy the Dirichlet boundary condition on $\partial\Omega_i \cap \partial\Omega_D$. We define the union of the subdomain boundaries by

$$\Gamma = \left(\bigcup_{i=1}^N \partial\Omega_i \right) \setminus \partial\Omega.$$

The triangulations $\{T_i\}_{i=1}^N$ may not match across the subdomain boundaries. We further introduce the finite element space $\mathbf{W}_i$ that is the trace space of $\mathbf{X}_i$ on $\partial\Omega_i \cap \Gamma$.

We note that the nodal unknowns corresponding to nodes at $\partial\Omega_N$ are considered as unknowns at the interior of the subdomains.

We denote the interface of two subdomains Ω_i and Ω_j by F_{ij}, which can be only part of a face of Ω_i or Ω_j. Among the subdomain faces, we select nonmortar faces F_l such that

$$\bigcup_l \overline{F}_l = \bigcup_{i,j} \overline{F}_{ij}, \quad F_l \cap F_k = \emptyset,\ l \neq k.$$

Here each F_l is a full face of a subdomain that we call the nonmortar subdomain of F_l. We call the subdomains on the other part across F_l the mortar subdomains.

Since the subdomain partition can be geometrically nonconforming, a single nonmortar face $F_l \subset \partial\Omega_i$ may intersect several subdomain boundaries $\partial\Omega_j$. This provides F_l with a partition,

$$\overline{F}_l = \bigcup_j \overline{F}_{ij}, \quad F_{ij} = \partial\Omega_i \cap \partial\Omega_j.$$

DEFINITION 3.2. *An interface $F_{ij} = \partial\Omega_i \cap \partial\Omega_j$ is substantial if it is shape regular and its diameter is comparable to H_i and H_j.*

Assumption 3.3. Let TOL_G be a given constant. On each interface $F_{ij} = \partial\Omega_i \cap \partial\Omega_j$, the nonmortar side Ω_i and the mortar side Ω_j are selected so that

$$\frac{G_i}{G_j} \leq TOL_G.$$

Remark 3.4. It is clearly possible to construct a set of subdomains such that Assumption 3.3 is not satisfied in geometrically nonconforming cases. We note that this problem arises from the specification of the mortar method itself.

We now introduce the finite element space

$$\overset{\circ}{\mathbf{W}}(F_l) = \left\{ \mathbf{w} \in [H_0^1(F_l)]^3 \ :\ \mathbf{w} = \mathbf{v}|_{F_l} \text{ for } \mathbf{v} \in \mathbf{X}_i \right\}, \tag{3.1}$$

where i denotes the index of the nonmortar subdomain of the face F_l and $\mathbf{v}|_{F_l}$ denotes the trace of $\mathbf{v}$ on the face F_l. This space is spanned by the nodal basis $\{\boldsymbol{\phi}_k\}_{k=1}^n$ of the nodes in F_l given by the triangulation T_i. Based on the space $\overset{\circ}{\mathbf{W}}(F_l)$, we construct a dual Lagrange multiplier space $\mathbf{M}(F_l)$ with a basis $\{\boldsymbol{\psi}_k\}_{k=1}^n$ satisfying

$$\int_{F_l} \boldsymbol{\phi}_m \cdot \boldsymbol{\psi}_k \, ds = \delta_{mk} \int_{F_l} \boldsymbol{\phi}_m \, ds \quad \forall\, m, k = 1, \ldots, n.$$

We refer to [13, 29] for a detailed construction of such a dual Lagrange multiplier space. The result of our paper is also applicable to the standard Lagrange multiplier space that was introduced in [2] for three dimensions. However, the dual Lagrange multiplier space leads to a computationally more efficient algorithm and also makes the implementation easier than it was with the older version. We note that both spaces contain the constant functions. The mortar matching condition is then written as

$$\int_{F_l} (\mathbf{v}_i - \boldsymbol{\phi}) \cdot \boldsymbol{\lambda} \, ds = 0 \quad \forall \boldsymbol{\lambda} \in \mathbf{M}(F_l),\ \forall F_l, \tag{3.2}$$

where $\mathbf{v}_i$ is a function from the nonmortar side and $\boldsymbol{\phi}$ is a function from the corresponding mortar parts. More precisely, $\boldsymbol{\phi} = \mathbf{v}_j$ on each $F_{ij} \subset F_l$.

We next introduce several finite element spaces,

$$\mathbf{X} = \prod_{i=1}^{N} \mathbf{X}_i, \tag{3.3}$$

$$\mathbf{W} = \prod_{i=1}^{N} \mathbf{W}_i,$$

$$\mathbf{W}_n = \prod_{F_l,\text{nonmortar}} \mathring{\mathbf{W}}(F_l),$$

$$\mathbf{M} = \prod_{F_l,\text{nonmortar}} \mathbf{M}(F_l).$$

Here the spaces $\mathbf{W}_n$ and $\mathbf{M}$ consist of functions defined on the nonmortar faces, while the spaces $\mathbf{W}$ and $\mathbf{X}$ consist of functions defined on both mortar and nonmortar faces, with elements that can be discontinuous across the interfaces. The space $\mathbf{W}$ can be considered as the trace space of the space $\mathbf{X}$ on the subdomain boundaries. In addition, we define the mortar finite element space,

$$\widehat{\mathbf{X}} = \{\mathbf{v} \in \mathbf{X} \,:\, \mathbf{v} \text{ satisfies (3.2)}\}.$$

The mortar discretization looks for an approximate solution of (2.1) in the space $\widehat{\mathbf{X}}$. In what follows, we derive a FETI-DP algorithm for the mortar discretization of (2.1).

3.2. Primal constraints in the FETI-DP formulation. We will build a FETI-DP algorithm that solves the model problem (2.1) in the space X defined in (3.3) by enforcing the mortar matching condition (3.2) across subdomain interfaces in dual and primal ways. We select some constraints from the mortar matching condition and enforce them strongly for the functions in the space X. We call them the primal constraints. The other constraints will be imposed weakly using Lagrange multipliers.

A proper selection of primal constraints is important for obtaining a scalable FETI-DP algorithm. In the work by Klawonn and Widlund [22] on elasticity problems, edge average and edge moment constraints and vertex constraints are selected. Furthermore, the concepts of an acceptable face path and an acceptable vertex path are introduced in an attempt to reduce the number of primal constraints.

We now select primal constraints from the mortar matching condition. We consider a nonmortar face $F \subset \partial\Omega_i$ and its partition $\{F_{ij}\}_j$ by mortar neighbors, i.e., $F_{ij} = \partial\Omega_i \cap \partial\Omega_j$. We define the space $\mathbf{M}(F_{ij})$ as a subspace of $\mathbf{M}(F)$ of functions that are supported in $\overline{F}_{ij}$. Six primal constraints are now introduced for each F_{ij} in the following way.

On each face F_{ij}, we first consider the six rigid body motions $\{\mathbf{r}_l\}_{l=1}^{6}$ as in (2.2) and (2.3), with H the diameter of the face F_{ij} and $\widehat{\mathbf{x}}$ a point on F_{ij}. We define a nodal interpolant $I_{M(F_{ij})} : [C(\overline{F}_{ij})]^3 \to \mathbf{M}(F_{ij})$ by

$$I_{M(F_{ij})}(\mathbf{v})(x) = \mathbf{v}(x), \quad x \in \mathbf{M}_{ij}^h,$$

where $\mathbf{M}_{ij}^h$ is the set of nodes corresponding to the Lagrange multiplier space $\mathbf{M}(F_{ij})$ and $C(\overline{F}_{ij})$ denotes the set of continuous functions on F_{ij}; see Figure 1 for a two dimensional case. We now select six primal constraints using the interpolated rigid

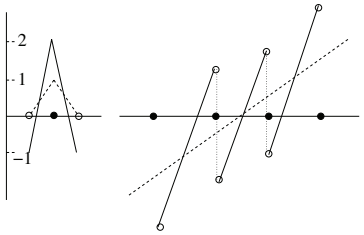


FIG. 1. *A dual Lagrange basis element (left, the solid line), a nodal basis element (left, the dashed line), and the nodal interpolant $I_{M(F_{ij})}(v)$ (right, the solid lines) to the linear function v (the dashed line).*

body motions,

$$\int_{F_{ij}} (\mathbf{v}_i - \mathbf{v}_j) \cdot I_{M(F_{ij})}(\mathbf{r}_l)\, ds = 0 \quad \forall l = 1, \ldots, 6.$$

Remark 3.5. When F_{ij} is the whole nonmortar face F, $\mathbf{M}(F_{ij})$ contains constant functions. The constraints with $\{I_{M(F_{ij})}(\mathbf{r}_l)\}_{l=1}^{3}$ are then nothing but the average matching condition across F_{ij} because $I_{M(F_{ij})}(\mathbf{r}_l) = \mathbf{r}_l$, for $l = 1, 2, 3$. The remaining constraints with $\{I_{M(F_{ij})}(\mathbf{r}_l)\}_{l=4}^{6}$, are similar to the moment matching constraints which were introduced for fully primal edges in [22] except that our constraints use the interpolated rotational rigid body motions and are enforced across faces. We call the constraints based on $\{I_{M(F_{ij})}(\mathbf{r}_l)\}_{l=4}^{6}$ the moment constraints.

Even though we have introduced a set of primal constraints to make the FETI-DP method more efficient, the enlarged coarse problem can be a bottleneck for the computation. To reduce the size of the coarse problem, we will select some interfaces as primal and impose the six constraints over them only. For the remaining constraint, the nonprimal faces, we assume that they each satisfy an acceptable face path condition. This assumption leads to a FETI-DP method, with primal constraints only for the primal interfaces, that has a condition number bound comparable to one obtained when the six primal constraints are imposed on all interfaces.

We now define an acceptable face path. Here, H_{ij} denotes the diameter of an interface F_{ij}.

DEFINITION 3.6 (acceptable face path). *Let L and TOL_P be given constants. For a pair of subdomains (Ω_i, Ω_j) having a common interface F_{ij}, an acceptable face path is a path*

$$\{\Omega_i, \Omega_{k_1}, \ldots, \Omega_{k_n}, \Omega_j\}$$

from Ω_i to Ω_j such that the coefficient G_{k_l} of Ω_{k_l} satisfies the condition

$$TOL_P * (1 + log(H_{ij}/h_i))^{-1} (1 + log(H_i/h_i))^2 * G_{k_l} \geq \min(G_i, G_j).$$

Moreover, the path from one subdomain to another must always be through a primal face, and the number of subdomains appearing in the path is bounded by the constant L.

3.3. The FETI-DP formulation with a change of basis. Let A_i be the stiffness matrix obtained from the finite element discretization of the bilinear form in the space $\mathbf{X}_i$,

$$a_i(\mathbf{u}_i, \mathbf{v}_i) := G_i \int_{\Omega_i} \varepsilon(\mathbf{u}_i) : \varepsilon(\mathbf{v}_i)\, dx + G_i \beta_i \int_{\Omega_i} \nabla \cdot \mathbf{u}_i \nabla \cdot \mathbf{v}_i\, dx \quad \forall \mathbf{u}_i, \mathbf{v}_i \in \mathbf{X}_i.$$

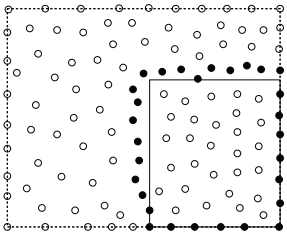


FIG. 2. *A primal interface F (solid line rectangle) of a face (dashed line rectangle) of subdomain Ω_i. The supports of the basis elements corresponding to the black nodes and the white nodes inside F intersect F; the values at the black nodes are unaffected by the transform T_F.*

Let S_i be the Schur complement of the matrix A_i that is obtained by eliminating the interior unknowns. We then write the mortar matching condition for $\mathbf{w} = (\mathbf{w}_1, \ldots, \mathbf{w}_N) \in W$ as

$$\sum_{i=1}^{N} B_i \mathbf{w}_i = 0.$$

We will express the matrices S_i and B_i in a new set of unknowns after a change of basis (unknowns). This idea was first presented in [21, 24] and algorithmic details were later described by Klawonn and Widlund [22]. The change of unknowns leads to a much simpler presentation of the algorithm as well as a more robust implementation; see [19, 18].

Let $\mathbf{w}_i$ denote the unknowns (or a function) in the space $\mathbf{W}_i$. We consider a primal interface $F \subset \partial\Omega_i$. Since the partition can be geometrically nonconforming, F might be only part of a face of Ω_i. Let $\mathbf{w}_F$ denote the restriction of a function $\mathbf{w}$ to a primal face $F \subset \Omega_i$. In other words, $\mathbf{w}_F$ is the vector of unknowns corresponding to the nodal basis elements with supports that intersect F; see Figure 2. For each primal face $F \subset \partial\Omega_i$, we define a transformation T_F by

$$\mathbf{w}_F = T_F \begin{pmatrix} \mathbf{w}_{F,\Pi} \\ \mathbf{w}_{F,\Delta} \end{pmatrix}, \tag{3.4}$$

where the six components of $\mathbf{w}_{F,\Pi}$ are given by

$$(\mathbf{w}_{F,\Pi})_l = \frac{\int_F \mathbf{w}_F \cdot I_{M(F)}(\mathbf{r}_l)\, ds}{H_F^2},\ l = 1, \ldots, 6, \tag{3.5}$$

and $\mathbf{v} = T_F\left(\begin{smallmatrix} 0 \\ \mathbf{w}_{F,\Delta} \end{smallmatrix}\right)$ has the six components of zero, i.e.,

$$\int_F \mathbf{v} \cdot I_{M(F)}(\mathbf{r}_l)\, ds = 0,\ l = 1, \ldots, 6. \tag{3.6}$$

Here H_F denotes the diameter of F. In addition, the transformation T_F retains the unknowns at the nodes other than those inside F. Since only the unknowns at the nodes inside F have been changed, each transform corresponding to an individual primal interface F can be applied independently.

We build such a transform T_F as follows. We first obtain six vectors $\{\boldsymbol{p}_l\}_{l=1}^6$ such that

$$\boldsymbol{p}_l^t \boldsymbol{w} = \frac{1}{H_F^2} \int_F \boldsymbol{w} \cdot I_{M(F)}(\mathbf{r}_l)\, ds,\ l = 1, \ldots, 6.$$

We then select six primal unknowns $\{w_{\Pi_l}\}_{l=1}^6$ among the unknowns of $\boldsymbol{w}$ at the interior of F so that $\boldsymbol{p}_l^t \boldsymbol{e}_{\Pi_l} \neq 0$, $l = 1, \ldots, 6$. Here $\boldsymbol{e}_{\Pi_l}$ is the coordinate unit vector with its value one at the location of the primal unknown v_{Π_l} and zero at the other unknowns. We then find six vectors $\{\boldsymbol{t}_{\Pi_l}\}_{l=1}^6$ in the space spanned by the vectors $\{\boldsymbol{e}_{\Pi_l}\}_{l=1}^6$ that are orthonormal to $\{\boldsymbol{p}_l\}_{l=1}^6$, i.e., $\boldsymbol{p}_l^t \boldsymbol{t}_{\Pi_k} = \delta_{lk}$, $l, k = 1, \ldots, 6$.

For the other unknowns $\{w_k\}_k$ of $\boldsymbol{w}$ at the interior of F, we build vectors $\mathbf{s}_k$ such that $(\boldsymbol{s}_k)_k = 1$, $(\boldsymbol{s}_k)_{\Pi_l} = a_{k,\Pi_l}$, $l = 1, \ldots, 6$, and the other components of $\boldsymbol{s}_k$ are zero. In other words, the components of $\boldsymbol{s}_k$ are zero except at the kth row and at the six rows corresponding to the primal unknowns. The values $\{a_{k,\Pi_l}\}_{l=1}^6$ are determined to satisfy $\boldsymbol{p}_l^t \boldsymbol{s}_k = 0$, $l = 1, \ldots, 6$. To build T_F, we first let T_F be the identity matrix and then replace each column of T_F by $\{\boldsymbol{t}_{\Pi_l}\}_{l=1}^6$ and $\{\boldsymbol{s}_k\}_k$ at the corresponding locations. We can see that this transform satisfies all of the requirements in (3.4)–(3.6) with $\{w_{\Pi_l}\}_{l=1}^6$ and $\{w_k\}_k$ as the six primal unknowns and the remaining dual unknowns, respectively. The resulting transform T_F is different from the identity matrix at the six rows and at the six columns corresponding to the primal unknowns.

The face-based primal constraints give change of basis involving unknowns at subdomain faces. Compared to the edge-based primal constraints in [22], our approach gives less sparse matrices after the change of basis. However, it does not significantly destroy the sparsity, since the transform T_F that we just introduced is different from the identity matrix at the six columns and the six rows corresponding to the primal unknowns.

After the change of unknowns, we order the unknowns $\mathbf{w}_i$ and the matrices S_i and B_i as

$$\mathbf{w}_i = \begin{pmatrix} \mathbf{w}_\Delta^{(i)} \\ \mathbf{w}_\Pi^{(i)} \end{pmatrix}, \quad S_i = \begin{pmatrix} S_{\Delta\Delta}^{(i)} & S_{\Delta\Pi}^{(i)} \\ S_{\Pi\Delta}^{(i)} & S_{\Pi\Pi}^{(i)} \end{pmatrix}, \quad B_i = \begin{pmatrix} B_\Delta^{(i)} & B_\Pi^{(i)} \end{pmatrix}, \tag{3.7}$$

where Π denotes the primal unknowns and Δ denotes the others, which we call the dual displacement unknowns. Any interface of a pair of subdomains is either a primal face or has an acceptable face path, so that each subdomain is connected to at least one neighbor through a primal face. This fact ensures that the matrix $S_{\Delta\Delta}^{(i)}$ is invertible, since the six primal constraints are linearly independent; a proof will be provided in Lemma 4.2.

According to the separation of the unknowns, the space $\mathbf{W}_i$ is decomposed into

$$\mathbf{W}_i = \mathbf{W}_\Delta^{(i)} \times \mathbf{W}_\Pi^{(i)},$$

where $\mathbf{W}_\Delta^{(i)}$ contains the dual unknowns and $\mathbf{W}_\Pi^{(i)}$ the primal unknowns. We further decompose the space $\mathbf{W}_\Delta^{(i)}$ of dual unknowns into

$$\mathbf{w}_\Delta^{(i)} = (\mathbf{w}_{\Delta,n}^{(i)}, \mathbf{w}_{\Delta,m}^{(i)}) \in \mathbf{W}_{\Delta,n}^{(i)} \times \mathbf{W}_{\Delta,m}^{(i)},$$

where n denotes the unknowns of the nonmortar faces and m denotes the remaining unknowns.

After enforcing the primal constraints on each primal face, we define the following space:

$$\widetilde{\mathbf{W}} = \{\mathbf{w} \in \mathbf{W} \,:\, \mathbf{w} \text{ satisfies the primal constraints across the primal faces}\}.$$

It can be decomposed into

$$\widetilde{\mathbf{W}} = \mathbf{W}_\Delta \times \mathbf{W}_\Pi,$$

where $\mathbf{W}_\Delta = \prod_{i=1}^N \mathbf{W}_\Delta^{(i)}$ and $\mathbf{W}_\Pi$ is the space of global primal unknowns. The space $\mathbf{W}_\Delta$ is further decomposed into

$$\mathbf{W}_\Delta = \mathbf{W}_{\Delta,n} \times \mathbf{W}_{\Delta,m},$$

where n and m refer to the nonmortar and the other part.

Throughout this paper, we use the same notation for a function and the corresponding vector of unknowns representing the function. Thus, $\mathbf{w}_\Delta^{(i)}$ denotes a vector of dual unknowns of $\mathbf{w}_i$ or the corresponding finite element function. The same convention applies to the spaces $\mathbf{W}_{\Delta,n}$, $\mathbf{W}$, $\mathbf{M}$, etc.

We now consider the mortar matching matrices B_i. When the mortar matching constraints are imposed on $\mathbf{w}$ in $\widetilde{\mathbf{W}}$, they are redundant, since the six primal constraints on each primal interface already have been enforced on $\mathbf{w}$. We therefore eliminate six equations from the rows of B_i for each primal interface $F \subset \partial\Omega_i$ and obtain the mortar matching constraints for the space $\widetilde{\mathbf{W}}$ that are nonredundant. We will use the same notation B_i after we have made them nonredundant.

By introducing Lagrange multipliers $\boldsymbol{\lambda}$ for the mortar matching constraints, we obtain the following mixed formulation of the problem (2.1):

$$\begin{pmatrix} S_{\Delta\Delta} & S_{\Delta\Pi} & B_\Delta^t \\ S_{\Pi\Delta} & S_{\Pi\Pi} & B_\Pi^t \\ B_\Delta & B_\Pi & 0 \end{pmatrix} \begin{pmatrix} \mathbf{w}_\Delta \\ \mathbf{w}_\Pi \\ \boldsymbol{\lambda} \end{pmatrix} = \begin{pmatrix} \mathbf{g}_\Delta \\ \mathbf{g}_\Pi \\ 0 \end{pmatrix}, \tag{3.8}$$

where each of the block matrices are obtained from subassembly of blocks of S_i and B_i in (3.7) at the global primal unknowns and at the dual unknowns.

The FETI-DP algorithm solves this mixed problem iteratively after eliminating all unknowns other than $\boldsymbol{\lambda}$. The elimination of the unknowns $\mathbf{w}_\Delta$ and $\mathbf{w}_\Pi$ leads to

$$F_{DP}\boldsymbol{\lambda} = \boldsymbol{d}. \tag{3.9}$$

We note that the matrix F_{DP} is symmetric and positive definite and satisfies the well-known relation (see [25, Lemma 4.3])

$$\langle F_{DP}\boldsymbol{\lambda}, \boldsymbol{\lambda}\rangle = \max_{\mathbf{w}\in\widetilde{\mathbf{W}}} \frac{\langle B\mathbf{w}, \boldsymbol{\lambda}\rangle^2}{\langle S\mathbf{w}, \mathbf{w}\rangle}, \tag{3.10}$$

where

$$S = \begin{pmatrix} S_{\Delta\Delta} & S_{\Delta\Pi} \\ S_{\Pi\Delta} & S_{\Pi\Pi} \end{pmatrix}, \quad B = \begin{pmatrix} B_\Delta & B_\Pi \end{pmatrix}.$$

The FETI-DP algorithm solves (3.9) for $\boldsymbol{\lambda}$ using a preconditioned conjugate gradient method with an appropriate preconditioner.

We now introduce a preconditioner $\widehat{M}^{-1}$ given by

$$\langle \widehat{M}\boldsymbol{\lambda}, \boldsymbol{\lambda}\rangle = \max_{\mathbf{w}_{\Delta,n}\in\mathbf{W}_{\Delta,n}} \frac{\langle BE(\mathbf{w}_{\Delta,n}), \boldsymbol{\lambda}\rangle^2}{\langle SE(\mathbf{w}_{\Delta,n}), E(\mathbf{w}_{\Delta,n})\rangle}, \tag{3.11}$$

where $E(\mathbf{w}_{\Delta,n})$ is the zero extension of $\mathbf{w}_{\Delta,n}$ into the space $\widetilde{\mathbf{W}}$. To be more precise,

$$E(\mathbf{w}_{\Delta,n}) = (\mathbf{w}_{\Delta,n}, 0, 0) \in \mathbf{W}_{\Delta,n} \times \mathbf{W}_{\Delta,m} \times \mathbf{W}_\Pi = \widetilde{\mathbf{W}}.$$

We then obtain

(3.12)
$$\langle \widehat{M}\boldsymbol{\lambda},\boldsymbol{\lambda}\rangle = \max_{\mathbf{w}_{\Delta,n}\in\mathbf{W}_{\Delta,n}} \frac{\langle BE(\mathbf{w}_{\Delta,n}),\boldsymbol{\lambda}\rangle^2}{\langle SE(\mathbf{w}_{\Delta,n}),E(\mathbf{w}_{\Delta,n})\rangle} \le \max_{\mathbf{w}\in\widetilde{\mathbf{W}}} \frac{\langle B\mathbf{w},\boldsymbol{\lambda}\rangle^2}{\langle S\mathbf{w},\mathbf{w}\rangle} = \langle F_{DP}\boldsymbol{\lambda},\boldsymbol{\lambda}\rangle.$$

Therefore the lower bound of the FETI-DP algorithm is bounded from below by 1.

The explicit form of the preconditioner

$$\widehat{M}^{-1} = \sum_{i=1}^{N} B_i D_i S_i D_i B_i^t \tag{3.13}$$

is similar to other FETI-DP preconditioners except that the weight matrix D_i is given differently. The weight matrix D_i to the preconditioner $\widehat{M}^{-1}$ in (3.11) is expressed by

$$D_i = \begin{pmatrix} ((B_{\Delta,n}^{(i)})^t B_{\Delta,n}^{(i)})^{-1} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

where $B_{\Delta,n}^{(i)}$ is the matrix with the columns of $B_{\Delta}^{(i)}$ corresponding to the unknowns in the space $\mathbf{W}_{\Delta,n}^{(i)}$; see (3.7). We note that $B_{\Delta,n}^{(i)}$ is square and invertible and that the weight matrix D_i is applied to the unknowns in the space $\mathbf{W}_i = \mathbf{W}_{\Delta,n}^{(i)} \times \mathbf{W}_{\Delta,m}^{(i)} \times \mathbf{W}_{\Pi}^{(i)}$. We can further express the preconditioner in a much simpler form,

$$\widehat{M}^{-1} = \sum_{i=1}^{N} \begin{pmatrix} \left(B_{\Delta,n}^{(i)\ t}\right)^{-1} & 0 & 0 \end{pmatrix} S_i \begin{pmatrix} B_{\Delta,n}^{(i)\ -1} \\ 0 \\ 0 \end{pmatrix}. \tag{3.14}$$

This form shows that multiplying $\widehat{M}^{-1}$ by a vector involves solving local elasticity problems with natural boundary condition on the nonmortar faces and the zero Dirichlet boundary condition on the other part. We call $\widehat{M}^{-1}$ the Neumann–Dirichlet preconditioner.

4. Condition number analysis. In this section, we will analyze the condition number bound of the proposed FETI-DP algorithm. First, we will construct functionals $\{f_l\}_{l=1}^6$, dual to the space $\mathbf{ker}(\varepsilon)$, that satisfy

$$f_m(\mathbf{r}_k) = \delta_{mk}, \quad m,k = 1,\dots,6,$$

$$|f_m(\mathbf{w})|^2 \le C\frac{\|\mathbf{w}\|_{0,F}^2}{H^2} \quad \text{for } \mathbf{w}\in[L^2(F)]^3. \tag{4.1}$$

Here $\{\mathbf{r}_k\}_{k=1}^6$ is a basis of $\mathbf{ker}\ (\varepsilon)$ with six rigid body motions scaled with respect to a face $F\subset\partial\Omega_i$; this means that we take $\widehat{\mathbf{x}}\in F$ and $H=\operatorname{diam}(F)$ in (2.2) and (2.3). Such a dual basis was introduced by Klawonn and Widlund [22].

We will now introduce six functionals which are closely related to the primal constraints applied across a primal face F,

$$g_l(\mathbf{w}) = \frac{\int_F \mathbf{w}\cdot I_{M(F)}(\mathbf{r}_l)\,ds}{H^2} \quad \text{for } \mathbf{w}\in[L^2(F)]^3,\ l=1,\dots,6. \tag{4.2}$$

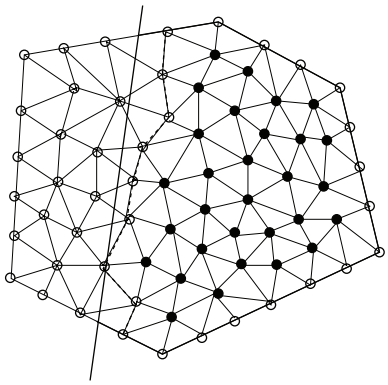


FIG. 3. *The face F, to the right of the solid line, is part of a nonmortar face, the whole face. The Lagrange multiplier basis to the black nodes are supported in F; $M(F)$ is the space spanned by these basis elements. The union of their supports is F_I, a subset of F that is to the right of the dashed lines.*

Here, $I_{M(F)}(\mathbf{r}_l)$ is the nodal interpolant to the Lagrange multiplier space $\mathbf{M}(F)$ provided for the face F and H is the diameter of F. For any real valued function g, we use the same notation $I_{M(F)}(g)$ for the nodal interpolant to the Lagrange multiplier space $M(F)$. We note that the vector valued Lagrange multiplier space is given by

$$\mathbf{M}(F) = [M(F)]^3.$$

LEMMA 4.1. *For any bounded function f and any linear function g, we have*

$$\left| \int_F f \left(g - I_{M(F)}(g) \right) \, ds \right| \leq ChH \|f\|_\infty \|g\|_\infty,$$

where H is the diameter of the face F and h is the mesh size of $M(F)$.

Proof. We consider the case when F is a part of a nonmortar face as in Figure 3. The subset F_I of F is the union of the supports of the basis elements in $M(F)$. Since g is linear, we obtain, at each nodal point $x \in F_I$,

$$|g(x) - I_{M(F)}(g)(x)| \leq Ch|g'| \leq C\frac{h}{H}\|g\|_\infty. \tag{4.3}$$

The above bound also holds for any $x \in F_I$, because both g and $I_{M(F)}(g)$ are linear in each triangle in F_I. We note that $I_{M(F)}(g)$ vanishes outside F_I.

We now have

$$\int_F f \left(g - I_{M(F)}(g) \right) \, ds = \int_{F_I} f \left(g - I_{M(F)}(g) \right) \, ds + \int_{F \setminus F_I} fg \, ds.$$

From the above bound and (4.3), using $|F_I| \leq CH^2$ and $|F \setminus F_I| \leq ChH$, the required bound then follows. ∎

LEMMA 4.2. *The functionals $\{g_l\}_{l=1}^6$ are linearly independent in the space $(ker(\varepsilon))'$, when h is sufficiently small.*

Proof. It suffices to show that the matrix G with the following entries is invertible:

$$G_{jk} = \int_F \mathbf{r}_j \cdot I_{M(F)}(\mathbf{r}_k) \, ds, \quad j,k = 1, \ldots, 6.$$

We first select $\widehat{x}_i$ in $\{\mathbf{r}_k\}_{k=4}^6$ so that

$$\int_F \frac{1}{H}(x_i - \widehat{x}_i)\, ds = 0, \quad i = 1, 2, 3.$$

The rigid body motions $\{\mathbf{r}_k\}_{k=4}^6$ are then orthogonal to $\{\mathbf{r}_k\}_{k=1}^3$ with respect to the inner product,

$$(\mathbf{r}, \boldsymbol{q})_F = \int_F \mathbf{r} \cdot \boldsymbol{q}\, ds.$$

We define $\widetilde{\mathbf{r}}_5$ and $\widetilde{\mathbf{r}}_6$ by

$$\widetilde{\mathbf{r}}_5 = \mathbf{r}_5 - a_4\mathbf{r}_4, \quad \widetilde{\mathbf{r}}_6 = \mathbf{r}_6 - b_4\mathbf{r}_4 - b_5\widetilde{\mathbf{r}}_5,$$

where

$$a_4 = \frac{(\mathbf{r}_4, \mathbf{r}_5)_F}{(\mathbf{r}_4, \mathbf{r}_4)_F},\ b_4 = \frac{(\mathbf{r}_4, \mathbf{r}_6)_F}{(\mathbf{r}_4, \mathbf{r}_4)_F},\ b_5 = \frac{(\widetilde{\mathbf{r}}_5, \mathbf{r}_6)_F}{(\widetilde{\mathbf{r}}_5, \widetilde{\mathbf{r}}_5)_F}.$$

The constants a_4, b_4, and b_5 are invariant to scaling. In other words, they are independent of H and h. We denote by $\{\widetilde{\mathbf{r}}_i\}_{i=1}^6$ the rigid body motions $\{\mathbf{r}_i\}_{i=1}^6$ with $\mathbf{r}_5$ and $\mathbf{r}_6$ replaced by $\widetilde{\mathbf{r}}_5$ and $\widetilde{\mathbf{r}}_6$. The rigid body motions $\{\widetilde{\mathbf{r}}_i\}_{i=1}^6$ are then orthogonal and the values of $(\widetilde{\mathbf{r}}_i, \widetilde{\mathbf{r}}_i)_F$ are scaling invariant. In other words, they are constants independent of H and h.

Using Lemma 4.1, we obtain

$$\left| \int_F \widetilde{\mathbf{r}}_j \cdot \widetilde{\mathbf{r}}_k\, ds - \int_F \widetilde{\mathbf{r}}_j \cdot I_{M(F)}(\widetilde{\mathbf{r}}_k)\, ds \right| \le ChH, \tag{4.4}$$

where C is a constant independent of H and h. We now consider a matrix $\widetilde{G}$ with entries

$$\widetilde{G}_{jk} = \int_F \widetilde{\mathbf{r}}_j \cdot I_{M(F)}(\widetilde{\mathbf{r}}_k), \quad j, k = 1, \ldots, 6.$$

Since the $\widetilde{\mathbf{r}}_i$ is orthogonal, we see, using the bound in (4.4), that $\widetilde{G}$ is diagonally dominant when h is sufficiently small. Therefore, $\widetilde{G}$ is invertible when h is small enough. The matrix G can be obtained from the invertible matrix $\widetilde{G}$ by using certain column and row operations and is therefore invertible. □

Remark 4.3. The proof of Lemma 4.2 also holds for an interface F that is not a flat surface.

Since the six functionals $\{g_l\}_{l=1}^6$ are linearly independent, they provide a basis of the dual space $(\mathbf{ker}(\varepsilon))'$. Thus there exists $\{\beta_{ml}\}_{m,l=1}^6$ such that

$$f_m = \sum_{l=1}^6 \beta_{ml} g_l, \quad m = 1, \ldots, 6. \tag{4.5}$$

Using Lemma 4.1, for any linear functions f and g we have

$$\left| \int_F fg\, ds - \int_F I_{M(F)}(f) I_{M(F)}(g)\, ds \right| \le C\|f\|_\infty \|g\|_\infty hH.$$

When h is small enough, we then find

$$(I_{M(F)}(\mathbf{r}_k), I_{M(F)}(\mathbf{r}_k))_F \leq C(\mathbf{r}_k, \mathbf{r}_k)_F \leq CH^2. \tag{4.6}$$

From (4.2), (4.6), and the Cauchy–Schwarz inequality, we obtain

$$|g_l(\mathbf{w})|^2 \leq C\frac{\|\mathbf{w}\|^2_{0,F}}{H^2}.$$

The inequality (4.1) follows from (4.5) and the above bound. We denote the dual functionals described above by $\{f_l^F\}_{l=1}^6$ for the given face F. We can then express any rigid body motion $\mathbf{r} \in \mathbf{ker}(\varepsilon)$ as a linear combination using the dual basis

$$\mathbf{r} = \sum_{l=1}^{6} f_l^F(\mathbf{r})\mathbf{r}_l.$$

In the following, we will provide several lemmas which will be used to provide an upper bound of the FETI-DP algorithm equipped with the preconditioner $\widehat{M}^{-1}$ given in (3.11); see also (3.14). We note that a lower bound with the constant 1 is provided in (3.12).

For a face $F \subset \partial\Omega_i$, the space $H_{00}^{1/2}(F)$ consists of the functions whose zero extension to the whole boundary $\partial\Omega_i$ belongs to the space $H^{1/2}(\partial\Omega_i)$; it is equipped with the norm

$$\|w\|_{H_{00}^{1/2}(F)} := \left(|w|^2_{1/2,F} + \int_F \frac{w(x)^2}{\operatorname{dist}(x,\partial F)}\, ds(x)\right)^{1/2},$$

where

$$|w|^2_{1/2,F} = \int_F\int_F \frac{|w(x)-w(y)|^2}{|x-y|^3}\, ds(x)ds(y).$$

In addition, a norm for the space $H^{1/2}(F)$ is defined by

$$\|w\|_{1/2,F} = \left(|w|^2_{1/2,F} + \frac{1}{H_F}\|w\|^2_{0,F}\right)^{1/2},$$

where H_F is the diameter of F.

These norms can be extended to the product spaces $[H_{00}^{1/2}(F)]^3$ and $[H^{1/2}(F)]^3$ by using the usual product norms. Similarly, we can extend the edge and face lemmas given below to product spaces. These lemmas can be found in Toselli and Widlund [27, Lemmas 4.24 and 4.25]. In the following, $W(F)$ denotes any conforming P_1-finite element space provided for a face F with its mesh size comparable to that of its nonmortar subdomain. Since the face F can be only part of a full face of its nonmortar subdomain Ω_i, the triangulation equipped for the face F can be different from that of Ω_i. We use $\mathbf{W}(F)$ to denote the corresponding vector valued finite element space, i.e., $\mathbf{W}(F) = [W(F)]^3$.

LEMMA 4.4 (edge lemma). *Let E be an edge of a face $F_{ij} \subset \partial\Omega_i$ and F_{ij} has Ω_i as its nonmortar subdomain. Then, for any $\mathbf{w} \in \mathbf{W}(F_{ij})$, we have*

$$\|\mathbf{w}\|^2_{0,E} \leq C\left(1+\log\frac{H_{ij}}{h_i}\right)\|\mathbf{w}\|^2_{1/2,F_{ij}}.$$

Let $C(\overline{F})$ be the space of continuous functions defined on F. For any subset $A \subset \overline{F}$, we define an interpolant $I_A : C(\overline{F}) \to W(F)$ by

$$I_A(w)(x) = \begin{cases} w(x) & \text{for } x \in A \cap N^h, \\ 0 & \text{for the other nodes,} \end{cases} \tag{4.7}$$

where N^h is the set of nodes of the finite element space $W(F)$. We note that $I_F(w)$, when $A = F$, vanishes at the boundary of F. We can extend the interpolant to a vector valued function $\mathbf{w} \in [C(\overline{F})]^3$ and we also denote it by $I_A(\mathbf{w})$.

LEMMA 4.5 (face lemma). *Let F_{ij} be a face of $\partial\Omega_i$ with Ω_i as its nonmortar subdomain. Then, for any $\mathbf{w} \in \mathbf{W}(F_{ij})$, we have*

$$\|I_{F_{ij}}(\mathbf{w})\|^2_{\mathbf{H}_{00}^{1/2}(F_{ij})} \le C\left(1 + \log\frac{H_{ij}}{h_i}\right)^2 \|\mathbf{w}\|^2_{1/2,F_{ij}}.$$

We now derive several inequalities for the mortar projection of functions. We recall that the spaces $\overset{\circ}{\mathbf{W}}(F_l)$ and $\mathbf{M}(F_l)$ are given on the nonmortar face F_l; see section 3.1.

DEFINITION 4.6 (mortar projection). *The mortar projection $\pi_l : [L^2(F_l)]^3 \to \overset{\circ}{\mathbf{W}}(F_l)$ is defined by*

$$\int_{F_l} (\pi_l(\mathbf{w}) - \mathbf{w}) \cdot \boldsymbol{\psi}\, ds = 0 \quad \forall \boldsymbol{\psi} \in \mathbf{M}(F_l).$$

It is known that the mortar projection is bounded in both the L^2- and the $H_{00}^{1/2}$-norms; see [1, 29].

For any function w defined on $F_{ij} \subset F_l$, we define $\chi_{F_{ij}} w$ by

$$\chi_{F_{ij}} w(x) = \begin{cases} w(x), & x \in F_{ij}, \\ 0, & x \in F_l \setminus F_{ij}. \end{cases}$$

LEMMA 4.7. *Let $F_l \subset \partial\Omega_i$ be a nonmortar face with its partition $\{F_{ij}\}_j$, and let $\mathbf{w}$ be a function in $[H^{1/2}(\partial\Omega_k)]^3$, $k = i, j$. When F_{ij} is substantial, we obtain*

$$\|\pi_l(\chi_{F_{ij}}\mathbf{w})\|^2_{H_{00}^{1/2}(F_l)} \le C\left(1 + \log\frac{H_{ij}}{h_i}\right)^2 \|\mathbf{w}\|^2_{1/2,\partial\Omega_k}, \quad k = i, j.$$

Proof. We may assume that $\mathbf{w} \in [H^{1/2}(\partial\Omega_j)]^3$. For the other case, the proof is the same. On the interface F_{ij}, we consider a quasiuniform triangulation of which mesh size is comparable to h_i, that of its nonmortar subdomain Ω_i, and denote the corresponding conforming P_1-finite element space by $\mathbf{W}(F_{ij})$. We then define the L^2-projection,

$$Q : [L^2(F_{ij})]^3 \to \mathbf{W}(F_{ij}). \tag{4.8}$$

We obtain

$$\|\pi_l(\chi_{F_{ij}}\mathbf{w})\|^2_{H_{00}^{1/2}(F_l)} \le 2\|\pi_l(\chi_{F_{ij}}(\mathbf{w}-Q\mathbf{w}))\|^2_{H_{00}^{1/2}(F_l)} + 2\|\pi_l(\chi_{F_{ij}}(Q\mathbf{w}))\|^2_{H_{00}^{1/2}(F_l)}. \tag{4.9}$$

The first term in the previous inequality is estimated by

$$\begin{aligned}\|\pi_l(\chi_{F_{ij}}(\mathbf{w}-Q\mathbf{w}))\|^2_{H^{1/2}_{00}(F_l)} &\le Ch_i^{-1}\|\pi_l(\chi_{F_{ij}}(\mathbf{w}-Q\mathbf{w}))\|^2_{0,F_l}\\ &\le Ch_i^{-1}\|(\mathbf{w}-Q\mathbf{w})\|^2_{0,F_{ij}}\\ &\le C\|\mathbf{w}\|^2_{1/2,\partial\Omega_j}. \qquad (4.10)\end{aligned}$$

Here we have used an inverse inequality, the continuity of π_l in the L^2-norm, and the approximation property of the projection Q combined with an interpolation argument (see [4, Chapter II])

$$\|\mathbf{w}-Q\mathbf{w}\|_{0,F_{ij}} \le Ch_i\|\mathbf{w}\|_{1,F_{ij}}, \quad \|Q\mathbf{w}\|_{1,F_{ij}} \le C\|\mathbf{w}\|_{1,F_{ij}}.$$

We now decompose $Q\mathbf{w}$ into interior and boundary parts, using the interpolant $I_A(\mathbf{w})$ defined in (4.7) to the space $\mathbf{W}(F_{ij})$,

$$Q\mathbf{w} = I_{F_{ij}}(Q\mathbf{w}) + I_{\partial F_{ij}}(Q\mathbf{w}).$$

Since $I_{F_{ij}}(Q\mathbf{w})$ is zero at the boundary of F_{ij}, we have

$$\chi_{F_{ij}}(I_{F_{ij}}(Q\mathbf{w})) \in [H^{1/2}_{00}(F_l)]^3, \quad I_{F_{ij}}(Q\mathbf{w}) \in [H^{1/2}_{00}(F_{ij})]^3.$$

Since F_{ij} is substantial, the definition of the $H^{1/2}_{00}$-norm states that

$$\|\chi_{F_{ij}}(I_{F_{ij}}(Q\mathbf{w}))\|_{H^{1/2}_{00}(F_l)} \le \|I_{F_{ij}}(Q\mathbf{w})\|_{H^{1/2}_{00}(F_{ij})}.$$

The second term in (4.9) is then bounded by

$$\begin{aligned}&\|\pi_l(\chi_{F_{ij}}(Q\mathbf{w}))\|^2_{H^{1/2}_{00}(F_l)}\\ &\le 2\|\pi_l(\chi_{F_{ij}}(I_{F_{ij}}(Q\mathbf{w})))\|^2_{H^{1/2}_{00}(F_l)} + 2\|\pi_l(\chi_{F_{ij}}(I_{\partial F_{ij}}(Q\mathbf{w})))\|^2_{H^{1/2}_{00}(F_l)}\\ &\le C\left(\|\chi_{F_{ij}}(I_{F_{ij}}(Q\mathbf{w}))\|^2_{H^{1/2}_{00}(F_l)} + h_i^{-1}\|(I_{\partial F_{ij}}(Q\mathbf{w}))\|^2_{0,F_{ij}}\right)\\ &\le C\left(\|I_{F_{ij}}(Q\mathbf{w})\|^2_{H^{1/2}_{00}(F_{ij})} + h_i^{-1}h_i\|(I_{\partial F_{ij}}(Q\mathbf{w}))\|^2_{0,\partial F_{ij}}\right). \qquad (4.11)\end{aligned}$$

Here we have used an inverse inequality and the continuity of π_l in the L^2- and $H^{1/2}_{00}$-norms.

By applying Lemmas 4.5 and 4.4 to the two terms in (4.11) and using that F_{ij} is substantial, we obtain

$$\begin{aligned}\|\pi_l(\chi_{F_{ij}}(Q\mathbf{w}))\|^2_{H^{1/2}_{00}(F_l)} &\le C\left(1+\log\frac{H_{ij}}{h_i}\right)^2\|Q\mathbf{w}\|^2_{1/2,F_{ij}}\\ &\le C\left(1+\log\frac{H_i}{h_i}\right)^2\|\mathbf{w}\|^2_{1/2,\partial\Omega_j}. \qquad (4.12)\end{aligned}$$

We obtain the required bound from (4.9)–(4.12). ∎

Remark 4.8. In our recent work [15], also on geometrically nonconforming partitions, progress has been made. The same bound, $C(1+\log(H/h))^2$, was proved for

the interfaces F_{ij} that do not need to be substantial, i.e., can be a narrow interface or only a small part of its nonmortar face F_l.

The following lemma is a simple modification of a result in Dryja, Smith, and Widlund [6, Lemma 4.4].

LEMMA 4.9. *Let F_{ij} be a face and $I_{F_{ij}}(\phi)$ be the nodal interpolant of ϕ to the finite element space equipped for F_{ij}. For a linear function ϕ, we have*

$$\|I_{F_{ij}}(\phi)\|^2_{H^{1/2}_{00}(F_{ij})} \leq C\left(1+\log\frac{H_{ij}}{h_i}\right)H_{ij}\|\phi\|^2_{\infty,F_{ij}},$$

where h_i is the mesh size and H_{ij} is the diameter of F_{ij}.

LEMMA 4.10. *Let $F_{ij}\subset F_l$, where F_l is a nonmortar face of $\partial\Omega_i$. For a linear function ϕ and a substantial interface F_{ij}, we have*

$$\|\pi_l(\chi_{F_{ij}}\phi)\|^2_{H^{1/2}_{00}(F_l)} \leq C\left(1+\log\frac{H_{ij}}{h_i}\right)H_{ij}\|\phi\|^2_{\infty,F_{ij}}.$$

Proof. By applying the estimate in (4.11) of Lemma 4.7 to ϕ, we have

$$\|\pi_l(\chi_{F_{ij}}(Q\phi))\|^2_{H^{1/2}_{00}(F_l)} \leq C\left(\|I_{F_{ij}}(Q\phi)\|^2_{H^{1/2}_{00}(F_{ij})}+\|I_{\partial F_{ij}}(Q\phi)\|^2_{0,\partial F_{ij}}\right),$$

where Q is the L^2-projection described in (4.8). Since $Q\phi=\phi$, the required bound follows by using Lemma 4.9. ∎

LEMMA 4.11. *Let $F_l\subset\partial\Omega_i$ be a nonmortar face with a partition $\{F_{ij}\}_j$ by its mortar neighbors. We assume that all interfaces F_{ij} are substantial and each interface F_{ij} is either primal or has an acceptable face path. In addition, the subdomain partition satisfies Assumption* 3.1. *Then, for $\mathbf{w}=(\mathbf{w}_1,\dots,\mathbf{w}_N)\in\widetilde{\mathbf{W}}$, we have*

$$\begin{aligned}G_i\|\pi_l(\mathbf{w}_i-\boldsymbol{\phi})\|^2_{H^{1/2}_{00}(F_l)} \leq C\Bigg\{&\left(1+\log\frac{H_i}{h_i}\right)^2|\mathbf{w}_i|^2_{S_i}\\ &+\sum_j L(F_{ij})*\sum_{k\in A(i,j)}\left(1+\log\frac{H_{ij}}{h_i}\right)\frac{G_i}{G_k}|\mathbf{w}_k|^2_{S_k}\\ &+\sum_j\frac{G_i}{G_j}\left(1+\log\frac{H_i}{h_i}\right)^2|\mathbf{w}_j|^2_{S_j}\Bigg\}.\end{aligned}$$

Here the restriction of $\boldsymbol{\phi}$ to F_{ij} is $\mathbf{w}_j$, $A(i,j)$ is the set of subdomain indices of the acceptable face path of F_{ij}, the constant $L(F_{ij})$ is the number of the subdomains in the path, and $|\mathbf{w}_i|^2_{S_i}:=\langle S_i\mathbf{w}_i,\mathbf{w}_i\rangle$.

Proof. Since $\{F_{ij}\}_j$ is a partition of the nonmortar face F_l, we can write

$$\mathbf{w}_i-\boldsymbol{\phi}=\sum_j\chi_{F_{ij}}\left(\mathbf{w}_i-\mathbf{w}_j\right),$$

and it suffices to estimate each term in the above expression.

Let $\{\Omega_i,\Omega_{k_1},\dots,\Omega_{k_n},\Omega_j\}$ be the related acceptable face path of F_{ij} which passes through the primal interfaces $\{F_{ik_1},F_{k_1k_2},\dots,F_{k_nj}\}$. When F_{ij} is primal, the path is simply $\{\Omega_i,\Omega_j\}$. Let $\{\mathbf{r}^{ik_1}_m\}$, $\{\mathbf{r}^{k_1k_2}_m\},\dots,\{\mathbf{r}^{k_nj}_m\}$ be bases of $\mathbf{ker}(\varepsilon)$ scaled with respect

to the faces F_{ik_1}, $F_{k_1k_2},\ldots,F_{k_nj}$, respectively. We denote the dual basis to $\{\mathbf{r}_m^{lk}\}_m$ by $\{f_m^{lk}\}_m$. We introduce the notation,

$$f_{lk}(\mathbf{w}) = \sum_{m=1}^{6} f_m^{lk}(\mathbf{w})\mathbf{r}_m^{lk}.$$

We note that $f_{lk}(\mathbf{r}) = \mathbf{r}$ for any rigid body motion $\mathbf{r}$ and that $f_{lk}(\mathbf{w}_l) = f_{lk}(\mathbf{w}_k)$, since $\mathbf{w}$ satisfies the primal constraints on the primal interface F_{lk}.

We then have

$$\begin{aligned}\mathbf{w}_i - \mathbf{w}_j &= (\mathbf{w}_i - \mathbf{r}_i) - f_{ik_1}(\mathbf{w}_i - \mathbf{r}_i) + f_{ik_1}(\mathbf{w}_{k_1} - \mathbf{r}_{k_1}) - f_{k_1k_2}(\mathbf{w}_{k_1} - \mathbf{r}_{k_1})\\ &\quad + f_{k_1k_2}(\mathbf{w}_{k_2} - \mathbf{r}_{k_2}) - f_{k_2k_3}(\mathbf{w}_{k_2} - \mathbf{r}_{k_2}) + \cdots\\ &\quad + f_{k_nj}(\mathbf{w}_j - \mathbf{r}_j) - (\mathbf{w}_j - \mathbf{r}_j),\end{aligned} \tag{4.13}$$

where $\mathbf{r}_k$ denotes any rigid body motions.

We apply Lemma 4.7 to the first and last terms of the above equation and obtain

$$\begin{aligned}&G_i\left(\|\pi_l(\chi_{F_{ij}}(\mathbf{w}_i - \mathbf{r}_i))\|^2_{H_{00}^{1/2}(F_l)} + \|\pi_l(\chi_{F_{ij}}(\mathbf{w}_j - \mathbf{r}_j))\|^2_{H_{00}^{1/2}(F_l)}\right)\\ &\le C\left(1+\log\frac{H_{ij}}{h_i}\right)^2\left(G_i\|\mathbf{w}_i - \mathbf{r}_i\|^2_{1/2,\partial\Omega_i} + G_i\|\mathbf{w}_j - \mathbf{r}_j\|^2_{1/2,\partial\Omega_j}\right)\\ &\le C\left(1+\log\frac{H_{ij}}{h_i}\right)^2\left(|\mathbf{w}_i|^2_{E(\partial\Omega_i)} + \frac{G_i}{G_j}|\mathbf{w}_j|^2_{E(\partial\Omega_j)}\right)\\ &\le C\left(1+\log\frac{H_i}{h_i}\right)^2\left(|\mathbf{w}_i|^2_{S_i} + \frac{G_i}{G_j}|\mathbf{w}_j|^2_{S_j}\right).\end{aligned}$$

Here we have also used Lemmas 2.2 and 2.3 combined with the bound

$$|\mathbf{w}_i|^2_{E(\partial\Omega_i)} \le |\mathbf{w}_i|^2_{S_i},$$

and the assumption that the interface F_{ij} is substantial.

We next consider the other terms in (4.13). From (4.1) and Lemmas 4.10 and 2.3, each term can be estimated as

$$\begin{aligned}&G_i|f_m^{k_1k_2}(\mathbf{w}_{k_2} - \mathbf{r}_{k_2})|^2\ \|\pi_l(\chi_{F_{ij}}\mathbf{r}_m^{k_1k_2})\|^2_{H_{00}^{1/2}(F_l)}\\ &\le CG_i\frac{\|\mathbf{w}_{k_2} - \mathbf{r}_{k_2}\|^2_{0,\partial\Omega_{k_2}}}{H^2_{k_1k_2}}\left(1+\log\frac{H_{ij}}{h_i}\right)H_{ij}\|\mathbf{r}_m^{k_1k_2}\|^2_{\infty,F_{ij}}\\ &\le C*\frac{G_i}{G_{k_2}}\left(1+\log\frac{H_{ij}}{h_i}\right)|\mathbf{w}_{k_2}|^2_{S_{k_2}}\frac{H_{k_2}H_{ij}}{H^2_{k_1k_2}}\|\mathbf{r}_m^{k_1k_2}\|^2_{\infty,F_{ij}}.\end{aligned}$$

We will show that the factor $(H_{k_2}H_{ij}/H^2_{k_1k_2})\|\mathbf{r}_m^{k_1k_2}\|^2_{\infty,F_{ij}}$ is bounded by a constant that depends only on $L(F_{ij})$. The assumption of the acceptable face path gives that $L(F_{ij}) \le L$ for a given constant L. Since the length of the face path is less than or equal to L and the subdomain partition is locally quasiuniform, the diameter of

the subdomains in the path is comparable to H_i. In addition, the diameter of any primal interface in the path is comparable to H_i, because it is substantial. From these observations, we obtain

$$\|\mathbf{r}_m^{k_1k_2}\|_{\infty,F_{ij}} \leq CL(F_{ij}), \quad \frac{H_{k_2}H_{ij}}{H_{k_1k_2}^2} \leq C\frac{H_{ij}}{H_i},$$

and have proved the bound

$$\frac{H_{k_2}H_{ij}}{H_{k_1k_2}^2}\|\mathbf{r}_m^{k_1k_2}\|_{\infty,F_{ij}}^2 \leq CL(F_{ij})^2. \tag{4.14}$$

The remaining terms in (4.13) can be bounded in a similar way leading to the required bound of $G_i\|\pi_l(\mathbf{w}_i - \boldsymbol{\phi})\|^2_{H_{00}^{1/2}(F_l)}$. ∎

Remark 4.12. The proof for the bound in (4.14) suggests that it is beneficial that the diameter of the nonprimal interface F_{ij} is smaller than those of the primal interfaces in the acceptable face path. In addition, the definition of an acceptable face path shows that a smaller diameter of F_{ij} provides more chances of finding an acceptable path of F_{ij} for a given TOL_P and L.

Remark 4.13. From our recent work [15], the same bound in Lemma 4.11 can also be obtained when F_{ij} is not substantial, i.e., can be a narrow strip or only a small part of its nonmortar face F_l.

We note that the acceptable face path in Definition 3.6 and Assumption 3.3 gives

$$\left(1+\log\frac{H_{ij}}{h_i}\right)\frac{G_i}{G_{k_l}} \leq TOL_P * \left(1+\log\frac{H_i}{h_i}\right)^2, \quad L(F_{ij}) \leq L, \quad \frac{G_i}{G_j} \leq TOL_G,$$

where $L(F_{ij})$ is the number of subdomains in the acceptable face path of F_{ij}. Therefore the bound in Lemma 4.11 is reduced to

$$G_i\|\pi_l(\mathbf{w}_i - \boldsymbol{\phi})\|^2_{H_{00}^{1/2}(F_l)} \leq C\left(1+\log\frac{H_i}{h_i}\right)^2 \sum_{l\in N_l}|\mathbf{w}_l|^2_{S_l}, \tag{4.15}$$

where N_l is the set of subdomain indices that appear on the acceptable face path of $F_{ij} \subset F_l$. The constant C depends on TOL_G, TOL_P, and L but does not depend on any mesh parameters nor on the coefficients G_i.

LEMMA 4.14. *Assume that every nonprimal interface satisfies the acceptable face path condition with given TOL_P, L, and coefficients G_i, and that all interfaces are substantial. In addition, Assumptions* 3.1 *and* 3.3 *hold with a given TOL_G. We then obtain*

$$\max_{\mathbf{w}\in\widetilde{\mathbf{W}}}\frac{\langle B\mathbf{w},\boldsymbol{\lambda}\rangle^2}{\langle S\mathbf{w},\mathbf{w}\rangle} \leq C\max_{i=1,\ldots,N}\left\{\left(1+log\frac{H_i}{h_i}\right)^2\right\}\langle\widehat{M}\boldsymbol{\lambda},\boldsymbol{\lambda}\rangle,$$

where the constant C depends on the TOL_G, TOL_P, and L, but not on any mesh parameters and coefficients G_i.

Proof. We consider

$$\langle B\mathbf{w},\boldsymbol{\lambda}\rangle^2 = \left(\sum_{F_l,\text{nonmortar}}\int_{F_l}(\mathbf{w}_i-\boldsymbol{\phi})\cdot\boldsymbol{\lambda}\,ds\right)^2$$

$$= \left(\sum_{F_l,\ \text{nonmortar}}\int_{F_l}\pi_l(\mathbf{w}_i-\boldsymbol{\phi})\cdot\boldsymbol{\lambda}\,ds\right)^2.$$

Since $\mathbf{w}\in\widetilde{\mathbf{W}}$, $\mathbf{w}$ satisfies the primal constraints on any primal interface F_{ij}, i.e.,

$$\int_{F_{ij}}(\mathbf{w}_i-\boldsymbol{\phi})\cdot I_{M(F_{ij})}(\mathbf{r}_m)\,ds=0,\quad m=1,\dots,6.$$

This implies that $\pi_l(\mathbf{w}_i-\boldsymbol{\phi})$ also satisfies the primal constraints

$$\int_{F_{ij}}\pi_l(\mathbf{w}_i-\boldsymbol{\phi})\cdot I_{M(F_{ij})}(\mathbf{r}_m)\,ds=0,\quad m=1,\dots,6,$$

because $I_{M(F_{ij})}(\mathbf{r}_m)$ belongs to the Lagrange multiplier space $\mathbf{M}(F_l)$. Therefore, $\mathbf{z}_n$, defined on each nonmortar face F_l by

$$\mathbf{z}_n|_{F_l}=\pi_l(\mathbf{w}_i-\boldsymbol{\phi}),$$

belongs to $\mathbf{W}_{\Delta,n}$. In other words, $\mathbf{z}_n$ has all its six primal components zero on each primal face. We then define by $\mathbf{z}=E(\mathbf{z}_n)$ the extension of $\mathbf{z}_n$ to $\widetilde{\mathbf{W}}(=\mathbf{W}_{\Delta,n}\times\mathbf{W}_{\Delta,m}\times\mathbf{W}_{\Pi})$ by zero.

By the definition of $\widehat{M}$ in (3.11), we find that

$$\langle B\mathbf{w},\boldsymbol{\lambda}\rangle^2=\langle B\mathbf{z},\boldsymbol{\lambda}\rangle^2$$

$$\le\langle\widehat{M}\boldsymbol{\lambda},\boldsymbol{\lambda}\rangle\langle S\mathbf{z},\mathbf{z}\rangle.$$

It suffices to show that

$$\langle S\mathbf{z},\mathbf{z}\rangle\le C\max_{i=1,\dots,N}\left\{\left(1+\log\frac{H_i}{h_i}\right)^2\right\}\langle S\mathbf{w},\mathbf{w}\rangle. \tag{4.16}$$

We now consider

$$\langle S\mathbf{z},\mathbf{z}\rangle=\sum_{i=1}^{N}\langle S_i\mathbf{z}_i,\mathbf{z}_i\rangle$$

$$\le C\sum_{i=1}^{N}\sum_{F_l,\ \text{nonmortar}}G_i|\mathcal{H}^i\left(\pi_l(\mathbf{w}_i-\boldsymbol{\phi})\right)|^2_{1,\Omega_i}$$

$$\le C\sum_{i=1}^{N}\sum_{F_l,\ \text{nonmortar}}G_i\|\pi_l(\mathbf{w}_i-\boldsymbol{\phi})\|^2_{H_{00}^{1/2}(F_l)}, \tag{4.17}$$

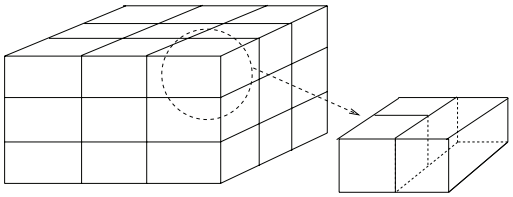


FIG. 4. *Partition of the domain into uniform parallelepipes when $N = 3$ (left), partition of each parallelepipe into geometrically nonconforming subdomains with good aspect ratio of interfaces (right).*

where $\mathcal{H}^i$ is the discrete harmonic extension into X_i and π_l is the mortar projection corresponding to the nonmortar face F_l. Here we have used that (see Lemma 2.2 and (2.4))

$$\langle S_i \mathbf{z}_i, \mathbf{z}_i \rangle \leq C G_i |\mathcal{H}^i(\mathbf{z}_i)|^2_{1,\Omega_i}.$$

From the bound (4.15) and (4.17), we obtain (4.16) with a constant C which depends on TOL_G, TOL_P, and L. □

The lower bound in (3.12) and the bound in Lemma 4.14, combined with (3.10), lead to the following condition number bound.

THEOREM 4.15. *We assume that the assumptions in Lemma* 4.14 *hold. We then obtain the condition number bound*

$$\kappa(\widehat{M}^{-1} F_{DP}) \leq C \max_{i=1,\ldots,N} \left\{ \left(1 + log \frac{H_i}{h_i} \right)^2 \right\}.$$

Here the constant C is independent of the mesh parameters and the coefficients G_i, but depends on the given constants TOL_G, TOL_P, and L.

Remark 4.16. We note that the result can be extended to the general case when all interfaces do not need to be substantial; see [15] and numerical results provided in the following section.

5. Numerical results. We perform numerical tests for the model elasticity problem (2.1) in a domain $\Omega = [0,1] \times [0,1] \times [0,1/2]$ and the material parameter $\beta(x) = 1$. The coefficient $G(x)$ can be heterogeneous. We let $\mathbf{F}$ be zero so that the exact solution of our model problem is zero. We start the FETI-DP iteration with a nonzero initial and terminate the iteration when the relative residual norm becomes less than a factor of 10^6.

We first test the performance of the suggested algorithm in regards to high contrast of the coefficients for the subdomain partitions of which interfaces are substantial and have a good aspect ratio. We obtain such geometrically nonconforming partitions in the following way. For a given N, we partition Ω into N^3 uniform parallelepipes and then partition each parallelepipe into geometrically nonconforming subdomains as in Figure 4. All interfaces are then substantial. We set $G(x) = G_i$ in each subdomain Ω_i. The values of G_i can be 1, 100, or 1000.

We consider the uniform coefficient case and the high contrast case as in Figure 5. For the high contrast case, we select the nonmortar and mortar sides to satisfy Assumption 3.3 or not for a given TOL_G. For the three cases, the number of iterations and condition numbers are presented in Table 1 as an increase of the number of subdomains and in Table 2 as an increase of the local problem size. We can see that

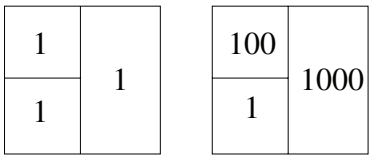


FIG. 5. *Values of G_i: uniform coefficients (left), high contrast of coefficients (right).*

TABLE 1

Performance of the FETI-DP algorithm regarding various coefficient selection as the increase of the number of subdomains (N) with good aspect ratio of interfaces and with a fixed local problem size ($n = 6$); Cond (condition number), Iter (the number of iterations).

	Uniform const.		With Assumption 3.3		Without Assumption 3.3	
N	Cond	Iter	Cond	Iter	Cond	Iter
3	21.54	26	20.79	25	7.542 x10^3	383
4	22.69	27	23.97	27	7.695 x10^3	404
5	23.24	27	26.12	28	7.751 x10^3	408
6	23.48	27	27.36	29	7.777 x10^3	411

TABLE 2

Performance of the FETI-DP algorithm regarding various coefficient selection as the increase of the local problem size n in a fixed subdomain partition with $N = 3$ and with good aspect ratio of interfaces; Cond (condition number), Iter (the number of iterations).

	Uniform const.		With Assumption 3.3		Without Assumption 3.3	
n	Cond	Iter	Cond	Iter	Cond	Iter
6	21.54	26	20.79	25	7.542 x10^3	383
8	27.59	31	25.34	27	8.869 x10^3	494
10	32.26	34	29.20	29	9.809 x10^3	542
12	36.16	36	32.41	31	1.055 x10^4	578

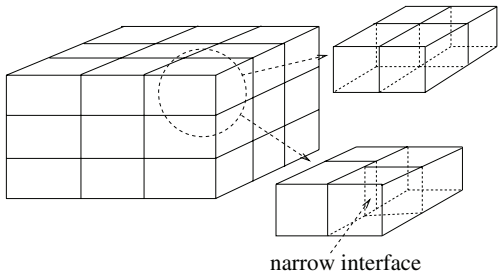


FIG. 6. *Partition of the domain into uniform parallelepipes when $N = 3$ (left), partition of each parallelepipe into geometrically conforming subdomains (right up), and into geometrically nonconforming subdomains with a bad aspect ratio of interfaces, i.e., narrow interfaces (right down).*

Assumption 3.3 is essential to achieving a good performance. In addition, the high contrast case gives results as good as those from the uniform case once its selection of nonmortar and mortar faces fulfills Assumption 3.3. The condition numbers agree well to our theory provided in section 4. In fact, these results show that the estimate of the condition number bound is optimal.

In our second test, we verify that our algorithm is also robust to partitions with a bad aspect ratio of interfaces that are not substantial. For a given N, we partition Ω into N^3 uniform parallelepipes and then partition each parallelepipe into four

TABLE 3

Performance of the FETI-DP algorithm when $G(x) = 1$ as the increase of n; conforming partition with good aspect ratio ($\alpha = 1.0$) of interfaces vs. geometrically nonconforming partition with bad aspect ratio ($\alpha = 3/(n-1)$); Cond (condition number), Iter (the number of iterations), nonprimal (bad aspect ratio interfaces are nonprimal), primal (bad aspect ratio interfaces are primal).

	Good aspect ratio		Bad aspect ratio (nonprimal)		Bad aspect ratio (primal)	
n	Cond	Iter	Cond	Iter	Cond	Iter
8	27.63	27	30.22	30	29.82	30
9	30.08	29	32.10	32	31.68	31
10	32.44	31	33.90	33	33.45	32
11	34.34	32	35.61	33	35.10	33
12	36.23	33	37.26	34	36.65	33

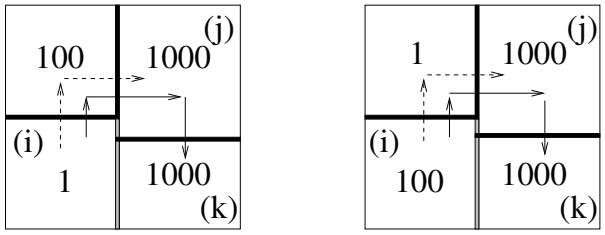


FIG. 7. *The acceptable face path assumption satisfied with a smaller TOL_P (left) and with a larger TOL_P (right); black interfaces are primal and gray interfaces are nonprimal; acceptable face path of $F_{ij} = \partial\Omega_i \cap \partial\Omega_j$ (dashed arrows), acceptable face path of $F_{ik} = \partial\Omega_i \cap \partial\Omega_k$ (solid arrows).*

TABLE 4

Performance of the FETI-DP algorithm depending on TOL_P in the acceptable face path assumption as the increase of the local problem size (n) in a geometrically nonconforming partition with bad aspect ratio ($\alpha = 2/(n-1)$); Cond (condition number), Iter (the number of iterations).

	Smaller TOL_P		Larger TOL_P	
n	Cond	Iter	Cond	Iter
8	26.40	30	57.58	38
9	27.94	31	60.30	40
10	29.35	32	62.71	41
11	30.65	33	64.89	41
12	31.85	34	66.88	42

subdomains that are geometrically conforming and geometrically nonconforming; see Figure 6. In the geometrically conforming case, the aspect ratio is always 1, while the aspect ratio gets worse as the increase of the local problem size in the geometrically nonconforming case. Here, the coefficients are uniform, i.e., $G(x) = 1$. Table 3 shows the number of iterations and condition numbers as the increase of the local problem size. We note that the aspect ratio gets worse as the local problem size increases. In the geometrically nonconforming case, we test our algorithm for the two cases when the narrow interfaces are nonprimal and when they are primal, respectively. From the result, we can see that the narrow interfaces are harmless to the performance of our algorithm regardless of selecting them as nonprimal interfaces or primal interfaces. Moreover, the narrow interface case also shows the optimal behavior of the condition number bound as the increase of the local problem size. In fact, there is only a slight increase in the number of iterations and condition numbers compared to the case of good aspect ratio of interfaces.

In the final test, we investigate whether or not the face path assumption is nec-

essary for subdomain partitions with bad aspect ratio of interfaces and with a high contrast of coefficients. We consider two cases that satisfy the face path assumption with a smaller TOL_P and with a larger TOL_P, respectively; see Figure 7 and Definition 3.6. Numerical results in Table 4 show that the larger value of TOL_P in the face path assumption gives a worse performance. Therefore, the face path assumption is in fact necessary to handle the bad aspect ratio interfaces with the high contrast of coefficients.

Acknowledgment. The author is deeply grateful to Professor Olof B. Widlund at Courant Institute for his support and encouragement, as well as valuable discussions.

REFERENCES

[1] F. B. Belgacem, *The mortar finite element method with Lagrange multipliers*, Numer. Math., 84 (1999), pp. 173–197.

[2] F. B. Belgacem and Y. Maday, *The mortar element method for three dimensional finite elements*, RAIRO Model. Math. Anal. Numer., 31 (1997), pp. 289–302.

[3] C. Bernardi, Y. Maday, and A. T. Patera, *A new nonconforming approach to domain decomposition: The mortar element method*, in Nonlinear Partial Differential Equations and Their Applications, Collège de France Seminar, Vol. XI (Paris, 1989–1991), Pitman Res. Notes Math. Ser. 299, Longman Sci. Tech., Harlow, 1994, pp. 13–51.

[4] D. Braess, *Finite Elements*, Theory, Fast Solvers, and Applications in Solid Mechanics, Cambridge University Press, Cambridge, 1997; translated from the 1992 German original by Larry L. Schumaker.

[5] Y.-W. Chang, H. H. Kim, and C.-O. Lee, *Preconditioners for the dual-primal FETI methods on nonmatching grids: Numerical study*, Comput. Math. Appl., 51 (2006), pp. 697–712.

[6] M. Dryja, B. F. Smith, and O. B. Widlund, *Schwarz analysis of iterative substructuring algorithms for elliptic problems in three dimensions*, SIAM J. Numer. Anal., 31 (1994), pp. 1662–1694.

[7] M. Dryja and O. B. Widlund, *A FETI-DP method for a mortar discretization of elliptic problems*, in Recent Developments in Domain Decomposition Methods (Zürich, 2001), Lect. Notes Comput. Sci. Eng. 23, Springer, Berlin, 2002, pp. 41–52.

[8] M. Dryja and O. B. Widlund, *A generalized FETI-DP method for a mortar discretization of elliptic problems*, in Domain Decomposition Methods in Science and Engineering (Cocoyoc, Mexico, 2002), UNAM, Mexico City, 2003, pp. 27–38.

[9] C. Farhat, M. Lesoinne, P. LeTallec, K. Pierson, and D. Rixen, *FETI-DP: A dual-primal unified FETI method*. I. *A faster alternative to the two-level FETI method*, Internat. J. Numer. Methods Engrg., 50 (2001), pp. 1523–1544.

[10] C. Farhat, M. Lesoinne, and K. Pierson, *A scalable dual-primal domain decomposition method*, Numer. Linear Algebra Appl., 7 (2000), pp. 687–714.

[11] C. Farhat and J. Mandel, *The two-level FETI method for static and dynamic plate problems*. I. *An optimal iterative solver for biharmonic systems*, Comput. Methods Appl. Mech. Engrg., 155 (1998), pp. 129–151.

[12] C. Farhat and F.-X. Roux, *A method of finite element tearing and interconnecting and its parallel solution algorithm*, Internat. J. Numer. Methods Engrg., 32 (1991), pp. 1205–1227.

[13] C. Kim, R. D. Lazarov, J. E. Pasciak, and P. S. Vassilevski, *Multiplier spaces for the mortar finite element method in three dimensions*, SIAM J. Numer. Anal., 39 (2001), pp. 519–538.

[14] H. H. Kim, *A FETI-DP preconditioner for mortar methods in three dimensions*, Electron. Trans. Numer. Anal., 26 (2007), pp. 103–120.

[15] H. H. Kim, M. Dryja, and O. B. Widlund, *A BDDC algorithm for problems with mortar discretization using a transformation of basis*, SIAM J. Numer. Anal., submitted.

[16] H. H. Kim and C.-O. Lee, *A preconditioner for the FETI-DP formulation with mortar methods in two dimensions*, SIAM J. Numer. Anal., 42 (2005), pp. 2159–2175.

[17] H. H. Kim and C.-O. Lee, *A Neumann-Dirichlet preconditioner for a FETI-DP formulation of the two-dimensional Stokes problem with mortar methods*, SIAM J. Sci. Comput., 28 (2006), pp. 1133–1152.

[18] A. Klawonn and O. Rheinbach, *A parallel implementation of dual-primal FETI methods for*

three-dimensional linear elasticity using a transformation of basis, SIAM J. Sci. Comput., 28 (2006), pp. 1886–1906.

[19] A. Klawonn and O. Rheinbach, *Robust FETI-DP methods for heterogeneous three dimensional elasticity problems*, Comput. Methods Appl. Mech. Engrg., 196 (2007), pp. 1400–1414.

[20] A. Klawonn and O. Rheinbach, *Some computational results for robust FETI-DP methods applied to heterogeneous elasticity problems in 3D*, in Domain Decomposition Methods in Science and Engineering XVI, Lect. Notes Comput. Sci. Eng. 55, Springer, Berlin, 2007, pp. 389–396.

[21] A. Klawonn and O. B. Widlund, *Dual and dual-primal FETI methods for elliptic problems with discontinuous coefficients in three dimensions*, in Domain Decomposition Methods in Sciences and Engineering (Chiba, 1999), DDM.org, Augsburg, Germany, 2001, pp. 29–39.

[22] A. Klawonn and O. B. Widlund, *Dual-primal FETI methods for linear elasticity*, Comm. Pure Appl. Math., 59 (2006), pp. 1523–1572.

[23] A. Klawonn, O. B. Widlund, and M. Dryja, *Dual-primal FETI methods for three-dimensional elliptic problems with heterogeneous coefficients*, SIAM J. Numer. Anal., 40 (2002), pp. 159–179.

[24] A. Klawonn, O. B. Widlund, and M. Dryja, *Dual-primal FETI methods with face constraints*, in Recent Developments in Domain Decomposition Methods (Zürich, 2001), Lect. Notes Comput. Sci. Eng. 23, Springer, Berlin, 2002, pp. 27–40.

[25] J. Mandel and R. Tezaur, *On the convergence of a dual-primal substructuring method*, Numer. Math., 88 (2001), pp. 543–558.

[26] D. Stefanica, *A numerical study of FETI algorithms for mortar finite element methods*, SIAM J. Sci. Comput., 23 (2001), pp. 1135–1160.

[27] A. Toselli and O. Widlund, *Domain Decomposition Methods—Algorithms and Theory*, Springer Series in Computational Mathematics 34, Springer-Verlag, Berlin, 2005.

[28] B. I. Wohlmuth, *A mortar finite element method using dual spaces for the Lagrange multiplier*, SIAM J. Numer. Anal., 38 (2000), pp. 989–1012.

[29] B. I. Wohlmuth, *Discretization methods and iterative solvers based on domain decomposition*, Lect. Notes Comput. Sci. Eng. 17, Springer, Berlin, 2001.

SIAM J. NUMER. ANAL.
Vol. 46, No. 5, pp. 2371–2392
© 2008 Society for Industrial and Applied Mathematics

CONVERGENCE RATE OF MONOTONE NUMERICAL SCHEMES FOR HAMILTON–JACOBI EQUATIONS WITH WEAK BOUNDARY CONDITIONS*

KNUT WAAGAN†

Abstract. We study a class of monotone numerical schemes for time-dependent Hamilton–Jacobi equations with weak Dirichlet boundary conditions. We get a convergence rate of $\frac{1}{2}$ under some usual assumptions on the data, plus an extra assumption on the Hamiltonian $H(Du, x)$ at the boundary $\partial\Omega$. More specifically the mapping $p \to H(p, x)$ must satisfy a monotonicity condition for all p in a certain subset of $\mathbf{R}^n$ given by Ω. This condition allows the use of the interior subsolution conditions at the boundary in the comparison arguments. We also prove a comparison result and Lipschitz regularity of the exact solution. As an example we construct a Godunov-type scheme that can handle the weakened boundary conditions.

Key words. Hamilton–Jacobi equations, viscosity solutions, weak boundary conditions, numerical approximation

AMS subject classifications. 35F30, 49L25, 65M12

DOI. 10.1137/050632932

1. Introduction.

Let the scalar function $u(x, t)$ solve the Hamilton–Jacobi equation

$$u_t + H(Du, x) = 0 \tag{1.1}$$

for x in an open and bounded set $\Omega \in \mathbf{R}^n$ and t in a bounded interval $(0, T)$. We will show well-posedness and Lipschitz regularity and will study numerical approximations to this equation under the Cauchy–Dirichlet boundary condition

$$\begin{aligned} u(x, 0) &= u_0(x), \quad x \in \bar{\Omega}, \\ u(x, t) &= g(x), \quad x \in \partial\Omega,\ 0 \le t < T. \end{aligned} \tag{1.2}$$

Note that $g = u_0|_{\partial\Omega}$. The mapping $p \to H(p, x)$ is assumed to tend uniformly to $+\infty$ as $|p| \to \infty$ and to do so in a later specified monotone way. In addition, some Lipschitz estimates on the data and some geometric conditions on the domain will be prescribed.

Classical solutions to (1.1) may develop discontinuities in the first derivative in finite time, hence viscosity solutions, as defined in [10], are more natural to study, since they are only continuous. Overviews of this subject are found in [9], [8], and [2]. The term "viscosity solution" comes from the vanishing viscosity method, in which a solution to the first order equation is sought by solving $u_t + H(Du, x) = \epsilon\,\Delta u$ and letting $\epsilon \to 0$, typically resulting in a viscosity solution. The viscosity solution concept also ensures causality in the sense that characteristics are not allowed to emanate from the singularities in Du. These facts reveal a connection to the entropy solution of scalar conservation laws which is more than superficial, but the two theories are

*Received by the editors May 30, 2005; accepted for publication (in revised form) August 25, 2007; published electronically June 6, 2008. This work was completed while the author was at the Department of Applied Mathematics, Würzburg University, Germany.
http://www.siam.org/journals/sinum/46-5/63293.html
†Center of Mathematics for Applications, University of Oslo, P.O. Box 1053 Blindern, NO-0316 Oslo, Norway (knut.waagan@cma.uio.no).

independent. However, numerical schemes for Hamilton–Jacobi equations are often derived from numerical schemes for conservation laws.

Since characteristics may be directed out of the boundary, the Cauchy–Dirichlet problem may be overdetermined both classically and in the original viscosity solutions setting, hence a weakened boundary condition is called for. The papers [5] and [14] both describe what has turned out to be the natural way to perform such a generalization, since the consistency with both control theory and vanishing viscosity approximation is maintained. This theory also includes other types of boundary data, as seen in, for example, [14] and [9], with Neumann conditions as another special case.

Before continuing, we need to specify what we mean by a viscosity solution. First, we state that by $C^1(\bar{\Omega} \times [0,T])$ we mean the class of all functions ψ such that ψ is a restriction to $\bar{\Omega} \times [0,T]$ of a function in $C^1(R^n \times R)$.

We define a viscosity solution to (1.1)–(1.2) as follows.

DEFINITION 1. *Let $\psi \in C^1(\bar{\Omega} \times [0,T])$. A function $u(x,t)$ is a subsolution of* (1.1)–(1.2) *if whenever $u^* - \psi$ has a local maximum at $(x,t) \in \bar{\Omega} \times [0,T]$, then*

$$\begin{aligned} \psi_t(x,t) + H(D\psi(x,t),x) &\le 0, \quad x \in \Omega,\ 0 < t \le T, \\ \min\left(u^*(x,t) - g(x), \psi_t(x,t) + H(D\psi(x,t),x)\right) &\le 0, \quad x \in \partial\Omega,\ t > 0, \\ u^*(x,0) &\le u_0(x), \quad x \in \bar{\Omega}. \end{aligned} \tag{1.3}$$

A function v is a supersolution of (1.1)–(1.2) *if whenever $v_* - \psi$ has a local minimum at $(x,t) \in \bar{\Omega} \times [0,T]$, then*

$$\begin{aligned} \psi_t(x,t) + H(D\psi(x,t),x) &\ge 0, \quad x \in \Omega,\ 0 < t \le T, \\ \max\left(v_*(x,t) - g(x), \psi_t(x,t) + H(D\psi(x,t),x)\right) &\ge 0, \quad x \in \partial\Omega,\ t > 0, \\ v_*(x,0) &\ge u_0(x), \quad x \in \bar{\Omega}. \end{aligned} \tag{1.4}$$

A function u is a viscosity solution of (1.1)–(1.2) *if it is both a sub- and a supersolution.*

The notations $\cdot^*$ and $\cdot_*$ refer to the usual upper and lower semicontinuous envelope, that is,

$$u^*(x,t) = \limsup_{y \in \bar{\Omega} \to x,\ s \in [0,T] \to t} u(y,s) \quad \text{and} \quad u_*(x,t) = \liminf_{y \in \bar{\Omega} \to x,\ s \in [0,T] \to t} u(y,s). \tag{1.5}$$

We have chosen to simply include the endpoint $t = T$ in our definition, but leaving the time interval open would have been equivalent. This is due to the monotone dependence on u_t in (1.1), which also makes weakening the initial condition redundant (see [5]).

The uniqueness of viscosity solutions follows if one can prove that $u \le v$ for all subsolutions u and supersolutions v. This is referred to as a comparison result.

In [5] there is a comparison result for (1.1)–(1.2) under some common continuity assumptions on the data and a nondegeneracy condition at the boundary. This nondegeneracy condition has uniform coercivity of H as a special case. We prefer to give a new comparison proof here since the proof in [5] requires uniform continuity of $p \to H(p,x)$, which can be avoided in our setting.

In [7] a uniform coercivity condition on $p \to H(p,x)$ is shown to imply $u^*|_{\partial\Omega} \le g$ and a comparison result for subsolutions that are lower Lipschitz bounded in time. The authors then construct a solution where this holds for initial data such that

$H(Du_0, \cdot) \leq 0$ using vanishing viscosity arguments. They also prove Lipschitz continuity of subsolutions that are nondecreasing in time under assumptions which are no stronger than will be made here.

The solution given by Definition 1 is the optimal value function for a finite horizon control problem where the cost stops running once a trajectory reaches $\partial\Omega$. This problem is covered in [13]. When the boundary condition is only weakly satisfied at $x \in \partial\Omega$, it is because the optimal trajectory starting at x does not exit Ω immediately.

The uniform convergence of monotone approximation schemes for the Cauchy problem was studied in [11] and [16]. The solution remains Lipschitz continuous if it is initially, and the convergence rate is $\frac{1}{2}$. In [6], locally uniform convergence for quite general boundary value problems is proved for the more general case of second order degenerate elliptic equations. The main condition in the proof is that the approximated problem satisfies a comparison principle. No explicit error estimate is given, and since schemes that form boundary layers are allowed, a uniform convergence rate would not be possible. Putting these results into practice for first order equations is the subject of [1]. Triangular grids are considered, and a Godunov scheme is extended to the boundary by reinterpreting the Hamiltonian as the Legendre transform of a control problem. Furthermore a Lax–Friedrichs scheme is also extended to the boundary, and certain nonconvex cases are also treated. Uniform coercivity of the Hamiltonian is shown to imply Lipschitz continuity of approximations to the stationary equation. No control theory will be used in this article, but our extension of the Godunov scheme to the boundary is very similar to that of [1].

Our main result is an estimate of $\frac{1}{2}$ for the convergence rate of monotone numerical schemes for problem (1.1)–(1.2). To achieve this result, we add conditions on H and Ω. Roughly speaking, we will require that as long as $p \cdot n < 0$ for the inward normal vector n, then $p \to H(p, x)$ is nondecreasing in the direction of $-n$. Special cases included are radially symmetric H on convex Ω, and H given by $H(u_{x_1}^2, \ldots, u_{x_n}^2)$, where H is nondecreasing in each argument and Ω is a rectangle oriented like $(0, 1)^n$. The proof is built on the treatment of the Cauchy problem in [11] and [16].

After stating some assumptions and preliminary results, we will specify the requirements at the boundary. Next a comparison result is given, from which existence and regularity follows. A generic numerical scheme is then defined, for which we state the main result on convergence rate. This is formulated in the tradition of [11] and [16], with the necessary extensions made to include the boundary conditions.

Finally we formulate a Godunov type scheme as an example. This turns out to have some theoretical difficulties other than just stating what happens at the grid points, since how to interpolate between the boundary and interior grid points is generally not obvious. Hence only stating what happens at the grid points is insufficient in this setting, although it may make sense for the Cauchy problem. The simple numerical boundary condition $u(x, t) = g(x)$ is not included in the analysis. Of course, we would not generally expect a uniform estimate for such a scheme since boundary layers could form, as they do for vanishing viscosity approximations. Our numerical results demonstrate convergence properties that are similar to what is known for pure Cauchy problems, as our convergence estimate predicts.

2. Assumptions and preliminary results. This section sums up some assumptions, definitions, and results we will need. The L^∞-norm of a function f is denoted $\|f\|$, and hence $\|Df\|$ denotes the Lipschitz seminorm of f, since a Lipschitz continuous function is differentiable almost everywhere by Rademacher's theorem.

We will make the following Lipschitz assumptions.

ASSUMPTION 1. *The mapping $p \to H(p,x)$ is continuous, and for any $R > 0$ there exists $C_R > 0$ such that*

$$|H(p,x) - H(p,y)| \le C_R|x-y| \quad \forall x, y \in \bar{\Omega},\ |p| \le R. \tag{2.1}$$

The initial value u_0 is Lipschitz continuous, so one may define

$$L_0 = \max(\|Du_0\|, p_0), \quad \text{where } p_0 = \sup_{x\in\Omega}\{|p| : H(p,x) = 0\} < \infty. \tag{2.2}$$

By uniform coercivity, we mean, as in [3], [4], and [7] the following.

ASSUMPTION 2. *The mapping $p \to H(p,x)$ is uniformly coercive, that is, for any $C \in \mathbf{R}$ there exists $r > 0$ such that $H(p,x) > C$ for all $|p| > r$ and $x \in \bar{\Omega}$.*

This is used to get Lipschitz continuity of u, and it also gives the following proposition. It requires yet another assumption, this one on Ω, which we state first. It is an exterior sphere condition given locally.

ASSUMPTION 3. *For each $x \in \partial\Omega$, there exist $r_x > 0$ and $z_x \in \mathbf{R}^n$ defining an open ball $B(z_x, r_x)$ such that $\bar{B}(z_x, r_x) \cap \bar{\Omega} = \{x\}$.*

PROPOSITION 2.1. *If Assumptions 2 and 3 hold, then*

$$u^*(x,t) \le g(x) \quad \forall x \in \partial\Omega,\ t \in [0,T] \tag{2.3}$$

holds for any subsolution u of (1.1)–(1.2) that is bounded above.

Our proof essentially copies the one in [7], which treated C^1 boundaries. A more intuitive motivation for this result is the fact that superdifferentials become unbounded on $\partial\Omega$.

Proof. Assume in order to obtain a contradiction that there exists $(x_0, t_0) \in \partial\Omega \times (0,T]$ such that $u^*(x_0,t_0) > g(x_0)$. Consider

$$\Psi(x,t) = u^*(x,t) - \frac{(t-t_0)^2}{2\beta^2} - \frac{1}{\alpha}\left(|x - z_{x_0}| - |x_0 - z_{x_0}|\right), \quad \alpha,\ \beta > 0. \tag{2.4}$$

The point z_{x_0} is given by Assumption 3, hence the last term of Ψ is nonnegative. Denote the maximum point of Ψ in $\bar{\Omega} \times [0,T]$ by $(\hat{x}, \hat{t})$. Now,

$$u^*(x_0,t_0) = \Psi(x_0,t_0) \le \Psi(\hat{x},\hat{t}) \le u^*(\hat{x},\hat{t}) - C\left(|\hat{x} - z_{x_0}| - |x_0 - z_{x_0}|\right) - \frac{(\hat{t}-t_0)^2}{2\beta^2}. \tag{2.5}$$

Since u^* is bounded by, say, $R > 0$, this means that $(|\hat{x} - z_{x_0}| - |x_0 - z_{x_0}|) \le 2R\alpha$ and $|\hat{t} - t_0| \le 2\sqrt{R}\beta$. This implies $u^*(\hat{x},\hat{t}) \to u^*(x_0,t_0)$ as α and β go to zero; hence, taking α and β small enough, we have $u^*(\hat{x},\hat{t}) > g(\hat{x})$. That means we may use the differential equation, so

$$\frac{\hat{t} - t_0}{\beta^2} + H\left(\frac{\hat{x} - z_{x_0}}{\alpha|\hat{x} - z_{x_0}|}, \hat{x}\right) \le 0. \tag{2.6}$$

By uniform coercivity of H, this implies

$$\frac{1}{\alpha} \le \Gamma_\beta \tag{2.7}$$

for some $\Gamma_\beta < \infty$. So choosing $\alpha < \frac{1}{\Gamma_\beta}$ produces the contradiction. ∎

2.1. Subsolutions on the boundary. In this section we state the monotonicity requirement of H at $\partial\Omega$. The aim is to give conditions such that the interior subsolution condition holds on all $\bar{\Omega}$, at least for test functions of the form typically used in a doubling of variables argument. The key observation is that the gradient of such a test function points out of the domain for certain choices of H and Ω. Hence, monotonicity of $p \to H(p,x)$ only needs to hold for p in a subset of $\mathbf{R}^n$. This is similar to the results of [12] and [17] with the exception that these require global monotonicity of H. In the simplest case that $\partial\Omega$ is C^1 and convex, with inward normal vector $n(x)$, we need

$$H(k(x-z)-\mu n(x),x) \geq H(k(x-z),x) \quad \forall\, k,\mu>0,\ x\in\partial\Omega,\ z\in\bar{\Omega}. \tag{2.8}$$

However, to allow corners, we require the following.

ASSUMPTION 4. *There exists a nonnegative function $d \in C^1(\Omega)\cap C^0(\bar{\Omega})$ such that $d(x)=0$ if and only if $x\in\partial\Omega$. Furthermore, for every $x_0\in\partial\Omega$, $k>0$, positive function $\mu:\Omega\to\mathbf{R}$, and $z\in\bar{\Omega}$,*

$$\liminf_{x\to x_0} H(k(x-z)-\mu(x)Dd(x),x) \geq H(k(x_0-z),x_0). \tag{2.9}$$

We cannot simply state the monotonicity property at the point x_0, since $Dd(x_0)$ will be 0 at any corner, and because we need to allow $\mu(x)$ to go to infinity at $\partial\Omega$. Note, however, that where $\partial\Omega$ is C^1, $Dd(x)$ must be parallel to the inward normal vector $n(x)$, and hence (2.8) suffices.

We could have reduced the domain of z by restricting $|z-x_0|$, but that would be a bit technically complicated to use in the proof of convergence rate. The assumption becomes useful through the following lemma.

LEMMA 2.2. *Let Assumption 4 hold. Let u be a bounded subsolution of (1.1)–(1.2) such that*

$$u^*(x,t) = \limsup_{y\in\Omega\to x,\, s\in(0,T)\to t} u(y,s). \tag{2.10}$$

Suppose the mapping

$$(x,t) \to u^*(x,t) - \frac{k}{2}(x-z)^2 - \eta(t)$$

attains its maximum over $\bar{\Omega}\times[0,T]$ at (x_0,t_0) with $x_0\in\partial\Omega$, $t_0>0$, $k>0$, $z\in\bar{\Omega}$, and $\eta\in C^1((0,T))$. Then

$$\eta_t(t_0) + H(k(x_0-z),x_0) \leq 0. \tag{2.11}$$

Proof. Without losing generality, we may assume the maximum is strict. Consider

$$\Psi(x,t) = u^*(x,t) - \frac{k}{2}(x-z)^2 - \eta(t) - \frac{\epsilon}{d(x)}. \tag{2.12}$$

Because of (2.10), Ψ will have a local maximum (x_ϵ,t_ϵ) with $x_\epsilon\in\Omega$ such that $(x_\epsilon,t_\epsilon)\to(x_0,t_0)$ as ϵ goes to 0. Hence

$$0 \geq \eta_t(t_\epsilon) + H\left(k(x_\epsilon-z) - \epsilon\frac{Dd(x_\epsilon)}{d(x_\epsilon)^2}, x_\epsilon\right). \tag{2.13}$$

Letting $\epsilon\to 0$, the lemma follows from Assumption 4. ∎

2.1.1. Examples. An example where Assumption 4 holds is if $p \to H(p,x)$ is radially symmetric with a unique minimum at $p=0$ (i.e., $H(p,x)=h(|p|,x)$, where $h:[0,\infty)\times\mathbf{R}^n\to\mathbf{R}$ is nondecreasing in the first variable), Ω is convex, and $\partial\Omega$ is piecewise C^1. The function d can be chosen as the product of the distances to each C^1 subset of $\partial\Omega$.

As a second example consider $\Omega=(0,1)^n$. Then the assumption holds for any Hamiltonian $H(p,x)=H_2\left(p_1^2,p_2^2,\dots,p_n^2,x\right)$ if $p_i^2\to H_2$ is nondecreasing for all $1\le i\le n$. We can choose $d(x_1,\dots,x_n)=\Pi_1^n x_i(1-x_i)$.

An example where the Lemma fails is if $H=H(p):\mathbf{R}^2\to\mathbf{R}$ has level sets that are ellipses with no axis parallel to the normal vector Dd. To demonstrate this we consider for simplicity $\Omega=\{x=(\xi,\eta):\xi>0\}\in\mathbf{R}^2$ (hence $d(x)=\xi$ is a reasonable choice of d) and take

$$H(Du)=\hat{H}(u_\xi,u_\eta)=(u_\xi-u_\eta)^2+\frac{1}{4}(u_\xi+u_\eta)^2-1. \tag{2.14}$$

For instance, we may choose k, x_0, and z in Lemma 2.2 such that $k(x_0-z)=(0,-1)$ and get

$$H(k(x_0-z))=\hat{H}\,(0,-1)>\hat{H}(-\mu,-1)=H\left(k(x_0-z)-\mu Dd(x_0)\right) \tag{2.15}$$

for all $\mu\in(0,\frac{6}{5})$. Hence Assumption 4 fails at the line $\xi=0$ for $H=\hat{H}$.

2.1.2. Translation of the zero gradient. In Assumption 4, the value $p=0$ is always included in the region of $\mathbf{R}^n$ where $p\to H(p,x)$ has to be monotone along Dd, but such a restriction may be remedied by a translation. Consider the situation that Assumption 4 holds for the Hamiltonian $p\to H(p-m(x),x)$ and that there exists a differentiable f such that $Df=m$. Then $w(x,t)=u(x,t)+f(x)$ will solve

$$\begin{aligned} w_t+H(Dw-m(x),x)=0,\quad w(x,0)=u_0(x)+f(x)\quad \forall x\in\Omega,\\ w(x,t)=g(x)+f(x)\quad \forall x\in\partial\Omega \end{aligned} \tag{2.16}$$

in the viscosity sense.

As a special case consider $\Omega=(0,1)\in\mathbf{R}$. Then, according to the above, $H(p,0)$ must be decreasing for $p<a$ and $H(p,1)$ must be increasing for $p>b$ for some constants a and b.

3. Uniqueness, existence, and regularity. In this section we prove the following theorem.

THEOREM 3.1. *Let Assumptions* 1–4 *hold. Then there exists one and only one viscosity solution* u *of* (1.1)–(1.2) *in* $C^0(\bar{\Omega}\times[0,T])$. *Furthermore,*

$$\|u_t\|\le L_t \ \text{ and } \ \|Du\|\le L, \tag{3.1}$$

where

$$L_t=\sup_{|p|\le L_0,x\in\Omega}|H(p,x)|$$

and L *is any number such that* $H(p,x)>L_t$ *for all* $|p|>L$ *and* $x\in\Omega$.

Note that $L\ge L_0$ and that L may be finite because of the uniform coercivity condition.

We first state that it is enough to prove the theorem for problem (1.1)–(1.2) with a truncated Hamiltonian, more specifically, the problem

$$\begin{aligned} u_t + H_R(Du, x) &= 0 \qquad \forall x \in \Omega,\, t \in (0,T), \\ u(x,0) &= u_0(x), \quad x \in \bar{\Omega}, \\ u(x,t) &= g(x), \quad x \in \partial\Omega,\, 0 \le t < T, \end{aligned} \tag{3.2}$$

where $H_R(p,x) = \min(H(p,x), R)$ for some finite $R > L_t$. The advantage is that H_R is globally Lipschitz in the x-variable. Clearly the solution of (3.2) also solves the original problem (1.1)–(1.2) because of (3.1). The uniqueness statement of Theorem 3.1 follows since we also show that comparison holds for (1.1)–(1.2) provided either the subsolution or the supersolution is Lipschitz continuous. The Hamiltonian H_R satisfies Assumption 4 but not Assumption 2, and hence Proposition 2 does not apply. However, it is enough to prove Theorem 3.1 for the problem (3.2) under the explicit assumption that the solution u of (3.2) satisfies $u|_{\partial\Omega} \le g$.

The crucial point is the following comparison result.

THEOREM 3.2. *Assume Assumptions* 1 *and* 4 *hold and that one of the following assumptions holds:*

1. *Let u be a bounded subsolution and v a bounded supersolution of* (3.2)*. Let u satisfy* (2.10)*, and let $u^*|_{\partial\Omega} \le g$.*
2. *Let u be a bounded subsolution and v a bounded supersolution of* (1.1)–(1.2)*. Let u satisfy* (2.10) *and assume either u or v is Lipschitz continuous.*

Then

$$u^* \le v_* \tag{3.3}$$

in $\bar{\Omega} \times [0,T]$.

Note that provided $u^*|_{\partial\Omega} \le g$, the function $\limsup_{y \in \Omega \to x, s \in (0,T) \to t} u(y,s)$ is a subsolution. Hence the comparison result implies that given a viscosity solution u of (3.2) such that $u^*|_{\partial\Omega} \le g$, we have that u_* is in $C^0(\bar{\Omega} \times [0,T])$ and is a viscosity solution of (3.2). However, u can be discontinuous at the boundary in general.

The proof is standard due to Lemma 2.2; the only new problem is handling the boundary condition for the supersolution, which can be done quite easily.

Proof. Define

$$M = \sup(u^*(x,t) - v_*(x,t) - 2\delta t) \tag{3.4}$$

We assume $M > 0$ and proceed to show that this leads to a contradiction for any constant $\delta > 0$. Let

$$\Psi(x,t,y,s) = u^*(x,t) - v_*(y,s) - \delta(t+s) - \frac{(x-y)^2}{2\epsilon^2} - \frac{(t-s)^2}{2\epsilon^2}. \tag{3.5}$$

Let

$$(x_\epsilon, t_\epsilon, y_\epsilon, s_\epsilon) \in \left(\bar{\Omega} \times [0,T]\right)^2 \tag{3.6}$$

be a maximum point of Ψ in $\left(\bar{\Omega} \times [0,T]\right)^2$.

Then,

$$\begin{aligned} &u^*(x_\epsilon, t_\epsilon) - v_*(y_\epsilon, s_\epsilon) - \delta(t_\epsilon + s_\epsilon) - \frac{(x_\epsilon - y_\epsilon)^2}{2\epsilon^2} - \frac{(t_\epsilon - s_\epsilon)^2}{2\epsilon^2} \\ &\qquad = \Psi(x_\epsilon, t_\epsilon, y_\epsilon, s_\epsilon) \ge \Psi(x,t,x,t) \ge u^*(x,t) - v_*(x,t) - 2\delta t. \end{aligned} \tag{3.7}$$

Taking the supremum of the last expression, we see that

$$u^*(x_\epsilon, t_\epsilon) - v_*(y_\epsilon, s_\epsilon) - \delta(t_\epsilon + s_\epsilon) \geq M, \tag{3.8}$$

and by standard arguments (see, for example, [2]), and taking a subsequence if necessary, we also get

$$x_\epsilon \to x^*,\ y_\epsilon \to x^*, \quad t_\epsilon \to t^*,\ s_\epsilon \to t^*, \tag{3.9}$$

where the supremum in (3.4) is obtained at (x^*, t^*). If $x^* \in \Omega$, or if $t^* = 0$, the proof is similar to the standard case.

If $x^* \in \partial\Omega$, we need to consider the possibilities that x_ϵ or y_ϵ are in $\partial\Omega$ as $\epsilon \to 0$. The case $x_\epsilon \in \partial\Omega$ can be treated in the same way as if $x_\epsilon \in \Omega$ due to Lemma 2.2. If $y_\epsilon \in \partial\Omega$, there are essentially two cases. First, if there are arbitrarily small ϵ such that $v_*(y_\epsilon, s_\epsilon) < g(y_\epsilon)$, we may argue as above. Otherwise, we may consider ϵ such that $v_*(y_\epsilon, s_\epsilon) \geq g(y_\epsilon)$, in which case (3.8) implies

$$\begin{aligned} M &\leq u^*(x_\epsilon, t_\epsilon) - v_*(y_\epsilon, s_\epsilon) \\ &\leq u^*(x_\epsilon, t_\epsilon) - g(y_\epsilon), \quad y_\epsilon \in \partial\Omega. \end{aligned} \tag{3.10}$$

Then, by (3.9) and $u^*|_{\partial\Omega} \leq g$,

$$M \leq \limsup_{\epsilon \to 0} \{u^*(x_\epsilon, t_\epsilon) - u_0(y_\epsilon)\} \leq u^*(x^*, t^*) - u_0(x^*) \leq 0. \qquad \square \tag{3.11}$$

We can prove the existence of a viscosity solution such that $u|_{\partial\Omega} \leq g$ by a standard application of Perron's method as adapted to viscosity solutions; see, for example, [2]. Note that as a supersolution we can take $u_0(x) + 2L\text{dist}(x)$, where $\text{dist}(x)$ is the distance function to $\partial\Omega$, and as a subsolution we can take $u_0(x) - L_t t$.

From the comparison result we easily get the following.

PROPOSITION 3.3. *Let Assumptions* 1 *and* 4 *hold, and let* $u \in C^0(\bar{\Omega} \times [0, T])$ *be a viscosity solution of* (3.2) *such that* $u|_{\partial\Omega} \leq g$. *Then*

$$\|u(\cdot, t)\| \leq \|u_0\| + C_0 t, \quad C_0 = \max_{x \in \bar{\Omega}} |H(0, x)|, \tag{3.12}$$

and

$$|u(x, t) - u_0(x)| \leq L_t t. \tag{3.13}$$

Proof. This follows from $\|u_0\| + C_0 t$ and $u_0(x) + L_t t$ being supersolutions and $-\|u_0\| - C_0 t$ and $u_0(x) - L_t t$ being subsolutions. $\square$

We now turn to Lipschitz regularity.

THEOREM 3.4. *Let Assumptions* 1 *and* 4 *hold, and let* $u \in C^0(\bar{\Omega} \times [0, T])$ *be a viscosity solution of* (3.2) *such that* $u|_{\partial\Omega} \leq g$. *Then*

$$u(x, t + \tau) - u(x, t) \leq L_t \tau \tag{3.14}$$

for $\tau > 0$.

Proof. This follows from

$$\tilde{u}(x, t) = u(x, t + \tau) - L_t \tau \tag{3.15}$$

being a subsolution. This is easily seen for $t > 0$ and $x \in \Omega$, and by Proposition 3.3,

$$\tilde{u}(x, 0) = u(x, \tau) - L_t \tau \leq u_0(x). \tag{3.16}$$

Finally, at the boundary points $x \in \partial\Omega$, we have $\tilde{u}(x, t) \leq u(x, t) \leq g(x)$. $\square$

The lower bound on u_t follows similarly by

$$(x,t) \mapsto u(x,t+\tau) + L_t\tau \tag{3.17}$$

being a subsolution. A coercivity argument could also have been used for the upper bound.

The Lipschitz continuity in x follows from Assumption 2, since it implies that

$$\underline{u}(x,t) = \sup_{y\in\overline{\Omega}}\{u(y,t) - k|y-x|\} \tag{3.18}$$

is a subsolution whenever $k > L$ with L defined as in Theorem 3.1. Consider first $t = 0$. Clearly there exists y_0 such that

$$u_0(y) - k|x-y| \le \underline{u}(x,0) = u_0(y_0) - k|x-y_0|.$$

More specifically, $u_0(x) \le u_0(y_0) - k|x-y_0|$; hence, unless $y_0 = x$, we get the contradiction $k \le L_0 \le L$, so $\underline{u}(x,0) = u_0(x)$. Next, consider $\underline{u}(x,t) - \psi(x,t)$ having a maximum point (x_0,t_0) with $t_0 > 0$. Then there exists y_0 such that

$$u(y,t) - k|y-x| - \psi(x,t) \le u(y_0,t_0) - k|y_0-x_0| - \psi(x_0,t_0). \tag{3.19}$$

If $x_0 \neq y_0$, we get

$$\psi_t(x_0,t_0) + H_R\left(k\frac{x_0-y_0}{|x_0-y_0|}, y_0\right) \le 0, \tag{3.20}$$

but by the Lipschitz continuity in time $|\psi_t(x_0,t_0)| \le L_t$, which contradicts $k > L$. Hence $x_0 = y_0$, so, by (3.19),

$$\psi_t(x_0,t_0) + H_R\left(D\psi(x_0,t_0), x_0\right) \le 0. \tag{3.21}$$

Because of Lemma 2.2, this also holds if $y_0 \in \partial\Omega$.

4. The convergence result. The convergence result may now be stated. We consider a scheme $S^\lambda(\rho) : B_g \to B_g$, where B_g is a class of functions in $L^\infty(\bar{\Omega})$ such that all Lipschitz continuous functions U satisfying $U|_{\partial\Omega} \le g$ are included in B_g. Furthermore, $S^\lambda(\rho)$ is defined for all $\lambda > 0$ sufficiently small, and $0 \le \rho \le \lambda$.

We define our approximate solution $U(x,t) : \bar{\Omega}\times[0,T] \to \mathbf{R}$ inductively by setting

$$U(x,0) = u(x,0), \quad x \in \bar{\Omega}, \tag{4.1}$$

and for $t \in (t_n, t_{n+1})$, where $t_n = n\Delta t$ with $0 \le \Delta t \le \lambda$:

$$U(x,t) = S^\lambda(t-t_n)U(\cdot,t_n)(x), \quad x \in \bar{\Omega}. \tag{4.2}$$

The time step Δt must be chosen such that for $\rho < \Delta t$, $S^\lambda(\rho)$ satisfies the following conditions:

1.

$$\left\|S^\lambda(\rho)U\right\| \le \|U\| + \tilde{C}_0\rho \tag{4.3}$$

for a constant $\tilde{C}_0$ independent of λ.

2.

$$\left\|S^\lambda(\rho)U - U\right\| \leq C_S(\|DU\|)\rho, \tag{4.4}$$

where $C_S(\cdot) \in C^0 : \mathbf{R}_0^+ \to \mathbf{R}_0^+$ and independent of λ.

3. Let $\psi(x) = \frac{k}{2}(x-z)^2 + K$ and $\theta(x) = -\frac{k}{2}(x-z)^2 + K$ for some $z \in \bar{\Omega}$ and constants $k, K \geq 0$. Then there is a $C > 0$ such that the following holds for all $x \in \Omega$: given $U \in B_g$, we have either that

$$\frac{(S^\lambda(\rho)\theta)_*(x) - \theta(x)}{\rho} + H(D\theta, x) \geq -C\left(1 + \|D\theta\| + \left\|D^2\theta\right\|\right)\lambda \tag{4.5}$$

whenever $\theta \leq U$ or that $(\xi, \tau) \to S^\lambda(\tau)U(\xi)$ is a supersolution of (1.1) in a neighborhood of (x, ρ).

4. There exists an operator $S_b^\lambda : L^\infty(\bar{\Omega}) \to L^\infty(\bar{\Omega})$ such that for any $U \in B_g$, $S^\lambda(\rho)U \leq S_b^\lambda(\rho)U$ and

$$S^\lambda(\rho)U(x) = \min\left(g(x), S_b^\lambda(\rho)U(x)\right) \quad \forall\, x \in \partial\Omega. \tag{4.6}$$

Furthermore, S_b^λ satisfies

$$\begin{aligned} &\frac{(S_b^\lambda(\rho)\theta)_*(x) - \theta(x)}{\rho} + H(D\theta(x), x) \geq -C\left(1 + \|D\theta\| + \left\|D^2\theta\right\|\right)\lambda, \\ &\frac{(S_b^\lambda(\rho)\psi)^*(x) - \psi(x)}{\rho} + H(D\psi(x), x) \leq C\left(1 + \|D\psi\| + \left\|D^2\psi\right\|\right)\lambda \end{aligned} \tag{4.7}$$

for all $x \in \bar{\Omega}$.

5. S^λ and S_b^λ are monotone in the sense that if $U_1(x) \leq U_2(x)$ for all $x \in \bar{\Omega}$, then

$$S^\lambda(\rho)U_1(x) \leq S^\lambda(\rho)U_2(x) \quad \text{and} \quad S_b^\lambda(\rho)U_1(x) \leq S_b^\lambda(\rho)U_2(x)$$

for all $x \in \bar{\Omega}$.

6. S^λ is also monotone in g in the following sense: for any function $\bar{g} \geq g$, there exists an operator $S_{\bar{g}}^\lambda : B_{\bar{g}} \to B_{\bar{g}}$ with $B_g \subseteq B_{\bar{g}}$ such that the above assumptions hold with g replaced by $\bar{g}$, and

$$S^\lambda(\rho)U \leq S_{\bar{g}}^\lambda(\rho)U \tag{4.8}$$

for any $U \in B_g$.

7. There exists $L_S \geq L_0$ such that one of the following two conditions holds: (i) For all $y \in \partial\Omega$ and $k \geq L_S$,

$$S_{\bar{g}}^\lambda(\rho)[\, g(y) + k|x-y|] \;\leq\; g(y) + k|x-y|, \tag{4.9}$$

where $\bar{g}$ here denotes $k|x-y| + g(y)$ restricted to $\partial\Omega$. (ii) We have the Lipschitz estimate

$$\left\|DS^\lambda(\rho)U\right\| \leq \|DU\| + L_S(1 + \|DU\|)\rho. \tag{4.10}$$

Our main result follows.

THEOREM 4.1. *Let Assumptions* 1–4 *hold. Then* $U(x,t)$ *converges uniformly to the unique viscosity solution* u *of* (1.1)–(1.2) *in* $C^0(\bar{\Omega}\times[0,T])$, *and*

$$|U(x,t)-u(x,t)| \le K(\lambda^{\frac{1}{2}}+\lambda) \quad \forall (x,t)\in \bar{\Omega}\times[0,T] \tag{4.11}$$

for a constant K *given by* $\|u_0\|, C, C_0, \tilde{C}_0, C_{2L}, C_S(L_0), L_t, L, L_S,$ *and* T.

Requiring both monotonicity and (4.7) means that schemes like S^λ exist only for a certain class of Hamiltonians and domains. However, this class should coincide reasonably well with the one specified by Assumption 4, since S_b^λ needs to be consistent only for gradients of the form $k(x-z)$ for which H is monotone. Unfortunately, Lemma 2.2 does not translate directly to a corresponding lemma for approximations.

Before proving the theorem, we show how monotonicity and consistency of S^λ imply that U is in a sense almost a sub- and a supersolution (at least with respect to the test functions we consider).

LEMMA 4.2. *Define*

$$\psi(x,t)=\frac{k}{2}(x-z)^2+\eta(t)$$

with $k\ge 0$, $z\in\bar{\Omega}$, *and* $\eta\in C^2([0,T])$. *If* $U^*(x,t)-\psi(x,t)$ *has a global maximum in* $(x_0,t_0)\in\bar{\Omega}\times(0,T]$ *for* $t_0>0$, *then*

$$\eta_t(x_0,t_0)+H(k(x_0-z),x_0)\le \big(\|\eta_{tt}\|+C(1+k\,\mathrm{diam}(\Omega)+k)\big)\lambda, \tag{4.12}$$

where $\mathrm{diam}(\Omega)=\sup_{x,y\in\Omega}|x-y|$.

Proof. By assumption,

$$U(x,t_n)\le U^*(x,t_n)\le \psi(x,t_n)+U^*(x_0,t_0)-\psi(x_0,t_0). \tag{4.13}$$

We may assume that $\psi(x_0,t_0)=U^*(x_0,t_0)$, since only the derivatives of ψ are relevant. Monotonicity of S_b^λ then implies

$$S^\lambda(t_0-t_n)U(\cdot,t_n)(x)\le S_b^\lambda(t_0-t_n)\psi(\cdot,t_n)(x)+U^*(x_0,t_0)-\psi(x_0,t_0) \tag{4.14}$$

with $t_n<t_0\le t_{n+1}$. If we take the upper semicontinuous envelope on both sides, and set $x=x_0$, this becomes

$$0\le [S_b^\lambda(t_0-t_n)\psi(\cdot,t_n)]^*(x_0)-\psi(x_0,t_0), \tag{4.15}$$

which can be rewritten as

$$0\le \frac{[S_b^\lambda(t_0-t_n)\psi(\cdot,t_n)]^*(x_0)-\psi(x_0,t_n)}{t_0-t_n}+\frac{\psi(x_0,t_n)-\psi(x_0,t_0)}{t_0-t_n}, \tag{4.16}$$

and hence from consistency (4.7):

$$\begin{aligned} 0\le&\frac{[S_b^\lambda(t_0-t_n)\psi(\cdot,t_n)]^*(x_0)-\psi(x_0,t_0)}{t_0-t_n}-\psi_t(x_0,t_0)-\|\psi_{tt}\|(t_0-t_n)\\ \le&-H(D\psi(x_0,t_0),x_0)-\psi_t(x_0,t_0)+\|\psi_{tt}\|\lambda+C(1+\|D\psi\|+\|D^2\psi\|)\lambda. \end{aligned} \tag{4.17}$$

□

Analogously, we have the next lemma.

LEMMA 4.3. *Let $\theta(x,t) = -\frac{k}{2}(x-z)^2 + \eta(t)$ with $k \geq 0$, $z \in \bar{\Omega}$, and $\eta \in C^2[0,T]$, and let $U_*(y,s) - \theta(y,s)$ have a global minimum in $(y_0,s_0) \in \bar{\Omega} \times (0,T]$. Assume also $U(y_0,s_0) < g(y_0)$ if $y_0 \in \partial\Omega$. Then*

$$\theta_s(y_0,s_0) + H(D\theta(y_0,s_0),y_0) \geq -\Big(\|\theta_{ss}\| + C\left(1 + \|D\theta\| + \left\|D^2\theta\right\|\right)\Big)\lambda. \tag{4.18}$$

Note that in the case $y_0 \in \partial\Omega$ and $U(y_0,s_0) = g(y_0)$, this fails to hold. However, then the Lipschitz type estimate (4.22) given below is sufficient for our purpose.

Proof. Following the arguments (4.14)–(4.17) (note that we have $U(x,t_n) \geq \theta(x,t_n)$), we get

$$\begin{aligned} 0 \geq & -H(D\theta(0,s_0),0) - \theta_s(0,s_0) \\ & - \|\theta_{ss}\|\,(s_0 - s_n) - C(1 + \|D\theta\| + \left\|D^2\theta\right\|)\lambda. \end{aligned} \tag{4.19}$$

For $y_0 \in \partial\Omega$, use (4.7) instead of (4.5). □

Before proving Theorem 4.1, we need some regularity estimates on U at the boundary and initially.

PROPOSITION 4.4. *Let U be defined by (4.1)–(4.2). Then, for all $(x,t) \in \bar{\Omega} \times [0,T]$ and $y \in \partial\Omega$,*

$$|U(x,t)| \leq \|u_0\| + C_0 t, \tag{4.20}$$

$$|U(x,t) - u_0(x)| \leq C_S(L_0)t, \tag{4.21}$$

$$U^*(x,t) \leq g(y) + \tilde{L}_S|x-y|, \tag{4.22}$$

where $\tilde{L}_S \in [0,\infty)$ is given by L_0, L_S, and T.

Proof. We first prove (4.21). The proof of (4.20) is similar and easier. Clearly this holds for $t \leq \Delta t$. Assume for induction that

$$U(x,t_n) \geq u_0(x) - C_S(L_0)t_n.$$

Then, for $t_n < t \leq t_{n+1}$,

$$\begin{aligned} U(x,t) &= S^\lambda(t-t_n)U(\cdot,t_n)(x) \geq S^\lambda(t-t_n)[u_0 - C_S(L_0)t_n](x) \\ &\geq u_0(x) - C_S(L_0)t, \end{aligned} \tag{4.23}$$

since $\|D(u_0 - C_S(L_0)t_n)\| \leq L_0$. To prove the opposite inequality, assume that $U(x,t_n) \leq u_0(x) + C_S(L_0)t_n$. Now, S^λ applied to the right-hand side is not defined since $(u_0 + C_S(L_0)t_n)|_{\partial\Omega} > g$. However, taking $\bar{g} = g + C_S(L_S)T$, we can use (4.8) to get

$$\begin{aligned} U(x,t) &= S^\lambda(t-t_n)U(\cdot,t_n)(x) \leq S^\lambda_{\bar{g}}(t-t_n)U(\cdot,t_n)(x) \\ &\leq S^\lambda_{\bar{g}}(t-t_n)[u_0 + C_S(L_0)t_n](x) \leq u_0(x) + C_S(L_0)t, \quad t_n < t \leq t_{n+1}, \end{aligned} \tag{4.24}$$

which is the desired estimate.

To prove (4.22) we start with the inequality

$$u_0(x,t) \leq k|x-y| + g(y), \tag{4.25}$$

where $k \geq L_S \geq L_0$. If (4.9) holds, we get that

$$S^\lambda(\rho)u_0(x,t) \leq S^\lambda_{\bar{g}}(\rho)u_0(x,t) \leq S^\lambda_{\bar{g}}(\rho)(k|x-y| + g(y)) \leq k|x-y| + g(y), \tag{4.26}$$

where $\bar{g}$ denotes $k|x-y|+g(y)$ restricted to $\partial\Omega$. It hence follows by induction that

$$U^*(x,t) \le g(y) + L_S|x-y|. \tag{4.27}$$

If on the other hand the assumption (4.10) holds, we get that

$$\|DU\| \le (L_0+1)\exp L_S t - 1. \tag{4.28}$$

To see this, observe that (4.10) applied inductively constitutes an Euler approximation to the ordinary differential equation $\|U_x\|_t = L_S(1+\|U_x\|)$. From this the result holds with

$$\tilde{L}_S = (L_0+1)\exp L_S T - 1. \qquad \square \tag{4.29}$$

4.1. Proof of Theorem 4.1. For completeness, we include the whole doubling of variables argument, although most of it is standard procedure due to Lemmas 4.2 and 4.18. The exception comes when we have a supersolution assuming the boundary condition strongly, in which case we use the regularity of u or (4.22).

We want to estimate $U(x,t)-u(x,t)$ from above first, so we define

$$M = \sup(U^*(x,t)-u(x,t)) \tag{4.30}$$

and assume $M>0$. From here we follow the doubling of variables strategy as in [11] or [16]. Let

$$\Psi(x,t,y,s) = U^*(x,t) - u(y,s) - \frac{(x-y)^2}{2\epsilon^2} - \frac{(t-s)^2}{2\epsilon^2} - \frac{t+s}{4T}M. \tag{4.31}$$

By upper semicontinuity Ψ attains a maximum in $\left(\bar{\Omega}\times[0,T]\right)^2$. Let

$$(x_0,t_0,y_0,s_0) \in \left(\bar{\Omega}\times[0,T]\right)^2 \tag{4.32}$$

be this maximum point.

Then we have

$$\begin{aligned} &U^*(x_0,t_0) - u(y_0,s_0) - \frac{(x_0-y_0)^2}{2\epsilon^2} \\ &\qquad \ge \Psi(x_0,t_0,y_0,s_0) \ge \Psi(x,t,x,t) \ge U^*(x,t) - u(x,t) - \frac{M}{2}. \end{aligned} \tag{4.33}$$

Taking the supremum of the last expression, we see that

$$U^*(x_0,t_0) - u(y_0,s_0) \ge M/2. \tag{4.34}$$

Using the Lipschitz continuity of u, we can show that

$$|x_0-y_0| \le 2L\epsilon^2, \quad |t_0-s_0| \le \left(L_t + \frac{R}{2T}\right)\epsilon^2 \tag{4.35}$$

with $R = \|u_0\| + \max(C_0,\tilde{C}_0)T$. Clearly, $|\frac{t_0-s_0}{2\epsilon^2} - \frac{M}{4T}| \le L_t$ by (3.1), and hence the last estimate follows from $M \le 2R$. To see the first, note that

$$u(y_0,s_0) + \frac{(x_0-y_0)^2}{2\epsilon^2} \le u(x_0,s_0), \tag{4.36}$$

which gives

$$\frac{(x_0-y_0)^2}{2\epsilon^2} \le u(x_0,s_0)-u(y_0,s_0) \le L|x_0-y_0|. \tag{4.37}$$

In the following let

$$\begin{aligned}\theta(y,s) &= -\frac{(x_0-y)^2}{2\epsilon^2}-\frac{(t_0-s)^2}{2\epsilon^2}-\frac{t_0+s}{4T}M,\\ \psi(x,t) &= \frac{(x-y_0)^2}{2\epsilon^2}+\frac{(t-s_0)^2}{2\epsilon^2}+\frac{t+s_0}{4T}M\end{aligned} \tag{4.38}$$

and observe that

$$D\theta(y_0,s_0)=D\psi(x_0,t_0) \quad \text{and} \quad |D\psi(x_0,t_0)|\le 2L. \tag{4.39}$$

Since u is a supersolution, and $u-\theta$ takes its minimum at (y_0,s_0),

$$\theta_s(y_0,s_0)+H(\theta(y_0,s_0),y_0)\ge 0 \tag{4.40}$$

unless $y_0\in\partial\Omega$ and $u(y_0,s_0)=g(y_0)$. Lemma 4.2 applies for $U^*-\psi$, so

(4.41)

$$\begin{aligned}\frac{M}{2T} &= \psi_t(x_0,t_0)-\theta_t(y_0,t_0)\\ &\le -H(D\psi(x_0,t_0),x_0)+H(D\theta(y_0,s_0),y_0)+\left(\frac{1}{\epsilon^2}+C\left(1+\frac{\operatorname{diam}(\Omega)}{\epsilon^2}+\frac{1}{\epsilon^2}\right)\right)\lambda\\ &\le C_{2L}|x_0-y_0|+\left(\frac{1}{\epsilon^2}+C\left(1+\frac{\operatorname{diam}(\Omega)}{\epsilon^2}+\frac{1}{\epsilon^2}\right)\right)\lambda\\ &\le C_{2L}2L\epsilon^2+\frac{1+C(1+\operatorname{diam}(\Omega)+\epsilon^2)}{\epsilon^2}\lambda\end{aligned}$$

whenever $t_0,s_0>0$.

If $t_0=0$ or $s_0=0$ we may use (3.1), (4.21), (4.34), and (4.35) to prove

(4.42)

$$M\le\left(C_S(L_0)\left(L_t+\frac{R}{T}\right)+2L_0L\right)\epsilon^2 \quad \text{or} \quad M\le\left(L_t\left(L_t+\frac{R}{T}\right)+2L_0L\right)\epsilon^2.$$

We consider $t_0\ge 0$ and $s_0=0$, which gives the first inequality of (4.42). The second is similar. From (4.34) we have

$$M/2\le U^*(x_0,t_0)-u(y_0,0)=U^*(x_0,t_0)-u_0(y_0). \tag{4.43}$$

Hence

$$M/2\le |U^*(x_0,t_0)-u_0(x_0)|+|u_0(x_0)-u_0(y_0)|\le C_S(L_0)t_0+L_0|x_0-y_0|, \tag{4.44}$$

which implies the first inequality by (4.35).

It remains to consider the case $y_0\in\partial\Omega$ and $u(y_0,s_0)=g(y_0)=u_0(y_0)$. From (4.34), (4.22), and (4.35) we have

$$\begin{aligned}M/2 &\le U^*(x_0,t_0)-u(y_0,s_0)\le u_0(y_0)+\tilde{L}_S|x_0-y_0|-u_0(y_0)\\ &\le 2\tilde{L}_SL\epsilon^2.\end{aligned} \tag{4.45}$$

Putting these estimates together and setting $\epsilon = \lambda^{1/4}$ gives

$$M \le K\Big(\|u_0\|, L_0, L, L_S, C, C_0, \tilde{C}_0, C_{2L}, L_t, C_S(L_0), T\Big)(\lambda^{\frac{1}{2}} + \lambda). \tag{4.46}$$

To estimate $U - u$ from below we consider

$$\Psi(x,t,y,s) = u(x,t) - U_*(y,s) - \frac{(x-y)^2}{2\epsilon^2} - \frac{(t-s)^2}{2\epsilon^2} - \frac{t+s}{4T}M \tag{4.47}$$

and use essentially the same argument.

5. Example. The proposed abstract scheme is not of much value if a numerically (or theoretically) useful one does not exist. In this section we will develop a Godunov-type scheme for an important example.

We will make the additional assumption on H that there exists $M_R > 0$ such that

$$|H(p,x) - H(q,x)| \le M_R|p-q| \qquad \forall x \in \bar{\Omega},\ \{p,q \in R^n :\ |p|,|q| \le R\}. \tag{5.1}$$

Also, we replace Assumption 1 with the existence of a $C_H \ge 0$ such that

$$|H(p,x) - H(p,y)| \le C_H(1+|p|)|x-y|; \tag{5.2}$$

however, we use this only for deriving a global bound on $|DU|$. Several monotone schemes for the Cauchy problem are given in [15]. Among them the Godunov scheme seems the most convenient to generalize to boundary problems.

5.1. One spatial dimension. Let us first concentrate on the one-dimensional case $\Omega = (0,1)$, $g(x) = 0$. The mapping $p \to H(p,x)$ is assumed to have a unique local minimum at $p = 0$ for all x. The Godunov scheme $\hat{S}^h(\rho)$ is defined here by

$$\hat{S}^h(\rho)u(x) = u(x) - \rho\hat{H}(\delta^- u(x), \delta^+ u(x), x) \tag{5.3}$$

for $h \le x \le 1-h$, where

$$\hat{H}(\alpha,\beta,x) = \begin{cases} \min_{\alpha \le p \le \beta} H(p,x), \alpha \le \beta, \\ \max_{\beta \le p \le \alpha} H(p,x), \beta \le \alpha, \end{cases} \tag{5.4}$$

and we have employed the one-sided difference operators

$$\delta^- u(x) = \frac{u(x) - u(x-h)}{h}, \quad \delta^+ u(x) = \frac{u(x+h) - u(x)}{h}. \tag{5.5}$$

More explicitly, for the Hamiltonian we study

$$\hat{H}(\alpha,\beta,x) = \begin{cases} H(\alpha,x), & \alpha,\beta \ge 0, \\ H(\beta,x), & \alpha,\beta > 0, \\ H(0,x), & \alpha < 0 < \beta, \\ \max(H(\alpha,x), H(\beta,x)), & \beta < 0 < \alpha. \end{cases} \tag{5.6}$$

At $x = 0$, we take

$$\hat{S}^h(\rho)u(0) = \min\left(u(0) - \rho H^-(\delta^+ u(0), 0),\ 0\right), \tag{5.7}$$

where $H^-(p,x) = H(\min(0,p),x) = \hat{H}(0,p,x)$. At $x=1$ we similarly take

$$\hat{S}^h(\rho)u(1) = \min\left(u(1) - \rho H^+(\delta^- u(1),1),\ 0\right), \tag{5.8}$$

where $H^+(p,x) = H(\max(0,p),x) = \hat{H}(p,0,x)$.

In order to use the convergence theory, we need to define $\hat{S}^h$ for all $0 \le x \le 1$. We will do this by extending u to $[-h, 1+h]$ and then applying the interior operator given by (5.3), from now on denoted by $\hat{S}_0^h$. We consider only $x=0$; the case $x=1$ is entirely similar.

We let $u(x) = u(0)$ for $-h \le x < 0$. This choice implies that $\hat{S}_0^h(\rho)u(0) = \hat{S}^h(\rho)u(0)$ whenever $\hat{S}^h(\rho)u(0) < 0$. The scheme $\hat{S}_0^h$ then satisfies all conditions of Theorem 4.1 except (4.6) and (4.9). In order to take the boundary condition into account, we take for $0 \le x < h$

$$\hat{S}^h(\rho)u(x) = \min\left(\hat{S}_0^h(\rho)u(x), \hat{L}_S x\right), \quad \tilde{L}_S = L_0 = \max\left(\|(u_0)_x\|, p_0\right). \tag{5.9}$$

If $\hat{S}_0^h(\rho)u(x) > \hat{L}_S x$, we can use that $\hat{L}_S x$ is a supersolution to (1.1). We can let B_g consist only of functions $u(x) \le U_S$ with $U_S(x) = \hat{L}_S \min(x, 1-x)$, since $u_0(x) \le U_S$, and $\hat{S}^h(\rho)U_S \le U_S$.

PROPOSITION 5.1. *The approximation scheme $\hat{S}^h : C^{0,1}([0,1]) \to C^{0,1}$ satisfies the conditions of Theorem* 4.1 *with $\lambda = h$, $C_S(L) = \sup_{|p|\le L,\, x\in\Omega} |H(p,x)|$, $\tilde{C}_0 = C_0, L_S = L_0$, provided*

$$\Delta t M_R \le h \quad \textit{for} \quad R = \|u_x\|. \tag{5.10}$$

Furthermore

$$\left\|(\hat{S}^h u)_x\right\| \le \max(\|u_x\|, L_0) + C_H\left(1 + \max(\|u_x\|, L_0)\right)\Delta t. \tag{5.11}$$

Proof. This result is standard except for the Lipschitz continuity (5.11). For that we need the next lemma.

LEMMA 5.2. *The Godunov Hamiltonian $\hat{H}$ satisfies*

$$|\hat{H}(\alpha,\beta,y) - \hat{H}(\alpha,\beta,x)| \le C_H(1 + \max(|\alpha|,|\beta|))|y-x|. \tag{5.12}$$

This follows easily from (5.4). Using our extension of u to $[-h,1+h]$, we first find a bound $\|(\hat{S}_0^h u)_x\|$ for the interior operator $\hat{S}_0^h$ for all $x \in [0,1]$. Define the translation operator τ_δ by $\tau_\delta u(x) = u(x+\delta)$. By Lemma 5.2

$$|\tau_\delta \hat{S}_0^h u(x) - \hat{S}_0^h \tau_\delta u(x)| \le C_H\left(1 + \max(\|u_x\|, p_0)\right)\Delta t\delta \tag{5.13}$$

whenever $x, x+\delta \in [0,1]$. Also $\hat{S}_0^h$ commutes with addition of constants ($\hat{S}_0^h(u+k) = \hat{S}_0^h(u)+k$), which from $u \le v + \|u-v\|$ and monotonicity implies $\|\hat{S}_0^h u - \hat{S}_0^h v\| \le \|u-v\|$. We can then conclude

$$\begin{aligned} |\hat{S}_0^h u(x+\delta) - \hat{S}_0^h u(x)| &\le |\hat{S}_0^h \tau_\delta u(x) - \hat{S}_0^h u(x)| + |\tau_\delta \hat{S}_0^h u(x) - \hat{S}_0^h \tau_\delta u(x)| \\ &\le \|u(x+\delta) + u(x)\| + C_H\left(1 + \max(\|u_x\|, p_0)\right)\Delta t\delta. \end{aligned} \tag{5.14}$$

The estimate (5.11) then follows immediately. ∎

As these considerations demonstrate, finding a scheme with the right properties in the vicinity of the boundary is a theoretical, but not computational, problem. For

the analysis, we just need to know that a monotone and consistent scheme exists everywhere.

We conclude as follows.

PROPOSITION 5.3. *Let $\Omega = (0,1)$, and let the mapping $p \to H(p,x)$ have a unique minimum at $p = 0$ for all x. Then (1.1)–(1.2) has a unique viscosity solution u in $C^0(\bar{\Omega} \times [0,T])$. Furthermore, the scheme $\hat{S}^h$ defines an approximation $U(x,t)$ such that $\|U - u\| \le O(\sqrt{h})$ if the CFL-condition $\Delta t M_R \le h$ holds with*

$$R = \|U_x(\cdot,t)\| \le (L_0 + 1)\exp C_H t - 1. \tag{5.15}$$

The estimate on (5.15), U_x, follows from (5.11). Observe that (5.11) applied inductively constitutes an Euler approximation to the ordinary differential equation

$$\max(\|U_x\|, L_0)_t = C_H(1 + \max(\|U_x\|, L_0)).$$

5.2. Two spatial dimensions. We will try to produce a two-dimensional example scheme using the same strategy. Consider now the spatial variables $(x,y) \in \Omega = (0,1)^2$ and a Hamiltonian

$$H(u_x, u_y) = h(u_x^2, u_y^2), \tag{5.16}$$

where h is increasing in both variables. Assumption 4 then clearly holds. As boundary condition we consider $g(x) = 0$. The Godunov scheme may be generalized to

$$\hat{S}^h(\rho)u = u - \rho\hat{H}(\delta_x^- u, \delta_x^+ u, \delta_y^- u, \delta_y^+ u) \tag{5.17}$$

with

$$\hat{H}(p_-,p_+,q_-,q_+) = \begin{cases} \min_{q_- \le q \le q_+} & \begin{cases} \min_{p_- \le p \le p_+} H(p,q), & p_- \le p_+ \\ \max_{p_+ \le p \le p_-} H(p,q), & p_+ \le p_- \end{cases}, q_- \le q_+, \\ \max_{q_+ \le q \le q_-} & \begin{cases} \min_{p_- \le p \le p_+} H(p,q), & p_- \le p_+ \\ \max_{p_+ \le p \le p_-} H(p,q), & p_+ \le p_- \end{cases}, q_+ \le q_-. \end{cases} \tag{5.18}$$

For the H given here, this becomes

$$\hat{H}((p_-,p_+,q_-,q_+) = H(p,q), \tag{5.19}$$

where

$$p = \begin{cases} p_+, & p_-, p_+ \le 0, \\ p_-, & p_-, p_+ \ge 0, \\ 0, & p_- < 0 < p_+, \\ \max(p_-,p_+), & p_+ < 0 < p_-, \end{cases} \qquad q = \begin{cases} q_+, & q_-, q_+ \le 0, \\ q_-, & q_-, q_+ \ge 0, \\ 0, & q_- < 0 < q_+, \\ \max(q_-,q_+), & q_+ < 0 < q_-. \end{cases} \tag{5.20}$$

At the boundaries we may simply replace the undefined finite differences with 0 in $\hat{H}$. That is,

$$\hat{S}_b^h(\rho)u = u - \rho \begin{cases} \hat{H}(0, \delta_x^+ u, \delta_y^- u, \delta_y^+ u), & x = 0,\ 0 < y < 1, \\ \hat{H}(\delta_x^- u, 0, \delta_y^- u, \delta_y^+ u), & x = 1,\ 0 < y < 1, \\ \hat{H}(\delta_x^- u, \delta_x^+ u, 0, \delta_y^+ u), & y = 0,\ 0 < x < 1, \\ \hat{H}(\delta_x^- u, \delta_x^+ u, \delta_y^- u, 0), & y = 1,\ 0 < x < 1, \end{cases} \tag{5.21}$$

and take

$$\hat{S}^h = \min(0, \hat{S}_b^h). \tag{5.22}$$

At the corners we have even less information, so we take

$$\hat{S}_b^h(\rho)u = u - \rho \begin{cases} \hat{H}(0, \delta_x^+ u, 0, \delta_y^+ u), & x = 0,\ y = 0, \\ \hat{H}(\delta_x^- u, 0, 0, \delta_y^+ u), & x = 1,\ y = 0, \\ \hat{H}(\delta_x^- u, 0, \delta_y^- u, 0), & x = 1,\ y = 1, \\ \hat{H}(0, \delta_x^+ u, \delta_y^- u, 0), & x = 0,\ y = 1. \end{cases} \tag{5.23}$$

This defines an approximation for $(x, y) \in [h, 1-h]^2 \cup \partial\Omega$. As in one dimension we need to do some work near the boundary. However, our strategy there carries over straightforwardly.

Given a function $u(x, y)$ defined on $[0, 1]^2$, we want to extend u to $(x, y) \in [-h, 0) \times [0, 1]$. We set $u(x, y) = u(0, y)$ there, and similarly around the rest of Ω, and apply the interior operator $\hat{S}_0^h$. In order to take the boundary condition into account, we take for $0 \le x < h$

$$\hat{S}^h(\rho)u(x, y) = \min\left(\hat{S}_0^h(\rho)u(x, y),\, \hat{U}_S(x, y)\right) \tag{5.24}$$

with

$$\hat{U}_S(x, y) = \hat{L}_S \min(x, y, 1-x, 1-y), \qquad \hat{L}_S = L_0 = \max(p_0, L_0), \tag{5.25}$$

which is a supersolution to (1.1)–(1.2). Since $\hat{S}^h(\rho)U_S \le U_S$, and $u_0 \le U_S$, we can let B_g only consist of functions such that $u \le U_S$. We may use the same method as in one dimension to estimate $\|D\hat{S}^h(\rho)u\|$.

Now we can conclude that the conditions of Theorem 4.1 also holds in this case, hence the next proposition.

PROPOSITION 5.4. *Let $\Omega = (0, 1)^2$ and H be defined by* (5.16). *Then* (1.1)–(1.2) *has a unique viscosity solution u in $C^0(\bar{\Omega} \times [0, T])$. Furthermore, the scheme $\hat{S}^h$ defines an approximation $U(x, t)$ such that $\|U - u\| \le O(\sqrt{h})$ if the CFL-condition $\Delta t C_{CFL} \le h$ holds for*

$$C_{CFL} = \max_{|p| \le \|U_x\|,\, |q| \le \|U_y\|} \left(|H_p(p, q)| + |H_q(p, q)| \right) \tag{5.26}$$

with

$$\max\left(\|U_x\|, \|U_y\|\right) \le (L_0 + 1) \exp C_H t - 1. \tag{5.27}$$

5.3. Numerical results. This two-dimensional scheme was implemented and tested. The domain was still $\Omega = (0, 1)^2$, while H and the boundary conditions varied. Some effects specific to the Cauchy–Dirichlet problem were noted, but in all cases we found that the error is governed by already well-known numerical phenomena.

5.3.1. Elliptic paraboloid. As an illustrative example, we chose $H(u_x, u_y) = u_x^2 + \frac{1}{4}u_y^2 - 1$ and $u_0(x, y) = -2u_s(x)$, where u_s is the stationary solution given by $u_s = \min(x, 1-x, 2y, 2(1-y))$. Figure 5.1 shows the results with discretization parameters

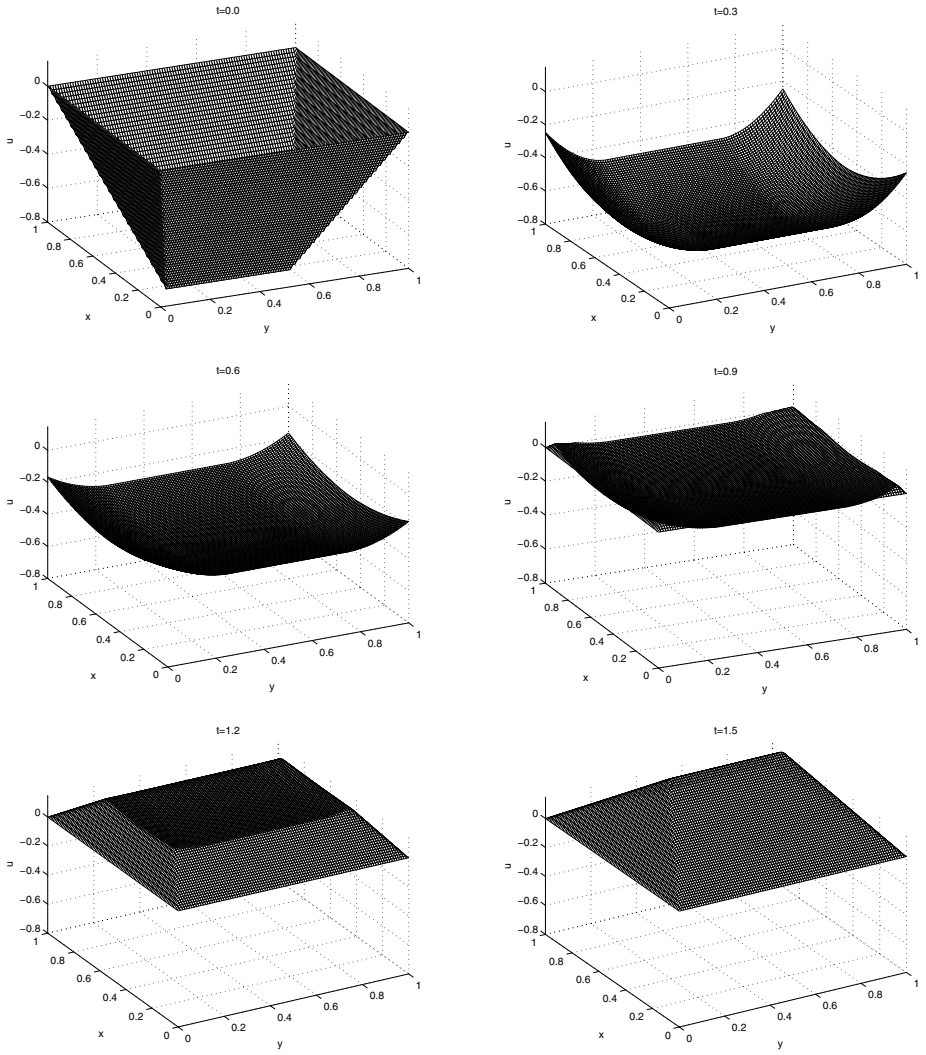


FIG. 5.1.

$h = 0.01$ and $\Delta t = \frac{h}{5}$. At first U satisfies the boundary condition only weakly, but at $t = 1.7$ it has reached the approximate stationary solution. The results look reasonable, and there is no sign that the weak boundary statement adds significantly to the error expected for a pure Cauchy or pure Dirichlet problem.

5.3.2. Piecewise linear H. The Hamiltonian $H(p,q) = |p| + |q| - 1$ allows simple explicit solutions. First, $u_s = \min(x, 1-x, y, (1-y))$ is the stationary solution. Starting with $u_0 = -u_s$, the solution will be $u(x,y,t) = \max(u_0(x,y), t - 0.5)$ until $t = 0.5$, as shown in Figure 5.2 to the left. At most points this is easy to check, but it should be proved for the line segment given by $x = y$ and $x \le 0.5$. At $x < t - 0.5$, it follows by the subdifferential

$$D^- u(x,x,t) = \{(p,q) : |p| + |q| = 1, q \le 0, p \le 0\}. \tag{5.28}$$

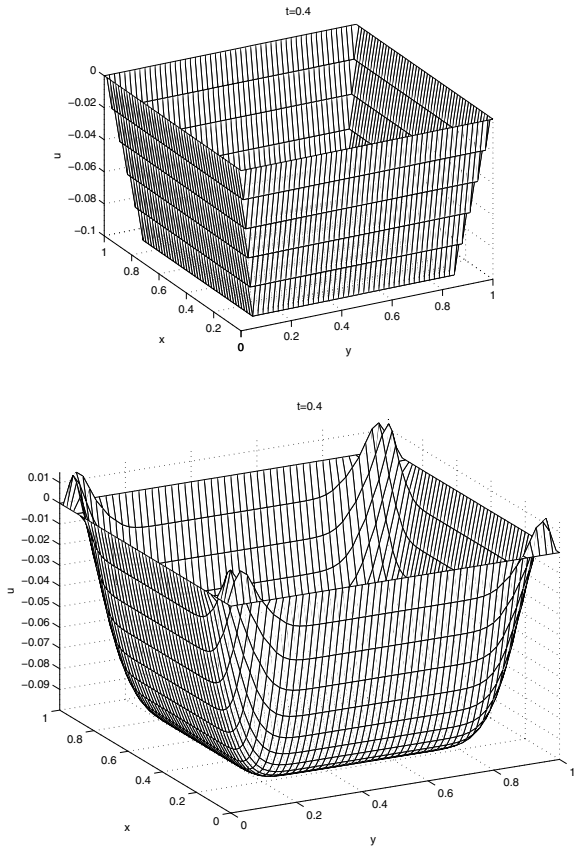


FIG. 5.2.

At $x = y = t - 0.5 = \hat{x}$, we have

$$u(\hat{x},\hat{x},t) - \psi(\hat{x},\hat{x},t) \le u(\hat{x}+h,\hat{x}+h,t-h) - \psi(\hat{x}+h,\hat{x}+h,t-h), \tag{5.29}$$

implying

$$\psi(\hat{x}+h,\hat{x}+h,t-h) - \psi(\hat{x},\hat{x},t) \le -h. \tag{5.30}$$

Letting $h \to 0$, we can conclude that

$$-\psi_x(\hat{x},t) - \psi_y(x_h,t) + \psi_t(\hat{x},t) - 1 \ge 0. \tag{5.31}$$

Obviously, $D^- u(\hat{x},\hat{x},t) \subseteq (-\infty,0)^2$, and hence the supersolution condition holds. In the numerical experiments, however, numerical smoothing causes the stationary solution u_s to appear prematurely near the corners. On the right in Figure 5.2 this is seen for $h = 0.02$ and $t = 0.01$.

With more grid points, the area where this effect occurs shrinks. Hence, the pointwise error is eventually unaffected by the boundary. Figure 5.3 shows one corner from a simulation with $h = 0.0025$ and $\Delta t = h/2$. These data were simulated for several h (with $\Delta t = h/2$), and the $\log_2$-values of the maximal error over the grid points at $t = 0.4$ are given in Table 5.1. The numbers fit well with the theoretical convergence rate of $\frac{1}{2}$.

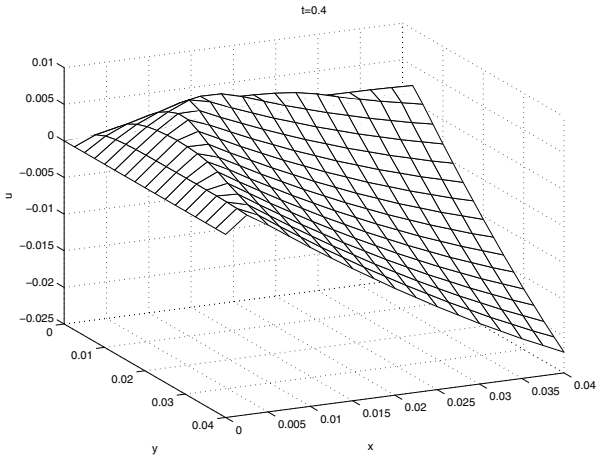


FIG. 5.3.

TABLE 5.1

h	Maximal error over grid	Observed convergence rate
0.04	0.0705	-
0.02	0.0501	0.493
0.01	0.0356	0.493
0.005	0.0252	0.498
0.0025	0.0178	0.502

5.3.3. Numerical diffusion of boundary values. Let $H(p,q) = |p| + |q| - 1$ as above. Then $u_s = \min(x,y)$ is a stationary solution if also $g(x,y) = \min(x,y)$. Figure 5.4 shows a simulation with u_s as initial data. The discretization parameters were $h = 0.02$ and $\Delta t = 0.01$. Due to numerical smoothing at the line $x = y$, the Dirichlet condition is not strongly satisfied in the discrete case.

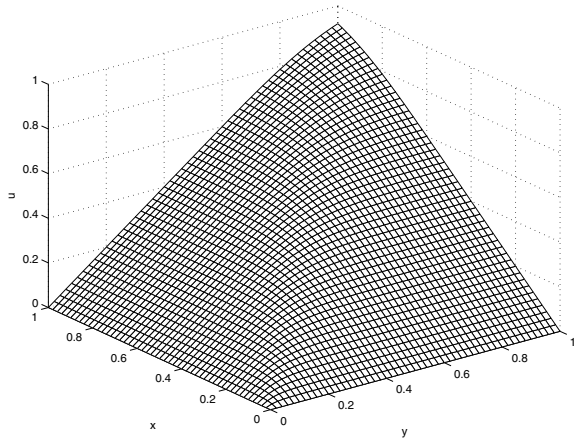


FIG. 5.4.

REFERENCES

[1] R. Abgrall, *Numerical discretization of boundary conditions for first order Hamilton–Jacobi equations*, SIAM J. Numer. Anal., 41 (2003), pp. 2233–2261.

[2] M. Bardi and I. Capuzzo-Dolcetta, *Optimal Control and Viscosity Solutions of Hamilton–Jacobi–Bellman Equations*, in Systems & Control: Foundations & Applications, Birkhäuser, Boston, 1997. With appendices by Maurizio Falcone and Pierpaolo Soravia.

[3] G. Barles, *Regularity results for first order Hamilton–Jacobi equations*, Differential Integral Equations, 3 (1990), pp. 103–125.

[4] G. Barles, *Uniqueness and regularity results for first-order Hamilton–Jacobi equations*, Indiana Univ. Math. J., 39 (1990), pp. 443–466.

[5] G. Barles and B. Perthame, *Comparison principle for Dirichlet-type Hamilton–Jacobi equations and singular perturbations of degenerated elliptic equations*, Appl. Math. Optim., 21 (1990), pp. 21–44.

[6] G. Barles and P. E. Souganidis, *Convergence of approximation schemes for fully nonlinear second order equations*, Asympt. Anal., 4 (1991), pp. 271–283.

[7] I. Capuzzo Dolcetta and F. Leoni, *On the Vanishing Viscosity Approximation of a Time Dependent Hamilton–Jacobi Equation*, in Recent Trends in Nonlinear Analysis, Birkhäuser, Basel, 2000, pp. 59–75.

[8] M. G. Crandall, *Viscosity Solutions: A Primer*, in Viscosity Solutions and Applications, Springer, Berlin, 1997, pp. 1–43.

[9] M. G. Crandall, H. Ishii, and P.-L. Lions, *User's guide to viscosity solutions of second order partial differential equations*, Bull. Amer. Math. Soc. (N.S.), 27 (1992), pp. 1–67.

[10] M. G. Crandall and P.-L. Lions, *Viscosity solutions of Hamilton–Jacobi equations*, Trans. Amer. Math. Soc., 277 (1983), pp. 1–42.

[11] M. G. Crandall and P.-L. Lions, *Two approximations of solutions of Hamilton–Jacobi equations*, Math. Comp., 43 (1984), pp. 1–19.

[12] M. G. Crandall and R. Newcomb, *Viscosity solutions of Hamilton–Jacobi equations at the boundary*, Proc. Amer. Math. Soc., 94 (1985), pp. 283–290.

[13] W. H. Fleming and H. M. Soner, *Controlled Markov Processes and Viscosity Solutions*, Appl. Math. 25, Springer-Verlag, New York, 1993.

[14] H. Ishii, *A boundary value problem of the Dirichlet type for Hamilton–Jacobi equations*, Ann. Scuola Norm. Sup. Pisa Cl. Sci., 16 (1989), pp. 105–135.

[15] S. Osher and C.-W. Shu, *High-order essentially nonoscillatory schemes for Hamilton–Jacobi equations*, SIAM J. Numer. Anal., 28 (1991), pp. 907–922.

[16] P. E. Souganidis, *Approximation schemes for viscosity solutions of Hamilton–Jacobi equations*, J. Differential Equations, 59 (1985), pp. 1–43.

[17] P. E. Souganidis, *A remark about viscosity solutions of Hamilton–Jacobi equations at the boundary*, Proc. Amer. Math. Soc., 96 (1986), pp. 323–329.

SIAM J. NUMER. ANAL.
Vol. 46, No. 5, pp. 2393–2410
© 2008 Society for Industrial and Applied Mathematics

FINITE ELEMENT APPROXIMATION OF THE FAST DIFFUSION AND THE POROUS MEDIUM EQUATIONS*

CARSTEN EBMEYER† AND W. B. LIU‡

Abstract. The fast diffusion equation $u_t = \Delta(|u|^{m-1}u)(0 < m < 1)$ and the porous medium equation $(1 < m < \infty)$ are studied in a parabolic cylinder $\Omega \times (0, T)$. A fully discrete Galerkin approximation is considered using C^0-piecewise linear finite elements in space and the backward Euler time discretization. A priori error estimates in quasi norms and rates of convergence are proved.

Key words. error estimates, fully discrete Galerkin scheme, free boundary problem, fast diffusion, slow diffusion, porous media

AMS subject classifications. 65M15, 65M60

DOI. 10.1137/060657728

1. Introduction.

Let $\Omega \subset \mathbb{R}^d$, $d \geq 2$, be a bounded polyhedral domain and $T < \infty$. In this paper we are concerned with the finite element approximation of the degenerate parabolic equation

$$\begin{aligned} &u_t = \Delta(|u|^{m-1}u) && \text{in } \Omega \times (0, T], \\ &u(x, t) = 0 && \text{on } \partial\Omega \times (0, T], \\ &u(x, 0) = u_0(x) && \text{in } \Omega, \end{aligned} \tag{1.1}$$

where $u(x, t) : \Omega \times [0, T] \to \mathbb{R}$. We treat the *fast diffusion* case $(0 < m < 1)$ and the *slow diffusion* case $(1 < m < \infty)$, that is, the fast diffusion equation and the porous medium equation. We consider a fully discrete Galerkin approximation using C^0-piecewise linear finite elements in space and the backward Euler discretization in time. Our aim is to prove a priori error estimates and rates of convergence.

We may rewrite (1.1) in the form $u_t = \operatorname{div}(m|u|^{m-1}\nabla u)$. In the fast diffusion case $(0 < m < 1)$, the modulus of ellipticity $m|u|^{m-1}$ blows up whenever u vanishes. Hence, the fast diffusion equation is a singular equation. The following phenomenon is known: Each weak solution decays to zero in some finite time; see, e.g., [5, 9]. In particular, weak solutions may have singularities near $\partial\Omega$ or at the extinction time.

The porous medium equation $(1 < m < \infty)$ is not uniformly elliptic as it degenerates at unknown points where $u = 0$. An interesting feature is that the support of $u(\cdot, t)$ has a free boundary with a finite speed of propagation; see, e.g., [6, 33]. The porous medium equation is used to model a process obeying a Darcy law. It is one of the prototypes of degenerate parabolic equations modeling a free boundary with a finite speed of propagation. Further typical examples are the parabolic p–Laplace equation for $2 < p < \infty$,

$$u_t = \operatorname{div}(|\nabla u|^{p-2}\nabla u), \quad u|_{\partial\Omega} = 0, \quad u(x, 0) = u_0(x),$$

or the doubly nonlinear parabolic equations; cf. [14, 15, 16] and the references given therein. Even if the data are smooth, the free boundary may have singularities. In particular, weak solutions are not smooth at the free boundary.

*Received by the editors April 20, 2006; accepted for publication (in revised form) June 19, 2007; published electronically June 6, 2008.
http://www.siam.org/journals/sinum/46-5/65772.html
†Mathematisches Seminar, Universität Bonn, Nussallee 15, D-53115 Bonn, Germany (ebmeyer.c@uni-bonn.de).
‡CBS & IMS, University of Kent, Canterbury, CT2 7NF, England (W.B.Liu@kent.ac.uk).

For such degenerate or singular equations many existing techniques in the finite element method do not work well. The orders of a priori error bounds are frequently suboptimal and often tend to zero when the equations' parameters (like m) approximate some special values like zero or ∞. Furthermore, there are gaps between the regularities required for deriving error bounds and the known regularity achievable for sufficiently smooth data; see section 4 for more detailed explanations. In this sense all the known a priori estimates for (1.1) are suboptimal. First error estimates for fully discrete schemes for the porous medium equation can be found in [30]; regularization or linearization procedures are used in [28, 30] and [20, 21, 27, 29]; schemes for the degenerate equation itself are considered in [1, 12, 22, 31]. Semidiscretizations in time for the porous medium equation and the fast diffusion equation are studied in [17] and [23, 24]; fully discrete schemes for the fast diffusion equation are discussed in [22, 31]. We will give more detailed reviews later.

In this paper we introduce a quasi norm and prove a priori error estimates of the finite element approximation error in this quasi norm. Further, due to new regularity results in fractional order spaces (see [13]), for suitable data we obtain, for example, the convergence rate

$$\left(\int_0^T \int_\Omega |u-U||w-W|\right)^{\frac{1}{2}} + \left\| \int_0^T (w-W) \right\|_{H^1(\Omega)} \le c(\tau+h),$$

where $w := |u|^{m-1}u$, U and W are finite element approximations of u and w, τ is the step size, and h is the mesh size.

Such a quasi norm naturally arises in degenerate problems. In fact, the quasi-norm approach is one of the significant developments in error estimation for degenerate equations. For example, it provides optimal error estimates for the p-Laplacian; cf. [2, 3, 10, 14].

The outline of this paper is as follows. In section 2 the assumptions on the data are given and some properties of weak solutions are discussed. The discrete problem is stated in section 3. The quasi norm is introduced in section 4, and some useful inequalities are shown. The a priori error estimates are proved in section 5, and the convergence rate is estimated in section 6.

2. The continuous problem. The following assumptions are made on the data.

(H1) $\Omega \subset \mathbb{R}^d$, $d \ge 2$, is a bounded, convex polyhedral domain.

(H2) $\partial\Omega = \bigcup_{1\le k\le M} \overline{\Gamma_k}$, where each Γ_k is a $(d-1)$-dimensional polyhedron, and $\Gamma_i \cap \Gamma_k = \emptyset$ for $i \ne k$.

(H3) $\partial\Gamma_{k1} \cap \ldots \cap \partial\Gamma_{kj} = \emptyset$ if $j > d$ and $k_1 < \cdots < k_j$ (that is, there are at most d adjacent faces Γ_k).

(H4) $u_0 \in L^\infty(\Omega)$.

We call $u(x,t)$ a weak solution of (1.1) if

$$u \in L^\infty(0,T;L^\infty(\Omega)), \quad |u|^{m-1}u \in L^2(0,T;W_0^{1,2}(\Omega)),$$

and

$$-\int_0^T \int_\Omega u\,\varphi_t + \int_0^T \int_\Omega \nabla(|u|^{m-1}u)\nabla\varphi = \int_\Omega u_0\varphi_0 \tag{2.1}$$

for all $\varphi \in L^2(0,T;W_0^{1,2}(\Omega)) \cap W^{1,1}(0,T;L^1(\Omega))$ satisfying $\varphi(\cdot,T) \equiv 0$ in Ω, where $\varphi_0 := \varphi(\cdot,0)$.

It is well known that under these assumptions there exists a unique weak solution; see, e.g., [18]. Let us note that $u \in L^\infty(0,T;L^\infty(\Omega))$ follows from a comparison argument [18]. Now we discuss the regularity of the weak solution under the assumption that $w_0 \in L^\infty(\Omega) \cap H_0^1(\Omega)$, where $w := |u|^{m-1}u$. It is known that

$$w \in L^\infty(0,T;W^{1,2}(\Omega)) \qquad \forall m > 0 \tag{2.2}$$

and

$$\partial_t(|u|^{\frac{m-1}{2}}u) \in L^2(0,T;L^2(\Omega)) \qquad \forall m > 0; \tag{2.3}$$

see [18]. Further, in [13] it is shown that

$$\int_0^t w(\cdot,\tau)dt \in L^\infty(0,T;W^{2,2}(\Omega)) \qquad \forall m > 0; \tag{2.4}$$

thus, $\int_\Omega(\int_0^T \Delta w)^2 < \infty$. Furthermore, in [13] regularity of w in fractional order Nikolskii spaces $\mathcal{N}^{s,p}(\Omega)$ is proved: It holds that

$$w \in L^{\frac{m+1}{m}}(0,T;\mathcal{N}^{\frac{2m}{m+1},\frac{m+1}{m}}(\Omega)) \qquad \forall m \in (0,1) \tag{2.5}$$

and

$$w \in L^{\frac{m+1}{m}}(\delta,T;\mathcal{N}^{\frac{2m}{m+1},\frac{m+1}{m}}(\Omega)) \qquad \forall m > 1, \forall \delta > 0 \quad \text{if } u_0 \geq 0. \tag{2.6}$$

Moreover, in [4] it is shown that

$$w \in L^{\frac{m+1}{m}-\varepsilon}(\delta,T;W^{2,\frac{m+1}{m}-\varepsilon}(\Omega)) \qquad \forall \delta,\varepsilon > 0 \tag{2.7}$$

if $m > 1$ and $u_0 \geq 0$. This is an L^p-property of the second derivatives of w.

Remark. The imbedding theorem of Nikolskii spaces into Sobolev spaces

$$\mathcal{N}^{s,p}(\Omega) \to W^{s-\varepsilon,p}(\Omega) \qquad \forall \varepsilon > 0$$

implies regularity in fractional order Sobolev spaces. For instance, from (2.5) it follows that $w \in L^{\frac{m+1}{m}}(0,T;W^{\frac{2m}{m+1}-\varepsilon,\frac{m+1}{m}}(\Omega))$ for all $\varepsilon > 0$ if $m < 1$.

3. The discrete problem. We deal with a fully discrete scheme using continuous piecewise linear finite elements in space and the backward Euler time discretization. Let $\tau > 0$ be the step size of the time steps. We use the notation $t_n = n\tau (n = 0,\dots,N)$, $T = t_N$, $I_n = (t_{n-1},t_n)$, $v_n(x) = v(x,t_n)$, $\bar{v}^0(x) = v_0(x)$, and

$$\bar{v}^n(x) = \tau^{-1}\int_{I_n} v(x,s)ds \qquad \text{for } n \geq 1.$$

The domain Ω is a polyhedral domain. Let T_h be a family of decompositions of Ω into closed d-simplices where h denotes the mesh-size. We suppose that T_h is a regular triangulation; i.e., the intersection of two nondisjoint nonidentical elements of T_h is a common vertex or edge or surface, and there is a constant $c > 0$ such that

$$|K| \geq c(\operatorname{diam} K)^d$$

for all simplices $K \in T_h$. We introduce the space

$$S_h(\Omega) = \{\phi_h \in C^0(\overline{\Omega}) : \phi_h \text{ is piecewise linear w.r.t. } T_h, \phi_h = 0, \text{ on } \partial\Omega\}.$$

Let $\Pi_h v \in S_h$ denote the C^0-piecewise linear interpolant of the function v. Further, let $P_h : H^1(\Omega) \to S_h$ be the H^1-projection onto S_h defined by

$$(\nabla(v - P_h v), \nabla\chi) = 0 \qquad \forall \chi \in S_h.$$

Now we define the finite element approximation, Let $w := |u|^{m-1}u$. Thus, $\partial_t u = \Delta w$, and, in fact,

$$\partial_t \psi(w) = \Delta w, \tag{3.1}$$

where $\psi(s) := |s|^{\frac{1-m}{m}} s$ $(s \in \mathbb{R})$. Let the functions $W_n \in S_h$, $n = 1, 2, \ldots$, be the solutions of the algebraic system

$$\left(\frac{\psi(W_n) - \psi(W_{n-1})}{\tau}, \chi_n\right) + (\nabla W_n, \nabla\chi_n) = 0 \qquad \forall \chi \in S_h, \tag{3.2}$$

where $W_0 := \psi^{-1}(\Pi_h \psi(w_0))$.

Thus, the functions $W_n(x)$ are successively defined for $n = 0, 1, \ldots$, we now define the finite element approximation $W(x,t)$ of $w(x,t)$ by setting

$$W(x,t) = \begin{cases} W_0(x) & \text{if } t = 0, \\ W_n(x,t) & \text{if } t \in (t_{n-1}, t_n], 1 \le n \le N. \end{cases} \tag{3.3}$$

Let us note that the function $W(x,t)$ is piecewise constant with respect to t and discontinuous in t at the time points $t_0, \ldots, t_{N-1}$.

We conclude this section by proving the following results concerning existence, uniqueness, and regularity of W.

LEMMA 3.1. *There exist unique functions $W_1, \ldots, W_N \in S_h$ solving* (3.2).

Proof. Due to the theory of monotone operators it is easy to see that there exist solutions W_n for each $n \ge 1$; cf. [32]. Moreover, the functions W_n are unique. Indeed, suppose that W_n and $\hat{W}_n (n \ge 1)$ are solutions of (3.2). Putting $\chi_n = \tau^2 \sum_{k=n}^N (W_k - \hat{W}_k)$ in (3.2) and summing up it follows that

$$\tau \sum_{n=1}^N (\psi(W_n) - \psi(\hat{W}_n), W_n - \hat{W}_n) \le \left(\psi(W_0) - \psi(\hat{W}_0), \tau \sum_{k=1}^N (W_k - \hat{W}_k)\right).$$

Noting that $\psi(W_0) = \psi(\hat{W}_0)$,

$$(\psi(W_n) - \psi(\hat{W}_n), W_n - \hat{W}_n) \ge c \left\| \psi(W_n) - \psi(\hat{W}_n) \right\|_{L^{m+1}(\Omega)}^{\frac{m+1}{m}}$$

if $m > 1$, and

$$(\psi(W_n) - \psi(\hat{W}_n), W_n - \hat{W}_n) \ge c \left\| W_n - \hat{W}_n \right\|_{L^{\frac{m+1}{m}}(\Omega)}^{\frac{m+1}{m}}$$

if $0 < m < 1$, the uniqueness follows. ∎

LEMMA 3.2. *It holds that*

$$\sup_{1 \le n \le N} \|W_n\|_{L^{\frac{m+1}{m}}(\Omega)}^{\frac{m+1}{m}} + \tau \sum_{n=1}^N \|\nabla W_n\|_{L^2(\Omega)}^2 \le c.$$

Proof. Multiplying the nth equation of (3.2) by $\chi_n = W_n$ and summing over n, we get

$$\sum_{n=1}^{N}(\psi(W_n), W_n) + \tau\sum_{n=1}^{N}(\nabla W_n, \nabla W_n) = \sum_{n=1}^{N}(\psi(W_{n-1}), W_n).$$

Young's inequality yields $|\psi(W_{n-1})Wn| \le \frac{1}{m+1}|W_{n-1}|^{\frac{1}{m+1}} + \frac{m}{m+1}|W_n|^{\frac{m+1}{m}}$. Noting that $\psi(W_n)W_n = |W_n|^{\frac{m+1}{m}}$, we obtain

$$\tfrac{1}{m+1}\|W_N\|_{\frac{m+1}{m}}^{\frac{m+1}{m}} + \tau\sum_{n=1}^{N}\|\nabla W_n\|_2^2 \le \tfrac{1}{m+1}\|W_0\|_{\frac{m+1}{m}}^{\frac{m+1}{m}}.$$

Clearly, $\|W_0\|_{\frac{m+1}{m}} \le c$. Thus, the assertion follows. ∎

4. A quasi norm and motivations. We first try to explain why for degenerate problems only suboptimal error estimates (either a priori or a posteriori) can be obtained via the standard approaches. Let, for instance,

$$a(u,v) = \int_\Omega (|\nabla u|^{p-2}\nabla u, \nabla v),\ (f,v) = \int_\Omega fv.$$

Then the weak formulation of the p–Laplace equation reads

$$a(u,v) = (f,v) \qquad \forall v \in W_0^{1,p}(\Omega).$$

In order to obtain optimal error estimates, the standard approaches depend on the facts (in the underlaying space norm)

$$a(u,u-v) - a(v,u-v) \ge c\|u-v\|^2,$$

$$|a(u,w) - a(v,w)| \le C\|u-v\|\|w\|,$$

which may be true for the nondegenerate equations but are untrue for the degenerate equations; see [7] and [26].

The so called *quasi-norm approach* looks for a quasi norm $|u-v|_u$ (or its power), which often depends on the solutions u to measure the approximation error so that

$$a(u,u-v) - a(v,u-v) \ge c|u-v|_u^2 \tag{4.1}$$

and

$$|a(u,w) - a(v,w)| \le C(\gamma|u-v|_u^2 + \delta(\gamma)|w|_w^2) \qquad \forall\gamma > 0. \tag{4.2}$$

This implies that

$$a(u,u-v) - a(v,u-v) \cong |u-v|_u^2.$$

With this quasi norm one can derive sharp a priori and a posteriori error estimates. For instance, one has the following basic a priori error bound for the finite element approximation of the p-Laplacian (see [2]):

$$|u-u_h|_u \le \inf_{v_h\in V^h} C|u-v_h|_u. \tag{4.3}$$

Then one can further obtain sharp error estimates with different norms; see [25] and [26]. For example, in the case of the p-Laplacian it is known (cf. [26]) that for all $x, y \in \mathbb{R}^n$

$$(|x|^{p-2}x - |y|^{p-2}y, x - y) \geq c(|x| + |x - y|)^{p-2}|x - y|^2,$$

$$||x|^{p-2}x - |y|^{p-2}||x - y)| \leq C(|x| + |x - y|)^{p-2}|x - y|^2.$$

This is the motivation for introducing the following quasi norm:

$$|u - v|_u^2 = \int_\Omega (|\nabla u| + |\nabla(u - v)|)^{p-2}|\nabla(u - v)|^2,$$

and with some technical inequalities one can show (4.1)–(4.2).

Deriving explicit sharp error bounds from (4.3) is often nontrivial as it needs sharp interpolation error estimates in the quasi norms, which are quite technical to obtain. Moreover, error bounds of the optimal order can be obtained only under sufficient regularity assumptions for the solutions. However, for degenerate PDEs like the p-Laplacian, regularities have upper limits, and higher regularity may not be achievable even for infinitely smooth data; see [2]. In [14], from (4.3) we obtained the optimal estimates in terms of both error order and regularity requirements via the quasi-norm approach. Let u and u_h be the solution of the p-Laplacian and its piecewise linear finite element approximation, respectively. Then

$$|u - u_h|_u^2 \leq Ch^2 \int_\Omega |\nabla u|^{p-2}|D^2 u|^2,$$

where the right integral is finite for sufficiently smooth data. In the same manner, for the parabolic p-Laplacian very recently it was shown that

$$||u - U||^2_{L^\infty(I,L^2(\Omega))} + \int_0^T |u - U|_u^2 dt \leq C(\tau^2 + h^2)$$

under appropriate assumptions on the data; cf. [10].

In order to treat the fast diffusion and the porous medium equation, we shall apply a similar idea to derive some a priori error bounds in a quasi norm, which is defined as follows: For $v_1, v_2 \in L^p(\Omega)$ and $p > 1$ let

$$||v_2||^2_{(v_1,p)} := \int_\Omega (|v_1| + |v_2|)^{p-2}|v_2|^2.$$

This is a quasi norm. In fact, it holds that $||v_2||_{(v_1,p)} \geq 0$, $||v_2||_{(v_1,p)} = 0$ if and only if $v_2 = 0$, and $||v_1 + v_2||_{(v_3,p)} \leq c(||v_1||_{(v_3,p)} + ||v_2||_{v_3,p})$ for any $v_1, v_2, v_3 \in L^p(\Omega)$.

Furthermore, it is easy to see that

$$c||v_2||^p_{L^p(\Omega)} \leq ||v_2||_{(v_1,p)} \leq ||v_2||^{p/2}_{L^p(\Omega)} \qquad \text{for } 1 < p < 2$$

and

$$||v_2||^{p/2}_{L^p(\Omega)} \leq ||v_2||_{(v_1,p)} \leq c'||v_2||^p_{L^p(\Omega)} \qquad \text{for } 2 < p < \infty,$$

Where $c, c' > 0$ are constants depending on $||v_1||_{L^p(\Omega)}$ and $||v_2||_{L^p(\Omega)}$.

In particular, we have (see Lemma 4.2 below)

$$\int_0^T \|w-v\|^2_{(w,\frac{m+1}{m})} \cong \int_0^T (\psi(w)-\psi(v), w-v)$$

$$\cong \int_0^T \|\psi(w)-\psi(v)\|^2_{(\psi(w),m+1)} = \int_0^T \|u-\psi(v)\|^2_{(u,m+1)} \tag{4.4}$$

for all $v \in L^{\frac{m+1}{m}}(\Omega\times[0,T])$, where $u=\psi(w)$.

Let us prove some useful inequalities. We write $A(r,s) \cong B(r,s)$ if there are constants $0<c_1<c_2$ independent of r and s such that $c_1A(r,s) \le B(r,s) \le c_2A(r,s)$.

LEMMA 4.1. *For all $r,s\in\mathbb{R}$ there hold the equivalences*

$$(\psi(r))-\psi(s))(r-s) \cong (|r|+|s|)^{\frac{1-m}{m}}|r-s|^2 \tag{4.5}$$

$$\cong \left||r|^{\frac{1-m}{2m}}r - |s|^{\frac{1-m}{2m}}s\right|^2. \tag{4.6}$$

Proof. Utilizing the Taylor expansion, we get

$$(\psi(r)-\psi(s))(r-s) = |r-s|^2\int_0^1 \psi(tr+(1-t)s)dt.$$

In view of the fact that (see [19])

$$\int_0^1 |tr+(1-t)s|^{\frac{1-m}{m}}dt \cong (|r|+|s|)^{\frac{1-m}{m}}, \tag{4.7}$$

the equivalence (4.5) follows.

Next, let $\phi(s) := |s|^{\frac{1-m}{m}}s$. The Taylor expansion yields

$$\phi(r)-\phi(s) = (r-s)\int_0^1 \phi'(tr+(1-t)s)dt.$$

It is easy to see that $\phi'(s)\cong|s|^{\frac{1-m}{2m}}$ for all $s\in\mathbb{R}$. Noting that $\int_0^1 tr+(l-t)s)|^{\frac{1-m}{2m}}dt \cong (|r|+|s|)^{\frac{1-m}{2m}}$, we obtain the equivalence (4.6). ∎

LEMMA 4.2. *For all $v_1,v_2\in L^{\frac{m+1}{m}}(0,T;L^{\frac{m+1}{m}}(\Omega))$ it holds that*

$$\int_0^T \|v_1-v_2\|^2_{(v_1,\frac{m+1}{m})} \cong \int_0^T (\psi(v_1)-\psi(v_2), v_1-v_2) \tag{4.8}$$

$$\cong \int_0^T \|\psi(v_1)-\psi(v_2)\|^2_{(\psi(v_1),m+1)}. \tag{4.9}$$

Proof. Utilizing the Taylor expansion and (4.7), we obtain for $r,s\in\mathbb{R}$

$$|\psi(r)-\psi(s)| \cong (|r|+|s|)^{\frac{1-m}{m}}|r-s|.$$

Noting that $(|r|+|s|)\cong(|r|+|r-s|)$, we get

$$(\psi(r)-\psi(s))(r-s) \cong (|r|+|r-s|)^{\frac{1-m}{m}}|r-s|^2.$$

Due to

$$\int_0^T \|v_1 - v_2\|^2_{(v_1, \frac{m+1}{m})} = \int_0^T \int_\Omega (|v_1| + |v_1 - v_2|)^{\frac{1-m}{m}} |v_1 - v_2|^2$$

the equivalence (4.8) follows.

Next, let $\bar{r} = \psi(r)$ and $\bar{s} = \psi(s)$. Then $(\psi(r) - \psi)(s))(r - s) = (\bar{r} - \bar{s})$ $(\psi^{-1}(\bar{r}) - \psi^{-1}(\bar{s}))$, where $\psi^{-1}(\bar{r}) = |\bar{r}|^{m-1}\bar{r}$. Arguing as above, we deduce equivalence (4.9). ∎

LEMMA 4.3. *There is a Constant c such that*

$$\begin{aligned} &(\psi(r_1) - \psi(r_3))(r_1 - r_3) \\ &\le c[(\psi(r_1) - \psi(r_2))(r_1 - r_2) + (\psi(r_2) - \psi(r_3))(r_2 - r_3)] \end{aligned} \tag{4.10}$$

for all $r_1, r_2, r_3 \in \mathbb{R}$.

Proof. The triangle inequality yields

$$\begin{aligned} &\left| |r_1|^{\frac{1-m}{2m}} r_1 - |r_3|^{\frac{1-m}{2m}} r_3 \right|^2 \\ &\le c\left(\left| |r_1|^{\frac{1-m}{2m}} r_1 - |r_2|^{\frac{1-m}{2m}} \right|^2 + \left| |r_2|^{\frac{1-m}{2m}} r_2 - |r_3|^{\frac{1-m}{2m}} r_3 \right|^2 \right). \end{aligned}$$

Due to (4.5) and (4.6) we obtain the assertion. ∎

LEMMA 4.4. *For any $\delta > 0$ there is a constant c_δ such that for all $r_1, r_2, r_3 \in \mathbb{R}$*

$$\begin{aligned} |\psi(r_1) - \psi(r_2)||r_1 - r_3| &\le \delta(\psi(r_1) - \psi(r_2))(r_1 - r_2) \\ &\quad + c_\delta(\psi(r_1) - \psi(r_3))(r_1 - r_3). \end{aligned} \tag{4.11}$$

Proof. It is well known that

$$(s_1 + s_2)^q s_2 s_3 \le \delta(s_1 + s_2)^q s_2^2 + c_\delta(s_1 + s_3)^q s_3^2$$

for all $s_1, s_2, s_3 \ge 0$ and $q > -1$; see, e.g., [3]. Putting $s_1 = |r_1|$, $s_2 = |r_1 - r_2|$, and $s_3 = |r_1 - r_3|$ and noting that

$$|\psi(r) - \psi(s)| \cong (|r| + |r - s|)^{\frac{1-m}{m}} |r - s|$$

the assertion follows. ∎

Let us conclude this section by discussing some approximation properties of quasi norms. Consider the elliptic problem

$$\psi(v) - \Delta v = f \quad \text{in } \Omega, \qquad v = 0 \quad \text{on } \partial\Omega,$$

where f is sufficiently smooth. Clearly, there is a unique weak solution satisfying

$$(\psi(v), \varphi) + (\Delta v, \Delta\varphi) = (f, \varphi) \qquad \forall \varphi \in W_0^{1,2}(\Omega). \tag{4.12}$$

Let $V \in S_h$ be the finite element approximation given by

$$(\psi(V), \chi) + (\nabla V, \nabla\chi) = (f, \chi) \qquad \forall \chi \in S_h. \tag{4.13}$$

Then there is a constant independent of h such that

$$\begin{aligned} &\|v - V\|^2_{(v, \frac{m+1}{m})} + \|\nabla v - \nabla V\|^2_{L^2(\Omega)} \\ &\le c \inf_{v_h \in S_h} \left(\|v - v_h\|^2_{(v, \frac{m+1}{m})} + \|\nabla v - \nabla v_h\|^2_{L^2(\Omega)} \right). \end{aligned} \tag{4.14}$$

This may be seen as follows. Let us choose $\varphi = \chi$ in (4.12). Taking the difference between (4.12) and (4.13), we get

$$(\psi(v) - \psi(V), \chi) + (\nabla v - \nabla V, \nabla \chi) = 0. \tag{4.15}$$

Let $v_h \in S_h$ be arbitrary. Utilizing (4.15), we deduce

$$\begin{aligned}&(\psi(v) - \psi(V), v - V) + (\nabla v - \nabla V, \nabla v - \nabla V)\\ &= (\psi(v) - \psi(V), v - v_h) + (\nabla v - \nabla V, \nabla v - \nabla v_h).\end{aligned}$$

Inequality (4.11) implies that

$$(\psi(v) - \psi(V), v - v_h) \le \delta(\psi(v) - \psi(V), v - V) + c_\delta(\psi(v) - \psi(v_h), v - v_h).$$

Further, we have

$$(\nabla v - \nabla V, \nabla v - \nabla v_h) \le \delta(\nabla v - \nabla V, \nabla v - \nabla V) + c_\delta(\nabla v - \nabla v_h, \nabla v - \nabla v_h).$$

Thus (4.14) follows. Therefore, *V is the best approximation of v with respect to the quasi norm* $(\|\cdot\|^2_{(v,\frac{m+1}{m})} + |\cdot|^2_{W^{1,2}(\Omega)})^{\frac{1}{2}}$.

5. The a priori error estimate. The aim of this section is to prove an a priori error estimate of $w - W$ in a suitable quasi norm. Our result is as follows.

THEOREM 5.1. *For any $m > 0$ there is a constant c independent of h and τ such that*

$$\begin{aligned}&\int_0^T \|w - W\|^2_{(w,\frac{m+1}{m})} + \|\nabla(\bar w - \overline{W})\|^2_{L^2(\Omega)}\\ &\le c\left(\sum_{n=1,}^{N}\int_{I_n}\|w_n - w\|^2_{(w,\frac{m+1}{m})} + \int_0^T \|w - P_h w\|^2_{(w,\frac{m+1}{m})}\right.\\ &\qquad \left. + \left\|\tau\sum_{n=1}^{N}(\nabla \bar w^n - \nabla P_h \bar w^n)\right\|^2_{L^2(\Omega)} + \|\psi(w_0) - \Pi_h\psi(w_0)\|^2_{L^2(\Omega)}\right),\end{aligned}$$

where $\bar w(x) = \int_0^T w(x,t)dt$ and $\overline{W}(x) = \int_0^T W(x,t)dt$.

Proof. Integrating (3.1) over I_n, we find

$$\tau^{-1}(\psi(w_n) - \psi(w_{n-1})) - \Delta \bar w^n = 0. \tag{5.1}$$

Let $\chi_n \in S_h$. We multiply (5.1) and (3.2) by $\tau\chi_n$, take the difference, and sum over n. We get

$$\sum_{n-1}^{N}(\psi(w_n) - \psi(W_n) - (\psi(w_{n-1}) - \psi(W_{n-1})), \chi_n) + \sum_{n=1}^{N}\tau(\nabla \bar w^n - \nabla W_n, \nabla \chi n) = 0.$$

Utilizing the identity $\sum_{n=1}^{N}(a_n - a_{n-1})b_n = a_N b_N - a_0 b_0 + \sum_{n=o}^{N-1} a_n(b_n - b_{n+1})$, we obtain

$$\begin{aligned}&(\psi(w_N) - \psi(W_N), \chi_N) + \sum_{n=0}^{N-1}(\psi(w_n) - \psi(W_n), \chi_n - \chi_{n+1})\\ &\quad + \sum_{n=1}^{N}\tau(\nabla \bar w^n - \nabla W_n, \nabla \chi_n) = (\psi(w_0) - \psi(W_0), \chi_0).\end{aligned}$$

We put $\chi_n = \tau \sum_{k=n}^{N} (P_h \bar{w}^k - \Pi_h W_k)$ for $0 \le n \le N$. Let us note that $\Pi_h W_k = W_k$ for all $k \ge 1$ and $\chi_n - \chi_{n+1} = \tau(P_h \bar{w}^n - \Pi_h W_n)$. We get

$$\begin{aligned}
&\tau \sum_{n=0}^{N} (\psi(w_n) - \psi(W_n), P_h \bar{w}^n - \Pi_h W_n) \\
&\quad + \tau \sum_{n=1}^{N} (\nabla \bar{w}^n - \nabla W_n, \tau \sum_{k=n}^{N} (\nabla P_h \bar{w}^k - \nabla W_k)) \\
&= (\psi(w_0) - \psi(W_0), \tau \sum_{n=0}^{N} (P_h \bar{w}^n - \Pi_h W_n)). \qquad (5.2)
\end{aligned}$$

Now we subtract $\tau(\psi(w_0) - \psi(W_0), P_h \bar{w}^0 - \Pi_h W_0)$ on both sides of (5.2). Notice that

$$\begin{aligned}
(\psi(w_n) - \psi(W_n), P_h \bar{w}^n - W_n) &= (\psi(w_n) - \psi(W_n), P_h \bar{w}^n - w_n) \\
&\quad + (\psi(w_n) - \psi(W_n), w_n - W_n).
\end{aligned}$$

Further, in view of the identity

$$\sum_{n=1}^{N} \left(a_n \sum_{k=n}^{N} a_k \right) = \frac{1}{2} \left(\sum_{n=1}^{N} a_n \right)^2 + \frac{1}{2} \sum_{n=1}^{N} (a_n)^2$$

and the fact that P_h is the H^1-projection, we have

$$\begin{aligned}
&\tau \sum_{n=1}^{N} \left(\nabla \bar{w}^n - \nabla W_n, \tau \sum_{k=n}^{N} (\nabla P_h \bar{w}^k - \nabla W_k) \right) \\
&= \tau \sum_{n=1}^{N} \left(\nabla P_h \bar{w}^n - \nabla W_n, \tau \sum_{k=n}^{N} (\nabla P_h \bar{w}^k - \nabla W_k) \right) \\
&\ge \frac{1}{2} \int_\Omega \left| \tau \sum_{n=1}^{N} (\nabla P_h \bar{w}^n - \nabla W_n) \right|^2.
\end{aligned}$$

Hence, from (5.2) we deduce that

$$\begin{aligned}
J_1 + J_2 &:= \tau \sum_{n=1}^{N} (\psi(w_n) - \psi(W_n), w_n - W_n) + \frac{1}{2} \int_\Omega \left| \tau \sum_{n=1}^{N} (\nabla P_h w^n - \nabla W_n) \right|^2 \\
&\le \tau \sum_{n=1}^{N} (\psi(w_n) - \psi(W_n), w_n - P_h \bar{w}^n) \\
&\quad + (\psi(w_0) - \psi(W_0), \tau \sum_{n=1}^{N} (P_h \bar{w}^n - W_n)) \\
&=: J_3 + J_4. \qquad (5.3)
\end{aligned}$$

Utilizing inequality (4.11), we find

$$J_3 = \sum_{n=1}^{N} \int_{I_n} (\psi(w_n) - \psi(W_n), w_n - P_h w)$$

$$\begin{aligned}&\le \delta \sum_{n=1}^{N} \int_{I_n} (\psi(w_n) - \psi(W_n), w_n - W_n) \\ &\quad + c_\delta \sum_{n=1}^{N} \int_{I_n} (\psi(w_n) - \psi(P_h w), w_n - P_h w) \\ &=: J_{31} + J_{32}.\end{aligned}$$

For $\delta > 0$ sufficiently small we may absorb J_{31} into the left-hand side of (5.3). Due to (4.10) we get

$$J_{32} \le c \left[\sum_{n=1}^{N} \int_{I_n} (\psi(w_n) - \psi(w), w_n - w) + \sum_{n=1}^{N} \int_{I_n} (\psi(w) - \psi(P_h w), w - P_h w) \right].$$

Further, Hölder's and Young's inequalities yield for $\delta > 0$

$$J_4 \ge c_\delta \|\psi(w_0) - \psi(W_0)\|^2_{L^n(\Omega)} + \delta \left\| \tau \sum_{n=1}^{N} (P_h \bar{w}^n - W_n) \right\|^2_{L^2(\Omega)}.$$

Due to Poincaré's inequality we may absorb $\delta \|\tau \sum (P_h \bar{w}^n - W_n)\|_2^2$. Moreover, noting that $\psi(W_0) = \Pi_h \psi(w_0)$ it holds that $\|\psi(w_0) - \psi(W_0)\|_2^2 = \|\psi(w_0) - \Pi_h \psi(w_0)\|_2^2$. Next, inequality (4.10) yields

$$(\psi(w) - \psi(w_n), w - W_n) \le c[(\psi(w) - \psi(w_n), w - w_n) + (\psi(w_n) - \psi(W_n), w_n - W_n)];$$

hence,

$$J_1 \ge c \left(\sum_{n=1}^{N} \int_{I_n} (\psi(w) - \psi(W_n), w - W_n) - \sum_{n=1}^{N} \int_{I_n} (\psi(w) - \psi(w_n), w - w_n) \right).$$

Further, we estimate

$$\begin{aligned}J_2 &= \frac{1}{2} \int_\Omega \left| \tau \sum_{n=1}^{N} (\nabla P_h \bar{w}^n - \nabla W_n) \right|^2 \\ &\ge \frac{1}{4} \int_\Omega \left| \tau \sum_{n=1}^{N} (\nabla \bar{w}^n - \nabla W_n) \right|^2 - \frac{1}{2} \int_\Omega \left| \tau \sum_{n=1}^{N} (\nabla \bar{w}^n - \nabla P_h \bar{w}^n) \right|^2\end{aligned}$$

and note that

$$\frac{1}{4} \int_\Omega \left| \tau \sum_{n=1}^{N} (\nabla \bar{w}^n - \nabla W_n) \right|^2 = \frac{1}{4} \int_\Omega |\nabla(\bar{w} - \overline{W})|^2.$$

Altogether, we deduce that

$$\begin{aligned}&\int_0^T (\psi(w) - \psi(W), w - W) + \int_\Omega |\nabla(\bar{w} - \overline{W})|^2 \\ &\le c \left(\sum_{n=1}^{N} \int_{I_n} (\psi(w_n) - \psi(w), w_n - w) + \sum_{n=1}^{N} \int_{I_n} (\psi(w) - \psi(P_h w), w - P_h w) \right. \\ &\qquad \left. + \int_\Omega \left| \tau \sum_{n=1}^{N} (\nabla \bar{w}^n - \nabla P_h \bar{w}^n) \right|^2 + \|\psi(w_0) - \Pi_h \psi(w_0)\|^2_{L^2\Omega} \right).\end{aligned}$$

Thus, the assertion follows. ∎

Let us note that the proof also implies the following result.

COROLLARY 5.2. *There is a constant c independent of h and τ such that*

$$\int_0^T \|w - W\|^2_{(w,\frac{m+1}{m})} \le c\left(\sum_{n=1}^N \int_{In} \|w_n - w\|^2_{(w,\frac{m+1}{m})} + \int_0^T \|w - P_h w\|^2_{(w,\frac{m+1}{m})} + \|\psi(w_0) - \Pi_h\psi(w_0)\|^2_{L^2(\Omega)}\right).$$

Proof. The assertion follows from the estimate

$$\int_0^T (\psi(w) - \psi(W), w - W) + \int_\Omega \left|\tau \sum_{n=1}^N (\nabla P_h \bar{w}^n - \nabla W_n)\right|^2$$
$$\le c\left(\sum_{n=1}^N \int_{I_n} (\psi(w_n) - \psi(w), w_n - w) + \sum_{n=1}^N \int_{I_n} (\psi(w) - \psi(P_h w), w - P_h w) + \|\psi(w_0) - \Pi_h\psi(w_0)\|^2_{L^2(\Omega)}\right). \quad ∎$$

6. The convergence rate. In this section we discuss the rates of convergence of the finite element approximation. For $m < 1$ we prove the following result.

THEOREM 6.1. *Let $0 < m < 1$ and $w_0 \in L^\infty(\Omega) \cap \in W^{1,2}(\Omega)$. Then there is a constant c independent of h and τ such that*

$$\left(\int_o^T \|w - W\|^2_{(w,\frac{m+1}{m})} + \sup_{1\le n\le N} \left\|\int_0^{t_n} (w - W)\right\|^2_{H^1(\Omega)}\right)^{\frac{1}{2}} \le c(\tau + h). \tag{6.1}$$

COROLLARY 6.2. *Let the assumptions of Theorem* 6.1 *be satisfied. Then there is a constant c independent of h and τ such that*

$$\|u - U\|_{L^{m+1}(0,T;L^{m+1}(\Omega))} \le c(\tau + h),$$

where $U = \psi(W)$.

Our next theorem concerns the case of $m > 1$. We treat only the case in which w satisfies either (2.6) for $\delta = 0$ or (2.7) for $\delta, \varepsilon = 0$.

THEOREM 6.3. *Let $m > 1$, $\psi(w_0) \in L^\infty(\Omega) \cap W^{1,2}(\Omega)$, and $\psi(w_0) \ge 0$.*

(a) *Let w satisfy* (2.6) *for $\delta = 0$. Then there is a constant c independent of h and τ such that*

$$\left(\int_0^T \|w - W\|^2_{(w,\frac{m+1}{m})} + \sup_{1\le n\le N} \left\|\int_0^{t_n} (w - W)\right\|^2_{H^1(\Omega)}\right)^{\frac{1}{2}} \le c(\tau + h). \tag{6.2}$$

(b) *Let w satisfy* (2.7) *for $\delta, \varepsilon = 0$ and $\psi(w_0) \in W^{\frac{m+1}{m},2}(\Omega)$. Then there is a constant c independent of h and τ such that*

$$\left(\int_0^T \|w - W\|^2_{(w,\frac{m+1}{m})}\right)^{\frac{1}{2}} \le c(\tau + h^{\frac{m+1}{m}}). \tag{6.3}$$

COROLLARY 6.4. *Let the assumption of Theorem* 6.3(a) *be satisfied. Then there is a constant* c *independent of* h *and* τ *such that*

$$\|w - W\|_{L^{\frac{m+1}{m}}(0,T;L^{\frac{m+1}{m}}(\Omega))} \le c(\tau + h).$$

Based on the results above, we have the following corollary.

COROLLARY 6.5. *Let the assumptions in either Theorem* 6.1 *or Theorem* 6.3 *hold. Then*

$$\left(\int_o^T \int_\Omega |u - U\|w - W|\right)^{\frac{1}{2}} + \left\|\int_0^t (w - W)\right\|_{L^\infty(0,T;H^1(\Omega))} \le c(\tau + h).$$

Proof of Theorem 6.1. Let us set $v := |u|^{\frac{m-1}{2}}u$. Utilizing the estimates (4.8) and (4.6), we get

$$\int_{I_n} \|w_n - w\|^2_{(w,\frac{m+1}{m})} \le c\int_{I_n} (\psi(w_n) - \psi(w), w_n - w) \le c\int_{I_n}\int_\Omega |v_n - v|^2.$$

Notice that

$$\int_{I_n} |v_n - v|^2 = \int_{I_n}\left|\int_t^{t_n} v_t\right|^2 \le \int_{I_n}\int_t^{t_n} |v_t|^2|I_n| \le \tau^2\int_{I_n}|v_t|^2$$

and $|v_t|^2 = |\partial_t(|u|^{\frac{m-1}{2}}u)|^2$. In view of (2.3), we deduce

$$\sum_{n=1}^N \int_{I_n} \|w_n - w\|^2_{(w,\frac{m+1}{m})} \le c\tau^2\left\|\partial_t u^{\frac{m+1}{2}}\right\|^2_{L^2(0,T;L^2(\Omega))} \le c\tau^2. \tag{6.4}$$

Utilizing (2.5), we estimate

$$\begin{aligned}
&\int_0^T \|w - P_h w\|^2_{(w,\frac{m+1}{m})} = \int_0^T\int_\Omega (|w| + |w - P_h w|)^{\frac{1-m}{m}}|w - P_h w|^2\\
&\le c\left(\|w\|^{\frac{1-m}{m}}_{L_\infty(0,T;L_\infty(\Omega))}\|w - P_h w\|^2_{L^2(0,T;L^2(\Omega))} + \|w - P_h w\|^{\frac{m+1}{m}}_{L^{\frac{m+1}{m}}(0,T;L^{\frac{m+1}{m}}(\Omega))}\right)\\
&\le ch^2\left(\|\nabla w\|^2_{L^2(0,T;L^2(\Omega))} + \|w\|^{\frac{m+1}{m}}_{L^{\frac{m+1}{m}}(0,T;\mathcal{N}^{\frac{2m}{m+1},\frac{m+1}{m}}(\Omega))}\right)\\
&\le ch^2.
\end{aligned} \tag{6.5}$$

Here we have used an interpolation error estimate in the Nikolskii spaces. This may be proved in the same way as in [8]. Due to (2.4) we find

$$\|\nabla(\bar w - P_h\bar w)\|^2_{L^2(\Omega)} \le ch^2\|\nabla^2\bar w\|^2_{L^2(\Omega)} \le ch^2. \tag{6.6}$$

Further, $w_0 \in L^\infty(\Omega)\cap W^{1,2}(\Omega)$ implies that $\psi(w_0) \in W^{1,2}(\Omega)$; thus,

$$\|\psi(w_0) - \Pi_h\psi(w_0)\|^2_{L^2(\Omega)} \le ch^2\|\nabla\psi(w_0)\|^2_{L^2(\Omega)} \le ch^2. \tag{6.7}$$

From Theorem 5.1 and (6.4)–(6.7) we conclude that

$$\int_0^T \|w - W\|^2_{(w,\frac{m+1}{m})} + \|\nabla(\bar{w} - \overline{W})\|^2_{L^2(\Omega)} \le c(\tau^2 + h^2), \tag{6.8}$$

where $\bar{w}(x) = \int_0^T w(x,t)dt, \overline{W}(x) = \int_0^T W(x,t)dt$, and $T = t_N$. In the same manner we can estimate (6.8) for all $n \in \{1, \dots, N\}$. Noting that the Poincaré inequality yields a constant c such that

$$\left\| \int_0^{t_n} (w - W) \right\|^2_{L^2(\Omega)} \le c \left\| \nabla \int_0^{t_n} (w - W) \right\|^2_{L^2(\Omega)}$$

the assertion follows. □

Proof of Corollary 6.2. The Hölder inequality yields (with $p_1 = \frac{2}{1-m}$ and $p_2 = \frac{2}{m+1}$)

$$\begin{aligned}
&\int_0^T \int_\Omega |u - U|^{m+1} \\
&= \int_0^T \int_\Omega (|u| + |u - U|)^{\frac{(1-m)(m+1)}{2}} (|u| + |u - U|)^{\frac{(m-1)(m+1)}{2}} |u - U|^{m+1} \\
&\le \left[\int_0^T \int_\Omega (|u| + |u - U|)^{m+1} \right]^{\frac{1-m}{2}} \left[\int_0^T \int_\Omega (|u| + |u - U|)^{m-1} |u - U|^2 \right]^{\frac{m+1}{2}}.
\end{aligned}$$

The first integral on the right-hand side in bounded. Further, due to Lemma 4.2 we have

$$\int_0^T \int_\Omega (|u| + |u - U|)^{m-1} |u - U|^2 \cong \int_0^T \|w - W\|^2_{(w,\frac{m+1}{m})}.$$

Applying Theorem 6.1, we obtain

$$\|u - U\|^{m+1}_{L^{m+1}(0,T;L^{m+1}(\Omega))} \le c(\tau + h)^{m+1}.$$

This yields the assertion. □

Proof of Theorem 6.3. (a) Let us note that $\psi(w_0) \in L^\infty(\Omega) \cap W^{1,2}(\Omega)$ implies $w_0 \in W^{1,2}(\Omega)$. Due to (2.6) we estimate

$$\begin{aligned}
\int_0^T \|w - P_h w\|^2_{(w,\frac{m+1}{m})} &\le \int_o^T \|w - P_h w\|^{\frac{m+1}{m}}_{L^{\frac{m+1}{m}}(\Omega)} \\
&\le ch^2 \|w\|^{\frac{m+1}{m}}_{L^{\frac{m+1}{m}}(0,T;\mathcal{N}^{\frac{2m}{m+1},\frac{m+1}{m}}(\Omega))} \\
&\le ch^2. \qquad (6.9)
\end{aligned}$$

Utilizing Theorem 5.1, (6.4), (6.6), (6.7), and (6.9) the assertion (6.2) follows.

(b) Using (2.7), we find

$$\begin{aligned}
\int_0^T \|w - P_h w\|^2_{(w,\frac{m+1}{m})} &\le \int_0^T \|w - P_h w\|^{\frac{m+1}{m}}_{L^{\frac{m+1}{m}}(\Omega)} \\
&\le h^{2\frac{m+1}{m}} \|\nabla^2 w\|^{\frac{m+1}{m}}_{L^{\frac{m+1}{m}}(0,T;L^{\frac{m+1}{m}}(\Omega))} \\
&\le ch^{2\frac{m+1}{m}}. \qquad (6.10)
\end{aligned}$$

Moreover,

$$\|\psi(w_0) - \Pi_h \psi(w_0)\|^2_{L^2(\Omega)} \le c h^{2\frac{m+1}{m}} \|\psi(w_0)\|^2_{W^{\frac{m+1}{m},2}(\Omega)} \le c h^{2\frac{m+1}{m}}. \tag{6.11}$$

Due to Corollary 5.2, (6.4), (6.10), and (6.11) we get the assertion (6.3). □

Proof of Corollary 6.4. Utilizing the Hölder inequality (with $p_1 = \frac{2m}{m-1}$ and $p_2 = \frac{2m}{m+1}$ and (6.2), we obtain

$$\begin{aligned}
&\int_0^T \int_\Omega |w - W|^{\frac{m+1}{m}} \\
&= \int_0^T \int_\Omega (|w| + |w - W|)^{\frac{(m+1)(m-1)}{2m^2}} (|w| + |w - W|)^{\frac{(m+1)(1-m)}{2m^2}} |w - W|^{\frac{m+1}{m}} \\
&\le \left[\int_0^T \int_\Omega (|w| + |w - W|)^{\frac{m+1}{m}} \right]^{\frac{m-1}{2m}} \left[\int_0^T \int_\Omega (|w| + |w - W|)^{\frac{1-m}{m}} |w - W|^2 \right]^{\frac{m+1}{2m}} \\
&\le c(\tau + h)^{\frac{m+1}{m}}.
\end{aligned}$$

This yields the assertion. □

Proof of Corollary 6.5. Let $T_0 \in (t_{n-1}, t_n)$. Nothing that W is constant with respect to t in $(t_{n-1}, t_n]$, we may rewrite the Nth equation of (3.2) as follows:

$$\left(\frac{\psi(W(T_0)) - \psi(W(t_{N-1}))}{\tau}, \chi_N \right) + (\nabla W(T_0), \nabla \chi_N) = 0 \qquad \forall \chi_N \in S_h.$$

Hence, replace t_N by T_0 and W_N by $W(T_0)$ in the proofs of the Theorems 5.1, 6.1, and 6.3, we obtain

$$\left\| \int_0^{T_0} (w - W) \right\|_{H^1(\Omega)} \le c(h + \tau).$$

Thus, we may conclude the assertion. □

Remark. The porous medium equation has a well-known family of self-similar solutions, the Barenblatt solutions

$$u_*(x,t) = (t+\tau)^{-k} \left\{ \left[a^2 - \frac{k(m-1)}{2dm} \frac{|x|^2}{(t+\tau)^{2k/d}} \right]^+ \right\}^{\frac{1}{m-1}},$$

Where $a, \tau > 0$ and $k = (m - 1 + 2/d)^{-1}$. Let us note that, in general, weak solutions of the porous medium equation may have singularities at points where parts of the free boundary collide with each other or at a *close in*; that is, a hole in the support fills up. However, the Barenblatt solutions are more regular, for they have a *smooth* free boundary. Due to the regularity of these solutions we expect a convergence rate of at most

$$\left(\int_0^T \|w_* - W_*\|^2_{(w, \frac{m+1}{m})} \right)^{\frac{1}{2}} \le c(\tau + h^{\frac{m}{m-1}})$$

for $m \ge 2$. Let us note that this rate is very close to that given in (6.3).

The error estimates obtained in the above theorems and corollaries are optimal in their approximation orders, and the regularities required are known to be achievable

for sufficiently smooth data—at least for $\delta, \varepsilon > 0$. In the literature, various error estimates for, say, nonnegative solutions of the porous medium equation are given, although direct comparisons with our results may not be straightforward. The results in [31] are well known; a fully discrete scheme was examined:

$$\tau^{-1}(U^{n+1} - U^n, \chi) + (\nabla W^{n+1}, \nabla \chi) = (f^{n+1/2}, \chi) \qquad \forall \chi \in S_h,$$

where $W^{n+1} \in S_h$ and $U^{n+1} = P_{0h}\psi(W^{n+1})$. In that paper it was shown that

$$||u - U||_{L^\infty(0,T;H^{-1}(\Omega))} + ||w - W||_{L^2(0,T;L^2(\Omega))} \le c(\tau + (\ln(1/h))^\alpha h)$$

for $\alpha = \frac{1}{2m}$. This estimate is optimal in the approximation order with respect to τ, although the regularity requirement for t seems too high, because Lipschitz continuity of u in $L^1(\Omega)$ is required, which is not achievable at $t = 0$.

The same fully discrete scheme was used in [12], and the following estimates were obtained by exploiting the nondegeneracy property of the free boundary [11]: There is a function $\sigma(\varepsilon, T, h)$ such that

$$||w - W||_{L^2(0,T;L^2(\Omega))} \le c\sigma.$$

Putting $\tau = h^{\frac{5m-1}{2m-2}}$ and $\varepsilon = ch^{\frac{2m^2-2}{3m^2+m}}$, we get $O(h^{\frac{m+1}{2m}})$. Similarly, we obtain

$$||u - U||_{L^2(0,T;L^2(\Omega))} \le ch^\alpha$$

for $\alpha = \frac{m^2+4m+3}{6m^2+2m} \ge \frac{1}{6}$.

In [22], the following fully discrete scheme was used:

$$\tau^{-1}((U^{n+1}) - (U^n), \chi) + (\nabla(|U^{n+1}|^{m-1}U^{n+1}), \nabla \chi) = 0 \qquad \forall \chi \in S_h,$$

where $U^{n+1} \in S_h$. Let $U(t)$ be piecewise linear in t; then

$$||u - U||_{L^{m+1}(0,T;L^{m+1}(\Omega))} \le c(\tau^{\frac{1}{2}} + h + h^{\frac{1}{m}(d\frac{m-1}{2m+2}+1)})^{\frac{1}{m+1}}.$$

In [28] the following fully discrete scheme using numerical integration was examined:

$$\tau^{-1}(\Pi_h\psi_\varepsilon(W^{n+1}) - \Pi_h\psi_\varepsilon(W^n), \chi)_h + (\nabla W^{n+1}, \nabla \chi) = 0 \qquad \forall \chi \in S_h,$$

where $W^{n+1} \in S_h, \Pi_h$ is the interpolant, and ψ_ε is regularized such that it is linear near 0. The following estimates were obtained:

$$||u - U||_{L^\infty(0,T;H^{-1}(\Omega))} + ||w - W||_{L^2(0,T;L^2(\Omega))} + \left\| \int_0^t (w - W) \right\|_{L^\infty(0,T;H^1(\Omega))} \le c\sigma,$$

where $\sigma = \sigma(\varepsilon, \tau, h) = (\frac{h^2}{\varepsilon} + \frac{h^4}{\varepsilon^2\tau} + \tau)^{1/2}$. Putting $\tau = ch^{\frac{m+1}{m}}$ and $\varepsilon = ch^{\frac{m-1}{m}}$, one gets the convergence rate $O(h^{\frac{m+1}{2m}})$.

In [30], a regularized version of our approximation scheme was used. The following suboptimal result was shown, choosing the regularization parameter in an appropriate way:

$$\left(\tau \sum_n ||u(t_n) - U^n||^{m+1}_{L^{m+1}(\Omega)}\right)^{\frac{1}{m+1}} \le c(\tau^{\frac{1}{m}} + (\ln(1/h))^{\frac{1}{m(m+1)}} h^{\frac{2}{m}})$$

for $d = 2, 3$, under the assumption that $u_t \in L^{\frac{m+1}{m}}(\Omega \times (0, T))$.

Furthermore, $||w-W||_{L^2(0,T;L^2(\Omega))} + ||\int_0^t (w-W)||_{L^\infty(0,T;H^1(\Omega))} \le c\tau^{\frac{1}{4}}$ was obtained in [20], where a linearized semidiscrete scheme was considered. Moreover, an a priori error estimate for $||(u-U)(w-W)||_{L^1(0,T;L^1(\Omega))}$ was proved in [1]. Due to the low regularity of u this estimate implies a convergence rate which tends to 0 for $m \to \infty$. In [17] it was shown that

$$||u-U||_{L^\infty(0,T;H^{-1}(\Omega)} + \left(\int_0^T \int_\Omega |u-U||w-W|\right)^{\frac{1}{2}} + \left|\left|\int_0^T (w-W)\right|\right|^2_{H^1(\Omega)} \le c\tau^{\frac{1}{2}},$$

where a semidiscrete scheme was studied. In [29], for a regularized semidiscrete scheme,

$$||w-W||_{L^2(0,T;L^2(\Omega))} \le c\tau^{\frac{1}{2}}$$

was obtained.

For the case in which $m < 1$, there are only a few results. For the fully discrete schemes in [22] and [31], it was shown that

$$||u-U||_{L^2(0,T;L^{m+1}(\Omega))} \le c(\tau^{1/2} + h + h^{d\frac{m-1}{2m+2}+1})^{1/2} \qquad (\tfrac{2d}{d+1} < m < 1)$$

and

$$||u-U||_{L^\infty(0,T;H^{-1}(\Omega))} \le c(\tau + (\ln(1/h))^\alpha h)$$

for some $\alpha > 0$, respectively.

REFERENCES

[1] T. Arbogast, M. F. Wheeler, and N.-Y. Zhang, *A nonlinear mixed finite element method for a degenerate parabolic equation arising in flow in porous media*, SIAM J. Numer. Anal., 33 (1996), pp. 1669–1687.

[2] J. W. Barrett and W. B. Liu, *Finite element approximation of the p-Laplacian,* Math. Comp. 61 (1993), pp. 523–537.

[3] J. W. Barrett and W. B. Liu, *Finite element approximation of the parabolic p-Laplacian*, SIAM J. Numer. Anal., 31 (1994), pp. 413–428.

[4] P. Bénilan, *A strong regularity L^p for solution of the porous media equation*, in Contributions to Nonlinear Partial Differential Equations (Madrid, 1981), Res. Notes Math. 89, Pitman, Boston, 1983, pp. 39–58.

[5] J. G. Berryman and C. J. Holland, *Stability of the separable solution for fast diffusion,* Arch. Ration. Mech. Anal., 74 (1980), pp. 379–388.

[6] L. A. Caffarelli and A. Friedman, *Regularity of the free boundary of a gas flow in an n-dimensional porous medium,* Indiana Univ. Math. J., 29 (1980), pp. 361–391.

[7] S. S. Chow, *Finite element error estimates for non-linear elliptic equations of monotone type*, Numer. Math., 54 (1988), pp. 373–393.

[8] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North–Holland, Amsterdam, 1978.

[9] E. DiBenedetto, Y. Kwong, and V. Vespri, *Local space-analyticity of solutions of certain singular parabolic equations,* Indiana Univ. Math. J., 40 (1991), pp. 741–765.

[10] L. Diening, C. Ebmeyer, and M. Růžička, *Optimal convergence for the implicit space-time discretization of parabolic systems with p-structure,* SIAM J. Numer. Anal., 45 (2007), pp. 457–472.

[11] C. Ebmeyer, *A non-degeneracy property for a class of degenerate parabolic equations,* Z. Anal. Anwendungen, 15 (1996), pp. 637–650.

[12] C. Ebmeyer, *Error estimates for a class of degenerate parabolic equations,* SIAM J. Numer. Anal., 35 (1998), pp. 1095–1112.

[13] C. Ebmeyer, *Regularity in Sobolev spaces for the fast diffusion and the porous medium equation,* J. Math. Anal. Appl., 307 (2005), pp. 134–152.

[14] C. Ebmeyer and W. B. Liu, *Quasi-norm interpolation error estimates for finite element approximations of problems with p-structure*, Numer. Math., 100 (2005), pp. 233–258.
[15] C. Ebmeyer and J. M. Urbano, *Regularity in Sobolev spaces for doubly nonlinear parabolic equations,* J. Differential Equations, 187 (2003), pp. 357–390.
[16] C. Ebmeyer and J. M. Urbano, *The smoothing property for a class of doubly nonlinear parabolic equations,* Trans. Amer. Math. Soc., 357 (2005), pp.3239–3253.
[17] A. Eden, B. Michaux, and J.-M. Rakotoson, *Semi-discretized nonlinear evolution equations as discrete dynamical systems and error analysis,* Indiana Univ. Math. J., 39 (1990), pp. 737–783.
[18] j. Filo, *A nonlinear diffusion equation with nonlinear boundary conditions: Method of lines*, Math. Slovaca, 38 (1988), pp. 273–296.
[19] E. Giusti, *Direct Methods in the Calculus of Variations,* World Scientific, Singapore, 2003.
[20] A. Handlovicová, *Error estimates of a linear approximation scheme for nonlinear diffusion problems,* Acta Math. Univ. Comenian, 61 (1992), pp. 27–39.
[21] W. Jäger and J. Kacur, *Solution of porous medium type systems by linear approximation schemes*, Numer. Math., 60 (1991), pp. 407–427.
[22] L. Lefton and D. Wei, *A priori L^α error estimates for Galerkin approximations to porous medium and fast diffusion equations*, Math. Comput., 68 (1999), pp. 971–989.
[23] M.-N. Le Roux, *Semi-discretization in time of a fast diffusion equation*, J. Math. Anal. Appl., 137 (1989), pp. 354–370.
[24] M.-N. Le Roux and P.-E. Mainge, *Numerical solution of a fast diffusion equation*, Math. Comp., 68 (1999), pp. 461–485.
[25] W. B. Liu and J. W. Barrett, *A further remark on the regularity of the solutions of the p-Laplacian and its applications to their finite element approximation*, J. Nonlinear Anal., 21 (1993), pp. 379–387.
[26] W. B. Liu and J. W. Barrett, *Finite element approximation of some degenerate monotone quasi-linear elliptic systems*, SIAM J. Numer. Anal., 33 (1996), pp. 88–106.
[27] R. H. Nochetto, *Finite element methods for parabolic free boundary problems*, in Advances in Numerical Analysis, Volume I (Lancaster, 1990), Oxford Sci. Publ., Oxford University Press, New York, 1991, pp. 34–95.
[28] R. H. Nochetto and C. Verdi, *Approximation of degenerate parabolic problems using numerical integration*, SIAM J. Numer. Anal., 25 (1988), pp. 784–814.
[29] I. S. Pop and W.-A. Yong, *A numerical approach to degenerate parabolic equations*, Numer. Math., 92 (2002), pp. 357–381.
[30] M. E. Rose, *Numerical methods for flows through porous media I*, Math. Comp., 40 (1983), pp. 435–467.
[31] J. Rulla and N. J. Walkington, *Optimal rates of convergence for degenerate parabolic problems in two dimensions*, SIAM J. Numer. Anal., 33 (1996), pp. 56–67.
[32] R. E. Showalter, *Monotone Operators in Banach Space and Nonlinear Partial Differential Equations*, Math. Surveys Monogr. 49, AMS, Providence, RI, 1997.
[33] J. L. Vazquez, *An introduction to the mathematical theory of the porous medium equation*, in Shaper Optimization and Free Boundaries, M. C. Delfour and G. Sabidussi, eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 1992, pp. 347–389.

SIAM J. NUMER. ANAL.
Vol. 46, No. 5, pp. 2411–2442
© 2008 Society for Industrial and Applied Mathematics

AN ANISOTROPIC SPARSE GRID STOCHASTIC COLLOCATION METHOD FOR PARTIAL DIFFERENTIAL EQUATIONS WITH RANDOM INPUT DATA*

F. NOBILE†, R. TEMPONE‡, AND C. G. WEBSTER§

Abstract. This work proposes and analyzes an anisotropic sparse grid stochastic collocation method for solving partial differential equations with random coefficients and forcing terms (input data of the model). The method consists of a Galerkin approximation in the space variables and a collocation, in probability space, on sparse tensor product grids utilizing either Clenshaw–Curtis or Gaussian knots. Even in the presence of nonlinearities, the collocation approach leads to the solution of uncoupled deterministic problems, just as in the Monte Carlo method. This work includes a priori and a posteriori procedures to adapt the anisotropy of the sparse grids to each given problem. These procedures seem to be very effective for the problems under study. The proposed method combines the advantages of isotropic sparse collocation with those of anisotropic full tensor product collocation: the first approach is effective for problems depending on random variables which weigh approximately equally in the solution, while the benefits of the latter approach become apparent when solving highly anisotropic problems depending on a relatively small number of random variables, as in the case where input random variables are Karhunen–Loève truncations of "smooth" random fields. This work also provides a rigorous convergence analysis of the fully discrete problem and demonstrates (sub)exponential convergence in the asymptotic regime and algebraic convergence in the preasymptotic regime, with respect to the total number of collocation points. It also shows that the anisotropic approximation breaks the *curse of dimensionality* for a wide set of problems. Numerical examples illustrate the theoretical results and are used to compare this approach with several others, including the standard Monte Carlo. In particular, for moderately large-dimensional problems, the sparse grid approach with a properly chosen anisotropy seems to be very efficient and superior to all examined methods.

Key words. collocation techniques, PDEs with random data, differential equations, finite elements, uncertainty quantification, anisotropic sparse grids, Smolyak sparse approximation, multivariate polynomial approximation

AMS subject classifications. 65N30, 65N35, 65N12, 65N15, 65C20

DOI. 10.1137/070680540

Introduction.

Stochastic formulations provide a natural way to include uncertainty quantification in practical applications. Uncertainty usually appears through

*Received by the editors January 22, 2007; accepted for publication (in revised form) February 13, 2008; published electronically June 6, 2008.
http://www.siam.org/journals/sinum/46-5/68054.html

†MOX, Dipartimento di Matematica "F. Brioschi," Politecnico di Milano, 20133 Milano, Italy (fabio.nobile@ polimi.it). This author's research was partially supported by M.U.R.S.T. PRIN 2005 "Numerical Modeling for Scientific Computing and Advanced Applications" and SANDIA project 523695.

‡Department of Mathematics and School of Computational Science, Florida State University, 400 Dirac Science Library, Tallahassee, FL 32306-4120, and School of Computer Sciences and Communication, KTH, S–100 44 Stockholm, Sweden (rtempone@scs.fsu.edu). This author's research was partially supported by startup funds from the School of Computation Science (SCS) at Florida State University, the Dahlquist fellowship at the Royal Institute of Technology, Stockholm, UdelaR in Uruguay, and SANDIA project 523695.

§Optimization and Uncertainty Quantification, Sandia National Laboratories, P.O. Box 5800, MS 1318, Albuquerque, NM 87185-1318 (cgwebst@sandia.gov). This author's research was supported by the School of Computational Science (SCS) at Florida State University and by the John von Neumann Fellowship at the Computer Science Research Institute (CSRI), Sandia National Laboratories. Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE AC04-94-AL85000.

the input data of the problem, such as model coefficients, forcing terms, boundary conditions, geometry, etc. Once the probability distribution of the input random data is known, the computational goal is the prediction of statistical moments of the solution (mean value, variance, covariance, etc.) or statistics of some given responses of the system (sometimes called quantities of physical interest which are real valued functionals of the solution). Examples of quantities of interest could include solution values in a given region, fluxes across given boundaries, etc.

In this work we focus on the solution of partial differential equations (PDEs) with random coefficients and forcing terms (input data of the model), and we especially address the situation where the input data are assumed to depend on a moderately large number of random variables.

The Monte Carlo method described in [11] is the classical and most widely applied technique for approximating expected values of quantities of interest depending on the solution of a PDE with random inputs. The algorithm approximates the desired expectation by a sample average of independent identically distributed (i.i.d.) realizations. When solving PDEs with random inputs this method employs the solution of one deterministic differential equation for each realization of the input parameters. This makes the method simple to implement, allows for maximal code reusability, and is straightforward to parallelize. The asymptotic numerical error of the Monte Carlo method is approximately $\mathcal{O}(1/\sqrt{M})$, where M is the number of realizations. The advantage of utilizing this approach is that the rate does not deteriorate with respect to the number of random variables in the problem, making the method very attractive for problems with large-dimensional random inputs. On the other hand, when solving large-scale applications, the exponent $1/2$ in the rate of convergence generates a huge amount of computational work required to achieve accurate solutions. Other ensemble-based methods like quasi-Monte Carlo, latin hypercube sampling, etc. (see, e.g., [23, 18] and the references therein), have been devised to produce faster convergence, $\mathcal{O}(\log(M)^r/M)$, where the coefficient $r > 0$ becomes larger with the dimension of the random input.

Another possible substitute for the Monte Carlo method is the so-called spectral Galerkin method. For more details see, e.g., [17]. It employs standard approximations in space (finite elements, finite volumes, spectral or h-p finite elements, etc.) and polynomial approximation in the probability domain, either by full polynomial spaces [30, 22, 16], tensor product polynomial spaces [2, 12, 25], or piecewise polynomial spaces [8, 2, 19, 28, 21]. This family of methods exploits the regularity of the solution to acquire faster convergence rates. However, in general, this technique requires solving a system of equations that couples all degrees of freedom in the approximation to the stochastic solution.

A new numerical technique, which has gained much attention recently by the computational community, is stochastic collocation (see [1, 13, 24, 20, 29]), which can be based on either full or sparse tensor product approximation spaces. As shown in [1], stochastic collocation essentially preserves the fast convergence of the spectral Galerkin method, even coinciding in particular cases, while maintaining an ensemble-based approach, just as Monte Carlo. Hence, stochastic collocation seems to be an ideal method for computing statistics from solutions of PDEs with random input data. The practical challenge to the use of stochastic collocation becomes evident when solving problems depending on a relatively large number of input random variables: any simple-minded approach requires unaffordable amounts of computational effort. This is a symptom of the *curse of dimensionality*, and it is our goal, therefore, to investi-

gate variants of this procedure to attain highly accurate solutions while significantly reducing the computational cost.

In particular, isotropic sparse approximations based on the Smolyak construction have been proposed and analyzed in [4] for the approximation of real valued functions and were first applied to PDEs with random coefficients in [29]. Our previous paper [24] provided a full convergence analysis for this approximation scheme under the assumption that the solution depends analytically on the input parameters, which is valid in many elliptic and parabolic problems.

Furthermore, the work [24] revealed that the isotropic sparse collocation algorithm is effective for problems whose input data depend on a moderate number of random variables, which "weigh equally" in the solution. For such isotropic situations the displayed convergence is faster than standard collocation techniques built upon full tensor product spaces. On the other hand, the convergence rate of the isotropic sparse collocation algorithm deteriorates for highly anisotropic problems, such as those appearing when the input random variables come, for example, from Karhunen–Loève-type truncations of "smooth" random fields. In such cases, a full anisotropic tensor product approximation [2, 1] may still be more effective for a small or modest number of random variables. However, if the number of random variables is large, the construction of the full tensor product spaces becomes infeasible, since the dimension of the approximating space grows exponentially fast with respect to the number of random variables in the problem.

The main contribution of this work is to propose and analyze a novel *anisotropic* sparse grid stochastic collocation method that is based on a weighted version of the Smolyak algorithm. In particular, we develop an original procedure, based on either a priori or a posteriori estimates for choosing the anisotropy of the sparse grid, and we show both theoretically and numerically the effectiveness of these methods for several problems. This approach is particularly attractive in the case of truncated expansions of random fields, since the anisotropy can be tuned to the decay properties of the expansion.

In this respect, the proposed method differs from the dimension-adaptive method developed in [15] and applied in the context of PDEs in [13], where the sparse approximation structure has to be uncovered only by successive enrichment, and does not exploit any a priori information of the solution.

A related approach is proposed in [26], where sparse wavelet chaos subspaces are constructed for the approximation of the solution of linear scalar elliptic PDEs with random coefficients. Those approximations are also anisotropic and use a priori regularity from the exact solution. The main difference with our approach is that the sparse approximation in [26] is the same for all problems, i.e., does not depend on the problem data, and it is used mainly to prove convergence rates.

This work provides a rigorous convergence analysis of our anisotropic sparse stochastic collocation method in the case where the solution depends analytically on the input parameters, and demonstrates (sub)exponential convergence in the asymptotic regime and algebraic convergence in the preasymptotic regime with respect to the total number of collocation points. Numerical examples illustrate the theoretical results and are used to compare this approach with the standard Monte Carlo method as well as the anisotropic full tensor product approach. In particular, for moderately large-dimensional problems, our approach with properly chosen anisotropy seems to be very efficient and superior to all examined methods. We also show that this method is able to break the curse of dimensionality in certain cases where the input random

variables come from a truncated Karhunen–Loève-type expansion of a smooth random field.

The outline of the paper is the following: In section 1 we introduce the mathematical problem and the main notation used throughout. Here, we also state assumptions on the parametrization of the random inputs, which is useful when later transforming the original stochastic problem into a deterministic parametric one, and on the problem solution's regularity, which is used to later prove the error estimates. In section 2 we introduce the anisotropic sparse collocation method as well as two interpolation techniques based on Clenshaw–Curtis or Gaussian abscissas. We also provide a priori and a posteriori procedures for tuning the anisotropy of our sparse grid method to the problem at hand. In section 3 we provide a detailed error analysis. Finally, in section 4 we present numerical results, including a comparison with the Monte Carlo method and the anisotropic full tensor product approach, showing the efficiency of the proposed method.

1. Problem setting. We begin by focusing our attention on an elliptic operator $\mathcal{L}$, linear or nonlinear, defined on a domain $D \subset \mathbb{R}^d$, which depends on some coefficients $a(\omega, x)$ with $x \in D$, $\omega \in \Omega$ and $(\Omega, \mathcal{F}, P)$ a complete probability space. Here Ω is the set of outcomes, $\mathcal{F} \subset 2^\Omega$ is the σ-algebra of events, and $P : \mathcal{F} \to [0,1]$ is a probability measure. Analogously, the forcing term $f = f(\omega, x)$ can be assumed to be random as well.

Consider the stochastic elliptic boundary value problem: Find a random function, $u : \Omega \times \overline{D} \to \mathbb{R}$, such that P-a.e. in Ω, or in other words almost surely (a.s.), the following equation holds:

$$\mathcal{L}(a)(u) = f \quad \text{in } D, \tag{1.1}$$

equipped with suitable boundary conditions. Before introducing some assumptions we denote by $W(D)$ a Banach space of functions $v : D \to \mathbb{R}$ and define, for $q \in [1, \infty]$, the stochastic Banach spaces

$$L_P^q(\Omega; W(D)) := \Bigg\{ v : \Omega \to W(D) \mid v \text{ is strongly measurable and } \int_\Omega \|v(\omega, \cdot)\|_{W(D)}^q dP(\omega) < +\infty \Bigg\}$$

and

$$L_P^\infty(\Omega; W(D)) := \Bigg\{ v : \Omega \to W(D) \mid v \text{ is strongly measurable and } P - \operatorname*{ess\,sup}_{\omega \in \Omega} \|v(\omega, \cdot)\|_{W(D)}^2 < +\infty \Bigg\}.$$

Of particular interest is the space $L_P^2(\Omega; W(D))$, consisting of Banach valued functions that have finite second moments.

We will now make the following assumptions:

(A_1) The solution to (1.1) has realizations in the Banach space $W(D)$, i.e., $u(\cdot, \omega) \in W(D)$ a.s. and $\forall \omega \in \Omega$

$$\|u(\cdot, \omega)\|_{W(D)} \le C \|f(\cdot, \omega)\|_{W^*(D)},$$

where we denote $W^*(D)$ to be the dual space of $W(D)$, and C is a constant independent of the realization $\omega \in \Omega$.

(A_2) The forcing term $f \in L^2_P(\Omega; W^*(D))$ is such that the solution u is unique and bounded in $L^2_P(\Omega; W(D))$.

Here we give two example problems that are posed in this setting.

Example 1.1. The linear problem

$$\begin{cases} -\nabla \cdot (a(\omega,\cdot)\nabla u(\omega,\cdot)) = f(\omega,\cdot) \text{ in } \Omega \times D, \\ u(\omega,\cdot) = 0 \text{ on } \Omega \times \partial D, \end{cases} \tag{1.2}$$

with $a(\omega,\cdot)$ uniformly bounded and coercive, i.e.,

$$\text{there exists } a_{min}, a_{max} \in (0,+\infty)$$
$$\text{such that } P(\omega \in \Omega : \ a(\omega,x) \in [a_{min}, a_{max}] \, \forall x \in \overline{D}) = 1$$

and $f(\omega,\cdot)$ square integrable with respect to P, satisfies assumptions A_1 and A_2 with $W(D) = H^1_0(D)$ (see [1]).

Example 1.2. Similarly, for $k \in \mathbb{N}^+$, the nonlinear problem

$$\begin{cases} -\nabla \cdot (a(\omega,\cdot)\nabla u(\omega,\cdot)) + u(\omega,\cdot)|u(\omega,\cdot)|^k = f(\omega,\cdot) \quad \text{in } \Omega \times D, \\ u(\omega,\cdot) = 0 \quad \text{on } \Omega \times \partial D, \end{cases} \tag{1.3}$$

with $a(\omega,\cdot)$ uniformly bounded and coercive and $f(\omega,\cdot)$ square integrable with respect to the probability measure, satisfies assumptions A_1 and A_2 with $W(D) = H^1_0(D) \cap L^{k+2}(D)$.

1.1. On finite-dimensional noise. In some applications, the coefficient a and the forcing term f appearing in (1.1) can be described by a random vector $[Y_1, \ldots, Y_N] : \Omega \to \mathbb{R}^N$. In these cases, we will emphasize such dependence by writing a_N and f_N.

An example is given by the case where a (and/or f) is a random field and a_N (and/or f_N) is obtained as a truncated Karhunen–Loève expansion (see [24] for details).

In what follows, we will focus on problems whose coefficients and forcing terms are described by a finite number of random variables (*finite-dimensional noise assumption*) and seek a random field $u_N : \Omega \times \overline{D} \to \mathbb{R}$, such that a.s. the following equation holds:

$$\mathcal{L}(a_N)(u_N) = f_N \quad \text{in } D. \tag{1.4}$$

Further, we assume that (1.4) admits a unique solution $u_N \in L^2_P(\Omega; W(D))$. We then have that the solution u_N of the stochastic elliptic boundary value problem (1.4) can be described by $u_N = u_N(\omega, x) = u_N(Y_1(\omega), \ldots, Y_N(\omega), x)$. We underline that the coefficients a_N and f_N in (1.4) may be either an exact or an approximate representation of the input data. In the latter case, the solution u_N will also be an approximation of the exact solution u in (1.1), and the truncation error $u - u_N$ has to be properly estimated; see, for instance, [24, section 4.2].

Remark 1.3. It is usual to have f_N and a_N independent, because the forcing terms and the parameters in the operator $\mathcal{L}$ are seldom related. In such a situation we have $a_N(Y(\omega), x) = a_N(Y_a(\omega), x)$ and $f_N(Y(\omega), x) = f_N(Y_f(\omega), x)$, with $Y = [Y_a, Y_f]$ and the vectors Y_a, Y_f are independent.

In this work we will adopt the following notation: we denote $\Gamma_n \equiv Y_n(\Omega)$ the image of Y_n, where we assume $Y_n(\omega)$ to be bounded. Without loss of generality we

can assume $\Gamma_n = [-1, 1]$. We also let $\Gamma^N = \prod_{n=1}^N \Gamma_n$ and assume that the random variables $[Y_1, Y_2, \ldots, Y_n]$ have a joint probability density function $\rho : \Gamma^N \to \mathbb{R}_+$, with $\rho \in L^\infty(\Gamma^N)$. Thus, the goal is to approximate the function $u_N = u_N(y, x)$ for any $y \in \Gamma^N$ and $x \in \overline{D}$ (see [1, 2]).

1.2. Regularity. Before discussing various collocation techniques and going through the convergence analysis of such methods, we need to state some regularity assumptions on the data of the problem and consequent regularity results for the exact solution u_N. We will perform a one-dimensional analysis in each direction y_n, $n = 1, \ldots, N$. For this, we denote $\Gamma_n^* = \prod_{\substack{j=1 \\ j \neq n}}^N \Gamma_j$, and let $\mathbf{y}_n^*$ be an arbitrary element of Γ_n^*. We require the solution to problem (1.1) to satisfy the following estimate.

Assumption 1.4. For each $y_n \in \Gamma_n$, there exists $\tau_n > 0$ such that the function $u_N(y_n, \mathbf{y}_n^*, x)$ as a function of y_n, $u_N : \Gamma_n \to C^0(\Gamma_n^*; W(D))$ admits an analytic extension $u(z, \mathbf{y}_n^*, x)$, $z \in \mathbb{C}$, in the region of the complex plane

$$\Sigma(\Gamma_n; \tau_n) \equiv \{z \in \mathbb{C}, \ \ \mathrm{dist}(z, \Gamma_n) \leq \tau_n\}. \tag{1.5}$$

Moreover, $\forall z \in \Sigma(\Gamma_n; \tau_n)$,

$$\|u_N(z)\|_{C^0(\Gamma_n^*; W(D))} \leq \lambda, \tag{1.6}$$

with λ a constant independent of n.

The previous assumption should be verified for each particular application. In particular, this has implications on the allowed regularity of the input data, e.g., coefficients, loads, etc., with respect to $y_n, n = 1, 2, \ldots, N$, of the PDE under study.

Please observe that this assumption does not preclude the possibility of having singularities in the solution with respect to the spatial variables.

Remark 1.5 (unbounded random variables). Following the approach in [1] we can also state and verify related regularity assumptions for the case where the input data depend on unbounded random variables, such as Gaussian or exponential ones. These regularity assumptions can be used to produce convergence estimates of anisotropic Smolyak approximations for such cases. For the sake of simplicity in this work we focus our study on bounded random variables only.

We recall now some theoretical results, which were proved in [1, section 3], for the linear problem introduced in Example 1.1. Once we have the input random fields described by a finite set of random variables, i.e., $a(\omega, x) = a_N(Y_1(\omega), \ldots, Y_N(\omega), x)$, and similarly for $f(\omega, x)$, the "finite-dimensional" version of the stochastic formulation (1.2) has a "deterministic" equivalent, which is the following: Find $u_N \in L^2_\rho(\Gamma^N; H^1_0(D))$ such that

$$\int_{\Gamma^N} \rho \, (a_N \nabla u_N, \nabla v)_{L^2(D)} \, dy = \int_{\Gamma^N} \rho \, (f_N, v)_{L^2(D)} \, dy \quad \forall \, v \in L^2_\rho(\Gamma^N; H^1_0(D)). \tag{1.7}$$

Observe that in this work the gradient notation, ∇, always means differentiation with respect to $x \in D$ only, unless otherwise stated. The stochastic boundary value problem (1.2) now becomes a deterministic Dirichlet boundary value problem for an elliptic PDE with an N-dimensional parameter. For convenience, we consider the solution u_N as a Banach valued function $u_N : \Gamma^N \to H^1_0(D)$ and use the notation $u_N(y)$ whenever we want to highlight the dependence on the parameter y. We use similar notation for the coefficient a_N and the forcing term f_N. Then it can be shown

that problem (1.7) is equivalent to

$$\int_D a_N(y)\nabla u_N(y)\cdot\nabla\phi\,dx = \int_D f_N(y)\phi\,dx \quad \forall\phi\in H_0^1(D),\ \ \rho\text{-a.e. in } \Gamma^N. \tag{1.8}$$

For our convenience, we will suppose that the coefficient a_N and the forcing term f_N admit a smooth extension on the ρ-zero measure sets. Then (1.8) can be extended a.e. in Γ^N with respect to the Lebesgue measure (instead of the measure ρdy).

It has been proved in [1] that problem (1.8) satisfies the analyticity result stated in Assumption 1.4. For instance, if we take the diffusivity coefficient as the truncated nonlinear expansion

$$a(\omega,x)\approx a_N(\omega,x)=a_{min}+\exp\left\{b_0(x)+\sum_{n=1}^{N}\sqrt{\lambda_n}b_n(x)Y_n(\omega)\right\}, \tag{1.9}$$

where (λ_n,b_n) are eigenpairs of the covariance operator associated to the random field $\log(a-a_{min})$ (see [24] for details), then a suitable analyticity region $\Sigma(\Gamma_n;\tau_n)$ is given by

$$\tau_n=\frac{1}{4\sqrt{\lambda_n}\|b_n\|_{L^\infty(D)}}. \tag{1.10}$$

Observe that, as $\sqrt{\lambda_n}\|b_n\|_{L^\infty(D)}\to 0$ for a regular enough covariance function (see [12]), the analyticity region increases as n increases. This fact naturally introduces an anisotropic behavior with respect to the "direction" n. This effect will be exploited in the numerical method proposed in the next sections.

2. Anisotropic sparse grid stochastic collocation. Stochastic collocation entails the sampling of approximate values to the solution u_N of (1.4) on a suitable set of abscissas $y_k\in\Gamma^N$. For that, we introduce some standard approximation spaces, namely:

- $W_h(D)\subset W(D)$ is a standard finite element space of dimension N_h, which contains continuous piecewise polynomials defined on regular triangulations $\mathcal{T}_h$ that have a maximum mesh-spacing parameter $h>0$. We suppose that W_h has the following deterministic approximability property: for a given function $\varphi\in W(D)$,

$$\min_{v\in W_h(D)}\|\varphi-v\|_{W(D)}\le C(s;\varphi)\,h^s, \tag{2.1}$$

where s is a positive integer determined by the smoothness of φ and the degree of the approximating finite element subspace and $C(s;\varphi)$ is independent of h.

Example 2.1. Let D be a convex polygonal domain and $W(D)=H_0^1(D)$. For piecewise linear finite element subspaces we have

$$\min_{v\in W_h(D)}\|\varphi-v\|_{H_0^1(D)}\le c\,h\,\|\varphi\|_{H^2(D)}.$$

That is, $s=1$ and $C(s;\varphi)=\|\varphi\|_{H^2(D)}$; see, for example, [5].

We will also assume that there exists a finite element operator $\pi_h: W(D)\to W_h(D)$ with the optimality condition

$$\|\varphi-\pi_h\varphi\|_{W(D)}\le C_\pi\min_{v\in W_h(D)}\|\varphi-v\|_{W(D)}\quad\forall\varphi\in W(D), \tag{2.2}$$

where the constant C_π is independent of the mesh size h.

- We also need to introduce global polynomial spaces

$$\mathcal{P}_{p_n}(\Gamma_n) = \text{span}(y_n^k,\, k = 0, \ldots, p_n), \quad n = 1, \ldots, N,$$

  which will be used as building blocks for the sparse approximation of the solution u_N with respect to the random variables.

Example 2.2. If we examine the linear problem (1.2), we can introduce the semidiscrete approximation $u_h^N : \Gamma^N \to W_h(D)$, obtained by projecting (1.8) onto the subspace $W_h(D)$ for each $y \in \Gamma^N$, i.e.,

$$\int_D a_N(y)\nabla u_h^N(y) \cdot \nabla \phi_h \, dx = \int_D f_N(y)\phi_h \, dx \quad \forall \phi_h \in W_h(D) \text{ for a.e. } y \in \Gamma^N. \tag{2.3}$$

Notice that the finite element functions $u_h^N(y)$ satisfy the optimality condition (2.2) $\forall y \in \Gamma^N$, with a constant C_π that might depend on y, i.e., $C_\pi = C_\pi(y)$.

The construction of a fully discrete approximation, $u_{h,\mathrm{w}}^N \in C^0(\Gamma^N; W_h(D))$, is based on a suitable interpolation of the sampled values. That is,

$$u_{h,\mathrm{w}}^N(y, \cdot) = \sum_k u_h^N(y_k, \cdot) l_k^{\mathrm{w}}(y), \tag{2.4}$$

where, for instance, the functions l_k^{w} can be taken as the Lagrange polynomials (see [24] and section 2.1). Here w is a suitable discretization parameter. This formulation can be used to approximate the mean value or variance of u, as

$$E[u] \approx \overline{u}_h^N \equiv \sum_k u_h^N(y_k, \cdot) \int_{\Gamma^N} l_k^{\mathrm{w}}(y)\rho(y)dy$$

and

$$\mathrm{Var}[u] \approx \sum_k \left(u_h^N(y_k, \cdot)\right)^2 \int_{\Gamma^N} l_k^{\mathrm{w}}(y)\rho(y)dy - \left(\overline{u}_h^N\right)^2.$$

Several choices are possible for the interpolation abscissas. We will discuss two of them, namely, Clenshaw–Curtis and Gaussian ones, in the next subsection. Regardless of the choice of interpolating abscissas, the approximation can be constructed by using either full tensor product polynomials [1] or the space of sparse polynomials [24].

2.1. Anisotropic Smolyak formula. To define the anisotropic Smolyak formula, let us briefly recall interpolation based on Lagrange polynomials. Let $i \in \mathbb{N}_+$ and $\{y_1^i, \ldots, y_{m_i}^i\} \subset [-1, 1]$ be a sequence of abscissas for Lagrange interpolation on $[-1, 1]$. For a univariate function $u \in C^0(\Gamma^1; W(D))$, with $\Gamma^1 = [-1, 1]$, we introduce a sequence of one-dimensional Lagrange interpolation operators $\mathscr{U}^i : C^0(\Gamma^1; W(D)) \to V_{m_i}(\Gamma^1; W(D))$,

$$\mathscr{U}^i(u)(y) = \sum_{j=1}^{m_i} u(y_j^i) \cdot l_j^i(y) \quad \forall u \in C^0(\Gamma^1; W(D)), \tag{2.5}$$

where $l_j^i \in \mathcal{P}_{m_i - 1}(\Gamma^1)$ are Lagrange polynomials of degree $p_i = m_i - 1$ and

$$V_m(\Gamma^1; W(D)) = \left\{ v \in C^0(\Gamma^1; W(D)) \,:\, v(y, x) = \sum_{k=1}^{m} \widetilde{v}_k(x) l_k(y),\, \{\widetilde{v}_k\}_{k=1}^m \in W(D) \right\}.$$

Here of course we have, for $i \in \mathbb{N}_+$, $l_j^i(y) = \prod_{\substack{k=1 \\ k \neq j}}^{m_i} \frac{(y - y_k^i)}{(y_j^i - y_k^i)}$. Using the convention $\mathscr{U}^0 = 0$, define

$$\Delta^i := \mathscr{U}^i - \mathscr{U}^{i-1} \quad \text{for } i = 1, 2, \ldots . \tag{2.6}$$

Next, we let $\mathbf{i} \in \mathbb{N}_+^N$ be a multi-index and, similar to [13, 14, 15], we let $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_N) \in \mathbb{R}_+^N$ be an N-dimensional weight vector for the different stochastic dimensions. Furthermore, we denote $\underline{\alpha} := \min_{1 \leq n \leq N} \alpha_n$ and consider the following index set:

$$X_{\boldsymbol{\alpha}}(\mathrm{w}, N) = \left\{ \mathbf{i} \in \mathbb{N}_+^N, \mathbf{i} \geq \mathbf{1} \,:\, \sum_{n=1}^{N} (i_n - 1)\alpha_n \leq \mathrm{w}\underline{\alpha} \right\}. \tag{2.7}$$

Then, using (2.6) and (2.7), we introduce the anisotropic Smolyak formula, $\mathscr{A}_{\boldsymbol{\alpha}}(\mathrm{w}, N)$, as

$$\mathscr{A}_{\boldsymbol{\alpha}}(\mathrm{w}, N) = \sum_{\mathbf{i} \in X_{\boldsymbol{\alpha}}(\mathrm{w}, N)} \left(\Delta^{i_1} \otimes \cdots \otimes \Delta^{i_N} \right) \tag{2.8}$$

for a given *level* $\mathrm{w} \in \mathbb{N}$. Equivalently, (2.8) can be written as

$$\mathscr{A}_{\boldsymbol{\alpha}}(\mathrm{w}, N) = \sum_{\mathbf{i} \in Y_{\boldsymbol{\alpha}}(\mathrm{w}, N)} c_{\boldsymbol{\alpha}}(\mathbf{i}) (\mathscr{U}^{i_1} \otimes \cdots \otimes \mathscr{U}^{i_N}), \tag{2.9}$$

with

$$c_{\boldsymbol{\alpha}}(\mathbf{i}) = \sum_{\substack{\mathbf{j} \in \{0,1\}^N \\ \mathbf{i} + \mathbf{j} \in X_{\boldsymbol{\alpha}}(\mathrm{w}, N)}} (-1)^{|\mathbf{j}|}$$

and

$$Y_{\boldsymbol{\alpha}}(\mathrm{w}, N) = X_{\boldsymbol{\alpha}}(\mathrm{w}, N) \setminus X_{\boldsymbol{\alpha}} \left(\mathrm{w} - \frac{|\boldsymbol{\alpha}|}{\underline{\alpha}}, N \right). \tag{2.10}$$

Similarly to the isotropic case (see [24]), to compute $\mathscr{A}_{\boldsymbol{\alpha}}(\mathrm{w}, N)(u)$, one only needs to know function values on the *anisotropic sparse grid*

$$\mathscr{H}_{\boldsymbol{\alpha}}(\mathrm{w}, N) = \bigcup_{\mathbf{i} \in Y_{\boldsymbol{\alpha}}(\mathrm{w}, N)} \left(\vartheta^{i_1} \times \cdots \times \vartheta^{i_N} \right). \tag{2.11}$$

We note that the isotropic Smolyak method studied in [24, 29, 4] is a special case of the anisotropic algorithm. This can be observed by simply taking the components of the weight vector to be equal, i.e., $\alpha_1 = \alpha_2 = \cdots = \alpha_N$. We now describe two choices of abscissas to be used in (2.8).

Clenshaw–Curtis abscissas. These abscissas are the extrema of Chebyshev polynomials (see [7]) and, for any choice of $m_i > 1$, are given by

$$y_j^i = -\cos\left(\frac{\pi(j-1)}{m_i - 1} \right), \quad j = 1, \ldots, m_i. \tag{2.12}$$

In addition, one sets $y_1^i = 0$ if $m_i = 1$ and lets the number of abscissas m_i in each level grow according to the following formula:

$$m_1 = 1 \quad \text{and} \quad m_i = 2^{i-1} + 1 \text{ for } i > 1. \tag{2.13}$$

With this particular choice, one obtains nested sets of abscissas, i.e., $\vartheta^i \subset \vartheta^{i+1}$ and thereby $\mathscr{H}_{\boldsymbol{\alpha}}(\mathrm{w}, N) \subset \mathscr{H}_{\boldsymbol{\alpha}}(\mathrm{w}+1, N)$. It is important to choose $m_1 = 1$ if we are interested in optimal approximation in relatively large N, because in all other cases the number of points used by $\mathscr{A}_{\boldsymbol{\alpha}}(\mathrm{w}, N)$ increases too fast with N.

Gaussian abscissas. These abscissas are the zeros of the orthogonal polynomials with respect to some positive weight. However, they are in general not nested. Regardless, as in the Clenshaw–Curtis case, we choose the number m_i of abscissas that are used by $\mathscr{U}^i$ as in (2.13). See the work [27] for an insightful comparison of quadrature formulas based on Clenshaw–Curtis and Gaussian abscissas. The natural choice of the weight should be the probability density function ρ of the random variables $Y_i(\omega)$, $i = 1, \ldots, N$. Yet, in the general joint multivariate case, if the random variables Y_i are not independent, the density ρ does not factorize, i.e., $\rho(y_1, \ldots, y_n) \neq \prod_{n=1}^{N} \rho_n(y_n)$. To this end, we first introduce an auxiliary probability density function $\hat{\rho} : \Gamma^N \to \mathbb{R}^+$ that can be seen as the joint probability of N independent random variables; i.e., it factorizes as

$$\hat{\rho}(y_1, \ldots, y_n) = \prod_{n=1}^{N} \hat{\rho}_n(y_n) \quad \forall y \in \Gamma^N \qquad \text{and is such that} \quad \left\| \frac{\rho}{\hat{\rho}} \right\|_{L^\infty(\Gamma^N)} < \infty. \tag{2.14}$$

For each dimension $n = 1, \ldots, N$, let the m_n Gaussian abscissas be the roots of the m_n degree polynomial that is $\hat{\rho}_n$-orthogonal to all polynomials of degree $m_n - 1$ on the interval $[-1, 1]$. The auxiliary density $\hat{\rho}$ should be chosen as close as possible to the true density ρ, so as to have the quotient $\rho/\hat{\rho}$ not too large. Indeed, such quotient will appear in the final error estimate (see section 3.1.2).

In Figure 1, for $N = 2$ and $\mathrm{w} = 7$, we show the anisotropic Smolyak sparse grids obtained utilizing the Clenshaw–Curtis abscissas and corresponding to the anisotropy ratio $\alpha_2/\alpha_1 = 1$ (isotropic grid), $\alpha_2/\alpha_1 = 3/2$, and $\alpha_2/\alpha_1 = 2$, respectively. We also show the indices $\mathbf{i} \in X_{\boldsymbol{\alpha}}(7, 2)$ that were used to construct the anisotropic sparse interpolant $\mathscr{A}_{\boldsymbol{\alpha}}(7, 2)$. In accordance with the nested structure of the Clenshaw–Curtis abscissas, in Figure 1 we point out the difference between active points/indices and previously computed points/indices.

2.2. Selection of the $\boldsymbol{\alpha}$ weights. The ability to evaluate the stochastic dimensions differently is a necessity since many practical problems exhibit highly anisotropic behavior, as illustrated by the linear problem described in section 1, where the size τ_n of the analyticity region (1.10) associated to each random variable Y_n increases with n.

We now assume that the solution to our problem has analytic dependence with respect to each of the random variables; i.e., it satisfies Assumption 1.4. In such a case, we know that if we approximate the dependence on each random variable with polynomials, the *best approximation error* decays exponentially fast with respect to the polynomial degree. More precisely, for a univariate analytic function, we recall the following lemma, whose proof can be found in [1, Lemma 7] and which is an immediate extension of the result given in [9, Chapter 7, section 8].

LEMMA 2.3. *Given a function $v \in C^0(\Gamma^1; W(D))$ which admits an analytic extension in the region of the complex plane $\Sigma(\Gamma^1; \tau) = \{z \in \mathbb{C}, \ \mathit{dist}(z, \Gamma^1) \leq \tau\}$ for some $\tau > 0$, there holds*

$$E_{m_i} \equiv \min_{w \in V_{m_i}} \|v - w\|_{C^0(\Gamma^1; W(D))} \leq \frac{2}{e^{2g} - 1} e^{-2m_i g} \max_{z \in \Sigma(\Gamma^1; \tau)} \|v(z)\|_{W(D)},$$

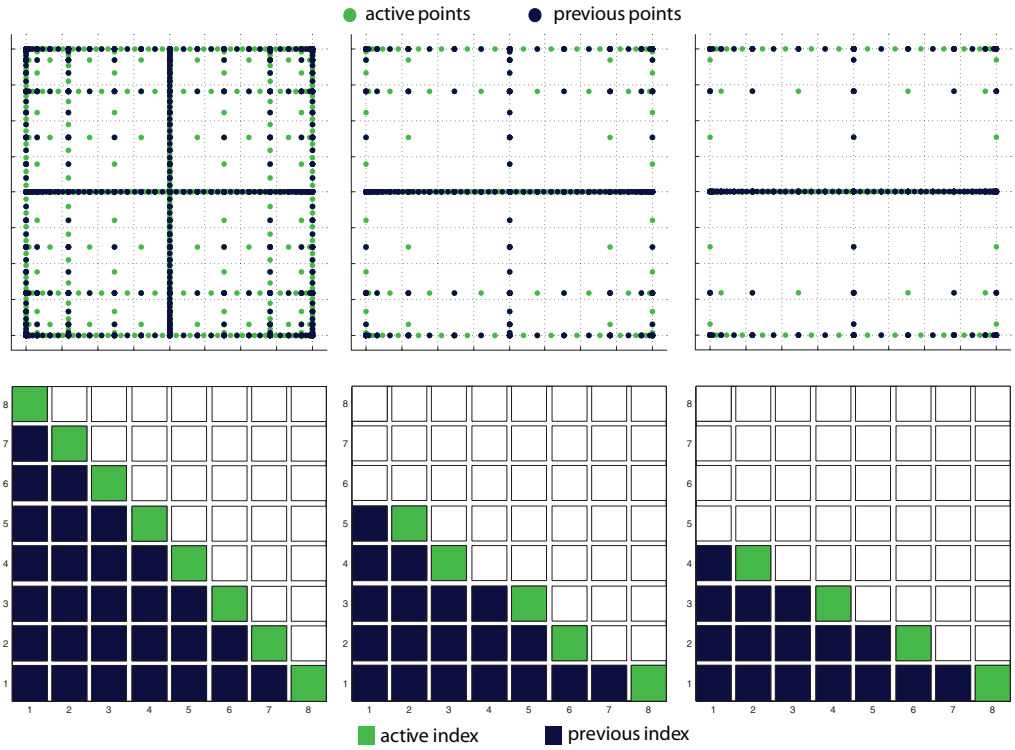


FIG. 1. *For a finite-dimensional* Γ^N *with* $N = 2$ *and* $\mathrm{w} = 7$ *we plot, in the top row, the anisotropic sparse grid with* $\alpha_2/\alpha_1 = 1$ *(isotropic),* $\alpha_2/\alpha_1 = 3/2$*, and* $\alpha_2/\alpha_1 = 2$*, utilizing the Clenshaw–Curtis abscissas. In the bottom row, we plot the corresponding indices* $(i_1, i_2) \in X_{\boldsymbol{\alpha}}(7, 2)$.

with

$$0 < g = \frac{1}{2} \log \left(\frac{2\tau}{|\Gamma^1|} + \sqrt{1 + \frac{4\tau^2}{|\Gamma^1|^2}} \right). \tag{2.15}$$

Remark 2.4 (approximation with unbounded random variables). A related result with weighted norms holds for unbounded random variables whose probability density decays as the Gaussian density at infinity (see [1]).

In the multivariate case, the size of the analyticity region, τ, will depend, in general, on the direction n (that is, $\tau = \tau_n$); cf. (1.10). As a consequence, the decay coefficient g will also depend on the direction, $n = 1, \ldots, N$, and will be denoted by $g(n)$.

The key idea of our anisotropic sparse stochastic collocation method is to place more points in those directions with slower convergence rate, i.e., with smaller $g(n)$. In particular, we propose to link the weights α_n with the rate of exponential convergence in the corresponding direction in the following way:

$$\alpha_n = g(n) \quad \forall n = 1, 2, \ldots, N. \tag{2.16}$$

In what follows we will use the notation

$$\underline{\alpha} = \underline{g} = \min_{1 \le n \le N} \{g(n)\} \quad \text{and} \quad \mathscr{G}(N) = \sum_{n=1}^{N} g(n). \tag{2.17}$$

As we will show in Remark 3.12, the choice $\boldsymbol{\alpha} = \mathbf{g}$ is optimal with respect to the error bound derived in Theorem 3.8. Observe that we have now transformed the problem of choosing $\boldsymbol{\alpha}$ into one of estimating the decay coefficients $\mathbf{g} = (g(1), \ldots, g(N))$. In the remainder of the section we will present two possible estimation strategies. The first uses a priori knowledge on the error decay in each direction, while the second approach uses a posteriori information from computations and fits the values of $\mathbf{g}$.

A priori selection. The estimation of the $\boldsymbol{\alpha} = \mathbf{g} \in \mathbb{R}_+^N$ coefficients can be achieved using a priori information, i.e., by first estimating suitable values for τ_n, and then taking $g(n)$ as in (2.15):

$$g(n) = \frac{1}{2} \log \left(\frac{2\tau_n}{|\Gamma_n|} + \sqrt{1 + \frac{4{\tau_n}^2}{|\Gamma_n|^2}} \right)$$

for $n = 1, 2, \ldots, N$. In the numerical examples presented in section 4 we have used the simpler and more conservative relation

$$g(n) = \frac{1}{2} \log \left(1 + \frac{2\tau_n}{|\Gamma_n|} \right). \tag{2.18}$$

A priori estimates for τ_n can be derived for certain classes of problems; see, for instance, the estimate (1.10), which applies to the linear problem (1.2) with random diffusivity coefficient as in (1.9). In section 4, we implement and test numerically this a priori choice of $\boldsymbol{\alpha}$ when solving problem (1.2).

A posteriori selection. In some practical applications it may not be possible to sharply estimate the vector $\mathbf{g}$ using a priori information. In this case, a computational alternative is given by a posteriori estimates. To do so, we can approximate the dependence of the solution on the nth parameter y_n by polynomial interpolation using either Clenshaw–Curtis or Gaussian abscissas, while freezing the other parameters y_n^* (for instance, we take $y_n^* = \mathbb{E}[Y_n^*]$, i.e., the mean value of the corresponding random variables). From Lemma 2.3, we expect the interpolation error to decay as

$$\varepsilon_n \approx d_n e^{-2p_n g(n)}, \tag{2.19}$$

where p_n is the number of collocation points in the direction n. Hence,

$$\log_{10}(\varepsilon_n) \approx \log_{10}(d_n) - 2p_n \log_{10}(e) g(n).$$

The error ε_n can be easily estimated by comparing two consecutive interpolants with p_n and $p_n + 1$ collocation points, respectively. Then, if we plot $\log_{10}(\varepsilon_n)$ versus p_n for each direction $n = 1, 2, \ldots, N$ and use a linear least squares approximation to fit the data, the slope of each line will give an estimate of $g(n)$.

In section 4 we will analyze the linear problem (1.2) with random diffusivity coefficient of the type (1.9) representing a Gaussian random field with a given correlation length L_c. Here we anticipate the results of the fitting of the coefficients $\mathbf{g}$ for a case with $N = 11$ random variables and correlation length $L_c = 1/64$ and $L_c = 1/2$. The former corresponds to an almost isotropic situation, while the latter corresponds to a highly anisotropic situation. Figure 2 shows the $L^2(D)$ approximation error in the expected value, $\mathbb{E}[\varepsilon_n]$. The results for $g(n)$, $n = 1, 2, \ldots, N = 11$, can be seen in Table 1, which reveals that the a posteriori selection of $\boldsymbol{\alpha} \in \mathbb{R}_+^N$ performs well at detecting the anisotropic feature of the problem in the two cases $L_c = 1/2$ and $1/64$.

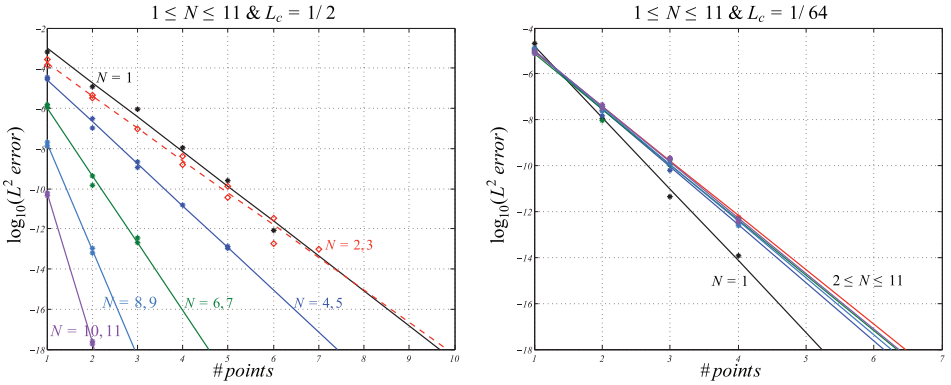


FIG. 2. *A linear least square approximation to fit* $\log_{10}(\|E[\varepsilon_n]\|_{L^2(D)})$ *versus* p_n *with* ε_n *defined by* (2.19). *For* $n = 1, 2, \ldots, N = 11$ *we plot the highly anisotropic case* $L_c = 1/2$ *(left) and the isotropic case* $L_c = 1/64$ *(right).*

TABLE 1

The $N = 11$ *values of the coefficients* $g(n)$ *constructed from a posteriori information. The values* $g(n)$, $n = 1, 2, \ldots, N$, *are the slopes of* N *linear least squares fits to univariate interpolation errors, when solving problem* (1.2) *with correlation lengths* $L_c = 1/2$ *and* $1/64$.

	$g(1)$	$g(2) = g(3)$	$g(4) = g(5)$	$g(6) = g(7)$	$g(8) = g(9)$	$g(10) = g(11)$
$L_c = 1/2$	0.85	0.80	1.0	1.6	2.6	3.7
$L_c = 1/64$	1.5	1.2	1.2	1.2	1.2	1.2

In the former case the vector $\boldsymbol{\alpha}$ weighs heavily in the higher dimensions, as opposed to the latter, which approximately weighs equally in all directions. Also, both cases $L_c = 1/2$ and $1/64$ reveal that $\underline{\alpha} = \alpha_2 = \alpha_3$.

Observe that in general the rate $g(n)$ also depends on the values $\mathbf{y}_n^* \in \Gamma_n^*$ of the other random variables. Our way of estimating the decay coefficients $g(n)$ is not conservative since we estimate them only at the single point $\mathbb{E}[Y_n^*]$. A more conservative estimate would consider the worst value of $\mathbf{y}_n^*$, i.e., the one that minimizes $g(n)$. This may be critical for nearly singular cases.

Finally, observe that the a posteriori selection of $\boldsymbol{\alpha} \in \mathbb{R}_+^N$, described above, is not restricted to random fields related to Karhunen–Loève-type expansions, and can be easily applied to any other case where the input data are described by a finite number of random variables.

3. Error analysis. Collocation methods can be used to approximate the solution $u_N \in C^0(\Gamma^N; W(D))$ using finitely many function values. By Assumption 1.4, u_N admits an analytic extension. Further, each function value will be computed by means of a finite element technique. In general, the semidiscrete finite element solution $\pi_h u_N$ also satisfies the regularity Assumption 1.4. In particular this is true for the linear problem (1.2) with random coefficients as in (1.9). Let the fully discrete approximation be $\mathscr{A}_{\boldsymbol{\alpha}}(\mathrm{w}, N)\pi_h u_N$. Our aim is to give a priori estimates for the total error

$$\epsilon = u - \mathscr{A}_{\boldsymbol{\alpha}}(\mathrm{w}, N)\pi_h u_N,$$

where the operator $\mathscr{A}_{\boldsymbol{\alpha}}(\mathrm{w}, N)$ is described by (2.8) and π_h is the finite element projection operator described by (2.2). We will investigate the error

(3.1)
$$\|u - \mathscr{A}_{\boldsymbol{\alpha}}(\mathrm{w}, N)\pi_h u_N\| \leq \underbrace{\|u - u_N\|}_{(I)} + \underbrace{\|u_N - \pi_h u_N\|}_{(II)} + \underbrace{\|\pi_h u_N - \mathscr{A}_{\boldsymbol{\alpha}}(\mathrm{w}, N)\pi_h u_N\|}_{(III)}$$

evaluated in the natural norm $L^2_P(\Omega; W(D))$. Since the error functions in (II) and (III) are finite-dimensional, the natural norm is equivalent to the $L^2_\rho(\Gamma^N; W(D))$ norm. By controlling the error in this natural norm we also control the error in the expected value of the solution, for example,

$$\|E[\epsilon]\|_{W(D)} \leq E\left[\|\epsilon\|_{W(D)}\right] \leq \|\epsilon\|_{L^2_P(\Omega; W(D))}.$$

The quantity (I) controls the truncation error for the case where the input data a_N and f_N are suitable truncations of random fields. This contribution to the total error was considered in [24, section 4.2]. The quantity (I) is otherwise zero if the representation of a_N and f_N is exact. The second term (II) controls the convergence with respect to h, i.e., the finite element error, which will be dictated by standard approximability properties of the finite element space $W_h(D)$, given by (2.1), and the regularity in space of the solution u (see, e.g., [6, 5]). Specifically,

$$\|u_N - \pi_h u_N\|_{L^2_\rho(\Gamma^N; W(D))} \leq h^s \left(\int_{\Gamma^N} C_\pi(y) C(s; u(y))^2 \rho(y)\, dy\right)^{1/2}$$

by the finite element approximability property (2.1).

In what follows, our primary concern will be to analyze the approximation error (III), namely,

$$\|\pi_h u_N - \mathscr{A}_{\boldsymbol{\alpha}}(\mathrm{w}, N)\pi_h u_N\|_{L^2_\rho(\Gamma^N; W(D))}, \tag{3.2}$$

for both the Clenshaw–Curtis and Gaussian versions of the anisotropic Smolyak algorithm.

Under the very reasonable assumption that the semidiscrete finite element solution $\pi_h u_N$ admits an analytic extension as described in Assumption 1.4 with the same analyticity region as for u_N, the behavior of the error (3.2) will be analogous to $\|u_N - \mathscr{A}_{\boldsymbol{\alpha}}(\mathrm{w}, N)u_N\|_{L^2_\rho(\Gamma^N; W(D))}$. For this reason, in the next sections we will analyze the latter.

3.1. Analysis of the anisotropic Smolyak approximation error. In order to get error bounds for the anisotropic Smolyak algorithm in the multidimensional case, we will connect the general case to the case $N = 1$ and then apply successively Lemma 2.3.

As stated in section 2.1, the anisotropic sparse grid construction treats all directions differently and relates the maximum number of collocation points used in each direction with the corresponding exponential convergence rate. We should expect, therefore, a faster convergence rate when compared to the isotropic Smolyak algorithm described in [29, 24] for such problems that exhibit strong anisotropic effects. We will analyze separately the two cases of Clenshaw–Curtis and Gaussian abscissas.

3.1.1. Interpolation estimates for the Clenshaw–Curtis abscissas. In this section we develop L^∞ error estimates for the anisotropic Smolyak interpolant

based on Clenshaw–Curtis abscissas, applied to analytic functions $u \in C^0(\Gamma^N; W(D))$ that satisfy Assumption 1.4. We remind the reader that even though in the global estimate (3.1) it is enough to bound the approximation error (III) in the $L^2_\rho(\Gamma^N; W(D))$ norm, we will still work with the more stringent $L^\infty(\Gamma^N; W(D))$ norm. In our notation the norm $\|\cdot\|_{\infty,N}$ is shorthand for $\|\cdot\|_{L^\infty(\Gamma^N;W(D))}$ and will be used henceforth.

We begin by letting E_m be the best approximation error, as in Lemma 2.3, to functions $u \in C^0(\Gamma_1; W(D))$ by polynomial functions $w \in V_m$. Since the interpolation formula $\mathscr{U}^{i_n}$, $n = 1, \dots, N$, is exact on polynomials in $V_{m_{i_n}-1}$, we can apply the general formula

$$\left\|u - \mathscr{U}^{i_n}(u)\right\|_{\infty,1} \le E_{m_{i_n}-1}(u) \cdot \left(1 + \Lambda_{m_{i_n}}\right), \tag{3.3}$$

where Λ_m is the Lebesgue constant corresponding to the abscissas (2.12). In this case, it is known that

$$\Lambda_m \le \frac{2}{\pi} \log(m-1) + 1 \tag{3.4}$$

for $m \ge 2$; see [10]. On the other hand, using Lemma 2.3 the best approximation to functions $u \in C^0(\Gamma_n; W(D))$ that admit an analytic extension as described by Assumption 1.4 is bounded by

$$E_{m_{i_n}}(u) \le \frac{2}{e^{2g(n)} - 1} e^{-2m_i g(n)} M(u), \tag{3.5}$$

where $M(u) = \max_{1\le n\le N} \max_{y_n^* \in \Gamma_n^*} \max_{z \in \Sigma(\Gamma_n;\tau_n)} \|u(z)\|_{W(D)}$. For $n = 1, 2, \dots, N$ denote the one-dimensional identity operator $I_1^{(n)} : C^0(\Gamma^n; W(D)) \to C^0(\Gamma^n; W(D))$, and use (3.3)–(3.5) to estimate

$$\begin{aligned} \left\|(I_1^{(n)} - \mathscr{U}^{i_n})(u)\right\|_{\infty,1} &\le \frac{4}{e^{2g(n)} - 1}\, i_n e^{-g(n)2^{i_n}}\, M(u), \\ \left\|(\Delta^{i_n})(u)\right\|_{\infty,1} &\le \frac{8}{e^{2g(n)} - 1}\, i_n e^{-g(n)2^{i_n - 1}} M(u). \end{aligned} \tag{3.6}$$

Since the value $M(u)$ affects the error estimates as a multiplicative constant, from here on we assume it to be one without any loss of generality.

The convergence proof will be split into several steps, the main results being given in Theorems 3.4 and 3.8, which state the convergence rates in terms of either the level w or, more interestingly, the total number of collocation points. For the purpose of error analysis we first introduce the set

$$\tilde{X}_{\boldsymbol{\alpha}}(\mathrm{w}, d) := \left\{ \mathbf{i} \in \mathbb{N}_+^d, \mathbf{i} \ge \mathbf{1} \,:\, \left(\sum_{n=1}^{d} (i_n - 1)\alpha_n - \underline{\alpha}\mathrm{w} \right) \in [0, \alpha_d] \right\}.$$

Moreover, we denote by I_d the identity operator applicable to functions which depend on the first d variables $y_1, \dots, y_d$. Then the following two lemmas, whose proofs are similar to the one presented in [24, Lemma 3.3], hold.

LEMMA 3.1 (error representation). *The following error representation holds for the Smolyak formula* (2.8)*:*

$$I_N - \mathscr{A}_{\boldsymbol{\alpha}}(\mathrm{w}, N) = \sum_{d=2}^{N} \left[\tilde{R}(\mathrm{w}, d) \bigotimes_{n=d+1}^{N} I_1^{(n)} \right] + \left(I_1^{(1)} - \mathscr{A}_{\boldsymbol{\alpha}}(\mathrm{w}, 1) \right) \bigotimes_{n=2}^{N} I_1^{(n)},$$

where, for $2 \le d \le N$, we have used the auxiliary notation

$$\tilde{R}(\mathrm{w}, d) = \sum_{\mathbf{i} \in X_{\boldsymbol{\alpha}}(\mathrm{w}, d-1)} \bigotimes_{n=1}^{d-1} \Delta^{i_n} \otimes \left(I_1^{(d)} - \mathscr{U}^{\hat{i}_d - 1} \right)$$

and, for any $(i_1, \ldots, i_{d-1}) \in X_{\boldsymbol{\alpha}}(\mathrm{w}, d-1)$, we have set $\hat{i}_d = \lfloor 2 + \mathrm{w}\frac{\alpha}{\alpha_d} - \sum_{n=1}^{d-1}(i_n - 1)\frac{\alpha_n}{\alpha_d} \rfloor$. Here $\lfloor \cdot \rfloor$ denotes the integer part of a real number.

LEMMA 3.2 (recursive estimate). *Let $u \in C^0(\Gamma^N; W(D))$ satisfy the assumptions of Lemma 2.3, and let the anisotropic Smolyak formula (2.8) be based on Clenshaw–Curtis abscissas. Then we have*

$$\left\| \left(I_N - \mathscr{A}_{\boldsymbol{\alpha}}(\mathrm{w}, N) \right)(u) \right\|_{L^\infty(\Gamma^N; W(D))} \le \sum_{d=1}^{N} R(\mathrm{w}, d), \tag{3.7}$$

with

$$R(\mathrm{w}, d) := \frac{8^d}{2} \sum_{\mathbf{i} \in \tilde{X}_{\boldsymbol{\alpha}}(\mathrm{w}, d)} \left(\prod_{n=1}^{d} \frac{i_n}{e^{2g(n)} - 1} \right) e^{-h(\mathbf{i}, d)} \tag{3.8}$$

and

$$h(\mathbf{i}, d) = \sum_{n=1}^{d} g(n) 2^{i_n - 1}. \tag{3.9}$$

LEMMA 3.3 (bound for $R(\mathrm{w}, d)$). *Let $\delta > 0$. Consider functions $u \in C^0(\Gamma^N; W(D))$ satisfying the assumption of Lemma 2.3 and the anisotropic Smolyak formula (2.8) based on Clenshaw–Curtis abscissas and weights $\alpha_n = g(n)$. Then the following bound holds for the term $R(\mathrm{w}, d)$, $d = 1, \ldots, N$, defined in (3.8):*

$$R(\mathrm{w}, d) \le C_1(\mathbf{g}, d, \delta) \exp\left\{ -\mathscr{G}(d)\, 2^{\mathrm{w}\frac{g}{\mathscr{G}(d)}} + \frac{d\delta}{2} \left\lfloor \mathrm{w}\frac{g}{\mathscr{G}(d)} \right\rfloor \right\}, \tag{3.10}$$

with

$$C_1(\mathbf{g}, d, \delta) := \frac{1}{2} \exp\left(\sum_{n=1}^{d} \left\{ \frac{1}{2} \left(\frac{1}{g(n)\log^2(2)} + \frac{1}{\log(2)\sqrt{2g(n)}} \right) + \frac{1}{\log(2)} \sqrt{\frac{\pi}{2g(n)}} \right\} \right) \times \prod_{n=1}^{d} \left(\frac{32 \exp(\delta)}{\delta e(\exp(2g(n)) - 1)} \right). \tag{3.11}$$

Proof. We follow closely the arguments in [24, Lemma 3.4]. First observe that $\forall \mathbf{i} \in \tilde{X}_{\boldsymbol{\alpha}}(\mathrm{w}, d)$ we have, denoting $\mathbf{i}^* = \left(1 + \mathrm{w}\frac{g}{\mathscr{G}(d)}, \ldots, 1 + \mathrm{w}\frac{g}{\mathscr{G}(d)}\right)$, the inequality

$$h(\mathbf{i}, d) \ge h(\mathbf{i}^*, d) + \sum_{n=1}^{d} \frac{g(n)\log^2(2)}{2} \left(i_n - \left(1 + \mathrm{w}\frac{g}{\mathscr{G}(d)} \right) \right)^2.$$

Therefore, combining the previous with (3.8) yields

$$R(\mathrm{w}, d) \le \frac{8^d}{2} e^{-h(\mathbf{i}^*, d)} \sum_{\mathbf{i} \in \tilde{X}_{\boldsymbol{\alpha}}(\mathrm{w}, d)} \left(\prod_{n=1}^{d} \frac{i_n}{e^{2g(n)} - 1} \right) \exp\left\{ -\sum_{n=1}^{d} \frac{g(n)\log^2(2)}{2} \left(i_n - \left(1 + \mathrm{w}\frac{g}{\mathscr{G}(d)} \right) \right)^2 \right\}.$$

Now recall that since $\mathbf{i} \in \tilde{X}_{\boldsymbol{\alpha}}(\mathrm{w}, d)$ we have

$$i_n \le i_{n,\max} = 1 + \left\lfloor \frac{\underline{\alpha}}{\alpha_n}\mathrm{w} + \frac{\alpha_d}{\alpha_n} \right\rfloor, \quad n = 1, \dots, d,$$

and this implies

$$R(\mathrm{w}, d) \le \frac{8^d}{2\prod_{n=1}^d (e^{2g(n)} - 1)} e^{-h(\mathbf{i}^*, d)} \prod_{n=1}^d \left(\underbrace{\sum_{i_n=1}^{i_{n,\max}} i_n \exp\left(-\frac{g(n)\log^2(2)}{2} \left(i_n - \left(1 + \mathrm{w}\frac{\underline{g}}{\mathscr{G}(d)} \right) \right)^2 \right)}_{=T(n)} \right).$$

Using the result in [24, Corollary A.4], we have the bound

$$T(n) \le \underbrace{2\left(\frac{1}{g(n)\log^2(2)} + \frac{1}{\log(2)\sqrt{2g(n)}} \right)}_{C_1} + 2\left(2 + \left\lfloor \mathrm{w}\frac{\underline{g}}{\mathscr{G}(d)} \right\rfloor \right) \left(1 + \underbrace{\frac{1}{\log(2)}\sqrt{\frac{\pi}{2g(n)}}}_{C_2} \right),$$

and

$$\begin{aligned} \prod_{n=1}^d T(n) &\le 4^d \prod_{n=1}^d \left(\frac{1}{2}C_1 + \left(1 + \frac{1}{2}\left\lfloor \mathrm{w}\frac{\underline{g}}{\mathscr{G}(d)} \right\rfloor \right)(1 + C_2) \right) \\ &\le 4^d \left(1 + \frac{1}{2}\left\lfloor \mathrm{w}\frac{\underline{g}}{\mathscr{G}(d)} \right\rfloor \right)^d \prod_{n=1}^d \left(\frac{1}{2}C_1 + (1 + C_2) \right) \\ &\le \left(\prod_{n=1}^d \frac{4e^\delta}{\delta e} \right) \exp\left(\frac{\delta d}{2}\left\lfloor \mathrm{w}\frac{\underline{g}}{\mathscr{G}(d)} \right\rfloor \right) \exp\left(\sum_{n=1}^d \left\{ \frac{1}{2}C_1 + C_2 \right\} \right). \end{aligned}$$

Observe that we have used the inequality $1 + x \le \exp(x)$ in the last step. Combining this bound with the previous estimate for $R(\mathrm{w}, d)$ leads to the desired result. ∎

THEOREM 3.4. *Consider functions $u \in C^0(\Gamma^N; W(D))$ satisfying the assumption of Lemma 2.3. The anisotropic Smolyak formula (2.8) based on Clenshaw–Curtis abscissas and weights $\alpha_n = g(n)$ satisfies the error estimate*

$$\left\| \left(I_N - \mathscr{A}_{\boldsymbol{\alpha}}(\mathrm{w}, N) \right)(u) \right\|_{L^\infty(\Gamma^N; W(D))} \le \hat{C}(\mathbf{g}, N) e^{\mathrm{w}\frac{\underline{g}}{2} - \lambda(\mathrm{w}, N)}, \tag{3.12}$$

where

$$\lambda(\mathrm{w}, N) := \begin{cases} \mathrm{w}\underline{g}\log(2)e & \text{if } 0 \le \mathrm{w} \le \frac{\mathscr{G}(N)}{\underline{g}\log(2)}, \\ \mathscr{G}(N) 2^{\mathrm{w}\frac{\underline{g}}{\mathscr{G}(N)}} & \text{otherwise}, \end{cases} \tag{3.13}$$

and the function $\hat{C}(\mathbf{g}, N) = \sum_{d=1}^{N} C_1(\mathbf{g}, d, \frac{\mathscr{G}(d)}{d})$ *does not depend on* w.

Proof. From Lemmas 3.2 and 3.3 and using constants $\delta_d > 0$, $d = 1, \ldots, N$, we obtain the following bound for the interpolation error:

$$\begin{aligned}
&\left\| \left(I_N - \mathscr{A}_{\boldsymbol{\alpha}}(\mathrm{w}, N)\right)(u) \right\|_{L^\infty(\Gamma^N; W(D))} \\
&\qquad \leq \sum_{d=1}^{N} C_1(\mathbf{g}, d, \delta_d) \exp\left(-\mathscr{G}(d)\, 2^{\mathrm{w}\frac{\underline{g}}{\mathscr{G}(d)}} + \frac{d\delta_d}{2} \left\lfloor \mathrm{w} \frac{\underline{g}}{\mathscr{G}(d)} \right\rfloor \right).
\end{aligned}$$

Now choose $\delta_d = \frac{\mathscr{G}(d)}{d}$ for $d = 1, \ldots, N$, yielding

$$\begin{aligned}
&\left\| \left(I_N - \mathscr{A}_{\boldsymbol{\alpha}}(\mathrm{w}, N)\right)(u) \right\|_{L^\infty(\Gamma^N; W(D))} \\
&\qquad \leq \sum_{d=1}^{N} C_1\left(\mathbf{g}, d, \frac{\mathscr{G}(d)}{d}\right) \exp\left(-\mathscr{G}(d) \left\{ 2^{\mathrm{w}\frac{\underline{g}}{\mathscr{G}(d)}} - \frac{1}{2} \left\lfloor \mathrm{w} \frac{\underline{g}}{\mathscr{G}(d)} \right\rfloor \right\} \right).
\end{aligned}$$

We now turn our attention to finding a minimum for $\widehat{\lambda}(\mathrm{w}, d) := \mathscr{G}(d) 2^{\mathrm{w}\frac{\underline{g}}{\mathscr{G}(d)}}$ for $1 \leq d \leq N$. Let us define, for $s \in \left[1, \frac{\mathscr{G}(N)}{\underline{g}}\right]$, the function

$$p(s) := s\, 2^{\frac{\mathrm{w}}{s}},$$

so that $\widehat{\lambda}(\mathrm{w}, d) = \underline{g} p\left(\frac{\mathscr{G}(d)}{\underline{g}}\right)$ and

$$\min_{1 \leq d \leq N} \widehat{\lambda}(\mathrm{w}, d) \geq \min_{s \in [1, \mathscr{G}(N)/\underline{g}]} \underline{g} p(s).$$

We have

$$\frac{dp}{ds} = 2^{\frac{\mathrm{w}}{s}} \left(1 - \left(\frac{\mathrm{w}}{s}\right) \log(2)\right) = 0,$$

yielding $s = \mathrm{w}\log(2)$. For w sufficiently large, the minimum of $p(s)$ falls outside the interval $\left[1, \frac{\mathscr{G}(N)}{\underline{g}}\right]$ and the function $p(s)$ is decreasing on this interval. Therefore, there are two cases to consider. The first is the situation when $\mathrm{w} > \frac{\mathscr{G}(N)}{\underline{g}\log(2)}$ and the second when $0 \leq \mathrm{w} \leq \frac{\mathscr{G}(N)}{\underline{g}\log(2)}$. In either case

$$\min_{1 \leq d \leq N} \widehat{\lambda}(\mathrm{w}, d) \geq \lambda(\mathrm{w}, N),$$

and hence

$$\max_{1 \leq d \leq N} e^{-\widehat{\lambda}(\mathrm{w}, d)} \leq e^{-\lambda(\mathrm{w}, N)}.$$

From this the result (3.12) follows by taking

$$\hat{C}(\mathbf{g}, N) = \sum_{d=1}^{N} C_1\left(\mathbf{g}, d, \frac{\mathscr{G}(d)}{d}\right). \tag{3.14}$$

∎

Remark 3.5 (behavior of $\hat{C}(\mathbf{g}, N)$ with respect to N). Denoting $k_1 = \frac{1}{2\log^2(2)}$ and $k_2 = \frac{1}{2\log(2)\sqrt{2}}(1+\sqrt{\pi})$, from (3.14) and (3.11) we observe that the constant $\hat{C}(\mathbf{g}, N)$ appearing in the convergence estimates of Theorem 3.4 has the form

(3.15)
$$\begin{aligned}\hat{C}(\mathbf{g}, N) &= \frac{1}{2}\sum_{d=1}^{N}\prod_{n=1}^{d}\left(\exp\left(k_1\frac{1}{g(n)} + k_2\frac{1}{\sqrt{g(n)}}\right)\frac{32\exp\left(g(n)\right)}{\frac{\mathscr{G}(d)}{d}e(\exp(2g(n))-1)}\right)\\ &= \frac{1}{2}\sum_{d=1}^{N}\left(\frac{16}{\frac{\mathscr{G}(d)}{d}e}\right)^{d}\exp\left(\sum_{n=1}^{N}\left\{k_1\frac{1}{g(n)} + k_2\frac{1}{\sqrt{g(n)}} - \log(\sinh(g(n)))\right\}\right).\end{aligned}$$

In the presence of anisotropic behavior, it is natural to assume that the monotone sequence $g(n)$ goes to infinity, which implies $\frac{\mathscr{G}(d)}{d} \to +\infty$. Clearly, in this situation the sum in (3.15) is convergent in N, so $\hat{C}(\mathbf{g}, N)$ does not deteriorate as the dimension of the problem grows to infinity. In particular, the more anisotropic the sequence $\mathbf{g}$ becomes (i.e., the faster $g(n)$ grows), the smaller the constant $\hat{C}(\mathbf{g}, N)$ becomes.

Observe that $\hat{C}(\mathbf{g}, N)$ goes to infinity when $\underline{g}$ tends to zero. We should note that in such case we lose convergence anyway. Sharper error estimates corresponding to small values of $g(n)$ are given in [24, Remark 3.5].

Now we relate the number of collocation points $\eta = \eta(\mathrm{w}, N) = \#\mathscr{H}_{\boldsymbol{\alpha}}(\mathrm{w}, N)$ to the level w of the anisotropic Smolyak algorithm. We state the result in the following lemma.

LEMMA 3.6. *Using the anisotropic Smolyak formula described by* (2.8)*, based on Clenshaw–Curtis abscissas, the total number of points required at level* w *satisfies the following bounds:*

$$2^{\mathrm{w}-1} \le \eta \le 2^{\mathrm{w}} e^{\mathrm{w}\sum_{n=1}^{N}\underline{\alpha}/\alpha_n}. \tag{3.16}$$

Moreover, as a direct consequence of (3.16) *we get that*

$$\frac{\log(\eta)}{\sum_{n=1}^{N}\frac{\underline{\alpha}}{\alpha_n} + \log(2)} \le \mathrm{w} \le \log_2(2\eta). \tag{3.17}$$

Proof. By using formula (2.8) and exploiting the nested structure of the Clenshaw–Curtis abscissas, we can count the number of points $\eta = \eta(\mathrm{w}, N) = \#\mathscr{H}_{\boldsymbol{\alpha}}(\mathrm{w}, N)$ in the following way:

$$\eta = \sum_{\mathbf{i}\in X_{\boldsymbol{\alpha}}(\mathrm{w},N)}\prod_{n=1}^{N} r(i_n), \quad \text{where } r(i) := \begin{cases} 1 & \text{if } i = 1,\\ 2 & \text{if } i = 2,\\ 2^{i-2} & \text{if } i > 2.\end{cases} \tag{3.18}$$

Begin by noticing that $\forall n = 1, 2, \ldots, N$ the following bound holds:

$$2^{i_n-2} \le r(i_n) \le 2^{i_n-1}. \tag{3.19}$$

Next, let $m \in [1, N]$ be the index corresponding to the minimum α, i.e., $\alpha_m = \min_{1\le n\le N}\{\alpha_n\} = \underline{\alpha}$. A lower bound on the number η of points can be obtained considering only the contribution from the tensor grid of indices $i_n = 1$ for $n \ne m$ and $i_m = \mathrm{w}+1$. On the other hand, it is easy to see that $|\mathbf{i}-\mathbf{1}| = \sum_{n=1}^{N}(i_n - 1) \le \mathrm{w}$,

so the following bounds hold:

$$2^{\mathrm{w}-1} \leq \eta = \sum_{\mathbf{i}\in X_{\boldsymbol{\alpha}}(\mathrm{w},N)} \prod_{n=1}^{N} r(i_n) \leq \sum_{\mathbf{i}\in X_{\boldsymbol{\alpha}}(\mathrm{w},N)} 2^{|\mathbf{i}-\mathbf{1}|} \leq 2^{\mathrm{w}} \# X_{\boldsymbol{\alpha}}(\mathrm{w},N).$$

We now need a bound for the cardinality of the set $X_{\boldsymbol{\alpha}}(\mathrm{w},N)$. Let us observe that the maximum index i_n^{max} allowable in each direction in the set $X_{\boldsymbol{\alpha}}(\mathrm{w},N)$ is $i_n^{max} = 1 + \lfloor \frac{\underline{\alpha}\mathrm{w}}{\alpha_n} \rfloor$, $n = 1\ldots,N$. It follows immediately that the cardinality of the set can be bound by

$$\# X_{\boldsymbol{\alpha}}(\mathrm{w},N) \leq \prod_{n=1}^{N} \left(\mathrm{w}\frac{\underline{\alpha}}{\alpha_n} + 1 \right). \tag{3.20}$$

Moreover we have

$$\# X_{\boldsymbol{\alpha}}(\mathrm{w},N) \leq \exp\left\{ \mathrm{w} \sum_{n=1}^{N} \frac{\underline{\alpha}}{\alpha_n} \right\}, \tag{3.21}$$

and the inequalities (3.16) and (3.17) follow. ∎

Remark 3.7. The bound for the cardinality of the set $\# X_{\boldsymbol{\alpha}}(\mathrm{w},N)$ given in (3.20) is not sharp when $\mathrm{w} \to \infty$, and actually one has instead the asymptotic behavior

$$\# X_{\boldsymbol{\alpha}}(\mathrm{w},N) \lesssim \frac{\mathrm{w}^N}{N!} \prod_{n=1}^{N} \frac{\underline{\alpha}}{\alpha_n},$$

which is consistent with the isotropic result given in [24]. Moreover, if we use Stirling's approximation for the factorial term, the previous bound becomes

$$\# X_{\boldsymbol{\alpha}}(\mathrm{w},N) \lesssim \frac{1}{\sqrt{N}} \prod_{n=1}^{N} \frac{e\mathrm{w}}{N} \frac{\underline{\alpha}}{\alpha_n} \leq \frac{1}{\sqrt{N}} \exp\left\{ \mathrm{w} \left(\frac{e}{N} \sum_{n=1}^{N} \frac{\underline{\alpha}}{\alpha_n} \right) \right\},$$

which greatly improves (3.21).

The next theorem provides an error bound in terms of the total number η of collocation points. The proof follows directly from the results in Theorem 3.4 and Lemma 3.6 and is therefore omitted.

THEOREM 3.8. *For functions $u \in L^2_\rho(\Gamma^N; W(D))$ satisfying the assumption of Lemma* 2.3 *the anisotropic Smolyak formula* (2.8)*, based on Clenshaw–Curtis abscissas and weights $\alpha_n = g(n)$, satisfies the following:*

- Algebraic convergence $\left(0 \leq \mathrm{w} \leq \frac{\mathscr{G}(N)}{\underline{g}\log(2)}\right)$:

$$\left\| \left(I_N - \mathscr{A}_{\boldsymbol{\alpha}}(\mathrm{w},N)\right)(u) \right\|_{L^\infty(\Gamma^N;W(D))} \leq \hat{C}(\mathbf{g},N)\eta^{-\mu_1}, \quad \mu_1 = \frac{\underline{g}\,(\log(2)e - 1/2)}{\log(2) + \sum_{n=1}^{N} \underline{g}/g(n)}. \tag{3.22}$$

- Subexponential convergence $\left(\mathrm{w} > \frac{\mathscr{G}(N)}{\underline{g}\log(2)}\right)$:

$$\left\| \left(I_N - \mathscr{A}_{\boldsymbol{\alpha}}(\mathrm{w},N)\right)(u) \right\|_{L^\infty(\Gamma^N;W(D))} \leq \hat{C}(\mathbf{g},N)\,\eta^{\frac{\mu_2}{2}}\, \exp\left(-\mathscr{G}(N)\eta^{\frac{\log(2)}{\mathscr{G}(N)}\mu_2}\right),$$

$$\textit{with} \qquad \mu_2 = \frac{\underline{g}}{\left(\log(2) + \sum_{n=1}^{N} \underline{g}/g(n)\right)}, \tag{3.23}$$

and constant $\hat{C}(\mathbf{g}, N)$ *defined in* (3.14) *and independent of* η.

Remark 3.9. The estimates given in (3.23) may be improved when $\mathrm{w} \to \infty$. Such an asymptotic estimate is obtained using the better counting result described in Remark 3.7.

Remark 3.10. We observe that subexponential rate of convergence is always faster than the algebraic one when $\mathrm{w} > \mathscr{G}(N)/(\underline{g}\log(2))$. Yet, this estimate is of little practical relevance since in practical computations such a high level w is seldom reached.

Remark 3.11 (on the curse of dimensionality). Suppose now that the stochastic input data are truncated expansions of random fields and that we are able to estimate the values $\{g(n)\}_{n=1}^{\infty}$. Whenever the sum $\sum_{n=1}^{\infty} \underline{g}/g(n)$ is finite then the algebraic exponent in (3.22) does not deteriorate as the truncation dimension N increases. This condition is satisfied, for instance, by the problem presented in the numerical section. This is a clear advantage with respect to the isotropic Smolyak method studied in [24]: since in this case we have $g(n) \to +\infty$, Remark 3.5 tells us that $\hat{C}(\mathbf{g}, N)$ does not deteriorate with N (i.e., it is bounded) and therefore the method does not suffer from the curse of dimensionality. In fact, in such a case we can work directly with the anisotropic Smolyak formula (2.8) in infinite dimensions: here the indices vectors become natural sequences and the constraint that defines the index set $X_{\boldsymbol{\alpha}}(\mathrm{w}, \infty)$ is $\sum_{n=1}^{\infty}(i_n - 1)g(n) \leq \mathrm{w}\underline{g}$.

The condition $\sum_{n=1}^{\infty} \underline{g}/g(n) < \infty$ is clearly sufficient to break the curse of dimensionality. In that case, even an anisotropic full tensor approximation also breaks the curse of dimensionality. The algebraic exponent for the convergence of the anisotropic full tensor approximation again deteriorates with the value of $\sum_{n=1}^{\infty} \underline{g}/g(n)$, but the constant of such convergence is $\sum_{n=1}^{N} \frac{2}{e^{2g(n)}-1}$. This constant is worse than that corresponding to the anisotropic Smolyak approximation, $\hat{C}(\mathbf{g}, N)$.

On the other hand, by considering the case where all $g(n)$ are equal and the results derived in [24], we see that in this work we have not estimated sharply the algebraic convergence exponent. We expect the anisotropic Smolyak method to break the curse of dimensionality in a wider set of problems; i.e., the condition $\sum_{n=1}^{\infty} \underline{g}/g(n) < \infty$ does not seem necessary to break the curse of dimensionality. This is in agreement with Remark 3.9.

Remark 3.12 (optimal choice of $\boldsymbol{\alpha}$). Looking at the exponential term $e^{-h(\mathbf{i},d)}$ in (3.8), which is the term determining the rate of convergence, we may try to choose the weight $\boldsymbol{\alpha}$ for $X_{\boldsymbol{\alpha}}(\mathrm{w}, N)$ as the solution to the optimization problem

$$\max_{\substack{\boldsymbol{\alpha}\in\mathbb{R}_+^d \\ |\boldsymbol{\alpha}|=1}} \min_{\mathbf{i}\in\tilde{X}_{\boldsymbol{\alpha}}(\mathrm{w},d)} h(\mathbf{i}, d).$$

This problem has the solution $\boldsymbol{\alpha} = \mathbf{g}$, and hence our choice of weights (2.16) is optimal.

3.1.2. Approximation estimates for Gaussian abscissas. By using a similar approach to that of our previous work [24, section 4.1.2] we can develop error estimates for approximating functions $u \in C^0(\Gamma^N; W(D))$ that admit an analytic extension as described by Assumption 1.4, using the anisotropic Smolyak formula based on Gaussian abscissas. We remind the reader that in the global estimate (3.1) we need to bound the approximation error (III) in the norm $L^2_\rho(\Gamma^N; W(D))$. However, the Gaussian abscissas described in section 2.1 are constructed for the auxiliary density

$\hat{\rho} = \prod_{n=1}^{N} \hat{\rho}_n$, and we have the following useful bound:

$$\|v\|_{L^2_\rho(\Gamma^N;W(D))} \leq \left\| \frac{\rho}{\hat{\rho}} \right\|_{L^\infty(\Gamma^N)} \cdot \|v\|_{L^2_{\hat{\rho}}(\Gamma^N;W(D))} \quad \forall v \in C^0(\Gamma^N; W(D)).$$

In what follows we will use the shorthand notation $\|\cdot\|_{\hat{\rho},N}$ for $\|\cdot\|_{L^2_{\hat{\rho}}(\Gamma^N;W(D))}$ and restrict our analysis to the case where $W(D)$ is a Hilbert space. Following [24], the one-dimensional interpolation error for Gaussian abscissas satisfies, $\forall i \in \mathbb{N}_+$,

$$\left\| (I_1^{(n)} - \mathscr{U}^{i_n})(u) \right\|_{\hat{\rho},1} \leq 2C_{\hat{\rho}} E_{m_{i_n}}(u) \leq \frac{4C_{\hat{\rho}}}{e^{2g(n)}-1} e^{-g(n)2^{i_n}} M(u),$$

$$C_{\hat{\rho}} = \max_{n=1,\ldots,N} \sqrt{\int_{\Gamma_n} \hat{\rho}_n(y)\, dy},$$

$$\left\| (\Delta^{i_n})(u) \right\|_{\hat{\rho},1} \leq 2C_{\hat{\rho}} (E_{m_{i_n}}(u) + E_{m_{i_{n-1}}}(u)) \leq \frac{8C_{\hat{\rho}}}{e^{2g(n)}-1} e^{-\frac{g(n)}{2}2^{i_n}} M(u).$$

The following theorem states the rate of convergence for the anisotropic Smolyak formula based on Gaussian abscissas. Since the proof is mostly similar to the one presented in the previous section for the Clenshaw–Curtis abscissas, we will provide a sketch of it and only the main differences will be highlighted.

THEOREM 3.13. *Let the function $u \in L^2_\rho(\Gamma^N; W(D))$ satisfy the assumptions of Lemma 2.3 and let the anisotropic Smolyak formula (2.8) be based on Gaussian abscissas. Then, choosing the anisotropy weights as $\alpha_n = g(n)$, we have the following:*

- Algebraic convergence $\left(0 \leq \mathrm{w} \leq \frac{\mathscr{G}(N)}{\underline{g}\log(2)}\right)$:

(3.24)
$$\|(I_N - \mathscr{A}_{\boldsymbol{\alpha}}(\mathrm{w}, N))(u)\|_{L^2_\rho(\Gamma^N;W(D))} \leq \hat{C}^G(\mathbf{g}, N)\eta^{-\mu_1}, \quad \mu_1 = \frac{\underline{g}\log(2)e}{2\log(2) + \sum_{n=1}^{N} \underline{g}/g(n)}.$$

- Subexponential convergence $\left(\mathrm{w} > \frac{\mathscr{G}(N)}{\underline{g}\log(2)}\right)$:

(3.25)
$$\|(I_N - \mathscr{A}_{\boldsymbol{\alpha}}(\mathrm{w}, N))(u)\|_{L^2_\rho(\Gamma^N;W(D))} \leq \hat{C}^G(\mathbf{g}, N) e^{-\frac{\mathscr{G}(N)}{2}\eta^{\mu_2}},$$
$$\textit{with} \qquad \mu_2 = \frac{\underline{g}\log(2)}{\mathscr{G}(N)\left(2\log(2) + \sum_{n=1}^{N} \underline{g}/g(n)\right)},$$

and the constant $\hat{C}^G(\mathbf{g}, N) := \frac{1}{2}\left\|\frac{\rho}{\hat{\rho}}\right\| \sum_{d=1}^{N} \exp\left(\frac{1}{\log(2)} \sum_{n=1}^{d} \sqrt{\frac{\pi}{2g(n)}}\right) \prod_{n=1}^{d} \frac{16C_{\hat{\rho}}}{e^{2g(n)}-1}$.

Proof. The recursion formula (3.7) still holds with the term $R(\mathrm{w}, d)$ defined now as

$$R(\mathrm{w}, d) = \left\|\frac{\rho}{\hat{\rho}}\right\| \frac{8^d}{2} \sum_{\mathbf{i} \in \tilde{X}_{\boldsymbol{\alpha}}(\mathrm{w},d)} e^{-h(\mathbf{i},d)} \prod_{n=1}^{d} \frac{1}{e^{2g(n)}-1}.$$

Notice that the negative effect of the Lebesgue constant for the Clenshaw–Curtis abscissas, which was responsible for the term $\prod_{n=1}^{d} i_n$ in the definition of $R(\mathrm{w}, d)$ (see (3.8)), is not present here. Then, following the guidelines of the proof of Lemma 3.3, this term can be bounded as

$$R(\mathrm{w}, d) \leq C_1^G(\mathbf{g}, d) \exp\left\{-\mathscr{G}(d) 2^{\mathrm{w}\frac{g}{\mathscr{G}(d)}}\right\},$$

with constant

$$C_1^G(\mathbf{g},d) = \left\|\frac{\rho}{\hat{\rho}}\right\| \exp\left(\frac{1}{\log(2)}\sum_{n=1}^{d}\sqrt{\frac{\pi}{2g(n)}}\right)\frac{1}{2}\prod_{n=1}^{d}\frac{16C_{\hat{\rho}}}{e^{2g(n)}-1}.$$

This leads to the estimate in terms of w,

$$\|(I_N - \mathscr{A}_{\boldsymbol{\alpha}}(\mathrm{w},N))(u)\|_{L^2_\rho(\Gamma^N;W(D))} \le \hat{C}^G(\mathbf{g},N)e^{-\lambda(\mathrm{w},N)}, \tag{3.26}$$

with the same $\lambda(\mathrm{w},N)$ as in (3.13) and $\hat{C}^G(\mathbf{g},N) = \sum_{d=1}^{N} C_1^G(\mathbf{g},d)$. Again, notice the great improvement with respect to the bound (3.12) holding for Clenshaw–Curtis abscissas. Finally, we observe that for a given level w, the number of Gauss points is larger than the number of Clenshaw–Curtis points (due to the nonnested structure), and, following [24], it is given by

$$\eta = \sum_{\mathbf{i}\in Y_{\boldsymbol{\alpha}}(\mathrm{w},N)}\prod_{n=1}^{N}\tilde{r}(i_n), \qquad \text{with} \quad \tilde{r}(i) := \begin{cases} 1 & \text{for } i=1,\\ 2^{i-1}+1 & \text{for } i>1, \end{cases} \tag{3.27}$$

where the set $Y_{\boldsymbol{\alpha}}(\mathrm{w},N)$ is defined in (2.10). Since

$$\prod_{n=1}^{N}\tilde{r}(i_n) \le \prod_{n=1}^{N}(2^{i_n-1}+1) \le 2^{2|i-1|} \le 2^{2\mathrm{w}},$$

the number of points can be bounded as

$$2^{\mathrm{w}-1} \le \eta \le 2^{2\mathrm{w}}\#Y_{\boldsymbol{\alpha}}(\mathrm{w},N) \le 2^{2\mathrm{w}}\#X_{\boldsymbol{\alpha}}(\mathrm{w},N),$$

which, substituted in (3.26), gives the desired result. ∎

Similar conclusions to those outlined in Remark 3.11 hold in this case. We can conclude, therefore, that this method breaks the curse of dimensionality as long as the sequence $g(n)$ goes to infinity sufficiently fast, meaning that the sum $\sum_{n\ge1}\frac{g}{g(n)}$ is convergent.

4. Numerical examples. This section illustrates the convergence of the anisotropic sparse collocation method for solving the stochastic linear elliptic problem described in Example 1.1 in two spatial dimensions. We will exemplify the algorithm using both the a priori and the a posteriori selection of the $\boldsymbol{\alpha}$ weights, described in section 2.2. The computational results are in accordance with the convergence rates predicted by the theory. Actually, we observe a faster convergence than stated in (3.22) and (3.24), which hints that the current estimates may still be improved.

We will also use this problem to compare the convergence of the anisotropic Smolyak method with other ensemble-based methods such as the isotropic Smolyak method described in [24], the anisotropic adaptive full tensor product method described in [3, section 9], and finally the well-known Monte Carlo method.

The problem is to solve

$$\begin{cases} -\nabla\cdot(a(\omega,\cdot)\nabla u(\omega,\cdot)) = f(\omega,\cdot) & \text{in } D\times\Omega,\\ u(\omega,\cdot) = 0 & \text{on } \partial D\times\Omega, \end{cases} \tag{4.1}$$

with $D=[0,d]^2$ and $d=1$. For this numerical example we take a deterministic load $f(\omega,x,z)=\cos(x)\sin(z)$ and construct the random diffusion coefficient $a_N(\omega,x)$ with

one-dimensional spatial dependence as

$$\log(a_N(\omega,x)-0.5) = 1 + \mathrm{Y}_1(\omega)\left(\frac{\sqrt{\pi}L}{2}\right)^{1/2} + \sum_{n=2}^{N} \zeta_n\,\varphi_n(x)\,\mathrm{Y}_n(\omega), \tag{4.2}$$

where

$$\zeta_n := \left(\sqrt{\pi}L\right)^{1/2} \exp\left(\frac{-\left(\lfloor \frac{n}{2} \rfloor \pi L\right)^2}{8}\right) \quad \text{if } n>1 \tag{4.3}$$

and

$$\varphi_n(x) := \begin{cases} \sin\left(\frac{\lfloor \frac{n}{2} \rfloor \pi x}{L_p}\right) & \text{if } n \text{ even}, \\ \cos\left(\frac{\lfloor \frac{n}{2} \rfloor \pi x}{L_p}\right) & \text{if } n \text{ odd}. \end{cases} \tag{4.4}$$

For $x \in [0,d]$ let L_c be a desired physical correlation length for the random field a, meaning that the random variables $a(x)$ and $a(y)$ become essentially uncorrelated for $|x-y| \gg L_c$. Then the parameter L_p in (4.4) can be taken as $L_p = \max\{d, 2L_c\}$ and the parameter L in (4.2) and (4.3) is $L = L_c/L_p$. Expression (4.2) represents a possible truncation of a one-dimensional random field with stationary covariance

$$\operatorname{cov}[\log(a_N - 0.5)](x_1,x_2) = \exp\left\{-\frac{(x_1-x_2)^2}{L_c^2}\right\}.$$

In this example, the random variables $\{Y_n(\omega)\}_{n=1}^{\infty}$ are independent, have zero mean and unit variance, i.e., $\mathbb{E}[Y_n]=0$ and $\mathbb{E}[Y_nY_m]=\delta_{nm}$ for $n,m \in \mathbb{N}_+$, and are uniformly distributed in the interval $[-\sqrt{3},\sqrt{3}]$.

We first consider an a priori selection of the weights $g(n)$ based on formula (2.18). For that, from (1.10) we can estimate the parameters τ_n by

$$\tau_n = \frac{1}{4\zeta_n \|\varphi\|_\infty} = \begin{cases} \sqrt{\frac{1}{8\sqrt{\pi}L}} & \text{for } n=1, \\ \sqrt{\frac{1}{16\sqrt{\pi}L}} \exp\left(\frac{\lfloor \frac{n}{2} \rfloor^2 \pi^2 L^2}{8}\right) & \text{for } n>1. \end{cases}$$

Then, using (2.18), the weight vector $\mathbf{g}$ becomes

$$g(n) = \begin{cases} \frac{1}{2}\log\left(1+\sqrt{\frac{1}{24\sqrt{\pi}L}}\right) & \text{for } n=1, \\ \frac{1}{2}\log\left(1+\sqrt{\frac{1}{48\sqrt{\pi}L}} \exp\left(\frac{\lfloor \frac{n}{2} \rfloor^2 \pi^2 L^2}{8}\right)\right) & \text{for } n>1. \end{cases} \tag{4.5}$$

In Table 2 we show the function values $g(n)$ for $n = 1,2,\ldots,N = 11$ using (4.5). With this a priori information we can construct the simplices $X_{\boldsymbol{\alpha}}(\mathrm{w},N)$ based on the weighting vector $\boldsymbol{\alpha} = \mathbf{g}$, which will be used for solving problem (4.1) with correlation lengths $L_c = 1/2, 1/4, 1/16$, and $1/64$. This table also yields insight into the anisotropic behavior of each problem. In the case of small correlation lengths, i.e., $L_c = 1/64$, we observe an almost equal weighing of all y-directions, except for the first one. The opposite behavior can be seen as we increase the correlation length. For example, when $L_c = 1/2$ the ratio between $g(11)$ and $\underline{g} = g(2) = g(3)$ is approximately $30:1$. Similar observations are made following Table 1, in section 2.2, where we utilized the a posteriori selection of the weights $g(n)$.

TABLE 2

The $N = 11$ values of the function $g(n)$ constructed from a priori information given by (4.5) for correlation lengths $L_c = 1/2, 1/4, 1/16, 1/64$.

	$g(1)$	$g(2) = g(3)$	$g(4) = g(5)$	$g(6) = g(7)$	$g(8) = g(9)$	$g(10) = g(11)$
$L_c = 1/2$	0.10	0.09	0.21	0.62	1.55	2.90
$L_c = 1/4$	0.13	0.10	0.13	0.18	0.28	0.45
$L_c = 1/16$	0.24	0.18	0.18	0.18	0.19	0.20
$L_c = 1/64$	0.39	0.31	0.31	0.31	0.31	0.31

Since the random variables Y_n are uniformly distributed, in this case the Gaussian abscissas correspond to the root of the Legendre polynomials. The finite element space for the spatial discretization is the span of continuous functions that are piecewise polynomials with degree two over a uniform triangulation of D with 4225 unknowns.

To study the convergence of the anisotropic Smolyak algorithm we consider a problem with a fixed dimension N and investigate the behavior as the level w increases. The computational results, using the a priori approach, for the $L^2(D)$ approximation error to the expected value, $\mathbb{E}[u]$, are shown in Figure 3. Here we consider

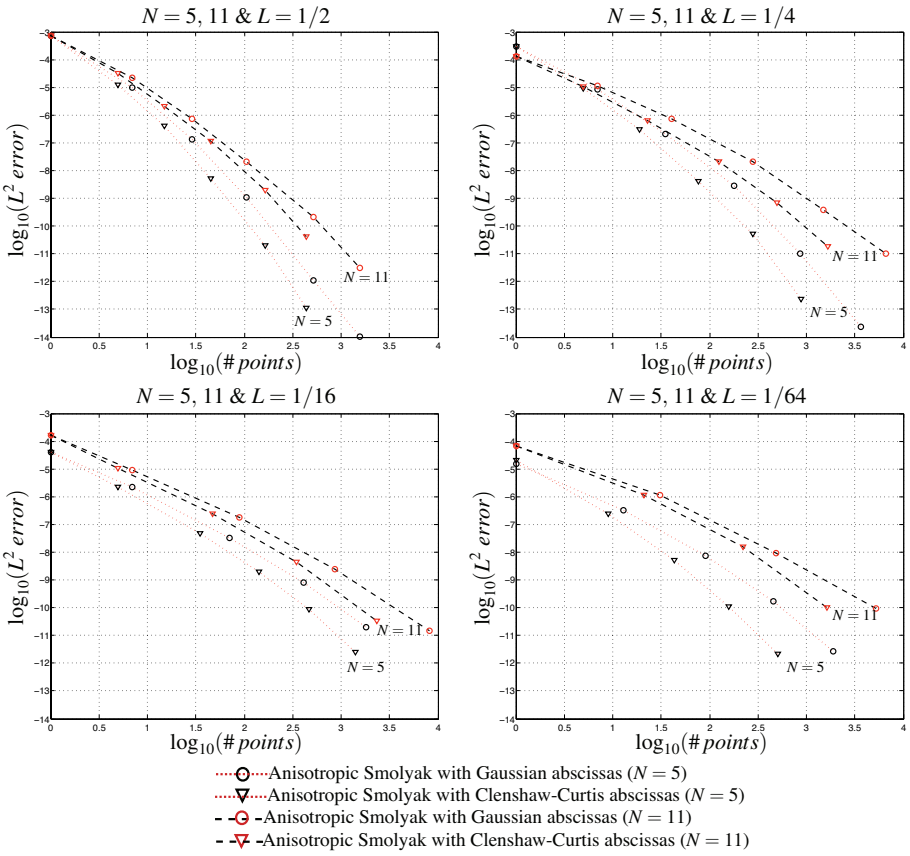


FIG. 3. *The convergence of the anisotropic Smolyak algorithm, with a priori selection of the $\boldsymbol{\alpha}$ weights, for solving problem (4.1) with given correlation lengths $L_c = 1/2, 1/4, 1/16$, and $1/64$ using both the Gaussian and Clenshaw–Curtis abscissas. For a finite-dimensional probability space Γ^N with $N = 5$ and $N = 11$ we plot the $L^2(D)$ approximation error in the expected value versus the number of collocation points.*

TABLE 3

The $N = 11$ components of the maximum indices $\boldsymbol{\kappa}(\boldsymbol{\alpha}, \mathrm{w})$ and $\widehat{\boldsymbol{\kappa}} = \boldsymbol{\kappa}(\widehat{\boldsymbol{\alpha}}, \overline{\mathrm{w}}+1)$, defined by (4.6) and (4.7), respectively, used for solving problem (4.1) with a correlation length $L_c = 1/2$.

w	$\kappa_1(\boldsymbol{\alpha}, \mathrm{w})$	$\kappa_2{=}\kappa_3(\boldsymbol{\alpha}, \mathrm{w})$	$\kappa_4{=}\kappa_5(\boldsymbol{\alpha}, \mathrm{w})$	$\kappa_6{=}\kappa_7(\boldsymbol{\alpha}, \mathrm{w})$	$\kappa_8{=}\kappa_9(\boldsymbol{\alpha}, \mathrm{w})$	$\kappa_{10}{=}\kappa_{11}(\boldsymbol{\alpha}, \mathrm{w})$
0	1	1	1	1	1	1
1	1	2	1	1	1	1
2	2	3	1	1	1	1
3	3	4	2	1	1	1
4	4	5	2	1	1	1
5	5	6	3	1	1	1
$\widehat{\boldsymbol{\kappa}}$	6	7	4	2	2	2

TABLE 4

The $N = 11$ components of the maximum indices $\boldsymbol{\kappa}(\boldsymbol{\alpha}, \mathrm{w})$ and $\widehat{\boldsymbol{\kappa}} = \boldsymbol{\kappa}(\widehat{\boldsymbol{\alpha}}, \overline{\mathrm{w}})$, defined by (4.6) and (4.7), respectively, used for solving problem (4.1) with a correlation length $L_c = 1/64$.

w	$\kappa_1(\boldsymbol{\alpha}, \mathrm{w})$	$\kappa_2{=}\kappa_3(\boldsymbol{\alpha}, \mathrm{w})$	$\kappa_4{=}\kappa_5(\boldsymbol{\alpha}, \mathrm{w})$	$\kappa_6{=}\kappa_7(\boldsymbol{\alpha}, \mathrm{w})$	$\kappa_8{=}\kappa_9(\boldsymbol{\alpha}, \mathrm{w})$	$\kappa_{10}{=}\kappa_{11}(\boldsymbol{\alpha}, \mathrm{w})$
0	1	1	1	1	1	1
1	1	2	2	2	2	2
2	2	3	3	3	3	3
3	3	4	4	4	4	4
4	4	5	5	5	5	5
$\widehat{\boldsymbol{\kappa}}$	5	6	6	6	6	6

the truncated probability space to have dimensions $N = 5$ and $N = 11$ and compute approximate solutions up to a certain level $\overline{\mathrm{w}}$. To estimate the computational error in the level w, for $0 \le \mathrm{w} \le \overline{\mathrm{w}}$, we have computed a "reference" solution in the following way: Let us denote the maximum index utilized by $\mathscr{A}_{\boldsymbol{\alpha}}(\mathrm{w}, N)$ in each stochastic direction, $\boldsymbol{\kappa}(\boldsymbol{\alpha}, \mathrm{w}) \in \mathbb{N}_+^N$, given componentwise as

$$\kappa_n(\boldsymbol{\alpha}, \mathrm{w}) \equiv \max_{\mathbf{i} \in X_{\boldsymbol{\alpha}}(\mathrm{w}, N)} \{i_n\} = 1 + \left\lfloor \frac{\underline{\alpha}\mathrm{w}}{\alpha_n} \right\rfloor. \tag{4.6}$$

Then we introduce the enriched (reference) solution,

$$\mathscr{A}_{\widehat{\boldsymbol{\alpha}}}(\overline{\mathrm{w}}+1, N)\pi_h u_N, \quad \text{with } \widehat{\alpha}_n = \left(\frac{\kappa_n(\boldsymbol{\alpha}, \overline{\mathrm{w}}) - 1}{\kappa_n(\boldsymbol{\alpha}, \overline{\mathrm{w}})}\right)\left(\frac{\alpha_n}{\underline{\alpha}}\right), \tag{4.7}$$

and approximate the computational error for $\mathrm{w} = 0, 1, 2, \ldots, \overline{\mathrm{w}}$ as

$$\|\mathbb{E}[\epsilon]\| \approx \|\mathbb{E}[\mathscr{A}_{\boldsymbol{\alpha}}(\mathrm{w}, N)\pi_h u_N - \mathscr{A}_{\widehat{\boldsymbol{\alpha}}}(\overline{\mathrm{w}}+1, N)\pi_h u_N]\|. \tag{4.8}$$

By construction, the enriched solution possesses a maximum index $\boldsymbol{\kappa}(\widehat{\boldsymbol{\alpha}}, \overline{\mathrm{w}}+1)$ which is larger by one in each direction, i.e., $\kappa_n(\widehat{\boldsymbol{\alpha}}, \overline{\mathrm{w}}+1) = \kappa_n(\boldsymbol{\alpha}, \overline{\mathrm{w}})+1$, $n = 1, \ldots, N$. For $N = 11$, Tables 3 and 4 show the values $\boldsymbol{\kappa}(\boldsymbol{\alpha}, \mathrm{w})$ for computing $\mathscr{A}_{\boldsymbol{\alpha}}(\mathrm{w}, N)$, with $\mathrm{w} = 0, 1, 2, \ldots, \overline{\mathrm{w}}$, as well as the value $\boldsymbol{\kappa}(\widehat{\boldsymbol{\alpha}}, \overline{\mathrm{w}})$ for computing the enriched solution, $\mathscr{A}_{\widehat{\boldsymbol{\alpha}}}(\overline{\mathrm{w}}, N)$, for the cases $L_c = 1/2$ and $L_c = 1/64$, respectively.

The convergence plots in Figure 3 show that the error decreases more than algebraically as the level w increases linearly. We observe that the convergence rate is dimension-dependent and deteriorates slightly as N increases. The figure also shows the performance of the algorithm by varying the correlation length L_c. We notice that the larger correlation lengths have positive effects on the rate of convergence.

This can be explained by examining $g(n)$ defined by (4.5). From this we see that as L_c becomes large the higher dimensions weigh less, which greatly reduces the number of function evaluations required by the anisotropic sparse collocation method. On the other hand, the effect of decreasing L_c is a deterioration of the rate of convergence, due to the equal weighing of all directions. In this case, our anisotropic Smolyak algorithm and the isotropic Smolyak algorithm obtain an equivalent convergence rate (see also Figure 4).

We focus our interest next on comparing the anisotropic sparse collocation method with its isotropic counterpart, as well as with the anisotropic adaptive full tensor product method, proposed in [3, 12], and the standard Monte Carlo method. We will use the a priori selection of the $\boldsymbol{\alpha}$ weights in the anisotropic Smolyak method to make this comparison.

The isotropic Smolyak algorithm for solving problem (4.1) was analyzed in [24] and can be constructed by equally weighing all stochastic directions, i.e., $\boldsymbol{\alpha} = \mathbf{1} \in \mathbb{R}^N_+$. To estimate the computational error in the level w we approximate $\|\mathbb{E}[\epsilon]\| \approx \|\mathbb{E}[\mathscr{A}(\mathrm{w}, N)\pi_h u_N - \mathscr{A}(\mathrm{w}+1, N)\pi_h u_N]\|$ using either Gaussian or Clenshaw–Curtis abscissas.

The anisotropic full tensor product algorithm relies on a full tensor polynomial approximation. Let $\mathcal{P}_{\mathbf{p}}(\Gamma^N) = \bigotimes_{n=1}^N \mathcal{P}_{p_n}(\Gamma_n)$ be the span of tensor product polynomials with degree at most $\mathbf{p} = (p_1, \ldots, p_N)$. Its dimension is $N_p = \prod_{n=1}^N (p_n + 1)$. The adaptive algorithm proposed in [3] can be described in the following way: Given a tolerance tol, the method computes a multi-index $\mathbf{p} = (p_1, p_2, \ldots, p_N)$, corresponding to the order of the approximating polynomial spaces $\mathcal{P}_{\mathbf{p}}(\Gamma^N)$. This adaptive algorithm increases the tensor polynomial degree with an anisotropic strategy: it increases the order of approximation in one direction as much as possible before considering the next direction. Table 5 and Table 6 show the values of components of the 11-dimensional multi-index $\mathbf{p}$ for different values of tol, corresponding to $L_c = 1/2$ and $L_c = 1/64$, respectively. These tables can also give insight into the anisotropic behavior of each particular problem and should be compared with Table 3 and Table 4, respectively. Observe, in particular, that for the case $L_c = 1/64$ the algorithm predicts a multi-index $\mathbf{p}$ which is equal in all directions, i.e., an isotropic tensor product space. A convergence plot for $L_c = 1/2$ and $L_c = 1/64$ can be constructed by examining each row of Table 5 and Table 6, respectively, and plotting the number of points in the tensor product grid versus the error in expectation. We estimate the error in expectation by $\|\mathbb{E}[\epsilon]\| \approx \|\mathbb{E}[u^N_{h,\mathbf{p}} - u^N_{h,\widetilde{\mathbf{p}}}]\|$, with $\widetilde{\mathbf{p}} = (p_1 + 1, p_2 + 1, \ldots, p_N + 1)$. This entails an additional computational cost, which is bounded by the factor $\exp\big(\sum_{n=1}^N 1/p_n\big)$

TABLE 5

The $N = 11$ components of the multi-index $\mathbf{p}$ computed by the anisotropic adaptive full tensor product algorithm when solving problem (4.1) with a correlation length $L_c = 1/2$.

tol	$N = 1$	$N = 2, 3$	$N = 4, 5$	$N = 6, 7$	$N = 8, 9$	$N = 10, 11$
1.0e-04	1	1	1	1	1	1
1.0e-05	2	1	1	1	1	1
1.0e-06	2	2	1	1	1	1
1.0e-07	3	2	2	1	1	1
1.0e-08	4	3	2	1	1	1
1.0e-09	4	4	3	1	1	1
1.0e-10	5	5	3	2	1	1
1.0e-11	5	5	4	2	1	1
1.0e-12	5	6	4	2	1	1

TABLE 6

The $N = 11$ components of the multi-index $\mathbf{p}$ computed by the anisotropic adaptive full tensor product algorithm when solving problem (4.1) with a correlation length $L_c = 1/64$.

tol	$N = 1$	$N = 2, 3$	$N = 4, 5$	$N = 6, 7$	$N = 8, 9$	$N = 10, 11$
1.0e-03	1	1	1	1	1	1
1.0e-06	2	2	2	2	2	2
1.0e-09	3	3	3	3	3	3
1.0e-12	4	4	4	4	4	4

times the work to compute $\mathbb{E}[u^N_{h,\mathbf{p}}]$.

The standard Monte Carlo finite element method is the most common choice when approximating statistics of the solution to (4.1). If, for instance, the aim is to approximate the solution's expected value, $\mathbb{E}[u]$, one approximates it numerically by a sample average of numerical solutions corresponding to i.i.d. realizations of the stochastic input data. Given a number of realizations, $M \in \mathbb{N}_+$, we compute the sample average as follows: For each $k = 1, \ldots, M$, sample i.i.d. realizations of $a(\omega_k, \cdot)$ and $f(\omega_k, \cdot)$, solve problem (4.1), and construct finite element approximations $u^N_h(\omega_k, \cdot)$. We note that once we have fixed the realization, $\omega = \omega_k$, the problem is completely deterministic and may be solved by standard methods as in the collocation approach. Finally, approximate $\mathbb{E}[u]$ by the sample average: $\overline{\mathbb{E}}[u^N_{h,k}; M](\cdot) := \frac{1}{M}\sum_{k=1}^{M} u^N_h(\omega_k, \cdot)$. For the cases $L_c = 1/2, 1/4, 1/16$, and $1/64$ we take $M = 2^i$, $i = 0, 1, 2, \ldots, 11$, realizations and compute the approximation to the error in expectation by $\|\mathbb{E}[\epsilon]\| \approx \|\overline{E}[u^N_{h,k}; M] - \mathbb{E}[\mathscr{A}_{\widehat{\boldsymbol{\alpha}}}(\overline{\mathrm{w}} + 1, N)\pi_h u_N]\|$, where $\mathscr{A}_{\widehat{\boldsymbol{\alpha}}}(\overline{\mathrm{w}} + 1, N)$ is the enriched anisotropic sparse solution defined previously.

To study the advantages of an anisotropic sparse approximation as opposed to the isotropic sparse or anisotropic full tensor product approximations, we show in Figure 4 the convergence of these methods when solving problem (4.1), using correlation lengths $L_c = 1/2, 1/4, 1/16$, and $L_c = 1/64$ with $N = 11$. We also include 5 ensembles of the Monte Carlo method described previously. Figure 4 reveals that for the isotropic case with $L_c = 1/64$ the anisotropic and isotropic Smolyak methods obtain a comparable convergence rate, both faster than the anisotropic full tensor product method. On the contrary, opposite behavior can be observed for $L_c = 1/2$. Since, in this case, the rate of decay in the expansion (4.2) is faster, the anisotropic full tensor method weighs heavily the important modes and, therefore, achieves a faster convergence than the isotropic Smolyak method. Similar conclusions can be made for the anisotropic Smolyak: the increased convergence for this method comes from the fact that it combines an optimal treatment of the problem anisotropy while reducing the curse of dimensionality via the use of sparse grids.

In all four cases we observe that all three methods outperform the Monte Carlo method. We know that the amount of work to reach the accuracy ϵ in the Monte Carlo approach can be approximated by $\epsilon \approx O(M^{-1/2})$ times the amount of work per sample, where M is the number of samples. This is affected only by the problem dimension through the increase of the work per sample. Nevertheless, the convergence rate is quite slow and a high level of accuracy is achieved only with a large amount of function evaluations. This can be seen from Figure 4, where we include reference lines with slopes $-1/2$ and -1, respectively, or in Table 7, where for $N = 11$ we compare the work required by each method to decrease the original error by a factor of 10^4, for all four correlation lengths $L_c = 1/2, 1/4, 1/16$, and $1/64$. Observe that the work is proportional to the number of samples (in the Monte Carlo method) or collocation points (for the stochastic collocation methods).

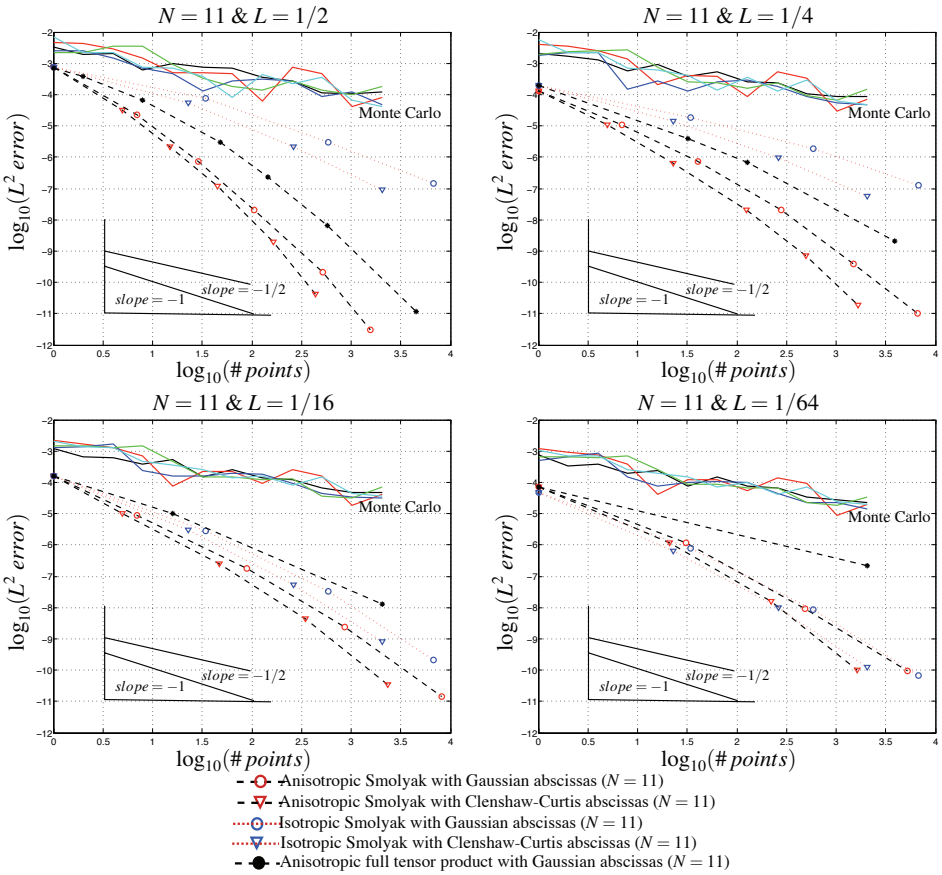


FIG. 4. *An* 11*-dimensional comparison of the anisotropic Smolyak method, with a priori selection of the* $\boldsymbol{\alpha}$ *weights, the isotropic Smolyak method, the anisotropic full tensor product algorithm, and the Monte Carlo approach for solving problem* (4.1) *with correlation lengths* $L_c = 1/2, 1/4, 1/16$, *and* $1/64$. *We plot the* $L^2(D)$ *approximation error in the expected value versus the number of collocation points (or samples of the Monte Carlo method).*

TABLE 7

For $N = 11$, *we compare the number of function evaluations required by the anisotropic Smolyak method (AS) using Clenshaw–Curtis abscissas, the anisotropic full tensor product method (AF) using Gaussian abscissas, the isotropic Smolyak (IS) using Clenshaw–Curtis abscissas, and the Monte Carlo method (MC) using random abscissas, to reduce the original error by a factor of* 10^4.

L_c	AS	AF	IS	MC
1/2	0.5×10^2	2.5×10^2	2.5×10^3	5.0×10^9
1/4	1.6×10^2	1.2×10^3	4.0×10^3	2.0×10^9
1/16	2.0×10^2	2.0×10^3	5.0×10^2	1.6×10^9
1/64	3.1×10^2	2.0×10^5	3.6×10^2	1.3×10^9

As shown in section 3.1, the ability to adaptively assess each stochastic dimension while maintaining the properties of sparse grid approximation has allowed us to make a substantial reduction in the curse of dimensionality. Therefore, our final numerical investigation concerns the convergence of the anisotropic sparse grid method when solving problem (4.1) for an increasing number of random variables (dimensions),

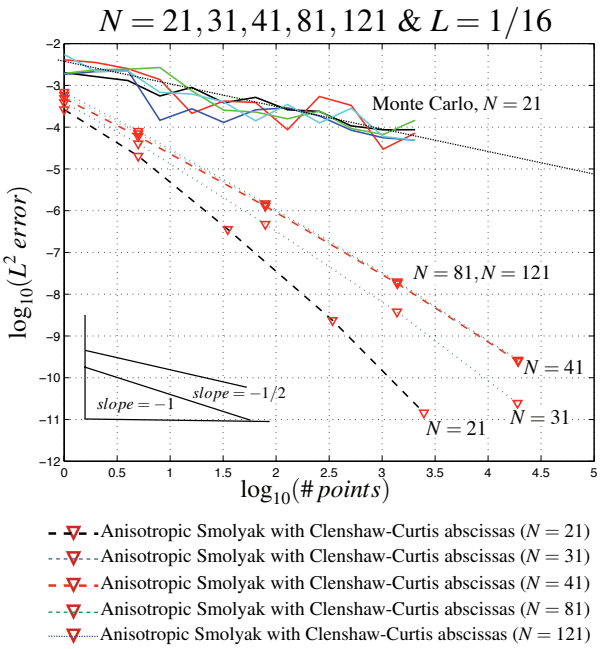


FIG. 5. *A* 21*-,* 31*-,* 41*-,* 81*-,* 121*-dimensional comparison of the anisotropic Smolyak method, with a posteriori selection of the* $\boldsymbol{\alpha}$ *weights, and a* 21*-dimensional Monte Carlo approach for solving problem* (4.1) *with correlation length* $L_c = 1/16$*. We plot the* $L^2(D)$ *approximation error in the expected value versus the number of collocation (sample) points.*

namely, $N = 21, 31, 41, 81$, and 121. In particular, we aim to confirm numerically the theoretical conclusion drawn in Remark 3.11, namely, that when the stochastic input data are truncated, expansions of random fields, the algebraic exponent in (3.22), does not deteriorate as the dimension N increases.

The computational comparison will be made with the Monte Carlo method only, since the isotropic Smolyak and anisotropic full tensor product methods require an excessively high number of points to compute accurate solutions in such high dimensions. The explosion in computational effort for these two methods is a symptom of the curse of dimensionality.

In this example we use the a posteriori selection of the $\boldsymbol{\alpha}$ weights to illustrate their practical use and consider only the case $L_c = 1/16$. The computational results are shown in Figure 5. Observe the remarkable performance of the anisotropic Smolyak method.

The convergence rate is dimension-dependent, but due to our algorithm's inherent capability to adapt itself to the eigenvalue decay of (4.2), we notice that this dependence is no longer evident as we increase the dimension from $N = 41$ to $N = 81$ and further to $N = 121$. Furthermore, the dominance of the anisotropic sparse grid method is displayed when comparing its convergence in 121 stochastic dimensions to that of Monte Carlo in only 21 dimensions.

5. Conclusions. This work proposed and analyzed a novel weighted Smolyak method, describing an optimal choice for the weight parameters. These weights tune the anisotropy of the approximation to each given problem. Their systematic choice can be based on both a priori or a posteriori information and is motivated by the regularity of the solution and the error estimates derived in this work.

The method proposed here can be viewed as a natural extension of previous works: the full tensor stochastic collocation method [1] and the isotropic sparse grid stochastic collocation method [29, 24].

The new technique consists of a Galerkin approximation in physical space and an anisotropic collocation in probability space at the zeros of sparse tensor product polynomials utilizing either Clenshaw–Curtis or Gaussian knots. As a consequence of the collocation approach our techniques naturally lead to the solution of uncoupled deterministic problems that are trivially parallelizable, as in the Monte Carlo method.

The error estimates derived in this work predict a rate of convergence that is at least algebraic with respect to the total number of collocation points. Similarly to the Monte Carlo method, the number of collocation points is directly proportional to the computational work required by the algorithm. Furthermore, the derived estimates provide a glimpse into future directions in this area. We observe that for problems possessing coefficients that are truncated expansions with proper decay in their random inputs, there may not be a deterioration in the exponent of the algebraic convergence with respect to the truncation dimension. This is a substantial advantage when compared to the previous works [1] and [29, 24].

The numerical examples included in this work gave computational ground to the theoretical results and suggest that the actual convergence rate of the method may be superior to the prediction yielded by the current error estimates. The numerical results include a comparison of the new weighted Smolyak method with the methods from [1, 24] and the standard Monte Carlo method. In particular, for moderately large-dimensional problems, the anisotropic sparse grid approach seems to be very efficient and superior to all examined methods.

Acknowledgments. The third author would like to thank Prof. Max Gunzburger for his insight, guidance, and many helpful discussions, and also Dr. John Burkardt for his assistance in implementing these numerical methods as well as his many useful visualizations.

REFERENCES

[1] I. M. Babuška, F. Nobile, and R. Tempone, *A stochastic collocation method for elliptic partial differential equations with random input data*, SIAM J. Numer. Anal., 45 (2007), pp. 1005–1034.

[2] I. M. Babuška, R. Tempone, and G. E. Zouraris, *Galerkin finite element approximations of stochastic elliptic partial differential equations*, SIAM J. Numer. Anal., 42 (2004), pp. 800–825.

[3] I. M. Babuška, R. Tempone, and G. E. Zouraris, *Solving elliptic boundary value problems with uncertain coefficients by the finite element method: The stochastic formulation*, Comput. Methods Appl. Mech. Engrg., 194 (2005), pp. 1251–1294.

[4] V. Barthelmann, E. Novak, and K. Ritter, *High dimensional polynomial interpolation on sparse grids*, Adv. Comput. Math., 12 (2000), pp. 273–288.

[5] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, Springer-Verlag, New York, 1994.

[6] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North–Holland, New York, 1978.

[7] C. W. Clenshaw and A. R. Curtis, *A method for numerical integration on an automatic computer*, Numer. Math., 2 (1960), pp. 197–205.

[8] M. K. Deb, I. M. Babuška, and J. T. Oden, *Solution of stochastic partial differential equations using Galerkin finite element techniques*, Comput. Methods Appl. Mech. Engrg., 190 (2001), pp. 6359–6372.

[9] R. A. DeVore and G. G. Lorentz, *Constructive Approximation*, Grundlehren Math. Wiss. 303, Springer-Verlag, Berlin, 1993.

[10] V. K. Dzjadyk and V. V. Ivanov, *On asymptotics and estimates for the uniform norms of the Lagrange interpolation polynomials corresponding to the Chebyshev nodal points*, Anal. Math., 9 (1983), pp. 85–97.
[11] G. S. Fishman, *Monte Carlo: Concepts, Algorithms, and Applications*, Springer Ser. Oper. Res., Springer-Verlag, New York, 1996.
[12] P. Frauenfelder, C. Schwab, and R. A. Todor, *Finite elements for elliptic problems with stochastic coefficients*, Comput. Methods Appl. Mech. Engrg., 194 (2005), pp. 205–228.
[13] B. Ganapathysubramanian and N. Zabaras, *Sparse grid collocation schemes for stochastic natural convection problems*, J. Comput. Phys., 225 (2007), pp. 652–685.
[14] J. Garcke and M. Griebel, *Classification with anisotropic sparse grids using simplicial basis functions*, Intelligent Data Anal., 6 (2002), pp. 483–502.
[15] T. Gerstner and M. Griebel, *Dimension-adaptive tensor-product quadrature*, Computing, 71 (2003), pp. 65–87.
[16] R. Ghanem, *Ingredients for a general purpose stochastic finite elements implementation*, Comput. Methods Appl. Mech. Engrg., 168 (1999), pp. 19–34.
[17] R. G. Ghanem and P. D. Spanos, *Stochastic Finite Elements: A Spectral Approach*, Springer-Verlag, New York, 1991.
[18] J. C. Helton and F. J. Davis, *Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems*, Reliab. Engrg. System Safety, 81 (2003), pp. 23–69.
[19] O. P. Le Maître, O. M. Knio, H. N. Najm, and R. G. Ghanem, *Uncertainty propagation using Wiener-Haar expansions*, J. Comput. Phys., 197 (2004), pp. 28–57.
[20] L. Mathelin, M. Y. Hussaini, and T. A. Zang, *Stochastic approaches to uncertainty quantification in CFD simulations*, Numer. Algorithms, 38 (2005), pp. 209–236.
[21] L. Mathelin and O. Le Maitre, *Dual-based a posteriori error estimation for stochastic finite element methods*, Commun. Appl. Math. Comput. Sci., 2 (2007), pp. 83–115.
[22] H. G. Matthies and A. Keese, *Galerkin methods for linear and nonlinear elliptic stochastic partial differential equations*, Comput. Methods Appl. Mech. Engrg., 194 (2005), pp. 1295–1331.
[23] H. Niederreiter, *Random Number Generation and Quasi-Monte Carlo Methods*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 63, SIAM, Philadelphia, 1992.
[24] F. Nobile, R. Tempone, and C. G. Webster, *A sparse grid stochastic collocation method for partial differential equations with random input data*, SIAM J. Numer. Anal., 46 (2008), pp. 2309–2345.
[25] L. J. Roman and M. Sarkis, *Stochastic Galerkin method for elliptic SPDEs: A white noise approach*, Discrete Contin. Dyn. Syst. Ser. B, 6 (2006), pp. 941–955.
[26] R. A. Todor and C. Schwab, *Convergence rates for sparse chaos approximations of elliptic problems with stochastic coefficients*, IMA J. Numer. Anal., 27 (2007), pp. 232–261.
[27] L. N. Trefethen, *Is Gauss quadrature better than Clenshaw–Curtis?*, SIAM Rev., 50 (2008), pp. 67–87.
[28] X. Wan and G. E. Karniadakis, *Beyond Wiener-Askey expansions: Handling arbitrary PDFs*, J. Sci. Comput., 27 (2006), pp. 455–464.
[29] D. Xiu and J. S. Hesthaven, *High-order collocation methods for differential equations with random inputs*, SIAM J. Sci. Comput., 27 (2005), pp. 1118–1139.
[30] D. Xiu and G. E. Karniadakis, *Modeling uncertainty in steady state diffusion problems via generalized polynomial chaos*, Comput. Methods Appl. Mech. Engrg., 191 (2002), pp. 4927–4948.

SIAM J. NUMER. ANAL.
Vol. 46, No. 5, pp. 2443–2468
© 2008 Society for Industrial and Applied Mathematics

SEMI-IMPLICIT TIME-DISCRETIZATION SCHEMES FOR THE BIDOMAIN MODEL*

MARC ETHIER† AND YVES BOURGAULT‡

Abstract. The bidomain model is a system of partial differential equations used to model the propagation of electrical potential waves in the myocardium. It is composed of coupled parabolic and elliptic partial differential equations, as well as at least one ordinary differential equation to model the ion activity through the cardiac cell membranes. The purpose of this paper is to propose and analyze several implicit, semi-implicit, and explicit time-stepping methods to solve that model, in particular to avoid the expensive resolution of a nonlinear system through the Newton–Raphson method. We identify necessary stability conditions on the time step Δt for the proposed methods through a theoretical analysis based on energy estimates. We next compare the methods for one- and two-dimensional test cases, in terms of both stability and accuracy of the numerical solutions. The theoretical stability conditions are seen to be consistent with those observed in practice. Our analysis allows us to recommend using either the Crank–Nicolson/Adams–Bashforth method or the second order semi-implicit backward differention method. These semi-implicit methods produce a good numerical solution; unlike the explicit methods, their stability does not depend on the spatial grid size; and unlike the implicit methods, they do not require the resolution of a system of nonlinear equations.

Key words. bidomain model, electrophysiology, finite elements, time-stepping schemes, stability analysis

AMS subject classifications. 65M60, 65M12, 65F10

DOI. 10.1137/070680503

1. Introduction.

The bidomain model is used in electrophysiology to model the propagation of electrical potential waves in the myocardium. It is obtained by homogenization over the discrete cells of the myocardium, in particular using the regular arrangement of these in fibers [7, 14]. At each point in the computational domain, two electrical potentials, namely, the intracellular potential u_i and the extracellular potential u_e, are recovered, representing the average of the electrical potential over the extracellular and the intracellular space, respectively, in the vicinity of that point.

There are many ways to write the bidomain model's equations. Many of them are introduced in [11], where their respective merits are also discussed. We will use the following formulation:

$$\frac{\partial u}{\partial t} = \frac{1}{\epsilon} f(u,v) + \nabla \cdot (\sigma_i \nabla u) + \nabla \cdot (\sigma_i \nabla u_e)\,, \tag{1.1}$$

$$\nabla \cdot (\sigma_i \nabla u + (\sigma_i + \sigma_e) \nabla u_e) = 0, \tag{1.2}$$

$$\frac{\partial v}{\partial t} = \epsilon g(u,v), \tag{1.3}$$

where $u = u_i - u_e$ is the action or transmembrane potential, σ_i and σ_e are second order tensors representing, respectively, the intracellular and the extracellular tissue's

*Received by the editors January 19, 2007; accepted for publication (in revised form) January 25, 2008; published electronically June 11, 2008. This work was partly supported by an NSERC scholarship and an NSERC operating grant.

http://www.siam.org/journals/sinum/46-5/68050.html

†Department of Mathematics and Statistics, University of Ottawa, Ottawa, ON, K1N 6N5, Canada (marcdethier@gmail.com).

‡Corresponding author. Department of Mathematics and Statistics, University of Ottawa, Ottawa, ON, K1N 6N5, Canada (ybourg@uottawa.ca).

electrical conductivity in each spatial direction, v is a lumped ionic variable, and ϵ is a parameter linked to the ratio between the repolarization rate and the tissue excitation rate [21]. It has been proved that the bidomain model has a unique solution [4, 7]. In [7] the bidomain model is formulated as a system of degenerate parabolic inequalities, while in [4] the analysis is based on the more standard theory of nonlinear parabolic PDEs. We chose to use this last formulation of the bidomain model as the basis of our numerical method because of its connection with standard parabolic and elliptic PDEs where the theory of finite element methods is well developed. We refer to [11, 19] for a more complete discussion on the different formulations of the bidomain model and their use in numerical simulations.

The equation (1.3) is not properly part of the bidomain model but rather models the ionic activity across the cellular membrane, responsible for the electrical activation of the tissue. Several models may be used for this task, the most widely known being the Hodgkin–Huxley model, originally introduced for the propagation of nerve influx inside neurons. Other models more adapted to the heart muscle have been derived, such as the Noble and Luo–Rudy models [14]. We will use the FitzHugh–Nagumo equations, which are a simplification of the Hodgkin–Huxley equations. We can therefore write the functions f and g as

$$f(u,v) = u - \frac{u^3}{3} - v, \tag{1.4}$$

$$g(u,v) = u + \beta - \gamma v, \tag{1.5}$$

where γ is a parameter controlling the ion transport and β is linked to cell excitability.

After an appropriate discretization in space, (1.1)–(1.3) read as a system of algebraic-differential equations, simply because the second equation (1.2) comes with no time derivative. The choice of a time-stepping method for that system of algebraic-differential equations has a strong impact on the stability, computational time, memory use, and quality of the numerical solution. Until recently, the numerical resolution of the bidomain model by low order explicit time-stepping methods had been very popular [11, 18, 22]. Explicit methods are easy to program and inexpensive in terms of memory use. Each time step costs typically little in terms of computational time with these methods. On the other hand, they suffer from strong limitations on the time step Δt to ensure the stability of the solution. Therefore, some of the efforts for solving the bidomain model migrated towards fully implicit [5, 11, 17], semi-implicit [9, 11, 13, 19], and operator-splitting [15, 16, 25, 26] methods. Implicit methods have much weaker conditions on the time step for their stability, but if the system to solve is nonlinear, as is the bidomain model with any ion kinetics, they require the solution of a large system of nonlinear equations at every time step. Semi-implicit methods—also called semi-explicit or implicit-explicit (IMEX)—combine explicit and implicit methods: they solve the linear terms implicitly and the nonlinear terms explicitly [1, 2, 23]. These methods are more stable than explicit methods but less so than implicit methods, and, while they require the solution of a system of equations at every time step, it is a system of linear equations.

So far, the semi-implicit methods used for the bidomain model have been only first order in time [9, 11, 13, 19], even if some of these rest on a splitting of the second order terms according to the Crank–Nicolson scheme. This is simply due to the fact that the ionic currents in the first equation (1.1) are taken explicitly as for the forward Euler scheme, lowering the global order of these semi-implicit methods. In [1, 2, 23], higher order semi-implicit schemes are proposed and analyzed. The goal of this paper

is precisely to study the use of these higher order semi-implicit schemes to solve the bidomain model.

The paper is organized as follows: we will introduce several time-stepping schemes in section 2 and then exhibit theoretical stability conditions on the time step in section 3. This will require an extension of the usual stability estimates for such semi-implicit schemes, such as done in [1], because the bidomain model can be rewritten as a system of nonlinear reaction-diffusion equations (see [4] for details) but is not solved as a standard reaction-diffusion equation. Indeed, the elliptic equation (1.2) is solved at each time step. We will obtain theoretical stability conditions showing the necessity of taking implicitly a portion of the second order terms in u and u_e, for time-stepping methods of all orders, to avoid time steps in $O(h^2)$. In section 4, numerical results will be shown with an emphasis on the stability and accuracy of the proposed schemes.

2. Numerical methods. The goal of this section is to introduce the different numerical methods studied in this article.

2.1. Space discretization. We first introduce a spatial semidiscretization of the bidomain model through first order simplicial Lagrange finite elements. Let us assume that the domain Ω can be covered by a regular partition $\mathscr{T}$ of simplexes—edges in one dimension, triangles in two dimensions, and tetrahedra in three dimensions—of maximal diameter h, with $N+1$ nodes, noted x_0 to x_N. Consider the space P_h^1 of continuous linear finite elements on $\mathscr{T}$ and the usual basis of hat functions $\Phi_0^h, \dots, \Phi_N^h$ attached to the nodes $x_0, \dots, x_N$, respectively. Here we suppose that homogeneous Neumann boundary conditions are used on $\partial\Omega$ for both u and u_e.

The semidiscrete bidomain problem then reads as: find a $[u, u_e, v] \in \mathscr{C}([0,T]; P_h^1)^3$ solution of the following variational equations:

$$\int_\Omega u_t \Phi_j^h \,\mathrm{d}\mathbf{x} = \frac{1}{\epsilon}\int_\Omega f(u,v)\Phi_j^h \,\mathrm{d}\mathbf{x} - \int_\Omega (\sigma_i \nabla(u+u_e))\cdot\nabla\Phi_j^h \,\mathrm{d}\mathbf{x}, \tag{2.1}$$

$$\int_\Omega (\sigma_i \nabla u)\cdot\nabla\Phi_j^h \,\mathrm{d}\mathbf{x} + \int_\Omega ((\sigma_i+\sigma_e)\nabla u_e)\cdot\nabla\Phi_j^h \,\mathrm{d}\mathbf{x} = 0, \tag{2.2}$$

$$\int_\Omega v_t \Phi_j^h \,\mathrm{d}\mathbf{x} = \epsilon\int_\Omega g(u,v)\Phi_j^h \,\mathrm{d}\mathbf{x} \tag{2.3}$$

for all $j = 0, \dots, N$.

By setting $u(t) = \sum_{i=0}^N u_i\Phi_i^h$, $u_e(t) = \sum_{i=0}^N u_{e,i}\Phi_i^h$, and $v(t) = \sum_{i=0}^N v_i\Phi_i^h$, we can rewrite these equations under matrix form:

$$M\mathbf{u}_t = \frac{1}{\epsilon}F(\mathbf{u},\mathbf{v}) - A_i(\mathbf{u}+\mathbf{u_e}), \tag{2.4}$$

$$A_i\mathbf{u} + (A_i+A_e)\mathbf{u_e} = \mathbf{0}, \tag{2.5}$$

$$M\mathbf{v}_t = \epsilon\, G(\mathbf{u},\mathbf{v}), \tag{2.6}$$

where $\mathbf{u} = [u_0, \dots, u_N]^T$, $\mathbf{u_e} = [u_{e,0}, \dots, u_{e,N}]^T$, $\mathbf{v} = [v_0, \dots, v_N]^T$, A_i and A_e are stiffness matrices obtained by integrating the second order terms, including, respectively, the tensors σ_i and σ_e, M is the mass matrix such that $M(i,j) = \int_\Omega \Phi_i^h\Phi_j^h \,\mathrm{d}\mathbf{x}$, and F and G are operators obtained by the integration of the zeroth order terms. For example, by using mass lumping to simplify the nonlinear operator F, we can define G and F as

$$G(\mathbf{u},\mathbf{v}) = M(\mathbf{u}-\gamma\mathbf{v}) + \beta\left[\int_\Omega \Phi_0^h \,\mathrm{d}\mathbf{x}, \dots, \int_\Omega \Phi_N^h \,\mathrm{d}\mathbf{x}\right]^T$$

and

$$F(\mathbf{u},\mathbf{v}) = M\left(\mathbf{u} - \frac{\mathbf{u^3}}{3} - \mathbf{v}\right),$$

where $\mathbf{u^3} = [u_0^3, \ldots, u_N^3]^T$.

2.2. Time discretization. We have considered a variety of explicit, implicit, and IMEX finite difference methods, using a constant time step Δt, for the time discretization. Note that the presence of (2.5) in the system has the effect that it is necessary to solve a system of equations at every time step, even for usually completely explicit methods such as the forward Euler scheme. Therefore, the following equation is part of every one of our time-discretization schemes:

$$A_i \mathbf{u^{n+1}} + (A_i + A_e)\mathbf{u_e^{n+1}} = \mathbf{0}.$$

That equation will not be repeated below, but it is understood that it is always part of the resulting fully discrete bidomain problem. We now present all of the schemes.

2.2.1. First order methods.

(i) forward Euler:

$$M\frac{\mathbf{u^{n+1}} - \mathbf{u^n}}{\Delta t} = \frac{1}{\epsilon}F(\mathbf{u^n},\mathbf{v^n}) - A_i(\mathbf{u^n} + \mathbf{u_e^n}), \tag{2.7}$$
$$M\frac{\mathbf{v^{n+1}} - \mathbf{v^n}}{\Delta t} = \epsilon G(\mathbf{u^n},\mathbf{v^n}).$$

(ii) forward-backward Euler:

$$M\frac{\mathbf{u^{n+1}} - \mathbf{u^n}}{\Delta t} = \frac{1}{\epsilon}F(\mathbf{u^n},\mathbf{v^n}) - A_i\left(\mathbf{u^{n+1}} + \mathbf{u_e^{n+1}}\right), \tag{2.8}$$
$$M\frac{\mathbf{v^{n+1}} - \mathbf{v^n}}{\Delta t} = \epsilon G(\mathbf{u^n},\mathbf{v^n}).$$

(iii) Crank–Nicolson–forward Euler (Crank–Nicolson (CN)):

$$M\frac{\mathbf{u^{n+1}} - \mathbf{u^n}}{\Delta t} = \frac{1}{\epsilon}F(\mathbf{u^n},\mathbf{v^n}) - \frac{1}{2}A_i\left(\mathbf{u^{n+1}} + \mathbf{u_e^{n+1}}\right) - \frac{1}{2}A_i(\mathbf{u^n} + \mathbf{u_e^n}), \tag{2.9}$$
$$M\frac{\mathbf{v^{n+1}} - \mathbf{v^n}}{\Delta t} = \epsilon G(\mathbf{u^n},\mathbf{v^n}).$$

(iv) IMEX first order Gear:

$$M\frac{\frac{3}{2}\mathbf{u^{n+1}} - 2\mathbf{u^n} + \frac{1}{2}\mathbf{u^{n-1}}}{\Delta t} = \frac{1}{\epsilon}F(\mathbf{u^n},\mathbf{v^n}) - A_i\left(\mathbf{u^{n+1}} + \mathbf{u_e^{n+1}}\right), \tag{2.10}$$
$$M\frac{\frac{3}{2}\mathbf{v^{n+1}} - 2\mathbf{v^n} + \frac{1}{2}\mathbf{v^{n-1}}}{\Delta t} = \epsilon G\left(\mathbf{u^{n+1}},\mathbf{v^{n+1}}\right).$$

(v) backward Euler:

$$M\frac{\mathbf{u^{n+1}} - \mathbf{u^n}}{\Delta t} = \frac{1}{\epsilon}F\left(\mathbf{u^{n+1}},\mathbf{v^{n+1}}\right) - A_i\left(\mathbf{u^{n+1}} + \mathbf{u_e^{n+1}}\right), \tag{2.11}$$
$$M\frac{\mathbf{v^{n+1}} - \mathbf{v^n}}{\Delta t} = \epsilon G\left(\mathbf{u^{n+1}},\mathbf{v^{n+1}}\right).$$

2.2.2. Second order methods.

(i) Crank–Nicolson/Adams–Bashforth (CNAB):

$$\begin{aligned} M\frac{\mathbf{u^{n+1}}-\mathbf{u^n}}{\Delta t} &= \frac{1}{\epsilon}\left(\frac{3}{2}F\left(\mathbf{u^n},\mathbf{v^n}\right)-\frac{1}{2}F\left(\mathbf{u^{n-1}},\mathbf{v^{n-1}}\right)\right) \\ &\quad -\frac{1}{2}A_i\left(\mathbf{u^{n+1}}+\mathbf{u_e^{n+1}}\right)-\frac{1}{2}A_i\left(\mathbf{u^n}+\mathbf{u_e^n}\right), \\ M\frac{\mathbf{v^{n+1}}-\mathbf{v^n}}{\Delta t} &= \epsilon\left(\frac{3}{2}G\left(\mathbf{u^n},\mathbf{v^n}\right)-\frac{1}{2}G\left(\mathbf{u^{n-1}},\mathbf{v^{n-1}}\right)\right). \end{aligned} \tag{2.12}$$

(ii) modified Crank–Nicolson/Adams–Bashforth (MCNAB):

$$\begin{aligned} M\frac{\mathbf{u^{n+1}}-\mathbf{u^n}}{\Delta t} &= \frac{1}{\epsilon}\left(\frac{3}{2}F\left(\mathbf{u^n},\mathbf{v^n}\right)-\frac{1}{2}F\left(\mathbf{u^{n-1}},\mathbf{v^{n-1}}\right)\right)-\frac{9}{16}A_i\left(\mathbf{u^{n+1}}+\mathbf{u_e^{n+1}}\right) \\ &\quad -\frac{3}{8}A_i\left(\mathbf{u^n}+\mathbf{u_e^n}\right)-\frac{1}{16}A_i\left(\mathbf{u^{n-1}}+\mathbf{u_e^{n-1}}\right), \\ M\frac{\mathbf{v^{n+1}}-\mathbf{v^n}}{\Delta t} &= \epsilon\left(\frac{3}{2}G\left(\mathbf{u^n},\mathbf{v^n}\right)-\frac{1}{2}G\left(\mathbf{u^{n-1}},\mathbf{v^{n-1}}\right)\right). \end{aligned} \tag{2.13}$$

(iii) second order semi-implicit backward differentiation (SBDF):

$$\begin{aligned} M\frac{\frac{3}{2}\mathbf{u^{n+1}}-2\mathbf{u^n}+\frac{1}{2}\mathbf{u^{n-1}}}{\Delta t} &= \frac{1}{\epsilon}\left(2F\left(\mathbf{u^n},\mathbf{v^n}\right)-F\left(\mathbf{u^{n-1}},\mathbf{v^{n-1}}\right)\right) \\ &\quad -A_i\left(\mathbf{u^{n+1}}+\mathbf{u_e^{n+1}}\right), \\ M\frac{\frac{3}{2}\mathbf{v^{n+1}}-2\mathbf{v^n}+\frac{1}{2}\mathbf{v^{n-1}}}{\Delta t} &= \epsilon\left(2G\left(\mathbf{u^n},\mathbf{v^n}\right)-G\left(\mathbf{u^{n-1}},\mathbf{v^{n-1}}\right)\right). \end{aligned} \tag{2.14}$$

(iv) implicit Gear:

$$\begin{aligned} M\frac{\frac{3}{2}\mathbf{u^{n+1}}-2\mathbf{u^n}+\frac{1}{2}\mathbf{u^{n-1}}}{\Delta t} &= \frac{1}{\epsilon}F\left(\mathbf{u^{n+1}},\mathbf{v^{n+1}}\right)-A_i\left(\mathbf{u^{n+1}}+\mathbf{u_e^{n+1}}\right), \\ M\frac{\frac{3}{2}\mathbf{v^{n+1}}-2\mathbf{v^n}+\frac{1}{2}\mathbf{v^{n-1}}}{\Delta t} &= \epsilon G\left(\mathbf{u^{n+1}},\mathbf{v^{n+1}}\right). \end{aligned} \tag{2.15}$$

2.2.3. Third order method. Third order SBDF:

$$\begin{aligned} M\frac{\frac{11}{6}\mathbf{u^{n+1}}-3\mathbf{u^n}+\frac{3}{2}\mathbf{u^{n-1}}-\frac{1}{3}\mathbf{u^{n-2}}}{\Delta t} &= \frac{1}{\epsilon}\big(3F\left(\mathbf{u^n},\mathbf{v^n}\right)-3F\left(\mathbf{u^{n-1}},\mathbf{v^{n-1}}\right) \\ &\quad +F\left(\mathbf{u^{n-2}},\mathbf{v^{n-2}}\right)\big)-A_i\left(\mathbf{u^{n+1}}+\mathbf{u_e^{n+1}}\right), \\ M\frac{\frac{11}{6}\mathbf{v^{n+1}}-3\mathbf{v^n}+\frac{3}{2}\mathbf{v^{n-1}}-\frac{1}{3}\mathbf{v^{n-2}}}{\Delta t} &= \epsilon\big(3G\left(\mathbf{u^n},\mathbf{v^n}\right)-3G\left(\mathbf{u^{n-1}},\mathbf{v^{n-1}}\right) \\ &\quad +G\left(\mathbf{u^{n-2}},\mathbf{v^{n-2}}\right)\big). \end{aligned} \tag{2.16}$$

By looking at (2.1)–(2.3), one can see that u_e appears in these equations only in the form of its gradient. Because of the Neumann boundary conditions, the variable u_e can be determined only modulo a constant. In other words, the numerical schemes

(2.7)–(2.16) give rise to singular systems of equations. There exist many ways to solve this problem [3]; we chose to fix u_e at a specific value on a particular node for every time step. Note that the indefiniteness on u is naturally removed by the nonlinear term in f in the first equation of the bidomain model. See [4] for details on that indefiniteness of the system, its physical significance, and its mathematical treatment for the continuous problem.

Even after removing the indefiniteness, the resulting linear systems are still very ill conditioned. The choice of linear solver is therefore hard. In the one-dimensional (1D) case, it is possible to use a direct solver such as the LU factorization. Because of memory constraints, this becomes much less possible in higher dimensions. Using an iterative method [10, 24] becomes necessary. Our numerical tests have convinced us that the conjugate gradient method is most appropriate in this case. We have also used the incomplete LU factorization to precondition the system. Our computer memory constraints drive the level of fill we may use with this preconditioner: in 2D with moderately fine meshes, we have been able to use two levels of fill, but finer meshes, as well as 3D meshes, would probably require us to use only one level of fill or none at all.

3. Numerical analysis of the schemes. This section is dedicated to demonstrating that the previously seen numerical schemes remain stable under appropriate conditions on the time step Δt. We also analyze the semidiscrete problem introduced in section 2.1 and prove that its solution is unconditionally stable.

In the rest of the paper, $\|\cdot\|_0$ refers to the L^2 norm over Ω and $|\cdot|_1$ to the H^1 seminorm over Ω. The constants m_i and m_e denote the infimum of the eigenvalues of σ_i and σ_e, respectively, and M_i and M_e the supremum of these eigenvalues, with the infimum and supremum taken all over Ω as well.

3.1. Stability of the semidiscrete method. We analyze the semidiscrete formulation (2.1)–(2.3) of the bidomain model. We first redefine the function g as $g(u,v) = u - \gamma v$ and set $S = \epsilon\beta$. This ensures that the graph of f and g goes through the origin. The term S then goes into the right-hand side of the PDE.

By using as test functions $\Phi_j^h = u(t)$, $u_e(t)$, and $v(t)$ in (2.1), (2.2), and (2.3), respectively, adding the three equations so obtained, and using the following lower bound for the diffusive terms:

$$\begin{aligned}\int_\Omega (\sigma_i\nabla(u+u_e))\cdot\nabla(u+u_e)\,\mathrm{d}\mathbf{x} + \int_\Omega (\sigma_e\nabla u_e)\cdot\nabla u_e\,\mathrm{d}\mathbf{x} &\geq m_i|u+u_e|_1^2 + m_e|u_e|_1^2 \\ &\geq \bar{m}(|u|_1^2 + |u_e|_1^2),\end{aligned} \tag{3.1}$$

where $\bar{m}$ depends on m_i and m_e by equivalence of seminorms, one easily obtains the following inequalities:

$$\begin{aligned}&\frac{d}{dt}\left[\|u\|_0^2 + \|v\|_0^2\right] + 2\bar{m}\left(|u|_1^2 + |u_e|_1^2\right) \\ &\qquad\leq \frac{2}{\epsilon}\left(\int_\Omega u^2\,\mathrm{d}\mathbf{x} - \frac{1}{3}\int_\Omega u^4\,\mathrm{d}\mathbf{x} - \int_\Omega uv\,\mathrm{d}\mathbf{x}\right) \\ &\qquad\quad + 2\epsilon\left(\int_\Omega uv\,\mathrm{d}\mathbf{x} - \gamma\int_\Omega v^2\,\mathrm{d}\mathbf{x}\right) + 2\int_\Omega Sv\,\mathrm{d}\mathbf{x} \\ &\qquad\leq \frac{2}{\epsilon}\|u\|_0^2 - 2\epsilon\gamma\|v\|_0^2 + \frac{\left|\frac{1}{\epsilon}-\epsilon\right|}{k}\|u\|_0^2 + \left|\frac{1}{\epsilon}-\epsilon\right| k\|v\|_0^2 + \frac{1}{r}\|S\|_0^2 + r\|v\|_0^2 \\ &\qquad\leq C_1\left[\|u\|_0^2 + \|v\|_0^2\right] + C_2\|S\|_0^2,\end{aligned} \tag{3.2}$$

where the constants k, $r > 0$, $C_1 = \max\{2/\epsilon + |1/\epsilon - \epsilon|/k, |1/\epsilon - \epsilon|k + r - 2\epsilon\gamma\}$, and $C_2 = 1/r$, owing to the fact that, for $a, b \in \mathbb{R}$ and for $k > 0$, $ab \le a^2/2k + b^2k/2$.

Applying Gronwall's lemma [8, Lemma 6.9, p. 284] to the inequality (3.2) gives:

$$\|u(t)\|_0^2 + \|v(t)\|_0^2 \le e^{C_1 t}\left[\|u(0)\|_0^2 + \|v(0)\|_0^2\right] + \int_0^t e^{C_1(t-\tau)}\left[C_2\|S\|_0^2 - 2\bar{m}\left(|u|_1^2 + |u_e|_1^2\right)\right]\mathrm{d}\tau \tag{3.3}$$

$$\Rightarrow \max_{t\in[0,T]}\left[\|u(t)\|_0^2 + \|v(t)\|_0^2\right] \le e^{C_1 T}\left[\|u_0\|_0^2 + \|v_0\|_0^2\right] + C_2\int_0^T e^{C_1(T-\tau)}\|S\|_0^2\,\mathrm{d}\tau,$$

where u_0 and $v_0 \in P_h^1$ are the initial data for the semidiscrete formulation. In other words, u and v remain bounded in $L^\infty(0,T;L^2(\Omega))$.

From (3.3), we obtain

$$2\bar{m}\int_0^T e^{C_1(T-\tau)}\left(|u|_1^2 + |u_e|_1^2\right)\mathrm{d}\tau \le e^{C_1 T}\left[\|u_0\|_0^2 + \|v_0\|_0^2\right] + C_2\int_0^T e^{C_1(T-\tau)}\|S\|_0^2\,\mathrm{d}\tau$$

$$\Rightarrow \int_0^T \left(|u|_1^2 + |u_e|_1^2\right)\mathrm{d}\tau \le \frac{e^{C_1 T}}{2\bar{m}}\left[\|u_0\|_0^2 + \|v_0\|_0^2\right] + \frac{C_2}{2\bar{m}}\int_0^T e^{C_1(T-\tau)}\|S\|_0^2\,\mathrm{d}\tau. \tag{3.4}$$

This shows that the $H^1(\Omega)$ seminorm of u and u_e remains L^2 bounded on any interval $[0,T]$.

By putting together the bounds obtained above, we find that the solution of the semidiscrete problem remains bounded in the following spaces:

$$\begin{aligned} u &\in L^\infty(0,T;L^2(\Omega)) \cap L^2(0,T;H^1(\Omega)),\\ u_e &\in L^2(0,T;H^1(\Omega)/\mathbb{R}),\\ v &\in L^\infty(0,T;L^2(\Omega)). \end{aligned} \tag{3.5}$$

In particular, the extracellular potential u_e is bounded only in $L^2(0,T;H^1(\Omega)/\mathbb{R})$, not in any $L^p(0,T;L^2(\Omega))$, since we can determine u_e only up to a constant. It is important to note that the relations (3.5) make use of the fact that the FitzHugh–Nagumo model is used for the ion activity. A nice feature is that the constants in the estimates (3.5) are independent from the cell size h of the space discretization. The semidiscrete problem is unconditionally stable. This is a main argument behind the existence proof given in [4].

3.2. Stability of the discrete methods. Let us now investigate the stability of some of the fully discrete numerical schemes introduced in section 2. We keep track as much as we can of the constants in the stability estimates to be able to identify the relevant parameters in the bidomain model controlling the stability of the numerical schemes.

3.2.1. Backward Euler. We consider the backward Euler method (2.11) as a representative of the implicit methods we have used. By using as test function Φ the

solutions u^{n+1}, u_e^{n+1}, and v^{n+1} in their respective variational equations, we obtain:

$$\int_\Omega \left[\frac{u^{n+1}-u^n}{\Delta t} - \frac{1}{\epsilon} f(u^{n+1}, v^{n+1})\right] u^{n+1}\,\mathrm{d}\mathbf{x} + \int_\Omega \left(\sigma_i \nabla(u^{n+1}+u_e^{n+1})\right)\cdot\nabla u^{n+1}\,\mathrm{d}\mathbf{x} = 0,$$

$$\int_\Omega (\sigma_i \nabla u^{n+1})\cdot\nabla u_e^{n+1}\,\mathrm{d}\mathbf{x} + \int_\Omega \left((\sigma_i+\sigma_e)\nabla u_e^{n+1}\right)\cdot\nabla u_e^{n+1}\,\mathrm{d}\mathbf{x} = 0,$$

$$\int_\Omega \left[\frac{v^{n+1}-v^n}{\Delta t} - \epsilon g(u^{n+1}, v^{n+1})\right] v^{n+1}\,\mathrm{d}\mathbf{x} = \int_\Omega S^n v^{n+1}\,\mathrm{d}\mathbf{x}. \tag{3.6}$$

By adding these three equations, using the identity

$$2(a^{n+1}-a^n)a^{n+1} = (a^{n+1})^2 + (a^{n+1}-a^n)^2 - (a^n)^2,$$

and proceeding as for the semidiscrete method, we get:

$$\begin{aligned}\|u^{n+1}\|_0^2 + \|u^{n+1}-u^n\|_0^2 - \|u^n\|_0^2 + \|v^{n+1}\|_0^2 + \|v^{n+1}-v^n\|_0^2 - \|v^n\|_0^2 \\ + 2\Delta t\,\bar m\left(|u^{n+1}|_1^2 + |u_e^{n+1}|_1^2\right) \\ \le \Delta t\, C_1\left[\|u^{n+1}\|_0^2 + \|v^{n+1}\|_0^2\right] + \Delta t\, C_2 \|S^n\|_0^2,\end{aligned} \tag{3.7}$$

where C_1 and C_2 are defined as they were in the semidiscrete case. Now denote by M the index of the final time step (i.e., $T = M\Delta t$), and sum the previous inequality for n going from 0 to $m-1$, $1 \le m \le M$, to obtain:

$$\begin{aligned}(1-\Delta t\, C_1)\left[\|u^m\|_0^2 + \|v^m\|_0^2\right] + 2\Delta t\bar m \sum_{n=1}^{m}\left(|u^n|_1^2 + |u_e^n|_1^2\right) \\ \le \|u^0\|_0^2 + \|v^0\|_0^2 + \Delta t\, C_1 \sum_{n=0}^{m-1}\left[\|u^n\|_0^2 + \|v^n\|_0^2\right] + \Delta t\, C_2 \sum_{n=0}^{m-1}\|S^n\|_0^2.\end{aligned} \tag{3.8}$$

Now, from the discrete Gronwall lemma [20, Lemma 1.4.2, p. 14], by assuming that $\Delta t < 1/C_1$, it can be concluded that

$$\begin{aligned}\max_{n=1,\dots,M}\left\{\|u^n\|_0^2 + \|v^n\|_0^2\right\} \le \Bigg(\frac{1}{1-\Delta t\, C_1}\left[\|u^0\|_0^2 + \|v^0\|_0^2\right] \\ + \frac{C_2}{1-\Delta t\, C_1}\sum_{n=0}^{M-1}\Delta t\,\|S^n\|_0^2\Bigg) e^{\frac{TC_1}{1-\Delta t C_1}}.\end{aligned} \tag{3.9}$$

Let us extend the numerical solutions at discrete times t_n to functions u and v piecewise constant in time and such that $u(n\Delta t) = u^n, v(n\Delta t) = v^n$ for $n = 0, \dots, M-1$. The last inequality means that $u, v \in L^\infty(0,T;L^2(\Omega))$. This property is true as long as $\sum_{n=0}^{M-1}\Delta t\|S^n\|_0^2$ remains bounded, which holds if $S \in L^2(0,T;L^2(\Omega))$.

Now return to (3.8). We can use that inequality to bound the other left-hand term:

$$\begin{aligned}\sum_{n=1}^{M}\Delta t\left(|u^n|_1^2 + |u_e^n|_1^2\right) \le \frac{1}{2\bar m}\left(\|u^0\|_0^2 + \|v^0\|_0^2\right) + \frac{TC_1}{2\bar m}\max_n\left\{\|u^n\|_0^2 + \|v^n\|_0^2\right\} \\ + \frac{C_2}{2\bar m}\sum_{n=0}^{M-1}\Delta t\,\|S^n\|_0^2.\end{aligned} \tag{3.10}$$

The left-hand term in this last inequality is nothing more—extending the solution u and u_e to piecewise functions as before—than the $L^2(0,T)$ norm of $|u|_1$ and $|u_e|_1$. We can thus conclude that the solutions u, u_e, and v of the fully discrete backward Euler method live in the same functional spaces as in the semidiscrete case. The difference is that we have here a constraint on the time step Δt. To ensure the stability of the method, we must have $\Delta t < 1/C_1$. Since $C_1 = O(1/\epsilon)$ for $\epsilon \ll 1$, the stability of the backward Euler scheme is guaranteed for $\Delta t = O(\epsilon)$. Note that this condition is independent of h. The parameter ϵ controls the thickness of the depolarization/repolarization fronts, meaning that the time step Δt must be small enough to resolve that front to ensure stability.

3.2.2. Second order SBDF. We shall now consider the stability of the second order SBDF method (2.14). With u^{n+1}, u_e^{n+1}, and v^{n+1} as test functions, the associated variational equations read as follows:

$$\begin{aligned}
&\int_\Omega \left[\frac{3u^{n+1}-4u^n+u^{n-1}}{2\Delta t} - \frac{1}{\epsilon}\left(2f(u^n,v^n)-f(u^{n-1},v^{n-1})\right)\right]u^{n+1}\,\mathrm{d}\mathbf{x} \\
&\qquad + \int_\Omega \left(\sigma_i\nabla(u^{n+1}+u_e^{n+1})\right)\cdot\nabla u^{n+1}\,\mathrm{d}\mathbf{x} = 0, \\
&\int_\Omega (\sigma_i\nabla u^{n+1})\cdot\nabla u_e^{n+1}\,\mathrm{d}\mathbf{x} + \int_\Omega \left((\sigma_i+\sigma_e)\nabla u_e^{n+1}\right)\cdot\nabla u_e^{n+1}\,\mathrm{d}\mathbf{x} = 0, \\
&\int_\Omega \left[\frac{3v^{n+1}-4v^n+v^{n-1}}{2\Delta t} - \epsilon\left(2g(u^n,v^n)-g(u^{n-1},v^{n-1})\right)\right]v^{n+1}\,\mathrm{d}\mathbf{x} \\
&\qquad = \int_\Omega S^n v^{n+1}\,\mathrm{d}\mathbf{x}.
\end{aligned} \tag{3.11}$$

Since the reaction terms are taken explicitly, a milder dependency of the reaction terms on the variables u and v must be assumed; e.g., we suppose that f and g satisfy the following Lipschitz condition:

$$\begin{aligned}
\|f(u,v)\|_0 &\le L_f\left[\|u\|_0+\|v\|_0\right], \\
\|g(u,v)\|_0 &\le L_g\left[\|u\|_0+\|v\|_0\right],
\end{aligned} \tag{3.12}$$

where L_f, $L_g > 0$.

Remark 3.1. The functions $f(u,v) = u - \frac{u^3}{3} - v$ and $g(u,v) = u - \gamma v$ in the FitzHugh–Nagumo ionic model are not Lipschitzian. However, if we make the hypothesis that $u \in L^\infty(\Omega)$ and that there exists a constant $\mathscr{C} > 0$ such that $\|u\|_\infty \le \mathscr{C}$ for all time $t \in [0,T]$, we have

$$\begin{aligned}
\|f(u,v)\|_0 \le \|u\|_0 + \frac{1}{3}\|u^3\|_0 + \|v\|_0 &\le \|u\|_0 + \frac{1}{3}\|u\|_\infty^2\|u\|_0 + \|v\|_0 \\
&\le \left(\frac{\mathscr{C}^2}{3}+1\right)\|u\|_0 + \|v\|_0 \\
&\le L_f\left[\|u\|_0+\|v\|_0\right],
\end{aligned}$$

where $L_f = \frac{\mathscr{C}^2}{3}+1$. We can also bound the function g:

$$\|g(u,v)\|_0 \le \|u\|_0 + \gamma\|v\|_0 \le L_g\left[\|u\|_0+\|v\|_0\right],$$

where $L_g = \max\{1,\gamma\}$. We do observe numerically that, provided that the initial solution $(u_0(x), v_0(x))$ at all points $x \in \Omega$ is in a properly chosen box in the (u,v)-plane,

$(u(x,t), v(x,t))$ remains in that box for all time. In other words, u remains bounded in $L^\infty(\Omega)$, and the constants L_f and L_g can be practically identified. Basically it would mean that we can replace the cubic function f of the FitzHugh–Nagumo model by an affine function $\tilde{f}$ outside of that box without changing the solution, the function $\tilde{f}$ now being Lipschitzian. It should be noted though that, as far as we know, such a maximum principle has not been proven for the bidomain model with FitzHugh–Nagumo ion kinetics.

By adding the equations of system (3.11) and applying the identity

$$\begin{aligned}(6a^{n+1} - 8a^n + 2a^{n-1})a^{n+1} = (a^{n+1})^2 + (2a^{n+1} - a^n)^2 + (\delta_{tt}a^{n+1})^2 \\ - (a^n)^2 - (2a^n - a^{n-1})^2,\end{aligned}$$

where $\delta_{tt}a^{n+1} = a^{n+1} - 2a^n + a^{n-1}$ to the time-discretization term, we obtain

$$\begin{aligned}&\|u^{n+1}\|_0^2 + \|2u^{n+1} - u^n\|_0^2 - \|u^n\|_0^2 - \|2u^n - u^{n-1}\|_0^2 + \|v^{n+1}\|_0^2 \\ &\quad + \|2v^{n+1} - v^n\|_0^2 - \|v^n\|_0^2 - \|2v^n - v^{n-1}\|_0^2 + 4\Delta t\,\bar{m}(|u^{n+1}|_1^2 + |u_e^{n+1}|_1^2) \\ &\quad \le \Delta t\,C_1\|u^{n+1}\|_0^2 + \Delta t\,C_2\|v^{n+1}\|_0^2 + \Delta t\,C_3\|u^n\|_0^2 + \Delta t\,C_4\|v^n\|_0^2 \\ &\quad + \Delta t\,C_5\|u^{n-1}\|_0^2 + \Delta t\,C_6\|v^{n-1}\|_0^2 + \Delta t\,C_7\|S^n\|_0^2. \qquad (3.13)\end{aligned}$$

Now sum this inequality for n going from 0 to $m-1$, where m is an integer between 1 and M. We will see the values u^{-1} and v^{-1} appear. They are used during the first time step using the second order SBDF scheme; we will take them equal to u^0 and v^0, respectively. This yields the following result:

$$\begin{aligned}&(1 - \Delta t\,C)\left[\|u^m\|_0^2 + \|v^m\|_0^2\right] + 4\Delta t\,\bar{m}\sum_{n=1}^{m}\left(|u^n|_1^2 + |u_e^n|_1^2\right) \\ &\le 2\left(\|u^0\|_0^2 + \|v^0\|_0^2\right) + \Delta t\,K\sum_{n=0}^{m-1}\left(\|u^n\|_0^2 + \|v^n\|_0^2\right) + \Delta t\,C_7\sum_{n=0}^{m-1}\|S^n\|_0^2, \qquad (3.14)\end{aligned}$$

where $C = \max\{C_1, C_2\}$ and $K = \max\{C_1 + C_3 + 2C_5, C_2 + C_4 + 2C_6\}$. Choose Δt such that $\Delta t < 1/C$, and, by applying the discrete Gronwall lemma to (3.14), obtain

$$\begin{aligned}\max_{n=1,\dots,M}\left\{\|u^n\|_0^2 + \|v^n\|_0^2\right\} \le \Bigg(&\frac{2}{1 - \Delta t\,C}\left[\|u^0\|_0^2 + \|v^0\|_0^2\right] \\ &+ \frac{C_7}{1 - \Delta t\,C}\sum_{n=0}^{M-1}\Delta t\,\|S^n\|_0^2\Bigg)e^{TK/(1-\Delta tC)}. \qquad (3.15)\end{aligned}$$

By combining (3.15) and (3.14), we write

$$\begin{aligned}\sum_{n=1}^{M}\Delta t\left(|u^n|_1^2 + |u_e^n|_1^2\right) \le \frac{1}{2\bar{m}}\left(\|u^0\|_0^2 + \|v^0\|_0^2\right) + \frac{TK}{4\bar{m}}\max_n\left\{\|u^n\|_0^2 + \|v^n\|_0^2\right\} \\ + \frac{C_7}{4\bar{m}}\sum_{n=0}^{M-1}\Delta t\,\|S^n\|_0^2. \qquad (3.16)\end{aligned}$$

This proves that the solutions u, u_e, and v remain bounded in the spaces described in (3.5) for the second order SBDF method, with the condition that $\Delta t < 1/C$ is met. C depends on C_1 and C_2, $C_1 = O(\frac{L_f}{\epsilon})$, and $C_2 = O(\epsilon L_g)$, so we can say that $C = O\left(\max\left\{L_f/\epsilon, \epsilon L_g\right\}\right)$. The stability of the method is guaranteed with $\Delta t = O(\min\{\frac{\epsilon}{L_f}, \frac{1}{\epsilon L_g}\}) = O(\epsilon/L_f)$ when $\epsilon \ll 1$ or $L_f \gg 1$.

3.2.3. Crank–Nicolson/Adams–Bashforth. Take $u^{n+1}+u^n$ and $v^{n+1}+v^n$ as test functions in (2.12) for the Crank–Nicolson/Adams–Bashforth method, and take $\frac{1}{2}(u_e^{n+1}+u_e^n)$ as a test function in (2.2) at times $t=t_n$ and $t=t_{n+1}$. The four resultant equations can then be summed, and by using the Lipschitz condition (3.12) and the equivalence of seminorms we get

$$\begin{aligned}\|u^{n+1}\|_0^2-\|u^n\|_0^2+\|v^{n+1}\|_0^2-\|v^n\|_0^2+\frac{\Delta t}{2}\,\bar{m}(|u^{n+1}+u^n|_1^2+|u_e^{n+1}+u_e^n|_1^2)\\ \le \Delta t\,C_1\|u^{n+1}\|_0^2+\Delta t\,C_2\|v^{n+1}\|_0^2+\Delta t\,C_3\|u^n\|_0^2+\Delta t\,C_4\|v^n\|_0^2\\ +\Delta t\,C_5\|u^{n-1}\|_0^2+\Delta t\,C_6\|v^{n-1}\|_0^2+\Delta t\,C_7\|S^n\|_0^2.\end{aligned}\tag{3.17}$$

By proceeding as we did for the second order SBDF scheme and using the discrete Gronwall lemma,

$$\begin{aligned}\max_{n=1,\dots,M}\left\{\|u^n\|_0^2+\|v^n\|_0^2\right\}\le\Bigg(\frac{1}{1-\Delta t\,C}\left[\|u^0\|_0^2+\|v^0\|_0^2\right]\\ +\frac{C_7}{1-\Delta t\,C}\sum_{n=0}^{M-1}\Delta t\,\|S^n\|_0^2\Bigg)e^{TK/(1-\Delta tC)}\end{aligned}\tag{3.18}$$

and

$$\begin{aligned}\sum_{n=0}^{M-1}\Delta t\big(|u^{n+1}+u^n|_1^2+|u_e^{n+1}+u_e^n|_1^2\big)\le\frac{2}{\bar{m}}\left(\|u^0\|_0^2+\|v^0\|_0^2\right)\\ +\frac{2TK}{\bar{m}}\max_n\left\{\|u^n\|_0^2+\|v^n\|_0^2\right\}+\frac{2C_7}{\bar{m}}\sum_{n=0}^{M-1}\Delta t\,\|S^n\|_0^2,\end{aligned}\tag{3.19}$$

with C and K defined as before and with $\Delta t<1/C$. This time, we need to extend our numerical solutions to functions u, u_e, and v piecewise affine in time, which shows that the solutions u, u_e, and v remain bounded in the spaces in (3.5) with the same conditions as for the second order SBDF scheme.

3.2.4. Forward Euler. The use of explicit methods, such as the forward Euler method considered here, requires more conditions on the time step Δt due to the fact that the diffusion terms are taken explicitly. We use u^{n+1} and v^{n+1} as test functions in (2.12) and u_e^{n+1} in (2.2) at time t_{n+1}, as with the backward Euler and second order SBDF methods, and by adding the resultant variational equations, we get

$$\begin{aligned}\|u^{n+1}\|_0^2+\|u^{n+1}-u^n\|_0^2+\|v^{n+1}\|_0^2+\|v^{n+1}-v^n\|_0^2+2\Delta t\,\bar{m}\left(|u^{n+1}|_1^2+|u_e^{n+1}|_1^2\right)\\ -2\Delta t\,M_i\left[|u^{n+1}-u^n|_1+|u_e^{n+1}-u_e^n|_1\right]|u^{n+1}|_1\\ \le\|u^n\|_0^2+\|v^n\|_0^2+\Delta t\,C_1\|u^{n+1}\|_0^2+\Delta t\,C_2\|v^{n+1}\|_0^2+\Delta t\,C_3\|u^n\|_0^2\\ +\Delta t\,C_4\|v^n\|_0^2+\Delta t\,C_5\|S^n\|_0^2.\end{aligned}\tag{3.20}$$

We must now bound the term with a negative sign in the left-hand side of the previous inequality. Let us rewrite the second variational equation of the system as

$$\begin{aligned}\int_\Omega\left((\sigma_i+\sigma_e)\nabla u_e^{n+1}\right)\cdot\nabla u_e^{n+1}\mathrm{d}\mathbf{x}=-\int_\Omega(\sigma_i\nabla u^{n+1})\cdot\nabla u_e^{n+1}\mathrm{d}\mathbf{x}\\ \Rightarrow|u_e^{n+1}|_1\le\frac{M_i}{m_i+m_e}|u^{n+1}|_1\\ \Rightarrow|u_e^{n+1}-u_e^n|_1\le\frac{M_i}{m_i+m_e}|u^{n+1}-u^n|_1\end{aligned}\tag{3.21}$$

by using the linearity of the second equation of the bidomain model.

By applying an inverse inequality [6, Theorem 3.2.6, p. 140] and the upper bound (3.21) to the inequality (3.20), we obtain the majoration for any constant $s > 0$:

$$\begin{aligned}\|u^{n+1}\|_0^2 + \left(1 - \Delta t \frac{\widetilde{\mathscr{C}}^2 M_i^2 \left(1 + \frac{M_i}{m_i+m_e}\right)^2}{h^2 s}\right) \|u^{n+1} - u^n\|_0^2 + \|v^{n+1}\|_0^2 \\ + \Delta t(2\bar{m} - s)\left(|u^{n+1}|_1^2 + |u_e^{n+1}|_1^2\right) \le \|u^n\|_0^2 + \|v^n\|_0^2 + \Delta t\, C_1 \|u^{n+1}\|_0^2 \\ + \Delta t\, C_2 \|v^{n+1}\|_0^2 + \Delta t\, C_3 \|u^n\|_0^2 + \Delta t\, C_4 \|v^n\|_0^2 + \Delta t\, C_5 \|S^n\|_0^2. \end{aligned} \tag{3.22}$$

To ensure the positivity of all of the terms on the left-hand side of the last inequality, we restrict s and Δt so that $s < 2\bar{m}$ and $1 - \Delta t\, \widetilde{\mathscr{C}}^2 M_i^2 (1 + \frac{M_i}{m_i+m_e})^2 / h^2 s \ge 0$; in other words,

$$\Delta t \le \frac{h^2 s}{\widetilde{\mathscr{C}}^2 M_i^2 \left(1 + \frac{M_i}{m_i+m_e}\right)^2} < \frac{2h^2 \bar{m}}{\widetilde{\mathscr{C}}^2 M_i^2 \left(1 + \frac{M_i}{m_i+m_e}\right)^2}. \tag{3.23}$$

Let $k = 2\bar{m} - s$, $C = \max\{C_1, C_2\}$, and $K = \max\{C_1 + C_3, C_2 + C_4\}$. Proceeding as with the previous methods yields

$$\begin{aligned}\max_{n=1,\dots,M} \left\{\|u^n\|_0^2 + \|v^n\|_0^2\right\} \le \Bigg(\frac{1}{1 - \Delta t\, C}\left[\|u^0\|_0^2 + \|v^0\|_0^2\right] \\ + \frac{C_5}{1 - \Delta t\, C} \sum_{n=0}^{M-1} \Delta t\, \|S^n\|_0^2\Bigg) e^{TK/(1-\Delta t C)}, \end{aligned} \tag{3.24}$$

$$\begin{aligned}\sum_{n=1}^{M} \Delta t \left(|u^n|_1^2 + |u_e^n|_1^2\right) \le \frac{1}{k}\left(\|u^0\|_0^2 + \|v^0\|_0^2\right) + \frac{TK}{k} \max_n \left\{\|u^n\|_0^2 + \|v^n\|_0^2\right\} \\ + \frac{C_5}{k} \sum_{n=0}^{M-1} \Delta t\, \|S^n\|_0^2. \end{aligned} \tag{3.25}$$

Therefore, if the conditions we have imposed on Δt are met, u, u_e, and v remain stable in the spaces of (3.5). These conditions are $\Delta t < 1/C$, with $C = O(\max\{L_f/\epsilon, \epsilon L_g\})$ and the condition (3.23). Clearly Δt must be $O(h^2)$, which limits the applicability of the forward Euler method.

3.2.5. Discussion on the stability conditions. We first want to quantify how small the time step must be to guarantee the stability of fully explicit methods, such as the forward Euler method. We need to estimate the constants in the stability condition (3.23), in particular, $\bar{m}$ and $\widetilde{\mathscr{C}}$.

The last inequality in majoration (3.1) can be rewritten as

$$\begin{aligned} m_i|x+y|^2 + m_e|y|^2 &= m_i|x|^2 + 2m_i\langle x, y\rangle + (m_i + m_e)|y|^2 \\ &= \begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} m_i & m_i \\ m_i & m_i + m_e \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \\ &\ge \bar{m}\left(|x|^2 + |y|^2\right), \end{aligned} \tag{3.26}$$

with $\bar{m} > 0$ the smallest eigenvalue of the above matrix. A simple computation gives

$$\bar{m} = \frac{2m_i + m_e - \sqrt{4m_i^2 + m_e^2}}{2} = O(\min\{m_i, m_e\}). \tag{3.27}$$

Let us now estimate the constant $\widetilde{\mathscr{C}}$ that appears in the inverse inequality in [6, Theorem 3.2.6, p. 140]. We will start by evaluating $\widetilde{\mathscr{C}}$ in the 1D case. Consider the function $v_h(x) = a + (b-a)\,x/h \in P_h^1$ on the element $K = [0,h]$. When evaluating $\|v_h\|_0$ and $|v_h|_1$ on K for that function, one must have for the inverse inequality to be valid that:

$$\widetilde{\mathscr{C}} \geq h\frac{|v_h|_1}{\|v_h\|_0} = \left(\frac{3(a-b)^2}{a^2+ab+b^2}\right)^{1/2}. \tag{3.28}$$

It is easy to find that $\widetilde{\mathscr{C}} = 2\sqrt{3}$ is the lowest upper bound for the above rational function.

Now consider the 2D case. Let K be the linear triangular element adjacent to the nodes at $(0,0)$, $(0,h)$, and $(h/2,h/2)$. This is not the most general case, since $\widetilde{\mathscr{C}}$ is likely to depend on the shape of K, but it is nevertheless representative of the meshes we will work with in section 4. Let v_h be a function in P_h^1 on this element with the property that $v_h(0,0) = a$, $v_h(0,h) = b$, and $v_h(h/2,h/2) = c$, i.e., $v_h(x,y) = a + (2c-(a+b))\,x/h + (b-a)\,y/h$. Now let us compute $\|v_h\|_0$ and $|v_h|_1$ on K:

$$\|v_h\|_0 = \left(\int_K v_h^2\,\mathrm{d}\mathbf{x}\right)^{1/2} = \left(\frac{h^2}{24}\left(a^2+b^2+c^2+ab+ac+bc\right)\right)^{1/2}, \tag{3.29}$$

$$|v_h|_1 = \left(\int_K |\nabla v_h|^2\,\mathrm{d}\mathbf{x}\right)^{1/2} = \left(\frac{a^2}{2}+\frac{b^2}{2}+c^2-ac-bc\right)^{1/2}. \tag{3.30}$$

We must choose a value of $\widetilde{\mathscr{C}}$ such that

$$\widetilde{\mathscr{C}} \geq h\frac{|v_h|_1}{\|v_h\|_0} = \left(\frac{12a^2+12b^2+24c^2-24ac-24bc}{a^2+b^2+c^2+ab+ac+bc}\right)^{1/2}. \tag{3.31}$$

In this case, the optimal value for $\widetilde{\mathscr{C}}$ may be computed to be $6\sqrt{2}$.

We consider two cases for the conductivity values.

Case 1: Suppose that $m_i = m_e$ and $M_i = \lambda m_i$ for a certain value $\lambda > 1$. This is in fact a case where the bidomain model reduces to a monodomain model [7, 14]. By using the constants calculated above, the condition on the time step reads as

$$\begin{aligned}\Delta t &< \frac{(3-\sqrt{5})h^2}{\widetilde{\mathscr{C}}^2\lambda^2 m_i\left(1+\lambda+\frac{\lambda^2}{4}\right)}\\ \Rightarrow \Delta t &< \frac{\widetilde{K}h^2}{\lambda^4 m_i} = \frac{\widetilde{K}m_i^3}{M_i^4}h^2,\end{aligned}$$

where $\widetilde{K}$ is a constant that depends on $1/\widetilde{\mathscr{C}}^2$. In that case,

$$\Delta t = O\left(\min\left\{\frac{\epsilon}{L_f}, \frac{1}{\epsilon L_g}, \frac{m_i^3}{M_i^4}h^2\right\}\right).$$

When h becomes small, the last term is the most stringent with a time step that decreases also with the cube of the ratio $m_i/M_i < 1$ of the minimal to maximal intracellular conductivities on the domain.

TABLE 3.1
Stability conditions on Δt and on the model used for several methods.

Method	Δt	Model
Forward Euler	$O\left(\min\left\{\frac{\epsilon}{L_f}, \frac{m_i^3}{M_i^4}h^2\right\}\right)$	f, g Lipschitz
Second order SBDF	$O\left(\frac{\epsilon}{L_f}\right)$	f, g Lipschitz
Backward Euler	$O(\epsilon)$	FitzHugh–Nagumo

Case 2: Take values that approach the physical values found in the heart, for example, those seen in [12], i.e., $m_i = 0.0263$, $m_e = 0.1087$, and $M_i = 0.263$. By substituting these and the constants computed above in (3.23), this would mean in 1D that $\Delta t < 0.00256\,h^2$ and in 2D that $\Delta t < 0.000167\,h^2$. Not only must the maximal time step Δt for stability be in $O(h^2)$, but in practice the constant appearing in that bound is very small. In other words, an explicit method cannot be efficient, especially in higher dimensions or with fine meshes, due to the excessively large number of time steps needed.

Remark 3.2. Suppose that we change the conductivity over the domain while keeping it proportional to what it was before, i.e., replace the conductivity tensors $\sigma_{i,e}$ by $\mu\,\sigma_{i,e}$, just to see how the critical time step scales with the magnitude of the conductivity. Clearly, from condition (3.23), the time step Δt will have to be taken μ times smaller. Therefore increasing the conductivity of the domain will reduce the stability of the forward Euler method.

Table 3.1 summarizes our stability analysis for the numerical time-stepping schemes considered in this section, by supposing that $\epsilon \ll 1$. The stability of implicit and semi-implicit time-stepping schemes is controlled by the time scale of the ion kinetics without regard to the space step size, while for fully explicit methods the space step size also controls the stability through the second order conduction terms.

4. Numerical results. This section presents numerical tests for both 1D and 2D test cases.

4.1. Comparison of the methods in 1D. We have used in the 1D case a simple finite element code written in Matlab. Our linear systems are solved with the LU factorization included in Matlab. The ease of implementation of a 1D code and the ease of use of Matlab have allowed us to test a significant number of methods.

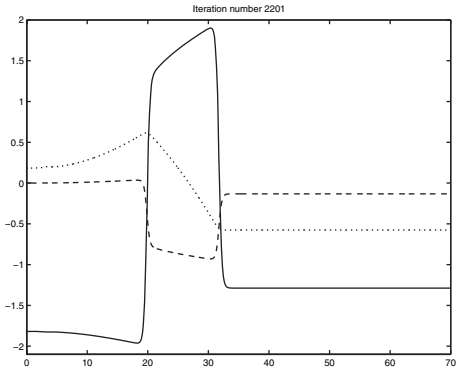
4.1.1. Problem. The 1D bidomain model can be written as

$$\frac{\partial u}{\partial t} = \frac{1}{\epsilon} f(u,v) + \frac{\partial}{\partial x}\left(\sigma_i \frac{\partial u}{\partial x}\right) + \frac{\partial}{\partial x}\left(\sigma_i \frac{\partial u_e}{\partial x}\right), \tag{4.1}$$

$$\frac{\partial}{\partial x}\left(\sigma_i \frac{\partial u}{\partial x}\right) + \frac{\partial}{\partial x}\left((\sigma_i + \sigma_e)\frac{\partial u_e}{\partial x}\right) = 0, \tag{4.2}$$

$$\frac{\partial u}{\partial t} = \epsilon g(u,v), \tag{4.3}$$

where $f(u,v) = u - \frac{u^3}{3} - v$ and $g(u,v) = u + \beta - \gamma v$, with Neumann boundary conditions $\partial_x u(0) = \partial_x u(L) = 0$ and $\partial_x u_e(0) = \partial_x u_e(L) = 0$. Recall that this system determines only u_e modulo a constant. To avoid trying to solve a singular system, we add the condition $u_e(0) = 0$. Our tests have been done over the domain $[0, L]$ with several values of L. We choose as parameter values $\epsilon = 0.1$, $\beta = 1.0$, and $\gamma = 0.5$ and



Solid: u; dashed: u_e; dotted: v

FIG. 4.1. *Solution of the* $1D$ *bidomain model's equations at* $t = 11$.

as conductivity tensors $\sigma_i = \sigma_e = 1.0$. Note that all of the parameter values and space and time measures we will use are nondimensional. For the initial condition, we take u and v constant at the equilibrium value of the system $f(u, v) = 0$ and $g(u, v) = 0$ and u_e constant at 0, except on the interval $[0, \frac{1}{20}L]$ where we fix u at a supercritical value, that is, $u(x) = 2$. This creates a traveling pulse wave that propagates through the domain at a constant speed. Figure 4.1 shows a typical wave form at time $t = 11$. Except where specifically mentioned, all tests in this section are done with $L = 70$, and the final time is $T = 30$.

4.1.2. Results. Given the impossibility of obtaining an exact solution to the problem, we have computed a very precise numerical reference solution (mesh resolution $h = 0.007$ and time step $\Delta t = 10^{-4}$) with the second order SBDF scheme (2.14). It is worth noting that, while it is easy to obtain a numerical solution that qualitatively looks like this reference solution, the numerical error may actually be important. The reason is that a space mesh that is too coarse, or a time step that is too large, often produces a traveling wave whose speed is incorrect. Therefore, we consider the relative error between the speed of the reference and numerical solutions as a good measure of the quality of the solution. We will also consider the L^2 error between both solutions.

Table 4.1 shows the value of the critical time step Δt_c, that is, the largest time step for which the solution remains bounded, for several methods and three space resolutions. We have noticed that, for values of Δt slightly inferior to Δt_c, oscillations often appear in regions where the solution has a sharp gradient, but these oscillations remain bounded and do not grow with time. This is probably due to the nonlinear reaction term, and it might be a loss of monotonicity rather than a loss of stability.

As expected from the theoretical stability analysis, the dependence of Δt_c on the mesh resolution varies between the different methods. The forward Euler method's critical time step seems proportional to h^2, in agreement with the analysis in section 3.2.4. By calculating the theoretical critical time steps for that method using the constants obtained in that section and $m_{i,e} = M_{i,e} = 1$, one gets $\Delta t_{c,th} = 5.55 \times 10^{-4}$, 1.39×10^{-4}, and 3.47×10^{-5} for $h = 0.14$, 0.07, and 0.035 ($N = 500$, 1000, and 2000), respectively. The theoretical critical time steps are about 6 times smaller than the experimental critical time steps, but this is not surprising as the theoretical stability conditions obtained above are only sufficient.

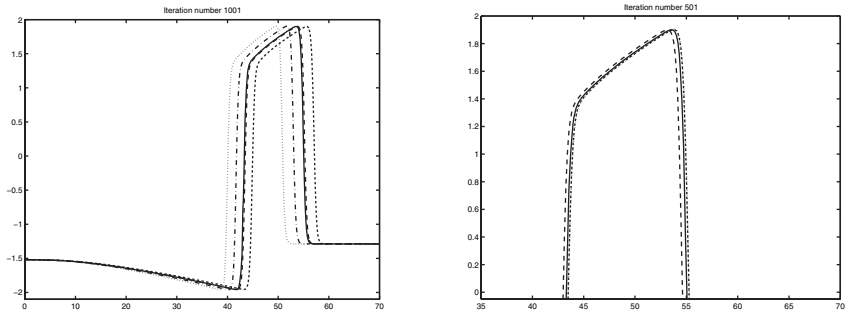
TABLE 4.1
Size of Δt_c for the numerical methods used.

Methods	$h = 0.14$	$h = 0.07$	$h = 0.035$
Forward Euler	0.003206	0.000813	0.000204
Forward-backward Euler	0.1395	0.1422	0.1415
Crank–Nicolson	0.1299	0.1044	0.08386
IMEX first order Gear	0.1594	0.1594	0.1594
Backward Euler	> 0.2	> 0.2	> 0.2
CNAB	0.06772	0.04951	0.04027
MCNAB	0.07538	0.075	0.075
Second order SBDF	0.09375	0.09404	0.09494
Implicit Gear	> 0.25	> 0.25	> 0.25
Third order SBDF	0.06834	0.07026	0.07126

For all of the other methods—the implicit, IMEX first order Gear, MCNAB, and second and third order SBDF methods—the critical time step shows limited dependence, if any, on the space step h. The fully implicit methods require the largest Δt_c, followed by the first order IMEX methods and then the higher order IMEX schemes, which require the smallest Δt_c. For the backward Euler scheme, we estimate that $\Delta t_{c,th} \approx 0.04$ by picking the optimal parameter k to equilibrate the leading order terms in $1/\epsilon$ in the stability condition of section 3.2.1. The theoretical step $\Delta t_{c,th}$ is about 5 times smaller than the experimental step Δt_c. Again the theoretical stability condition is only sufficient, which may explain the discrepancy between both values. For fully implicit methods, we have indicated a lower bound for Δt_c because, as Δt gets larger, the nonlinear system is harder to solve with Newton's method, leading to the stagnation of the iterative solver and making an accurate evaluation of Δt_c difficult. Only two semi-implicit methods required smaller Δt_c with smaller space steps h, namely, the CN and CNAB schemes. For those two methods, the dependence of Δt_c on the mesh resolution is at most of sublinear order, unless the asymptotic behavior of Δt_c when h goes to 0 has not yet been reached even with $h = 0.035$.

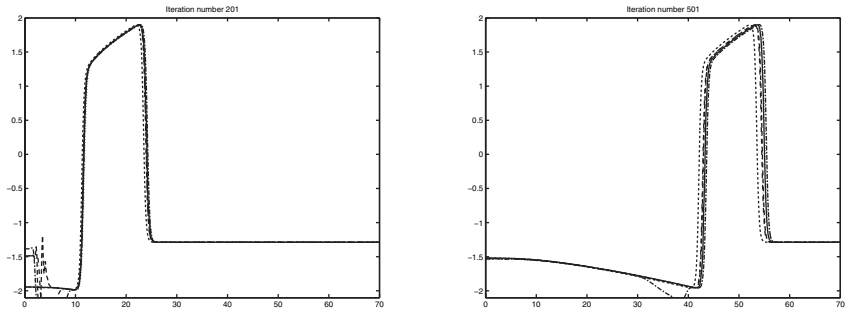
Figure 4.2 shows the position of the u wave at different times for different methods. The two top figures show the position of the waves at time $t = 20$ computed with all first order methods compared to the position of the reference solution. The bottom figure at the right shows the same thing but for second order methods. For second order methods, we used a time step twice as large as for first order methods to have a comparable workload at each time step and globally to reach a solution at a given time t. Clearly, for a comparable computational effort, second order methods offer greater accuracy with less phase lag compared to the reference solution. The figure at the bottom left shows the solutions obtained with the same second order methods at an earlier time. The phase error between the reference and numerical solution grows with time, as a result of the different wave speeds for the different time integration schemes. We will thus use the wave speed as a measure of error.

Table 4.2 shows the signed relative error between the speed of the numerical wave c_{num} and a reference value c_{ref}, where c_{num} is defined as the average speed at which the level curve for $u = 1$ moves on the interval $[25, 50]$. We have computed the traveling wave speed for our reference solution to be $c_{ref} = 2.577444$. For the section of Table 4.2 dealing with the forward Euler method, we have chosen to show the wave speed for a number of time steps nt approximately equal to the number of time steps necessary to ensure the stability and monotonicity of the solution, as well as for 1.5 and 2 times that number of time steps. For all of the other methods, we could use



$t = 20$ $t = 20$, zoom

Solid: reference solution
Dashed: CNAB ($\Delta t = 0.02$)
Dash-dot: Crank–Nicolson ($\Delta t = 0.02$)
Dotted: IMEX Gear ($\Delta t = 0.02$)
Short-dashed: Backward Euler ($\Delta t = 0.02$)
Dashed: CNAB ($\Delta t = 0.04$)
Dotted: Implicit Gear ($\Delta t = 0.08$)
Short-dashed: Forward Euler ($\Delta t = 0.0025$)



$t = 8$ $t = 20$

Solid: reference solution
Dashed: CNAB ($\Delta t = 0.04$)
Dash-dot: MCNAB ($\Delta t = 0.04$)
Short-dashed: 2^{nd} order SBDF ($\Delta t = 0.04$)
Dash-dash: 3^{rd} order SBDF ($\Delta t = 0.04$)
Dotted: Implicit Gear ($\Delta t = 0.08$)

FIG. 4.2. *Position of the u wave at different t for several methods ($h = 0.14$).*

the same standardized time steps, sufficiently small to ensure the stability of all of the methods.

Table 4.2 clearly shows that with an underresolved grid one cannot reliably compute waves with the right speed, irrespective of the time integration scheme used. It is only for grids with $h \leq 0.14$ that the wave speed error is below 1 %, at least for higher order methods. It had already been observed in [5] that, with a second order fully implicit Gear scheme, 2D action potential waves could travel with the wrong speed due to underresolved spatial grids. Here we emphasize that the grid resolution is only one aspect of the problem; the order of the time integration scheme matters. Finer grids give better accuracy with all methods, but the semi-implicit or implicit first order schemes (Crank–Nicolson, IMEX Gear, and forward-backward Euler in the table) can hardly provide a wave speed error below 1% even on the finest grids considered. It is only by reducing the time step to values comparable to what is used for the forward Euler method that these first order schemes can provide accurate results, at the price of a much higher workload to reach a solution at a given time than for second order methods. Indeed the forward Euler scheme gave the most accurate wave

TABLE 4.2
Relative error (%) between c_{ref} and c_{num}.

Crank–Nicolson/Adams–Bashforth					
	$h = 0.7$	$h = 0.35$	$h = 0.14$	$h = 0.07$	$h = 0.035$
$\Delta t = 0.04$	9.702	1.567	−1.041	−1.591	−1.605
$\Delta t = 0.02$	10.43	2.406	−0.02937	−0.1765	−0.3810
$\Delta t = 0.01$	10.43	2.512	0.07294	−0.07387	−0.07387
$\Delta t = 0.005$	10.36	2.512	0.1242	−0.02250	−0.07387

Modified Crank–Nicolson/Adams–Bashforth					
	$h = 0.7$	$h = 0.35$	$h = 0.14$	$h = 0.07$	$h = 0.035$
$\Delta t = 0.04$	9.702	1.152	−1.440	−1.866	−1.866
$\Delta t = 0.02$	10.43	2.195	−0.02937	−0.3810	−0.4553
$\Delta t = 0.01$	10.43	2.512	0.07294	−0.07387	−0.07387
$\Delta t = 0.005$	10.43	2.565	0.1242	−0.02250	−0.07387

Second order SBDF					
	$h = 0.7$	$h = 0.35$	$h = 0.14$	$h = 0.07$	$h = 0.035$
$\Delta t = 0.04$	8.284	−0.4727	−3.005	−3.544	−3.544
$\Delta t = 0.02$	9.942	1.985	−0.6389	−0.7876	−0.7876
$\Delta t = 0.01$	10.30	2.406	−0.02937	−0.1765	−0.1765
$\Delta t = 0.005$	10.43	2.565	0.1242	−0.07387	−0.07387

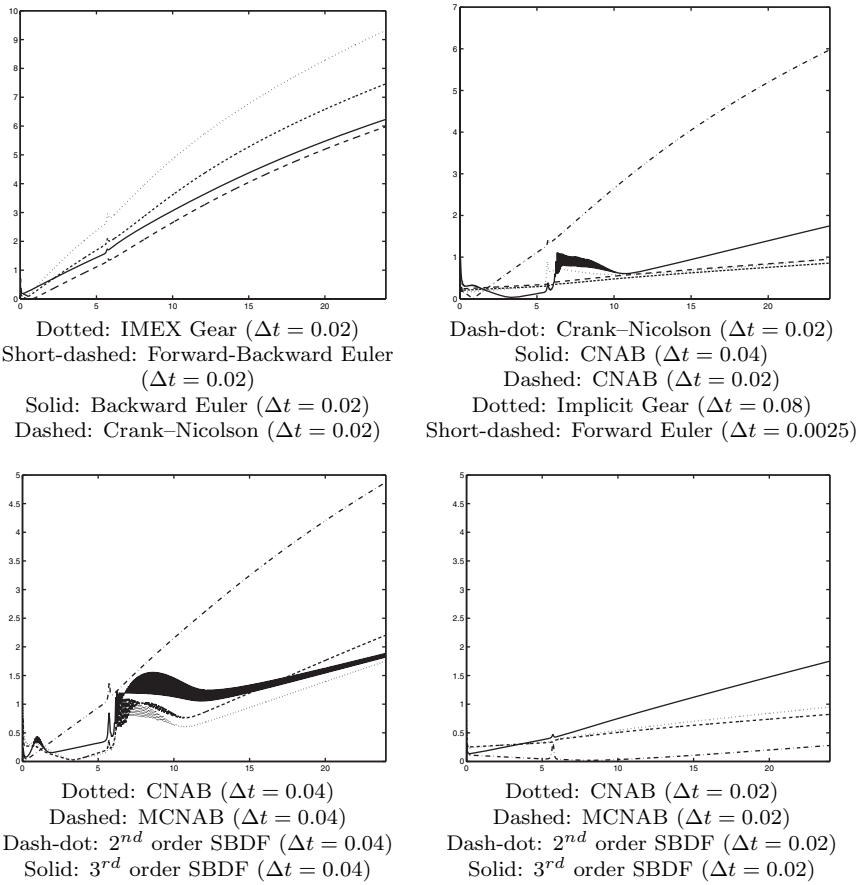
Third order SBDF					
	$h = 0.7$	$h = 0.35$	$h = 0.14$	$h = 0.07$	$h = 0.035$
$\Delta t = 0.04$	11.16	3.260	1.003	0.1617	0.1617
$\Delta t = 0.02$	10.67	2.831	0.3812	0.02890	0.09517
$\Delta t = 0.01$	10.55	2.618	0.1755	0.02890	0.02890
$\Delta t = 0.005$	10.43	2.565	0.1755	−0.02250	−0.02250

Crank–Nicolson					
	$h = 0.7$	$h = 0.35$	$h = 0.14$	$h = 0.07$	$h = 0.035$
$\Delta t = 0.04$	2.121	−6.128	−8.454	−8.620	−8.748
$\Delta t = 0.02$	5.999	−1.852	−4.146	−4.491	−4.491
$\Delta t = 0.01$	8.167	0.2318	−2.131	−2.283	−2.381
$\Delta t = 0.005$	9.225	1.411	−0.9911	−1.241	−1.241

IMEX Gear					
	$h = 0.7$	$h = 0.35$	$h = 0.14$	$h = 0.07$	$h = 0.035$
$\Delta t = 0.04$	−5.201	−13.35	−15.13	−15.54	−15.54
$\Delta t = 0.02$	1.708	−6.305	−8.454	−8.620	−8.620
$\Delta t = 0.01$	5.776	−2.046	−4.520	−4.678	−4.678
$\Delta t = 0.005$	7.993	0.1811	−2.229	−2.430	−2.430

Forward-backward Euler					
	$h = 0.7$	$h = 0.35$	$h = 0.14$	$h = 0.07$	$h = 0.035$
$\Delta t = 0.04$	−1.483	−8.889	−11.12	−10.96	−11.21
$\Delta t = 0.02$	4.240	−3.571	−5.808	−5.969	−6.100
$\Delta t = 0.01$	7.243	−0.6721	−3.005	−3.159	−3.159
$\Delta t = 0.005$	8.752	0.8949	−1.490	−1.690	−1.690

Forward Euler					
	$h = 0.7$	$h = 0.35$	$h = 0.14$	$h = 0.07$	$h = 0.035$
nt	770	1900	10000	37300	148000
Δt	0.0390	0.0158	0.003	0.000804	0.000203
% (nt)	−0.03216	0.2905	−0.3046	−0.2089	−0.1023
% $(1.5\,nt)$	1.542	0.8274	−0.1723	−0.1704	−0.09331
% $(2\,nt)$	2.348	1.179	−0.09067	−0.1471	−0.08881



Dotted: IMEX Gear ($\Delta t = 0.02$)
Short-dashed: Forward-Backward Euler ($\Delta t = 0.02$)
Solid: Backward Euler ($\Delta t = 0.02$)
Dashed: Crank–Nicolson ($\Delta t = 0.02$)

Dash-dot: Crank–Nicolson ($\Delta t = 0.02$)
Solid: CNAB ($\Delta t = 0.04$)
Dashed: CNAB ($\Delta t = 0.02$)
Dotted: Implicit Gear ($\Delta t = 0.08$)
Short-dashed: Forward Euler ($\Delta t = 0.0025$)

Dotted: CNAB ($\Delta t = 0.04$)
Dashed: MCNAB ($\Delta t = 0.04$)
Dash-dot: 2^{nd} order SBDF ($\Delta t = 0.04$)
Solid: 3^{rd} order SBDF ($\Delta t = 0.04$)

Dotted: CNAB ($\Delta t = 0.02$)
Dashed: MCNAB ($\Delta t = 0.02$)
Dash-dot: 2^{nd} order SBDF ($\Delta t = 0.02$)
Solid: 3^{rd} order SBDF ($\Delta t = 0.02$)

FIG. 4.3. *Evolution in time of the L^2 error for several methods ($h = 0.14$).*

speed on coarse grids and nearly as accurate results as second order schemes on fine grids, but this was done with a prohibitive number of time steps.

Except for circumstantial very accurate results (as with CNAB using $h = 0.07$ and $\Delta t = 0.005$), the most accurate wave speeds have been obtained with the third order SBDF scheme on fine enough grids. It is worth noting, however, that using methods of third order and over isn't necessarily an advantage. The effort in implementing a third order scheme is larger for a marginal gain in accuracy compared to second order schemes. The third order SBDF scheme was also subject to much more stringent numerical stability conditions than its companion second order scheme (see Table 4.1), and even when using a time step size below Δt_c the third order SBDF scheme was more subject to numerical artifacts (Figure 4.2, bottom right). We will come back to this point below when recommending methods.

Finally, Figure 4.3 shows the progression in time of the L^2 error between the reference and numerical solutions for $h = 0.14$. The top two images compare the first order methods to each other and to the second order methods. The two bottom images compare the second and third order methods for two values of Δt. We see that, after a certain time interval in which the action potential wave takes form, the L^2 error then seems to grow linearly in time, although with different slopes for different methods. Among the first order methods, the Crank–Nicolson scheme is the

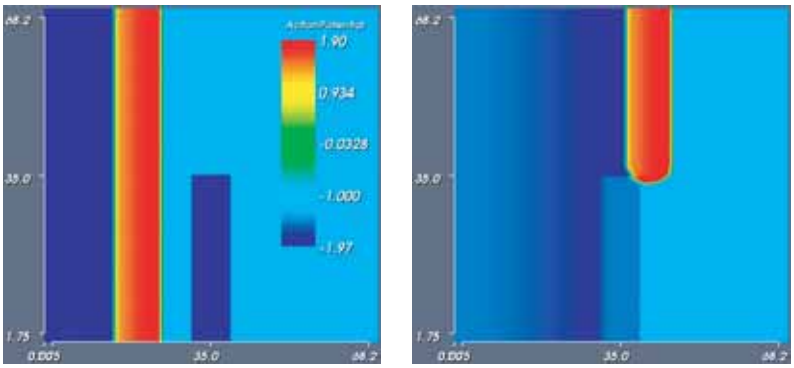
most accurate. Figure 4.3 confirms the conclusion reached with Table 4.2 that higher order methods are necessary to solve this problem. The error is consistently lower with higher order than first order methods, except for the forward Euler method, which produces results with a L^2 error about the size of the CNAB and implicit Gear schemes but with a much smaller time step than for these second order methods.

Looking back at Figure 4.2 at the bottom, small oscillations are observed in the recovery zone behind the repolarization front for the third order SBDF method. Note that these oscillations are also seen with the CNAB and MCNAB schemes, but they get smoothed out with the advance of time. The third order SBDF scheme gives a bounded solution, but more work must be done to make the instabilities disappear. This method, in fact, does not give better results when $h = 0.14$, as illustrated in Figure 4.3. We may therefore conclude that second order methods are in fact the optimal choice. Among these, we will concentrate on the IMEX schemes, due to the large cost of every time step with the implicit Gear scheme. The second order SBDF method appears to be the most interesting one. As seen in Table 4.1, it is more stable than the CNAB and MCNAB methods—which in fact produce very similar solutions. While it is true (see Figure 4.2 and Table 4.2) that for larger time steps (for example, $\Delta t = 0.04$) the second order SBDF does not preserve the wave speed quite as well as the other two, when smaller time steps are used, it produces an as-good solution. This method also has the advantage of requiring the computation of the diffusion term only at time step $n+1$. For all of these reasons, the second order SBDF method is recommended for solving the bidomain model.

4.2. Comparison of the methods in 2D. In order to solve the 2D bidomain model, we have used an object-oriented finite element code written in C++ based on the solver library PETSc [27]. This library allows the user to choose between many direct and iterative solvers and between many preconditioners. The tests we did on coarse grids have convinced us that the LU factorization is ideal for this problem, subject to memory constraints. When finer grids are used, iterative methods become necessary. Our tests have shown that, of all available solvers and preconditioners, the conjugate gradient method with an incomplete LU preconditioner (see [24]) use the least computation time. The tests shown in this section have been done with this solver and the incomplete LU preconditioner with two levels of fill.

4.2.1. Problem. We consider the 2D formulation of the bidomain model, including Neumann boundary conditions. The domain $[0, 70] \times [0, 70]$ is used—these values again being nondimensional. It is meshed with $4N^2$ triangular elements obtained by dividing the $N \times N$ squares into four congruent triangles. For all of our tests, $N = 300$ ($h = 0.233$). As in the 1D case, we fix the parameters $\epsilon = 0.1$, $\beta = 1.0$, and $\gamma = 0.5$. The degeneracy of the system is removed by fixing the value of u_e somewhere on the domain; we have chosen $u_e(0, 0) = 0$.

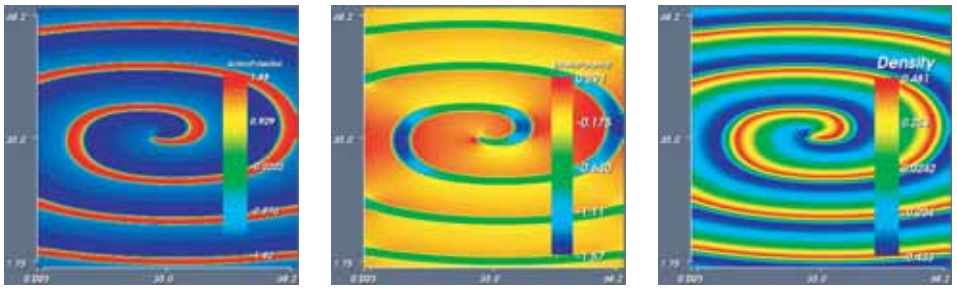
The initial condition used in all comparisons done below is obtained numerically as follows. We initiate a spiral (reentrant) wave by setting u and v at their equilibrium value and $u_e = 0$, except on $[0, 3.5] \times [0, 70]$, where we fix $u = 2$ as explained in section 4.1.1, and also on $[31, 39] \times [0, 35]$, where we fix $v = 2$, which creates a recovery zone. The result is a planar wave that travels at constant speed until hitting the recovery zone and then rolls around this zone, forming a spiral wave. Figure 4.4 illustrates the way this spiral wave is formed for the variable u. That spiral wave is then evolved during enough time steps to go through a transient initial stage and reach an established (unsteady) regime. The solution at a given time in that established regime is used as an initial condition. The initial condition is shown on Figure 4.5.



$t = 10$ $t = 20$

$t = 30$ $t = 40$

FIG. 4.4. *Solution for u of the 2D bidomain equations at different times.*



u u_e v

FIG. 4.5. *Initial condition for the 2D problem.*

Due to time and memory constraints, it is very difficult to obtain a precise reference solution in 2D, as was done with 10,000 grid points in 1D. Instead of increasing the number of grid points to a very large value in each direction to reach the grid independence of the numerical solution, we chose to use only $N = 300$ but pick values of the conductivity tensors that are large enough to avoid mesh effects by ensuring a thick enough de-/repolarization wave front. The conductivity tensors are thus taken as $\sigma_i = 25\,\mathrm{diag}\{0.263, 0.0263\}$ and $\sigma_e = 25\,\mathrm{diag}\{0.263, 0.1087\}$. These tensors

TABLE 4.3
Vectors used by different numerical schemes.

	Time derivative		Reaction term		Diffusion term
	t_n	t_{n-1}	t_n	t_{n-1}	t_n
CNAB	X		X	X	X
SBDF2	X	X	X	X	
CN	X		X		X

are proportional to the values suggested by Hooks et al. [12] for the human heart conductivities.

4.2.2. Results. We compare the second order SBDF and the CNAB methods, given the good results they gave in the 1D tests, as well as the Crank–Nicolson–forward Euler method. We use as a comparison point between the different methods the trajectory of the spiral tip. The spiral tip is defined as the point where the level curves with value 0 for u and v meet. We will check if the form of the trajectory of the spiral tip is the same for the numerical solutions and for the reference solution and if there is a drift of the spiral tip.

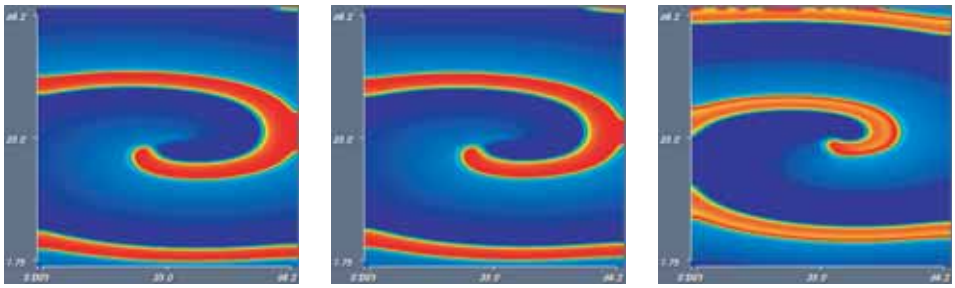
To speed up computations at each time step, we compute and store the linear system and preconditioner matrices before the first time step and leave them unchanged afterwards. Note that we solve all three equations of the system simultaneously, resulting in matrices with $3 \times 4N^2$ rows for the grid used. These matrices involve the submatrices M, A_i, and A_e defined in section 2.1. Another strategy to reduce the memory requirements would be to solve each equation of the bidomain model separately and do subiterations to solve the coupled linear system at each time step.

The three methods that we have tested require only the two matrices mentioned above. However, they differ in the number of vectors that need assembling at each time step. Table 4.3 shows which vectors are used by each of these three schemes. We suppose that the next time step is the one that will compute the solution at time t_{n+1}. This table shows that the second order SBDF method forces an extra vector—the solution at time t_{n-1}—to be kept in memory but also approximates the diffusion term only at time t_{n+1}, which is included in the linear system matrix, while the other methods require computing the diffusion term at time t_n, needing reassembly at each time step. The assembly of the right-hand term is therefore more costly in terms of computation time for the Crank–Nicolson/Adams–Bashforth and Crank–Nicolson–forward Euler methods than for the second order SBDF method.

We have tested each second order method with three different time steps: $\Delta t = \frac{1}{50}$, $\Delta t = \frac{1}{30}$, and Δt near Δt_c. The Crank–Nicolson method has been tested only with $\Delta t = \frac{1}{50}$ and $\Delta t = \frac{1}{30}$. We have found our values of Δt close to Δt_c with tests with decreasing values of Δt until the method becomes stable. We are thus able to say that, for the initial solution shown in Figure 4.5, $\frac{1}{12} > \Delta t_c \geq \frac{1}{14}$ for the second order SBDF method and $\frac{1}{17} > \Delta t_c \geq \frac{1}{18}$ for the CNAB method. We therefore choose $\Delta t = \frac{1}{18}$ with CNAB and $\Delta t = \frac{1}{14}$ with SBDF. We notice that, just as in the 1D case (see Table 4.1), the SBDF method is slightly more stable than the CNAB method.

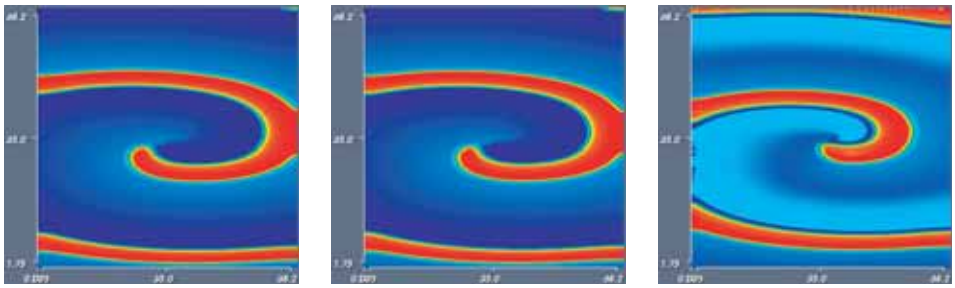
We first note that our largest values of Δt are not sufficient to ensure that the solution does not contain oscillations. This is obvious on the third image at the right on the first two lines of Figure 4.6. We already noticed in 1D that the potential wave lags behind when Δt is too large, for both the second order SBDF and CNAB methods (see Table 4.2). We can see in Figure 4.6 that this property remains true in 2D: the solutions for the largest values of Δt do not seem as advanced as for smaller Δt. We

2^{nd} order SBDF method



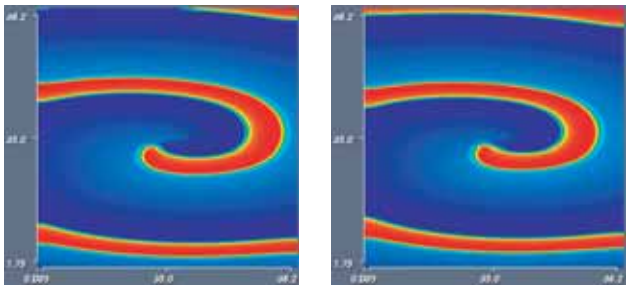
$\Delta t = \frac{1}{50}$ $\Delta t = \frac{1}{30}$ $\Delta t = \frac{1}{14}$

Crank–Nicolson/Adams–Bashforth method



$\Delta t = \frac{1}{50}$ $\Delta t = \frac{1}{30}$ $\Delta t = \frac{1}{18}$

Crank–Nicolson–Forward Euler method



$\Delta t = \frac{1}{50}$ $\Delta t = \frac{1}{30}$

FIG. 4.6. *Solution for u at $t = 80$ with several methods. The legend is the same as for Figure* 4.5.

also remark that the spiral tip drift is very important when Δt is too large, as shown in Figure 4.7(a). The dashed and short-dashed curves, respectively, CNAB with $\Delta t = \frac{1}{18}$ and SBDF2 with $\Delta t = \frac{1}{14}$, get further and further from the solid curve (CNAB with $\Delta t = \frac{1}{50}$) as the solution advances in time. The foot of the trajectory is found at the left in Figures 4.7(a)–(c), while for Figure 4.7(d) it is in the upper right corner.

As in the 1D case, we have been able to determine that first order methods such as the Crank–Nicolson–forward Euler method (2.9) are not really appropriate to solve this particular problem. Even when $\Delta t = \frac{1}{50}$, as seen in the first image of the third

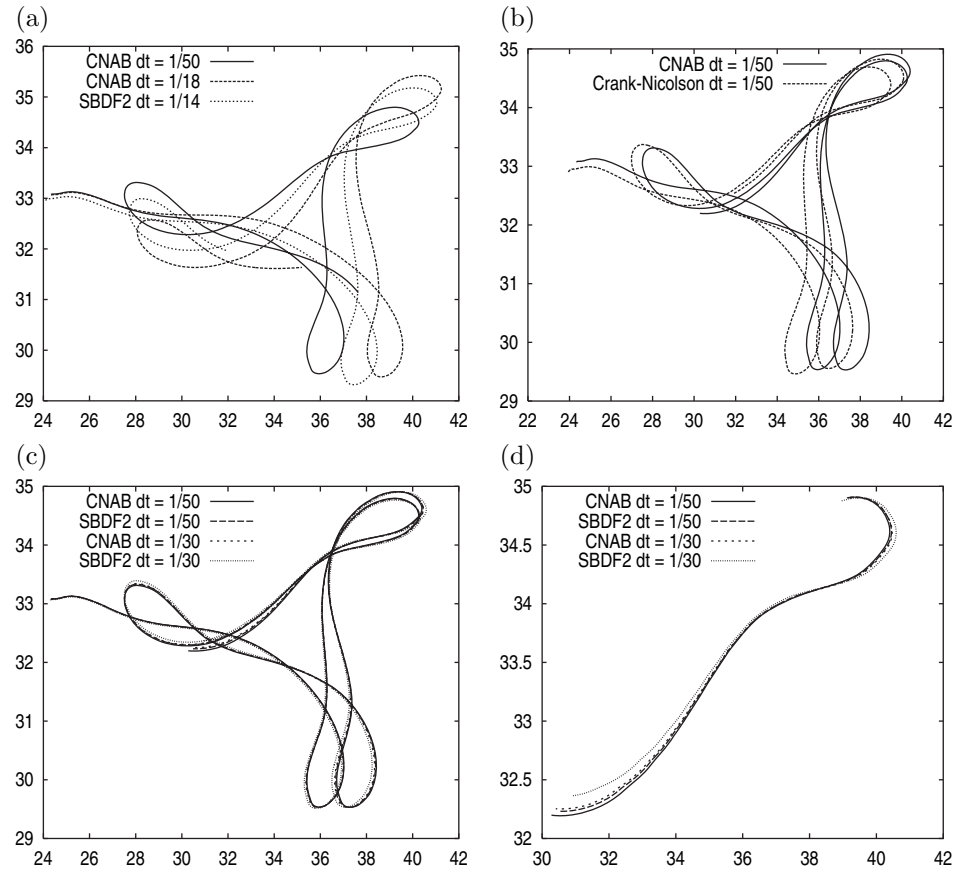


FIG. 4.7. *Trajectory of the spiral tip for:* (a) *second order methods and different values of* Δt, $t \in [10, 55]$*;* (b) *CNAB and CN for* $\Delta t = 1/50$, $t \in [10, 80]$*;* (c) *second order methods and different values of* Δt, $t \in [10, 80]$*;* (d) *second order methods and different values of* Δt, $t \in [70, 80]$.

row of Figure 4.6, the solution already shows an important phase error when compared with the second order methods with $\Delta t = \frac{1}{50}$ or even $\frac{1}{30}$. The error is also visible in Figure 4.7(b), which shows the trajectory of the spiral tip for Crank–Nicolson and CNAB with $\Delta t = \frac{1}{50}$. The error on the spiral tip has already shown up in the 10 time units preceding this image and then increases rapidly.

Figures 4.7(c)–(d) show the evolution of the spiral tip for both second order methods tested, with $\Delta t = \frac{1}{50}$ and $\Delta t = \frac{1}{30}$. The solutions are quite similar regardless of which method is used, but we can note that the solution obtained with the second order SBDF method with $\Delta t = \frac{1}{30}$ drifts slightly compared with the other more accurate results. This fact is more easily seen in Figure 4.7(d). On the other hand, the solution obtained with the CNAB method and $\Delta t = \frac{1}{30}$ follows very closely the more accurate solutions obtained with $\Delta t = \frac{1}{50}$. We recall that in our 1D tests the second order SBDF method produced a wave slightly lagging behind, while the wave computed with the CNAB method was more precisely located for larger values of Δt. Each time step is significantly more costly when solving the 2D problem than the 1D problem, so it is a good idea to use a method that tolerates larger values of Δt. Therefore, the Crank–Nicolson/Adams–Bashforth method seems a more optimal

choice when compared with the second order SBDF method. That being said, the second order SBDF method is almost as accurate as the CNAB method and presents the advantage, as already stated, that each time step is less costly than with the CNAB method, and it is likely that, just as it was with our 1D results, with values of Δt of the order of $\frac{1}{50}$ or less, the second order SBDF method would produce a better solution.

5. Conclusion. In this paper, we have proposed and studied several implicit, explicit, and IMEX time-stepping methods to solve the bidomain model. We have noted that the critical time steps for stability obtained through numerical tests, as seen in Table 4.1, are consistent with the stability analysis of section 3. Indeed, the numerical value of Δt_c depends on h^2 for the forward Euler method, while it has little or no dependence on h for the IMEX and implicit methods. We reached the conclusion that higher order IMEX methods are the most appropriate for solving the bidomain model. Indeed, they are quite stable, their stability does not depend on h, and they produce a good numerical solution even when the time step Δt is not so small. As shown with our numerical results in section 4.1.2, first order schemes, such as the Euler or the Crank–Nicolson methods commonly used for the bidomain model, provide the required accuracy only at the expense of very small time steps, while the second order schemes studied easily provided an error (well) below 1% on the wave speed. It is certainly one of the main conclusions of our work: a second order time-stepping scheme is required, and the Crank–Nicolson scheme the way it is usually implemented for the bidomain model with a fully explicit occurrence of the reaction term does not provide second order accuracy.

Another main advantage of these IMEX methods is that they also require only the solution of a linear system of equations, with both the system and preconditioner matrices constant over the time steps. Second order fully implicit methods are available and slightly more stable, but the matrices involved must be reassembled and factored during Newton–Raphson iterations at each time step.

We have also been able to discriminate between the higher order IMEX methods considered. Two of them have been especially impressive: the CNAB method and the second order SBDF method. We reached the conclusion that, while the SBDF method is more stable than the CNAB method, both methods require a time step about half of its critical value to produce an acceptable solution. With such a value of Δt, the SBDF method produces a solution whose delay compared to the reference solution is slightly larger than the solution obtained with the CNAB method. With yet smaller values of Δt the SBDF method does produce solutions closer to the reference solution than the CNAB method. However, the SBDF method also costs less in terms of computation time than the CNAB method, since it does not require reassembling the diffusion term every time step.

REFERENCES

[1] G. Akrivis, M. Crouzeix, and C. Makridakis, *Implicit-explicit multistep finite element methods for nonlinear parabolic problems*, Math. Comp., 67 (1998), pp. 457–477.

[2] U. M. Ascher, S. J. Ruuth, and B. T. R. Wetton, *Implicit-explicit methods for time-dependent partial differential equations*, SIAM J. Numer. Anal., 32 (1995), pp. 797–823.

[3] P. Bochev and R. B. Lehoucq, *On the finite element solution of the pure Neumann problem*, SIAM Rev., 47 (2005), pp. 50–66.

[4] Y. Bourgault, Y. Coudière, and C. Pierre, *Existence and uniqueness of the solution for the bidomain model used in cardiac electrophysiology*, Nonlinear Anal. Real World Appl., to appear.

[5] Y. Bourgault, M. Ethier, and V. G. LeBlanc, *Simulation of electrophysiological waves with an unstructured finite element method*, Math. Model. Numer. Anal., 37 (2003), pp. 649–661.
[6] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, Stud. Math. Appl. 4, North–Holland, Amsterdam, 1978.
[7] P. C. Franzone and G. Savaré, *Degenerate evolution systems modeling the cardiac electric field at micro and macroscopic level*, in Progr. Nonlinear Differential Equations Appl., 50, Birkhauser, Boston, Cambridge, MA, 2002, pp. 49–78.
[8] A. Ern and J.-L. Guermond, *Theory and Practice of Finite Elements*, Appl. Math. Sci. 159, Springer-Verlag, New York, 2004.
[9] P. C. Franzone and L. F. Pavarino, *A parallel solver for reaction-diffusion systems in computational electrocardiology*, Math. Models Methods Appl. Sci., 14 (2004), pp. 883–912.
[10] A. Greenbaum, *Iterative Methods for Solving Linear Systems*, SIAM, Philadelphia, 1997.
[11] N. Hooke, C. S. Henriquez, P. Lanzkron, and D. Rose, *Linear algebraic transformations of the bidomain equations: Implications for numerical methods*, Math. Biosci., 120 (1994), pp. 127–145.
[12] D. A. Hooks, K. A. Tomlinson, S. G. Marsden, I. J. LeGrice, B. H. Smaill, A. J. Pullan, and P. J. Hunter, *Cardiac microstructure: Implications for electrical propagation and defibrillation in the heart*, Circulation Res., 91 (2002), pp. 331–338.
[13] J. P. Keener and K. Bogar, *A numerical method for the solution of the bidomain equations in cardiac tissue*, Chaos, 8 (1998), pp. 234–241.
[14] J. Keener and J. Sneyd, *Mathematical Physiology*, Interdiscip. Appl. Math. 8, Springer-Verlag, New York, 1998.
[15] G. T. Lines, M. L. Buist, P. Grottum, A. J. Pullan, J. Sundnes, and A. Tveito, *Mathematical models and numerical methods for the forward problem in cardiac electrophysiology*, Comput. Vis. Sci., 5 (2003), pp. 215–239.
[16] G. T. Lines, P. Grottum, and A. Tveito, *Modeling the electrical activity of the heart: A bidomain model of the ventricles embedded in a torso*, Comput. Vis. Sci., 5 (2003), pp. 195–213.
[17] M. Murillo and X.-C. Cai, *A fully implicit parallel algorithm for simulating the non-linear electrical activity of the heart*, Numer. Linear Algebra Appl., 11 (2004), pp. 261–277.
[18] R. C. Penland, D. M. Harrild, and C. S. Henriquez, *Modeling impulse propagation and extracellular potential distribution in anisotropic cardiac tissue using a finite volume element discretization*, Comput. Vis. Sci., 4 (2002), pp. 215–226.
[19] M. Pennacchio and V. Simoncini, *Efficient algebraic solution of reaction-diffusion systems for the cardiac excitation process*, J. Comput. Appl. Math., 145 (2002), pp. 49–70.
[20] A. Quarteroni, R. Sacco, and F. Saleri, *Numerical Mathematics*, Texts Appl. Math. 37, Springer-Verlag, New York, 2000.
[21] B. J. Roth, *Frequency locking of meandering spiral waves in cardiac tissue*, Phys. Rev. E, 57 (1998), pp. R3735–R3738.
[22] B. J. Roth, *Meandering of spiral waves in anisotropic cardiac tissue*, Phys. D, 150 (2001), pp. 127–136.
[23] S. J. Ruuth, *Implicit-explicit methods for reaction-diffusion problems in pattern formation*, J. Math. Biol., 34 (1995), pp. 148–176.
[24] Y. Saad, *Iterative Methods for Sparse Linear Systems*, available online from http://www-users.cs.umn.edu/∼saad/books.html, 2000.
[25] R. W. D. Santos, G. Plank, S. Bauer, and E. J. Vigmond, *Preconditioning techniques for the bidomain equations*, in Proceedings of the 15th International Conference on Domain Decomposition Methods, Lect. Notes Comput. Sci. Eng. 40, Springer-Verlag, New York, 2003, pp. 571–580.
[26] J. Sundnes, G. T. Lines, and A. Tveito, *An operator splitting method for solving the bidomain equations coupled to a volume conductor model for the torso*, Math. Biosci., 2 (2005), pp. 233–248.
[27] See the documentation at http://www-unix.mcs.anl.gov/petsc/petsc-2/.

SIAM J. NUMER. ANAL.
Vol. 46, No. 5, pp. 2469–2483
© 2008 Society for Industrial and Applied Mathematics

CONVERGENCE OF A FINITE ELEMENT-BASED SPACE-TIME DISCRETIZATION IN ELASTODYNAMICS*

ANDREAS PROHL†

Abstract. We study a finite element-based space-time discretization of the elastodynamics equation $u^\varepsilon_{tt} - \operatorname{div}\sigma(\nabla u^\varepsilon) - \varepsilon\Delta u^\varepsilon_t = 0$ for $\varepsilon \geq 0$, where $\sigma = D\phi$ and ϕ is a nonconvex function. The convergence for regularization parameters $\varepsilon > 0$ of iterates to weak solutions and for the limiting problem $\varepsilon = 0$ of iterates towards generalized solutions is shown in a general setting of data. Computational experiments are included to motivate formation and propagation of two-dimensional microstructures for decreasing values of the regularization parameter.

Key words. convergence analysis, finite element method, elastodynamics

AMS subject classifications. 65N12, 65N15, 35G25, 73G25

DOI. 10.1137/070685166

1. Introduction.

We show the convergence of a fully discrete scheme to approximate generalized solutions $u^\varepsilon : \Omega_T \to \mathbb{R}^m$ of the nonlinear hyperbolic system

$$u^\varepsilon_{tt} - \operatorname{div}\sigma(\nabla u^\varepsilon) - \varepsilon\Delta u^\varepsilon_t = 0 \qquad \text{in } \Omega_T := (0,T)\times\Omega, \tag{1.1}$$

$$u^\varepsilon = 0 \qquad \text{on } \partial\Omega_T := (0,T)\times\partial\Omega, \tag{1.2}$$

$$u^\varepsilon(0,\cdot) = u_0, \quad u^\varepsilon_t(0,\cdot) = v_0 \qquad \text{on } \Omega \tag{1.3}$$

for $\varepsilon \geq 0$. Here, $\Omega \subset \mathbb{R}^n$ is a bounded Lipschitz domain, and $T > 0$. The stress $\sigma = D\phi$ is a nonmonotone function, with nonconvex potential $\phi : \mathbb{R}^{m\times n} \to \mathbb{R}$.

Problem (1.1)–(1.3) has been studied in several analytical contributions [1, 5, 6, 7, 8, 9, 10, 12, 13, 15, 16, 17]. The existence of weak solutions for $\varepsilon > 0$ has been verified in [8], where the implicit Euler method is used to semidiscretize the problem in time: Let $I_k = \{t_j\}_{j=1}^J$ be an equidistant net of mesh width $k = t_j - t_{j-1}$ to discretize $[0,T]$, with $j \geq 1$. We set

$$u^{-1}_\varepsilon := u_0 - kv_0 \qquad \text{and} \qquad u^0_\varepsilon := u_0. \tag{1.4}$$

For $j \geq 1$, and given $u^{j-i}_\varepsilon : \Omega \to \mathbb{R}^m$, for $i = 1, 2$, then $u^j_\varepsilon : \Omega \to \mathbb{R}^m$ solves

$$u^j_\varepsilon - 2u^{j-1}_\varepsilon + u^{j-2}_\varepsilon - k^2 \operatorname{div}\sigma(\nabla u^j_\varepsilon) - \varepsilon k\,\Delta\big[u^j_\varepsilon - u^{j-1}_\varepsilon\big] = 0 \quad \text{in } \Omega, \tag{1.5}$$

$$u^j_\varepsilon = 0 \quad \text{on } \partial\Omega. \tag{1.6}$$

Provided that ϕ satisfies certain smoothness and asymptotic growth assumptions, together with the Andrews–Ball condition (H3) given below, weak solutions of (1.1)–(1.3) may be constructed as proper limits of existing solutions to the semidiscretized problem. The uniqueness of weak solutions in the case of globally Lipschitz-continuous σ is also verified in [8]. Qualitative studies, including long-time dynamics, are performed in [9, 10, 1].

*Received by the editors March 13, 2007; accepted for publication (in revised form) February 4, 2008; published electronically June 11, 2008.
http://www.siam.org/journals/sinum/46-5/68516.html
†Mathematisches Institut, Universität Tübingen, Auf der Morgenstelle 10, D-72076 Tübingen, Germany (prohl@na.uni-tuebingen.de).

The existence of generalized solutions to (1.1)–(1.3) in the case $\varepsilon = 0$ is studied in [5, 6, 12]. In the first references, a time discretization by the implicit Euler method is again employed to construct Young-measure-valued solutions for every t_j ($j \in \mathbb{N}$), using relaxation theory. This direct approach, however, requires convexity of the relaxed energy to validate the crucial energy decay property for corresponding first moments of the Young-measure-valued solutions, to afterwards identify limits for $k \to 0$ as generalized solutions of (1.1)–(1.3). A different approach is proposed in [12], where generalized solutions to (1.1)–(1.3) ($\varepsilon = 0$) are constructed as proper limits of corresponding weak solutions for $\varepsilon > 0$; here, no (poly)convexity assumption of related relaxed energies as in [7] is needed, and the Andrews–Ball condition is sufficient to verify solvability.

So far, numerical analysis for (1.1)–(1.3) is only rare. In [2], the authors establish strong convergence with optimal rates for a fully discrete version of (1.5)–(1.6) in the case $\varepsilon = 1$, using low-order finite elements. However, the error analysis requires additional regularity assumptions for solutions of (1.1)–(1.3), which are not known to hold, in general, and it is pointed out in [2, Remarks (4) and (6) on pp. 76–77] that convergence towards weak solutions is not known and that "a mild constraint on discretization parameters in space and time might be necessary to guarantee convergence" in this context [2, Remark (5) on p. 77]. Moreover, usage of the discrete Gronwall's lemma leads to an exponential factor which grows in terms of $\varepsilon^{-1}T$ for general $\varepsilon > 0$, thus leaving unclear how to efficiently balance temporal and spatial discretization parameters with the regularization parameter to converge to solutions of (1.5)–(1.6) for $\varepsilon = 0$.

In [3], a fully discrete scheme is proposed for $\varepsilon = 0$, where a discrete Young-measure-valued solution which solves the problem is shown to exist. As is argued in [3, Remark 3.7], by first driving the spatial mesh size h to zero, and afterwards tending $k \to 0$, measure-valued solutions to the limiting problem (1.1)–(1.3) may be recovered. However, this leaves open the practically relevant question on whether solutions of the proposed scheme converge to those of (1.1)–(1.3) when spatial and temporal discretization parameters simultaneously tend to zero and whether a mesh constraint $F(k,h) \geq 0$ applies to make convergence hold for $(k,h) \to 0$ or not.

The goal of this paper is to approximate weak ($\varepsilon > 0$) and generalized ($\varepsilon = 0$) solutions of (1.1)–(1.3). In a first step, we study a finite element-based discretization (see scheme A in section 3) for values $\varepsilon > 0$, verify a discrete energy decay property for corresponding iterates, and establish convergence in Theorem 3.1 for $(k,h) \to 0$ towards weak solutions of (1.1)–(1.3), provided triangulations $\mathcal{T}_h$ of $\Omega \subset \mathbb{R}^n$ meet the requirements of [4]—which are satisfied in the case of quasi-uniform meshes, for example. This result answers the questions from above raised in [2] in the sense that discretization parameters may independently tend to zero in the construction process of weak solutions. Then we prove a discrete energy decay property for iterates of scheme A in the case that a mild constraint $F(\varepsilon,k) \geq 0$ is additionally met and eventually show a convergence towards generalized solutions of (1.1)–(1.3) (for $\varepsilon = 0$) when all (ε,k,h) simultaneously tend to zero; see Theorem 3.2. Hence, the latter result justifies a convergence of solutions of the proposed scheme A towards generalized solutions of (1.1)–(1.3) (for $\varepsilon = 0$) as an alternative approach to a direct discretization which uses discrete Young-measure-valued solutions.

The remainder of the paper is as follows: Section 2 provides necessary background material for problem (1.1)–(1.3) as well as further details on used notation. Section 3 proposes a fully discrete scheme and studies both stability and convergence for $(k,h) \to 0$ and $(\varepsilon,k,h) \to 0$. Section 4 reports on computational experiments to

study formation and propagation of two-dimensional (2D) microstructures in the case of positive and vanishing values of $\varepsilon \geq 0$ in (1.1)–(1.3).

2. Preliminaries and notation.

2.1. Weak solutions ($\varepsilon > 0$) (resp.) generalized solutions ($\varepsilon = 0$), of (1.1)–(1.3). For $\varepsilon > 0$, weak solutions to (1.1)–(1.3) are constructed in [8], provided $\sigma = D\phi$ satisfies the following hypotheses:

(H1) $\phi \in C^2(\mathbb{R}^{m\times n})$.

(H2) There exist constants $C_i > 0$ and $2 \leq p < \infty$ such that

$$C_1\,|F|^p - C_2 \leq \phi(F) \leq C_2\,(|F|^p + 1)\,,$$
$$|\sigma^{(i)}(F)| \leq C_2\left(|F|^{p-1-i} + 1\right) \quad \text{for } i = 0, 1.$$

(H3) There exists $K > 0$ such that

$$\langle \sigma(F_1) - \sigma(F_2), F_1 - F_2\rangle \geq -K|F_1 - F_2|^2 \quad \forall\, F_1, F_2 \in \mathbb{R}^{m\times n}.$$

Note that some polynomial asymptotic growth behavior for first and second order derivatives of ϕ is assumed in (H2). Hypothesis (H3) is often referred to as the Andrews–Ball condition. In the following, let $E[u, v] = \int_\Omega [\phi(\nabla u) + \frac{1}{2}|v|^2]\,\mathrm{d}x$ denote the corresponding energy functional.

DEFINITION 2.1. *Let $\Omega \subset \mathbb{R}^n$ be a bounded Lipschitz domain, fix $T > 0$ finite, and let ϕ satisfy assumptions* (H1), (H2), *and* (H3). *Suppose that $(u_0, v_0) \in W_0^{1,p}(\Omega, \mathbb{R}^m)\times L^2(\Omega, \mathbb{R}^m)$. A function $u^\varepsilon : \Omega_T \to \mathbb{R}^m$ is called a* weak solution *of* (1.1)–(1.3) *for $\varepsilon > 0$ if there hold*

(i) $u^\varepsilon \in L^\infty(0, T; W_0^{1,p}(\Omega, \mathbb{R}^m)) \cap W^{1,\infty}(0, T; L^2(\Omega, \mathbb{R}^m)) \cap W^{1,2}(0, T; W_0^{1,2}(\Omega, \mathbb{R}^m))$;

(ii) *initial data are attained, i.e., for $t \to 0$,*

$$u^\varepsilon(t, \cdot) \to u_0 \quad \text{in } L^2(\Omega, \mathbb{R}^m), \qquad u_t^\varepsilon(t, \cdot) \rightharpoonup v_0 \quad \text{in } L^2(\Omega, \mathbb{R}^m);$$

(iii) *for all $\xi \in C_0^\infty([0, T) \times \Omega; \mathbb{R}^m)$ there holds*

$$\int_0^T \int_\Omega [-u_t \cdot \xi_t + \sigma(\nabla u^\varepsilon) \cdot \nabla\xi + \varepsilon\,\nabla u_t^\varepsilon \cdot \nabla\xi]\,\mathrm{d}x\mathrm{d}t = \int_\Omega v_0 \cdot \xi(0, \cdot)\,\mathrm{d}x;$$

(iv) *for almost every $t \in [0, T]$,*

$$E\left[u^\varepsilon(t, \cdot), u_t^\varepsilon(t, \cdot)\right] - E\left[u_0, v_0\right] \leq -\varepsilon \int_0^t \int_\Omega |\nabla u_t^\varepsilon|^2\,\mathrm{d}x\mathrm{d}s.$$

The existence of weak solutions to (1.1)–(1.3) ($\varepsilon > 0$) has been shown in [8]; moreover, weak solutions are unique provided that σ is globally Lipschitz. For $\varepsilon = 0$, gradients of solutions of (1.1)–(1.3) are measure-valued, which leads to the following notion of generalized solutions. In the following, we employ Radon measures $\nu \equiv \{\nu_{t,x}\} \in L_w^\infty(\Omega_T, \mathcal{M}(\mathbb{R}^{m\times n})) := (L^1(\Omega_T; C_0(\mathbb{R}^{m\times n}))'$.

DEFINITION 2.2. *Let $\Omega \subset \mathbb{R}^n$ be a bounded Lipschitz domain, fix $T > 0$ finite, and let ϕ satisfy assumptions* (H1), (H2), *and* (H3). *Suppose that $(u_0, v_0) \in W_0^{1,p}(\Omega, \mathbb{R}^m)\times W^{1,2}(\Omega, \mathbb{R}^m)$. A tuple (u, ν) is called a* generalized solution *of* (1.1)–(1.3) *for $\varepsilon = 0$ if there hold*

(i) $(u, \nu) \in [L^\infty(0, T; W_0^{1,p}(\Omega, \mathbb{R}^m)) \cap W^{1,\infty}(0, T; L^2(\Omega, \mathbb{R}^m))] \times L_w^\infty(\Omega_T, \mathcal{M}(\mathbb{R}^{m\times n}))$, *where $\nu = \{\nu_{t,x}\}_{t,x}$ is a probability measure;*

(ii) *initial data are attained, i.e., for* $t \to 0$,

$$u(t,\cdot) \to u_0 \quad \text{in } L^2(\Omega,\mathbb{R}^m), \qquad u_t(t,\cdot) \rightharpoonup v_0 \quad \text{in } L^2(\Omega,\mathbb{R}^m);$$

(iii) *for all* $\xi \in C_0^\infty([0,T)\times\Omega;\mathbb{R}^m)$,

$$\int_0^T \int_\Omega \left[-u_t(t,x)\cdot\xi_t(t,x) + \nabla\xi(t,x) \cdot \int_{\mathbb{R}^{m\times n}} \sigma(\lambda)\,\mathrm{d}\nu_{t,x}(\lambda) \right] \mathrm{d}x\mathrm{d}t = \int_\Omega v_0\cdot\xi(0,\cdot)\,\mathrm{d}x$$

and

$$\nabla u(t,x) = \int_{\mathbb{R}^{n\times m}} \lambda\,\mathrm{d}\nu_{t,x}(\lambda) \quad \text{a.e. in } \Omega_T.$$

The existence of generalized solutions to (1.1)–(1.3) has been shown in [5, 6] for cases where the quasi convexification of ϕ is convex. In [12], generalized solutions are obtained as proper limits of weak solutions of (1.1)–(1.3) for $\varepsilon \to 0$, for which assumptions (H1), (H2), and (H3) are sufficient.

2.2. Finite element spaces, discrete time derivatives, and interpolation. Throughout this paper we assume that $\mathcal{T}_h$ is a regular triangulation of the polygonal- or polyhedral-bounded Lipschitz domain $\Omega \subset \mathbb{R}^n$ into triangles or tetrahedra of maximal diameter $h > 0$, for $n = 2, 3$, which satisfies the requirements stated in [4]. Let

$$V_h := \{v_h \in C(\overline{\Omega},\mathbb{R}^m) : v\big|_K \text{ affine function } \forall\, K \in \mathcal{T}_h\} \subset W^{1,p}(\Omega)$$

denote the lowest-order conforming finite element space. Given the set of all nodes (or vertices) $\mathcal{N}_h$ in $\mathcal{T}_h$ and letting $\{\varphi_z : z \in \mathcal{N}_h\}$ denote the nodal basis in V_h, we define the nodal interpolation operator $\mathcal{I}_h : C(\overline{\Omega},\mathbb{R}^m) \to V_h$ by $\mathcal{I}_h\psi := \sum_{z\in\mathcal{N}_h} \psi(z)\varphi_z$ for $\psi \in C(\overline{\Omega},\mathbb{R}^m)$. Projections $P_X : X \to V_h$, for $X = L^2(\Omega), W_0^{1,2}(\Omega)$ are defined via $(u - P_X u, \Phi)_X = 0$ for all $\Phi \in V_h$; in particular, there holds $P_X u \to u$ in X for $h \to 0$.

Given a time-step size $k > 0$ and a sequence $\{\varphi^j\}_{j\ge 0}$ in some vector space X, we set, for $j \ge 1$,

$$d_t\varphi^{j+1} := k^{-1}\left(\varphi^{j+1} - \varphi^j\right) \quad \text{and} \quad d_t^2\varphi^{j+1} := d_t\left(d_t\varphi^{j+1}\right).$$

The piecewise constant interpolations of φ^j are defined on $0 \le t \le Jk$ such that $t \in [jk,(j+1)k)$ for some $j \in \{0,..,J-1\}$ by

$$\varphi_k^-(t,\cdot) := \varphi^j \quad \text{and} \quad \varphi_k^+(t,\cdot) := \varphi^{j+1},$$

and a piecewise affine interpolation is defined through

$$\hat{\varphi}_k(t,\cdot) := \frac{t-jk}{k}\varphi^{j+1} + \frac{(j+1)k-t}{k}\varphi^j.$$

Note that $\|\varphi_k^+ - \hat{\varphi}_k\|_X + \|\varphi_k^- - \hat{\varphi}_k\|_X \le 2k\|\hat{\varphi}_t\|_X$.

Throughout the paper, let $C > 0$ be a generic constant which depends on given data Ω, T, u_0, v_0, and ϕ and geometric properties of $\mathcal{T}_h$.

3. A finite element-based discretization of (1.1)–(1.3) with $\varepsilon \geq 0$. In [8], the authors introduce and study the semidiscretization in time given by (1.4)–(1.6) for the case $\varepsilon = 1$, where iterates $u_\varepsilon^j : \Omega \to \mathbb{R}^m$ for $j \geq 1$ minimize $E_k^j : W_0^{1,p}(\Omega, \mathbb{R}^m) \to \mathbb{R}$, with

$$E_k^j[w] := \int_\Omega \left[\phi(\nabla w) + \frac{\varepsilon}{2k} |\nabla(w - u_\varepsilon^{j-1})|^2 + \frac{1}{2k^2} |w - 2u_\varepsilon^{j-1} + u_\varepsilon^{j-2}|^2 \right] \mathrm{d}x.$$

The attainment of minimizers follows from the convexity of $F \mapsto \tilde{\phi}(F) := \phi(F) + \frac{\varepsilon}{2k}|F|^2$ for $k \leq \frac{\varepsilon}{K}$, since (H3) implies that

$$\left\langle D\tilde{\phi}(F_1) - D\tilde{\phi}(F_2), F_1 - F_2 \right\rangle \geq \left(-K + \frac{\varepsilon}{k} \right) |F_1 - F_2|^2 \qquad \forall\, F_1, F_2 \in \mathbb{R}^{m \times n}.$$

As a consequence, $E_k^j : W_0^{1,p}(\Omega, \mathbb{R}^m) \to \mathbb{R}$ is weakly lower-semicontinuous for every $j \geq 1$.

The second step in their construction is to derive a (semi)discrete energy inequality to control iterates of (1.4)–(1.6). For every $\delta \in (0,1)$, and all $k \leq 2\frac{\delta\varepsilon}{K}$, there holds

$$\sup_{j \in \mathbb{N}_0} E\left[u_\varepsilon^j, d_t u_\varepsilon^j\right] + k \sum_{j=1}^{\infty} \int_\Omega \left[\varepsilon(1-\delta) |\nabla d_t u_\varepsilon^j|^2 + \frac{\varepsilon k}{2} |d_t^2 u^j|^2 \right] \mathrm{d}x \leq E[u_\varepsilon^0, v_\varepsilon^0]. \tag{3.1}$$

Therefore, uniform bounds are available, and the convergence of globally continuous, piecewise affine interpolations $\{\hat{u}_k\}$ to a limit $u^\varepsilon : \Omega_T \to \mathbb{R}^m$ for $k \to 0$ may be concluded, as well as

$$\left| \int_{\Omega_T} ((\hat{u}_k)_t - u_t^\varepsilon)\, \xi_t \, \mathrm{d}x\mathrm{d}t \right| + \varepsilon \left| \int_{\Omega_T} \nabla\left((\hat{u}_k)_t - u_t^\varepsilon\right) \cdot \nabla \xi \, \mathrm{d}x\mathrm{d}t \right| \to 0 \quad \forall\, \xi \in C_0^\infty\left(\Omega_T, \mathbb{R}^m\right).$$

The crucial last step now is to identify limits in the nonlinear term to obtain the second term in (iii) of Definition 2.1.

For this purpose, strong convergence $\nabla \hat{u}_k \to \nabla u^\varepsilon$ in $L^2(\Omega_T, \mathbb{R}^m)$ for $k \to 0$ is shown, which ensures enough compactness of sequences to pass to the limit in the second term in (iii) of Definition 2.1 and then proves $u^\varepsilon : \Omega_T \to \mathbb{R}^m$ to be a weak solution of (1.1)–(1.3).

The goal of this paper is to study the convergence properties of a fully discrete version of (1.4)–(1.6), where lowest-order conforming finite elements are used. Below, we drop the index ε such that $U_\varepsilon^j \equiv U^j \in V_h$ for every $j \geq -1$ and $V^j := k^{-1}(U^j - U^{j-1})$ for every $j \geq 0$. The practical scheme reads as follows for a given positive triple $(\varepsilon, k, h) > 0$.

Scheme A. Let $U^0 = P_{L^2} u_0 \in V_h$ and $V^0 = P_{L^2} v_0 \in V_h$. For $j \geq 1$, and given $U^{j-1}, U^{j-2} \in V_h$, find $U^j \in V_h$ such that for all $\Phi \in V_h$ there holds

$$(d_t^2 U^j, \Phi) + \left(\sigma(\nabla U^j), \nabla \Phi\right) + \varepsilon\left(\nabla d_t U^j, \nabla \Phi\right) = 0 \qquad \forall\, \Phi \in V_h. \tag{3.2}$$

Similar to above, solutions $U^j \in V_h$ minimize $E_{k,h}^j : V_h \to \mathbb{R}$, with

$$E_{k,h}^j[W] := \int_\Omega \left[\phi(\nabla W) + \frac{\varepsilon}{2k} |\nabla(W - U^{j-1})|^2 + \frac{1}{2k^2} |W - 2U^{j-1} + U^{j-2}|^2 \right] \mathrm{d}x.$$

The scheme has been proposed in [2]. The existence of solutions $\{U^j\}_{j \geq 1} \subset V_h$ follows from continuity and coercivity properties of $E_{k,h}^j$. The following lemma states

a discrete energy inequality for iterates $\{U^j\}_{j\geq -1} \subset V_h$ in the case $k = \mathcal{O}(\varepsilon)$. Let $V^j := d_t U^j \in V_h$ for $j \geq 1$.

LEMMA 3.1. *Let $\{U^j\}_{j\geq -1} \subset V_h$ solve scheme A. Given $\delta \in (0,1)$, for all $k \leq \frac{\delta\varepsilon}{K}$ there holds*

$$\sup_{j\in\mathbb{N}_0} E\left[U^j, V^j\right] + k\sum_{j=1}^{\infty}\int_\Omega \left[\varepsilon(1-\delta)|\nabla V^j|^2 + \frac{k}{2}|d_t V^j|^2\right] \mathrm{d}x \leq E[U^0, V^0]. \tag{3.3}$$

Proof. The proof is the same as in the context of semidiscretization [8, Lemma 2.2], which combines convexity properties of the above $\tilde{\phi}$ and uses (3.2) with $\Phi = d_t U^j$. ∎

We remark that the dissipative character of scheme A is the reason for the last term on the left-hand side of (3.3). As will be evident below, this control is necessary to verify convergence for the scheme.

We use $V^j = d_t U^j \in V_h$ to restate scheme A for $j \geq 0$ as follows:

$$(d_t V^j, \Phi) + \big(\sigma(\nabla U^j), \Phi\big) + \varepsilon\,(\nabla d_t U^j, \nabla\Phi) = 0, \quad (V^j, \Phi) - (d_t U^j, \Phi) = 0 \qquad \forall\,\Phi \in V_h. \tag{3.4}$$

We now use $(U_k^+, \hat{U}_k)$ (resp.) $(V_k^+, \hat{V}_k)$, which are constant and linear interpolations in time of the sequences $\{U^j\}$ (resp.) $\{V^j\}$. For notational brevity, most of the time we omit subindices k and use $(U^+, \hat{U})$ and $(V^+, \hat{V})$ to stand for $(U_k^+, \hat{U}_k)$ and $(V_k^+, \hat{V}_k)$, respectively.

In the next lemma, we restate scheme A for $\hat{U}, \hat{V} \in C([0,T]; V_h)$.

LEMMA 3.2. *Suppose that the assumptions of Lemma 3.1 are valid. For all $T > 0$, and $\Psi \in W^{1,1}(0,T;V_h) \cap L^2(0,T;V_h)$, there holds*

$$\begin{aligned}
&\left| \int_0^T \left[-(\hat{U}_t, \Psi_t) + \Big(\sigma(\nabla\hat{U}), \nabla\Psi\Big) + \varepsilon\,(\nabla\hat{U}_t, \nabla\Psi)\right] \mathrm{d}t \right. \\
&\quad \left. + \Big(\hat{V}(T,\cdot), \Psi(T,\cdot)\Big) - \big(V^0, \Psi(0,\cdot)\big) \right| \\
&\leq \left| \int_0^T (\hat{V} - V^+, \Psi_t)\,\mathrm{d}t \right| + \left| \int_0^T \Big(\sigma(\nabla\hat{U}) - \sigma(\nabla U^+), \nabla\Psi\Big)\,\mathrm{d}t \right|.
\end{aligned} \tag{3.5}$$

Proof. We rewrite (3.4) as follows: For all $\Psi, \Phi \in W^{1,1}(0,T;V_h) \cap L^2(0,T;V_h)$,

$$\int_0^T \left[(\hat{V}_t, \Psi) + \Big(\sigma(\nabla U^+), \nabla\Psi\Big) + \varepsilon\,(\nabla\hat{U}_t, \nabla\Psi)\right] \mathrm{d}s = 0, \tag{3.6}$$

$$\int_0^T \Big[(V^+, \Phi) - (\hat{U}_t, \Phi)\Big]\,\mathrm{d}s = 0. \tag{3.7}$$

Integration by parts in time in the first term of (3.6) leads to

$$\begin{aligned}
&\int_0^T \left[-(\hat{V}, \Psi_t) + \Big(\sigma(\nabla U^+), \nabla\Psi\Big) + \varepsilon\,(\nabla\hat{U}_t, \nabla\Psi)\right] \mathrm{d}s \\
&\qquad + \Big(\hat{V}(T,\cdot), \Psi(T,\cdot)\Big) - \Big(\hat{V}(0,\cdot), \Psi(0,\cdot)\Big) = 0.
\end{aligned}$$

Because of (3.7), we obtain

$$\begin{aligned}
&\int_0^T \left[-(\hat{U}_t, \Psi_t) + \Big(\sigma(\nabla\hat{U}), \nabla\Psi\Big) + \varepsilon\,(\nabla\hat{U}_t, \nabla\Psi)\right] \mathrm{d}s + \big(V^0, \Psi(0,\cdot)\big) \\
&\quad = \int_0^T \left[\Big(\hat{V} - V^+, \Psi_t\Big) + \Big(\sigma(\nabla\hat{U}) - \sigma(\nabla U^+), \nabla\Psi\Big)\right] \mathrm{d}s.
\end{aligned}$$

∎

The following result gives uniform bounds for $\{\hat{V}\}$.

LEMMA 3.3. *Let $p \geq 2$, and suppose that the assumptions of Lemma* 3.1 *are valid. Then*

$$\|\hat{V}_t\|_{L^2(0,T;W^{-1,p'})} + \|\hat{V}\|_{L^2(0,T;W^{1,2})} \leq C.$$

Proof. Use (3.6), (H2), and [4, Theorems 3 and 4] to conclude that

$$\begin{aligned}\|\hat{V}_t(s,\cdot)\|_{W^{-1,p'}} &:= \sup_{\varphi \in W_0^{1,p}} \frac{(\hat{V}_t(s,\cdot), P_{L^2}\varphi)}{\|\varphi\|_{W^{1,p}}} \\ &\leq C\left(1 + \|\nabla U^+(s,\cdot)\|_{L^p}^{p-1}\right) + C\varepsilon\|\nabla \hat{U}_t(s,\cdot)\|_{L^2}.\end{aligned}$$

Estimate (3.3) then implies the first part of the assertion. The second part follows from (3.3), and

$$\|\hat{V}\|_{L^2(0,T;W^{1,2})} \leq \|V^-\|_{L^2(0,T;W^{1,2})} + \|V^+\|_{L^2(0,T;W^{1,2})}. \qquad \square$$

It follows from Lemmas 3.1 and 3.3 and Hypothesis (H2) that, under the given mesh constraint, there exist $u^\varepsilon \in L^\infty(0,T;W_0^{1,p}(\Omega,\mathbb{R}^m)) \cap W^{1,\infty}(0,T;L^2(\Omega,\mathbb{R}^m)) \cap W^{1,2}(0,T;W_0^{1,2}(\Omega))$ and a convergent subsequence of $\{\hat{U}\}_{k,h}$ such that, for $(k,h) \to 0$,

$$\begin{aligned}
&\hat{U} \overset{*}{\rightharpoonup} u^\varepsilon \quad \text{in } L^\infty\left(0,T;W_0^{1,p}(\Omega,\mathbb{R}^m)\right) \cap W^{1,\infty}\left(0,T;L^2(\Omega,\mathbb{R}^m)\right), \\
&\hat{U} \rightharpoonup u^\varepsilon \quad \text{in } W^{1,2}\left(0,T;W_0^{1,2}(\Omega,\mathbb{R}^m)\right), \\
&\sigma(\nabla\hat{U}) \overset{*}{\rightharpoonup} \tilde{\sigma} \quad \text{in } L^\infty\left(0,T;L^{p'}(\Omega,\mathbb{R}^{m\times n})\right), \\
&\hat{V} \rightharpoonup u_t^\varepsilon \quad \text{in } W^{1,2}\left(0,T;W^{-1,p'}(\Omega,\mathbb{R}^m)\right) \cap L^2\left(0,T;W_0^{1,2}(\Omega,\mathbb{R}^m)\right), \\
&\hat{U} \to u^\varepsilon \quad \text{in } W^{1,2}\left(0,T;L^2(\Omega,\mathbb{R}^m)\right),
\end{aligned} \tag{3.8}$$

thanks to (3.7), $(3.8)_4$ is a consequence of the control

$$\|V^+ - \hat{V}\|_{L^2(\Omega_T)} \leq \sqrt{k}\left[\sqrt{k}\|\hat{V}_t\|_{L^2(\Omega_T)}\right] \leq \sqrt{k}E[U_0,V_0], \tag{3.9}$$

and $(3.8)_5$ follows from an Aubin–Lions compactness result. As can be concluded from, e.g., [14, Proposition 1.2, p. 106], $u^\varepsilon \in C([0,T];L^2(\Omega,\mathbb{R}^m))$. We are ready to prove the first main convergence theorem of this paper.

THEOREM 3.1. *Let $T > 0$, $\varepsilon > 0$, and $(u_0,v_0) \in W_0^{1,p}(\Omega,\mathbb{R}^m) \times L^2(\Omega,\mathbb{R}^m)$. Let assumptions* (H1), (H2), *and* (H3) *be satisfied, and let $k \leq \frac{\varepsilon}{2K}$. Then there exist $u^\varepsilon : \Omega_T \to \mathbb{R}^m$ and a subsequence of solutions $\{\hat{U}\}_{k,h}$ (denoted by the same notation) of scheme A such that for $(k,h) \to 0$*

$$\hat{U} \overset{*}{\rightharpoonup} u^\varepsilon \quad \text{in } L^\infty\left(0,T;W_0^{1,p}(\Omega,\mathbb{R}^m)\right) \cap W^{1,\infty}\left(0,T;L^2(\Omega,\mathbb{R}^m)\right).$$

Moreover, u^ε is a weak solution of (1.1)–(1.3).

This result complements the one in [2], because it verifies the convergence of solutions of scheme A to weak solutions of (1.1)–(1.3) and hence answers questions raised in [2, Remarks (5) and (6) on pp. 76–77]. Note that the result in the theorem is for a subsequence $\{\hat{U}\}_{k,h}$; weak convergence of the whole sequence holds in the

case of unique weak solutions, which is, e.g., valid for globally Lipschitz-continuous σ. Finally, thanks to [4] there holds

$$\big(U^0, V^0\big) \to (u_0, v_0) \quad \text{in } W_0^{1,p}(\Omega,\mathbb{R}^m) \times L^2(\Omega,\mathbb{R}^m) \quad (h \to 0).$$

Proof. Step 1: We may use $(3.8)_2$ and $(3.8)_4$ to verify for $(k,h) \to 0$

$$\begin{aligned} &\left| \int_0^T (\hat{U}_t - u_t^\varepsilon, \Psi_t) \mathrm{d}s \right| + \varepsilon \left| \int_0^T (\nabla[\hat{U}_t - u_t^\varepsilon], \nabla\Psi)\, \mathrm{d}s \right| \\ &\qquad + \big|\big(V^0 - v_0, \Psi(0,\cdot)\big)\big| + \left| \Big(\hat{V}(T,\cdot) - u_t^\varepsilon(T,\cdot), \Psi(T,\cdot)\Big) \right| \to 0 \end{aligned} \tag{3.10}$$

for all $\Psi(t,\cdot) = \mathcal{I}_h \xi(t,\cdot)$, and $\xi \in W^{1,1}(0,T;C_0^\infty(\Omega,\mathbb{R}^m)) \cap L^2(0,T;C_0^\infty(\Omega,\mathbb{R}^m))$. Since $\mathcal{I}_h\xi(t,\cdot) \to \xi(t,\cdot)$ in any $W^{1,p}(\Omega)$ $(1 \le p \le \infty)$ for $h \to 0$, again using (3.8), we may conclude that (3.10) holds for any $\Psi \in W^{1,1}(0,T;C_0^\infty(\Omega,\mathbb{R}^m)) \cap L^2(0,T;C_0^\infty(\Omega,\mathbb{R}^m))$.

Terms on the right-hand side of (3.5) vanish, which is a consequence of Lemma 3.1. We compute

$$\begin{aligned} \left| \int_0^T (V^+ - \hat{V}, \Psi_t)\, \mathrm{d}s \right| &\le C\sqrt{k} \left(k \int_0^T \|\hat{V}_t\|_{L^2}^2\, \mathrm{d}s \right)^{1/2} \left(\int_0^T \|\Psi_t\|_{L^2}^2\, \mathrm{d}s \right)^{1/2} \\ &\le C\sqrt{k} \left(\int_0^T \|\Psi_t\|_{L^2}^2\, \mathrm{d}s \right)^{1/2}. \end{aligned} \tag{3.11}$$

For the second term on the right-hand side of (3.5), we use (H2) and Lemma 3.1; without restriction, assume that $p > 2$ for the following calculation:

$$\begin{aligned} &\left| \int_0^T \Big(\sigma(\nabla\hat{U}) - \sigma(\nabla U^+), \nabla\Psi \Big)\, \mathrm{d}s \right| \\ &\quad \le C \int_{\Omega_T} \Big[1 + |\nabla\hat{U}|^{p-2} + |\nabla U^+|^{p-2} \Big] |\nabla[\hat{U} - U^+]| |\nabla\Psi|\, \mathrm{d}x\mathrm{d}s \\ &\quad \le C \left[\int_0^T \left(\int_\Omega \Big[1 + |\nabla\hat{U}|^p + |\nabla U^+|^p \Big]\, \mathrm{d}x \right)^{\frac{p-2}{p}} \left(\int_\Omega |\nabla[\hat{U} - U^+]|^{\frac{p}{2}}\, \mathrm{d}x \right)^{2/p} \mathrm{d}s \right] \\ &\qquad\qquad \|\Psi\|_{L^\infty(0,T;W^{1,\infty})} \\ &\quad \le C \|\Psi\|_{L^\infty(0,T;W^{1,\infty})} \int_0^T \|\nabla[\hat{U} - U^+]\|_{L^{p/2}}\, \mathrm{d}s. \end{aligned} \tag{3.12}$$

The following interpolation result is needed only for $p > 4$:

$$\|\nabla[U^+ - \hat{U}]\|_{L^{p/2}} \le C\, \|\nabla[U^+ - \hat{U}]\|_{L^p}^{\frac{[p-4]_+}{p-2}} \|\nabla[U^+ - \hat{U}]\|_{L^2}^{\min\{1, \frac{2}{p-2}\}},$$

where $[\cdot]_+ := \max\{\cdot, 0\}$. Hence, the last integral term in (3.12) may be controlled as follows, again using Lemma 3.1:

$$\begin{aligned} \int_0^T \|\nabla[U^+ - \hat{U}]\|_{L^{p/2}}\, \mathrm{d}s &\le C\, \|\nabla U^+\|_{L^\infty(0,T;L^p)}^{\frac{[p-4]_+}{p-2}} k^{\min\{1, \frac{2}{p-2}\}} \int_0^T \|\nabla\hat{U}_t\|_{L^2}^{\min\{1, \frac{2}{p-2}\}}\, \mathrm{d}s \\ &\le C k^{\min\{1, \frac{2}{p-2}\}} \|\nabla U^+\|_{L^\infty(0,T;L^p)}^{\frac{[p-4]_+}{p-2}} \|\nabla\hat{U}_t\|_{L^2(\Omega_T)}^{\min\{1, \frac{2}{p-2}\}} \\ &\le C k^{\min\{1, \frac{2}{p-2}\}} \varepsilon^{\max\{-\frac{1}{2}, -\frac{1}{p-2}\}}. \end{aligned} \tag{3.13}$$

Consequently, terms on the right-hand sides of (3.11) and (3.12) tend to zero for $k \to 0$, and hence the right-hand side (RHS) of (3.5) tends to zero. By (3.10), we may replace $\hat{U}$ and Ψ in terms one, three, four, and five on the left-hand side of (3.5) by corresponding ones with u^ε and ξ. Therefore it remains to verify that

$$\lim_{k,h\to 0} \int_0^T \Big(\sigma(\nabla \hat{U}) - \sigma(\nabla u^\varepsilon), \nabla \xi\Big)\, \mathrm{d}s = 0$$
$$\forall\, \xi \in W^{1,1}\left(0,T; C_0^\infty(\Omega,\mathbb{R}^m)\right) \cap L^2\left(0,T; C_0^\infty(\Omega,\mathbb{R}^m)\right),$$

which is done in the next step.

Step 2: In order to identify $\tilde{\sigma} = \sigma(\nabla u^\varepsilon)$ almost everywhere in Ω_T, we verify strong convergence $\hat{U} \to u^\varepsilon$ in $L^\infty(0,T,W^{1,2}(\Omega,\mathbb{R}^m))$ for $(k,h) \to 0$. So far, we have

$$\int_0^T \left[-(u_t^\varepsilon, \xi_t) + (\tilde{\sigma}, \nabla\xi) + \varepsilon\,(\nabla u_t^\varepsilon, \nabla\xi)\right] \mathrm{d}t + (u_t^\varepsilon(T,\cdot), \xi(T,\cdot)) - (v_0, \xi(0,\cdot)) = 0$$
$$\forall\, \xi \in W^{1,1}\left(0,T; C_0^\infty(\Omega,\mathbb{R}^m)\right) \cap L^2\left(0,T; C_0^\infty(\Omega,\mathbb{R}^m)\right). \tag{3.14}$$

Note that, by density and (3.8), identity (3.14) holds for all

$$\xi \in X := W^{1,1}\left(0,T; W_0^{1,p}(\Omega,\mathbb{R}^m)\right) \cap L^2\left(0,T; W_0^{1,2}(\Omega,\mathbb{R}^m)\right). \tag{3.15}$$

We subtract (3.14) from the left-hand side of (3.5) and choose $\xi = \Psi \in X \cap W^{1,1}(0,T; V_h) \cap L^2(0,T;V_h)$, where, for all $t \in [0,T]$, we take $\Psi(t,\cdot) = P_{L^2} e(t,\cdot)$, with $e := u^\varepsilon - \hat{U}$. We use $(e_t - P_{L^2} e_t, P_{L^2} e_t) = 0$ and

$$(\nabla e_t, \nabla P_{L^2} e) = (\nabla e_t, \nabla e) + (\nabla e_t, \nabla[P_{L^2} e - e]) = \frac{1}{2}\frac{d}{dt}\|\nabla e\|^2 - (\nabla e_t, \nabla[u^\varepsilon - P_{L^2} u^\varepsilon])$$

to conclude that

$$\begin{aligned} &\left| \int_0^T \left[-\|P_{L^2} e_t\|^2 + \Big(\tilde{\sigma} - \sigma(\nabla\hat{U}), \nabla P_{L^2} e\Big) + \frac{\varepsilon}{2}\frac{d}{dt}\|\nabla e\|^2 \right] \mathrm{d}t \right. \\ &\qquad \left. + (e_t(T,\cdot), P_{L^2} e(T,\cdot)) - \big(v_0 - V^0, P_{L^2} e(0,\cdot)\big) \right| \\ &\qquad \le |\mathrm{RHS}(P_{L^2} e)| + \left| \varepsilon \int_0^T (\nabla e_t, \nabla[u^\varepsilon - P_{L^2} u^\varepsilon])\, \mathrm{d}s \right|, \end{aligned} \tag{3.16}$$

where

$$\mathrm{RHS}(\Psi) := \left| \int_0^T (\hat{V} - V^+, \Psi_t)\, \mathrm{d}t \right| + \left| \int_0^T (\sigma(\nabla\hat{U}) - \sigma(\nabla U^+), \nabla\Psi)\, \mathrm{d}t \right|.$$

The left-hand side of (3.16) consists of five terms I, II, ..., V, which we now study independently, for $(k,h) \to 0$. The first tends to zero, owing to the L^2-stability of P_{L^2} and $(3.8)_5$.

$$\begin{aligned} \mathrm{II} &= \int_0^T \Big(\tilde{\sigma} - \sigma(\nabla\hat{U}), \nabla[P_{L^2} u^\varepsilon - u^\varepsilon]\Big)\, \mathrm{d}t \\ &+ \int_0^T \Big(\sigma(\nabla u^\varepsilon) - \sigma(\nabla\hat{U}), \nabla[u^\varepsilon - \hat{U}]\Big)\, \mathrm{d}t + \int_0^T \Big(\tilde{\sigma} - \sigma(\nabla u^\varepsilon), \nabla[u^\varepsilon - \hat{U}]\Big)\, \mathrm{d}t \\ &=: \mathrm{II}_a + \mathrm{II}_b + \mathrm{II}_c. \end{aligned} \tag{3.17}$$

We employ $[P_{L^2}-\mathrm{Id}]\varphi\to 0$ in $W^{1,p}(\Omega)$ $(h\to 0)$ for $\varphi\in W^{1,p}(\Omega)$, which follows from $W^{1,p}(\Omega)$-stability of the $L^2(\Omega)$-projection from [4, Theorems 3 and 4], and

$$\begin{aligned}\|\varphi-P_{L^2}\varphi\|_{W^{1,p}}&\le\inf_{W\in V_h}[\|\varphi-W\|_{W^{1,p}}+\|W-P_{L^2}\varphi\|_{W^{1,p}}]\\&\le C\inf_{W\in V_h}\|\varphi-W\|_{W^{1,p}}\end{aligned}$$

in combination with (H2), Lemma 3.1, $(3.8)_3$, and the bound $\||\nabla\hat U(t,\cdot)|^{p-1}\|_{L^{p'}}\le\|\nabla\hat U(t,\cdot)\|_{L^p}^{p-1}$ to conclude that $\lim_{h\to 0}\mathrm{II}_a=0$,

$$\begin{aligned}\mathrm{II}_a&\le\int_0^T\left[\|\tilde\sigma\|_{L^{p'}}+\|\sigma(\nabla\hat U)\|_{L^{p'}}\right]\|P_{L^2}u^\varepsilon-u^\varepsilon\|_{W^{1,p}}\,\mathrm{d}t\\&\le C_2\int_0^T\left[\|\tilde\sigma\|_{L^{p'}}+\||\nabla\hat U|^{p-1}\|_{L^{p'}}+1\right]\|P_{L^2}u^\varepsilon-u^\varepsilon\|_{W^{1,p}}\,\mathrm{d}t\\&\le CC_2\|P_{L^2}u^\varepsilon-u^\varepsilon\|_{L^1(0,T;W^{1,p})}\to 0\qquad(h\to 0).\end{aligned}$$

For the second term, we use (H3) to conclude that $\mathrm{II}_b\ge -K\int_0^T\|\nabla[u^\varepsilon-\hat U]\|_{L^2}^2\,\mathrm{d}s$. Because of $\nabla\hat U-\nabla u^\varepsilon\overset{*}{\rightharpoonup}0$ in $L^\infty(0,T;L^p(\Omega,\mathbb{R}^{m\times n}))$ for $(k,h)\to 0$, $\tilde\sigma,\sigma(\nabla u^\varepsilon)\in L^\infty(0,T;L^{p'}(\Omega,\mathbb{R}^{m\times n}))$ due to $(3.8)_{1,3}$, and (H2), this implies that $\lim_{k,h\to 0}\mathrm{II}_c=0$.

We show that term IV vanishes for $(k,h)\to 0$; for this purpose, we use the stability of the $L^2(\Omega)$-projection to conclude that $|(e_t(T,\cdot),P_{L^2}e(T,\cdot))|\le\|e_t(T,\cdot)\|_{L^2}\|e(T,\cdot)\|_{L^2}$ and $(3.8)_5$, which implies, for $(k,h)\to 0$, that

$$\|e(T,\cdot)\|_{L^2}\le\sqrt{T}\|e_t\|_{L^2(\Omega_T)}+\|e(0,\cdot)\|_{L^2}\to 0.$$

Finally, $\lim_{h\to 0}\mathrm{V}=0$ follows from $[P_{L^2}-\mathrm{id}]v_0,[P_{L^2}-\mathrm{id}]u_0\to 0$ in $L^2(\Omega)$ $(h\to 0)$.

We collect these results to restate estimate (3.16) in the form

$$\begin{aligned}&\frac{\varepsilon}{2}\|\nabla[u^\varepsilon(t,\cdot)-\hat U(t,\cdot)]\|_{L^2}^2\\&\le\frac{\varepsilon}{2}\|\nabla[u^\varepsilon(0,\cdot)-\hat U(0,\cdot)]\|_{L^2}^2+\eta(k,h)+K\int_0^t\|\nabla[u^\varepsilon(s,\cdot)-\hat U(s,\cdot)]\|_{L^2}^2\,\mathrm{d}s\end{aligned}\tag{3.18}$$

for almost every $t\in[0,T]$. For brevity, we introduce $0\le\eta(k,h)\le 1$ such that $\lim_{k,h\to 0}\eta(k,h)=0$. By Gronwall's inequality, this verifies that $\hat U\to u^\varepsilon$ in $L^\infty(0,T;W^{1,2}(\Omega,\mathbb{R}^m))$ for $(k,h)\to 0$.

But then we may identify $\tilde\sigma=\sigma(\nabla u^\varepsilon)$ a.e. in Ω_T, which then establishes assertion (iii) of Definition 2.1.

Step 3: The energy estimate given in Lemma 3.1 may be restated as follows:

$$\sup_{t\ge 0}E\left[U^+(t,\cdot),\hat U_t(t,\cdot)\right]+\varepsilon\int_0^t(1-\delta)|\nabla\hat U_t|^2\,\mathrm{d}x\le E\left[U^0V^0\right],\tag{3.19}$$

for all $t\ge 0$, and for all $k\le\frac{\delta\varepsilon}{K}$, given $\delta\in(0,1)$. A problem arises when passing to the limit $(k,h)\to 0$, due to the lack of a weakly lower-semicontinuity property on $W^{1,p}$

of $u \mapsto \int_\Omega \phi(\nabla u)\,\mathrm{d}x$; the following argument is taken from [8, p. 377]: Because of (H3), we exploit the convexity of $F \mapsto \hat{\phi}(F) := \phi(F) + \frac{K}{2}|F|^2$, to conclude, for $t \geq 0$, that

$$
\begin{aligned}
&\int_\Omega \phi\left(\nabla u^\varepsilon(t,\cdot)\right)\,\mathrm{d}x \\
&\quad = \int_\Omega \left[\hat{\phi}\left(\nabla u^\varepsilon(t,\cdot)\right) - \frac{K}{2}|\nabla u^\varepsilon(t,\cdot)|^2\right]\,\mathrm{d}x \\
&\quad \leq \liminf_{k,h\to 0} \int_\Omega \left[\hat{\phi}\left(\nabla \hat{U}(t,\cdot)\right) - \frac{K}{2}|\nabla u^\varepsilon(t,\cdot)|^2\right]\,\mathrm{d}x \\
&\quad \leq \liminf_{k,h\to 0} \int_\Omega \phi\left(\nabla \hat{U}(t,\cdot)\right)\,\mathrm{d}x + \liminf_{k,h\to 0} \frac{K}{2}\int_\Omega \left||\nabla \hat{U}(t,\cdot)|^2 - |\nabla u^\varepsilon(t,\cdot)|^2\right|\,\mathrm{d}x \qquad (3.20)\\
&\quad = \liminf_{k,h\to 0} \int_\Omega \phi\left(\nabla \hat{U}(t,\cdot)\right)\,\mathrm{d}x,
\end{aligned}
$$

thanks to $\hat{U} \to u^\varepsilon$ in $L^\infty(0,T;W^{1,2}(\Omega,\mathbb{R}^m))$ for $(k,h) \to 0$. Inserting this into (3.19) and using (3.8) yields to assertion (iv) of Definition 2.1.

Step 4: The initial data are attained in the sense of Definition 2.1(ii): Because of $u^\varepsilon \in C([0,T];L^2(\Omega,\mathbb{R}^m))$; (see below (3.9)), the function $u^\varepsilon(t,\cdot)$ is well-defined for each $t \geq 0$, and $u^\varepsilon(t,\cdot) \to u_0$ in $L^2(\Omega,\mathbb{R}^m)$ for $t \to 0$.

Convergence $u_t^\varepsilon(t,\cdot) \rightharpoonup v_0$ in $L^2(\Omega,\mathbb{R}^m)$ $(t \to 0)$ follows from the absolute continuity property of $u_t^\varepsilon(t,\cdot) : W^{-1,p'}(\Omega,\mathbb{R}^m) \to W^{-1,p'}(\Omega,\mathbb{R}^m)$, which is a consequence of $(3.8)_4$, and $u_t^\varepsilon \in L^\infty(0,T;L^2(\Omega,\mathbb{R}^m))$; see, e.g., [18, Lemma 1.4, p. 178]. ∎

Next, we study how to relate ε and (k,h), for all three parameters tending to zero, in order to construct generalized solutions in the sense of Definition 2.2 from scheme A. For this purpose, no strong convergence of gradients $\{\nabla \hat{U}_\varepsilon\}_{k,h}$ is needed if generalized solutions to (1.1)–(1.3) for $\varepsilon = 0$ are approximated and the strictest mesh constraint comes from (3.9). Then the following result shows that, for simultaneously tending $(\varepsilon,k,h) \to 0$, solutions $\hat{U} : \Omega_T \to \mathbb{R}^m$ to scheme A approximate generalized solutions to (1.1)–(1.3) for $\varepsilon = 0$.

THEOREM 3.2. *Let $T > 0$ and $(u_0,v_0) \in W_0^{1,p}(\Omega,\mathbb{R}^m) \times L^2(\Omega,\mathbb{R}^m)$. Let assumptions* (H1), (H2), *and* (H3) *be satisfied, and let $k = o(\varepsilon)$ in scheme A. Then there exist $u : \Omega_T \to \mathbb{R}^m$ and a subsequence of solutions $\{\hat{U}\}_{\varepsilon,k,h}$ (denoted by the same notation) of scheme A such that as $(\varepsilon,k,h) \to 0$*

$$
\begin{aligned}
\hat{U} &\overset{*}{\rightharpoonup} u \quad \text{in } L^\infty\left(0,T;W_0^{1,p}(\Omega,\mathbb{R}^m)\right) \cap W^{1,\infty}\left(0,T;L^2(\Omega,\mathbb{R}^m)\right),\\
\delta_{\nabla \hat{U}} &\overset{*}{\rightharpoonup} \nu \quad \text{in } L_w^\infty\left(\Omega_T;\mathcal{M}(\mathbb{R}^{m\times n})\right).
\end{aligned}
$$

Moreover, the tuple (u,ν) is a generalized solution to (1.1)–(1.3) *for $\varepsilon = 0$.*

Proof. Step 1: It follows from Lemmas 3.1 and 3.3 and Hypothesis (H2) that, under the given mesh constraint, there exist

$$
(u,\nu) \in \left[L^\infty\left(0,T;W_0^{1,p}(\Omega,\mathbb{R}^m)\right) \cap W^{1,\infty}\left(0,T;L^2(\Omega,\mathbb{R}^m)\right)\right] \times L_w^\infty\left(\Omega_T,\mathcal{M}(\mathbb{R}^{m\times n})\right)
$$

and a convergent subsequence of $\{\hat{U}\}_{\varepsilon,k,h}$ such that for $(\varepsilon,k,h)\to 0$

$$
\begin{aligned}
\hat{U} &\overset{*}{\rightharpoonup} u \quad \text{in } L^\infty\left(0,T;W_0^{1,p}(\Omega,\mathbb{R}^m)\right)\cap W^{1,\infty}\left(0,T;L^2(\Omega,\mathbb{R}^m)\right),\\
\sigma(\nabla\hat{U}) &\overset{*}{\rightharpoonup} \overline{\sigma} \quad \text{in } L^\infty\left(0,T;L^{p'}(\Omega,\mathbb{R}^{m\times n})\right),\\
\hat{V} &\rightharpoonup u_t \quad \text{in } W^{1,2}\left(0,T;W^{-1,p'}(\Omega,\mathbb{R}^m)\right),\\
\hat{U} &\to u \quad \text{in } L^2\left(\Omega_T,\mathbb{R}^m\right),\\
\delta_{\nabla\hat{U}} &\overset{*}{\rightharpoonup} \nu \quad \text{in } L^\infty_w\left(\Omega_T;\mathcal{M}(\mathbb{R}^{m\times n})\right).
\end{aligned}
\tag{3.21}
$$

Step 2: We compute the limit $\varepsilon\to 0$ of the third term on the left-hand side of (3.5):

$$
\left|\varepsilon\int_0^T\left(\nabla\hat{U}_t,\nabla\Psi\right)\,\mathrm{d}t\right|\le\sqrt{\varepsilon}\left[\sqrt{\varepsilon}\|\nabla\hat{U}_t\|_{L^2(\Omega_T)}\right]\|\nabla\Psi\|_{L^2(\Omega_T)}\to 0
$$

$$
\forall\,\Psi\in L^2\left(0,T;V_h\cap W_0^{1,2}(\Omega,\mathbb{R}^m)\right).
$$

Thanks to [11, Theorem 6.2], there holds

$$
\overline{\sigma}(t,x)=\int_{\mathbb{R}^{m\times n}}\sigma(\lambda)\,\mathrm{d}\nu_{t,x}(\lambda)\quad\text{a.e. in }\Omega_T.
$$

To identify limits in the remaining terms of (3.5) for $(\varepsilon,k,h)\to 0$ is as in the first step of the proof of Theorem 3.1; the arguments given there in step four may be adapted as well, which settles the proof. ∎

4. Computational experiments. In this section, we report on computational studies for fixed and decreasing values of $\varepsilon>0$ in (1.1)–(1.3) performed with scheme A. The nonlinear systems of equations in each time step were approximately solved by a simple fixed-point method. The scheme is implemented in MATLAB with a direct solver for linear systems of equations.

The first example is taken from the literature to allow for direct comparison of our computational results with those obtained with a finite difference method in [16, section 6.2]; we remark that no results of convergence are given in [16] to theoretically justify their scheme.

Example 4.1. For $\Omega=(1,2)^2$, choose $u_0(x,y)=x^2(x-1)y(y-1)$, $v_0(x,y)=0$, and $\phi(F)=\frac{1}{4}(|F|^2-1)^2$. The evolution (1.1)–(1.3) is approximated by scheme A for values $\varepsilon=0.25$, and $(k,h)=(5.\times 10^{-4},2^{-6})$.

Snapshots are displayed in Figure 1, together with the evolution of kinetic, elastic, and total energies, where

$$
E_{\mathrm{kin}}\left[d_tU^j\right]=\frac{1}{2}\int_\Omega|d_tU^j|^2\,\mathrm{d}x,\quad E_{\mathrm{ela}}\left[\nabla U^j\right]=\int_\Omega\phi(\nabla U^j)\,\mathrm{d}x.
$$

The results are in good agreement with those reported in [16, section 6.2].

In the second example, we study dynamics in the same setup of data for values of ε tending to zero.

Example 4.2. Choose data as in Example 4.1 for values $\varepsilon=0.1\cdot 2^{-\ell}$, $0\le\ell\le 2$.

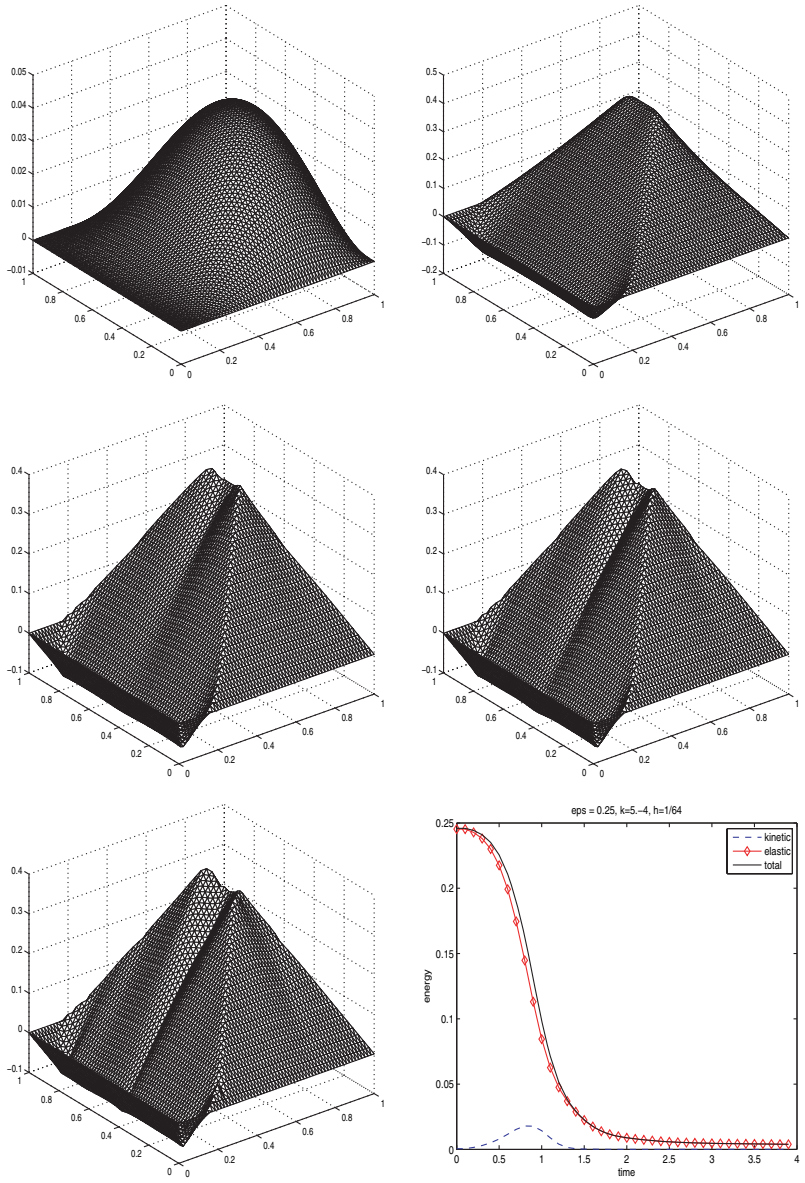


FIG. 1. *Example* 4.1*: Snapshots of* $t \mapsto U(t,\cdot)$*, for* $t = 0.1, 2, 4, 6, 8$*, and* $\varepsilon = 0.25$*. Evolution of kinetic, elastic, and total energy.*

As is known from [5, p. 149], the generalized solutions $u : \Omega_T \to \mathbb{R}$ of (1.1)–(1.3) ($\varepsilon = 0$) solve $u_{tt} = \operatorname{div} p(\nabla u)$, where $p := D\phi^{**}$, for the existing convexification

$$\phi^{**} = \sup\{f : f \le \phi,\ f \text{ convex}\}.$$

For our choice of ϕ, we easily compute $\phi^{**}(F) = \frac{1}{4}([|F|^2 - 1]_+)^2$, where $[\cdot]_+ = \max\{0, \cdot\}$. For the chosen initial data, we then find $u \equiv u_0$ in Ω_T. The snapshots in Figure 2 for decreasing values of ε show 2D microstructures which develop and propagate in time. The qualitative evolution of different energy contributions is comparable to the one depicted in Figure 1.

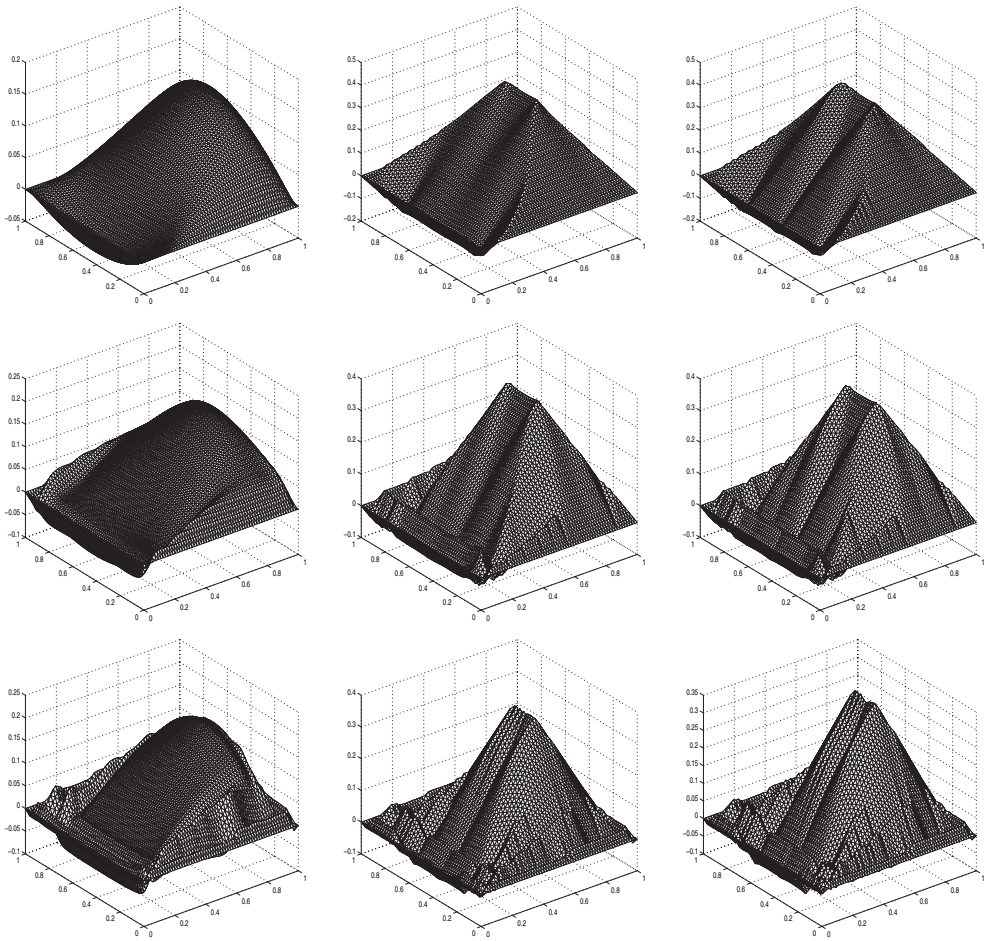


FIG. 2. *Example* 4.2*: Snapshots of* $t \mapsto U(t,\cdot)$ *at times* $t = 1, 2, 4$ *(left to right) for values* $\varepsilon = 0.1, 0.05, 0.025$ *(top to bottom).*

REFERENCES

[1] J. M. BALL, P. J. HOLMES, R. D. JAMES, R. L. PEGO, AND P. J. SWART, *On the dynamics of fine structure*, J. Nonlinear Sci., 1 (1991), pp. 17–70.

[2] C. CARSTENSEN AND G. DOLZMANN, *Time-space discretization of the nonlinear hyperbolic system* $u_{tt} = \operatorname{div}(\sigma(Du) + Du_t)$, SIAM J. Numer. Anal., 42 (2004), pp. 75–89.

[3] C. CARSTENSEN AND M. O. RIEGER, *Young-measure approximations for elastodynamics with non-monotone stress-strain relations*, M2AN Math. Model. Numer. Anal., 38 (2004), pp. 397–418.

[4] M. CROUZEIX AND V. THOMÉE, *The stability in* L_p *and* W_p^1 *of the* L_2*-projection onto finite element function spaces*, Math. Comp., 48 (1987), pp. 521–532.

[5] S. DEMOULINI, *Young measure solutions for nonlinear evolutionary systems of mixed type*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 14 (1997), pp. 143–162.

[6] S. DEMOULINI, *Weak solutions for a class of nonlinear systems of viscoelasticity*, Arch. Ration. Mech. Anal., 155 (2000), pp. 299–334.

[7] S. DEMOULINI, D.M.A. STUART, AND A. E. TZAVARAS, *A variational approximation scheme for three-dimensional elastodynamics with polyconvex energy*, Arch. Ration. Mech. Anal., 157 (2001), pp. 325–344.

[8] G. FRIESECKE AND G. DOLZMANN, *Implicit time discretization and global convergence for a quasi-linear evolution equation with nonconvex energy*, SIAM J. Math. Anal., 28 (1997), pp. 363–380.

[9] G. Friesecke and J. B. McLeod, *Dynamics as a mechanism preventing the formation of finer and finer microstructure*, Arch. Ration. Mech. Anal., 133 (1996), pp. 199–247.
[10] G. Friesecke and J. B. McLeod, *Dynamic stability of non-minimizing phase mixtures*, Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 453 (1997), pp. 2427–2436.
[11] P. Pedregal, *Parametrized Measures and Variational Principles*, Birkhäuser–Verlag, Basel, Switzerland, 1997.
[12] M. O. Rieger, *Young measure solutions for nonconvex elastodynamics*, SIAM J. Math. Anal., 34 (2003), pp. 1380–1398.
[13] C. Rohde, *Phase transitions and sharp-interface limits for the* $1d$*-elasticity system with nonlocal energy*, Interfaces Free Bound., 7 (2005), pp. 107–129.
[14] R. E. Showalter, *Monotone Operators in Banach Space and Nonlinear Partial Differential Equations*, AMS, Providence, RI, 1997.
[15] M. Slemrod, *Dynamics of measured valued solutions to a backward-forward heat equation*, J. Dynam. Differential Equations, 3 (1991), pp. 1–28.
[16] P. J. Swart and P. J. Holmes, *Energy minimization and the formation of microstructure in dynamic anti-plane shear*, Arch. Ration. Mech. Anal., 121 (1992), pp. 37–85.
[17] C. Lattanzio and A. E. Tzavaras, *Structural properties of stress relaxation and convergence from viscoelasticity to polyconvex elastodynamics*, Arch. Ration. Mech. Anal., 180 (2006), pp. 449–492.
[18] R. Temam, *Navier-Stokes Equations—Theory and Numerical Analysis*, AMS Chelsea Publishing, Providence, RI, 2001, reprinted version.

SIAM J. NUMER. ANAL.
Vol. 46, No. 5, pp. 2484–2504
© 2008 Society for Industrial and Applied Mathematics

AN ANALYSIS OF A FETI–DP ALGORITHM ON IRREGULAR SUBDOMAINS IN THE PLANE*

AXEL KLAWONN†, OLIVER RHEINBACH†, AND OLOF B. WIDLUND‡

Abstract. In the theory for domain decomposition algorithms of the iterative substructuring family, each subdomain is typically assumed to be the union of a few coarse triangles or tetrahedra. This is an unrealistic assumption, in particular if the subdomains result from the use of a mesh partitioner, in which case they might not even have uniformly Lipschitz continuous boundaries. The purpose of this study is to derive bounds for the condition number of these preconditioned conjugate gradient methods which depend only on a parameter in an isoperimetric inequality, two geometric parameters characterizing John and uniform domains, and the maximum number of edges of any subdomain. A related purpose is to explore to what extent well-known technical tools previously developed for quite regular subdomains can be extended to much more irregular subdomains. Some of these results are valid for any John domain, while an extension theorem, which is needed in this study, requires that the subdomains have complements which are uniform. The results, so far, are complete only for problems in two dimensions. Details are worked out for a FETI–DP algorithm and numerical results support the findings. Some of the numerical experiments illustrate that care must be taken when selecting the scaling of the preconditioners in the case of irregular subdomains.

Key words. domain decomposition, preconditioners, iterative substructuring, dual-primal FETI, John and uniform domains, fractal subdomains

AMS subject classifications. 65F10, 65N30, 65N55

DOI. 10.1137/070688675

1. Introduction.

In the theory of domain decomposition methods of iterative substructuring type, we typically assume that each subdomain is quite regular, e.g., the union of a small set of coarse triangles or tetrahedra; see, e.g., [36, Assumption 4.3]. However, such an assumption is unlikely to hold especially if the subdomains result from using a mesh partitioner, in which case the subdomain boundaries might not even be uniformly Lipschitz continuous, i.e., the number of patches which cover $\partial\Omega$, and in each of which the boundary is the graph of a Lipschitz continuous function, might not be uniformly bounded independently of the finite element mesh size. We also note that the shape of the subdomains is likely to change if the mesh size is altered and a mesh partitioner is used. The purpose of this paper is to develop a theory for domain decomposition methods under much weaker assumptions on the partitioning and to categorize the rate of convergence in terms of a few geometric parameters. We will denote the nonoverlapping subdomains by Ω_i and the interface between them by Γ.

So far, complete results have been obtained only for problems in the plane. To simplify our presentation, we will also focus on scalar elliptic problems of the following

*Received by the editors April 18, 2007; accepted for publication (in revised form) February 22, 2008; published electronically June 11, 2008.
http://www.siam.org/journals/sinum/46-5/68867.html

†Fachbereich Mathematik, Universität Duisburg-Essen, Universitätsstr. 3, 45177 Essen, Germany (axel.klawonn@uni-due.de, oliver.rheinbach@uni-due.de, http://www.numerik.uni-due.de). The work of these authors was supported in part by the Deutsche Forschungsgemeinschaft (DFG) under research grant KL 2094/1-1.

‡Courant Institute, 251 Mercer Street, New York, NY 10012 (widlund@cims.nyu.edu, http://www.cs.nyu.edu/cs/faculty/widlund). This author's work was supported in part by the U.S. Department of Energy under contracts DE-FG02-06ER25718 and DE-FC02-01ER25482 and in part by National Science Foundation grant DMS-0513251.

form:

$$-\mathrm{div}\,(\rho(x)\nabla u(x)) = f(x) \quad x \in \Omega, \tag{1.1}$$

with a homogeneous Dirichlet boundary condition on a measurable part, $\partial\Omega_D$, of $\partial\Omega$, the boundary of Ω, and a Neumann boundary condition on its complement $\partial\Omega_N$. The coefficient $\rho(x)$ is strictly positive and assumed to be a constant ρ_i for $x \in \Omega_i$. As for many other algorithms, our results hold equally well for compressible elasticity problems in two dimensions; see [36, Chapter 8] for a discussion, [24] for further details on FETI–DP for linear elasticity, and the discussion below.

We use linear continuous finite elements and triangulations with shape regular triangles, i.e., the diameter of an element is bounded uniformly by a constant times the radius of the largest inscribed circle. We assume that each subdomain is a union of elements. We denote the finite element space on Ω by $W^h(\Omega)$ and assume that all elements of $W^h(\Omega)$ vanish on $\partial\Omega_D$.

The analysis will be carried out for a FETI–DP algorithm but the results immediately carry over to the corresponding BDDC algorithm (see, e.g., [26]) as well as to a class of inexact FETI–DP methods; see [22]. The tools that we will develop can also be used in the analysis of a number of other domain decomposition algorithms, in particular, those of [12, 13, 14]; our study originated within this collaborative project with Clark Dohrmann. For a collection of auxiliary results used in the analysis of iterative substructuring algorithms, in the case of regular subdomains, see [36, section 4.6].

Thus, our study requires the generalization of some technical tools used in the proof of bounds of the convergence rate of the FETI–DP algorithm. We also have to modify the basic line of reasoning in the proof of our main result; a proof for triangular subdomains and constant coefficients was first given in [29]; see also [32, Chapter 2] for a proof more similar to ours. Four auxiliary results, namely a Poincaré inequality, a Sobolev-type inequality for finite element functions, bounds for certain edge terms, and a finite element extension theorem, will be required in our proof; see Lemmas 2.3, 4.3, 4.4, and 4.5. We will work with *John domains* and *uniform domains* (see Definitions 2.1 and 2.4); the latter are also known as Jones or (ε, ∞)-domains. We will express our bounds on the rate of convergence of our algorithm in terms of a few parameters identified in Definitions 2.1 and 2.4 and Lemma 2.2.

We will also consider the problem of linear, isotropic elasticity in two dimensions:

$$-\mathrm{div}\,(2\mu\varepsilon(\mathbf{u}) + \lambda \mathrm{tr}\,(\varepsilon(\mathbf{u}))I) = f \quad \text{in } \Omega \subset \mathbb{R}^2. \tag{1.2}$$

We assume zero Dirichlet conditions on one part of the boundary of $\partial\Omega$ with nonvanishing measure and traction conditions on the remaining part of the boundary. FETI–DP methods have been analyzed for linear elasticity and regular subdomains in [24]; see [21] and [23] for numerical results.

A crucial tool in any such study is the second Korn inequality, Lemma 2.6; for a proof for uniform domains, see Durán and Muschietti [16], and for John domains, see Acosta, Durán, and Muschietti [1]. The proof in [16] has many details in common with Jones' proof of extension theorems for Sobolev spaces, [20], and the constant in the inequality depends only on the same geometric parameters as the result of Lemma 2.5.

The remainder of this article is organized as follows. In section 2, we provide a Poincaré inequality for general domains and introduce John and uniform domains. In section 3, we describe our FETI–DP algorithm. In section 4, we develop a convergence theory for our FETI–DP algorithm for subdomains with complements which

are uniform. Finally, in section 5, we present numerical results which confirm our theoretical findings. We note that we find a striking difference in the performance of the iterative method between the scaling based on our theory and another based on diagonal elements of the stiffness matrix in cases when the subdomain boundaries are highly irregular.

2. A Poincaré inequality, John and uniform domains. We will first introduce John domains and then consider a Poincaré inequality for such domains. Subsequently, we will introduce uniform domains; the latter are needed in order to prove a finite element extension theorem. We note that this final assumption on the subdomains is not required in our analysis of domain decomposition algorithms based on overlapping subdomains; see [14, 13, 12].

We next give the definition of a John domain; see [19] and the references therein. In the proofs of several of our auxiliary results, we will assume that the subdomains belong to this class.

DEFINITION 2.1 (John domain). *A domain $\Omega \subset \mathbb{R}^n$—an open, bounded, and connected set—is a John domain if there exists a constant $C_J \geq 1$ and a distinguished central point $x_0 \in \Omega$ such that each $x \in \Omega$ can be joined to it by a rectifiable curve $\gamma : [0,1] \to \Omega$ with $\gamma(0) = x_0,\ \gamma(1) = x$, and $|x - \gamma(t)| \leq C_J \cdot \mathit{distance}(\gamma(t), \partial\Omega)$ for all $t \in [0,1]$.*

This condition can be viewed as a twisted cone condition. We note that certain snowflake curves with fractal boundaries are John domains (see section 5) and that the boundary of a John domain can be arbitrarily much longer than its diameter.

We note that for any choice of the point x_0, there is a point $x \in \Omega$ at a distance of at least $\mathit{diameter}(\Omega)/2$. We find that $\mathit{diameter}(\Omega) \leq 2C_J r_\Omega$, where r_Ω is the radius of the largest ball inscribed in Ω and centered at x_0. Conditions on the boundary are of course also imposed.

In any analysis of any domain decomposition method with a second, coarse level, we need a Poincaré inequality. This inequality is closely related to an isoperimetric inequality. The next lemma is due to Maz'ja [30] and Federer and Fleming [18]; see also Lin and Yang [27, Theorem 5.3.2] or Maz'ja [31].

LEMMA 2.2 (isoperimetric inequality). *Let $\Omega \subset \mathbb{R}^n$ be a domain, and let u be sufficiently smooth. Then,*

$$\inf_{c \in R} \left(\int_\Omega |u - c|^{n/(n-1)}\, dx \right)^{(n-1)/n} \leq \gamma(\Omega, n) \int_\Omega |\nabla u|\, dx$$

if and only if

$$[\min(|A|, |B|)]^{1-1/n} \leq \gamma(\Omega, n) |\partial A \cap \partial B|. \tag{2.1}$$

Here, $A \subset \Omega$ is an arbitrary open set, and $B = \Omega \setminus \bar{A}$; $\gamma(\Omega, n)$ is the best possible constant, and $|A|$ is the measure of the set A, etc.

We note that the domain does not need to be star-shaped or Lipschitz. For $n = 2$, the best choice of $c = \bar{u}_\Omega$, the average of u over the domain. A small value of $\gamma(\Omega, n)$ is desirable for our purposes.

It is known that any simply connected plane domain with a finite Poincaré parameter $\gamma(\Omega, 2)$ is a John domain; see [9]. It is also known (see [6]) that any John domain has a bounded Poincaré parameter $\gamma(\Omega, n)$. Since we have been unable to find a simple estimate of $\gamma(\Omega, n)$ in terms of C_J in the literature, we have chosen to express our results in terms of both of these geometric parameters. In two dimensions, we

immediately obtain the standard Poincaré inequality by using the Cauchy–Schwarz inequality.

LEMMA 2.3 (Poincaré's inequality). *Let Ω be a John domain in the plane. Then,*

$$\|u - \bar{u}_\Omega\|^2_{L_2(\Omega)} \leq (\gamma(\Omega, 2))^2 |\Omega| \|\nabla u\|^2_{L_2(\Omega)} \quad \forall u \in H^1(\Omega).$$

For $n = 3$ such a bound is obtained by using Hölder's inequality several times. In Lemma 2.3, we then should replace $|\Omega|$ by $|\Omega|^{2/3}$.

Throughout, we will use a weighted $H^1(\Omega_i)$-norm defined by

$$\|u\|^2_{H^1(\Omega_i)} := \int_{\Omega_i} \nabla u \cdot \nabla u dx + 1/H_i^2 \int_{\Omega_i} |u|^2 dx = |u|^2_{H^1(\Omega_i)} + 1/H_i^2 \int_{\Omega_i} |u|^2 dx.$$

Here H_i is the diameter of Ω_i. The weight for the L_2-term results from the standard H^1-norm on a domain with diameter one and a dilation. We will use Lemma 2.3 to remove L_2-terms in our estimates.

We finally consider uniform domains also known as (ϵ, ∞) domains. It is known, and easy to see, that any uniform domain is a John domain. It is also easy to construct John domains that are not uniform domains. According to Jones [20, Theorem 4], they form the largest class of finitely connected domains for which an extension theorem holds in two dimensions; see Lemma 2.5 below. According to [20, Theorem 2], every uniform domain in $\mathbb{R}^n$ allows for a bounded extension with respect to the seminorm of H^1. It also follows from [20, Theorem C] that if a simply connected domain, with a simply connected complement, is uniform, so is its complement.

DEFINITION 2.4 (uniform domains). *A domain $\Omega \subset \mathbb{R}^n$ is a uniform domain if there exists a constant C_U such that any pair of points $x_1 \in \Omega$ and $x_2 \in \Omega$ can be joined by a rectifiable curve $\gamma(t) : [0, 1] \to \Omega$ with $\gamma(0) = x_1$, $\gamma(1) = x_2$, and where the Euclidean arc length of $\gamma \leq C_U |x_1 - x_2|$ and $\min_{i=1,2} |x_i - \gamma(t)| \leq C_U \cdot$ distance$(\gamma(t), \partial\Omega)$ for all $t \in [0, 1]$.*

Remark 1. The complement of a bounded set is of course always infinite. What matters in our work, with any uniform domain, is its boundary. We can therefore, whenever required, remove the part of an infinite domain that is outside a sufficiently large circle and effectively work only with bounded domains.

We note that domains with boundaries formed by von Koch curves, as in section 5, are also uniform domains.

We will use a main result for uniform domains; see Jones [20, Theorem 2].

LEMMA 2.5. *Let $\Omega \subset \mathbb{R}^n$ be a uniform domain, and let P_0 be the space of constants. There then exists a bounded, linear operator $E_\Omega : H^1(\Omega)/P_0 \to H^1(\mathbb{R}^n)/P_0$, which extends any element in $H^1(\Omega)$ to one defined for all of $\mathbb{R}^n$, i.e., $(E_\Omega u)_{|\Omega} = u$ for all $u \in H^1(\Omega)/P_0$. The norm of this operator depends only on $C_U(\Omega)$.*

As previously noted, for a result on linear elasticity, we also need a second Korn inequality for uniform domains; see Durán and Muschietti [16, Corollary 2.8].

LEMMA 2.6 (Korn inequality for uniform domains). *Let $\Omega \subset \mathbb{R}^n$ be a bounded uniform domain. Then, there exists a constant $C > 0$, which depends only on the uniformity constant $C_U(\Omega)$ and the dimension n, such that*

$$|\mathbf{u}|^2_{H^1(\Omega)} \leq C \|\varepsilon(\mathbf{u})\|^2_{L_2(\Omega)}$$

for all $\mathbf{u} \in \{\mathbf{u} \in \mathbf{H^1}(\Omega) : \int_\Omega (\frac{\partial u_i}{\partial x_j} - \frac{\partial u_j}{\partial x_i}) dx = 0,\ i, j = 1, \ldots, n\}$.

3. The FETI–DP algorithm. Let a domain $\Omega \subset \mathbb{R}^2$ be decomposed into N nonoverlapping subdomains Ω_i of diameter H_i, each of which is a union of finite

elements with the finite element nodes on the boundaries of neighboring subdomains matching across the interface $\Gamma := \bigcup_{i \neq j} \partial\Omega_i \cap \partial\Omega_j$. Here $\partial\Omega_i$ and $\partial\Omega_j$ are the boundaries of Ω_i and Ω_j, respectively. The interface Γ is the union of edges and vertices. The nodes on an edge are shared by exactly two subdomains, and the edges are open subsets of Γ. The vertices are endpoints of the edges. For a more detailed definition of faces, edges, and vertices in two and three dimensions, see [24, section 3] and [21, section 2].

For each subdomain Ω_i, $i = 1, \ldots, N$, we assemble the subdomain stiffness matrices $K^{(i)}$ and load vectors $f^{(i)}$ from the contributions of the individual elements. We denote the vector of unknowns of the subdomain Ω_i by $u^{(i)}$ and then partition the unknowns of $u^{(i)}$ into a vector of primal variables $u_\Pi^{(i)}$ and a vector of nonprimal variables $u_B^{(i)}$. We can obtain fast convergence by choosing only vertices as primal variables, since we are considering only two dimensional problems; see [17, 29, 32]. To simplify the description of the algorithm, we will assume that all vertices are primal. The nonprimal variables are those of $u_I^{(i)}$, associated with the interior nodes of the subdomain, and the dual displacement variables of the vector $u_\Delta^{(i)}$, associated with the remaining edge nodes. We will enforce the continuity of the solution in the primal unknowns of $u_\Pi^{(i)}$ by making them global; we subassemble the subdomain stiffness matrices $K^{(i)}$ with respect to this set of variables and denote the resulting matrix by $\widetilde{K}$. For all other interface variables, i.e., those of the $u_\Delta^{(i)}$, we will introduce Lagrange multipliers to enforce continuity.

Here are some more details: we partition the stiffness matrices according to the different sets of unknowns and obtain

$$K^{(i)} = \begin{bmatrix} K_{BB}^{(i)} & K_{\Pi B}^{(i)\,T} \\ K_{\Pi B}^{(i)} & K_{\Pi\Pi}^{(i)} \end{bmatrix}, \quad K_{BB}^{(i)} = \begin{bmatrix} K_{II}^{(i)} & K_{\Delta I}^{(i)\,T} \\ K_{\Delta I}^{(i)} & K_{\Delta\Delta}^{(i)} \end{bmatrix},$$

$$\text{and} \qquad f^{(i)T} = [f_B^{(i)T}\ f_\Pi^{(i)T}], \quad f_B^{(i)^T} = [f_I^{(i)^T}\ f_\Delta^{(i)^T}].$$

Next we define the block diagonal matrices

$$K_{BB} = \operatorname{diag}_{i=1}^N(K_{BB}^{(i)}), \quad K_{\Pi B} = \operatorname{diag}_{i=1}^N(K_{\Pi B}^{(i)}), \quad K_{\Pi\Pi} = \operatorname{diag}_{i=1}^N(K_{\Pi\Pi}^{(i)}),$$

and load vectors $f_B^T = [f_B^{(1)\,T}, \ldots, f_B^{(N)\,T}]$, $f_\Pi^T = [f_\Pi^{(1)\,T}, \ldots, f_\Pi^{(N)\,T}]$.

Assembling the local subdomain matrices and load vectors with respect to the primal variables, we obtain the partially assembled global stiffness matrix $\widetilde{K}$ and the load vector $\tilde{f}$,

$$\widetilde{K} = \begin{bmatrix} K_{BB} & \widetilde{K}_{\Pi B}^T \\ \widetilde{K}_{\Pi B} & \widetilde{K}_{\Pi\Pi} \end{bmatrix}, \qquad \tilde{f} = \begin{bmatrix} f_B \\ \tilde{f}_\Pi \end{bmatrix},$$

where a tilde marks an assembled quantity. It is easy to see that, with all vertices primal, the matrix $\widetilde{K}$ is positive definite.

To enforce the continuity of the remaining interface variables $u_\Delta^{(i)}$, we introduce a jump operator B_B with entries $0, -1$, or 1 and a vector of Lagrange multipliers λ. The FETI–DP saddlepoint problem is then given by

$$\begin{bmatrix} K_{BB} & \widetilde{K}_{\Pi B}^T & B_B^T \\ \widetilde{K}_{\Pi B} & \widetilde{K}_{\Pi\Pi} & 0 \\ B_B & 0 & 0 \end{bmatrix} \begin{bmatrix} u_B \\ \tilde{u}_\Pi \\ \lambda \end{bmatrix} = \begin{bmatrix} f_B \\ \tilde{f}_\Pi \\ 0 \end{bmatrix}. \tag{3.1}$$

By eliminating u_B and $\tilde{u}_\Pi$ from (3.1), we obtain

$$F\lambda = d, \tag{3.2}$$

with

$$F = B_B K_{BB}^{-1} B_B^T + B_B K_{BB}^{-1} \widetilde{K}_{\Pi B}^T \widetilde{S}_{\Pi\Pi}^{-1} \widetilde{K}_{\Pi B} K_{BB}^{-1} B_B^T, \quad \widetilde{S}_{\Pi\Pi} = \widetilde{K}_{\Pi\Pi} - \widetilde{K}_{\Pi B} K_{BB}^{-1} \widetilde{K}_{\Pi B}^T,$$

$$\text{and} \qquad d = B_B K_{BB}^{-1} f_B - B_B K_{BB}^{-1} \widetilde{K}_{\Pi B}^T \widetilde{S}_{\Pi\Pi}^{-1} (\widetilde{f}_\Pi - \widetilde{K}_{\Pi B} K_{BB}^{-1} f_B).$$

To define the Dirichlet preconditioner M, we need to introduce a scaled variant of the jump operator B_B, which we denote by $B_{B,D} = [B_{B,D}^{(1)}, \ldots, B_{B,D}^{(N)}]$. The matrix $B_{B,D}^{(i)}$ is associated with the subdomain Ω_i and defined as follows: each row of $B_B^{(i)}$ with a nonzero entry corresponds to a Lagrange multiplier connecting the subdomain Ω_i with a neighboring subdomain Ω_j at a node $x \in \partial\Omega_{i,h} \cap \partial\Omega_{j,h}$. We obtain $B_{B,D}^{(i)}$ by multiplying each such row of $B_B^{(i)}$ by

$$\delta_i^\dagger := \frac{\rho_j}{\rho_j + \rho_i}; \tag{3.3}$$

see [24] or [36, section 6.4] for further details. We will refer to this scaling as the ρ-scaling. The Dirichlet preconditioner is then given by

$$M^{-1} = B_{B,D} (R_\Delta^B)^T S_{\Delta\Delta} R_\Delta^B B_{B,D}^T, \qquad \text{where}$$

$$S_{\Delta\Delta} = \operatorname{diag}_{i=1}^N (S_{\Delta\Delta}^{(i)}),\ S_{\Delta\Delta}^{(i)} = K_{\Delta\Delta}^{(i)} - K_{\Delta I}^{(i)} (K_{II}^{(i)})^{-1} K_{\Delta I}^{(i)T},\ \text{and } R_\Delta^B = \operatorname{diag}_{i=1}^N (R_\Delta^{B\,(i)}).$$

The matrices $R_\Delta^{B\,(i)}$ are restriction operators with entries 0 or 1 which restrict the nonprimal degrees of freedom $u_B^{(i)}$ of a subdomain to its dual part $u_\Delta^{(i)}$.

Our FETI–DP method is the preconditioned conjugate gradient algorithm applied to the symmetric, positive definite system (3.2), using the preconditioner M^{-1}. We note that, as usual for iterative substructuring methods, the local Schur complements $S_{\Delta\Delta}^{(i)}$ do not have to be built explicitly. Instead local linear systems are solved in each iteration step.

4. Convergence theory. We recall that $W^h(\Omega_i)$ is the standard finite element space of continuous, piecewise linear functions on Ω_i which vanish on $\partial\Omega_D$. The corresponding finite element trace spaces are denoted by $W^{(i)} := W^h(\partial\Omega_i \cap \Gamma), i = 1, \ldots, N$. The product space of the $W^{(i)}$ is denoted by W and its partially assembled subspace, with the primal variables global, is denoted by $\widetilde{W}$. We will denote by h_i the smallest diameter of the finite elements in the subdomain Ω_i.

We can now formulate our main result, which is also valid for compressible elasticity with Lamé parameters which are constant in each subdomain.

THEOREM 4.1 (condition number estimate). *Let the domain $\Omega \subset \mathbb{R}^2$ be partitioned into subdomains Ω_i, which are partitioned into shape regular elements and which have complements $\mathcal{C}\Omega_i$ that are uniform in the sense of Definition* 2.4. *Then, with M the Dirichlet preconditioner, F the FETI–DP operator, and with h_i the smallest diameter of any element in Ω_i, the condition number of the conjugate gradient method satisfies*

$$\kappa(M^{-1}F) \leq C \max_i (1 + \log(H_i/h_i))^2.$$

Here C is a constant which depends only on the parameters $C_J(\Omega_i)$ and $C_U(\mathcal{C}\Omega_i)$ of Definitions 2.1 *and* 2.4*, the Poincaré parameters $\gamma(\Omega_i, 2)$ of the subdomains, the maximum number of edges of any subdomain, and the shape regularity of the finite elements.*

We will denote by $\mathcal{H}$ the discrete harmonic extension operator; $\mathcal{H}(v)$ is the minimal energy extension of the restriction of the finite element function v to the interface Γ. For each edge $\mathcal{E}^{ij}$, we also need to define an edge cut-off function $\theta_{\mathcal{E}^{ij}}$, which is the discrete harmonic function which equals 1 at all nodes on the edge $\mathcal{E}^{ij}$, common to Ω_i and Ω_j, and which vanishes at all other interface nodes.

In order to prove Theorem 4.1, we need only establish the following result; see [24, section 8] for a proof of a three-dimensional counterpart for the case of tetrahedral subdomains. We denote the restriction operator from $\widetilde{W}$ to $W^{(i)}$ by $R^{(i)}, i = 1, \ldots, N$.

LEMMA 4.2. *Let $\mathcal{E}^{ij}$ be an edge common to the boundaries of Ω_i and Ω_j. For all $\tilde{w} \in \widetilde{W}$ and with $w^{(i)} := R^{(i)}w, w^{(j)} := R^{(j)}w$, we have*

$$\begin{aligned} \rho_i|\mathcal{H}(\theta_{\mathcal{E}^{ij}}\delta_i^\dagger(w^{(i)} - w^{(j)}))|^2_{H^1(\Omega_i)} \quad &\leq \quad C(1+\log(H_i/h_i))^2\rho_i|w^{(i)}|^2_{H^1(\Omega_i)} \\ &\quad + C(1+\log(H_j/h_j))^2\rho_j|w^{(j)}|^2_{H^1(\Omega_j)}. \end{aligned} \tag{4.1}$$

Remark 2. We note that when this lemma is used to prove Theorem 4.1, the maximum number of edges of any subdomain will enter the estimate.

To prove this lemma, we need three auxiliary results, in addition to Poincaré's inequality. The first is a discrete Sobolev inequality. This inequality, (4.2), is well known in the theory of iterative substructuring methods. Proofs for domains satisfying an interior cone condition are given in [7] and [8, section 4.9], and a different proof is given in [36, page 102]. For a proof for John domains, see [12].

LEMMA 4.3 (discrete Sobolev inequality). *Let $\Omega_i \subset \mathbb{R}^2$ be a John domain. Then*

$$\|u - \bar{u}_{\Omega_i}\|^2_{L_\infty(\Omega_i)} \leq C(1+\log(H_i/h_i))|u|^2_{H^1(\Omega_i)} \tag{4.2}$$

for all $u \in W^h(\Omega_i)$. The constant C depends only on the John parameter $C_J(\Omega_i)$ of Ω_i and the shape regularity of the finite elements.

Another important tool provides estimates of certain edge terms. For regular subdomains in two dimensions, this lemma was first given in [15].

LEMMA 4.4 (edge lemma). *Let Ω_i be a John domain, $\mathcal{E}^{ij} \subset \partial\Omega_i$ be an edge, and $\theta_{\mathcal{E}^{ij}} \in W^h(\Omega_i)$ be a finite element function which equals 1 at all nodes of $\mathcal{E}^{ij}$, vanishes at the other nodes on $\partial\Omega_i$, and is discrete harmonic in Ω_i. Then, for any $u \in W^h(\Omega_i)$, we have*

$$|\mathcal{H}(\theta_{\mathcal{E}^{ij}}u)|^2_{H^1(\Omega_i)} \leq C\,(1+\log(H_i/h_i))^2\|u\|^2_{H^1(\Omega_i)}, \tag{4.3}$$

$$|\theta_{\mathcal{E}^{ij}}|^2_{H^1(\Omega_i)} \leq C(1+\log(H_i/h_i)), \tag{4.4}$$

and

$$\|\theta_{\mathcal{E}^{ij}}\|^2_{L_2(\Omega_i)} \leq CH_i^2(1+\log(H_i/h_i)). \tag{4.5}$$

Here, C depends only on the John parameter $C_J(\Omega_i)$ of Ω_i and the shape regularity of the finite elements. The logarithmic factor in (4.5) *can be removed if all angles of the triangulation are acute.*

Proof. We will establish inequality (4.3) by using ideas similar to those of [36, Proofs of Lemmas 4.24 and 4.25.]. We will construct a function $\vartheta_{\mathcal{E}^{ij}}$ which has the

same boundary values as $\theta_{\mathcal{E}^{ij}}$ and which satisfies the two inequalities (4.3) and (4.4). Since $\theta_{\mathcal{E}^{ij}}$ and $\mathcal{H}(\theta_{\mathcal{E}^{ij}}u)$ are discrete harmonic, the two inequalities (4.3) and (4.4) will then hold. We note that $\vartheta_{\mathcal{E}^{ij}}$, which we will construct, will not be a finite element function, and that the same is true for $\vartheta_{\mathcal{E}^{ij}}u$. But it is easy to see that the gradient of the finite element interpolant of $\vartheta_{\mathcal{E}^{ij}}$ will satisfy the same bound as $\nabla\vartheta_{\mathcal{E}^{ij}}$. We then use [36, Lemma 4.31] to complete the argument for the linear interpolant of $\vartheta_{\mathcal{E}^{ij}}u$.

We first construct two curves, as in Definition 2.1, which connect the endpoints of the edge $\mathcal{E}^{ij}$ to the point x_0.

For each point of the curves, it is now possible to construct a circular disk, contained in Ω_i, with a radius as in Definition 2.1. We will replace this set by a union of rectangles contained in the union of a smaller family $\{B_k\}$ of disks creating two polygonal curves connecting the two endpoints of the edge. These two curves partition Ω_i into three parts. We will make $\vartheta_{\mathcal{E}_{ij}}$ identically 1 in the set next to the edge and make it vanish in the set farthest from the edge and we will bring the values of this function from 1 to 0 across the third, interior set.

The two John curves that meet in x_0 can be quite complicated. We will therefore replace them by the union of line segments which connect the centers of the family of circular disks $\{B_k\}$ for each of the John curves. These disks, in turn, will be used to construct the family of rectangles $\{R_{k,k+1}\}$. We are able to estimate the lengths of the sides of these rectangles as a function of the distance r to the closest endpoint of the edge $\mathcal{E}^{ij}$, and we will use the union of these rectangles to construct a function $\vartheta_{\mathcal{E}^{ij}}$ with a bound on its gradient which is inversely proportional to r.

The circular disks are constructed as in [12, section 3] in which a number of properties are also derived; they are collected in Lemma 3.3 of that paper. All of this material is essentially borrowed from [19]. We will recall some of these results.

Let $\gamma(t)$ be the chosen John curve from the central point $x_0 \in \Omega_i$ to x which is one of the endpoints of the edge $\mathcal{E}^{ij}$. A family of circular disks B_k, $k \geq 0$ is introduced as follows: the first disk is $B_0 := B(x_0, dist(x_0, \partial\Omega)/4)$, the open set centered at x_0 and with radius $r_0 = dist(x_0, \partial\Omega)/4$. The other disks are defined similarly as $B_k := B(x_k, r_k)$, centered at x_k and with radius $r_k := |x - x_k|/(4C_J), k \geq 1$. All the x_k lie on the John curve and all the B_k are subsets of Ω_i. Given B_k, we select $x_{k+1} \in \partial B_k$ as the last point of exit of $\gamma(t)$ from B_k.

The following properties are established in [12, Lemma 3.3]: there exists a constant $M(C_J)$ such that

1. no point in Ω belongs to more than M disks;
2. the number of x_k that are at a distance larger than r from x is bounded by $M\log(H/r)$;
3. $r_{k+1} \geq (4C_J - 1)/(4C_J + 1)\, r_k$.

Let the rectangle $R_{k,k+1}$ be the largest rectangle inscribed in the closure of $B_k \cup B_{k+1}$ and with two sides parallel to the line segment between x_k and x_{k+1}; see Figure 4.1. Since x_{k+1} is on ∂B_k, one side of the rectangle is of length $2r_k$. It is also easy to show, by using the lower bound on r_{k+1} given in 3. above, that the other is bounded from below by cr_k, where c is a positive constant independent of k. Therefore, the distance from any point on $[x_k, x_{k+1}]$ to the boundary of the rectangle exceeds $(c/2)r_k$, and the same is clearly true for the union of the rectangles.

We denote by $N(h)$ the index of the first x_k which lies in the union of the elements which have x, one of the endpoints of the edge $\mathcal{E}^{ij}$, as a vertex. We take the union of all rectangles $R_{k,k+1}$ with $k < N(h)$ and those elements next to x. We then triangulate the resulting region using only nodes on the boundary of this set. We assign the value

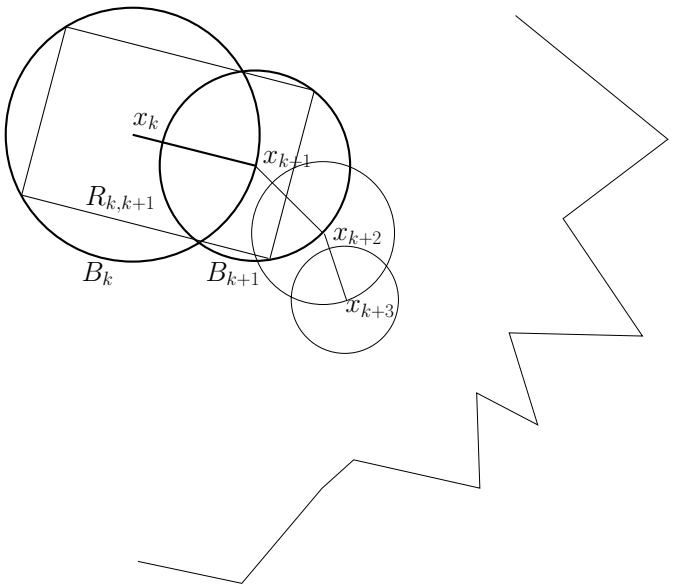


FIG. 4.1. *The circular disks B_k and B_{k+1}, their centers x_k and x_{k+1}, and the rectangle $R_{k,k+1}$.*

1 to $\vartheta_{\mathcal{E}^{ij}}$ on the side of this set which is closest to the edge $\mathcal{E}^{ij}$ and let this function vanish on the other side. In between, we use linear interpolation on the triangulation, just constructed, to define $\vartheta_{\mathcal{E}^{ij}}$.

For any rectangle $R_{k,k+1}$ and any triangulation of the domain, just constructed, we can estimate the gradient of $\vartheta_{\mathcal{E}^{ij}}$ by C/r_k, where C is constant, which depends only on C_J. Using Lemma 4.3, we then find that

$$|\vartheta_{\mathcal{E}^{ij}} u|^2_{H^1(R_{k,k+1})} \leq C(1+\log(H/h))\|u\|^2_{H^1(\Omega_i)}.$$

The contributions from the elements which have x as a vertex can be estimated quite easily. Inequality (4.4) now follows by using the upper bound $M\log(H/r)$ on the number of x_k.

The estimate of $|\vartheta_{\mathcal{E}^{ij}}|_{H^1(\Omega_i)}$ follows from the same type of arguments.

For the $L_2(\Omega_i)$-bound, we note that $\vartheta_{\mathcal{E}^{ij}} - \theta_{\mathcal{E}^{ij}}$ vanishes on the subdomain boundary, and that we can estimate the $L_2(\Omega_i)$-norm of $\vartheta_{\mathcal{E}^{ij}} - \theta_{\mathcal{E}^{ij}}$ in terms of the smallest eigenvalue of the Dirichlet problem for the Laplacian and the $H^1(\Omega_i)$-bound just established. The bound for $\theta_{\mathcal{E}^{ij}}$ then follows from the trivial bound $\|\vartheta_{\mathcal{E}^{ij}}\|_{L_2(\Omega_i)} \leq |\Omega_i|^{1/2}$.

If all angles of the triangulation are acute, then the finite element discretization will satisfy a maximum principle; see Ciarlet and Raviart [10]. Then $0 \leq \theta_{\mathcal{E}^{ij}} \leq 1$, and we can easily estimate the L_2-norm of $\theta_{\mathcal{E}^{ij}}$ in the same way as for $\vartheta_{\mathcal{E}^{ij}}$. ∎

The next lemma was proven for Lipschitz domains and quite general conforming finite elements in Widlund [37], using a technique by Astrakhatsev [2]; see also Toselli and Widlund [36] for a different proof. Here, we present a proof for more general domains by different means. We note that, in essence, this is a result, for a quite irregular interface, for the classical Dirichlet–Neumann algorithm for two subdomains as introduced in [36, section 1.3.3].

LEMMA 4.5 (extension lemma). *Let Ω_i and Ω_j, subsets of $\mathbb{R}^n$, be two subdomains with a common $(n-1)$-dimensional interface $\mathcal{E}^{ij}$. Furthermore, let Ω_i be a domain*

with a uniform complement, let

$$V_i^h = \{v_h \in W^h(\Omega_i) : v_h(x) = 0 \ \text{ at all nodes of } \ \partial\Omega_i \setminus \mathcal{E}^{ij}\},$$

and let

$$V_j^h = \{v_h \in W^h(\Omega_j) : v_h(x) = 0 \ \text{ at all nodes of } \ \partial\Omega_j \setminus \mathcal{E}^{ij}\}.$$

Then, there exists an extension operator

$$E_{ji}^h : V_j^h \longrightarrow V_i^h$$

with the following properties:

1. $(E_{ji}^h u_h)_{|\Omega_j} = u_h \quad \forall u_h \in V_j^h$,
2. $|E_{ji}^h u_h|_{H^1(\Omega_i)} \leq C|u_h|_{H^1(\Omega_j)} \quad \forall u_h \in V_j^h$,

where the constant $C > 0$ depends only on the uniformity parameter $C_U(\mathcal{C}\Omega_i)$ of the complement of Ω_i and the shape regularity of the finite elements and is otherwise independent of the finite element mesh sizes h_i and h_j and the diameters H_i and H_j.

Proof. The proof uses Lemma 2.5 and a result on finite element interpolation of nonsmooth functions given in Scott and Zhang [35]; see also Brenner and Scott [8, section 4.8]. Their results represent a refinement of well-known work by Clément [11].

From Lemma 2.5 and the fact that by assumption the complement of Ω_i is uniform, we know that there exists an extension operator

$$E_{\mathcal{C}\Omega_i} : H^1(\mathcal{C}\Omega_i) \longrightarrow H^1(\mathbb{R}^n)$$

and a positive constant C, which depends only on the uniformity constant $C_U(\mathcal{C}\Omega_i)$, such that

$$|E_{\mathcal{C}\Omega_i} u|_{H^1(\mathbb{R}^n)} \leq C\,|u|_{H^1(\mathcal{C}\Omega_i)} \quad \forall u \in H^1(\mathcal{C}\Omega_i), \tag{4.6}$$

with $(E_{\mathcal{C}\Omega_i} u)_{|\mathcal{C}\Omega_i} = u$ for $u \in H^1(\mathcal{C}\Omega_i)$.

Let $\Omega_{ij} = \Omega_i \cup \Omega_j \cup \mathcal{E}^{ij}$. It is shown in Scott and Zhang [35] that there exists a finite element interpolation operator

$$\Pi^h : H_0^1(\Omega_{ij}) \longrightarrow W^h(\Omega_{ij}) \cap H_0^1(\Omega_{ij}),$$

$$\text{with} \qquad \Pi^h v_h = v_h \quad \forall v_h \in W^h(\Omega_{ij}) \cap H_0^1(\Omega_{ij}), \tag{4.7}$$

$$\text{and} \qquad |\Pi^h v|_{H^1(\Omega_{ij})} \leq C\,|v|_{H^1(\Omega_{ij})} \quad \forall v \in H^1(\Omega_{ij}), \tag{4.8}$$

which preserves the boundary values of any finite element function on each individual element. The constant C is independent of the mesh parameters and H_i and H_j.

In their proof, Scott and Zhang construct a dual basis for the standard nodal basis of the finite element space in terms of integrals over elements or $(n-1)$-simplices, i.e., edges in two dimensions and faces in three dimensions. There is a great deal of flexibility and, in particular, for each node—and its associated degree or degrees of freedom—one can choose, in the construction, any element of which it is a vertex. For all nodes in $\overline{\Omega}_j$ we select elements of Ω_j and for the nodes in $\overline{\Omega}_i \setminus \mathcal{E}^{ij}$ elements of Ω_i. This will guarantee that the values in $\overline{\Omega}_j$ will not be changed when the operator Π^h is applied; cf. (4.7).

Any finite element function $u_h \in V_j^h$ can now be extended by zero outside of Ω_{ij}. This extended function, $\widetilde{u}$, has the same H^1-seminorm as u_h. We next define

$$E_{ji}^h u_h := \Pi^h(E_{\mathcal{C}\Omega_i}\widetilde{u}),$$

and obtain, by using (4.6) and (4.8),

$$|E_{ji}^h u_h|_{H^1(\Omega_i)} \leq C\,|E_{\mathcal{C}\Omega_i}\widetilde{u}|_{H^1(\mathbb{R}^n)} \leq C\,|\widetilde{u}|_{H^1(\mathcal{C}\Omega_i)} = C\,|u_h|_{H^1(\Omega_j)}.$$

The constant C depends only on the shape regularity of the finite elements and the uniformity constant $C_U(\mathcal{C}\Omega_i)$. By construction, we also have

$$(E_{ji}^h u_h)_{|\Omega_j} = u_h \quad \forall u_h \in V_j^h. \qquad \square$$

We now present a proof of Lemma 4.2.

Proof of Lemma 4.2. Consider an arbitrary $w \in \widetilde{W}$, and let $w^{(i)} := R^{(i)}w$ and $w^{(j)} := R^{(j)}w$. Our scaling factors (3.3) satisfy the following elementary estimate (see [24, Lemma 8.4])

$$\rho_i(\delta_i^\dagger)^2 = \rho_i \frac{\rho_j^2}{(\rho_i+\rho_j)^2} \leq \min\{\rho_i,\rho_j\}.$$

Proceeding as in [24] or [32], we find

$$\begin{aligned}
&\rho_i|\mathcal{H}(\theta_{\mathcal{E}^{ij}}\delta_i^\dagger(w^{(i)}-w^{(j)}))|^2_{H^1(\Omega_i)}\\
&= \rho_i|\mathcal{H}(\theta_{\mathcal{E}^{ij}}\delta_i^\dagger((w^{(i)}-\overline{w}_{i,\Omega_i})-(w^{(j)}-\overline{w}_{j,\Omega_j})+(\overline{w}_{i,\Omega_i}-\overline{w}_{j,\Omega_j})))|^2_{H^1(\Omega_i)}\\
&\leq 3\min\{\rho_i,\rho_j\}\Big(|\mathcal{H}(\theta_{\mathcal{E}^{ij}}(w^{(i)}-\overline{w}_{i,\Omega_i}))|^2_{H^1(\Omega_i)} + |\mathcal{H}(\theta_{\mathcal{E}^{ij}}(w^{(j)}-\overline{w}_{j,\Omega_j}))|^2_{H^1(\Omega_i)}\\
&\quad + |\theta_{\mathcal{E}^{ij}}(\overline{w}_{i,\Omega_i}-\overline{w}_{j,\Omega_j})|^2_{H^1(\Omega_i)}\Big).
\end{aligned}$$

We have added and subtracted the averages $\overline{w}_{i,\Omega_i} = \int_{\Omega_i} w^{(i)}dx/\int_{\Omega_i}1dx$ and $\overline{w}_{j,\Omega_j} = \int_{\Omega_j} w^{(j)}dx/\int_{\Omega_j}1dx$. We can estimate the first term above using Lemmas 2.3 and 4.4 and find that

$$\begin{aligned}
&\min\{\rho_i,\rho_j\}|\mathcal{H}(\theta_{\mathcal{E}^{ij}}(w^{(i)}-\overline{w}_{i,\Omega_i}))|^2_{H^1(\Omega_i)}\\
&\leq C(1+\log(H_i/h_i))^2\rho_i\|w^{(i)}-\overline{w}_{i,\Omega_i}\|^2_{H^1(\Omega_i)}\\
&\leq C(1+\log(H_i/h_i))^2\rho_i|w^{(i)}|^2_{H^1(\Omega_i)}.
\end{aligned}$$

We have the same type of bound for the second term. By using that the harmonic extension always has the least energy of all possible extensions, and Lemma 4.5, we obtain

$$\begin{aligned}
|\mathcal{H}(\theta_{\mathcal{E}^{ij}}(w^{(j)}-\overline{w}_{j,\Omega_j}))|^2_{H^1(\Omega_i)} &\leq |E_{ji}^h(\mathcal{H}(\theta_{\mathcal{E}^{ij}}(w^{(j)}-\overline{w}_{j,\Omega_j})))|^2_{H^1(\Omega_i)}\\
&\leq C\,|\mathcal{H}(\theta_{\mathcal{E}^{ij}}(w^{(j)}-\overline{w}_{j,\Omega_j}))|^2_{H^1(\Omega_j)}.
\end{aligned}$$

We can then proceed exactly as for the first term.

There remains to estimate the third term:

$$|\theta_{\mathcal{E}^{ij}}(\overline{w}_{i,\Omega_i} - \overline{w}_{j,\Omega_j})|^2_{H^1(\Omega_i)} = |\theta_{\mathcal{E}^{ij}}|^2_{H^1(\Omega_i)}|\overline{w}_{i,\Omega_i} - \overline{w}_{j,\Omega_j}|^2.$$

The energy of $\theta_{\mathcal{E}^{ij}}$ is estimated in Lemma 4.4. Adding and subtracting the common value $w_{i,\mathcal{V}^{ik}} = w_{j,\mathcal{V}^{ik}}$ at a primal vertex $\mathcal{V}^{ik}$, which is an endpoint of $\mathcal{E}^{ij}$, we find that

$$|\overline{w}_{i,\Omega_i} - \overline{w}_{j,\Omega_j}|^2 \leq 2|\overline{w}_{i,\Omega_i} - w_{i,\mathcal{V}^{ik}}|^2 + 2|\overline{w}_{j,\Omega_j} - w_{j,\mathcal{V}^{ik}}|^2.$$

For the first term on the right-hand side, we obtain, using the discrete Sobolev inequality (4.2),

$$\begin{aligned} |w_{i,\mathcal{V}^{ik}} - \overline{w}_{i,\Omega_i}|^2 &\leq \|w^{(i)} - \overline{w}_{i,\Omega_i}\|^2_{L_\infty(\Omega_i)} \\ &\leq C(1+\log(H_i/h_i))|w^{(i)}|^2_{H^1(\Omega_i)}. \end{aligned} \tag{4.9}$$

We can proceed in the same way with the second term. ∎

5. Numerical results. In this section, we present numerical results for some irregular (and regular) subdomains for the Poisson and linear elasticity problems in two dimensions using linear finite elements. In our FETI–DP algorithm, all subdomain vertices are primal and there are no additional primal variables. We solve the linear systems of equations by the conjugate gradient method until a relative residual reduction of rtol=$1e-10$ is reached. Our FETI–DP implementation [21] uses PETSc [4, 3, 5].

5.1. Geometries. We examine the effect of three different types of geometries. We first consider the case of the unit square, discretized using a structured mesh and decomposed into 4×4 subdomains. We consider a regular (see Figure 5.2 (upper)) and an irregular, ragged decomposition (see Figure 5.2 (lower)) of this mesh. The irregular decomposition is constructed in the following way. We start with quadrilateral elements and the unit square decomposed into square subdomains. We then reassign some elements to neighboring subdomains in the following way: all elements, except those touching the subdomain vertices, are reassigned in an alternating way. Each quadrilateral is then replaced by two triangles. Such ragged decompositions have previously been considered by Mandel, Dohrmann, and Tezaur; see, e.g., [28].

We then consider 3×3 subdomains and an unstructured mesh. We consider a standard, regular decomposition (see Figure 5.4) and one where the interface Γ is defined in a recursive way (see Figure 5.3) as a fractal snowflake. These subdomains are obtained by first partitioning the unit square into smaller squares. We then replace the middle third of each edge by two other edges of an equilateral triangle, increasing the length of the interface by a factor 4/3. The middle third of each of the resulting shorter edges is then replaced in the same way, and this process is repeated until the desired refinement level is reached. This procedure gives us an interface which is snowflake-like; see Figure 5.1.

5.2. Poisson's equation. We first consider the Poisson problem for our different geometries using homogeneous Dirichlet boundary conditions and different values of H/h. For unstructured meshes, we approximate H/h by $(\text{dof})^{1/2}$, the square root of the total number of degrees of freedom of a single subdomain. We compare the iteration numbers and the condition numbers for two different scalings. The ρ-scaling is the standard scaling as in (3.3) which uses the local coefficient of the PDE; see also [36, 25, 24]. The results of this paper are valid only for this type of scaling.

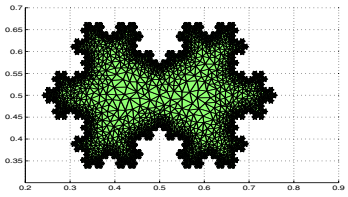


FIG. 5.1. *A snowflake-shaped subdomain, recursion level* 4.

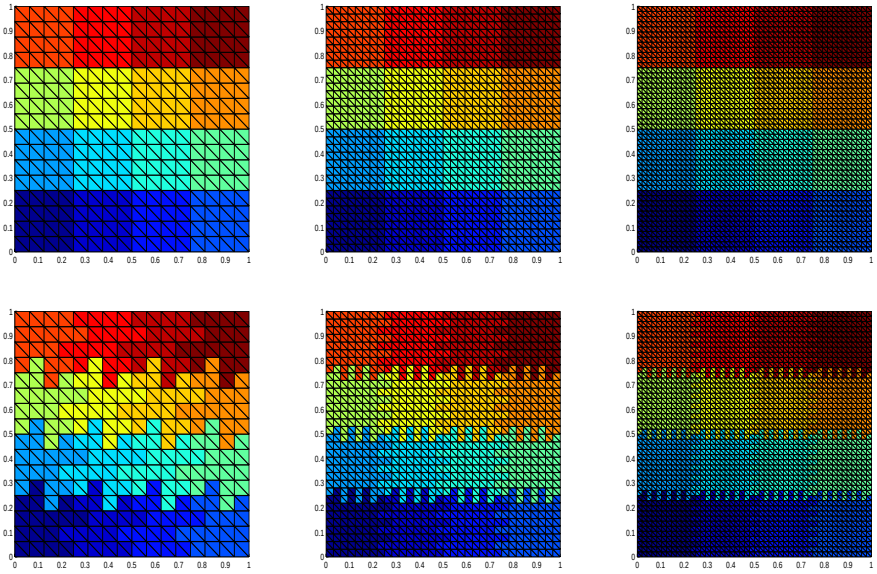


FIG. 5.2. *Upper: Structured mesh, regular subdomains,* $N = 16$, $H/h = 4, 8, 12$. *Lower: Structured mesh irregular (ragged) subdomains,* $N = 16$, $H/h = 4, 8, 12$.

Since we treat only problems with constant coefficients, the ρ-scaling reduces to the multiplicity scaling of [33, 34]. We can therefore describe this scaling as follows: the multiplicity of a node is the number of subdomains it belongs to. The multiplicity scaling then scales the contribution from and to each node by the inverse of the multiplicity.

The stiffness scaling, cf. [33, 34], refers to a common practice in iterative substructuring codes, which has been introduced for problems with variable coefficients to approximate the ρ-scaling. In the stiffness scaling, the diagonals of the local stiffness matrices are used to approximate the local coefficients. If we have constant coefficients, a structured mesh, and a regular decomposition into subdomains, then the stiffness scaling reduces to the multiplicity scaling, but this is not the case for an irregular decomposition.

Indeed, in the case of a structured mesh and a regular decomposition (see Figure 5.2 (upper)) we get the same results for the two types of scalings (see Table 5.1) and the expected $C(1+\log(H/h))^2$ behavior of the condition number (see Figure 5.6).

However, for an irregular decomposition as in Figure 5.2 (lower), a significant deterioration of the condition number is found and the rate of convergence for increasing

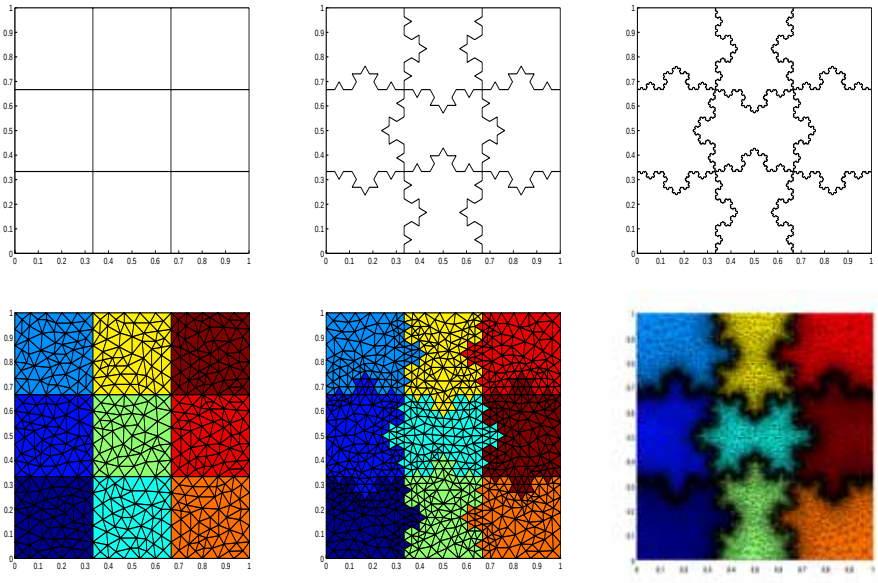


FIG. 5.3. *Unstructured mesh, irregular subdomains (snowflakes),* $N = 9$. *Level* 0, 2, *and* 4.

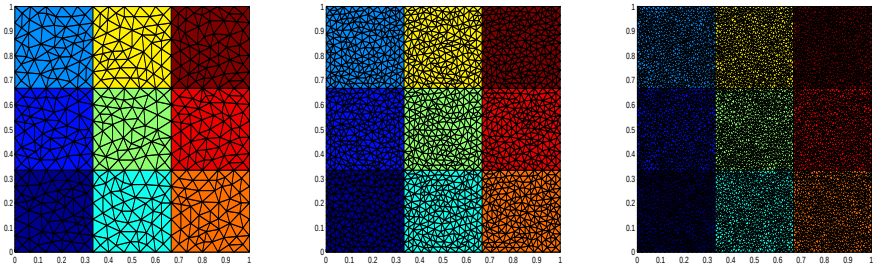


FIG. 5.4. *Unstructured mesh, regular subdomains,* $N = 9$. *Refinement level* 0, 1, *and* 2.

values of H/h; see Table 5.2 and Figure 5.5.

From Figure 5.5, we can see that for the ρ-scaling, we can again numerically confirm the expected $C(1 + \log(H/h))^2$ bound for the condition number. However, we see a linear growth in H/h of the condition number if the stiffness scaling is used. The growth of the condition number is also reflected in the iteration count.

In order to verify that the condition number is independent of the number of subdomains for a fixed H/h also in the irregular case, we choose $H/h = 16$ in Table 5.3 and increase the number of subdomains N. We see that the iteration count as well as the condition numbers are indeed uniformly bounded.

We then consider the case where we have an interface which is similar to a fractal snowflake. The finite element meshes are unstructured. We note that the number of nodes on the interface grows rapidly with the level of refinement of the snowflakes; cf. Tables 5.4 and 5.5. In spite of this, Tables 5.4 and 5.5, show that the condition number of the preconditioned FETI–DP operator only exceeds that of the trivial case by a factor of 3 resulting in a doubling of the number of conjugate gradient iterations

TABLE 5.1
Poisson: Structured mesh, regular subdomains. ρ-scaling versus stiffness scaling, $N = 16$.

Poisson: Structured regular						
			ρ		stiff	
H/h	total dof	$\|\Gamma\|$	it	cond	it	cond
4	289	84	4	1.63	4	1.63
8	1 089	180	5	2.22	5	2.22
16	4 225	372	6	2.96	5	2.96
32	16 641	756	7	3.84	7	3.84
64	66 049	1 524	7	4.85	7	4.85
128	263 169	3 060	8	6.02	8	6.02
192	591 361	4 596	9	6.76	9	6.76
256	1 050 625	6 132	8	7.31	8	7.31
320	1 640 961	7 668	11	7.75	11	7.75
368	2 362 369	9 204	11	8.13	11	8.13
448	3 214 849	10 740	11	8.45	11	8.45
512	4 198 401	12 276	10	8.73	10	8.73
576	5 313 025	13 812	11	8.99	11	8.99

TABLE 5.2
Poisson: Structured mesh, irregular subdomains. ρ-scaling versus stiffness scaling, $N = 16$.

Poisson: Structured irregular (ragged)						
			ρ		stiff	
H/h	total dof	$\|\Gamma\|$	it	cond	it	cond
4	289	180	19	4.95	15	3.21
8	1 089	468	23	6.50	18	5.27
16	4 225	1 044	24	7.45	21	9.75
32	16 641	2 196	25	8.47	25	20.36
64	66 049	4 500	26	9.61	32	44.74
128	263 169	9 108	27	10.89	41	100.07
192	591 361	13 716	27	11.70	48	159.88
256	1 050 625	18 324	27	12.32	54	222.50
320	1 640 961	22 932	28	12.82	60	287.20
368	2 362 369	27 540	28	13.24	65	353.54
448	3 214 849	32 148	28	13.60	70	421.25
512	4 198 401	36 756	28	13.92	74	490.13
576	5 313 025	41 364	28	14.20	77	560.02

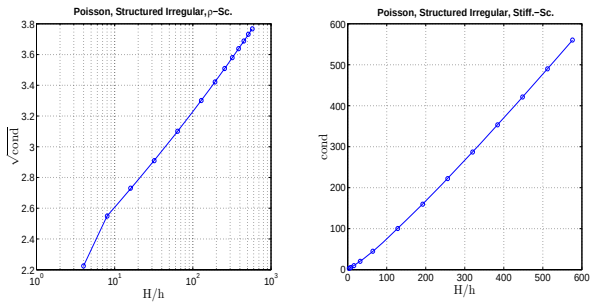


FIG. 5.5. *Poisson (Table 5.2): Structured irregular (ragged), $N = 16$. Left: ρ-scaling. Semilog Plot of $\sqrt{cond}$ versus H/h. Right: Stiffness scaling. Plot of cond versus H/h.*

if ρ-scaling is used. Figure 5.7 indicates that the square root of the condition number grows linearly with the logarithm of the number of degrees of freedom of individual subdomains. We also get such a bound for the unstructured regular case; see Table 5.5

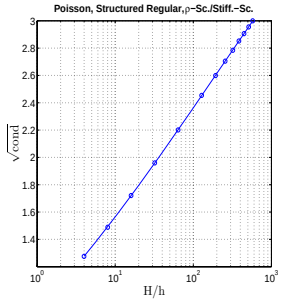


FIG. 5.6. *Poisson (Table 5.1): Structured regular, $N = 16$. Semilog plot of $\sqrt{cond}$ versus H/h. ρ-scaling and stiffness scaling give identical results.*

TABLE 5.3
Poisson: Structured regular/irregular, condition number, $H/h = 16$.

N			16	64	256	1 024	4 096
Regular	ρ	cond	2.96	3.28	3.35	3.38	3.38
		it	6	14	16	16	15
	stiff	cond	2.96	3.28	3.35	3.38	3.38
		it	6	14	16	16	15
Irregular	ρ	cond	7.45	8.01	8.13	8.12	8.10
		it	24	27	30	29	29
	stiff	cond	9.75	10.62	10.60	10.75	10.73
		it	21	27	31	31	30

TABLE 5.4
Poisson: Unstructured mesh, irregular subdomains, $N = 9$, ρ-scaling versus stiffness scaling.

Poisson: Unstructured irregular (snowflakes)						
			ρ		stiff	
level	total dof	$\|\Gamma\|$	it	cond	it	cond
0	498	56	9	1.90	9	1.89
1	528	92	14	2.66	12	2.80
2	841	188	16	3.38	21	7.49
3	4 166	764	19	5.15	37	37.58
4	19 160	3 068	21	7.37	70	190.40
5	81 340	12 284	23	10.08	128	940.87
6	331 895	49 148	25	13.29	240	4 494.98

TABLE 5.5
Poisson: Unstructured mesh, regular subdomains, $N = 9$, ρ-scaling versus stiffness scaling.

Poisson: Unstructured regular					
		ρ		stiff	
total dof	$\|\Gamma\|$	it	cond	it	cond
398	56	8	1.90	9	1.89
1 688	104	8	2.40	10	2.39
5 430	200	10	3.04	12	3.07
21 513	404	11	3.83	13	3.95
85 641	812	12	4.75	15	4.94
341 745	1 628	13	5.79	16	6.06

and Figure 5.8.

In contrast, in the case where stiffness scaling is applied, the condition number is superlinear in the square root of the number of degrees of freedom.

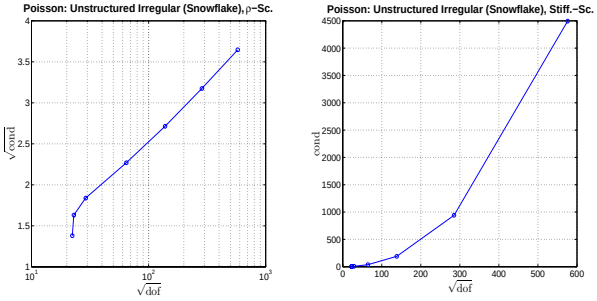


FIG. 5.7. *Poisson (Table 5.4): Unstructured irregular (snowflakes), $N = 9$. Left: ρ-scaling. Semilog plot of $\sqrt{cond}$ versus $\sqrt{dof}$. Right: Stiffness scaling. Plot of cond versus $\sqrt{dof}$.*

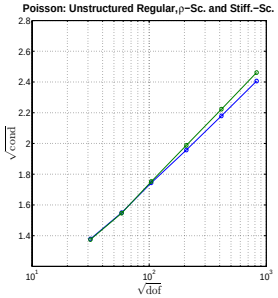


FIG. 5.8. *Poisson (Table 5.5): Unstructured regular, $N = 9$. Semilog plot of $\sqrt{cond}$ versus $\sqrt{dof}$. ρ-scaling (lower curve) and stiffness scaling (upper curve) give similar results.*

5.3. Elasticity. We have repeated all the experiments of section 5.2 for the problem of linear elasticity. We essentially get the same results; cf. Figures 5.9, 5.10, 5.11, 5.12, and Tables 5.6, 5.7, 5.8, 5.9, 5.10. We confirm the $C(1+\log(H/h))^2$ bound for all cases where the ρ-scaling is applied, and we see linear or even superlinear behavior in H/h of the condition number for the cases where the stiffness scaling is used with unstructured interfaces.

TABLE 5.6
Elasticity: Structured mesh, regular subdomains, ρ-scaling versus stiffness scaling, $N = 16$.

Elasticity: Structured regular						
			ρ		stiff	
H/h	total dof	$\|\Gamma\|$	it	cond	it	cond
4	578	168	12	2.11	12	2.11
8	2 178	360	14	2.97	14	2.97
16	8 450	744	17	4.03	17	4.03
32	33 282	1 512	20	5.30	20	5.30
64	132 098	2 048	22	6.78	22	6.78
128	526 338	6 120	24	8.73	24	8.72
192	1 182 722	9 192	26	10.32	26	10.32
256	2 101 250	12 264	27	11.35	27	11.35
320	3 281 922	15 336	29	12.19	29	12.19
368	4 724 738	18 408	30	12.90	30	12.90
448	6 429 698	21 480	31	13.52	31	13.52
512	8 396 802	24 552	30	14.07	30	14.07
576	10 626 050	27 624	32	14.56	32	14.56

TABLE 5.7
Elasticity: Structured mesh, irregular subdomains, ρ-scaling versus stiffness scaling, $N = 16$.

Elasticity: Structured irregular (ragged)						
			ρ		stiff	
H/h	total dof	$\|\Gamma\|$	it	cond	it	cond
4	578	360	39	19.42	27	10.98
8	2 178	936	60	33.93	42	19.45
16	8 450	2 088	66	38.87	51	33.11
32	33 282	4 392	69	41.69	61	65.72
64	132 098	9 000	70	45.06	77	139.13
128	526 338	18 216	71	48.87	103	299.05
192	1 182 722	27 432	72	51.29	121	468.50
256	2 101 250	36 648	72	53.07	138	644.49
320	3 281 922	45 864	73	54.52	156	825.07
368	4 724 738	55 080	73	55.71	169	1 009.14
448	6 429 698	64 296	74	56.77	180	1 196.07
512	8 396 802	73 512	74	57.67	195	1 385.60
576	10 626 050	82 728	74	58.50	204	1 577.28

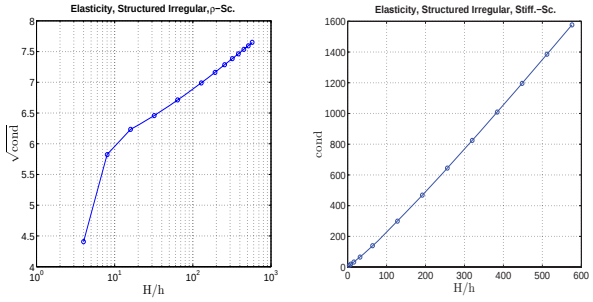


FIG. 5.9. *Elasticity (Table 5.7): Structured irregular (ragged), $N = 16$. Left: ρ-scaling. Semilog plot of $\sqrt{cond}$ versus H/h. Right: Stiffness scaling. Plot of cond versus H/h.*

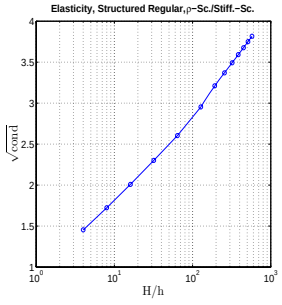


FIG. 5.10. *Elasticity (Table 5.6): Structured regular, $N = 16$. Semilog plot of $\sqrt{cond}$ versus H/h. ρ-scaling and stiffness scaling give identical results.*

TABLE 5.8
Elasticity: Structured regular/irregular, condition number, $H/h = 16$.

N			16	64	256	1 024	4 096
Regular	ρ	cond	4.03	4.98	5.76	6.20	6.40
		it	17	22	24	26	26
	stiff	cond	4.03	4.98	5.76	6.20	6.40
		it	17	22	24	26	26
Irregular	ρ	cond	38.87	43.41	47.77	50.28	51.60
		it	66	72	76	78	79
	stiff	cond	33.11	47.56	55.81	62.81	67.12
		it	51	69	80	85	89

TABLE 5.9
Elasticity: Unstructured mesh, irregular subdomains, $N = 9$, ρ-scaling versus stiffness scaling.

Elasticity: Unstructured irregular (snowflakes)						
			ρ		stiff	
level	total dof	$\|\Gamma\|$	it	cond	it	cond
0	996	112	13	2.35	13	2.38
1	1 056	184	24	5.94	19	4.04
2	1 682	376	32	9.08	33	14.75
3	8 332	1 528	38	13.20	70	94.40
4	38 320	6 136	42	18.69	143	520.77
5	162 680	24 568	47	26.05	287	2 733.43
6	663 790	98 296	52	35.12	564	13 434.80

TABLE 5.10
Elasticity: Unstructured mesh, regular subdomains, $N = 9$, ρ-scaling versus stiffness scaling.

Elasticity: Unstructured regular					
		ρ		stiff	
total dof	$\|\Gamma\|$	it	cond	it	cond
992	112	13	2.35	13	2.38
3 376	208	15	3.05	15	3.10
10 860	400	17	3.95	17	4.12
43 026	808	19	5.17	20	5.49
171 282	1 624	21	6.72	22	7.16
683 490	3 256	24	8.54	25	9.12

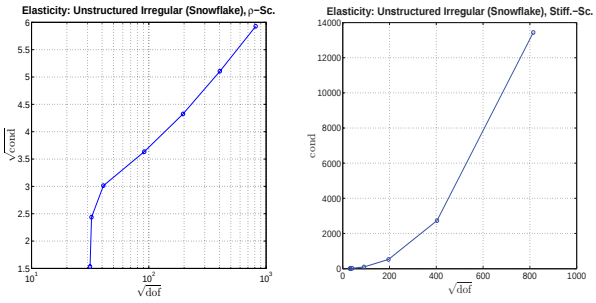


FIG. 5.11. *Elasticity (Table 5.9): Unstructured irregular (snowflakes), $N = 9$. Left: ρ-scaling. Semilog plot of $\sqrt{cond}$ versus $\sqrt{dof}$. Right: Stiffness scaling. Plot of cond versus $\sqrt{dof}$.*

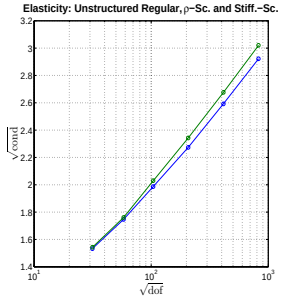


FIG. 5.12. *Elasticity (Table* 5.10*): Unstructured regular,* $N = 9$. *Semilog plot of* $\sqrt{cond}$ *versus* $\sqrt{dof}$. *ρ-scaling (lower curve) and stiffness scaling (upper curve) give similar results.*

Acknowledgments. The authors are grateful to Professor Fanghua Lin of the Courant Institute for introducing the third author to John domains and Poincaré's inequality for very general domains. They also wish to thank Professors Piotr Hajłasz and Marcus Sarkis for interesting comments on our work.

REFERENCES

[1] G. ACOSTA, R. G. DURÁN, AND M. A. MUSCHIETTI, *Solutions of the divergence operator on John domains*, Adv. Math., 206 (2006), pp. 373–401.

[2] G. P. ASTRAKHATSEV, *Method of fictitious domains for a second-order elliptic equation with natural boundary conditions*, U.S.S.R. Computational Math. and Math. Phys., 18 (1978), pp. 114–121.

[3] S. BALAY, K. BUSCHELMAN, V. EIJKHOUT, W. D. GROPP, D. KAUSHIK, M. G. KNEPLEY, L. C. MCINNES, B. F. SMITH, AND H. ZHANG, *PETSc users manual*, Tech. Report ANL-95/11—Revision 2.1.5, Argonne National Laboratory, Argonne, IL, 2004.

[4] S. BALAY, K. BUSCHELMAN, W. D. GROPP, D. KAUSHIK, M. G. KNEPLEY, L. C. MCINNES, B. F. SMITH, AND H. ZHANG, *PETSc Web page*, available online at http://www.mcs.anl.gov/petsc, 2001.

[5] S. BALAY, W. D. GROPP, L. C. MCINNES, AND B. F. SMITH, *Efficient management of parallelism in object oriented numerical software libraries*, in Modern Software Tools in Scientific Computing, E. Arge, A. M. Bruaset, and H. P. Langtangen, eds., Birkhäuser Press, Basel, Switzerland, 1997, pp. 163–202.

[6] B. BOJARSKI, *Remarks on Sobolev imbedding inequalities*, in Complex Analysis, Joensuu 1987, Lecture Notes in Math. 1351, Springer, Berlin, 1988, pp. 52–68.

[7] J. H. BRAMBLE, J. E. PASCIAK, AND A. H. SCHATZ, *The construction of preconditioners for elliptic problems by substructuring.* I, Math. Comp., 47 (1986), pp. 103–134.

[8] S. C. BRENNER AND L. R. SCOTT, *The Mathematical Theory of Finite Element Methods*, Texts in Applied Mathematics 15, 2nd ed., Springer-Verlag, New York, 2002.

[9] S. BUCKLEY AND P. KOSKELA, *Sobolev-Poincaré implies John*, Math. Res. Lett., 2 (1995), pp. 577–593.

[10] P. G. CIARLET AND P.-A. RAVIART, *Maximum principle and uniform convergence for the finite element method*, Comput. Methods Appl. Mech. Engrg., 2 (1973), pp. 17–31.

[11] P. CLÉMENT, *Approximation by finite element functions using local regularization*, Rev. Française Automat. Informat. Recherche Opérationnelle Sér. RAIRO Analyse Numérique, 9 (1975), pp. 77–84.

[12] C. R. DOHRMANN, A. KLAWONN, AND O. B. WIDLUND, *Domain decomposition for less regular subdomains: Overlapping Schwarz in two dimensions*, SIAM J. Numer. Anal., 46 (2008), pp. 2153–2168.

[13] C. R. DOHRMANN, A. KLAWONN, AND O. B. WIDLUND, *Extending theory for domain decomposition algorithms to irregular subdomains*, in Proceedings of the 17th International Conference on Domain Decomposition Methods in Science and Engineering, Strobl, Austria, 2006, U. Langer, M. Discacciati, D. Keyes, O. B. Widlund, and W. Zulehner, eds., Lect. Notes Comput. Sci. Engrg. 60, Springer-Verlag, 2007, pp. 255–261.

[14] C. R. Dohrmann, A. Klawonn, and O. B. Widlund, *A family of energy minimizing coarse spaces for overlapping Schwarz preconditioners*, in Proceedings of the 17th International Conference on Domain Decomposition Methods in Science and Engineering, Strobl, Austria, 2006, U. Langer, M. Discacciati, D. Keyes, O. B. Widlund, and W. Zulehner, eds., Lect. Notes Comput. Sci. Engrg. 60, Springer-Verlag, 2007, pp. 247–254.

[15] M. Dryja and O. B. Widlund, *Some domain decomposition algorithms for elliptic problems*, in Iterative Methods for Large Linear Systems, L. Hayes and D. Kincaid, eds., Academic Press, 1990, pp. 273–291.

[16] R. G. Durán and M. A. Muschietti, *The Korn inequality for Jones domains*, Electron. J. Differential Equations, 2004 (2004), pp. 1–10.

[17] C. Farhat, M. Lesoinne, P. Le Tallec, K. Pierson, and D. Rixen, *FETI-DP: A dual-primal unified FETI method—Part* I: *A faster alternative to the two-level FETI method*, Internat. J. Numer. Methods Engrg., 50 (2001), pp. 1523–1544.

[18] H. Federer and W. H. Fleming, *Normal and integral currents*, Ann. of Math. (2), 72 (1960), pp. 458–520.

[19] P. Hajłasz, *Sobolev inequalities, truncation method, and John domains*, in Papers on Analysis, Rep. Univ. Jyväskylä Dep. Math. Stat. 83, Univ. Jyväskylä, Jyväskylä, Finland, 2001, pp. 109–126.

[20] P. W. Jones, *Quasiconformal mappings and extendability of functions in Sobolev spaces*, Acta Math., 147 (1981), pp. 71–88.

[21] A. Klawonn and O. Rheinbach, *A parallel implementation of dual-primal FETI methods for three-dimensional linear elasticity using a transformation of basis*, SIAM J. Sci. Comput., 28 (2006), pp. 1886–1906.

[22] A. Klawonn and O. Rheinbach, *Inexact FETI-DP methods*, Internat. J. Numer. Methods Engrg., 69 (2007), pp. 284–307.

[23] A. Klawonn and O. Rheinbach, *Robust FETI-DP methods for heterogeneous three dimensional elasticity problems*, Comput. Methods Appl. Mech. Engrg., 196 (2007), pp. 1400–1414.

[24] A. Klawonn and O. B. Widlund, *Dual-primal FETI methods for linear elasticity*, Comm. Pure Appl. Math., 59 (2006), pp. 1523–1572.

[25] A. Klawonn, O. B. Widlund, and M. Dryja, *Dual-primal FETI methods for three-dimensional elliptic problems with heterogeneous coefficients*, SIAM J. Numer. Anal., 40 (2002), pp. 159–179.

[26] J. Li and O. B. Widlund, *FETI–DP, BDDC, and block Cholesky methods*, Internat. J. Numer. Methods Engrg., 66 (2006), pp. 250–271.

[27] F. Lin and X. Yang, *Geometric Measure Theory—An Introduction*, Advanced Mathematics, (Beijing/Boston), 1, Science Press, Beijing, 2002.

[28] J. Mandel, C. R. Dohrmann, and R. Tezaur, *An algebraic theory for primal and dual substructuring methods by constraints*, Appl. Numer. Math., 54 (2005), pp. 167–193.

[29] J. Mandel and R. Tezaur, *On the convergence of a dual-primal substructuring method*, Numer. Math., 88 (2001), pp. 543–558.

[30] V. G. Maz'ja, *Classes of domains and imbedding theorems for function spaces*, Soviet Math. Dokl., 1 (1960), pp. 882–885.

[31] V. G. Maz'ja, *Sobolev spaces*, Springer Series in Soviet Mathematics, Springer-Verlag, Berlin, 1985 (translated from the Russian by T. O. Shaposhnikova).

[32] O. Rheinbach, *Parallel Scalable Iterative Substructuring: Robust Exact and Inexact FETI-DP Methods with Applications to Elasticity*, Ph.D. thesis, Department of Mathematics, University of Duisburg-Essen, Essen, Germany, 2006.

[33] D. Rixen, *Substructuring and Dual Methods in Structural Analysis*, Ph.D. thesis, University of Liège, Liège, Belgium, 1997.

[34] D. Rixen and C. Farhat, *A simple and efficient extension of a class of substructure based preconditioners to heterogeneous structural mechanics problems*, Internat. J. Numer. Methods Engrg., 44 (1999), pp. 489–516.

[35] L. R. Scott and S. Zhang, *Finite element interpolation of nonsmooth functions satisfying boundary conditions*, Math. Comp., 54 (1990), pp. 483–493.

[36] A. Toselli and O. B. Widlund, *Domain decomposition methods—algorithms and theory*, in Springer Series in Computational Mathematics, 34, Springer-Verlag, Berlin, 2005.

[37] O. B. Widlund, *An extension theorem for finite element spaces with three applications*, pp. 110–122. Proceedings of the Second GAMM-Seminar on Numerical Techniques in Continuum Machanics, Kiel, Germany, 1986. W. Hackbusch and K. Witsch, eds., Friedr. Vieweg und Sohn, Braunschweig/Wiesbaden, 1987.

SIAM J. NUMER. ANAL.
Vol. 46, No. 5, pp. 2505–2523
© 2008 Society for Industrial and Applied Mathematics

COMPUTING A^α, $\log(A)$, AND RELATED MATRIX FUNCTIONS BY CONTOUR INTEGRALS*

NICHOLAS HALE†, NICHOLAS J. HIGHAM‡, AND LLOYD N. TREFETHEN†

Abstract. New methods are proposed for the numerical evaluation of $f(\mathbf{A})$ or $f(\mathbf{A})b$, where $f(\mathbf{A})$ is a function such as $\mathbf{A}^{1/2}$ or $\log(\mathbf{A})$ with singularities in $(-\infty, 0]$ and $\mathbf{A}$ is a matrix with eigenvalues on or near $(0, \infty)$. The methods are based on combining contour integrals evaluated by the periodic trapezoid rule with conformal maps involving Jacobi elliptic functions. The convergence is geometric, so that the computation of $f(\mathbf{A})b$ is typically reduced to one or two dozen linear system solves, which can be carried out in parallel.

Key words. Cauchy integral, conformal map, contour integral, matrix function, quadrature, rational approximation, trapezoid rule

AMS subject classifications. 65F30, 65D30

DOI. 10.1137/070700607

1. Introduction.

It is well known that an analytic function f of a square matrix $\mathbf{A}$ can be represented as a contour integral,

$$f(\mathbf{A}) = \frac{1}{2\pi i} \int_\Gamma f(z)\,(z\mathbf{I} - \mathbf{A})^{-1}\,dz, \tag{1.1}$$

where Γ is a closed contour lying in the region of analyticity of f and winding once around the spectrum $\sigma(\mathbf{A})$ in the counterclockwise direction. However, this idea has not often been exploited for numerical computation. Here we propose efficient ways to make use of (1.1) in the case where $\mathbf{A}$ has eigenvalues on or near the positive real axis $(0, \infty)$, and $f(z)$ is a function such as z^α or $\log z$ that is analytic apart from singularities or a branch cut on or near the negative real axis $(-\infty, 0]$. To be precise, it is not the matrix $f(\mathbf{A})$ that we usually compute but the vector $f(\mathbf{A})b$ for a given vector b. The first method we propose is to transplant the problem by a conformal map to an annulus and then apply the trapezoid rule, which converges geometrically (section 2). This procedure will be especially effective in cases where $\mathbf{A}$ is what might be called a "backslash matrix": a large sparse matrix for which systems of equations $(z\mathbf{I} - \mathbf{A})x = b$ can be solved efficiently by sparse direct methods but Krylov subspace iterations and Schur reduction to triangular form are impractical. An example is the matrix corresponding to the standard 5-point finite difference discretization of the Laplacian in two dimensions.

We then consider two modifications to this technique to take advantage of a particular structure of f: first, if f has a branch cut but no singularities on $(-\infty, 0)$ (section 3); second, if $f(z) = z^{1/2}$ (section 4). (Here and throughout this paper, the notations $z^{1/2}$, z^α, and $\log z$ refer to the standard branches of these functions,

*Received by the editors August 20, 2007; accepted for publication (in revised form) February 21, 2008; published electronically June 11, 2008.
http://www.siam.org/journals/sinum/46-5/70060.html

†Oxford Computing Laboratory, Wolfson Bldg., Parks Rd., Oxford OX1 3QD, UK (nich@comlab.ox.ac.uk, http://www.comlab.ox.ac.uk/nick.hale/, LNT@comlab.ox.ac.uk, http://web.comlab.ox.ac.uk/nick.trefethen/).

‡School of Mathematics, The University of Manchester, Manchester, M13 9PL, UK (higham@ma.man.ac.uk, http://www.ma.man.ac.uk/∼higham/). This author's work was supported by a Royal Society-Wolfson Research Merit Award.

and $\mathbf{A}^{1/2}$ and so on to the corresponding principal branches of the matrix functions.) Significant improvements can be achieved in such cases, including the avoidance of complex arithmetic.

All our methods based on quadrature formulas can be interpreted as making implicit use of rational approximations $f(z) \approx r(z)$ in regions of the complex plane. In section 5 we examine how these implicitly constructed rational functions compare with those obtained by solving a rational approximation problem directly. For the case $f(z) = z^{1/2}$, we find that the combination of our conformal map with the trapezoid rule reproduces a best rational approximation discovered by Zolotarev in 1877.

In section 6 we consider the effect of complex eigenvalues on the convergence of our methods, and section 7 examines three further numerical examples.

Before beginning, we shall slightly change the starting point. We have found that for the problems considered here a better way to obtain $f(\mathbf{A})$ is often by computing $\mathbf{A} \cdot \mathbf{A}^{-1} f(\mathbf{A})$, replacing (1.1) with the formula

$$f(\mathbf{A}) = \frac{\mathbf{A}}{2\pi i} \int_\Gamma z^{-1} f(z) \, (z\mathbf{I} - \mathbf{A})^{-1} \, dz. \tag{1.2}$$

The reason is that the dz factor becomes large as Γ swings out into the complex plane, and the factor z^{-1} counters this effect. For the first two methods derived in sections 2 and 3, the use of (1.2) instead of (1.1) typically improves accuracy by a digit or two. For the method of section 4, its use is essential since the contour passes through $z = \infty$.

For general information about the theory and computation of functions of matrices, see [11].

2. Method 1: Conformal map from an annulus. Let $\mathbf{A}$ be a real matrix whose eigenvalues lie in the interval $(0, \infty)$. In many applications $\mathbf{A}$ will be symmetric, but we do not require this. The assumption that $\mathbf{A}$ is real will save a factor of two by allowing trapezoid sums to exploit symmetry. The assumption that the eigenvalues are real can be relaxed; our methods do not degrade very much if the eigenvalues move slightly off the real axis, as illustrated in section 6.

Let m and M be the minimum and maximum eigenvalues of $\mathbf{A}$, respectively. If $\mathbf{A}$ is symmetric, or more generally normal, then M/m is its 2-norm condition number, whereas for general $\mathbf{A}$ the condition number may be larger. We assume at the outset that m and M are known, although in practice it would often be necessary first to estimate them. We also suppose that $\sigma(\mathbf{A})$ more or less fills $[m, M]$ in the sense that we do not attempt to exploit any gaps in the spectrum.

The assumption we make initially about f is that it is analytic in the slit complex plane $\mathbb{C}\backslash(-\infty, 0]$. This is true, for example, if $f(z) = \log z$ or $f(z) = z^\alpha = e^{\alpha \log z}$ (for any $\alpha \in \mathbb{R}$) for the standard branches. It is also true of more complicated functions such as $f(z) = \Gamma(z)$ or $f(z) = \tanh(z^{1/2})$.

One numerical approach to (1.2) is to surround $[m, M]$ by a circle in the right half-plane and apply the trapezoid rule on this circle to approximate the integral, as sketched in Figure 1. However, this method will be terribly inefficient if $\mathbf{A}$ is ill-conditioned, requiring $O(M/m)$ or more linear system solves to get any accuracy at all [4], because Γ is contained in only a very narrow annulus of analyticity, the annulus centered at $(M + m)/2$ with outer radius $(M + m)/2$ and inner radius $(M - m)/2$. Our strategy will be to improve the problem by a change of variables in the complex plane, which will bring the count down to $O(\log(M/m))$. The idea of combining the trapezoid rule with a change of variables was put forward for real integrals by Sag and Szekeres in 1964 [20] and Schwartz in 1969 [22] and developed more fully in the

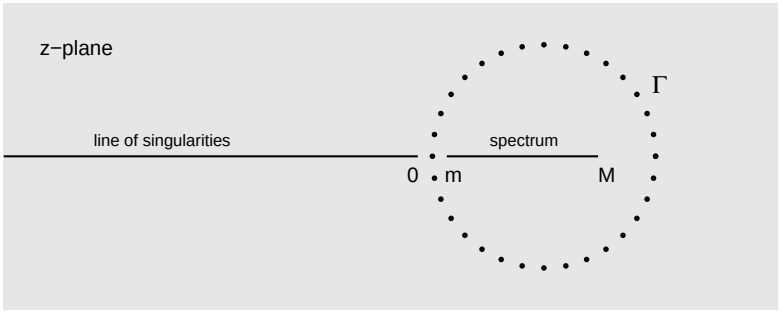


FIG. 1. *Depiction of the trapezoid rule applied directly to* (1.2)*, with the contour* Γ *taken as a circle. This method is highly inefficient if* $M \gg m$.

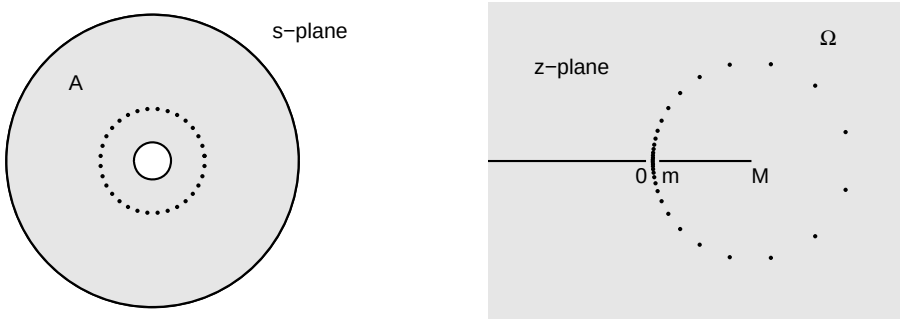


FIG. 2. *Conformal map for Method* 1. *First, transplant the entire doubly connected region of analyticity* Ω *conformally to an annulus* $\mathcal{A}$*, with the line of singularities mapping to the outer boundary circle and the interval containing the spectrum to the inner one. Then apply the trapezoid rule over a circle in the annulus. This will correspond to a nonuniform distribution of quadrature points in the z-plane. In this figure* $M/m = 8$ *and the annulus is quite thick. As* M/m *increases, it becomes thinner only logarithmically. The dots show quadrature nodes for* $N = 16$.

1970s by Iri, Moriguti, and Takasawa [12], Mori [17], Takahasi and Mori [28, 29], and Stenger [25, 26, 27], among others. For inverse Laplace transform contour integrals, such ideas were developed by Talbot in 1979 [30] and have been exploited by a variety of authors [10, 16, 23, 31], and the parameters have recently been optimized in various senses by Weideman and Trefethen [36], Weideman [35], and Trefethen, Weideman, and Schmelzer [33]. However, we are not aware of previous papers on such methods for the matrix function problems considered here.

Our aim is to transplant the region of analyticity of f and $(z\mathbf{I}-\mathbf{A})^{-1}$ conformally to an annulus

$$\mathcal{A} = \{z \in \mathbb{C} : r < |z| < R\}.$$

According to results going back in part to Poisson in the 1820s and first fully worked out by Davis in the 1950s [5], [6, section 4.6.5], the trapezoid rule applied over a circle within $\mathcal{A}$ will converge geometrically as $N \to \infty$, where N is the number of sample points. Moreover, this will be a thicker annulus than was available before the conformal map (Figure 1), which will make the convergence constant quite favorable.

This strategy is sketched in Figure 2. The region in question is the doubly connected set

$$\Omega = \mathbb{C}\backslash\big((-\infty, 0] \cup [m, M]\big).$$

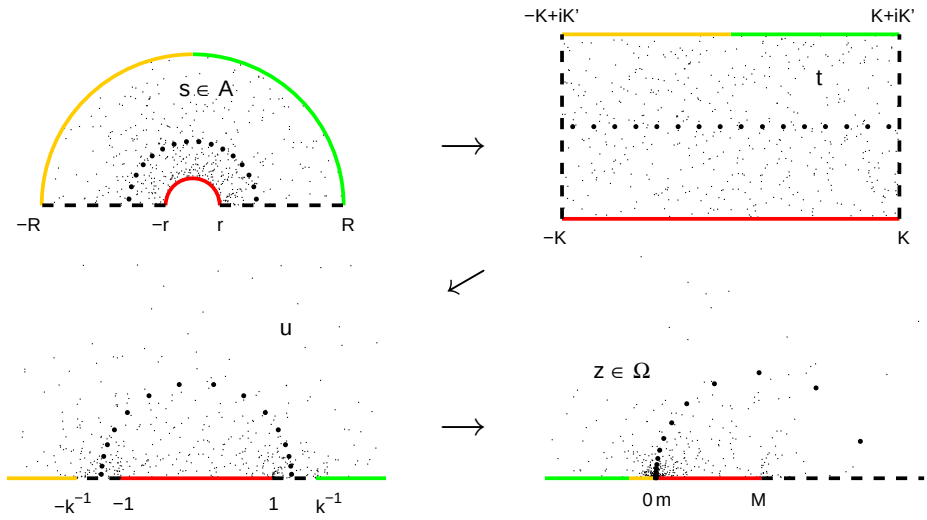


FIG. 3. *The map of Figure* 2, *constructed in three steps* $s \to t \to u \to z$. *Here* $m = 1$, $M = 20$. *The boundaries are shown in various colors and line styles together with* 500 *random interior points in the rectangle and a typical set of* $N = 16$ *quadrature points for the trapezoid rule in the* t*-plane. Notice how the narrow gap between* 0 *and* m *in the* z*-plane broadens to a comfortable ratio* $R/r \approx 5.8$ *in the annulus.*

We thus face a conformal mapping problem: How to map this domain Ω onto an annulus $\mathcal{A}$? More precisely it is the map from $\mathcal{A}$ to Ω that we shall need to compute. The radii R and r are not entirely at our disposal: The formulas below imply that R/r is determined by M/m.

The map can be carried out in three steps, as shown in Figure 3. First, the function

$$t = \frac{2Ki}{\pi} \log(-is/r)$$

carries the upper half of $\mathcal{A}$ to the rectangle with vertices $\pm K$ and $\pm K + iK'$. The numbers K and K' are complete elliptic integrals whose values are fixed by the next step of the conformal map. In this second step the Jacobi elliptic function

$$u = \mathrm{sn}(t) = \mathrm{sn}(t|k^2), \qquad k = \frac{\sqrt{M/m} - 1}{\sqrt{M/m} + 1}, \tag{2.1}$$

maps the rectangle to the upper half-plane, with the ends mapping to $[-k^{-1}, -1]$ and $[1, k^{-1}]$ (see [1, 9]). Finally, the Möbius transformation

$$z = \sqrt{mM} \left(\frac{k^{-1} + u}{k^{-1} - u} \right) \tag{2.2}$$

carries the upper half-plane to itself in such a way that $[-k^{-1}, -1]$ and $[1, k^{-1}]$ are sent to $[0, m]$ and $[M, \infty]$, respectively. We have now mapped the upper half of $\mathcal{A}$ to the upper half of Ω. Reflection across the interval (r, R) (using the Schwarz reflection principle) extends this map to all of $\mathcal{A}$ and all of Ω.

Conceptually, this map from $\mathcal{A}$ to Ω shows the essence of our method most clearly. To apply the trapezoid rule and derive corresponding theorems, however, we can

bypass $\mathcal{A}$ and work with the map from the K, K' rectangle in the t-plane to Ω. (Thus, although R/r is implicitly determined by K and K', we never make a choice of R or r individually, and they do not appear in our MATLAB programs.) By composing (2.2) and (2.1) we get

$$z = \sqrt{mM}\left(\frac{k^{-1}+\mathrm{sn}(t)}{k^{-1}-\mathrm{sn}(t)}\right), \qquad k = \frac{\sqrt{M/m}-1}{\sqrt{M/m}+1}. \tag{2.3}$$

The integral (1.2) can accordingly be written

$$f(\mathbf{A}) \;=\; \frac{-\mathbf{A}}{2\pi i}\int_{-K+iK'/2}^{3K+iK'/2} z^{-1} f(z(t))\,(z(t)-\mathbf{A})^{-1}\,\frac{dz}{du}\,\frac{du}{dt}\,dt;$$

the limits $-K+iK'/2$ to $K+iK'/2$ would correspond to the part of Γ in the upper half-plane (the dots in the fourth panel of Figure 3), and extending this to $3K+iK'/2$ brings in the part in the lower half-plane. With the aid of the identities

$$\frac{dz}{du} = \frac{2k^{-1}\sqrt{mM}}{(k^{-1}-u)^2}, \qquad \frac{du}{dt} = \mathrm{sn}'(t) = \sqrt{1-k^2u^2}\,\sqrt{1-u^2} = \mathrm{cn}(t)\,\mathrm{dn}(t),$$

where cn and dn are further Jacobi elliptic functions in standard notation [1], this becomes

$$f(\mathbf{A}) \;=\; \frac{-\mathbf{A}\sqrt{mM}}{\pi i k}\int_{-K+iK'/2}^{3K+iK'/2} \frac{z^{-1} f(z(t))\,(z(t)-\mathbf{A})^{-1}\,\mathrm{cn}(t)\,\mathrm{dn}(t)}{(k^{-1}-u)^2}\,dt. \tag{2.4}$$

Now since $\mathbf{A}$ is real, the integrand is real-symmetric, i.e., the values it takes at two points z and $\bar{z}$ are complex conjugate. It follows that $f(\mathbf{A})$ is twice the real part of the value obtained by integrating over the first half of the contour, or if we cancel the i in the denominator,

$$f(\mathbf{A}) = \frac{-2\mathbf{A}\sqrt{mM}}{\pi k}\;\mathrm{Im}\int_{-K+iK'/2}^{K+iK'/2} \frac{z^{-1} f(z(t))\,(z(t)-\mathbf{A})^{-1}\,\mathrm{cn}(t)\,\mathrm{dn}(t)}{(k^{-1}-u)^2}\,dt, \tag{2.5}$$

with $z(t)$ given by (2.3).

We now apply the trapezoid rule with N equally spaced points in $(-K+iK'/2, K+iK'/2)$,

$$t_j = -K + \frac{iK'}{2} + 2\frac{(j-\frac{1}{2})K}{N}, \quad 1 \le j \le N. \tag{2.6}$$

(Actually this choice should more precisely be called the midpoint rule. One could make it a true trapezoid rule by shifting the sample points to include the endpoints with weights 1/2.) The result becomes $f(\mathbf{A}) \approx f_N(\mathbf{A})$ with

$$f_N(\mathbf{A}) \;=\; \frac{-4K\mathbf{A}\sqrt{mM}}{\pi N k}\;\mathrm{Im}\sum_{j=1}^{N} \frac{f(z(t_j))\,(z(t_j)\mathbf{I}-\mathbf{A})^{-1}\,\mathrm{cn}(t_j)\,\mathrm{dn}(t_j)}{z(t_j)\,(k^{-1}-u(t_j))^2}. \tag{2.7}$$

(If A is not real, then extend the limits in (2.6) and (2.7) from N to $2N$, and in (2.7), replace $4K$ by $2K$ and multiply by $-i$ instead of extracting the imaginary part.)

The convergence of the method we have described is geometric, with the constant of convergence worsening only logarithmically as $M/m \to \infty$.

THEOREM 2.1. *Let* $\mathbf{A}$ *be a real matrix with eigenvalues in* $[m, M]$, $0 < m < M$, *and let* f *be a function analytic in* $\mathbb{C}\backslash(-\infty, 0]$. *Then the* N*-point conformally transplanted quadrature formula* (2.7) *converges at the rate*

$$\|f(\mathbf{A}) - f_N(\mathbf{A})\| = O\big(e^{\varepsilon - \pi K' N/(2K)}\big) \tag{2.8}$$

for any $\varepsilon > 0$ *as* $N \to \infty$, *and the constant in the exponent is asymptotically* $\pi K'/(2K) \sim \pi^2/\log(M/m)$ *as* $M/m \to \infty$. *For any* M *and* m *we have*

$$\|f(\mathbf{A}) - f_N(\mathbf{A})\| = O\big(e^{-\pi^2 N/(\log(M/m)+3)}\big). \tag{2.9}$$

Proof. The formula (2.7) for $f_N(\mathbf{A})$ comes from applying the trapezoid rule to a function in the t-plane that is analytic in a strip of half-width $a = K'/2$, with grid spacing $\Delta t = 2K/N$. Standard results on the trapezoid rule for periodic integrands imply convergence in such a context at the rate $O(e^{\varepsilon - 2\pi a/\Delta t})$ for any $\epsilon > 0$ [5, 6], and combining these numbers gives (2.8). The claim about asymptotics follows from the relationship $K'/K \sim \pi/(\log(16) - \log(1-k^2)) \sim -\pi/\log(1-k)$ as $k \to 1$ for Jacobi elliptic functions [1, eq. (17.3.26)] together with the relationship $\log(M/m) \sim -2\log(1-k)$ implied by (2.1). The bound $\pi^2/(\log(M/m)+3) < (\pi/2)(K'/K)$ needed to derive (2.9) from (2.8) was established numerically. ∎

We can illustrate Method 1 by the following MATLAB script. Standard MATLAB includes functions to evaluate $\mathrm{sn}(z)$ for real arguments only. This program needs complex arguments, and for this purpose it calls the functions `ellipkkp` and `ellipjc` from Driscoll's Schwarz–Christoffel Toolbox [7, 8], which is freely available online. The mathematical basis of these codes is outlined in [9].

This test code `method1`, like `method2` and `method3` to follow, computes the whole matrix function $f(\mathbf{A})$ rather than just a vector $f(\mathbf{A})b$. Our purpose is to illustrate convergence rates as a function of the number of sample points N, which we can do just as well for a small matrix that can be fully inverted. The matrix illustrated here is the 5×5 Pascal triangle matrix `pascal(5)`, with $M/m \approx 10^4$. The code begins by computing the minimal and maximal eigenvalues m and M, although, in practice, some kind of estimation would normally be used.

```
% method1.m - evaluate f(A) by contour integral.  The functions
%             ellipkkp and ellipjc are from Driscoll's SC Toolbox.

  f = @sqrt;                    % change this for another function f
  A = pascal(5);                % change this for another matrix A
  X = sqrtm(A);                 % change this if f is not sqrt
  I = eye(size(A));
  e = eig(A); m = min(e); M = max(e);   % use only for toy problems!
  k = (sqrt(M/m)-1)/(sqrt(M/m)+1);
  L = -log(k)/pi;
  [K,Kp] = ellipkkp(L);
  for N = 5:5:40
      t = .5i*Kp - K + (.5:N)*2*K/N;
      [u cn dn] = ellipjc(t,L);
      z = sqrt(m*M)*((1/k+u)./(1/k-u));
      dzdt = cn.*dn./(1/k-u).^2;
      S = zeros(size(A));
      for j = 1:N
          S = S + (f(z(j))/z(j))*inv(z(j)*I-A)*dzdt(j);
```

```
    end
    S = -4*K*sqrt(m*M)*imag(S)*A/(k*pi*N);
    error = norm(S-X)/norm(X);
    fprintf('%4d %10.2e\n', N, error)
end
```

Here are the relative errors $\|f(\mathbf{A}) - f_N(\mathbf{A})\|/\|f(\mathbf{A})\|$ printed by this program ($\|\cdot\|$ is the 2-norm), showing 15-digit precision for $N \approx 40$. If M/m were only 10^2 (assuming $\mathbf{A}$ is symmetric or more generally normal), then $N \approx 20$ would be enough to achieve the same accuracy. If $z^{1/2}$ is replaced by $\log z$, then the performance is about the same.

```
>> method1
   5   3.03e-02
  10   4.74e-04
  15   7.29e-06
  20   1.12e-07
  25   1.73e-09
  30   2.66e-11
  35   4.11e-13
  40   7.07e-15
```

In a sequence of computations like this, one could reuse certain data points. For example, the resolvents evaluated for $N = 10$ are evaluated again for $N = 30$. We have not attempted to exploit this redundancy; if one were doing so, it would be advantageous to switch from the midpoint to the true trapezoid formulation, as discussed above, to make the overlaps occur at multiples of 2 rather than 3.

We conclude this section with an example involving a more challenging function f, with singularities all along $(-\infty, 0]$, and a nonnormal matrix with eigenvalues at $(3 \pm \sqrt{5})/2$:

$$\mathbf{A} = \begin{pmatrix} 1 & \frac{1}{2} \\ 2 & 2 \end{pmatrix}, \qquad f(z) = \Gamma(z).$$

The gamma function has poles at all the nonpositive integers, but that is not a problem for Method 1. Taking $N = 42$ is enough to give a result accurate to ten digits,

$$\Gamma(\mathbf{A}) = \begin{pmatrix} 2.0835578979 & -0.1960182234 \\ -0.7840728935 & 1.6915214512 \end{pmatrix},$$

as is readily checked by a diagonalization of $\mathbf{A}$. Of course the point of our method is that it is applicable also in cases where the dimension of $\mathbf{A}$ is such that diagonalization is impractical, or indeed, if $\mathbf{A}$ is not diagonalizable at all. Another approach to the computation of $\Gamma(\mathbf{A})$ based on a Hankel integral has been discussed in [21].

3. Method 2: When $(-\infty, 0)$ is just a branch cut. The assumption of the last section was that f might have arbitrary singularities on $(-\infty, 0]$. In practice, many functions of interest, including z^α and $\log z$, have a singularity at $z = 0$ but just a branch cut on $(-\infty, 0)$. This means that $f(z)$ can be analytically continued along any path that avoids $z = 0$; the difficulty for $z \neq 0$ is only that the value of $f(z)$ will change if the path winds around the origin. Thus we may regard f either as a single-valued function on the slit plane $\mathbb{C}\backslash(-\infty, 0]$ or as a multivalued function on a Riemann surface associated with the punctured plane $C\backslash\{0\}$.

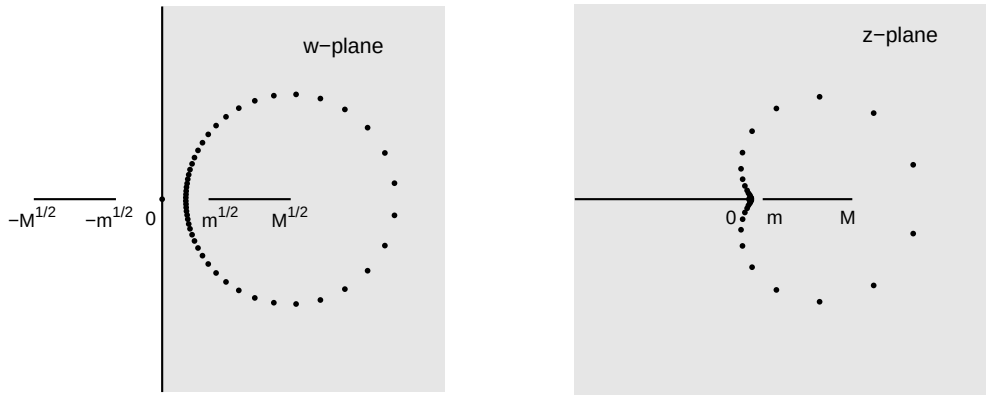


FIG. 4. *In Method 2, for problems where $f(z)$ has only a branch cut on $(-\infty, 0)$, we apply the same quadrature points seen in the fourth panel of Figure 3 to the function $f(w^2)$ in the w-plane (left). This square-roots the condition number and approximately doubles the rate of convergence. In the z-plane, the new method corresponds to a quadrature rule with the nodes shown on the right.*

For functions of this kind, Method 1 is not as efficient as it might be, since it employs a contour that avoids $[m, M]$ and all of $(-\infty, 0]$ rather than just $[m, M]$ and $\{0\}$. One way to improve matters is sketched in Figure 4. If we introduce a new variable $w = z^{1/2}$, $dz = 2w\,dw$, then (1.2) becomes

$$f(\mathbf{A}) = \frac{\mathbf{A}}{\pi i} \int_{\Gamma_w} w^{-1} f(w^2) \, (w^2 - \mathbf{A})^{-1} dw. \tag{3.1}$$

Under the square root, the branch cut of f along $(-\infty, 0]$ in the z plane unfolds to the imaginary axis in the w-plane. The assumption on $f(z)$ then implies that $f(w^2)$ can be analytically continued to an analytic function throughout the slit w-plane $\mathbb{C}\backslash(-\infty, 0]$. Thus we have a new contour integral problem, where the contour should lie in the slit plane and enclose $[m^{1/2}, M^{1/2}]$. The geometry is the same as in the last section, but with $[m, M]$ improved to $[m^{1/2}, M^{1/2}]$. Accordingly, we may use the same method as before. The formula for the new method applied to a real matrix $\mathbf{A}$ is

$$f_N(\mathbf{A}) = \frac{-8K\mathbf{A}(mM)^{1/4}}{\pi N k} \, \mathrm{Im} \sum_{j=1}^{N} \frac{f(w(t_j)^2) \, (w(t_j)^2 \mathbf{I} - \mathbf{A})^{-1} \, \mathrm{cn}(t_j) \mathrm{dn}(t_j)}{w(t_j) \, (k^{-1} - u(t_j))^2}, \tag{3.2}$$

where the t_j are again given by (2.6), but now (2.3) is modified to

$$w = (mM)^{1/4} \left(\frac{k^{-1} + \mathrm{sn}(t)}{k^{-1} - \mathrm{sn}(t)} \right), \qquad k = \frac{(M/m)^{1/4} - 1}{(M/m)^{1/4} + 1}, \tag{3.3}$$

and the values of K and K' are the complete elliptic integrals associated with this new parameter k. For a complex matrix $\mathbf{A}$ the modifications are as for Method 1.

Here is a MATLAB code.

```
% method2.m - Variant of method1.m for evaluating f(A) by contour
%             integration, assuming only a branch cut along (-inf,0).

  f = @sqrt;                    % change this for another function f
```

```
A = pascal(5);                % change this for another matrix A
X = sqrtm(A);                 % change this if f is not sqrt
I = eye(size(A));
e = eig(A); m = min(e); M = max(e);   % only for toy problems
k = ((M/m)^(1/4)-1)/((M/m)^(1/4)+1);
L = -log(k)/pi;
[K,Kp] = ellipkkp(L);
for N = 5:5:25
    t = .5i*Kp - K + (.5:N)*2*K/N;
    [u cn dn] = ellipjc(t,L);
    w = (m*M)^(1/4)*((1/k+u)./(1/k-u));
    dzdt = cn.*dn./(1/k-u).^2;
    S = zeros(size(A));
    for j = 1:N
        S = S + (f(w(j)^2)/w(j))*inv(w(j)^2*I-A)*dzdt(j);
    end
    S = -8*K*(m*M)^(1/4)*imag(S)*A/(k*pi*N);
    error = norm(S-X)/norm(X);
    fprintf('%4d %10.2e\n', N, error)
end
```

The new code converges a good deal faster than the previous one, as expected, getting close to full precision for $N \approx 25$ for this matrix with $M/m \approx 10^4$. Again the results are similar if $z^{1/2}$ is replaced by $\log z$.

```
>> method2
   5   2.97e-03
  10   5.51e-07
  15   7.03e-10
  20   4.88e-12
  25   7.29e-15
```

The same arguments as before give us a convergence theorem. Asymptotically as $M/m \to \infty$, Method 2 is twice as fast as Method 1, though the improvement is less than a factor of 2 for finite M/m.

THEOREM 3.1. *Let* $\mathbf{A}$ *be a real matrix with eigenvalues in* $[m, M]$*, and let* f *be a function analytic in* $\mathbb{C}\backslash(-\infty, 0]$ *that can be continued analytically across* $(-\infty, 0)$ *from the upper half-plane to the lower half-plane. Then the* N*-point improved formula* (3.2) *converges at the rate*

$$\|f(\mathbf{A}) - f_N(\mathbf{A})\| = O\big(e^{\varepsilon - \pi K' N/(2K)}\big) \tag{3.4}$$

for any $\varepsilon > 0$*, where* K *and* K' *are the complete elliptic integrals associated with the parameter* k *of* (3.3)*, and the constant in the exponent is asymptotically* $\pi K'/(2K) \sim 2\pi^2/\log(M/m)$ *as* $M/m \to \infty$*. For any* M *and* m *we have*

$$\|f(\mathbf{A}) - f_N(\mathbf{A})\| = O\big(e^{-2\pi^2 N/(\log(M/m)+6)}\big). \tag{3.5}$$

4. Method 3: Special treatment for $f(z) = z^{1/2}$. The function $f(w^2)$ in the w-plane in the last section, as sketched in Figure 4, has a singularity in general at the origin; that is why the quadrature points must pass between 0 and $m^{1/2}$. For the particular case $f(z) = z^{1/2} = w$, however, there is no singularity, because the change

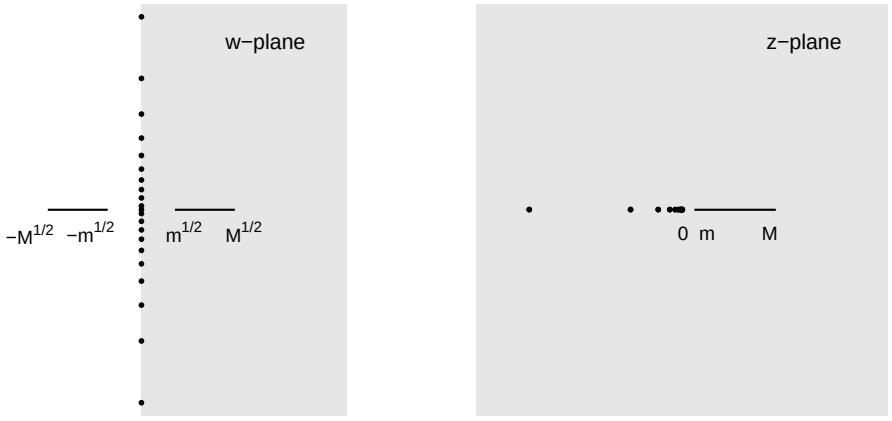


FIG. 5. *Quadrature points for Method* 3. *With* $f(z) = z^{1/2}$, *we can now fully exploit the symmetry in the w-plane. The points in the z-plane lie in* $(-\infty, 0]$, *so all of the matrix arithmetic is real.*

of variables from z to w has eliminated it. Equation (3.1) becomes simply

$$\mathbf{A}^{1/2} = \frac{\mathbf{A}}{\pi i} \int_{\Gamma_w} (w^2 - \mathbf{A})^{-1} dw, \tag{4.1}$$

where Γ_w is a closed contour surrounding $[m^{1/2}, M^{1/2}]$ but not $[-M^{1/2}, -m^{1/2}]$. This makes some further improvements possible. First of all, the quadrature points can now safely approach or even pass through 0. Second, we can use the symmetry to put them on the imaginary axis, corresponding to the negative real axis in the z-plane, thereby eliminating the complex arithmetic in the linear algebra problems. Indeed, since the integrand is of size $O(w^{-2})$, (4.1) is equivalent to an integral over the imaginary axis alone,

$$\mathbf{A}^{1/2} = \frac{i\mathbf{A}}{\pi} \int_{-i\infty}^{i\infty} (w^2 - \mathbf{A})^{-1} dw. \tag{4.2}$$

(Essentially the same formula appears as equation (6.1) in [11].) Third, we can save a factor of two in discretizing this integral since $(w^2-\mathbf{A})^{-1} = ((-w)^2-\mathbf{A})^{-1}$, regardless of the properties of $\mathbf{A}$, so it will no longer matter whether or not $\mathbf{A}$ is real.

Figure 5 shows how Figure 4 changes in this special case. Figure 6 shows the conformal map we can use to compute this transformation, which is essentially the middle step of Figure 3, in which a rectangle is mapped to the upper half-plane. The rectangle in the t-plane is the same as before, and the function $w = m^{1/2}\,\mathrm{sn}(t \mid k^2)$ with

$$k = m^{1/2}/M^{1/2} \tag{4.3}$$

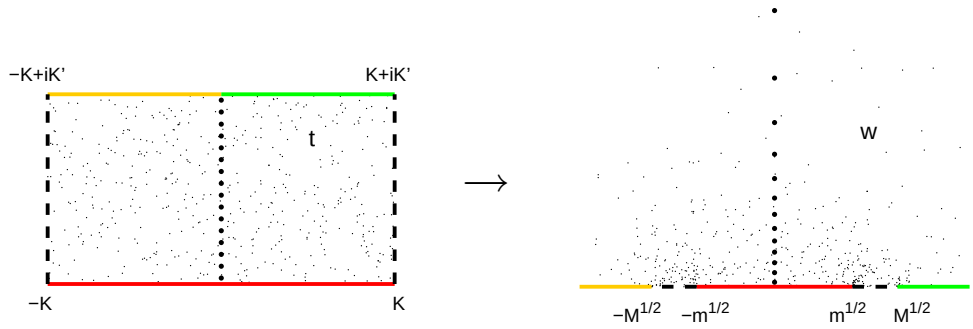


FIG. 6. *The conformal map for the quadrature points of Figure* 5 *is the scaled Jacobi sine function* $w = m^{1/2}\,\mathrm{sn}(t\,|\,k^2)$ *with* $k = m^{1/2}/M^{1/2}$. *Compare Figure* 3.

maps it onto the upper half w-plane. This time, however, it is the vertical midline of the rectangle, the imaginary interval $[0, iK']$, where the trapezoid rule is applied. This gives us the approximation

$$f_N(\mathbf{A}) \;=\; \frac{-2K'm^{1/2}\mathbf{A}}{\pi N}\;\sum_{j=1}^{N}(w(t_j)^2\mathbf{I}-\mathbf{A})^{-1}\,\mathrm{cn}(t_j)\,\mathrm{dn}(t_j) \tag{4.4}$$

with

$$t_j = i(j-\tfrac{1}{2})K'/N, \quad 1 \le j \le N. \tag{4.5}$$

Equation (4.4) works well, but like the algorithms of the past two sections, it has the disadvantage of requiring evaluation of elliptic functions for complex arguments. In this case, however, the arguments (4.5) are purely imaginary. This allows us to take advantage of equations (16.20.1)–(16.20.3) of [1], and the code below implements (4.4) requiring only MATLAB's standard elliptic routines `ellipke` and `ellipj`. For $M/m > 10^6$, however, this approach can be unstable, and better numerical stability is achieved by returning to `ellipkkp` and `ellipjc` from the SC Toolbox. The code has been constructed in such a way that this alteration can be carried out by replacing the line beginning `Kp = ⋯` by the two lines `L = -.5*log(k2)/pi; [K,Kp] = ellipkkp(L);` and the two lines beginning `[sn cn dn] = ⋯` and `cn = ⋯` by the single line `[sn cn dn] = ellipjc(t,L);`.

```
% method3.m - Variant of method1.m and method2.m for evaluating sqrt(A)

  A = pascal(5);                          % change this for another matrix A
  X = sqrtm(A);                           % this cannot be changed
  I = eye(size(A)); e = eig(A);
  m = min(e); M = max(e);                 % only for toy problems
  k2 = m/M;                               % elliptic functions parameter k^2
  Kp = ellipke(1-k2);
  for N = 5:5:20
      t = 1i*(.5:N)*Kp/N;
      [sn cn dn] = ellipj(imag(t),1-k2);
      cn = 1./cn; dn = dn.*cn; sn = 1i*sn.*cn;
      w = sqrt(m)*sn;
```

```
    dzdt = cn.*dn;
    S = zeros(size(A));
    for j = 1:N
        S = S - inv(A-w(j)^2*I)*dzdt(j);
    end
    S = (-2*Kp*sqrt(m)/(pi*N))*A*S;
    error(N) = norm(S-X)/norm(X);
    fprintf('%4d %10.2e\n', N, error(N))
end
```

Method 3 converges very quickly. For our matrix with condition number $M/m \approx 10^4$, nearly full precision is achieved with $N \approx 20$, and we get 6-digit precision with just 10 backslashes.

```
>> method3
   5   9.47e-04
  10   2.24e-07
  15   5.30e-11
  20   1.10e-14
```

The map involved in Method 3 is essentially the same as that of Method 1, except with z improved to $w = z^{1/2}$. Accordingly, the same arguments as before lead to the following analogue of Theorem 2.1 with convergence rates twice as fast.

THEOREM 4.1. *Let* $\mathbf{A}$ *be a real or complex matrix with eigenvalues in* $[m, M]$. *Then the formulas* (4.3)–(4.5) *converge to* $\mathbf{A}^{1/2}$ *at the rate*

$$\|\mathbf{A}^{1/2} - f_N(\mathbf{A})\| = O\big(e^{\varepsilon - 2\pi KN/K'}\big) \tag{4.6}$$

for any $\varepsilon > 0$ *for the values of* K *and* K' *associated with the parameter* (4.3), *and the constant in the exponent is asymptotically* $\pi K'/(2K) \sim 2\pi^2/\log(M/m)$ *as* $M/m \to \infty$. *For any* M *and* m *we have*

$$\|\mathbf{A}^{1/2} - f_N(\mathbf{A})\| = O\big(e^{-2\pi^2 N/(\log(M/m)+3)}\big). \tag{4.7}$$

For $\mathbf{A} = \texttt{pascal(5)}$, for example, the theorem guarantees convergence at the rate $O(5^{-N})$, and for a matrix with $M/m = 10$ this rate increases to $O(41^{-N})$. Even with $M/m = 10^{10}$ the error is $O(2.1^{-N})$.

Figure 7 compares Methods 1, 2, and 3 for the function for which they are all applicable, $\mathbf{A}^{1/2}$, for two matrices with $M/m \approx 62.0$ and 2.06×10^7. Method 3 converges twice as fast as Method 1 in both cases, and it does not lose a factor of 2 if $\mathbf{A}$ is complex. (Its difficulties with rounding errors for the ill-conditioned matrix can be fixed by the use of the alternative formulation described above.) Method 2 has essentially the same convergence rate for the ill-conditioned matrix but falls between the other two methods for the well-conditioned one.

5. Connection with rational approximation. Since the time of Gauss it has been known that quadrature formulas are associated with rational functions: The nodes are the poles, and the weights are the residues. Thus the methods we have proposed can be interpreted as implicit descriptions of rational approximations

$$f(x) \approx r(x), \quad x \in [m, M]. \tag{5.1}$$

One might ask how close these approximations are to optimal. Conversely, if one constructs the optimal rational approximations, do they lead to good methods for

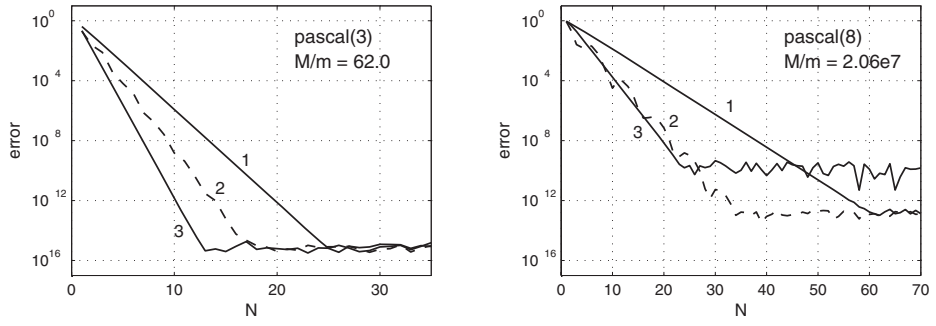


FIG. 7. *Convergence of Methods* 1, 2, *and* 3 *for computing* $\mathbf{A}^{1/2}$ *with* $\mathbf{A}$ = `pascal(3)` *and* `pascal(8)`. *Methods* 1 *and* 2 *are also applicable to more general functions than the square root. The loss of final accuracy in Method* 3 *can be fixed by the complex formulation described just above the listing of* `method3.m`.

evaluating $f(\mathbf{A})$? The same questions are considered in the context of Talbot contours for inverse Laplace transforms in [33].

The link to rational functions goes as follows. Suppose that by some quadrature procedure we derive an approximation

$$f(\mathbf{A}) \approx \mathbf{A} \sum_{j=1}^{N} \gamma_j (z_j - \mathbf{A})^{-1} \tag{5.2}$$

as in (2.7), (3.2), or (4.4), where f is assumed to have spectrum in $[m, M]$. This approximation defines a rational function of type (N, N) (numerator and denominator of degree $\le N$),

$$r(z) = \sum_{j=1}^{N} \frac{\gamma_j z}{z_j - z}. \tag{5.3}$$

It is easily seen that if the difference of the two sides of (5.2) has matrix norm ε, then $|f(\lambda) - r(\lambda)| \le \varepsilon$ must hold for any λ which is an eigenvalue of $\mathbf{A}$. Thus (5.1) certainly holds for the eigenvalues λ; presumably also for other $x \in [m, M]$ if the approximation was derived without reference to particular eigenvalues of $\mathbf{A}$. These comments apply whether A is normal or nonnormal; the difference in the nonnormal case is that whereas (5.2) implies (5.1) in the sense just discussed, the converse does not hold.

Figure 8 examines our three approximations from this point of view for the special case $f(z) = z^{1/2}$, $[m, M] = [1, 100]$, $N = 6$. The error curves $f(x) - r(x)$ are plotted for Methods 1, 2, and 3, and the fourth panel of the figure shows the error curve for the best supremum-norm approximation, which equioscillates 14 times. (Computing the best approximation for such problems is not a trivial matter. We used the Carathéodory–Fejér method [32], as outlined in [33].) An examination of the axis scales for these curves reveals that the approximations are very good, with that of Method 3 in particular quite close to optimal.

A deeper analysis of Figure 8 points to some interesting mathematics. The methods we have put forward in this paper are general, combining conformal mapping and trapezoid rule quadrature to handle general functions $f(z)$. The third panel of the figure, however, reveals an unexpected optimality property of the Method 3 variant for

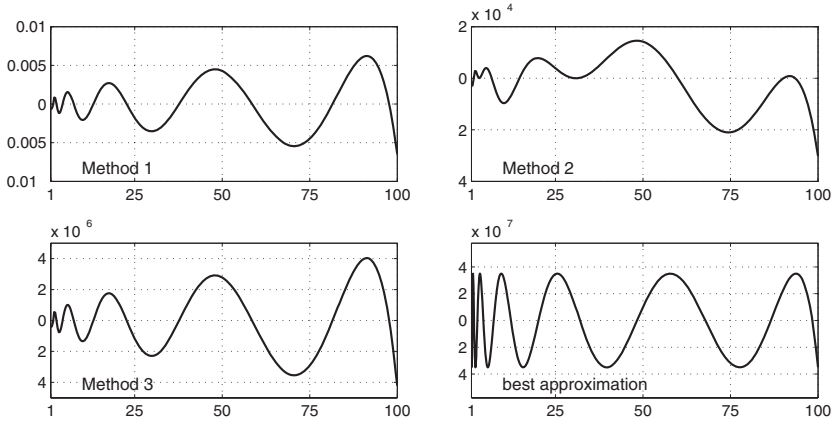


FIG. 8. *Errors* $z^{1/2} - r(z)$ *for various type* (N, N) *rational approximations to* $z^{1/2}$ *on* $[1, 100]$ *with* $N = 6$.

$f(z) = z^{1/2}$. If one scales this error curve by $z^{-1/2}$, it is found to perfectly equioscillate. From this equioscillation one can infer that (5.3) is in this case the optimal type (N, N) approximation to $z^{1/2}$ on $[m, M]$ with respect to the maximum norm weighted by $z^{-1/2}$. Is this best rational approximation already known? Yes indeed—it was discovered by Zolotarev in 1877, who showed that several approximation problems of this kind, though not the unweighted problem of Figure 8, can be solved by the use of Jacobi elliptic functions.[1] The poles for these best approximations, he found, are evenly spaced with respect to the variable we have called t [2, App. E]. (It is this even spacing that leads to our method getting the same result; we do not know a deeper reason why our approximations "had" to come out optimal for this particular weight function.) For more on the connection of Zolotarev's work to matrix functions, see [13, 14, 34].

In other words, the trapezoid rule has rediscovered for us one of Zolotarev's results from 130 years ago! For computing $\mathbf{A}^{1/2}$ when $\mathbf{A}$ has a positive real spectrum, it follows that there is probably no advantage in using the trapezoid rule rather than explicit rational approximations. Incidentally, the use of rational approximations to $\mathbf{A}^{1/2}$ becomes particularly interesting when $\mathbf{A}$ is singular, with eigenvalues in an interval $[0, M]$. It was a celebrated discovery of Newman in 1964 that whereas polynomial approximations to $z^{1/2}$ on such an interval converge only algebraically, rational approximations converge at a rate $O(\exp(-C\sqrt{N}))$ [18]. (Newman's paper deals with approximation of $|z|$, but the result can be carried over to $z^{1/2}$ by a change of variables.) It would be interesting to see whether this behavior can be recovered by a conformal map and the trapezoid rule, but we do not pursue this idea here as it is tied to the special case $f(z) = z^{1/2}$.

The advantage of the trapezoid rule technique is that it is general, applying to arbitrary functions f and to matrices whose spectra may not be all real, yet still readily yielding theorems about asymptotic rates of convergence. We now consider some of these matters.

[1] Egor Ivanovich Zolotarev was a student of Chebyshev, and he learned the theory of elliptic functions from Weierstrass during a visit to Berlin in 1872. Zolotarev made significant contributions to number theory as well as approximation theory before dying at age 31 as a result of being hit by a train at what is now called the Pushkin station outside St. Petersburg [24].

FIG. 9. *On the left, repetition of the right panel of Figure* 4*, now with curves shown inside the quadrature points. If the eigenvalues of* $\mathbf{A}$ *are not confined to the real interval* $[m, M]$ *(solid line) but lie within the innermost curve, for example, then Method* 2 *will still converge exponentially with a rate* 90% *as fast as before. The next curve corresponds to* 80%*, then* 70%*, and so on down to* 10%*. On the right, a closeup of the same figure near the origin (marked by a cross).*

6. Complex eigenvalues and eigenvalue estimates. The methods we have described are flexible. Though we have assumed that the spectrum of $\mathbf{A}$ lies in an interval $[m, M]$ and that m and M are known, the convergence rate degrades only slowly as these conditions are relaxed.

Figure 9 gives an indication of the mathematical reason for this robustness. The general explanation is that in an integral like (1.1), there is no need for the spectrum to be real, provided it is enclosed by the contour Γ. For a quantitative analysis we note that each of our Theorems 2.1, 3.1, and 4.1 are proved by reducing the original problem in the z-plane to a function analytic in a rectangle in the t-plane. The width of the rectangle determines the convergence rate, and if the spectrum of $\mathbf{A}$ moves off the real axis, this narrows the rectangle in a predictable way. For example, in the case of Theorem 2.1, the rectangle has half-width $K'/2$ and is shown in the second panel of Figure 3, where the bottom edge is the portion of the boundary that maps conformally to $[m, M]$. If the spectrum lies in a region in the z-plane larger than $[m, M]$, then this will narrow the rectangle of analyticity. The curves in the figure accordingly correspond to images in the z-plane of horizontal lines in the t-plane at distances $90\%, 80\%, \ldots, 10\%$ of the original distance $K'/2$ below the midline.

Perhaps it is instructive to focus on the middle of the nine curves in Figure 9. So long as the spectrum of $\mathbf{A}$ lies within this quite generous region, the convergence of Method 2 will slow down by at most a factor of 2. Similarly ample regions govern the convergence of Methods 1 and 3. In the case of Method 3, it can be shown that the 50% curve is exactly the circle of radius $M(1 - m/M)^{1/2}$ about the point $z = M$. For $M \gg m$ this is approximately the circle about $z = M$ that passes through the point $z = m/2$. Thus the loss if the spectrum widens from an interval to a disk, and a bigger disk at that, is at worst a factor of 2.

The arguments just made apply to our numerical methods with no adjustments of parameters, but if the spectrum of $\mathbf{A}$ is complex, such adjustments may be advantageous. For example, in the second panel of Figure 3, if the lower half of the rectangle must be narrowed by an amount α because $\mathbf{A}$ has a complex spectrum, then one can halve the impact on the convergence rate by moving the line of quadrature points upwards a distance $\alpha/2$.

Here is an example of the success of these methods for matrices with complex spectra. In MATLAB, `gallery('parter',32)` constructs the 32×32 nonsymmetric Toeplitz matrix with entries $a_{ij} = (i - j + \frac{1}{2})^{-1}$. The norm of $\mathbf{A}$ matches π to 15 digits, and the eigenvalues are complex numbers of absolute value slightly less than

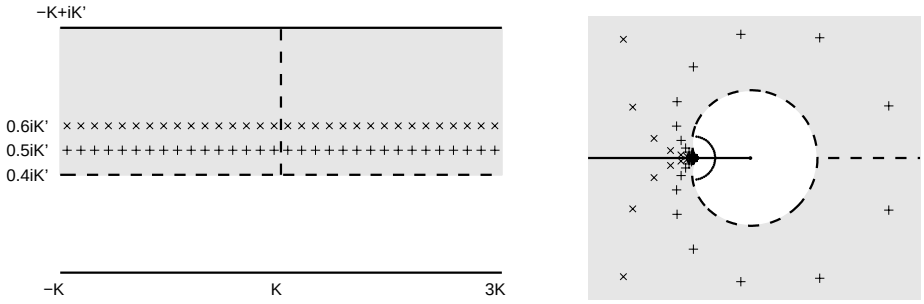


FIG. 10. *Example of the* 32×32 *Parter matrix. The plot on the left shows the rectangle in the t-plane with* 32 *quadrature points at heights* $0.5K'$ *(plusses) and* $0.6K'$ *(crosses). The plot on the right shows their images in the z-plane under the Method* 2 *map with parameters* $m = 0.25$, $M = 8$. *The dashed curve on the right encloses the eigenvalues of* $\mathbf{A}$, *which appear fused in a semicircle, and corresponds to the line at height* $0.4K'$ *on the left. Above this line, in the shaded strip, the integrand will be analytic, so convergence is guaranteed at a rate corresponding to the lower half-width* $0.1K'$ *for the plusses and* $0.2K'$ *for the crosses, that is,* 20% *and* 40% *of the rate given by Theorem* 3.1 *for a matrix with eigenvalues in* $[0.25, 8]$.

π lying approximately on a semicircle in the right half-plane. If Method 2 is applied to compute $\log(\mathbf{A})$ with parameters $m = 0.25$ and $M = 8$ (these parameters were obtained experimentally, not on the basis of analysis of the problem), and with the imaginary part of the quadrature points in (2.6) shifted from $0.5K'$ to $0.6K'$, then here are the results:

```
 5   1.31e-02
10   3.99e-05
15   3.53e-07
20   1.58e-09
25   2.76e-12
30   2.08e-14
```

We see that although the spectrum is complex, thanks to the better condition number, the convergence is nearly as fast as for the example of section 3. Figure 10 explains what is going on. If the imaginary part is fixed at $0.5K'$, then the convergence is about half as fast, becoming about three-quarters as fast in that case if M is increased to 64. If it is fixed at $0.7K'$, then the convergence can be improved further, but in this case the quadrature points in the z-plane cross the branch cut on the negative real axis, so for a successful computation certain values must be negated in accordance with the nonstandard branch of the square root.

This example illustrates that the methods we have put forward can be applied broadly, but also that parameter choices are involved. If these methods are to become the basis of general tools, then it will be necessary to develop systematic ways to estimate eigenvalues or related information from which to derive such parameters. For example, our programs `method1`–`method3` choose m and M by computing the eigenvalues of A and then finding their minimum and maximum. In practice this would usually be impractical, and in any case the codes as written will give the wrong result if the eigenvalues are complex.

7. Three more examples. We conclude with three more numerical examples illustrating a highly nonnormal matrix of small dimension, a fractional matrix power

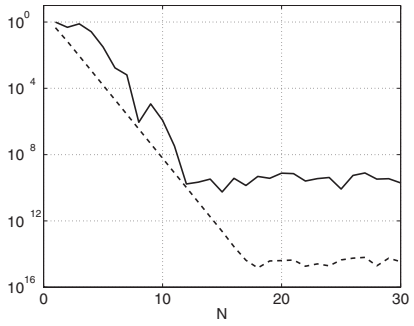


FIG. 11. *Convergence of Method* 3 *in computing* $\mathbf{A}^{1/2}$ *for* $\mathbf{A}$ = `gallery('frank',12)` *(solid curve) and for a diagonal matrix with the same eigenvalues (dashed).*

of medium dimension, and the square root of a sparse matrix of large dimension.

By a highly nonnormal matrix we mean a matrix whose eigenvectors, though a complete set may exist, are far from orthogonal. For an example of this kind we take a Frank matrix, an upper-Hessenberg matrix with integer entries $a_{ij} = n + 1 - j$ on and above the diagonal and $a_{ij} = n + 1 - i$ on the subdiagonal: MATLAB's `gallery('frank',n)`. All of the eigenvalues are real and positive, and it is well known that the condition numbers of the smaller eigenvalues increase rapidly as n increases. Thus the matrix is highly nonnormal. For the 12×12 Frank matrix we computed the square root using a modified version of `method3.m`, with parameters set to their true values $m \approx 0.031$, $M \approx 32.2$. For the exact square root we took the matrix computed in 64-digit arithmetic via diagonalization using MATLAB's symbolic math toolbox. The convergence shown in Figure 11 is rapid, the error reaching 1.7×10^{-10} at $N = 12$ and thereafter levelling off due to the effects of rounding errors. Convergence need not be monotonic, as is confirmed by the sharp increase in error at $N = 9$. It is interesting to note that the error in the square root computed by `sqrtm` is 2×10^{-9}, so contour integration is giving about one more digit of accuracy than the Schur method, which is the standard method of choice if a Schur decomposition can be computed [11].

The dashed curve in Figure 11, corresponding to a diagonal matrix with the same eigenvalues as the Frank matrix, shows that in this example the nonnormality has had a modest effect on the convergence and a great effect on the final accuracy. The reason for the latter is that Method 3 requires the computation of various matrix inverses $(w(t_j)^2\mathbf{I}-\mathbf{A})^{-1}$ (or solution of corresponding systems of equations, if computing $\mathbf{A}^{1/2}b$ instead of $\mathbf{A}^{1/2}$). The numbers $w(t_j)^2$ are far from the spectrum, but not from the ε-pseudospectra for $\varepsilon \ll 1$, so these matrix problems are well conditioned in the normal case but highly ill conditioned in the nonnormal case, causing great amplification of rounding errors. The right solution would be to employ an integration contour that avoids the pseudospectra as well as the spectrum.

For a second example, we use Method 2 to compute $\mathbf{A}^{1/7}b$, where b is a random vector and $\mathbf{A}$ is a 598×598 real nonsymmetric matrix with positive eigenvalues and $m \approx 2.5$, $M \approx 6.1 \times 10^9$. To be precise, $\mathbf{A}$ is the Chebyshev spectral differentiation matrix obtained by the MATLAB commands `A = gallery('chebspec',600)^2` followed by `A = A(2:end-1,2:end-1)`. One way to compute $\mathbf{A}^{1/7}b$ is by means of the identity $\mathbf{A}^{1/7} = \exp(\log(\mathbf{A})/7)$, or `expm(logm(A)/7)` in MATLAB; on a 2003 laptop, this takes 32 seconds. Another approach is to first orthogonally reduce $\mathbf{A}$ to Hessenberg form, then solve N Hessenberg systems of equations in the algorithm of

Method 2. With $N = 40$ this gets comparable accuracy in about 4 seconds. By increasing the dimension, one could increase the disparity in timings.

Our final example (Table 1) applies Method 3 to compute $\mathbf{A}^{1/2}b$, where $\mathbf{A}$ is the matrix generated by the MATLAB command `gallery('poisson',n)`, an $n^2 \times n^2$ matrix corresponding to a standard 5-point finite difference discretization of the Laplacian, and b is the vector of dimension n^2 of all ones. For m and M we take the estimates $2\pi^2/(n+1)^2$ and 8, respectively. This example shows that the methods described here are capable of handling large sparse matrices. The final column of the table compares against MATLAB's built-in method of forming `sqrtm(A)` (the full matrix) and then multiplying by b, a method that is impractical for larger dimensions.

TABLE 1

Computation of $\mathbf{A}^{1/2}b$ by Method 3 for discrete Laplacians of various dimensions. In each case N is chosen large enough for 10 digits of relative accuracy, and time is measured in seconds on a 2003 laptop.

n^2	N	M/m	time	time (`sqrtm`)
16	8	10.1	0.001	0.0006
64	9	32.8	0.02	0.005
256	10	117.	0.04	0.2
1024	12	441.	0.2	21.
4096	14	1712.	1.0	26 minutes
16384	15	6744.	6.0	$<$ 1 day?

REFERENCES

[1] M. ABRAMOWITZ AND I. A. STEGUN, *Handbook of Mathematical Functions*, Dover, New York, 1972.

[2] N. I. ACHIESER, *Theory of Approximation*, Frederick Ungar, New York, 1956.

[3] M. BENZI AND G. H. GOLUB, *Bounds for the entries of matrix functions with applications to preconditioning*, BIT, 39 (1999), pp. 417–438.

[4] P. I. DAVIES AND N. J. HIGHAM, *Computing $f(A)b$ for matrix functions f*, in QCD and Numerical Analysis III, A. Boriçi, A. Frommer, B. Joó, A. Kennedy, and B. Pendleton, eds., Lect. Notes Computat. Sci. Engrg. 47, Springer-Verlag, Berlin, 2005, pp. 15–24.

[5] P. J. DAVIS, *On the numerical integration of periodic analytic functions*, in On Numerical Integration: Proceedings of a Symposium, R. E. Langer, ed., Madison, WI, 1958, Math. Res. Ctr., U. of Wisconsin, Madison, WI, 1959, pp. 45–59.

[6] P. J. DAVIS AND P. RABINOWITZ, *Methods of Numerical Integration*, 2nd ed., Academic Press, Orlando, FL, 1984.

[7] T. A. DRISCOLL, *The Schwarz–Christoffel toolbox*, available online at http://www.math.udel.edu/∼driscoll/software/SC/.

[8] T. A. DRISCOLL, *Algorithm 843: Improvements to the Schwarz–Christoffel toolbox for MATLAB*, ACM Trans. Math. Software, 31 (2005), pp. 239–251.

[9] T. A. DRISCOLL AND L. N. TREFETHEN, *Schwarz–Christoffel Mapping*, Cambridge University Press, Cambridge, UK, 2002.

[10] I. P. GAVRILYUK AND V. L. MAKAROV, *Exponentially convergent parallel discretization methods for the first order evolution equations*, Comput. Methods Appl. Math., 1 (2001), pp. 333–355.

[11] N. J. HIGHAM, *Functions of Matrices: Theory and Computation*, SIAM, Philadelphia, 2008.

[12] M. IRI, S. MORIGUTI, AND Y. TAKASAWA, *On a certain quadrature formula*, Kokyuroku Res. Inst. Math. Sci., Kyoto Univ., 91 (1970), pp. 82–118 (in Japanese, translated into English in J. Comput. Appl. Math., 17 (1987), pp. 3–20).

[13] A. D. KENNEDY, *Approximation theory for matrices*, Nucl. Phys. B, 128 (2004), pp. 107–116.

[14] C. S. KENNEY AND A. J. LAUB, *The matrix sign function*, IEEE Trans. Automat. Control, 40 (1995), pp. 1330–1348.

[15] E. MARTENSEN, *Zur numerischen Auswertung uneigentlicher Integrale*, Z. Angew. Math. Mech., 48 (1968), pp. T83–T85.

[16] W. McLean and V. Thomée, *Time discretization of an evolution equation via Laplace transforms*, IMA J. Numer. Anal., 24 (2004), pp. 439–463.
[17] M. Mori, *Discovery of the double exponential transformation and its developments*, Publ. Res. Inst. Math. Sci., Kyoto Univ., 41 (2005), pp. 897–935.
[18] D. J. Newman, *Rational approximation to $|x|$*, Michigan Math. J., 11 (1964), pp. 11–14.
[19] S.-D. Poisson, *Sur le calcul numérique des intégrales définies*, Mémoires de L'Académie Royale des Sciences de L'Institut de France, 4 (1827), pp. 571–602 (written in 1823).
[20] T. W. Sag and G. Szekeres, *Numerical evaluation of high-dimensional integrals*, Math. Comp., 18 (1964), pp. 245–253.
[21] T. Schmelzer and L. N. Trefethen, *Computing the gamma function using contour integrals and rational approximations*, SIAM J. Numer. Anal., 45 (2007), pp. 558–571.
[22] C. Schwartz, *Numerical integration of analytic functions*, J. Comput. Phys., 4 (1969), pp. 19–29.
[23] D. Sheen, I. H. Sloan, and V. Thomée, *A parallel method for time-discretization of parabolic problems based on contour integral representation and quadrature*, Math. Comp., 69 (1999), pp. 177–195.
[24] K.-G. Steffens, *The History of Approximation Theory: From Euler to Bernstein*, Birkhäuser Boston, Boston, 2006.
[25] F. Stenger, *Integration formulae based on the trapezoidal formula*, J. Inst. Math. Appl., 12 (1973), pp. 103–114.
[26] F. Stenger, *Numerical methods based on Whittaker cardinal, or sinc functions*, SIAM Rev., 23 (1981), pp. 165–224.
[27] F. Stenger, *Numerical Methods Based on Sinc and Analytic Functions*, Springer-Verlag, New York, 1993.
[28] H. Takahasi and M. Mori, *Quadrature formulas obtained by variable transformation*, Numer. Math., 21 (1973), pp. 206–219.
[29] H. Takahasi and M. Mori, *Double exponential formulas for numerical integration*, Publ. Res. Inst. Math. Sci., Kyoto Univ., 9 (1974), pp. 721–741.
[30] A. Talbot, *The accurate numerical inversion of Laplace transforms*, J. Inst. Math. Appl., 23 (1979), pp. 97–120.
[31] N. M. Temme, *Special Functions*, John Wiley and Sons, New York, 1996.
[32] L. N. Trefethen and M. H. Gutknecht, *The Carathéodory–Fejér method for real rational approximation*, SIAM J. Numer. Anal., 20 (1983), pp. 420–436.
[33] L. N. Trefethen, J. A. C. Weideman, and T. Schmelzer, *Talbot quadratures and rational approximation*, BIT, 46 (2006), pp. 653–670.
[34] J. van den Eshof, A. Frommer, Th. Lippert, K. Schilling, and H. A. van der Vorst, *Numerical methods for the QCD overlap operator*. I. *Sign-function and error bounds*, Comput. Phys. Commun., 146 (2002), pp. 203–224.
[35] J. A. C. Weideman, *Optimizing Talbot's contours for the inversion of the Laplace transform*, SIAM J. Numer. Anal., 44 (2006), pp. 2342–2362.
[36] J. A. C. Weideman and L. N. Trefethen, *Parabolic and hyperbolic contours for computing the Bromwich integral*, Math. Comp., 76 (2007), pp. 1341–1356.

SIAM J. NUMER. ANAL.
Vol. 46, No. 5, pp. 2524–2550
© 2008 Society for Industrial and Applied Mathematics

QUASI-OPTIMAL CONVERGENCE RATE FOR AN ADAPTIVE FINITE ELEMENT METHOD*

J. MANUEL CASCON†, CHRISTIAN KREUZER‡, RICARDO H. NOCHETTO§, AND KUNIBERT G. SIEBERT‡

Abstract. We analyze the simplest and most standard adaptive finite element method (AFEM), with any polynomial degree, for general second order linear, symmetric elliptic operators. As is customary in practice, the AFEM marks exclusively according to the error estimator and performs a minimal element refinement without the interior node property. We prove that the AFEM is a contraction, for the sum of the energy error and the scaled error estimator, between two consecutive adaptive loops. This geometric decay is instrumental to derive the optimal cardinality of the AFEM. We show that the AFEM yields a decay rate of the energy error plus oscillation in terms of the number of degrees of freedom as dictated by the best approximation for this combined nonlinear quantity.

Key words. error reduction, convergence, optimal cardinality, adaptive algorithm

AMS subject classifications. 65N30, 65N50, 65N15, 65N12, 41A25

DOI. 10.1137/07069047X

1. Introduction.

Let Ω be a bounded, polyhedral domain in $\mathbb{R}^d$, $d \geq 2$. We consider a homogeneous Dirichlet boundary value problem for a self-adjoint second order elliptic partial differential equation (PDE)

$$\begin{aligned} \mathcal{L}u := -\operatorname{div}(\mathbf{A}\nabla u) + c\,u &= f \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \partial\Omega. \end{aligned} \tag{1.1}$$

The choice of boundary condition is made for ease of presentation, since similar results are valid for other boundary conditions. Precise conditions on given data $\mathbf{D} := (\mathbf{A}, c)$ and f of $\mathcal{L}$ are stated in section 2.1. We refer to [5] for more general operators.

We analyze here a standard adaptive finite element method (AFEM) of the form

$$\mathsf{SOLVE} \to \mathsf{ESTIMATE} \to \mathsf{MARK} \to \mathsf{REFINE}. \tag{1.2}$$

Even though adaptivity has been a fundamental tool of engineering and scientific computing for about three decades, the convergence analysis is rather recent. It started with Dörfler [8], who introduced a crucial marking, from now on called

*Received by the editors May 3, 2007; accepted for publication (in revised form) January 15, 2008; published electronically June 18, 2008.
http://www.siam.org/journals/sinum/46-5/69047.html

†Departamento de Matemáticas, Universidad de Salamanca, 37008 Salamanca, Spain (casbar@usal.es, http://matematicas.fis.usal.es/~casbar). This author's research was partially supported by grants CGL2004-06171-C03-03/CLI, Ministerio de Ciencia y Tecnología (Spain), and SA078A05, Junta de Castilla y León (Spain), FEDER funds (European Union), and NSF grant DMS-0505454.

‡Institut für Mathematik, Universität Augsburg, Universitätsstraße 14, 86159 Augsburg, Germany (kreuzer@math.uni-augsburg.de, http://www.math.uni-augsburg.de/de/prof/lam/mitarbeiter/christian_kreuzer, siebert@math.uni-augsburg.de, http://scicomp.math.uni-augsburg.de/siebert/). The research of these authors was partially supported by DAAD grant "Efficient Finite Element Methods for Solid and Fluid Mechanics Computations."

§Department of Mathematics and Institute of Physical Science and Technology, University of Maryland, College Park, MD 20742 (rhn@math.umd.edu, http://www.math.umd.edu/~rhn). This author's research was partially supported by NSF grant DMS-0505454.

Dörfler's marking, and proved strict energy error reduction for the Laplacian provided the initial mesh $\mathcal{T}_0$ satisfies a fineness assumption. Morin, Nochetto, and Siebert [17, 18] showed that such strict energy error reduction cannot be expected in general. By introducing the concept of data oscillation and the interior node property, they proved convergence of the AFEM without restrictions on $\mathcal{T}_0$. The latter result, however, is valid only for $\mathbf{A}$ in (1.1) piecewise constant on $\mathcal{T}_0$ and vanishing c. Inspired by the work by Chen and Feng [6], Mekchay and Nochetto [16] extended this result to general second order elliptic operators upon dealing with the new concept of *total error*, namely, the sum of the energy error plus oscillation. They proved that the AFEM is a contraction for the total error, a property that will turn out to be essential in this paper as well. Recently, Diening and Kreuzer [7] proved a similar property for the p-Laplacian but avoided marking for oscillation by using the fact that oscillation appearing in the lower bound is dominated by the estimator.

In this paper we go back to the basis of the AFEM philosophy and consider the simplest possible approach to adaptivity for (1.1). In section 2 we first state precisely the problem and describe the modules of the AFEM: SOLVE computes the Ritz–Galerkin approximation; i.e., we assume an exact solution and integration; ESTIMATE computes the standard residual estimator; MARK resorts to Dörfler marking based solely on the estimator; and REFINE utilizes bisectioning of elements with the minimal refinement condition that marked elements are bisected at least once. In this respect the AFEM is a really standard algorithm in that it avoids marking for oscillation and circumvents the interior node property of [16, 17, 18] for marked elements. We would like to stress that removing these ingredients is very important from a practical point of view. The interior node property enforces six bisections of marked elements in three space dimensions and thus increases the number of DOFs between two iterations drastically. Additionally, since oscillation is not used by the AFEM, it does not have to be computed. Although computing time for oscillation may be negligible, this strongly improves on implementation requirements.

The first convergence result for such a standard AFEM is due to Morin, Siebert, and Veeser [20]. They proved recently, for a larger problem class and more general marking strategies, plain convergence of the AFEM but without an error reduction property. By relying on Dörfler marking we are able to prove a stronger result for self-adjoint elliptic operators of the form (1.1), namely, contraction of the quasi-error and quasi-optimal cardinality of the AFEM.

To summarize the main results, let $\{\mathcal{T}_k, \mathbb{V}_k, U_k, \eta_k, \operatorname{osc}_k\}_{k\geq 0}$ be the sequence of meshes, finite element spaces, discrete solutions, estimators, and oscillations produced by the AFEM in the kth step. Even though the energy error is monotone, strict error reduction fails when $U_{k+1} = U_k$, i.e., $|\!|\!| u - U_k |\!|\!|_\Omega = |\!|\!| u - U_{k+1} |\!|\!|_\Omega$; see [17, 18] for further details. On the other hand, the residual estimator $\eta_k = \eta_k(U_k)$ exhibits a strict reduction when $U_{k+1} = U_k$ but no monotone behavior in general; this is shown in section 3. The new insight of this paper is that the sum of the energy error and the scaled estimator, the so-called *quasi error* $(|\!|\!| u - U_k |\!|\!|_\Omega^2 + \gamma\, \eta_k^2)^{1/2}$, is strictly reduced by the AFEM even though each term may not be. In fact, we prove in section 4 that the AFEM is a contraction for this new error notion (see Theorem 4.1).

MAIN RESULT 1. There exist constants $\gamma > 0$ and $0 < \alpha < 1$ such that

$$|\!|\!| u - U_{k+1} |\!|\!|_\Omega^2 + \gamma\, \eta_{k+1}^2 \leq \alpha^2 \Big(|\!|\!| u - U_k |\!|\!|_\Omega^2 + \gamma\, \eta_k^2 \Big).$$

Quasi-optimal convergence rates for the AFEM, expressing energy error decay in terms of the number of DOFs as dictated by nonlinear approximation theory, were

first proved by Binev, Dahmen, and DeVore [3]. They resorted to a crucial, but somewhat artificial, coarsening step. Coarsening was later removed by Stevenson [23], who developed an optimality theory for a much more realistic AFEM but still included an inner loop to deal with oscillation. Both papers [3, 23] are restricted to the Laplace operator and rely on suitable marking by oscillation and the interior node property.

To derive convergence rates we need to seek a suitable error quantity and an associated approximation class $\mathbb{A}_s$. Since all decisions of the AFEM in MARK are based on the estimator η_k, a decay rate for the true error is closely related to the quality of the estimator, which is described by the *global lower bound*

$$\eta_k^2 \preccurlyeq |\!|\!| u - U_k |\!|\!|_\Omega^2 + \operatorname{osc}_k^2;$$

hereafter, *total error* refers to the square root of the right-hand side (see [16]). The lower bound states that the estimator is dominated by the error up to an oscillation term and overestimation may occur only when oscillation is large. Furthermore we observe that the total error is bounded again by the estimator—thanks to the upper bound $|\!|\!| u - U_k |\!|\!|_\Omega \preccurlyeq \eta_k$ and $\operatorname{osc}_k \le \eta_k$—and the estimator itself is an upper bound for the quasi error, which is the quantity reduced by the AFEM:

$$|\!|\!| u - U_k |\!|\!|_\Omega^2 + \operatorname{osc}_k^2 \;\approx\; \eta_k^2 \;\approx\; |\!|\!| u - U_k |\!|\!|_\Omega^2 + \gamma\, \eta_k^2.$$

In summary, the performance of the AFEM is intrinsically bonded to the total error, which measures both the approximability of u as well as of data, encoded in the oscillation term. For piecewise polynomial coefficients $(\mathbf{A}, c)$, oscillation reduces to approximation of the right-hand side f of (1.1). In the general case, approximation of data $(\mathbf{A}, c)$ in osc_k couples in a nonlinear fashion with the discrete solution U_k. In section 5 we start with the definition of an approximation class $\mathbb{A}_s$ based on the total error. Roughly speaking, $(u, f, \mathbf{D})$ belongs to $\mathbb{A}_s$ if the total error can be approximated within any tolerance $\epsilon > 0$ with $O(\epsilon^{-s})$ DOFs. We then prove the *quasi-optimal* convergence rate for the AFEM in terms of DOFs by assuming certain restrictions on the initial triangulation and the marking parameter $\theta \in (0, \theta_*)$ (see Theorem 5.11).

MAIN RESULT 2. If $(u, f, \mathbf{D}) \in \mathbb{A}_s$, then there exists a constant C such that

$$\left\{ |\!|\!| u - U_k |\!|\!|_\Omega^2 + \operatorname{osc}_k^2 \right\}^{1/2} \le C \left(\#\mathcal{T}_k - \#\mathcal{T}_0 \right)^{-s}.$$

Since the estimator is equivalent to the total error, we obtain the same decay rate for the estimator η_k. Besides the contraction property of the AFEM, crucial ingredients of the proof are a quasi-monotonicity property of oscillation as well as a localized upper bound, both proved in section 3. The latter refines a similar bound by Stevenson [23]. In contrast to former optimality proofs [3, 23], the present analysis stays within the class of conforming meshes. This is a necessary framework for the analysis when dealing with oscillation in the jump residual.

As already mentioned above, for piecewise polynomial coefficients $(\mathbf{A}, c)$ one can separate approximation of u and data f. When doing so and marking with respect to two quantities such as the estimator and oscillation, the role of marking becomes critical for proving optimality. To shed light on this issue, we conclude this article by discussing in section 6 the simultaneous adaptive approximation of two functions with distinct asymptotic error decays. The discussion reveals that separate marking might be suboptimal and explains why the AFEM designed by Morin, Nochetto, and Siebert [17, 18] is altered in former optimality proofs [3, 23].

2. Problem and adaptive finite element method. We first introduce the underlying problem and state assumptions on given data. We then describe the refinement framework and the AFEM along with its modules.

2.1. Weak formulation. Let Ω be a bounded, polyhedral domain in $\mathbb{R}^d$, $d \geq 2$, that is triangulated by a conforming triangulation $\mathcal{T}_0$. We assume that data of (1.1) have the following properties:

(a) $\mathbf{A}\colon \Omega \mapsto \mathbb{R}^{d\times d}$ is piecewise Lipschitz over $\mathcal{T}_0$ and is symmetric positive definite with eigenvalues in $0 < a_* \leq a^* < \infty$, i.e.,

$$a_*(x)\,|\xi|^2 \leq \mathbf{A}(x)\xi\cdot\xi \leq a^*(x)\,|\xi|^2 \quad \forall \xi \in \mathbb{R}^d,\ x \in \Omega;$$

(b) $c \in L^\infty(\Omega)$ is nonnegative, i.e., $c \geq 0$ in Ω;

(c) $f \in L^2(\Omega)$.

Now we turn to the weak formulation of (1.1). For any set $\omega \subset \mathbb{R}^d$ with a nonempty interior we denote by $H^1(\omega)$ the usual Sobolev space of functions in $L^2(\omega)$ whose first derivatives are also in $L^2(\omega)$, endowed with the norm

$$\|u\|_{H^1(\omega)} := \left(\|u\|^2_{L^2(\omega)} + \|\nabla u\|^2_{L^2(\omega)}\right)^{1/2}.$$

Moreover, we denote by $\langle .\,,.\rangle_\omega$ the $L^2(\omega)$ scalar product. Finally we let $\mathbb{V} := H^1_0(\Omega)$ be the space of functions in $H^1(\Omega)$ with vanishing trace on $\partial\Omega$. A weak solution of (1.1) is a function u satisfying

$$u \in \mathbb{V}: \qquad \mathcal{B}[u,v] = \langle f, v\rangle_\Omega \qquad \forall v \in \mathbb{V}, \tag{2.1}$$

where the bilinear form is defined to be

$$\mathcal{B}[u,v] := \langle \mathbf{A}\nabla u, \nabla v\rangle_\Omega + \langle c\,u, v\rangle_\Omega \qquad \forall u, v \in \mathbb{V}.$$

In view of the Poincaré–Friedrichs inequality, one has *coercivity* on $\mathbb{V}$

$$\mathcal{B}[v,v] = \langle \mathbf{A}\nabla v, \nabla v\rangle_\Omega + \langle c\,v, v\rangle_\Omega \geq \int_\Omega a_*\,|\nabla v|^2 + c\,v^2 \geq c_B\,\|v\|^2_{H^1(\Omega)},$$

and c_B depends only on data and Ω. The bilinear form $\mathcal{B}$ induces the so-called *energy norm*:

$$|\!|\!|v|\!|\!|_\Omega := \mathcal{B}[v,v]^{1/2} \qquad \forall v \in \mathbb{V}.$$

Note that the bilinear form also fulfills the *local continuity*

$$\mathcal{B}[v,w] \leq \sqrt{C_B}\;|\!|\!|v|\!|\!|_\omega\,\|w\|_{H^1(\omega)} \qquad \forall v, w \in H^1_0(\Omega),\ \mathrm{supp}(w) \subset \omega \subset \Omega,$$

with $|\!|\!|v|\!|\!|_\omega := \langle \mathbf{A}\nabla v, \nabla v\rangle_\omega + \langle c\,v, v\rangle_\omega$. This local continuity is essential in deriving *local lower bounds* in the a posteriori error analysis. Furthermore it implies continuity of $\mathcal{B}[\cdot,\cdot]$ on $H^1(\Omega)$ at once. Thanks to the coercivity and continuity of $\mathcal{B}$, the norm $|\!|\!|\cdot|\!|\!|_\Omega$ is equivalent to $\|\cdot\|_{H^1(\Omega)}$ on $H^1_0(\Omega)$. The existence and uniqueness of (2.1) thus follow from the Lax–Milgram theorem [9]. The restriction to a symmetric bilinear form and $c \geq 0$ can be relaxed, provided (2.1) admits a unique solution. This extension will be studied in [5].

2.2. Refinement framework. Refinement is based on shape-regular bisection of single elements. Any given simplex is subdivided into two subsimplices of the same size such that the minimal angle is uniformly bounded from below; we refer to [2, 13, 14, 15, 24, 25] or the monograph [21] and the references therein. In two dimensions, this is the newest vertex bisection. Bisectioning creates a unique *master forest* $\mathbb{F}$ of binary trees with infinite depth, where each node is a simplex, its two successors are the two children created by bisection, and the roots of the binary trees are the elements of the initial conforming triangulation $\mathcal{T}_0$. The master forest $\mathbb{F}$ contains full information of all possible subdivisions created from $\mathcal{T}_0$ by bisection, i.e., information about vertices, neighbors, refinement edges, etc., for any simplex that can be generated.

A finite subset $\mathcal{F} \subset \mathbb{F}$ is called a *forest* if $\mathcal{T}_0 \subset \mathcal{F}$ and the nodes satisfy:

(a) all nodes in $\mathcal{F} \setminus \mathcal{T}_0$ have a predecessor;

(b) all nodes in $\mathcal{F}$ have either two successors or none.

Any node of $\mathcal{F}$ is thus uniquely connected with a node of the initial triangulation $\mathcal{T}_0$. Furthermore, any forest may have interior nodes, i.e., nodes with successors, and does have leaf nodes, i.e., nodes without successors. Finally, each forest $\mathcal{F}$ corresponds one to one to a subdivision $\mathcal{T}(\mathcal{F})$ of Ω into simplices by defining $\mathcal{T}(\mathcal{F})$ as the set of leaf nodes of $\mathcal{F}$. Note that such a subdivison may be nonconforming.

If $\mathcal{F} \subset \mathcal{F}_*$ are two forests, we call $\mathcal{T}_* = \mathcal{T}(\mathcal{F}_*)$ a refinement of $\mathcal{T} = \mathcal{T}(\mathcal{F})$ and denote this by $\mathcal{T} \leq \mathcal{T}_*$. We define

$$\mathcal{R}_{\mathcal{T}\to\mathcal{T}_*} := \mathcal{T} \setminus (\mathcal{T}_* \cap \mathcal{T})$$

as the set of *refined elements*; i.e., $\mathcal{R}_{\mathcal{T}\to\mathcal{T}_*}$ is the set of leaf nodes of $\mathcal{F}$ that are interior nodes of $\mathcal{F}_*$. The class of all *conforming* refinements by bisection of $\mathcal{T}_0$ is

$$\mathbb{T} := \{\mathcal{T}(\mathcal{F}) \mid \mathcal{F} \subset \mathbb{F} \text{ is a forest and } \mathcal{T}(\mathcal{F}) \text{ is a conforming triangulation of } \Omega\}.$$

Given two forests $\mathcal{F}, \mathcal{F}_* \subset \mathbb{F}$ corresponding to subdivisions $\mathcal{T}$ and $\mathcal{T}_*$ of Ω, we define $\mathcal{F} \cup \mathcal{F}_*$ to be the union of the nodes of $\mathcal{F}$ and $\mathcal{F}_*$. Obviously, $\mathcal{F} \cup \mathcal{F}_* \subset \mathbb{F}$ and $\mathcal{T}_0 \subset \mathcal{F} \cup \mathcal{F}_*$. By construction, all nodes of $(\mathcal{F} \cup \mathcal{F}_*) \setminus \mathcal{T}_0$ have a predecessor, and all nodes of the union have either two successors or none: this implies that $\mathcal{F} \cup \mathcal{F}_*$ is a forest. We call the resulting unique subdivision of Ω the *overlay of* $\mathcal{T}$ *and* $\mathcal{T}_*$:

$$\mathcal{T} \oplus \mathcal{T}_* := \mathcal{T}(\mathcal{F} \cup \mathcal{F}_*).$$

For two conforming triangulations $\mathcal{T}, \mathcal{T}_* \in \mathbb{T}$ we prove in Lemma 3.7 below that $\mathcal{T} \oplus \mathcal{T}_* \in \mathbb{T}$ is the smallest conforming refinement of $\mathcal{T}, \mathcal{T}_*$ and satisfies

$$\#(\mathcal{T} \oplus \mathcal{T}_*) \leq \#\mathcal{T} + \#\mathcal{T}_* - \#\mathcal{T}_0. \tag{2.2}$$

We finally introduce some notations related to triangulations. For $T \in \mathcal{T}$ we denote by $h_T := |T|^{\frac{1}{d}}$ the local mesh size and by ω_T the union of all elements in $\mathcal{T}$ sharing one side with T. We further denote by $\mathcal{S}$ the skeleton of $\mathcal{T}$, i.e., the union of the interelement sides, and for an interior side $\sigma \in \mathcal{S}$ we let ω_σ be the union of the two adjacent elements sharing σ.

Thanks to the properties of bisection, all constants depending on shape regularity of $\mathcal{T} \in \mathbb{T}$ are uniformly bounded by a constant solely depending on $\mathcal{T}_0$ [14, 25].

2.3. The module SOLVE. Given any conforming triangulation $\mathcal{T} \in \mathbb{T}$, we define the finite element space

$$\mathbb{V}(\mathcal{T}) := \{V \in \mathbb{V} \mid V_{|T} \in \mathbb{P}_n(T),\ T \in \mathcal{T}\},$$

where $n \in \mathbb{N}$ is a fixed polynomial degree and $\mathbb{P}_n$ denotes the space of all polynomials of degree $\leq n$. Since the continuity and coercivity of $\mathcal{B}$ are inherited by any subspace of $\mathbb{V}$, the Lax–Milgram theorem implies the existence and uniqueness of the Ritz–Galerkin approximation in $\mathbb{V}(\mathcal{T})$ defined by

$$U \in \mathbb{V}(\mathcal{T}): \qquad \mathcal{B}[U, V] = \langle f, V \rangle_\Omega \qquad \forall\, V \in \mathbb{V}(\mathcal{T}). \tag{2.3}$$

We suppose that the module SOLVE outputs the exact Ritz–Galerkin solution on $\mathcal{T}$:

$$U = \mathsf{SOLVE}(\mathcal{T});$$

i.e., U is computed via exact linear algebra and exact integration. Optimal multilevel solvers can be incorporated as in [23], but quadrature is more delicate.

We observe that for any pair $\mathcal{T}, \mathcal{T}_* \in \mathbb{T}$, with $\mathcal{T} \leq \mathcal{T}_*$, there holds $\mathbb{V}(\mathcal{T}) \subset \mathbb{V}(\mathcal{T}_*)$; i.e., the discrete spaces are *nested*. This turns out to be a crucial property in the subsequent analysis in that the following *orthogonality* relation holds:

$$\|\!|u - U|\!\|_\Omega^2 = \|\!|u - U_*|\!\|_\Omega^2 + \|\!|U_* - U|\!\|_\Omega^2 . \tag{2.4}$$

This is a consequence of $\mathcal{B}[u - U_*, U_* - U] = 0$ or, equivalently, that $u - U_*$ and $U_* - U$ are orthogonal with respect to the scalar product induced by $\mathcal{B}[\cdot, \cdot]$.

2.4. The module ESTIMATE. For $\mathcal{T} \in \mathbb{T}$ and $V \in \mathbb{V}(\mathcal{T})$ we define the *element residual* and *jump residual* for V by

$$R(V)_{|T} := (f + \mathcal{L}\, V)_{|T}, \quad T \in \mathcal{T}, \qquad \text{and} \qquad J(V)_{|\sigma} := ([\![\mathbf{A}\nabla V]\!] \cdot \boldsymbol{\nu})_{|\sigma}, \quad \sigma \in \mathcal{S}.$$

Hereafter, $[\![q]\!]$ is the jump of q across an interior side σ, and $\boldsymbol{\nu}$ denotes a unit normal vector associate to side σ. The error indicator for V on $T \in \mathcal{T}$ is given by

$$\eta_{\mathcal{T}}^2(V, T) := h_T^2 \, \|R(V)\|_{L^2(T)}^2 + h_T \, \|J(V)\|_{L^2(\partial T \cap \Omega)}^2 , \tag{2.5}$$

where we recall that $h_T = |T|^{1/d}$. We assume that, given a triangulation $\mathcal{T}$ and the Ritz–Galerkin solution $U \in \mathbb{V}(\mathcal{T})$, the module ESTIMATE outputs the indicators

$$\{\eta_{\mathcal{T}}(U, T)\}_{T \in \mathcal{T}} = \mathsf{ESTIMATE}(U, \mathcal{T}).$$

The lower bound involves oscillation, which we define next. For $m \in \mathbb{N}_0$, we denote by Π_m^p the L^p-best approximation operator onto the set of discontinuous polynomials of degree $\leq m$ over either $T \in \mathcal{T}$ or $\sigma \in \mathcal{S}$, depending on the context. If $P_m^p := \mathrm{id} - \Pi_m^p$, then we define oscillation of $V \in \mathbb{V}(\mathcal{T})$ to be

$$\mathrm{osc}_{\mathcal{T}}^2(V, T) := h_T^2 \left\| P_{2n-2}^2 R(V) \right\|_{L^2(T)}^2 + h_T \left\| P_{2n-1}^2 J(V) \right\|_{L^2(\partial T \cap \Omega)}^2 .$$

Finally, for any subset $\mathcal{T}' \subset \mathcal{T}$ we set

$$\eta_{\mathcal{T}}^2(V, \mathcal{T}') := \sum_{T \in \mathcal{T}'} \eta_{\mathcal{T}}^2(V, T) \qquad \text{and} \qquad \mathrm{osc}_{\mathcal{T}}^2(V, \mathcal{T}') := \sum_{T \in \mathcal{T}'} \mathrm{osc}_{\mathcal{T}}^2(V, T).$$

Remark 2.1. Let $\mathcal{T} \in \mathbb{T}$ and $V \in \mathbb{V}(\mathcal{T})$ be given. We first observe that the indicator $\eta_{\mathcal{T}}(V, T)$ *dominates* oscillation $\mathrm{osc}_{\mathcal{T}}(V, T)$, i.e., $\mathrm{osc}_{\mathcal{T}}(V, T) \leq \eta_{\mathcal{T}}(V, T)$ for all $T \in \mathcal{T}$. In addition, the definitions of the error indicator and oscillation are fully localized to T; i.e., for any triangulation $\mathcal{T}_* \in \mathbb{T}$, with $T \in \mathcal{T}_*$, there holds

$\eta_{\mathcal{T}}(V,T) = \eta_{\mathcal{T}_*}(V,T)$ and $\operatorname{osc}_{\mathcal{T}}(V,T) = \operatorname{osc}_{\mathcal{T}_*}(V,T)$. Moreover, if $\mathcal{T}_* \geq \mathcal{T}$ is any refinement of $\mathcal{T}$, then, by combining the monotonicity of local mesh sizes and properties of the local L^2 projection, we deduce the *monotonicity* properties

$$\eta_{\mathcal{T}_*}(V,\mathcal{T}_*) \leq \eta_{\mathcal{T}}(V,\mathcal{T}) \quad \text{and} \quad \operatorname{osc}_{\mathcal{T}_*}(V,\mathcal{T}_*) \leq \operatorname{osc}_{\mathcal{T}}(V,\mathcal{T}) \qquad \forall V \in \mathbb{V}(\mathcal{T}).$$

We now recall the well-known upper and lower bounds for the energy error in terms of the residual-type estimator [1, 16, 17, 18, 23, 27].

LEMMA 2.2 (global a posteriori upper and lower bounds). *Let $u \in \mathbb{V}$ be the solution of (2.1), $\mathcal{T} \in \mathbb{T}$, and $U \in \mathbb{V}(\mathcal{T})$ be the Ritz–Galerkin solution (2.3).*

Then there exist a constant C_1, solely depending on $\mathcal{T}_0$ and $1/c_B$, such that

$$|||u - U|||_\Omega^2 \leq C_1\, \eta_{\mathcal{T}}^2(U,\mathcal{T}), \tag{2.6}$$

and a constant C_2, solely depending on $\mathcal{T}_0$ and $1/C_B$, such that

$$C_2\, \eta_{\mathcal{T}}^2(U,\mathcal{T}) \leq |||u - U|||_\Omega^2 + \operatorname{osc}_{\mathcal{T}}^2(U,\mathcal{T}). \tag{2.7}$$

2.5. The module MARK. In the selection of elements we rely on Dörfler marking. Given a grid $\mathcal{T}$, the set of indicators $\{\eta_{\mathcal{T}}(U,T\}_{T\in\mathcal{T}}$, and marking parameter $\theta \in (0,1]$, we suppose that MARK outputs a subset of *marked elements* $\mathcal{M} \subset \mathcal{T}$, i.e.,

$$\mathcal{M} = \mathsf{MARK}(\{\eta_{\mathcal{T}}(U,T\}_{T\in\mathcal{T}}, \mathcal{T}, \theta),$$

such that $\mathcal{M}$ satisfies Dörfler property

$$\eta_{\mathcal{T}}(U,\mathcal{M}) \geq \theta\, \eta_{\mathcal{T}}(U,\mathcal{T}). \tag{2.8}$$

2.6. The module REFINE. We suppose that a function REFINE is at our disposal that implements iterative or recursive bisection; see [2, 13, 14, 15, 24, 25]. When relying on recursive bisection, the distribution of refinement edges has to fulfill some compatibility conditions on $\mathcal{T}_0$. Given a fixed number $b \geq 1$, for any $\mathcal{T} \in \mathbb{T}$ and a subset $\mathcal{M} \subset \mathcal{T}$ of marked elements,

$$\mathcal{T}_* = \mathsf{REFINE}(\mathcal{T},\mathcal{M})$$

outputs a conforming triangulation $\mathcal{T}_* \in \mathbb{T}$, where at least all elements of $\mathcal{M}$ are bisected b times. In particular, this implies that $\mathcal{M} \subset \mathcal{R}_{\mathcal{T}\to\mathcal{T}_*}$.

To ensure conformity of $\mathcal{T}_*$ one usually has to refine additional elements in $\mathcal{T}\setminus\mathcal{M}$ once or several times. In general, the number of these additionally refined elements is not controlled by $\#\mathcal{M}$ that is, $\#\mathcal{T}_* - \#\mathcal{T}$ cannot be bounded by $C\,\#\mathcal{M}$ with a constant C independent of $\mathcal{T}$; C may depend on the refinement level. On the other hand, by arguing with the entire sequence $\{\mathcal{T}_k\}_{k\geq 0}$ of refinements, Binev, Dahmen, and DeVore showed in two dimensions that the *cumulative* number of elements added by conformity does not inflate the total number of marked elements [3, Theorem 2.4]. Stevenson generalized this result to higher dimensions [24, Theorem 6.1].

LEMMA 2.3 (complexity of REFINE). *Assume that $\mathcal{T}_0$ verifies condition* (b) *of section* 4 *in* [24]. *For $k \geq 0$ let $\{\mathcal{T}_k\}_{k\geq 0}$ be any sequence of refinements of $\mathcal{T}_0$ where $\mathcal{T}_{k+1}$ is generated from $\mathcal{T}_k$ by $\mathcal{T}_{k+1} = \mathsf{REFINE}(\mathcal{T}_k, \mathcal{M}_k)$ with a subset $\mathcal{M}_k \subset \mathcal{T}_k$.*

Then there exists a constant C_0 solely depending on $\mathcal{T}_0$ and b such that

$$\#\mathcal{T}_k - \#\mathcal{T}_0 \leq C_0 \sum_{j=0}^{k-1} \#\mathcal{M}_j \qquad \forall k \geq 1.$$

The original result is proved for a sequence of meshes generated by a loop

$$\mathsf{Mark} \to \mathsf{Subdivision} \to \mathsf{Complete},$$

where Mark selects certain elements, Subdivision bisects marked elements only once ($b = 1$), and Complete adds some additional subdivisions to generate a conforming triangulation. To apply the above result when REFINE performs $b > 1$ bisections, the set $\mathcal{M}_k$ is to be understood as a sequence of *single* bisections recorded in sets $\{\mathcal{M}_k(j)\}_{j=1}^b$, which belong to intermediate triangulations between $\mathcal{T}_k$ and $\mathcal{T}_{k+1}$ with $\#\mathcal{M}_k(j) \le 2^{j-1}\#\mathcal{M}_k$, $j = 1, \dots, b$. Then Lemma 2.3 is a direct consequence of [24, Theorem 6.1] because

$$\sum_{j=1}^{b} \#\mathcal{M}_k(j) \le \sum_{j=1}^{b} 2^{j-1}\#\mathcal{M}_k = (2^b - 1)\#\mathcal{M}_k.$$

2.7. Adaptive algorithm. In this section we now collect the modules described in the previous sections into the adaptive finite element method. In doing this, we replace the dependence on the actual triangulation $\mathcal{T}$ by the iteration counter $k \ge 0$, for instance, $\mathbb{V}_k = \mathbb{V}(\mathcal{T}_k)$. The basic loop of AFEM is then given by the following iteration:

AFEM

Given the initial grid $\mathcal{T}_0$ and marking parameter $0 < \theta \le 1$
set $k := 0$ and iterate

(1) $U_k = \mathsf{SOLVE}(\mathcal{T}_k)$;
(2) $\{\eta_k(U_k, T)\}_{T\in\mathcal{T}_k} = \mathsf{ESTIMATE}(U_k, \mathcal{T}_k)$;
(3) $\mathcal{M}_k = \mathsf{MARK}(\{\eta_k(U_k, T)\}_{T\in\mathcal{T}_k}, \mathcal{T}_k, \theta)$;
(4) $\mathcal{T}_{k+1} = \mathsf{REFINE}(\mathcal{M}_k, \mathcal{T}_k)$; $k := k + 1$.

We want to stress that the AFEM is really a *standard algorithm* in that it employs only the error indicators $\{\eta_k(U_k, T)\}_{T\in\mathcal{T}_k}$ and does not use the oscillation indicators $\{\mathrm{osc}_k(U_k, T)\}_{T\in\mathcal{T}_k}$. In addition, the AFEM relies only on a minimal refinement; i.e., an interior node property is not enforced for marked elements.

Several remarks are now in order:

- Recently, Morin, Siebert, and Veeser proved convergence of the above iteration for general marking strategies, including maximum and equidistribution besides Dörfler strategy [20]. The main result is a plain convergence result, i.e.,

$$\lim_{k\to\infty} \|u - U_k\|_{L^2(\Omega)} = \lim_{k\to\infty} \eta_k(U_k, \mathcal{T}_k) = 0.$$

  This does not provide a strict total error reduction between two successive iterations, which is one key ingredient in the optimality proof of section 5.
- By using similar techniques as Diening and Kreuzer for the p-Laplacian [7], we prove in section 4 a contraction property for $\|u - U_k\|^2_{L^2(\Omega)} + \gamma\, \eta_k^2(U_k, \mathcal{T}_k)$ between two successive iterations, with $\gamma > 0$ a suitable scaling factor; we observe that [7] still requires the interior node property. This in turn implies *linear convergence*. Crucial ingredients are the a posteriori *global upper bound* and *Dörfler marking*.

- We prove in section 5 that $\mathcal{T}_k$ exhibits *quasi-optimal cardinality* provided that
  (a) the marking parameter θ satisfies $\theta < \theta_*$, where θ_*, defined explicitly in Assumption 5.8, depends on the ratio $\sqrt{C_2/C_1}$ of constants in (2.6) and (2.7);
  (b) $\mathcal{M}_k$ satisfies (2.8) with minimal cardinality;
  (c) $\mathcal{T}_{k+1}$ is the smallest refinement of $\mathcal{T}_k$ such that $\mathcal{M}_k \subset \mathcal{R}_{\mathcal{T}_k \to \mathcal{T}_{k+1}}$.

  The proof hinges on the a posteriori *lower bound* and a *localized upper bound*.

3. Auxiliary results. In this section we prove a quasi-reduction property of the estimator, which is instrumental for convergence. We also show a perturbation result of oscillation and a localized upper bound, both between two discrete functions, that play an essential role in the optimality proof. Finally we discuss the overlay of two meshes and thereby prove (2.2).

Throughout this paper we use the notation $a \preccurlyeq b$ to indicate $a \le C\, b$, with generic constants C solely depending on shape regularity and thus on $\mathcal{T}_0$, the number of bisections b of marked elements, and the polynomial degree n. Dependence on data $\mathbf{D} = (\mathbf{A}, c)$ and f of (1.1) is traced explicitly. We denote $a \preccurlyeq b \preccurlyeq a$ by $a \approx b$.

3.1. Reduction of the error estimator and oscillation. We relate the error indicators and oscillation of two nested triangulations to each other. The link involves weighted maximum-norms of the coefficient functions $\mathbf{D}$ or their oscillation. These results are the basis for our analyses in sections 4 and 5.

We start by defining the weighted maximum-norm of the coefficients $\mathbf{D}$ and their oscillation. For $m \in \mathbb{N}_0$, $\mathcal{T} \in \mathbb{T}$, and $v \in L^\infty(\Omega)$, we recall that $\Pi_m^\infty v$ is the best $L^\infty(\Omega)$-approximation in the space of discontinuous polynomials of degree $\le m$. We further set $\Pi_{-1}^\infty v = 0$, $P_m^\infty v = v - \Pi_m^\infty v$, and

$$\eta_{\mathcal{T}}^2(\mathbf{D}, T) := h_T^2 \left(\|\operatorname{div} \mathbf{A}\|_{L^\infty(T)}^2 + h_T^{-2} \|\mathbf{A}\|_{L^\infty(\omega_T)}^2 + \|c\|_{L^\infty(T)}^2 \right),$$

$$\operatorname{osc}_{\mathcal{T}}^2(\mathbf{D}, T) := h_T^2 \Big(\left\| P_{n-1}^\infty \operatorname{div} \mathbf{A} \right\|_{L^\infty(T)}^2 + h_T^{-2} \|P_n^\infty \mathbf{A}\|_{L^\infty(\omega_T)}^2 + h_T^2 \left\| P_{n-2}^\infty c \right\|_{L^\infty(T)}^2 + \left\| P_{n-1}^\infty c \right\|_{L^\infty(T)}^2 \Big);$$

note that P_m^∞ is defined elementwise. For any subset $\mathcal{T}' \subset \mathcal{T}$ we finally set

$$\eta_{\mathcal{T}}(\mathbf{D}, \mathcal{T}') := \max_{T \in \mathcal{T}'} \eta_{\mathcal{T}}(\mathbf{D}, T) \qquad \text{and} \qquad \operatorname{osc}_{\mathcal{T}}(\mathbf{D}, \mathcal{T}') := \max_{T \in \mathcal{T}'} \operatorname{osc}_{\mathcal{T}}(\mathbf{D}, T).$$

Remark 3.1 (monotonicity). The use of best approximation in L^∞ in the definitions of $\eta_{\mathcal{T}}(\mathbf{D}, \mathcal{T})$ and $\operatorname{osc}_{\mathcal{T}}(\mathbf{D}, \mathcal{T})$ implies the following monotonicity property: for all $\mathcal{T}, \mathcal{T}_* \in \mathbb{T}$, with $\mathcal{T} \le \mathcal{T}_*$, there holds

$$\eta_{\mathcal{T}_*}(\mathbf{D}, \mathcal{T}_*) \le \eta_{\mathcal{T}}(\mathbf{D}, \mathcal{T}) \quad \text{and} \quad \operatorname{osc}_{\mathcal{T}_*}(\mathbf{D}, \mathcal{T}_*) \le \operatorname{osc}_{\mathcal{T}}(\mathbf{D}, \mathcal{T}).$$

The following lemma allows us to avoid any smoothness assumption on the coefficients $\mathbf{D}$ of the PDE.

LEMMA 3.2 (implicit interpolation). *Let ω be either a d or a $(d-1)$ simplex. For $\ell \in \mathbb{N}$ denote by $\Pi_m^p \colon L^p(\omega, \mathbb{R}^\ell) \to \mathbb{P}_m(\omega, \mathbb{R}^\ell)$ the operator of best L^p-approximation in ω, and set $P_m^p = \mathrm{id} - \Pi_m^p$. Then for all $v \in L^\infty(\omega, \mathbb{R}^\ell)$, $V \in \mathbb{P}_n(\omega, \mathbb{R}^\ell)$ and $m \ge n$, there holds*

$$\left\| P_m^2(vV) \right\|_{L^2(\omega)} \le \left\| P_{m-n}^\infty v \right\|_{L^\infty(\omega)} \|V\|_{L^2(\omega)}.$$

Proof. Since $V \in \mathbb{P}_n$ we obtain $\Pi^\infty_{m-n}(v)\cdot V \in \mathbb{P}_m$, and thus the orthogonality of the L^2-projection yields

$$\begin{aligned}\int_\omega (vV - \Pi^2_m(vV))(vV - \Pi^2_m(vV)) &= \int_\omega (vV - \Pi^2_m(vV))(v - \Pi^\infty_{m-n}(v))\cdot V \\ &\le \left\|P^2_m(vV)\right\|_{L^2(\omega)} \left\|P^\infty_{m-n} v\right\|_{L^\infty(\omega)} \|V\|_{L^2(\omega)},\end{aligned}$$

which finishes the proof. ∎

PROPOSITION 3.3 (local perturbation). *Let $\mathcal{T} \in \mathbb{T}$. For all $T \in \mathcal{T}$ and for any pair of discrete functions $V, W \in \mathbb{V}(\mathcal{T})$, we have*

$$\eta_{\mathcal{T}}(V,T) \le \eta_{\mathcal{T}}(W,T) + \bar{\Lambda}_1\, \eta_{\mathcal{T}}(\mathbf{D},T)\, \|V - W\|_{H^1(\omega_T)}, \tag{3.1}$$

$$\operatorname{osc}_{\mathcal{T}}(V,T) \le \operatorname{osc}_{\mathcal{T}}(W,T) + \bar{\Lambda}_1 \operatorname{osc}_{\mathcal{T}}(\mathbf{D},T)\|V - W\|_{H^1(\omega_T)}, \tag{3.2}$$

where ω_T is the union of elements in $\mathcal{T}$ sharing a side with T. The constant $\bar{\Lambda}_1 > 0$ depends only on the shape regularity of $\mathcal{T}_0$ and the polynomial degree n.

Proof. We prove only (3.2); estimate (3.1) is somewhat simpler and can be derived similarly. Recall the definition of the element residual $R(V) = f + \mathcal{L}(V)$ and the notation $P^p_m = \mathrm{id} - \Pi^p_m$. Since the L^2-projection Π^2_m is linear, by adding and subtracting W and using the triangle inequality, we obtain

$$\operatorname{osc}_{\mathcal{T}}(V,T) \le \operatorname{osc}_{\mathcal{T}}(W,T) + h_T \left\|P^2_{2n-2}\mathcal{L}(E)\right\|_{L^2(T)} + h_T^{1/2} \left\|P^2_{2n-1} J(E)\right\|_{L^2(\partial T)},$$

with $E := V - W$. It remains to show that the second and third terms are bounded by $\operatorname{osc}_{\mathcal{T}}(\mathbf{D},T)$ times the local H^1-norm of E. For $\mathcal{L}(E) = \operatorname{div}\mathbf{A}\nabla E - c\,E$ we have

$$\left\|P^2_{2n-2}\mathcal{L}(E)\right\|_{L^2(T)} \le \left\|P^2_{2n-2}(\operatorname{div}\mathbf{A}\nabla E)\right\|_{L^2(T)} + \left\|P^2_{2n-2}(cE)\right\|_{L^2(T)}.$$

We split the divergence term as

$$\operatorname{div}(\mathbf{A}\nabla E) = \operatorname{div}\mathbf{A}\cdot\nabla E + \mathbf{A}\colon D^2 E,$$

where D^2E is the Hessian of E. By invoking Lemma 3.2 with $\omega = T$ and observing that the polynomial degree of ∇E is $n-1$, we infer for the first term that

$$\left\|P^2_{2n-2}(\operatorname{div}\mathbf{A}\cdot\nabla E)\right\|_{L^2(T)} \le \left\|P^\infty_{n-1}\operatorname{div}\mathbf{A}\right\|_{L^\infty(T)} \|\nabla E\|_{L^2(T)}.$$

Since D^2E is a polynomial of degree $\le n-2$, by applying Lemma 3.2 again in conjunction with an inverse inequality, we have for the second term

$$\left\|P^2_{2n-2}(\mathbf{A} : D^2E)\right\|_{L^2(T)} \preccurlyeq h_T^{-1} \|P^\infty_n \mathbf{A}\|_{L^\infty(T)} \|\nabla E\|_{L^2(T)}.$$

To analyze the reaction term we write

$$\left\|P^2_{2n-2}(c\,E)\right\|_{L^2(T)} \le \left\|P^2_{2n-2}(c\,\Pi^2_0 E)\right\|_{L^2(T)} + \left\|P^2_{2n-2}(c\,P^2_0 E)\right\|_{L^2(T)}.$$

By applying again Lemma 3.2 we have for the first term

$$\left\|P^2_{2n-2}(c\,\Pi^2_0 E)\right\|_{L^2(T)} \le \left\|P^\infty_{2n-2} c\right\|_{L^\infty(T)} \|E\|_{L^2(T)}$$

and for the last one

$$\left\|P^2_{2n-2}(c\,P^2_0 E)\right\|_{L^2(T)} \preccurlyeq h_T \left\|P^\infty_{n-2} c\right\|_{L^\infty(T)} \|\nabla E\|_{L^2(T)}.$$

We now deal with the jump residual. Let $T' \in \mathcal{T}$ share an interior side σ with T. We write $J(E) = ((\mathbf{A}\nabla E)|_T - (\mathbf{A}\nabla E)|_{T'}) \cdot \boldsymbol{\nu}$ and use the linearity of Π^2_{2n-1}, Lemma 3.2 with $\omega = \sigma$, and the inverse inequality $\|\nabla E\|_{L^2(\sigma)} \preccurlyeq h_T^{-1/2} \|\nabla E\|_{L^2(T)}$ to deduce that

$$\left\|P^2_{2n-1}(\mathbf{A}\nabla E)|_T \cdot \boldsymbol{\nu}\right\|_{L^2(\sigma)} \preccurlyeq h_T^{-\frac{1}{2}} \|P^\infty_n \mathbf{A}\|_{L^\infty(T)} \|\nabla E\|_{L^2(T)}.$$

The same argument holds for T', and, since $\mathcal{T}$ is shape-regular, we can replace $h_{T'}$ by h_T. Finally, collecting the above estimates for T and all of its neighbors yields the assertion (3.2). ∎

The following two corollaries are global forms of the above lemma.

COROLLARY 3.4 (estimator reduction). *For $\mathcal{T} \in \mathbb{T}$ and $\mathcal{M} \subset \mathcal{T}$ let $\mathcal{T}_* \in \mathbb{T}$ be given by $\mathcal{T}_* := \mathsf{REFINE}(\mathcal{T}, \mathcal{M})$. If $\Lambda_1 := (d+1)\bar{\Lambda}_1^2/c_B$, with $\bar{\Lambda}_1$ from Proposition 3.3 and $\lambda := 1 - 2^{-\frac{b}{d}} > 0$, then there holds for all $V \in \mathbb{V}(\mathcal{T})$ and $V_* \in \mathbb{V}(\mathcal{T}_*)$ and any $\delta > 0$*

$$\begin{aligned}\eta^2_{\mathcal{T}_*}(V_*, \mathcal{T}_*) \le (1+\delta)\left\{\eta^2_{\mathcal{T}}(V, \mathcal{T}) - \lambda\, \eta^2_{\mathcal{T}}(V, \mathcal{M})\right\} \\ + (1+\delta^{-1})\Lambda_1 \eta^2_{\mathcal{T}_0}(\mathbf{D}, \mathcal{T}_0)\, |\!|\!| V_* - V |\!|\!|^2_\Omega .\end{aligned}$$

Proof. By applying Proposition 3.3 with $V_*, V \in \mathbb{V}(\mathcal{T}_*)$ over $T \in \mathcal{T}_*$ and using Young's inequality with parameter δ, we derive

$$\eta^2_{\mathcal{T}_*}(V_*, T) \le (1+\delta)\eta^2_{\mathcal{T}_*}(V, T) + (1+\delta^{-1})\bar{\Lambda}_1^2\, \eta^2_{\mathcal{T}_*}(\mathbf{D}, T)\, \|V_* - V\|^2_{H^1(\omega_T)}.$$

By summing over all elements $T \in \mathcal{T}_*$, using the finite overlap property of patches ω_T, and the equivalence of the H^1-norm and the energy norm in Ω, we have the following direct by-product:

$$\eta^2_{\mathcal{T}_*}(V_*, \mathcal{T}_*) \le (1+\delta)\eta^2_{\mathcal{T}_*}(V, \mathcal{T}_*) + (1+\delta^{-1})\Lambda_1 \eta^2_{\mathcal{T}_*}(\mathbf{D}, \mathcal{T}_*)\, |\!|\!| V_* - V |\!|\!|^2_\Omega .$$

For a marked element $T \in \mathcal{M}$, we set $\mathcal{T}_{*,T} := \{T' \in \mathcal{T}_* \mid T' \subset T\}$. Since $V \in \mathbb{V}(\mathcal{T})$ and $\mathbf{A}$ jumps only across sides of $\mathcal{T}_0$, we see that $J(V) = 0$ on sides of $\mathcal{T}_{*,T}$ in the interior of T. We then obtain

$$\sum_{T' \in \mathcal{T}_{*,T}} \eta^2_{\mathcal{T}_*}(V, T') \le 2^{-\frac{b}{d}} \eta^2_{\mathcal{T}}(V, T),$$

because refinement by bisection implies that $h_{T'} = |T'|^{\frac{1}{d}} \le (2^{-b}|T|)^{\frac{1}{d}} \le 2^{-\frac{b}{d}} h_T$ for all $T' \in \mathcal{T}_{*,T}$. For an element $T \in \mathcal{T} \setminus \mathcal{M}$, instead, Remark 2.1 yields $\eta_{\mathcal{T}_*}(V, T) \le \eta_{\mathcal{T}}(V, T)$. Hence, by summing over all $T \in \mathcal{T}_*$, we arrive at

$$\eta^2_{\mathcal{T}_*}(V, \mathcal{T}_*) \le \eta^2_{\mathcal{T}}(V, \mathcal{T} \setminus \mathcal{M}) + 2^{-\frac{b}{d}} \eta^2_{\mathcal{T}}(V, \mathcal{M}) = \eta^2_{\mathcal{T}}(V, \mathcal{T}) - \lambda \eta^2_{\mathcal{T}}(V, \mathcal{M}).$$

The assertion finally follows from the monotonicity $\eta_{\mathcal{T}_*}(\mathbf{D}, \mathcal{T}_*) \le \eta_{\mathcal{T}_0}(\mathbf{D}, \mathcal{T}_0)$ stated in Remark 3.1. ∎

COROLLARY 3.5 (perturbation of oscillation). *Let $\mathcal{T}, \mathcal{T}_* \in \mathbb{T}$, with $\mathcal{T} \le \mathcal{T}_*$, and let $\Lambda_1 = (d+1)\bar{\Lambda}_1^2/c_B$ be as in Corollary 3.4. For all $V \in \mathbb{V}(\mathcal{T})$ and $V_* \in \mathbb{V}(\mathcal{T}_*)$, we have*

$$\operatorname{osc}^2_{\mathcal{T}}(V, \mathcal{T} \cap \mathcal{T}_*) \le 2 \operatorname{osc}^2_{\mathcal{T}_*}(V_*, \mathcal{T} \cap \mathcal{T}_*) + 2\Lambda_1 \operatorname{osc}^2_{\mathcal{T}_0}(\mathbf{D}, \mathcal{T}_0)\, |\!|\!| V - V_* |\!|\!|^2_\Omega .$$

Proof. Remark 2.1 yields $\operatorname{osc}_{\mathcal{T}}(V, T) = \operatorname{osc}_{\mathcal{T}_*}(V, T)$ for all $T \in \mathcal{T} \cap \mathcal{T}_*$, whence

$$\operatorname{osc}^2_{\mathcal{T}}(V, T) \le 2 \operatorname{osc}^2_{\mathcal{T}_*}(V_*, T) + 2\bar{\Lambda}_1^2 \operatorname{osc}^2_{\mathcal{T}_*}(\mathbf{D}, T) \|V - V_*\|^2_{H^1(\omega_T)},$$

by (3.2) and Young's inequality. We now sum over $T \in \mathcal{T} \cap \mathcal{T}_*$ and use the equivalence of H^1 and energy norms in Ω, as well as the monotonicity property $\operatorname{osc}_{\mathcal{T}_*}(\mathbf{D}, \mathcal{T}_*) \le \operatorname{osc}_{\mathcal{T}_0}(\mathbf{D}, \mathcal{T}_0)$ stated in Remark 3.1, to prove the assertion. ∎

3.2. Localized upper bound. To prove the optimality of the AFEM, we need a localized upper bound for the distance between two nested solutions. This slightly improves a similar result by Stevenson [23] in the sense that the error can be estimated here by using only the indicators of refined elements without a buffer layer.

LEMMA 3.6 (localized upper bound). *For $\mathcal{T}, \mathcal{T}_* \in \mathbb{T}$, with $\mathcal{T} \leq \mathcal{T}_*$, let $\mathcal{R} = \mathcal{R}_{\mathcal{T}\to\mathcal{T}_*}$ be the set of refined elements. Let $U \in \mathbb{V}(\mathcal{T})$ and $U_* \in \mathbb{V}(\mathcal{T}_*)$ be the discrete solutions of* (2.1)*. Then the following localized upper bound is valid:*

$$|||U_* - U|||_\Omega^2 \leq C_1\, \eta_T^2(U, \mathcal{R}).$$

Proof. Let $\Omega_* = \bigcup\{T\colon T \in \mathcal{R}\}$ be the union of refined elements, and denote by Ω_i, $i = 1, \dots, I$, the connected components of its interior. For $i \in \{1, \dots, I\}$ we set $\mathcal{T}_i = \{T \in \mathcal{T} : T \subset \bar{\Omega}_i\}$ and let $\mathbb{V}(\mathcal{T}_i)$ be the restriction of $\mathbb{V}(\mathcal{T})$ to Ω_i. By using the Scott–Zhang interpolation operator $P_i\colon H^1(\Omega_i) \to \mathbb{V}(\mathcal{T}_i)$ over the triangulation $\mathcal{T}_i$, the following interpolation estimate holds true for all $v \in H^1(\mathcal{T}_i)$:

$$\sum_{T\in\mathcal{T}_i} h_T^{-2}\|v - P_i v\|_{L_2(T)}^2 + h_T^{-1}\|v - P_i v\|_{L_2(\partial T)}^2 \preccurlyeq \|\nabla v\|_{L^2(\Omega_i)}^2. \tag{3.3}$$

The constant hidden in "$\preccurlyeq$" does not depend on Ω_i but only on the shape-regularity of the underlying triangulation $\mathcal{T}_i$ and thus solely on $\mathcal{T}_0$ [22]. By construction, P_i preserves conforming boundary values, i.e., $P_i v = v$ on $\partial\Omega_i$ whenever $v = V$ on $\partial\Omega_i$ for some $V \in \mathbb{V}(\mathcal{T}_i)$.

For the error $E_* := U_* - U \in \mathbb{V}(\mathcal{T}_*)$ we construct an approximation $V \in \mathbb{V}(\mathcal{T})$ by setting

$$V := \begin{cases} P_i E_* & \text{in } \Omega_i, \\ E_* & \text{elsewhere.} \end{cases}$$

Since E_* has conforming boundary values on $\partial\Omega_i$ in $\mathbb{V}(\mathcal{T}_i)$, we conclude that V is continuous in Ω implying that $V \in \mathbb{V}(\mathcal{T})$ and V is an H^1-stable approximation to E_*.

Nesting of spaces $\mathbb{V}(\mathcal{T}) \subset \mathbb{V}(\mathcal{T}_*)$ implies Galerkin orthogonality $\mathcal{B}[E_*, E_*] = \mathcal{B}[E_*, E_* - V]$. In conjunction with $E_* - V = 0$ in $\Omega \setminus \Omega_*$ we thus obtain by standard arguments

$$\mathcal{B}[E_*, E_*] = \sum_{T\in\mathcal{R}} \langle R(U), E_* - V\rangle_T + \tfrac{1}{2}\langle J(U), E_* - V\rangle_{\partial T} \preccurlyeq \eta_{\mathcal{T}}(U, \mathcal{R})\|\nabla E_*\|_{L^2(\Omega_*)}.$$

In the last step we have used (3.3) on all connected components Ω_i of Ω_* with $v = E_*$. This, in conjunction with the coercivity of $\mathcal{B}$, proves the proposition. ∎

3.3. Overlay of meshes. We finish the auxiliary results with a counting argument for the overlay $\mathcal{T}_1 \oplus \mathcal{T}_2$ of two triangulations $\mathcal{T}_1$ and $\mathcal{T}_2$ from [23]. For the convenience of the reader we include a proof of this crucial result. As a consequence of the following lemma we see that for two conforming triangulations $\mathcal{T}_1, \mathcal{T}_2 \in \mathbb{T}$ the overlay is the smallest conforming triangulation $\mathcal{T} \in \mathbb{T}$, with $\mathcal{T}_1, \mathcal{T}_2 \leq \mathcal{T}$.

LEMMA 3.7 (overlay of meshes). *For $\mathcal{T}_1, \mathcal{T}_2 \in \mathbb{T}$ the overlay $\mathcal{T} := \mathcal{T}_1 \oplus \mathcal{T}_2$ is conforming, i.e., $\mathcal{T} \in \mathbb{T}$, and satisfies*

$$\#\mathcal{T} \leq \#\mathcal{T}_1 + \#\mathcal{T}_2 - \#\mathcal{T}_0.$$

Proof. Assume that $\mathcal{T}$ contains a nonconforming vertex z. Then there exist $T_1, T_2 \in \mathcal{T}$ with a common edge such that z is a vertex of T_1 and $z \in T_2$ but z is no

vertex of T_2. Without loss of generality let $T_1 \in \mathcal{T}_1$. Since $\mathcal{T}_1$ is conforming, there exists a $T' \in \mathcal{T}_1$, $T' \subset T_2$ such that z is a vertex of T'. Hence, T_2 cannot be a leaf node of $\mathcal{F}(\mathcal{T})$, i.e., $T_2 \notin \mathcal{T}$, a contradiction.

For $T \in \mathcal{T}_0$ and $i = 1, 2$ we denote by $\mathcal{F}_i(T) \subset \mathcal{F}(\mathcal{T})$ the binary trees with root T corresponding to $\mathcal{T}_i$ and let $\mathcal{T}_i(T)$ be the triangulation given by the leaf nodes of $\mathcal{F}_i(T)$. Since $\mathcal{T}(T) \subset \mathcal{T}_1(T) \cup \mathcal{T}_2(T)$, we infer that $\#\mathcal{T}(T) \leq \#\mathcal{T}_1(T) + \#\mathcal{T}_2(T)$. We now show that $\#\mathcal{T}(T) \leq \#\mathcal{T}_1(T) + \#\mathcal{T}_2(T) - 1$ by distinguishing two cases.

Case 1: $\mathcal{T}_1(T) \cap \mathcal{T}_2(T) \neq \emptyset$. Then there exists $T' \in \mathcal{T}_1(T) \cap \mathcal{T}_2(T)$, and so $T' \in \mathcal{T}(T)$. By counting T' only once in $\#(\mathcal{T}_1(T) \cup \mathcal{T}_2(T))$ we get $\#\mathcal{T}(T) \leq \#\mathcal{T}_1(T) + \#\mathcal{T}_2(T) - 1$.

Case 2: $\mathcal{T}_1(T) \cap \mathcal{T}_2(T) = \emptyset$. Then there exists $T' \in \mathcal{T}_1(T)$ (resp., $T' \in \mathcal{T}_2(T)$) so that $T' \notin \mathcal{T}(T)$, for otherwise $T' \in \mathcal{T}_2(T)$ (resp., $T' \in \mathcal{T}_1(T)$), thereby contradicting the assumption. We obtain again $\#\mathcal{T}(T) \leq \#\mathcal{T}_1(T) + \#\mathcal{T}_2(T) - 1$.

Finally, since $\mathcal{T}_i = \bigcup_{T \in \mathcal{T}_0} \mathcal{T}_i(T)$, the assertion follows by adding over $T \in \mathcal{T}_0$. ∎

4. Contraction property of the AFEM. We now prove that the AFEM is a contraction with respect to the sum of the energy error plus the scaled error estimator, the so-called *quasi error*. Consequently, the quasi error is reduced by a fixed rate at every step. This can be motivated heuristically as follows: in light of (2.4) the energy error $|||u - U_k|||_\Omega$ decreases strictly, unless $U_{k+1} = U_k$, in which case the estimator $\eta_k(U_k, \mathcal{T}_k)$ does according to Corollary 3.4. The theorem below makes this observation quantitative.

THEOREM 4.1 (contraction property). *Let $\theta \in (0, 1]$, and let $\{\mathcal{T}_k, \mathbb{V}_k, U_k\}_{k \geq 0}$ be the sequence of meshes, finite element spaces, and discrete solutions produced by the AFEM.*

Then there exist constants $\gamma > 0$ and $0 < \alpha < 1$, depending solely on the shape-regularity of $\mathcal{T}_0$, b, and the marking *parameter $0 < \theta \leq 1$, such that*

$$|||u - U_{k+1}|||_\Omega^2 + \gamma\, \eta_{k+1}^2(U_{k+1}, \mathcal{T}_{k+1}) \leq \alpha^2 \Big(|||u - U_k|||_\Omega^2 + \gamma\, \eta_k^2(U_k, \mathcal{T}_k) \Big).$$

Proof. For convenience, we use the notation

$$e_k := |||u - U_k|||_\Omega\,, \qquad E_k := |||U_{k+1} - U_k|||_\Omega\,,$$
$$\eta_k := \eta_k(U_k, \mathcal{T}_k), \quad \eta_k(\mathcal{M}_k) := \eta_k(U_k, \mathcal{M}_k), \quad \eta_0(\mathbf{D}) := \eta_0(\mathbf{D}, \mathcal{T}_0).$$

We combine the orthogonality (2.4) with Corollary 3.4 to write

$$e_{k+1}^2 + \gamma\, \eta_{k+1}^2 \leq e_k^2 - E_k^2 + (1 + \delta)\, \gamma \left(\eta_k^2 - \lambda\, \eta_k^2(\mathcal{M}_k) \right) + (1 + \delta^{-1})\, \gamma\, \Lambda_1\, \eta_0^2(\mathbf{D}) E_k^2.$$

We choose γ dependent on δ to be

$$\gamma := \frac{1}{(1 + \delta^{-1})\, \Lambda_1\, \eta_0^2(\mathbf{D})} \qquad \Leftrightarrow \qquad \gamma\,(1 + \delta) = \frac{\delta}{\Lambda_1\, \eta_0^2(\mathbf{D})} \tag{4.1}$$

to obtain

$$e_{k+1}^2 + \gamma\, \eta_{k+1}^2 \leq e_k^2 + (1 + \delta)\, \gamma\, \eta_k^2 - (1 + \delta)\, \lambda\, \gamma\, \eta_k^2(\mathcal{M}_k).$$

By invoking Dörfler marking (2.8), we deduce that

$$e_{k+1}^2 + \gamma\, \eta_{k+1}^2 \leq e_k^2 + (1 + \delta)\, \gamma\, \eta_k^2 - (1 + \delta)\, \lambda\, \theta^2\, \gamma\, \eta_k^2.$$

We rewrite this inequality as follows with any $\beta \in (0,1)$:

$$e_{k+1}^2 + \gamma\,\eta_{k+1}^2 \leq e_k^2 + (1+\delta)\,\gamma\,\eta_k^2 - \beta\,(1+\delta)\,\lambda\,\theta^2\,\gamma\,\eta_k^2 - (1-\beta)\,(1+\delta)\,\lambda\,\theta^2\,\gamma\,\eta_k^2,$$

apply the upper bound (2.6), and replace γ according to (4.1) to obtain

$$e_{k+1}^2 + \gamma\,\eta_{k+1}^2 \leq \alpha_1^2(\delta,\beta)\,e_k^2 + \gamma\,\alpha_2^2(\delta,\beta)\,\eta_k^2,$$

with

$$\alpha_1^2(\delta,\beta) := 1 - \beta\,\frac{\lambda\,\theta^2}{C_1\Lambda_1\,\eta_0^2(\mathbf{D})}\,\delta, \qquad \alpha_2^2(\delta,\beta) := (1+\delta)\big(1-(1-\beta)\,\lambda\,\theta^2\big).$$

Now choosing $\delta > 0$ small enough yields

$$\alpha^2 := \max\{\alpha_1^2,\,\alpha_2^2\} < 1,$$

which is the desired result. ∎

Remark 4.2 (ingredients for convergence). We stress that this new proof of linear convergence relies exclusively on the upper bound (2.6), the orthogonality relation (2.4), the error estimator reduction property of Corollary 3.4, and the Dörfler marking (2.8) for the estimator. It does not need any marking due to oscillation, which turns out to be problematic for optimality in light of the discussion in section 6. Equality (2.4) is used only to cancel, via (4.1), the contribution involving $\eta_0(\mathbf{D},\mathcal{T}_0)$ in Corollary 3.4. Its role is much less prominent than in [8, 16, 17, 18, 23].

Moreover, we use neither the lower bound (2.7) nor a discrete lower bound for proving convergence. The latter hinges on the rather demanding *interior node property*: every element of $\mathcal{M}_k$, as well as its adjacent elements, contains a node of $\mathcal{T}_{k+1}$ in their interior as well as in the interior of their common sides. However, the global lower bound (2.7) will be instrumental to prove optimality in section 5.

The treatment of oscillation and the interior node property is essential in [16, 17, 18] and so in [3, 23]. Our new approach simplifies the analysis and directly applies to any polynomial degree $n \geq 1$.

Remark 4.3 (optimal contraction factor α). Consider

$$D := \{(\delta,\beta) \in \mathbb{R}_+ \times [0,1] : \quad 0 \leq \alpha_1^2(\delta,\beta), \alpha_2^2(\delta,\beta) \leq 1\}.$$

This set is nonempty, according to the proof of Theorem 4.1, closed, and bounded; thus D is compact. Since α_1^2 and α_2^2 are continuous functions in D, $\alpha^2 = \max\{\alpha_1^2, \alpha_2^2\}$ attains its minimum in D. It turns out that $\alpha^2|_{\partial D} = 1$ by definition of D and $\alpha^2 < 1$ in the interior of D from the proof of Theorem 4.1. Consequently, α^2 attains an absolute minimum smaller than 1 and satisfies

$$\alpha^2 = \alpha_1^2 = \alpha_2^2.$$

To see this assume that $\alpha_1^2 < \alpha_2^2$, and decrease the value of δ slightly. Since α_1^2 increases whereas α_2^2 decreases, this yields a contradiction to the minimality of α^2. The case $\alpha_1^2 > \alpha_2^2$ is similar. In principle, this optimal value of α^2 can be computed explicity.

Remark 4.4 (range of γ and α). We see from (4.1) that $\gamma \approx \eta_0^{-2}(\mathbf{D},\mathcal{T}_0)$ provided that $\eta_0(\mathbf{D},\mathcal{T}_0)$ is large; this provides a lower bound for γ. An upper bound results from the condition $\alpha_1^2 > 0$ and $\delta \leq 1$, namely,

$$\alpha_1^2 = 1 - \frac{(1+\delta)\beta\lambda\theta^2}{C_1}\gamma \qquad \Rightarrow \qquad \gamma < \frac{C_1}{2\beta\lambda\theta^2} \approx C_1.$$

On the other hand, it is clear from the definitions of α_1 and α_2 that α deteriorates if either $\eta_0(\mathbf{D}, \mathcal{T}_0)$ increases or θ decreases. In fact, take $\beta = \frac{1}{2}$ and $\delta = \frac{1}{2}\lambda\theta^2$ to get

$$\alpha_1^2 = 1 - \frac{1}{C_1\Lambda_1\eta_0^2(\mathbf{D}, \mathcal{T}_0)}\Big(\frac{\lambda\theta^2}{2}\Big)^2, \qquad \alpha_2^2 = 1 - \Big(\frac{\lambda\theta^2}{2}\Big)^2.$$

Thus by the definition of α there exist constants $c, C > 0$ not depending on θ, with

$$1 - C\,\theta^4 \le \alpha^2 \le 1 - c\,\theta^4.$$

5. Quasi-optimal cardinality of the AFEM. Building on fundamental ideas of Binev, Dahmen, and DeVore [3] and Stevenson [23], used in the first optimality proofs for variants of the AFEM for the Poisson equation, we prove now the optimal cardinality of a *standard* AFEM for the general symmetric elliptic problem (1.1). We thus improve and extend the results of [3, 23].

5.1. Approximation class. Procedure MARK selects elements for refinement based on information provided by the error indicators $\eta_{\mathcal{T}}(U, T)$ for all $T \in \mathcal{T}$. Hence, a decay rate of the AFEM in terms of DOFs is closely related to the behavior of the estimator $\eta_{\mathcal{T}}(U, \mathcal{T})$. In order to link the estimator to the energy error $|\!|\!| u - U |\!|\!|_\Omega$, we observe that, according to the upper bound (2.6)

$$|\!|\!| u - U |\!|\!|_\Omega^2 \le C_1\,\eta_{\mathcal{T}}^2(U, \mathcal{T}),$$

a decay rate for the estimator directly implies a decay rate for the error. However, the error may decay faster than the estimator whenever overestimation occurs. The quality of the upper bound is described by the global lower bound (2.7)

$$C_2\,\eta_{\mathcal{T}}^2(U, \mathcal{T}) \le |\!|\!| u - U |\!|\!|_\Omega^2 + \operatorname{osc}_{\mathcal{T}}^2(U, \mathcal{T}).$$

Hence, the quality of the estimator strongly depends on the resolution of data $(\mathbf{D}, f)$ as well as the interaction of $\mathbf{D}$ with the discrete solution U, encoded in the oscillation term $\operatorname{osc}_{\mathcal{T}}(U, \mathcal{T})$. The latter cannot be removed as the following example shows.

Remark 5.1 (role of oscillation). The presence of $\operatorname{osc}_{\mathcal{T}}(U, \mathcal{T})$ in the lower bound is the price to pay for having a simple and computable estimator $\eta_{\mathcal{T}}(U, \mathcal{T})$. To see that oscillation cannot be removed, we let $\epsilon = 2^{-K}$ for K integer and extend the function $\frac{1}{2}x(\epsilon - |x|)$ defined on $(-\epsilon, \epsilon)$ to a 2ϵ-periodic C^1 function u on $\Omega = (-1, 1)$. We then consider u for data $\mathbf{D} = (1, 0)$ to be the exact solution, whence the corresponding right-hand side f is 2ϵ-periodic and piecewise constant with values $+1$ and -1 that change at multiples of ϵ.

Let $\mathcal{T}$ be a uniform mesh with mesh size $h = 2^{-k}$, with $k \ll K$. We consider piecewise linear approximation ($n = 1$), so that f is orthogonal to the space of piecewise constants, i.e., $\Pi_0^2 f = 0$. It is easy to verify that $U = 0$ and

$$|\!|\!| u - U |\!|\!|_\Omega = |\!|\!| u |\!|\!|_\Omega = \frac{\epsilon}{\sqrt{6}} = \frac{2^{-K}}{\sqrt{6}} \ll 2^{-k} = h = \|hf\|_\Omega = \operatorname{osc}_{\mathcal{T}}(U, \mathcal{T}) = \eta_{\mathcal{T}}(U, \mathcal{T}).$$

Therefore, the ratio between $|\!|\!| u - U |\!|\!|_\Omega$ and $\eta_{\mathcal{T}}(U, \mathcal{T})$ can be made arbitrarily small by increasing K/k, and $\operatorname{osc}_{\mathcal{T}}(U, \mathcal{T})$ accounts for the discrepancy. Moreover, if we were to express the oscillation in $H^{-1}(\Omega)$ as suggested in [3, 23], namely, $\|f - \Pi_0^2 f\|_{H^{-1}(\Omega)} = \|f\|_{H^{-1}(\Omega)} \approx \epsilon$, then $\operatorname{osc}_{\mathcal{T}}(U, \mathcal{T})$ would overestimate this quantity. This would correspond to evaluating the estimator in $H^{-1}(\Omega)$, which is impractical.

According to Remark 2.1 the estimator dominates oscillation, i.e., $\operatorname{osc}_{\mathcal{T}}^2(U,\mathcal{T}) \leq \eta_{\mathcal{T}}^2(U,\mathcal{T})$, which in combination with the upper bound yields

$$\|\!|u-U|\!\|_\Omega^2 + \operatorname{osc}_{\mathcal{T}}^2(U,\mathcal{T}) \leq (C_1+1)\,\eta_{\mathcal{T}}^2(U,\mathcal{T}).$$

We thus realize that

$$\eta_{\mathcal{T}}(U,\mathcal{T}) \approx \Big(\|\!|u-U|\!\|_\Omega^2 + \operatorname{osc}_{\mathcal{T}}^2(U,\mathcal{T})\Big)^{1/2},$$

where the right-hand side is the so-called *total error*. Since all decisions of the AFEM are based on the estimator, a decay rate for the AFEM can be characterized only by properties of the total error, and this has to be reflected in the definition of a suitable approximation class. By invoking the upper bound once again, we also see that the total error is equivalent to the quasi error

$$\|\!|u-U|\!\|_\Omega^2 + \operatorname{osc}_{\mathcal{T}}^2(U,\mathcal{T}) \approx \|\!|u-U|\!\|_\Omega^2 + \eta_{\mathcal{T}}^2(U,\mathcal{T}),$$

which is strictly reduced by the AFEM. Finally, the total error satisfies Cea's lemma.

LEMMA 5.2 (quasi optimality of the total error). *Let u be the solution of* (2.1)*, and for $\mathcal{T}\in\mathbb{T}$ let $U\in\mathbb{V}(\mathcal{T})$ be the Ritz–Galerkin approximation of* (2.3)*.*

Then there exists a constant $C_{\mathbf{D}}$ depending only on data $\mathbf{D}$ and shape-regularity of $\mathcal{T}_0$ such that

$$\|\!|u-U|\!\|_\Omega^2 + \operatorname{osc}_{\mathcal{T}}^2(U,\mathcal{T}) \leq C_{\mathbf{D}} \inf_{V\in\mathbb{V}(\mathcal{T})} \Big(\|\!|u-V|\!\|_\Omega^2 + \operatorname{osc}_{\mathcal{T}}^2(V,\mathcal{T})\Big).$$

Proof. For $\epsilon>0$ choose $V_\epsilon\in\mathbb{V}(\mathcal{T})$, with

$$\|\!|u-V_\epsilon|\!\|_\Omega^2 + \operatorname{osc}_{\mathcal{T}}^2(V_\epsilon,\mathcal{T}) \leq (1+\epsilon) \inf_{V\in\mathbb{V}(\mathcal{T})} \Big(\|\!|u-V|\!\|_\Omega^2 + \operatorname{osc}_{\mathcal{T}}^2(V,\mathcal{T})\Big).$$

Corollary 3.5 with $\mathcal{T}_*=\mathcal{T}$, $V=U$, and $V_*=V_\epsilon$ yields

$$\operatorname{osc}_{\mathcal{T}}^2(U,\mathcal{T}) \leq 2\,\operatorname{osc}_{\mathcal{T}}^2(V_\epsilon,\mathcal{T}) + 2\,\Lambda_1\,\operatorname{osc}_{\mathcal{T}_0}^2(\mathbf{D},\mathcal{T}_0)\,\|\!|U-V_\epsilon|\!\|_\Omega^2\,.$$

Since U is the Galerkin solution, $\|\!|u-U|\!\|_\Omega^2 + \|\!|U-V_\epsilon|\!\|_\Omega^2 = \|\!|u-V_\epsilon|\!\|_\Omega^2$, whence

$$\begin{aligned}\|\!|u-U|\!\|_\Omega^2 + \operatorname{osc}_{\mathcal{T}}^2(U,\mathcal{T}) &\leq \big(1+2\,\Lambda_1\,\operatorname{osc}_{\mathcal{T}_0}^2(\mathbf{D},\mathcal{T}_0)\big)\,\|\!|u-V_\epsilon|\!\|_\Omega^2 + 2\,\operatorname{osc}_{\mathcal{T}}^2(V_\epsilon,\mathcal{T})\\ &\leq (1+\epsilon)\,C_{\mathbf{D}}^2 \inf_{V\in\mathbb{V}(\mathcal{T})} \Big(\|\!|u-V|\!\|_\Omega^2 + \operatorname{osc}_{\mathcal{T}}^2(V,\mathcal{T})\Big),\end{aligned}$$

with $C_{\mathbf{D}} = \max\big\{2, 1+2\,\Lambda_1\,\operatorname{osc}_{\mathcal{T}_0}^2(\mathbf{D},\mathcal{T}_0)\big\}$, and the assertion follows from $\epsilon\to 0$. ∎

After collecting the above observations we are ready to define an approximation class $\mathbb{A}_s$ based on the total error. Let $\mathbb{T}_N\subset\mathbb{T}$ be the set of all possible conforming triangulations generated from $\mathcal{T}_0$ with at most N elements more than $\mathcal{T}_0$:

$$\mathbb{T}_N := \{\mathcal{T}\in\mathbb{T} \mid \#\mathcal{T} - \#\mathcal{T}_0 \leq N\}.$$

The quality of the best approximation to the total error in the set $\mathbb{T}_N$ is given by

$$\sigma(N;v,f,\mathbf{D}) := \inf_{\mathcal{T}\in\mathbb{T}_N} \inf_{V\in\mathbb{V}(\mathcal{T})} \Big(\|\!|v-V|\!\|_\Omega^2 + \operatorname{osc}_{\mathcal{T}}^2(V,\mathcal{T})\Big)^{1/2},$$

and for $s > 0$ we define the nonlinear approximation class $\mathbb{A}_s$ to be

$$\mathbb{A}_s := \Big\{(v, f, \mathbf{D}) \mid |v, f, \mathbf{D}|_s := \sup_{N>0} \big(N^s\, \sigma(N; v, f, \mathbf{D})\big) < \infty\Big\}.$$

Thanks to Lemma 5.2, the solution u of (1.1) with data $(\mathbf{D}, f)$ satisfies

$$\sigma(N; u, f, \mathbf{D}) \approx \inf_{\mathcal{T}\in\mathbb{T}_N} \big\{\eta_{\mathcal{T}}(U, \mathcal{T}) : U = \mathsf{SOLVE}(\mathcal{T})\big\}. \tag{5.1}$$

An important pending issue is the characterization of $\mathbb{A}_s$, which is beyond the scope of this paper. Nevertheless, we present a preliminary discussion now. We start by stressing that $\mathbb{A}_s$ is not a typical approximation class for functions, as known from approximation theory, because there is a nonlinear interaction between data $\mathbf{D} = (\mathbf{A}, c)$ and U through $\operatorname{osc}_{\mathcal{T}}(U, \mathcal{T})$. To avoid such an interaction, we first assume that coefficients $\mathbf{A}$ and c are piecewise polynomials of degree at most n and $n-2$ over $\mathcal{T}_0$, respectively. In this case, oscillation reduces to

$$\operatorname{osc}_{\mathcal{T}}(U, \mathcal{T}) = \big\|h(f - \Pi^2_{2n-2} f)\big\|_{L^2(\Omega)} \qquad \forall\, \mathcal{T} \in \mathbb{T},$$

which is independent of U. We have the following characterization of $\mathbb{A}_s$ in terms of standard approximation classes [3, 4, 23]:

$$\mathcal{A}_s := \Big\{v \in \mathbb{V} \mid \sup_{N>0} \big(N^s \inf_{\mathcal{T}\in\mathbb{T}_N} \inf_{V\in\mathbb{V}(\mathcal{T})} |\!|\!| v - V |\!|\!|_\Omega\big) < \infty\Big\},$$
$$\bar{\mathcal{A}}_s := \Big\{g \in L^2(\Omega) \mid \sup_{N>0} \big(N^s \inf_{\mathcal{T}\in\mathbb{T}_N} \big\|h(g - \Pi^2_{2n-2} g)\big\|_{L^2(\Omega)}\big) < \infty\Big\}.$$

LEMMA 5.3 (equivalence of classes). *If data $\mathbf{D} = (\mathbf{A}, c)$ are piecewise polynomials over $\mathcal{T}_0$, as specified above, then $(u, f, \mathbf{D}) \in \mathbb{A}_s$ if and only if $(u, f) \in \mathcal{A}_s \times \bar{\mathcal{A}}_s$.*

Proof. It is obvious that $(u, f, \mathbf{D}) \in \mathbb{A}_s$ implies that $(u, f) \in \mathcal{A}_s \times \bar{\mathcal{A}}_s$. In order to prove the reverse implication, let $(u, f) \in \mathcal{A}_s \times \bar{\mathcal{A}}_s$ and $\mathcal{T}_1, \mathcal{T}_2 \in \mathbb{T}_N$ be meshes so that $|\!|\!| u - U |\!|\!|_\Omega \preccurlyeq N^{-s}$ and $\big\|f - \Pi^2_{2n-2} f\big\|_{L^2(\Omega)} \preccurlyeq N^{-s}$. According to Lemma 3.7, the overlay $\mathcal{T} = \mathcal{T}_1 \oplus \mathcal{T}_2$ is a mesh with $\#\mathcal{T} - \#\mathcal{T}_0 \le 2N$, whence $\mathcal{T} \in \mathbb{T}_{2N}$ and

$$|\!|\!| u - U |\!|\!|_\Omega + \operatorname{osc}_{\mathcal{T}}(U, \mathcal{T}) \preccurlyeq N^{-s}.$$

This implies that $(u, f, \mathbf{D}) \in \mathbb{A}_s$, as asserted. ∎

A similar approximability condition $(u, f) \in \mathcal{A}_s \times \bar{\mathcal{A}}_s$ is imposed in [3, 23] to obtain quasi-optimal cardinality in terms of DOFs. In contrast to our approach, however, the two classes are handled separately (see section 6.2 for separate marking).

We now turn to the nonlinear interaction encoded in $\operatorname{osc}_{\mathcal{T}}(U, \mathcal{T})$. We first show, by using a priori arguments, that $(u, f, \mathbf{D}) \in \mathbb{A}_s$ for some $s > 0$.

LEMMA 5.4 (decay rate of oscillation). *The oscillation term $\operatorname{osc}_{\mathcal{T}}(U, \mathcal{T})$ has a decay rate of order at least $-1/d$, i.e.,*

$$\inf_{\mathcal{T}\in\mathbb{T}_N} \big\{\operatorname{osc}_{\mathcal{T}}(U, \mathcal{T}) \mid U = \mathsf{SOLVE}(\mathcal{T})\big\} \preccurlyeq N^{-1/d}. \tag{5.2}$$

Therefore, $(u, f, \mathbf{D}) \in \mathbb{A}_{r/d}$ for some $r \in (0, 1]$ depending on the regularity of $\mathbf{D}$ and Ω. Furthermore, if $u \in \mathcal{A}_s$ for some $s \in (0, 1/d]$, then $(u, f, \mathbf{D}) \in \mathbb{A}_s$.

Proof. Let $\mathcal{T} \in \mathbb{T}$ be generated by quasi-uniform refinement of $\mathcal{T}_0$ with $\#\mathcal{T} \approx N$. By applying (3.2) with $V = U$ and $W = 0$, we obtain for the oscillation term

$$\operatorname{osc}_{\mathcal{T}}(U, T) \le h_T \|P^2_{2n-2} f\|_T + \bar{\Lambda}_1 \operatorname{osc}_{\mathcal{T}}(\mathbf{D}, T) \|U\|_{H^1(\omega_T)} \preccurlyeq h_T \big(\|f\|_T + \|U\|_{H^1(\omega_T)}\big)$$

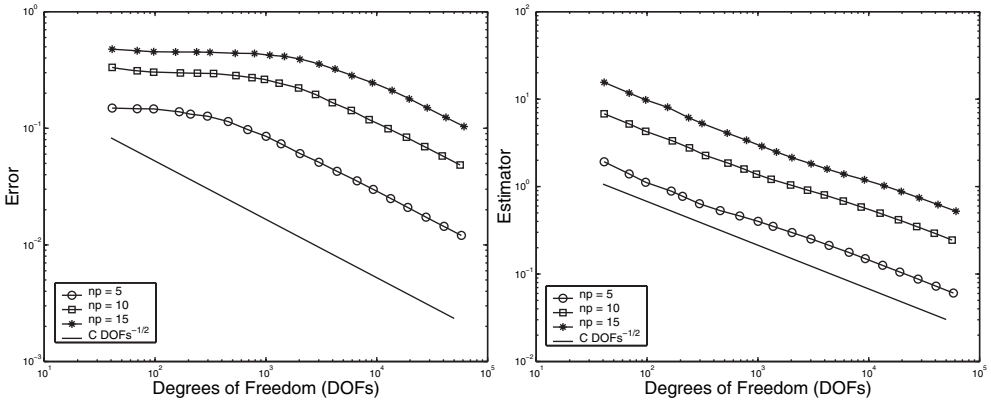


FIG. 5.1. *Decay of the energy error (left) and the estimator (right) for the smooth solution u_S of section 6.1 with frequencies 5, 10, and 15. The energy error exhibits a frequency-dependent plateau in the preasymptotic regime and later an optimal decay. This behavior is allowed by $\mathbb{A}_s$.*

by the definition of oscillation $\operatorname{osc}_{\mathcal{T}}(\mathbf{D}, T)$ and $\|P_n^\infty \mathbf{A}\|_{L^\infty(T)} \leq h_T \|\nabla \mathbf{A}\|_{L^\infty(T)}$. Quasi-uniform refinement yields the relation $h_T \approx N^{-1/d}$ for all $T \in \mathcal{T}$ which in turn implies (5.2). In the same vein, regularity $u \in H^{1+r}(\Omega)$ for some $0 < r \leq 1$ in any dimension d [11, Theorem 3] implies that a $\|\!|u - U|\!\|_\Omega \preccurlyeq N^{-r/d}$ by standard a priori estimates. Combining these estimates yields a decay rate $-r/d$ for the total error.

Finally, let $u \in \mathcal{A}_s$ and $\mathcal{T}_1 \in \mathbb{T}_N$ be such that $\|\!|u - U|\!\|_\Omega \preccurlyeq N^{-s}$. In view of (5.2), we find another grid $\mathcal{T}_2 \in \mathbb{T}_N$ with $\operatorname{osc}_{\mathcal{T}}(U, \mathcal{T}) \preccurlyeq N^{-1/d} \leq N^{-s}$. According to Lemma 3.7, the overlay $\mathcal{T} = \mathcal{T}_1 \oplus \mathcal{T}_2$ is a mesh with $\#\mathcal{T} - \#\mathcal{T}_0 \leq 2N$, i.e., $\mathcal{T} \in \mathbb{T}_{2N}$. The decay rate of total error is N^{-s} over $\mathcal{T}$, whence $(u, f, \mathbf{D}) \in \mathbb{A}_s$. ∎

Remark 5.5 (error notion). Lemma 5.4 states that for $s \leq 1/d$ oscillation exhibits a decay faster than the energy error; so the situation described in Remark 5.1 shows that the class $\mathbb{A}_s$ may fail to describe the behavior of $\|\!|u - U_k|\!\|_\Omega$ in the early stages of adaptivity. In fact, we recall that $\operatorname{osc}_k(U_k, \mathcal{T}_k) = \|h(f - \Pi_0^2 f)\|_\Omega$, the discrete solution $U_k = 0$, and $\|\!|u - U_k|\!\|_\Omega \approx 2^{-K}$ is constant for as many steps $k \leq K$ as desired. In contrast, $\eta_k(U_k, \mathcal{T}_k) = \operatorname{osc}_k(U_k, \mathcal{T}_k) = \|hf\|_\Omega$ reduces strictly for $k \leq K$ but overestimates $\|\!|u - U_k|\!\|_\Omega$. On the other hand, $\operatorname{osc}_k(U_k, \mathcal{T}_k) = 0$, and $\eta_k(U_k, \mathcal{T}_k) \approx \|\!|u - U_k|\!\|_\Omega$ dictates the asymptotics for $k > K$. This is consistent with Lemma 5.4 and $(u, f, \mathbf{D}) \in \mathbb{A}_1$. However, the fact that the preasymptotic regime $k \leq K$ for the energy error could be made arbitrarily long would be problematic if we were to focus exclusively on $\|\!|u - U_k|\!\|_\Omega$.

In practice, this effect is typically less dramatic because f is not orthogonal to $\mathbb{V}(\mathcal{T}_k)$. Figure 5.1 displays the behavior of the AFEM for the smooth solution u_S of section 6.1 with frequencies 5, 10, and 15. We can see that the error exhibits a frequency-dependent plateau in the preasymptotic regime and later an optimal decay. In contrast, the estimator decays always with the optimal rate. Since all decisions of the AFEM are based on the estimator, this behavior has to be expected and is consistent with our notion of the approximation class $\mathbb{A}_s$, which can be characterized just by the estimator; compare (5.1).

According to Lemma 5.4, the new approximation class $\mathbb{A}_s$ is completely determined by the standard approximation class $\mathcal{A}_s$ for $s \leq 1/d$. Furthermore, in practice oscillation is often much smaller than the true error, whence the asymptotic rate N^{-s} would be determined by $u \in \mathcal{A}_s$ for any $s > 0$. We point out the upper bound $s \leq n/d$ for polynomial degree $n \geq 1$ in dimension d, which can be easily seen for

uniform refinement; thus $s = 1/d$ is the highest attainable order with piecewise linears $n = 1$. A near characterization of $\mathcal{A}_s$ for $d = 2$ in terms of Besov spaces can be found in [3, 4]: $u \in \mathcal{A}_s$ implies that $u \in B_p^{2s+1}(L^p(\Omega))$ for $p = \frac{2}{2s+1}$ [3, Theorem 9.3]; $u \in B_p^{2s+1}(L^p(\Omega))$ for $p > \frac{2}{2s+1}$ implies that $u \in \mathcal{A}_s$ [3, Theorem 9.1]. Note that $p < 1$ for $s > 1/2$.

To motivate the case $n = 1$, $d = 2$, $s = 1/2$, we consider the following heuristic argument. Assume that a mesh $\mathcal{T} \in \mathbb{T}$ equidistributes the quantity $\Lambda = \|D^2 u\|_{L^1(T)}$. Polynomial interpolation theory then yields

$$\|\nabla(u - I_T u)\|_T \preccurlyeq \|D^2 u\|_{L^1(\omega_T)} \approx \Lambda \qquad \forall T \in \mathcal{T},$$

and thus

$$\|\nabla(u - I_{\mathcal{T}} u)\|_\Omega^2 = \sum_{T \in \mathcal{T}} \|\nabla(u - I_T u)\|_T^2 \preccurlyeq \Lambda^2 \#\mathcal{T}.$$

Since $\|D^2 u\|_{L^1(\Omega)} \approx \Lambda \#\mathcal{T}$, we deduce the error decay

$$\|\nabla(u - I_{\mathcal{T}} u)\|_\Omega \preccurlyeq \|D^2 u\|_{L^1(\Omega)} (\#\mathcal{T})^{-1/2}.$$

What this argument does not show is the existence of such a mesh $\mathcal{T} \in \mathbb{T}$ created by bisection from $\mathcal{T}_0$. This explains the need for a strict inequality in the characterization above. However, we still have the following optimal result.

LEMMA 5.6 (membership in $\mathbb{A}_{1/2}$). *Let $d = 2$, $n = 1$, and u be the solution of (1.1). Let $\Omega, \mathbf{A}, c$, and f satisfy the assumptions in section 2.1, and let $\mathbf{A}$ be globally Lipschitz. Then $(u, \mathbf{D}, f) \in \mathbb{A}_{1/2}$.*

Proof. By applying [10, Theorem 5.2.2], we deduce that $u \in W_p^2(\Omega)$ for some $p > 1$. The assertion follows upon combining [3, Theorem 9.1] for $s = 1/2$ with Lemma 5.4. ∎

A complete characterization of $\mathbb{A}_s$ for general d and n is still missing. It is important to realize that the nonlinear interaction between data $\mathbf{D}$ and U must be accounted for, thereby leading to a new concept of approximation class $\mathbb{A}_s$, which generalizes those in [3, 4, 23].

Remark 5.7 (conforming vs. nonconforming meshes). In contrast to [3, 23], our approach relies exclusively on conforming triangulations. When $\mathbf{D} = (\mathbf{I}, 0)$ and $n = 1$, the approximation class $\mathbb{A}_s$ is the same regardless of conformity [3]. The situation is quite different, however, in dealing with oscillation of a jump residual unless the depth of the nonconforming refinement is restricted beforehand.

Assume that $(u, f, \mathbf{D}) \in \mathbb{A}_s$ for some $s > 0$, where typically $s \le n/d$. We now prove that the approximation U_k generated by the AFEM converges to u with the same rate $(\#\mathcal{T}_k - \#\mathcal{T}_0)^{-s}$ as the best approximation described by $\mathbb{A}_s$ up to a multiplicative constant. We need to count elements marked by the estimator (the cardinality of $\mathcal{M}_k$) as well as those added to keep mesh conformity (see Lemma 2.3). To this end, we impose more stringent requirements than for convergence of the AFEM.

ASSUMPTION 5.8 (optimality). *We assume the following properties of the AFEM:*

(a) *The marking parameter θ satisfies $\theta \in (0, \theta_*)$, with*

$$\theta_*^2 = \frac{C_2}{1 + C_1(1 + 2\Lambda_1 \operatorname{osc}_{\mathcal{T}_0}^2(\mathbf{D}, \mathcal{T}_0))}.$$

(b) *Procedure* MARK *selects a set $\mathcal{M}_k$ of marked elements with minimal cardinality.*

(c) *The distribution of refinement edges on* $\mathcal{T}_0$ *satisfies condition (b) of section* 4 *in* [24].

The limit value θ_* depends on the ratio $\sqrt{C_2/C_1} \leq 1$, which quantifies the quality of estimator $\eta_{\mathcal{T}_k}(U_k, \mathcal{T}_k)$, as well as the oscillation $\operatorname{osc}_{\mathcal{T}_0}(\mathbf{D}, \mathcal{T}_0)$ of coefficients of the PDE on $\mathcal{T}_0$.

5.2. Cardinality of $\mathcal{M}_k$. The following lemma establishes a link between nonlinear approximation theory and the AFEM through the Dörfler marking strategy. Roughly speaking we prove that, if an approximation satisfies a suitable total error reduction from $\mathcal{T}$ to $\mathcal{T}_* \geq \mathcal{T}$, the error indicators of the coarser solutions must satisfy a Dörfler property on the set $\mathcal{R}_{\mathcal{T}\to\mathcal{T}_*}$. In other words, Dörfler marking and total error reduction are intimately connected.

LEMMA 5.9 (optimal marking). *Assume that the marking parameter* θ *verifies* (a) *of Assumption* 5.8. *Let* $\mathcal{T} \in \mathbb{T}$ *and* $U \in \mathbb{V}(\mathcal{T})$ *be the discrete solution of* (2.3). *Set* $\mu := \frac{1}{2}(1 - \frac{\theta^2}{\theta_*^2}) > 0$, *and let* $\mathcal{T}_* \in \mathbb{T}$ *be any refinement of* $\mathcal{T}$, *i.e.,* $\mathcal{T} \leq \mathcal{T}_*$, *such that the discrete solution* $U_* \in \mathbb{V}(\mathcal{T}_*)$ *satisfies*

$$|||u - U_*|||_\Omega^2 + \operatorname{osc}_{\mathcal{T}_*}^2(U_*, \mathcal{T}_*) \leq \mu \left\{ |||u - U|||_\Omega^2 + \operatorname{osc}_{\mathcal{T}}^2(U, \mathcal{T}) \right\}. \tag{5.3}$$

Then the set $\mathcal{R} := \mathcal{R}_{\mathcal{T}\to\mathcal{T}_*}$ *satisfies the Dörfler property*

$$\eta_{\mathcal{T}}(U, \mathcal{R}) \geq \theta\, \eta_{\mathcal{T}}(U, \mathcal{T}). \tag{5.4}$$

Proof. We first combine the lower bound (2.7) with (5.3) to obtain

$$\begin{aligned} (1 - 2\mu)\, C_2\, \eta_{\mathcal{T}}^2(U, \mathcal{T}) &\leq (1 - 2\mu) \left(|||u - U|||_\Omega^2 + \operatorname{osc}_{\mathcal{T}}^2(U, \mathcal{T}) \right) \\ &\leq |||u - U|||_\Omega^2 - |||u - U_*|||_\Omega^2 + \operatorname{osc}_{\mathcal{T}}^2(U, \mathcal{T}) - 2 \operatorname{osc}_{\mathcal{T}_*}^2(U_*, \mathcal{T}_*). \end{aligned} \tag{5.5}$$

We estimate separately the error and oscillation terms. We invoke the orthogonality (2.4) and the localized upper bound of Lemma 3.6 to arrive at

$$|||u - U|||_\Omega^2 - |||u - U_*|||_\Omega^2 = |||U_* - U|||_\Omega^2 \leq C_1\, \eta_{\mathcal{T}}^2(U, \mathcal{R}). \tag{5.6}$$

For the oscillation terms we argue according to whether an element $T \in \mathcal{T}$ belongs to the set of refined elements $\mathcal{R}$ or not. For $T \in \mathcal{R}$ we use the dominance

$$\operatorname{osc}_{\mathcal{T}}^2(U, T) \leq \eta_{\mathcal{T}}^2(U, T)$$

of Remark 2.1. For $T \in \mathcal{T} \cap \mathcal{T}_*$, Corollary 3.5 with $V = U$ and $V_* = U_*$ yields

$$\operatorname{osc}_{\mathcal{T}}^2(U, \mathcal{T} \cap \mathcal{T}_*) - 2 \operatorname{osc}_{\mathcal{T}_*}^2(U_*, \mathcal{T} \cap \mathcal{T}_*) \leq 2\Lambda_1 \operatorname{osc}_{\mathcal{T}_0}^2(\mathbf{D}, \mathcal{T}_0)\, |||U_* - U|||_\Omega^2.$$

By combining these two estimates with (5.6) we infer that

$$\operatorname{osc}_{\mathcal{T}}^2(U, \mathcal{T}) - 2 \operatorname{osc}_{\mathcal{T}_*}^2(U_*, \mathcal{T}_*) \leq (1 + 2\, C_1 \Lambda_1\, \operatorname{osc}_{\mathcal{T}_0}^2(\mathbf{D}, \mathcal{T}_0))\, \eta_{\mathcal{T}}^2(U, \mathcal{R}). \tag{5.7}$$

By collecting (5.5), (5.6), and (5.7), we finally deduce that

$$\eta_{\mathcal{T}}^2(U, \mathcal{R}) \geq \frac{(1 - 2\mu)\, C_2}{1 + C_1(1 + 2\,\Lambda_1 \operatorname{osc}_{\mathcal{T}_0}^2(\mathbf{D}, \mathcal{T}_0))}\, \eta_{\mathcal{T}}^2(U, \mathcal{T}) = \theta^2\, \eta_{\mathcal{T}}^2(U, \mathcal{T})$$

in light of the definitions of θ_*, θ, and μ. This concludes the proof. ∎

The key to relate the best mesh with AFEM triangulations is the fact that procedure MARK selects the marked set $\mathcal{M}_k$ with *minimal* cardinality. This crucial idea, due to Stevenson [23], is used next.

LEMMA 5.10 (cardinality of $\mathcal{M}_k$). *Assume that the marking parameter θ verifies* (a) *and procedure MARK satisfies* (b) *of Assumption* 5.8. *Let u be the solution of* (1.1), *and let $\{\mathcal{T}_k, \mathbb{V}_k, U_k\}_{k\geq 0}$ be the sequence of meshes, finite element spaces, and discrete solutions produced by the AFEM.*

If $(u, f, \mathbf{D}) \in \mathbb{A}_s$, then the following estimate is valid:

$$\#\mathcal{M}_k \preccurlyeq \left(1 - \frac{\theta^2}{\theta_*^2}\right)^{-\frac{1}{2s}} |u, f, \mathbf{D}|_s^{\frac{1}{s}} C_{\mathbf{D}}^{\frac{1}{2s}} \left\{ |\!|\!| u - U_k |\!|\!|_\Omega^2 + \operatorname{osc}_{\mathcal{T}_k}^2(U_k, \mathcal{T}_k) \right\}^{-\frac{1}{2s}}.$$

Proof. We set $\epsilon^2 := \mu\, C_{\mathbf{D}}^{-1}(|\!|\!| u - U_k |\!|\!|_\Omega^2 + \operatorname{osc}_{\mathcal{T}_k}^2(U_k, \mathcal{T}_k))$, where $\mu = \frac{1}{2}(1 - \frac{\theta^2}{\theta_*^2}) > 0$ and $C_{\mathbf{D}}$ is the constant in Lemma 5.2. Since $(u, f, \mathbf{D}) \in \mathbb{A}_s$, there exists a $\mathcal{T}_\epsilon \in \mathbb{T}$ and a $V_\epsilon \in \mathbb{V}(\mathcal{T}_\epsilon)$ such that

$$\#\mathcal{T}_\epsilon - \#\mathcal{T}_0 \preccurlyeq |u, f, \mathbf{D}|_s^{1/s} \epsilon^{-1/s}, \tag{5.8}$$

$$|\!|\!| u - V_\epsilon |\!|\!|_\Omega^2 + \operatorname{osc}_{\mathcal{T}_\epsilon}^2(V_\epsilon, \mathcal{T}_\epsilon) \leq \epsilon^2. \tag{5.9}$$

Let $\mathcal{T}_* := \mathcal{T}_\epsilon \oplus \mathcal{T}_k$ be the overlay of $\mathcal{T}_\epsilon$ and $\mathcal{T}_k$, and let $U_* \in \mathbb{V}(\mathcal{T}_*)$ be the discrete solution of (2.3) on $\mathcal{T}_*$. To show that there is a reduction by a factor μ of the total error between U_* and U_k, since $\mathcal{T}_* \geq \mathcal{T}_\epsilon$, we argue as in Lemma 5.2 and obtain

$$\begin{aligned} |\!|\!| u - U_* |\!|\!|_\Omega^2 + \operatorname{osc}_{\mathcal{T}_*}^2(U_*, \mathcal{T}_*) &\leq C_{\mathbf{D}} \left\{ |\!|\!| u - V_\epsilon |\!|\!|_\Omega^2 + \operatorname{osc}_{\mathcal{T}_\epsilon}^2(V_\epsilon, \mathcal{T}_\epsilon) \right\} \\ &\leq C_{\mathbf{D}}\, \epsilon^2 = \mu \left\{ |\!|\!| u - U_k |\!|\!|_\Omega^2 + \operatorname{osc}_k^2(U_k, \mathcal{T}_k) \right\}. \end{aligned}$$

Hence, we deduce from Lemma 5.9 that the subset $\mathcal{R} := \mathcal{R}_{\mathcal{T}_k \to \mathcal{T}_*} \subset \mathcal{T}_k$ verifies the Dörfler property (5.4) for $\theta < \theta_*$. The fact that procedure MARK selects a subset $\mathcal{M}_k \subset \mathcal{T}_k$ with minimal cardinality satisfying the same property translates into

$$\#\mathcal{M}_k \leq \#\mathcal{R} \leq \#\mathcal{T}_* - \#\mathcal{T}_k \leq \#\mathcal{T}_\epsilon - \#\mathcal{T}_0, \tag{5.10}$$

where Lemma 3.7 has been employed in the last step. Finally, by combining (5.10), (5.8), and the definition of ϵ, we end up with

$$\#\mathcal{M}_k \leq \#\mathcal{T}_\epsilon - \#\mathcal{T}_0 \preccurlyeq \mu^{-\frac{1}{2s}} |u, f, \mathbf{D}|_s^{\frac{1}{s}} C_{\mathbf{D}}^{\frac{1}{2s}} \left\{ |\!|\!| u - U_k |\!|\!|_\Omega^2 + \operatorname{osc}_k^2(U_k, \mathcal{T}_k) \right\}^{-\frac{1}{2s}},$$

which is the asserted estimate. ∎

5.3. Quasi optimality. As a consequence of the previous estimates and the fact that the AFEM is a contraction for the quasi error, we obtain quasi optimality of the total error.

THEOREM 5.11 (quasi optimality). *Let Assumption* 5.8 *be satisfied by the AFEM. Let u be the solution of* (1.1), *and let $\{\mathcal{T}_k, \mathbb{V}_k, U_k\}_{k\geq 0}$ be the sequence of meshes, finite element spaces, and discrete solutions produced by the AFEM.*

Let $(u, f, \mathbf{D}) \in \mathbb{A}_s$ and $\Theta(\theta, s) := \theta^{-4s}(1 - \frac{\theta^2}{\theta_*^2})^{-1/2}$ describe the asymptotics of the AFEM as $\theta \to \theta_*, 0$ or $s \to 0$. Then there exists a constant C, depending on data, the refinement depth b, and $\mathcal{T}_0$, but independent of s, such that

$$\left\{ |\!|\!| u - U_k |\!|\!|_\Omega^2 + \operatorname{osc}_k^2(U_k, \mathcal{T}_k) \right\}^{1/2} \leq C\, \Theta(\theta, s) |u, f, \mathbf{D}|_s \big(\#\mathcal{T}_k - \#\mathcal{T}_0\big)^{-s}.$$

Proof. By combining Lemmas 2.3 and 5.10 we deduce that

$$\#\mathcal{T}_k - \#\mathcal{T}_0 \preccurlyeq \sum_{j=0}^{k-1} \#\mathcal{M}_j \preccurlyeq M \sum_{j=0}^{k-1} \Big\{ |\!|\!| u - U_j |\!|\!|_\Omega^2 + \operatorname{osc}_j^2(U_j, \mathcal{T}_j) \Big\}^{-\frac{1}{2s}}, \tag{5.11}$$

with $M := (1 - \frac{\theta^2}{\theta_*^2})^{-\frac{1}{2s}} |u, f, \mathbf{D}|_s^{\frac{1}{s}} C_{\mathbf{D}}^{\frac{1}{2s}}$. We infer from the lower bound (2.7) that

$$\begin{aligned} |\!|\!| u - U_j |\!|\!|_\Omega^2 + \gamma \ \operatorname{osc}_j^2(U_j, \mathcal{T}_j) &\le |\!|\!| u - U_j |\!|\!|_\Omega^2 + \gamma \, \eta_j^2(U_j, \mathcal{T}_j) \\ &\le \left(1 + \frac{\gamma}{C_2}\right) \Big\{ |\!|\!| u - U_j |\!|\!|_\Omega^2 + \operatorname{osc}_j^2(U_j, \mathcal{T}_j) \Big\}. \end{aligned} \tag{5.12}$$

On the other hand, the linear rate $\alpha = \alpha(\theta) < 1$ of convergence of Theorem 4.1 for the sum of the energy error and the scaled error estimator implies that afor $0 \le j \le k-1$

$$|\!|\!| u - U_k |\!|\!|_\Omega^2 + \gamma \, \eta_k^2(U_k, \mathcal{T}_k) \le \alpha^{2(k-j)} \Big\{ |\!|\!| u - U_j |\!|\!|_\Omega^2 + \gamma \, \eta_j^2(U_j, \mathcal{T}_j) \Big\}. \tag{5.13}$$

We combine (5.11), (5.12), and (5.13) to obtain

$$\#\mathcal{T}_k - \#\mathcal{T}_0 \preccurlyeq M \left(1 + \frac{\gamma}{C_2}\right)^{\frac{1}{2s}} \Big\{ |\!|\!| u - U_k |\!|\!|_\Omega^2 + \gamma \, \eta_k^2(U_k, \mathcal{T}_k) \Big\}^{-\frac{1}{2s}} \sum_{j=1}^{k} \alpha^{\frac{j}{s}}.$$

Since $\alpha < 1$, the geometric series is bounded by the constant $S_\theta = \alpha^{1/s}(1 - \alpha^{1/s})^{-1}$. By recalling that the element residual dominates the oscillation, we end up with

$$\#\mathcal{T}_k - \#\mathcal{T}_0 \preccurlyeq S_\theta M \left(1 + \frac{\gamma}{C_2}\right)^{\frac{1}{2s}} \Big\{ |\!|\!| u - U_k |\!|\!|_\Omega^2 + \gamma \operatorname{osc}_k^2(U_k, \mathcal{T}_k) \Big\}^{-\frac{1}{2s}}. \tag{5.14}$$

To examine the asymptotics as $\theta, s \to 0$, hidden in S_θ, we use Remark 4.4 to get $\alpha \le (1-t)^{\frac{1}{2}}$, with $t = c\theta^4$, and observe that $\lim_{t \to 0} \frac{t}{1 - (1-t)^{\frac{1}{2s}}} = 2s$. Therefore

$$S_\theta^s \le \frac{\alpha}{t^s} \left(\frac{t}{1 - (1-t)^{1/2s}} \right)^s \approx \alpha s^s \theta^{-4s},$$

whence $S_\theta^s M^s \approx \Theta(\theta, s) |u, f, \mathbf{D}|_s \, C_{\mathbf{D}}^{1/2}$. By raising (5.14) to the sth power and reordering, we obtain the desired estimate for the total error with $C > 0$ independent of s. ∎

The same decay rate is obtained for the error estimator due to its equivalence to the total error.

COROLLARY 5.12. *Under the same requirements as in Theorem* 5.11*, there exists a constant C such that*

$$\eta_k \le C \, \Theta(\theta, s) |u, f, \mathbf{D}|_s \big(\#\mathcal{T}_k - \#\mathcal{T}_0 \big)^{-s}.$$

The constant C may differ from the one in Theorem 5.11.

6. The critical role of marking for optimality. We conclude the article by revisiting former optimality proofs by Binev, Dahmen, and DeVore [3] and Stevenson [23], which are based on the AFEM designed by Morin, Nochetto, and Siebert [17, 18], from now on called MNS. This is an adaptive loop of the form (1.2), which marks

separately for both the estimator and the oscillation in each iteration, which in turn yields a contraction for the energy error. It turns out that separate marking appears to be problematic for proving optimality, and thus variants of MNS are studied instead [3, 23]. The key issue is the delicate choice of marking parameters for the estimator and oscillation. To shed light on this intrinsic difficulty we first present some numerical experiments, then discuss the effect of separate marking, and finally comment on marking for an optimal AFEM.

6.1. Separate marking and MNS. The procedure ESTIMATE of MNS calculates both the error and oscillation indicators $\{\eta_k(U_k,T),\operatorname{osc}_k(U_k,T)\}_{T\in\mathcal{T}_k}$, and the procedure MARK uses Dörfler marking for both the estimator and oscillation. More precisely, the routine MARK of MNS is of the form: *given parameters* $0<\theta_{\rm est},\theta_{\rm osc}<1$,

(6.1a) *mark any subset* $\mathcal{M}_k\subset\mathcal{T}_k$ *such that* $\eta_k(U_k,\mathcal{M}_k)\geq\theta_{\rm est}\eta_k(U_k,\mathcal{T}_k)$;

(6.1b) *if necessary enlarge* $\mathcal{M}_k$ *to satisfy* $\operatorname{osc}_k(U_k,\mathcal{M}_k)\geq\theta_{\rm osc}\operatorname{osc}_k(U_k,\mathcal{T}_k)$.

Since oscillation is generically of higher order than the estimator, the issue at stake is whether elements added by oscillation, even though immaterial relative to the error, could ruin the optimality of MNS observed in experiments [17, 18, 16]. If $\eta_k(U_k,\mathcal{T}_k)$ has large indicators in a small area, then Dörfler marking for the estimator (6.1a) could select a set $\mathcal{M}_k$ with a small number of elements relative to $\mathcal{T}_k$. However, if $\operatorname{osc}_k(U_k,\mathcal{T}_k)$ were globally distributed in $\mathcal{T}_k$, then MNS would require additional marking of a large percentage of all elements to satisfy (6.1b); i.e., $\#\mathcal{M}_k$ could be large relative to $\#\mathcal{T}_k$.

To explore this idea computationally, we consider a simple modification of the two-dimensional Example 5.3 in [12, 17]

$$-\operatorname{div}\mathbf{A}\nabla u=f\qquad\text{in }\Omega,\qquad u=g\qquad\text{on }\Omega,$$

where $\Omega=(-1,1)^2$ and $\mathbf{A}:=a_i\mathbf{I}$ is a piecewise constant (checkerboard pattern), with $a_1=a_3$ in the first and third quadrants and $a_2=a_4$ in the second and fourth quadrants. The exact solution u is given as a sum: $u=u_K+u_S$. Function $u_K\in H^1(\Omega)$ is the weak solution of $\operatorname{div}(\mathbf{A}\nabla u_K)=0$ for parameters [12, 17, 18]

$$a_1=a_3=161.4476387975881,\qquad a_2=a_4=1;$$

the singularity $u_K\approx r^{0.1}$ at the origin is so extreme that it typically leads to marking of a handful of elements per step to satisfy (6.1a) (see [17, 18] for details). Function u_S is given in each quadrant by

$$u_S(x,y)=10^{-2}a_i^{-1}(x^2+y^2)\sin^2(4\pi x)\sin^2(4\pi y),\qquad 1\leq i\leq 4,$$

is smooth, and is of comparable magnitude with u_K, while the corresponding $f=-\operatorname{div}\mathbf{A}\nabla u_S$ exhibits an increasing amount of data oscillation away from the origin. Procedure MARK takes the usual value of $\theta_{\rm est}=0.5$ [8, 17, 18, 21], and procedure REFINE subdivides all elements in $\mathcal{M}_k$ by using two bisections.

The behavior of MNS for several values of $\theta_{\rm osc}$ is depicted in Figure 6.1. We can visualize the sensitivity of MNS with respect to parameter $\theta_{\rm osc}$. For values of $\theta_{\rm osc}\leq 0.4$ the rate of convergence appears to be quasi-optimal. However, beyond this threshold the curves for both the error and the estimator flatten out, thereby showing lack of optimality. The threshold value $\theta_{\rm osc}=0.4$, even though consistent with the practice of MNS, is tricky to find in general since it is problem-dependent.

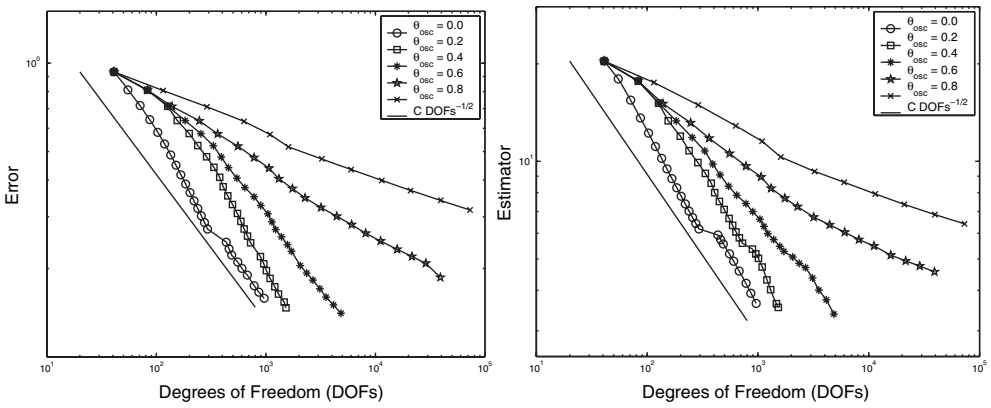


FIG. 6.1. *Decay of the error (left) and the estimator (right) vs. DOFs for* $\theta_{\mathrm{est}} = 0.5$ *and values* $\theta_{\mathrm{osc}} = 0.0, 0.2, 0.4, 0.6,$ *and* 0.8. *For values of* $\theta_{\mathrm{osc}} \leq 0.4$ *the rate of convergence is quasi-optimal, but for* $\theta_{\mathrm{osc}} > 0.4$ *the curves flatten out, thereby showing lack of optimality.*

Lemma 6.1 below provides theoretical insight on this matter. Altogether this indicates that an optimality proof for MNS is out of reach without restrictions on θ_{osc} and makes marking by oscillation (6.1b) questionable.

6.2. Analysis of separate marking. In this section we focus on the adaptive approximation of two given functions in an idealized scenario. We show that separate marking, similar to (6.1), may lead to suboptimal meshes in general. However, a suitable choice of marking parameters may restore optimality. The experiments of section 6.1 confirm this theoretical insight in the nonidealized case of MNS.

For the discussion, we assume that we have two functions u_i, $i = 1, 2$, and have access to their local approximation error

$$e_{\mathcal{T}}(u_i; T) = |u_i - I_{\mathcal{T}} u_i|_{i;T} \qquad \forall T \in \mathcal{T}$$

and global error $e^2_{\mathcal{T}}(u_i) = \sum_{T \in \mathcal{T}} e^2_{\mathcal{T}}(u_i; T)$; hereafter $|\cdot|_i$ are unspecified norms, and $I_{\mathcal{T}}$ is a local interpolation operator over $\mathcal{T} \in \mathbb{T}$. We define the *total error* to be

$$e^2_{\mathcal{T}} := e^2_{\mathcal{T}}(u_1) + e^2_{\mathcal{T}}(u_2)$$

and are interested in its asymptotic decay. If $\mathcal{T} = \mathcal{T}_k$, then we denote $e_k = e_{\mathcal{T}_k}$.

To explore the use of (6.1), we examine the effect of separate marking for $e_k(u_i)$ on a sequence of meshes $\mathcal{T}^i_k$ for $i = 1, 2$. We put ourselves in an idealized, but plausible, situation governed by the following three simplifying assumptions:

(6.2a) *Independence:* $\mathcal{T}^1_k$ *and* $\mathcal{T}^2_k$ *are generated from* $\mathcal{T}_0$ *and are independent of each other;*

(6.2b) *Marking: Separate Dörfler marking with parameters* $\theta_i \in (0, 1)$ *implies that* $e_k(u_i) \approx \alpha^k_i$ *on* $\mathcal{T}^i_k$, *with* $\alpha_i \in (0, 1)$;

(6.2c) *Approximability:* $e_k(u_i) \approx (\#\mathcal{T}^i_k - \#\mathcal{T}_0)^{-s_i}$, *with* $s_1 \leq s_2$ *maximal.*

We are interested in the decay of the total error e_k on the overlay $\mathcal{T}_k := \mathcal{T}^1_k \oplus \mathcal{T}^2_k$. This scenario is a simplification of the more realistic approximation of u_1 and u_2 with separate Dörfler marking on the same sequence of grids $\mathcal{T}_k$ but avoids the complicated interaction of the two marking procedures.

LEMMA 6.1 (separate marking). *Let assumptions* (6.2) *be satisfied.*

Then the decay of the total error e_k on the overlay $\mathcal{T}_k = \mathcal{T}_k^1 \oplus \mathcal{T}_k^2$ for separate marking is always suboptimal except when α_1 and α_2 satisfy

$$\alpha_2 \le \alpha_1 \le \alpha_2^{s_1/s_2}.$$

Proof. Assumption (6.2b) on the average reduction rate implies for the total error that

$$e_k \approx e_k(u_1) + e_k(u_2) \approx \max\{e_k(u_1), e_k(u_2)\} \approx \max\{\alpha_1^k, \alpha_2^k\}. \tag{6.3}$$

Combining (6.2b) and (6.2c) yields $\alpha_i^k \approx (\#\mathcal{T}_k^i - \#\mathcal{T}_0)^{-s_i}$, whence

$$\#\mathcal{T}_k^1 - \#\mathcal{T}_0 \approx \alpha_1^{-k/s_1} = \beta^k \alpha_2^{-k/s_2} \approx \beta^k (\#\mathcal{T}_k^2 - \#\mathcal{T}_0), \tag{6.4}$$

with $\beta = \alpha_1^{-1/s_1}\alpha_2^{1/s_2}$. In view of (2.2), this gives for the overlay $\mathcal{T}_k = \mathcal{T}_k^1 \oplus \mathcal{T}_k^2$

$$\#\mathcal{T}_k - \#\mathcal{T}_0 \approx \begin{cases} \#\mathcal{T}_k^1 - \#\mathcal{T}_0, & \beta \ge 1, \\ \#\mathcal{T}_k^2 - \#\mathcal{T}_0, & \beta < 1. \end{cases} \tag{6.5}$$

The optimal decay of total error e_k corresponds to $e_k \approx (\#\mathcal{T}_k - \#\mathcal{T}_0)^{-s_1}$ because $s_1 \le s_2$. In analyzing the relation of e_k to the number of elements $\#\mathcal{T}_k$ in the overlay $\mathcal{T}_k$, we distinguish three cases and employ (6.3), (6.4), and (6.5).

Case 1: $\alpha_1 < \alpha_2$. We note that $\alpha_1 < \alpha_2$ and $s_1 \le s_2$ yields $\beta \ge 1$. We thus deduce that

$$\begin{aligned} e_k &\approx \max\{\alpha_1^k, \alpha_2^k\} = \alpha_2^k = (\alpha_2/\alpha_1)^k \, \alpha_1^k \\ &\approx (\alpha_2/\alpha_1)^k \, (\#\mathcal{T}_k^1 - \#\mathcal{T}_0)^{-s_1} \approx (\alpha_2/\alpha_1)^k \, (\#\mathcal{T}_k - \#\mathcal{T}_0)^{-s_1}. \end{aligned}$$

Since $\alpha_2/\alpha_1 > 1$, the approximation of e_k on $\mathcal{T}_k$ is suboptimal.

Case 2: $\alpha_1 \ge \alpha_2$ and $\beta < 1$. We obtain

$$\begin{aligned} e_k &\approx \max\{\alpha_1^k, \alpha_2^k\} = \alpha_1^k \approx (\#\mathcal{T}_k^1 - \#\mathcal{T}_0)^{-s_1} \\ &\approx \beta^{-ks_1} (\#\mathcal{T}_k^2 - \#\mathcal{T}_0)^{-s_1} \approx \beta^{-ks_1} (\#\mathcal{T}_k - \#\mathcal{T}_0)^{-s_1}, \end{aligned}$$

whence the approximation of the total error on $\mathcal{T}_k$ is again suboptimal.

Case 3: $\alpha_1 \ge \alpha_2$ and $\beta \ge 1$. We infer that

$$e_k \approx \max\{\alpha_1^k, \alpha_2^k\} = \alpha_1^k \approx (\#\mathcal{T}_k^1 - \#\mathcal{T}_0)^{-s_1} \approx (\#\mathcal{T}_k - \#\mathcal{T}_0)^{-s_1}$$

and that $\mathcal{T}_k$ exhibits optimal cardinality. This exceptional case corresponds to the assertion and concludes the proof. ∎

We learn from Lemma 6.1 that separate marking requires a critical choice of parameters θ_i to obtain optimal cardinality of grids with respect to the total error e_k. By revisiting MNS in light of Lemma 6.1, we could identify the estimator η_k with the error $e_k(u_1)$ and oscillation osc_k with the error $e_k(u_2)$. We observe that $\mathrm{osc}_k \le \eta_k$ combined with (6.2b) implies that $\alpha_2 \le \alpha_1$ and that osc_k is generically of higher order than η_k, thereby yielding $s_1 < s_2$.

We wonder whether or not the optimality condition $\alpha_1 \le \alpha_2^{s_1/s_2}$ is valid. Note that $\alpha_2^{s_1/s_2}$ increases as the gap between s_1 and s_2 increases. Since the oscillation reduction estimate of [17] reveals that α_2 increases as θ_{osc} decreases, we see that

separate marking may be optimal for a wide range of marking parameters $\theta_{\mathrm{est}}, \theta_{\mathrm{osc}}$; this is confirmed by the numerical experiments in section 6.1 even though it is unclear whether η_k and osc_k satisfy (6.2). However, choosing marking parameters $\theta_{\mathrm{est}}, \theta_{\mathrm{osc}}$ is rather tricky in practice because neither the explicit dependence of average reduction rates α_1, α_2 on $\theta_{\mathrm{est}}, \theta_{\mathrm{osc}}$ nor the optimal exponents s_1, s_2 are known. In contrast to MNS, the standard AFEM of section 2.7 marks solely according to the estimator $\eta_{\mathcal{T}}(U, \mathcal{T})$ and thus avoids separate marking.

We conclude by reviewing how an alternative optimal AFEM for the Poisson problem compensates for the lack of optimality of separate marking. Binev, Dahmen, and DeVore [3] added a coarsening step to MNS to prove optimality. Veeser [26] was the first to mark oscillation relative to the error estimator to prove convergence of the AFEM for the p-Laplacian. More recently, Stevenson [23] resorted to a similar marking to prove optimality of the AFEM at the expense of an additional inner loop to reduce data oscillation. Since for the more general elliptic PDE (1.1) oscillation $\mathrm{osc}_{\mathcal{T}}(U, \mathcal{T})$ interacts nonlinearly with the discrete solution U, it is not clear how to follow the approaches in [3, 23] with separate marking.

Acknowledgments. We thank the editor and the anonymous referees for their constructive comments that helped to improve the presentation.

REFERENCES

[1] M. Ainsworth and J. T. Oden, *A Posteriori Error Estimation in Finite Element Analysis*, John Wiley & Sons, New York, 2000.
[2] E. Bänsch, *Local mesh refinement in* 2 *and* 3 *dimensions*, IMPACT Comput. Sci. Engrg., 3 (1991), pp. 181–191.
[3] P. Binev, W. Dahmen, and R. DeVore, *Adaptive finite element methods with convergence rate*, Numer. Math., 97 (2004), pp. 219–268.
[4] P. Binev, W. Dahmen, R. DeVore, and P. Petrushev, *Approximation classes for adaptive methods*, Serdica Math. J., 28 (2002), pp. 1001–1026.
[5] J. M. Cascón, C. Kreuzer, R. H. Nochetto, and K. G. Siebert, *Quasi-optimal AFEM for General Elliptic Operators*, manuscript.
[6] Z. Chen and J. Feng, *An adaptive finite element algorithm with reliable and efficient error control for linear parabolic problems*, Math. Comp., 73 (2004), pp. 1167–1193.
[7] L. Diening and C. Kreuzer, *Linear convergence of an adaptive finite element method for the p-Laplacian equation*, SIAM J. Numer. Anal., 46 (2008), pp. 614–638.
[8] W. Dörfler, *A convergent adaptive algorithm for Poisson's equation*, SIAM J. Numer. Anal., 33 (1996), pp. 1106–1124.
[9] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin, 1983.
[10] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Pitman, Boston, 1985.
[11] F. Jochmann, *An H^s regularity result for the gradient of solutions to elliptic equations with mixed boundary conditions*, J. Math. Anal. Appl., 238 (1999), pp. 429–450.
[12] R. B. Kellogg, *On the Poisson equation with intersecting interfaces*, Appl. Anal., 4 (1975), pp. 101–129.
[13] I. Kossaczký, *A recursive approach to local mesh refinement in two and three dimensions*, J. Comput. Appl. Math., 55 (1994), pp. 275–288.
[14] J. Maubach, *Local bisection refinement for n-simplicial grids generated by reflection*, SIAM J. Sci. Comput., 16 (1995), pp. 210–227.
[15] W. F. Mitchell, *A comparison of adaptive refinement techniques for elliptic problems*, ACM Trans. Math. Software, 15 (1989), pp. 326–347.
[16] K. Mekchay and R. H. Nochetto, *Convergence of adaptive finite element methods for general second order linear elliptic PDEs*, SIAM J. Numer. Anal., 43 (2005), pp. 1803–1827.
[17] P. Morin, R. H. Nochetto, and K. G. Siebert, *Data oscillation and convergence of adaptive FEM*, SIAM J. Numer. Anal., 38 (2000), pp. 466–488.
[18] P. Morin, R. H. Nochetto, and K. G. Siebert, *Convergence of adaptive finite element methods*, SIAM Rev., 44 (2002), pp. 631–658.

[19] P. Morin, R. H. Nochetto, and K. G. Siebert, *Local problems on stars: a posteriori error estimation, convergence, and performance*, Math. Comp., 72 (2003), pp. 1067–1097.
[20] P. Morin, K.G. Siebert, and A. Veeser, *A basic convergence result for conforming adaptive finite elements*, Math. Models Methods Appl., 18 (2008), pp. 707–737.
[21] A. Schmidt and K. G. Siebert, *Design of Adaptive Finite Element Software: The Finite Element Toolboox ALBERTA*, Lect. Notes Comput. Sci. Eng. 42, Springer-Verlag, Berlin, 2005.
[22] L. R. Scott and S. Zhang, *Finite element interpolation of nonsmooth functions satisfying boundary conditions*, Math. Comp., 54 (1990), pp. 483–493.
[23] R. Stevenson, *Optimality of a standard adaptive finite element method*, Found. Comput. Math., 7 (2007), pp. 245–269.
[24] R. Stevenson, *The completion of locally refined simplicial partitions created by bisection*, Math. Comp., 77 (2008), pp. 227–241.
[25] C. T. Traxler, *An algorithm for adaptive mesh refinement in n dimensions*, Computing, 59 (1997), pp. 115–137.
[26] A. Veeser, *Convergent adaptive finite elements for the nonlinear Laplacian*, Numer. Math., 92 (2002), pp. 743–770.
[27] R. Verfürth, *A Review of A Posteriori Error Estimation and Adaptive Mesh-Refinement Technique*, Wiley–Teubner, Chichester, 1996.

SIAM J. NUMER. ANAL.
Vol. 46, No. 5, pp. 2551–2566
© 2008 Society for Industrial and Applied Mathematics

STABILITY OF FD–TD SCHEMES FOR MAXWELL–DEBYE AND MAXWELL–LORENTZ EQUATIONS*

BRIGITTE BIDÉGARAY-FESQUET†

Abstract. The stability of five finite difference–time domain (FD–TD) schemes coupling Maxwell equations to Debye or Lorentz models has been analyzed in [P. Petropoulos, *Stability and phase error analysis of FD–TD in dispersive dielectrics*, IEEE Trans. Antennas Propag., 42 (1994), pp. 62–69], where numerical evidence for specific media have been used. We use von Neumann analysis to give necessary and sufficient stability conditions for these schemes for any medium, in accordance with the partial results of [P. Petropoulos, *Stability and phase error analysis of FD–TD in dispersive dielectrics*, IEEE Trans. Antennas Propag., 42 (1994), pp. 62–69]. To make this approach tractable for two-dimensional and three-dimensional models, we have developed a computer algebra environment which has a wider range of applicability.

Key words. stability analysis, Maxwell–Debye, Maxwell–Lorentz

AMS subject classifications. 65M06, 65M12, 78M20

DOI. 10.1137/060671255

1. Introduction.

To describe the propagation of an electromagnetic wave through a dispersive medium, some extensions of Maxwell equations are used. They involve time differential equations which account for the constitutive laws of the material that link the displacement $\mathbf{D}$ to the electric field $\mathbf{E}$ or, equivalently, the polarization $\mathbf{P}$ to $\mathbf{E}$. We focus on two of these models (the Debye and Lorentz models) which are addressed in [10] in view of specific applications to the interaction of an electromagnetic wave with a human body. In [10], specific values for the physical and numerical constants are chosen and numeric calculations are performed to conclude stability or not. A survey of numerical couplings between Maxwell equations and various matter models may be found in [16], where stability conditions are given for Maxwell–Debye models, but there is no proof for some of them, and Maxwell–Lorentz formulae are considered too complex to be studied. In contrast with these two references, we are able to treat any medium which is described by these models. To this end we use von Neumann analysis.

1.1. Maxwell–Debye and Maxwell–Lorentz models.

In our context (no magnetization) we use different formulations of Maxwell equations where Faraday equations always read

$$\partial_t \mathbf{B}(t,\mathbf{x}) = -\ \mathrm{curl}\ \mathbf{E}(t,\mathbf{x}), \tag{1.1}$$

where $\mathbf{x} \in \mathbb{R}^N$. On the contrary, the Ampère equation may be cast using variables $\mathbf{B}$ and $\mathbf{D}$:

$$\partial_t \mathbf{D}(t,\mathbf{x}) = \frac{1}{\mu_0}\ \mathrm{curl}\ \mathbf{B}(t,\mathbf{x}); \tag{1.2a}$$

*Received by the editors October 2, 2006; accepted for publication (in revised form) February 13, 2008; published electronically June 18, 2008.
http://www.siam.org/journals/sinum/46-5/67125.html
†Laboratoire Jean Kuntzmann, Université de Grenoble et CNRS UMR 5224, B.P. 53, 38041 Grenoble Cedex 9, France (brigitte.bidegaray@imag.fr).

or the polarization $\mathbf{P}$:

$$\varepsilon_0\varepsilon_\infty\partial_t\mathbf{E}(t,\mathbf{x}) = \frac{1}{\mu_0}\ \mathrm{curl}\ \mathbf{B}(t,\mathbf{x}) - \partial_t\mathbf{P}(t,\mathbf{x}), \tag{1.2b}$$

where $\mathbf{P}(t,\mathbf{x}) = \mathbf{D}(t,\mathbf{x}) - \varepsilon_0\varepsilon_\infty\mathbf{E}(t,\mathbf{x})$ and ε_∞ is the relative infinite frequency permittivity. Denoting by $\mathbf{J}$ the time derivative of $\mathbf{P}$, we also have

$$\varepsilon_0\varepsilon_\infty\partial_t\mathbf{E}(t,\mathbf{x}) = \frac{1}{\mu_0}\ \mathrm{curl}\ \mathbf{B}(t,\mathbf{x}) - \mathbf{J}(t,\mathbf{x}). \tag{1.2c}$$

The system of Faraday and Ampère equations is closed by a linear constitutive law,

$$\mathbf{D}(t,\mathbf{x}) = \varepsilon_0\varepsilon_\infty\mathbf{E}(t,\mathbf{x}) + \varepsilon_0\int_{-\infty}^{t}\mathbf{E}(t-\tau,\mathbf{x})\chi(\tau)d\tau, \tag{1.3}$$

where χ is the linear susceptibility. The discretization of the integral expression (1.3) leads to recursive schemes (see, e.g., [7], [17]). However, differentiating (1.3) leads to a time differential equation for $\mathbf{D}$ which depends on the specific form of χ. For a Debye medium,

$$t_\mathrm{r}\partial_t\mathbf{D} + \mathbf{D} = t_\mathrm{r}\varepsilon_0\varepsilon_\infty\partial_t\mathbf{E} + \varepsilon_0\varepsilon_\mathrm{s}\mathbf{E}, \tag{1.4a}$$

where $t_\mathrm{r} > 0$ is the relaxation time and $\varepsilon_\mathrm{s} \geq \varepsilon_\infty$ is the relative static permittivity. An equivalent form of (1.4a) using variable $\mathbf{P}$ is

$$t_\mathrm{r}\partial_t\mathbf{P} + \mathbf{P} = \varepsilon_0(\varepsilon_\mathrm{s} - \varepsilon_\infty)\mathbf{E}, \tag{1.4b}$$

which can be coupled with the Ampère equation cast as (1.2b). For a Lorentz medium with one resonant frequency ω_1, we likewise have

$$\partial_t^2\mathbf{D} + \nu\partial_t\mathbf{D} + \omega_1^2\mathbf{D} = \varepsilon_0\varepsilon_\infty\partial_t^2\mathbf{E} + \varepsilon_0\varepsilon_\infty\nu\partial_t\mathbf{E} + \varepsilon_0\varepsilon_\mathrm{s}\omega_1^2\mathbf{E}, \tag{1.5a}$$

where $\nu \geq 0$ is a damping coefficient, or, equivalently,

$$\partial_t^2\mathbf{P} + \nu\partial_t\mathbf{P} + \omega_1^2\mathbf{P} = \varepsilon_0(\varepsilon_\mathrm{s} - \varepsilon_\infty)\omega_1^2\mathbf{E}. \tag{1.5b}$$

1.2. Yee-based numerical schemes. A classical and very efficient way to compute the Maxwell equations is the Yee scheme [14]. We restrict our study to existing Yee-based schemes. In contrast to the recursive schemes, we are interested in direct integration schemes which are based on the finite difference–time domain (FD–TD) discretization of (1.4a) to (1.5b) (see [5], [4], [15]). Other space discretizations may be found in the literature in the context of Maxwell–Debye and Maxwell–Lorentz equations; see, e.g., [3] for pseudo-spectral schemes or [12] for finite element–time domain (FE–TD) schemes.

The Yee scheme consists of discretizing $\mathbf{E}$ and $\mathbf{B}$ on staggered grids in space and time. This allows one to use only centered discrete differential operators. We denote by h the space step (supposed here to be the same in all directions in the case of multi-dimensional equations) and by k the time step. In space dimension 1, we consider only the dependence in the space variable z; classically, two polarizations for the field may be decoupled. For example, the transverse electric polarization involves only $E \equiv E_x$ and $B \equiv B_y$. The discretized variables are $E_j^n \simeq E(nk, jh)$ (and similar

notations for $D \equiv D_x$) and $B^{n+\frac{1}{2}}_{j+\frac{1}{2}} \simeq B((n+\frac{1}{2})k, (j+\frac{1}{2})h)$, and the Yee scheme for the Maxwell equation reads

$$\frac{B^{n+\frac{1}{2}}_{j+\frac{1}{2}} - B^{n-\frac{1}{2}}_{j+\frac{1}{2}}}{k} = -\frac{E^n_{j+1} - E^n_j}{h}, \tag{1.6}$$

coupled with one of the following:

$$\frac{D^{n+1}_j - D^n_j}{k} = -\mu_0 \frac{B^{n+\frac{1}{2}}_{j+\frac{1}{2}} - B^{n+\frac{1}{2}}_{j-\frac{1}{2}}}{h}, \tag{1.7a}$$

$$\varepsilon_0\varepsilon_\infty \frac{E^{n+1}_j - E^n_j}{k} = -\mu_0 \frac{B^{n+\frac{1}{2}}_{j+\frac{1}{2}} - B^{n+\frac{1}{2}}_{j-\frac{1}{2}}}{h} - \frac{P^{n+1}_j - P^n_j}{k}, \tag{1.7b}$$

$$\varepsilon_0\varepsilon_\infty \frac{E^{n+1}_j - E^n_j}{k} = -\mu_0 \frac{B^{n+\frac{1}{2}}_{j+\frac{1}{2}} - B^{n+\frac{1}{2}}_{j-\frac{1}{2}}}{h} - J^{n+\frac{1}{2}}_j. \tag{1.7c}$$

Usual Maxwell equations consist of taking $J^{n+\frac{1}{2}}_j \equiv 0$ in (1.7c) or, equivalently, $D^n_j = \varepsilon_0\varepsilon_\infty E^n_j$ in (1.7a) and lead to a stable second order scheme under a Courant–Friedrichs–Lewy (CFL) stability condition. Namely, if $c_\infty = 1/\sqrt{\varepsilon_0\varepsilon_\infty\mu_0}$ denotes the infinite frequency light speed, then the CFL condition reads $c_\infty k \le h/\sqrt{N}$.

1.3. Scheme naming. Since we deal with many schemes, we have to distinguish between them and name them. Numbers or names of first contributor(s) are not very meaningful, and we prefer here to have a descriptive name. Our description gives both the variables used (e.g., coupling (1.2a) and (1.4a) uses variables **B**, **E**, and **D**) and the location of discretized variables. Space occurs only as a parameter in Debye or Lorentz equations. To avoid interpolation, the reasonable choice is to always locate variables **D**, **P**, and **J** on the same space grid as **E**. The different schemes therefore differ only in the time location of variables: in (1.7a) and (1.7b) integer times are chosen; in (1.7c) half-integer times are chosen.

Our naming convention is the following: we separate variables at half-integer times from variables at integer times by an underscore sign, e.g., "B_ED" when coupling (1.6) and (1.7a). In Table 1.1 we give the correspondence between the article where the schemes have been derived first, the terminology of the survey [16] and our description.

TABLE 1.1
Correspondence between different namings.

Contributor(s)	Name in [16]	Description	Name in [16]	Description
	Debye		Lorentz	
Joseph et al. [4]	D-DIM 3	B_ED	L-DIM 3	B_ED
Kashiwa et al. [5]	D-DIM 2	B_EP	L-DIM 2	B_EPJ
Young [15]	D-DIM 1	BP_E	L-DIM 1	BJ_EP

1.4. Outline. The von Neumann stability analysis is recalled in section 2. We also describe the sketch of our proofs which is common for all of the schemes. In section 3, three one-dimensional direct integration schemes for Debye media are presented and analyzed, carefully pointing out the physical properties needed to ensure stability and the specific cases which have to be handled separately. Numerical applications to

physical media are also given. The same point of view is carried out for Lorentz media in section 4. To address two- and three-dimensional schemes which have too many variables to compute by hand, we have developed a computer algebra environment based on MAPLE, which is described in section 5. Two- and three dimensional results are given in section 6.

2. Principles of the von Neumann analysis. The von Neumann analysis allows one to localize roots of certain classes of polynomials, which proves to be crucial here. We recall the main principles of this technique. Details and proofs of theorems may be found in [13].

2.1. Schur and simple von Neumann polynomials. We define two families of polynomials: Schur polynomials and simple von Neumann polynomials.

DEFINITION 2.1. *A polynomial is a Schur polynomial if all its roots, r, satisfy $|r| < 1$.*

DEFINITION 2.2. *A polynomial is a simple von Neumann polynomial if all its roots, r, lie on the unit disk ($|r| \leq 1$) and its roots on the unit circle are simple roots.*

If a polynomial has a high degree or sophisticated coefficients, it may be difficult to locate its roots. However, there is a way to split this difficult problem into many simpler ones. For this, we construct a sequence of polynomials of decreasing degree. Let ϕ be written as

$$\phi(z) = c_0 + c_1 z + \cdots + c_p z^p,$$

where $c_0, c_1, \ldots, c_p \in \mathbb{C}$ and $c_p \neq 0$. We define its conjugate polynomial ϕ^* by

$$\phi^*(z) = c_p^* + c_{p-1}^* z + \cdots + c_0^* z^p.$$

Given a polynomial ϕ_0, we may define a sequence of polynomials

$$\phi_{m+1}(z) = \frac{\phi_m^*(0)\phi_m(z) - \phi_m(0)\phi_m^*(z)}{z}.$$

It is clear that $\deg\phi_{m+1} < \deg\phi_m$, if $\phi_m \not\equiv 0$.

THEOREM 2.3. *A polynomial ϕ_m is a Schur polynomial of exact degree d if and only if ϕ_{m+1} is a Schur polynomial of exact degree $d-1$ and $|\phi_m(0)| \leq |\phi_m^*(0)|$.*

THEOREM 2.4. *A polynomial ϕ_m is a simple von Neumann polynomial if and only if*

- *ϕ_{m+1} is a simple von Neumann polynomial and $|\phi_m(0)| \leq |\phi_m^*(0)|$*

or

- *ϕ_{m+1} is identically zero and ϕ_m' is a Schur polynomial.*

The main ingredient in the proof of both theorems is the Rouché theorem (see [9], [13]). To analyze ϕ_0, at each step m, conditions should be checked (leading coefficient is nonzero, $|\phi_m(0)| \leq |\phi_m^*(0)|$, ...) until a definitive negative answer arises or the degree is 1.

2.2. Stability analysis. The models we deal with are linear models. They may therefore be analyzed in the frequency domain. We can thus assume that the scheme handles a single vector-valued variable $U_{\mathbf{j}}^n$ with spatial dependence

$$U_{\mathbf{j}}^n = U^n \exp(i\boldsymbol{\xi} \cdot \mathbf{j}),$$

where $\boldsymbol{\xi}$ and $\mathbf{j} \in \mathbb{R}^N$, $N = 1, 2, 3$. The amplification matrix G is the matrix such that $U^{n+1} = GU^n$. We assume that G does not depend on time or on h and k separately

but only on the ratio h/k. Let ϕ_0 be the characteristic polynomial of G; then we have a sufficient stability condition.

THEOREM 2.5. *A sufficient stability condition is that ϕ_0 be a simple von Neumann polynomial.*

This condition is not necessary. A scheme is stable if and only if the sequence $(U^n)_{n\in\mathbb{N}}$ is bounded. Since we assume that G does not depend on time, then $U^n = G^nU^0$ and stability is also the boundedness of $(G^n)_{n\in\mathbb{N}}$. If the eigenvalues of G, i.e., the roots r of ϕ_0, lie inside the unit circle ($|r| < 1$), then $\lim_{n\to\infty} G^n = 0$ and the sequence is bounded. If any root lies outside the unit circle, then G^n grows exponentially and the scheme is unstable. The intermediate case when some roots may be on the unit circle (and the others inside) may lead to different situations. Consider, for example, the case when G is the identity. Then $U^n = U^0$ and the scheme is clearly stable. However, there are other examples of matrices with multiple roots on the unit circle that lead either to bounded or unbounded sequences $(G^n)_{n\in\mathbb{N}}$. It is clearly a property of the amplification matrix and not of its characteristic polynomial. If the dimension of eigensubspace associated with a root is equal to its multiplicity, then G^n is bounded (an example for this is the identity: $\mathrm{Id}^n = \mathrm{Id}$ remains bounded). In the opposite case G^n grows linearly. A paradigm for this is

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}^n = \begin{pmatrix} 1 & n \\ 0 & 1 \end{pmatrix},$$

which grows linearly with iterations, and only one eigenvector can be found. Such cases (which occur for our schemes) should therefore be handled specifically.

We will stick here to locating roots in the unit circle. Another way has already been developed in, e.g., [6], [11]. It consists of using a conformal map to locate roots in the left half-plane. This changes the problem in a Routh–Hurwitz problem, which reads as a list of sufficient conditions for the characteristic polynomial to be von Neumann ($|r| \le 1$). This has been implemented in the FIDE REDUCE package.

2.3. Sketch of proofs. In the next sections, we will not give the proofs, but only list in tables the arguments used for each situation. We describe here the general plan and give names to specific final arguments used. The detailed proofs may be found in [1] for space dimensions 1 and 2. The three-dimensional case has been performed *via* computer algebra simulations [2].

Usually the coupled system is given in an implicit form. The first step consists of writing it in an explicit form. This yields the amplification matrix G. Then we compute its characteristic polynomial ϕ_0. In order to perform a von Neumann analysis, we compute the series (ϕ_m). In the general case, under the assumption that the stability condition cannot be better than Yee's, we can apply either Theorem 2.3 (*Theorem* 2.3 argument) or Theorem 2.4 (*Theorem* 2.4 argument), checking estimates at each level until ϕ_m is a one degree polynomial. Special cases arise when $\varepsilon_s = \varepsilon_\infty$, $\sin(\xi/2) = 0$ or ± 1, and sometimes for limit values of physical coefficients. In these cases, different points of view have to be considered:

- Theorem 2.4 has to be used instead of Theorem 2.3.
- Some eigenvalues lie on the unit circle (mostly ± 1 or $\pm i$). Polynomial ϕ_0 is at best a simple von Neumann polynomial and we have to study the other eigenvalues which are roots of a lower degree polynomial (*subpolynomial* argument).
- Some eigenvalues lie on the unit circle and are not simple. Besides the study of the other eigenvalues (*subpolynomial* argument), we have to compare the

multiplicity of the root to the number of eigenvectors found. If they are equal (stable case), then this is usually checked directly on the form of matrix G (*G form* argument). The unstable case necessitates the computation of eigenvectors (*eigenvectors* argument).

If $\varepsilon_s = \varepsilon_\infty$, then Maxwell equations decouple from the other equations. We study this limit as a hint of how the numerical scheme may behave when ε_s is very close to ε_∞. For the same reason, we will study the undamped oscillator ($\nu = 0$) for Lorentz media.

3. Debye media. We address successively the three discretizations of Maxwell–Debye equations mentioned in Table 1.1.

3.1. Debye B_ED scheme. In [4], Joseph, Hagness, and Taflove close system (1.6)–(1.7a) by a discretization for (1.4a), namely

$$\varepsilon_0\varepsilon_\infty t_\mathrm{r}\frac{E_j^{n+1}-E_j^n}{k}+\varepsilon_0\varepsilon_\mathrm{s}\frac{E_j^{n+1}+E_j^n}{2}=t_\mathrm{r}\frac{D_j^{n+1}-D_j^n}{k}+\frac{D_j^{n+1}+D_j^n}{2}.$$

The resulting system may be cast in an explicit form which handles the variable (^{t} denotes transposition)

$$U_j^n=\left(c_\infty B_{j+\frac12}^{n-\frac12},E_j^n,D_j^n/\varepsilon_0\varepsilon_\infty\right)^t,$$

and the amplification matrix G reads

$$\begin{pmatrix} 1 & -\lambda(e^{i\xi}-1) & 0\\ -\frac{(1+\delta)\lambda(1-e^{-i\xi})}{1+\delta\eta_\mathrm{s}} & \frac{(1-\delta\eta_\mathrm{s})+(1+\delta)\lambda^2(e^{i\xi}-2+e^{-i\xi})}{1+\delta\eta_\mathrm{s}} & \frac{2\delta}{1+\delta\eta_\mathrm{s}}\\ -\lambda(1-e^{-i\xi}) & \lambda^2(e^{i\xi}-2+e^{-i\xi}) & 1 \end{pmatrix},$$

where $\lambda = c_\infty k/h$ is the CFL constant, $\delta = k/2t_\mathrm{r} > 0$ is the normalized time step, and $\eta_\mathrm{s} = \varepsilon_\mathrm{s}/\varepsilon_\infty \geq 1$ denotes the normalized static permittivity. Moreover, we define $q = -\lambda^2(e^{i\xi}-2+e^{-i\xi}) = 4\lambda^2\sin^2(\xi/2)$. The characteristic polynomial is proportional to

$$\phi_0(Z) = [1+\delta\eta_\mathrm{s}]Z^3 - [3+\delta\eta_\mathrm{s}-(1+\delta)q]Z^2 + [3-\delta\eta_\mathrm{s}-(1-\delta)q]Z - [1-\delta\eta_\mathrm{s}].$$

The proof arguments are summed up in Table 3.1, and we deduce that the stability condition is $q \leq 4$ if $\varepsilon_\mathrm{s} > \varepsilon_\infty$ and $q < 4$ if $\varepsilon_\mathrm{s} = \varepsilon_\infty$.

TABLE 3.1
Proof arguments and results for the Debye B_ED and B_EP schemes.

q	ε_s	Argument	Result
$]0,4[$	$> \varepsilon_\infty$	Theorem 2.3	stable
$]0,4[$	$= \varepsilon_\infty$	Theorem 2.4	stable
0	$\geq \varepsilon_\infty$	G form	stable
4	$> \varepsilon_\infty$	Theorem 2.4	stable
4	$= \varepsilon_\infty$	eigenvectors	unstable

3.2. Debye B_EP scheme. In [5], Kashiwa, Yoshida, and Fukai close system (1.6)–(1.7b) by a discretization for (1.4b):

$$t_{\mathrm{r}}\frac{P_j^{n+1}-P_j^n}{k}=-\frac{P_j^{n+1}+P_j^n}{2}+\varepsilon_0(\varepsilon_{\mathrm{s}}-\varepsilon_\infty)\frac{E_j^{n+1}+E_j^n}{2}.$$

The system now handles the variable

$$U_j^n=\left(c_\infty B_{j+\frac{1}{2}}^{n-\frac{1}{2}},E_j^n,P_j^n/\varepsilon_0\varepsilon_\infty\right)^t,$$

the amplification matrix G reads

$$\begin{pmatrix} 1 & -\lambda(e^{i\xi}-1) & 0\\ -\frac{(1+\delta)\lambda(1-e^{-i\xi})}{1+\delta\eta_{\mathrm{s}}} & \frac{(1-\delta\eta_{\mathrm{s}})+(1+\delta)\lambda^2(e^{i\xi}-2+e^{-i\xi})+2\delta}{1+\delta\eta_{\mathrm{s}}} & \frac{2\delta}{1+\delta\eta_{\mathrm{s}}}\\ -\frac{\delta(\eta_s-1)\lambda(1-e^{-i\xi})}{1+\delta\eta_{\mathrm{s}}} & \frac{(\lambda^2(e^{i\xi}-2+e^{-i\xi})+2)(\eta_s-1)\delta}{1+\delta\eta_{\mathrm{s}}} & \frac{1+\eta_s\delta-2\delta}{1+\delta\eta_{\mathrm{s}}} \end{pmatrix},$$

and the characteristic polynomial is proportional to the same polynomial as for the B_ED scheme. All proofs are the same except those on matrix G directly, but even in these cases the conclusions are the same. Table 3.1 is also valid for the B_EP scheme.

3.3. Debye BP_E scheme. In [15], Young chooses to close system (1.6)–(1.7c) by two discretizations for (1.4b), namely

$$t_{\mathrm{r}}\frac{P_j^{n+\frac{1}{2}}-P_j^{n-\frac{1}{2}}}{k}=-\frac{P_j^{n+\frac{1}{2}}+P_j^{n-\frac{1}{2}}}{2}+\varepsilon_0(\varepsilon_{\mathrm{s}}-\varepsilon_\infty)E_j^n,$$

$$t_{\mathrm{r}}J_j^{n+\frac{1}{2}}=-P_j^{n+\frac{1}{2}}+\varepsilon_0(\varepsilon_{\mathrm{s}}-\varepsilon_\infty)\frac{E_j^{n+1}+E_j^n}{2}.$$

Although $J_j^{n+\frac{1}{2}}$ is used for the computations, this not a genuine variable for the full system which handles the variable

$$U_j^n=\left(c_\infty B_{j+\frac{1}{2}}^{n-\frac{1}{2}},E_j^n,P_j^{n-\frac{1}{2}}/\varepsilon_0\varepsilon_\infty\right)^t,$$

and the amplification matrix G reads

$$\begin{pmatrix} 1 & -\lambda(e^{i\xi}-1) & 0\\ -\frac{\lambda(1-e^{-i\xi})}{1+\delta\alpha} & \frac{1+\delta-\delta\alpha+3\delta^2\alpha-(1+\delta)q}{(1+\delta)(1+\delta\alpha)} & \frac{1-\delta}{1+\delta}\frac{2\delta}{1+\delta\alpha}\\ 0 & \frac{2\delta\alpha}{1+\delta} & \frac{1-\delta}{1+\delta} \end{pmatrix}$$

with the same notations as above and $\alpha=\eta_{\mathrm{s}}-1\geq 0$. The characteristic polynomial is proportional to

$$\phi_0(Z)=[(1+\delta\alpha)(1+\delta)]Z^3-[3+\delta+\delta\alpha+3\delta^2\alpha-(1+\delta)q]Z^2$$
$$+[3-\delta-\delta\alpha+3\delta^2\alpha-(1-\delta)q]Z-[(1-\delta\alpha)(1-\delta)].$$

Table 3.2 gathers the different arguments used. The stability condition is $q\leq 4$ and $\delta\leq 1$ if $\varepsilon_{\mathrm{s}}>\varepsilon_\infty$ and $q<4$ if $\varepsilon_{\mathrm{s}}=\varepsilon_\infty$.

TABLE 3.2
Proof arguments and results for the Debye BP_E scheme.

q	ε_s	δ	Argument	Result
$]0,4]$	$>\varepsilon_\infty$	$]0,1[$	Theorem 2.3	stable
$]0,4[$	$=\varepsilon_\infty$	>0	Theorem 2.4	stable
0	$\geq\varepsilon_\infty$	>0	G form	stable
$]0,4]$	$>\varepsilon_\infty$	1	subpolynomial	stable
4	$=\varepsilon_\infty$	>0	eigenvectors	unstable

3.4. Conclusion for one-dimensional Debye schemes. If $\varepsilon_s > \varepsilon_\infty$, then the pure CFL condition $q \leq 4$ is the same for both models. It is exactly the condition for Maxwell equations. However, the BP_E scheme necessitates another condition, $\delta \leq 1$, which corresponds to a sufficient discretization of Debye equation (1.4b). Even if we are interested here in stability properties, such conditions are to be considered to ensure equations to be correctly taken into account. Results are given in physical variables in Table 3.3.

TABLE 3.3
Stability of Debye models for $\varepsilon_s > \varepsilon_\infty$ and $\varepsilon_s = \varepsilon_\infty$.

Scheme		Dimension 1		Dimension 1
	$\varepsilon_s > \varepsilon_\infty$		$\varepsilon_s = \varepsilon_\infty$	
B_ED or B_EP	$q \leq 4$	$k \leq h/c_\infty$	$q < 4$	$k < h/c_\infty$
BP_E	$q \leq 4$, $\delta \leq 1$	$k \leq \min(h/c_\infty, 2t_r)$	$q < 4$	$k < h/c_\infty$

To compare conditions on q and δ, let us consider a simple physical case. We assume that a matter with $\varepsilon_\infty = 1$ (and thus $c_\infty \simeq 3\times10^8\,\mathrm{m\,s^{-1}}$) is lighted by a wave of, say, wavelength 1 cm. The space step h has to be smaller than this wavelength, and therefore $q < 4$ reads at least $k < \frac{1}{3}\times10^{-10}\,\mathrm{s}$. In a Debye medium, relaxation times t_r are of the order of a picosecond (or even a nanosecond) which maybe of the same order than the previous bound. A typical example is water, for which $\varepsilon_\infty = 1.8$, $\varepsilon_s = 81.0$, and $t_r = 9.4\times10^{-12}\,\mathrm{s}$ [17]. Condition $k \leq 2t_r$ comes to $k \leq 1.88\times10^{-11}\,\mathrm{s}$. Condition $q \leq 4$ yields a similar condition if $h = 4.2\times10^{-3}\,\mathrm{m}$. For a very coarse space grid and with the BP_E scheme, the CFL condition comes from $k \leq 2t_r$; otherwise, for a fine grid and/or B_ED or B_EP schemes, it comes from $q \leq 4$.

A quite different material is, for example, the 0.25-dB loaded foam given in [8] for which $\varepsilon_\infty = 1.01$, $\varepsilon_s = 1.16$, and $t_r = 6.497\times10^{-10}\,\mathrm{s}$. Condition $k \leq 2t_r$ comes to $k \leq 1.3\times10^{-9}\,\mathrm{s}$ and $q \leq 4$ yields a similar condition if $h = 3.9\times10^{-1}\,\mathrm{m}$. In practice h has to be smaller to describe a wave with 1 cm wavelength, and the stability condition for this foam is $q < 4$ for both schemes.

In conclusion, the stability condition may depend on the material in classical applications, leading us to prefer the B_ED scheme, when t_r is small (picosecond). The result announced in [10] was $q \leq 4$ for the B_ED scheme and for water, which is consistent with our result.

4. Lorentz media. We now address the three discretizations of Maxwell–Lorentz equations mentioned in Table 1.1.

Each of these schemes reads the same in the undamped ($\nu = 0$) or damped ($\nu > 0$) cases. However, the analysis will differ greatly since $\phi_1 \equiv 0$ for all the schemes in the undamped case.

4.1. Lorentz B_ED scheme. In [4], system (1.6)–(1.7a) is closed by a discretization for (1.5a), namely

$$\varepsilon_0\varepsilon_\infty\frac{E_j^{n+1}-2E_j^n+E_j^{n-1}}{k^2}+\nu\varepsilon_0\varepsilon_\infty\frac{E_j^{n+1}-E_j^{n-1}}{2k}+\varepsilon_0\varepsilon_{\rm s}\omega_1^2\frac{E_j^{n+1}+E_j^{n-1}}{2}$$
$$=\frac{D_j^{n+1}-2D_j^n+D_j^{n-1}}{k^2}+\nu\frac{D_j^{n+1}-D_j^{n-1}}{2k}+\omega_1^2\frac{D_j^{n+1}+D_j^{n-1}}{2}.$$

The explicit version of the subsequent system does not explicitly use the value of D_j^{n-1}, and therefore this system handles the variable

$$U_j^n=\left(c_\infty B_{j+\frac12}^{n-\frac12},E_j^n,E_j^{n-1},D_j^n/\varepsilon_0\varepsilon_\infty\right)^t.$$

The amplification matrix G reads

$$\begin{pmatrix}1&-\lambda(e^{i\xi}-1)&0&0\\ -\frac{2\delta\lambda(1-e^{-i\xi})}{1+\delta+\omega\eta_{\rm s}}&\frac{2-q(1+\delta+\omega)}{1+\delta+\omega\eta_{\rm s}}&\frac{1-\delta+\omega\eta_{\rm s}}{1+\delta+\omega\eta_{\rm s}}&\frac{2\omega}{1+\delta+\omega\eta_{\rm s}}\\ 0&1&0&0\\ -\lambda(1-e^{-i\xi})&-q&0&1\end{pmatrix},$$

where $\delta=\nu k/2\geq 0$ is the new normalized time step, and $\omega=\omega_1^2k^2/2>0$ denotes the normalized squared frequency. The other notations used for the Debye model remain valid. The characteristic polynomial is proportional to

$$\phi_0(Z)=[1+\delta+\omega\eta_{\rm s}]Z^4-[4+2\delta+2\omega\eta_{\rm s}-(1+\delta+\omega)q]Z^3$$
$$+[6+2\omega\eta_{\rm s}-2q]Z^2-[4-2\delta+2\omega\eta_{\rm s}-(1-\delta+\omega)q]Z+[1-\delta+\omega\eta_{\rm s}].$$

The proofs are summed up in Table 4.1 for the damped and the undamped case.

TABLE 4.1
Proof arguments and results for the Lorentz B_ED scheme.

q	$\varepsilon_{\rm s}$	Argument	Result	Argument	Result
		damped: $\nu>0$		undamped: $\nu=0$	
$]0,2[$	$>\varepsilon_\infty$	Theorem 2.3	stable	Theorem 2.4	stable
$]0,2]$	$=\varepsilon_\infty$	Theorem 2.4	stable	subpolynomial	unstable
0	$\geq\varepsilon_\infty$	G form	stable	G form	stable
2	$\geq\varepsilon_\infty$	subpolynomial	stable	subpolynomial	stable

In the damped case the stability condition is $q\leq 2$ for all $\varepsilon_{\rm s}\geq\varepsilon_\infty$. The $\varepsilon_{\rm s}=\varepsilon_\infty$ undamped case needs some explanation. For $q\in]0,2]$, ϕ_0 may be cast as the product of two second order polynomials. The roots are two couples of conjugate complex roots of modulus 1. For the specific value $q=2\omega/(1+\omega)$, which always lies in the interval $]0,2]$, the two couples degenerate in one double couple with only two eigendirections, which is the unstable case. To avoid this instability one may think to bound q and say that the scheme is stable provided $q\in[0,2\omega/(1+\omega)[$. But if we come back to the original variables, we see that this is not an upper bound on k but rather a lower bound on h, which we surely do not want. It is therefore better to avoid using the B_ED scheme in this very specific case, $\varepsilon_{\rm s}=\varepsilon_\infty$ and $\nu=0$. A better scheme for this case is provided next.

4.2. Lorentz B_EPJ scheme. In [5], Kashiwa, Yoshida, and Fukai close system (1.6)–(1.7b) by a discretization for (1.5b), namely

$$\frac{P_j^{n+1}-P_j^n}{k}=\frac{J_j^{n+1}+J_j^n}{2},$$

$$\frac{J_j^{n+1}-J_j^n}{k}=-\nu\frac{J_j^{n+1}+J_j^n}{2}+\omega_1^2(\varepsilon_{\rm s}-\varepsilon_\infty)\varepsilon_0\frac{E_j^{n+1}+E_j^n}{2}-\omega_1^2\frac{P_j^{n+1}+P_j^n}{2}.$$

The explicit version of the system obtained handles the variable

$$U_j^n=\left(c_\infty B_{j+\frac12}^{n-\frac12},E_j^n,P_j^n/\varepsilon_0\varepsilon_\infty,kJ_j^n/\varepsilon_0\varepsilon_\infty\right)^t,$$

and the amplification matrix G reads

$$\begin{pmatrix} 1 & -\lambda(e^{i\xi}-1) & 0 & 0\\ \frac{-\lambda(1-e^{-i\xi})(\Delta-\frac12\omega\alpha)}{\Delta} & \frac{\Delta-q\Delta-(2-q)\frac12\omega\alpha}{\Delta} & \frac{\omega}{\Delta} & \frac{-1}{\Delta}\\ \frac{-\lambda(1-e^{-i\xi})\frac12\omega\alpha}{\Delta} & \frac{(2-q)\frac12\omega\alpha}{\Delta} & \frac{\Delta-\omega}{\Delta} & \frac{1}{\Delta}\\ \frac{-\lambda(1-e^{-i\xi})\omega\alpha}{\Delta} & \frac{(2-q)\omega\alpha}{\Delta} & \frac{-2\omega}{\Delta} & \frac{2-\Delta}{\Delta} \end{pmatrix},$$

where together with the previously defined notations, $\Delta=1+\delta+\omega\eta_{\rm s}/2$. The characteristic polynomial is proportional to

$$\phi_0(Z)=\left[1+\delta+\frac12\omega\eta_{\rm s}\right]Z^4-\left[4+2\delta-\left(1+\delta+\frac12\omega\right)q\right]Z^3$$
$$+[6-\omega\eta_{\rm s}+(\omega-2)q]Z^2-\left[4-2\delta-\left(1-\delta+\frac12\omega\right)q\right]Z+\left[1-\delta+\frac12\omega\eta_{\rm s}\right].$$

The proofs are summed up in Table 4.2. Both in the damped and undamped cases, the stability condition is $q<4$, which is much better than the previous scheme since we gain a factor 2 on k, and we have no problem when $\varepsilon_{\rm s}=\varepsilon_\infty$ and $\nu=0$ as for the previous model.

TABLE 4.2
Proof arguments and results for the Lorentz B_EPJ scheme.

q	$\varepsilon_{\rm s}$	Argument	Result	Argument	Result
		damped: $\nu>0$		undamped: $\nu=0$	
$]0,4[$	$>\varepsilon_\infty$	Theorem 2.3	stable	Theorem 2.4	stable
$]0,4[$	$=\varepsilon_\infty$	Theorem 2.4	stable	Theorem 2.4	stable
0	$\geq\varepsilon_\infty$	G form	stable	G form	stable
4	$\geq\varepsilon_\infty$	eigenvectors	unstable	eigenvectors	unstable

4.3. Lorentz BJ_EP scheme. In [15], system (1.6)–(1.7c) is closed by a discretization for (1.5b), namely

$$\frac{P_j^{n+1}-P_j^n}{k}=J^{n+\frac12},$$

$$\frac{J_j^{n+\frac12}-J_j^{n-\frac12}}{k}=-\nu\frac{J_j^{n+\frac12}+J_j^{n-\frac12}}{2}+\omega_1^2(\varepsilon_{\rm s}-\varepsilon_\infty)\varepsilon_0E_j^n-\omega_1^2P_j^n.$$

The explicit version of the system handles the variable

$$U_j^n = \left(c_\infty B_{j+\frac{1}{2}}^{n-\frac{1}{2}}, E_j^n, P_j^n/\varepsilon_0\varepsilon_\infty, kJ_j^{n-\frac{1}{2}}/\varepsilon_0\varepsilon_\infty\right)^t,$$

and the amplification matrix G reads

$$\begin{pmatrix} 1 & -\lambda(e^{i\xi}-1) & 0 & 0 \\ -\lambda(1-e^{-i\xi}) & \frac{(1-q)(1+\delta)-2\omega\alpha}{1+\delta} & \frac{2\omega}{1+\delta} & -\frac{1-\delta}{1+\delta} \\ 0 & \frac{2\omega\alpha}{1+\delta} & \frac{1+\delta-2\omega}{1+\delta} & \frac{1-\delta}{1+\delta} \\ 0 & \frac{2\omega\alpha}{1+\delta} & \frac{-2\omega}{1+\delta} & \frac{1-\delta}{1+\delta} \end{pmatrix}.$$

The characteristic polynomial is proportional to

$$\phi_0(Z) = [1+\delta]Z^4 - [4+2\delta-2\omega\eta_s-(1+\delta)q]Z^3 + 2[3-2\omega\eta_s+(\omega-1)q]Z^2 - [4-2\delta-2\omega\eta_s-(1-\delta)q]Z + [1-\delta].$$

TABLE 4.3
Proof arguments and results for the Lorentz BJ_EP scheme.

q	ε_s	ω	Argument	Result
damped: $\nu > 0$				
$]0,2[$	$> \varepsilon_\infty$	$\leq 2/(2\eta_s-1)$	Theorem 2.3	stable
2	$> \varepsilon_\infty$	$< 2/(2\eta_s-1)$		
$]0,2]$	$= \varepsilon_\infty$	< 2	Theorem 2.4	stable
$]0,2]$	$= \varepsilon_\infty$	$= 2$	subpolynomial	stable
2	$> \varepsilon_\infty$	$= 2/(2\eta_s-1)$	Theorem 2.4	stable
0	$\geq \varepsilon_\infty$	$\leq 2/(2\eta_s-1)$	G form	stable
q	ε_s	ω	Argument	Result
undamped: $\nu = 0$				
$]0,2[$	$> \varepsilon_\infty$	$\leq 2/(2\eta_s-1)$	Theorem 2.4	stable
2	$> \varepsilon_\infty$	$< 2/(2\eta_s-1)$		
$]0,2]$	$= \varepsilon_\infty$	< 2	eigenvectors	unstable
$]0,2]$	$= \varepsilon_\infty$	$= 2$	Theorem 2.4	stable
2	$> \varepsilon_\infty$	$= 2/(2\eta_s-1)$	eigenvectors	unstable
0	$> \varepsilon_\infty$	$\leq 2/(2\eta_s-1)$	G form	stable
0	$= \varepsilon_\infty$	$< 2/(2\eta_s-1)$		
0	$= \varepsilon_\infty$	$= 2/(2\eta_s-1)$	eigenvectors	unstable

The proofs are summed up in Table 4.3. This scheme combines three drawbacks we have already encountered. First, as for the Debye model, there is an extra condition on the time step: $\omega < 2/(2\eta_s - 1)$. This will have to be compared to the condition on q for physical examples. Second, as for the Lorentz B_ED scheme, we need a twice smaller k than for raw Maxwell equations: $q \leq 2$ instead of $q \leq 4$. Last, and also as for the Lorentz B_ED scheme, the $\varepsilon_s = \varepsilon_\infty$ and $\nu = 0$ lead to an instability. This is exactly the same story. This time $q = 2\omega$ leads to double couples of conjugate complex roots of modulus 1, with only two eigendirections. If $\omega > 1$, then this value of q is, however, never reached. Else $q = 2\omega$ is rather a condition on h and therefore not a proper stability condition. As for the Lorentz B_ED, it seems better to avoid using this scheme if $\varepsilon_s = \varepsilon_\infty$ and $\nu = 0$.

4.4. Conclusion for one-dimensional Lorentz schemes. We summarize all our results for Lorentz schemes in Table 4.4. For the undamped BJ_EP scheme, if $\varepsilon_s > \varepsilon_\infty$, then the condition is slightly better since $q = 2$ and $\omega < 2/(2\eta_s - 1)$, or $q < 2$ and $\omega = 2/(2\eta_s - 1)$ also yield stable schemes.

Contrarily to Debye materials, for which all schemes compete, the B_EPJ scheme seems to overcome others for Lorentz material. First, there is a gain in CFL condition ($q < 4$ is twice better as $q \leq 2$), second, there are no instabilities for limiting values of the physical coefficients, and last, there is no extra condition on the time step. In practice, an extra condition is, however, needed to account for the dynamics of the Lorentz equation, but not for stability reasons.

However, we can compare the relative strength of the different conditions on k for the B_ED and BJ_EP schemes, and for optical waves of, say, wavelength $1\,\mu$m. The values used in [10] are $\varepsilon_\infty = 1$, $\varepsilon_s = 2.25$, $\omega_1 = 4 \times 10^{16}\,\mathrm{rad\,s^{-1}}$, and $\nu = 0.56 \times 10^{16}\,\mathrm{rad\,s^{-1}}$. Condition $\omega \leq 2/\sqrt{2\eta_s - 1}$ comes to $k \leq 2.7 \times 10^{-17}\,\mathrm{s}$, which is very small and corresponds to $h = 1.13 \times 10^{-8}\,\mathrm{m}$ in the $q < 2$ condition. This space step is more than sufficient to discretize optical waves. For such a material the extra condition imposed by the B_ED scheme is stronger than the basic CFL condition. The B_EPJ model is then more advisable.

TABLE 4.4
Stability of damped and undamped Lorentz models for $\varepsilon_s > \varepsilon_\infty$ and $\varepsilon_s = \varepsilon_\infty$.

Scheme		Dimension 1
damped: $\nu > 0$, and $\varepsilon_s \geq \varepsilon_\infty$		
B_ED	$q \leq 2$	$k \leq h/\sqrt{2}c_\infty$
B_EPJ	$q < 4$	$k < h/c_\infty$
BJ_EP	$q \leq 2$, $\omega \leq 2/(2\eta_s - 1)$	$k \leq \min(h/\sqrt{2}c_\infty, 2/\omega_1\sqrt{2\eta_s - 1})$
undamped: $\nu = 0$, and $\varepsilon_s > \varepsilon_\infty$		
B_ED	$q \leq 2$	$k \leq h/\sqrt{2}c_\infty$
B_EPJ	$q < 4$	$k < h/c_\infty$
BJ_EP	$q < 2$, $\omega < 2/(2\eta_s - 1)$	$k < \min(h/\sqrt{2}c_\infty, 2/\omega_1\sqrt{2\eta_s - 1})$
undamped: $\nu = 0$, and $\varepsilon_s = \varepsilon_\infty$		
B_ED	$q < 2\omega/(1+\omega)$	condition on h
B_EPJ	$q < 4$	$k < h/c_\infty$
BJ_EP	$q < 2$, $\omega < 2/(2\eta_s - 1)$	$k < \min(h/\sqrt{2}c_\infty, 2/\omega_1\sqrt{2\eta_s - 1})$ if $\omega > 1$; condition on h else

In [15] there is a totally different material for which $\varepsilon_\infty = 1.5$, $\varepsilon_s = 3$, $\omega_1 = 2\pi \times 5 \times 10^{10}\,\mathrm{rad\,s^{-1}}$, and $\nu = 10^{10}\,\mathrm{rad\,s^{-1}}$ (these round values certainly refer to a model material). In this case $\omega \leq 2/\sqrt{2\eta_s - 1}$ comes to $k \leq 3.6 \times 10^{-12}\,\mathrm{s}$, which corresponds to $h = 1.9 \times 10^{-3}\,\mathrm{m}$ in the $q < 2$ condition. For this material condition $q < 2$ is the strongest for optical waves. The B_EPJ scheme is, however, more advisable, since it allows $q < 4$ instead of $q \leq 2$.

The results obtained in [10] were obtained for our first cited material and for the B_ED and BJ_EP schemes. Petropoulos observed instabilities for $\xi > \frac{\pi}{2}$. We note that if $\xi \leq \frac{\pi}{2}$, then $\sin(\xi/2) \leq 1/\sqrt{2}$ and $q \leq 2$ instead of $q \leq 4$. This is exactly our result. He found also the B_EPJ scheme to be stable for $q \leq 4$.

5. Automation *via* computer algebra. For a three-dimensional Lorentz medium, ϕ_0 is typically a 12th degree polynomial with polynomial coefficients of

degree 6 in the different parameters. The previous procedure becomes awful if made by hand. A computer algebra environment based on MAPLE has been developed specifically to automate all the computational steps which may be the source of errors [2]. It is still dedicated only to electromagnetic models but could be extended in the future to other applications.

The schemes are defined by four parameters:

1. the space dimension `Dim` (1, 2 or 3),
2. the polarization `Polar` (TE or TM in dimension 2),
3. the physical model `Model` (e.g., Debye),
4. the variables used `Formula` (e.g., `B_EP`).

Maxwell equations have been written once and for all and just have to be "called":

```
> Faraday(Eq, Dim, Polar):
> Ampere(Eq, Dim, Polar, Formula):
```

For our applications, in the other equations space is only a parameter. Such equations are written once with no spatial dependence:

```
> NewEq := tr*(P[n+1]-P[n])/dt
         + 1/2*(P[n+1]+P[n])
         - eps0*(epss-epsinfini)*(E[n+1]+E[n])/2:
> CreateEq(Eq, NewEq, Dim, Polar):
```

and propagated to all the useful coordinates with the right indexes on the staggered grid, according to the space dimension and polarization.

Then changes of variables are automatically performed to have dimensionless variables (specific to the model), no redundant variables (specific to the scheme), and an explicit scheme in the frequency domain. This yields the amplification matrix G. The computation of polynomial ϕ_0 is then performed as well as the computational part of the von Neumann analysis: computation of the sequence of polynomials and factorizations. On these forms the user of the toolbox can easily see which are the specific cases to consider separately.

The comparisons $|\phi_m(0)| \leq |\phi_m^*(0)|$ are the real difficult points from the computer algebra point of view; therefore it is necessary to evaluate the sign of a polynomial in many variables (4 for a Lorentz medium) and of total degree of order say 6 for ϕ_0, about 10 for $\phi_1, \ldots$ knowing some variables are positive (like $\eta_s - 1$ or δ) and others lie within an interval (like q). This is also automated:

```
> SignCheck(phi, Z, [0 < delta, 1 < etas, 0 < q, q < 4]),
```

but sometimes MAPLE does not yield a totally explicit answer. This might lead us to migrate the whole toolbox in a C code to make use of some existing softwares specific for the solving of interval arithmetic problems.

Finally, tools are defined to compare the number of eigenvectors and the multiplicity of eigenvalues in the degenerate cases.

6. Two- and three-dimensional results.

6.1. Equation setting. In a two-dimensional context where unknowns depend only on space variables x and y, the Maxwell system may be split in two decoupled systems corresponding to the transverse electric (TE) (B_x, B_y, E_z) and the transverse magnetic (TM) (B_z, E_x, E_y) polarizations. In the one-dimensional case, Maxwell–Debye equations were represented by three equations and Maxwell–Lorentz by four equations. In the TE polarization, one more Faraday equation is added; therefore, we have four equations for Maxwell–Debye and five equations for Maxwell–Lorentz. In the TM polarization for the Maxwell–Debye model, one Ampère equation and one Debye equation have to be added, leading to five equation systems. For the

Maxwell–Lorentz model, there is one more Ampère equation and two more Lorentz equations, so the system consists of seven equations.

In the three-dimensional context, equations do not decouple any more, and systems consist of nine equations for the Maxwell–Debye schemes and twelve equations for the Maxwell–Lorentz schemes.

The principle of the stability analysis is exactly the same, but we now have larger polynomials to study. We, however, found out that the one-dimensional polynomials (which we denote by $\phi_0^{1D}(Z)$) are a factor in two- and three-dimensional polynomials, which reduces the formal calculations. More precisely we now denote by h_x, h_y, and h_z the space steps in the x-, y-, and z-directions, respectively, and by q the quantity

$$q = q_x + q_y = 4c_\infty^2 \left(\frac{k^2}{h_x^2} \sin^2(\xi_x/2) + \frac{k^2}{h_y^2} \sin^2(\xi_y/2) \right)$$

or

$$q = q_x + q_y + q_z = 4c_\infty^2 \left(\frac{k^2}{h_x^2} \sin^2(\xi_x/2) + \frac{k^2}{h_y^2} \sin^2(\xi_y/2) + \frac{k^2}{h_z^2} \sin^2(\xi_z/2) \right)$$

according to the space dimension (recall $q = 4c_\infty^2 \frac{k^2}{h_x^2} \sin^2(\xi_x/2)$ in one dimension).

6.2. Results. In the two-dimensional TE polarization, the characteristic polynomial reads

$$\phi_0^{2D,TE}(Z) = [Z-1]\phi_0^{1D}(Z)$$

for all the Maxwell–Debye and Maxwell–Lorentz schemes we study here. This could be a problem if 1 is already a root of $\phi_0^{1D}(Z)$, i.e., if $q = 0$, but it happens that it is never a problem: minimal stable subspaces are always one-dimensional.

In the TM polarization, the same factorization occurs but the remaining polynomial is slightly more complicated, namely

$$\phi_0^{2D,TM}(Z) = [Z-1]\psi_0(Z)\phi_0^{1D}(Z),$$

where $\psi_0(Z)$ is equal to:

- Debye B_ED and B_EP: $[(1+\delta\eta_s)Z - (1-\delta\eta_s)]$,
- Debye BP_E: $[(1+\alpha)(1+\delta\alpha)Z - (1-\alpha)(1-\delta\alpha)]$,
- Lorentz B_ED: $[(1+\delta+\omega\eta_s)Z^2 - 2Z + (1-\delta+\omega\eta_s)]$,
- Lorentz B_EPJ: $[(1+\delta+\frac{1}{2}\omega\eta_s)Z^2 - (2-\omega\eta_s)Z + (1-\delta+\frac{1}{2}\omega\eta_s)]$,
- Lorentz BJ_EP: $[(1+\delta)Z^2 - 2(1-\omega\eta_s)Z + (1-\delta)]$.

As for the TE polarization the extra eigenvalue 1 is never a source of instability. The other extra eigenvalues always lie inside or on the unit circle (conjugate complex roots). The only problem is when modulus 1 eigenvalues are also eigenvalues of the one-dimensional polynomial. This occurs only for the Lorentz B_ED scheme when $\varepsilon_s = \varepsilon_\infty$, and $q = 2\omega/(1+\omega)$, which is a resonant value we have already pointed out in the undamped case for this scheme.

In the three-dimensional context, the characteristic polynomial reads

$$\phi_0^{3D}(Z) = [Z-1]^2\psi_0(Z)\left(\phi_0^{1D}(Z)\right)^2$$

for all the schemes we have studied. In addition to extra eigenvalues 1, we have to check cases when we found out that $\phi_0^{1D}(Z)$ is a von Neumann polynomial but not a

Schur polynomial. If $\varepsilon_s = \varepsilon_\infty$, then we systematically have instabilities. Other extra instable cases depend on the scheme: $q = 4$ for the Debye B_ED and B_EP schemes, $\delta = 1$ for the Debye BP_E scheme; $q = 2$ for the Lorentz D_ED and BJ_EP schemes, $\nu = 0$ for all Lorentz schemes.

We do not have any explanation for these factorizations. This is a property of the characteristic polynomial and not necessarily of the amplification matrix, i.e., this may occur even if variables are not decoupled. However, there is probably some underlying block-triangular structure, which is still to be found.

We shall not duplicate Tables 3.3 and 4.4 for two- and three-dimensional schemes. In the two-dimensional case, if $h_x = h_y = h_z \equiv h$, condition $q \leq 4$ becomes $k \leq h/(\sqrt{2}c_\infty)$ and condition $q \leq 2$ becomes $k \leq h/(2c_\infty)$ in the physical variables. Besides, the Lorentz B_ED model, which was leading to a lower bound on h in the undamped case, leads also to such a bound in the damped case. These are the only differences with Tables 3.3 and 4.4.

In short, three-dimensional conditions are also essentially the same as one-dimensional conditions but inequalities always have to be strict and limiting (nonphysical) cases $\varepsilon_s = \varepsilon_\infty$ and $\nu = 0$ cannot be dealt with properly with a three-dimensional model.

7. Conclusion. We have studied a class of FD–TD schemes for dispersive materials based on the Yee scheme for Maxwell equations and compared them from the stability point of view. This study was inspired by Petropoulos [10] who performs the same analysis but uses specific values for the physical and numerical constants, and uses numeric routines to locate eigenvalues of the amplification matrix. Here we have general results which yield the constraint on numerical constants (k and h) for any Debye or Lorentz material. Our results confirm those of Petropoulos.

For usual Debye media, both studied schemes are stable under the same condition as the Yee scheme for nanosecond delay materials. The B_ED and B_EP schemes overcome the BP_E scheme in terms of stability condition for picosecond delay materials. Among the studied schemes for Lorentz media, the B_EPJ scheme clearly ranks first as far as stability is concerned. Its stability condition is also that of the Yee scheme. However, to properly take into account the Lorentz model, a smaller time step may have to be chosen, independently of stability issues. The two examples do not help us to deduce a general strategy to locate variables in time, in order to treat other physical models.

The computer algebra system that has been developed to handle the three-dimensional schemes has a much wider range of application. First, it can be extended with very small effort to other electromagnetic linear models such as cold plasmas, collisionless warm plasmas, or magnetic ferrites. With some more work (no computation skill but knowledge of dimensionless variables), it can be practically extended to the analysis of any linear finite-difference scheme occurring in areas other than electromagnetism.

Acknowledgments. The author thanks Françoise Jung and Jean-Guillaume Dumas for their assistance in mastering the tricks of MAPLE.

REFERENCES

[1] B. BIDÉGARAY-FESQUET, *Von Neumann Stability Analysis of Finite Difference Schemes for Maxwell–Debye and Maxwell–Lorentz Equations*, English translation of Technical report 1077-M, LMC-IMAG, available online at http://hal.ccsd.cnrs.fr/ccsd-00005369 (2005).

[2] B. Bidégaray-Fesquet, *NAUtil: von Neumann Analysis Utilities*, Maple Toolbox, available online at http://ljk.imag.fr/membres/Brigitte.Bidegaray/NAUtil/ (2006).

[3] M. Feise, J. Schneider, and P. Bevelacqua, *Finite-difference and pseudospactral time–domain methods applied to backward-wave metamaterials*, IEEE Trans. Antennas Propag., 52 (2004), pp. 2955–2962.

[4] R. M. Joseph, S. C. Hagness, and A. Taflove, *Direct time integration of Maxwell's equations in linear dispersive media with absorption for scattering and propagation of femtosecond electromagnetic pulses*, Opt. Lett., 16 (1991), pp. 1412–1414.

[5] T. Kashiwa, N. Yoshida, and I. Fukai, *A treatment by the FD–TD method of the dispersive characteristics associated with orientation polarization*, IEICE Trans., E73 (1990), pp. 1326–1328.

[6] R. Liska and L. Drsda, *FIDE; a REDUCE package for automation of FInite difference method for solving pDE*, in Proceedings of the International Symposium on Symbolic and Algebraic Computation, Tokyo, Japan, 1990, ACM Press, 1990, pp. 169–176.

[7] R. Luebbers, F. Hunsberger, K. Kunz, R. Standler, and M. Schneider, *A frequency-dependent finite-difference time–domain formulation for dispersive materials*, IEEE Trans. Electromagn. Compat., 32 (1990), pp. 222–227.

[8] R. Luebbers, D. Steich, and K. Kunz, *FDTD calculation of scattering from frequency-dependent materials*, IEEE Trans. Antennas Propag., 41 (1993), pp. 1249–1257.

[9] J. J. H. Miller, *On the location of zeros of certain classes of polynomials with applications to numerical analysis*, IMA J. Appl. Math., 8 (1971), pp. 397–406.

[10] P. Petropoulos, *Stability and phase error analysis of FD–TD in dispersive dielectrics*, IEEE Trans. Antennas Propag., 42 (1994), pp. 62–69.

[11] S. Steinberg and R. Liska, *Stability analysis by quatifier elimination*, Math. Comput. Simulation, 42 (1996), pp. 629–638.

[12] N. Stoykov, T. Kuiken, M. Lowery, and A. Taflove, *Finite-element time–domain algorithms for modeling linear Debye and Lorentz dielectric dispersions at low frequencies*, IEEE Trans. Biomed. Eng., 50 (2003), pp. 1100–1107.

[13] J. Strikwerda, *Finite Difference Schemes and Partial Differential Equations*, Wadsworth & Brooks/Cole, Wadsworth, CA, 1989.

[14] K. Yee, *Numerical solution of initial boundary value problems involving Maxwell's equations in isotropic media*, IEEE Trans. Antennas Propag., 14 (1966), pp. 302–307.

[15] J. L. Young, *Propagation in linear dispersive media: Finite difference time–domain methodologies*, IEEE Trans. Antennas Propag., 43 (1995), pp. 422–426.

[16] J. L. Young, *Finite difference time–domain methodologies for electromagnetic wave propagation in complex media*, in Frontiers in Electromagnetics, D. H. Werner and R. Mittra, eds., IEEE Press, Piscataway, NJ, 2000, pp. 666–707.

[17] J. L. Young, A. Kittichartphayak, Y. Kwok, and D. Sullivan, *On the dispersion errors related to FD^2 TD type schemes*, IEEE Trans. Antennas Propag., 43 (1995), pp. 1902–1910.

SIAM J. NUMER. ANAL.
Vol. 46, No. 5, pp. 2567–2592
© 2008 Society for Industrial and Applied Mathematics

JACOBI CORRECTION EQUATION, LINE SEARCH, AND CONJUGATE GRADIENTS IN HERMITIAN EIGENVALUE COMPUTATION I: COMPUTING AN EXTREME EIGENVALUE*

E. E. OVTCHINNIKOV†

Abstract. This paper is concerned with the convergence properties of iterative algorithms of conjugate gradient type applied to the computation of an extreme eigenvalue of a Hermitian operator via the optimization of the Rayleigh quotient functional. The algorithms in focus employ the line search for the extremum of the Rayleigh quotient functional in the direction that is a linear combination of the gradient and the previous search direction. An asymptotically explicit equation for the reduction in the eigenvalue error after the line search step as a function of the search direction is derived and applied to the analysis of the local convergence properties (i.e., the error reduction after an iteration) of the algorithms in question. The local (stepwise) asymptotic equivalence of various conjugate gradient algorithms and the generalized Davidson method is proved, and a new convergence estimate for conjugate gradient iterations is derived, showing the reduction in the eigenvalue error after any two consecutive iterations. The paper's analysis extensively employs remarkable properties of the operator of the so-called Jacobi orthogonal complement correction equation.

Key words. Hermitian eigenvalue computation, convergence estimates, conjugate gradients, generalized Davidson method, Jacobi orthogonal complement correction equation

AMS subject classifications. 65F15, 65N12, 65N22, 65N30

DOI. 10.1137/070688742

1. Introduction.

The conjugate gradient (CG) method was originally introduced as a method for solving linear systems [7]. It was later reinterpreted as a functional minimization method and is widely used as such in the smooth nonlinear optimization; see, e.g., [6, 28, 3, 21]. In its most familiar form, CG uses the following two-term recurrent formulas to generate a sequence of approximations x^i to the minimum point x of a functional $\phi(v)$:

$$x^{i+1} = x^i - \alpha_i y^i, \tag{1.1}$$

$$y^{i+1} = g^{i+1} + \beta_i y^i, \quad i = 0, 1, \ldots, \quad y^0 = g^0, \tag{1.2}$$

where g^i is the gradient of $\phi(v)$ at $v = x^i$. The parameter α_i in (1.1) minimizes $\phi(x^{i+1})$; the role of the parameter β_i is discussed below. We observe that (1.1) represents the line search for the minimum of $\phi(v)$ in the direction of $-y^i$; hence, we refer to the latter as the search direction (vector). Step (1.2) is referred to as the conjugation of g^{i+1} to y^i, or else as the conjugation of the search directions y^{i+1} and y^i.

For $\beta_i = 0$, the iterative scheme (1.1)–(1.2) becomes that of the steepest descent method, and the conjugation can be viewed as a way to improve the search direction. The case of the quadratic functional $\phi(v) = (Lv, v) - (f, v) - (v, f)$, where L is a Hermitian positive definite operator, is unique in that for each i there exists the optimal β_i, i.e., such that x^{i+2} is the minimum of $\phi(v)$ in the affine space $x^0 +$

*Received by the editors April 19, 2007; accepted for publication (in revised form) March 14, 2008; published electronically June 18, 2008.
http://www.siam.org/journals/sinum/46-5/68874.html
†Harrow School of Computer Science, University of Westminster, London HA1 3TP, UK (ovtchie@westminster.co.uk).

$\mathrm{span}\{y^0,\dots,y^{i+1}\} = x^0 + \mathrm{span}\{g^0,\dots,g^{i+1}\} = x^0 + \mathrm{span}\{g^0, Lg^0,\dots,L^{i+1}g^0\}$ that contains all x^i, $i = 2,\dots,i+2$, that can be generated by (1.1)–(1.2) for arbitrary $\beta_1,\dots,\beta_i$. In the case of a general nonquadratic $\phi(v)$, no choice of β_i is known to deliver x^{i+2} that is optimal in this or a similar sense. Various ways to compute β_i considered in the literature (see, e.g., [6, 28, 3, 21, 20]) essentially reuse various mathematically equivalent expressions for β_i derived for the quadratic case, and the respective versions of CG are usually referred to as CG schemes.

The fact that the smallest and the largest eigenvalues of a Hermitian operator L acting in a Euclidean space are, respectively, the minimum and the maximum of the Rayleigh quotient functional $\lambda(v) = (Lv, v)/\|v\|^2$ opens the way to using CG for computing extreme eigenvalues. Since the case of the largest eigenvalue simply mirrors that of the smallest, in this paper we focus on the latter. Specific properties of the Rayleigh quotient functional allow certain modifications of the general CG scheme: the new vector iterate x^{i+1} is usually normalized, and the gradient g^i replaced by the residual vector $r^i = r(x^i) = Lx^i - \lambda(x^i)x^i$, which is collinear to g^i. The line search can be replaced with the minimization in the two-dimensional subspace $\mathrm{span}\{x^i,\ y^i\}$, for which the Rayleigh–Ritz procedure can be employed, or else with the minimization on the sphere $\|v\| = 1$ (see [4]). Certain precautions need to be taken with some CG schemes, however (cf. [2]), as the Hessian of $\lambda(v)$ is not positive definite.

It should be noted that a number of algorithms for computing eigenvalues employ an inner-outer iteration approach whereby CG is used for solving certain auxiliary linear systems (see, e.g., [24, 14, 26] or [1]). Such algorithms are not discussed in this paper, which is entirely focused on the one-level approach just described.

A number of CG schemes for eigenvalue computation are discussed in [5, 29] and [4]; the first two papers also present numerical comparisons between different schemes. A general conclusion at which the cited papers arrive is that most schemes are similar in convergence, one exception being the Perdon–Gambolati scheme that uses L-conjugation of search directions (cf. (3.4) in section 3) and is shown by [5] to be less efficient than the rest. A plausible explanation of the observed similarity of convergence behavior is the asymptotic closeness of the schemes in question pointed out by [5] (based on informal arguments) and, in the framework of the minimization on the sphere $\|v\| = 1$, by [4]. The present paper complements the comparative analysis of [5] and [4] by studying local convergence properties of CG schemes for eigenvalue computation and providing explicit upper bounds for the difference between new approximate eigenvalues $\lambda^{i+1} = \lambda(x^{i+1})$ computed by various CG schemes based on the same x^i and y^{i-1}.

In our local convergence analysis we use an approach based on the so-called *Jacobi orthogonal complement correction* equation of [27]. The operator of this equation maps the error vector into the residual vector, and, consequently, the associated energy norm of the former coincides with the respective dual energy norm of the latter; remarkably, for the normalized approximate eigenvector, both coincide with the eigenvalue error (see Lemma 2.5). In the paper we show how this and some other properties of the Jacobi correction operator can be used in the convergence analysis of methods employing the line search, in particular, the steepest descent and CG methods. We derive an equation that expresses the reduction in the eigenvalue error after the line search step as a function of the search direction vector, and we use this equation in our comparison of various CG schemes and the generalized Davidson method whereby the new vector iterate x^{i+1} minimizes the Rayleigh quotient in the subspace spanned by $x^0, g^0,\dots,g^i$. Further, we derive an asymptotic equation that relates the eigenvalue error reduction on two consecutive CG steps with that of the steepest descent step.

Based on this equation and available estimates for the steepest descent method, we derive a new convergence estimate for CG that explicitly shows the acceleration of convergence compared to the steepest descent method due to the conjugation of two consecutive search directions.

The sequel [19] to this paper shows how the Jacobi correction equation technique can be extended to the convergence analysis of a block version of the iterative scheme (1.1)–(1.2) that computes several eigenvalues simultaneously.

In this paper the eigenvalues λ_j of L are enumerated in increasing order, each one counted as many times as the multiplicity, and the eigenvectors are denoted by x_j. The first (leftmost) eigenvalue λ_1 is assumed to be simple. For a Hermitian positive definite A we denote by $(\cdot,\cdot)_A$ and $\|\cdot\|_A$ the associated energy scalar product and norm, i.e., $(u,v)_A = (Au,v)$ and $\|u\|_A^2 = (u,u)_A$. All of the results of this paper apply to the generalized eigenvalue problem after simple modifications (cf. the introduction in [18] for the relevant arguments; cf. also Remark 1 in section 2.1).

2. Jacobi correction operator.

2.1. Jacobi orthogonal complement correction equation and its operator. Correction equations, i.e., equations for the difference between an exact and an approximate eigenvector, have been employed by a number of authors in their design of algorithms for computing eigenvalues (see, e.g., [27, 14, 26]) or the comparative analysis of such algorithms (see, e.g., [30]). In the present paper we employ a particular kind of a correction equation that is solved approximately in the Jacobi–Davidson method of [27], and we start this section with the analysis of some elementary properties of the latter correction equation, or rather its operator, most of which do not appear to have been previously investigated to a comparable level of detail.

We recall that $\lambda(u)$ stands for the Rayleigh quotient of u, i.e., $\lambda(u) = (Lu,u)/\|u\|^2$. If u approximates an eigenvector, then $\lambda(u)$ is an approximation to the corresponding eigenvalue that is optimal in the sense that $\|Lu - \lambda(u)u\| \le \|Lu - \tilde{\lambda}u\|$ for any alternative approximate eigenvalue $\tilde{\lambda}$; i.e., $\lambda(u)$ minimizes the residual vector. In the paper we denote $r(u) = Lu - \lambda(u)u$ and refer to this minimal residual vector as simply the residual. We denote by P_u the orthogonal projector onto $\mathrm{span}\{u\}$, i.e., $P_u v = \|u\|^{-2}u(v,u)$, and we denote $R(u) = (I - P_u)LP_u$, where I is the identity operator. Owing to a simple relationship $P_u L P_u v = \|u\|^{-2}u(LP_u v,u) = \|u\|^{-4}u(v,u)(Lu,u) = \lambda(u)P_u$, the operator $R(u)$ is related to the residual vector $r(u)$ as follows: $R(u)v = LP_u v - \lambda(u)P_u v = \|u\|^{-2}(v,u)(Lu - \lambda(u)u) = \|u\|^{-2}(v,u)r(u)$, which implies $\|R(u)\| = \|r(u)\|$. It is referred to as the residual operator in this paper.

The correction equation in question can be derived in the following manner. Let $\tilde{x}_1$ be an approximation to an eigenvector x_1 corresponding to λ_1 and denote by e_1 the corresponding correction vector, i.e., $\tilde{x}_1 + e_1 = x_1$. Since $(L - \lambda_1 I)x_1 = 0$, we have

$$(L - \lambda_1 I)e_1 = -(L - \lambda_1 I)\tilde{x}_1.$$

Applying to both sides of the above equation the operator $\tilde{P}_1^{\perp} = I - P_{\tilde{x}_1}$, we obtain

$$\tilde{P}_1^{\perp}(L - \lambda_1 I)e_1 = -\tilde{P}_1^{\perp}(L - \lambda_1 I)\tilde{x}_1 = -\tilde{P}_1^{\perp}(L - \lambda(\tilde{x}_1)I)\tilde{x}_1 = -\tilde{P}_1^{\perp}r(\tilde{x}_1) = -r(\tilde{x}_1).$$

If $\tilde{x}_1$ is not orthogonal to x_1, we can normalize the latter so that $(\tilde{x}_1, x_1) = \|\tilde{x}_1\|^2$, in which case $(e_1, \tilde{x}_1) = 0$, i.e., $e_1 = \tilde{P}_1^{\perp}e_1$, and we arrive at what is called the *Jacobi orthogonal complement correction* (JOCC) equation in [27]:

$$J(\lambda_1, \tilde{x}_1)e_1 = -r(\tilde{x}_1). \tag{2.1}$$

Here and in what follows, we use the notation

$$J(\mu, u) = (I - P_u)(L - \mu)(I - P_u). \tag{2.2}$$

If we subtract $(\lambda(\tilde{x}_1) - \lambda_1 I)e_1$ from both sides (2.1), to arrive at

$$J(\lambda(\tilde{x}_1), \tilde{x}_1)e_1 = -r(\tilde{x}_1) - (\lambda(\tilde{x}_1) - \lambda_1)e_1, \tag{2.3}$$

and drop the second term in the right-hand side, we end up with the equation

$$J(\lambda(\tilde{x}_1), \tilde{x}_1)\tilde{e}_1 = -r(\tilde{x}_1) \tag{2.4}$$

that is solved approximately in the Jacobi–Davidson method of [27].

Remark 1. In the case of the generalized eigenvalue problem $Lx = \lambda Mx$ with Hermitian positive definite M, the definition of the operator $J(\mu, u)$ needs to be modified as follows:

$$J(\mu, u) = (I - P^*_{u,M})(L - \mu M)(I - P_{u,M}), \tag{2.5}$$

where $P_{u,M}$ is the orthogonal projector onto $\mathrm{span}\{u\}$ in the sense of the scalar product $(\cdot, \cdot)_M = (M\cdot, \cdot)$. To apply any of the results of this paper to the generalized case, one simply needs to replace the standard norm with $\|\cdot\|_M = \sqrt{(M\cdot, \cdot)}$ and use the Rayleigh quotient $\lambda(u) = (Lu, u)/(Mu, u)$ and the residual $r(u) = Lu - \lambda(u)Mu$.

Auxiliary results below reveal some elementary properties of the operator $J(\mu, y)$ that are used in the rest of the paper. Below we denote by $\mathrm{span}\{u\}^\perp$ the orthogonal complement to $\mathrm{span}\{u\}$, i.e., the subspace of vectors orthogonal to u. The orthogonal complement correction vector for a general nonzero u is denoted by $e(u)$, i.e.,

$$e(u) = \frac{\|u\|^2}{(x_1, u)} x_1 - u. \tag{2.6}$$

We start with a simple algebraic expression that relates $J(\mu, u)$ to the shifted operator $L - \mu$.

LEMMA 2.1. $J(\mu, u) = L - \mu I - R(u) - R(u)^* - (\lambda(u) - \mu)P_u$.

Proof. $L - \mu = (I - P_u + P_u)(L - \mu I)(I - P_u + P_u) = J(\mu, u) + (I - P_u)(L - \mu)P_u + P_u(L - \mu I)(I - P_u) + P_u(L - \mu I)P_u = J(\mu, u) + (I - P_u)LP_u + P_uL(I - P_u) + P_uLP_u - \mu P_u = L - \mu I - R(u) - R(u)^* - (\lambda(u) - \mu)P_u$. ∎

The second result shows that $J(\mu, u)$ is positive definite in the subspace $\mathrm{span}\{u\}^\perp$ provided that μ and $\lambda(u)$ are close enough to λ_1 (cf. Lemma 3.1 in [14]).

LEMMA 2.2. *For any* $v \in \mathrm{span}\{u\}^\perp$, $(J(\mu, u)v, v) \geq (\lambda_1 + \lambda_2 - \mu - \lambda(u))\|v\|^2$.

Proof. The scaling of u and v does not affect the inequality in question; hence, assume that $\|u\| = \|v\| = 1$. The Ritz values $\hat{\lambda}_1$ and $\hat{\lambda}_2$ of L in $\mathrm{span}\{u, v\}$ are the eigenvalues of the matrix

$$\hat{L} = \begin{bmatrix} (Lu, u) & (Lv, u) \\ (Lu, v) & (Lv, v) \end{bmatrix}. \tag{2.7}$$

We have $\hat{\lambda}_1 + \hat{\lambda}_2 = Tr(\hat{L}) = (Lu, u) + (Lv, v) = \lambda(u) + \lambda(v)$ and, by the minimax principle, $\hat{\lambda}_i \geq \lambda_i$. Hence, $\lambda(v) \geq \lambda_1 + \lambda_2 - \lambda(u)$. Since $(J(\mu, u)v, v) = (\lambda(v) - \mu)\|v\|^2$, we arrive at the inequality of the lemma. ∎

COROLLARY 2.3. $(J(\mu, u)v, v) \geq (\lambda_1 + \lambda_2 - \mu - \lambda(u))\|(I - P_u)v\|^2$.

The next result shows that the $J(\mu,u)$-energy seminorms are equivalent for various μ that are close enough to λ_1.

LEMMA 2.4. *For any* $\mu \le \nu < \lambda_1 + \lambda_2 - \lambda(u)$

$$\frac{\lambda_1+\lambda_2-\lambda(u)-\nu}{\lambda_1+\lambda_2-\lambda(u)-\mu}(J(\mu,u)v,v) \le (J(\nu,u)v,v) \le (J(\mu,u)v,v). \tag{2.8}$$

Proof. We have $J(\nu,u) = J(\mu,u) - (\nu-\mu)(I-P_u)$, and therefore $(J(\nu,u)v,v) = (J(\mu,u)v,v) - (\nu-\mu)\|(I-P_u)v\|^2$. Hence, using Corollary 2.3, we arrive at inequality (2.8). ∎

The last auxiliary result of this section relates to the operator $J(\lambda_1,u)$ that features in the JOCC equation (2.1); in what follows, we denote for brevity $J_1(u) = J(\lambda_1,u)$, and we call $J_1(u)$ the Jacobi correction operator. We observe that $J_1(u)$ is positive semidefinite; hence we can use the seminorm $\|\cdot\|_{J_1(u)} = \sqrt{(J_1(u)\cdot,\cdot)}$. We note also that the restriction of $J_1(u)$ to $\mathrm{span}\{u\}^\perp$ is equal to that of $L-\lambda_1$, and hence is positive definite if u is not orthogonal to x_1 since λ_1 is assumed to be simple.

LEMMA 2.5. *If* $(u,x_1) \ne 0$, *then* $(J_1(u)e(u),e(u)) = (J_1(u)^{-1}r(u),r(u)) = (\lambda(u)-\lambda_1)\|u\|^2$, *where* $J_1(u)^{-1}$ *is the inverse of the restriction of* $J_1(u)$ *to* $\mathrm{span}\{u\}^\perp$.

Proof. Denoting $\alpha = \|u\|^2/(x_1,u)$, we have

$$\begin{aligned}(J_1(u)e(u),e(u)) &= ((L-\lambda_1 I)(u-\alpha x_1),u-\alpha x_1)\\ &= ((L-\lambda_1 I)u,u) = (\lambda(u)-\lambda_1)\|u\|^2.\end{aligned}$$

To prove the second identity of the lemma, we observe that by (2.1) we have $J_1(u)e(u) = -r(u)$. As noted above, for u that is not orthogonal to x_1, $J_1(u)$ is positive definite, and hence invertible, in $\mathrm{span}\{u\}^\perp$. Thus, $e(u) = -J_1(u)^{-1}r(u)$, and we arrive at the identity in question. ∎

The remarkable simplicity of the relationship between the eigenvalue error and $J_1(u)$-based norms of the correction and residual vectors exposed by Lemma 2.5 suggests that these norms are "natural" ones in the a posteriori error analysis. The same can be said about the a priori error analysis. To give a simple illustration, let us obtain an estimate for the difference $\tilde e_1 - e_1$, where e_1 and $\tilde e_1$ are the solutions of the exact and approximate correction equations (2.3) and (2.4), respectively, assuming that $\lambda(\tilde x_1) < (\lambda_1+\lambda_2)/2$ and that $\|\tilde x_1\| = 1$ (cf. the related results in [15] that make use of a similar energy norm). Let us denote $J = J(\lambda_1,\tilde x_1)$ and $\tilde J = J(\lambda(\tilde x_1),\tilde x_1)$. From (2.3) and (2.4) we have $J(\tilde e_1 - e_1) = (\lambda(\tilde x_1)-\lambda_1)\tilde e_1$. By Lemma 2.2, the minimal eigenvalue of J restricted to the subspace $\mathrm{span}\{\tilde x_1\}^\perp$ (to which e_1, $\tilde e_1$ and their J- and $\tilde J$-images belong) is not less than $\lambda_2 - \lambda(\tilde x_1)$, which implies that

$$\|\tilde e_1 - e_1\|_J \le \frac{\lambda(\tilde x_1)-\lambda_1}{\lambda_2-\lambda(\tilde x_1)}\|\tilde e_1\|_J. \tag{2.9}$$

Using Lemma 2.4 and the relationship $(\tilde J\tilde e_1,\tilde e_1) = -(r(\tilde x_1),\tilde e_1) = (Je_1,\tilde e_1)$, we obtain

$$\|\tilde e_1\|_J^2 \le \frac{\lambda_2-\lambda(\tilde x_1)}{\lambda_1+\lambda_2-2\lambda(\tilde x_1)}\|\tilde e_1\|_{\tilde J}^2, \quad \|\tilde e_1\|_{\tilde J}^2 \le \|\tilde e_1\|_J\|e_1\|_J \le \frac{\lambda_2-\lambda(\tilde x_1)}{\lambda_1+\lambda_2-2\lambda(\tilde x_1)}\|e_1\|_J^2,$$

which, together with (2.9), yield a remarkably simple estimate

$$\|\tilde e_1 - e_1\|_J \le \frac{\lambda(\tilde x_1)-\lambda_1}{\lambda_1+\lambda_2-2\lambda(\tilde x_1)}\|e_1\|_J = \frac{\|e_1\|_J^3}{\lambda_1+\lambda_2-2\lambda(\tilde x_1)}. \tag{2.10}$$

In the remaining part of this section we demonstrate how the Jacobi correction operator can be used in the convergence analysis of methods employing the line search, in particular, the steepest descent and CG methods.

2.2. Jacobi correction operator and the line search. In this section we obtain some useful formulae for the reduction in the eigenvalue error achieved by the search for the local minimum of the Rayleigh quotient on a line.

Unlike in the case of a quadratic functional, the line search for the minimum of the Rayleigh quotient may fail in the sense that the minimum of $\lambda(w)$ on the line $w(\tau) = u + \tau v$ is not attained for any finite value of the parameter τ. It is not difficult to see that this happens if v is the Ritz vector of L in the subspace $\operatorname{span}\{u, v\}$ corresponding to the smaller Ritz value. Let us show first that the line search can be made fail-safe by orthogonalization of the search direction v to the current vector iterate u—provided that v is not orthogonal to the gradient of $\lambda(w)$ at $w = u$, i.e., not tangential to the level surface of the Rayleigh quotient.

Let u and v be two orthonormal vectors. The minimum of $\lambda(w)$ in $\operatorname{span}\{u, v\}$ is the leftmost Ritz value of L, i.e., the leftmost eigenvalue of the matrix $\hat{L}$ given by (2.7). Let $\tilde{x}_1 = \alpha u + \beta v$ be the corresponding normalized Ritz vector. We have $(Lv, u) = (v, Lu) = (v, Lu - \lambda(u)u) = (v, r(u))$. Hence, if v is not orthogonal to the gradient of $\lambda(w)$ at $w = u$, which is collinear to $r(u)$, then the off-diagonal element (Lv, u) of $\hat{L}$ is nonzero, and it is easy to see that neither α nor β can be zero. Thus, $\tilde{x}_1$ is collinear to $u + \tau_v v$, where $\tau_v = \beta/\alpha$.

We note that by the above calculation, the normalization of $u + \tau_v v$ produces $\tilde{x}_1$; i.e., the line search in the orthogonal direction with subsequent normalization (which is typically used in eigenvalue computation) is equivalent to the Rayleigh–Ritz procedure in $\operatorname{span}\{u, v\}$ as far as the leftmost Ritz value is concerned.

The next two results reveal an intimate link between the Jacobi correction operator and the line search technique for the Rayleigh quotient.

THEOREM 2.6. *Let u be a nonzero vector and let v be such that $(v, r(u)) \neq 0$. Let $u_v = u + \tau_v(I - P_u)v$, where P_u is the orthogonal projector onto* $\operatorname{span}\{u\}$ *and*

$$\tau_v \equiv \arg\min_{\tau} \lambda(u + \tau(I - P_u)v),$$

and denote $J_{u,v} = J(\lambda(u_v), u)$. The following relationships hold:

$$\tau_v = -\frac{(r(u), v)}{(J_{u,v}v, v)}, \tag{2.11}$$

$$\lambda(u_v) = \lambda(u) - \frac{|(r(u), v)|^2}{\|u\|^2 (J_{u,v}v, v)}, \tag{2.12}$$

$$(J_{u,v}(u - u_v), u - u_v) = (\lambda(u) - \lambda(u_v))\|u\|^2, \tag{2.13}$$

$$r(u_v) = r(u) + J_{u,v}(u_v - u). \tag{2.14}$$

If $\lambda(u) < (\lambda_1 + \lambda_2)/2$, then

$$\|r(u_v)\|^2_{J^{-1}_{u,v}} = \min_{\tau} \|r(u) + \tau J_{u,v}v\|^2_{J^{-1}_{u,v}} = \|r(u)\|^2_{J^{-1}_{u,v}} - (\lambda(u) - \lambda(u_v))\|u\|^2, \tag{2.15}$$

where $J^{-1}_{u,v}$ is the inverse of the restriction of $J_{u,v}$ to $\operatorname{span}\{u\}^{\perp}$.

Proof. Denote $v^\perp = (1 - P_u)v$. We have

$$r(u + \tau_v v^\perp) = r(u) + (\lambda(u) - \lambda(u + \tau_v v^\perp))u + (L - \lambda(u + \tau_v v^\perp)I)\tau_v v^\perp. \tag{2.16}$$

Since $\lambda(u + \tau_v v^\perp) = \min_\tau \lambda(u + \tau v^\perp)$, the gradient of $\lambda(w)$ at $w = u_v = u + \tau_v v^\perp$ is orthogonal to $v^\perp$; hence $(r(u_v), v^\perp) = 0$. Together with $(r(u_v), u_v) = 0$, this implies $(r(u_v), u) = 0$, i.e., $P_u r(u_v) = 0$. Applying $1 - P_u$ to both sides of (2.16), we obtain (2.14); multiplying (2.14) by v, we obtain (2.11). Multiplying $u_v = u + \tau_v v^\perp$ by $r(u)$, we obtain, on the one hand,

$$\begin{aligned}(u_v, r(u)) &= (u_v, Lu - \lambda(u)u) = (u_v, Lu - \lambda(u_v)u) + (\lambda(u_v) - \lambda(u))(u_v, u)\\ &= ((L - \lambda(u_v)I)u_v, u) + (\lambda(u_v) - \lambda(u))\|u\|^2 = (\lambda(u_v) - \lambda(u))\|u\|^2,\end{aligned}$$

and, on the other hand, $(u_v, r(u)) = (u + \tau_v v^\perp, r(u)) = \tau_v(v^\perp, r(u)) = \tau_v(v - P_u v, r(u)) = \tau_v(v, r(u))$, which, together with (2.11), yields (2.12) and (2.13). Turning to (2.15), we observe first that the assumption of the theorem made just before (2.15) implies that $J_{u,v}$ is positive definite in $\mathrm{span}\{u\}^\perp$ (see Lemma 2.2); hence it is invertible there. Owing to $0 = (r(u_v), v) = (r(u_v), J_{u,v}v)_{J_{u,v}^{-1}}$, we have $\|r(u) + \tau J_{u,v}v\|^2_{J_{u,v}^{-1}} = \|r(u_v)\|^2_{J_{u,v}^{-1}} + (\tau - \tau_v)^2\|v\|^2_{J_{u,v}}$, and we arrive at the first equation in (2.15). Setting $\tau = 0$ results in the second equation, owing to (2.13). ∎

We observe that the relationships of Theorem 2.6 are essentially the same as in the case of the standard line search minimization step for the quadratic functional $\phi(u) = (Lu, u) - (u, f) - (f, u)$, with L replaced by $J_{u,v}$ and ϕ with λ.

The next result provides an explicit (modulo asymptotically small term) expression for the relative reduction in the eigenvalue error as a function of the search direction.

THEOREM 2.7. *Let u be a nonzero vector nonorthogonal to x_1, and let v be such that $(v, r(u)) \neq 0$. Let u_v be the point of minimum of $\lambda(w)$ on the line $w(\tau) = u + \tau(I - P_u)v$. The following relationship is valid:*

$$\frac{\lambda(u_v) - \lambda_1}{\lambda(u) - \lambda_1} = \sin^2\theta(v) - \epsilon(v)\cos^2\theta(v), \tag{2.17}$$

where $\theta(v)$ is the angle between v and the correction vector $e(u)$ given by (2.6) *measured in the semiscalar product $(\cdot,\cdot)_{J_1(u)}$ associated with the Jacobi correction operator $J_1(u) = J(\lambda_1, u)$; i.e.,*

$$\cos\theta(v) = \frac{|(e(u), v)_{J_1(u)}|}{\|e(u)\|_{J_1(u)}\|v\|_{J_1(u)}}, \tag{2.18}$$

and

$$\epsilon(v) = \frac{\lambda(u_v) - \lambda_1}{\lambda(u) - \lambda(u_v)}\frac{\|u_v - u\|^2}{\|u\|^2}. \tag{2.19}$$

Proof. We start by noting that $(v, r(u)) \neq 0$ implies that v is not collinear to u; hence, $\|v\|_{J_1(u)} > 0$.

Using (2.12), (2.1), (2.18), and Lemma 2.5, we obtain

$$\begin{aligned}\lambda(u) - \lambda(u_v) &= \frac{|(r(u), v)|^2}{\|u\|^2(J_{u,v}v, v)} = \frac{\|v\|^2_{J_1(u)}}{(J_{u,v}v, v)}\frac{|(e(u), v)_{J_1(u)}|^2}{\|v\|^2_{J_1(u)}\|e(u)\|^2_{J_1(u)}}\frac{\|e(u)\|^2_{J_1(u)}}{\|u\|^2}\\ &= (1 + \epsilon(v))\cos^2\theta(v)(\lambda(u) - \lambda_1),\end{aligned}$$

where $\epsilon(v) = \|v\|^2_{J_1(u)}/(J_{u,v}v, v) - 1$. Hence, trivial calculation yields (2.17). To obtain (2.19), we observe that $J(\lambda_1, u) - J_{u,v} = (\lambda(u_v) - \lambda_1)(I - P_u)$ and hence $\|v\|^2_{J_1(u)} - (J_{u,v}v, v) = (\lambda(u_v) - \lambda_1)\|(I - P_u)v\|^2$. Therefore,

$$\epsilon(v) = \frac{(\lambda(u_v) - \lambda_1)\|(I - P_u)v\|^2}{(J_{u,v}v, v)} = \frac{(\lambda(u_v) - \lambda_1)\|u_v - u\|^2}{(J_{u,v}(u_v - u), u_v - u)},$$

and, taking into account (2.13), we obtain (2.19). □

We note that the second term in the right-hand side of (2.17) is asymptotically (as $u \to x_1$) smaller than the first in the following sense: if $\lambda(u) < (\lambda_1 + \lambda_2)/2$, then, applying Lemma 2.4 to the ratio $\|v\|^2_{J_1(u)}/(J_{u,v}v, v) = \epsilon(v) + 1$, one can estimate $\epsilon(v)$ as follows:

$$\epsilon(v) \le \frac{\lambda_1 + \lambda_2 - \lambda(u) - \lambda_1}{\lambda_1 + \lambda_2 - \lambda(u) - \lambda(u_v)} - 1 = \frac{\lambda(u_v) - \lambda_1}{\lambda_1 + \lambda_2 - \lambda(u) - \lambda(u_v)}$$

$$\le \frac{\lambda(u) - \lambda_1}{\lambda_1 + \lambda_2 - 2\lambda(u)} \frac{\lambda(u_v) - \lambda_1}{\lambda(u) - \lambda_1} \le \frac{\lambda(u) - \lambda_1}{\lambda_1 + \lambda_2 - 2\lambda(u)} \sin^2 \theta(v).$$

We observe that the factor in front of $\sin^2 \theta(v)$ is asymptotically close to the eigenvalue error ratio $(\lambda(u) - \lambda_1)/(\lambda_2 - \lambda_1)$.

2.3. Jacobi correction operator and the steepest descent. For the sake of illustration, in this section we apply Theorem 2.7 to the steepest descent method, and in the next to its preconditioned variant.[1]

The steepest descent method employs the line search in the direction that is opposite to that of the gradient of the Rayleigh quotient which, in turn, is collinear to the residual vector. Hence, by Theorem 2.7, the relative reduction in the eigenvalue error after a steepest descent step is asymptotically equal to the square of the sine of the $J_1(u)$-angle between the error $e(u)$ and $v = -r(u) = J_1(u)e(u)$ (cf. (2.1)). Applying Lemma A.1 from Appendix A to the operator $J_1(u)$ acting in the space $\mathrm{span}\{u\}^\perp$ endowed with the scalar product $(\cdot,\cdot)_{J_1(u)} = (J_1(u)\cdot, \cdot)$, we arrive at the estimate

$$\sin \angle_{J_1(u)}\{v, e(u)\} = \sin \angle_{J_1(u)}\{J_1(u)e(u), e(u)\} \le \frac{cond_\perp(J_1(u)) - 1}{cond_\perp(J_1(u)) + 1}, \tag{2.20}$$

where $\angle_{J_1(u)}\{v, w\}$ stands for the angle between v and w measured in $J_1(u)$-scalar product, and $cond_\perp(J_1(u))$ is the condition number of $J_1(u)$ restricted to $\mathrm{span}\{u\}^\perp$. Since the latter is the invariant subspace of $J_1(u)$ corresponding to positive eigenvalues, the said condition number is equal to the ratio $\iota_{-1}(u)/\iota_2(u)$ of the largest to the second eigenvalue of $J_1(u)$. As $u \to x_1$, $J_1(u) \to L - \lambda_1$ and $\iota_j(u) \to \lambda_j - \lambda_1$. Hence, the relative eigenvalue error reduction after a steepest descent step is asymptotically estimated by the square of the ratio

$$\gamma_1 = \frac{1 - \xi_1}{1 + \xi_1}, \quad \xi_1 = \frac{\lambda_2 - \lambda_1}{\lambda_{-1} - \lambda_1} \tag{2.21}$$

(cf. [9]), where λ_{-1} is the rightmost eigenvalue of L.

[1] These two sections may be skipped by a reader who is familiar with the preconditioned steepest descent and is not interested in its theoretical aspects.

2.4. Jacobi correction operator and preconditioning. In many applications, the spread of the spectrum of L, i.e., $\lambda_{-1} - \lambda_1$, is very large compared to $\lambda_2 - \lambda_1$. The estimate of the previous section (which is asymptotically sharp owing to the sharpness of (2.20); cf. also [18]) indicates that the convergence of steepest descent iterations for such problems may be very slow, and, indeed, this is the case in practical computation, the same being true for a closely related CG method. In this section we briefly discuss a technique, known as *preconditioning*, that can be employed to accelerate the convergence, and apply the results of section 2.2 to the convergence analysis of the *preconditioned* steepest descent.

Preconditioning is currently widely used for solving the linear system $Lx = f$ with a poorly conditioned Hermitian positive definite operator L. In the latter context, preconditioning is commonly conceived of as the multiplication of the system at hand by a Hermitian positive definite operator K such that the condition number of KL is considerably smaller than that of L. Since KL is Hermitian positive definite in energy scalar products $(L\cdot,\cdot)$ and $(K^{-1}\cdot,\cdot)$, the preconditioned system $KLx = Kf$ can be solved by any iterative algorithm for solving Hermitian positive definite systems, in particular, by the steepest descent iterations $x^{i+1} = x^i - \tau_i(KLx^i - Kf^i) = x^i - \tau_i Kr^i$.

Alternatively, preconditioning can be interpreted as the use of a different scalar product for computing the gradient, notably, $(K^{-1}\cdot,\cdot)$.

Yet another interpretation is that preconditioning is a way to reduce the angle between the search and the error direction, the latter being obviously the optimal search direction. If τ_i minimizes the quadratic functional $\phi(w) = (Lw, w) - (f, w) - (w, f)$ on the line $w(\tau) = x^i - \tau Kr^i$, then the gradient of $\phi(w)$ at $w = x^{i+1}$ is orthogonal to the search direction $y^i = Kr^i$, i.e., $(r^{i+1}, y^i) = 0$. Since $r^{i+1} = -Le^{i+1}$, where $e^{i+1} = x - x^{i+1}$ is the error vector, the two error vectors e^i, e^{i+1} and the search direction y^i form a triangle with the right L-angle between y^i and e^{i+1}. Hence, the ratio of the L-norms of the two errors is equal to the sine of the angle between $y^i = Kr^i = KLe^i$ and e^i measured in L-norm. On the intuitive level, one expects that if KL is close to a multiple of the identity, i.e., if $cond(KL)$ is close to 1, then the angle between y^i and e^i is small and the convergence is fast. Lemma A.1 supports such a conclusion by providing an explicit estimate for the sine of the L-angle between $y^i = Ae^i$, where $A = KL$, and e^i:

$$\sin \angle_L\{y^i, e^i\} = \sin \angle_L\{Ae^i, e^i\} \le \frac{cond(A) - 1}{cond(A) + 1}, \quad A = KL. \tag{2.22}$$

The same considerations apply to the steepest descent for the Rayleigh quotient minimization. Bearing in mind that the search direction v should preferably be orthogonal to u, let us use $v = -(I - P_u)Kr(u)$ in place of $v = -r(u)$ as the descent direction. We observe that $v = -(I - P_u)K(I - P_u)r(u) = -(I - P_u)K(I - P_u)J_1(u)e(u) = -A(u)e(u)$ and that in $\mathrm{span}\{u\}^\perp$ the operator $A(u) = (I - P_u)K(I - P_u)J_1(u)$ is Hermitian positive definite in the scalar product $(J_1(u)\cdot,\cdot)$. By Theorem 2.7, the reduction in the eigenvalue error is determined by the $J_1(u)$-angle between $A(u)e(u)$ and $e(u)$. Hence, the closer $A(u)$ is to a multiple of the identity, the greater the reduction, and one has one-step convergence if K is the inverse of $J_1(u)$ restricted to $\mathrm{span}\{u\}^\perp$; i.e., $v = Kr(u)$ solves the correction equation $J_1(u)v = r(u)$. For a general K, we have, by Lemma A.1, the estimate

$$\sin \angle_{J_1(u)}\{A(u)e(u), e(u)\} \le \frac{cond_\perp(A(u)) - 1}{cond_\perp(A(u)) + 1}, \tag{2.23}$$

where $cond_\perp(A(u))$ is the condition number of $A(u)$ restricted to $\mathrm{span}\{u\}^\perp$.

Together with Theorem 2.7, the estimate (2.23) yields the following estimate:

$$\frac{\min_\tau \lambda(u+\tau v)-\lambda_1}{\lambda(u)-\lambda_1} \le \left(\frac{cond_\perp(A(u))-1}{cond_\perp(A(u))+1}\right)^2. \tag{2.24}$$

The derivation of the above convergence result for the preconditioned steepest descent for the Rayleigh quotient minimization is remarkably simple, compared to much more complicated proofs of similar results that can be found in the literature; see, e.g., [8, 13, 12] and the relevant results in [16, 17, 18].[2] Admittedly, (2.24) is given in different terms. In order to reformulate this result in terms similar to, e.g., those used by the respective results in [8, 9], let us assume for simplicity of notation that L is positive definite.

We start by observing that $A(u) = ((I-P_u)K(I-P_u)+\kappa P_u)J_1(u) \equiv K_\kappa(u)J_1(u)$ for any scalar κ, and that $K_\kappa(u)$ is positive definite if κ is positive, which we assume from now on. Hence, the eigenvalues $\alpha_j(u)$ of $A(u)$ are the eigenvalues of the generalized problem $J_1(u)v = \alpha K_\kappa(u)^{-1}v$. As $u \to x_1$, $J_1(u) \to L - \lambda_1 I$ and $K_\kappa(u) \to K_\kappa \equiv (I-P_1)K(I-P_1)+\kappa P_1$, where $P_1 = P_{x_1}$. Denoting by κ_1 and κ_{-1} the minimal and maximal eigenvalues of KL, respectively, we have $\kappa_1 L^{-1} \le K \le \kappa_{-1}L^{-1}$, and, hence, assuming $\kappa_1 \le \kappa \le \kappa_{-1}$, $\kappa_1 L^{-1} \le K_\kappa \le \kappa_{-1}L^{-1}$. Thus, $\kappa_1(\lambda_j-\lambda_1)/\lambda_j \le \alpha_j(x_1) \le \kappa_{-1}(\lambda_j-\lambda_1)/\lambda_j$, and the relative error reduction is estimated by the square of γ_1 given by (2.21) with ξ_1 replaced by

$$\xi_1 = \frac{\kappa_1}{\kappa_{-1}}\frac{\lambda_2-\lambda_1}{\lambda_2}\frac{\lambda_{-1}}{\lambda_{-1}-\lambda_1} = \frac{1}{cond(KL)}\frac{\lambda_2-\lambda_1}{\lambda_2}\frac{\lambda_{-1}}{\lambda_{-1}-\lambda_1} \tag{2.25}$$

(cf. [9]), where $cond(KL)$ is the condition number of KL. We observe that a large λ_{-1} is no longer a problem and that we have a considerable improvement over the nonpreconditioned case if K is a good preconditioner for L; i.e., $cond(KL)$ is considerably smaller than $cond(L)$.

2.5. Jacobi correction operator and conjugate gradients. CG achieves significant acceleration of convergence compared to the steepest descent by making use of the previous search direction y^{i-1} alongside with the gradient direction (collinear to r^i or Kr^i). Here we show that the conjugation of search directions employed by CG can be viewed as another way of reducing the angle between the search direction $-y^i$ and the error direction e^i.

Turning to the quadratic functional $\phi(w) = (Lw,w)-(f,w)-(w,f)$ for guidance, we observe that the error $e^i = x - x^i$ is L-orthogonal to the previous search direction y^{i-1} owing to the fact that x^i is the minimum point of $\phi(w)$ on the line $w(\tau) = x^{i-1} - \tau y^{i-1}$ and hence $0 = (r^i, y^{i-1}) = -(Le^i, y^{i-1})$, where $r^i = Lx^i - f$. The L-orthogonalization of the gradient to y^{i-1} employed by CG projects the search direction onto the subspace L-orthogonal to y^{i-1}, where e^i belongs, and hence makes the L-angle between y^i and e^i as minimal as possible among all directions that can be formed by combining the gradient and y^{i-1}. It remains to note that the ratio of L-norms of e^{i+1} and e^i is equal to the sine of this angle, owing to L-orthogonality of e^{i+1} and y^i.

The same reasoning can be applied to CG in eigenvalue computation. Assume that x^i is not orthogonal to x_1 and $r^i = r(x^i) \ne 0$, and assume that the line search did not fail on the previous iteration; i.e., y^{i-1} is not collinear to x^i. Since x^i is the

[2] It should be noted, however, that [13, 12] and [18] present sharp estimates, whereas (2.24) is only asymptotically sharp.

minimum point of $\lambda(w)$ on the line $w(\tau) = x^{i-1} - \tau y^{i-1}$, we have $0 = (r^i, y^{i-1}) = (J_1(x^i)e(x^i), y^{i-1})$; i.e., $e(x^i)$ lies in the subspace that is $J_1(x^i)$-orthogonal to y^{i-1} (note that $J_1(x^i)y^{i-1}$ is nonzero since x^i is not orthogonal to x_1). We observe that y^i is not collinear to x^i for any β_{i-1}, owing to $(y^i, r^i) = (Kr^i, r^i) > 0$. Hence we are in a position to use the $J_1(x^i)$-seminorm for measuring the angle between y^i and $e(x^i)$, which, by Theorem 2.7, determines the reduction in the eigenvalue error. By simple geometrical arguments, this angle is minimal if y^i is $J_1(x^i)$-orthogonal to y^{i-1}; i.e.,

$$(J_1(x^i)y^i, y^{i-1}) = 0. \tag{2.26}$$

The analysis of section 3 suggests that in (2.26) the operator $J_1(x^i)$ can be replaced by $J(\mu^i, x^i)$ with $\mu^i \le \lambda(x^i)$ that is asymptotically close to λ_1, e.g., $\mu^i = \lambda^i$.

In the next section we compare the conjugation scheme just described, which we refer to as the *Jacobi* scheme, with other conjugation schemes that can be found in the literature, and show that most of them employ asymptotically equivalent conjugation.

3. Asymptotic equivalence of CG schemes.

3.1. CG schemes. We recall that the generic CG iterative scheme is given by (1.1)–(1.2), where g^i is the gradient of the minimized functional at $v = x^i$ and the line search parameter α_i minimizes this functional at x^{i+1}. In this section we analyze comparatively CG schemes for the minimization of the Rayleigh quotient functional that use the following formulas for the conjugation parameter β_i in (1.2):

$$\beta_i = \frac{(g^{i+1}, g^{i+1})}{(g^i, g^i)} \qquad \text{(Bradbury–Fletcher);} \tag{3.1}$$

$$\beta_i = \frac{(g^{i+1},\ g^{i+1} - g^i)}{(g^i, g^i)} \qquad \text{(Polak–Ribiére);} \tag{3.2}$$

$$\beta_i = -\frac{(H(x^{i+1})g^{i+1},\ y^i)}{(H(x^{i+1})y^i,\ y^i)} \qquad \text{(Daniel);} \tag{3.3}$$

$$\beta_i = -\frac{(Lg^{i+1},\ y^i)}{(Ly^i, y_i)} \qquad \text{(Perdon–Gambolati);} \tag{3.4}$$

$$\beta_i = -\frac{(J(\mu^{i+1}, x^{i+1})g^{i+1},\ y^i)}{(J(\mu^{i+1}, x^{i+1})y^i,\ y^i)} \qquad \text{(Jacobi)} \tag{3.5}$$

(cf., respectively, [2, 21, 3, 20] and section 2.5). In (3.3), $H(x^{i+1})$ is the Hessian of $\lambda(u)$ at $u = x^{i+1}$ or a multiple thereof, e.g.,

$$H(u) = (I - 2P_u)(L - \lambda(u))(I - 2P_u), \tag{3.6}$$

where P_u is the orthogonal projector onto $\operatorname{span}\{u\}$ (see, e.g., [30]). In (3.4), L is assumed to be positive definite. In (3.5), $\mu^{i+1} \le \lambda(x^{i+1})$ is any value that approximates λ_1, e.g., $\mu^{i+1} = \lambda(x^{i+1})$.

In yet another CG scheme suggested by Takahashi [28], β_i is chosen to minimize the functional at hand at $x^{i+2} = x^{i+1} - \alpha_{i+1}(g^{i+1} + \beta_i y^i)$. It is not difficult to see that in this case the new vector iterate x^{i+2} minimizes the functional in question in the two-dimensional affine space $x^{i+1} + \operatorname{span}\{y^i,\ y^{i+1}\} = x^{i+1} + \operatorname{span}\{y^i,\ g^{i+1}\}$, and, in the case of the Rayleigh quotient functional, in the three-dimensional subspace

$\mathrm{span}\{x^{i+1},\ g^{i+1},\ y^i\} = \mathrm{span}\{x^{i+1},\ g^{i+1},\ x^i\}$. The implied implementation of the Takahashi scheme is known as *locally optimal* CG ([23, 10, 11]).

In (3.3), (3.4), and (3.5), one can replace the gradients g^i with the residual vectors $r^i = r(x^i)$ or, if preconditioning is used, with preconditioned residuals Kr^i. The same can be done in (3.1) and (3.2), provided that the new vector iterate x^{i+1} is normalized after the line search step (1.1), in which case $g^i = 2r^i$ or $g^i = 2Kr^i$ (note, however, that unlike other schemes, these two are not scale-invariant because of the quadratic dependence of β_i on g^{i+1}; hence the normalization of x^{i+1} results in a nonequivalent scheme). In the preconditioned case, the K^{-1}-scalar product and norm must be used in (3.1) and (3.2), i.e., $\beta_i = (Kr^{i+1}, r^{i+1})/(Kr^i, r^i)$ and $\beta_i = (Kr^{i+1}, r^{i+1} - r^i)/(Kr^i, r^i)$, respectively. For the sake of generality, from now on we consider the preconditioned schemes.

In this section we compare the new approximate eigenvalues $\lambda^{i+2} = \lambda(x^{i+2})$ produced by the CG schemes in question based on given x^{i+1} and y^i assuming that these two vectors are not collinear, $r(x^{i+1})$ is nonzero, and $\lambda^i < (\lambda_1 + \lambda_2)/2$. To simplify the notation, we denote $u = x^{i+1}$, $\lambda = \lambda(x^{i+1})$, $\mu = \mu^{i+1}$, $r = r(x^{i+1})$, $v = y^{i+1}$, $\alpha = \alpha_i$, $u_p = x^i$, $v_p = y^i$, $r_p = r^i$, $\alpha_p = \alpha_{i-1}$. We observe that in all schemes $(r, v_p) = 0$, owing to the fact that u is the minimum point of $\lambda(w)$ in $\mathrm{span}\{u_p, v_p\}$, and hence $(r, v) = (r, Kr + \beta_i v_p) = (r, Kr) > 0$. Thus $(r, v) = (r, (1 - P_u)v) \neq 0$ for $r \neq 0$, and the line search in the direction $(1 - P_u)v$ never fails (cf. section 2.2). We observe also that u does not belong to $\mathrm{span}\{v, v_p\}$ (if $u = \sigma v + \tau v_p$, then the multiplication by $r(u)$ yields $0 = \sigma(Kr, r)$, and hence $u = x^{i+1}$ is collinear to $v_p = y^i$, which contradicts our assumption).

3.2. Locally optimal scheme. Let u_o and λ_o be the new eigenvector and eigenvalue iterates computed by the locally optimal (Takahashi) scheme; i.e., $\lambda_o = \lambda(u_o)$ is the minimum of $\lambda(w)$ in the subspace $\mathrm{span}\{u, Kr, v_p\}$. Since the new approximate eigenvector computed by any of the schemes discussed here belongs to this subspace, the corresponding approximate eigenvalue is not less than λ_o. In order to compare new approximate eigenvalues produced by different schemes, in this section we derive an upper bound for the difference $\lambda_v - \lambda_o$, where λ_v is the minimum of $\lambda(w)$ on the line $w(\tau) = u + \tau(1 - P_u)v$, for a general $v = Kr + \beta v_p$ as a function of v.

We start by showing that the locally optimal scheme implicitly employs a Jacobi conjugation scheme with $\mu = \lambda_o$, Since λ_o is not greater than the minimum of $\lambda(w)$ in $\mathrm{span}\{u, Kr\}$, which by (2.12) is strictly less than $\lambda = \lambda(u)$, we have $\lambda_o < \lambda$. This implies that $u_o = \sigma u + \tau Kr + \rho v_p$, where $\tau \neq 0$ (otherwise λ_o would be the minimum of $\lambda(w)$ in $\mathrm{span}\{u, v_p\}$, which is equal to λ). Let us show that $(u_o, u) \neq 0$. Multiplying u_o by r, we have, on the one hand, $(u_o, r) = \tau(r, Kr) \neq 0$, and, on the other hand, $(u_o, r) = (u_o, (L - \lambda)uI) = ((L - \lambda I)u_o, u) = (r(u_o), u) - (\lambda - \lambda_o)(u_o, u) = -(\lambda - \lambda_o)(u_o, u)$, owing to the fact that the gradient of $\lambda(w)$ at $w = u_o$ is orthogonal to the subspace $\mathrm{span}\{u, Kr, v_p\}$. Thus, u_o is not orthogonal to u and hence can be normalized so that $u_o - u$ is orthogonal to u. Assuming this normalization, we can write $u_o = u + (I - P_u)(-\alpha Kr - \gamma v_p) = u - \alpha(I - P_u)v_o, v_o = Kr + \beta_o v_p$. By (2.14), we have $r(u_o) = r(u) - \alpha J(\lambda_o, u)v_o$, and hence $(J(\lambda_o, u)v_o, v_p) = 0$. Thus, the locally optimal CG is a Jacobi scheme with $\mu = \lambda_o$, and it can be equivalently represented by the following two-term recurrent relationships for y^i and r^i:

$$y^{i+1} = Kr^i + \beta_i y^i, \qquad \beta_i :\ (J_i y^{i+1}, y^i) = 0, \tag{3.7}$$

$$r^{i+1} = r^i - \alpha_i J_i y^{i+1}, \quad \alpha_i :\ (r^{i+1}, y^i) = 0, \tag{3.8}$$

where $J_i = J(\lambda^{i+1}, x^i)$ and we shifted the superscript of y^i and subscript of β_i. We note that in (3.7), β_i minimizes the J_i-norm of the new search direction vector, and in (3.8), α_i minimizes the J_i^{-1}-norm of the new residual (see Theorem 2.6).

Remark 2. If in (3.7)–(3.8) we replace J_i with $J(\lambda^0, x^0)$, then (3.7)–(3.8) become the standard CG iterations with preconditioning (PCG) for solving the Jacobi correction equation $J(\lambda^0, x^0)h = -r^0$. If we replace J_i with $J_\infty = L - \lambda_1 I$, we arrive at PCG iterations for solving $(L - \lambda_1 I)h = 0$. (In both cases the meaning of r^i changes accordingly.)

Let us now estimate $\lambda_v - \lambda_o$, where λ_v is specified at the beginning of this subsection, assuming $\lambda < (\lambda_1 + \lambda_2)/2$. Denote $J_o = J(\lambda_o, u)$. We have observed at the beginning of this subsection that the new approximate eigenvector computed by any of the schemes discussed here belongs to the subspace $\mathcal{V} = \operatorname{span}\{u, Kr, v_p\}$. Let us denote by P the orthogonal projector onto $\mathcal{V}$ and by $\hat{L}$ the restriction of the operator PLP to $\mathcal{V}$. The leftmost eigenvalue of $\hat{L}$ is λ_o; hence, the Jacobi correction operator for $\hat{L}$ is $\hat{J}_1(u) = (I - P_u)(\hat{L} - \lambda_o)(I - P_u)$. Since $PP_u = P_u P = P_u$, this operator is related to J_o as $\hat{J}_1(u) = PJ_oP$. Hence, in $\mathcal{V}$ these two operators coincide. Applying Theorem 2.7, we have $(\lambda_v - \lambda_o)/(\lambda - \lambda_o) = \sin^2\theta_o(v) - \epsilon_o(v)\cos^2\theta_o(v)$, where

$$\cos\theta_o(v) = \frac{|(v_o, v)_{\hat{J}_1(u)}|}{\|v_o\|_{\hat{J}_1(u)}\|v\|_{\hat{J}_1(u)}} = \frac{|(v_o, v)_{J_o}|}{\|v_o\|_{J_o}\|v\|_{J_o}}, \quad \epsilon_o(v) = \frac{\lambda_v - \lambda_o}{\lambda - \lambda_v}\frac{\|u_v - u\|^2}{\|u\|^2}.$$

Since $\epsilon_o(v) \geq 0$ and v_o is J_o-orthogonal to v_p, the above relationship yields the desired estimate

$$\lambda_v - \lambda_o \leq (\lambda - \lambda_o)\cos^2\theta_p(v), \tag{3.9}$$

where $\theta_p(v)$ is the J_o-angle between v and v_p.

3.3. Jacobi scheme. Since the locally optimal scheme is a particular case of the Jacobi scheme, the comparative analysis of the two is relatively simple. Denoting by $v_{_J}$ the new search direction of the Jacobi scheme, we have $(J_o v_{_J}, v_p) = (J(\mu, u)v_{_J}, v_p) + (\mu - \lambda_o)((I - P_u)v_{_J}, v_p) = (\mu - \lambda_o)((I - P_u)v_{_J}, v_p)$. Hence, applying Corollary 2.3, we obtain

$$|(J_o v_{_J}, v_p)| \leq |\mu - \lambda_o|\|(I - P_u)v_{_J}\|\|(I - P_u)v_p\| \leq \frac{|\mu - \lambda_o|}{\lambda_1 + \lambda_2 - \lambda - \lambda_o}\|v_{_J}\|_{_{J_o}}\|v_p\|_{_{J_o}},$$

and thus

$$\lambda_{_J} - \lambda_o \leq \frac{(\lambda - \lambda_o)(\mu - \lambda_o)^2}{(\lambda_1 + \lambda_2 - \lambda - \lambda_o)^2} \leq \frac{(\lambda - \lambda_o)^3}{(\lambda_1 + \lambda_2 - \lambda - \lambda_o)^2}, \tag{3.10}$$

where $\lambda_{_J}$ is the new approximate eigenvalue produced by the Jacobi scheme.

3.4. Daniel scheme. The asymptotic closeness of the Daniel scheme to those just discussed is implied by the following simple relationship between $H(u)$ and J_o that is obtained in the same way as in Lemma 2.1:

$$H(u) = J(\lambda(u), u) - R(u) - R(u)^* = J_o - (\lambda - \lambda_o)(I - P_u) - R(u) - R(u)^*.$$

The search direction $v_{_D}$ of the Daniel scheme is $H(u)$-conjugate to v_p; i.e., $(H(u)v_{_D}, v_p) = 0$. Hence, $(J_o v_{_D}, v_p) = ((J(\lambda_o, u) - H(u))v_{_D}, v_p) = (\lambda - \lambda_o)((I - P_u)v_{_D}, v_p) + (R(u)v_{_D}, v_p) + (v_{_D}, R(u)v_p) = (\lambda - \lambda_o)((I - P_u)v_{_D}, v_p) + \|u\|^{-2}(u, v_p)(v_{_D}, r)$, where

we have used the fact that $R(u)v_{_D}$ is collinear to vector r that is orthogonal to v_p. Applying Corollary 2.3, we obtain

$$((I-P_u)v_{_D},v_p)| \le \|(I-P_u)v_{_D}\|\|(1-P_u)v_p\| \le \frac{\|v_{_D}\|_{J_o}\|v_p\|_{J_o}}{\lambda_1+\lambda_2-\lambda-\lambda_o}$$

and

$$|(u,v_p)| = \cos\{u;v_p\}\|u\|\|v_p\| = \frac{\cos\{u;v_p\}}{\sin\{u;v_p\}}\|u\|\|(I-P_u)v_p\| \le \frac{\cot\{u;v_p\}\|u\|\|v_p\|_{J_o}}{\sqrt{\lambda_1+\lambda_2-\lambda-\lambda_o}}.$$

Since $|(v_{_D},r)| \le \|v_{_D}\|_{J_o}\|r\|_{J_o^{-1}}$, it remains to estimate the J_o^{-1}-norm of r via the eigenvalue error, for which we employ first Lemma 2.4 and then Lemma 2.5 to arrive at

$$\|r\|^2_{J_o} \le \frac{\lambda_2-\lambda}{\lambda_1+\lambda_2-\lambda-\lambda_o}\|r\|^2_{J_1(u)} = \frac{(\lambda_2-\lambda)(\lambda-\lambda_1)}{\lambda_1+\lambda_2-\lambda-\lambda_o}\|u\|^2. \tag{3.11}$$

Put together, the above estimates provide an estimate for the J_o-cosine of the angle between $v_{_D}$ and v_p that, by (3.9), implies the following estimate for the difference between the new approximate eigenvalue $\lambda_{_D}$ produced by the Daniel scheme and λ_o:

$$\lambda_{_D}-\lambda_o \le (\lambda-\lambda_o)\left(\frac{\lambda-\lambda_o+\cot\{u;v_p\}\sqrt{(\lambda_2-\lambda)(\lambda-\lambda_1)}}{\lambda_1+\lambda_2-\lambda-\lambda_o}\right)^2. \tag{3.12}$$

The obtained upper bound is larger than that in (3.10) in the general case; however, if we modify the Daniel scheme by making v_p orthogonal to u, then (3.12) coincides with (3.10). The same is true for the Daniel scheme employed in the framework of the trace minimization CG of [4], whereby v_p is translated along the geodesics on the unit sphere and hence is made orthogonal to u. In fact, the Jacobi and Daniel schemes are identical in this framework because preconditioning is not used, and thus $v = Kr + \beta_i v_p = r + \beta_i v_p$ is also orthogonal to u; hence, $(H(u)v,v_p) = ((I-P_u)H(u)(I-P_u)v,v_p) = (J(u)v,v_p)$.

Yet another modification that makes the upper bound in (3.12) comparable to that in (3.10) is to orthogonalize v_p to u_p before the line search step (this modification has an additional benefit of making the line search step failsafe (see section 2.2), and we note also that it does not affect λ but generally affects $v_{_D}$). Indeed, estimating the standard norm of $u-u_p$ via the $J(\lambda,u_p)$-norm with the help of Corollary 2.3, and the latter norm via $\lambda_p-\lambda$ with the help of (2.13), we arrive at

$$\cot^2\{u;v_p\} = \cot^2\{u;u-u_p\} = \frac{\|u-u_p\|^2}{\|u_p\|^2} \le \frac{\lambda_p-\lambda}{\lambda_1+\lambda_2-\lambda-\lambda_p}. \tag{3.13}$$

We observe that the same estimate obviously holds in the case of yet another normalization that is commonly used in practical computation, notably, $\|u\| = \|u_p\| = 1$.

3.5. Polak–Ribiére scheme. The comparison with the Polak–Ribiére scheme is considerably more difficult. In order to simplify the calculations, let us employ one of the modifications of the standard CG scheme mentioned in the previous section, notably, the one where v_p is orthogonalized to u_p. In this case, by (2.14) we have

$$r = r_p - \alpha_p J_p v_p, \quad J_p = J(\lambda,u_p),$$

and we observed previously that $(v_p, r) = 0$ and $\|r_p\|_K^2 = (v_p, r_p)$. Hence $\|r_p\|_K^2 = (v_p, r_p - r)$, and we have for β_i computed by the Polak–Ribiére scheme the following relationship:

$$\beta_i = \frac{(r,\ r - r_p)_K}{\|r_p\|_K^2} = \frac{(Kr,\ r - r_p)}{(v_p, r_p - r)} = -\frac{(Kr, J_p v_p)}{(v_p, J_p v_p)}. \tag{3.14}$$

Thus, denoting by v_{PR} the new search direction, we have $(J_p v_{PR}, v_p) = 0$; i.e., in the scheme in question, the search directions are also conjugate with respect to a Jacobi correction operator, namely, $J_p = J(\lambda, u_p)$. The continuous dependence of the operator $J(\mu, u)$ on its arguments ensures that this scheme is also asymptotically close to the locally optimal scheme. In order to obtain an explicit estimate for the difference between the new approximate eigenvalue λ_{PR} computed by the Polak–Ribiére scheme and λ_o, we observe that

$$J_o - J_p = (\lambda - \lambda_o)I + R(u_p) - R(u) + R(u_p)^* - R(u)^* + (\lambda_p - \lambda)P_{u_p} - (\lambda - \lambda_o)P_u$$

(cf. Lemma 2.1). Since v_p is orthogonal to u_p, we have $P_{u_p} v_p = 0$ and $R(u_p)v_p = 0$, and since v_p is orthogonal to r, it is also orthogonal to $R(u)w = \|u\|^{-2}(w, u)r$ for any w. Hence, $((J_o - J_p)v_p, v_p) = (\lambda - \lambda_o)((I - P_u)v_p, v_p)$, and $(J_o v_p, v_p) = (\lambda - \lambda_o)\|(I - P_u)v_p\|^2 + (J_p v_p, v_p) \geq (J_p v_p, v_p)$. Since $(J_p v_{PR}, v_p) = 0$, we have $(J_o v_{PR}, v_p) = ((J_o - J_p)v_{PR}, v_p) = (\lambda - \lambda_o)((I - P_u)v_{PR}, v_p) + (R(u_p)v_{PR}, v_p) - (v_{PR}, R(u)v_p)$, where we have used the fact that $R(u_p)v_p = 0$ and $(R(u)v_{PR}, v_p) = 0$. The term $(\lambda - \lambda_o)((I - P_u)v_{PR}, v_p)$ has already been estimated in the previous section. For $(R(u_p)v_{PR}, v_p)$ we have

$$(R(u_p)v_{PR}, v_p) = \frac{(v_{PR}, u_p)(r_p, v_p)}{\|u_p\|^2} = \frac{(v_{PR}, (I - P_u)(u_p - u)) + (v_{PR}, P_u u_p)}{\|u_p\|^2}(r_p, v_p).$$

With the help of Corollary 2.3 and (3.13), the first term in the numerator is estimated by

$$\|(I - P_u)v_{PR}\|\|u_p - u\| \leq \frac{\|v_{PR}\|_{J_o}}{\sqrt{\lambda_1 + \lambda_2 - \lambda - \lambda_o}} \frac{\sqrt{\lambda_p - \lambda}\|u_p\|}{\sqrt{\lambda_1 + \lambda_2 - \lambda - \lambda_p}}.$$

The second term is estimated by

$$\|P_u v_{PR}\|\|u_p\| = \|(I - P_u)v_{PR}\|\|u_p\| \cot\{u; v_{PR}\} \leq \frac{\|v_{PR}\|_{J_o}\|u_p\| \cot\{u; v_{PR}\}}{\sqrt{\lambda_1 + \lambda_2 - \lambda - \lambda_o}},$$

and (r_p, v_p) is estimated as

$$|(r_p, v_p)| \leq \|r_p\|_{J_p^{-1}}\|v_p\|_{J_p} \leq \frac{\sqrt{(\lambda_2 - \lambda_p)(\lambda_p - \lambda_1)}\|u_p\|\|v_p\|_{J_o}}{\sqrt{\lambda_1 + \lambda_2 - \lambda - \lambda_p}},$$

where we have used the estimate for the J_p^{-1}-norm of r via $\lambda_p - \lambda_1$ similarly to (3.11). Finally $(v_{PR}, R(u)v_p) = \|u\|^{-2}(u, v_p)(v_{PR}, r)$, where (v_{PR}, r) is estimated in the same way as (v_D, r) in the previous section, and so is (u, v_p), with $\cot\{u; v_p\}$ estimated via (3.13), and we arrive at

$$\begin{aligned}\lambda_{PR} - \lambda_o \leq (\lambda - \lambda_o)\Bigg(&\frac{\lambda - \lambda_o}{\delta(\lambda + \lambda_o)} + \frac{\sqrt{(\lambda_2 - \lambda)(\lambda - \lambda_1)(\lambda_p - \lambda)}}{\delta(\lambda + \lambda_o)\sqrt{\delta(\lambda + \lambda_p)}} \\ &+ \frac{\sqrt{(\lambda_2 - \lambda_p)(\lambda_p - \lambda_1)}}{\sqrt{\delta(\lambda + \lambda_o)\delta(\lambda + \lambda_p)}}\left(\frac{\sqrt{\lambda_p - \lambda}}{\sqrt{\delta(\lambda + \lambda_p)}} + \cot\{u; v_{PR}\}\right)\Bigg)^2,\end{aligned} \tag{3.15}$$

where $\delta(\mu) = \lambda_1 + \lambda_2 - \mu$. The obtained upper bound is generally larger than that for the previous two schemes (assuming the orthogonality of v_p to u_p in the Daniel scheme) and cannot be reduced by orthogonalizing v_{PR} to u because this would generally destroy J_p-orthogonality of v_{PR} and v_p. We note, however, that if preconditioning is not used, then $(v_{PR}, u_p) = (r + \beta_i v_p, u_p) = 0$ and the third term in brackets in (3.15) disappears altogether, making this upper bound comparable to those for the other two schemes.

3.6. Bradbury–Fletcher scheme. Since u_o is the point of minimum of $\lambda(w)$ in the subspace $\operatorname{span}\{u, Kr, v_p\}$, we have $(r(u_o), Kr) = 0$. Hence, if v_p is the locally optimal search direction, then $(r, Kr_p) = 0$, and the Polak–Ribiére and Bradury–Fletcher schemes produce the same new search direction, which, as we have just shown, is asymptotically close to the new locally optimal one. Thus, in this particular sense, the estimate (3.15) applies to the Bradbury–Fletcher scheme as well.

We note that the asymptotic relationship between the Bradbury–Fletcher and Polak–Ribiére schemes without preconditioning has also been discussed in [4] based on the relationship $(r, r_p) = \mathcal{O}\left(\epsilon^3\right)$ assumed to be valid in the ϵ-vicinity of x_1. It is not difficult to show that for the scheme considered in the previous subsection and in the absence of preconditioning one actually has $(r, r_p) = \mathcal{O}\left(\epsilon^4\right)$. Indeed, by the calculations of section 3.5, the angle between the search directions of the locally optimal and Polak–Ribiére schemes is estimated by the expression inside the brackets in (3.15), which, in the absence of preconditioning, is of the order $\mathcal{O}\left(\lambda_p - \lambda_1\right) = \mathcal{O}\left(\epsilon^2\right)$. Hence, the difference between the new approximate eigenvectors u_o and u_{PR} computed by the locally optimal and Polak–Ribiére schemes is of the order $\mathcal{O}\left(\epsilon^2\right) \|u_o - u\| = \mathcal{O}\left(\epsilon^3\right)$, which implies $\|Pr(u_{PR})\| = \|Pr(u_o)\| + \mathcal{O}\left(\epsilon^3\right) = \mathcal{O}\left(\epsilon^3\right)$, where P is the orthogonal projector onto $\operatorname{span}\{v_p, r\}$. The J_p-orthogonality of $v_{PR} = r + \beta_i v_p$ to v_p implies that $\|\beta_i v_p\|_{J_p} \leq \|r\|_{J_p}$, and thus $\|\beta_i v_p\| = \mathcal{O}\left(\epsilon\right)$. Hence, $(r(u_{PR}), r) = (r(u_{PR}), v_{PR} - \beta_i v_p) = -(r(u_{PR}), \beta_i v_p) = -(Pr(u_{PR}), \beta_i v_p) = \mathcal{O}\left(\epsilon^4\right)$.

3.7. Perdon–Gambolati scheme. In the Perdon–Gambolati scheme, $(v, v_p)_L = 0$; i.e., the cosine of the L-angle between v and v_p (the angle measured via $\|\cdot\|_L$) is zero. In the general case, the cosine of the J_o-angle is not close to that of the L-angle, and hence the new search direction produced by the Pedon–Gambolati scheme may deviate significantly from the locally optimal one and therefore, by the calculations of this section, from those produced by other schemes in question. A similar comment is made by [4], and the numerical tests in [5] demonstrate that, indeed, the convergence of the Perdon–Gambolati CG is considerably slower than that of the Bradbury–Fletcher and Polak–Ribiére schemes.

4. New local convergence estimate for conjugate gradients. Owing to the inclusion $\operatorname{span}\{u, Kr, v_p\} \supset \operatorname{span}\{u, Kr\}$, the new approximate eigenvalue λ_o produced by the locally optimal CG step is not greater than that produced by the (preconditioned) steepest descent step for the same u. The calculations of the previous section imply that the same can be said, modulo asymptotically small terms, about the other schemes investigated there, save for the Perdon–Gambolati scheme. In practice, CG converges considerably faster than the steepest descent, but in eigenvalue computation this practical observation has not yet been supported by theoretical convergence estimates that would demonstrate convergence acceleration compared to the steepest descent. One way to obtain such estimates is to follow [4] in treating the Rayleigh quotient minimization as a minimization of a quadratic functional (Lv, v) constrained to the sphere $\|v\| = 1$. By introducing a suitable coordinate system on

this sphere, the latter minimization problem can be reformulated as a nonquadratic unconstrained problem in $(n-1)$-dimensional space, the main advantage over the original nonquadratic problem being the fact that the Hessian is positive definite, and hence the convergence results of, e.g., [3] or [22] are applicable. The described approach has a potential to produce convergence estimates for CG in eigenvalue computation similar to well-known estimates for linear systems and estimates of Kaniel and Saad for the Lanczos method. We note, however, that the mentioned results from nonquadratic CG minimization assume the initial vector x^0 to be close enough to the minimum point x and suggest upper bounds for the error reduction after i iterations of the form $q_i + c_i\|x^0 - x\|$, where $q_i < 1$ is related to the condition number of the Hessian at the minimum point in the same way as in the quadratic case and c_i grows rapidly with the iteration number i. Such results are meaningful only for the case where the iterations are frequently restarted, and after a sufficient number of restarts. We note also that the practical implementation of the approach of [4] to the generalized eigenvalue problem $Lx = \lambda Mx$ involves the application of M^{-1} to a vector (see section 4.5 of the cited paper), which hinders its application to generalized problems of large size, and it is not clear how to properly and efficiently employ preconditioning in the framework of this approach.

In this section we take a different approach that pursues a less ambitious goal of obtaining convergence results that would clearly demonstrate the advantages of conjugation—but would remain meaningful after any number of iterations. First, we derive an asymptotic equation that relates the reduction in the eigenvalue error achieved by the locally optimal CG to that which could be achieved by the locally optimal steepest descent for the same current eigenvector iterate x^i. Based on this equation, we then derive a convergence estimate for the general Jacobi scheme that explicitly shows the acceleration delivered by the conjugation of two consecutive search directions.

We start with a few simple observations regarding the equivalence of $J(\mu, u)$- and K^{-1}-energy norms. Below we assume that $\lambda_1 \le \mu \le \lambda(u) < (\lambda_1 + \lambda_2)/2$.

For any Hermitian positive definite K we have

$$(J(\mu, u)v, v) \le (J_1(u)v, v) \le \kappa_{-1}(u)(K^{-1}v, v), \tag{4.1}$$

where $J_1(u) = J(\lambda_1, u)$ and $\kappa_{-1}(u)$ is the rightmost eigenvalue of $KJ_1(u)$. By Lemma 2.2, the operator $J(u) = J(\lambda(u), u)$ is positive definite in $\mathrm{span}\{u\}^\perp$; hence for $\mu \le \lambda(u)$ and $v \in \mathrm{span}\{u\}^\perp$

$$(J(\mu, u)v, v) \ge (J(u)v, v) \ge \kappa_2(u)(K^{-1}v, v), \tag{4.2}$$

where $\kappa_2(u) > 0$ is second leftmost eigenvalue of $KJ(u)$. We note that as $u \to x_1$, $\kappa_2(u) \to \kappa_2$ and $\kappa_{-1}(u) \to \kappa_{-1}$, where κ_2 and κ_{-1} are the second leftmost and the rightmost eigenvalues of $K(L - \lambda_1)$.

Denoting $K(u) = (I - P_u)K(I - P_u) + \kappa P_u$, where $\kappa > 0$, and assuming $\mu + \lambda(u) < \lambda_1 + \lambda_2$, we observe that for $v \in \mathrm{span}\{u\}^\perp$

$$(J(\mu, u)v, v) \ge \kappa_2(\mu, u)(K(u)^{-1}v, v), \tag{4.3}$$

where $\kappa_2(\mu, u)$ is the second leftmost eigenvalue of $K(u)J(\mu, u)$. Lemma 2.2 implies that $\kappa_2(\mu, u) > 0$, and it is not difficult to see that $\kappa_2(\mu, u) \to \kappa_2$ as $\{\mu, u\} \to \{\lambda_1, x_1\}$, assuming $\kappa \ge \kappa_2$. Since $(J(\mu, u)Kr(u), Kr(u)) = (J(\mu, u)K(u)r(u), K(u)r(u))$ and $K(u)r(u) \in \mathrm{span}\{u\}^\perp$, by (4.3) we have

$$(J(\mu, u)Kr(u), Kr(u)) \ge \kappa_2(\mu, u)(r(u), K(u)r(u)) = \kappa_2(\mu, u)(r(u), Kr(u)). \tag{4.4}$$

Since $r(u) \in \operatorname{span}\{u\}^{\perp}$, we also have

$$(Kr(u), r(u)) = (K(u)r(u), r(u)) \le \kappa_{-1}(\mu, u)(J(\mu, u)^{-1}r(u), r(u)), \tag{4.5}$$

$$(Kr(u), r(u)) \ge \kappa_2(\mu, u)(J(\mu, u)^{-1}r(u), r(u)), \tag{4.6}$$

where $J(\mu, u)^{-1}$ is the inverse of the restriction of $J(\mu, u)$ to $\operatorname{span}\{u\}^{\perp}$ and $\kappa_{-1}(\mu, u)$ is the rightmost eigenvalue of $K(u)J(\mu, u)$.

We observe that $\kappa_{-1}(u)$, $\kappa_2(u)$, and $\kappa_2(\mu, u)$ continuously depend on u and μ. Hence, for $\lambda_1 \le \mu \le \lambda(u) \le \lambda^0 < (\lambda_1+\lambda_2)/2$ we have uniform (in μ and u) equivalence of the K^{-1}- and J-energy norms involved.

Below we return to the notation of section 3.1.

The first result of this section establishes an asymptotic relationship between the eigenvalue error reductions achieved on two consecutive locally optimal CG iterations and the error reduction achievable with the same data by the steepest descent.

THEOREM 4.1. *Assuming that $\lambda^{i-1} \le \nu < (\lambda_1 + \lambda_2)/2$ and $r^i \ne 0$, the following asymptotic relationship is valid:*

$$\frac{1}{\lambda^i - \lambda^{i+1}} + \frac{1}{\lambda^{i-1} - \lambda^i} = \frac{1 + \mathcal{O}(\epsilon_i)}{\lambda^i - \lambda_s^{i+1}}, \quad \epsilon_i = \sqrt{\lambda^{i-1} - \lambda_1}, \tag{4.7}$$

where $\lambda^k = \lambda(x^k)$ are computed by the locally optimal CG and λ_s^{i+1} by the steepest descent; i.e., λ_s^{i+1} minimizes $\lambda(w)$ in $\operatorname{span}\{x^i, g^i\}$.

Proof. Let us denote $J_i = J(\lambda^{i+1}, x^i)$ and $J_i^s = J(\lambda_s^{i+1}, x^i)$. In section 3.2 we have shown that the locally optimal scheme is a particular case of the Jacobi scheme with $\mu^i = \lambda^{i+1}$; i.e., the locally optimal scheme employs the line search in the direction $y^i = Kr^i + \beta_{i-1}y^{i-1}$, where β_{i-1} is such that $(J_i y^i, y^{i-1}) = 0$; i.e., $\beta_{i-1} = -(J_i Kr^i, y^{i-1})/(J_i y^{i-1}, y^{i-1})$. The scheme in question is scale-invariant; hence let us assume that $\|x^i\| = 1$ and $(x^{i+1} - x^i, x^i) = 0$.

By Theorem 2.6,

$$\lambda^i - \lambda^{i+1} = \frac{|(r^i, y^i)|^2}{\|y^i\|_{J_i}^2} = \frac{\|r^i\|_K^4}{\|y^i\|_{J_i}^2}, \quad \lambda^i - \lambda_s^{i+1} = \frac{\|r^i\|_K^4}{\|Kr^i\|_{J_i^s}^2}. \tag{4.8}$$

Thus,

$$\begin{aligned}\frac{\lambda^i - \lambda_s^{i+1}}{\lambda^i - \lambda^{i+1}} &= \frac{\|y^i\|_{J_i}^2}{\|Kr^i\|_{J_i^s}^2} = \frac{\|Kr^i\|_{J_i}^2 - |\beta_{i-1}|^2\|y^{i-1}\|_{J_i}^2}{\|Kr^i\|_{J_i^s}^2} \\ &= \frac{\|Kr^i\|_{J_i}^2}{\|Kr^i\|_{J_i^s}^2} - \frac{|(J_i Kr^i, y^{i-1})|^2}{\|Kr^i\|_{J_i^s}^2\|y^{i-1}\|_{J_i}^2} = \frac{\|Kr^i\|_{J_i}^2}{\|Kr^i\|_{J_i^s}^2} - \frac{|(J_i Kr^i, h^{i-1})|^2}{\|Kr^i\|_{J_i^s}^2\|h^{i-1}\|_{J_i}^2},\end{aligned} \tag{4.9}$$

where $h^{i-1} = x^i - x^{i-1}$. By Lemma 2.4, the ratio of the J_i- to the J_i^s- norm of any vector noncollinear to x^i is $1 + \mathcal{O}\left(\lambda_s^{i+1} - \lambda^{i+1}\right) = 1 + \mathcal{O}\left(\epsilon_i^2\right)$, and the continuity of $J(\mu, u)$ with respect to both arguments implies that in the last term in (4.9) J_i can be replaced with J_{i-1} at the cost of adding an $\mathcal{O}(\epsilon_i)$ term, leading to (cf. calculations in Appendix B)

$$\begin{aligned}\frac{|(J_i Kr^i, h^{i-1})|^2}{\|Kr^i\|_{J_i^s}^2\|h^{i-1}\|_{J_i}^2} &= \frac{|(J_{i-1} Kr^i, h^{i-1})|^2}{\|Kr^i\|_{J_i^s}^2\|h^{i-1}\|_{J_{i-1}}^2} + \mathcal{O}(\epsilon_i) \\ &= \frac{|(Kr^i, r^{i-1} - r^i)|^2}{\|Kr^i\|_{J_i^s}^2(\lambda^{i-1} - \lambda^i)} + \mathcal{O}(\epsilon_i),\end{aligned} \tag{4.10}$$

where we have used the relationships $r^i = r^{i-1} - J_{i-1}h^{i-1}$ and $\|h^{i-1}\|^2_{J_{i-1}} = (\lambda^{i-1} - \lambda^i)\|x^{i-1}\|^2$ (cf. (2.14) and (2.13)). A crucial observation is that $(Kr^i, r^{i-1}) = (r^i, Kr^{i-1}) = (r^i, y^{i-1} - \beta_{i-2}y^{i-2}) = 0$, and hence (4.8), (4.9), and (4.10) yield

$$\frac{\lambda^i - \lambda_s^{i+1}}{\lambda^i - \lambda^{i+1}} = 1 - \frac{\|r^i\|^4}{\|Kr^i\|^2_{J_i^s}(\lambda^{i-1} - \lambda^i)} + \mathcal{O}(\epsilon_i) = 1 - \frac{\lambda^i - \lambda_s^{i+1}}{\lambda^{i-1} - \lambda^i} + \mathcal{O}(\epsilon_i),$$

and we arrive at (4.7). ∎

We stress that in Theorem 4.1, $\mathcal{O}(\epsilon_i)$ stands for a quantity that is not greater than ϵ_i times a constant that does not depend on the iteration number.

The main merit of the result just obtained is that it explicitly and precisely (modulo an asymptotically small term) shows the convergence acceleration due to the conjugation of search direction. It strengthens considerably the trivial observation that the locally optimal CG is not worse than the steepest descent mentioned at the beginning of this section. Furthermore, (4.7) allows one to benefit from available convergence results for a relatively much better studied (as far as the eigenvalue computation is concerned) steepest descent method to arrive, in particular, at the following convergence estimate.

THEOREM 4.2. *Assuming $\lambda^{i-1} < (\lambda_1 + \lambda_2)/2$ and $\lambda_1 \le \mu^i \le \lambda^i$, the following asymptotic convergence estimate for Jacobi CG is valid:*

$$\lambda^{i+1} - \lambda_1 \le \left(\frac{q + \delta_i}{1 + \delta_i}\right)^2 (\lambda^{i-1} - \lambda_1), \quad q = T_2\left(\frac{\kappa + 1}{\kappa - 1}\right)^{-1}, \tag{4.11}$$

where $T_2(x) = 2x^2 - 1$, κ is the ratio of the rightmost to leftmost positive eigenvalue of $K(L - \lambda_1 I)$, and $\delta_i = \mathcal{O}\left(\sqrt{\lambda^{i-1} - \lambda_1}\right)$.

Proof. Consider first the locally optimal scheme, i.e., the particular case $\mu^i = \lambda^{i+1}$. Let us denote $q_i = (\lambda^{i+1} - \lambda_1)/(\lambda^i - \lambda_1)$ and $p_i = (\lambda_{sd}^{i+1} - \lambda_1)/(\lambda^i - \lambda_1)$. By the asymptotic estimate of [25] (cf. also section 2.4), $p_i \le p + \mathcal{O}\left(\lambda^{i-1} - \lambda_1\right)$, where $p = ((\kappa - 1)/(\kappa + 1))^2$. From (4.7) we have

$$\frac{1}{1 - q_i} + \frac{q_{i-1}}{1 - q_{i-1}} = \frac{1 + \mathcal{O}(\epsilon_i)}{1 - p_i} = \frac{1}{1 - p_i} + \varepsilon_i \le \frac{1}{1 - p} + \varepsilon_i,$$

where $\varepsilon_i = \mathcal{O}(\epsilon_i) = \mathcal{O}\left(\sqrt{\lambda^{i-1} - \lambda_1}\right)$. Subtracting 1 from both sides yields

$$\frac{q_i}{1 - q_i} + \frac{q_{i-1}}{1 - q_{i-1}} \le \frac{p}{1 - p} + \varepsilon_i, \tag{4.12}$$

or $u_i + u_{i-1} \le v_i$, where $u_i = q_i/(1 - q_i)$ and v_i is the right-hand side of (4.12). We have $q_i = u_i/(u_i + 1)$, and hence

$$q_i q_{i-1} = \frac{u_i u_{i-1}}{u_i u_{i-1} + v_i + 1} \le \frac{\frac{1}{4}v_i^2}{\frac{1}{4}v_i^2 + v_i + 1} = \frac{v_i^2}{(v_i + 2)^2}$$

$$= \frac{(p + (1 - p)\varepsilon_i)^2}{(2 - p + (1 - p)\varepsilon_i)^2} = \left(\frac{\frac{p}{2-p} + \delta_i}{1 + \delta_i}\right)^2.$$

Simple calculation shows that $(2-p)/p = T_2((\kappa+1)/(\kappa-1))$, and we arrive at (4.11).

To prove (4.11) in the general case, we need to revisit the proof of Theorem 4.1. Denoting $J_i^c = J(\mu^i, x^i)$, we observe that in the case at hand the search directions

are J_i^c-orthogonal, and hence $\beta_{i-1} = -(J_i^c Kr^i, y^{i-1})/(J_i^c y^{i-1}, y^{i-1})$, and in place of (4.9) we now have

$$\frac{\lambda^i - \lambda_s^{i+1}}{\lambda^i - \lambda^{i+1}} = \frac{\|y^i\|_{J_i}^2}{\|y^i\|_{J_i^c}^2} \frac{\|y^i\|_{J_i^c}^2}{\|Kr^i\|_{J_i^s}^2} = \frac{\|y^i\|_{J_i}^2}{\|y^i\|_{J_i^c}^2} \left(\frac{\|Kr^i\|_{J_i}^2}{\|Kr^i\|_{J_i^s}^2} - \frac{|(J_i^c Kr^i, h^{i-1})|^2}{\|Kr^i\|_{J_i^s}^2 \|h^{i-1}\|_{J_i^c}^2} \right).$$

The factor in front of the brackets and the first term inside them are of the order $1 + \mathcal{O}\left(|\mu^i - \lambda^{i+1}|\right)$. Regarding the second term in the brackets, it is not difficult to verify that in (4.10), J_i can be replaced with J_i^c. However, this time (Kr^i, r^{i-1}) is generally nonzero. Let us show that this scalar product is asymptotically insignificant.

We start by observing that the norm equivalencies noted at the beginning of this section together with the formula for $\lambda^i - \lambda_s^{i+1}$ in (4.8) imply $\lambda^i - \lambda_s^{i+1} \sim \|r\|_K^2 \sim \|Kr^i\|_{J_i^s}^2 \sim \lambda^i - \lambda_1 \sim \lambda^i - \lambda^{i+1}$. Next, we observe that

$$\begin{aligned}(r^i, Kr^{i-1}) &= (r^i, y^{i-1} - \beta_{i-2} y^{i-2}) = -(r^i, \beta_{i-2} y^{i-2}) \\ &= (-r^{i-1} + \alpha_{i-1} J(\lambda^i, x^{i-1}) y^{i-1}, \beta_{i-2} y^{i-2}) \\ &= (\mu^{i-1} - \lambda^i)((I - P_{x^{i-1}})\alpha_{i-1} y^{i-1}, (I - P_{x^{i-1}})\beta_{i-2} y^{i-2}),\end{aligned}$$

where we have used the $J(\mu^{i-1}, x^{i-1})$-orthogonality of y^{i-1} and y^{i-2}. By Corollary 2.3, the norms of $(I - P_{x^{i-1}})\alpha_{i-1} y^{i-1}$ and $(I - P_{x^{i-1}})\beta_{i-2} y^{i-2}$ are bounded by their $J(\lambda^i, x^{i-1})$-norms. By Theorem 2.6, the former norm is $\mathcal{O}(\sqrt{\lambda^{i-1} - \lambda^i})$, and the $J(\mu^{i-1}, x^{i-1})$-orthogonality of $y^{i-1} = Kr^{i-1} + \beta_{i-2} y^{i-2}$ and $\beta_{i-2} y^{i-2}$ implies that the latter norm is not greater than the same norm of Kr^{i-1}, which is equivalent to $\sqrt{\lambda^{i-1} - \lambda^i}$. Thus,

$$|(r^i, Kr^{i-1})| = \mathcal{O}\left(|\mu^{i-1} - \lambda^i|\right)(\lambda^{i-1} - \lambda^i) = \mathcal{O}\left((\lambda^i - \lambda_1)(\lambda^{i-1} - \lambda^i)\right).$$

Now, if $\|Kr^i\|_{J_i^s} < \lambda^{i-1} - \lambda_1$, then $\lambda^i - \lambda_1 = \mathcal{O}\left(\|r^i\|_K^2\right) = \mathcal{O}\left((\lambda^{i-1} - \lambda_1)^2\right)$, and we have (4.11); otherwise the ratio of $|(r^i, Kr^{i-1})|$ to $\|Kr^i\|_{J_i^s}\sqrt{\lambda^{i-1} - \lambda^i}$ is of the order $\mathcal{O}(\sqrt{\lambda^{i-1} - \lambda^i})$, and via (4.10) we again arrive at (4.7) and thus at (4.11). ∎

Remark 3. A similar result for the particular case $i = 1$ is reported in [10] (cf. also [16, 17]) under the assumption that the initial approximate eigenvalue λ^0 is close enough to λ_1, which makes it meaningful only for the first two iterations of the restarted CG after a sufficient number of restarts.

It is not difficult to verify that (4.7) applies to the case of CG-minimization of the quadratic functional $\phi(u) = (Lu, u) - (u, f) - (f, u)$, with λ^i replaced by $\phi(x^i)$ and $\epsilon_i = 0$, and hence so does (4.11) with $\delta_i = 0$ and κ standing for the condition number of KL. In the quadratic case, it can be shown that the number of CG iterations in (4.11) can be increased from 2 to arbitrary $k > 2$, with T_2 replaced with the Chebyshev polynomial of degree k (for $i = 1$ this is well known). It is plausible that a similar estimate is valid in the case at hand as well. Hence, one way to improve (4.11) is to try to extend the local convergence analysis of this section to more than two consecutive CG iterations.

An alternative way is to exploit the asymptotic stepwise closeness of the CG and the generalized Davidson method discussed in the next section.

5. Asymptotic local equivalence of CG and generalized Davidson methods. In the generalized Davidson method, the new eigenvector iterate x^{i+1} is defined as the minimum point of $\lambda(w)$ in the subspace $\mathcal{D}^i$ spanning the preconditioned residual (gradient) $g^i = Kr^i$ and all the previous iterates; i.e.,

$$x^{i+1} \in \mathcal{D}^i \ : \quad \lambda^{i+1} \equiv \lambda(x^{i+1}) = \min_{w \in \mathcal{D}^i} \lambda(w), \tag{5.1}$$

where $\mathcal{D}^i = \operatorname{span}\{x^0, \dots, x^i, g^i\} = \operatorname{span}\{g^0, \dots, g^i, x^i\}$. Since the computational expenses of the minimization in this $(i+2)$-dimensional subspace grow rapidly with i, in practical computation iterations are restarted when the dimension of $\mathcal{D}^i$ reaches a certain limit. In this section we show that after a sufficient number of restarts, the minimization in $\mathcal{D}^i$ becomes asymptotically equivalent to the line search employed by CG.

Our comparison of the generalized Davidson and CG below employs the same reasoning as that used earlier to compare the locally optimal CG with other CG schemes (cf. section 3.2). Let $P_{(i)}$ be the orthogonal projector onto $\mathcal{D}^i$, and let $\hat{L}_i$ be the restriction of the operator $P_{(i)}LP_{(i)}$ to $\mathcal{D}^i$. Since the leftmost eigenvalue of $\hat{L}_i$ is λ^{i+1}, the Jacobi correction operator for $\hat{L}_i$ is $\hat{J}_1(x^i) = (I - P_i)(\hat{L}_i - \lambda^{i+1}I)(I - P_i)$, where P_i is the orthogonal projector onto x^i. Since $P_iP = PP_i = P_i$, we have $\hat{J}_1(x^i) = P_{(i)}J_iP_{(i)}$, and thus $\hat{J}_1(x^i)$ and J_i coincide in $\mathcal{D}^i$. It remains to invoke Theorem 2.7 to arrive at the following estimate for the difference between λ^{i+1} and the approximate eigenvalue computed by the line search in an arbitrary direction y such that $(y, r^i) \neq 0$:

$$\lambda_y^{i+1} - \lambda^{i+1} \leq (\lambda^i - \lambda^{i+1}) \sin^2 \theta(y), \quad \lambda_y^{i+1} = \min_{\tau} \lambda(x^i - \tau y), \tag{5.2}$$

where $\theta(y)$ is the J_i-angle between y and h^i.

Let us now apply (5.2) to estimating $\lambda_o^{i+1} - \lambda^{i+1}$, where λ_o^{i+1} is computed by the locally optimal scheme based on x^i and x^{i-1}; i.e., λ_o^{i+1} is the minimum of $\lambda(w)$ in $\operatorname{span}\{x^{i-1}, x^i, Kr^i\}$. Below we assume that $\lambda^0 < (\lambda_1 + \lambda_2)/2$.

The generalized Davidson method is obviously scale-invariant, and, by the same arguments as with the locally optimal CG (see section 3.2), x^{i+1} is not orthogonal to x^i. Hence, we may assume $\|x^0\| = 1$ and assume the normalization of x^j for which $h^j = x^{j+1} - x^j$ is orthogonal to x^j. The vector x^{i+1} minimizes the Rayleigh quotient in $\mathcal{D}^i$ and hence on the line $x(\tau) = x^i + \tau h^i$ as well. By (2.14), this implies

$$r^{i+1} = r^i - J_i h^i, \quad J_i \equiv J(\lambda^{i+1}, x^i). \tag{5.3}$$

Since r^{i+1} is collinear to the gradient of the Rayleigh quotient at the point of minimum in $\mathcal{D}^i$, it is orthogonal to $x^0, \dots, x^{i+1}$ and hence to $h^0, \dots, h^i$. In view of (5.3), this implies

$$(J_i h^i, h^j) = 0, \quad j = 0, \dots, i-1. \tag{5.4}$$

The locally optimal search direction can be expressed as $y_o^i = Kr^i + \beta_o h^{i-1}$, where β_o is such that y_o^i and h^{i-1} are $J(\lambda_o^{i+1}, x^i)$-orthogonal (cf. section 3.2). We observe that $(y_o^i, r^i) = (Kr^i, r^i) > 0$; hence (5.2) applies. Let us denote $z^j = \|h^j\|_{J_j}^{-1} h^j$, $j = 0, \dots, i-1$. Employing $h^0, \dots, h^i, x^i$ as the basis of $\mathcal{D}^i$, we have

$$y_o^i = Kr^i + \beta_o h^{i-1} = \sum_{j=0}^{i-1} \alpha_j z^j + \alpha h^i + \gamma x^i = z_o^i + \alpha h^i + \gamma x^i, \tag{5.5}$$

and one observes that $\sin\theta(y_o^i) \le \|y_o^i - \alpha h^i\|_{J_i}/\|y_o^i\|_{J_i} = \|z_o^i\|_{J_i}/\|y_o^i\|_{J_i}$. In order to estimate the latter ratio, we multiply (5.5) by $J_i z^k$, $k = 0, \dots, i-1$, to arrive at the linear system $Au = v$, where A is an i-by-i matrix with entries $(J_i z^j, z^k)$ and the components of vectors u and v are α_k and $(J_i y_o^i, z^k) = (Kr^i + \beta_o h^{i-1}, J_i z^k)$, $k = 0, \dots, i-1$, respectively. Simple calculations (see section C) show that for $i > j \ge k$

$$|(J_i z^j, z^k) - (J_j z^j, z^k)| \le \sigma\sqrt{\epsilon_j}\sqrt{\epsilon_k}, \quad \epsilon_m = \frac{\lambda^m - \lambda_1}{\lambda_1 + \lambda_2 - 2\lambda^0}, \tag{5.6}$$

whereas calculations of section B adapted to the case at hand imply that for $j < i-1$

$$|(J_i K r^i, z^j)| \le \rho \|Kr^i\|_{J_i}\sqrt{\epsilon_j}, \tag{5.7}$$

where σ and ρ are constants independent of i. Together with (C.1), (5.6) implies $\|A - I\| = \mathcal{O}(\epsilon_0)$, where I is an i-by-i identity matrix, which implies $\|u\| = \|v\|(1 + \mathcal{O}(\epsilon_o))$. Since $\|z_o^i\|_{J_i}^2 = u^* A u = u^* v = \|v\|^2(1 + \mathcal{O}(\epsilon_0))$, it remains to estimate $\|v\|$ via $\|y_o^i\|_{J_i}$. For the v's first component $(J_i y_o^i, h^{i-1})$ we have an estimate similar to that in section 3.3:

$$|(J_i y_o^i, z^{i-1})| = |(\lambda_o^{i+1} - \lambda^{i+1})(y_o^i, z^{i-1})| \le (\lambda^i - \lambda_1)\|y_o^i\|\|(I - P_i)z^{i-1}\| \le \epsilon_i \|y_o^i\|_{J_i}.$$

For $j < i-1$ from (5.6) and (5.7) we obtain

$$|(J_i y_o^i, z^j)| = |(J_i K r^i, z^j) + \beta_o (J_i h^{i-1}, z^j)| \le (\rho\|Kr^i\|_{J_i} + \sigma|\beta_o| h^{i-1}\|_{J_i}\sqrt{\epsilon_{i-1}})\sqrt{\epsilon_j}.$$

Since $(Kr^i, h^{i-1})_{K^{-1}} = (r^i, h^{i-1}) = 0$, we have $\|Kr^i\|_{J_i} \sim \|Kr^i\|_{K^{-1}} \le \|y_o^i\|_{K^{-1}} \sim \|y_i^o\|_{J_i}$. Thus, the jth component of v is of the order $\mathcal{O}\left(\sqrt{\epsilon_j}\right)\|y_i^o\|_{J_i}$, and via (C.1) we arrive at $\sin\theta(y_o^i) = \mathcal{O}\left(\sqrt{\epsilon_0}\right)$, which by (5.2) implies

$$\lambda_o^{i+1} - \lambda^{i+1} = \mathcal{O}\left(\frac{\lambda^0 - \lambda_1}{\lambda_1 + \lambda_2 - 2\lambda^0}\right)(\lambda^i - \lambda^{i+1}). \tag{5.8}$$

The established asymptotic relationship between the CG and Davidson methods can be used as a starting point in the derivation of nonlocal convergence estimates for CG in eigenvalue computation. In [16, 17] the following convergence estimate for the restarted generalized Davidson method is obtained:

$$\lambda^k - \lambda^0 \le \left(T_k\left(\frac{\kappa_{-1} + \kappa_2}{\kappa_{-1} - \kappa_2}\right)^{-2} + c_k\sqrt{\lambda^0 - \lambda_1}\right)(\lambda^0 - \lambda_1), \tag{5.9}$$

where $T_k(x)$ is Chebyshev polynomial of degree k; i.e., $T_k(x) = \cos(k \arccos(x))$ for $x \in [-1, 1]$. We observe that for $k = 2$, (5.9) has the same asymptotics as (4.11) for $i = 1$. The calculations of this section suggest that it might be possible to obtain an estimate similar to (5.9) for the restarted CG. In this paper we refrain from using such an approach to the convergence analysis of CG because it is likely to produce an estimate of the form (5.9) with c_k rapidly growing in k (cf. convergence estimates of [3] and [22] and the discussion at the beginning of section 4).

6. Conclusion. The results of this paper amply demonstrate that the Jacobi correction operator is a useful and versatile tool for the analysis of iterative methods for eigenvalue computation employing the line search technique, in particular, steepest

descent and CG methods. It allows an analyst to relate the eigenvalue and eigenvector errors and residuals via identities rather than inequalities, and thus opens the way to establishing precise (modulo asymptotically small terms) relationships between different methods, such as (4.7) or the identity that yielded the estimate (3.9). It simplifies the convergence analysis of the line search methods and reveals inherent similarities between convergence acceleration techniques as different and mutually independent as preconditioning and conjugation. A challenging task is to extend the Jacobi correction operator technique to block iterative methods for computing eigenvalues; this task is faced by the second part [19] of the present paper. Yet another challenge, still to be faced, is to extend the two-step convergence analysis of the CG method presented here to several CG steps.

Appendix A. A sharp estimate for the angle between u and Au.

LEMMA A.1. *Let A be a Hermitian positive definite operator in a finite-dimensional linear space with the scalar product. The following inequality is valid for the angle $\angle\{u, Au\}$ between a nonzero u and Au:*

$$\sin\angle\{u, Au\} \le \frac{cond(A)-1}{cond(A)+1}, \tag{A.1}$$

where $cond(A)$ is the condition number of A, i.e., the ratio of the maximal to the minimal eigenvalue of A. The estimate (A.1) *is sharp; i.e., the right-hand side cannot be made smaller without additional assumptions on u.*

Proof. Denote by λ_1 and λ_{-1} the minimal and the maximal eigenvalue of A, respectively, and set $\tau_* = 2/(\lambda_1+\lambda_{-1})$. The norm of $1-\tau_* A$ is equal to the maximum of the absolute value of its eigenvalues, i.e., $(\lambda_{-1}-\lambda_1)/(\lambda_{-1}+\lambda_1)$, which coincides with the right-hand side of (A.1). Since

$$\sin\angle\{u, Au\} = \min_\tau \frac{\|u-\tau Au\|}{\|u\|} \le \frac{\|u-\tau_* Au\|}{\|u\|} \le \|1-\tau_* A\|,$$

we arrive at (A.1). Simple calculation shows that the minimum point of $\sigma(\tau) = \|u - \tau Au\|/\|u\|$ is $\tau_{min}(u) = (u, Au)/\|Au\|^2$ and that $\tau_{min}(u) = \tau_*$ for $u = \alpha_1 x_1 + \alpha_{-1} x_{-1}$, where x_1 and x_{-1} are normalized eigenvectors corresponding to λ_1 and λ_{-1} and $\alpha_1^2\lambda_1 = \alpha_{-1}^2\lambda_{-1}$. Since x_1 and x_{-1} are eigenvectors of $1-\tau_* A$ corresponding to eigenvalues $\pm\|1-\tau_* A\|$, $\|u - \tau_{min}(u)Au\| = \|u-\tau_* Au\| = \|1-\tau_* A\|\|u\|$, and both sides of (A.1) coincide. ∎

Appendix B. Auxiliary estimates for section 4. This section uses the notation of the proof of Theorem 4.1. Additionally, we denote $P_i = P_{x_i}$ and $R_i = R(x_i) = (1-P_i)LP_i$.

From Lemma 2.1 we have a relationship for $J_i - J_{i-1}$ that essentially coincides with that for $J_o - J_p$ in section 3.5:

$$\begin{aligned} J_i - J_{i-1} &= (\lambda^i - \lambda^{i+1})I + R_{i-1} - R_i + R_{i-1}^* - R_i^* \\ &\quad + (\lambda^{i-1} - \lambda^i)P_{i-1} - (\lambda^i - \lambda^{i+1})P_i. \end{aligned} \tag{B.1}$$

For exactly the same reasons as in section 3.5 we have $P_{i-1}h^{i-1} = 0$, $R_{i-1}h^{i-1} = 0$, and $(R_i w, h^{i-1}) = 0$ for any w. Hence, $\|h^{i-1}\|_{J_i}^2 = \|h^{i-1}\|_{J_{i-1}}^2 + (\lambda^i - \lambda^{i+1})\|(1-P_i)h^{i-1}\|^2$, and, employing Corollary 2.3, we obtain $\|h^{i-1}\|_{J_i}^2 = \|h^{i-1}\|_{J_{i-1}}^2(1+\mathcal{O}\left(\epsilon_i^2\right))$.

Further, by (2.13), $\|h^{i-1}\|^2_{J_{i-1}} = (\lambda^{i-1} - \lambda^i)\|x^{i-1}\|^2 = (\lambda^{i-1} - \lambda^i)(1 - \|h^{i-1}\|^2) = (\lambda^{i-1} - \lambda^i)(1 - \mathcal{O}\left(\lambda^{i-1} - \lambda^i\right))$.

Turning now to (4.10), we observe first that by Lemma 2.4, $\|Kr^i\|^2_{J_i^s} = \|Kr^i\|^2_{J_i}(1+\mathcal{O}\left(\epsilon_i^2\right)$. Hence, the left-hand side coincides with the cosine of the J_i-angle between Kr^i and h^{i-1} up to a quantity of the order $\mathcal{O}\left(\epsilon_i^2\right)$ and thus is not greater than $1+\mathcal{O}\left(\epsilon_i^2\right)$. Together with the relationships for the J_i- and J_{i-1}-norms of h^{i-1} just obtained, this implies that $\|h^{i-1}\|^2_{J_i}$ can be replaced with $\|h^{i-1}\|^2_{J_{i-1}}$ and, further, with $\lambda^{i-1} - \lambda^i$ at the cost of adding an asymptotic term of the order $\mathcal{O}\left(\epsilon_i^2\right)$.

Further, $(R_i Kr^i, h^{i-1}) = 0$, and hence (B.1) implies

$$((J_i - J_{i-1})Kr^i, h^{i-1}) = (\lambda^i - \lambda^{i+1})((I - P_i)Kr^i, h^{i-1}) + (R_{i-1}Kr^i, h^{i-1}) - (Kr^i, R_i h^{i-1}).$$

It is easy to see that the first term in the right-hand side of the above identity is of the order $\mathcal{O}\left(\lambda^i - \lambda^{i+1}\right)\|Kr^i\|_{J_i^s}\|h^{i-1}\|_{J_i}$. For the second term we have

$$|(R_{i-1}Kr^i, h^{i-1})| = \frac{|(Kr^i, x^{i-1})||(r^{i-1}, h^{i-1})|}{\|x^{i-1}\|^2} \le \frac{\|Kr^i\|\|r^{i-1}\|_{J_{i-1}^{-1}}\|h^{i-1}\|_{J_{i-1}}}{\|x^{i-1}\|}$$

$$\le \|K\|^{\frac{1}{2}}\|r^i\|_K\|r^{i-1}\|_{J_{i-1}^{-1}}\sqrt{\lambda^{i-1} - \lambda^i} \sim \|Kr^i\|_{J_i^s}\sqrt{\lambda^{i-1} - \lambda_1}\sqrt{\lambda^{i-1} - \lambda^i},$$

and for the third

$$(Kr^i, R_i h^{i-1}) = \frac{(Kr^i, r^i)(x^i, h^{i-1})}{\|x^{i-1}\|^2} = \frac{\|r^i\|^2_K\|h^{i-1}\|^2}{\|x^{i-1}\|^2}$$

$$= \mathcal{O}\left(\|r^i\|_K\right)\|Kr^i\|_{J_i^s}(\lambda^{i-1} - \lambda^i).$$

Thus, $|((J_i - J_{i-1})Kr^i, h^{i-1})| = \mathcal{O}\left(\epsilon_i\right)\|Kr^i\|_{J_i^s}\sqrt{\lambda^{i-1} - \lambda^i}$, which implies (4.10).

Appendix C. Auxiliary estimates for section 5. In this section we use the notation of section 5.

Since $\mathcal{D}^i \supset \operatorname{span}\{x^i, Kr^i\}$, on each iteration of the generalized Davidson method the error reduction is not worse than with the steepest descent method. The upper bound provided for the latter by (2.24) is given in terms of the condition number of the operator $(I-P_i)K(I-P_i)J_1(x^i)$, which is uniformly bounded for x^i in the vicinity of x_1 (cf. section 4). Hence, there exists $q < 1$ such that $\lambda^{i+1} - \lambda_1 \le q(\lambda^i - \lambda_1) \le \cdots \le q^{i+1}(\lambda^0 - \lambda_1)$, which implies that for any $p > 0$ we have

$$\sum_{i=j}^{k}(\lambda^i - \lambda_1)^p \le \frac{1}{1-q^p}(\lambda^j - \lambda_1)^p. \tag{C.1}$$

Since $(h^i, x^i) = 0$, the application of (2.13) yields

$$\|x^{i+1}\|^2 = \|x^i\|^2 + \|h^i\|^2 \le \|x^i\|^2 + \frac{\|h^i\|^2_{J_i}}{\delta(\lambda^i + \lambda^{i+1})} = \|x^i\|^2\left(1 + \frac{\lambda^i - \lambda^{i+1}}{\delta(\lambda^i + \lambda^{i+1})}\right),$$

where $\delta(\mu) = \lambda_1 + \lambda_2 - \mu$, which implies $\|x^i\| = 1 + \mathcal{O}\left(\lambda^0 - \lambda^{i+1}\right) = 1 + \mathcal{O}\left(\lambda^0 - \lambda_1\right)$.

By (5.4), $(J_j z^j, z^k) = 0$ for $k < j$, and by the definition of z^j, $(J_j z^j, z^j) = 1$. Employing Lemma 2.1, we obtain

$$\begin{aligned}(J_i z^j, z^k) &= ((J_i - J_j)z^j, z^k) = (\lambda^{j+1} - \lambda^{i+1})(z^j, z^k) + (R_j z^j, z^k) + (z^j, R_j z^k) \\ &\quad -(R_i z^j, z^k) - (z^j, R_i z^k) + (\lambda^j - \lambda^{j+1})(P_j z^j, z^k) - (\lambda^i - \lambda^{i+1})(P_i z^j, z^k) \\ &= (\lambda^{j+1} - \lambda^{i+1})(z^j, z^k) + (z^j, R_j z^k) - (\lambda^i - \lambda^{i+1})(P_i z^j, z^k),\end{aligned}$$

where $R_m = R(x^m)$, and we have used the fact that $P_j z^j = 0$ and $R_j z^j = 0$ and that for any z, $R_i z$ is collinear to r^i, which is orthogonal to z^j and z^k, and hence $(R_i z^j, z^k) = (R_i z^k, z^j) = 0$. With the help of Lemma 2.2, one finds that the first term in the right-hand side of the above identity is bounded by $\epsilon_{j+1} < \sqrt{\epsilon_j \epsilon_k}$, and the last term by a still smaller quantity ϵ_i. For the middle term we have $|(z^j, R_j z^k)| = |(z^j, r^j)(x^j, z^k)|$, and a calculation similar to (2.15) shows that $|(z^j, r^j)| \le \|r^j\|_{J_j^{-1}} \le \sqrt{\lambda_2 - \lambda_1}\sqrt{\epsilon_j}\|x^j\|$, whereas $(x^j, z^k) = (x^j - P_k x^j, z^k)| \le \|x^j - P_k x^j\|_{J_k}\|z^k\|_{J_k}/\delta(\lambda^k + \lambda^{k+1})$. Since x^j is the minimum point of $\lambda(v)$ in $\mathcal{D}^{j-1}$, it is also the minimum point on the line $v(\tau) = P_k x^j + \tau(x^j - P_k x^j)$. Hence, by (2.13), $\|x^j - P_k x^j\|_{J_k} = \sqrt{\lambda^k - \lambda^j}\|P_k x^j\| \le \sqrt{\lambda^k - \lambda^j}\|x^j\| \le \sqrt{\lambda^k - \lambda_1}\|x^j\|$, and we arrive at (5.6).

REFERENCES

[1] P.-A. Absil, C. G. Baker, and K. A. Gallivan, *A truncated-CG style method for symmetric generalized eigenvalue problems*, J. Comput. Appl. Math., 189 (2006), pp. 274–285.

[2] W. W. Bradbury and R. Fletcher, *New iterative methods for solution of the eigenproblem*, Numer. Math., 9 (1966), pp. 259–267.

[3] J. W. Daniel, *The conjugate gradient method for linear and nonlinear operators equations*, SIAM J. Numer. Anal., 4 (1967), pp. 10–26.

[4] A. Edelman, T. A. Arias, and S. T. Smith, *The geometry of algorithms with orthogonality constraints*, SIAM J. Matrix Anal. Appl., 20 (1998), pp. 303–353.

[5] Y. T. Feng and D. R. Owen, *Conjugate gradient methods for solving the smallest eigenpair of large symmetric eigenvalue problems*, Internat. J. Numer. Methods Engrg., 39 (1996), pp. 2209–2229.

[6] R. Fletcher and C. M. Reeves, *Function minimization by conjugate gradients*, Comput. J., 7 (1964), pp. 149–154.

[7] M. R. Hestenes and E. Stiefel, *Methods of conjugate gradients for solving linear systems*, J. Research Nat. Bur. Standards, 49 (1952), pp. 409–436.

[8] A. V. Knyazev, *Computation of Eigenvalues and Eigenvectors for Mesh Problems: Algorithms and Error Estimates*, Department of Numerical Mathematics, USSR Academy of Sciences, Moscow, 1986 (in Russian).

[9] A. V. Knyazev, *Convergence rate estimates for iterative methods for a mesh symmetric eigenvalue problem*, Soviet J. Numer. Anal. Math. Modelling, 2 (1987), pp. 371–396.

[10] A. V. Knyazev, *A preconditioned conjugate gradient method for eigenvalue problems and its implementation in a subspace*, Internat. Ser. Numer. Math., 96 (1991), pp. 143–154.

[11] A. V. Knyazev, *Toward the optimal preconditioned eigensolver: Locally optimal block preconditioned conjugate gradient method*, SIAM J. Sci. Comput., 23 (2001), pp. 517–541.

[12] A. V. Knyazev and K. Neymeyr, *A geometric theory for preconditioned inverse iteration.* III: *A short and sharp convergence estimate for generalized eigenvalue problems*, Linear Algebra Appl., 358 (2003), pp. 95–114.

[13] K. Neymeyr, *A geometric theory for preconditioned inverse iteration*, Linear Algebra Appl., 322 (2001), pp. 61–104.

[14] Y. Notay, *Combination of Jacobi–Davidson and conjugate gradients for the partial symmetric eigenproblem*, Numer. Linear Algebra Appl., 9 (2002), pp. 21–44.

[15] Y. Notay, *Convergence analysis of inexact Rayleigh quotient iteration*, SIAM J. Matrix. Anal. Appl., 24 (2003), pp. 627–644.

[16] E. Ovtchinnikov, *Convergence estimates for the generalized Davidson method for symmetric eigenvalue problems* I: *The preconditioning aspect,* SIAM J. Numer. Anal., 41 (2003), pp. 258–271.

[17] E. Ovtchinnikov, *Convergence estimates for the generalized Davidson method for symmetric eigenvalue problems* II: *The subspace acceleration,* SIAM J. Numer. Anal., 41 (2003), pp. 272–286.

[18] E. E. Ovtchinnikov, *Sharp convergence estimates for the preconditioned steepest descent method for Hermitian eigenvalue problems*, SIAM. J. Numer. Anal., 43 (2006), pp. 2668–2689.

[19] E. E. Ovtchinnikov, *Jacobi correction equation, line search, and conjugate gradients in Hermitian eigenvalue computation* II: *Computing several extreme eigenvalues*, SIAM J. Numer. Anal., 46 (2008), pp. 2593–2619.

[20] A. Perdon and G. Gambolati, *Extreme eigenvalues of large sparse matrices by Rayleigh quotient and modified conjugate gradients*, Comput. Methods Appl. Mech. Engrg., 56 (1986), pp. 251–264.

[21] E. Polak, *Computational Methods in Optimization: A Unified Approach*, Academic Press, New York, London, 1971.

[22] B. T. Polyak, *The conjugate gradient method in extremal problems*, USSR Comput. Math. Math. Phys., 9 (1969), pp. 94–112.

[23] B. T. Polyak, *Introduction to Optimization*, Optimization Software, New York, 1987.

[24] A. Sameh and Z. Tong, *The trace minimization method for the symmetric generalized eigenvalue problem*, J. Comput. Appl. Math., 123 (2000), pp. 155–175.

[25] B. A. Samokish, *The steepest descent method for an eigenvalue problem with semi-bounded operators*, Izv. Vyssh. Uchebn. Zaved. Mat., 5 (1958), pp. 105–114 (in Russian).

[26] V. Simonchini and L. Eldén, *Inexact Rayleigh quotient-type methods for eigenvalue computations*, BIT, 42 (2002), pp. 159–182.

[27] G. L. G. Sleijpen and H. A. van der Vorst, *A Jacobi–Davidson iteration method for linear eigenvalue problems*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 401–425.

[28] I. Takahashi, *A note on the conjugate gradient method*, Inform. Process. Japan, 5 (1965), pp. 45–49.

[29] X. Yang, T. K. Sarkar, and E. Arvas, *A survey of conjugate gradient algorithms for solution of extreme eigen-problems of a symmetric matrix*, IEEE Trans. Acoust. Speech Signal Process., 37 (1989), pp. 1550–1556.

[30] Y. Zhou, *Studies on Jacobi–Davidson, Rayleigh quotient iteration, inverse iteration generalized Davidson and Newton updates*, Numer. Linear Algebra Appl., 13 (2006), pp. 621–642.

SIAM J. NUMER. ANAL.
Vol. 46, No. 5, pp. 2593–2619
© 2008 Society for Industrial and Applied Mathematics

JACOBI CORRECTION EQUATION, LINE SEARCH, AND CONJUGATE GRADIENTS IN HERMITIAN EIGENVALUE COMPUTATION II: COMPUTING SEVERAL EXTREME EIGENVALUES*

E. E. OVTCHINNIKOV†

Abstract. This paper addresses the question of how to efficiently adapt the conjugate gradient (CG) method to the computation of several leftmost or rightmost eigenvalues and corresponding eigenvectors of Hermitian problems. A generic block CG algorithm instantiated by some available block CG algorithms is considered whereby the new approximate eigenpairs are computed by applying the Rayleigh–Ritz procedure in the trial subspace spanning current approximate eigenvectors and the search direction vectors, each of the latter being a linear combination of the respective gradient of the Rayleigh quotient and all search directions from the previous iteration. An approach related to the so-called *Jacobi orthogonal complement correction* equation is exploited in the local convergence analysis of this CG algorithm. Based on theoretical considerations, a new block conjugation scheme (a way to compute search directions) is suggested that enjoys a certain kind of optimality and has proved to be competitive in practical eigenvalue computation.

Key words. eigenvalue computation, convergence estimates, block conjugate gradients

AMS subject classifications. 65F15, 65N12, 65N22, 65N30

DOI. 10.1137/070688754

1. Introduction.

This is the second part of a paper that addresses the computation of eigenvalues of Hermitian operators by iterative schemes based on the conjugate gradient (CG) method. The first part [17] is concerned with the computation of one extreme eigenvalue; here we tackle the case of several extreme eigenvalues. Just as in [17], we focus on the case of leftmost eigenvalues, to which the case of rightmost eigenvalues is totally symmetric, and consider, for simplicity of notation, the standard eigenvalue problem.

One way to compute several leftmost eigenvalues of a Hermitian problem by CG is to use it for computing the leftmost eigenvalue and corresponding eigenvector and then to apply CG in the subspace of vectors orthogonal to the computed eigenvector, which will produce the second leftmost eigenvalue, etc. (the computational scheme just described is a particular implementation of the technique known as *deflation*). The present paper deals with an alternative approach where eigenpairs are computed simultaneously rather than successively.

The CG method is a functional minimization method that employs, on each iteration, the line search for the minimum of the functional in question in the direction of a linear combination of the gradient and previous search direction. It can be used for computing the leftmost eigenvalue of a Hermitian operator L acting in a Euclidean space $\mathcal{E}$, owing to the fact that this eigenvalue is the minimum of the Rayleigh quotient functional $\lambda(v) = (Lv, v)/\|v\|^2$. Specific properties of the Rayleigh quotient allow the replacement of the line search step with the Rayleigh–Ritz procedure (see section A.1 for the summary of this procedure and related terminology) in the two-dimensional

*Received by the editors April 19, 2007; accepted for publication (in revised form) March 14, 2008; published electronically June 18, 2008.
http://www.siam.org/journals/sinum/46-5/68875.html
†Harrow School of Computer Science, University of Westminster, London HA1 3TP, UK (ovtchie@westminster.co.uk).

trial subspace spanning the current approximate eigenvector and the search vector, with the Ritz vector corresponding to the left Ritz value taken as the new approximate eigenvector. A straightforward generalization of such a technique to simultaneous computation of several eigenvalues is to apply the Rayleigh–Ritz procedure in the subspace spanning all current approximate eigenvectors and search vectors and to take Ritz vectors corresponding to leftmost Ritz values as the new approximate eigenvectors. Here we refer to algorithms employing such a technique as *block* algorithms.

It should be stressed that block CG algorithms are generally not optimization algorithms in the standard sense, since most of them do not minimize a scalar function of several arguments, which prevents the direct use of convergence results and conjugation schemes available to date for CG for the function minimization.

What appears to be the first block CG algorithm was considered in [12]; however, it was discarded there, apparently for lack of justification for the way the search directions were defined and conjugated. (In the block CG algorithm in question, the jth search direction, where j enumerates approximate eigenpairs, was computed by conjugating the jth gradient to the jth previous search direction by the Bradbury–Fletcher scheme. An implicit assumption that can be seen behind this conjugation scheme is that on each iteration the jth Ritz vector represents an ever improving approximation of the same exact eigenvector, which is generally not true.) Another block CG scheme, where the search directions are computed in a similar way but using just one conjugation parameter for all pairs, can be found in section 4.4 of [5]. The motivation for such conjugation is transplanted from a totally different approach to the use of CG for simultaneous eigenvalue computation known as the trace minimization. We note that the latter approach was also considered and discarded by [12], and we refer the reader to [18] for further details and discussion of this and some other approaches not mentioned here.

Turning back to single-vector CG algorithms for instruction, one notices that the conjugation issue can be bypassed by adding previous search directions to the trial subspace, as done in the locally optimal CG (cf. part I of this paper [17]). It is not difficult to see that in the block CG described above, each search direction, however it is defined, that essentially participates in the computation of new eigenvector approximations (in the sense that discarding the former would affect the latter) is a linear combination of current and new approximate eigenvectors. Hence, if the current search directions are linear combinations of gradients and previous search directions, then the enlarged trial subspace can be defined as that spanning current approximate eigenvectors, respective gradients, and either previous approximate eigenvectors or some linear combinations of current and previous approximate eigenvectors. The resulting block CG, known as locally optimal block preconditioned conjugate gradient (LOBPCG) [8, 9], does not involve conjugation of the search direction—or else, one might say that the conjugation is implicit.

Bypassing the conjugation problem comes at a price: in LOBPCG one has to perform the Rayleigh–Ritz procedure in a larger trial subspace, and computationally expensive precautions must be taken to avoid problems caused by almost linearly dependent search directions (see [10]). In an apparent attempt to reduce the size of the trial subspace while preserving good convergence, [1] reinvented the block CG scheme suggested in [12], albeit in the framework of a somewhat different algorithm and with preconditioning, and showed that it works reasonably well. Judging by the numerical results of [1], this attempt has not been quite successful: the reinvented block CG scheme failed to outperform LOBPCG, despite being less computationally expensive,

because of a slower convergence. Numerical tests in [18] demonstrate that the convergence of the method in question could have been improved had the Bradbury–Fletcher conjugation scheme employed by [1] been replaced by the Polak–Ribiére scheme (cf. (2.3) and (2.4) in section 2). The resulting method outperforms LOBPCG in terms of the CPU time in the tests of [18], but the convergence of the former still lags considerably behind that of the latter. Still, the improvement brought about by the change of the conjugation scheme indicates that the potential of the approach reluctantly suggested by [12] and implemented by [1] has not yet been fully exploited, and more could be achieved if a "proper" conjugation were used.

The purpose of conjugation in the standard CG algorithm for the minimization of a quadratic functional with positive definite Hessian is to achieve optimal convergence. Outside the abovementioned particular optimization area, no results demonstrating optimality of any known CG algorithm are available at present. The Rayleigh quotient minimization has the additional difficulty of the Hessian not being positive definite, and the situation is still worse with block CG algorithms for simultaneous eigenvalue computation, where none of the known explicit conjugation schemes has any theoretical justification, save for the trace minimization scheme of [5] without preconditioning, which has certain practical limitations discussed in [18]. Thus, the question of "proper" conjugation in block CG algorithms remains wide open. In the present paper we tackle this question by studying the local convergence properties (the reduction in the error after an iteration) of a generic block CG algorithm instantiated by the abovementioned block CG algorithms. Based on the results of this investigation, we suggest a block conjugation scheme that is optimal in a certain sense. Numerical tests of [18] demonstrate that the convergence of the resulting block CG method is remarkably similar to that of LOBPCG, with lower computational cost per iteration. The two main ingredients of our analysis are (i) a technique, employed previously in [14], that reduces the convergence analysis of a block iterative method to that of the line search applied simultaneously to all approximate eigenpairs, and (ii) the so-called Jacobi correction equation [23] technique for the convergence analysis of methods employing the line search developed in part I [17], suitably generalized here based on the generalization of the Jacobi correction equation suggested in [4].

The plan of the paper is as follows.

In section 2 we formulate a generic block CG algorithm, of which the block CG algorithms of [12] and [1] and the block variant of the trace minimization scheme of [5] are particular cases. In section 3, we introduce a generalization of the Jacobi correction operator of part I [17]. We study the properties of this operator, the most important of which is the fact that for a given group of approximate eigenpairs, the total adjoint energy norm of the residuals coincides with the total eigenvalue error up to asymptotically insignificant terms (Theorem 3.2), a result of independent theoretical value. We emphasize that dealing with eigenpair groups rather than individual eigenpairs ensures that the results presented here are not adversely affected by eigenvalue clustering (cf. [14]). In section 3.4, we apply the results of section 3 to the simultaneous line search; in particular, we derive an asymptotic equation for the reduction in the total eigenvalue error of a group as a function of the respective search directions. In section 4 we apply the results of section 3.4 to the local convergence analysis of the block CG scheme of section 2 to arrive at a conjugation scheme that is locally optimal in the sense of the optimality of the related simultaneous line search. We derive a preliminary convergence estimate for this scheme similar to that derived for LOBPCG and some other block eigensolvers in [14].

Appendix A summarizes the Rayleigh–Ritz procedure and related results that are used in the paper. We note that some of the results of section A.2, notably, the formula (A.6) for the Lehmann optimal operator, the fact that Lehmann bounds are eigenvalues of matrices (A.8) or (A.9) (depending on the position of the Lehmann pole), and the residual bounds (A.10) and (A.11), do not appear to have been obtained elsewhere. Finally, Appendix B contains proofs that are too long and technical to be placed in the main text.

The convergence analysis of block iterative methods for simultaneous eigenvalue computation is often complicated and requires cumbersome and plentiful notation. The present paper is not exceptional in this respect; however, behind the sizeable amount of notation used here lie just a few principles. Capital calligraphic letters are used for linear spaces (except for $\mathcal{I}$ and $\mathcal{J}$, which are used for sets of indices in the appendices). Noncalligraphic capital letters are used for linear operators and matrices. Lowercase roman letters are used for vectors in $\mathcal{E}$, except for i, j, k, l, m, and n, used for indices and dimensions (cf. implicit integers in FORTRAN), and the letter c, traditionally used for coefficients. Small Greek letters are used for scalars; one exception to this rule is the notation for the Rayleigh quotient and residual functions $\lambda(\cdot)$ and $r(\cdot)$ whose values are determined by the argument (see section 2 for explicit formulae). The letter i always refers to the iteration number, and j, k, l to the exact or approximate eigenpair numbers. The eigenvalues λ_j of L are enumerated in increasing order, each one counted as many times as the multiplicity, and the eigenvectors are denoted by x_j. In sections dealing with CG iterations, we place the superscript i on approximate eigenvalues and eigenvectors and related quantities, and in sections not involving the iteration number, we put a tilde on top of them. The remaining more specific notation is introduced wherever it is needed.

All of the results of the paper apply to the generalized eigenvalue problem after simple modifications (see the introduction and Remark 1 in [17] for the modifications needed).

2. A block CG scheme for eigenvalue computation. In what follows, we denote by $\mathcal{R}R_m(\mathcal{V})$ the matrix whose columns are the Ritz vectors of L in the trial subspace $\mathcal{V}$ of dimension $\dim \mathcal{V} \geq m$ corresponding to the m leftmost Ritz values (see Appendix A for the summary of the Rayleigh–Ritz procedure), and we extend the notation for the Rayleigh quotient and residual to matrices with linear independent columns as follows:

$$\lambda(V) = (V^*V)^{-1}V^*LV, \quad r(V) = LV - V\lambda(V).$$

Consider the following iterative scheme for computing the n_l leftmost eigenpairs of L:

$$X^{i+1} = \mathcal{R}R_{n_b}(\operatorname{span}\{X^i, Y^i\}), \tag{2.1}$$

$$Y^{i+1} = KR^{i+1} + Y^iB_i, \tag{2.2}$$

where X^i and Y^i are matrices with $n_b \geq n_l$ columns, $R^i = r(X^i)$, K is a Hermitian positive definite operator that is used as the preconditioner (cf. part I [17]), KR^{i+1} stands for the matrix whose columns are K-images of the columns of R^{i+1}, and B_i is an n_b-by-n_b matrix.

We observe that the block CG scheme of [12] and [1] is a particular case of (2.1)–(2.2) with a diagonal B_i, the diagonal entries $\beta_j^{(i)}$, $j = 1, \ldots, n_b$, being computed by

the Bradbury–Fletcher conjugation scheme; i.e.,

$$\beta_j^{(i)} = \frac{(Kr_j^{i+1}, r_j^{i+1})}{(Kr_j^i, r_j^i)}, \tag{2.3}$$

where r_j^i are the columns of R^i, i.e., residual vectors (note that in [12], K is the identity; i.e., preconditioning is not used). A similar scheme, tested numerically in [18], uses Polak–Ribiére conjugation:

$$\beta_j^{(i)} = \frac{(Kr_j^{i+1}, r_j^{i+1} - r_j^i)}{(Kr_j^i, r_j^i)}. \tag{2.4}$$

We further observe that the Ritz vectors corresponding to the n_b leftmost Ritz values minimize the sum of the Rayleigh quotients over all sets of n_b orthogonal vectors in the trial subspace; i.e., X^{i+1} minimizes the trace of V^*LV over all V's with n_b orthonormal columns belonging to $\mathcal{V}$. Hence, (2.1)–(2.2) can be used as a trace minimization scheme, in which case $B_i = \beta^{(i)} I_{n_b}$, where I_{n_b} stands for the n_b-by-n_b identity and $\beta^{(i)}$ is a scalar that can be computed by a chosen conjugation scheme in $n_b n$-dimensional space, e.g.,

$$\beta^{(i)} = \frac{\langle KR^{i+1}, R^{i+1} - R^i \rangle}{\langle KR^i, R^i \rangle}, \tag{2.5}$$

where $\langle X, Y \rangle = \operatorname{Tr} Y^*X$ and $\operatorname{Tr} A$ stands for the trace of A (cf. [5]; note that here we have added preconditioning).

In this paper we investigate the local convergence properties of (2.1)–(2.2) (i.e., the reduction in eigenvalue error after an iteration) to come up with a suggestion for B_i that in practical computation delivers considerably better convergence than with any of B_i given above.

As mentioned in the introduction, the convergence analysis of the paper follows a technique employed by [14] for the convergence analysis of block preconditioned steepest descent and some related block methods. We observe that the trial subspace $\mathcal{V}^i = \operatorname{span}\{X^i, Y^i\}$ contains the following auxiliary subspace:

$$\mathcal{W}_m^i = \operatorname{span}\{x_j^i - \alpha_j^{(i)}(1 - P_i)y_j^i\}_{j=1}^m, \tag{2.6}$$

where $m \le n_b$ is such that $\lambda_m^i < \lambda_{m+1}$, P_i is the orthogonal projector onto $\mathcal{X}^i = \operatorname{span}\{X^i\}$, and $\alpha_j^{(i)}$ are scalar parameters. By the minimax principle, the m leftmost Ritz values in $\mathcal{V}^i$ are not greater than the respective Ritz values in $\mathcal{W}_m^i$. At the same time, it is not difficult to construct an example where the two coincide. We derive asymptotic equations for the Ritz values in $\mathcal{W}_m^i$ for a specific (locally optimal) $\alpha_j^{(i)}$, and, based on these equations, suggest a conjugation scheme that minimizes these Ritz values. According to [18], the convergence of the resulting two-term algorithm (2.1)–(2.2) is remarkably similar to that of LOBPCG, which suggests that the "core" subspace $\mathcal{W}_m^i$ contained in the trial subspaces of both methods might be the main component controlling their convergence, a hypothesis yet to be investigated.

We emphasize that the "core" subspace given by (2.6) is a purely theoretical object that is employed in this paper for the convergence analysis of the block CG scheme (2.1)–(2.2), and not a part of a new block algorithm suggested as an alternative to the latter.

The outlined approach requires the investigation of the efficiency of the line search in the direction of $(1 - P_i)y_j^i$ as a function of y_j^i. As mentioned in the introduction, the present paper employs for this task a technique based on the so-called Jacobi correction operator. In the next section we study the properties of a generalization of this operator to the case at hand.

3. Jacobi correction operator and simultaneous line search.

3.1. The Jacobi correction equation for invariant subspaces and related operators. Denote $X = [x_j]_{j=1}^m$, $\Lambda = \operatorname{diag}\{\lambda_j\}_{j=1}^m$, and $\mathcal{X} = \operatorname{span}\{X\}$. Let $\tilde{x}_j$ be linearly independent vectors approximating x_j, $j = 1, \dots, m$, and denote $\tilde{X} = [\tilde{x}_j]_{j=1}^m$, $\tilde{\mathcal{X}} = \operatorname{span}\{\tilde{X}\}$, and $R = r(\tilde{X})$. We assume that the angle between $\tilde{\mathcal{X}}$ and $\mathcal{X}$ is less than $\pi/2$; i.e., $\tilde{\mathcal{X}}$ does not contain vectors orthogonal to $\mathcal{X}$. Owing to this assumption, the m-by-m matrix $X^*\tilde{X}$ is nondegenerate. Denoting $X' = XS$, where $S = (\tilde{X}^*X)^{-1}\tilde{X}^*\tilde{X}$, we observe that $\tilde{X}^*(X' - \tilde{X}) = 0$, i.e., $X' = \tilde{X} + E$, where $\tilde{X}^*E = 0$.

Since $\mathcal{X}$ is an invariant subspace, and X' is the basis of $\mathcal{X}$, we have

$$LX' = X'\lambda(X'). \tag{3.1}$$

The substitution of $\tilde{X} + E$ for X' yields

$$LE - E\lambda(X') = -L\tilde{X} + \tilde{X}\lambda(X').$$

Let $\tilde{P}$ be the orthogonal projector onto $\tilde{\mathcal{X}}$. Applying $\tilde{P}_\perp = 1 - \tilde{P}$ to both sides of the above equation and noting that $\tilde{P}E$ and $\tilde{P}R$ vanish, we arrive at

$$\begin{aligned}\tilde{P}_\perp L\tilde{P}_\perp E - \tilde{P}_\perp E\lambda(X') &= -\tilde{P}_\perp(L\tilde{X} - \tilde{X}\lambda(X')) = -\tilde{P}_\perp(L\tilde{X} - \tilde{X}\lambda(\tilde{X}))\\ &= -\tilde{P}_\perp R = -R + \tilde{P}R = -R\end{aligned}$$

or

$$J(\lambda(X'), \tilde{\mathcal{X}})E = -R, \tag{3.2}$$

where

$$J(M, \mathcal{Y})V = (1 - P_{\mathcal{Y}})L(1 - P_{\mathcal{Y}})V - (1 - P_{\mathcal{Y}})VM \tag{3.3}$$

and $P_{\mathcal{Y}}$ is the orthogonal projector onto $\mathcal{Y}$. We observe that in the particular case $m = 1$, (3.2) becomes the Jacobi orthogonal complement correction equation of [23], and $J(M, \mathcal{Y})$ becomes what is for brevity called "Jacobi correction operator" in [17].

Assuming $\tilde{x}_i$ to be orthonormal, we observe that $\tilde{X}^*X'$ is the m-by-m identity matrix, and since $R^*\tilde{X} = 0$, we obtain

$$\begin{aligned}R^*E = R^*(X' - \tilde{X}) = R^*X' &= \tilde{X}^*LX' - \lambda(\tilde{X})\tilde{X}^*X'\\ &= \tilde{X}^*X'\lambda(X') - \lambda(\tilde{X})\tilde{X}^*X' = \lambda(X') - \lambda(\tilde{X}),\end{aligned}$$

i.e.,

$$\lambda(X') = \lambda(\tilde{X}) + R^*E, \tag{3.4}$$

and hence 3.2 can be rewritten as the following nonlinear Riccati equation for E derived in different terms in [4]:[1]

$$J(\tilde{X})E = -R + ER^*E, \tag{3.5}$$

where $J(Y) = J(\lambda(Y), \text{span}\{Y\})$. In [4], various algorithms for computing invariant subspaces via the solution of (3.5) are suggested; in the present paper we use this equation for purely theoretical purposes.

If $\tilde{x}_j$ are Ritz vectors, then $\lambda(\tilde{X})$ is a diagonal matrix with the corresponding Ritz values $\tilde{\lambda}_j$ on the diagonal, and we have

$$J(\tilde{X})E = [J(\tilde{\lambda}_j, \tilde{\mathcal{X}})e_j]_{j=1}^m, \tag{3.6}$$

where e_j is the jth column of E and $J(\mu,\mathcal{U})$ is given by

$$J(\mu,\mathcal{U}) = (1 - P_{\mathcal{U}})(L - \mu)(1 - P_{\mathcal{U}}). \tag{3.7}$$

3.2. Elementary properties of the operator $J(\mu,\mathcal{U})$. Some properties of the operator $J(\mu,\mathcal{U})$ in the case $\dim\mathcal{U} > 1$ are quite similar to its properties in the case $\dim\mathcal{U} = 1$ studied in part I [17]; e.g., the following analogue of Lemma 2.2 from [17] holds. Below $\mathcal{U}^\perp$ denotes the subspace of vectors orthogonal to $\mathcal{U}$.

LEMMA 3.1. *Let $\lambda_j^{\mathcal{U}}$, $j = 1,\dots,m$, be the Ritz values of L in the trial subspace $\mathcal{U}$ of dimension m. For any $v \in \mathcal{U}^\perp$*

$$(J(\mu,\mathcal{U})v, v) \geq \gamma(\mu,\mathcal{U})\|v\|^2, \tag{3.8}$$

where

$$\gamma(\mu,\mathcal{U}) = \lambda_{m+1} - \mu - \delta_\lambda(\mathcal{U}), \quad \delta_\lambda(\mathcal{U}) = \sum_{j=1}^m (\lambda_j^{\mathcal{U}} - \lambda_j). \tag{3.9}$$

Proof. Let $u_1,\dots,u_m$ be orthonormal vectors in $\mathcal{U}$, and denote $U = [u_1,\dots,u_m]$. The scaling of v does not affect 3.8; hence, assume that $\|v\| = 1$. The Ritz values $\hat{\lambda}_j$, $j = 1,\dots,m+1$, of L in $\text{span}\{u_1,\dots,u_m,v\}$ are the eigenvalues of the matrix

$$\hat{L} = \begin{bmatrix} U^*LU & U^*Lv \\ v^*LU & v^*Lv \end{bmatrix}.$$

We have $\hat{\lambda}_1 + \cdots + \hat{\lambda}_{m+1} = \text{Tr}(\hat{L}) = \text{Tr}(U^*LU) + v^*Lv = \lambda_1^{\mathcal{U}} + \cdots + \lambda_m^{\mathcal{U}} + \lambda(v)$ and, by the minimax principle, $\hat{\lambda}_i \geq \lambda_i$. Hence, $\lambda(v) \geq \lambda_1 + \cdots + \lambda_{m+1} - \lambda_1^{\mathcal{U}} - \cdots - \lambda_m^{\mathcal{U}}$. Since $(J(\mu,\mathcal{U})v, v) = (\lambda(v) - \mu)\|v\|^2$, we arrive at 3.8. ∎

Lemma 3.1 implies that $J(\mu,\mathcal{U})$ is positive definite in $\mathcal{U}^\perp$, provided that the total error of the Rayleigh–Ritz approximations to the first m eigenvalues, i.e., the sum of $\lambda_j^{\mathcal{U}} - \lambda_j$, is less than $\lambda_{m+1} - \mu$. One corollary of this fact is that, under the latter assumption on μ and $\mathcal{U}$, the operator $J(\mu,\mathcal{U})$ is invertible in $\mathcal{U}^\perp$ and the inverse of its restriction to $\mathcal{U}^\perp$, denoted here as $J(\mu,\mathcal{U})^{-1}$, satisfies

$$(J(\mu,\mathcal{U})^{-1}v, v) \leq \gamma(\mu,\mathcal{U})^{-1}\|v\|^2 \leq \gamma(\mu,\mathcal{U})^{-2}(J(\mu,\mathcal{U})v, v). \tag{3.10}$$

[1] We note that both derivations do not assume L to be Hermitian.

Another corollary is that the $J(\mu,\mathcal{U})$-energy seminorms for different μ's satisfying this assumption are equivalent; notably, if $\gamma(\nu,\mathcal{U}) > 0$, then for any $\mu < \nu$

$$(J(\nu,\mathcal{U})v,v) \le (J(\mu,\mathcal{U})v,v) \le \frac{\gamma(\mu,\mathcal{U})}{\gamma(\nu,\mathcal{U})}(J(\nu,\mathcal{U})v,v) \tag{3.11}$$

(cf. Lemma 2.4 of [17]), which implies

$$(J(\mu,\mathcal{U})^{-1}v,v) \le (J(\nu,\mathcal{U})^{-1}v,v) \le \frac{\gamma(\mu,\mathcal{U})}{\gamma(\nu,\mathcal{U})}(J(\mu,\mathcal{U})^{-1}v,v). \tag{3.12}$$

3.3. Asymptotic relationships between errors and residuals. It is not difficult to verify that the norm of E is the tangent of the angle between $\mathcal{X}$ and $\tilde{\mathcal{X}}$:

$$\tan\{\mathcal{X};\tilde{\mathcal{X}}\} = \max_{w\in\mathcal{X}} \frac{\|\tilde{P}_\perp w\|}{\|\tilde{P}w\|} = \max_{\hat{w}\in\mathcal{E}^m} \frac{\|\tilde{P}_\perp X'\hat{w}\|}{\|\tilde{P}X'\hat{w}\|} = \max_{\hat{w}\in\mathcal{E}^m} \frac{\|E\hat{w}\|}{\|\tilde{X}\hat{w}\|} = \max_{\hat{w}\in\mathcal{E}^m} \frac{\|E\hat{w}\|}{\|\hat{w}\|} = \|E\|,$$

where $\mathcal{E}^m$ stands for the m-dimensional Euclidean space of the column vector, and $\|E\|$ is the norm of E viewed as a linear operator acting from $\mathcal{E}^m$ to $\mathcal{E}$ (the above identity is valid for arbitrary orthonormal $\tilde{x}_j$). Furthermore, for the Ritz vectors $\tilde{x}_j$, the columns e_j of E are asymptotically (as $\|E\| \to 0$) close to the error vectors $x_j - \tilde{x}_j$ (see [7]). In this section we derive asymptotic relationships between norms of the "error vectors" e_j and the corresponding residual norms and eigenvalue errors. Below we assume that

$$\gamma(\tilde{\lambda}_m, \tilde{\mathcal{X}}) \ge \gamma > 0. \tag{3.13}$$

In order to eliminate possible adverse effects of eigenvalue clustering and to allow for multiple eigenvalues, the relationships in question are derived here for a group of leftmost Ritz values and vectors, assuming that the eigenvalues of L approximated by these Ritz values are well separated from the rest of the spectrum (cf. [15, 16]).

Owing to 3.13, the operators $J_j = J(\tilde{\lambda}_j, \tilde{\mathcal{X}})$ are positive definite in $\tilde{\mathcal{X}}^\perp$, and below we denote by J_j^{-1} the inverse of the restriction of J_j to $\tilde{\mathcal{X}}^\perp$. For any $v \in \tilde{\mathcal{X}}^\perp$ we denote $\|v\|_{J_j}^2 = (J_j v, v)$ and $\|v\|_{J_j^{-1}}^2 = (J_j^{-1} v, v)$. The result below, a generalization of Lemma 2.5 from [17], suggests that the two introduced energy norms are "natural" ones for the "error" and residual vectors, respectively.

THEOREM 3.2. *Assuming* 3.13, *for* $l \le m$ *such that* $\lambda_l < \lambda_{l+1}$ *the following inequalities are valid:*

$$\left| \sum_{j=1}^{l} \|e_j\|_{J_j}^2 - \sum_{j=1}^{l} (\tilde{\lambda}_j - \lambda_j) \right| \le c_e \frac{\delta_\lambda(\tilde{\mathcal{X}})}{\lambda_{m+1} - \lambda_m} \sum_{j=1}^{l} (\tilde{\lambda}_j - \lambda_j), \tag{3.14}$$

$$\left| \sum_{j=1}^{l} \|r_j\|_{J_j^{-1}}^2 - \sum_{j=1}^{l} (\tilde{\lambda}_j - \lambda_j) \right| \le c_r \frac{\delta_\lambda(\tilde{\mathcal{X}})}{\lambda_{m+1} - \lambda_m} \sum_{j=1}^{l} (\tilde{\lambda}_j - \lambda_j), \tag{3.15}$$

where $\delta_\lambda(\cdot)$ *is given in* 3.9 *and the coefficients* c_e *and* c_r *depend only on the ratios* $\rho_m = (\lambda_{m+1} - \lambda_m)/\gamma$, $\delta_m = (\lambda_{m+1} - \lambda_m)/(\lambda_{m+1} - \lambda_1)$ *and* $\delta_l = (\lambda_{l+1} - \lambda_l)/(\lambda_{m+1} - \lambda_1)$.

Proof. See section B.1. □

The inequalities of Theorem 3.2 reveal the asymptotic identity of J_j-based residual and error norms and eigenvalue errors:

$$\sum_{j=1}^{l} \|e_j\|_{J_j}^2 \simeq \sum_{j=1}^{l} \|r_j\|_{J_j^{-1}}^2 \simeq \sum_{j=1}^{l} (\tilde{\lambda}_j - \lambda_j), \tag{3.16}$$

where $a \simeq b$ stands for $a = (1 + \mathcal{O}(\delta_\lambda(\tilde{\mathcal{X}})))b$. In the paper we use these asymptotic relationships in order to obtain asymptotic identities related to the eigenvalue error reduction after a simultaneous line search applied to m vectors as a function of search directions, which, in turn, lead to a suggestion on a "proper" choice of the conjugation matrix B_i in (2.2).

3.4. Convergence properties of the simultaneous line search. In this section we study the properties of the simultaneous line search step whereby one starts with Ritz vectors $\tilde{x}_j$, $j = 1, \ldots, m$, and then applies the Rayleigh–Ritz procedure in the trial subspace

$$\mathcal{W} = \operatorname{span}\{w_j\}_{j=1}^m, \quad w_j = \tilde{x}_j - \alpha_j y_j, \quad \alpha_j = \arg\min_\alpha \lambda(\tilde{x}_j - \alpha y_j), \tag{3.17}$$

where y_j is orthogonal to $\tilde{\mathcal{X}}$ and nonorthogonal to $r_j = r(\tilde{x}_j)$. The orthogonality of y_j to $\tilde{x}_j$ ensures that the line searches in 3.17 do not fail (cf. [17, section 2.2]), and the orthogonality to $\tilde{\mathcal{X}}$ ensures that vectors w_j are linearly independent.

The first result of this section, a (partial) generalization of Theorem 2.6 of [17], expresses the difference between approximate eigenpairs before and after an individual line search step as a function of the search vector.

THEOREM 3.3. *Let u be a nonzero vector, let $\mathcal{U}$ be a subspace that contains u and is orthogonal to $r(u)$, and let v be orthogonal to $\mathcal{U}$ and such that $(v, r(u)) \neq 0$. Let $u_v = u + \tau_v v$, where $\tau_v \equiv \arg\min_\tau \lambda(u + \tau v)$, and denote $J_{u,v} = J(\lambda(u_v), \mathcal{U})$. The following relationships are valid:*

$$\tau_v = -\frac{(r(u), v)}{(J_{u,v} v, v)}, \tag{3.18}$$

$$\lambda(u) - \lambda(u_v) = \frac{|(r(u), v)|^2}{\|u\|^2 (J_{u,v} v, v)}, \tag{3.19}$$

$$(J_{u,v}(u - u_v), u - u_v) = (\lambda(u) - \lambda(u_v))\|u\|^2. \tag{3.20}$$

Proof. We observe that $(J_{u,v} w, w) = (J(\lambda(u_v), \mathcal{U})w, w) = (J(\lambda(u_v), \operatorname{span}\{u\})w, w)$ for any w that is orthogonal to $\mathcal{U}$. Hence, Theorem 2.6 of part I [17] implies (3.18)–(3.20). ∎

In the rest of this section we assume 3.13 to be satisfied.

In the case of the steepest descent line search, i.e., with $y_j = Kr_j$, (3.19) implies

$$\tilde{\lambda}_j - \mu_j = \frac{|(r_j, y_j)|^2}{(J(\mu_j, \tilde{\mathcal{X}}) y_j, y_j)} = \frac{|(r_j, Kr_j)|^2}{(J(\mu_j, \tilde{\mathcal{X}}) Kr_j, Kr_j)} \geq \frac{1}{\omega_j} \|r_j\|_K^2 \geq \frac{\alpha_j}{\omega_j} \|r_j\|_{J(\mu_j, \tilde{\mathcal{X}})^{-1}}^2,$$

where $\mu_j = \lambda(w_j)$, and α_j and ω_j are the minimal and maximal eigenvalues of the operator $A_j = J(\mu_j, \tilde{\mathcal{X}})K$, restricted to $\tilde{\mathcal{X}}^\perp$ (note that A_j is Hermitian positive definite in the scalar products $(\cdot, \cdot)_K$ and $(\cdot, \cdot)_{J(\mu_j, \tilde{\mathcal{X}})^{-1}}$). Below we assume that the

search directions y_j are "not much worse" than the steepest descent directions in the sense that a similar inequality is valid; notably, for some $\eta > 0$, we have

$$\|r_j\|^2_{J_j^{-1}} \le \eta(\tilde{\lambda}_j - \mu_j). \tag{3.21}$$

The second result demonstrates that the difference between the sum of the Ritz values in $\mathcal{W}$ approximating a well-separated group of $l \le m$ leftmost eigenvalues and the respective sum of the Rayleigh quotients $\lambda(w_j)$ is asymptotically insignificant compared to the total error reduction achieved by the line searches in (3.17).

THEOREM 3.4. *Denote $\mu_j = \lambda(w_j)$, and denote by ν_j the Ritz values of L in the trial subspace $\hat{\mathcal{X}}$ given by* (3.17). *Let $l \le m$ be such that $\tilde{\lambda}_l < \lambda_{l+1}$. Assuming* (3.13) *and* (3.21), *the following inequality is valid:*

$$\left|\sum_{j=1}^{l}(\nu_j - \mu_j)\right| \le c\frac{\delta_\lambda(\tilde{\mathcal{X}})}{\lambda_{m+1} - \lambda_m}\sum_{j=1}^{l}(\tilde{\lambda}_j - \mu_j), \tag{3.22}$$

where the coefficient c depends only on η and the ratios listed in Theorem 3.2, *with δ_l replaced by $\tilde{\delta}_l = (\lambda_{l+1} - \tilde{\lambda}_l)/(\lambda_{m+1} - \lambda_1)$.*

Proof. For the proof, see section B.2. ∎

Put together, the two results obtained in this subsection allow one to express the total error reduction for $\lambda_1, \dots, \lambda_l$, achieved by moving from $\tilde{X}$ to $\mathcal{W}$ as an explicit, modulo an asymptotically small term, function of the corresponding search directions. In the next section we use such an asymptotic expression in order to optimize search directions y_j in the framework of the iterative scheme (2.1)–(2.2).

4. Local convergence properties of the block CG. Let us now apply the results of the previous section to the investigation of local convergence properties of the block CG scheme (2.1)–(2.2) and the related issue of "proper" conjugation. Below we are using the notation of section 2. We assume that all residuals r_j^i are nonzero, and we assume that[2] (cf. (3.9) and (3.13))

$$\gamma(\lambda_m^i, \mathcal{X}_m^i) \ge \gamma > 0, \quad \mathcal{X}_m^i = \operatorname{span}\{x_j^i\}_{j=1}^m, \quad m \le n_b, \tag{4.1}$$

which, by Lemma 3.1, implies that the Jacobi operators $J(\lambda_j^i, \mathcal{X}_m^i)$ are positive definite in the orthogonal complement to $\mathcal{X}_m^i$. Since $(J(\lambda_j^i, \mathcal{X}_m^i)w, w) = (J(\lambda_j^i, \mathcal{X}^i)w, w)$ for any w that is orthogonal to $\mathcal{X}^i = \operatorname{span}\{X^i\}$, the operators $J(\lambda_j^i, \mathcal{X}^i)$, $j = 1, \dots, m$, are positive definite in the orthogonal complement to $\mathcal{X}^i$.

We remind that the new approximate eigenvalues λ_j^{i+1} produced by (2.1) are the Ritz values in the trial subspace $\mathcal{Y}^i = \operatorname{span}\{X^i, Y^i\}$. Since $(r_j^i, y_k^{i-1}) = 0$, owing to the fact that x_j^i are Ritz vectors in the trial subspace containing y_k^{i-1}, we have

[2] This assumption is somewhat stronger than $\lambda_{m+1} - \lambda_m^i > 0$, assumed in, e.g., [7, 14] (the weakest possible one, in terms of exact and approximate eigenvalues only, that ensures the convergence to the first m eigenpairs of L). The assumption we take here, that $\lambda_{m+1} - \lambda_m^i$ is greater than the total error in m leftmost eigenvalues, is actually not much stronger because (i) the spectrum of L typically consists of clusters of eigenvalues separated by lacunae of the size that is larger than the average distance between the eigenvalues at the margin of the spectrum in focus, and (ii) looking at the convergence curves for eigenvalues (see, e.g., [18]), one often finds that the convergence of a given group of approximate eigenvalues is rather erratic until the eigenvalue errors approach the level of the distance between the corresponding eigenvalues and the rest of the spectrum (hence useful theoretical estimates above this level do not seem to be likely) after which the convergence curves for the logarithm of the error become virtually straight.

$(r_j^i, y_j^i) = (r_j^i, Kr_j^i) = \|r_j^i\|_K^2$; hence, y_j^i does not belong to $\mathcal{X}^i$, and we can perform the line search in the direction $(1 - P_i)y_j^i$, where P_i is the orthogonal projector onto $\mathcal{X}^i$. Let us denote by ν_j^i the Ritz values in the subspace $\mathcal{W}_m^i$ given by (2.6) with $\alpha_j^{(i)}$ that minimize the Rayleigh quotients for $w_j^i = x_j^i - \alpha_j^{(i)}(1 - P_i)y_j^i$. By the minimax principle, $\lambda_j^{i+1} \le \nu_j^i$, and it is not difficult to show that these inequalities are sharp by considering the case where x_j^i and y_j^i belong to $\operatorname{span}\{x_j, x_{j+n_b}, x_{j+2n_b}, \ldots\}$. Indeed, in this case the triples w_j^i, x_j^i, y_j^i belong to different invariant subspaces of L for different j, and hence the residual $r(w_j^i)$ is orthogonal to x_k^i and y_k^i with $k \ne j$; it is also orthogonal to $(1 - P_i)y_j^i$, owing to the locally optimal $\alpha_j^{(i)}$, and hence to x_j^i. Thus, $r(w_j^i)$ is orthogonal to $\mathcal{V}^i$; i.e., w_j^i are Ritz vectors in $\mathcal{V}^i$, and the assumption (4.1) implies that the corresponding Ritz values are the leftmost ones. Hence $\lambda_j^{i+1} = \nu_j^i$.

The above considerations suggest that the search directions should be conjugated in such a way that ν_j^i are as minimal as possible. The results of the previous subsection imply the asymptotic identity

$$\sum_{j=1}^{l}(\lambda_j^i - \nu_j^i) = (1 + \varepsilon_{il})\sum_{j=1}^{l}\frac{\|r_j^i\|_K^4}{(J(\lambda_j^i, \mathcal{X}^i)y_j^i, y_j^i)}, \quad |\varepsilon_{il}| \le c\delta_\lambda(\mathcal{X}_m^i), \tag{4.2}$$

where we have used the fact that $(r_j^i, y_j^i) = (r_j^i, Kr_j^i)$, $(J(\lambda_j^i, \mathcal{X}_m^i)(1 - P_i)y_j^i, (1 - P_i)y_j^i) = (J(\lambda_j^i, \mathcal{X}^i)y_j^i, y_j^i)$, and the coefficient c has the same dependencies as that in Theorem 3.4. Thus, the conjugation scheme we are after should be aimed at minimizing $J(\lambda_j^i, \mathcal{X}^i)$-norms of y_j^i. Since y_j^i is the sum of Kr_j^i and a linear combination of the previous search directions, the $J(\lambda_j^i, \mathcal{X}^i)$-norm of y_j^i is minimal when y_j^i is $J(\lambda_j^i, \mathcal{X}^i)$-orthogonal to the previous search directions; i.e., B_i should be such that

$$(J(\lambda_j^i, \mathcal{X}^i)y_j^i, y_k^{i-1}) = 0. \tag{4.3}$$

The conjugation scheme just described is a generalization of the Jacobi scheme of [17], and here we refer to it as the block Jacobi scheme.[3] By construction, this scheme is "quasi-optimal" in the sense that it minimizes the Ritz values in the "core subspace" (2.6), which, as shown above, provide sharp upper bounds for the Ritz values in $\operatorname{span}\{X^i, Y^i\}$. One may also say that the search directions y_j^i are "individually locally optimal": by Theorem 3.3, the line search in the direction $(1 - P_i)y_j^i$ produces w_j^i such that

$$\delta_j^i = \lambda(x_j^i) - \lambda(w_j^i) = \frac{\|r_j^i\|_K^4}{\left(J(\lambda(w_j^i), \mathcal{X}^i)y_j^i, y_j^i\right)} = \frac{\|r_j^i\|_K^4}{(J(\lambda_j^i, \mathcal{X}^i)y_j^i, y_j^i)}\left(1 + \mathcal{O}\left(\delta_j^i\right)\right).$$

The Jacobi conjugation minimizes the denominator in the right-hand side, and thus $(1 - P_i)y_j^i$ is asymptotically the best possible line search direction for minimizing the Rayleigh quotient $\lambda(x_j^i - \alpha(1 - P_i)y_j^i)$ that can be expressed as a linear combination of Kr_j^i and $y_1^{i-1}, \ldots, y_{n_b}^{i-1}$.

In order to provide explicit expressions for the entries of the conjugation matrix B_i, let us consider an alternative formulation of the block Jacobi scheme. Denote for brevity[4] $J_j = J(\lambda_j^{i+1}, \mathcal{X}^{i+1})$. Since $0 = (J_j y_j^{i+1}, y_k^i) = (J_j y_j^{i+1}, (1 - P_{i+1})y_k^i)$,

[3] Just as in [17], λ_j^i may be replaced with asymptotically close values; here we stick to λ_j^i for simplicity.

[4] The iteration superscript is shifted for consistency with section 2.

the conjugation matrix B_i does not change if we replace y_k^i by $(1-P_{i+1})y_k^i$ in 2.2; neither does this orthogonalization affect the new trial subspace $\operatorname{span}\{X^{i+1}, Y^{i+1}\}$. Furthermore, vectors $(1-P_{i+1})y_k^i$ can obviously be replaced by the basis vectors in $\mathcal{Y}_\perp^i = \operatorname{span}\{(1-P_{i+1})Y^i\}$. Noting that $\mathcal{Y}_\perp^i$ is orthogonal to the n_b Ritz vectors in $\mathcal{V}^i = \operatorname{span}\{X^i, Y^i\}$ corresponding to the leftmost Ritz values, we observe that $\mathcal{Y}_\perp^i$ belongs to the linear span of the remaining Ritz vectors corresponding to the rightmost Ritz values. Hence, if $\dim \mathcal{Y}^\perp = \dim \mathcal{V}^i - n_b$, then we can equivalently use these Ritz vectors, denoted here by z_k^{i+1}, in place of y_k^i. If $\dim \mathcal{Y}^\perp < \dim \mathcal{V}^i - n_b$, then we can still do the same because the conjugation to a larger set of vectors reduces $(J_j y_j^{i+1}, y_j^{i+1})$ still further and thus improves the individual local optimality. Denoting $\mu_j^{i+1} = \lambda(z_j^{i+1})$, we observe that $(J_j z_k^{i+1}, z_l^{i+1}) = 0$ for $k \neq l$ and $(J_j z_k^{i+1}, z_k^{i+1}) = (\mu_k^{i+1} - \lambda_j^{i+1})\|z_k^{i+1}\|^2$. Finally, we note that $((L-\lambda_j^{i+1})x_l^{i+1}, z_k^{i+1}) = 0$, and therefore $(J_j K r_j^{i+1}, z_k^{i+1}) = ((L-\lambda_j^{i+1})K r_j^{i+1}, z_k^{i+1})$. Hence, multiplying

$$y_j^{i+1} = K r_j^{i+1} + \sum_{l=1}^{m_i} \beta_{lj}^{(i)} z_l^{i+1}$$

by $J_j z_k^{i+1}$, we obtain, assuming $\lambda_{n_b}^{i+1} < \mu_1^{i+1}$,

$$\beta_{kj}^{(i)} = -\frac{((L-\lambda_j^{i+1})K r_j^{i+1}, z_k^{i+1})}{(\mu_k^{i+1} - \lambda_j^{i+1})\|z_k^{i+1}\|^2}, \tag{4.4}$$

and if $\lambda_{n_b}^{i+1} = \mu_1^{i+1}$, then we set $\beta_{kj}^{(i)} = 0$ for any k such that $\mu_k^{i+1} - \lambda_j^{i+1} = 0$.

Thus, we arrive at the following iterative scheme:

$$[X^{i+1}, Z^{i+1}] = \mathcal{R} R_{\dim \mathcal{V}^i} \mathcal{V}^i, \quad \mathcal{V}^i = \operatorname{span}\{X^i, Y^i\}, \tag{4.5}$$

$$Y^{i+1} = K R^{i+1} + Z^{i+1} B_i, \tag{4.6}$$

where $Z^{i+1} = [z_1^{i+1}, \dots, z_{m_i}^{i+1}]$, $m_i = \dim \mathcal{V}^i - n_b$, and the entries $\beta_{kj}^{(i)}$ of the conjugation matrix B_i are given by (4.4) (practical considerations can be found in [19]).

Unlike in the case $m = 1$, there is apparently no asymptotical equivalence between this scheme and other schemes mentioned in section 2; however, the numerical test of [18] demonstrates remarkable similarity of the convergence behavior of the block Jacobi scheme and LOBPCG, a phenomenon yet to be explained.

A preliminary convergence result given below shows that the local convergence rate of the block CG (2.1)–(2.2) can be estimated in terms of the condition numbers κ_{ij} of operators $A_{ij} = J(\lambda_j^i, \mathcal{X}_m^i)K$, restricted to the subspace $(\mathcal{X}_m^i)^\perp$. We note that A_{ij} is Hermitian in K- and $J(\lambda_j^i, \mathcal{X}_m^i)^{-1}$-scalar products and that the following estimate for κ_{ij} is valid:

$$\kappa_{ij} \le \hat{\kappa}_{ij} = \kappa \left(1 + \lambda_{m+1} \sum_{k=1}^{m} \left(\frac{1}{\lambda_k} - \frac{1}{\lambda_k^i}\right)\right) \frac{\lambda_{m+1} - \delta_\lambda(\mathcal{X}_m^i)}{\lambda_{m+1} - \delta_\lambda(\mathcal{X}_m^i) - \lambda_j^i}, \tag{4.7}$$

where L is assumed to be positive definite for simplicity of notation[5] and κ is the condition number of KL (see section B.4 for the proof). As $\delta_\lambda(\mathcal{X}_m^i) \to 0$, the right-hand side of 4.7 decreases to $\kappa\lambda_{m+1}/(\lambda_{m+1} - \lambda_j)$; hence, $\kappa_{ij} \le \hat{\kappa}_{i_0 j} \le \hat{\kappa}_{i_0 l}$.

[5]To obtain an estimate for a nonpositive L, apply (4.7) to $L - \lambda_0$ with $\lambda_0 < \lambda_1$.

THEOREM 4.1. *Assume that* $\lambda_l^{i_0} < \lambda_{l+1}$ *for some* $l \le m$ *and iteration number* i_0, *and assume* (4.1) *for* $i = i_0$. *The following estimate is valid at each iteration* $i \ge i_0$ *of the block CG* (2.1)–(2.2) *with the search directions satisfying* (4.3)*:*

$$\sum_{j=1}^{l}(\lambda_j^{i+1} - \lambda_j) \le \frac{q_{il} + \epsilon_{il}}{1 + \epsilon_{il}} \sum_{j=1}^{l}(\lambda_j^i - \lambda_j), \tag{4.8}$$

where

$$q_{il} = \max_{1\le j\le l}\left(\frac{\kappa_{ij} - 1}{\kappa_{ij} + 1}\right)^2, \quad \epsilon_{il} \le c\frac{\delta_\lambda(\mathcal{X}_m^i)}{\lambda_{m+1} - \lambda_m}, \tag{4.9}$$

κ_{ij} *is the condition number of the operator* $A_{ij} = J(\lambda_j^i, \mathcal{X}_m^i)K$, *restricted to* $(\mathcal{X}_m^i)^\perp$, $\delta_\lambda(\cdot)$ *is given in* (3.9), *and the coefficient* c *depends on* $\kappa_l = \hat\kappa_{i_0 l}$, *where* $\hat\kappa_{ij}$ *is given in* (4.7), $\rho_m = (\lambda_{m+1} - \lambda_m)/\gamma$, $\delta_m = (\lambda_{m+1} - \lambda_m)/(\lambda_{m+1} - \lambda_1)$, *and* $\delta_l = (\lambda_{l+1} - \lambda_l^{i_0})/(\lambda_{m+1} - \lambda_1)$ *only.*

Proof. Let us first show that the search directions of the scheme in question satisfy (3.21). Below we denote by α_{ij} and ω_{ij} the minimal and maximal eigenvalues of A_{ij}, and $J_{ij} = J(\lambda_j^i, \mathcal{X}_m^i)$.

Let $\alpha_j^{(i)}$ be such that $\mu_j^i = \lambda(x_j^i - \alpha_j^{(i)}(1 - P_i)y_j^i) = \min_\alpha \lambda(x_j^i - \alpha(1 - P_i)y_j^i)$. By Theorem 3.3 we have (cf. (3.18))

$$|\alpha_j^{(i)}| = \frac{|(r_j^i, y_j^i)|}{(J'_{ij}y_j^i, y_j^i)} = \frac{\|r_j^i\|_K^2}{\|y_j^i\|_{J'_{ij}}^2}, \quad J'_{ij} = J(\mu_j^i, \mathcal{X}_m^i).$$

Denoting $\rho_1 = (\lambda_{m+1} - \lambda_1)/\gamma = \rho_m/\delta_m$, by (3.11) we have $\|y_j^i\|_{J'_{ij}}^2 \le \rho_1\|y_j^i\|_{J_{ij}}^2$. The conjugation condition (4.3) implies $\|y_j^i\|_{J_{ij}}^2 \le \|Kr_j^i\|_{J_{ij}}^2$. Since $P_i r_j^i = 0$, we have $\|Kr_j^i\|_{J_{ij}}^2 = (Kr_j^i, J_{ij}Kr_j^i) = (r_j^i, A_{ij}r_j^i)_K \le \omega_{ij}\|r_j^i\|_K^2$. Thus, $|\alpha_j^{(i)}|^{-1} \le \rho_1\omega_{ij}$. Finally, $\|r_j^i\|_{J_{ij}^{-1}}^2 = (J_{ij}^{-1}r_j^i, r_j^i) = (A_{ij}^{-1}r_j^i, r_j^i)_K \le \alpha_{ij}^{-1}\|r_j^i\|_K^2 = \alpha_{ij}^{-1}|\alpha_j^{(i)}|\|y_j^i\|_{J'_{ij}}^2 = \alpha_{ij}^{-1}|\alpha_j^{(i)}|^{-1}\|h_j^i\|_{J'_{ij}}^2 \le \rho_1^2\kappa_l\|h_j^i\|_{J_{ij}}^2 \le \rho_1^2\kappa_l\|h_j^i\|_{J'_{ij}}^2 = \rho_1^2\kappa_l(\lambda_j^i - \mu_j^i)$. Thus, (3.21) is satisfied, and Theorem 3.4 applies.

We have $\lambda_j^{i+1} - \lambda_j \le \nu_j^i - \lambda_j = \lambda_j^i - \lambda_j - (\lambda_j^i - \mu_j^i) - (\mu_j^i - \nu_j^i)$, where ν_j^i are the Ritz values in the trial subspace $\mathcal{W}_m^i = \operatorname{span}\{x_j^i - \alpha_j^{(i)}(1 - P_i)y_j^i\}$. By (3.22), the summation over $j = 1, \dots, l$ yields

$$\sum_{j=1}^{l}(\lambda_j^{i+1} - \lambda_j) \le \sum_{j=1}^{l}(\lambda_j^i - \lambda_j) - \left(1 - c\frac{\delta(\mathcal{X}_m^i)}{\lambda_{m+1} - \lambda_m}\right)\sum_{j=1}^{l}(\lambda_j^i - \mu_j^i).$$

In turn, for $\lambda_j^i - \mu_j^i$ we have, using (3.19) and then (3.11),

$$\lambda_j^i - \mu_j^i = \frac{|(r_j^i, Kr_j^i)|^2}{(J(\mu_j^i, \mathcal{X}^i)y_j^i, y_j^i)} \ge \frac{|(r_j^i, Kr_j^i)|^2}{\|y_j^i\|_{J_{ij}}^2}\left(1 - \frac{\lambda_j^i - \mu_j^i}{\gamma}\right).$$

The J_{ij}-orthogonality of y_j^i and z_k^{i-1} implies $\|y_j^i\|_{J_{ij}} \le \|Kr_j^i\|_{J_{ij}}$, and hence

$$\frac{|(r_j^i, Kr_j^i)|^2}{\|y_j^i\|_{J_{ij}}^2} \ge \frac{|(r_j^i, Kr_j^i)|^2}{\|K_i r_j^i\|_{J_{ij}}^2} = \frac{(r_j^i, A_{ij}r_j^i)_{J_{ij}^{-1}}^2}{\|A_{ij}r_j^i\|_{J_{ij}^{-1}}^2} = \|r_j^i\|_{J_{ij}^{-1}}^2 \cos^2_{J_{ij}^{-1}} \angle\{r_j^i, A_{ij}r_j^i\}.$$

By Lemma A.1 from part I [17] the sine of the angle between r_j^i and $A_{ij}r_j^i$ is bounded by $(\kappa_{ij}-1)/(\kappa_{ij}+1)$. As shown in the proof of Theorem 3.4, the sum of $\lambda_j^i - \mu_j^i$ is not greater than $8\rho_1^2\delta_\lambda(\mathcal{X}_m^i)$. Hence, simple manipulations with the three inequalities displayed just above yield

$$\sum_{j=1}^{l}(\lambda_j^{i+1}-\lambda_j) \le q_{il}\sum_{j=1}^{l}(\lambda_j^i-\lambda_j), + \frac{c_0\delta_\lambda(\mathcal{X}_m^i)}{\lambda_{m+1}-\lambda_m}\sum_{j=1}^{l}\|r_j^i\|^2_{J_{ij}^{-1}}, \tag{4.10}$$

where c_0 has the same dependencies as c in (4.8). Finally, $\|r_j^i\|^2_{J_{ij}^{-1}} \le \alpha_{ij}^{-1}\|r_j^i\|_K^2$ and (3.11) implies $(J(\lambda_1,\mathcal{X}_m^i)y_j^i,y_j^i) \le \rho_1(J_{ij}y_j^i,y_j^i)$, and thus $(J(\lambda_1,\mathcal{X}_m^i)y_j^i,y_j^i) \le \rho_1(J_{ij}Kr_j^i,Kr_j^i) \le \rho_1\omega_{ij}(Kr_j^i,r_j^i)$. Consulting Lemma B.1, we conclude that

$$\sum_{j=1}^{l}\|r_j^i\|^2_{J_{ij}^{-1}} \le \rho_1\kappa_{ij}\sum_{j=1}^{l}(\lambda_j^i-\lambda_j^{i+1}) = \rho_1\kappa_{ij}\left(\sum_{j=1}^{l}(\lambda_j^i-\lambda_j) - \sum_{j=1}^{l}(\lambda_j^{i+1}-\lambda_j),\right),$$

which, together with (4.10), yields (4.8). ∎

We stress that the above convergence result should be viewed as a preliminary one, being rather "pessimistic": it merely shows that the convergence of the block Jacobi CG is asymptotically not worse than that of the block steepest descent method (cf. convergence estimates for the block steepest descent of [13, 14], which, owing to the minimax principle, apply to LOBPCG—but not to (2.1)–(2.2)). In practice, the convergence of algorithms using this and other block CG schemes is considerably faster: the numerical tests of [18] suggest that in (4.9), κ_{ij} can be replaced with its square root (cf. a similar hypothesis in [9]). Still, the result in question adequately portrays the observed convergence behavior of the block Jacobi scheme in three important respects: (i) the convergence is not adversely affected by eigenvalue clustering, (ii) neither is it affected by the spread of the spectrum of L, provided that a good enough preconditioner K is used, and (iii) the convergence to leftmost eigenvalues is faster than to those deeper in the spectrum; hence the performance of the algorithm benefits substantially from the deflation technique (unlike with the trace minimization scheme (2.1), (2.2), (2.5); cf. [18]).

Appendix A. The Rayleigh–Ritz approximation. For the reader's convenience, section A.1 summarizes the Rayleigh–Ritz procedure; section A.2 presents new residual error bounds for the Rayleigh–Ritz approximation that are used in the proofs of Appendix B.

A.1. The Rayleigh–Ritz procedure. Denote by $P_{\mathcal{V}}$ the orthogonal projector onto a subspace $\mathcal{V}$. If $\mathcal{V}$ has a good approximation to x_j, then $\tilde{x}_j = P_{\mathcal{V}}x_j$ is close to x_j, and the norm of the residual $r_j = L\tilde{x}_j - \lambda(\tilde{x}_j)\tilde{x}_j$ is small, and hence so is the norm of $r_j^{\mathcal{V}} = P_{\mathcal{V}}r_j = P_{\mathcal{V}}L\tilde{x}_j - \lambda(\tilde{x}_j)\tilde{x}_j$. Denoting by $L_{\mathcal{V}}$ the restriction of the operator $P_{\mathcal{V}}LP_{\mathcal{V}}$ to $\mathcal{V}$, we observe that for $v \in \mathcal{V}$ the Rayleigh quotient $(L_{\mathcal{V}}v,v)/\|v\|^2$ coincides with $\lambda(v)$, and thus $r_j^{\mathcal{V}} = L_{\mathcal{V}}\tilde{x}_j - \lambda(\tilde{x}_j)\tilde{x}_j$ is the residual of $\tilde{x}_j$ viewed as an approximation to an eigenvector of $L_{\mathcal{V}}$. By the well-known results from eigenvalue approximation theory (see, e.g., [20]), the smallness of $r_j^{\mathcal{V}}$ implies that $\tilde{x}_j$ is close to an invariant subspace of $L_{\mathcal{V}}$. Thus, if some eigenvectors of L are close to $\mathcal{V}$, then one can find approximations to those eigenvectors in the invariant subspaces of $L_{\mathcal{V}}$.

The eigenvectors $x_j^{\mathcal{V}}$ of $L_{\mathcal{V}}$ are called *Ritz vectors* of L in the *trial subspace* $\mathcal{V}$, and the corresponding eigenvalues $\lambda_j^{\mathcal{V}} = \lambda(x_j^{\mathcal{V}})$ are called Ritz values. Since $0 =$

$L_{\mathcal{V}}x_j^{\mathcal{V}} - \lambda_j^{\mathcal{V}}x_j^{\mathcal{V}} = P_{\mathcal{V}}(Lx_j^{\mathcal{V}} - \lambda(x_j^{\mathcal{V}})x_j^{\mathcal{V}})$, the gradient of $\lambda(v)$ at $v = x_j^{\mathcal{V}}$ is orthogonal to $\mathcal{V}$. Conversely, if the gradient of $\lambda(v)$ is orthogonal to $\mathcal{V}$, then $L_{\mathcal{V}}v - \lambda(v)v = P_{\mathcal{V}}(Lv - \lambda(v)v) = 0$; i.e., v is a Ritz vector. Hence, Ritz vectors can be equivalently defined as the vectors on which the Rayleigh quotient restricted to $\mathcal{V}$ is stationary. Further, since $x_j^{\mathcal{V}}$ are eigenvectors of $L_{\mathcal{V}}$, $(x_j^{\mathcal{V}}, x_k^{\mathcal{V}}) = 0$ and $(Lx_j^{\mathcal{V}}, x_k^{\mathcal{V}}) = (L_{\mathcal{V}}x_j^{\mathcal{V}}, x_k^{\mathcal{V}}) = 0$; again, it is not difficult to verify that these two orthogonality properties, shared with the eigenvectors of L, represent yet another equivalent definition of Ritz vectors.

In practical terms, Ritz vectors and values are computed by what is known as the Rayleigh–Ritz procedure. A basis $\{v_1, \ldots, v_m\}$ is introduced in $\mathcal{V}$, and two m-by-m Gram matrices $\hat{A}$ and $\hat{B}$ with entries (Lv_j, v_k) and (v_j, v_k) are computed. Since $x_j^{\mathcal{V}} \in \mathcal{V}$, we have $x_j^{\mathcal{V}} = V\hat{x}_j$, where $V = [v_1, \ldots, v_m]$ and $\hat{x}_j$ is a column vector of dimension m. The orthogonality of $r(x_j^{\mathcal{V}})$ to $\mathcal{V}$ implies $0 = V^*(LV\hat{x}_j - \lambda(x_j^{\mathcal{V}})V\hat{x}_j) = \hat{A}\hat{x}_j - \lambda_j^{\mathcal{V}}\hat{B}\hat{x}_j$; i.e., the Ritz values $\lambda_j^{\mathcal{V}}$ are the eigenvalues of the generalized problem

$$\hat{A}\hat{x} = \lambda^{\mathcal{V}}\hat{B}\hat{x}, \tag{A.1}$$

and the Ritz vectors $x_j^{\mathcal{V}}$ are related to the eigenvectors $\hat{x}_j$ of this problem as $x_j^{\mathcal{V}} = V\hat{x}_j$.

It remains to note that the Rayleigh–Ritz procedure can also be applied to the generalized problem $Lx = \lambda Mx$ with Hermitian positive definite M, the only difference being the definition of the entries of the Gram matrix $\hat{B}$ as (Mv_j, v_k); we note that in this case Ritz vectors are M-orthogonal. Finally, we note that the scalar product that has been used in this section does not have to be the standard Euclidean one: any scalar product in which L and M are Hermitian can be used instead.

A.2. Lehmann intervals and related residual error bounds. By the minimax principle, Ritz values are not less than the leftmost eigenvalues with respective indices. Hence, one way to obtain estimates for the difference between the two is to find lower bounds for the eigenvalues of L, for which a method suggested by Lehmann can be employed (see, e.g., [11, 20, 2, 7]).

Lehmann's method produces, for a given subspace $\mathcal{V}$ and a given value α that is not a Ritz value in $\mathcal{V}$, a set of values

$$\cdots \le \lambda_{-2}^{\mathcal{V}}(\alpha) \le \lambda_{-1}^{\mathcal{V}}(\alpha) < \alpha < \lambda_1^{\mathcal{V}}(\alpha) \le \lambda_2^{\mathcal{V}}(\alpha) \le \cdots \tag{A.2}$$

such that each semiclosed interval $[\lambda_{-k}^{\mathcal{V}}(\alpha), \alpha)$ or $(\alpha, \lambda_k^{\mathcal{V}}(\alpha)]$ contains at least k eigenvalues of L; here we refer to α as the Lehmann pole and to $\lambda_k^{\mathcal{V}}(\alpha)$ as Lehmann values. The number of Lehmann values left and right of the pole coincides with the respective number of Ritz values, each Lehmann value being further away from the pole than the respective Ritz value. The described properties of Lehmann values imply that if $\lambda_m^{\mathcal{V}} < \alpha \le \lambda_{m+1}$, then $\lambda_{i-m-1}^{\mathcal{V}}(\alpha) \le \lambda_i \le \lambda_i^{\mathcal{V}}$, $i = 1, \ldots, m$.

Lehmann's method is a product of the following reasoning (cf. [11]). Let α be any value that is not a Ritz value of L in $\mathcal{V}$, and assume for a moment that neither is it an eigenvalue of L. The eigenvalues of the shifted operator $L - \alpha$ are $\lambda_j - \alpha$, and they coincide with those of the generalized eigenvalue problem

$$(L - \alpha)^2 y = \nu(L - \alpha)y, \tag{A.3}$$

which is equivalent to

$$(L - \alpha)y = \mu(L - \alpha)^2 y, \tag{A.4}$$

the eigenvalues of the latter being the inverses of those of the former. We observe that both problems are definite owing to the assumptions on α. Let us enumerate eigenvalues μ_j of (A.4) as follows: positive eigenvalues are $\mu_1 \geq \mu_2 \geq \cdots$ (starting from the rightmost) and negative are $\mu_{-1} \leq \mu_{-2} \leq \cdots$ (starting from the leftmost). Denote by $\mu_k^{\mathcal{V}}$ the Ritz values of (A.4) in $\mathcal{V}$ enumerated in the same way. By the minimax principle, $\mu_k^{\mathcal{V}} \leq \mu_k$ for positive k and $\mu_k^{\mathcal{V}} \geq \mu_k$ for negative k. Recalling that the eigenvalues of (A.4) are the inverses of the eigenvalues of (A.3), which in turn are the eigenvalues of L shifted by α, for $\lambda_k^{\mathcal{V}}(\alpha) = 1/\mu_k^{\mathcal{V}} + \alpha$ we have

$$\lambda_{(k)} \leq \lambda_k^{\mathcal{V}}(\alpha), \quad k > 0; \quad \lambda_{(k)} \geq \lambda_k^{\mathcal{V}}(\alpha), \quad k < 0, \tag{A.5}$$

where $\lambda_{(1)} \leq \lambda_{(2)} \leq \cdots$ are eigenvalues of L right of α and $\lambda_{(-1)} \geq \lambda_{(-2)} \geq \cdots$ are those left of α.

Observing that $\lambda_k^{\mathcal{V}}(\alpha)$ continuously depend on α in any interval not containing Ritz values of L in $\mathcal{V}$, we conclude that (A.5) is also valid when α is an eigenvalue of L (for as long as it is not a Ritz value at the same time, i.e., for as long as $\mathcal{V}$ does not contain the corresponding eigenvector).

Denote $X_{\mathcal{V}} = [x_1^{\mathcal{V}}, \ldots, x_m^{\mathcal{V}}]$, $R_{\mathcal{V}} = [r_1^{\mathcal{V}}, \ldots, r_m^{\mathcal{V}}]$, and $\Lambda_{\mathcal{V}} = \operatorname{diag}\{\lambda_1^{\mathcal{V}}, \ldots, \lambda_m^{\mathcal{V}}\}$, where $x_j^{\mathcal{V}}$ and $\lambda_j^{\mathcal{V}}$ are normalized Ritz vectors and Ritz values of L in $\mathcal{V}$, and $r_j^{\mathcal{V}} = Lx_j^{\mathcal{V}} - \lambda_j^{\mathcal{V}} x_j^{\mathcal{V}}$. Simple calculation shows that the eigenvalues of the operator

$$L(\alpha) = \alpha + [X_{\mathcal{V}}, R_{\mathcal{V}}] \begin{bmatrix} \Lambda_{\mathcal{V}} - \alpha & I \\ I & (\Lambda_{\mathcal{V}} - \alpha)^{-1} \end{bmatrix} \begin{bmatrix} X_{\mathcal{V}}^* \\ R_{\mathcal{V}}^* \end{bmatrix} \tag{A.6}$$

other than α coincide with its Lehmann values, and that $L(\alpha)w = Lw$ for any $w \in \mathcal{V}$. Hence, Lehmann bounds are sharp; i.e., they cannot be improved if the only information available is $\mathcal{V}$ and L-images of vectors from $\mathcal{V}$. We observe also that Lehmann values can be alternatively defined as eigenvalues of $L(\alpha)$ other than α.

It is not difficult to verify that $L(\alpha_1) \leq L(\alpha_2)$ for $\alpha_1 \leq \alpha_2$ (where $A \leq B$ means $B - A$ is positive semidefinite), which implies that Lehmann values are nondecreasing functions of α. For this reason, assuming $\lambda_m^{\mathcal{V}} < \lambda_{m+1}$, the optimal position of the pole, as far as the m leftmost eigenvalues are concerned, is $\alpha = \lambda_{m+1}$, as this is the rightmost value of α that provides guaranteed left bounds for $\lambda_1, \ldots, \lambda_m$.

From the practical point of view, Lehmann values can be computed using Ritz values and residuals. Indeed, applying to (A.4) the Rayleigh–Ritz procedure with Ritz vectors as the basis, ones gets, after simple calculation, the generalized eigenvalue problem (cf. [2])

$$(\Lambda_{\mathcal{V}} - \alpha)\hat{x} = \hat{\mu}\left((\Lambda_{\mathcal{V}} - \alpha)^2 + R_{\mathcal{V}}^* R_{\mathcal{V}}\right)\hat{x}. \tag{A.7}$$

Simple manipulations with (A.7) show that in the case $\alpha > \lambda_m^{\mathcal{V}}$, Lehmann values are the eigenvalues of the matrix

$$\hat{L}_{\alpha,-} = \Lambda_{\mathcal{V}} - S_{\mathcal{V},\alpha,-}^* S_{\mathcal{V},\alpha,-}, \quad S_{\mathcal{V},\alpha,-} = R_{\mathcal{V}}(\alpha - \Lambda_{\mathcal{V}})^{-1/2}, \tag{A.8}$$

and in the case $\alpha < \lambda_1^{\mathcal{V}}$, of the matrix

$$\hat{L}_{\alpha,+} = \Lambda_{\mathcal{V}} + S_{\mathcal{V},\alpha,+}^* S_{\mathcal{V},\alpha,+}, \quad S_{\mathcal{V},\alpha,+} = R_{\mathcal{V}}(\Lambda_{\mathcal{V}} - \alpha)^{-1/2}. \tag{A.9}$$

We observe that $\hat{L}_{+\infty,-} = \hat{L}_{-\infty,+} = \Lambda_{\mathcal{V}}$; i.e., $\lambda_{j-m-1}^{\mathcal{V}}(+\infty) = \lambda_j^{\mathcal{V}}(-\infty) = \lambda_j^{\mathcal{V}}$.

The established relationship between Lehmann values and (weighted) residuals facilitates the derivation of various residual error bounds for Ritz values. For example, assuming $\lambda_m^{\mathcal{V}} < \lambda_{m+1}$ and taking the trace of both sides of (A.8) for $\alpha = \lambda_{m+1}$, we arrive at the following total eigenvalue error estimate:

$$\sum_{j=1}^{m}(\lambda_j^{\mathcal{V}} - \lambda_j) \le \sum_{j=1}^{m}\lambda_j^{\mathcal{V}} - \sum_{j=-m}^{-1}\lambda_j^{\mathcal{V}}(\lambda_{m+1}) = \sum_{j=1}^{m}\frac{\|r_j^{\mathcal{V}}\|^2}{\lambda_{m+1} - \lambda_j^{\mathcal{V}}}. \tag{A.10}$$

The eigenvalues of the diagonal block of $\hat{L}_{\lambda_{m+1},-}$ on the intersection of several last rows and columns are not greater than the respective rightmost eigenvalues of this matrix, which implies that in (A.10) the summation can actually start at any value $1 \le k \le m$; i.e.,

$$\sum_{j=k}^{m}(\lambda_j^{\mathcal{V}} - \lambda_j) \le \sum_{j=k}^{m}\frac{\|r_j^{\mathcal{V}}\|^2}{\lambda_{m+1} - \lambda_j^{\mathcal{V}}}. \tag{A.11}$$

If α is placed to the right of the spectrum of L, then all eigenvalues of (A.4) are negative, with those closest to zero corresponding to the leftmost eigenvalues of L. The fact that the Ritz values for (A.4) do not exceed corresponding eigenvalues implies that the Lehmann values in this case are not smaller than the corresponding eigenvalues of L, whereas by (A.8) they are not greater than the corresponding Ritz values. Hence, taking the trace of both sides in (A.8), we arrive at the following estimate of [16]:

$$\sum_{j=1}^{m}\frac{\|r_j\|^2}{\alpha - \tilde{\lambda}_j} \le \sum_{j=1}^{m}(\tilde{\lambda}_j - \lambda_j) \tag{A.12}$$

that is valid for arbitrary α to the right of the spectrum of L; by continuity, it remains valid for α at the rightmost point of the spectrum as well. Just as with (A.10) and (A.11), the number of items in both sums in (A.12) can be simultaneously reduced, this time by lowering the upper summation limit.

It remains to note that the above residual bounds for eigenvalue errors can be applied to the generalized eigenvalue problem $Lx = \lambda Mx$ with Hermitian positive definite M as well. One just needs to consider an equivalent standard eigenvalue problem for the operator $A = M^{-1}L$ and use the scalar product $(\cdot,\cdot)_M = (M\cdot,\cdot)$, in which A is Hermitian. It is not difficult to verify that the Ritz values and vectors of A coincide with those of the generalized problem in question, and that the M-norm of the residual vector $A\tilde{x} - \tilde{\lambda}\tilde{x}$ is equal to the M^{-1}-norm of the residual vector $L\tilde{x} - \tilde{\lambda}M\tilde{x}$; hence, in the case at hand, we need to use M^{-1}-norms of residuals in the above estimates.

Appendix B. Proofs. Below, for any m-by-m matrix A and any $\mathcal{I}, \mathcal{J} \subset \{1, \dots, m\}$, we denote by $A_{\mathcal{I},\mathcal{J}}$ the block (submatrix) of A on the intersection of rows with indices from $\mathcal{I}$ and columns with indices from $\mathcal{J}$. For any matrix U with m columns, $U_{\mathcal{J}}$ stands for the matrix formed by columns of U with indices from $\mathcal{J}$.

In order to make the calculations more compact, we use the following notation. For any matrices U and V of the same size we denote $\langle U, V\rangle = \mathrm{Tr}V^*U$. We observe that the latter bilinear form is a scalar product (it is actually the standard Euclidean scalar product of two vectors formed by columns of U and V), and the corresponding norm $\langle U\rangle$ is just the Frobenius norm of U. Elementary calculation shows that for any matrices A and B for which the product AUB makes sense, one has $\langle AUB\rangle \le$

$\|A\|\langle U\rangle\|B\|$. Here and below, $\|A\|$ stands for the operator norm of A, i.e., $\|A\| = \sup_u \|Au\|/\|u\|$; we note that $\|A\| \le \langle A\rangle$. For any square matrix H of the same size as the height of U and V we denote $\langle U, V\rangle_H = \mathrm{Tr} V^* H U$. If H is Hermitian positive definite, then the latter is again a scalar product, and we denote by $\langle U\rangle_H$ the corresponding norm (often referred to as the energy norm). Since $\langle U\rangle_H = \langle H^{1/2}U\rangle$, we have $\langle UA\rangle_H \le \langle U\rangle_H \|A\|$ for any A that can be multiplied by U, and $|\langle U, V\rangle| \le \langle U\rangle_H \langle V\rangle_{H^{-1}}$.

For matrices U and V of the same size as E and columns orthogonal to $\tilde{\mathcal{X}}$, we denote

$$\langle U_{\mathcal{I}}, V_{\mathcal{I}}\rangle_J = \sum_{j\in\mathcal{I}} (J_j u_j, v_j), \quad \langle U_{\mathcal{I}}, V_{\mathcal{I}}\rangle_{J^{-1}} = \sum_{j\in\mathcal{I}} (J_j^{-1} u_j, v_j),$$

where $\mathcal{I} \subset \{1, \dots, m\}$ and J_j^{-1} is the inverse of the restriction of J_j to the subspace $\tilde{\mathcal{X}}^\perp$ (orthogonal complement to $\tilde{\mathcal{X}}$): here and below we assume (3.13) to hold. Again, the above are two scalar products in the latter subspace, and we denote by $\langle U\rangle_J$ and $\langle U\rangle_{J^{-1}}$ the corresponding norms. Owing to the equivalence of J_j-norms (cf. (3.11)), these norms are equivalent to J_j and J_j^{-1} norms of U, respectively, for any $1 \le j \le m$, and it is not difficult to verify that

$$\langle U\rangle_{J_j} \langle V\rangle_{J_j^{-1}} \le \sqrt{\rho_1} \langle U\rangle_J \langle V\rangle_{J^{-1}}, \tag{B.1}$$

where $\rho_1 = (\lambda_{m+1} - \lambda_1)/\gamma = \rho_m/\delta_m$. Finally, we observe that for any matrices U and V of the same height as E we have

$$\|V^*U\| \le \langle V^*U\rangle = \langle (J_1^{\frac12} V)^* J_1^{-\frac12} U\rangle \le \langle V\rangle_{J_1} \langle U\rangle_{J_1^{-1}} \le \sqrt{\rho_1} \langle V\rangle_J \langle U\rangle_{J^{-1}}. \tag{B.2}$$

Remark 1. It is not difficult to verify that (B.1) and (B.2) remain valid if we replace J_j with $J(\mu_j, \tilde{\mathcal{X}})$, where $\lambda_1 \le \mu_j \le \tilde{\lambda}_m$, everywhere.

B.1. Proof of Theorem 3.2. First we note that the assumptions of the theorem together with the inequality $\sin\{\mathcal{X}; \tilde{\mathcal{X}}\} \le (\tilde{\lambda}_m - \lambda_1)/(\lambda_{m+1} - \lambda_1)$ (cf., e.g., [7]) ensure that $\sin\{\mathcal{X}; \tilde{\mathcal{X}}\} < 1$, and hence we are in a position to use the notation and relationships of section 3.1.

The proof of inequalities (3.14) and (3.15) is based on the following two observations: (i) the eigenvalues of $\lambda(X')$ are $\lambda_1, \dots, \lambda_m$, and those of $\lambda(\tilde{X})$ are $\tilde{\lambda}_1, \dots, \tilde{\lambda}_m$, and hence (3.4) implies

$$\sum_{j=1}^m (\tilde{\lambda}_j - \lambda_j) = \mathrm{Tr}(\lambda(\tilde{X}) - \lambda(X')) = -\mathrm{Tr} R^* E = -\sum_{j=1}^m (e_j, r_j), \tag{B.3}$$

and (ii) the Riccati equation (3.5) is a set of m linear systems

$$(J_j - E R^*) e_j = -r_j, \tag{B.4}$$

linked by the asymptotically small operator perturbation ER^*; hence $J_j e_j \approx -r_j$, which leads, via calculations given below, to the asymptotic identities of the theorem.

Below we denote $\mathcal{I} = \{1, \dots, l\}$, $\delta_\lambda = \delta_\lambda(\tilde{\mathcal{X}})$, and $\delta_{\lambda,\mathcal{I}} = \sum_{j\in\mathcal{I}} (\tilde{\lambda}_j - \lambda_j)$.

To guide the reader through the rather long and technical proof below, we split it into several steps.

Step 1. *We show that* $\|R\|_{J^{-1}}^2 = \mathcal{O}(\delta_\lambda)$. In order to obtain the estimate in question, we apply the results of section A.2 to an auxiliary eigenvalue problem

$-L_0^{-1}x = \omega x$, where $L_0 = L - \lambda_0$ and $\lambda_0 < \lambda_1$ is specified later on. The reason for introducing this auxiliary problem rather than applying the estimates of section A.2 to the original problem is that in the latter case we would end up with coefficients that depend on the spread of L (cf. (A.12)), which is very large in many applications. Working the auxiliary problem just introduced allows us to eliminate this quantity from our calculation.

We observe that the eigenvectors of $-L_0^{-1}$ coincide with those of L and the corresponding eigenvalues are $\omega_j = 1/(\lambda_0 - \lambda_j)$. Replacing the standard scalar product with L_0-energy product $(\cdot,\cdot)_{L_0} = (L_0\cdot,\cdot)$, we observe that $-L_0^{-1}$ is Hermitian in the new product, the Ritz vectors of $-L_0^{-1}$ in $\tilde{\mathcal{X}}$ are $\tilde{x}_j$, and the corresponding Ritz values $\tilde{\omega}_j = 1/(\lambda_0 - \tilde{\lambda}_j)$. Hence, denoting $s_j = -L_0^{-1}\tilde{x}_j - \tilde{\omega}_j\tilde{x}_j$ (and bearing in mind that $\tilde{x}_j$ are not L_0-normalized), we derive from (A.12) with $\alpha = 0$ the inequality

$$-\sum_{j=1}^{m} \frac{\|s_j\|_{L_0}^2}{\tilde{\omega}_j \|\tilde{x}_j\|_{L_0}^2} = \sum_{j=1}^{m} \tilde{\omega}_j^2 \|r_j\|_{L_0^{-1}}^2 \le \sum_{j=1}^{m} (\tilde{\omega}_j - \omega_j), \tag{B.5}$$

which, after switching ω's to λ's, yields

$$\langle R\rangle_{L_0^{-1}} = \sum_{j=1}^{m} \|r_j\|_{L_0^{-1}}^2 \le \left(\frac{\lambda_{m+1} - \lambda_0}{\lambda_1 - \lambda_0}\right)^2 \sum_{j=1}^{m} (\tilde{\lambda}_j - \lambda_j). \tag{B.6}$$

Setting $\lambda_0 = \lambda_1 - (\lambda_{m+1} - \lambda_1)$, we arrive at the estimate $\langle R\rangle_{L_0^{-1}}^2 \le 4\delta_\lambda$. In order to obtain an estimate for the J^{-1}-norm of R, consider two auxiliary systems $L_0 u = r$ and $J_0 v = r$, where $J_0 = J(\lambda_0, \tilde{\mathcal{X}})$ and $v \in \tilde{\mathcal{X}}^\perp$, for an arbitrary $r \in \tilde{\mathcal{X}}^\perp$. Since $(J_0 v, v) = (L_0 v, v)$, we have $(L_0(v-u), v) = 0$, which implies $(L_0 v, v) \le (L_0 u, u)$. Recalling that $u = L_0^{-1} r$ and $v = J_0^{-1} r$, we arrive at $\|r\|_{J_0^{-1}} \le \|r\|_{L_0^{-1}}$, and, thus, via (3.12), at

$$\langle R\rangle_{J^{-1}}^2 \le 2\rho_1 \langle R\rangle_{J_0^{-1}}^2 \le 8\rho_1 \delta_\lambda. \tag{B.7}$$

Step 2. *We prove* (3.14) *and* (3.15) *for* $l = m$. Let us denote $\theta = \|(1-\tilde{P})P\|$ and $\theta_F = \langle (1-\tilde{P})P\rangle$, where P and $\tilde{P}$ are orthogonal projectors onto $\mathcal{X}$ and $\tilde{\mathcal{X}}$ respectively. We observe that θ is the sine of the (maximal) angle between $\mathcal{X}$ and $\tilde{\mathcal{X}}$, and we recall that in section 3.3 we have shown that $\|E\|$ is the tangent of this angle; i.e., $\|E\|^2 = \theta^2/(1-\theta^2)$, which implies $\|X'\|^2 = 1/(1-\theta^2)$. Lemma 4 of [15] implies

$$\theta_F^2 \le \frac{\delta_\lambda}{\lambda_{m+1} - \lambda_m}.$$

Owing to the assumption (3.13), $\delta_\lambda \le \lambda_{m+1} - \lambda_m - \gamma$, and, hence, $\theta^2 \le \theta_F^2 \le 1 - \gamma/(\lambda_{m+1} - \lambda_m)$. Thus,

$$\|E\|^2 \le \langle E\rangle^2 = \langle (1-\tilde{P})X'\rangle^2 = \langle (1-\tilde{P})PX'\rangle^2 \le \theta_F^2 \|X'\|^2 = \frac{\theta_F^2}{1-\theta^2} \le \frac{\delta_\lambda}{\gamma}.$$

From (3.5) we have $\langle E\rangle_J^2 = -\langle R, E\rangle + \langle ER^*E, E\rangle$. We have shown above (see (B.3)) that the first term in the right-hand side is equal to δ_λ. The second can be estimated as follows:

$$|\langle ER^*E, E\rangle| = |\operatorname{Tr} E^*ER^*E| = |\langle E, RE^*E\rangle| \le \langle E\rangle_{J_1} \langle RE^*E\rangle_{J_1}^{-1}$$

$$\le \langle E\rangle_{J_1} \langle R\rangle_{J_1}^{-1} \|E\|^2 \le \sqrt{\rho_1}\langle E\rangle_J \langle R\rangle_{J^{-1}} \frac{\delta_\lambda}{\gamma} \le \frac{1}{2}\langle E\rangle_J^2 + 4\rho_1^2 \frac{\delta_\lambda^3}{\gamma^2},$$

and thus,

$$\text{(B.8)}\quad \langle E\rangle_J^2 \le 2\left(1+4\rho_1^2\frac{\delta_\lambda^2}{\gamma^2}\right)\delta_\lambda \le 2\left(1+4\rho_1^2\rho_m^2\right)\delta_\lambda = 2\left(1+4\frac{\rho_m^4}{\delta_m^2}\right)\delta_\lambda \equiv c_1\delta_\lambda,$$

which implies

$$|\langle ER^*E,E\rangle| \le \sqrt{\rho_1}\langle E\rangle_J\langle R\rangle_{J^{-1}}\frac{\delta_\lambda}{\gamma} \le 2\rho_1\sqrt{2c_1}\frac{\delta_\lambda^2}{\gamma} = 2\frac{\rho_m^2}{\delta_m}\sqrt{2c_1}\frac{\delta_\lambda^2}{\lambda_{m+1}-\lambda_m},$$

and we arrive at (3.14) with $l=m$.

In order to obtain the respective estimate for $\langle R\rangle_{J^{-1}}$, we multiply (B.4) by $-J_j^{-1}r_j$ and take the sum over $j=1,\dots,m$ to arrive at the identity $-\langle E,R\rangle = \langle R\rangle_{J^{-1}} - \langle ER^*E,R\rangle_{J^{-1}}$. The left-hand side of the above identity is equal to δ_λ, and the second term in the right-hand side can be estimated, with the help of (3.10) and (B.2), as follows:

$$|\langle ER^*E,R\rangle_{J^{-1}}| \le \frac{1}{\gamma}\langle ER^*E\rangle\langle R\rangle_{J^{-1}} \le \frac{1}{\gamma}\langle E\rangle\|R^*E\|\langle R\rangle_{J^{-1}} \le \frac{\sqrt{\rho_1}}{\gamma}\langle E\rangle\langle E\rangle_J\langle R\rangle_{J^{-1}}^2,$$

and, via previously obtained estimates for $\langle E\rangle$, $\langle E\rangle_J$, and $\langle R\rangle_{J^{-1}}^2$, we arrive at (3.15) with $l=m$.

Step 3. *We show that* $\langle E_{\mathcal{I}}\rangle_J = \mathcal{O}(\delta_{\lambda,\mathcal{I}})$. In order to obtain such an estimate, we take advantage of the results of [15] applied to the auxiliary eigenvalue problem for the shifted operator L_0 introduced earlier (again, our motivation here is to avoid any dependency on the spread of L). Let us denote by $\tilde P_0$ the orthogonal projector onto $\tilde{\mathcal{X}}$ in the sense of the scalar product $(\cdot,\cdot)_{L_0}$. For $\theta_0 = \|(1-\tilde P_0)P\|_{L_0}$, where P is the same as above (note that P is orthogonal in both the standard and the L_0-product), we have

$$\theta_0^2 \le \frac{\tilde\omega_m-\omega_1}{\omega_{m+1}-\omega_1} < 1, \quad 1-\theta_0^2 \ge \frac{\omega_{m+1}-\tilde\omega_m}{\omega_{m+1}-\omega_1} = \frac{\lambda_{m+1}-\tilde\lambda_m}{\tilde\lambda_m-\lambda_0} \ge \frac{\gamma}{2(\lambda_{m+1}-\lambda_1)} = \frac{1}{2\rho_1}.$$

Observing that $\|\tilde P_0 x\|_{L_0}^2 \ge (1-\theta_0^2)\|x\|_{L_0}^2$ for any $x\in\mathcal{X}$, we conclude that the restriction of $\tilde P_0$ to $\mathcal{X}$ is invertible, and the norm of its inverse, which we denote by Q_0, is not greater than $1/\sqrt{1-\theta^2} \le \sqrt{2\rho_1}$. Denote $\mathcal{X}_{\mathcal{I}} = \operatorname{span}\{x_j\}_{j\in\mathcal{I}}$, $\tilde{\mathcal{X}}_{\mathcal{I}} = \operatorname{span}\{x_j\}_{j\in\mathcal{I}}$, and $\mathcal{Y}_{\mathcal{I}} = \operatorname{span}\{(1-\tilde P_0)x_j\}_{j\in\mathcal{I}}$, and denote by $P_{\mathcal{I}}$, $\tilde P_{\mathcal{I}}$, and S the L_0-orthogonal projectors onto these three subspaces. Lemma 4 of [15] implies the estimate

$$\text{(B.9)}\quad \langle(1-P_{\mathcal{I}})\tilde P_{\mathcal{I}}\rangle_{L_0}^2 = \langle(1-\tilde P_{\mathcal{I}})P_{\mathcal{I}}\rangle_{L_0}^2 \le \frac{1}{\omega_{l+1}-\omega_l}\sum_{j=1}^{l}(\tilde\omega_j-\omega_j) \le \frac{4\delta_{\lambda,\mathcal{I}}}{\lambda_{l+1}-\lambda_l}.$$

In order to take advantage of these estimates, we observe that

$$E_{0,\mathcal{I}} \equiv (Q_0-1)\tilde X_{\mathcal{I}} = (Q_0-1)S\tilde X_{\mathcal{I}} + (Q_0-1)(1-S)\tilde X_{\mathcal{I}}.$$

By the definition of Q_0, $Q_0-1 = (1-\tilde P_0)Q_0$, and since $\mathcal{Y}_{\mathcal{I}}$ is the projection of $\mathcal{X}_{\mathcal{I}}$, we have $Q_0S = P_{\mathcal{I}}Q_0S$. Thus, $(Q_0-1)S\tilde X_{\mathcal{I}} = (1-\tilde P_0)P_{\mathcal{I}}Q_0S\tilde X_{\mathcal{I}}$, and hence

$$\langle(Q_0-1)S\tilde X_{\mathcal{I}}\rangle_{L_0} \le \langle(1-\tilde P_0)P_{\mathcal{I}}\rangle_{L_0}\|Q_0\|_{L_0}\|S\|_{L_0}\|\tilde X_{\mathcal{I}}\|_{L_0} \le 4\frac{\sqrt{\rho_1(\lambda_{m+1}-\lambda_1)\delta_{\lambda,\mathcal{I}}}}{\sqrt{\lambda_{l+1}-\lambda_l}},$$

where we have used the fact that $\|u\|_{L_0}^2 \le (\tilde\lambda_m - \lambda_0)\|u\|^2 \le 2(\lambda_{m+1}-\lambda_1)$ for $u \in \tilde{\mathcal{X}}$. Further, $Q_0 - 1 = (1-\tilde P_0)Q_0 = (1-\tilde P_0)PQ_0$, and hence

$$\begin{aligned}\langle (Q_0-1)(1-S)\tilde X_{\mathcal I}\rangle_{L_0} &\le \|(1-\tilde P_0)P\|_{L_0}\|Q_0\|_{L_0}\langle (1-S)\tilde X_{\mathcal I}\rangle_{L_0}\\ &\le \sqrt{2\rho_1}\langle (1-S)\tilde X_{\mathcal I}\rangle_{L_0}.\end{aligned}$$

In order to estimate $\langle (1-S)\tilde X_{\mathcal I}\rangle_{L_0}$, we observe that for any $u, v \in \tilde{\mathcal X}_{\mathcal I}$, $\|(1-S)u\|_{L_0} \le \|u-v\|_{L_0}$, which implies $\|(1-S)u\|_{L_0} \le \|(1-SP_{\mathcal I})u\|_{L_0}$, and hence $\langle (1-S)\tilde X_{\mathcal I}\rangle_{L_0} \le \langle (1-SP_{\mathcal I})\tilde X_{\mathcal I}\rangle_{L_0}$. Recalling again that $\mathcal{Y}_{\mathcal I}$ is the projection of $\mathcal{X}_{\mathcal I}$, we observe that $SP_{\mathcal I} = \tilde P_0 P_{\mathcal I}$. Hence, $1 - SP_{\mathcal I} = 1 - P_{\mathcal I} + (1-S)P_{\mathcal I} = 1 - P_{\mathcal I} + (1-\tilde P_0)P_{\mathcal I}$, and thus $U = (1-SP_{\mathcal I})\tilde X_{\mathcal I} = (1-P_{\mathcal I})\tilde X_{\mathcal I} + (1-\tilde P_0)P_{\mathcal I}\tilde X_{\mathcal I} = (1-P_{\mathcal I})\tilde X_{\mathcal I} + V$. Observing that $V^* L_0 U = 0$, we conclude that $\langle (1-SP_{\mathcal I})\tilde X_{\mathcal I}\rangle_{L_0} \le \langle (1-P_{\mathcal I})\tilde X_{\mathcal I}\rangle_{L_0}$. Thus,

$$\langle (1-S)\tilde X_{\mathcal I}\rangle_{L_0} \le \langle (1-P_{\mathcal I})\tilde X_{\mathcal I}\rangle_{L_0} \le \langle (1-P_{\mathcal I})\tilde P_{\mathcal I}\rangle \|\tilde X_{\mathcal I}\|_{L_0} \le \sqrt{8\frac{\lambda_{m+1}-\lambda_1}{\lambda_{l+1}-\lambda_l}\delta_{\lambda,\mathcal I}},$$

and we arrive at the estimate

$$\langle E_{0,\mathcal I}\rangle_{L_0}^2 \le c_2 \delta_{\lambda,\mathcal I}, \tag{B.10}$$

where the coefficient c_2 depends on ρ_m, δ_m, and δ_l only. Let us now estimate $\langle E_{\mathcal I} - E_{0,\mathcal I}\rangle_{L_0}$. It is easy to see that $\tilde P = \tilde X \tilde X^*$ and $\tilde P_0 = \tilde X \tilde\Lambda_0^{-1}\tilde X^* L_0$, where $\tilde\Lambda_0$ is a diagonal matrix with $\tilde\lambda_j - \lambda_0$ on the diagonal. Since $L_0\tilde X - \tilde X\tilde\Lambda_0 = R$, we have $\tilde P_0 - \tilde P = \tilde X \tilde\Lambda_0^{-1} R^*$. Denoting by Q the inverse of the restriction of $\tilde P$ to $\mathcal{X}$, we observe that $X' = Q\tilde X$, and hence $E_{\mathcal I} - E_{0,\mathcal I} = (Q - Q_0)\tilde X_{\mathcal I} = Q(\tilde P_0 - \tilde P)Q_0\tilde X_{\mathcal I} = Q(\tilde P_0 - \tilde P)E_{0,\mathcal I} = Q\tilde X\tilde\Lambda_0^{-1}R^* E_{0,\mathcal I}$, and

$$\langle E_{\mathcal I} - E_{0,\mathcal I}\rangle_{L_0} \le \|X'\|_{L_0}\langle \tilde\Lambda_0^{-1}R\rangle_{L_0^{-1}}\langle E_{0,\mathcal I}\rangle_{L_0} \le \sqrt{\lambda_m - \lambda_0}\|X'\|\frac{\sqrt{\delta_\lambda}}{\lambda_1-\lambda_0}\langle E_{0,\mathcal I}\rangle_{L_0},$$

where we have used (B.5) for estimating $\langle \tilde\Lambda_0^{-1}R\rangle_{L_0^{-1}}$. Since $\|X'\| = 1/\sqrt{1-\theta^2} \le \sqrt{\rho_m}$ and $\delta_\lambda \le \lambda_{m+1} - \lambda_m \le \lambda_{m+1} - \lambda_1 = \lambda_1 - \lambda_0$, we end up with $\langle E_{\mathcal I} - E_{0,\mathcal I}\rangle_{L_0} \le \sqrt{2\rho_m}\langle E_{0,\mathcal I}\rangle_{L_0}$, which implies

$$\langle E_{\mathcal I}\rangle_J^2 \le 2\rho_1\langle E_{\mathcal I}\rangle_{L_0}^2 \le 2\rho_1\left(1+\sqrt{2\rho_m}\right)^2 c_2\delta_{\lambda,\mathcal I} \equiv c_3\delta_{\lambda,\mathcal I}. \tag{B.11}$$

Step 4. *We show that* $\langle R_{\mathcal I}\rangle_{J^{-1}} = \mathcal{O}(\delta_{\lambda,\mathcal I})$. We start by observing that (3.5) implies (cf. also (B.4))

$$\langle R_{\mathcal I}\rangle_{J^{-1}}^2 = -\langle E_{\mathcal I}, R_{\mathcal I}\rangle + \langle ER^*E_{\mathcal I}, R_{\mathcal I}\rangle_{J^{-1}} \le \langle E_{\mathcal I}\rangle_J\langle R_{\mathcal I}\rangle_{J^{-1}} + \langle ER^*E_{\mathcal I}\rangle_{J^{-1}}\langle R_{\mathcal I}\rangle_{J^{-1}},$$

and hence $\langle R_{\mathcal I}\rangle_{J^{-1}} \le \langle E_{\mathcal I}\rangle_J + \langle ER^*E_{\mathcal I}\rangle_{J^{-1}} \le \langle E_{\mathcal I}\rangle_J + \frac{1}{\sqrt{\gamma}}\langle ER^*E_{\mathcal I}\rangle$. With the help of (B.2), we have for $\langle ER^*E_{\mathcal I}\rangle$

$$\langle ER^*E_{\mathcal I}\rangle \le \langle E\rangle\|R^*E_{\mathcal I}\| \le \sqrt{\rho_1}\langle E\rangle\langle R\rangle_{J^{-1}}\langle E_{\mathcal I}\rangle_J \le \frac{\rho_1\sqrt{8}}{\sqrt{\gamma}}\delta_\lambda\langle E_{\mathcal I}\rangle_J, \tag{B.12}$$

and thus

$$\langle R_{\mathcal I}\rangle_{J^{-1}}^2 \le \left(1+\rho_1\sqrt{8}\frac{\delta_\lambda}{\gamma}\right)^2\langle E_{\mathcal I}\rangle_J^2 \le \left(1+\sqrt{8}\frac{\rho_m^2}{\delta_m}\right)^2 c_3\delta_{\lambda,\mathcal I} \equiv c_4\delta_{\lambda,\mathcal I}. \tag{B.13}$$

Step 5. *We show that* $\langle E_{\mathcal{I}}\rangle_J^2$, $\langle R_{\mathcal{I}}\rangle_{J^{-1}}^2$, *and* $-\langle E_{\mathcal{I}}, R_{\mathcal{I}}\rangle$ *are asymptotically identical.* From the correction equation (3.5) we have

$$\left|\langle E_{\mathcal{I}}\rangle_J^2 + \langle E_{\mathcal{I}}, R_{\mathcal{I}}\rangle\right| = |\langle ER^*E_{\mathcal{I}}, E_{\mathcal{I}}\rangle| \le \langle ER^*E_{\mathcal{I}}\rangle\langle E_{\mathcal{I}}\rangle \le \frac{1}{\sqrt{\gamma}}\langle ER^*E_{\mathcal{I}}\rangle\langle E_{\mathcal{I}}\rangle_J,$$

$$\left|\langle R_{\mathcal{I}}\rangle_{J^{-1}}^2 + \langle E_{\mathcal{I}}, R_{\mathcal{I}}\rangle\right| = |\langle ER^*E_{\mathcal{I}}, R_{\mathcal{I}}\rangle_{J^{-1}}| \le \frac{1}{\sqrt{\gamma}}\langle ER^*E_{\mathcal{I}}\rangle\langle R_{\mathcal{I}}\rangle_{J^{-1}}.$$

Hence, using (B.12) we obtain

$$\left|\langle E_{\mathcal{I}}\rangle_J^2 + \langle E_{\mathcal{I}}, R_{\mathcal{I}}\rangle\right| \le \sqrt{8}\frac{\rho_m^2}{\delta_m}c_3\frac{\delta_\lambda}{\lambda_{m+1}-\lambda_m}\delta_{\lambda,\mathcal{I}}, \tag{B.14}$$

$$\left|\langle R_{\mathcal{I}}\rangle_{J^{-1}}^2 + \langle E_{\mathcal{I}}, R_{\mathcal{I}}\rangle\right| \le \sqrt{8}\frac{\rho_m^2}{\delta_m}\sqrt{c_3c_4}\frac{\delta_\lambda}{\lambda_{m+1}-\lambda_m}\delta_{\lambda,\mathcal{I}}. \tag{B.15}$$

Step 6. *We show that* $-\langle E_{\mathcal{I}}, R_{\mathcal{I}}\rangle$ *and* $\delta_{\lambda,\mathcal{I}}$ *are asymptotically identical.* The identity in question has already been shown in the case $\mathcal{I} = \{1,\dots,m\}$ (see (B.3)). In what follows we assume the set $\mathcal{J} = \{1,\dots,m\}\setminus\mathcal{I}$ to be nonempty.

Denote $A = (X')^*LX'$, $B = (X')^*X'$, and $\Lambda' = \lambda(X')$. We observe that $B = I + E^*E \ge I$, where I is the m-by-m identity matrix; $\Lambda' = \tilde{\Lambda} + R^*E$, where $\tilde{\Lambda} = \lambda(\tilde{X})$ (see (3.4)); and $A = B\Lambda'$. Hence, $B_{\mathcal{I},\mathcal{J}} = E_{\mathcal{I}}^*R_{\mathcal{J}}$, $\Lambda'_{\mathcal{J},\mathcal{I}} = R_{\mathcal{J}}^*E_{\mathcal{I}}$, and

$$A_{\mathcal{I},\mathcal{I}} = B_{\mathcal{I},\mathcal{I}}\Lambda'_{\mathcal{I},\mathcal{I}} + B_{\mathcal{I},\mathcal{J}}\Lambda'_{\mathcal{J},\mathcal{I}} = B_{\mathcal{I},\mathcal{I}}\Lambda'_{\mathcal{I},\mathcal{I}} + E_{\mathcal{I}}^*E_{\mathcal{J}}R_{\mathcal{J}}^*E_{\mathcal{I}}. \tag{B.16}$$

Since $B_{\mathcal{I},\mathcal{I}} = I_{\mathcal{I},\mathcal{I}} + E_{\mathcal{I}}^*E_{\mathcal{I}} \ge I_{\mathcal{I},\mathcal{I}}$, we have $\|B_{\mathcal{I},\mathcal{I}}^{-1}\| \le 1$. Hence,

$$\begin{aligned}
\operatorname{Tr}(B_{\mathcal{I},\mathcal{I}}^{-1}A_{\mathcal{I},\mathcal{I}} - \Lambda'_{\mathcal{I},\mathcal{I}}) &= \operatorname{Tr}B_{\mathcal{I},\mathcal{I}}^{-1}E_{\mathcal{I}}^*E_{\mathcal{J}}R_{\mathcal{J}}^*E_{\mathcal{I}} = \left|\langle E_{\mathcal{I}}, R_{\mathcal{J}}E_{\mathcal{J}}^*E_{\mathcal{I}}B_{\mathcal{I},\mathcal{I}}^{-1}\rangle\right| \\
&\le \langle E_{\mathcal{I}}\rangle_{J_1}\langle E_{\mathcal{I}}, R_{\mathcal{J}}E_{\mathcal{J}}^*E_{\mathcal{I}}B_{\mathcal{I},\mathcal{I}}^{-1}\rangle_{J_1^{-1}} \le \langle E_{\mathcal{I}}\rangle_{J_1}\langle R_{\mathcal{J}}\rangle_{J_1^{-1}}\|E_{\mathcal{J}}^*E_{\mathcal{I}}B_{\mathcal{I},\mathcal{I}}^{-1}\| \\
&\le \sqrt{\rho_1}\langle E_{\mathcal{I}}\rangle_J\langle R_{\mathcal{J}}\rangle_{J^{-1}}\langle E_{\mathcal{J}}\rangle\langle E_{\mathcal{I}}\rangle \le \sqrt{\rho_1}\langle E_{\mathcal{I}}\rangle_J\langle R\rangle_{J^{-1}}\langle E\rangle\langle E_{\mathcal{I}}\rangle \le \hat{c}_1\frac{\delta_\lambda}{\lambda_{m+1}-\lambda_m}\delta_{\lambda,\mathcal{I}},
\end{aligned}$$

where $\hat{c}_1$ depends on ρ_m, δ_m, and δ_l only.

The rest of the proof employs the fact that the eigenvalues of $B_{\mathcal{I},\mathcal{I}}^{-1}A_{\mathcal{I},\mathcal{I}}$ are Ritz values of the generalized problem $Av = \lambda Bv$ in the trial subspace spanned by the columns of I with indices from $\mathcal{I}$. Denote by $\Lambda'_{\mathcal{I}}$ the diagonal matrix with the eigenvalues of $B_{\mathcal{I},\mathcal{I}}^{-1}A_{\mathcal{I},\mathcal{I}}$ on the diagonal and by $V_{\mathcal{I}}$ and $R'_{\mathcal{I}}$ the matrices whose columns are the corresponding Ritz and residual vectors, respectively. We have

$$\begin{aligned}
V_{\mathcal{I},\mathcal{I}}\Lambda'_{\mathcal{I}} = B_{\mathcal{I},\mathcal{I}}^{-1}A_{\mathcal{I},\mathcal{I}}V_{\mathcal{I}.\mathcal{I}} &= (\Lambda'_{\mathcal{I},\mathcal{I}} + B_{\mathcal{I},\mathcal{I}}^{-1}E_{\mathcal{I}}^*E_{\mathcal{J}}R_{\mathcal{J}}^*E_{\mathcal{I}})V_{\mathcal{I},\mathcal{I}} \\
&= (\tilde{\Lambda}_{\mathcal{I},\mathcal{I}} + R_{\mathcal{I}}^*E_{\mathcal{I}} + B_{\mathcal{I},\mathcal{I}}^{-1}E_{\mathcal{I}}^*E_{\mathcal{J}}R_{\mathcal{J}}^*E_{\mathcal{I}})V_{\mathcal{I},\mathcal{I}},
\end{aligned}$$

i.e., $\tilde{\Lambda}_{\mathcal{I},\mathcal{I}}V_{\mathcal{I},\mathcal{I}} - V_{\mathcal{I},\mathcal{I}}\Lambda'_{\mathcal{I}} = -(R_{\mathcal{I}}^*E_{\mathcal{I}} + B_{\mathcal{I},\mathcal{I}}^{-1}E_{\mathcal{I}}^*E_{\mathcal{J}}R_{\mathcal{J}}^*E_{\mathcal{I}})V_{\mathcal{I},\mathcal{I}}$. By Kahan's theorem (see, e.g., [20]), the eigenvalues of $\Lambda'_{\mathcal{I}}$ are not further away from respective eigenvalues of $\tilde{\Lambda}_{\mathcal{I},\mathcal{I}}$ than

$$\frac{1}{\sigma}\|\tilde{\Lambda}_{\mathcal{I},\mathcal{I}}V_{\mathcal{I},\mathcal{I}} - V_{\mathcal{I},\mathcal{I}}\Lambda'_{\mathcal{I}}\| \le \sqrt{\rho_m\rho_1}\left(\langle R_{\mathcal{I}}\rangle_{J^{-1}}\langle E_{\mathcal{I}}\rangle_J + \langle E_{\mathcal{I}}\rangle\langle E_{\mathcal{J}}\rangle\langle R_{\mathcal{J}}\rangle_{J^{-1}}\langle E_{\mathcal{I}}\rangle_J\right),$$

where σ is the minimal singular value of $V_{\mathcal{I},\mathcal{I}}$, for which we have, owing to $I_{\mathcal{I},\mathcal{I}} = V_{\mathcal{I},\mathcal{I}}^* B_{\mathcal{I},\mathcal{I}} V_{\mathcal{I},\mathcal{I}} \le (1 + \|E_{\mathcal{I}}\|^2) V_{\mathcal{I},\mathcal{I}}^* V_{\mathcal{I},\mathcal{I}}$, the lower bound $\sqrt{1-\theta^2} \ge 1/\sqrt{\rho_m}$. Since $\tilde{\lambda}_j - \lambda_j \le \delta_{\lambda,\mathcal{I}}$ for $j \in \mathcal{I}$, using previously obtained estimates for $\langle E_{\mathcal{I}}\rangle_J$ and $\langle R_{\mathcal{I}}\rangle_{J^{-1}}$ we conclude that the eigenvalues of $\Lambda'_{\mathcal{I}}$ are not further away from $\lambda_1, \dots, \lambda_l$ than $\hat{c}_2 \delta_{\lambda,\mathcal{I}}$, where $\hat{c}_2$ has the same dependencies as $\hat{c}_1$. Let us assume for a while that $\delta_{\lambda,\mathcal{I}} \le \delta_{\lambda,0} = \delta_{\mathcal{I},\mathcal{J}}/(2\hat{c}_2)$, where $\delta_{\mathcal{I},\mathcal{J}} = \lambda_{l+1} - \lambda_l$. Under this assumption, the Ritz values in question, i.e., the eigenvalues of $B_{\mathcal{I},\mathcal{I}}^{-1} A_{\mathcal{I},\mathcal{I}}$, lie in the interval that does not contain eigenvalues of L other than $\lambda_k, \dots, \lambda_l$, and we are in a position to apply residual bounds of section A.2.

Since the residual vectors are orthogonal to the trial subspace, the only nonzero rows of $R'_{\mathcal{I}}$ are those with indices from $\mathcal{J}$, and it is not difficult to see that these rows form the matrix $R'_{\mathcal{J},\mathcal{I}} = A_{\mathcal{J},\mathcal{I}} V_{\mathcal{I},\mathcal{I}} - B_{\mathcal{J},\mathcal{I}} V_{\mathcal{I},\mathcal{I}} \Lambda'_{\mathcal{I}}$, where $V_{\mathcal{I},\mathcal{I}}$ is the block of $V_{\mathcal{I}}$ formed by rows with indices from $\mathcal{I}$. Since

$$A = B\Lambda' = (I + E^*E)(\tilde{\Lambda} + R^*E) = \tilde{\Lambda} + R^*E + E^*E\tilde{\Lambda} + E^*ER^*E,$$

we have $A_{\mathcal{J},\mathcal{I}} = R_{\mathcal{J}}^* E_{\mathcal{I}} + E_{\mathcal{J}}^* E_{\mathcal{I}} \tilde{\Lambda}_{\mathcal{I},\mathcal{I}} + E_{\mathcal{J}}^* E R^* E_{\mathcal{I}}$, and hence

$$R'_{\mathcal{J},\mathcal{I}} = R_{\mathcal{J}}^* E_{\mathcal{I}} V_{\mathcal{I},\mathcal{I}} + E_{\mathcal{J}}^* E_{\mathcal{I}} \tilde{\Lambda}_{\mathcal{I},\mathcal{I}} + E_{\mathcal{J}}^* E R^* E_{\mathcal{I}} V_{\mathcal{I},\mathcal{I}} - E_{\mathcal{J}}^* E_{\mathcal{I}} V_{\mathcal{I},\mathcal{I}} \Lambda'_{\mathcal{I}}.$$

Assuming the Ritz vectors in question to be normalized by $V_{\mathcal{I}}^* B V_{\mathcal{I}} = V_{\mathcal{I},\mathcal{I}}^* B_{\mathcal{I},\mathcal{I}} V_{\mathcal{I},\mathcal{I}} = I_{\mathcal{I},\mathcal{I}}$, we have $V_{\mathcal{I},\mathcal{I}}^* V_{\mathcal{I},\mathcal{I}} \le V_{\mathcal{I},\mathcal{I}}^* B_{\mathcal{I},\mathcal{I}} V_{\mathcal{I},\mathcal{I}} = I_{\mathcal{I},\mathcal{I}}$, and hence $\|V_{\mathcal{I},\mathcal{I}}\| \le 1$. Therefore,

$$\begin{aligned}\langle R'_{\mathcal{I}}\rangle_{B^{-1}} \le \langle R'_{\mathcal{I}}\rangle = \langle R'_{\mathcal{J},\mathcal{I}}\rangle &\le \langle R_{\mathcal{J}}\rangle_{J_1^{-1}} \langle E_{\mathcal{I}}\rangle_{J_1} + \langle E_{\mathcal{J}}\rangle \langle E_{\mathcal{I}}\rangle \|\tilde{\Lambda}_{\mathcal{I},\mathcal{I}} V_{\mathcal{I},\mathcal{I}} - V_{\mathcal{I},\mathcal{I}} \Lambda'_{\mathcal{I}}\| \\ &+ \langle E_{\mathcal{J}}\rangle \langle E\rangle \langle R\rangle_{J_1^{-1}} \langle E_{\mathcal{I}}\rangle_{J_1} \le \sqrt{\rho_1}\,\big(\langle R_{\mathcal{J}}\rangle_{J^{-1}} \langle E_{\mathcal{I}}\rangle_J + \langle E_{\mathcal{J}}\rangle \langle E_{\mathcal{I}}\rangle \langle R_{\mathcal{I}}\rangle_{J^{-1}} \langle E_{\mathcal{I}}\rangle_J \\ &+ \langle E_{\mathcal{J}}\rangle^2 \langle E_{\mathcal{I}}\rangle^2 \langle R_{\mathcal{J}}\rangle_{J^{-1}} \langle E_{\mathcal{I}}\rangle_J + \langle E_{\mathcal{J}}\rangle \langle E\rangle_J \langle R\rangle_{J^{-1}} \langle E_{\mathcal{I}}\rangle_J\big) \le \sqrt{\rho_1 \hat{c}_3 \delta_\lambda \delta_{\lambda,\mathcal{I}}},\end{aligned}$$

where $\hat{c}_3$ depends on ρ_m, δ_m, δ_k (for $k > 1$), and δ_l only. Applying (A.11), we obtain

$$|\operatorname{Tr}\Lambda'_{\mathcal{I}} - \Lambda_{\mathcal{I},\mathcal{I}}| \le \frac{2}{\delta_{\mathcal{I},\mathcal{J}}} \langle R'_{\mathcal{I}}\rangle_{B^{-1}}^2 \le \frac{2\rho_1 \hat{c}_3}{\delta_{\mathcal{I},\mathcal{J}}} \delta_\lambda \delta_{\lambda,\mathcal{I}}.$$

Thus,

$$\begin{aligned}|\delta_{\lambda,\mathcal{I}} + \langle R_{\mathcal{I}}^* E_{\mathcal{I}}\rangle| &= \Big|\operatorname{Tr}\tilde{\Lambda}_{\mathcal{I},\mathcal{I}} + \operatorname{Tr} R_{\mathcal{I}}^* E_{\mathcal{I}} - \operatorname{Tr}\Lambda_{\mathcal{I},\mathcal{I}}\Big| = \big|\operatorname{Tr}\Lambda'_{\mathcal{I},\mathcal{I}} - \operatorname{Tr}\Lambda_{\mathcal{I},\mathcal{I}}\big| \\ &\le \big|\operatorname{Tr}\Lambda'_{\mathcal{I},\mathcal{I}} - \operatorname{Tr}\Lambda'_{\mathcal{I}}\big| + |\operatorname{Tr}\Lambda'_{\mathcal{I}} - \operatorname{Tr}\Lambda_{\mathcal{I},\mathcal{I}}| \le \left(\hat{c}_1 \frac{\delta_\lambda}{\lambda_{m+1} - \lambda_m} + 2\rho_1 \hat{c}_3 \frac{\delta_\lambda}{\delta_{\mathcal{I},\mathcal{J}}}\right) \delta_{\lambda,\mathcal{I}} \\ &\le \left(\hat{c}_1 + 2\hat{c}_3 \rho_m \frac{\lambda_{m+1} - \lambda_1}{\delta_{\mathcal{I},\mathcal{J}}}\right) \frac{\delta_\lambda \delta_{\lambda,\mathcal{I}}}{\lambda_{m+1} - \lambda_m} \le (\hat{c}_1 + 2\hat{c}_3 \rho_m \hat{c}_4) \frac{\delta_\lambda \delta_{\lambda,\mathcal{I}}}{\lambda_{m+1} - \lambda_m},\end{aligned}$$

where $\hat{c}_4 = 2\delta_l^{-1}$.

It remains to note that if $\delta_{\lambda,\mathcal{I}} > \delta_{\lambda,0}$, then a similar estimate holds, notably,

$$\begin{aligned}|\delta_{\lambda,\mathcal{I}} + \langle R_{\mathcal{I}}^* E_{\mathcal{I}}\rangle| &\le \delta_{\lambda,\mathcal{I}} + \langle R_{\mathcal{I}}\rangle_{J^{-1}} \langle E_{\mathcal{I}}\rangle_J \le (1 + \sqrt{c_3 c_4})\, \frac{\delta_\lambda}{\delta_{\lambda,0}} \delta_{\lambda,\mathcal{I}} \\ &\le 2\hat{c}_2\,(1 + \sqrt{c_3 c_4})\, \frac{\delta_\lambda}{\delta_{\mathcal{I},\mathcal{J}}} \delta_{\lambda,\mathcal{I}} \le 2\hat{c}_2 \hat{c}_4\,(1 + \sqrt{c_3 c_4})\, \frac{\delta_\lambda \delta_{\lambda,\mathcal{I}}}{\lambda_{m+1} - \lambda_m}.\end{aligned}$$

Together with (B.14) and (B.15), the estimates just obtained yield (3.14) and (3.15).

B.2. Proof of Theorem 3.4. The result in question is shift-invariant; hence below we assume L to be shifted so that $\lambda_1 = 0$.

In addition to the notation of section 3 and section B.1, let us denote $H = [h_1, \dots, h_m]$. The Ritz values ν_j are the eigenvalues of $B^{-1}A$, where $A = (\tilde{X} - H)^* L(\tilde{X} - H) = \tilde{\Lambda} - \tilde{X}^* L H - H^* L \tilde{X} - H^* L H = \tilde{\Lambda} - R^* H - H^* R - H^* L H$, $\tilde{\Lambda}$ is a diagonal matrix with $\tilde{\lambda}_j$ on the diagonal, $B = (\tilde{X} - H)^*(\tilde{X} - H) = I + H^* H$, and I is the m-by-m identity matrix. Let us denote $\delta A = A - \tilde{\Lambda}$ and $\delta B = B - I$. A direct calculation shows that (cf. Lemma 2 in Appendix A of [14])

$$\operatorname{Tr}(B^{-1}A) = \operatorname{Tr}(B_d^{-1}A_d) - \operatorname{Tr}A_1 + \operatorname{Tr}A_2 + \operatorname{Tr}A_3,$$

where $A_1 = B_d^{-1}\delta B B_d^{-1}\delta A$, $A_2 = B_d^{-1}\delta B_d B_d^{-1}\delta A_d$, $A_3 = B_d^{-1}B_o B^{-1} B_o B_d^{-1} A$, M_d stands for the diagonal matrix with the same diagonal as in M, and $M_o = M - M_d$. We observe that the diagonal entries of $B_d^{-1}A_d$ are μ_j; hence, to prove the theorem for $l = m$, it remains to estimate the traces of A_1, A_2, and A_3. Since $\operatorname{Tr}(ST) \le \operatorname{Tr}S\|T\|$ for any Hermitian positive definite S and Hermitian T (see Appendix A in [14]), we have $|\operatorname{Tr}(ST)| \le \operatorname{Tr}S\|T\| \le \operatorname{Tr}S\langle T\rangle$, and hence

$$|\operatorname{Tr}A_1| \le \operatorname{Tr}(B_d^{-1}\delta B B_d^{-1})\|\delta A\| \le \operatorname{Tr}(\delta B)\langle \delta A\rangle, \quad |\operatorname{Tr}A_2| \le \operatorname{Tr}(\delta B_d)\langle \delta A\rangle,$$

$$|\operatorname{Tr}A_3| \le \operatorname{Tr}(B_d^{-1}B_o B^{-1} B_o B_d^{-1})\|A\| = \langle B^{-\frac{1}{2}} B_o B_d^{-1}\rangle^2 \|A\| \le \langle \delta B\rangle^2 \|A\|,$$

where we have used the fact that $B \ge I$ and $\langle B_o\rangle = \langle \delta B_o\rangle \le \langle \delta B\rangle$. It is not difficult to see that $\operatorname{Tr}(\delta B) = \operatorname{Tr}(\delta B_d) = \langle H\rangle^2 \le \gamma^{-1}\langle H\rangle_J^2$, $\langle \delta A\rangle \le 2\langle R\rangle_{J_1^{-1}}\langle H\rangle_{J_1} + \langle H\rangle_L^2 \le 2\sqrt{\rho_1}\langle R\rangle_{J^{-1}}\langle H\rangle_J + \rho_1\langle H\rangle_J^2$, $\langle \delta B\rangle \le \langle H\rangle^2$, and $\|A\| \le \tilde{\lambda}_m - \lambda_1 + \|\delta A\| \le \lambda_{m+1} - \lambda_1 + \langle \delta A\rangle$. In view of (3.20) and (3.21), $\|h_j\|_{J_j}^2 \le \tilde{\lambda}_j - \mu_j$ and $\|r_j\|_{J_j^{-1}}^2 \le \eta(\tilde{\lambda}_j - \mu_j)$. Thus, for $l = m$ we have (3.22) with $c = c_0$, where c_0 depends on ρ_1, ρ_m, and η only.

Turning now to the general case, denote $\mathcal{W}_{\mathcal{I}} = \operatorname{span}\{w_j\}_{j\in\mathcal{I}}$, where $\mathcal{I} = \{1, \dots, l\}$, and denote by $\nu_j^{\mathcal{I}}$ the Ritz values of L in $\mathcal{W}_{\mathcal{I}}$. Repeating the above calculations with $A_{\mathcal{I},\mathcal{I}}$ and $B_{\mathcal{I},\mathcal{I}}$ substituted for A and B, we arrive at the estimate

$$\left|\sum_{j=1}^{l}(\nu_j^{\mathcal{I}} - \mu_j)\right| \le \frac{c_0\,(\delta_{\mu,\mathcal{I}})^2}{\lambda_{m+1} - \lambda_m}, \quad \delta_{\mu,\mathcal{I}} = \sum_{j=1}^{l}(\tilde{\lambda}_j - \mu_j). \tag{B.17}$$

We observe that $\tilde{\lambda}_j - \mu_j = (J(\mu_j, \tilde{\mathcal{X}})h_j, h_j) \le \rho_1\|h_j\|_{J_j}^2 \le \rho_1\|r_j\|_{J_j^{-1}}^2$, and hence $\delta_{\mu,\mathcal{I}} \le c_1\delta_{\lambda,\mathcal{I}} \le c_1\delta_\lambda \le c_1(\lambda_{m+1} - \lambda_m)$, where $\delta_{\lambda,\mathcal{I}}$ is the sum of $\tilde{\lambda}_j - \lambda_j$, $j = 1, \dots, l$, and $c_1 = 8\rho_1^2$ (cf. (B.7)). Further, (3.21) implies, via (A.10), $\delta_{\lambda,\mathcal{I}} \le c_2\delta_{\mu,\mathcal{I}}$, where c_2 depends on ρ_1, $\tilde{\delta}_l$, and η only. By the minimax principle, $\lambda_j \le \nu_j \le \nu_j^{\mathcal{I}}$. Thus,

$$\left|\sum_{j=1}^{l}(\nu_j - \mu_j)\right| \le \max\left\{\left|\sum_{j=1}^{l}(\lambda_j - \mu_j)\right|, \left|\sum_{j=1}^{l}(\nu_j^{\mathcal{I}} - \mu_j)\right|\right\}$$

$$\le \max\{\delta_{\lambda,\mathcal{I}} + \delta_{\mu,\mathcal{I}}, c_0c_1\delta_{\mu,\mathcal{I}}\} \le \max\{c_2 + 1, c_0c_1\}\delta_{\mu,\mathcal{I}},$$

which trivially implies (3.22) for $\delta_\lambda > \delta_0 = (\lambda_{m+1} - \lambda_m)/c_3$, where c_3 is an arbitrary positive value depending on the same parameters as c. It remains to prove (3.22) for $\delta_\lambda \le \delta_0$.

The estimate (B.17) suggests that in order to complete the proof, we need to estimate the difference between μ_j and $\nu_j^{\mathcal{I}}$. We observe that the latter are the eigenvalues

of $B_{\mathcal{I},\mathcal{I}}^{-1}A_{\mathcal{I},\mathcal{I}}$ and hence are the Ritz values of the generalized problem $Av = \lambda Bv$ in the subspace spanned by the columns of the m-by-m identity matrix with indices from $\mathcal{I}$. Denote by $V_{\mathcal{I}}$ and $R'_{\mathcal{I}}$ the matrices whose columns are the corresponding Ritz and residual vectors, respectively, and observe that the only nonzero rows of $R'_{\mathcal{I}}$ are those with indices from $\mathcal{J} = \{l+1, \dots, m\}$ and that they form the matrix

$$A_{\mathcal{J},\mathcal{I}}V_{\mathcal{I},\mathcal{I}} - B_{\mathcal{J},\mathcal{I}}V_{\mathcal{I},\mathcal{I}}\Lambda'_{\mathcal{I}} = (-R^*_{\mathcal{J}}H_{\mathcal{I}} - H^*_{\mathcal{J}}R_{\mathcal{I}} + H^*_{\mathcal{J}}LH_{\mathcal{I}})V_{\mathcal{I},\mathcal{I}} - H^*_{\mathcal{J}}H_{\mathcal{I}}V_{\mathcal{I},\mathcal{I}}\Lambda'_{\mathcal{I}},$$

where $V_{\mathcal{I},\mathcal{I}}$ is the block of $V_{\mathcal{I}}$ formed by rows with indices from $\mathcal{I}$, and $\Lambda'_{\mathcal{I}}$ is a diagonal matrix with $\lambda_j^{\mathcal{X}'_{\mathcal{I}}}$ on the diagonal. Assuming the Ritz vectors in question to be normalized by $V^*_{\mathcal{I},\mathcal{I}}B_{\mathcal{I},\mathcal{I}}V_{\mathcal{I},\mathcal{I}} = I_{\mathcal{I},\mathcal{I}}$, we have $\|V_{\mathcal{I},\mathcal{I}}\| \le 1$; owing to this and to (3.20) and (3.21) we have $\langle R'_{\mathcal{I}}\rangle \le c_4\delta_{\lambda,\mathcal{I}}\delta_{\lambda,\mathcal{J}}$, where c_4 has the same dependencies as c and $\delta_{\lambda,\mathcal{J}}$ is the sum of $\tilde{\lambda}_j - \lambda_j$, $j = l+1, \dots, m$. In order to apply the residual estimates of section A.2, let us find c_3 that ensures $\lambda_{l+1} - \nu_l^{\mathcal{I}} \ge (\lambda_{l+1} - \lambda_l)/2$. Since $\nu_l^{\mathcal{I}}$ is the largest eigenvalue of $B_{\mathcal{I},\mathcal{I}}^{-1}A_{\mathcal{I},\mathcal{I}}$, it is also the largest eigenvalue of $B_{\mathcal{I},\mathcal{I}}^{-1/2}A_{\mathcal{I},\mathcal{I}}B_{\mathcal{I},\mathcal{I}}^{-1/2}$ and hence is equal to $\|B_{\mathcal{I},\mathcal{I}}^{-1/2}A_{\mathcal{I},\mathcal{I}}B_{\mathcal{I},\mathcal{I}}^{-1/2}\| \le \|A_{\mathcal{I},\mathcal{I}}\| \le \|\tilde{\Lambda}_{\mathcal{I},\mathcal{I}}\| + \|\delta A_{\mathcal{I},\mathcal{I}}\| \le \tilde{\lambda}_l + \langle \delta A\rangle \le \tilde{\lambda}_l + c_5\delta_\lambda$. Therefore, $\lambda_{l+1} - \nu_l^{\mathcal{I}} = \lambda_{l+1} - \tilde{\lambda}_l + \tilde{\lambda}_l - \nu_l^{\mathcal{I}} \ge \lambda_{l+1} - \tilde{\lambda}_l - c_5\delta_\lambda \ge \lambda_{l+1} - \lambda_l - (1+c_5)\delta_\lambda$. Thus, c_3 should be set to $2(\lambda_{m+1} - \lambda_m)$ divided by $(\lambda_{l+1} - \lambda_l)/(1+c_5)$. It remains to apply (A.11) to arrive at

$$\left|\sum_{j=1}^{l}(\nu_j - \nu_j^{\mathcal{I}})\right| \le \frac{2c_4\delta_{\lambda,\mathcal{I}}\delta_{\lambda,\mathcal{J}}}{\lambda_{l+1} - \lambda_l} \le \frac{2c_4c_1\delta_\lambda}{\lambda_{l+1} - \lambda_l}\sum_{j=1}^{l}(\tilde{\lambda}_j - \mu_j),$$

which, together with (B.17), yields (3.22).

B.3. An auxiliary estimate for residuals.

LEMMA B.1. *Assuming that the search directions y_j^i employed by (2.1)–(2.2) satisfy $(J(\lambda_1, \mathcal{X}^i)y_j^i, y_j^i) \le \sigma(Kr_j^i, r_j^i)$ for some $\sigma > 0$, the following inequality is valid:*

$$\sum_{j=1}^{l}\|r\|_K^2 \le \sigma\sum_{j=1}^{l}(\lambda_j^i - \lambda_j^{i+1}). \tag{B.18}$$

Proof. The result in question is shift-invariant; hence we assume, in order to simplify the calculations, that $\lambda_1 = 0$. Denote $X_l^i = [x_j^i]_{j=1}^l$; $Y_l^i = [(1-P_i)y_j^i]_{j=1}^l$, where P_i is the orthogonal projector onto $\mathcal{X}^i$; $R_l^i = [r_j^i]_{j=1}^l$; $\Lambda_l^i = \operatorname{diag}\{\lambda_j^i\}_{j=1}^l$. Let ν_j^i be the Ritz values of L in $\mathcal{U}^i = \operatorname{span}\{X_l^i - \tau Y_l^i\}$, where $\tau = \sigma^{-1}$. Since $\mathcal{U}^i \subset \mathcal{V}^i = \operatorname{span}\{X^i, Y^i\}$, the minimax principle implies $\lambda_j^{i+1} \le \nu_j^i$. The Ritz values ν_j^i are the eigenvalues of the generalized problem $Au = \nu Bu$, where $A = (X_l^i)^*LX_l^i - \tau(X_l^i)^*LY_l^i - \tau(Y_l^i)^*LX_l^i + \tau^2(Y_l^i)^*LY_l^i = \Lambda_l^i - \tau(R_l^i)^*Y_l^i - \tau(Y_l^i)^*R_l^i + \tau^2(Y_l^i)^*LY_l^i = \Lambda_l^i - 2\tau(R_l^i)^*KR_l^i + \tau^2(Y_l^i)^*LY_l^i$; $B = I + \tau^2(Y_l^i)^*Y_l^i$; and I is l-by-l identity. Since $A \ge 0$ and $B \ge I$, the Ritz values ν_j^i are not greater than the eigenvalues of A. The assumption $(J(\lambda_1, \mathcal{X}^i)y_j^i, y_j^i) = (L(1-P_i)y_j^i, (1-P_i)y_j^i) \le \sigma(Kr_j^i, r_j^i)$ implies $\tau^2(Y_l^i)^*LY_l^i \le \tau(R_l^i)^*KR_l^i$, and hence the eigenvalues of A are not greater than the eigenvalues of $\hat{A} = \Lambda_l^i - \tau(R_l^i)^*KR_l^i$. Thus, λ_j^{i+1}, $j = 1, \dots, l$, are not greater than the eigenvalues of $\hat{A}$, and considering the trace of $\hat{A}$ results in (B.18). ∎

B.4. Proof of the estimate (4.7). The eigenvalue problem for A_{ij} is equivalent to the generalized eigenvalue problem $J_{ij}KJ_{ij}u = \xi J_{ij}u$, and the proof of (4.7) exploits this fact and the following energy norm equivalences: (i) the equivalence of the K- and L^{-1}-norms, responsible for the first factor in (4.7), (ii) the equivalence of the L^{-1}- and $J(0,\mathcal{X}^i)^{-1}$-norms, responsible for the second (bracketed) factor, and (iii) the equivalence of the $J(0,\mathcal{X}^i)^{-1}$ and J_{ij}^{-1}-norms, responsible for the last factor.

Denote by α and ω the minimal and the maximal eigenvalue of KL. We have $\alpha L^{-1} \le K \le \omega L^{-1}$, and hence for the eigenvalues ξ_k of A_{ij} we have $\alpha\eta_k \le \xi_k \le \omega\eta_k$, where η_k are the eigenvalues of the problem $J_{ij}L^{-1}J_{ij}v = \eta J_{ij}v$. As we have shown in section B.1, $\|r\|_{J_0^{-1}} \le \|r\|_{L^{-1}}$ for any $r \in (\mathcal{X}^i)^\perp$, where $J_0 = J(0,\mathcal{X}^i)$ (cf. the argument just above (B.7)). Thus, $(J_{ij}L^{-1}J_{ij}u,u) = \|J_{ij}u\|^2_{L^{-1}} \ge \|J_{ij}u\|^2_{J_0^{-1}} \ge (\gamma_{ij}/\gamma_i)\|J_{ij}u\|^2_{J_{ij}^{-1}} = (\gamma_{ij}/\gamma_i)\|u\|^2_{J_{ij}}$, where $\gamma_i = \lambda_{m+1} - \delta_\lambda(\mathcal{X}^i)$, $\gamma_{ij} = \lambda_{m+1} - \delta_\lambda(\mathcal{X}^i) - \lambda^i_j$, and we have used (3.12), which implies $\eta_k \ge \gamma_{ij}/\gamma_i$. In order to establish an upper bound for η_k, consider an auxiliary linear system $Lu = J_{ij}v$. We have $\|u\|^2_L = (J_{ij}v,u) \le \|v\|_{J_{ij}}\|u\|_{J_{ij}} \le \|v\|_{J_{ij}}\|u\|_{J_0}$. Denote by Q_i the $(\cdot,\cdot)_L$-orthogonal projector onto $\mathcal{X}^i$. As we have seen in section B.1 (Step 3), $Q_i = P_i + X^i\Lambda_i^{-1}(R^i)^*$, where Λ_i is an m-by-m diagonal matrix with λ^i_j on the diagonal. Since $((1-Q_i)u,(Q_i-P_i)u)_L = 0$, we have $\|u\|^2_{J_0} = \|(1-P_i)u\|_L = \|(1-Q_i)u\|^2_L + \|(Q_i-P_i)u\|^2_L \le \|u\|^2_L + \|X^i\Lambda_i^{-1}(R^i)^*u\|^2_L \le \|u\|^2_L + \tilde\lambda_m\|R_i\Lambda_i^{-1}\|^2_{L^{-1}}\|u\|^2_L \le (1+\lambda_{m+1}\|R_i\Lambda_i^{-1}\|^2_{L^{-1}})\|u\|^2_L \equiv \pi_m\|u\|^2_L$, and hence $\|u\|^2_L \le \pi_m\|v\|^2_{J_{ij}}$. Observing that $\|u\|_L = \|L^{-1}J_{ij}v\|_L = \|J_{ij}v\|_{L^{-1}}$, we conclude that $\|J_{ij}v\|^2_{L^{-1}} = (J_{ij}L^{-1}J_{ij}v,v) \le \pi_m(J_{ij}v,v)$. It remains to apply the estimate (B.5) for $\langle R^i\Lambda_i^{-1}\rangle_{L^{-1}} \ge \|R^i\Lambda_i^{-1}\|_{L^{-1}}$ to obtain an upper bound for π_m and hence for η_k. The obtained bounds for η_k yield the bounds for ξ_k that, in turn, yield (4.7).

REFERENCES

[1] P. Arbenz, U. L. Hetmaniuk, R. B. Lehoucq, and R. S. Tuminaro, *A comparison of eigensolvers for large-scale 3D modal analysis using AMG-preconditioned iterative methods*, Internat J. Numer. Methods Engrg., 64 (2005), pp. 204–236.

[2] C. Beattie, *Harmonic Ritz and Lehmann bounds*, Electron. Trans. Numer. Anal., 7 (1998), pp. 18–39.

[3] W. W. Bradbury and R. Fletcher, *New iterative methods for solution of the eigenproblem*, Numer. Math., 9 (1966), pp. 259–267.

[4] J. Brandts, *The Riccati algorithm for eigenvalues and invariant subspaces of matrices with inexpensive action*, Linear Algebra Appl., 358 (2003), pp. 335–365.

[5] A. Edelman, T. A. Arias, and S. T. Smith, *The geometry of algorithms with orthogonality constraints*, SIAM J. Matrix Anal. Appl., 20 (1998), pp. 303–353.

[6] Z. Fu and E. M. Dowling, *Conjugate gradient eigenstructure tracking for adaptive spectral estimation*, IEEE Trans. Signal Process., 43 (1995), pp. 1151–1160.

[7] A. V. Knyazev, *Convergence rate estimates for iterative methods for a mesh symmetric eigenvalue problem*, Soviet J. Numer. Anal. Math. Modelling, 2 (1987), pp. 371–396.

[8] A. V. Knyazev, *Preconditioned eigensolvers—an oxymoron?*, Electron. Trans. Numer. Anal., 7 (1998), pp. 104–123.

[9] A. V. Knyazev, *Toward the optimal preconditioned eigensolver: Locally optimal block preconditioned conjugate gradient method*, SIAM J. Sci. Comput., 23 (2001), pp. 517–541.

[10] A. V. Knyazev, M. E. Argentati, I. Lashuk, and E. E. Ovtchinnikov, *Block Locally Optimal Preconditioned Eigenvalue Xolvers (BLOPEX) in hypre and PETSc*, SIAM J. Sci. Comput., 29 (2007), pp. 2224–2239.

[11] N. J. Lehmann, *Optimal eigenvalue localization in the solution of symmetric matrix problems*, Numer. Math., 5 (1963), pp. 246-272.

[12] D. E. Longsine and S. F. McCormick, *Simultaneous Rayleigh-quotient minimization methods for $Ax = \lambda Bx$*, Linear Algebra Appl., 34 (1980), pp. 195–234.

[13] K. Neymeyr, *A geometric theory for preconditioned inverse iteration applied to a subspace*, Math. Comput., 71 (2002), pp. 197–216.
[14] E. E. Ovtchinnikov, *Cluster robustness of preconditioned gradient subspace iteration eigensolvers*, Linear Algebra Appl., 415 (2006), pp. 140–166.
[15] E. E. Ovtchinnikov, *Cluster robust error estimates for the Rayleigh–Ritz approximation* I: *Estimates for invariant subspaces*, Linear Algebra Appl., 415 (2006), pp. 167–187.
[16] E. E. Ovtchinnikov, *Cluster robust error estimates for the Rayleigh–Ritz approximation* II: *Estimates for eigenvalues*, Linear Algebra Appl., 415 (2006), pp. 188–209.
[17] E. E. Ovtchinnikov, *Jacobi correction equation, line search, and conjugate gradients in Hermitian eigenvalue computation* I: *Computing an extreme eigenvalue*, SIAM J. Numer. Anal., 46 (2008), pp. 2567–2592.
[18] E. E. Ovtchinnikov, *Computing several eigenpairs of Hermitian problems by conjugate gradient iterations*, J. Comput. Phys., submitted.
[19] E. E. Ovtchinnikov, J. K. Reid, and J. A. Scott, *The Design of a Preconditioned Conjugate Gradient Code for Solving Hermitian Eigenvalue Problems*, RAL Technical report, Rutherford Appleton Laboratory, Didcot, UK, 2008.
[20] B. N. Parlett, *The Symmetric Eigenvalue Problem*, Prentice–Hall, Englewood Cliffs, NJ, 1980.
[21] A. Perdon and G. Gambolati, *Extreme eigenvalues of large sparse matrices by Rayleigh quotient and modified conjugate gradients*, Comput. Methods Appl. Mech. Engrg., 56 (1986), pp. 251–264.
[22] F. Sartoretto, G. Pini, and G. Gambolati, *Accelerated simultaneous iterations*, J. Comput. Phys., 81 (1989), pp. 53–69.
[23] G. L. G. Sleijpen and H. A. van der Vorst, *A Jacobi–Davidson iteration method for linear eigenvalue problems*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 401–425.

SIAM J. NUMER. ANAL.
Vol. 46, No. 5, pp. 2620–2635
© 2008 Society for Industrial and Applied Mathematics

RECOVERY OF EDGES FROM SPECTRAL DATA WITH NOISE—A NEW PERSPECTIVE*

SHLOMO ENGELBERG† AND EITAN TADMOR‡

Abstract. We consider the problem of detecting edges—jump discontinuities in piecewise smooth functions from their N-degree spectral content, which is assumed to be corrupted by noise. There are three scales involved: the "smoothness" scale of order $1/N$, the noise scale of order $\sqrt{\eta}$, and the $\mathcal{O}(1)$ scale of the jump discontinuities. We use concentration factors which are adjusted to the standard deviation of the noise $\sqrt{\eta} \gg 1/N$ in order to detect the underlying $\mathcal{O}(1)$-edges, which are separated from the noise scale $\sqrt{\eta} \ll 1$.

Key words. piecewise smoothness, edge detection, noisy data, concentration kernels, constrained optimization, separation of scales

AMS subject classifications. 42A10, 42A50, 65T10

DOI. 10.1137/070689899

1. Introduction and statement of main results.

We consider the detection of jump discontinuities in piecewise smooth data from its Fourier projection

$$S_N f(x) = \sum_{|k|\le N} \widehat{f}(k) e^{ikx}, \qquad \widehat{f}(k) := \frac{1}{2\pi}\int_{-\pi}^{\pi} f(\xi) e^{-i\xi x}.$$

Our approach for edge detection—we use the terminology of edges or jump discontinuities interchangeably—is based on the technique of *concentration kernels* advocated in [3, 4] and the closely related techniques described in [2]. The technique presented makes use of the fact that, if f is discontinuous, its Fourier coefficients contain a slowly decaying part associated with the jumps of f. This part decays much more slowly than the rapidly decaying smooth part of f. For example, if f is smooth except for a single jump discontinuity at $x = z$ of size $[f](z) := f(z+) - f(z-)$, the jump discontinuity is associated with slowly decaying Fourier coefficients (here and below we use $X \lesssim Y$ to denote the estimate $X \le CY$, where C is a constant which is independent of the variables $k, N, \dots$)

$$\widehat{f}(k) = [f](z)\frac{e^{-ikz}}{2\pi ik} + \widehat{s}(k), \qquad |\widehat{s}(k)| \lesssim \frac{1}{|k|^2}, \quad [f](x) := f(x+) - f(x-),$$

where $\widehat{s}(k)$ are the Fourier coefficients of s, the smooth part of f; the smoother s is, the faster is the decay of $\widehat{s}(k)$. Concentration kernels succeed in separating the two sets of coefficients. To this end one computes

$$\mathcal{K}_N^\sigma * (S_N f)(x) = \pi i \sum_{|k|\le N} \mathrm{sgn}(k)\sigma_N\left(\frac{|k|}{N}\right)\widehat{f}(k)e^{ikx}, \qquad \mathrm{sgn}(k) := \begin{cases} k/|k| & k \ne 0, \\ 0 & k = 0. \end{cases}$$

*Received by the editors May 7, 2007; accepted for publication (in revised form) January 22, 2008; published electronically July 2, 2008.
http://www.siam.org/journals/sinum/46-5/68989.html

†Department of Electronics, Jerusalem College of Technology, P.O.B. 16031, Jerusalem, Israel (shlomoe@jct.ac.il, http://www.cc.jct.ac.il/~shlomoe/).

‡Department of Mathematics, Institute for Physical Science and Technology, and Center of Scientific Computation and Mathematical Modeling (CSCAMM), University of Maryland, College Park, MD 20742 (tadmor@cscamm.umd.edu, http://www.cscamm.umd.edu/~tadmor). This author's research was supported in part by NSF grant 07-07949 and ONR grant N00014-91-J-1076.

Here, $\sigma_N(\xi)$ can be drawn from a large family of properly normalized concentration factors that is at our disposal. The resulting function tends to zero in regions in which f is smooth and tends to the amplitude of the jumps at points where the function has jump discontinuities, e.g., [8, section 4],

$$\mathcal{K}_N^\sigma * (S_N f)(x) = [f](x) + \mathcal{O}\left(\frac{\log N}{N}\right).$$

Thus, jump discontinuities, or edges, are detected by *separation of scales*: the quantity on the right is of a vanishing order $\lesssim (\log N)/N$, in regions of smoothness of f, and it is $\gg (\log N)/N$, in the neighborhood of jump discontinuities of f.

In this paper we utilize concentration kernels to detect edges from spectral information which is corrupted by *white noise*. In this context we observe that there are three scales involved—edges of order $\mathcal{O}(1)$, noise with variance $\mathcal{O}(\eta)$, and the smooth part of f which is resolved within order $\mathcal{O}(1/N)$ or smaller. Here, we can separate the noisy part $n(x)$ from the smooth part

$$s(x) \to s(x) + n(x), \qquad E(|\widehat{n}(k)|^2) = \eta;$$

if $\sqrt{\eta} \lesssim 1/N$, then the noisy part could be identified with (or below) the $\mathcal{O}(1/N)$-variation of the smooth part of f. In this case, there are essentially two scales, and edges can be detected using the usual framework of concentration kernels advocated in [3, 4]. Thus, our main focus in this paper is when the smoothness scale is dominated by the scale of the noise which is still well separated from the $\mathcal{O}(1)$-scale of the jumps,[1]

$$\frac{1}{N} \lesssim \sqrt{\eta} \ll 1.$$

The spectral information is now corrupted by white noise, affecting both low and high frequencies:

$$\widehat{f}(k) = [f](z)\frac{e^{-ikz}}{2\pi ik} + \widehat{s}(k) + \widehat{n}(k).$$

In order to separate edges from the noisy scale, the edge detector $K_N^\sigma f(x)$ must be properly adapted to the presence of white noise. We show how to design edge detectors that optimally compensate for noise and for the effects of the smooth part of the signal.

The paper is organized as follows. In section 2 we discuss the general framework of edge detection based on concentration kernels. We revisit the results of [4, 8], providing a simpler proof for the concentration property for a large family of concentration factors. In particular, we trace the precise dependence of the error on the regularity of the associated concentration factor. This will prove useful when we deal with noisy data in section 3. Here, we introduce our new perspective, where concentration factors are derived by a constrained minimization while taking into account the two main ingredients of our data—jump discontinuities and the noisy parts of the data. Numerical results are demonstrated in section 4. In section 5 we extend our construction of concentration factors to include the *three* ingredients of the data—taking into account the smooth part of the data in addition to the edges and noisy parts; numerical results are presented in section 6.

[1] Here and below, $X \ll Y$ indicates that X is an "order of magnitude smaller" than Y, so that $\lim_{N\to\infty} X/Y \to 0$, $X \sim Y$, indicates that X and Y are "of the same order of magnitude," namely, $X \lesssim Y \lesssim X$, and $X \approx Y$ indicates that $\lim_{N\to\infty} |X - Y| = 0$.

2. Detection of edges—concentration kernels. Consider an f which is *piecewise smooth* in the sense that it is sufficiently smooth except for finitely many jump discontinuities, say, at $z_1 < z_2, \dots < z_m$, where

$$[f](z_j) := f(z_j+) - f(z_j-) \neq 0, \quad j = 1, 2, \dots m.$$

Given the Fourier coefficients $\left\{\widehat{f}(k)\right\}_{k=-N}^{N}$, we are interested in detecting the edges of the underlying piecewise smooth f, namely, to detect their location $z_1, \dots, z_m$ and their amplitudes $[f](z_1), \dots, [f](z_m)$.

We utilize edge detection based on *concentration kernels*

$$\mathcal{K}_N^\sigma(y) := -\sum_{k=1}^{N} \sigma_N\left(\frac{k}{N}\right) \sin ky, \qquad \frac{\sigma_N(\xi)}{\xi} \in C^1[0,1]. \tag{2.1a}$$

We shall need the kernel $-\mathcal{K}_N^\sigma$ to have (approximately) unit mass

$$\int_0^\pi \mathcal{K}_N^\sigma(y)dy \approx -1.$$

To this end, we require that $\sigma(\xi) \equiv \sigma_N(\xi)$ be a properly normalized concentration factor so that $\sigma(\xi)$ satisfies

$$\int_0^1 \frac{\sigma(\xi)}{\xi} d\xi = 1. \tag{2.1b}$$

Indeed, the rectangular quadrature rule yields

$$\begin{aligned}\int_0^\pi \mathcal{K}_N^\sigma(y)dy &= \sum_{k=1}^{N} \sigma\left(\frac{k}{N}\right) \frac{(-1)^k - 1}{k} \\ &= -\sum_{k \text{ odd}\geq 1}^{N} \frac{\sigma(\xi_k)}{\xi_k} \frac{2}{N} = -\int_0^1 \frac{\sigma(\xi)}{\xi} d\xi + \delta_N, \quad \xi_k := \frac{k}{N},\end{aligned}$$

with an error estimate, e.g., [6, 1],

$$\left| \int_0^\pi \mathcal{K}_N^\sigma(y)dy + 1 \right| \lesssim \delta_N, \qquad \delta_N := \frac{1}{N} \left\| \frac{\sigma_N(\xi)}{\xi} \right\|_{BV}. \tag{2.2}$$

We set[2]

$$K_N^\sigma f(x) := \mathcal{K}_N^\sigma * f(x) = \pi i \sum_{|k|\leq N} \operatorname{sgn}(k) \sigma_N\left(\frac{|k|}{N}\right) \widehat{f}(k) e^{ikx}.$$

Our purpose is to choose the *concentration factors* $\sigma_N(|k|/N)$ such that $K_N^\sigma f(x) \approx [f](x)$. Thus, $K_N^\sigma f(x)$ will detect the edges $[f](z_j)$, $j = 1, \dots, m$, by concentrating near these $\mathcal{O}(1)$-edges, which are to be separated from a much smaller scale of order $K_N^\sigma f(x) \approx 0$ in regions of smoothness. In the following theorem we present a rather

[2]We observe that $K_N^\sigma f$ is the *operator* associated with, but otherwise different from, the concentration kernel $\mathcal{K}_N^\sigma(x)$.

general framework for edge detectors based on concentration factors. In particular, we track the precise dependence of the scale separation on the behavior of σ.

THEOREM 2.1 (concentration kernels). *Assume that $f(\cdot)$ is piecewise smooth such that the first variation of f is of locally bounded variation (BV)*

$$\omega_x(y) \equiv \omega_f(y,x) := \frac{f(x+y) - f(x-y) - [f](x)}{y} \in BV[-\pi,\pi]. \tag{2.3}$$

Let $\mathcal{K}_N^\sigma(x)$ be an admissible concentration kernel (2.1) *such that*

$$\delta_N := \frac{1}{N}\left\| \frac{\sigma_N(\xi)}{\xi} \right\|_{BV} \ll 1, \tag{2.4a}$$

$$\varepsilon_1(N) := \frac{1}{N}|\sigma_N(1)| \ll 1, \tag{2.4b}$$

$$\varepsilon_2(N) := \frac{1}{N}\int_{\xi\sim\frac{1}{N}}^{1} \frac{|\sigma_N'(\xi)|}{\xi} d\xi \ll 1, \tag{2.4c}$$

$$\varepsilon_3(N) := \left|\sigma_N\left(\frac{1}{N}\right)\right| \ll 1. \tag{2.4d}$$

Set

$$\varepsilon_N := \varepsilon_1(N) + \varepsilon_2(N) + \varepsilon_3(N). \tag{2.5}$$

Then, the conjugate sum

$$K_N^\sigma f(x) = \pi i \sum_{|k|\le N} \operatorname{sgn}(k)\sigma_N\left(\frac{|k|}{N}\right)\widehat{f}(k)e^{ikx} \tag{2.6}$$

satisfies the concentration property

$$|K_N^\sigma f(x) - [f](x)| \lesssim \delta_N + \varepsilon_N \|\omega_x(\cdot)\|_{BV}. \tag{2.7}$$

Proof. We simplify the proof in [4, 8]. The key to the proof is to observe that $\mathcal{K}_N^\sigma$ is an appropriately normalized *derivative* of the delta function; in particular, since $\mathcal{K}_N^\sigma(\cdot)$ is odd

$$\begin{aligned}\mathcal{K}_N^\sigma * f(x) &= -\int_0^\pi K_N^\sigma(y)\Big(f(x+y) - f(x-y)\Big)dy \\ &= -\int_0^\pi \mathcal{K}_N^\sigma(y)\Big(f(x+y) - f(x-y) - [f](x)\Big)dy - [f](x)\times\int_0^\pi \mathcal{K}_N^\sigma(y)dy.\end{aligned}$$

Since σ is assumed normalized, the error estimate (2.2) tells us[3] that

$$\int_0^\pi \mathcal{K}_N^\sigma(y)dy = -1 + \mathcal{O}(\delta_N), \qquad \delta_N = \frac{1}{N}\left\| \frac{\sigma_N(\xi)}{\xi} \right\|_{BV},$$

and we end up with the error estimate

$$\left|K_N^\sigma f(x) - [f](x)\right| \lesssim \left|\int_0^\pi \mathcal{K}_N^\sigma(y) y\omega_x(y)dy\right| + \mathcal{O}(\delta_N). \tag{2.8}$$

[3]In our case, the errors for the trapezoidal and rectangular rules provided, respectively, in [1] and [6] coincide modulo the $\mathcal{O}(\varepsilon_1(N) + \varepsilon_3(N))$ terms.

To upper bound the expression on the right, we use the following identity, which is derived by a straightforward summation by parts:

$$2\sin(y/2)\mathcal{K}_N^\sigma(y) \equiv \overbrace{\sigma_N(1)\cos\left(N+\frac{1}{2}\right)y}^{I_1(y)} + \sum_{k=1}^{N-1}\overbrace{\Big(\sigma_N(\xi_k)-\sigma_N(\xi_{k+1})\Big)\cos\left(k+\frac{1}{2}\right)y}^{I_{2k}(y)} - \overbrace{\sigma_N(\xi_1)\cos\left(\frac{y}{2}\right)}^{I_3(y)}.$$

The usual cancellation estimate $|\int_0^\pi \cos((k+\frac{1}{2})y)w(y)dy| \lesssim \|w\|_{BV}/|k+\frac{1}{2}|$ implies that

$$\left|\int_0^\pi I_1(y)\frac{y/2}{\sin(y/2)}\omega_x(y)dy\right| \lesssim |\sigma_N(1)|\cdot\frac{1}{N}\|\omega_x(\cdot)\|_{BV} \le \varepsilon_1(N)\|\omega_x(\cdot)\|_{BV},$$
$$\left|\int_0^\pi I_{2k}(y)\frac{y/2}{\sin(y/2)}\omega_x(y)dy\right| \lesssim |\sigma(\xi_{k+1})-\sigma(\xi_k)|\cdot\frac{1}{k+\frac{1}{2}}\|\omega_x(\cdot)\|_{BV},$$
$$\left|\int_0^\pi I_3(y)\frac{y/2}{\sin(y/2)}\omega_x(y)dy\right| \lesssim \varepsilon_3(N)\cdot\|\omega_x(\cdot)\|_{L^\infty}.$$

We conclude that

$$\begin{aligned}\left|\int_0^\pi \mathcal{K}_N^\sigma(y)y\omega_x(y)dy\right| &= \left|\int_0^\pi\left(I_1(y)+\sum_k I_{2k}(y)+I_3(y)\right)\frac{y/2}{\sin(y/2)}\omega_x(y)dy\right| \\ &\lesssim \left(\varepsilon_1(N)+\sum_{k=1}^{N-1}\frac{|\sigma(\xi_{k+1})-\sigma(\xi_k)|}{\xi_{k+\frac{1}{2}}}\frac{1}{N}+\varepsilon_3(N)\right)\left\|\frac{y/2}{\sin(y/2)}\omega_x(y)\right\|_{BV} \\ &\lesssim \Big(\varepsilon_1(N)+\varepsilon_2(N)+\varepsilon_3(N)\Big)\|\omega_x(\cdot)\|_{BV}.\end{aligned} \tag{2.9}$$

The result (2.7) follows from (2.8) and (2.9). ∎

As an example, we consider the noise-free case with a concentration factor for which

$$\frac{\sigma_N(\xi)}{\xi}\in C^1. \tag{2.10}$$

Clearly $\delta_N \lesssim 1/N$, and, since σ is bounded, $\varepsilon_1(N)\lesssim 1/N$. Moreover, since $\sigma_N'(\xi)$ is bounded,

$$\varepsilon_2(N)\lesssim\frac{1}{N}\int_{\xi\sim\frac{1}{N}}^1\frac{1}{\xi}d\xi\lesssim\frac{\log(N)}{N}.$$

Finally, since $|\sigma_N(\xi)|\lesssim\xi$, then $\varepsilon_3(N)\lesssim 1/N$. Thus, Theorem 2.1 applies with $\delta_N=1/N$ and $\varepsilon_N=\log(N)/N$, which we summarize in the following corollary.

COROLLARY 2.2 (see [8, Theorem 4.1]). *Assume that $f(\cdot)$ is a piecewise smooth function whose first variation is BV bounded, so that* (2.3) *holds. Let $\mathcal{K}_N^\sigma(x)$ be a*

normalized concentration kernel (2.1) *with* $\frac{\sigma(\xi)}{\xi} \in C^1$. *Then* $K_N^\sigma f(x)$ *satisfies the concentration property*

$$|K_N^\sigma f(x) - [f](x)| \lesssim \delta_N + \frac{\log N}{N}\|\omega_x(\cdot)\|_{BV}, \qquad \omega_x(\cdot) \equiv \omega_f(\cdot, x). \tag{2.11}$$

Corollary 2.2 tells us that if x is sufficiently close to z_j, then $K_N^\sigma f(x)$ is close to $[f](z_j)$; but how close should x be? This is determined by the behavior of $\omega_f(\cdot, x)$, which in turn depends on the behavior of $f(\cdot)$ in the neighborhood of x. Our next theorem converts (2.11) into a precise error statement which depends solely on the position of x relative to the jump discontinuities but otherwise *is independent of the behavior of* f *near* x.

THEOREM 2.3 (concentration kernels revisited). *Assume that* $f(\cdot) \in BV[-\pi, \pi]$ *is a piecewise* C^2*-smooth function so that*

$$M_2 := \|f\|_{C^2\left(\cup_j (z_j, z_{j+1})\right)} < \infty.$$

Let $\mathcal{K}_N^\sigma(x)$ *be a normalized concentration kernel* (2.1) *with* $\sigma(\xi)/\xi \in C^1$. *Then, there exist constants depending on* M_2 *but otherwise independent of* N *such that* $K_N^\sigma f(x)$ *satisfies the concentration property*

$$K_N^\sigma f(x) = \begin{cases} [f](z_j) + \mathcal{O}\left(\dfrac{\log(N)}{N}\right) & \text{if } |x - z_j| \lesssim \dfrac{\log(N)}{N}, \quad j = 1, \dots, m, \\ \mathcal{O}\left(\dfrac{\log(N)}{N d_x}\right) & \text{if } d_x := dist\Big\{x, \{z_1, \dots, z_m\}\Big\} \gg \dfrac{1}{N}. \end{cases} \tag{2.12}$$

Remark 2.4. Theorem 2.3 tells us that admissible kernels $\mathcal{K}_N^\sigma(x)$ tend to concentrate in the immediate $\mathcal{O}(1/N)$-neighborhood of the jump discontinuities z_j's, where they approach the amplitude of these jumps $[f](z_j)$, while they decay to zero away from these jumps, where $\log(N)/Nd_x \sim 0$ as $d_x \gg 1/N$. We note in passing that for the concentration property $K_N^\sigma f(x) \to [f](x)$ to hold, it suffices to require the integrability $\omega_f(\cdot, x) \in L^1 \ \forall x$, e.g., [3, 7, 5], and the references therein. The BV regularity of $\omega_f(\cdot, x)$ enables us to derive the first-order concentration rate stated in (2.11), and, when f satisfies the stronger assumption of piecewise C^2 smoothness, we have the improved first-order rate (2.12). In practice, the decay rate may be even *faster* than first order, depending on the fine properties of σ; consult [4, p. 1396].

The proof of Theorem 2.3 is based on the following sharp upper-bound:

$$\|\omega_x(\cdot)\|_{BV} \lesssim d_x \|f\|_{C^2(I_x)} + \frac{1}{d_x}\|f\|_{BV}, \tag{2.13}$$

here and below, I_x denotes the largest *punctured* interval of smoothness enclosing x, and d_x denotes the distance to the nearest jump, that is,[4]

$$\begin{aligned} &I_x = (x - d_x, x) \cup (x, x + d_x), \\ &d_x := \begin{cases} \min\limits_{0 \le j \le m+1} |x - z_j| > 0 & \text{if } x \notin \{z_1, \dots, z_m\}, \\ \min\limits_{\pm}\{|z_j - z_{j\pm 1}|\} > 0 & \text{if } x = z_j, \quad j = 1, 2, \dots m. \end{cases} \end{aligned} \tag{2.14}$$

[4]By periodicity, $z_0 := z_m - 2\pi$ and $z_{m+1} := z_1 + 2\pi$.

We postpone the proof of (2.13) to the end of this section (consult Lemma 2.5 below), and we turn to the proof of Theorem 2.3, where we distinguish between three cases.

• *Case* (i). If x is a discontinuity point, say, $x = z_j$, then (2.13) with $d_k := d_{z_k}$ states that

$$\|\omega_{z_j}(\cdot)\|_{BV} \lesssim C_{11} := \max_k \left\{ d_k \|f\|_{C^2(I_{z_k})} + \frac{1}{d_k}\|f\|_{BV} \right\}, \qquad j = 1, 2, \ldots, m,$$

and Corollary 2.2 implies that

$$|K_N^\sigma f(z_j) - [f](z_j)| \le C_{11}\frac{\log N}{N}, \qquad j = 1, 2, \ldots, m. \tag{2.15a}$$

• *Case* (ii). Assume that x is near a discontinuity $|x - z_j| \lesssim \log(N)/N$. We note that $\mathcal{K}_N^\sigma$ is uniformly bounded:

$$\|\mathcal{K}_N^\sigma\|_{L^\infty} \lesssim \sum_{|k|\le N} \sigma\left(\frac{|k|}{N}\right)|\widehat{f}(k)| \lesssim \sum_{k=1}^{N} \frac{\sigma(\xi_k)}{\xi_k}\frac{1}{N}\|f\|_{BV} \lesssim C_{12}, \qquad C_{12} = (1+\delta_N)\|f\|_{BV},$$

which in turn implies the Lipchitz continuity of $K_N^\sigma f(x)$:

$$|K_N^\sigma f(x_2) - K_N^\sigma f(x_1)| \le \|\mathcal{K}_N^\sigma\|_{L^\infty} \int |f(x_2 - y) - f(x_1 - y)|dy \lesssim C_{12}|x_2 - x_1| \cdot \|f\|_{BV}. \tag{2.15b}$$

Combining (2.15a) and (2.15b) yields the first half of (2.12)

$$\begin{aligned} |K_N^\sigma f(x) - [f](z_j)| &\le |K_N^\sigma f(x) - K_N^\sigma f(z_j)| + |K_N^\sigma f(z_j) - [f](z_j)| \\ &\le C_{12}|x - z_j| \cdot \|f\|_{BV} + C_{11}\frac{\log N}{N} \lesssim \frac{\log N}{N}, \qquad |x - z_j| \lesssim \frac{\log(N)}{N}. \end{aligned}$$

• *Case* (iii). Finally, we assume that x is bounded away from the discontinuities $d_x \gg 1/N$. Then (2.13) and (2.11) yield the second half of (2.12)

$$|K_N^\sigma f(x)| \lesssim \frac{\log(N)}{N}\left(d_x\|f\|_{C^2(I_x)} + \frac{1}{d_x}\|f\|_{BV}\right) \lesssim \frac{\log(N)}{N d_x}, \qquad d_x \gg \frac{\log(N)}{N}.$$

We close this section with the following BV bound, which is the heart of the matter.

LEMMA 2.5 (BV bound). *Assume that $f(\cdot)$ is piecewise C^2-smooth, and let $I_x = (x - d_x) \cup (x, x + d_x)$ in* (2.14) *be the largest punctured interval of smoothness enclosing x. Then*

$$\|\omega_x(\cdot)\|_{BV} \lesssim d_x\|f\|_{C^2(I_x)} + \frac{1}{d_x}\|f\|_{BV}.$$

Proof. We decompose $\omega_x(y)$ into two symmetric parts:

$$\omega_x(y) = \frac{f(x + y) - f(x+)}{y} - \frac{f(x-) - f(x - y)}{y} =: \omega_+(y) - \omega_-(y).$$

To upper bound the variation of $\omega_+(\cdot)$ we compute

$$\begin{aligned} \omega_+'(y) &= -\frac{f(x + y) - f(x+) - yf'(x + y)}{y^2} \\ &= \frac{-1}{y^2}\left(\int_{x+}^{x+y} f'(\xi)d\xi - yf'(x + y)\right) \\ &= \frac{1}{y^2}\int_{x+}^{x+y} (\xi - x)f''(\xi)d\xi, \qquad x < y < x + d_x. \end{aligned}$$

The variation of ω_+ to the right of x is therefore upper bounded by

$$\begin{aligned}\|\omega_+(\cdot)\|_{BV(x+,\pi)} &\le \int_{y=0}^{d_x}\left|\int_{\xi=x+}^{x+y}\frac{(\xi-x)}{y^2}f''(\xi)d\xi\right|dy\\ &\quad+\int_{y=d_x}^{\pi}\left|\frac{f'(x+y)}{y}\right|+\left|\frac{f(x+y)-f(x+)}{y^2}\right|dy\\ &\le \frac{1}{2}\int_{y=0}^{d_x}\max_{\xi\in(x,x+d_x)}|f''(\xi)|dy\\ &\quad+\frac{1}{d_x}\int_{y=d_x}^{\pi}|f'(x+y)|+\left|\frac{f(x+y)-f(x+)}{y}\right|dy\\ &\le \frac{d_x}{2}\|f\|_{C^2(x,x+d_x)}+\frac{2}{d_x}\|f\|_{BV}.\end{aligned}$$

Similarly, the variation of ω_+ to the left of x does not exceed $\|\omega_+(\cdot)\|_{BV(-\pi,x-)} \le \frac{d_x}{2}\|f\|_{C^2(x-d_x,x)}+\frac{2}{d_x}\|f\|_{BV}$. We then have

$$\|\omega_+(\cdot)\|_{BV}\le [f](x)+d_x\|f\|_{C^2(I_x)}+\frac{4}{d_x}\|f\|_{BV}\lesssim d_x\|f\|_{C^2(I_x)}+\frac{1}{d_x}\|f\|_{BV}.$$

A similar bound holds for $\|\omega_-(\cdot)\|_{BV}$, and the lemma follows. ∎

3. Noisy data—a new perspective based on constrained minimization. Assume that f experiences a single jump discontinuity at location z of height $[f](z)$. This dictates a first-order decay of the Fourier coefficients

$$\widehat{f}(k)=[f](z)\frac{e^{-ikz}}{2\pi ik}+\widehat{s}(k)+\widehat{n}(k). \tag{3.1}$$

Here, $\widehat{s}(k)$ are associated with the regular part of f after extracting the jumps $[f](z_j)$; their decay is of order $\sim |k|^{-2}$ or faster, depending on the smoothness of the regular part $s(\cdot)$. The new aspect of the problem enters through the $\widehat{n}(k)$'s, which are the Fourier coefficients of the noisy part corrupting the smooth part of the data. We assume $n(\cdot)$ to be white noise whose mean-square power at each frequency is $E(|\widehat{n}(k)|^2)=\eta$. With (3.1), the conjugate sum (2.6) becomes

$$\begin{aligned}K_N^\sigma f(x)=[f](z)\times 2\pi i\sum_{k=1}^{N}\frac{\sigma_N\left(\frac{k}{N}\right)}{k}\cos k(x-z)-2\pi\sum_{k=1}^{N}\sigma\left(\frac{k}{N}\right)\widehat{s}(k)\sin kx\\ -2\pi\sum_{k=1}^{N}\sigma\left(\frac{k}{N}\right)\widehat{n}(k)\sin kx.\end{aligned}$$

We quantify the "energy" of each of the three sums on the right. E_J and E_R are associated with the discontinuous and regular parts of f:

$$\sum_{k=1}^{N}\left(\frac{\sigma\left(\frac{k}{N}\right)}{k}\right)^2\approx\frac{1}{N}\int_0^1\left(\frac{\sigma(\xi)}{\xi}\right)^2d\xi=:E_J(\sigma), \tag{3.2a}$$

$$\sum_{k=1}^{N}\sigma^2\left(\frac{k}{N}\right)|\widehat{s}(k)|^2\sim\frac{1}{N^3}\int_0^1\frac{\sigma^2(\xi)}{\xi^4}d\xi=:E_R(\sigma), \tag{3.2b}$$

and $E_\eta(\sigma)$ is associated with the noisy part of f, which was assumed to have variance η:

$$\sum_{k=1}^{N} \sigma^2\left(\frac{k}{N}\right) E(|\widehat{n}(k)|^2) \approx \eta N \int_0^1 \sigma^2(\xi)d\xi =: E_\eta(\sigma). \tag{3.2c}$$

Our perspective for construction of edge detectors for such noisy data is to treat the problem as a *constrained minimization*. We seek a function $\sigma(\xi)$ which minimizes the total energy, thus making the conjugate sum $K_N^\sigma f$ as localized as possible, subject to prescribed normalization constraint (2.1b),

$$\min\left\{ a_J E_J(\sigma) + a_\eta E_\eta(\sigma) + a_R E_R(\sigma) \;\middle|\; \int_0^1 \frac{\sigma(\xi)}{\xi} d\xi = 1 \right\}. \tag{3.3}$$

This yields

$$\sigma(\xi) = C_\sigma \frac{\frac{1}{\xi}}{a_J \frac{1}{N\xi^2} + a_\eta \eta N + a_R \frac{1}{N^3\xi^4}} = \frac{C_\sigma N^3 \xi^3}{a_J N^2 \xi^2 + a_\eta \eta N^4 \xi^4 + a_R}.$$

We ignore the relatively negligible contribution of the regular part which becomes even smaller as $s(\cdot)$ becomes smoother. Setting $a_R = 0$ we end up with concentration factors of the form

$$\sigma(\xi) = \frac{C_\sigma}{a_J} \cdot \frac{N\xi}{1 + \eta\beta^2 N^2 \xi^2}, \quad \beta := \sqrt{\frac{a_\eta}{a_J}}. \tag{3.4}$$

The corresponding concentration kernel depends only on the *relative* size of the amplitudes $\beta^2 = a_\eta / a_J$. Indeed, the normalization of $\sigma(\xi)/\xi$ (2.1b) causes the constant C_σ to satisfy

$$\int_0^1 \frac{\sigma(\xi)}{\xi} d\xi = \frac{C_\sigma}{a_J \sqrt{\eta}\beta} \tan^{-1}\left(\sqrt{\eta}\beta N\right) = 1,$$

and we end up with the *normalized* concentration factor

$$\sigma(\xi) \equiv \sigma_\eta(\xi) = \frac{1}{\tan^{-1}\left(\sqrt{\eta}\beta N\right)} \cdot \frac{\sqrt{\eta}\beta N \xi}{1 + \eta\beta^2 N^2 \xi^2}, \qquad \beta = \sqrt{\frac{a_\eta}{a_J}}. \tag{3.5}$$

The corresponding edge detector then takes the form

$$K_N^{\sigma_\eta} f(x) = \frac{\pi\sqrt{\eta}\beta}{\tan^{-1}\left(\sqrt{\eta}\beta N\right)} \sum_{|k|\le N} \frac{ik}{1 + \eta\beta^2 k^2} \widehat{f}(k) e^{ikx}. \tag{3.6}$$

The concentration factor $\sigma = \sigma_\eta$ now involves three factors: the ratio β, the noise variance η, and the number of modes N. The concentration kernel (3.6) tends to de-emphasize both the low frequencies which are "corrupted" by the jump discontinuity(ies) and the high frequencies which are corrupted by the noise. Different procedures yield different policies for the choice of $\beta = \beta(\eta)$; one will be discussed in the next subsection. It is worth noting the essential dependence of $\sigma_\eta(\xi)$ on the variance of the noise η. There are three scales involved—the small "smoothness" of order $\sim 1/N$, the noise scale of order $\sim \sqrt{\eta}$, and the $\mathcal{O}(1)$-scale of jump discontinuities. We distinguish between two cases. If $\sqrt{\eta} \ll 1/N$ so that $\sqrt{\eta}\beta N \ll 1$, then the

noise can be considered part of the smooth variation of f and $\sigma_\eta(\xi) \approx \xi$ recovers the usual concentration factor for noise-free data. Indeed, $\sigma_\eta(\xi) = \xi$ at the limit of $\eta \downarrow 0$. Otherwise, when the $\mathcal{O}(1/N)$-smoothness scale is dominated by the $\mathcal{O}(\sqrt{\eta})$-noise scale in the sense that $\sqrt{\eta}\beta \gtrsim 1/N$ in which we assume the noise to be still well below the $\mathcal{O}(1)$-scale of the jumps,

$$\frac{1}{N} \lesssim \sqrt{\eta}\beta \ll \mathcal{O}(1).$$

In this case, we can ignore the bounded factor $1/\tan^{-1}(\sqrt{\eta}\beta N)$, and we compute the small scale dictated by Theorem 2.1. Setting $\zeta(\xi) := \sqrt{\eta}\beta N\xi$ we find

$$\begin{aligned}
\delta_N &= \frac{1}{N}\left\|\frac{\sigma_\eta(\xi)}{\xi}\right\|_{BV} \lesssim \frac{\sqrt{\eta}\beta N}{N}\int \frac{1}{(1+\zeta^2)^2}d\zeta \lesssim \sqrt{\eta}\beta,\\
\varepsilon_1(N) &= \frac{1}{N}|\sigma_\eta(1)| \lesssim \frac{1}{\sqrt{\eta}\beta N^2},\\
\varepsilon_2(N) &= \frac{1}{N}\int_{1/N}^1 \frac{1-\zeta^2}{\xi(1+\zeta^2(\xi))^2}d\xi \lesssim \sqrt{\eta}\beta\log\big(\sqrt{\eta}\beta\big),\\
\varepsilon_3(N) &= \left|\sigma_\eta\left(\frac{1}{N}\right)\right| \lesssim \sqrt{\eta}\beta.
\end{aligned}$$

Hence (2.4) holds with

$$\varepsilon_N \equiv \varepsilon_\eta := \sqrt{\eta}\beta\log\big(\sqrt{\eta}\beta\big).$$

It is remarkable to see how the small scale of smoothness in the noiseless case $\mathcal{O}(\log(N)/N)$ is now replaced by the small scale of noise $\varepsilon_\eta = \mathcal{O}(\sqrt{\eta}\beta\log(\sqrt{\eta}\beta))$. We now appeal to (2.7): since $\sqrt{\eta} \ll 1$, Theorem 2.1 implies that $K_N^{\sigma_\eta} f$ separates the $\mathcal{O}(1)$-scale of the edges from the noise scale of order $\varepsilon_\eta \ll 1$.

THEOREM 3.1 (edge detection in noisy data). *Assume that $f(\cdot)$ is piecewise smooth in the sense that* (2.3) *holds. Assume that its spectral data contain white noise with variance $\eta \ll 1$. Let $K_N^{\sigma_\eta} f(x)$ be a normalized concentration kernel* (3.6)

$$K_N^{\sigma_\eta} f(x) = \frac{\pi\sqrt{\eta}\beta}{\tan^{-1}(\sqrt{\eta}\beta N)} \sum_{|k|\le N} \frac{ik}{1+\eta\beta^2k^2}\widehat{f}(k)e^{ikx}$$

associated with the concentration factor

$$\sigma_\eta(\xi) = \frac{1}{\tan^{-1}(\sqrt{\eta}\beta N)} \cdot \frac{\sqrt{\eta}\beta N\xi}{1+\eta\beta^2N^2\xi^2}.$$

We distinguish between two cases:

(i) *if $\sqrt{\eta}\beta N \ll 1$, we set the small scale $\varepsilon = \varepsilon_N := \log(N)/N$;*

(ii) *if $1/N \lesssim \sqrt{\eta}\beta \ll 1$, we set the small scale $\varepsilon = \varepsilon_\eta := \sqrt{\eta}\beta\log\big(\sqrt{\eta}\beta\big)$. Then, $K_N^{\sigma_\eta} f(x)$ satisfies the following concentration property:*

$$K_N^{\sigma_\eta} f(x) = \begin{cases} [f](z_j) + \mathcal{O}\big(\varepsilon\big), & |x - z_j| \lesssim \varepsilon, \quad j = 1, \dots, m,\\ \mathcal{O}\left(\dfrac{\varepsilon}{d_x}\right), & d_x = dist\Big\{x, \{z_1, \dots, z_m\}\Big\} \gg \varepsilon. \end{cases} \tag{3.7}$$

Remark 3.2. We conclude that the concentration factor advocated in Theorem 3.1 is a rescaling of

$$\sigma(\xi) = \frac{\xi}{1+\xi^2}.$$

Thus, $\sigma_\eta(\xi) = C_\sigma \sigma(\sqrt{\eta}\beta N \xi)$, where C_σ is the proper normalization factor.

4. Numerical results. I.

4.1. Noiseless data. To illustrate the results of the previous section, we present two sets of numerical results. We begin with the case of *noiseless data* depicted in Figure 4.1—a standard periodic sawtooth function with a single jump continuity at $x = 1$. We still use the concentration factors advocated in Theorem 3.1, $C_\sigma \sigma(\sqrt{\eta}\beta N \xi)$. We set $\beta = 1$, corresponding to equal weights for the errors due to the noise and the discontinuous parts of the signal. The edge detector concentrates near the unit jump discontinuity at $x = 1$, and it decays to zero away from it. Since the data are noiseless, η is treated as an extra parameter. As η gets smaller, $K_N^{\sigma_\eta} f(x)$ tends to zero faster for x away from 1. The results of Figure 4.1 show that, by decreasing the value of η by a factor of 10, the decay of $K_N^{\sigma_\eta} f(x)$ away from $x = 1$ is accelerated by a factor of $0.00365/0.00137 = 2.66 \approx \sqrt{10}$; thus, the decay ratio in the smooth regime is of order $\sqrt{\eta}$, as it should be.

4.2. Noisy data—balancing the different types of errors. We now turn to the case of *noisy data*. In order to choose the free parameter β, it is important to know how β influences the error at the output of our edge detector (3.6). Let us

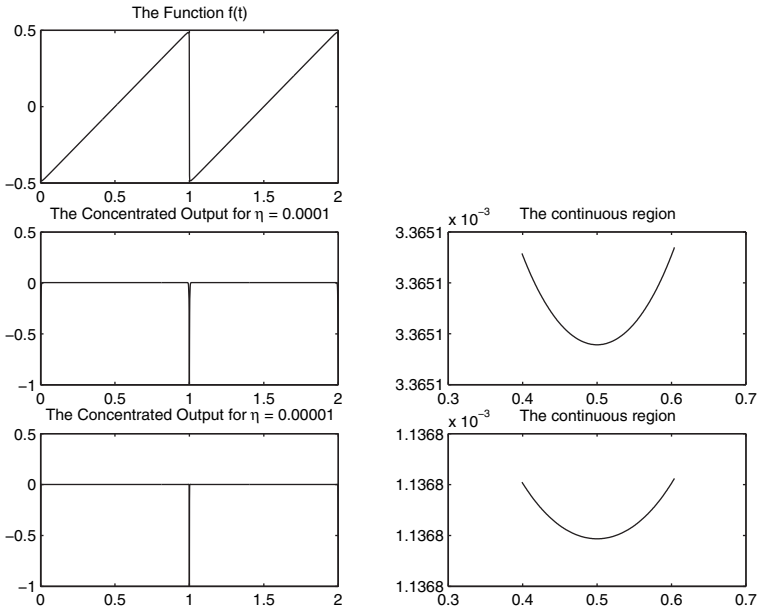


FIG. 4.1. *The output for various values of η when there is actually no noise at the input.*

consider E_η. We have seen that

$$\begin{aligned} E_\eta &= \eta N \int_0^1 \sigma^2(\xi)\, d\xi \\ &= \frac{\eta N}{(\tan^{-1}(\sqrt{\eta}\beta N))^2} \int_0^1 \frac{\eta\beta^2 N^2 \xi^2}{(1+\eta\beta^2 N^2\xi^2)^2}\, d\xi \\ &= \frac{\sqrt{\eta}}{\beta(\tan^{-1}(\sqrt{\eta}\beta N))^2} \int_0^{\sqrt{\eta}\beta N} \frac{\zeta^2}{(1+\zeta^2)^2}\, d\zeta \overset{\sqrt{\eta}\beta N \gtrsim 1}{\sim} \frac{\sqrt{\eta}}{\beta}. \end{aligned}$$

E_η is approximately the variance of the contribution due to noise to the edge detector. If we want to consider the *size* of the noise, we should consider the standard deviation of this contribution. It is customary to bound the noise by some constant multiple of the standard deviation—we will use the factor 2. We define the *effective size* of the noise's contribution to be

$$E_{\eta,eff} \equiv 2\sqrt{E_\eta} \sim \frac{\eta^{1/4}}{\sqrt{\beta}}.$$

By (3.2a), the effective contribution from the jump is of order

$$E_{J,eff} = E_J(\sigma_\eta) \sim \eta^{1/2}\beta.$$

Minimizing $E_{\eta,eff} + E_{J,eff}$ with respect to β yields $\beta \sim \eta^{-1/6}$. Figure 4.2 demonstrates the edge detected in noisy data using the concentration kernel (3.6) with the advocated $\beta = \pi\eta^{-1/6}$.

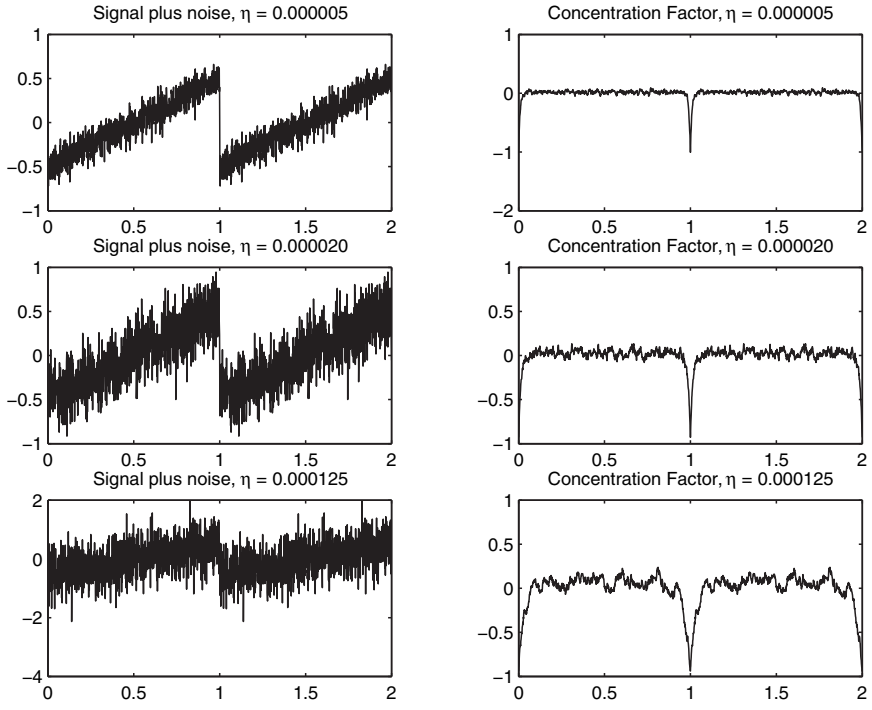


FIG. 4.2. *Detection of edges in a noisy sawtooth function corrupted with various values of* η *using the concentration kernel* (3.6) *with* $\beta = \pi\eta^{-1/6}$.

5. Noisy data and smoothness—concentration kernels revisited. As an alternative approach to the L^2-minimization offered in section 3, we now replace the L^2-"averaged" effect of the regular part taken in (3.2b) by the BV-like quantity

$$E_R(\sigma) \approx \sum_{k=1}^{N} \left| \sigma\left(\frac{k}{N}\right) \right| \cdot |\widehat{s}(k)|, \tag{5.1a}$$

where the regular part is sufficiently smooth that

$$|\hat{s}(k)| \sim \frac{1}{k^2}. \tag{5.1b}$$

As in (3.3), we consider the constrained minimization

$$\min \left\{ J(\sigma) \;\middle|\; \int_0^1 \frac{\sigma(\xi)}{\xi}\, d\xi = 1 \right\}, \tag{5.2a}$$

with $J(\sigma) = a_J E_J(\sigma) + a_\eta E_\eta(\sigma) + a_R E_R(\sigma)$, where $E_J(\sigma)$ and $E_\eta(\sigma)$ are given by (3.2a) and (3.2c), respectively, but with an alternative expression for the "energy" of the regular part motivated by (5.1): $E_R(\sigma) := \int_0^1 |\sigma(\xi)|\xi^{-2} d\xi/N$. We find that

$$J(\sigma) := \frac{a_J}{N} \int_0^1 \frac{\sigma^2(\xi)}{\xi^2}\, d\xi + a_\eta \eta N \int_0^1 \sigma^2(\xi)\, d\xi + \frac{a_R}{N} \int_0^1 \frac{|\sigma(\xi)|}{\xi^2}\, d\xi. \tag{5.2b}$$

Proceeding formally, the solution for the first variation of (5.2) leads to

$$\frac{2a_J\sigma(\xi)}{N\xi^2} + 2\eta a_\eta N\sigma(\xi) + \frac{a_R \cdot \operatorname{sgn}(\sigma(\xi))}{N\xi^2} = \frac{\lambda}{\xi}, \qquad \operatorname{sgn}(\sigma) = \begin{cases} 1, & \sigma > 0, \\ 0, & \sigma = 0, \\ -1, & \sigma < 0. \end{cases}$$

We will show that the resulting optimal concentration factor is given by

$$\sigma(\xi) = C_\sigma \frac{(N\xi - k_0)_+}{1 + \eta\beta^2 N^2 \xi^2}, \qquad k_0 := \frac{a_R}{\lambda}, \quad \beta = \sqrt{\frac{a_\eta}{a_J}}, \quad C_\sigma = \frac{\lambda}{2a_J}. \tag{5.3}$$

Indeed, to justify the passage to (5.3), one may consider a *regularized* version of the variational statement (5.2) $\min J_\varepsilon(\sigma)$, where

$$J_\varepsilon(\sigma) := \frac{a_J}{N} \int_0^1 \frac{\sigma^2(\xi)}{\xi^2}\, d\xi + a_\eta \eta N \int_0^1 \sigma^2(\xi)\, d\xi + \frac{a_R}{N} \int_0^1 \frac{|\sigma(\xi)|_\varepsilon}{\xi^2}\, d\xi$$

involves a *mollified* absolute value function

$$|\sigma|_\varepsilon := \begin{cases} |\sigma|, & |\sigma| \geq \epsilon, \\ \dfrac{\sigma^2}{2\epsilon} + \dfrac{\epsilon}{2}, & |\sigma| \leq \epsilon. \end{cases}$$

The solution of the corresponding regularized first variation yields the minimizer

$$\sigma(\xi) = C_\sigma \frac{N\xi - k_0 \cdot \operatorname{sgn}_\varepsilon(\sigma(\xi))}{\eta\beta^2 N^2 \xi^2 + 1}, \qquad \operatorname{sgn}_\varepsilon(\sigma) := \begin{cases} 1, & \sigma > \varepsilon, \\ \sigma/\varepsilon, & |\sigma| \leq \varepsilon, \\ -1, & \sigma < -\varepsilon. \end{cases}$$

Thus, we end up with the optimal concentration factor $\sigma(\xi) = \sigma_\varepsilon(\xi)$,

$$\sigma_\varepsilon(\xi) = \begin{cases} C_\sigma \dfrac{N\xi - k_0}{1 + \eta\beta^2 N^2 \xi^2}, & N\xi - k_0 > \varepsilon, \\ C_\sigma \dfrac{\varepsilon N \xi}{\varepsilon(1 + \eta\beta^2 N^2 \xi^2) + C_\sigma k_0}, & N\xi - k_0 \le \varepsilon, \end{cases}$$

and (5.3) is recovered by letting $\varepsilon \downarrow 0$. Clearly, the resulting optimal concentration factor is nonnegative.

It remains to calculate the normalization factor C_σ, for which

$$C_\sigma \int_{k_0/N}^1 \frac{1}{\xi} \frac{N\xi - k_0}{\eta\beta^2 N^2 \xi^2 + 1}\, d\xi = 1.$$

The integral on the left is found to be

$$\begin{aligned} &\int_{k_0/N}^1 \left(\frac{-k_0}{\xi} + \frac{k_0 \eta\beta^2 N^2 \xi + N}{\eta\beta^2 N^2 \xi^2 + 1} \right) d\xi \\ &\quad = k_0 \log\left(\frac{k_0}{N}\right) + \frac{k_0}{2} \log\left(\frac{\eta\beta^2 N^2 + 1}{\eta\beta^2 k_0^2 + 1}\right) + \frac{1}{\sqrt{\eta}\beta} \Big(\tan^{-1}(\sqrt{\eta}\beta N) - \tan^{-1}(\sqrt{\eta}\beta k_0) \Big). \end{aligned}$$

We focus our attention on the "noisy" case when $1/N \lesssim \sqrt{\eta}\beta \ll 1$ so that the fourth term on the right is negligible while the second term on the right is approximated by

$$\frac{k_0}{2} \log\left(\frac{\eta\beta^2 N^2 + 1}{\eta\beta^2 k_0^2 + 1}\right) \approx k_0 \log(\sqrt{\eta}\beta N).$$

We end up with an approximated integral

$$\int_{k_0/N}^1 \Big(\dots \Big) d\xi \approx k_0 \log(\sqrt{\eta}\beta N) + \frac{1}{\sqrt{\eta}\beta} \tan^{-1}(\sqrt{\eta}\beta N).$$

The balance between these two terms depends on the specific policy for β and the detailed balance between $\sqrt{\eta}\beta$ and N. Our normalized concentration factor takes the form

$$\sigma_\eta(\xi) = \frac{1}{\sqrt{\eta}\beta k_0 \log(\sqrt{\eta}\beta N) + \tan^{-1}(\sqrt{\eta}\beta N)} \cdot \frac{\sqrt{\eta}\beta (N\xi - k_0)_+}{1 + \eta\beta^2 N^2 \xi^2}. \tag{5.4a}$$

We can simplify this concentration factor in several ways; we mention two here.

(i) When N is large enough, we have $\tan^{-1}(\sqrt{\eta}\beta N) \approx \pi/2$ yielding

$$\sigma_\eta(\xi) = \frac{1}{k_0 \log(\sqrt{\eta}\beta N) + \pi/(2\sqrt{\eta}\beta)} \cdot \frac{(N\xi - k_0)_+}{1 + \eta\beta^2 N^2 \xi^2}. \tag{5.4b}$$

(ii) Observe that $\sigma_\eta(\xi)$ is rapidly decreasing at $\xi \approx 1$, with

$$\sigma_\eta(\xi) \sim \frac{1}{\eta\beta^2 N \log N}, \qquad \xi \approx 1,$$

so $\sigma_\eta(\xi)$ can be set to zero for $\xi \approx 1$ when N is large enough. In order to properly normalize the resulting concentration factor, N must be replaced by N_0. This leads us to the following:

$$\sigma_\eta\Big(\frac{k}{N}\Big) = \begin{cases} 0, & k < k_0, \\ \dfrac{1}{k_0 \log(\sqrt{\eta}\beta N_0) + \pi/(2\sqrt{\eta}\beta)} \cdot \dfrac{(k-k_0)_+}{1+\eta\beta^2 k^2}, & k_0 < k < N_0, \\ 0, & N_0 < k < N. \end{cases} \tag{5.4c}$$

6. Numerical results. II. We consider two examples depicted in Figure 6.1. In the first case, we have a noise of variance $\eta = 2\times 10^{-5}$ to be detected out of the first $N \gg 1000$ modes. The use of the edge detector (5.4) depends on three parameters, namely, β, k_0, and N_0. For β we use, as before, $\beta = \pi\eta^{-1/6} \sim 15$. The value of k_0 was chosen as $k_0 = 8\pi$: observe that, according to (5.3), $k_0 C_\sigma$ equals the ratio $a_R/2a_J$, and we found optimal results when the ratio of these amplitudes is set to be of order 10, yielding $k_0 = 10/0.3984 \sim 8\pi$. Finally, N_0 was taken as $N_0 = 1000$ since $\sigma_\eta(k/N) \le 0.1$ for $N \ge N_0$. With these parameters we find

$$\sigma_\eta\left(\frac{k}{N}\right) = \begin{cases} 0, & k = 1, \dots, 24, 1001, 1002, \dots, \\ 0.3985 \cdot \dfrac{(k-8\pi)_+}{1+\eta(15k)^2}, & 25 \le k \le 1000. \end{cases}$$

In the second case, the noise variance is $\eta = 4.5 \times 10^{-5}$. This led us to the choice of $\beta = \pi\eta^{-1/6} \sim 13$. Setting $k_0 = 6\pi$ (as before, $k_0 \sim 10/C_\sigma = 10/0.5070$) and $N_0 = 1000$, we have

$$\sigma_\eta\left(\frac{k}{N}\right) = \begin{cases} 0, & k = 1, \dots, 19, 1001, 1002, \dots, \\ 0.5070 \cdot \dfrac{(k-6\pi)_+}{1+\eta(13k)^2}, & 20 \le k \le 1000. \end{cases}$$

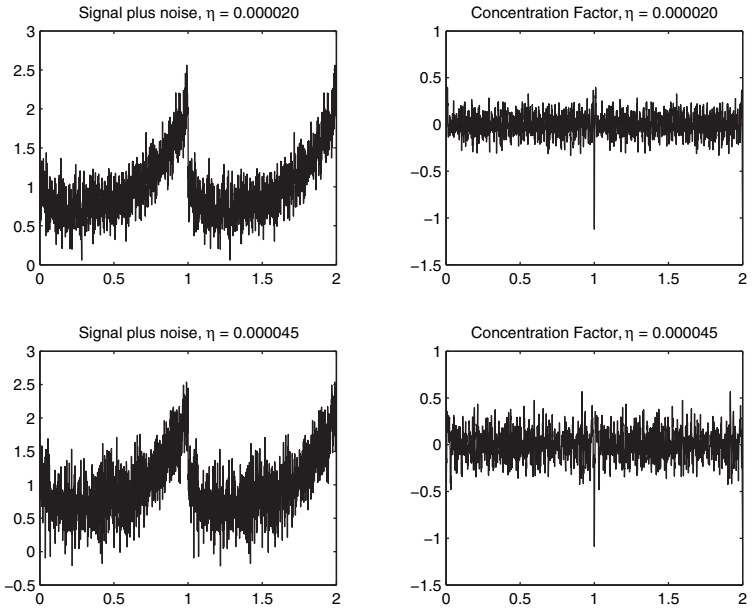


FIG. 6.1. *The output for various values of η when the input consists of a piecewise smooth function and noise (of the specified power spectral density).*

Note that in calculating the constants we made use of the *exact* normalization factor C_σ in (5.4b). $N_0 = 1000$ is not large enough to make the approximate normalization factor C_σ given in (5.4c) useful.

Note that, even with a large amount of white noise and a smooth signal, the location of the jump discontinuity is still clear. When considering jumps "corrupted" by low frequency data, we avoid low frequency signals by not using low frequency data. This helps keep the smooth signal from corrupting our results. On the other hand, because the jump discontinuity has most of its energy at low frequencies as well, our technique will increase the noise's effect. Comparing Figures 4.2 and 6.1, we find that the latter is not as clean as the former in the subfigure where the strength of the noise is the same.

Acknowledgment. Part of this research was carried out while the second author was visiting the Weizmann Institute, and he is grateful for their hospitality.

REFERENCES

[1] D. Cruz-Uribe and C. J. Neugebauer, *Sharp error bounds for the trapezoidal rule and Simpson's rule*, J. Inequal. Pure Appl. Math., 3 (2002), article 39.
[2] S. Engelberg, *Edge detection using Fourier coefficients*, Amer. Math. Monthly, to appear.
[3] A. Gelb and E. Tadmor, *Detection of edges in spectral data*, Appl. Comput. Harmon. Anal., 7 (1999), pp. 101–135.
[4] A. Gelb and E. Tadmor, *Detection of edges in spectral data* II. *Nonlinear enhancement*, SIAM J. Numer. Anal., 38 (2000), pp. 1389–1408.
[5] L. Hu and X. Shi, *Concentration factors for functions with harmonic bounded mean variation*, Acta Math. Hungar., 116 (2007), pp. 89–103.
[6] G. Polya and G. Szego, *Problems and Theorems in Analysis*, Vol. I, Springer-Verlag, New York, 1972.
[7] Q. Shi and X. Shi, *Determination of jumps in terms of spectral data*, Acta Math. Hungar., 110 (2006), pp. 193–206.
[8] E. Tadmor, *Filters, mollifiers and the computation of the Gibbs phenomenon*, Acta Numer., 16 (2007), pp. 305–378.

SIAM J. NUMER. ANAL.
Vol. 46, No. 5, pp. 2636–2661
© 2008 Society for Industrial and Applied Mathematics

ON NUMERICAL APPROXIMATIONS OF FORWARD-BACKWARD STOCHASTIC DIFFERENTIAL EQUATIONS*

JIN MA†, JIE SHEN‡, AND YANHONG ZHAO‡

Abstract. A numerical method for a class of forward-backward stochastic differential equations (FBSDEs) is proposed and analyzed. The method is designed around the *four step scheme* [J. Douglas, Jr., J. Ma, and P. Protter, *Ann. Appl. Probab.*, 6 (1996), pp. 940–968] but with a Hermite-spectral method to approximate the solution to the decoupling quasi-linear PDE on the whole space. A rigorous synthetic error analysis is carried out for a fully discretized scheme, namely a first-order scheme in time and a Hermite-spectral scheme in space, of the FBSDEs. Equally important, a systematical numerical comparison is made between several schemes for the resulting decoupled forward SDE, including a stochastic version of the Adams–Bashforth scheme. It is shown that the stochastic version of the Adams–Bashforth scheme coupled with the Hermite-spectral method leads to a convergence rate of $\frac{3}{2}$ (in time) which is better than those in previously published work for the FBSDEs.

Key words. forward-backward stochastic differential equations, four step scheme, Hermite-spectral method, convergence rate

AMS subject classifications. Primary, 60H10; Secondary, 34F05, 93E03

DOI. 10.1137/06067393X

1. Introduction.

Since the seminal work of Pardoux and Peng [32] in early 1990s, the theory of backward stochastic differential equations (BSDEs) and forward-backward stochastic differential equations (FBSDEs) has grown into a ubiquitous tool in the fields of stochastic optimizations and mathematical finance (see, for example, the books of El Karoui and Mazliak [14] and Ma and Yong [28] and the survey of El Karoui, Peng, and Quenez [15] for a detailed account of both theory and applications of such SDEs). In the meantime, after the earlier works of Bally [1] and Douglas, Ma, and Protter [13], finding an efficient numerical scheme for both BSDEs and FBSDEs has also become an independent but integral part of the theory. Tremendous efforts have been made during the past decade to circumvent the fundamental difficulties caused by the combination of the "backward" nature of the SDEs and the associated decoupling techniques for FBSDEs. In the "pure backward" (or "decoupled" forward-backward) case, various methods have been proposed. These include the PDE method in the Markovian case (e.g., Chevance [8] and Zhang and Zheng [36]), random walk approximations (e.g., Briand, Delyon, and Mémin [6] and Ma et al. [26]), Malliavin calculus and Monte-Carlo method (e.g., Zhang [35], Ma and Zhang [29], and Bouchard and Touzi [5]), the quantization method (e.g., Bally and Pagès [2] and Bally, Pagès, and Printems [3]), and recently the regression-based Monte-Carlo method (Lemor, Gobet, and Warin [23] and [24]).

In the case of coupled FBSDEs, however, the results are very limited, largely due to the lack of a solution method itself in such cases. It has been well understood

*Received by the editors November 2, 2006; accepted for publication (in revised form) March 10, 2008; published electronically July 2, 2008. This work was supported in part by NSF grants DMS-0311915, DMS-0505427, and DMS-0610646 and a Purdue Research Foundation grant.
http://www.siam.org/journals/sinum/46-5/67393.html
†Department of Mathematics, University of Southern California, Los Angeles, CA 90089-2532 (jinma@usc.edu).
‡Department of Mathematics, Purdue University, West Lafayette, IN 47907-1395 (shen@math.purdue.edu, zhao6@purdue.edu).

that in order to solve an FBSDE in an arbitrary duration, a (numerically) tractable method is to utilize the "decoupling PDE," based on the so-called four step scheme initiated in Ma and Yong [28]. Such an idea has led directly to most of the existing numerical results for FBSDEs: from the early work [13] to the recent improvements by Milstein and Tretyakov [31] and Delarue and Menozzi [12]. Extending such a decoupling idea and combining it with an optimal control method as well as Monte-Carlo simulation, Cvitanić and Zhang [9] and Bender and Zhang [4] recently proposed a numerical scheme for FBSDEs without attacking the associated PDE directly. We should note that the Monte-Carlo method is most effective only when a single value of the solution is concerned at each computing cycle, which is quite different from the original problem where essentially the distribution of the solution at each time point was sought.

There are two main technical obstacles in developing numerical schemes for FBSDEs: the *dimensionality* and the *rate of convergence*. The former is a natural consequence of the close relation between the FBSDE and its decoupling quasi-linear PDE, where the notorious "curse of dimensionality" is still a formidable difficulty for any numerical method. In fact, to the best of our knowledge, there is still no efficient numerical method for high dimensional PDEs that is directly applicable to our case. We note that almost all the existing numerical schemes for FBSDEs are constructed with convergence rate $\frac{1}{2}$, except in [31], where the Euler scheme for the forward SDE is replaced by Milstein's first order scheme so the rate of convergence is improved to 1. We should mention that although higher order approximation is possible for forward SDEs (cf., e.g., [21]), it is by no means clear whether this can be extended to the coupled FBSDEs, even when all the coefficients are assumed to be *smooth*(!), given the intrinsic difficulties arising from the current numerical methods.

This then raises an interesting question: Is it possible to design a numerical scheme that has better than first-order convergence rate, and at the same time is applicable (in the practical sense) to high-dimensional cases? This paper is an effort towards this goal. We shall revisit the four step scheme again, but will replace the usual finite difference method for the PDE by a *Hermite-spectral method*. Several main features of the Hermite-spectral method are worth noting: (i) the PDE solution is approximated directly by Hermite functions as an orthonormal basis on the whole space, without using ad hoc artificial boundary conditions as in previous approaches using finite differences; (ii) the rate of convergence is related directly to the regularity of the coefficients; to wit, the higher regularity implies the higher convergence rate. We should remark that this last feature marks the main difference between a spectral method and traditional finite difference and finite element methods. In fact, as we pointed out before, even given smooth coefficients, none of the existing schemes for FBSDEs seems to be able to achieve higher than first-order convergence rate.

Given the aforementioned spectral accuracy in space, it seems hopeful that one can produce arbitrarily higher-order global error by applying a higher-order scheme in SDEs in, e.g., [21], at least when the coefficients are smooth. We shall carry out some numerical simulations to validate this point. To compare with the existing results, we shall use an example proposed in Milstein and Tretyakov [31], and we test several methods for the forward SDEs. These include the standard Euler scheme, the Milstein scheme, the Platen–Wagner scheme, and the *stochastic Adam–Bashforth* (SAB) scheme proposed in [16]. By the nature of these schemes, we expect the Euler scheme to be 1/2-order convergent, the Milstein first-order scheme to be first-order convergent, and the Platen–Wagner and SAB schemes to be 3/2-order convergent. Our simulation results give numerical evidence that this is exactly the case. Also, this

example shows that in the one-dimensional case our scheme has the best performance (in terms of computing time). Due to the apparent complexity of the error analysis and the length of the paper, in this paper we only give a rigorous proof of the rate of convergence with the Euler scheme for the forward SDE. A rigorous proof for higher order schemes would be quite similar, although conceivably much more tedious. We hope our numerical results are sufficiently convincing for this purpose.

As a final remark, we would like to point out that the numerical method in this paper, albeit technical, can be extended to higher-dimensional cases with a tensor-product approach. However, such a method quickly becomes nonfeasible for spatial dimensions higher than three, unless some non–tensor-product method, such as those based on sparse grid (cf. [7] and [34]) or lattice rules (cf. [25], [19], and [22]), is introduced. We plan to address the higher dimensional issue in a forthcoming work.

The rest of the paper is organized as follows. In section 2 we give the necessary preliminaries; in section 3 we introduce the Hermite-spectral method for the PDE and perform an error analysis. In section 4 we study the synthetic error analysis of the full numerical scheme, and in section 5 we carry out some numerical experiments.

2. Problem formulation and preliminaries. Throughout this paper we assume that $(\Omega, \mathcal{F}, P)$ is a complete probability space, on which is defined a d-dimensional Brownian motion $W = \{W(t) : t \geq 0\}$. We shall denote $\mathbf{F}^W = \{\mathcal{F}_t^W : t \geq 0\}$ to be the natural filtration generated by W, with usual P-augmentation so that it is right continuous and contains all the P-null sets in $\mathcal{F}$.

We consider the following FBSDE: for $t \in [0, T]$,

$$\begin{cases} X(t) = x + \displaystyle\int_0^t b(s, \Theta(s))ds + \int_0^t \sigma(s, X(s), Y(s))dW(s), \\ \\ Y(t) = \varphi(X(T)) + \displaystyle\int_t^T g(s, \Theta(s))ds - \int_t^T Z(s)dW(s), \end{cases} \tag{2.1}$$

where $\Theta \stackrel{\triangle}{=} (X, Y, Z)$. To simplify presentation, in what follows we shall assume that all processes involved are one dimensional. Moreover, we shall make use of the following *standing assumptions* in what follows.

(A1) The functions b, σ, g, and φ are continuously differentiable in all variables. Moreover, if we denote these functions by a generic one ψ, then there exists a constant $K > 0$ such that, for any $t \in [0, T]$, ψ satisfies the uniform Lipschitz condition:

$$|\psi(t, x, y, z) - \psi(t, x', y', z')| \leq K(|x - x'| + |y - y'| + |z - z'|).$$

(A2) The functions b, g, and φ satisfies the following growth conditions: for some $K > 0$,

$$|g(t, x, y, z)| \leq K(1 + |y| + |z|),$$

$$|\varphi(x)|, |b(t, x, y, z)| \leq K.$$

(A3) Function σ has bounded second derivatives, and there exist constants $0 < c < C$ such that

$$c \leq \sigma(t, x, y) \leq C \quad \forall\, (t, x, y) \in [0, T] \times \mathbb{R}^2. \tag{2.2}$$

Remark 2.1. We remark that assumptions (A1)–(A3) are stronger than is necessary for the well-posedness of FBSDE (2.1). In fact, in [10] it was shown that under (A1)–(A3) but without the differentiability assumption on the coefficients the FBSDE (2.1) already possesses a unique adapted solution over an arbitrarily prescribed time duration. The extra smoothness condition is needed only for our numerical scheme, and therefore is not essential, in principle. Note also, however, that even with the added differentiability conditions, our assumptions are still much weaker than that of [13].

Four step scheme. In [27] (see also [10]) it was shown that the *unique* adapted solution of FBSDE (2.1) can be obtained by the following steps:

Step 1. Define a function $z:\ [0,T]\times\mathbb{R}^3\ \to\ \mathbb{R}$ by

$$z(t,x,y,p) = p\,\sigma(t,x,y) \quad \forall\ (t,x,y,p).$$

Step 2. Using the function z above, solve the quasi-linear parabolic PDE :

$$\begin{cases} u_t + \dfrac{1}{2}\sigma^2(t,x,u)u_{xx} + b(t,x,u,z(t,x,u,u_x))u_x \\ \quad + g(t,x,u,z(t,x,u,u_x)) = 0, \qquad (t,x)\in(0,T)\times\mathbb{R}, \\ u(T,x) = \varphi(x), \quad x\in\mathbb{R}. \end{cases} \tag{2.3}$$

Step 3. Using the functions u and z, solve the forward SDE:

$$X(t) = x + \int_0^t \tilde{b}(s,X(s))ds + \int_0^t \tilde{\sigma}(s,X(s))dW(s), \tag{2.4}$$

where $\tilde{b} = b(t,x,u(t,x),z(t,x,u,u_x))$ and $\tilde{\sigma} = \sigma(t,x,u(t,x))$.

Step 4. Set

$$Y(t) = u(t,X(t)) \quad \text{and} \quad Z(t) = \sigma(t,X(t),u(t,X(t)))u_x(t,X(t)).$$

Then, (X,Y,Z) is the adapted solution to (2.1).

It is readily seen that if a numerical scheme is designed along the lines of four step scheme, then one essentially has to deal with two separate discretization schemes: one for the PDE (2.3) and the other for the (forward) SDE (2.4). We shall adopt a Hermite–Galerkin method with numerical integration. To do this we find it convenient to rewrite PDE (2.3) in a divergence form, mainly for the sake of numerical stability:

$$\begin{cases} u_t + \partial_x\left(\dfrac{1}{2}\sigma^2(t,x,u)u_x\right) + \hat{b}(t,x,u,u_x)u_x \\ \quad + g(t,x,u,\sigma(t,x,u)u_x) = 0, \qquad (t,x)\in(0,T)\times\mathbb{R}, \\ u(T,x) = \varphi(x), \quad x\in\mathbb{R}, \end{cases} \tag{2.5}$$

where

$$\hat{b}(t,x,y,z) = b(t,x,y,\sigma(t,x,y)z) - \sigma(t,x,y)(\sigma_x(t,x,y) + \sigma_y(t,x,y)z).$$

Remark 2.2. We note here that the main difference between b and $\hat{b}$ is that the latter now has a linear growth in the variable z, and thus it no longer satisfies assumptions (A1) and (A2). We shall make the following adjustment in what follows

to facilitate the error analysis. Note that under (A1)–(A3) it is known (cf. [11]) that the solution to the PDE (2.3), whence (2.5), has a globally bounded gradient u_x. Thus in light of the four step scheme, if we introduce the "cut-off" version of the coefficient function $\hat{b}$,

$$\hat{b}_K(t,x,y,z) = b(t,x,y,\sigma(t,x,y)z) - \sigma(t,x,y)(\sigma_x(t,x,y) + \sigma_y(t,x,y)\gamma_K(z)),$$

where $\gamma_K(z) \in C^\infty(\mathbb{R})$, $\gamma_K(z) = z$ for $|z| \le K$, and $|\gamma_K(z)| \le K+1$, $|\gamma_K'(z)| \le 1$, then for K large enough, the solution of the PDE (2.5) with coefficient $\hat{b}_K$ will coincide with the original solution. In other words, we can simply consider the "truncated" coefficients $\hat{b}_K$ for the purpose of approximation, which again satisfies (A1) and (A2). For the sake of simplicity, in what follows we shall still use $\hat{b}$ to denote $\hat{b}_K$.

Hermite polynomials, Hermite functions, and their properties. Throughout this paper we shall denote by $L^2(\mathbb{R})$ the space of all square integrable functions $u : \mathbb{R} \mapsto \mathbb{R}$, with the inner product

$$\langle u, v\rangle \;=\; \int_{\mathbb{R}} u(x)v(x)dx \qquad\qquad \forall\, u, v \in L^2(\mathbb{R})$$

and the norm $\|u\|^2 \triangleq \int_{\mathbb{R}} |u(x)|^2 dx$. For simplicity, in what follows, without further specification we shall always denote by $\|\cdot\|$ the $L^2(\mathbb{R})$ norm.

Next, for each integer $k \ge 1$, we denote the differential operator $\partial^k \triangleq \frac{d^k}{dx^k}$. Then the Sobolev space $H^m(\mathbb{R})$ is defined by

$$H^m(\mathbb{R}) = \{u \mid \partial^k u \in L^2(\mathbb{R}) \quad \forall\, 0 \le k \le m\},$$

with the seminorm $|u|_m \triangleq \|\partial^m u\|$, and the norm $\|u\|_m^2 \triangleq \sum_{k=0}^m \|\partial^k u\|^2$. We note that the operator ∂^k can be easily extended to the higher-dimensional case, namely, as a partial differential operator, in an obvious way. We define $H^{-1}(\mathbb{R})$ as the dual space of $H^1(\mathbb{R})$, and its norm will be denoted by $\|\cdot\|_{-1}$.

We recall that (cf., e.g., [17]) the *Hermite polynomials* $\{H_n(x)\}_{n\ge 0}$ are defined by

$$H_n(x) = (-1)^n \exp(x^2)\frac{d^n}{dx^n}\exp(-x^2), \qquad n = 0, 1, \ldots . \tag{2.6}$$

One can easily check that H_n is the solution to the recursive equations

$$\begin{cases} H_{n+1}(x) = 2xH_n(x) - 2nH_{n-1}(x), \\ H_n'(x) = 2nH_{n-1}(x), \\ H_0(x) = 1, \quad H_1(x) = 2x, \end{cases}$$

and the following orthogonality condition holds:

$$\int_{-\infty}^{\infty} H_m(x)H_n(x)e^{-x^2}dx = \gamma_n\delta_{mn}, \qquad m, n \ge 1, \tag{2.7}$$

where $\gamma_n = \sqrt{\pi}2^n n!$ and δ_{mn} is the Kronecker delta. For each $n \ge 1$, $H_n(x)$ is called the "Hermite polynomial of degree n," and the *Hermite function of degree* n is defined by

$$\widehat{H}_n(x) = \frac{1}{\sqrt{2^n n!}} e^{-x^2/2} H_n(x).$$

It follows immediately that the Hermite functions $\widehat{H}_n(x)$ enjoy the following recurrence and orthogonal relations:

$$\widehat{H}_{n+1}(x) = x\sqrt{\frac{2}{n+1}}\widehat{H}_n(x) - \sqrt{\frac{n}{n+1}}\widehat{H}_{n-1}(x), \quad n \geq 1, \tag{2.8}$$

$$\frac{d}{dx}\widehat{H}_n(x) = \sqrt{\frac{n}{2}}\widehat{H}_{n-1}(x) - \sqrt{\frac{n+1}{2}}\widehat{H}_{n+1}(x), \tag{2.9}$$

$$\int_{-\infty}^{\infty} \widehat{H}_m(x)\widehat{H}_n(x)dx = \sqrt{\pi}\delta_{mn}. \tag{2.10}$$

Moreover, we derive from $H_n' = 2nH_{n-1}$ that $\widehat{H}_n' + x\widehat{H}_n = \sqrt{2n}\widehat{H}_{n-1}$. Hence, by introducing a new operator $Du \stackrel{\triangle}{=} \partial_x u + xu$, we obtain the following recursive relation:

$$D\widehat{H}_n = \sqrt{2n}\widehat{H}_{n-1}. \tag{2.11}$$

Using the operator D we can define a Hilbert space similar to the Sobolev space $H^m(\mathbb{R})$, which will be important for our error analysis. To begin with, for any positive integer N, let $\mathcal{P}_N$ be the space of polynomials of degree less than or equal to N. We then define

$$\mathcal{H}_N = e^{-x^2/2}\mathcal{P}_N \stackrel{\triangle}{=} \{e^{-x^2/2}p(x) :\ p(x) \in \mathcal{P}_N\}. \tag{2.12}$$

Next, for each $N \in \mathbb{N}$ let us denote $\{x_i\}_{i=0}^N$ to be the roots of the polynomial $H_{N+1}(x)$ and define

$$w_j = \frac{\sqrt{\pi}2^N N!}{(N+1)[H_N(x_j)]^2}, \quad 0 \leq j \leq N. \tag{2.13}$$

The pairs $\{(x_i, \omega_i)\}_{i=0}^N$ are known as the *Hermite–Gauss quadrature points and weights*, respectively (cf., e.g., [33]). Using the Hermite–Gauss quadrature points and weights, we can then define a discrete scalar product defined on $C(\mathbb{R})$ by

$$\langle f, g\rangle_N \stackrel{\triangle}{=} \sum_{j=0}^{N} f(x_j)g(x_j)\hat{w}_j \qquad \forall f, g \in C(\mathbb{R}), \tag{2.14}$$

where $\hat{w}_j$'s are the normalized Hermite–Gauss quadrature weights defined by

$$\hat{w}_j \stackrel{\triangle}{=} w_j\ e^{x_j{}^2} = \frac{\sqrt{\pi}}{(N+1)\widehat{H}_N^2(x_j)}, \qquad 0 \leq j \leq N.$$

A direct consequence of the Hermite–Gauss quadrature is

$$\langle u, v\rangle\ =\ \langle u, v\rangle_N \qquad \forall\ u \in \mathcal{H}_m,\ v \in \mathcal{H}_n,\ n + m \leq 2N + 1. \tag{2.15}$$

For any integer $m \geq 0$, let us define

$$H_D^m(\mathbb{R}) \stackrel{\triangle}{=} \{u \mid D^m u \in L^2(\mathbb{R})\},$$

equipped with norm $\|D^m u\|$. Moreover, for any real number $r \geq 0$, the space $H_D^r(\mathbb{R})$ and its norm can be defined by the usual space interpolation.

In order to introduce our numerical approximation we need the following two operators:

- the "*orthogonal projection operator*" $P_N : L^2(\mathbb{R}) \to \mathcal{H}_N$, defined by
$$\langle u - P_N u, v\rangle = 0 \quad \forall u \in L^2(\mathbb{R}),\ v \in \mathcal{H}_N;$$
- the "*Hermite–Gauss interpolation operator*" $I_N : C(\mathbb{R}) \to \mathcal{H}_N$, defined by
$$I_N f \in \mathcal{H}_N \quad \text{such that} \quad I_N f(x_j) = f(x_j), \quad j = 0, 1, \ldots, N. \tag{2.16}$$

The following basic approximation results for P_N can be found in, e.g., [18].

LEMMA 2.3. *For any $u \in H^r_D(\mathbb{R})$, $N \in \mathbb{N}$, and $0 \le s \le r$, there exists a constant $C > 0$, independent of N and the function u, such that*

$$\|D^s(u - P_N u)\| \le C N^{\frac{s-r}{2}} \|D^r u\| \tag{2.17}$$

and

$$\|\partial_x(u - P_N u)\| \le C N^{\frac{1-r}{2}} \|D^r u\|. \tag{2.18}$$

3. Numerical schemes. In this section we present our numerical schemes, both for the PDE (2.5) and for the forward SDE (2.4). Although the scheme for the general quasi-linear PDE (2.5) can be treated the same way with more complicated notation, to simplify presentation we shall assume that the coefficients b and g are both independent of variable z. We note that under such a simplification the PDE (2.5) becomes, with a slight abuse of notation in light of Remark 2.2,

$$\begin{cases} u_t + \partial_x\left(\frac{1}{2}\sigma^2(t,x,u)u_x\right) + \hat{b}(t,x,u,u_x)u_x + g(t,x,u) = 0, \\ \qquad\qquad\qquad\qquad\qquad\qquad (t,x) \in (0,T)\times\mathbb{R}, \\ u(T,x) = \varphi(x), \quad x \in \mathbb{R}, \end{cases} \tag{3.1}$$

where $\hat{b}(t,x,y,z) = b(t,x,y) - \sigma(t,x,y)(\sigma_x(t,x,y) + \sigma_y(t,x,y)\gamma_{\bar{K}}(z))$, for some $\bar{K} > 0$. We shall design a numerical scheme for (3.1) based on a semi-implicit discretization in time and the Hermite-collocation method in space, and discuss three types of schemes for the corresponding forward SDE (2.4). The error analysis for the combined scheme will be carried out in the next two sections.

Before we proceed further, we should note that the constant $\bar{K} > 0$ above is determined by the *true* solution, which is independent of the numerical scheme, and is useful only for error analysis. In what follows we shall denote all the constants depending on K in (A1)–(A3) and $\bar{K}$ above by a generic one, and still denote it by K. Then, from (A1) we see that $|\sigma_x(t,x,y)| \le K$ and $|\sigma_y(t,x,y)| \le K$ for all $(t,x,y) \in [0,T]\times\mathbb{R}^2$ follow easily. Furthermore, by (A3) it implies that σ_y satisfies the Lipschitz condition as

$$|\sigma_y(t,x,y) - \sigma_y(t,x,y')| \le K\,|y - y'|.$$

Hence,

$$\begin{aligned} &|\sigma(t,x,y)\sigma_y(t,x,y)\gamma_{_K}(z) - \sigma(t,x,y')\sigma_y(t,x,y')\gamma_{_K}(z')| \\ &\le |\sigma(t,x,y)\sigma_y(t,x,y)\gamma_{_K}(z) - \sigma(t,x,y')\sigma_y(t,x,y)\gamma_{_K}(z)| \\ &\quad + |\sigma(t,x,y')\sigma_y(t,x,y)\gamma_{_K}(z) - \sigma(t,x,y')\sigma_y(t,x,y')\gamma_{_K}(z)| \\ &\quad + |\sigma(t,x,y')\sigma_y(t,x,y')\gamma_{_K}(z) - \sigma(t,x,y')\sigma_y(t,x,y')\gamma_{_K}(z')| \\ &\le K^2(K+1)|y-y'| + CK(1+K)|y-y'| + CK|z-z'|. \end{aligned}$$

Similarly, we have

$$|\sigma(t,x,y)\sigma_x(t,x,y) - \sigma(t,x,y')\sigma_x(t,x,y')| \le (K^2 + CK)|y - y'|.$$

For notational simplicity, we often allow the generic constant K to vary from line to line; then by the above arguments, the function $\hat{b}$ satisfies the Lipschitz condition

$$|\hat{b}(t,x,y,z) - \hat{b}(t,x,y',z')| \le K(|y - y'| + |z - z'|). \tag{3.2}$$

On the other hand, $\hat{b}$ satisfies the bounded property by boundedness of b and $\gamma_{_K}(z)$; i.e.,

$$|\hat{b}(t,x,y,z)| \le K. \tag{3.3}$$

3.1. Numerical scheme for the PDE (3.1). For any $M \in \mathbb{N}^*$, let $h = \frac{T}{M}$ be the length of the time step and set $t_k = kh$, $k = 0, 1, \ldots, M$.

We begin with a first-order semi-implicit time discretization: Let $u^M(x) = \varphi(x)$. For $k = M-1, \ldots, 0$, we compute u^k by solving the following discretized version of (3.1):

$$\begin{aligned} &\frac{u^{k+1} - u^k}{h} + \frac{1}{2}\partial_x\Big(\sigma^2(t_k, x, u^{k+1})u_x^k\Big) \\ &\quad + \hat{b}(t_k, x, u^{k+1}, u_x^{k+1})u_x^{k+1} + g(t_k, x, u^{k+1}) = 0. \end{aligned} \tag{3.4}$$

In other words, at each time step, the problem is reduced to solving the following elliptic equation:

$$u^k - \frac{h}{2}\partial_x\Big(\sigma^2(t_k, x, u^{k+1})u_x^k\Big) = f_h^{k+1}(x), \quad \lim_{x\to\pm\infty} u^k(x) = 0, \tag{3.5}$$

where

$$f_h^{k+1}(x) = u^{k+1} + h\Big\{\hat{b}(t_k, x, u^{k+1}, u_x^{k+1})u_x^{k+1} + g(t_k, x, u^{k+1})\Big\}.$$

Since the problem is set on the whole domain, it is natural to consider a *Hermite-spectral method*. To be more precise, let $\{x_j\}_{j=0}^N$ be the Hermite–Gauss quadrature points and let $\mathcal{H}_N$ be defined in (2.12). A Hermite-collocation method for (3.5) is as follows:

Let $u_N^M(x) = I_N\varphi$. For $k = M-1, \ldots, 0$, we compute $u_N^k \in \mathcal{H}_N$ such that

$$u_N^k(x_j) - \frac{h}{2}\partial_x\Big(I_N(\sigma^2(t_k, x, u_N^{k+1})\partial_x u_N^k)\Big)(x_j) = f_h^{k+1}(x_j), \quad j = 0, 1, \ldots, N. \tag{3.6}$$

Using the identity (2.15) and integration by parts, we can rewrite (3.6) in the following variational formulation: Find $u_N^k \in \mathcal{H}_N$ such that

$$\langle u_N^k,\, v_N\rangle_N + \frac{h}{2}\langle\sigma^2(t_k, x, u_N^{k+1})\partial_x u_N^k,\, \partial_x v_N\rangle_N = \langle f_{h,N}^{k+1},\, v_N\rangle_N \quad \forall v_N \in \mathcal{H}_N, \tag{3.7}$$

where $f_{h,N}^{k+1} \in \mathcal{H}_N$ and $f_{h,N}^{k+1}(x_j) = f^{k+1}(x_j)$, $j = 0, 1, \ldots, N$.

We remark here that while the first-order (time discretization) scheme is simple to use and analyze, in practice one often prefers higher-order schemes. These schemes could be easily constructed and often implemented in a straightforward manner, but

their error analysis becomes rather tedious. In fact, as we will see in the next sections, the error analysis for the first-order semi-implicit scheme presented in (3.7) is already unpleasantly lengthy. As an example, we present the following second-order backward difference formula (BDF) Adam–Bashforth scheme which will be used in our numerical experiments without theoretical error analysis.

The second-order BDF Adam–Bashforth scheme.

Let $u_N^M(x) = \varphi(x)$ and let $u_N^{M-1}(x)$ be computed from (3.4).

For $k = M-2, \dots, 1, 0$, we compute u^k from

$$\frac{-3u_N^{k+2} + 4u_N^{k+1} - u_N^k}{2h} + \partial_x\Big(\frac{\sigma^2}{2}\partial_x u_N^k\Big) + (2\partial_x u_N^{k+1} - \partial_x u_N^{k+2})\hat{b} + g = 0, \tag{3.8}$$

where

$$\sigma^2 \triangleq \sigma^2(t_k, x, 2u_N^{k+1} - u_N^{k+2}), \quad g \triangleq g(t_k, x, 2u_N^{k+1} - u_N^{k+2}),$$

$$\hat{b} \triangleq \hat{b}(t_k, x, 2u_N^{k+1} - u_N^{k+2}, 2\partial_x u_N^{k+1} - \partial_x u_N^{k+2}).$$

We note that this scheme still leads to an elliptic equation of the form (3.5) for u^k at each time step. In particular, this second-order scheme should be used together with the $\frac{3}{2}$-order SDE scheme presented below.

3.2. Numerical schemes for the SDE (2.4). We now turn our attention to the discretization of the forward SDE (2.4). We begin with the simplest one, known as the "*forward Euler scheme*." Assuming that for each $N \in \mathbb{N}$ we have obtained the approximating solution to the PDE (3.1), denoted by u_N^k, $k = 0, 1, \dots, M$, at each time $t_k = kh$ we define recursively the approximate solution to the SDE (2.4) by

$$\begin{cases} X_0^N = x, \\ X_{k+1}^N = X_k^N + b(t_k, X_k^N, u_N^k(X_k^N))h + \sigma(t_k, X_k^N, u_N^k(X_k^N))\Delta_k W, \ k = 0, 1, \dots, M, \end{cases} \tag{3.9}$$

where $\Delta_k W \triangleq W(t_k + h) - W(t_k)$. For notational simplicity in what follows we shall simply denote $X_k = X_k^N$ when the context is clear.

It is well understood that the Euler scheme is easy to implement and has the minimum requirements on the coefficients. Furthermore, if we denote the true solution to the PDE (3.1) by u, denote $u^k(x) \triangleq u(t_k, x)$, and define an intermediate approximation $\hat{X}$ by

$$\begin{cases} \hat{X}_0 = x, \\ \hat{X}_{k+1} = \hat{X}_k + b(t_k, \hat{X}_k, u^k(\hat{X}_k))h + \sigma(t_k, \hat{X}_k, u^k(\hat{X}_k))\Delta_k W, \quad k = 0, 1, \dots, M, \end{cases} \tag{3.10}$$

then by the *fundamental convergence theorem* (cf., e.g., [21]), the mean-square error of the Euler scheme is $E|X(t_k) - \hat{X}_k|^2 \sim h$, where X is the true solution of (2.4) (namely, the rate of convergence is $\frac{1}{2}$).

In order to obtain a higher-order rate of convergence, one has to use a higher order scheme for the forward SDE. However, we should note that while in theory arbitrarily high-order schemes for (2.4) can be constructed with Taylor–Itô-type expansions (see, e.g., [21]), the complexity of these schemes increases drastically. Consequently it often

becomes too "expensive" in computational terms to implement. We will consider the following higher-order schemes in our numerical experiments to test the numerical accuracy and to compare our method with existing results.

A. *Milstein scheme* (cf. [21], [31]).

This is a well-known scheme with a first-order convergence rate. The recursive relation, adapted to our case, is given as follows:

(3.11)
$$\begin{cases} X_0 = x, \\ X_{k+1} = X_k + b(t_k, X_k, u_N^k(X_k))h + \sigma(t_k, X_k, u_N^k(X_k))\Delta_k W \\ \qquad +\dfrac{1}{2}\sigma(t_k, X_k, u_N^k(X_k))\{\sigma_x(t_k, X_k, u_N^k(X_k)) + \sigma_u(t_k, X_k, u_N^k(X_k))\partial_x u_N^k(X_k)\} \\ \qquad \times(\Delta_k^2 W - h), \qquad k = 0, 1, \ldots, M-1, \end{cases}$$

where $\Delta_k^2 W \stackrel{\triangle}{=} (W(t_{k+1}) - W(t_k))^2$.

The next two schemes require higher-order regularity of the coefficients. We shall assume that all such requirements, whenever needed, are fulfilled without further specification.

B. *Platen–Wagner scheme* (cf. [21]).

Using the idea of Taylor–Itô expansion up to order $\frac{3}{2}$, and assuming that the coefficients are actually twice continuously differentiable, Platen and Wagner proposed the following scheme: for $k = 0, 1, \ldots, M-1$,

(3.12)
$$\begin{cases} X_0 = x, \\ X_{k+1} = X_k + bh + \sigma\Delta_k W + \dfrac{1}{2}\sigma\sigma'\{\Delta_k^2 W - h\} + b'\sigma\Delta_k\Psi + \dfrac{1}{2}\Big(bb' + \dfrac{1}{2}\sigma^2 b''\Big)h^2 \\ \qquad +\Big(b\sigma' + \dfrac{1}{2}\sigma^2\sigma''\Big)\{h\Delta_k W - \Delta_k W\} + \dfrac{1}{2}\sigma\Big(\sigma\sigma'' + (\sigma')^2\Big)\left\{\dfrac{1}{3}\Delta_k^2 W - h\right\}\Delta_k W, \end{cases}$$

where $\Delta_k^2 W \stackrel{\triangle}{=} (W(t_{k+1}) - W(t_k))^2$, $\Delta_k\Psi \stackrel{\triangle}{=} \int_{t_k}^{t_{k+1}}[W(t) - W(t_k)]dt$, $b = b(t_k, X_k, u_N^k(X_k))$, $\sigma = \sigma(t_k, X_k, u_N^k(X_k))$, and, for $\varphi = b$, σ,

$$\varphi' = \varphi_x + \varphi_u\partial_x u_N^k, \quad \varphi'' \stackrel{\triangle}{=} \varphi_{xx} + 2\varphi_{xu}\partial_x u_N^k + \varphi_{uu}(\partial_x u_N^k)^2 + \varphi_u\partial_x^2 u_N^k. \tag{3.13}$$

Here the partial derivatives of b and σ should be evaluated at $(t_k, X_k, u_N^k(X_k))$, and the approximate functions u_N^k, $\frac{du_N^k}{dx}$, and $\frac{d^2u_N^k}{dx^2}$ should be evaluated at point (X_k).

We should note that $\Delta_k\Psi = \int_{t_k}^{t_{k+1}} W(t) - W(t_k)dt \sim N(0, \frac{1}{3}h^3)$, and the covariance of $\Delta_k W$ and $\Delta_k\Psi$ is $\mathbb{E}\Big(\Delta_k W\Delta_k\Psi\Big) = \frac{1}{2}h^2$. The presence of the $\Delta_k\Psi$ obviously complicates the analysis of the scheme. This will be even more so when the order of expansion increases. Treating these terms effectively will become more important.

C. *The stochastic Adam–Bashforth (SAB) scheme* (cf. [16]).

It is easily seen from (3.13) that the Platen–Wagner $\frac{3}{2}$-order scheme requires that b and σ are both twice differentiable. Recently, Ewald and Temam [16] constructed a stochastic version of the Adam–Bashforth scheme which does not involve the second

derivatives of coefficient function b but still achieves the $\frac{3}{2}$-order convergence rate. This SAB scheme, adapted for a one-dimensional diffusion of the form

$$dX_t = b(t, X_t)dt + \sigma(t, X_t)dW(t),$$

can be rewritten as follows:

(3.14)
$$X_{k+2} = X_{k+1} + \frac{\Delta t}{2}[3b(t_{k+1}, X_{k+1}) - b(t_k, X_k)] - \frac{3}{2}\Delta t A_k(t_k, X_k) + B_k(t_k, X_k),$$

where

(3.15) $A_k(t, x) = [\sigma b_x](t, x)\Delta_k W,$

$$B_k(t, x) = \sigma(t, x)\Delta_k W + \left[\sigma_t + b\sigma_x + \frac{1}{2}\sigma^2\sigma_{xx}\right](t, x)I_{(0,1)} + [\sigma b_x](t, x)I_{(1,0)},$$

$$+ [\sigma\sigma_x](t, x)I_{(1,1)} + [\sigma((\sigma_x)^2 + \sigma\sigma_{xx})](t, x)I_{(1,1,1)},$$

with $\Delta_k W = W_{t_{k+1}} - W_{t_k}$ and $h = t_{k+1} - t_k$ as before, and the random coefficients I's are defined by

$$I_{(0,1)} = 2h\Delta_{k+1}W - \int_{t_{k+1}}^{t_{k+2}} [W(s) - W(t_{k+1})]ds,$$

(3.16)
$$I_{(1,0)} = h\Delta_k W + \int_{t_{k+1}}^{t_{k+2}} [W(s) - W(t_{k+1})]ds,$$

and $I(1,1)$, $I(1,1,1)$ are the iterated Itô integrals

(3.17)
$$I_{(1,1)} = \int_{t_k}^{t_{k+2}} \int_{t_k}^{t} dW(s)dW(t) - \int_{t_k}^{t_{k+1}} \int_{t_k}^{t} dW(s)dW(t),$$

$$I_{(1,1,1)} = \int_{t_k}^{t_{k+2}} \int_{t_k}^{t} \int_{t_k}^{s} dW(r)dW(s)dW(t) - \int_{t_k}^{t_{k+1}} \int_{t_k}^{t} \int_{t_k}^{s} dW(r)dW(s)dW(t).$$

To calculate the iterated Itô integrals $I_{(1,\dots,1)}$, the useful formula of Itô [20] is often employed. More precisely, recall the scaled Hermite polynomials h_n defined by

$$h_n(x) = (-1)^n e^{\frac{1}{2}x^2} \frac{d^n}{dx^n}(e^{-\frac{1}{2}x^2}), \quad n = 0, 1, 2, \dots .$$

Let $H_n^*(x, z) = z^{n/2}h_n(x/\sqrt{z})$. For any $g \in L^2([a, b])$, let $X_t = \int_a^t g(s)dW_s$. Applying Ito's formula repeatedly to $H_n^*(X_t, \langle X\rangle_t)$ from a to b and using the fact that $(\frac{\partial}{\partial_z} + 1/2\frac{\partial^2}{\partial_{x^2}})H_n^*(x, z) = 0$, we can derive

(3.18)
$$n! \int_a^b \int_a^{t_n} \cdots \int_a^{t_2} g(t_1)g(t_2)\cdots g(t_n)dW_{t_1}\cdots dW_{t_n} = \|g\|^n h_n\left(\frac{\theta}{\|g\|}\right),$$

where

$$\|g\| = \|g\|_{L^2([a,b])} \quad \text{and} \quad \theta = \int_a^b g(t)dW_t.$$

Thus the coefficients $I_{(1,1)}$ and $I_{(1,1,1)}$ in (3.17) can be calculated explicitly as

$$I_{(1,1)} = \frac{1}{2}\Big(\Delta_{k+1}^2 W + 2\Delta_{k+1}W\Delta_k W - h\Big), \tag{3.19}$$

$$I_{(1,1,1)} = \frac{1}{6}\Big((\Delta_{k+1}W + \Delta_k W)^3 - (\Delta_k W)^3 - 6h\Delta_{k+1}W - 3h\Delta_k W\Big).$$

Using the above relations, we can rewrite the $\frac{3}{2}$-order SAB scheme for the one-dimensional forward SDE (2.4) as

$$\begin{aligned}
&X_0 = x,\\
&X_{k+2} = X_{k+1} + \frac{h}{2}\Big[3b(t_{k+1}, X_{k+1}, u_N^{k+1}(X_{k+1})) - b\Big]\\
&\quad -\frac{3}{2}h\sigma b'\Delta_k W + \sigma\Delta_k W + \frac{1}{2}\sigma\sigma'\Big\{(\Delta_{k+1}W)^2 + 2(\Delta_{k+1}W)(\Delta_k W) - h\Big\}\\
&\quad +\Big(\sigma_t + b\sigma' + \frac{1}{2}\sigma^2\sigma''\Big)\{2h\Delta_{k+1}W - \Delta_{k+1}\Psi\} + \sigma b'\{h\Delta_k W + \Delta_{k+1}\Psi\}\\
&\quad +\frac{1}{6}\sigma\Big(\sigma\sigma'' + (\sigma')^2\Big)\Big\{(\Delta_{k+1}W + \Delta_k W)^3 - (\Delta_k W)^3 - 6h\Delta_{k+1}W - 3h\Delta_k W\Big\},
\end{aligned}$$

where $k = 0, 1, \ldots, M-1$. Similarly,

$$\begin{aligned}
b &= b(t_k, X_k, u_N^k(X_k)),\\
\sigma &= \sigma(t_k, X_k, u_N^k(X_k)),\\
b' &= \frac{\partial b}{\partial x} + \frac{\partial b}{\partial u}\frac{du_N^k}{dx},\\
\sigma' &= \frac{\partial \sigma}{\partial x} + \frac{\partial \sigma}{\partial u}\frac{du_N^k}{dx},\\
\sigma'' &= \frac{\partial^2\sigma}{\partial x^2} + 2\frac{\partial^2\sigma}{\partial x\partial u}\frac{du_N^k}{dx} + \frac{\partial^2\sigma}{\partial u^2}\left(\frac{du_N^k}{dx}\right)^2 + \frac{\partial\sigma}{\partial u}\frac{d^2u_N^k}{dx^2},
\end{aligned}$$

where the partial derivatives of b and σ are evaluated at $(t_k, X_k, u_N^k(X_k))$ and the approximate functions u_N^k, $\frac{du_N^k}{dx}$, and $\frac{d^2u_N^k}{dx^2}$ are evaluated at point (X_k).

Finally, the components Y and Z of the solution to (2.1) are approximated as

$$Y_k = u_N^k(X_k), \quad Z_k = \sigma(t_k, X_k, Y_k)\frac{\partial u_N^k}{\partial x}(X_k). \tag{3.20}$$

4. Error analysis for the PDE approximation. In this section we carry out an error analysis for the approximation of the PDE (3.1). In order to simplify the presentation, we shall consider only the following *Hermite–Galerkin scheme* for (3.4).

Let $u_N^M(x) = P_N\varphi$. For $k = M-1, \dots, 0$, we compute $u_N^k \in \mathcal{H}_N$ such that

$$\left\langle \frac{u_N^{k+1} - u_N^k}{h}, v_N \right\rangle + \frac{1}{2}\left\langle \partial_x\Big(\sigma^2(t_k, \cdot, u_N^{k+1})(u_N^k)_x\Big), v_N \right\rangle \tag{4.1}$$

$$+ \langle \hat{b}(t_k, \cdot, u_N^{k+1}, (u_N^{k+1})_x)\partial_x(u_N^{k+1}) + g(t_k, \cdot, u_N^{k+1}), v_N\rangle = 0 \qquad \forall v_N \in \mathcal{H}_N.$$

We note that the only difference between this Hermite–Galerkin scheme and the Hermite-collocation scheme (3.7) is that in the latter the continuous inner product is replaced by the discrete inner product. It is well known that for N sufficiently large, the differences between the two approaches are negligible (cf. [18]).

Hereafter, we shall use "$A \lesssim B$" to mean that there exists a constant $C > 0$ independent of N or h such that $A \le CB$.

Our main result in this section is the following theorem.

THEOREM 4.1. *Assume* (A1)–(A3). *Let u and u_N^k be the solutions of* (3.1) *and* (4.1), *respectively. Assume further that $u \in C([0,T]; H_D^m(\mathbb{R}))$, that $u_t \in L^2(0,T; H_D^m(\mathbb{R}))$ with some $m > 1$, and that $u_{tt} \in L^2(0,T; H^{-1}(\mathbb{R}))$. Then, there exists $h_0 > 0$ such that for $h \le h_0$ the scheme* (3.4) *is stable, and the following error estimates hold: for each $k = 0, 1, \dots, M$,*

$$\|u(t_k,\cdot) - u_N^k\| + \left(h\sum_{j=k}^{M}\|\partial_x(u(t_j,\cdot) - u_N^j)\|^2\right)^{\frac{1}{2}}$$

$$\lesssim h\Big(\|u_{tt}\|_{L^2(0,T;H^{-1}(\mathbb{R}))} + \|u_t\|_{L^2(0,T;H^1(\mathbb{R}))}\Big)$$

$$+ N^{\frac{1-m}{2}}\Big(\|u\|_{C([0,T];H_D^m)} + \|u_t\|_{L^2(0,T;H_D^m)}\Big).$$

Proof. Let E^k, $k = M-1, \dots, 1, 0$, be the consistency error defined by

$$E^k(\cdot) \triangleq \frac{1}{h}\Big(u(t_{k+1},\cdot) - u(t_k,\cdot)\Big) + \hat{b}(t_k,\cdot,u(t_{k+1},\cdot),\partial_x u(t_{k+1},\cdot))\,\partial_x u(t_{k+1},\cdot)$$

$$+ \partial_x\left(\frac{1}{2}\sigma^2(t_k,\cdot,u(t_{k+1},\cdot))\partial_x u(t_k,\cdot)\right) + g(t_k,\cdot,u(t_{k+1},\cdot)), \tag{4.2}$$

where $u(t_k,\cdot)$ is the exact solution to (3.1). First, we shall obtain the estimate for the consistency error E^k. By using the integral residue form of the Taylor series, and the fact that $u(t_k,\cdot)$ is the exact solution to (3.1), we can easily rewrite $E^k(\cdot)$ as

$$E^k(\cdot) = \frac{1}{h}\int_{t_k}^{t_{k+1}} u_{tt}(t,\cdot)(t-t_k)dt$$

$$+ g(t_k,\cdot,u(t_{k+1},\cdot)) - g(t_k,\cdot,u(t_k,\cdot))$$

$$+ \hat{b}(t_k,\cdot,u(t_{k+1},\cdot),\partial_x u(t_{k+1},\cdot))\,\partial_x u(t_{k+1},\cdot)$$

$$- \hat{b}(t_k,\cdot,u(t_{k+1},\cdot),\partial_x u(t_{k+1},\cdot))\,\partial_x u(t_k,\cdot)$$

$$+ \hat{b}(t_k,\cdot,u(t_{k+1},\cdot),\partial_x u(t_{k+1},\cdot))\,\partial_x u(t_k,\cdot) - \hat{b}(t_k,\cdot,u(t_k,\cdot),\partial_x u(t_k,\cdot))\,\partial_x u(t_k,\cdot)$$

$$+\partial_x\left(\frac{1}{2}(\sigma^2(t_k,\cdot,u(t_{k+1},\cdot))-\sigma^2(t_k,\cdot,u(t_k,\cdot)))\partial_x u(t_k,\cdot)\right).$$

Under assumptions (A1)–(A3) (see also Remark 2.2), it is known that the solution to (3.1) must have bounded first-order derivative $\partial_x u$; i.e., $\|\partial_x u\|_{L^\infty((0,T)\times\mathbb{R})}\le C$ (see, e.g., [10]). We can then derive from assumptions (A1)–(A3) (see also Remark 2.2) that

$$\begin{aligned}\|E^k(\cdot)\|_{-1}\lesssim{}&\frac{1}{h}\int_{t_k}^{t_{k+1}}\|u_{tt}(t,\cdot)\|_{-1}(t-t_k)dt+h^{\frac{1}{2}}\left(\int_{t_k}^{t_{k+1}}\|\partial_t u(t,\cdot)\|^2dt\right)^{\frac{1}{2}}\\&+h^{\frac{1}{2}}\left(\int_{t_k}^{t_{k+1}}\|\partial_{tx}u(t,\cdot)\|^2dt\right)^{\frac{1}{2}}\\&+\|(\sigma^2(t_k,\cdot,u(t_{k+1},\cdot))-\sigma^2(t_k,\cdot,u(t_k,\cdot)))\partial_x u(t_k,\cdot)\|.\end{aligned}$$

Applying the Schwarz inequality and using the boundedness of $\partial_x u$, we find

$$\|E^k(\cdot)\|_{-1}^2\lesssim h\int_{t_k}^{t_{k+1}}\|u_{tt}(t,\cdot)\|_{-1}^2dt+h\int_{t_k}^{t_{k+1}}\|u_t(t,\cdot)\|_1^2dt,$$

which implies that

$$h\sum_{k=0}^{M}\|E^k(\cdot)\|_{-1}^2\lesssim h^2\|u_{tt}\|^2_{L^2(0,T;H^{-1}(\mathbb{R}))}+h^2\|u_t\|^2_{L^2(0,T;H^1(\mathbb{R}))}.\tag{4.3}$$

In what follows we set $\hat{e}_N^k=P_Nu(t_k,\cdot)-u_N^k$, $\tilde{e}_N^k=u(t_k,\cdot)-P_Nu(t_k,\cdot)$, and thus we have $e_N^k=u(t_k,\cdot)-u_N^k=\hat{e}_N^k+\tilde{e}_N^k$.

Next, multiplying by $v_N\in\mathcal{H}_N$ on both sides of (4.2), using integration by parts, and then subtracting (4.1), we obtain that

$$\begin{aligned}\langle E^k,v_N\rangle=\frac{1}{h}\langle e_N^{k+1}-e_N^k\,,v_N\rangle&-\left\langle\frac{1}{2}\sigma^2(t_k,\cdot,u(t_{k+1},\cdot))\partial_x u(t_k,\cdot)\,,\partial_x v_N\right\rangle\\&+\left\langle\frac{1}{2}\sigma^2(t_k,\cdot,u_N^{k+1})\partial_x u_N^k\,,\partial_x v_N\right\rangle\\&+\langle\hat{b}(t_k,\cdot,u(t_{k+1},\cdot),\partial_x u(t_{k+1},\cdot)))\,\partial_x u(t_{k+1},\cdot)\\&\quad-\hat{b}(t_k,\cdot,u_N^{k+1},\partial_x u_N^{k+1})\partial_x u_N^{k+1}\,,v_N\rangle\\&+\langle g(t_k,\cdot,u(t_{k+1},\cdot))-g(t_k,\cdot,u_N^{k+1})\,,v_N\rangle.\end{aligned}$$

Or, equivalently,

$$\begin{aligned}\langle E^k,v_N\rangle=\frac{1}{h}\langle e_N^{k+1}-e_N^k\,,v_N\rangle&-\left\langle\frac{1}{2}\sigma^2(t_k,\cdot,u_N^{k+1})\partial_x e_N^k\,,\partial_x v_N\right\rangle\\&-\left\langle\partial_x u(t_k,\cdot)\,,\frac{1}{2}\sigma^2(t_k,\cdot,u(t_{k+1},\cdot))\partial_x v_N-\frac{1}{2}\sigma^2(t_k,\cdot,u_N^{k+1})\partial_x v_N\right\rangle\\&+\langle\hat{b}(t_k,\cdot,u(t_{k+1},\cdot),\partial_x u(t_{k+1},\cdot))\,\partial_x u(t_{k+1},\cdot)\\&\quad-\hat{b}(t_k,\cdot,u_N^{k+1},\partial_x u_N^{k+1})\partial_x u_N^{k+1}\,,v_N\rangle\\&+\langle g(t_k,\cdot,u(t_{k+1},\cdot))-g(t_k,\cdot,u_N^{k+1})\,,v_N\rangle.\end{aligned}$$

Since $e_N^k = \tilde{e}_N^k + \hat{e}_N^k$, we have

(4.4)
$$\begin{aligned}
&\frac{1}{h}\langle \hat{e}_N^{k+1} - \hat{e}_N^k, v_N\rangle - \left\langle \frac{1}{2}\sigma^2(t_k,\cdot,u_N^{k+1})\partial_x \hat{e}_N^k \,,\partial_x v_N \right\rangle \\
&\qquad = -\frac{1}{h}\langle \tilde{e}_N^{k+1} - \tilde{e}_N^k \,, v_N\rangle + \left\langle \frac{1}{2}\sigma^2(t_k,\cdot,u_N^{k+1})\partial_x \tilde{e}_N^k \,,\partial_x v_N \right\rangle \\
&\qquad + \left\langle \partial_x u(t_k,\cdot)\,, \frac{1}{2}\sigma^2(t_k,\cdot,u(t_{k+1},\cdot))\partial_x v_N - \frac{1}{2}\sigma^2(t_k,\cdot,u_N^{k+1})\partial_x v_N \right\rangle \\
&\qquad\quad - \langle \hat{b}(t_k,\cdot,u(t_{k+1},\cdot),\partial_x u(t_{k+1},\cdot))\ \partial_x u(t_{k+1},\cdot) \\
&\qquad - \hat{b}(t_k,\cdot,u_N^{k+1},\partial_x u_N^{k+1})\partial_x u_N^{k+1}\ , v_N\rangle \\
&\qquad - \langle g(t_k,\cdot,u(t_{k+1},\cdot)) - g(t_k,\cdot,u_N^{k+1})\ , v_N\rangle + \langle E^k(\cdot)\ , v_N\rangle.
\end{aligned}$$

Denote

$$I_1 = -\langle g(t_k,\cdot,u(t_{k+1},\cdot)) - g(t_k,\cdot,u_N^{k+1})\ , v_N\rangle,$$

$$I_2 = -\langle \hat{b}(t_k,\cdot,u(t_{k+1},\cdot),\partial_x u(t_{k+1},\cdot))\ \partial_x u(t_{k+1},\cdot) - \hat{b}(t_k,\cdot,u_N^{k+1},\partial_x u_N^{k+1})\partial_x u_N^{k+1}\ , v_N\rangle,$$

$$I_3 = \left\langle \partial_x u(t_k,\cdot)\,, \frac{1}{2}\sigma^2(t_k,\cdot,u(t_{k+1},\cdot))\partial_x v_N - \frac{1}{2}\sigma^2(t_k,\cdot,u_N^{k+1})\partial_x v_N \right\rangle,$$

$$I_4 = \left\langle \frac{1}{2}\sigma^2(t_k,\cdot,u_N^{k+1})\partial_x \tilde{e}_N^k \,,\partial_x v_N \right\rangle,$$

$$I_5 = -\frac{1}{h}\langle \tilde{e}_N^{k+1} - \tilde{e}_N^k \,, v_N\rangle,$$

$$I_6 = \langle E^k(\cdot)\ , v_N\rangle.$$

We shall estimate these terms separately under assumptions (A1)–(A3) (see also Remark 2.2). To begin with, noting the Lipschitz property of g and applying the Cauchy–Schwarz inequality, we have, for some constant $C_1 > 0$ and any $\varepsilon > 0$,

$$\begin{aligned}
(4.5)\quad |I_1| &\le C_1\langle |u(t_{k+1},\cdot) - u_N^{k+1}|\ , |v_N|\rangle = C_1\langle |e_N^{k+1}|\ , |v_N|\rangle \\
&\le \frac{C_1 h}{2\varepsilon}\|e_N^{k+1}\|^2 + \frac{C_1\varepsilon}{2h}\|v_N\|^2 \le \frac{C_1 h}{\varepsilon}\|\hat{e}_N^{k+1}\|^2 + \frac{C_1 h}{\varepsilon}\|\tilde{e}_N^{k+1}\|^2 + \frac{C_1\varepsilon}{2h}\|v_N\|^2.
\end{aligned}$$

Similarly, we have

$$(4.6)\qquad |I_4| \le C_4\langle |\partial_x \tilde{e}_N^k|\ , |\partial_x v_N|\rangle \le \frac{C_4 h}{2\varepsilon}\|\partial_x \tilde{e}_N^k\|^2 + \frac{C_4\varepsilon}{2h}\|\partial_x v_N\|^2;$$

$$
\begin{aligned}
|I_2| &\le |\langle \hat{b}(t_k,\cdot,u(t_{k+1},\cdot),\partial_x u(t_{k+1},\cdot))\ \partial_x u(t_{k+1},\cdot) \\
&\qquad - \hat{b}(t_k,\cdot,u_N^{k+1},\partial_x u_N^{k+1})\partial_x u(t_{k+1},\cdot)\ , v_N\rangle| \\
&\quad + |\langle \hat{b}(t_k,\cdot,u_N^{k+1},\partial_x u_N^{k+1}))(\partial_x u(t_{k+1},\cdot) - \partial_x u_N^{k+1})\ , v_N\rangle| \\
&\le |\langle \|\partial_x u\|_{L^\infty((0,T)\times\mathbb{R})} K(|e_N^{k+1}| + |\partial_x e_N^{k+1}|),\ |v_N|\rangle| + K|\langle |\partial_x e_N^{k+1}|, |v_N|\rangle| \\
&\le C_2\langle |e_N^{k+1}| + |\partial_x e_N^{k+1}|, |v_N|\rangle + C_2\langle |\partial_x e_N^{k+1}|, |v_N|\rangle \qquad (4.7)\\
&\le C_2\langle |\hat{e}_N^{k+1}| + |\partial_x \hat{e}_N^{k+1}| + |\tilde{e}_N^{k+1}| + |\partial_x \tilde{e}_N^{k+1}|\ , |v_N|\rangle \\
&\quad + C_2\langle |\partial_x \hat{e}_N^{k+1}| + |\partial_x \tilde{e}_N^{k+1}|\ , |v_N|\rangle \\
&\le \frac{C_2 h}{\varepsilon}(\|\hat{e}_N^{k+1}\|^2 + \|\tilde{e}_N^{k+1}\|^2 + \|\partial_x \tilde{e}_N^{k+1}\|^2) \\
&\quad + \frac{C_2\varepsilon}{h}\|v_N\|^2 + C_2\varepsilon_1 h\|\partial_x \hat{e}_N^{k+1}\|^2 + \frac{C_2}{\varepsilon_1 h}\|v_N\|^2,
\end{aligned}
$$

where C_2 depends only on $K, \|\partial_x u\|_{L^\infty((0,T)\times\mathbb{R})}$ and ε_1 is to be determined later;

$$
\begin{aligned}
|I_3| &\le CK\ \|\partial_x u\|_{L^\infty((0,T)\times\mathbb{R})}\ \langle |e_N^{k+1}|\ , |\partial_x v_N|\rangle \le \frac{C_3 h}{2\varepsilon}\|e_N^{k+1}\|^2 + \frac{C_3\varepsilon}{2h}\|\partial_x v_N\|^2 \qquad (4.8)\\
&\le \frac{C_3 h}{\varepsilon}(\|\hat{e}_N^{k+1}\|^2 + \|\tilde{e}_N^{k+1}\|^2) + \frac{C_3\varepsilon}{2h}\|\partial_x v_N\|^2,
\end{aligned}
$$

where $C_3 = CK\ \|\partial_x u\|_{L^\infty((0,T)\times\mathbb{R})}$. Finally, let $\tilde{E}_N^k = -\frac{1}{h}(\tilde{e}_N^{k+1} - \tilde{e}_N^k)$; then we have

$$
|I_5| \le \frac{h}{2\varepsilon}\|\tilde{E}_N^k\|^2 + \frac{\varepsilon}{2h}\|v_N\|^2 \quad \text{and} \quad |I_6| \le \frac{h}{2\varepsilon}\|E^k(\cdot)\|_{-1}^2 + \frac{\varepsilon}{2h}\|v_N\|_1^2. \qquad (4.9)
$$

To obtain the desired estimate, let us now look at the left-hand side of (4.4) with $v_N = -h\ \hat{e}_N^k$. Note that

$$
\begin{aligned}
&\frac{1}{h}\langle \hat{e}_N^{k+1} - \hat{e}_N^k\ , -h\ \hat{e}_N^k\rangle = \frac{1}{2}(\|\hat{e}_N^k\|^2 - \|\hat{e}_N^{k+1}\|^2 + \|\hat{e}_N^k - \hat{e}_N^{k+1}\|^2), \\
&-\left\langle \partial_x \hat{e}_N^k\ , \frac{1}{2}\sigma^2(t_k,\cdot,u_N^{k+1})\partial_x v_N \right\rangle \ge ch\|\partial_x \hat{e}_N^k\|^2, \qquad (4.10)
\end{aligned}
$$

where $c > 0$ is a lower bound of σ^2, with a slight abuse of notation (compare to the constant $c > 0$ in (A3)). Thanks to the Schwarz inequality, and denoting $C = 5\max_{1\le i\le 4}\{C_i\} + 1$, we derive from (4.5)–(4.10) that

$$
\begin{aligned}
\|\hat{e}_N^k\|^2 + ch\|\partial_x \hat{e}_N^k\|^2 &\le \left(\frac{Ch}{\varepsilon} + \frac{1}{2}\right)\|\hat{e}_N^{k+1}\|^2 + C\varepsilon_1 h\|\partial_x \hat{e}_N^{k+1}\|^2 \qquad (4.11)\\
&\quad + \left(\frac{Ch}{\varepsilon_1} + C\varepsilon h + \varepsilon h + \frac{1}{2}\right)\|\hat{e}_N^k\|^2 \\
&\quad + C\varepsilon h\|\partial_x \hat{e}_N^k\|^2 + \frac{Ch}{\varepsilon}(\|\tilde{e}_N^{k+1}\|^2 + \|\partial_x \tilde{e}_N^{k+1}\|^2) \\
&\quad + \frac{Ch}{\varepsilon}\|\partial_x \tilde{e}_N^k\|^2 + \frac{h}{\varepsilon}\|E^k(\cdot)\|_{-1}^2 + \frac{h}{\varepsilon}\|\tilde{E}_N^k\|^2.
\end{aligned}
$$

Therefore, (4.11) can now be rewritten as

$$\left(\frac{1}{2} - h\left(C\varepsilon + \varepsilon + \frac{C}{\varepsilon_1}\right)\right)\|\hat{e}_N^k\|^2 + (c - C\varepsilon)h\|\partial_x \hat{e}_N^k\|^2$$

$$\leq \left(\frac{Ch}{\varepsilon} + \frac{1}{2}\right)\|\hat{e}_N^{k+1}\|^2 + C\varepsilon_1 h\|\partial_x \hat{e}_N^{k+1}\|^2 \tag{4.12}$$

$$+ \frac{h}{\varepsilon}\Big(C\|\tilde{e}_N^{k+1}\|^2 + C\|\partial_x \tilde{e}_N^k\|^2 + C\|\partial_x \tilde{e}_N^{k+1}\|^2 + \|\tilde{E}_N^k\|^2 + \|E^k(\cdot)\|_{-1}^2\Big).$$

We choose ε small enough so that $c - C\varepsilon > 0$. Multiplying both sides of (4.12) by 2, choosing $\varepsilon_1 = \frac{c - C\varepsilon}{2C}$, setting $C_5 \triangleq C + 1$, $C_\varepsilon \triangleq 2(C\varepsilon + \varepsilon + \frac{C}{\varepsilon_1})$, and assuming that the time step h satisfies the conditions

$$h < h_0 = \min\left(\frac{1}{2C_\varepsilon}, \frac{\varepsilon}{2C}\right), \tag{4.13}$$

inequality (4.12) becomes

(4.14)

$$(1 - hC_\varepsilon)\|\hat{e}_N^k\|^2 + 2(c - C\varepsilon)h\|\partial_x \hat{e}_N^k\|^2 \leq \left(\frac{2Ch}{\varepsilon} + 1\right)\|\hat{e}_N^{k+1}\|^2 + (c - C\varepsilon)h\|\partial_x \hat{e}_N^{k+1}\|^2$$

$$+ \frac{hC_5}{\varepsilon}\Big(\|\tilde{e}_N^{k+1}\|^2 + \|\partial_x \tilde{e}_N^k\|^2 + \|\partial_x \tilde{e}_N^{k+1}\|^2 + \|\tilde{E}_N^k\|^2 + \|E^k(\cdot)\|_{-1}^2\Big).$$

We now apply a standard discrete backward Gronwall inequality and, noticing that $\hat{e}_N^M = 0$, we obtain

$$\|\hat{e}_N^k\|^2 + h\sum_{j=k}^{M}\|\partial_x \hat{e}_N^j\|^2 \lesssim h\sum_{j=k}^{M}\Big(\|\tilde{e}_N^j\|^2 + \|\partial_x \tilde{e}_N^j\|^2 + \|\tilde{E}_N^j\|^2 + \|E^j(\cdot)\|_{-1}^2\Big). \tag{4.15}$$

We now estimate the terms on the right-hand side. First, by virtue of Lemma 2.3 with $s = 0, 1$, respectively, and $r = m$, we have

$$\|\tilde{e}_N^{k+1}\|^2 + \|\tilde{e}_N^k\|^2 \leq C_6 N^{-m}(\|D^m u(t_{k+1}, .)\|^2 + \|D^m u(t_k, .)\|^2),$$

$$\|\partial_x \tilde{e}_N^{k+1}\|^2 \leq C_6 N^{1-m}\|D^m u(t_{k+1}, .)\|^2. \tag{4.16}$$

It remains to estimate $\|\tilde{E}_N^k\|$. Let $\tilde{e}_N(t, \cdot) = u(t, \cdot) - P_N u(t, \cdot)$. Then,

$$\tilde{E}_N^k = -\frac{1}{h}(\tilde{e}_N^{k+1} - \tilde{e}_N^k) = -\frac{1}{h}\int_{t_k}^{t_{k+1}} \partial_t \tilde{e}_N(t, \cdot)dt.$$

Applying the Schwarz inequality and (2.17) again, we obtain that

$$\|\tilde{E}_N^k\|^2 \leq \frac{1}{h}\int_{t_k}^{t_{k+1}} \|\partial_t \tilde{e}_N(t, \cdot)\|^2 \, dt. \tag{4.17}$$

We then derive from the above and (4.15) that

$$\|\hat{e}_N^k\|^2 + h\sum_{j=k}^{M}\|\partial_x \hat{e}_N^j\|^2 \lesssim h^2\Big(\|u_{tt}\|_{L^2(0,T;H^{-1}(\mathbb{R}))}^2 + \|u_t\|_{L^2(0,T;H^1(\mathbb{R}))}^2\Big)$$

$$+ N^{1-m}\Big(\|u\|_{C([0,T];H_D^m)}^2 + \|u_t\|_{L^2(0,T;H_D^m)}^2\Big),$$

proving the theorem. □

Remark 4.2. We would like to point out that one of the main purposes of Theorem 4.1 is to display a significant feature of the spectral method, that is, that the rate of convergence of the scheme increases as the smoothness of the solutions (i.e., m in the assumption of the above theorem) increases. Assumptions (A1)–(A3) are used only to guarantee the convergence of the numerical schemes proposed in this paper (in fact, all but the Platen–Wagner scheme where the second derivative of the drift b is required). The assumption on smoothness of the solution is used to prove the rate of convergence, which should always be true if the coefficients are sufficiently regular. We should note, however, that the explicit relation between the regularities of the coefficients and that of the solution seems to be a more subtle issue due to the quasi linearity of the PDE. We thus prefer making the exact smoothness requirement needed on the solution rather than on the coefficients.

We note also that by using a tensor product approach, the results in the above theorem can be directly extended to the high-dimensional case, although such a tensor product approach will become prohibitively expensive for space dimension ≥ 4.

5. The synthetic error analysis. In this section we present an error analysis for the synthesized numerical scheme. Our main result is the following.

THEOREM 5.1. *Assume that* (A1)–(A3) *and the assumptions in Theorem* 4.1 *hold. Let* (X, Y, Z) *be the adapted solution to the FBSDE* (2.1)*, and let* $\{X_k\}_{k=0}^M$ *be the solution to the Euler scheme* (3.9)*. Define*

$$Y_k \triangleq u_N^k(X_k), \qquad Z_k \triangleq \sigma(t_k, X_k, Y_k)\partial_x(u_N^k)(X_k), \qquad k = 0, 1, \ldots, M, \tag{5.1}$$

where $\{u_N^k\}_{k=0}^M$ *is the solution to* (4.1)*. Then, the following error estimate holds:*

$$\begin{aligned}\sup_{0\leq k\leq M} \mathbb{E}\Big[(X(t_k) - X_k)^2 + (Y(t_k) - Y_k)^2 + (Z(t_k) - Z_k)^2\Big]^{1/2} \\ \lesssim \sqrt{h} + h\Big(\|u_{tt}\|_{L^2(0,T;H^{-1}(\mathbb{R}))} + \|u_t\|_{L^2(0,T;L^2(\mathbb{R}))}\Big) \\ + N^{\frac{1-m}{2}}\Big(\|u\|_{C([0,T];H_D^m)} + \|u_t\|_{L^2(0,T;H_D^m)}\Big).\end{aligned} \tag{5.2}$$

Proof. Recall the intermediate Euler scheme (3.10):

$$\begin{cases} \hat{X}_0 = x, \\ \hat{X}_{k+1} = \hat{X}_k + b(t_k, \hat{X}_k, u^k(\hat{X}_k))h + \sigma(t_k, \hat{X}_k, u^k(\hat{X}_k))\Delta_k W, \end{cases} \tag{5.3}$$

where $u^k(x) = u(t_k, x)$ and $u(\cdot, \cdot)$ is the solution to the original PDE (3.1). Since the true solution u is at least uniformly Lipschitz under (A1)–(A3), we can apply the *fundamental convergence theorem* for SDEs (cf. [21] or [30]) to conclude that

$$\mathbb{E}|X(t_k) - \hat{X}_k|^2 \leq Ch \quad \forall\, k = 0, 1, \ldots, M. \tag{5.4}$$

Thus it remains to evaluate $\mathbb{E}|\hat{X}_k - X_k|^2$. To this end, note that

$$\begin{aligned}\hat{X}_{k+1} - X_{k+1} = \hat{X}_k - X_k + [b(t_k, \hat{X}_k, u(t_k, \hat{X}_k)) - b(t_k, X_k, u_N^k(X_k))]h \\ + [\sigma(t_k, \hat{X}_k, u(t_k, \hat{X}_k)) - \sigma(t_k, X_k, u_N^k(X_k))]\Delta_k W.\end{aligned} \tag{5.5}$$

Since $\Delta_k W = W(t_{k+1}) - W(t_k)$ is independent of $\mathcal{F}_{t_k}$ with $\mathbb{E}[\Delta_k W] = 0$ and $\mathbb{E}|\Delta_k W|^2 = h$, one has $\mathbb{E}[(\Delta_k W)F] = 0$ and $\mathbb{E}[(\Delta_k W)F]^2 = h\mathbb{E}[F]^2$ for all $F \in L^2(\mathcal{F}_{t_k};\mathbb{R})$. Therefore, squaring both sides of (5.5), applying the Cauchy–Schwartz inequality, and then using the Lipschitz assumption on the coefficients b and σ, we obtain that

$$\begin{aligned}
&\mathbb{E}|\hat{X}_{k+1} - X_{k+1}|^2 \\
&\quad \le (1+h)\mathbb{E}|\hat{X}_k - X_k|^2 + (h^2+h)\mathbb{E}|b(t_k,\hat{X}_k,u^k(\hat{X}_k)) - b(t_k,X_k,u_N^k(X_k))|^2 \\
&\quad\quad + h\,\mathbb{E}|\sigma(t_k,\hat{X}_k,u^k(\hat{X}_k)) - \sigma(t_k,X_k,u_N^k(X_k))|^2 \qquad (5.6)\\
&\quad \le (1+h+2K^2(h^2+h))\mathbb{E}|\hat{X}_k - X_k|^2 + 2K^2(h^2+h)\mathbb{E}|u^k(\hat{X}_k) - u_N^k(X_k)|^2.
\end{aligned}$$

Now, we write

$$u^k(\hat{X}_k) - u_N^k(X_k) = [u^k(\hat{X}_k) - u^k(X_k)] + [u^k(X_k) - u_N^k(X_k)] \qquad (5.7)$$

and note that

$$\mathbb{E}|u^k(\hat{X}_k) - u^k(X_k)|^2 \le K^2\mathbb{E}|\hat{X}_k - X_k|^2. \qquad (5.8)$$

We need only to estimate $\mathbb{E}|u^k(X_k) - u_N^k(X_k)|^2$. To this end, let us first establish the following simple inequality:

$$\|f\|_{L^\infty(\mathbb{R})} \le \|f\|_{L^2(\mathbb{R})} + \|f'\|_{L^2(\mathbb{R})} \qquad \forall\, f \in\ H^1(\mathbb{R}). \qquad (5.9)$$

Indeed, we first observe that $f \in H^1(\mathbb{R})$ implies that $f \in C(\mathbb{R})$. Then, for any unit interval $[n, n+1)$, $n \in \mathbb{Z}$, there must exist $x_n \in [n, n+1)$ such that $|f(x_n)| \le \|f\|_{L^2(n,n+1)} \le \|f\|_{L^2(\mathbb{R})}$. Hence, for any given $x \in [n, n+1)$, we have

$$|f(x)| \le |f(x) - f(x_n)| + |f(x_n)| = \left|\int_{x_n}^x f'(t)dt\right| + |f(x_n)| \le \|f\|_{L^2(\mathbb{R})} + \|f'\|_{L^2(\mathbb{R})}.$$

Thanks to (5.9), we obtain that

$$\begin{aligned}
\mathbb{E}|u^k(X_k) - u_N^k(X_k)|^2 &= \int_{-\infty}^{\infty} |u(t_k,x) - u_N^k(x)|^2 f_{X_k}(x)dx \\
&\le \|u(t_k,\cdot) - u_N^k(\cdot)\|^2_{L^\infty(\mathbb{R})} \int_{-\infty}^{\infty} f_{X_k}(x)dx \qquad (5.10)\\
&\le 2\|u(t_k,\cdot) - u_N^k\|^2 + 2\|\partial_x(u(t_k,\cdot) - u_N^k)\|^2,
\end{aligned}$$

where f_{X_k} is the probability density function of random variable X_k.

Combining (5.8)–(5.10), noting (5.7), and assuming without loss of generality that $h < 1$, it follows easily from (5.6) that, for $k = 1, 2, \ldots, M$,

$$\begin{aligned}
\mathbb{E}|\hat{X}_{k+1} - X_{k+1}|^2 &\le (1+Ch)\mathbb{E}|\hat{X}_k - X_k|^2 \\
&\quad + Ch\Big(\|u(t_k,\cdot) - u_N^k\|^2 + \|\partial_x(u(t_k,\cdot) - u_N^k)\|^2\Big).
\end{aligned}$$

We now apply a standard forward discrete Gronwall inequality and, noticing that $\hat{X}_0 = X_0 = x$ and using Theorem 4.1, we obtain, for $k = 1, 2, \ldots, M$,

$$\mathbb{E}|\hat{X}_k - X_k|^2 \lesssim \mathbb{E}|\hat{X}_0 - X_0|^2 + h\sum_{j=1}^{k}\Big(\|u(t_j,\cdot) - u_N^j\|^2 + \|\partial_x(u(t_j,\cdot) - u_N^j)\|^2\Big)$$

$$\lesssim h^2\Big(\|u_{tt}\|^2_{L^2(0,T;H^{-1}(\mathbb{R}))} + \|u_t\|^2_{L^2(0,T;L^2(\mathbb{R}))}\Big)$$

$$+ N^{1-m}\Big(\|u\|^2_{C([0,T];H_D^m)} + \|u_t\|^2_{L^2(0,T;H_D^m)}\Big).$$

This, together with (5.4), yields the final mean-square error estimate:

$$\mathbb{E}|X(t_k) - X_k|^2 \lesssim h + h^2\Big(\|u_{tt}\|^2_{L^2(0,T;H^{-1}(\mathbb{R}))} + \|u_t\|^2_{L^2(0,T;L^2(\mathbb{R}))}\Big)$$

$$+ N^{1-m}\Big(\|u\|^2_{C([0,T];H_D^m)} + \|u_t\|^2_{L^2(0,T;H_D^m)}\Big).$$

The error estimates for $\mathbb{E}|Y(t_k) - Y_k|^2$ and $\mathbb{E}|Z(t_k) - Z_k|^2$ then follow easily from the relation (3.20), (5.1), and the estimate (5.11). The proof is now complete. ∎

Remark 5.2. From the proof of Theorem 5.1 we see that if we combine the higher-order schemes for both PDE (2.5) and the resulting decoupled SDE (2.4), then we will obtain a higher-order rate of convergence. The proof would be completely similar, but conceivably much more lengthy and tedious. We shall discuss this issue with extensive numerical examples in the next section.

Unlike the results in Theorem 4.1, the results in the above theorem cannot be directly extended to the high-dimensional case due (only) to the fact that the embedding (5.9) is no longer valid in a high space dimension.

6. Numerical simulations and conclusions.

6.1. Implementation details. Let $\{h_j(x)\}_{j=0,1,\ldots,N}$ be the Lagrange interpolation polynomials in P_N associated with the Hermite–Gauss points $\{x_j\}_{j=0,1,\ldots,N}$, which are zeros of $H_{N+1}(x)$. We define $\hat{h}_j(x) = h_j(x)\frac{e^{-x^2/2}}{e^{-x_j^2/2}}$; then

$$\hat{h}_j(x_i) = \delta_{ij}, \quad \mathcal{H}_N = \text{span}\{\hat{h}_j : j = 0, 1, \ldots, N\},$$

so any function $g \in \mathcal{H}_N$ can be written as

$$g(x) = \sum_{j=0}^{N} g(x_j)\hat{h}_j(x).$$

We derive easily from $h_j^{'}(x_i)$ that

$$\hat{h}_j^{'}(x_i) = \frac{e^{-x_i^2/2}}{e^{-x_j^2/2}}(h_j^{'}(x_i) - x_i\delta_{ij}) = \begin{cases} \frac{\widehat{H}_N(x_i)}{(x_i - x_j)\widehat{H}_N(x_j)}, & i \neq j, \\ 0, & i = j. \end{cases}$$

Hence, $\hat{h}_j^{'}(x_i)$ can be computed in a stable way for N large.

Let us denote

$$\bar{u} = (u_N^k(x_0), u_N^k(x_1), \dots, u_N^k(x_N))^T, \quad \bar{f} = (f^{k+1}(x_0), f^{k+1}(x_1), \dots, f^{k+1}(x_N))^T$$

$$s_{ij} = (\alpha(x)\hat{h}_j^{'}, \hat{h}_i^{'})_N = \sum_{l=0}^{N} \alpha(x_l)\hat{h}_j^{'}(x_l)\hat{h}_i^{'}(x_l)\hat{w}_l$$

with $\alpha(x_i) = \frac{h}{2}\sigma^2(t_k, x_i, u_N^{k+1}(x_i))$. Then, (3.7) is reduced to the following linear system:

$$(S + W)\bar{u} = W\bar{f},$$

where

$$W = \mathrm{diag}(\hat{w}_0, \hat{w}_1, \dots, \hat{w}_N), \quad S = (s_{ij})_{i,j=0,1,\dots,N}.$$

6.2. Numerical results and discussions. In order to make sensible comparisons to the existing results, we shall consider an example proposed by Milstein and Tretyakov [31], use our spectral method to carry out the numerical approximation for $u(t, x)$, and then use different stochastic numerical schemes for the resulting decoupled SDE. In particular, we shall use the standard Euler scheme, first-order Milstein scheme, $\frac{3}{2}$-order Platen–Wagner strong scheme, as well as the $\frac{3}{2}$-order stochastic SAB scheme, respectively.

Example (Milstein–Tretyakov [31]). Consider the following FBSDE:

(6.1)
$$\begin{cases} dX = \dfrac{X(1+X^2)}{(2+X^2)^3}dt + \dfrac{1+X^2}{2+X^2}\sqrt{\dfrac{1+2Y^2}{1+Y^2+\exp(-\frac{2X^2}{t+1})}}dW(t), \quad X(0) = x, \\ \\ dY = -g(t,X,Y)dt - f(t,X,Y)Zdt + ZdW(t), \quad Y(T) = \exp\left(-\dfrac{X^2(T)}{T+1}\right), \end{cases}$$

where

$$g(t,x,u) = \frac{1}{t+1}\exp\left(-\frac{x^2}{t+1}\right)\left[\frac{4x^2(1+x^2)}{(2+x^2)^3} + \left(\frac{1+x^2}{2+x^2}\right)^2\left(1 - \frac{2x^2}{t+1}\right) - \frac{x^2}{t+1}\right],$$

$$f(t,x,u) = \frac{x}{(2+x^2)^2}\sqrt{\frac{1+u^2+\exp(-\frac{2x^2}{t+1})}{1+2u^2}}\,.$$

Then the corresponding Cauchy problem has the following form, for $t < T, x \in \mathbb{R}$:

(6.2)
$$\begin{cases} 0 = \dfrac{\partial u}{\partial t} + \dfrac{1}{2}\left(\dfrac{1+x^2}{2+x^2}\right)^2 \dfrac{1+2u^2}{1+u^2+\exp(-\frac{2x^2}{t+1})}\dfrac{\partial^2 u}{\partial x^2} + \dfrac{2x(1+x^2)}{(2+x^2)^3}\dfrac{\partial u}{\partial x} \\ \\ \qquad + \dfrac{1}{t+1}\exp\left(-\dfrac{x^2}{t+1}\right)\left[\dfrac{4x^2(1+x^2)}{(2+x^2)^3} + \left(\dfrac{1+x^2}{2+x^2}\right)^2\left(1 - \dfrac{2x^2}{t+1}\right) - \dfrac{x^2}{t+1}\right], \\ \\ u(T,x) = \exp(-\frac{x^2}{T+1}). \end{cases}$$

It is easy to verify that the solution to problem (6.2) is

$$u(t,x) = \exp\left(-\frac{x^2}{t+1}\right). \tag{6.3}$$

Since $Y(t) = u(t, X(t))$, plugging $Y(t)$ into the forward equation in (6.1), we get

$$dX = \frac{X(1+X^2)}{(2+X^2)^3}dt + \frac{1+X^2}{2+X^2}dW(t), \quad X(0) = x. \tag{6.4}$$

One can then easily check that the solution to this equation can be expressed as

$$X(t) = \Lambda(x + \arctan x + W(t)),$$

where the function $\Lambda(z)$ is defined by the equation

$$\Lambda(z) + \arctan \Lambda(z) = z. \tag{6.5}$$

Differentiating (6.5) with respect to z, we get

$$\Lambda'(z) = \frac{1+\Lambda^2(z)}{2+\Lambda^2(z)} > 0, \quad \forall\ z \in \mathbb{R}, \quad \text{and} \quad \Lambda''(z) = \frac{2\Lambda(1+\Lambda^2)}{(2+\Lambda^2)^3}, \tag{6.6}$$

which implies that $\Lambda(z)$ is a one-to-one function. Hence $X(0) = \Lambda(x + \arctan x) = x$. Furthermore, due to the Itô formula, we have

$$\begin{aligned} dX &= \Lambda'(x + \arctan x + W(t))dW(t) + \frac{1}{2}\Lambda''(x + \arctan x + W(t))dW(t) \\ &= \frac{X(1+X^2)}{(2+X^2)^3}dt + \frac{1+X^2}{2+X^2}dW(t). \end{aligned}$$

Hence, the solution to (6.1) is

$$\begin{aligned} X(t) &= \Lambda(x + \arctan x + W(t)), \\ Y(t) &= \exp\left(-\frac{X^2}{t+1}\right), \\ Z(t) &= -\frac{2X(1+X^2)}{(t+1)(2+X^2)} \exp\left(-\frac{X^2}{t+1}\right). \end{aligned}$$

We first carry out numerical tests on the parabolic equation (6.2) using the second-order (in time) scheme (3.8). In Tables 1–3, we tabulate the errors $\max_k \|u(t_k,x) - u_N^k(x)\|_{L^2(\mathbb{R})}$, $\max_k \|\frac{\partial u}{\partial x}(t_k,x) - \frac{d}{dx}u_N^k(x)\|_{L^2(\mathbb{R})}$, and $\max_k \|\frac{\partial^2 u}{\partial x^2}(t_k,x) - \frac{d^2}{dx^2}u_N^k(x)\|_{L^2(\mathbb{R})}$, respectively. We note that it is important to measure the errors of the approximation $u_N^k(x)$ to the first and second derivatives of $u(t_k,x)$ since the first-order Milstein scheme (3.11) and $\frac{3}{2}$-order Platen–Wagner scheme (3.12) need to use these derivatives. In the last column of these tables, the rates of convergence in time with $N = 150$ are reported. We observe that essentially second-order accuracy in time and exponential convergence in space are achieved for all three quantities.

Next, we examine the errors of the full simulation for the FBSDE with four different schemes presented in section 3 for the forward SDE and with the Hermite-collocation scheme (3.7). For the sake of comparison with the results in [31], we

TABLE 1
$\max_k \|u(t_k,x) - u_N^k(x)\|_{L^2(\mathbb{R})}$.

$h\backslash N$	64	70	100	128	150	Rate
0.5	0.0049	0.0049	0.0049	0.0049	0.0049	
0.2	0.0010	0.0010	0.0010	0.0010	0.0010	1.7344
0.05	1.5391e-004	9.0229e-005	7.8519e-005	7.8554e-005	7.8555e-005	1.8351
0.02	1.5389e-004	8.3270e-005	1.3177e-005	1.3167e-005	1.3168e-005	1.9492
0.005	1.5388e-004	8.3266e-005	3.9669e-006	8.4446e-007	8.4360e-007	1.9822

TABLE 2
$\max_k \|\frac{\partial u}{\partial x}(t_k,x) - \frac{d}{dx}u_N^k(x)\|_{L^2(\mathbb{R})}$.

$h\backslash N$	64	70	100	128	150	Rate
0.5	0.0075	0.0074	0.0074	0.0074	0.0074	
0.2	0.0018	0.0017	0.0017	0.0017	0.0017	1.6052
0.05	0.0017	9.6178e-004	1.3741e-004	1.3685e-004	1.3685e-004	1.8174
0.02	0.0017	9.6179e-004	5.4862e-005	2.3165e-005	2.3152e-005	1.9391
0.005	0.0017	9.6178e-004	5.4859e-005	3.7980e-006	1.4925e-006	1.9777

TABLE 3
$\max_k \|\frac{\partial^2 u}{\partial x^2}(t_k,x) - \frac{d^2}{dx^2}u_N^k(x)\|_{L^2(\mathbb{R})}$.

$h\backslash N$	64	70	100	128	150	Rate
0.5	0.0230	0.0226	0.0224	0.0224	0.0224	
0.2	0.0188	0.0112	0.0055	0.0055	0.0055	1.5326
0.05	0.0188	0.0112	7.7070e-004	4.5192e-004	4.5174e-004	1.8029
0.02	0.0188	0.0112	7.7061e-004	7.8042e-005	7.7116e-005	1.9293
0.005	0.0188	0.0112	7.7058e-004	6.0488e-005	8.0385e-006	1.6310

also computed the same problem with a Monte Carlo simulation (with $S = 1000$ independent realizations of $X(T)$ and X_N) for the forward SDE. The results of the Euler scheme with Monte Carlo simulation are reported in Table 4. The averages presented in the table are computed as follows:

$$E(X(T) - X_N)^2 = \frac{1}{S}\sum_{k=1}^{S}\left(X^{(k)}(T) - X_N^{(k)}\right)^2 \pm 2\sqrt{\frac{D_S}{S}},$$

where

$$D_S = \frac{1}{S}\sum_{k=1}^{S}\left(X^{(k)}(T) - X_N^{(k)}\right)^4 - \left[\frac{1}{S}\sum_{k=1}^{S}\left(X^{(k)}(T) - X_N^{(k)}\right)^2\right]^2.$$

Hence,

$$\begin{aligned}\left[E(X(T) - X_N)^2\right]^{1/2} &= \sqrt{\frac{1}{S}\sum_{k=1}^{S}\left(X^{(k)}(T) - X_N^{(k)}\right)^2} \\ &\quad + \sqrt{\frac{1}{S}\sum_{k=1}^{S}\left(X^{(k)}(T) - X_N^{(k)}\right)^2 \pm 2\sqrt{\frac{D_S}{S}}} \\ &\quad - \sqrt{\frac{1}{S}\sum_{k=1}^{S}\left(X^{(k)}(T) - X_N^{(k)}\right)^2},\end{aligned}$$

where $X^{(k)}(T)$ and $X_N^{(k)}$ are independent realizations of $X(T)$ and X_N, respectively $(k = 1, 2, \ldots, S = 1000)$.

TABLE 4
Euler scheme with 1000 *Monte Carlo simulation realizations with* $N = 150$.

h	$\left[E(X(T) - X_N)^2\right]^{1/2}$	Rate	$\left[E(Y(T) - Y_N)^2\right]^{1/2}$	$\left[E(Z(T) - Z_N)^2\right]^{1/2}$
0.5	0.2453 ± 0.0121		0.0325 ± 0.0022	$0.0121 \pm 7.9404e-004$
0.2	0.1603 ± 0.0083	0.4643	0.0213 ± 0.0015	$0.0079 \pm 5.0520e-004$
0.05	0.0745 ± 0.0037	0.5527	0.0101 ± 0.0007	$0.0037 \pm 2.2556e-004$
0.02	0.0479 ± 0.0024	0.482	0.0064 ± 0.0004	$0.0023 \pm 1.4899e-004$
0.005	0.0259 ± 0.0013	0.4435	0.0036 ± 0.0002	$0.0013 \pm 7.6865e-005$

For conciseness, we now omit the Monte Carlo errors, which, as can be seen from Table 4, are significantly smaller than the approximation errors. In what follows, we denote $[E(X(T) - X_N)^2]^{1/2}$ by X, $[E(Y(T) - Y_N)^2]^{1/2}$ by Y, and $[E(Z(T) - Z_N)^2]^{1/2}$ by Z.

In the Tables 5–7, we report these errors by using the three schemes with 1000 independent realizations of $X(T)$ and X_N.

TABLE 5
First-order SDE scheme with $N = 150$.

h	X	Rate	Y	Rate	Z	Rate
0.5	0.0937		0.0131		0.0048	
0.2	0.0394	0.9455	0.0052	1.0084	0.0021	0.9022
0.05	0.0095	1.0261	0.0013	1.000	4.7994e-004	1.0647
0.02	0.0039	0.9717	5.3451e-004	0.9700	1.9852e-004	0.9634
0.005	9.5023e-004	1.0186	1.2996e-004	1.0201	4.8413e-005	1.0179

TABLE 6
$\frac{3}{2}$*-order strong scheme with* $N = 200$.

h	X	Rate	Y	Rate	Z	Rate
0.5	0.0425		0.0041		0.0025	
0.2	0.0115	1.4266	0.0015	1.0974	5.4051e-004	1.6714
0.05	0.0015	1.4693	1.7410e-004	1.5535	7.5482e-005	1.4201
0.02	3.3583e-004	1.6333	3.6401e-005	1.7080	1.7780e-005	1.5779
0.005	3.9962e-005	1.5355	7.6776e-006	1.1226	2.2442e-006	1.4930

TABLE 7
$\frac{3}{2}$*-order SAB scheme with* $N = 200$.

h	X	Rate	Y	Rate	Z	Rate
0.5	0.1301		0.0177		0.0063	
0.2	0.0400	1.2872	0.0053	1.3160	0.0020	1.2522
0.05	0.0048	1.5294	6.6087e-004	1.5018	2.3552e-004	1.5430
0.02	0.0013	1.4256	1.6757e-004	1.4975	6.5992e-005	1.3885
0.005	1.5749e-004	1.5226	2.0907e-005	1.5014	7.9073e-006	1.5305

Observe that all three schemes produce expected convergence rates. The $\frac{3}{2}$-order SAB scheme is slightly less accurate than the $\frac{3}{2}$-order strong scheme, but the convergence rates of these two schemes are essentially the same.

7. Concluding remarks. We presented a numerical method for a class of forward-backward stochastic differential equations (FBSDEs). The method is based on the *four step scheme* with a Hermite-spectral method to approximate the solution to the decoupling quasi-linear PDE on the whole space. The use of the Hermite-spectral method not only avoids the use of artificial far field boundary conditions but also leads to spectrally accurate results in space. We carried out a rigorous error analysis for a fully discretized scheme for the FBSDEs with a first-order scheme in time and a Hermite-spectral scheme in space, and indicated that similar analysis can be extended to higher-order schemes in time. We presented detailed numerical comparisons between several schemes for the resulting decoupled forward SDE and showed that the stochastic version of the Adams–Bashforth scheme coupled with the Hermite-spectral method leads to a convergence rate of $\frac{3}{2}$ (in time).

Although the analysis and computation is performed for the one-dimensional case, the numerical scheme, and most of the analysis, can be extended to higher-dimensional cases. More precisely, the Hermite-spectral method for the PDE can be extended to high-dimensional cases in a straightforward matter using a tensor-product approach. Furthermore, the results in Theorem 4.1 still hold for high-dimensional cases; unfortunately, it appears to be very difficult to establish the results in Theorem 5.1 for high-dimensional cases due mainly to the fact that the Sobolev-type inequality (5.9) is no longer valid for higher-dimensional cases. On the other hand, the aforementioned tensor-product approach is feasible only for two- or three-dimensional problems, as the number of unknowns grows exponentially fast as the dimension increases, and the computational cost quickly becomes prohibitive for problems of four or more dimensions. In a forthcoming work, we plan to introduce new elliptic solvers based on lattice rules and sparse grids which would allow us to handle FBSDEs in relatively large space dimensions.

Acknowledgment. The authors would like to thank an anoymous referee whose comments and suggestions have led to numerous improvements to the paper.

REFERENCES

[1] V. Bally, *Approximation scheme for solutions of BSDE*, in Backward Stochastic Differential Equations (Paris, 1995–1996), Pitman Res. Notes Math. Ser. 364, Longman, Harlow, UK, 1997, pp. 177–191.

[2] V. Bally and G. Pagès, *A quantization algorithm for solving multi-dimensional discrete-time optimal stopping problems*, Bernoulli, 9 (2003), pp. 1003–1049.

[3] V. Bally, G. Pagès, and J. Printems, *A quantization tree method for pricing and hedging multidimensional American options*, Math. Finance, 15 (2005), pp. 119–168.

[4] C. Bender and J. Zhang, *Time discretization and Markovian iteration for coupled FBSDEs*, Ann. Appl. Probab., 18 (2008), pp. 143–177.

[5] B. Bouchard and N. Touzi, *Discrete-time approximation and Monte-Carlo simulation of backward stochastic differential equations*, Stochastic Process. Appl., 111 (2004), pp. 175–206.

[6] P. Briand, B. Delyon, and J. Mémin, *Donsker-type theorem for BSDEs*, Electron. Comm. Probab., 6 (2001), pp. 1–14.

[7] H.-J. Bungartz and T. Dornseifer, *Sparse grids: Recent developments for elliptic partial differential equations*, in Multigrid Methods V (Stuttgart, 1996), Lect. Notes Comput. Sci. Eng. 3, Springer, Berlin, 1998, pp. 45–70.

[8] D. Chevance, *Numerical methods for backward stochastic differential equations*, in Numerical Methods in Finance, Publ. Newton Inst., Cambridge University Press, Cambridge, UK, 1997, pp. 232–244.

[9] J. Cvitanić and J. Zhang, *The steepest descent method for forward-backward SDEs*, Electron. J. Probab., 10 (2005), pp. 1468–1495.

[10] F. Delarue, *On the existence and uniqueness of solutions to FBSDEs in a non-degenerate case*, Stochastic Process. Appl., 99 (2002), pp. 209–286.
[11] F. Delarue, *Estimates of the solutions of a system of quasi-linear PDEs. A probabilistic scheme*, in Séminaire de Probabilités XXXVII, Lecture Notes in Math. 1832, Springer, Berlin, 2003, pp. 290–332.
[12] F. Delarue and S. Menozzi, *A forward-backward stochastic algorithm for quasi-linear PDEs*, Ann. Appl. Probab., 16 (2006), pp. 140–184.
[13] J. Douglas, Jr., J. Ma, and P. Protter, *Numerical methods for forward-backward stochastic differential equations*, Ann. Appl. Probab., 6 (1996), pp. 940–968.
[14] N. El Karoui and L. Mazliak, eds., *Backward Stochastic Differential Equations* (Paris, 1995–1996), Pitman Res. Notes Math. Ser. 364, Longman, Harlow, UK, 1997.
[15] N. El Karoui, S. Peng, and M. C. Quenez, *Backward stochastic differential equations in finance*, Math. Finance, 7 (1997), pp. 1–71.
[16] B. D. Ewald and R. Témam, *Numerical analysis of stochastic schemes in geophysics*, SIAM J. Numer. Anal., 42 (2005), pp. 2257–2276.
[17] I. M. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, 6th ed., Academic Press, San Diego, CA, 2000.
[18] B.-Y. Guo, J. Shen, and C.-L. Xu, *Spectral and pseudospectral approximations using Hermite functions: Application to the Dirac equation*, Adv. Comput. Math., 19 (2003), pp. 35–55.
[19] F. J. Hickernell, *Lattice rules: How well do they measure up?*, in Random and Quasi-Random Point Sets, Lecture Notes in Statist. 138, Springer, New York, 1998, pp. 109–166.
[20] K. Itô, *Multiple Wiener integral*, J. Math. Soc. Japan, 3 (1951), pp. 157–169.
[21] P. E. Kloeden and E. Platen, *Numerical Solution of Stochastic Differential Equations*, Springer, Berlin, 1992.
[22] F. Y. Kuo, I. H. Sloan, and H. Woźniakowski, *Lattice rules for multivariate approximation in the worst case setting*, in Monte Carlo and Quasi-Monte Carlo Methods 2004, Springer, Berlin, 2006, pp. 289–330.
[23] J.-P. Lemor, E. Gobet, and X. Warin, *A regression-based Monte Carlo method for backward stochastic differential equations*, Ann. Appl. Probab., 15 (2005), pp. 2172–2202.
[24] J.-P. Lemor, E. Gobet, and X. Warin, *Rate of convergence of an empirical regression method for solving generalized backward stochastic differential equations*, Bernoulli, 12 (2006), pp. 889–916.
[25] D. Li and F. J. Hickernell, *Trigonometric spectral collocation methods on lattices*, in Recent Advances in Scientific Computing and Partial Differential Equations (Hong Kong, 2002), Contemp. Math. 330, AMS, Providence, RI, 2003, pp. 121–132.
[26] J. Ma, P. Protter, J. San Martin, and S. Torres, *Numerical method for backward stochastic differential equations*, Ann. Appl. Probab., 12 (2002), pp. 302–316.
[27] J. Ma, P. Protter, and J. Yong, *Solving forward-backward stochastic differential equations explicitly—a four step scheme*, Probab. Theory Related Fields, 98 (1994), pp. 339–359.
[28] J. Ma and J. Yong, *Forward-backward stochastic differential equations and their applications*, Lecture Notes in Math. 1702, Springer, Berlin, 1999.
[29] J. Ma and J. Zhang, *Representations and regularities for solutions to BSDEs with reflections*, Stochastic Process. Appl., 115 (2005), pp. 539–569.
[30] G. N. Milstein and M. V. Tretyakov, *Stochastic Numerics for Mathematical Physics*, Springer, Berlin, 2004.
[31] G. N. Milstein and M. V. Tretyakov, *Numerical algorithms for forward-backward stochastic differential equations*, SIAM J. Sci. Comput., 28 (2006), pp. 561–582.
[32] É. Pardoux and S. Peng, *Backward stochastic differential equations and quasilinear parabolic partial differential equations*, in Stochastic Partial Differential Equations and Their Applications (Charlotte, NC, 1991), Lecture Notes in Control and Inform. Sci. 176, Springer, Berlin, 1992, pp. 200–217.
[33] G. Szegő, *Orthogonal Polynomials*, 4th ed., Amer. Math. Soc. Colloq. Publ. XXIII, AMS, Providence, RI, 1975.
[34] B. Volker, N. Erich, and R. Klaus, *High dimensional polynomial interpolation on sparse grids*, Adv. Comput. Math., 12 (2000), pp. 273–288.
[35] J. Zhang, *A numerical scheme for BSDEs*, Ann. Appl. Probab., 14 (2004), pp. 459–488.
[36] Y. Zhang and W. Zheng, *Discretizing a backward stochastic differential equation*, Int. J. Math. Math. Sci., 32 (2002), pp. 103–116.

SIAM J. NUMER. ANAL.
Vol. 46, No. 5, pp. 2662–2687
© 2008 Society for Industrial and Applied Mathematics

CONVERGENCE OF IMPLICIT MONOTONE SCHEMES WITH APPLICATIONS IN MULTIPHASE FLOW IN POROUS MEDIA*

FELIX KWOK† AND HAMDI A. TCHELEPI‡

Abstract. Phase-based upstreaming, which is a commonly used spatial discretization for multiphase flow in reservoir simulation, naturally gives rise to implicit monotone schemes when implicit time-stepping is used. The nonlinear Gauss–Seidel and Jacobi algorithms are shown to converge to a unique bounded solution when applied to the resulting system of equations. Thus, for one-dimensional problems, we obtain an alternate, constructive proof that such schemes are well-defined and converge to the entropy solution of the conservation law for arbitrary CFL numbers. The accuracy of phase-based upstream solutions is studied for various spatial and temporal grids, revealing the importance of unconditional stability when nonuniform grids and/or variable porosity is involved.

Key words. conservation laws, multiphase flow, reservoir simulation, upstream weighting, implicit monotone schemes

AMS subject classifications. 35L65, 65H10, 65M06, 65M12

DOI. 10.1137/070703922

1. Introduction.

In the simulation of multiphase flow in petroleum reservoirs, the most commonly used model consists of a system of n conservation laws, where n is the number of immiscible fluid phases. Each conservation law is defined for all $\mathbf{x} \in \Omega \subset \mathbb{R}^k$ $(1 \le k \le 3)$ and has the form

$$\frac{\partial(\phi \rho_j S_j)}{\partial t} + \nabla \cdot (\rho_j \mathbf{v}_j) = \rho_j q_j, \qquad j = 1, \ldots, n, \tag{1.1}$$

where $\phi = \phi(\mathbf{x})$ is the porosity of the medium (with $0 < \phi_{min} \le \phi \le 1$), $K = K(\mathbf{x}) \ge K_{min} > 0$ is the absolute permeability; for each phase j, $S_j = S_j(\mathbf{x}, t)$ is the saturation (i.e., the volume fraction occupied by phase j in the neighborhood of $\mathbf{x}$), $\rho_j > 0$ is the density, $q_j = q_j(\mathbf{x})$ is the source or sink term, and $\mathbf{v}_j$ is the phase velocity, which is given by generalized Darcy's law:

$$\mathbf{v}_j = -K\lambda_j \left[\nabla p_j - \rho_j \mathbf{g}\right]. \tag{1.2}$$

Here $\lambda_j = \lambda_j(S_1, \ldots, S_n)$ is the phase mobility, p_j is the phase pressure, and $\mathbf{g} \in \mathbb{R}^k$ is a constant vector representing the gravitational acceleration. In addition, we have the following algebraic relations:

$$\text{Saturation constraint:} \qquad \sum S_j = 1, \tag{1.3}$$

$$\text{Capillary pressure constraint:} \quad p_j - p_{j+1} = P_{cj}(S_1, \ldots, S_n), \quad j = 1, \ldots, n-1. \tag{1.4}$$

*Received by the editors September 27, 2007; accepted for publication (in revised form) March 18, 2008; published electronically July 2, 2008. This research was supported by the SUPRI-B Reservoir Simulation Affiliates program at Stanford University.

http://www.siam.org/journals/sinum/46-5/70392.html

†Section de Mathématiques, Université de Genéve, 2–4 rue du Liévre, CP 64, CH-1211 Geneva, Switzerland (Felix.Kwok@math.unige.ch). This author's research was done while he was a student at the Institute of Computational Mathematics and Engineering at Stanford University.

‡Energy Resources Engineering, Stanford University, Stanford, CA 94305 (tchelepi@stanford.edu).

In this paper we restrict our attention to incompressible flow with zero capillarity; i.e., we assume that the densities ρ_j are constant with respect to time and space, and that $p_1 = \cdots = p_n \equiv p$. The resulting system of PDEs exhibits a mixed hyperbolic-parabolic character, which becomes apparent when we consider the various limiting cases. If we multiply (1.1) by $1/\rho_j$ and sum over $j = 1, \ldots, N$, then we obtain $\sum_j \nabla \cdot \mathbf{v}_j = \sum_j q_j$, where the $\partial/\partial t$ terms cancel because of the saturation constraint. We have

$$-\nabla \cdot \left[K\lambda_T \nabla p - K\mathbf{g}\textstyle\sum_j \lambda_j \rho_j \right] = \sum_j q_j, \tag{1.5}$$

where $\lambda_T = \sum_j \lambda_j$ is the total mobility, which is assumed to be positive and uniformly bounded away from zero. Thus, for a given saturation distribution, the pressure field satisfies an elliptic PDE. On the other hand, if we define the total velocity $\mathbf{v}_T$ by $\mathbf{v}_T = \sum_j \mathbf{v}_j$, then we can rewrite $\mathbf{v}_j$ as

$$\mathbf{v}_j = \frac{\lambda_j}{\lambda_T} \left[\mathbf{v}_T - K\mathbf{g}\textstyle\sum_\ell \lambda_\ell (\rho_\ell - \rho_j) \right]. \tag{1.6}$$

This means when $\mathbf{v}_T$ is constant over time (which is the case for flow in a one-dimensional (1D) porous medium), the phase velocities $\mathbf{v}_j$ become functions of $\mathbf{x}$ and $S_1, \ldots, S_n$ only, so when we substitute (1.6) into (1.1), we get

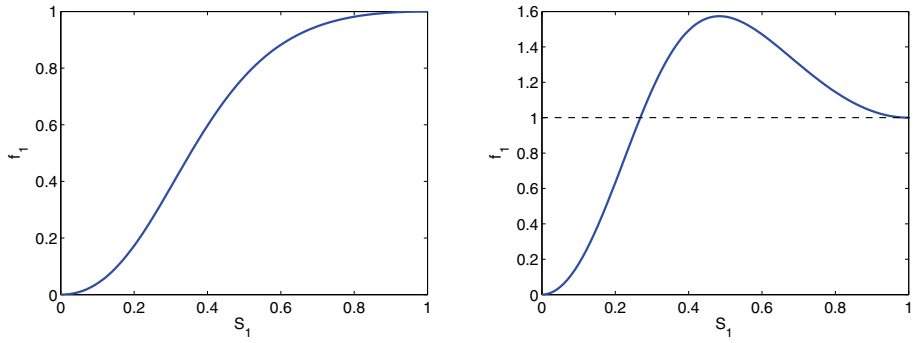
$$\phi \frac{\partial S_j}{\partial t} + \nabla \cdot \mathbf{v}_j(\mathbf{x}, S_1, \ldots, S_n) = 0, \qquad j = 1, \ldots, n-1. \tag{1.7}$$

As a result, saturation behaves like the solution to a system of first-order hyperbolic conservation laws, so one should expect discontinuous saturation profiles. In higher dimensions, $\mathbf{v}_T$ generally varies over time, and a strong coupling exists between pressure and saturation because of the saturation dependence of λ_j and λ_T in (1.5), as well as the dependence of $\mathbf{v}_T$ on the pressure field. However, one can still observe the same kind of discontinuities in the saturation profile in such cases.

The vastly different behavior between saturation and pressure means these variables need to be treated differently in a numerical method. The shock-forming nature of (1.7) requires the use of a finite-volume method, whereas the elliptic nature of (1.5) means pressure variables must be treated implicitly to avoid a time-step restriction of the form $\Delta t = \mathcal{O}(\Delta x^2)$. One can treat saturation variables either explicitly, for which a Courant–Friedrichs–Lewy (CFL) condition of the form $\Delta t = \mathcal{O}(\Delta x)$ applies, or implicitly, where no such time-step limit exists. In problems of practical interest, the porosity ϕ and permeability K are highly oscillatory, nonsmooth functions of $\mathbf{x}$, and $K(\mathbf{x})$ can vary by several orders of magnitude over the domain Ω. Thus, local CFL numbers can also exhibit large spatial variations. This means the time-step limit of an explicit method, which is determined by the maximum local CFL number, is often overly restrictive compared to the *average* CFL number. For this reason, the discretization of choice for most reservoir simulators is the *fully implicit method* (FIM), which uses finite volume in space and backward Euler in time. The numerical flux function $F_{j,il}$, which approximates the continuous flux $\int_{\partial V_{il}} \rho_j \mathbf{v}_j \cdot \mathbf{n}_{il}\, ds$, is given by

$$F_{j,il} = |\partial V_{il}| \rho_j K_{il} \lambda_j(S^*) \left(-\frac{(p_l - p_i)(\mathbf{x}_l - \mathbf{x}_i)}{|\mathbf{x}_l - \mathbf{x}_i|^2} + \rho_j \mathbf{g} \right) \cdot \mathbf{n}_{il}, \tag{1.8}$$

where $|\partial V_{il}|$ is the area of the interface between cells i and l, $\mathbf{x}_i$ and $\mathbf{x}_l$ are the locations of the centers of cells i and l, and $\mathbf{n}_{il}$ is the unit normal to the cell interface, pointing



(a) (b)

FIG. 1.1. *Flux functions for* 1*D incompressible two-phase flow:* (a) *Cocurrent flow (no buoyancy effects);* (b) *countercurrent flow due to gravity.*

from cell i to cell l. The mobility λ_j is evaluated at the *upstream* saturation:

$$S = \begin{cases} S(x_i) & \text{if } \left(-\dfrac{(p_l - p_i)(\mathbf{x}_l - \mathbf{x}_i)}{|\mathbf{x}_l - \mathbf{x}_i|^2} + \rho_j \mathbf{g}\right) \cdot \mathbf{n}_{il} \geq 0, \\ S(x_l) & \text{otherwise.} \end{cases} \tag{1.9}$$

As we will see in section 4, the resulting numerical flux functions are different from those used in classical CFD (e.g., the Godunov and Engquist–Osher schemes). Despite being only first-order accurate, phase-based upstreaming is the preferred upwind method in reservoir simulation because it is physically intuitive, and because it is generally easier to verify a consistency condition such as (1.9) than to locate potential sonic points, which vary over the domain and are strong functions of permeability and total velocity. This is especially true for the FIM because the total velocity at time t^{n+1} is usually unknown.

Notice that in (1.9) it is possible for each phase to have a different upstream direction when the densities ρ_j are different, i.e., when buoyancy forces are significant; this is known as *countercurrent flow* in reservoir engineering literature. In 1D porous media, countercurrent flow manifests itself through the presence of sonic points in the flux function v_j; thus, the flux function for a countercurrent flow problem would typically look like the one shown in Figure 1.1(b), whereas without countercurrent flow it would look more like Figure 1.1(a). A detailed treatment of phase-based upstreaming is given in [3], in which the authors show that, when explicit saturations are used on a two-phase flow problem, phase-based upstreaming leads to a monotone difference scheme, as long as the appropriate CFL condition is satisfied. This in turn implies that solutions obtained from an explicit scheme converge to the entropy solution of the two-phase equations

$$\frac{\partial S_1}{\partial t} + \frac{\partial f_1}{\partial x} = 0, \tag{1.10}$$

$$f_1(x, S_1) = v_1(x, S_1) = \frac{\lambda_1(S_1)}{\lambda_1(S_1) + \lambda_2(S_1)} \left[v_T + K(x) g \lambda_2(S_1)(\rho_1 - \rho_2)\right] \tag{1.11}$$

as $\Delta t, \Delta x \to 0$ while satisfying the CFL condition. The goal of this paper is to extend this result for the fully implicit case. This leads us to study the more general problem of implicit monotone schemes, of which the multiphase flow problem is a special case.

The use of implicit time stepping leads to a (typically large) system of nonlinear algebraic equations that must be solved for each time step. Moreover, the residual functions are generally nondifferentiable due to upstreaming criteria of the form (1.9); thus, the existence of a unique solution to these systems of equations is not immediately obvious. For implicit monotone schemes for 1D scalar conservation laws, Lucier [9] showed that a unique solution to the discrete problem exists whenever the initial data is bounded and has bounded total variation. The proof of existence, which relies heavily on Crandall–Liggett theory [5], proceeds along the following lines (see [6, Ch. 3] for more details). First, one shows that the residual function R for the numerical scheme defines an m-accretive operator in the L^1 norm. Then by the Crandall–Ligett theorem, the ODE

$$\frac{du}{dt} = -Ru, \quad u(0) = x \tag{1.12}$$

has a unique solution for $t \in [0, \infty)$ for any initial point x. Let $u(t; x)$ denote the solution of (1.12) with starting point x. Then one shows that the *Poincaré operator* P_ω, which maps the point x to the point $u(\omega, x)$, is strictly contractive. Then by Banach's fixed point theorem, P_ω has a unique fixed point x_0. One then proceeds to prove that $u(t; x_0) = x_0$ for all $0 \le t \le \omega$; thus, $du/dt = 0$, which implies $Rx_0 = 0$.

While this argument does prove the existence and uniqueness of a solution to the discretized problem, the proof does not suggest a practical algorithm for finding the solution. In section 3, we present an alternate constructive proof of existence by showing that the classical Gauss–Seidel and Jacobi iterations converge for this class of problems. In fact, we show that the iterative methods converge whenever the initial data for the discrete problem is bounded, so the implicit scheme is well-defined even when the initial data does not have bounded variation in $\mathbb{R}$. The well-definedness of the numerical scheme, together with the total variation diminishing (TVD) property and the existence of a discrete entropy inequality, imply that the numerical scheme converges to the entropy solution as the mesh is refined (i.e., as $\Delta x \to 0$). This result holds for any mesh ratio $\lambda = \Delta t/\Delta x$ (i.e., for any Courant number).

The remainder of this paper is organized as follows. Section 2 states the necessary assumptions of our framework. Section 3 contains the main result of this paper, which asserts that the nonlinear Gauss–Seidel and Jacobi processes converge when applied to implicit monotone schemes. This leads to a constructive proof of well-definedness of the numerical scheme, from which convergence of the numerical scheme under grid refinement follows [14]. In section 4, we derive the numerical flux functions for an implicit monotone scheme that is equivalent to the coupled problem (1.1) in the 1D case, which would allow us to establish the well-definedness and convergence of the coupled problem. We also discuss the applicability of the above analysis for multidimensional problems. Finally, we discuss the convergence behavior and accuracy issues for implicit schemes with phase-based upstreaming in section 5.

2. Implicit monotone schemes. We consider the following fully implicit, finite volume discretization

$$\phi_i(u_i^{n+1} - u_i^n) + \lambda(F_{i+1/2}^{n+1} - F_{i-1/2}^{n+1}) = 0, \qquad \lambda = \Delta t/\Delta x, \quad i \in \mathbb{Z}, \tag{2.1}$$

where $F_{i+1/2}$ denotes the numerical flux across the interface between cells i and $i+1$. The above scheme approximates the 1D nonlinear conservation law

$$\phi(x)u_t + f(x,u)_x = 0, \qquad (x,t) \in \mathbb{R} \times \mathbb{R}^+, \tag{2.2}$$

$$u(x,0) = u^0(x), \tag{2.3}$$

which generalizes the problem (1.10), (1.11) to the variable porosity and permeability case. For simplicity, we assume a three-point scheme $F^{n+1}_{i+1/2} = F_{i+1/2}(u^{n+1}_i, u^{n+1}_{i+1})$; thus, the implicit stencil at cell i contains the value at cell i at time t^n, as well as the values at cells $i-1$, i, and $i+1$ at the *future* time t^{n+1}. We also ignore the effects of boundary conditions by limiting ourselves to Cauchy problems for the moment; in section 4 we explain how certain boundary conditions can be incorporated. Assume that f and F are both locally Lipschitz continuous (but not necessarily differentiable), and that the numerical flux function $F_{i+1/2}$ is *consistent* with f in the sense that $F_{i+1/2}(u,u) = f(x_{i+1/2}, u)$. Given we are interested in handling flux functions of the type shown in Figure 1.1(b), we cannot assume that the flux function $f(x,u)$ is monotonic in u, so sonic points may be present. However, the numerical flux function is assumed to satisfy the following condition, which also appears in [14] and is central to all our arguments in this paper.

Assumption 1 (monotonic fluxes). For all $i \in \mathbb{Z}$, the numerical flux function $F_{i+1/2}$ is nondecreasing in the first argument and nonincreasing in the second argument; i.e., for any w, we have $F_{i+1/2}(u,w) \le F_{i+1/2}(v,w)$ and $F_{i+1/2}(w,u) \ge F_{i+1/2}(w,v)$ whenever $u \le v$.

For two-phase flow problems in one dimension, the numerical flux functions obtained by phase-based upstreaming have the form

$$F_{i+1/2}(S_{1,i}, S_{1,i+1}) = \frac{\lambda_1(S_{1,i})}{\lambda_1(S_{1,i}) + \lambda_2(S_1^*)}[v_T + K\lambda_2 g(\rho_1 - \rho_2)], \tag{2.4}$$

where $g(\rho_1 - \rho_2) \ge 0$ and

$$S_1^* = \begin{cases} S_{1,i} & \text{if } v_T - K\lambda_1(S_i)g(\rho_1-\rho_2) \ge 0, \\ S_{1,i+1} & \text{otherwise.} \end{cases} \tag{2.5}$$

The detailed derivation is deferred to section 4. The following theorem, which summarizes several results by Brenier and Jaffré [3] in the case of two-phase flow, guarantees that $F_{i+1/2}(S_i, S_{i+1})$ has the required properties.

THEOREM 2.1. *Assume that the mobility of phase j is increasing with S_j and decreasing with the saturation of the other phase for $j = 1, 2$ (e.g., water and oil). Then the numerical fluxes obtained from phase-based upstreaming defined by* (2.4), (2.5) *are consistent, are Lipschitz continuous, and satisfy Assumption* 1.

The hypothesis on phase mobilities is physically realistic [2]. Assumption 1 ensures that (2.1) is an implicit monotone scheme, in the sense that the resulting residual function defines an m-accretive operator in $\ell^1(\mathbb{Z})$ (see [7] for a proof). We do not use m-accretivity directly in this work. Instead, we show that Assumption 1 also implies that the residual function is an M-function in the sense of Rheinboldt [11]; this is generally not equivalent to m-accretivity [8] but nonetheless allows us to prove the existence and uniqueness of solutions. Once we establish that the implicit scheme is well-defined, we can apply the following theorem of Sanders [14, Theorem II] to conclude that the solutions of (2.1) converge to the unique entropy solution of (2.2)

when f does not depend explicitly on x, i.e., when $f = f(u)$. Note that Assumption 1 implies that (2.1) is an E-scheme [10], so it is at most first-order accurate.

THEOREM 2.2 (see Sanders [14]). *Let $\mathbb{R} = \cup \mathcal{I}_i^n$ with $\mathcal{I}_i^n = [x_i, x_{i+1}) \times [t^n, t^{n+1})$ and $\delta = \sup_{i,n} |x_{i+1} - x_i| + |t^{n+1} - t^n|$. Suppose the conservation law (2.2), (2.3) with $f = f(u)$ is discretized using (2.1), where $F_{i+1/2}(u_i, u_{i+1}) \equiv F(u_i, u_{i+1})$ for all i. Assume that F is locally Lipschitz continuous, consistent, and monotonic. Define the step function v_δ such that $v_\delta(x,t) = u_i^n$ when $(x,t) \in \mathcal{I}_i^n$, and suppose the initial data (2.3) is discretized via the averaging operator T_δ,*

$$T_\delta(u_0)(x) = \frac{1}{|x_{i+1} - x_i|} \int_{x_i}^{x_{i+1}} u_0(\xi) d\xi,$$

when $x \in [x_i, x_{i+1})$. Then $v_\delta(x,t)$ converges in $L^\infty(L^1(\mathbb{R}); [0,T])$ to the unique entropy solution of (2.2), (2.3) as δ tends to zero.

3. Existence and uniqueness of solutions for the discretized problem. In this section, we show that the nonlinear Jacobi and Gauss–Seidel processes, when applied to the infinite system of nonlinear equations (2.1), converge to a unique bounded solution. Thus, we obtain an alternate constructive proof of the well-definedness of implicit monotone schemes. In addition, we show that Jacobi and Gauss–Seidel converge for any starting point that is bounded by the initial data, leading to a practical algorithm for computing the solution. Finally, we show how to extend the analysis to deal with problems with spatially varying coefficients, as well as problems for which the flux function $f(u)$ is only defined over a finite interval rather than all of $\mathbb{R}$. An extension to multidimensional problems is considered in section 4.2. (Note: Throughout this section, x and t are generic variables and do not denote space or time.)

3.1. Nonlinear Jacobi and Gauss–Seidel processes. Suppose we want to solve a nonlinear system of algebraic equations $R(x) = 0$ for $x \in \ell^\infty(\mathbb{N})$, where $R = (r_1, r_2, \dots)^T : \ell^\infty(\mathbb{N}) \to \ell^\infty(\mathbb{N})$. Then we can consider the *nonlinear Gauss–Seidel process*:

$$\begin{aligned} &\text{solve} \quad r_i(x_1^{k+1}, \dots, x_{i-1}^{k+1}, x_i^*, x_{i+1}^k, x_{i+2}^k, \dots) = 0 \text{ for } x_i^*, \\ &\text{set} \qquad x_i^{k+1} = x_i^*, \quad i = 1, 2, \dots, \quad k = 0, 1, 2, \dots, \end{aligned} \tag{3.1}$$

as well as the *nonlinear Jacobi process*:

$$\begin{aligned} &\text{solve} \quad r_i(x_1^k, \dots, x_{i-1}^k, x_i^*, x_{i+1}^k, x_{i+2}^k, \dots) = 0 \text{ for } x_i^*, \\ &\text{set} \qquad x_i^{k+1} = x_i^*, \quad i = 1, 2, \dots, \quad k = 0, 1, 2, \dots . \end{aligned} \tag{3.2}$$

The only difference between (3.1)/(3.2) and classical Gauss–Seidel/Jacobi processes is that each Gauss–Seidel/Jacobi "sweep" now involves infinitely many variables and equations. To ensure that the iterations (3.1)/(3.2) make sense, we need the following assumptions on the residual function R.

Assumption 2 (preservation of bounded sets). $R : \ell^\infty(\mathbb{N}) \to \ell^\infty(\mathbb{N})$ is a mapping between bounded sequences for which there exists an increasing function $\zeta : [0, \infty) \to [0, \infty)$ such that $\|x\|_\infty \le B$ implies $\|R(x)\| \le \zeta(B)$.

Assumption 3 (finite number of dependencies). For each i, the residual function $r_i(x_1, x_2, \dots)$ is nonconstant with respect to at most a finite number of x_j.

In other words, the residual functions must come from a compact stencil and must preserve boundedness. Assumption 3 ensures that the iterative processes are

well-defined, since it guarantees that for any given $i, k \in \mathbb{N}$, the value of x_i^{k+1} can be calculated using a finite number of univariate solves. It also guarantees that whenever R is continuous and (3.1)/(3.2) converges, the limit point x^* must satisfy $R(x^*) = 0$. Assumption 2 then allows us to extend Rheinboldt's analysis [11] to the infinite-dimensional case and prove that (3.1)/(3.2) converges to a unique bounded solution when R is an M-function.

Remark. Even though the nonlinear iterations (3.1)/(3.2) is well-defined in theory, a practical implementation would require that we either solve a finite-dimensional problem by supplying appropriate boundary conditions, or that we use delayed evaluation [1] to compute only those values of x_i^k that are needed.

3.2. *M*-function theory. M-functions are essentially generalizations of M-matrices in linear algebra. In the linear setting, it is well known [13] that the Gauss–Seidel method applied to $Ax = b$ converges for any right-hand side b and starting point x_0 if A is a nonsingular M-matrix. M-functions have similar properties with respect to the nonlinear Gauss–Seidel process, which is the subject of investigation in [11]. Here we provide extensions to the relevant definitions and theorems in [11] that would allow us to prove the existence and uniqueness of bounded solutions to (2.1).

For the remainder of this section, the natural partial ordering on $\ell^\infty(\mathbb{N})$ is written as $x \le y$, i.e., $x \le y$ iff $x_i \le y_i$ for all $i \in \mathbb{N}$. We denote by e^i the unit basis vectors with the ith component one and all others zero. The following definitions are essentially identical to those in [11], except the domain of definition has been changed from $\mathbb{R}^n$ to $\ell^\infty(\mathbb{N})$ to handle vectors of infinite length.

DEFINITION 3.1. *Let $R : \ell^\infty(\mathbb{N}) \to \ell^\infty(\mathbb{N})$.*

1. *R is* isotone *(or* antitone*) if, for all $x, y \in \ell^\infty(\mathbb{N})$, $x \le y$ implies $R(x) \le R(y)$ (or $R(x) \ge R(y)$). It is* strictly isotone *(or antitone) if $x < y$ implies $R(x) < R(y)$ (or $R(x) > R(y)$).*
2. *R is* inverse isotone *if, for all $x, y \in \ell^\infty(\mathbb{N})$, $R(x) \le R(y)$ implies $x \le y$.*
3. *R is (strictly)* diagonally isotone *if, for all $x \in \ell^\infty(\mathbb{N})$, the functions*

$$\rho_{ii} : \mathbb{R} \to \mathbb{R}, \quad \rho_{ii}(t) = r_i(x + te^i), \quad i = 1, 2, \dots, \tag{3.3}$$

   are (strictly) isotone.
4. *R is* off-diagonally antitone *if, for any $x \in \ell^\infty(\mathbb{N})$ the functions*

$$\rho_{ij} : \mathbb{R} \to \mathbb{R}, \quad \rho_{ij}(t) = r_i(x + te^j), \quad i \ne j, \quad i, j = 1, 2, \dots, \tag{3.4}$$

   are antitone.
5. *R is an M-*function *if R is inverse isotone and off-diagonally antitone.*

One characterization of M-functions is given by the following theorem.

THEOREM 3.2. *Let $R : \ell^\infty(\mathbb{N}) \to \ell^\infty(\mathbb{N})$ be off-diagonally antitone and satisfy Assumptions 2 and 3. Then R is an M-function if, for each $B > 0$, there exists a positive sequence $\{w_i^B\}$ such that*

1. *$\sum_{i=1}^\infty w_i^B < \infty$,*
2. *for any $\|x\|_\infty < B$, the function $Q(t) = (q_1(t), q_2(t), \dots)$ defined by*

$$q_i(t) = \sum_{j=1}^\infty w_j^B r_j(x + te^i)$$

is strictly isotone over the interval $t \in (t_{min}, t_{max})$, where $t_{min} = -B - \inf_i x_i$ and $t_{max} = B - \sup_i x_i$.

Proof. The proof is based on [11, Theorem 5.1], suitably modified to handle the infinite-dimensional case. Suppose $R(x) \le R(y)$ for some $x, y \in \ell^\infty(\mathbb{N})$. Define the sets

$$N^- = \{i \in \mathbb{N} \mid y_i < x_i\}; \quad N^+ = \{i \in \mathbb{N} \mid y_i \ge x_i\}.$$

Suppose N^- is nonempty. For each $i \in N^-$, let $\gamma_i = (x_i - y_i)e^i$. We consider two cases:

1. If $|N^-| < \infty$, let $i_1 < i_2 < \cdots < i_m$ be the elements of N^-, and define

$$z^0 = y, \quad z^1 = y + \gamma_{i_1}, \dots, \quad z^m = y + \gamma_{i_1} + \cdots + \gamma_{i_m},$$

   and let $z^k = z^m = z$ for all $k > m$.
2. If $|N^-| = \infty$, let $i_1 < i_2 < \cdots$ be the elements of N^-, and define

$$z^0 = y, \quad z^1 = y + \gamma_{i_1}, \dots, \quad z^k = y + \gamma_{i_1} + \cdots + \gamma_{i_k}, \dots,$$

   and let $z = \{z_i\}$ be such that $z_i = \max\{x_i, y_i\}$.

Define $R^k := R(z^k)$ and $R^\infty = R(z)$. In either case, we have the following properties:

1. $\|z^k\|_\infty < B$ and $\|z\|_\infty < B$, where $B = \max\{\|x\|_\infty, \|y\|_\infty\}$. Hence, by Assumption 2, $\|R^k\|_\infty \le \zeta(B)$ for all k (similarly for R^∞).
2. For each i, $z_i^k = z_i$ for large enough k, so by Assumption 3, $R_j^k \to R_j^\infty$ *pointwise* for each j.

Thus, the sequence $R^k = (R_j^k)_{j=1}^\infty$ is dominated by $G = (\zeta(B), \zeta(B), \dots)$ for each $k \in \mathbb{N}$. Moreover, $\sum_{j=1}^\infty w_j^B G_j < \infty$, so by the dominated convergence theorem [12],

$$\sum_{j=1}^\infty w_j^B R_j^k \to \sum_{j=1}^\infty w_j^B R_j^\infty \quad \text{as } k \to \infty.$$

By the strict isotonicity of Q, we have

$$\sum_{j=1}^\infty w_j^B R_j^0 \le \sum_{j=1}^\infty w_j^B R_j^1 \le \cdots$$

with at least one strict inequality (since N^- is nonempty). Thus, we must have

$$\sum_{j=1}^\infty w_j^B r_j(y) = \sum_{j=1}^\infty w_j^B R_j^0 < \sum_{j=1}^\infty w_j^B R_j^\infty = \sum_{j=1}^\infty w_j^B r_j(z). \tag{3.5}$$

Now split the last sum into two parts

$$\sum_{j=1}^\infty w_j^B r_j(z) = \sum_{j \in N^-} w_j^B r_j(z) + \sum_{j \in N^+} w_j^B r_j(z), \tag{3.6}$$

where the summation over N^+ may be empty. Then by off-diagonal antitonicity of R (and invoking the dominated convergence theorem whenever necessary), we can show similarly that

$$\sum_{j \in N^-} w_j^B r_j(z) \le \sum_{j \in N^-} w_j^B r_j(x), \qquad \sum_{j \in N^+} w_j^B r_j(z) \le \sum_{j \in N^+} w_j^B r_j(y), \tag{3.7}$$

using the fact that $z - x$ and $z - y$ vanish on N^- and N^+, respectively. Combining (3.5)–(3.7) gives

$$\sum_{j=1}^{\infty} w_j^B r_j(y) < \sum_{j\in N^-} w_j^B r_j(x) + \sum_{j\in N^+} w_j^B r_j(y), \tag{3.8}$$

which implies $\sum_{j\in N^-} w_j^B r_j(y) < \sum_{j\in N^-} w_j^B r_j(x)$. Thus, we must have $r_j(y) < r_j(x)$ for some $j \in N^-$, which contradicts the hypothesis $R(x) \le R(y)$. Hence N^- must be empty, so $x \le y$. ∎

COROLLARY 3.3. *Let R satisfy the hypotheses of Theorem* 3.2. *Let $z \in \ell^\infty(\mathbb{N})$. Then there is at most one bounded solution to the equation $R(x) = z$.*

Remark. In the context of discretized PDEs one normally assumes tacitly that the solution of interest must be bounded; this can be regarded as a boundary condition "at infinity." However, since such boundary conditions are not explicitly stated in the definition of M-functions, one must be careful to exclude any parasitic unbounded solutions that may arise. In fact, the solution is not necessarily unique if we allow unbounded solutions. Consider the linear function $R = (r_1, r_2, \dots)$ defined by $r_i(x) = x_i - \alpha x_{i+1}$ for $|\alpha| < 1$. Then for any $\|x\|_\infty < \infty$, we have $\|R(x)\|_\infty \le (1+\alpha)\|x\|_\infty$, so that Assumption 2 is satisfied. Assumption 3 (finitely many dependencies) is also satisfied because each r_i is only nonconstant with respect to two components of x. Finally, if we let $w_j^B = \beta^j$ for any $|\alpha| < \beta < 1$, then $\sum_j \beta^j < \infty$ and

$$q_i(t) = \sum_{j=1}^{\infty} \beta^j \left[x_j + t\delta_{ij} - \alpha(x_{j+1} + t\delta_{i,j+1})\right] = (\beta-\alpha)\beta^{i-1}t + \beta x_1 + (\beta-\alpha)\sum_{j=2}^{\infty} \beta^{j-1} x_j,$$

so $q_i(t)$ is well-defined and is strictly increasing with respect to t whenever $\|x\|_\infty < \infty$. So the hypotheses of Theorem 3.2 are satisfied, and hence $x = 0$ is the only bounded solution of $R(x) = 0$. However, unbounded solutions of the form $y = \{K\alpha^{-i}\}$, $K \ne 0$ also satisfy $R(y) = 0$, so the theorem does not preclude these possibilities.

3.3. Convergence of nonlinear Jacobi and Gauss–Seidel. It turns out the hypotheses of Theorem 3.2 are enough to ensure convergence of nonlinear Jacobi and Gauss–Seidel for certain starting points, which will be described below. The following result is essentially Theorem 3.1 in [11], with modified hypotheses to accommodate ℓ^∞-bounded vectors with infinitely many components. The proof in [11] goes through verbatim but is reproduced here because similar arguments also appear in the proof of Theorem 3.7. Note that by Assumption 3, each r_i depends on only finitely many arguments, so the standard arguments on limits, continuity, and antitonicity hold without additional complications when they are used on individual components of R.

THEOREM 3.4 (see Rheinboldt [11]). *Let $R : \ell^\infty(\mathbb{N}) \to \ell^\infty(\mathbb{N})$ satisfy the hypotheses of Theorem* 3.2. *Suppose for some $z \in \ell^\infty(\mathbb{N})$ there exist $x^0, y^0 \in \ell^\infty(\mathbb{N})$ such that*

$$x^0 \le y^0, \qquad R(x^0) \le z \le R(y^0).$$

Then the nonlinear Gauss–Seidel and Jacobi iterates $\{y^k\}$ and $\{x^k\}$, given by (3.1) *and* (3.2), *and starting from y^0 and x^0, respectively, are uniquely defined and satisfy*

$$x^0 \le x^k \le x^{k+1} \le y^{k+1} \le y^k \le y^0, \qquad R(x^k) \le z \le R(y^k) \tag{3.9}$$

for $k = 0, 1, \dots$. In addition, the pointwise limits

$$\lim_{k\to\infty} x^k = \lim_{k\to\infty} y^k = x^* \tag{3.10}$$

exist, and $R(x^*) = z$.

First, we need the following lemma (which is part of Theorem 2.10 in [11]).

LEMMA 3.5. *Let* $R : \ell^\infty(\mathbb{N}) \to \ell^\infty(\mathbb{N})$ *be an M-function. Then* R *is strictly diagonally isotone.*

Proof. Suppose that for some $x \in \ell^\infty(\mathbb{N})$, $s, t \in \mathbb{R}$, $s > t$ and index i we have $r_i(x + se^i) \le r_i(x + te^i)$. Off-diagonal antitonicity then implies that

$$r_j(x + se^i) \le r_j(x + te^i), \qquad j \ne i,$$

or, altogether, that $R(x + se^i) \le R(x + te^i)$. By inverse isotonicity, this leads to the contradiction $s \le t$, which shows that R must be strictly diagonally isotone. ∎

Proof of Theorem 3.4. We present only the proof for convergence of Gauss–Seidel; the proof for Jacobi is similar. We proceed by induction and suppose that for some $k \ge 0$ and $i \ge 1$

$$x^0 \le x^k \le y^k \le y^0, \qquad R(x^k) \le z \le R(y^k), \tag{3.11a}$$

$$x_j^k \le x_j^{k+1} \le y_j^{k+1} \le y_j^k, \qquad j = 1, \dots, i-1, \tag{3.11b}$$

where for $i = 1$ the relation (3.11b) is vacuous. Clearly, (3.11) is valid for $k = 0$ and $i = 1$. Define the functions

$$\alpha(s) = r_i(x_1^{k+1}, \dots, x_{i-1}^{k+1}, s, x_{i+1}^k, x_{i+2}^k, \dots),$$

$$\beta(s) = r_i(y_1^{k+1}, \dots, y_{i-1}^{k+1}, s, y_{i+1}^k, x_{i+2}^k, \dots)$$

for $s \in [x_i^0, y_i^0]$. Then (3.11) and the off-diagonal antitonicity of R yield

$$\beta(s) \le \alpha(s), \qquad s \in [x_i^0, y_i^0], \tag{3.12}$$

$$\beta(x_i^k) \le \alpha(x_i^k) \le r_i(x^k) \le z_i \le r_i(y^k) \le \beta(y_i^k) \le \alpha(y_i^k). \tag{3.13}$$

Since R is strictly diagonally isotone, α and β are both continuous and strictly increasing, so (3.13) implies the existence of unique $\hat{y}_i^k$ and $\hat{x}_i^k$ for which

$$\beta(\hat{y}_i^k) = z_i = \alpha(\hat{x}_i^k), \qquad x_i^k \le \hat{x}_i^k \le \hat{y}_i^k \le y_i^k,$$

where the relation $\hat{x}_i^k \le \hat{y}_i^k$ is a consequence of (3.12). But $x_i^{k+1} = \hat{x}_i^k$ and $y_i^{k+1} = \hat{y}_i^k$ by definition, so we have proved (3.11b) for $j = 1, \dots, i$. Hence by induction (3.11b) holds for all $i \in \mathbb{N}$, and hence $x^k \le x^{k+1} \le y^{k+1} \le y^k$. From this it follows again from off-diagonal antitonicity that

$$r_i(y^{k+1}) \ge r_i(y_1^{k+1}, \dots, y_i^{k+1}, y_{i+1}^k, y_{i+2}^k \dots) = z_i$$

and similarly that

$$r_i(x^{k+1}) \le r_i(x_1^{k+1}, \dots, x_i^{k+1}, x_{i+1}^k, x_{i+2}^k \dots) = z_i.$$

This completes the induction on k and hence the proof of (3.9). Applying the monotone convergence theorem for sequences, we conclude that the *pointwise* limits

$$\lim_{k\to\infty} x_j^k = x_j^* \le y_j^* = \lim_{k\to\infty} y_j^k$$

exist for each j, which allows us to define $x^* = \{x_j^*\}$ and $y^* = \{y_j^*\}$. Since each r_i is continuous and depends on only finitely many arguments, the definition of the Gauss–Seidel process then implies $r_i(x^*) = r_i(y^*) = z_i$ for each i, and hence $R(x^*) = R(y^*) = z$. Since both x^* and y^* are bounded, Corollary 3.3 implies that they are equal, completing the proof. ∎

3.4. Well-definedness of implicit monotone schemes. Using the theory in the previous two sections, we can now prove that implicit monotone schemes (i.e., implicit schemes whose flux functions satisfy Assumption 1) are well-defined for bounded initial conditions. What we need to show is that the residual functions satisfy the hypotheses of Theorem 3.4. In the interest of clarity, in this section we show only convergence of the iterative schemes for problems whose coefficients do not vary in space (i.e., corresponding to the conservation law $u_t + f(u)_x = 0$, discretized on a uniform spatial grid). In the next section, we state the additional assumptions on ϕ_i and $F_{i+1/2}$ that are required for the spatially varying case.

THEOREM 3.6. *Consider the numerical scheme* (2.1) *with*

$$F_{i+1/2}^{n+1} = F(u_i^{n+1}, u_{i+1}^{n+1}),$$

where $F : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ *is locally Lipschitz continuous and satisfies Assumption* 1*, i.e., nondecreasing in the first argument and nonincreasing in the second. Assume that the initial condition* $\{u_i^0\}_{i=-\infty}^{\infty}$ *is bounded. Then* (2.1) *has a unique bounded solution* $\{u_i^{n+1}\}$ *for* $n = 0, 1, 2, \dots$*. Moreover, this bounded solution satisfies the estimate*

$$\inf_{j\in\mathbb{Z}} u_j^n \le u_i^{n+1} \le \sup_{j\in\mathbb{Z}} u_j^n \qquad \forall i \in \mathbb{Z}. \tag{3.14}$$

Proof. First, we need to define an ordering for the Gauss–Seidel sweeps, i.e., to permute the equations and variables so that the spatial indices go from 1 to ∞ rather than from $-\infty$ to ∞. After that, it suffices to check that all of the hypotheses of Theorem 3.4 are satisfied for this ordering.

1. For $j = 1, 2, \dots$, define $\sigma(j) = (-1)^j \lfloor j/2 \rfloor$, i.e., σ maps $\{1, 2, 3, 4, 5, \dots\}$ to $\{0, 1, -1, 2, -2, \dots\}$. Let τ be the inverse map, such that $\tau(\sigma(j)) = j$. Define $R : \ell^\infty(\mathbb{N}) \to \ell^\infty(\mathbb{N})$ to be the reordered (and rescaled) set of residual equations, i.e.,

$$r_j(v) = \frac{v_j - u_{\sigma(j)}^n}{\lambda} + F(v_j, v_{\tau(\sigma(j)+1)}) - F(v_{\tau(\sigma(j)-1)}, v_j), \tag{3.15}$$

where $v_j = u_{\sigma(j)}^{n+1}$.

2. Since F is locally Lipschitz continuous, it is Lipschitz continuous over any compact set, so for any $B > 0$ there exists K_B (which can be chosen to be increasing with B) such that for any $(x, y) \in [-B, B] \times [-B, B]$,

$$|F(x, y) - F(0, 0)| \le K_B(|x| + |y|) \le 2K_B \cdot B.$$

Thus, for any $\|v\|_\infty \le B$, we have $|r_j(v)| \le \zeta(B)$ for all j, where

$$\zeta(B) = \Big(\frac{1}{\lambda} + 4K_B\Big)B + \frac{1}{\lambda}\|u^n\|_\infty.$$

Hence Assumption 2 is satisfied. Moreover, since each r_j depends only on v_j, $v_{\tau(\sigma(j)-1)}$ and $v_{\tau(\sigma(j)+1)}$, Assumption 3 (finite number of dependencies) is also satisfied.

3. By Assumption 1 (monotonic fluxes), F is clearly off-diagonally antitone. To satisfy the remaining hypotheses of Theorem 3.2, let $\{w_j^B\}$ take the form $w_j^B = \beta^{|\sigma(j)|}$ for some $0 < \beta < 1$, so that $\sum_{j=1}^{\infty} w_j^B < \infty$. An easy calculation shows that

$$q_i(t) := \sum_{j=1}^{\infty} w_j^B r_j(v + te^i) = \tilde{q}_i(t) + \sum_{j=1}^{\infty} w_j^B r_j(v),$$

where

$$\tilde{q}_i(t) = w_i^B t/\lambda + (w_i^B - w^B_{\tau(\sigma(i)+1)})\big[F(v_i + t, v_{\tau(\sigma(i)+1)}) - F(v_i, v_{\tau(\sigma(i)+1)})\big] + (w^B_{\tau(\sigma(i)-1)} - w_i^B)\big[F(v_{\tau(\sigma(i)-1)}, v_i + t) - F(v_{\tau(\sigma(i)-1)}, v_i)\big].$$

By the definition of w_i^B, we see that

$$|w_i^B - w^B_{\tau(\sigma(i)\pm 1)}| \le \beta^{|\sigma(i)|-1}(1-\beta),$$

which, when combined with the local Lipschitz continuity of F, gives

$$\beta^{|\sigma(i)|-1}\big[\beta t/\lambda - 2(1-\beta)K_B|t|\big] \le \tilde{q}_i(t) \le \beta^{|\sigma(i)|-1}\big[\beta t/\lambda + 2(1-\beta)K_B|t|\big].$$

Hence, $\tilde{q}_i(t)$ is strictly isotone whenever $\beta/\lambda > 2(1-\beta)K_B$, so picking

$$\frac{2\lambda K_B}{1+2\lambda K_B} < \beta < 1 \tag{3.16}$$

ensures isotonicity for $\tilde{q}_i(t)$ (and hence $q_i(t)$) for all i, as required in Theorem 3.2. (Note that the choice of β depends on B.)

4. We need to choose starting points x^0 and y^0 that satisfy the requirements of Theorem 3.4. Let x^0 and y^0 both be constant sequences with

$$x_i^0 = \inf_{j\in\mathbb{Z}} u_j^n, \qquad y_i^0 = \sup_{j\in\mathbb{Z}} u_j^n \qquad \forall i \in \mathbb{N}.$$

Then clearly $x^0 \le y^0$, and for all $i \in \mathbb{N}$,

$$r_i(x^0) = \frac{1}{\lambda}\left(x_i^0 - u^n_{\sigma(i)}\right) = \frac{1}{\lambda}\left(\inf_{j\in\mathbb{Z}} u_j^n - u^n_{\sigma(i)}\right) \le 0,$$

$$r_i(y^0) = \frac{1}{\lambda}\left(y_i^0 - u^n_{\sigma(i)}\right) = \frac{1}{\lambda}\left(\sup_{j\in\mathbb{Z}} u_j^n - u^n_{\sigma(i)}\right) \ge 0,$$

so $R(x^0) \le 0 \le R(y^0)$. Thus, by Theorem 3.4, the nonlinear Gauss–Seidel iterates $\{y^k\}$ and $\{x^k\}$ both converge (pointwise) to the unique solution x^* with $R(x^*) = 0$; hence, a unique solution to (2.1) exists, i.e., $u_i^{n+1} = x^*_{\tau(i)}$. Moreover, we know that $x^0 \le x^* \le y^0$, which immediately implies (3.14). ∎

Remark. The initial condition $\{u_i^0\}_{i=-\infty}^{\infty}$ is not assumed to be in ℓ^1 nor in BV, so this result is somewhat more general than results that use Crandall–Liggett theory.

Remark. The definition of an M-function is invariant under symmetric permutations; i.e., $R(x)$ is an M-function if and only if $\sigma R(\sigma x)$ is also an M-function for any permutation $\sigma : \mathbb{N} \to \mathbb{N}$. Thus, the Gauss–Seidel process must converge regardless of the way the ordering is chosen in step 1 of the proof. However, the *rate* of convergence is sensitive to the ordering [8].

In fact, one can show that the nonlinear Jacobi and Gauss–Seidel processes converge for any starting point $\{z_i^{(0)}\}$ that is bounded by the initial data $\{u_i^n\}$. (In what follows, superscripts in brackets indicate iterates within the Gauss–Seidel process, and superscripts without brackets indicate the time level in the numerical scheme.)

THEOREM 3.7. *Assume the hypotheses of Theorem* 3.6. *Suppose the initial guess* $\{z_i^{(0)}\}$ *satisfies*

$$\inf_{j\in\mathbb{Z}} u_j^n \le z_i^{(0)} \le \sup_{j\in\mathbb{Z}} u_j^n \tag{3.17}$$

for all $i \in \mathbb{Z}$. Then the nonlinear Jacobi and Gauss–Seidel processes (3.1) *and* (3.2) *are well-defined and converge to the unique bounded solution of* (2.1).

Proof. Again we show only convergence for the Gauss–Seidel process, since the proof for Jacobi is similar. Denote $\underline{u} = \inf_{j\in\mathbb{Z}} u_j^n$ and $\overline{u} = \sup_{j\in\mathbb{Z}} u_j^n$. First, we show that the Gauss–Seidel iterates are well-defined and that $\underline{u} \le z_j^{(k)} \le \overline{u}$ for all j, k. At each step we need to solve

$$r_j(z_j^*) = \frac{1}{\lambda}\left(z_j^* - u_j^n\right) + F(z_j^*, z_{j+1}) - F(z_{j-1}, z_j^*) = 0, \tag{3.18}$$

where $z_{j\pm1} = z_{j\pm1}^{(k)}$ or $z_{j\pm1}^{(k+1)}$ depending on the ordering of the Gauss–Seidel sweep, which by induction must lie between $\underline{u}$ and $\overline{u}$. But

$$\begin{aligned} r_j(\underline{u}) &= \frac{1}{\lambda}\left(\underline{u} - u_j^n\right) + F(\underline{u}, z_{j+1}) - F(z_{j-1}, \underline{u}) \\ &\le 0 + F(\underline{u}, \underline{u}) - F(\underline{u}, \underline{u}) = 0, \end{aligned}$$

where the inequality follows from Assumption 1. Similarly one obtains $r_j(\overline{u}) \ge 0$, so by continuity of F (and hence r_j) there must exist a solution z_j^* to (3.18), which by Lemma 3.5 must be unique. Hence, by induction, the Gauss–Seidel iterates are well-defined and are bounded above and below by $\overline{u}$ and $\underline{u}$, respectively.

Now consider the Gauss–Seidel iterates $\{x_j^{(k)}\}$ and $\{y_j^{(k)}\}$ with initial guess $x_j^{(0)} = \underline{u}$ and $y_j^{(0)} = \overline{u}$ for all j. By Theorem 3.6 these iterates converge pointwise to the same solution $\{x_j^*\}$. We show inductively that $x^{(k)} \le z^{(k)} \le y^{(k)}$ for all k, which would imply that $z_j^{(k)} \to x_j^*$ pointwise. Using the same reordering as in Theorem 3.6, assume that for some $k \ge 0$ and $i \ge 1$ we have

$$y^{(k)} \ge z^{(k)} \ge x^{(k)}, \qquad y_j^{(k+1)} \ge z_j^{(k+1)} \ge x_j^{(k+1)}, \qquad j = 1, \dots, i-1,$$

which is valid for $k = 0$ and $i = 1$. Then by the same boundedness and antitonicity arguments as in Theorem 3.4, we have

$$\begin{aligned} r_i\left(y_1^{(k+1)}, \dots, y_{i-1}^{(k+1)}, y_i^{(k+1)}, y_{i+1}^{(k)}, \dots\right) = 0 = r_i\left(z_1^{(k+1)}, \dots, z_{i-1}^{(k+1)}, z_i^{(k+1)}, z_{i+1}^{(k)}, \dots\right) \\ \ge r_i\left(y_1^{(k+1)}, \dots, y_{i-1}^{(k+1)}, z_i^{(k+1)}, y_{i+1}^{(k)}, \dots\right), \end{aligned}$$

which, together with the strict diagonal isotonicity of r_i, implies that $y_i^{(k+1)} \ge z_i^{(k+1)}$. Similarly it follows that $z_i^{(k+1)} \le x_i^{(k+1)}$. This completes the induction, and hence $z_j^{(k)} \to x_j^*$ pointwise. ∎

In particular, the nonlinear Gauss–Seidel and Jacobi processes converge if we use $\{u_j^n\}$ (i.e., the solution from the previous time step) as an initial guess. For small to moderate time-step sizes, one generally expects the solutions between consecutive time steps to be close to each other, so using $\{u_j^n\}$ often results in much faster convergence than either $\underline{u}$ or $\overline{u}$ as an initial guess.

3.5. Extensions. In this section we show how to extend the results of Theorems 3.6 and 3.7 to deal with conservation laws with nonuniform spatial grids and/or spatially varying flux functions, as well as flux functions that are only defined over a closed interval $I \subset \mathbb{R}$.

3.5.1. Nonuniform grids and spatially varying flux functions. Consider again the fully implicit discretization (2.1):

$$\phi_i(u_i^{n+1} - u_i^n) + \lambda(F_{i+1/2}^{n+1} - F_{i-1/2}^{n+1}) = 0, \qquad \lambda = \Delta t/\Delta x, \;\; i \in \mathbb{Z},$$

where ϕ_i and $F_{i+1/2}$ can vary over the spatial index i. We assume that $0 < \phi_i \leq 1$. Note that the nonuniform grid case is automatically included: For any nonuniform discretization of the form

$$\frac{\tilde{\phi}_i(u_i^{n+1} - u_i^n)}{\Delta t} + \frac{F_{i+1/2}^{n+1} - F_{i-1/2}^{n+1}}{\Delta x_i} = 0, \tag{3.19}$$

we can multiply (3.19) by $\Delta t \Delta x_i/\Delta x_{max}$ to recover the form of (2.1) with

$$\phi_i = \tilde{\phi}_i \Delta x_i/\Delta x_{max}, \qquad \lambda = \Delta t/\Delta x_{max}.$$

To ensure convergence of the Jacobi and Gauss–Seidel processes, we need the following assumptions:

1. The family of flux functions $\{F_{i+1/2}\}_{i=-\infty}^{\infty}$ is equicontinuous (cf. [12]) with the same local Lipschitz constant K_B.
2. $\{\phi_i\}$ is uniformly bounded away from zero; i.e., there exists $\phi_{min} > 0$ such that $\phi_i \geq \phi_{min}$ for all $i \in \mathbb{Z}$.

While the equicontinuity condition may appear severe, it is usually satisfied in practice because the spatially varying coefficients (e.g., $K(x)$ in (1.11)) tend to be uniformly bounded, ensuring equicontinuity in the flux functions. With the above assumptions, we can mimic Theorem 3.6 exactly by replacing λ with λ/ϕ_i. Then the proof goes through verbatim, except for (3.16), which must be modified to

$$\frac{2\lambda K_B}{\phi_{min} + 2\lambda K_B} < \beta < 1. \tag{3.20}$$

3.5.2. Bounded admissible solutions. Formally, Theorem 3.6 requires the discrete flux function $F(u_i, u_{i+1})$ to be defined on $\mathbb{R} \times \mathbb{R}$. In practice one may want to solve problems for which the flux function f is defined only on an interval $[u_{min}, u_{max}]$ rather than on all of $\mathbb{R}$, because states outside these bounds are unphysical. For instance, in the two-phase flow problem, we must have $S_i \in [0, 1]$ for all i, and the flux function $f(S)$ in (1.11) is not even defined outside this range. Fortunately, the estimate (3.14) ensures that as long as the initial conditions are within physical bounds, so will the solution for subsequent time steps $n > 0$, as well as any intervening Jacobi or Gauss–Seidel iterate. Thus, to apply Theorem 3.6 to these problems, one can *formally extend* the domain of the flux function f to $\mathbb{R}$ by defining, for instance,

$$\tilde{f}(u) = \begin{cases} f(u_{min}), & u < u_{min}, \\ f(u), & u_{min} \leq u \leq u_{max}, \\ f(u_{max}), & u > u_{max}, \end{cases}$$

and similarly for the discrete flux $F(u, v)$. Since all Gauss–Seidel iterates $\{y^k\}$ and $\{x^k\}$ satisfy the bound $x^0 \leq x^k \leq y^k \leq y^0$, the exact manner in which the extension is defined is unimportant as long as Assumption 1 (monotonicity) is valid.

4. Applications to porous media flow.

4.1. 1D Buckley–Leverett problem with gravity. Consider a 1D incompressible two-phase flow problem with a constant-rate injection boundary condition on the left, a pressure boundary condition on the right, and no internal sources or sinks:

$$\phi(x)\frac{\partial S_j}{\partial t} + \frac{\partial v_j}{\partial x} = 0, \quad (x,t) \in (x_L, x_R) \times \mathbb{R}^+, \tag{4.1}$$

$$S_1(x,0) = S^0(x), \quad x \in (x_L, x_R), \tag{4.2}$$

$$v_j(x_L,t) = v_{j,L}, \quad p(x_R,t) = p_R, \quad t \in \mathbb{R}^+, \tag{4.3}$$

for $j = 1, 2$, where $v_j = -K(x)\lambda_j(S_1)(\frac{dp}{dx} - \rho_j \mathbf{g}\cdot\mathbf{i})$, $p \equiv p_1 = p_2$ (i.e., zero capillary pressure), and $\mathbf{i}$ is the unit vector along the x-direction. We assume that the injection velocities $v_{1,L}$ and $v_{2,L}$ are nonnegative, and that the total velocity $v_{T,L} := v_{1,L} + v_{2,L}$ is strictly positive. (These assumptions cover the most interesting cases, such as oil recovery by water flooding.) This formulation, which contains pressure variables, is known as the *parabolic form* of the problem, since it represents the incompressible limit of a parabolic problem. We can also derive the *hyperbolic* or "fractional flow" form of the problem by eliminating the pressure variables as follows. The discretized PDEs can be written as

$$\frac{\phi_i(S_{1,i} - S_{1,i}^{\text{old}})}{\Delta t} + \frac{F_{1,i+1/2} - F_{1,i-1/2}}{\Delta x} = 0, \tag{4.4a}$$

$$\frac{\phi_i(S_{1,i}^{\text{old}} - S_{1,i})}{\Delta t} + \frac{F_{2,i+1/2} - F_{2,i-1/2}}{\Delta x} = 0, \tag{4.4b}$$

where

$$F_{j,i+1/2} = K_{i+1/2}\lambda_{j,i+1/2}\left(\frac{p_i - p_{i+1}}{\Delta x} + \rho_j g\right) \tag{4.5}$$

for $j = 1, 2$, $i = 1, \ldots, N$, with $g = \mathbf{g}\cdot\mathbf{i}$. The numerical boundary conditions become

$$F_{j,1/2} = v_{j,L} \quad (j = 1, 2), \qquad p_{N+1} = 2p_R - p_N. \tag{4.6}$$

Assume, without loss of generality, that $g(\rho_1 - \rho_2) \geq 0$. To eliminate the pressure variables p_i, first note that summing (4.4a) and (4.4b) yields, after some manipulation,

$$F_{1,i+1/2} + F_{2,i+1/2} = F_{1,i-1/2} + F_{2,i-1/2} = v_1 + v_2 =: v_T.$$

In other words, the total flux v_T is constant across any interface, which must then be equal to $v_{T,L}$. We can express the pressure gradient $(p_i - p_{i+1})/\Delta x$ in terms of v_T by summing (4.5) through $j = 1, 2$:

$$v_T = K_{i+1/2}\left[\lambda_{T,i+1/2}\frac{p_i - p_{i+1}}{\Delta x} + (\lambda_{1,i+1/2}\,\rho_1 g + \lambda_{2,i+1/2}\,\rho_2 g)\right],$$

where $\lambda_{T,i+1/2} = \lambda_{1,i+1/2} + \lambda_{2,i+1/2}$. Thus,

$$\frac{p_i - p_{i+1}}{\Delta x} = \frac{v_T - K_{i+1/2}(\lambda_{1,i+1/2}\,\rho_1 g + \lambda_{2,i+1/2}\,\rho_2 g)}{K_{i+1/2}(\lambda_{1,i+1/2} + \lambda_{2,i+1/2})}. \tag{4.7}$$

Substituting into (4.5) for $j = 1$ gives

$$F_{1,i+1/2} = \frac{\lambda_{1,i+1/2}}{\lambda_{T,i+1/2}} \left[v_T + K_{i+1/2} \lambda_{2,i+1/2} (\rho_1 - \rho_2) g \right] \tag{4.8}$$
$$= F_{1,i+1/2}(S_{1,i}, S_{1,i+1}).$$

This, together with (4.4a), gives

$$\phi_i (S_{1,i} - S_{1,i}^{\text{old}}) + \frac{\Delta t}{\Delta x} \left(F_{1,i+1/2} - F_{1,i-1/2} \right) = 0, \tag{4.9}$$

leading to a numerical scheme that is identical to (2.1) except for the boundary conditions.Clearly, the treatment of boundary conditions will significantly affect the stability and accuracy of the numerical scheme. However, in order to understand the behavior of the numerical scheme at interior points, we will simply replace the initial-boundary value problem (4.1)–(4.3) with an initial value problem on an infinite domain with appropriate initial conditions. In particular, we replace the injection boundary condition with

$$S_1^0(x) = S_{1,L}, \qquad x < x_L, \tag{4.10}$$

where $S_{1,L}$ satisfies $f_1(x_L, S_{1,L}) = v_{1,L}/v_T$, with $f_1(x, S_1)$ given by (1.10). We also replace the pressure boundary condition with

$$S_1^0(x) = S_1^0(x_R), \qquad x > x_R. \tag{4.11}$$

The modified continuous problem will yield a solution identical to (4.1)–(4.3) for $0 < t < T_{BT}$, where T_{BT} is the breakthrough time (i.e., the time at which the shock front arrives at the pressure boundary). Note that f_1 is one-to-one over the interval $I = \{S : 0 \le f_1(S) < 1\}$ (see Figure 1.1), and $v_{1,L} \le v_T$ by assumption; thus, (4.10) is well-defined unless $v_{2,L} = 0$. (If $v_{2,L} = 0$, we define $S_1^0(x) = \inf f_1^{-1}(v_T)$, where f_1^{-1} denotes the inverse image.) For the remainder of this section, we will drop the phase subscript and denote $S_{1,i}$ and $F_{1,i+1/2}$ by S_i and $F_{i+1/2}$, respectively.

Phase-based upstreaming. Recall from section 1 (cf. (1.9)) that the mobilities $\lambda_{j,i+1/2}$ are evaluated using the upstream saturations with respect to the flow direction of phase j:

$$\lambda_{j,i+1/2} = \begin{cases} \lambda_j(S_i) & \text{if } \frac{1}{\Delta x}(p_i - p_{i+1}) + \rho_j g \ge 0, \\ \lambda_j(S_{i+1}) & \text{otherwise.} \end{cases} \tag{4.12}$$

In light of (4.7), we can rewrite the upstream conditions as

$$\lambda_{j,i+1/2} = \begin{cases} \lambda_j(S_i) & \text{if } v_T + K_{i+1/2} \lambda_{j',i+1/2} (\rho_j - \rho_{j'}) g \ge 0, \\ \lambda_j(S_{i+1}) & \text{otherwise,} \end{cases} \tag{4.13}$$

where the subscript $j' := 3 - j$ denotes the phase other than phase j. Even though pressure dependence has been eliminated, (4.13) still does not explicitly define the upstream direction for λ_j, since the latter is defined in terms of the (yet undetermined) mobility of the other phase $\lambda_{j',i+1/2}$. For explicit numerical schemes, Brenier and Jaffré have shown in [3] how to explicitly determine the upstream direction for each

phase for a given saturation profile $\{S_i^n\}$. In the special case of two-phase flow, they define the following quantities:

$$\theta_{1,i+1/2} = v_T + K_{i+1/2} g(\rho_1 - \rho_2)\lambda_2(S_{i+1}^n),$$
$$\theta_{2,i+1/2} = v_T - K_{i+1/2} g(\rho_1 - \rho_2)\lambda_1(S_i^n).$$

These quantities correspond precisely to the condition in (4.13), but the condition is evaluated at S_{i+1}^n for θ_1 and S_i^n for θ_2. Clearly $\theta_{1,i+1/2} > 0$, since $g(\rho_1 - \rho_2) \geq 0$. The correct upstream directions are then given by

$$\lambda_{1,i+1/2}^n = \lambda_1(S_i^n), \qquad \lambda_{2,i+1/2}^n = \lambda_2(S_i^n) \qquad \text{if } 0 \leq \theta_{2,i+1/2} \leq \theta_{1,i+1/2},$$
$$\lambda_{1,i+1/2}^n = \lambda_1(S_i^n), \qquad \lambda_{2,i+1/2}^n = \lambda_2(S_{i+1}^n) \qquad \text{if } \theta_{2,i+1/2} \leq 0 \leq \theta_{1,i+1/2}.$$

Thus, the numerical fluxes are well-defined if one uses an explicit method with respect to saturation. However, this analysis does not address implicit-saturation schemes (such as FIM), which require the upstream directions to be consistent with saturation values *at the end of the time step*, i.e., with the saturation profile $\{S_i^{n+1}\}$. Because of this consistency requirement, it is not clear a priori that a solution to the parabolic form of the problem (4.4) even exists. Our approach to proving existence is to rely on the hyperbolic form of the problem (4.8)–(4.11). From the above derivation, it is evident that if $\{(S_i, p_i)\}_{i=1}^N$ is any solution to the parabolic form (4.4)–(4.6), then $\{S_i\}_{i=1}^N$ must be a solution to the hyperbolic problem. Thus, the key idea is to first find the correct saturation profile $\{S_i\}$ using (4.8)–(4.11) with a numerical flux that automatically ensures consistency with the upstream directions; once the $\{S_i\}$ are known, we can easily solve for the pressure part because the pressure equation is linear. We distinguish two cases:

1. If $K_{i+1/2} g(\rho_1 - \rho_2)\lambda_{1,\max} \leq v_T$, then $\theta_{2,i+1/2} \geq 0$ always, so we revert to a single-point upstream scheme $F_{i+1/2} = F_{i+1/2}(S_i)$.

2. If $K_{i+1/2} g(\rho_1 - \rho_2)\lambda_{1,\max} > v_T$, then by the monotonicity of $\lambda_1(S)$, there exists a unique $0 < S_c < 1$ such that $K_{i+1/2} g(\rho_1 - \rho_2)\lambda_1(S_c) = v_T$. Then the numerical flux, which is to be evaluated at time t^{n+1}, is defined as

(4.14)
$$F_{i+1/2}(S_i, S_{i+1}) = \begin{cases} \dfrac{\lambda_1(S_i)\left[v_T + K_{i+1/2} g(\rho_1 - \rho_2)\lambda_2(S_i)\right]}{\lambda_1(S_i) + \lambda_2(S_i)} & \text{if } 0 \leq S_i \leq S_c, \\ \dfrac{\lambda_1(S_i)\left[v_T + K_{i+1/2} g(\rho_1 - \rho_2)\lambda_2(S_{i+1})\right]}{\lambda_1(S_i) + \lambda_2(S_{i+1})} & \text{if } S_c < S_i \leq 1. \end{cases}$$

Figure 4.1 shows a plot of $F(S_i, S_{i+1})$ in the latter case, which corresponds to the continuous flux function in Figure 1.1(b). Note that we are able to recover $f(S)$ from the numerical flux $F(S_i, S_{i+1})$ because consistency implies $F(S, S) = f(S)$. Even though $f(S)$ itself is nonmonotonic, the plot clearly shows that $F(S_i, S_{i+1})$ is an increasing function of S_i and a decreasing function of S_{i+1}, as predicted by Theorem 2.1. Also, the numerical flux is independent of the downstream saturation S_{i+1} inside the cocurrent region ($0 \leq S_i \leq S_c \approx 0.27$) but becomes a function of both variables when $S_i > S_c$. Finally, $F(S_i, S_{i+1})$ is Lipschitz continuous, but nondifferentiable, along the line $S_i = S_c$ due to the upstream condition (4.14). Since the numerical flux satisfies the monotonicity assumption, the analysis in the last two

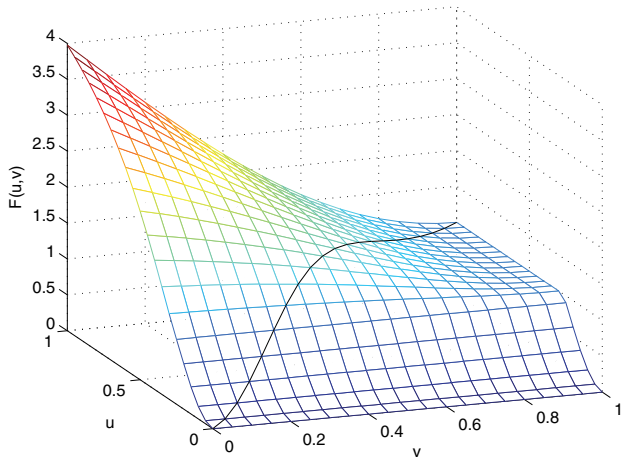


FIG. 4.1. *The numerical flux function $F(u,v)$ corresponding to the fractional flow in Figure* 1.1(b). *The black curve along the diagonal indicates the value of $F(u,u) = f(u)$.*

sections shows that the hyperbolic problem with implicit time-stepping has a unique solution $\{S_i^{n+1}\}$, which must also be the correct saturation profile for the parabolic problem. To solve for pressure, we use (4.7) and (4.6):

$$\frac{p_i - p_{i+1}}{\Delta x} = \frac{v_T - K_{i+1/2}(\lambda_{1,i+1/2}\, g\rho_1 + \lambda_{2,i+1/2}\, g\rho_2)}{K_{i+1/2}(\lambda_{1,i+1/2} + \lambda_{2,i+1/2})}, \qquad i = 1, \ldots, N, \tag{4.15}$$

$$p_{N+1} = 2p_R - p_N. \tag{4.16}$$

Since $\{S_i^{n+1}\}$ is now known, the right-hand side of (4.7) is also completely determined. Thus, the vector p of pressures satisfies $Ap = b$, where A is an $N \times N$ upper triangular matrix with a nonzero diagonal. So A is nonsingular, which means there is a unique pressure profile $\{p_i^{n+1}\}$ that satisfies (4.7) and (4.6). It is easy to see that this pressure profile is consistent with the upstream condition (4.12): because of (4.7), this upstream condition is equivalent to (4.13), and the conditions therein are precisely the ones we use to define the numerical flux function (4.14) for the hyperbolic problem. Hence, we have shown that the parabolic form (4.4)–(4.6) has a unique solution, given by the above $\{(S_i^{n+1}, p_i^{n+1})\}$.

4.2. Multidimensional considerations. In multiple dimensions, one can no longer eliminate pressure variables as shown above, because the total velocity $\mathbf{v}_T$ is generally a function of space and time. Thus, the system of PDEs (1.1), (1.2) does not reduce to a purely hyperbolic problem, which means we cannot directly apply our existence and uniqueness results to the FIM in this case. Nonetheless, our analysis does apply to a related numerical scheme known as the *sequential-implicit method* (SEQ). At each time step in SEQ, we first solve the discrete version of the (linear) elliptic equation (1.5), in which the saturation-dependent coefficients are taken at time t^n. In other words, we solve for p^{n+1} via

$$-\nabla \cdot \left[K\lambda_T(S^n)\nabla p^{n+1} - K\mathbf{g}\textstyle\sum_j \rho_j \lambda_j(S^n)\right] = \sum_j q_j. \tag{4.17}$$

Next, we compute the total velocity

$$\mathbf{v}_T^* = \textstyle\sum_j \mathbf{v}_j^* = -\sum_j K\lambda_j(S^n)(\nabla p^{n+1} - \rho_j \mathbf{g}). \tag{4.18}$$

Finally, we compute the saturations S_j^{n+1} $(j = 1, \ldots, n-1)$ by solving the discrete version of (1.6) and (1.7) with *implicit time-stepping*:

$$\phi \frac{\partial S_j}{\partial t} + \nabla \cdot \mathbf{v}_j(x, S_1, \ldots, S_n) = 0, \tag{4.19}$$

$$\mathbf{v}_j = \frac{\lambda_j}{\lambda_T}\left(\mathbf{v}_T^* - K\mathbf{g}\textstyle\sum_\ell \lambda_\ell(\rho_\ell - \rho_j)\right). \tag{4.20}$$

Essentially, the SEQ method decouples the system into an elliptic and a hyperbolic subproblem. For two-phase flow problems, one can readily extend the convergence results of Theorems 3.6 and 3.7 to the hyperbolic subproblem above, as long as the spatial grid satisfies certain shape and connectivity requirements. Discretizing (4.19) yields the multidimensional analogue of (2.1):

$$\phi_i(S_i^{n+1} - S_i^n) + \sum_{l \in \mathrm{adj}(i)} \lambda_{il} F_{il}(S_i^{n+1}, S_l^{n+1}) = 0. \tag{4.21}$$

Here, F_{il} is the flux (or velocity) from cell i to cell l, and $\lambda_{il} = \Delta t|\partial V_{il}|/|V_i|$, where $|\partial V_{il}|$ is the area of the surface separating cells i and l, $|V_i|$ is the volume of cell i, and Δt is the time step. For a conservative scheme we must have $F_{il}(S_i, S_l) = -F_{li}(S_l, S_i)$, and for monotonicity we require that F_{il} be nondecreasing with respect to the first argument and nonincreasing with respect to the second. This requirement is satisfied for two-phase flow problems, since we can reproduce the derivation in section 4.1 to obtain the flux function

$$F_{il} = \frac{\lambda_{1,il}}{\lambda_{T,il}}\left[v_{il} + K_{il}g_{il}(\rho_1 - \rho_2)\lambda_{2,il}\right]$$

and the upstream condition

$$\lambda_{j,il} = \begin{cases} \lambda_j(S_i) & \text{if } v_{il} + K_{il}g_{il}(\rho_j - \rho_{j'})\lambda_{j',il} \geq 0, \\ \lambda_j(S_l) & \text{otherwise,} \end{cases}$$

for $j = 1, 2$, where $v_{il} = \mathbf{v}_T^* \cdot \mathbf{n}_{il}$ and $g_{il} = \mathbf{g} \cdot \mathbf{n}_{il}$. In order to mimic the proof of Theorem 3.6, we need the following assumptions on the grid:

1. The number of cells (control volumes) adjacent to cell i is bounded for all i.
2. The ratio $|\partial V_{il}|/|V_i|$ is bounded for all pairs of adjacent cells (i, l).
3. The quantity $\phi_i|V_i|$ is uniformly bounded away from zero for all i.
4. For any cell i, the total number of cells reachable from i in k steps is $O(k^p)$ for some fixed $p > 0$ (i.e., grows at most polynomially in k).

The above assumptions are easily satisfied by regular Cartesian grids, and also by most unstructured grids of practical interest. We also need the following assumption on the numerical flux F_{il} (which is analogous to the assumption in section 3.5.1):

5. F_{il} is equicontinuous with the same Lipschitz constant for all pairs of adjacent cells (i, l).

These assumptions ensure that the residual functions are all bounded and have the same Lipschitz constant over the set $\{u \in \ell^\infty(N) \,|\, \|u\|_\infty < B\}$. The polynomial

growth assumption (4) allows us to assign the weights $\{w_i^B\}$ to each cell i in the following manner: pick any node i_0 and let $w_i^B = \beta^{d(i_0,i)}$, where $d(i,j)$ is the shortest distance between nodes i and j in the graph-theoretic sense. Since the number of cells within k steps of i_0 grows polynomially in k, the series $\sum_i w_i^B$ converges for any $0 < \beta < 1$, so β can be chosen the same way as in step 3 of the proof of Theorem 3.6 and the same argument will hold. Hence, we can conclude that the hyperbolic subproblem in the SEQ method is well-defined for any time-step size Δt, and the nonlinear Gauss–Seidel and Jacobi processes are guaranteed to converge for these problems.

5. Accuracy of phase-based upstreamed solutions. In this section, we investigate the accuracy of the numerical solution obtained from implicit phase-based upstreaming when we vary the spatial and temporal grid. Our 1D test case consists of a countercurrent flow problem inside the domain $\Omega = [0,1]$, and our two-dimensional (2D) example is a cocurrent flow problem in a heterogeneous reservoir. For the 1D problem, water is injected at the boundary $x = 0$ and a pressure boundary condition is maintained at $x = 1$. The hyperbolic form of the problem is described by (1.10), (1.11). The flux function $f_1(S)$, which is independent of x, is shown in Figure 1.1(b), with a sonic point at $S = 0.49$; countercurrent flow occurs whenever $S \geq S_c \approx 0.27$. The initial saturation profile is a step function with

$$S^0(x) = \begin{cases} 1, & 0 \leq x < 0.2, \\ 0, & 0.2 < x \leq 1. \end{cases}$$

The numerical solution is compared with the analytical solution at time $t = 0.15$. Because of the sonic point, the solution contains two shocks connected by a rarefaction; one shock moves to the right with a velocity of 3.9, and the other travels to the left with a velocity of -1.2. When considering the accuracy of a numerical solution, two error measures are shown:

- The L^1-*error*, which is the difference between the numerical and the analytical solution in the L^1-norm.
- The *front dispersion*, which is the distance between analytical shock front and the leftmost point for which the numerical solution becomes zero.

We also measure how difficult the nonlinear problem is by showing, for each test case, the average number of nonlinear Gauss–Seidel iterations required for convergence at each time step. We remark that this measure is only useful for problems with countercurrent flow; in the cocurrent case, the flux function $F_{i+1/2}$ is a function of S_i only, which means Gauss–Seidel will always converge in one iteration if we solve the single-cell equations in the order $1, 2, \ldots, N$ (i.e., from upstream to downstream). In the countercurrent case, convergence is generally linear, and the rate of convergence is a function of the time step [8].

5.1. Refinement under fixed mesh ratio. Here we refine the grid at a fixed mesh ratio $\Delta x/\Delta t$ to maintain a fixed CFL number of 4.10, which is above the CFL limit for explicit schemes. Figure 5.1 shows the plots for $N = 50, 100, 200, 400$, and Table 5.1 shows the L^1-error and front dispersion data. The plots show that the numerical solution converges to the analytical solution even though the CFL number is greater than 1, which confirms our analysis. Moreover, both the L^1-error and the front dispersion are decreasing to zero a bit worse than linearly, with a ratio of about 0.61 and about 0.58, respectively, for every refinement by a factor of two. Also note the poor resolution near the left boundary for $N = 50, 100$, where instead of approaching

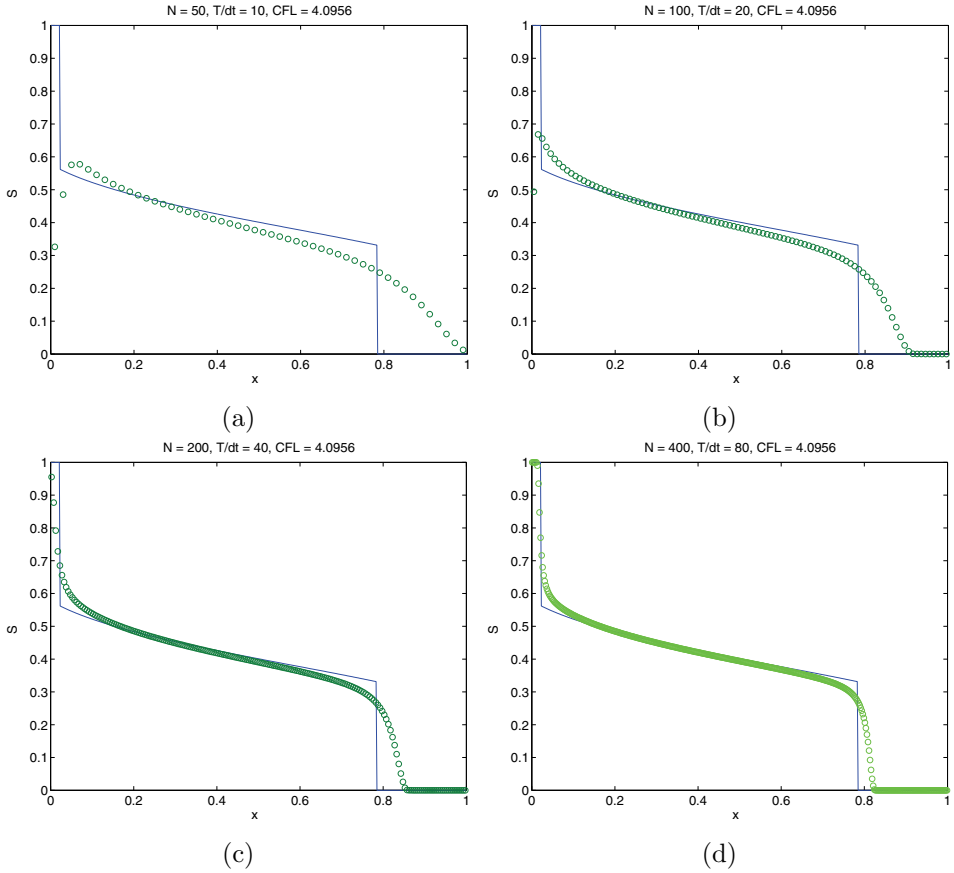


FIG. 5.1. *Numerical solution at different resolutions, CFL* $= 4.10$, $t = 0.15$.

$S = 1$, the solution is closer to $S_c \approx 0.27$ at the left boundary. For these coarser grids, the numerical solution cannot decide whether the left-moving wave has reached the boundary, which is maintained at $S(0,t) = S_c$ (see (4.10)). For higher resolutions ($N = 200, 400$), the artifact has disappeared and the numerical solution reproduces the back end of the saturation profile quite accurately. The average number of Gauss–Seidel iterations required for convergence are all similar, so refining the grid at a fixed mesh ratio does not increase the difficulty of the problem for the nonlinear solver.

5.2. Spatial refinement for fixed time steps. Here, we refine the spatial grid only while fixing the time-step size. Figure 5.2 and Table 5.2 show the results for $N = 25, 50, 100, 200, 400$ and a time-step size of $\Delta t = 0.0075$; i.e., we use 20 time steps to integrate up to $t = 0.15$. We see that even though the $N = 25$ case has a

TABLE 5.1
Accuracy of numerical solutions for a fixed CFL number.

N	$t/\Delta t$	CFL	L^1-error	Front dispersion	Average # iterations
50	10	4.10	0.0665	> 0.215	4.9
100	20	4.10	0.0444	0.116	4.4
200	40	4.10	0.0273	0.066	4.2
400	80	4.10	0.0168	0.039	4.1

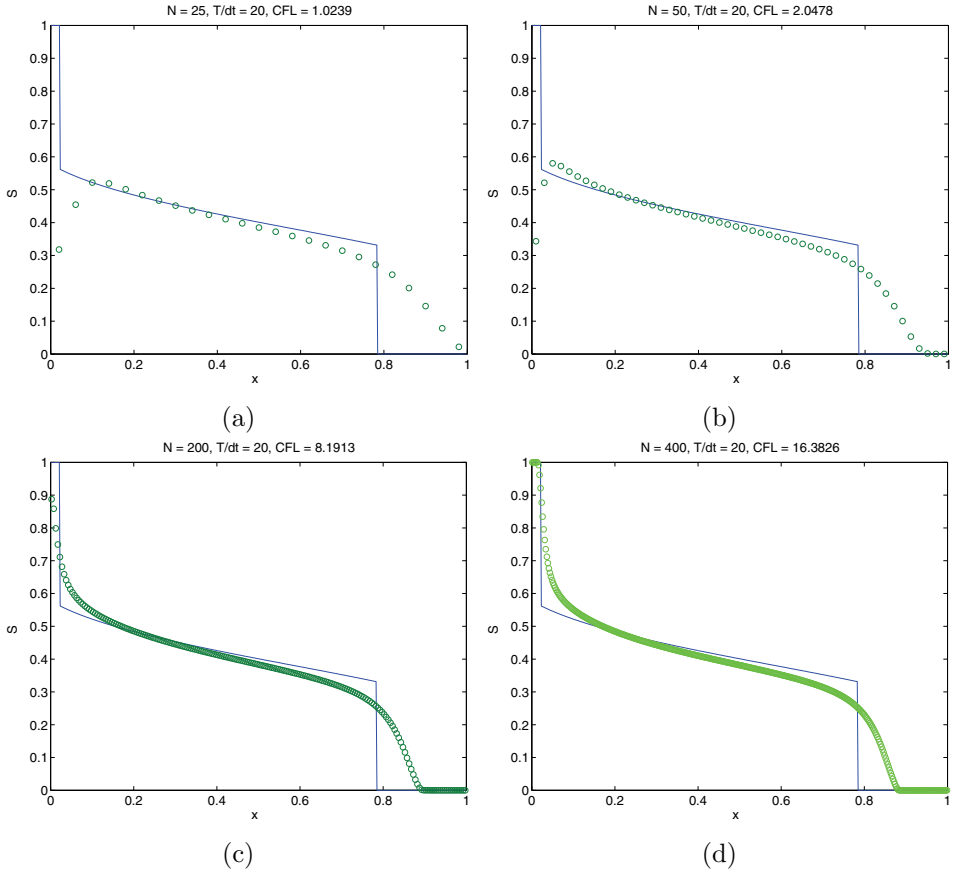


FIG. 5.2. *Numerical solution for different spatial grids,* 20 *time steps,* $t = 0.15$. *The* $N = 100$ *case is identical to Figure* 5.1(b) *and hence omitted.*

CFL number close to 1, the grid is clearly too coarse, and the shock fronts are very poorly resolved. The accuracy increases substantially when the spatial grid is refined to $N = 50, 100$, even though the CFL number becomes progressively larger; thus, the CFL number by itself is not a good measure of solution quality. However, the improvement due to spatial grid refinement becomes negligible for $N > 100$, since errors due to time discretization are now the dominant source of error. In addition, the average number of iterations required to attain convergence increases with each refinement: as we refine the grid, we are solving increasingly difficult problems, even though the improvement in solution accuracy will stagnate beyond a certain point. Thus, even though the FIM can tolerate arbitrarily large CFL numbers, one should not expect the solution accuracy to improve indefinitely simply by using a finer spatial grid, without making a corresponding reduction in time-step size.

5.3. Nonuniform grids. The real advantage of the FIM over an explicit scheme lies in its efficiency when applied to a heterogeneous problem, where the porosity $\phi(x)$ and permeability $K(x)$ can vary by orders of magnitude over the domain. In these problems, the CFL condition is determined by the minimum porosity in the domain, which can be much smaller than the average porosity. To illustrate this point, we show an example in which the spatial grid is nonuniform (which, from section 3.5.1,

is equivalent to the spatially varying porosity case). The nonuniform grid contains 50 gridblocks, with $\Delta x_{\max}/\Delta x_{\min} = 96$. Table 5.3 and Figure 5.3 compare the numerical solutions obtained from this grid to the uniform-grid solutions. We see that the solutions are qualitatively (from the plots) and quantitatively (from the L^1-error and front dispersion) not very different, even though the CFL number is 50 times larger in the nonuniform case. Thus, an explicit integrator would have to take unacceptably small time steps, whereas an implicit method allows time steps that are much more reasonable. In addition, the average number of iterations required for convergence is roughly the same for both cases, so the equations resulting from a nonuniform grid are not harder to solve, despite the large CFL numbers.

5.4. 2D example. The goal of this example is to compare the unconditionally stable SEQ method with the implicit-pressure-explicit-saturation (IMPES) method, which carries a time-step restriction $\Delta t = \mathcal{O}(\Delta x)$. We use a 110×30 grid to discretize a 2D horizontal reservoir (i.e., no gravity). The porosity ϕ is constant throughout the reservoir, whereas the permeability field $K(\mathbf{x})$ is taken from the SPE 10 test set [4] and ranges from a minimum of 0.0052 to a maximum of 1219 (see Figure 5.4(b)). The reservoir is initially saturated with oil. Starting from $t = 0$, water is injected at a constant rate into cell $(1, 1)$, and fluid is produced at constant pressure from cell $(110, 30)$. We simulate the reservoir until $t = 30$ (0.182 pore volumes injected), which is roughly when breakthrough occurs at the outlet boundary. For IMPES, we take the largest time step allowed by the CFL criterion, whereas for SEQ we use two time-stepping strategies:

- Small Δt: $t = 1, 2, 3, 4$; after $t = 4$, $\Delta t = 2$ until $t = 30$.
- Large Δt: $t = 1, 3, 6, 10$; after $t = 10$, $\Delta t = 5$ until $t = 30$.

Figures 5.4(a), (c), and (e) show the saturation profiles obtained from IMPES and SEQ for the two time-step sizes, and Table 5.4 shows the error and running times for each case. Note that all three solutions are very similar; the largest discrepancies occur near flood fronts, where the SEQ solutions are noticeably more diffuse than the IMPES solution. However, the IMPES profile is sharp only because the method is forced to take tiny time steps to satisfy a very severe CFL condition. As a result, IMPES runs 65 times slower than SEQ for the large Δt case, and 31 times slower for the small Δt case. Table 5.4 indicates that the maximum difference between the

TABLE 5.2
Accuracy of numerical solutions for a fixed time step size.

N	$t/\Delta t$	CFL	L^1-error	Front dispersion	Average # iterations
25	20	1.02	0.0673	> 0.215	2.6
50	20	2.05	0.0529	0.156	3.3
100	20	4.10	0.0444	0.116	4.4
200	20	8.20	0.0378	0.101	6.4
400	20	16.40	0.0366	0.094	9.2

TABLE 5.3
Accuracy of numerical solutions for a nonuniform grid.

	N	$t_D/\Delta t$	CFL	L^1-error	Front disp.	Avg. # iters.
Nonuniform	50	20	105.80	0.0566	0.180	3.2
	50	50	42.30	0.0475	0.132	2.2
Uniform	50	20	2.05	0.0529	0.156	3.3
	50	50	0.82	0.0435	0.116	2.1

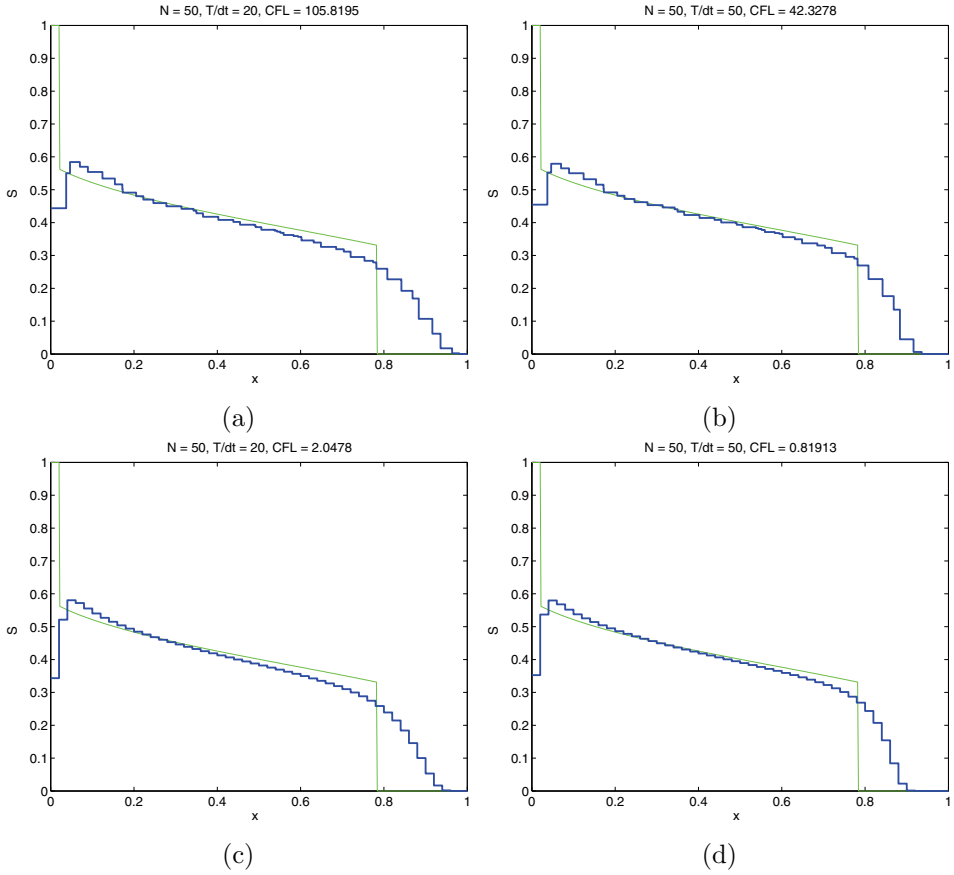
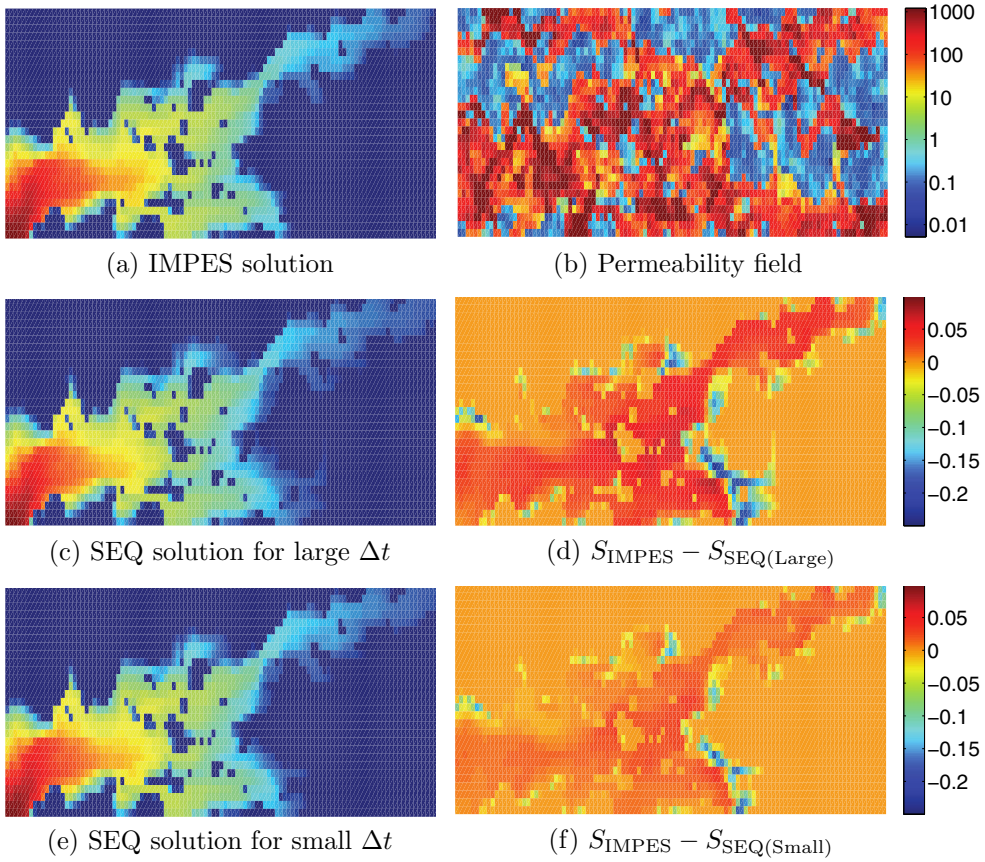


FIG. 5.3. *Numerical solutions obtained from a nonuniform grid ((*a*) and* (b)*) and their uniform-grid counterparts ((*c*) and* (d)*),* $t = 0.15$.

TABLE 5.4
Performance of different numerical methods on the 2D example.

	IMPES	SEQ (large Δt)	SEQ (small Δt)
No. of time steps	1350	8	17
Maximum Δt	0.0222	5	2
Maximum CFL	1	225	90
Breakthrough time (days)	29.1	24.6	27.7
$\|S_{\text{IMPES}} - S_{\text{SEQ}}\|_{L^1(\Omega)}$	–	0.0188	0.0091
$\|S_{\text{IMPES}} - S_{\text{SEQ}}\|_{L^\infty(\Omega)}$	–	0.2348	0.2004
Solution time (sec)	66017	1016	2130

IMPES and SEQ solutions remains nearly constant when the time step is reduced. This is not surprising, because the numerical solution cannot converge uniformly when discontinuities are present in the solution. However, we do observe a decrease in the L^1-error, as well as a later breakthrough time, when Δt is reduced. This is consistent with our 1D results, where refining the grid reduces the L^1 and front dispersion errors, but the L^∞-error does not go to zero because of smearing across the shock front. This example shows that, for problems of practical interest, an implicit method can produce solutions of comparable quality at much lower computational costs than an explicit



(a) IMPES solution (b) Permeability field

(c) SEQ solution for large Δt (d) $S_{\text{IMPES}} - S_{\text{SEQ(Large)}}$

(e) SEQ solution for small Δt (f) $S_{\text{IMPES}} - S_{\text{SEQ(Small)}}$

FIG. 5.4. *Permeability of the 2D reservoir, as well as saturation profiles produced by the different numerical methods.*

method.

6. Conclusion. We have shown that, for any residual function that arises from the implicit monotone discretization of a scalar hyperbolic conservation law, the nonlinear Gauss–Seidel and Jacobi processes converge to the unique bounded solution whenever the initial guess is bounded. This provides an alternate, constructive proof of the well-definedness of monotone implicit schemes, for which a solution algorithm is easily implementable. Convergence to the entropy solution for arbitrary CFL numbers follows immediately from the properties of the flux functions. These results are applicable to the fully coupled two-phase flow problem in one dimension, and to the hyperbolic subproblem in the sequential-implicit method in higher dimensions. Finally, we studied the accuracy of phase-based upstream solutions under different grid refinement schemes, and the importance of unconditional stability became evident when a nonuniform grid and/or a variable porosity field is used.

REFERENCES

[1] H. ABELSON, J. SUSSMAN, AND J. SUSSMAN, *Structure and Interpretation of Computer Programs*, MIT Press, Cambridge, MA, 1984.

[2] K. AZIZ AND A. SETTARI, *Petroleum Reservoir Simulation*, Applied Science Publishers, New

York, 1979.
[3] Y. Brenier and J. Jaffré, *Upstream differencing for multiphase flow in reservoir simulation*, SIAM J. Numer. Anal., 28 (1991), pp. 685–696.
[4] M. A. Christie and M. J. Blunt, *Tenth SPE comparative solution project: A comparison of upscaling techniques*, SPE Reser. Eval. Eng., 4 (2001), pp. 308–317.
[5] M. G. Crandall and T. M. Liggett, *Generation of semi-groups of nonlinear transformations on general Banach spaces*, Amer. J. Math., 93 (1971), pp. 265–298.
[6] K. Deimling, *Nonlinear Functional Analysis*, Springer-Verlag, Berlin, 1985.
[7] S. Evje and K. H. Karlsen, *Degenerate convection-diffusion equations and implicit monotone difference schemes*, in Hyperbolic Problems: Theory, Numerics, Applications, Internat. Ser. Numer. Math. 129, M. Fey and R. Jeltsch, eds., Birkhäuser Verlag, Basel, 1999, pp. 285–294.
[8] F. Kwok, *Scalable Linear and Nonlinear Algorithms for Multiphase Flow in Porous Media*, Ph.D. thesis, Stanford University, Stanford, CA, 2007.
[9] B. J. Lucier, *On nonlocal monotone difference schemes for scalar conservation laws*, Math. Comp., 47 (1986), pp. 19–36.
[10] S. Osher, *Riemann solvers, the entropy condition, and difference approximations*, SIAM J. Numer. Anal., 21 (1984), pp. 217–235.
[11] W. C. Rheinboldt, *On M-functions and their application to nonlinear Gauss-Seidel iterations and to network flows*, J. Math. Anal. Appl., 32 (1970), pp. 274–307.
[12] H. L. Royden, *Real Analysis*, Prentice–Hall, Englewood Cliffs, NJ, 1988.
[13] Y. Saad, *Iterative Methods for Sparse Linear Systems*, 2nd ed., SIAM, Philadelphia, 2003.
[14] R. Sanders, *On convergence of monotone finite difference schemes with variable spatial differencing*, Math. Comp., 40 (1983), pp. 91–106.

SIAM J. NUMER. ANAL.
Vol. 46, No. 5, pp. 2688–2714
© 2008 Society for Industrial and Applied Mathematics

THE l^1-ERROR ESTIMATES FOR A HAMILTONIAN-PRESERVING SCHEME FOR THE LIOUVILLE EQUATION WITH PIECEWISE CONSTANT POTENTIALS*

XIN WEN† AND SHI JIN‡

Abstract. We study the l^1-error of a Hamiltonian-preserving scheme, developed in [S. Jin and X. Wen, *Commun. Math. Sci.*, 3 (2005), pp. 285–315], for the Liouville equation with a piecewise constant potential in one space dimension. This problem has important applications in computations of the semiclassical limit of the linear Schrödinger equation through barriers, and of the high frequency waves through interfaces. We use the l^1-error estimates established in [X. Wen and S. Jin, *J. Comput. Math.*, 26 (2008), pp. 1–22; X. Wen, *Convergence of an Immersed Interface Upwind Scheme for Linear Advection Equations with Piecewise Constant Coefficients* II: *Some Related Binomial Coefficient Inequalities*, preprint, 2008] for the immersed interface upwind scheme to the linear advection equations with piecewise constant coefficients. We prove that the scheme with the Dirichlet incoming boundary conditions is l^1-convergent for a class of bounded initial data and derive the half order l^1-error bounds with *explicit* coefficients. The initial conditions can be satisfied by applying the decomposition technique proposed in [S. Jin, H. L. Liu, S. Osher, and R. Tsai, *J. Comput. Phys.*, 205 (2005), pp. 222–241] for solving the Liouville equation with measure-valued initial data, which arises in the semiclassical limit of the linear Schrödinger equation.

Key words. Liouville equations, Hamiltonian preserving schemes, piecewise constant potentials, error estimate, half order error bound, semiclassical limit

AMS subject classifications. 65M06, 65M12, 65M25, 35L45, 70H99

DOI. 10.1137/070700681

1. Introduction.

In [11], we constructed a class of numerical schemes for the d-dimensional Liouville equation in classical mechanics:

$$f_t + \mathbf{v}\cdot\nabla_{\mathbf{x}} f - \nabla_{\mathbf{x}} V\cdot\nabla_{\mathbf{v}} f = 0\,, \quad t>0, \quad \mathbf{x},\mathbf{v}\in R^d\,, \tag{1.1}$$

where $f(t,\mathbf{x},\mathbf{v})$ is the density distribution of a classical particle at position $\mathbf{x}$ and time t and traveling with velocity $\mathbf{v}$. $V(\mathbf{x})$ is the potential. The main interest is in the case of a discontinuous potential $V(\mathbf{x})$, corresponding to a potential barrier. When V is discontinuous, the Liouville equation (1.1) is a linear hyperbolic equation with a measure-valued coefficient. Such a problem cannot be understood mathematically using the renormalized solution by DiPerna and Lions for linear advection equations with discontinuous coefficients [5] (see also [2]). Our approach in [11, 12] to such problems was to provide an interface condition to couple the Liouville equation (1.1) on both sides of the barrier or interface. The interface condition accounts for particle or wave transmission and reflection. Based on this notion of the solution, the so-called Hamiltonian-preserving schemes were constructed in [11, 12], which build this interface condition into the numerical flux. Schemes so constructed provide solutions that are physically relevant for particle or wave reflection and transmission through the barriers or interfaces.

*Received by the editors August 21, 2007; accepted for publication (in revised form) March 21, 2008; published electronically July 2, 2008. This research was supported in part by the Knowledge Innovation Project of the Chinese Academy of Sciences grants K5501312S1, K5502212F1, K7290312G7, and K7502712F7, NSFC grant 10601062, NSF grant DMS-0608720, and NSAF grant 10676017.
http://www.siam.org/journals/sinum/46-5/70068.html

†LSEC, Institute of Computational Mathematics, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, P.O. Box 2719, Beijing 100190, China (wenxin@amss.ac.cn).

‡Department of Mathematics, University of Wisconsin, Madison, WI 53706 (jin@math.wisc.edu).

The Liouville equation is the phase space representation of Newton's second law:

$$\frac{d\mathbf{x}}{dt} = \mathbf{v}\,, \quad \frac{d\mathbf{v}}{dt} = -\nabla_{\mathbf{x}} V\,,$$

which is a Hamiltonian system with the Hamiltonian

$$H = \frac{1}{2}|\mathbf{v}|^2 + V(\mathbf{x}).$$

It is known from classical mechanics that the Hamiltonian remains constant across a potential barrier. This is one of the main ingredients in the Hamiltonian-preserving schemes developed in [11, 12]. The two schemes developed in [11]—one based on a finite difference formulation (called Scheme I) and the other on a finite volume formulation (called Scheme II)—were proved, in one space dimension with a piecewise constant potential, to be positive and l^1- and l^∞-stable for suitable initial value problems and under a hyperbolic CFL condition (see also [29]).

The more sophisticated issue of the l^1-stability of Scheme I was studied in [29]. We proved in [29] that, in the case of a step function potential, Scheme I with the homogeneous Dirichlet incoming boundary conditions is l^1-stable under a certain condition on the initial data. In particular, we showed that such an initial condition is satisfied when choosing the numerical initial data as the cell averages of bounded analytical initial data. Since the exact solution, in the aforementioned notion using the interface condition, to the same problem is l^1-contracting, an l^1-convergent scheme should be l^1-stable. Thus the results established in [29] imply that Scheme I meets the necessary condition for the l^1-convergence in the case of the homogeneous Dirichlet boundary conditions and bounded initial data.

In this paper we will study the l^1-convergence of Scheme I in the case of a step function potential and the Dirichlet incoming boundary conditions. We will prove that in this case Scheme I is l^1-convergent for a class of bounded initial data. The initial conditions can be satisfied when applying the decomposition technique proposed in [10] for solving the Liouville equation with measure-valued initial data arising in the semiclassical limit of the linear Schrödinger equation.

The Liouville equation with a step function (or more generally piecewise constant) potential belongs to hyperbolic equations with singular (discontinuous or measure-valued) coefficients. For the discontinuous coefficient case, the convergence of numerical schemes was widely studied [6, 9, 14, 15, 4, 17, 25, 19, 21, 7, 26, 27, 16, 13, 1, 20]. However, in most cases the convergence rate estimates for numerical schemes were not studied. For the measure-valued coefficient case, recently we have obtained the half order l^1-error estimates for the immersed interface upwind scheme which builds the interface condition into the numerical flux; see [30, 28]. In this paper, we extend this result to the more interesting case of the Liouville equation with a step function potential. Since the Liouville equation is defined in the phase space, extra efforts are needed to deal with the discretization in the velocity space.

To derive the desired error estimates, we split the computational domain into several parts. In each subdomain, we will introduce a number of linear advection equations with step function coefficients. The immersed interface upwind schemes for these linear advection equations yield the same numerical solution as Scheme I for the Liouville equation with the step function potential. Then the l^1-error estimates for Scheme I can be achieved by applying the l^1-error estimates for the immersed interface upwind schemes and estimating the l^1-errors between the exact solutions of the linear

advection equations with step function coefficients and that of the Liouville equation with the step function potential. The first part of each error estimate is obtained by applying the results established in [30, 28] with the aid of the condition on the initial data. The second part of the error is analyzed also utilizing the condition on the initial data. With this approach, we can derive the half order l^1-error bound with *explicit* coefficients for Scheme I.

Our proof is based on the step function potential. This result clearly can be generated to piecewise constant potentials in a straightforward but tedious way. The same approach should also be applicable to analyzing the l^1-error estimates for the finite volume Scheme II developed in [11]. But we will not pursue this in this paper.

This paper is organized as follows. In section 2 we review the Hamiltonian-preserving scheme proposed in [11] for the Liouville equation with a discontinuous potential in one space dimension. In section 3 we recall the l^1-error estimates for the immersed interface upwind scheme established in [30, 28] for the linear advection equation with a step function wave speed. In section 4 we set up the problem and state the main theorem of this paper which is proved in section 5 using the results presented in section 3. We conclude the paper in section 6.

2. A Hamiltonian-preserving scheme. In this section we review the Hamiltonian-preserving scheme proposed in [11] for the Liouville equation in one space dimension

$$f_t + \xi f_x - V_x f_\xi = 0 \tag{2.1}$$

with a discontinuous potential $V(x)$.

Consider a uniform mesh with grid points at $x_{i+\frac{1}{2}}$, $i \in \mathbb{Z}$, in the x-direction and $\xi_{j+\frac{1}{2}}$, $j \in \mathbb{Z}$, in the ξ-direction. The cells are centered at $(x_i, \xi_j), (i,j) \in \mathbb{Z}^2$ with $x_i = \frac{1}{2}(x_{i+\frac{1}{2}} + x_{i-\frac{1}{2}})$ and $\xi_j = \frac{1}{2}(\xi_{j+\frac{1}{2}} + \xi_{j-\frac{1}{2}})$. The mesh size is denoted by $\Delta x = x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}}, \Delta\xi = \xi_{j+\frac{1}{2}} - \xi_{j-\frac{1}{2}}$. We also assume a uniform time step Δt, and the discrete times are given by $t_n = n\Delta t$, $n \in \mathbb{N} \cup \{0\}$. We assume that the computation is performed in a bounded rectangular domain

$$\{(x,y) | x_{\frac{1}{2}} \le x \le x_{N+\frac{1}{2}},\ \xi_{\frac{1}{2}} \le \xi \le \xi_{M+\frac{1}{2}}\}. \tag{2.2}$$

Let the cell averages of f be

$$f_{ij} = \frac{1}{\Delta x \Delta\xi} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \int_{\xi_{j-\frac{1}{2}}}^{\xi_{j+\frac{1}{2}}} f(x,\xi,t) d\xi dx.$$

The one-dimensional average quantity $f_{i+1/2,j}$ is defined as

$$f_{i+1/2,j} = \frac{1}{\Delta\xi} \int_{\xi_{j-\frac{1}{2}}}^{\xi_{j+\frac{1}{2}}} f(x_{i+1/2}, \xi, t) d\xi\,.$$

$f_{i,j+1/2}$ is defined similarly.

In classical mechanics, a particle will either cross a potential barrier with a changing momentum or be reflected, depending on its momentum and the strength of the potential barrier.

Figure 2.1 shows the typical situations when a particle moves from left to right at a potential barrier. If the potential is reduced, the particle will cross it with

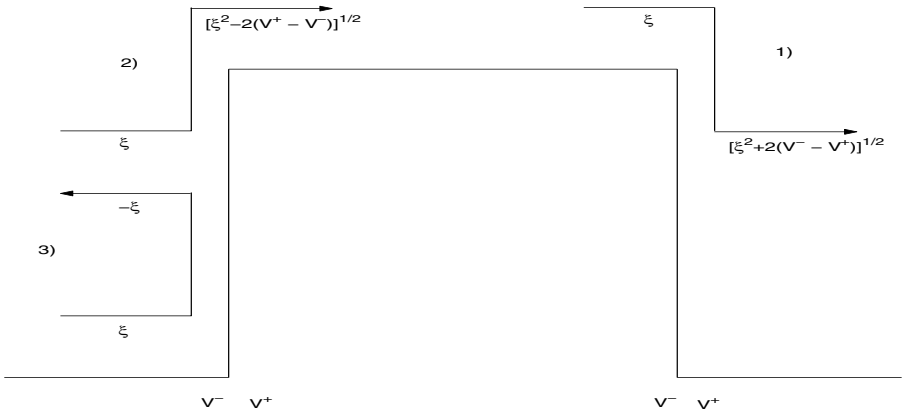


FIG. 2.1. *Transmission and reflection of a particle at a potential barrier.*

an accelerated speed. If the potential is increased, the particle is reflected if its momentum is not big enough to overcome the barrier or otherwise is transmitted with a reduced velocity. *The Hamiltonian* $H = \frac{1}{2}\xi^2 + V$ *is preserved across the potential barrier:*

$$\frac{1}{2}(\xi^+)^2 + V^+ = \frac{1}{2}(\xi^-)^2 + V^-, \tag{2.3}$$

where the superscripts $\pm$ indicate the right and left limits of the quantity at the potential barrier. This property was used in [11] to provide the interface condition for (2.1) at the barrier:

$$f(t, x^+, \xi^+) = f(t, x^-, \xi^-) \quad \text{for transmission,} \tag{2.4}$$

$$f(t, x^\pm, \xi^\pm) = f(t, x^\pm, -\xi^\pm) \quad \text{for reflection,} \tag{2.5}$$

where ξ^+ and ξ^- are related by the constant Hamiltonian condition (2.3) in the case of transmission. With such an interface condition, we established the well-posedness of the initial value problem to (1.1) with a piecewise constant potential in [12].

The main ingredient in the Hamiltonian-preserving schemes developed in [11], like the earlier work for shallow-water equations [22], was to build into the numerical flux the interface conditions (2.4), (2.5) at the barrier.

We now present the first Hamiltonian-preserving scheme, called *Scheme* I in [11].

Assume that the discontinuous points of the potential V are located at the grid points. Let the left and right limits of V at point $x_{i+1/2}$ be $V^-_{i+\frac{1}{2}}$ and $V^+_{i+\frac{1}{2}}$, respectively. Note that if V is continuous at $x_{j+1/2}$, then $V^-_{i+\frac{1}{2}} = V^+_{i+\frac{1}{2}}$. We approximate V by a piecewise linear function

$$V(x) \approx V^+_{i-1/2} + \frac{V^-_{i+1/2} - V^+_{i-1/2}}{\Delta x}(x - x_{i-1/2})\,.$$

The flux-splitting, semidiscrete scheme (with time continuous) reads

$$\partial_t f_{ij} + \xi_j \frac{f^-_{i+\frac{1}{2},j} - f^+_{i-\frac{1}{2},j}}{\Delta x} - \frac{V^-_{i+\frac{1}{2}} - V^+_{i-\frac{1}{2}}}{\Delta x}\frac{f_{i,j+\frac{1}{2}} - f_{i,j-\frac{1}{2}}}{\Delta \xi} = 0,$$

where the numerical fluxes $f_{i,j+\frac{1}{2}}$ are defined using the upwind discretization. Since the characteristics of the Liouville equation may be different on the two sides of a barrier, the corresponding numerical fluxes should also be different. The essential part of the algorithm is to define the split numerical fluxes $f^-_{i+\frac{1}{2},j}, f^+_{i-\frac{1}{2},j}$ at each cell interface. Conditions (2.4), (2.5) will be used to define these fluxes.

Assume V is discontinuous at $x_{i+1/2}$. Consider the case $\xi_j > 0$. Using upwind scheme, $f^-_{i+\frac{1}{2},j} = f_{ij}$. However,

$$f^+_{i+1/2,j} \equiv f(x^+_{i+1/2}, \xi^+) = f(x^-_{i+1/2}, \xi^-),$$

where ξ^- is obtained from $\xi^+ = \xi_j$ from (2.3). Since ξ^- may not be a grid point, we have to define it approximately. The first approach is to locate the two cell centers that bound this velocity and then use a linear interpolation to evaluate the needed numerical flux at ξ^-. The case of $\xi_j < 0$ is treated similarly. The algorithm to generate the numerical flux is given in [11]. Here we present the simplified algorithm for the case $V^-_{i+\frac{1}{2}} > V^+_{i+\frac{1}{2}}$ being discussed in this paper.

ALGORITHM I.

- $\xi_j > 0$

  $f^-_{i+\frac{1}{2},j} = f_{ij}$,

  ❑ if $\xi_j > \sqrt{2\left(V^-_{i+\frac{1}{2}} - V^+_{i+\frac{1}{2}}\right)}$,

  $\xi^- = \sqrt{\xi_j^2 + 2\left(V^+_{i+\frac{1}{2}} - V^-_{i+\frac{1}{2}}\right)}$

  if $\xi_k \le \xi^- < \xi_{k+1}$ for some k,

  then $f^+_{i+\frac{1}{2},j} = \frac{\xi_{k+1}-\xi^-}{\Delta\xi} f_{ik} + \frac{\xi^- - \xi_k}{\Delta\xi} f_{i,k+1}$

  ❑ else

  $f^+_{i+\frac{1}{2},j} = f_{i+1,k}$, where $\xi_k = -\xi_j$

  ❑ end

- $\xi_j < 0$

  $f^+_{i+\frac{1}{2},j} = f_{i+1,j}$,

  $\xi^+ = -\sqrt{\xi_j^2 + 2\left(V^-_{i+\frac{1}{2}} - V^+_{i+\frac{1}{2}}\right)}$

  if $\xi_k \le \xi^+ < \xi_{k+1}$ for some k,

  then $f^-_{i+\frac{1}{2},j} = \frac{\xi_{k+1}-\xi^+}{\Delta\xi} f_{i+1,k} + \frac{\xi^+ - \xi_k}{\Delta\xi} f_{i+1,k+1}$.

After the spatial discretization is specified, one can use any time discretization for the time derivative.

In [11] we proved that, when the first order upwind scheme is used spatially and the forward Euler method is used in time, and the potential V has a single jump, Scheme I is positive and l^∞-contracting under the CFL condition

$$\Delta t \left[\frac{\max_j |\xi_j|}{\Delta x} + \frac{\max_i \left| \frac{V^-_{i+\frac{1}{2}} - V^+_{i-\frac{1}{2}}}{\Delta x} \right|}{\Delta\xi} \right] < 1. \tag{2.6}$$

In [29] we proved that when the potential is a step function, the same scheme is l^1-stable under the CFL condition (2.6) and a suitable condition on the numerical initial data.

Note that the quantity $\left|\frac{V^-_{i+\frac{1}{2}} - V^+_{i-\frac{1}{2}}}{\Delta x}\right|$ represents the gradient of the potential at its

smooth point, which has a *finite* upper bound. Thus the scheme satisfies a hyperbolic CFL condition.

3. The l^1-error estimates for the immersed interface upwind scheme. In this section we present the l^1-error estimates for the immersed interface upwind scheme established in [30, 28] for the linear advection equation

$$\frac{\partial u}{\partial t} + c(x)\frac{\partial u}{\partial x} = 0, \qquad t > 0,\ x \in \mathbb{R}, \tag{3.1}$$

$$u|_{t=0} = u_0(x), \tag{3.2}$$

with a step function wave speed

$$c(x) = \begin{cases} c^-, & x < 0, \\ c^+, & x > 0. \end{cases} \tag{3.3}$$

We assume that $c(x)$ has a definite sign. This avoids the mathematical difficulties of dealing with nonuniqueness or measure-valued solutions; see, for example, [3, 23]. Without loss of generality we assume $c(x) > 0$.

We consider the following interface condition for (3.1)–(3.3) requiring u to be continuous across the interface:

$$u(0^+, t) = u(0^-, t). \tag{3.4}$$

In [30] we have considered a more general class of interface conditions including (3.4). In this paper we present only the results corresponding to the interface condition (3.4), which will be used to derive the l^1-error estimates for Scheme I.

The exact solution of (3.1)–(3.3) with the interface condition (3.4) can be constructed using the method of characteristics:

$$u(x,t) = \begin{cases} u_0(x - c^- t), & x < 0, \\ u_0\left(\frac{c^-}{c^+}x - c^- t\right), & 0 < x < c^+ t, \\ u_0(x - c^+ t), & x > c^+ t. \end{cases} \tag{3.5}$$

Consider the uniform mesh introduced in section 2 in the x- and t-directions. We assume that the interface $x = 0$ is located at a grid point. We introduce quantities $\lambda^- = c^-\frac{\Delta t}{\Delta x},\ \lambda^+ = c^+\frac{\Delta t}{\Delta x}$. The condition $0 < \lambda^-, \lambda^+ < 1$ is the CFL condition.

The immersed interface upwind scheme which builds the interface condition (3.4) into the upwind difference scheme (with the forward Euler time discretization) for (3.1)–(3.3) reads

$$v_i^{n+1} = (1 - \lambda^-)v_i^n + \lambda^- v_{i-1}^n \quad \text{if } x_i < 0, \tag{3.6}$$

$$v_i^{n+1} = (1 - \lambda^+)v_i^n + \lambda^+ v_{i-1}^n \quad \text{if } x_i > 0, \tag{3.7}$$

where

$$v_i^0 = \frac{1}{\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} u_0(x)dx. \tag{3.8}$$

To compare the numerical solution computed from (3.6)–(3.8) with the exact solution (3.5), we introduce

$$v(x,t) = v_i^n \qquad \text{for } (x,t) \in [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}) \times [t_n, t_{n+1}).$$

The following theorem was proved in [30].

THEOREM 3.1. *The upwind difference scheme* (3.6)–(3.8), *under the CFL condition* $0 < \lambda^-, \lambda^+ < 1$, *has the following* l^1*-error to the exact solution* (3.5)*:*

$$\|v(\cdot,t_n) - u(\cdot,t_n)\|_{l^1(\mathbb{R})} \le \|u_0\|_{BV(\mathbb{R})} \left[\gamma_m \frac{c^M}{c^m}\sqrt{\Delta x} + \left(1 + \frac{c^M}{c^m} + \max\left\{\frac{c^+}{c^-}, 1\right\}\right)\Delta x\right], \tag{3.9}$$

where

$$c^m = \min\{c^-, c^+\}, \quad c^M = \max\{c^-, c^+\}, \quad \gamma_m = \sqrt{\frac{2}{e}\, c^m \left(1 - c^m \frac{\Delta t}{\Delta x}\right) t_{n+1}},$$

and the definition of the BV norm on $\mathbb{R}$ *is given by*

$$\|u_0\|_{BV(\mathbb{R})} = \sup_{|h|\neq 0} \frac{1}{|h|} \|u_0(\cdot + h) - u_0(\cdot)\|_{l^1(\mathbb{R})}. \tag{3.10}$$

Remark 3.1. In the case of $c^- = c^+$, namely, a constant wave speed, the coefficient before Δx in (3.9) can be sharpened. In this paper we will also use (3.9) for the constant wave speed case since this will not affect the leading order coefficient in the l^1-error estimates for Scheme I.

4. Setup of the problem and the main result. In this section we establish the l^1-error estimates for Scheme I (with the first order numerical flux and the forward Euler method in time) under suitable conditions on the initial data. We consider the case when $V(x)$ is a step function, with a jump $-D, D > 0$ at $x = 0$. Namely,

$$V(0^-) - V(0^+) = D.$$

Let the computational domain be confined in the rectangular domain (2.2). We employ the uniform mesh introduced in section 2. Define mesh ratios $\lambda_x^t = \frac{\Delta t}{\Delta x}$, $\lambda_x^\xi = \frac{\Delta \xi}{\Delta x}$, assumed to be fixed. Let the potential barrier $x = 0$ be at a grid point $x_{m+\frac{1}{2}}$. Then the pointwise values of V satisfy

$$V^-_{m+\frac{1}{2}} - V^+_{m+\frac{1}{2}} = D, \quad V^\pm_{i+\frac{1}{2}} = V^-_{m+\frac{1}{2}},\ i < m, \quad V^\pm_{i+\frac{1}{2}} = V^+_{m+\frac{1}{2}},\ i > m,$$

where the superscripts $-, +$ represent the left and right limits at $x = 0$.

We consider the typical situation that $\xi_1 < -\sqrt{2D}$, $\xi_M > \sqrt{2D}$, so that all possible particle behaviors, including both transmission and reflection, are included. To simplify the discussion, we choose the mesh such that 0 and $\pm\sqrt{2D}$ are grid points in the ξ-direction. Define the indices I_0, I_+ satisfying

$$\xi_{I_0+\frac{1}{2}} = 0, \quad \xi_{I_++\frac{1}{2}} = \sqrt{2D}.$$

Define the domain

$$D_b = \left\{(x,y) | x < 0, \xi < -\sqrt{\left(\xi_{\frac{1}{2}}\right)^2 - 2D}\right\}.$$

Due to the velocity change across the barrier at $x = 0$, D_b represents the area where particles come from outside of the domain $[x_{\frac{1}{2}}, x_{N+\frac{1}{2}}] \times [\xi_{\frac{1}{2}}, \xi_{M+\frac{1}{2}}]$. In order to implement Scheme I conveniently, we need to exclude this domain from the computational domain. Define the index I_b satisfying

$$\xi_{I_b-\frac{3}{2}} < -\sqrt{\left(\xi_{\frac{1}{2}}\right)^2 - 2D} \le \xi_{I_b-\frac{1}{2}}$$

and the domain

$$\widehat{D}_b = \left\{(x,\xi) | x < 0, \xi < \xi_{I_b-\frac{1}{2}}\right\}.$$

Then we choose the computational domain as

$$D_C = \{(x,\xi) | x_{\frac{1}{2}} \le x \le x_{N+\frac{1}{2}},\ \xi_{\frac{1}{2}} \le \xi \le \xi_{M+\frac{1}{2}}\} \setminus \widehat{D}_b.$$

Figure 4.1 depicts D_C and $\widehat{D}_b$.

We consider the Dirichlet boundary conditions at the incoming boundaries and assume that the initial data satisfy these boundary conditions:

$$f(x_{\frac{1}{2}}, \xi, t) = f(x_{\frac{1}{2}}, \xi, 0), \quad 0 < \xi < \xi_{M+\frac{1}{2}}, \tag{4.1}$$

$$f(x_{N+\frac{1}{2}}, \xi, t) = f(x_{N+\frac{1}{2}}, \xi, 0), \quad \xi_{\frac{1}{2}} < \xi < 0. \tag{4.2}$$

Remark 4.1. When V is a step function, (2.1) becomes

$$f_t + \xi f_x = 0,$$

where ξ serves only as a parameter. However, when implementing the interface conditions (2.4), (2.5) numerically, the mesh in ξ is needed. This is the main new difficulty when comparing with the problem presented in section 3.

4.1. The initial data assumption. We now impose assumptions on the initial data. We assume that the initial data are given on the rectangular domain (2.2). We have the following assumption.

Assumption 4.1. The initial data $f(x,\xi,0)$ have bounded variation in the x-direction and are Lipschitz continuous in the ξ-direction. Namely,

$$\|f(.,\xi,0)\|_{BV([x_{\frac{1}{2}}, x_{N+\frac{1}{2}}])} \le A \quad \forall \xi \in [\xi_{\frac{1}{2}}, \xi_{M+\frac{1}{2}}], \tag{4.3}$$

$$|f(x,\xi',0) - f(x,\xi'',0)| \le B|\xi' - \xi''| \quad \forall x \in \left[x_{\frac{1}{2}}, x_{N+\frac{1}{2}}\right],\ \xi', \xi'' \in \left[\xi_{\frac{1}{2}}, \xi_{M+\frac{1}{2}}\right]. \tag{4.4}$$

Remark 4.2. When arising in the semiclassical limit of the linear Schrödinger equation, the Liouville equation is supplied with the measure-valued initial data [8, 18], which does not satisfy Assumption 4.1. However, in [10], a decomposition of the initial data was introduced, which allows one to solve the semiclassical limit problem with bounded initial data which do satisfy Assumption 4.1. Specifically, if one needs to solve the Liouville equation (2.1) with the measure-valued initial data

$$f(x,\xi,0) = \rho_0(x)\delta(\xi - v_0(x)), \tag{4.5}$$

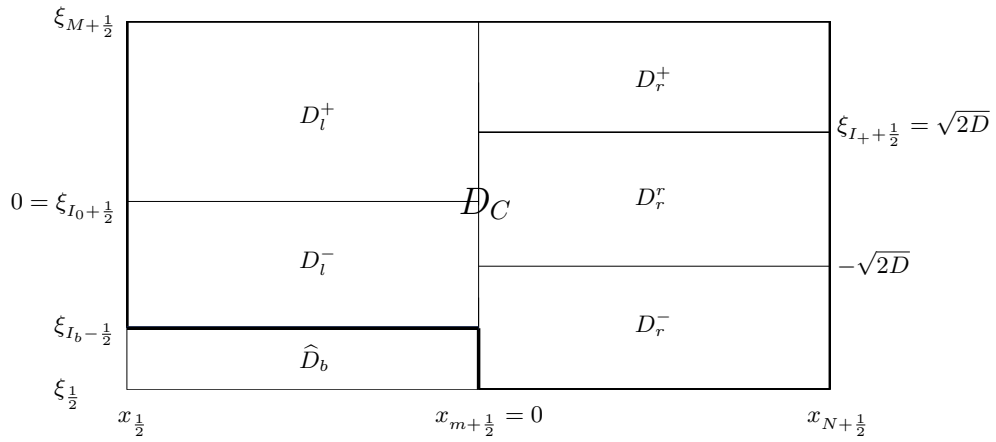


FIG. 4.1. *Sketch of partition of D_C and $\widehat{D}_b$.*

the decomposition technique proposed in [10] suggests that one just solves two functions satisfying the same Liouville equation with initial data

$$\phi(x,\xi,0) = \rho_0(x)\,, \quad \psi(x,\xi,0) = \xi - v_0(x)\,. \tag{4.6}$$

Then the measure-valued solution to the Liouville equation with the initial data (4.5) is simply

$$f(x,\xi,t) = \phi(x,\xi,t)\delta(\psi(x,\xi,t)).$$

The evaluations of the moments of f then resort to the numerical approximations to the delta function integrals.

In solving for ϕ, ψ, the initial data (4.6) satisfy Assumption 4.1 if the initial density and velocity $\rho_0(x), v_0(x)$ have bounded variations. For this situation we have established the l^1-error estimates for Scheme I.

Remark 4.3. It can be checked that the initial data satisfying Assumption 4.1 are bounded in both l^∞- and l^1-norms on D_C. When the numerical initial data are chosen to be cell averages of such initial data, the l^1-stability of Scheme I with homogeneous Dirichlet boundary conditions is established [29].

4.2. The exact solution. For our proof, we need to introduce the partition of D_C as

$$D_l^+ = \left\{(x,\xi) \;\middle|\; x_{\frac{1}{2}} < x < 0,\ 0 < \xi < \xi_{M+\frac{1}{2}}\right\},$$

$$D_l^- = \left\{(x,\xi) \;\middle|\; x_{\frac{1}{2}} < x < 0,\ \xi_{I_b-\frac{1}{2}} < \xi < 0\right\},$$

$$D_r^+ = \left\{(x,\xi) \;\middle|\; 0 < x < x_{N+\frac{1}{2}},\ \sqrt{2D} < \xi < \xi_{M+\frac{1}{2}}\right\},$$

$$D_r^- = \left\{(x,\xi) \;\middle|\; 0 < x < x_{N+\frac{1}{2}},\ \xi_{\frac{1}{2}} < \xi < -\sqrt{2D}\right\},$$

$$D_r^r = \left\{(x,\xi) \;\middle|\; 0 < x < x_{N+\frac{1}{2}},\ -\sqrt{2D} < \xi < \sqrt{2D}\right\}$$

and the extension of the initial data

(4.7)
$$\widehat{f}_0(x,\xi) = \begin{cases} f(x,\xi,0), & x_{\frac{1}{2}} \le x \le x_{N+\frac{1}{2}} \\ f(x_{\frac{1}{2}},\xi,0), & x < x_{\frac{1}{2}} \\ f(x_{N+\frac{1}{2}},\xi,0), & x > x_{N+\frac{1}{2}} \end{cases} \quad \text{for } x \in \mathbb{R},\ \xi_{\frac{1}{2}} \le \xi \le \xi_{M+\frac{1}{2}}.$$

Figure 4.1 shows a sketch of the partition of D_C.

The exact solution of (2.1) with the step function potential $V(x)$, when using the interface conditions (2.4), (2.5), can be obtained from the initial data $f(x,\xi,0)$ and boundary conditions (4.1), (4.2) by the method of characteristics:

(1) for $(x,\xi) \in D_l^+ \cup D_r^-$,

$$f(x,\xi,t_n) = \widehat{f}_0(x-\xi t_n,\xi)\,; \tag{4.8}$$

(2) for $(x,\xi) \in D_r^r$,

$$f(x,\xi,t_n) = \begin{cases} \widehat{f}_0(x-\xi t_n,\xi), & x > \xi t_n, \\ \widehat{f}_0(-x+\xi t_n,-\xi), & x < \xi t_n; \end{cases} \tag{4.9}$$

(3) for $(x,\xi) \in D_r^+$,

(4.10)
$$f(x,\xi,t_n) = \begin{cases} \widehat{f}_0\left(\frac{\sqrt{\xi^2-2D}}{\xi}\,x - \sqrt{\xi^2-2D}\,t_n, \sqrt{\xi^2-2D}\right), & 0 < x < \xi t_n, \\ \widehat{f}_0\left(x-\xi t_n,\xi\right), & x > \xi t_n; \end{cases}$$

(4) for $(x,\xi) \in D_l^-$,

(4.11)
$$f(x,\xi,t_n) = \begin{cases} \widehat{f}_0\left(-\frac{\sqrt{\xi^2+2D}}{\xi}\,x + \sqrt{\xi^2+2D}\,t_n, -\sqrt{\xi^2+2D}\right), & \xi t_n < x < 0, \\ \widehat{f}_0\left(x-\xi t_n,\xi\right), & x < \xi t_n. \end{cases}$$

4.3. The numerical solution. Denote

$$\mu_j = \lambda_x^t\,|\xi_j|\,, \qquad 1 \le j \le M. \tag{4.12}$$

Under the CFL condition (2.6), $\mu_j < 1$ for $1 \le j \le M$.

Since $V_x(x) = 0$ except at $x = x_{m+1/2}$, with the boundary conditions (4.1), (4.2), Scheme I on D_C is given by

(1) if $0 < \xi_j < \xi_{M+\frac{1}{2}}, i \ne m+1$,

$$g_{ij}^{n+1} = (1-\mu_j)g_{ij}^n + \mu_j g_{i-1,j}^n\,; \tag{4.13}$$

(2) if $\xi_{I_b-\frac{1}{2}} < \xi_j < 0,\ i < m$, or $\xi_{\frac{1}{2}} < \xi_j < 0,\ i > m$,

$$g_{ij}^{n+1} = (1-\mu_j)g_{ij}^n + \mu_j g_{i+1,j}^n\,; \tag{4.14}$$

(3) if $\sqrt{2D} < \xi_j < \xi_{M+\frac{1}{2}}$,

$$g^{n+1}_{m+1,j} = (1-\mu_j) g^n_{m+1,j} + \mu_j \left(c_{j,1} g^n_{m,d_j} + c_{j,2} g^n_{m,d_j+1} \right); \tag{4.15}$$

(4) if $0 < \xi_j < \sqrt{2D}$,

$$g^{n+1}_{m+1,j} = (1-\mu_j) g^n_{m+1,j} + \mu_j g^n_{m+1,d_j}; \tag{4.16}$$

(5) if $\xi_{I_b-\frac{1}{2}} < \xi_j < 0$,

$$g^{n+1}_{mj} = (1-\mu_j) g^n_{mj} + \mu_j \left(c_{j,1} g^n_{m+1,d_j} + c_{j,2} g^n_{m+1,d_j+1} \right), \tag{4.17}$$

where $0 \le c_{j,1}, c_{j,2} \le 1$ and $c_{j,1} + c_{j,2} = 1$. d_j's in (4.15)–(4.17) are determined according to Algorithm I by

$$\xi_{d_j} \le \sqrt{(\xi_j)^2 - 2D} < \xi_{d_j+1} \qquad \text{for } d_j \text{ in (4.15)}, \tag{4.18}$$

$$\xi_{d_j} = -\xi_j \qquad \text{for } d_j \text{ in (4.16)}, \tag{4.19}$$

$$\xi_{d_j} \le -\sqrt{(\xi_j)^2 + 2D} < \xi_{d_j+1} \qquad \text{for } d_j \text{ in (4.17)}. \tag{4.20}$$

The initial and incoming boundary values of the numerical solution are given by

$$g^0_{ij} = \frac{1}{\Delta x \Delta \xi} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \int_{\xi_{j-\frac{1}{2}}}^{\xi_{j+\frac{1}{2}}} f(x,\xi,0) d\xi dx, \quad (x_i,\xi_j) \in D_C, \tag{4.21}$$

$$g^n_{0,j} = \frac{1}{\Delta \xi} \int_{\xi_{j-\frac{1}{2}}}^{\xi_{j+\frac{1}{2}}} f(x_{\frac{1}{2}},\xi,0) d\xi, \quad 0 < \xi_j < \xi_{M+\frac{1}{2}}, \tag{4.22}$$

$$g^n_{N+1,j} = \frac{1}{\Delta \xi} \int_{\xi_{j-\frac{1}{2}}}^{\xi_{j+\frac{1}{2}}} f(x_{N+\frac{1}{2}},\xi,0) d\xi, \quad \xi_{\frac{1}{2}} < \xi_j < 0. \tag{4.23}$$

To compare the numerical solution computed from (4.13)–(4.23) with the exact solution (4.8)–(4.11), we introduce

$$g(x,\xi,t) = g^n_{ij} \qquad \text{for } (x,\xi,t) \in [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}) \times [\xi_{j-\frac{1}{2}}, \xi_{j+\frac{1}{2}}) \times [t_n, t_{n+1}), \quad (x_i,\xi_j) \in D_C.$$

We now state the main theorem of this paper.

THEOREM 4.1. *Under Assumption* 4.1 *on the initial data, the CFL condition* (2.6)*, and the mesh size restriction*

$$\Delta \xi \le \frac{3-2\sqrt{2}}{2} \sqrt{2D}, \tag{4.24}$$

Scheme I (4.13)–(4.23) *has the following* l^1*-error bound compared to the exact solution*

(4.8)–(4.11):

$$
\begin{aligned}
&\|g(\cdot,\cdot,t_n)-f(\cdot,\cdot,t_n)\|_{l^1(D_C)}\\
&\quad\le\left[\left(\xi_{M+\frac12}+\left|\xi_{\frac12}\right|+\sqrt{2D}\right)A+4DB\right]\sqrt{\frac{t_{n+1}}{2e\lambda_x^t}}\sqrt{\Delta x}\\
&\qquad+\left(4A+2\sqrt{2D}B\right)\sqrt{\frac{t_{n+1}}{2e}}\\
&\qquad\times\left(\frac{\xi_{M+\frac12}}{(2D)^{\frac14}}\sqrt{\frac{\xi_{M+\frac12}-\sqrt{2D}}{\lambda_x^t}}+2\left|\xi_{\frac12}\right|\left[\left(\xi_{\frac12}\right)^2-2D\right]^{\frac14}\right)\sqrt{\Delta x}\\
&\qquad+\left(2A+\sqrt{2D}B\right)\left[\frac{2\left(\xi_{M+\frac12}\right)^2t_n\lambda_x^\xi}{\sqrt{D}}+\left|\xi_{\frac12}\right|\right]\ln\left(\frac{1}{\Delta x}\right)\Delta x+O(\Delta x).
\end{aligned}
\tag{4.25}
$$

5. The proof of Theorem 4.1. This section is devoted to the proof of Theorem 4.1.

The l^1-error in (4.25) can be split according to the partition of D_C:

$$
\begin{aligned}
&\|g(\cdot,\cdot,t_n)-f(\cdot,\cdot,t_n)\|_{l^1(D_C)}\\
&=\|g(\cdot,\cdot,t_n)-f(\cdot,\cdot,t_n)\|_{l^1(D_l^+)}+\|g(\cdot,\cdot,t_n)-f(\cdot,\cdot,t_n)\|_{l^1(D_l^-)}\\
&\quad+\|g(\cdot,\cdot,t_n)-f(\cdot,\cdot,t_n)\|_{l^1(D_r^+)}+\|g(\cdot,\cdot,t_n)-f(\cdot,\cdot,t_n)\|_{l^1(D_r^r)}\\
&\quad+\|g(\cdot,\cdot,t_n)-f(\cdot,\cdot,t_n)\|_{l^1(D_r^-)}\\
&\equiv E_l^+ + E_l^- + E_r^+ + E_r^r + E_r^-.
\end{aligned}
\tag{5.1}
$$

We will estimate the five terms in (5.1). As stated in section 1, the strategy is to introduce a number of linear advection equations with step function coefficients in each partition whose immersed interface upwind schemes yield the same numerical solution as Scheme I. Then we apply the error estimates for the immersed interface upwind schemes given in section 3 and compare the exact solutions of the linear advection equations and that of the Liouville equation to derive the desired estimates.

5.1. The upper bounds for E_l^+ and E_r^-. For E_l^+ one has

$$
E_l^+=\sum_{j=I_0+1}^{M}\int_{\xi_{j-\frac12}}^{\xi_{j+\frac12}}\int_{x_{\frac12}}^{0}|g(x,\xi,t_n)-f(x,\xi,t_n)|\,dxd\xi.
\tag{5.2}
$$

For $I_0+1 \le j \le M$, consider the linear advection equation

$$\frac{\partial u^j_{l,+}}{\partial t} + \xi_j \frac{\partial u^j_{l,+}}{\partial x} = 0, \qquad t>0,\ x\in\mathbb{R}, \tag{5.3}$$

$$u^j_{l,+}|_{t=0} = \begin{cases} \frac{1}{\Delta\xi}\int_{\xi_{j-\frac12}}^{\xi_{j+\frac12}} \widehat{f}_0(x,\xi)d\xi, & x<0, \\ \frac{1}{\Delta\xi}\int_{\xi_{j-\frac12}}^{\xi_{j+\frac12}} \widehat{f}_0(0,\xi)d\xi, & x\ge 0. \end{cases} \tag{5.4}$$

The exact solution to (5.3), (5.4) is simply

$$u^j_{l,+}(x,t) = u^j_{l,+}(x-\xi_j t, 0). \tag{5.5}$$

Consider the upwind scheme for (5.3):

$$v_i^{n+1} = (1-\mu_j)v_i^n + \mu_j v_{i-1}^n, \tag{5.6}$$

$$v_i^0 = \frac{1}{\Delta x}\int_{x_{i-\frac12}}^{x_{i+\frac12}} u^j_{l,+}(x,0)dx, \tag{5.7}$$

where μ_j is defined in (4.12).

Define

$$v^j_{l,+}(x,t) = v_i^n \qquad \text{for } (x,t)\in[x_{i-\frac12}, x_{i+\frac12})\times[t_n, t_{n+1}).$$

Applying Theorem 3.1, the l^1-error between $v^j_{l,+}(x,t_n)$ and $u^j_{l,+}(x,t_n)$ is given by

$$\|v^j_{l,+}(\cdot,t_n) - u^j_{l,+}(\cdot,t_n)\|_{l^1(\mathbb{R})} \le \|u^j_{l,+}(\cdot,0)\|_{BV(\mathbb{R})}\left[\gamma_j\sqrt{\Delta x} + 3\Delta x\right], \tag{5.8}$$

where

$$\gamma_j = \sqrt{\frac{2}{e}\,\xi_j\left(1-\xi_j\frac{\Delta t}{\Delta x}\right)t_{n+1}}. \tag{5.9}$$

From definition (4.7) and condition (4.3) one has

$$\|\widehat{f}_0(\cdot,\xi)\|_{BV(\mathbb{R})} \le A, \quad \xi_{\frac12}\le\xi\le\xi_{M+\frac12}. \tag{5.10}$$

Definition (5.4) and condition (5.10) imply

$$\|u^j_{l,+}(\cdot,0)\|_{BV(\mathbb{R})} \le A. \tag{5.11}$$

γ_j in (5.9) satisfies

$$\gamma_j \le \sqrt{\frac{t_{n+1}}{2e}\frac{\Delta x}{\Delta t}} \equiv \gamma. \tag{5.12}$$

Substituting (5.11), (5.12) into (5.8) gives

$$\|v^j_{l,+}(\cdot,t_n) - u^j_{l,+}(\cdot,t_n)\|_{l^1(\mathbb{R})} \le A\left[\gamma\sqrt{\Delta x} + 3\Delta x\right]. \tag{5.13}$$

In comparison with schemes (4.13), (4.21), (4.22) and (5.6), (5.7), (5.4), (4.7), one can check that, for $I_0+1 \le j \le M$,

$$g(x,\xi,t_n) = v^j_{l,+}(x,t_n) \quad \text{for } \xi_{j-\frac12} < \xi < \xi_{j+\frac12},\ \ x_{\frac12} < x < 0. \tag{5.14}$$

From (5.2), (5.14), (5.13), (5.5), (5.4), and (4.8) one has, by the triangle inequality,

$$\begin{aligned} E^+_l \le & \sum_{j=I_0+1}^{M} \int_{\xi_{j-\frac12}}^{\xi_{j+\frac12}} \int_{x_{\frac12}}^{0} \left| v^j_{l,+}(x,t_n) - u^j_{l,+}(x,t_n) \right| dx d\xi \\ & + \sum_{j=I_0+1}^{M} \int_{\xi_{j-\frac12}}^{\xi_{j+\frac12}} \int_{x_{\frac12}}^{0} \left| u^j_{l,+}(x,t_n) - f(x,\xi,t_n) \right| dx d\xi \equiv E^{+,1}_l + E^{+,2}_l, \end{aligned} \tag{5.15}$$

where

$$E^{+,1}_l \le \xi_{M+\frac12} A \left[\gamma \sqrt{\Delta x} + 3\Delta x \right], \tag{5.16}$$

$$\begin{aligned} E^{+,2}_l &= \sum_{j=I_0+1}^{M} \int_{\xi_{j-\frac12}}^{\xi_{j+\frac12}} \int_{x_{\frac12}}^{0} \left| \frac{1}{\Delta\xi} \int_{\xi_{j-\frac12}}^{\xi_{j+\frac12}} \widehat{f}_0(x-\xi_j t_n, \xi') d\xi' - \widehat{f}_0(x-\xi t_n, \xi) \right| dx d\xi \\ &\le \sum_{j=I_0+1}^{M} \int_{\xi_{j-\frac12}}^{\xi_{j+\frac12}} \int_{x_{\frac12}}^{0} \left| \frac{1}{\Delta\xi} \int_{\xi_{j-\frac12}}^{\xi_{j+\frac12}} \left[\widehat{f}_0(x-\xi_j t_n, \xi') - \widehat{f}_0(x-\xi_j t_n, \xi) \right] d\xi' \right| dx d\xi \\ &\quad + \sum_{j=I_0+1}^{M} \int_{\xi_{j-\frac12}}^{\xi_{j+\frac12}} \int_{x_{\frac12}}^{0} \left| \widehat{f}_0(x-\xi_j t_n, \xi) - \widehat{f}_0(x-\xi t_n, \xi) \right| dx d\xi. \end{aligned} \tag{5.17}$$

According to the definition of the BV norm (3.10) and condition (5.10), one has

$$\int_{x_{\frac12}}^{0} \left| \widehat{f}_0(x-\xi_j t_n, \xi) - \widehat{f}_0(x-\xi t_n, \xi) \right| dx \le \| \widehat{f}_0(\cdot,\xi) \|_{BV(\mathbb{R})} |\xi - \xi_j| t_n \le A |\xi - \xi_j| t_n. \tag{5.18}$$

Applying the Lipschitz condition (4.4) and (5.18) to (5.17) yields

$$E^{+,2}_l \le \xi_{M+\frac12} \left| x_{\frac12} \right| B \Delta\xi + \sum_{j=I_0+1}^{M} \int_{\xi_{j-\frac12}}^{\xi_{j+\frac12}} |\xi - \xi_j|\, t_n A d\xi \le \xi_{M+\frac12} \left[\left| x_{\frac12} \right| B + \frac{t_n}{2} A \right] \Delta\xi. \tag{5.19}$$

Combining (5.15), (5.16), and (5.19) leads to

$$\begin{aligned} E^+_l &\le \xi_{M+\frac12} A \left[\gamma \sqrt{\Delta x} + 3\Delta x \right] + \xi_{M+\frac12} \left[\left| x_{\frac12} \right| B + \frac{t_n}{2} A \right] \Delta\xi \\ &= \xi_{M+\frac12} A \gamma \sqrt{\Delta x} + O(\Delta x). \end{aligned} \tag{5.20}$$

Similarly, for E_r^- one can deduce

$$E_r^- \le \left|\xi_{\frac{1}{2}} + \sqrt{2D}\right| A\left[\gamma\sqrt{\Delta x} + 3\Delta x\right] + \left|\xi_{\frac{1}{2}} + \sqrt{2D}\right|\left[x_{N+\frac{1}{2}}B + \frac{t_n}{2}A\right]\Delta\xi$$

$$= \left|\xi_{\frac{1}{2}} + \sqrt{2D}\right| A\gamma\sqrt{\Delta x} + O(\Delta x). \tag{5.21}$$

5.2. The upper bound for E_r^r. For E_r^r one has

$$E_r^r = \sum_{j=I_0+1}^{I_+} \int_0^{x_{N+\frac{1}{2}}} \left(\int_{\xi_{j-\frac{1}{2}}}^{\xi_{j+\frac{1}{2}}} |g(x,\xi,t_n) - f(x,\xi,t_n)|\, d\xi \right. \tag{5.22}$$

$$\left. + \int_{-\xi_{j+\frac{1}{2}}}^{-\xi_{j-\frac{1}{2}}} |g(x,\xi,t_n) - f(x,\xi,t_n)|\, d\xi \right) dx.$$

For $0 \le \xi \le \sqrt{2D}$, define the function

$$f_1(x,\xi) = \begin{cases} \widehat{f}_0(x,\xi), & x \ge 0 \\ \widehat{f}_0(-x,-\xi), & x < 0 \end{cases} \qquad \text{for } x \in \mathbb{R},\ 0 \le \xi \le \sqrt{2D}. \tag{5.23}$$

For $I_0 + 1 \le j \le I_+$, consider the linear advection equation

$$\frac{\partial u_{r,r}^j}{\partial t} + \xi_j \frac{\partial u_{r,r}^j}{\partial x} = 0, \qquad t > 0,\ x \in \mathbb{R}, \tag{5.24}$$

$$u_{r,r}^j|_{t=0} = \frac{1}{\Delta\xi} \int_{\xi_{j-\frac{1}{2}}}^{\xi_{j+\frac{1}{2}}} f_1(x,\xi) d\xi. \tag{5.25}$$

The exact solution to (5.24), (5.25) is simply

$$u_{r,r}^j(x,t) = u_{r,r}^j(x - \xi_j t, 0).$$

The upwind scheme for (5.24) is

$$v_i^{n+1} = (1-\mu_j)v_i^n + \mu_j v_{i-1}^n, \tag{5.26}$$

$$v_i^0 = \frac{1}{\Delta x} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} u_{r,r}^j(x,0) dx. \tag{5.27}$$

Define

$$v_{r,r}^j(x,t) = v_i^n \qquad \text{for } (x,t) \in [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}) \times [t_n, t_{n+1}).$$

Applying Theorem 3.1, the l^1-error between $v_{r,r}^j(x,t_n)$ and $u_{r,r}^j(x,t_n)$ is given by

$$\|v_{r,r}^j(\cdot,t_n) - u_{r,r}^j(\cdot,t_n)\|_{l^1(\mathbb{R})} \le \|u_{r,r}^j(\cdot,0)\|_{BV(\mathbb{R})} \left[\gamma_j\sqrt{\Delta x} + \Delta x\right], \tag{5.28}$$

where γ_j is given by (5.9).

From definition (5.23) and conditions (4.3), (4.4) one has, for $0 \le \xi \le \sqrt{2D}$,

$$\begin{aligned}\|f_1(\cdot,\xi)\|_{BV(\mathbb{R})} &\le \|\widehat{f}_0(.,\xi)\|_{BV([0,x_{N+\frac{1}{2}}])} + \|\widehat{f}_0(.,-\xi)\|_{BV([0,x_{N+\frac{1}{2}}])} \\ &\quad + \left|\widehat{f}_0(0,-\xi) - \widehat{f}_0(0,\xi)\right| \\ &\le 2A + 2\sqrt{2D}B. \end{aligned} \tag{5.29}$$

Therefore

$$\|u^j_{r,r}(\cdot,0)\|_{BV(\mathbb{R})} \le 2A + 2\sqrt{2D}B. \tag{5.30}$$

Combining (5.28), (5.30), and (5.12) gives

$$\|v^j_{r,r}(\cdot,t_n) - u^j_{r,r}(\cdot,t_n)\|_{l^1(\mathbb{R})} \le \left(2A + 2\sqrt{2D}B\right)\left[\gamma\sqrt{\Delta x} + \Delta x\right] \tag{5.31}$$

with γ defined in (5.12).

In comparison with schemes (4.13), (4.14), (4.16), (4.19), (4.21), (4.23) and (5.26), (5.27), (5.25), (5.23), (4.7), one can check that, for $I_0 + 1 \le j \le I_+$,

(5.32)
$$g(x,\xi,t_n) = \begin{cases} v^j_{r,r}(x,t_n) & \text{for } 0 < x < x_{N+\frac{1}{2}},\ \xi_{j-\frac{1}{2}} < \xi < \xi_{j+\frac{1}{2}}, \\ v^j_{r,r}(-x,t_n) & \text{a.e. for } 0 < x < x_{N+\frac{1}{2}},\ -\xi_{j+\frac{1}{2}} < \xi < -\xi_{j-\frac{1}{2}}. \end{cases}$$

The second part of (5.32) holds except possibly at the cell interfaces.

From (5.22), (5.32), (4.9), and (5.23), one arrives at

$$\begin{aligned} E^r_r &= \sum_{j=I_0+1}^{I_+} \int_{\xi_{j-\frac{1}{2}}}^{\xi_{j+\frac{1}{2}}} \int_0^{x_{N+\frac{1}{2}}} \left|v^j_{r,r}(x,t_n) - f_1(x-\xi t_n,\xi)\right| dxd\xi \\ &\quad + \sum_{j=I_0+1}^{I_+} \int_{-\xi_{j+\frac{1}{2}}}^{-\xi_{j-\frac{1}{2}}} \int_0^{x_{N+\frac{1}{2}}} \left|v^j_{r,r}(-x,t_n) - f_1(-x+\xi t_n,-\xi)\right| dxd\xi \\ &= \sum_{j=I_0+1}^{I_+} \int_{\xi_{j-\frac{1}{2}}}^{\xi_{j+\frac{1}{2}}} \int_{-x_{N+\frac{1}{2}}}^{x_{N+\frac{1}{2}}} \left|v^j_{r,r}(x,t_n) - f_1(x-\xi t_n,\xi)\right| dxd\xi. \end{aligned}$$

From definition (5.23) and condition (4.4), $f_1(x,\xi)$ satisfies the Lipschitz condition in ξ with the Lipschitz constant B. Utilizing this and (5.29), (5.31), similarly to the deduction from (5.15) to (5.20), one obtains

$$\begin{aligned} E^r_r &\le \sqrt{2D}\left(2A + 2\sqrt{2D}B\right)\left[\gamma\sqrt{\Delta x} + \Delta x\right] \\ &\quad + \sqrt{2D}\left[2x_{N+\frac{1}{2}}B + t_n\left(A + \sqrt{2D}B\right)\right]\Delta\xi \\ &= \sqrt{2D}\left(2A + 2\sqrt{2D}B\right)\gamma\sqrt{\Delta x} + O(\Delta x). \end{aligned} \tag{5.33}$$

5.3. The upper bound for E_r^+. For E_r^+ one has

$$E_r^+ = \sum_{j=I_++1}^{M} \int_{\xi_{j-\frac{1}{2}}}^{\xi_{j+\frac{1}{2}}} \int_0^{x_{N+\frac{1}{2}}} |g(x,\xi,t_n) - f(x,\xi,t_n)|\, dx d\xi. \tag{5.34}$$

For $I_+ + 1 \le j \le M$, consider two linear advection equations

$$\frac{\partial u^j_{r,+,p}}{\partial t} + c_{+,p} \frac{\partial u^j_{r,+,p}}{\partial x} = 0, \qquad t > 0,\ x \in \mathbb{R}, \tag{5.35}$$

$$c_{+,p}(x) = \begin{cases} \xi_{J_{j,p}}, & x < 0, \\ \xi_j, & x > 0, \end{cases} \tag{5.36}$$

$$u^j_{r,+,p}|_{t=0} = \frac{1}{\Delta\xi} \int_{\xi_{j-\frac{1}{2}}}^{\xi_{j+\frac{1}{2}}} f^j_{r,+,p}(x,\xi) d\xi, \tag{5.37}$$

$$f^j_{r,+,p}(x,\xi) = \begin{cases} \widehat{f}_0(x, \xi - \xi_j + \xi_{J_{j,p}}), & x < 0, \\ \widehat{f}_0(x,\xi), & x \ge 0 \end{cases} \tag{5.38}$$

for $p = 1, 2$, where

$$J_{j,p} = d_j + p - 1, \quad p = 1, 2, \tag{5.39}$$

with d_j defined in (4.18).

The exact solutions to (5.35)–(5.38) are given by

$$u^j_{r,+,p}(x,t) = \begin{cases} u^j_{r,+,p}\left(x - \xi_{J_{j,p}} t, 0\right), & x < 0, \\ u^j_{r,+,p}\left(\frac{\xi_{J_{j,p}}}{\xi_j} x - \xi_{J_{j,p}} t, 0\right), & 0 \le x < \xi_j t, \quad p = 1, 2. \\ u^j_{r,+,p}\left(x - \xi_j t, 0\right), & x \ge \xi_j t, \end{cases} \tag{5.40}$$

The upwind schemes for (5.35) are

$$v_i^{n+1,j,p} = \begin{cases} (1 - \mu_{J_{j,p}}) v_i^{n,j,p} + \mu_{J_{j,p}} v_{i-1}^{n,j,p} & \text{if } x_i < 0, \\ (1 - \mu_j) v_i^{n,j,p} + \mu_j v_{i-1}^{n,j,p} & \text{if } x_i > 0, \end{cases} \tag{5.41}$$

$$v_i^{0,j,p} = \frac{1}{\Delta x} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} u^j_{r,+,p}(x,0) dx \tag{5.42}$$

for $p = 1, 2$.

Define

$$v^j_{r,+,p}(x,t) = v_i^{n,j,p} \qquad \text{for } (x,t) \in [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}) \times [t_n, t_{n+1}),\ \ p = 1, 2.$$

We present some lemmas before giving estimates for E_r^+.

LEMMA 5.1. *If the mesh size satisfies* (4.24)*, then*

$$\frac{\xi_j}{\xi_{J_{j,p}}} \le \frac{\xi_{M+\frac{1}{2}}}{(2D)^{\frac{1}{4}} \sqrt{(j - I_+ - \frac{1}{2})\Delta\xi}}, \quad p = 1, 2, \tag{5.43}$$

$$\frac{\xi_j}{\sqrt{(\xi_j)^2 - 2D}} \le \frac{\xi_{M+\frac{1}{2}}}{\sqrt{2}(2D)^{\frac{1}{4}} \sqrt{(j - I_+ - \frac{1}{2})\Delta\xi}} \tag{5.44}$$

for $I_+ + 1 \le j \le M$.

Proof. We give the proof for (5.43); the proof for (5.44) is similar. One has for $I_+ + 1 \le j \le M$

$$\frac{\xi_j}{\xi_{J_{j,p}}} \le \frac{\sqrt{2D} + (j - I_+ - \frac{1}{2})\Delta\xi}{\sqrt{2\sqrt{2D}(j - I_+ - \frac{1}{2})\Delta\xi + \left[(j - I_+ - \frac{1}{2})\Delta\xi\right]^2} - \Delta\xi}$$

$$\le \frac{\xi_{M+\frac{1}{2}}}{\sqrt{2\sqrt{2D}(j - I_+ - \frac{1}{2})\Delta\xi} - \Delta\xi}.$$

Utilizing the mesh size condition (4.24), one then obtains (5.43). ∎

LEMMA 5.2.

(5.45)

$$\left|\sqrt{\xi^2 - 2D} - \xi'\right| \le \frac{\xi_j \Delta\xi}{\sqrt{(\xi_j)^2 - 2D}} + \frac{3}{2}\Delta\xi \quad \forall \xi_{j-\frac{1}{2}} \le \xi \le \xi_{j+\frac{1}{2}},\ \xi_{d_j-\frac{1}{2}} \le \xi' \le \xi_{d_j+\frac{3}{2}}$$

for $I_+ + 1 \le j \le M$.

Proof. Observe that, for all ξ, $\xi_{j-\frac{1}{2}} \le \xi \le \xi_{j+\frac{1}{2}}$,

$$\left|\sqrt{\xi^2 - 2D} - \sqrt{(\xi_j)^2 - 2D}\right| \le \sqrt{(\xi_j)^2 - 2D} - \sqrt{\left(\xi_{j-\frac{1}{2}}\right)^2 - 2D}$$

$$\le \frac{(\xi_j)^2 - \left(\xi_j - \frac{1}{2}\Delta\xi\right)^2}{\sqrt{(\xi_j)^2 - 2D}} \le \frac{\xi_j \Delta\xi}{\sqrt{(\xi_j)^2 - 2D}}.$$

Therefore

$$\left|\sqrt{\xi^2 - 2D} - \xi_{J_{j,p}}\right| \le \frac{\xi_j \Delta\xi}{\sqrt{(\xi_j)^2 - 2D}} + \Delta\xi, \quad p = 1, 2. \tag{5.46}$$

Thus (5.45) can be deduced. ∎

LEMMA 5.3. *Define functions*

$$H_\xi^{+,j,p}(x) = \begin{cases} \left(\frac{\sqrt{\xi^2 - 2D}}{\xi}\, x - \sqrt{\xi^2 - 2D}\, t_n + \xi_{J_{j,p}} t_n\right) \frac{\xi_j}{\xi_{J_{j,p}}}, & x < \xi t_n, \\ x - \xi t_n + \xi_j t_n, & x \ge \xi t_n. \end{cases}$$

Then

$$\left|H_\xi^{+,j,p}(x) - x\right| \le E_H^j \tag{5.47}$$

for $x \ge 0$, $\xi_{j-\frac{1}{2}} \le \xi \le \xi_{j+\frac{1}{2}}$, $I_+ + 1 \le j \le M$, $p = 1, 2$, *where*

$$E_H^j = \frac{5}{2}\frac{t_n \xi_{M+\frac{1}{2}}}{(2D)^{\frac{1}{4}}}\frac{\Delta\xi}{\sqrt{(j - I_+ - \frac{1}{2})\Delta\xi}} + \frac{\sqrt{2}\left(\xi_{M+\frac{1}{2}}\right)^2 t_n}{\sqrt{2D}}\frac{\Delta\xi}{(j - I_+ - \frac{1}{2})\Delta\xi}. \tag{5.48}$$

Proof. Clearly,

$$\left|H_\xi^{+,j,p}(x)-x\right| \le \frac{\Delta\xi}{2}t_n \quad \text{for } x \ge \xi t_n,\ \xi_{j-\frac{1}{2}} \le \xi \le \xi_{j+\frac{1}{2}}. \tag{5.49}$$

For $0 \le x < \xi t_n$,

$$\left|H_\xi^{+,j,p}(x)-x\right| = \left|\frac{\sqrt{\xi^2-2D}}{\xi}\,x - \frac{\xi_{J_{j,p}}}{\xi_j}\,x + \left(\xi_{J_{j,p}} - \sqrt{\xi^2-2D}\right)\,t_n\right| \frac{\xi_j}{\xi_{J_{j,p}}}$$

$$\le \left[\left(\frac{\xi_{J_{j,p}}\,|\xi-\xi_j|}{\xi\xi_j} + \frac{\left|\sqrt{\xi^2-2D}-\xi_{J_{j,p}}\right|}{\xi}\right)x + \left|\sqrt{\xi^2-2D}-\xi_{J_{j,p}}\right|\,t_n\right]\frac{\xi_j}{\xi_{J_{j,p}}}$$

$$\le \left(2\left|\sqrt{\xi^2-2D}-\xi_{J_{j,p}}\right| + |\xi-\xi_j|\right)t_n\frac{\xi_j}{\xi_{J_{j,p}}}. \tag{5.50}$$

Combining (5.49) and (5.50), applying (5.46) and Lemma 5.1 to (5.50) then leads to the estimates (5.47). ∎

LEMMA 5.4. *Define functions*

$$F_p^j(\xi) = \int_0^{x_{N+\frac{1}{2}}} \left|\widetilde{f}_{r,+,p}^{\,j}(x,\xi,t_n) - \widehat{f}_{r,+}(x,\xi,t_n)\right| dx \tag{5.51}$$

for $\xi_{j-\frac{1}{2}} \le \xi \le \xi_{j+\frac{1}{2}},\ I_+ + 1 \le j \le M,\ p = 1,2,$ *where*

$$\widetilde{f}_{r,+,p}^{\,j}(x,\xi,t) = \begin{cases} \widehat{f}_0\left(\frac{\xi_{J_{j,p}}}{\xi_j}x - \xi_{J_{j,p}}t, \sqrt{\xi^2-2D}\right), & x < \xi_j t, \\ \widehat{f}_0\left(x-\xi_j t,\xi\right), & x \ge \xi_j t, \end{cases} \tag{5.52}$$

$$\widehat{f}_{r,+}(x,\xi,t) = \begin{cases} \widehat{f}_0\left(\frac{\sqrt{\xi^2-2D}}{\xi}\,x - \sqrt{\xi^2-2D}\,t, \sqrt{\xi^2-2D}\right), & x < \xi t, \\ \widehat{f}_0\left(x-\xi t,\xi\right), & x \ge \xi t. \end{cases} \tag{5.53}$$

Then

$$F_p^j(\xi) \le 2\left(2A + \sqrt{2D}B\right)E_H^j \tag{5.54}$$

with E_H^j *defined in* (5.48).

Proof. From the definition (5.52) and the conditions (4.3), (4.4), one obtains

$$\|\widetilde{f}_{r,+,p}^{\,j}(\cdot,\xi,t)\|_{BV(\mathbb{R})} \le 2A + \sqrt{2D}B. \tag{5.55}$$

We use the fact that

$$\widehat{f}_{r,+}(x,\xi,t_n) = \widetilde{f}_{r,+,p}^{\,j}\left(H_\xi^{+,j,p}(x),\xi,t_n\right), \quad x \in \mathbb{R},\ p = 1,2. \tag{5.56}$$

From (5.51), (5.56), and (5.55), applying Lemma 5.3 and Lemma A.1 in the appendix then leads to (5.54). ∎

We now give estimates for E_r^+. In comparison with schemes (4.13), (4.15), (4.18), (4.21), (4.22) and (5.41), (5.42), (5.37), (5.38), (4.7), one can check that, for $I_+ + 1 \le j \le M$,

$$g(x,\xi,t_n) = \sum_{p=1}^{2} c_{j,p} v^j_{r,+,p}(x,t_n) \quad \text{for } \xi_{j-\frac{1}{2}} < \xi < \xi_{j+\frac{1}{2}},\ 0 < x < x_{N+\frac{1}{2}}. \tag{5.57}$$

Together with (5.34), (5.57) and the expression (4.10) of $f(x,\xi,t_n)$ in D_r^+, one obtains by triangle inequality

$$E_r^+ \le E_r^{+,1} + E_r^{+,2} + E_r^{+,3}, \tag{5.58}$$

where

$$E_r^{+,1} = \sum_{j=I_++1}^{M} \int_{\xi_{j-\frac{1}{2}}}^{\xi_{j+\frac{1}{2}}} \int_0^{x_{N+\frac{1}{2}}} \sum_{p=1}^{2} c_{j,p} \left| v^j_{r,+,p}(x,t_n) - u^j_{r,+,p}(x,t_n) \right| dx d\xi, \tag{5.59}$$

$$E_r^{+,2} = \sum_{j=I_++1}^{M} \int_{\xi_{j-\frac{1}{2}}}^{\xi_{j+\frac{1}{2}}} \int_0^{x_{N+\frac{1}{2}}} \sum_{p=1}^{2} c_{j,p} \left| u^j_{r,+,p}(x,t_n) - \widetilde{f}^j_{r,+,p}(x,\xi,t_n) \right| dx d\xi, \tag{5.60}$$

$$E_r^{+,3} = \sum_{j=I_++1}^{M} \int_{\xi_{j-\frac{1}{2}}}^{\xi_{j+\frac{1}{2}}} \int_0^{x_{N+\frac{1}{2}}} \sum_{p=1}^{2} c_{j,p} \left| \widetilde{f}^j_{r,+,p}(x,\xi,t_n) - \widehat{f}_{r,+}(x,\xi,t_n) \right| dx d\xi. \tag{5.61}$$

We then estimate these three terms.

Applying Theorem 3.1, the l^1-error between $v^j_{r,+,p}$ and $u^j_{r,+,p}$, $p = 1, 2$, is given by

$$\|v^j_{r,+,p}(\cdot,t_n) - u^j_{r,+,p}(\cdot,t_n)\|_{l^1(\mathbb{R})} \le \|u^j_{r,+,p}(\cdot,0)\|_{BV(\mathbb{R})} \left[\Delta x + \left(\gamma\sqrt{\Delta x} + 2\Delta x \right) \frac{\xi_j}{\xi_{J_{j,p}}} \right], \tag{5.62}$$

where γ is given by (5.12).

For $I_+ + 1 \le j \le M$, the index d_j defined in (4.18) satisfies

$$\xi_j - \sqrt{2D} \le \xi_{d_j} < \xi_{d_j+1} \le \xi_j. \tag{5.63}$$

From definitions (5.38) and conditions (4.3), (4.4), (5.63) one has, for $p = 1, 2$,

$$\|f^j_{r,+,p}(\cdot,\xi)\|_{BV(\mathbb{R})} \le 2A + \sqrt{2D}B, \qquad \xi_{j-\frac{1}{2}} \le \xi \le \xi_{j+\frac{1}{2}}.$$

Therefore

$$\|u^j_{r,+,p}(\cdot,0)\|_{BV(\mathbb{R})} \le 2A + \sqrt{2D}B. \tag{5.64}$$

Utilizing (5.62), (5.64), and (5.43) for (5.59), by summation of j and using the inequality

$$\sum_{j=I_++1}^{M} \frac{\Delta\xi}{\sqrt{(j - I_+ - \frac{1}{2})\Delta\xi}} \le 2\sqrt{\xi_{M+\frac{1}{2}} - \sqrt{2D}}, \tag{5.65}$$

one obtains

$$E_r^{+,1} \le \left(4A + 2\sqrt{2D}B\right) \frac{\xi_M}{(+\frac{1}{2}2D)^{\frac{1}{4}}} \sqrt{\xi_{M+\frac{1}{2}} - \sqrt{2D}}\gamma\sqrt{\Delta x} + O(\Delta x). \tag{5.66}$$

Using Lemma 5.2, the fact that

$$|\xi - \xi'| \le \Delta\xi \quad \forall\, \xi_{j-\frac{1}{2}} \le \xi,\ \xi' \le \xi_{j+\frac{1}{2}}, \tag{5.67}$$

and condition (4.4), from definitions (5.37), (5.38), (5.40), and (5.52) one has

$$\left|u^j_{r,+,p}(x,t_n) - \widetilde{f}^j_{r,+,p}(x,\xi,t_n)\right| \le B\left[\frac{\xi_j\Delta\xi}{\sqrt{(\xi_j)^2 - 2D}} + \frac{3}{2}\Delta\xi\right] \tag{5.68}$$

for $x \ge 0,\ \xi_{j-\frac{1}{2}} \le \xi \le \xi_{j+\frac{1}{2}}$.

Utilizing (5.68) and (5.44) for (5.60) and using the inequality (5.65), one obtains

$$E_r^{+,2} \le x_{N+\frac{1}{2}} B\left[\left(\xi_{M+\frac{1}{2}} - \sqrt{2D}\right)\frac{3}{2} + \frac{\xi_{M+\frac{1}{2}}}{\sqrt{2}(2D)^{\frac{1}{4}}} 2\sqrt{\xi_{M+\frac{1}{2}} - \sqrt{2D}}\right]\Delta\xi. \tag{5.69}$$

Applying Lemma 5.4 to (5.61), using the inequality (5.65) and

$$\sum_{j=I_++1}^{M} \frac{\Delta\xi}{(j - I_+ - \frac{1}{2})\Delta\xi} \le 2 + \ln\left(\xi_{M+\frac{1}{2}} - \sqrt{2D}\right) + \ln\left(\frac{1}{\Delta\xi}\right), \tag{5.70}$$

one obtains

$$E_r^{+,3} \le \left(4A + 2\sqrt{2D}B\right) \frac{\left(\xi_{M+\frac{1}{2}}\right)^2 t_n \lambda_x^{\xi}}{\sqrt{D}} \ln\left(\frac{1}{\Delta x}\right)\Delta x + O(\Delta x). \tag{5.71}$$

The inequalities (5.65) and (5.70) are obtained using the fact that the midpoint rule is less than the exact integral when the integrand is convex. The right-hand side of the inequality (5.70) contains the $\ln(\cdot)$ term which results in the $\ln(\cdot)$ term appearing in the estimate (4.25) in Theorem 4.1.

Combining (5.58), (5.66), (5.69), (5.71) gives

$$\begin{aligned} E_r^{+} \le{}& \left(4A + 2\sqrt{2D}B\right) \frac{\xi_{M+\frac{1}{2}}}{(2D)^{\frac{1}{4}}} \sqrt{\xi_{M+\frac{1}{2}} - \sqrt{2D}}\gamma\sqrt{\Delta x} \\ &+ \left(4A + 2\sqrt{2D}B\right) \frac{\left(\xi_{M+\frac{1}{2}}\right)^2 t_n \lambda_x^{\xi}}{\sqrt{D}} \ln\left(\frac{1}{\Delta x}\right)\Delta x + O(\Delta x). \end{aligned} \tag{5.72}$$

5.4. The upper bound for E_l^-. For E_l^- one has

$$E_l^- = \sum_{j=I_b}^{I_0} \int_{\xi_{j-\frac{1}{2}}}^{\xi_{j+\frac{1}{2}}} \int_{x_{\frac{1}{2}}}^{0} |g(x,\xi,t_n) - f(x,\xi,t_n)|\, dx d\xi. \tag{5.73}$$

For $I_b \le j \le I_0$, consider two linear advection equations

$$\frac{\partial u^j_{l,-,p}}{\partial t} + c_{-,p}\frac{\partial u^j_{l,-,p}}{\partial x} = 0, \qquad t>0,\ x\in\mathbb{R}, \tag{5.74}$$

$$c_{-,p}(x) = \begin{cases} \left|\xi_{J_{j,p}}\right|, & x<0,\\ |\xi_j|, & x>0, \end{cases} \tag{5.75}$$

$$u^j_{l,-,p}|_{t=0} = \frac{1}{\Delta\xi}\int_{\xi_{j-\frac12}}^{\xi_{j+\frac12}} f^j_{l,-,p}(x,\xi)d\xi, \tag{5.76}$$

$$f^j_{l,-,p}(x,\xi) = \begin{cases} \widehat{f_0}(-x,\xi-\xi_j+\xi_{J_{j,p}}), & x<0,\\ \widehat{f_0}(-x,\xi), & x\ge 0 \end{cases} \tag{5.77}$$

for $p=1,2$, where $J_{j,p}$ is defined in (5.39) with d_j defined in (4.20).

The exact solutions to (5.74)–(5.77) are given by

$$u^j_{l,-,p}(x,t) = \begin{cases} u^j_{l,-,p}\left(x-\left|\xi_{J_{j,p}}\right|t,0\right), & x<0,\\ u^j_{l,-,p}\left(\left|\frac{\xi_{J_{j,p}}}{\xi_j}\right|x-\left|\xi_{J_{j,p}}\right|t,0\right), & 0\le x<|\xi_j|\,t, \qquad p=1,2.\\ u^j_{l,-,p}\left(x-|\xi_j|\,t,0\right), & x\ge|\xi_j|\,t, \end{cases} \tag{5.78}$$

The upwind schemes for (5.74) are

$$v_i^{n+1,j,p} = \begin{cases} (1-\mu_{J_{j,p}})v_i^{n,j,p} + \mu_{J_{j,p}}v_{i-1}^{n,j,p} & \text{if } x_i<0,\\ (1-\mu_j)v_i^{n,j,p} + \mu_j v_{i-1}^{n,j,p} & \text{if } x_i>0, \end{cases} \tag{5.79}$$

$$v_i^{0,j,p} = \frac{1}{\Delta x}\int_{x_{i-\frac12}}^{x_{i+\frac12}} u^j_{l,-,p}(x,0)dx \tag{5.80}$$

for $p=1,2$.

Define

$$v^j_{l,-,p}(x,t) = v_i^{n,j,p} \qquad \text{for } (x,t)\in[x_{i-\frac12},x_{i+\frac12})\times[t_n,t_{n+1}),\ \ p=1,2.$$

We have the following lemmas.

LEMMA 5.5.

$$\left|\sqrt{\xi^2+2D}+\xi'\right| \le 2\Delta\xi \quad \forall \xi_{j-\frac12}\le\xi\le\xi_{j+\frac12},\ \ \xi_{d_j-\frac12}\le\xi'\le\xi_{d_j+\frac32} \tag{5.81}$$

for $I_b\le j\le I_0$.

This lemma can be directly checked, and the proof is omitted.

LEMMA 5.6. *Define functions*

$$H_\xi^{-,j,p}(x) = \begin{cases} -\left(\frac{\xi_{J_{j,p}}}{\xi_j}x-\xi_{J_{j,p}}t_n-\sqrt{\xi^2+2D}\,t_n\right)\frac{\xi}{\sqrt{\xi^2+2D}}, & x>\xi_j t_n,\\ x-\xi_j t_n+\xi t_n, & x\le\xi_j t_n. \end{cases}$$

Then

$$\left|H_{\xi}^{-,j,p}(x)-x\right| \le \frac{7}{2}\Delta\xi t_n \tag{5.82}$$

for $x \le 0,\ \xi_{j-\frac{1}{2}} \le \xi \le \xi_{j+\frac{1}{2}},\ I_b \le j \le I_0,\ p=1,2.$

This lemma can be proved using a technique similar to that used for proving Lemma 5.3, and we omit the detailed proof.

LEMMA 5.7. *Define functions*

$$G_p^j(\xi) = \int_{x_{\frac{1}{2}}}^{0} \left|\widetilde{f}_{l,-,p}^{j}(x,\xi,t_n) - \widehat{f}_{l,-}(x,\xi,t_n)\right| dx \tag{5.83}$$

for $\xi_{j-\frac{1}{2}} \le \xi \le \xi_{j+\frac{1}{2}},\ I_b \le j \le I_0,\ p=1,2,$ *where*

(5.84)

$$\widetilde{f}_{l,-,p}^{j}(x,\xi,t) = \begin{cases} \widehat{f}_0\left(\frac{\xi_{J_{j,p}}}{\xi_j}x - \xi_{J_{j,p}}t, -\sqrt{\xi^2+2D}\right), & x > \xi_j t, \\ \widehat{f}_0\left(x-\xi_j t,\xi\right), & x \le \xi_j t, \end{cases}$$

(5.85)

$$\widehat{f}_{l,-}(x,\xi,t) = \begin{cases} \widehat{f}_0\left(-\frac{\sqrt{\xi^2+2D}}{\xi}\,x + \sqrt{\xi^2+2D}\,t, -\sqrt{\xi^2+2D}\right), & x > \xi t, \\ \widehat{f}_0\left(x-\xi t,\xi\right), & x \le \xi t. \end{cases}$$

Then

$$G_p^j(\xi) \le \left(2A+\sqrt{2D}B\right) 7\Delta\xi t_n. \tag{5.86}$$

Proof. From definition (5.85) and conditions (4.3), (4.4) one obtains

$$\|\widehat{f}_{l,-}(\cdot,\xi,t)\|_{BV(\mathbb{R})} \le 2A+\sqrt{2D}B. \tag{5.87}$$

We use the fact that

$$\widetilde{f}_{l,-,p}^{j}(x,\xi,t_n) = \widehat{f}_{l,-}\left(H_{\xi}^{-,j,p}(x),\xi,t_n\right), \quad x\in\mathbb{R},\ p=1,2. \tag{5.88}$$

From (5.83), (5.88), and (5.87), applying Lemmas 5.6 and A.1 gives (5.86). ∎

We now give estimates for E_l^-. In comparison with schemes (4.14), (4.17), (4.20), (4.21), (4.23) and (5.79), (5.80), (5.76), (5.77), (4.7), one can check that, for $I_b \le j \le I_0$,

$$g(x,\xi,t_n) = \sum_{p=1}^{2} c_{j,p} v_{l,-,p}^{j}(-x,t_n) \quad \text{a.e. for } \xi_{j-\frac{1}{2}} < \xi < \xi_{j+\frac{1}{2}},\ x_{\frac{1}{2}} < x < 0. \tag{5.89}$$

Using (5.73), (5.89), and the expression (4.11) of $f(x,\xi,t_n)$ in D_l^-, one obtains

$$E_l^- \le E_l^{-,1} + E_l^{-,2} + E_l^{-,3}, \tag{5.90}$$

where

$$E_l^{-,1} = \sum_{j=I_b}^{I_0} \int_{\xi_{j-\frac{1}{2}}}^{\xi_{j+\frac{1}{2}}} \int_{x_{\frac{1}{2}}}^{0} \sum_{p=1}^{2} c_{j,p} \left| v_{l,-,p}^j(-x,t_n) - u_{l,-,p}^j(-x,t_n) \right| dx d\xi, \tag{5.91}$$

$$E_l^{-,2} = \sum_{j=I_b}^{I_0} \int_{\xi_{j-\frac{1}{2}}}^{\xi_{j+\frac{1}{2}}} \int_{x_{\frac{1}{2}}}^{0} \sum_{p=1}^{2} c_{j,p} \left| u_{l,-,p}^j(-x,t_n) - \widetilde{f}_{l,-,p}^j(x,\xi,t_n) \right| dx d\xi, \tag{5.92}$$

$$E_l^{-,3} = \sum_{j=I_b}^{I_0} \int_{\xi_{j-\frac{1}{2}}}^{\xi_{j+\frac{1}{2}}} \int_{x_{\frac{1}{2}}}^{0} \sum_{p=1}^{2} c_{j,p} \left| \widetilde{f}_{l,-,p}^j(x,\xi,t_n) - \widehat{f}_{l,-}(x,\xi,t_n) \right| dx d\xi. \tag{5.93}$$

Applying Theorem 3.1, the l^1-error between $v_{l,-,p}^j$ and $u_{l,-,p}^j$, $p = 1, 2$, is given by

$$\begin{aligned} &\|v_{l,-,p}^j(\cdot,t_n) - u_{l,-,p}^j(\cdot,t_n)\|_{l^1(\mathbb{R})} \\ &\quad \le \|u_{l,-,p}^j(\cdot,0)\|_{BV(\mathbb{R})} \left[\sqrt{\frac{2t_{n+1}}{e}} \frac{|\xi_{J_{j,p}}|}{\sqrt{|\xi_j|}} \sqrt{\Delta x} + \left(2 + \left| \frac{\xi_{J_{j,p}}}{\xi_j} \right| \right) \Delta x \right]. \end{aligned} \tag{5.94}$$

Similar to the derivation of (5.64), one has

$$\|u_{l,-,p}^j(\cdot,0)\|_{BV(\mathbb{R})} \le 2A + \sqrt{2D}B. \tag{5.95}$$

Applying (5.94) and (5.95) to (5.91), by summation of j and using the inequalities (5.65) and (5.70), one obtains

$$\begin{aligned} E_l^{-,1} \le &\left(4A + 2\sqrt{2D}B \right) \sqrt{\frac{2t_{n+1}}{e}} \left| \xi_{\frac{1}{2}} \right| \left[\left(\xi_{\frac{1}{2}} \right)^2 - 2D \right]^{\frac{1}{4}} \sqrt{\Delta x} \\ &+ \left(2A + \sqrt{2D}B \right) \left| \xi_{\frac{1}{2}} \right| \ln \left(\frac{1}{\Delta x} \right) \Delta x + O(\Delta x). \end{aligned} \tag{5.96}$$

Using Lemma 5.5, (5.67), and condition (4.4), from the definitions (5.76), (5.77), (5.78), and (5.84), one has

$$\left| u_{l,-,p}^j(-x,t_n) - \widetilde{f}_{l,-,p}^j(x,\xi,t_n) \right| \le 2B\Delta\xi$$

for $x \le 0$, $\xi_{j-\frac{1}{2}} \le \xi \le \xi_{j+\frac{1}{2}}$.

Thus one has

$$E_l^{-,2} \le \left| \xi_{I_b - \frac{1}{2}} \right| \left| x_{\frac{1}{2}} \right| 2B\Delta\xi. \tag{5.97}$$

Applying Lemma 5.7 to (5.93) yields

$$E_l^{-,3} \le 7 \left(2A + \sqrt{2D}B \right) \left| \xi_{I_b - \frac{1}{2}} \right| t_n \Delta\xi. \tag{5.98}$$

Combining (5.90), (5.96), (5.97), (5.98) leads to

$$\begin{aligned} E_l^{-} \le &\left(4A + 2\sqrt{2D}B \right) \sqrt{\frac{2t_{n+1}}{e}} \left| \xi_{\frac{1}{2}} \right| \left[\left(\xi_{\frac{1}{2}} \right)^2 - 2D \right]^{\frac{1}{4}} \sqrt{\Delta x} \\ &+ \left(2A + \sqrt{2D}B \right) \left| \xi_{\frac{1}{2}} \right| \ln \left(\frac{1}{\Delta x} \right) \Delta x + O(\Delta x). \end{aligned} \tag{5.99}$$

Finally, combining (5.1), (5.20), (5.21), (5.33), (5.72), (5.99), (5.12) completes the proof for Theorem 4.1.

Remark 5.1. The error terms derived in (4.25) include the half order terms and the $\ln\left(\frac{1}{\Delta x}\right)\Delta x$ terms. Thus the leading error terms are of half order, and Scheme I has a half order convergence rate. Note that this is sharp, since even for the discontinuous solution to linear hyperbolic equations with a smooth coefficient, one cannot expect a better convergence order in the l^1-norm [24].

6. Conclusion. In this paper we derived the l^1-error estimates for a Hamiltonian-preserving scheme, developed in [11], for the Liouville equation with a piecewise constant potential in one space dimension. The Hamiltonian-preserving scheme is designed by incorporating into the numerical fluxes the particle behavior—transmission and reflection—at the potential barrier. We proved that, with the Dirichlet incoming boundary conditions and for a class of bounded initial data, the numerical solution by the Hamiltonian-preserving scheme converges in the l^1-norm to the solution of the Liouville equation—defined by using the interface condition that accounts for particle transmission and reflection. The initial data conditions can be satisfied by applying the decomposition technique proposed in [10] for solving the Liouville equation with measure-valued initial data arising in the semiclassical limit of the linear Schrödinger equation.

The strategy for the error analysis in this paper is to apply the l^1-error estimates established in [30, 28] for the immersed interface upwind scheme to the linear advection equations with piecewise constant coefficients. This is a natural approach since the solution of the Liouville equation with a step function potential satisfies linear advection equations with piecewise constant coefficients on the bicharacteristics. To apply this strategy, we split the computational domain into several parts. In each subdomain, we introduced a number of linear advection equations with step function coefficients for which the immersed interface upwind schemes yield the same numerical solution as the Hamiltonian-preserving scheme for the Liouville equation with the step function potential. Then the l^1-error estimates for the Hamiltonian-preserving scheme were derived by applying the l^1-error estimates for the immersed interface upwind schemes and comparing the exact solutions of the linear advection equations and that of the Liouville equation in each of these subdomains. The conditions on the initial data were used in deriving these estimates. As a result, we obtained the half order l^1-error bound with explicit coefficients for the Hamiltonian-preserving scheme.

The problem under study has important applications in the computation of the semiclassical limit of the linear Schrödinger equation through barriers and, more generally, in the computation of high frequency waves through interfaces.

Appendix. In this appendix we prove a property of BV functions on $\mathbb{R}$.

LEMMA A.1. *Let $f(x)$ be a BV function on $\mathbb{R}$ and $H(x)$ be a function on $[a,b]$ satisfying*

$$|H(x)-x| \le H_C \quad \text{for } x \in [a,b],$$

where H_C is a positive constant. Then

$$\|f(\cdot)-f(H(\cdot))\|_{L^1([a,b])} \le 2H_C\|f\|_{BV(\mathbb{R})}. \tag{A.1}$$

Proof. For all $x \in [a,b]$, one has

$$|f(x)-f(H(x))| \le \|f\|_{BV([x-H_C,x+H_C])}.$$

Since $f(x)$ has a bounded variation on $\mathbb{R}$, $f'(x)$ exists a.e., and

$$\|f\|_{BV([x-H_C,x+H_C])} = \int_{x-H_C}^{x+H_C} |f'(y)|\,dy.$$

Thus

$$\|f(\cdot) - f(H(\cdot))\|_{L^1([a,b])} \le \int_a^b \int_{x-H_C}^{x+H_C} |f'(y)|\,dydx. \tag{A.2}$$

Introduce the function

$$g(x,y) = \begin{cases} |f'(y)| & \text{if } |x-y| \le H_C, \\ 0 & \text{else.} \end{cases}$$

Then $g(x,y) \in L^1([a,b] \times [a-H_C, b+H_C])$. Applying Fubini's theorem gives

$$\int_a^b \int_{x-H_C}^{x+H_C} |f'(y)|\,dydx = \int_a^b \int_{a-H_C}^{b+H_C} g(x,y)dydx$$

$$= \int_{a-H_C}^{b+H_C} \int_a^b g(x,y)dxdy \le \int_{a-H_C}^{b+H_C} \int_{y-H_C}^{y+H_C} |f'(y)|\,dxdy$$

$$= 2H_C \int_{a-H_C}^{b+H_C} |f'(y)|\,dy \le 2H_C \|f\|_{BV(\mathbb{R})}. \tag{A.3}$$

Combining (A.2) and (A.3) gives (A.1). ∎

REFERENCES

[1] J. Jaffré Adimurthi and G. D. V. Gowda, *Godunov-type methods for conservation laws with a flux function discontinuous in space*, SIAM J. Numer. Anal., 42 (2004), pp. 179–208.

[2] L. Ambrosio, *Transport equation and Cauchy problem for BV vector fields*, Invent. Math., 158 (2004), pp. 227–260.

[3] F. Bouchut and F. James, *One-dimensional transport equations with discontinuous coefficients*, Nonlinear Anal., 32 (1998), pp. 891–933.

[4] G. M. Coclite and N. H. Risebro, *Conservation laws with time dependent discontinuous coefficients*, SIAM J. Math. Anal., 36 (2005), pp. 1293–1309.

[5] R. J. DiPerna and P.-L. Lions, *Ordinary differential equations, transport theory, and Sobolev spaces*, Invent. Math., 98 (1989), pp. 511–547.

[6] T. Gimse, *Conservation laws with discontinuous flux functions*, SIAM J. Math. Anal., 24 (1993), pp. 279–289.

[7] L. Gosse and F. James, *Numerical approximations of one-dimensional linear conservation equations with discontinuous coefficients*, Math. Comp., 69 (2000), pp. 987–1015.

[8] P. Gérard, P. A. Markowich, N. J. Mauser, and F. Poupaud, *Homogenization limits and Wigner transforms*, Comm. Pure Appl. Math., 50 (1997), pp. 321–377.

[9] T. Gimse and N. H. Risebro, *Solution of the Cauchy problem for a conservation law with a discontinuous flux function*, SIAM J. Math. Anal., 23 (1992), pp. 635–648.

[10] S. Jin, H. L. Liu, S. Osher, and R. Tsai, *Computing multivalued physical observables for the semiclassical limit of the Schrödinger equation*, J. Comput. Phys., 205 (2005), pp. 222–241.

[11] S. Jin and X. Wen, *The Hamiltonian-preserving schemes for the Liouville equation with discontinuous potentials*, Commun. Math. Sci., 3 (2005), pp. 285–315.

[12] S. Jin and X. Wen, *A Hamiltonian-preserving scheme for the Liouville equation of geometrical optics with partial transmissions and reflections*, SIAM J. Numer. Anal., 44 (2006), pp. 1801–1828.

[13] K. H. Karlsen, C. Klingenberg, and N. H. Risebro, *A relaxation scheme for conservation laws with a discontinuous coefficient*, Math. Comp., 73 (2004), pp. 1235–1259.
[14] C. Klingenberg and N. H. Risebro, *Convex conservation laws with discontinuous coefficients. Existence, uniqueness and asymptotic behavior*, Comm. Partial Differential Equations, 20 (1995), pp. 1959–1990.
[15] C. Klingenberg and N. H. Risebro, *Stability of a resonant system of conservation laws modeling polymer flow with gravitation*, J. Differential Equations, 170 (2001), pp. 344–380.
[16] K. H. Karlsen, N. H. Risebro, and J. D. Towers, *Upwind difference approximations for degenerate parabolic convection-diffusion equations with a discontinuous coefficient*, IMA J. Numer. Anal., 22 (2002), pp. 623–664.
[17] K. H. Karlsen and J. D. Towers, *Convergence of the Lax-Friedrichs scheme and stability for conservation laws with a discontinuous space-time dependent flux*, Chinese Ann. Math. Ser. B, 25 (2004), pp. 287–318.
[18] P. L. Lions and T. Paul, *Sur les measures de Wigner*, Rev. Mat. Iberoamericana, 9 (1993), pp. 553–618.
[19] L. Lin, J. B. Temple, and J. Wang, *A comparison of convergence rates for Godunov's method and Glimm's method in resonant nonlinear systems of conservation laws*, SIAM J. Numer. Anal., 32 (1995), pp. 824–840.
[20] S. Mishra, *Convergence of upwind finite difference schemes for a scalar conservation law with indefinite discontinuities in the flux function*, SIAM J. Numer. Anal., 43 (2005), pp. 559–577.
[21] D. N. Ostrov, *Viscosity solutions and convergence of monotone schemes for synthetic aperture radar shape-from-shading equations with discontinuous intensities*, SIAM J. Appl. Math., 59 (1999), pp. 2060–2085.
[22] B. Perthame and C. Simeoni, *A kinetic scheme for the Saint-Venant system with a source term*, Calcolo, 38 (2001), pp. 201–231.
[23] F. Poupaud and M. Rascle, *Measure solutions to the linear multi-dimensional transport equation with non-smooth coefficients*, Comm. Partial Differential Equations, 22 (1997), pp. 337–358.
[24] T. Tang and Z. H. Teng, *The sharpness of Kuznetsov's $O(\sqrt{\Delta x})$ L1-error estimate for monotone difference schemes*, Math. Comp., 64 (1995), pp. 581–589.
[25] B. Temple, *Global solution of the Cauchy problem for a class of 2×2 nonstrictly hyperbolic conservation laws*, Adv. in Appl. Math., 3 (1982), pp. 335–375.
[26] J. D. Towers, *Convergence of a difference scheme for conservation laws with a discontinuous flux*, SIAM J. Numer. Anal., 38 (2000), pp. 681–698.
[27] J. D. Towers, *A difference scheme for conservation laws with a discontinuous flux: The nonconvex case*, SIAM J. Numer. Anal., 39 (2001), pp. 1197–1218.
[28] X. Wen, *Convergence of an Immersed Interface Upwind Scheme for Linear Advection Equations with Piecewise Constant Coefficients* II: *Some Related Binomial Coefficient Inequalities*, preprint, 2008.
[29] X. Wen and S. Jin, *The l^1-stability of a Hamiltonian-preserving scheme for the Liouville equation with discontinuous potentials*, J. Comput. Math., to appear.
[30] X. Wen and S. Jin, *Convergence of an immersed interface upwind scheme for linear advection equations with piecewise constant coefficients* I: *L^1-error estimates*, J. Comput. Math., 26 (2008), pp. 1–22.

SIAM J. NUMER. ANAL.
Vol. 46, No. 5, pp. 2715–2741
© 2008 Society for Industrial and Applied Mathematics

AN ERROR ESTIMATE FOR A NEW SCHEME FOR MEAN CURVATURE MOTION*

N. FORCADEL†

Abstract. In this work, we propose a new numerical scheme for the anisotropic mean curvature equation. The solution of the scheme is not unique, but for all numerical solutions, we provide an error estimate between the continuous solution and the numerical approximation. Our scheme is also applicable to computing the solution to dislocation dynamics equations. We also provide some numerical simulations.

Key words. mean curvature motion, error estimate, numerical scheme, dislocation dynamics

AMS subject classifications. 65M06, 65M12, 65M15, 49L25, 53C44

DOI. 10.1137/060678282

1. Introduction.

Mean curvature motion has been largely studied in terms of both theory and computation, in particular due to the large number of applications such as front propagation, image processing, and fluid dynamics (see, for instance, Sethian [30] and Osher and Paragios [27]).

The level set framework has been used in both theoretical and numerical problems. Theoretically, the mean curvature equation has been well understood using the framework of viscosity solutions by Chen, Giga, and Goto [10] and Evans and Spruck [20]. However, this equation has serious problems for the question of numerical approximations. Nevertheless, there are several works on this question. First, let us mention the numerical method of Osher and Sethian [28]. This method is often used in practice, but, as far as we know, there is no convergence result. Another algorithm is the Merriman, Bence, and Osher scheme [25] in which motion by mean curvature is viewed as the singular limit of a diffusion equation with threshold. The convergence of this scheme has been proved by Barles and Georgelin [4] and by Evans [19] (see also Ishii [21]; Ishii, Pires, and Souganidis [22]; and Chambolle and Novaga [9]). A class of convergent schemes for nonlinear parabolic equations including mean curvature motion has been proposed by Crandall and Lions [12] (see also Deckelnick [14] for an error analysis of this scheme). Let us mention also the recent work of Oberman [26]. In the last two works, two different scales are used. The first is the space step Δx, and the second is the size of the stencil ε. As pointed out in [12], these two scales are very important in approximating degenerate equations such as the mean curvature equation. Some other approaches are also used to approximate the mean curvature motion. Let us mention the parametrization of surfaces used by Deckelnick and Dziuk [15, 16]. We also refer to the work of Walkington [33] in which the author used a finite element algorithm to approximate the level set solution of the mean curvature motion.

An overview of the above methods together with an extensive list of references can be found in the survey articles by Elliott [18] and Deckelnick, Dziuk, and Elliott [17].

*Received by the editors December 21, 2006; accepted for publication (in revised form) March 26, 2008; published electronically July 2, 2008. This research was supported by contract JC 1025, "ACI jeunes chercheuses et jeunes chercheurs," of the French Ministry of Research (2003-2006).
http://www.siam.org/journals/sinum/46-5/67828.html
†CERMICS, Ecole des Ponts, Paris Tech, 6 et 8 avenue Blaise Pascal, Cité Descartes, Champs-sur-Marne, 77455 Marne-la-Vallée Cedex 2, France (forcadel@cermics.enpc.fr).

The goal of our work is to propose a new scheme for mean curvature motion and to prove an error estimate between the continuous solution and its numerical approximation. This error estimate is not optimal, but as far as we know, it is the first one for a complete discretized scheme for mean curvature type motion.

The idea is to use a recent work of Da Lio, Forcadel, and Monneau [13] concerning the convergence of dislocation dynamics to mean curvature motion. Dislocations are linear defects which move in crystals. Their dynamics can be represented by a nonlocal first order Hamilton–Jacobi equation (see Alvarez et al. [3]). The first goal of this work is to prove an error estimate between the solution of dislocation dynamics and the solution of mean curvature motion. To do this, we will use in a more quantitative way the definition of viscosity solutions for mean curvature motion proposed by Barles and Georgelin [4] by considering a regularization with quartics.

The second goal of this work is to propose a numerical scheme for dislocation dynamics. The main properties of this scheme are that it is implicit (so there is no CFL condition) and that it uses two different scales. Moreover, this scheme is not monotone and does not admit a unique solution. It is only "almost monotone" (see Lemma 5.2). Nevertheless, the fact that it is implicit and the velocity is monotone allows us to "freeze" the velocity and to prove a Crandall–Lions-type [11] error estimate for any solutions (we refer to Alvarez et al. [1, 2] for error estimates for dislocation dynamics in the nonmonotone case). It is also possible to make the scheme explicit. In this case, we are able to prove the convergence of the scheme under a CFL condition, but error estimates in this case are still open. This comes from the fact that there is no consistency error for the scheme (see Proposition 5.6).

Finally, let us mention some works on error estimates for numerical schemes for Hamilton–Jacobi equations. For the first order case, we refer to Souganidis [32]. For the second order case, when the Hamiltonian is convex, we refer to Krylov [23, 24] and Barles and Jakobsen [5, 6]. When the Hamiltonian is uniformly elliptic, we refer to Caffarelli and Souganidis [8].

Let us now explain how this paper is organized. In section 2, we state the main results of this work. In section 3, we prove the main result concerning mean curvature-type motion. Section 4 is devoted to proving the error estimate between the solution of dislocation dynamics and that of mean curvature motion. In section 5, we study the numerical scheme for dislocation dynamics and we prove an error estimate between the continuous solution and its numerical approximation. Some numerical simulations are provided in section 6.

2. Main results.

2.1. Error estimate for mean curvature motion. In order to propose a numerical scheme for anisotropic mean curvature motion, we will use the work of Da Lio, Forcadel, and Monneau [13] that we briefly recall here. Given a function g defined on the unit sphere $\mathbf{S}^{N-1}$of $\mathbb{R}^N$ by

$$g \in \mathrm{Lip}(\mathbf{S}^{N-1}), \quad g(-\theta) = g(\theta) \geq 0 \quad \forall \theta \in \mathbf{S}^{N-1}, \tag{2.1}$$

we consider kernels $c_0 \in L^\infty(\mathbb{R}^N) \cap W^{1,1}(\mathbb{R}^N)$ satisfying

$$\begin{cases} c_0(x) = \dfrac{1}{|x|^{N+1}}\ g\left(\dfrac{x}{|x|}\right) & \text{if} \quad |x| \geq 1, \\ c_0(-x) = c_0(x) \geq 0 \quad \forall x \in \mathbb{R}^N, \end{cases} \tag{2.2}$$

and we use the following scaling for $0 < \varepsilon < \frac{1}{e}$:

$$c_0^\varepsilon(x) = \frac{1}{\varepsilon^{N+1}|\ln\varepsilon|}\ c_0\left(\frac{x}{\varepsilon}\right). \tag{2.3}$$

The term $|\ln\varepsilon|$ comes from the bad decay at infinity of the kernel c_0 (see [13, section 4.1]).

We then consider the auxiliary problem

$$\begin{cases} u_t^\varepsilon(x,t) = \left((c_0^\varepsilon \star 1_{\{u^\varepsilon(\cdot,t)\geq u^\varepsilon(x,t)\}})(x) - \dfrac{1}{2}\displaystyle\int_{\mathbb{R}^N} c_0^\varepsilon\right)|Du^\varepsilon(x,t)| & \text{in} \quad \mathbb{R}^N\times(0,T), \\ u^\varepsilon(\cdot,0) = u_0^\varepsilon & \text{in} \quad \mathbb{R}^N, \end{cases} \tag{2.4}$$

where u_t^ε denotes the derivative with respect to the time variable, Du^ε indicates the gradient of u^ε with respect to the space variables, the convolution is done in space only, and $1_{\{u^\varepsilon(\cdot,t)\geq u^\varepsilon(x,t)\}}$ is the characteristic function of the set $\{u^\varepsilon(\cdot,t)\geq u^\varepsilon(x,t)\}$ (which is equal to 1 on the set and 0 outside). This equation arises in the theory of dislocation dynamics (see Alvarez et al. [3]) and uses the Slepčev formulation [31] to consider the simultaneous evolutions of all the level sets of the function u^ε. We refer to Da Lio, Forcadel, and Monneau [13] for the study of this model and in particular to [13, Definition 2.1] for the definition of viscosity solutions for problem (2.4). In particular, we recall that we take the indicatrix of $\{u^\varepsilon(\cdot,t)\geq u^\varepsilon(x,t)\}$ for the subsolution and the indicatrix of $\{u^\varepsilon(\cdot,t) > u^\varepsilon(x,t)\}$ for the supersolution.

The main result of [13] is that, if $u_0^\varepsilon = u_0$, then the unique solution u^ε of (2.4) converges locally uniformly on compact sets to the solution u^0 of the limit equation

$$\begin{cases} u_t^0 - \text{trace}\ \left(D^2u^0\cdot A\left(\dfrac{Du^0}{|Du^0|}\right)\right) = 0 & \text{in} \quad \mathbb{R}^N\times(0,T), \\ u^0(\cdot,0) = u_0 & \text{in} \quad \mathbb{R}^N, \end{cases} \tag{2.5}$$

with

$$A\left(\frac{p}{|p|}\right) = \int_{\theta\in\mathbf{S}^{N-2}=\mathbf{S}^{N-1}\cap\left\{\langle x,\frac{p}{|p|}\rangle=0\right\}} \left(\frac{1}{2}g(\theta)\ \theta\otimes\theta\right)\ d\theta \tag{2.6}$$

and where $M\cdot A$ and $\langle\cdot,\cdot\rangle$ denote, respectively, the product between the two matrices and the usual scalar product.

Remark 2.1. In the particular case where $g\equiv 1$, we get $A = \frac{|\mathbf{S}^{N-2}|}{2(N-1)}Id_{\{x,\ <x,p>=0\}}$, where $|\mathbf{S}^{N-2}|$ is the Lebesgue measure of $\mathbf{S}^{N-2}$, and then

$$\text{trace}\left(M\cdot A\left(\frac{p}{|p|}\right)\right) = \frac{-|\mathbf{S}^{N-2}|}{2(N-1)}\ \text{trace}\left(\left(Id - \frac{p}{|p|}\otimes\frac{p}{|p|}\right)\cdot M\right).$$

We recover the classical mean curvature motion up to the factor $|\mathbf{S}^{N-2}|/2(N-1)$.

Our first main result is an error estimate between u^ε and u^0.

THEOREM 2.2 (error estimate for mean curvature motion). *Let $N\geq 1$ and $T\leq 1$. Assume that $u_0, u_0^\varepsilon\in \text{Lip}(\mathbb{R}^N)$, $g\in\text{Lip}(\mathbf{S}^{N-1})$, and that c_0 given in* (2.2) *satisfies $c_0\in W^{1,1}(\mathbb{R}^N)$. Then, there exists a constant K_1 depending on $N, \sup_{\mathbb{R}^N} c_0$,*

$|Dg|_{L^\infty(\mathbf{S}^{N-1})}$, and $|Du_0|_{L^\infty(\mathbb{R}^N)}$ such that the difference between the solution u^ε of (2.4) *and the solution u^0 of* (2.5) *is given by*

$$\sup_{\mathbb{R}^N\times(0,T)} |u^\varepsilon - u^0| \leq K_1 \left(\frac{T}{|\ln\varepsilon|}\right)^{\frac{1}{6}} + \sup_{\mathbb{R}^N} |u_0^\varepsilon - u_0|. \tag{2.7}$$

2.2. Discrete-continuous error estimate for dislocation dynamics. To approximate the solution u^0 of (2.5), we then have to approximate the solution u^ε of (2.4). Up to a change of variable (see Corollary 3.1), it suffices to approximate the solution u of

(2.8)
$$\begin{cases} u_t(x,t) = \left((c_0 \star 1_{\{u(\cdot,t)\geq u(x,t)\}})(x) - \dfrac{1}{2}\displaystyle\int c_0\right)|Du(x,t)| & \text{in} \quad \mathbb{R}^N\times(0,T), \\ u(\cdot,0) = \bar{u}_0 & \text{in} \quad \mathbb{R}^N. \end{cases}$$

Given a mesh size $\Delta x_1,\dots,\Delta x_N,\Delta t$ and a lattice $Q_T^\Delta = Q^\Delta\times\{0,\dots,(\Delta t)N_T\}$, where $Q^\Delta = \{(i_1\Delta x_1,\dots,i_N\Delta x_N),\, I=(i_1,\dots,i_N)\in\mathbb{Z}^N\}$ and N_T is the integer part of $T/\Delta t$, we will denote by $(x_1,\dots,x_N,t_n) = (x_I,t_n)$ the node $(i_1\Delta x_1,\dots,i_N\Delta x_N,$ $n\Delta t)$ and by v_I^n the value of a numerical approximation of the exact solution $u(x_I,t_n)$. We set $\Delta x = \sqrt{\Delta x_1^2+\cdots+\Delta x_N^2}$ the space mesh size. We shall assume throughout that $\Delta x + \Delta t \leq 1$.

The discrete solution v is computed iteratively by solving the implicit scheme

$$v_I^0 = \tilde{u}_0(x_I), \qquad \frac{v_I^{n+1}-v_I^n}{\Delta t} = c^\Delta[v]_I^{n+1} G(v^{n+1})_I, \tag{2.9}$$

where $\tilde{u}_0$ is an approximation of $\bar{u}_0$ and $G(v^{n+1})_I$ is a suitable approximation of the gradient of v^{n+1} taken at point x_I. The nonlocal velocity is the discrete convolution

$$c^\Delta[v]_I^{n+1} = c^\Delta[v](x_I,t_{n+1}) = \sum_{J\in\mathbb{Z}^N} \bar{c}^0_{I-J} 1_{\{v_J^{n+1}\geq v_I^{n+1}\}}\Delta x_1,\dots,\Delta x_N \tag{2.10}$$
$$-\frac{1}{2}\sum_{J\in\mathbb{Z}^N} \bar{c}^0_J \Delta x_1,\dots,\Delta x_N$$

with

$$\bar{c}^0_I = \frac{1}{|Q_I|}\int_{Q_I} c_0(x)\,dx, \tag{2.11}$$

where Q_I is the square cell centered in x_I,

$$Q_I = [x_{i_1} - \Delta x_1/2, x_{i_1} + \Delta x_1/2]\times\cdots\times[x_{i_N} - \Delta x_N/2, x_{i_N} + \Delta x_N/2]. \tag{2.12}$$

Finally, let us define

$$v_\#(y,t_n) = \sum_I v(x_I,t_n)\chi_{Q_I}(y), \tag{2.13}$$

where χ_{Q_I} is the indicator function of Q_I.

The approximation of the gradient is obtained using the Osher–Sethian scheme [28] (we can also use the one proposed by Rouy and Tourin [29]). It is monotone, consistent and depends on the sign of the nonlocal velocity. Its precise definition is recalled in section 5.

Since the velocity c^Δ is nonlocal and not continuous, we have to give a sense to the equality in the scheme (2.9). In fact, we will use the analogue of the Slepčev formulation [31] for discrete sub- and supersolutions (see Definition 5.1) and we will use a discrete version of Perron's method to construct a discrete solution. The solution of the scheme is not unique, but, for any solutions, we have the following Crandall–Lions-type [11] error estimate.

THEOREM 2.3 (discrete-continuous error estimate for (2.8)). *Let $N \geq 1$ and $T \leq 1$. Assume that $\bar{u}_0, \tilde{u}_0 \in W^{1,\infty}(\mathbb{R}^N)$ and that c_0 given in (2.2) satisfies $c_0 \in W^{1,1}(\mathbb{R}^N)$. Assume that $\Delta x + \Delta t \leq 1$. Then there exists a constant $K_2 > 0$ depending only on N, $|c_0|_{W^{1,1}(\mathbb{R}^N)}$, $|D\bar{u}_0|_{L^\infty(\mathbb{R}^N)}$, and $|D\tilde{u}_0|_{L^\infty(\mathbb{R}^N)}$ such that the error estimate between the continuous solution u of (2.8) and any discrete solution v of the finite difference scheme (2.9) is given by*

$$\sup_{\mathbb{R}^N \times \{0,\dots,t_{N_T}\}} |u - v_\#| \leq K_2 \sqrt{T}\,(\Delta x + \Delta t)^{1/2} + \sup_{\mathbb{R}^N} |\bar{u}_0 - (\tilde{u}_0)_\#|,$$

provided that $\Delta x + \Delta t \leq \frac{1}{K_2^2}$.

Remark 2.4. It is possible to make the computation of the gradient explicit, i.e., to replace the term $G(v^{n+1})_I$ by $G(v^n)_I$ in the scheme (2.9) and to consider the solution v of

$$v_I^0 = \tilde{u}_0(x_I), \qquad \frac{v_I^{n+1} - v_I^n}{\Delta t} = c^\Delta[v]_I^{n+1} G(v^n)_I. \tag{2.14}$$

In this case, as usual, we have to satisfy a CFL condition, for instance,

$$\Delta t \leq \frac{\Delta x}{2|c_0|_{L^1(\mathbb{R}^N)}},$$

for the Osher–Sethian discretization of the gradient. Under this additional assumption, Theorem 2.3 remains true with v the solution of the scheme (2.14).

2.3. Discrete-continuous error estimate for mean curvature motion. Using the above results, in section 3 we will prove the following theorem.

THEOREM 2.5 (discrete-continuous error estimate for the mean curvature motion). *Let $N \geq 1$ and $T \leq 1$. Let us denote by v^ε a solution of (2.9)–(2.11) with initial condition $\tilde{u}_0^\varepsilon$ (which is an approximation of u_0) and with c_0^ε in the place of c_0. Assume that $u_0, \tilde{u}_0^\varepsilon \in W^{1,\infty}(\mathbb{R}^N)$, $g \in \mathrm{Lip}(\mathbf{S}^{N-1})$, and that c_0 given in (2.2) satisfies $c_0 \in W^{1,1}(\mathbb{R}^N)$. Assume also that $\Delta x + \Delta t \leq 1$. Then there exists a constant $K_3 > 0$ depending only on N, $\sup_{\mathbb{R}^N} c_0$, $|Dg|_{L^\infty(\mathbf{S}^{N-1})}$, $|c_0|_{W^{1,1}(\mathbb{R}^N)}$, $|Du_0|_{L^\infty(\mathbb{R}^N)}$, and $|D\tilde{u}_0^\varepsilon|_{L^\infty(\mathbb{R}^N)}$ such that the error estimate between the continuous solution u^0 of (2.5) and its numerical approximation v^ε is given by*

$$\sup_{\mathbb{R}^N \times \{0,..,t_{N_T}\}} |u^0 - v_\#^\varepsilon| \leq K_3 \left(\frac{T}{|\ln \varepsilon|}\right)^{1/6} + \sup_{\mathbb{R}^N} |u_0 - (\tilde{u}_0^\varepsilon)_\#|,$$

where $\varepsilon \geq K_3(\Delta x + \sqrt{\Delta t})$.

Remark 2.6. If $T \geq 1$, since $|Du^0|_{L^\infty(\mathbb{R}^N\times(0,T))} \leq |Du_0|_{L^\infty(\mathbb{R}^N)}$ and "$|Dv^\varepsilon|_{L^\infty(\mathbb{R}^N)}$" $\leq |D\tilde{u}_0^\varepsilon|_{L^\infty(\mathbb{R}^N)}$ (see Proposition 5.4 for the exact setting), we can iterate the process and get a linear estimate in T, i.e.,

$$\sup_{\mathbb{R}^N\times\{0,..,t_{N_T}\}} |u^0 - v^\varepsilon_\#| \leq \frac{K_3}{|\ln\varepsilon|^{1/6}} T + \sup_{\mathbb{R}^N} |u_0 - (\tilde{u}_0^\varepsilon)_\#|.$$

Remark 2.7. We can truncate the kernel c_0 at infinity and consider

$$\tilde{c}_0^R(x) = \begin{cases} c_0(x) & \text{if } |x| \leq R, \\ 0 & \text{else.} \end{cases}$$

In this case, we make an error of order $\int_{\mathbb{R}^N\setminus B_R(0)} c_0 \leq \frac{K}{R}$ and we can make the computation on a finite stencil, even if Δx goes to zero. This is possible if we choose ε of the same order as Δx. The condition $\varepsilon \geq K(\Delta x + \sqrt{\Delta t})$ in Theorem 2.5 then implies that we need to impose a CFL condition $\Delta t \leq K\Delta x^2$ (which is classical for second order equations).

On the other hand, if we do not impose any CFL condition, we can choose Δt larger than Δx^2, but we have to choose ε of the same order of $\sqrt{\Delta t}$ and thus make the convolution on larger and larger stencils as Δx goes to zero.

As we mention in the introduction, it is also possible to make the computation of the velocity explicit in the scheme (2.9) and to consider

$$v_I^0 = \tilde{u}_0(x_I), \qquad \frac{v_I^{n+1} - v_I^n}{\Delta t} = c^\Delta[v]_I^n G(v^n)_I, \tag{2.15}$$

with c_0^ε in the place of c_0 in the definition of c^Δ.

We can prove the convergence of the scheme (see Theorem 2.8) under the CFL condition

$$\Delta t \leq \frac{\varepsilon|\ln\varepsilon|}{2|c_0|_{L^1(\mathbb{R}^N)}}\Delta x. \tag{2.16}$$

But, in this case, we are not able to prove an error estimate. The reason is that when we make the computation of the velocity implicit, it "freezes" it, and then we can use the consistency error on the scheme with given velocity. On the other hand, for the explicit scheme, we have to control the consistency error of the velocity, which is not possible, as we point out in Proposition 5.6.

We define

$$v^{\varepsilon,\delta}(x,t) = v^\varepsilon(x_I, t_n) \quad \text{if} \quad x \in Q_I,\ t \in [t_n, t_{n+1}),$$

where $\delta = (\Delta x, \Delta t)$ and v^ε is the solution of the scheme (2.15), (2.10), (2.11) with the kernel c_0^ε in the place of c_0. Then we have the following convergence result.

THEOREM 2.8 (convergence of the explicit scheme). *Assume* (2.16). *Under the assumptions of Theorem* 2.5, *the function $v^{\varepsilon,\delta}$ converges locally uniformly on compact sets as $\delta \to 0$ to the solution u^ε of* (2.4).

Remark 2.9. In Theorem 2.8, if we take the limit $\delta \to 0$ and $\varepsilon \to 0$ with $\delta \ll \varepsilon$, we will approach the solution of (2.5). The condition $\delta \ll \varepsilon$ implies that, in practice, we have to make the convolution in larger and larger stencils as $\delta \to 0$.

Remark 2.10. In the scheme (2.15), it is possible to make the computation of the gradient implicit and thus to withdraw the CFL condition (2.16) to get the same result as Theorem 2.8.

Notation. In what follows, we will denote by K a generic constant, which will then satisfy $K + K = K$, $K \cdot K = K$, and so on.

3. Numerical scheme for mean curvature motion.

3.1. Proof of Theorem 2.5. Using a rescaling argument, we will prove the following corollary of Theorem 2.3.

COROLLARY 3.1 (discrete-continuous error estimate for (2.4)). *Let us denote by v^ε a solution of* (2.9)–(2.11) *with initial condition $\tilde{u}_0^\varepsilon$ (which is an approximation of u_0^ε) and with c_0^ε in the place of c_0. Under the assumptions of Theorem* 2.3, *there exists a constant $K > 0$ depending only on $|c_0|_{W^{1,1}(\mathbb{R}^N)}$, $|Du_0^\varepsilon|_{L^\infty(\mathbb{R}^N)}$, and $|D\tilde{u}_0^\varepsilon|_{L^\infty(\mathbb{R}^N)}$ such that the error estimate between the continuous solution u^ε of* (2.4) *and its numerical approximation v^ε is given by*

$$\sup_{\mathbb{R}^N\times\{0,\dots,t_{N_T}\}} |u^\varepsilon - v_\#^\varepsilon| \le K\frac{1}{\varepsilon}\sqrt{\frac{T}{|\ln\varepsilon|}}\sqrt{\varepsilon\Delta x + \frac{\Delta t}{|\ln\varepsilon|}} + \sup_{\mathbb{R}^N}|u_0^\varepsilon - (\tilde{u}_0^\varepsilon)_\#|,$$

provided that $\varepsilon\Delta x + \dfrac{\Delta t}{|\ln\varepsilon|} \le \dfrac{\varepsilon^2}{K^2}$.

Proof of Corollary 3.1. First we remark that a simple change of variable gives (for u and v solutions of (2.8) and (2.9), respectively)

$$u^\varepsilon(y,\tau) = \varepsilon u\left(\frac{y}{\varepsilon}, \frac{\tau}{\varepsilon^2|\ln\varepsilon|}\right), \qquad u_0^\varepsilon(y) = \varepsilon\bar{u}_0\left(\frac{y}{\varepsilon}\right),$$

$$v^\varepsilon(y,\tau) = \varepsilon v\left(\frac{y}{\varepsilon}, \frac{\tau}{\varepsilon^2|\ln\varepsilon|}\right), \qquad \tilde{u}_0^\varepsilon(y) = \varepsilon\tilde{u}_0\left(\frac{y}{\varepsilon}\right).$$

Let us denote $x = \dfrac{y}{\varepsilon}$, $t = \dfrac{\tau}{\varepsilon^2|\ln\varepsilon|}$, $\Delta x = \dfrac{\Delta y}{\varepsilon}$, $\Delta t = \dfrac{\Delta\tau}{\varepsilon^2|\ln\varepsilon|}$, $T = \dfrac{\Gamma}{\varepsilon^2|\ln\varepsilon|}$, $s_n = n\Delta\tau$, and by N_Γ the integer part of $\Gamma/\Delta\tau$. Then, the following inequality holds:

$$\begin{aligned}\sup_{R^N\times\{0,\dots,s_{N_\Gamma}\}} |u^\varepsilon - v_\#^\varepsilon| &= \varepsilon \sup_{R^N\times\{0,\dots,t_{N_T}\}} |u - v_\#| \\ &\le K\frac{\sqrt{\Gamma}}{\sqrt{|\ln\varepsilon|}}\sqrt{\frac{\Delta y}{\varepsilon} + \frac{\Delta\tau}{\varepsilon^2|\ln\varepsilon|}} + \varepsilon\sup_{\mathbb{R}^N}|\bar{u}_0 - (\tilde{u}_0)_\#| \\ &\le K\frac{\sqrt{\Gamma}}{\varepsilon\sqrt{|\ln\varepsilon|}}\sqrt{\varepsilon\Delta y + \frac{\Delta\tau}{|\ln\varepsilon|}} + \sup_{\mathbb{R}^N}|u_0^\varepsilon - (\tilde{u}_0^\varepsilon)_\#|,\end{aligned}$$

provided that

$$\frac{\Delta y}{\varepsilon} + \frac{\Delta\tau}{\varepsilon^2|\ln\varepsilon|} \le \frac{1}{K^2},$$

i.e.,

$$\varepsilon\Delta y + \frac{\Delta\tau}{|\ln\varepsilon|} \le \frac{\varepsilon^2}{K^2}.$$

This ends the proof of the corollary. ∎

Proof of Theorem 2.5. The proof of Theorem 2.5 is now very easy. Indeed, using Theorem 2.2 and Corollary 3.1, we deduce that

$$|u^0 - v^\varepsilon| \le |u^0 - u^\varepsilon| + |u^\varepsilon - v^\varepsilon|$$
$$\le K\left(\frac{T}{|\ln\varepsilon|}\right)^{\frac{1}{6}} + \frac{K}{\varepsilon\sqrt{|\ln\varepsilon|}}\sqrt{T}\sqrt{\varepsilon\Delta x + \Delta t} + \sup_{\mathbb{R}^N}|u_0^\varepsilon - (\tilde{u}_0^\varepsilon)_\#| + \sup_{\mathbb{R}^N}|u_0^\varepsilon - u_0|.$$

Taking $\varepsilon \ge K(\Delta x + \sqrt{\Delta t})$ implies the prevalence of the first term with respect to the second, and so this implies the desired estimate for the choice of initial condition $u_0^\varepsilon = u_0$.

3.2. Proof of Theorem 2.8. The idea of the proof is borrowed from Barles and Souganidis [7]. Let us set

$$\overline{v}^\varepsilon = \limsup_{\delta\to 0}{}^*\, v^{\varepsilon,\delta}, \quad \underline{v}^\varepsilon = \liminf_{\delta\to 0}{}_*\, v^{\varepsilon,\delta}.$$

We now prove that $\overline{v}^\varepsilon$ and $\underline{v}^\varepsilon$ are, respectively, the sub- and the supersolution of (2.4). Let $\varphi \in C^\infty(\mathbb{R}^N \times [0,T))$ such that $\overline{v}^\varepsilon - \varphi$ reaches a strict maximum at the point (x_0,t_0) with $t_0 \in (0,T)$. Then, there exists $\delta \to 0$, $(x^\delta,t^\delta) \to (x_0,t_0)$ such that $v^{\varepsilon,\delta} - \varphi$ reaches a maximum at (x^δ,t^δ). We denote by $(x_I^\delta,t_n^\delta) \in Q_T^\Delta$ the node such that $v^{\varepsilon,\delta}(x^\delta,t^\delta) = v^{\varepsilon,\delta}(x_I^\delta,t_n^\delta)$. This implies that, for all (x,t),

$$v^{\varepsilon,\delta}(x,t) \le v^{\varepsilon,\delta}(x_I^\delta,t_n^\delta) - \varphi(x^\delta,t^\delta) + \varphi(x,t).$$

Using the monotony of the scheme for a given velocity and the fact that v is a solution of the scheme, we get

$$\frac{\varphi(x^\delta,t^\delta) - \varphi(x^\delta,t^\delta - \Delta t)}{\Delta t}$$
$$\le c^\varepsilon[v^{\varepsilon,\delta}](x_I^\delta,t^\delta - \Delta t)\begin{cases} G^+(D^+\varphi(x^\delta,t^\delta-\Delta t), D^-\varphi(x^\delta,t^\delta-\Delta t)) \\ \qquad \text{if} \quad c[v^{\varepsilon,\delta}](x_I^\delta,t^\delta-\Delta t) \ge 0, \\ G^-(D^+\varphi(x^\delta,t^\delta-\Delta t), D^-\varphi(x^\delta,t^\delta-\Delta t)) \\ \qquad \text{if} \quad c[v^{\varepsilon,\delta}](x_I^\delta,t^\delta-\Delta t) \le 0 \end{cases}$$

with

$$c^\varepsilon[u](x,t) = c_0^\varepsilon \star 1_{\{u(\cdot,t)\ge u(x,t)\}}(x) - \frac{1}{2}\int_{\mathbb{R}^N} c_0^\varepsilon,$$

where we have used Lemma 5.5 to obtain the velocity. Sending $\delta \to 0$, and using the Slepčev lemma [31, eq. (5)] and the consistency of the scheme for a given velocity, we get

$$\varphi_t(x_0,t_0) \le \left(c_0^\varepsilon \star 1_{\{\overline{v}^\varepsilon(\cdot,t_0)\ge \overline{v}^\varepsilon(x_0,t_0)\}}(x_0) - \frac{1}{2}\int_{\mathbb{R}^N} c_0^\varepsilon\right)|D\varphi(x_0,t_0)|.$$

This proves that $\overline{v}^\varepsilon$ is a subsolution of (2.4). The proof for $\underline{v}^\varepsilon$ is the same, and we skip it. Moreover, using a barrier argument (using the equivalent of Proposition 5.4), we get that $\overline{v}^\varepsilon(\cdot,0) = \underline{v}^\varepsilon(\cdot,0) = u_0^\varepsilon$. Then we have that $v^\varepsilon = \overline{v}^\varepsilon = \underline{v}^\varepsilon$ is solution of (2.4). This ends the proof of the theorem.

Remark 3.2. In the scheme (2.15), we have made the choice to take the velocity $c^{\Delta}[v]_I^n$. Another possibility is to take

$$\tilde{c}^{\Delta}[v]_I^{n+1} = \sum_{J\in\mathbb{Z}^N} \bar{c}^0_{I-J} 1_{\{v_J^{n+1} > v_I^{n+1}\}} \Delta x_1, \dots, \Delta x_N - \frac{1}{2} \sum_{J\in\mathbb{Z}^N} \bar{c}^0_J \Delta x_1, \dots, \Delta x_N.$$

This constructs two discrete solutions v and $\tilde{v}$. Thus, we can take any discrete function contained between $\tilde{v}$ and v. In fact, this is equivalent to defining a discrete solution as a sub- and a supersolution as in Definition 5.1. With this definition, v will be the greater subsolution and $\tilde{v}$ will be the lower supersolution and Theorem 2.8 will be true for every solution.

3.3. What happens if we change the kernel? A natural question is what happens if we change the kernel and if we take a kernel which decreases more quickly at infinity. For this kind of kernel K_0, the natural scaling is the following (see [13, section 4.1]):

$$K_0^{\varepsilon}(x) = \frac{1}{\varepsilon^{n+1}} K_0\left(\frac{x}{\varepsilon}\right).$$

In fact, using the same argument as the one we use for the proof of Theorem 2.2, we can prove the same kind of error estimate. For example, for $K_0(x) = \dfrac{1}{|x|^{n+p}}$ for $|x| \geq 1$ with $p \geq 3$, we get

$$|u^{\varepsilon} - u^0| \leq K\sqrt{\varepsilon} T^{1/4}.$$

This is the best estimate for ε small that we can obtain for a general kernel, as we can see in Step 3, Case 1 of the proof of Theorem 2.2.

The main difference is that the estimate of Corollary 3.1 is replaced by

$$\sup_{\mathbb{R}^N\times\{0,\dots,t_{N_T}\}} |u^{\varepsilon} - v^{\varepsilon}_{\#}| \leq K\frac{\sqrt{T}}{\varepsilon}\sqrt{\varepsilon\Delta x + \Delta t} + \sup_{\mathbb{R}^N} |u_0^{\varepsilon} - (\tilde{u}_0^{\varepsilon})_{\#}| \quad \text{for } \varepsilon\Delta x + \Delta t \leq \frac{\varepsilon^2}{K^2}.$$

Finally, we obtain for the choice $u_0^{\varepsilon} = u_0$

$$\sup_{\mathbb{R}^N\times\{0,\dots,t_{N_T}\}} |u^0 - (v^{\varepsilon})_{\#}| \leq K\sqrt{\varepsilon} T^{\frac{1}{4}} + \sup_{\mathbb{R}^N} |u_0 - \tilde{u}_0^{\varepsilon}|$$

if $\varepsilon \geq K(\sqrt{\Delta x} + (\Delta t)^{\frac{1}{3}})$. This implies that $\frac{\Delta x}{\varepsilon} \to 0$ as $\Delta x \to 0$. Thus, when $\Delta x \to 0$, we have to make the convolution on all the space, which can be very expensive in practice. This approach can also be used for the Gaussian kernel and should give the same estimates. In particular, this could give an error estimate for the equivalent version of the classical Bence–Merriman–Osher algorithm [25].

4. Proof of Theorem 2.2. Before proving Theorem 2.2, we need some notation. Let us define

$$F(M,p) = \operatorname{trace}\left(M \cdot A\left(\frac{p}{|p|}\right)\right) \quad \text{and} \quad G(M,p) = \frac{-1}{|p|} F(M,p),$$

where we recall that $A(p)\cdot p = 0$ for all $p \in \mathbf{S}^{N-1}$.

Then we have the following fundamental estimate for balls.

LEMMA 4.1 (error estimate for a ball). *Let $\varphi \in C^2(\mathbb{R}^N)$ with $D\varphi(x_0) \neq 0$ be such that the set $\{\varphi(x) \geq \varphi(x_0)\}$ is a ball of radius R. For $c_0^\varepsilon(\cdot) = \frac{1}{\varepsilon^{n+1}|\ln\varepsilon|}\ c_0\left(\frac{\cdot}{\varepsilon}\right)$, let us define*

$$c^\varepsilon = (c_0^\varepsilon \star 1_{\{\varphi(\cdot)>\varphi(x_0)\}})(x_0) - \frac{1}{2}\int_{\mathbb{R}^n} c_0^\varepsilon.$$

Then, there exists a constant $K = K(N, \sup_{\mathbb{R}^n} c_0, |Dg|_{L^\infty(\mathbf{S}^{N-1})}) > 0$ such that for $0 < \varepsilon < \delta$ with $0 < \delta \leq R/2$, we have

$$|c^\varepsilon + G(D^2\varphi(x_0), D\varphi(x_0))| \leq K \cdot e\,(\varepsilon, \delta, R)$$

with

$$e\,(\varepsilon, \delta, R) = \frac{1}{|\ln\varepsilon|}\left(\frac{1}{\delta} + \frac{1}{R}|\ln\delta|\right) + \frac{\delta}{R^2}.$$

This is a straightforward consequence of Da Lio, Forcadel, and Monneau [13, Proposition 4.1], where their Lemma 4.3 is replaced by a direct estimate for an explicit function $h(x') = R - \sqrt{R^2 - |x'|^2}$ with $x' \in \mathbb{R}^{N-1}$ and $|x'| \leq R$.

Proof of Theorem 2.2. Let u^0 be the solution of the mean curvature motion (2.5). The idea of the proof is to regularize the function u^0 by a kind of sup-convolution but using quartic penalization and then to plug the regularized function $u^{0,\alpha}$ into (2.4). This regularization allows us to control quantitatively the first and second space derivatives of $u^{0,\alpha}$. This is a quantitative version of the definition of Barles and Georgelin [4] for viscosity solutions of mean curvature motion, where they used this kind of regularization to prove that one can take test functions such that $D\varphi \neq 0$ or $D\varphi = 0$ and $D^2\varphi = 0$ in the definition of viscosity solutions. This kind of argument is also used to obtain the comparison principle for mean curvature-type equations.

The proof is decomposed into five steps:

Step 1. *Regularization of u^0.*

By classical estimates for mean curvature-type equations, we have

$$\|Du^0\|_{L^\infty(\mathbb{R}^N\times(0,T))} \leq \|Du_0\|_{L^\infty(\mathbb{R}^N)}. \tag{4.1}$$

We regularize u^0 by considering a spatial sup-convolution by quartics:

$$u^{0,\alpha}(x,t) = \sup_{y\in\mathbb{R}^N}\left\{u^0(x-y,t) - \frac{1}{4\alpha}|y|^4\right\} \tag{4.2}$$

$$= \sup_{z\in\mathbb{R}^N}\left\{u^0(z,t) - \frac{1}{4\alpha}|z-x|^4\right\}. \tag{4.3}$$

Since u^0 is Lipschitz continuous, the supremum is reached.

Step 2. *Estimate between u^0 and $u^{0,\alpha}$.*

For $x \in \mathbb{R}^N$ and $t \in (0,T)$, we denote by $\bar{x} = \bar{x}(x,t)$ a point where the maximum is reached in (4.3). Then

$$u^{0,\alpha}(x,t) = u^0(\bar{x},t) - \frac{1}{4\alpha}|\bar{x}-x|^4 \geq u^0(x,t). \tag{4.4}$$

Therefore

$$\frac{1}{4\alpha}|\bar{x}-x|^4 \le u^0(\bar{x},t)-u^0(x,t) \le K|\bar{x}-x|, \tag{4.5}$$

and so

$$|\bar{x}-x| \le K\alpha^{1/3}. \tag{4.6}$$

Moreover, by (4.4), we get

$$0 \le u^{0,\alpha}(x,t)-u^0(x,t) \le u^0(\bar{x},t)-u^0(x,t) \le K|\bar{x}-x|. \tag{4.7}$$

Using (4.6) and (4.7), we then deduce that, for $x \in \mathbb{R}^N$ and $t \in (0,T)$,

$$|u^{0,\alpha}(x,t)-u^0(x,t)| \le K\alpha^{1/3}. \tag{4.8}$$

Before continuing the proof, we need some notation. We now want to compare u^ε and $u^{0,\alpha}$. To do that, we will prove that $u^{0,\alpha} \pm (f(\varepsilon,\alpha)t + K\alpha^{\frac{1}{3}} + \sup_{\mathbb{R}^N}|u_0^\varepsilon - u_0|)$ are the super- and subsolutions, respectively, of problem (2.4) for

$$f(\varepsilon,\alpha) = K\left(\frac{1}{\sqrt{|\ln\varepsilon|}} + \frac{\ln|\ln\varepsilon|}{\alpha^{1/3}|\ln\varepsilon|} + \frac{1}{\alpha^{2/3}}\left(\frac{1}{\sqrt{|\ln\varepsilon|}}+\varepsilon\right)\right).$$

Let $\psi \in C^\infty(\mathbb{R}^N \times (0,T))$ such that $u^{0,\alpha}-\psi$ reaches a global strict maximum at $(x_0,t_0) \in \mathbb{R}^N \times (0,T)$. Since $u^{0,\alpha}$ is semiconvex in the space variable, the function $u^{0,\alpha}$ is derivable with respect to x at the point (x_0,t_0).

Let us denote

$$\varphi_z(x) = u^0(z,t_0) - \frac{1}{4\alpha}|z-x|^4.$$

Then

$$u^{0,\alpha}(x,t_0) = \sup_{z\in\mathbb{R}^N} \varphi_z(x).$$

We denote by $\bar{x}_0$ a point where the maximum is reached for $x = x_0$ and we also denote $\varphi = \varphi_{\bar{x}_0}$. We then deduce that

$$\psi(x,t_0) \ge u^{0,\alpha}(x,t_0) \ge \varphi(x) \tag{4.9}$$

with equality for $x = x_0$. This implies that

$$D\varphi(x_0) = Du^{0,\alpha}(x_0,t_0) = D\psi(x_0,t_0) \tag{4.10}$$

and

$$D^2\varphi(x_0) \le D^2\psi(x_0,t_0). \tag{4.11}$$

Moreover, with $h = |\bar{x}_0 - x_0|$, we have (by (4.6))

$$h \le K\alpha^{\frac{1}{3}}, \tag{4.12}$$

$$|D\varphi(x_0)| = \frac{h^3}{\alpha}, \tag{4.13}$$

and the set $\{\varphi(x) \geq \varphi(x_0)\}$ is a ball of radius $R = h$. Let us define the error

$$e(\varphi) = |c^\varepsilon[\varphi](x_0)|D\varphi(x_0)| - F(D^2\varphi(x_0), D\varphi(x_0))|,$$

where

$$c^\varepsilon[\varphi](x_0) = \left(c_0^\varepsilon \star 1_{\{\varphi(\cdot)\geq\varphi(x_0)\}}\right)(x_0) - \frac{1}{2}\int_{\mathbb{R}^N} c_0^\varepsilon.$$

Step 3. $e(\varphi) \leq f(\varepsilon, \alpha)$.

We distinguish three cases:

Case 1. $h \leq 2\varepsilon$. In this case, we directly have the estimate (using (4.13))

$$|c^\varepsilon[\varphi](x_0)|D\varphi(x_0)|| \leq |c_0^\varepsilon|_{L^1(\mathbb{R}^N)}\frac{h^3}{\alpha} \leq \frac{h^2|c_0|_{L^1(\mathbb{R}^N)}}{\varepsilon|\ln\varepsilon|}\frac{h}{\alpha} \leq \frac{K}{\alpha^{2/3}}\frac{\varepsilon}{|\ln\varepsilon|},$$

where we have also used (4.12). Using, moreover, the fact that

$$|F(D^2\varphi(x_0), D\varphi(x_0))| \leq \frac{K}{h}|D\varphi(x_0)| \leq K\frac{h^2}{\alpha},$$

we then deduce that

$$e(\varphi) \leq \frac{K}{\alpha^{2/3}}\varepsilon \leq f(\varepsilon, \alpha).$$

Case 2. $2\varepsilon < h \leq \frac{2}{\sqrt{|\ln\varepsilon|}}$. Using (4.13) and Lemma 4.1, with $\delta = h/2$, we then deduce that

$$\begin{aligned} e(\varphi) &\leq \frac{Kh^3}{\alpha}\left(\frac{1}{|\ln\varepsilon|}\left(\frac{1}{h} + \frac{1}{h}|\ln\delta|\right) + \frac{1}{h}\right) \\ &\leq \frac{Kh}{\alpha^{2/3}} \leq \frac{K}{\alpha^{2/3}}\frac{1}{\sqrt{|\ln\varepsilon|}} \leq f(\varepsilon, \alpha). \end{aligned}$$

Case 3. $h > \frac{2}{\sqrt{|\ln\varepsilon|}}$. We set $\delta = \frac{1}{\sqrt{|\ln\varepsilon|}}$. We then have $\varepsilon < \delta < \frac{h}{2}$. So by Lemma 4.1, we deduce that

$$\begin{aligned} e(\varphi) &\leq \frac{Kh^3}{\alpha}\left(\frac{1}{\sqrt{|\ln\varepsilon|}} + \frac{1}{h}\frac{\ln|\ln\varepsilon|}{|\ln\varepsilon|} + \frac{1}{h^2}\frac{1}{\sqrt{|\ln\varepsilon|}}\right) \\ &\leq K\left(\frac{1}{\sqrt{|\ln\varepsilon|}} + \frac{1}{\alpha^{1/3}}\frac{\ln|\ln\varepsilon|}{|\ln\varepsilon|} + \frac{1}{\alpha^{2/3}}\frac{1}{\sqrt{|\ln\varepsilon|}}\right) \\ &\leq f(\varepsilon, \alpha). \end{aligned}$$

Step 4. *Estimate between* $u^{0,\alpha}$ *and* u^ε.

By construction, we have, since $u^{0,\alpha}$ is a subsolution of (2.5) (but does not satisfy the initial condition),

$$\begin{aligned} \psi_t(x_0, t_0) &\leq F(D^2\psi(x_0, t_0), D\psi(x_0, t_0)) \\ &\leq c^\varepsilon[\varphi](x_0)|D\psi(x_0, t_0)| + F(D^2\varphi(x_0), D\varphi(x_0)) - c^\varepsilon[\varphi](x_0)|D\varphi(x_0)| \\ &\leq c^\varepsilon[u^{0,\alpha}(\cdot, t_0)](x_0)|D\psi(x_0, t_0)| + f(\varepsilon, \alpha), \end{aligned}$$

where we have used (4.9) for the last line. On the other hand, we have

$$u^{0,\alpha}(\cdot,0) - K\alpha^{1/3} - \sup_{\mathbb{R}^N} |u_0^\varepsilon - u_0| \le u_0^\varepsilon,$$

where we have used (4.8) at the limit $t = 0$. We then deduce that $\tilde{u}^{0,\alpha}(x,t) = u^{0,\alpha}(x,t) - f(\varepsilon,\alpha)t - K\alpha^{1/3} - \sup_{\mathbb{R}^N} |u_0^\varepsilon - u_0|$ is a subsolution of (2.4). This implies that $\tilde{u}^{0,\alpha} \le u^\varepsilon$ and so

$$u^{0,\alpha} - u^\varepsilon \le f(\varepsilon,\alpha)t + K\alpha^{1/3} + \sup_{\mathbb{R}^N} |u_0^\varepsilon - u_0|. \tag{4.14}$$

Step 5. *Estimate between* u^0 *and* u^ε.

Using (4.8) and (4.14), we then get

$$u^0 - u^\varepsilon = u^0 - u^{0,\alpha} + u^{0,\alpha} - u^\varepsilon \le f(\varepsilon,\alpha)t + K\alpha^{1/3} + \sup_{\mathbb{R}^N} |u_0^\varepsilon - u_0|.$$

But

$$\begin{aligned} &f(\varepsilon,\alpha)t + K\alpha^{1/3} \\ &\quad = K\left(\frac{1}{\sqrt{|\ln\varepsilon|}} + \frac{\ln|\ln\varepsilon|}{\alpha^{1/3}|\ln\varepsilon|} + \frac{1}{\alpha^{2/3}}\left(\frac{1}{\sqrt{|\ln\varepsilon|}} + \varepsilon\right)\right)t + K\alpha^{1/3} \\ &\quad \le K\left(\frac{T}{\sqrt{|\ln\varepsilon|}}\right)^{\frac{1}{3}} \end{aligned}$$

for $\alpha = \frac{Kt}{\sqrt{|\ln\varepsilon|}}$ and $t \le T \le 1$. We then get

$$u^0 - u^\varepsilon \le K\left(\frac{T}{\sqrt{|\ln\varepsilon|}}\right)^{\frac{1}{3}} + \sup_{\mathbb{R}^N} |u_0^\varepsilon - u_0|.$$

A lower bound is proved similarly (using the same result as Lemma 4.1, where the set $\{\varphi(x) \ge \varphi(x_0)\}$ is replaced by $\{\varphi(x) > \varphi(x_0)\}$). This implies (2.7) for $T \le 1$. This ends the proof of the theorem.

5. Numerical scheme for dislocation dynamics.

5.1. Definitions and preliminary results. We recall here the notation used in the scheme. The discrete solution v is computed iteratively by solving the implicit scheme

$$v_I^0 = \tilde{u}_0(x_I), \qquad \frac{v_I^{n+1} - v_I^n}{\Delta t} = c^\Delta[v]_I^{n+1} G(v^{n+1})_I, \tag{5.1}$$

where the nonlocal velocity is defined in (2.10). The approximation of the gradient of v^{n+1} at the point x_I is given by

$$G(v^{n+1})_I = \begin{cases} G^+(D^+v_I^{n+1}, D^-v_I^{n+1}) & \text{if} \quad c^\Delta[v]_I^{n+1} \ge 0, \\ G^-(D^+v_I^{n+1}, D^-v_I^{n+1}) & \text{if} \quad c^\Delta[v]_I^{n+1} < 0, \end{cases}$$

where $G^{\pm}$ is a suitable approximation of the Euclidean norm and $D^{\pm}v^n(x_I) = (D^{\pm}_{x_1}v^n(x_I), \dots, D^{\pm}_{x_N}v^n(x_I))$ are the discrete gradients. The terms $D^{\pm}_{x_i}v^n(x_I)$ are the standard forward and backward first order differences; i.e., for a general function $f(x_I)$

$$D^+_{x_i}f(x_I) = \frac{f(x_{I^{i,+}}) - f(x_I)}{\Delta x_i},$$

$$D^-_{x_i}f(x_I) = \frac{f(x_I) - f(x_{I^{i,-}})}{\Delta x_i},$$

where

$$I^{k,\pm} = (i_1, \dots, i_{k-1}, i_k \pm 1, i_{k+1}, \dots, i_N). \tag{5.2}$$

The approximations of the Euclidean norm $G^{\pm}$ are those proposed by Osher and Sethian [28] (we can also use the ones proposed by Rouy and Tourin [29]):

$$G^+(P,Q) = \left(\sum_{i=1,..,N} \max(P_i,0)^2 + \sum_{i=1,..,N} \min(Q_i,0)^2 \right)^{\frac{1}{2}}, \tag{5.3}$$

$$G^-(P,Q) = \left(\sum_{i=1,..,N} \min(P_i,0)^2 + \sum_{i=1,..,N} \max(Q_i,0)^2 \right)^{\frac{1}{2}}.$$

We recall that the functions $G^{\pm}$ are Lipschitz continuous with respect to the discrete gradients, i.e.,

$$|G^{\pm}(P_1,P_2) - G^{\pm}(P_1',P_2')| \le (|P_1 - P_1'| + |P_2 - P_2'|). \tag{5.4}$$

They are consistent with the Euclidean norm

$$G^{\pm}(P,P) = |P|, \tag{5.5}$$

and $G^{\pm} = G^{\pm}(P_1^+, \dots, P_N^+, P_1^-, \dots, P_N^-)$ enjoy suitable monotonicity with respect to each variable

$$\frac{\partial G^+}{\partial p_i^+} \ge 0, \quad \frac{\partial G^+}{\partial p_i^-} \ge 0, \quad \frac{\partial G^-}{\partial p_i^+} \ge 0, \quad \frac{\partial G^-}{\partial p_i^-} \ge 0. \tag{5.6}$$

To define discrete solution, we need the following notation:

$$\tilde{c}^{\Delta}[v]_I^{n+1} = \sum_{J \in \mathbb{Z}^N} \bar{c}^0_{I-J} 1_{\{v_J^{n+1} > v_I^{n+1}\}} \Delta x_1, \dots, \Delta x_N - \frac{1}{2} \sum_{J \in \mathbb{Z}^N} \bar{c}^0_J \Delta x_1, \dots, \Delta x_N, \tag{5.7}$$

$$\tilde{G}(v^{n+1})_I = \begin{cases} G^+(D^+v_I^{n+1}, D^-v_I^{n+1}) & \text{if} \quad \tilde{c}^{\Delta}[v]_I^{n+1} \ge 0, \\ G^-(D^+v_I^{n+1}, D^-v_I^{n+1}) & \text{if} \quad \tilde{c}^{\Delta}[v]_I^{n+1} < 0, \end{cases}$$

where $\bar{c}_0$ is defined in (2.11). Finally, for simplicity of presentation, let us denote

$$s[v]_I^{n+1} = \mathrm{Sign}(c^{\Delta}[v]_I^{n+1}), \quad \tilde{s}[v]_I^{n+1} = \mathrm{Sign}(\tilde{c}^{\Delta}[v]_I^{n+1}).$$

DEFINITION 5.1 (numerical subsolution, supersolution, and solution of the scheme). *We say that v is a discrete subsolution (resp., supersolution) of the scheme* (2.9) *if, for all $I \in \mathbb{Z}^N$, $n \in \mathbb{N}$, we have*

$$v_I^{n+1} \le v_I^n + \Delta t\, c^\Delta[v]_I^{n+1} G(v^{n+1})_I$$

$$\left(\text{resp., } v_I^{n+1} \ge v_I^n + \Delta t\, \tilde{c}^\Delta[v]_I^{n+1} \tilde{G}(v^{n+1})_I\right).$$

We say that v is a discrete solution if and only if it is a sub- and a supersolution.

LEMMA 5.2 (almost monotonicity of the scheme). *The scheme* (2.9) *is almost monotone in the following sense. Let v, w be two discrete functions and assume that there is I such that $w_I = v_I$ and $w_J \ge v_J$ for $J \ne I$. Then*

$$c^\Delta[v]_I G(v)_I \le c^\Delta[w]_I G(w)_I$$

and

$$\tilde{c}^\Delta[v]_I G(v)_I \le \tilde{c}^\Delta[w]_I G(w)_I.$$

Proof of Lemma 5.2. First, we remark that $c^\Delta[v]_I \le c^\Delta[w]_I$. Then, there are three cases:

1. $c^\Delta[v]_I \le 0 \le c^\Delta[w]_I$. In this case, the result is trivial.
2. $0 \le c^\Delta[v]_I \le c^\Delta[w]_I$. Using the monotonicity of G, we get that $G(v)_I \le G(w)_I$. This implies the result.
3. $c^\Delta[v]_I \le c^\Delta[w]_I \le 0$. Using the monotonicity of G, we get that $G(v)_I \ge G(w)_I$. This implies the result.

The proof for $\tilde{c}^\Delta$ is the same, and we skip it. This ends the proof of the lemma.

The existence of a solution for the scheme is not trivial (since it is implicit and nonlocal). This is the subject of the following proposition.

PROPOSITION 5.3 (existence of a solution for the scheme (2.9)). *There exists at least one discrete solution v of the scheme* (2.9) *in the sense of Definition* 5.1.

Proof of Proposition 5.3. We assume that there exists a solution v^n at step n, and we will construct a solution v^{n+1} at step $n+1$. First, we remark that if $(v^{n+1,i})_i$ is a family indexed by i of a discrete subsolution at step $n+1$, then $v^{n+1} = \max\, v^{n+1,i}$ is still a subsolution (it suffices to use Lemma 5.2). Moreover, $w^+ = \sup_{Q^\Delta} |v^n|$ and $w^- = -\sup_{Q^\Delta} |v^n|$ are, respectively, a discrete supersolution and a subsolution of the scheme (2.9). Then, let us define

$$v^{n+1} = \max\{w \text{ subsolution at step } n+1 \text{ such that } w \le w^+\}.$$

Then v^{n+1} is a discrete subsolution at step $n+1$. Now let us prove that v^{n+1} is a supersolution. By contradiction, assume that there is I such that

$$\frac{v_I^{n+1} - v_I^n}{\Delta t} < \tilde{c}[v]_I^{n+1} G^{\tilde{s}[v]_I^n}(D^+ v_I^{n+1}, D^- v_I^{n+1}).$$

This implies in particular that $v_I^{n+1} < w^+$. Now let us consider the solution w_I (in a sense similar to Definition 5.1) of

$$\frac{w_I - v_I^n}{\Delta t} = c^\Delta_{v^{n+1}}[w]_I \begin{cases} G^+(D^+_{v^{n+1}} w_I, D^-_{v^{n+1}} w_I) & \text{if} \quad c^\Delta_{v^{n+1}}[w]_I \ge 0, \\ G^-(D^+_{v^{n+1}} w_I, D^-_{v^{n+1}} w_I) & \text{if} \quad c^\Delta_{v^{n+1}}[w]_I < 0, \end{cases}$$

where

$$c^{\Delta}_{v^{n+1}}[w]_I = \sum_{J\in\mathbb{Z}^N} \bar{c}^0_{I-J} 1_{\{v^{n+1}_J \geq w_I\}} \Delta x_1,\dots,\Delta x_N - \frac{1}{2}\sum_{J\in\mathbb{Z}^N} \bar{c}^0_J \Delta x_1,\dots,\Delta x_N$$

and $D^{\pm}_{v^{n+1}} w_I = (D^{\pm}_{x_1,v^{n+1}} w_I,\dots,D^{\pm}_{x_N,v^{n+1}} w_I)$ with

$$D^{+}_{x_i,v^{n+1}} w_I = \frac{v^{n+1}_{I^{i,+}} - w_I}{\Delta x_i},$$

$$D^{-}_{x_i,v^{n+1}} w_I = \frac{w_I - v^{n+1}_{I^{i,-}}}{\Delta x_i},$$

where $I^{k,\pm}$ is defined in (5.2). The existence of such a solution comes from the fact that the left-hand side is nondecreasing in w_I and the right-hand side is nonincreasing. Then it is easy to prove (using Lemma 5.2) that $w_I > v^{n+1}_I$ and that w defined by

$$w_J = \begin{cases} w_I & \text{if } J = I, \\ v^{n+1}_J & \text{otherwise} \end{cases}$$

is a discrete subsolution of (2.9) at step $n+1$. This contradicts the definition of v^{n+1} and ends the proof of the proposition.

PROPOSITION 5.4 (properties of the discrete solutions). *Assume that* $\tilde{u}_0 \in W^{1,\infty}(\mathbb{R}^N)$. *Let us denote* $L = |D\tilde{u}_0|_{L^\infty(\mathbb{R}^N)}$ *and*

(5.8)
$$\overline{v} = \sup\{v \text{ subsolution of (2.9) with initial condition } v^0 = \tilde{u}_0, v \leq |\tilde{u}_0|_{L^\infty(\mathbb{R}^N)}\}.$$

For $k \in \mathbb{Z}^N$ *such that* $k\cdot\Delta X > 0$ *(with* $\Delta X = (\Delta x_1,\dots,\Delta x_N)$*), the following estimates hold:*

$$\frac{\overline{v}(x_I,t_n) - \overline{v}(x_I + k\cdot\Delta X, t_n)}{k\cdot\Delta X} \leq L, \tag{5.9}$$

$$\frac{\overline{v}(x_I,t_n) - \overline{v}(x_I + k\cdot\Delta X, t_n)}{k\cdot\Delta X} \geq -L, \tag{5.10}$$

$$\left|\frac{\overline{v}^{n+1}_I - \overline{v}^n_I}{\Delta t}\right| \leq \frac{\sqrt{N}}{2}|c_0|_{L^1(\mathbb{R}^N)}\, L. \tag{5.11}$$

The same estimates hold for

(5.12)
$$\underline{v} = \inf\{v \text{ supersolution of (2.9) with initial condition } v^0 = \tilde{u}_0, v \geq -|u_0|_{L^\infty(\mathbb{R}^N)}\}.$$

Proof of Proposition 5.4. For $k \in \mathbb{Z}^N$, let us denote by

$$\tilde{u}_{0,k}(x_I) = \tilde{u}_0(x_I + k\cdot\Delta X) + Lk\cdot\Delta X \geq \tilde{u}_0(x_I)$$

and by $\overline{v}_k$ the greater subsolution, with initial condition $\tilde{u}_{0,k}$, i.e.,

$$\overline{v}_k = \sup\{v_k \text{ subsolution of (2.9) with initial condition } v^0_k = \tilde{u}_{0,k}, v_k \leq |u_0|_{L^\infty(\mathbb{R}^N)}\}.$$

First, since the equation does not see the constants, we remark that $\overline{v} - Lk \cdot \Delta X$ is a subsolution of (2.9) with initial condition $\overline{v}^0 - Lk \cdot \Delta X \leq \tilde{u}_{0,k}$. Thus, by definition of $\overline{v}_k$, we have

$$\overline{v} - Lk \cdot \Delta X \leq \overline{v}_k.$$

Moreover, using the fact that the scheme is invariant by translation in space, we deduce that $\overline{v}(x_I + k \cdot \Delta X, t_n)$ is the greater subsolution with initial condition $\overline{v}(x_I + k \cdot \Delta X, 0) = \tilde{u}_{0,k}(x_I)$. We then deduce that

$$\overline{v}_k(x_I, t_n) = \overline{v}(x_I + k \cdot \Delta X, t_n).$$

This implies that for $k \cdot \Delta X > 0$

$$\frac{\overline{v}(x_I, t_n) - \overline{v}(x_I + k \cdot \Delta X, t_n)}{k \cdot \Delta X} \leq L.$$

The proof of (5.10) is similar, and we skip it.

To prove (5.11), we use the fact that $\overline{v}$ is a solution and the two previous estimates. This implies

$$-\frac{\sqrt{N}}{2}|c_0|_{L^1(\mathbb{R}^N)}L \leq \frac{\overline{v}_I^{n+1} - \overline{v}_I^n}{\Delta t} \leq \frac{\sqrt{N}}{2}|c_0|_{L^1(\mathbb{R}^N)}L,$$

which is the desired result.

Before proving Theorem 2.3, we need the following lemma.

LEMMA 5.5 (equivalent formulation for the discrete velocity). *The discrete velocity $c^\Delta[v]$ can be rewritten as*

$$c^\Delta[v]_I^n = \left(c_0 \star 1_{\{v_\#(.,t_n) \geq v_\#(x_I,t_n)\}}\right)(x_I) - \frac{1}{2}|c_0|_{L^1(\mathbb{R}^N)},$$

where $v_\#$ is defined in (2.13).

Proof of Lemma 5.5. The idea of the proof is borrowed from Alvarez et al. [2]. Using the definition of the discrete velocity and $\bar{c}^0$, we get

$$\begin{aligned}
c^\Delta[v]_I^n &= \sum_{J \in \mathbb{Z}^N} \bar{c}^0_{I-J} 1_{\{v_J^n \geq v_I^n\}} \Delta x_1, \ldots, \Delta x_N - \frac{1}{2} \sum_{J \in \mathbb{Z}^N} \bar{c}^0_J \Delta x_1, \ldots, \Delta x_N \\
&= \sum_{J \in \mathbb{Z}^N} \int_{Q_{J-I}} c_0(y)\, dy\, 1_{\{v_J^n \geq v_I^n\}} - \frac{1}{2} \sum_{J \in \mathbb{Z}^N} \int_{Q_J} c_0(y)\, dy \\
&= \sum_{J \in \mathbb{Z}^N} \int_{Q_J} c_0(x_I - y)\, 1_{\{v_J^n \geq v_I^n\}}\, dy\, - \frac{1}{2}|c_0|_{L^1(\mathbb{R}^N)} \\
&= \sum_{J \in \mathbb{Z}^N} \int_{Q_J} c_0(x_I - y)\, 1_{\{v_\#(y,t_n) \geq v_\#(x_I,t_n)\}}\, dy\, - \frac{1}{2}|c_0|_{L^1(\mathbb{R}^N)} \\
&= \left(c_0 \star 1_{\{v_\#(\cdot,t_n) \geq v_\#(x_I,t_n)\}}\right)(x_I) - \frac{1}{2}|c_0|_{L^1(\mathbb{R}^N)}.
\end{aligned}$$

This ends the proof of the lemma.

We now prove Theorem 2.3.

Proof of Theorem 2.3. The proof is an adaptation of the proof of Crandall and Lions [11], revisited by Alvarez et al. [1]. Nevertheless, for the reader's convenience, we give the main steps in order to show the new difficulties due to the nonlocal term. The main idea of the proof is the same as that of comparison principles, i.e., to consider the maximum of $u-v_{\#}$, to duplicate the variable, and to use the viscosity inequalities to get the result.

The proof splits into four steps.

We first assume that

$$\bar{u}_0(x_I) \geq (\tilde{u}_0(x_I))_{\#} \quad \forall I \in \mathbb{Z}^N, \tag{5.13}$$

and we set

$$\mu^0 = \sup_{\mathbb{R}^N}(\bar{u}_0 - (\tilde{u}_0)_{\#}) \geq 0. \tag{5.14}$$

We denote throughout by K various constants depending only on N, $|c_0|_{W^{1,1}(\mathbb{R}^N)}$, $|D\bar{u}_0|_{L^\infty(\mathbb{R}^N)}$, $|D\tilde{u}_0|_{L^\infty(\mathbb{R}^N)}$, and μ^0.

Step 1: *Estimate on* v.

We have the following estimate for the discrete solution:

$$-Kt_n \leq \bar{u}_0(x) - v_{\#}(x,t_n) \leq Kt_n + \mu^0. \tag{5.15}$$

To show this, it suffices to use the estimate (5.11) of Proposition 5.4. This implies that, for $x \in Q_I$ (with the notation $\overline{v}$ and $\underline{v}$ defined in Proposition 5.4),

$$\begin{aligned}(\tilde{u}_0)_{\#}(x) - Kt_n = v_I^0 - Kt_n \leq \underline{v}_I^n \leq v_I^n \\ = v_{\#}(x,t_n) \leq \overline{v}_I^n \leq Kt_n + v_I^0 = Kt_n + (\tilde{u}_0)_{\#}(x).\end{aligned}$$

This implies the desired estimate.

Before continuing the proof, we need some notation. We put

$$\mu = \sup_{\mathbb{R}^N}(u - v_{\#}).$$

We want to bound from above μ by μ^0 plus a constant. For every $0 < \alpha \leq 1$, $0 < \gamma \leq 1$, and $0 < \eta \leq 1$, we set

$$M_\eta^{\alpha,\gamma} = \sup_{\mathbb{R}^N \times \mathbb{R}^N \times (0,T) \times \{0,\dots,t_{N_T}\}} \Psi_\eta^{\alpha,\gamma}(x,y,t,t_n)$$

with

$$\Psi_\eta^{\alpha,\gamma}(x,y,t,t_n) = u(x,t) - v_{\#}(y,t_n) - \frac{|x-y|^2}{2\gamma} - \frac{|t-t_n|^2}{2\gamma} - \eta t - \alpha(|x|^2 + |y|^2).$$

We shall drop the super- and subscripts on Ψ when no ambiguity arises concerning the value of the parameter.

The main difference with the classical Crandall–Lions proof is to consider the function $v_{\#}$ in the place of v in the separation of variables. This allows us to treat the nonlocal velocity.

Since u is Lipschitz continuous and $T \leq 1$, we have

$$|u(x,t)| \leq K(1+|x|). \tag{5.16}$$

Moreover, by Step 1, we have

$$\begin{aligned}|v(x_i,t_n)| &\le |v(x_i,t_n)-\bar{u}_0(x_i)|+|\bar{u}_0(x_i)|\\ &\le Kt_n+K(1+|x_i|)\\ &\le K(1+|x_i|).\end{aligned}$$

We then deduce that Ψ achieves its maximum at some point that we denote by (x^*,y^*,t^*,t_n^*).

Step 2: *Estimates for the maximum point of* Ψ.

The maximum point of Ψ enjoys the following estimates:

$$\alpha|x^*|+\alpha|y^*|\le K \tag{5.17}$$

and

$$|x^*-y^*|\le K\gamma,\quad |t^*-t_n^*|\le (K+2\eta)\gamma. \tag{5.18}$$

Indeed, by inequality $\Psi(x^*,y^*,t^*,t_n^*)\ge\Psi(0,0,0,0)\ge 0$, we obtain

$$\begin{aligned}\alpha|x^*|^2+\alpha|y^*|^2 &\le u(x^*,t^*)-v_\#(y^*,t_n^*)\le K(1+|x^*|+|y^*|)\\ &\le K+\frac{K^2}{\alpha}+\frac{\alpha}{2}|x^*|^2+\frac{\alpha}{2}|y^*|^2.\end{aligned}$$

This implies (5.17), since $\alpha\le 1$.

The first bound of (5.18) follows from the Lipschitz regularity in space of u, from the inequality $\Psi(x^*,y^*,t^*,t_n^*)\ge\Psi(y^*,y^*,t^*,t_n^*)$, and from (5.17). The second bound of (5.18) is obtained in the same way, using the inequality $\Psi(x^*,y^*,t^*,t_n^*)\ge\Psi(x^*,y^*,t_n^*,t_n^*)$.

Step 3: *A better estimate for the maximum point of* Ψ.

Inequality (5.17) can be strengthened to

$$\alpha|x^*|^2+\alpha|y^*|^2\le K. \tag{5.19}$$

Indeed, using the Lipschitz regularity of u, the inequality $\Psi(x^*,y^*,t^*,t_n^*)\ge\Psi(0,0,0,0)$, Step 1, and (5.18), we have

$$\begin{aligned}\alpha|x^*|^2+\alpha|x_i^*|^2 &\le u(x^*,t^*)-v_\#(y^*,t_n^*)+u_0(y^*)-u_0(y^*)\\ &\le K(|x^*-y^*|+t^*)+Kt_n^*+\mu^0\le K.\end{aligned}$$

Step 4: *Upper bound of* μ.

We have the bound $\mu\le K\sqrt{T}\,(\Delta x+\Delta t)^{\frac{1}{2}}+\mu^0$ if $\Delta x+\Delta t\le\frac{1}{K^2}$.

First, we claim that for η large enough, we have either $t^*=0$ or $t_n^*=0$. We argue by contradiction. Then the function $(x,t)\mapsto\Psi(x,y^*,t,t_n^*)$ achieves its maximum at a point of $\mathbb{R}^N\times(0,T)$. Then, using the fact that u is a subsolution of (2.8), we deduce that

$$\eta+p_t^*\le c[u](x^*,t^*)|p_x^*+2\alpha x^*|, \tag{5.20}$$

where

$$c[u](x,t)=c_0\star 1_{\{u(\cdot,t)\ge u(x,t)\}}(x)-\frac{1}{2}\int_{\mathbb{R}^N}c_0$$

and

$$p_t^* = \frac{t^* - t_n^*}{\gamma}, \qquad p_x^* = \frac{x^* - y^*}{\gamma}.$$

Since $t_n^* > 0$, we also have $\Psi(x^*, y^*, t^*, t_n^*) \geq \Psi(x^*, y, t^*, t_n)$ for $t_n \geq t_n^* - \Delta t$. This implies that

$$v_\#(y^*, t_n^*) - v_\#(y, t_n) \leq \varphi(y^*, t_n^*) - \varphi(y, t_n) \quad \text{for } t_n \geq t_n^* - \Delta t \tag{5.21}$$

for $\varphi(y, t_n) = -(\frac{|x^*-y|^2}{2\gamma} + \frac{|t^*-t_n|^2}{2\gamma} + \alpha|y|^2)$. We denote by x_I^* the node such that $y^* \in Q_{I^*}$ (the unit cube centered in x_I^*). We then deduce

$$\begin{aligned}
&\frac{\varphi(y^*, t_n^*) - \varphi(y^*, t_n^* - \Delta t)}{\Delta t} \\
&\quad \geq \frac{v_\#(y^*, t_n^*) - v_\#(y^*, t_n^* - \Delta t)}{\Delta t} \\
&\quad = \frac{v(x_I^*, t_n^*) - v(x_I^*, t_n^* - \Delta t)}{\Delta t} \\
&\quad \geq \tilde{c}^\Delta[v](x_I^*, t_n^*) G^{s[v](x_I^*, t_n^*)}(D^+ v(x_I^*, t_n^*), D^- v(x_I^*, t_n^*)) \\
&\quad \geq \tilde{c}^\Delta[v](x_I^*, t_n^*) G^{s[v](x_I^*, t_n^*)}(D^+ \varphi(y^*, t_n^*), D^- \varphi(y^*, t_n^*)),
\end{aligned}$$

where we have used the monotonicity of $G^\pm$ and (5.21) with $t_n = t_n^*$ for the last line. We denote $s^* = s[v](x_I^*, t_n^*)$. Straightforward computations of the discrete derivative of φ yield

$$\begin{aligned}
&p_t^* + \frac{\Delta t}{2\gamma} \\
&\quad \geq \tilde{c}^\Delta[v](x_I^*, t_n^*) G^{s*}\left(p_x^* - \frac{\Delta x}{2\gamma} - \alpha(2y^* + \Delta x),\ p_x^* + \frac{\Delta x}{2\gamma} - \alpha(2y^* - \Delta x)\right).
\end{aligned} \tag{5.22}$$

Subtracting the above inequality from (5.20) yields

$$\begin{aligned}
\eta &\leq \frac{\Delta t}{2\gamma} + c[u](x^*, t^*)|p_x^* + 2\alpha x^*| \\
&\quad - \tilde{c}^\Delta[v](x_I^*, t_n^*) G^{s^*}\left(p_x^* - \frac{\Delta x}{2\gamma} - \alpha(2y^* + \Delta x),\ p_x^* + \frac{\Delta x}{2\gamma} - \alpha(2y^* - \Delta x)\right) \\
&\leq \frac{\Delta t}{2\gamma} + (c[u](x^*, t^*) - \tilde{c}^\Delta[v](x_I^*, t_n^*))|p_x^*| + \alpha K|x^*| + \tilde{c}^\Delta[v](x_I^*, t_n^*) \\
&\quad \times \left| G^{s^*}\left(p_x^* - \frac{\Delta x}{2\gamma} - \alpha(2y^* + \Delta x),\ p_x^* + \frac{\Delta x}{2\gamma} - \alpha(2y^* - \Delta x)\right) - G^{s^*}(p_x^*, p_x^*) \right| \\
&\leq \frac{\Delta t}{2\gamma} + (c[u](x^*, t^*) - \tilde{c}^\Delta[v](x_I^*, t_n^*))|p_x^*| + \alpha K|x^*| \\
&\quad + K\left(\frac{\Delta x}{\gamma} + 2\alpha|y^*| + 2\alpha\Delta x\right).
\end{aligned}$$

We now have to bound the term

$$\begin{aligned}\mathcal{I} &= (c[u](x^*,t^*) - \tilde{c}^{\Delta}[v](x_I^*,t_n^*))|p_x^*| \\ &= (c_0 \star 1_{\{u(\cdot,t^*)\geq u(x^*,t^*)\}}(x^*) - c_0 \star 1_{\{v_\#(\cdot,t_n^*)>v_\#(y^*,t_n^*)\}}(x_I^*))|p_x^*|.\end{aligned}$$

We have to distinguish two cases:

Case 1. Assume that for $\gamma > 0$ fixed $\frac{|x^*-y^*|^2}{\gamma} \geq C_\gamma^2$ for some constant $C_\gamma > 0$ and for all α small.

By inequality $\Psi(x^*,y^*,t^*,t_n^*) \geq \Psi(x,x,t^*,t_n^*)$, we deduce that

$$\begin{aligned}u(x,t^*) - u(x^*,t^*) \leq\ & v_\#(x,t_n^*) - v_\#(y^*,t_n^*) \\ & - \frac{|x^*-y^*|^2}{2\gamma} - \alpha(|x^*|^2+|y^*|^2) + 2\alpha|x|^2.\end{aligned}$$

We then get the inclusion

$$\{u(\cdot,t^*) \geq u(x^*,t^*)\} \cap \{v_\#(\cdot,t_n^*) \leq v_\#(y^*,t_n^*)\} \subset \{|x|^2 \geq R_{\alpha,\gamma}^2\},$$

where $R_{\alpha,\gamma}^2 = \frac{1}{2\alpha}(\frac{|x^*-y^*|^2}{2\gamma} + \alpha(|x^*|^2+|y^*|^2))$. We also have

$$\{|x-x^*| \geq R_{\alpha,\gamma}\} \subset \{|x| \geq \tilde{R}_{\alpha,\gamma}\},$$

where $\tilde{R}_{\alpha,\gamma} = R_{\alpha,\gamma} - |x^*| \to \infty$ as $\alpha \to 0$ (see Da Lio, Forcadel, and Monneau [13, Lemma 2.5]). We then obtain

$$\begin{aligned}\mathcal{I} &\leq (c_0 \star 1_{\{v_\#(\cdot,t_n^*)>v_\#(y^*,t_n^*)\}}(x^*) - c_0 \star 1_{\{v_\#(\cdot,t_n^*)>v_\#(y^*,t_n^*)\}}(x_I^*) \\ &\qquad + c_0 \star 1_{B^c(0,R_{\alpha,\gamma})}(x^*))|p_x^*| \\ &\leq K\left(|Dc_0|_{L^1(\mathbb{R}^N)}|x^*-x_I^*| + \int_{B^c(0,\tilde{R}_{\alpha,\gamma})} c_0(x)dx\right) \\ &< K(\gamma+\Delta x) + o_\alpha(1),\end{aligned}$$

where we have used (5.18) to bound $|p_x^*|$ in the second line.

Case 2. Assume that there exists a subsequence $\alpha_n > 0$ which we still denote by α such that

$$\frac{|x^*-y^*|^2}{\gamma} \to 0 \quad \text{as } \alpha \to 0.$$

We then get that $p_x^* \to 0$ as $\alpha \to 0$. Since the velocities are bounded, we get that $\mathcal{I} = o_\alpha(1)$.

To sum up, we have shown that

$$\mathcal{I} < K(\gamma+\Delta x) + o_\alpha(1).$$

We then get that

$$\eta < K\left(\frac{\Delta x + \Delta t}{\gamma} + \gamma\right) + o_\alpha(1).$$

Putting $\eta^* = K(\frac{\Delta x+\Delta t}{\gamma} + \gamma) + o_\alpha(1)$, we then conclude that either $t^* = 0$ or $t_n^* = 0$ for $\eta \geq \eta^*$.
Whenever $t^* = 0$, we get for $\eta^* \leq \eta \leq 1$

$$M = \Psi(x^*, y^*, 0, t_n^*) \leq u_0(x^*) - v_\#(y^*, t_n^*)$$

$$\leq K|x^* - y^*| + Kt_n^* + \mu^0$$

$$\leq K\gamma + \mu^0.$$

The same result holds whenever $t_n^* = 0$. We then deduce that for $\eta = \eta^*$ and $(x, t_n) \in \mathbb{R}^N \times \{0, \dots, t_{N_T}\}$

$$u(x, t_n) - v_\#(x, t_n) - \eta^* T - 2\alpha|x_I|^2 \leq M \leq K\gamma + \mu^0.$$

Sending $\alpha \to 0$, taking the supremum over (x, t_n), and choosing $\gamma = T^{1/2}(\Delta x + \Delta t)^{1/2}$, we get

$$\mu \leq K(\Delta x + \Delta t)^{1/2}\sqrt{T} + \mu^0,$$

provided that $\Delta x + \Delta t \leq \frac{1}{K^2}$. Using the same arguments of Alvarez et al. [1, Theorem 2], we easily deduce the result in the general case. This ends the proof of Theorem 2.3.

We now point out that there is no consistency error for the scheme (2.15). If there is a consistency error, then there will be a consistency error for the velocities. This means that, for all $\varphi \in C^1(\mathbb{R}^N \times (0, T))$ such that $\|\varphi\|_{C^1(\mathbb{R}^N \times (0,T))} \leq C$, we have

$$c^\Delta[\varphi_{|Q_T^\Delta}]_I^n - c[\varphi](x_I, t_n) \leq f(\Delta x, \Delta t) \tag{5.23}$$

with $f(\Delta x, \Delta t) \to 0$ as $(\Delta x, \Delta t) \to 0$.

PROPOSITION 5.6 (no consistency error). *There is no consistency error for the scheme* (2.15)*; i.e.,* (5.23) *does not hold.*

Proof of Proposition 5.6. We have to prove that there exists a constant $C_0 > 0$ such that for all $\Delta x, \Delta t > 0$ there exists a function φ such that

$$c^\Delta[\varphi_{|Q_T^\Delta}]_I^n - c[\varphi](x_I, t_n) \geq C_0 \quad \text{and} \quad |\varphi|_{C^1(\mathbb{R}^N \times (0,T))} \leq C.$$

To see this, it suffices to take a function φ which oscillates as shown in Figure 1.

Remark 5.7. There is no consistency error, but the scheme is consistent. Indeed, if we fix the function φ in the previous proof, then using the Slepčev lemma [31, eq. (5)], we get

$$\lim_{\Delta x \to 0} c^\Delta[\varphi_{|Q_T^\Delta}]_I^n - c[\varphi](x_I, t_n) = 0.$$

We now point out that the two schemes (2.9) and (2.15) are not monotone.

PROPOSITION 5.8 (nonmonotonicity of the scheme). *The two schemes* (2.9) *and* (2.15) *are not monotone for general* $\bar{c}^0$.

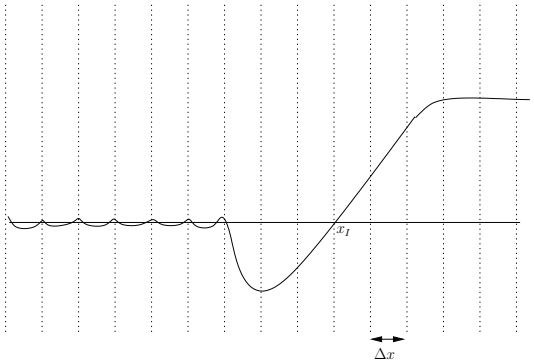


FIG. 1. *Graph of the function φ defined in $\mathbb{R}$.*

Proof of Proposition 5.8.

1. *The implicit scheme* (2.9).
   We give a simple counterexample in the one-dimensional case. We take $\tilde{u}_i^n = v_i^n = i\Delta x$. The goal is to construct a subsolution u^{n+1} and a supersolution v^{n+1} such that

$$u_i^{n+1} = i\Delta x + c_1,$$

$$v_i^{n+1} = i\Delta x + c_2.$$

   A simple computation gives

$$c_1 = \Delta t \sum_{j\geq 0} \bar{c}_j^0 \quad \text{and} \quad c_2 = \Delta t \sum_{j>0} \bar{c}_j^0.$$

   We then deduce that $c_1 - c_2 = \Delta t \bar{c}_0^0 > 0$ (if $\bar{c}_0^0 > 0$). This implies that we can construct two different solutions (the supremum of subsolutions and the infimum of supersolutions). This implies that the scheme (2.9) is not monotone.

2. *The explicit scheme* (2.15).
   We also give a counterexample in the one-dimensional case. We take

$$u_i^n = \begin{cases} -\delta & \text{if } i = 0, \\ 1-\delta & \text{if } i = -1, \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad v_i^n = \begin{cases} \delta & \text{if } i = 0, \\ 1-\delta & \text{if } i = -1, \\ 0 & \text{otherwise.} \end{cases}$$

   We then have $u_i^n \leq v_i^n$ for all $i \in \mathbb{Z}$. We then prove that $u_0^{n+1} > v_0^{n+1}$ for δ small enough. First we remark that $c^{\Delta}[v^n]_0 \leq 0$ (if $\sum_{i\neq 0,-1} \bar{c}_j^0 \geq \bar{c}_0^0 + \bar{c}_{-1}^0$), so $v_0^{n+1} \leq \delta$. Moreover,

$$\begin{aligned} u_0^{n+1} &= -\delta + \frac{\Delta t}{\Delta x}\left(\sum_j \bar{c}_j^0 - \frac{1}{2}\sum_j \bar{c}_j^0\right)\sqrt{1+\delta^2} \\ &\geq -\delta + \frac{\Delta t}{\Delta x}\left(\frac{1}{2}|c_0|_{L^1(\mathbb{R})}\right) \\ &> \delta \end{aligned}$$

if

$$\delta < \frac{\Delta t}{2\Delta x}\left(\frac{1}{2}|c_0|_{L^1(\mathbb{R})}\right).$$

This contradicts the monotonicity of the scheme.
This ends the proof of the proposition.

6. Numerical simulations. In this section, we provide some numerical simulations. In subsection 6.1, we explain a method to solve numerically the implicit scheme (2.9)–(2.11). Then we provide a simple simulation concerning a collapsing circle, with the implicit scheme, and another simulation, to highlight the fattening phenomena, with the explicit scheme.

6.1. How do we solve the implicit scheme? To solve (2.9)–(2.11), we will use an iterative process. Assume that we have a solution v^n at step n. We now want to compute a solution v^{n+1} at step $n+1$. To do this, for $\tilde{w} \in \mathbb{R}^{Q^\Delta}$, we denote by $w = \Phi(\tilde{w})$ the solution of the following auxiliary scheme:

$$\frac{w_I - v_I^n}{\Delta t} = c_{\tilde{w}}^{\Delta}[w]_I \begin{cases} G^+(D_{\tilde{w}}^+ w_I, D_{\tilde{w}}^- w_I) & \text{if} \quad c_{\tilde{w}}^{\Delta}[w]_I \geq 0, \\ G^-(D_{\tilde{w}}^+ w_I, D_{\tilde{w}}^- w_I) & \text{if} \quad c_{\tilde{w}}^{\Delta}[w]_I < 0, \end{cases} \tag{6.1}$$

where

$$c_{\tilde{w}}^{\Delta}[w]_I = \sum_{J\in\mathbb{Z}^N} \bar{c}_{I-J}^0 1_{\{\tilde{w}_J \geq w_I\}} \Delta x_1, \ldots, \Delta x_N - \frac{1}{2} \sum_{J\in\mathbb{Z}^N} \bar{c}_J^0 \Delta x_1, \ldots, \Delta x_N$$

and $D_{\tilde{w}}^{\pm} w_I = (D_{x_1,\tilde{w}}^{\pm} w_I, \ldots, D_{x_N,\tilde{w}}^{\pm} w_I)$ with

$$D_{x_i,\tilde{w}}^+ w_I = \frac{\tilde{w}_{I^{i,+}} - w_I}{\Delta x_i},$$

$$D_{x_i,\tilde{w}}^- w_I = \frac{w_I - \tilde{w}_{I^{i,-}}}{\Delta x_i},$$

where $I^{k,\pm}$ is defined in (5.2). Since the left-hand side of (6.1) is nondecreasing in w_I and the right-hand side is nonincreasing, there exists a unique solution (in a sense similar to Definition 5.1) which can be computed using a dichotomy process.

The important point is that if $\tilde{w}$ is a subsolution (resp., supersolution), then $w = (w_I)_I$ is still a subsolution (resp., supersolution) and satisfies $w_I \geq \tilde{w}_I$ (resp., $w_I \leq \tilde{w}_I$) (see the proof of Proposition 5.3).

The idea is then to define $w^0 = v^n - C$, where the constant C is such that w^0 is a subsolution of (2.9), and then to construct iteratively $w^{k+1} = \Phi(w^k)$. Setting $v^{n+1} = \lim_{k\to\infty} w^k$, we have that v^{n+1} is a solution of (2.9).

Remark 6.1. In practice, the kernel $\bar{c}^0$ will be truncated at infinity, and the convolution term will be approximated by a finite sum (see also Remark 2.7).

6.2. A collapsing circle. In this subsection, we provide a simple test with the implicit scheme concerning the evolution of a circle. The goal of this simple simulation is just to check that the circle will disappear at the expected time. We take a circle of radius 1. The parameters are $\Delta x = 0,05$, $\Delta t = 0,01$, and $\varepsilon = 0,3$. Moreover, we take the kernel

$$c_0^\varepsilon = \begin{cases} 1 & \text{if} \quad |x| \leq 0,05, \\ \frac{1}{|\ln \varepsilon||x|^3} & \text{if} \quad 0,05 \leq |x| \leq 2, \\ 0 & \text{if} \quad |x| \geq 2. \end{cases}$$

The initial condition is the distance to the circle. The result is shown in Figure 2.

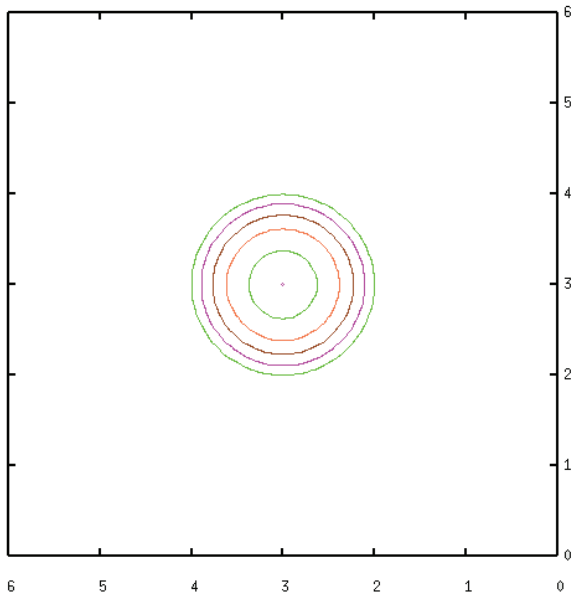


FIG. 2. *Evolution of a circle of radius* 1 *at time* 0, 0.1, 0.2, 0.3, 0.4, *and* 0.49.

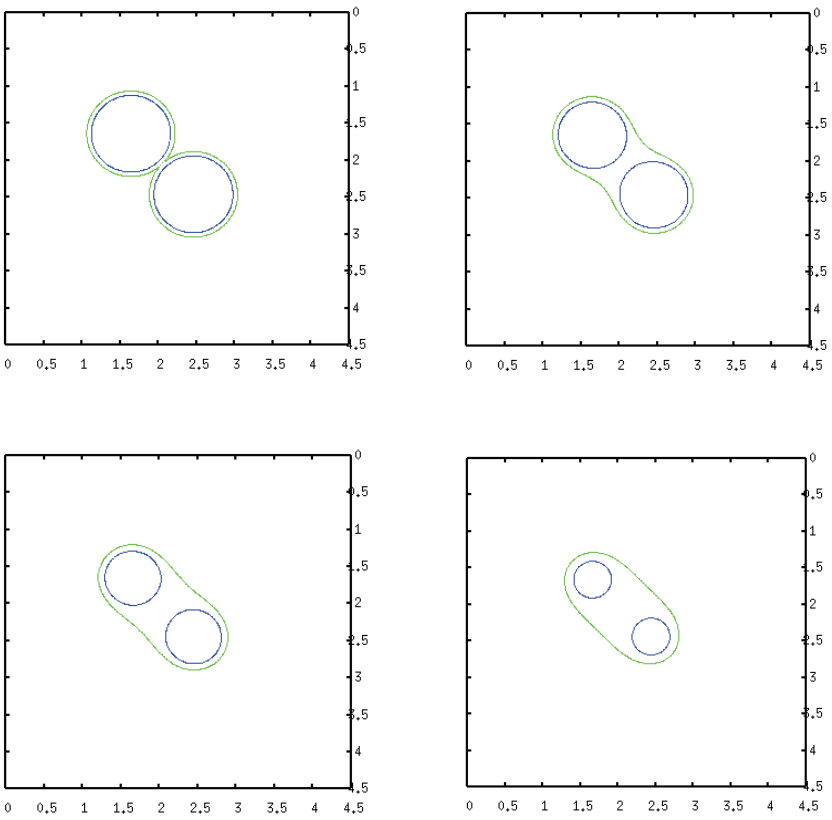


FIG. 3. *Evolution of the figure* 8 *shape at time* 0, 0.05, 0.1, *and* 0.15.

Numerically, the disappearing time is contained between 0.49 and 0.50, which corresponds to what we expect theoretically (the real time is 0.50).

6.3. Fattening phenomena. The second test is concerned with the evolution of the figure 8 shape to point out the fattening phenomena. This test has been made with the explicit scheme. We take two circles of radius $0,58$ such that they are tangent in one point and we look at the evolution of the level set 0 and $-0,06$. The parameters are $\Delta x_1 = \Delta x_2 = 0,01$, $\Delta t = 0,0001$, and $\varepsilon = 0,1$. Moreover, we take the kernel

$$c_0^\varepsilon(x) = \begin{cases} 0 & \text{if} \quad |x| \leq 0,3, \\ \frac{1}{|\ln \varepsilon||x|^3} & \text{if} \quad 0,05 \leq |x| \leq \frac{10}{3}, \\ 0 & \text{if} \quad |x| \geq \frac{10}{3}. \end{cases}$$

The results are provided in Figure 3.

Acknowledgments. The author would like to thank R. Monneau for proposing the problem and for fruitful discussions in the preparation of this paper, and E. Carlini for enlightening discussions about numerical simulations.

REFERENCES

[1] O. Alvarez, E. Carlini, R. Monneau, and E. Rouy, *Convergence of a first order scheme for a non-local eikonal equation*, Appl. Numer. Math., 56 (2006), pp. 1136–1146.

[2] O. Alvarez, E. Carlini, R. Monneau, and E. Rouy, *A convergent scheme for a nonlocal Hamilton Jacobi equation modelling dislocation dynamics*, Numer. Math., 104 (2006), pp. 413–444.

[3] O. Alvarez, P. Hoch, Y. Le Bouar, and R. Monneau, *Dislocation dynamics: Short time existence and uniqueness of the solution*, Arch. Ration. Mech. and Anal., 85 (2006), pp. 371–414.

[4] G. Barles and C. Georgelin, *A simple proof of convergence for an approximation scheme for computing motions by mean curvature*, SIAM J. Numer. Anal., 32 (1995), pp. 484–500.

[5] G. Barles and E. R. Jakobsen, *On the convergence rate of approximation schemes for Hamilton-Jacobi-Bellman equations*, M2AN Math. Model. Numer. Anal., 36 (2002), pp. 33–54.

[6] G. Barles and E. R. Jakobsen, *Error bounds for monotone approximation schemes for Hamilton–Jacobi–Bellman equations*, SIAM J. Numer. Anal., 43 (2005), pp. 540–558.

[7] G. Barles and P. E. Souganidis, *Convergence of approximation schemes for fully nonlinear second order equations*, Asymptot. Anal., 4 (1991), pp. 271–283.

[8] L. Caffarelli and P. E. Souganidis, *A rate of convergence for monotone finite difference approximations to fully nonlinear, uniformly elliptic PDEs*, Comm. Pure Appl. Math., 61 (2008), pp. 1–17.

[9] A. Chambolle and M. Novaga, *Approximation of the anisotropic mean curvature flow*, Math. Models Methods Appl. Sci., 17 (2007), pp. 833–844.

[10] Y. G. Chen, Y. Giga, and S. Goto, *Uniqueness and existence of viscosity solutions of generalized mean curvature flow equations*, J. Differential Geom., 33 (1991), pp. 749–786.

[11] M. G. Crandall and P.-L. Lions, *Two approximations of solutions of Hamilton-Jacobi equations*, Math. Comp., 43 (1984), pp. 1–19.

[12] M. G. Crandall and P.-L. Lions, *Convergent difference schemes for nonlinear parabolic equations and mean curvature motion*, Numer. Math., 75 (1996), pp. 17–41.

[13] F. Da Lio, N. Forcadel, and R. Monneau, *Convergence of a non-local eikonal equation to anisotropic mean curvature motion. Application to dislocations dynamics*, J. Eur. Math. Soc., to appear.

[14] K. Deckelnick, *Error bounds for a difference scheme approximating viscosity solutions of mean curvature flow*, Interfaces Free Bound., 2 (2000), pp. 117–142.

[15] K. Deckelnick and G. Dziuk, *Discrete anisotropic curvature flow of graphs*, M2AN Math. Model. Numer. Anal., 33 (1999), pp. 1203–1222.

[16] K. Deckelnick and G. Dziuk, *Error estimates for a semi-implicit fully discrete finite element scheme for the mean curvature flow of graphs*, Interfaces Free Bound., 2 (2000), pp. 341–359.

[17] K. Deckelnick, G. Dziuk, and C. M. Elliott, *Computation of geometric partial differential equations and mean curvature flow*, Acta Numer., 14 (2005), pp. 139–232.
[18] C. M. Elliott, *Approximation of curvature dependent interface motion*, in The State of the Art in Numerical Analysis (York, 1996), Inst. Math. Appl. Conf. Ser. New Ser. 63, Oxford University Press, New York, 1997, pp. 407–440.
[19] L. C. Evans, *Convergence of an algorithm for mean curvature motion*, Indiana Univ. Math. J., 42 (1993), pp. 533–557.
[20] L. C. Evans and J. Spruck, *Motion of level sets by mean curvature*. I, J. Differential Geom., 33 (1991), pp. 635–681.
[21] H. Ishii, *A generalization of the Bence, Merriman and Osher algorithm for motion by mean curvature*, in Curvature Flows and Related Topics (Levico, 1994), GAKUTO Internat. Ser. Math. Sci. Appl. 5, Gakkōtosho, Tokyo, 1995, pp. 111–127.
[22] H. Ishii, G. E. Pires, and P. E. Souganidis, *Threshold dynamics type approximation schemes for propagating fronts*, J. Math. Soc. Japan, 51 (1999), pp. 267–308.
[23] N. V. Krylov, *On the rate of convergence of finite-difference approximations for Bellman's equations*, Algebra i Analiz, 9 (1997), pp. 245–256.
[24] N. V. Krylov, *On the rate of convergence of finite-difference approximations for Bellman's equations with variable coefficients*, Probab. Theory Related Fields, 117 (2000), pp. 1–16.
[25] B. Merriman, J. Bence, and S. Osher, *Diffusion generated motion by mean curvature*, in Proceedings of the Computational Crystal Growers Workshop, J. Taylor, ed., AMS, Providence, RI, 1992, pp. 73–83.
[26] A. M. Oberman, *A convergent monotone difference scheme for motion of level sets by mean curvature*, Numer. Math., 99 (2004), pp. 365–379.
[27] S. Osher and N. Paragios, eds., *Geometric Level Set Methods in Imaging, Vision, and Graphics*, Springer-Verlag, New York, 2003.
[28] S. Osher and J. A. Sethian, *Fronts propagating with curvature-dependent speed: Algorithms based on Hamilton-Jacobi formulations*, J. Comput. Phys., 79 (1988), pp. 12–49.
[29] E. Rouy and A. Tourin, *A viscosity solutions approach to shape-from-shading*, SIAM J. Numer. Anal., 29 (1992), pp. 867–884.
[30] J. A. Sethian, *Level Set Methods and Fast Marching Methods. Evolving Interfaces in Computational Geometry, Fluid Mechanics, Computer Vision, and Materials Science*, Cambridge Monogr. Appl. Comput. Math. 3, 2nd ed., Cambridge University Press, Cambridge, UK, 1999.
[31] D. Slepčev, *Approximation schemes for propagation of fronts with nonlocal velocities and Neumann boundary conditions*, Nonlinear Anal., 52 (2003), pp. 79–115.
[32] P. E. Souganidis, *Approximation schemes for viscosity solutions of Hamilton-Jacobi equations*, J. Differential Equations, 59 (1985), pp. 1–43.
[33] N. J. Walkington, *Algorithms for computing motion by mean curvature*, SIAM J. Numer. Anal., 33 (1996), pp. 2215–2238.

SIAM J. NUMER. ANAL.
Vol. 46, No. 5, pp. 2742–2768
© 2008 Society for Industrial and Applied Mathematics

A MULTILEVEL PRECONDITIONER FOR THE INTERIOR PENALTY DISCONTINUOUS GALERKIN METHOD*

KOLJA BRIX†, MARTIN CAMPOS PINTO‡, AND WOLFGANG DAHMEN†

Abstract. In this article we present a multilevel preconditioner for interior penalty discontinuous Galerkin discretizations of second order elliptic boundary value problems that gives rise to uniformly bounded condition numbers without any additional regularity assumptions on the solution. The underlying triangulations are assumed only to be shape regular but may have hanging nodes subject to certain mild grading conditions. A key role is played by certain decompositions of the discontinuous trial space into a conforming subspace and a nonconforming subspace that is controlled by the jumps across the edges.

Key words. interior penalty method, energy-stable splittings, admissible averaging operators, frames, multilevel Schwarz preconditioners, discontinuous Galerkin methods

AMS subject classifications. 65F10, 65N55, 65N30

DOI. 10.1137/07069691X

1. Introduction.

An attractive feature of discontinuous Galerkin (DG) discretizations is their suitability for a variety of different problem types as well as the flexibility regarding local mesh refinement and even locally varying the order of the discretization. While initially the main focus has been on transport problems such as hyperbolic conservation laws, interest has shifted recently towards diffusion problems. One simple reason is that such problems naturally enter the picture in more complex applications such as the compressible or incompressible Navier–Stokes equations. By now numerous papers have been devoted to well-posedness and stability issues as well as to the derivation of (a priori and a posteriori) error estimation [2, 1, 13, 18, 19]. In comparison much less rigorous foundation of efficient solution methods for the resulting systems of equations seems to be currently available that covers the above-mentioned full scope of flexibility. An important step in this direction has been [11], which proposes a multigrid scheme (covering actually convection-diffusion problems as well) that does give rise to uniformly bounded condition numbers for diffusive problems provided that (i) the underlying hierarchy of meshes is *quasi-uniform* and (ii) the solution exhibits a certain (weak) extra regularity. Another successful approach, formulated as a two-level scheme in the sense of the auxiliary space method (see, e.g., [4, 17, 21]), was proposed in [10] and exhibits also mesh-independent convergence. However, its theoretical underpinning is also restricted only to quasi-uniform meshes with no hanging nodes.

On the other hand, the local nature of the method suggests a particular suitability for problems with singular behavior calling for inhomogeneous meshes and possibly

*Received by the editors July 11, 2007; accepted for publication (in revised form) February 8, 2008; published electronically July 3, 2008. This work was supported in part by the French-German PROCOPE contract 11418YB, by the European Commission Human Potential Programme under contract HPRN-CT-2002-00286, "Breaking Complexity," by the SFB 401, and by the Leibniz Programme funded by DFG.

http://www.siam.org/journals/sinum/46-5/69691.html

†Institut für Geometrie und Praktische Mathematik, RWTH Aachen, 52056 Aachen, Germany (brix@igpm.rwth-aachen.de, http://www.igpm.rwth-aachen.de/brix/; dahmen@igpm.rwth-aachen.de, http://www.igpm.rwth-aachen.de/dahmen/).

‡Institut de Recherche Mathématique Avancée, CNRS–Université Louis Pasteur, 67084 Strasbourg, France (campos@math.u-strasbg.fr, http://www.ann.jussieu.fr/~campos/).

weak regularity requirements. Therefore the main objective of this paper is to present a preconditioner that is optimal in the sense of providing uniformly bounded condition numbers even for meshes that are shape regular but only locally quasi-uniform and may have hanging nodes, arising, e.g., from local mesh refinements. Moreover, no extra regularity assumptions are needed. Our approach is quite different from [11] and uses stable splittings and concepts from the theory of multilevel Schwarz schemes. It can also be understood as a realization of the abstract setting in [17, 21].

In fact, one important by-product of our analysis, that could be reinterpreted in terms of the auxiliary space method, is that a simple splitting of any discontinuous, high order piecewise polynomial space $V_h := \mathbb{P}_k(\mathcal{T}_h)$ based on its first degree conforming subspace (the minimal conforming subspace) is stable with respect to the standard mesh-dependent norm $\|\!|\cdot|\!\|_h$ associated with the DG discretization of an H^1-elliptic boundary value problem (see (2.7)).

A special and somewhat simplified version of these findings can be formulated as follows.

THEOREM 1.1. *Let $\Phi_h := \{\phi_i : i \in \mathcal{I}_h\}$ be a standard local basis of V_h (i.e., each basis function is a piecewise polynomial supported in a single triangle), and let $V_{h,1}^{\mathrm{c}} := H_0^1(\Omega) \cap \mathbb{P}_1(\mathcal{T}_h) \subset V_h$ denote the space of continuous functions that are piecewise affine on $\mathcal{T}_h$. Then*

$$c\,|\!|\!| v |\!|\!|_h^2 \le \inf_{\substack{v^{\mathrm{c}} \in V_{h,1}^{\mathrm{c}},\ d_i \in \mathbb{R} \\ v = v^{\mathrm{c}} + \sum_{i \in \mathcal{I}_h} d_i \phi_i}} \|v^{\mathrm{c}}\|_1^2 + \sum_{i \in \mathcal{I}_h} |\!|\!| d_i \phi_i |\!|\!|_h^2 \le C\,|\!|\!| v |\!|\!|_h^2, \quad v \in V_h,$$

holds for constants c and C independent of $\mathcal{T}_h$, provided that the above-mentioned grading property holds.

However, when permitting hanging nodes due to local refinements, an essential (and necessary) requirement for the validity of such a result is that the triangulation $\mathcal{T}_h$ underlying V_h satisfies a certain mild *grading* property. Given such a grading property, one can prove a certain *local* Jackson-type estimate which is a crucial ingredient of the proof. In fact, a global version of such an estimate also appears in [10]. Thus, in summary, the primary goal of this paper is the rigorous foundation for mesh-independent iterative multilevel schemes for DG discretizations with local mesh refinements and hanging nodes. We are content here with additive schemes, although the quantitative performance of multiplicative versions is expected to be better; see [10, 11].

The layout of the paper is as follows. In section 2 we set notation, formulate the model problem, and recall some relevant facts about the *symmetric interior penalty discontinuous Galerkin method* (SIPG). In section 3 we briefly review the relevant aspects of additive Schwarz schemes that provide the framework for our preconditioners. Specifically, in section 3.1 we point out first that a certain naive attempt motivated by wavelet-type preconditioners does not work. This is unfortunate, since so-called multiwavelets would offer a simple tool in the present context. The explanation for why this fails is instructive in that it highlights the necessity of working with suitable splits of the trial spaces into a conforming part and a part that could be controlled solely by the jumps across element edges. Section 4 is devoted to the identification of suitable stable splittings that give rise to uniformly bounded condition numbers. There are essentially two crucial issues that are perhaps worth mentioning. First, the required grouping of stable splittings into conforming and nonconforming parts requires nontrivial conforming subspaces. When working with local mesh refinements with hanging vertices, this will be seen to boil down to certain mild *grading con-*

ditions on the meshes that are discussed in section 4.1. Second, given such graded meshes, suitable decompositions into conforming and nonconforming subspaces can be obtained through certain averaging projectors. Their essential qualifying property, namely, the *Jackson-type estimate* mentioned above, is developed in section 4.2. In fact, the main result of section 4 is that any split induced by such an admissible projector into a conforming subspace, which can be viewed as an auxiliary space in the sense of [21], gives rise to optimal Schwarz preconditioners by simply taking the union of two stable splittings for the respective subspaces; see Theorem 1.1. Section 5 is then devoted to the detailed concrete realization of a specific admissible averaging projector whose image is the space of continuous piecewise linear finite elements. In this case stable splittings for the conforming part can be invoked from the literature, while the discontinuous piecewise polynomial basis of the whole trial space turns out to be a stable splitting for the corresponding nonconforming subspace. In this case the conforming subspace is minimal. We emphasize though that this is only one possible realization among many other ones. In section 6 we present some first numerical experiments. We conclude with indicating alternatives, based on projectors onto a maximal conforming subspace, whose detailed analysis supported by more extensive numerical quantifications are deferred to a forthcoming paper; see [5].

We shall sometimes write $a \lesssim b$ to mean that a can be bounded by some constant multiple of b where the constant does not depend on any parameters on which a and b may depend. Likewise $a \sim b$ means that both $a \lesssim b$ and $b \lesssim a$ hold.

2. The symmetric interior penalty method—a model problem. Although the subsequent developments carry over to higher spatial dimensions, we shall confine the discussion for simplicity to the two-dimensional case. Thus, let Ω be a bounded polygonal domain in $\mathbb{R}^2$, and let

$$a(u,v) := \int_\Omega (A\nabla u)\cdot\nabla v + \int_\Omega buv, \tag{2.1}$$

where A is a uniformly positive definite 2×2 matrix and b is a nonnegative bounded function on Ω. To avoid quadrature issues, we shall always assume that both A and b are piecewise constant with respect to a fixed coarse *shape regular* conforming triangulation $\mathcal{T}^0 = \mathcal{T}^0(\Omega)$ of Ω. As usual this means that the ratio of the radii of the circumscribed and inscribed circles of all triangles remains uniformly bounded. It will be convenient to work with *closed* triangles; i.e., the triangles K will always be assumed to be closed.

We consider the following model problem: Given $f \in L_2(\Omega)$, find $u \in H_0^1(\Omega)$ (the usual Sobolev space of functions in $L_2(\Omega)$ whose first order weak derivatives belong also to $L_2(\Omega)$ and whose trace vanishes on $\partial\Omega$) such that

$$a(u,v) = \langle f, v\rangle, \quad v \in H_0^1(\Omega). \tag{2.2}$$

The Lax–Milgram theorem ensures the unique solvability of (2.2). In particular, we know that there exist positive constants c_a and C_a depending only on Ω, A, and b such that for $\|v\|_1^2 := \|v\|_{L_2(\Omega)}^2 + \|\nabla v\|_{L_2(\Omega)}^2$ one has

$$c_a\|v\|_1^2 \le a(v,v) \quad \text{and} \quad |a(v,w)| \le C_a\|v\|_1\|w\|_1, \quad v,w \in H_0^1(\Omega). \tag{2.3}$$

We shall discretize (2.2) with a discontinuous Galerkin scheme. To describe this it will be convenient to employ the following localized inner products. For any domain

$G \subseteq \Omega$ we set

$$\langle v, w\rangle_G := \int_G vw, \quad a(v,w)_G := \int_G (A\nabla u)\cdot\nabla v + \int_G buv.$$

2.1. The symmetric interior penalty method on locally refined triangulations. The DG discretization that we shall describe next will be based on triangulations $\mathcal{T}_h = \mathcal{T}_h(\Omega)$ that are generated from some initial shape regular conforming triangulation $\mathcal{T}^0$ through possibly local refinements. The subscript h therefore refers to a piecewise constant *mesh size function* given by $h(K) := \operatorname{diam}(K)$, $K \in \mathcal{T}_h$. In principle, any kind of local refinement is permitted that preserves shape regularity such as a partition based on "newest vertex insertion" or the standard partition of a triangle into four congruent subtriangles. One could think of closing the partitions arising from such local refinements to keep the triangulation conforming (which means that any two distinct closed triangles have an empty intersection or share a common face (edge or vertex)) while preserving shape regularity. We stress though that *nonconforming* triangulations are permitted as well; i.e., *hanging vertices* (sometimes called hanging nodes) may occur; see Figure 2.1. Since the occurrence of hanging vertices does complicate the construction of the preconditioner and its analysis somewhat and since conforming shape regular triangulations may be viewed as simpler special cases, we shall concentrate in the following on situations where hanging vertices occur. As usual we require the triangulations to be graded, meaning that an edge may contain at most one hanging vertex; see Figure 2.1.

Later we shall have to impose a slightly stronger grading property.

By denoting by $\mathbb{P}_k(G)$ the space of all polynomials of total degree at most k over the domain G, we associate with any (admissible) triangulation $\mathcal{T}_h$ the trial space

$$V_h := \mathbb{P}_k(\mathcal{T}_h) := \bigoplus_{K\in\mathcal{T}_h} \mathbb{P}_k(K) = \left\{ v = \sum_{K\in\mathcal{T}_h} \chi_K P_K : P_K \in \mathbb{P}_k(K) \right\}$$

of all (possibly discontinuous) piecewise polynomials of degree at most k on $\mathcal{T}_h$. Here χ_K is the standard indicator function, i.e., $\chi_K(x) = 1$ for $x \in K$ and $\chi_K(x) = 0$ otherwise. The subscript k is thus to be understood (in the same way as the mesh size function h) as a *degree function* taking a constant integer value $k = k(K)$ on each triangle $K \in \mathcal{T}_h$ that may vary for different K. However, we shall always assume that a fixed upper bound $\bar{k}$ is never exceeded, i.e., $k \le \bar{k}$ for all $K \in \mathcal{T}_h$.

For the derivation of the symmetric interior penalty method we refer, e.g., to [1, 2]. To formulate it in the present setting we need the following notation. Let $\mathcal{E}_h$ be the set of edges of the triangulation $\mathcal{T}_h$. Since hanging vertices are permitted we shall adopt the convention that whenever an edge has a hanging node it is replaced in $\mathcal{E}_h$ by its two halves. Moreover, $\mathcal{E}_{h,b}$ and $\mathcal{E}_{h,i}$ denote the subsets of boundary and interior edges, respectively. Since triangles are always assumed to be closed sets, for

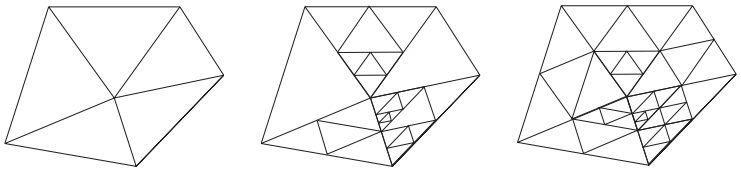


FIG. 2.1. *Coarse $\mathcal{T}^0$ (left), refinement (middle), and graded, shape regular $\mathcal{T}_h$ (right).*

any $e \in \mathcal{E}_{h,i}$ there exist $K, K' \in \mathcal{T}_h$ such that $e = K \cap K'$, and at least one of the two triangles has e as an edge. It will also be convenient to use the abbreviation

$$v^K := \chi_K v.$$

Thus for any $v \in V_h$ and $K \in \mathcal{T}_h$ it makes sense to denote by $v^K|_e$ the trace of the restriction of v to K on the subset e of ∂K. As usual the jump of v across e is then defined as

$$[v]_e := \mathbf{n}_{K,e}\, v^K|_e + \mathbf{n}_{K',e}\, v^{K'}|_e,$$

where $\mathbf{n}_{K,e} = -\mathbf{n}_{K',e}$ denotes the unit normal vector on e pointing to the exterior of K. When $e \in \mathcal{E}_{h,b}$ is a boundary edge we think of v being extended by zero outside Ω. Likewise, averages on e are defined as

$$\{v\}_e := \begin{cases} \frac{1}{2}(v^K|_e + v^{K'}|_e), & e \in \mathcal{E}_{h,i}, \\ v^K|_e, & e \in \mathcal{E}_{h,b} \end{cases}$$

(and we shall drop the subscript e when it is obvious from the context). Given $\mathcal{T}_h$, consider the bilinear form on V_h

(2.4)
$$a_h(v,w) := \sum_{K\in\mathcal{T}_h} a(v,w)_K - \sum_{e\in\mathcal{E}_h} \int_e \big(\{A\nabla w\}\cdot[v] + \{A\nabla v\}\cdot[w]\big) + \sum_{e\in\mathcal{E}_h} \frac{\gamma}{|e|}\int_e [w]\cdot[v].$$

Here γ is a positive (possibly piecewise) constant that has to be chosen appropriately to ensure stability of the method; see relation (2.7) below and, e.g., [19]. Its existence is usually argued under slightly different assumptions on the underlying meshes, but the argument easily carries over.

In fact, the only points to be addressed are perhaps the occurrence of a hanging vertex and a possible local variation of γ depending on the coefficients in (2.1). By applying a Cauchy–Schwarz inequality to (2.4), one gets

$$a_h(v,v) \geq \sum_{K\in\mathcal{T}_h} a(v,v)_K - 2\left(\sum_{e\in\mathcal{E}_h} \frac{|e|}{\sigma}\|\{A\nabla v\}\|^2_{L_2(e)}\right)^{1/2}\left(\sum_{e\in\mathcal{E}_h}\frac{\sigma}{|e|}\|[v]\|^2_{L_2(e)}\right)^{1/2} + \sum_{e\in\mathcal{E}_h}\frac{\gamma}{|e|}\|[v]\|^2_{L_2(e)}.$$

When $e = K \cap K'$ is an edge of K' but only half of an edge of K (see also Figure 4.1 in section 4), we can write for $v \in V_h$ the restriction v^K as $\sum_{J\in\boldsymbol{C}(K)} v^J$, where $\boldsymbol{C}(K)$ is the set of "children" triangles partitioning K; see section 4.1. Hence, by denoting by $\lambda_{\max\,/\,\min}(A)$ the maximal/minimal eigenvalue of A and defining for $\omega \subseteq \Omega$

$$\underline{\lambda}_\omega := \inf_{x\in\omega} \lambda_{\min}(A(x)), \quad \overline{\lambda}_\omega := \sup_{x\in\omega} \lambda_{\max}(A(x)), \tag{2.5}$$

the usual combination of a trace inequality with an inverse inequality yields

$$\begin{aligned} |e|\|\{A\nabla v\}\|^2_{L_2(e)} &\leq C(\|A\nabla v\|^2_{L_2(K')} + \|A\nabla v\|^2_{L_2(J)}) \\ &\leq C \max_{J'=K',J}\{\overline{\lambda}_{J'}\}\, a(v,v)_{K'\cup J}, \end{aligned} \tag{2.6}$$

where $J \subset K$ is the child of K adjacent to K' and the constant C depends on the shape properties of $\mathcal{T}^0$ and $\bar{k}$. Employing then Young's inequality in combination with (2.6) and a suitable choice of $\sigma > 0$ shows that there is indeed a positive γ (see Remark 2.1 below) such that

$$a_h(v,v) \geq \bar{c}\left(\sum_{K\in\mathcal{T}_h} a(v,v)_K + \sum_{e\in\mathcal{E}_h} \frac{\gamma}{|e|}\|[v]\|^2_{L_2(e)}\right) =: \bar{c}\,|\!|\!|v|\!|\!|_h^2\,, \quad v \in V_h, \tag{2.7}$$

for some uniform constant $\bar{c}$ that could actually be made $\bar{c} = \frac{1}{2}$, which yields the desired stability.

Remark 2.1. If we allow γ to vary, we should take $\gamma \geq 2\sigma$ where, for any edge of K, σ and hence γ should be of the order of $\overline{\lambda}_K$ up to constants depending on the shape properties of $\mathcal{T}^0$ and the degree k of the piecewise polynomials but not on h.

The same type of estimate establishes continuity

$$|a_h(v,w)| \leq \bar{C}\,|\!|\!|v|\!|\!|_h\,|\!|\!|w|\!|\!|_h\,, \quad v,w \in V_h, \tag{2.8}$$

for some uniform constant $\bar{C}$ that could actually be made $\bar{C} = \frac{3}{2}$ when taking γ sufficiently large but again of the order of $\overline{\lambda}_K$ locally. (In fact, continuity could be established even in a larger infinite-dimensional space; see, e.g., [13].) Hence, the problem

$$\text{find} \;\; u_h \in V_h \quad \text{such that} \;\; a_h(u_h,v) = \langle f, v\rangle \quad \forall v \in V_h \tag{2.9}$$

has a unique solution, where a_h is the symmetric bilinear form defined on V_h by (2.4).

Unfortunately, as in the case of conforming discretizations the corresponding linear systems of equations grow increasingly ill conditioned when the mesh size decreases. For instance, for quasi-uniform $\mathcal{T}_h$ the spectral condition numbers of stiffness matrices are known to grow like h^{-2}; see, e.g., [11].

3. Additive Schwarz schemes. It is now known that, for nonconforming discretizations, the abstract two-level method (see, e.g., [4, 17, 21]) allows us to reduce the task of preconditioning the resulting system to that of preconditioning an auxiliary system resulting from some auxiliary *conforming* discretization. In this section we shall explain why a "direct" classical multiwavelet approach fails to yield optimal preconditioners, and by following [16, 17] we shall formulate the more general framework for two-level methods and additive Schwarz schemes in the DG context that give rise to uniformly bounded condition numbers for the above type of discretizations under no additional regularity requirements on the solution u of (2.2).

3.1. A way that does not work. Schwarz-type preconditioners for conforming discretizations are closely related to change-of-basis preconditioners when energy space stable multiscale bases are available [7, 8]. To explain this, suppose that $\Psi = \{\psi_i : i \in \mathcal{I}\}$ is a basis of V_h, and recall that the spectral condition number $\kappa(\mathbf{L}_h^\Psi)$ of the stiffness matrix $\mathbf{L}_h^\Psi := (a_h(\psi_i,\psi_j))_{i,j\in\mathcal{I}}$ is given by

$$\kappa(\mathbf{L}_h^\Psi) = \left(\sup_{\mathbf{v}\neq\mathbf{0}} \frac{\mathbf{v}^T\mathbf{L}_h^\Psi\mathbf{v}}{\mathbf{v}^T\mathbf{v}}\right)\left(\inf_{\mathbf{v}\neq\mathbf{0}} \frac{\mathbf{v}^T\mathbf{L}_h^\Psi\mathbf{v}}{\mathbf{v}^T\mathbf{v}}\right)^{-1}.$$

Now according to (2.7) and (2.8), by writing $\mathbf{v} = (v_i)_{i\in\mathcal{I}} \neq \mathbf{0}$ we have, for $v = \sum_{i\in\mathcal{I}} v_i\psi_i$,

$$\frac{\mathbf{v}^T\mathbf{L}_h^\Psi\mathbf{v}}{\mathbf{v}^T\mathbf{v}} = \frac{a_h(v,v)}{\|\mathbf{v}\|^2_{\ell^2}} \sim \frac{|\!|\!|v|\!|\!|_h^2}{\|\mathbf{v}\|^2_{\ell^2}}.$$

This immediately yields the following fact.

Remark 3.1. If Ψ is stable in the sense that

$$\|\!|v|\!\|_h^2 \sim \sum_{i\in\mathcal{I}} |v_i|^2 \, \|\!|\psi_i|\!\|_h^2 \tag{3.1}$$

for any $v = \sum_{i\in\mathcal{I}} v_i\psi_i$ in V_h, then an optimal preconditioner is obtained by a symmetric diagonal scaling; namely, we have

$$\kappa(\mathbf{D}\mathbf{L}_h^{\Psi}\mathbf{D}) \sim 1, \quad \text{with} \quad \mathbf{D} = \operatorname{diag}\big(\, \|\!|\psi_i|\!\|_h^{-1} : i \in \mathcal{I}\big).$$

In the conforming case such energy space stable bases are typically derived from L_2-Riesz multilevel bases by rescaling. In the DG context it is very easy to construct L_2-stable multilevel bases for V_h by using the concept of multiwavelets. One might therefore be tempted to construct a change-of-basis preconditioner by suitably rescaling such a multiwavelet basis. For such a multilevel basis of V_h, a refinement $\mathcal{T}_{h'}$ of the mesh $\mathcal{T}_h$ entails expanding the current basis Ψ_h for V_h to the larger basis $\Psi_{h'}$ by adding a collection of new basis functions (on corresponding locally higher refinement levels) that span a complementary space W of V_h so that $V_{h'} = V_h \oplus W$. In the case of multiwavelets all basis functions would consist of globally discontinuous piecewise polynomials, and in the present context this turns out to cause a principal problem that we describe now.

Consider a *discontinuous* basis function ψ_i of Ψ, and assume that it belongs to some space V_h. Clearly there is at least one edge e in $\mathcal{E}_h$ for which $\|[\psi_i]\|_{L_2(e)}$ is positive. Now, if $\mathcal{T}_{h'}$ is a refinement of $\mathcal{T}_h$ where e is uniformly refined into 2^l pieces, we have

$$\|\!|\psi_i|\!\|_{h'}^2 \geq \sum_{e'\in\mathcal{E}_{h'}:e'\subset e} \frac{\gamma}{|e'|}\|[\psi_i]\|_{L_2(e')}^2 = 2^l\,\frac{\gamma}{|e|}\|[\psi_i]\|_{L_2(e)}^2.$$

Hence $\|\!|\psi_i|\!\|_{h'}$ tends to infinity as the edge e is further refined. As a first consequence, this shows that the ψ_i will not remain normalized when progressing to higher levels of refinements causing the mesh-dependent norm $\|\!|\cdot|\!\|_h$ to change.

This does not exclude yet stability in the sense of (3.1). On the other hand, the energy norm of a *continuous* $v \in V_h$ reduces to $a(v,v)^{1/2}$ and hence does not depend on the mesh. If ψ_i remains unchanged when expanding the basis of V_h to a basis of $V_{h'}$, the associated coordinate v_i does not change either. The desired stability of Ψ would thus yield

$$a(v,v) = a_h(v,v) = \|\!|v|\!\|_{h'}^2 \sim \sum_{i'\in\mathcal{I}_{h'}} |v_{i'}|^2 \, \|\!|\psi_{i'}|\!\|_{h'}^2 \geq |v_i|^2 \, \|\!|\psi_i|\!\|_{h'}^2 \,,$$

which clearly implies that $v_i = 0$, since otherwise the (constant) quantity $a(v,v)$ would have to tend to infinity with $h' \to 0$.

From this observation we infer that, when Ψ is a uniformly stable multilevel basis for the spaces V_h, it must have a subset Ψ^{c} that spans the conforming spaces $V_h^{\mathrm{c}} := V_h \cap H_0^1(\Omega)$. In particular, Ψ and V_h, respectively, decompose into

$$\Psi = \Psi^{\mathrm{c}} \cup \Psi^{\mathrm{nc}} \quad \text{and} \quad V_h = V_h^{\mathrm{c}} \oplus V_h^{\mathrm{nc}}$$

in such a way that subsets of Ψ^{c} and Ψ^{nc} are stable bases for V_h^{c} and V_h^{nc}, respectively, and that the splitting is stable with respect to the discrete energy norm $\|\!|\cdot|\!\|_h$.

This observation that appears to exclude standard multiwavelets will guide the subsequent developments.

3.2. Additive Schwarz preconditioners. It is of course not necessary to construct stable multilevel bases to produce efficient preconditioners for the problem (2.9) of finding $u_h \in V_h$ such that

$$a_h(u_h, v) = \langle f, v\rangle, \quad v \in V_h. \tag{3.2}$$

In fact, the theory of *two-level methods* and *additive Schwarz schemes* offers a more flexible framework based on so-called stable splittings for the respective energy space which in turn permit redundant systems, namely, frames. Such stable splittings for the DG setting will be constructed in section 4.3 below. First let us briefly recall the notion of such splittings and under which circumstances they give rise to good Schwarz preconditioners. The collection

$$\mathcal{S} = \{V_i\}_{i\in\mathcal{I}_h}, \qquad V_i \subset V_h,$$

is called a ($|||\cdot|||_h$-)stable splitting of V_h, uniformly with respect to all triangulations $\mathcal{T}_h$ of the specified type, if $\sum_{i\in\mathcal{I}_h} V_i = V_h$ and

$$c_S\,|||v|||_h^2 \le \inf_{\substack{v_i\in V_i\\ v=\sum_{i\in\mathcal{I}_h} v_i}} \left\{\sum_{i\in\mathcal{I}_h} |||v_i|||_h^2\right\} \le C_S\,|||v|||_h^2\,, \quad v \in V_h, \tag{3.3}$$

holds with constants c_S and C_S independent of $\mathcal{T}_h$.

With the aid of such stable splittings one can transform (3.2) into an equivalent problem which will be uniformly well conditioned.

To this end, the next ingredients are auxiliary inner products on the spaces V_i, namely,

$$b_i(\cdot,\cdot) : V_i \times V_i \to \mathbb{R}, \quad i \in \mathcal{I}_h,$$

yielding norms that are locally equivalent to $|||\cdot|||_h$, i.e., satisfy

$$c_b\,|||v_i|||_h^2 \le b_i(v_i, v_i) \le C_b\,|||v_i|||_h^2\,, \quad v_i \in V_i,\ i \in \mathcal{I}_h, \tag{3.4}$$

for constants c_b and C_b depending at most on the maximal degree $\bar{k}$, the shape properties of $\mathcal{T}^0$, and possibly the coefficients in the bilinear form $a(\cdot,\cdot)$ (see (4.37) below for concrete choices).

Finally, define the operators $P_i : V_h \to V_i$ and elements $f_i \in V_i$ by

$$b_i(P_i w, v_i) = a_h(w, v_i), \quad b_i(f_i, v_i) = \langle f, v_i\rangle, \quad v_i \in V_i,\ i \in \mathcal{I}_h. \tag{3.5}$$

Note that, whenever $V_i = \operatorname{span}(\theta_i)$ is a one-dimensional space, the application of P_i just amounts to solving a linear equation with a single unknown, namely,

$$P_i w = d_i\theta_i \quad\Longleftrightarrow\quad d_i = a_h(w,\theta_i)/b_i(\theta_i,\theta_i). \tag{3.6}$$

The central result we shall use reads then as follows; see, e.g., [16, 12].

THEOREM 3.1. *Define $P_h : V_h \to V_h$ and $\bar{f}_h \in V_h$ by*

$$P_h := \sum_{i\in\mathcal{I}_h} P_i, \qquad \bar{f}_h := \sum_{i\in\mathcal{I}_h} f_i. \tag{3.7}$$

Then the problem (3.2) *is equivalent to the operator equation*

$$P_h u_h = \bar{f}_h. \tag{3.8}$$

Moreover, if (3.3) *and* (3.4) *hold, then the spectral values of the symmetric positive definite (with respect to the inner product $a_h(\cdot,\cdot)$) operator P_h can be bounded by constants independent of $\mathcal{T}_h$, namely,*

(3.9)
$$\frac{\bar{c}}{C_b C_S} \leq \lambda_{\min}(P_h) := \inf_{\substack{v\in V_h\\ v\neq 0}} \frac{a_h(P_h v, v)}{a_h(v,v)}, \qquad \sup_{\substack{v\in V_h\\ v\neq 0}} \frac{a_h(P_h v, v)}{a_h(v,v)} =: \lambda_{\max}(P_h) \leq \frac{\bar{C}}{c_b c_S},$$

where $\bar{c} = \frac{1}{2}$ and $\bar{C} = \frac{3}{2}$ are the constants from (2.7) *and* (2.8)*, respectively. In particular, the spectral condition number of P_h is bounded independently of $\mathcal{T}_h$:*

$$\kappa(P_h) := \frac{\lambda_{\max}(P_h)}{\lambda_{\min}(P_h)} \leq \frac{\bar{C} C_b C_S}{\bar{c} c_b c_S}. \tag{3.10}$$

Obviously the complexity of applying the operator P_h in some iterative procedure has linear complexity when the dimension of the subspaces V_i is uniformly bounded. Such stable splittings will be constructed in sections 4.3 and 5 below; see in particular Corollary 5.3 for one specific realization.

Now, to interpret the above result, let us view (3.2) as the operator equation

$$L_h u_h = f_h, \quad \text{where } L_h \text{ is defined by } \quad \langle L_h w, v\rangle = a_h(w,v), \qquad w, v \in V_h,$$

and f_h is the L_2-orthogonal projection of f on V_h. Moreover, by defining $L_i : V_i \to V_i$ by $b_i(w_i, v_i) = \langle L_i w_i, v_i\rangle$ (i.e., L_i^{-1} is the approximate inverse of the restriction of L_h to V_i) and denoting by $Q_i : V_h \to V_i$ the V_i-orthogonal projection, we have by definition

$$P_i = L_i^{-1} Q_i L_h.$$

Therefore, we can write as usual

$$P_h u = \left(\sum_{i\in\mathcal{I}_h} L_i^{-1} Q_i\right) L_h u \quad \text{and} \quad \bar{f} = \left(\sum_{i\in\mathcal{I}_h} L_i^{-1} Q_i\right) f_h; \tag{3.11}$$

i.e., the operator

$$\mathcal{C}_h := \sum_{i\in\mathcal{I}_h} L_i^{-1} Q_i \tag{3.12}$$

can be viewed as a preconditioner.

4. Stable splittings for the DG method. This section is devoted to the construction and analysis of stable multilevel splittings for the DG method that can be used in Theorem 3.1. There are two ingredients that are perhaps worth emphasizing.

- The first one is that stable splittings for V_h can simply be assembled from stable splittings for an appropriate conforming and nonconforming part of V_h, respectively. The appropriateness will hinge on certain admissible averaging operators that are required to satisfy a Jackson estimate. In fact, this estimate would also allow one to conclude optimality of the preconditioners by utilizing the abstract framework of general two-level (also called "auxiliary space") methods in [4, 17, 21]. Indeed, this approach has been applied in [10] to a variety of DG schemes, including SIPG, but under the restriction that the meshes are quasi-uniform and have no hanging nodes.

- Second, the locality of such an estimate, however, turns out to impose a slightly stronger notion of gradedness (see Definition 4.1 below) for the underlying meshes (without actually inflating the complexity; see Remark 4.2 below) which we shall therefore address first.

4.1. Admissible triangulations—extended grading. We have already defined a standard grading requirement prohibiting an edge to have more than one hanging vertex. The grading that we shall have to impose on the triangulations will actually be slightly stronger but essentially keeps the complexity equivalent; see Remark 4.2 below.

To limit technicalities we shall consider only triangulations that are obtained by local subdivisions into four congruent subtriangles, starting from a conforming and shape regular coarse $\mathcal{T}^0$, as in Figure 2.1. As a consequence, every element of $\mathcal{T}_h$ has a *refinement level* $j(K)$, where $j(K) = 0$ means that $K \in \mathcal{T}^0$. From the shape regularity of $\mathcal{T}_h$ (inherited from that of $\mathcal{T}^0$), one then easily infers that

$$|K| := \operatorname{meas}(K) \sim (h|_K)^2 \sim 2^{-2j(K)} \quad \text{for any} \quad K \in \mathcal{T}_h, \tag{4.1}$$

with constants that depend only on the coarse $\mathcal{T}^0$.

DEFINITION 4.1 (graded triangulations). *A triangulation $\mathcal{T}_h$ obtained from refining a conforming coarse $\mathcal{T}^0$ as described above is called* graded *if every hanging vertex n is the midpoint of an edge spanned by two vertices, denoted by n' and n'' in Figure* 4.1 *(left) and if in such a case n' and n'' are never hanging vertices themselves.*

Note that this grading requirement implies the usual one; i.e., there is at most one hanging vertex per edge.

Remark 4.1 (necessity of the (extended) grading). The above grading requirement is necessary for Theorem 1.1 (and further similar splittings) to hold.

To see the validity of the above claim consider one triangulation $\mathcal{T}_h$ that has been obtained from one single coarse triangle $\mathcal{T}^0 = \{K_0\}$ by recursively refining the central child at each level up to some j_h, as shown in Figure 4.1 (right). Clearly, such a $\mathcal{T}_h$ satisfies the "classical" grading requirement permitting only one hanging vertex per edge but not the above stronger one. In particular, all of its vertices except those on $\partial\Omega$ are hanging, so that $H^1_0(\Omega) \cap \mathbb{P}_1(\mathcal{T}_h)$ contains only 0. Hence Theorem 1.1 would imply here that, for any basis $\{\phi_i : i \in \mathcal{I}_h\}$ of $V_h := \mathbb{P}_k(\mathcal{T}_h)$ consisting of local polynomial pieces normalized in L_∞, one has

$$\|\!|v|\!\|_h^2 \sim \sum_{i\in\mathcal{I}_h} |d_i|^2 \quad \text{for every} \quad v = \sum_{i\in\mathcal{I}_h} d_i\phi_i \in V_h, \tag{4.2}$$

with constants independent of the refinement level j_h. Consider now the piecewise constant function $v = \chi_K$, where K is the central triangle on level two. If the ϕ_i are

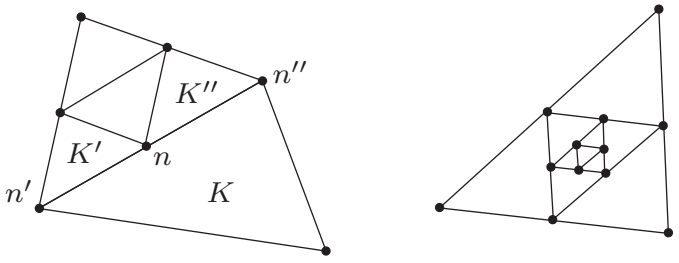


FIG. 4.1. *For $\mathcal{T}_h$ to be graded (the right one is not), n' and n'' must be regular vertices.*

interpolatory, such as in a local Lagrange basis, one has $d_i = 1$ for all but finitely many indices i and the right-hand side in (4.2) grows proportionally to j_h. On the other hand, since v is continuous inside K, it can be seen from the definition of the energy norm given in section 2.1 (see (2.7)) that $\|\!|v|\!\|_h$ does not depend on the further refinements of $\mathcal{T}_h$, which clearly contradicts (4.2).

We shall next give an algorithmic characterization of the grading property in Definition 4.1. Since the triangulations $\mathcal{T}_h$ of interest result from (possibly local) refinements, such a refinement history defines a tree $\widehat{\mathcal{T}}_h$, where $\mathcal{T}_h$ is the set of leaves (i.e., final nodes) of $\widehat{\mathcal{T}}_h$. We denote by $\boldsymbol{P}(K)$ and $\boldsymbol{C}(K)$ the *parent* and set of *children*, respectively, of a given triangle K. This means that any child $K' \in \boldsymbol{C}(K)$ is obtained by subdividing its parent $K = \boldsymbol{P}(K')$ once, according to the given rule. By $\boldsymbol{P}^2(K) = \boldsymbol{P}(\boldsymbol{P}(K))$ we denote the *grandfather* of K when $j(K) \geq 2$.

Also, by $\mathcal{T}^j$ we denote the jth uniform dyadic refinement of $\mathcal{T}^0$. Clearly, for any level j this is a partition of Ω. Given any triangle K in $\mathcal{T}_h$, we then let its *neighbors* $\boldsymbol{N}(K)$ be defined as the collection of triangles at level $j(K)$, i.e., belonging to $\mathcal{T}^{j(K)}$ (but not necessarily to $\widehat{\mathcal{T}}_h$) that share an edge with K. Finally, $\boldsymbol{N}^2(K) := \cup_{K' \in \boldsymbol{N}(K)} \boldsymbol{N}(K')$ consists of the neighbors of the neighbors of K. In these terms a constructive formulation of the grading property reads as follows.

Remark 4.2 (algorithmic formulation of the grading requirement). A triangle K in $\mathcal{T}_h$ is said to have the **G**-property if it is of level $j(K) \leq 1$ or if

- in the case that $\boldsymbol{P}(K)$ is not the central child of $\boldsymbol{P}^2(K)$, every triangle of $\boldsymbol{N}(\boldsymbol{P}(K))$ is in the tree $\widehat{\mathcal{T}}_h$, or
- in the case that $\boldsymbol{P}(K)$ is the central child of $\boldsymbol{P}^2(K)$, every triangle of $\boldsymbol{N}^2(\boldsymbol{P}(K))$ is in the tree $\widehat{\mathcal{T}}_h$.

Then the following properties hold, whose proofs are left to the reader:

1. A triangulation $\mathcal{T}_h$ is graded in the sense of Definition 4.1 if and only if all of its elements satisfy the **G**-property.
2. Given any multilevel triangulation $\mathcal{T}_h$, its smallest graded refinement is built by imposing recursively (starting from the highest levels) the **G**-property on all of its elements.

Therefore the **G**-property is a practical tool for constructing graded triangulations. Moreover, it can be checked that the set

$$\widehat{\mathcal{T}}_h^* := \widehat{\mathcal{T}}_h \cup \bigcup_{K \in \widehat{\mathcal{T}}_h} \boldsymbol{C}(\boldsymbol{N}^2(\boldsymbol{P}(K))) \tag{4.3}$$

is indeed a tree and that the corresponding set of leaves $\mathcal{T}_h^*$ is a refinement of $\mathcal{T}_h$ in which any triangle has the **G**-property. $\mathcal{T}_h^*$ is then a graded refinement of $\mathcal{T}_h$, and it is a standard exercise (see, for instance, Lemma 2.4 in [6]) to derive from (4.3) the following complexity estimates:

$$\#(\mathcal{T}_h^*) \sim \#(\widehat{\mathcal{T}}_h^*) \sim \#(\widehat{\mathcal{T}}_h) \sim \#(\mathcal{T}_h),$$

which hold with uniform constants. Thus, imposing our (extended) grading keeps the complexity of adaptive triangulations up to uniform constants.

We shall henceforth assume that $\mathcal{T}_h$ is graded in the sense of Definition 4.1.

4.2. Admissible conforming subspaces. Given a graded triangulation $\mathcal{T}_h$, it remains to construct energy-stable splittings to be used in Theorem 3.1. Motivated by the observations made in section 3.1, we shall first look at decomposing V_h into a conforming and a nonconforming part V_h^{c} and V_h^{nc}, respectively. However, it turns

out that, in the more general framework offered by Schwarz preconditioners, even the meaning of V_h^c can be relaxed in that it need not exhaust all of $V_h \cap H_0^1(\Omega)$. Different conforming subspaces will be seen to lead to different versions of preconditioners that are all asymptotically optimal in the sense of uniformly bounded condition numbers.

What matters is that the conforming part is the range of a suitable *averaging projector* $\mathcal{A}$ into some subspace of $V_h \cap H_0^1(\Omega)$. In this section we formulate the key property that qualifies the averaging projector as suitable, namely, a certain local Jackson-type estimate involving for each $K \in \mathcal{T}_h$ the localized energy norms

$$|\!|\!| v |\!|\!|_{h,\omega}^2 := \sum_{K \in \mathcal{T}_h : K \subset \omega} a(v,v)_K + \sum_{e \in \mathcal{E}_h : e \subset \omega} \frac{\gamma}{|e|} \|[v]\|_{L_2(e)}^2$$

for certain neighborhoods $\omega = \omega(K) \subset \Omega$ of any $K \in \mathcal{T}_h$.

To define $\omega(K)$ we need some further preparations and notational conventions. The set of vertices, viz., *first degree nodes*, will be denoted by

$$\mathcal{N}_1(K) := \{\text{vertices of } K\} \quad \text{and} \quad \mathcal{N}_{h,1} := \bigcup_{K \in \mathcal{T}_h} \mathcal{N}_1(K).$$

In connection with higher order elements we shall have to consider later corresponding sets of higher kth degree nodes that will be denoted by $\mathcal{N}_k(K)$; see section 4.3.

Moreover, it will be convenient to set for any closed domain D

$$\mathcal{T}_h(D) := \{K \in \mathcal{T}_h : K \cap D \neq \emptyset\}.$$

Similarly we define the sets $\mathcal{N}_{h,1}(D)$ and $\mathcal{E}_h(D)$ for mesh elements touching D. For instance, $\mathcal{T}_h(n)$ consists of the triangles that share the vertex n (including hanging vertices). Note that, when an edge of a triangle K contains a hanging vertex, the collections $\mathcal{N}_1(K)$ and $\mathcal{N}_{h,1}(K)$ differ by that hanging vertex. Unfortunately, due to hanging vertices, straightforward neighborhoods based on these notions do not suffice, and the construction of the domains $\omega(K)$ appearing in the announced Jackson estimate will actually rely on some *extended* sets of mesh quantities.

To this end, recall that for any vertex n there are two possibilities: Either it is hanging (i.e., there is one triangle K in $\mathcal{T}_h(n)$ for which n is not a vertex), and then n is the midpoint of one edge $e = [n', n'']$ of K (see Figure 4.1), or it is not. With these notations we first define

$$\mathcal{N}_{h,1}^*(n) := \begin{cases} \{n, n', n''\} & \text{if } n \text{ is hanging,} \\ \{n\} & \text{otherwise, i.e., if } n \text{ is regular.} \end{cases} \tag{4.4}$$

Recall that the grading we impose on $\mathcal{T}_h$ precisely means that both n' and n'' are regular; see Definition 4.1. For any triangle $K \in \mathcal{T}_h$ we then set

$$\mathcal{N}_{h,1}^*(K) := \bigcup_{n \in \mathcal{N}_1(K)} \mathcal{N}_{h,1}^*(n) \quad \subset \mathcal{N}_{h,1}, \tag{4.5}$$

$$\mathcal{E}_h^*(K) := \bigcup_{n \in \mathcal{N}_{h,1}^*(K)} \mathcal{E}_h(n) \quad \subset \mathcal{E}_h, \tag{4.6}$$

and

$$\mathcal{T}_h^*(K) := \bigcup_{n \in \mathcal{N}_{h,1}^*(K)} \mathcal{T}_h(n) \quad \subset \mathcal{T}_h \tag{4.7}$$

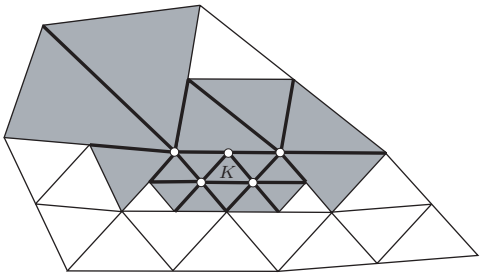


FIG. 4.2. *The neighborhood $\omega(K)$ consists of the union of the (gray) triangles in $\mathcal{T}_h^*(K)$.*

and finally define the domain

$$\omega(K) := \bigcup_{K' \in \mathcal{T}_h^*(K)} K' \quad \subset \Omega \tag{4.8}$$

as the union of triangles that are in contact with the extended set of vertices $\mathcal{N}_{h,1}^*(K)$. An illustration is presented in Figure 4.2, where the sets $\mathcal{N}_{h,1}^*(K)$, $\mathcal{E}_h^*(K)$, and $\mathcal{T}_h^*(K)$ are represented by white vertices, bold edges, and gray triangles, respectively.

Note in particular that the above definitions yield

$$\sum_{K' \in \mathcal{T}_h^*(K)} a(v,v)_{K'} + \sum_{e \in \mathcal{E}_h^*(K)} \frac{\gamma}{|e|} \|[v]\|_{L_2(e)}^2 \le |\!|\!| v |\!|\!|_{h,\omega(K)}^2 . \tag{4.9}$$

We are now prepared to formulate the key property of averaging projectors $\mathcal{A}$ that will be crucial for the subsequent developments.

PROPERTY 4.3. *A linear projector $\mathcal{A}$ that takes V_h into $V_h \cap H_0^1(\Omega)$ is called admissible if for any $v \in V_h$ one has the following Jackson estimate:*

$$\|(I - \mathcal{A})v\|_{L_2(K)} \le C^* 2^{-j(K)} |\!|\!| v |\!|\!|_{h,\omega(K)} , \qquad K \in \mathcal{T}_h, \tag{4.10}$$

with a constant C^* that may depend on the shape properties of $\mathcal{T}^0$ and on $\bar{k}$ (and on concrete realizations of γ). Note that the domains $\omega(K)$ defined by (4.8) *are local (closed) neighborhoods of the triangles $K \in \mathcal{T}_h$ satisfying a bounded overlapping property, i.e.,*

$$\sup_{K \in \mathcal{T}_h} \#\big(\{K' \in \mathcal{T}_h : \omega(K) \cap \omega(K') \neq \emptyset\}\big) \lesssim 1 \tag{4.11}$$

holds with a constant that depends only on $\mathcal{T}^0$.

We postpone the concrete construction of admissible $\mathcal{A}$ and the verification of Property 4.3 to section 4 and discuss first the consequences regarding stable splittings.

The significance of Property 4.3 lies in the following two consequences, summarized in Propositions 4.2 and 4.3. The first one concerns the energy stability of an admissible $\mathcal{A}$.

PROPOSITION 4.2. *Assume that $\mathcal{A}$ satisfies the Jackson estimate* (4.10). *Then one has*

$$|\!|\!| (I - \mathcal{A})v |\!|\!|_h \lesssim |\!|\!| v |\!|\!|_h , \quad |\!|\!| \mathcal{A}v |\!|\!|_h \lesssim |\!|\!| v |\!|\!|_h , \quad v \in V_h, \tag{4.12}$$

where the constants depend on $\bar{k}$ and the shape properties of $\mathcal{T}^0$ as well as on C^* and $\bar{C}_a$ defined in (4.14) *below.*

As a consequence we have

$$\|\!|v|\!\|_h \le \|\!|\mathcal{A}v|\!\|_h + \|\!|(I-\mathcal{A})v|\!\|_h \le C_{\mathcal{A}}\,\|\!|v|\!\|_h\,, \tag{4.13}$$

with the above dependence of the constant $C_{\mathcal{A}}$ in the upper bound.

Proof. Since the jumps of $(I-\mathcal{A})v$ equal the jumps of v we need only estimate $\sum_{K\in\mathcal{T}_h} a((I-\mathcal{A})v,(I-\mathcal{A})v)_K$. To keep track of the dependence of the various constants, recall that $\overline{\lambda}_K$ is defined by (2.5), and let

$$C_{a,K} := 2\max\{\overline{\lambda}_K, 2^{-2j(K)}\|b\|_{L_\infty(K)}\} \ \text{ and } \ \bar{C}_a := \max_{K\in\mathcal{T}_h} C_{a,K} \ \lesssim \ C_a; \tag{4.14}$$

see (2.3). Since $(I-\mathcal{A})v$ is a polynomial on K, a standard inverse estimate yields

$$a((I-\mathcal{A})v,(I-\mathcal{A})v)_K \le cC_{a,K}2^{2j(K)}\|(I-\mathcal{A})v\|^2_{L_2(K)} \ \lesssim \ C^*C_{a,K}\,\|\!|v|\!\|^2_{h,\omega(K)}\,,$$

where c depends on k and $\mathcal{T}^0$ and where we have used (4.10) again in the last step. On account of the bounded overlap of the $\omega(K), K\in\mathcal{T}_h$ (see (4.11)), and since $\|\!|\mathcal{A}v|\!\|_h \le \|\!|v|\!\|_h + \|\!|(I-\mathcal{A})v|\!\|_h$, this concludes the proof. ∎

Thus, by defining

$$V_h^{\rm c} := \mathcal{A}V_h \quad \text{and} \quad V_h^{\rm nc} := (I-\mathcal{A})V_h \tag{4.15}$$

as possible candidates for conforming and nonconforming parts of V_h, respectively, the above observation (4.13) already indicates that a stable splitting for all of V_h of the form (3.3) can be composed of stable splittings for the individual subspaces $V_h^{\rm c}$ and $V_h^{\rm nc}$, respectively. In fact, this would allow us to invoke known splittings for conforming spaces; see, e.g., [3, 7, 12, 16], and section 5.1.1.

Now the second consequence of Property 4.3 is a localization of $V_h^{\rm nc}$ that will simplify the identification of stable splittings for the nonconforming part. In what follows we shall make frequent use of the following simple consequences of standard trace inequalities, rescaling arguments, and the fact that all norms on a fixed finite-dimensional space are equivalent.

Remark 4.4. For any polynomial $P\in\mathbb{P}_k$, any $K\in\mathcal{T}_h$, and any edge e of K, the following relations hold:

(i) One has

$$\|P\|_{L_\infty(e)} \sim |e|^{-1/2}\|P\|_{L_2(e)}, \tag{4.16}$$

where the constants depend only on the degree k of P and on $\mathcal{T}_0$.

(ii) There exists a constant c depending only on the degree k of P and on $\mathcal{T}^0$ such that

$$\|\!|\chi_K P|\!\|_h \le c(C_{a,K}+\gamma)^{\frac12}\left\{\begin{array}{l} 2^{j(K)}\|P\|_{L_2(K)},\\ \|P\|_{L_\infty(K)},\end{array}\right. \tag{4.17}$$

where the constant c depends only on k and on the shape properties of $\mathcal{T}^0$ and $C_{a,K}$ is given by (4.14).

Observe then that splittings for $V_h^{\rm nc}$ will indeed take a simple form due to localization.

PROPOSITION 4.3. *Assume that $\mathcal{A}$ is admissible, i.e., satisfies* (4.10). *Then there exists a constant C_0, depending only on $\bar{k}$, on the shape properties of $\mathcal{T}^0$, on $\bar{C}_a$ defined in* (4.14), *and on the constant C^* in* (4.10) *such that for $w^K := \chi_K w$*

$$\|\!|w|\!\|_h^2 \le \sum_{K\in\mathcal{T}_h} \|\!|w^K|\!\|_h^2 \le C_0\,\|\!|w|\!\|_h^2\,, \quad w\in V_h^{\rm nc} := (I-\mathcal{A})V_h. \tag{4.18}$$

Proof. The lower inequality in (4.18) is due to the fact that for any edge $e \subset K \cap K'$ one has

$$\|[w]\|^2_{L_2(e)} \leq \|w^K\|^2_{L_2(e)} + \|w^{K'}\|^2_{L_2(e)}.$$

As for the upper inequality, since $w \in V_h^{\rm nc}$ and w^K is a polynomial on K, we infer from (4.17) and (4.10) (noticing that $(I-\mathcal{A})w = w$) that

$$\left|\!\left|\!\left| w^K \right|\!\right|\!\right|_h^2 \lesssim (C_{a,K}+\gamma)2^{2j(K)}\|w^K\|^2_{L_2(K)} \lesssim (C_{a,K}+\gamma)C^* \left|\!\left|\!\left| w \right|\!\right|\!\right|^2_{h(\omega(K))}, \tag{4.19}$$

where $C_{a,K}$ can be bounded by C_a from (2.3). The upper inequality follows now again from (4.11). ∎

4.3. Energy-stable splittings for V_h. As mentioned above, the stability of $\mathcal{A}$ asserted by Proposition 4.2 implies that stable splittings for the individual parts $V_h^{\rm c}$ and $V_h^{\rm nc}$ give rise to stable splittings for all of V_h. Moreover, the localization in $V_h^{\rm nc}$ given by Proposition 4.3 together with the fact that the frame for the nonconforming part need actually not be contained in $V_h^{\rm nc}$ greatly simplifies finding such a complementary frame as reflected by the following result (see [5] for placing this in the context of general two-level methods).

THEOREM 4.4. *Let $\{V_i^{\rm c}\}_{i\in\mathcal{I}_h^{\rm c}}$ be a stable splitting for $V_h^{\rm c}$, i.e.,*

$$c_1 a(v,v) \leq \inf_{\substack{v_i\in V_i^{\rm c}\\ v=\sum_{i\in\mathcal{I}_h^{\rm c}} v_i}} \left\{ \sum_{i\in\mathcal{I}_h^{\rm c}} a(v_i,v_i) \right\} \leq C_1 a(v,v), \quad v \in V_h^{\rm c}, \tag{4.20}$$

holds for some constants c_1 and C_1 independent of $\mathcal{T}_h$. Then one has for any $v \in V_h$

$$c_S \left|\!\left|\!\left| v \right|\!\right|\!\right|_h^2 \leq \inf_{\substack{v_K,\ v_i\\ v=\sum_{K\in\mathcal{T}_h} v_K+\sum_{i\in\mathcal{I}_h^{\rm c}} v_i}} \left\{ \sum_{K\in\mathcal{T}_h} \left|\!\left|\!\left| v_K \right|\!\right|\!\right|_h^2 + \sum_{i\in\mathcal{I}_h^{\rm c}} a(v_i,v_i) \right\} \leq C_S \left|\!\left|\!\left| v \right|\!\right|\!\right|_h^2, \tag{4.21}$$

where the v_K and v_i belong to $\chi_K\mathbb{P}_k(K)$ and $V_i^{\rm c}$, respectively, $c_S = \frac{1}{2}\min\{1,c_1\}$, and $C_S = C_{\mathcal{A}}\max\{C_0,C_1\}$ (C_0 being defined in Proposition 4.3).

Thus the collection

$$\mathcal{S} = \{\chi_K\mathbb{P}_k(K) : K \in \mathcal{T}_h\} \cup \{V_i^{\rm c} : i \in \mathcal{I}_h^{\rm c}\} \tag{4.22}$$

is a stable splitting for V_h in the sense of (3.3).

Proof. By definition of the infimum we have

$$\begin{aligned}
&\inf_{\substack{v_K,\ v_i\\ v=\sum_{K\in\mathcal{T}_h} v_K+\sum_{i\in\mathcal{I}_h^{\rm c}} v_i}} \left\{ \sum_{K\in\mathcal{T}_h} \left|\!\left|\!\left| v_K \right|\!\right|\!\right|_h^2 + \sum_{i\in\mathcal{I}_h^{\rm c}} a(v_i,v_i) \right\}\\
&\qquad \leq \inf_{\substack{v_K\\ (I-\mathcal{A})v=\sum_{K\in\mathcal{T}_h} v_K}} \left\{ \sum_{K\in\mathcal{T}_h} \left|\!\left|\!\left| v_K \right|\!\right|\!\right|_h^2 \right\} + \inf_{\substack{v_i\\ \mathcal{A}v=\sum_{i\in\mathcal{I}_h^{\rm c}} v_i}} \left\{ \sum_{i\in\mathcal{I}_h^{\rm c}} a(v_i,v_i) \right\}\\
&\qquad \leq C_0 \left|\!\left|\!\left| (I-\mathcal{A})v \right|\!\right|\!\right|_h^2 + C_1 \left|\!\left|\!\left| \mathcal{A}v \right|\!\right|\!\right|_h^2 \leq C_S \left|\!\left|\!\left| v \right|\!\right|\!\right|_h^2,
\end{aligned} \tag{4.23}$$

where we have used Proposition 4.3 and (4.20) in the second step (remember that $a(\cdot,\cdot)$ and $\left|\!\left|\!\left| \cdot \right|\!\right|\!\right|_h^2$ coincide on $V_h^{\rm c}$) and Proposition 4.2 in the last step. Thus $C_S = C_{\mathcal{A}}\max\{C_0,C_1\}$. This confirms the upper bound in (4.21).

In order to confirm the other direction consider any expansion

$$v = \sum_{K\in\mathcal{T}_h} \tilde{v}_K + \sum_{i\in\mathcal{I}_h^c} \tilde{v}_i =: v_1 + v_2,$$

with $\tilde{v}_K \in \chi_K \mathbb{P}_k(K)$ and $\tilde{v}_i \in V_i^c$. By writing first

$$\|\!|v|\!\|_h^2 \le 2(\|\!|v_1|\!\|_h^2 + \|\!|v_2|\!\|_h^2), \tag{4.24}$$

we note that by (4.20)

$$\|\!|v_2|\!\|_h^2 = a(v_2,v_2) \le c_1^{-1} \inf_{v_2=\sum_{i\in\mathcal{I}_h^c} v_i} \left\{ \sum_{i\in\mathcal{I}_h^c} a(v_i,v_i) \right\} \le c_1^{-1} \sum_{i\in\mathcal{I}_h^c} a(\tilde{v}_i,\tilde{v}_i), \tag{4.25}$$

while as in the proof of Proposition 4.3 one has for any $v_1 = \sum_{K\in\mathcal{T}-h} \tilde{v}_K \in V_h$

$$\|\!|v_1|\!\|_h^2 \le \sum_{K\in\mathcal{T}_h} \|\!|\tilde{v}_K|\!\|_h^2 . \tag{4.26}$$

Thus, we can combine (4.25) and (4.26) to conclude that

$$\|\!|v|\!\|_h^2 \le 2 \left(\sum_{K\in\mathcal{T}_h} \|\!|\tilde{v}_K|\!\|_h^2 + c_1^{-1} \sum_{i\in\mathcal{I}_h^c} a(\tilde{v}_i,\tilde{v}_i) \right).$$

Since the decomposition of v was arbitrary we have shown the lower inequality in (4.21) for $\frac{1}{2}\min\{1,c_1\}$. ∎

Theorem 4.4 covers various possible specifications of splittings which we shall begin to discuss now. Recall first that, in principle, energy-stable splittings for conforming finite element spaces are well understood. Of course, a concrete identification of collections $\{V_i^c\}_{i\in\mathcal{I}_h^c}$ depends on the concrete realization of an admissible projector $\mathcal{A}$, i.e., on the specific subspace $\mathcal{A}V_h$. We postpone this issue to section 5.

At this point we note first that the subspaces $V_K^{nc} := \chi_K \mathbb{P}_k(K)$ are of uniformly bounded finite dimension depending only on $\bar{k}$ and therefore could be further broken down. In fact, any (reasonable) basis for V_h will be seen to be an energy-stable frame for V_h^{nc}. To describe such a basis it is convenient to introduce the principal lattice of order k (also called the canonical k-mesh)

$$\mathcal{N}_k(K) := \left\{ p = p_\beta := \frac{1}{k} \sum_{j=0}^{2} \beta_j n_j : \beta \in \mathbb{Z}_+^3, \sum_{j=0}^{2} \beta_j = k \right\}, \tag{4.27}$$

induced on the triangle K; see Figure 4.3. Of course, for $k = 1$ we simply have

$$\mathcal{N}_1(K) = \{n_0, \ldots, n_2\},$$

the set of vertices of K itself, as used before. Note that the β/k are the barycentric coordinates of the mesh points $p = p_\beta$, a fact that will be used later again. Clearly we have $\dim(\mathbb{P}_k) = \#\mathcal{N}_k(K)$.

Now let $\{P_p : p \in \mathcal{N}_k(\hat{K})\}$ be a fixed basis for $\mathbb{P}_k$, where $\hat{K}$ is the standard triangle with vertices $(0,0)$, $(1,0)$, and $(0,1)$, and define for $K = F\hat{K} \in \mathcal{T}_h$, $F = F_K$ affine,

$$\phi_{K,p} := \chi_K P_p \circ F_K^{-1}, \quad p \in \mathcal{N}_k(K). \tag{4.28}$$

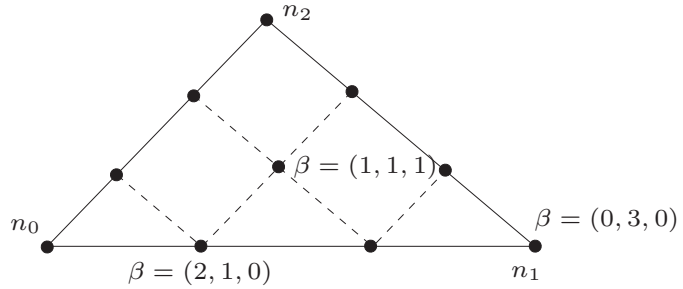


FIG. 4.3. *Principal lattice for $k = 3$, $\mathcal{N}_3(K)$.*

Obviously, $\Phi_h := \{\phi_{K,p} : p \in \mathcal{N}_k(K), K \in \mathcal{T}_h\}$ is a basis for V_h. Later, the reference polynomials P_p will be (essentially) normalized in L_∞ (or in H^1, which is equivalent in $\mathbb{R}^2$). Thus one has

$$\|P_p\|_{L_\infty(\hat{K})} \sim 1, \quad \|\{a_p\}_{p\in\mathcal{N}_k(\hat{K})}\|_{\ell_\infty} \sim \Big\| \sum_{p\in\mathcal{N}_k(\hat{K})} a_p P_p \Big\|_{L_\infty(\hat{K})}, \tag{4.29}$$

with constants depending only on k (and the specific choice of the polynomial basis).

Theorem 4.4 yields then the following general result, where it is seen that a complementary frame is simply given by a union of local polynomial basis functions that do not need to belong to the nonconforming part $V_h^{\rm nc}$ but only to span it.

COROLLARY 4.5. *Assume that $\mathcal{A}$ is admissible (i.e., satisfies the Jackson estimate* (4.10)*), and let $\{V_i^{\rm c} : i \in \mathcal{I}_h^{\rm c}\}$ be any energy-stable splitting for $V_h^{\rm c} := \mathcal{A}V_h$. Moreover, let the basis functions $\phi_{K,p}$ be constructed as above, i.e., satisfy* (4.28)*, and assume that the set of indices*

$$\mathcal{I}_h^{\rm nc} \subseteq \{i = (K,p) : p \in \mathcal{N}_k(K), K \in \mathcal{T}_h\} \tag{4.30}$$

satisfies

$$V_h^{\rm nc} := (I - \mathcal{A})V_h \subseteq \operatorname{span}\{\phi_i : i \in \mathcal{I}_h^{\rm nc}\} =: \operatorname{span}\Phi_h^{\rm nc}. \tag{4.31}$$

Then, by setting $V_i^{\rm nc} := \operatorname{span}\phi_i$, the collection

$$\mathcal{S} = \{V_i^{\rm nc} : i \in \mathcal{I}_h^{\rm nc}\} \cup \{V_i^{\rm c} : i \in \mathcal{I}_h^{\rm c}\} \tag{4.32}$$

is an energy-stable splitting for V_h in the sense of (3.3).

Proof. In view of Theorem 4.4, it suffices to estimate the local contributions $\|\!|w^K|\!\|_h$, $w^K = \chi_K w$, for $w \in V_h^{\rm nc}$. More precisely, we shall prove that

$$\sum_{K\in\mathcal{T}_h} \|\!|w^K|\!\|_h^2 \lesssim \sum_{(K,p)\in\mathcal{I}_h^{\rm nc}} \|\!|w_{K,p}\phi_{K,p}|\!\|_h^2 \lesssim \sum_{K\in\mathcal{T}_h} \|\!|w^K|\!\|_h^2 \tag{4.33}$$

holds with constants depending only on γ, $\bar{k}$, the shape properties of $\mathcal{T}^0$, and the constants C^*, c_a, and C_a in (4.10) and (2.3). Indeed by writing

$$w^K = \sum_{(K,p)\in\mathcal{I}_h^{\rm nc}} w_{K,p}\phi_{K,p},$$

we note from (4.28) and (4.29) that $\|\phi_{K,p}\|_{L_\infty(K)} \sim 1$. Hence, by (4.17) we have $\|\!|w_{K,p}\phi_{K,p}|\!\|_h^2 \lesssim (C_{a,K}+\gamma)|w_{K,p}|^2$ with a further constant factor depending on $\bar{k}$ and the shape properties of $\mathcal{T}^0$. By using the fact that only a fixed finite number (depending on $\bar{k}$) of basis functions are supported on each K, we conclude that

$$\begin{aligned}\|\!|w^K|\!\|_h^2 &\lesssim \sum_{(K,p)\in\mathcal{I}_h^{\mathrm{nc}}} \|\!|w_{K,p}\phi_{K,p}|\!\|_h^2 \lesssim \sum_{(K,p)\in\mathcal{I}_h^{\mathrm{nc}}} |w_{K,p}|^2 \lesssim \left\|\{w_{K,p}\}_{(K,p)\in\mathcal{I}_h^{\mathrm{nc}}}\right\|_{\ell_\infty}^2 \\ &\lesssim \left\|\sum_{(K,p)\in\mathcal{I}_h^{\mathrm{nc}}} w_{K,p}\phi_{K,p}\right\|_{L_\infty(K)}^2 = \|w\|_{L_\infty(K)}^2, \qquad (4.34)\end{aligned}$$

where the constants depend only on $\bar{k}$ as well as on $C_{a,K}$ and γ. Now we use again a scaling/norm equivalence argument and (4.10) to obtain

$$\|w\|_{L_\infty(K)} \lesssim 2^{j(K)}\|w\|_{L_2(K)} \lesssim C^* \|\!|w|\!\|_{h(\omega(K))}, \qquad (4.35)$$

where the remaining unspecified constant factor depends only on $\bar{k}$ and the shape properties of $\mathcal{T}^0$. Finally, we note again that for any edge $e \subset K' \cap K''$ one has $\|[w]\|_{L_2(e)}^2 \le \|w^{K'}\|_{L_2(e)}^2 + \|w^{K''}\|_{L_2(e)}^2$ and hence

$$\|\!|w|\!\|_{h(\omega(K))}^2 \le \sum_{J\in\mathcal{T}_h : J\subset\omega(K)} \|\!|w^J|\!\|_h^2. \qquad (4.36)$$

Combining now (4.34), (4.35), and (4.36) and using again (4.11) yields (4.33) and completes the proof. ∎

For the concrete realization of the preconditioned problem (3.8), it now remains to identify viable choices of the auxiliary bilinear forms $b_i(\cdot,\cdot)$ for $i \in \mathcal{I}_h := \mathcal{I}_h^{\mathrm{c}} \cup \mathcal{I}_h^{\mathrm{nc}}$. Here $\mathcal{I}_h^{\mathrm{c}}$ and $\mathcal{I}_h^{\mathrm{nc}}$ are the index sets for stable splittings $\{V_i^{\mathrm{c}}\}$ and $\{V_i^{\mathrm{nc}}\}$ of V_h^{c} and V_h^{nc}, respectively, as discussed above. Likewise, the conforming splitting typically contains the full coarse grid space. In general, let $\omega(i)$ denote the union of all of the supports of the nodal basis functions spanning the subspace V_i^{c} (which could be all of Ω when i refers to a full coarse level space).

Some admissible choices for the $b_i(v,w)$ are the following:

$$\begin{cases} a(v,w)_K + \sum_{e\in\mathcal{E}_h, e\subset\partial K} \frac{\gamma}{|e|}\int_e [v][w] & \text{or} \\ |K|^{-1}\langle v,w\rangle_K + \sum_{e\in\mathcal{E}_h, e\subset\partial K} \frac{\gamma}{|e|}\int_e [v][w] & \text{when } \ i=(K,p)\in\mathcal{I}_h^{\mathrm{nc}}, \\ a(v,w)_{\omega(i)} & \text{or} \\ |\omega(i)|^{-1}\langle v,w\rangle_{\omega(i)} & \text{when } \ i\in\mathcal{I}_h^{\mathrm{c}}. \end{cases} \qquad (4.37)$$

Remark 4.5. In all of the above cases in (4.37) there exist constants c_b and C_b depending only on the degree k, the shape properties of $\mathcal{T}^0$, and possibly the coefficients in the bilinear form $a(\cdot,\cdot)$ such that (3.4) holds.

Thus we can summarize these findings as follows.

THEOREM 4.6. *Assume that $\mathcal{A}$ is admissible. Then the additive Schwarz preconditioner based on the the above splittings (for the respective realizations of V_h^{c}; see section* 5.1.1*) and for any of the bilinear forms from* (4.37) *is asymptotically optimal in the sense of Theorem* 3.1.

The robustness of the scheme with respect to the diffusion coefficients in $a(\cdot,\cdot)$ will be discussed in more detail elsewhere.

5. Projecting onto the minimal conforming subspace. In principle the results of the previous section offer a general framework for a family of stable splittings and associated preconditioners. In this paper we are content with one specific realization of an admissible projector $\mathcal{A}$ and the identification of corresponding stable splittings for the conforming part. Specifically, we choose $\mathcal{A} = \mathcal{A}_1$ so as to map V_h into continuous *piecewise affine elements*—in this sense the smallest possible subspace of H^1_0-conforming trial functions, i.e.,

$$V^{\mathrm{c}}_{h,1} := \mathcal{A}_1 V_h = H^1_0(\Omega) \cap \mathbb{P}_1(\mathcal{T}_h). \tag{5.1}$$

In general it differs from $V_h \cap H^1_0(\Omega)$ whenever higher degrees are permitted.

An alternative would be to have $\mathcal{A} = \mathcal{A}_k$ produce globally continuous finite elements of the *same degree* as the original discontinuous elements and in that sense to split off the *maximal* conforming subspace in V_h, i.e.,

$$V^{\mathrm{c}}_{h,k} := \mathcal{A}_k V_h = V_h \cap H^1_0(\Omega). \tag{5.2}$$

This is elaborated on in [5].

In principle, such averaging operators are familiar in other contexts concerning DG discretizations; see, e.g., [13]. The form of such projectors in the presence of hanging vertices is to our knowledge less familiar. Since hanging vertices do appear to add some complication we shall give in what follows an explicit self-contained construction.

5.1. The construction of $\mathcal{A}_1$. The construction of $\mathcal{A}_1$ is inspired by the results in [13, 15] but has to cope with hanging nodes. For this reason, let us remember that we call *regular* those vertices that are not hanging. In particular, every vertex on the boundary $\partial\Omega$ is regular. Since boundary vertices are subject to a zero boundary condition in $H^1_0(\Omega)$, we can define $\mathcal{A}_1 v$ by prescribing its nodal values at every interior regular vertex n, following

$$(\mathcal{A}_1 v)(n) := \frac{1}{\#\big(\mathcal{T}_h(n)\big)} \sum_{K \in \mathcal{T}_h(n)} v^K(n). \tag{5.3}$$

Note that the same operator (denoted $\tilde{Q}_1$) is used in [10] to build a two-level method, but there the authors establish the Jackson estimate (4.10) only for piecewise affine DG spaces and in the case where the mesh is quasi-uniform and has no hanging nodes, as far as we understand.

Since $V^{\mathrm{c}}_{h,1}$ is always a subset of V_h (that is, for any degree distribution k), the space

$$V^{\mathrm{nc}}_{h,1} := (I - \mathcal{A}_1) V_h \tag{5.4}$$

stands as a possible candidate for complementing $V^{\mathrm{c}}_{h,1}$; i.e., we have

$$V_h = V^{\mathrm{c}}_{h,1} \oplus V^{\mathrm{nc}}_{h,1}.$$

5.1.1. Energy-stable splittings for $V^{\mathrm{c}}_{h,1}$. To describe an energy frame $\Phi^{\mathrm{c}}_{h,1}$ for $V^{\mathrm{c}}_{h,1}$ we can essentially resort to known results; see, e.g., [20, 22]. In fact $\Phi^{\mathrm{c}}_{h,1}$ has multilevel structure and is comprised by all nodal continuous basis functions associated with all predecessors of $\mathcal{T}_h$ within the tree $\widehat{\mathcal{T}}_h$. To describe this, let us denote again by $\mathcal{T}^j$ the jth uniform dyadic refinement of $\mathcal{T}^0$ and let

$$\mathcal{T}^j_h := \{K \in \mathcal{T}^j : \exists \;\; K' \in \mathcal{T}_h \;\; \text{such that} \;\; K' \subseteq K\} = \mathcal{T}^j \cap \widehat{\mathcal{T}}_h; \tag{5.5}$$

i.e. $\mathcal{T}_h^j$ is comprised of all level j triangles that appear in the refinement history leading to $\mathcal{T}_h$. Moreover, let

$$\Omega_h^j := \bigcup \{K : K \in \mathcal{T}_h^j\}$$

be the area covered by the elements in $\mathcal{T}_h^j$. The subspace $V_{h,1}^{\mathrm{c}} := \mathcal{A}_1 V_h$ consists of all continuous piecewise linear functions on $\mathcal{T}_h$. Consider now $\mathcal{N}_{h,1}^{j,\mathrm{c}}$ the set of those vertices that belong to a triangle in $\mathcal{T}_h^j$ and lie in the interior of Ω_h^j, i.e.,

$$\mathcal{N}_{h,1}^{j,\mathrm{c}} := \bigcup \{\mathcal{N}_1(K) \setminus \partial\Omega_h^j : K \in \mathcal{T}_h^j\}.$$

Note that, since hanging vertices are always located on some boundary $\partial\Omega_h^j$, the vertices in $\mathcal{N}_{h,1}^{j,\mathrm{c}}$ are neither hanging nor do they lie in the boundary of the corresponding domains.

Now define the multilevel index set

$$\mathcal{I}_{h,1}^{\mathrm{c}} := \bigcup \{i = (j,n) : j = 0, \dots, j_h,\ n \in \mathcal{N}_{h,1}^{j,\mathrm{c}}\}, \tag{5.6}$$

where j_h is the maximal level appearing in $\mathcal{T}_h$. With each $i = (j,n) \in \mathcal{I}_{h,1}^{\mathrm{c}}$ we associate now the standard nodal (piecewise affine) hat function φ_i^{c} at the vertex n supported on the star of triangles in $\mathcal{T}_h^j$ sharing n. Clearly φ_i^{c} is continuous, and it immediately follows from (4.17) that

$$\|\!|\varphi_i^{\mathrm{c}}|\!\|_h = a(\varphi_i^{\mathrm{c}}, \varphi_i^{\mathrm{c}})^{1/2} \sim 1, \quad i \in \mathcal{I}_{h,1}^{\mathrm{c}}, \tag{5.7}$$

holds with constants depending only on the constants in (2.3) and on the shape properties of $\mathcal{T}^0$. Now set

$$\Phi_{h,1}^{\mathrm{c}} := \{\varphi_i^{\mathrm{c}} : i \in \mathcal{I}_{h,1}^{\mathrm{c}}\}.$$

It is easy to see that $\Phi_{h,1}^{\mathrm{c}}$ spans $V_{h,1}^{\mathrm{c}}$, and it is known (see, e.g., [16, 12, 7]) that this collection is an energy-stable splitting, i.e.,

$$a(v,v) \sim \inf_{\substack{v_i \in \mathbb{R} \\ v = \sum_{i \in \mathcal{I}_{h,1}^{\mathrm{c}}} v_i \varphi_i^{\mathrm{c}}}} \left\{ \sum_{i \in \mathcal{I}_{h,1}^{\mathrm{c}}} v_i^2 a(\varphi_i^{\mathrm{c}}, \varphi_i^{\mathrm{c}}) \right\}, \quad v \in V_{h,1}^{\mathrm{c}}, \tag{5.8}$$

where the constants depend only on $\mathcal{T}^0$ and the constants c_a and C_a.

5.1.2. The Jackson estimate for $\mathcal{A}_1$. We shall confirm now the validity of Property 4.3 for $\mathcal{A} = \mathcal{A}_1$. To this end, let us first establish the following intermediate result that is valid only for *piecewise affine* functions. Remember that the different sets $\mathcal{E}_h(n)$, $\mathcal{N}_{h,1}^*(n)$, $\mathcal{T}_h(n), \dots$, which denote local neighborhoods made of mesh elements, are defined in section 4.2.

LEMMA 5.1. *If $v \in \mathbb{P}_1(\mathcal{T}_h)$ is a piecewise affine function, then we have for any node $n \in \mathcal{N}_{h,1}$*

$$\left|(v - \mathcal{A}_1 v)^K(n)\right| \le \sum_{n' \in \mathcal{N}_{h,1}^*(n)} \sum_{e \in \mathcal{E}_h(n')} \|[v]\|_{L_\infty(e)} \quad \forall K \in \mathcal{T}_h. \tag{5.9}$$

Proof. We begin with the following basic estimate, independent of the operator $\mathcal{A}_1$: At any vertex $n \in \mathcal{N}_{h,1}$, we have

$$|(v^K - v^{K'})(n)| \le \sum_{e \in \mathcal{E}_h(n)} \|[v]\|_{L_\infty(e)}, \quad K,\ K' \in \mathcal{T}_h(n), \tag{5.10}$$

for any $v \in V_h$. Indeed, since K and K' do not necessarily share an edge, write $\mathcal{T}_h(n) = \{K_1, \ \ldots, K_M\}$ the (closed) triangles containing n, in such a way that K_i and K_{i+1} have the edge $e_i \in \mathcal{E}_h(n)$ as their intersection, and assume that $K = K_M$ and $K' = K_m$, with $m < M$ (observe from the definition of the set $\mathcal{E}_h$ that this is possible even when n is a hanging vertex). It is then easily seen that

$$|(v^K - v^{K'})(n)| \le \sum_{i=m}^{M-1} |(v^{K_{i+1}} - v^{K_i})(n)| \le \sum_{i=m}^{M-1} \|[v]\|_{L_\infty(e_i)},$$

which clearly gives (5.10).

Actually, because the jumps over boundary edges are defined as if v was extended to 0 outside Ω, this argument yields the following estimate when n is on $\partial\Omega$:

$$|v^K(n)| \le \sum_{e \in \mathcal{E}_h(n)} \|[v]\|_{L_\infty(e)}, \qquad n \in \partial\Omega, \ K \in \mathcal{T}_h(n). \tag{5.11}$$

Now to prove the assertion of the lemma we consider $n \in K$ and distinguish three cases: (i) n is on the boundary $\partial\Omega$, (ii) n is a regular vertex in the interior of Ω, and finally (iii) n is hanging; i.e., there is (only) one triangle that contains n and for which n is not a vertex. Note that in the first two cases we have $\mathcal{N}^*_{h,1}(n) = \{n\}$.

In case (i), (5.11) readily yields Lemma 5.1 since then we have $\mathcal{A}_1 v = 0$.

In case (ii) where n is interior and regular, by definition (5.3) of $\mathcal{A}_1$, we have

$$\big|(v - \mathcal{A}_1 v)^K(n)\big| \le \frac{1}{\#(\mathcal{T}_h(n))} \sum_{K' \in \mathcal{T}_h(n)} \big|(v^K - v^{K'})(n)\big|, \tag{5.12}$$

so that Lemma 5.1 now follows from (5.10).

Finally, in case (iii), where n is hanging, denote by K' the unique triangle in $\mathcal{T}_h(n)$ for which n is not a vertex and by $[n', n'']$ the edge of K' for which n is the midpoint. Since both v and $\mathcal{A}_1 v$ are affine on K', we have

$$\big|(v - \mathcal{A}_1 v)^{K'}(n)\big| \le \frac{1}{2}\Big(\big|(v - \mathcal{A}_1 v)^{K'}(n')\big| + \big|(v - \mathcal{A}_1 v)^{K'}(n'')\big|\Big). \tag{5.13}$$

Now from the grading of $\mathcal{T}_h$, we know that n' and n'' are never hanging themselves, so that (5.12) applies, providing, on account of (5.10),

$$\big|(v - \mathcal{A}_1 v)^{K'}(\check{n})\big| \le \sum_{e \in \mathcal{E}_h(\check{n})} \|[v]\|_{L_\infty(e)}, \quad \check{n} = n' \text{ or } n''. \tag{5.14}$$

By writing then

$$\big|(v - \mathcal{A}_1 v)^K(n)\big| \le \big|v^K(n) - v^{K'}(n)\big| + \big|(v - \mathcal{A}_1 v)^{K'}(n)\big|,$$

we prove (5.9) in case (iii) by combining the estimates (5.13), (5.10), and (5.14), since then we have by definition $\mathcal{N}^*_{h,1}(n) = \{n, n', n''\}$; see (4.4). ∎

We are now ready to verify the validity of Property 4.3 for $\mathcal{A}_1$.

PROPOSITION 5.2. *The projector $\mathcal{A}_1$ has Property* 4.3 *where C^* in* (4.10) *depends only on the shape properties of $\mathcal{T}^0$, on $\bar{k}$, and on* $\max_{K' \subset \omega(K)} \underline{\lambda}_{K'}^{-1}$.

Proof. To estimate the projection error $(I - \mathcal{A}_1)v$ consider the interpolation operator $P_1 : V_h \to \mathbb{P}_1(\mathcal{T}_h)$ given by

$$\{P_1 v\}^K(n) := v^K(n), \qquad K \in \mathcal{T}_h, \ n \in \mathcal{N}_1(K), \tag{5.15}$$

for which the estimate

$$\|(I-P_1)v\|_{L_2(K)} \lesssim 2^{-j(K)}\|\nabla v\|_{L_2(K)} \tag{5.16}$$

is well known with a constant depending only on the shape properties of $\mathcal{T}^0$. In view of (5.3), it is readily seen that $\mathcal{A}_1 P_1 = \mathcal{A}_1$, so that the projection error reads

$$(I-\mathcal{A}_1)v = (I-P_1)v + (I-\mathcal{A}_1)P_1 v.$$

By writing $w := (I-\mathcal{A}_1)P_1 v \in \mathbb{P}_1(\mathcal{T}_h)$ it follows that

$$\|(I-\mathcal{A}_1)v\|_{L_2(K)} \lesssim 2^{-j(K)}\|\nabla v\|_{L_2(K)} + \|w\|_{L_2(K)}. \tag{5.17}$$

Now denote by $\{\phi^1_{(K,n)} : n \in \mathcal{N}_1(K)\}$ the discontinuous nodal basis of $\mathbb{P}_1(K)$. On any triangle $K \in \mathcal{T}_h$ we have

$$\|w\|_{L_2(K)} \le \sum_{n\in\mathcal{N}_1(K)} |w^K(n)|\, \|\phi^1_{(K,n)}\|_{L_2} \lesssim \sum_{n\in\mathcal{N}_1(K)} 2^{-j(K)}|w^K(n)|.$$

By applying then Lemma 5.1 to $P_1 v$ and recalling that the set $\mathcal{E}_h^*(K)$ is given by (4.6), we find

$$\|w\|_{L_2(K)} \lesssim 2^{-j(K)} \sum_{e\in\mathcal{E}_h^*(K)} \|[P_1 v]\|_{L_\infty(e)},$$

where the constant depends only on the shape properties of $\mathcal{T}^0$.

Now assume for the moment that for any edge $e \in \mathcal{E}_h$ there is one triangle K_e for which $e \subset \partial K_e$ and such that the inequality

$$\|[P_1 v]\|_{L_\infty(e)} \lesssim \|[v]\|_{L_\infty(e)} + \|\nabla v\|_{L_2(K_e)} \tag{5.18}$$

holds with a constant that depends only on the shape properties of $\mathcal{T}^0$. Since for every e in $\mathcal{E}_h^*(K)$, any such K_e is in $\mathcal{T}_h^*(K)$ (see (4.6) and (4.7)), this would yield

$$\|w\|_{L_2(K)} \lesssim 2^{-j(K)} \left(\sum_{e\in\mathcal{E}_h^*(K)} \|[v]\|_{L_\infty(e)} + \sum_{K'\in\mathcal{T}_h^*(K)} \|\nabla v\|_{L_2(K')} \right). \tag{5.19}$$

Now $[v]$ is polynomial on any e, so (4.16) reads $\|[v]\|_{L_\infty(e)} \sim |e|^{-1/2}\|[v]\|_{L_2(e)}$. Moreover, each term $\|\nabla v\|^2_{L_2(K')}$ can be bounded by $\underline{\lambda}_{K'}^{-1} a(v,v)_{K'}$, and indeed we have

$$a(v,v)_{K'} = \int_{K'} \left((A\nabla v)\cdot v + bv^2\right) \ge \|A^{1/2}\nabla v\|^2 \ge \underline{\lambda}_{K'}\|\nabla v\|^2_{L_2(K')}.$$

By using the fact that the sets $\mathcal{E}_h^*(K)$ and $\mathcal{T}_h^*(K)$ are of uniformly bounded cardinality (depending only on $\mathcal{T}^0$), we therefore find by combining (5.17), (5.19), and (4.9) that

$$\|(I-\mathcal{A}_1)v\|_{L_2(K)} \le C^* 2^{-j(K)} \,|\!|\!|v|\!|\!|_{h(\omega(K))},$$

where

$$C^* = C\Big(\max\big\{\{\gamma^{-1}\} \cup \{\underline{\lambda}_{K'}^{-1} : K' \in \mathcal{T}_h^*(K)\}\big\}\Big)^{1/2}$$

and where the unspecified C depends only on $\bar{k}$ and the shape properties of $\mathcal{T}^0$. This completes the proof of Lemma 4.3.

Thus, it remains only to prove (5.18). Here three situations may occur: Either (1) the edge e is on the boundary $\partial\Omega$, or this is not the case and we can write $e = K_e \cap K'_e$, with $K_e, K'_e \in \mathcal{T}_h$. Now either (2) e is an edge of both K_e and K'_e, or again (3) this is not the case. In the two first cases, we observe that $[P_1 v]_e = P_1([v]_e)$, so that

$$\|[P_1 v]\|_{L_\infty(e)} \le \|[v]\|_{L_\infty(e)}$$

is obvious. Now according to the structure of graded triangulations, the only possible configuration in case (3) is represented in Figure 4.1 with $e = [n', n]$. By setting then K_e and K'_e to be the coarser and finer triangles, respectively (i.e., K and K' in Figure 4.1), we observe that $P_1 v$ is affine on K_e. Therefore

$$\|[P_1 v]\|_{L_\infty(e)} = \max\left\{ |v^{K_e}(n') - v^{K'_e}(n')|, \left| \frac{1}{2}\big(v^{K_e}(n') + v^{K_e}(n'')\big) - v^{K'_e}(n) \right| \right\}.$$

By writing

$$\frac{1}{2}\big(v^{K_e}(n') + v^{K_e}(n'')\big) = \frac{1}{2}\big(v^{K_e}(n') - v^{K_e}(n)\big) + \frac{1}{2}\big(v^{K_e}(n) - v^{K_e}(n'')\big) + v^{K_e}(n),$$

we find (with an absolute constant)

$$\|[P_1 v]\|_{L_\infty(e)} \lesssim \|[v]\|_{L_\infty(e)} + |v^{K_e}(n') - v^{K_e}(n)| + |v^{K_e}(n) - v^{K_e}(n'')|.$$

By using a scaling argument together with the fact that v is polynomial on K_e, one finally verifies that

$$|v^{K_e}(n) - v^{K_e}(\check{n})| \lesssim \|\nabla v\|_{L_2(K_e)}, \quad \check{n} = n' \text{ or } n'',$$

holds with a constant that depends only on the shape properties of $\mathcal{T}^0$. By collecting the above estimates, one confirms (5.18) and completes the proof. ∎

Combining Corollary 4.5, (5.8), and Proposition 5.2 yields the following result.

COROLLARY 5.3. *Let $\mathcal{I}^c_{h,1}$ and φ^c_i be defined as in section* 5.1.1. *Then, for any basis $\{\phi_{K,p} : K \in \mathcal{T}_h,\ p \in \mathcal{N}_k(K)\}$ of V_h constructed as in* (4.28) *(such as a piecewise Lagrange basis), the collection*

$$\mathcal{S}_1 := \{\mathrm{span}(\varphi^c_i) : i \in \mathcal{I}^c_{h,1}\} \cup \{\mathrm{span}(\phi_{K,p}) : K \in \mathcal{T}_h,\ p \in \mathcal{N}_k(K)\}$$

is a stable splitting of V_h in the sense of (3.3). *In particular, the corresponding problem* (3.8) *is equivalent to the original equation* (3.2) *and has a condition number bounded by a constant independent of $\mathcal{T}_h$.*

Remark 5.1 (proof of Theorem 1.1). Theorem 1.1 is an immediate consequence of Corollary 4.5 and Proposition 5.2, observing that V^c_h can always be regarded as a splitting of itself. ∎

6. Numerical experiments. We now present some first numerical experiments for the simple model problem $a(u, v) = \int_\Omega \nabla u \cdot \nabla v$, where Ω is an L-shaped domain in $\mathbb{R}^2$ obtained by cutting out the upper right square $(1, 2]^2$ from $[0, 2]^2$. In the first two examples we consider only the first degree case $\bar{k} = 1$ on a uniformly refined grid. At this point we are mainly interested in the quantitative influence of the specific choice of the auxiliary bilinear forms $b_i(\cdot, \cdot)$, of the stabilization parameter γ, and of the low-dimensional subspaces V_i in the nonconforming part.

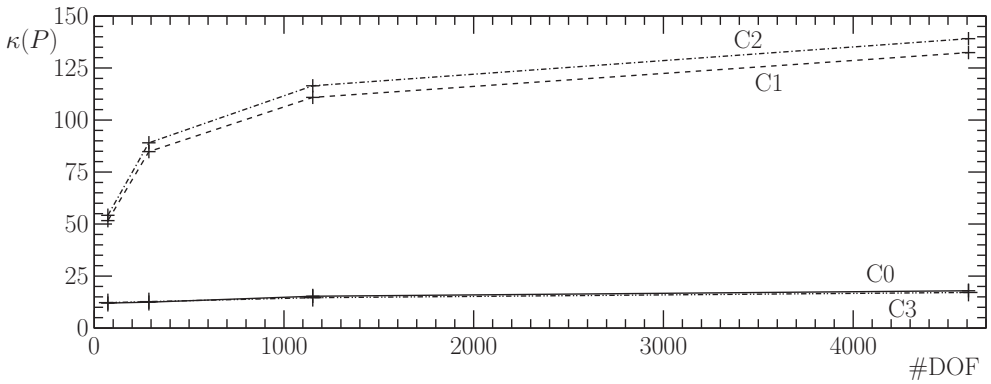


FIG. 6.1. *Condition numbers for the cases C0, C1, C2, and C3 for $\gamma = 10$ and $d_i = 1$.*

We shall use the following notation: $d_i := \dim V_i$ denotes the dimension of the subspaces in the nonconforming splitting; $d_i = 3$ refers to the case that V_i is spanned by the three nodal basis functions on each triangle in $\mathcal{T}_h$.

The case C0 refers to the choice

$$b_i(v,w) = \begin{cases} a(v,w)_K + \sum_{e\in\mathcal{E}_h, e\in\partial K} \frac{\gamma}{|e|}\int_e [v][w] & \text{for} \quad i = (K,p) \in \mathcal{I}_h^{\mathrm{nc}}, \\ a(v,w)_{\omega(i)} & \text{for} \quad i = (j,p) \in \mathcal{I}_h^{\mathrm{c}} \end{cases} \tag{6.1}$$

of the auxiliary inner products (see (4.37)), while C1 corresponds to

$$b_i(v,w) = \begin{cases} |K|^{-1}\langle v,w\rangle_K + \sum_{e\in\mathcal{E}_h, e\in\partial K} \frac{\gamma}{|e|}\int_e [v][w] & \text{for} \quad i = (K,p) \in \mathcal{I}_h^{\mathrm{nc}}, \\ 2^{2j}\,\langle v,w\rangle_{\omega(i)} & \text{for} \quad i = (j,p) \in \mathcal{I}_h^{\mathrm{c}}. \end{cases} \tag{6.2}$$

In order to test the preconditioning effect we have calculated the condition numbers. Figure 6.1 shows the dependence of the condition number of the preconditioned operator $\kappa(P_h)$ on the number of degrees of freedom for four selections of $b_i(\cdot,\cdot)$. For completeness, in addition to C0 and C1 we here also consider the cases C2 and C3, which are given by combining the first line of (6.1) with the second line of (6.2) and vice versa, respectively. Qualitatively for all choices of $b_i(\cdot,\cdot)$ the condition numbers of the preconditioned operator P_h show the same bounded behavior. The influence of the choice of the auxiliary bilinear form for the conforming part on the condition number dominates the choice for the nonconforming part. Selecting the bilinear form of the problem as the auxiliary bilinear form for the nonconforming part in cases C0 and C3 yields much smaller condition numbers than the scaled L_2 scalar product. The results in cases C0 and C3 show a very slight difference favoring the use of the scaled L_2 scalar product for the nonconforming part.

Recall that $\gamma = 10$ is usually considered as an appropriate choice. We consider first the moderate values $\gamma = 5$ and $\gamma = 7.5$ in the case $d_i = 1$.

Level	#DOF	κ for preconditioner (case/γ)					
		None/5	C0/5	C1/5	None/7.5	C0/7.5	C1/7.5
1	72	29.56	11.22	50.31	42.57	10.96	48.15
2	288	103.88	13.04	81.22	153.64	11.55	78.52
3	1152	403.17	16.31	105.98	601.26	14.50	102.56
4	4608	1602.08	18.98	126.37	2395.59	16.99	122.41

The results for $\gamma = 10$ and $\gamma = 15$ are shown in the next table.

Level	#DOF	κ for preconditioner (case/γ)					
		None/10	C0/10	C1/10	None/15	C0/15	C1/15
1	72	55.44	12.02	51.66	80.92	13.80	58.02
2	288	203.15	12.46	84.80	301.78	14.33	95.77
3	1152	799.07	15.28	110.87	1194.24	16.79	125.31
4	4608	3188.93	17.97	132.40	4775.39	19.84	149.72

The choice C0 is superior to C1 in both cases. We turn now to the case $d_i = 3$ for $\gamma = 5$ and $\gamma = 7.5$

Level	#DOF	κ for preconditioner (case/γ)					
		None/5	C0/5	C1/5	None/7.5	C0/7.5	C1/7.5
1	72	29.56	8.55	47.71	42.57	8.09	47.08
2	288	103.88	12.36	78.13	153.64	10.82	76.17
3	1152	403.17	15.60	101.91	601.26	13.76	99.48
4	4608	1602.08	18.24	121.52	2395.59	16.20	118.73

and further again for $\gamma = 10$ and $\gamma = 15$

Level	#DOF	κ for preconditioner (case/γ)					
		None/10	C0/10	C1/10	None/15	C0/15	C1/15
1	72	55.44	8.58	49.53	80.92	9.51	54.42
2	288	203.15	11.21	80.52	301.78	12.09	88.96
3	1152	799.07	14.35	105.24	1194.24	15.58	116.39
4	4608	3188.93	16.95	125.68	4775.39	18.47	139.06

which shows a similar effect as for $d_i = 1$ before, but the increase of d_i slightly decreases the condition numbers in all cases.

Next, for significantly larger problem sizes, we inspect the number of CG iterations needed to reduce the error below a certain tolerance. To derive an adequate stopping and accuracy criterion consider any basis $\Psi = \{\psi_i : i \in \mathcal{I}\}$ of V_h and the stiffness matrix $\mathbf{L}_h^\Psi := (a_h(\psi_i, \psi_j))_{i,j\in\mathcal{I}}$. Given u_h as in (3.2) and the kth CG iterate u_h^k with coefficient vectors $\mathbf{u}_h$ and $\mathbf{u}_h^k$ in Ψ, respectively, we define the residual vector $\mathbf{r}_h^k := \mathbf{f}_h - \mathbf{L}_h^\Psi \mathbf{u}_h^k$, where the ith entry of $\mathbf{f}_h$ is given by $\langle f, \psi_i\rangle$.

Consider the norm $\|\mathbf{v}\|_{\boldsymbol{\mathcal{C}}_h^\Psi} := \langle \mathbf{v}, \boldsymbol{\mathcal{C}}_h^\Psi \mathbf{v}\rangle^{\frac{1}{2}}$ for the coefficient arrays, induced by the preconditioner $\mathcal{C}_h$ with symmetric and positive definite matrix $\boldsymbol{\mathcal{C}}_h^\Psi := (\langle \psi_i, \mathcal{C}_h\psi_j\rangle)_{i,j\in\mathcal{I}}$. Because of (3.9) the eigenvalues of P_h are bounded independently of h. Therefore we can apply (2.7), (2.8), and Lemmas 8.23 and 8.25 from [9] to conclude that

$$\left|\!\left|\!\left| u_h^k - u_h \right|\!\right|\!\right|_h \sim a(u_h^k - u_h, u_h^k - u_h)_h^{\frac{1}{2}} = \left(\left(\mathbf{u}_h^k - \mathbf{u}_h\right)^T \mathbf{L}_h^\Psi \left(\mathbf{u}_h^k - \mathbf{u}_h\right) \right)^{\frac{1}{2}} \sim \|\mathbf{r}_h^k\|_{\boldsymbol{\mathcal{C}}_h^\Psi}.$$

Thus, in the following tests we stop the CG method when the residual $\|\mathbf{r}_h^k\|_{\boldsymbol{\mathcal{C}}_h^\Psi}$, controlling the error $\left|\!\left|\!\left| u_h^k - u_h \right|\!\right|\!\right|_h$, has been reduced by a factor of 10^{-5}.

Table 6.1 displays the corresponding number of iterations for $\gamma = 10$ and $f = 1$ on Ω as the right-hand side and zero as the initial guess for the conjugate gradient method, again for several choices of auxiliary bilinear forms.

We conclude with an example on an adaptive mesh. As trial functions we use Lagrange polynomials of degree $\bar{k} = 3$ corresponding to the principal lattice. Since the test case C0 with maximum d_i showed the best results in the previous examples, we now select this case with $d_i = 10$ and set $\gamma = 15$, which performed best in preliminary tests.

TABLE 6.1

Level	#DOF	Number of iterations for preconditioner (case/d_i)				
		None	C0/1	C0/3	C1/1	C1/3
1	72	26	19	15	20	18
2	288	55	19	17	33	29
3	1152	108	20	19	44	42
4	4608	208	21	20	46	45
5	18432	418	22	21	47	42
6	73728	849	23	22	48	44
7	294912	1718	24	22	48	42
8	1179648	3504	24	22	48	43
9	4718592	7137	24	23	47	41

Contrary to the last example, we now apply the preconditioner in a nested iteration scheme. In each loop we solve the linear system, apply the a posteriori error estimator [13], and capture a ϑ-fraction of the total estimated error by using a standard bulk chasing scheme, where $\vartheta = 0.6$ seems to be a good choice. The CG method stops in the initial loop on the coarsest level when the residual $\|\mathbf{r}_h^k\|_{\mathcal{C}_h^\Psi}$ reaches 10^{-15} and in the following loops when the the residual falls below $\frac{1}{1000}$ of the total estimated error.

Loop	#DOF	#its	Estimated error
0	240	78	1.70E-01
2	360	11	1.13E-01
4	480	10	8.16E-02
6	660	12	6.39E-02
8	990	12	3.66E-02
10	1380	13	2.36E-02
12	2010	12	1.23E-02
14	2760	13	7.42E-03
16	4140	13	4.26E-03
18	6150	12	2.43E-03
20	8670	13	1.28E-03
22	13110	13	7.20E-04
24	20010	12	4.09E-04
26	28230	14	2.28E-04
28	39960	14	1.29E-04
30	59820	13	7.53E-05
32	87330	13	4.43E-05
34	119280	14	2.58E-05
36	166440	14	1.53E-05

The quantitative performance of multiplicative schemes is known to be usually better than that of additive counterparts which is confirmed when comparing the above tests with [10, 14], whose analysis, however, is restricted to quasi-uniform meshes without hanging vertices. Nevertheless, the above comparisons show that the preconditioning effect is dominated by the treatment of the conforming subspace, namely, by the choice of the auxiliary bilinear forms in the above additive context, again supporting the favorable effect of employing a multiplicative version for this component.

7. Concluding remarks. We have reduced the problem of devising multilevel Schwarz preconditioners for the symmetric interior penalty DG formulation to the availability of admissible averaging projectors provided that the underlying meshes satisfy a certain grading property. We have specified one such projector for which uniformly bounded condition numbers are proved under no extra regularity conditions for

any graded triangulations and fixed highest polynomial degree. First numerical experiments illustrate the theoretical findings for a simple model problem and adaptive refinements. Alternative splittings for higher order discretizations will be explored in a forthcoming paper together with more extensive numerical experiments for variable coefficients.

REFERENCES

[1] D. N. Arnold, *An interior penalty finite element method with discontinuous elements*, SIAM J. Numer. Anal., 19 (1982), pp. 742–760.

[2] D. N. Arnold, F. Brezzi, B. Cockburn, and L. D. Marini, *Unified analysis of discontinuous Galerkin methods for elliptic problems*, SIAM J. Numer. Anal., 39 (2002), pp. 1749–1779.

[3] J. H. Bramble, J. E. Pasciak, and J. Xu, *Parallel multilevel preconditioners*, Math. Comp., 55 (1990), pp. 1–22.

[4] S. C. Brenner, *Two-level additive Schwarz preconditioners for nonconforming finite element methods*, Math. Comp., 65 (1996), pp. 897–921.

[5] K. Brix, M. Campos Pinto, W. Dahmen, and R. Massjung, *Multilevel preconditioners for the interior penalty discontinuous Galerkin method* II - *Quantitative Studies*, IGPM report 282, RWTH-Aachen, 2007, Commun. Comput. Phys., to appear.

[6] W. Dahmen, *Adaptive approximation by multivariate smooth splines*, J. Approx. Theory, 36 (1982), pp. 119–140.

[7] W. Dahmen, *Wavelet and multiscale methods for operator equations*, Acta Numer., 6 (1997), pp. 55–228.

[8] W. Dahmen and A. Kunoth, *Multilevel preconditioning*, Numer. Math., 63 (1992), pp. 315–344.

[9] P. Deuflhard and A. Hohmann, *Numerical Analysis in Modern Scientific Computing. An Introduction*, Texts Appl. Math. 43, Springer, Berlin, 2003.

[10] V. A. Dobrev, R. D. Lazarov, P. S. Vassilevski, and L. T. Zikatanov, *Two-level preconditioning of discontinuous Galerkin approximations of second-order elliptic equations*, Numer. Linear Algebra Appl., 13 (2006), pp. 753–770.

[11] J. Gopalakrishnan and G. Kanschat, *A multilevel discontinuous Galerkin method*, Numer. Math., 95 (2003), pp. 527–550.

[12] M. Griebel and P. Oswald, *On the abstract theory of additive and multiplicative Schwarz algorithms*, Numer. Math., 70 (1995), pp. 163–180.

[13] P. Houston, D. Schötzau, and T. P. Wihler, *Energy norm a-posteriori error estimation of hp-adaptive discontinuous Galerkin methods for elliptic problems*, Math. Models Methods Appl. Sci., 17 (2007), pp. 33–62.

[14] G. Kanschat, *Multilevel methods for discontinuous Galerkin FEM on locally refined meshes*, Comput. & Structures, 82 (2004), pp. 2437–2445.

[15] O. Karakashian and F. Pascal, *A posteriori error estimates for a discontinuous Galerkin approximation of second-order elliptic problems*, SIAM J. Numer. Anal., 41 (2003), pp. 2374–2399.

[16] P. Oswald, *Multilevel Finite Element Approximation - Theory and Applications*, Teubner Skripten zur Numerik, B. G. Teubner, Stuttgart, 1994.

[17] P. Oswald, *Preconditioners for nonconforming discretizations*, Math. Comp., 65 (1996), pp. 923–941.

[18] E. Süli, C. Schwab, and P. Houston, *hp-DGFEM for partial differential equations with non-negative characteristic form*, in Discontinuous Galerkin Methods: Theory, Computation and Applications, Lect. Notes Comput. Sci. Eng. 11, B. Cockburn, G. E. Karniadakis, and C.-W. Shu, eds., Springer, Berlin, 2000, pp. 211–230.

[19] T. P. Wihler, P. Frauenfelder, and C. Schwab, *Exponential convergence of the hp-DGFEM for diffusion problems*, Comput. Math. Appl., 46 (2003), pp. 183–205.

[20] J. Xu, *Iterative methods by space decomposition and subspace correction*, SIAM Rev., 34 (1992), pp. 581–613.

[21] J. Xu, *The auxiliary space method and optimal multigrid preconditioning techniques for unstructured grids*, Computing, 56 (1996), pp. 215–235.

[22] H. Yserentant, *On the multi-level splitting of finite element spaces*, Numer. Math., 49 (1986), pp. 379–412.

SIAM J. NUMER. ANAL.
Vol. 46, No. 6, pp. 2769–2794
© 2008 Society for Industrial and Applied Mathematics

A NEW APPROACH TO NUMERICAL SOLUTION OF DEFECTIVE BOUNDARY VALUE PROBLEMS IN INCOMPRESSIBLE FLUID DYNAMICS*

LUCA FORMAGGIA†, ALESSANDRO VENEZIANI‡, AND CHRISTIAN VERGARA†

Abstract. We consider the incompressible Navier–Stokes equations where on a part of the boundary flow rate and mean pressure boundary conditions are prescribed. There are basically two strategies for solving these defective boundary problems: the *variational approach* (see J. Heywood, R. Rannacher, and S. Turek, *Internat. J. Numer. Methods Fluids*, 22 (1996), pp. 325–352) and the *augmented formulation* (see L. Formaggia, J. F. Gerbeau, F. Nobile, and A. Quarteroni, *SIAM J. Numer. Anal.*, 40 (2002), pp. 376–401, and A. Veneziani and C. Vergara, *Internat. J. Numer. Methods Fluids*, 47 (2005), pp. 803–816). However, both of these approaches present some drawbacks. For the flow rate problem, the former resorts to nonstandard functional spaces, which are quite difficult to discretize. On the other hand, for the mean pressure problem, it yields exact solutions only in very special cases. The latter is applicable only to the flow rate problem, since for the mean pressure problem it provides unfeasible boundary conditions. In this paper, we propose a new strategy based on a reformulation of the problems at hand in terms of the minimization of an appropriate functional. This approach allows us to treat the two kinds of problems (flow rate and mean pressure) successfully within the same framework, which can be useful in view of a mixed problem where the two conditions are simultaneously prescribed on different artificial boundaries. Moreover, it is more versatile, being prone to be extended to other kinds of defective conditions. We analyze the problems obtained with this approach and propose some algorithms for their numerical solution. Several numerical results are presented supporting the effectiveness of our approach.

Key words. Navier–Stokes equations, flow rate boundary condition, mean pressure boundary condition, computational hemodynamics

AMS subject classification. 65N99

DOI. 10.1137/060672005

1. Introduction.

In computational fluid dynamics the available boundary data are sometimes incomplete, in particular with reference to the *artificial boundaries*, i.e., the boundaries created just to limit the computational domain and not corresponding to a physical interface. On these boundaries, measurements or results coming from other computations (such as in the *geometrical multiscale approach*—see [5, 13]) typically provide mean or, more in general, incomplete data. This kind of problem has been addressed in [9] and in [4, 15, 16, 17]. In particular, by denoting with Γ an artificial portion of the boundary, two conditions have been considered on Γ, namely, *the mean pressure problem*

$$\frac{1}{|\Gamma|}\int_\Gamma p\, d\gamma = P(t), \tag{1.1}$$

*Received by the editors October 10, 2006; accepted for publication (in revised form) February 26, 2008; published electronically July 3, 2008. This work was supported by the EU Commision through the *Haemodel project* HPRN-CT-2002-00270, by the INDAM grant "Integration of Complex Systems in Biomedicine," and by the Italian Ministry for Research (MURST) through the COFIN05 program.
http://www.siam.org/journals/sinum/46-6/67200.html

†MOX, Department of Mathematics, Politecnico di Milano, Piazza L. da Vinci 32, 20133 Milano, Italy (luca.formaggia@polimi.it, christian.vergara@polimi.it).

‡MOX, Department of Mathematics, Politecnico di Milano, Milano, Italy (alessandro.veneziani@polimi.it), and Department of Mathematics and Computer Science, Emory University, Atlanta, GA, 30322 (ale@mathcs.emory.edu).

where $p = p(t, \boldsymbol{x})$ is the fluid pressure and $P = P(t)$ is a given function of time, and the *flow rate problem*, where we have

$$\rho \int_\Gamma \boldsymbol{u} \cdot \boldsymbol{n} \, d\gamma = Q(t), \tag{1.2}$$

where $\boldsymbol{u} = \boldsymbol{u}(t, \boldsymbol{x})$ is the fluid velocity, $\boldsymbol{n}$ the outward unit vector, and ρ the constant fluid density and $Q = Q(t)$ is a given function.

In [9] a variational approach is proposed for finding an appropriate numerical solution of these problems. A well posedness analysis of the resulting boundary value problems is given, too. In [12] the same conditions on fluxes and pressures are considered as *asymptotic conditions* on infinite domains. In this approach, the artificial boundaries are replaced by far field conditions to be prescribed in an infinite domain. The analysis of this problem is by far not trivial; however, in [12] it is proved that flux and mean pressure conditions regarded as asymptotic conditions under appropriate assumptions yield well-posed problems. In practical problems, however, the computational domain needs to be truncated, so an appropriate method for prescribing such conditions in finite domains is needed.

The variational approach advocated in [9] for the mean pressure problem gives reliable results only in the case of special geometries. On the other hand, in the case of flow rate problems, the proposed variational formulation of [9] resorts to nonstandard functional spaces. The construction of finite-dimensional subspaces is actually quite problematic. For this reason, a different approach is proposed in [4], based on an augmented formulation, where condition (1.2) is regarded as a constraint for the solution to be forced by means of a Lagrange multiplier (see also [15]). Unfortunately, the technique cannot treat the mean pressure problem, as it yields inaccurate solutions for the velocity (see [4, 17]).

In this paper, we propose a new general approach for the defective boundary data problems, which does not suffer from the limitations of the previous ones, so that both mean pressure and flow rate problems can be solved successfully within the same framework. This approach relies on the introduction of an appropriate functional which quantifies the difference between the boundary solution and the prescribed data (see, in a different context, [1]). By resorting to a technique similar to that of a control problem, we introduce control variables related to the fluid velocity and pressure on the boundary. By acting on these control variables, we seek the minimum of the functional at hand. This approach is general and can be applied in principle to all kinds of defective boundary problems. The idea of using boundary data as control variables to force some properties of the solution has been already exploited: We quote, for instance, [7, 10, 2] and more recently [3]. In [6] it has been used for an effective splitting of velocity-pressure computation in standard unsteady Stokes problems. Boundary control variables were used here to force the incompressibility constraint. In general, in these papers the basic idea is to use pointwise velocity data on the boundary for driving the solution to some desired behavior. Here we still use boundary data (not necessarily the velocity) for forcing the "optimal" fulfillment of the incomplete boundary conditions. We consider both the flow rate and the mean pressure problem. We discuss the effectiveness of this approach in comparison with the variational approach proposed in [9] and with the augmented formulation addressed in [4, 15]. Numerical results presented here confirm the flexibility of our method.

The outline of the paper is as follows. Section 2 is devoted to the flow rate problem. The approach based on using normal stresses as control variables is first presented for the generalized Stokes problem and then extended to the nonlinear case.

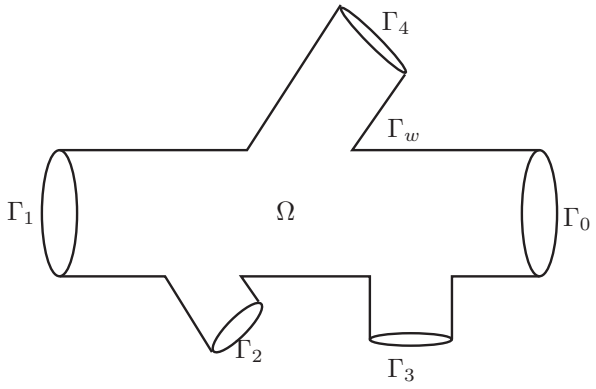


FIG. 1.1. *Reference domain* Ω.

For both cases a proof of well posedness is given. In section 3 we illustrate the approach for the mean pressure problem. We propose different formulations, featuring different sets of control variables. Well posedness analysis is in this case nontrivial and will be carried out in future work, yet the effectiveness of the methods is demonstrated here by numerical tests. Section 4 is devoted to the presentation of some algorithms for the solution of the proposed problems and to their convergence analysis. Finally, in section 5 we present and discuss some numerical results.

1.1. Basic notation. Let us denote by $\Omega \in \mathbb{R}^d$ a domain filled by an incompressible fluid, whose boundary can be split into two parts. The former is denoted by Γ_w and corresponds to a physical wall. The latter is given by the union of the artificial sections Γ_j $(j = 0, \ldots, m)$ which limit the domain of interest but do not have a direct physical significance. A possible domain of this sort is given in Figure 1.1. We assume that the walls are rigid, so that the velocity field is zero on Γ_w, and that the fluid is Newtonian. This domain is representative of incompressible flow problems in a network of pipes, such as in computational hemodynamics, which motivated the present study. The extension to the case of compliant vessels, which is of paramount interest for blood flow problems, requires a specific analysis still to be carried out. The Navier–Stokes equations for the problem at hand are

$$\begin{cases} \rho \dfrac{\partial \boldsymbol{u}}{\partial t} - \mu \triangle \boldsymbol{u} + \rho(\boldsymbol{u} \cdot \nabla)\boldsymbol{u} + \nabla p = \boldsymbol{f} & \text{in } (0,T) \times \Omega, \\ \nabla \cdot \boldsymbol{u} = 0 & \text{in } (0,T) \times \Omega, \end{cases} \tag{1.3}$$

where $\boldsymbol{f}(t, \boldsymbol{x})$ is a given forcing term and μ is the constant fluid viscosity. In what follows, for the sake of simplicity and without loss of generality we will set $\rho = 1$. Moreover, we have *no-slip* boundary conditions on Γ_w, i.e.,

$$\boldsymbol{u}|_{\Gamma_w} = \boldsymbol{0} \qquad \text{in } (0,T), \tag{1.4}$$

as well as the initial condition

$$\boldsymbol{u}|_{t=0} = \boldsymbol{u}_0 \qquad \text{in } \Omega, \tag{1.5}$$

where $\boldsymbol{u}_0 = \boldsymbol{u}_0(\boldsymbol{x})$ is a given function, regular enough. Boundary conditions still have to specified on Γ_j. On those boundaries, we will consider the two possibilities addressed in the introduction.

In what follows we will refer to the functional spaces

$$L^2(\Omega) = \Big\{ v : \Omega \to \mathbb{R} : \int_\Omega v^2 \, d\omega < \infty \Big\}, \qquad H^1(\Omega) = \{ v \in L^2(\Omega) : \nabla v \in L^2(\Omega) \},$$
$$\boldsymbol{V} = \{ \boldsymbol{v} \in \boldsymbol{H}^1(\Omega) : \boldsymbol{v}|_{\Gamma_w} = \boldsymbol{0} \},$$

where $\boldsymbol{H}^1(\Omega) = (H^1(\Omega))^d$, and use the following notation for scalar functions s, $q \in H^1(\Omega)$ and for vector functions $\boldsymbol{v}, \boldsymbol{w} \in \boldsymbol{H}^1(\Omega)$:

$$(s,q) = \int_\Omega s\, q \, d\omega, \quad (\boldsymbol{v}, \boldsymbol{w}) = \int_\Omega \boldsymbol{v} \cdot \boldsymbol{w} \, d\omega = \sum_{i=1}^d \int_\Omega v_i w_i \, d\omega,$$
$$(s,q)_{H^1} = (s,q) + (\nabla s, \nabla q),$$

$$(\nabla \boldsymbol{v}, \nabla \boldsymbol{w}) = \int_\Omega \nabla \boldsymbol{v} : \nabla \boldsymbol{w} \, d\omega = \sum_{i,j=1}^d \int_\Omega \frac{\partial v_i}{\partial x_j} \frac{\partial w_i}{\partial x_j} \, d\omega, \quad (\boldsymbol{v}, \boldsymbol{w})_{H^1} = (\boldsymbol{v}, \boldsymbol{w}) + (\nabla \boldsymbol{v}, \nabla \boldsymbol{w}),$$

$$a(\boldsymbol{v}, \boldsymbol{w}) = \alpha(\boldsymbol{v}, \boldsymbol{w}) + \mu(\nabla \boldsymbol{v}, \nabla \boldsymbol{w}), \qquad c(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}) = ((\boldsymbol{u} \cdot \nabla)\boldsymbol{v}, \boldsymbol{w}),$$

$\alpha \geq 0$ being a given parameter. We recall that $a(\cdot, \cdot)$ is coercive on the space $\boldsymbol{V}$ $\forall \alpha \geq 0$. Moreover, given a vector $\boldsymbol{q} \in \mathbb{R}^m$, we set

$$\|\boldsymbol{q}\|_p = \left(\sum_{j=1}^m |q_j|^p \right)^{1/p}.$$

Finally, given m Hilbert spaces $W_1, \ldots, W_m$, let $\boldsymbol{W} \equiv W_1 \times W_2 \times \cdots \times W_m$, $N : \boldsymbol{W} \to \mathbb{R}$, such that $(y_1, \ldots, y_m) \in \boldsymbol{W} \to N(y_1, \ldots, y_m) \in \mathbb{R}$ and $\langle \cdot, \cdot \rangle$ be the duality pairing between W' and $\boldsymbol{W}$. We denote

$$\langle dN_{y_j}[w_1, \ldots, w_m], z \rangle = \lim_{\varepsilon \to 0} \left(\frac{N(y_1, \ldots, y_j + \varepsilon z, \ldots, y_m) - N(y_1, \ldots, y_j, \ldots, y_m)}{\varepsilon} \right) \Bigg|_{\boldsymbol{y} = \boldsymbol{w}}$$

the Gateaux differential of N, with the respect of y_j computed in the point $\boldsymbol{w} = (w_1, \ldots, w_m) \in \boldsymbol{W}$ along the direction $z \in W_j$.

2. Flow rate boundary conditions. We start by considering the following boundary conditions applied to the artificial boundaries Γ_i:

$$\text{(2.1)} \qquad \begin{cases} (p\boldsymbol{n} - \mu \nabla \boldsymbol{u}\, \boldsymbol{n})|_{\Gamma_0} = \boldsymbol{0} & \text{in } (0,T), \\ \displaystyle\int_{\Gamma_i} \boldsymbol{u} \cdot \boldsymbol{n} \, d\gamma = Q_i, \qquad i = 1, 2, \ldots, m, & \text{in } (0,T), \end{cases}$$

where $Q_i = Q_i(t)$ are given functions of time. Since Γ_w is a rigid boundary, the incompressibility constraint implies that the flow rate on Γ_0 is $Q_0 = -\sum_{i=1}^m Q_i$. Conditions $(2.1)_2$ are not enough to make the problem at hand well-posed in the classical setting, since d "pointwise" boundary data should be prescribed. The approach proposed in [9] to overcome this lack of data is based on the introduction of a variational formulation in the "null fluxes" subspace of $\boldsymbol{V}$ and by lifting the solution by means of an appropriate set of functions called *flux carriers*. It is possible to prove (see, e.g., [14]) that

the boundary conditions implicitly prescribed on Γ_i by this formulation are constant-in-space Neumann boundary conditions. From the numerical viewpoint this has the drawback of dealing with a nonstandard functional space, whose finite-dimensional discretization is not easy to construct. For this reason, a different approach, based on Lagrange multipliers and on the reformulation of the problem in terms of a constraint minimization, has been investigated and numerically tested in [4] for the steady Stokes problem and then extended to the unsteady Navier–Stokes problem in [15, 16, 17]. The idea is to regard the flow rate boundary condition as a constraint to be forced by means of Lagrange multipliers. The *augmented variational formulation* associated with this approach leads to the following problem.

PROBLEM 1. *Given* $\boldsymbol{u}_0 \in \boldsymbol{V}$, *with* $\nabla \cdot \boldsymbol{u}_0 = 0$, $\boldsymbol{f} \in L^2(0,T;\boldsymbol{L}^2(\Omega))$, *and a vector of functions* $\boldsymbol{Q} \in (C^0([0,T]))^m$, *satisfying the conditions* $\int_{\Gamma_j} \boldsymbol{u}_0 \cdot \boldsymbol{n}\, d\gamma = Q_j(0)$ *for* $j = 1, \dots, m$, *find* $\boldsymbol{u} \in L^2(0,T;\boldsymbol{V}) \cap L^\infty(0,T;\boldsymbol{L}^2(\Omega))$, $p \in L^2(0,T;L^2(\Omega))$, *and* $\boldsymbol{\lambda} \in (L^2(0,T))^m$ *such that at any* $t \in (0,T)$

$$
\begin{cases}
\left(\dfrac{\partial \boldsymbol{u}}{\partial t}, \boldsymbol{v}\right) + \mu(\nabla \boldsymbol{u}, \nabla \boldsymbol{v}) + ((\boldsymbol{u}\cdot\nabla)\boldsymbol{u}, \boldsymbol{v}) - (p, \nabla\cdot\boldsymbol{v}) \\
\qquad\qquad + \displaystyle\sum_{j=1}^{m} \lambda_j \int_{\Gamma_j} \boldsymbol{v}\cdot\boldsymbol{n}\, d\gamma = (\boldsymbol{f}, \boldsymbol{v}) & \text{in } \Omega, \\
(q, \nabla\cdot\boldsymbol{u}) = 0 & \text{in } \Omega, \\
\displaystyle\int_{\Gamma_i} \boldsymbol{u}\cdot\boldsymbol{n}\, d\gamma = Q_i, & i = 1, \dots, m,
\end{cases}
$$

for all $\boldsymbol{v} \in \boldsymbol{V}$ *and* $q \in L^2(\Omega)$ *and with* $\boldsymbol{u} = \boldsymbol{u}_0$ *for* $t = 0$.

It has been proven in [4, 15] that this approach leads to the same natural boundary conditions on the artificial boundaries as the variational approach presented in [9]. However, in this case we deal with the standard functional space $\boldsymbol{V}$.

In the next section we show a different variational formulation, still prompt to numerical discretization. For the sake of simplicity we start considering the generalized Stokes problem. Then, in section 2, we extend the approach to the nonlinear case.

2.1. The generalized Stokes problem. In order to address conditions $(2.1)_2$, let us consider the following generalized Stokes problem with complete (i.e., nondefective) boundary conditions:

$$
\begin{cases}
\alpha\boldsymbol{u}(\boldsymbol{k}) - \mu\triangle\boldsymbol{u}(\boldsymbol{k}) + \nabla p(\boldsymbol{k}) = \boldsymbol{f} & \text{in } \Omega, \\
\nabla\cdot\boldsymbol{u}(\boldsymbol{k}) = 0 & \text{in } \Omega, \\
\boldsymbol{u}(\boldsymbol{k})|_{\Gamma_w} = \boldsymbol{0}, & \\
(-p(\boldsymbol{k})\boldsymbol{n} + \mu\nabla\boldsymbol{u}(\boldsymbol{k})\,\boldsymbol{n})|_{\Gamma_0} = \boldsymbol{0}, & \\
(-p(\boldsymbol{k})\boldsymbol{n} + \mu\nabla\boldsymbol{u}(\boldsymbol{k})\,\boldsymbol{n})|_{\Gamma_i} = -k_i\boldsymbol{n}, & i = 1, \dots, m,
\end{cases}
\tag{2.2}
$$

where $\boldsymbol{k} = (k_1, \dots, k_m) \in \mathbb{R}^m$ denotes the vector of the constant values of the normal stresses, $\boldsymbol{f} \in \boldsymbol{L}^2(\Omega)$ is given, and $\alpha \geq 0$ is a given parameter. For the solution of unsteady problems, α is related to the time step and to the time-advancing scheme (see section 2.1.1). As our notation suggests, we regard velocity and pressure fields as a function of the vector $\boldsymbol{k}$. More precisely, for any given $\boldsymbol{k} \in \mathbb{R}^m$, $\boldsymbol{u}(\boldsymbol{k})$ and $p(\boldsymbol{k})$ denote the velocity and the pressure field, respectively, obtained by solving (2.2). In this context, we consider $\boldsymbol{k}$ as a *control variable* to be set such that $\boldsymbol{u} = \boldsymbol{u}(\boldsymbol{k})$ fulfills

the constraint $(2.1)_2$ in some sense to be specified. To this aim we introduce the functional $J_Q : \boldsymbol{V} \to \mathbb{R}^+$:

$$J_Q(\boldsymbol{w}) = \frac{1}{2}\sum_{i=1}^{m}\Big(\int_{\Gamma_i} \boldsymbol{w}\cdot\boldsymbol{n}\, d\gamma - Q_i\Big)^2, \tag{2.3}$$

which is clearly minimal and equal to zero when $(2.1)_2$ is fulfilled. For each $\boldsymbol{w} \in \boldsymbol{V}$, $s \in L^2(\Omega)$, and $\boldsymbol{\eta} \in \mathbb{R}^m$, we build the following Lagrange functional, where (2.2) plays the role of *constraints* for the solution:

$$\begin{aligned}\mathcal{L}(\boldsymbol{w}, s; \boldsymbol{\lambda}_w, \lambda_s; \boldsymbol{\eta}) &= J_Q(\boldsymbol{w}) + a(\boldsymbol{w}, \boldsymbol{\lambda}_w) - (s, \nabla\cdot\boldsymbol{\lambda}_w)\\ &\quad + \sum_{i=1}^{m}\int_{\Gamma_i} \eta_i\boldsymbol{\lambda}_w\cdot\boldsymbol{n}\, d\gamma - (\boldsymbol{f}, \boldsymbol{\lambda}_w) - (\lambda_s, \nabla\cdot\boldsymbol{w}).\end{aligned} \tag{2.4}$$

Here $\boldsymbol{\lambda}_w \in \boldsymbol{V}$ and $\lambda_s \in L^2(\Omega)$ are the *adjoint* variables associated to $\boldsymbol{w}$ and s, respectively. In order to find the corresponding Euler equations, we force the Gateaux differentials of $\mathcal{L}$ evaluated for any test function to vanish in correspondence to the solution $\boldsymbol{\chi} = [\boldsymbol{u}, p; \boldsymbol{\lambda}_u, \lambda_p; \boldsymbol{k}]$. That is, we will consider the following problem, where for the sake of notation we omit to specify that the differentials are computed in $\boldsymbol{\chi}$.

PROBLEM 2. *Given $\boldsymbol{f} \in \boldsymbol{L}^2(\Omega)$ and $\boldsymbol{Q} \in \mathbb{R}^m$, find $\boldsymbol{k} \in \mathbb{R}^m$, $\boldsymbol{u}(\boldsymbol{k}) \in \boldsymbol{V}, p(\boldsymbol{k}) \in L^2(\Omega), \boldsymbol{\lambda}_u \in \boldsymbol{V}$, and $\lambda_p \in L^2(\Omega)$ such that, for all $\boldsymbol{v} \in \boldsymbol{V}$, $q \in L^2(\Omega)$, and $\nu \in \mathbb{R}$:*

$$\begin{cases}
\text{(P)} \begin{cases} \langle d\mathcal{L}_{\boldsymbol{\lambda}_w}, \boldsymbol{v}\rangle = a(\boldsymbol{u}, \boldsymbol{v}) - (p, \nabla\cdot\boldsymbol{v}) + \displaystyle\sum_{i=1}^{m}\int_{\Gamma_i} k_i\boldsymbol{v}\cdot\boldsymbol{n}\, d\gamma - (\boldsymbol{f}, \boldsymbol{v}) = 0, \\ \langle d\mathcal{L}_{\lambda_s}, q\rangle = -(q, \nabla\cdot\boldsymbol{u}) = 0, \end{cases} \\
\text{(A)} \begin{cases} \langle d\mathcal{L}_{\boldsymbol{w}}, \boldsymbol{v}\rangle = a(\boldsymbol{v}, \boldsymbol{\lambda}_u) + \displaystyle\sum_{i=1}^{m}\left(\int_{\Gamma_i} \boldsymbol{u}\cdot\boldsymbol{n}\, d\gamma - Q_i\right)\int_{\Gamma_i} \boldsymbol{v}\cdot\boldsymbol{n}\, d\gamma - (\lambda_p, \nabla\cdot\boldsymbol{v}) = 0, \\ \langle d\mathcal{L}_s, q\rangle = -(q, \nabla\cdot\boldsymbol{\lambda}_u) = 0, \end{cases} \\
\text{(C}_\text{j}\text{)} \quad \langle d\mathcal{L}_{\eta_j}, \nu\rangle = \nu\displaystyle\int_{\Gamma_j} \boldsymbol{\lambda}_u\cdot\boldsymbol{n}\, d\gamma = 0, \quad j = 1, \dots, m.
\end{cases} \tag{2.5}$$

This system couples a generalized Stokes problem (P) with its adjoint (A) and m scalar equations (*optimality conditions*), denoted by (C_j). Observe that the latter force the adjoint variable $\boldsymbol{\lambda}_u$ to have null flux on the artificial boundaries.

By exploiting the symmetry of the bilinear form $a(\cdot,\cdot)$, the strong formulation of the adjoint problem (A) can be readily deduced, yielding

$$\begin{cases}
\alpha\boldsymbol{\lambda}_u - \mu\triangle\boldsymbol{\lambda}_u + \nabla\lambda_p = \boldsymbol{0} & \text{in } \Omega, \\
\nabla\cdot\boldsymbol{\lambda}_u = 0 & \text{in } \Omega, \\
\boldsymbol{\lambda}_u|_{\Gamma_w} = \boldsymbol{0}, & \\
(-\lambda_p\boldsymbol{n} + \mu\nabla\boldsymbol{\lambda}_u\ \boldsymbol{n})|_{\Gamma_0} = \boldsymbol{0}, & \\
(-\lambda_p\boldsymbol{n} + \mu\nabla\boldsymbol{\lambda}_u\ \boldsymbol{n})|_{\Gamma_i} = -\left(\displaystyle\int_{\Gamma_i} \boldsymbol{u}\cdot\boldsymbol{n}\, d\gamma - Q_i\right)\boldsymbol{n}, & i = 1, \dots, m.
\end{cases} \tag{2.6}$$

Numerical solution of Problem 2 can be pursued by means of iterative techniques, as we will see in section 4. It is worth noting that, if the iterative process converges, at the limit (i.e., when $J_Q(\boldsymbol{u}) = 0$), the fulfillment of (A) and (C_j) implies that $\boldsymbol{\lambda}_u = \boldsymbol{0}$ and $\lambda_p = 0$. This is promptly verified by bearing in mind conditions (C_j) and by

selecting $\boldsymbol{v} = \boldsymbol{\lambda}_u$ in (A). The adjoint variables are, however, needed to drive the iterative scheme to the optimal solution, as will be illustrated in section 4.

A general theory of well posedness of problems arising in flow control is presented in [7, Chap. 6]. Here, we prove a well posedness results strongly based on the specific feature of the problems at hand.

PROPOSITION 2.1. *Problem* 2 *admits a unique solution* $[\boldsymbol{u}(\boldsymbol{k}), p(\boldsymbol{k}); \boldsymbol{\lambda}_u, \lambda_p; \boldsymbol{k}]$.

Proof. We introduce the following operators:

1. $\mathcal{P}_{\boldsymbol{f}} : \mathbb{R}^m \to \boldsymbol{V}$: Given a vector $\boldsymbol{h} \in \mathbb{R}^m$, it associates it with the function $\boldsymbol{w}$, where $[\boldsymbol{w}, s]$ is the solution of

$$\begin{cases} a(\boldsymbol{w}, \boldsymbol{v}) - (s, \nabla \cdot \boldsymbol{v}) = -\displaystyle\sum_{j=1}^{m} h_j \int_{\Gamma_j} \boldsymbol{v} \cdot \boldsymbol{n}\, d\gamma + (\boldsymbol{f}, \boldsymbol{v}) & \forall \boldsymbol{v} \in \boldsymbol{V}, \\ (q, \nabla \cdot \boldsymbol{w}) = 0 & \forall q \in L^2(\Omega). \end{cases} \tag{2.7}$$

   In particular, if $\boldsymbol{f} = \boldsymbol{0}$, we will write $\mathcal{P}_{\boldsymbol{0}}$. We point out that operator $\mathcal{P}_{\boldsymbol{f}}$ is well defined since, for a given $\boldsymbol{h}$, (2.7) admits an unique solution, being a generalized Stokes system with a constant mixed Neumann and Dirichlet boundary condition (see, e.g., [9]). By exploiting the linearity of the problem we have $\boldsymbol{w} = \mathcal{P}_{\boldsymbol{f}}(\boldsymbol{h}) = \boldsymbol{w}_1 + \boldsymbol{w}_2$, where

$$\boldsymbol{w}_1 = \mathcal{P}_{\boldsymbol{f}}(\boldsymbol{0}), \quad \boldsymbol{w}_2 = \mathcal{P}_{\boldsymbol{0}}(\boldsymbol{h}).$$

2. $\mathcal{A} : \boldsymbol{V} \to \mathbb{R}^m$ such that

$$\mathcal{A}\boldsymbol{w} = \left(\int_{\Gamma_1} \boldsymbol{w} \cdot \boldsymbol{n}\, d\gamma, \int_{\Gamma_2} \boldsymbol{w} \cdot \boldsymbol{n}\, d\gamma, \ldots, \int_{\Gamma_m} \boldsymbol{w} \cdot \boldsymbol{n}\, d\gamma \right). \tag{2.8}$$

3. We define $\mathcal{B}_{\boldsymbol{f}} = \mathcal{A} \circ \mathcal{P}_{\boldsymbol{f}}$ and $\mathcal{B}_{\boldsymbol{0}} = \mathcal{A} \circ \mathcal{P}_{\boldsymbol{0}}$, respectively. Operator $\mathcal{B}_{\boldsymbol{f}}$ is an affine operator from $\mathbb{R}^m$ to $\mathbb{R}^m$, while $\mathcal{B}_{\boldsymbol{0}}$ is linear. Observe that

$$\mathcal{B}_{\boldsymbol{f}}(\boldsymbol{h}) = \mathcal{A}\,(\boldsymbol{w}_1 + \boldsymbol{w}_2) = \mathcal{B}_{\boldsymbol{0}}\boldsymbol{h} + \mathcal{B}_{\boldsymbol{f}}(\boldsymbol{0}). \tag{2.9}$$

We point out that Problem 2 is equivalent to solving the following linear system for the unknown $\boldsymbol{k}$:

$$\mathcal{B}_{\boldsymbol{0}}[\mathcal{B}_{\boldsymbol{f}}(\boldsymbol{k}) - \boldsymbol{Q}] = \boldsymbol{0}. \tag{2.10}$$

Let us prove that $\mathcal{B}_{\boldsymbol{0}}$ is an invertible operator by providing its explicit representation. This is a linear operator from $\mathbb{R}^m$ to $\mathbb{R}^m$, so it can be represented by a matrix. Let us denote by $\boldsymbol{w}_j$ the solution of (2.7) for $\boldsymbol{f} = \boldsymbol{0}$ and $\boldsymbol{h} = \boldsymbol{e}_j$, $\boldsymbol{e}_j$ being the jth unit vector. For a generic vector $\boldsymbol{h} = \sum_{j=1}^m h_j \boldsymbol{e}_j$, we have $\mathcal{P}_{\boldsymbol{0}}(\boldsymbol{h}) = \sum_{j=1}^m h_j \boldsymbol{w}_j$.

Let us now exploit $(2.7)_2$ in $(2.7)_1$ by taking $\boldsymbol{v} = \boldsymbol{w}_j$ and $\boldsymbol{h} = \boldsymbol{e}_i$. We obtain

$$a(\boldsymbol{w}_i, \boldsymbol{w}_j) = -\int_{\Gamma_i} \boldsymbol{w}_j \cdot \boldsymbol{n} d\gamma.$$

This implies that

$$[\mathcal{B}_{\boldsymbol{0}}\boldsymbol{h}]_i = -\sum_{j=1}^{m} a(\boldsymbol{w}_i, \boldsymbol{w}_j) h_j,$$

so we finally have

$$[\mathcal{B}_{\boldsymbol{0}}]_{ij} = -a(\boldsymbol{w}_i, \boldsymbol{w}_j).$$

Since $a(\cdot,\cdot)$ is coercive, symmetric bilinear form, matrix $\mathcal{B}_{\mathbf{0}}$ is symmetric and negative definite. Therefore, it follows that (2.10) reduces to

$$\mathcal{B}_{\boldsymbol{f}}(\boldsymbol{k}) = \boldsymbol{Q}.$$

Thanks to the linearity, this in turn can be rewritten as

$$\mathcal{B}_{\mathbf{0}}\boldsymbol{k} = \boldsymbol{Q} - \mathcal{B}_{\boldsymbol{f}}(\mathbf{0}).$$

Again, thanks to the nonsingularity of $\mathcal{B}_{\mathbf{0}}$, we conclude that solution $\boldsymbol{k}$ exists and is unique. The corresponding $\boldsymbol{u}$ and p are obtained by solving (2.7) with $\boldsymbol{h} = \boldsymbol{k}$ and satisfy the constraints by construction. The associated adjoint variables $\boldsymbol{\lambda}_u$ and λ_p are zero. ∎

2.1.1. The unsteady case. When one wants to solve the unsteady Stokes problem

$$\text{(2.11)}\qquad \begin{cases} \dfrac{\partial \boldsymbol{u}(\boldsymbol{k})}{\partial t} - \mu\triangle\boldsymbol{u}(\boldsymbol{k}) + \nabla p(\boldsymbol{k}) = \boldsymbol{f} & \text{in } (0,T)\times\Omega, \\ \nabla\cdot\boldsymbol{u}(\boldsymbol{k}) = 0 & \text{in } (0,T)\times\Omega, \\ \boldsymbol{u}(\boldsymbol{k})|_{t=0} = \boldsymbol{u}_0 & \text{in } \Omega, \\ \boldsymbol{u}(\boldsymbol{k})|_{\Gamma_w} = \mathbf{0} & \text{in } (0,T), \\ (-p(\boldsymbol{k})\boldsymbol{n} + \mu\nabla\boldsymbol{u}(\boldsymbol{k})\,\boldsymbol{n})|_{\Gamma_0} = \mathbf{0} & \text{in } (0,T), \\ (-p(\boldsymbol{k})\boldsymbol{n} + \mu\nabla\boldsymbol{u}(\boldsymbol{k})\,\boldsymbol{n})|_{\Gamma_i} = -k_i(t)\boldsymbol{n}, \quad i=1,\dots,m, & \text{in } (0,T), \end{cases}$$

with $\boldsymbol{f} \in L^\infty(0,T;\boldsymbol{L}^2(\Omega))$, $\boldsymbol{u}_0 \in \boldsymbol{L}^2(\Omega)$, $\nabla\cdot\boldsymbol{u}_0 = 0$, and $k_i(t)$, $i = 1,\dots,m$, chosen such that (2.4) is minimal at each time step, a possible approach would rely on a Lagrangian functional for the time-dependent problem. However, this implies a time-dependent adjoint equation with a final time condition. The solution of the associated control problem in the time interval $(0,T)$ would lead to iteratively solving (and storing) the solution of the primal problem in order to find the time-reversed adjoint solution. A different, *discretize then optimize* approach consists of applying our control technique on the time discretization of (2.11). See, in different contexts, [2, 11]. In this way, we avoid the solution of the time adjoint problem, since we already know its exact solution, which is zero. In fact, here the adjoint problem is just a tool for enforcing the mean flux boundary conditions. In particular, by denoting $t^n = n\Delta t$, with Δt the time discretization step and referring to a backward difference formula (BDF) time-advancing scheme, we obtain, for each n, problem (2.2) and the Lagrangian (2.4) with $\alpha = \beta_0/\Delta t$ and a forcing term equal to $\boldsymbol{f}^n + \sum_{i=1}^r \beta_i/\Delta t\,\boldsymbol{u}^{n-i}$, where $\boldsymbol{k}^n = \boldsymbol{k}(t^n)$, $\boldsymbol{u}^n = \boldsymbol{u}^n(\boldsymbol{k}^n) = \boldsymbol{u}(t^n,\boldsymbol{x})$, $p^n = p^n(\boldsymbol{k}^n) = p(t^n,\boldsymbol{x})$, $\boldsymbol{f}^n = \boldsymbol{f}(t^n,\boldsymbol{x})$, $\boldsymbol{u}^{n-i}$ is the approximation of $\boldsymbol{u}(t^{n-i},\boldsymbol{x})$, and β_i $(i = 0,1,\dots,r \le n)$ are the coefficients of the time discretization. Therefore, with this notation, we can interpret Problem 2 as a technique for the solution of a *discrete-in-time* unsteady flow rate Stokes problem. In this case the strong formulation of the adjoint problem (A) is given by (2.6) with $\alpha = \beta_0/\Delta t$. It may be noted that in this case system (2.6) is exactly the BDF discretization at time $t = t^n$ of the backward-in-time adjoint problem when the exact solution $(\boldsymbol{\lambda}_u^{n+k} = \mathbf{0}, \lambda_p^{n+k} = 0)$ for $k = 1,2,\dots,r$ is used.

2.2. The nonlinear case. We focus now on the system arising when the nonlinear convective term is present in the fluid equations. More precisely, we consider the following Navier–Stokes problem:

$$\text{(2.12)}\qquad \alpha\boldsymbol{u}(\boldsymbol{k}) - \mu\triangle\boldsymbol{u}(\boldsymbol{k}) + (\boldsymbol{u}(\boldsymbol{k})\cdot\nabla)\boldsymbol{u}(\boldsymbol{k}) + \nabla p(\boldsymbol{k}) = \boldsymbol{f} \quad \text{in } \Omega$$

together with $(2.2)_{2-5}$. To minimize (2.3) with the constraint given by (2.12) together with $(2.2)_{2-5}$, we have to find the stationary point of the following Lagrangian functional:

$$\mathcal{L}(\boldsymbol{w}, s; \boldsymbol{\lambda}_w, \lambda_s; \boldsymbol{\eta}) = J_Q(\boldsymbol{w}) + a(\boldsymbol{w}, \boldsymbol{\lambda}_w) + c(\boldsymbol{w}, \boldsymbol{w}, \boldsymbol{\lambda}_w) - (s, \nabla \cdot \boldsymbol{\lambda}_w)$$

$$+\sum_{i=1}^{m} \int_{\Gamma_i} \eta_i \boldsymbol{\lambda}_w \cdot \boldsymbol{n}\, d\gamma - (\boldsymbol{f}, \boldsymbol{\lambda}_w) - (\lambda_p, \nabla \cdot \boldsymbol{w}).$$

This leads to the following.

PROBLEM 3. *Given $\boldsymbol{f} \in \boldsymbol{L}^2(\Omega)$ and $\boldsymbol{Q} \in \mathbb{R}^m$, find $\boldsymbol{k} \in \mathbb{R}^m$, $\boldsymbol{u}(\boldsymbol{k}) \in \boldsymbol{V}, p(\boldsymbol{k}) \in L^2(\Omega), \boldsymbol{\lambda}_u \in \boldsymbol{V}$, and $\lambda_p \in L^2(\Omega)$ such that, for all $\boldsymbol{v} \in \boldsymbol{V}$, $q \in L^2(\Omega)$, and $\nu \in \mathbb{R}$:*

$$\begin{cases} \text{(P)} \begin{cases} \langle d\mathcal{L}_{\boldsymbol{\lambda}_u}, \boldsymbol{v} \rangle = a(\boldsymbol{u}, \boldsymbol{v}) + c(\boldsymbol{u}, \boldsymbol{u}, \boldsymbol{v}) - (p, \nabla \cdot \boldsymbol{v}) + \displaystyle\sum_{i=i}^{m} \int_{\Gamma_i} k_i \boldsymbol{v} \cdot \boldsymbol{n}\, d\gamma - (\boldsymbol{f}, \boldsymbol{v}) = 0, \\ \langle d\mathcal{L}_{\lambda_p}, q \rangle = -(q, \nabla \cdot \boldsymbol{u}) = 0, \end{cases} \\ \text{(A)} \begin{cases} \langle d\mathcal{L}_{\boldsymbol{u}}, \boldsymbol{v} \rangle = a(\boldsymbol{v}, \boldsymbol{\lambda}_u) + c(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{\lambda}_u) + c(\boldsymbol{v}, \boldsymbol{u}, \boldsymbol{\lambda}_u) \\ \qquad\qquad -(\lambda_p, \nabla \cdot \boldsymbol{v}) + \displaystyle\sum_{i=1}^{m} \left(\int_{\Gamma_i} \boldsymbol{u} \cdot \boldsymbol{n}\, d\gamma - Q_i \right) \int_{\Gamma_i} \boldsymbol{v} \cdot \boldsymbol{n}\, d\gamma = 0, \\ \langle d\mathcal{L}_p, q \rangle = -(q, \nabla \cdot \boldsymbol{\lambda}_u) = 0, \end{cases} \\ \text{(C}_\text{j}\text{)} \quad \langle d\mathcal{L}_{k_j}, \nu \rangle = \displaystyle\int_{\Gamma_j} \nu \boldsymbol{\lambda}_u \cdot \boldsymbol{n}\, d\gamma = 0, \quad j = 1, \ldots, m. \end{cases}$$

In order to extend Proposition 2.1 to the nonlinear case, we introduce the following operators:

1. $\mathcal{F}_{\boldsymbol{f}} : \mathbb{R}^m \to \boldsymbol{V}$ associates to a vector $\boldsymbol{h} \in \mathbb{R}^m$ the function $\boldsymbol{w}$, where $[\boldsymbol{w}, s]$ is the solution of the Navier–Stokes problem:

(2.13)
$$\begin{cases} a(\boldsymbol{w}, \boldsymbol{v}) + c(\boldsymbol{w}, \boldsymbol{w}, \boldsymbol{v}) - (s, \nabla \cdot \boldsymbol{v}) = -\displaystyle\sum_{j=1}^{m} h_j \int_{\Gamma_j} \boldsymbol{v} \cdot \boldsymbol{n}\, d\gamma + (\boldsymbol{f}, \boldsymbol{v}) & \forall \boldsymbol{v} \in \boldsymbol{V}, \\ (q, \nabla \cdot \boldsymbol{w}) = 0 & \forall q \in L^2(\Omega). \end{cases}$$

2. Given $\boldsymbol{z} \in \boldsymbol{V}$, $\mathcal{M}[\boldsymbol{z}] : \mathbb{R}^m \to \boldsymbol{V}$ associates to a vector $\boldsymbol{h} \in \mathbb{R}^m$ the function $\boldsymbol{w}$, where $[\boldsymbol{w}, s]$ is the solution of the problem:

(2.14)
$$\begin{cases} a(\boldsymbol{w}, \boldsymbol{v}) + c(\boldsymbol{z}, \boldsymbol{v}, \boldsymbol{w}) + c(\boldsymbol{v}, \boldsymbol{z}, \boldsymbol{w}) - (s, \nabla \cdot \boldsymbol{v}) = -\displaystyle\sum_{j=1}^{m} h_j \int_{\Gamma_j} \boldsymbol{v} \cdot \boldsymbol{n}\, d\gamma & \forall \boldsymbol{v} \in \boldsymbol{V}, \\ (q, \nabla \cdot \boldsymbol{w}) = 0 & \forall q \in L^2(\Omega). \end{cases}$$

Let us denote with $\boldsymbol{Q}^0 = \mathcal{A} \circ \mathcal{F}_{\boldsymbol{f}}(\boldsymbol{0})$ the vector of flow rates (an entry for each artificial boundary) of the solution of the problem (2.13) with $h_j = 0$ for any j. $\mathcal{A}$ has been defined in (2.8).

PROPOSITION 2.2. *If the viscosity μ is large enough, and $\sum_i |Q_i - Q_i^0|$ and $\|\boldsymbol{f}\|$ are small enough, then Problem 3 admits a solution $[\boldsymbol{u}(\boldsymbol{k}), p(\boldsymbol{k}); \boldsymbol{\lambda}_u, \lambda_p; \boldsymbol{k}]$.*

Proof. Assume that the bilinear form $b_{\boldsymbol{z}}(\boldsymbol{w}, \boldsymbol{v}) := a(\boldsymbol{w}, \boldsymbol{v}) + c(\boldsymbol{z}, \boldsymbol{v}, \boldsymbol{w}) + c(\boldsymbol{v}, \boldsymbol{z}, \boldsymbol{w})$ is continuous and coercive for any $\boldsymbol{z} \in \boldsymbol{V}$. This is actually possible if the first term

dominates the other two, namely, if the viscosity μ is large enough. In this case, problem $\boldsymbol{w} = \mathcal{M}[\boldsymbol{z}](\boldsymbol{h})$ admits an unique solution for any $\boldsymbol{z} \in \boldsymbol{V}$. Let us introduce the following operators from $\mathbb{R}^m$ to $\mathbb{R}^m$: $\mathcal{C}[\boldsymbol{z}] = \mathcal{A} \circ \mathcal{M}[\boldsymbol{z}]$ and $\mathcal{D}_{\boldsymbol{f}} = \mathcal{A} \circ \mathcal{F}_{\boldsymbol{f}}$. Problem 3 is then equivalent to solving

$$\mathcal{C}[\mathcal{F}_{\boldsymbol{f}}(\boldsymbol{k})](\mathcal{D}_{\boldsymbol{f}}(\boldsymbol{k}) - \boldsymbol{Q}) = \boldsymbol{0}. \tag{2.15}$$

Now, for any $\boldsymbol{z} \in \boldsymbol{V}$, $\mathrm{Ker}(\mathcal{C}[\boldsymbol{z}]) = \{\boldsymbol{0}\}$. Indeed, $\mathcal{C}[\boldsymbol{z}](\boldsymbol{h})$ corresponds to computing the volume fluxes on Γ_i, $i = 1, \ldots, m$, of the velocity field solution of (2.14). $\mathcal{C}[\boldsymbol{z}]$ being a linear operator, it is in fact represented by matrix whose elements are $[\mathcal{C}[\boldsymbol{z}]]_{ij} = -b_{\boldsymbol{z}}(\boldsymbol{w}_j, \boldsymbol{w}_i)$ and where the velocity fields $\boldsymbol{w}_s$ are defined as in Proposition 2.1. Since $b_z(\cdot, \cdot)$ is coercive, we conclude that $\mathcal{C}[\boldsymbol{z}]$ is invertible. Therefore, system (2.15) is equivalent to the nonlinear problem

$$\mathcal{D}_{\boldsymbol{f}}(\boldsymbol{k}) = \boldsymbol{Q}, \tag{2.16}$$

that of finding suitably the k_j's such that the velocity field solution of system (P) in Problem 3 satisfies the flow rate condition $(2.1)_2$. This problem coincides with the augmented formulation proposed in [4, 15] with the difference that there the k_j were treated as Lagrange multipliers while here they act as control variables. In [15] it is shown that if $\sum_i |Q_i|$ is small enough, then an unique solution $[\boldsymbol{u}, p, \boldsymbol{k}]$ of the augmented problem exists in $[0, T]$, for a suitable $T > 0$, in the case $\boldsymbol{f} = \boldsymbol{0}$. By using similar arguments we can extend this result to the case $\boldsymbol{f} \neq \boldsymbol{0}$ by asking that $\sum_i |Q_i - Q_i^0|$ and $\|\boldsymbol{f}\|$ are small enough. Under these assumptions, problem (2.16) admits (at least) a solution $\boldsymbol{k}$ in $[0, T]$.

The corresponding $\boldsymbol{u}$ and p are obtained by solving system (P) in Problem 3 and satisfy the constraints by construction. The associated adjoint variables $\boldsymbol{\lambda}_u$ and λ_p are zero. ∎

The hypothesis on the smallness of $\sum_i |Q_i - Q_i^0|$ requires in fact that the prescribed fluxes are sufficiently close to the ones featured by the solution of the Navier–Stokes problem fulfilling homogeneous Neumann conditions in correspondence of the artificial boundaries.

3. Mean pressure boundary conditions. We consider now the mean pressure problem. Still referring to Figure 1.1, we consider (1.3) together with the boundary condition (1.4), the initial condition (1.5), and the following defective condition on the artificial sections Γ_j at any $t \in (0, T)$:

$$\frac{1}{|\Gamma_j|} \int_{\Gamma_j} p \, d\gamma = P_j, \quad j = 0, 1, \ldots, m, \tag{3.1}$$

where the $P_j = P_j(t)$ are given functions of time. We will specify later on the regularity of the data that we address. Conditions (3.1) are still not sufficient to make the problem at hand well-posed, too.

The variational approach for this type of problem advocated in [9] is based on the forcing of implicit natural homogeneous boundary conditions leading in this case to constant (in space) normal stresses and to zero tangential stresses on each Γ_i. Thus, this formulation is in fact just an approximation of the mean pressure problem and gives the expected results only for specific cases (see section 3.2). On the other hand, the Lagrange multiplier approach is not suited in this case. Its direct application indeed implies a constant normal velocity on each artificial section, which in general

not compatible with the conditions on Γ_w. Indeed, by referring, for the sake of simplicity, to the linear case, let us write the Lagrangian functional

$$\mathcal{M}(\boldsymbol{w}, s, \eta_0, \dots, \eta_m) = \frac{1}{2} a(\boldsymbol{w}, \boldsymbol{w}) + \sum_{j=0}^{m} \eta_j \int_{\Gamma_j} (s - P_j)\, d\gamma + (\nabla s, \boldsymbol{w}) - (\boldsymbol{f}, \boldsymbol{w}),$$

where $\boldsymbol{\eta} \in \mathbb{R}^{m+1}$ is the vector of the Lagrange multipliers related to the boundary conditions (3.1). By introducing the function

$$\Psi(\varepsilon, \varepsilon, \alpha_0, \dots, \alpha_m) = \mathcal{M}(\boldsymbol{w} + \varepsilon \boldsymbol{v}, s + \varepsilon q, \eta_0 + \alpha_0 \chi_0, \dots, \eta_m + \alpha_m \chi_m),$$

which takes into account variations of the velocity, of the pressure, and of the Lagrange multipliers along the directions $\boldsymbol{v}$, q, and χ_j, respectively, we impose $\frac{d\psi}{d\varepsilon}(0, 0, 0, \dots, 0) = 0$ and $\frac{d\psi}{d\alpha_j}(0, 0, 0, \dots, 0) = 0$, $j = 0, \dots, m$, and obtain the following augmented problem.

For all $\boldsymbol{v} \in \boldsymbol{V}$ and $q \in H^1(\Omega)$, find $\boldsymbol{u} \in V$, $p \in H^1(\Omega)$, and $\boldsymbol{\lambda} \in \mathbb{R}^{m+1}$ such that

$$\begin{cases} a(\boldsymbol{u}, \boldsymbol{v}) + (\nabla p, \boldsymbol{v}) = (\boldsymbol{f}, \boldsymbol{v}), \\ (\nabla q, \boldsymbol{u}) + \displaystyle\sum_{j=0}^{m} \lambda_j \int_{\Gamma_j} q\, d\gamma = 0, \\ \dfrac{1}{|\Gamma_j|} \displaystyle\int_{\Gamma_j} p\, d\gamma = P_j, \quad j = 0, 1, \dots, m. \end{cases}$$

The choice of the space $H^1(\Omega)$ for the pressures is done to guarantee the existence of their trace (see section 3.1). The second equation of the previous system gives

$$-(q, \nabla \cdot \boldsymbol{u}) + \int_{\partial\Omega} q\, \boldsymbol{u} \cdot \boldsymbol{n}\, d\gamma + \sum_{j=0}^{m} \lambda_j \int_{\Gamma_j} q\, d\gamma = 0,$$

that is, $(\boldsymbol{u} \cdot \boldsymbol{n})|_{\Gamma_j} = \lambda_j \, \forall j = 0, \dots, m$, meaning that the normal component of the velocity is constant on each Γ_j. These conditions are not in general compatible with the no-slip boundary condition (1.4) on Γ_w.

Here we propose as a possible alternative an extension of the approach introduced in the previous section. To this aim, it is crucial to select an appropriate control variable. For the moment, the (constant) normal stresses $\boldsymbol{k}$ will be retained as control variables.

3.1. Basic approach. Let us consider the generalized Stokes problem (2.2) where also at Γ_0 we prescribe a condition such as $(2.2)_5$. The control variables $\boldsymbol{k}$ are determined so that the following functional is minimized:

$$J_P(s) = \frac{1}{2} \left(\sum_{i=0}^{m} \frac{1}{|\Gamma_i|} \int_{\Gamma_i} s\, d\gamma - P_i \right)^2. \tag{3.2}$$

We refer to the following Lagrangian functional:

$$\begin{aligned} \mathcal{L}(\boldsymbol{w}, s; \boldsymbol{\lambda}_w, \lambda_s; \boldsymbol{\eta}) = J_P(s) + a(\boldsymbol{w}, \boldsymbol{\lambda}_w) - (s, \nabla \cdot \boldsymbol{\lambda}_w) \\ + \sum_{i=0}^{m} \int_{\Gamma_i} \eta_i \boldsymbol{\lambda}_w \cdot \boldsymbol{n}\, d\gamma - (\boldsymbol{f}, \boldsymbol{\lambda}_w) - (\lambda_s, \nabla \cdot \boldsymbol{w}) \end{aligned} \tag{3.3}$$

for all $\boldsymbol{w} \in \boldsymbol{V}$, $s \in H^1(\Omega)$, $\boldsymbol{\lambda}_w \in \boldsymbol{V}$, $\lambda_s \in H^1(\Omega)$, and $\boldsymbol{\eta} \in \mathbb{R}^{m+1}$. We point out that we take $s \in H^1(\Omega)$ so that its trace on Γ_i, $i = 0, \dots, m$, is meaningful. Of course, we assume that Ω is sufficiently regular. The stationary point of $\mathcal{L}$ fulfills the following.

PROBLEM 4. *Given $\boldsymbol{f} \in \boldsymbol{L}^2(\Omega)$ and $\boldsymbol{P} \in \mathbb{R}^{m+1}$, find $\boldsymbol{k} \in \mathbb{R}^{m+1}$, $\boldsymbol{u}(\boldsymbol{k}) \in \boldsymbol{V}, p(\boldsymbol{k}) \in H^1(\Omega), \boldsymbol{\lambda}_u \in \boldsymbol{V}$, and $\lambda_p \in H^1(\Omega)$ such that, for all $\boldsymbol{v} \in \boldsymbol{V}$, $q \in H^1(\Omega)$, and $\nu \in \mathbb{R}$,*

$$
\begin{cases}
\text{(P)} \begin{cases} \langle d\mathcal{L}_{\boldsymbol{\lambda}_u}, \boldsymbol{v} \rangle = a(\boldsymbol{u}, \boldsymbol{v}) - (p, \nabla \cdot \boldsymbol{v}) + \displaystyle\sum_{i=0}^{m} \int_{\Gamma_i} k_i \boldsymbol{v} \cdot \boldsymbol{n} \, d\gamma - (\boldsymbol{f}, \boldsymbol{v}) = 0, \\ \langle d\mathcal{L}_{\lambda_p}, q \rangle = -(q, \nabla \cdot \boldsymbol{u}) = 0, \end{cases} \\
\text{(A)} \begin{cases} \langle d\mathcal{L}_{\boldsymbol{u}}, \boldsymbol{v} \rangle = a(\boldsymbol{v}, \boldsymbol{\lambda}_u) - (\lambda_p, \nabla \cdot \boldsymbol{v}) = 0, \\ \langle d\mathcal{L}_p, q \rangle = \displaystyle\sum_{i=0}^{m} \left(\frac{1}{|\Gamma_i|} \int_{\Gamma_i} p \, d\gamma - P_i \right) \frac{1}{|\Gamma_i|} \int_{\Gamma_i} q \, d\gamma - (q, \nabla \cdot \boldsymbol{\lambda}_u) = 0, \end{cases} \\
\text{(C}_\text{j}\text{)} \quad \langle d\mathcal{L}_{k_i}, \nu \rangle = \displaystyle\int_{\Gamma_i} \nu \boldsymbol{\lambda}_u \cdot \boldsymbol{n} \, d\gamma = 0, \quad i = 0, \dots, m.
\end{cases}
$$

In this case, we obtain a system coupling a generalized Stokes (P) and a fluid problem featuring a nonzero divergence velocity (A). These two problems considered separately are well-posed in the spaces $[\boldsymbol{V}, H^1(\Omega)]$. Indeed, the inf-sup condition holds even if we restrict to the test functions $q \in H^1(\Omega)$, where $H^1(\Omega) \subset L^2(\Omega)$. Moreover, thanks to the density of $H^1(\Omega)$ in $L^2(\Omega)$, given $g \in L^2(\Omega)$, the equality $(q, \nabla \cdot \boldsymbol{v}) = (q, g) \, \forall q \in H^1(\Omega)$ implies that $\nabla \cdot \boldsymbol{v} = g$ in a weak sense.

The analysis of this control problem seems, however, more complex than that of Problem 2. The explicit definition of the operator mapping the given data to the control variables resorts to pseudodifferential operators involving the pressure. On the other hand, general theoretical results on the convergence of Lagrange multipliers, such as the ones reported in [7, Chap. 6], cannot be applied, since the functional at hand does not contain the control variables explicitly. The well posedness of this formulation is therefore still an open problem. Numerical results suggest, however, that the problem can in fact be solved (see section 4.2).

3.2. More complex functionals. In solving defective boundary problems, it is sometimes worthwhile to include into the formulation of the problem specific features of the solution that are not represented in terms of boundary conditions. However, the knowledge of these features can be useful in completing a defective boundary data set. The approach presented here is quite versatile and prone to include these features. In fact, the functional to be minimized can be properly designed for including these data. Consequently, the set of control variables for the optimization process can be driven by the definition of a modified functional. For the sake of clarity, let us consider a domain given by an almost rectilinear pipe, such as the one depicted in Figure 3.1, where we prescribe a mean value for the pressure on an artificial boundary not normal to the pipe axis. Geometrical considerations suggest that the flow is mainly directed along the axial direction **a**. On the contrary, a solution to Problem 4 would in general provide a significant transverse velocity component on the artificial boundary. Problem 4 indeed forces zero tangential stresses on Γ_i as a natural boundary condition. We can, however, modify the functional in such a way that the minimization process could include the transverse velocity component. This can be done, for instance, by assuming that

$$
J_P(\boldsymbol{w}, s) = \frac{1}{2} \sum_{i=0}^{m} \left(\frac{1}{|\Gamma_i|} \int_{\Gamma_i} s \, d\gamma - P_i \right)^2 + \mathcal{S}(\boldsymbol{w}, s, \Omega), \tag{3.4}
$$

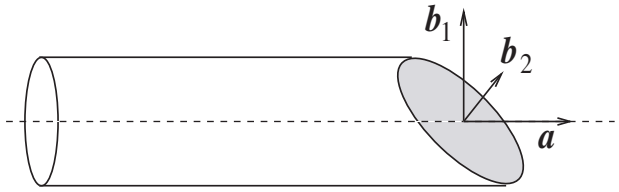


FIG. 3.1. *Artificial section Γ_i, axial direction $\boldsymbol{a}$, and orthogonal directions $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$.*

where $\mathcal{S}$ depends also on the velocity field and on the domain. For example, for the case at hand a possible expression for $\mathcal{S}$ is

$$\mathcal{S} = \mathcal{S}_1(\boldsymbol{w}, \boldsymbol{b}_1, \dots, \boldsymbol{b}_{d-1}) = \frac{1}{2} \sum_{l=1}^{d-1} \sum_{i=0}^{m} \int_{\Gamma_i} |\boldsymbol{w} \cdot \boldsymbol{b}_l|^2 \, d\gamma, \tag{3.5}$$

where $\boldsymbol{b}_l$, $l = 1, \dots, d-1$, are, for each section Γ_i, the orthogonal unit vectors to the axial direction $\boldsymbol{a}$ (see Figure 3.1). In order to improve the effectiveness of the control procedure on the boundary velocity, we augment the set of control variables. For instance, we do not fix a specific direction for the normal stresses and set

$$(-p\boldsymbol{n} + \mu \nabla \boldsymbol{u}\, \boldsymbol{n})|_{\Gamma_j} = -\boldsymbol{K}_j(\boldsymbol{x}), \quad j = 0, \dots, m, \tag{3.6}$$

where, for instance, the $\boldsymbol{K}_j \in (L^2(\Gamma_j))^d$, $j = 0, \dots, m$, are the control variables. From the Lagrangian functional associated with (3.4) and (3.5), with the constraints $(2.2)_1$, $(2.2)_2$, $(2.2)_3$, and (3.6), we obtain the following.

PROBLEM 5. *Given $\boldsymbol{f} \in \boldsymbol{L}^2(\Omega)$ and $\boldsymbol{P} \in \mathbb{R}^{m+1}$, find $\boldsymbol{K}_j \in (L^2(\Gamma_j))^d$, $j = 0, \dots, m$, $\boldsymbol{u}(\boldsymbol{K}_0, \dots, \boldsymbol{K}_m) \in \boldsymbol{V}$, $p(\boldsymbol{K}_0, \dots, \boldsymbol{K}_m) \in H^1(\Omega)$, $\boldsymbol{\lambda}_u \in \boldsymbol{V}$, and $\lambda_p \in H^1(\Omega)$ such that, for all $\boldsymbol{v} \in \boldsymbol{V}$, $q \in H^1(\Omega)$, and $\boldsymbol{\nu} \in (L^2(\Gamma_i))^d$, $i = 0, \dots, m$:*

$$\begin{cases} \text{(P)} \begin{cases} \langle d\mathcal{L}_{\boldsymbol{\lambda}_u}, \boldsymbol{v} \rangle = a(\boldsymbol{u}, \boldsymbol{v}) - (p, \nabla \cdot \boldsymbol{v}) + \displaystyle\sum_{i=0}^{m} \int_{\Gamma_i} \boldsymbol{K}_i \cdot \boldsymbol{v} \, d\gamma - (\boldsymbol{f}, \boldsymbol{v}) = 0, \\ \langle d\mathcal{L}_{\lambda_p}, q \rangle = -(q, \nabla \cdot \boldsymbol{u}) = 0, \end{cases} \\ \text{(A)} \begin{cases} \langle d\mathcal{L}_{\boldsymbol{u}}, \boldsymbol{v} \rangle = \displaystyle\sum_{l=1}^{d-1} \sum_{i=0}^{m} \int_{\Gamma_i} (\boldsymbol{u} \cdot \boldsymbol{b}_l)(\boldsymbol{v} \cdot \boldsymbol{b}_l) \, d\gamma + a(\boldsymbol{v}, \boldsymbol{\lambda}_u) - (\lambda_p, \nabla \cdot \boldsymbol{v}) = 0, \\ \langle d\mathcal{L}_p, q \rangle = \displaystyle\sum_{i=0}^{m} \left(\frac{1}{|\Gamma_i|} \int_{\Gamma_i} p \, d\gamma - P_i \right) \frac{1}{|\Gamma_i|} \int_{\Gamma_i} q \, d\gamma - (q, \nabla \cdot \boldsymbol{\lambda}_u) = 0, \end{cases} \\ \text{(C}_\text{i}\text{)} \quad \langle d\mathcal{L}_{\boldsymbol{k}_i}, \boldsymbol{\nu} \rangle = \displaystyle\int_{\Gamma_i} \boldsymbol{\nu} \cdot \boldsymbol{\lambda}_u \, d\gamma = 0, \quad i = 0, \dots, m. \end{cases}$$

Observe that in this case the optimality conditions (C$_\text{i}$) imply that $\boldsymbol{\lambda}_u|_{\Gamma_i} = \boldsymbol{0}$, $i = 0, \dots, m$. The conditions of fulfillment of the optimal state are more restrictive than the ones in Problem 4 having a larger control variable set.

Other forms for the perturbing term $\mathcal{S}$ can be considered as well. For instance, we could minimize the functional

$$\mathcal{S}_2(\boldsymbol{w}, \boldsymbol{a}) = \frac{1}{2} \sum_{i=0}^{m} \int_{\Gamma_i} \|\nabla \boldsymbol{w}\, \boldsymbol{a}\|_2^2 \, d\gamma. \tag{3.7}$$

Here $\| \cdot \|_2$ is the Euclidean norm made on the components of the vector $\nabla \boldsymbol{w}\, \boldsymbol{a}$. In this way the variation of the velocity along $\boldsymbol{a}$ on the artificial boundary is minimized.

Due to the incompressibility constraint, a variation of the velocity along $\boldsymbol{a}$ will be compensated by variations of the components orthogonal to $\boldsymbol{a}$. For this reason, this form of $\mathcal{S}$ can be considered as an indirect way of forcing null tangential velocity. This leads to the following.

PROBLEM 6. *Given* $\boldsymbol{f} \in \boldsymbol{L}^2(\Omega)$ *and* $\boldsymbol{P} \in \mathbb{R}^{m+1}$, *find* $\boldsymbol{K}_j \in (L^2(\Gamma_j))^d$, $j = 0, \ldots, m$, $\boldsymbol{u}(\boldsymbol{K}_0, \ldots, \boldsymbol{K}_m) \in \boldsymbol{V}$, $p(\boldsymbol{K}_0, \ldots, \boldsymbol{K}_m) \in H^1(\Omega)$, $\boldsymbol{\lambda}_u \in \boldsymbol{V}$, *and* $\lambda_p \in H^1(\Omega)$ *such that, for all* $\boldsymbol{v} \in \boldsymbol{V}$, $q \in H^1(\Omega)$, *and* $\boldsymbol{\nu} \in (L^2(\Gamma_i))^d$, $i = 0, \ldots, m$:

$$
\begin{cases}
\text{(P)} \begin{cases} \langle d\mathcal{L}_{\boldsymbol{\lambda}_u}, \boldsymbol{v} \rangle = a(\boldsymbol{u}, \boldsymbol{v}) - (p, \nabla \cdot \boldsymbol{v}) + \displaystyle\sum_{i=0}^{m} \int_{\Gamma_i} \boldsymbol{K}_i \cdot \boldsymbol{v} \, d\gamma - (\boldsymbol{f}, \boldsymbol{v}) = 0, \\ \langle d\mathcal{L}_{\lambda_p}, q \rangle = -(q, \nabla \cdot \boldsymbol{u}) = 0, \end{cases} \\
\text{(A)} \begin{cases} \langle d\mathcal{L}_{\boldsymbol{u}}, \boldsymbol{v} \rangle = \displaystyle\sum_{i=0}^{m} \int_{\Gamma_i} (\nabla \boldsymbol{u}\, \boldsymbol{a}) \cdot (\nabla \boldsymbol{v}\ \boldsymbol{a}) \, d\gamma + a(\boldsymbol{v}, \boldsymbol{\lambda}_u) - (\lambda_p, \nabla \cdot \boldsymbol{v}) = 0, \\ \langle d\mathcal{L}_p, q \rangle = \displaystyle\sum_{i=0}^{m} \left(\frac{1}{|\Gamma_i|} \int_{\Gamma_i} p \, d\gamma - P_i \right) \frac{1}{|\Gamma_i|} \int_{\Gamma_i} q \, d\gamma - (q, \nabla \cdot \boldsymbol{\lambda}_u) = 0, \end{cases} \\
\text{(C}_\text{i}\text{)} \quad \langle d\mathcal{L}_{\boldsymbol{K}_i}, \boldsymbol{\nu} \rangle = \displaystyle\int_{\Gamma_i} \boldsymbol{\nu} \cdot \boldsymbol{\lambda}_u \, d\gamma = 0, \quad i = 0, \ldots, m.
\end{cases}
$$

We point out that it is possible to extend the previous control problems to the nonlinear case.

Remark 1. It is well known that the correct formulation of the diffusive term in the weak form of the momentum equation is $\int_\Omega \mu(\nabla \boldsymbol{u} + \nabla \boldsymbol{u}^T)\nabla \boldsymbol{v}$. This is crucial, for instance, in fluid structure interaction problems in correct computing of the stresses exchanged by the fluid and structure. On the other hand, in [9] it is observed that the complete diffusive term can have some drawbacks in applying the variational formulation for solving pressure drop problems due to boundary tangential velocity components induced by this approach. These drawbacks in the present framework can be dominated by resorting to functionals (3.4) and (3.5) or (3.4) and (3.7). This aspect will be investigated in a forthcoming paper in the case of problems in deformable domains.

3.3. Using the flow rates as control variables. For solving the mean pressure problem we can also pursue a sort of "dual" approach to the one proposed in section 2 for the flow rate problem. Here the control variables are given by the flow rates Q_j on Γ_j, $j = 1, \ldots, m$. More precisely, we consider the generalized Stokes problem

$$
\begin{cases}
\alpha \boldsymbol{u}(\boldsymbol{Q}) - \mu \triangle \boldsymbol{u}(\boldsymbol{Q}) + \nabla p(\boldsymbol{Q}) = \boldsymbol{f} & \text{in } \Omega, \\
\nabla \cdot \boldsymbol{u}(\boldsymbol{Q}) = 0 & \text{in } \Omega, \\
\boldsymbol{u}(\boldsymbol{Q})|_{\Gamma_w} = \boldsymbol{0}, & \\
(-p(\boldsymbol{Q})\boldsymbol{n} + \mu \nabla \boldsymbol{u}(\boldsymbol{Q})\, \boldsymbol{n})|_{\Gamma_0} = -P_0 \boldsymbol{n}, & \\
\displaystyle\int_{\Gamma_j} \boldsymbol{u}(\boldsymbol{Q}) \cdot \boldsymbol{n} \, d\gamma = Q_j, & j = 1, \ldots, m,
\end{cases} \tag{3.8}
$$

where P_0 is a given constant. Thanks to the Neumann condition on Γ_0, there is no need for compatibility conditions on the flow rate data. When we formulate the corresponding control problem where (3.2) (or its possible modifications) is the functional to be minimized, while (3.8) are the state equations, the latter can be solved according to the augmented Lagrangian approach (Problem 1). Therefore, we need to introduce

the following Lagrangian functional:

$$\begin{aligned}\mathcal{L}(\boldsymbol{w}, s, \boldsymbol{\xi}; \boldsymbol{\lambda}_w, \lambda_s, \boldsymbol{\lambda}_\xi; \boldsymbol{\eta}) = \quad & J_P(s) + a(\boldsymbol{w}, \boldsymbol{\lambda}_w) - (s, \nabla \cdot \boldsymbol{\lambda}_w) + \sum_{i=1}^{m} \xi_i \int_{\Gamma_i} \boldsymbol{\lambda}_w \cdot \boldsymbol{n}\, d\gamma \\ & -(\boldsymbol{f}, \boldsymbol{\lambda}_w) - (\lambda_s, \nabla \cdot \boldsymbol{w}) + \sum_{i=1}^{m} \lambda_{\xi_i} \left(\int_{\Gamma_i} \boldsymbol{w} \cdot \boldsymbol{n}\, d\gamma - \eta_i \right),\end{aligned}$$

where the $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_m)$ are the Lagrange multipliers of the augmented formulation (see [4, 15]). The quantities $\boldsymbol{\lambda}_w, \lambda_s$, and $\lambda_{\xi_1}, \ldots, \lambda_{\xi_m}$ are the adjoint variables related to $\boldsymbol{w}, s$, and $\xi_1, \ldots, \xi_m$, respectively. By forcing the derivatives of $\mathcal{L}$ with respect to all of the variables in correspondence of the solution $[\boldsymbol{u}, p, \boldsymbol{\zeta}; \boldsymbol{\lambda}_u, \lambda_p, \boldsymbol{\lambda}_\zeta; \boldsymbol{Q}]$ to vanish, we obtain the following.

PROBLEM 7. *Given* $\boldsymbol{f} \in \boldsymbol{L}^2(\Omega)$ *and* $\boldsymbol{P} \in \mathbb{R}^m$, *find* $\boldsymbol{Q} \in \mathbb{R}^m$, $\boldsymbol{u}(\boldsymbol{Q}) \in \boldsymbol{V}, p(\boldsymbol{Q}) \in H^1(\Omega), \zeta_j(\boldsymbol{Q}) \in \mathbb{R}$, $j = 1, \ldots, m$, $\boldsymbol{\lambda}_u \in \boldsymbol{V}, \lambda_p \in H^1(\Omega)$, *and* $\lambda_{\zeta_j} \in \mathbb{R}$, $j = 1, \ldots, m$, *such that, for all* $\boldsymbol{v} \in \boldsymbol{V}$, $q \in H^1(\Omega)$, *and* $\nu \in \mathbb{R}$:

$$\begin{cases} \text{(P)} \begin{cases} \langle d\mathcal{L}_{\boldsymbol{\lambda}_u}, \boldsymbol{v} \rangle = a(\boldsymbol{u}, \boldsymbol{v}) - (p, \nabla \cdot \boldsymbol{v}) + \displaystyle\sum_{i=1}^{m} \zeta_i \int_{\Gamma_i} \boldsymbol{v} \cdot \boldsymbol{n}\, d\gamma - (\boldsymbol{f}, \boldsymbol{v}) = 0, \\ \langle d\mathcal{L}_{\lambda_p}, q \rangle = -(q, \nabla \cdot \boldsymbol{u}) = 0, \\ \langle d\mathcal{L}_{\lambda_{\zeta_i}}, \nu \rangle = \left(\displaystyle\int_{\Gamma_i} \boldsymbol{u} \cdot \boldsymbol{n}\, d\gamma - Q_i \right) \nu = 0, \end{cases} \\ \text{(A)} \begin{cases} \langle d\mathcal{L}_{\boldsymbol{u}}, \boldsymbol{v} \rangle = a(\boldsymbol{v}, \boldsymbol{\lambda}_u) - (\lambda_p, \nabla \cdot \boldsymbol{v}) + \displaystyle\sum_{i=1}^{m} \lambda_{\zeta_i} \int_{\Gamma_i} \boldsymbol{v} \cdot \boldsymbol{n}\, d\gamma = 0, \\ \langle d\mathcal{L}_p, q \rangle = \displaystyle\sum_{i=1}^{m} \left(\frac{1}{|\Gamma_i|} \int_{\Gamma_i} p\, d\gamma - P_i \right) \frac{1}{|\Gamma_i|} \int_{\Gamma_i} q\, d\gamma - (q, \nabla \cdot \boldsymbol{\lambda}_u) = 0, \\ \langle d\mathcal{L}_{\zeta_i}, \nu \rangle = \left(\displaystyle\int_{\Gamma_i} \boldsymbol{\lambda}_u \cdot \boldsymbol{n}\, d\gamma \right) \nu = 0, \end{cases} \\ \text{(C}_\text{i}\text{)} \quad \langle d\mathcal{L}_{Q_i}, \nu \rangle = -\lambda_{\zeta_i} \nu = 0, \quad i = 1, \ldots, m. \end{cases}$$

The compatibility conditions in this case yield $\lambda_{\zeta_i} = 0, i = 1, \ldots, m$, which amounts to forcing normal stress in the adjoint problem to be zero on each artificial section (see [4]). This is the dual situation to what we found for the flow rate problems of section 2, in which fluxes of the adjoint problem were zero. Also in this case, in correspondence to the solution of Problem 7, the adjoint problem features the (unique) zero solution.

The previous approach can be extended to the nonlinear case in analogy with section 2.

Also for solving Problem 7 we resort to an iterative scheme, where the optimality conditions (C$_\text{i}$) and $\lambda_{\zeta_i} = 0$ drive the convergence (see [7]).

4. Numerical algorithms. We present now some numerical procedures for the solution of the problems introduced in the previous sections. In particular, in section 4.1 we investigate the flow rate problem. We first introduce a general strategy applied to Problem 2; then we present a special algorithm for the prescription of just one flow rate condition. In section 4.2 we illustrate numerical algorithms for solving Problem 4 as well, as an example of mean pressure problem. Methods presented can be extended to different defective boundary problems.

4.1. Flow rate problems. Let us consider Problem 2. For its numerical solution, we can resort to an iterative method such that at each iteration we solve separately problems (P) and (A) and check condition (C$_\text{j}$) until convergence.

Let us notice that the adjoint problem (A) in Problem 2 depends linearly on the values of the natural boundary condition on Γ_j. Therefore, let $(\boldsymbol{\lambda}_{u,i}, \boldsymbol{\lambda}_{p,i})$ be the *reference solutions* of the problems featuring unit normal stresses:

$$\begin{cases} a(\boldsymbol{v}, \boldsymbol{\lambda}_{u,i}) - (\lambda_{p,i}, \nabla \cdot \boldsymbol{v}) + \displaystyle\int_{\Gamma_i} \boldsymbol{v} \cdot \boldsymbol{n}\, d\gamma = 0 & \forall \boldsymbol{v} \in \boldsymbol{V}, \\ (q, \nabla \cdot \boldsymbol{\lambda}_{u,i}) = 0 & \forall q \in L^2(\Omega) \end{cases} \tag{4.1}$$

for all $i = 1, \ldots, m$; then the solution of (A) in Problem 2 can be written as a linear combination of such reference solutions

$$\boldsymbol{\lambda}_u = \sum_{i=1}^{m} \left(\int_{\Gamma_i} \boldsymbol{u} \cdot \boldsymbol{n}\, d\gamma - Q_i \right) \boldsymbol{\lambda}_{u,i}, \qquad \lambda_p = \sum_{i=1}^{m} \left(\int_{\Gamma_i} \boldsymbol{u} \cdot \boldsymbol{n}\, d\gamma - Q_i \right) \lambda_{p,i}.$$

Hence, the computational cost of the adjoint problem reduces to the m problems (4.1), which can be computed preliminarily. Obviously, this is not true anymore for the nonlinear case.

Observe first of all that if we introduce the affine map $\mathcal{T}_\tau : \mathbb{R}^m \to \mathbb{R}^m$:

$$\mathcal{T}_\tau(\boldsymbol{k}) := \boldsymbol{k} + \tau \mathcal{B}_{\mathbf{0}}(\mathcal{B}_{\boldsymbol{f}}(\boldsymbol{k}) - \boldsymbol{Q}),$$

where $\mathcal{B}_{\boldsymbol{f}}$ and $\mathcal{B}_{\mathbf{0}}$ have been introduced in section 2.1 and τ is a given real coefficient, Problem 2 can be reformulated as the computation of a fixed point solution:

$$\hat{\boldsymbol{k}} = \mathcal{T}_\tau(\hat{\boldsymbol{k}}). \tag{4.2}$$

This naturally leads to the numerical method based on the fixed point iteration

$$\boldsymbol{k}^{l+1} = \mathcal{T}_\tau(\boldsymbol{k}^l),$$

where l denotes the iteration index. The following algorithm reflects this scheme.

ALGORITHM 1.

1. Solve for $j = 1, \ldots, m$ problems (4.1) discretized in space, giving solutions $\boldsymbol{\lambda}_{u,j,h}$ and $\lambda_{p,j,h}$ for $j = 1, \ldots, m$.
2. Loop: Given $k^1_{j,h}$, $j = 1, \ldots, m$, and ε, set $l = 1$ and do until convergence
   - Solve

$$\begin{cases} a(\boldsymbol{u}_h^l, \boldsymbol{v}_h) - (p_h^l, \nabla \cdot \boldsymbol{v}_h) + \displaystyle\sum_{i=1}^{m} \int_{\Gamma_i} k_{i,h}^l \boldsymbol{v}_h \cdot \boldsymbol{n}\, d\gamma - (\boldsymbol{f}, \boldsymbol{v}_h) = 0 & \forall \boldsymbol{v}_h \in \boldsymbol{V}_h, \\ (q_h, \nabla \cdot \boldsymbol{u}_h^l) = 0 & \forall q_h \in Q_h. \end{cases}$$

   - Compute the adjoint solutions

$$\boldsymbol{\lambda}_{u,h}^l = \sum_{i=1}^{m} \left(\int_{\Gamma_i} \boldsymbol{u}_h^l \cdot \boldsymbol{n}\, d\gamma - Q_i \right) \boldsymbol{\lambda}_{u,i,h},$$

$$\lambda_{p,h}^l = \sum_{i=1}^{m} \left(\int_{\Gamma_i} \boldsymbol{u}_h^l \cdot \boldsymbol{n}\, d\gamma - Q_i \right) \lambda_{p,i,h}.$$

   - Convergence test: If $|\int_{\Gamma_j} \boldsymbol{\lambda}_{u,h}^l \cdot \boldsymbol{n}\, d\gamma| < \varepsilon \, \forall j = 1, \ldots, m$, then break

`else` $k_{j,h}^{l+1} = k_{j,h}^{l} + \tau^l \int_{\Gamma_j} \boldsymbol{\lambda}_{u,h}^{l} \cdot \boldsymbol{n}\, d\gamma, \quad \forall j = 1, \dots, m,$ `and set` $l = l + 1$.

`end`

We have the following proposition.

PROPOSITION 4.1. *For an appropriate choice of* τ*, Algorithm* 1 *converges to the solution of Problem* 2.

Proof. Let us prove that $\mathcal{T}_\tau$ is a contraction for suitable values of τ. For a given increment $\delta\boldsymbol{k}$, thanks to the linearity of the involved operators, we obtain

$$\mathcal{T}_\tau(\boldsymbol{k} + \delta\boldsymbol{k}) - \mathcal{T}_\tau(\boldsymbol{k}) = (\mathcal{I} + \tau\mathcal{B}_{\mathbf{0}}^2)\delta\boldsymbol{k}, \tag{4.3}$$

where $\mathcal{I} + \tau\mathcal{B}_{\mathbf{0}}^2$ is a square matrix, $\mathcal{I}$ being the identity. $\mathcal{B}_{\mathbf{0}}$ being symmetric and negative definite (see Proposition 2.1), T_τ is a contraction for

$$-\frac{2}{\rho(\mathcal{B}_{\mathbf{0}}^2)} < \tau < 0,$$

where $\rho(\mathcal{B}_{\mathbf{0}})$ is the spectral radius of $\mathcal{B}_{\mathbf{0}}$. For any τ in this range, Algorithm 1 converges to the solution of Problem 2. The optimal value for the coefficient τ, which guarantees the highest convergence rate, is given by

$$\tau_{opt} = -\frac{2}{\rho(\mathcal{B}_0^2) + \mu_{min}},$$

where μ_{min} is the minimum eigenvalue of $\mathcal{B}_0^2$. ∎

The knowledge of the optimal value of τ is not exploitable in practice, as it is based on the knowledge of the spectrum of $\mathcal{B}_{\mathbf{0}}$, which is not in general available. Dynamic strategies for the selection of the parameter τ, regarded as a function of the iteration l, can be considered. In particular, by denoting by $\boldsymbol{\chi}_h^l$ the numerical solution obtained at iteration l, we can resort to the choice

$$\tau^l = \tau_N^l = -\frac{J_Q(\boldsymbol{u}_h^l)}{\|\mathcal{L}_{\boldsymbol{k}}(\boldsymbol{\chi}_h^l)\|_2^2} \tag{4.4}$$

that amounts to updating the k_j as follows:

$$k_{j,h}^{l+1} = k_{j,h}^{l} - \frac{\frac{1}{2}\sum_{j=1}^{m}\left(\int_{\Gamma_j} \boldsymbol{u}_h^l \cdot \boldsymbol{n}\, d\gamma - Q_j\right)^2 \int_{\Gamma_j} \boldsymbol{\lambda}_{u,h}^{l} \cdot \boldsymbol{n}\, d\gamma}{\sum_{j=1}^{m} |\int_{\Gamma_j} \boldsymbol{\lambda}_{u,h}^{l} \cdot \boldsymbol{n}\, d\gamma|^2}.$$

This choice stems from the application of the classical Newton method for the equation $J_Q(\boldsymbol{k}) = 0$. Indeed, it can be shown that $\langle J_Q'(\boldsymbol{k}), \boldsymbol{\nu}\rangle = \langle \mathcal{L}_{\boldsymbol{k}}(\boldsymbol{s}), \boldsymbol{\nu}\rangle \, \forall \boldsymbol{\nu} \in \mathbb{R}^m$. A further improvement can be yielded by observing that J_Q is a quadratic functional and the associate solution is supposed to be of multiplicity 2, so that we select $\tau^l = 2\tau_N^l$. As we will see in section 5 this is an affective choice.

Algorithm 1 can be suitably extended to nonlinear Problem 3.

4.1.1. Prescription of a single flow rate. In the particular case of only one prescribed flux, we can resort to a different algorithm based on the variational formulation of (2.2) (or (2.12) for the nonlinear case). For instance, considering (2.12) together with $(2.2)_{2-5}$ and $(2.1)_2$ and with $m = 1$, by setting $\boldsymbol{v} = \boldsymbol{u}$ we obtain:

$$a(\boldsymbol{u}, \boldsymbol{u}) + ((\boldsymbol{u} \cdot \nabla)\boldsymbol{u}, \boldsymbol{u}) + \int_\Gamma k\boldsymbol{u} \cdot \boldsymbol{n}\, d\gamma = (\boldsymbol{f}, \boldsymbol{u}),$$

and then, by exploiting the fact that k is constant, if $Q \neq 0$, we can calculate directly

$$k = \frac{(\boldsymbol{f}, \boldsymbol{u}(k)) - \|\boldsymbol{u}(k)\|_a^2 - ((\boldsymbol{u}(k)\cdot\nabla)\boldsymbol{u}(k), \boldsymbol{u}(k))}{Q} = \phi(k), \tag{4.5}$$

where we have set $\|\boldsymbol{v}\|_a^2 := a(\boldsymbol{v}, \boldsymbol{v})$. A fixed point iteration may then be set up as follows.

ALGORITHM 2. `Given` k_h^1, $\boldsymbol{u}_h^0$, `and` ε, `set` $l = 1$ `and do until convergence`

1. `Solve`

$$\begin{cases} a(\boldsymbol{u}_h^l, \boldsymbol{v}_h) + ((\boldsymbol{u}_h^{l-1}\cdot\nabla)\boldsymbol{u}_h^l, \boldsymbol{v}_h) - (p_h^l, \nabla\cdot\boldsymbol{v}_h) + \displaystyle\int_\Gamma k_h^l \boldsymbol{v}_h\cdot\boldsymbol{n}\, d\gamma = (\boldsymbol{f}, \boldsymbol{v}_h) & \forall \boldsymbol{v}_h \in \boldsymbol{V}_h, \\ (q_h, \nabla\cdot\boldsymbol{u}_h^l) = 0 & \forall q_h \in Q_h. \end{cases}$$

2. `Compute`

$$k_h^{l+1} = \frac{(\boldsymbol{f}, \boldsymbol{u}_h^l) - \|\boldsymbol{u}_h^l\|_a^2 - ((\boldsymbol{u}_h^{l-1}\cdot\nabla)\boldsymbol{u}_h^l, \boldsymbol{u}_h^l)}{Q}.$$

3. `Convergence test: if` $|k_h^{l+1} - k_h^l| < \varepsilon$, `then exit`
4. `else` $l = l + 1$

`end`

Remark 2. In the case of an unsteady problem, Algorithm 2 fails if there exists a $\widetilde{t}$ such that $Q(\widetilde{t})$=0.

In the particular case of a linear problem with $\boldsymbol{f} = \boldsymbol{0}$, we do not even need to resort to a fixed point iteration. Indeed, in this case the velocity $\boldsymbol{u}$ depends linearly on k. Therefore, by denoting with $\widetilde{\boldsymbol{u}}$ the solution obtained for $k = 1$, $\boldsymbol{u}$ is computed thanks to $\boldsymbol{u} = k\widetilde{\boldsymbol{u}}$. Therefore, from (4.5), we obtain $k = -k^2\|\widetilde{\boldsymbol{u}}\|_a^2/Q$, yielding the following explicit expression for k:

$$k = -\frac{Q}{\|\widetilde{\boldsymbol{u}}\|_a^2}.$$

4.2. Mean pressure problem. Algorithm 1 can be properly extended to the mean pressure problem presented in section 3. If the control variable is a vector of constants and the problem at hand is linear (like in Problems 4 and 7), the adjoint problem can be still solved effectively by solving a suitable problem out of the iterative loop. For instance, if we refer to Problem 4, and we denote by $(\widetilde{\boldsymbol{\lambda}}_{u,i}, \widetilde{\lambda}_{p,i})$ the solutions of

$$\begin{cases} a(\widetilde{\boldsymbol{\lambda}}_{u,i}, \boldsymbol{v}) - (\widetilde{\lambda}_{p,i}, \nabla\cdot\boldsymbol{v}) = 0 & \forall \boldsymbol{v} \in \boldsymbol{V}, \\ -(q, \nabla\cdot\widetilde{\boldsymbol{\lambda}}_{u,i}) = -\displaystyle\int_{\Gamma_i} q\, d\gamma & \forall q \in H^1(\Omega), \end{cases} \tag{4.6}$$

then we can obtain the solution to the adjoint problem (A) by setting

$$\boldsymbol{\lambda}_u = \sum_{i=0}^{m} \left(\frac{1}{|\Gamma_i|}\int_{\Gamma_i} p(\boldsymbol{k})\, d\gamma - P_i \right) \widetilde{\boldsymbol{\lambda}}_{u,i}, \qquad \lambda_p = \sum_{i=0}^{m} \left(\frac{1}{|\Gamma_i|}\int_{\Gamma_i} p(\boldsymbol{k})\, d\gamma - P_i \right) \widetilde{\lambda}_{p,i}.$$

Let $\boldsymbol{V}_h \subset \boldsymbol{V}$ and $Q_h^1 \subset H^1(\Omega)$, respectively, be a couple of inf-sup compatible finite-dimensional spaces. We denote by $\{\boldsymbol{\varphi}_i\}$ $(i = 1, \ldots, n_{vel})$ and $\{\psi_i\}$ $(i = 1, \ldots, n_p)$ the basis function set of $\boldsymbol{V}_h$ and Q_h^1, respectively, n_{vel} and n_p being their dimensions.

For solving Problem 4, we propose the following algorithm.

ALGORITHM 3.

1. `Solve, for` $i = 0, \dots, m$, `problems (4.6) discretized in space, giving the solutions` $\widetilde{\boldsymbol{\lambda}}_{u,i,h}$ `and` $\widetilde{\lambda}_{p,i,h}$.
2. `Loop: Given` $k^1_{j,h}$, $j = 0, \dots, m$, `and` ε, `set` $l = 1$ `and do until convergence`
   - `Solve`

$$\begin{cases} a(\boldsymbol{u}^l_h, \boldsymbol{v}_h) - (p^l_h, \nabla \cdot \boldsymbol{v}_h) + \displaystyle\sum_{i=0}^{m} \int_{\Gamma_i} k^l_{i,h} \boldsymbol{v}_h \cdot \boldsymbol{n}\, d\gamma - (\boldsymbol{f}, \boldsymbol{v}_h) = 0 & \forall \boldsymbol{v}_h \in \boldsymbol{V}_h \\ (q_h, \nabla \cdot \boldsymbol{u}^l_h) = 0 & \forall q_h \in Q^1_h. \end{cases}$$

   - `Compute the adjoint solutions`

$$\boldsymbol{\lambda}^l_{u,h} = \sum_{i=0}^{m} \left(\frac{1}{|\Gamma_i|} \int_{\Gamma_i} p^l_h \, d\gamma - P_i \right) \widetilde{\boldsymbol{\lambda}}_{u,i,h},$$

$$\lambda^l_{p,h} = \sum_{i=0}^{m} \left(\frac{1}{|\Gamma_i|} \int_{\Gamma_i} p^l_h \, d\gamma - P_i \right) \widetilde{\lambda}_{p,i,h}.$$

   - `Convergence test: If` $|\int_{\Gamma_j} \boldsymbol{\lambda}^l_{u,h} \cdot \boldsymbol{n}\, d\gamma| < \varepsilon \, \forall j = 0, \dots, m$, `then exit`
   - `else` $k^{l+1}_{j,h} = k^l_{j,h} + \tau^l \int_{\Gamma_j} \boldsymbol{\lambda}^l_{u,h} \cdot \boldsymbol{n}\, d\gamma \quad \forall j = 0, \dots, m$ `and set` $l = l + 1$
3. `end`

In order to investigate the properties of this algorithm, we introduce the following operators.

1. $\mathcal{N}_{\boldsymbol{f},h} : \mathbb{R}^{m+1} \to Q^1_h$. This operator maps a vector $\boldsymbol{h} \in \mathbb{R}^{m+1}$ to the function s_h, where $[\boldsymbol{w}_h, s_h]$ is the unique solution of

$$\begin{cases} a(\boldsymbol{w}_h, \boldsymbol{v}_h) - (s_h, \nabla \cdot \boldsymbol{v}_h) = -\displaystyle\sum_{j=0}^{m} h_j \int_{\Gamma_j} \boldsymbol{v}_h \cdot \boldsymbol{n}\, d\gamma + (\boldsymbol{f}, \boldsymbol{v}_h) & \forall \boldsymbol{v}_h \in \boldsymbol{V}_h, \\ (q_h, \nabla \cdot \boldsymbol{w}_h) = 0 & \forall q_h \in Q^1_h. \end{cases}$$

   Let A denote the $n_{vel} \times n_{vel}$ matrix with entries $a_{ij} = a(\boldsymbol{\varphi}_j, \boldsymbol{\varphi}_i)$, B the $n_{vel} \times n_p$ matrix $b_{ij} = -(\psi_j, \boldsymbol{\varphi}_i)$, G the matrix with entries $g_{ij} = \int_{\Gamma_i} \boldsymbol{\varphi}_j \cdot \boldsymbol{n}\, d\gamma$, and $\widetilde{\mathbf{f}}$ the vector whose components are $(\boldsymbol{f}, \boldsymbol{\varphi}_i)$. Then the algebraic counterpart of $\mathcal{N}_{\boldsymbol{f},h}$ reads:

$$\mathbf{s} = \left(BA^{-1}B^T\right)^{-1} \left(-BA^{-1}G^T\mathbf{h} + BA^{-1}\widetilde{\mathbf{f}}\right).$$

   We observe, by the way, that G is the algebraic representation of the finite-dimensional operator $\mathcal{A}_h$ obtained by taking the analogous of operator $\mathcal{A}$ defined in (2.8) from the subspace $\mathbf{V}_h$ to $\mathbb{R}^{m+1}$.
2. $\widetilde{\mathcal{G}}_h : Q^1_h \to \mathbb{R}^{m+1}$ such that $\widetilde{\mathcal{G}}_h(s_h) = (\frac{1}{|\Gamma_0|}\int_{\Gamma_0} s_h \, d\gamma, \frac{1}{|\Gamma_1|}\int_{\Gamma_1} s_h \, d\gamma, \dots, \frac{1}{|\Gamma_m|}\int_{\Gamma_m} s_h \, d\gamma)$. The algebraic counterpart of this operator is given by the matrix denoted by $\widetilde{G}$ with entries $\widetilde{g}_{ij} = \frac{1}{|\Gamma_i|}\int_{\Gamma_i} \psi_j \, d\gamma$.
3. $\mathcal{S}_h : \mathbb{R}^{m+1} \to \boldsymbol{V}_h$ maps a vector $\mathbf{r}$ in $\boldsymbol{w}_h$, where $[\boldsymbol{w}_h, s_h]$ is the unique solution of the generalized Stokes problem

$$\begin{cases} a(\boldsymbol{w}_h, \boldsymbol{v}_h) - (s_h, \nabla \cdot \boldsymbol{v}_h) = 0 & \forall \boldsymbol{v}_h \in \boldsymbol{V}_h, \\ -(q_h, \nabla \cdot \boldsymbol{w}_h) = -\displaystyle\sum_{i=0}^{m} r_i \frac{1}{|\Gamma_i|} \int_{\Gamma_i} q_h \, d\gamma & \forall q_h \in Q^1_h. \end{cases} \tag{4.7}$$

The algebraic formulation of this operator reads:

$$\mathbf{w} = A^{-1}B^T \left(BA^{-1}B^T\right)^{-1} \widetilde{G}^T \mathbf{r}.$$

Now set $\mathcal{H}_h = \mathcal{A}_h \circ \mathcal{S}_h$ and $\mathcal{E}_{\boldsymbol{f},h} = \widetilde{\mathcal{G}}_h \circ \mathcal{N}_{\boldsymbol{f},h}$. Then it is promptly verified that the discrete control problem at hand reads:

$$\mathcal{H} \circ (\mathcal{E}_{\boldsymbol{f},h}(\mathbf{k}) - \mathbf{P}) = 0. \tag{4.8}$$

Correspondingly, observe that the algorithm proposed above can be regarded as the Richardson iteration: Given $\mathbf{k}^l$, find $\mathbf{k}^{l+1}$ such that

$$\mathbf{k}^{l+1} = \mathbf{k}^l + \tau^l \mathcal{H}_h \circ \left(\mathbf{P} - \mathcal{E}_{\boldsymbol{f},h}(\mathbf{k}^l)\right).$$

Unfortunately, an analysis proving the definiteness of the residual operator on the right-hand side in this case, which would imply the convergence of the iterative scheme, is still missing.

We limit ourselves to point out that in algebraic terms if $\Sigma = BA^{-1}B^T$ is the Schur complement associated to the Stokes problem, (4.8) can be rewritten as:

$$GA^{-1}B^T\Sigma^{-1}\widetilde{G}^T \left(\widetilde{G}\Sigma^{-1}BA^{-1}\left(-G^T\mathbf{k} + \mathbf{f}\right) - \mathbf{P}\right) = 0.$$

We set $Z = \widetilde{G}\Sigma^{-1}BA^{-1}G^T$ so that, thanks to the symmetry of Σ and A, the matrix problem can be rewritten:

$$Z^T Z\mathbf{k} = Z^T \left(\widetilde{G}\Sigma^{-1}BA^{-1}\mathbf{f} - \mathbf{P}\right).$$

This formulation highlights that the matrix at hand is semidefinite positive. This excludes that the algorithm could yield divergent sequences, even if the convergence is still not ensured. This issue will be investigated in a future work. Numerical evidence suggests, however, that with a proper choice of the parameter τ the numerical solution converges.

Remark 3. A similar algorithm can be devised for Problem 7. In this case both the adjoint problem out of the iterative cycle and the problem in the iterative cycle are solved with the augmented formulation, in particular, by resorting to the *GMRES+Schur complement* scheme introduced in [15].

5. Numerical results. In this section we present several numerical results to validate the algorithms introduced in section 4. In particular, in section 5.1 we focus on the flow rate problem and in section 5.2 on the mean pressure problem.

5.1. Flow rate problems. All simulations of this section and of section 5.2 have been implemented by using the two-dimensional (2D) finite element library *Freefem++* (see [8]), with a discretization time step $\Delta t = 0.01$ s and using conforming $\mathbb{P}_2 - \mathbb{P}_1$ elements. In the first simulations we present, the computational domain is a rectangle R of 6×1 cm (see Figure 5.1, left), while the fluid kinematic viscosity is $\mu = 0.035$ cm^2/s. This is the usual value adopted for blood viscosity in computational hemodynamics. We prescribed both a steady ($Q = 0.1$ cm^2/s) and a pulsatile ($Q = 0.15 + 0.1\cos(2\pi t)$ cm^2/s) flow rate at the inlet Γ_{in} of R. We have used an almost uniform triangular grid with spacing $h = 0.05$ cm. We have considered in this first test case the Stokes problem.

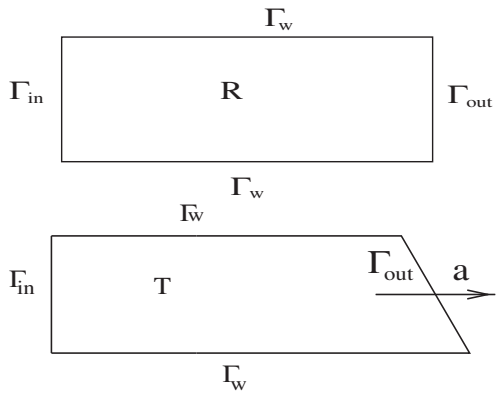


FIG. 5.1. *Computational domains R and T.*

TABLE 5.1
Relative errors of the numerical solutions obtained with Algorithms 1 *and* 2*, compared to the reference solution.*

	Time	D_{I}	D_{II}
Steady case		0.0000003	0.0000050
Unsteady case	$t = 1.25$ s	0.000230	0.000160
	$t = 1.30$ s	0.000154	0.000156
	$t = 1.60$ s	0.000348	0.000348

In order to verify the accuracy of the solutions obtained with the two algorithms, we provide comparisons with the analytical Poiseuille solution in the steady case (see Table 5.1, where $D_{\mathrm{I}} = \|\boldsymbol{u}_{pois} - \boldsymbol{u}_{\mathrm{I}}\|_{L^2(\Gamma_{in})}/\|\boldsymbol{u}_{pois}\|_{L^2(\Gamma_{in})}$ and $D_{\mathrm{II}} = \|\boldsymbol{u}_{pois} - \boldsymbol{u}_{\mathrm{II}}\|_{L^2(\Gamma_{in})}/\|\boldsymbol{u}_{pois}\|_{L^2(\Gamma_{in})}$, $\boldsymbol{u}_{pois}, \boldsymbol{u}_{\mathrm{I}}$ and $\boldsymbol{u}_{\mathrm{II}}$ are the Poiseuille solution and the numerical solutions obtained with Algorithms 1 and 2, respectively).

In the unsteady case we compare the results with the solution obtained by the augmented formulation (see [15, 17]). In particular, this formulation has been faced with the *GMRES+Schur complement* (GS) algorithm. In Table 5.1 we report $D_{\mathrm{I}} = \|\boldsymbol{u}_{augm} - \boldsymbol{u}_{\mathrm{I}}\|_{L^2(\Gamma_{in})}/\|\boldsymbol{u}_{augm}\|_{L^2(\Gamma_{in})}$ and $D_{\mathrm{II}} = \|\boldsymbol{u}_{augm} - \boldsymbol{u}_{\mathrm{II}}\|_{L^2(\Gamma_{in})}/\|\boldsymbol{u}_{augm}\|_{L^2(\Gamma_{in})}$, where $\boldsymbol{u}_{augm}$ is the numerical solutions obtained with the augmented formulation. In Figure 5.2, the axial velocity at the inlet Γ_{in}, computed with the two algorithms in the steady (top, left) and in the unsteady case, is compared with the reference solution. Differences among the different solutions are clearly small.

The number of iterations for the three algorithms is shown in Table 5.2. In the unsteady case, we give the mean values over the entire computation in time. For Algorithm 1 we compare different possible choices of τ. An Aitken acceleration procedure has been successfully implemented to speed up convergence. Algorithm 2 has been considered only in the case of Aitken procedure implementation.

In the steady case, as expected from section 4.1.1, Algorithm 2 is very fast. In the unsteady case, Algorithm 1 with $\tau = 2\tau_N^l$ gives the best performances. The number of iterations is in fact comparable with the one of the GMRES algorithm for the augmented problem. The latter is supposed (in exact arithmetic) to converge in $m + 1$ iterations per time step, where m is the number of prescribed flow rates (see [15]). In these two cases, we also compute the (mean) CPU time used to perform an iteration. These values are shown in Table 5.2 in the brackets, and they highlight

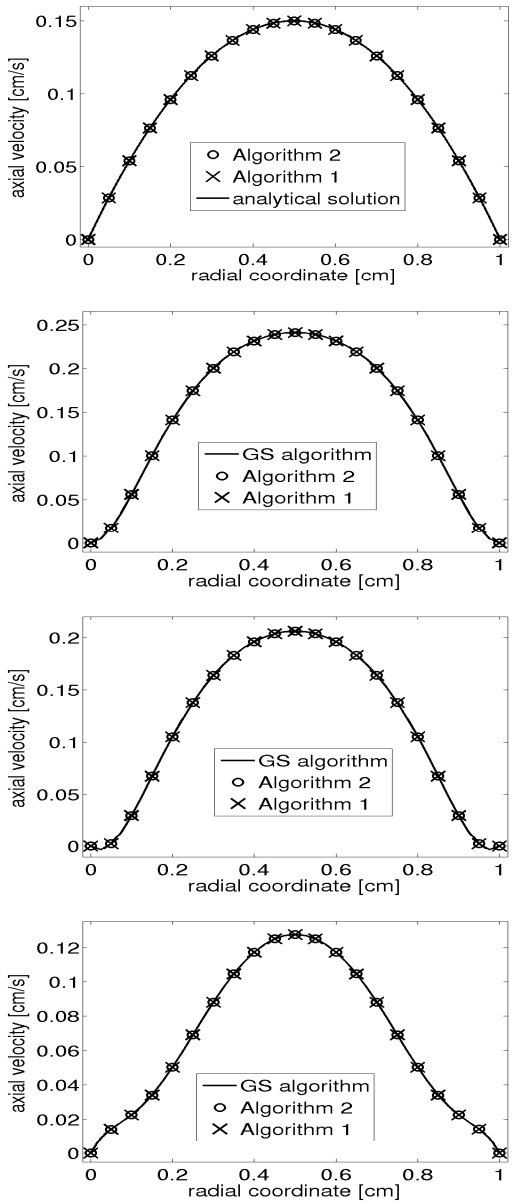


FIG. 5.2. *Axial velocity computed with Algorithms* 1 *and* 2 *and the reference solution—steady simulation (top, left) and unsteady simulation,* $t = 1.25\,s$ *(top, right),* $t = 1.30\,s$ *(bottom, left), and* $t = 1.60\,s$ *(bottom, right);* $toll = 10^{-7}$.

that also the computational costs of Algorithm 1 with $\tau = 2\tau_N^l$ and of the GMRES algorithm for the augmented problem are comparable.

In the second test case, we consider the computational domain depicted in Figure 5.3 representing a 2D simplified model of a bypass anastomosis (see, e.g., [17]). The space discretization step is of $h = 0.1$ cm. We solve the Navier–Stokes equations by extending Algorithm 1 to the nonlinear case. In this case the adjoint problem needs to be solved in the time-advancing loop. We prescribe the flow rates $Q_1 = \cos(2\pi t)$

TABLE 5.2

Number of iterations for the convergence of Algorithm 1, *Algorithm* 2, *and the GS algorithm (in average for the unsteady case) and, in brackets, CPU times in seconds for Algorithm* 1 *with* $\tau = 2\tau_N^l$ *and for the GS algorithm.*

	Alg. 1 $\tau = -1$	Alg. 1 $\tau = \tau_N^l$	Alg. 1 $\tau = \tau_N^l$+Aitken	Alg. 1 $\tau = 2\tau_N^l$	Alg.2 +Aitken	GS
Steady case	87	22	4	2	1	2
Unsteady case	-	5.08	3.81	1.98 (6.37 s)	4.50	2.00 (6.41 s)

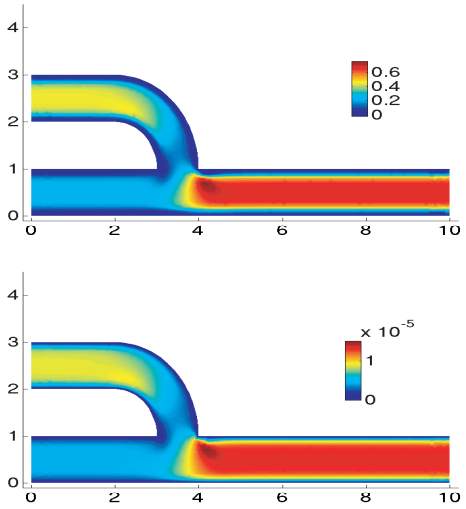


FIG. 5.3. *Anastomosis simulation—Algorithm* 1—$t = 1.2$ *s; on the left the axial velocity and on the right the difference with the GS algorithm; toll* $= 10^{-9}$.

cm^2/s on the upper inlet and $Q_2 = 0.5\cos(2\pi t)$ cm^2/s on the lower inlet. We show in Figure 5.3 the axial velocity on the left and the difference with the solution obtained by using the augmented formulation on the right. The two numerical results are in excellent agreement with the difference bounded by 10^{-5}. Moreover, we point out that this error is independent on the mesh size. The mean number of iterations per time step required in this case by Algorithm 1 with the choice $\tau = 2\tau_N^l$ is 18.22 versus the $m + 1 = 3$ iterations needed by the GS algorithm. The (mean) CPU times per iteration are 169.2 s and 16.1 s, respectively. The slow convergence of Algorithm 1 is due to the ill-conditioning of the problem when prescribing 2 flow rates. In this case it is well known that the convergence of the steepest descent method can be quite slow. One possibility that will be investigated in future work is to search for conjugate directions.

5.2. Mean pressure problems. In this section, we illustrate the performances of Algorithm 3 and its extensions to solve Problems 5, 6, and 7. In the first test case, we prescribe the mean pressure $P = \sin(2\pi t)$ $\text{g}/(\text{s}^2\ \text{cm})$ at the outlet Γ_{out} of the domain R (with $h = 0.1$ cm) of Figure 5.1 left. We compare the two strategies of using the normal component of the normal stress as a control variable (strategy PS; see Problem 4) and the one using flow rates (strategy PF; see Problem 7). Computations are performed for the Stokes problem.

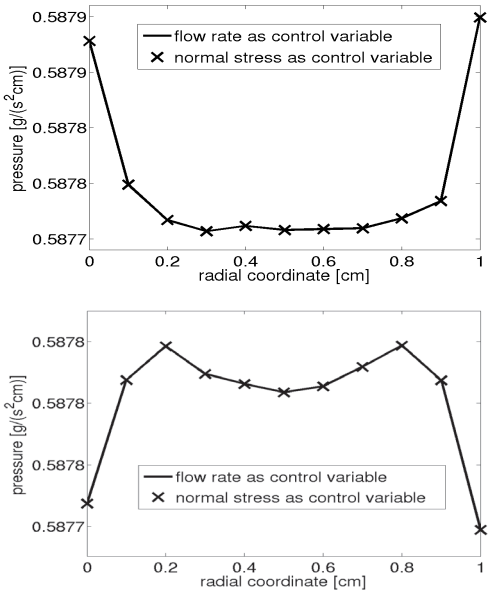


FIG. 5.4. *Mean pressure prescribed at the outlet at $t = 1.1$ s, $P = 0.5878\, g/(s^2\, cm)$ (left) and at $t = 1.4$ s, $P = 0.5878\, g/(s^2\, cm)$ (right); toll $= 10^{-7}$.*

TABLE 5.3
Relative errors in the $L^2(\Gamma)$ norm of the solutions obtained using the flow rate (PF) and the normal stress (PS), respectively, as a control variable.

Time	PF	PS
$t = 1.10\, s$	$9.327 \cdot 10^{-5}$	$9.329 \cdot 10^{-5}$
$t = 1.40\, s$	$3.569 \cdot 10^{-5}$	$3.578 \cdot 10^{-5}$

Figure 5.4 and Table 5.3 show that the two strategies give close results. However, the PF strategy is more problematic in terms of convergence of the numerical scheme. As a matter of fact, we observed that it diverges for the nonstationary choices of τ that we tested. Finally, we resorted to set a static value $\tau = -10^{-6}$. The average number of iterations in this case has been 31.31. It is worthwhile to observe that this strategy is more expensive also in terms of CPU times, since for each iteration it requires the solution of $m + 1 = 2$ generalized Stokes problems for solving the augmented formulation.

On the contrary, convergence with the PS strategy is obtained with an average of 2.00 iterations (with $\tau = 2\tau_N^l$). Each iteration requires the solution of only one generalized Stokes problem.

We conclude that, if the control variables can be chosen, strategy PS is definitely the best one.

In the second set of simulations, we prescribe the mean pressure $P = 1$ g/(s^2 cm) at the outlet Γ_{out} of the domain T (see Figure 5.1, right). We indicate with $\boldsymbol{a}$ the axial direction. By minimizing functional (3.2), an undesirable transverse velocity and, consequently, an incorrect axial velocity at the outlet occur (Figure 5.5, top). To avoid these effects, we solve Problems 5 and 6; i.e., we minimize functional (3.4) with (3.5) and (3.7), respectively. We consider the same iterative scheme of Algorithm

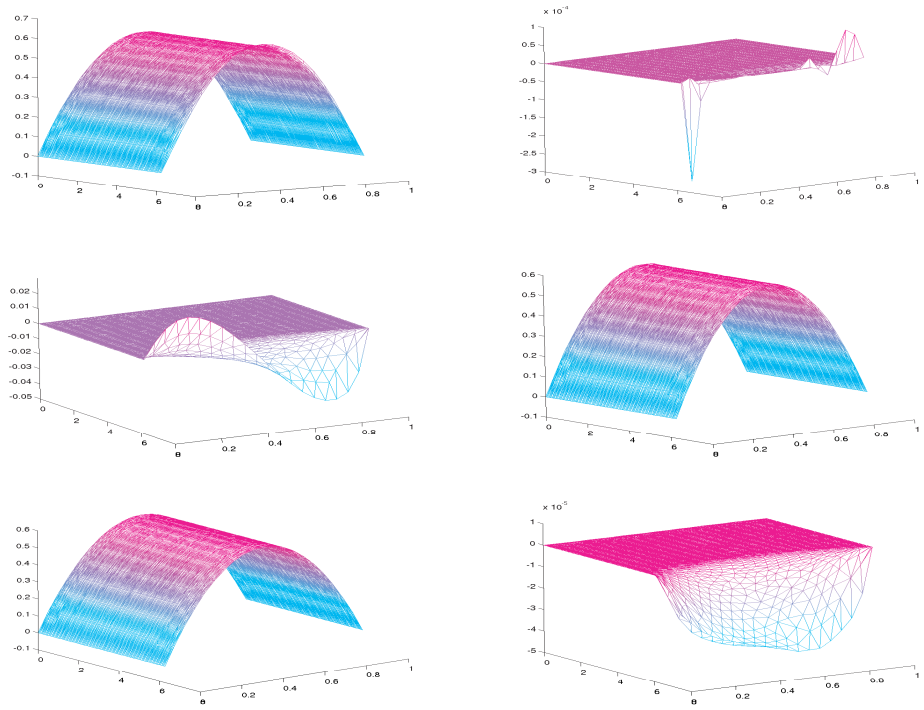


FIG. 5.5. *Axial (left) and tangential (right) velocity in* cm^2/s *obtained by prescribing* $P = 1\,g/(s^2\,cm)$ *at the outlet of* T *minimizing* (3.2) *(top),* (3.4) *with* (3.5) *(middle), and* (3.4) *with* (3.7) *(bottom); toll* $= 10^{-7}$.

3. Figure 5.5 illustrates the effectiveness of these strategies, yielding errors of about 3×10^{-4} in the former case and 5×10^{-5} in the latter one.

Acknowledgments. The authors thank Luca Dedè, Jean-Frederic Gerbeau, and Carlo D'Angelo for their useful comments and suggestions.

REFERENCES

[1] R. Becker, H. Kapp, and R. Rannacher, *Adaptive finite element methods for optimal control of Partial Differential Equations: Basic concept*, SIAM J. Control Optim., 39 (2000), pp. 113–132.

[2] K. Chrysafinos, *Error estimates for time-discretizations for the velocity tracking problem for Navier-Stokes flows by penalty methods*, Discrete Contin. Dyn. Syst. Ser. B, 6 (2006), pp. 1077–1096.

[3] L. Dedé and A. Quarteroni, *Optimal control and numerical adaptivity for advection-diffusion equations*, M2AN, Math. Model. Numer. Anal., 39 (2005), pp. 1019–1040.

[4] L. Formaggia, J. F. Gerbeau, F. Nobile, and A. Quarteroni, *Numerical treatment of defective boundary conditions for the Navier–Stokes equation*, SIAM J. Numer. Anal., 40 (2002), pp. 376–401.

[5] L. Formaggia, F. Nobile, A. Quarteroni, and A. Veneziani, *Multiscale modelling of the vascular system: A preliminary analysis*, Comput. Vis. Sci., 2 (1999), pp. 75–83.

[6] R. Glowinski and O. Pironneau, *On mixed finite element approximation of the Stokes problem*, Numer. Math., 33 (1979), pp. 397–424.

[7] M. Gunzburger, *Perspectives in Flow Control and Optimizations*, SIAM, Philadelphia, 2003.

[8] F. Hecht, O. Pironneau, F. Le Hyaric, and K. Ohtsuka, *Freefem++*, www.freefem.org.

[9] J. Heywood, R. Rannacher, and S. Turek, *Artificial boundary and flux and pressure conditions for the incompressible Navier-Stokes equations*, Internat. J. Numer. Methods Fluids,

22 (1996), pp. 325–352.

[10] W. Liu and N. Yan, *A posteriori error estimates for control problems governed by Stokes equations*, SIAM J. Numer. Anal., 40 (2002), pp. 1850–1869.

[11] V. Martin, F. Clément, A. Decoene, and J. F. Gerbeau, *Parameter identification for a one-dimensional blood flow model*, in Proceedings of Cemracs 2004, ESAIM, Les Ulis, Vol. 14, 2005, pp. 174–200.

[12] S. A. Nazarov and K. Pileckas, *Asymptotic conditions at infinity for the Stokes and Navier-Stokes problems in domains with cylindrical outlets to inifnity*, in Advances in Fluid Dynamics, P. Maremonti, ed., Quaderni di Matematica Seconda Università di Napoli, Aracne ed. Caserta, 1999.

[13] A. Quarteroni and A. Veneziani, *Analysis of a geometrical multiscale model based on the coupling of ODE and PDE for blood flow simulations*, Multiscale Model. Simul., 1 (2003), pp. 173–195.

[14] A. Veneziani, *Boundary conditions for blood flow problems*, in Proceedings of ENUMATH, Rannacher et al. eds., World Scientific Publishing, River Edge, NJ, 1998.

[15] A. Veneziani and C. Vergara, *Flow rate defective boundary conditions in haemodinamics simulations*, Internat. J. Numer. Methods Fluids, 47 (2005), pp. 803–816.

[16] A. Veneziani and C. Vergara, *An approximate method for solving incompressible Navier-Stokes problem with flow rate conditions*, Comput. Methods Appl. Mech. Engrg., 196 (2007), pp. 1685–1700.

[17] C. Vergara, *Numerical Modeling of Defective Boundary Problems in Incompressible Fluid-Dynamics—Applications to Computational Haemodynamics*, Ph.D. thesis, Politecnico di Milano, 2006; available online from http://mox.polimi.it.

SIAM J. NUMER. ANAL.
Vol. 46, No. 6, pp. 2795–2817
© 2008 Society for Industrial and Applied Mathematics

LEAPFROG FOR OPTIMAL CONTROL*

C. YALÇIN KAYA† AND J. LYLE NOAKES‡

Abstract. The leapfrog algorithm, so called because of its geometric nature, for solving a class of optimal control problems is proposed. Initially a feasible trajectory is given and subdivided into smaller pieces. In each subdivision, with the assumption that local optimal controls can easily be calculated, a piecewise-optimal trajectory is obtained. Then the junctions of these smaller pieces of optimal control trajectories are updated through a scheme of midpoint maps. Under some broad assumptions the sequence of trajectories is shown to converge to a trajectory that satisfies the maximum principle. The main advantages of the leapfrog algorithm are that (i) it does not need an initial guess for the costates and (ii) the piecewise-optimal trajectory generated in each iteration is feasible. These are illustrated through a numerical implementation of leapfrog on a problem involving the van der Pol system.

Key words. optimal control, two-point boundary-value problem, multiple shooting, geodesics, numerical methods, van der Pol system

AMS subject classifications. 49J15, 34B10, 65L10, 53C22

DOI. 10.1137/060675034

1. Introduction.

Computational techniques for finding an optimal control solution of a nonlinear system can be classified into two major categories, namely, *direct* and *indirect* methods [37]. In a direct method, the problem is discretized and the resulting (usually large scale) finite-dimensional optimization problem is solved by using nonlinear programming techniques [4, 36, 6, 3, 19]. The solution obtained this way is an approximation of the true solution, unless the control function is of a special class, for example, of the bang-bang type [18, 14, 28]. On the other hand, in an indirect method, one tackles the two-point boundary-value problem (TPBVP) arising from the Pontryagin maximum principle (PMP). Accurate solutions can be obtained through indirect methods; however, available solution techniques for TPBVPs often face serious convergence difficulties, mainly because of the lack of a good initial guess. If the set of controls is bounded, the problem is even more difficult.

The leapfrog algorithm for optimal control, so called because of its geometric nature, is an extension of a similar iterative algorithm given by Noakes [30] for finding a geodesic joining two given points on a Riemannian manifold. The main advantages of leapfrog are that a good initial guess is not needed and the trajectories generated in each iteration are feasible. Under fairly general conditions, the algorithm is convergent to a trajectory that satisfies the PMP. The algorithm presented here, as well as that given in [30], can be viewed as a continuous-time analogue of the so-called two-stage approximation technique described by Zuo [39] for discrete-time optimal control.

Methods that are used for the solution of TPBVPs arising in optimal control usually incorporate some combination of (a) multiple shooting, (b) collocation methods,

*Received by the editors November 16, 2006; accepted for publication (in revised form) March 7, 2008; published electronically July 3, 2008.

http://www.siam.org/journals/sinum/46-6/67503.html

†School of Mathematics and Statistics, University of South Australia, Mawson Lakes, S.A. 5095, Australia (yalcin.kaya@unisa.edu.au). The author gratefully acknowledges support by a fellowship from CAPES, Ministry of Education, Brazil (grant 0138-11/04), for his visit to Department of Systems and Computing at the Federal University of Rio de Janeiro, where a part of this research was carried out.

‡School of Mathematics and Statistics, The University of Western Australia, Nedlands, W.A. 6009, Australia (lyle@maths.uwa.edu.au).

and (c) continuation (homotopy) methods. Various expositions of multiple shooting can be found in [20, 32, 2, 35, 31, 11, 10, 26, 25, 23, 22]. Multiple shooting is a very accurate method but does still require good initial guesses in the subintervals taken. As the number of subdivisions is increased in order to exploit simple shooting more efficiently in each subinterval, the number of variables to be determined is substantially increased. With these problems in mind, multiple shooting is sometimes used in conjunction with some other method, for example, the method of collocation [38]. The method of collocation is used to obtain an approximate solution, in terms of cubic splines, for example, which is then fed into multiple shooting to achieve the required solution accurately. This twofold procedure may eventually need long computational times. The continuation method is also a method used in conjunction with multiple shooting [9, 7]. Usually, the nonlinear system is partitioned into its linear and nonlinear parts, and, by using the initial linear system solution and continuation techniques, the nonlinear system's solution is obtained. This method can also take long computational times depending on the nonlinearity of the system.

The leapfrog algorithm resembles multiple shooting in that the time interval is subdivided and local optimal controls are found separately over each subinterval. A significant difference is that leapfrog uses overlapping subintervals. Because of the use of subdivisions the algorithm is not expected to be troubled by the nonlinear dynamics. Three apparent advantages of the leapfrog algorithm over multiple shooting can be listed:

- Once a feasible trajectory is constructed, affine approximation of the subproblem in a subdivision (between nearby points) provides a good initial guess needed for simple shooting in that subdivision. So leapfrog does not depend critically on the supply of good initial guesses along the feasible trajectory.
- State trajectories generated by leapfrog are feasible at each step of the iteration, which is not in general the case in multiple shooting. Continuing improvements are obtained on the feasible trajectories, resulting in suboptimal trajectories. This may in particular be useful in real-time (online) applications, where time allowed to compute a solution is restricted.
- In each leapfrog iteration the number of subdivisions can be adjusted along the feasible trajectory. This allows one to eventually reduce the number of subdivisions as the solution is approached.

A drawback of leapfrog is that, as the solution is approached, convergence of leapfrog slows down; therefore the number of subdivisions q has to be progressively reduced to 2 (which is the case of simple shooting) so as to obtain a solution in a reasonable number of iterations. The effort to see, in each iteration, whether a reduction in q is possible or not incurs an additional computational cost.

Although leapfrog needs an initial feasible trajectory, in an implementation one can as well start with a set of nearby subdivision points through which concatenated local optimal solutions are obtained. Generally speaking, feasible trajectories can be established much more easily than optimal ones.

We list below some additional features of the leapfrog algorithm.

- Parallel computing is one of the efficiency concerns in the TPBVP literature (see [1, 29, 33]). Multiple shooting is particularly suited to doing computations in each subdivision independently, and for this reason it is sometimes referred to as *parallel shooting*. By using a slight variation of a leapfrog iteration, those subproblems whose time intervals are not overlapping can be solved independently. Therefore leapfrog is also suited to parallel computations.

- Suppose that an optimal control problem is given with n state equations, where the initial and terminal states are prescribed and the terminal time is fixed. For finding a local optimal control in a subdivision, the leapfrog algorithm updates only $2n$ real variables (states and costates) at a time. This amounts to a total of $[(2q-4)n]$ updates in one leapfrog iteration. With q subdivisions, multiple shooting updates $[(2q-1)n]$ variables at a time in one iteration.
- After a subproblem is solved, if two or more subdivision points (i.e., some of the junctions of the piecewise-optimal trajectories) become the same, they are referred to as *multiple points*. Multiple points occur rarely in practice. However, if they occur, leapfrog can eliminate them, yielding fewer subdivisions to deal with along a (piecewise-optimal) feasible trajectory. Elimination of multiple points does not have a clear advantage with multiple shooting.

In this paper, we show that the leapfrog algorithm converges under some general assumptions, such as the uniqueness of the local optimal control and costate variables in a given subdivision. In [15, 16], we gave some background and a survey of ideas leading to the development of leapfrog for solving the geodesic problem. We also reported an implementation of leapfrog (with fixed terminal time) for optimal control, without elaboration. In the present paper, we provide a careful and detailed account of a numerical implementation of leapfrog. We also gave some preliminaries of the theory in an earlier publication [17], which did not have the detailed setting and approach of the present article.

The present paper is organized as follows. In section 2 we state the optimal control problem and give definitions and assumptions, verifications of which are provided through example problems. In section 3 we describe the leapfrog algorithm. In section 4 we address multiple points and provide further definitions and facts. We prove convergence of the leapfrog algorithm in section 5. A computational illustration of the algorithm is given in section 6 by solving a problem involving the van der Pol system.

2. Preliminaries. Consider the control system

$$\dot{x}(t) = f(x(t), u(t)), \tag{1}$$

where the state $x(t) \in \mathbb{R}^n$ and the control $u : \mathbb{R} \longrightarrow \mathbb{R}^m$ is an arbitrary piecewise-continuous function, which, as described in [34], has discontinuity at finitely many points and has finite right- and left-hand limits at the points of discontinuity. The function $f : \mathbb{R}^n \times \mathbb{R}^m \longrightarrow \mathbb{R}^n$ is C^1 in x and u.

The following optimal control problem is considered:

$$\mathrm{P} : \begin{cases} \text{minimize} & \displaystyle\int_{t_0}^{t_f} f_0(x(t), u(t))\, dt \\ \text{subject to} & \dot{x}(t) = f(x(t), u(t)), \\ & x(t_0) = x_0 \quad \text{and} \quad x(t_f) = x_f, \end{cases}$$

where the function $f_0 : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is C^1 and nonnegative and the final time $t_f > t_0$ is free. A *state trajectory* corresponding to a control $u(\cdot)$ is a continuous curve $x(\cdot)$ solving $\dot{x}(t) = f(x(t), u(t))$ for almost all t. An *optimal state trajectory* $x(\cdot)$ and a corresponding *optimal control* $u(\cdot)$ are those which solve problem (P).

The PMP gives necessary conditions of optimality [34] for large classes of problems including problem (P). In these problems the so-called Hamiltonian $H : \mathbb{R}^n \times \mathbb{R}^n \times$

$\mathbb{R}^m \longrightarrow \mathbb{R}$ is defined as

$$H(x, \psi, u) = \psi_0 \, f_0(x, u) + \langle \psi, f(x, u) \rangle,$$

where $\psi_0(t) \in \mathbb{R}$ and $\psi(t) \in \mathbb{R}^n$ are the adjoint or costate variables and $\langle \cdot, \cdot \rangle$ is the Euclidean inner product. The PMP states that if a state trajectory $x(\cdot)$ and a corresponding control $u(\cdot)$ are optimal, then there exists a nontrivial pair $(\psi_0, \psi) : [t_0, t_f] \longrightarrow \mathbb{R}^{n+1}$ such that, for almost every $t \in [t_0, t_f]$,

$$\dot{x} = \partial H / \partial \psi = f(x, u), \tag{2}$$

$$\dot{\psi} = -\partial H / \partial x, \tag{3}$$

$$H(x, \psi, u) = \min_{v \in \mathcal{U}} H(x, \psi, v), \tag{4}$$

$$H(x, \psi, u) \equiv 0, \tag{5}$$

with $x(t_0) = x_0$ and $x(t_f) = x_f$. Furthermore PMP requires that $\psi_0(t) \equiv c$, where c is some nonnegative constant. The costate variable ψ_0 accounts for the case when cost is treated as an additional state coordinate. Problems that yield $c = 0$ are referred to as *abnormal* in the literature, for which the necessary conditions in (2)–(5) are independent of the cost and therefore insufficiently informative. A discussion of the necessary conditions of optimality for abnormal trajectories can be found in [24] and the references therein.

ASSUMPTION 1. *Problem* (P) *is normal. Without loss we set* $\psi_0 = 1$.

We refer to the conditions given in (2)–(5) as the *optimality system*. A trajectory of the optimality system corresponding to a control $u(\cdot)$ is a continuous pair of curves $(x(\cdot), \psi(\cdot))$ which solve (2)–(5). If the trajectory pair $(x(\cdot), \psi(\cdot))$ corresponding to control $u(\cdot)$ satisfies the optimality system, then it is said to be a *critical trajectory* (*pair*). We will refer to the trajectory $(x(\cdot), \psi(\cdot))$ as optimal if the state trajectory $x(\cdot)$ along with a corresponding control $u(\cdot)$ is optimal. Note that a critical trajectory is not necessarily optimal.

Remark 2.1. Condition (5) stands because the terminal time t_f is free—t_f is an unknown of the problem. In the case when t_f is prescribed, however, condition (5) is not part of the optimality system anymore. In this case one has $H(x, \psi, u) \equiv a$, where a is any real constant, but this condition is usually not needed for computations, resulting in a simpler optimality system.

Remark 2.2. Although leapfrog is described for free terminal time problems, we can also tackle problems with a fixed terminal time. Given a problem with fixed t_f, say, $t_f = 1$, we can define a new state variable $x_{n+1}(t) = t$ such that $\dot{x}_{n+1}(t) = 1$, with the end points $x_{n+1}(0) = 0$ and $x_{n+1}(t_f) = 1$, where we now let t_f be free. The equation for x_{n+1} is trivial to solve. Therefore, in a practical implementation, there would be no need to augment the state equations in this way.

Next we describe the subproblems employed within leapfrog. Let $x(\cdot)$ be a continuous piecewise-C^1 trajectory such that $x(t_0) = x_0$ and $x(t_f) = x_f$. The point x_f is not necessarily close to x_0. The trajectory $x(\cdot)$ is simply a curve between x_0 and x_f. The aim is to find a critical trajectory between x_0 and x_f. We initially partition this feasible trajectory so as to obtain q pieces with roughly comparable costs. Let the corresponding partition times be $t_0 < t_1 < \cdots < t_{q-1} < t_q = t_f$. The number q of partitions is chosen so that the partition points $z_{i-1} := x(t_{i-1})$ and $z_i := x(t_i)$, $i = 1, \ldots, q$, are sufficiently close to each other that there exists a local optimal solution from z_{i-1} to z_{i+1}, $i = 1, \ldots, q-1$, which is easily computable. Namely, we

assume that local solutions to the following subproblems can be easily obtained:

$$\mathrm{P}_i : \begin{cases} \text{minimize} & \displaystyle\int_{t_{i-1}}^{t_{i+1}} f_0(x(t), u(t))\, dt \\ \text{subject to} & \dot{x}(t) = f(x(t), u(t)), \\ & x(t_{i-1}) = z_{i-1} \quad \text{and} \quad x(t_{i+1}) = z_{i+1}, \end{cases}$$

where t_{i-1} is fixed by the solution of problem (P_{i-1}) and t_{i+1} is free. Note that $z_0 = x_0$ and $z_q = x_f$. Through the PMP, the optimality system associated with problem (P_i) can be written as in (2)–(5) for $t \in [t_{i-1}, t_{i+1}]$, with end points $x(t_{i-1}) = z_{i-1}$ and $x(t_{i+1}) = z_{i+1}$. Solution of the optimality system provides an initial value of the costate variable, namely, $\psi(t_{i-1}) =: \lambda_{i-1}$, for problem (P_i).

Assumption 2. *There exists a.e. a (locally) unique optimal control* $u : [t_{i-1}, t_{i+1}] \longrightarrow \mathbb{R}^m$ *taking the system from* z_{i-1} *to* z_{i+1}.

Assumption 3. *The normalized costate vector* $\psi(t)$ *associated with problem* (P_i) *is determined uniquely by the locally optimal control* $u(t)$ *and the initial point* z_{i-1} *of the optimal trajectory.*

In what follows we discuss two example problems, albeit relatively simple, to illustrate Assumptions 1–3. In these problems the terminal time is fixed: As pointed out earlier they can easily be converted into an equivalent free terminal time form used for leapfrog.

Example 1 (geodesics). Let $\{g_{ij} : 1 \le i, j \le n\}$ define a Riemannian metric on an open subset U of $\mathbb{R}^n$, namely, the matrix $[g_{ij}(x)]$ is symmetric positive definite, and C^∞ in x. The problem of finding geodesics between two given points can be formulated as an optimal control problem as follows:

$$\mathrm{P}_g : \begin{cases} \text{minimize} & \displaystyle\int_0^1 \sum_{i,j} g_{ij}(x) u_i u_j \, dt \\ \text{subject to} & \dot{x} = u, \\ & x(0) = x_0 \quad \text{and} \quad x(1) = x_f \ , \end{cases}$$

where the initial and terminal points x_0 and x_f are in $\mathbb{R}^n$. The control $u(t)$ is also in $\mathbb{R}^n$. The Hamiltonian for the problem is

$$H(x, u, \psi) = \psi_0 \sum_{i,j} g_{ij}(x)\, u_i u_j + \langle \psi, u \rangle.$$

Condition (4) in this case implies that $\partial H / \partial u_i = 0$, $i = 1, \ldots, n$, which gives

$$\psi_i = -2\, \psi_0 \sum_j g_{ij}(x)\, u_j \ . \tag{6}$$

Suppose that $\psi_0 = 0$. Then $\psi_i = 0$, $i = 1, \ldots, n$, i.e., (ψ_0, ψ) is trivial, which is a contradiction. This verifies Assumption 1. Set $\psi_0 = 1$ from this point onwards.

Given u_i, $i = 1, \ldots, n$, ψ_i are determined uniquely from (6). By differentiating ψ_i one gets

$$\dot{\psi}_i = -2 \left(\sum_{j,k} \frac{\partial g_{ij}}{\partial x_k} u_j u_k + \sum_j g_{ij}(x)\, \dot{u}_j \right) \ .$$

On the other hand, another condition (3) of PMP yields

$$\dot{\psi}_i = -\sum_{j,k} \frac{\partial g_{jk}}{\partial x_i} u_j u_k.$$

By equating the two expressions for $\dot{\psi}_i$ and rearranging,

$$\dot{u}_i = -\sum_{j,k} \Gamma^i_{jk}(x)\, u_j u_k, \tag{7}$$

where

$$\Gamma^i_{jk}(x) = \frac{1}{2} \sum_s g^{is}(x) \left(2\frac{\partial g_{sj}}{\partial x_k} - \frac{\partial g_{jk}}{\partial x_s} \right).$$

In the above equations Γ^i_{jk} are the so-called Christoffel symbols [5], and g^{is} is the isth entry of the inverse of the matrix $[g_{is}]$. A direct substitution from the state equations give the so-called geodesic equation which has a unique solution for x_f sufficiently close to x_0:

$$\ddot{x}_i = -\sum_{j,k} \Gamma^i_{jk}(x)\, \dot{x}_j \dot{x}_k \ . \tag{8}$$

Then locally the control u has a unique solution, which verifies Assumption 2; in turn, Assumption 3 is verified by (6).

Example 2 (LTI systems). Consider the minimum energy control of a linear time-invariant (LTI) system:

$$\mathrm{P}_{LTI} : \begin{cases} \text{minimize} & \dfrac{1}{2}\displaystyle\int_0^1 u^T R u \, dt \\ \text{subject to} & \dot{x}(t) = Ax(t) + Bu(t), \\ & x(0) = x_0 \quad \text{and} \quad x(1) = x_f \ , \end{cases}$$

where $x(t) \in \mathbb{R}^n$, $u : \mathbb{R} \longrightarrow \mathbb{R}^m$, and the constant matrices A and B have appropriate dimensions. The constant matrix R is symmetric and positive definite. The system is assumed to be controllable. The Hamiltonian is

$$H(x, \psi, u) = \frac{1}{2}\, \psi_0\, u^T R u + \psi^T (Ax + Bu) \ .$$

The condition $\partial H / \partial u = 0$ gives

$$\psi_0\, Ru = -B^T \psi. \tag{9}$$

Furthermore (3) of PMP yields

$$\dot{\psi}^T = -\psi^T A \ . \tag{10}$$

By differentiating (9) up to $(n-1)$ times and using (10) one gets

$$\begin{aligned} \psi^T B &= \psi_0\, Ru, \\ \psi^T AB &= \psi_0\, R\dot{u}, \\ &\cdots \\ \psi^T A^{n-1} B &= \psi_0\, Ru^{(n-1)}, \end{aligned} \tag{11}$$

where $u^{(i)} = d^i u / dt^i$, $i = 1, \ldots, n-1$, with $\dot{u} = u^{(1)}$.

Suppose that $\psi_0 = 0$. Since the system is controllable, the controllability matrix $[B|AB|\dots|A^{n-1}B]$ is of rank n, so $\psi = \mathbf{0}$, which contradicts the nontriviality of (ψ_0, ψ) in PMP. This verifies Assumption 1.

We set $\psi_0 = 1$. Again because the system is controllable, ψ can be expressed uniquely from (11) in terms of u and its derivatives. This verifies Assumption 3.

It can be shown (by using $R = R^{1/2}R^{1/2}$ and by considering a generalization of the controllability Gramian in Exercise 9.2-11 of [13]) that the optimal control $u(t)$ is given by the expression

$$u(t) = -R^{-1}B^T e^{-A^T t} M^{-1} \left(x(0) - e^{-A}x(1)\right), \tag{12}$$

where

$$M = \int_0^1 e^{-sA} B R^{-1} B^T e^{-sA^T} ds \ ,$$

and the nonsingularity of M is furnished by the controllability of the system. As can be seen, the optimal control $u(t)$ is unique by the choice of $x(0)$ and $x(1)$, which verifies Assumption 2.

A point $\widetilde{x}$ is said to be *reachable* from a point x_0 with cost less than $\Delta > 0$ if and only if, for some admissible control, there exists a trajectory $x(\cdot)$ defined on an interval $[t_0, t]$, $t > t_0$, such that $x(t_0) = x_0$, $x(t) = \widetilde{x}$, and $\int_{t_0}^{t_f} f_0(x(t), u(t))\, dt < \Delta$. The trajectory $x(\cdot)$ of concern here does not necessarily satisfy PMP; namely, it is not necessarily critical. The set of all such points $\widetilde{x}$ is denoted by $\mathcal{R}(x_0, < \Delta)$. The reachable set $\mathcal{R}(x_0, < \Delta)$ is said to be a *small-cost reachable set* when Δ is sufficiently small.

Let D denote the set of all ordered pairs (z_{i-1}, z_{i+1}) such that z_{i+1} can be reached from z_{i-1} with a cost less than 2δ, namely,

$$D = \{(z_{i-1}, z_{i+1}) \in \mathbb{R}^n \times \mathbb{R}^n : z_{i+1} \in \mathcal{R}(z_{i-1}, < 2\delta)\} \ , \tag{13}$$

where $\delta > 0$ is so chosen that there exists a local optimal control from z_{i-1} to z_{i+1}, which is easily computable. Namely, the choice of δ ensures that the points z_{i-1} and z_{i+1} are nearby so that affine approximations and the simple shooting method can be used to get a local solution for problem (P_i) efficiently.

Let $\tau : D \longrightarrow \mathbb{R}$ denote the *local cost function* such that, given $(z_{i-1},\ z_{i+1})$ in D, $\tau(z_{i-1}, z_{i+1})$ is the infimum of the cost to get from z_{i-1} to z_{i+1}. Note that $\tau(z_{i-1}, z_{i+1}) < 2\delta$.

Clearly both D and τ depend on δ; however, for brevity, we will not show this dependence explicitly. The associated value of δ should be clear from the context.

3. Leapfrog algorithm. In this section we give a description of the leapfrog algorithm. Suppose that a feasible trajectory $\mu_z : [t_0, t_f] \longrightarrow \mathbb{R}^n$ is given between the initial and terminal states $z_0 = \mu_z(t_0)$ and $z_q = \mu_z(t_f)$, which is divided into q trajectory segments, such that $z_i = \mu_z(t_i)$, $i = 1, \dots, q-1$. The junction points of μ_z are represented by the $(q+1)$-tuple $z = (z_0, z_1, z_2, \dots, z_q)$. Through the leapfrog scheme, while the initial and terminal states z_0 and z_q, and t_0, are kept fixed, each z_i, $i = 1, \dots, q-1$, and t_i, $i = 1, \dots, q$, are adjusted in the following fashion.

For $(z_{i-1}, z_{i+1}) \in D$, let $\gamma_{z_{i-1}, z_{i+1}} : [t_{i-1}, t_{i+1}] \longrightarrow \mathbb{R}^n$ be the unique *local optimal trajectory* (by Assumption 2) such that $\gamma_{z_{i-1}, z_{i+1}}(t_{i-1}) = z_{i-1}$ and $\gamma_{z_{i-1}, z_{i+1}}(t_{i+1}) = z_{i+1}$. The *midpoint* $\widehat{z}_i := \gamma_{z_{i-1}, z_{i+1}}(\widehat{t}_i)$, $t_{i-1} \le \widehat{t}_i \le t_{i+1}$, between z_{i-1} and z_{i+1} is

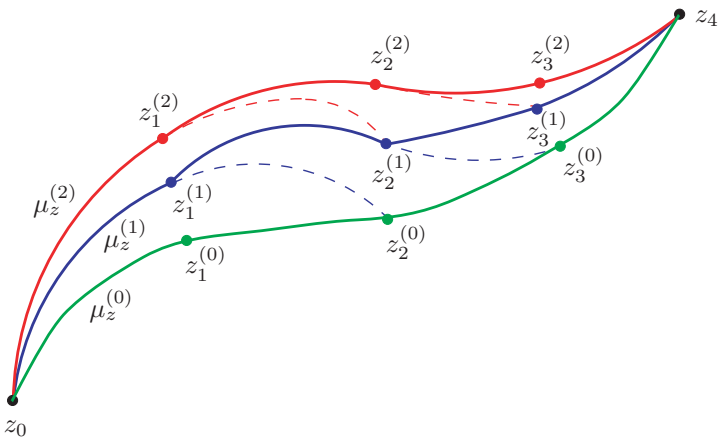


FIG. 1. *An illustration of the leapfrog algorithm in state space.*

defined by $\tau(z_{i-1},\widehat{z}_i) = \tau(\widehat{z}_i, z_{i+1}) = \tau(z_{i-1}, z_{i+1})/2$. For convenience, we also set $\widehat{t}_0 := t_0$ and $\widehat{t}_q := t_q$. The point z_i is moved to the midpoint $\widehat{z}_i$, while the time t_i is updated as $\widehat{t}_i$.

Moving z_i onto an optimal trajectory from z_{i-1} to z_{i+1} (obtained by solving problem (P_i)) achieves locally the largest possible decrease in cost while keeping other junction points fixed. The point where z_i goes in the optimal trajectory segment is taken to be the point such that the cost of getting from z_{i-1} to that point is half the cost of getting from z_{i-1} to z_{i+1}. However, somewhat different positioning of z_i is also permissible.

The leapfrog algorithm consists of updating z_i and t_i in this way infinitely often where i cycles through $1,\ldots,q-1$. This generates a sequence

$$\mathcal{M}_z = \{\mu_z^{(k)} : [t_0, t_f^{(k)}] \longrightarrow \mathbb{R}^n : k > 0\}$$

of piecewise-optimal (state) trajectories. Along each consecutive trajectory $\mu_z^{(k)}$ the cost decreases. This decrease precludes convergence to a global maximum. A schematic illustration of the leapfrog algorithm is given in state space in Figure 1. The initial set of points $z_i^{(0)}$, $i = 1, 2, 3$, are the junctions chosen on an initial feasible trajectory. Two leapfrog iterations are depicted. The piecewise-optimal trajectory $\mu_z^{(1)}$ is obtained after the first iteration from the initial feasible trajectory. The second iteration yields the piecewise-optimal trajectory $\mu_z^{(2)}$.

Let us outline the leapfrog algorithm by using the definitions we have just given.

Step 0. Choose the number of subdivisions $q > 0$ and final time $t_f > t_0$.

Take an initial feasible trajectory $\mu_z^{(0)}$ given between $z_0 = x_0$ and $z_q = x_f$.

Subdivide $\mu_z^{(0)}$ into q segments with subdivision points

$$z_0^{(0)},\ z_1^{(0)}, \ldots, z_{q-1}^{(0)},\ z_q^{(0)}\ .$$

Set $t_0^{(0)} = t_0$ and $k = 0$.

Step k. Perform Steps $k.1$ and $k.2$ below for $i = 1, 2, \ldots, q-1$.

Step $k.1$. Given $z_{i-1}^{(k)}$, $z_{i+1}^{(k)}$ and $t_{i-1}^{(k)}$: Solve problem (P_i) with $x(t_{i-1}) = z_{i-1}^{(k)}$, $x(t_{i+1}) = z_{i+1}^{(k)}$, and $t_{i-1} = t_{i-1}^{(k)}$.

Step k.2. Let $z_i^{k+1} = x(t_i)$ and $t_i^{k+1} = t_i$ such that

$$\tau(x(t_{i-1}), x(t_i)) = \frac{1}{2}\,\tau(x(t_{i-1}), x(t_{i+1}))\ .$$

Set $k = k + 1$ and repeat Step k.

For proving convergence to a critical trajectory, it is necessary to consider the local critical trajectory pairs $(x(\cdot), \psi(\cdot))$ obtained by solving the subproblems (P_i). For each update of z_i, an update is also obtained for $\lambda_{i-1} := \psi(t_{i-1})$ and t_i, $i = 1, \ldots, q-1$, through the solution of (P_i) and the associated optimality system. The updates for λ_{q-1}, λ_q, and $t_q = t_f$ are given by the trajectories of the optimality system associated with problem (P_{q-1}).

Now let $y_i = (z_i, \lambda_i) = (x(t_i), \psi(t_i))$, and consider the $(q+1)$-tuple $y = (y_0, y_1, y_2, \ldots, y_q)$ and the sequence

$$\mathcal{M} = \{\mu^{(k)} : [t_0, t_f^{(k)}] \longrightarrow \mathbb{R}^{2n} : k > 0\}\ ,$$

where $\mu^{(k)}(\cdot) = (x^{(k)}(\cdot), \psi^{(k)}(\cdot))$, with $\mu^{(k)}(t_0) = (z_0, \lambda_0^{(k)})$ and $\mu^{(k)}(t_q) = (z_q, \lambda_q^{(k)})$. In each iteration of leapfrog the states $x(t_0) = z_0$ and $x(t_f) = z_q$ at the end points of $\mu^{(k)}$ are fixed, but the costates $\psi(t_0) = \lambda_0$ and $\psi(t_f) = \lambda_q$, as well as the terminal time t_f, are updated when the new local optimal controls are found and concatenated. Under the assumptions we pose, notably under the uniqueness of the local optimal control and costates, the sequence $\{\mu^{(k)} : k > 0\}$ will be shown to converge to a critical trajectory.

4. Further definitions and facts.

4.1. Elimination of multiple points. If two or more of z_i become the same point, then they are referred to as *multiple points*. Clearly there are no multiple points along the initial trajectory, because the trajectory is not self-intersecting and along each trajectory segment the cost is positive. Multiple points may occur, albeit rarely, in leapfrog, but they can be eliminated through a relabeling procedure described below.

Suppose that there are $\tilde{q}$ subdivisions. Fix i, $0 < i < \tilde{q}$. Let $z_i^{(k+1)}$ be the midpoint along the local optimal trajectory from $z_{i-1}^{(k+1)}$ to $z_{i+1}^{(k)}$.

(M1) If $z_i^{(k+1)} = z_\ell^{(k+1)}$ for some ℓ where $0 \le \ell < i$, then relabel $z_r^{(k)}$, $r = i+1, \ldots, \tilde{q}$, such that

$$z_{i+s}^{(k)} = z_{\ell+s}^{(k)}, \quad s = 1, \ldots, \tilde{q} - i.$$

(M2) If $z_i^{(k+1)} = z_\ell^{(k)}$ for some ℓ where $i < \ell \le \tilde{q}$, then relabel $z_r^{(k)}$, $r = \ell, \ldots, \tilde{q}$, such that

$$z_{\ell+s}^{(k)} = z_{i+s}^{(k)}, \quad s = 0, 1, \ldots, \tilde{q} - \ell.$$

Note that if (M1) applies, the earlier piecewise-optimal trajectory from $z_\ell^{(k+1)}$ to $z_{i+1}^{(k)}$ is replaced by the new local optimal trajectory from $z_i^{(k+1)}$ to $z_{i+1}^{(k)}$. If (M2) applies, the earlier piecewise-optimal trajectory from $z_{i-1}^{(k+1)}$ to $z_\ell^{(k)}$ is replaced by the new local optimal trajectory from $z_{i-1}^{(k+1)}$ to $z_i^{(k)}$. In either case the number of subdivisions $\tilde{q}$ is reduced by $|\ell - i|$. The removal of trajectory segments also reduces total cost.

Either (M1) or (M2) is repeated as often as there are multiple points. Because $\tilde{q}$ is finite, the elimination procedure is applied only a finite number of times, until no more multiple points occur in the remaining leapfrog iterations, yielding a constant number of subdivisions q such that $1 \leq q \leq \tilde{q}$.

It can be shown as a special case that in steps (M1) and (M2) one cannot have $\ell = i - 1$ or $i + 1$. We state this observation as a lemma below.

LEMMA 4.1. *$z_i^{(k+1)} \neq z_{i-1}^{(k+1)}$ and $z_i^{(k+1)} \neq z_{i+1}^{(k)}$.*

Proof. Suppose that $z_i^{(k+1)} = z_{i-1}^{(k+1)}$. Then there exist two local optimal trajectories from $z_{i-1}^{(k+1)}$ to $z_{i+1}^{(k)}$: the local optimal trajectory emanating from $z_{i-1}^{(k+1)}$, revisiting $z_{i-1}^{(k+1)} = z_i^{(k+1)}$, and terminating at $z_{i+1}^{(k)}$ and the segment of the same trajectory after revisiting $z_{i-1}^{(k+1)}$. This contradicts the uniqueness assumption of the local optimal trajectories. Therefore $z_i^{(k+1)} \neq z_{i-1}^{(k+1)}$. Suppose that $z_i^{(k+1)} = z_{i+1}^{(k)}$. A similar argument leads to a similar contradiction. Hence $z_i^{(k+1)} \neq z_{i+1}^{(k)}$. ∎

It should be noted that, in practice, sufficiently nearby z_i may be treated as multiple points. This is highly desirable because it would result in a reduction in both $\tilde{q}$ and cost along the current leapfrog trajectory.

In the sections to follow we give more tools and assumptions for the convergence analysis of leapfrog with constant q after the elimination process described above is completed.

4.2. Continuity of the local cost function. The following assumption will be used in proving the continuity of τ stated in Lemma 4.2, which will in turn be useful in the convergence analysis.

ASSUMPTION 4. *Given $\epsilon > 0$, there exists $\kappa > 0$ such that, for any $z_i, \widetilde{z}_i \in \mathbb{R}^n$,*

$$\|z_i - \widetilde{z}_i\| < \kappa \implies \tau(z_i, \widetilde{z}_i) < \epsilon \ .$$

Example 3. Recall the optimal control problem with the LTI system in Example 2. With $x_0 = z_i$ and $x(1) = \widetilde{z}_i$, it can be shown that

$$\tau(z_i, \widetilde{z}_i) = \frac{1}{2} \left(z_i - e^{-A}\widetilde{z}_i\right)^T M^{-1} \left(z_i - e^{-A}\widetilde{z}_i\right) \ ,$$

where

$$M = \int_0^1 e^{-sA} B R^{-1} B^T e^{-sA^T} ds \ .$$

Because the LTI system is controllable, the Gramian M is positive definite. Therefore $\tau(z_i, \widetilde{z}_i) < \epsilon$ defines a region enclosed by an ellipsoid containing z_i, and thus some $\kappa > 0$ can be found to enclose an open ball of radius $\kappa > 0$ centered at z_i within the ellipsoid. This justifies Assumption 4. In fact, one can conclude from the above expression that the local cost function τ is continuous.

LEMMA 4.2. *The local cost function τ is continuous.*

Proof. Consider the set D given in (13) with some $\delta > 0$ and the open ball $B_\kappa(z_{i-1}, z_{i+1})$ with radius $\kappa > 0$ centered at $(z_{i-1}, z_{i+1}) \in D$, namely,

$$B_\kappa(z_{i-1}, z_{i+1}) = \{(\widetilde{z}_{i-1}, \widetilde{z}_{i+1}) \in D : \|(z_{i-1}, z_{i+1}) - (\widetilde{z}_{i-1}, \widetilde{z}_{i+1})\| < \kappa\}, \tag{14}$$

where $\|\cdot\|$ is the Euclidean norm. The inequality in (14) implies that

$$\|z_{i-1} - \widetilde{z}_{i-1}\| < \kappa \qquad \text{and} \qquad \|z_{i+1} - \widetilde{z}_{i+1}\| < \kappa \ .$$

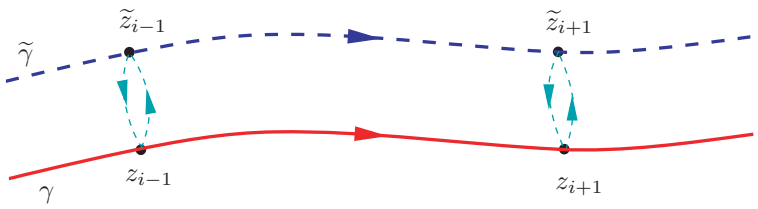


FIG. 2. *Two neighboring optimal trajectories γ and $\widetilde{\gamma}$.*

Because $(z_{i-1}, z_{i+1}) \in D$, there is a local optimal trajectory γ passing through z_{i-1} and z_{i+1}. Similarly because $(\widetilde{z}_{i-1}, \widetilde{z}_{i+1}) \in D$, one has a local optimal trajectory $\widetilde{\gamma}$ passing through $\widetilde{z}_{i-1}$ and $\widetilde{z}_{i+1}$. The trajectories γ and $\widetilde{\gamma}$ are illustrated in Figure 2.

By Assumption 4, for any $\epsilon > 0$, there exist $\kappa_1 > 0$ and $\kappa_2 > 0$ such that

$$\|z_i - \widetilde{z}_i\| < \kappa_1 \implies \tau(z_i, \widetilde{z}_i) < \epsilon/2$$

and

$$\|\widetilde{z}_i - z_i\| < \kappa_2 \implies \tau(\widetilde{z}_i, z_i) < \epsilon/2 \ .$$

Let $\kappa = \min\{\kappa_1, \kappa_2\}$. Then, for any $\epsilon > 0$,

$$\|z_i - \widetilde{z}_i\| < \kappa \implies \tau(z_i, \widetilde{z}_i) < \epsilon/2 \quad \text{and} \quad \tau(\widetilde{z}_i, z_i) < \epsilon/2 \ .$$

This statement is depicted in Figure 2 for the pairs $(z_{i-1}, \widetilde{z}_{i-1})$ and $(z_{i+1}, \widetilde{z}_{i+1})$. Because γ is optimal,

$$\begin{aligned} \tau(z_{i-1}, z_{i+1}) &\le \tau(z_{i-1}, \widetilde{z}_{i-1}) + \tau(\widetilde{z}_{i-1}, \widetilde{z}_{i+1}) + \tau(\widetilde{z}_{i+1}, z_{i+1}) \\ &\le \tau(\widetilde{z}_{i-1}, \widetilde{z}_{i+1}) + \epsilon \ . \end{aligned} \tag{15}$$

Similarly, because $\widetilde{\gamma}$ is optimal,

$$\begin{aligned} \tau(\widetilde{z}_{i-1}, \widetilde{z}_{i+1}) &\le \tau(\widetilde{z}_{i-1}, z_{i-1}) + \tau(z_{i-1}, z_{i+1}) + \tau(z_{i+1}, \widetilde{z}_{i+1}) \\ &\le \tau(z_{i-1}, z_{i+1}) + \epsilon \ . \end{aligned} \tag{16}$$

Inequalities (15)–(16) furnish continuity of τ. ∎

4.3. Midpoint maps and total cost. Consider $(z_{i-1}, z_{i+1}) \in D$, where D is as in (13). We define a *midpoint map* to be any $M : \mathbb{R}^n \times \mathbb{R}^n \longrightarrow \mathbb{R}^n$ such that $M(z_{i-1}, z_{i+1}) = \widehat{z}_i$, where $\widehat{z}_i$ is the midpoint between z_{i-1} and z_{i+1}.

Let $X \subseteq \mathbb{R}^{2n}$ and Y be the set of all $(q+1)$-tuples $y = (y_0, y_1, \ldots, y_q) \in X^{q+1}$ such that $y_i = (z_i, \lambda_i)$ and $\tau(z_i, z_{i+1}) \le \delta$ for all $i = 0, 1, \ldots, q-1$.

For $1 \le p < q$ define $G_p : Y \longrightarrow X^{q+1}$ by

$$G_p(y) = (y_0, \ldots, y_{p-1}, \widehat{y}_p, y_{p+1}, \ldots, y_q),$$

where $\widehat{y}_p = (\widehat{z}_p, \widehat{\lambda}_p)$, $\widehat{z}_p = M(z_{p-1}, z_{p+1})$, and $\widehat{\lambda}_p$ is given by one of the two cases below.

(i) For $p = 0, 1, \ldots, (q-3)$, $\widehat{\lambda}_p = \psi(\widehat{t}_p)$, the initial costate associated with problem (P_{p+1}).

(ii) For $p = (q-2), (q-1), q$, $\widehat{\lambda}_p = \psi(\widehat{t}_p)$, costates associated with problem (P_{q-1}). It should be noted that the update $\widehat{y}_p$, $p = 0, 1, \ldots, (q-3)$, is obtained by solving two subproblems; namely, problem (P_p) is solved to get $\widehat{z}_p$, and problem (P_{p+1}) is solved to get $\widehat{\lambda}_p$.

LEMMA 4.3. *$G_p(y) \in Y$.*

Proof. $\tau(z_{p-1}, \widehat{z}_p) = \tau(z_{p-1}, z_{p+1})/2 \leq (\tau(z_{p-1}, z_p) + \tau(z_p, z_{p+1}))/2 \leq (2\delta)/2$. ∎

So $G_p : Y \longrightarrow Y$, where $1 \leq p < q$. Define $F : Y \longrightarrow Y$ as the composite

$$F = G_{q-1} \circ G_{q-2} \circ \cdots \circ G_1 \ .$$

The $(q+1)$-tuple $\widehat{y} = (\widehat{y}_0, \widehat{y}_1, \ldots, \widehat{y}_q) = F(y) \in Y$ can be defined alternatively by

(i) $\widehat{y}_0 = (\widehat{z}_0, \widehat{\lambda}_0)$, where $\widehat{z}_0 = z_0$ and $\widehat{\lambda}_0 = \psi(t_0)$ is given by the optimal control from z_0 to z_2;

(ii) $\widehat{y}_i = (\widehat{z}_i, \widehat{\lambda}_i)$, where $\widehat{z}_i = M(\widehat{z}_{i-1}, z_{i+1})$ and $\widehat{\lambda}_i = \psi(\widehat{t}_i)$ is given by the optimal control from $\widehat{z}_i$ to z_{i+2} for $1 \leq i \leq q-2$;

(iii) $\widehat{y}_{q-1} = (\widehat{z}_{q-1}, \widehat{\lambda}_{q-1})$, where $\widehat{z}_{q-1} = M(\widehat{z}_{q-2}, z_q)$ and $\widehat{\lambda}_{q-1} = \psi(\widehat{t}_{q-1})$ is given by the optimal control from $\widehat{z}_{q-2}$ to z_q;

(iv) $\widehat{y}_q = (\widehat{z}_q, \widehat{\psi}_q)$, where $\widehat{z}_q = z_q$ and $\widehat{\psi}_q = \psi(\widehat{t}_q)$ is given by the optimal control from $\widehat{z}_{q-2}$ to z_q.

Let Z be the set of all $(q+1)$-tuples $z = (z_0, z_1, \ldots, z_q)$. The $(q+1)$-tuple z is said to have cost $\alpha(z)$, where $\alpha : Z \longrightarrow \mathbb{R}$ is defined by

$$\alpha(z) = \sum_{i=1}^{q} \tau(z_{i-1}, z_i) \ .$$

Note that $\alpha(z) = \int_{t_0}^{t_f} f_0(x, u)\, dt$ is the cost incurred along the piecewise-optimal trajectory defined by the $(q+1)$-tuple z. The function α is continuous, because τ is continuous.

Let the local cost function $\tau : D \longrightarrow \mathbb{R}$ be extended as $\widetilde{\tau} : \widetilde{D} \longrightarrow \mathbb{R}$, where

$$\widetilde{D} = \{(x, \widetilde{x}) \in \mathbb{R}^n \times \mathbb{R}^n : \widetilde{x} \in \mathcal{R}(x, < \Delta)\}$$

and Δ is not necessarily small, such that $\widetilde{\tau}(x, \widetilde{x})$ is the infimum of the cost to get from x to $\widetilde{x}$. Then

$$\widetilde{\tau}(z_0, z_q) \leq \alpha(y) \ .$$

Define the state-space projection map $\pi_1 : Y \longrightarrow Z$ such that

$$\pi_1 \circ G_p(y) = (z_0, z_1, \ldots, \widehat{z}_p, \ldots, z_q)$$

and that $\pi_1 \circ F(y) = (\widehat{z}_0, \widehat{z}_1, \ldots, \widehat{z}_q) =: \widehat{z}$.

LEMMA 4.4. *For $1 \leq p < q$ and all $y \in Y$*

$$\alpha(\pi_1 \circ G_p(y)) \leq \alpha(z) \ .$$

Proof.

$$\begin{aligned} \alpha(z) - \alpha(\pi_1 \circ G_p(y)) &= \tau(z_{p-1}, z_p) + \tau(z_p, z_{p+1}) - \tau(z_{p-1}, \widehat{z}_p) - \tau(\widehat{z}_p, z_{p+1}) \\ &= \tau(z_{p-1}, z_p) + \tau(z_p, z_{p+1}) - \tau(z_{p-1}, z_{p+1}) \\ &\geq 0. \end{aligned}$$ ∎

If one considers the composite $F = G_{q-1} \circ G_{q-2} \circ \cdots \circ G_1$ in Lemma 4.4 instead of G_p, then the following lemma is obtained.

LEMMA 4.5. *For all* $y \in Y$

$$\alpha(\pi_1 \circ F(y)) \leq \alpha(z) \ .$$

5. Convergence analysis.

5.1. Compactness of the state-costate space. Define the set $C(x_0, r)$ of all final states $x_f = x(t_f)$ and costates $\psi_f = \psi(t_f)$ such that the states x_f can be reached from $x_0 = x(t_0)$ over $[t_0, t_f]$ with a minimum cost less than or equal to $r \geq 0$; namely,

$$C(x_0, r) = \{(x_f, \psi_f) \in \mathbb{R}^{2n} : \widetilde{\tau}(x_0, x_f) \leq r\} \ .$$

For each (x_0, x_f, ψ_f) above, the optimal trajectory pair $(x(t), \psi(t))$ and a corresponding control trajectory $u(t)$ satisfy conditions (2)–(5) with $x(t_0) = x_0$, $x(t_f) = x_f$, and $\psi(t_f) = \psi_f$.

Let α_0 be the cost incurred along the initial feasible trajectory.

ASSUMPTION 5. *The set* $C(x_0, \alpha_0)$ *is bounded.*

ASSUMPTION 6. *The optimal terminal time* t_f *is finite.*

The following two examples verify Assumptions 5 and 6.

Example 4. On a Riemannian manifold N, minimal geodesics between two sufficiently nearby given points in a compact subset S of N are contained within a bounded set $W \supset S$. The costates associated with the geodesic problem are also bounded by virtue of (6) because the controls $u_i = \dot{x}_i$ are bounded. Furthermore $t_f = 1$ is fixed.

Example 5. In the case of linear time-varying control systems the right-hand side of (1) is not bounded. However, suppose that the optimal control can be expressed as a linear state feedback and that it is substituted into the system equation to get a system dependent only on the state. It is a well-known fact that there is no *finite-escape time* for linear time-varying systems [21]. So the states are bounded. The associated costates are bounded likewise. The simple case of a fixed time horizon is an example to Assumption 6.

Let $y \in Y$, and define $y^{(k)} = F^k(y)$ for $k \geq 1$.

LEMMA 5.1. Y *is compact.*

Proof. Since $C(x_0, \alpha_0)$ is bounded (by Assumption 5), the set Y is also bounded. Now one needs to show that Y is closed as well. Consider $y = (y_0, y_1, \ldots, y_q) \in Y$. Note that $z_i^{(k)} \in \mathcal{R}(z_{i-1}^{(k)}, \leq \delta)$ for the kth iterate $y^{(k)} = F(y^{(k-1)})$. Since the closure $\overline{C}(x_0, \alpha_0)$ is compact, one can take a convergent subsequence $\{y^{(k_j)} : j \geq 1\}$ of the sequence $\{y^{(k)} : k \geq 0\} \subset Y$. Write

$$\lim_{j \longrightarrow \infty} y^{(k_j)} = y^{(\infty)} \ , \tag{17}$$

and so

$$\lim_{j \longrightarrow \infty} z^{(k_j)} = z^{(\infty)} \ ,$$

where

$$y^{(\infty)} = \left(y_0^{(\infty)}, y_1^{(\infty)}, \ldots, y_q^{(\infty)}\right)$$

and

$$z^{(\infty)} = \left(z_0^{(\infty)}, z_1^{(\infty)}, \ldots, z_q^{(\infty)}\right) \ .$$

Recall that for $1 \leq i \leq q$

$$\tau(z_{i-1}^{(k)}, z_i^{(k)}) \leq \delta \ .$$

By taking limits and using the continuity of τ, we get

$$\tau(z_{i-1}^{(\infty)}, z_i^{(\infty)}) \leq \delta,$$

which implies that $y^{(\infty)} \in Y$. ∎

LEMMA 5.2. *Each sequence* $\{t_i^{(k)} : k \geq 0\}$, $i = 1, \ldots, q$, *has a convergent subsequence with the same index set as that of* $\{y^{(k_j)} : j \geq 1\}$ *such that* $\lim_{j \longrightarrow \infty} y^{(k_j)} = y^{(\infty)}$.

Proof. The conclusion follows from Assumption 6 and the limit expression in (17). ∎

LEMMA 5.3. *The sequence* $\{\alpha(z^{(k)}) : k \geq 1\}$ *converges to its infimum* $\alpha^{(\infty)}$. *In particular*

$$0 < \alpha(z^{(\infty)}) = \alpha^{(\infty)} \in [\widetilde{\tau}(z_0, z_q), q\delta].$$

Proof. The result follows from the continuity of α and Lemmas 5.1 and 4.5. ∎

5.2. Extreme points and leapfrog splicing. Recall that by Assumption 2 the curve $\gamma_{z_{i-1}, z_{i+1}} : [t_{i-1}, t_{i+1}] \longrightarrow \mathbb{R}^n$ is the unique optimal trajectory from z_{i-1} to z_{i+1}. The point $y_i = (z_i, \lambda_i)$ is said to be *between* y_{i-1} and y_{i+1} if z_i lies in the image of $\gamma_{z_{i-1}, z_{i+1}}$. Note that, trivially, y_{i-1} and y_{i+1} are between y_{i-1} and y_{i+1}.

The $(q+1)$-tuple $y \in Y$ is said to be *extreme* when y_i is between y_{i-1} and y_{i+1} for all $1 < i < q$.

Let $\mu_i : [t_i, t_{i+1}] \longrightarrow \mathbb{R}^{2n}$ be the optimal trajectory pair $(\gamma_{z_{i-1}, z_i}(\cdot), \psi(\cdot))$ solving the optimality system associated with problem (P_i) from z_{i-1} to z_i, $i = 1, \ldots, q$. Then $\mu_y : [t_0, t_f] \longrightarrow \mathbb{R}^{2n}$ is defined as the concatenation (or track sum) of $\mu_1, \mu_2, \ldots, \mu_q$ in the given order such that $\mu_y(t_0) = \mu_1(t_0)$, $\mu_y(t_i) = \mu_i(t_i) = \mu_{i+1}(t_i)$, $i = 1, \ldots, q-1$, and $\mu_y(t_f) = \mu_q(t_q)$. It is convenient to use a special symbol, for example, $\smile$, for this kind of concatenation. The trajectory μ_y can then be expressed by

$$\mu_y := \mu_1 \smile \mu_2 \smile \cdots \smile \mu_q \ .$$

LEMMA 5.4. *If* y *is extreme, then* μ_y *is a critical trajectory.*

Proof. Suppose that $y \in Y$ is extreme. We will use induction on what we call the *leapfrog splicing* of trajectories, which is illustrated in Figure 3, to prove the lemma. We define

$$\mu_y^{(p)} := \mu_1 \smile \mu_2 \smile \cdots \smile \mu_p$$

and aim to prove the following: $\mu_y^{(p)}$ is critical with the fixed end points $x(t_0) = z_0$ and $x(t_p) = z_p$ for $2 \leq p \leq q$, where q is provided.

Suppose that $p = 2$. Then $\mu_y^{(2)} = \mu_1 \smile \mu_2$. Because y_1 is between y_0 and y_2, z_1 is in the image of the optimal trajectory γ_{z_0, z_2}. Therefore there exists some nontrivial $\psi(\cdot)$ such that $\mu_y^{(2)} = (\gamma_{z_0, z_2}(\cdot), \psi(\cdot))$ solves the associated optimality system. So $\mu_y^{(2)}$ is critical.

Assume that $\mu_y^{(k)} = \mu_1 \smile \mu_2 \smile \cdots \smile \mu_k$ is critical. The next leapfrog step from z_{k-1} to z_{k+1} gives the local optimal trajectory segment $\gamma_{z_{k-1}, z_{k+1}}(\cdot)$ and the associated

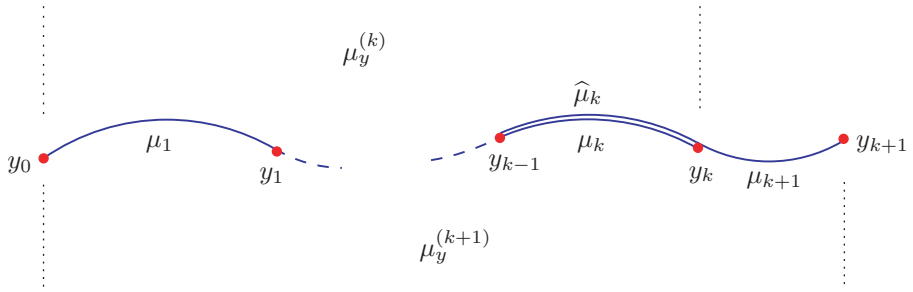


FIG. 3. *Concatenation of the local optimal trajectories μ_i and the leapfrog splicing.*

costate trajectory $\widehat{\psi}(\cdot)$. So $(\widehat{\gamma}_{z_{k-1},z_{k+1}}(\cdot), \widehat{\psi}(\cdot)) = \widehat{\mu}_k \smile \mu_{k+1}$. By the induction hypothesis, $\mu_k = (\gamma_{z_{k-1},z_k}(\cdot), \psi(\cdot))$ is optimal. Because y_k is between y_{k-1} and y_{k+1}, $\widehat{\mu}_k = (\widehat{\gamma}_{z_{k-1},z_k}(\cdot), \widehat{\psi}(\cdot))$ is also optimal. By the local uniqueness of the controls and costates posed in Assumptions 2 and 3, $\gamma_{z_{k-1},z_k}(\cdot) = \widehat{\gamma}_{z_{k-1},z_k}(\cdot)$ and $\psi(\cdot) = \widehat{\psi}(\cdot)$. Therefore $\mu_k = \widehat{\mu}_k$. By "splicing" the trajectory segments $\mu_1 \smile \mu_2 \smile \cdots \smile \mu_k$ and $\widehat{\mu}_k \smile \mu_{k+1}$, by means of the overlap $\widehat{\mu}_k = \mu_k$ between y_{k-1} and y_k (see Figure 3), we obtain the trajectory

$$\mu_y^{(k+1)} = \mu_1 \smile \mu_2 \smile \cdots \smile \mu_k \smile \mu_{k+1},$$

which satisfies the optimality system associated with the optimal control problem with the end conditions $x(t_0) = z_0$ and $x(t_{k+1}) = z_{k+1}$. So $\mu_y^{(k+1)}$ is critical. ∎

Remark 5.1. The converse of Lemma 5.4 holds if μ_y is not only critical but also (locally) optimal.

LEMMA 5.5. (a) *If y is extreme and all z_i, $i = 1, \ldots, q-1$, are midpoints, then $F(y) = y$.*

(b) *If $F(y) = y$, then y is extreme.*

Proof. (a) Suppose that y is extreme and all z_p, $p = 1, \ldots, q-1$, are midpoints, that is, $z_p = \widehat{z}_p = M(z_{p-1}, z_{p+1})$. One further has $\lambda_p = \widehat{\lambda}_p$, because, by using the notation of the proof of Lemma 5.4, $\widehat{\mu}_{p-1} \smile \widehat{\mu}_p = \mu_{p-1} \smile \mu_p$ by the local uniqueness of optimal trajectories. Therefore $y_p = \widehat{y}_p = G_p(y)$, $p = 1, \ldots, q-1$. The composition $F = G_{q-1} \circ G_{q-2} \circ \cdots \circ G_1$ then gives $F(y) = \widehat{y} = y$.

(b) Suppose that $F(y) = y$. Recall that $\widehat{z}_0 = y_0$, $\widehat{z}_p = M(\widehat{z}_{p-1}, z_{p+1})$ for $p = 1, \ldots, q-1$, and $\widehat{z}_q = z_q$, where M is the midpoint map. Now since $F(y) = y$,

$$z_p = \widehat{z}_p = M(z_{p-1}, z_{p+1}) \ .$$

So y_p is between y_{p-1} and y_{p+1}, $p = 1, \ldots, q-1$. ∎

LEMMA 5.6. *y is extreme if and only if $\mu_{F(y)} = \mu_y$.*

Proof. Suppose that y is extreme. Because y_p is between y_{p-1} and y_{p+1}, $\widehat{y}_p = G_p(y)$ is also between y_{p-1} and y_{p+1}. So $\mu_{G_p(y)} = \mu_y$, by the uniqueness of $\gamma_{z_{i-1},z_{i+1}}$. The composition $F = G_{q-1} \circ G_{q-2} \circ \cdots \circ G_1$ gives $\mu_{F(y)} = \mu_y$.

Suppose that $\mu_{F(y)} = \mu_y$. Then $\mu_{G_1(y)} = \mu_y$. By using the notation of the proof of Lemma 5.4, $\widehat{\mu}_1 \smile \widehat{\mu}_2 = \mu_1 \smile \mu_2$. So $\widehat{y}_1$ is between y_0 and y_1. The composition $F = G_{q-1} \circ G_{q-2} \circ \cdots \circ G_1$ yields that y is extreme. ∎

LEMMA 5.7. *y is extreme if and only if $\alpha(\pi_1 \circ F(y)) = \alpha(z)$.*

Proof. Suppose that y is extreme. Then $\alpha(\pi_1 \circ F(y)) = \alpha(z)$ by Lemma 5.6. Suppose that $\alpha(\pi_1 \circ F(y)) = \alpha(z)$. Recall that

$$\alpha(z) = \tau(z_0, z_1) + \cdots + \tau(z_{p-1}, z_p) + \tau(z_p, z_{p+1}) + \cdots + \tau(z_{q-1}, z_q) \ .$$

For $0 < p < q$,

$$\alpha(\pi_1 \circ G_p(y)) = \tau(z_0, z_1) + \cdots + \tau(z_{p-1}, \widehat{z}_p) + \tau(\widehat{z}_p, z_{p+1}) + \cdots + \tau(z_{q-1}, z_q) \ .$$

Then

$$\tau(z_{p-1}, \widehat{z}_p) + \tau(\widehat{z}_p, z_{p+1}) = \tau(z_{p-1}, z_p) + \tau(z_p, z_{p+1}) \ . \tag{18}$$

By the definition of the midpoint map, $\widehat{z}_p$ lies in the image of $\gamma_{z_{p-1}, z_{p+1}}$. Equation (18) implies that z_p lies in a trajectory of equal cost, and this trajectory can only be $\gamma_{z_{p-1}, z_{p+1}}(\cdot)$, by uniqueness. Therefore y_p is between y_{p-1} and y_{p+1}. The composition $F = G_{q-1} \circ G_{q-2} \circ \cdots \circ G_1$ then provides that y is extreme. ∎

5.3. Convergence to a critical trajectory.

LEMMA 5.8. *$y^{(\infty)}$ is extreme.*

Proof. By Lemmas 4.5 and 5.3, $\alpha(\pi_1 \circ F(y^{(\infty)})) = \alpha^{(\infty)} = \alpha(\pi_1(y^{(\infty)}))$. Then the result follows by Lemma 5.7. ∎

LEMMA 5.9. *$\mu_{y^{(\infty)}}$ is a critical trajectory.*

Proof. Lemmas 5.8 and 5.4 furnish the lemma. ∎

THEOREM 5.1. *The leapfrog iterations converge to a critical trajectory $\mu_{y^{(\infty)}}$.*

Proof. Lemmas 5.8 and 5.6 imply that $\mu_{F(y^{(\infty)})} = \mu_{y^{(\infty)}}$, which together with Lemma 5.9 proves the theorem. ∎

Although Theorem 5.1 guarantees convergence of the trajectory iterates $\mu_{y^{(k)}}$, no assertion has yet been made about convergence of the iterates $y^{(k)}$. Convergence of $y^{(k)}$ is particularly important in a numerical implementation. If it is known that $y^{(k)}$ is convergent, too, then, instead of using an infinite-dimensional norm (for example, the L^∞ norm) to measure the change from the trajectory iterate $\mu_{y^{(k)}}$ to $\mu_{y^{(k+1)}}$, a finite-dimensional norm (for example, the Euclidean norm) can be used to measure the change from the iterate $y^{(k)}$ to $y^{(k+1)}$, which is much less demanding computationally. In what follows, we will give a lemma and Theorem 5.2, which guarantees convergence of the iterates $y^{(k)}$.

LEMMA 5.10. *If y is extreme, then all $z_i^{(\infty)}$, $i = 1, \ldots, q-1$, are uniformly distributed with respect to cost, namely,*

$$\tau(z_{i-1}^{(\infty)}, z_i^{(\infty)}) = \tau(z_i^{(\infty)}, z_{i+1}^{(\infty)}) \ ,$$

$i = 1, \ldots, q-1$.

Proof. Let y be extreme. Then by Lemma 5.4, μ_y is critical. Suppose that μ_y is parameterized proportionally to the cost so that $\mu_y : [0,1] \longrightarrow \mathbb{R}^{2n}$, with $\mu_y(0) = y_0$ and $\mu_y(1) = y_q$. Note that, by Lemma 5.6, $\mu_{F(y)} = \mu_y$. Then it suffices to show that the sequence $\{F^k(y) : k \geq 1\}$ converges to the uniformly distributed $(q+1)$-tuple

$$(y_0, \mu_y(1/q), \mu_y(2/q), \ldots, \mu_y(i/q), \ldots, y_q) \ .$$

The rest of the proof follows similar lines as those given by [30, Lemma 3.2], which we adapt here by using our notation and definitions.

If $q = 2$, the result is clear since $F(y)$ is independent of y_1 and the limit is achieved immediately as $F(y)$. For $q > 2$ we give some definitions first and then prove the convergence to the required $(q+1)$-tuple.

Let $y^{(0)} = y$ and $y^{(k)} = F^k(y)$ for $k = 1, 2, \ldots$. Since $\mu_{y^{(k)}} = \mu_{y^{(0)}}$ by Lemma 5.6,

$$y^{(k)} = (\mu(0), \mu(t_1^{(k)}), \mu(t_2^{(k)}), \ldots, \mu(t_{q-1}^{(k)}), \mu(t_q^{(k)}))$$

for $k = 0, 1, 2, \ldots$, where $t_i^{(k)}$, $i = 1, \ldots, q-1$, are the junction times. Since μ_y is parameterized proportionally to the cost, application of F gives the recursion

$$t_i^{(k)} = (t_{i-1}^{(k)} + t_{i+1}^{(k-1)})/2$$

for $i = 1, \ldots, q-1$, where $t_0^{(k)} = 0$, $t_q^{(k)} = 1$. Write

$$t^{(k)} = [t_2^{(k)}, t_3^{(k)}, \ldots, t_{q-1}^{(k)}]^T \in \mathbb{R}^{q-2} \ .$$

Then $t^{(k)} = A\, t^{(k-1)} + b$, where

$$A = \begin{bmatrix} 1/2^2 & 1/2 & 0 & \cdots & 0 \\ 1/2^3 & 1/2^2 & 1/2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1/2^{q-1} & 1/2^{q-2} & 1/2^{q-3} & \cdots & 1/2^2 \end{bmatrix}$$

and $b = [0, 0, \ldots, 0, 1/2]^T \in \mathbb{R}^{q-2}$. It is straightforward to obtain

$$t^{(k)} = b + Ab + A^2 b + \cdots + A^{k-1} b + A^k t^0 \ ,$$

which converges since $\|A\| < 1$. So the limit $t^{(\infty)}$ satisfies $t^{(\infty)} = A\, t^{(\infty)} + b$. It is seen that

$$t^{(\infty)} = [2, 3, \ldots, q-1]^T / q$$

is the unique solution, which completes the proof. □

THEOREM 5.2. *$\{F(y^{(k)}) : k \geq 0\}$ has a unique accumulation point.*

Proof. By Lemmas 5.8, 5.10, and 5.5(a), $F(y^{(\infty)}) = y^{(\infty)}$. □

6. Numerical implementation. We will illustrate a numerical implementation of leapfrog on an optimal control problem with a fixed terminal time. Recall Remark 2.2. For a fixed terminal time problem, one solves the subproblems (P_i) with both t_{i-1} and t_{i+1} fixed. At the start of the algorithm, the values $t_{i-1}^{(0)}$ and $t_{i+1}^{(0)}$ are given by a partition between $x_{n+1}(0) = 0$ and $x_{n+1}(t_f) = 1$: One typically takes a regular partition of the time interval $[0, 1]$. Then, given $t_{i-1}^{(k+1)}$ and $t_{i+1}^{(k)}$, the time iterate $t_i^{(k+1)}$ is determined through the solution of (P_i) as the time at which the midpoint $z_i^{(k+1)}$ is placed.

6.1. The van der Pol system. The van der Pol system has been used in earlier control applications, including time-optimal control problems where the control regime is bang–bang [18, 27]. Here we consider the van der Pol system with unbounded control

$$\dot{x}_1(t) = x_2(t) \ , \tag{19}$$

$$\dot{x}_2(t) = -x_1(t) - (x_1^2(t) - 1)\, x_2(t) + u(t) \ , \tag{20}$$

where the problem is to get from $x(0) = (-2, -3)$ to $x(1) = (5, 0)$, minimizing the quadratic cost

$$\frac{1}{2} \int_0^1 (x_1^2(t) + x_2^2(t) + u^2(t)) dt \ .$$

The Hamiltonian is

$$H(x,\psi,u) = (x_1^2 + x_2^2 + u^2)/2 + \psi_1 x_2 + \psi_2[-x_1 - (x_1^2 - 1)x_2 + u] \ .$$

The maximum principle yields, in addition to the system equations (19)–(20), the following costate equations:

$$\dot{\psi}_1(t) = -x_1(t) + (1 + 2x_1(t)x_2(t))\,\psi_2(t) \ , \tag{21}$$

$$\dot{\psi}_2(t) = -x_2(t) - \psi_1(t) + (x_1^2(t) - 1)\,\psi_2(t) \ . \tag{22}$$

Furthermore optimal control has to satisfy

$$u(t) = -\psi_2(t) \ . \tag{23}$$

Substitution of (23) gives rise to the following TPBVP:

$$\dot{x}_1 = x_2 \ , \tag{24}$$

$$\dot{x}_2 = -x_1 - (x_1^2 - 1)\,x_2 - \psi_2 \ , \tag{25}$$

$$\dot{\psi}_1 = -x_1 + (1 + 2x_1x_2)\,\psi_2 \ , \tag{26}$$

$$\dot{\psi}_2 = -x_2 - \psi_1 + (x_1^2 - 1)\,\psi_2 \ , \tag{27}$$

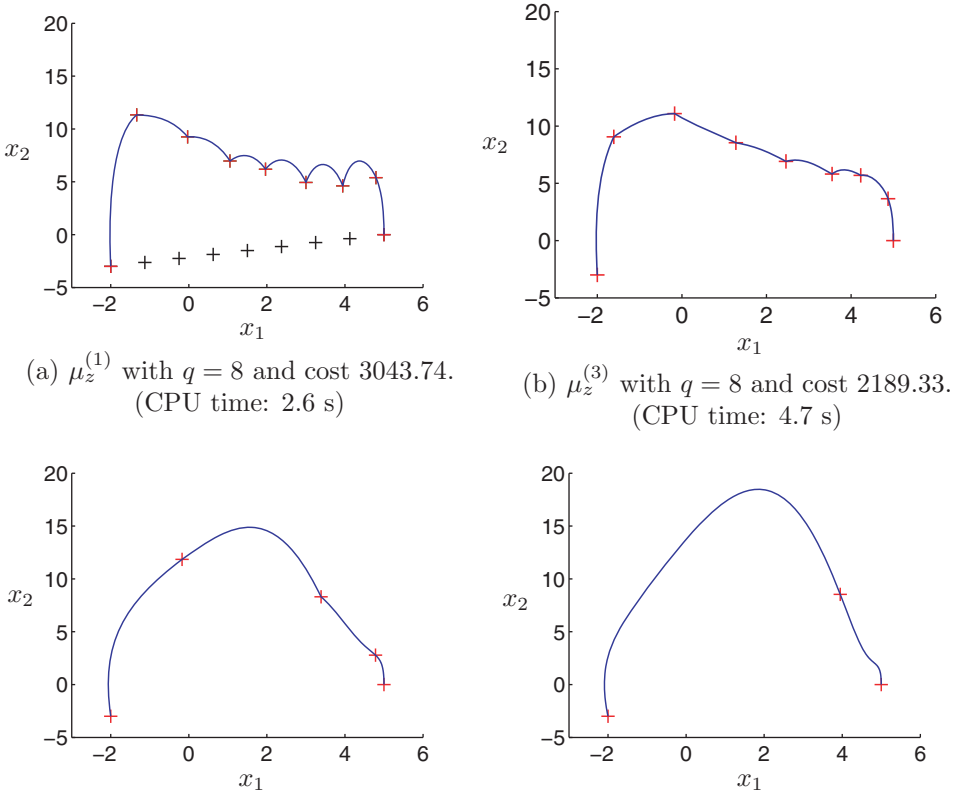
$$\text{with } x(0) = (-2,-3) \quad \text{and} \quad x(1) = (5,0) \ .$$

6.1.1. Solution with leapfrog. The leapfrog algorithm owes its working largely to the solutions of subproblems (P_i). These subproblems for the van der Pol problem have been reduced to (24)–(27) with the accompanying boundary conditions. A solution to these equations can be found by using simple shooting, because the leapfrog partition points are initially chosen close enough to one another so that affine approximations provide good initial guesses for simple shooting.

A MATLAB code has been written to implement the leapfrog algorithm. For solving the ODEs (24)–(27), the MATLAB function `ode45` is utilized with the absolute and relative tolerances of 10^{-8}. The simple shooting technique that we have programmed incorporates Newton's method with backtracking [8]. This increases the chances of simple shooting to converge; however, if the number of iterations within simple shooting exceeds 10, then we regard simple shooting to have failed. In the final iteration of leapfrog the relative and absolute tolerances for `ode45` are sharpened to 10^{-10}.

Iterations of the leapfrog algorithm in the state space are shown in Figure 4. Figure 4(a) displays both the initial partition with which we start and the concatenation of the locally optimal trajectories $\mu_z^{(1)}$ obtained as a result of the first leapfrog iteration. The initial points $z_i^{(0)}$ have been chosen to be equally spaced along the straight line between x_0 and x_T. For ease of bookkeeping in the code we consider powers of 2 for the number of subdivisions; i.e., we use $q = 2, 4, 8$, etc. In each iteration we use the smallest possible q with which each subproblem (P_i) is solved successfully. This eventually leads to a final iteration where $q = 2$, and simple shooting is used to achieve the solution.

With leapfrog, different positioning of the midpoints is permissible, as long as the subintervals of the consequent subproblems overlap. So, in the implementation, the midpoint $z_i^{(k+1)}$ is chosen as the point which is reached in roughly half the time along the optimal trajectory from $z_{i-1}^{(k+1)}$ to $z_{i+1}^{(k)}$. This choice of midpoints is computationally more economical than the choice of midpoint where *exactly* half the cost is incurred.



(a) $\mu_z^{(1)}$ with $q = 8$ and cost 3043.74. (CPU time: 2.6 s)

(b) $\mu_z^{(3)}$ with $q = 8$ and cost 2189.33. (CPU time: 4.7 s)

(c) $\mu_z^{(6)}$ with $q = 4$ and cost 1151.34. (CPU time: 8.0 s)

(d) $\mu_z^{(10)}$ with $q = 2$ and cost 993.67. (CPU time: 14.2 s)

FIG. 4. *Locally optimal state trajectories $\mu_z^{(k)}$ during leapfrog iterations for the van der Pol system. Crosses in iteration k represent partition points $z_i^{(k)}$, $i = 0, 1, \ldots, q$.*

The first to fifth iterations could all be successfully completed by taking at least eight subdivisions. The sixth to ninth iterations could be carried out with a minimum of four subdivisions, where the new partition is formed in the sixth iteration by taking the points $z_2^{(5)}$, $z_4^{(5)}$, and $z_6^{(5)}$ from iteration 5. The trajectory $\mu_z^{(3)}$ and the updated partition points $z_i^{(3)}$, $i = 1, 2, 3$, are illustrated in Figure 4(c). The tenth iteration, after reducing the number of partitions to two, or simply applying simple shooting between x_0 and x_T with the initial guess value $\psi(0) = \lambda_0^{(9)}$, yields the solution trajectory for the problem, which is depicted in Figure 4(d). The costs along each of the leapfrog trajectory iterates $\mu_z^{(k)}$, $k = 1, 3, 6, 10$, are reported in Figures 4(a)–(d), respectively. We also report the CPU times, cumulatively at each iteration, as measured within MATLAB, running on the (single user) Windows XP Professional (Version 5.1) operating system with a 2.00 GHz Intel Pentium M processor and 1 GB of RAM.

The first leapfrog iteration can also be viewed as one where a trajectory is constructed in a simple way and verified to be feasible. Because of the Newton backtracking, this iteration takes more than twice longer than each of the rest of the iterations except the last one. The last iteration, where we reduce the whole process to simple shooting, also takes longer because of the Newton backtracking and because we require a higher accuracy.

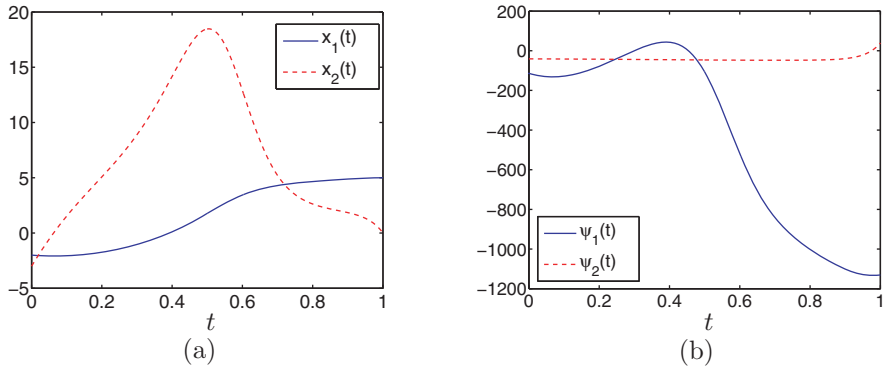


FIG. 5. *The solution state and costate trajectories for the van der Pol system.*

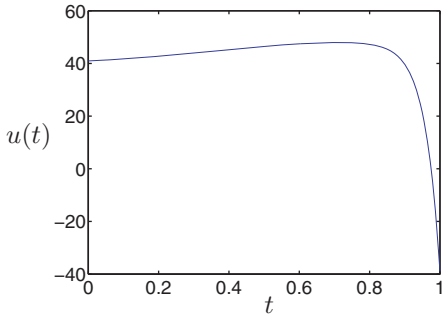


FIG. 6. *The solution control trajectory for the van der Pol system.*

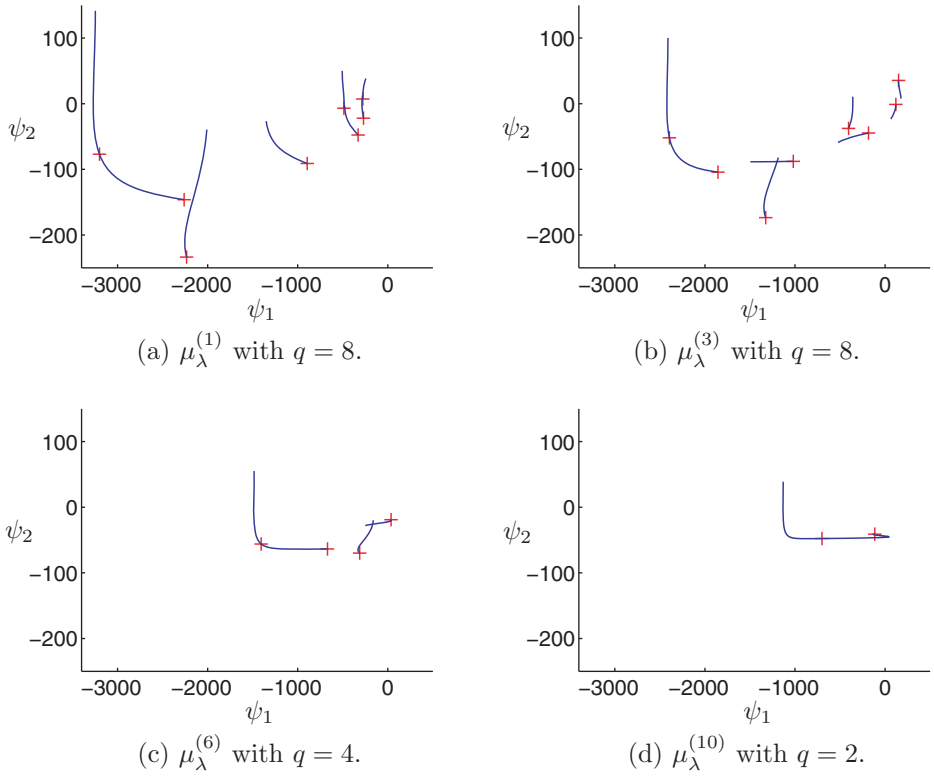
The solution trajectories of the state, costate, and control variables as functions of time can be seen in Figures 5(a)–(b) and 6. In Figure 6, simply $u(t) = -\psi_2(t)$. The initial costate vector $\psi(0) = (-113.817881, -40.950820)$ generates an accurate solution.

Some of the locally optimal costate trajectories obtained in the leapfrog iterations are shown in Figure 7.

Overall it appears that there are pros and cons with using leapfrog. Numerical evidence arising from our application of leapfrog to the van der Pol problem point to certain advantages and disadvantages of leapfrog.

ADVANTAGE 1. With leapfrog, one has only to choose the starting values of the states. These values can usually be chosen in a practical manner (for example, along a straight line joining x_0 and x_T), which could then generate a feasible trajectory. In the leapfrog scheme, it is not necessary to guess the starting values of costates. For solving the subproblems in leapfrog, affine approximations usually give reasonable starting values for the costates.

ADVANTAGE 2. At each iteration, leapfrog generates a locally optimal trajectory $\mu_z^{(k)}$, which is feasible. Suppose that the time allowed to design an optimal control for the system is not long enough to compute a solution. Then leapfrog may provide at least a suboptimal control. For example, if it is necessary to obtain a solution in less than 10 seconds, then, within the computational means we have in this article, we may offer the suboptimal solution depicted in Figure 4(c), which is far better than the feasible solution given in Figure 4(a).



(a) $\mu_\lambda^{(1)}$ with $q = 8$. (b) $\mu_\lambda^{(3)}$ with $q = 8$.

(c) $\mu_\lambda^{(6)}$ with $q = 4$. (d) $\mu_\lambda^{(10)}$ with $q = 2$.

FIG. 7. *Locally optimal costate trajectories during leapfrog iterations for the van der Pol system. Crosses in iteration k represent partition points $\lambda_i^{(k)}$, $i = 0, 1, \ldots, q-1$.*

One restriction on the applicability of the suboptimal solution found by leapfrog is that the control $u(t) = -\psi(t)$ is discontinuous, because the piecewise optimal $\psi(t)$ obtained in the third iteration is discontinuous, as can be seen in Figure 7(c). However, many dynamical systems allow discontinuous controls; so the suboptimal control generated by leapfrog can possibly be implemented.

An apparent drawback of leapfrog is that, as the solution is approached, it is observed from the numerical experiments that, with a fixed number of partitions $q > 2$, the leapfrog updates get sluggish. Therefore, q needs to be reduced progressively (eventually down to $q = 2$) for convergence in a reasonable number of iterations. This effort may in general incur an additional computational cost.

6.2. Further comments on implementation of leapfrog. We observe that the solution for the problem in the previous subsection is obtained in ten iterations. It is worthwhile to note that the same problem with different initial and target states may require a different number of subdivisions at the start, and reducing the number of subdivisions may take a lot longer than what is presented here. The performance of the algorithm would depend not only on the difficulty of the problem but also on the way the algorithm is implemented. Two particular issues seem to be of main concern in any implementation of leapfrog, namely, (i) the efficiency of the simple shooting technique employed in subproblems (P_i) and (ii) a dynamic way of doing the partitioning along a leapfrog trajectory iterate.

Because we were concerned only with providing an illustration of the leapfrog algorithm in this particular study, we have coded a rather straightforward simple shooting scheme. A more sophisticated simple shooting scheme may be necessary for more difficult problems. In particular, one should note that `ode45` is not necessarily the method of choice for the integration of the state-costate equations; it merely serves to illustrate an implementation of leapfrog. Otherwise, in a more elaborate implementation, a certain approximation order for both the states and the costates can be achieved only if certain conditions are satisfied [12].

In the case when $q > 2$ is fixed, convergence can usually not be obtained in a finite number of iterations. Furthermore if q is large, then the numerical experiments show that convergence becomes rather slow. So it is important to employ some kind of *dynamic* partitioning. A simple dynamic partitioning scheme was prescribed in the example application in section 6.1. However, more sophisticated schemes can also be employed. One can increase or decrease the number of partitions *locally* in a smaller segment of the leapfrog trajectory, as the need arises. Given a trajectory, it may be appropriate to increase the number of subdivisions in one segment of the trajectory, while it is appropriate to decrease the number in some other segment. The ultimate aim would still remain to reduce the number of iterations to just two, in which case simple shooting, if successful, would deliver the required solution.

Acknowledgment. The authors offer their warm thanks to the anonymous reviewer whose comments and suggestions improved the manuscript.

REFERENCES

[1] U. M. Ascher and S. Y. P. Chan, *On parallel methods for boundary value ODEs*, Computing, 46 (1991), pp. 1–17.

[2] U. M. Ascher, R. M. M. Mattheij, and R. D. Russell, *Numerical Solution of Boundary Value Problems for Ordinary Differential Equations*, SIAM, Philadelphia, 1995.

[3] J. T. Betts, *Practical Methods for Optimal Control Using Nonlinear Programming*, SIAM, Philadelphia, 2001.

[4] H. G. Bock and K. J. Plitt, *A multiple shooting algorithm for direct solution of optimal control problems*, in Proceedings of the 9th IFAC World Congress, Vol. 9, Budapest, Hungary, 1984, pp. 242–247.

[5] W. M. Boothby, *An Introduction to Differentiable Manifolds and Riemannian Geometry*, Academic Press, London, 1986.

[6] C. Büskens, *Optimierungsmethoden und Sensitivitätsanalyse für optimale Steuerprozesse mit Steuer- und Zustands-Beschränkungen*, dissertation, Institut für Numerische Mathematik, Universität Münster, Germany, 1998.

[7] Y. Chen and A. A. Desrochers, *Minimum-time control laws for robotic manipulators*, Internat. J. Control, 57 (1993), pp. 1–27.

[8] J. E. Dennis, Jr., and R. B. Schnabel, *Numerical Methods For Unconstrained Optimization and Nonlinear Equations*, SIAM, Philadelphia, 1996.

[9] P. Deuflhardt, *A stepsize control for continuation methods and its special application to multiple shooting techniques*, Numer. Math., 33 (1979), pp. 115–146.

[10] G. Fraser-Andrews, *A multiple shooting technique for optimal control*, J. Optim. Theory Appl., 102 (1999), pp. 299–313.

[11] W. Grimm and A. Markl, *Adjoint estimation from a direct multiple shooting method*, J. Optim. Theory Appl., 92 (1997), pp. 263–283.

[12] W. W. Hager, *Runge-Kutta methods in optimal control and the transformed adjoint system*, Numer. Math., 87 (2000), pp. 247–282.

[13] T. Kailath, *Linear Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1980.

[14] C. Y. Kaya, S. K. Lucas, and S. T. Simakov, *Computations for bang–bang constrained optimal control using a mathematical programming formulation*, Optimal Control Appl. Methods, 25 (2004), pp. 295–308.

[15] C. Y. Kaya and J. L. Noakes, *Geodesics and an optimal control algorithm*, in Proceedings of the 36th IEEE Conference on Decision and Control (CDC), San Diego, CA, 1997, pp. 4918–4919.

[16] C. Y. Kaya and J. L. Noakes, *The leap-frog algorithm and optimal control: Background and demonstration*, in Proceedings of International Conference on Optimization and Applications (ICOTA '98), Perth, Australia, 1998, pp. 835–842.

[17] C. Y. Kaya and J. L. Noakes, *The leap-frog algorithm and optimal control: Theoretical aspects*, in Proceedings of International Conference on Optimization and Applications (ICOTA '98), Perth, Australia, 1998, pp. 843–850.

[18] C. Y. Kaya and J. L. Noakes, *Computational method for time-optimal switching control*, J. Optim. Theory Appl., 117 (2003), pp. 69–92.

[19] C. Y. Kaya and J. M. Martínez, *Euler discretization and inexact restoration for optimal control*, J. Optim. Theory Appl., 134 (2007), pp. 191–206.

[20] H. B. Keller, *Numerical Methods for Two-Point Boundary-Value Problems*, Blaisdell, London, 1968.

[21] H. K. Khalil, *Nonlinear Systems*, Maxwell Macmillan International, New York, 1992.

[22] M. E. Kramer and R. M. M. Mattheij, *Application of global methods in parallel shooting*, SIAM J. Numer. Anal., 30 (1993), pp. 1723–1739.

[23] P. Kunkel, V. Mehrmann, and R. Stöver, *Multiple shooting for unstructured nonlinear differential-algebraic equations of arbitrary index*, SIAM J. Numer. Anal., 42 (2005), pp. 2277–2297.

[24] U. Ledzewicz and H. Schättler, *A high-order generalized local maximum principle*, SIAM J. Control Optim., 38 (2000), pp. 823–854.

[25] D. B. Leineweber, I. Bauer, H. G. Bock, and J. P. Schlöder, *An efficient multiple shooting based reduced SQP strategy for large-scale dynamic process optimization. Part* 1: *Theoretical aspects*, Computers and Chemical Engineering, 27 (2003), pp. 157–166.

[26] H. Maurer, *Numerical solution of singular control problems using multiple shooting techniques*, J. Optim. Theory Appl., 18 (1976), pp. 235–257.

[27] H. Maurer and N. P. Osmolovskii, *Second order sufficient conditions for time-optimal bang–bang control*, SIAM J. Control Optim., 42 (2004), pp. 2239–2263.

[28] H. Maurer, C. Büskens, J.-H. R. Kim, and C. Y. Kaya, *Optimization methods for the verification of second-order sufficient conditions for bang–bang controls*, Optimal Control Appl. Methods, 26 (2005), pp. 129–156.

[29] A. Miele and T. Wang, *Parallel computation of two-point boundary-value problems via particular solutions*, J. Optim. Theory Appl., 79 (1993), pp. 5–29.

[30] J. L. Noakes, *A global algorithm for geodesics*, J. Aust. Math. Soc., Ser. A, 65 (1998), pp. 37–50.

[31] H. J. Oberle and W. Grimm, *BNDSCO - A Program for the Numerical Solution of Optimal Control Problems*, report 515, Institute for Flight System Dynamics, Oberpfaffenhofen, German Aerospace Research Establishment (DLR), 1989.

[32] M. R. Osborne, *On shooting methods for boundary value problems*, J. Math. Anal. Appl., 29 (1969), pp. 417–433.

[33] H. Pasic, *Multi-point boundary-value solution of two-point boundary-value problems*, J. Optim. Theory Appl., 100 (1999), pp. 397–416.

[34] L. S. Pontryagin, V. G. Boltyanskii, R. V. Gamkrelidze, and E. F. Mishchenko, *The Mathematical Theory of Optimal Processes*, Interscience, New York, 1962.

[35] J. Stoer and R. Bulirsch, *Introduction to Numerical Analysis*, 2nd ed., Springer-Verlag, New York, 1993.

[36] K. L. Teo, C. J. Goh, and K. H. Wong, *A Unified Computational Approach to Optimal Control Problems*, Longman Scientific and Technical, New York, 1991.

[37] O. von Stryk and R. Bulirsch, *Direct and indirect methods for trajectory optimization*, Ann. Oper. Res., 37 (1992), pp. 357-373.

[38] O. von Stryk and M. Schlemmer, *Optimal control of the industrial robot Manutec r*3, in Computational Optimal Control, R. Bulirsch and D. Kraft, eds., Birkhäuser-Verlag, Basel, 1994, pp. 367–382.

[39] Z.-Q. Zuo, *Two new techniques for optimal control*, IEEE Trans. Automat. Control, 36 (1991), pp. 1307–1310.

SIAM J. NUMER. ANAL.
Vol. 46, No. 6, pp. 2818–2842
© 2008 Society for Industrial and Applied Mathematics

PRIMAL AND DUAL INTERFACE CONCENTRATED ITERATIVE SUBSTRUCTURING METHODS*

SVEN BEUCHLER†, TINO EIBNER‡, AND ULRICH LANGER§

Abstract. This paper is devoted to the fast solution of interface concentrated finite element equations. The interface concentrated finite element schemes are constructed on the basis of a nonoverlapping domain decomposition where a conforming boundary concentrated finite element approximation is used in every subdomain. Similar to boundary element domain decomposition methods, the total number of unknowns per subdomain behaves like $O((H/h)^{(d-1)})$, where H, h, and d denote the usual scaling parameter of the subdomains, the average discretization parameter of the subdomain boundaries, and the spatial dimension, respectively. We propose and analyze primal and dual substructuring iterative methods which asymptotically exhibit the same or at least almost the same complexity as the number of unknowns. In particular, the so-called all-floating finite element tearing and interconnecting solvers are highly parallel and very robust with respect to large coefficient jumps.

Key words. solution of discretized equations, domain decomposition, finite elements

AMS subject classifications. 65N30, 65N22, 65N55

DOI. 10.1137/070691723

1. Introduction.

Boundary concentrated finite element methods (BC-FEM), introduced and investigated by Khoromskij and Melenk in [22], allow us to solve potential problems with smooth coefficients and right-hand sides as accurately as the corresponding standard finite element methods (FEMs) but with a significantly smaller number of unknowns. Indeed, the total number of unknowns in the BC-FEM is proportional to $O(h^{-(d-1)})$, whereas the total number of unknowns in the standard FEM behaves like $O(h^{-d})$, where h and d denote the average discretization parameter of the boundary and the spatial dimension, respectively. This idea has also been used in the related boundary element method (BEM); see the recent paper [20]. Another advantage is the lower condition number of the corresponding finite element stiffness matrices in comparison to the usual FEM, at least, for the Dirichlet problem. This property was already discovered by Yserentant [48].

Efficient solution methods for the BC-FEM have been designed in [22, 23, 10, 11]. In [23] it is shown how the Cholesky factorization of the stiffness matrix for two-dimensional problems can be performed with the complexity $O(h^{-1}\log^8 h^{-1})$. The paper [22] briefly suggests a preconditioner based on the fast realization of $H^{1/2}(\partial\Omega)$-like norms on the space of piecewise linear functions. Moreover, the paper [22] also

*Received by the editors May 15, 2007; accepted for publication (in revised form) April 2, 2008; published electronically August 1, 2008. This work was supported by the Johann Radon Institute for Computational and Applied Mathematics (RICAM) of the Austrian Academy of Sciences and by the Austrian Science Fund FWF under grant SFB F013 (subproject F1306).

http://www.siam.org/journals/sinum/46-6/69172.html

†Institute of Computational Mathematics, Johannes Kepler University Linz, Altenbergerstraße 69, A-4040 Linz, Austria (sven.beuchler@jku.at).

‡Fakultät für Mathematik, TU Chemnitz, D-09107 Chemnitz, Germany (tino.eibner@mathematik.tu-chemnitz.de). This author's work was supported by the SFB 393 "Parallel Numerical Simulation in Physics and Continuum Mechanics" of the DFG and the DFG-project "Adaptive and parallel BPX solver" under grant ME 1224/6-1.

§Institute of Computational Mathematics, Johannes Kepler University Linz and Johann Radon Institute for Compuational and Applied Mathematics, Austrian Academy of Sciences, Altenbergerstraße 69, A-4040 Linz, Austria (ulanger@numa.uni-linz.ac.at).

points out the case of Dirichlet boundary conditions as being special, since the condition number of the diagonally scaled stiffness matrix is shown to be bounded by $O(\log^{\beta} N)$ for some $\beta \geq 0$ depending on the choice of shape functions. However, these preconditioners lead to condition numbers that grow polylogarithmically in the problem size instead of the desirable optimal $O(1)$. The first preconditioners for BC-FEM that achieve optimal condition numbers are presented in [10, 11]. These preconditioners are based on the additive Schwarz framework; see, e.g., [38, 41, 47]. They are dimension independent and applicable to a wide range of boundary conditions.

Interface concentrated finite element methods (IC-FEM) are a combination of nonoverlapping domain decomposition (DD) methods with the BC-FEM that is used in the subdomains Ω_k in which the initial domain Ω is decomposed. Now we can admit piecewise smooth coefficients in the PDE that we are going to solve. An efficient method for the solution of symmetric and positive definite (SPD) systems of linear algebraic equations arising from the finite element discretization of elliptic boundary value problems is the preconditioned conjugate gradient (PCG) method with DD preconditioners. The nonoverlapping Dirichlet–Dirichlet DD preconditioning technique and the finite-element-tearing and interconnecting (FETI) method are typical representatives of primal and dual nonoverlapping DD approaches, respectively. The Dirichlet–Dirichlet DD preconditioner requires preconditioners for the local Dirichlet problems on the subdomains, a Schur complement preconditioner related to the interfaces and approximate discrete harmonic extension operators, or, more precisely, bounded discrete extension operators acting from the subdomain boundaries into the interior; see, e.g., [34, 38, 17, 18, 29] and the references therein.

Today, for preconditioning the local Dirichlet problems, a lot of very efficient techniques are available from preconditioning standard h, p, or hp finite element equations; see, e.g., [29]. The first preconditioners for the Schur complement were proposed by Dryja for some special cases in [8] and [9]. The BPS preconditioners proposed and developed by Bramble, Pasciak, and Schatz in a series of papers can be seen as a generalization to more general decompositions and discretizations [3]. During the last 20 years many alternative Schur complement preconditioners have been proposed in the literature. Here let us mention only the Bramble–Pasciak–Xu (BPX) Schur complement preconditioner [44], techniques borrowed from the BEM [32], and $\mathcal{H}$-matrix techniques [21]. In the case of uniformly refined meshes, two basic techniques for the fast evaluation of the approximate discrete harmonic extension of a finite element function on the boundary to the interior exist, namely, an averaging technique [38], [35] and hierarchical decomposition techniques [18, 15, 19].

The FETI technology introduced by Farhat and Roux in [14] has been developed to a very efficient and robust parallel solver technique for large-scale finite element equations. There are a lot of modifications and new versions of the classical one-level FETI method like the FETI-DP (dual-primal) [13] and BDDC (balanced domain decomposition by constraints) methods [6, 33]. The main building blocks of the FETI methods are solvers or preconditioners for the local Dirichlet and the local Neumann problems in the subdomains and a component managing the global information exchange. Preconditioners of Neumann–Neumann-type and FETI preconditioners for hp-FEM tensor product discretizations of elliptic boundary value problems have been proposed in [45, 46]. We again refer the reader to the monograph [47] for more information on and references to FETI methods.

In this paper, we propose and investigate fast, robust, and highly parallel DD solvers for large-scale interface concentrated finite element equations. The polynomials

of high order are removed by an overlapping preconditioner with inexact subproblem solvers which was proposed by Pavarino [42]; see also [36, 11]. This preconditioner requires

- a solver for each patch corresponding to a node, and
- a coarse space solver for $p = 1$ on the computational domain.

The first problem is treated by a direct solver, whereas the second problem is treated by two different types of DD preconditioners. The first preconditioner is a nonoverlapping DD preconditioner of Dirichlet–Dirichlet type with inexact subproblem solvers. The second type of preconditioners are all-floating (AF)-FETI preconditioners. The AF or total tearing and interconnecting techniques have independently been introduced by Of [40] and by Dostál, Horák, and Kučera [7]. We will prove that the design of both preconditioners leads to an almost optimal solver for the corresponding system of algebraic finite element equations. The complexity is almost proportional to the number of unknowns corresponding to the skeleton of the DD. Moreover, we show that the FETI preconditioners are robust with respect to jumps in the coefficients of the PDE. The preconditioners also work in the three-dimensional case. For the sake of simplicity, we restrict ourselves to the two-dimensional case. A brief discussion of the three-dimensional case is presented at the end of this paper.

The rest of the paper is organized as follows: In section 2, we formulate the boundary value problem and describe its discretization with the help of the IC-FEM. Section 3 is devoted to the primal DD preconditioners. In section 4, we investigate the FETI preconditioners. Several numerical experiments are presented in section 5.

2. Model problem and finite element discretization.

2.1. Model problem. For a polygonal Lipschitz domain Ω with

$$\overline{\Omega} = \bigcup_{i=1}^{s} \overline{\Omega}_i \subset \mathbb{R}^2, \qquad \Omega_i \cap \Omega_j = \emptyset \ \text{ with } i \neq j, \tag{2.1}$$

an SPD matrix function $\mathcal{D} \in \mathbb{R}^{2,2}$ which is piecewise constant on Ω_j and a right-hand side $f \in L^2(\Omega)$ analytic on $\overline{\Omega}_j$, we consider the following Dirichlet problem, given in weak formulation.

PROBLEM 2.1 (model problem). *Find $u \in H_0^1(\Omega)$ such that*

$$a(u,v) := \int_\Omega \nabla u(x) \cdot \mathcal{D}(x) \nabla v(x)\, dx = \int_\Omega f(x) v(x)\, dx =: l(v) \qquad \forall\, v \in H_0^1(\Omega). \tag{2.2}$$

For simplicity, throughout this paper, the subdomains Ω_j are assumed to be polygonal Lipschitz domains. Of course, weaker assumptions are possible (e.g., piecewise smooth Lipschitz domains).

2.2. Discretization: The geometric mesh, the linear degree vector, and the finite element space. We will restrict our considerations to γ-shape-regular triangulations τ of Ω consisting of affine triangles; i.e., each element $K \in \tau$ is the image $F_K(\hat{K})$ of the reference triangle $\hat{K}$, where the mapping F_K satisfies the inequality

$$h_K^{-1} \|F_K'\|_{L^\infty(K)} + h_K \|(F_K')^{-1}\|_{L^\infty(K)} \leq \gamma \qquad \forall\, K \in \tau.$$

Here h_K denotes the diameter of the element K. Moreover, we assume $\tau = \cup_{j=1}^{s} \tau^{(j)}$ with $\tau^{(j)} := \{K \in \tau \mid K \subset \Omega_j\}$ to be a geometric mesh with boundary mesh size h (see also Figure 5.1). The precise description is given in the following definition.

DEFINITION 2.2 (geometric mesh). *There exist $c_1, c_2 > 0$ s.t. for all $K \in \tau^{(j)}$:*
1. *if $\overline{K} \cap \partial\Omega_j \neq \emptyset$, then $h \leq h_K \leq c_2 h$, and*
2. *if $\overline{K} \cap \partial\Omega_j = \emptyset$, then $c_1 \inf_{x\in K} \operatorname{dist}(x, \partial\Omega_j) \leq h_K \leq c_2 \sup_{x\in K} \operatorname{dist}(x, \partial\Omega_j)$.*

In order to define *hp*-FEM spaces on the mesh τ, we associate a polynomial degree $p_K \in \mathbb{N}$ with each element $K \in \tau$ and collect these p_K in the polynomial degree vector $\mathbf{p} := (p_K)_{K\in\tau}$. Furthermore, we associate a polynomial degree

$$p_e := \min\{p_K \mid \text{e is an edge of element } K\} \tag{2.3}$$

with each edge e of the triangulation. We denote the vector containing the polynomial distribution of the triangle $K \in \tau$ with edges $\{e_i \mid i = 1, 2, 3\}$ by

$$\mathbf{p}(K) := (p_{e1}, p_{e2}, p_{e3}, p_K). \tag{2.4}$$

In conjunction with geometric meshes, a particularly useful polynomial degree distribution is the linear degree vector.

DEFINITION 2.3 (linear degree vector). *Let $\overline{\Omega} = \cup_{j=1}^s \overline{\Omega}_j$. For all $j = 1, \ldots, s$, let $\tau^{(j)}$ be a geometric mesh on Ω_j with boundary mesh size h, and let $\tau = \cup_{j=1}^s \tau^{(j)}$ be a mesh on Ω. A polynomial degree vector $\mathbf{p} = (p_K)_{K\in\tau}$ is said to be a linear degree vector with slope $\alpha > 0$ if*

$$1 + \alpha c_1 \log\frac{h_K}{h} \leq p_K \leq 1 + \alpha c_2 \log\frac{h_K}{h} \tag{2.5}$$

for some $c_1, c_2 > 0$.

Now we are in a position to define our *hp*-FEM spaces.

DEFINITION 2.4 (FEM spaces). *Let $\overline{\Omega} = \cup_{j=1}^s \overline{\Omega}_j$. For all $j = 1, \ldots, s$, let $\tau^{(j)}$ be a geometric mesh on Ω_j with boundary mesh size h, and let $\tau = \cup_{j=1}^s \tau^{(j)}$ be a mesh on Ω. Let $\mathbf{p}$ be a linear degree vector. Furthermore, for all edges e, let p_e be given by* (2.3) *and for all $K \in \tau$ let $\mathbf{p}(K)$ be given by* (2.4)*. Then we set*

$$\begin{aligned}
\mathbb{S}_j^{\mathbf{p}} = \mathbb{S}^{\mathbf{p}}(\Omega_j, \tau^{(j)}) &:= \{u \in H^1(\Omega_j) \mid u \circ F_K \in \mathcal{P}_{\mathbf{p}(K)}(\hat{K}) \quad \forall K \in \tau^{(j)}\},\\
\mathbb{S}_{j,0}^{\mathbf{p}} = \mathbb{S}_0^{\mathbf{p}}(\Omega_j, \tau^{(j)}) &:= \mathbb{S}_j^{\mathbf{p}} \cap H_0^1(\Omega_j),\\
\mathbb{V}^{\mathbf{p}} = \mathbb{V}^{\mathbf{p}}(\Omega, \tau) &:= \{u \in H^1(\Omega_j) \mid u|_{\Omega_j} \in \mathbb{S}_j^{\mathbf{p}} \quad \forall j = 1, \ldots, s\},\\
\mathbb{V}_0^{\mathbf{p}} &:= \mathbb{V}^{\mathbf{p}} \cap H_0^1(\Omega),
\end{aligned}$$

where

$$\mathcal{P}_{\mathbf{p}(K)}(\hat{K}) := \{u \in \mathcal{P}_{p_K}(\hat{K}) \mid u|_{e_i} \in \mathcal{P}_{p_{e_i}},\ i = 1, 2, 3\}.$$

The finite element discretization of Problem 2.1 then reads as follows.

PROBLEM 2.5 (finite element approximation). *Find $u_h \in \mathbb{V}_0^{\mathbf{p}}$ such that*

$$a(u_h, v_h) = l(v_h) \qquad \forall\, v_h \in \mathbb{V}_0^{\mathbf{p}}.$$

Considering the approximation properties of the BC-*hp*-FEM, the optimality of the convergence order can be proved in the two-dimensional case.

THEOREM 2.6. *Let $\Omega_i \subset \mathbb{R}^2$ and τ_i be a geometric mesh with boundary mesh size h on Ω_i. Let $\mathbf{p}$ be a linear degree vector with slope α sufficiently large. Then for $u \in H^{1+\delta}(\Omega_i)$, $\delta \in (0, 1)$, we have*

$$\inf_{v_h\in\mathbb{S}_i^{\mathbf{p}}} \|u - v_h\|_{H^1(\Omega_i)} \leq Ch^{\delta}.$$

Proof. See [22]. ∎

2.3. The system of finite element equations. To solve Problem 2.5 numerically, we equip the space $\mathbb{V}_0^{\mathbf{p}}$ with a basis

$$\Phi = \{\Phi_B, \Phi_L, \Phi_H\} = \{\phi_i \mid i = 1, \dots, N\} \tag{2.6}$$

in the following way.

The functions $\phi_1, \dots, \phi_{N_B}$ are the usual nodal hat functions with $\operatorname{supp}\phi_i \not\subset \overline{\Omega}_j$ for all $j = 1, \dots, s$, for all $i = 1, \dots, N_B$; i.e., the support is not contained in one subdomain $\overline{\Omega}_j$ only. The functions $\phi_{N_B+1}, \dots, \phi_{N_L}$ are the usual nodal hat functions with $\operatorname{supp}\phi_i \subset \overline{\Omega}_j$ for some j, for all $i = N_B + 1, \dots, N_L$. The remaining functions $\phi_{N_L+1}, \dots, \phi_{N_H}$ are all polynomials of degree $p > 1$. Since $h_K \sim h$ for all elements with $\overline{K} \cap \partial\Omega_j \neq \emptyset$, we can assume that $p = 1$ for all these boundary elements by (2.5). With this definition, the basis functions ϕ_i can be divided into three groups:

- the boundary functions $\Phi_B = \{\phi_i \mid i = 1, \dots, N_B\}$,
- the interior functions of low order $\Phi_L = \{\phi_i \mid i = N_B + 1, \dots, N_L\}$, and
- the interior functions of high order $\Phi_H = \{\phi_i \mid i = N_L + 1, \dots, N_H\}$.

In the same way, the basis $\Phi_i = \{\phi_{j,i}\}_j$ on the subdomains Ω_i, $i = 1, \dots, s$, are introduced. Analogously, this basis can be partitioned into $\Phi_{B,i}$, $\Phi_{L,i}$, and $\Phi_{H,i}$, i.e., boundary functions (B), interior functions of low order (L), and interior functions of high order (H). More precisely, let

$$\Phi_i = \Phi\,|_{\Omega_i}\, R_i \quad \text{and} \quad \Phi_{B,i} = \Phi_B\,|_{\Omega_i}\, R_{B,i} \tag{2.7}$$

with the corresponding restriction matrices R_i and $R_{B,i}$. Now, the solution of Problem 2.5 is equivalent to solving the system of linear equations

$$\mathcal{K}\underline{u} = \underline{f}, \quad \text{where} \quad \mathcal{K} = [a(\phi_j, \phi_k)]_{k,j=1}^N, \qquad \underline{f} = [l(\phi_j)]_{j=1}^N, \qquad \underline{u} = [u_k]_{k=1}^N. \tag{2.8}$$

Then, $u_h = \sum_{k=1}^N u_k \phi_k$ provides the solution of Problem 2.5. Moreover, let $\mathcal{K}_i$ be the stiffness matrix restricted to Φ_i, $i = 1, \dots, s$, i.e.,

$$\mathcal{K}_i = [a(\phi_{j,i}, \phi_{k,i})]_{k,j=1}^{N_i}. \tag{2.9}$$

Then, a consequence of (2.7)–(2.9) is the formula

$$\mathcal{K} = \sum_{i=1}^{s} R_i \mathcal{K}_i R_i^\top \tag{2.10}$$

with the restriction matrix introduced in (2.7). In the following, we are interested in constructing a fast solver for the system of linear algebraic equations (2.8).

3. Primal iterative substructuring methods.

3.1. Block partitioning of the stiffness matrix. We partition the stiffness matrix into three-by-three blocks, which corresponds to the partition of the shape functions into Φ_B, Φ_L, and Φ_H (see (2.6)),

$$\mathcal{K} = \begin{bmatrix} \mathcal{K}_B & \mathcal{K}_{B,L} & \mathcal{K}_{B,H} \\ \mathcal{K}_{L,B} & \mathcal{K}_L & \mathcal{K}_{L,H} \\ \mathcal{K}_{H,B} & \mathcal{K}_{H,L} & \mathcal{K}_H \end{bmatrix} = \begin{bmatrix} \mathcal{K}_B & \mathcal{K}_{B,D} \\ \mathcal{K}_{D,B} & \mathcal{K}_D \end{bmatrix},$$

with $D = L \cup H$. This two-by-two block matrix can be factorized into

$$\mathcal{K} = \begin{bmatrix} I & \mathcal{K}_{B,D}\mathcal{K}_D^{-1} \\ \mathbf{0} & I \end{bmatrix} \begin{bmatrix} \mathcal{S}_B & \mathbf{0} \\ \mathbf{0} & \mathcal{K}_D \end{bmatrix} \begin{bmatrix} I & \mathbf{0} \\ \mathcal{K}_D^{-1}\mathcal{K}_{D,B} & I \end{bmatrix} \tag{3.1}$$

with the Schur complement

$$\mathcal{S}_B = \mathcal{K}_B - \mathcal{K}_{B,D}\mathcal{K}_D^{-1}\mathcal{K}_{D,B}. \tag{3.2}$$

In this section, we will investigate preconditioners of the type

$$\mathcal{C} = \begin{bmatrix} I & -\mathcal{E}_{DB}^\top \\ \mathbf{0} & I \end{bmatrix} \begin{bmatrix} \mathcal{C}_S & \mathbf{0} \\ \mathbf{0} & \mathcal{C}_D \end{bmatrix} \begin{bmatrix} I & \mathbf{0} \\ -\mathcal{E}_{DB} & I \end{bmatrix}, \tag{3.3}$$

where $\mathcal{C}_D$ is a preconditioner for $\mathcal{K}_D$, $\mathcal{E}_{DB}$ is the matrix representation of the extension operators E_j acting from the interface $\Gamma = \cup_{i=1}^s \partial\Omega_i \backslash \partial\Omega$ into the interior of the domains Ω_j, and $\mathcal{C}_S$ is a preconditioner for the Schur complement $\mathcal{S}_B$ or the inexact Schur complement

$$\mathcal{S}_B + \mathcal{T}_B := \mathcal{S}_B + (\mathcal{E}_{DB}^\top - \mathcal{H}_{DB}^\top)\mathcal{K}_D(\mathcal{E}_{DB} - \mathcal{H}_{DB}) \tag{3.4}$$

with $\mathcal{H}_{DB} = \mathcal{K}_D^{-1}\mathcal{K}_{DB}$.

The following result is the key in analyzing the preconditioner $\mathcal{C}$, which generalizes previous estimates given in [16, 39].

LEMMA 3.1. *Let $\mathcal{C}_S$ and $\mathcal{C}_D$ be SPD preconditioners for $\mathcal{S}_B$ and $\mathcal{K}_D$, i.e.,*

$$c_S\,(\mathcal{C}_S\underline{v},\underline{v}) \le (\mathcal{S}_B\underline{v},\underline{v}) \quad \forall \underline{v} \in \mathbb{R}^{M_c}, \tag{3.5}$$

$$c_I\,(\mathcal{C}_D\underline{v},\underline{v}) \le (\mathcal{K}_D\underline{v},\underline{v}) \le C_I\,(\mathcal{C}_D\underline{v},\underline{v}) \quad \forall \underline{v} \in \mathbb{R}^{M_i}. \tag{3.6}$$

Moreover, let

$$\left(\mathcal{K}\begin{bmatrix} I \\ \mathcal{E}_{DB} \end{bmatrix}\underline{g}, \begin{bmatrix} I \\ \mathcal{E}_{DB} \end{bmatrix}\underline{g}\right) \le c_E^2\left(\mathcal{C}_S\underline{g},\underline{g}\right) \quad \forall \underline{g} \in \mathbb{R}^{M_c}. \tag{3.7}$$

Then, $c_E^2 \ge c_S$ and the inequalities

$$c\,(\mathcal{C}\underline{v},\underline{v}) \le (\mathcal{K}\underline{v},\underline{v}) \le C\,(\mathcal{C}\underline{v},\underline{v}) \quad \forall \underline{v} \in \mathbb{R}^M$$

hold with $c = 0.5\,c_I \min\{1, c_S/(c_I + c_E^2 - c_S)\}$ and $C = 2\max\{c_E^2, C_I\}$.

Proof. The proof has been given in the report [2]. ∎

Remark 3.2. A direct consequence of (3.7) is the spectral equivalence of the Schur complement $\mathcal{S}_B$ and the perturbed Schur complement $\mathcal{S}_B + \mathcal{T}_B$; see [16].

Summarizing, we have to find the preconditioners $\mathcal{C}_D$ and $\mathcal{C}_S$ as well as the extension $\mathrm{E} \leftrightarrow \mathcal{E}_{BD}$ in order to derive a good preconditioner $\mathcal{C}$.

For the AF-FETI2 and AF-FETI3 preconditioners (see section 4), a preconditioner $\mathcal{C}_i$ for $\mathcal{K}_i$ (cf. (2.9)), i.e., for the Neumann problem on domain Ω_i, is required. Due to (2.7), the block partitioning for $\mathcal{K}$ in (3.1) can be done for the subdomain stiffness matrices, too. A simple computation shows

$$\mathcal{K}_i = \begin{bmatrix} I & \mathcal{K}_{B,D,i}\mathcal{K}_{D,i}^{-1} \\ \mathbf{0} & I \end{bmatrix} \begin{bmatrix} \mathcal{S}_{B,i} & \mathbf{0} \\ \mathbf{0} & \mathcal{K}_{D,i} \end{bmatrix} \begin{bmatrix} I & \mathbf{0} \\ \mathcal{K}_{D,i}^{-1}\mathcal{K}_{D,B,i} & I \end{bmatrix} \tag{3.8}$$

with the local Schur complement

$$\mathcal{S}_{B,i} = \mathcal{K}_{B,i} - \mathcal{K}_{B,D,i}\mathcal{K}_{D,i}^{-1}\mathcal{K}_{D,B,i}. \tag{3.9}$$

The global Schur complement can be computed via the local Schur complements, i.e.,

$$\mathcal{S}_B = \sum_{i=1}^{s} R_{B,i}\mathcal{S}_{B,i}R_{B,i}^\top \tag{3.10}$$

with the assembling matrices $R_{B,i}$. The DD approach can also be applied to $\mathcal{K}_i$ instead of $\mathcal{K}$. In the following subsections, we will develop the preconditioners for the subproblems.

3.2. An overlapping preconditioner for the Dirichlet subproblems. In this subsection, we present the preconditioner for the matrix $\mathcal{K}_D$. Since $\mathcal{K}_D = \text{blockdiag}\,[\mathcal{K}_{D,j}]_j$, we restrict ourselves to the case of one subdomain Ω_j.

Here, we use an overlapping preconditioner which has been developed in [36]; see also [11, 42]. The preconditioner is based on the additive Schwarz framework (see, e.g., [47]). We decompose the *hp*-FEM space $\mathbb{S}^{\mathbf{p}}_{j,0} = \sum_{i=0}^{\#K} \mathbb{V}_i$ into subspaces $\mathbb{V}_i$ that will be specified below. This splitting defines an additive Schwarz method (ASM) preconditioner. This eliminates all high order degrees of freedom. Moreover, the properties of this preconditioner can be characterized by the two numbers C_0 and $\rho(\epsilon)$ defined below. As a well-known fact from ASM theory, the condition number of the preconditioned system is bounded by $C_0^2(1+\rho(\epsilon))$.

In order to keep the notation simple, we skip the index j for the subdomain j.

LEMMA 3.3 (Clement interpolation). *Let τ be a γ-shape-regular affine triangulation. For $K \in \tau$ define the patch*

$$\omega_K := \cup\{K' \in \tau \mid \overline{K} \cap \overline{K'} \neq \emptyset\}.$$

Then there exist $C > 0$ depending solely on γ and a linear operator $I^C : H^1(\Omega) \mapsto \mathbb{S}^{\mathbf{1}}(\tau)$ such that

$$\|u - I^C u\|_{L^2(K)} \leq C h_K \|\nabla u\|_{L^2(\omega_K)} \quad \text{and} \quad \|\nabla I^C u\|_{L^2(K)} \leq C \|\nabla u\|_{L^2(\omega_K)}.$$

Proof. The proof can be found in several papers; see, e.g., [12]. ∎

LEMMA 3.4. *Let $\hat{K}$ be the reference triangle and $\mathbf{p}(K) = (p_{e_1}, p_{e_2}, p_{e_3}, p_K)$ with $p_{e_i} \leq p_K$, $i = 1, \ldots, 3$, a polynomial degree distribution. Let $\mathcal{J}_p$ be the one-dimensional Gauss–Lobatto interpolation operator. Then there exist a constant $C > 0$ and a linear operator*

$$I_{\mathbf{p}(\mathbf{K})} : \mathcal{P}_{2p_K}(\hat{K}) \mapsto \mathcal{P}_{\mathbf{p}}(\hat{K}) \tag{3.11}$$

such that

- $I_{\mathbf{p}(\mathbf{K})} u = u$ *for all* $u \in \mathcal{P}_{\mathbf{p}}(\hat{K})$,
- $\|I_{\mathbf{p}(\mathbf{K})} u\|_{H^1(\hat{K})} \leq C \|u\|_{H^1(\hat{K})}$ *for all* $u \in \mathcal{P}_{2p_K}(\hat{K})$, *and*
- $(I_{\mathbf{p}(\mathbf{K})} u)|_{\Gamma_i} = \mathcal{J}_{p_{e_i},\Gamma_i} u$ *for* $i = 1, \ldots, 3$.

Proof. The proof is given in [12, Theorem 4.4]. ∎

DEFINITION 3.5. *Let τ be a γ-shape-regular triangulation with polynomial degree distribution $\mathbf{p} = (p_K)_{K\in\tau}$. For each $K \in \tau$ let the operator $I_{\mathbf{p}(\mathbf{K})}$ be given by* (3.11). *Then we define*

$$I_{\mathbf{p}} : \mathbb{S}^{2\mathbf{p}}(\tau) \mapsto \mathbb{S}^{\mathbf{p}}(\tau) \qquad \text{by} \qquad (I_{\mathbf{p}} u)|_K = (I_{\mathbf{p}}(u|_K \circ F_K)) \circ F_K^{-1}.$$

For a finite element mesh τ, we denote the set of its vertices by $V(\tau)$ and the polynomial degree vector by $\mathbf{p} = (p_K)_{K\in\tau}$. In order to define our ASM preconditioner, we associate with each vertex v_j a patch

$$\omega_j := \cup\{K \in \tau \mid v_j \text{ is a vertex of } K\}$$

and decompose the space $\mathbb{S}^{\mathbf{p}}(\tau)$ into the following subspaces:

1. the space of piecewise linear polynomials $\mathbb{V}_0 := \mathbb{S}^{\mathbf{1}}(\tau)$, and
2. local higher dimensional spaces $\mathbb{V}_j := \{u \in \mathbb{S}^{\mathbf{p}}(\tau) \mid \text{supp}\, u \subset \overline{\omega}_j\}$.

To analyze the properties of the decomposition of $\mathbb{S}^{\mathbf{p}}(\tau)$, let C_0 be a constant such that

$$\min\left\{\sum_{j=0}^{\#V(\tau)}\|u_j\|_A^2 \;\middle|\; u=u_0+\sum_{j=1}^{\#V(\tau)}u_j,\;\; u_j\in\mathbb{V}_j\right\}\le C_0^2\|u\|_A^2 \tag{3.12}$$

holds, where $\|u\|_A^2 = a(u,u)$ denotes the energetic norm. Moreover, we define a symmetric matrix $\epsilon\in\mathbb{R}^{(\#V(\tau))\times(\#V(\tau))}$ with entries $\epsilon_{ik}=0$ if ω_i and ω_k are disjoint and $\epsilon_{ik}=1$ otherwise. Let $\rho(\epsilon)$ be the spectral radius of ϵ. Thus

$$|a(u_i,u_k)|\le\epsilon_{ik}\|u_i\|_A\|v_k\|_A \qquad \forall u_i\in\mathbb{V}_i,\ \forall u_k\in\mathbb{V}_k. \tag{3.13}$$

THEOREM 3.6. *Let τ be a γ-shape-regular triangulation and $\mathbf{p}$ a corresponding polynomial degree distribution. Let $\mathbb{S}^{\mathbf{p}}(\tau)=\mathbb{V}_0+\sum_{j=1}^{\#V(\tau)}\mathbb{V}_j$, and let $a(\cdot,\cdot):H^1\times H^1\mapsto\mathbb{R}$ be an SPD bilinear form. Then there exists a constant $C>0$ depending solely on the bilinear form $a(\cdot,\cdot)$ and γ such that $C_0\le C$ and $\rho(\epsilon)\le C$, where C_0 and ϵ are from* (3.12) *and* (3.13)*, respectively.*

Proof. Since there exists $M>0$, depending only on γ, such that

$$\#\{v'\in V\mid\overline{\omega}_v\cap\overline{\omega}_{v'}\neq\emptyset\}\le M \qquad \forall v\in V,$$

we have $\rho(\epsilon)\le\|\epsilon\|_\infty\le M$. In order to bound C_0, we exploit the properties of the operators I^C and $I_{\mathbf{p}}$ and decompose $u\in\mathbb{S}^{\mathbf{p}}(\tau)$ as

$$u=I^Cu+I_{\mathbf{p}}(u-I^Cu) \;=\; I^Cu+I_{\mathbf{p}}\left(\sum_{j=1}^{\#V(\tau)}\phi_j(u-I^Cu)\right),$$

where ϕ_j denotes the usual nodal hat function corresponding to the vertex $v_j\in V(\tau)$. Thus we have

$$u=u_0+\sum_{j=1}^{\#V(\tau)}u_j \qquad \text{with } u_0:=I^C(u)\in\mathbb{V}_0,\quad u_j:=I_{\mathbf{p}}(\phi_j(u-I^Cu))\in\mathbb{V}_j.$$

Now we want to bound $\|u_0\|_A^2+\sum_{j=1}^{\#V(\tau)}\|u_j\|_A^2$. Lemma 3.3 and the Poincaré–Friedrichs inequality $\|\cdot\|_A\sim\|\cdot\|_{H^1}$ imply the estimate

$$\|u_0\|_A^2=\left\|I^C(u)\right\|_A^2\le C\|u\|_A^2.$$

Next we consider a fixed $j\in1,\dots,\#V(\tau)$ and a fixed $K\subset\omega_j$ such that

$$\|u_j|_K\|_A^2\le C\|u_j\|_{H^1(K)}^2=\left\|I_{\mathbf{p}}(\phi_j(u-I^Cu))\right\|_{H^1(K)}^2.$$

Denoting $\tilde{u}:=(u-I^Cu)$, $\hat{u}:=\tilde{u}|_K\circ F_K$, and $\hat{\phi}:=\phi_j|_K\circ F_K$ and applying Lemmas 3.4 and 3.3, we get the estimates

$$\begin{aligned}
\|u_j\|_{H^1(K)}^2 &\le C\left(h_K^2\|I_{\mathbf{p}}(\hat{\phi}\hat{u})\|_{L^2(\hat{K})}^2+\|\nabla I_{\mathbf{p}}(\hat{\phi}\hat{u})\|_{L^2(\hat{K})}^2\right)\\
&\le C\|I_{\mathbf{p}}(\hat{\phi}\hat{u})\|_{H^1(\hat{K})}^2\\
&\le C\|\hat{\phi}\hat{u}\|_{H^1(\hat{K})}^2\\
&\le C\|\hat{u}\|_{H^1(\hat{K})}^2\\
&\le C\left(h_K^{-2}\|u-I^Cu\|_{L^2(K)}^2+\|\nabla(u-I^Cu)\|_{L^2(K)}^2\right)\\
&\le C\|\nabla u\|_{L^2(\omega_j)}^2.
\end{aligned}$$

Thus, the inequalities

$$\|u_j\|^2_{H^1(\omega_j)} \le C \sum_{K\subset\omega_j} \|\nabla u\|^2_{L^2(\omega_j)} \text{ and } \sum_{j=1}^{\#V(\tau)} \|u_j\|^2_{H^1(\omega_j)} \le C\|\nabla u\|^2_{L^2(\Omega)} \le C\|u\|^2_A$$

follow immediately. ∎

Now we are able to define the following preconditioner for $\mathcal{K}_{D,i}$. Let

$$\mathcal{C}_D^{-1} = \sum_{j=1}^{\#V(\tau)} Q_j^\top \mathcal{K}_{P,j}^{-1} Q_j + Q_0^\top (\mathcal{D}_L)^{-1} Q_0, \tag{3.14}$$

where

- $\mathcal{K}_{P,j}$ denotes the stiffness matrix restricted to $\mathbb{V}_j$, $j = 1, \dots, \#V(\tau)$,
- Q_j, $j = 0, \dots, \#V(\tau)$, is the corresponding finite element restriction matrix onto $\mathbb{V}_j$, and
- $\mathcal{D}_L$ is the diagonal part of the matrix $\mathcal{K}_L$.

THEOREM 3.7. *Let $\mathcal{C}_D$ be defined via* (3.14). *Then, we have $\mathcal{C}_D \sim \mathcal{K}_D$.*

Proof. Due to [48] and [22], the condition number of $\mathcal{K}_L$ does not depend on the discretization parameter. This gives $\mathcal{D}_L \sim \mathcal{K}_L$. Now, the result is a consequence of Theorem 3.6. ∎

3.3. Schur complement preconditioner and extension operator. Next, we consider the preconditioner $\mathcal{C}_S$ for $\mathcal{S}_B$ (3.2). We will prove that the norm induced by the Schur complement $\mathcal{S}_B$ (3.2) is equivalent to a suitable seminorm. Therefore, the extension operators E for the nearly discrete harmonic extension of a function $g \in H^{1/2}(\partial\tilde{\Omega})$ to a function $u = \mathrm{E}g \in H^1(\tilde{\Omega})$ have to be investigated. We follow the approach of Haase and Nepomnyaschikh [19].

In a first step, we describe the extension of Haase and Nepomnyaschikh. Then, we prove that this type of extension can also be used for boundary concentrated meshes with $p = 1$ or $p > 1$ defined via Definition 2.3. Finally, the equivalence of the Schur complement is shown.

The technique of Haase and Nepomnyaschikh was originally derived for the h-version of the FEM with $p = 1$, using uniform refined meshes and the hierarchy of the nested finite element spaces. Let $\tau_0 \subset \tau_1 \subset \tau_2 \subset \cdots \subset \tau_L := \tau$ be a sequence of nested triangular ($d = 2$) or tetrahedral ($d = 3$) finite element meshes with nodes $x_i^{(l)}$, $i = 1, \dots, N_l$, $l = 0, \dots, L$, on a domain $\tilde{\Omega}$. Here τ_0 is the coarse mesh and τ_l is obtained from τ_{l-1} by subdividing each element of τ_{l-1} into 2^d new congruent elements.

Due to Definition 2.4, $\mathbb{S}_l^{\mathbf{1}}$ denotes the space of all functions which are linear on each element of τ_l and belong to $H^1(\tilde{\Omega})$. The corresponding trace space is denoted by $\mathbb{B}_l$, i.e., $\mathbb{B}_l = \mathbb{S}_l^{\mathbf{1}}\,|_{\partial\tilde{\Omega}}$. This space is spanned by nodal basis functions $\psi_{i,l}$, i.e.,

$$\mathbb{B}_l = \operatorname{span}\{\psi_{i,l}\}_{i=1}^{N_l} = \operatorname{span}\Psi_l. \tag{3.15}$$

Moreover, let $\mathbb{W}_l^{(i)} = \operatorname{span}\{\psi_{i,l}\}$ be the one-dimensional subspace corresponding to the node $x_i^{(l)}$. Let $g \in \mathbb{B}_l \subset H^{1/2}(\partial\tilde{\Omega})$. In order to define a function $u = \mathrm{E}g$ such that

$$\| u \|_{H^1(\tilde{\Omega})} \le c \, \| g \|_{H^{1/2}(\partial\tilde{\Omega})}, \quad u\,|_{\partial\tilde{\Omega}} = g, \tag{3.16}$$

Haase and Nepomnyaschikh used a BPX-like decomposition of the function g. Let $Q_{i,l} : L_2(\partial\tilde{\Omega}) \mapsto \mathbb{W}_l^{(i)}$ be the L_2-like projection from $L_2(\partial\tilde{\Omega})$ onto $\mathbb{W}_l^{(i)}$, i.e.,

$$Q_{i,l}g = d_{i,l}(g)\psi_{i,l} \quad \forall g \in \mathbb{B}_l.$$

Possible choices of $d_{i,l}$ are $d_{i,l}(g) = \frac{(g,\psi_{i,l})_{L_2(\partial\tilde{\Omega})}}{(\psi_{i,l},\mathbf{1})_{L_2(\partial\tilde{\Omega})}}$ or $d_{i,l}(g) = \frac{(g,\mathbf{1})_{L_2(\operatorname{supp}\psi_{i,l})}}{(\mathbf{1},\mathbf{1})_{L_2(\operatorname{supp}\psi_{i,l})}}$. Moreover, let

$$Q_l = \sum_{i=1}^{N_l} Q_{i,l}, \quad l = 0, \ldots, L, \quad \text{and} \quad Q_{-1} = \mathbf{0}.$$

Now, the extension of a function $g \in \mathbb{B}_L$ into $\tilde{\Omega}$ can be defined.

ALGORITHM 3.8.

1. *Let* $\psi_l = (Q_l - Q_{l-1})g \in \mathbb{B}_l$, $l = 0, \ldots, L$.
2. *Define the function* $u_l \in \mathbb{S}_l^{\mathbf{1}}$ *via*

$$u_l(x_i^{(0)}) = \begin{cases} \psi_l(x_i^{(0)}) & \text{if} \quad x_i^{(0)} \in \partial\tilde{\Omega}, \\ \overline{g} & \text{if} \quad x_i^{(0)} \notin \partial\tilde{\Omega}, \end{cases} \tag{3.17}$$

$$u_l(x_i^{(l)}) = \begin{cases} \psi_l(x_i^{(l)}) & \text{if} \quad x_i^{(l)} \in \partial\tilde{\Omega}, \\ 0 & \text{if} \quad x_i^{(l)} \notin \partial\tilde{\Omega}, \end{cases} \tag{3.18}$$

   where $\overline{g}$ *is defined in Remark* 3.9.
3. *Set* $u = \mathrm{E}g = u_0 + u_1 + \cdots + u_L$.

Remark 3.9. Note that the function u_l in (3.17) and (3.18) is uniquely defined by setting the function values in the nodes $x_i^{(l)}$ of the mesh τ_l. For $\overline{g}$, a coarse grid problem on the mesh τ_0 has to be solved. Alternatively, the mean value of g over $\partial\tilde{\Omega}$ can be taken.

Now, the following result has been shown.

THEOREM 3.10. *Let* $\tau_0 \subset \cdots \subset \tau_L = \tau$ *be a family of uniformly refined meshes, and let* $\mathbb{S}_l^{\mathbf{1}}$, $l = 0, \ldots, L$, *be the space of the piecewise linear nodal basis functions on* τ_l. *Let* $\mathrm{E} : \mathbb{B}_L \mapsto \mathbb{S}^{\mathbf{1}} = \mathbb{S}_L^{\mathbf{1}}$ *be the extension of Algorithm* 3.8. *Then, we have* $\mathrm{E}g \mid_{\partial\tilde{\Omega}} = g$. *Moreover, there is a constant* c_2 *which is independent of* L *such that*

$$\| \mathrm{E}g \|_{H^1(\tilde{\Omega})} \leq c_2 \| g \|_{H^{1/2}(\partial\tilde{\Omega})} \quad \forall g \in \mathbb{B}_L.$$

Proof. The proof is given in [19]. ∎

Next, we consider a family of boundary concentrated finite element meshes and the corresponding finite element spaces $\mathbb{S}_l^{\mathbf{p}}$.

THEOREM 3.11. *Let* $\tau_0 \subset \cdots \subset \tau_L = \tau$ *be a family of meshes on a domain* $\tilde{\Omega}$ *which are geometrically refined to the boundary. Let* $\mathbb{S}_l^{\mathbf{p}}$, $l = 0, \ldots, L$, *be the finite element space of Definition* 2.4 *for the mesh* τ_l. *Then,* $\mathbb{S}^{\mathbf{p}} \mid_{\partial\tilde{\Omega}} = \mathbb{B}_L$. *Moreover, let* $\mathrm{E} : \mathbb{B}_L \mapsto \mathbb{S}^{\mathbf{p}}$ *be the extension operator of Algorithm* 3.8. *Then, there exists a constant* c_2 *which is independent of* L *(and* N*) such that*

$$\| \mathrm{E}g \|_{H^1(\tilde{\Omega})} \leq c_2 \| g \|_{H^{1/2}(\partial\tilde{\Omega})} \quad \forall g \in \mathbb{B}_L. \tag{3.19}$$

Proof. Due to (2.5), the polynomial degree on the elements which touch $\partial\tilde{\Omega}$ can be assumed to be 1. Thus, $\mathbb{S}^{\mathbf{p}} \mid_{\partial\tilde{\Omega}} = \mathbb{B}_L$. This proves the first assertion.

Due to (3.17) and (3.18), the image of the extension operator described in Algorithm 3.8 for uniformly refined meshes belongs to the space $\mathbb{S}^{\mathbf{1}}$, i.e., $\mathrm{E}g \in \mathbb{S}^{\mathbf{1}}$. Theorem 3.10 implies the estimate

$$\| \mathrm{E}g \|_{H^1(\tilde{\Omega})} \leq c_2 \| g \|_{H^{1/2}(\partial\tilde{\Omega})} \quad \forall g \in \mathbb{B}_L$$

with the constant of Theorem 3.10. Since $\mathbb{S}^{\mathbf{1}} \subset \mathbb{S}^{\mathbf{p}}$, the second assertion follows. ∎

Remark 3.12. If $\int_{\partial\tilde{\Omega}} g = 0$, or $g\,|_{\tilde{\Gamma}} = 0$ for $\tilde{\Gamma} \subset \partial\tilde{\Omega}$ with $\mathrm{meas}(\tilde{\Gamma}) > 0$, relation (3.19) remains valid if the seminorms are used instead of the norms in (3.19). This is a consequence of the Poincaré–Friedrichs-type inequalities; see [47, Lemma A17] or [37].

Hence, we have shown that the extension of Haase and Nepomnyaschikh for uniformly refined meshes can also be used for geometrically refined meshes with grading factor 0.5.

In the next theorem, we show that the norm induced by the Schur complement $\mathcal{S}_{B,i}$ (3.9) is equivalent to the $H^{1/2}(\partial\Omega_i)$ norm and that the norm induced by the Schur complement $\mathcal{S}_B$ (3.2) is equivalent to the following norm on the interface $\Gamma = \cup_{i=1}^{s} \partial\Omega_i \backslash \partial\Omega$:

$$\| g \|^2_{H^{1/2}(\Gamma)} := \| g \|^2_{L^2(\Gamma)} + |g|^2_{H^{1/2}(\Gamma)}, \quad \text{where} \quad |g|^2_{H^{1/2}(\Gamma)} = \inf_{\substack{u|_\Gamma = g \\ u|_{\partial\Omega} = 0}} |u|^2_{H^1(\Omega)}.$$

THEOREM 3.13. *Let $\tau_0^{(i)} \subset \cdots \subset \tau_L^{(i)} = \tau^{(i)}$ be a family of geometrically refined meshes on the domains Ω_i, $i = 1, \ldots, s$, and let $\mathbb{B}_i$ and $\mathbb{B}_\Gamma$ be the trace spaces of $\mathbb{S}^{\mathbf{p}}$ and $\mathbb{V}_0$ onto $\partial\Omega_i$ and the interface $\Gamma = \cup_{i=1}^{s} \partial\Omega_i \backslash \partial\Omega$, respectively. Let $\Phi_{B,i}$ and Φ_B be defined via* (2.6) *and* (2.7)*, respectively.*

Moreover, let $\mathcal{S}_{B,i}$ be defined via (3.9)*. If $\partial\Omega \cap \partial\Omega_i \neq \emptyset$, we have*

(3.20)
$$c_4 \, |g_i|^2_{H^{1/2}(\partial\Omega_i \cap \Gamma)} \leq (\mathcal{S}_{B,i} \underline{g}_i, \underline{g}_i) \leq c_5 \, |g_i|^2_{H^{1/2}(\partial\Omega_i \cap \Gamma)} \quad \forall g_i = \Phi_{B,i} \underline{g}_i \in \mathbb{B}_i, g\,|_{\partial\Omega} = 0.$$

If $\partial\Omega \cap \partial\Omega_i = \emptyset$, we have

$$c_4 \, |g_i|^2_{H^{1/2}(\partial\Omega_i \cap \Gamma)} \leq (\mathcal{S}_{B,i} \underline{g}_i, \underline{g}_i) \leq c_5 \, |g_i|^2_{H^{1/2}(\partial\Omega_i \cap \Gamma)} \quad \forall g_i = \Phi_{B,i} \underline{g}_i \in \mathbb{B}_i. \tag{3.21}$$

Moreover, the relation

$$c_6 \| g \|^2_{H^{1/2}(\Gamma)} \leq (\mathcal{S}_B \underline{g}, \underline{g}) \leq c_7 \| g \|^2_{H^{1/2}(\Gamma)} \quad \forall g = \Phi_B \underline{g} \in \mathbb{B}_\Gamma \tag{3.22}$$

holds. The constants c_4, c_5, c_6, and c_7 are independent of L and N.

Proof. The proof of (3.20) is adapted from [47]; see also [1]. By the trace theorem and the Poincaré–Friedrichs inequality, one can conclude that

$$c_{14} c_T^{-2} |g_i|^2_{H^{1/2}(\partial\Omega_i)} \leq c_{14} \| u \|^2_{H^1(\Omega_i)} \leq c_{15} |u|^2_{H^1(\Omega_i)} \leq \min_{\underline{u}_D} (\mathcal{K}_i \underline{u}, \underline{u}) = (\mathcal{S}_{B,i} \underline{g}_i, \underline{g}_i),$$

with $\underline{u}^\top = \left[\begin{array}{cc} \underline{g}_i^\top & \underline{u}_D^\top \end{array} \right]$ and $u = \Phi_i \underline{u}$. Due to Theorem 3.11 and Remark 3.12, there exists a function $u \in H^1(\Omega_i)$, $u\,|_{\partial\Omega_i} = g$, such that $|u|_{H^1(\Omega_i)} \leq c_2 |g|_{H^{1/2}(\partial\Omega_i)}$. Thus, we can conclude that

$$(\mathcal{S}_{B,i} \underline{g}_i, \underline{g}_i) \leq (\mathcal{K}_i \underline{u}, \underline{u}) \leq c_{12} |u|^2_{H^1(\Omega_i)} \leq c_{12} c_2^2 |g_i|^2_{H^{1/2}(\partial\Omega_i)}.$$

This proves (3.20).

To prove (3.21), we have

(3.23)
$$c_4\,|g_i|^2_{H^{1/2}(\partial\Omega_i\cap\Gamma)} \le (\mathcal{S}_{B,i}\underline{g}_i,\underline{g}_i) \le c_5\,|g_i|^2_{H^{1/2}(\partial\Omega_i\cap\Gamma)} \quad \forall g_i=\Phi_{B,i}\underline{g}_i\in\mathbb{B}_i, \int_{\partial\Omega_i} g_i = 0$$

by the same arguments.

Let $g_i\in\mathbb{B}_i$ with $\overline{g}=\int_{\partial\Omega} g_i$. Thus, one obtains $g_i=\overline{g}\cdot\mathbf{1}+g_{i,0}$ with $\int_{\partial\Omega} g_{i,0}=0$. Since $|g_i|_{H^{1/2}(\partial\Omega_i)}=|g_{i,0}|_{H^{1/2}(\partial\Omega_i)}$ and $(\mathcal{S}_{B,i}\underline{g}_i,\underline{g}_i)=(\mathcal{S}_{B,i}\underline{g}_{0,i},\underline{g}_{0,i})$ for $g_{i,0}=\Phi_{B,i}\underline{g}_i$, relation (3.21) is a consequence of (3.23).

To prove (3.22), we have

$$\| g \|^2_{H^{1/2}(\Gamma)} = \sum_{i=1}^{s} \| g \|^2_{H^{1/2}(\partial\Omega_i\cap\Gamma)}\,. \tag{3.24}$$

Since we have homogeneous Dirichlet boundary conditions on $\partial\Omega$, i.e., $\partial\Omega=\Gamma_D$, the $H^{1/2}(\Gamma)$ seminorm is equivalent to the $H^{1/2}(\Gamma)$ norm. This is implied by the Poincaré–Friedrichs inequality. Using (3.24), (3.10), (3.21), and (3.20), the assertion follows. ∎

3.4. Final condition number estimates. Now, we are in a position to summarize the results of subsections 3.2 and 3.3. In a first step, we propose the primal DD preconditioner for $\mathcal{K}_i$ (3.8). Let

$$\mathcal{C}_i = \begin{bmatrix} I & -\mathcal{E}^\top_{DB,i} \\ \mathbf{0} & I \end{bmatrix} \begin{bmatrix} \mathcal{C}_{S,i} & \mathbf{0} \\ \mathbf{0} & \mathcal{C}_{D,i} \end{bmatrix} \begin{bmatrix} I & \mathbf{0} \\ -\mathcal{E}_{DB,i} & I \end{bmatrix}, \tag{3.25}$$

where

- $\mathcal{C}_{D,i}$ is the preconditioner (3.14) for $\mathcal{K}_{D,i}$,
- $\mathcal{C}_{S,i}$ is a preconditioner for $\mathcal{S}_{B,i}$ with $\mathcal{C}_{S,i}\sim\mathcal{S}_{B,i}$, and
- $\mathcal{E}_{DB,i}$ is the matrix representation of the extension operator $\mathrm{E}_i:\partial\Omega_i\mapsto\Omega_i$ in Algorithm 3.8.

Note that the matrix $\mathcal{C}_i$ is positive definite if $\partial\Omega\cap\partial\Omega_i\neq\emptyset$. If $\partial\Omega\cap\partial\Omega_i=\emptyset$, this matrix is positive semidefinite with $\ker\mathcal{C}_i=\ker\mathcal{K}_i=[\underline{u}_B,\underline{u}_L,\underline{u}_H]^\top=\left[\underline{1}^\top_{i,B},\underline{1}^\top_{i,L},\underline{0}^\top_{i,H}\right]^\top$. The choice of the preconditioners $\mathcal{C}_{S,i}$ is specified in the next subsection.

THEOREM 3.14. *Let $\mathcal{C}_i$ be defined via* (3.25). *Then, the spectral inequalities*

$$c_8\,(\mathcal{C}_i\underline{v},\underline{v}) \le (\mathcal{K}_i\underline{v},\underline{v}) \le c_9\,(\mathcal{C}_i\underline{v},\underline{v}) \quad \forall\underline{v}\in\mathbb{R}^{N_i} \tag{3.26}$$

are valid, where the spectral constants c_8 and c_9 do not depend on N_i.

Proof. We show that the assumptions of Lemma 3.1 are satisfied. Due to Theorem 3.7, we have $\mathcal{C}_{D,i}\sim\mathcal{K}_{D,i}$. Our assumptions imply the estimate $\mathcal{C}_{S,i}\sim\mathcal{S}_{B,i}$. Using Theorem 3.13, Theorem 3.11, and $|u|^2_{H^1(\Omega_i)}\sim(\mathcal{K}_i\underline{u},\underline{u})$ for all $u=\Phi_i\underline{u}\in\mathbb{S}^{\mathbf{1}}_{L,i}$, the relation

$$\left(\mathcal{K}_i\begin{bmatrix} I \\ \mathcal{E}_{BD,i}\end{bmatrix}\underline{g},\begin{bmatrix} I \\ \mathcal{E}_{BD,i}\end{bmatrix}\underline{g}\right) \le c_E^2\,(\mathcal{C}_{S,i}\underline{g},\underline{g}) \quad \forall\underline{g}\in\mathbb{R}^{M_i}$$

holds with a constant which is independent of N_i for the matrix representation of the extension operator in Algorithm 3.8. ∎

Now, we will present the preconditioner for the global stiffness matrix $\mathcal{K}$. Let

$$\mathcal{C} = \begin{bmatrix} I & -\mathcal{E}^\top_{DB} \\ \mathbf{0} & I \end{bmatrix} \begin{bmatrix} \mathcal{C}_S & \mathbf{0} \\ \mathbf{0} & \mathcal{C}_D \end{bmatrix} \begin{bmatrix} I & \mathbf{0} \\ -\mathcal{E}_{DB} & I \end{bmatrix}, \tag{3.27}$$

where $\mathcal{C}_D = \mathrm{diag}(\mathcal{C}_{D,i})_{i=1,\dots,s}$, $\mathcal{C}_S$ is a preconditioner for $\mathcal{S}_B$ which satisfies $\mathcal{C}_S \sim \mathcal{S}_B$, and $\mathcal{E}_{DB}$ is the matrix representation of the extension operators $\mathrm{E}_{L,i} : \partial\Omega_i \mapsto \Omega_i$ in Algorithm 3.8.

THEOREM 3.15. *Let $\mathcal{C}$ be defined via* (3.27). *Then,*

$$c_{10}\,(\mathcal{C}\underline{v},\underline{v}) \le (\mathcal{K}\underline{v},\underline{v}) \le c_{11}\,(\mathcal{C}\underline{v},\underline{v}) \quad \forall \underline{v} \in \mathbb{R}^N.$$

The constants do not depend on N.

Proof. The proof is similar to the proof of Theorem 3.14. ∎

3.5. Computational aspects. In the previous subsection, we have proved the optimality of the condition number of $\mathcal{C}^{-1}\mathcal{K}$. The purpose of this subsection is to show that the operation $\mathcal{C}^{-1}\underline{r}$ can be performed in optimal arithmetical complexity.

The preconditioner $\mathcal{C}$ (3.27) consists of three ingredients, namely, the preconditioner $\mathcal{C}_D$ for $\mathcal{K}_D$, the Schur complement preconditioner $\mathcal{C}_S$, and the extension operator $\mathcal{E}_{DB}$.

For the preconditioning operation $\mathcal{C}_D^{-1}\underline{r}_D$ (3.14), we have to solve subproblems in the high order subspaces $\mathbb{V}_j$, $j = 1,\dots,n$. Those problems can be treated by direct solvers since $\dim(\mathbb{V}_j) \preceq (1 + \log \frac{h_K}{h})^2$; see (2.5). On the subspace $\mathbb{V}_0$, we have to multiply with a diagonal matrix. Hence, the cost for $\mathcal{C}_D^{-1}\underline{r}_D$ is $\mathcal{O}(N)$.

Furthermore, we need a good preconditioner C_S for $\mathcal{S}_B$. Due to Theorem 3.13, the norm induced by the Schur complement is equivalent to the $H^{1/2}$ norm on the interface. Hence, any preconditioner which is known from the h-version of the FEM using uniformly refined meshes can be used as a preconditioner for the Schur complements $\mathcal{S}_{B,i}$ or the assembled Schur complement $\mathcal{S}_B$ (see, e.g., [29] or [47]). Indeed, today many preconditioners $\mathcal{C}_{S,i}$ for $\mathcal{S}_{B,i}$ are available such that $\kappa({\mathcal{C}_{S,i}}^{-\frac{1}{2}}\mathcal{S}_{B,i}{\mathcal{C}_{S,i}}^{-\frac{1}{2}}) = \mathcal{O}(1)$ and the solution of a system $\mathcal{C}_{S,i}\underline{w} = \underline{r}$ requires $\mathcal{O}(N_i)$ floating point operations. Similar results are valid for the assembled Schur complement $\mathcal{S}_B$. In our numerical experiments presented in section 5, we use the so-called Schur complement BPX preconditioner proposed in [44].

The last ingredient is the matrix representation of the extension operator $\mathcal{E}_{DB}$. Due to Algorithm 3.8, the operations $\mathcal{E}_{DB}\underline{w}_B$ and $\mathcal{E}_{DB}^T\underline{w}_D$ involve only BPX-like basis transformations on the boundary or the interface and in the interior, which are proportional to the numbers of unknowns in the interface or the interior. Therefore, both operations require $\mathcal{O}(N)$ floating point operations.

Summarizing, the total cost for the evaluation of $\mathcal{C}^{-1}\underline{r}$ is $\mathcal{O}(N)$ flops.

4. All-floating interface concentrated finite element tearing and interconnecting methods.

4.1. AF-FETI formulations. In order to avoid assembled matrices and vectors, we tear the global potential vector $\underline{u}$ on the subdomain boundaries $\Gamma_i = \partial\Omega_i$ by introducing the individual local unknowns $\underline{u}_i = R_i\underline{u}$; cf. also (2.7).

In contrast to the preceding sections, the vector $\underline{u}$ of nodal parameters now also contains those nodal parameters belonging to the Dirichlet boundary Γ_D that coincides with $\partial\Omega$ for our model problem (2.1). The global continuity of the potentials and the Dirichlet boundary conditions are now enforced by the constraints

$$\sum_{i=1}^{s} \mathcal{B}_i\underline{u}_i = \underline{0} \tag{4.1}$$

interconnecting the local potential vectors across the subdomain boundaries and the nodal parameters on the boundary Γ_D with the corresponding Dirichlet data,

which are zero for our model problem. Each row of the matrix $\mathcal{B} = (\mathcal{B}_1, \ldots, \mathcal{B}_s) = ((\mathcal{B}_{1,B}, \mathbf{0}), \ldots, (\mathcal{B}_{s,B}, \mathbf{0}))$ is connected with a pair of matching nodes across the subdomain boundaries or with a Dirichlet node. The entries of the former rows are 1 and -1 for the indices corresponding to the matching nodes on the interface (coupling boundaries) $\Gamma = \Gamma_S \setminus \Gamma_D$ and 0 otherwise, whereas an entry corresponding to a Dirichlet node on Γ_D is 1 and again 0 otherwise, where $\Gamma_S = \cup_{i=1}^{s} \Gamma_i$ denotes the skeleton. This method is called all floating (AF)-FETI. Therefore, (4.1) implies that the corresponding finite element functions $u_{i,h}$ are continuous across the interface Γ, i.e., $u_{i,h} = u_{j,h}$ on $\Gamma_i \cap \Gamma_j \neq \emptyset$, and are vanishing on $\Gamma_i \cap \Gamma_D$. We assume here that the number of constraints at some matching node is equal to the number of matching subdomains minus one. This method resulting in a minimal number of multipliers is called nonredundant; see, e.g., [28] for the use of redundant constraints.

By introducing Lagrange multipliers $\underline{\lambda} \in \mathbb{R}^L$, the linear system (2.8) is equivalent to the extended system

$$\begin{pmatrix} \mathcal{K}_1 & & & \mathcal{B}_1^\top \\ & \ddots & & \vdots \\ & & \mathcal{K}_s & \mathcal{B}_s^\top \\ \mathcal{B}_1 & \ldots & \mathcal{B}_s & \mathbf{0} \end{pmatrix} \begin{pmatrix} \underline{u}_1 \\ \vdots \\ \underline{u}_s \\ \underline{\lambda} \end{pmatrix} = \begin{pmatrix} \underline{f}_1 \\ \vdots \\ \underline{f}_s \\ \underline{0} \end{pmatrix}, \tag{4.2}$$

where now all matrices $\mathcal{K}_i$ correspond to local Neumann problems given by the Neumann bilinear form $\int_{\Omega_i} \nabla u \cdot \mathcal{D} \nabla v \, \mathrm{d}x$ on $H^1(\Omega_i)$. Due to the choice of the Lagrange multipliers, all matrices $\mathcal{K}_i$ are singular with kernels spanned by the vector $\left[\underline{1}_{i,B}^\top, \underline{1}_{i,L}^\top, \underline{0}_{i,H}^\top\right]^\top$, as was already mentioned in subsection 3.4.

Eliminating all interior unknowns $\underline{u}_{i,D}$, $i = 1, \ldots, s$, from system (4.2), we arrive at the equivalent reduced system

$$\begin{pmatrix} \mathcal{S}_{B,1} & & & \mathcal{B}_{B,1}^\top \\ & \ddots & & \vdots \\ & & \mathcal{S}_{B,s} & \mathcal{B}_{B,s}^\top \\ \mathcal{B}_{B,1} & \ldots & \mathcal{B}_{B,s} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \underline{u}_{B,1} \\ \vdots \\ \underline{u}_{B,s} \\ \underline{\lambda} \end{pmatrix} = \begin{pmatrix} \underline{g}_{B,1} \\ \vdots \\ \underline{g}_{B,s} \\ \underline{0} \end{pmatrix}, \tag{4.3}$$

with the singular Schur complements

$$\mathcal{S}_{B,i} = \mathcal{K}_{B,i} - \mathcal{K}_{BD,i} \mathcal{K}_{D,i}^{-1} \mathcal{K}_{DB,i}, \quad i = 1, \ldots, s, \tag{4.4}$$

and the corresponding right-hand sides

$$\underline{g}_{B,i} = \underline{f}_{B,i} - \mathcal{K}_{BD,i} \mathcal{K}_{D,i}^{-1} \underline{f}_{D,i}, \quad i = 1, \ldots, s. \tag{4.5}$$

The kernels of the matrices $\mathcal{S}_{B,i}$ are spanned by the vector $\underline{1}_{B,i}$. We now replace the singular Neumann–Schur complements $\mathcal{S}_{B,i}$ by the regularized matrices

$$\widetilde{\mathcal{S}}_{B,i} = \mathcal{S}_{B,i} + \alpha_i \underline{1}_{B,i} \underline{1}_{B,i}^\top, \quad i = 1, \ldots, s. \tag{4.6}$$

This corresponds to the regularized Neumann bilinear form

$$\int_{\Omega_i} \nabla u(x) \cdot \mathcal{D}(x) \nabla v(x) \, \mathrm{d}x + \left(\int_{\Gamma_i} u(x) \, \mathrm{d}s \right) \left(\int_{\Gamma_i} v(x) \, \mathrm{d}s \right) \tag{4.7}$$

on $H^1(\Omega_i)\times H^1(\Omega_i)$ with a positive regularization parameter α_i, e.g., $\alpha_i = h_i^2$. Taking into account the solvability conditions for local Neumann problems

$$(\underline{g}_{B,i} - \mathcal{B}_{B,i}^\top \underline{\lambda}, \underline{1}_{B,i}) = 0 \tag{4.8}$$

and the representation of the vectors $\underline{u}_{B,i}$ as

$$\underline{u}_{B,i} = \underline{v}_{B,i} + \gamma_i \underline{1}_{B,i} \quad \text{with} \quad (\underline{v}_{B,i}, \underline{1}_{B,i}) = 0, \tag{4.9}$$

we get the two-fold saddle point problem

$$\begin{pmatrix} \widetilde{\mathcal{S}}_{B,1} & & & \mathcal{B}_{B,1}^\top & \mathbf{0} \\ & \ddots & & \vdots & \\ & & \widetilde{\mathcal{S}}_{B,s} & \mathcal{B}_{B,s}^\top & \mathbf{0} \\ \mathcal{B}_{B,1} & \dots & \mathcal{B}_{B,s} & \mathbf{0} & \mathcal{G} \\ \mathbf{0} & \dots & \mathbf{0} & \mathcal{G}^\top & \mathbf{0} \end{pmatrix} \begin{pmatrix} \underline{v}_{B,1} \\ \vdots \\ \underline{v}_{B,s} \\ \underline{\lambda} \\ \underline{\gamma} \end{pmatrix} = \begin{pmatrix} \underline{g}_{B,1} \\ \vdots \\ \underline{g}_{B,s} \\ \underline{0} \\ \underline{e} \end{pmatrix} \tag{4.10}$$

with the notation $\underline{\gamma} = (\gamma_i)_{i=1,\dots,s} \in \mathbb{R}^s$, $\mathcal{G} = (\mathcal{B}_{B,1}\underline{1}_{B,1}, \dots, \mathcal{B}_{B,s}\underline{1}_{B,s})$, and $\underline{e} = (e_i)_{i=1,\dots,s} = ((\underline{g}_{B,i}, \underline{1}_{B,i}))_{i=1,\dots,s} \in \mathbb{R}^s$. The solution $\underline{u}_{B,i}$ can be recovered via formula (4.9).

Applying the orthogonal projection

$$\mathcal{P} = \mathcal{I} - \mathcal{G}(\mathcal{G}^\top\mathcal{G})^{-1}\mathcal{G}^\top \tag{4.11}$$

from the space $\mathbf{\Lambda} := \mathbb{R}^L$ onto the subspace $\mathbf{\Lambda}_0 = \ker \mathcal{G}^\top = (\operatorname{range}\mathcal{G})^\perp$ with respect to the scalar product $(\cdot,\cdot) = (\cdot,\cdot)_{\mathbf{\Lambda}} = (\cdot,\cdot)_{\mathbb{R}^L}$ to the penultimate equation of system (4.10), we can exclude $\underline{\gamma}$ from the first $s+1$ block equations of (4.10). Moreover, let us represent $\underline{\lambda}$ in the form

$$\underline{\lambda} = \mathcal{L}_0\underline{\lambda}_0 + \underline{\lambda}_e \tag{4.12}$$

with known $\underline{\lambda}_e = \mathcal{G}(\mathcal{G}^\top\mathcal{G})^{-1}\underline{e} \in (\ker\mathcal{G}^\top)^\perp = \operatorname{range}\mathcal{G}$, fulfilling the constraints $\mathcal{G}^\top\underline{\lambda}_e = \underline{e}$, and unknown $\mathcal{L}_0\underline{\lambda}_0 \in \ker\mathcal{G}^\top$, i.e., $\mathcal{G}^\top\mathcal{L}_0\underline{\lambda}_0 = \underline{0}$, where $\underline{\lambda}_0 \in \mathbb{R}^{L_0}$ and $L_0 = \dim\Lambda_0$. It is clear that the columns of $\mathcal{L}_0$ span a basis of Λ_0. Let

$$\widetilde{\mathcal{B}}_i = \begin{pmatrix} \mathcal{L}_0^\top\mathcal{P}^\top\mathcal{B}_{B,i} & \mathbf{0} \end{pmatrix}. \tag{4.13}$$

Now we can define $\underline{v}_{B,1}, \dots, \underline{v}_{B,s}$ and $\underline{\lambda}_0$ from the one-fold saddle point problem

$$\begin{pmatrix} \widetilde{\mathcal{S}}_{B,1} & & & \widetilde{\mathcal{B}}_1^\top \\ & \ddots & & \vdots \\ & & \widetilde{\mathcal{S}}_{B,s} & \widetilde{\mathcal{B}}_s^\top \\ \widetilde{\mathcal{B}}_1 & \dots & \widetilde{\mathcal{B}}_s & \mathbf{0} \end{pmatrix} \begin{pmatrix} \underline{v}_{B,1} \\ \vdots \\ \underline{v}_{B,s} \\ \underline{\lambda}_0 \end{pmatrix} = \begin{pmatrix} \underline{d}_{B,1} \\ \vdots \\ \underline{d}_{B,s} \\ \underline{0} \end{pmatrix}, \tag{4.14}$$

where $\underline{d}_{B,i} = \underline{g}_{B,i} - \mathcal{B}_{1,B}^\top\underline{\lambda}_e$. Once the vectors $\underline{v}_{B,1}, \dots, \underline{v}_{B,s}$ and $\underline{\lambda}_0$ are defined from (4.14), we get $\underline{\lambda}$ from (4.12), $\underline{\gamma}$ from the penultimate equation of (4.10), i.e.,

$$\underline{\gamma} = -(\mathcal{G}^\top\mathcal{G})^{-1}\mathcal{G}^\top(\mathcal{B}_{B,1}\underline{v}_{B,1} + \dots + \mathcal{B}_{B,s}\underline{v}_{B,s}), \tag{4.15}$$

and, finally, $\underline{u}_{B,1}, \dots, \underline{u}_{B,s}$ from (4.9). We will call the one-fold saddle point problem (4.14) an AF-FETI2 system.

Remark 4.1. We mention that the iteration updates for $\underline{\lambda}$ completely live in the subspace $\mathbf{\Lambda}_0$ if the initial guess is chosen as $\underline{\lambda}_e$. Therefore, a basis of $\mathbf{\Lambda}_0$, forming the columns of $\mathcal{L}_0$, is not explicitly required in the computation. We further mention that in the case of large jumps in the coefficients of the PDE the scalar product in $\mathbf{\Lambda}$ has to be changed from $(\cdot,\cdot)_{\mathbb{R}^L}$ to $(\mathcal{Q}\cdot,\cdot)_{\mathbb{R}^L}$ with some SPD matrix $\mathcal{Q}$ according to the proposals made in [28]; see also [5] and [47]. Of course, the change of the scalar product changes the orthogonal projection $\mathcal{P}$ from (4.11) to $\mathcal{P} = \mathcal{I} - \mathcal{Q}\mathcal{G}(\mathcal{G}^\top \mathcal{Q}\mathcal{G})^{-1}\mathcal{G}^\top$.

Eliminating the unknowns $\underline{v}_{B,1}, \ldots, \underline{v}_{B,s}$ from the AF-FETI2 system (4.14), we get the FETI Schur complement problem

$$\hat{\mathcal{F}}\underline{\lambda}_0 = \hat{\underline{d}}_0 \tag{4.16}$$

with the FETI Schur complement

$$\hat{\mathcal{F}} = \mathcal{L}_0^\top \mathcal{P}^\top \mathcal{F}\mathcal{P}\mathcal{L}_0 \quad \text{with} \quad \mathcal{F} = \sum_{i=1}^{s} \mathcal{B}_{B,i}\widetilde{\mathcal{S}}_{B,i}^{-1}\mathcal{B}_{B,i}^\top \tag{4.17}$$

and the corresponding right-hand side

$$\hat{\underline{d}}_0 = \mathcal{L}_0^\top \mathcal{P}^\top \underline{d}_0 - \mathcal{L}_0^\top \mathcal{P}^\top \sum_{i=1}^{s} \mathcal{B}_{B,i}\widetilde{\mathcal{S}}_{B,i}^{-1}\underline{d}_{B,i}. \tag{4.18}$$

The SPD system (4.16) is called the AF-FETI1 system.

On the other hand, we can again unfold the AF-FETI2 system (4.14), arriving at a saddle point system that is similar to the saddle point system (4.2), but now with regular diagonal blocks

$$\widetilde{\mathcal{K}}_i = \begin{bmatrix} \widetilde{\mathcal{K}}_{B,i} & \mathcal{K}_{BD,i} \\ \mathcal{K}_{DB,i} & \mathcal{K}_{D,i} \end{bmatrix} \tag{4.19}$$

with the regularized matrices $\widetilde{\mathcal{K}}_{B,i} = \mathcal{K}_{B,i} + \alpha_i \underline{1}_{B,i}\underline{1}_{B,i}^\top$. Then, the resulting system, called the AF-FETI3 system, can be written in the form

$$\begin{pmatrix} \widetilde{\mathcal{K}}_1 & \mathbf{0} & \ldots & \mathbf{0} & \widetilde{\mathcal{B}}_1^\top \\ \mathbf{0} & \widetilde{\mathcal{K}}_2 & & \vdots & \widetilde{\mathcal{B}}_2^\top \\ \vdots & & \ddots & \mathbf{0} & \vdots \\ \mathbf{0} & \ldots & \mathbf{0} & \widetilde{\mathcal{K}}_s & \widetilde{\mathcal{B}}_s^\top \\ \widetilde{\mathcal{B}}_1 & \ldots & & \widetilde{\mathcal{B}}_s & \mathbf{0} \end{pmatrix} \begin{pmatrix} \underline{\widetilde{u}}_1 \\ \vdots \\ \vdots \\ \underline{\widetilde{u}}_s \\ \underline{\lambda}_0 \end{pmatrix} = \begin{pmatrix} \underline{\widetilde{f}}_1 \\ \vdots \\ \vdots \\ \underline{\widetilde{f}}_s \\ \underline{0} \end{pmatrix}, \tag{4.20}$$

where

$$\underline{\widetilde{u}}_i = \begin{pmatrix} \underline{v}_{B,i} \\ \underline{u}_{D,i} \end{pmatrix}, \quad \underline{\widetilde{f}}_i = \begin{pmatrix} \underline{f}_{B,i} - \mathcal{B}_{B,i}^\top \underline{\lambda}_e \\ \underline{f}_{D,i} \end{pmatrix}, \quad \text{and} \quad \tilde{B}_i$$

are defined via (4.13).

Let us mention that all AF-FETI systems derived in this subsection are equivalent. In the following subsection we propose and analyze fast and robust solvers for the AF-FETI1–AF-FETI3 systems (4.16), (4.14), and (4.20).

4.2. AF-FETI solvers and preconditioners. Throughout this subsection, we assume that $\mathcal{D}(x) = a_i\mathcal{I}$ on Ω_i. We are interested in the robustness of our FETI preconditioners with respect to large jumps of these coefficients across the interfaces. We note that all results remain valid if the spectral condition number $\kappa(\mathcal{D}) = \lambda_{max}(\mathcal{D})/\lambda_{min}(\mathcal{D})$ of the diffusion matrix is uniformly small on Ω_i.

4.2.1. AF-FETI1 solvers. Since the system matrix $\hat{\mathcal{F}}$ of the AF-FETI1 system (4.16) is SPD, it can be efficiently solved with the help of the PCG method with an appropriate FETI preconditioner $\hat{\mathcal{C}}_F$. It is clear that we have only to construct a preconditioner $\mathcal{C}_F$ for $\mathcal{F}$ on the subspace $\mathbf{\Lambda}_0 = \ker \mathcal{G}^\top = (\operatorname{range} \mathcal{G})^\perp$. Candidates for $\mathcal{C}_F$ are the following preconditioners:

- Scaled exact Dirichlet FETI preconditioner (see, e.g., [47]):

$$\mathcal{C}_F^{-1} = \mathcal{A}\mathcal{S}_B\mathcal{A}^\top, \quad \mathcal{S}_B = \operatorname{diag}(\widetilde{\mathcal{S}}_{B,i}). \tag{4.21}$$

- Scaled inexact Dirichlet FETI preconditioner:

$$\mathcal{C}_F^{-1} = \mathcal{A}(\mathcal{S}_B + \mathcal{T}_B)\mathcal{A}^\top, \quad \mathcal{S}_B + \mathcal{T}_B = \operatorname{diag}(\widetilde{\mathcal{S}}_{B,i} + \mathcal{T}_{B,i}), \tag{4.22}$$

  where the exact Schur complement $\mathcal{S}_B$ is replaced by the perturbed Schur complement $\mathcal{S}_B + \mathcal{T}_B$; see (3.4).

- Scaled data-sparse hypersingular BETI preconditioner:

$$\mathcal{C}_F^{-1} = \mathcal{A}\mathcal{W}\mathcal{A}^\top, \quad \mathcal{W} = \operatorname{diag}(\widetilde{\mathcal{W}}_{B,i}), \tag{4.23}$$

  with the matrices $\widetilde{\mathcal{W}}_{B,i}$ arising from an appropriate data-sparse approximation of the hypersingular operator on $\partial\Omega_i$ [31].

The matrix $\mathcal{A} = (\mathcal{B}\tilde{\mathcal{D}}^{-1}\mathcal{B}^\top)^{-1}\mathcal{B}\tilde{\mathcal{D}}^{-1}$ is defined by the interconnecting matrix $\mathcal{B} = \operatorname{diag}(\mathcal{B}_{B,i})$ and by the scaling matrix $\tilde{\mathcal{D}} = \operatorname{diag}(\tilde{\mathcal{D}}_{B,i})$ with appropriate diagonal matrices $\tilde{\mathcal{D}}_{B,i}$. The diagonal entries depend on the coefficients a_i; see [28, 5, 47] for further details.

THEOREM 4.2. *Let $\mathcal{C}_F$ be one of the FETI preconditioners defined by* (4.21), (4.22), *or* (4.23). *Then, there exist positive constants $\underline{c}_F$ and $\overline{c}_F$ such that the spectral inequalities*

$$\underline{c}_F\,(\mathcal{C}_F\underline{\lambda},\underline{\lambda}) \le (\mathcal{F}\underline{\lambda},\underline{\lambda}) \le \overline{c}_F\,(1+\log(H/h))^2\,(\mathcal{C}_F\underline{\lambda},\underline{\lambda}) \tag{4.24}$$

hold for all $\underline{\lambda} \in \mathbf{\Lambda}_0 = \ker \mathcal{G}^\top$, where the constants $\underline{c}_F$ *and* $\overline{c}_F$ are *independent of h_i, H_i, and the a_i's (coefficient jumps), $H/h = \max H_i/h_i$.*

Proof. By Theorem 3.13, the matrices $\widetilde{\mathcal{S}}_{B,i}$, $\widetilde{\mathcal{S}}_{B,i} + \mathcal{T}_{B,i}$, and $\widetilde{\mathcal{W}}_{B,i}$ are spectrally equivalent to an auxiliary Schur complement $\widetilde{\mathcal{S}}_{B,i}^{FEM}$ arising from a standard finite element discretization with linear triangular elements on an auxiliary quasi-uniform triangular mesh generated from the given quasi-uniform boundary mesh on $\partial\Omega_i$ with the average mesh size h [31]. Since the spectral inequalities (4.24) were proved for $\widetilde{\mathcal{S}}_{B,i}^{FEM}$ (see [28], [5], or [47]), they are also valid if we replace $\widetilde{\mathcal{S}}_{B,i}^{FEM}$ in $\mathcal{F}$ and $\mathcal{C}_F^{-1}$ by the matrices $\widetilde{\mathcal{S}}_{B,i}$, $\widetilde{\mathcal{S}}_{B,i}+\mathcal{T}_{B,i}$, or $\widetilde{\mathcal{W}}_{B,i}$. This completes the proof of the theorem. ∎

Now we can solve the SPD system (4.16) via the PCG method, which can be efficiently performed on the subspace $\mathbf{\Lambda}_0$ with one of the preconditioners $\mathcal{C}_F^{-1}$ proposed above. Theorem 4.2 immediately implies that we need at most $I(\varepsilon) = O((1+\log(H/h))\log\varepsilon^{-1})$ iterations in order to reduce the initial error by the factor $\varepsilon \in (0,1)$ in a parallel regime. The matrix-vector product $\mathcal{F}\underline{\lambda}$, or $\widehat{\mathcal{F}}\underline{\lambda}_0$, requires the (parallel) solution of the systems

$$\widetilde{\mathcal{S}}_{B,i}\underline{w}_{B,i} = \underline{r}_{B,i}, \tag{4.25}$$

which are equivalent to the solution of the regularized finite element Neumann problems

$$\begin{pmatrix} \widetilde{\mathcal{K}}_{B,i} & \mathcal{K}_{BD,i} \\ \mathcal{K}_{DB,i} & \mathcal{K}_{D,i} \end{pmatrix} \begin{pmatrix} \underline{w}_{B,i} \\ \underline{w}_{D,i} \end{pmatrix} = \begin{pmatrix} \underline{r}_{B,1} \\ \underline{0}_{D,1} \end{pmatrix}, \quad i = 1, \ldots, s. \tag{4.26}$$

The discrete Neumann problem (4.26) can now be solved either by means of a sparse direct method with the matrix factorization in a preprocessing step or with the help of the PCG method with the Neumann preconditioner (3.25) proposed in subsection 3.4. The preconditioning step $\mathcal{C}_F^{-1}\underline{\rho}$ requires

- the (parallel) solution of local discrete Dirichlet problems with the system matrices $\mathcal{K}_{D,i}$,
- the (parallel) matrix-vector multiplications $(\widetilde{\mathcal{S}}_{B,i} + \mathcal{T}_{B,i})\underline{w}_{B,i}$ with a cost of $O((H_i/h_i))$ arithmetical operations, or
- the (parallel) multiplications $\widetilde{\mathcal{W}}_{B,i}\underline{w}_{B,i}$ with a cost of $O((H_i/h_i)(\log(H_i/h_i))^2)$ arithmetical operations

for the preconditioners (4.21), (4.22), or (4.23), respectively. Thus, the scaled inexact Dirichlet FETI preconditioner (4.22) is the cheapest FETI preconditioner.

4.2.2. AF-FETI2 solvers. The saddle point problem (4.14) can be solved by the Bramble–Pasciak PCG [4] or by any other suitable Krylov subspace iterative solver [43], e.g., the preconditioned conjugate residual method [27]; see also [25] and [26] for inexact FETI-DP as possible alternative. In any case, we need

- efficient preconditioners $\mathcal{C}_{B,i}$ for the regularized local Schur complements $\widetilde{\mathcal{S}}_{B,i}$ and
- a FETI preconditioner $\mathcal{C}_F$,

which are available from section 3 (cf. (3.25)) and (4.21)–(4.23), respectively. It follows from Theorem 4.2, Theorem 3.7, and the results of [4] and [49] that the number $I(\varepsilon)$ of iterations of the Bramble–Pasciak PCG behaves like $O((1+\log(H/h))\log\varepsilon^{-1})$.

The matrix-vector multiplication $\widetilde{\mathcal{S}}_{B,i}\underline{v}_{B,i}$ requires the solution of the local Dirichlet systems

$$\mathcal{K}_{D,i}\underline{w}_{D,i} = \underline{r}_{D,i}, \tag{4.27}$$

which can be done in parallel and which can be performed by means of a sparse direct method with the matrix factorization in a preprocessing step or with the help of the PCG method with the Dirichlet preconditioner $\mathcal{C}_{D,i}$ proposed by relation (3.14) in subsection 3.2.

4.2.3. AF-FETI3 solvers. The AF-FETI3 system (4.20) is again a one-fold saddle point problem which can be solved by the solvers mentioned above. In any case, we need

- efficient preconditioners $\mathcal{C}_i$ for the regularized local Neumann matrices $\tilde{K}_i$ (4.19) and
- a FETI preconditioner $\mathcal{C}_F$,

which are available from section 3 and (4.21)–(4.23), respectively.

THEOREM 4.3. *Assume that the AF-FETI3 system* (4.20) *is solved by means of the Bramble–Pasciak PCG method with the scaled inexact Dirichlet FETI preconditioner* (4.22) *and an appropriately scaled local Neumann preconditioner* (3.25) *composed of the optimal components* $\mathcal{C}_{D,i}$, $\mathcal{C}_{S,i}$, *and* $\mathcal{E}_{DB,i}$ *given in section* 3. *Then not more than* $I(\varepsilon) = O((1+\log(H/h))\log\varepsilon^{-1})$ *iterations and ops*$(\varepsilon) = O((H/h)(1+$

$\log(H/h))\log\varepsilon^{-1})$ *arithmetical operations are required in order to reduce the initial error by the factor* $\varepsilon \in (0,1)$ *in a parallel regime, where* $H/h = \max H_i/h_i$. *The number of iterations* $I(\varepsilon)$ *is robust with respect to the jumps in the coefficients across the interfaces. Moreover, at most* $O(H/h)$ *storage units are needed per processor.*

Proof. The estimate for the number $I(\varepsilon)$ of iterations is a direct consequence of results given in [4] (see [49] for improved estimates) and of the spectral estimates (3.14) and (4.24). The complexity estimates for the arithmetical costs and the memory demand follow from the corresponding complexity estimates for the block matrices involved in the matrix-vector multiplications and in the preconditioning. ∎

We emphasize that we can solve the interface concentrated finite element equations with the same complexity (even with a slightly better complexity) as data-sparse boundary element DD equations (cf. [30]). In contrast to boundary element technology, we can easily include source terms $f(\cdot)$ and we can even admit nonconstant but smooth coefficients $a_i(\cdot)$ in the subdomains Ω_i.

5. Numerical experiments. In this section, we first present some numerical experiments for the primal DD solver proposed in section 3. Finally, we investigate the behavior of the FETI preconditioner studied in section 4.

We consider problem (2.2) in $\Omega = (-1,1)^2$ using a piecewise constant coefficient matrix $\mathcal{D}(x,y)$ with coefficients $a_i\mathcal{I}$, $i = 1,2,3,4$, and a right-hand side $f(x,y) = 1$; see Figure 5.1. The problem is discretized by an IC-FEM with $p = 1$ in all elements. The interfaces are at the lines $x = 0$ and $y = 0$, whereas the boundaries are situated at $x = \pm 1$ and $y = \pm 1$. Here, we choose homogeneous Dirichlet boundary conditions.

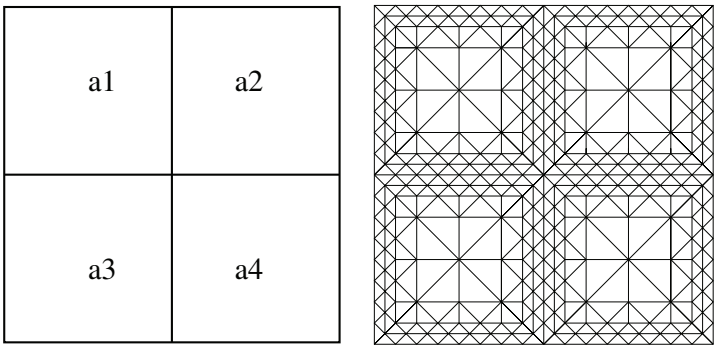


FIG. 5.1. *Coefficient matrix (left) and IC-concentrated mesh (right).*

In all experiments, we use the preconditioner $\mathcal{C}$ (3.27), where $\mathcal{E}_{DB}$ is the matrix representation of the extension operator of Algorithm 3.8 and $\mathcal{C}_S = \{\mathcal{S}_B, \mathcal{C}_{B,BPX}\}$ and $\mathcal{C}_D = \{\mathcal{K}_L, \mathcal{D}_L, \mathcal{C}_{D,BPX}, RR^T\}$ are the preconditioners for the Schur complement and the interior block, respectively, where

- $\mathcal{C}_{B,BPX}$ and $\mathcal{C}_{D,BPX}$ denote the BPX preconditioners for $\mathcal{S}_B$ and $\mathcal{K}_D$, respectively,
- R denotes the Cholesky factor of the incomplete Cholesky factorization for $\mathcal{K}_D$, i.e., $\mathcal{C}_{ichol} = RR^T$, and
- $\mathcal{D}_L$ denotes the diagonal part of $\mathcal{K}_L$.

Since $p = 1$ on all elements, the block $\mathcal{K}_H$ does not exist. The block $\mathcal{K}_H$ has been removed by the arguments of the overlapping preconditioner $\mathcal{C}_D$ which is defined via (3.14). This preconditioner is adapted from [12]; see also [36] for the three-dimensional

case. In both papers, the performance of the preconditioners is checked numerically for different applications. The experiments show the efficiency of this preconditioner. The aim of our experiments is to present efficiency of our preconditioner, which corresponds to a solver for the low order part in (3.14). Nevertheless, we present some numerical experiments for the AF-FETI1 preconditioner with $p > 1$.

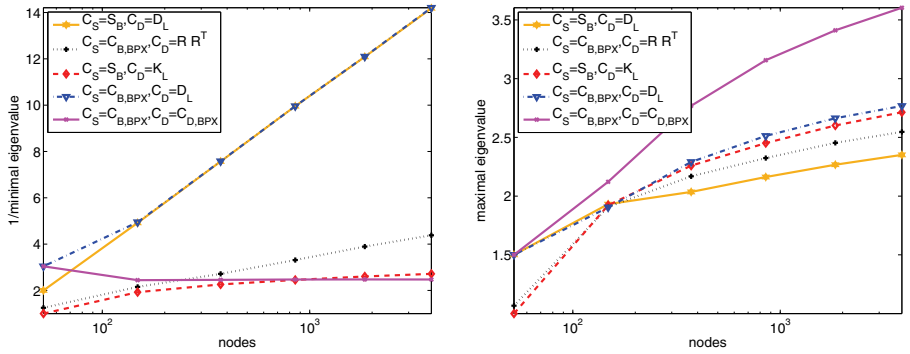


FIG. 5.2. *Eigenvalues of $\mathcal{C}^{-1}\mathcal{K}$ for moderate jumping coefficients: Minimal eigenvalue left; maximal eigenvalue right.*

Figure 5.2 displays the maximal and minimal eigenvalues of $\mathcal{C}^{-1}\mathcal{K}$ for several variants of $\mathcal{C}_S$ and $\mathcal{C}_D$ in the case of moderate jumping coefficients $a_i = i$, $i = 1, 2, 3, 4$. The optimality of the condition number can be seen in all cases. The usage of $\mathcal{C}_D = C_{D,BPX}$ or $\mathcal{C}_D = RR^T$ reduces the constants in comparison to $\mathcal{C}_D = \mathcal{D}_L$.

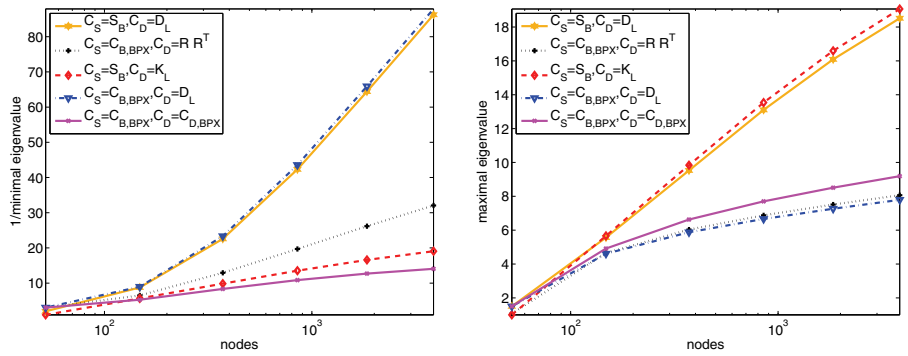


FIG. 5.3. *Eigenvalues of $\mathcal{C}^{-1}\mathcal{K}$ for strongly jumping coefficients: Minimal eigenvalue left; maximal eigenvalue right.*

In a second experiment, we have chosen $a_i = 10^i$, $i = 1, 2, 3, 4$. From the experiments, it can seen that the primal DD preconditioner is asymptotically optimal with respect to the number of unknowns. However, it is not robust with respect to the jumps of the diffusion coefficient; see Figure 5.3.

Moreover, we determine the numbers of iterations of the PCG method for the solution of $\mathcal{K}\underline{u} = \underline{f}$ using the preconditioner $\mathcal{C}$ (3.27) with $\mathcal{C}_S = \mathcal{C}_{B,BPX}$ and $\mathcal{C}_D = \mathcal{C}_{D,BPX}$ and a relative accuracy of 10^{-5}. Table 5.1 displays the numbers of iterations for the diffusion coefficients $a_i = i$ and $a_i = 10^i$, $i = 1, 2, 3, 4$, respectively.

In both cases, the numbers of PCG iterations are bounded in accordance with

TABLE 5.1
Numbers of PCG iterations for the solution of $\mathcal{K}\underline{u} = \underline{f}$ with the preconditioner $\mathcal{C}$.

Level	0	1	2	3	4	5	6	7	8	9
N	41	129	337	785	1713	3601	7409	15057	30385	61073
$a_i = i$	5	13	16	17	17	18	18	19	19	19
$a_i = 10^i$	5	23	35	41	53	57	57	55	56	56

our theory.

In the next experiment, we investigate the FETI preconditioner $\mathcal{C}_F$ (4.21) for $\mathcal{F}$ (4.16). Again, the finite element mesh of Figure 5.1 with $p = 1$ is used. We consider five different distributions for the coefficients $\underline{mat} = [a_i]_{i=1}^4$ in the subdomains Ω_i, $i = 1, 2, 3, 4$. The maximal eigenvalue of $\mathcal{C}_F^{-1}\mathcal{F}$ is displayed in Figure 5.4, whereas the minimal eigenvalue is 1 in all experiments. Both eigenvalues are robust with respect to material jumps. Moreover, a polynomial logarithmic growth of the condition number can be seen.

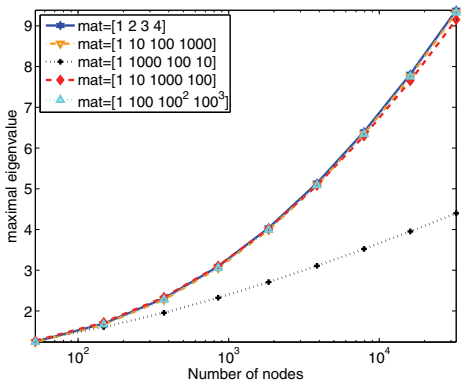


FIG. 5.4. *Maximal eigenvalue of the FETI preconditioner.*

In a last experiment, the solution of (4.16) with the FETI preconditioner $\mathcal{C}_F$ (4.21) is studied for more complex DDs. We consider problem (2.2) in $\Omega = (-1, 1)^2$ using a piecewise constant coefficient matrix $\mathcal{D}(x, y)$ with coefficients a_i, $i = 1, 2, 3, 4$, and a right-hand side $f(x, y) = 1$; see Figure 5.1. This problem is discretized by the IC-*hp*-FEM with $p > 1$, where Ω is decomposed into $s = 4^k$ congruent squares, $k = 1, 2, 3, 4$; see Figure 5.5 for $k = 2$. As in the previous examples, we compare moderate and highly jumping coefficients, i.e., $a_i = i$ and $a_i = 10^i$, $i = 1, 2, 3, 4$, respectively. Note that each coefficient a_i is chosen in 4^{k-1} subdomains. Figure 5.6 displays the condition number of $\mathcal{C}_F^{-1}\hat{\mathcal{F}}$ which is estimated by the Lanzcos method for the maximal eigenvalue and by one for the minimal eigenvalue.

In the case of highly jumping coefficients $a_i = 10^i$, the AF-FETI1 system (4.16) has been solved by the PCG method using the FETI preconditioner $\mathcal{C}_F^{-1}$ (4.21). Table 5.2 displays the PCG iteration number using a relative accuracy of 10^{-8}. The largest problem corresponding to level 12 and $s = 256$ subdomains has 36330497 unknowns. We mention that a uniformly refined grid would yield about 4.3 billions unknowns. The solution of the problem with 36330497 unknowns requires less than 40 minutes using the 8 processors of our 4 × Dualcore Intel Xeon computer (CPU 3.40GHz, 16MB processor cache, 64GB RAM in total). We also determined the number of

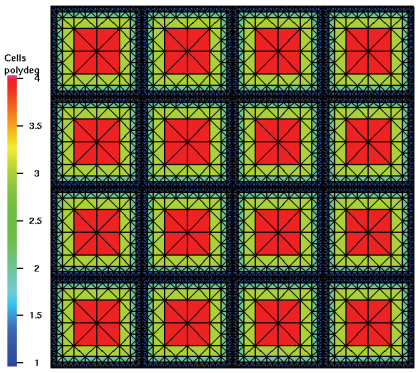


FIG. 5.5. *IC finite element mesh with* 16 *subdomains.*

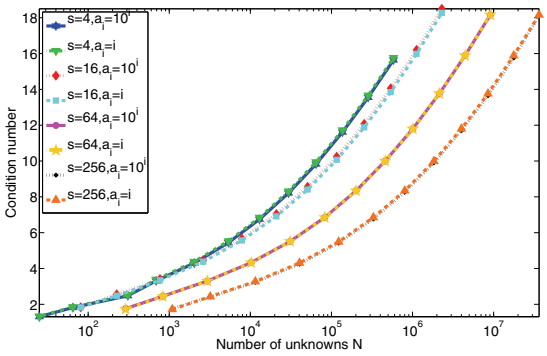


FIG. 5.6. *Maximal eigenvalue (condition number) of* $C_F^{-1}\hat{\mathcal{F}}$ *for several test problems.*

TABLE 5.2
Number of PCG-iterations for the AF-FETI1 system (4.16) *with the FETI-preconditioner* (4.21).

Level	p_{max}	$s = 4$	$s = 16$	$s = 64$	$s = 256$
1	1	2	9	10	10
2	1	4	12	12	12
3	2	4	13	15	14
4	3	5	15	17	17
5	4	5	17	19	19
6	5	5	18	21	21
7	6	6	20	23	23
8	7	6	21	24	25
9	8	6	22	25	26
10	9	6	22	27	27
11	10	8	24	29	29
12	11	8	25	30	31

PCG iterations for the case of an interface concentrated finite element discretization with polynomial degree $p = 1$ on all elements. The number of iterations for $p > 1$ and $p = 1$ differs only for the cases $l = 5, s = 16$ (17 instead of 16) and $l = 12, s = 64$ (30 instead of 31) by one iteration!

Remark 5.1. In real applications, the numbers of unknowns on the subdomains Ω_i can be quite different. Thus, a mesh partitioning where each processor obtains one

domain Ω_i is not suitable in the sense of a good load balancing. In order to obtain a good load balancing, a mesh partitioner is used in order to decompose the mesh into subdomains $\tilde{\Omega}_j$ such that $\cup_j \overline{\tilde{\Omega}}_j = \overline{\Omega}_i$ or $\cup_i \overline{\Omega}_i = \overline{\tilde{\Omega}}_j$ if the number of unknowns on Ω_i is too large or too small, respectively. Material or source interfaces are not cut with this strategy. The preconditioners are not changed, but the data structures have to be adapted. We mention that the component $\mathcal{D}_L$ of the overlapping preconditioner (3.14) is a diagonal matrix. However, the component $\mathcal{K}_{P,j}$ of the preconditioner (3.14) requires data exchanges of the unknowns on elements between different processors.

6. Conclusions and possible generalizations. The solution of potential problems with piecewise smooth coefficients can be approximated by the IC-FEM with the same asymptotical accuracy as the standard FEM but with a considerably smaller number of unknowns. We have proposed primal and dual DD preconditioners resulting in asymptotically optimal or almost optimal solvers for the interface concentrated finite element equations. The first class of preconditioners is based on inexact Dirichlet–Dirichlet preconditioning techniques. The second class of preconditioners is constructed on the basis of the so-called all-floating finite element tearing and interconnecting (AF-FETI) technology. The latter class of preconditioners is very robust with respect to jumps in the coefficients of the PDE and highly efficient in parallel computing. It is clear that instead of AF-FETI technology, the FETI-DP technique can also be used. The number of unknowns involved in the IC-FEM is comparable to the number of unknowns living on the skeleton of the DD. Thus, the coupling of IC-FEM with data-sparse boundary element methods (BEM) seems to be very attractive for many applications. This coupling has also been suggested for exterior boundary value problems in [22] based on the results in [24]. In contrast to the standard FEM, the IC-FEM will not perturb the complexity of data-sparse BEM. We refer the reader to the survey paper [32] for the standard symmetric BEM-FEM coupling.

Our results have been presented for the two-dimensional case. The results for the development of the preconditioners can also be applied to the three-dimensional case. The key ingredients, the overlapping preconditioner of subsection 3.2 (see [36]) and the Haase–Nepomnyaschikh extension are available. However, an approximation theorem such as Theorem 2.6 is still missing in the literature. It might be possible to prove such a result as in [22], but the proof seems to be very technical.

Acknowledgments. The authors thank Clemens Pechstein (Linz) for the numerical experiments with the FETI preconditioner and the anonymous referees for improving the paper.

REFERENCES

[1] S. Beuchler, *Extension operators on tensor product structures in two and three dimensions*, SIAM J. Sci. Comput., 26 (2005), pp. 1776–1795.

[2] S. Beuchler, T. Eibner, and U. Langer, *Primal and dual boundary concentrated iterative substructuring methods*, Technical report 2007-07, RICAM, Linz, Austria, 2007.

[3] J. Bramble, J. Pasciak, and A. Schatz, *The construction of preconditioners for elliptic problems by substructuring* I, Math. Comp., 47 (1986), pp. 103–134.

[4] J. H. Bramble and J. E. Pasciak, *A preconditioning technique for indefinite systems resulting from mixed approximations of elliptic problems*, Math. Comp., 50 (1988), pp. 1–17.

[5] S. C. Brenner, *An additive Schwarz preconditioner for the FETI method*, Numer. Math., 94 (2003), pp. 1–31.

[6] C. R. Dohrmann, *A preconditioner for substructuring based on constrained energy minimization*, SIAM J. Sci. Comput., 25 (2003), pp. 246–258.

[7] Z. Dostál, D. Horák, and R. Kučera, *Total FETI—an easier implementable variant of the FETI method for numerical solution of elliptic PDE*, Comm. Numer. Methods Engrg., 22 (2006), pp. 1155–1162.

[8] M. Dryja, *A capacitance matrix method for Dirichlet problem on polygon region*, Numer. Math., 39 (1982), pp. 51–64.

[9] M. Dryja, *A finite element-capacitance matrix method for elliptic problems on regions partitioned into substructures*, Numer. Math., 44 (1984), pp. 153–168.

[10] T. Eibner, *Randkonzentrierte und adaptive hp-FEM*, Ph.D. thesis, TU Chemnitz, Chemnitz, Germany, 2006.

[11] T. Eibner and J. M. Melenk, *Multilevel preconditioning for the boundary concentrated hp-fem*, Comp. Methods Appl. Mech. Engrg., 196 (2007), pp. 3713–3725.

[12] T. Eibner and J. M. Melenk, *A local error analysis of the boundary-concentrated hp-FEM*, IMA J. Numer. Anal., 27 (2007), pp. 752–778.

[13] C. Farhat, M. Lesoinne, and K. Pierson, *A scalable dual-primal domain decomposition method*, Numer. Linear Algebra Appl., 7 (2000), pp. 687–714.

[14] C. Farhat and F.-X. Roux, *A method of finite element tearing and interconnecting and its parallel solution algorithm*. Internat. J. Numer. Meth. Engrg., 32 (1991), pp. 1205–1227.

[15] G. Haase, *Hierarchical extension operators plus smoothing in domain decomposition preconditioners*, Appl. Numer. Math., 20 (1996), pp. 1–20.

[16] G. Haase, U. Langer, and A. Meyer, *The approximate Dirichlet decomposition method* I. *An algebraic approach*, Computing, 47 (1991), pp. 137–151.

[17] G. Haase, U. Langer, and A. Meyer, *The approximate Dirichlet domain decomposition method* II. *Applications to* 2*nd-order elliptic BVPs*, Computing, 47 (1991), pp. 153–167.

[18] G. Haase, U. Langer, A. Meyer, and S. V. Nepomnyaschikh, *Hierarchical extension operators and local multigrid methods in domain decomposition preconditioners*, East-West J. Numer. Math., 2 (1994), pp. 173–193.

[19] G. Haase and S. V. Nepomnyaschikh, *Extension explicit operators on hierarchical grids*, East-West J. Numer. Math., 5 (1997), pp. 231–248.

[20] W. Hackbusch, B. N. Khoromskij, and S. Sauter, *Adaptive Galerkin boundary element methods with panel clustering*, Numer. Math., 105 (2007), pp. 603–631.

[21] W. Hackbusch, B. N. Khoromskij, and R. Kriemann, *Direct Schur complement method by domain decomposition based on $\mathcal{H}$-matrix approximation*, Comput. Vis. Sci., 8 (2005), pp. 179–188.

[22] B. N. Khoromskij and J. M. Melenk, *Boundary concentrated finite element methods*, SIAM J. Numer. Anal., 41 (2003), pp. 1–36.

[23] B. N. Khoromskij and J. M. Melenk, *An efficient direct solver for the boundary concentrated FEM in* 2*D*, Computing, 69 (2002), pp. 91–117.

[24] B. N. Khoromskij and S. Prössdorf, *Fast computations with the harmonic Poincaré-Steklov operators on nested refined meshes*, Adv. Comput. Math., 8 (1998), pp. 111–135.

[25] A. Klawonn, *An iterative substructuring method with Lagrange multipliers for elasticity problems using approximate Neumann subdomain solvers*, in Multifield Problems, Springer-Verlag, Berlin, 2000, pp. 193–200.

[26] A. Klawonn and O. Rheinbach, *Inexact FETI-DP methods*, Internat. J. Numer. Meth. Engrg., 69 (2007), pp. 284–307.

[27] A. Klawonn and O. B. Widlund, *A domain decomposition method with Lagrange multipliers and inexact solvers for linear elasticity*, SIAM J. Sci. Comput., 22 (2000), pp. 1199–1219.

[28] A. Klawonn and O. B. Widlund, *FETI and Neumann–Neumann iterative substructuring methods: Connections and new results*, Comm. Pure Appl. Math., 54 (2001), pp. 57–90.

[29] V. G. Korneev and U. Langer, *Domain decomposition and preconditioning*, in Encyclopedia of Computational Mechanics, Vol. 1, E. Stein, R. de Borst, and Th.J.R. Hughes, eds., John Wiley & Sons, New York, 2004, pp. 617–647.

[30] U. Langer, G. Of, O. Steinbach, and W. Zulehner, *Inexact data-sparse boundary element tearing and interconnecting methods*, SIAM J. Sci. Comput., 29 (2007), pp. 290–314.

[31] U. Langer and O. Steinbach, *Boundary element tearing and interconnecting method*, Computing., 71 (2003), pp. 205–228.

[32] U. Langer and O. Steinbach, *Coupled finite and boundary element domain decomposition methods*, in Boundary Element Analysis: Mathematical Aspects and Application, M. Schanz and O. Steinbach, eds., Lect. Notes Appl. Comput. Mech. 29, Springer-Verlag, Berlin, 2007, pp. 61–95.

[33] J. Mandel and C. R. Dohrmann, *Convergence of a balancing domain decomposition by constraints and energy minimization*, Numer. Linear Algebra Appl., 10 (2003), pp. 639–659.

[34] A. M. Matsokin and S. V. Nepomnyaschikh, *The Schwarz alternation method in a subspace*, Izv. Vyssh. Uchebn. Zaved. Mat., 29 (1985), pp. 61–66, 85 (in Russian).
[35] A. M. Matsokin and S. V. Nepomnyaschikh, *The fictitious space method and explicit extension operators*, Comput. Maths Math. Phys., 33 (1993), pp. 45–59.
[36] J. M. Melenk, C. Pechstein, J. Schöberl, and S. Zaglmayr, *Additive Schwarz preconditioning for p-version triangular and tetrahedral finite elements*, IMA J. Numer. Anal., 28 (2008), pp. 1–24.
[37] J. Nečas, *Les méthodes directes en théorie des équations elliptiques*, Masson et Cie, Éditeurs, Paris, 1967.
[38] S. V. Nepomnyaschikh, *Domain Decomposition and Schwarz Methods in a Subspace for the Approximate Solution of Elliptic Boundary Value Problems*, Ph.D. thesis, Computing Center of Siberian Division of Russian Academy of Sciences, Moscow, 1986.
[39] S. V. Nepomnyaschikh, *Method of splitting into subspaces for solving elliptic boundary value problems in compllex-form domains*, Soviet. J. Numer. Anal. Math. Modelling, 6 (1991), pp. 151–168.
[40] G. Of, *BETI-Gebietszerlegungsmethoden mit schnellen Randelementverfahren und Anwendungen*, Ph.D. thesis, Universität Stuttgart, Stuttgart, Germany, 2006.
[41] P. Oswald, *Multilevel Finite Element Approximation*, Teubner Skripten zur Numerik, Teubner, Stuttgart, Germany, 1994.
[42] L. F. Pavarino, *Additive Schwarz methods for the p-version finite element method*, Numer. Math., 66 (1994), pp. 493–515.
[43] Y. Saad, *Iterative Methods for Sparse Linear Systems*, 2nd ed., SIAM, Philadelphia, 2003.
[44] C. H. Tong, T. F. Chan, and C. C. J. Kuo, *A domain decomposition preconditioner based on a change to a multilevel nodal basis*, SIAM J. Sci. Statist. Comput., 12 (1991), pp. 1486–1495.
[45] A. Toselli and X. Vasseur, *A numerical study on Neumann-Neumann and FETI methods for hp approximations on geometrically refined boundary layer meshes in two dimensions*, Comput. Methods Appl. Mech. Engrg., 192 (2003), pp. 4551–4579.
[46] A. Toselli and X. Vasseur, *Domain decomposition preconditioners of Neumann-Neumann type for hp-approximations on boundary layer meshes in three dimensions*, IMA J. Numer. Anal., 24 (2004), pp. 123–156.
[47] A. Toselli and O. Widlund, *Domain Decoposition Methods: Algorithms and Theory*, Springer Ser. Comput. Math. 34, Springer-Verlag, Berlin, 2004.
[48] H. Yserentant, *Coarse grid spaces for domains with a complicated boundary*, Numer. Algorithms, 21 (1999), pp. 387–392.
[49] W. Zulehner, *Analysis of iterative methods for saddle point problems: A unified approach*, Math. Comp., 71 (2002), pp. 479–505.

SIAM J. NUMER. ANAL.
Vol. 46, No. 6, pp. 2843–2866
© 2008 Society for Industrial and Applied Mathematics

ANALYSIS OF FINITE ELEMENT DOMAIN EMBEDDING METHODS FOR CURVED DOMAINS USING UNIFORM GRIDS*

SHENG ZHANG†

Abstract. We analyze the error of a finite element domain embedding method for elliptic equations on a domain ω with curved boundary. The domain is embedded in a rectangle R on which uniform mesh and linear continuous elements are employed. The numerical scheme is based on an extension of the differential equation from ω to R by regularization with a small parameter ϵ (for Neumann and Robin problems), or penalty with a large parameter ϵ^{-1} (for the Dirichlet problem), or a mixture of these (for a mixed boundary value problem). For Neumann and Robin problems, we prove that when $\epsilon \leq h$ (the mesh size), the error in the $H^1(\omega)$ norm is of the optimal order $\mathcal{O}(h)$. For the Dirichlet problem, when $\epsilon \leq h^{1/2}$, the error is $\mathcal{O}(h^{1/2})$ that is not optimal. If the mesh is adjusted around $\partial\omega$ to fit it, then the optimal convergence rate $\mathcal{O}(h)$ holds for the Dirichlet problem if ω is convex and $\epsilon \leq h$. If ω is not convex, then the convergence rate can only be improved to $\mathcal{O}(h^{2/3})$ by such mesh adjustment, with the parameter being $\epsilon = h^{2/3}$. In this latter case, a parameter smaller than $h^{2/3}$ thwarts the convergence rate, which is verified by a numerical result.

Key words. domain embedding, fictitious domain, curved boundary, uniform mesh, fast Poisson solver

AMS subject classifications. 65N30, 65N45, 65T50

DOI. 10.1137/060671681

1. Introduction.

In recent years, domain embedding, or fictitious domain, methods have developed into a general methodology of the numerical computations of partial differential equations that has the advantage of efficiently dealing with complex domain boundaries by using uniform, or slightly adjusted, grids on a larger rectangular domain [11]. For a second order elliptic boundary value problem on a complicated domain, the ultimate goal of domain embedding methods is to quickly obtain an accurate numerical solution by mainly solving the Poisson equation on uniform meshes on a rectangle by fast solvers, through, say, preconditioned iterations. A first step in these methods is to extend the boundary value problem to a larger rectangle. There are many ways to do this, one of which is using regularization for Neumann or Robin problem, or penalty for Dirichlet problem, or a combination of these for a mixed boundary value problem [14, 15, 17, 23]. A more general approach, including some additional treatments on the original domain boundary, can be found in [2]. This step results in an extended boundary value problem on the rectangle with large jumps in the coefficients of the differential equation. A finite element domain embedding method then is obtained as a straightforward discretization of the extended equation on a uniform mesh on the rectangle. The error of the finite domain embedding method has two sources. One is due to the replacement of the original problem by the one on the larger rectangular domain. This has been analyzed by many authors, and in [23, 24] the sharp estimates on the regularization/penalty error can be found. The other is due to the discretization of the extended equation, which seems not to have been completely understood. In the literature, there are numerous

*Received by the editors October 9, 2006; accepted for publication (in revised form) April 7, 2008; published electronically August 1, 2008. This work was partially supported by NSF grant DMS-0513559.

http://www.siam.org/journals/sinum/46-6/67168.html

†Department of Mathematics, Wayne State University, Detroit, MI 48202 (sheng@math.wayne.edu).

works devoted to this subject. In [17], a thorough analysis is presented under the assumption that the original domain boundary aligns with the uniform mesh. This assumption seems to be a stringent restriction on the arbitrariness of the domain. For example, the domain boundary cannot be smoothly curved. The regularity of the solution and the alignment of the domain boundary with the uniform mesh required in this analysis are often not compatible. Furthermore, the results of [17] are actually not valid for general curved domains for Dirichlet problems. In [14, 15], an analysis for the Neumann problem can be found, which did not impose such restriction on the domain. But the error estimates established there are not sharp enough. A difficulty in this kind of error analysis for curved and complex domains is that the solution of the extended equation is not globally smooth. It has jumps in its normal derivative across the original domain boundary that inevitably cuts through some element if the mesh is uniform. In this paper, we present an analysis on such discretization errors for various boundary value problems. We first establish the estimates under the mere assumption that the original domain is Lipschitz. We then assume the domain boundary is smoothly curved, and we establish sharp estimates and use the aforementioned regularization/penalty error estimate to determine the balance between the mesh size and the regularization/penalty parameter such that the overall accuracy of the finite element domain embedding method is optimized [20, 19]. We shall see that the balance is rather delicate in some cases.

While the primary objective of this paper is analyzing the accuracy of the finite element domain embedding methods, there is, however, an important question that must be addressed before the methods can be claimed competitive. A domain embedding finite element method results in a discrete system that involves large or small parameters. The question is whether the fast Poisson solver can effectively precondition such a system so that a preconditioned iterative method could converge uniformly with respect to both the finite element mesh sizes and the large/small parameters. While the uniformity of such convergence with respect to the mesh sizes is obvious, the uniformity with respect to the parameters is by no means trivial. We shall not go into details with this issue, but refer to [5] for a strategy to resolve this. This strategy suggests to start the preconditioned iteration with some special initial vectors such that the whole process of the iteration is confined in a subspace in which the parameters have no effect. In our case, we need to choose initial vectors that are zero on the finite element nodes outside the original domain.

In view of the preconditioning, a slight distortion of the uniform mesh is acceptable, as long as the discrete Laplacian on the distorted mesh is spectrally equivalent to that on the uniform mesh. It turns out that no such adjustment is needed for Neumann or Robin problems, since the full accuracy is already achieved by uniform meshes. But for the Dirichlet problem, due to a locking phenomena, the accuracy offered by uniform meshes is rather poor. A slight adjustment of the uniform mesh around the original domain boundary by moving the nearby nodes onto the boundary efficiently reduces the locking effect and renders the full accuracy to the domain embedding finite element method when the domain is convex, or significantly improves the accuracy when the domain is not convex around the concave portion of accuracy.

For simplicity, we only present the results for the Poisson equation defined on a two-dimensional domain and for linear continuous finite elements. The results of this paper are, however, valid for general self-adjoint second order elliptic equations. Similar results can be derived for finite elements of higher orders. The theory is also possible to extend to some higher order equations like those of clamped and free (but not simply supported) plates. In the remainder of this introduction, we briefly

summarize our results and compare them with existing results in the literature.

Let $\omega \subset \mathbb{R}^2$ be a bounded connected domain such that its boundary Γ is of C^2 class [13]. For a given function $f \in L^2(\omega)$, we consider the boundary value problem

$$-\Delta u = f \quad \text{in} \quad \omega$$

subject to the homogeneous Dirichlet boundary condition $u = 0$ on Γ; or the homogeneous Neumann condition $\partial u/\partial n = 0$, with n being the unit outward normal of Γ; or a Robin condition $ku + \partial u/\partial n = 0$, with k being a smooth, bounded, and a strictly positive function on Γ; or a mixed boundary condition. The homogeneity is not a real restriction for Neumann and Robin conditions. A nonhomogeneous Dirichlet condition could be reduced to a homogeneous one by defining a function satisfying the boundary condition and subtracting it from u.

The finite element domain embedding method is based on an approximation to u obtained by solving an ϵ-dependent boundary value problem on a larger rectangular domain $R \subset \mathbb{R}^2$ such that $\omega \subset R$. The formulation of the problem on R depends on the boundary condition in the original problem. The finite element domain embedding method is then a straightforward discretization of the extended one on R with mesh size h. We shall assume that $\omega \subset\subset R$. Then the boundary condition on ∂R is at one's disposal. We impose the homogeneous Dirichlet condition. However, Neumann, Robin, mixed, or a periodical boundary condition on ∂R works equally well. (If $\partial R \cap \Gamma \neq \emptyset$, the condition on ∂R must be subject to some restrictions [24].) We let $\Omega = R \setminus \overline{\omega}$ be the fictitious domain, and we extend the function f to a function $\bar{f}$ on R by defining $\bar{f} = 0$ on Ω. Henceforth, we shall use the notation $\mathcal{P} \lesssim \mathcal{Q}$, which means that there exists a constant C independent of $\mathcal{P}$, $\mathcal{Q}$, ϵ, and h such that $\mathcal{P} \leq C\mathcal{Q}$. The notation $\mathcal{P} \simeq \mathcal{Q}$ means $\mathcal{P} \lesssim \mathcal{Q}$ and $\mathcal{Q} \lesssim \mathcal{P}$. We use $H^s(D)$ to denote the L^2-based Sobolev space of order s on a domain D, in which the norm and semi-norm are denoted by $\|\cdot\|_{s,D}$ and $|\cdot|_{s,D}$, respectively. In particular, the L^2 norm is denoted by $\|\cdot\|_{0,D}$ or $|\cdot|_{0,D}$.

We start with the Neumann problem of which the weak formulation seeks $u \in H^1(\omega)/\mathbb{R}$ such that

$$(\nabla u, \nabla v)_{[L^2(\omega)]^2} = \int_\omega f v dx \quad \forall\ v \in H^1(\omega). \tag{1.1}$$

Here ∇ is the gradient operator, and the parentheses stand for the inner products in Hilbert spaces indicated by the subscripts. This problem is well posed under the assumption that $\int_\omega f dx = 0$. For $\epsilon > 0$, we determine that $u^\epsilon \in H_0^1(R)$ by a regularization such that

$$(\nabla u^\epsilon, \nabla v)_{[L^2(\omega)]^2} + \epsilon(\nabla u^\epsilon, \nabla v)_{[L^2(\Omega)]^2} = \int_R \bar{f} v dx \quad \forall\ v \in H_0^1(R). \tag{1.2}$$

This is an elliptic problem that yields a unique $u^\epsilon \in H_0^1(R)$. As $\epsilon \to 0$, u^ϵ converges to a limit $u^0 \in H_0^1(R)$ at the sharp rate of ϵ as long as ω is a Lipschitz domain (which is assured by our assumption on Γ). The limit u^0 is harmonic on Ω, and $u^0|_\omega$ solves the Neumann problem (1.1). The equivalent estimate [24]

$$\|u^\epsilon - u^0\|_{1,R} \simeq |u^\epsilon - u^0|_{1,\omega} \simeq \epsilon$$

holds with the rare exception that $u^\epsilon \equiv u^0$, which occurs if and only if $u^0 = 0$ on Γ, i.e., a solution of the homogeneous Neumann problem also satisfies the homogeneous

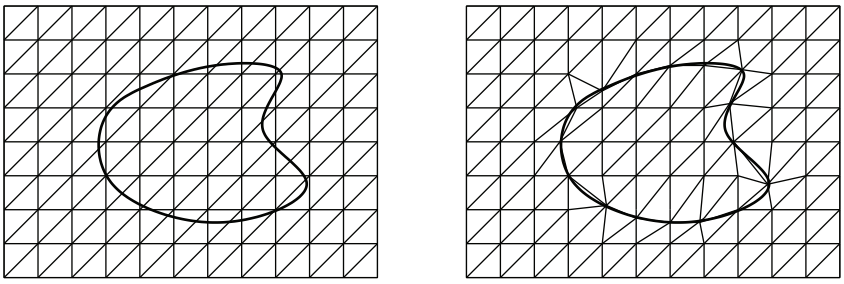


FIG. 1. *A domain ω embedded in a rectangle R with a uniform triangulation (left), and a triangulation adjusted around Γ (right).*

Dirichlet condition on Γ. Incidentally, this is also equivalent to $u^0 \in H^2(R)$. In this case, we can take $\epsilon = 1$ in (1.2), and the Poisson equation on R produces the exact solution of the original equation on ω. Except for this special case, we do not have $u^0 \in H^2(R)$, since there is a jump in its normal derivative across Γ.

We introduce a uniform triangulation of mesh size h on R; see Figure 1 (left), and let $H_h \subset H_0^1(R)$ be the subspace of piecewise linear continuous functions subordinate to this triangulation. The finite element domain embedding method is a straightforward discretization of (1.2). It determines $u_h^\epsilon \in H_h$ such that

$$(\nabla u_h^\epsilon, \nabla v_h)_{[L^2(\omega)]^2} + \epsilon(\nabla u_h^\epsilon, \nabla v_h)_{[L^2(\Omega)]^2} = \int_R \bar{f} v_h dx \quad \forall\ v \in H_h. \tag{1.3}$$

Then the restriction of u_h^ϵ on ω is the numerical solution offered by the finite domain embedding method. We prove that

$$|u_h^\epsilon - u^0|_{1,\omega} \lesssim \epsilon + \inf_{v_h \in H_h} \left(|u^0 - v_h|_{1,\omega} + \sqrt{\epsilon}|u^0 - v_h|_{1,\Omega} \right).$$

This estimate is valid as long as ω is a Lipschitz domain. When $\Gamma \in C^2$, we have the piecewise smoothness property of u^0 such that $u^0|_\omega \in H^2(\omega)$ and $u^0|_\Omega \in H^2(\Omega)$. Using this regularity, we construct an interpolation v_h of u^0 such that $|u^0 - v_h|_{1,\omega} \lesssim h$ and $|u^0 - v_h|_{1,\Omega} \lesssim \sqrt{h}$; thus to prove that

$$|u_h^\epsilon - u^0|_{1,\omega} \lesssim \epsilon + h + \sqrt{\epsilon\, h}. \tag{1.4}$$

From this we see that when $\epsilon \lesssim h$, we have a domain embedding finite element method that achieves the full accuracy of linear continuous finite elements with an error of $\mathcal{O}(h)$. This result is better than that of [14, 15], in which an estimate of the form $\sqrt{\epsilon} + h + \sqrt{\epsilon}$, instead of that in (1.4), was established. They, therefore, require $\epsilon = Ch^2$ to achieve the full accuracy. Taking smaller ϵ does not hurt the accuracy of the finite element domain embedding methods. From a computational point of view, one would like to avoid very small ϵ. In this sense, $\epsilon = Ch$ seems the best choice.

The piecewise smoothness of u^0 is a consequence of the assumption that $\Gamma \in C^2$ and $f \in L^2(\omega)$ [13]. It is not valid when ω is a polygon, since either $u^0|_\omega \in H^2(\omega)$ or $u^0|_\Omega \in H^2(\Omega)$ will be broken, depending on the convexity of ω. However, when ω is convex and Γ aligns with the mesh line, as assumed in [17], the result (1.4) remains valid, which needs a different argument.

If we fix h and let $\epsilon \to 0$, we have $u_h^\epsilon \to u_h^0 \in H_h$. The restriction of the limit u_h^0 on ω is identical to the solution given by a method proposed by Friedrichs and Keller

in 1966 [12], of which an analogue for the Robin problem (see below) was used in the 1980s by the Boeing program `tranair` to circumvent the mesh generation problem around a complete aircraft and was also included in the software `freefem3d` [11]. The positive value of ϵ renders to the equation (1.3) the advantages of better data structure and faster resolution by using fast Poisson solvers on the rectangle R as preconditioners. This method also offers an alternative of dealing with the singularity of a discretized Neumann problem. See [6], where a survey of such various techniques can be found.

The weak formulation of the Robin problem seeks $u \in H^1(\omega)$ such that

$$(\nabla u, \nabla v)_{[L^2(\omega)]^2} + (ku, v)_{L^2(\Gamma)} = \int_\omega f v dx \quad \forall\ v \in H^1(\omega). \tag{1.5}$$

This equation can be extended to R in the following way. We determine $u^\epsilon \in H_0^1(R)$ such that

$$(\nabla u^\epsilon, \nabla v)_{[L^2(\omega)]^2} + (ku^\epsilon, v)_{L^2(\Gamma)} + \epsilon(\nabla u^\epsilon, \nabla v)_{[L^2(\Omega)]^2} = \int_R \bar{f} v dx \quad \forall\ v \in H_0^1(R). \tag{1.6}$$

This, once again, is a well-defined problem on R, and $\lim_{\epsilon \to 0} u^\epsilon = u^0 \in H_0^1(R)$. The limit u^0 is harmonic on Ω, and $u^0|_\omega$ solves the Robin problem (1.5). The equivalent estimate

$$\|u^\epsilon - u^0\|_{1,R} \simeq \|u^\epsilon - u^0\|_{1,\omega} + \|u^\epsilon - u^0\|_{0,\Gamma} \simeq \epsilon \tag{1.7}$$

holds as long as ω is Lipschitz [24]. There is an exception to the equivalent estimate (1.7) in which $u^\epsilon \equiv u^0$. This happens if and only if $u^0 = 0$ on Γ. This is the only case in which $u^0 \in H^2(R)$. Otherwise, u^0 has only the piecewise smoothness as for the Neumann problem. The finite element domain embedding method that determines $u_h^\epsilon \in H_h$ is a straightforward discretization of (1.6) on a uniform mesh. Using the very same argument as for Neumann problem, we obtain that

$$\|u_h^\epsilon - u^0\|_{1,\omega} \lesssim \epsilon + h + \sqrt{\epsilon\, h}.$$

The weak formulation of the Dirichlet problem seeks $u \in H_0^1(\omega)$ such that

$$(\nabla u, \nabla v)_{[L^2(\omega)]^2} = \int_\omega f v dx \quad \forall\ v \in H_0^1(\omega).$$

If ω is simply connected, then, on the continuous level, the domain embedding method determines $u^\epsilon \in H_0^1(R)$ such that

$$(\nabla u^\epsilon, \nabla v)_{[L^2(\omega)]^2} + \epsilon^{-1}(\nabla u^\epsilon, \nabla v)_{[L^2(\Omega)]^2} = \int_R \bar{f} v dx \quad \forall\ v \in H_0^1(R). \tag{1.8}$$

We have that $\lim_{\epsilon \to 0} u^\epsilon = u^0 \in H_0^1(R)$, that the limit is characterized by $u^0|_\omega = u$ and $u^0|_\Omega = 0$, and that the equivalent estimate

$$\|u^\epsilon - u^0\|_{1,R} \simeq \|u^\epsilon\|_{1,\Omega} \simeq \epsilon \tag{1.9}$$

holds, under the assumption that Ω is Lipschitz [24]. (This is true since we have assumed $\Gamma \in C^2$ and $\omega \subset\subset R$; it is also true when, for example, a L-shaped polygon is tightly embedded in a rectangle. It, however, is not true when a circular domain

is tightly embedded in a square.) Furthermore, if we let $u^1 \in H_0^1(R)$ be the unique function determined by $\Delta u^1 = 0$ in ω, and let

$$(\nabla u^1, \nabla v)_{[L^2(\Omega)]^2} = \int_\omega f v dx - (\nabla u^0, \nabla v)_{[L^2(\omega)]^2} \quad \forall\ v \in H_0^1(R). \tag{1.10}$$

Then we have

$$\|u^\epsilon - u^0 - \epsilon\, u^1\|_{1,R} \simeq \epsilon^2\,. \tag{1.11}$$

An exception to (1.9) and (1.11) is that $u^\epsilon \equiv u^0$ and $u^1 = 0$. This occurs if and only if u also satisfies the homogeneous Neumann condition on Γ. This is equivalent to, say, $u^0 \in H^2(R)$. Except for this special case, we do not have $u^0 \in H^2(R)$, since the normal derivative of u^0 is not continuous across Γ.

The finite element domain embedding method determines $u_h^\epsilon \in H_h$ such that

$$(\nabla u_h^\epsilon, \nabla v_h)_{[L^2(\omega)]^2} + \epsilon^{-1}(\nabla u_h^\epsilon, \nabla v_h)_{[L^2(\Omega)]^2} = \int_R \bar{f} v_h dx \quad \forall\ v \in H_h. \tag{1.12}$$

We prove that

$$\begin{aligned}\|u_h^\epsilon - u\|_{1,\omega} \lesssim \epsilon + \inf_{v_h^1 \in H_h} \left(\epsilon\, \|u^1 - v_h^1\|_{1,\omega} + \sqrt{\epsilon}\|u^1 - v_h^1\|_{1,\Omega}\right) \\ + \inf_{v_h^0 \in H_h} \left(\|u^0 - v_h^0\|_{1,\omega} + \frac{1}{\sqrt{\epsilon}}\|v_h^0\|_{1,\Omega}\right).\end{aligned} \tag{1.13}$$

We construct a v_h^1 to interpolate u^1 in the same way as interpolating u^0 in a Neumann problem and bound the above second term by $\epsilon\sqrt{h} + h\sqrt{\epsilon}$. Our assumption of $\Gamma \in C^2$ and $f \in L^2(\omega)$ assures us that $u = u^0|_\omega \in H^2(\omega) \cap H_0^1(\omega)$. We construct an interpolation v_h^0 in such a way that it is zero on any open triangular element that is not entirely contained in ω. This makes $\|v_h^0\|_{1,\Omega} = 0$ and $\|u^0 - v_h^0\|_{1,\omega} \lesssim \sqrt{h}$. The above third term is then bounded by $\sqrt{h}$, which is a sharp estimate. We thus have

$$\|u_h^\epsilon - u\|_{1,\omega} \lesssim \epsilon + \epsilon\sqrt{h} + h\sqrt{\epsilon} + \sqrt{h}. \tag{1.14}$$

So the accuracy of the finite element domain embedding method is limited by $\mathcal{O}(\sqrt{h})$, which is assured when $\epsilon \lesssim \sqrt{h}$. The rather low order of accuracy is a kind of locking phenomena that is not unusual in the numerical computation of equations involving large parameters, like the Reissner–Mindlin plate and nearly incompressible elasticity [3].

The accuracy of the finite element domain embedding method for Dirichlet problem can be significantly improved by a slight adjustment of the nodes around Γ. One can move the nearby nodes onto Γ and reconnect some of the affected nodes such that no mesh-line segment has one end in ω and the other in Ω. This way, a polygonal interpolation of Γ will be formed in the mesh; see Figure 1 (right). This can be done by using Börger's algorithm [7] that results in a triangulation that is quasi uniform and shape regular, on which the discrete Laplacian is spectrally equivalent to that on the uniform mesh. On the adjusted triangulation, we then construct v_h^0 by simply interpolating u^0 at all of the nodes. When ω is convex, we have $\|u^0 - v_h^0\|_{1,\omega} \lesssim h$; see [10], and $\|v_h^0\|_{1,\Omega} = 0$ in (1.13). Thus, to obtain the estimate

$$\|u_h^\epsilon - u\|_{1,\omega} \lesssim \epsilon + \epsilon\sqrt{h} + h\sqrt{\epsilon} + h. \tag{1.15}$$

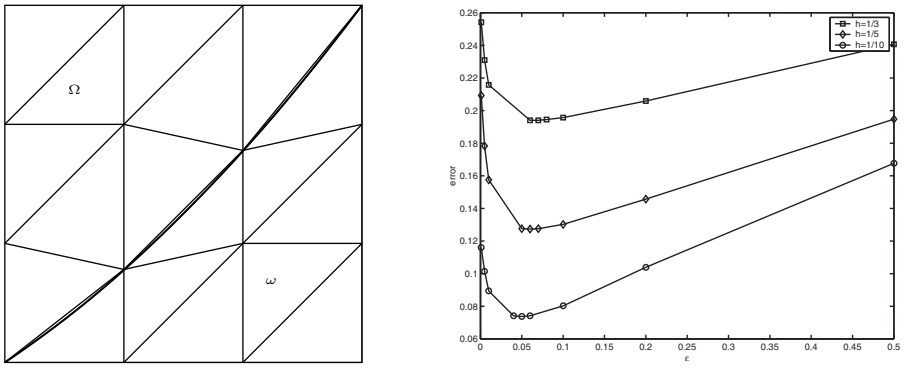


FIG. 2. *A nonconvex ω embedded in the unit square with adjusted mesh around its boundary (left), and the dependence of the error on ϵ and h (right).*

So when $\epsilon \lesssim h$ the error of the finite element domain embedding method in the $H^1(\omega)$ norm is of the optimal order $\mathcal{O}(h)$. If ω is not convex, we still have $\|u^0 - v_h^0\|_{1,\omega} \lesssim h$. However, $\|v_h^0\|_{1,\Omega}$ is no longer zero, but it is bounded by $\mathcal{O}(h)$. The error estimate becomes

$$\|u_h^\epsilon - u\|_{1,\omega} \lesssim \epsilon + \epsilon\sqrt{h} + h\sqrt{\epsilon} + h + \frac{h}{\sqrt{\epsilon}}. \tag{1.16}$$

This estimate is also sharp. The dependence of the error of a finite element domain embedding method on h and ϵ is rather complicated. From the above estimate we see that if we take $\epsilon = Ch^{2/3}$, the domain embedding finite element method has the error estimate $\mathcal{O}(h^{2/3})$ in $H^1(\omega)$ norm. It is interesting to note that, in this case, a smaller ϵ thwarts the convergence rate. This should be compared with the Neumann problem and the Dirichlet problem on convex domains, in which smaller ϵ does not improve but does not hurt the convergence rate of the finite element domain embedding method either. To confirm this observation, we include a numerical example here. We choose R to be the unit square $(0,1) \times (0,1)$ on the xy-plane, in which ω is the curvilinear triangle under the curved diagonal $y = x(x+2)/3$, which is not convex. The exact solution of the Dirichlet problem with a homogeneous boundary condition is $u(x,y) = [3y - x(x+2)]y(x-1)$. We first divide R by a uniform mesh of mesh size h, and then we move some nodes near the curved boundary of ω onto it to form the adjusted mesh; see Figure 2 (left), in which $h = 1/3$. On this mesh, we perform the computation with the method (1.12). Figure 2 (right) shows the dependence of $|u - u_h^\epsilon|_{1,\omega}$ on ϵ for $h = 1/3$, $1/5$, and $1/10$, respectively. It clearly shows that there is an optimal value of ϵ for each h, which decreases with h, and smaller ϵ thwarts the accuracy.

In conclusion, a slight adjustment of the finite element mesh to accommodate Γ appears to be worthwhile for the Dirichlet problem, which enhances the convergence rate from $\sqrt{h}$ to h if ω is convex, and from $\sqrt{h}$ to $h^{2/3}$ if ω is not convex.

It is important to note that the discrete system (1.12) based on such adjusted mesh can be efficiently preconditioned by using the discrete Laplacian on the uniform mesh with the preconditioned iteration initiated with a vector of zero entries corresponding to nodes in $R \setminus \omega$ [7, 5]. If one wishes to achieve full accuracy for nonconvex ω, then the mesh could be refined around the concave portion of Γ. A local mesh size $\mathcal{O}(h^{1.5})$ would lead to the full accuracy of the order $\mathcal{O}(h)$ of the domain embedding finite

element method in which $\epsilon = Ch$. This refinement is needed only around the concave portion of the boundary of ω. See [18] for such an example.

If ω is polygonal and Γ aligns with the finite element mesh, then if $u^0|_\omega \in H^2(\omega)$ (which is true when ω is a convex polygon, but is only an assumption otherwise), we construct v_h^0 and v_h^1 by nodal interpolation to u^0 and u^1 in (1.13) to obtain the estimate

$$\|u_h^\epsilon - u\|_{1,\omega} \lesssim \epsilon + \epsilon\sqrt{h} + \sqrt{\epsilon\, h} + h. \tag{1.17}$$

Here we have used the fact that $u^1|_\omega \in H^{1.5}(\omega)$ and $u^1|_\Omega \in H^{1.5}(\Omega)$ [16]. Thus when $\epsilon \lesssim h$, one has a domain embedding finite element method that has the optimal accuracy of $\mathcal{O}(h)$. This was mentioned in [17], where a rigorous analysis was presented for the Neumann problem under this alignment assumption.

If ω is not simply connected, a term like $\epsilon^{-1}(u^\epsilon, v)_{L^2(\Omega)}$ needs to be added to the left-hand side of (1.8). Then the estimate (1.9) holds, and $u^0|_\omega = u$ regardless of the connectivity of ω. Actually, in case that ω is multiply connected, adding a term of the form $\epsilon^{-1}(u^\epsilon, v)_{L^2(\Omega^0)}$ in the left-hand side of (1.8) is sufficient to guarantee the validity of the domain embedding method on the continuous level. Here Ω^0 is the union of the isolated components of Ω. The finite element domain embedding methods should then be based on such a modification of (1.8).

Finally, we make some remarks on the mixed boundary value problem. To make the presentation sufficiently general, we assume that Γ is split into three parts such that $\Gamma = \overline{\Gamma}_D \cup \overline{\Gamma}_N \cup \overline{\Gamma}_R$, and on Γ_D, Γ_N, and Γ_R, homogeneous Dirichlet, Neumann, and Robin conditions are imposed, respectively. Corresponding to this splitting, we divide Ω into three parts such that $\overline{\Omega} = \overline{\Omega}_D \cup \overline{\Omega}_N \cup \overline{\Omega}_R$, and $\partial\Omega_D \cap \Gamma = \overline{\Gamma}_D$, $\partial\Omega_N \cap \Gamma = \overline{\Gamma}_N$, and $\partial\Omega_R \cap \Gamma = \overline{\Gamma}_R$. The weak formulation of the mixed problem determines $u \in H_D^1(\omega)$ (that is, the space of $H^1(\omega)$ functions that vanish on Γ_D) such that

$$(\nabla u, \nabla v)_{[L^2(\omega)]^2} + (ku, v)_{L^2(\Gamma_R)} = \int_\omega f v dx \quad \forall\ v \in H_D^1(\omega). \tag{1.18}$$

On the continuous level, the domain embedding method determines that $u^\epsilon \in H_0^1(R)$ such that

$$\begin{aligned}(\nabla u^\epsilon, \nabla v)_{[L^2(\omega)]^2} &+ (ku^\epsilon, v)_{L^2(\Gamma_R)} \\ &+ \epsilon^{-1}(\nabla u^\epsilon, \nabla v)_{[L^2(\Omega_D)]^2} + \epsilon(\nabla u^\epsilon, \nabla v)_{[L^2(\Omega\setminus\overline{\Omega}_D)]^2} = \int_R \bar{f} v dx \quad \forall\ v \in H_0^1(R).\end{aligned} \tag{1.19}$$

When $\epsilon \to 0$, u^ϵ converges to a limit $u^0 \in H_0^1(R)$. The limit satisfies (1) $u^0|_\omega$ solves (1.18) (assuming ω is simply connected), (2) $u^0 = 0$ on Ω_D, and (3) u^0 is harmonic on Ω_N and Ω_R. Under the assumption that *both* Ω_D *and* $\omega \cup \Gamma_D \cup \Omega_D$ *are Lipschitz domains*, the equivalent estimate

$$\|u^\epsilon - u^0\|_{1,R} \simeq \|u^\epsilon - u^0\|_{1,\omega} + \|u^\epsilon\|_{1,\Omega_D} \simeq \epsilon \tag{1.20}$$

holds, with the exception that $u^\epsilon \equiv u^0$, which occurs when and only when $u^0 \in H^2(R)$ [23]. The finite element domain embedding method then is again a straightforward discretization of (1.18) that determines $u_h^\epsilon \in H_h$ such that

$$\begin{aligned}(\nabla u_h^\epsilon, \nabla v_h)_{[L^2(\omega)]^2} &+ (ku_h^\epsilon, v_h)_{L^2(\Gamma_R)} \\ &+ \epsilon^{-1}(\nabla u_h^\epsilon, \nabla v_h)_{[L^2(\Omega_D)]^2} + \epsilon(\nabla u_h^\epsilon, \nabla v_h)_{[L^2(\Omega\setminus\overline{\Omega}_D)]^2} = \int_R \bar{f} v_h dx \quad \forall\ v_h \in H_h.\end{aligned} \tag{1.21}$$

The convergence rate estimates on $\|u^0 - u_h^\epsilon\|_{1,\omega}$ can be obtained under some regularity assumption on the solution of the original problem (1.18), which is not guaranteed by our assumptions on Γ and f. For uniform meshes, Γ_D generally cuts through some elements, and the ends of Γ_D lie in the interior of some elements; the rate is only bounded by $\sqrt{h}$. By moving nearby nodes onto Γ_D and its ends, and reconnecting some mesh lines, as for the Dirichlet problem, we can improve the convergence rate to $\mathcal{O}(h)$ or $\mathcal{O}(h^{2/3})$, depending on the convexity of Γ_D with respect to ω. As for Neumann and Robin problems, no treatment is needed to accommodate Γ_N or Γ_R.

The paper is organized as follows. In section 2, as a preparation, we derive some estimates on the finite element interpolation error on a strip around Γ that cuts through some elements. We also include a result about the ϵ-dependent variational problem, which is a cornerstone in estimating the regularization/penalty error. This result is also needed in the discretization error estimate for the Dirichlet problem. In section 3, we estimate the error of the finite element domain embedding methods for Neumann and Dirichlet problems, and we determine the balance between the value of the regularization/penalty parameter and the finite element mesh size. Since no new technique is needed for the Robin problem and the mixed problem, we shall not go into details to prove the results mentioned above.

2. Technical preliminaries. We need to estimate the error of piecewise linear interpolation to functions whose restriction on both ω and Ω are in H^2, but the functions themselves do not belong to $H^2(R)$. Their normal derivatives are not continuous across the curve Γ, and Γ cuts through some of the triangular elements. For this purpose we need to estimate the H^1 norms on thin strips of piecewise H^2 functions and their finite element interpolations. The following lemmas will be used in the next section.

Let $y = \gamma(x)$ be a C^2 function on $[0, L]$. We consider a strip domain γ_δ that is bounded by $x = 0$, $x = L$, $y = \gamma(x)$, and $y = \gamma(x) + \delta$; see Figure 3. We assume that D is a given domain such that $\gamma_\delta \subset\subset D$. We have the following estimates on functions restricted on γ_δ.

LEMMA 2.1. *Let $w \in H^1(D)$. Then we have that*

$$\|w\|^2_{L^2(\gamma_\delta)} \le C\delta \|w\|^2_{1,D}.$$

Here, C is a constant independent of δ and w.

Proof. In terms of the curvilinear coordinates $X = x, Y = y - \gamma(x)$, we have that $\|w\|^2_{0,\gamma_\delta} \simeq \int_0^\delta \int_0^L w^2 dX dY$. For each $Y \in (0, \delta)$, by a trace theorem [1], we have that $\int_0^L w^2 dX \le C\|w\|^2_{1,D}$. The constant C can be chosen to be independent of Y. Integrating this estimate, with respect to Y from 0 to δ, gives the desired estimate. ∎

From this result, the following lemma easily follows.

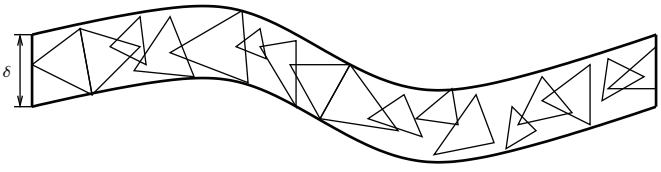


FIG. 3. *A strip γ_δ with a cluster of shape regular triangles.*

LEMMA 2.2. *Let $w \in H^2(D)$. Then we have that*

$$\|w\|^2_{1,\gamma_\delta} \le C\delta \|w\|^2_{2,D}.$$

Here, C is independent of δ and w.

LEMMA 2.3. *Let $\mathcal{T}_\delta$ be the set of a cluster of shape regular, but not necessarily quasi uniform, open triangles contained in γ_δ, and suppose that any point in γ_δ can only be covered by at most a certain number of triangles in $\mathcal{T}_\delta$. Let $w \in H^2(D)$. For each $\tau \in \mathcal{T}_\delta$, we let $I_\tau w$ be the linear interpolation of w on the vertices of τ. Then we have that*

$$\sum_{\tau \in \mathcal{T}_\delta} |I_\tau w|^2_{1,\tau} \le C\delta \|w\|^2_{2,D}.$$

Here, C is a constant independent of δ and w.

Proof. Let $\tau \in \mathcal{T}_\delta$ be one such triangle. We affine map it onto a standard reference triangle T of unit size through the mapping F_τ. The restriction on τ of a function $w \in H^2(D)$ is mapped to a function $\hat{w}$ on T such that $\hat{w} \circ F_\tau = w$. Let I be the linear interpolation operator on the vertices of T. For any constant p_0, we have that

$$|I_\tau w|^2_{1,\tau} = |I_\tau(w - p_0)|^2_{1,\tau} \le C|I(\hat{w} - p_0)|^2_{1,T} \le C\|\hat{w} - p_0\|^2_{L^\infty(T)} \le C\|\hat{w} - p_0\|^2_{2,T}.$$

Thus, by the Bramble–Hilbert Lemma

$$|I_\tau w|^2_{1,\tau} \le C(|\hat{w}|^2_{1,T} + |\hat{w}|^2_{2,T}).$$

Scaling the right-hand side back from T to τ, we get that

$$|I_\tau w|^2_{1,\tau} \le C(|w|^2_{1,\tau} + h^2_\tau |w|^2_{2,\tau}).$$

Here h_τ is the diameter of τ, which does not exceed δ. Summing up the above inequality for all triangles in the cluster $\mathcal{T}_\delta$, we get that

$$\sum_{\tau \in \mathcal{T}_\delta} |I_\tau w|^2_{1,\tau} \le C \sum_{\tau \in \mathcal{T}_\delta} |w|^2_{1,\tau} + C\delta^2 \sum_{\tau \in \mathcal{T}_\delta} |w|^2_{2,\tau} \le C|w|^2_{1,\gamma_\delta} + C\delta^2 |w|^2_{2,\gamma_\delta}.$$

In the last step, we used the assumption that any point in γ_δ can only be covered by at most a certain number of triangles in $\mathcal{T}_\delta$. The desired result follows from Lemma 2.2. ∎

We shall apply these estimates to strips that can be covered by a finite number of (translated and rotated) strips as defined here. The strips of width δ referred to in the later sections are understood in this sense.

As we mentioned in the introduction, there are two sources of errors in the finite element domain embedding methods. The first one is due to the extension of the original boundary value problem from ω to R. The sharp estimates on this kind of error can be obtained by using the following equivalent estimates on an abstract ϵ-dependent equation. This estimate will also be directly used in estimating the finite element interpolation error for the Dirichlet problem in the next section.

Let H, U, and V be Hilbert spaces, $A : H \to U$ be a bounded linear operator, and let $B : H \to V$ be a bounded linear operator *with closed range*. We assume that

$$\|A\boldsymbol{v}\|_U + \|B\boldsymbol{v}\|_V \simeq \|\boldsymbol{v}\|_H \quad \forall\ \boldsymbol{v} \in H. \tag{2.1}$$

Thus the bilinear form

$$(\boldsymbol{u},\boldsymbol{v})_{\mathcal{H}} := (A\boldsymbol{u},A\boldsymbol{v})_U + (B\boldsymbol{u},B\boldsymbol{v})_V$$

defines an equivalent inner product on H. Furnished with this new inner product, the space H will be denoted by $\mathcal{H}$. Let $\ker B \subset H$ be the kernel of the operator B. Let $\boldsymbol{f}$ be a linear continuous functional on H such that $\boldsymbol{f}|_{\ker B} = 0$. There is a unique $\boldsymbol{u}^\epsilon \in H$ such that

$$\epsilon(A\boldsymbol{u}^\epsilon, A\boldsymbol{v})_U + (B\boldsymbol{u}^\epsilon, B\boldsymbol{v})_V = \langle \boldsymbol{f},\boldsymbol{v}\rangle \ \forall \boldsymbol{v} \in H. \tag{2.2}$$

Without loss of generality, we assume that the operator B maps H onto V. (If necessary, we replace V by the range of B in it.) Then B is an isomorphism between $(\ker B)^\perp_{\mathcal{H}}$ (the orthogonal complement of $\ker B$ with respect to the $\mathcal{H}$-norm) and V. Since we have assumed that $\boldsymbol{f}|_{\ker B} = 0$, according the the closed range theorem and the Riesz representation theorem, there exists a unique $\boldsymbol{u}^0 \in (\ker B)^\perp_{\mathcal{H}}$ such that

$$(B\boldsymbol{u}^0, B\boldsymbol{v})_V = \langle \boldsymbol{f},\boldsymbol{v}\rangle \ \ \forall\ \boldsymbol{v} \in H. \tag{2.3}$$

LEMMA 2.4. *Under the assumptions that $\boldsymbol{f}|_{\ker B} = 0$ and B has a closed range in V as $\epsilon \to 0$, $\boldsymbol{u}^\epsilon$, the solution of (2.2) converges to the limit $\boldsymbol{u}^0 \in (\ker B)^\perp_{\mathcal{H}}$ defined by (2.3), and we have the equivalence estimate*

$$\|\boldsymbol{u}^\epsilon - \boldsymbol{u}^0\|_H \simeq \|B(\boldsymbol{u}^\epsilon - \boldsymbol{u}^0)\|_V \simeq \epsilon\,\|A\boldsymbol{u}^0\|_U. \tag{2.4}$$

Therefore, if $A\boldsymbol{u}^0 = 0$, then $\boldsymbol{u}^\epsilon \equiv \boldsymbol{u}^0$. Otherwise $\boldsymbol{u}^\epsilon$ converges to $\boldsymbol{u}^0$ at the sharp rate of ϵ.

Proof. From (2.2) and (2.3), we see that

$$\epsilon(A(\boldsymbol{u}^\epsilon - \boldsymbol{u}^0), A\boldsymbol{v})_U + (B(\boldsymbol{u}^\epsilon - \boldsymbol{u}^0), B\boldsymbol{v})_V = -\,\epsilon(A\boldsymbol{u}^0, A\boldsymbol{v})_U \ \forall \boldsymbol{v} \in H. \tag{2.5}$$

Taking $\boldsymbol{v} = \boldsymbol{u}^\epsilon - \boldsymbol{u}^0$, we get that

$$\epsilon\,\|A(\boldsymbol{u}^\epsilon - \boldsymbol{u}^0)\|_U^2 + \|B(\boldsymbol{u}^\epsilon - \boldsymbol{u}^0)\|_V^2 = -\,\epsilon(A\boldsymbol{u}^0, A(\boldsymbol{u}^\epsilon - \boldsymbol{u}^0))_U. \tag{2.6}$$

It is easy to see that $\boldsymbol{u}^\epsilon$ lies in the subspace $(\ker B)^\perp_{\mathcal{H}}$. So does $\boldsymbol{u}^\epsilon - \boldsymbol{u}^0$. Because B is an isomorphism between $(\ker B)^\perp_{\mathcal{H}}$ and V, we have that

$$\|A(\boldsymbol{u}^\epsilon - \boldsymbol{u}^0)\|_U \lesssim \|\boldsymbol{u}^\epsilon - \boldsymbol{u}^0\|_H \simeq \|B(\boldsymbol{u}^\epsilon - \boldsymbol{u}^0)\|_V. \tag{2.7}$$

By the Cauchy–Schwarz inequality, there exists a constant C such that

$$|\,\epsilon(A\boldsymbol{u}^0, A(\boldsymbol{u}^\epsilon - \boldsymbol{u}^0))_U| \le C\,\epsilon^2\,\|A\boldsymbol{u}^0\|_U^2 + \frac{1}{2}\|B(\boldsymbol{u}^\epsilon - \boldsymbol{u}^0)\|_V^2.$$

This estimate and the equation (2.6) show that

$$\|B(\boldsymbol{u}^\epsilon - \boldsymbol{u}^0)\|_V \lesssim \epsilon\,\|A\boldsymbol{u}^0\|. \tag{2.8}$$

The equivalence (2.1) and (2.7) then lead to $\|\boldsymbol{u}^\epsilon - \boldsymbol{u}^0\|_H \lesssim \epsilon\,\|A\boldsymbol{u}^0\|_U$.

To see the lower bound, in (2.5) we take $\boldsymbol{v} = \boldsymbol{u}^0$ to get

$$\epsilon(A(\boldsymbol{u}^\epsilon - \boldsymbol{u}^0), A\boldsymbol{u}^0)_U + (B(\boldsymbol{u}^\epsilon - \boldsymbol{u}^0), B\boldsymbol{u}^0)_V = -\,\epsilon\,\|A\boldsymbol{u}^0\|_U^2.$$

If $A\boldsymbol{u}^0 \neq 0$, then

$$\epsilon\,\|A\boldsymbol{u}^0\|_U \lesssim \epsilon\,\|A(\boldsymbol{u}^\epsilon - \boldsymbol{u}^0)\|_U + \frac{\|B\boldsymbol{u}^0\|_V}{\|A\boldsymbol{u}^0\|_U}\|B(\boldsymbol{u}^\epsilon - \boldsymbol{u}^0)\|_V \lesssim \|B(\boldsymbol{u}^\epsilon - \boldsymbol{u}^0)\|_V.$$

The equivalence (2.4) then follows. ∎

3. Analysis of domain embedding methods. In this section, we derive the error estimates for the finite element domain embedding methods for various boundary conditions, as described in the Introduction. The main issues appear in Neumann and Dirichlet boundary value problems. In subsection 3.1, we analyze the Neumann problem and assume that the grid on R is uniform. However, all of the results hold for shape regular triangulations. We analyze the Dirichlet problem with uniform mesh in subsection 3.2. The results of this subsection motivate mesh adjustment around Γ for the Dirichlet problem, to which we devote subsection 3.4.

3.1. Neumann boundary condition. We first consider the Neumann problem. Let $\omega \subset \mathbb{R}^2$ be a bounded domain whose boundary Γ is in the C^2 class. Let $f \in L^2(\omega)$. We seek a function u on ω solving the homogeneous Neumann problem whose weak formulation seeks $u \in H^1(\omega)/\mathbb{R}$ such that

$$(\nabla u, \nabla v)_{[L^2(\omega)]^2} = \int_\omega f v dx \quad \forall\ v \in H^1(\omega). \tag{3.1}$$

We assume that $\int_\omega f(x)dx = 0$, so that this is well posed in the quotient space $H^1(\omega)/\mathbb{R}$.

Let $R \subset \mathbb{R}^2$ be a larger rectangular domain such that $\omega \subset\subset R$. Let $\Omega = R \setminus \overline{\omega}$ be the fictitious domain. We extend the function f to a function $\bar{f}$ on R by defining $\bar{f} = 0$ on Ω. The first step in the finite element domain embedding method is replacing (3.1) by the following problem defined on the rectangle R. For $\epsilon > 0$, one determines $u^\epsilon \in H_0^1(R)$ such that

$$(\nabla u^\epsilon, \nabla v)_{[L^2(\omega)]^2} + \epsilon(\nabla u^\epsilon, \nabla v)_{[L^2(\Omega)]^2} = \int_R \bar{f} v dx \ \forall\ v \in H_0^1(R). \tag{3.2}$$

This equation fits in the form of (2.2) in an obvious manner. The condition that B has closed range is assured as long as ω is a Lipschitz domain. All of the other conditions of Lemma 2.4 are also satisfied; see [24] for details. We have that $\lim_{\epsilon\to 0} u^\epsilon = u^0$. The limit $u^0 \in H_0^1(R)$ is characterized by the following two equations:

$$(\nabla u^0, \nabla v)_{[L^2(\omega)]^2} = \int_\omega f v dx \quad \forall\ v \in H^1(\omega), \tag{3.3}$$

i.e., the restriction of u^0 on ω solves the Neumann problem (3.1), and

$$\Delta u^0 = 0 \text{ in } \Omega, \quad \left\langle \frac{\partial u^0}{\partial n^-}, 1 \right\rangle_{H^{-1/2}(\Gamma)\times H^{1/2}(\Gamma)} = 0. \tag{3.4}$$

Here n^- is the outward normal to Γ viewed as a boundary of Ω. By Lemma 2.4 we have that

$$\|u^\epsilon - u^0\|_{1,R} \simeq |u^\epsilon - u^0|_{1,\omega} \simeq \epsilon\, |u^0|_{1,\Omega}. \tag{3.5}$$

It is easy to see that $|u^0|_{1,\Omega} = 0$ if and only if $u^0 = 0$ on Γ, in which case $u^\epsilon \equiv u^0$. This is the exceptional case in which $u^0 \in H^2(R)$. Generally, u^0 is only piecewise smooth, with a normal derivative jump across Γ.

We introduce a uniform triangulation of mesh size h on R; see Figure 1 (left), and let $H_h \subset H_0^1(R)$ be the subspace of piecewise linear continuous functions subordinate to this triangulation. The finite element domain embedding method is a straightforward discretization of (1.2). It determines $u_h^\epsilon \in H_h$ such that

$$(\nabla u_h^\epsilon, \nabla v_h)_{[L^2(\omega)]^2} + \epsilon(\nabla u_h^\epsilon, \nabla v_h)_{[L^2(\Omega)]^2} = \int_R \bar{f} v_h dx \quad \forall\ v \in H_h. \tag{3.6}$$

The numerical approximation of the solution of (3.1) offered by the finite element domain embedding method is the restriction of u_h^ϵ on ω. We have the following theorem.

THEOREM 3.1. *Let $u^0 \in H_0^1(R)$ be the limit of u^ϵ that is determined by the domain embedding equation* (3.2). *Then $u^0|_\omega$ solves the Neumann problem* (3.1). *Let $u_h^\epsilon \in H_h$ be the solution of the finite element domain embedding equation* (3.6). *We have the following estimate on the difference between $u_h^\epsilon|_\omega$ and $u^0|_\omega$:*

$$|u_h^\epsilon - u^0|_{1,\omega} \le C\,\epsilon\,|u^0|_{1,\Omega} + C \inf_{v_h \in H_h} \left(|u^0 - v_h|_{1,\omega} + \sqrt{\epsilon}|u^0 - v_h|_{1,\Omega}\right). \tag{3.7}$$

Here, C is a constant independent of h, ϵ, and f.

Proof. Since u_h^ϵ is the projection of u^ϵ in the subspace H_h with respect to the inner product defined by the bilinear form in the left-hand side of the domain embedding equation (3.2), we have that

$$|u^\epsilon - u_h^\epsilon|_{1,\omega} + \sqrt{\epsilon}|u^\epsilon - u_h^\epsilon|_{1,\Omega} \le |u^\epsilon - v_h|_{1,\omega} + \sqrt{\epsilon}|u^\epsilon - v_h|_{1,\Omega} \quad \forall\, v_h \in H_h.$$

Thus by (3.5)

$$\begin{aligned}
|u^0 - u_h^\epsilon|_{1,\omega} &\le |u^0 - u^\epsilon|_{1,\omega} + |u^\epsilon - u_h^\epsilon|_{1,\omega} \\
&\le \epsilon\,|u^0|_{1,\Omega} + |u^\epsilon - v_h|_{1,\omega} + \sqrt{\epsilon}|u^\epsilon - v_h|_{1,\Omega} \\
&\le \epsilon\,|u^0|_{1,\Omega} + |u^0 - v_h|_{1,\omega} + |u^\epsilon - u^0|_{1,\omega} + \sqrt{\epsilon}|u^0 - v_h|_{1,\Omega} + \sqrt{\epsilon}|u^\epsilon - u^0|_{1,\Omega} \\
&\le (2\,\epsilon + \epsilon^{3/2})|u^0|_{1,\Omega} + |u^0 - v_h|_{1,\omega} + \sqrt{\epsilon}|u^0 - v_h|_{1,\Omega} \quad \forall\, v_h \in H_h.
\end{aligned}$$

The estimate thus follows. ∎

This theorem is valid as long as ω is Lipschitz. Based on this result, we can establish the following error estimate on the finite element domain embedding method for the Neumann problem under our assumptions that $\Gamma \in C^2$ and $f \in L^2(\omega)$. Under these assumptions, by the regularity of elliptic differential equation [13], we have that $u^0|_\omega \in H^2(\omega)$ and $\|u^0\|_{2,\omega} \lesssim \|f\|_{0,\omega}$. Since $u^0|_\Omega$ is harmonic and it shares the value with $u^0|_\omega$ on Γ, we have that $\|u^0\|_{2,\Omega} \lesssim \|u^0\|_{1.5,\Gamma} \lesssim \|u^0\|_{2,\omega} \lesssim \|f\|_{0,\omega}$. We can extend $u^0|_\omega$ to R; see [21], to obtain a function $\bar{u}^0 \in H^2(R)$ such that $\bar{u}^0|_\omega = u^0|_\omega$ and $\|\bar{u}^0\|_{2,R} \lesssim \|u^0\|_{2,\omega}$. Similarly, we extend $u^0|_\Omega$ to R to obtain a function $\underline{u}^0 \in H^2(R)$ such that $\underline{u}^0|_\Omega = u^0|_\Omega$ and $\|\underline{u}^0\|_{2,R} \lesssim \|u^0\|_{2,\Omega}$. Note that $\bar{u}^0|_\Omega \ne u^0|_\Omega$, $\underline{u}^0|_\omega \ne u^0|_\omega$, and $\bar{u}^0 = \underline{u}^0 = u^0$ on Γ.

THEOREM 3.2. *Under the assumption that $\Gamma \in C^2$ and $f \in L^2(\omega)$, there is a positive constant C that is independent of ϵ, h, and f such that the following error estimate of the finite element domain embedding method for the Neumann problems holds:*

$$|u_h^\epsilon - u^0|_{1,\omega} \le C\left(\epsilon + h + \sqrt{\epsilon\,h}\right)\|f\|_{0,\omega}. \tag{3.8}$$

Before proving this theorem, we introduce some notations to describe triangles, vertices, and their relative relations with Γ. Let $\mathcal{T}$ be the set of all the open triangular elements of the finite element partition of R. We divide $\mathcal{T}$ into three distinctive subsets such that $\mathcal{T} = \mathcal{T}_\omega \cup \mathcal{T}_\Gamma \cup \mathcal{T}_\Omega$, and

$$\mathcal{T}_\Gamma = \{\tau \in \mathcal{T}; \tau \cap \Gamma \ne \emptyset\},\ \mathcal{T}_\omega = \{\tau \in \mathcal{T}; \tau \subset \omega\},\ \text{and}\ \mathcal{T}_\Omega = \{\tau \in \mathcal{T}; \tau \subset \Omega\}.$$

Among the triangles of $\mathcal{T}_\omega$, we let $\mathcal{T}_\omega^0$ be the subset of those that share a vertex with some triangle in $\mathcal{T}_\Gamma$, and we define $\mathcal{T}_\Omega^1$ as a subset of $\mathcal{T}_\Omega$ similarly. See Figure 4 in

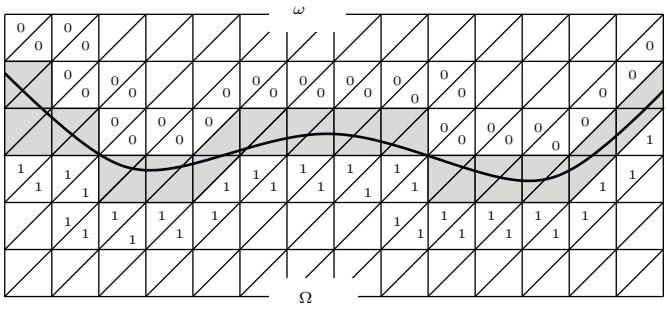


FIG. 4. *A portion of uniform mesh around* Γ.

which triangles in $\mathcal{T}_\Gamma$ are shaded, those in $\mathcal{T}_\omega^0$ are marked by 0, and those in $\mathcal{T}_\Omega^1$ are marked by 1.

With a slight abuse of notations, we shall also use $\mathcal{T}_\Gamma$ to denote the domain formed by the union of all triangles in $\mathcal{T}_\Gamma$, which is the interior of $\cup_{\tau\in\mathcal{T}_\Gamma}\bar{\tau}$. We use $\mathcal{T}$, $\mathcal{T}_\omega$, $\mathcal{T}_\Omega$, $\mathcal{T}_\omega^0$, and $\mathcal{T}_\Omega^1$ in the same way. Thus $\mathcal{T} = R$, and $\omega \subset \overline{\mathcal{T}}_\omega \cup \mathcal{T}_\Gamma$, etc.

Let $\mathcal{V}$ be the set of all the vertices in R of the triangulation. We divide $\mathcal{V}$ into three distinctive subsets such that $\mathcal{V} = \mathcal{V}_\omega \cup \mathcal{V}_\Gamma \cup \mathcal{V}_\Omega$, in which

$$\begin{aligned}\mathcal{V}_\Gamma &= \{\nu \in \mathcal{V}; \nu \text{ is a vertex of a triangle of } \mathcal{T}_\Gamma, \text{ or } \nu \in \Gamma\},\\ \mathcal{V}_\omega &= \{\nu \in \mathcal{V}; \nu \in \omega, \text{ and } \nu \text{ is not a vertex in } \mathcal{V}_\Gamma\},\\ \mathcal{V}_\Omega &= \{\nu \in \mathcal{V}; \nu \in \Omega, \text{ and } \nu \text{ is not a vertex in } \mathcal{V}_\Gamma\}.\end{aligned}$$

Note that $\mathcal{T}_\Gamma$ could be empty, for example, when Γ is straight and aligns with the triangulation, in which case $\mathcal{V}_\Gamma$ is composed of nodes on Γ. We let

$$\begin{aligned}\mathcal{V}_\omega^0 &= \{\nu \in \mathcal{V}_\omega;\ \nu \text{ is a vertex of a triangle in } \mathcal{T}_\omega^0\},\\ \mathcal{V}_\Omega^1 &= \{\nu \in \mathcal{V}_\Omega;\ \nu \text{ is a vertex of a triangle in } \mathcal{T}_\Omega^1\}.\end{aligned}$$

Proof of Theorem 3.2. What we need to do is to construct a finite element interpolation $v_h \in H_h$ to u^0 such that $|u^0 - v_h|_{1,\omega} \lesssim h$ and $|u^0 - v_h|_{1,\Omega} \lesssim \sqrt{h}$. We define v_h by requiring $v_h(\nu) = \bar{u}^0(\nu)$ for $\nu \in \mathcal{V}_\omega \cup \mathcal{V}_\Gamma$, and $v_h(\nu) = u^0(\nu)$ for $\nu \in \mathcal{V}_\Omega$. Note that the defined interpolation interpolates u^0 on all of the vertices except those in $\mathcal{V}_\Gamma \cap \Omega$. By the standard argument of finite element interpolation, we have that

$$|v_h - \bar{u}^0|_{1,\tau}^2 \le Ch^2 |\bar{u}^0|_{2,\tau}^2 \quad \forall\ \tau \in \mathcal{T}_\omega \cup \mathcal{T}_\Gamma.$$

Summing up these estimates, we get that

$$|v_h - \bar{u}^0|_{1,\omega\cup\mathcal{T}_\Gamma}^2 \le Ch^2 |\bar{u}^0|_{2,\omega\cup\mathcal{T}_\Gamma}^2 \le Ch^2 |\bar{u}^0|_{2,R}^2 \le Ch^2 |u^0|_{2,\omega}^2 \le Ch^2 \|f\|_{0,\omega}^2.$$

Thus

$$|v_h - u^0|_{1,\omega}^2 = |v_h - \bar{u}^0|_{1,\omega}^2 \le Ch^2 \|f\|_{0,\omega}^2. \tag{3.9}$$

By the standard argument of finite element interpolation, we have that

$$|v_h - u^0|_{1,\tau}^2 \le Ch^2 |u^0|_{2,\tau}^2 \quad \forall\ \tau \in \mathcal{T}_\Omega \setminus \mathcal{T}_\Omega^1.$$

Summing up, we get that

$$|v_h - u^0|^2_{1,\mathcal{T}_\Omega \setminus \overline{\mathcal{T}}^1_\Omega} \le Ch^2 |u^0|^2_{2,\Omega} \le Ch^2 \|f\|^2_{0,\omega}. \tag{3.10}$$

The remaining part of Ω, i.e., $\mathcal{T}^1_\Omega \cup (\Omega \cap \mathcal{T}_\Gamma)$, is covered by a strip of width $\mathcal{O}(h)$, which meets the definition of a γ_δ of section 2, with $\delta = \mathcal{O}(h)$. Thus, according to Lemma 2.2, we have that

$$|u^0|^2_{1,\mathcal{T}^1_\Omega \cup (\Omega \cap \mathcal{T}_\Gamma)} \le Ch \|\underline{u}^0\|^2_{2,R} \le Ch \|f\|^2_{0,\omega}. \tag{3.11}$$

We are left with estimating $|v_h|^2_{1,\mathcal{T}^1_\Omega \cup (\Omega \cap \mathcal{T}_\Gamma)}$. We consider this separately on $\Omega \cap \mathcal{T}_\Gamma$ and $\mathcal{T}^1_\Omega$. We see that $|v_h|^2_{1,\Omega \cap \mathcal{T}_\Gamma} \le |v_h|^2_{1,\mathcal{T}_\Gamma}$. Since v_h interpolates $\bar{u}^0$ at every vertex of triangles in $\mathcal{T}_\Gamma$, which itself is a cluster of shape regular triangles contained in a strip of width $\mathcal{O}(h)$, by Lemma 2.3, we have that $|v_h|^2_{1,\mathcal{T}_\Gamma} \le Ch \|\bar{u}^0\|^2_{2,R}$. Thus

$$|v_h|^2_{1,\Omega \cap \mathcal{T}_\Gamma} \le Ch \|\bar{u}^0\|^2_{2,R} \le Ch \|f\|^2_{0,\omega}. \tag{3.12}$$

To estimate $|v_h|^2_{1,\mathcal{T}^1_\Omega}$, we introduce an auxiliary interpolation $I_{\mathcal{T}^1_\Omega} u^0$ that interpolates u^0 on every vertex of triangles in $\mathcal{T}^1_\Omega$. Since $\mathcal{T}^1_\Omega$ is composed of a cluster of shape regular triangles covered by a strip of width h, by Lemma 2.3, we have that

$$|I_{\mathcal{T}^1_\Omega} u^0|^2_{1,\mathcal{T}^1_\Omega} \le Ch \|\underline{u}^0\|^2_{2,R} \le Ch \|f\|^2_{0,\omega}. \tag{3.13}$$

From the definition we see that $v_h(\nu) - I_{\mathcal{T}^1_\Omega} u^0(\nu) = 0$ for all $\nu \in \mathcal{V}^1_\Omega$, and $v_h(\nu) - I_{\mathcal{T}^1_\Omega} u^0(\nu) = \bar{u}^0(\nu) - u^0(\nu)$ for all $\nu \in \mathcal{V}_\Gamma \cap \Omega$. Thus we have that

$$|v_h - I_{\mathcal{T}^1_\Omega} u^0|^2_{1,\mathcal{T}^1_\Omega} \le C \sum_{\nu \in \mathcal{V}_\Gamma \cap \Omega} (v_h(\nu) - I_{\mathcal{T}^1_\Omega} u^0(\nu))^2 = C \sum_{\nu \in \mathcal{V}_\Gamma \cap \Omega} (\bar{u}^0(\nu) - u^0(\nu))^2.$$

For each $\nu \in \mathcal{V}_\Gamma \cap \Omega$, we construct a shape regular triangle τ_ν, with ν being one of its vertices, and the other two vertices sitting on Γ. Let $I_{\tau_\nu}(\bar{u}^0 - \underline{u}^0)$ be the linear interpolation of $\bar{u}^0 - \underline{u}^0$ on τ_ν. Since $\bar{u}^0 = \underline{u}^0$ on Γ, we have that

$$|I_{\tau_\nu}(\bar{u}^0 - \underline{u}^0)|^2_{1,\tau_\nu} \simeq (\bar{u}^0(\nu) - u^0(\nu))^2.$$

Thus

$$\sum_{\nu \in \mathcal{V}_\Gamma \cap \Omega} (\bar{u}^0(\nu) - u^0(\nu))^2 \le C \sum_{\nu \in \mathcal{V}_\Gamma \cap \Omega} |I_{\tau_\nu}(\bar{u}^0 - \underline{u}^0)|^2_{1,\tau_\nu}.$$

The cluster of triangles $\{\tau_\nu; \nu \in \mathcal{V}_\Gamma \cap \Omega\}$ satisfies the condition of Lemma 2.3, with $\delta = h$. We therefore have that

$$\sum_{\nu \in \mathcal{V}_\Gamma \cap \Omega} |I_{\tau_\nu}(\bar{u}^0 - \underline{u}^0)|^2_{1,\tau_\nu} \le Ch \|\bar{u}^0 - \underline{u}^0\|^2_{2,R}.$$

From these estimates, we see that

$$|v_h - I_{\mathcal{T}^1_\Omega} u^0|^2_{1,\mathcal{T}^1_\Omega} \le Ch \|\bar{u}^0 - \underline{u}^0\|^2_{2,R} \le Ch \|f\|^2_{0,\omega}. \tag{3.14}$$

Therefore, by (3.13) and (3.14),

$$|v_h|^2_{1,\mathcal{T}^1_\Omega} \le |I_{\mathcal{T}^1_\Omega} u^0|^2_{1,\mathcal{T}^1_\Omega} + |v_h - I_{\mathcal{T}^1_\Omega} u^0|^2_{1,\mathcal{T}^1_\Omega} \le Ch(\|\underline{u}^0\|^2_{2,R} + \|\bar{u}^0 - \underline{u}^0\|^2_{2,R}) \le Ch \|f\|^2_{0,\omega}.$$

Combining this and (3.12), we get that

$$|v_h|^2_{1,\mathcal{T}^1_\Omega \cup (\Omega \cap \mathcal{T}_\Gamma)} \leq Ch\|f\|^2_{0,\omega}. \tag{3.15}$$

The theorem now follows from (3.9), (3.10), (3.11), and (3.15). ∎

From this theorem we see that, when $\Gamma \in C^2$ and $f \in L^2(\omega)$ and if we take $\epsilon = Ch$, we have a domain embedding finite element method on uniform mesh that achieves the full accuracy with an error of $\mathcal{O}(h)$. Although the triangulation was assumed to be uniform, it is clear that Theorem 3.2 holds for any triangulation that is shape regular, (but not necessarily quasi uniform,) with a maximum mesh size h. The result of Theorem 3.1 is valid as long as ω is a Lipschitz domain, which does not guarantee the piecewise regularity of u^0. For example, when ω is a convex polygon, then $u^0|_\omega \in H^2(\omega)$, but $u^0|_\Omega$ cannot be expected to have the H^2 regularity. If ω is a convex polygon and Γ aligns with the mesh, one can define v_h by simply requiring it to interpolate u^0 at every node in $\mathcal{V}$. Then one uses the regularity $u^0|_\omega \in H^2(\omega)$ and $u^0|_\Omega \in H^{1.5}(\Omega)$ [16] to show the validity of the estimate in Theorem 3.2. If ω is a nonconvex polygon, then $u^0|_\omega \in H^2(\omega)$ may not hold, in which case, a local refinement may be needed at reentrant corners of ω to make $|u^0 - v_h|_{1,\omega} \lesssim h$. This can be done by taking a small rectangle around a corner, which conforms to the uniform mesh, and on which one uses a fine uniform mesh. Thus doing, fast Fourier transforms on the global uniform mesh and local uniform fine mesh can be combined in preconditioning the discrete system, in view of the additive multilevel Schwarz method or the Chimera method [9].

For a Robin boundary condition, the theory is similar to that of the Neumann problem, so we will be very brief. We seek a function u on ω solving the homogeneous Robin problem

$$-\Delta u = f \text{ on } \omega, \quad \frac{\partial u}{\partial n} + ku = 0 \text{ on } \Gamma. \tag{3.16}$$

Here k is a smooth, bounded, and positively valued function on Γ. The weak formulation is

$$\begin{gathered}(\nabla u, \nabla v)_{[L^2(\omega)]^2} + \int_\Gamma kuvds = \int_\omega fvdx \ \ \forall\ v \in H^1(\omega), \\ u \in H^1(\omega).\end{gathered} \tag{3.17}$$

On the continuous level, the domain embedding method determines $u^\epsilon \in H^1_0(R)$ such that

$$(\nabla u^\epsilon, \nabla v)_{[L^2(\omega)]^2} + \int_\Gamma kuvds + \epsilon(\nabla u^\epsilon, \nabla v)_{[L^2(\Omega)]^2} = \int_R \bar{f}vdx \ \forall\ v \in H^1_0(R). \tag{3.18}$$

This equation also fits in the form of (2.2), in which the B operator has closed range in V if ω is Lipschitz. Therefore, as $\epsilon \to 0$, u^ϵ converges to a limit $u^0 \in H^1_0(R)$. The limit is harmonic on Ω, and $u^0|_\omega$ solves the Robin problem (3.17). By Lemma 2.4 we have that

$$\|u^\epsilon - u^0\|_{1,R} \simeq |u^\epsilon - u^0|_{1,\omega} + \|u^\epsilon - u^0\|_{0,\Gamma} \simeq \epsilon\,|u^0|_{1,\Omega}. \tag{3.19}$$

The finite element domain embedding method is a straightforward discretization of (3.18) on a uniform mesh on R. The same results of Theorems 3.1 and 3.2 hold for the Robin problem, and the argument is exactly the same.

The Robin problem provides an alternative approach to designing domain embedding for the Dirichlet problem, other than the one we analyze below. In the equation (3.18), we can fix ϵ and take k as a parameter. When $k \to \infty$, the solution of (3.18) then converges to a function (at a rate much lower than $1/k$ [22]) whose restriction on ω solves the homogeneous Dirichlet problem. This method is essentially that of [4]. But the above theory of the domain embedding method for the Robin problem does not extend to this approach for the Dirichlet problem.

3.2. Dirichlet boundary condition. We seek a function u on ω solving the homogeneous Dirichlet problem whose weak formulation is

$$(\nabla u, \nabla v)_{[L^2(\omega)]^2} = \int_\omega f v dx \quad \forall\ v \in H_0^1(\omega), \qquad u \in H_0^1(\omega). \tag{3.20}$$

Under the assumption that ω is simply connected, on the continuous level, the domain embedding method determines $u^\epsilon \in H_0^1(R)$ such that

$$(\nabla u^\epsilon, \nabla v)_{[L^2(\omega)]^2} + \epsilon^{-1}(\nabla u^\epsilon, \nabla v)_{[L^2(\Omega)]^2} = \int_R \bar{f} v dx \quad \forall\ v \in H_0^1(R). \tag{3.21}$$

This is a penalty formulation. We extend u to a function u^0 on R such that $u^0|_\omega = u$ and $u^0|_\Omega = 0$. It is easy to see that, for all $v \in H_0^1(R)$,

$$\epsilon(\nabla(u^\epsilon - u^0), \nabla v)_{[L^2(\omega)]^2} + (\nabla(u^\epsilon - u^0), \nabla v)_{[L^2(\Omega)]^2} = \epsilon \left[\int_R \bar{f} v dx - (\nabla u^0, \nabla v)_{[L^2(\omega)]^2} \right]. \tag{3.22}$$

This equation fits in the form of (2.2) by properly defining the operators and spaces. We see that

$$\ker B = \{v \in H_0^1(R); v = 0 \text{ on } \Omega\},$$

and, more importantly, the right-hand side of (3.22), as a functional, annihilates this subspace. Thus there is a unique $u^1 \in H_0^1(R)$ such that $\Delta u^1 = 0$ on ω (since $u^1 \in (\ker B)^\perp_{\mathcal{H}}$), and

$$(\nabla u^1, \nabla v)_{[L^2(\Omega)]^2} = \left[\int_R \bar{f} v dx - (\nabla u^0, \nabla v)_{[L^2(\omega)]^2} \right] \quad \forall\ v \in H_0^1(R). \tag{3.23}$$

In strong form, this function satisfies

$$-\Delta u^1 = 0 \text{ in } \Omega, \quad \frac{\partial u^1}{\partial n^-} = \frac{\partial u^0}{\partial n} \text{ on } \Gamma, \quad \text{and } u^1 = 0 \text{ on } \partial R. \tag{3.24}$$

Here n is the unit outward normal to Γ viewed as a boundary of ω, and n^- is the outward normal of Ω, which is opposite to n. When Ω is a Lipschitz domain, the operator B arising from fitting (3.22) into (2.2) has closed range. By Lemma 2.4 we have that

$$\|u^\epsilon - u^0 - \epsilon\, u^1\|_{1,R} \simeq |u^\epsilon - u^0 - \epsilon\, u^1|_{1,\Omega} \simeq \epsilon^2\, |u^1|_{1,\omega}. \tag{3.25}$$

From (3.24) we see that $|u^1|_{1,\omega} = 0$, which is equivalent to $|u^1|_{1,\Omega} = 0$ if and only if u^0 satisfies the homogeneous Neumann condition on Γ, or equivalently, $u^0 \in H^2(R)$. In this case $u^\epsilon \equiv u^0$, otherwise u^ϵ converges to u^0 at the sharp rate of ϵ.

Remark. If ω is not simply connected, then Ω could have isolated components whose union is, say, Ω_0. In this case, the limit u^0 of u^ϵ does not satisfy the homogeneous Dirichlet condition on Γ, and the domain embedding method (3.21) fails. A remedy is adding, for example, the term $\epsilon^{-1}(u^\epsilon, v)_{L^2(\Omega_0)}$ or $\epsilon^{-1}(u^\epsilon, v)_{L^2(\Gamma)}$ (or something else as long as $\ker B$ can be identified with $H_0^1(\omega)$) to the left-hand side of the domain embedding equation (3.21). Then the method would be convergent, and the convergence rate would be retained.

The finite element domain embedding method is a straightforward discretization of (3.21) in the finite element space $H_h \subset H_0^1(R)$. In this subsection, as for the Neumann and Robin problems, we assume the triangulation is uniform. The numerical method determines $u_h^\epsilon \in H_h$ such that

$$(\nabla u_h^\epsilon, \nabla v_h)_{[L^2(\omega)]^2} + \epsilon^{-1}(\nabla u_h^\epsilon, \nabla v_h)_{[L^2(\Omega)]^2} = \int_R \bar{f} v_h dx \quad \forall\ v \in H_h. \tag{3.26}$$

We have the following estimate on the error of this method.

THEOREM 3.3. *Let $u^0 \in H_0^1(R)$ be the limit of u^ϵ determined by the domain embedding equation* (3.21)*. Then $u^0|_\omega$ solves the Dirichlet problem* (3.20)*, and $u^0|_\Omega = 0$. Let $u_h^\epsilon \in H_h$ be the solution of the finite element domain embedding equation* (3.26)*. We have that*

$$\begin{aligned}\|u_h^\epsilon - u^0\|_{1,\omega} \le C\,\epsilon\,|u^1|_{1,R} + C\sqrt{\epsilon} \inf_{v_h^1 \in H_h} \left(\sqrt{\epsilon}|u^1 - v_h^1|_{1,\omega} + |u^1 - v_h^1|_{1,\Omega}\right) \\ + C \inf_{v_h^0 \in H_h} \left(|u^0 - v_h^0|_{1,\omega} + \frac{1}{\sqrt{\epsilon}}|v_h^0|_{1,\Omega}\right).\end{aligned} \tag{3.27}$$

Here u^1 is defined by (3.23)*, and C is a constant independent of h, ϵ, and f.*

Proof. Since u_h^ϵ is the projection of u^ϵ in the subspace H_h with respect to the inner product defined by the bilinear form in (3.21), we have that

$$|u^\epsilon - u_h^\epsilon|_{1,\omega} + \frac{1}{\sqrt{\epsilon}}|u^\epsilon - u_h^\epsilon|_{1,\Omega} \le |u^\epsilon - v_h|_{1,\omega} + \frac{1}{\sqrt{\epsilon}}|u^\epsilon - v_h|_{1,\Omega} \quad \forall\ v_h \in H_h.$$

Thus, by using (3.25), we see that

$$\begin{aligned}\|u^0 - u_h^\epsilon\|_{1,\omega} le\ C|u^0 - u_h^\epsilon|_{1,R} &\le C\Big(|u^0 - u^\epsilon|_{1,R} + |u^\epsilon - u_h^\epsilon|_{1,R}\Big) \\ &\le C\left(\epsilon\,|u^1|_{1,R} + |u^\epsilon - v_h|_{1,\omega} + \frac{1}{\sqrt{\epsilon}}|u^\epsilon - v_h|_{1,\Omega}\right) \\ &\le C\Bigg(\epsilon\,|u^1|_{1,R} + |u^\epsilon - u^0 - \epsilon\,u^1|_{1,\omega} + |u^0 + \epsilon\,u^1 - v_h|_{1,\omega} \\ &\quad + \frac{1}{\sqrt{\epsilon}}(|u^\epsilon - u^0 - \epsilon\,u^1|_{1,\Omega} + |\,\epsilon\,u^1 - v_h|_{1,\Omega})\Bigg) \quad \forall\ v_h \in H_h.\end{aligned}$$

We write any $v_h \in H_h$ as $v_h = v_h^0 + \epsilon\,v_h^1$ and use (3.25) again to obtain that

$$\begin{aligned}\|u^0 - u_h^\epsilon\|_{1,\omega} \le C[(\epsilon + \epsilon^{1.5} + \epsilon^2)|u^1|_{1,R} + \epsilon\,|u^1 - v_h^1|_{1,\omega} + \sqrt{\epsilon}|u^1 - v_h^1|_{1,\Omega} \\ + |u^0 - v_h^0|_{1,\omega} + \frac{1}{\sqrt{\epsilon}}|v_h^0|_{1,\Omega}].\end{aligned}$$

The desired estimate then follows. ∎

This result is valid as long as Ω is a Lipschitz domain. In the following, we assume that $\omega \subset\subset R$, $\Gamma \in C^2$, and $f \in L^2(\omega)$. Under this assumption, we have the regularity that $u^0|_\omega \in H^2(\omega) \cap H_0^1(\omega)$. Thus $\partial u^0/\partial n \in H^{1/2}(\Gamma)$. This in turn means that $\partial u^1/\partial n^- \in H^{1/2}(\Gamma)$; cf. (3.24). Therefore, $u^1|_\Omega \in H^2(\Omega)$. Since $u^1|_\omega$ is harmonic and shares values with $u^1|_\Omega$ on Γ, so $u^1|_\omega \in H^2(\omega)$. Furthermore, we have that the estimates $\|u^0\|_{2,\omega} \le C\|f\|_{0,\omega}$, $\|u^1\|_{2,\omega} \le C\|f\|_{0,\omega}$, and $\|u^1\|_{2,\Omega} \le C\|f\|_{0,\omega}$. We shall also use the fact that $u^0|_\omega$ has a H^2 extension onto R [21], denoted by $\bar{u}^0$, such that $\|\bar{u}^0\|_{2,R} \le C\|u^0\|_{2,\omega} \le C\|f\|_{0,\omega}$.

THEOREM 3.4. *Under the assumption that $\omega \subset\subset R$, $\Gamma \in C^2$, and $f \in L^2(\omega)$, for the Dirichlet problem, the following error estimate of the finite element domain embedding method based on uniform mesh holds:*

$$\|u^0 - u_h^\epsilon\|_{1,\omega} \le C(\epsilon + \epsilon\sqrt{h} + h\sqrt{\epsilon} + \sqrt{h})\|f\|_{0,\omega}.$$

Here, C is a constant independent of ϵ, h, and f.

Proof. We need to estimate the three terms in the right-hand side of (3.27). The first one is trivial. For the other two, we need to construct finite element interpolation v_h^1 to u^1 and v_h^0 to u^0, respectively. The construction of v_h^1 aims to estimate the infimum

$$\inf_{v_h^1 \in H_h} \left(\sqrt{\epsilon}|u^1 - v_h^1|_{1,\omega} + |u^1 - v_h^1|_{1,\Omega}\right).$$

We observe that this is in the same form as the second term in the right-hand side of the estimate for the Neumann problem; c.f. (3.7), except that the roles of Ω and ω are switched. We use exactly the same technique of proving Theorem 3.2 to estimate this term. We define $v_h^1 \in H_h$ such that $v_h^1(\nu) = \underline{u}^1(\nu)$ for all $\nu \in \mathcal{V}_\Omega \cup \mathcal{V}_\Gamma$, and $v_h^1(\nu) = u^1(\nu)$ for all $\nu \in \mathcal{V}_\omega$. Here $\underline{u}^1 \in H^2(R)$ is an extension of $u^1|_\Omega$ to R such that $\underline{u}^1|_\Omega = u^1|_\Omega$. With such defined interpolation, we have that

$$|u^1 - v_h^1|_{1,\Omega} \le Ch\|f\|_{0,\omega} \text{ and } |u^1 - v_h^1|_{1,\omega} \le C\sqrt{h}\|f\|_{0,\omega}.$$

Thus

$$\inf_{v_h^1 \in H_h} \left(\sqrt{\epsilon}|u^1 - v_h^1|_{1,\omega} + |u^1 - v_h^1|_{1,\Omega}\right) \le C(\sqrt{\epsilon\, h} + h)\|f\|_{0,\omega}. \tag{3.28}$$

We now turn to the last term in (3.27); i.e.,

$$\inf_{v_h^0 \in H_h} \left(|u^0 - v_h^0|_{1,\omega} + \frac{1}{\sqrt{\epsilon}}|v_h^0|_{1,\Omega}\right). \tag{3.29}$$

The best choice of the interpolation to estimate this one is taking $v_h^0 \in H_h$ such that

$$v_h^0(\nu) = u^0(\nu) \ \ \forall\ \nu \in \mathcal{V}_\omega, \quad v_h^0(\nu) = 0 \ \ \forall\ \nu \in \mathcal{V}_\Gamma \cup \mathcal{V}_\Omega.$$

(See the remark below for the reason why this is the best.) This choice makes $|v_h^0|_{1,\Omega} = 0$. To estimate the other term in (3.29), we observe that

$$|u^0 - v_h^0|_{1,\omega}^2 = |u^0 - v_h^0|_{1,\mathcal{T}_\omega \setminus \mathcal{T}_\omega^0}^2 + |u^0 - v_h^0|_{1,\mathcal{T}_\omega^0}^2 + |u^0|_{1,\omega \cap \mathcal{T}_\Gamma}^2. \tag{3.30}$$

Note that v_h^0 interpolates u^0 at all of the vertices of triangles in $\mathcal{T}_\omega \setminus \mathcal{T}_\omega^0$. On each of the triangle $\tau \in \mathcal{T}_\omega \setminus \mathcal{T}_\omega^0$, we have that $|u^0 - v_h^0|_{1,\tau}^2 \le Ch^2|u^0|_{2,\tau}^2$. Summing up, we get that

$$|u^0 - v_h^0|_{1,\mathcal{T}_\omega \setminus \mathcal{T}_\omega^0}^2 \le Ch^2|u^0|_{2,\mathcal{T}_\omega \setminus \mathcal{T}_\omega^0}^2 \le Ch^2|u^0|_{2,\omega}^2. \tag{3.31}$$

Since $\mathcal{T}_\omega^0$ is covered by a strip of width $\mathcal{O}(h)$ that satisfies the condition of Lemma 2.2, so we have that

$$|u^0|^2_{1,\mathcal{T}_\omega^0} \le Ch\|\bar{u}^0\|^2_{2,R}. \tag{3.32}$$

Here $\bar{u}^0$ is an H^2 extension of $u^0|_\omega$ to R such that $\|\bar{u}^0\|_{2,R} \le C\|u^0\|_{2,\omega}$. Note that a vertex of any triangle in $\mathcal{T}_\omega^0$ either belongs to $\mathcal{V}_\omega^0$ or belongs to $\omega \cap \mathcal{V}_\Gamma$. Since $v_h^0(\nu) = u^0(\nu)$ for all $\nu \in \mathcal{V}_\omega^0$ and $v_h^0(\nu) = 0$ for all $\nu \in \omega \cap \mathcal{V}_\Gamma$, we have that $|v_h^0|^2_{1,\mathcal{T}_\omega^0} \le C\sum_{\nu\in\mathcal{V}_\omega^0}(u^0(\nu))^2$. For each $\nu \in \mathcal{V}_\omega^0$, we construct a shape regular triangle τ_ν with ν being a vertex, and the other two vertices sitting on Γ. Let $I_{\tau_\nu}\bar{u}^0$ be the linear interpolation of $\bar{u}^0$ on this triangle. We have that $(u^0(\nu))^2 \simeq |I_{\tau_\nu}\bar{u}^0|^2_{1,\tau_\nu}$. Thus

$$|v_h^0|^2_{1,\mathcal{T}_\omega^0} \le C\sum_{\nu\in\mathcal{V}_\omega^0} |I_{\tau_\nu}\bar{u}^0|^2_{1,\tau_\nu} \le Ch\|\bar{u}^0\|^2_{2,R}. \tag{3.33}$$

In the last step, we used Lemma 2.3. Finally, according to Lemma 2.2, we have that

$$|u^0|^2_{1,\omega\cap\mathcal{T}_\Gamma} \le Ch\|\bar{u}^0\|^2_{2,R}. \tag{3.34}$$

The estimates (3.32), (3.33), and (3.34) together lead to

$$|u^0 - v_h^0|^2_{1,\mathcal{T}_\omega^0} + |u^0|^2_{1,\omega\cap\mathcal{T}_\Gamma} \le Ch\|\bar{u}^0\|^2_{2,R}. \tag{3.35}$$

This and (3.31) show that $|u^0 - v_h^0|_{1,\omega} \le C\sqrt{h}\|\bar{u}^0\|_{2,R}$, which together with (3.27) and (3.28) complete the proof. ∎

3.3. Remark on the sharpness of the estimate on (3.29). The estimated upper bound $C\sqrt{h}$ on (3.29) is rather big. This estimate was established by taking the interpolation v_h^0 to u^0. One of the reasons for the estimate to be so big is that the chosen v_h^0 does not offer any approximation to u^0 on $\omega \cap \mathcal{T}_\Gamma$, where v_h^0 is identically equal to zero. Even so, the estimate is sharp. To see this, we consider an example of a homogeneous Dirichlet boundary value problem in which ω is the unit circle, $f = 1$, and the solution is $u^0|_\omega = (1 - x^2 - y^2)/2$. We embed ω in a larger rectangle R, on which we introduce a uniform triangulation of mesh size h. The unit circle inevitably cuts through Ch triangles. The area of $\omega \cap \mathcal{T}_\Gamma$ is of the magnitude h. We have that $|u^0|_{1,\omega\cap\mathcal{T}_\Gamma} \simeq \sqrt{h}$. The interpolation v_h^0, defined in the proof, is zero on $\omega \cap \mathcal{T}_\Gamma$, which makes the error bound to be at least $C\sqrt{h}$. But the estimate $C\sqrt{h}$ is sharp. One may try to construct a "better" piecewise linear approximation to u^0 to make $|u^0 - v_h^0|_{1,\omega\cap\mathcal{T}_\Gamma}$ smaller. However, if the v_h^0 were to offer any approximation to u^0, one would at least have that $|v_h^0|_{1,\omega\cap\mathcal{T}_\Gamma} \simeq |u^0|_{1,\omega\cap\mathcal{T}_\Gamma}$. Since v_h^0 is piecewise linear on $\mathcal{T}_\Gamma$, we have that $|v_h^0|_{1,\Omega\cap\mathcal{T}_\Gamma} \simeq |v_h^0|_{1,\omega\cap\mathcal{T}_\Gamma} \simeq |u^0|_{1,\omega\cap\mathcal{T}_\Gamma} \simeq \sqrt{h}$. The bound for (3.29) thus obtained would exceed $\sqrt{h}/\sqrt{\epsilon}$, which is much bigger than $\sqrt{h}$.

This argument shows that the estimate of Theorem 3.4 on the error of the finite element domain embedding method for the Dirichlet problem is sharp. The optimal convergence rate $\mathcal{O}(\sqrt{h})$ can be achieved by taking the penalty parameter as $\epsilon = \mathcal{O}(\sqrt{h})$. Smaller ϵ does not help in improving the accuracy, but it does not hurt the accuracy either. From a computational point of view, the best value of ϵ should be $\epsilon = C\sqrt{h}$. Since this accuracy is rather poor, it seems desirable to find ways to enhance it.

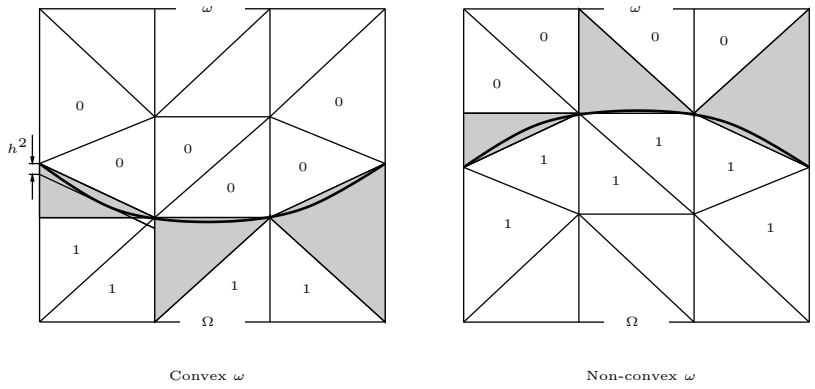


FIG. 5. *A portion of the adjusted mesh around* Γ.

3.4. Dirichlet boundary condition with adjusted mesh. One way to reduce the magnitude of (3.29) is to adjust the uniform mesh around Γ by moving the nearby nodes onto Γ and reconnecting some mesh lines to form a polygonal interpolation to Γ in the mesh; see Figure 1. The adjusted mesh must fulfill the following requirements: (1) The total number of nodes keeps unchanged. (2) No mesh-line segment has one end in ω and the other in Ω. (3) The adjusted triangulation is shape regular and quasi uniform. Fulfilling these requirements ensures, on one hand, that the discrete Laplacian on the adjusted mesh is spectrally equivalent to the one on the uniform triangulation thus retains the effectiveness of preconditioning the discrete system by fast Poisson solvers. On the other hand, the magnitude of (3.29) is significantly reduced, and thus the accuracy of the finite element domain embedding method is accordingly enhanced. The Börgers's algorithm [7] exactly carries out such adjustment. It is interesting to note that the algorithm was proposed in a context of solving a Neumann problem by domain embedding methods [8]. Our above analysis shows that the mesh adjustment is actually not necessary for Neumann problems, while it has a significant effect for the Dirichlet problem.

It turns out that the effect of this mesh adjustment is different whether ω is convex. For convex ω, we prove that the full accuracy of $\mathcal{O}(h)$ is achieved by the domain embedding finite element method on such adjusted mesh if one takes $\epsilon = \mathcal{O}(h)$. Smaller ϵ does not hurt the accuracy. If ω is not convex, then the accuracy can only be improved to $\mathcal{O}(h^{2/3})$ by taking $\epsilon = Ch^{2/3}$. But in this case, smaller ϵ could diminish the improvement and sets the accuracy back to that of the uniform mesh. This is a case in which the balance between ϵ and h is delicate.

We keep the notations of triangles and vertices introduced for uniform triangulation. For example, $\mathcal{T}_\Gamma$ still comprises those triangles that intersect Γ, which are shaded in Figure 5. In addition, we shall use $\mathcal{P}$ to denote the set of all triangles enclosed by the polygonal interpolation of Γ. In consistence to our earlier convention, $\mathcal{P}$ also stands for the enclosed polygonal domain.

THEOREM 3.5. *We assume that* $\omega \subset\subset R$, $\Gamma \in C^2$, *and* $f \in L^2(\omega)$. *The triangulation* $\mathcal{T}$ *of* R *is obtained by adjusting a uniform mesh around* Γ *in the way described above. The finite element domain embedding method* (3.26) *is a straightforward discretization of the penalty formulation* (3.21) *on the adjusted mesh. Then if* ω *is convex, we have that*

$$\|u^0 - u_h^\epsilon\|_{1,\omega} \le C(\epsilon + \epsilon\sqrt{h} + h\sqrt{\epsilon} + h)\|f\|_{0,\omega}. \tag{3.36}$$

If ω is not convex, we have that

$$\|u^0 - u_h^\epsilon\|_{1,\omega} \le C\left(\epsilon + \epsilon\sqrt{h} + h\sqrt{\epsilon} + h + \frac{h}{\sqrt{\epsilon}}\right)\|f\|_{0,\omega}. \tag{3.37}$$

Here, C is a constant independent of ϵ, h, and f.

Proof. The proof is based on the estimate (3.27) that was established in Theorem 3.3, which is still valid even when $\mathcal{T}$ is not uniform. There is nothing new in estimating the first two terms in the right-hand side of (3.27); i.e., we have $\epsilon\,|u^1|_{1,R} \le C\,\epsilon\,\|f\|_{0,R}$ and $\sqrt{\epsilon}\inf_{v_h^1\in H_h}\left(\sqrt{\epsilon}|u^1 - v_h^1|_{1,\omega} + |u^1 - v_h^1|_{1,\Omega}\right) \le C(\epsilon\sqrt{h} + h\sqrt{\epsilon})\|f\|_{0,R}$. We need only to analyze the third term

$$\inf_{v_h^0\in H_h}\left(|u^0 - v_h^0|_{1,\omega} + \frac{1}{\sqrt{\epsilon}}|v_h^0|_{1,\Omega}\right). \tag{3.38}$$

We let $v_h^0 \in H_h$ be such a piecewise linear function that interpolates u^0 for all the vertices $\nu \in \mathcal{V}$. When ω is convex, this interpolation is consistent with that defined in the proof of Theorem 3.4. It is, however, different if ω is not convex, since the latter was required to be zero on the shaded $\mathcal{T}_\Gamma$; see Figure 5.

When ω is convex, we have $\mathcal{P} \subset \omega$, and

$$|u^0 - v_h^0|_{1,\omega}^2 = |u^0|_{1,\omega\setminus\overline{\mathcal{P}}}^2 + |u^0 - v_h^0|_{1,\mathcal{P}}^2.$$

Since $\omega \setminus \overline{\mathcal{P}}$ is covered by a strip γ_δ, with $\delta = Ch^2$; see Figure 5 (left), by Lemma 2.2, we have that $|u^0|_{1,\omega\setminus\overline{\mathcal{P}}}^2 \le Ch^2\|\bar{u}^0\|_{2,R}^2$. Here $\bar{u}^0 \in H^2(R)$ is a H^2 extension of $u^0|_\omega$ such that $\|\bar{u}^0\|_{2,R} \le C\|u^0\|_{2,\omega}$. Since v_h^0 interpolates u^0 at the vertices of every triangle $\tau \in \mathcal{P}$, by the standard argument, we see that $|u^0 - v_h^0|_{1,\mathcal{P}}^2 \le Ch^2|u^0|_{2,\mathcal{P}}^2$. Therefore, $|u^0 - v_h^0|_{1,\omega} \le Ch\|f\|_{0,\omega}$. Since $v_h^0 = 0$ on Ω, we get that

$$|u^0 - v_h^0|_{1,\omega} + \frac{1}{\sqrt{\epsilon}}|v_h^0|_{1,\Omega} \le Ch\|f\|_{0,\omega}.$$

The estimate (3.36) then follows from this and Theorem 3.3.

When ω is not convex, we have that

$$|u^0 - v_h^0|_{1,\omega}^2 = |u^0|_{1,\omega\cap\mathcal{P}^c}^2 + |u^0 - v_h^0|_{1,\omega\cap\mathcal{P}}^2.$$

Here $\mathcal{P}^c = R\setminus\overline{\mathcal{P}}$. Since $\omega\cap\mathcal{P}^c$ is contained in a strip of width $\mathcal{O}(h^2)$, by Lemma 2.2, we have that $|u^0|_{1,\omega\cap\mathcal{P}^c}^2 \le Ch^2\|\bar{u}^0\|_{2,R}^2$. We see that $|u^0 - v_h^0|_{1,\omega\cap\mathcal{P}}^2 \le |\bar{u}^0 - v_h^0|_{1,\mathcal{P}}^2$. Since v_h^0 interpolates $\bar{u}^0$ on the vertices of every triangle in $\mathcal{P}$, by the standard argument of finite element interpolation, we have that $|\bar{u}^0 - v_h^0|_{1,\mathcal{P}}^2 \le Ch^2\|\bar{u}^0\|_{2,\mathcal{P}}^2$. Therefore,

$$|u^0 - v_h^0|_{1,\omega} \le Ch\|f\|_{0,\omega}. \tag{3.39}$$

Now, we consider $|v_h^0|_{1,\Omega}$. Since v_h^0 is zero on $\Omega\cap\mathcal{P}^c$, we have that $|v_h^0|_{1,\Omega}^2 = |v_h^0|_{1,\Omega\cap\mathcal{P}}^2$. Let $\mathcal{P}_\Omega = \{\tau \in \mathcal{P}; \tau\cap\Omega \ne \emptyset\}$, which is a cluster of triangles contained in a strip of width $\mathcal{O}(h)$ around Γ. On every $\tau \in \mathcal{P}_\Omega$, v_h^0 interpolates $\bar{u}^0$. According to Lemma 2.3, we have that $\sum_{\tau\in\mathcal{P}_\Omega}|v_h^0|_{1,\tau}^2 \le Ch\|\bar{u}^0\|_{2,R}^2$. For each $\tau \in \mathcal{P}_\Omega$, the portion $\tau\cap\Omega$ is either contained in a trapezoid of height $\mathcal{O}(h^2)$ by one side of τ, or a smaller triangle of diameter $\mathcal{O}(h^2)$ at a vertex of τ. In any case, we have that $|v_h^0|_{1,\tau\cap\Omega}^2 \le Ch|v_h^0|_{1,\tau}^2$. We thus have that

$$|v_h^0|_{1,\Omega\cap\mathcal{P}}^2 = \sum_{\tau\in\mathcal{P}_\Omega}|v_h^0|_{1,\tau\cap\Omega}^2 \le Ch\sum_{\tau\in\mathcal{P}_\Omega}|v_h^0|_{1,\tau}^2 \le Ch^2\|\bar{u}^0\|_{2,R}^2.$$

Therefore, $|v_h^0|_{1,\Omega} \le Ch\|\bar{u}^0\|_{2,R}$. This together with (3.39) shows that

$$\inf_{v_h^0 \in H_h} \left(|u^0 - v_h^0|_{1,\omega} + \frac{1}{\sqrt{\epsilon}} |v_h^0|_{1,\Omega} \right) \le C \left(h + \frac{h}{\sqrt{\epsilon}} \right) \|\bar{u}^0\|_{2,R}.$$

The proof of (3.37) is complete. □

The estimates of this theorem are sharp. The sharpness argument is similar to that which follows Theorem 3.4.

From (3.36) we see that when ω is convex, the finite element domain embedding method (3.26) with the adjusted mesh achieves the full accuracy $\mathcal{O}(h)$ in the H^1 norm by taking $\epsilon = \mathcal{O}(h)$. From a computational point of view, the best value then is $\epsilon = Ch$. But smaller ϵ won't hurt the accuracy.

From (3.37) we see that if ω is not convex, then the full accuracy of $\mathcal{O}(h)$ cannot be achieved by the mesh adjustment. In this case, there is an optimal value for ϵ. It is $\epsilon = Ch^{2/3}$, which leads to an overall accuracy of order $h^{2/3}$ in the H^1 norm. Unlike the Neumann problem with a uniform mesh or the Dirichlet problem on a convex domain with adjusted mesh, for this case, a smaller ϵ thwarts the accuracy of the finite element domain embedding. It could reduce the order back to $h^{1/2}$, that is, the order of accuracy achieved by the uniform mesh.

REFERENCES

[1] R.A. Adams, *Sobolev Spaces*, Academic Press, New York, 1975.

[2] P. Angot, *A unified fictitious domain model for general embedded boundary conditions*, C. R. Acad. Sci. Paris, Ser. I, 341 (2005), pp. 683–688.

[3] D.N. Arnold and R. Falk, *A uniformly accurate finite element method for the Reissner–Mindlin plate*, SIAM J. Numer. Anal., 26 (1989), pp. 1276–1290.

[4] I. Babuška, *The finite element method with penalty*, Math. Comp., 27 (1973), pp. 221–228.

[5] N.S. Bakhvalov and A.V. Knyazev, *Preconditioned iterative methods in a subspace for linear algebraic equations with large jumps in the coefficients*, in Domain Decomposition Methods in Science and Engineering, Contemp. Math. 180, D. Keyes and J. Xu, eds., AMS, Providence, RI, 1994.

[6] P. Bochev and R.B. Lehoucq, *On the finite element solution of the pure Neumann problem*, SIAM Rev., 47 (2005), pp. 50–66.

[7] C. Börgers, *A triangulation algorithm for fast elliptic solvers based on domain embedding*, SIAM J. Numer. Anal., 27 (1990), pp. 1187–1196.

[8] C. Börgers and O.B. Widlund, *On finite element domain imbedding methods*, SIAM J. Numer. Anal., 27 (1990), pp. 963–978.

[9] F. Brezzi, J.-L. Lions, and O. Pironneau, *The Chimera method for a model problem*, in Numerical Mathematics and Advanced Applications -ENUMATH 2001, F. Brezzi, A. Buffa, S. Corsaro, and A. Murli, eds., Springer, New York, 2003, pp. 817–826.

[10] P.G. Ciarlet, *Finite Element Method for Elliptic Equations*, North-Holland, Amsterdam, 1978.

[11] S. Del Pino and O. Pironneau, *A fictitious domain based general PDE solver*, in Numerical Methods for Scientific Computing: Variational Problems and Applications, Y. Kuznetsov, P. Neittanmaki, and O. Pironneau, eds., Internat. Center Numer. Methods Eng., Barcelona, 2003.

[12] K.O. Friedrichs and H.B. Keller, *A finite difference scheme for generalized Neumann problems*, in Numerical Solutions of Partial Differential Equations, J.H. Bramble, ed., Academic Press, New York and London, 1966.

[13] D. Gilbarg and N.S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer, Berlin, 1983.

[14] R. Glowinski and T.W. Pan, *Error estimate for fictitious domain/penalty/finite element methods*, Calcolo, 29 (1992), pp. 125–141.

[15] R. Glowinski, T.W. Pan, R. O. Wells, Jr., and X. Zhou, *Wavelet and finite element solutions for the Neumann problem using fictitious domains*, J. Comput. Phys., 126 (1996), pp. 40–51.

[16] P. Grisvard, *Singularities in Boundary Value Problems*, Masson, Springer, New York, 1992.
[17] A. Knyazev and O. Widlund, *Lavrentiev regularization + Ritz approximation = uniform finite element error estimates for differential equations with rough coefficients*, Math. Comp., 72 (2003), pp. 17–40.
[18] Yu. A. Kuznetsov, *Domain decomposition and fictitious domain methods with distributed Lagrange multipliers*, in Domain Decomposition Methods in Science and Engineering, N. Debit, M. Garbey, R. Hoppe, D. Keyes, Y. Kuznetsov, and J. Periaux, eds., Internat. Center Numer. Methods Eng., Barcelona, 2002.
[19] R. Lazarov, S. Lu, and S.V. Pereverzev, *On the balancing principle for some problems of numerical analysis*, Numer. Math., 106 (2007), pp. 659–689.
[20] S. Pereverzev and E. Schock, *On the adaptive selection of the parameter in regularization of ill-posed problems*, SIAM J. Numer. Anal., 43 (2005), pp. 2060–2076.
[21] E.M. Stein, *Singular Integrals and Differentiability Properties of Functions*, Princeton University Press, Princeton, NJ, 1970.
[22] S. Zhang, *Equivalence estimates for a class of singular perturbation problems*, C. R. Acad. Sci. Paris Ser. I, 342 (2006), pp. 285–288.
[23] S. Zhang, *A domain embedding method for mixed boundary value problems*, C. R. Acad. Sci. Paris Ser. I, 343 (2006), pp. 287–290.
[24] S. Zhang, *Sharp Convergence Rate of Domain Embedding Methods for Various Boundary Conditions*, http://www.math.wayne.edu/~sheng/embedding.pdf (July 14, 2008).

SIAM J. NUMER. ANAL.
Vol. 46, No. 6, pp. 2867–2891
© 2008 Society for Industrial and Applied Mathematics

REDUCED-ORDER OPTIMAL CONTROL BASED ON APPROXIMATE INERTIAL MANIFOLDS FOR NONLINEAR DYNAMICAL SYSTEMS*

KAZUFUMI ITO† AND KARL KUNISCH‡

Abstract. A reduced-order method for optimal control problems in infinite dimensions based on approximate inertial manifolds is developed. Convergence of the cost, optimal controls, and optimal states of the finite dimensional, reduced-order, optimal control problems to the original optimal control problem is analyzed. Special attention is given to the particular case when the dynamics are described by the Navier–Stokes equations in dimension two.

Key words. reduced-order methods, approximate inertial manifold, nonlinear Galerkin, decomposition of state space, infinite dimensional system

AMS subject classifications. 93B11, 93B52, 49N35

DOI. 10.1137/060666421

1. Introduction.

In this research we consider optimal control problems governed by partial differential equations. For such problems there has recently been an increased interest in developing reduced-order control methods. This is motivated by computational needs and by systems science considerations as well. For nonlinear distributed parameter control systems, let us specifically mention model reduction based on proper orthogonal decomposition approach [1, 15, 16, 24] and the reduced-basis method [11, 12, 13, 17]. The key issue for the reduction consists in selecting basis elements which are rich in information in the sense that they capture well the essential dynamical properties of the original control system. After the basis elements are selected by these methods the standard Galerkin approach is applied to obtain a reduced-order control system. For linear control systems many alternative reduction methods were proposed and analyzed, including the Hankel-norm approximation [7] and the LQG-balanced truncation realization [2]. The LQG-balanced truncation method was introduced in the finite dimensional literature by [14] and other interpretations followed in [19, 20]. For the infinite dimensional theory we refer to [3]. For linear systems, model reduction based on specific versions of proper orthogonal decomposition and of balanced truncation coincide [22]. The selection of references here is by no means complete and we refer to the citations in the quoted papers for references.

In this paper we discuss an order-reduction method based on the concept of inertial manifolds. Let us recall that an important feature of inertial manifolds consists in the description of the small-scale dynamics as a graph of the large-scale dynamics. The inertial manifold is a finite dimensional invariant manifold that attracts all orbits exponentially. Thus it is natural to expect that controlling the inertial manifold dynamics results in the ability to control the full underlying control system. However,

*Received by the editors July 31, 2006; accepted for publication (in revised form) March 7, 2008; published electronically August 6, 2008.
http://www.siam.org/journals/sinum/46-6/66642.html
†Department of Mathematics, North Carolina State University, Raleigh, NC 27695-8205 (kito@math.ncsu.edu). This author's research was partially supported by the Army Research Office under DAAD19-02-1-0394.
‡Institut für Mathematik und wissenschaftliches Rechnen, Karl-Franzens-Universität Graz, A-8010 Graz, Austria, and Radon Institute, A-4040 Linz, Austria (karl.kunisch@uni-graz.at).

there are some technical difficulties associated with this approach: (1) the existence of the inertial manifold can only be proved for a limited class of systems, (2) the construction of the inertial manifold is highly involved and technical, and (3) the analysis of the behavior of the closed-loop system has not been fully addressed.

Here we therefore consider approximate inertial manifolds which are not necessarily invariant under the dynamics but which approximate all orbits starting from a bounded set with any desired accuracy. We use nonlinear Galerkin approximations as proposed, e.g., in [5, 18], and the manifold is represented by a stationary graph determined by the residual dynamics. If one would simply truncate the residual dynamics (the flat manifold), then this approach coincides with the standard Galerkin approximation. In [10] we demonstrated the effectiveness of the approximate inertial manifold approach for the linear quadratic regulator problem.

Error estimates for the uncontrolled solution based on the approximate inertial manifold method were investigated extensively. We refer, e.g., to [5, 18], where the estimates are obtained for $t \geq t^* > 0$ for some sufficiently large t^*. In this paper we establish error estimates for control systems based on the reduced-order systems governed by approximate inertial manifolds. Such estimates are obtained by the necessary optimality condition for the optimal control problem, a second order sufficient optimality condition, and the gap estimate for the approximate inertial manifold.

The outline of this paper is as follows. In section 2 we describe the general inertial manifold approach. An upper bound for the performance of the reduced-order control as well as an error estimate for the reduced-order control to the infinite dimensional control in terms of the gap estimate of the approximate inertial manifold are established. In section 3 we discuss error estimates for the two-dimensional Navier–Stokes equations in $L^2(0,T;X)$, where X denotes the state space. Section 4 is devoted to the necessary optimality condition, a second order sufficient optimality condition as well as a first concrete estimate for the finite dimensional optimal controls to the infinite dimensional one. In section 5 a sharper estimate of the convergence rate of the reduced-order control is established on the basis of a sensitivity analysis for the necessary optimality condition.

2. Approximate inertial manifold and reduced-order control system. In this section we describe a general strategy for the approximation of nonlinear control systems by nonlinear Galerkin schemes. Throughout this paper, X and U denote separable Hilbert spaces. We consider the optimal control problem

$$\min_{u \in L^2(0,T;U)} \quad J(x,u) = \int_0^T \ell(x(t)) + h(u(t))\, dt, \tag{2.1}$$

subject to

$$\frac{d}{dt}x(t) = A_0 x(t) + F(x(t)) + Bu(t), \quad x(0) = x_0 \in X, \tag{2.2}$$

where $T > 0$, $F : X \to X^*$ is a nonlinear mapping, A_0 is a linear self-adjoint negative definite operator in X with dense domain denoted by $dom(A_0)$, and B is a bounded linear operator from U to X. We further set $V = dom(-A_0)^{\frac{1}{2}}$ endowed with $|v|_V = \sqrt{\langle -A_0 v\,, v\rangle}$ as norm. We assume that (2.2) is a well-posed control system, i.e., given any $x_0 \in X$ and $u \in L^2(0,T;U)$, there exists a unique weak solution $x = x(t;x_0,u) \in C(0,T;X) \cap L^2(0,T;V)$ to (2.2) which depends continuously on $(x_0,u) \in X \times L^2(0,T;U)$. For example, we can formulate the control problem (2.1)–(2.2) in a

Gelfand triple formulation or as semilinear control systems; see, e.g., [9, 21]. Further, $\ell : X \to \mathbb{R}$ and $h : U \to \mathbb{R}$ are C^1 mappings, ℓ is supposed to be uniformly Lipschitz continuous on bounded sets of X, and h is such that the associated substitution operator from $L^2(0,T;U)$ to $\mathbb{R}$ is well-defined.

We assume that $-A_0$ has eigenpairs (λ_i, ϕ_i) in ascending order and that $\{\phi_i\}_{i=1}^{\infty}$ forms an orthonormal basis of X. Let P_1 be the orthogonal projection of X onto

$$X_1 = \text{ span } \{\phi_i : 1 \le i \le N\}.$$

Further we set $X_2 = (I - P_1)X$. Expressing x as

$$x = p + q, \quad \text{with } p(t) = P_1 x(t) \text{ and } q(t) = P_2 x(t),$$

where $P_2 = I - P_1$, we have from (2.2)

$$\begin{aligned} \frac{d}{dt} p(t) &= A_0 p(t) + P_1 F(p(t) + q(t)) + P_1 B u(t), \\ \frac{d}{dt} q(t) &= A_0 q(t) + P_2 F(p(t) + q(t)) + P_2 B u(t). \end{aligned} \tag{2.3}$$

In the linear Galerkin approach, the higher order modes q are neglected. This results in the control system

$$\frac{d}{dt} \hat{x}_1(t) = A_0 \hat{x}_1(t) + P_1 F(\hat{x}_1(t)) + P_1 B u(t).$$

In the nonlinear Galerkin approach, $\frac{d}{dt}\, q(t)$ is assumed to be negligible compared to $A_0\, q(t)$ in (2.3). This suggests considering the control system

$$\begin{aligned} \frac{d}{dt} \hat{x}_1(t) &= A_0 \hat{x}_1(t) + P_1 F(\hat{x}_1(t) + \hat{x}_2(t)) + P_1 B u(t), \\ 0 &= A_0 \hat{x}_2(t) + P_2 F(\hat{x}_1(t) + \hat{x}_2(t)) + P_2 B u(t). \end{aligned} \tag{2.4}$$

We set

$$\hat{x} = \hat{x}_1 + \hat{x}_2.$$

and the first equation in (2.4) can be expressed as

$$\frac{d}{dt} P_1 \hat{x}(t) = A_0 \hat{x}(t) + P_1 F(\hat{x}(t)) + P_1 B u(t).$$

The second equation in (2.4) is nonlinear and is replaced in the approach that follows by the linear equation

$$A_0 \hat{x}_2(t) + P_2 F(\hat{x}_1) = 0,$$

where the coupling in the nonlinearity and $P_2 B u(t)$ are neglected. An alternative to this may be to consider

$$A_0 \hat{x}_2 + P_2(F(\hat{x_1}) + F'(\hat{x_1})\hat{x}_2) = 0,$$

which should lead to better approximation properties. The reduced-order system that we consider in this paper is given by

$$\frac{d}{dt}\hat{x}_1(t) = A_0\hat{x}_1(t) + P_1F(\hat{x}_1(t)+\hat{x}_2(t)) + P_1Bu(t), \tag{2.5}$$

$$0 = A_0\hat{x}_2(t) + P_2F(\hat{x}_1(t)).$$

Given $\hat{x}_1 \in X_1 = P_1X$, we denote the unique solution to the second equation in (2.5) by

$$\hat{x}_2 = \Phi(\hat{x}_1), \quad \text{where} \quad \Phi(\phi) = -A_0^{-1}P_2F(\phi). \tag{2.6}$$

In this way we obtain the finite dimensional reduced-order control system in X_1:

$$\min_{u\in L^2(0,T;U)} \quad \int_0^T \ell\big(z(t)+\Phi(z(t))\big) + h(u(t))\,dt \tag{2.7}$$

subject to

$$\frac{d}{dt}z(t) = A_0z(t) + P_1F(z(t)+\Phi(z(t))) + P_1Bu(t), \quad z(0) = P_1x_0; \tag{2.8}$$

i.e., we set $z = \hat{x}_1$ with $\hat{x}_1$ as in (2.5). Here we assume the existence of an optimal control u to (2.7) as well as u^* to (2.1). To obtain error estimates for $u-u^*$, we first establish an error estimate for the reduced-order equation. This involves an estimate of the gap

$$\Delta(t) = |q(t) - \Phi(p(t))|,$$

where $p(t)$ and $q(t)$ are defined in (2.3); see, e.g., [5] and section 3. Then, we can argue that under appropriate conditions

$$|z(t)-p(t)|^2 \le M\int_0^t |q(s)-\Phi(p(s))|^2\,ds \text{ for } t\in[0,T] \tag{2.9}$$

for some $M>0$. In fact, for $S=\{v\in V : |v|\le K\}$, with $K\ge 0$, let us assume the properties

$$\langle F(x)-F(y),\phi\rangle \le c\,|x-y||\phi|_V \quad \text{for all} \quad x,y,\phi\in S\subset V, \tag{2.10}$$

and

$$x(t)\in S \text{ and } z(t)+\Phi(z(t))\in S \text{ for a.e. } t\in(0,T).$$

From (2.3) and (2.8) we have

(2.11)

$$\tfrac{d}{dt}(z(t)-p(t)) = A_0(z(t)-p(t)) + P_1\,F(z(t)+\Phi(z(t))) - P_1\,F(p(t)+q(t)),$$

$$z(0)-p(0) = 0.$$

By (2.10) we find for $x(t)\in S$ and $z(t)+\Phi(z(t))\in S$ that

$$\begin{aligned}&\langle P_1\,F\big(z(t)+\Phi(z(t))\big) - P_1\,F\big(p(t)+q(t)\big), z(t)-p(t)\rangle\\ &\le c\big(|z(t)-p(t)| + |\Phi(z(t))-\Phi(p(t))| + |q(t)-\Phi(p(t))|\,\big)|z(t)-p(t)|_V\\ &\le c\big((1+\|\Phi\|)|z(t)-p(t)| + |q(t)-\Phi(p(t))|\big)|z(t)-p(t)|_V,\end{aligned} \tag{2.12}$$

where $\|\Phi\|$ is the Lipschitz constant of Φ on S. Taking the inner product of (2.11) with $z(t) - p(t)$ and using (2.12) results in

$$\frac{1}{2}\frac{d}{dt}|z(t) - p(t)|^2 \le -|z(t) - p(t)|_V^2 + c\big((1 + \|\Phi\|)\,|z(t) - p(t)| + |q(t) - \Phi(p(t))|\big)\,|z(t) - p(t)|_V,$$

which implies, using $ab \le a^2 + \frac{1}{4}b^2$, that

$$\frac{1}{2}\frac{d}{dt}|z(t) - p(t)|^2 \le \frac{c^2}{2}\big((1 + \|\Phi\|)^2|z(t) - p(t)|^2 + |q(t) - \Phi(p(t))|^2\big).$$

Estimate (2.9) with $M = c^2 exp\big(c^2(1 + \|\Phi\|)^2 T\big)$ follows from Gronwall's inequality since $z(0) = p(0)$.

Once (2.9) is established, it implies the error estimate

$$\begin{aligned} |\hat{x}(t) - x(t)| &= |z(t) + \Phi(z(t)) - x(t)| \\ &\le |z(t) - p(t)| + |\Phi(z(t)) - \Phi(p(t))| + |q(t) - \Phi(p(t))| \\ &\le (1 + \|\Phi\|)\,|z(t) - p(t)| + |q(t) - \Phi(p(t))|, \end{aligned} \tag{2.13}$$

for $t \in [0, T]$.

To demonstrate the use of the a priori estimate (2.13) in the context of the optimal control problem, let us assume the existence of optimal controls u and u^* to (2.7)–(2.8) and (2.1)–(2.2), respectively. Moreover, let x and x^* in $C(0, T; X) \cap L^2(0, T; V)$ be the solutions to (2.2) with u and u^*, respectively. Thus x^* is the optimal trajectory to (2.1)–(2.2), whereas x is the trajectory of the infinite dynamical system controlled by the optimal control computed on the bases of the reduced-order system. Let

$$\hat{x} = z + \Phi(z) \ \text{ and } \hat{x}^* = z^* + \Phi(z^*),$$

where z and z^* are the solutions in $C(0, T; X) \cap L^2(0, T; V)$ to (2.8), corresponding to u and u^*, respectively. Then, for the finite horizon control on $[0, T]$ we have

$$\begin{aligned} 0 &\le J(x, u) - J(x^*, u^*) \\ &= J(x, u) - J(\hat{x}, u) + J(\hat{x}, u) - J(\hat{x}^*, u^*) + J(\hat{x}^*, u^*) - J(x^*, u^*) \\ &\le J(x, u) - J(\hat{x}, u) + J(\hat{x}^*, u^*) - J(x^*, u^*), \end{aligned}$$

where we used that $J(\hat{x}, u) - J(\hat{x}^*, u^*) \le 0$. Thus

$$0 \le J(x, u) - J(x^*, u^*) \le C\,(|x^* - \hat{x}^*|_{L^2(0,T;X)} + |x - \hat{x}|_{L^2(0,T;X)}), \tag{2.14}$$

with C the Lipschitz constant of $\ell : X \to R$ on the ball determined by the $C(0, T; X)$ norm of $x, \hat{x}, x^*$, and $\hat{x}^*$. The error estimate (2.9) now provides an upper bound for the performance of the suboptimal control u based on the reduced-order control problem (2.7)–(2.8) imposed onto the original control problem (2.1)–(2.2). In fact, if

$$x(t),\ \hat{x}(t),\ x^*(t),\ \hat{x}^*(t) \ \text{ are in } S \text{ for a.e. } t \in (0, T)$$

with $\hat{x} = z + \Phi(z)$ and $\hat{x}^* = z^* + \Phi(z^*)$, then by (2.9) and (2.13)–(2.14)

$$0 \le J(x, u) - J(x^*, u^*) \le \hat{C}(|q - \Phi(p)|_{L^2(0,T;X)} + |q^* - \Phi(p^*)|_{L^2(0,T;X)}), \tag{2.15}$$

where $\hat{C} = \hat{C}(\|\Phi\|, c, C)$ and (p, q), (p^*, q^*) are the orthogonal projections of x and x^* according to (2.3).

Suppose, moreover, that for some $\sigma > 0$ the second order sufficient optimality condition

$$J(x(v), v) - J(x^*, u^*) \geq \sigma \, |v - u^*|^2_{L^2(0,T;U)} \tag{2.16}$$

holds for all $v \in L(0, T; U)$ in a neighborhood of u^*, where $x(v)$ denotes the solution to (2.2) with u replaced by v. Then, if u is contained in this neighborhood for second order sufficient optimality, we have the estimate

$$\begin{aligned} |u - u^*|^2_{L^2(0,T;U)} &\leq \tfrac{C}{\sigma} \left(|x - \hat{x}|_{L^2(0,T;X)} + |x^* - \hat{x}^*|_{L^2(0,T;X)} \right) \\ &\leq \tfrac{\hat{C}}{\sigma} (|q - \Phi(p)|_{L^2(0,T;X)} + |q^* - \Phi(p^*)|_{L^2(0,T;X)}), \end{aligned} \tag{2.17}$$

and thus the optimal controls for the finite and the original optimal control systems are estimated in terms of the gap $q - \Phi(p)$ and $q^* - \Phi(p^*)$.

3. Basic estimates for two-dimensional Navier–Stokes systems. In this section we analyze the case where the dynamical system (2.2) is given by the Navier–Stokes equations in dimension two. We establish estimates for q and the gap $\Delta = q - \Phi(p)$ following [5]. Our estimate here, however, differs from those in [5] in that we consider the transient situation on the time horizon $[0, T]$, as opposed to the case where t is sufficiently large as in [5]. We then turn to analyzing properties of the mapping Φ and derive error estimates for the reduced-order system.

Let us consider the two-dimensional incompressible Navier–Stokes equations:

$$\begin{aligned} &v_t + (v \cdot \nabla)v + \nabla \rho = \nu \, \Delta v + Bu && \text{for } (t, x) \in (0, T] \times \Omega, \\ &\nabla \cdot v = 0 && \text{in } (0, T] \times \Omega, \\ &v(x, 0) = v_0(x) \text{ in } \Omega, \quad v(t, x) = 0 && \text{for } (t, x) \in (0, T] \times \Gamma, \end{aligned} \tag{3.1}$$

where Ω is a bounded open domain in $\mathbb{R}^2$ with sufficiently smooth boundary Γ, $v = v(t, x) \in \mathbb{R}^2$ is the velocity field, $\rho = \rho(t, x) \in \mathbb{R}$ is the pressure, and $\nu > 0$ is the normalized viscosity. For convenience, in order to exclude special cases in the estimate we also assume that $\nu \leq 1$. Let

$$V = \{\phi \in H^1_0(\Omega)^2 : \nabla \cdot \phi = 0\},$$

and let

$$X = \{\phi \in L^2(\Omega) : \nabla \cdot \phi = 0, \;\; n \cdot \phi = 0 \text{ on } \partial\Omega\}$$

denote the closure of $\{\phi \in \mathcal{D}(\Omega) : \nabla \cdot \phi = 0\}$ in $H^1_0(\Omega)^2$ and $L^2(\Omega)^2$, respectively. Let P stand for the orthogonal projection of $L^2(\Omega)^2$ onto the closed subspace X, and let $A = P\Delta$ in X denote the Stokes operator defined by

$$(A\phi, \psi) = -(\nabla\phi, \nabla\psi)_{L^2} \text{ for } \phi, \; \psi \in V.$$

It is a closed self-adjoint operator with $dom(A) = H^2(\Omega)^2 \cap V$ and $dom((-A)^{\frac{1}{2}}) = V$. The operator $(A)^{-1}$ is compact from X to $dom(A)$ and hence the eigenvalue-eigenvector pairs of $-A$ define a complete orthonormal basis of X.

Let $B \in \mathcal{L}(U, X)$, and define $A_0 = \nu A$, and $F : X \to X^*$

$$-F(\phi) = b(\phi, \phi) = P(\phi \cdot \nabla \phi).$$

Then (3.1) can equivalently be expressed in abstract form as (2.2). For results on the Navier–Stokes equations that we utilize here, we refer to [23]. Note that we chose to endow V with $\sqrt{\langle -A_0 v\,, v\rangle}$ as norm in section 2, whereas $\sqrt{\langle -Av\,, v\rangle}$ is the norm for V in this and the following sections in order to specify the dependency of our estimates on $\nu > 0$ as in [5]. We use the following properties [5, 23] of the bilinear form $b(\phi, \psi) = P(\phi \cdot \nabla \psi)$: There exist constants c_i such that for all ϕ, ψ, w in $dom\, A$

$$
\begin{aligned}
&|b(\phi,\psi)|_{L^2} \le c_1 \begin{cases} |\phi|^{\frac{1}{2}} |\phi|_V^{\frac{1}{2}} |\psi|_V^{\frac{1}{2}} |A\psi|^{\frac{1}{2}}, \\ |\phi|^{\frac{1}{2}} |A\phi|^{\frac{1}{2}} |\psi|_V, \end{cases} \\
&|(b(\phi,\psi), w)| \le c_2\, |\phi|^{\frac{1}{2}} |\phi|_V^{\frac{1}{2}} |\psi|_V |w|^{\frac{1}{2}} |w|_V^{\frac{1}{2}}, \\
&|\phi|_{L^\infty(\Omega)} \le c_3\, |\phi|_V \left(1 + \log\left(\frac{|A\phi|^2}{\lambda_1 |\phi|_V^2}\right)\right)^{\frac{1}{2}}, \\
&|b(\phi,\psi)|_{L^2} \le c_4 \begin{cases} |\phi|_V |\psi|_V \left(1 + \log\left(\frac{|A\phi|^2}{\lambda_1 |\phi|_V^2}\right)\right)^{\frac{1}{2}}, \\ |\phi|\, |A\psi| \left(1 + \log\left(\frac{|A^{\frac{3}{2}}\psi|^2}{\lambda_1 |A\psi|^2}\right)\right)^{\frac{1}{2}}, \end{cases}
\end{aligned}
\tag{3.2}
$$

and satisfies

$$(b(\phi, \psi), \psi) = 0 \quad \text{for } \phi, \psi \in V. \tag{3.3}$$

It follows from [23, proof of Theorem III.3.10], that for $x_0 \in V$ and $u \in L^2(0, T; U)$ there exists a unique solution to (2.2),

$$x(t) \in C(0, T; V) \cap L^2(0, T; dom\,(A)),$$

and

$$|x|_{C(0,T;V)\cap L^2(0,T;dom\,(A))} \le C(|x_0|_V, |u|_{L^2(0,T;U)}),$$

for a constant C depending continuously on its arguments. We let

$$M_0 = \sup_{t\in[0,T]} |x(t)| \text{ and } M_1 = \sup_{t\in[0,T]} |x(t)|_V.$$

Throughout the remainder of this paper we assume that the cut off for the subspace X_1 is such that

$$\lambda_{N+1} \ge \left(\frac{4c_2 M_1}{\nu}\right)^2. \tag{3.4}$$

Note that

$$|A\phi|^2 \le \lambda_N\, |\phi|_V^2 \text{ for } \phi \in X_1, \text{ and } |\phi|_V^2 \ge \lambda_{N+1} |\phi|^2 \text{ for } \phi \in X_2. \tag{3.5}$$

Taking the inner product of (2.2) with $q(t) \in X_2$ we obtain

$$\frac{1}{2}\frac{d}{dt}|q|^2 + \nu\, |q|_V^2 = -(b(p+q, p+q), q) + (P_2 B u, q) \text{ for a.e. } t \in (0, T).$$

From (3.3)

$$(b(p+q,p+q),q) = (b(p,p),q) + (b(q,p),q).$$

Together with (3.2)–(3.5) it follows that

$$|(b(p+q,p+q),q)| \le |(b(p,p),q)| + |(b(q,p),q)| \le c_4 L^{\frac{1}{2}} |p|_V^2 |q| + c_2 |q||q|_V |p|_V,$$

at a.e. $t \in (0,T)$, where

$$L = 1 + \log \frac{\lambda_{N+1}}{\lambda_1}. \tag{3.6}$$

Referring to (3.5) once again we obtain

$$\begin{aligned}\frac{d}{dt}|q|^2 + (2\nu - 2c_2\lambda_{N+1}^{-\frac{1}{2}} M_1|)\,|q|_V^2 &\le 2|P_2Bu||q| + 2c_4 M_1 L^{\frac{1}{2}}\,|p|_V|q| \\ &\le \tfrac{4}{\nu\lambda_{N+1}}(c_4^2 L M_1^2 |p|_V^2 + |P_2Bu|^2) + \tfrac{\nu}{2}|q|_V^2.\end{aligned}$$

By (3.4) the above inequality implies that

$$\frac{d}{dt}|q|^2 + \frac{3\nu}{2}\,|q|_V^2 \le \frac{\nu}{2}|q|_V^2 + \frac{4}{\nu\lambda_{N+1}}(|P_2Bu|^2 + c_4^2 M_1^2 L|p|_V^2)$$

and

$$\frac{d}{dt}|q|^2 + \nu\lambda_{N+1}\,|q|^2 \le \frac{4}{\nu\lambda_{N+1}}(|P_2Bu|^2 + c_4^2 M_1^2 L|x|_V^2).$$

Multiplying this inequality with $e^{\nu\lambda_{N+1}t}$ and integrating over $(0,t)$ implies that

$$|q(t)|^2 \le e^{-\nu\lambda_{N+1}t}|q(0)|^2 + \frac{4}{\nu\lambda_{N+1}}\int_0^t e^{\nu\lambda_{N+1}(s-t)}(|P_2Bu(s)|^2 + c_4^2 M_1^4 L)\,ds, \tag{3.7}$$

for every $t \in [0,T]$. Consequently,

$$\int_0^T |q|^2\,dt \le \frac{1}{\nu\lambda_{N+1}}|q(0)|^2 + \frac{4}{\nu^2\lambda_{N+1}^2}\int_0^T (|P_2Bu|^2 + c_4^2 M_1^4 L)\,dt$$

and this implies the existence of a constant $k_1 = k_1(M_1, c_4, T)$ such that

$$\int_0^T |q|^2\,dt \le \frac{1}{\nu^2\lambda_{N+1}^2}(|q(0)|_V^2 + 4|P_2Bu|^2_{L^2(0,T;X)} + k_1\,L). \tag{3.8}$$

Taking the inner product of (2.2) with Aq we find that

$$\frac{1}{2}\frac{d}{dt}|q|_V^2 + \nu\,|Aq|^2 = -(b(p+q,p+q),Aq) + (P_2Bu,Aq).$$

From (3.2)–(3.5)

$$\begin{aligned}&|(b(p+q,p+q),Aq)| + (P_2Bu,Aq) \\ &\le c_4 L^{\frac{1}{2}}|p|_V|Aq|(|p|_V + |q|_V) + c_1\,|q|^{\frac{1}{2}}|Aq|^{\frac{3}{2}}(|p|_V + |q|_V) + (P_2Bu,Aq).\end{aligned}$$

Expressing $|Aq|^{\frac{3}{2}}(|p|_V + |q|_V)$ as $\nu^{\frac{3}{4}}|Aq|^{\frac{3}{2}}\frac{1}{\nu^{\frac{3}{4}}}(|p|_V + |q|_V)$ and analogously for the remaining terms which contain $A(q)$, we have

$$\begin{aligned}&|(b(p+q,p+q),Aq)| + (P_2Bu,Aq)\\ &\quad\le \frac{\nu}{2}|Aq|^2 + \frac{1}{\nu}|P_2Bu| + \frac{\tilde{c}_3M_1^2L\,|x|_V^2}{\nu} + \frac{\tilde{c}_1M_0^2|x|_V^4}{\nu^3},\end{aligned}$$

where $\tilde{c}_1, \tilde{c}_3$ are independent of ν, L, x, and u. Hence

$$\frac{d}{dt}|q|_V^2 + \nu\,|A\,q|^2 \le \frac{2}{\nu}|P_2Bu|^2 + \frac{2\tilde{c}_3M_1^2L\,|x|_V^2}{\nu} + \frac{2\tilde{c}_1M_0^2|x|_V^4}{\nu^3},$$

and

$$|q(t)|_V^2 \le e^{-\nu\lambda_{N+1}t}|q_0|_V^2 + \int_0^t e^{\nu\lambda_{N+1}(s-t)}\left(\frac{2}{\nu}|P_2Bu(s)| + \frac{2\tilde{c}_3M_1^4L}{\nu} + \frac{2\tilde{c}_1M_0^2M_1^4}{\nu^3}\right).$$

Thus there exists a constant $k_2 = k_2(M_1)$ such that

$$\int_0^T |q(t)|_V^2\,dt \le \frac{1}{\nu\lambda_{N+1}}\left(|q(0)|_V^2 + \frac{2}{\nu}|P_2Bu|_{L^2}^2 + \frac{k_2L}{\nu} + \frac{k_2}{\nu^3}\right). \tag{3.9}$$

For this estimate the assumption $\lambda_{N+1} \ge (\frac{4c_2M_1}{\nu})^2$ is not needed. We now summarize these results.

THEOREM 3.1. *Assume that $x_0 \in V$, $u \in L^2(0,T;U)$, and that $|x|_{C(0,T;V)} \le M_1$ and $\lambda_{N+1} \ge (\frac{4c_2M_1}{\nu})^2$. Then there exist constants $k_1 = k_1(M_1)$ and $k_2 = k_2(M_1)$ such that*

$$\int_0^T |q(t)|^2\,dt \le \delta^2(|q(0)|_V^2 + 4|P_2Bu|_{L^2(0,T;X)}^2 + k_1L)$$

and

$$\int_0^T |q(t)|_V^2\,dt \le \delta\left(|q(0)|_V^2 + \frac{2}{\nu}|P_2Bu|_{L^2(0,T;X)}^2 + \frac{k_2L}{\nu} + \frac{k_2}{\nu^3}\right),$$

where $\delta = \frac{1}{\nu\lambda_{N+1}}$ and $L = 1 + \log\frac{\lambda_{N+1}}{\lambda 1}$.

If, moreover, $u \in L^\infty(0,T;U)$ and $x_0 \in dom(A_0)$, then for constants $k_3 = k_3(M_1)$ and $k_4 = k_4(M_1)$

$$|q(t)|^2 \le \delta^2(|A_0q(0)|^2 + 4|P_2Bu|_{L^\infty(0,T;X)}^2 + k_3L),$$

and

$$|q(t)|_V^2 \le \delta\left(|A_0q(0)|^2 + \frac{2}{\nu}|P_2Bu|_{L^\infty(0,T;X)}^2 + \frac{k_4L}{\nu} + \frac{k_4}{\nu^3}\right),$$

for all $t \in [0,T]$.

The second part of the theorem follows from (3.7) and the equation above (3.9). For the optimal solution u to (2.1)–(2.2), the assumption that $u \in L^\infty(0,T;U)$ is not restrictive as can be seen from the associated optimality system which will be given below.

3.1. Gap estimate. In this subsection we prove an estimate for the gap

$$\Delta(t) = q(t) - \Phi(p(t)) \text{ for } t \in [0,T].$$

Throughout we assume that $x(t) \in C(0,T;V) \cap L^2(0,T;dom\,(A_0))$, and we again set $M_1 = \sup_{t\in[0,T]} |x(t)|_V$. By (2.3) and (2.6) we have

$$-A_0\Delta + b(p,q) + b(q,p) + b(q,q) + \frac{d}{dt}\, q - P_2Bu = 0 \text{ for a.e. } t \in (0,T). \tag{3.10}$$

Using (3.2), (3.3), and (3.6) we find that

$$\begin{aligned} &|b(p,q) + b(q,p) + b(q,q)| \\ &\le (c_4 L^{\frac12} + c_1)|p|_V|q|_V + \tfrac{c_1}{\sqrt{\lambda_{N+1}}}|q|_V|Aq| \quad \text{for a.e. } t \in (0,T). \end{aligned} \tag{3.11}$$

Taking the inner product of (2.2) with $\frac{d}{dt}q$,

$$\frac{\nu}{2}\frac{d}{dt}|q|_V^2 + |\frac{d}{dt}q|^2 = -\left(b(p+q,p+q)\,,\frac{d}{dt}q\right) + \left(P_2Bu, \frac{d}{dt}q\right),$$

and therefore

$$\begin{aligned} &\tfrac{\nu}{2}\tfrac{d}{dt}|q|_V^2 + |\tfrac{d}{dt}q|^2 \\ &\le |b(p,q) + b(q,p) + b(q,q)|^2 + |P_2(b(p,p) + Bu)|^2 + \tfrac12|\tfrac{d}{dt}q|^2. \end{aligned}$$

This implies that

$$\begin{aligned} &\nu|q(T)|_V^2 + \textstyle\int_0^T |\frac{d}{dt}q|^2\,dt \le \nu|q(0)|_V^2 \\ &+2\textstyle\int_0^T |b(p,q) + b(q,p) + b(q,q)|^2\,dt + 2\int_0^T |P_2(b(p,p) + Bu)|^2\,dt \\ &\le \nu|q(0)|_V^2 + 4\textstyle\int_0^T \left((c_4L^{\frac12} + c_1)^2M_1^2|q|_V^2 + \frac{c_1^2}{\lambda_{N+1}}M_1^2|Aq|^2\right)dt \\ &+2\textstyle\int_0^T |P_2(b(p,p) + Bu)|^2\,dt. \end{aligned} \tag{3.12}$$

From (3.10)–(3.12) there exists a constant $\hat{k}$ such that

$$\begin{aligned} \textstyle\int_0^T |A_0\Delta|^2\,dt \le \hat{k}\big[\textstyle\int_0^T [(c_4L^{\frac12} + c_1)^2M_1^2|q|_V^2 + \frac{c_1^2}{\lambda_{N+1}}M_1^2|A\,q|^2]\,dt \\ +\nu|q(0)|_V^2 + \textstyle\int_0^T |P_2(b(p,p))|^2\,dt + \int_0^T |P_2Bu|^2\,dt\big]. \end{aligned}$$

Hence using (3.5) and $L \ge 1$ there exists a constant k_5 independent of ν such that

$$\begin{aligned} &\textstyle\int_0^T |A_0\Delta|^2\,dt \le k_5\frac{L}{\lambda_{N+1}}[|Aq|^2_{L^2(0,T;X)} \\ &\quad +\nu|Aq(0)|^2 + |P_2Bu|^2_{L^2(0,T;V)}] + k_5|P_2b(p,p)|^2_{L^2(0,T;X)}. \end{aligned} \tag{3.13}$$

We consider the term $P_2b(p,p)$. Recall that $b(p,p) = P\,(p\cdot\nabla)p$. Since $p \in X_1$ and hence, in particular, $p \in dom(A)$, then we have

$$g := (p\cdot\nabla)p \;\in\; H_0^1(\Omega).$$

Then Pg satisfies

$$P\,g = g - \nabla\tilde{p} \in X,$$

where $\tilde{p} \in H^1(\Omega)/\mathbb{R}$ is given by

$$\begin{cases} \Delta\tilde{p} = \nabla \cdot g, \\ \frac{\partial\tilde{p}}{\partial n} = 0. \end{cases}$$

Note that $\tilde{p} \in H^2(\Omega)$ and $|\nabla\tilde{p}|_{H^1} \le c\,|g|_{H^1}$. Hence

$$|Pg|_{H^1} = |g - \nabla\tilde{p}|_{H^1} \le c\,|g|_{H^1}$$

and $Pg \in H^1(\Omega)\cap X$, with $Pg\cdot n = 0$. (But Pg is not in V since $\tau\cdot Pg \neq 0$, in general.) Recall from [6] that $dom((-A)^\alpha) = dom((-\Delta)^\alpha)\cap X$ for $\alpha \in (0,1)$, where Δ denotes the vector Laplacian. We further refer to [4, Chapter 8], for characterizations and properties of the domains of $(-\Delta)^\alpha$. It follows from [4] that for $\epsilon > 0$ there exists C_ϵ such that

$$|Pg|_{dom((-A)^{\frac{1}{4}-\epsilon})} \le C_\epsilon |Pg|_{H^1(\Omega)}$$

and hence we have

$$|P_2 b(p,p)|_X = |P_2 P(p\cdot\nabla)p|_X \le \frac{1}{\lambda_{N+1}^{\frac{1}{4}-\epsilon}} |(-A)^{\frac{1}{4}-\epsilon} P((p\cdot\nabla)p)| \le \frac{cC_\epsilon}{\lambda_{N+1}^{\frac{1}{4}-\epsilon}} |(p\cdot\nabla)p|_{H^1}.$$

Here and below C_ϵ is a constant independent of N and p. By (3.2) we have

$$|(p\cdot\nabla)p|_{H^1}^2 \le |(p\cdot\nabla)p|^2 + |\nabla(p\cdot\nabla)p|^2 \le |(p\cdot\nabla)p|^2 + 2\big(|(p\cdot\nabla)p_{x_1}|^2$$
$$+|(p\cdot\nabla)p_{x_2}|^2 + |(p_{x_1}\cdot\nabla)p|^2 + |(p_{x_2}\cdot\nabla)p|^2\big) \le C|p|_V^2 |Ap|^2 L.$$

Combining these estimates, we find

$$|P_2\, b(p,p)|_X \le \frac{C_\epsilon\, L^{1/2}}{\lambda_{N+1}^{\frac{1}{4}-\epsilon}} |p|_V |Ap| \le \frac{C_\epsilon\, L^{1/2}}{\lambda_{N+1}^{\frac{1}{4}-\epsilon}} M_1 |Ap|.$$

This estimate can now be used for the last term in (3.13).

For Neumann boundary conditions $V = \{\phi \in H^1(\Omega)^2 : \nabla\cdot\phi = 0\}$. Since $p \in H^2(\Omega) \subset L^\infty(\Omega)$ $g = (p\cdot\nabla)p \in H^1(\Omega)$ and $Pg = g - \nabla\tilde{p}$, where $\tilde{p} \in H^1(\Omega)/\mathbb{R}$ with

$$\begin{cases} \Delta\tilde{p} = \nabla \cdot g, \\ \frac{\partial\tilde{p}}{\partial n} = n\cdot g. \end{cases}$$

Hence $\tilde{p} \in H^2(\Omega)$, $Pg \in H^1(\Omega)$, and $Pg \in V$. In this case, as well as for periodic boundary conditions, $\lambda_{N+1}^{\frac{1}{4}-\epsilon}$ and the above estimates can be replaced by $\lambda_{N+1}^{\frac{1}{2}}$.

We summarize these estimates in the following theorem.

THEOREM 3.2. *If $x_0 \in dom\,(A_0), P_2 Bu \in L^2(0,T;V)$, and $\epsilon > 0$, then there exists a constant k_7 such that*

$$\int_0^T |A_0(q - \Phi(p))|^2\, dt \le k_7\, \frac{L}{(\lambda_{N+1})^\alpha},$$

where $\alpha = 1/2 - \epsilon$ in the case of Dirichlet boundary conditions and $\alpha = 1$ for Neumann or periodic boundary conditions. Here k_7 depends on ϵ and also on $k_5, |Aq|_{L^2(0,T;X)}$, $|Aq(0)|_X, |P_2 Bu|_{L^2(0,T;V)}, \nu$.

3.2. Lipschitz continuity of Φ. We derive Lipschitz estimates for

$$e = \Phi(p_1) - \Phi(p_2),$$

in X and V, where $p_1 \in X_1$ and $p_2 \in X_1$ satisfy $|p_1|_V,\ |p_2|_V \le M_1$. From (2.6) and (3.2), (3.5) we have

$$\nu(Ae, e) = -\big(b(p_1, p_1 - p_2) + b(p_1 - p_2, p_2), e\big) \le c_4(|p_1|_V + |p_2|_V)|p_1 - p_2|L^{\frac{1}{2}}|e|_V$$

and hence

$$|\Phi(p_1) - \Phi(p_2)|_X \le k_8\,|p_1 - p_2|_X, \quad \text{where } k_8 = \frac{2c_4 L^{\frac{1}{2}} M_1}{\nu\sqrt{\lambda_{N+1}}}. \tag{3.14}$$

Note that $k_8 = k_8(N) \to 0$ for $N \to \infty$. We also find

$$\begin{aligned}\nu|e|_V^2 &\le -\big(b(p_1, p_1 - p_2) + b(p_1 - p_2, p_2), e\big)\\ &\le c_4\,(|p_1|_V + |p_2|_V)\,|p_1 - p_2|_V L^{\frac{1}{2}}|e| \le \tfrac{2}{\sqrt{\lambda_{N+1}}} c_4 M_1 L^{\frac{1}{2}}\ |p_1 - p_2|_V |e|_V,\end{aligned}$$

and hence

$$|\Phi(p_1) - \Phi(p_2)|_V \le k_8\,|p_1 - p_2|_V. \tag{3.15}$$

LEMMA 3.1. *Let $|p_1|_V \le M_1, |p_2|_V \le M_1$ and set $k_8 = \frac{2c_4 L^{\frac{1}{2}} M_1}{\nu\sqrt{\lambda_{N+1}}}$. Then*

$$|\Phi(p_1) - \Phi(p_2)|_V \le k_8|p_1 - p_2|_V,\ \textit{and}\ |\Phi(p_1) - \Phi(p_2)|_X \le k_8|p_1 - p_2|_X.$$

3.3. Error analysis for reduced-order system. We derive error estimates for $z - p$, where z and p are the components of the reduced systems given by (2.8) and the first equation in (2.3). The following a priori estimates will be needed.

LEMMA 3.2. *Let $u \in L^2(0,T;U)$ and $x_0 \in V$, and let x denote the solution to (2.2). Then for all N sufficiently large, the solution $\hat{x}$ to (2.5) exists. Moreover, there exists a constant M_1, such that $|x(t)|_V \le M_1$ and $|\hat{x}(t)|_V \le M_1$ for all $t \in [0,T]$, and this constant is uniform with respect to (u, x_0) in bounded sets of $L^2(0,T;U) \times V$ and all N sufficiently large.*

An outline for the proof is given in the appendix.

From (2.8) we have

$$\begin{aligned}&\tfrac{d}{dt}(z(t) - p(t)) - \nu\, A(z(t) - p(t))\\ &\quad + P_1 b(\hat{x}(t), \hat{x}(t) - x(t)) + P_1 b(\hat{x}(t) - x(t), x(t)) = 0, \quad z(0) - p(0) = 0,\end{aligned}$$

where

$$x(t) = p(t) + q(t), \quad \hat{x} = z(t) + \Phi(z(t)).$$

Taking the inner product with $z(t) - p(t)$,

$$\begin{aligned}&\tfrac{1}{2}\tfrac{d}{dt}|z(t) - p(t)|^2 + \nu|z(t) - p(t)|_V^2\\ &\le |(b(\hat{x}, \hat{x} - x), z - p)| + |(b(\hat{x} - x, x), z - p)|\\ &\le |(b(\hat{x}, \Phi(z) - q), z - p)| + |(b(z - p, x), z - p)| + |(b(\Phi(z) - q, x), z - p)|,\end{aligned}$$

where we suppressed the dependence on t on the right-hand side and used

$$(b(\hat{x}(t),\hat{x}(t)-x(t)),z(t)-p(t)) = (b(\hat{x}(t),\Phi(z(t))-q(t)),z(t)-p(t)).$$

From (3.2) and the fact that $\big(b(\hat{x},\hat{x}-x),z-p\big) = -\big(b(\hat{x},\hat{x}-x),z-p\big)$, we find that

$$\begin{aligned}
&\tfrac{1}{2}\tfrac{d}{dt}|z(t)-p(t)|^2+\nu|z(t)-p(t)|_V^2\\
&\quad\le |\hat{x}|_{L^\infty}|z(t)-p(t)|_V\big(|z(t)-p(t)|_X+|q(t)-\Phi(p(t))|_X\big)\\
&\quad+M_1\big(c_2|z(t)-p(t)|_X+c_3L^{\frac{1}{2}}(k_8|z(t)-p(t)|_X+|q(t)-\Phi(p(t))|_X)\big)|z(t)-p(t)|_V.
\end{aligned}$$

Concerning the term $|\hat{x}(t)|_{L^\infty}$, note that for $z\in V$

$$|A_0\Phi(z)|\le c_4L^{\frac{1}{2}}|z|_V^2,$$

and hence

$$\nu|A\Phi(z)|\le c_4\,L^{\frac{1}{2}}|z|_V^2.$$

Moreover, for every $\epsilon>0$ there exists $c>0$ such that

$$|\phi|_{L^\infty}\le c|A^{\frac{1}{2}+\epsilon}\phi| \text{ for all } \phi\in dom(A).$$

Consequently,

$$|\Phi(z(t))|_{L^\infty}\le c\frac{c_4}{\nu}\lambda_{N+1}^{-\frac{1}{2}+\epsilon}L^{\frac{1}{2}}|z(t)|_V^2 \quad\text{and}\quad |z(t)|_{L^\infty}\le c_3L^{\frac{1}{2}}|z(t)|_V,$$

where (3.2) was used for the second estimate. Since $L=L(N)\sim\log\lambda_{N+1}$, we can assume without loss of generality that

$$|\hat{x}(t)|_{L^\infty}=|z(t)+\Phi(z(t))|_{L^\infty}\le\bar{c}L^{\frac{1}{2}}M_1^2 \quad \text{for } t\in[0,T],$$

where $\bar{c}=\max(c_3,\frac{c_4}{\nu})$. Without loss of generality we also assume that $M_1\ge 1$ and, as before, that $\nu\le 1$. Combining these estimates we find that

$$\begin{aligned}
&\tfrac{1}{2}\tfrac{d}{dt}|z(t)-p(t)|^2+\nu|z(t)-p(t)|_V^2\\
&\quad\le M_1^2(c_2+2k_8\bar{c}L^{\frac{1}{2}})\,|z(t)-p(t)|\,|z(t)-p(t)|_V+2M_1^2\bar{c}L^{\frac{1}{2}}\,|q(t)\\
&\qquad-\Phi(p(t))|\,|z(t)-p(t)|_V
\end{aligned}$$

and thus

$$\begin{aligned}
&\tfrac{d}{dt}|z(t)-p(t)|^2+\nu\,|z(t)-p(t)|_V^2\\
&\quad\le 2\frac{(c_2M_1^2+2k_8M_1^2\bar{c}L^{\frac{1}{2}})^2}{\nu}|z(t)-p(t)|^2+\frac{8\bar{c}^2LM_1^4}{\nu}|q(t)-\Phi(p(t)|^2.
\end{aligned}$$

Gronwall's lemma now implies the following theorem.

THEOREM 3.3. *Let z and p denote the solutions to* (2.8) *and the first equation in* (2.2)*, with $x_0\in V$. Then for all N sufficiently large*

$$|z(t)-p(t)|^2+\nu\int_0^t|z(s)-p(s)|_V^2\,dt\le\frac{8\bar{c}^2LM_1^4}{\nu}\int_0^t e^{c(t-s)}\,|q(s)-\Phi(p(s))|^2\,ds,$$

where $\bar{c} = \max(c_3, \frac{c_4}{\nu})$ *and* $c = 2\frac{(c_2 M_1^2 + 2k_8 M_1^2 \bar{c} L^{\frac{1}{2}})^2}{\nu}$, *and* $t \in [0, T]$.

Considering the definitions of k_8 and $\bar{c}$, we deduce that c behaves like $\frac{L^2}{\lambda_{N+1}}$, where $L(N) = 1 + \log\frac{\lambda_{N+1}}{\lambda_1}$, for $N \to \infty$ and, in particular, is bounded with respect to N for $N \to \infty$.

At the end of section 2 we addressed the question of the performance of the finite dimensional controller applied to the infinite dimensional optimal control problem, measured in term of the difference of the control-costs. For the two-dimensional Navier–Stokes equations we obtain, using the results of this section, the following performance indicator in terms of the cost.

COROLLARY 3.1. *Let* $x_0 \in V$, J *as in section* 2, *and let* u^* *denote an optimal solution to* (2.1)–(2.2) *and* u *an optimal solution to* (2.7)–(2.8), *with dynamics given by the Navier–Stokes equations. Then there exists a constant* $k_9 = k_9(C, \nu, \bar{c}, M_1, c, k_7, k_8)$ *independent of* N, *such that*

$$0 \le J(x,u) - J(x^*,u^*) \le k_9 \, \frac{L}{(\lambda_{N+1})^{1+\frac{\alpha}{2}}},$$

where x *denotes the solution to the infinite dimensional system* (2.2) *controlled by* u.

Proof. Let $x^* = x(u^*)$ and $x = x(u)$ denote the solutions to (2.2) with controls u and u^*, respectively. Then by Lemma 3.2 there exists M_1 such that $x^*, x, \hat{x}(u^*)$, and $\hat{x}(u)$ are all bounded in $C([0,T];V)$ by M_1. By (2.14)

$$0 \le J(x,u) - J(x^*,u^*) \le C\big(|x^* - \hat{x}^*|_{L^2(0,T;X)} + |x - \hat{x}|_{L^2(0,T;X)}\big).$$

From (2.13) and Lemma 3.1 we have with $\hat{C} = C(1 + k_8)$ that

$$\begin{aligned} J(x,u) - J(x^*,u^*) \le \hat{C}\big(&|p(u_N) - z(u_N)|_{L^2(0,T;X)} + |q(u_N) - \Phi(z(u_N))|_{L^2(0,T;X)} \\ &+ |p(u^*) - z(u^*)|_{L^2(0,T;X)} + |q(u^*) - \Phi(z(u^*))|_{L^2(0,T;X)}\big). \end{aligned}$$

In the following estimates k denotes a generic constant depending continuously on the indicated arguments but independent of N. From Theorem 3.3, (3.5), and Theorem 3.2, we have

$$\begin{aligned} &J(x,u) - J(x^*,u^*) \\ &\quad \le k(\hat{C}, \nu, \bar{c}, M_1, c)\sqrt{L}\big(|q(u) - \Phi(z(u))|_{L^2(0,T;X)} + |q(u^*) - \Phi(z(u^*))|_{L^2(0,T;X)}\big) \\ &\quad \le k(\hat{C}, \nu, \bar{c}, M_1, c)\tfrac{\sqrt{L}}{\nu\lambda_{N+1}} \\ &\qquad \big(|A_0(q(u) - \Phi(z(u)))|_{L^2(0,T;X)} + |A_0(q(u^*) - \Phi(z(u^*)))|_{L^2(0,T;X)}\big) \\ &\quad \le k(\hat{C}, \nu, \bar{c}, M_1, c, k_7)\frac{L}{(\lambda_{N+1})^{1+\frac{\alpha}{2}}}. \qquad \square \end{aligned}$$

In section 4 these estimates will be carried further to obtain estimates on $u_N - u^*$, where u_N is the solution to (2.7)–(2.8), as $N \to \infty$.

4. Optimality condition. In this section we discuss a direct approach to obtain necessary and sufficient optimality conditions for (2.1)–(2.2), with the dynamical system given by (3.1). As above, let $V = dom\,((-A)^{\frac{1}{2}})$ and $-\langle A\phi, \phi\rangle = |\phi|_V^2$ for $\phi \in V$. Let (x^*, u^*) be an optimal solution to (2.1)–(2.2) with

$$J(x,u) = \frac{1}{2}\int_0^T (Q(x(t) - \bar{x}(t)), x(t) - \bar{x}(t))\,dt + \frac{\beta}{2}\int_0^T |u(t)|_U^2, \tag{4.1}$$

where Q is symmetric and positive definite on X, $\bar{x} \in L^2(0,T;X)$, and $\beta > 0$. We shall use the following properties:

- Given $(x_0, u) \in X \times L^2(0,T;U)$, there exists a unique weak solution x to (2.2) in $W = L^2(0,T;V) \times H^1(0,T;V^*)$ satisfying $x(0) = x_0$ and

$$\left\langle \frac{d}{dt}x(t), \psi \right\rangle = \langle A_0 x(t) + F(x(t)) + Bu(t), \psi \rangle \text{ for all } \psi \in V.$$

- $F : V \to V^*$ is continuously differentiable and there exists a $c > 0$ such that

$$\begin{aligned} &\langle F(y) - F(x), y - x \rangle \le c\,|y-x||y-x|_V|x|_V \\ &\langle F'(x)y, y \rangle \le c\,|y||y|_V|x|_V \\ &|F(y) - F(x) - F'(x)(y-x)|_{V^*} \le c\,|y-x||y-x|_V \end{aligned}$$

for all $x, y \in V$.

The second item follows from (3.2).

4.1. Necessary optimality condition. Since the bilinear form a on $V \times V$ defined by

$$t \to a(t,\phi,\psi) = -\langle A_0\phi, \psi\rangle - \langle \phi, F'(x^*(t))\psi\rangle$$

satisfies

$$a(t,\phi,\phi) \ge \nu|\phi|_V^2 - c|x^*(t)|_V|\phi||\phi|_V \ge \frac{\nu}{2}|\phi|_V^2 - \frac{c^2\,|x^*(t)|_V^2}{2\nu}|\phi|^2,$$

with $t \to |x^*(t)|_V^2 \in L^1(0,T)$, there exists a unique solution $\chi^* \in H^1(0,T;V^*) \cap L^2(0,T;V)$ to the adjoint equation

$$-\frac{d}{dt}\chi^*(t) = A_0\chi^*(t) + F'(x^*(t))^*\chi^*(t) + Q(x^*(t) - \bar{x}(t)), \quad \chi^*(T) = 0. \tag{4.2}$$

Recall that for $x \in W$, we have $x \in C(0,T;X)$, and for $x,\ \chi \in W$ it follows that

$$\frac{d}{dt}(x(t), \chi(t))_X = \left\langle \frac{d}{dt}x(t), \chi(t) \right\rangle + \left\langle x(t), \frac{d}{dt}\chi(t) \right\rangle$$

for a.e. $t \in (0,T)$. Denoting by $x \in W$ the weak solution corresponding to $u \in L^2(0,T;X)$, and using $a\,b \le \frac{1}{4}a^2 + b^2$ we have

$$\begin{aligned} \frac{1}{2}\frac{d}{dt}|x(t) - x^*(t)|^2 + \nu|x - x^*|_V^2 &= \langle F(x) - F(x^*), x - x^*\rangle + \langle B(u-u^*), x - x^*\rangle \\ &\le c\,|x^*|_V|x-x^*|\,|x-x^*|_V + \|B\|_{\mathcal{L}(U,V^*)}|x-x^*|_V|u-u^*| \\ &\le \frac{\nu}{2}|x-x^*|_V^2 + \frac{c^2}{\nu}|x^*|_V^2|x-x^*|^2 + \frac{1}{\nu}\|B\|_{\mathcal{L}(U,V^*)}^2|u-u^*|^2. \end{aligned}$$

By Gronwall's inequality

$$|x(t) - x^*(t)|^2 \le \frac{2}{\nu^2}\|B\|_{\mathcal{L}(U,V^*)}^2\, exp\left(\frac{2}{\nu}c^2\int_0^t |x^*(s)|_V^2\,ds\right)\int_0^t |u(s) - u^*(s)|^2\,ds \tag{4.3}$$

and

$$\int_0^T |x(t)-x(t)|_V^2\,dt \le M\int_0^T |u(t)-u^*(t)|^2\,dt \tag{4.4}$$

with

$$M=\frac{2}{\nu^2}\|B\|^2_{\mathcal{L}(U,V^*)}\left(1+\frac{2}{\nu}c^2\left(\int_0^T |x^*(t)|_V^2\,dt\right)e^{\frac{2}{\nu}c^2\int_0^T |x^*(t)|_V^2\,dt}\right).$$

Note that

$$\begin{aligned}J(u)-J(u^*)&=\beta(u-u^*,u^*)_{L^2(0,T;U)}+(Q(x^*-\bar x),x-x^*)_{L^2(0,T;X)}\\&\quad+\frac{1}{2}\left(\beta|u-u^*|^2_{L^2(0,T;U)}+(Q(x-x^*),(x-x^*))_{L^2(0,T;X)}\right),\end{aligned} \tag{4.5}$$

and

$$\begin{aligned}&\tfrac{d}{dt}(x(t)-x^*(t),\chi^*(t))=\langle F(x)-F(x^*)-F'(x^*)(x-x^*),\chi^*\rangle\\&\quad-(Q(x^*-\bar x),x-x^*)+\langle B(u-u^*),\chi^*\rangle.\end{aligned}$$

Integrating this expression on $[0,T]$,

$$\int_0^T (Q(x^*-\bar x),x-x^*)\,dt=\int_0^T (B^*\chi^*,u-u^*)+(H(x-x^*,x-x^*),\chi^*)\,dt,$$

where

$$H(x-x^*,x-x^*)=F(x)-F(x^*)-F'(x^*)(x-x^*).$$

Then, from (4.5)

(4.6)

$$\begin{aligned}J(u)-J(u^*)&=(\beta u^*+B^*\chi^*,u-u^*)_{L^2(0,T;U)}+\int_0^T (H(x-x^*,x-x^*),\chi^*)\,dt\\&\quad+\frac{1}{2}\left(\beta|u-u^*|^2_{L^2(0,T;U)}+(Q(x-x^*),(x-x^*))_{L^2(0,T;X)}\right),\end{aligned}$$

where

$$\begin{aligned}&\left|\int_0^T (H(x-x^*,x-x^*),\chi^*)\,dt\right|\\&\quad\le c\left(\int_0^T |x-x^*|^2|x-x^*|_V^2\,dt\right)^{\frac12}\left(\int_0^T |\chi^*|_V^2\,dt\right)^{\frac12}.\end{aligned} \tag{4.7}$$

Let $u=u^*+\tau v,\ v\in L^2(0,T;U)$. Then

$$0\le J(u^*+\tau v)-J(u^*)=(\beta u^*+B^*\chi^*,\tau v)_{L^2(0,T;U)}+o(\tau).$$

Dividing by τ and letting τ tend to 0^+ implies that $(\beta u^*+B^*\chi^*,v)\le 0$ for all $v\in L^2(0,T;X)$. Hence the necessary optimality condition is given by

$$\beta u^*+B^*\chi^*=0\quad\text{in } L^2(0,T;U), \tag{4.8}$$

and the optimality system consists of the primal equation (2.2), the adjoint equation (4.2), and the optimality condition (4.8).

4.2. Second order sufficient optimality. Suppose that there exists a constant $\sigma > 0$ such that

$$\begin{aligned}(4.9)\qquad &\frac{1}{2}\left(|u-u^*|^2_{L^2(0,T;U)} + (Q(x-x^*),(x-x^*))_{L^2(0,T;X)}\right)\\ &\quad + \int_0^T (H(x-x^*,x-x^*),\chi^*)\,dt \geq \sigma\,|u-u^*|^2_{L^2(0,T;U)}\end{aligned}$$

in a neighborhood of u^*, where $x = x(u)$ denote the solution to (2.2). Then the second order sufficient optimality condition (2.16) follows. This follows from (4.6) and (4.8) above.

Turning to (4.9), we first note that

$$-\frac{1}{2}\frac{d}{dt}|\chi^*(t)|^2 \leq -\nu|\chi^*(t)|^2_V + c\,|x^*(t)|_V|\chi^*(t)|\,|\chi^*(t)|_V + (Q(x^*(t)-\bar{x}(t)),\chi^*(t)),$$

and hence

$$-\frac{1}{2}\frac{d}{dt}|\chi^*(t)|^2 + \frac{\nu}{2}|\chi^*(t)|^2_V \leq \frac{c^2}{\nu}\,|x^*(t)|^2_V\,|\chi^*(t)|^2 + \frac{1}{\nu}|Q(x^*(t)-\bar{x}(t))|^2_{V^*}.$$

By Gronwall's inequality applied backwards with respect to time and terminal condition $\chi^*(T) = 0$, we find that

$$\int_0^T |\chi^*(t)|^2 \leq \frac{2}{\nu}\int_0^T |Q(x^*-\bar{x})|^2_{V^*}\exp\left(\int_0^T \frac{2c^2}{\nu}|x^*(s)|^2_V\right).$$

Hence there exists a constant $M = M(c, \frac{1}{\nu})$ such that

$$(4.10)\qquad \int_0^T |\chi^*|^2_V \leq M^2\int_0^T |Q(x^*-\bar{x})|^2_{V^*}\,dt.$$

By (4.7) and (4.10)

$$\begin{aligned}&\frac{1}{2}\left(|u-u^*|^2_{L^2(U)} + (Q(x-x^*),(x-x^*))_{L^2(X)}\right) + \int_0^T (H(x-x^*,x-x^*),\chi^*)\,dt\\ &\quad \geq \tfrac{1}{2}\left(|u-u^*|^2_{L^2(U)} + (Q(x-x^*),(x-x^*))_{L^2(X)}\right)\\ &\qquad -cM|x-x^*|_{C(X)}\,|x-x^*|_{L^2(V)}\,|Q(\bar{x}-x^*)|_{L^2(V^*)},\end{aligned}$$

where $L^2(U)$ stands for $L^2(0,T;U)$. Let $B(r)$ denote the ball with center 0 and radius r in $L^2(0,T;U)$. Let $r > 0$ and $x_0 \in V$. Then by Lemma 3.2 estimate (4.9) holds for all $u \in B$, provided that

$$(4.11)\qquad |Q(\bar{x}-x^*)|_{L^2(V^*)}$$

is sufficiently small.

These results can now be utilized to argue convergence of the reduced optimal controls u to an optimal control u^* of (2.1)–(2.2).

THEOREM 4.1. *Let $x_0 \in V$, J as in* (4.1)*, and let u^* denote an optimal solution to* (2.1)–(2.2) *and u_N an optimal solution to* (2.7)–(2.8)*, with dynamics given by the Navier–Stokes equations. Then, if $|Q(\bar{x}-x^*)|_{L^2(V^*)}$ is sufficiently small, there exists*

a constant $k_{11} = k_{11}(C, \sigma, \nu, \bar{c}, M_1, c, k_7, k_8)$*, independent of* N*, such that for all* N *sufficiently large*

$$|u_N - u^*|^2_{L^2(0,T;U)} \le k_{11}\ \frac{L}{(\lambda_{N+1})^{1+\frac{\sigma}{2}}}.$$

Proof. Following the notation already used at the end of section 2, we denote by

$$x(u) \text{ the solution to (2.2) with } u,$$

$$z(u) \text{ the solution to (2.8) with } u.$$

Note that $J(z(u_N) + \Phi(z(u_N)), u_N) \le J(z(0) + \Phi(z(0)), 0)$. By Lemma 3.2 the right-hand side is bounded and hence there exists $r \in \mathbb{R}$ such that $u_N \in B(r)$ for all N. Thus, if (4.11) is chosen such that the second order sufficient optimality condition (4.9) is satisfied, then from (2.16), (2.14), (2.13), and Lemma 3.1 we have with $\hat{C} = \frac{C}{\sigma}(1 + k_8)$ that

$$\begin{aligned}|u_N - u^*|^2_{L^2(0,T;U)} &\le \hat{C}\big(|p(u_N) - z(u_N)|_{L^2(X)} \\ &+|q(u_N) - \Phi(z(u_N))|_{L^2(X)} + |p(u^*) - z(u^*)|_{L^2(X)} + |q(u^*) - \Phi(z(u^*))|_{L^2(X)}\big),\end{aligned}$$

where again $L^2(X)$ is short for $L^2(0, T; X)$. The proof can now be completed by Theorem 3.3, (3.5), and Theorem 3.2 in the same manner as that of Corollary 3.1. ∎

5. Convergence rate analysis. In the previous section convergence and a first rate of convergence estimate were obtained on the basis of the second order sufficient optimality condition. In this section a rate estimate for the solutions of the reduced problem to the original problem is derived by means of the optimality systems. For the sake of simplicity we assume throughout this section that J is of the form given in (4.1) with $Q = I$ and that

$$\text{range}\,(B) \subset X_1. \tag{5.1}$$

Further let (x^*, u^*) denote a solution to (2.1)–(2.2). Then, (x^*, u^*) with $x^*(t) = p(t) + q(t)$ satisfies

$$\frac{d}{dt}p(t) = A_0 p(t) + P_1\, F\big(p(t) + \Phi(p(t)) + \Delta(t)\big) + Bu^*(t), \tag{5.2}$$

where

$$\Delta(t) = q(t) - \Phi(p(t)) \in X_2.$$

Let

$$\tilde{F}(z) = P_1 F(z + \Phi(z)) \text{ for } z \in X_1$$

and recall that $\Phi : X_1 \to X_2$ is C_1 regular.

The first order necessary optimality condition for (2.7), (2.8) is given by

$$\begin{cases} \dfrac{d}{dt}z(t) = A_0 z(t) + \tilde{F}(z(t)) + Bu(t), \\ -\dfrac{d}{dt}\xi(t) = A_0\xi(t) + \tilde{F}'(z(t))^*\xi(t) + (z(t) - P_1\bar{x}) + \Phi'(z(t))^*(\Phi(z(t)) - P_2\bar{x}), \\ \beta\, u(t) + B^*\xi(t) = 0, \end{cases} \tag{5.3}$$

where $z(0) = P_1 x_0$, $\xi(T) = 0$, and $\hat{x}(t) = z(t) + \Phi(z(t))$. Observe that $\beta u(t) + B^* P_1 \xi(t) = 0$, and

$$\begin{aligned} \tilde{F}'(z)h &= P_1 F'(\hat{x})(h + \Phi'(z)h) = F'_{11}(\hat{x})h + F'_{12}(\hat{x})\Phi'(z)h, \\ \tilde{F}'(z)^* h &= F'_{11}(\hat{x})^* h + \Phi'(z)^* F'_{12}(\hat{x})^* h, \end{aligned} \tag{5.4}$$

for $h \in X_1$. Here $F_{ij} = P_i F P_j$. The adjoint equation split into the X_1 and X_2 components and the necessary optimality condition for (2.1)–(2.2) are given by

$$\begin{aligned} -\tfrac{d}{dt}\eta &= A_0\eta + F'_{11}(x^*)^*\eta + F'_{21}(x^*)^*\zeta + p(t) - P_1\bar{x}, \\ -\tfrac{d}{dt}\zeta &= A_0\zeta + F'_{12}(x^*)^*\eta + F'_{22}(x^*)^*\zeta + q(t) - P_2\bar{x}, \\ \beta u^*(t) &+ B^*\eta(t) = 0, \end{aligned} \tag{5.5}$$

where $\eta(T) = 0$, $\zeta(T) = 0$. Note that for $\hat{x} = z + \Phi(z) \in X_1 + X_2$ we have

$$\begin{aligned} \Phi'(z)h_1 &= -A_0^{-1}F'_{21}(z)h_1, \quad h_1 \in X_1, \\ \Phi'(z)^* h_2 &= -F'_{21}(\hat{x})^* A_0^{-1} h_2, \quad h_2 \in X_2. \end{aligned} \tag{5.6}$$

From the second equation of (5.5)

$$\zeta(t) = -A_0^{-1}(F'_{12}(x^*)^*\eta + q(t) - P_2\bar{x}) - E(t)$$

with

$$E(t) = A_0^{-1}\left(\frac{d}{dt}\zeta + F'_{22}(x^*)^*\zeta\right). \tag{5.7}$$

This quantity E will appear in the final estimate. So let us discuss the effect of the terms which enter into the second equation in (5.5), which is the equation for the fast modes in the adjoint system. The input to the system is $F'_{12}(x^*)^*\eta + q(t) - P_2\bar{x}$ which is small if the effect of the X_2-component of the primal equation and the target state $\bar{x}$, as well as the coupling of the nonlinearity from X_1 to X_2 are small. This results in smallness of the term $\int_0^T |\frac{d}{dt}\zeta + F'_{22}(x^*)^*\zeta|_X^2\, dt$ is small. Further A_0 acts on the fast components ζ of the adjoint state and thus

$$|E(t)|_X \le \frac{1}{\lambda_{N+1}}\left|\frac{d}{dt}\zeta + F'_{22}(x^*)^*\zeta\right|_X.$$

Inserting E into the first equation of (5.5) we find that

$$-\frac{d}{dt}\eta = A_0\eta + F'_{11}(x^*)^*\eta - F'_{21}(x^*)^*\big(A_0^{-1}(F'_{12}(x^*)^*\eta + q(t) - P_2\bar{x}) + E\big) + p(t) - P_1\bar{x},$$

and thus by the second equations of (5.4) and (5.6)

(5.8)
$$
\begin{aligned}
-\frac{d}{dt}(\xi-\eta) &= A_0(\xi-\eta)+\tilde F'(z)^*(\xi)-F_{11}'(x^*)^*\eta+z-p+\Phi'(z)^*(\Phi(z)-P_2\bar x)\\
&\quad +(F_{21}'(x^*)^*A_0^{-1})(F_{12}'(x^*)^*\eta+q-P_2\bar x)+F_{21}'(x^*)^*E\\
&= A_0(\xi-\eta)+\tilde F'(z)^*(\xi-\eta)+(\tilde F'(z)^*-\tilde F'(p)^*)\eta+\tilde F'(x^*)^*\eta-F_{11}'(x^*)^*\eta+z-p\\
&\quad +\Phi'(z)^*\Phi(z)+F_{21}'(\hat x)^*A_0^{-1}(q-P_2\bar x)+(F_{21}'(x^*)^*-F_{21}'(\hat x)^*)A_0^{-1}(q-P_2\bar x)\\
&\quad +F_{21}'(x^*)^*A_0^{-1}F_{12}(x^*)^*\eta-\Phi'(z)^*P_2\bar x+F_{21}'(x^*)^*E\\
&= A_0(\xi-\eta)+\tilde F'(z)^*(\xi-\eta)+(\tilde F'(z)^*-\tilde F'(p)^*)\eta+z-p\\
&\quad +\Phi'(z)^*(\Phi(z)-\Phi(p)-\Delta)+(F_{21}'(x^*)^*-F_{21}'(\hat x)^*)A_0^{-1}(q-P_2\bar x)+F_{21}'(x^*)^*E.
\end{aligned}
$$

Similarly we find

(5.9)
$$
\begin{aligned}
\frac{d}{dt}(z-p) &= A_0(z-p)+\tilde F'(z)(z-p)+B(u-u^*)\\
&\quad +P_1F(\hat x)-P_1F(x^*)-P_1F'(\hat x)(\hat x-x^*)\\
&\quad +F_{12}'(\hat x)\big(\Phi(z)-\Phi(p)-\Phi'(\hat x)(z-p)-\Delta\big).
\end{aligned}
$$

Here we used that

$$
\begin{aligned}
P_1F'(\hat x)(\hat x-x^*) &= F_{11}'(\hat x)(z-p)+F_{12}'(\hat x)(\Phi(z)-q)\\
&= F_{11}'(\hat x)(z-p)+F_{12}'(\hat x)(\Phi(z)-\Phi(p)-\Delta),
\end{aligned}
$$

$$
\tilde F'(z)(z-p)=F_{11}'(\hat x)(z-p)+F_{12}'(\hat x)(\Phi'(\hat x)(z-p)).
$$

For the following discussion let $f : L^2(0,T;U)\times L^2(0,T;X_2)\times L^2(0,T;X_2)\to L^2(0,T;U)$ be defined by the forward integration of the control dynamics for $p(t)$, then the backward integration of the adjoint equation for $\xi(t)$, followed by the application of B^*, where $(u,(\Delta,E))\in L^2(0,T;U)\times L^2(0,T;X_2)\times L^2(0,T;X_2)$ are given. Thus f is the composite $u\to p(u)\to\xi(p)\to B^*\xi=B^*P_1\xi$. Then the optimality system for (2.1)–(2.2) can be expressed as

$$
\beta\, u^*+f(u^*,\Delta,E)=0,
$$

and, using (5.8), (5.9) the optimality system for (2.7), (2.8) is

$$
\beta\, u+f(u,0,0)=0.
$$

Now, if f is C^1 and

$$
\beta\, I+H,\quad H=f_u(u,0,0)
$$

is an isomorphism on $L^2(0,T;U)$, then an argument based on the implicit function theorem implies that

$$
|u^*-u|_{L^2(0,T;U)}\le k_{12}\,(|\Delta|_{L^2(0,T;X_2)}+|E|_{L^2(0,T;X_2)}),
$$

provided that $u = u_N$ is sufficiently close to u^*. Sufficient conditions for convergence of u_N to u^* were already given in Theorem 4.1. We prove such an estimate directly. First, observe that

$$\beta\,(u-u^*, u-u^*) + (B^*(\xi-\eta), u-u^*) = 0. \tag{5.10}$$

From (5.8), (5.9), and (5.4) we find that

$$\begin{aligned}\frac{d}{dt}(z-p,\xi-\eta) &= \langle F(\hat{x}) - F(x^*) - F'(\hat{x})(\hat{x}-x^*), \xi(t)-\eta(t)\rangle \\ &\quad -\langle \tilde{F}'(z) - \tilde{F}'(p)(z-p), \eta\rangle - |z-p|^2 - (\Phi'(z)^*(\Phi(z)-\Phi(p)-\Delta), z-p) \\ &\quad +((\Phi'(z)^* - \Phi'(p)^*)(q-P_2\bar{x}), z-p) - \langle F_{21}'(x^*)^*E, z-p\rangle + (B^*(\xi-\eta), u-u^*) \\ &\quad +\langle F_{12}'(x)(\Phi(z)-\Phi(p)-\Phi'(\hat{x})(z-p)-\Delta), \xi-\eta\rangle.\end{aligned}$$

Integrating this expression onto $[0,T]$,

$$\begin{aligned}&\int_0^T (B^*(\xi-\eta), u-u^*)\,dt \\ &= \int_0^T \big[|z-p|^2 + \Phi'(z)^*(\Phi(z)-\Phi(p)-\Delta, z-p) \\ &\qquad +(\Phi'(z)^* - (\Phi'(p)^*)(q-P_2\bar{x}), z-p)\big]\,dt \\ &\quad -\int_0^T [\langle F(\hat{x}) - F(x^*) - F'(\hat{x})(\hat{x}-x^*), \xi-\eta\rangle + \langle(\tilde{F}'(z) - \tilde{F}'(p)(z-p), \eta\rangle \\ &\quad +\langle F_{21}'(x^*)^*E, z-p\rangle]\,dt - \textstyle\int_0^T \langle F_{12}'(x)(\Phi(z)-\Phi(p)-\Phi'(z)(z-p)-\Delta), \xi-\eta\rangle\,dt.\end{aligned} \tag{5.11}$$

Let us now assume that the assumptions of Theorem 4.1 are satisfied, and let U denote a neighborhood of u^* in $L^2(0,T;U)$. Let M_1 be a bound for $|\hat{x}|_{C(0,T;V)}$ with $u \in U$, and let N be sufficiently large such that $u = u_N \in U$. Using the properties of F specified at the beginning of section 4, we verify from (5.9) with techniques as in the proof of Theorem 3.3 that for some constant k_1,

$$|z(t)-p(t)|^2 + \int_0^T |z(s)-p(s)|_V^2\,ds \le k_1 \int_0^T (|u(s)-u^*(s)|^2 + L\,|\Delta(s)|^2)\,ds \tag{5.12}$$

for all $t \in [0,T]$ and N sufficiently large. Similarly by (5.8), if $x^* \in L^2(0,T;dom(-A))$, then there exists a constant k_2 such that for all N sufficiently large so that $u \in U$,

$$\begin{aligned}&|\xi(t)-\eta(t)|^2 + \int_0^T |\xi(s)-\eta(s)|_V^2\,ds \\ &\qquad \le k_2 \textstyle\int_0^T \big(|u(s)-u^*(s)|_U^2 + L\,|\Delta(s)|_V^2 + |E(s)|^2\big)\,ds.\end{aligned} \tag{5.13}$$

Applying the last two estimates to (5.11) we find with Lemma 3.1 and (3.2) the existence of a constant k_3 such that

$$\begin{aligned}\int_0^T (B^*(\xi-\eta), u-u^*)\,dt \le k_3 \int_0^T \big(&|u(s)-u^*(s)|_U^2 \\ &+ L\,|\Delta(s)|^2 + |E(s)|^2\big)\,ds,\end{aligned} \tag{5.14}$$

for u sufficiently close to u^*. The detailed estimates are quite similar to those necessary for proving (5.12) and (5.13). Together with (5.10) we obtain the following result.

THEOREM 5.1. *Under the assumptions of Theorem* 4.1, *if β is sufficiently large, then there exists a constant k_4 such that*

$$|u_N - u^*|_{L^2(0,T;U)} \le \frac{k_4}{\beta}(L^{\frac{1}{2}} |\Delta|_{L^2(0,T;X_2)} + |E|_{L^2(0,T;X_2)}),$$

where Δ defines the gap and E is defined by (5.7).

Referring to (5.11) let us discuss qualitatively which terms enter into the assumption that β is sufficiently large. These are clearly the first and second terms, without Δ, on the right-hand side of (5.11). The third term is quadratic in $|z-p|_{L^2(0,T;V)}$ multiplied by $|q - P_2\bar{x}|_{L^2(0,T;V)}$, which is small for small residue problems. Similarly, the fifth term is a quadratic multiplied by $|\eta|_{L^2(0,T;V)}$, which in view of (5.5) is again small for small residue problems when p is close to $P_1\bar{x}$. The fourth and seventh terms are of cubic nature. The remaining terms involving Δ and E imply no assumption on size of β.

6. Appendix. Here we sketch the proof for Lemma 3.2. The estimate for x is well known [23]. We turn to the estimate for $\hat{x}$. Let $f = Bu$ and $r = \Phi(z)$. Then $\hat{x} = z + r$ satisfies

$$\left(\frac{d}{dt}z, \psi_1\right) - \nu(Az, \psi_1) + (b(z+r, z+r), \psi_1) = (f, \psi_1),$$
$$\nu\,(Ar, \psi_2) - (b(z,z), \psi_2) = 0$$

for all $\psi_1 \in X_1$ and $\psi_2 \in X_2$. Letting $\psi_1 = z$ and $\psi_2 = r$, we have

$$\frac{1}{2}\frac{d}{dt}|z|^2 + \nu\,(|z|_V^2 + |r|_V^2) + b(r, r, z) = (f, z), \tag{6.1}$$

where we used

$$(b(z+r, z+r), z) + (b(z,z), r) = (b(r,r), z).$$

Assume that $|z(t)|_V \le M$ on $[0,T]$. The definition of M will be given at the end of the proof. From (3.2)

(6.2)
$$|r|_V \le c_4 L^{\frac{1}{2}} M \frac{1}{\nu\sqrt{\lambda_{N+1}}}|z|_V, \quad |r|_X \le \frac{1}{\sqrt{\lambda_{N+1}}}|r|_V, \quad |r|_X \le c_4 L^{\frac{1}{2}} M \frac{1}{\nu\sqrt{\lambda_{N+1}}}|z|_X.$$

Since $|(b(r,r), z)| \le c_2 |r||r|_V |z|_V$,

$$|b(r,r,z)| \le c_2 c_4^2 L M^3 \frac{1}{\nu^2(\lambda_{N+1})^{3/2}}|z|_V^2.$$

We select N so that $c_2 c_4^2 L M^3 \dfrac{1}{\nu^2(\lambda_{N+1})^{3/2}} \le \dfrac{\nu}{2}$. Then from (6.1)

$$|z(t)|^2 + \nu\int_0^t |z|_V^2\, ds \le e^t |z(0)|^2 + \int_0^t e^{t-s}|f(s)|^2\, dx \le C, \tag{6.3}$$

for $t \in [0,T]$, where

$$C = e^T |z(0)|^2 + \int_0^T e^{T-s} |f(s)|^2 \, dx. \tag{6.4}$$

Next, letting $\psi_1 = -Az$ and $\psi_2 = -Ar$,

$$\begin{aligned} &\frac{1}{2}\frac{d}{dt}|z|_V^2 + \nu\,|Az|^2 - (b(z+r, z+r), Az) = (f, Az), \\ &\nu\,|Ar|^2 - (b(z,z), Ar) = 0. \end{aligned} \tag{6.5}$$

From the second equation

$$\nu\,|Ar| \le c_4 L^{\frac{1}{2}} |z|_V^2.$$

Thus,

$$|(b(z+r, r), Az)| \le |z+r|_{L^\infty} |r|_V |Az| \le \bar{c} c_4 L M^3 \frac{1}{\nu\sqrt{\lambda_{N+1}}} |z|_V |Az|,$$

where, as shown in section 3, we can assume that $\bar{c} = \max c_3, \frac{c_4}{\nu}$ is chosen such that

$$|z(t) + r(t)|_{L^\infty} \le \bar{c} M^2 L^{\frac{1}{2}}.$$

From (3.2) and (6.2) we deduce that

$$|(b(z+r, z), Az)| \le 2c_1 |z|^{\frac{1}{2}} |z|_V |Az|^{\frac{3}{2}},$$

where we assumed

$$c_4 L^{\frac{1}{2}} M \frac{1}{\nu\sqrt{\lambda_{N+1}}} \le 1.$$

From the first equation in (6.5) we obtain

$$\frac{1}{2}\frac{d}{dt}|z|_V^2 + \frac{\nu}{2}|Az|^2 \le \left(\frac{2}{\nu} d^2 + \frac{108}{\nu^3} c_1^4 |z|^2 |z|_V^2\right) |z|_V^2 + \frac{2}{\nu}|f|^2,$$

where

$$d = c_4^2 L M^3 \frac{1}{\nu^2\sqrt{\lambda_{N+1}}}.$$

We select N such that $d^2 = d(N)^2 \le \frac{\nu}{4}$. Then

$$\frac{1}{2}\frac{d}{dt}|z|_V^2 + \frac{\nu}{2}|Az|^2 \le \left(1 + \frac{216}{\nu^3} c_1^4 C |z|_V^2\right) |z|_V^2 + \frac{4}{\nu}|f|^2.$$

By Gronwall's lemma we obtain

$$|z(t)|_V^2 + \nu \int_0^t |Az|^2 \, ds \le exp\left(\int_0^T \left(1 + \frac{216}{\nu^3} c_1^4 C |z|_V^2\right) ds\right) \left(|z(0)|_V^2 + \frac{4}{\nu}\int_0^T |f|^2 \, ds\right),$$

and hence

(6.6)
$$|z(t)|_V^2 + \nu \int_0^t |Az|^2 \, ds \leq \exp\left(T + \frac{216}{\nu^4} c_1^4 C^2\right) \left(|z(0)|_V^2 + \frac{4}{\nu} \int_0^T |f|^2 \, ds\right) =: M.$$

Note that the estimate for $|z(t)|_X$ by C in (6.3) involves $|z(0)|$ and $|f|_{L^2(0,T;X)} = |B(u)|_{L^2(0,T;X)}$ but is independent of M for all N sufficiently large. Similarly, the estimate for $|z(t)|_V$ in (6.6) involves C, $|z(0)|$, and $|B(u)|_{L^2(0,T;X)}$, but is independent of M for all N sufficiently large. Hence, the theory of ordinary differential equations on the ball with radius M in $V \cap X_1$ allows one to construct a solution to (2.5) with a priori bound for the $z(t)$ component given by M. The uniform bound for the $\Phi(z(t))$ component follows immediately from (6.2).

REFERENCES

[1] J. A. Atwell and B. B. King, *Proper orthogonal decomposition for reduced basis feedback controllers for parabolic equations*, Math. Comput. Modelling, 33 (2001), pp. 1–19.

[2] R. F. Curtain, *Model reduction for control design for distributed parameter systems*, in Research Directions in Distributed Parameter Systems, R. Smith and M. Demetriou, eds., Frontiers Appl. Math. 27, SIAM, Philadelphia, 2003, pp. 95–121.

[3] R. F. Curtain and H. Zwart, *An Introduction to Infinite-Dimensional Linear Systems Theory*, Springer-Verlag, New York, 1995.

[4] H. O. Fattorini, *Infinite-Dimensional Optimization and Control Theory*, Cambridge University Press, Cambribge, UK, 1999.

[5] C. Foias, O. Manley, and R. Temam, *Modeling of the interaction of small and large eddies in two dimensional turbulent flows*, RAIRO Math. Mod. Num. Anal., 22 (1988), pp. 93–114.

[6] H. Fujita and H. Morimoto, *On fractional powers of the Stokes operator*, Proc. Japan Acad., 16 (1970), pp. 1141–1143.

[7] K. Glover, R. F. Curtain, and J. R. Partington, *Realisation and approximation of linear infinite-dimensional systems with error bounds*, SIAM J. Control Optim., 26 (1988), pp. 863–898.

[8] K. Ito, *Strong convergence and convergence rates of approximating solutions for algebraic Riccati equation in Hilbert spaces*, Proc. Conf. Control and Identification of Distributed Parameter Systems (Vorau, Austria, 1986), Lecture Notes in Control and Inform. Sci. 102, Springer-Verlag, 1987, pp. 153–166.

[9] K. Ito and K. Kunisch, *Receding horizon optimal control for infinite dimensional systems*, ESAIM Control Optim. Calc. Var., 8 (2002), pp. 741–760.

[10] K. Ito and K. Kunisch, *Reduced order control based on approximate inertial manifolds*, Linear Algebra Appl., 415 (2006), pp. 531–541.

[11] K. Ito and S. S. Ravindran, *A reduced order method for simulation and control of fluid flows*, J. Comput. Phys., 143 (1998), pp. 403–425.

[12] K. Ito and S. S. Ravindran, *Optimal control of thermally convected fluid flow*, SIAM J. Sci. Comput., 19 (1998), pp. 1847–1869.

[13] K. Ito and J. Schroeter, *Reduced order feedback synthesis for viscous incompressible flows*, Math. Comput. Modelling, 33 (2001), pp. 173–192.

[14] E. A. Jonckheere and L. M. Silverman, *A new set of invariants for linear systems—application to reduced order compensator design*, IEEE Trans. Automat. Control, 28 (1983), pp. 953–964.

[15] K. Kunisch and S. Volkwein, *Galerkin proper orthogonal decomposition methods for parabolic problems*, Numer. Math., 90 (2001), pp. 117–148.

[16] H. V. Ly and H. T. Tran, *Proper orthogonal decomposition for flow calculation and optimal control in a horizontal CVD reactor*, Quart. Appl. Math., 60 (2002), pp.631–656.

[17] L. Machiels, Y. Maday, and A. T. Patera, *Output bounds for reduced-order approximations of elliptic partial differential equations*, Comput. Methods Appl. Mech. Engrg., 190 (2001), pp. 3413-3426.

[18] M. Marion and R. Temam, *Nonlinear Galerkin methods*, SIAM J. Numer. Anal., 26 (1989), pp. 1139–1157.

[19] D. G. Meyer, *Fractional balanced reduction: Model reduction via fractional representation*, IEEE Trans. Automat. Control, 35 (1990), pp. 1341–1345.
[20] R. J. Ober and D. C. MacFarlane, *Balanced canonical forms for minimal systems: A normalized coprime factor approach*, Linear Algebra Appl., 122 (1989), pp. 23–64.
[21] A. Pazy, *Semigroup of Linear Operators and Applications to Partial Differential Equation*, Springer-Verlag, New York, 1983.
[22] C. W. Rowley, *Model reduction for fluids, using balanced proper orthogonal decomposition*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 15 (2003), pp. 997–1013.
[23] R. Temam, *Navier-Stokes Equations*, North–Holland, Amsterdam, 1979.
[24] K. Willcox and J. Peraire, *Balanced model reduction via the proper orthogonal decomposition*, AIAA Journal, 40 (2002), pp. 2323–2330.

SIAM J. NUMER. ANAL.
Vol. 46, No. 6, pp. 2892–2919
© 2008 Society for Industrial and Applied Mathematics

EFFICIENT STRONG INTEGRATORS FOR LINEAR STOCHASTIC SYSTEMS*

GABRIEL LORD†, SIMON J. A. MALHAM†, AND ANKE WIESE†

SJAM would like to dedicate this paper to the memory of Nairo Aparicio, a friend and collaborator who passed away on 20th June 2005

Abstract. We present numerical schemes for the strong solution of linear stochastic differential equations driven by an arbitrary number of Wiener processes. These schemes are based on the Neumann (stochastic Taylor) and Magnus expansions. First, we consider the case when the governing linear diffusion vector fields commute with each other, but not with the linear drift vector field. We prove that numerical methods based on the Magnus expansion are more accurate in the mean-square sense than corresponding stochastic Taylor integration schemes. Second, we derive the maximal rate of convergence for arbitrary multidimensional stochastic integrals approximated by their conditional expectations. Consequently, for general nonlinear stochastic differential equations with noncommuting vector fields, we deduce explicit formulae for the relation between error and computational costs for methods of arbitrary order. Third, we consider the consequences in two numerical studies, one of which is an application arising in stochastic linear-quadratic optimal control.

Key words. linear stochastic differential equations, strong numerical methods, Magnus expansion, stochastic linear-quadratic control

AMS subject classifications. 60H10, 60H35, 93E20

DOI. 10.1137/060656486

1. Introduction.

We are interested in designing efficient numerical schemes for the strong approximation of linear Stratonovich stochastic differential equations of the form

$$y_t = y_0 + \sum_{i=0}^{d} \int_0^t a_i(\tau)\, y_\tau \,\mathrm{d}W_\tau^i, \tag{1.1}$$

where $y \in \mathbb{R}^p$, $W^0 \equiv t$, $(W^1, \dots, W^d)$ is a d-dimensional Wiener process and $a_0(t)$ and $a_i(t)$ are given $p \times p$ coefficient matrices. We call $a_0(t)\, y$ the linear *drift* vector field and $a_i(t)\, y$ for $i = 1, \dots, d$ the linear *diffusion* vector fields. We can express the stochastic differential equation (1.1) more succinctly in the form

$$y = y_0 + \mathsf{K} \circ y, \tag{1.2}$$

where $\mathsf{K} \equiv \mathsf{K}_0 + \mathsf{K}_1 + \cdots + \mathsf{K}_d$ and $(\mathsf{K}_i \circ y)_t \equiv \int_0^t a_i(\tau)\, y_\tau \,\mathrm{d}W_\tau^i$. The solution of the integral equation for y is known as the *Neumann series*, *Peano–Baker series*, *Feynman–Dyson path ordered exponential*, or *Chen–Fleiss series*

$$y_t = (\mathsf{I} - \mathsf{K})^{-1} \circ y_0 \equiv (\mathsf{I} + \mathsf{K} + \mathsf{K}^2 + \mathsf{K}^3 + \cdots) \circ y_0.$$

*Received by the editors April 6, 2006; accepted for publication (in revised form) April 10, 2008; published electronically August 6, 2008. The research in this paper was supported by the EPSRC first grant GR/S22134/01.

http://www.siam.org/journals/sinum/46-6/65648.html

†Maxwell Institute for Mathematical Sciences and School of Mathematical and Computer Sciences, Heriot-Watt University, Edinburgh EH14 4AS, UK (G.J.Lord@hw.ac.uk, S.J.Malham@ma.hw.ac.uk, A.Wiese@hw.ac.uk). A significant portion of this paper was completed while SJAM was visiting the Isaac Newton Institute in the spring of 2007.

The flow-map, or fundamental solution matrix, S_t maps the initial data y_0 to the solution $y_t = S_t\, y_0$ at time $t > 0$. It satisfies an analogous matrix valued stochastic differential equation to (1.2) with the $p \times p$ identity matrix as initial data. The logarithm of the Neumann expansion for the flow-map is the *Magnus expansion*. We can thus write the solution to the stochastic differential equation (1.1) in the form

$$y_t = (\exp \sigma_t)\, y_0,$$

where

$$\sigma_t = \ln\big((\mathsf{I} - \mathsf{K})^{-1} \circ I\big) \equiv \mathsf{K} \circ I + \mathsf{K}^2 \circ I - \tfrac{1}{2}(\mathsf{K} \circ I)^2 + \cdots . \tag{1.3}$$

See Magnus [41], Yamato [60], Kunita [35], Azencott [3], Ben Arous [5], Strichartz [54], Castell [12], Burrage [9], Burrage and Burrage [7], and Baudoin [4] for the derivation and convergence of the original and also stochastic Magnus expansion; Iserles et al. [30] for a deterministic review; Lyons [38] and Sipiläinen [53] for extensions to rough signals; Lyons and Victoir [40] for a recent application to probabilistic methods for solving partial differential equations; and Sussmann [56] for a related product expansion.

In the case when the coefficient matrices $a_i(t) = a_i$, $i = 0, \ldots, d$ are constant and noncommutative, the solution to the linear problem (1.1) is nontrivial and given by the Neumann series or stochastic Taylor expansion (see Kloeden and Platen [33])

$$y_t^{\text{neu}} = \sum_{\ell=0}^{\infty} \sum_{\alpha \in \mathbb{P}_\ell} J_{\alpha_\ell \cdots \alpha_1}(t)\, a_{\alpha_1} \cdots a_{\alpha_\ell}\, y_0, \tag{1.4}$$

where

$$J_{\alpha_\ell \cdots \alpha_1}(t) \equiv \int_0^t \int_0^{\xi_1} \cdots \int_0^{\xi_{\ell-1}} \mathrm{d}W_{\xi_\ell}^{\alpha_\ell} \cdots \mathrm{d}W_{\xi_2}^{\alpha_2}\, \mathrm{d}W_{\xi_1}^{\alpha_1}.$$

Here $\mathbb{P}_\ell$ is the set of all combinations of multi-indices $\alpha = \{\alpha_1, \ldots, \alpha_\ell\}$ of length ℓ with $\alpha_k \in \{0, 1, \ldots, d\}$ for $k = 1, \ldots, \ell$. There are some special noncommutative cases where we can write down an explicit analytical solution. For example the stochastic differential equation $\mathrm{d}y_t = a_1 y_t\, \mathrm{d}W_t^1 + y_t a_2\, \mathrm{d}W_t^2$ with the identity matrix as initial data has the explicit analytical solution $y_t = \exp(a_1 W_t^1) \cdot \exp(a_2 W_t^2)$. However, in general we cannot express the Neumann solution series (1.4) in such a closed form.

Classical numerical schemes such as the Euler–Maruyama and Milstein methods correspond to truncating the stochastic Taylor expansion to generate global strong order 1/2 and order 1 schemes, respectively. Stochastic Runge–Kutta numerical methods have also been derived—see Kloeden and Platen [33] and Talay [57]. At the linear level, the Neumann, stochastic Taylor, and Runge–Kutta type methods are equivalent. In the stochastic context, Magnus integrators have been considered by Castell and Gaines [13], Burrage [9], Burrage and Burrage [7], and Misawa [46].

We present numerical schemes based on truncated Neumann and Magnus expansions. Higher order multiple Stratonovich integrals are approximated across each time-step by their expectations conditioned on the increments of the Wiener processes on suitable subdivisions (see Newton [49] and Gaines and Lyons [22]). What is new in this paper is that we:

1. Prove the strong convergence of the truncated stochastic Magnus expansion for small stepsize.

2. Derive uniformly accurate higher order stochastic integrators based on the Magnus expansion in the case of commuting linear diffusion vector fields.
3. Prove the maximal rate of convergence for arbitrary multidimensional stochastic integrals approximated by their conditional expectations.
4. Derive explicit formulae for the relation between error and computational costs for methods of arbitrary order in the case of general nonlinear, noncommuting governing vector fields.

Our results can be extended to nonlinear stochastic differential equations with analogous conditions on the governing nonlinear vector fields, where the exponential Lie series (replacing the Magnus expansion) can be evaluated using the Castell–Gaines approach.

In the first half of this paper, sections 2–5, we focus on proving the convergence of the truncated Magnus expansion and establishing Magnus integrators that are more accurate than Neumann (stochastic Taylor) schemes of the same order. The numerical schemes we present belong to the important class of *asymptotically efficient* schemes introduced by Newton [49]. Such schemes have the optimal minimum leading error coefficient among all schemes that depend on increments of the underlying Wiener process only. Castell and Gaines [13, 14] prove that the order $1/2$ Magnus integrator driven by a d-dimensional Wiener process and a modified order 1 Magnus integrator driven by a 1-dimensional Wiener process are asymptotically efficient. We extend this result of Castell and Gaines to an arbitrary number of driving Wiener processes. We prove that if we assume the linear diffusion vector fields commute, then an analogously modified order 1 Magnus integrator and a new order $3/2$ Magnus integrator are globally more accurate than their corresponding Neumann integrators.

There are several potential sources of cost contributing to the overall computational effort of a stochastic numerical integration scheme. The main ones are the efforts associated with:

- Evaluation: computing (and combining) the individual terms and special functions such as the matrix exponential.
- Quadrature: the accurate representation of multiple Stratonovich integrals.

There are usually fewer terms in the Magnus expansion compared to the Neumann expansion to the same order, but there is the additional computational expense of computing the matrix exponential. When the cost of computing the matrix exponential is not significant, due to their superior accuracy we expect Magnus integrators to be preferable to classical stochastic numerical integrators. This will be the case for systems that are small (see Moler and Van Loan [47] and Iserles and Zanna [31]) or for large systems when we have only to compute the exponential of a large sparse matrix times given vector data for which we can use Krylov subspace methods (see Moler and Van Loan [47] and Sidje [52]). Magnus integrators are also preferable when using higher order integrators (applied to nonsparse systems of any size) when high accuracies are required. This is because, in this scenario, quadrature computational cost dominates integrator effort.

In the second half of this paper, sections 6–8, we focus on the quadrature cost associated with approximating multiple Stratonovich integrals to a degree of accuracy commensurate with the order of the numerical method implemented. Our conclusions apply generally to the case of nonlinear, noncommuting governing vector fields. The governing set of vector fields and driving path process $(W^1, \dots, W^d)$ generate the unique solution process $y \in \mathbb{R}^p$ to the stochastic differential equation (1.1). For a scalar driving Wiener process W the Itô map $W \mapsto y$ is continuous in the topology of uniform

convergence. For a d-dimensional driving process with $d \geq 2$ the universal limit theorem implies that the Itô map $(W^1, \ldots, W^d) \mapsto y$ is continuous in the p-variation topology, in particular for $2 \leq p < 3$ (see Lyons [38], Lyons and Qian [39], and Malliavin [43]). Since Wiener paths with $d \geq 2$ have finite p-variation for $p > 2$, approximations to y constructed using successively refined approximations to the driving path will converge only to the correct solution y if we include information about the Lévy chordal areas of the driving path (the L^2-norm of the 2-variation of a Wiener process is finite though). Hence if we want to implement a scheme using adaptive stepsize we should consider order 1 or higher pathwise stochastic numerical methods (see Gaines and Lyons [22]).

However, simulating multiple Stratonovich integrals accurately is costly! For classical accounts of this limitation on applying higher order pathwise stochastic numerical schemes, see Kloeden and Platen [33, p. 367], Milstein [44, p. 92], and Schurz [51, p. 58], and for more recent results, see Gaines and Lyons [21, 22], Wiktorsson [58], Cruzeiro, Malliavin, and Thalmaier [17], Stump and Hill [55], Gilseng and Shardlow [26], and Giles [24, 25].

Taking a leaf from Gaines and Lyons [22], we consider whether it is computationally cheaper to collect a set of sample data over a given time interval and then evaluate the solution (conditioned on that sample data) than it is to evaluate the solution frequently, say at every sample time. The resounding result here is that of Clark and Cameron [16] who prove that when the multiple Stratonovich integral J_{12} is approximated by its expectation conditioned on intervening sample points, the maximal rate of L^2-convergence is of order $h/Q^{1/2}$, where h is the integration steplength and Q is the sampling rate. We extend this result to multiple Stratonovich integrals $J_{\alpha_1,\ldots,\alpha_\ell}$ of arbitrary order approximated by their expectation conditioned on intervening information sampled at the rate Q. Indeed we prove that the maximal rate of convergence is $h^{\ell/2}/Q^{1/2}$ when $\alpha_1, \ldots, \alpha_\ell$ are nonzero indices (and an improved rate of convergence if some of them are zero). In practice the key information is how the accuracy achieved scales with the effort required to produce it on the global interval of integration say $[0, T]$, where $T = Nh$. We derive an explicit formula for the relation between the global error and the computational effort required to achieve it for a multiple Stratonovich integral of arbitrary order when the indices $\alpha_1, \ldots, \alpha_\ell$ are distinct. This allows us to infer the effectiveness of strong methods of arbitrary order for systems with noncommuting vector fields. For a given computational effort which method delivers the best accuracy? The answer not only relies on methods that are more accurate at a given order, it is also influenced by three regimes for the stepsize that are distinguished as follows. In the first large stepsize regime the evaluation effort is greater than the quadrature effort; higher order methods produce superior performance for given effort. Quadrature effort exceeds evaluation effort in the second smaller stepsize regime. We show that in this regime when $d = 2$, or when $d \geq 3$ and the order of the method $M \leq 3/2$, then the global error scales with the computational effort with an exponent of $-1/2$. Here more accurate higher order methods still produce superior performance for given effort; but not at an increasing rate as the stepsize is decreased. However, when $d \geq 3$ for strong methods with $M \geq 2$ the global error verses computational effort exponent is worse than $-1/2$ and this distinguishes the third very small stepsize regime. The greater exponent means that eventually lower order methods will deliver greater accuracy for a given effort.

We have chosen to approximate higher order integrals over a given time step by their expectations conditioned on the increments of the Wiener processes on suitable

subdivisions. This is important for adaptive time-step schemes (Gaines and Lyons [22]) and filtering problems where the driving processes (say W^1 and W^2) are observed signals. However, it should be noted that Wiktorsson [58] has provided a practical method for efficiently sampling the set of multiple Stratonovich integrals $\{J_{ij} : i, j = 1, \dots, d\}$ across a given time-step associated with a d-dimensional driving process (see Gilsing and Shardlow [23] for a practical implementation). Wiktorsson simulates the tail distribution in a truncated Karhunen–Loève Fourier series approximation of these integrals which produces a convergence rate of order $hd^{3/2}/Q$, where Q is analogously the number of required independent normally distributed samples.

Other potential sources of computational effort might be path generation and memory access. Path generation effort depends on the application context. This cost is at worst proportional to the quadrature effort where we could subsume it. Memory access efforts depend on the processing and access memory environment. To reveal higher order methods (which typically require more path information) in the best light possible, we have ignored this effect.

Our paper is outlined as follows. We start in section 2 by proving that the exponential of every truncation of the Magnus series converges to the solution of our linear stochastic differential equation (1.1). In section 3 we define the strong error measures we use and how to compute them. Using these, we explicitly compare the local and then global errors for the Magnus and Neumann integrators in section 4 and thus establish our stated results for uniformly accurate Magnus integrators. In section 5 we show that when the linear diffusion vector fields do not commute we cannot expect the corresponding order 1 Magnus integrator to in general be globally more accurate than the order 1 Neumann integrator. We then turn our attention in section 6 to the method of approximating multiple Stratonovich integrals by their conditional expectations. We prove the maximal rate of convergence for an arbitrary multiple Stratonovich integral in section 6. We then use this result in section 7 to show how the global error scales with the computational effort for numerical schemes of arbitrary order. The shuffle algebra of multiple Stratonovich integrals generated by integration by parts allows for different representations and therefore bases for the solution of a stochastic differential equation. Some choices of basis representation are more efficiently approximated than others and in section 8 we investigate the impact of this choice. In section 9 we present numerical experiments that reflect our theoretical results. To illustrate the superior accuracy of the uniformly accurate Magnus methods we apply them to a stochastic Riccati differential system that can be reformulated as a linear system which has commuting diffusion vector fields. Since for the linear system expensive matrix-matrix multiplications can be achieved independent of the path, the Neumann method performs better than an explicit Runge–Kutta type method applied directly to the nonlinear Riccati system. We also numerically solve an explicit linear system with governing linear vector fields that do not commute for two and also three driving Wiener processes—Magnus integrators also exhibit superior accuracy in practice in these cases also. Last, in section 10 we outline how to extend our results to nonlinear stochastic differential equations and propose further extensions and applications.

2. Strong convergence of truncated Magnus series. We consider here the case when the stochastic differential equation (1.1) is driven by d Wiener processes with constant coefficient matrices $a_i(t) = a_i$, $i = 0, 1, \dots, d$. The Neumann expansion has the form shown in (1.4). We construct the Magnus expansion by taking the logarithm of this Neumann series as in (1.3). In the appendix we explicitly give the

Neumann and Magnus expansions up to terms with L^2-norm of order 3/2. Let $\hat{\sigma}_t$ denote the truncated Magnus series

$$\hat{\sigma}_t = \sum_{\alpha \in \mathbb{Q}_m} J_\alpha(t)\, c_\alpha, \tag{2.1}$$

where $\mathbb{Q}_m$ denotes the finite set of multi-indices α for which $\|J_\alpha\|_{L^2}$ is of order up to and including t^m. Note that here m is a half-integer index, $m = 1/2, 1, 3/2, \ldots$. The terms c_α are linear combinations of finitely many (more precisely exactly length α) products of the a_i, $i = 0, 1, \ldots, d$. Let $|\mathbb{Q}_m|$ denote the cardinality of $\mathbb{Q}_m$.

THEOREM 2.1 (convergence). *For any $t \le 1$, the exponential of the truncated Magnus series, $\exp \hat{\sigma}_t$, is square-integrable. Further, if y_t is the solution of the stochastic differential equation* (1.1)*, there exists a constant $C(m)$ such that*

$$\left\| y_t - \exp \hat{\sigma}_t \cdot y_0 \right\|_{L^2} \le C(m)\, t^{m+1/2}. \tag{2.2}$$

Proof. First, we show that $\exp \hat{\sigma}_t \in L^2$. Using the expression (2.1) for $\hat{\sigma}_t$, we see that for any number k, $(\hat{\sigma}_t)^k$ is a sum of $|\mathbb{Q}_m|^k$ terms, each of which is a k-multiple product of terms $J_\alpha c_\alpha$. It follows that

$$\left\| (\hat{\sigma}_t)^k \right\|_{L^2} \le \Big(\max_{\alpha \in \mathbb{Q}_m} \|c_\alpha\|_{\mathrm{op}} \Big)^k \cdot \sum_{\substack{\alpha_i \in \mathbb{Q}_m \\ i = 1, \ldots, k}} \|J_{\alpha(1)} J_{\alpha(2)} \cdots J_{\alpha(k)}\|_{L^2}. \tag{2.3}$$

Note that the maximum of the operator norm $\|\cdot\|_{\mathrm{op}}$ of the coefficient matrices is taken over a finite set. Repeated application of the product rule reveals that the product $J_{\alpha(i)} J_{\alpha(j)}$, where $\alpha(i)$ and $\alpha(j)$ are multi-indices of length $\ell(\alpha(i))$ and $\ell(\alpha(j))$, is a linear combination of $2^{\ell(\alpha(i)) + \ell(\alpha(j)) - 1}$ multiple Stratonovich integrals. Since $\ell(\alpha(i)) \le 2m$ for $i = 1, \ldots, k$, each term $J_{\alpha(1)} J_{\alpha(2)} \cdots J_{\alpha(k)}$ in (2.3) is thus the sum of at most 2^{2mk-1} Stratonovich integrals J_β. We also note that $k \le \ell(\beta) \le 2mk$.

From equation (5.2.34) in Kloeden and Platen [33], every multiple Stratonovich integral J_β can be expressed as a finite sum of at most $2^{\ell(\beta)-1}$ multiple Itô integrals I_γ with $\ell(\gamma) \le \ell(\beta)$. Further, from Remark 5.2.8 in Kloeden and Platen [33], $\ell(\gamma) + \mathrm{n}(\gamma) \ge \ell(\beta) + \mathrm{n}(\beta)$, where $\mathrm{n}(\beta)$ and $\mathrm{n}(\gamma)$ denote the number of zeros in β and γ, respectively. From Lemma 5.7.3 in Kloeden and Platen [33],

$$\|I_\gamma\|_{L^2} \le 2^{\ell(\gamma) - \mathrm{n}(\gamma)}\, t^{(\ell(\gamma) + \mathrm{n}(\gamma))/2}.$$

Noting that $\ell(\gamma) \le \ell(\beta) \le 2mk$ and $\ell(\gamma) + \mathrm{n}(\gamma) \ge k$, it follows that for $t \le 1$, we have $\|J_\beta\|_{L^2} \le 2^{4mk-1}\, t^{k/2}$. Since the right-hand side of (2.3) consists of $|\mathbb{Q}_m|^k\, 2^{2mk-1}$ Stratonovich integrals J_β, we conclude that

$$\left\| \left(\hat{\sigma}_t\right)^k \right\|_{L^2} \le \Big(\max_{\alpha \in \mathbb{Q}_m} \|c_\alpha\|_{\mathrm{op}} \cdot |\mathbb{Q}_m| \cdot 2^{6m} \cdot t^{1/2} \Big)^k.$$

Hence $\exp \hat{\sigma}_t$ is square-integrable.

Second, we prove (2.2). Let $\hat{y}_t$ denote Neumann series solution (1.4) truncated to include terms of order up to and including t^m. We have

$$\left\| y_t - \exp \hat{\sigma}_t \cdot y_0 \right\|_{L^2} \le \left\| y_t - \hat{y}_t \right\|_{L^2} + \left\| \hat{y}_t - \exp \hat{\sigma}_t \cdot y_0 \right\|_{L^2}. \tag{2.4}$$

We know $y_t \in L^2$ (see Gihman and Skorohod [23] or Arnold [2]). Furthermore, for any order m, $\hat{y}_t$ corresponds to the truncated Taylor expansion involving terms of order

up to and including t^m. Hence $\hat{y}_t$ is a strong approximation to y_t to that order with the remainder consisting of $\mathcal{O}(t^{m+1/2})$ terms (see Proposition 5.9.1 in Kloeden and Platen [33]). It follows from the definition of the Magnus series as the logarithm of the flow-map Neumann series that the terms of order up to and including t^m in $\exp\hat{\sigma}_t \cdot y_0$ correspond with $\hat{y}_t$; the error consists of $\mathcal{O}(t^{m+1/2})$ terms. □

Convergence of approximations based on truncations of the stochastic Taylor expansion has been studied in Kloeden and Platen [33]; see Propositions 5.10.1, 5.10.2, and 10.6.3. Ben Arous [5] and Castell [12] prove the remainder of the exponential of any truncation of the Magnus series is bounded in probability as $t \to 0$ (in the full nonlinear case). Burrage [9] shows that the first terms up to and including order 3/2 Magnus expansion coincide with the terms in the Taylor expansion of the same order. Our result holds for any order in L^2 for sufficiently small t. A more detailed analysis is needed to establish results concerning the convergence radius. Similar arguments can be used to study the nonconstant coefficient case with suitable conditions on the coefficient matrices (see Proposition 5.10.1 in Kloeden and Platen [33] for the corresponding result for the Taylor expansion).

Note that above and in subsequent sections, one may equally consider a stochastic differential equation starting at time $t_0 > 0$ with square-integrable $\mathcal{F}_{t_0}$-measurable initial data y_0. Here $(\mathcal{F}_t)_{t\geq 0}$ denotes the underlying filtration.

3. Global and local error. Suppose $S_{t_n,t_{n+1}}$ and $\hat{S}_{t_n,t_{n+1}}$ are the exact and approximate flow-maps across the interval $[t_n, t_{n+1}]$, respectively; both satisfying the usual flow-map semigroup property: composition of flow-maps across successive intervals generates the flow-map across the union of those intervals. We call the difference between the exact and approximate flow-maps

$$R_{t_n,t_{n+1}} \equiv S_{t_n,t_{n+1}} - \hat{S}_{t_n,t_{n+1}} \tag{3.1}$$

the *local flow remainder*. For an approximation $\hat{y}_{t_{n+1}}$ across the interval $[t_n, t_{n+1}]$ the local remainder is thus $R_{t_n,t_{n+1}} y_{t_n}$. Our goal here is to see how the leading order terms in the local remainders accumulate and contribute to the global error.

Definition 3.1 (strong global error). *We define the* strong global error *associated with an approximate solution $\hat{y}_T$ to the stochastic differential equation* (1.1) *over the global interval of integration $[0,T]$ as $\mathcal{E} \equiv \|y_T - \hat{y}_T\|_{L^2}$.*

The global error can be decomposed additively into two components, the global truncation error due to truncation of higher order terms, and the global quadrature error due to the approximation of multiple Stratonovich integrals retained in the approximation. If $[0,T] = \cup_{n=0}^{N-1}[t_n, t_{n+1}]$, where $t_n = nh$, then for small stepsize $h = T/N$ we have

$$\begin{aligned}\mathcal{E} &= \left\| \left(\prod_{n=N-1}^{0} S_{t_n,t_{n+1}} - \prod_{n=N-1}^{0} \hat{S}_{t_n,t_{n+1}} \right) y_0 \right\|_{L^2} \\ &= \left\| \left(\sum_{n=0}^{N-1} \hat{S}_{t_{n+1},t_N} R_{t_n,t_{n+1}} \hat{S}_{t_0,t_n} \right) y_0 \right\|_{L^2} + \mathcal{O}\big(\max_n \|R_{t_n,t_{n+1}}\|_{L^2}^{3/2}\, h^{-3/2}\big). \end{aligned} \tag{3.2}$$

The local flow remainder has the following form in the case of constant coefficients a_i, $i = 1, \ldots, d$, (see, for example, the integrators in the appendix):

$$R_{t_n,t_{n+1}} = \sum_{\alpha} J_\alpha(t_n, t_{n+1})\, c_\alpha.$$

Here α is a multi-index and the terms c_α represent products or commutations of the constant matrices a_i. The J_α represent Stratonovich integrals or linear combinations, of the same order, of products of integrals including permutations of α. The global error $\mathcal{E}^2$ at leading order in the stepsize is thus

$$y_0^{\mathrm{T}} \sum_{\alpha,\beta} \Bigg(\sum_n \mathbb{E}\big(J_\alpha(t_n,t_{n+1})\, J_\beta(t_n,t_{n+1})\big) \mathbb{E}\Big(\big(\hat{S}_{t_{n+1},t_N} c_\alpha \hat{S}_{t_0,t_n}\big)^{\mathrm{T}} \big(\hat{S}_{t_{n+1},t_N} c_\beta \hat{S}_{t_0,t_n}\big) \Big)$$

$$+ \sum_{n\neq m} \mathbb{E}\big(J_\alpha(t_n,t_{n+1})\big)\, \mathbb{E}\big(J_\beta(t_m,t_{m+1})\big) \mathbb{E}\Big(\big(\hat{S}_{t_{n+1},t_N} c_\alpha \hat{S}_{t_0,t_n}\big)^{\mathrm{T}} \big(\hat{S}_{t_{m+1},t_N} c_\beta \hat{S}_{t_0,t_m}\big) \Big) \Bigg) y_0.$$

Hence in the global truncation error we distinguish between the *diagonal sum* consisting of the first sum on the right-hand side above, and the *off-diagonal sum* consisting of the second sum above with $n \neq m$.

Suppose we include in our integrator all terms with local L^2-norm up to and including $\mathcal{O}(h^M)$. The leading terms $J_\alpha c_\alpha$ in $R_{t_n,t_{n+1}}$ thus have L^2-norm $\mathcal{O}(h^{M+1/2})$. Those with zero expectation will contribute to the diagonal sum, generating $\mathcal{O}(h^M)$ terms in the global error, consistent with a global order M integrator. However, those with nonzero expectation contribute to the off-diagonal double sum. They will generate $\mathcal{O}(h^{M-1/2})$ terms in the global error. We must thus either include them in the integrator, or more cheaply, include only their expectations—the corresponding terms $\big(J_\alpha - \mathbb{E}(J_\alpha)\big)c_\alpha$ of order $h^{M+1/2}$ in $R_{t_n,t_{n+1}}$ will then have zero expectation and contribute only through the diagonal sum—see, for example, Milstein [44, p. 12]. This also guarantees that the next order term in the global error estimate (3.2), whose largest term has the upper bound $\max_n \|R_{t_n,t_{n+1}}\|_{L^2}^{3/2} h^{-3/2}$, involves only higher order contributions to the leading $\mathcal{O}(h^M)$ error.

Note that high order integrators may include multiple Stratonovich integral terms. We approximate these multiple integrals by their conditional expectations to the local order of approximation $h^{M+1/2}$ of the numerical method. Hence terms in the *integrator* of the form $J_\alpha c_\alpha$ are in fact approximated by $\mathbb{E}(J_\alpha|\mathcal{F}_Q)c_\alpha$, their expectation conditioned on intervening path information $\mathcal{F}_Q$ (see section 6 for more details). This generates terms of the form $\big(J_\alpha - \mathbb{E}(J_\alpha|\mathcal{F}_Q)\big)c_\alpha$ in the local flow remainder, which have zero expectation and hence contribute to the global error through the diagonal sum generating $\mathcal{O}(h^M)$ terms.

4. Uniformly accurate Magnus integrators. Our goal is to identify a class of Magnus integrators that are more accurate than Neumann (stochastic Taylor) integrators of the same order for any governing set of linear vector fields (for the integrators of order 1 and 3/2, we assume the diffusion vector fields commute). We thus compare the local accuracy of the Neumann and Magnus integrators through the leading terms of their remainders. We consider the case of constant coefficient matrices a_i, $i = 0, 1, \ldots, d$.

The local flow remainder of a Neumann integrator R^{neu} is simply given by the terms not included in the flow-map Neumann approximation. Suppose $\hat{\sigma}$ is the truncated Magnus expansion and that ρ is the corresponding remainder, i.e., $\sigma = \hat{\sigma} + \rho$.

Then the local flow remainder R^{mag} associated with the Magnus approximation is

$$\begin{aligned} R^{\mathrm{mag}} &= \exp\sigma - \exp\hat{\sigma} \\ &= \exp\bigl(\hat{\sigma}+\rho\bigr) - \exp\hat{\sigma} \\ &= \rho + \tfrac{1}{2}(\hat{\sigma}\rho + \rho\hat{\sigma}) + \mathcal{O}(\hat{\sigma}^2\rho). \end{aligned} \tag{4.1}$$

Hence the local flow remainder of a Magnus integrator R^{mag} is the truncated Magnus expansion remainder ρ and higher order terms $\frac{1}{2}(\hat{\sigma}\rho+\rho\hat{\sigma})$ that can contribute to the global error at leading order through their expectations. For the integrators considered in this section, these higher order terms do not contribute in this way; however, for the order 1 integrator we consider in the next section that they do.

DEFINITION 4.1 (uniformly accurate Magnus integrators). *When the linear diffusion vector fields commute so that* $[a_i,a_j]=0$ *for all* $i,j\neq 0$*, we define the order* 1 *and order* $3/2$ uniformly accurate Magnus integrators *by*

$$\hat{\sigma}^{(1)}_{t_n,t_{n+1}} = J_0a_0 + \sum_{i=1}^{d}\bigl(J_ia_i + \tfrac{h^2}{12}[a_i,[a_i,a_0]]\bigr),$$

and

$$\hat{\sigma}^{(3/2)}_{t_n,t_{n+1}} = J_0a_0 + \sum_{i=1}^{d}\bigl(J_ia_i + \tfrac{1}{2}(J_{i0}-J_{0i})[a_0,a_i] + \tfrac{h^2}{12}[a_i,[a_i,a_0]]\bigr).$$

By *uniformly* we mean for any given set of governing linear vector fields (or equivalently coefficient matrices a_i, $i=0,1,\ldots,d$) for which the diffusion vector fields commute, and for any initial data y_0.

THEOREM 4.2 (global error comparison). *For any initial condition* y_0 *and sufficiently small fixed stepsize* $h=t_{n+1}-t_n$*, the order* $1/2$ *Magnus integrator is globally more accurate in* L^2 *than the order* $1/2$ *Neumann integrator. If, in addition, we assume the linear diffusion vector fields commute so that* $[a_i,a_j]=0$ *for all* $i,j\neq 0$*, then the order* 1 *and* $3/2$ uniformly accurate Magnus integrators *are globally more accurate in* L^2 *than the corresponding Neumann integrators. In other words, if* $\mathcal{E}^{\mathrm{mag}}$ *denotes the global error of the order* $1/2$ *Magnus integrator or the uniformly accurate Magnus integrators of order* 1 *or order* $3/2$*, respectively, and* $\mathcal{E}^{\mathrm{neu}}$ *is the global error of the Neumann integrators of the corresponding order, then at each of those orders*

$$\mathcal{E}^{\mathrm{mag}} \leq \mathcal{E}^{\mathrm{neu}}. \tag{4.2}$$

Proof. Let R^{mag} and R^{neu} denote the local flow remainders corresponding to the Magnus and Neumann approximations across the interval $[t_n,t_{n+1}]$ with $t_n=nh$. A direct calculation reveals that

$$\mathbb{E}\bigl((R^{\mathrm{neu}})^{\mathrm{T}}R^{\mathrm{neu}}\bigr) = \mathbb{E}\bigl((R^{\mathrm{mag}})^{\mathrm{T}}R^{\mathrm{mag}}\bigr) + D_M + \mathcal{O}\bigl(h^{2M+1/2}\bigr),$$

where if we set $\hat{R}\equiv R^{\mathrm{neu}}-R^{\mathrm{mag}}$, then

$$D_M \equiv \mathbb{E}\bigl(\hat{R}^{\mathrm{T}}R^{\mathrm{mag}}\bigr) + \mathbb{E}\bigl((R^{\mathrm{mag}})^{\mathrm{T}}\hat{R}\bigr) + \mathbb{E}\bigl(\hat{R}^{\mathrm{T}}\hat{R}\bigr). \tag{4.3}$$

We now show explicitly in each of the three cases of the theorem that at leading order R^{mag} and $\hat{R}$ are uncorrelated and hence D_M is positive semidefinite. This implies that

the local remainder for the Neumann expansion is larger than that of the Magnus expansion. Hereafter assume the indices $i, j, k, l \in \{1, \dots, d\}$.

For the *order* $1/2$ *integrators* we have to leading order (see the form of the expansions in the appendix)

$$R^{\mathrm{mag}} = \sum_{i<j} \tfrac{1}{2}(J_{ij} - J_{ji})[a_j, a_i],$$

and

$$\hat{R} = \sum_i \big(J_{ii} - \tfrac{1}{2}h\big)a_i^2 + \sum_{i<j} \tfrac{1}{2}(J_{ij} + J_{ji})(a_j a_i + a_i a_j),$$

which are uncorrelated by direct inspection.

We henceforth assume $[a_i, a_j] = 0$ for all i, j. For the *uniformly accurate order* 1 *integrator* we have to leading order (again see the appendix)

$$R^{\mathrm{mag}} = \sum_i \tfrac{1}{2}(J_{i0} - J_{0i})[a_0, a_i],$$

and

$$\hat{R} = \tfrac{1}{2}h^2 a_0^2 + \sum_i \tfrac{1}{2}\big(J_{i0} + J_{0i}\big)(a_0 a_i + a_i a_0) + \tfrac{1}{4}h^2(a_i^2 a_0 + a_0 a_i^2)$$

$$+ \sum_{i,j,k} \big(J_{ijk} - \mathbb{E}(J_{ijk})\big)a_k a_j a_i,$$

which again by direct inspection are uncorrelated.

For the *uniformly accurate order* $3/2$ *integrator* the local flow remainders are

$$R^{\mathrm{neu}} = \sum_i (J_{ii0} - \tfrac{1}{4}h^2)a_0 a_i^2 + J_{i0i}a_i a_0 a_i + (J_{0ii} - \tfrac{1}{4}h^2)a_i^2 a_0$$

$$+ \sum_{i<j} \Big((J_{0ji} + J_{0ij})a_i a_j a_0 + J_{j0i}a_i a_0 a_j$$

$$+ J_{i0j}a_j a_0 a_i + (J_{ji0} + J_{ij0})a_0 a_i a_j \Big)$$

$$+ \sum_{i,j,k,l} \big(J_{ijkl} - \mathbb{E}(J_{ijkl})\big)a_l a_k a_j a_i,$$

and

$$R^{\mathrm{mag}} = \sum_i \tfrac{1}{12}(J_i^2 J_0 - h^2 - 6J_{i0i})[a_i, [a_i, a_0]]$$

$$+ \sum_{i<j} \tfrac{1}{6}\big(J_0 J_i J_j - 3(J_{i0j} + J_{j0i})\big)[a_j, [a_i, a_0]].$$

Consequently, we have

$$\begin{aligned}\hat{R} = \tfrac{1}{12}\sum_i \Big(&(6J_{i0}J_i - J_i^2J_0 - 2h^2)a_0a_i^2 + 2(J_i^2J_0 - h^2)a_ia_0a_i \\ &+ (6J_{0i}J_i - J_i^2J_0 - 2h^2)a_i^2a_0\Big) \\ + \tfrac{1}{6}\sum_{i<j}\Big(&\big(3(J_iJ_{0j} + J_jJ_{0i}) - J_0J_iJ_j\big)a_ia_ja_0 \\ &+ \big(J_0J_iJ_j + 3(J_{i0j} - J_{j0i})\big)a_ja_0a_i \\ &+ \big(J_0J_iJ_j + 3(J_{j0i} - J_{i0j})\big)a_ia_0a_j \\ &+ \big(3(J_iJ_{j0} + J_jJ_{i0}) - J_0J_iJ_j\big)a_0a_ia_j\Big) \\ + \sum_{i,j,k,l} &\big(J_{ijkl} - \mathbb{E}(J_{ijkl})\big)a_la_ka_ja_i.\end{aligned}$$

First, we note that the terms in $\hat{R}$ of the form $J_{ijkl}a_la_ka_ja_i$, for which at least three of the distinct indices, are uncorrelated with any terms in R^{mag}. We thus focus on the terms of this form with at most two distinct indices, namely

$$\sum_{i<j}(J_{ii}J_{jj} - \tfrac{1}{4}h^2)a_j^2a_i^2 + \sum_{i\neq j}J_{iii}J_ja_ja_i^3 + \sum_i(J_{iiii} - \tfrac{1}{8}h^2)a_i^4$$

and other remaining terms in $\hat{R}$. Since $\mathbb{E}\big[J_{i0i}|J_i\big] \equiv h(J_i^2 - h)/6$ and $\mathbb{E}\big[J_{i0j}|J_i,J_j\big] \equiv hJ_iJ_j/6$ for $i \neq j$, the following conditional expectations are immediate:

$$\begin{aligned}\mathbb{E}\big[(J_i^2J_0 - h^2)(J_i^2J_0 - h^2 - 6J_{i0i})|J_i\big] &= 0,\\ \mathbb{E}\big[(J_{iiii} - \tfrac{1}{8}h^2)(J_i^2J_0 - h^2 - 6J_{i0i})|J_i\big] &= 0,\\ \mathbb{E}\big[(J_{ii}J_{jj} - \tfrac{1}{4}h^2)(J_i^2J_0 - h^2 - 6J_{i0i})|J_i,J_j\big] &= 0,\\ \mathbb{E}\big[\big(J_0J_iJ_j + 3(J_{i0j} - J_{j0i})\big)\big(J_0J_iJ_j - 3(J_{i0j} + J_{j0i})\big)|J_i,J_j\big] &= 0,\\ \mathbb{E}\big[(J_{ii}J_{jj} - \tfrac{1}{4}h^2)\big(J_0J_iJ_j - 3(J_{i0j} + J_{j0i})\big)|J_i,J_j\big] &= 0,\\ \mathbb{E}\big[J_{iii}J_j\big(J_0J_iJ_j - 3(J_{i0j} + J_{j0i})\big)|J_i,J_j\big] &= 0.\end{aligned}$$

Hence the expectations of the terms shown are also zero. Second, direct computation of the following expectations reveals

$$\begin{aligned}\mathbb{E}\big((6J_{0i}J_i - J_i^2J_0 - 2h^2)(J_i^2J_0 - h^2 - 6J_{i0i})\big) &= 0,\\ \mathbb{E}\Big(\big(3(J_iJ_{0j} + J_jJ_{0i}) - J_0J_iJ_j\big)\big(J_0J_iJ_j - 3(J_{i0j} + J_{j0i})\big)\Big) &= 0.\end{aligned}$$

Hence R^{mag} and $\hat{R}$ are uncorrelated.

The corresponding global comparison results (4.2) now follow directly in each of the three cases above using that the local errors accumulate and contribute to the

global error as diagonal terms in the standard manner described in detail at the end of section 3. Note that we include the terms $\frac{1}{12}h^2[a_i,[a_i,a_0]]$ in the order 1 uniformly accurate Magnus integrator. These terms appear at leading order in the global remainder where they would otherwise generate a nonpositive definite contribution to D_M in (4.3). ∎

Note that the Magnus integrator $\hat{\sigma}^{(1)}$ is an order 1 integrator without the terms $\frac{1}{12}h^2[a_i,[a_i,a_0]]$. However, they are cheap to compute and including them in the integrator guarantees global superior accuracy independent of the set of governing coefficient matrices.

5. Noncommuting linear diffusion vector fields. What happens when the linear diffusion vector fields do not commute, i.e., we have $[a_i,a_j]\neq 0$ for nonzero indices $i\neq j$? Hereafter, assume $i,j,k\in\{1,\dots,d\}$. Consider the case of order 1 integrators. The local flow remainders are

$$R^{\mathrm{neu}}=\sum_i (J_{i0}a_0a_i+J_{0i}a_ia_0)+\sum_{i,j,k}J_{kji}a_ia_ja_k$$

and

$$\begin{aligned}R^{\mathrm{mag}}=&\sum_i \tfrac{1}{2}(J_{i0}-J_{0i})[a_0,a_i]+\sum_{i\neq j}\big(J_{iij}-\tfrac{1}{2}J_iJ_{ij}+\tfrac{1}{12}J_i^2J_j\big)[a_i,[a_i,a_j]]\\&+\sum_{i<j<k}\big((J_{ijk}+\tfrac{1}{2}J_jJ_{ki}+\tfrac{1}{2}J_kJ_{ij}-\tfrac{2}{3}J_iJ_jJ_k)[a_i,[a_j,a_k]]\\&\qquad+(J_{jik}+\tfrac{1}{2}J_iJ_{kj}+\tfrac{1}{2}J_kJ_{ji}-\tfrac{2}{3}J_iJ_jJ_k)[a_j,[a_i,a_k]]\big).\end{aligned}$$

Computing D_M in (4.3) gives

$$D_M=h^2\sum_{i\neq j}U_{iij}^{\mathrm{T}}BU_{iij}+h^2\sum_{i\neq j\neq k}U_{ijk}^{\mathrm{T}}CU_{ijk},$$

where $U_{iij}=(a_ja_i^2,a_ia_ja_i,a_i^2a_j,a_j^3,a_0a_j,a_ja_0)^{\mathrm{T}}\in\mathbb{R}^{6p\times p}$ and, in addition, we have that $U_{ijk}=(a_ka_ja_i,a_ja_ka_i,a_ka_ia_j,a_ia_ka_j,a_ja_ia_k,a_ia_ja_k)^{\mathrm{T}}\in\mathbb{R}^{6p\times p}$. Here $B,C\in\mathbb{R}^{6p\times 6p}$ consist of $p\times p$ diagonal blocks of the form $b_{k\ell}I_p$ and $c_{k\ell}I_p$, where

$$b=\tfrac{1}{144}\begin{pmatrix}31&10&1&18&12&24\\10&4&10&0&0&0\\1&10&31&18&24&12\\18&0&18&60&36&36\\12&0&24&36&36&36\\24&0&12&36&36&36\end{pmatrix},$$

and

$$c=\tfrac{1}{36}\begin{pmatrix}4&1&1&1&1&-2\\1&4&1&-2&1&1\\1&1&4&1&-2&1\\1&-2&1&4&1&1\\1&1&-2&1&4&1\\-2&1&1&1&1&4\end{pmatrix}.$$

Again, c is positive semidefinite with eigenvalues $\{\frac{1}{6},\frac{1}{6},\frac{1}{6},\frac{1}{6},0,0\}$. However, b has eigenvalues $\{\frac{1}{6},\frac{1}{48}(5+\sqrt{41}),\frac{1}{48}(5-\sqrt{41}),0.94465,0.0943205,-0.03897\}$, where the final three values are approximations to the roots of $288x^3-2888x^2+14x+1$.

The eigenvalues of b and c, respectively, are multiple eigenvalues for the matrices B and C, respectively. This implies that there are certain matrix combinations and initial conditions, for which the order 1 Taylor approximation is more accurate in the mean-square sense than the Magnus approximation. However, the two negative values are small in absolute value compared to the positive eigenvalues. For the majority of systems, one can thus expect the Magnus aproximation to be more accurate (as already observed by Sipiläinen [53], Burrage [9], and Burrage and Burrage [7]). For any given linear system of stochastic differential equations, the scheme that is more accurate can be identified using the results above.

In this case there are terms from $\frac{1}{2}(\hat{\sigma}\rho+\rho\hat{\sigma})$ in the Magnus remainder expression (4.1) that appear at leading order in the local flow remainder. These terms are of the form $\frac{1}{12}\big(a_i[a_j,[a_j,a_i]]+a_j[a_i,[a_i,a_j]]\big)h^2$. They make a negative definite contribution to global error, though this can be negated by including them as cheaply computable terms in the Magnus integrator (indeed we recommend doing so).

6. Quadrature. We start by emphasizing that there are two inherent scales:

- *Quadrature scale* Δt on which the discrete Wiener paths are generated.
- *Evaluation scale* h on which the stochastic differential equation is advanced.

The idea is to approximate multiple Stratonovich integrals by their corresponding expectations conditioned on the σ-algebra representing intervening knowledge of the Wiener paths (Clark and Cameron [16], Newton [49], and Gaines and Lyons [22]). Hence we approximate $J_\alpha(t_n,t_{n+1})$ by $\mathbb{E}\big(J_\alpha(t_n,t_{n+1})|\,\mathcal{F}_Q\big)$, where

$$\mathcal{F}_Q=\{\Delta W^i_{t_n+q\,\Delta t}\colon i=1,\dots,d;\,q=0,\dots,Q-1;\,n=0,\dots,N-1\},$$

with $\Delta W^i_{t_n+q\,\Delta t}\equiv W^i_{t_n+(q+1)\,\Delta t}-W^i_{t_n+q\,\Delta t}$, and $Q\Delta t\equiv h$, i.e., Q is the number of Wiener increments. We extend the result of Clark and Cameron [16] on the maximum rate of convergence to arbitrary order multiple Stratonovich integrals.

LEMMA 6.1 (quadrature error). *Suppose at least two of the indices in the multi-index* $\alpha=\{\alpha_1,\dots,\alpha_\ell\}$ *are distinct. Define*

$$j^*\equiv\min\{j\le\ell-1\colon\alpha_i=\alpha_l\;\forall\,i,l>j\}.$$

Let $\mathrm{n}(\alpha)$ *denote the number of zeros in* α, *and* $\mathrm{n}^*(\alpha)=\mathrm{n}(\{\alpha_{j^*},\dots,\alpha_\ell\})$. *Then the* L^2 *error in the multiple Stratonovich integral approximation* $\mathbb{E}\big(J_\alpha(t_n,t_{n+1})|\,\mathcal{F}_Q\big)$ *is*

$$\big\|J_\alpha(t_n,t_{n+1})-\mathbb{E}\big(J_\alpha(t_n,t_{n+1})|\,\mathcal{F}_Q\big)\big\|_{L^2}=\mathcal{O}\bigg(\frac{h^{(\ell+\mathrm{n}(\alpha))/2}}{Q^{(\mathrm{n}^*(\alpha)+1)/2}}\bigg).\tag{6.1}$$

Proof. For any s,t, we write for brevity

$$\widehat{J}_\alpha(s,t)=\mathbb{E}\big(J_\alpha(s,t)\,|\mathcal{F}_Q\big).$$

We define α^{-k} as the multi-index obtained by deleting the last k indices, that is, $\alpha^{-k} = \{\alpha_1, \ldots, \alpha_{\ell-k}\}$. We set $\tau_q \equiv t_n + q\Delta t$, where $q = 0, \ldots, Q-1$. Then

$$\begin{aligned} J_\alpha(t_n, t_{n+1}) &= \sum_{q=0}^{Q-1} \int_{\tau_q}^{\tau_{q+1}} J_{\alpha^{-1}}(t_n, \tau)\, \mathrm{d}W_\tau^{\alpha_\ell} \\ &= \sum_{q=0}^{Q-1} \left(\sum_{k=1}^{\ell-1} J_{\alpha^{-k}}(t_n, \tau_q) J_{\alpha_{\ell-k+1},\ldots,\alpha_\ell}(\tau_q, \tau_{q+1}) + J_\alpha(\tau_q, \tau_{q+1}) \right). \end{aligned}$$

Thus we have

$$\begin{aligned} &J_\alpha(t_n, t_{n+1}) - \widehat{J}_\alpha(t_n, t_{n+1}) \\ &\quad = \sum_{q=0}^{Q-1} \left(\sum_{k=1}^{\ell-1} \big(J_{\alpha^{-k}}(t_n, \tau_q) - \widehat{J}_{\alpha^{-k}}(t_n, \tau_q)\big) J_{\alpha_{\ell-k+1},\ldots,\alpha_\ell}(\tau_q, \tau_{q+1}) \right. \\ &\qquad\qquad + \widehat{J}_{\alpha^{-k}}(t_n, \tau_q) \left(J_{\alpha_{\ell-k+1},\ldots,\alpha_\ell}(\tau_q, \tau_{q+1}) - \widehat{J}_{\alpha_{\ell-k+1},\ldots,\alpha_\ell}(\tau_q, \tau_{q+1})\right) \\ &\qquad\qquad \left. + J_\alpha(\tau_q, \tau_{q+1}) - \widehat{J}_\alpha(\tau_q, \tau_{q+1}) \right). \end{aligned} \tag{6.2}$$

We prove the assertion by induction over ℓ. For $\ell = 2$, the first two terms in the sum (6.2) are zero and

$$\begin{aligned} \Big\| \sum_{q=0}^{Q-1} \big(J_\alpha(\tau_q, \tau_{q+1}) - \widehat{J}_\alpha(\tau_q, \tau_{q+1})\big) \Big\|_2^2 &= \mathcal{O}\big(Q(\Delta t)^{\ell+\mathrm{n}(\alpha)}\big) \\ &= \mathcal{O}\left(\frac{h^{\ell+\mathrm{n}(\alpha)}}{Q^{\ell+\mathrm{n}(\alpha)-1}} \right) \\ &= \mathcal{O}\left(\frac{h^{\ell+\mathrm{n}(\alpha)}}{Q^{\mathrm{n}^*(\alpha)+1}} \right). \end{aligned}$$

Here we have used that fact that $\ell = 2$, $j^* = 1$ and thus $\mathrm{n}(\alpha) = \mathrm{n}^*(\alpha)$.

Assume now that $\ell > 2$. We will investigate the order of each of the three types of terms in (6.2) separately. If at least two indices in α^{-k} are distinct, then by induction hypothesis

$$\big\| J_{\alpha^{-k}}(t_n, \tau_q) - \widehat{J}_{\alpha^{-k}}(t_n, \tau_q) \big\|_2^2 = \mathcal{O}\left(\frac{(\tau_q - t_n)^{\ell-k+\mathrm{n}(\alpha^{-k})}}{q^{\mathrm{n}^*(\alpha^{-k})+1}} \right),$$

and since $\tau_q - t_n = q\Delta t$ and $Q\,\Delta t = h$, we have for each $k = 1, \ldots, \ell-1$,

$$\begin{aligned} &\sum_{q=0}^{Q-1} \big\| \big(J_{\alpha^{-k}}(t_n, \tau_q) - \widehat{J}_{\alpha^{-k}}(t_n, \tau_q)\big)\, J_{\alpha_{\ell-k+1},\ldots,\alpha_\ell)}(\tau_q, \tau_{q+1}) \big\|_2^2 \\ &\qquad = \mathcal{O}\left(\frac{h^{\ell+\mathrm{n}(\alpha)}}{Q^{\mathrm{n}^*(\alpha^{-k})+\mathrm{n}(\alpha)-\mathrm{n}(\alpha^{-k})+k}} \right). \end{aligned}$$

Note that we have

$$\mathrm{n}^*(\alpha^{-k}) + \mathrm{n}(\alpha) - \mathrm{n}(\alpha^{-k}) + k \geq \mathrm{n}^*(\alpha) + k \geq \mathrm{n}^*(\alpha) + 1. \tag{6.3}$$

If all indices in α^{-k} are equal, then $\big\|J_{\alpha^{-k}}(t_n,\tau_q) - \widehat{J}_{\alpha^{-k}}(t_n,\tau_q)\big\|_2^2 = 0$.

For the second term in (6.2) we have for each $k = 1,\dots,\ell-1$,

$$\sum_{q=0}^{Q-1} \big\|\widehat{J}_{\alpha^{-k}}(t_n,\tau_q)\left(J_{\alpha_{\ell-k+1},\dots,\alpha_\ell}(\tau_q,\tau_{q+1}) - \widehat{J}_{\alpha_{\ell-k+1},\dots,\alpha_\ell}(\tau_q,\tau_{q+1})\right)\big\|_2^2$$

$$= \begin{cases} 0 & \text{if } \alpha_\ell = \cdots = \alpha_{\ell-k+1}, \\ \mathcal{O}\Big(\frac{h^{\ell+\mathrm{n}(\alpha)}}{Q^{\mathrm{n}(\alpha)-\mathrm{n}(\alpha^{-k})+k-1}}\Big) & \text{otherwise.} \end{cases}$$

Note that in the second of these cases

$$\mathrm{n}(\alpha) - \mathrm{n}(\alpha^{-k}) + k - 1 \geq \mathrm{n}^*(\alpha) + 1. \tag{6.4}$$

For the third term we have

$$\sum_{q=0}^{Q-1} \big\|J_\alpha(\tau_q,\tau_{q+1}) - \widehat{J}_\alpha(\tau_q,\tau_{q+1})\big\|_2^2 = \mathcal{O}\Big(\frac{h^{\ell+\mathrm{n}(\alpha)}}{Q^{\ell+\mathrm{n}(\alpha)-1}}\Big).$$

Again, note that we have

$$\ell + \mathrm{n}(\alpha) - 1 \geq \mathrm{n}^*(\alpha) + 1. \tag{6.5}$$

Equality holds in at least one of (6.3) to (6.5). To see this we distinguish the case when the last two indices are equal, $\alpha_{\ell-1} = \alpha_\ell$, and the case when the last two indices are distinct, $\alpha_{\ell-1} \neq \alpha_\ell$. If $\alpha_{\ell-1} = \alpha_\ell$, then

$$\mathrm{n}^*(\alpha^{-1}) + \mathrm{n}(\alpha) - \mathrm{n}(\alpha^{-1}) = \mathrm{n}^*(\alpha).$$

Since in this case at least two indices in α^{-1} are distinct, equality holds for $k = 1$ in (6.3). If $\alpha_{\ell-1} \neq \alpha_\ell$, then

$$\mathrm{n}^*(\alpha) = \mathrm{n}(\alpha) - \mathrm{n}(\alpha^{-2}),$$

and thus equality holds for $k = 2$ in (6.4). Hence the lemma follows. □

Note that each multiple Stratonovich integral $J_\alpha(t_n,t_{n+1})$ can be thought of as an ℓ-dimensional topologically conical volume in $(W^1,\dots,W^\ell)$-space. The surface of the conical volume is *panelled* with each panel distinguished by a double Stratonovich integral term involving two consecutive indices from α. The edges between the panels are distinguished by a triple integral and so forth. The conditional expectation approximation $\mathbb{E}\big(J_\alpha(t_n,t_{n+1})\big|\,\mathcal{F}_Q\big)$ can also be decomposed in this way. In the L^2-error estimate for this approximation, the leading terms are given by sums over the panels which also confirm the estimate (6.1).

Approximations of multiple Stratonovich integrals constructed using their conditional expectations are intimately linked to those constructed using paths W_t^i that are approximated by piecewise linear interpolations of the intervening sample points. The difference between the two approaches are asymptotically smaller terms. For more details, see Wong and Zakai [59], Kloeden and Platen [33], Hofmann and Müller-Gronbach [29], and Gyöngy and Michaletzky [27].

TABLE 7.1

Slopes of the logarithm of the global error $\mathcal{E}$ verses the logarithm of the quadrature effort $\mathcal{U}$, i.e., the exponent $-M/\beta(M,\ell)$, for different values of ℓ and M when $J_{\alpha_1\cdots\alpha_\ell}$ has distinct nonzero indices.

	$\ell=2$	$\ell=3$	$\ell=4$	$\ell=5$	$\ell=6$
No zero index					
$M=1$	$-1/2$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$M=3/2$	$-1/2$	$-1/2$	$\cdots$	$\cdots$	$\cdots$
$M=2$	$-1/2$	$-2/5$	$-1/2$	$\cdots$	$\cdots$
$M=5/2$	$-1/2$	$-5/14$	$-5/14$	$-1/2$	$\cdots$
$M=3$	$-1/2$	$-1/3$	$-3/10$	$-1/3$	$-1/2$

7. Global error vs computational effort. We examine in detail the step-size/accuracy regimes for which higher order stochastic integrators are feasible and also when they become less efficient than lower order schemes. In any strong simulation there are two principle sources of computational effort. First, there is evaluation effort $\mathcal{U}^{\text{eval}}$ associated with evaluating the vector fields, their compositions, and any functions such as the matrix exponential. Second, there is the quadrature effort $\mathcal{U}^{\text{quad}}$ associated with approximating multiple stochastic integrals to an accuracy commensurate with the order of the method. For a numerical approximation of order M the computational evaluation effort measured in flops over $N = Th^{-1}$ evaluation steps is

$$\mathcal{U}^{\text{eval}} = (c_M p^2 + c_E)\, Th^{-1}.$$

Here p is the size of the system, c_M represents the number of scalar-matrix multiplications and matrix-matrix additions for the order M truncated Magnus expansion, and c_E is the effort required to compute the matrix exponential. Note that if we implement an order $1/2$ method there is no quadrature effort. Hence since $\mathcal{E} = \mathcal{O}(h^{1/2})$, we have $\mathcal{E} = \mathcal{O}\big((\mathcal{U}^{\text{eval}})^{-1/2}\big)$.

Suppose we are required to simulate $J_{\alpha_1\cdots\alpha_\ell}(t_n, t_{n+1})$ with all of the indices distinct with a global error of order h^M; naturally $\ell \geq 2$ and $M \geq \ell/2$.

LEMMA 7.1. *The quadrature effort $\mathcal{U}$ measured in flops required to approximate $J_{\alpha_1\cdots\alpha_\ell}(t_n, t_{n+1})$ with a global error of order h^M when all of the indices are distinct and nonzero, is to leading order in h*

$$\mathcal{U} = \mathcal{O}\big(h^{-\beta(M,\ell)}\big),$$

where

$$\beta(M,\ell) = (\ell-1)(2M+1-\ell)+1.$$

Since we stipulate the global error associated with the multiple integral approximation to be $\mathcal{E} = \mathcal{O}(h^M)$, we have

$$\mathcal{E} = \mathcal{O}\big(\mathcal{U}^{-M/\beta(M,\ell)}\big).$$

Proof. The quadrature effort required to construct $\mathbb{E}\big(J_{\alpha_1\cdots\alpha_\ell}(t_n, t_{n+1})\,|\,\mathcal{F}_Q\big)$, which is an $(\ell-1)$-multiple sum, over $[0,T]$ is $\mathcal{U} = \mathcal{O}(Q^{\ell-1}N)$ with $N = Th^{-1}$. Using Lemma 6.1 to achieve a global accuracy of order h^M and therefore local L^2-norm of order $h^{M+1/2}$ for this integral, requires that $Q = h^{\ell-1-2M}$. ∎

In Table 7.1 we quote values for the exponent $-M/\beta(M,\ell)$ for different values of M and ℓ in the case when all of the distinct indices $\alpha_1, \ldots, \alpha_\ell$ are nonzero.

Suppose we are given a stochastic differential equation driven by a d-dimensional Wiener process with noncommuting governing vector fields. To successfully implement a strong numerical method of order M we must guarantee that the global error associated with each multiple integral present in the integrator that is approximated by its conditional expectation is also of order M. If we implement a numerical method of order $M \leq d/2$, we will *in general* be required to simulate multiple Stratonovich integrals with distinct indices of length ℓ with $2 \leq \ell \leq 2M \leq d$. We will also have to simulate multiple integrals with repeated indices of length $\ell_r \leq 2M$. These integrals will require the same or fewer quadrature points than those with distinct indices as we can take advantage of the repeated indices—see section 8 for more details. Such integrals therefore represent lower order corrections to the quadrature effort. Similarly multiple integrals with distinct indices that involve a zero index have index length that is one less than similar order multiple integrals with distinct nonzero indices. Hence they will also represent lower order corrections to the quadrature effort.

To examine the scaling exponent in the relation between the global error $\mathcal{E}$ and the quadrature effort $\mathcal{U}^{\text{quad}}$, which is the sum total of the efforts required to approximate all of the required multiple integrals to order h^M, we use Table 7.1 as a guide for the dominant scalings. For methods of order $M \leq d/2$, if $d = 2$ and we implement a method of order $M = 1$, then the dominant exponent is $-1/2$. Similarly if $d = 3$, then order 1 and $3/2$ methods also invoke a dominant scaling exponent of $-1/2$ for the integrals of length $\ell = 2$ and $\ell = 3$. If $d = 4$, then methods of order 1 and $3/2$ have the same scaling exponent of $-1/2$; however, the method of order 2 involves multiple integrals with three indices which are all distinct and the dominant scaling exponent for them is $-2/5$.

If we implement a method of order $M > d/2$, then we will be required to simulate multiple Stratonovich integrals with distinct indices of length ℓ with $2 \leq \ell \leq d$. We must also simulate higher order multiple integrals with indices of length ℓ_r involving repetitions with $d \leq \ell_r \leq 2M$; these may be cheaper to simulate than multiple integrals of the same length with distinct indices (again see section 8). When $d = 2$ the dominant scaling exponent is $-1/2$ for all orders. For $d \geq 3$ the dominant scaling exponent is *at best* $-1/2$, and so forth.

Last, we give an estimate for the critical stepsize h_{cr} below which the quadrature effort dominates the evaluation effort. Since $\ell = M + 1$ minimizes $\beta(M, \ell)$, we have the following estimate.

COROLLARY 7.2. *For the case of general noncommuting governing vector fields and a numerical approximation of order M, we have $\mathcal{U}^{\text{eval}} \geq \mathcal{U}^{\text{quad}}$ if and only if $h \geq h_{\text{cr}}$, where the critical stepsize*

$$h_{\text{cr}} = \mathcal{O}\Big(\big(T(c_M p^2 + c_E)\big)^{-1/(1-\beta(M,\ell_{\max}))} \Big),$$

where $\ell_{\max} = \max\{d, M + 1\}$.

In practice when we implement numerical methods for stochastic differential equations driven by a d-dimensional Wiener process, we expect that for $h \geq h_{\text{cr}}$ the evaluation effort dominates the compuational cost. In this scenario integrators of order M scale like their deterministic counterparts. Consider what we might expect to see in a log-log plot of global error versus computational cost. As a function of increasing computational cost, we expect the global error for each method to fan out with slope $-M$, with higher order methods providing superior accuracy for a given effort. However, once the quadrature effort starts to dominate, the scaling exponents described above take over. When $d = 2$, for example, and all methods dress themselves with the

scaling exponent $-1/2$, we then expect to see parallel graphs with higher order methods still providing superior accuracy for a given cost. However, higher order methods that assume a scaling exponent worse than $-1/2$ will eventually reintersect the graphs of their lower order counterparts, and past that regime they should not be used.

Note that in the case when all of the diffusion vector fields commute, methods of order 1 do not involve any quadrature effort and hence $\mathcal{E} = \mathcal{O}\big((\mathcal{U}^{\text{eval}})^{-1}\big)$. Using Lemma 6.1, we can, by analogy with the arguments in the proof of Lemma 7.1, determine the dominant scaling exponents for methods of order $M \geq 3/2$. For example, the L^2-error associated with approximating J_{0i} by its expectation conditioned on intervening information is of order $h^{3/2}/Q$. Hence we need only choose $Q = h^{-1/2}$ to achieve a global error of order $3/2$. In this case the dominant scaling exponent is -1. However, the L^2-error associated with approximating J_{0ij} for $i \neq j$ is of order $h^2/Q^{1/2}$, whereas for J_{i0j} and J_{ij0} it is of order h^2/Q. For the case of diffusing vector fields, we do not need to simulate J_{0ij}, and so for a method of order 2 the dominant scaling exponent is still -1. However, more generally the effort associated with approximating J_{0ij} dominates the effort associated with the other two integrals.

8. Efficient quadrature bases. When multiple Stratonovich integrals contain repeated indices, are they as cheap to compute as the corresponding lower dimensional integrals with an equal number of distinct indices (none of them repeated)?

Let $i \cdots i_p$ denote the multi-index with p copies of the index i. Repeated integration by parts yields the formulae

$$J_{i\cdots i_p j i\cdots i_q} = \sum_{k=1}^{q} (-1)^{k+1} J_{i\cdots i_k} J_{i\cdots i_p j i\cdots i_{q-k}} + (-1)^{q+2} \int J_{i\cdots i_p} J_{i\cdots i_q} \mathrm{d}J_j, \tag{8.1a}$$

$$J_{i\cdots i_p j\cdots j_q} = \sum_{k=1}^{q-1} (-1)^{k+1} J_{j\cdots j_k} J_{i\cdots i_p j\cdots j_{q-k}} + (-1)^{q+1} \int J_{i\cdots i_p} \mathrm{d}J_{j\cdots j_q}. \tag{8.1b}$$

The first relation (8.1a) suggests that any integral of the form $J_{i\cdots i_p j i\cdots i_q}$ can always be approximated by a single sum. This last statement is true for $q = 1$. If we assume it is true for $q-1$ and apply the relation (8.1a), then we establish by induction that $J_{i\cdots i_p j i\cdots i_q}$ can be approximated by a single sum. A similar induction argument using (8.1b) then also establishes that any integral of the form $J_{i\cdots i_p j\cdots j_q}$ can also be approximated by a single sum. Hence in both cases the quadrature effort is proportional to QN.

Implicit in the relations (8.1a) and (8.1b) is the natural underlying shuffle algebra created by integration by parts (see Gaines [20, 19], Kawksi [32], and Munthe–Kaas and Wright [48]). Two further results are of interest. First, we remark that by integration by parts we have the following two shuffle product results:

$$J_{i_1 i_2 i_3} J_{i_4} = J_{i_1 i_2 i_3 i_4} + J_{i_1 i_2 i_4 i_3} + J_{i_1 i_4 i_2 i_3} + J_{i_4 i_1 i_2 i_3}, \tag{8.2a}$$

$$J_{i_1 i_2} J_{i_3 i_4} = J_{i_1 i_2 i_3 i_4} + J_{i_1 i_3 i_2 i_4} + J_{i_3 i_1 i_2 i_4} + J_{i_3 i_4 i_1 i_2} + J_{i_3 i_1 i_4 i_2} + J_{i_1 i_3 i_4 i_2}. \tag{8.2b}$$

If we replace $\{i_1, i_2, i_3, i_4\}$ by $\{i, i, j, j\}$ in (8.2b) and (8.2a) and then by $\{i, j, i, j\}$ in (8.2a) and (8.2b), respectively, we obtain the linear system of equations

$$\begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 2 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 2 \end{pmatrix} \begin{pmatrix} J_{jiji} \\ J_{ijji} \\ J_{jiij} \\ J_{ijij} \end{pmatrix} = \begin{pmatrix} J_{ii} J_{jj} - J_{iijj} - J_{jjii} \\ J_{iji} J_j - J_{jjii} \\ J_{iij} J_j - 2 J_{iijj} \\ J_{ij} J_{ij} - 4 J_{iijj} \end{pmatrix}. \tag{8.3}$$

By direct inspection the coefficient matrix on the left-hand side has rank 4 and so all of the multiple Stratonovich integrals J_{jiji}, J_{ijji}, J_{jiij}, and J_{ijij} can be expressed in terms of J_{iijj} and J_{jjii} and products of lower order integrals, all of which can be approximated by single sums (note that $J_{iji} \equiv J_{ij}J_i - 2J_{iij}$).

Now consider the set of multiple Stratonovich integrals

$$\mathfrak{J} = \big\{J_{i_1i_2i_3i_4i_5} : \{i_1,i_2,i_3,i_4,i_5\} \in \text{perms}\{i,i,i,j,j\}\big\} \backslash \big\{J_{iiijj}, J_{jjiii}\big\},$$

where we exclude the elements J_{iiijj} and J_{jjiii} which we know can be approximated by single sums from (8.1b). By considering the shuffle relations generated by products of the form $J_{i_1i_2i_3i_4}J_{i_5}$, $J_{i_1i_2i_3}J_{i_4i_5}$, $J_{i_1i_2}J_{i_3i_4i_5}J_{i_5}$ and $J_{i_1}J_{i_2}J_{i_3}J_{i_4}J_{i_5}$ and substituting in the ten elements with indices from perms$\{i,i,i,j,j\}$ we obtain a linear system of equations analogous to (8.3) with 50 equations for the 8 unknowns in $\mathfrak{J}$. However, direct calculation shows that the corresponding coefficient matrix has rank 7. In particular, all of the multiple integrals in $\mathfrak{J}$ can be expressed in terms of J_{iiijj}, J_{jjiii}, and J_{jiji}. Hence the set of multiple integrals with indices from perms$\{i,i,i,j,j\}$ cannot all be approximated by single sums, but in fact require a double sum to approximate J_{jiji}.

For simplicity, assume $d = 1$. Consider numerical schemes of increasing order M. If $3/2 \le M \le 3$, then all of the necessary multiple integrals can be approximated by single sums—at the highest order in this range indices involving permutations of $\{1,1,1,1,0\}$ and $\{1,1,0,0\}$ are included for which the corresponding integrals can be approximated by single sums. For methods of order $M \ge 7/2$, we require at least double sums to approximate the necessary multiple integrals.

When $d = 2$, for methods of order $M = 1, 3/2$ the integrals involved can be approximated by single sums, but for $M = 2$ integrals involving indices with permutations of $\{2,1,0\}$ are included which can be approximated only by double sums. If there were no drift vector field, then for $1 \le M \le 2$ the necessary multiple integrals can all be approximated by single sums, but for $M = 5/2$ we need to include multiple integrals involving indices with permutations of $\{2,2,1,1,1\}$ which require approximation by double sums. We can in principle extend these results to higher values M; however, methods of order $M \ge d/2$ for $d \ge 3$ are not commonly implemented!

9. Numerical simulations.

9.1. Riccati system. Our first application is for stochastic Riccati differential systems—some classes of which can be reformulated as linear systems (see Freiling [18] and Schiff and Shnider [50]). Such systems arise in stochastic linear-quadratic optimal control problems, for example, mean-variance hedging in finance (see Bobrovnytska and Schweizer [6] and Kohlmann and Tang [34])—though often these are backward problems (which we intend to investigate in a separate study). Consider, for example, Riccati equations of the form

$$u_t = u_0 + \sum_{i=0}^{d} \int_0^t \big(u_\tau A_i(\tau) u_\tau + B_i(\tau) u_\tau + u_\tau C_i(\tau) + D_i(\tau)\big)\, \mathrm{d}W_\tau^i.$$

If $y = (U\ V)^{\mathrm{T}}$ satisfies the linear stochastic differential system (1.1), with

$$a_i(t) \equiv \begin{pmatrix} B_i(t) & D_i(t) \\ -A_i(t) & -C_i(t) \end{pmatrix},$$

then $u = UV^{-1}$ solves the previous Riccati equation.

We consider here a Riccati problem with two additive Wiener processes, W^1 and W^2, and coefficient matrices

$$D_0 = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ 0 & 1 \end{pmatrix}, \qquad D_1 = \begin{pmatrix} 0 & 1 \\ -\frac{1}{2} & -\frac{51}{200} \end{pmatrix}, \qquad \text{and} \qquad D_2 = \begin{pmatrix} 1 & 1 \\ 1 & \frac{1}{2} \end{pmatrix}, \tag{9.1}$$

and

$$A_0 = \begin{pmatrix} -1 & 1 \\ -\frac{1}{2} & -1 \end{pmatrix} \qquad \text{and} \qquad C_0 = \begin{pmatrix} -\frac{1}{2} & 0 \\ -1 & -1 \end{pmatrix}.$$

All other coefficient matrices are zero. The initial data is the 2×2 identity matrix, i.e., $u_0 = I_2$ and therefore $U_0 = I_2$ and $V_0 = I_2$ also. We found the paper by Higham [28] a very useful starting point for our MATLAB simulations.

Note that for this example the coefficient matrices a_1 and a_2 are upper right block triangular and therefore nilpotent of degree 2, and also that $a_1 a_2$ and $a_2 a_1$ are identically zero so that, in particular, $[a_1, a_2] = 0$. The number of terms in each integrator at either order 1 or 3/2 is roughly equal, and so for a given stepsize the uniformly accurate Magnus integrators should be more expensive to compute due to the cost of computing the 4×4 matrix exponential—we used a $(6, 6)$ Padé approximation with scaling to compute the matrix exponential; see Moler and Van Loan [47] and also Iserles and Zanna [31] (the computational cost is roughly 6 times the system size cubed). Also note that the order 1 integrators do not involve quadrature effort whilst the order 3/2 integrators involve the quadrature effort associated with approximating J_{10} and J_{20}. For comparison, we use a nonlinear Runge–Kutta type order 3/2 scheme for the case of two additive noise terms (from Kloeden and Platen [33, p. 383]) applied directly to the original Riccati equation:

$$\begin{aligned} S_{t_n,t_{n+1}} &= S_{t_n} + f\big(S_{t_n}\big)h + D_1 J_1 + D_2 J_2 \\ &\quad + \tfrac{h}{4}\big(f(Y_1^+) + f(Y_1^-) + f(Y_2^+) + f(Y_2^-) - 4f\big(S_{t_n}\big)\big) \\ &\quad + \tfrac{1}{2\sqrt{h}}\big(\big(f(Y_1^+) - f(Y_1^-)\big)J_{10} + \big(f(Y_2^+) - f(Y_2^-)\big)J_{20}\big), \end{aligned} \tag{9.2}$$

where $Y_j^{\pm} = S_{t_n} + \frac{h}{2} f\big(S_{t_n}\big) \pm D_j \sqrt{h}$ and $f(S) = SA_0S + B_0S + SC_0 + D_0$.

In Figure 9.1 we show how the global error scales with stepsize and also CPU clocktime for this Riccati problem. Note that, as anticipated, for the same stepsize (compare respective plot points starting from the left), the order 1 Magnus integrator is more expensive to compute and more accurate than the order 1 Neumann integrator. Now compare the order 3/2 integrators. For the nonlinear scheme (9.2), we must evaluate $f(S)$ five times per step per path costing $20p^3 + 54p^2$ flops—here $p = 2$ refers to the size of the original Riccati system. For the Neumann and Magnus integrators the evaluation costs are $16(2p \times p) = 32p^2$ and $6(2p)^3 + 11(2p)^2 = 48p^3 + 44p^2$ flops, respectively (directly counting from the schemes). Hence for large stepsize we expect the Neumann integrator to be cheapest and the Magnus and nonlinear Runge–Kutta integrators to be more expensive. However, for much smaller stepsizes the quadrature effort should start to dominate. The efforts of all the order 3/2 integrators will not be much different and the Magnus integrator then outperforms the other two due to its superior accuracy.

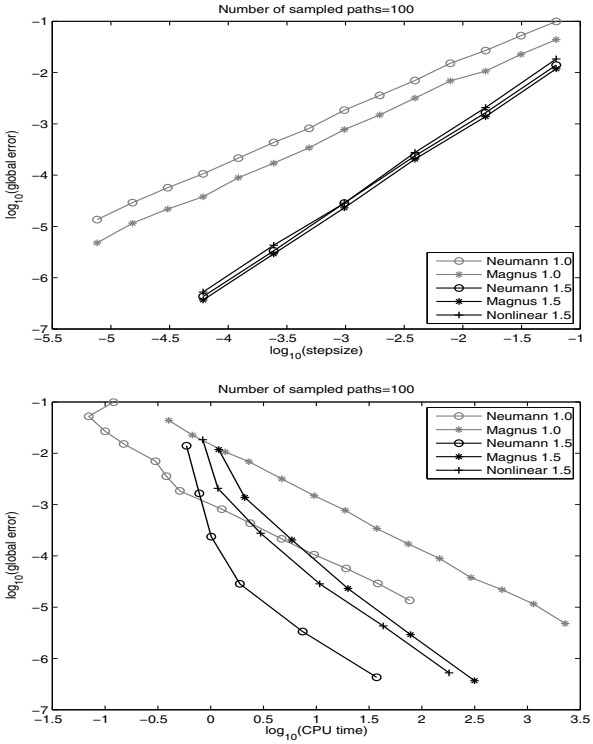


FIG. 9.1. *Global error vs stepsize (top) and vs CPU clocktime (bottom) for the Riccati problem at time* $t = 1$. *The Magnus integrators of order* 1 *and* 3/2 *shown are the uniformly accurate Magnus integrators from section* 4.

9.2. Linear system. Our second numerical example is for a homogeneous and constant coefficient linear problem involving two Wiener processes with coefficient matrices $a_i \equiv D_i$, where the D_i, $i = 0, 1, 2$ are given in (9.1) and do *not commute*, and initial data $y_0 = (\frac{1}{2}\ 1)^T$. In Figure 9.2 we show how the error scales with stepsize and CPU clocktime. We see that the superior accuracy of the Magnus integrators is achieved for the same computational cost. Note that in the case $M = 1$ we have $d = \ell = 2$. For the case when $h \leq h_{\rm cr}$, when computational cost is dominated by quadrature effort, the relation between the global error $\mathcal{E}_1$ and computational cost $\mathcal{U}$ is, ignoring $\mathcal{U}^{\rm eval}$ and taking the logarithm,

$$\log \mathcal{E}_1 \approx \log K_1 - \tfrac{1}{2} \log \mathcal{U}.$$

For the order 1/2 Magnus method the computational cost $\mathcal{U}$ is given solely by the evaluation effort and therefore we have

$$\log \mathcal{E}_{1/2} = \log K_{1/2} + \tfrac{1}{2} \log T(c_{1/2} p^2 + c_E) - \tfrac{1}{2} \log \mathcal{U}.$$

Further note that $K_{1/2}$ and K_1 are strictly order 1, and that $T = 1$, $p = 2$, $c_{1/2} = 5$, and $c_1 = 7$ (counting directly from the integration scheme). In addition, $c_E \approx 6p^3 = 48$ flops using the $(6, 6)$ Padé approximation with scaling. Substituting these values into the difference of these last two estimates reveals that

$$\log \mathcal{E}_1 - \log \mathcal{E}_{1/2} \approx -\tfrac{1}{2} \log 68 \approx -0.9,$$

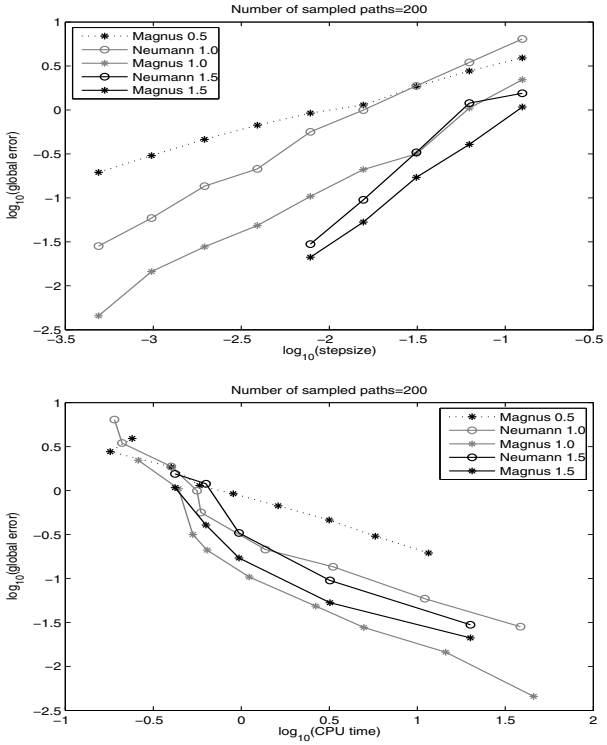


FIG. 9.2. *Global error vs stepsize (top) and vs CPU clocktime (bottom) for the model problem at time* $t = 1$ *with* 2 *driving Wiener processes. The error corresponding to the largest stepsize takes the shortest time to compute.*

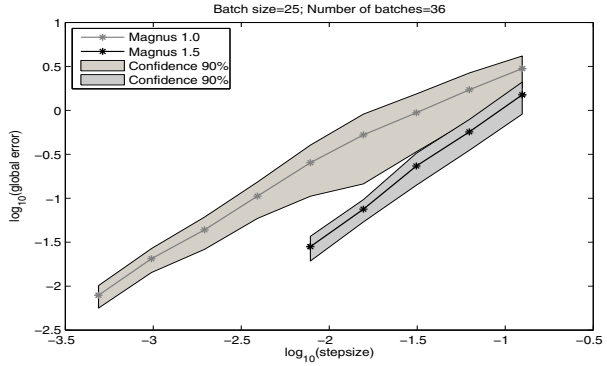


FIG. 9.3. *Confidence intervals for the global errors of the uniformly accurate Magnus integrators for the model problem at time* $t = 1$ *with* 2 *driving Wiener processes.*

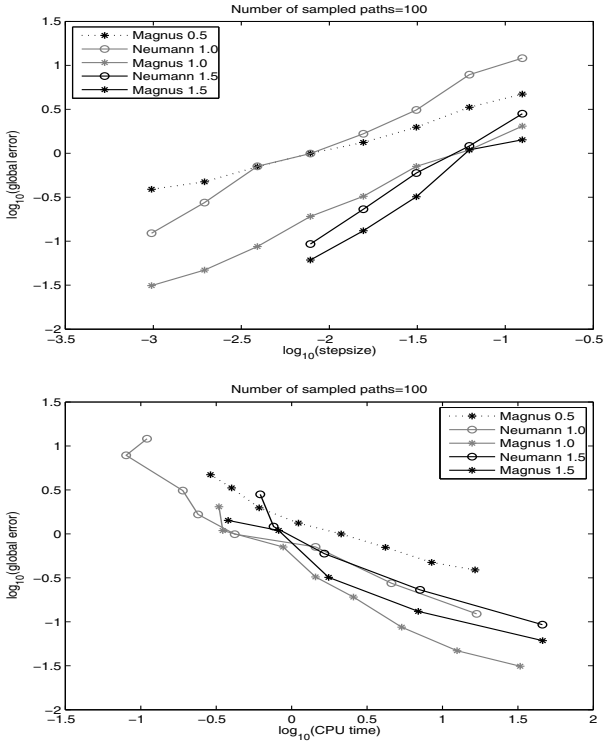


FIG. 9.4. *Global error vs stepsize (top) and vs CPU clocktime (bottom) for the model problem at time* $t = 1$ *with* 3 *driving Wiener processes. The error corresponding to the largest step size takes the shortest time to compute.*

which is in good agreement with the difference shown in Figure 9.2.

Also for this example, we have shown 90% confidence intervals in Figure 9.3 for the global errors of the uniformly accurate Magnus integrators of orders 1 and 3/2. The confidence intervals become narrower as the stepsize decreases and order of the method increases, as expected—see Kloeden and Platen [33, pp. 312–316].

In Figure 9.4 we consider a linear stochastic differential system driven by *three* independent scalar Wiener processes, with the same vector fields as in the linear system with two Wiener processes just considered but with an additional linear diffusion vector field characterized by the coefficient matrix

$$a_3 = \begin{pmatrix} \frac{1}{4} & \frac{2}{5} \\ \frac{1}{6} & \frac{1}{7} \end{pmatrix},$$

which does not commute with a_0, a_1, or a_2. Using Table 7.1 we would expect to see for small stepsizes for the order 1 and 3/2 Magnus and Neumann methods, that the global error scales with the computational effort with exponent $-1/2$. This can be seen in Figure 9.4.

10. Concluding remarks. Our results can be extended to nonlinear stochastic differential equations. Consider a stochastic differential system governed by $(d+1)$ nonlinear autonomous vector fields $V_i(y)$—instead of the linear vector fields $a_i(t)y$ in (1.1)—and driven by a d-dimensional Wiener process $(W^1, \ldots, W^d)$. If we take

the logarithm of the stochastic Taylor series for the flow-map, then we obtain the exponential Lie series (see Chen [15] and Strichartz [54])

$$\sigma_t = \sum_{i=0}^{d} J_i(t) V_i + \sum_{j>i=0}^{d} \tfrac{1}{2}(J_{ij} - J_{ji})(t)[V_i, V_j] + \cdots .$$

Here $[\cdot\,,\cdot]$ is the Lie–Jacobi bracket on the Lie algebra of vector fields defined on $\mathbb{R}^p$. The solution y_t of the system at time $t > 0$ is given by $y_t = \exp \sigma_t \circ y_0$ (see, for example, Ben Arous [5] or Castell and Gaines [13, 14]). Across the interval $[t_n, t_{n+1}]$, let $\hat{\sigma}_{t_n,t_{n+1}}$ denote the exponential Lie series truncated to a given order, with multiple integrals approximated by their expectations conditioned on intervening sample points. Then

$$\hat{y}_{t_{n+1}} = \exp\bigl(\hat{\sigma}_{t_n,t_{n+1}}\bigr) \circ y_{t_n}$$

is an approximation to the exact solution $y_{t_{n+1}}$ at the endpoint of the interval. The truncated and conditioned exponential Lie series $\hat{\sigma}_{t_n,t_{n+1}}$ is itself an ordinary vector field. Hence to compute $\hat{y}_{t_{n+1}}$, we solve the ordinary differential system

$$u'(\tau) = \hat{\sigma}_{t_n,t_{n+1}} \circ u(\tau)$$

with $u(0) = y_{t_n}$ across the interval $\tau \in [0, 1]$ (see Castell and Gaines [13, 14]). If we use a sufficiently accurate ordinary differential integrator commensurate with the order of the truncation of the exponential Lie series, then $u(1) \approx \hat{y}_{t_{n+1}}$ to that order.

Hence the order 1/2 exponential Lie series integrator is more accurate than the Euler–Maruyama method. Further in the case of commuting diffusion vector fields, the uniformly accurate exponential Lie series integrators of order 1 and 3/2 are more accurate than the stochastic Taylor approximations of the corresponding order. This generalization is discussed in Malham and Wiese [42]. An important future investigation to justify the viability of these schemes would be the relation between global error and computational cost, which must take into account the additional computational effort associated with the ordinary differential solver.

An important application for our results that we have in mind for the future are large order problems driven by a large number of Wiener processes. High-dimensional problems occur in many financial applications, for example in portfolio optimization or risk management and in the context of option pricing when high-dimensional models are used, for example for the pricing of interest rate options or rainbow options. Large order problems also arise when numerically solving stochastic parabolic partial differential equations driven by a multiplicative noise term which is white noise in time and spatially smooth (see, for example, Lord and Shardlow [37]). Here we think of projecting the system onto a finite spatial basis set which results in a large system of coupled ordinary stochastic differential equations each driven by a multiplicative noise term. The high dimension d of the driving Wiener process will now be an important factor in the computational cost for order $M \geq 1$ as, for example, we will need to simulate $\frac{1}{2}d(d-1)$ multiple integrals J_{ij}; though the results of Wiktorsson [58] suggest this can be improved upon. Krylov subspace methods for computing large matrix exponentials would be important for efficient implementation of our methods for this case (see Moler and Van Loan [47] and Sidje [52]).

Last, extensions of our work that we also intend to investigate further are: (1) implementing a variable step scheme following Gaines and Lyons [22], Lamba, Mattingly and Stuart [36], Burrage and Burrage [10], and Burrage, Burrage, and Tian [8]—by

using analytic expressions for the local truncation errors (see Aparicio, Malham, and Oliver [1]); (2) pricing path-dependent options; (3) deriving strong symplectic numerical methods (see Milstein, Repin, and Tretyakov [45]); and (4) constructing numerical methods based on the nonlinear Magnus expansions of Casas and Iserles [11].

11. Appendix. Neumann and Magnus integrators. We present Neumann and Magnus integrators up to global order $3/2$ in the case of a d-dimensional Wiener process $(W^1,\dots,W^d)$, and with constant coefficient matrices a_0 and a_i, $i=1,\dots,d$. The Neumann expansion for the solution of the stochastic differential equation (1.1) over an interval $[t_n,t_{n+1}]$, where $t_n=nh$, is

$$y^{\text{neu}}_{t_n,t_{n+1}} \approx \big(I+S_{1/2}+S_1+S_{3/2}\big)\,y_{t_n},$$

where (the indices i,j,k run over the values $1,\dots,d$)

$$\begin{aligned}
S_{1/2} &= J_0a_0+\sum_i J_ia_i+\sum_i J_{ii}a_i^2,\\
S_1 &= \sum_{i\neq j} J_{ij}a_ia_j,\\
S_{3/2} &= \sum_i (J_{i0}a_0a_i+J_{0i}a_ia_0)+\sum_{i,j,k} J_{kji}a_ia_ja_k\\
&\quad+\sum_i (J_{0ii}a_i^2a_0+J_{ii0}a_0a_i^2)+\sum_{i,j} J_{iijj}a_i^2a_j^2+J_{00}a_0^2.
\end{aligned}$$

The corresponding approximation using the Magnus expansion is

$$y^{\text{mag}}_{t_n,t_{n+1}} \approx \exp\big(s_{1/2}+s_1+s_{3/2}\big)\,y_{t_n},$$

where, with $[\cdot,\cdot]$ as the matrix commutator,

$$\begin{aligned}
s_{1/2} &= J_0a_0+\sum_i J_ia_i,\\
s_1 &= \sum_{i<j} \tfrac{1}{2}(J_{ji}-J_{ij})[a_i,a_j],\\
s_{3/2} &= \sum_i \tfrac{1}{2}(J_{i0}-J_{0i})[a_0,a_i]+\sum_{i\neq j}(J_{iij}-\tfrac{1}{2}J_iJ_{ij}+\tfrac{1}{12}J_i^2J_j)[a_i,[a_i,a_j]]\\
&\quad+\sum_{i<j<k}\big((J_{ijk}+\tfrac{1}{2}J_jJ_{ki}+\tfrac{1}{2}J_kJ_{ij}-\tfrac{2}{3}J_iJ_jJ_k)[a_i,[a_j,a_k]]\\
&\qquad+(J_{jik}+\tfrac{1}{2}J_iJ_{kj}+\tfrac{1}{2}J_kJ_{ji}-\tfrac{2}{3}J_iJ_jJ_k)[a_j,[a_i,a_k]]\big)\\
&\quad+\sum_i (J_{ii0}-\tfrac{1}{2}J_iJ_{i0}+\tfrac{1}{12}J_i^2J_0)[a_i,[a_i,a_0]].
\end{aligned}$$

To obtain a numerical scheme of global order M using the Neumann or Magnus expansion, we must use all of the terms up to and including S_M or s_M, respectively. Leading order terms of order $M+1/2$ with nonzero expectation can be replaced by their expectations (as detailed at the end of section 3). Further, extending these solution series to the nonhomogeneous and/or the nonconstant coefficient case is straightforward.

Acknowledgments. We thank Kevin and Pamela Burrage, Sandy Davie, Terry Lyons, Per-Christian Moan, Nigel Newton, Tony Shardlow, Josef Teichmann, and Michael Tretyakov for stimulating discussions. We would also like to thank the anonymous referees for their very helpful comments and suggestions.

REFERENCES

[1] N. D. Aparicio, S. J. A. Malham, and M. Oliver, *Numerical evaluation of the Evans function by Magnus integration*, BIT, 45 (2005), pp. 219–258.
[2] L. Arnold, *Stochastic Differential Equations*, John Wiley & Sons, New York, 1974.
[3] R. Azencott, *Formule de Taylor stochastique et développement asymptotique d'intégrales de Feynman*, Seminar on Probability XVI, Lecture Notes in Math., 921, Springer, Berlin, 1982, pp. 237–285.
[4] F. Baudoin, *An Introduction to the Geometry of Stochastic flows*, Imperial College Press, London, 2004.
[5] G. Ben Arous, *Flots et séries de Taylor stochastiques*, Probab. Theory Related Fields, 81 (1989), pp. 29–77.
[6] O. Bobrovnytska and M. Schweizer, *Mean-variance hedging and stochastic control: Beyond the Brownian setting*, IEEE Trans. Automat. Control, 49 (2004), pp. 396–408.
[7] K. Burrage and P. M. Burrage, *High strong order methods for non-commutative stochastic ordinary differential equation systems and the Magnus formula*, Phys. D, 133 (1999), pp. 34–48.
[8] K. Burrage, P. M. Burrage, and T. Tian, *Numerical methods for strong solutions of stochastic differential equations: An overview*, Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 460 (2004), pp. 373–402.
[9] P. M. Burrage, *Runge–Kutta methods for stochastic differential equations*, Ph.D. thesis, University of Queensland, Queensland, Australia, 1999.
[10] P. M. Burrage and K. Burrage, *A variable stepsize implementation for stochastic differential equations*, SIAM J. Sci. Comput., 24 (2002), pp. 848–864.
[11] F. Casas and A. Iserles, *Explicit Magnus expansions for nonlinear equations*, Technical report NA2005/05, DAMTP, University of Cambridge, Cambridge, UK, 2005.
[12] F. Castell, *Asymptotic expansion of stochastic flows*, Probab. Theory Related Fields, 96 (1993), pp. 225–239.
[13] F. Castell and J. Gaines, *An efficient approximation method for stochastic differential equations by means of the exponential Lie series*, Math. Comput. Simulation, 38 (1995), pp. 13–19.
[14] F. Castell and J. Gaines, *The ordinary differential equation approach to asymptotically efficient schemes for solution of stochastic differential equations*, Ann. Inst. H. Poincaré Probab. Statist., 32 (1996), pp. 231–250.
[15] K.-T. Chen, *Integration of paths, geometric invariants and a generalized Baker–Hausdorff formula*, Ann. of Mathematics, 65 (1957), pp. 163–178.
[16] J. M. C. Clark and R. J. Cameron, *The maximum rate of convergence of discrete approximations for stochastic differential equations*, in Lecture Notes in Control and Information Sci. 25, 1980, pp. 162–171.
[17] A. B. Cruzeiro, P. Malliavin, and A. Thalmaier, *Geometrization of Monte-Carlo numerical analysis of an elliptic operator: Strong approximation*, C. R. Math. Acad. Sci. Paris, 338 (2004), pp. 481–486.
[18] G. Freiling, *A survey of nonsymmetric Riccati equations*, Linear Algebra Appl., 351/352 (2002), pp. 243–270.
[19] J. G. Gaines, *The algebra of iterated stochastic integrals*, Stochastics Stochastics Rep., 49 (1994), pp. 169–179.
[20] J. G. Gaines, *A basis for iterated stochastic integrals*, Math. Comput. Simulation, 38 (1995), pp. 7–11.
[21] J. G. Gaines and T. J. Lyons, *Random generation of stochastic area integrals*, SIAM J. Appl. Math., 54 (1994), pp. 1132–1146.
[22] J. G. Gaines and T. J. Lyons, *Variable step size control in the numerical solution of stochastic differential equations*, SIAM J. Appl. Math., 57 (1997), pp. 1455–1484.
[23] I. I. Gihman and A. V. Skorohod, *The Theory of Stochastic Processes* III, Springer-Verlag, Berlin, 1979.
[24] M. B. Giles, *Multi-level Monte Carlo path simulation*, OUCL, Report no. 06/03, Oxford, UK.

[25] M. B. Giles, *Improved multilevel Monte Carlo convergence using the Milstein scheme*, OUCL, Report no. 06/22, Oxford, UK.
[26] H. Gilsing and T. Shardlow, *SDELab: Stochastic differential equations with MATLAB*, MIMS EPrint: 2006.1, ISSN 1749-9097, Manchester, UK.
[27] I. Gyöngy and G. Michaletzky, *On Wong–Zakai approximations with δ-martingales*, Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 460 (2004), pp. 309–324.
[28] D. J. Higham, *An algorithmic introduction to numerical simulation of stochastic differential equations*, SIAM Rev., 43 (2001), pp. 525–546.
[29] N. Hofmann and T. Müller-Gronbach, *On the global error of Itô–Taylor schemes for strong approximation of scalar stochastic differential equations*, J. Complexity, 20 (2004), pp. 732–752.
[30] A. Iserles, H. Z. Munthe-Kaas, S. P. Nørsett, and A. Zanna, *Lie-group methods*, Acta Numer., (2000), pp. 215–365.
[31] A. Iserles and A. Zanna, *Efficient computation of the matrix exponential by generalized polar decompositions*, SIAM J. Numer. Anal., 42 (2005), pp. 2218–2256.
[32] M. Kawski, *The combinatorics of nonlinear controllability and noncommuting flows*, ICTP Lect. Notes VIII, Trieste, Abdus Salam Int. Cent. Theoret. Phys., 2002, pp. 223–311.
[33] P. E. Kloeden and E. Platen, *Numerical Solution of Stochastic Differential Equations*, Springer-Verlag, Berlin, 1992.
[34] M. Kohlmann and S. Tang, *Multidimensional backward stochastic Riccati equations and applications*, SIAM J. Control Optim., 41 (2003), pp. 1696–1721.
[35] H. Kunita, *On the representation of solutions of stochastic differential equations*, Lecture Notes in Math. 784, Springer–Verlag, Berlin, 1980, pp. 282–304.
[36] H. Lamba, J. C. Mattingly, and A. M. Stuart, *An adaptive Euler-Maruyama scheme for SDEs: Convergence and stability*, IMA J. Numer. Anal., 27 (2007), pp. 479–506.
[37] G. J. Lord and T. Shardlow, *Postprocessing for stochastic parabolic partial differential equations*, SIAM J. Numer. Anal., 45 (2007), pp. 870–889.
[38] T. Lyons, *Differential equations driven by rough signals*, Rev. Mat. Iberoamericana, 14 (1998), pp. 215–310.
[39] T. Lyons and Z. Qian, *System Control and Rough Paths*, Oxford University Press, Oxford, UK, 2002.
[40] T. Lyons and N. Victoir, *Cubature on Wiener space*, Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 460 (2004), pp. 169–198.
[41] W. Magnus, *On the exponential solution of differential equations for a linear operator*, Comm. Pure Appl. Math., 7 (1954), pp. 649–673.
[42] S. J. A. Malham and A. Wiese, *Stochastic Lie group integrators*, SIAM J. Sci. Comput., 30 (2008), pp. 597–617.
[43] P. Malliavin, *Stochastic Analysis*, Grundlehren Math. Wiss. 313, Springer-Verlag, Berlin, 1997.
[44] G. N. Milstein, *Numerical Integration of Stochastic Differential Equations*, Mathematics and Its Applications, Kluwer Academic Publishers, Dordrecht, 1994.
[45] G. N. Milstein, Yu. M. Repin, and M. V. Tretyakov, *Numerical methods for stochastic systems preserving symplectic structure*, SIAM J. Numer. Anal., 40 (2002), pp. 1583–1604.
[46] T. Misawa, *A Lie algebraic approach to numerical integration of stochastic differential equations*, SIAM J. Sci. Comput., 23 (2001), pp. 866–890.
[47] C. Moler and C. Van Loan, *Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later*, SIAM Rev., 45 (2003), pp. 3–49.
[48] H. Z. Munthe–Kaas and W. M. Wright, *On the Hopf algebraic structure of Lie group integrators*, Found. Comput. Math., 8 (2008), pp. 227–257.
[49] N. J. Newton, *Asymptotically efficient Runge–Kutta methods for a class of Itô and Stratonovich equations*, SIAM J. Appl. Math., 51 (1991), pp. 542–567.
[50] J. Schiff and S. Shnider, *A natural approach to the numerical integration of Riccati differential equations*, SIAM J. Numer. Anal., 36 (1999), pp. 1392–1413.
[51] H. Schurz, *A brief introduction to numerical analysis of (ordinary) stochastic differential equations without tears*, in Handbook of Stochastic Analysis and Applications, V. Lakshmikantham and D. Kannan, eds., Marcel Dekker, Basel, 2002, pp. 237–359.
[52] R. B. Sidje, *Expokit: A software package for computing matrix exponentials*, ACM Trans. Math. Software, 24 (1998), pp. 130–156.
[53] E.-M. Sipiläinen, *A pathwise view of solutions of stochastic differential equations*, Ph.D. thesis, University of Edinburgh, Edinburgh, Scotland, 1993.
[54] R. S. Strichartz, *The Campbell–Baker–Hausdorff–Dynkin formula and solutions of differential equations*, J. Funct. Anal., 72 (1987), pp. 320–345.

[55] D. M. Stump and J. M. Hill, *On an infinite integral arising in the numerical integration of stochastic differential equations*, Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 461 (2005), pp. 397–413.

[56] H. J. Sussmann, *Product expansions of exponential Lie series and the discretization of stochastic differential equations*, in Stochastic Differential Systems, Stochastic Control Theory and Applications, W. Fleming and J. Lions, eds., IMA Math. Appl. Vol. 10, Springer, New York, 1988, pp. 563–582.

[57] D. Talay, *Simulation and numerical analysis of stochastic differential systems*, in Probabilistic Methods in Applied Physics, P. Kree and W. Wedig, eds., Lecture Notes in Phys. 451, Springer, Berlin, 1995, pp. 63–106.

[58] M. Wiktorsson, *Joint characteristic function and simultaneous simulation of iterated Itô integrals for multiple independent Brownian motions*, Ann. Appl. Probab., 11 (2001), pp. 470–487.

[59] E. Wong and M. Zakai, *On the relation between ordinary and stochastic differential equations*, Internat. J. Engrg Sci., 3 (1965), pp. 213–229.

[60] Y. Yamato, *Stochastic differential equations and nilpotent Lie algebras*, Z. Wahrsch. Verw. Gebiete, 47 (1979), pp. 213–229.

SIAM J. NUMER. ANAL.
Vol. 46, No. 6, pp. 2920–2952
© 2008 Society for Industrial and Applied Mathematics

CONVERGENCE OF A GENERALIZED FAST-MARCHING METHOD FOR AN EIKONAL EQUATION WITH A VELOCITY-CHANGING SIGN*

E. CARLINI†, M. FALCONE†, N. FORCADEL‡, AND R. MONNEAU‡

Abstract. We present a new fast-marching algorithm for an eikonal equation with a velocity-changing sign. This first order equation models a front propagation in the normal direction. The algorithm is an extension of the fast-marching method in two respects. The first is that the new scheme can deal with a *time-dependent* velocity, and the second is that there is *no restriction on its change in sign*. We analyze the properties of the algorithm, and we prove its convergence in the class of discontinuous viscosity solutions. Finally, we present some numerical simulations of fronts propagating in $\mathbb{R}^2$.

Key words. Hamilton–Jacobi equations, fast-marching scheme, convergence, viscosity solutions

AMS subject classifications. 65M06, 65M12, 49L25

DOI. 10.1137/06067403X

1. Introduction.

The goal of this paper is to propose and analyze a numerical scheme to compute the evolution of a front driven by its normal velocity $c(x,t)$ under very general assumptions on c. In particular, we will remove the usual assumption which assigns to c a constant sign during the evolution. This means that the front can oscillate and pass several times over the same point. The initial front is the boundary of an open set Ω_0, which is represented by a characteristic function $1_{\Omega_0} - 1_{\Omega_0^c}$, defined equal to 1 on Ω_0 and -1 on its complementary set. The use of characteristic functions instead of continuous functions has the advantage to avoid the situation when the representation function of the front has a very low derivative around its 0-level set. In fact, in that situation it can be very difficult to compute the front by the level set due to numerical errors, and this is why some (rather expensive) reinitialization techniques have been proposed (see, e.g., [18, 26]) particularly in the framework of fluid dynamics. However, it has been observed [14] that such reinitialization is obviously a disagreement between the theory of the level set method and its implementation and in many cases moves the front from its exact location. On the contrary the representation of the front by characteristics functions has always a jump at the front, and no reinitialization is needed; naturally, we pay a price in terms of accuracy (we will come back on this point in section 6). The above considerations explain why we are interested in the discontinuous viscosity solution $\theta(x,t)$ of the following equation:

$$\begin{cases} \theta_t(x,t) = c(x,t)|D\theta(x,t)| & \text{on} \quad \mathbb{R}^N \times (0,T), \\ \theta(\cdot,0) = 1_{\Omega_0} - 1_{\Omega_0^c}. \end{cases} \tag{1.1}$$

Here the support of the discontinuities of the function θ localizes the front in which we are interested. This work is motivated by the applications to dislocations dynamics

*Received by the editors November 3, 2006; accepted for publication (in revised form) March 7, 2008; published electronically August 20, 2008.
http://www.siam.org/journals/sinum/46-6/67403.html

†Dipartimento di Matematica, Università di Roma "La Sapienza," P. Aldo Moro, 2, 00185 Roma, Italy (carlini@mat.uniroma1.it, falcone@mat.uniroma1.it).

‡CERMICS, Ecole Nationale des Ponts et Chaussées, 6 et 8 avenue Blaise Pascal, Cité Descartes, Champs-sur-Marne, 77455 Marne-la-Vallée Cedex 2, France (forcadel@cermics.enpc.fr, monneau@cermics.enpc.fr).

where the velocity of the front depends on an integral term and can change sign (see Alvarez et al. [3] and Rodney et al. [19]).

A very popular method to describe the evolution of a front is the level sets method (see the seminal papers by Osher and Sethian [17] and Barles et al. [6] as well as the books [21, 22, 13]), where the discontinuous solution θ is replaced by a continuous function, and the equation is discretized by using a finite difference method with a CFL condition of the type $\Delta t\|c\|_\infty \leq \Delta x$ for explicit schemes, where Δx is the space step and Δt is the time step. In the case when the normal velocity c depends on t and can change its sign, it is absolutely necessary to solve the evolutive problem like (1.1) and track the front by its 0-level set. On the other hand, when the velocity c depends only on x and it is constant in sign (positive or negative), the evolution of the front is monotone (increasing or decreasing, respectively), and the problem can be solved via an associated stationary problem corresponding to a generalized eikonal equation (see [12] for more details and for the relations with the minimum time problem). Once the problem is reduced to the stationay eikonal equation we can use the fast-marching method (FMM) (see Sethian [23, 20]), where the unknown of the problem is the time $T(x)$ the front reaches the point x. This method works for nonnegative (nonpositive) velocities and provides a very efficient scheme which concentrates the computational effort on a neighborhood of the front. To be more precise, keeping in mind the viewpoint of discontinuous solutions, in the usual FMM we define the accepted region (A_+) as the discretization of the region $\{\theta = 1\}$ and the narrow band (NB_-) as the discretization of the boundary $\partial\{\theta = 1\}$, which is at the discrete level contained in the region $\{\theta = -1\}$. The algorithm computes the new values only at the nodes belonging to the narrow band and accepts just one of them, the one corresponding to the minimum value (see Kim [15] for a faster implementation). If c cannot change sign, we have a monotone (increasing or decreasing) evolution, and the front passes just one time on every point of the computational domain. The corresponding arrival time of the front is univalued so that the evolutive problem reduces to a stationary problem (the eikonal equation). Note that in this method there is no time step, because the time is itself the unknown of the problem so that the original evolutive problem (1.1) reduces to a stationary problem as remarked in [12, 16], i.e.,

$$|\nabla T| = \frac{1}{c(x)}. \tag{1.2}$$

To set this paper into perspective, let us recall that the FMM was initially developed for (1.2) with time-independent velocities $c(x) > 0$ (see Sethian [20] and Tsitsiklis [27]). This FMM scheme has been proved to be convergent, using a relation between the FMM solution and the numerical solution to finite difference schemes for the level sets formulation, for which it is known that these schemes are convergent (see Sethian and Vladimirsky [24] and Cristiani and Falcone [11]). More recently, the method has been extended to a more general Hamilton–Jacobi equation by Sethian and Vladimirsky [24, 25], and it has also been adapted to the case of time-dependent nonnegative velocities $c(x,t) \geq 0$ by Vladimirsky [28]. We also refer to Chopp [9] for some results for nonmonotone propagation but with time-independent velocity. However, to our knowledge, no proof of convergence has been given for the variable sign velocity case.

As we said, the goal of this paper is to propose a generalized fast-marching method (GFMM) which works for general velocities $c(x,t)$ without sign restrictions. This implies that the evolution is not necessarily monotone and that the time of arrival of the front can be multivalued. Then in our GFMM it is natural to introduce two

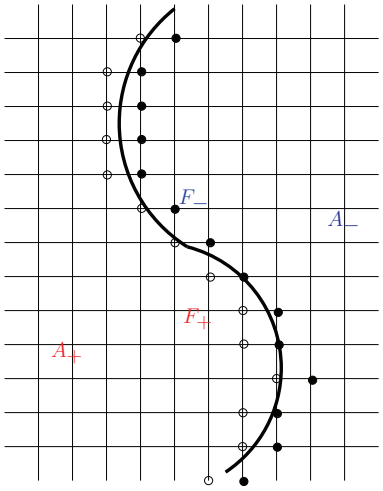


FIG. 1. *The narrow bands F_+ and F_-.*

accepted regions (A_+) and (A_-) and two narrow bands (F_+) and (F_-) in order to be able to take into account the changes of sign of the velocity. The typical picture is Figure 1. We track two fronts: one moving with positive velocity and one moving with negative velocity. A preliminary version of this new scheme has been proposed in [8]; however, in that first version no proof was given, and some small but very important details, which make the scheme work in the general case, were missing.

Our GFMM has a great potential for several future developments. Let us mention the application to dislocations dynamics where the velocity depends on the front itself via a convolution term. The corresponding level set model leads to a nonlocal Hamilton–Jacobi equation which has been analyzed in [3], a numerical approximation has been proposed in [2].

Let us underline some differences between the classical FMM method and our algorithm. One feature of the classical FMM algorithm is to generate a sequence of times t_n. Let us observe that there are several subtilities that do not appear in the usual FMM for $c(x) > 0$. These new features seem necessary to make the scheme work for general $c(x,t)$. Let us list a few of them. First, where the velocity changes sign in space, we need somehow to regularize it (see Definition 2.2) to avoid instabilities (in time) of the front. Second, because the time step is the difference $\Delta t_n = t_{n+1} - t_n$ between two computed times, we need our algorithm to ensure that this time step remains bounded from above by a given time step Δt, i.e.,

$$\Delta t_n \leq \Delta t. \tag{1.3}$$

Indeed, this is not a CFL condition. Here Δt can be much bigger than Δx divided by the maximum of the velocity. Condition (1.3) is necessary only to ensure the convergence of the scheme when Δt and Δx go to 0 (especially in the region where the velocity is very close to zero). If bound (1.3) is not respected, the algorithm may generate a sequence of time steps Δt_n nonconvergent to zero as Δx goes to zero. Third, in the classical FMM, the computed time $\tilde{t}_{n+1}$ associated to an accepted point satisfies $\tilde{t}_{n+1} > t_n$, and then we can choose as useful time $t_{n+1} = \tilde{t}_{n+1}$. One of the difficulties that we have to face in our GFMM is that it may happen that $\tilde{t}_{n+1} < t_n$ (if, for instance, the velocity is always equal to zero except at time t_n). In that case,

we cannot choose $t_{n+1} = \tilde{t}_{n+1}$, but we choose $t_{n+1} = t_n$ (see Step 8 of the algorithm). Fourth, when the front is close to a given point we have to choose carefully if we update or not the value of the time at this given point. This really depends on the position of the discrete front at time t_n and at time t_{n+1} and on the definition of the new accepted points (see Step 12 of the algorithm).

The main result of this paper is Theorem 2.5, which shows the convergence of our GFMM algorithm. When the discontinuous solution is unique, this result states that the numerical solution converges to the discontinuous viscosity solution as $\Delta x, \Delta t$ go to zero. In the case where the discontinuous viscosity solution is not unique, the result claims only that the upper semicontinuous envelope (obtained by a limsup*; see (2.7) for a definition) of the numerical solution is a discontinuous viscosity subsolution and, conversely, that the lower semicontinuous envelope (obtained by the liminf$_*$) is a supersolution.

Another novelty is the proof of convergence of this GFMM algorithm. In fact, we cannot use the relation with the usual schemes for the eikonal equation as in the case of nonnegative velocities $c(x) > 0$, and we need a *direct proof*. It is interesting to remark that, even in the case of nonnegative velocity, our proof is new. However, the idea of our proof is inspired by the paper by Barles and Georgelin [5] on fronts driven by mean curvature, where they prove convergence for a scheme in the framework of discontinuous viscosity solutions (we also refer to Barles and Souganidis [7] for convergence in the framework of continuous viscosity solutions). Basically, it is sufficient to consider a test function touching the upper semicontinuous envelope of the numerical solution (obtained as $\Delta x, \Delta t$ go to zero) which violates the subsolution property and to derive from this some properties of the discrete solution for nonzero $\Delta x, \Delta t$. This corresponds to considering test functions touching the discrete analogue of the discontinuous function θ in order to get a contradiction with the basic properties of the algorithm.

The paper is organized as follows. In section 2 we introduce our notation and present our GFMM algorithm and the main result of this paper, i.e., the convergence of the algorithm (Theorem 2.5). Section 3 is devoted to proving comparison principles and symmetry for GFMM. In section 4, several preliminary results are presented, focusing on properties of discrete times and on the geometry of the level sets of test functions. In section 5, we use the results of section 4 to prove the subsolution property of the limsup* envelope of the numerical solution, while the comparison principle of section 3 is used to prove this subsolution property at the initial time. The main result of section 5 is the proof of our main theorem (Theorem 2.5). In section 6 we present some numerical simulations and comment about these results in connection with our theoretical results.

2. The GFMM algorithm and the main result. In this section we give details for our GFMM algorithm for unsigned velocity. Let us start by introducing our definitions and notation.

Let us consider a lattice $Q \equiv \{x_I = (x_{i_1}, \dots, x_{i_N}) = (i_1 \Delta x, \dots, i_N \Delta x),\ I = (i_1, \dots, i_N) \in \mathbb{Z}^N\}$, with space step $\Delta x > 0$. We will also use a time step $\Delta t > 0$ (which does not satisfy any CFL condition).

The following definitions will be useful in what follows.

DEFINITION 2.1. *The* neighborhood of the node $I \in \mathbb{Z}^N$ *is the set*

$$V(I) \equiv \{J \in \mathbb{Z}^N : |J - I| \leq 1\}.$$

DEFINITION 2.2. *Given the speed $c_I^n \equiv c(x_I, t_n)$, we define the function*

$$\widehat{c}_I^n \equiv \begin{cases} 0 & \text{if there exists } J \in V(I) \text{ such that } (c_I^n c_J^n < 0 \text{ and } |c_I^n| \le |c_J^n|), \\ c_I^n & \text{otherwise.} \end{cases}$$

DEFINITION 2.3. *The* numerical boundary *∂E of a set $E \subset \mathbb{Z}^N$ is*

$$\partial E \equiv V(E) \backslash E,$$

with

$$V(E) = \left\{ J \in \mathbb{Z}^N, \quad \exists I \in E, \quad J \in V(I) \right\}.$$

DEFINITION 2.4. *Given a field θ_I^n with values $+1$ and -1, we define the two phases*

$$\Theta_\pm^n \equiv \{ I : \theta_I^n = \pm 1 \}$$

and the fronts $F_\pm^n \equiv \partial \Theta_\mp^n$ and $F^n \equiv F_+^n \cup F_-^n$.

In the description of the algorithm we will use the following notations:

$$\pm g \ge 0 \ \text{ for } I \in F_\pm \tag{2.1}$$

means

$$+g \ge 0 \ \text{ for } I \in F_+ \quad \text{and} \quad -g \ge 0 \text{ for } I \in F_-. \tag{2.2}$$

Moreover,

$$\min_\pm \{0, g_\pm\} \equiv \min\{0, g_+, g_-\} \quad \text{and} \quad \max_\pm \{0, g_\pm\} \equiv \max\{0, g_+, g_-\}. \tag{2.3}$$

2.1. The algorithm step by step. We describe now our GFMM algorithm for unsigned velocity. As in the classical FMM, our GFMM will generate a nondecreasing sequence of times $(t_n)_{n\in\mathbb{N}}$, with $t_0 = 0$. Contrary to the FMM, we will need a phase parameter θ_I^n with values $+1$ or -1 defined at each step n of the algorithm and for any $I \in \mathbb{Z}^N$. This θ_I^n should be thought of as a discretization of the solution θ of (1.1) at time t_n and at point x_I.

Similarly to the classical FMM, we will need to introduce a time u_I^n defined for I in the whole front F^n at the nth iteration of the algorithm. In the FMM, this time u_I^n can be interpreted as the time when the front F^n reaches the point I (this interpretation is essentially true for our GFMM algorithm except in the more delicate case where the velocity vanishes).

To write the details of the algorithm, we will need to introduce further variables:

- $\hat{u}_I^n$ corresponds to the time u_I^n at which the front passes through I, or ∞ if I is not an "upwind" point; this information is needed to compute the candidate time.
- $\tilde{u}_I^n$ is the candidate time the front may reach the point I.
- $\tilde{t}_n$ is simply the minimum of such $\tilde{u}_I^n$ for I in the front F^n. It is not a nondecreasing sequence. In general we accept $\tilde{t}_n$ as the time the front passes through at least one grid point; i.e., we have $t_n = \tilde{t}_n$, except if $\tilde{t}_n < t_{n-1}$ (nonincreasing sequence) or if $\tilde{t}_n$ is too large (speed close to zero), i.e., $\tilde{t}_n > t_{n-1} + \Delta t$ for some arbitrary (but given) time step $\Delta t > 0$.
- $\hat{t}_n$ is an intermediate time to ensure that $t_{n-1} - t_n < \Delta t$.

Initialization

1. Set $n = 1$.
2. *Initialize the field* θ^0 as
$$\theta_I^0 = \begin{cases} 1 & \text{for } x_I \in \Omega_0, \\ -1 & \text{elsewhere.} \end{cases}$$
3. *Initialize the time on* F^0
$u_I^0 = 0$ for all $I \in F^0$, $t_0 = 0$.

Main cycle

4. *Initialize* $\widehat{u}^{n-1}$ *everywhere on the grid*
$$\widehat{u}_{\pm,J}^{n-1} = \begin{cases} u_J^{n-1} & \text{for } J \in F_\pm^{n-1}, \\ \infty & \text{elsewhere.} \end{cases}$$

5. *Compute* $\tilde{u}^{n-1}$ *on* F^{n-1} *as follows:*
Let $I \in F_\pm^{n-1}$, and then
   1. if $\pm\widehat{c}_I^{n-1} \geq 0$, $\tilde{u}_I^{n-1} = \infty$;
   2. if $\pm\widehat{c}_I^{n-1} < 0$, we compute $\tilde{u}_I^{n-1}$ as the solution of the following second order equation:

$$\begin{aligned} &\sum_{k=1}^{N} \left(\max_{\pm} \left(0, \tilde{u}_I^{n-1} - \widehat{u}_{+,I^{k,\pm}}^{n-1} \right) \right)^2 = \frac{(\Delta x)^2}{|\widehat{c}_I^{n-1}|^2} \quad \text{if } I \in F_-^{n-1}, \\ &\sum_{k=1}^{N} \left(\max_{\pm} \left(0, \tilde{u}_I^{n-1} - \widehat{u}_{-,I^{k,\pm}}^{n-1} \right) \right)^2 = \frac{(\Delta x)^2}{|\widehat{c}_I^{n-1}|^2} \quad \text{if } I \in F_+^{n-1}, \end{aligned} \tag{2.4}$$

   where $I^{k,\pm} = (i_1, \ldots, i_{k-1}, i_k \pm 1, i_{k+1}, \ldots, i_N)$.
6. $\tilde{t}_n = \min\left\{ \tilde{u}_I^{n-1},\ I \in F^{n-1} \right\}$.
7. $\widehat{t}_n = \min\{\tilde{t}_n, t_{n-1} + \Delta t\}$.
8. $t_n = \max(t_{n-1}, \widehat{t}_n)$.
9. If $t_n = t_{n-1} + \Delta t$ and $t_n < \tilde{t}_n$, go to Step 4 with $n := n + 1$ and
$$\begin{cases} u_I^n = u_I^{n-1} \text{ for all } I \in F^n := F^{n-1}, \\ \theta_I^n = \theta_I^{n-1} \text{ for all } I \in \mathbb{Z}^N. \end{cases}$$

10. *Initialize the new accepted point*
$NA_\pm^n = \{I \in F_\pm^{n-1},\ \tilde{u}_I^{n-1} = \tilde{t}_n\}$, $NA^n = NA_+^n \cup NA_-^n$.
11. *Reinitialize* θ^n
$$\theta_I^n = \begin{cases} -1 & \text{for} \quad I \in NA_+^n, \\ 1 & \text{for} \quad I \in NA_-^n, \\ \theta_I^{n-1} & \text{elsewhere.} \end{cases}$$

12. *Reinitialize* u^n *on* F^n:
   1. If $I \in F^n \backslash V(NA^n)$, then $u_I^n = u_I^{n-1}$.
   2. If $I \in NA^n$, then $u_I^n = t_n$.
   3. If $I \in (F^{n-1} \cap V(NA^n)) \backslash NA^n$, then $u_I^n = u_I^{n-1}$.
   4. If $I \in V(NA^n) \backslash F^{n-1}$, then $u_I^n = t_n$.
13. Set $n := n + 1$ and go to Step 4.

Let us describe a few features of this new algorithm:

1. We know, at each time step, the time u_I^{n-1} on the fronts, i.e., on both sides of the front. This is necessary to allow the changes of the velocity sign in time.

2. In Step 5, we use the regularized velocity $\hat{c}$ and not c in order to stabilize the front. Indeed, if we do not do that, this typically leads to a duplication of the front.
3. Step 7 avoids large jumps in time and guarantees that $t_n - t_{n-1} \leq \Delta t$ with Δt small enough. Notice that this restriction on Δt is not a CFL condition but is here only to ensure that $t_n - t_{n-1} \to 0$ when we refine the mesh (i.e., decrease Δt). In particular, Δt can be chosen completely independently of Δx.
4. Step 9 allows us to increase the time. For example, if at time step n, we have $\widehat{c}_I^{n-1} = 0$ for all $I \in F^{n-1}$, then there will be no new accepted points, the time will not change, and the algorithm will be blocked without Steps 7 and 9.
5. The sequence of time $(\tilde{t}_n)$ is not necessarily nondecreasing in time. Indeed, if the velocity increases in time, we can have $\tilde{t}_n < \tilde{t}_{n-1}$, and so we have to do something to have an increasing sequence of time. In fact, Step 8 guarantees that the physical time t_n does not decrease.
6. In Step 12, for the reinitialization of u_I^n, we change its value only if a point of the neighborhood of the point I has been accepted. Moreover when u_I^n is updated, we use the physical time t_n and not $\tilde{t}_n$ or $\hat{t}_n$.

These choices, which can appear strange with respect to the classical FMM scheme, seem necessary to guarantee the convergence of our algorithm.

2.2. The main result. The scheme approximates the evolution of the fronts by a double *narrow band* and the physical time by the sequence $\{t_k, k \in \mathbb{N}\}$, defined at Step 8 in the algorithm. Such a sequence is nondecreasing, and we can extract a subsequence $\{t_{k_n}, n \in \mathbb{N}\}$ strictly increasing such that

$$t_{k_n} = t_{k_n+1} = \cdots = t_{k_{n+1}-1} < t_{k_{n+1}}.$$

We denote by S_I^n the square cell $S_I^n = [x_I, x_I + \Delta x[\times[t_{k_n}, t_{k_{n+1}}[$, with

$$[x_I, x_I + \Delta x[= \Pi_{\alpha=1}^N [x_{i_\alpha}, x_{i_\alpha} + \Delta x[,$$

and by ε the couple

$$\varepsilon = (\Delta x, \Delta t).$$

Let us define the following functions:

(2.5)
$$\theta^\varepsilon(x,t) = \begin{cases} \sup\{\theta_I^m : k_n \leq m \leq k_{n+1}-1\} & \text{if } (x,t) \in S_I^n \text{ and } c(x_I, t_{k_n}) > 0, \\ \inf\{\theta_I^m : k_n \leq m \leq k_{n+1}-1\} & \text{if } (x,t) \in S_I^n \text{ and } c(x_I, t_{k_n}) < 0, \\ \theta_I^m \ \forall m : \ k_n \leq m \leq k_{n+1}-1 & \text{if } (x,t) \in S_I^n \text{ and } c(x_I, t_{k_n}) = 0. \end{cases}$$

This definition will be useful in the proof but is indeed equivalent to the following:

$$\theta^\varepsilon(x,t) = \theta_I^{k_{n+1}-1} \text{ if } (x,t) \in S_I^n. \tag{2.6}$$

We define the half-relaxed limits

$$\overline{\theta}^0(x,t) = \text{limsup}_{\varepsilon \to 0, y \to x, s \to t} \theta^\varepsilon(y,s), \quad \underline{\theta}^0(x,t) = \text{liminf}_{\varepsilon \to 0, y \to x, s \to t} \theta^\varepsilon(y,s). \tag{2.7}$$

We make the following assumption:

(A) The velocity $c \in W^{1,\infty}(\mathbb{R}^N \times [0,T])$; for some constant $L > 0$ we have $|c(x',t') - c(x,t)| \leq L(|x'-x| + |t'-t|)$; and Ω_0 is a C^2 open set, with bounded boundary $\partial\Omega_0$.

THEOREM 2.5 (convergence result). *Under assumption* (A), $\overline{\theta}^0$ *(resp.,* $\underline{\theta}^0$*) is a viscosity subsolution (resp., supersolution) of* (1.1). *In particular, if* (1.1) *satisfies a comparison principle, then* $\overline{\theta}^0 = (\underline{\theta}^0)^*$ *and* $(\overline{\theta}^0)_* = \underline{\theta}^0$ *is the unique viscosity solution of* (1.1).

Remark 2.6. When the uniqueness holds, this is up to the upper and lower semicontinuous envelopes.

Remark 2.7. Note that when $c > 0$ our GFMM algorithm is a modified FMM algorithm where the time on the narrow band is computed by using *only* the accepted points. In this monotone case the viscosity solution of (1.1) is unique, and our result provides a convergence result (see also Test 3 in the last section).

Remark 2.8. The Lipschitz continuity in time of the velocity could be relaxed to continuity but is assumed here to simplify the presentation of the proofs, which are already quite complicated.

2.3. Computational complexity and remarks on the implementation. In this subsection, we give some rough asymptotic bounds on the computational complexity of our GFMM algorithm. Let us assume that the velocity is constant on each time interval $[k\Delta T, (k+1)\Delta T)$ for some ΔT. Of course, the velocity is not Lipschitz in time but can always be seen as the discretization of some Lipschitz velocity. We work on a (spatial) grid box of width $M^{\frac{1}{N}}$ in dimension N and therefore with a total number of grid points equal to M. We assume that the velocity is normalized $|c| \simeq 1$, and then the time T for the front to pass one time on the whole grid box is roughly $T \simeq 1$. Moreover, we normalize the space step with $\Delta x = 1$. The typical size (as a number of grid points) of the front is $M^{\frac{N-1}{N}}$ in the box. We can distinguish several cases depending on the value of ΔT.

Case 1 (constant in time velocity (i.e., $\Delta T = +\infty$)). Here the situation is very similar to the one of the classical FMM, and we can use a binary heap. Because the velocity is independent on time, we need only to recompute the value of the time at the points I whose neighbors have been accepted (i.e., $I \in V(NA^n)$). This means in practice to slightly modify point 5b of our algorithm GFMM. This implies that our GFMM is equivalent to two FMM algorithms, and so we recover the complexity in $O(M \log M)$.

Case 2 ($O(\frac{1}{M^{\frac{1}{N}}}) \leq \Delta T < +\infty$). In spirit, it is equivalent to Case 1 (indeed on the time interval $[k\Delta T, (k+1)\Delta T)$). Since the time T for the front to pass one time on the whole grid box is roughly $T \simeq 1$, we deduce that the number of time interval $[k\Delta T, (k+1)\Delta T)$ is $\frac{T}{\Delta T} = \frac{1}{\Delta T}$. On each interval $[k\Delta T, (k+1)\Delta T)$, the complexity is $M_k \log M_k$ (as in Case 1), where M_k is the number of points crossed by the front during this interval of time (with $\sum_k M_k = M$). This gives a complexity

$$\sum_{k=1}^{\frac{1}{\Delta T}} M_k \log M_k.$$

Moreover, at each $k\Delta T$, we have to recompute the candidate times and the binary heap (since the velocity changes). The complexity for these operations is

$O(M^{\frac{N-1}{N}} \log M)$. Therefore, the total complexity is then

$$\sum_{k=1}^{\frac{1}{\Delta T}} M_k \log M_k + \sum_{k=1}^{\frac{1}{\Delta T}} M^{\frac{N-1}{N}} \log M \leq M \log M + \frac{1}{\Delta T} M^{\frac{N-1}{N}} \log M.$$

If $\Delta T \geq \frac{1}{M^{\frac{1}{N}}}$, we then get a complexity in $O(M \log M)$ (as in the classical case).

Case 3 (variable velocity (i.e., $\Delta T = 0$) or $0 < \Delta T < O(\frac{1}{M^{\frac{1}{N}}})$). From a complexity point of view, it is not interesting to use a binary heap to sort the time of the points on the front. Because the velocity changes at each step n of the algorithm, it is much more efficient to recompute all of the times on the front and extract the minimum of these times at each iteration. The complexity for these operations is $O(M^{\frac{N-1}{N}})$. On the time necessary for the front to pass one time on the whole grid box, we need to do this computation $O(M)$ times. Therefore, the total complexity is $O(M^{\frac{2N-1}{N}})$.

In this case, it seems that the narrow band level set method can be more interesting from the complexity point of view (see [1]) although it is rather difficult to make a precise statement on this point.

We can then implement our algorithm depending on the variability of the velocity in time (see Cases 1, 2, and 3 above). Finally, we want to point out that Case 2, where ΔT is not so small, is acceptable in practice if the velocity is Lipschitz in time with a reasonable Lipschitz constant (and this is what we assume theoretically in our convergence theorem).

3. Comparison principles for the GFMM algorithm. As we said in the introduction, our convergence result will be proved in the framework of discontinuous viscosity solutions. To this end the role of comparison principles is crucial.

In this section, we first present a property of symmetry of the algorithm and then present some comparison principles in some special cases.

3.1. Symmetry of the algorithm. The following lemma claims that if we change the sign of the velocity and the sign of the two phases at the initial time, then the GFMM algorithm computes the same front.

LEMMA 3.1 (symmetry of the GFMM algorithm). *We denote by* $\overline{\theta}^0[\theta^0, c]$ *and* $\underline{\theta}^0[\theta^0, c]$ *the functions constructed by the GFMM algorithm with initial condition* θ^0 *and velocity* c. *Then we have*

$$\underline{\theta}^0[\theta^0, c] = -\overline{\theta}^0[-\theta^0, -c].$$

Proof of Lemma 3.1. With the same kind of notation, we remark that $\theta_I^n[-\theta^0, -c] = -\theta_I^n[\theta^0, c]$. We then have, for $x \in [x_I, x_I + \Delta x[$, $t \in [t_{k_n}, t_{k_{n+1}}[$, and $c(x_I, t_{k_n}) > 0$,

$$\begin{aligned}\theta^\varepsilon[\theta^0, c](x,t) &= \sup\{\theta_I^k[\theta^0, c],\ k_n \leq k \leq k_{n+1} - 1\} \\ &= -\inf\{-\theta_I^k[\theta^0, c],\ k_n \leq k \leq k_{n+1} - 1\} \\ &= -\inf\{\theta_I^k[-\theta^0, -c],\ k_n \leq k \leq k_{n+1} - 1\} = -\theta^\varepsilon[-\theta^0, -c].\end{aligned}$$

The result is similar for $c(x_I, t_{k_n}) \leq 0$. Therefore, $\underline{\theta}^0[\theta^0, c] = -\overline{\theta}^0[-\theta^0, -c]$. ∎

3.2. Comparison principles.

PROPOSITION 3.2 (comparison principle for the time). *We denote by* u_I^n *(resp.,* v_I^n*) the numerical solution at the point* (x_I, t_n) *of the GFMM algorithm with velocity*

c_u (resp., c_v). We assume that there exists $T > 0$ such that for all $(x,t) \in \mathbb{R}^N \times [0,T]$

$$\inf_{s\in[t-\Delta t,t],\ s\geq 0} c_v(x,s) \geq \sup_{s\in[t-\Delta t,t],\ s\geq 0} (c_u(x,s))^+,$$

where $(f)^+$ is the positive part of f. We assume that

$$\{\theta_u^0 = 1\} \subset \{\theta_v^0 = 1\} \quad \text{and} \quad v^0 = u^0 = 0.$$

We define $\bar{m}$ and $\bar{k}$ such that

$$\begin{cases} t_{\bar{m}} \leq T < t_{\bar{m}+1}, \\ s_{\bar{k}} \leq T < s_{\bar{k}+1}, \end{cases}$$

where $(t_m)_m$ and $(s_m)_m$ are the sequence of time constructed by the GFMM algorithm with velocity c_u and c_v, respectively. We then consider

$$v_I = \begin{cases} v_I^0 & \text{if} \quad \theta_{v,I}^0 = 1, \\ v_I^k & \text{if} \quad I \in NA_v^k \quad \text{for some} \quad k \leq \bar{k}+1, \\ s_{\bar{k}+1} & \text{if} \quad \theta_{v,I}^{\bar{k}} = -1. \end{cases}$$

Then, $\forall l \leq \bar{m}$, $\forall I \in NA_u^l$, we have $v_I \leq u_I^l$.

Remark 3.3. Here the notations for $\theta_u, \theta_v, NA_u^l$, and further notation in what follows are obvious and are not explained. Moreover we also remark that the front for v passes at most one time at a given point because $c_v \geq 0$.

Proof of Proposition 3.2. We argue by contradiction. We denote by $m(u)$ the first index such that there exists $I \in NA_u^{m(u)}$ such that

$$u_I^{m(u)} < v_I. \tag{3.1}$$

We define $k(v)$ such that $I \in NA_v^{k(v)}$ with the convention that $k(v) = \bar{k}+1$ if $\theta_{v,I}^{\bar{k}} = -1$. This implies that $t_{m(u)} = u_I^{m(u)} < v_I = s_{k(v)}$. The proof distinguishes two cases:

1. $I \in NA_{-,u}^{m(u)} \subset F_{-,u}^{m(u)-1}$. We claim that, for all $J \in V(I)\setminus\{I\}$, we have

$$\widehat{u}_{+,J}^{m(u)-1} \geq \widehat{v}_{+,J}^{k(v)-1}. \tag{3.2}$$

Indeed assume that $\widehat{u}_{+,J}^{m(u)-1} < \infty$ (if $\widehat{u}_{+,J}^{m(u)-1} = \infty$, then (3.2) holds); then $J \in F_{+,u}^{m(u)-1}$, and we have

$$t_{m(u)} \geq \widehat{u}_{+,J}^{m(u)-1} \geq v_J.$$

It just remains to show that $v_J = \widehat{v}_{+,J}^{k(v)-1}$. We argue by contradiction. Assume that $\widehat{v}_{+,J}^{k(v)-1} = \infty$, i.e., $J \in \{\theta_v^{k(v)-1} = -1\}$. Then $v_J \geq s_{k(v)}$. This contradicts the fact that $v_J \leq t_{m(u)} < s_{k(v)}$ and proves (3.2).

We define

$$k^* := \sup\{k, s_k \leq t_{m(u)}\} < k(v).$$

In particular, we have $t_{m(u)} - \Delta t \leq s_{k^*} \leq t_{m(u)}$. Since for all $J \in V(I) \cap F_{+,u}^{m(u)-1}$

$$t_{m(u)} \geq \widehat{u}_{+,J}^{m(u)-1} \geq \widehat{v}_{+,J}^{k(v)-1},$$

we deduce that

$$s_{k^*} \geq \widehat{v}_{+,J}^{k(v)-1}. \tag{3.3}$$

Indeed, $+\infty > \widehat{v}_{+,J}^{k(v)-1} > s_{k^*}$ would imply that there exists $k' > k^*$ such that $t_{m(u)} \geq \widehat{v}_{+,J}^{k(v)-1} = s_{k'}$, which contradicts the definition of k^*.

Then we claim that for all $J \in V(I) \cap F_{+,u}^{m(u)-1}$

$$\widehat{v}_{+,J}^{k(v)-1} = \widehat{v}_{+,J}^{k^*}. \tag{3.4}$$

We now prove the claim (3.4). First, because we have $\widehat{v}_{+,J}^{k(v)-1} < +\infty$, we deduce that $\theta_{v,J}^{k(v)-1} = 1$; then there exists $k \leq k(v)-1$ such that, if $k \geq 1$, then $J \in NA_v^k$ and $\widehat{v}_{+,J}^{k(v)-1} = v_J^k = s_k$, and, if $k = 0$, then $\theta_{v,J}^0 = 1$ and $\widehat{v}_{+,J}^{k(v)-1} = v_J^0 = 0$.

Assume by contradiction that $k > k^*$. Then

$$\widehat{v}_{+,J}^{k(v)-1} = v_J^k = s_k \geq s_{k^*+1} > s_{k^*}$$

in contradiction with (3.3). Therefore $k \leq k^*$. Now we have $\theta_{v,I}^{k(v)} = 1$ and $\theta_{v,I}^m = -1$ for $m \leq k(v)-1$. Therefore $J \in F_{+,v}^{k^*}$ and

$$\widehat{v}_{+,J}^{k(v)-1} = v_J^k = \widehat{v}_{+,J}^{k^*},$$

which ends the proof of the claim (3.4). We deduce that

$$\widehat{v}_{+,J}^{k(v)-1} = \widehat{v}_{+,J}^{k^*} \leq \widehat{u}_{+,J}^{m(u)-1},$$

where we have used (3.2). We define the following function:

$$f_{\widehat{u}^m}^2(t) = \sum_{k=1}^{N} \left(\max_{\pm} \left(0, t - \widehat{u}_{+,I^{k,\pm}}^m \right) \right)^2.$$

We then have, by using the fact that $\tilde{v}_I^{k^*} \geq s_{k^*+1} > s_{k^*}$,

$$f_{\widehat{v}^{k^*}}(s_{k^*+1}) \leq f_{\widehat{v}^{k^*}}(\tilde{v}_I^{k^*}) = \left| \frac{\Delta x}{\widehat{c}_{I,v}^{k^*}} \right| \leq \left| \frac{\Delta x}{\widehat{c}_{I,u}^{m(u)-1}} \right| = f_{\widehat{u}^{m(u)-1}}(\tilde{u}_I^{m(u)-1}) \leq f_{\widehat{v}^{k^*}}(\tilde{u}_I^{m(u)-1}).$$

We then deduce that $s_{k^*+1} \leq \tilde{u}_I^{m(u)-1} \leq u_I^{m(u)} = t_{m(u)}$. This is absurd.

2. $I \in NA_{+,u}^{m(u)} \subset F_{+,u}^{m(u)-1}$. We consider the following subcases:
   1. $I \in \{\theta_v^0 = 1\}$. Then $v_I = v_I^0 = 0 = u_I^0 \leq u_I^{m(u)}$. This is absurd.
   2. $I \in \{\theta_v^0 = -1\}$. Then $\theta_{u,I}^0 = -1$, and so there exists $n < m(u)$ such that

$$\theta_{u,I}^{n-1} = -1 \quad \text{and} \quad \theta_{u,I}^n = 1.$$

This implies that $u_I^n \geq v_I > u_I^{m(u)} \geq u_I^n$. This is absurd. ∎

Remark 3.4. If we imply the computation of the gradient, i.e., the computation of $\tilde{u}$ in step 5, the situation seems better, and one could expect to prove a general comparison principle without restriction on the velocity.

We now rephrase this comparison principle for the functions θ^ε and prove it.

COROLLARY 3.5 (comparison principle with a nonnegative velocity). *Under the assumptions of Proposition* 3.2*, we have for all* $(x,t) \in \mathbb{R}^N \times [0,T]$

$$\theta_u^\varepsilon(x,t) \le \theta_v^\varepsilon(x,t).$$

Proof of Corollary 3.5. By contradiction, assume that there exist x_I and t such that

$$\theta_u^\varepsilon(x_I,t) = 1 \quad \text{and} \quad \theta_v^\varepsilon(x_I,t) = -1. \tag{3.5}$$

We denote by t the first time such that (3.5) holds. We then have, since $c_v \ge 0$,

$$\theta_u^\varepsilon(x_I,s) = -1 \quad \text{if} \quad s < t.$$

We then deduce that there exists $m(u)$ such that $t = t_{m(u)}$, $I \in NA_u^{m(u)}$, and $u_I^{m(u)} = t_{m(u)} = t$. Moreover, since the index I has not been already accepted for v, we have $v_I > t = u_I^{m(u)}$. This is absurd. ∎

COROLLARY 3.6 (comparison principle for a nonpositive velocity). *We denote by* u_I^n *(resp.,* v_I^n*) the numerical solution at the point* (x_I,t_n) *of the GFMM algorithm with velocity* c_u *(resp.,* c_v*). We assume that there exists* $T > 0$ *such that for all* $(x,t) \in \mathbb{R}^N \times [0,T]$

$$\sup_{s\in[t-\Delta t,t],\ s\ge 0} c_u(x,s) \le \inf_{s\in[t-\Delta t,t],\ s\ge 0} -(c_v(x,s))^-,$$

where $(f)^- \ge 0$ *is the negative part of* f*. We assume that*

$$\{\theta_v^0 = -1\} \subset \{\theta_u^0 = -1\} \quad \text{and} \quad v^0 = u^0 = 0.$$

Then, for all $(x,t) \in \mathbb{R}^N \times [0,T]$*, we have*

$$\theta_u^\varepsilon(x,t) \le \theta_v^\varepsilon(x,t).$$

Proof of Corollary 3.6. This is a straightforward consequence of Corollary 3.5 and the fact that $\theta^\varepsilon[-\theta^0,-c] = -\theta^\varepsilon[\theta^0,c]$ (with the notation of Lemma 3.1). ∎

4. Preliminary results on the discrete time and on the level sets of test functions. The GFMM algorithm described in section 2 has several properties which fit the physics of the problem that we want to solve. We present in this section several results that will be used in the proof of Proposition 5.1, which is crucial for the proof of our main result of convergence.

In a first subsection, we present some properties of the various times $\hat{u}, t, \tilde{t}$ appearing in our algorithm, and in a second subsection we give some geometrical consequences of the existence of test functions tangent from above to our function θ^ε.

4.1. Preliminary results on the discrete time.

LEMMA 4.1 (time character of the $\widehat{u}$). *Assume that there exists* $\delta > 0$ *and* $(I,n) \in \mathbb{Z}^N \times \mathbb{N}$ *such that* $c(x_I,t_n) \ge \delta > 0$*,* $\theta_I^{n-1} = -1$*, and* $\theta_I^n = 1$ *(resp.,* $c(x_I,t_n) \le -\delta < 0$*,* $\theta_I^{n-1} = 1$*; and* $\theta_I^n = -1$*) then for any* $J \in V(I) \cap F_+^{n-1}$ *(resp.,* $J \in V(I) \cap F_-^{n-1}$*), we have for* $\Delta x \le \frac{\delta^2}{16L}$

$$\widehat{u}_{+,J}^{n-1} = \sup\{t_m \le t_{n-1},\ \theta_J^{m-1} = -1, \theta_J^p = 1 \quad \textit{for} \quad m \le p \le n-1\} > t_n - \frac{4\Delta x}{\delta},$$

with the convention that $\widehat{u}^{n-1}_{+,J} = 0$ *if* $\theta^p_J = 1$ *for* $0 \le p \le n-1$

(*resp.,* $\widehat{u}^{n-1}_{-,J} = \sup\{t_m \le t_{n-1},\ \theta^{m-1}_J = 1, \theta^p_J = -1 \quad for \quad m \le p \le n-1\} > t_n - \dfrac{4\Delta x}{\delta}$,

with the convention that $\widehat{u}^{n-1}_{+,J} = 0$ *if* $\theta^p_J = -1$ *for* $0 \le p \le n-1$*).*

This lemma claims in fact that the $\widehat{u}^{n-1}_{+,J}$ is defined as the last time at which the front passed through J. Intuitively, this comes from the fact that, since the velocity is locally nonnegative and since the front has crossed the node x_I at time t_n, it has crossed the node x_J at a time closed to t_n.

Proof of Lemma 4.1. We do only the proof in the case $c > 0$ (the case $c < 0$ is similar). By assumption, c is Lipschitz-continuous with constant L, and there exists $\delta_0 \le \delta/(4L)$ such that, for all $(x_J, t_m) \in B_{\delta_0}(x_I) \times [t_n - \delta_0, t_n + \delta_0]$, we have

$$\widehat{c}^m_J \ge \frac{\delta}{2}.$$

This implies that

$$\theta^m_I = -1 \text{ for all } m \text{ such that } \ t_n - \delta_0 \le t_m \le t_{n-1}. \tag{4.1}$$

Let $J \in V(I) \cap F^{n-1}_+$. We define $m_J = \sup\{m \le n-1,\ \theta^{m-1}_J = -1, \theta^m_J = 1\}$. We claim that, for all $J \in V(I) \cap F^{n-1}_+$, we have $t_{m_J} > t_n - \delta_0$ for Δx small enough. Indeed, by contradiction, assume that there exists $J \in V(I) \cap F^{n-1}_+$ such that $t_{m_J} \le t_n - \delta_0$. Let us define $p \ge 0$ such that $t_n = \cdots = t_{n-p} > t_{n-p-1}$. We then have $\widehat{u}^{n-p-1}_{+,J} \le t_n - \delta_0$ and $\tilde{u}^{n-p-1}_I \ge t_{n-p} = t_n$. By using the fact that

$$\sum_{k=1}^{N} \left(\max_{\pm} \left(0, \tilde{u}^{n-p-1}_I - \widehat{u}^{n-p-1}_{+,I^{k,\pm}}\right)\right)^2 = \left(\frac{\Delta x}{\widehat{c}^{n-p-1}_I}\right)^2,$$

we then deduce that

$$\delta_0 = t_{n-p} - (t_n - \delta_0) \le \tilde{u}^{n-p-1}_I - \widehat{u}^{n-p-1}_J \le \frac{2\Delta x}{\delta}.$$

This is absurd for the choice $\delta_0 = \frac{4\Delta x}{\delta} \le \frac{\delta}{4L}$ which is valid for Δx small enough. Moreover, by using (4.1), we deduce that $J \in F^m$ for all $m_J \le m \le n-1$. This implies that $\widehat{u}^{n-1}_{+,J} = u^{n-1}_J = u^{m_J}_J = t_{m_J}$. ∎

The following lemma is concerned with the fact that we can control the decay of the time $\tilde{t}_n$ given by the GFMM algorithm, by the variations in time of the velocity.

LEMMA 4.2 (error estimate between t_n and $\tilde{t}_n$). *Assume that there exists* $I \in NA^n$ *such that* $|\widehat{c}^{n-1}_I| \ge \delta > 0$. *Then the following estimate holds:*

$$\left(t_n - \tilde{t}_n\right)^+ \le \frac{2L}{\delta^2}\Delta x \Delta t \quad \textit{if} \quad \Delta t \le \frac{\delta}{2L}.$$

Proof of Lemma 4.2. We treat only the case $c^{n-1}_I \ge \delta > 0$ (the other case is similar). Assume that $\tilde{t}_n < t_n$, and then necessarily $t_n = t_{n-1}$. We define $p > 0$ such that $t_{n-p-1} < t_{n-p} = \cdots = t_{n-1} = t_n$. In particular, we have

$$t_{n-p} \le \tilde{t}_{n-p} \le \tilde{u}^{n-p-1}_J \quad \forall J \in F^{n-p-1}$$

and

$$\tilde{t}_n = \tilde{u}_I^{n-1} \le \tilde{u}_J^{n-1} \quad \forall J \in F^{n-1}.$$

We claim that $I \in F_-^{n-p-1}$. Indeed, assume that $I \notin F_-^{n-p-1}$. By using the fact that $\theta_I^{n-p-1} = -1$ (since $\widehat{c}_I > 0$), we deduce that, for all $J \in V(I) \cap F_+^{n-1}$, we have $\theta_J^{n-p-1} = -1$ and so $\widehat{u}_{+,J}^{n-1} = t_n$; this means that also the node J has been accepted at the physical time t_n . This implies that $\tilde{u}_I^{n-1} > t_n$, and this is absurd.

Moreover, because $t_{n-p} - t_{n-p-1} \le \Delta t$, we have $\widehat{c}_I^{n-p-1} \ge \frac{\delta}{2}$ for $\Delta t \le \frac{\delta}{2L}$. We then have

$$\sum_{k=1}^{N} \left(\max_{\pm} \left(0, \tilde{u}_I^{n-p-1} - \widehat{u}_{+,I^{k,\pm}}^{n-p-1} \right) \right)^2 = \left(\frac{\Delta x}{\widehat{c}_I^{n-p-1}} \right)^2 \tag{4.2}$$

and

$$\sum_{k=1}^{N} \left(\max_{\pm} \left(0, \tilde{u}_I^{n-1} - \widehat{u}_{+,I^{k,\pm}}^{n-1} \right) \right)^2 = \left(\frac{\Delta x}{\widehat{c}_I^{n-1}} \right)^2. \tag{4.3}$$

Let us compare $\widehat{u}_{+,J}^{n-1}$ and $\widehat{u}_{+,J}^{n-p-1}$ for $J \in V(I) \cap F_+^{n-1}$. If $J \notin F_+^{n-p-1}$, then u_J changes values during the iterations $n-p \le m \le n-1$, and for such m we have $\widehat{u}_{+,J}^{n-1} = u_J^m = t_m = t_n$. Since $\tilde{t}_n < t_n$, then this node $J \in V(I)$ does not contribute to the evaluation of (4.3) and

$$\sum_{k=1}^{N} \left(\max_{\pm} \left(0, \tilde{t}_n - \widehat{u}_{+,I^{k,\pm}}^{n-p-1} \right) \right)^2 = \sum_{k=1}^{N} \left(\max_{\pm} \left(0, \tilde{t}_n - \widehat{u}_{+,I^{k,\pm}}^{n-1} \right) \right)^2. \tag{4.4}$$

Let us denote that

$$f_q(v) = \left\{ \sum_{k=1}^{N} \left(\max_{\pm} \left(0, v - \widehat{u}_{+,I^{k,\pm}}^{q} \right) \right)^2 \right\}^{1/2}.$$

The function f_q verifies for any $q \in \mathbb{N}$ such that $I \in F_-^q$:

$$f_q(\tilde{u}_I^q) = \frac{\Delta x}{|\widehat{c}_I^q|}, \quad f_q'(v) \ge 1.$$

Then

$$\begin{aligned} t_n - \tilde{t}_n &\le \tilde{u}_I^{n-p-1} - \tilde{t}_n \le f_{n-p-1}(\tilde{u}_I^{n-p-1}) - f_{n-p-1}(\tilde{t}_n) \\ &= f_{n-p-1}(\tilde{u}_I^{n-p-1}) - f_{n-1}(\tilde{t}_n) = \Delta x \left(\frac{1}{|\widehat{c}_I^{n-p-1}|} - \frac{1}{|\widehat{c}_I^{n-1}|} \right) \\ &\le \Delta x \frac{|\widehat{c}_I^{n-p-1} - \widehat{c}_I^{n-p}|}{|\widehat{c}_I^{n-p-1}||\widehat{c}_I^{n-p}|} \le \frac{\Delta x |\partial_t c|_{L^\infty} |t_{n-p} - t_{n-p-1}|}{|\widehat{c}_I^{n-p-1}||\widehat{c}_I^{n-p}|} \le \frac{2\Delta x |\partial_t c|_{L^\infty} \Delta t}{\delta^2}. \end{aligned}$$

∎

4.2. Preliminary results on the level sets of test functions. In this subsection, we give some geometric results concerning the test function for θ^ε (see Figures 2 and 3).

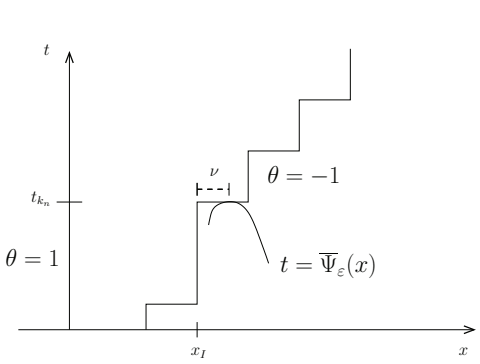


FIG. 2. *Test function from below.*

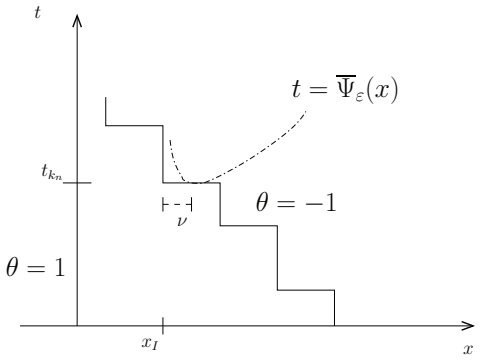


FIG. 3. *Test function from above.*

LEMMA 4.3 (separation of the phases of θ^ε by the level set of a test function). *Let $\varphi \in C^2$ in a neighborhood V of (x_0, t_0) such that $\varphi_t(x_0, t_0) > 0$ (resp., $\varphi_t(x_0, t_0) < 0$). There exist $\delta_0 > 0$, $r > 0$, $\tau > 0$ such that, if $\max_{\overline{V}}((\theta^\varepsilon)^* - \varphi)$ is reached at $(x_\varepsilon, t_\varepsilon) \in B_{\delta_0}(x_0, t_0) \subset V$ with $(\theta^\varepsilon)^*(x_\varepsilon, t_\varepsilon) = 1$, then there exists $\Psi_\varepsilon \in C^2(B_r(x_0), (t_0 - \tau, t_0 + \tau))$ such that*

(i) *for all $(x_J, t_m) \in Q_{r,\tau}(x_0, t_0) = B_r(x_0) \times (t_0 - \tau, t_0 + \tau)$*

$$\theta^\varepsilon(x_J, t_m) = 1 \quad \Longrightarrow \quad t_m \geq \Psi_\varepsilon(x_J) \text{ (resp., } t_m \leq \Psi_\varepsilon(x_J)).$$

(ii) *There exists $(I, n) \in \mathbb{Z}^N \times \mathbb{N}$ such that*

$$(x_\varepsilon, t_\varepsilon) \in \bar{Q}_I^n = [x_I, x_I + \Delta x] \times [t_{k_n}, t_{k_{n+1}}], \quad (\theta^\varepsilon)^*(x_I, t_{k_n}) = 1, \quad t_{k_n} = \Psi_\varepsilon(x_I),$$

and $\theta_I^{\overline{n}} = 1$, $\theta_I^m = -1$, $m_0 \leq m \leq \overline{n} - 1$ (resp., $\theta_I^{\overline{n}} = -1$, $\theta_I^m = 1$, $m_0 \leq m \leq \overline{n} - 1$), where $\overline{n} = \inf\{k, k_n \leq k \leq k_{n+1} - 1, \theta_I^k = 1$ (resp., $\theta_I^k = -1)\}$ and $m_0 = \inf\{m,\ t_m \geq t_0 - \tau\}$.

(iii) *The following Taylor expansion holds:*

$$\Psi_\varepsilon(x_J) = \Psi_\varepsilon(x_I) - \frac{D\varphi(x_0, t_0)}{\varphi_t(x_0, t_0)}(x_J - x_I) + (\Delta x)\, O\left(\Delta x + |x_I - x_0| + |t_{k_n} - t_0|\right).$$

(iv) *If $\varphi_t(x_0, t_0) < 0$, then for all $(x_J, t_{k_n}) \in Q_{r,\tau}(x_0, t_0) = B_r(x_0) \times (t_0 - \tau, t_0 + \tau)$*

$$\theta^\varepsilon(x_J, t_{k_{n-1}}) = 1 \quad \text{and} \quad \theta^\varepsilon(x_J, t_{k_n}) = -1 \quad \Longrightarrow \quad t_{k_n} \leq \Psi_\varepsilon(x_J).$$

Proof of Lemma 4.3. We consider the case $\varphi_t(x_0, t_0) > 0$. The other case can be treated in a similar way. We define $\varphi_\varepsilon = \varphi + ((\theta^\varepsilon)^* - \varphi)(x_\varepsilon, t_\varepsilon)$. In particular, we have $(\theta^\varepsilon)^* \leq \varphi_\varepsilon$ and $(\theta^\varepsilon)^*(x_\varepsilon, t_\varepsilon) = \varphi_\varepsilon(x_\varepsilon, t_\varepsilon) = 1$. We start by proving (i) and (ii). The proof is decomposed in several steps.

Step 1. We have $t_\varepsilon = t_{k_n}$.

Indeed, assume that $t_\varepsilon \in (t_{k_n}, t_{k_{n+1}})$. By using the fact that $(\theta^\varepsilon)^*(x_\varepsilon, t_\varepsilon) = 1$, we deduce that $(\theta^\varepsilon)^*(x_\varepsilon, t) = 1$ for $t_{k_n} \leq t \leq t_{k_{n+1}}$ and so $\varphi_t(x_\varepsilon, t_\varepsilon) = 0$. This is absurd for δ_0 small enough since $\varphi_t(x_0, t_0) > 0$.

Step 2. We have $(\theta^\varepsilon)^* = -1$ on all $Q_J^{n-1} =]x_J, x_J + \Delta x[\times]t_{k_{n-1}}, t_{k_n}[$ such that $(x_\varepsilon, t_{k_n}) \in \bar{Q}_J^{n-1}$.

Indeed, since $\varphi_\varepsilon(x_\varepsilon, t_{k_n}) = 1$ and $(\varphi_\varepsilon)_t > 0$, we deduce that $\varphi_\varepsilon(x_\varepsilon, t) < 1$ if $t < t_{k_n}$. Using the fact that $(\theta^\varepsilon)^* - \varphi_\varepsilon$ reaches a maximum in (x_ε, t_{k_n}) yields

$$(\theta^\varepsilon)^*(x_\varepsilon, t) \leq \varphi_\varepsilon(x_\varepsilon, t) < 1 \quad \text{if } t < t_{k_n}$$

and so

$$(\theta^\varepsilon)^*(x_\varepsilon, t) = -1 \quad \text{if } t < t_{k_n}.$$

By using the semicontinuity of $(\theta^\varepsilon)^*$, one deduce that

$$(\theta^\varepsilon)^* = -1 \quad \text{on all } Q_J^{n-1} \text{ such that } (x_\varepsilon, t_{k_n}) \in \bar{Q}_J^{n-1}.$$

Step 3. There exists $I \in \mathbb{Z}^N$ such that $(x_\varepsilon, t_{k_n}) \in \bar{Q}_I^n$ and $(\theta^\varepsilon)^* = 1$ on Q_I^n.

By contradiction, assume that, on all cubes Q_J^n such that $(x_\varepsilon, t_{k_n}) \in \bar{Q}_J^n$, we have $(\theta^\varepsilon)^* = -1$. Then, by using Step 2, we deduce that $(\theta^\varepsilon)^* = -1$ in a neighborhood of (x_ε, t_{k_n}). This is absurd since $(\theta^\varepsilon)^*(x_\varepsilon, t_{k_n}) = 1$.

Before continuing the proof, we need a few notations. We set

$$\overline{n} = \inf\{k, \quad k_n \le k \le k_{n+1} - 1,\ \theta_I^k = 1\}.$$

In particular, we have $\theta_I^{\overline{n}} = 1$ and $\theta_I^{\overline{n}-1} = -1$.

Since $(\varphi_\varepsilon)_t(x_\varepsilon, t_{k_n}) > 0$ for ε small enough, by the implicit function theorem, there exists a neighborhood V_ε of $(x_\varepsilon, t_\varepsilon)$ and a function $\bar{\Psi}_\varepsilon$ such that

$$\{\varphi_\varepsilon(x,t) < 1\} \Leftrightarrow \{t < \bar{\Psi}_\varepsilon(x)\}$$

in V_ε. Using the fact that $(\theta^\varepsilon)^* \le \varphi_\varepsilon$ yields

$$\{(\theta^\varepsilon)^* = 1\} \subset \{t \ge \bar{\Psi}_\varepsilon(x)\}. \tag{4.5}$$

Moreover, for δ_0 small enough, i.e., for $(x_\varepsilon, t_\varepsilon)$ closed enough to (x_0, t_0), we can assume that $V_\varepsilon \supset Q_{r,\tau}(x_0, t_0) = B_r(x_0) \times (t_0 - \tau, t_0 + \tau)$. We define $\nu = x_\varepsilon - x_I \in [0, \Delta x)^N$ and $\Psi_\varepsilon(x) = \bar{\Psi}_\varepsilon(x + \nu)$. In particular, we have $\Psi_\varepsilon(x_I) = t_{k_n}$.

Step 4. For all $(x_J, t_{k_m}) \in Q_{r,\tau}(x_0, t_0) = B_r(x_0) \times (t_0 - \tau, t_0 + \tau)$

$$\theta^\varepsilon(x_J, t_{k_m}) = 1 \quad \Longrightarrow \quad t_{k_m} \ge \Psi_\varepsilon(x_J).$$

To prove this, we consider the collection of nodes

$$\mathcal{C} = \{(x_J, t_{k_m}) \in Q_{r,\tau}(x_0, t_0) \cap \{\theta^\varepsilon = 1\}\}.$$

By inclusion (4.5), we then have $t_{k_m} \ge \bar{\Psi}_\varepsilon(x_J)\ \forall (x_J, t_{k_m}) \in \mathcal{C}$. We deduce $\forall (x_J, t_{k_m}) \in \mathcal{C}$ that

$$\bar{Q}_J^m = [x_J, x_J + \Delta x] \times [t_{k_m}, t_{k_{m+1}}] \subset \{t \ge \bar{\Psi}_\varepsilon(x)\}.$$

This implies that

$$(x_J + (x_\varepsilon - x_I), t_{k_m}) \in \{t \ge \bar{\Psi}_\varepsilon(x)\}$$

and so

$$(x_J, t_{k_m}) \in \{t \ge \Psi_\varepsilon(x)\},$$

which implies (i) because any $t_{m'}$ can be written t_{k_m} for a suitable m.

Step 5. We have $\theta_I^m = -1$ for $m_0 \le m \le \overline{n} - 1$, where $m_0 = \inf\{m,\ t_m \ge t_0 - \tau\}$.

By contradiction, suppose that there exists $m_0 \le m \le \overline{n} - 1$ such that $\theta_I^m = 1$. We then define m_1 as

$$m_1 = \sup\{m \le \overline{n} - 1,\ \theta_I^m = 1\}.$$

In particular, we have $\theta_I^{m_1+1} = -1$ (since $\theta_I^{\overline{n}-1} = -1$). Two cases may occur:

1. $t_{m_1} = t_{k_n} = t_{\overline{n}}$. In this case, we have $\widehat{c}_I^{m_1} = \widehat{c}_I^{\overline{n}-1} > 0$ (since $\theta_I^{\overline{n}-1} = -1$ and $\theta_I^{\overline{n}} = 1$). This contradicts the fact that $\theta_I^{m_1} = 1$ and $\theta_I^{m_1+1} = -1$.
2. $t_{m_1} < t_{k_n} = t_{\overline{n}}$. In this case, we have $\theta^\varepsilon(x_I, t_{m_1}) = 1$ and $t_{m_1} < t_{k_n} = \Psi_\varepsilon(x_I)$. This contradicts Step 4.

We now prove (iii).

By the implicit functions theorem, we have $\varphi_\varepsilon(x, \bar{\Psi}_\varepsilon(x)) = 1$. Deriving yields

$$\varphi_t(x, \bar{\Psi}_\varepsilon(x)) D\bar{\Psi}_\varepsilon(x) + D\varphi(x, \bar{\Psi}(x)) = 0.$$

Taking $x = x_\varepsilon$ yields

$$D\Psi_\varepsilon(x_I) = -\frac{D\varphi(x_\varepsilon, \bar{\Psi}_\varepsilon(x_\varepsilon))}{\varphi_t(x_\varepsilon, \bar{\Psi}_\varepsilon(x_\varepsilon))} = -\frac{D\varphi(x_I, t_{k_n})}{\varphi_t(x_I, t_{k_n})} + O(\Delta x)$$

and so

$$D\Psi_\varepsilon(x_I) = -\frac{D\varphi(x_0, t_0)}{\varphi_t(x_0, t_0)} + O(|x_I - x_0| + |t_{k_n} - t_0| + \Delta x).$$

Moreover, by Taylor expansion, we get, if $|\varphi(x_\varepsilon, t_\varepsilon) - 1|$ is small enough, for all $J \in V(I)$

$$\begin{aligned}\Psi_\varepsilon(x_J) =& \Psi_\varepsilon(x_I) + (x_J - x_I) \cdot D\Psi_\varepsilon(x_I) + O(|\Delta x|^2) \\ =& \Psi_\varepsilon(x_I) - \frac{D\varphi(x_0, t_0)}{\varphi_t(x_0, t_0)} \cdot (x_J - x_I) + (\Delta x)\ O\left(\Delta x + |x_I - x_0| + |t_{k_n} - t_0|\right),\end{aligned}$$

where "the O is uniform in ε." This ends the proof of (iii).

It just remains to show that, if $\varphi_t(x_0, t_0) < 0$, then for all $(x_J, t_{k_n}) \in Q_{r,\tau}(x_0, t_0) = B_r(x_0) \times (t_0 - \tau, t_0 + \tau)$

$$\theta^\varepsilon(x_J, t_{k_{n-1}}) = 1 \quad \text{and} \quad \theta^\varepsilon(x_J, t_{k_n}) = -1 \quad \Longrightarrow \quad t_{k_n} \le \Psi_\varepsilon(x_J).$$

In this case, inclusion (4.5) is replaced by

$$\{(\theta^\varepsilon)^* = 1\} \subset \{t \le \bar{\Psi}_\varepsilon(x)\}.$$

By the definition of θ^ε, for all $y \in [x_J, x_J + \Delta x]$, we have $(\theta^\varepsilon)^*(y, t_{k_n}) = 1$. By taking $y = x_J + \nu$, we then deduce that

$$t_{k_n} \le \bar{\Psi}_\varepsilon(y) = \bar{\Psi}_\varepsilon(x_J + \nu) = \Psi_\varepsilon(x_J). \qquad \square$$

LEMMA 4.4 (approximate horizontal level set in the i-direction for negative velocity). *Under the notation and assumptions of Lemma 4.3 with $\varphi_t(x_0, t_0) < 0$, let us suppose that there exists $\delta_0 > 0$ such that $c < -\delta < 0$ on $B_{\delta_0}(x_0, t_0)$.*

Let us assume moreover that $(x_I, t_{\overline{n}}) \in B_{\delta_0}(x_0, t_0)$, $\theta_I^{\overline{n}-1} = 1$, and $\theta_I^{\overline{n}} = -1$. If for some fixed $i \in \{1, \dots, N\}$ we have

$$\tilde{u}_I^{\overline{n}-1} - \widehat{u}_{-,I^{i,+}}^{\overline{n}-1} < 0 \quad \text{and} \quad \tilde{u}_I^{\overline{n}-1} - \widehat{u}_{-,I^{i,-}}^{\overline{n}-1} < 0,$$

then

$$\left| \frac{D\varphi(x_0, t_0)}{\varphi_t(x_0, t_0)} \cdot e_i \right| \le o(1).$$

Proof. We first prove that if $\tilde{u}_I^{\overline{n}-1} - \widehat{u}_{-,J}^{\overline{n}-1} < 0$ for some $J \in V(I)\backslash\{I\}$, then

$$\Psi_\varepsilon(x_I) - \Psi_\varepsilon(x_J) \leq o(\Delta x).$$

There are two cases: $\widehat{u}_{-,J}^{\overline{n}-1} = \infty$ or $\widehat{u}_{-,J}^{\overline{n}-1} < \infty$.

If $\widehat{u}_{-,J}^{\overline{n}-1} < \infty$, then $J \in F_-^{\overline{n}-1}$. By Lemma 4.1 it results that

$$\widehat{u}_{-,J}^{\overline{n}-1} = \sup\{t_m \leq t_{\overline{n}-1},\ \theta_J^{m-1} = 1, \theta_J^p = -1 \quad \text{for} \quad m \leq p \leq \overline{n}-1\},$$

and by Lemma 4.3(iv) we have $\widehat{u}_{-,J}^{\overline{n}-1} \leq \Psi_\varepsilon(x_J)$.

We then deduce that

$$0 > \tilde{u}_I^{\overline{n}-1} - \widehat{u}_{-,J}^{\overline{n}-1} \geq \tilde{t}_{\overline{n}} - \Psi_\varepsilon(x_J) = \Psi_\varepsilon(x_I) - \Psi_\varepsilon(x_J) - (t_{\overline{n}} - \tilde{t}_{\overline{n}}).$$

We apply Lemma 4.2, and we obtain

$$\Psi_\varepsilon(x_I) - \Psi_\varepsilon(x_J) \leq o(\Delta x).$$

If $\widehat{u}_{-,J}^{\overline{n}-1} = \infty$, then necessarily $\theta_J^{\overline{n}-1} = 1$, and now either $\theta_J^{\overline{n}} = 1$ or $\theta_J^{\overline{n}} = -1$. Then we can apply Lemma 4.3(i), (resp., (iv)), and we get $t_{\overline{n}} \leq \Psi_\varepsilon(x_J)$. We deduce then that

$$\Psi_\varepsilon(x_I) - \Psi_\varepsilon(x_J) \leq t_{\overline{n}} - t_{\overline{n}} \leq 0.$$

By using Lemma 4.3(iii) for $J = I^{i,\pm}$, we deduce that

$$\pm\Delta x \frac{D\varphi(x_0,t_0)}{\varphi_t(x_0,t_0)} \cdot e_i \leq o(\Delta x). \qquad \square$$

LEMMA 4.5 (decay of θ^ε in the gradient direction of a test function). *Let φ be C^2 in a neighborhood V of (x_0,t_0), and let us suppose that there exist $\delta_0 > 0$ such that $\max_{\overline{V}}((\theta^\varepsilon)^* - \varphi) = (\theta^\varepsilon)^*(x_\varepsilon,t_\varepsilon) - \varphi(x_\varepsilon,t_\varepsilon)$, with $(x_\varepsilon,t_\varepsilon) \in B_{\delta_0}(x_0,t_0) \subset V$ and $(\theta^\varepsilon)^*(x_\varepsilon,t_\varepsilon) = 1$. Then there exists a node $(I,n) \in \mathbb{Z}^N \times \mathbb{N}$ such that $\theta_I^{k_{n+1}-1} = 1$ with $(x_\varepsilon,t_\varepsilon) \in \partial Q_I^n = \partial(]x_I, x_I + \Delta x[\times]t_{k_n}, t_{k_{n+1}}[)$ such that if $\mp e_i \cdot D\varphi(x_0,t_0) > 0$, then*

$$\theta^\varepsilon(x,t) = -1 \qquad \text{in } Q_{I^{i,\pm}}^n =]x_{I^{i,\pm}}, x_{I^{i,\pm}} + \Delta x[\times]t_{k_n}, t_{k_{n+1}}[.$$

Proof of Lemma 4.5. Since $(\theta^\varepsilon)^*(x_\varepsilon,t_\varepsilon) = 1$, there exists a node $(I,n) \in \mathbb{Z}^N \times \mathbb{N}$ such that $\theta_I^{k_{n+1}-1} = 1$ with $(x_\varepsilon,t_\varepsilon) \in \partial Q_I^n = \partial(]x_I, x_I + \Delta x[\times]t_{k_n}, t_{k_{n+1}}[)$.

Assume, for example, that

$$e_i \cdot D\varphi(x_0,t_0) < 0,$$

and let us suppose by contradiction that $\theta^\varepsilon = 1$ in $Q_{I^{i,+}}^n =]x_{I^{i,+}}, x_{I^{i,+}} + \Delta x[\times]t_{k_n}, t_{k_{n+1}}[$.

We define $\varphi_\varepsilon = \varphi + ((\theta^\varepsilon)^* - \varphi)(x_\varepsilon,t_\varepsilon)$. In particular, we have $(\theta^\varepsilon)^* \leq \varphi_\varepsilon$ and $(\theta^\varepsilon)^*(x_\varepsilon,t_\varepsilon) = \varphi_\varepsilon(x_\varepsilon,t_\varepsilon) = 1$. Since $(\theta^\varepsilon)^* \leq \varphi_\varepsilon$, the following inclusion holds:

$$\{(\theta^\varepsilon)^* = 1\} \subset \{\varphi_\varepsilon \geq 1\}.$$

We define $x_\varepsilon^{i,\lambda} = x_\varepsilon + \lambda e_i$, with $0 \leq \lambda \leq \Delta x$ such that $(\theta^\varepsilon)^*(x_\varepsilon^{i,\lambda}, t_\varepsilon) = 1$. Then $\varphi_\varepsilon(x_\varepsilon^{i,\lambda}, t_\varepsilon) \geq 1$ and

$$\frac{\varphi_\varepsilon(x_\varepsilon^{i,\lambda}, t_\varepsilon) - \varphi_\varepsilon(x_\varepsilon, t_\varepsilon)}{\lambda} \geq 0.$$

By taking the limit for $\lambda \to 0$, we obtain

$$e_i \cdot D\varphi(x_\varepsilon, t_\varepsilon) = e_i \cdot D\varphi_\varepsilon(x_\varepsilon, t_\varepsilon) \geq 0.$$

This ends the proof, since it contradicts the assumption. ∎

LEMMA 4.6 (bound on $|t_\varepsilon - t_{\overline{m}_0}|$ for negative velocity). *Under the notation and assumptions of Lemma* 4.5*, if we suppose that there exist* $\delta > 0$ *and* $\delta_0 > 0$ *such that* $c(x,t) < -\delta < 0$ *in* $(x,t) \in B_{2\delta_0}(x_0,t_0) \subset V$*, then the following estimate holds:*

$$|t_\varepsilon - t_{\overline{m}_0}| \leq \frac{\Delta x}{\delta},$$

with

$$t_{\overline{m}_0} = \sup\{t_m \leq t_{k_n} \ : \ \theta^{m-1}_{I^{i,+}} = 1,\ \theta^m_{I^{i,+}} = -1\} \quad \textit{if we assume} \quad -e_i \cdot D\varphi(x_0,t_0) > 0$$

(*resp.,* $t_{\overline{m}_0} = \sup\{t_m \leq t_{k_n} : \theta^{m-1}_{I^{i,-}} = 1,\ \theta^m_{I^{i,-}} = -1\}$ *if we assume* $+e_i \cdot D\varphi(x_0,t_0) > 0$),

where I *is defined in Lemma* 4.5.

Proof of Lemma 4.6. Let us define

$$\overline{m}_0 = \sup\{\ m \leq k_{n+1} - 1,\ \theta^{m-1}_{I^{i,\pm}} = 1, \theta^m_{I^{i,\pm}} = -1\}.$$

For $\Delta x, \Delta t$ small enough, we can assume that $(x_K, t_m) \in B_{2\delta_0}(x_0,t_0)$ for $K = I, I^{i,\pm}$ and $\overline{m}_0 \leq m \leq k_{n+1}$. Since $c < 0$ in $B_{2\delta_0}(x_0,t_0)$, $\theta^{k_{n+1}-1}_I = 1$ implies $\theta^m_I = 1$ for all $\overline{m}_0 \leq m \leq k_{n+1}-1$ and, by the definition of $\overline{m}_0$, $\theta^m_{I^{i,\pm}} = -1$ for all $\overline{m}_0 \leq m \leq k_{n+1}-1$.

This means that $I^{i,\pm} \in F^m_-$ for all $\overline{m}_0 \leq m \leq k_{n+1} - 1$ and so

$$\widehat{u}^m_{-,I^{i,\pm}} = t_{\overline{m}_0} \text{ for } \overline{m}_0 \leq m \leq k_{n+1} - 1. \tag{4.6}$$

In particular, $\widehat{u}^{k_{n+1}-1}_{-,I^{i,\pm}} = t_{\overline{m}_0}$, and by the definition of the $\widehat{t}_{k_{n+1}}$ it results that $\tilde{u}^{k_{n+1}-1}_I \geq \widehat{t}_{k_{n+1}}$, with $\widehat{t}_{k_{n+1}} = t_{k_{n+1}}$, since $t_{k_{n+1}} > t_{k_n}$.

By the equation

$$\sum_{k=1}^{N} \left(\max_{\pm} \left(0, \tilde{u}^{k_{n+1}-1}_I - \widehat{u}^{k_{n+1}-1}_{-,I^{k,\pm}}\right)\right)^2 = \left(\frac{\Delta x}{\widehat{c}^{k_{n+1}-1}_I}\right)^2,$$

we conclude that

$$t_\varepsilon - t_{\overline{m}_0} \leq t_{k_{n+1}} - t_{\overline{m}_0} \leq \tilde{u}^{k_{n+1}-1}_I - \widehat{u}^{k_{n+1}-1}_{-,I^{i,\pm}} \leq \frac{(\Delta x)}{|\widehat{c}^{k_{n+1}-1}_I|} \leq \frac{\Delta x}{\delta}. \qquad \blacksquare$$

5. Proof of Theorem 2.5. This section is dedicated to the proof of the main theorem, which is preceded by two important propositions.

The first proposition will show that the limit function $\overline{\theta}^0$ is a subsolution in all of the domain excepted for the initial time, whereas the second proposition will show that the limit function $\overline{\theta}^0$ is a subsolution at the initial time. The reason why we need to treat apart the initial condition is that the proof of the first proposition is based on the definition of a discontinuous viscosity subsolution (see Barles [4] and Crandall, Ishii, and Lions [10]) consisting in testing the equation by smooth functions, but this definition does not hold at the initial time. Then we treat the initial condition by using the technique of barriers.

At the end of this section, we give the main proof using both results.

PROPOSITION 5.1 (subsolution property of the limit). *The function* $\overline{\theta}^0$ *is a subsolution of the equation*

$$\theta_t(x,t) = c(x,t)|D\theta(x,t)|$$

on $\mathbb{R}^N \times (0,T)$.

Proof of Proposition 5.1. By contradiction, assume that there are (x_0,t_0) and $\varphi \in C^2$ such that $\overline{\theta}^0 - \varphi$ reaches a strict maximum at (x_0,t_0), with $\overline{\theta}^0(x_0,t_0) = \varphi(x_0,t_0)$ and

$$\varphi_t(x_0,t_0) = \alpha + c(x_0,t_0)|D\varphi(x_0,t_0)|, \tag{5.1}$$

with $\alpha > 0$. Since the maximum of $\overline{\theta}^0 - \varphi$ is strict, there exists $(x_\varepsilon,t_\varepsilon) \to (x_0,t_0)$ as $\Delta x \to 0$ such that

$$\max((\theta^\varepsilon)^* - \varphi) = ((\theta^\varepsilon)^* - \varphi)(x_\varepsilon,t_\varepsilon).$$

In particular, we have $(\theta^\varepsilon)^*(x_\varepsilon,t_\varepsilon) = 1$ for $\Delta x, \Delta t$ small enough. Indeed, by contradiction, suppose that $(\theta^\varepsilon)^*(x_\varepsilon,t_\varepsilon) = -1$. By using the fact that $(\theta^\varepsilon)^*$ is upper semi-continuous, we obtain $(\theta^\varepsilon)^* = -1$ a neighborhood of $(x_\varepsilon,t_\varepsilon)$. We then deduce that $\varphi_t(x_\varepsilon,t_\varepsilon) = D\varphi(x_\varepsilon,t_\varepsilon) = 0$ and so

$$0 = \varphi_t(x_\varepsilon,t_\varepsilon) - c(x_\varepsilon,t_\varepsilon)|D\varphi(x_\varepsilon,t_\varepsilon)| \to \varphi_t(x_0,t_0) - c(x_0,t_0)|D\varphi(x_0,t_0)| = \alpha.$$

This is absurd.

If $|D\varphi(x_0,t_0)| \neq 0$, we note that we can rewrite inequality (5.1) as

$$\varphi_t(x_0,t_0) = \bar{c}|D\varphi(x_0,t_0)|, \quad \text{with } \bar{c} > c(x_0,t_0). \tag{5.2}$$

We denote that

$$\vec{n}_0 = \frac{D\varphi(x_0,t_0)}{|D\varphi(x_0,t_0)|}. \tag{5.3}$$

To continue the proof, we have to distinguish several cases:

1. $c(x_0,t_0) > 0$. In this case, we have in particular $\varphi_t(x_0,t_0) > 0$. Then we can apply Lemma 4.3, and we deduce that there exist $\Psi_\varepsilon \in C^2$ and $(I,n) \in \mathbb{Z}^N \times \mathbb{N}$ such that $(x_I,t_{k_n}) \to (x_0,t_0)$ as $\varepsilon = (\Delta x, \Delta t) \to 0$,

$$(\theta^\varepsilon)^*(x_I,t_{k_n}) = 1, \quad t_{k_n} = \Psi_\varepsilon(x_I),$$

and $\theta_I^{\overline{n}} = 1, \quad \theta_I^{\overline{n}-1} = -1$, where $\bar{n}$ is defined in Lemma 4.3. By using Lemmas 4.1 and 4.3(i), we deduce also that, for all $J \in V(I)\setminus\{I\}$ such that $\theta_J^{\overline{n}-1} = 1$, we have

$$\widehat{u}_{+,J}^{\overline{n}-1} \geq \Psi_\varepsilon(x_J).$$

This implies for all $J \in V(I) \cap F_+^{\overline{n}-1}$, using also the (general) fact that $\tilde{u}_I^{\overline{n}-1} \leq t_{\overline{n}} = t_{k_n}$, that

$$\tilde{u}_I^{\overline{n}-1} - \widehat{u}_{+,J}^{\overline{n}-1} \leq t_{\overline{n}} - \widehat{u}_{+,J}^{\overline{n}-1} = \Psi_\varepsilon(x_I) - \widehat{u}_{+,J}^{\overline{n}-1} \leq \Psi_\varepsilon(x_I) - \Psi_\varepsilon(x_J). \tag{5.4}$$

By the GFMM algorithm (Step 5), $\tilde{u}_I^{\overline{n}-1}$ is a solution of the equation

$$\left(\frac{\Delta x}{c(x_I, t_{\overline{n}-1})}\right)^2 = \sum_{i=1}^{N} \left(\max_{\pm}\left(0, \tilde{u}_I^{\overline{n}-1} - \widehat{u}_{+,I^{i,\pm}}^{\overline{n}-1}\right)\right)^2.$$

If $|D\varphi(x_0,t_0)| \neq 0$, by adding (5.4) for $J = I^{i,\pm}$ on all directions $i \in \mathcal{C} \subset \{1,\dots,N\}$ such that

$$\tilde{u}_I^{\overline{n}-1} - \widehat{u}_{+,I^{i,+}}^{\overline{n}-1} \geq 0 \ \text{ or } \ \tilde{u}_I^{\overline{n}-1} - \widehat{u}_{+,I^{i,-}}^{\overline{n}-1} \geq 0$$

and by using Lemma 4.3(iii), we can estimate

$$\begin{aligned}\left(\frac{\Delta x}{c(x_I, t_{\overline{n}-1})}\right)^2 &= \sum_{i\in\mathcal{C}} \left(\max_{\pm}\left(\tilde{u}_I^{\overline{n}-1} - \widehat{u}_{+,I^{i,\pm}}^{\overline{n}-1}\right)\right)^2 \leq \sum_{i\in\mathcal{C}} \max_{\pm}\left(\Psi_\varepsilon(x_I) - \Psi_\varepsilon(x_{I^{i,\pm}})\right)^2 \\ &\leq \frac{(\Delta x)^2}{\bar{c}^2} \sum_{i\in\mathcal{C}} (\vec{n_0}\cdot e_i)^2 + (\Delta x)^2 O(\Delta x + |x_I - x_0| + |t_{k_n} - t_0|) \\ &\leq \frac{(\Delta x)^2}{\bar{c}^2} + (\Delta x)^2 O(\Delta x + |x_I - x_0| + |t_{k_n} - t_0|),\end{aligned}$$

where $\bar{c}$ and $\vec{n}_0$ are defined in (5.2) and (5.3), respectively.

It follows that

$$\frac{1}{c^2(x_I, t_{\overline{n}-1})} - \frac{1}{\bar{c}^2} \leq O(\Delta x + |x_I - x_0| + |t_{k_n} - t_0|).$$

By taking the limit $\varepsilon = (\Delta x, \Delta t) \to 0$, we obtain a contradiction.

If $D\varphi(x_0,t_0) = 0$, we get in the same way

$$\frac{1}{c^2(x_I, t_{\overline{n}-1})} \leq O(\Delta x + |x_I - x_0| + |t_{k_n} - t_0|).$$

By taking the limit $\varepsilon \to 0$, since we have assumed that $c(x_0,t_0) > 0$, we obtain a contradiction.

2. $c(x_0,t_0) < 0$. In this case, we have no information on the sign of φ_t, so we have to distinguish several cases:

1. $\varphi_t(x_0,t_0) < 0$. Note that, in this case, $|D\varphi(x_0,t_0)| \neq 0$ and (5.2) holds with $0 > \bar{c} > c(x_0,t_0)$.
   Then we can apply Lemma 4.3 and we deduce that there exist $\Psi_\varepsilon \in C^2$ and $(I,n) \in \mathbb{Z}^N \times \mathbb{N}$ such that

   $$(\theta^\varepsilon)^*(x_I, t_{k_n}) = 1, \quad t_{k_n} = \Psi_\varepsilon(x_I),$$

   and

   $$\theta_I^{\overline{n}} = -1, \quad \theta_I^{\overline{n}-1} = 1,$$

   where $\bar{n}$ is defined in Lemma 4.3. By using Lemmas 4.1 and 4.3(iv), we deduce also that, for all $J \in V(I)\setminus\{I\}$ such that $\theta_J^{\overline{n}-1} = -1$, we have

   $$\widehat{u}_{-,J}^{\overline{n}-1} \leq \Psi_\varepsilon(x_J).$$

This implies that for all $J \in V(I) \cap F_-^{\overline{n}-1}$

$$\tilde{u}_I^{\overline{n}-1} - \widehat{u}_{-,J}^{\overline{n}-1} \geq \tilde{t}_{\overline{n}} - \Psi_\varepsilon(x_J) = t_{\overline{n}} - \Psi_\varepsilon(x_J) + (\tilde{t}_{\overline{n}} - t_{\overline{n}}) = \Psi_\varepsilon(x_I) - \Psi_\varepsilon(x_J) + (\tilde{t}_{\overline{n}} - t_{\overline{n}}).$$

Since $c(x_0, t_0) \neq 0$, there exist $\delta,\ \delta_0 > 0$ such that $|c| \geq \delta > 0$ on $B_{\delta_0}(x_0, t_0)$, and we can apply Lemma 4.2 to get

$$\tilde{u}_I^{\overline{n}-1} - \widehat{u}_{-,J}^{\overline{n}-1} \geq \Psi_\varepsilon(x_I) - \Psi_\varepsilon(x_J) + o(\Delta x).$$

Using Lemma 4.3(iii) yields

$$\tilde{u}_I^{\overline{n}-1} - \widehat{u}_{-,J}^{\overline{n}-1} \geq \frac{1}{\bar{c}} \vec{n_0} \cdot (x_J - x_I) + (\Delta x)\ O(\Delta x + |x_I - x_0| + |t_{k_n} - t_0|) + o(\Delta x).$$

By adding the previous equation for $J = I^{i,\pm}$ on all directions $i \in \mathcal{C} \subset \{1, \ldots, N\}$ such that

$$\tilde{u}_I^{\overline{n}-1} - \widehat{u}_{-,I^{i,+}}^{\overline{n}-1} \geq 0 \ \text{ or } \ \tilde{u}_I^{\overline{n}-1} - \widehat{u}_{-,I^{i,-}}^{\overline{n}-1} \geq 0,$$

we obtain, since $|D\varphi(x_0, t_0)| \neq 0$,

$$\begin{aligned} \left(\frac{\Delta x}{c(x_I, t_{\overline{n}-1})}\right)^2 &= \sum_{i=1}^{N} \left(\max_{\pm}\left(0, \tilde{u}_I^{\overline{n}-1} - \widehat{u}_{-,I^{i,\pm}}^{\overline{n}-1}\right)\right)^2 \\ &= \sum_{i \in \mathcal{C}} \left(\tilde{u}_I^{\overline{n}-1} - \widehat{u}_{-,I^{i,\pm}}^{\overline{n}-1}\right)^2 \\ &\geq \frac{(\Delta x)^2}{\bar{c}^2} \sum_{i \in \mathcal{C}} (\vec{n_0} \cdot e_i)^2 + (\Delta x)^2\ O(\Delta x + |x_I - x_0| \\ &\qquad + |t_{k_n} - t_0|) + o(\Delta x)^2. \end{aligned} \tag{5.5}$$

If $i \notin \mathcal{C}$ (i.e., $\tilde{u}_I^{\overline{n}-1} - \widehat{u}_{-,I^{i,+}}^{\overline{n}-1} < 0$ and $\tilde{u}_I^{\overline{n}-1} - \widehat{u}_{-,I^{i,-}}^{\overline{n}-1} < 0$), then by Lemma 4.4 we deduce that

$$\left|\frac{1}{\bar{c}} \Delta x\ \vec{n_0} \cdot e_i\right| = o(\Delta x). \tag{5.6}$$

By combining (5.5) and (5.6), we get

$$\begin{aligned} \left(\frac{\Delta x}{c(x_I, t_{\overline{n}-1})}\right)^2 &\geq \frac{(\Delta x)^2}{\bar{c}^2} \sum_{i=1}^{N} (\vec{n_0} \cdot e_i)^2 + (\Delta x)^2\ O(\Delta x + |x_I - x_0| \\ &\qquad + |t_{k_n} - t_0|) + o(\Delta x^2) \\ &= \frac{(\Delta x)^2}{\bar{c}^2} + (\Delta x)^2\ O(\Delta x + |x_I - x_0| + |t_{k_n} - t_0|) + o(\Delta x^2). \end{aligned}$$

This implies that

$$\frac{1}{c^2(x_I, t_{\overline{n}-1})} - \frac{1}{\bar{c}^2} \geq O(\Delta x + |x_I - x_0| + |t_{k_n} - t_0|) + o(1).$$

By taking the limit $\varepsilon = (\Delta x, \Delta t) \to 0$, we get the contradiction since $|c(x_0, t_0)| > |\bar{c}|$.

2. $\varphi_t(x_0,t_0) > 0$. Since $c(x_0,t_0) < 0$, we have by the algorithm that $\frac{\partial(\theta^\varepsilon)^*}{\partial t} \le 0$. We define $\varphi_\varepsilon = \varphi + ((\theta^\varepsilon)^* - \varphi)(x_\varepsilon, t_\varepsilon)$. In particular, we have $(\theta^\varepsilon)^* \le \varphi_\varepsilon$ and
$$(\theta^\varepsilon)^*(x_\varepsilon,t_\varepsilon) = \varphi_\varepsilon(x_\varepsilon,t_\varepsilon) = 1.$$
We have $t_\varepsilon = t_{k_n}$. Indeed, assume that $t_\varepsilon \in (t_{k_n}, t_{k_{n+1}})$. Bu using the fact that $(\theta^\varepsilon)^*(x_\varepsilon,t_\varepsilon) = 1$, we deduce that $(\theta^\varepsilon)^*(x_\varepsilon,t) = 1$ for $t_{k_n} \le t \le t_{k_{n+1}}$ and so $\varphi_t(x_\varepsilon,t_\varepsilon) = 0$. This is absurd for ε small enough since $\varphi_t(x_0,t_0) > 0$.
By using the fact that $(\varphi_\varepsilon)_t > 0$, we deduce that $(\theta^\varepsilon)^*(x_\varepsilon,t) \le \varphi_\varepsilon(x_\varepsilon,t) < 1$ for $t < t_{k_n}$. This is absurd since $\frac{\partial(\theta^\varepsilon)^*}{\partial t} \le 0$.
3. $\varphi_t(x_0,t_0) = 0$. Since (5.1) holds with $\alpha > 0$, we have, in particular, $|D\varphi(x_0,t_0)| \neq 0$. Then there exists a direction $\pm e_i$ such that $\mp e_i \cdot D\varphi(x_0,t_0) > 0$. By using Lemma 4.5, we deduce that there exists $(I,n) \in \mathbb{Z}^N \times \mathbb{N}$ such that $\theta_I^{k_{n+1}-1} = 1$ and $\theta^\varepsilon = -1$ on $Q^n_{I^{i,\pm}} =]x_{I^{i,\pm}}, x_{I^{i,\pm}} + \Delta x[\times]t_{k_n}, t_{k_{n+1}}[$. We define $t_{\overline{m}_0}$ such that
$$\overline{m}_0 = \sup\{m \,:\, t_m \le t_{k_n},\ \theta^{m-1}_{I^{i,\pm}} = 1,\ \theta^m_{I^{i,\pm}} = -1\}.$$
In particular, $(\theta^\varepsilon)^*(x, t_{\overline{m}_0}) = 1$ for all $x \in [x_{I^{i,\pm}}, x_{I^{i,\pm}} + \Delta x]$.
We define $\varphi_\varepsilon = \varphi + ((\theta^\varepsilon)^* - \varphi)(x_\varepsilon,t_\varepsilon)$. In particular, we have $(\theta^\varepsilon)^* \le \varphi_\varepsilon$ and
$$(\theta^\varepsilon)^*(x_\varepsilon,t_\varepsilon) = \varphi_\varepsilon(x_\varepsilon,t_\varepsilon) = 1.$$
Since the following inclusion $\{(\theta^\varepsilon)^* = 1\} \subset \{\varphi_\varepsilon \ge 1\}$ holds, $\varphi_\varepsilon(x, t_{\overline{m}_0}) \ge 1$ for all $x \in [x_{I^{i,\pm}}, x_{I^{i,\pm}} + \Delta x]$.
Let $\nu \in [0,\Delta x]^N$ be such that $x_\varepsilon = x_I + \nu$, and let us define $y \equiv x_{I^{i,\pm}} + \nu$ and $\overline{\varphi}(\cdot,\cdot) \equiv \varphi_\varepsilon(\cdot + \nu, \cdot)$. Then it yields $\overline{\varphi}(x_I,t_\varepsilon) = \varphi_\varepsilon(x_\varepsilon,t_\varepsilon) = 1$, and $\overline{\varphi}(x_{I^{i,\pm}}, t_{\overline{m}_0}) = \varphi_\varepsilon(y, t_{\overline{m}_0}) \ge 1$.
To obtain the contradiction, we consider the expansion of $\overline{\varphi}$ up to the first order
$$\begin{aligned} 0 &\le \overline{\varphi}(x_{I^{i,\pm}}, t_{\overline{m}_0}) - \overline{\varphi}(x_I,t_\varepsilon) \\ &\le (x_{I^{i,\pm}} - x_I)\cdot D\overline{\varphi}(x_I,t_\varepsilon) + (t_{\overline{m}_0} - t_\varepsilon)\partial_t\overline{\varphi}(x_I,t_\varepsilon) + O((\Delta x)^2 + |t_\varepsilon - t_{\overline{m}_0}|^2). \end{aligned}$$
Now by Lemma 4.6 and using the fact that $\partial_t\varphi(x_0,t_0) = 0$ we obtain
$$\pm e_i \cdot D\overline{\varphi}(x_I,t_\varepsilon)\Delta x + o(\Delta x) \ge 0,$$
which is absurd, since by assumption $\pm e_i \cdot D\varphi(x_0,t_0) < 0$.
3. $c(x_0,t_0) = 0$. In this case, we have
$$\varphi_t = \alpha > 0,$$
and we can apply Lemma 4.3. Hence, there exist $r,\ \tau > 0$, a function $\Psi_\varepsilon \in C^2(B_r(x_0),(t_0-\tau,t_0+\tau))$, and a node $(I,n) \in \mathbb{Z}^N \times \mathbb{N}$ such that
$$(\theta^\varepsilon)^*(x_I,t_{k_n}) = 1, \quad t_{k_n} = \Psi_\varepsilon(x_I),$$
and for all $J \in V(I),\ t_m \in (t_0-\tau, t_0+\tau)$, we have
$$\theta^\varepsilon(x_J,t_m) = 1 \implies t_m \ge \Psi_\varepsilon(x_J). \tag{5.7}$$
We define m_0 such that $t_{m_0-1} < t_0 - \tau \le t_{m_0}$.
For all $J \in (V(I)\setminus\{I\}) \cap \{\theta^{\overline{n}-1} = 1\}$ (with $\bar{n}$ defined in Lemma 4.3), we define
$$m_J = \sup\{k \le \overline{n},\ \theta^{k-1}_J = -1\}.$$
We distinguish two cases:

1. There exists $J \in (V(I)\setminus\{I\}) \cap \{\theta^{\overline{n}-1} = 1\}$ such that $m_J < m_0$.
   By using the fact that $\theta_I^k = -1$ for $m_0 \le k \le \overline{n}-1$ (see Lemma 4.3(ii)), we have that $J \in F_+^k \ \forall\, m_0 \le k \le \overline{n}-1$, and we deduce that

$$\widehat{u}_{+,J}^{\overline{n}-1} = u_J^{\overline{n}-1} \le t_{m_0} \quad \text{and} \quad \theta^\varepsilon(x_J, t_{m_0}) = 1.$$

   By (5.7), we then have $t_{m_0} \ge \Psi_\varepsilon(x_J)$.
   We now assume that $|D\varphi| \neq 0$ (the case $|D\varphi| = 0$ can be treated in a similar way). By using Lemma 4.3, we deduce that

$$t_{m_0} \ge \Psi_\varepsilon(x_J) = t_{k_n} - \frac{1}{\bar{c}}\,\vec{n}_0 \cdot (x_J - x_I) + (\Delta x)\ O(\Delta x + |x_I - x_0| + |t_{k_n} - t_0|),$$

   and so

$$\begin{aligned} t_{k_n} - t_{m_0} \le & \frac{1}{\bar{c}}\,\vec{n}_0 \cdot (x_J - x_I) + (\Delta x)\ O(\Delta x + |x_I - x_0| + |t_{k_n} - t_0|) \\ \le & \frac{\Delta x}{\bar{c}} + (\Delta x)\ O(\Delta x + |x_I - x_0| + |t_{k_n} - t_0|). \end{aligned}$$

   Sending $\Delta x, \Delta t$ to 0 yields $t_0 - (t_0 - \tau) = \tau \le 0$. This is absurd.
2. For all $J \in (V(I)\setminus\{I\}) \cap \{\theta^{\overline{n}-1} = 1\}$, $m_J \ge m_0$.
   We then have $\theta^\varepsilon(x_J, t_{m_J}) = 1$, and so by (5.7) we have $\widehat{u}_{+,J}^{\overline{n}-1} = t_{m_J} \ge \Psi(x_J)$.
   We now assume that $|D\varphi| \neq 0$ (the case $|D\varphi| = 0$ can be treated in a similar way). By using Lemma 4.3, we deduce that

$$\widehat{u}_{+,J}^{\overline{n}-1} \ge \Psi(x_J) = t_{k_n} - \frac{1}{\bar{c}}\vec{n}_0 \cdot (x_J - x_I) + (\Delta x)\ O(\Delta x + |x_I - x_0| + |t_{k_n} - t_0|),$$

   and so

$$\begin{aligned} \tilde{u}_I^{\overline{n}-1} - \widehat{u}_{+,J}^{\overline{n}-1} \le & \ t_{k_n} - \widehat{u}_{+,J}^{\overline{n}-1} \\ \le & \ \frac{1}{\bar{c}}\vec{n}_0 \cdot (x_J - x_I) + (\Delta x)\ O(\Delta x + |x_I - x_0| + |t_{k_n} - t_0|). \end{aligned}$$

   By adding for $J = I^{i,\pm}$ on all directions $i \in \mathcal{C} \subset \{1, \dots, N\}$ such that

$$\tilde{u}_I^{\overline{n}-1} - \widehat{u}_{+,I^i}^{\overline{n}-1} = \max(\tilde{u}_I^{\overline{n}-1} - \widehat{u}_{+,I^{i,+}}^{\overline{n}-1}, \tilde{u}_I^{\overline{n}-1} - \widehat{u}_{+,I^{i,-}}^{\overline{n}-1}) \ge 0,$$

   we deduce that

$$\left(\frac{\Delta x}{\widehat{c}_I^{\overline{n}-1}}\right)^2 = \sum_{i\in\mathcal{C}} \left(\tilde{u}_I^n - \widehat{u}_{+,I^i}^{n-1}\right)^2 \le \left(\frac{\Delta x}{\bar{c}}\right)^2 + (\Delta x)^2\, O(\Delta x + |x_I - x_0| + |t_{k_n} - t_0|),$$

   i.e.,

$$\frac{1}{|\widehat{c}_I^{\overline{n}-1}|^2} \le \frac{1}{\bar{c}^2} + O(\Delta x + |x_I - x_0| + |t_{k_n} - t_0|).$$

   Sending $\Delta x, \Delta t$ to 0 yields a contradiction since $\bar{c} > c(x_0, t_0) = 0$. ∎

We construct a barrier subsolution, and we prove that $\overline{\theta}^0$ defined by (2.7) satisfies the initial condition of (1.1).

PROPOSITION 5.2 (initial condition). *We have the following inequality:*

$$\overline{\theta}^0(\cdot, 0) \le \left(1_{\Omega_0} - 1_{\Omega_0^c}\right)^*. \tag{5.8}$$

Proof of Proposition 5.2. For $\alpha > 0$ which will be made precise later, we consider the following function:

$$v(x) = \alpha \operatorname{dist}(x, \Omega_0), \tag{5.9}$$

and we define for all $I \in \mathbb{Z}^N$

$$v_I = v(x_I).$$

We then define for $x_I \in \Omega_0^c$ a velocity $\infty > c_{v,I} > 0$ by solving

$$\sum_{k=1}^{N} (\max_{\pm}(0, v_I - \widehat{v}_{I^{k,\pm}}))^2 = \left(\frac{\Delta x}{c_{v,I}}\right)^2,$$

where

$$\widehat{v}_J = \begin{cases} v_J & \text{if} \quad v_J \le v_I, \\ \infty & \text{if} \quad v_J > v_I. \end{cases}$$

This define a GFMM with velocity $c_{v,I}$ and whose solution is v_I. On the one hand, using the fact that $|v_I - v_J| \le \alpha \Delta x$ yields for $J \in V(I)$

$$c_{v,I} \ge \frac{1}{\alpha\sqrt{N}}. \tag{5.10}$$

On the other hand, the C^2 regularity of $\partial\Omega_0$ implies that $c_{v,I}$ is uniformly bounded as $\Delta x \to 0$ in a neighborhood of $\partial\Omega_0$.

Moreover, we can define θ_v^ε in the following way:

$$\theta_v^\varepsilon(x,t) = \begin{cases} 1 & \text{if} \quad x \in [x_I, x_I + \Delta x[\text{ and } t \ge v_I, \\ -1 & \text{if} \quad x \in [x_I, x_I + \Delta x[\text{ and } t < v_I. \end{cases}$$

We denote by u the solution of the GFMM algorithm with velocity $c(x,t)$. We then have

$$\theta_{u,I}^0 = 1 \Rightarrow x_I \in \Omega_0 \Rightarrow v_I = 0 \Rightarrow \theta_{v,I}^0 = 1,$$

and so $\{\theta_u^0 = 1\} \subset \{\theta_v^0 = 1\}$. Moreover, by using (5.10), we deduce that, for α small enough, we have, for all $t \ge 0$, $c_{v,I} \ge (c(x_I,t))^+$. By using the comparison principle Corollary 3.5, we deduce that $\theta_v^\varepsilon(x,t) \ge \theta^\varepsilon(x,t)$. We denote that $v^\varepsilon(x) = \sup_{y \in [x-\Delta x, x]} v(y)$ and $\theta_{v^\varepsilon}(x,t) = 1_{\{v^\varepsilon(x) \ge t\}} - 1_{\{v^\varepsilon(x) < t\}}$. It is easy to check that

$$(\theta_{v^\varepsilon})^*(x,t) \ge (\theta_v^\varepsilon)^*(x,t) \ge (\theta^\varepsilon)^*(x,t).$$

By passing to the limit $\varepsilon \to 0$, we then obtain for $t > 0$

$$1_{\{v(x) \ge t\}} - 1_{\{v(x) < t\}} = \theta_v(x,t) \ge \overline{\theta}^0(x,t)$$

and so $(1_{\Omega_0} - 1_{\Omega_0^c})^* \ge \overline{\theta}^0(x,0)$. This implies that $\overline{\theta}^0$ satisfies the initial condition (5.8). □

Proof of Theorem 2.5. The proof of Theorem 2.5 is now quite simple. Indeed, by using Theorem 5.1 and Proposition 5.2, we get that $\overline{\theta}^0$ is a viscosity subsolution of (1.1).

For the supersolution property of $\underline{\theta}^0$, it suffices to use the symmetry of $\overline{\theta}^0$ and $\underline{\theta}^0$ (see Lemma 3.1). Indeed, by contradiction, assume that there are (x_0, t_0) and $\varphi \in C^2$ such that $\underline{\theta}^0 - \varphi$ reaches a strict minimum at (x_0, t_0), with

$$\varphi_t(x_0, t_0) = -\alpha + c(x_0, t_0)|D\varphi(x_0, t_0)|,$$

with $\alpha > 0$ and $t_0 > 0$. Let us define $c_1 = -c$, $\varphi_1 = -\varphi$, and $\overline{\theta}_1^0 = \overline{\theta}^0[-\theta^0, -c]$. Then, by using Lemma 3.1, we get that $\overline{\theta}_1^0 - \varphi_1$ reaches a strict maximum at (x_0, t_0) with $\overline{\theta}_1^0(x_0, t_0) = \varphi_1(x_0, t_0)$ and

$$(\varphi_1)_t(x_0, t_0) = \alpha + c_1(x_0, t_0)|D\varphi(x_0, t_0)|.$$

This contradicts the subsolution property of $\overline{\theta}_1^0$. For the initial condition, we use the same arguments as those of Proposition 5.2.

Moreover, if (1.1) satisfies a comparison principle, then $\overline{\theta}^0 \leq (\underline{\theta}^0)^*$ and $(\overline{\theta}^0)_* \leq \underline{\theta}^0$. Since, by definition, $\overline{\theta}^0 \geq \underline{\theta}^0$, we get that $\overline{\theta}^0 = (\underline{\theta}^0)^*$ and $(\overline{\theta}^0)_* = \underline{\theta}^0$ is a solution of (1.1). This exactly means that $\overline{\theta}^0$ and $\underline{\theta}^0$ are solutions, which are then unique (when the comparison principle holds for a special choice of the initial data), up to the upper and the lower semicontinuous envelopes. ∎

6. Numerical tests. We are going to verify our algorithm by some numerical tests in dimension $N = 2$.

First we will give in two cases the representation formula of the solution so that we will be able to obtain numerical errors comparing it with the numerical solution obtained by the GFMM algorithm.

Representation formulas for hyperplanes and spheres propagating with linear speed. We verify that hyperplanes and spheres in $\mathbb{R}^N$, which propagate with a linear speed along the normal direction, keep their shapes during the evolution remaining, respectively, hyperplanes and spheres.

These manifolds can be characterized by the level set of a polynomial $P(x) : \mathbb{R}^N \to \mathbb{R}$ of degree 1 and 2. We denote by $P(x, t)$ the polynomials with coefficients depending on t.

Each point x such that (s.t.) $P(x, t_0) = 0$ verifies the following dynamics:

$$\begin{cases} \dot{y}(t) &= -c(y(t), t)\frac{DP(y(t),t)}{|DP(y(t),t)|}, \\ y(t_0) &= x, \end{cases}$$

since they propagate with speed c along the unit normal to the manifold. These trajectories are known as *characteristics*. Then we just need to check that the evolution of each point of the manifold verifies the equation $P(y(t), t) = 0$, i.e., derive with respect to t

$$P_t(y(t), t) - |DP(y(t), t)|c(y(t), t) = 0 \tag{6.1}$$

for any linear speed $c(x, t) = a(t)x + b(t)$ and for any $P(x, t)$ representing hyperplanes or spheres.

Hyperplanes: $P(x, t) = \alpha(t)x + \beta(t)$. It results that $P_t(x, t) = \dot{\alpha}(t)x + \dot{\beta}(t)$ and $|DP(x(t), t)| = |\alpha(t)|$, and then $P(x, t)$ verifies (6.1) with coefficients such that:

$$\begin{cases} \dot{\alpha}(t) = |\alpha(t)|a(t), \\ \dot{\beta}(t) = |\alpha(t)|b(t). \end{cases}$$

Spheres: $P(x,t) = R(t)^2 - |x - x_0(t)|^2$. It results that $P_t(x,t) = 2(x - x_0(t))\dot{x}_0(t) + 2R(t)\dot{R}(t)$ and $|DP(x(t),t)| = 2|x - x_0(t)|$, and then $P(x,t)$ verifies (6.1) with coefficients such that:

$$\begin{cases} \dot{x}_0(t) = a(t)R(t), \\ \dot{R}(t) = x_0(t)a(t) + b(t). \end{cases}$$

Test 1: *A rotating line.* We choose as initial data a line $P(x,0) = x_2 + 1.5x_1$ and then as representing function:

$$\theta(x,0) = \begin{cases} 1 & \text{if } x_2 + 1.5x_1 > 0, \\ -1 & \text{otherwise.} \end{cases} \tag{6.2}$$

We choose as velocity $c(x,t) = x_1$. We have proved that a line propagating with linear speed stays a line. By applying the result of the previous section, we obtain that $P(x,t) = \alpha(t)x + \beta(t)$ has coefficients solving the following ODE:

$$\begin{cases} \dot{\alpha}_1(t) = \sqrt{1 + \alpha_1(t)^2}, \\ \alpha_1(0) = 1.5, \end{cases} \qquad \begin{cases} \dot{\alpha}_2(t) = 0, \\ \alpha_2(0) = 1, \end{cases}$$

and then we get $P(x,t) = \sinh(t + \operatorname{arcsinh}(\alpha_1(0)))x_1 + x_2$.

We compute the discrete solution in the numerical domain $D = [-1,1] \times [-1,1]$. The evaluation of the error is a delicate point. We decided to evaluate the error at the final time $T = 1$ in terms of the difference between the area $\mathcal{A}(\cdot)$ of the set $\Omega_T^+ = \{x \in \mathbb{R}^2 : P(x,T) > 0\}$ and its numerical approximation $\Omega_m^+ = \{x \in \mathbb{R}^2 : I[\theta^m](x) > 0\}$, where m is the number of iterations to reach the final time T and $I[\theta^m]$ is the linear interpolation of the discrete numerical solution $(\theta_I^m)_I$. In both cases we were able to compute the area exactly.

Note that a priori a small error on the areas does not guarantee that the two fronts (exact and approximated) are close each other since, for example, positive errors in the area on one piece of the boundary can be compensated by negative errors on another piece of the boundary. Naturally, the more accurate way to evaluate the error at $T = 1$ is to compute the Hausdorff distance $\mathcal{H}(\cdot,\cdot)$ between the exact front $\mathcal{C} = \{x \in \mathbb{R}^2 : P(x,T) = 0\}$ and the approximated front $\tilde{\mathcal{C}} = \{x \in \mathbb{R}^2 : I[\theta^m](x) = 0\}$. Although that distance cannot be computed exactly, a good approximation can be obtained by computing it on only a finite number of points belonging to the fronts.

We measure the amount of inhomogeneity of the speed by the ratio

$$\max_{I \in F_+^n} |\widehat{c}_I^n| / \min_{I \in F_-^n} |\widehat{c}_I^n|.$$

Since in this test there are points in the front with speed 0, the coefficient results to be infinity.

The speed is constant in time; then we are in the situation described in Case 1 in section 2.3, and we can implement the algorithm without updating the values of $\tilde{u}$ on each node of the front at each iteration. We need only to recompute the value of $\tilde{u}$ at the node I whose neighbors have been accepted ($I \in V(NA^n)$); i.e., after one iteration the algorithm should be modified in Step 5 as follows:

for $n \geq 2$

5. *Compute* $\tilde{u}^{n-1}$ *on* $F_\pm^{n-1}$ *as follows*:
   1. If $I \notin V(NA^{n-1})$, then $\tilde{u}_I^{n-1} = \tilde{u}_I^{n-2}$;

TABLE 1
Area and Hausdorff distances: GFMM-Case 1 *versus FD, for Test* 1 *with constant time speed.*

	GFMM			FD		
Δx	$\lvert\mathcal{A}(\Omega_T^+) - \mathcal{A}(\Omega_m^+)\rvert$	$\mathcal{H}(\mathcal{C}, \tilde{\mathcal{C}})$	CPU	$\lvert\mathcal{A}(\Omega_T^+) - \mathcal{A}(\Omega_n^+)\rvert$	$\mathcal{H}(\mathcal{C}, \tilde{\mathcal{C}})$	CPU
0.04	$1.62 \cdot 10^{-1}$	$5.08 \cdot 10^{-2}$	0.19 s	$1.15 \cdot 10^{-1}$	$4.10 \cdot 10^{-2}$	1.82 s
0.02	$8.26 \cdot 10^{-2}$	$2.72 \cdot 10^{-2}$	0.73 s	$5.16 \cdot 10^{-2}$	$2.05 \cdot 10^{-2}$	13.2 s
0.01	$4.05 \cdot 10^{-2}$	$1.35 \cdot 10^{-2}$	3.98 s	$2.80 \cdot 10^{-2}$	$1.03 \cdot 10^{-2}$	102 s
0.005	$2.05 \cdot 10^{-2}$	$6.80 \cdot 10^{-3}$	76 s	$9.00 \cdot 10^{-3}$	$2.60 \cdot 10^{-3}$	810 s

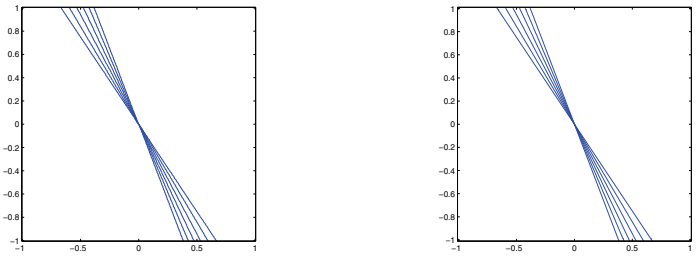


FIG. 4. *A rotating line by the GFMM algorithm (left) and by the FD algorithm (right).*

2. if $I \in V(NA^{n-1})$, then
   (a) if $\pm\widehat{c}_I^{n-1} \geq 0$, $\tilde{u}_I^{n-1} = \infty$;
   (b) if $\pm\widehat{c}_I^{n-1} < 0$, compute $\tilde{u}_I^{n-1}$ as the solution of (2.4).

We compare the GFMM method with the iterative finite difference (FD) scheme described in [17], in terms of complexity and accuracy. In this case the approximated set is $\Omega_n^+ = \{x \in \mathbb{R}^2 : I[v^n](x) > 0\}$, where $(v_I^n)_I$ is the solution of the FD scheme approximating the continuous solution $P(x,T)$, $I[v^n]$ its linear interpolation, and $\tilde{\mathcal{C}} = \{x \in \mathbb{R}^2 : I[v^n](x) = 0\}$ the approximated front.

Table 1 shows the error for the tests run with 50, 100, 200, 400 and the number of nodes for each side of the square domain.

Table 1 shows a better performance in terms of accuracy for the FD scheme; this can be explained by observing that in the scheme the front is represented by the zero level set of a continuous function and this convergence is uniform. On the other hand, the GFMM algorithm represents the front by the interface of a discontinuous function (i.e., the front is where there is a jump between -1 and 1). The advantage here to use the GFMM scheme is in terms of CPU time; the GFMM results are approximately 10 times faster than the full matrix approach.

Figure 4 (left) shows the interface between $\{\theta^\varepsilon = 1\}$ and $\{\theta^\varepsilon = -1\}$ at each time interval 0.1. The line is rotating clockwise, and it will reach in infinite time the x_2 axis. Figure 4 (right) shows the same test computed by the FD scheme; here the lines are the 0-level set of the discrete function $(v_I^n)_{I,n}$. In both cases the test has computed with $\Delta x = 0.01$, and the line has been plotted at times $t_n = n0.1$, $n = 1, 2, 3 \ldots$.

Now let us consider a speed depending on time and changing sign during the evolution $c(x,t) = \sin(2\pi t)x_1$. This means that we are in Case 2 of section 2.3. We obtain the exact solution by solving (6.1). The line is rotating around the origin, changing its direction of rotation after every time interval of length 1/2. We compute the discrete solution in the numerical domain $D = [-1,1] \times [-1,1]$, we evaluate the error at final time $T = 1$, and we choose $\Delta T = 4\Delta x > \frac{1}{\sqrt{M}}$ as the time step to update

TABLE 2
Area and Hausdorff distances: GFMM-Case 2 versus FD for Test 1 with time-dependent speed.

	GFMM			FD		
Δx	$\lvert\mathcal{A}(\Omega_T^+) - \mathcal{A}(\Omega_m^+)\rvert$	$\mathcal{H}(\mathcal{C}, \tilde{\mathcal{C}})$	CPU	$\lvert\mathcal{A}(\Omega_T^+) - \mathcal{A}(\Omega_n^+)\rvert$	$\mathcal{H}(\mathcal{C}, \tilde{\mathcal{C}})$	CPU
0.04	$1.60 \cdot 10^{-1}$	$5.21 \cdot 10^{-2}$	0.52 s	$9.76 \cdot 10^{-2}$	$4.82 \cdot 10^{-2}$	1.82 s
0.02	$8.66 \cdot 10^{-2}$	$3.07 \cdot 10^{-2}$	1.71s	$4.76 \cdot 10^{-2}$	$2.46 \cdot 10^{-2}$	13.3 s
0.01	$4.27 \cdot 10^{-2}$	$1.54 \cdot 10^{-2}$	10.5 s	$2.37 \cdot 10^{-2}$	$1.35 \cdot 10^{-2}$	102 s
0.005	$2.14 \cdot 10^{-2}$	$9.00 \cdot 10^{-3}$	130 s	$6.88 \cdot 10^{-3}$	$7.00 \cdot 10^{-3}$	842 s

the speed and the values of the time on the front; i.e., Step 5 in the algorithm is computed only on $F^n \cap V(NA^n)$ and each ΔT time step on all of the nodes of F^n.

By measuring the amount of inhomogeneity as before, we still get infinity for the coefficient, since also in this test there are points in the front with zero speed.

Table 2 shows an increase of the CPU time for the GFMM method with respect to the previous test. This was expected since we update all of the values of $\tilde{u}$ on the fronts after every time interval ΔT. The FD scheme is still more accurate, but in terms of CPU time the GFMM method wins since it is 6 times faster (in the last test) than the FD scheme. All tables show that for smooth speed the GFMM algorithm has approximately order of convergence 1.

All of the tests have been computed on a laptop with a Intel Centrino Duo processor.

Test 2: *Propagation of a circle.* We choose as initial data a circle $P(x, 0) = x_1^2 + x_2^2 - 1$ and then as representing function:

$$\theta(x,0) = \begin{cases} 1 & x_1^2 + x_2^2 - 1 < 0, \\ -1 & \text{otherwise.} \end{cases} \tag{6.3}$$

We choose as velocity $c(x,t) = 0.1t - x_1$. We have proved that a circle propagating with linear speed stays a circle. By applying the result of the previous section, we obtain that $P(x,t) = (x_1 - x_{0,1}(t))^2 + (x_2 - x_{0,2}(t))^2 - R(t)^2$ has coefficients solving the following ODE:

$$\begin{cases} \dot{x}_{0,1}(t) = -R(t), \\ x_{0,1}(0) = 0, \end{cases} \quad \begin{cases} \dot{x}_{0,2}(t) = 0, \\ x_{0,2}(0) = 0, \end{cases} \quad \begin{cases} \dot{R}(t) = -x_{0,1}(t) + 0.1t, \\ R(0) = 1. \end{cases}$$

By solving, we obtain $x_{0,1}(t) = 1/20(2t + 11(\exp(-t) - \exp(t))$ and $R(t) = 1/20(-2 + 11(\exp(t) + \exp(-t)))$.

We compute the discrete solution in the numerical domain $D = [-2, 2] \times [-2, 2]$. The amount of inhomogeneity as in the previous test is infinity, because again in this test there are points in the front with zero speed.

We evaluate the error at the final time $T = 0.5$ by comparing the areas $\mathcal{A}(\cdot)$ of the sets Ω_T^+ and Ω_m^+ and computing the approximated Hausdorff distance between the exact and approximated fronts, as defined in the previous test.

Table 3 shows the error for the tests run with 50, 100, 200, 400 and the number of nodes for each side of the square domain. As can be seen in the table, the order of convergence is approximately 1, which seems to be a very good result for an algorithm which adopts a discontinuous representation of the front.

TABLE 3
Area and Hausdorff distances for Test 2 with GFMM.

Δx	$\lvert\mathcal{A}(\Omega_T^+) - \mathcal{A}(\Omega_m^+)\rvert$	$\mathcal{H}(\mathcal{C}, \tilde{\mathcal{C}})$
0.08	$1.26 \cdot 10^{-1}$	$8.52 \cdot 10^{-1}$
0.04	$6.56 \cdot 10^{-2}$	$4.42 \cdot 10^{-2}$
0.02	$3.34 \cdot 10^{-2}$	$2.41 \cdot 10^{-2}$
0.01	$1.41 \cdot 10^{-2}$	$1.24 \cdot 10^{-2}$

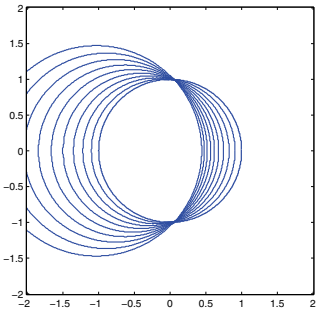


FIG. 5. *A propagating circle.*

Figure 5 shows the interface between $\{\theta^\varepsilon = 1\}$ and $\{\theta^\varepsilon = -1\}$ at each time interval 0.1 obtained with space step 0.01; the circle is expanding, and its center is propagating on the left.

Test 3: *Comparison between the FMM and GFMM algorithms.* When the evolution is monotone (with time-independent velocity), i.e., $c(x) > 0$, there exists a link between the evolutive and the stationary equation (see [12, 16]):

$$\begin{cases} c(x)|DT(x)| = 1 & x \in \Omega, \\ T(x) = 0 & x \in \partial\Omega. \end{cases}$$

In this case the discrete function u_I^n, computed by the GFMM algorithm, approximates the solution $T(x)$ outside the set Ω.

The two schemes, the FMM and the GFMM, are run in the case where the speed is $c(x,t) = 1$ with initial set Ω a circle centered in the origin with radius 0.5. The amount of inhomogeneity is obviously one, because in this test the speed is a constant.

For this choice of speed, the solution $T(x)$ corresponds to the distance function of the point x from the set Ω (see Figure 6).

We compare the two schemes by computing the errors in the $||\cdot||_\infty$ discrete norm:

$$\|T(x_I) - u_I\|_\infty \equiv \sup_{\{I:x_I \in D\}} |T(x_I) - u_I|.$$

As one can see in Table 4, the GFMM scheme produces in this particular case almost the same results as the FMM scheme (as implemented in the HJpack library [29]). The results are slightly different in particular because the time computed in the narrow band in the classical FMM uses not only the accepted points but also the points of the narrow band.

TABLE 4
Numerical errors for Test 3.

Δx	FMM	GFMM
0.08	$6.5 \cdot 10^{-2}$	$7.8 \cdot 10^{-2}$
0.04	$3.3 \cdot 10^{-2}$	$3.9 \cdot 10^{-2}$
0.02	$2.0 \cdot 10^{-2}$	$1.8 \cdot 10^{-2}$
0.01	$1.0 \cdot 10^{-2}$	$1.0 \cdot 10^{-2}$

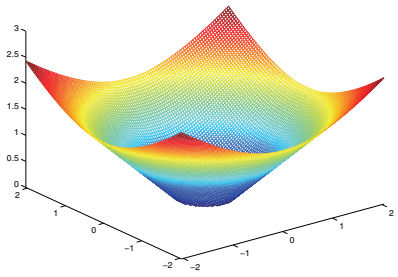


FIG. 6. *The discrete time u of a propagating circle with positive constant speed.*

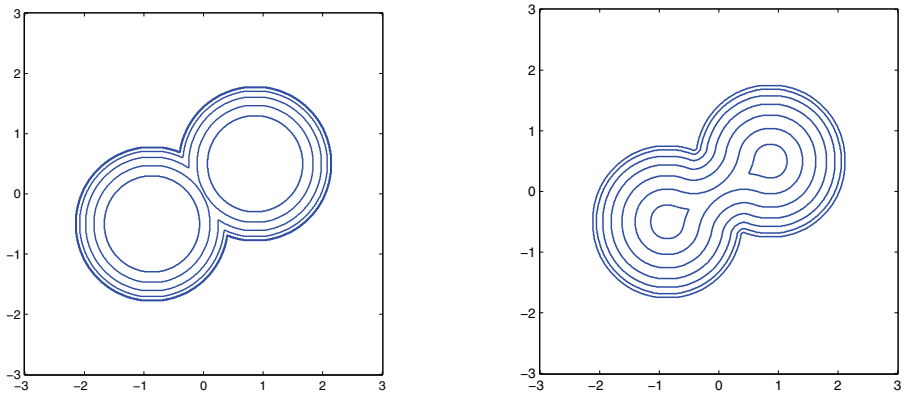


FIG. 7. *Two propagating circles.*

Test 4: *Two collapsing circles.* We choose as initial data two circles and as velocity $c(x,t) = 1 - t$. The amount of inhomogeneity is one, since the speed is not depending on space variables.

The two circles grow as far as the speed is positive. At $t = 1$, when the velocity changes sign, they start to decrease. Figure 7 on the left shows the interface between $\{\theta^\varepsilon = 1\}$ and $\{\theta^\varepsilon = -1\}$ obtained with $\Delta x = 0.01$ at each time interval 0.2 until $t = 1$, and Figure 7 on the right shows the interface between $\{\theta^\varepsilon = 1\}$ and $\{\theta^\varepsilon = -1\}$ at each time interval 0.2 for the time interval $[1.2, 2.4]$.

Acknowledgments. This work was supported by the project ACI 1025 (2003–2007) from the French Ministry of Research and by the Italian MIUR 2005 project

"Modellistica numerica per il calcolo scientifico ed applicazioni avanzate." The first author has been supported by INdAM (Istituto Nazionale di Alta Matematica). The authors thank Y. Le Bouar and A. Finel for fruitful discussions and A. Vladimirsky for valuable comments which have improved the citation of the prior works and the presentation of the manuscript.

REFERENCES

[1] D. Adalsteinsson and J. A. Sethian, *A fast level set method for propagating interfaces*, J. Comput. Phys., 118 (1995), pp. 269–277.
[2] O. Alvarez, E. Carlini, R. Monneau, and E. Rouy, *A convergent scheme for a non local Hamilton-Jacobi equation modelling dislocation dynamic*, Numer. Math., 104 (2006), pp. 413–444.
[3] O. Alvarez, P. Hoch, Y. Le Bouar, and R. Monneau, *Dislocation dynamics: Short time existence and uniqueness of the solution*, Arch. Ration. Mech. Anal., 181 (2006), pp. 449–504.
[4] G. Barles, *Solutions de viscosité des équations de Hamilton-Jacobi*, in Mathématiques et Applications, Vol. 17, Springer-Verlag, Paris, 1994.
[5] G. Barles and C. Georgelin, *A simple proof of convergence for an approximation scheme for computing motions by mean curvature*, SIAM J. Numer. Anal., 32 (1995), pp. 484–500.
[6] G. Barles, H. M. Soner, and P. E. Souganidis, *Front propagation and phase field theory*, SIAM J. Control Optim., 31 (1993), pp. 439–469.
[7] G. Barles and P. E. Souganidis, *Convergence of approximation schemes for fully nonlinear second order equations*, Asymptot. Anal., 4 (1991), pp. 271–283.
[8] E. Carlini, E. Cristiani, and N. Forcadel, *A non-monotone fast marching scheme for a Hamilton-Jacobi equation modeling dislocation dynamics*, in Numerical Mathematics and Advanced Applications - ENUMATH2005, A. Bermudez de Castro, D. Gomez, P. Quintela, and P. Salgado, eds., Springer-Verlag, New York, 2006, pp. 724–731.
[9] D. L. Chopp, *Handbook of Biomedical Image Analysis*, Vol. 1, Kluwer Academic, Norwell, MA, 2005.
[10] M. G. Crandall, H. Ishii, and P. L. Lions, *User's guide to viscosity solutions of second order partial differential equations*, Bull. Amer. Math. Soc. (N.S.), 27 (1992), pp. 1–67.
[11] E. Cristiani and M. Falcone, *Fast semi-Lagrangian schemes for the Eikonal equation and applications*, SIAM J. Numer. Anal., 45 (2007), pp. 1979–2011.
[12] M. Falcone, *The minimum time problem and its applications to front propagation*, in Motion by Mean Curvature and Related Topics, A. Visintin and G. Buttazzo, eds., De Gruyter Verlag, Berlin, 1994.
[13] R. Fedkiw and S. Osher, *Level Set Methods and Dynamic Implicit Surfaces*, Appl. Math. Sci. 153, Springer-Verlag, New York, 2003.
[14] J. Gomes and O. Faugeras, *Reconciling distance functions and level sets*, J. Visual Communic. and Imag. Representation, 11 (2000), pp. 209–223.
[15] S. Kim, *An O(N) level set method for eikonal equations*, SIAM J. Sci. Comput., 22 (2001), pp. 2178–2193.
[16] S. Osher, *A level set formulation for the solution of the Dirichlet problem for Hamilton-Jacobi equations*, SIAM J. Math. Anal., 24 (1993), pp. 1145–1152.
[17] S. Osher and J. Sethian, *Fronts propagating with curvature-dependent speed: Algorithms based on Hamilton-Jacobi formulations*, J. Comput. Phys., 79 (1988), pp. 12–49.
[18] D. Peng, B. Merriman, S. Osher, H. Zhao, and M. Kang, *A PDE-based fast local level set method*, J. Comput. Phys., 155 (1999), pp. 410–438.
[19] D. Rodney, Y. Le Bouar, and A. Finel, *Phase field methods and dislocations*, Acta Materiala, 51 (2003), pp. 17–31.
[20] J. A. Sethian, *Fast marching methods*, SIAM Rev., 41 (1999), pp. 199–235.
[21] J. A. Sethian, *Level Set Methods, Evolving Interfaces in Geometry, Fluid Mechanics, Computer Vision, and Materials Science*, Cambridge University Press, London, 1996.
[22] J. A. Sethian, *Level Set Methods and Fast Marching Methods: Evolving Interfaces in Computational Geometry, Fluid Mechanics, Computer Vision and Materials Science*, Cambridge University Press, London, 1999.
[23] J. A. Sethian, *A fast marching level set method for monotonically advancing fronts*, Proc. Natl. Acad. Sci. USA, 93 (1996), pp. 1591–1595.
[24] J. A. Sethian and A. Vladimirsky, *Ordered upwind methods for static Hamilton-Jacobi equations*, Proc. Natl. Acad. Sci. USA, 98 (2001), pp. 11069–11074.

[25] J. A. Sethian and A. Vladimirsky, *Ordered upwind methods for static Hamilton–Jacobi equations: Theory and algorithms*, SIAM J. Numer. Anal., 41 (2003), pp. 325–363.
[26] M. Sussman and E. Fatemi, *An efficient, interface-preserving level set redistancing algorithm and its application to interfacial incompressible fluid flow*, SIAM J. Sci. Comput., 20 (1999), pp. 1165–1191.
[27] J. N. Tsitsiklis, *Efficient algorithms for globally optimal trajectories*, IEEE Trans. Automat. Control, 40 (1995), pp. 1528–1538.
[28] A. Vladimirsky, *Static PDEs for time-dependent control problems*, Interfaces Free Bound., 8/3 (2006), pp. 281–300.
[29] http://www.caspur.it/hjpack/.

SIAM J. NUMER. ANAL.
Vol. 46, No. 6, pp. 2953–2977
© 2008 Society for Industrial and Applied Mathematics

ENO MULTIRESOLUTION SCHEMES WITH GENERAL DISCRETIZATIONS*

PASCAL GETREUER† AND FRANÇOIS G. MEYER‡

Abstract. Harten's framework is a nonlinear generalization of the wavelet framework. Previously, the choice of discretization (scaling function) in Harten multiresolution schemes has been limited to point-value, cell-average, and hat-based discretization. This paper shows how to construct multiresolution schemes consistent with Harten's framework for a variety of discretizations. The construction here begins with the discrete operators and deduces the corresponding continuous operators, reversing the order of the usual approach. This construction yields as a special case essentially nonoscillatory (ENO) multiresolution schemes for any order of spline discretization and also has the flexibility to define multiresolution schemes with nonspline discretizations. An error-control strategy is also developed.

Key words. ENO interpolation, multiscale decomposition, wavelets

AMS subject classifications. 41A05, 65D05, 65N55, 42C40

DOI. 10.1137/060663763

1. Introduction.

Harten's multiresolution framework [5] generalizes wavelet transforms. Harten's framework requires that the decimation (coarsening) operator be linear, but the prediction (refining) operator may be nonlinear.

One effective choice for this prediction operator is essentially nonoscillatory (ENO) interpolation, an edge-adaptive interpolation technique originating in the numerical methods of hyperbolic conservation laws [12]. ENO multiresolution schemes based on ENO interpolation achieve efficient, edge-adaptive decompositions and are especially successful on piecewise-smooth functions, as shown by the numerical experiments in [2].

1.1. Earlier work.

Much work on nonlinear multiresolution in the wavelet literature has been inspired by Sweldens's wavelet lifting scheme [19]. In [8], Claypoole et al. used a nonlinear prediction stage to create an edge-adaptive transform. Heijmans and Goutsias [15] developed the general morphological wavelets framework for constructing nonlinear wavelets.

Many have worked to develop Harten multiresolution, including:

- Harten [13, 14] developed the foundation of Harten's framework and ENO multiresolution schemes.
- Cohen, Dyn, and Mateï [10] studied the convergence of ENO prediction schemes and other properties of recursive prediction using the machinery of subdivision refinement schemes.
- Sonar, Iske, Cecil, Qian, and Osher [7, 16, 18] proposed the extensions of ENO beyond one dimension using radial basis functions. Cohen and Mateï [9] extended ENO to two dimensions with rotating step functions.

*Received by the editors June 27, 2006; accepted for publication (in revised form) March 28, 2008; published electronically August 20, 2008. This work was supported by National Science Foundation VIGRE grant DMS-9810751.

http://www.siam.org/journals/sinum/46-6/66376.html

†Department of Mathematics, University of California Los Angeles, Los Angeles, CA 90095 (getreuer@gmail.com).

‡Department of Electrical Engineering, University of Colorado at Boulder, Boulder, CO 80309 (francois.meyer@colorado.edu).

- Aràndiga, Donat, and Harten [6] constructed an ENO transform satisfying Harten's framework using hat-based discretization.

1.2. This paper's contribution. Harten's framework is in principle more general than the wavelet framework, yet most development so far has focused on the point-value and cell-average discretizations (analogous to the lazy and Haar wavelets). Aràndiga, Donat, and Harten [6] developed a third, the hat-based discretization. In wavelet nomenclature, only three possible scaling functions have been thoroughly considered.

The goal of this paper is to extend these constructions to all discretizations where the decimation filter is finite impulse response (FIR). This class of discretizations includes those induced by FIR wavelets as well as the three previously developed discretizations. Previously, Harten constructions start in the continuous domain by defining a discretization operator $\mathcal{D}_k$ and a reconstruction operator $\mathcal{R}_k$. Approaching the construction from the discrete domain, however, many operators can be defined without $\mathcal{D}_k$ and $\mathcal{R}_k$. We construct the decimation and prediction operators first and then deduce $\mathcal{D}_k$ and a choice of $\mathcal{R}_k$.

1.3. Outline. Section 2 reviews the construction of Harten's framework. Section 3 describes ENO interpolation (section 3.1), point-value discretization ENO schemes (section 3.2), and cell-average and hat-based discretization ENO schemes (section 3.3). Sections 4 and 5 develop for any FIR decimation filter h,

- a valid prediction operator (section 4.3),
- detail operators for a nonredundant scheme (section 4.1),
- $\mathcal{D}_k$ and $\mathcal{R}_k$ (section 4.4),
- an error-control strategy (section 5).

Section 6 develops two examples of this construction: Schemes with any spline discretization (section 6.1) and an example of nonspline discretization (section 6.2). Section 7 is the conclusion.

1.4. Notation. Define a *discrete signal* as any real or complex-valued sequence $v = (v_n)_{n\in\mathbb{Z}}$. An FIR filter is a discrete signal that is nonzero for finitely many n. Define its bilateral Z-transform $v(z)$,

$$v(z) \stackrel{\text{def}}{=} \mathcal{Z}\{v\}(z) = \sum_{n\in\mathbb{Z}} v_n z^{-n}. \tag{1.1}$$

Define dyadic downsampling and upsampling on v as

$$(\downarrow v)_n = v_{2n}, \qquad (\uparrow v)_n = \begin{cases} v_{n/2} & \text{if } n \text{ even}, \\ 0 & \text{if } n \text{ odd}. \end{cases}$$

The Z-transform of $\uparrow\downarrow v$ is

$$\tfrac{1}{2}\big(v(z) + v(-z)\big). \tag{1.2}$$

Denote by $\hat{\cdot}$ the Fourier transform,

$$\hat{f}(\omega) \stackrel{\text{def}}{=} \int_{-\infty}^{+\infty} f(x)\, \mathrm{e}^{-i\omega x}\, dx.$$

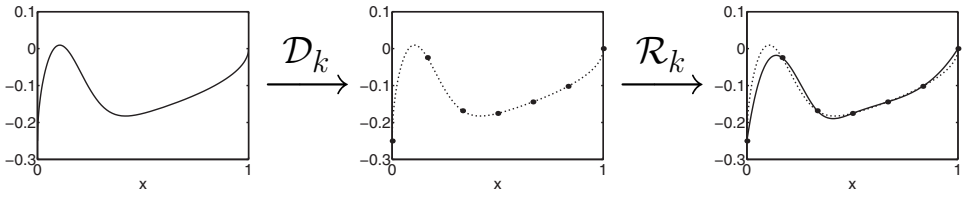


FIG. 2.1. *Left:* $f(x) = \mathrm{e}^{-25x^2}\sqrt{x} - \frac{1}{4}\sqrt{1-x}$. *Center:* $\mathcal{D}_k f$ *with point-value discretization,* $(\mathcal{D}_k f)_n = f(n/6)$. *Right:* $\mathcal{R}_k \mathcal{D}_k f$, *where* $\mathcal{R}_k$ *is cubic spline interpolation.*

2. Harten's framework. Harten's framework [5] builds from two operators: *Discretization* $\mathcal{D}_k$ and *reconstruction* $\mathcal{R}_k$. These operators map between functions and discrete signals at resolution k. The index k denotes the resolution level, where increasing k implies finer resolution.

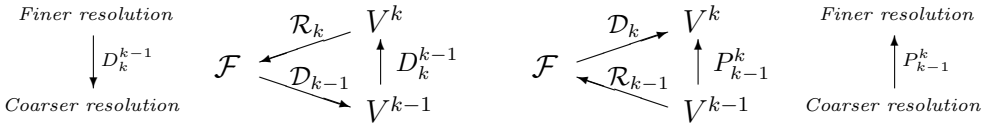
Let $\mathcal{F}$ be space of functions, and let V^k be a vector space of discrete signals. Discretization $\mathcal{D}_k : \mathcal{F} \to V^k$ is a *linear* operator mapping $f \in \mathcal{F}$ to a discrete signal $v^k = \mathcal{D}_k f$. Reconstruction $\mathcal{R}_k : V^k \to \mathcal{F}$ is a (generally) *nonlinear* operator mapping discrete signals to functions. This nonlinearity is where Harten's framework is more general than the wavelet framework. Figure 2.1 shows an example of discretization and reconstruction, where $\mathcal{D}_k$ is point-value discretization and $\mathcal{R}_k$ is cubic spline interpolation.

Harten's framework requires that the discretization and reconstruction operators satisfy a consistency relationship

$$\mathcal{D}_k \mathcal{R}_k = I_{V^k}, \tag{2.1}$$

where I_{V^K} denotes identity on V^k. The reconstruction $\mathcal{R}_k v^k$ must be consistent with the discrete information in v^k.

To construct a multiresolution scheme, define a decimation operator $D_k^{k-1} = \mathcal{D}_{k-1}\mathcal{R}_k$ and a prediction operator $P_{k-1}^k = \mathcal{D}_k \mathcal{R}_{k-1}$ as follows:



Decimation reduces the discrete signal v^k to v^{k-1}; $v^{k-1} = D_k^{k-1} v^k$. The prediction operator predicts v^k from v^{k-1}; $P_{k-1}^k v^{k-1}$ is an approximation of v^k. Define the prediction error as

$$e^k = v^k - P_{k-1}^k v^{k-1}.$$

The operators D_k^{k-1} and P_{k-1}^k can be used to construct a multiresolution pyramid (see Figure 2.2). If v^k is the input, then one stage of decomposition outputs the

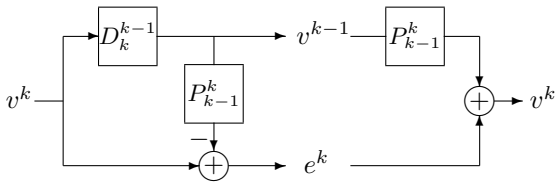


FIG. 2.2. *Pyramid decomposition under Harten's framework.*

decimated signal v^{k-1} and the prediction error e^k. Because e^k is at the same sample rate as v^k, the decomposition is 50% oversampled, that is, (v^{k-1}, e^k) redundantly represents v^k. This single decomposition stage is iterated on the decimated signal for a multiresolution representation $(v^{k-L}, e^{k-L+1}, \dots, e^k)$.

The discretization sequence $(\mathcal{D}_k)$ is *nested* if

$$\mathcal{D}_k f = 0 \;\Rightarrow\; \mathcal{D}_{k-1} f = 0 \qquad \text{for all } f \in \mathcal{F},\ k \in \mathbb{Z}. \tag{2.2}$$

If $(\mathcal{D}_k)$ is nested, then D^k_{k-1} has no dependence on $\mathcal{R}_k$ despite its definition, and it must be a linear operator [5]. This property, along with the consistency requirement (2.1), implies that $\mathcal{D}_m \mathcal{R}_k \mathcal{D}_k = \mathcal{D}_m$ for $m \le k$ and a discrete analogy of the consistency relationship

$$D^{k-1}_k P^k_{k-1} = \mathcal{D}_{k-1}\mathcal{R}_k\mathcal{D}_k\mathcal{R}_{k-1} = \mathcal{D}_{k-1}\mathcal{R}_{k-1} = I_{V^{k-1}}. \tag{2.3}$$

Consequently, e^k is in the nullspace of D^{k-1}_k:

$$\begin{aligned} D^{k-1}_k e^k &= D^{k-1}_k \left(I_{V^k} - P^k_{k-1} D^{k-1}_k\right) v^k \\ &= \left(D^{k-1}_k - D^{k-1}_k (P^k_{k-1} D^{k-1}_k)\right) v^k \;=\; 0. \end{aligned} \tag{2.4}$$

By (2.4), it is possible to design a nonredundant (critically sampled) multiresolution decomposition. Let G_k be a detail encoder such that $d^k = G_k e^k$ is at half the sample rate of e^k, and let $\tilde{G}_k$ be the corresponding decoder such that $\tilde{G}_k G_k e^k = e^k$ for any e^k in the nullspace of D^{k-1}_k. Then (v^{k-1}, d^k) is a nonredundant representation of v^k. This single stage is iterated on the decimated signal for a multiresolution representation $(v^{k-L}, d^{k-L+1}, \dots, d^k)$.

In summary, a multiresolution scheme within Harten's framework is characterized by six operators: The fundamental discretization and reconstruction operators $\mathcal{D}_k$ and $\mathcal{R}_k$, the decimation operator D^{k-1}_k, the prediction operator P^k_{k-1}, and the detail operators G_k and $\tilde{G}_k$. The next section shows how ENO multiresolution schemes fit into the Harten framework.

3. ENO multiresolution schemes. ENO multiresolution schemes are Harten multiresolution schemes where the reconstruction operator is ENO interpolation. ENO interpolation is an effective edge-adaptive strategy for piecewise polynomial interpolation (essentially) without oscillatory artifacts (see Figure 3.1).

3.1. ENO interpolation. Let $v_n = f(x_n)$ be samples from an underlying function f. ENO interpolation approximates f from the point-values v by a piecewise polynomial model.

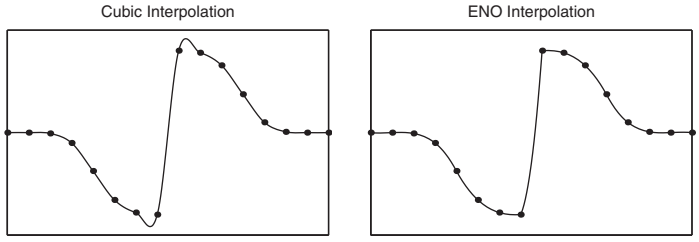


FIG. 3.1. *An interpolation of piecewise-smooth data. Cubic interpolation produces oscillations around the discontinuity, but ENO interpolation does not.*

On each subinterval $[x_{n-1}, x_n]$, a polynomial interpolant $q_n(x)$ is constructed based on a stencil $\mathcal{S}_n$ such that $q_n(x) = f(x)$ for all $x \in \mathcal{S}_n$. For example, cubic interpolation finds the cubic polynomial q_n satisfying the point-values at $\mathcal{S}_n = \{x_{n-2}, x_{n-1}, x_n, x_{n+1}\}$ to interpolate $[x_{n-1}, x_n]$. In ENO interpolation, the $\mathcal{S}_n$ are selected to adapt to the v^k.

The accuracy of such an interpolant depends heavily on the stencil. Let $\mathcal{S}$ be a stencil, and denote by $\widehat{\mathcal{S}}$ its convex hull. If $\mathcal{S}$ has $M+1$ points and $f \in C^{M+1}(\widehat{\mathcal{S}})$, the interpolation error is

$$f(x) - q(x) = \frac{f^{(M+1)}\big(\xi(x)\big)}{(M+1)!} \prod_{x_j \in \mathcal{S}} (x - x_j) \quad \text{for some } \xi(x) \in \widehat{\mathcal{S}}. \tag{3.1}$$

If f is locally smooth, the interpolation error is small. However, if f has a jump or derivative singularity in $\widehat{\mathcal{S}}$, the error can be much greater [5].

ENO interpolation attempts to choose stencils that do not cross jumps or derivative discontinuities. To construct an interpolant $q_n(x)$ of degree M, consider the stencils $\mathcal{S}_n = \{x_{k_n-1}, \ldots, x_{k_n+M-1}\}$, $n - M + 1 \le k_n \le n$. Each stencil has $M+1$ points and includes x_{n-1} and x_n. The interpolation error associated with a particular stencil is estimated with the Mth-order divided difference of the stencil samples $f[\mathcal{S}]$:

$$|f[\mathcal{S}]| = \begin{cases} 0, & f \text{ is locally polynomial,} \\ O([f^{(p)}])/h^{M-p}, & \text{the stencil crosses a discontinuity in } f^{(p)}, \\ O(\|f^{(M)}\|), & \text{otherwise,} \end{cases} \tag{3.2}$$

where $[f^{(p)}]$ is the size of the jump and $\|f^{(M)}\|$ is the max-norm of $f^{(M)}$ over $\widehat{\mathcal{S}}$. For example, if $M = 2$, (3.2) has the form

$$|f[\mathcal{S}_i]| = \frac{1}{2h^2} \left| f(x_{k-1}) - 2f(x_k) + f(x_{k+1}) \right|, \quad k = i-1 \text{ or } i.$$

This estimate distinguishes between intervals where f is locally linear and intervals containing jumps or first derivative discontinuities. In general, the Mth-order error estimate can detect discontinuities in up to the $M-1$ derivative.

In [12], Harten et al. consider two methods for choosing the stencil shifts k_n:

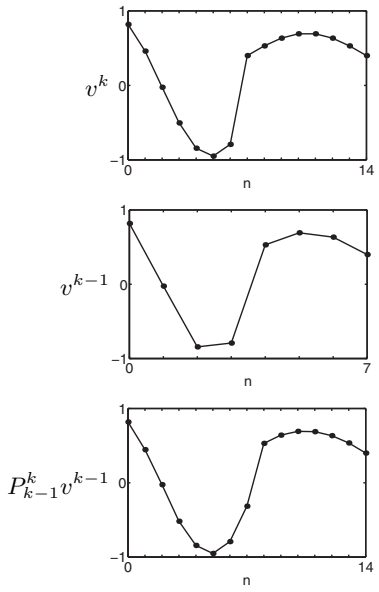
Hierarchical stencil selection

For each n
  $k_n := n$
  for $j = 0, \ldots, M-2$
    if $|f[x_{k_n-2}, \ldots, x_{k_n+j}]| < |f[x_{k_n-1}, \ldots, x_{k_n+j+1}]|$, **then** $k_n := k_n - 1$
  end.

Nonhierarchical stencil selection

For each n, choose k_n such that
$$|f[x_{k_n-1}, \ldots, x_{k_n+M-1}]| = \min_{n-M+1 \le k \le n} |f[x_{k-1}, \ldots, x_{k+M-1}]|.$$

The hierarchical method has the disadvantage that it can produce singularity-crossing stencils for discontinuities in f'' or higher derivatives regardless of M. The nonhierarchical method avoids this problem, but it is biased by $f^{(M)}$ [5, 12]. The hierarchical method is usually preferred; however, for the multiresolution constructions



v^k	v^{k-1}	$P^k_{k-1}v^{k-1}$	e^k	d^k
0.8185	0.8185	0.8185	0	
0.4603		0.4460	0.0143	0.0143
−0.0253	−0.0253	−0.0253	0	
−0.5039		−0.5177	0.0138	0.0138
−0.8430	−0.8430	−0.8430	0	
−0.9487		−0.9510	0.0024	0.0024
−0.7916	−0.7916	−0.7916	0	
0.4000		−0.3146	0.7146	0.7146
0.5302	0.5302	0.5302	0	
0.6345		0.6416	−0.0070	−0.0070
0.6925	0.6925	0.6925	0	
0.6925		0.6883	0.0042	0.0042
0.6345	0.6345	0.6345	0	
0.5302		0.5366	−0.0065	−0.0065
0.4000	0.4000	0.4000	0	

FIG. 3.2. *An example of point-value ENO decomposition using cubic ENO interpolation. The largest error* $e^k_7 = 0.7146$ *occurs in the interval containing the discontinuity.*

developed in this paper, the nonhierarchical method is often the better choice. This will be discussed further in section 6.1.

Define $\mathcal{I}^{\mathrm{ENO}}_k : V^k \to \mathcal{F}$ as ENO interpolation. For uniformly spaced x_n, define

$$(P^{\mathrm{ENO}}v)_n = (\mathcal{I}^{\mathrm{ENO}}_k v)(x_n/2).$$

P^{ENO} has the property $\downarrow P^{\mathrm{ENO}} = I$. We call an operator with this property an *interpolatory* operator.

3.2. Point-value discretization. ENO interpolation is the reconstruction operator in point-value ENO multiresolution. Define $\mathcal{D}_k$ as sampling point values:

$$v^k_n = (\mathcal{D}_k f)_n = f(x^k_n), \quad x^k_n = 2^{-k}n.$$

The interpolant $\mathcal{R}_k v^k = \mathcal{I}^{\mathrm{ENO}}_k v^k$ is equal to f at the point-values $f(x^k_n)$, so the consistency relationship $\mathcal{D}_k\mathcal{R}_k = I_{V^k}$ is trivially satisfied. The decimation operator is downsampling, $v^{k-1} = D^k_{k-1}v^k = \downarrow v^k$. The prediction operator $P^k_{k-1} = \mathcal{D}_k\mathcal{R}_{k-1}$ is ENO interpolatory prediction,

$$(P^k_{k-1}v^{k-1})_n = (\mathcal{I}^{\mathrm{ENO}}_k v^{k-1})(x^k_n) = (P^{\mathrm{ENO}}v^{k-1})_n.$$

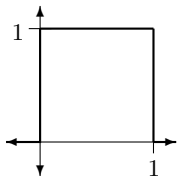
The prediction error $e^k = v^k - P^k_{k-1}v^{k-1}$ is nonzero only for odd n. For a nonredundant representation, the detail is encoded by keeping samples at odd n; $d^k_n = e^k_{2n+1}$ (see Figure 3.2).

3.3. Cell-average and hat-based discretizations. In most multiresolution schemes, smoothing is applied before downsampling to avoid aliasing in the coarser subbands. The cell-average and hat-based discretizations use local averages to achieve this smoothing. Suppose $f \in L^1_{\mathrm{loc}}$ is discretized by computing weighted averages

$$(\mathcal{D}_k f)_n = \int 2^k \phi(n - 2^k x)\, f(x)\, dx,$$

where ϕ is a compactly supported weighting function. Equivalently, $(\mathcal{D}_k f)_n = (\phi_k * f)(x_n^k)$, with $\phi_k(x) = 2^k\phi(2^k x)$. For cell-average discretization, ϕ is the Haar scaling function

$$\phi(x) = \begin{cases} 1, & 0 \le x < 1, \\ 0, & \text{otherwise.} \end{cases}$$



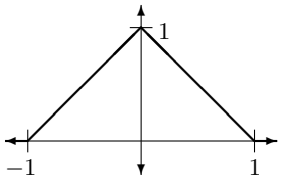
The discretization $(\mathcal{D}_k f)_n$ is the average over the cell $c_n^k = (\,2^{-k}(n-1), 2^{-k}n\,)$,

$$(\mathcal{D}_k f)_n = 2^k \int_{c_n^k} f(x)\,dx,$$

hence the name "cell-average." The dilation equation of the Haar scaling function $\phi(x) = \phi(2x)+\phi(2x-1)$ implies that $v_n^{k-1} = \frac{1}{2}v_{2n}^k + \frac{1}{2}v_{2n-1}^k$. Therefore, the decimation operator is $(D_k^{k-1}v^k)_n = \frac{1}{2}v_{2n}^k + \frac{1}{2}v_{2n-1}^k$.

For hat-based discretization, ϕ is the hat function

$$\phi(x) = \begin{cases} 1+x, & -1 \le x \le 0, \\ 1-x, & 0 \le x \le 1, \\ 0, & \text{otherwise.} \end{cases}$$



As with cell-average discretization, the dilation equation of ϕ determines D_k^{k-1}:

$$\begin{aligned} \phi(x) &= \tfrac{1}{2}\phi(2x-1) + \phi(2x) + \tfrac{1}{2}\phi(2x+1), \\ (D_k^{k-1}v^k)_n &= \tfrac{1}{4}v_{2n-1}^k + \tfrac{1}{2}v_{2n}^k + \tfrac{1}{4}v_{2n+1}^k. \end{aligned}$$

Consider the reconstruction operators for these discretization operators. To satisfy the consistency requirement $\mathcal{D}_k\mathcal{R}_k = I_{V^k}$, one approach is to modify the point-value ENO interpolation described in section 3.1 such that the reconstruction function attains averages rather than point-values [1].

Harten's approach to constructing $\mathcal{R}_k$ is *reconstruction via primitive function* [6, 14]. The idea is to reduce reconstruction to interpolation (that is, to point-value reconstruction). For cell-average discretization, the relationship between f and its primitive $\mathring{f}$ is

$$\mathring{f}(x) = \int_0^x f(y)\,dy, \qquad f(x) = \frac{d}{dx}\mathring{f}(x).$$

Set $\mathring{v}_{-1}^k = 0$, then the relationship between v^k and $\mathring{v}^k$ is

$$\mathring{v}_n^k = 2^{-k}\sum_{m=0}^{n} v_m^k, \quad v_n^k = 2^k(\mathring{v}_n^k - \mathring{v}_{n-1}^k), \quad n = 0, \ldots J_k - 1. \tag{3.3}$$

Let $\mathcal{I}_k$ be any interpolatory operator, for example, ENO interpolation. Define the reconstruction operator as

$$(\mathcal{R}_k v^k)(x) = \frac{d}{dx}\mathcal{I}_k\mathring{v}^k(x). \tag{3.4}$$

This $\mathcal{R}_k$ satisfies the consistency relationship (2.1): For any $v^k \in V^k$,

$$(\mathcal{D}_k \mathcal{R}_k v^k)_n = 2^k \int_{x^k_{n-1}}^{x^k_n} \frac{d}{dx}(\mathcal{I}_k \mathring{v}^k)\, dx = 2^k(\mathring{v}^k_n - \mathring{v}^k_{n-1}) = v^k_n.$$

The derivative in (3.4) should be interpreted in the weak sense. When $\mathcal{I}_k$ is ENO interpolation, $\mathcal{I}_k \mathring{v}^k$ is piecewise differentiable, so $\mathcal{R}_k v^k$ is piecewise continuous.

Using (3.3) to obtain $\mathring{v}^k$ from v^k, the prediction operator $P^k_{k-1} = \mathcal{D}_k \mathcal{R}_{k-1}$ is

$$(P^k_{k-1} v^{k-1})_n = 2^k \left[(\mathcal{I}_{k-1}\mathring{v}^{k-1})(x^k_n) - (\mathcal{I}_{k-1}\mathring{v}^{k-1})(x^k_{n-1})\right]. \tag{3.5}$$

The approach is similar for hat-based discretization. As developed in [6], the relationship between f and its primitive is

$$\mathring{f}(x) = \int_0^x \int_0^y f(z)\, dz\, dy, \qquad f(x) = \frac{d^2}{dx^2}\mathring{f}(x).$$

Setting $\mathring{v}^k_n = 0$ for $n < 0$, there is a bijection between v^k and $\mathring{v}^k$:

$$\mathring{v}^k_n = 4^{-k} \sum_{m=0}^{n} \sum_{j=0}^{m} v^k_j, \quad v^k_n = 4^k(\mathring{v}^k_{n+1} - 2\mathring{v}^k_n + \mathring{v}^k_{n-1}), \quad n = 0, \dots J_k - 1.$$

Define the reconstruction operator analogously to (3.4),

$$(\mathcal{R}_k v^k)(x) = \frac{d^2}{dx^2}\mathcal{I}_k \mathring{v}^k(x). \tag{3.6}$$

The second derivative should be interpreted in the weak sense. While the first derivative in (3.4) can produce jump discontinuities, (3.6) can produce Dirac measures.

It is true that (3.6) satisfies $\mathcal{D}_k \mathcal{R}_k = I_{V^k}$, but verifying this is more cumbersome than in the cell-average case; see [6]. The prediction operator $P^k_{k-1} = \mathcal{D}_k \mathcal{R}_{k-1}$ is

$$\begin{aligned}(P^k_{k-1} v^{k-1})_n = 4^k \big[&(\mathcal{I}_{k-1}\mathring{v}^{k-1})(x^k_{n+1}) \\ &-2(\mathcal{I}_{k-1}\mathring{v}^{k-1})(x^k_n) + (\mathcal{I}_{k-1}\mathring{v}^{k-1})(x^k_{n-1})\big].\end{aligned} \tag{3.7}$$

Detail encoder and decoder operators can be found using the property that $D^{k-1}_k e^k = 0$. For cell-average discretization, this is $e^k_{2n} = -e^k_{2n-1}$; thus a choice of detail encoder and decoder operators is

$$d^k_n = (G_k e^k)_n = e^k_{2n-1}, \qquad \begin{cases} e^k_{2n-1} = (\tilde{G}_k d^k)_{2n-1} = d^k_n, \\ e^k_{2n} = (\tilde{G}_k d^k)_{2n} = -d^k_n. \end{cases}$$

For hat-based discretization, $e^k_{2n} = -\frac{1}{2}(e^k_{2n-1} + e^k_{2n+1})$. As proposed in [6], the detail operators can be

$$d^k_n = (G_k e^k)_n = e^k_{2n-1}, \qquad \begin{cases} e^k_{2n-1} = (\tilde{G}_k d^k)_{2n-1} = d^k_n, \\ e^k_{2n} = (\tilde{G}_k d^k)_{2n} = -\frac{1}{2}(d^k_n + d^k_{n+1}). \end{cases}$$

Cell-average and hat-based discretizations are two members of a family of spline-based discretizations. Denote the order of a spline discretization by N, with cell-average discretization as $N = 1$ and hat-based discretization as $N = 2$. Point-value discretization fits into this classification as $N = 0$.

For general N, the weight function ϕ is the B-spline of order $N-1$, and $\mathring{f}$ is related to f through an Nth-order integral. The next sections develop the construction of the framework operators for a variety of discretizations, including spline-based discretization of any order (section 6.1).

4. Construction with general discretizations. Harten's framework is defined starting with a discretization operator $\mathcal{D}_k$ and a reconstruction operator $\mathcal{R}_k$ in the continuous domain. These operators are then used to define the decimation and prediction operators in the discrete domain. In this section, the reverse is done: Construction begins with the discrete operators.

Through this section, let h be an FIR filter, and define the decimation operator $D_k^{k-1}v^k = \downarrow(h * v^k)$.

4.1. Detail encoding. Considering h as a wavelet lowpass filter, let $\tilde{h}$, g, and $\tilde{g}$ be corresponding dual lowpass, highpass, and dual highpass wavelet filters, and define $H_k v = D_k^{k-1} v$, $G_k v = \downarrow(g * v)$, $\tilde{H}_k v = \tilde{h} *\uparrow v$, $\tilde{G}_k = \tilde{g} *\uparrow v$.

Suppose this wavelet transform is applied to e^k. Since $H_k e^k = D_k^{k-1} e^k = 0$ by (2.4), the lowpass component is zero (see Figure 4.1). Therefore, G_k and $\tilde{G}_k$ from this wavelet transform are a choice of detail operators for nonredundant decomposition.

Since G_k and $\tilde{G}_k$ are linear, the nonredundant decomposition shown in Figure 4.2 is equivalent to a one-stage lifting scheme (see Figure 4.3).

Remark 1. More generally, any linear operators G_k from V^k to V^{k-1} and $\tilde{H}_k$ and $\tilde{G}_k$ from V^{k-1} to V^k satisfying $\tilde{H}_k D_k^{k-1} + \tilde{G}_k G_k = I_{V^k}$ constitute a lifting scheme as shown in Figure 4.3.

To construct the detail operators for a given decimation filter h, a wavelet must be constructed with this same filter h as its lowpass filter. The perfect reconstruction of the wavelet scheme implies the conditions

$$h(z)\,\tilde{h}(z) \;+\; g(z)\,\tilde{g}(z) = 2,$$
$$h(z)\,\tilde{h}(-z) \;+\; g(z)\,\tilde{g}(-z) = 0.$$

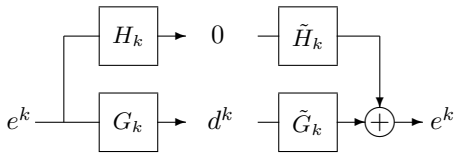


FIG. 4.1. *A wavelet transform of* e^k.

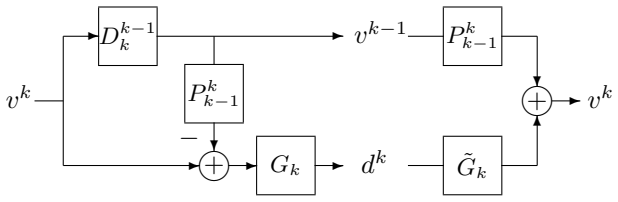


FIG. 4.2. *Nonredundant decomposition using detail encoding.*

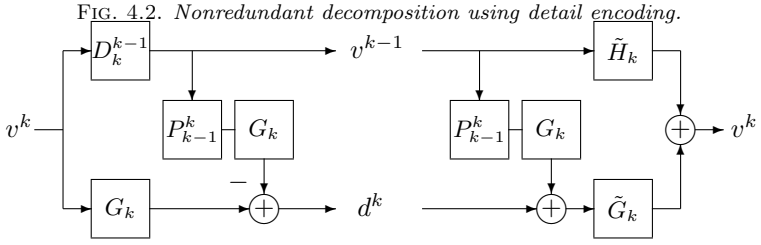


FIG. 4.3. *Nonredundant decomposition as a one-stage lifting scheme.*

If $g(z) = z^{-1}\,\tilde{h}(-z)$ and $\tilde{g}(z) = z\,h(-z)$, the second condition is satisfied. The first condition then reduces to $h(z)\,\tilde{h}(z) + h(-z)\,\tilde{h}(-z) = 2$, which is true if $(h * \tilde{h})_0 = 1$ and $(h * \tilde{h})_n = 0$ for nonzero even n. Thus $\tilde{h}$ must satisfy the linear system

$$\sum_j \tilde{h}_j\, h_{2n-j} = \begin{cases} 1 & \text{if } n = 0, \\ 0 & \text{if } n \neq 0 \end{cases} \qquad \text{for all } n. \tag{4.1}$$

For example, consider $h(z) = az^2 + bz + c + dz^{-1} + ez^{-2}$. If $\tilde{h}$ is assumed to have the form $\tilde{h}(z) = \tilde{h}_{-1}\,z + \tilde{h}_0 + \tilde{h}_1\,z^{-1}$, (4.1) becomes the matrix equation

$$\begin{bmatrix} b & a & 0 \\ d & c & b \\ 0 & e & d \end{bmatrix} \begin{bmatrix} \tilde{h}_{-1} \\ \tilde{h}_0 \\ \tilde{h}_1 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}.$$

Provided a nonzero determinant $\kappa = bcd - ad^2 - b^2e \neq 0$, the solution is $\tilde{h}(z) = \frac{1}{\kappa}(-adz + bd - bez^{-1})$ and

$$\begin{aligned} h(z) &= az^2 + bz + c + dz^{-1} + ez^{-2}, \quad & g(z) &= \frac{1}{\kappa}\left(ad + bdz^{-1} + bez^{-2}\right), \\ \tilde{h}(z) &= \frac{1}{\kappa}\left(-adz + bd - bez^{-1}\right), & \tilde{g}(z) &= az^3 - bz^2 + cz - d + ez^{-1}. \end{aligned} \tag{4.2}$$

It is also possible to select filters from existing wavelet families. Given a number of reconstruction vanishing moments N and decomposition vanishing moments $\tilde{N}$ where $N + \tilde{N}$ is even, the spline wavelet filters are

$$h(z) = z^{\lfloor N/2 \rfloor} \left(\frac{1 + z^{-1}}{2}\right)^N, \tag{4.3}$$

$$\tilde{h}(z) = 2z^{\lceil \tilde{N}/2 \rceil} \left(\frac{1 + z^{-1}}{2}\right)^{\tilde{N}} \sum_{n=0}^{M} \binom{M+n}{n} (-4)^{-n} (z - 2 + z^{-1})^n, \tag{4.4}$$

where $M = \frac{1}{2}(N + \tilde{N}) - 1$, $\lfloor \cdot \rfloor$ denotes the floor function, and $\lceil \cdot \rceil$ denotes the ceiling function [11]. The corresponding highpass filters are $g(z) = \frac{1}{2}z^{-1}\tilde{h}(-z)$ and $\tilde{g}(z) = 2zh(-z)$. The lowpass filter h is the decimation filter corresponding to the spline discretization of order N. If the prediction operator can predict locally polynomial signals, then vanishing moments in g are unnecessary, and $\tilde{N}$ should be 0 (or 1, if N is odd) to minimize the support of g.

Setting $N = 2$, $\tilde{N} = 0$ yields the hat filters

$$h(z) = z\left(\frac{1 + z^{-1}}{2}\right)^2 = \tfrac{1}{4}z + \tfrac{1}{2} + \tfrac{1}{4}z^{-1}, \qquad g(z) = z^{-1},$$

$$\tilde{h}(z) = 2\sum_{n=0}^{0} (-4)^{-n}(z - 2 + z^{-1})^n = 2, \qquad \tilde{g}(z) = -\tfrac{1}{2}z^2 + z - \tfrac{1}{2}.$$

These lead to the same encoding and decoding operators as before in section 3.3:

$$\begin{aligned} &\textit{Encoding} \qquad d^k = \downarrow(e^k * g) \quad \Leftrightarrow \quad d^k_n = e^k_{2n-1}, \\ &\textit{Decoding} \qquad e^k = (\uparrow d^k) * \tilde{g} \quad \Leftrightarrow \quad \begin{cases} e^k_{2n-1} = d^k_n, \\ e^k_{2n} = -\frac{1}{2}(d^k_n + d^k_{n+1}). \end{cases} \end{aligned}$$

Section 6.1 will expand upon this example to construct spline schemes for $N > 2$.

Remark 2. If h, $\tilde{h}$, g, and $\tilde{g}$ satisfy the perfect reconstruction conditions, then so do the scaled and shifted filters

$$h'(z) = \alpha z^{2j} h(z), \qquad g'(z) = \beta z^{2k} g(z),$$
$$\tilde{h}'(z) = \frac{1}{\alpha} z^{-2j} \tilde{h}(z), \qquad \tilde{g}'(z) = \frac{1}{\beta} z^{-2k} \tilde{g}(z).$$

Notice that the detail encoding and decoding operators are not uniquely determined by h. Consider the filter $h(z) = \frac{1}{8}z + \frac{3}{8} + \frac{3}{8}z^{-1} + \frac{1}{8}z^{-2}$. With the spline wavelet approach with $N = 3$ and $\tilde{N} = 1$, the filters g and $\tilde{g}$ are

$$g(z) = \tfrac{1}{4}z + \tfrac{3}{4} - \tfrac{3}{4}z^{-1} - \tfrac{1}{4}z^{-2}, \qquad \tilde{g}(z) = \tfrac{1}{4}z^2 - \tfrac{3}{4}z + \tfrac{3}{4} - \tfrac{1}{4}z^{-1}.$$

Using (4.2), the filters are

$$g(z) = 3z^{-1} + z^{-2}, \qquad \tilde{g}(z) = -\tfrac{1}{8}z^2 + \tfrac{3}{8}z - \tfrac{3}{8} + \tfrac{1}{8}z^{-1}.$$

There is a lot of flexibility in designing the detail operators. Harten's framework does not even require that they are linear, only that $\tilde{G}_k G_k e^k = e^k$ for all $e^k \in \mathcal{N}(D_k^{k-1})$.

4.2. Diagonalizing the decimation operator. Following the "via primitive" technique (section 3.3), our next step is to choose maps under which D_k^{k-1} becomes pure downsampling.

THEOREM 4.1. *Suppose $D_k^{k-1} : V^k \to V^{k-1}$ is any linear and surjective operator, then there exists bijections T_k on V^k and $\tilde{T}_{k-1}$ on V^{k-1} such that*

$$\tilde{T}_{k-1}^{-1} D_k^{k-1} T_k = \downarrow . \tag{4.5}$$

For any such maps $\tilde{T}_{k-1}$ and T_k, P_{k-1}^k is a consistent prediction operator if and only if $T_k^{-1} P_{k-1}^k \tilde{T}_{k-1}$ is an interpolatory operator.

Proof. One way to construct T_k and $\tilde{T}_{k-1}$ is to interpret D_k^{k-1} as a $J_{k-1} \times J_k$ matrix, and let $D_k^{k-1} = U\Sigma V^*$ be a singular value decomposition (SVD). Since D_k^{k-1} is surjective, the singular values $\sigma_1, \ldots, \sigma_{J_{k-1}}$ are all positive. Let S be the $J_{k-1} \times J_{k-1}$ diagonal matrix $S_{ii} = \sigma_i$, and let P be the $J_k \times J_k$ permutation matrix such that $S^{-1}\Sigma P = \left[I \,\middle|\, 0 \right] P = \downarrow$. Then

$$D_k^{k-1} \;=\; US(S^{-1}\Sigma P)P^* V^* = US \downarrow (VP)^*.$$

So a choice for the maps is $T_k = VP$ and $\tilde{T}_{k-1} = US$.

For any interpolatory operator Q_{k-1}^k, the prediction operator

$$P_{k-1}^k = T_k Q_{k-1}^k \tilde{T}_{k-1}^{-1} \tag{4.6}$$

is consistent: $D_{k-1}^k T_k Q_{k-1}^k \tilde{T}_{k-1}^{-1} = \tilde{T}_{k-1} \downarrow Q_{k-1}^k \tilde{T}_{k-1}^{-1} = I_{V^{k-1}}$. Conversely, if P_{k-1}^k is consistent, $\downarrow T_k^{-1} P_{k-1}^k \tilde{T}_{k-1} = \tilde{T}_{k-1}^{-1} D_k^{k-1} P_{k-1}^k \tilde{T}_{k-1} = I_{V^{k-1}}$. Thus P_{k-1}^k is a consistent prediction operator if and only if $T_k^{-1} P_{k-1}^k \tilde{T}_{k-1}$ is interpolatory. ∎

Remark 3. Furthermore, if an SVD of D_k^{k-1} is available, then $E_k v = [0 | I_{J_{k-1} \times J_{k-1}}] V^* v$, and $F_k d = E_k^* d$ are valid detail operators.

Given $\tilde{H}_k$, G_k, and $\tilde{G}_k$ such that $\tilde{H}_k D_k^{k-1} + \tilde{G}_k G_k = I_{V^k}$, a practical choice for the diagonalization maps is

$$T_k = \left[\tilde{H}_k \; \tilde{G}_k \right] P^*, \qquad \tilde{T}_{k-1} = I_{V^{k-1}}, \tag{4.7}$$

where P is the permutation matrix in the proof of Theorem 4.1. Then prediction takes the form $P_{k-1}^k v^{k-1} = \tilde{H}_k v^{k-1} + \tilde{G}_k w$, where w may be any element of V^{k-1}.

Moreover, (4.7) yields a strategy for correcting the consistency of a nonconsistent prediction operator. Let $\hat{P}_{k-1}^k : V^{k-1} \to V^k$ be any operator, then

$$P_{k-1}^k v^{k-1} = \tilde{H}_k v^{k-1} + \tilde{G}_k G_k \hat{P}_{k-1}^k v^{k-1} \tag{4.8}$$

is consistent and if $\hat{P}_{k-1}^k$ is already consistent, then $P_{k-1}^k = \hat{P}_{k-1}^k$. Such a prediction operator may be efficiently implemented in lifting scheme form as in Figure 4.3:

$$\textit{Encoding} \begin{cases} v^{k-1} = D_k^{k-1} v^k, \\ d^k = G_k v^k - G_k \hat{P}_{k-1}^k v^{k-1}, \end{cases} \qquad \textit{Decoding} \begin{cases} \tilde{d}^k = d^k + G_k \hat{P}_{k-1}^k v^{k-1}, \\ v^k = \tilde{H}_k v^{k-1} + \tilde{G}_k \tilde{d}^k. \end{cases}$$

The maps T_k and $\tilde{T}_{k-1}$ diagonalize D_k^{k-1} to the operation of pure downsampling. In the cell-average and hat-based constructions, these maps are seen as a relationship between the original signal v^{k-1} and its primitive $\mathring{v}^{k-1}$:

$$v^{k-1} \xrightarrow{\tilde{T}_{k-1}^{-1}} \mathring{v}^{k-1}, \qquad \mathring{v}^k \xrightarrow{T_k} v^k.$$

At least for spline discretizations, there are diagonalization maps such that $\tilde{T}_k = T_k$. However, $\tilde{T}_k$ and T_k need not be equal in general.

4.3. Prediction for convolution-based decimation. The preceding discussion applies to constructions with any linear and surjective D_{k-1}^k. In the rest of our construction we restrict to decimation of the form $D_{k-1}^k v^k = \downarrow(h * v^k)$, where h is an FIR filter.

In both the cell-average and hat-based constructions, the relationship between v^k and $\mathring{v}^k$ is linear shift-invariant, which suggests diagonalization maps of the form

$$T_k \mathring{v}^k = a * \mathring{v}^k, \quad \tilde{T}_{k-1}^{-1} v^{k-1} = b * v^{k-1}.$$

Moreover, from a design point of view, this is a convenient choice of maps since if Q_{k-1}^k is shift-invariant, then so is P_{k-1}^k.

Ignoring boundary issues for the moment, consider the scheme on the infinite grid. Discrete consistency $D_k^{k-1} P_{k-1}^k = I_{V^{k-1}}$ can be written using (1.2) as

$$\tfrac{1}{2}\left[h(z)\mathcal{Z}\{P_{k-1}^k v^{k-1}\}(z) + h(-z)\mathcal{Z}\{P_{k-1}^k v^{k-1}\}(-z)\right] = v^{k-1}(z^2).$$

Consider $P_{k-1}^k v^{k-1} = a * Q_{k-1}^k (b * v^{k-1})$, with $a(z) = h(-z)$. Then

$$\begin{aligned} \tfrac{1}{2}\left[h(z)h(-z)\mathcal{Z}\{Q_{k-1}^k(b * v^{k-1})\}(z) + h(-z)h(z)\mathcal{Z}\{Q_{k-1}^k(b * v^{k-1})\}(-z)\right] &= v^{k-1}(z^2) \\ h(z)h(-z)\tfrac{1}{2}\left[\mathcal{Z}\{Q_{k-1}^k(b * v^{k-1})\}(z) + \mathcal{Z}\{Q_{k-1}^k(b * v^{k-1})\}(-z)\right] &= v^{k-1}(z^2). \end{aligned}$$

Let $\mathring{v}^{k-1} = b * v^{k-1}$, then since Q_{k-1}^k is interpolatory,

$$\tfrac{1}{2}\left[\mathcal{Z}\{Q_{k-1}^k \mathring{v}^{k-1}\}(z) + \mathcal{Z}\{Q_{k-1}^k \mathring{v}^{k-1}\}(-z)\right] = \mathcal{Z}\{\uparrow\downarrow(Q_{k-1}^k \mathring{v}^{k-1})\}(z) = \mathring{v}^{k-1}(z^2).$$

Thus

$$h(z)h(-z)\left[b(z^2)v^{k-1}(z^2)\right] = v^{k-1}(z^2) \quad \Rightarrow \quad b(z^2) = \frac{1}{h(z)h(-z)}.$$

Since $h(z)h(-z)$ is even, $b(z)$ is well defined by this expression. Therefore, for any interpolatory operator Q^k_{k-1},

$$P^k_{k-1}v^{k-1} = a * Q^k_{k-1}(b * v^{k-1}) \tag{4.9}$$

is a consistent prediction operator with

$$a(z) = h(-z), \quad b(z^2) = \frac{1}{h(z)h(-z)}.$$

For finite-length signals, the consistency of the prediction operator (4.9) does not hold near the boundaries. The boundaries can be corrected using (4.8).

Remark 4. If c satisfies $c(z) = c(-z)$, the prediction operator $P^k_{k-1}v^{k-1} = a * Q^k_{k-1}(b * v^{k-1})$ with filters

$$b(z^2) = \frac{1}{c(z)h(z)h(-z)}, \qquad a(z) = c(z)h(-z)$$

is also consistent. In particular, b and a may be rescaled with $c(z) = \kappa$.

Beware that b is typically an unstable filter. If $h(z)$ has a zero at r, then $b(z)$ has a pole at r^2. Depending on the rate of unstable growth and the signal length, a direct implementation of (4.9) may yet be numerically tractable. Otherwise, (4.9) should be considered conceptually and then implemented using other diagonalization maps (for example, (4.7)).

This construction reproduces the cell-average and hat-based discretization strategies in section 3.3. Let $\mathcal{I}_k$ be an interpolatory operator, and define Q^k_{k-1} by $(Q^k_{k-1}v^{k-1})_n = (\mathcal{I}_{k-1}v^{k-1})(x^k_n)$. If h is the cell-average decimation filter $h(z) = \frac{1}{2} + \frac{1}{2}z^{-1}$ and $\kappa = 2^{k+1}$,

$$\begin{aligned} b(z^2) = \frac{1}{\kappa\, h(z)h(-z)} &= \frac{2^{1-k}}{1-z^{-2}} \qquad \Rightarrow \qquad b(z) = \frac{2^{1-k}}{1-z^{-1}}, \\ a(z) = \kappa\, h(-z) &= 2^k(1-z^{-1}). \end{aligned}$$

Convolution with b is cumulative summation, and convolution with a is backward differencing. Define

$$\tilde{v}^{k-1}_n = (b * v^{k-1})_n = 2^{1-k}\sum_{m=1}^{n} v^{k-1}_m,$$

then the prediction operator is

$$\begin{aligned} (P^k_{k-1}v^{k-1})_n = (a * Q^k_{k-1}\tilde{v}^{k-1})_n &= 2^k\left[(Q^k_{k-1}\tilde{v}^{k-1})_n - (Q^k_{k-1}\tilde{v}^{k-1})_{n-1}\right] \\ &= 2^k\left[(\mathcal{I}_{k-1}\tilde{v}^{k-1})(x^k_n) - (\mathcal{I}_{k-1}\tilde{v}^{k-1})(x^k_{n-1})\right]. \end{aligned}$$

Compare this to the cell-average prediction described in section 3.3, which is

$$(P^k_{k-1}v^{k-1})_n = 2^k\left[(\mathcal{I}_{k-1}\mathring{v}^{k-1})(x^k_n) - (\mathcal{I}_{k-1}\mathring{v}^{k-1})(x^k_{n-1})\right], \quad \mathring{v}^{k-1}_n = 2^{1-k}\sum_{m=1}^{n} v^{k-1}_m.$$

Notice that $\tilde{v}^{k-1} = b*v^{k-1}$ has the same role as the primitive samples $\mathring{v}^{k-1}_n = \mathring{f}(x^{k-1}_n)$. Section 4.4 generalizes the notion of primitive through this connection.

For the hat-based prediction as in section 3.3, set $h(z) = \frac{1}{4}z + \frac{1}{2} + \frac{1}{4}z^{-1}$ and $\kappa = -4^{k+1}$:

$$b(z^2) = \frac{1}{\kappa\, h(z)h(-z)} = \frac{4^{1-k}z^{-2}}{(1-z^{-2})^2} \qquad \Rightarrow \qquad b(z) = \frac{4^{1-k}z^{-1}}{(1-z^{-1})^2},$$
$$a(z) = \kappa\, h(-z) = 4^k(z - 2 + z^{-1}).$$

The interpolatory operator Q^k_{k-1} must be sufficiently regular to compensate for applying the highpass filter a. For hat-based prediction, $a * Q^k_{k-1}\mathring{v}^{k-1}$ is the second difference of $Q^k_{k-1}\mathring{v}^{k-1}$, which has Dirac measures if Q^k_{k-1} is not smooth. One way to make a smoother version of Q^k_{k-1} is by defining

$$\tilde{Q}^k_{k-1}\mathring{v}^{k-1} = \operatorname*{arg\,min}_{q:\downarrow q = \mathring{v}^{k-1}} \left\|Q^k_{k-1}\mathring{v}^{k-1} - q\right\|_2^2 + \|\lambda \cdot (d * q)\|_2^2, \tag{4.10}$$

where d is a finite difference filter and λ_n weights the amount of regularization around point n. The minimization is equivalent to solving a sparse linear system.

Remark 5. There is no restriction that a prediction must be formulated in the same way for all scales k. The filters a and b may depend on k.

4.4. Operators $\mathcal{D}_k$ and $\mathcal{R}_k$. Knowing only the discrete operators D^{k-1}_k, P^k_{k-1}, G_k, and $\tilde{G}_k$ is sufficient for implementation. However, to complete Harten's framework, this section finds operators $\mathcal{D}_k$ and $\mathcal{R}_k$ consistent with the discrete operators. In this section, complications at the boundaries are ignored by working on the infinite grid.

Given a discrete signal g, define its delta sequence at resolution level k as

$$g_{\delta_k}(x) \stackrel{\text{def}}{=} \sum_{n\in\mathbb{Z}} g_n\, \delta(x - x^k_n), \quad x^k_n = 2^{-k}n.$$

Suppose that $h(z)$ has at least one root at $z = -1$ and is normalized as $h(1) = 1$. Analogous to the scaling function in wavelet theory, define

$$\phi(x) = 2\sum_n h_n\, \phi(2x - n). \tag{4.11}$$

This determines $\hat{\phi}$ up to a scale factor [17], which is chosen such that $\hat{\phi}(0) = 1$. Define the discretization operator

$$(\mathcal{D}_k f)_n = (\phi_k * f)(x^k_n), \qquad \phi_k(x) = 2^k\phi(2^k x). \tag{4.12}$$

Now consider the reconstruction operator $\mathcal{R}_k$. Let $\mathcal{I}_k$ be any interpolatory operator. Define the corresponding interpolatory operator $(Q^k_{k-1}v^{k-1})_n = (\mathcal{I}_{k-1}v^{k-1})(x^k_n)$. Section 4.3 defines the prediction operator as

$$\begin{aligned}(P^k_{k-1}v^{k-1})_n &= \big(a * Q^k_{k-1}(b * v^{k-1})\big)_n \\ &= \big(a_{\delta_k} * \mathcal{I}_{k-1}(b * v^{k-1})\big)(x^k_n),\end{aligned} \tag{4.13}$$

where $b(z^2) = 1/\big(h(z)h(-z)\big)$ and $a(z) = h(-z)$. Suppose that the reconstruction operator has the form $\mathcal{R}_k v^k = \xi_k * \mathcal{I}_k(b * v^k)$, where ξ_k is a distribution. Then the prediction operator is defined by the framework to be

$$\begin{aligned}(P^k_{k-1}v^{k-1})_n &= (\mathcal{D}_k\mathcal{R}_{k-1}v^{k-1})_n \\ &= \big(\phi_k * \xi_{k-1} * \mathcal{I}_{k-1}(p * v^{k-1})\big)(x^k_n).\end{aligned}$$

In order to agree with the construction of P_{k-1}^k in (4.13), ξ_{k-1} is defined implicitly by $\phi_k * \xi_{k-1} = a_{\delta_k}$. Since h is FIR, ϕ_k has a finite degree of continuity, and hence $\left|\hat{\xi}_{k-1}(\omega)\right| = |\hat{a}_{\delta_k}(\omega)| / |\hat{\phi}_k(\omega)| \le \|a\|_{\ell^1} / |\hat{\phi}_k(\omega)| = O(|\omega|^p)$ for some finite p. Therefore, ξ_{k-1} is a tempered distribution.

Let $w(z) = 1/b(z)$, then since $w = \downarrow(h * a)$ and $\phi_{k-1} = h_{\delta_k} * \phi_k$,

$$\phi_k * \xi_{k-1} = a_{\delta_k} \quad \Rightarrow \quad h_{\delta_k} * \phi_k * \xi_{k-1} = (h * a)_{\delta_k} \quad \Rightarrow \quad \phi_{k-1} * \xi_{k-1} = w_{\delta_{k-1}}.$$

So ξ_k is also given by $\phi_k * \xi_k = w_{\delta_k}$. Define the reconstruction operator as

$$\mathcal{R}_k v^k = \xi_k * \mathcal{I}_k(b * v^k), \qquad \phi_k * \xi_{k-1} = a_{\delta_k}. \tag{4.14}$$

For example, in the cell-average case, $\mathcal{R}_k$ should be $\mathcal{R}_k v^k = \frac{d}{dx}\mathcal{I}_k(b * v^k)$, with $b(z) = 2^{-k}(1 - z^{-1})^{-1}$ and $\xi_k = -\delta'$ (such that $\xi_k * (\cdot) = \frac{d}{dx}(\cdot)$).

This approach to the reconstruction operator provides a general notion of primitive. If the primitive samples are $\mathring{v}^k = b * v^k$, then

$$\begin{aligned} \mathring{f}^k(x_n^k) = \mathring{v}_n^k = (b * v^k)_n = (b_{\delta_k} * \phi_k * f)(x_n^k), \\ (\xi_k * \mathring{f}^k)(x_n^k) = (\xi_k * b_{\delta_k} * \phi_k * f)(x_n^k) = (b_{\delta_k} * w_{\delta_k} * f)(x_n^k) = f(x_n^k), \end{aligned}$$

which suggests defining a "generalized primitive" $\mathring{f}^k$ (see Figure 4.4) by

$$\mathring{f}^k = b_{\delta_k} * \phi_k * f, \qquad f = \xi_k * \mathring{f}^k.$$

In contrast to the primitive in the cell-average and hat-based formulations (section 3.3), the generalized primitive $\mathring{f}^k$ is a different function for each scale k. However, for spline discretizations, section 6.1 will show that there is a single primitive independent of k.

THEOREM 4.2. *The discretization and reconstruction operators defined in* (4.12) *and* (4.14) *satisfy the consistency relationship, and* $(\mathcal{D}_k)$ *is nested. Furthermore, the decimation operator* $D_k^{k-1} = \mathcal{D}_{k-1}\mathcal{R}_k$ *and the prediction operator* $P_{k-1}^k = \mathcal{D}_k\mathcal{R}_{k-1}$ *are*

$$D_k^{k-1} v^k = \downarrow(h * v^k), \qquad P_{k-1}^k v^{k-1} = a * Q_{k-1}^k(b * v^{k-1}),$$

where $(Q_{k-1}^k v^{k-1})_n = (\mathcal{I}_{k-1} v^{k-1})(x_n^k)$.

Proof. For any v^k, $(\mathcal{D}_k\mathcal{R}_k v^k)_n = (\phi_k * \xi_k * \mathcal{I}_k \mathring{v}^k)(x_n^k) = (w_{\delta_k} * \mathcal{I}_k \mathring{v}^k)(x_n^k)$. Because $\mathcal{I}_k$ is an interpolatory operator, $(\mathcal{I}_k \mathring{v}^k)(x_n^k) = \mathring{v}_n^k$. Also noting that w_{δ_k} is a delta sequence, the convolution is equivalently written as the discrete convolution $(w * \mathring{v}^k)_n$. Thus,

$$(\mathcal{D}_k\mathcal{R}_k v^k)_n \;=\; (w * \mathring{v}^k)_n \;=\; v_n^k.$$

So the consistency relationship is satisfied. Because $\phi_{k-1} = h_{\delta_k} * \phi_k$,

$$v_n^{k-1} \;=\; (\phi_{k-1} * f)(x_n^{k-1}) \;=\; (h_{\delta_k} * \phi_k * f)(x_{2n}^k) \;=\; \downarrow(h * v^k)_n.$$

If $\mathcal{D}_k f = v^k = 0$, then $\mathcal{D}_{k-1} f = v^{k-1} = \downarrow(h * v^k) = 0$. Therefore, $(\mathcal{D}_k)$ is nested. Furthermore, consistency and nestedness imply that

$$D_k^{k-1} v^k \;=\; \mathcal{D}_{k-1}\mathcal{R}_k\mathcal{D}_k f \;=\; \mathcal{D}_{k-1} f \;=\; v^{k-1} \;=\; \downarrow(h * v^k).$$

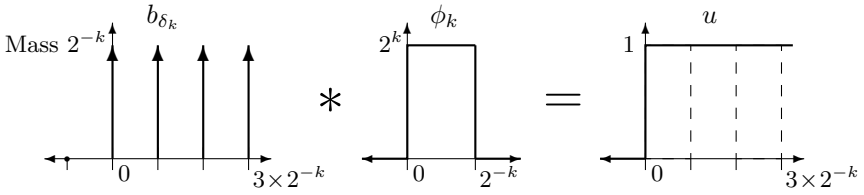


FIG. 4.4. *The cell-average primitive $\mathring{f}$ is related to f by convolution with the unit step $\mathring{f} = b_{\delta_k} * \phi_k * f = u * f$.*

The prediction operator is

$$\begin{aligned}(\mathcal{D}_k \mathcal{R}_{k-1} v^{k-1})_n &= (\phi_k * \xi_{k-1} * \mathcal{I}_{k-1} \mathring{v}^{k-1})(x_n^k) \\ &= \big(a_{\delta_k} * \mathcal{I}_{k-1}(b * v^{k-1})\big)(x_n^k) \\ &= \big(a * Q_{k-1}^k (b * v^{k-1})\big)_n. \qquad \square\end{aligned}$$

Remark 6. If c satisfies $c(z) = c(-z)$, Theorem 4.2 also holds for filters

$$b(z^2) = \frac{1}{c(z)h(z)h(-z)}, \qquad a(z) = c(z)h(-z).$$

Remark 7. Suppose that the interpolation $\mathcal{I}_k$ is linear, shift-invariant, and scale-independent, then

$$(\mathcal{I}_k \mathring{v}^k)(x) = \sum_n 2^k \theta(2^k x - n)\, \mathring{v}_n^k = (\theta_k * \mathring{v}_{\delta_k}^k)(x), \qquad \theta_k(x) = 2^k \theta(2^k x),$$

where the associated function θ satisfies $\theta(0) = 1$ and $\theta(n) = 0$ for all nonzero integers $n \in \mathbb{Z}\backslash\{0\}$. The reconstruction operator becomes $\mathcal{R}_k v^k = \xi_k * \theta_k * b_{\delta_k} * v_{\delta_k}^k$. Let $\tilde{\phi}_k = \xi_k * \theta_k * b_{\delta_k}$ such that $\mathcal{R}_k v^k = \tilde{\phi}_k * v^k$. Since $\phi_k * \xi_k = w_{\delta_k}$ and $w(z) = 1/b(z)$,

$$\phi_k * \tilde{\phi}_k = (\phi_k * \xi_k) * \theta_k * b_{\delta_k} = (w_{\delta_k} * b_{\delta_k}) * \theta_k = \theta_k,$$

so $\tilde{\phi}_k$ satisfies $\phi_k * \tilde{\phi}_k = \theta_k$. Reconstruction reduces to the wavelet reconstruction

$$(\mathcal{R}_k v^k)(x) = \sum_n 2^k \tilde{\phi}(2^k x - n)\, v_n^k,$$

where the dual scaling function $\tilde{\phi}$ satisfies $\phi * \tilde{\phi} = \theta$.

As an example of Theorem 4.2, consider a cell-average ENO with $h(z) = (1+z^{-1})$. Then if $\kappa = 2^k$, $a(z) = 2^{k-1}(1 - z^{-1})$ and $b(z) = 2^{-k}(1 - z^{-1})^{-1}$. The discretization operator is

$$(\mathcal{D}_k f)_n = (\phi_k * f)(x_n^k) = \int_{x_{n-1}^k}^{x_n^k} f(x)\, dx, \qquad \phi(x) = \begin{cases} 1 & \text{if } 0 \le x < 1, \\ 0 & \text{otherwise.} \end{cases}$$

The primitive $\mathring{f}$ is related to f by

$$\mathring{f}(x) = (b_{\delta_k} * \phi_k * f)(x) = (u * f)(x) = \int_0^x f(y)\, dy.$$

Since $\phi_{k+1} * \xi_k = a_{\delta_{k+1}}$,

$$\hat{\xi}_k(\omega) \;=\; \frac{\hat{a}_{\delta_{k+1}}(\omega)}{\hat{\phi}_{k+1}(\omega)} \;=\; \frac{1-\exp(-i\omega 2^{-k-1})}{\frac{1}{i\omega}\left(1-\exp(-i\omega 2^{-k-1})\right)} \;=\; i\omega.$$

That is, convolution with ξ_k is differentiation. Consistent with (3.4) in section 3.3, the reconstruction operator is $(\mathcal{R}_k v^k)(x) = (\xi_k * \mathcal{I}_k \mathring{v}^k)(x) = \frac{d}{dx}(\mathcal{I}_k \mathring{v}^k)(x)$, where $\mathring{v}^k_n = (b * v^k)_n = \mathring{f}(x^k_n)$.

5. Error control. An application of multiresolution decompositions is signal approximation. When the multiresolution transform is linear and orthogonal (or nearly orthogonal), an effective strategy is retaining the M largest transform coefficients and setting the rest to zero. This "M-term approximation" works, since the ℓ^2 error incurred by truncating a coefficient is proportional to its magnitude. However, for nonorthogonal or nonlinear schemes, another strategy is necessary.

Rather than fixing the number of nonzero coefficients, Harten proposed [13] a modified encoding procedure that produces a sparse representation $\hat{m}$ with approximation error less than a specified maximum error.

Define the (hard) thresholding operator

$$\mathrm{tr}(x,\epsilon) = \begin{cases} 0 & \text{if } |x| \le \epsilon, \\ x & \text{if } |x| > \epsilon. \end{cases}$$

Given v^L and a sequence of tolerances (ϵ_k), the modified encoding algorithm is

(5.1)
$$\begin{aligned}
&\textbf{for } k = L, \ldots, 1 \\
&\quad v^{k-1} = D^k_{k-1} v^k \\
&\textbf{end} \\
&\hat{v}^0 = v^0 \\
&\textbf{for } k = 1, \ldots, L \\
&\quad \tilde{d}^k = G_k(v^k - P^k_{k-1}\hat{v}^{k-1}) \\
&\quad \hat{d}^k = \mathrm{tr}(\tilde{d}^k, \epsilon_k) \\
&\quad \hat{v}^k = P^k_{k-1}\hat{v}^{k-1} + \tilde{G}_k \hat{d}^k \\
&\textbf{end} \\
&\hat{m} = (\hat{v}^0, \hat{d}^1, \ldots, \hat{d}^L).
\end{aligned}$$

The thresholds (ϵ_k) control the error of the approximation. Larger values of ϵ_k lead to sparser decompositions, but also larger error. Given a tolerance ϵ, thresholds (ϵ_k) can be selected such that the error is bounded:

$$J_L^{-1/p} \left\| v^L - \hat{v}^L \right\|_p \le \epsilon,$$

where J_k is the length of v^k and $\|\cdot\|_p$ denotes the usual ℓ^p norm.

Such error control has been developed for point-value, cell-average, and hat-based discretizations with zero-padded boundary handling [5, 6]. For point-value discretization, if $\epsilon_k = \epsilon$, then

$$J_L^{-1/p} \left\| v^L - \hat{v}^L \right\|_p \le \epsilon \qquad \text{for } p = 1, 2, \infty. \tag{5.2}$$

For cell-average discretization, if $\epsilon_k = \epsilon\, q^{L-k}$ with $0 < q < 1$, then

$$J_L^{-1/p} \left\| v^L - \hat{v}^L \right\|_p \le \begin{cases} \dfrac{\epsilon}{1-q} & \text{for } p = 1, \infty, \\ \dfrac{\epsilon}{\sqrt{1-q^2}} & \text{for } p = 2. \end{cases} \tag{5.3}$$

For hat-based discretization, if $\epsilon_k = \epsilon\, q^{L-k}$ with $0 < q \le \frac{1}{2}$, then

$$J_L^{-1/p} \left\| v^L - \hat{v}^L \right\|_p \le \begin{cases} \dfrac{\epsilon}{1-2q} & \text{if } 0 < q < \frac{1}{2}, \\ L\epsilon & \text{if } q = \frac{1}{2}, \end{cases} \qquad \text{for } p = 1, 2, \infty. \tag{5.4}$$

See [5, 6] for proofs. Theorem 5.1 provides error control for general discretizations.

THEOREM 5.1. *Let D_k^{k-1} and G_k be linear operators from V^k to V^{k-1}, and $\tilde{H}_k$ and $\tilde{G}_k$ be linear operators from V^{k-1} to V^k satisfying $\tilde{H}_k D_k^{k-1} + \tilde{G}_k G_k = I_{V^k}$, and let P_{k-1}^k be any operator from V^{k-1} to V^k satisfying $D_k^{k-1} P_{k-1}^k = I^{V_{k-1}}$.*

Let $\|\cdot\|$ be a norm, and let α, β be such that

$$\|\tilde{H}_k v\| \le \alpha \|v\| \quad \text{and} \quad \|\tilde{G}_k v\| \le \beta \|v\|_\infty \qquad \text{for all } v \in V^{k-1}, k \le L.$$

If the thresholds (ϵ_k) in the modified encoding algorithm (5.1) satisfy $\epsilon_k = \epsilon\, q^{L-k}$ for some $q \in (0, \alpha^{-1})$, then the reconstruction error is bounded:

$$\left\| v^L - \hat{v}^L \right\| \;<\; \frac{\beta\epsilon}{1-\alpha q}.$$

Proof. First, observe that

$$v^k - \hat{v}^k = (P_{k-1}^k v^{k-1} - P_{k-1}^k \hat{v}^{k-1}) + \tilde{G}_k (d^k - \hat{d}^k).$$

The detail term expands to $d^k - \hat{d}^k = (d^k - \tilde{d}^k) + (\tilde{d}^k - \hat{d}^k) = -G_k(P_{k-1}^k v^k - P_{k-1}^k \hat{v}^{k-1}) + (\tilde{d}^k - \hat{d}^k)$. Therefore,

$$\begin{aligned} v^k - \hat{v}^k &= (I_{V^k} - \tilde{G}_k G_k)(P_{k-1}^k v^{k-1} - P_{k-1}^k \hat{v}^{k-1}) + \tilde{G}_k(\tilde{d}^k - \hat{d}^k) \\ &= \tilde{H}_k D_k^{k-1} (P_{k-1}^k v^{k-1} - P_{k-1}^k \hat{v}^{k-1}) + \tilde{G}_k(\tilde{d}^k - \hat{d}^k) \\ &= \tilde{H}_k (D_k^{k-1} P_{k-1}^k v^{k-1} - D_k^{k-1} P_{k-1}^k \hat{v}^{k-1}) + \tilde{G}_k(\tilde{d}^k - \hat{d}^k) \\ &= \tilde{H}_k (v^{k-1} - \hat{v}^{k-1}) + \tilde{G}_k(\tilde{d}^k - \hat{d}^k). \end{aligned}$$

Now by definitions of α, β, and $\hat{d}^k = \mathrm{tr}(\tilde{d}^k, \epsilon_k)$, it follows that

$$\begin{aligned} \left\| v^k - \hat{v}^k \right\| &\le \left\| \tilde{H}_k (v^{k-1} - \hat{v}^{k-1}) \right\| + \left\| \tilde{G}_k (\tilde{d}^k - \hat{d}^k) \right\| \\ &\le \alpha \left\| v^{k-1} - \hat{v}^{k-1} \right\| + \beta\epsilon_k. \end{aligned}$$

Applying this bound recursively,

$$\left\| v^L - \hat{v}^L \right\| \;\le\; \beta \sum_{k=1}^{L} \alpha^{L-k} \epsilon_k \;=\; \beta\epsilon \frac{1 - (\alpha q)^L}{1 - \alpha q} \;<\; \frac{\beta\epsilon}{1-\alpha q}. \qquad \square$$

COROLLARY 5.2. *Let $\|v\|$ be the scaled ℓ^p norm $J_k^{-1/p} \|v\|_p$. Let $\tilde{h}$ and $\tilde{g}$ be wavelet filters as in section 4.1, and let $\tilde{h}_n^e = \tilde{h}_{2n}$ and $\tilde{h}_n^o = \tilde{h}_{2n+1}$ be the even and odd components of $\tilde{h}$, and similarly $\tilde{g}^e$ and $\tilde{g}^o$ be the even and odd components of $\tilde{g}$.*

Suppose that $\circledast$ is periodic convolution or, for convolution with other boundary handling, suppose that the wavelet filters have the form $h(z) = h_0 + zh^o(z^2)$ and $g(z) = z^{-1}$. Then $\tilde{H}_k v = \tilde{h} \circledast \uparrow v$ and $\tilde{G}_k v = \tilde{g} \circledast \uparrow v$.

(i) *If $\circledast$ is periodic convolution, then Theorem* 5.1 *applies with*

(5.5)
$$\alpha = \left(\frac{\|\tilde{h}^e\|_1^p + \|\tilde{h}^o\|_1^p}{2} \right)^{1/p}, \quad \beta = \left(\frac{\|\tilde{g}^e\|_1^p + \|\tilde{g}^o\|_1^p}{2} \right)^{1/p} \quad \text{if } 1 \le p < \infty,$$
$$\alpha = \|\tilde{h}^e\|_1 \vee \|\tilde{h}^o\|_1, \qquad \beta = \|\tilde{g}^e\|_1 \vee \|\tilde{g}^o\|_1 \qquad \text{if } p = \infty.$$

(ii) *If $\circledast$ is zero-padded convolution and $h(z) = h_0 + zh^o(z^2)$, $g(z) = z^{-1}$, then Theorem* 5.1 *applies with* (5.5).

(iii) *If $\circledast$ is convolution with symmetric or constant boundary extension and $h(z) = h_0 + zh^o(z^2)$, $g(z) = z^{-1}$, then Theorem* 5.1 *applies with* (5.5) *for $p = \infty$.*

Proof. (i) By Young's inequality on the periodic grid,

$$\begin{aligned} J_k^{-1/p} \left\| \tilde{h} \circledast \uparrow v \right\|_p &= J_k^{-1/p} \left[\left\| \tilde{h}^e \circledast v \right\|_p^p + \left\| \tilde{h}^o \circledast v \right\|_p^p \right]^{1/p} \\ &\le J_k^{-1/p} \left[\left(\left\| \tilde{h}^e \right\|_1^p + \left\| \tilde{h}^o \right\|_1^p \right) \|v\|_p^p \right]^{1/p} = \alpha J_{k-1}^{-1/p} \|v\|_p . \end{aligned}$$

(ii) and (iii) Let $v \in V^k$, and let v^{ext} be its extension to the infinite grid, as determined by the boundary-handling method, but with $v_n^{\text{ext}} = 0$ for $|n| > \operatorname{supp} \tilde{h} + \operatorname{supp} v$. Then $(\tilde{h} \circledast \uparrow v)_n = (\tilde{h} * \uparrow \tilde{v})_n$ for $n = 0, \dots J_k - 1$ and $\|v\|_p = \|v^{\text{ext}}\|_p$. By Young's inequality on the infinite grid,

$$\begin{aligned} J_k^{-1/p} \left\| \tilde{h} \circledast \uparrow v \right\|_p &\le J_k^{-1/p} \left\| \tilde{h} * \uparrow v^{\text{ext}} \right\|_p \\ &= J_k^{-1/p} \left[\left\| \tilde{h}^e * v^{\text{ext}} \right\|_p^p + \left\| \tilde{h}^o * v^{\text{ext}} \right\|_p^p \right]^{1/p} \\ &\le J_k^{-1/p} \left[\left(\left\| \tilde{h}^e \right\|_1^p + \left\| \tilde{h}^o \right\|_1^p \right) \left\| v^{\text{ext}} \right\|_p^p \right]^{1/p} = \alpha J_{k-1}^{-1/p} \|v\|_p . \end{aligned}$$

Similarly, β is such that $\left\| \tilde{g} \circledast v \right\| \le \beta \|v\| \le \beta \|v\|_\infty$. ∎

Remark 8. For nonperiodic boundary handling, the requirement $h(z) = h_0 + zh^o(z^2)$ and $g(z) = z^{-1}$ ensures a wavelet transform implementation with a single update lifting step. For wavelets with more than one lifting step, it is generally not true that $\tilde{H}_k v = \tilde{h} \circledast \uparrow v$ and $\tilde{G}_k v = \tilde{g} \circledast \uparrow v$, though these formulas do hold for points sufficient far from the boundaries.

However, presuming these boundary effects are small, the conclusions of Theorem 5.1 with (5.5) hold approximately for any wavelet and boundary handling.

For the point-value, cell-average, and hat-based discretizations, the wavelet filters have the form $h(z) = h_0 + zh^o(z^2)$ and $g(z) = z^{-1}$. For both the cell-average and hat-based discretizations with periodic or zero-padded boundary handling, $\alpha = 2^{1-1/p}$ and $\beta = 1$, and Theorem 5.1 yields

$$J_L^{-1/p} \left\| v^L - \hat{v}^L \right\|_p \le \frac{\epsilon}{1 - 2^{1-1/p} q} = \begin{cases} \dfrac{\epsilon}{1-q} & \text{for } p = 1, \\ \dfrac{\epsilon}{1-\sqrt{2}q} & \text{for } p = 2, \\ \dfrac{\epsilon}{1-2q} & \text{for } p = \infty. \end{cases}$$

For cell-average discretization, this bound agrees with (5.3) for $p = 1$ and is worse for $p = 2, \infty$. For hat-based discretization, it agrees with (5.4) for $p = \infty$ and is better for $p < \infty$.

For point-value discretization, Theorem 5.1 is overly conservative with error bound $\epsilon/(2^{1/p-1} - q)$, which is much worse than (5.2). However, this case is extreme, as all involved filters are one-tap.

6. Example constructions. In this section, two examples of Harten multiresolution schemes are constructed using the approach proposed in the previous sections. The first is a family of spline discretization schemes, where cell-average discretization is order $N = 1$ and hat-based discretization is $N = 2$. The second is a simple example with nonspline decimation filters.

6.1. Spline discretizations. This section develops Harten schemes with spline discretization. Let $h(z)$ be the Nth-order spline decimation filter

$$h(z) = z^{\lfloor N/2 \rfloor} \left(\frac{1 + z^{-1}}{2} \right)^N,$$

and define $D_k^{k-1} v^k = {\downarrow}(h * v^k)$. As in section 4.3, the prediction operator is $P_{k-1}^k v^{k-1} = a * Q_{k-1}^k (b * v^{k-1})$ for any interpolatory operator Q_{k-1}^k, with

$$b(z) = \frac{1}{z^{\lfloor N/2 \rfloor} (1 - z^{-1})^N}, \qquad a(z) = z^{\lfloor N/2 \rfloor} (1 - z^{-1})^N.$$

Choose $\tilde{N}$ such that $M = \frac{1}{2}(N + \tilde{N}) - 1$ is integer, and define $\tilde{h}$, g, and $\tilde{g}$ as in (4.4),

$$\tilde{h}(z) = 2 z^{\lceil \tilde{N}/2 \rceil} \left(\frac{1 + z^{-1}}{2} \right)^{\tilde{N}} \sum_{n=0}^{M} \binom{M+n}{n} (-4)^{-n} (z - 2 + z^{-1})^n,$$
$$g(z) = \tfrac{1}{2} z^{-1} \tilde{h}(-z), \qquad \tilde{g}(z) = 2 z h(-z).$$

The detail encoder operator is $G_k e^k = {\downarrow}(g * e^k)$, and the decoder is $\tilde{G}_k d^k = \tilde{g} * ({\uparrow} d^k)$. At this point, all of the necessary operators for a discrete implementation of the scheme are established.

Now consider the continuous operators $\mathcal{D}_k$ and $\mathcal{R}_k$. The function $\phi(x)$ satisfying dilation equation (4.11) is the B-spline

$$\phi(x) = \frac{1}{(N-1)!} \sum_{n=0}^{N} \binom{N}{n} (-1)^n \left[(x + \left\lfloor \frac{N}{2} \right\rfloor - n)^+ \right]^{N-1}. \tag{6.1}$$

The B-spline ϕ can be shown to satisfy the dilation equation in the Fourier domain. Let $\tau = 0$ if N is even and $\tau = 1$ if N is odd. Its Fourier transform is $\hat{\phi}(\omega) = \mathrm{e}^{-i\omega\tau/2} (\frac{\sin(\omega/2)}{\omega/2})^N$, so that

$$\begin{aligned}
2 \sum_n h_n \tfrac{1}{2} \hat{\phi}(\omega/2) \mathrm{e}^{-in\omega/2} &= 2 \sum_{n=-\lfloor N/2 \rfloor}^{\lceil N/2 \rceil} 2^{-N} \binom{N}{n + \lfloor \frac{N}{2} \rfloor} \tfrac{1}{2} \mathrm{e}^{-i\omega(\tau+2n)/4} \left(\frac{\sin(\omega/4)}{\omega/4} \right)^N \\
&= \mathrm{e}^{-i\omega(\tau - 2\lfloor N/2 \rfloor)/4} \left(\sum_{n=0}^{N} \binom{N}{n} \mathrm{e}^{-i\omega n/2} \right) \left(\frac{\sin(\omega/4)}{\omega/2} \right)^N \\
&= \mathrm{e}^{-i\omega\tau/2} \left(\frac{\sin(\omega/2)}{\omega/2} \right)^N = \hat{\phi}(\omega).
\end{aligned}$$

As in section 4.4, define the discretization operator as $(\mathcal{D}_k f)_n = (\phi_k * f)(x_n^k)$. The primitive $\mathring{f}^k$ is related to f by $\mathring{f}^k = b_{\delta_k} * \phi_k * f$. In the Fourier domain,

$$\hat{b}_{\delta_k}(\omega)\hat{\phi}_k(\omega) \;=\; \frac{\mathrm{e}^{-i\omega\tau/2^{k+1}}\left(\frac{\sin(\omega/2^{k+1})}{\omega/2^{k+1}}\right)^N}{\mathrm{e}^{i\omega(2\lfloor N/2\rfloor - N)/2^{k+1}}(2i\sin(\omega/2^{k+1}))^N} \;=\; 2^{kN}(i\omega)^{-N}. \tag{6.2}$$

Thus the primitive $\mathring{f}^k$ is related to f by an Nth-order integral with scale factor 2^{kN},

$$\mathring{f}^k(x) = 2^{kN}\int_0^x\int_0^{y_1}\cdots\int_0^{y_{N-1}} f(y_N)\,dy_N\cdots dy_2\,dy_1.$$

Let $\mathring{v}_n^k = \mathring{f}^k(x_n^k)$ be samples of $\mathring{f}^k$. In the Fourier domain, ξ_k is

$$\frac{\hat{a}_{\delta_{k+1}}(\omega)}{\hat{\phi}_{k+1}(\omega)} \;=\; \frac{\mathrm{e}^{i\omega(2\lfloor N/2\rfloor - N)/2^{k+2}}(2i\sin(\omega/2^{k+2}))^N}{\mathrm{e}^{-i\omega\tau/2^{k+2}}\left(\frac{\sin(\omega/2^{k+2})}{\omega/2^{k+2}}\right)^N} \;=\; 2^{-kN}(i\omega)^N. \tag{6.3}$$

Convolution with ξ_k is Nth-order differentiation, with a scale factor 2^{-kN}. Given an interpolatory operator $\mathcal{I}_k$, the reconstruction operator is

$$(\mathcal{R}_k v^k)(x) = \big(\xi_k * \mathcal{I}_k(b * v^k)\big)(x) = 2^{-kN}\frac{d^N}{dx^N}(\mathcal{I}_k\mathring{v}^k)(x).$$

If b is scaled by 2^{kN} and a by 2^{-kN}, then the 2^{kN} scale factor in (6.2) and (6.3) cancels, yielding a primitive independent of k:

$$\mathring{f}(x) = \int_0^x\int_0^{y_1}\cdots\int_0^{y_{N-1}} f(y_N)\,dy_N\cdots dy_2\,dy_1,$$
$$(\mathcal{R}_k v^k)(x) = \frac{d^N}{dx^N}(\mathcal{I}_k\mathring{v}^k)(x), \qquad \mathring{v}_n^k = \mathring{f}(x_n^k).$$

Since the primitive $\mathring{f}$ is related to f by Nth-order integration, predicting a polynomial of degree K requires interpolation, with an order of at least $K+N$.

To control regularity, $\mathcal{I}_k\mathring{v}^{k-1}$ should be at least N times weakly differentiable. An ENO interpolant has a weak first derivative, so $\mathcal{I}_k = \mathcal{I}_k^{\mathrm{ENO}}$ works directly for the point-value $N=0$ and the cell-average $N=1$ cases. In the hat-based case $N=2$, the second derivative of an ENO interpolant contains Dirac measures. Aràndiga, Donat, and Harten [6] control the manifestation of Dirac measures in their hat-based scheme using a careful subcell-resolution scheme.

Another solution is to follow P^{ENO} with a smoothing step (4.10). This modification controls the regularity without losing the prediction's edge-adaptive behavior, as demonstrated in Figure 6.1.

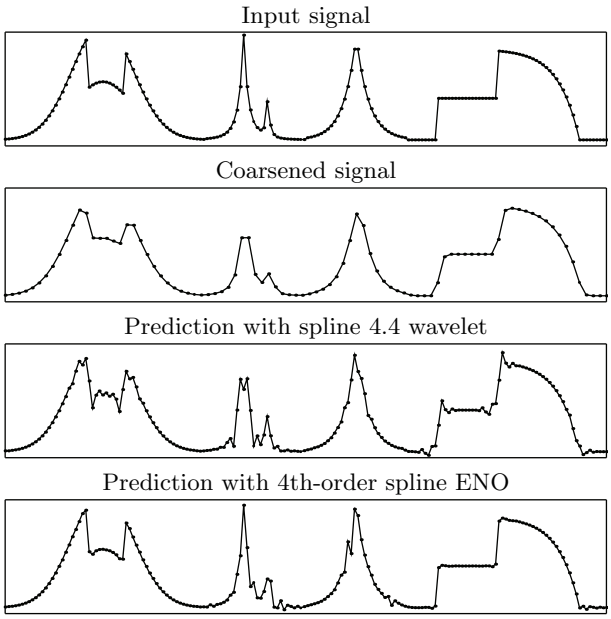


FIG. 6.1. *A piecewise-regular signal is first coarsened with 4th-order spline decimation, and then prediction with the spline 4.4 wavelet and 4th-order spline ENO both attempt to recover the original. ENO interpolation is followed by* (4.10) *to control regularity.*

The hierarchical stencil selection method (described in section 3.1) has the problem that it may select singularity-crossing stencils for discontinuities in $\mathring{f}''$ and higher derivatives [5]. Since a jump becomes a discontinuity in $\mathring{f}^{(N)}$, jump discontinuity in f is not correctly handled for $N \geq 2$. Nonhierarchical selection does not have this problem, so it is the better method for $N \geq 2$.

6.2. A nonspline scheme. Harten's multiresolution framework generalizes wavelets, including those with nonspline lowpass filters. This section constructs a simple Harten multiresolution scheme based on piecewise polyharmonic (PPH) interpolation and the Cohen–Daubechies–Feauveau 9/7 (CDF 9/7) wavelet.

Define the linear/cubic PPH interpolatory operator

$$(P^{\mathrm{PPH}} v)_{2n+1} = \frac{v_{n+1} + v_n}{2} - \frac{1}{4} \frac{(\Delta^2 v_{n+1} \Delta^2 v_n)^+}{\Delta^2 v_{n+1} + \Delta^2 v_n},$$

where $\Delta^2 v_n = v_{n+1} - 2v_n + v_{n-1}$. Amat et al. [3, 4] considered it as the prediction operator in point-value discretization schemes and have shown its nice properties and potential in nonlinear approximation.

Let H, G, $\tilde{H}$, and $\tilde{G}$ be the decomposition and reconstruction operators for the CDF 9/7 wavelet, and let $D_k^{k-1} = H$ be the decimation operator. As in (4.8), a consistent prediction based on P^{PPH} is

$$P_{k-1}^k v^{k-1} = \tilde{H} v^{k-1} + \tilde{G} G P^{\mathrm{PPH}} v^{k-1}. \tag{6.4}$$

The CDF 9/7 PPH scheme is applied to nonlinear approximation in Figure 6.2. An image is sparsely decomposed with 3% nonzero coefficients using the modified encoding algorithm (section 5) and then reconstructed. The same is done with the

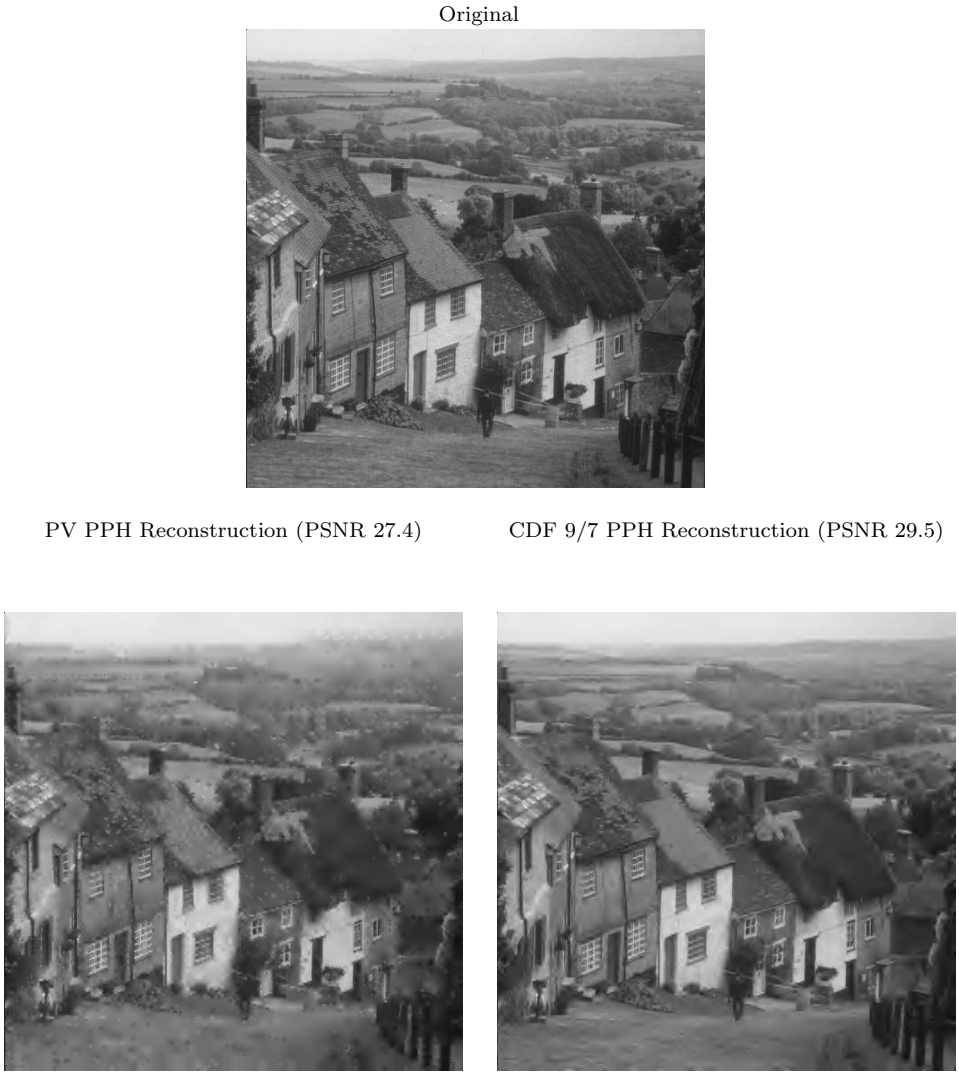


FIG. 6.2. *A nonlinear approximation of Goldhill using* 3% *nonzero coefficients. Left: Reconstruction using point-value discretization and PPH interpolation. Right: Reconstruction using CDF* 9/7 *wavelet discretization and* (6.4).

PPH scheme based on point-value discretization (PV PPH). Reconstruction quality is measured in terms of peak signal-to-noise (PSNR). Since the CDF 9/7 PPH scheme applies a smoothing filter before decimating, it can suppress aliasing artifacts and handle texture better.

7. Conclusion. This paper develops a new approach to the construction of Harten multiresolution schemes, where the decimation operator is based on an arbitrary FIR decimation filter h.

- It is possible to design all of the operators (D_k^{k-1}, P_{k-1}^k, G_k, $\tilde{G}_k$) necessary for implementing a nonredundant decomposition from h without ever defining $\mathcal{D}_k$ or $\mathcal{R}_k$ (section 4.3, section 4.1).

- $\mathcal{D}_k$ and $\mathcal{R}_k$ can be constructed (section 4.4) such that they are consistent with the decimation and prediction operators defined in section 4.3.
- The construction of $\mathcal{R}_k$ "via primitive function" [6, 14], and the notion of primitive generalizes to discretizations with any h (section 4.4).
- Schemes under this construction may be applied to nonlinear approximation using the error-control strategy developed in section 5.
- As a special case, the construction yields a family of Harten schemes with spline discretizations (section 6.1), where the point-value, cell-average, and hat-based discretizations are the first three members.
- It is also possible to construct Harten schemes with nonspline discretizations, as demonstrated by the example in section 6.2.

The presented construction expands the choice of discretization to any discretization with an FIR decimation filter. New Harten multiresolution schemes are possible using smoother and nonspline decimation filters.

Acknowledgments. We are grateful to the reviewers for their comments which helped improve the paper. We also thank Dr. James Curry and Dr. Anne Dougherty, who provided advice and guidance that focused the direction and ensured the quality of this work.

REFERENCES

[1] R. Abgrall and T. Sonar, *On the use of Mühlbach expansions in the recovery step of ENO methods*, Numer. Math., 76 (1997), pp. 1–25.
[2] S. Amat, F. Aràndiga, A. Cohen, R. Donat, G. Garcia, and M. Von Oehsen, *Data compression with ENO schemes: A case study*, Appl. Comput. Harmon. Anal., 11 (2002), pp. 273–288.
[3] S. Amat, R. Donat, J. Liandrat, and J.C. Trillo, *A fully adaptive PPH multi-resolution scheme for image processing*, Math. Comput. Modelling, 46 (2007), pp. 2–11.
[4] S. Amat, R. Donat, J. Liandrat, and J.C. Trillo, *Analysis of a new nonlinear subdivision scheme*, Applications in image processing, Found. Comp. Math., 6 (2006), pp. 193–225.
[5] F. Aràndiga and R. Donat, *Nonlinear multiscale decompositions: The approach of A. Harten*, Numer. Algorithms, 23 (2000), pp. 175–216.
[6] F. Aràndiga, R. Donat, and A. Harten, *Multiresolution based on weighted averages of the hat function* II: *Nonlinear reconstruction techniques*, SIAM J. Sci. Comput., 20 (1998), pp. 1053–1093.
[7] T. Cecil, J. Qian, and S. Osher, *Numerical methods for high dimensional Hamilton-Jacobi equations using radial basis functions*, J. Comput. Phys., 2003.
[8] R. Claypoole, G. Davis, W. Sweldens, and R. Baraniuk, *Nonlinear wavelet transforms for image coding*, in Proceedings of the 31st IEEE Asilomar Conference, Pacific Grove, CA, 1997.
[9] A. Cohen and B. Mateï, *Compact representation of images by edge adapted multiscale transforms*, in Proceedings of the International Conference on Image Processing, Thessaloniki, Greece, IEEE, 2001.
[10] A. Cohen, N. Dyn, and B. Mateï, *Quasilinear subdivision schemes with applications to ENO interpolation*, Appl. Comput. Harmon. Anal., 15 (2003), pp. 89–116.
[11] S. Godavarthy, *Generating spline wavelets*, in Proceedings in the 36th ACM Southeast Regional Conference, Marietta, GA, 1998, pp. 8–14.
[12] A. Harten, B. Engquist, S. Osher, and S. Chakravarthy, *Uniformly high order accurate essentially non-oscillatory schemes* III, J. Comput. Phys., 71 (1987), pp. 231–303.
[13] A. Harten, *Discrete multiresolution analysis and generalized wavelets*, J. Appl. Numer. Math., 12 (1993), pp. 152–192.
[14] A. Harten, *Multiresolution representation of data*: *A general framework*, SIAM J. Numer. Anal., 33 (1996), pp. 1205–1256.
[15] H. Heijmans and J. Goutsias, *Nonlinear multiresolution signal decomposition schemes–Part* II: *Morphological wavelets*, IEEE Trans. Image Process., 9 (2000), pp. 1897–1913.

[16] A. Iske and T. Sonar, *On the structure of function spaces in optimal recovery of point functionals for ENO-schemes by radial basis functions*, Numer. Math., 74 (1996), pp. 177–201.
[17] S. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, New York, 1998.
[18] T. Sonar, *On families of pointwise optimal finite volume ENO approximations*, SIAM J. Numer. Anal., 35 (1998), pp. 2350–2369.
[19] W. Sweldens, *The lifting scheme: A construction of second generation wavelets*, SIAM J. Math. Anal., 29 (1998), pp. 511–546.

SIAM J. NUMER. ANAL.
Vol. 46, No. 6, pp. 2978–3005
© 2008 Society for Industrial and Applied Mathematics

NUMERICAL APPROXIMATION OF RELATIVE EQUILIBRIA FOR EQUIVARIANT PDES*

VERA THÜMMLER†

Abstract. We prove convergence results for the numerical approximation of relative equilibria of parabolic systems in one space dimension. These systems are special examples of equivariant evolution equations. We use finite differences on a large interval with appropriately chosen boundary conditions. Moreover, we consider the approximation of isolated eigenvalues of finite multiplicity of the linear operator which arises from linearization at the equilibrium as well as the approximation of the corresponding invariant subspace. The results in this paper which are a generalization of the results in [V. Thümmler, *Numerical Analysis of the Method of Freezing Traveling Waves*, Ph.D. thesis, Bielefeld University, 2005] are illustrated by numerical computations for the cubic quintic Ginzburg–Landau equation.

Key words. equivariance, Lie groups, partial differential algebraic equations, unbounded domains, finite differences

AMS subject classifications. 65M99, 35K57

DOI. 10.1137/070698324

1. Introduction and basic setting.

Relative equilibria, i.e., solutions of partial differential equations which are equilibria in an appropriately comoving frame, frequently occur in systems with underlying symmetry [8], [16]. A basic class is formed by traveling waves which are stationary in a frame which moves with the velocity of the wave. The purpose of this paper is to analyze numerical methods for the approximation of relative equilibria of parabolic systems in one space dimension which are equivariant w.r.t. the action of a finite-dimensional Lie group [16], [18]. The numerical computation of the relative equilibrium amounts to solving a boundary value problem on the real line and to computing extra parameters (from the Lie algebra). These determine the motion of the comoving frame in the original coordinates. We analyze the combined effect of discretization with finite differences and the truncation of the infinite boundary value problem to a finite but large interval. This is achieved by using exponential dichotomies for the continous system which are inherited by the discretized equations. The truncation of the finite interval can then be handled via approximations of invariant subspaces. We prove the rate of convergence to the exact solution as well as error estimates for invariant subspaces corresponding to isolated eigenvalues of finite multiplicity of the linearization at the relative equilibrium.

A similar task has been accomplished for Galerkin approximations for traveling waves on cylindrical domains in [14]. In [17] a different approach for approximating homoclinic and heteroclinic orbits using Laguerre approximations has been proposed. In this paper we deal with the simplest numerical method—finite differences—and with a rather general set of problems—relative equilibria on the real line. The key difficulty is the simultaneous truncation of the real line to a finite interval together with the approximation of differential operators by finite differences. This is different

*Received by the editors July 25, 2007; accepted for publication (in revised form) April 29, 2008; published electronically August 20, 2008. This research was supported by CRC 701, Bielefeld University.

http://www.siam.org/journals/sinum/46-6/69832.html

†Fakultät für Mathematik, Universität Bielefeld, 33501 Bielefeld, Germany (thuemmle@math.uni-bielefeld.de).

from well-known results for boundary value problems on finite intervals as presented, e.g., in [1].

In this section we review the basic setting of the freezing approach which realizes the comoving frame and which has been developed in [5]. We fix the notations and provide an example. Moreover, we state the main spectral conditions on the linearization of a continuous system at the relative equilibrium. In section 2 we introduce the discretization and boundary conditions and state the main nonlinear approximation results Theorems 2.5 (approximation of relative equilibria) and 2.6 (approximation of isolated eigenvalues and invariant subspaces of the linearization). These are based on corresponding estimates for a linearized problem which are proven in section 3. In section 5 we illustrate the main results from section 2 by numerical computations for rotating pulses of the cubic quintic Ginzburg–Landau equation.

1.1. Equivariant evolution equations. We begin with a rather abstract setting for equivariant equations on Banach manifolds that covers the approaches in [5], [8], [23], [25] (see also [6]). Thereby the general structure of the boundary value problem to be solved becomes apparent. We show the necessary computations for a general relative equilibrium and in this way avoid repeating the calculation for different symmetries anew.

It will be applied later in Examples 1.3 and 1.4 with affine Banach spaces.

Let M be a manifold modeled over some Banach space X, and let N be a submanifold modeled over some dense subspace $Y \subset X$. Consider an evolutionary equation

$$u_t = F(u), \quad u(0) = u^0, \tag{1.1}$$

with a vector field $F : N \to TM$, where TM denotes the tangent bundle of M. In our applications (see section 1.2) we will have either Banach spaces $X = M,\ Y = N$ or affine spaces $M = \tilde{v} + X,\ N = \tilde{v} + Y$ such that the tangent spaces always satisfy $T_u M = X,\ T_v N = Y$ for all $u \in M, v \in N$.

Further we assume that the system is equivariant w.r.t. a finite-dimensional (possibly noncompact) Lie group G acting on M via

$$a : G \times M \to M,\ (\gamma, u) \mapsto a(\gamma)u,$$

where

$$a(\gamma_1 \circ \gamma_2)v = a(\gamma_1)[a(\gamma_2)v], \quad a(1)v = v, \quad 1 = \text{unit element in } G,$$

which has a tangent action Ta in TM, i.e., $Ta(\gamma) : T_v M \to T_{a(\gamma)v} M$. By equivariance we mean that the following relation holds:

$$\begin{aligned} a(\gamma)(N) &\subset N \quad \forall \gamma \in G, \\ F(a(\gamma)u) &= Ta(\gamma)F(u) \quad \forall u \in N, \gamma \in G. \end{aligned}$$

We assume that for any $v \in M$ the map $a(\cdot)v :\ G \to M, \quad \gamma \mapsto a(\gamma)v$ is continuous and is continuously differentiable for any $v \in N$ with its derivative denoted by

$$da(\gamma)v : T_\gamma G \to T_{a(\gamma)v} M, \quad \lambda \mapsto [da(\gamma)v]\lambda.$$

Finally, we denote the left translation by $L_\gamma :\ G \to G,\ g \mapsto \gamma \circ g$ and its derivative at g by $dL_\gamma(g) : T_g G \to T_{\gamma \circ g} G$.

In the following we give a constructive definition of relative equilibria which is appropriate from a numerical point of view [5].

DEFINITION 1.1. *A solution $\bar{u}$ of (1.1) is called a relative equilibrium if it has the form $\bar{u}(t) = a(\bar{\gamma}(t))\bar{v}$, where $\bar{\gamma} : [0, \infty) \to G$ is a smooth curve satisfying $\bar{\gamma}(0) = 1$ and $\bar{v} \in N$ does not depend on time.*

In general one calls the whole group orbit $\mathcal{O}(\bar{v}) = \{a(\gamma)\bar{v},\ \gamma \in G\}$ a relative equilibrium if it is invariant under the semiflow [8], [16]. For our purpose it is more convenient to select a special time orbit within this group orbit, since $\mathcal{O}(\bar{v})$ can always be generated by applying the group action a to the selected time orbit.

By using the ansatz $u(t) = a(\gamma(t))v(t)$, (1.1) can be transformed into a system for the unknowns $v(t) \in M,\ \gamma(t) \in G$, and $\mu(t) \in T_1G$ as follows (cf. [5], [22], [25]):

$$v_t = F(v) - [da(1)v]\mu, \qquad v(0) = u^0, \tag{1.2a}$$

$$\gamma_t = dL_\gamma(1)\mu, \qquad \gamma(0) = 1. \tag{1.2b}$$

If a relative equilibrium $\bar{u}(t) = a(\bar{\gamma}(t))\bar{v}$ is given such that $da(1)\bar{v}\mu \in T_{\bar{v}}M$ for all $\mu \in T_1G$ and $da(1)\bar{v} : T_1G \to T_{\bar{v}}M$ has full rank, then it can be shown that there exists $\bar{\mu} \in T_1G$ such that $(\bar{v}, \bar{\mu})$ is a stationary solution of (1.2), i.e.,

$$0 = F(\bar{v}) - [da(1)\bar{v}]\bar{\mu} \tag{1.3}$$

and $\bar{\gamma}_t = dL_\gamma(1)\bar{\mu} = \exp(t\bar{\mu})$. Conversely, if (1.3) holds for $(\bar{v}, \bar{\mu})$, then $a(\bar{\gamma}(t))\bar{v}$ is a relative equilibrium, where $\bar{\gamma}$ solves $\gamma_t = dL_\gamma(1)\bar{\mu} = \exp(t\bar{\mu})$.

Note that one can recover the relative equilibrium $\bar{u}(t) = a(\bar{\gamma}(t))\bar{v}$ exactly from a stationary solution $(\bar{v}, \bar{\mu})$ of (1.2a) by using the reconstruction equation (1.2b): For a solution $(\bar{v}, \bar{\mu})$ of (1.3), (1.2b) is an ODE which in many cases can be solved analytically (e.g., $\bar{\gamma}(t) = t\bar{\mu}$ in the traveling wave case; cf. Example 1.3). In this paper we are interested only in the approximation of $(\bar{v}, \bar{\mu})$ under discretization. However, the stability in time of $(\bar{v}, \bar{\mu})$ as an equilibrium of (1.2a) and the convergence towards the stationary solution is another interesting issue which has been dealt with in [26].

1.2. Parabolic equations. In the rest of the paper we are concerned with the special case of a parabolic PDE

$$u_t = Au_{xx} + f(u, u_x), \quad x \in \mathbb{R}, t > 0, u(x, t) \in \mathbb{R}^m, \tag{1.4}$$

where $A \in \mathbb{R}^{m,m}$ is positive definite and $f \in \mathcal{C}^1(\mathbb{R}^m \times \mathbb{R}^m, \mathbb{R}^m)$ is of the form

$$f(u, v) = f_1(u)v + f_2(u), \quad f_1 \in \mathcal{C}^1(\mathbb{R}^m, \mathbb{R}^{m,m}), \qquad f_2 \in \mathcal{C}^1(\mathbb{R}^m, \mathbb{R}^m), \tag{1.5}$$

and f_1, f_2, f_1', f_2' are globally Lipschitz.

Remark 1.2. The above conditions on f are also satisfied for "conservation laws" with special flux functions (e.g., Burgers equation). They imply that f_1' and f_2' are globally bounded, and with

$$D_1 f(u, w) = f_1'(u)(w, \cdot) + f_2'(u), \quad D_2 f(u, w) = f_1(u),$$

we obtain for $u, w, \delta_u, \delta_w \in \mathbb{R}^m$

$$\begin{aligned} &\|D_1 f(u + \delta_u, w + \delta_w) - D_1 f(u, w)\| \le \text{const}(\|\delta_u\| + \|\delta_w\|), \\ &\|D_2 f(u + \delta_u, w + \delta_w) - D_2 f(u, w)\| \le \text{const}\, \|\delta_u\|. \end{aligned} \tag{1.6}$$

In this case F in (1.1) reads

$$F(u) = Au'' + \bar{f}(u, u'),$$

where $\bar{f}(u,u')(x) = f(u(x),u'(x))$. Let a function $\tilde{v}:\mathbb{R}\to\mathbb{R}^m$ be given with

$$F(\tilde{v}) = A\tilde{v}'' + f(\tilde{v},\tilde{v}') \in \mathcal{L}_2. \tag{1.7}$$

Then we define $M = \tilde{v} + \mathcal{L}_2$ and $N = \tilde{v} + \mathcal{H}^2$, and with the properties of f we obtain that the condition $F : \tilde{v} + \mathcal{H}^2 \to \mathcal{L}_2$ is satisfied.

We choose a basis $\{e^1,\dots,e^p\}$ in the Lie algebra T_1G, where p is the dimension of G, write $\mu = \sum_{i=1}^p \mu_i e^i$, and define $S^i(v) = -\,da(1)ve^i$. Then the stationary equation reads

$$0 = Av'' + S(v)\mu + \bar{f}(v,v'), \quad x\in\mathbb{R},\ v(x)\in\mathbb{R}^m, \mu\in\mathbb{R}^p, \tag{1.8}$$

where we use the short notation $S(v)\mu = \sum_{i=1}^p S^i(v)\mu_i$.

Throughout the paper we need two additional assumptions:

A1. The operators S^i are linear differential operators of order ≤ 1 which can be written as

$$S^i(v)(x) = S^i_0 v(x) + S^i_1 v'(x), \quad S^i_{0,1}\in\mathbb{R}^{m,m} \tag{1.9}$$

and for which $S^i(\bar{v})\in\mathcal{L}_2$ for $i=1,\dots,p$. The last condition is necessary for the existence of a solution $\bar{v}\in\tilde{v}+\mathcal{H}^2$ of (1.8).

A2. There exist $\bar{v}_\pm\in\mathbb{R}^m$ such that

$$\lim_{x\to\pm\infty}\bar{v}(x) = \bar{v}_\pm. \tag{1.10}$$

In order to compensate for the additional p degrees of freedom which are obtained by introducing the parameter μ, we add a phase condition of the form

$$0 = \langle S^i(\hat{v}), v-\hat{v}\rangle_{\mathcal{L}_2}, \quad i=1,\dots,p, \tag{1.11}$$

to (1.8). Here $\hat{v}\neq 0$ is a given template function which satisfies $\hat{v}-\bar{v}\in\mathcal{H}^1$. The use of this phase condition is also called the method of slices [22]. In numerical calculations one can use the initial approximation of $\bar{v}$ as a template function.

In the following we present two characteristic examples, namely, traveling waves and rotating and traveling waves.

Example 1.3 (traveling waves). Consider (1.4), and let $\tilde{v}$ be a function with $\|\tilde{v}(x)-v_\pm\|\le \text{const}\,\mathrm{e}^{\pm\varrho x}$, where $f(v_\pm,0)=0$. Consider the shift action of $G=\mathbb{R}$, i.e., $[a(\gamma)u](x) = u(x-\gamma)$ on $M=\tilde{v}+\mathcal{L}_2 \supset N = \tilde{v}+\mathcal{H}^2$. Then we have $[da(1)v]e^1 = -v_x$, i.e., $S^1_1 = I$, $S^1_0 = 0$ and (1.2a), (1.11) read

$$\begin{aligned} v_t &= Av_{xx} + \lambda v_x + f(v,v_x),\\ 0 &= \langle \hat{v}', v-\hat{v}\rangle_{\mathcal{L}_2}. \end{aligned}$$

The relative equilibria are traveling waves $\bar{u}(x,t) = \bar{v}(x-\bar{\lambda}t)$ with stationary points $\lim_{x\to\pm\infty}\bar{v}(x) = v_\pm$.

Example 1.4 (rotating and traveling waves). Consider (1.4) for $\tilde{v}=0$, i.e., for $M=\mathcal{L}_2$ and $N=\mathcal{H}^2$. Let the Lie group be $G = S^1\times\mathbb{R}$ with $(\rho,\tau)=\gamma\in G$ and $(\rho,\tau)\circ(\tilde{\rho},\tilde{\tau}) = (\rho+\tilde{\rho},\tau+\tilde{\tau})$. Let the action $a: G\times\mathcal{L}_2\to\mathcal{L}_2$ be given for $u:\mathbb{R}\to\mathbb{R}^2$ by

$$[a(\gamma)u](x) = R_{-\rho}u(x-\tau), \quad R_\rho = \begin{pmatrix}\cos\rho & -\sin\rho\\ \sin\rho & \cos\rho\end{pmatrix}.$$

Then we have $[da(1)v]e^1 = -v_x$, $[da(1)v]e^2 = -R_{\frac{\pi}{2}}v$, i.e., $S_1^1 = I$, $S_0^2 = R_{\frac{\pi}{2}}$, $S_0^1 = S_1^2 = 0$ and (1.2a), (1.11) read with $\mu_\tau = \tau_t$, $\mu_\rho = \rho_t$,

$$\begin{aligned} v_t &= Av_{xx} + \mu_\tau v_x + \mu_\rho R_{\frac{\pi}{2}}v + f(v, v_x), \\ 0 &= \langle \hat{v}', v - \hat{v}\rangle_{\mathcal{L}_2}, \quad 0 = \langle R_{\frac{\pi}{2}}v, v - \hat{v}\rangle_{\mathcal{L}_2}. \end{aligned} \tag{1.12}$$

The relative equilibria are rotating and traveling waves

$$\bar{u}(x,t) = R_{-\bar{\mu}_\rho t}\bar{v}(x - \bar{\mu}_\tau t). \tag{1.13}$$

In the previous example our theory will apply only to pulses $\bar{v}$ ($\bar{v}_- = \bar{v}_+$) with arbitrary $\bar{\mu}_\rho$ and to fronts ($\bar{v}_- \neq \bar{v}_+$) with $\bar{\mu}_\rho = 0$. However, in the case of a front with $\bar{\mu}_\rho \neq 0$ we cannot expect (1.12) to be a well-posed problem in $\mathcal{L}_2$. On a formal level the condition $S^2(\bar{v}) = R_{\frac{\pi}{2}}\bar{v} \in \mathcal{L}_2$ of Assumption A1 is violated. On an intuitive level we expect that any $\mathcal{L}_2$ perturbation of the right-hand side of (1.12) will lead to an angular shift (in direction $R_{\frac{\pi}{2}}$) that does not decay at infinity and hence is not in $\mathcal{L}_2$.

1.3. Spectral properties of the continuous system. In order to deal with the truncation to a finite interval, we need some information of the behavior at $x \to \pm\infty$. More specifically we need to ensure the existence of exponential dichotomies for the linear operators obtained by linearizing the continuous equation (1.8) about the relative equilibrium $(\bar{v}, \bar{\mu})$. These dichotomies are then inherited by the discretized equations and allow us to deal with the influence of the boundary conditions.

The linearization of (1.8) w.r.t. v at $(\bar{v}, \bar{\mu})$ is given by

$$\Lambda v = Av'' + Bv' + Cv, \tag{1.14}$$

where

$$B(x) = D_2 f(\bar{v}(x), \bar{v}'(x)) + \sum_{i=1}^{p} \bar{\mu}_i S_1^i, \quad C(x) = D_1 f(\bar{v}(x), \bar{v}'(x)) + \sum_{i=1}^{p} \bar{\mu}_i S_0^i.$$

Assumption (1.10) implies with the invertibility of A that $\lim_{x\to\pm\infty} \bar{v}'(x) = 0$. Thus Λ converges for $x \to \pm\infty$ to constant coefficient operators

$$\Lambda_\pm v = Av'' + B_\pm v' + C_\pm v, \qquad B_\pm = \lim_{x\pm\infty} B(x), \qquad C_\pm = \lim_{x\pm\infty} C(x). \tag{1.15}$$

The main spectral assumption on Λ is the following.

Hypothesis 1.5.

1. *Eigenvalue condition*: The functions $S^i(\bar{v}) \in \mathcal{L}_2$, $i = 1, \dots, p$, are linearly independent and span the null space of $\Lambda : \mathcal{H}^2 \to \mathcal{L}_2$, i.e.,

$$\mathcal{N}(\Lambda) = \operatorname{span}\{S^1(\bar{v}), \dots, S^p(\bar{v})\}.$$

   Moreover, the algebraic and the geometric multiplicity of zero both equal p.
2. *Spectral condition*: The eigenvalue 0 lies in the connected component of $\mathbb{C} \setminus \{\Sigma_+ \cup \Sigma_-\}$ that contains a right half plane, where

$$\Sigma_\pm = \{s \in \mathbb{C} : \det(-\kappa^2 A + i\kappa B_\pm + C_\pm - sI) = 0 \text{ for some } \kappa \in \mathbb{R}\}.$$

Note that spectral properties of Λ determine asymptotic stability with an asymptotic phase of relative equilibria [11]. However, the spectral conditions given above are weaker than the standard stability conditions and allow us to discuss the approximation of unstable equilibria as well.

Example 1.6. For Example 1.4 the operator Λ reads

$$\Lambda v = Av'' + (\mu_\tau I + D_2 f(\bar v, \bar v'))v' + (\mu_\rho R_{\frac{\pi}{2}} + D_1 f(\bar v, \bar v'))v,$$

and its null space is spanned by $\bar v'$ and $R_{\frac{\pi}{2}}\bar v$.

The condition that zero is semisimple cannot be concluded from properties of the Lie group G. For the cubic Ginzburg–Landau equation which has the symmetry properties of Example 1.4 and is a special case of the system considered in section 5, a principal vector exists for a special parameter combination [2]; i.e., the algebraic multiplicity is 3 and the geometric multiplicity is 2.

2. Discretization. In this section we first introduce the discretization and the relevant notions for the double limit process of increasing the interval and decreasing the step size. Then we state the main nonlinear approximation theorems for the approximation of relative equilibria and the approximation of isolated eigenvalues and corresponding invariant subspaces.

2.1. Problem formulation. In order to compute numerical approximations of $(\bar v, \bar\mu)$ we define a discrete interval

$$J = [n_-, n_+] = \{n \in \mathbb{Z} :\ n_- \le n \le n_+,\ \text{where } n_\pm \in \mathbb{Z} \cup \{\pm\infty\}\ \}$$

and a corresponding equidistant grid with grid size $h > 0$

$$J_h = \{x_n :\ x_n = nh,\ n \in J\}.$$

We write $J_h \to \mathbb{R}$ if $h \to 0$ and simultaneously $h \cdot \min\{-n_-, n_+\} \to \infty$, i.e., $\pm n_\pm$ grows faster than h decreases, so that $[hn_-, hn_+] \to \mathbb{R}$.

We denote the Banach space of sequences in $\mathbb{R}^m$ which are indexed by J provided with the supremum norm $\|z\|_\infty = \sup_{n\in J}\|z_n\|$ by $\ell^J_\infty(\mathbb{R}^m)$ and write $\ell_\infty(\mathbb{R}^m)$ for $\ell^{\mathbb{Z}}_\infty(\mathbb{R}^m)$. If no confusion is possible, we drop the argument $\mathbb{R}^m$ and write just ℓ^J_∞ and ℓ_∞.

Let the standard finite difference operators on the extended grid

$$\hat J_h = \Big\{x_n :\ x_n = nh,\ n \in \hat J = [n_- - 1, n_+ + 1]\Big\}$$

be given by

$$\begin{aligned}
&\delta_0 : \ell^{\hat J}_\infty \to \ell^J_\infty, && (\delta_0 v)_n = \frac{1}{2h}(v_{n+1} - v_{n-1}),\\
&\delta_+ : \ell^{[n_-, n_+ + 1]}_\infty \to \ell^J_\infty, && (\delta_+ v)_n = \frac{1}{h}(v_{n+1} - v_n),\\
&\delta_- : \ell^{[n_- - 1, n_+]}_\infty \to \ell^J_\infty, && (\delta_- v)_n = \frac{1}{h}(v_n - v_{n-1}).
\end{aligned}$$

For sequences $u, v \in \ell^J_\infty$, $J = [n_-, n_+]$ we define the inner product and discrete Sobolev norms by

$$\begin{aligned}
&\langle u, v\rangle_{J_h} = \sum_{n=n_-}^{n_+} h u_n^T v_n, && \|u\|_{\mathcal{L}_{2,h}} = \sqrt{\langle u, u\rangle_{J_h}},\\
&\|u\|_{\mathcal{H}^1_h} = \|u\|_{\mathcal{L}_{2,h}} + \|\delta_+ u\|_{\mathcal{L}_{2,h}}, && \|u\|_{\mathcal{H}^2_h} = \|u\|_{\mathcal{H}^1_h} + \|\delta_+\delta_- u\|_{\mathcal{L}_{2,h}}.
\end{aligned}$$

One has to keep in mind that the summation is done on different sets for the different difference operators. If necessary, we embed each $u \in \ell_\infty^J$ in ℓ_∞ by setting $u_n = 0$ for $n \in \mathbb{Z} \setminus J$ without further notice. For matrices $U = [U^1, \dots, U^p]$ and $V = [V^1, \dots, V^p] \in (\ell_\infty^J(\mathbb{R}^m))^p$ we use the notation $\langle U, V \rangle_{J_h} = [\langle U^i, V^j \rangle_{J_h}]_{i=1,\dots,p}^{j=1,\dots,p} \in \mathbb{R}^{p,p}$. This notation is also adopted for the $\mathcal{L}_2$ inner product $\langle \cdot, \cdot \rangle_{\mathcal{L}_2}$ in an analogous fashion.

We discretize (1.8) and (1.11) and add linear boundary conditions

$$\mathcal{B}v = P_- v_{n_-} + Q_-(\delta_0 v)_{n_-} + P_+ v_{n_+} + Q_+(\delta_0 v)_{n_+}, \quad P_\pm, Q_\pm \in \mathbb{R}^{2m,m}. \tag{2.1}$$

This leads to the discrete boundary value problem

$$A(\delta_+\delta_- v)_n + \hat{S}_n(v)\mu + f(v_n, (\delta_0 v)_n) = 0, \quad n \in J, \tag{2.2a}$$

$$\mathcal{B}v = \eta, \tag{2.2b}$$

$$\langle \hat{S}^i(\hat{v}_{|_{J_h}}), v - \hat{v}_{|_{J_h}} \rangle_{J_h} = 0, \quad i = 1, \dots, p, \tag{2.2c}$$

where $\hat{S}_n^i(v) = S_0^i v_n + S_1^i(\delta_0 v)_n \in \mathbb{R}^m$ and $\hat{S}_n(v)\mu = \sum_{i=1}^p \mu_i \hat{S}_n^i(v)$. For the approximation theorem to hold, we need boundary conditions which are consistent with the behavior of the exact solution at $\pm\infty$; i.e., we will require $\eta = P_-\bar{v}_- + P_+\bar{v}_+$ in Hypothesis 2.4.

Linearizing the left-hand side of (2.2a) at a solution $(\tilde{v}, \tilde{\mu}) \in \ell_\infty^{\hat{J}} \times \mathbb{R}^p$ of (2.2) leads to the following discrete approximation of Λ:

$$\left(\tilde{\Lambda} v\right)_n = A(\delta_+\delta_- v)_n + \tilde{B}_n(\delta_0 v)_n + \tilde{C}_n v_n, \quad n \in J, \tag{2.3}$$

where

$$\tilde{B}_n = D_2 f(\tilde{v}_n, (\delta_0 \tilde{v})_n) + \sum_{i=1}^p \tilde{\mu}_i S_1^i, \quad \tilde{C}_n = D_1 f(\tilde{v}_n, (\delta_0 \tilde{v})_n) + \sum_{i=1}^p \tilde{\mu}_i S_0^i.$$

It is well known that a multiple eigenvalue will split into several eigenvalues under perturbations [12]. Therefore we approximate the whole invariant subspace which corresponds to zero instead of dealing with each eigenfunction separately. The spatial discretization of the invariant subspace equation together with boundary conditions and a normalizing condition leads for the unknowns (V, Σ) to the system

$$\left(\tilde{\Lambda} V\right)_n - V_n \Sigma = 0, \quad n \in J, \tag{2.4a}$$

$$\mathcal{B}V = 0, \tag{2.4b}$$

$$\left\langle \hat{X}_{|_{J_h}}, V \right\rangle_{J_h} = I. \tag{2.4c}$$

Here $\hat{X} = [\hat{X}^1, \dots, \hat{X}^p] : \mathbb{R} \to \mathbb{R}^{m,p}$ is a given normalizing function for which $\langle \hat{X}, S(\bar{v}) \rangle_{\mathcal{L}_2} = I$. The columns of $V \in (\ell_\infty^{\hat{J}})^p$ span the invariant subspace which corresponds to the eigenvalues of $\Sigma \in \mathbb{R}^{p,p}$. Thus for V to be a good approximation of the null space of Λ, the norm of Σ has to be small.

2.2. Main results. Before we can state the main approximation results, we have to collect the necessary hypotheses on the boundary conditions and the phase condition. The template function $\hat{v}$ and the normalizing function $\hat{X}$ are assumed to be in the following class of functions.

DEFINITION 2.1. *We define a function* $g : I \to \mathbb{R}^{m,p},\ I \subset \mathbb{R}$, *to be in* $\mathcal{E}_\varrho(I, \mathbb{R}^{m,p})$ *if there exists* $K > 0$ *such that* $\forall x \in I$*:*

$$\|g(x)\| \le K\mathrm{e}^{-\varrho|x|} \quad \textit{and} \quad \|g'(x)\| \le K\mathrm{e}^{-\varrho|x|}.$$

Similar to [29] we have the following lemma.

LEMMA 2.2. *Let* $g \in \mathcal{E}_\varrho(\mathbb{R}^+, \mathbb{R}^{m,p})$ *and* $\tilde{g} \in (\ell_\infty)^p$ *be given with*

$$\|g(x_n) - \tilde{g}_n\| \le \mathrm{const}\, h\mathrm{e}^{-\varrho x_n} \quad \forall n \in \mathbb{N}.$$

Then the following estimate holds:

$$\left\| \int_0^\infty g(x)\, dx - h \sum_{n=0}^{n_+-1} \tilde{g}_n \right\| \le \mathrm{const}(h + \mathrm{e}^{-\varrho x_{n_+}}). \tag{2.5}$$

A well-known determinant condition [3], which ensures the existence of the resolvent on finite intervals in the continuous case, uses the following determinant.

DEFINITION 2.3. *Define*

$$\mathcal{D} = \det\left(\begin{pmatrix} P_- & Q_- \end{pmatrix} \begin{pmatrix} Y_-^s \\ Y_-^s \Sigma_-^s \end{pmatrix} \quad \begin{pmatrix} P_+ & Q_+ \end{pmatrix} \begin{pmatrix} Y_+^u \\ Y_+^u \Sigma_+^u \end{pmatrix} \right), \tag{2.6}$$

where $(\Sigma_-^s, Y_-^s),\ (\Sigma_+^u, Y_+^u) \in \mathbb{R}^{m,m} \times \mathbb{R}^{m,m}$ *solve the quadratic eigenvalue problems*

$$AY\Sigma^2 + B_\pm Y\Sigma + C_\pm Y = 0 \tag{2.7}$$

with $\operatorname{Re}\sigma(\Sigma_-^s) < 0$ *and* $\operatorname{Re}\sigma(\Sigma_+^u) > 0$.

Then we can formulate the determinant condition and a consistency assumption for the boundary conditions as well as a regularity assumption for the phase condition.

Hypothesis 2.4.

1. *Boundary conditions*: The boundary condition (2.2b) is satisfied at the stationary points $\bar{v}_\pm$, i.e.,

   $$\eta = P_-\bar{v}_- + P_+\bar{v}_+,$$

   and the determinant $\mathcal{D}$ defined in (2.6) is nonzero.
2. *Phase condition*: The phase condition is satisfied by $\bar{v}$, i.e.,

   $$\langle S(\hat{v}), \bar{v} - \hat{v} \rangle_{\mathcal{L}_2} = 0, \tag{2.8}$$

   $\bar{v} - \hat{v} \in \mathcal{H}^1$, $S(\hat{v}) \in \mathcal{E}_\varrho(\mathbb{R}, \mathbb{R}^{m,p})$, and the $p \times p$ matrix

   $$\langle S(\hat{v}), S(\bar{v}) \rangle_{\mathcal{L}_2} = \int_{\mathbb{R}} [S(\hat{v})](x)^T [S(\bar{v})](x)\, dx \tag{2.9}$$

   is nonsingular.

The determinant condition on the boundary conditions may seem hard to check but is satisfied for Dirichlet, Neumann, and periodic boundary conditions (see the discussion in [3]). Condition (2.8) is a technical assumption that prevents the approximate solution from converging to a shifted version of $\bar{v}$, which can be dropped. Then one has to deal with an additional phase shift, which can be estimated in terms of $\bar{v}$ and $\hat{v}$, but we do not pursue this here.

Now we can formulate the main approximation results of the paper.

THEOREM 2.5 (approximation of the relative equilibrium). *Assume Hypotheses* 1.5 *and* 2.4. *Then there exist* $\varrho > 0$, $T > 0$, *and* $h_0 > 0$ *such that for* $h < h_0$ *and* $\pm h n_\pm > T$ *the boundary value problem* (2.2) *has a unique solution* $(\tilde{v}, \tilde{\mu})$ *in a neighborhood* $B_\varrho(\bar{v}, \bar{\mu}) = \{(v, \mu) \in \ell^{\hat{J}}_\infty \times \mathbb{R}^p : \|\bar{v}_{|J_h} - v\|_{\mathcal{H}^2_h} + \|\bar{\mu} - \mu\| < \varrho\}$. *This solution obeys the following estimate for some* $\alpha > 0$*:*

$$\|\bar{v}_{|J_h} - \tilde{v}\|_{\mathcal{H}^2_h} + \|\bar{\mu} - \tilde{\mu}\| \le \mathrm{const}(h^2 + \mathrm{e}^{-\alpha h \min\{-n_-, n_+\}}). \tag{2.10}$$

The second result is an approximation theorem for the zero eigenvalue of Λ and the corresponding invariant subspace.

THEOREM 2.6 (approximation of the zero eigenvalue). *Assume Hypothesis* 1.5 *and* $\hat{X} \in \mathcal{E}_\beta(\mathbb{R}, \mathbb{R}^{m,p})$ *for some* $\beta > 0$. *Then there exist* $\varrho > 0$, $T > 0$, *and* $h_0 > 0$ *such that for* $h < h_0$ *and* $\pm h n_\pm > T$ *the eigenvalue problem* (2.4) *has a unique solution* (X, Σ) *in a neighborhood* $B_\varrho(S(\bar{v})_{|J_h}, 0) := \{(X, \Sigma) \in ((\ell^{\hat{J}}_\infty)^p \times \mathbb{R}^{p,p} : \|S(\bar{v})_{|J_h} - X\|_{\mathcal{H}^2_h} + \|\Sigma\| < \varrho\}$. *This solution obeys the following estimate for some* $\alpha > 0$*:*

$$\|S(\bar{v})_{|J_h} - X\|_{\mathcal{H}^2_h} + \|\Sigma\| \le \mathrm{const}(h^2 + \mathrm{e}^{-\alpha \min\{-n_-, n_+\}}). \tag{2.11}$$

Remark 2.7.

- The above result can be easily generalized to invariant subspaces which are related to other isolated eigenvalues of finite multiplicity.
- The stated results hold for $\|\cdot\|_\infty$ as well, and estimates similar to (2.10) and (2.11) can be shown [25].
- Estimate (2.11) implies with standard perturbation results for eigenvalues [12] the convergence of the eigenvalues as $J_h \to \mathbb{R}$.

3. The linear equation. Before we can state the main linear theorem in this section, we have to collect some results from the theory of exponential dichotomies [19], [20]. First we transfer exponential dichotomies for the continous system to the discretized system. Particular attention is paid to the dichotomy constants, which should be uniform for both large intervals and small step sizes. This is relevant for constructing solutions of the linearized problem in section 3.2 and deriving estimates for them.

3.1. Exponential dichotomies and auxiliary results. Consider the linear differential operator

$$Lz = z' - M(\cdot)z, \quad x \in J \subset \mathbb{R},\ M(\cdot) : J \to \mathbb{R}^{k,k} \tag{3.1}$$

with solution operator $\mathcal{S}(x, \xi)$.

DEFINITION 3.1 (exponential dichotomy, continuous case). *The operator* L *defined in* (3.1) *has an exponential dichotomy in the interval* $J = [x_-, x_+]$, $x_\pm \in \mathbb{R} \cup \{\pm\infty\}$, *with data* (K, α, π) *if there exist a bound* $K > 0$, *a rate* $\alpha > 0$, *and a function* $\pi : J \ni x \mapsto \pi(x)$, $\pi(x)$ *a projector, such that the following holds:*

$$\mathcal{S}(x, \xi)\pi(\xi) = \pi(x)S(x, \xi), \tag{3.2}$$

and the Green's function

$$\mathcal{G}(x, \xi) = \begin{cases} \mathcal{S}(x, \xi)\pi(\xi), & x \ge \xi, \\ -\mathcal{S}(x, \xi)(I - \pi(\xi)), & x < \xi, \end{cases}$$

satisfies the exponential estimate

$$\|\mathcal{G}(x,\xi)\| \le K\mathrm{e}^{-\alpha|x-\xi|}, \quad x,\xi \in J. \tag{3.3}$$

In the following remarks we recall some well known facts about exponential dichotomies (see [19], [7]).

Remark 3.2. If L has an exponential dichotomy on $\mathbb{R}^- = (-\infty, 0]$ and $\mathbb{R}^+ = [0,\infty)$ with data (K_-, α_-, π_-) and (K_+, α_+, π_+), respectively, then the kernel of L is given by

$$\mathcal{N}(L) = \{\mathcal{S}(\cdot,0)z_0 : \ z_0 \in \mathcal{N}(\pi_-(0)) \cap \mathcal{R}(\pi_+(0))\}.$$

Remark 3.3. We transform Λ via $z = (v, v')$ into a first order operator

$$Lz = z' - M(\cdot)z, \qquad M(x) = \begin{pmatrix} 0 & I \\ -A^{-1}C(x) & -A^{-1}B(x) \end{pmatrix}. \tag{3.4}$$

Hypothesis 1.5 ensures together with the positive definiteness of A that L possesses exponential dichotomies on $\mathbb{R}^\pm$ with data $(K^\pm, \alpha^\pm, \pi^\pm)$. Moreover, the null space of L is given by

$$\mathcal{N}(L) = \operatorname{span}\{\phi^1, \dots, \phi^p\}, \quad \phi^i = (S^i(\bar v), S^i(\bar v'))^T. \tag{3.5}$$

Note that the columns of $X^s_- = \left(\begin{smallmatrix} Y^s_- \\ Y^s_-\Sigma^s_- \end{smallmatrix}\right)$ and $X^u_+ = \left(\begin{smallmatrix} Y^u_+ \\ Y^u_+\Sigma^u_+ \end{smallmatrix}\right)$, where (Y^s_-, Σ^s_-) and $(Y^u_+\Sigma^u_+)$ are the solutions of the characteristic equation (2.7) in Definition 2.3, span the stable invariant subspace of $M_- = \lim_{x\to-\infty} M(x)$ and the unstable invariant subspace of $M_+ = \lim_{x\to+\infty} M(x)$, respectively.

Remark 3.4. The fact that L defined in (3.4) has exponential dichotomies on $\mathbb{R}^-, \mathbb{R}^+$ implies that $(\bar v', \bar v'')$ is actually exponentially decaying for $x \to \pm\infty$, i.e.,

$$\|\bar v(x) - v_\pm\| \le \text{const}\,\mathrm{e}^{\mp\varrho x} \quad \text{as well as} \quad \|\bar v^{(k)}(x)\| \le \text{const}\,\mathrm{e}^{-\varrho|x|}, \ k = 1, 2,$$

for some $\varrho > 0$. Thus we have $\bar v' \in \mathcal{E}_\varrho(\mathbb{R}, \mathbb{R}^m)$ and $\lim_{x\to\infty}[S^i(\bar v')](x) = 0$ for $i = 1, \dots, p$. Moreover, from the eigenvalue condition we obtain $S^i(\bar v) \in \mathcal{H}^2$ and thus $\lim_{x\to\infty}[S^i(\bar v)](x) = 0$, $i = 1, \dots, p$.

Exponential dichotomies can be defined for a linear difference operator

$$(\hat L z)_n = z_{n+1} - \hat M_n z_n, \quad n \in J, \ \hat M_n \in \mathbb{R}^{k,k}, \tag{3.6}$$

in the following way. If the matrices $\hat M_n$ are invertible for all $n \in J$, then the map $\hat{\mathcal{S}} : J \times J \to \mathbb{R}^{k,k}$ given by

$$\hat{\mathcal{S}}(n,m) = \begin{cases} \hat M_{n-1} \cdot \ldots \cdot \hat M_m & \text{for } n > m, \\ I & \text{for } n = m, \\ \hat M_n^{-1} \cdot \ldots \cdot \hat M_{m-1}^{-1} & \text{for } n < m \end{cases}$$

is a solution operator for (3.6), which has the cocycle property

$$\hat{\mathcal{S}}(n,l)\hat{\mathcal{S}}(l,m) = \hat{\mathcal{S}}(n,m) \quad \forall\, l, m, n \in J.$$

Definition 3.5 (exponential dichotomy, discrete case). *The linear difference operator $\hat L$ defined in* (3.6) *has an exponential dichotomy with data (K, α, P) on $J \subset \mathbb{Z}$*

if $\hat{M}_n$ is invertible for all $n \in J$ and there exist a bound $K > 0$, a rate $\alpha > 0$, and projectors P_n such that the following holds:

$$\hat{\mathcal{S}}(n,m)P_m = P_n\hat{\mathcal{S}}(n,m), \tag{3.7}$$

and the Green's function

$$\hat{\mathcal{G}}(n,m) = \begin{cases} \hat{\mathcal{S}}(n,m)P_m & \text{for } n \geq m, \\ -\hat{\mathcal{S}}(n,m)(I-P_m) & \text{for } n < m \end{cases} \tag{3.8}$$

satisfies the exponential estimate

$$\|\hat{\mathcal{G}}(n,m)\| \leq K\mathrm{e}^{-\alpha|n-m|}, \quad n,m \in J. \tag{3.9}$$

A connection between an exponential dichotomy for an ODE and the corresponding difference equation is given via the time h-map of the flow [29].

LEMMA 3.6. *Let the linear differential operator L given by*

$$Lz = z' - Mz, \quad x \in J = \mathbb{R}^{\pm}, \quad M : J \to \mathbb{R}^{k,k}$$

have an exponential dichotomy with data (K,α,π) on $J = \mathbb{R}^{\pm}$ and define

$$\hat{\mathcal{S}}(n,m) = \mathcal{S}(x_n, x_m).$$

Then the difference operator

$$\hat{L}z = (z_{n+1} - \hat{\mathcal{S}}(n+1,n)z_n)_{n\in\hat{J}}$$

has an exponential dichotomy on $\hat{J} = \mathbb{Z}^{\pm}$ with data $(K, \alpha h, P)$, where $P_n = \pi(x_n)$.

3.2. The linear difference equation. The main result in this section deals with the existence of solutions $(Z,\Upsilon) \in (\ell_\infty^{[n_-,n_+ +1]}(\mathbb{R}^{2m}))^q \times \mathbb{R}^{p,q}$, $1 \leq q \leq p$, of the following linear inhomogeneous boundary value problem:

$$Z_{n+1} - \hat{M}_n Z_n - \hat{V}_n\Upsilon = R^{de}, \quad n \in J = [n_-, n_+], \tag{3.10a}$$

$$(P_-Q_-)Z_{n_-} + (P_+Q_+)Z_{n_+} = R^{bc} \in \mathbb{R}^{2m,q}, \tag{3.10b}$$

$$\langle\hat{\Psi}, Z\rangle_{J_h} = R^{pc} \in \mathbb{R}^{p,q}, \tag{3.10c}$$

where

$$\hat{M}_n = \mathcal{S}(x_{n+1}, x_n) \in \mathbb{R}^{2m,2m}, \tag{3.11}$$

$\mathcal{S}$ denotes the solution operator corresponding to L defined in (3.4), and $\hat{\Psi} \in \ell_\infty^J(\mathbb{R}^{2m,p})$ and $\hat{V}_n \in \mathbb{R}^p$ satisfy the following hypothesis.

Hypothesis 3.7.

1. *Phase condition*: There exists a continuous template function $\Psi \in \mathcal{E}_\varrho(\mathbb{R}, \mathbb{R}^{2m,p})$ such that $\|\hat{\Psi} - \Psi_{|_{J_h}}\|_{\mathcal{L}_{2,h}} \to 0$ as $J_h \to \mathbb{R}$ and $\langle\Psi,\Phi\rangle_{\mathcal{L}_{2,h}} \in \mathbb{R}^{p,p}$ is nonsingular, where

   $$\Phi = [\phi^1, \ldots, \phi^p] \tag{3.12}$$

   and ϕ^i, $i = 1, \ldots, p$, are the functions which span the null space of L, which have been defined in (3.5).

2. *Nondegeneracy condition*: The matrices $\hat{V}_n$ have the form

$$\hat{V}_n = hV(x_n) + \mathcal{O}(h^2) \in \mathbb{R}^{2m,p}, \tag{3.13}$$

where $V \in \mathcal{L}_2(\mathbb{R}, \mathbb{R}^{2m,p})$, $\|V\|_\infty < \infty$, $\operatorname{rank}(V) = p$, and $\mathcal{R}(L) \cap \mathcal{R}(V) = \{0\}$.

Remark 3.8. Define $\Xi = [\xi_1, \dots, \xi_p]$, where ξ_i, $i = 1, \dots, p$, span the null space of the adjoint operator

$$L^* : z \mapsto -z' - M^T(\cdot)z.$$

Then the eigenvalue condition in Hypothesis 1.5 implies that the Melnikov integral

$$\langle \Xi, V \rangle = \int_{\mathbb{R}} [\xi_1, \dots, \xi_p](x)^T V(x)\, dx \ \in \mathbb{R}^{p,p}$$

is nonsingular.

It can be shown that Hypothesis 3.7 implies the invertibility of the operator $\left(\begin{smallmatrix} L & -V \\ \Psi & 0 \end{smallmatrix}\right)$ by the bordering lemma [13].

Now we can formulate the main linear existence result from which we obtain the existence of solutions of (3.10) as well as corresponding estimates.

THEOREM 3.9. *Consider* (3.10)*, and let Hypothesis* 3.7 *be satisfied.*

Then there exist $h_0 > 0$ *and* $T > 0$ *such that for* $h < h_0$ *and* $\pm h n_\pm > T$ (3.10) *has a unique solution* $(\tilde{Z}, \tilde{\Upsilon}) \in (\ell_\infty^{[n_-, n_+ + 1]}(\mathbb{R}^{2m}))^q \times \mathbb{R}^{p,q}$ *for any* $R^{de} \in (\ell_\infty^J(\mathbb{R}^{2m}))^q$, $R^{bc} \in \mathbb{R}^{2m,q}$, $R^{pc} \in \mathbb{R}^{p,q}$. *Furthermore, the following estimate holds:*

$$\|\tilde{Z}\|_{\mathcal{H}_h^1} + \|\tilde{\Upsilon}\| \le \text{const} \left(\frac{1}{h} \|R^{de}\|_{\mathcal{L}_{2,h}} + \|R^{bc}\| + \|R^{pc}\| \right). \tag{3.14}$$

Note that in the traveling wave case we have $p = 1$ and $q = 1$, whereas for the approximation of general equilibria we have $p = \dim G$, and $q = 1$, and for the approximation of the null space we have $p = q = \dim G$.

The existence of exponential dichotomies ensures that certain boundary value problems can be solved that arise later in the construction of solutions of (3.10). We use a slightly adapted version of Lemma 1.1.6 in [4] or Lemma 2.7 in [20] in which the dependency on h is addressed.

LEMMA 3.10. *Let the linear difference operator*

$$\hat{L} : \ell_\infty^{[n_-, n_+ + 1]} \to \ell_\infty^J, z \mapsto \left(z_{n+1} - \hat{M}_n z_n \right)_{n \in J}$$

have an exponential dichotomy with data (K, β, P) *on* $J = [n_-, n_+] \subset \mathbb{Z}$, *where* $n_\pm = \pm\infty$ *is allowed.*

For each $r \in \ell_\infty^J$ *there exists a unique solution* $\tilde{z} \in \ell_\infty^{[n_-, n_+ + 1]}$ *of the inhomogeneous equation*

$$\begin{aligned} \left(\hat{L} z\right)_n &= r_n \quad n \in J, \\ P_{n_-} z_{n_-} &= \varrho_- \in \mathcal{R}(P_{n_-}) \qquad \text{if } n_- \in \mathbb{Z}, \\ (I - P_{n_+}) z_{n_+} &= \varrho_+ \in \mathcal{R}(I - P_{n_+}) \quad \text{if } n_+ \in \mathbb{Z}. \end{aligned}$$

It is given by

$$\tilde{z}_n = b_n^-(\varrho_-) + b_n^+(\varrho_+) + \hat{s}_n(r), \quad n \in J,$$
$$\tilde{z}_{n_+ +1} = \hat{M}_{n_+}\tilde{z}_{n_+} + r_{n_+},$$

where $\hat{s}$ is defined with the Green's function $\hat{\mathcal{G}}$ from (3.8) *as follows:*

$$\hat{s}_n(r) = \sum_{m=n_-}^{n_+-1} \hat{\mathcal{G}}(n, m+1)\ r_m, \quad n \in J, \tag{3.15}$$

and

$$b_n^\pm(\varrho) = \begin{cases} \hat{\mathcal{S}}(n, n_\pm)\varrho & \text{in case } \pm n_\pm < \infty, \\ 0 & \text{otherwise.} \end{cases}$$

Furthermore, the following estimate holds for $n \in J$:

$$\|\hat{s}_n(r)\| \le K\ C_\beta \|r\|_\infty, \quad \text{where } C_\beta = \frac{1+\mathrm{e}^{-\beta}}{1-\mathrm{e}^{-\beta}}. \tag{3.16}$$

In addition, if $r \in \mathcal{L}_{2,h}(J)$, then

$$\|\hat{s}_n(r)\| \le K\sqrt{\frac{C_{2\beta}}{h}}\ \|r\|_{\mathcal{L}_{2,h}} \quad \forall n \in J \quad \text{and} \quad \|\hat{s}(r)\|_{\mathcal{L}_{2,h}} \le KC_\beta\ \|r\|_{\mathcal{L}_{2,h}}. \tag{3.17}$$

In case $\pm n_\pm < \infty$ we obtain for the boundary terms the estimates

$$\|b_n^\pm(\varrho_\pm)\| \le K\mathrm{e}^{-\beta|n-n_\pm|}\|\varrho_\pm\|,\ n \in J, \quad \text{and} \quad \|b^\pm(\varrho_\pm)\|_{\mathcal{L}_{2,h}} \le K\sqrt{hC_{2\beta}}\|\varrho_\pm\|.$$

Note that C_β does not depend on the interval J but only on the dichotomy data and that $C_{h\alpha}$ is of order $\mathcal{O}(\frac{1}{h})$ for small h. Thus we can estimate the solutions $\tilde{z}$ for $\beta = \alpha h$ as follows.

COROLLARY 3.11. *If $\beta = \alpha h$, then the partial solutions $\tilde{z}$ defined in Lemma* 3.10 *obey the estimate*

$$\|\tilde{z}\|_{\mathcal{L}_{2,h}} \le \text{const}\left(\frac{1}{h}\|r\|_{\mathcal{L}_{2,h}} + \|\varrho_-\| + \|\varrho_+\|\right). \tag{3.18}$$

In the following we transfer the proof in [4] to the discrete case along the lines of the method used in [29] and [30].

Proof of Theorem 3.9. From the properties of L (see Remark 3.3) one obtains that L has exponential dichotomies on $\mathbb{R}^\pm$ with data $(K_\pm, \alpha_\pm, \pi^\pm)$. Then Lemma 3.6 implies that the operator $\hat{L} : \ell_\infty \to \ell_\infty$ defined by

$$\hat{L}z = \left(z_{n+1} - \hat{M}_n z_n\right)_{n\in\mathbb{Z}}$$

possesses exponential dichotomies on $\mathbb{Z}^\pm$ with data $(K_\pm, \alpha_\pm h, P^\pm)$ and

$$\mathcal{N}\left(\hat{L}\right) = \text{span}\{\hat{\phi}^1, \dots, \hat{\phi}^p\}, \quad \text{where } \hat{\phi}^i = \phi^i{}_{|_{J_h}} = \begin{pmatrix} S^i(\bar{v}) \\ S^i(\bar{v}') \end{pmatrix},\ i = 1, \dots, p. \tag{3.19}$$

We use Lemma 3.10 to define solutions of the boundary value problems

$$\begin{aligned}\left(\hat{L}Z\right)_n &= R^{de}_n, \quad n \in [n_-, -1],\\ P^-_{n_-} Z_{n_-} &= \rho_- \in (\mathcal{R}(P^-_{n_-}))^q,\\ (I - P^-_0)Z_0 &= Z^-_0 \in (\mathcal{N}(P^-_0))^q\end{aligned}$$

and

$$\begin{aligned}\left(\hat{L}Z\right)_n &= R^{de}_n, \quad n \in [0, n_+],\\ P^+_0 Z_0 &= Z^+_0 \in \left(\mathcal{R}\left(P^+_0\right)\right)^q,\\ (I - P^+_{n_+})Z_{n_+} &= \rho_+ \in \left(\mathcal{N}\left(P^+_{n_+}\right)\right)^q,\end{aligned}$$

where we use the notation $(\hat{L}Z)_n := [(\hat{L}Z^1)_n, \dots, (\hat{L}Z^q)_n] \in \mathbb{R}^{m,q}$. These are for each $R^{de} \in (\ell^J_\infty)^q$ defined by

$$\begin{aligned}\tilde{Z}^-_n &= \hat{s}^-_n \left(R^{de} + \hat{V}^\Upsilon\right) + \hat{\mathcal{S}}(n,0)Z^-_0 + \hat{\mathcal{S}}(n,n_-)\rho_-, \quad n \in [n_-, 0],\\ \tilde{Z}^+_n &= \hat{s}^+_n \left(R^{de} + \hat{V}^\Upsilon\right) + \hat{\mathcal{S}}(n,0)Z^+_0 + \hat{\mathcal{S}}(n,n_+)\rho_+, \quad n \in [0, n_+],\\ \tilde{Z}^+_{n_++1} &= \hat{M}_{n_+}\tilde{Z}^+_{n_+} + R^{de}_{n_+} \in \mathbb{R}^{2m,q},\end{aligned}$$

where $\hat{V}^\Upsilon_n = \hat{V}_n\Upsilon \in \mathbb{R}^{2m,q}$.

Then $\tilde{Z} \in (\ell^{[n_-,n_++1]}_\infty)^q$ defined by

$$\tilde{Z}_n = \begin{cases}\tilde{Z}^-_n & \text{for } n \in J^- := [n_-, -1],\\ \tilde{Z}^+_n & \text{for } n \in J^+ := [0, n_+ + 1]\end{cases} \tag{3.20}$$

is a solution of (3.10) if it solves the following system:

$$\tilde{Z}^-_0 = \tilde{Z}^+_0 \quad \in \mathbb{R}^{2m,q}, \tag{3.21a}$$

$$(P_-Q_-)\tilde{Z}_{n_-} + (P_+Q_+)\tilde{Z}_{n_+} = R^{bc} \quad \in \mathbb{R}^{2m,q}, \tag{3.21b}$$

$$\left\langle \hat{\Psi}, \tilde{Z} \right\rangle_{J_h} = R^{pc} \quad \in \mathbb{R}^{p,q}. \tag{3.21c}$$

Note that the dependency on $\Upsilon \in \mathbb{R}^{p,q}$ is hidden in the definition of $\tilde{Z}^\pm_n$ and is yet to be determined.

We decompose $\mathbb{R}^{2m}$ as follows: Let $\mathcal{W}_1 = \mathcal{R}(P^+_0) \cap \mathcal{N}(P^-_0)$. From (3.19) we obtain $\dim(\mathcal{W}_1) = p$, and thus we can complement $\mathcal{W}_1$ by subspaces $\mathcal{W}_2$ and $\mathcal{W}_3$ such that

$$\mathcal{R}\left(P^+_0\right) = \mathcal{W}_1 \oplus \mathcal{W}_2, \quad \mathcal{N}\left(P^-_0\right) = \mathcal{W}_1 \oplus \mathcal{W}_3.$$

Since $\dim(\mathcal{W}_1 \oplus \mathcal{W}_2 \oplus \mathcal{W}_3) = 2m - p$, there exists a subspace $\mathcal{W}_4$ with $\dim \mathcal{W}_4 = p$ such that $\mathcal{W}_1 \oplus \mathcal{W}_2 \oplus \mathcal{W}_3 \oplus \mathcal{W}_4 = \mathbb{R}^{2m}$ is a complete decomposition of $\mathbb{R}^{2m}$.

We can change the projectors $P^\pm_0$ in such a way that (see [20, Proposition 2.3])

$$\mathcal{N}\left(P^+_0\right) = \mathcal{W}_3 \oplus \mathcal{W}_4, \quad \mathcal{R}\left(P^-_0\right) = \mathcal{W}_2 \oplus \mathcal{W}_4,$$

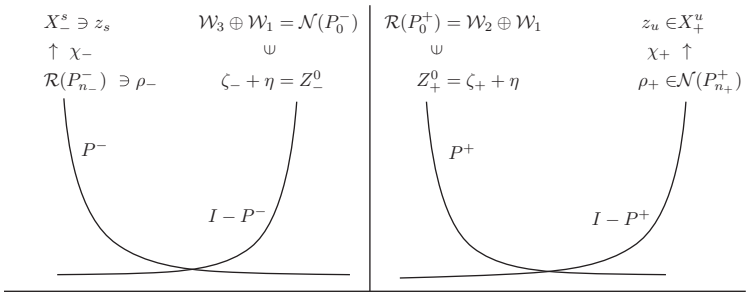


FIG. 3.1. *Overview over dichotomy estimates.*

increasing the dichotomy constants $K_\pm$ to $2K_\pm(1+K_\pm)$ and without changing $\mathcal{N}(P_0^-)$, $\mathcal{R}(P_0^+)$ or the dichotomy exponents $\alpha_\pm$. From the choice of Z_0^- and Z_0^+ follows

$$\begin{aligned} \left(I - P_0^-\right)\tilde{Z}_0^- = Z_0^- \quad &\in \left(\mathcal{N}\left(P_0^-\right)\right)^q = \mathcal{W}_3^q \oplus \mathcal{W}_1^q, \\ P_0^+ \tilde{Z}_0^+ = Z_0^+ \quad &\in \left(\mathcal{R}\left(P_0^+\right)\right)^q = \mathcal{W}_2^q \oplus \mathcal{W}_1^q \end{aligned}$$

(for an overview of the decomposition according to the different dichotomy data see Figure 3.1). We use the ansatz $Z_0^- = \zeta_- + \eta_-$, $Z_0^+ = \zeta_+ + \eta_+$, where $\zeta_- \in \mathcal{W}_3^q$, $\zeta_+ \in \mathcal{W}_2^q$, and $\eta_\pm \in \mathcal{W}_1$, and from (3.21a) we obtain $\eta_+ = \eta_- =: \eta$. Equation (3.21a) now reads

$$\zeta_- - \zeta_+ + \hat{\mathcal{S}}(0, n_-)\rho_- - \hat{\mathcal{S}}(0, n_+)\rho_+ + \left(\hat{s}_0^-\left(\hat{V}\right) - \hat{s}_0^+\left(\hat{V}\right)\right)\Upsilon = \hat{s}_0^+\left(R^{de}\right) - \hat{s}_0^-\left(R^{de}\right).$$

We transform the boundary values ρ_- and ρ_+ to coordinates (z_s, z_u) which are independent of J as follows: Denote by E_-^s the projector onto X_-^s along X_-^u and by E_+^u the projector onto X_+^u along X_+^s, where $X_\pm^s$ are the stable subspaces and $X_\pm^u$ are the unstable subspaces of $M_\pm$ (cf. Definition (2.3)).

We define the transformations

$$\chi_- : \mathcal{R}\left(P_{n_-}^-\right) \to X_-^s,\ \varrho_- \mapsto z_s, \qquad \chi_+ : \mathcal{N}\left(P_{n_+}^+\right) \to X_+^u,\ \varrho_+ \mapsto z_u$$

by

$$\chi_- = I + E_-^s - P_{n_-}^-, \qquad \chi_+ = I - E_+^s + P_{n_+}^+.$$

Now we apply $\chi_\pm$ columnwise. From the roughness theorem for exponential dichotomies [19] we have $\lim_{x\to\pm\infty} \pi^\pm(x) = E_\pm^s$. With $P_n^\pm = \pi^\pm(x_n)$ the invertibility of χ_- and χ_+ follows for $J_h \to \mathbb{R}$ together with the estimates

$$\left\|\chi_\pm^{-1}\right\| \le \frac{1}{1 - \left\|P_{n_\pm}^\pm - E_\pm^s\right\|} \le 2. \tag{3.22}$$

This implies for all $(z_s, z_u) \in X_-^s \times X_+^u$ that

$$\left\|(I - E_-^s)\chi_-^{-1} z_s\right\| \le \frac{\|P_{n_-}^- - E_-^s\|\|z_s\|}{1 - \|P_{n_-}^- - E_-^s\|}, \qquad \left\|E_+^s \chi_+^{-1} z_u\right\| \le \frac{\|P_{n_+}^+ - E_+^s\|\|z_u\|}{1 - \|P_{n_+}^+ - E_+^s\|}. \tag{3.23}$$

Define $\hat{\Phi} \in \ell_\infty(\mathbb{R}^{2m,p})$ by $\hat{\Phi}_n = \Phi(hn)$, where Φ has been defined in (3.12). Since the columns $\hat{\phi}_0^i$, $i = 1, \ldots, p$, of $\hat{\Phi}_0 = \Phi(0)$ span a basis of $\mathcal{N}(P_0^-) \cap \mathcal{R}(P_0^+)$ (see Remark 3.2), we can write $\eta \in \mathcal{W}_1^q$ as $\eta = \hat{\Phi}_0 \kappa$ for some $\kappa \in \mathbb{R}^{p,q}$.

Equation (3.21) can now be written as

$$T\begin{pmatrix}(\zeta_-,\zeta_+,\Upsilon)\\(z_s,z_u)\\\kappa\end{pmatrix}=\begin{pmatrix}\hat{s}_0^+(R^{de})-\hat{s}_0^-\left(R^{de}\right)\\R^{bc}-\left((P_+Q_+)\hat{s}_{n_+}^+\left(R^{de}\right)-(P_-Q_-)\hat{s}_{n_-}^-\left(R^{de}\right)\right)\\R^{pc}-\left\langle\hat{\Psi}_{|_{J^-}},\hat{s}^-\left(R^{de}\right)\right\rangle_{J_h}-\left\langle\hat{\Psi}_{|_{J^+}},\hat{s}^+\left(R^{de}\right)\right\rangle_{J_h}\end{pmatrix},\tag{3.24}$$

where $T:\ (\mathcal{W}_2^q\times\mathcal{W}_3^q\times\mathbb{R}^{p,q})\times(X_-^s)^q\times(X_+^u)^q\times\mathbb{R}^{p,q}\to(\mathcal{W}_2\oplus\mathcal{W}_3\oplus\mathcal{W}_4)^q\times\mathbb{R}^{2m,q}\times\mathbb{R}^{p,q}$ has the following structure:

$$T=\begin{pmatrix}\mathcal{X}&\sigma&0\\\Delta&\mathcal{Y}&\varrho\\\Theta&\Pi&\mathcal{Z}\end{pmatrix},$$

where

$$\begin{aligned}
\mathcal{X}(\zeta_-,\zeta_+,\lambda)&=\zeta_--\zeta_++\left(\hat{s}_0^-\left(\hat{V}\right)-\hat{s}_0^+\left(\hat{V}\right)\right)\Upsilon,\\
\sigma(z_s,z_u)&=\hat{\mathcal{S}}(0,n_-)\left[\chi_-^{-1}z_s^1,\dots,\chi_-^{-1}z_s^q\right]-\hat{\mathcal{S}}(0,n_+)\left[\chi_+^{-1}z_u^1,\dots,\chi_+^{-1}z_u^q\right],\\
\Delta(\zeta_-,\zeta_+,\Upsilon)&=(P_-Q_-)\left(\hat{\mathcal{S}}(n_-,0)\zeta_-+\hat{s}_{n_-}^-\left(\hat{V}\right)\Upsilon\right)\\
&\quad+(P_+Q_+)\left(\hat{\mathcal{S}}(n_+,0)\zeta_++\hat{s}_{n_+}^+\left(\hat{V}\right)\Upsilon\right),\\
\mathcal{Y}(z_s,z_u)&=(P_-Q_-)\left[\chi_-^{-1}z_s^1,\dots,\chi_-^{-1}z_s^q\right]+(P_+Q_+)\left[\chi_+^{-1}z_u^1,\dots,\chi_+^{-1}z_u^q\right],\\
\varrho(\kappa)&=\left((P_-Q_-)\hat{\mathcal{S}}(n_-,0)+(P_+Q_+)\hat{\mathcal{S}}(n_+,0)\right)\hat{\Phi}_0\kappa,\\
\Theta(\zeta_-,\zeta_+,\Upsilon)&=\left\langle\hat{\Psi}_{|_{J^-}},\hat{\mathcal{S}}(\cdot,0)\zeta_-+\hat{s}^-\left(\hat{V}\right)\Upsilon\right\rangle_{J_h}+\left\langle\hat{\Psi}_{|_{J^+}},\hat{\mathcal{S}}(\cdot,0)\zeta_++\hat{s}^+\left(\hat{V}\right)\Upsilon\right\rangle_{J_h},\\
\Pi(z_s,z_u)&=\left\langle\hat{\Psi}_{|_{J^-}},\hat{\mathcal{S}}(\cdot,n_-)\chi_-^{-1}\left[z_s^1,\dots,z_s^q\right]\right\rangle_{J_h}\\
&\quad+\left\langle\hat{\Psi}_{|_{J^+}},\hat{\mathcal{S}}(\cdot,n_+)\chi_+^{-1}\left[z_u^1,\dots,z_u^q\right]\right\rangle_{J_h},\\
\mathcal{Z}(\kappa)&=\left\langle\hat{\Psi},\hat{\Phi}\right\rangle_{J_h}\kappa.
\end{aligned}$$

We have to show the invertibility of T as well as an estimate of the inverse of T. The terms σ and ϱ can be estimated by using (3.22), the exponential dichotomy of $\hat{L}$ and $\mathcal{R}(\hat{\Phi}_0)=\mathcal{R}(P_0^+)\cap\mathcal{N}(P_0^-)$ by

$$\|\sigma\|,\|\varrho\|\le\text{const}\,\mathrm{e}^{-\alpha h\min(-n_-,n_+)}\quad\to 0\quad\text{as } J_h\to\mathbb{R}.$$

The boundedness of the operators Δ,Π, and Θ follows from the dichotomy estimates (3.9) and the properties of $\hat{V}$. The properties of $\hat{\Psi}$ in Hypothesis 3.7 ensure that the map $(\mathcal{L}_{2,h}(\mathbb{R}^{2m}))^q\to\mathbb{R}^{p,q},\ Z\mapsto\langle\hat{\Psi},Z\rangle_{J_h}$ is uniformly bounded.

Hypothesis 2.4 and (3.23) imply that $\mathcal{Y}$ has a uniformly bounded inverse. Lemma 2.2 and the properties of $\hat{V}$ imply that the other operators on the diagonal $\mathcal{X}$ and $\mathcal{Z}$ converge for $J_h\to\mathbb{R}$ to asymptotic operators $\bar{\mathcal{X}}$ and $\bar{\mathcal{Z}}$. These are defined as follows: Let $\mathcal{S}$ be the solution operator corresponding to L, and $\pi^\pm$ the stable and unstable dichotomy projectors. We define

$$\bar{\mathcal{X}}(\zeta_-,\zeta_+,\Upsilon)=\zeta_--\zeta_++(s^-(V)(0)-s^+(V)(0))\Upsilon,$$

where

$$s^-(V)(x) = \int_{-\infty}^{x} \mathcal{S}(x,\xi)\pi^-(\xi)V(\xi)\,d\xi - \int_{x}^{0} \mathcal{S}(x,\xi)(I-\pi^-(\xi))V(\xi)\,d\xi \quad \text{for } x \le 0,$$
$$s^+(V)(x) = \int_{0}^{x} \mathcal{S}(x,\xi)\pi^+(\xi)V(\xi)\,d\xi - \int_{x}^{\infty} \mathcal{S}(x,\xi)(I-\pi^+(\xi))V(\xi)\,d\xi \quad \text{for } x \ge 0,$$

and

$$\bar{\mathcal{Z}}(\kappa) = \langle \Psi, \Phi\rangle \kappa.$$

The invertibility of $\bar{\mathcal{Z}}$ is ensured by Hypothesis 3.7. Similarly the invertibility of $\bar{\mathcal{X}}$ follows from the nondegeneracy condition in Hypothesis 3.7 similar to [7], by multiplying the equation

$$0 = \zeta_- - \zeta_+ + (s^-(V)(0) - s^+(V)(0))\Upsilon$$

from the left by $\Xi(0)^T$ and using Remark 3.8.

Then we obtain that $\mathcal{X}$ and $\mathcal{Z}$ are invertible for h small enough and $\mp h n_\pm$ large enough with a uniform bound for the inverse.

By summing up the estimates for the right-hand side in (3.24) and using estimate (3.16) for $\hat{s}^\pm$ in Lemma 3.10 as well as the properties of $\hat{\Psi}$, one obtains for $J_h \to \mathbb{R}$

$$\|\zeta_-\| + \|\zeta_+\| + \|\Upsilon\| + \|z_s\| + \|z_u\| + \|\kappa\| \le \text{const}\left(\frac{1}{h}\|R^{de}\|_{\mathcal{L}_{2,h}} + \|R^{bc}\| + \|R^{pc}\|\right).$$

This implies with Corollary 3.11 and (3.22) for $\tilde{Z}^- \in (\ell_\infty^{J^-})^q$ and $\tilde{Z}^+ \in (\ell_\infty^{J^+})^q$ that

$$\|\tilde{Z}^\pm\|_{\mathcal{L}_{2,h}} \le \text{const}\left(\frac{1}{h}\|R^{de}\|_{\mathcal{L}_{2,h}} + \|R^{bc}\| + \|R^{pc}\|\right).$$

A similar estimate for the contribution of $\tilde{Z}_{n_++1}$ at $n_+ + 1$ leads to

$$\left\|\tilde{Z}\right\|_{\mathcal{L}_{2,h}} + \|\Upsilon\| \le \text{const}\left(\frac{1}{h}\|R^{de}\|_{\mathcal{L}_{2,h}} + \|R^{bc}\| + \|R^{pc}\|\right).$$

This can be improved for h small enough to the $\|\cdot\|_{\mathcal{H}_h^1}$ estimate (3.14) by using the difference equation (3.10a) and the following properties of $\hat{V}$ and $\hat{M}$:

$$\hat{V}_n = hV(x_n) + \mathcal{O}(h^2), \quad \hat{M}_n = I + \mathcal{O}(h),$$

which follow from (3.13) and (3.11). ∎

4. Proof of the main results. In the following section we use the following fixed point theorem (see [27] or [7, Lemma 3.1]) to prove the main approximation results.

THEOREM 4.1. *Let $F : \mathcal{Y} \supset B_\varrho(\bar{y}) \to \mathcal{Z}$ be a C^1 mapping between two Banach spaces $\mathcal{Y}$ and $\mathcal{Z}$, and let $(DF(\bar{y}))^{-1} \in L[\mathcal{Z},\mathcal{Y}]$ exist. Assume the following estimates for $\kappa, \sigma > 0$:*

$$\|DF(y) - DF(\bar{y})\|_{\mathcal{Y}\to\mathcal{Z}} \le \kappa < \sigma \le \frac{1}{\|DF(\bar{y})^{-1}\|_{\mathcal{Z}\to\mathcal{Y}}} \quad \forall y \in B_\varrho(\bar{y}), \tag{4.1}$$

$$\|F(\bar{y})\|_Z \le (\sigma - \kappa)\varrho. \tag{4.2}$$

Then F has a unique zero y_0 in $B_\varrho(\bar y) = \{y : \|y-\bar y\|_{\mathcal Y} \le \varrho\}$, and the following estimates hold:

$$\|y_0-\bar y\|_{\mathcal Y} \le \frac{1}{(\sigma-\kappa)}\|F(\bar y)\|_{\mathcal Z}, \tag{4.3}$$

$$\|y_1-y_2\|_{\mathcal Y} \le \frac{1}{(\sigma-\kappa)}\|F(y_1)-F(y_2)\|_{\mathcal Z} \quad \forall y_1,y_2 \in B_\varrho(\bar y). \tag{4.4}$$

The linear Theorem 3.9 is used to derive estimates of the inverse of the linearization $DF(\bar y)^{-1}$.

4.1. Approximation of relative equilibria.

Proof of Theorem 2.5. Set $w = v-\bar v$ and $\lambda = \mu - \bar\mu$. Then a solution of (2.2) is given by $(\bar w+\bar v, \bar\lambda+\bar\mu)$, where $(\bar w,\bar\lambda)$ is a zero of the operator $F : \ell^{\hat J}_\infty \times \mathbb{R}^p \to \ell^J_\infty \times \mathbb{R}^{2m}\times\mathbb{R}^p$, given by

$$F(w,\lambda) = \begin{pmatrix} \left(\left(\bar\Lambda w\right)_n + \hat S_n\left(\bar v_{|_{J_h}}\right)\lambda + g_n(w,\lambda) + \left(R^{de}_{J_h}\right)_n\right)_{n\in J} \\ \mathcal B w + R^{bc}_{J_h} \\ \left\langle \hat S\left(\hat v_{|_{J_h}}\right), w\right\rangle_{J_h} + R^{pc}_{J_h} \end{pmatrix},$$

where

$$\begin{aligned} \left(\bar\Lambda w\right)_n &= A(\delta_+\delta_- w)_n + B_n(\delta_0 w)_n + C_n w_n, \qquad B_n = B(x_n),\ C_n = C(x_n),\\ g_n(w,\lambda) &= f(\bar v_n + w_n, \delta_0(\bar v+w)_n) - f(\bar v_n,\delta_0\bar v_n)\\ &\quad - D_1 f(\bar v_n,\bar v_n')w_n - D_2 f(\bar v_n,\bar v_n')\delta_0 w_n + \hat S_n(w)\lambda, \quad n\in J, \end{aligned}$$

and

$$\begin{aligned} R^{de}_{J_h} &= A\delta_+\delta_-\bar v + \hat S(\bar v)\bar\mu + f(\bar v,\delta_0\bar v), \qquad R^{bc}_{J_h} = \mathcal B\bar v_{|_{J_h}} - (P_-\bar v_- + P_+\bar v_+),\\ R^{pc}_{J_h} &= \left\langle \hat S(\hat v_{|_{J_h}}), (\bar v-\hat v)_{|_{J_h}}\right\rangle_{J_h}. \end{aligned}$$

We will apply Theorem 4.1 to F with $\mathcal Y = (\ell^{\hat J}_\infty)\times\mathbb{R}^p$ and $\mathcal Z = (\ell^J_\infty)\times\mathbb{R}^{2m}\times\mathbb{R}^p$, at the approximative zero $\bar y = 0$ with norms

$$\|(v,\mu)\|_{\mathcal Y} = \|v\|_{\mathcal H^2_h} + \|\mu\|, \qquad \|(R^{de},R^{bc},R^{pc})\|_{\mathcal Z} = \|R^{de}\|_{\mathcal L_{2,h}} + \|R^{bc}\| + \|R^{pc}\|.$$

We have with the properties of $\bar v$ (see Remark 3.4)

$$\begin{aligned} \|R^{de}_{J_h}\|_{\mathcal L_{2,h}} &\le \|A\|\,\|\delta_+\delta_-\bar v - \bar v''\|_{\mathcal L_{2,h}} + \|f(\bar v,\delta_0\bar v) - f(\bar v,\bar v')\|_{\mathcal L_{2,h}}\\ &\qquad + \|\hat S(\bar v_{|_{J_h}}) - S(\bar v)\|_{\mathcal L_{2,h}}\|\bar\mu\|\\ &\le \operatorname{const} h^2. \end{aligned}$$

Similarly, we obtain with the assumptions on the phase condition in Hypothesis 2.4 and Lemma 2.2

$$\begin{aligned} \|R^{pc}_{J_h}\| &\le \left\|\left\langle S(\hat v)_{|_{J_h}}, (\bar v-\hat v)_{|_{J_h}}\right\rangle_{J_h}\right\| + \left\|\left\langle \hat S(\hat v_{|_{J_h}}) - S(\hat v)_{|_{J_h}}, (\bar v-\hat v)_{|_{J_h}}\right\rangle_{J_h}\right\|\\ &\le \operatorname{const}\left(h^2 + \mathrm{e}^{-\alpha h\min\{n_-,n_+\}}\right). \end{aligned}$$

By using the consistency of the boundary conditions in Hypothesis 2.4 we have

$$\begin{aligned}\|R^{bc}_{J_h}\| &\le \|P_-\|\|\bar v_{n_-} - \bar v_-\| + \|Q_-\|(\|\delta_0 \bar v_{n_-} - \bar v'(x_{n_-})\| + \|\bar v'(x_{n_-})\|)\\ &\quad + \|P_+\|\|\bar v_{n_+} - \bar v_+\| + \|Q_+\|(\|\delta_0 \bar v_{n_+} - \bar v'(x_{n_+})\| + \|\bar v'(x_{n_+})\|)\\ &\le \text{const}\left(h^2 + \mathrm{e}^{-\alpha h \min\{n_-, n_+\}}\right).\end{aligned}$$

These estimates together with $g(0,0) = 0$ imply the consistency estimate

$$\|F(0,0)\|_Z \le \|R^{de}\|_{\mathcal{L}_{2,h}} + \|R^{bc}\| + \|R^{pc}\| \le \text{const}\left(h^2 + \mathrm{e}^{-\alpha h \min\{-n_-, n_+\}}\right).$$

With

$$DF\left(\tilde w, \tilde\lambda\right)(w,\lambda) = \begin{pmatrix} (\bar\Lambda w)_n + \hat S_n\left(\bar v_{|_{J_h}}\right)\lambda + Dg_n\left(\tilde w, \tilde\lambda\right)(w,\lambda) \\ \mathcal{B}w \\ \left\langle \hat S\left(\hat v_{|_{J_h}}\right), w\right\rangle_{J_h}\end{pmatrix},$$

$Dg(0,0) = 0$, the properties of f (see Remark 1.2), and the discrete Sobolev embedding $\|v\|_\infty \le \text{const}\,\|v\|_{\mathcal{H}^1_h}$, we arrive at

$$\begin{aligned}\left\|\left(DF\left(\tilde w, \tilde\lambda\right) - DF(0,0)\right)(w,\lambda)\right\|_{\mathcal{Z}} &= \left\|Dg\left(\tilde w, \tilde\lambda\right)(w,\lambda)\right\|_{\mathcal{L}_{2,h}}\\ &\le \text{const}\left(\|\tilde w\|_{\mathcal{H}^2_h} + \|\tilde\lambda\|\right)\left(\|w\|_{\mathcal{H}^2_h} + \|\lambda\|\right).\end{aligned}$$

It remains to estimate the inverse of $DF(0,0)$. With the definitions

$$z_n = (v_n, w_n), \qquad w_n = (\delta_- v)_n, \qquad \text{and} \qquad \hat R^{de}_n = \begin{pmatrix} 0 \\ hR^{de}_n\end{pmatrix}, \tag{4.5}$$

the equation

$$DF(0,0)(w,\lambda) = (R^{de}, R^{bc}, R^{pc})^T \tag{4.6}$$

is transformed into

$$\tilde T(z,\lambda) = \left(\hat R^{de}, R^{bc}, R^{pc}\right)^T, \tag{4.7}$$

where $\tilde T : \ell^{[n_-, n_+ + 1]}_\infty(\mathbb{R}^{2m}) \times \mathbb{R}^p \to \ell^J_\infty(\mathbb{R}^{2m}) \times \mathbb{R}^{2m} \times \mathbb{R}^p$ is given by

(4.8)
$$\tilde T(z,\lambda) = \begin{pmatrix} \left(\tilde N_n z_{n+1} - \tilde K_n z_n - \tilde U_n \lambda\right)_{n\in J} \\ P_- v_{n_-} + Q_-\left(w_{n_-} + \frac{h}{2}\delta_+ w_{n_-}\right) + P_+ v_{n_+} + Q_+\left(w_{n_+} + \frac{h}{2}\delta_+ w_{n_+}\right) \\ \left\langle \tilde\Psi, z\right\rangle_{J_h}\end{pmatrix},$$

with

$$\tilde N_n = \begin{pmatrix} I & -hI \\ 0 & A + \frac{h}{2}B_n\end{pmatrix}, \quad \tilde K_n = \begin{pmatrix} I & 0 \\ -hC_n & A - \frac{h}{2}B_n\end{pmatrix}, \quad \tilde U_n = -\begin{pmatrix} 0 \\ h\hat S\left(\bar v_{|_{J_h}}\right)_n\end{pmatrix},$$

and

$$\tilde{\Psi} = \begin{pmatrix} \hat{S}\left(\hat{v}_{|_{J_h}}\right) \\ 0 \end{pmatrix}.$$

In order to use the properties of the continuous operator L, we consider the following approximation of $\tilde{T}$:

$$T(z,\lambda) = \begin{pmatrix} \left(N z_{n+1} - N\hat{M}_n z_n - U_n \lambda\right)_{n\in J} \\ (P_- \; Q_-) z_{n_-} + (P_+ \; Q_+) z_{n_+} \\ \left\langle \Psi_{|_{J_h}}, z \right\rangle_{J_h} \end{pmatrix},$$

where

$$N = \begin{pmatrix} I & 0 \\ 0 & A \end{pmatrix}, \qquad U_n = -\begin{pmatrix} 0 \\ h[S(\bar{v})](x_n) \end{pmatrix}, \qquad \Psi(x) = \begin{pmatrix} [S(\hat{v})](x) \\ 0 \end{pmatrix}. \tag{4.9}$$

From the definition of $\hat{M}_n$ we get

$$\hat{M} = \mathcal{S}(x_{n+1}, x_n) = I + hM(x_n) + \mathcal{O}(h^2)$$

as well as the estimates

$$\begin{aligned} \|\tilde{N}_n - N\| &\le \text{const}\, h, \\ \|N\hat{M}_n - \tilde{K}_n + \tilde{N}_n - N\| &\le \text{const}\, h^2, \\ \|\tilde{U} - U\|_{\mathcal{L}_{2,h}} &\le \text{const}\, h^2. \end{aligned}$$

These estimates imply that T is a perturbation of $\tilde{T}$ of order h w.r.t. the norms $\|(\hat{R}^{de}, R^{bc}, R^{pc})\|_h^* = \frac{1}{h}\|\hat{R}^{de}\|_{\mathcal{L}_{2,h}} + \|R^{bc}\| + \|R^{pc}\|$ and $\|(z,\lambda)\|_{\mathcal{H}_h^1}^* = \|z\|_{\mathcal{H}_h^1} + \|\lambda\|$ as follows. We have

$$\begin{aligned} &\left\| \left(\left(\tilde{N}_n - N\right) z_{n+1} - \left(\tilde{K}_n - N\hat{M}_n\right) z_n + \left(\tilde{U}_n - U_n\right)\lambda \right)_{n\in J} \right\|_{\mathcal{L}_{2,h}}^2 \\ &\quad \le \text{const} \sum_{n=n_-}^{n_+} h \left(h^2 \|\tilde{N}_n - N\|^2 \|\delta_+ z_n\|^2 + \|N\hat{M}_n - \tilde{K}_n + \tilde{N}_n - N\|^2 \|z_n\|^2 \right. \\ &\qquad\qquad \left. + \|\tilde{U}_n - U_n\|^2 \|\lambda\|^2 \right) \\ &\quad \le \text{const}\, h^4 \|\delta_+ z\|_{\mathcal{L}_{2,h}}^2 + h^4 \|z\|_{\mathcal{L}_{2,h}}^2 + \|\tilde{U} - U\|_{\mathcal{L}_{2,h}}^2 \|\lambda\|^2 \le \text{const}\, h^4 (\|z\|_{\mathcal{H}_h^1} + \|\lambda\|)^2 \end{aligned}$$

and with Lemma 2.2

$$\|\Psi_{|_{J_h}} - \tilde{\Psi}\|_{J_h} \le \left\| \hat{S}\left(\hat{v}_{|_{J_h}}\right) - S\left(\hat{v}\right)_{|_{J_h}} \right\|_{J_h} \le \text{const} \left(h^2 + \mathrm{e}^{-\alpha h \min\{-n_-, n_+}} \right).$$

Thus we obtain

$$\begin{aligned}\left\|\left(\tilde{T}-T\right)(z,\lambda)\right\|_h^* &= \frac{1}{h}\left\|\left(\left(\tilde{N}_n-N\right)z_{n+1}-\left(\tilde{K}_n-N\hat{M}_n\right)z_n+\left(\tilde{U}_n-U_n\right)\lambda\right)_{n\in J}\right\|_{\mathcal{L}_{2,h}}\\ &\quad+\left\|Q_-\left(\frac{h}{2}\delta_+ w_{n_-}\right)+Q_+\left(\frac{h}{2}\delta_+ w_{n_+}\right)\right\|+\left\|\left\langle\Psi_{|_{J_h}}-\tilde{\Psi},z\right\rangle_{J_h}\right\|\\ &\le \text{const}\,h(\|z\|_{\mathcal{H}_h^1}+\|\lambda\|)\\ &\quad+\frac{h}{2}(\|Q_-\|\|\delta_+ w_{n_-}\|+\|Q_+\|\|\delta_+ w_{n_+}\|)+\text{const}\,h^2\|z\|_{\mathcal{L}_{2,h}}\\ &\le \text{const}\left(h+\mathrm{e}^{-\alpha h\min\{-n_-,n_+\}}\right)\left(\|z\|_{\mathcal{H}_h^1}+\|\lambda\|\right).\end{aligned}$$

Define

$$V(x)=\begin{pmatrix}0\\ -A^{-1}[S(\bar{v})](x)\end{pmatrix}, \tag{4.10}$$

and set $\hat{V}_n=N^{-1}U_n=hV(x_n)$. Then the equation

$$T(z,\lambda)=\left(\hat{R}^{de},R^{bc},R^{pc}\right)^T \tag{4.11}$$

can be equivalently written as

$$\begin{aligned}z_{n+1}-\hat{M}_n z_n-\hat{V}_n\lambda &= N^{-1}\hat{R}_n^{de},\\ (P_-\ Q_-)z_{n_-}+(P_+\ Q_+)z_{n_+} &= R^{bc},\\ \left\langle\Psi_{|_{J_h}},z\right\rangle_{J_h} &= R^{pc}.\end{aligned} \tag{4.12}$$

In order to apply the linear Theorem 3.9 to (4.12), we show that Hypothesis 3.7 is satisfied. The nondegeneracy condition in Hypothesis 3.7 follows from the fact that the null space of Λ is spanned by $S^i(\bar{v})$, $i=1,\dots,p$, and (4.10).

The assumptions on the phase condition in Hypothesis 3.7 follow with the definition of Ψ in (4.9) directly from the corresponding assumptions in Hypothesis 2.4.

By applying Theorem 3.9 to (4.12) and multiplying with the bounded matrix N, we obtain the invertibility of T as well as an uniform bound for the inverse

$$\left\|T^{-1}\left(R^{de},R^{bc},R^{pc}\right)\right\|_{\mathcal{H}_h^1}^*\le\text{const}\left(\frac{1}{h}\|\hat{R}^{de}\|_{\mathcal{L}_{2,h}}+\|R^{bc}\|+\|R^{pc}\|\right).$$

By perturbation arguments the invertibility of $\tilde{T}$ follows with the same bound for a possibly different constant. These estimates can be transformed back into (2.10) by using the structure of the matrices $\tilde{N}_n,\tilde{K}_n$, and $\tilde{U}_n$ (see (4.9)) as well as the structure of the right-hand side $\hat{R}^{de}$ (see (4.5)) together with the estimate $\|z\|_{\mathcal{H}_h^1}\le\text{const}\,\|v\|_{\mathcal{H}_h^2}$. ∎

4.2. Approximation of the spectrum. The proof of Theorem 2.6 follows the same lines as the proof of Theorem 2.5.

Proof of Theorem 2.6. As in the previous proof we insert $X = V - S(\bar{v})_{|_{J_h}}$ into (2.4) and obtain $F(X,\Sigma) = 0$, where

$$F(X,\Sigma) = \begin{pmatrix} \left(\left(\tilde{\Lambda}X\right)_n - [S(\bar{v})](x_n)\Sigma - X_n\Sigma + \left(R^{de}_{J_h}\right)_n\right)_{n\in J} \\ \mathcal{B}X + R^{bc}_{J_h} \\ \left\langle \hat{X}_{|_{J_h}}, X\right\rangle_{J_h} + R^{pc}_{J_h} \end{pmatrix},$$

with $R^{de}_{J_h} = \tilde{\Lambda}S(\bar{v})_{|_{J_h}}$, $R^{bc}_{J_h} = \mathcal{B}S(\bar{v})_{|_{J_h}}$, and $R^{pc}_{J_h} = \langle \hat{X}_{|_{J_h}}, S(\bar{v})_{|_{J_h}}\rangle_{J_h} - I$.

We will apply Theorem 4.1 to F with $\mathcal{Y} = (\ell^{\hat{J}}_\infty)^p \times \mathbb{R}^{p,p}$ and $\mathcal{Z} = (\ell^J_\infty)^p \times \mathbb{R}^{2m,p} \times \mathbb{R}^{p,p}$, at the approximative zero $\bar{y} = 0$ with norms

$$\|(X,\Sigma)\|_{\mathcal{Y}} = \|X\|_{\mathcal{H}^2_h} + \|\Sigma\|, \qquad \|(R^{de},R^{bc},R^{pc})\|_{\mathcal{Z}} = \|R^{de}\|_{\mathcal{L}_{2,h}} + \|R^{bc}\| + \|R^{pc}\|.$$

In the same way as in the proof of Theorem 2.5 we have

$$\left\|R^{bc}_{J_h}\right\| + \left\|R^{pc}_{J_h}\right\| \le \text{const}\left(h^2 + \mathrm{e}^{-\alpha h \min\{-n_-,n_+\}}\right)$$

and

$$\begin{aligned}\left\|R^{de}_{J_h}\right\|_{\mathcal{L}_{2,h}} &\le \left\|\tilde{\Lambda}S(\bar{v})_{|_{J_h}} - (\Lambda S(\bar{v}))_{|_{J_h}}\right\|_{\mathcal{L}_{2,h}} + \left\|(\Lambda S(\bar{v}))_{|_{J_h}}\right\|_{\mathcal{L}_{2,h}} \\ &\le \text{const}\left(h^2 + \mathrm{e}^{-\alpha h \min\{-n_-,n_+\}}\right).\end{aligned}$$

The properties of $S(\bar{v})$ (see Remark 3.4) imply that $P_-S(\bar{v}_-) + P_+S(\bar{v}_+) = 0$ as well as $\lim_{x\to\infty}[S(\bar{v}')](x) = 0$. These estimates lead to

$$\begin{aligned}\|F(0,0)\|_Z &\le \left\|\tilde{\Lambda}S(\bar{v})_{|_{J_h}}\right\|_{\mathcal{L}_{2,h}} + \left\|\mathcal{B}S(\bar{v})_{|_{J_h}}\right\| + \left\|\left\langle \hat{X}_{|_{J_h}}, S(\bar{v})_{|_{J_h}}\right\rangle_{J_h} - I\right\| \\ &\le \text{const}\left(h^2 + \mathrm{e}^{-\alpha h \min\{-n_-,n_+\}}\right).\end{aligned}$$

With

$$DF\left(\tilde{X},\tilde{\Sigma}\right)(X,\Sigma) = \begin{pmatrix} \left(\left(\tilde{\Lambda}X\right)_n - [S(\bar{v})](x_n)\Sigma - \tilde{X}_n\Sigma - X_n\tilde{\Sigma}\right)_{n\in J} \\ \mathcal{B}X \\ \left\langle \hat{X}_{|_{J_h}}, X\right\rangle_{J_h} \end{pmatrix},$$

we obtain

$$\begin{aligned}\left\|\left(DF\left(\tilde{X},\tilde{\Sigma}\right) - DF(0,0)\right)(X,\Sigma)\right\|_{\mathcal{Z}} &= \left\|\left(\tilde{X}_n\Sigma + X_n\tilde{\Sigma}\right)_{n\in J}\right\|_{\mathcal{L}_{2,h}} \\ &\le \text{const}\left(\left\|\tilde{X}\right\|_{\mathcal{H}^2_h} + \left\|\tilde{\Sigma}\right\|\right)\left(\|X\|_{\mathcal{H}^2_h} + \|\Sigma\|\right).\end{aligned}$$

It remains to estimate the inverse of $DF(0,0)$. In the same way as before, we transform (4.6) into (4.7), where $\tilde{T} : (\ell^{[n_-,n_++1]}_\infty(\mathbb{R}^{2m}))^p \times \mathbb{R}^{p,p} \to (\ell^J_\infty(\mathbb{R}^{2m}))^p \times \mathbb{R}^{2m,p} \times \mathbb{R}^{p,p}$ is given by (4.8) with

$$\tilde{N}_n = \begin{pmatrix} I & -hI \\ 0 & A + \frac{h}{2}\tilde{B}_n \end{pmatrix}, \quad \tilde{K}_n = \begin{pmatrix} I & 0 \\ -h\tilde{C}_n & A - \frac{h}{2}\tilde{B}_n \end{pmatrix}, \quad \tilde{U}_n = -\begin{pmatrix} 0 \\ h[S(\bar{v})](x_n) \end{pmatrix},$$

$$\tilde{\Psi} = \begin{pmatrix} \hat{X} \\ 0 \end{pmatrix}.$$

Again we solve (4.11) in order to solve (4.7), where $U_n = \tilde{U}_n$ and $\Psi = \tilde{\Psi}$. With the properties of f (see (1.6)) and (2.10) we obtain

$$\begin{aligned}\left\|\tilde{B}_n - B(x_n)\right\| &\le \|D_2 f(\tilde{v}_n, (\delta_0 \tilde{v})_n) - D_2 f(\bar{v}_n, \bar{v}'_n)\| + \sum_{i=1}^{p} |\tilde{\mu}_i - \bar{\mu}_i| \|S_1^i\| \\ &\le \text{const}\left(h^2 + \mathrm{e}^{-\alpha h \min\{n_-, n_+\}}\right),\end{aligned}$$

and similarly

$$\left\|\tilde{C}_n - C(x_n)\right\| \le \text{const}\left(h^2 + \mathrm{e}^{-\alpha h \min\{-n_-, n_+\}}\right).$$

This implies that T is a perturbation of $\tilde{T}$ of order h. As in the previous proof, the assumptions on the phase condition in Hypothesis 3.7 follow with the definition of Ψ directly from the assumptions on $\hat{X}$. The rest of the arguments is the same as in the proof of Theorem 2.5. ∎

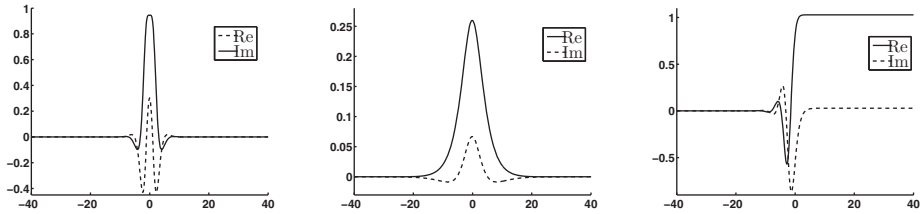
5. Numerical results. We illustrate the approximation results on the cubic quintic Ginzburg–Landau equation, a reaction-diffusion equation for which exact solutions are known. We compare the order of approximation for different grid sizes h and intervals J with the expected behavior from Theorems 2.5 and 2.6. Similar experiments have been conducted for traveling wave solutions of the Nagumo equation in [25]. The cubic quintic Ginzburg–Landau equation [24], [28], [9] reads

$$u_t = a u_{xx} + \delta u + g(u), \quad g(u) = \beta |u|^2 u + \gamma |u|^4 u, \ \delta \in \mathbb{R}, \ a, \beta, \gamma \in \mathbb{C}. \tag{5.1}$$

It shows a variety of coherent structures, such as stable pulse solutions, fronts, sources, sinks, etc.

For numerical computations we write (5.1) in real variables; then (5.1) has the equivariance properties given in Example 1.4. For the parameter set $a = 1$, $\delta = -0.1$, $\beta = 3 + i$, $\gamma = -2.75 + i$, which has been used in [24], this equation possesses a stable rotating pulse (Figure 5.1(a)), an unstable rotating pulse [24] (Figure 5.1(b)), and a rotating and traveling front (Figure 5.1(c)). All of these solutions can (in real notation) be written in the form (1.13) where $\bar{\mu}_\tau = 0$ in the case of the rotating pulse. Thus these solutions are relative equilibria in the sense of Definition 1.1 (cf. Example 1.4). Depending on the choice of initial conditions, a different type of solution is selected.

To solve the boundary value problem (2.2) we use Newton's method with initial data that in the stable cases 5.1(a) and 5.1(c) has been generated by using the method



(a) stable pulse (b) unstable pulse (c) front

FIG. 5.1. *Cubic quintic Ginzburg–Landau system: stable/unstable pulse and stable front.*

of freezing described in [5], [25], [6]. For the unstable case 5.1(b) there is an explicit formula that has been derived in [15] by using Painlevé methods, which reads (in complex notation)

$$u(x,t) = \mathrm{e}^{-i\bar{\mu}_\rho t}\ \bar{v}(x), \tag{5.2}$$

where

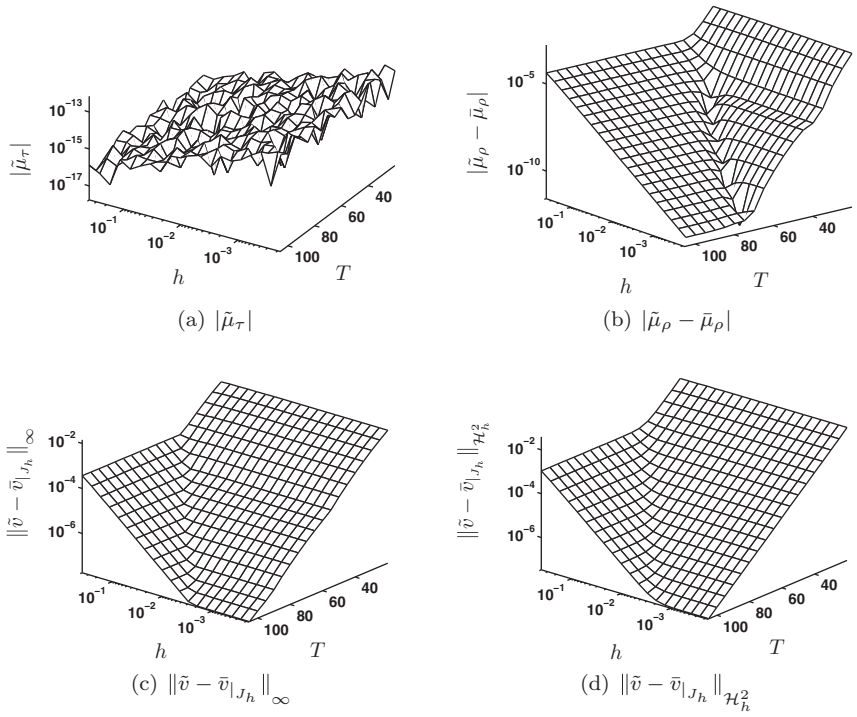
$$\bar{v}(\xi) = u_0 \mathrm{e}^{ia_0\theta_0\xi}(\cosh(k\xi) - \cosh(\rho))^{ia_0}\sqrt{\frac{k\sinh(\rho)}{\cosh(k\xi)+\cosh(\rho)}}\ .$$

The parameters $u_0, a_0, \theta_0, \rho, \bar{\mu}_\rho$, and k can be computed explicitly from a, δ, β, and γ by the formulas given in [15].

Figure 5.1 shows the numerical solutions on the interval $[-40, 40]$ with mesh size $h = 0.1$ using Dirichlet and Neumann boundary conditions.

In case 5.1(b) we can illustrate the results of Theorems 2.5 and 2.6 in sections 5.1 and 5.2. While the numerics seems to work in every case, Theorems 2.5 and 2.6 apply to cases 5.1(a) and 5.1(b) but not to 5.1(c). To some surprise the method seems to work as well, but it is hard to derive a numerical test to check the validity of the estimates of Theorems 2.5 and 2.6 in this case.

5.1. Approximation of the unstable pulse. We compare the approximation error of the solution of the boundary value problem (2.2) with the estimates (2.10) in Theorem 2.5. For the unstable pulse the exact solution is explicitly given by (5.2). Figure 5.2 shows the approximation error of the pulse for Dirichlet boundary



(a) $|\tilde{\mu}_\tau|$ (b) $|\tilde{\mu}_\rho - \bar{\mu}_\rho|$

(c) $\|\tilde{v} - \bar{v}_{|J_h}\|_\infty$ (d) $\|\tilde{v} - \bar{v}_{|J_h}\|_{\mathcal{H}_h^2}$

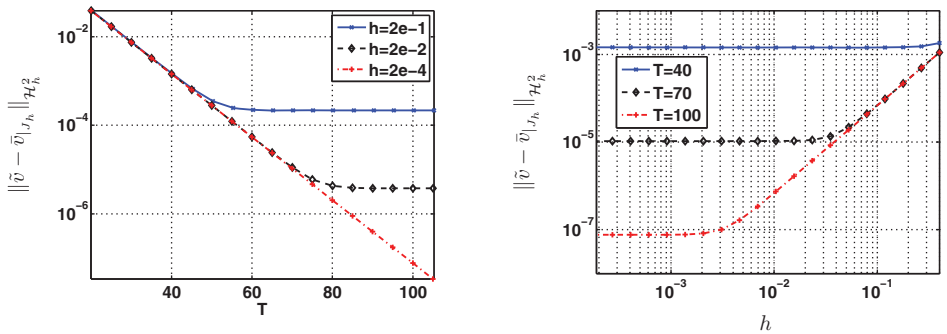
FIG. 5.2. *Approximation error for the unstable pulse.*

TABLE 5.1

Errors $\epsilon_u = \|\tilde{v} - \bar{v}_{|J_h}\|_{\mathcal{H}_h^2}$ *and* $\epsilon_{\mu_\rho} = |\tilde{\mu}_\rho - \bar{\mu}_\rho|$ *for fixed* $T = 105$ *and varying* h *(left) and for fixed* $h = 2 \cdot 10^{-4}$ *and varying* T *(right).*

$\log_{10} h$	$\log_{10} \epsilon_u$	$\log_{10} \epsilon_{\mu_\rho}$
−0.40	−2.96	−4.37
−0.75	−3.66	−5.08
−1.10	−4.36	−5.78
−1.45	−5.07	−6.49
−1.81	−5.78	−7.19
−2.16	−6.48	−7.89
−2.51	−7.13	−8.60
−2.86	−7.44	−9.30
−3.22	−7.47	−10.00
−3.57	−7.47	−10.71

T	$\log_{10} \epsilon_u$	$\log_{10} \epsilon_{\mu_\rho}$
20	−1.41	−2.77
30	−2.13	−4.38
40	−2.84	−6.24
50	−3.55	−7.06
60	−4.27	−8.47
70	−4.98	−10.49
80	−5.69	−10.94
90	−6.40	−11.06
100	−7.11	−11.06



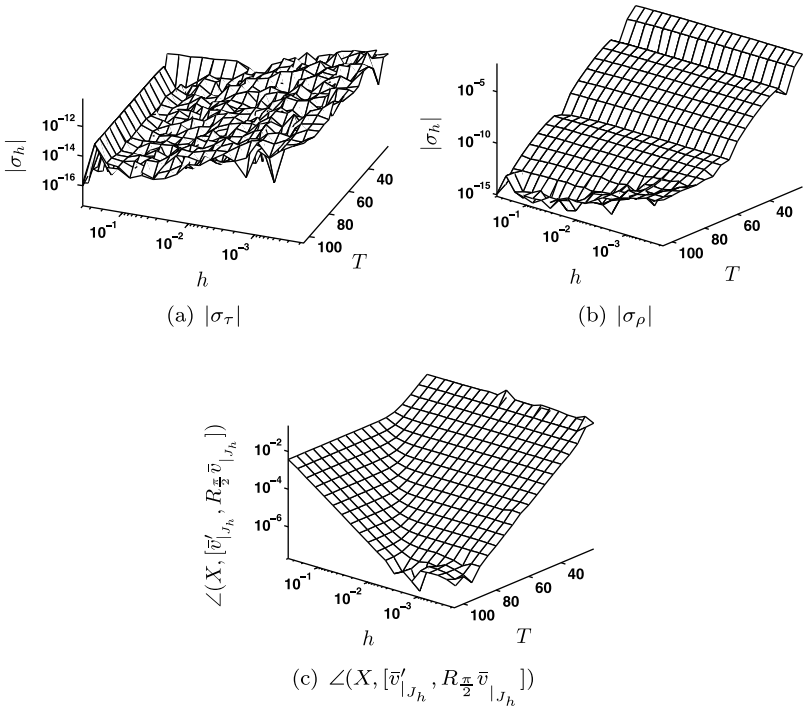
(a) Varying T, fixed h (b) Varying h, fixed T

FIG. 5.3. *Approximation error* $\epsilon_u = \|\tilde{v} - \bar{v}_{|J_h}\|_{\mathcal{H}_h^2}$ *for the unstable pulse for fixed* h *and fixed* T*, respectively.*

conditions. Here the interplay between the discretization error, which varies with the step size h, and the truncation error, which varies with the interval length T, is clearly visible. The grid size h is varied exponentially in $[10^{-4}, 10^{-1}]$ and the size T of the symmetric interval J linearly in $[20, 80]$.

As shown in Figure 5.2 the rate of convergence of μ_ρ to $\bar{\mu}_\rho \approx 0.0573$ is of order 2 in h and the exponential rate in T is $\alpha \approx 0.15$ (cf. Table 5.1). In contrast, μ_τ reaches quickly the range of machine precision where rounding errors dominate and the bad conditioning of the equations in the Newton iteration becomes prominent. The wave $\tilde{v}$ itself converges also as predicted with order 2 in h (cf. Table 5.1 and Figure 5.3(a)) and with $\alpha \approx 0.16$ in T (cf. Table 5.1 and Figure 5.3(b)). This behavior can be observed in $\|\cdot\|_{\mathcal{H}_h^2}$ as well as in $\|\cdot\|_\infty$ (see Figures 5.2(c) and 5.2(d)). Note that the exponent α of convergence w.r.t. the interval length T depends on the asymptotics of the exact solution for $x \to \pm\infty$. In all cases the overall behavior matches estimate (2.10) in Theorem 2.5.

5.2. Approximation of discrete eigenvalues. In this section we compare the approximation error of the solution of the eigenvalue problem (2.4) for the unstable pulse with the estimates (2.11) in Theorem 2.6. In order to solve the eigenvalue problem (2.4) we use a Newton iteration, starting from $\hat{X} = [\bar{v}'_{|J_h}, R_{\frac{\pi}{2}} \bar{v}_{|J_h}]$. Alternatively one can use an Arnoldi iteration to compute the two eigenvalues of smallest magnitude of the corresponding generalized eigenvalue problem. This leads to similar results.



(a) $|\sigma_\tau|$ (b) $|\sigma_\rho|$

(c) $\angle(X, [\bar{v}'_{|_{J_h}}, R_{\frac{\pi}{2}}\bar{v}_{|_{J_h}}])$

FIG. 5.4. *Approximation error for the double zero eigenvalue and the null space.*

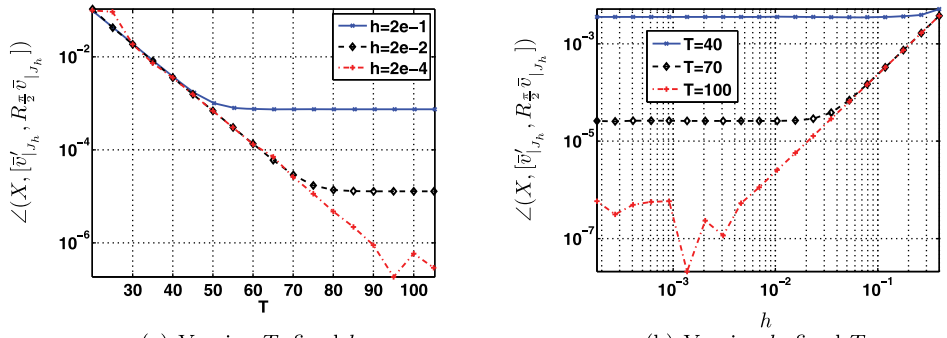
TABLE 5.2

Errors $\epsilon_v = \angle(X, [\bar{v}'_{|_{J_h}}, R_{\frac{\pi}{2}}\bar{v}_{|_{J_h}}])$ *and* $\epsilon_{\sigma_\rho} = |\sigma_\rho|$ *for fixed* $T = 100$ *and varying* h *(left) and for fixed* $h = 2 \cdot 10^{-4}$ *and varying* T *(right).*

$\log_{10} h$	$\log_{10} \epsilon_v$	$\log_{10} \epsilon_{\sigma_\rho}$
−0.40	−2.42	−14.89
−0.75	−3.13	−14.27
−1.10	−3.84	−14.82
−1.45	−4.54	−14.83
−1.81	−5.24	−13.48
−2.16	−5.95	−13.37
−2.51	−6.84	−12.31
−2.86	−6.99	−11.82
−3.22	−6.26	−11.37
−3.57	−6.33	−11.09

T	$\log_{10} \epsilon_v$	$\log_{10} \epsilon_{\sigma_\rho}$
20	−1.00	−2.26
30	−1.73	−5.526
40	−2.45	−5.06
50	−3.16	−6.37
60	−3.88	−8.08
70	−4.59	−9.70
80	−5.33	−10.48
90	−6.05	−10.59
100	−6.23	−10.67

The angle between the approximating subspace X and $\hat{X}$ (see [10]) and the absolute values of the two eigenvalues near zero are shown in Figure 5.4. Here σ_τ denotes the eigenvalue which belongs to the approximation of the translational eigenfunction $\bar{v}'$, and σ_ρ is the eigenvalue which belongs to the approximation of the eigenfunction $R_{\frac{\pi}{2}}\bar{v}$. It can be seen that the eigenvalue σ_τ is in the range of machine precision; thus the errors increase for decreasing h, since the condition of the eigenvalue problem gets worse. The error in the rotational eigenvalue σ_ρ is nearly constant for different h but decreases for increasing T, as expected (cf. Table 5.2). For very small h and large T the increase in error due to the conditioning becomes visible as well. The angle between the invariant subspace which belongs to σ_τ and σ_ρ and the span of the exact



(a) Varying T, fixed h (b) Varying h, fixed T

FIG. 5.5. *Approximation error* $\epsilon_v = \angle(X, [\bar{v}'|_{J_h}, R_{\frac{\pi}{2}}\bar{v}|_{J_h}])$ *for the invariant subspace for fixed* h *and fixed* T*, respectively.*

eigenfunctions $\bar{v}'$ and $R_{\frac{\pi}{2}}\bar{v}$ of Λ restricted to the grid shows the expected behavior. It decreases quadratically in h and exponentially in T with a rate of $\alpha \approx -0.32$ (cf. Table 5.2 and Figure 5.5) until the range of machine precision is reached.

6. Conclusions. This paper gives a detailed study of the numerical approximation of relative equilibria of parabolic equations on the real line. These can be obtained as a stationary solution of a suitable boundary value problem. This is derived in an abstract setting, which reveals their underlying structure and makes it possible to deal with different symmetry groups in a general fashion. The discretization of this boundary value problem with a widely used numerical discretization and rather general boundary conditions is analyzed in detail. In particular, error estimates are given that are uniform in the step size and in the length of the finite interval. Although the use of dynamical systems tools such as exponential dichotomies restricts the analysis to one space dimension, the general methodology described in section 1.1 also works in higer space dimensions [5], [6]. Moreover, many one-dimensional phenomena such as traveling and rotating waves have relevant applications to systems from biology, chemistry, and physics. We expect that the results of this paper as well as some of the techniques can be generalized to higher space dimensions and to more refined structures such as wave trains or modulated waves. This corresponds to point-to-periodic or periodic-to-periodic connecting orbits, for which approximation results for the truncation to finite intervals have been derived in [21].

Acknowledgments. We wish to thank the referees for their valuable comments, which helped improve the exposition of this paper.

REFERENCES

[1] U. M. ASCHER, R. M. M. MATTHEIJ, AND R. D. RUSSELL, *Numerical Solution of Boundary Value Problems for Ordinary Differential Equations*, Classics Appl. Math. 13, SIAM, Philadelphia, 1995.

[2] W.-J. BEYN, W. KLESS, AND V. THÜMMLER, *Continuation of Low-Dimensional Invariant Subspaces in Dynamical Systems of Large Dimension*, in Ergodic Theory, Analysis, and Efficient Simulation of Dynam. Systems, Springer-Verlag, Berlin, 2001, pp. 47–72.

[3] W.-J. BEYN AND J. LORENZ, *Stability of traveling waves: Dichotomies and eigenvalue conditions on finite intervals*, Numer. Funct. Anal. Optim., 20 (1999), pp. 201–244.

[4] W,-J. BEYN AND J.-M. KLEINKAUF, *The numerical computation of homoclinic orbits for maps* SIAM J. Numer. Anal., 34 (1997), pp. 1207–1236.

[5] W.-J. Beyn and V. Thümmler, *Freezing solutions of equivariant evolution equations*, SIAM J. Appl. Dyn. Syst., 3 (2004), pp. 85–116.
[6] W.-J. Beyn and V. Thümmler, *Phase Conditions, Symmetries and PDE Continuation*, in Numerical Continuation Methods for Dynamical Systems, Springer-Verlag, Berlin, 2007, pp. 301–330.
[7] W.-J. Beyn, *The numerical computation of connecting orbits in dynamical systems*, IMA J. Numer. Anal., 9 (1989), pp. 379–405.
[8] P. Chossat and R. Lauterbach, *Methods in Equivariant Bifurcations and Dynamical Systems*, Adv. Ser. Nonlinear Dynam. 15, World Scientific, River Edge, NJ, 2000.
[9] P. de Mottoni and M. Schatzman, *Asymptotics of the Thual-Fauve Pulse*, in Asymptotic and Numerical Methods for Partial Differential Equations with Critical Parameters (Beaune, 1992), NATO Sci. Ser. C Math. Phys. Sci. 384, Kluwer Academic, Dordrecht, 1993, pp. 225–239.
[10] G. H. Golub and C. F. Van Loan, *Matrix Computations*, Johns Hopkins Studies in the Mathematical Sciences, 3rd ed., Johns Hopkins University Press, Baltimore, MD, 1996.
[11] D. Henry, *Geometric Theory of Semilinear Parabolic Equations*, Lecture Notes in Math. 840, Springer-Verlag, Berlin, Heidelberg, New York, 1981.
[12] T. Kato, *Perturbation Theory for Linear Operators*, in Classics in Mathematics, Springer-Verlag, Berlin, 1995. Reprint of the 1980 edition.
[13] H. B. Keller, *Numerical Solution of Bifurcation and Nonlinear Eigenvalue Problems*, in Applications of Bifurcation Theory, Academic Press, New York, 1977, pp. 359–384.
[14] G. J. Lord, D. Peterhof, B. Sandstede, and A. Scheel, *Numerical computation of solitary waves in infinite cylindrical domains*, SIAM J. Numer. Anal., 37 (2000), pp. 1420–1454.
[15] P. Marcq, H. Chaté, and R. Conte, *Exact solutions of the one-dimensional quintic complex Ginzburg-Landau equation*, Phys. D, 73 (1994), pp. 305–317.
[16] J. E. Marsden and T. S. Ratiu, *Introduction to Mechanics and Symmetry*, Texts Appl. Math. 17, 2nd ed., Springer-Verlag, New York, 1999.
[17] G. Moore, *Laguerre approximation of stable manifolds with application to connecting orbits*, Math. Comp., 73 (2004), pp. 211–242.
[18] P. J. Olver, *Applications of Lie Groups to Differential Equations*, Grad. Texts Math. 107, 2nd ed., Springer-Verlag, New York, 1993.
[19] K. J. Palmer, *Exponential dichotomies and transversal homoclinic points*, J. Differential Equations, 55 (1984), pp. 225–256.
[20] K. J. Palmer, *Exponential Dichotomies, the Shadowing Lemma and Transversal Homoclinic Points*, in Dynamics Reported, Vol. 1, B. G. Teubner, Stuttgart, 1988, pp. 265–306.
[21] T. Pampel, *Numerical approximation of connecting orbits with asymptotic rate*, Numer. Math., 90 (2001), pp. 309–348.
[22] C. W. Rowley, I. G. Kevrekidis, J. E. Marsden, and K. Lust, *Reduction and reconstruction for self-similar dynamical systems*, Nonlinearity, 16 (2003), pp. 1257–1275.
[23] C. W. Rowley and J. E. Marsden, *Reconstruction equations and the Karhunen-Loève expansion for systems with symmetry*, Phys. D, 142 (2000), pp. 1–19.
[24] O. Thual and S. Fauve, *Localized structures generated by subcritical instabilities*, J. Phys. France, 49 (1988), pp. 1829–1833.
[25] V. Thümmler, *Numerical Analysis of the Method of Freezing Traveling Waves*, Ph.D. thesis, Bielefeld University, 2005; also available online from www.math.uni-bielefeld.de/∼beyn/AG_Numerik/downloads/preprints/misc/thphd.pdf.
[26] V. Thümmler, *The effect of freezing and discretization to the asymptotic stability of relative equilibria*, J. Dynam. Differential Equations, 20 (2008), pp. 425–477.
[27] G. Vainikko, *Funktionalanalysis der Diskretisierungsmethoden*, B. G. Teubner Verlag, Leipzig, 1976.
[28] W. van Saarloos and P. C. Hohenberg, *Fronts, pulses, sources and sinks in generalized complex Ginzburg-Landau equations*, Phys. D, 56 (1992), pp. 303–367.
[29] Y. Zou and W.-J. Beyn, *On manifolds of connectings orbits in discretizations of dynamical systems*, Nonlinear Anal., 52 (2003), pp. 1499–1520.
[30] Y. Zou and W.-J. Beyn, *On the existence of transversal heteroclinic orbits in discretized dynamical systems*, Nonlinearity, 17 (2004), pp. 2275–2292.

SIAM J. NUMER. ANAL.
Vol. 46, No. 6, pp. 3006–3031
© 2008 Society for Industrial and Applied Mathematics

POLYNOMIAL EXTENSION OPERATORS. PART I*

LESZEK DEMKOWICZ†, JAYADEEP GOPALAKRISHNAN‡, AND JOACHIM SCHÖBERL§

Abstract. In this series of papers, we construct operators that extend certain given functions on the boundary of a tetrahedron into the interior of the tetrahedron, with continuity properties in appropriate Sobolev norms. These extensions are novel in that they have certain polynomial preservation properties important in the analysis of high order finite elements. This part of the series is devoted to introducing our new technique for constructing the extensions, and its application to the case of polynomial extensions from $H^{1/2}(\partial K)$ into $H^1(K)$, for any tetrahedron K.

Key words. Sobolev, polynomial, extension, tangential, normal, trace

AMS subject classifications. 46E35, 46E39, 65N30, 47H60, 11C08, 31B10

DOI. 10.1137/070698786

1. Introduction.

This paper is the first in a series of papers that construct extension operators with certain polynomial preservation properties for the three basic first-order Sobolev spaces

$$H^1(D) = \{u \in L^2(D) : \ \mathbf{grad}\, u \in \boldsymbol{L}^2(D)\} \tag{1.1}$$

$$\boldsymbol{H}(\mathbf{curl}, D) = \{\boldsymbol{u} \in \boldsymbol{L}^2(D) : \ \mathbf{curl}\, \boldsymbol{u} \in \boldsymbol{L}^2(D)\} \tag{1.2}$$

$$\boldsymbol{H}(\mathrm{div}, D) = \{\boldsymbol{u} \in \boldsymbol{L}^2(D) : \ \mathrm{div}\, \boldsymbol{u} \in L^2(D)\}. \tag{1.3}$$

Here the derivatives are understood in the distributional sense, $L^2(D)$ denotes the set of square integrable functions (with respect to the Lebesgue measure) on an open subset D of the three-dimensional Euclidean space, and $\boldsymbol{L}^2(D)$ denotes the set of vector functions whose components are in $L^2(D)$. The domain into which our extensions are performed is a tetrahedron K.

Extension operators are right inverses of trace maps. To describe the traces of the spaces in (1.1)–(1.3), let ϕ be a smooth scalar function and $\boldsymbol{\phi}$ is a smooth vector function on K. Then three standard trace operators are

$$\begin{aligned}
&\mathrm{trc}\, \phi = \phi|_{\partial K}, && \text{(scalar trace)},\\
&\mathrm{trc}_\tau\, \boldsymbol{\phi} = \big(\boldsymbol{\phi} - (\boldsymbol{\phi}\cdot\boldsymbol{n})\boldsymbol{n}\big)\big|_{\partial K}, && \text{(tangential trace)},\\
&\mathrm{trc}_n\, \boldsymbol{\phi} = (\boldsymbol{\phi}\cdot\boldsymbol{n})\big|_{\partial K}, && \text{(normal trace)}.
\end{aligned}$$

where $\boldsymbol{n}$ denotes the outward unit normal on ∂K. It is well known that trc extends to a continuous operator (which we continue to call trc) from $H^1(K)$ onto $H^{1/2}(\partial K)$ [17]. Similarly, the tangential trace trc_τ is well defined on $\boldsymbol{H}(\mathbf{curl}, K)$ and its range is a

*Received by the editors July 30, 2007; accepted for publication (in revised form) April 24, 2008; published electronically September 4, 2008. This work was supported in part by the National Science Foundation under grants 0713833, 0619080, the Johann Radon Institute for Computational and Applied Mathematics (RICAM), and the FWF-Start-Project Y-192 "hp-FEM."
http://www.siam.org/journals/sinum/46-6/69878.html

†Institute of Computational Engineering and Sciences, University of Texas at Austin, 1 University Station, C0200, Austin, TX 78712 (leszek@ices.utexas.edu).

‡Department of Mathematics, University of Florida, Gainesville, FL 32611–8105 (jayg@math.ufl.edu).

§Center for Computational Engineering Science, RWTH Aachen, Pauwelstrasse 19, 52074 Aachen, Germany (joachim.schoeberl@rwth-aachen.de).

subspace of $\boldsymbol{H}^{-1/2}(\partial K)$, and the normal trace trc_n is well defined on $\boldsymbol{H}(\mathrm{div}, K)$ [6] with its range equal to $H^{-1/2}(\partial K)$. We want to construct continuous extension operators that map functions on the boundary of a tetrahedron lying in the range of $\mathrm{trc}, \mathrm{trc}_\tau$, and trc_n into $H^1(K), \boldsymbol{H}(\mathbf{curl}, K)$, and $\boldsymbol{H}(\mathrm{div}, K)$, respectively.

A *polynomial extension operator* is an extension operator with the additional property that whenever the function on ∂K to be extended is the trace of a polynomial on K, the extended function is also a polynomial.

1.1. Goal. Our aim in this series of papers is to construct three polynomial extension operators $\mathcal{E}_K^{\mathrm{grad}}, \mathcal{E}_K^{\mathrm{curl}}$, and $\mathcal{E}_K^{\mathrm{div}}$ on any tetrahedron K such that the following diagram commutes:

$$\begin{array}{ccccc} H^{1/2}(\partial K) & \xrightarrow{\mathbf{grad}_\tau} & \mathrm{trc}_\tau(\boldsymbol{H}(\mathbf{curl}, K)) & \xrightarrow{\mathrm{curl}_\tau} & \mathrm{trc}_n(\boldsymbol{H}(\mathrm{div}, K)) \\ \downarrow \mathcal{E}_K^{\mathrm{grad}} & & \downarrow \mathcal{E}_K^{\mathrm{curl}} & & \downarrow \mathcal{E}_K^{\mathrm{div}} \\ H^1(K) & \xrightarrow{\mathbf{grad}} & \boldsymbol{H}(\mathbf{curl}, K) & \xrightarrow{\mathbf{curl}} & \boldsymbol{H}(\mathrm{div}, K) \end{array} \tag{1.4}$$

The goal of this part of the series is to construct the first in the sequence of these operators, namely, $\mathcal{E}_K^{\mathrm{grad}}$. In the forthcoming parts [11, 12], we will construct the other two polynomial extension operators. Next, we state precisely the properties we require for $\mathcal{E}_K^{\mathrm{grad}}$.

1.2. The $H^1(K)$ polynomial extension problem. The problem we occupy ourselves with in this paper is that of constructing a linear operator $\mathcal{E}_K^{\mathrm{grad}} : H^{1/2}(\partial K) \mapsto H^1(K)$ with the following properties:

(1.5)

The trace of $\mathcal{E}_K^{\mathrm{grad}} u$ on ∂K coincides with u; i.e., $\mathrm{trc}\left(\mathcal{E}_K^{\mathrm{grad}} u\right) = u$.

(1.6)

$\mathcal{E}_K^{\mathrm{grad}}$ is a continuous map from $H^{1/2}(\partial K)$ into $H^1(K)$; i.e., there is a constant C_{grad} independent of u such that

$$\|\mathcal{E}_K^{\mathrm{grad}} u\|_{H^1(K)} \leq C_{\mathrm{grad}} \|u\|_{H^{1/2}(\partial K)} \qquad \text{for all } u \in H^{1/2}(\partial K).$$

(1.7)

If u is a polynomial of degree at most p on each face of K and continuous on ∂K, then $\mathcal{E}_K^{\mathrm{grad}} u$ is a polynomial of degree at most p on K.

The main result of this paper is Theorem 6.1, which solves the problem as stated above.

1.3. Existing work. The concept of extension operators is intimately related with the idea of the trace operators and it has been present in the vast literature on Sobolev spaces for a long time. For instance, the proof of subjectivity of trace operator for standard Sobolev spaces $H^s(\Omega)$ is based on the Lion's construction of a corresponding extension operator; see, e.g., [23, Lemma 3.36]. It is perhaps worth mentioning that, contrary to this trace operator, the extension operator does not break down at half-integers and that the same construction serves the whole Sobolev scale for $s \in \mathbb{R}$.

The subject of polynomial preserving extension operators originates from the convergence analysis for the *p*- and *hp*-versions of the Finite Element Method (FEM).

The first construction of such an operator is for the H^1-space on a two-dimensional region and is due to Babuška and Suri [3]. It contains the origins of many ideas that have been generalized and developed in subsequent contributions. Those include the definition of the primary extension operator (cf. section 2 of this contribution), the solution of a 'two-edge extension problem' using a system of two integral equations, and the analysis of continuity properties of 'edge-to-edge operators' necessary for the solution of an ultimate 'three edge extension' problem. The analysis was carried out first for a triangle and then extended to the case of a square element by using a bilinear map collapsing a square into a triangle. The construction from [3] appears again in a later paper [2] by Babuška, Craig, Mandel, and Pitkäranta, in the context of preconditioning for the p version of the FEM.

The two-dimensional polynomial preserving extension operator was subsequently utilized by Maday [21, 22] to demonstrate that interpolation between polynomial spaces equipped with Sobolev integer norms yields norms equivalent to the standard fractional Sobolev norms. A recent exposition of this subject is contained in [5]. Maday [22] also studied the continuity properties of the Babuška–Suri operator in weighted Sobolev spaces.

The first construction of a polynomial preserving extension operator in three space dimensions from $H^{1/2}(\partial U)$ into $H^1(U)$ was done for a cube U by Ben Belgacem in [4]. The construction utilized the earlier results of Maday mentioned above. The second construction was done for a tetrahedron by Muñoz-Sola [24]. The idea of Muñoz-Sola is rooted in the construction of "face bubble" shape functions for tetrahedral FEM. If ζ is the product of barycentric coordinates of the vertices of a face, then the face bubble functions have ζ as a factor. Muñoz-Sola applies a three-dimensional analogue of the Babuška–Suri extension to the quotient ϕ/ζ, and multiplies the factor ζ back into the extension. The resulting lift vanishes on the remaining faces and displays appropriate continuity properties. The Muñoz-Sola's construction was recently analyzed in [18] in the context of boundary elements, wherein it is shown that the norm of the extension as an operator from L^2 into $\tilde{H}^{1/2}$ grows with polynomial order p as $\log^{1/2} p$. (Here $\tilde{H}^{1/2}$ is the stronger intrinsic norm on the subspace of $H^{1/2}$ with weakly vanishing traces—see [18] for its definition.) Although the tetrahedral H^1 polynomial extension problem, as stated in (1.2), was solved by Muñoz-Sola [24], in this paper we want to present an alternate solution to the same problem. The main reason for presenting our new construction is that our techniques can be generalized to give polynomial extensions in the other two Sobolev spaces in (1.2) and (1.3), as will be amply evident from the subsequent parts of this series [11, 12]. Furthermore, with our techniques, we are able to provide polynomial extensions with the commutativity properties shown in (1.4), which we find aesthetic as well as useful. It is not clear to us if the technique in [24] can yield such extensions, but we do recognize the importance of [24], and indeed some of our arguments are motivated by it.

The need for polynomial preserving extension operators, at least in the analysis of p methods in H^1, is well known, as attested by the above-mentioned works. In fact, they are important also in other areas, such as in the theory of spectral methods even on one element as shown in [15], and for preconditioning as shown in [25]. The importance of polynomial extensions in the context of p and hp approximation theory for Maxwell equations was first recognized by Demkowicz and Babuška in [9], who considered the two-dimensional triangular case. For a triangle T, the construction of a polynomial preserving extension operator from $H^{-1/2}(\partial T)$ into $H(\mathrm{curl}, T)$ follows directly from the corresponding construction for the H^1-case, and presents no essential technical difficulties. The utility of the polynomial extension operators in $H^1(T)$ and

$H(\mathrm{curl}, T)$, forming a commuting de Rham diagram in developing approximations using "projection-based interpolation operators," is evident from [9]. The projection-based interpolation theory was generalized to three space dimensions (tetrahedra) by Demkowicz and Buffa in [10], under the conjecture of existence of commuting, polynomial preserving extension operators for the three spaces $H^1(K)$, $\boldsymbol{H}(\mathbf{curl}, K)$, and $\boldsymbol{H}(\mathrm{div}, K)$ (for a refined version of the theory in context of both tetrahedral and hexahedral elements, see [13]). The extensions we construct in this series of papers establish the truth of this conjecture, thus providing the missing link in the approximation theory of [10].

An alternate technique for constructing commuting $H^1(T)$ and $H(\mathrm{curl}, T)$ polynomial extensions deserves special mention. In two space dimensions, for any triangle T, Ainsworth and Demkowicz [1] constructed polynomial extensions explicitly by solving a system of three integral equations in the spirit of the system of two integral equations analyzed in [3]. The $H^1(T)$ extension operator has smaller norm than the two-dimensional version of the Muñoz-Sola's operator [24] and can be used for constructing optimal shape functions for a triangular element. The Ainsworth–Demkowicz operator was also shown to map $L^2(\partial K)$ into $H^{1/2}(K)$, with a constant independent of polynomial order p (cf. this result with that of [18] mentioned above). Unfortunately, our attempts to extend the Ainsworth–Demkowicz technique to three dimensions failed as the analysis reached prohibitive levels of complexity.

To the best of our knowledge, our commuting polynomial extensions for a tetrahedron is the first result of this kind. To compare with the most recent other work in this direction that we know of, the contribution of Costabel, Dauge, and Demkowicz [8] presents an analogous family of commuting extension operators defined on polynomial spaces on a cube. The construction mimics separation of variables for polynomial spaces and is based on the Maday's spectral equivalence results for fractional spaces mentioned earlier. However, these extension operators defined on polynomial spaces change with the polynomial degree, although their norms are shown to be independent of polynomial degree p. In contrast, the expressions defining our extensions do not vary with degree.

1.4. Overview of our techniques. Our approach to solve the problem stated above starts with a study of the simpler case of extending a given function from just one face. Then we will analyze how to modify this extension process to solve the case when data is given on more faces. We highlight the main new techniques in the construction of our extension operators:

1. *Primary extensions:* We start with processes that extend functions given on the plane $\mathbb{R}^2$, which we call "primary extensions". In constructing the final extension from ∂K, the first step is to pick a face of ∂K and apply the primary extension from that face. Of course, such an extended function in general will not have the needed traces in the remaining faces.
2. *Corrections:* The next step is to "correct" the above-mentioned incorrect traces on the remaining faces. This step is divided into the construction of "face correction operators," "edge correction operators," and a "vertex correction operator." Most of the technical aspects of the presented construction are in the design of these corrections.
3. *Commutativity:* Once we construct the first operator appearing in the commuting diagram (1.4), namely $\mathcal{E}_K^{\mathrm{grad}}$, then the construction of the succeeding operators is motivated by the commutativity properties in (1.4). Indeed, to obtain $\mathcal{E}_K^{\mathrm{curl}}$, we consider each of the primary and correction operators that

went into the design of the preceding operator $\mathcal{E}_K^{\text{grad}}$, and find corresponding $\boldsymbol{H}(\mathbf{curl})$ operators that commute with it. Similarly, the construction of the $\boldsymbol{H}(\text{div})$ extension will proceed by examining the steps in the construction of the $\boldsymbol{H}(\mathbf{curl})$ extension and finding their commuting $\boldsymbol{H}(\text{div})$ analogues. This will be clear from Parts II [11] and III [12] of this series.

4. *Regular decomposition of traces:* In the $\boldsymbol{H}(\mathbf{curl})$ and $\boldsymbol{H}(\text{div})$ cases where traces are in Sobolev spaces of negative index, we will characterize the trace space using a decomposition involving Sobolev spaces of positive indices only. This will feature in Parts II [11] and III [12] only, as the trace space of $H^1(K)$ is $H^{1/2}(\partial K)$, a Sobolev space of positive index.
5. *Weighted norm estimates:* The extension operators we construct are all integral operators with polynomial kernels. To establish their continuity properties in appropriate fractional Sobolev norms, we use weighted norm estimates. We are able to do this even when traces are in Sobolev spaces of negative index, because of the above-mentioned regular decomposition of the trace space. A number of weighted norm estimates used throughout this series are given in Appendix A of this part.

1.5. Notations. For ready reference, we quickly list here some notations that we will use throughout this series.

The "reference tetrahedron" $\hat{K}$ is defined by

$$\hat{K} = \{(x, y, z) : x \geq 0,\ y \geq 0,\ x + y + z \leq 1\},$$

with the following enumeration of vertices:

$$\hat{\boldsymbol{a}}_0 = (0, 0, 0),\ \hat{\boldsymbol{a}}_1 = (1, 0, 0),\ \hat{\boldsymbol{a}}_2 = (0, 1, 0),\ \text{and } \hat{\boldsymbol{a}}_3 = (0, 0, 1).$$

The face of $\hat{K}$ opposite to vertex $\hat{\boldsymbol{a}}_i$ is denoted by $\hat{F}_i$. The face $\hat{F}_3$ is distinguished in that the data for the primary extensions is given there, so we shall also denote it simply by $\hat{F}$. The edge connecting $\hat{\boldsymbol{a}}_i$ and $\hat{\boldsymbol{a}}_j$ is denoted by $\hat{E}_{ij}$.

We denote by K a generic tetrahedron (of positive volume). When the notations defined for $\hat{K}$ above are employed without the superscript (hat), they denote the corresponding geometrical objects on a general tetrahedron K; e.g., $\boldsymbol{a}_i$ denotes the vertices of K, and F_i is the face of K opposite to the vertex $\boldsymbol{a}_i$.

Let $\lambda_i(\boldsymbol{x})$, for $i = 0, \ldots 3$ be the linear function on K satisfying $\lambda_i(\boldsymbol{a}_j) = \delta_{ij}$. The *affine coordinates* (or barycentric coordinates) of a point in K are the values of λ_i at that point arranged into a 4-tuple $(\lambda_i, \lambda_j, \lambda_k, \lambda_l)$, where the order is not significant. Throughout this series, the indices i, j, k, l are a permutation of $0, 1, 2, 3$.

Many integral operators we consider will require us to integrate over subtriangles of a face F_l of K. We now express these subtriangles using the affine coordinates of F_l, namely $\lambda_m^{F_l} = \lambda_m|_{F_l}$ for $m = i, j$, or k. For any permutation $\{i, j, k, l\}$ of $\{0, 1, 2, 3\}$, we define

$$T_l(r_i, r_j, r_k) = \left\{\boldsymbol{x} \in F_l :\ \lambda_i^{F_l}(\boldsymbol{x}) \geq r_i,\ \lambda_j^{F_l}(\boldsymbol{x}) \geq r_j,\ \text{and } \lambda_k^{F_l}(\boldsymbol{x}) \geq r_k\right\}. \tag{1.8}$$

Note that the order of the arguments r_i, r_j, r_k in the notation $T_l(r_i, r_j, r_k)$ is not significant, but the subscripts of these arguments indicate to which affine coordinate it corresponds. This region is illustrated in Figure 1. Also, define

$$T_l(0, r_j, r_k) = \left\{\boldsymbol{x} \in F_l :\ \lambda_j^{F_l}(\boldsymbol{x}) \geq r_j,\ \text{and } \lambda_k^{F_l}(\boldsymbol{x}) \geq r_k\right\}, \tag{1.9}$$

$$T_l(0, 0, r_i) = \left\{\boldsymbol{x} \in F_l :\ \lambda_i^{F_l}(\boldsymbol{x}) \geq r_i\right\}. \tag{1.10}$$

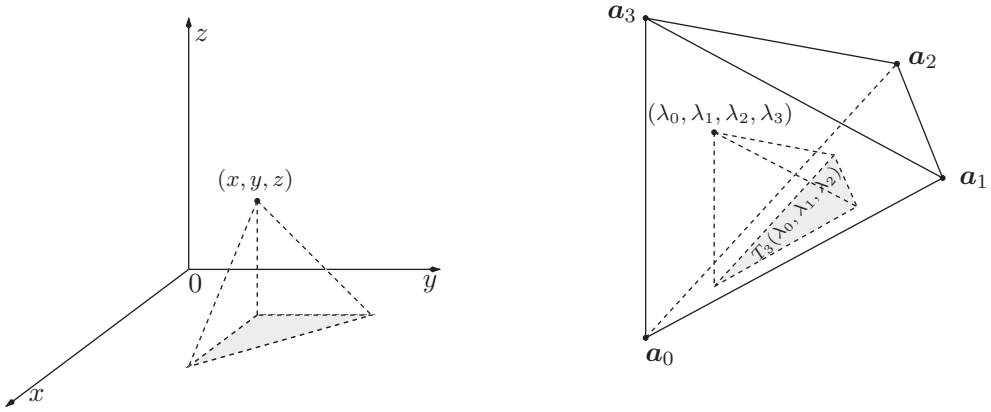


FIG. 1. *The regions of integration that define* $\mathcal{E}^{\mathrm{grad}}u$. *The left figure shows the integration region in gray for the extension from the x-y plane. The right figure shows the integration region for the mapped extension from a face of a general tetrahedron.*

Note that the definition of $T_l(0, r_j, r_k)$ is consistent with (1.8) when $r_i = 0$. Similarly the definition of $T_l(0, 0, r_i)$ is consistent with (1.8) when $r_j = r_k = 0$. In particular, from the indices of the arguments, we judge which barycentric coordinates are zero; e.g., in $T_l(0, r_k, 0)$, since l and k have already appeared, and since $\{i, j, k, l\}$ is a permutation of $\{0, 1, 2, 3\}$, we understand that the affine coordinates with indices that have not appeared, namely $\lambda_i^{F_l}$ and $\lambda_j^{F_l}$, are simply greater than or equal to zero: $T_l(0, r_k, 0) = \{\boldsymbol{x} \in F_l : \ \lambda_k^{F_l}(\boldsymbol{x}) \geq r_k, \ \lambda_i^{F_l}(\boldsymbol{x}) \geq 0, \ \lambda_j^{F_l}(\boldsymbol{x}) \geq 0\}$, which is consistent with (1.8). See Figures 1, 3, and 4 for illustrations of these subtriangles.

Our notation for Sobolev spaces is standard: Let $H^s(D)$ denote the standard Sobolev space of order s on domain D; e.g., when $s = 1$ the definition is as in (1.1). See, e.g., [14, 19, 23] for the more complex definitions in the case when s is fractional. The definitions of $\boldsymbol{H}(\mathbf{curl}, D)$ and $\boldsymbol{H}(\mathrm{div}, D)$ are already given in (1.2)–(1.3). To shorten the notation, when the domain is K, we often simply write $\boldsymbol{H}(\mathbf{curl})$ for $\boldsymbol{H}(\mathbf{curl}, K)$ (and similarly $\boldsymbol{H}(\mathrm{div})$).

In inequalities bounding function norms, we denote by C (or C with some subscript) a generic constant whose value at different occurrences may differ but is independent of the functions involved.

2. The primary extension operator. The primary extension operator for the $H^1(K)$ case follows from the well-known, two-dimensional polynomial extension from a line [2], as generalized to three space dimensions in [24]. Suppose $u(x, y)$ is a smooth function given on the face $\hat{F}$ of $\hat{K}$ (see section 1.5 for notation). The primary extension maps u to a function $\mathcal{E}^{\mathrm{grad}}u$ defined on $\hat{K}$ as follows:

$$\mathcal{E}^{\mathrm{grad}}u\,(x, y, z) = \frac{2}{z^2}\int_x^{x+z}\int_y^{x+y+z-\widetilde{x}} u(\widetilde{x}, \widetilde{y})\, d\widetilde{y}\, d\widetilde{x}. \tag{2.1}$$

Clearly, the value of the extension at the point (x, y, z) is determined by integrating u over the triangle with vertices $(x, y, 0), (x + z, y, 0)$, and $(x, y + z, 0)$, as shown in the first illustration of Figure 1. Note that $\mathcal{E}^{\mathrm{grad}}$ can be thought of either as an extension from the plane $\mathbb{R}^2$ into the adjacent infinite slab $\mathbb{R}^2 \times (0, 1) \equiv \{(x, y, z) : 0 < z < 1\}$, or as an extension from the face $\hat{F}$ into $\hat{K}$.

THEOREM 2.1. *The following statements apply to* $\mathcal{E}^{\mathrm{grad}}$*:*

(1) *The operator* $\mathcal{E}^{\mathrm{grad}}$ *defined for smooth functions* u *by* (2.1) *extends as a continuous operator from* $H^{1/2}(\hat{F})$ *to* $H^1(\hat{K})$*; i.e., there is a constant* $C > 0$ *such that*

$$\|\mathcal{E}^{\mathrm{grad}}u\|_{H^1(\hat{K})} \leq C\|u\|_{H^{1/2}(\hat{F})},$$

for all u *in* $H^{1/2}(\hat{F})$.

(2) *The trace of* $\mathcal{E}^{\mathrm{grad}}u$ *on* $\hat{F}$ *equals* u *for all* u *in* $H^{1/2}(\hat{F})$.

(3) *If* u *is a polynomial of degree at most* p*, then* $\mathcal{E}^{\mathrm{grad}}u$ *is also a polynomial of degree at most* p.

Proof. The proof of the continuity estimate in the first item proceeds by a standard technique involving the Fourier transform. We have included this proof in Appendix B. (In Part II [11], we will develop an alternate technique.)

The second statement of the theorem is immediately verified for smooth u, once we rewrite (2.1) by a change of variable as

$$\mathcal{E}^{\mathrm{grad}}u\,(x,y,z) = 2\int_0^1\int_0^{1-s} u(x+sz, y+tz)\,dt\,ds. \tag{2.2}$$

So, by the standard density argument, it follows for all $u \in H^{1/2}(\hat{F})$.

To prove the last statement of the theorem, observe that if u is a polynomial of degree at most p, then the integrand $u(x+sz, y+tz)$ in (2.2) is a polynomial of degree at most p in x, y, and z, with coefficients depending on s and t. After integrating over s and t, we continue to have a polynomial of degree at most p in x, y, and z. ∎

Since any tetrahedron can be mapped one-one onto $\hat{K}$ using an affine map, the above definition of $\mathcal{E}^{\mathrm{grad}}$ automatically defines an extension from a face for any tetrahedron. The expressions for such mapped extensions are greatly simplified by the use of affine coordinates (or barycentric coordinates) of the tetrahedron. Let K be any tetrahedron. Suppose u is a smooth function defined on a face F_l of K and we want to extend it to K. The extension operator on K now involves integration of u over a subtriangle of F_l. We express such subtriangles using the notations established in section 1.5. The extension from F_l into K is defined by

$$\mathcal{E}_l^{\mathrm{grad}}u\,(\lambda_i,\lambda_j,\lambda_k,\lambda_l) = \frac{2}{\lambda_l^2}\iint_{T_l(\lambda_i,\lambda_j,\lambda_k)} u(\boldsymbol{s})\,d\boldsymbol{s}, \tag{2.3}$$

where $\boldsymbol{s}$ is the two-dimensional variable of integration running over the subtriangle $T_l(\lambda_i,\lambda_j,\lambda_k)$ defined in (1.8) and $d\boldsymbol{s}$ is the standard (two-dimensional) Lebesgue measure on this subtriangle. (The region of integration is illustrated in Figure 1.) Clearly, the definition in (2.3) reduces to (2.1) if we choose $l = 3$ and $K = \hat{K}$. An analogue of Theorem 2.1 obviously holds for the operator in (2.3) by mapping.

3. Face corrections. From the definition of $\mathcal{E}_l^{\mathrm{grad}}u$, it is clear that when u is a smooth function on F_l which vanishes on the boundary ∂F_l, the extended function $\mathcal{E}_l^{\mathrm{grad}}u$, in general, does not have zero trace on $\partial K \setminus F_l$. Now we show how to add a correction to $\mathcal{E}_l^{\mathrm{grad}}u$ so that the corrected result will have zero trace on a face other than F_l.

For definiteness, we first consider the situation on the face $\hat{F}_1$ of the reference tetrahedron $\hat{K}$ after the extension $\mathcal{E}^{\mathrm{grad}}u$ has been performed. The correction term

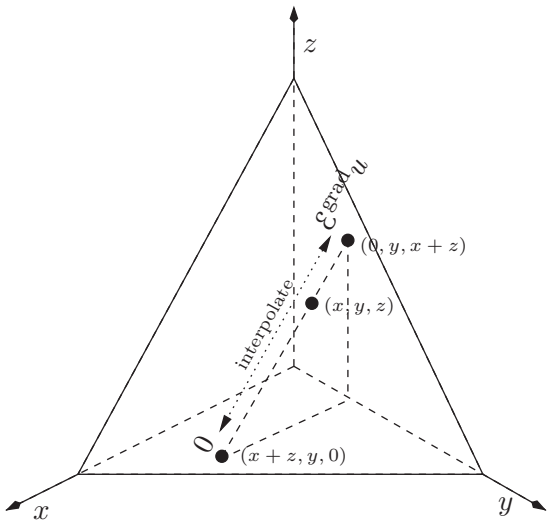


FIG. 2. *Illustration of the face correction process.*

for $\hat{F}_1$, namely, $\mathcal{E}^{\text{grad}}_{\hat{F}_1}u$, must be so that the corrected result $\mathcal{E}^{\text{grad}}u - \mathcal{E}^{\text{grad}}_{\hat{F}_1}u$ achieves zero trace on $\hat{F}_1$ without altering the trace on the original face $\hat{F}$. Therefore, we design the correction term by linear interpolation between the value of $\mathcal{E}^{\text{grad}}u$ on $\hat{F}_1$ and 0 along the lines where $x+z$ and y are constant (see Figure 2); i.e., we define

$$\mathcal{E}^{\text{grad}}_{\hat{F}_1}u\,(x,y,z) = \frac{z}{x+z}\,\mathcal{E}^{\text{grad}}u(0,y,x+z). \tag{3.1}$$

The interpolation process that we employed here has the following interesting continuity property in weighted spaces (whose proof appears in Appendix A). Our notation for weighted spaces is as follows: Let $L^2_w(D)$ denote the weighted space of all Lebesgue measurable functions f on a domain D (in $\mathbb{R}^n$ – in this paper $n = 1, 2$, or 3) such that $\|f\|^2_{L^2_w(D)} \equiv \int_D w\,|f|^2\,dx < \infty$. Here $w(x)$ is a nonnegative weight function on D and dx is the standard Lebesgue measure on D. Similarly, the weighted Sobolev space $H^1_w(D)$ consists of all functions in $L^2_w(D)$ whose first-order distributional derivatives are also in $L^2_w(D)$.

LEMMA 3.1. *The map $\mathcal{B}_{\hat{F}_1}$ defined by*

$$\phi(y,z) \;\mapsto\; \mathcal{B}_{\hat{F}_1}\phi := \frac{z}{x+z}\phi(y,x+z)$$

is continuous from $L^2_{1/z}(\hat{F}_1) \cap H^1_z(\hat{F}_1)$ into $H^1(\hat{K})$.

To study the face correction operator in (3.1), in addition to the above lemma, we shall need some more properties of the primary extension $\mathcal{E}^{\text{grad}}u$ involving weighted spaces.

LEMMA 3.2.

(1) *Let $\mathcal{R}^{\text{grad}}$ map smooth functions u on face $\hat{F}_3$ to functions on face $\hat{F}_1$ by*

$$u(x,y) \;\mapsto\; \mathcal{R}^{\text{grad}}u\,(y,z) := \mathcal{E}^{\text{grad}}u(0,y,z).$$

$\mathcal{R}^{\text{grad}}$ extends to a continuous map from $L^2_{1/x}(\hat{F}_3)$ into $L^2_{1/z}(\hat{F}_1) \cap H^1_z(\hat{F}_1)$.

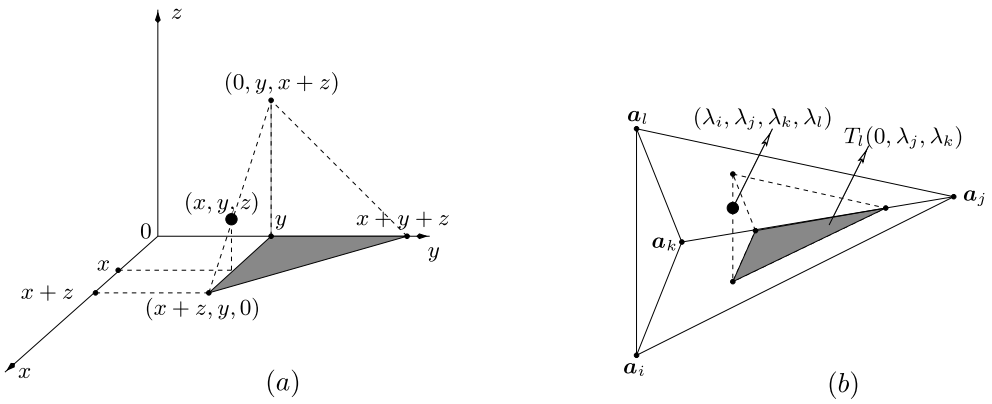


FIG. 3. *Domain of integration for face corrections* (*a*) *in the reference domain and* (*b*) *in a general tetrahedron.*

(2) *Let* $\mathcal{L}^{\mathrm{grad}}$ *map smooth functions* u *on face* $\hat{F}_3$ *to functions on edge* $\hat{E}_{03}$ *by*

$$u(x,y) \;\mapsto\; \mathcal{L}^{\mathrm{grad}}u\,(z) := \mathcal{E}^{\mathrm{grad}}u(0,0,z).$$

$\mathcal{L}^{\mathrm{grad}}$ extends to a continuous map from $L^2_{1/x}(\hat{F}_3)\cap L^2_{1/y}(\hat{F}_3)$ into $L^2(\hat{E}_{03})\cap H^1_{z^2}(\hat{E}_{03})$.

Proofs of this and all other lemmas appear in Appendix A. Note that although Theorem 2.1 needs u to be in $H^{1/2}(\hat{F})$ for $\mathcal{E}^{\mathrm{grad}}u$ to be in $H^1(\hat{K})$, the above lemma shows that the traces of $\mathcal{E}^{\mathrm{grad}}u$ on a face and an edge are well defined even when u is only in a weighted L^2 space.

Before stating the properties of $\mathcal{E}^{\mathrm{grad}}_{\hat{F}}$, let us extend its definition for any face of a general tetrahedron K, using the affine coordinate notation established in section 1.5. The face correction operator for a face F_i to correct the primary extension from face F_l is defined in affine coordinates by

$$\mathcal{E}^{\mathrm{grad}}_{F_i,l}u\,(\lambda_0,\lambda_1,\lambda_2,\lambda_3) = \frac{2\lambda_l}{(\lambda_i+\lambda_l)^3}\iint\limits_{T_l(0,\lambda_j,\lambda_k)} u(\boldsymbol{s})\,d\boldsymbol{s}, \tag{3.2}$$

where $T_l(0,r_j,r_k)$ is as defined in (1.9). The region of integration is illustrated in Figure 3.

Now suppose u is a smooth function on F_l which vanishes on one of its edges, say the edge connecting vertices $\boldsymbol{a}_j$ and $\boldsymbol{a}_k$, which we denote by E_{jk}. This edge is shared by F_i and F_l. The *two-face problem* is the problem of finding a polynomial extension of u from F_l that is zero on F_i. With the above-defined face correction, we are now able to solve the two-face problem. The extension operator that solves the F_i-F_l two-face problem is defined by

$$\mathcal{E}^{\mathrm{grad}}_{i,l}u = \mathcal{E}^{\mathrm{grad}}_{l}u - \mathcal{E}^{\mathrm{grad}}_{F_i,l}u. \tag{3.3}$$

It can be extended to an operator on

$$H^{1/2}_{0,i}(F_l) = H^{1/2}(F_l)\cap L^2_{1/\lambda_i}(F_l) \tag{3.4}$$

with the following properties:

PROPOSITION 3.1. *The two-face extension* $\mathcal{E}^{\mathrm{grad}}_{i,l}$ *satisfies the following:*

(1) $\mathcal{E}^{\mathrm{grad}}_{i,l}$ *is a continuous operator from* $H^{1/2}_{0,i}(F_l)$ *into* $H^1(K)$.
(2) *For all* $u \in H^{1/2}_{0,i}(F_l)$, *the trace of* $\mathcal{E}^{\mathrm{grad}}_{i,l}u$ *on* F_i *is zero, while its trace on* F_l *equals* u.
(3) *If* u *is a polynomial of degree at most* p *that vanishes on* E_{jk}, *then* $\mathcal{E}^{\mathrm{grad}}_{i,l}u$ *is a polynomial of degree at most* p.

Proof. To prove item (3.1), observe that since Theorem 2.1 yields

$$\|\mathcal{E}^{\mathrm{grad}}_{l}u\|_{H^1(K)} \le C\|u\|_{H^{1/2}(F_l)} \le C\|u\|_{H^{1/2}_{0,i}(F_l)},$$

it suffices to prove that $\mathcal{E}^{\mathrm{grad}}_{F_i,l}$ is a continuous map from $H^{1/2}_{0,i}(F_l)$ into $H^1(K)$. In fact, $\mathcal{E}^{\mathrm{grad}}_{F_i,l}$ is continuous on the larger space $L^2_{1/\lambda_i}(F_l)$, as we show now. Because $\mathcal{E}^{\mathrm{grad}}_{F_i,l}$ for any i is obtained by mapping $\mathcal{E}^{\mathrm{grad}}_{\hat{F}_1}$ from $\hat{K}$, it suffices to prove that $\mathcal{E}^{\mathrm{grad}}_{\hat{F}_1} : L^2_{1/x}(\hat{F}_3) \mapsto H^1(\hat{K})$ is continuous. By definition, $\mathcal{E}^{\mathrm{grad}}_{\hat{F}_1}$ is the composition

$$\mathcal{E}^{\mathrm{grad}}_{\hat{F}_1} = \mathcal{B}_{\hat{F}_1} \circ \mathcal{R}^{\mathrm{grad}}.$$

By Lemma 3.2, $\mathcal{R}^{\mathrm{grad}} : L^2_{1/x}(\hat{F}_3) \mapsto L^2_{1/z}(\hat{F}_1) \cap H^1_z(\hat{F}_1)$ is continuous, and by Lemma 3.1, $\mathcal{B}_{\hat{F}_1} : L^2_{1/z}(\hat{F}_1) \cap H^1_z(\hat{F}_1) \mapsto H^1(\hat{K})$ is continuous, hence the continuity of their composition follows.

Proof of (3.1): By (2.3) and (3.2),

$$\mathcal{E}^{\mathrm{grad}}_{i,l}u = \frac{2}{\lambda_l^2} \iint\limits_{T_l(\lambda_i,\lambda_j,\lambda_k)} u(\boldsymbol{s})\, d\boldsymbol{s} \;-\; \frac{2\lambda_l}{(\lambda_i+\lambda_l)^3} \iint\limits_{T_l(0,\lambda_j,\lambda_k)} u(\boldsymbol{s})\, d\boldsymbol{s}.$$

Since $\lambda_i = 0$ on face F_i, we immediately see after setting $\lambda_i = 0$ above, that the trace on F_i vanishes. The trace on F_l is u because the the last term above vanishes upon setting $\lambda_l = 0$, while the trace of $\mathcal{E}^{\mathrm{grad}}_{l}u$ on F_l is u (by Theorem 2.1).

Proof of (3.1): Going back to the reference tetrahedron, observe that if u vanishes on edge along the y-axis, then

$$u(x,y) = \, x\, u_{p-1}(x,y)$$

for some polynomial u_{p-1} of degree at most $p-1$. Using this in the face correction expression in (3.1), we have

$$\begin{aligned}
\left(\mathcal{E}^{\mathrm{grad}}_{\hat{F}_1}u\right)(x,y,z) &= \frac{z}{x+z}\int_0^1\int_0^{1-t} u(s(x+z), y+t(x+z))\, ds\, dt \\
&= \frac{z}{x+z}\int_0^1\int_0^{1-t} s(x+z)u_{p-1}(s(x+z), y+t(x+z))\, ds\, dt \\
&= z\int_0^1\int_0^{1-t} s u_{p-1}(s(x+z), y+t(x+z))\, ds\, dt.
\end{aligned}$$

The integral above is a polynomial of degree at most $p-1$ by the same arguments as in the proof of Theorem 2.1 (3.1); hence, the result follows. ∎

4. Edge corrections. If the function u (to be extended from face F_l) vanishes on ∂F_l, then we want the extended function to vanish on all faces other than F_l. After an application of the face correction operator for face F_i, the extension has zero trace on F_i. To obtain zero trace on another face, say F_j, we can consider applying the face correction operator for F_j. Unfortunately, after this second correction the resulting total trace on F_i will no longer be zero, in general. This necessitates the use of further correction operators which we discuss now.

Let us first consider the case of the reference tetrahedron $\hat{K}$ after the application of the face correction $\mathcal{E}^{\mathrm{grad}}_{\hat{F}_1}$. This operator alters the traces on face $\hat{F}_2$. In order to return this trace to its original setting, we use an additional correction operator whose trace coincides with that of $\mathcal{E}^{\mathrm{grad}}_{\hat{F}_1}u$ on $\hat{F}_2$. It is defined by linearly interpolating the value of $\mathcal{E}^{\mathrm{grad}}_{\hat{F}_1}u$ on this face down to zero along the lines where $y+z$ and x are constant:

$$\begin{aligned}\mathcal{E}^{\mathrm{grad}}_{\hat{E}_{03}}u(x,y,z) &= \frac{z}{y+z}\mathcal{E}^{\mathrm{grad}}_{\hat{F}_1}u(x,0,y+z),\\ &= \frac{z}{x+y+z}\mathcal{E}^{\mathrm{grad}}u(0,0,x+y+z), && \text{by (3.1)} \qquad (4.1)\\ &= \frac{2z}{(x+y+z)^3}\int_0^{x+y+z}\int_0^{x+y+z-\widetilde{x}} u(\widetilde{x},\widetilde{y})\,d\widetilde{y}\,d\widetilde{x}, && \text{by (2.1).} \qquad (4.2)\end{aligned}$$

We call this operator the "edge correction operator" for the $\hat{E}_{03}$ edge, because as is clear from (4.1), its action depends only on the value of $\mathcal{E}^{\mathrm{grad}}u$ along $\hat{E}_{03}$. The interpolation process from this edge to the tetrahedron implicit in (4.1) has the following continuity property:

LEMMA 4.1. *The map* $\mathcal{B}_{\hat{E}_{03}}$ *defined by*

$$\phi(z) \;\mapsto\; \mathcal{B}_{\hat{E}_{03}}\phi := \frac{z}{x+y+z}\phi(x+y+z)$$

is continuous from $L^2(\hat{E}_{03})\cap H^1_{z^2}(\hat{E}_{03})$ *into* $H^1(\hat{K})$.

Next, we generalize the expression in (4.2) to one in affine coordinates on a general tetrahedron. Let K be a general tetrahedron with the function to be extended given on face F_l. Let E_{il} denote the edge connecting the vertex $\boldsymbol{a}_i$ to $\boldsymbol{a}_l$. Then, the correction operator associated to the edge E_{il} is

$$\mathcal{E}^{\mathrm{grad}}_{E_{il},l}u\,(\lambda_0,\lambda_1,\lambda_2,\lambda_3) = \frac{2\lambda_l}{(1-\lambda_i)^3}\iint\limits_{T_l(0,0,\lambda_i)} u(\boldsymbol{s})\,d\boldsymbol{s}, \qquad (4.3)$$

where $T_l(0,0,r_i)$ is as defined in (1.10). The integration domain is illustrated in Figure 4.

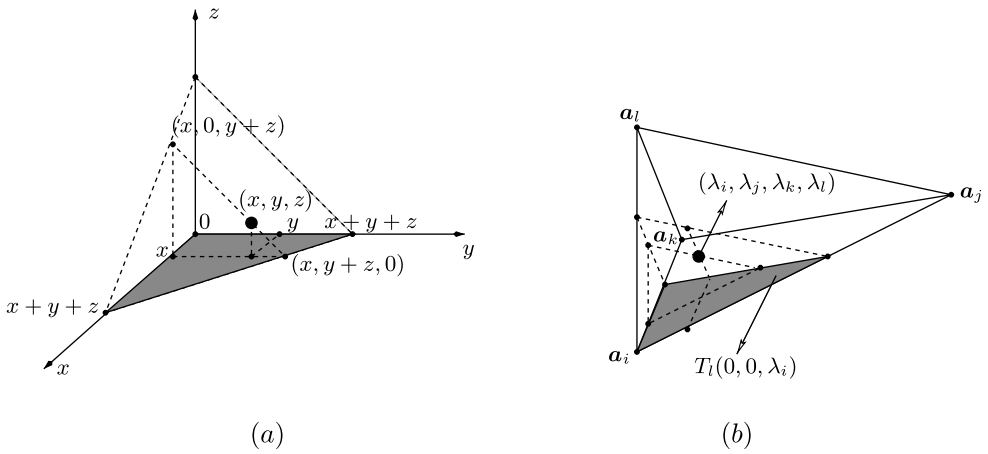
We can now solve the *three-face problem* of finding an extension of u from F_l that is zero on F_i and F_j whenever u is a smooth function that vanishes on edges E_{jk} and E_{ik}. The F_i-F_j-F_l three-face problem is solved by the extension operator

$$\mathcal{E}^{\mathrm{grad}}_{ij,l} = \mathcal{E}^{\mathrm{grad}}_{l} - \mathcal{E}^{\mathrm{grad}}_{F_i,l} - \mathcal{E}^{\mathrm{grad}}_{F_j,l} + \mathcal{E}^{\mathrm{grad}}_{E_{kl},l}, \qquad (4.4)$$

whose properties appear in the next proposition. The space to which $\mathcal{E}^{\mathrm{grad}}_{ij,l}$ can be extended continuously is

$$H^{1/2}_{0,ij}(F_l) = H^{1/2}(F_l)\;\cap\; L^2_{1/\lambda_i}(F_l)\;\cap\; L^2_{1/\lambda_j}(F_l) \qquad (4.5)$$

as seen below.



(a) (b)

FIG. 4. *Domain of integration for the edge correction operators in* (a) *reference domain, and* (b) *general tetrahedron.*

PROPOSITION 4.1. *The three-face extension* $\mathcal{E}^{\mathrm{grad}}_{ij,l}$ *satisfies the following:*

(1) $\mathcal{E}^{\mathrm{grad}}_{ij,l}$ *is a continuous operator from* $H^{1/2}_{0,ij}(F_l)$ *into* $H^1(K)$.

(2) *For all* $u \in H^{1/2}_{0,ij}(F_l)$ *the traces of the edge and face corrections satisfy*

$$\mathcal{E}^{\mathrm{grad}}_{F_i,l} u\big|_{F_k} = \mathcal{E}^{\mathrm{grad}}_{E_{jl},l} u, \tag{4.6}$$

$$\left(\mathcal{E}^{\mathrm{grad}}_{E_{jl},l} u\right)\big|_{F_j} = 2\lambda_l \iint_{F_l} u(\boldsymbol{s})\, d\boldsymbol{s}, \tag{4.7}$$

$$\left(\mathcal{E}^{\mathrm{grad}}_{E_{jl},l} u\right)\big|_{F_l} = 0, \tag{4.8}$$

$$\mathcal{E}^{\mathrm{grad}}_{ij,l} u\big|_{F_i \cup F_j} = 0. \tag{4.9}$$

(3) *If* u *is a polynomial of degree at most* p *that vanishes on* $E_{ij} \cup E_{jk}$, *then* $\mathcal{E}^{\mathrm{grad}}_{E_{jl},l} u$ *is a polynomial of degree at most* p *for all* $j = 0, 1, 2$.

Proof. To prove the first item, in view of Theorem 2.1(1) and Proposition 3.1(1), it suffices to prove that the last term in (4.4), namely, $\mathcal{E}^{\mathrm{grad}}_{E_{kl},l}$, is a continuous operator from $H^{1/2}_{0,ij}$ into $H^1(K)$. Since $\mathcal{E}^{\mathrm{grad}}_{E_{kl},l}$ is obtained by mapping $\mathcal{E}^{\mathrm{grad}}_{\hat{E}_{03}}$ from $\hat{K}$, it suffices to prove the continuity of $\mathcal{E}^{\mathrm{grad}}_{\hat{E}_{03}}$. But $\mathcal{E}^{\mathrm{grad}}_{\hat{E}_{03}} = \mathcal{B}_{\hat{E}_{03}} \circ \mathcal{L}^{\mathrm{grad}}$, so by Lemmas 3.2 and 4.1, we conclude that $\mathcal{E}^{\mathrm{grad}}_{\hat{E}_{03}}$ is continuous from $L^2_{1/x}(\hat{F}_3) \cap L^2_{1/y}(\hat{F}_3)$ into $H^1(\hat{K})$. Thus $\mathcal{E}^{\mathrm{grad}}_{E_{kl},l}$ is continuous from $H^{1/2}_{0,ij}$ into $H^1(K)$.

To prove (4.6), we start with

$$\mathcal{E}^{\mathrm{grad}}_{F_i,l} u - \mathcal{E}^{\mathrm{grad}}_{E_{jl},l} u = \frac{2\lambda_l}{(\lambda_i + \lambda_l)^3} \iint_{T_l(0,\lambda_j,\lambda_k)} u(\boldsymbol{s})\, d\boldsymbol{s} \;-\; \frac{2\lambda_l}{(1-\lambda_j)^3} \iint_{T_l(0,\lambda_j,0)} u(\boldsymbol{s})\, d\boldsymbol{s}.$$

Since $\lambda_k = 0$ on face F_k, by setting $\lambda_k = 0$ in the first integral above, and observing that $1 - \lambda_j = \lambda_i + \lambda_l$, we see that the right-hand side above is zero, thus proving (4.6). Proofs of (4.7) and (4.8) follow immediately by setting $\lambda_i = 0$ and $\lambda_l = 0$, respectively,

in (4.3). To prove (4.9), observe that

$$\mathcal{E}^{\mathrm{grad}}_{ij,l}u\big|_{F_i} = \mathcal{E}^{\mathrm{grad}}_{i,l}u\big|_{F_i} - (\mathcal{E}^{\mathrm{grad}}_{F_j,l}u - \mathcal{E}^{\mathrm{grad}}_{E_{kl},l}u)\big|_{F_i}.$$

We see that this trace is zero by using Proposition 3.1(2) and (4.6). By symmetry of $\mathcal{E}^{\mathrm{grad}}_{ij,l}$, the trace on F_j must also vanish.

Proof of (4.1): This is best seen using the reference tetrahedron expression (4.1). Since u vanishes on the edges of $\hat{K}$ along the x and y axes, we can write, for instance, $u(x,y) = xu_{p-1}(x,y)$ for some polynomials u_{p-1} of degree at most $p-1$. Then

$$\begin{aligned}\mathcal{E}^{\mathrm{grad}}_{\hat{E}_{03}}u &= \frac{2z}{x+y+z}\int_0^1\int_0^{1-s} u(s(x+y+z),t(x+y+z))\,dt\,ds \\ &= 2z\int_0^1\int_0^{1-s} s\,u_{p-1}(s(x+y+z),t(x+y+z))\,dt\,ds,\end{aligned}$$

so the polynomial preservation property follows. ∎

5. Extension of a face bubble. In this section, we consider the problem of extending functions in

$$H^{1/2}_{0,ijk}(F_l) = H^{1/2}(F_l)\cap L^2_{1/\lambda_i}(F_l)\cap L^2_{1/\lambda_j}(F_l)\cap L^2_{1/\lambda_k}(F_l) \tag{5.1}$$

in such a way that the extension is polynomial preserving and has zero traces on all faces other than F_l. Such extensions have applications in the design of optimal face shape functions for high order finite elements. Note that smooth functions that vanish on the boundary edges of F_l (the so called "face bubble" functions) are in $H^{1/2}_{0,ijk}(F_l)$.

The extension operator from $H^{1/2}_{0,ijk}(F_l)$ into $H^1(K)$ is given by

$$\mathcal{E}^{\mathrm{grad}}_{ijk,l}u = \mathcal{E}^{\mathrm{grad}}_{l}u \;-\; \mathcal{E}^{\mathrm{grad}}_{V_l}u \;-\; \sum_{m\in\{i,j,k\}}\big(\mathcal{E}^{\mathrm{grad}}_{F_m,l}u - \mathcal{E}^{\mathrm{grad}}_{E_{ml},l}u\big), \tag{5.2}$$

where $\mathcal{E}^{\mathrm{grad}}_{l}$ is the primary extension operator defined in (2.3), $\mathcal{E}^{\mathrm{grad}}_{F_i,l}$ is the face correction operator defined in (3.2), $\mathcal{E}^{\mathrm{grad}}_{E_{il},l}$ is the edge correction operator defined in (4.3), and

$$\mathcal{E}^{\mathrm{grad}}_{V_l}u\,(\lambda_0,\lambda_1,\lambda_2,\lambda_3) = 2\lambda_l\iint_{F_l} u(\boldsymbol{s})\,d\boldsymbol{s}.$$

This last operator may be thought of as a *vertex correction*, because the right-hand side above equals $\lambda_l\mathcal{E}^{\mathrm{grad}}_{l}u(\boldsymbol{a}_l)$; i.e., it depends only on the value of $\mathcal{E}^{\mathrm{grad}}_{l}u$ at a vertex. It is needed because the edge corrections alter the traces zeroed by the face corrections.

PROPOSITION 5.1. *The operator $\mathcal{E}^{\mathrm{grad}}_{ijk,l}$ satisfies the following:*

1. *$\mathcal{E}^{\mathrm{grad}}_{ijk,l}$ is a continuous map from $H^{1/2}_{0,ijk}(F_l)$ into $H^1(K)$.*
2. *The traces of $\mathcal{E}^{\mathrm{grad}}_{ijk,l}u$ on all faces of the tetrahedron are zero except for the face F_l, where the trace equals u.*
3. *If u is a polynomial of degree at most p, then $\mathcal{E}^{\mathrm{grad}}_{ijk,l}u$ is a polynomial of degree at most p.*

Proof. The continuity property of $\mathcal{E}^{\mathrm{grad}}_{ijk,l}$ follows immediately by the results established in Propositions 3.1(1) and 4.1(1), since the continuity of the vertex correction is obvious.

Let us now prove that the trace of $\mathcal{E}^{\mathrm{grad}}_{ijk,l}u$ on F_i is zero. Observe that by a rearrangement of the terms in (5.2),

$$\begin{aligned}\mathcal{E}^{\mathrm{grad}}_{ijk,l}u|_{F_i} = &\left(\mathcal{E}^{\mathrm{grad}}_{l}u - \mathcal{E}^{\mathrm{grad}}_{F_i,l}u\right)|_{F_i}\\ &- \left(\mathcal{E}^{\mathrm{grad}}_{F_j,l}u - \mathcal{E}^{\mathrm{grad}}_{E_{kl},l}u\right)|_{F_i} - \left(\mathcal{E}^{\mathrm{grad}}_{F_k,l}u - \mathcal{E}^{\mathrm{grad}}_{E_{jl},l}u\right)|_{F_i}\\ &+ \left(\mathcal{E}^{\mathrm{grad}}_{E_{il},l}u - \mathcal{E}^{\mathrm{grad}}_{V_l}u\right)|_{F_i}.\end{aligned}$$

The first term on the right-hand side is zero by Proposition 3.1(2). The second and third are also zero by Proposition 4.1—see (4.6). Furthermore, the last term also vanishes by (4.7). Thus the trace of $\mathcal{E}^{\mathrm{grad}}_{ijk,l}u$ on F_i is zero. The traces on F_j and F_k must also be zero because the expression for $\mathcal{E}^{\mathrm{grad}}_{ijk,l}u$ is symmetric in i,j,k. That the trace on F_l is u, as well as the polynomial preservation property, follows by collecting the results in Propositions 3.1 and 4.1. ∎

6. The total extension operator. We are now in a position to solve the extension problem as posed in the beginning of this paper in (1.5)–(1.7) by combining the primary extension $\mathcal{E}^{\mathrm{grad}}$ with the face, edge, and vertex corrections.

Let u be any function in $H^{1/2}(\partial K)$. We extend this function into K by selecting the faces of K in some order, say F_i, F_j, F_k, and F_l, and defining the following extensions from these faces:

$$\begin{aligned}U_i &= \mathcal{E}^{\mathrm{grad}}_{i}u,\\ U_j &= \mathcal{E}^{\mathrm{grad}}_{i,j}w_j, && \text{where } w_j = (u-U_i)|_{F_j},\\ U_k &= \mathcal{E}^{\mathrm{grad}}_{ij,k}w_k, && \text{where } w_k = (u-U_i-U_j)|_{F_k},\\ U_l &= \mathcal{E}^{\mathrm{grad}}_{ijk,l}w_l, && \text{where } w_l = (u-U_i-U_j-U_k)|_{F_l}.\end{aligned}$$

Here, the operators $\mathcal{E}^{\mathrm{grad}}_{i}$, $\mathcal{E}^{\mathrm{grad}}_{i,j}$, $\mathcal{E}^{\mathrm{grad}}_{ij,k}$, and $\mathcal{E}^{\mathrm{grad}}_{ijk,l}$ are as defined in (2.3), (3.3), (4.4), and (5.2), respectively. The total extension operator is then defined by

$$\mathcal{E}^{\mathrm{grad}}_{K}u = U_i + U_j + U_k + U_l. \tag{6.1}$$

This operator is well defined provided w_j, w_k, and w_l are in the domains of the operators $\mathcal{E}^{\mathrm{grad}}_{i,j}$, $\mathcal{E}^{\mathrm{grad}}_{ij,k}$, and $\mathcal{E}^{\mathrm{grad}}_{ijk,l}$, respectively, which is indeed the case as asserted by the following lemma:

LEMMA 6.1. *There is a constant $C>0$ independent of u such that the functions w_j, w_k, and w_l defined above satisfy*

$$\begin{aligned}\|w_j\|_{H^{1/2}_{0,i}(F_j)} &\le C\|u\|_{H^{1/2}(\partial K)},\\ \|w_k\|_{H^{1/2}_{0,ij}(F_k)} &\le C\|u\|_{H^{1/2}(\partial K)},\\ \|w_l\|_{H^{1/2}_{0,ijk}(F_l)} &\le C\|u\|_{H^{1/2}(\partial K)}.\end{aligned}$$

Collecting all the results we have established in the course of the construction of $\mathcal{E}^{\mathrm{grad}}_{K}$, we have the following theorem:

THEOREM 6.1. *The operator $\mathcal{E}^{\mathrm{grad}}_{K}$ defined by* (6.1) *satisfies*

(1) *the extension property* (1.5),
(2) *the continuity property* (1.6), *and*
(3) *the polynomial preservation property* (1.7).

Proof. All three properties follow from the previous propositions. For example, to prove the continuity property, we use the previously proved inequalities

$$\begin{aligned}
\|\mathcal{E}^{\mathrm{grad}}_{ijk,l} w_l\|_{H^1(K)} &\le C\|w_l\|_{H^{1/2}_{0,ijk}(F_l)}, && \text{by Proposition 5.1(1)},\\
\|\mathcal{E}^{\mathrm{grad}}_{ij,k} w_k\|_{H^1(K)} &\le C\|w_k\|_{H^{1/2}_{0,ij}(F_k)}, && \text{by Proposition 4.1(1)},\\
\|\mathcal{E}^{\mathrm{grad}}_{i,j} w_j\|_{H^1(K)} &\le C\|w_j\|_{H^{1/2}_{0,i}(F_j)}, && \text{by Proposition 3.1(1)},\\
\|\mathcal{E}^{\mathrm{grad}}_{i} u\|_{H^1(K)} &\le C\|u\|_{H^{1/2}(F_l)}, && \text{by Theorem 2.1(1)},
\end{aligned}$$

in

$$\|\mathcal{E}^{\mathrm{grad}}_{K} u\|_{H^1(K)} \le \|\mathcal{E}^{\mathrm{grad}}_{i} u\|_{H^1(K)} + \|\mathcal{E}^{\mathrm{grad}}_{i,j} w_j\|_{H^1(K)} + \|\mathcal{E}^{\mathrm{grad}}_{ij,k} w_k\|_{H^1(K)} + \|\mathcal{E}^{\mathrm{grad}}_{ijk,l} w_l\|_{H^1(K)},$$

and complete the estimate using Lemma 6.1. ∎

Appendix A. Proofs of the lemmas.

In this section, we prove all the previously stated lemmas in the order in which they appeared in sections 2–5. Before we start proving these lemmas, let us begin with some preliminary results which will turn useful in the proofs. These results are often stated in more generality than we need them in this paper, because we will need the general forms in later parts of this series.

There are two kind of operators that pervade the proofs, namely, the averaging type and the interpolating type. We first collect the continuity properties of some averaging type operators in Lemmas A.1 and A.2. The interpolatory operators are considered next in Lemmas A.3 and A.4, after which we begin proving the lemmas of the previous sections. We start with operators that map functions on the x-y face to functions on the y-z face.

LEMMA A.1 (Face-to-face maps). *Let $\theta(s,t)$ be a function in $C(\bar{\hat{F}})$. Define the the following maps for smooth $u(x,y)$:*

$$\begin{aligned}
u(x,y) &\longmapsto A^\theta_1 u\,(y,z) := \int_0^1 \theta(s,0)\; u(sz,y)\; ds,\\
u(x,y) &\longmapsto A^\theta_2 u\,(y,z) := \int_0^1 \theta(s,1-s)\; u(sz,y+(1-s)z)\; ds,\\
u(x,y) &\longmapsto A^\theta_3 u\,(y,z) := 2\int_0^1\int_0^{1-s} \theta(s,t)\; u(sz,y+tz)\; dt\, ds.
\end{aligned}$$

Then the following continuity properties hold:

1. *The maps A^θ_1, A^θ_2, and A^θ_3 extend to continuous operators from $L^2_{1/x}(\hat{F}_3)$ into $L^2_{1/z}(\hat{F}_1)$.*
2. *If, in addition, $\theta(s,t)$ is in $C^1(\bar{\hat{F}})$, then A^θ_3 is continuous under the a stronger norm, namely,*

$$\|A^\theta_3 u\|_{L^2_{1/z}(\hat{F}_1)\cap H^1_z(\hat{F}_1)} \le C\|u\|_{L^2_{1/x}(\hat{F}_3)}. \tag{A.1}$$

Proof. Because of the well-known [20] density of $C^\infty(\bar{\hat{F}}_3)\cap L^2_{1/x}(\hat{F}_3)$ in $L^2_{1/x}(\hat{F}_3)$, to prove item (1), it suffices to prove that there is a constant $C>0$ such that

$$\|A^\theta_i u\|_{L^2_{1/z}(\hat{F}_3)} \le C\|u\|_{L^2_{1/x}(\hat{F}_1)}$$

for all smooth functions $u(x,y)$.

Let us consider the operator A_1^θ acting on a smooth function u. We start with an application of Cauchy–Schwarz inequality to bound the required norm:

$$\begin{aligned}\|A_1^\theta u\|^2_{L^2_{1/z}(\hat{F}_1)} &= \int_0^1\int_0^{1-y} \frac{1}{z}\left|\int_0^1 \theta(s,0)\ u(sz,y)\ ds\right|^2 dz\ dy\\ &\le \int_0^1\int_0^{1-y} \frac{1}{z}\left(\int_0^1 |\theta(s,0)|^2 ds\right)\left(\int_0^1 |u(sz,y)|^2 ds\right) dz\ dy,\end{aligned}$$

by Cauchy–Schwarz inequality. Setting $C_\theta = \int_0^1 |\theta(s,0)|^2 ds$, we continue, and apply Fubini's theorem on a tetrahedral region, as follows.

$$\begin{aligned}\|A_1^\theta u\|^2_{L^2_{1/z}(\hat{F}_1)} &= C_\theta^2 \int_0^1\int_0^{1-y} \frac{1}{z}\int_0^z |u(x,y)|^2\ \frac{dx}{z}\ dz\ dy,\\ &= C_\theta^2 \int_0^1\int_0^{1-y} \left(\int_x^{1-y} \frac{1}{z^2}\ dz\right)|u(x,y)|^2\ dx\ dy,\\ &= C_\theta^2 \int_0^1\int_0^{1-y} \left(\frac{1}{x} - \frac{1}{1-y}\right)|u(x,y)|^2\ dx\ dy,\\ &\le C_\theta^2\ \|u\|^2_{L^2_{1/x}(\hat{F})}.\end{aligned}$$

This establishes the continuity of A_1^θ.

To prove the continuity of the next map,

$$\begin{aligned}\|A_2^\theta u\|^2_{L^2_{1/z}(\hat{F}_1)} &= \int_0^1\int_0^{1-y} \frac{1}{z}\left|\int_0^1 \theta(s,1-s)u(sz,y+(1-s)z)\ ds\right|^2 dz\,dy\\ &= \left(\int_0^1 |\theta(s,1-s)|^2 ds\right)\int_0^1\int_0^{1-y} \frac{1}{z}\int_0^1 |u(sz,y+(1-s)z)|^2 ds\,dz\,dy\\ &= C_\theta^2 \int_0^1\int_0^{1-y}\int_0^z \frac{1}{z}|u(x,y+z-x)|^2 \frac{dx}{z} dz\,dy,\end{aligned}$$

by the substitution $x = zs$. Here $C_\theta = \int_0^1 |\theta(s,1-s)|^2 ds$. Continuing,

$$\begin{aligned}\|A_2^\theta u\|^2_{L^2_{1/z}(\hat{F}_1)} &= C_\theta^2 \int_0^1\int_x^1\int_0^{1-z} \frac{1}{z^2}|u(x,y+z-x)|^2\ dy\,dz\,dx,\\ &\qquad\qquad \text{by Fubini's theorem,}\\ &= C_\theta^2 \int_0^1\int_x^1\int_{z-x}^{1-x} \frac{1}{z^2}|u(x,y')|^2 dy'\,dz\,dx,\\ &\qquad\qquad \text{by substitution } y' = y+z-x,\\ &\le C_\theta^2 \int_0^1\int_x^1\int_0^{1-x} |u(x,y')|^2 dy'\,dz\,dx\\ &= C_\theta^2 \int_0^1\int_0^{1-x} \left(\frac{1}{x} - 1\right)|u(x,y')|^2 dy'\,dx\\ &\le\ C_\theta^2\ \|u\|^2_{L^2_{1/x}(\hat{F})}.\end{aligned}$$

To prove the continuity of A_3^θ,

$$\left\|\frac{1}{2}A_3^1 u\right\|^2_{L^2_{1/z}(\hat{F}_1)} = \int_0^1\int_0^{1-y}\frac{1}{z}\left(\int_0^1\int_0^{1-s}\theta(s,t)\ u(sz,y+tz)\ \ dt\,ds\right)^2 dz\,dy$$
$$\leq \|\theta\|^2_{L^2(\hat{F})}\int_0^1\int_0^{1-y}\frac{1}{z}\left(\int_0^1\int_0^{1-s}|u(sz,y+tz)|^2\ dt\,ds\right)dz\,dy,$$

where we have applied the Cauchy–Schwarz inequality. Next, we change variables, and then interchange the order of integration, carefully considering the variable integration limits:

$$\left\|\frac{1}{2}A_3^1 u\right\|^2_{L^2_{1/z}(\hat{F}_1)} \leq \int_0^1\int_0^{1-y}\frac{1}{z^3}\int_0^z\int_y^{y+z-x'}|u(x',y')|^2\ \ dy'dx'\ \ dz\,dy$$
$$= \int_0^1\int_0^{1-y'}|u(x',y')|^2\int_0^{y'}\int_{x'+y'-y}^{1-y} z^{-3}\,dz\,dy\ \ dx'\,dy',$$

(In the last step we have applied Fubini's theorem to a four-dimensional region.) Now the inner z and y integrals can be evaluated and estimated by

$$2\int_0^{y'}\int_{x'+y'-y}^{1-y} z^{-3}\,dz\,dy = \int_0^{y'}\left((x'+y'-y)^{-2}-(1-y)^{-2}\right)dy$$
$$= -\frac{1}{1-y'}+\frac{1}{x'}+1-\frac{1}{x'+y'}$$
$$\leq \frac{1}{x'}.$$

Hence the continuity of A_3^θ from $L^2_{1/x}(\hat{F}_3)$ into $L^2_{1/z}(\hat{F}_1)$ follows.

It now only remains to prove (A.1) under the stronger assumption on θ. This is a consequence of the following two identities.

$$\partial_y\left(A_3^\theta u\right) = -\frac{1}{z}A_3^{\partial_t\theta}u+\frac{2}{z}\left(A_1^\theta u - A_2^\theta u\right), \tag{A.2}$$

$$\partial_z\left(A_3^\theta u\right) = -\frac{1}{z}A_3^{(s\partial_s\theta+t\partial_t\theta)}u+\frac{2}{z}\left(A_2^\theta u - A_3^\theta u\right). \tag{A.3}$$

These identities follow by variable changes and differentiation; e.g., to prove the first one,

$$\partial_y A_3^\theta v = \frac{2}{z^2}\frac{\partial}{\partial y}\left(\int_0^x\int_y^{y+z-x'}\theta\left(\frac{x'}{z},\frac{y'-y}{z}\right)\ u(x',y')\ dy'\,dx'\right)$$
$$= -\frac{1}{z}\int_0^x\int_y^{y+z-x'}\frac{\partial\theta}{\partial t}\left(\frac{x'}{z},\frac{y'-y}{z}\right)\ u(x',y')\ dy'\,dx'$$
$$+\frac{2}{z^2}\int_0^x\theta\left(\frac{x'}{z},\frac{y'-y}{z}\right)\ u(x',y+z-x')\,dx' - \frac{2}{z^2}\int_0^x\theta\left(\frac{x'}{z},0\right)\ v(x',y)\,dx'$$
$$= -\frac{1}{z}A_3^{\partial_t\theta}u+\frac{2}{z}\int_0^1\theta(s,1-s)\,u(sz,y+(1-s)z)\,ds-\frac{2}{z}\int_0^1\theta(s,0)\,u(sz,y)\,ds,$$

which is (A.2). Equation (A.3) is proved similarly. From (A.2) and (A.3), it is clear that if θ satisfies the additional smoothness assumptions, then we can apply continuity results already proved above to obtain (A.1). ∎

LEMMA A.2 (Face-to-edge maps). *Let $\theta(s,t)$ be a function in $C(\bar{\hat{F}})$. Define the the following maps for smooth $u(x,y)$.*

$$\begin{aligned} u(x,y) &\longmapsto \quad B_1^\theta u\,(z) := \int_0^1 \theta(s,1-s)\; u(sz,(1-s)z)\; ds, \\ u(x,y) &\longmapsto \quad B_2^\theta u\,(z) := 2\int_0^1\int_0^{1-s} \theta(s,t) u(sz,tz)\; dt\, ds. \end{aligned}$$

Then the following continuity properties hold:

1. *B_1^θ and B_2^θ extend to continuous operators from $L^2_{1/x}(\hat{F}_3)\cap L^2_{1/y}(\hat{F}_3)$ into $L^2(\hat{E}_{03})$.*
2. *If, in addition, $\theta(s,t)$ is in $C^1(\bar{\hat{F}})$, then B_2^θ is continuous in a stronger norm:*

$$\|B_2^\theta u\|_{H^2_{z^2}(\hat{E}_{03}\cap L^2(\hat{E}_{03})} \le C\|u\|_{L^2_{1/x}(\hat{F}_3)\cap L^2_{1/y}(\hat{F}_3)}. \tag{A.4}$$

Proof. As in Lemma A.1, it suffices to consider smooth $u(x,y)$ because of density. To prove the continuity of B_1^θ, set $C_\theta = \int_0^1 |\theta(s,1-s)|^2 ds$ and observe that

$$\begin{aligned} \left\|B_1^\theta u\right\|^2_{L^2(\hat{E}_{03})} &= \int_0^1 \left| \int_0^1 \theta(s,1-s)\; u(sz,(1-s)z)\; ds\right|^2 dz, \\ &\le C_\theta^2 \int_0^1\int_0^1 |u(sz,(1-s)z)|^2\, ds\, dz, \\ &= C_\theta^2 \int_0^1\int_0^z \frac{1}{z}|u(x',z-x')|^2\, dx'\, dz, \\ &= C_\theta^2 \int_0^1\int_0^{1-x} \frac{1}{x+y}|u(x,y)|^2\, dy\, dx. \end{aligned}$$

The result now follows from

$$\frac{1}{x+y} \le \frac{1}{2}\left(\frac{1}{x}+\frac{1}{y}\right).$$

The continuity of B_2 is proved as follows:

$$\begin{aligned} \left\|\frac{1}{2}B_2^\theta u\right\|^2_{L^2(\hat{E}_{03})} &= \int_0^1 \left| \int_0^1\int_0^{1-s} \theta(s,t)\; u(sz,tz)\; dt\, ds\right|^2 dz \\ &\le \|\theta\|^2_{L^2(\hat{F})} \int_0^1\int_0^1\int_0^{1-s} |u(sz,tz)|^2\; dt\, ds\, dz \\ &= \|\theta\|^2_{L^2(\hat{F})} \int_0^1\int_0^z\int_0^{z-x'} \frac{1}{z^2}|u(x',y')|^2\; dy'\, dx'\, dz. \end{aligned}$$

Now we apply Fubini's theorem over a tetrahedron to get

$$\begin{aligned} \left\|\frac{1}{2}B_2^\theta u\right\|^2_{L^2(\hat{E}_{03})} &= \int_0^1\int_0^{1-x'} |u(x',y')|^2 \int_{x'+y'}^1 \frac{1}{z^2}\; dz\;\; dy'\, dx' \\ &= \int_0^1\int_0^{1-x'} \left(\frac{1}{x'+y'}-1\right)|u(x',y')|^2\; dz\, dy'\, dx', \end{aligned}$$

from which the result follows since $\frac{1}{x'+y'}-1 \le \frac{1}{2}(\frac{1}{x'}+\frac{1}{y'})$.

It now only remains to prove (A.4). Differentiating,

$$\frac{d}{dz}B_2^\theta u = \frac{d}{dz}\left(\frac{2}{z^2}\int_0^z\int_0^{z-x'} u(x',y')\ dy'\,dx'\right)$$
$$= -\frac{2}{z}B_2^\theta u - \frac{1}{z}B_2^{s\partial_s\theta+t\partial_t\theta}u + \frac{2}{z^2}\int_0^z \left(\frac{x'}{z}, z-x'z\right)\ u(x',z-x')\,dx'.$$

In other words,

$$\frac{d}{dz}B_2^\theta u = -\frac{1}{z}B_2^{s\partial_s\theta+t\partial_t\theta}u + \frac{2}{z}\left(B_1^\theta u - B_2^\theta u\right). \tag{A.5}$$

Hence (A.4) follows from the already proved item (1). ∎

LEMMA A.3 (Face-to-tetrahedron maps). *Let $\theta(x,y,z)$ be a function in $L^\infty(\hat{K})$. The following map, defined for smooth $\phi(y,z)$,*

$$\phi(y,z) \longmapsto J_\theta\phi\,(x,y,z) := \theta(x,y,z)\,\phi(y,x+z)$$

extends to a continuous operator from $L^2_z(\hat{F}_1)$ into $L^2(\hat{K})$.

Proof. The proof is immediate once we make the change of variable $(x',y',z') = (x,y,x+z)$. Indeed,

$$\|J_\theta\phi\|^2_{L^2(\hat{K})} = \int_0^1\int_0^{1-x}\int_0^{1-x-z} |\theta(x,y,z)\,\phi(y,x+z)|^2\ dy\,dz\,dx$$
$$= \int_0^1\int_0^{1-y'}\int_0^{z'} \left|\theta(x',y',z'-x')|^2\,|\phi(y',z')\right|^2\ dx'\,dz'dy'$$
$$\le \int_0^1\int_0^{1-y'}\left(\int_0^{z'}\|\theta\|^2_{L^\infty(\hat{K})}dx'\right)|\phi(y',z')|^2\,dz'dy'$$
$$\le \|\theta\|^2_{L^\infty(\hat{K})}\|\phi\|^2_{L_z(\hat{F}_1)}. \qquad ∎$$

LEMMA A.4 (Edge-to-tetrahedron maps). *Let $\theta(x,y,z)$ be a function in $L^\infty(\hat{K})$. The following map, defined for smooth $\psi(z)$,*

$$\psi(z) \longmapsto L_\theta\psi\,(x,y,z) := \theta(x,y,z)\,\psi(x+y+z)$$

extends to continuous operators from $L^2_{z^2}(\hat{E}_{03})$ into $L^2(\hat{K})$.

Proof. Making the variable change $(x',y',z') = (x+y,y,x+y+z)$, we find that

$$\|L_\theta\psi\|^2_{L^2(\hat{K})} = \int_0^1\int_0^{z'}\int_0^{y'} |\theta(x'-y',y',z-x')\,\psi(z')|^2\ dx'\,dy'\,dz'$$
$$\le \int_0^1 \|\theta\|^2_{L^\infty(\hat{K})}\frac{(z')^2}{2}\,|\psi(z')|^2\ dz',$$

thus proving the continuity of L_θ. ∎

With the help of the above results, we now start proving all the lemmas that appeared in the previous sections.

Proof of Lemma 3.2. First, we need to prove that

$$\mathcal{R}^{\text{grad}}:\ L^2_{1/x}\left(\hat{F}_3\right) \longmapsto L^2_{1/z}\left(\hat{F}_1\right)\cap H^1_z\left(\hat{F}_1\right)$$

is continuous. But

$$\mathcal{R}^{\mathrm{grad}} u = A_3^\theta u,$$

with $\theta(s,t) \equiv 1$. Hence, we obtain the required continuity properties from Lemma A.1, specifically the estimate in (A.1).

Similarly, the continuity of $\mathcal{L}^{\mathrm{grad}} : L^2_{1/x}(\hat{F}_3) \cap L^2_{1/y}(\hat{F}_3) \longmapsto L^2(\hat{E}_{03}) \cap H^1_{z^2}(\hat{E}_{03})$ follows from Lemma A.2 because

$$\mathcal{L}^{\mathrm{grad}} u = B_2^\theta u,$$

with $\theta(s,t) \equiv 1$ (see (A.4)). ∎

Proof of Lemma 3.1. It suffices to prove that

$$\|\mathcal{B}_{\hat{F}_1}\phi\|_{H^1(\hat{K})} \le C\left(\|\phi\|_{L^2_{1/z}(\hat{F}_1)} + \|\phi\|_{H^1_z(\hat{F}_1)}\right) \tag{A.6}$$

for all $\phi(y,z)$ in $C^\infty(\bar{\hat{F}}_1)$ that vanish in a neighborhood of the y-axis, because it is proven in [16, Lemma 3.1] that such functions are dense in $L^2_{1/z}(\hat{F}_1) \cap H^1_z(\hat{F}_1)$. For such a ϕ, the function $\mathcal{B}_{\hat{F}_1}\phi$ is obviously in $H^1(\hat{K})$ and vanishes on $\hat{F}_3$. Hence, the Poincaré inequality gives a constant $\hat{C}$ depending only on $\hat{K}$ such that

$$\|\mathcal{B}_{\hat{F}_1}\phi\|_{L^2(\hat{K})} \le \hat{C}\|\,\mathbf{grad}\left(\mathcal{B}_{\hat{F}_1}\phi\right)\|_{L^2(\hat{K})}.$$

Moreover,

$$\begin{aligned}
\mathbf{grad}(\mathcal{B}_{\hat{F}_1}\phi) &= \frac{z}{x+z}\,\mathbf{grad}\Big(\phi(y,x+z)\Big) + \phi(y,x+z)\,\mathbf{grad}\left(\frac{z}{x+z}\right)\\
&= \frac{z}{x+z}\begin{pmatrix}\partial_z\phi(y,x+z)\\ \partial_y\phi(y,x+z)\\ \partial_z\phi(y,x+z)\end{pmatrix} + \frac{1}{(x+z)^2}\begin{pmatrix}-z\\ 0\\ x\end{pmatrix}\phi(y,x+z)\\
&= \begin{pmatrix}J_{\theta_2}(\partial_z\phi)\\ J_{\theta_2}(\partial_y\phi)\\ J_{\theta_2}(\partial_z\phi)\end{pmatrix} + \begin{pmatrix}-J_{\theta_2}(\phi/z)\\ 0\\ J_{\theta_1}(\phi/z)\end{pmatrix},
\end{aligned}$$

with

$$\theta_1 = \frac{x}{x+z} \quad \text{and} \quad \theta_2 = \frac{z}{x+z}.$$

Thus (A.6) follows from the continuity properties of J_θ established in Lemma A.3 (noting that both θ_1 and θ_2 take only values between 0 and 1). ∎

Proof of Lemma 4.1. Since $\mathcal{B}_{\hat{E}_{03}}\psi = L_{\theta_2}\psi$ with

$$\theta_2 = \frac{z}{x+y+z},$$

and since

$$\begin{aligned}
\mathbf{grad}(\mathcal{B}_{\hat{E}_{03}}\psi) &= \frac{1}{(x+y+z)^2}\begin{pmatrix}-z\\ -z\\ x+y\end{pmatrix}\psi(x+y+z) + \frac{z}{x+y+z}\begin{pmatrix}1\\ 1\\ 1\end{pmatrix}\psi'(x+y+z)\\
&= \begin{pmatrix}-L_{\theta_2}(\psi/z)\\ -L_{\theta_2}(\psi/z)\\ L_{\theta_1}(\psi/z)\end{pmatrix} + \begin{pmatrix}L_{\theta_2}(\psi')\\ L_{\theta_2}(\psi')\\ L_{\theta_2}(\psi')\end{pmatrix},
\end{aligned}$$

with

$$\theta_1 = \frac{x+y}{x+y+z},$$

the required estimate follows immediately from the continuity properties of L_θ established in Lemma A.4. ∎

Proof of Lemma 6.1. By a well-known characterization of $H^{1/2}$ norms on polyhedral surfaces (see [6, Theorem 2.5], [17, Lemma 1.3.2.6], or [24, Lemma 1]), there is a constant $C > 0$ such that for all v in $H^{1/2}(\partial K)$,

$$\|v\|^2_{L^2_{1/\lambda_i}(F_j)} \le C\|v\|^2_{H^{1/2}(F_i\cup F_j)} \qquad \text{if } v|_{F_i} = 0, \tag{A.7}$$

$$\|v\|^2_{L^2_{1/\lambda_i}(F_k)} + \|v\|^2_{L^2_{1/\lambda_j}(F_k)} \le C\|v\|^2_{H^{1/2}(F_i\cup F_j\cup F_k)} \qquad \text{if } v|_{F_i\cup F_j} = 0, \tag{A.8}$$

$$\|v\|^2_{L^2_{1/\lambda_i}(F_l)} + \|v\|^2_{L^2_{1/\lambda_j}(F_l)} + \|v\|^2_{L^2_{1/\lambda_k}(F_l)} \le C\|v\|^2_{H^{1/2}(\partial K)} \qquad \text{if } v|_{F_i\cup F_j\cup F_k} = 0. \tag{A.9}$$

Applying the first inequality above to the function $v = u - \mathcal{E}_i^{\mathrm{grad}} u|_{\partial K}$, which vanishes on F_i (by Theorem 2.1(2)), and coincides with the given function w_j on F_j, we find that

$$\begin{aligned}
\|w_j\|^2_{H^{1/2}_{0,i}(F_j)} &= \|w_j\|^2_{H^{1/2}(F_j)} + \|w_j\|^2_{L^2_{1/\lambda_i}(F_j)} \le C\|v\|^2_{H^{1/2}(F_i\cup F_j)} \\
&\le C\|u\|^2_{H^{1/2}(F_i\cup F_j)} + C\|\mathcal{E}_i^{\mathrm{grad}} u\|^2_{H^{1/2}(F_i\cup F_j)} \\
&\le C\|u\|^2_{H^{1/2}(F_i\cup F_j)} + C\|\mathcal{E}_i^{\mathrm{grad}} u\|^2_{H^1(K)} \qquad \text{(by trace theorem)} \\
&\le C\|u\|^2_{H^{1/2}(F_i\cup F_j)} + C\|u\|^2_{H^{1/2}(F_i)} \qquad \text{(by Theorem 2.1(1))},
\end{aligned}$$

which proves the first inequality of the lemma.

To prove the bound for $w_k = (u - U_i - U_j)|_{F_k}$, we apply (A.8) to the function

$$v = (I - \mathcal{E}_{i,j}^{\mathrm{grad}} R_j)(I - \mathcal{E}_i^{\mathrm{grad}})u,$$

where R_j denotes the restriction to face F_j (and I is the identity). Clearly, $w_k = R_k v$, and by Proposition 3.1(2), v vanishes on $F_i \cup F_j$, so (A.8) gives

$$\|w_k\|_{H^{1/2}_{0,ij}(F_k)} \le C\|v\|_{H^{1/2}(F_i\cup F_j\cup F_k)}.$$

The required bound for w_k follows as in the w_j case by using the trace theorem and the continuity of the operators $\mathcal{E}_{i,j}^{\mathrm{grad}}$ and $\mathcal{E}_i^{\mathrm{grad}}$. The final estimate for w_l is proved similarly, but using

$$v = \left(I - \mathcal{E}_{ij,k}^{\mathrm{grad}} R_k\right)\left(I - \mathcal{E}_{i,j}^{\mathrm{grad}} R_j\right)\left(I - \mathcal{E}_i^{\mathrm{grad}}\right)u$$

and (A.9). ∎

Appendix B. Two techniques to study the primary extension.

The starting point for constructing the tetrahedral extensions in any of the $H^1(K)$, $\boldsymbol{H}(\mathbf{curl})$, or $\boldsymbol{H}(\mathrm{div})$ case is the study of "primary extensions," i.e., extensions

into $H^1(\hat{K})$, $\boldsymbol{H}(\mathbf{curl}\,,\hat{K})$, and $\boldsymbol{H}(\mathrm{div},\hat{K})$ from one face:

$$\mathcal{E}^{\mathrm{grad}}u\,(x,y,z) = \frac{2}{z^2}\int_x^{x+z}\int_y^{x+y+z-\widetilde{x}} u(\widetilde{x},\widetilde{y})\,d\widetilde{y}\,d\widetilde{x}, \tag{B.1}$$

$$\mathcal{E}^{\mathrm{curl}}\boldsymbol{v}\,(x,y,z) = \frac{2}{z^3}\int_x^{x+z}\int_y^{x+y+z-\widetilde{x}} \begin{pmatrix} z & 0 \\ 0 & z \\ \widetilde{x}-x & \widetilde{y}-y \end{pmatrix} \boldsymbol{v}(\widetilde{x},\widetilde{y})\,d\widetilde{y}\,d\widetilde{x}, \tag{B.2}$$

$$\mathcal{E}^{\mathrm{div}}w\,(x,y,z) = \frac{2}{z^3}\int_x^{x+z}\int_y^{x+y+z-\widetilde{x}} \begin{pmatrix} \widetilde{x}-x \\ \widetilde{y}-y \\ -z \end{pmatrix} w(\widetilde{x},\widetilde{y})\,d\widetilde{y}\,d\widetilde{x}. \tag{B.3}$$

We have already seen the primary extension appearing in (B.1) in the previous sections. The other two operators above are the primary extensions for the $\boldsymbol{H}(\mathbf{curl}\,)$ and $\boldsymbol{H}(\mathrm{div})$ case (which will be discussed in more detail in Parts II [11] and III [12]).

The purpose of this appendix is to establish the continuity property of $\mathcal{E}^{\mathrm{grad}}$ stated in Theorem 2.1. But since the arguments required to prove the continuity of the other two operators $\mathcal{E}^{\mathrm{curl}}$ and $\mathcal{E}^{\mathrm{div}}$ are very similar, we establish the continuity of all of these primary extensions together here. We can exhibit two techniques to establish such continuity properties, one using the Fourier transform, and the other using Peetre's K-functional. The latter approach will be shown in Part II. In this appendix, we outline the more standard Fourier transform technique.

The Fourier transform technique proceeds by viewing the above operators as extending functions given on $\mathbb{R}^2$ (the entire x-y plane) into the infinite slab $\mathbb{R}^2 \times (0,1) \equiv \{(x,y,z) : 0 < z < 1\}$. Define the Fourier transform on $\mathbb{R}^2$ as usual by

$$\mathcal{F}u\,(\boldsymbol{\omega}) \equiv \widehat{u}(\boldsymbol{\omega}) := \iint_{\mathbb{R}^2} e^{-i2\pi\boldsymbol{\omega}\cdot\boldsymbol{x}}u(\boldsymbol{x})\,d\boldsymbol{x}. \tag{B.4}$$

We begin by noting some properties of the Fourier transform of an indicator function.

LEMMA B.1. *Let $\chi_{\triangle}$ denote the characteristic function of the unit triangle $\triangle = \{(x,y)\in\mathbb{R}^2 : x>0, y>0, x+y<1\}$, and $\eta(x,y)$ be a linear function. Then, for all $t>0$ and all unit vectors $\boldsymbol{\omega}$,*

(1) *$\widehat{\chi_{\triangle}\eta}\,(t\boldsymbol{\omega})$ is bounded uniformly in t and $\boldsymbol{\omega}$, and*

(2) *the integral $\displaystyle\int_0^\infty |\widehat{\chi_{\triangle}\eta}\,(t\boldsymbol{\omega})|^2\,dt$ is finite and bounded uniformly in $\boldsymbol{\omega}$.*

Proof. We begin with the case of a constant function $\eta\equiv 1$. For this case the first assertion is immediate as the exponential in (B.4) is uniformly bounded by one:

$$|\widehat{\chi_{\triangle}\eta}(\boldsymbol{\omega})| = |\widehat{\chi}_{\triangle}(\boldsymbol{\omega})| \le \iint_{\triangle} \left|e^{-i2\pi t\boldsymbol{\omega}\cdot\boldsymbol{x}}\right| d\boldsymbol{x} \le \frac{1}{2}. \tag{B.5}$$

To prove the second assertion, it will be convenient to work with a rotated system of coordinates $\widetilde{x},\widetilde{y}$ in such a way that the $\widetilde{x}$-axis passes through the point $\boldsymbol{\omega}$ (see Figure 5). We consider only the case when $\boldsymbol{\omega}$ is in the first quadrant as the other cases are similar. Then

$$\widehat{\chi_{\triangle}\eta}(\boldsymbol{\omega}) = \int_{\mathbb{R}}\int_{\mathbb{R}} e^{-i2\pi|\boldsymbol{\omega}|\widetilde{x}}\chi_{\triangle}(\widetilde{x},\widetilde{y})\,d\widetilde{x}\,d\widetilde{y} = \int_{\widetilde{y}_1}^{\widetilde{y}_2}\left[\frac{e^{-i2\pi|\boldsymbol{\omega}|\widetilde{x}}}{-i2\pi|\boldsymbol{\omega}|}\right]_{\widetilde{x}=a(\widetilde{y})}^{\widetilde{x}=b(\widetilde{y})} d\widetilde{y}\,,$$

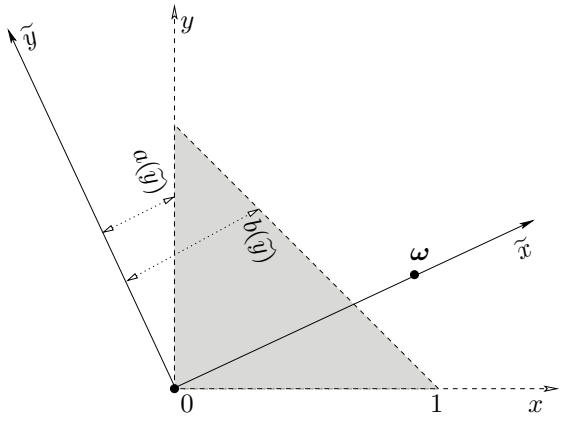


FIG. 5. *Rotated system of coordinates* $(\widetilde{x}, \widetilde{y})$ *in Lemma* B.1.

where $a(\widetilde{y})$ and $b(\widetilde{y})$ are as in Figure 5, $\widetilde{y}_1$ and $\widetilde{y}_2$ are such that the interval $[\widetilde{y}_1, \widetilde{y}_2]$ is the projection of $\triangle$ on the $\widetilde{y}$-axis, and the notation $[g(x)]_{x=p}^{x=q}$ denotes the difference $g(q) - g(p)$. Now, for $|\boldsymbol{\omega}| = 1$ and $t > 0$, the second assertion follows from

$$\int_1^\infty |\widehat{\chi}_\triangle(t\boldsymbol{\omega})|^2 \, dt = \int_1^\infty \left| \int_{\widetilde{y}_1}^{\widetilde{y}_2} \left[\frac{e^{-i2\pi t\widetilde{x}}}{-i2\pi t} \right]_{\widetilde{x}=a(\widetilde{y})}^{\widetilde{x}=b(\widetilde{y})} d\widetilde{y} \right|^2 dt$$
$$\leq \int_1^\infty \frac{1}{4\pi^2} \left| \frac{1}{t} \int_{-1}^1 \left[e^{-i2\pi t\widetilde{x}} \right]_{\widetilde{x}=a(\widetilde{y})}^{\widetilde{x}=b(\widetilde{y})} d\widetilde{y} \right|^2 dt$$

and the fact that the inner integral is bounded uniformly in t. Notice that integrability from 0 to 1 follows from the first assertion.

The proofs for $\eta(x, y) = x$ and $\eta(x, y) = y$ are similar. For instance, for $\eta(x, y) = x = (\widetilde{x}, -\widetilde{y})^t \cdot \boldsymbol{\omega}$, we have

$$\widehat{\chi_\triangle x}(t\boldsymbol{\omega}) = \int_{\widetilde{y}_1}^{\widetilde{y}_2} \left(\int_{a(\widetilde{y})}^{b(\widetilde{y})} e^{-i2\pi t\widetilde{x}} \begin{pmatrix} \widetilde{x} \\ -\widetilde{y} \end{pmatrix} \cdot \boldsymbol{\omega} \; d\widetilde{x} \right) d\widetilde{y},$$

so the first assertion follows because the integrand is uniformly bounded. The second assertion also follows because once the inner integral is evaluated, it is immediate that it can be bounded as before by C/t as $t \to \infty$. ∎

Remark B.1. In fact, Lemma B.1 extends to a large class of functions $\eta(x, y)$, including polynomials of arbitrary degree.

LEMMA B.2. *Let* $\eta(s, t)$ *be a linear polynomial. Then the map defined for smooth functions* $u(x, y)$ *on* $\mathbb{R}^2$ *by*

$$u(x, y) \;\longmapsto\; \mathcal{K}_\eta u\,(x, y, z) := \int_0^1 \int_0^{1-s} \eta(s, t)\, u(x + sz, y + tz)\, dt\, ds$$

extends to a continuous operator from $H^{-1/2}(\mathbb{R}^2)$ *into* $L^2\big(\mathbb{R}^2 \times (0, 1)\big)$.

Proof. We shall consider the case $\eta \equiv 1$ first. Notice that for each fixed z, the operator is in the form of a convolution in x and y variables:

$$\mathcal{K}_\eta u(x, y, z) = \frac{1}{z^2} \int_{-z}^0 \int_{-z-s'}^0 u(x - s', y - t')\, dt'\, ds' = \frac{1}{z^2} (\chi_{z\triangledown} * u),$$

where $\chi_{z\nabla}$ is the characteristic function of $\{(x,y) : x<0, y<0, x+y>-z\}$ (the unit triangle scaled by $-z$). In what follows, we assume that $u(x,y)$ is a test function from the Schwartz space (see, e.g., [19, 23]) and, upon establishing the continuity estimate, tacitly complete the proof using the standard density argument.

Applying the Fourier transform (on the x-y plane) and using its standard properties [19]

$$\widehat{\mathcal{K}_\eta u}\,(\boldsymbol{\omega},z) \;=\; \frac{1}{z^2}\,\widehat{\chi}_{z\nabla}(\boldsymbol{\omega})\,\widehat{u}(\boldsymbol{\omega}) \;=\; \widehat{\chi}_{\triangle}(-z\boldsymbol{\omega})\,\widehat{u}(\boldsymbol{\omega}).$$

By Parseval's identity,

$$\begin{aligned}\int_0^1 \iint\limits_{\mathbb{R}^2} |\mathcal{K}_\eta u(x,y,z)|^2\,dx\,dy\,dz &= \int_0^1 \iint\limits_{\mathbb{R}^2} |\widehat{\mathcal{K}_\eta u}(\boldsymbol{\omega},z)|^2\,d\boldsymbol{\omega}\,dz\\ &= \iint\limits_{\mathbb{R}^2}\left(\int_0^1 |\widehat{\chi}_{\triangle}(z\boldsymbol{\omega})|^2\,dz\right)|\widehat{u}(\boldsymbol{\omega})|^2\,d\boldsymbol{\omega}. \qquad \text{(B.6)}\end{aligned}$$

Now we split the integral over $\mathbb{R}^2$ into two integrals, one over the unit disk $D=\{\boldsymbol{\omega}\in\mathbb{R}^2:\ |\boldsymbol{\omega}|<1\}$ and the other over $\mathbb{R}^2\setminus D$, and analyze each separately.

Substituting $z|\boldsymbol{\omega}|=t$ and denoting $\widehat{\boldsymbol{\omega}}=\boldsymbol{\omega}/|\boldsymbol{\omega}|$,

$$\int_0^1 |\widehat{\chi}_{\triangle}(z\boldsymbol{\omega})|^2\,dz = \frac{1}{|\boldsymbol{\omega}|}\int_0^{|\boldsymbol{\omega}|} |\widehat{\chi}_{\triangle}(t\widehat{\boldsymbol{\omega}})|^2\,dt \le \frac{1}{2} \qquad \text{(B.7)}$$

by the first assertion of Lemma B.1 (see (B.5)). By the second assertion of Lemma B.1, we can get another bound for the same term:

$$\int_0^1 |\widehat{\chi}_{\triangle}(z\boldsymbol{\omega})|^2\,dz = \frac{1}{|\boldsymbol{\omega}|}\int_0^{|\boldsymbol{\omega}|} |\widehat{\chi}_{\triangle}(t\widehat{\boldsymbol{\omega}})|^2\,dt \le \frac{1}{|\boldsymbol{\omega}|}\int_0^{\infty} |\widehat{\chi}_{\triangle}(t\widehat{\boldsymbol{\omega}})|^2\,dt \le \frac{c}{|\boldsymbol{\omega}|}. \qquad \text{(B.8)}$$

Using these estimates in (B.6), specifically (B.7) in D and (B.8) in $\mathbb{R}^2\setminus D$,

$$\begin{aligned}\int_0^1 \iint\limits_{\mathbb{R}^2} |\mathcal{K}_\eta u(x,y,z)|^2\,dx\,dy\,dz &\le \iint\limits_{D} \frac{1}{2}|\widehat{u}(\boldsymbol{\omega})|^2\,d\boldsymbol{\omega} + \iint\limits_{\mathbb{R}^2\setminus D} \frac{c}{|\boldsymbol{\omega}|}|\widehat{u}(\boldsymbol{\omega})|^2\,d\boldsymbol{\omega}\\ &\le C\iint\limits_{\mathbb{R}^2}\left(1+|\boldsymbol{\omega}|^2\right)^{-1/2}|\widehat{u}(\boldsymbol{\omega})|^2\,d\boldsymbol{\omega}.\end{aligned}$$

Because of a well-known characterization of Sobolev norms via the Fourier transform [19, 23], the right-hand side above is the square of a norm equivalent to the $H^{-1/2}(\mathbb{R}^2)$ norm, so the proof for the $\eta\equiv 1$ case is finished.

The reasoning for cases $\eta(x,y)=x$ and $\eta(x,y)=y$ is fully analogous. For instance, for the former, we have

$$\mathcal{K}_\eta\,(x,y,z) = \frac{1}{z^2}\left(\frac{x}{z}\chi_{z\nabla} * u\right),$$

so

$$\widehat{\mathcal{K}_\eta u}(\boldsymbol{\omega},z) \;\; = \frac{1}{z^2}\,\widehat{\chi_{z\nabla}x}(\boldsymbol{\omega})\,\widehat{u}(\boldsymbol{\omega}) = \widehat{\chi_{\triangle}x}(z\boldsymbol{\omega})\widehat{u}(\boldsymbol{\omega}).$$

The rest of the estimation is fully analogous utilizing the boundedness properties for the Fourier transform of factor $x\chi_{\triangle}$ proved in Lemma B.1. ∎

THEOREM B.1. *The polynomial extension operators in* (B.1)–(B.3) *extend to continuous operators on the spaces below:*

$$\begin{aligned} \mathcal{E}^{\mathrm{grad}}: &\quad H^{1/2}(\hat{F}) &&\longmapsto &&H^1(\hat{K}) \\ \mathcal{E}^{\mathrm{curl}}: &\quad \boldsymbol{H}^{-1/2}(\mathrm{curl}_\tau, \hat{F}) &&\longmapsto &&\boldsymbol{H}(\mathbf{curl}, \hat{K}) \\ \mathcal{E}^{\mathrm{div}}: &\quad H^{-1/2}(\hat{F}) &&\longmapsto &&\boldsymbol{H}(\mathrm{div}, \hat{K}). \end{aligned}$$

Proof. We quickly sketch the arguments, considering the last operator first.

The $\boldsymbol{H}(\mathrm{div})$ primary extension: According to Lemma B.2, each of the three components of $\mathcal{E}^{\mathrm{div}}$ is a continuous map from $H^{-1/2}(\mathbb{R}^2)$ into $L^2(\mathbb{R}^2 \times (0,1))$. Moreover, $\mathrm{div}(\mathcal{E}^{\mathrm{div}} w) = 0$. Consequently, $\mathcal{E}^{\mathrm{div}}$ maps $H^{-1/2}(\mathbb{R}^2)$ into $\boldsymbol{H}(\mathrm{div}, \mathbb{R}^2 \times (0,1))$. Since there is a continuous extension operator [7] from $H^{-1/2}(\hat{F})$ into $H^{-1/2}(\mathbb{R}^2)$, and since the restriction operator from $\boldsymbol{H}(\mathrm{div}, \mathbb{R}^2 \times (0,1))$ into $\boldsymbol{H}(\mathrm{div}, \hat{K})$ is obviously continuous, the proof of this case is finished.

The $\boldsymbol{H}(\mathbf{curl})$ primary extension: By the same arguments as above, the three components of $\mathcal{E}^{\mathrm{curl}}$ are continuous maps from $H^{-1/2}(\hat{F})$ into $L^2(\hat{K})$. Moreover, by the readily established commutativity property

$$\mathbf{curl}\,(\mathcal{E}^{\mathrm{curl}} \boldsymbol{v}) = \mathcal{E}^{\mathrm{div}}(\mathrm{curl}_\tau\, \boldsymbol{v})$$

and the continuity of $\mathcal{E}^{\mathrm{div}}$ on $H^{-1/2}(\hat{F})$, we find that $\mathcal{E}^{\mathrm{curl}}$ is a continuous map from $\boldsymbol{H}^{-1/2}(\mathbf{curl}, \hat{F})$ into $\boldsymbol{H}(\mathbf{curl}, T)$.

The H^1 primary extension: The surface gradient $\mathbf{grad}_\tau$ is a continuous map [6] from $H^{1/2}(\hat{F})$ into $\boldsymbol{H}^{-1/2}(\hat{F})$. Combining this with the easily established commutativity property

$$\mathbf{grad}(\mathcal{E}^{\mathrm{grad}} u) = \mathcal{E}^{\mathrm{curl}}(\mathbf{grad}_\tau\, u),$$

and the continuity of $\mathcal{E}^{\mathrm{curl}}$, we find that the left-hand side above has its $L^2(\hat{K})$ norm bounded by the $H^{1/2}(\hat{F})$ norm of u. Since square integrability of the gradient $\mathbf{grad}(\mathcal{E}^{\mathrm{grad}} u)$ implies the square integrability of the function $\mathcal{E}^{\mathrm{grad}} u$ (by, e.g., a generalization of Stokes theorem [14, Theorem I.2.9]), and since constants are preserved by $\mathcal{E}^{\mathrm{grad}}$, the H^1 extension operator $\mathcal{E}^{\mathrm{grad}}$ is continuous from $H^{1/2}(\hat{F})$ into $H^1(\hat{K})$. ∎

REFERENCES

[1] M. AINSWORTH AND L. DEMKOWICZ, *Explicit polynomial preserving trace liftings on a triangle*, Math. Nachr., to appear. See also *ICES Report* 03-47.

[2] I. BABUŠKA, A. CRAIG, J. MANDEL, AND J. PITKÄRANTA, *Efficient preconditioning for the p-version finite element method in two dimensions*, SIAM J. Numer. Anal., 28 (1991), pp. 624–661.

[3] I. BABUŠKA AND M. SURI, *The h-p version of the finite element method with quasiuniform meshes*, RAIRO Mode'el. Math. Anal. Num'er., 21 (1987), pp. 199–238.

[4] F. B. BELGACEM, *Polynomial extensions of compatible polynomial traces in three dimensions*, Comput. Methods Appl. Mech. Engrg., 116 (1994), pp. 235–241. ICOSAHOM'92 (Montpellier, 1992).

[5] C. BERNARDI, M. DAUGE, AND Y. MADAY, *Polynomials in the sobolev world*, Technical report 07-14, IRMAR, University of Rennes, 2007.

[6] A. BUFFA AND P. CIARLET, JR., *On traces for functional spaces related to Maxwell's equations. I. An integration by parts formula in Lipschitz polyhedra*, Math. Methods Appl. Sci., 24 (2001), pp. 9–30.

[7] M. Cessenat, *Mathematical methods in electromagnetism. Linear theory and applications*, Ser. Adv. Math. Appl. Sci. 41, World Scientific Publishing Co. Pte. Ltd., Singapore, 1996.

[8] M. Costabel, M. Dauge, and L. Demkowicz, *Polynomial extension operators for H^1, $\boldsymbol{H}(\mathbf{curl})$, and $\boldsymbol{H}(\mathrm{div})$ spaces on a cube*, Technical report, IRMAR, Rennes, 2007.

[9] L. Demkowicz and I. Babuška, *Optimal p interpolation error estimates for edge finite elements of variable order in two dimensions*, SIAM J. Numer. Anal., 41 (2003), pp. 1195–1208 (electronic).

[10] L. Demkowicz and A. Buffa, *H^1, $H(\mathrm{curl})$ and $H(\mathrm{div})$-conforming projection-based interpolation in three dimensions. Quasi-optimal p-interpolation estimates*, Comput. Methods Appl. Mech. Engrg., 194 (2005), pp. 267–296.

[11] L. Demkowicz, J. Gopalakrishnan, and J. Schöberl, *Polynomial extension operators. Part* II, SIAM J. Numer. Anal., submitted.

[12] L. Demkowicz, J. Gopalakrishnan, and J. Schöberl, *Polynomial extension operators. Part* III, in preparation.

[13] L. Demkowicz, *Polynomial Exact Sequences and Projection-based Interpolation with Applications to Maxwell Equations*, in Mixed Finite Elements, Compatibility Conditions and Applications, D. Boffi and L. Gastaldi, eds., Lecture Notes in Mathematics, Springer-Verlag, 2007. See also ICES Report 06-12.

[14] V. Girault and P.-A. Raviart, *Finite Element Methods for Navier-Stokes Equations*, in Springer series in Computational Mathematics 5, Springer-Verlag, New York, 1986.

[15] J. Gopalakrishnan and L. F. Demkowicz, *Quasioptimality of some spectral mixed methods*, J. Comput. Appl. Math., 167 (2004), pp. 163–182.

[16] J. Gopalakrishnan and J. E. Pasciak, *The convergence of V-cycle multigrid algorithms for axisymmetric Laplace and Maxwell equations*, Math. Comp., 75 (2006), pp. 1697–1719.

[17] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, in Monographs and Studies in Mathematics 24, Pitman Advanced Publishing Program, Marshfield, Massachusetts, 1985.

[18] N. Heuer and F. Leydecker, *An extension theorem for polynomials on triangles*, Technical report, BICOM, Brunel University, Uxbridge, UB8 3PH, England, 2005.

[19] S. Kesavan, *Topics in Functional Analysis and Applications*, Wiley Eastern Limited, Bombay, 1989.

[20] A. Kufner, *Weighted Sobolev Spaces*, A Wiley-Interscience Publication, John Wiley & Sons Inc., New York, 1985. Translated from the Czech.

[21] Y. Maday, *Relèvements de traces polynomiales et interpolations hilbertiennes entre espaces de polynômes*, C. R. Acad. Sci. Paris. Sér. I Math., 309 (1989), pp. 463–468.

[22] Y. Maday, *Analysis of spectral projectors in one-dimensional domains*, Math. Comp., 55 (1990), pp. 537–562.

[23] W. McLean, *Strongly Elliptic Systems and Boundary Integral Equations*, Cambridge University Press, 2000.

[24] R. Muñoz-Sola, *Polynomial liftings on a tetrahedron and applications to the h-p version of the finite element method in three dimensions*, SIAM J. Numer. Anal., 34 (1997), pp. 282–314.

[25] J. Schöberl, J. Melenk, C. Pechstein, and S. Zaglmayr, *Additive Schwarz preconditioning for p-version triangular and tetrahedral finite elements*, IMA J. Numer. Anal., 28 (2008), pp. 1–24.

SIAM J. NUMER. ANAL.
Vol. 46, No. 6, pp. 3032–3070
© 2008 Society for Industrial and Applied Mathematics

FINITE VOLUME METHOD FOR 2D LINEAR AND NONLINEAR ELLIPTIC PROBLEMS WITH DISCONTINUITIES*

FRANCK BOYER† AND FLORENCE HUBERT†

Abstract. In this paper we study the approximation of solutions to linear and nonlinear elliptic problems with discontinuous coefficients in the discrete duality finite volume (DDFV) framework. This family of schemes allows very general meshes and inherits the main properties of the continuous problem. In order to take into account the discontinuities and to prevent consistency defect in the scheme, we propose to modify the definition of the numerical fluxes on the edges of the mesh where the discontinuity occurs. We first illustrate our approach by the study of the 1D situation. Then, we show how to design our new scheme, called m-DDFV, and we propose its analysis. We also describe an iterative solver, whose convergence is proved, which can be used to solve the nonlinear discrete equations defining the finite volume scheme. Finally, we provide numerical results which confirm that the m-DDFV scheme significantly improves the convergence rate of the usual DDFV method for both linear and nonlinear problems.

Key words. finite volume schemes, discontinuous coefficients, nonlinear elliptic problems

AMS subject classifications. 35J65, 65N15, 74S10

DOI. 10.1137/060666196

1. Introduction.

In this paper, we are concerned with the finite volume approximation of solutions to the following nonlinear diffusion problem with homogeneous Dirichlet boundary conditions:

$$\begin{cases} -\mathrm{div}(\varphi(z,\nabla u_e(z))) = f(z), & \text{in } \Omega, \\ u_e = 0, & \text{on } \partial\Omega, \end{cases} \tag{1.1}$$

where Ω is a given bounded polygonal domain in $\mathbb{R}^2$. We first recall the classical functional framework ensuring that the problem above is well-posed (see [16]). Let $p \in]1,\infty[$ and $p' = \frac{p}{p-1}$. The flux $\varphi : \Omega \times \mathbb{R}^2 \to \mathbb{R}^2$ in (1.1) is supposed to be a Caratheodory function, which is strictly monotonic with respect to $\xi \in \mathbb{R}^2$:

$$(\mathcal{H}_1) \qquad (\varphi(z,\xi) - \varphi(z,\eta), \xi - \eta) > 0, \text{ for all } \xi \neq \eta, \text{ for } a.e.\ z \in \Omega\,.$$

We also assume that there exists $C_\varphi > 0$ such that

$$(\mathcal{H}_2) \qquad (\varphi(z,\xi),\xi) \geq \frac{1}{C_\varphi}|\xi|^p - C_\varphi, \text{ for all } \xi \in \mathbb{R}^2, \text{ for } a.e.\ z \in \Omega,$$

$$(\mathcal{H}_3) \qquad |\varphi(z,\xi)| \leq C_\varphi \left(1 + |\xi|^{p-1}\right), \text{ for all } \xi \in \mathbb{R}^2, \text{ for } a.e.\ z \in \Omega.$$

These assumptions ensure that $u \mapsto -\mathrm{div}(\varphi(\,\cdot\,,\nabla u))$ is a Leray–Lions operator, and hence that Problem (1.1) has a unique solution in $W_0^{1,p}(\Omega)$ for any $f \in W^{-1,p'}(\Omega)$. Nevertheless, since we are particularly interested in proving error estimates for (piecewise) smooth enough solutions, we restrict our attention here to source terms f in $L^{p'}(\Omega)$.

*Received by the editors July 27, 2006; accepted for publication (in revised form) April 25, 2008; published electronically September 4, 2008.
http://www.siam.org/journals/sinum/46-6/66619.html
†LATP, Université de Provence, 39 rue F. Joliot Curie, 13453 Marseille Cedex 13, France (fboyer@cmi.univ-mrs.fr, fhubert@cmmi.univ-mrs.fr).

In the present work, we concentrate on the case where the flux φ defining the equation admits discontinuities with respect to the space variable z. This kind of transmission (or bimaterial) problems were, for instance, studied in the finite element framework in [11, 20] for $p = 2$ and in [17, 18] for $p \neq 2$. A typical physical situation where one can face such problems is the computation of an incompressible flow in an heterogeneous porous medium (with various layers of different kinds, for instance). In a linear regime (for small or moderate pressure gradients), the situation is modeled by a Darcy low whose permeability tensor is anisotropic and possibly discontinuous across interfaces between different material layers. In a nonlinear regime, the Darcy law may not be relevant anymore, and we may use a nonlinear Darcy law (like the Darcy–Forchheimer law, for instance) in each of the layers of the porous medium. In that situation, we exactly have to deal with Leray–Lions-type operators with discontinuous coefficients. Note that the general situation where the operator is linear in some subdomains and nonlinear in others does not exactly enter the scope of the present paper. Nevertheless, the schemes we propose here can easily be extended to this case and the analysis (which is much more intricate) can also be performed (see some details in [5]).

A key concept in this work is the notion of transmission condition. Indeed, suppose, for instance, that Ω is divided into two disjoint domains Ω_1 and Ω_2 and that φ is smooth with respect to the space variable z on each subdomain Ω_i, $i = 1, 2$. A function $u_e \in W_0^{1,p}(\Omega)$ solves (1.1) if and only if:

1. It solves the equation $-\mathrm{div}(\varphi(z, \nabla u_e(z))) = f(z)$ in the two domains Ω_1 and Ω_2.
2. The normal flux $\varphi(z, \nabla u_e(z)) \cdot \boldsymbol{\nu}$ is *continuous* accross $\Gamma = \partial\Omega_1 \cap \partial\Omega_2$, which means that the two traces of this flux (in the sense of $W^{-\frac{1}{p'},p'}(\Gamma)$) computed from the two domains Ω_1 and Ω_2 are the same.

In the context of flows in heterogeneous porous media briefly described above, the transmission condition is nothing but the mass conservation property across the interfaces between the different layers.

Our analysis mainly consists in taking into account such a transmission condition in a finite volume scheme accross each edge in the mesh.

Finite volume approximation of such nonlinear elliptic problems is a current research topic. We refer to [3, 4, 9] for the description and the analysis of the main available schemes up to now. More precisely, we proposed in [3] to approach the solution to (1.1) by using a discrete duality finite volume method (DDFV). This method (previously studied in the linear case in [8, 14, 15]) can be applied to a wide class of 2D meshes (note that 3D cases can also be treated, see [6, 19]) and inherits the main qualitative properties of the continuous problem: monotonicity, coercivity, variational formulation, etc... (see Lemmas 4.7, 4.8, 4.9). Hence, we succeeded in showing the convergence of such schemes and error estimates in the case where the flux φ and the exact solution u_e are assumed to be smooth enough. In the case where φ has discontinuous coefficients, our results in [3] show that the scheme is still convergent but the error analysis is no more valid.

Actually, it is known (even for a 1D linear equation) that such discontinuities in the coefficients imply a consistency defect in the numerical fluxes of usual finite volume schemes. In the linear case, this leads to a $\frac{1}{2}$ convergence rate in the discrete H^1 norm instead of the first order we may expect. The situation is the same for DDFV schemes, and it is needed to modify the scheme in order to take into account the jumps of the coefficients of the problem and then to recover a better convergence rate.

The aim of this work is to present a modified DDFV scheme in this framework—that we called m-DDFV—which enjoys a better convergence rate than the usual DDFV method. Then we provide the error analysis of this scheme. In particular, in the linear case, we prove the first-order convergence of the m-DDFV scheme. Hence, our analysis provides a theoretical confirmation of the behavior numerically observed in a particular case in [15].

Outline. In section 2, we propose to study a simple 1D problem where the flux φ has only one point of discontinuity. This section will let us introduce the main ideas of the method we propose and illustrate the way one can obtain the consistency estimate for the scheme under study.

In section 3, we recall the DDFV framework for the finite volume approximation of nonlinear elliptic problems on unstructured grids. We also recall the scheme introduced and analyzed in [3]. In section 4, we describe the m-DDFV scheme and its first properties.

Section 5 is devoted to the error analysis of this method in the case where the exact solution is assumed to be piecewise smooth enough. The main new difficulty in the analysis, compared to the ones already encountered in [3], is contained in the consistency estimate of the new discrete gradient operator introduced in section 4. As an illustration, we give some explicit examples of the schemes under study in section 6. Nevertheless, in general, the method is not explicit and then seems to be difficult to solve. That is the reason why in section 7 we propose an iterative explicit algorithm to compute the approximate solution for any given data and we prove its convergence.

Notice that we also introduce a so-called *hybrid* DDFV scheme, called h-DDFV, for which a better error estimate can be obtained in the very common case where the flux is, in fact, smooth enough on a finite number of subdomains covering the whole domain Ω.

We finally conclude this paper by showing, in section 8, some numerical results illustrating both the efficiency of the finite volume scheme and of the iterative solver.

2. A 1D finite volume method for a model problem.

2.1. The toy system. Let us consider in this section a model problem in 1D of the form (1.1) in order to illustrate the main steps we will follow in the sequel of this paper for 2D problems. We take $\Omega =]-1,1[$ (denoting here by x the space variable) and we define $\varphi(x,\xi) = \varphi_-(\xi)$ for $x < 0$ and $\varphi(x,\xi) = \varphi_+(\xi)$ for $x > 0$. We suppose that φ_- and φ_+ are two increasing maps from $\mathbb{R}$ to $\mathbb{R}$ such that $(\mathcal{H}_2)$ and $(\mathcal{H}_3)$ hold.

Problem (1.1) reads in this setting

$$\begin{cases} -\partial_x\big(\varphi_-(\partial_x u_e)\big) = f(x), \text{ for } x < 0, \\ -\partial_x\big(\varphi_+(\partial_x u_e)\big) = f(x), \text{ for } x > 0, \\ u_e(-1) = u_e(1) = 0, \\ \varphi_-(\partial_x u_e(0^-)) = \varphi_+(\partial_x u_e(0^+)), \end{cases} \tag{2.1}$$

the last equation being, in this simplified case, the transmission condition described in the introduction.

2.2. The 1D finite volume scheme. Suppose now that we are given a finite volume mesh $\mathcal{T}$ of the domain Ω compatible with the discontinuity point. More precisely, let $x_0 = -1 < \ldots < x_N = 0 < \ldots < x_{N+M} = 1$ a subdivision of $[-1,1]$. We denote by $\kappa_{i+\frac{1}{2}} = [x_i, x_{i+1}]$, $i \in \{0,\ldots,N+M-1\}$ the control volumes of this discretization and by $x_{i+\frac{1}{2}}$ their centers. The finite volume method associates to each

center $x_{i+\frac{1}{2}}$ an unknown value $u_{i+\frac{1}{2}}$. We denote by $u^{\mathcal{T}} = (u_{i+\frac{1}{2}})_{0\le i\le N+M-1}$ the whole approximate solution, and we define the usual difference quotients

$$\nabla_i u^{\mathcal{T}} = \frac{u_{i+\frac{1}{2}} - u_{i-\frac{1}{2}}}{x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}}},\ i \in \{0,\dots,N+M\},$$

where, conventionally, we set $x_{-\frac{1}{2}} = x_0 = -1$, $x_{N+M+\frac{1}{2}} = x_{N+M} = 1$, and $u_{-\frac{1}{2}} = u_{N+M+\frac{1}{2}} = 0$. To obtain the finite volume scheme, we integrate the problem (2.1) on each control volume

$$-\int_{\mathcal{K}_{i+\frac{1}{2}}} \partial_x\big(\varphi(x,\partial_x u_e)\big)\,dx = \int_{\mathcal{K}_{i+\frac{1}{2}}} f(x)\,dx,\quad \forall i \in \{1,\dots,N+M-1\}.$$

Integrating the first term by parts, the scheme reads

$$-F_{i+1} + F_i = \int_{\mathcal{K}_{i+\frac{1}{2}}} f(x)\,dx,\quad \forall i \in \{0,\dots,N+M-1\}, \tag{2.2}$$

where F_i, $i \in \{0, N+M\}$ is an approximation of the flux $\varphi(x_i, \partial_x u_e(x_i))$. This approximation can easily be obtained away from the discontinuity in the usual way:

$$\begin{cases} F_i = \varphi_-\left(\nabla_i u^{\mathcal{T}}\right),\quad \forall i \in \{0,\dots,N-1\},\\ F_i = \varphi_+\left(\nabla_i u^{\mathcal{T}}\right),\quad \forall i \in \{N+1,\dots,N+M\}. \end{cases} \tag{2.3}$$

The problem is: how do we choose the numerical flux F_N at the point $x_N = 0$ where the discontinuity occurs? One may imagine many naive ways to treat this problem. For instance, one can define F_N by

$$F_N = \varphi_-\left(\nabla_N u^{\mathcal{T}}\right),\ \text{or } F_N = \varphi_+\left(\nabla_N u^{\mathcal{T}}\right),\ \text{or } F_N = \frac{1}{2}\left(\varphi_-\left(\nabla_N u^{\mathcal{T}}\right) + \varphi_+\left(\nabla_N u^{\mathcal{T}}\right)\right).$$

In fact, it can be shown that all these choices lead, in general, to a nonconsistent approximation of the flux at x_N. This fact is well known even in the linear case (see, e.g., [10]). A good way to find out a consistent approximation of the flux is to introduce a new artificial unknown $\tilde{u}$ at the point of the discontinuity x_N so that we can define two different approximate gradients on both sides of the discontinuity

$$\nabla_N^+ u^{\mathcal{T}} = \frac{u_{N+\frac{1}{2}} - \tilde{u}}{h_N^+},\quad \text{and } \nabla_N^- u^{\mathcal{T}} = \frac{\tilde{u} - u_{N-\frac{1}{2}}}{h_N^-}, \tag{2.4}$$

where we set $h_N^+ = x_{N+\frac{1}{2}} - x_N$ and $h_N^- = x_N - x_{N-\frac{1}{2}}$. In fact, it is convenient (see Figure 2.1 and the discussion below) to look for $\tilde{u}$ under the form

$$\tilde{u} = \bar{u} + \delta,\quad \text{with } \bar{u} = \frac{h_N^- u_{N+\frac{1}{2}} + h_N^+ u_{N-\frac{1}{2}}}{h_N^- + h_N^+}.$$

The value $\bar{u}$ is the value at the point 0 of the affine interpolation between $(x_{N-\frac{1}{2}}, u_{N-\frac{1}{2}})$ and $(x_{N+\frac{1}{2}}, u_{N+\frac{1}{2}})$. From now on, δ is the new artificial unknown to be determined. It follows that

$$\nabla_N^+ u^{\mathcal{T}} = \nabla_N u^{\mathcal{T}} - \frac{\delta}{h_N^+},\quad \text{and } \nabla_N^- u^{\mathcal{T}} = \nabla_N u^{\mathcal{T}} + \frac{\delta}{h_N^-}.$$

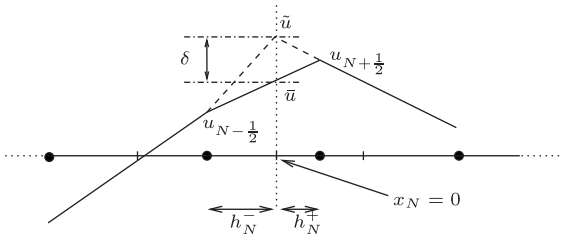


FIG. 2.1. *Illustration of the* 1*D case.*

Notice that we have

$$\nabla_N u^{\mathcal{T}} = \frac{1}{h_N^- + h_N^+} \left(h_N^- \nabla_N^- u^{\mathcal{T}} + h_N^+ \nabla_N^+ u^{\mathcal{T}} \right). \tag{2.5}$$

It is now necessary to eliminate the new unknown δ. This is done by imposing a discrete equivalent of the transmission condition in (2.1) which reads

$$\varphi_-(\nabla_N^- u^{\mathcal{T}}) = \varphi_+(\nabla_N^+ u^{\mathcal{T}}). \tag{2.6}$$

This equation uniquely defines δ as a function $\delta = \delta_N(\nabla_N u^{\mathcal{T}})$ of the usual difference quotient $\nabla_N u^{\mathcal{T}}$ since the map $\delta \mapsto \varphi_-(\nabla_N^- u^{\mathcal{T}}) - \varphi_+(\nabla_N^+ u^{\mathcal{T}})$ is strictly monotonic and tends to infinity at infinity. In the particular case where $\varphi_- = \varphi_+$, δ_N is then identically zero and $\nabla_N^+ u^{\mathcal{T}} = \nabla_N^- u^{\mathcal{T}} = \nabla_N u^{\mathcal{T}}$. Hence, we recover the generic situation without discontinuities in the coefficients of the equation. We can eventually define the approximate flux at the discontinuity by

$$F_N = \varphi_- \left(\nabla_N u^{\mathcal{T}} + \frac{\delta_N(\nabla_N u^{\mathcal{T}})}{h_N^-} \right) = \varphi_+ \left(\nabla_N u^{\mathcal{T}} - \frac{\delta_N(\nabla_N u^{\mathcal{T}})}{h_N^+} \right),$$

the last equality being true by definition of $\delta_N(\nabla_N u^{\mathcal{T}})$. In a more symmetric way we also have

$$F_N = \frac{h_N^- \varphi_- \left(\nabla_N u^{\mathcal{T}} + \frac{\delta_N(\nabla_N u^{\mathcal{T}})}{h_N^-} \right) + h_N^+ \varphi_+ \left(\nabla_N u^{\mathcal{T}} - \frac{\delta_N(\nabla_N u^{\mathcal{T}})}{h_N^+} \right)}{h_N^- + h_N^+}. \tag{2.7}$$

Example 2.1. *Let us consider the case where φ_- and φ_+ are two p-Laplacian-like fluxes given by*

$$\varphi_-(\xi) = k_- |\xi + G_-|^{p-2}(\xi + G_-), \ \forall \xi \in \mathbb{R},$$
$$\varphi_+(\xi) = k_+ |\xi + G_+|^{p-2}(\xi + G_+), \ \forall \xi \in \mathbb{R},$$

where $k_-, k_+ \in \mathbb{R}^+$ and $G_-, G_+ \in \mathbb{R}$. In this situation, all the computations can be made by hand. In particular, (2.6) *can be solved and we find*

$$\delta_N(\xi) = h_N^- h_N^+ \frac{\left[(k_+)^{\frac{1}{p-1}} - (k_-)^{\frac{1}{p-1}} \right] \xi + \left[(k_+)^{\frac{1}{p-1}} G_+ - (k_-)^{\frac{1}{p-1}} G_- \right]}{h_N^+ (k_-)^{\frac{1}{p-1}} + h_N^- (k_+)^{\frac{1}{p-1}}}.$$

Finally, the numerical flux at the discontinuity is found to be

$$F_N = \left(\frac{k_-^{\frac{1}{p-1}} k_+^{\frac{1}{p-1}} (h_N^- + h_N^+)}{h_N^+ k_-^{\frac{1}{p-1}} + h_N^- k_+^{\frac{1}{p-1}}} \right)^{p-1} \left| \nabla_N u^{\mathcal{T}} + \overline{G} \right|^{p-2} \left(\nabla_N u^{\mathcal{T}} + \overline{G} \right),$$

where $\overline{G}$ *is the arithmetic mean-value between* G_- *and* G_+ *defined by*

$$\overline{G} = \frac{h_N^- G_- + h_N^+ G_+}{h_N^- + h_N^+}.$$

Notice that the map $\nabla_N u^{\mathcal{T}} \mapsto F_N$ *is increasing and coercive. In the linear case (i.e.,* $p = 2$*) we recover the well-known harmonic mean-value formula between the two diffusion coefficients* k_- *and* k_+ *(see, for instance,* [10]*):*

$$F_N = \frac{k_- k_+ (h_N^- + h_N^+)}{h_N^+ k_- + h_N^- k_+} \left(\nabla_N u^{\mathcal{T}} + \overline{G} \right).$$

Let us sum up the previous study: we defined a monotonic map $\nabla_N u^{\mathcal{T}} \mapsto \delta_N(\nabla_N u^{\mathcal{T}})$ and a numerical flux F_N at the discontinuity, which is also a monotonic map with respect to $\nabla_N u^{\mathcal{T}}$. The finite volume scheme is then given by (2.2) with (2.3) and (2.7).

A very important remark is that the map δ_N is defined through the implicit relation (2.6) and hence, in general, cannot be computed explicitly like in Example 2.1. At first sight, it can be considered as a major drawback of our approach. Nevertheless, we will propose in section 7 a fully practical solver for this nonlinear scheme whose convergence is proved and whose computational cost is of the same order as in the case of continuous coefficient equations.

2.3. Consistency analysis. Let us analyze the consistency property of the flux F_N defined above. The following computations give the main ideas used in the analysis of the 2D scheme presented in the sequel of this paper. For simplicity we assume in this section that $p > 2$ and we suppose that φ_- and φ_+ satisfy the strong monotonicity assumption ($\mathcal{H}_{1'b}$) and the Hölder regularity assumption ($\mathcal{H}_{4b}$) described in section 4.1. Finally we suppose that the exact solution u_e of (2.1) is continuous on Ω and smooth on the two sides of the discontinuity point $x_N = 0$. In order to simplify the notations, assume that $h_N^- = h_N^+$ and denote this common value by h.

The error analysis for such a finite volume scheme basically relies on the estimate of the consistency error on the discrete gradient just defined above. This consistency error measures the distance between the gradient $\partial_x u_e$ of the exact solution and the discrete gradient applied to the projection $\mathbb{P}^{\mathcal{T}} u_e = (u_e(x_{i+\frac{1}{2}}))_{0 \le i \le N+M-1}$ of the exact solution on the mesh. More precisely, one has typically to estimate quantities like

$$R = \frac{1}{h} \int_{-h}^{0} |\partial_x u_e - \nabla_N^- \mathbb{P}^{\mathcal{T}} u_e|^p \, dx. \tag{2.8}$$

The corresponding result in the DDFV framework will be given in Proposition 5.7. It will imply the estimate of the consistency error on the fluxes (Proposition 5.8) and finally the *a priori* error estimates for the finite volume (Theorems 5.1 and 5.3).

Returning to our 1D example, let us derive an estimate for the term R defined in (2.8). Since u_e is smooth on $[-h, 0]$ we have

$$\begin{aligned} R &\le C \|\partial_x^2 u_e\|_\infty h^p + \frac{1}{h} \int_{-h}^{0} \left| \frac{u_e(x_N) - u_e(x_{N-\frac{1}{2}})}{h} - \frac{\tilde{u} - u_e(x_{N-\frac{1}{2}})}{h} \right|^p dx \\ &= C \|\partial_x^2 u_e\|_\infty h^p + \left| \frac{u_e(x_N) - \tilde{u}}{h} \right|^p, \end{aligned} \tag{2.9}$$

where $\tilde{u}$ is the artifical unknown (associated to $u^{\mathcal{T}} = \mathbb{P}^{\mathcal{T}} u_e$) defined in (2.6), that is, through the equation

$$\varphi_- \left(\frac{\tilde{u} - u_e(x_{N-\frac{1}{2}})}{h} \right) = \varphi_+ \left(\frac{u_e(x_{N+\frac{1}{2}}) - \tilde{u}}{h} \right). \tag{2.10}$$

Furthermore, since u_e is piecewise smooth, Taylor expansions yield

$$\frac{u_e(x_N) - u_e(x_{N-\frac{1}{2}})}{h} = \partial_x u_e(0^-) + T_1 h, \tag{2.11}$$

$$\frac{u_e(x_{N+\frac{1}{2}}) - u_e(x_N)}{h} = \partial_x u_e(0^+) + T_2 h, \tag{2.12}$$

where T_1 and T_2 are bounded with respect to h. Hence, using (2.11) and (2.12), the transmission condition in (2.1) reads

$$\varphi_- \left(\frac{u_e(x_N) - u_e(x_{N-\frac{1}{2}})}{h} - T_1 h \right) = \varphi_+ \left(\frac{u_e(x_{N+\frac{1}{2}}) - u_e(x_N)}{h} - T_2 h \right). \tag{2.13}$$

Finally, we estimate the second term in the right-hand side of (2.9) by using (2.10) and (2.13). To this end, we subtract (2.13) from (2.10) and we multiply by $\frac{u_e(x_N) - \tilde{u}}{h}$. It follows

$$\begin{aligned} &\left(\varphi_- \left(\frac{u_e(x_N) - u_e(x_{N-\frac{1}{2}})}{h} - T_1 h \right) - \varphi_- \left(\frac{\tilde{u} - u_e(x_{N-\frac{1}{2}})}{h} \right) \right) \frac{u_e(x_N) - \tilde{u}}{h} \\ &\quad + \left(\varphi_+ \left(\frac{u_e(x_{N+\frac{1}{2}}) - \tilde{u}}{h} \right) - \varphi_+ \left(\frac{u_e(x_{N+\frac{1}{2}}) - u_e(x_N)}{h} - T_2 h \right) \right) \frac{u_e(x_N) - \tilde{u}}{h} = 0. \end{aligned}$$

We add and subtract now the terms $T_1 h$ and $T_2 h$ in order to make appear expressions under the form $(\varphi_\pm(\xi) - \varphi_\pm(\eta))(\xi - \eta)$. We get

$$\begin{aligned} &\left(\varphi_- \left(\frac{u_e(x_N) - u_e(x_{N-\frac{1}{2}})}{h} - T_1 h \right) - \varphi_- \left(\frac{\tilde{u} - u_e(x_{N-\frac{1}{2}})}{h} \right) \right) \left(\frac{u_e(x_N) - \tilde{u}}{h} - T_1 h \right) \\ &+ \left(\varphi_+ \left(\frac{u_e(x_{N+\frac{1}{2}}) - \tilde{u}}{h} \right) - \varphi_+ \left(\frac{u_e(x_{N+\frac{1}{2}}) - u_e(x_N)}{h} - T_2 h \right) \right) \left(\frac{u_e(x_N) - \tilde{u}}{h} + T_2 h \right) \\ &\quad = -T_1 h \left(\varphi_- \left(\frac{u_e(x_N) - u_e(x_{N-\frac{1}{2}})}{h} - T_1 h \right) - \varphi_- \left(\frac{\tilde{u} - u_e(x_{N-\frac{1}{2}})}{h} \right) \right) \\ &\qquad + T_2 h \left(\varphi_+ \left(\frac{u_e(x_{N+\frac{1}{2}}) - \tilde{u}}{h} \right) - \varphi_+ \left(\frac{u_e(x_{N+\frac{1}{2}}) - u_e(x_N)}{h} - T_2 h \right) \right). \end{aligned}$$

Hence, using assumptions $(\mathcal{H}_{1'b})$ and $(\mathcal{H}_{4b})$ we deduce that

$$\begin{aligned} &\left| \frac{u_e(x_N) - \tilde{u}}{h} - T_1 h \right|^p + \left| \frac{u_e(x_N) - \tilde{u}}{h} + T_2 h \right|^p \\ &\qquad \leq Ch \left(1 + \left| \frac{u_e(x_N) - \tilde{u}}{h} \right|^{p-2} \right) \left(Ch + \left| \frac{u_e(x_N) - \tilde{u}}{h} \right| \right), \end{aligned}$$

and finally we have

$$\left|\frac{u_e(x_N)-\tilde{u}}{h}\right|^p \leq Ch^{\frac{p}{p-1}},$$

so that the consistency term R is finally bounded by

$$R \leq Ch^{\frac{p}{p-1}}.$$

When $p \to 2$ we recover the usual first-order estimate (that is, $R = O(h^2)$), whereas when p increases, this consistency order decreases.

3. The discrete duality finite volume framework.

3.1. Definition of the mesh. We recall the notations used in [3]. Let $\mathcal{T}$ be a triple $(\mathfrak{M}, \mathfrak{M}^*, \mathfrak{D})$ of meshes on Ω as follows (see Figure 3.1). The set $\mathfrak{M}$ is a set of disjoint open polygonal convex control volumes $\kappa \in \Omega$ such that $\cup\bar{\kappa} = \bar{\Omega}$. For all adjacent volume κ and $\mathcal{L}$, we assume that $\partial\kappa \cap \partial\mathcal{L}$ is a segment that we call *an edge of the mesh* and that we denote by $\sigma = \kappa|\mathcal{L}$. Let $\mathcal{E}_{int}$ denote the set of such edges, $\mathcal{E}_{ext}$ denotes the set of edges $\sigma = \partial\kappa \cap \partial\Omega$, and finally $\mathcal{E} = \mathcal{E}_{int} \cup \mathcal{E}_{ext}$. We associate to $\mathfrak{M}$ a family $\mathfrak{P}_{int}$ of points x_κ such that $x_\kappa \in \kappa$ and to the set $\mathcal{E}_{ext}$ a family $\mathfrak{P}_{ext}$ of points x_σ where x_σ is a point of $\sigma \in \mathcal{E}_{ext}$. Let $\mathfrak{P}^*$ be the set of vertices of the mesh $\mathfrak{M}$. The set $\mathfrak{P}^*$ can be decomposed into $\mathfrak{P}^* = \mathfrak{P}^*_{int} \cup \mathfrak{P}^*_{ext}$, where $\mathfrak{P}^*_{ext} \subset \partial\Omega$ and $\mathfrak{P}^*_{int} \cap \partial\Omega = \emptyset$. To any point $x_{\kappa^*} \in \mathfrak{P}^*_{int}$, we associate a polygon $\kappa^* \in \mathfrak{M}^*$ whose vertices are $\{x_\kappa \in \mathfrak{P}/x_{\kappa^*} \in \bar{\kappa},\ \kappa \in \mathfrak{M}\}$ sorted with respect to the clockwise order of the corresponding primal control volumes. Moreover $\mathcal{E}^*$ denotes the set of the edges of the mesh $\mathfrak{M}^*$.

For each $\sigma = \kappa|\mathcal{L} \in \mathcal{E}_{int}$, we can associate a diamond cell $\mathcal{D}$ where $\mathcal{D}$ is the quadrangle whose diagonals are $\sigma = (x_{\mathcal{K}^*}, x_{\mathcal{L}^*})$ and $\sigma^* = \kappa^*|\mathcal{L}^* = (x_\kappa, x_\mathcal{L})$ if $\sigma \in \mathcal{E}_{int}$, and if $\sigma = (x_{\mathcal{K}^*}, x_{\mathcal{L}^*}) \in \mathcal{E}_{ext} \cap \partial\bar{\kappa}$, $\mathcal{D}$ is the triangle defined by the points $x_\kappa, x_{\kappa^*}, x_{\mathcal{L}^*}$. The set of all diamond cells is noted $\mathfrak{D} = \mathfrak{D}_{\rm int} \cup \mathfrak{D}_{\rm ext}$. Note that $\mathfrak{D}$ forms a partition of Ω.

In this work, we assume that the diamond cells are all convex. Notice that this assumption is not necessary, in general, in order to define and analyze DDFV methods (see [3] and [8]).

3.2. Notations. For any one- (resp. two-) dimensional set $\mathcal{V}$, we denote by $|\mathcal{V}|$ its one-dimensional (resp. two-dimensional) Lebesgue measure.

For any control volume $\kappa \in \mathfrak{M}$, we define

- $\mathfrak{D}_\kappa = \{\mathcal{D} \in \mathfrak{D}/\kappa \cap \mathcal{D} \neq \emptyset\}$.

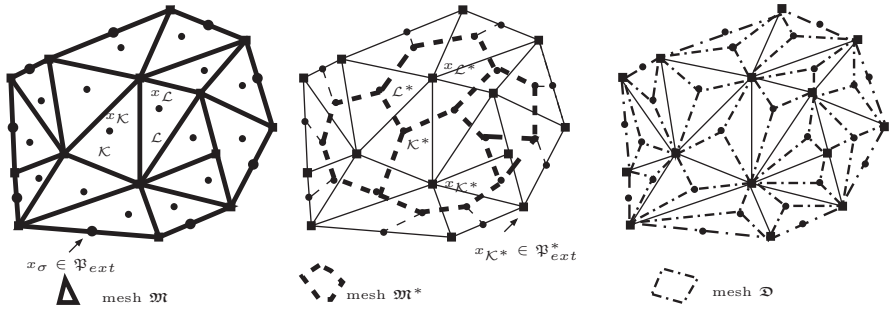


FIG. 3.1. *Example of a DDFV mesh.*

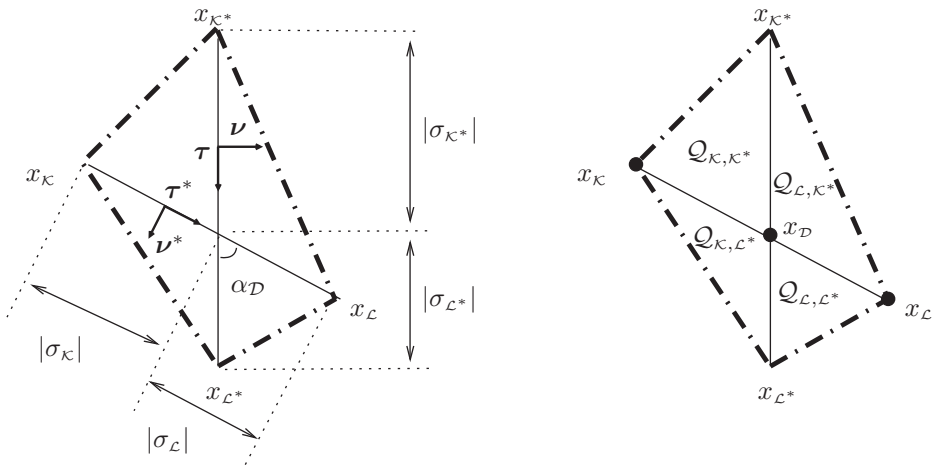


FIG. 3.2. *Notations in a diamond cell; quarter diamonds.*

- $\boldsymbol{\nu}_{\mathcal{K}}$, the outward unit normal vector to $\partial\mathcal{K}$.
- $\mathrm{d}_{\mathcal{K}}$, the diameter of $\mathcal{K}$.

In the same way, for a dual control volume $\mathcal{K}^* \in \mathfrak{M}^*$, we set

- $\mathfrak{D}_{\mathcal{K}^*} = \{\mathcal{D} \in \mathfrak{D} / \mathcal{K}^* \cap \mathcal{D} \neq \emptyset\}$.
- $\boldsymbol{\nu}_{\mathcal{K}^*}$, the outward unit normal vector to $\partial\mathcal{K}^*$.
- $\mathrm{d}_{\mathcal{K}^*}$, the diameter of $\mathcal{K}^*$.

For a diamond cell $\mathcal{D} \in \mathfrak{D}_{\mathrm{int}}$ (resp. $\mathcal{D} \in \mathfrak{D}_{\mathrm{ext}}$), recall that $(x_{\mathcal{K}}, x_{\mathcal{K}^*}, x_{\mathcal{L}}, x_{\mathcal{L}^*})$ are the vertices of $\mathcal{D}$ (resp. $(x_{\mathcal{K}}, x_{\mathcal{K}^*}, x_{\mathcal{L}^*})$ are vertices of $\mathcal{D}$ and $x_\sigma \in \partial\mathcal{D}$) and note:

- $\boldsymbol{\tau}$, the unit vector parallel to σ, oriented from $x_{\mathcal{K}^*}$ to $x_{\mathcal{L}^*}$.
- $\boldsymbol{\nu}$, the unit vector normal to σ, oriented from $x_{\mathcal{K}}$ to $x_{\mathcal{L}}$ (resp. from $x_{\mathcal{K}}$ to x_σ).
- $\boldsymbol{\tau}^*$, the unit vector parallel to σ^*, oriented from $x_{\mathcal{K}}$ to $x_{\mathcal{L}}$ (resp. from $x_{\mathcal{K}}$ to x_σ).
- $\boldsymbol{\nu}^*$, the unit vector normal to σ^*, oriented from $x_{\mathcal{K}^*}$ to $x_{\mathcal{L}^*}$.
- $\alpha_{\mathcal{D}}$, the angle between $\boldsymbol{\tau}$ and $\boldsymbol{\tau}^*$.
- $\mathrm{d}_{\mathcal{D}}$, the diameter of $\mathcal{D}$.
- $x_{\mathcal{D}}$ the intersection of $(x_{\mathcal{K}}, x_{\mathcal{L}})$ and $(x_{\mathcal{K}^*}, x_{\mathcal{L}^*})$ (resp. $x_{\mathcal{D}} = x_\sigma$).

As shown in Figure 3.2, each diamond cell $\mathcal{D} \in \mathfrak{D}_{\mathrm{int}}$ (resp. $\mathcal{D} \in \mathfrak{D}_{\mathrm{ext}}$) is naturally split along its diagonals into a set $\mathfrak{Q}_{\mathcal{D}}$ of four triangles (resp. two triangles), denoted by $\mathcal{Q}_{\mathcal{K},\mathcal{K}^*}$, $Q_{\mathcal{K},\mathcal{L}^*}$, $Q_{\mathcal{L},\mathcal{K}^*}$, and $Q_{\mathcal{L},\mathcal{L}^*}$ and satisfying $Q_{\mathcal{K},\mathcal{K}^*} \subset \overline{\mathcal{K}} \cap \overline{\mathcal{K}^*}$, etc ... The set of all the quarter diamonds in the domain is denoted by $\mathfrak{Q} = \cup_{\mathcal{D}\in\mathfrak{D}} \mathfrak{Q}_{\mathcal{D}}$, and $\mathrm{d}_{\mathcal{Q}}$ denotes the diameter of any $\mathcal{Q} \in \mathfrak{Q}$.

We denote by $\sigma_{\mathcal{K}}, \sigma_{\mathcal{L}}, \sigma_{\mathcal{K}^*}$, and $\sigma_{\mathcal{L}^*}$ the segments $(x_{\mathcal{K}}, x_{\mathcal{D}})$, $(x_{\mathcal{L}}, x_{\mathcal{D}})$, $(x_{\mathcal{K}^*}, x_{\mathcal{D}})$, and $(x_{\mathcal{L}^*}, x_{\mathcal{D}})$, so that $\sigma = \sigma_{\mathcal{K}^*} \cup \sigma_{\mathcal{L}^*}$ and $\sigma^* = \sigma_{\mathcal{K}} \cup \sigma_{\mathcal{L}}$ for $\mathcal{D} \in \mathfrak{D}_{\mathrm{int}}$. For $\mathcal{D} \in \mathfrak{D}_{\mathrm{ext}}$ we note abusively $\sigma^* = \sigma_{\mathcal{K}}$.We note $\mathcal{E}_{\mathcal{Q}}$ the set of such segments included in $\partial\mathcal{Q}$.

3.3. Regularity assumptions for the meshes. We note size$(\mathcal{T})$ as the maximum of the diameters of the diamond cells in $\mathfrak{D}$. The bounds follow:

$$
\begin{aligned}
&|\sigma| \leq \mathrm{size}(\mathcal{T}),\ \forall \sigma \in \mathcal{E}; \quad |\sigma^*| \leq \mathrm{size}(\mathcal{T}),\ \forall \sigma^* \in \mathcal{E}^*;\\
&|\mathcal{K}| \leq \pi\, \mathrm{size}(\mathcal{T})^2,\ \forall \mathcal{K} \in \mathfrak{M}; \quad |\mathcal{K}^*| \leq \pi\, \mathrm{size}(\mathcal{T})^2,\ \forall \mathcal{K}^* \in \mathfrak{M}^*;\\
&|\mathcal{Q}| \leq \frac{1}{2}\mathrm{size}(\mathcal{T})^2,\ \forall \mathcal{Q} \in \mathfrak{Q}.
\end{aligned}
$$

To measure how flat the diamond cells are, we introduce $\alpha_{\mathcal{T}}$ the unique real in $]0, \frac{\pi}{2}]$ such that $\sin\alpha_{\mathcal{T}} \stackrel{\text{def}}{=} \min_{\mathcal{D}\in\mathfrak{D}} |\sin\alpha_{\mathcal{D}}|$. We also need to control the ratio between the sizes of the quarter diamond cells inside each diamond $\mathcal{D}$. As a consequence, we will measure the regularity of the DDFV mesh by the following quantity:

$$\mathrm{reg}(\mathcal{T}) \stackrel{\text{def}}{=} \max\left(\frac{1}{\alpha_{\mathcal{T}}}, \max_{\mathcal{D}\in\mathfrak{D}} \max_{\mathcal{Q}\in\mathfrak{Q}_{\mathcal{D}}} \frac{\mathrm{d}_{\mathcal{D}}}{\sqrt{|\mathcal{Q}|}}, \max_{\substack{\mathcal{K}\in\mathfrak{M}\\ \mathcal{D}\in\mathfrak{D}_{\mathcal{K}}}} \frac{\mathrm{d}_{\mathcal{K}}}{\mathrm{d}_{\mathcal{D}}}, \max_{\substack{\mathcal{K}^*\in\mathfrak{M}^*\\ \mathcal{D}\in\mathfrak{D}_{\mathcal{K}^*}}} \frac{\mathrm{d}_{\mathcal{K}^*}}{\mathrm{d}_{\mathcal{D}}}\right).$$

In particular, there exists two constants C_1 and C_2 depending on $\mathrm{reg}(\mathcal{T})$ such that for any $\kappa \in \mathfrak{M}$, $\kappa^* \in \mathfrak{M}^*$, and $\mathcal{D} \in \mathfrak{D}$ such that $\mathcal{D} \cap \kappa \neq \emptyset$ and $\mathcal{D} \cap \kappa^* \neq \emptyset$ we have

$$C_1|\kappa| \leq |\mathcal{D}| \leq C_2|\kappa|, \quad C_1|\kappa^*| \leq |\mathcal{D}| \leq C_2|\kappa^*|.$$

3.4. Original DDFV approach for linear problems. The DDFV finite volume method associates to all primal control volume $\kappa \in \mathfrak{M}$ an unknown value $u_{\mathcal{K}}$ and to all dual control volume $\kappa^* \in \mathfrak{M}^*$ an unknown value $u_{\mathcal{K}^*}$. The approximate solution $u^{\mathcal{T}}$ is denoted by

$$u^{\mathcal{T}} = \left((u_{\mathcal{K}})_{\mathcal{K}\in\mathfrak{M}}, (u_{\mathcal{K}^*})_{\mathcal{K}^*\in\mathfrak{M}^*}\right).$$

The set of such unknowns $u^{\mathcal{T}}$ is denoted by $\mathbb{R}^{\mathcal{T}}$.

The method consists in introducing a discrete gradient operator $\nabla^{\mathcal{T}}$ (see [7, 8]) defined to be constant on each diamond cell

$$\nabla^{\mathcal{T}}_{\mathcal{D}} u^{\mathcal{T}} = \frac{1}{\sin\alpha_{\mathcal{D}}}\left(\frac{u_{\mathcal{L}} - u_{\mathcal{K}}}{|\sigma^*|}\boldsymbol{\nu} + \frac{u_{\mathcal{L}^*} - u_{\mathcal{K}^*}}{|\sigma|}\boldsymbol{\nu}^*\right), \tag{3.1}$$

where $u_{\mathcal{K}^*} = 0$ (resp. $u_{\mathcal{L}^*} = 0$) if $x_{\mathcal{K}^*} \in \mathfrak{P}^*_{ext}$ (resp. if $x_{\mathcal{L}^*} \in \mathfrak{P}^*_{ext}$) and $u_{\mathcal{L}} = 0$ if $\mathcal{D} \in \mathfrak{D}_{\mathrm{ext}}$. By elementary geometric arguments, we also have the useful equivalent formula given in [8]:

$$\nabla^{\mathcal{T}}_{\mathcal{D}} u^{\mathcal{T}} = \frac{1}{2|\mathcal{D}|}\Big(|\sigma|(u_{\mathcal{L}} - u_{\mathcal{K}})\boldsymbol{\nu} + |\sigma^*|(u_{\mathcal{L}^*} - u_{\mathcal{K}^*})\Big). \tag{3.2}$$

Then the discrete divergence operator is defined to be the adjoint of $\nabla^{\mathcal{T}}$, so that for linear equations (in particular the Laplace equation), we obtain a well-posed finite volume scheme which is, for instance, studied in [8, 14]. Notice that the discrete gradient operator $\nabla^{\mathcal{T}}$ was already used, e.g., in [1, 2, 7], but in these references the values $u_{\mathcal{K}^*}$ and $u_{\mathcal{L}^*}$ on the dual mesh were not considered as unknowns of the problem but were built *via* interpolation formulas between the values of the solution on the primal mesh. In this last approach, there are less unknowns, but the structure of the original equation (in particular, the symmetry) is lost. The main advantage of the DDFV method is that the discrete equations inherit from the properties of the continuous one, which is crucial in particular in the nonlinear setting.

3.5. The DDFV method for nonlinear elliptic problems. In [3], we have studied the generalization of the DDFV method to the nonlinear equation (1.1). We proved that all the tools used in the study of this equation (monotonicity, compactness, etc...) can be translated to the discrete level. The scheme we proposed consists in integrating the equation (1.1) on each $\kappa \in \mathfrak{M}$ and each $\kappa^* \in \mathfrak{M}^*$ and then to approximate fluxes $\int_\sigma (\varphi(s, \nabla u_e(s)), \boldsymbol{\nu})\, ds$ or $\int_{\sigma^*} (\varphi(s, \nabla u_e(s)), \boldsymbol{\nu}^*)\, ds$ by using the

discrete gradient $\nabla^{\mathcal{T}}$ operator defined above. The scheme now writes:

$$\left\{\begin{array}{l} -\displaystyle\sum_{\mathcal{D}\in\mathfrak{D}_{\mathcal{K}}} |\sigma|\left(\varphi_{\mathcal{D}}(\nabla^{\mathcal{T}}_{\mathcal{D}}u^{\mathcal{T}}),\boldsymbol{\nu}_{\mathcal{K}}\right) = |\mathcal{K}| f_{\mathcal{K}},\ \forall \mathcal{K}\in\mathfrak{M}, \\ -\displaystyle\sum_{\mathcal{D}\in\mathfrak{D}_{\mathcal{K}^*}} |\sigma^*|\left(\varphi_{\mathcal{D}}(\nabla^{\mathcal{T}}_{\mathcal{D}}u^{\mathcal{T}}),\boldsymbol{\nu}_{\mathcal{K}^*}\right) = |\mathcal{K}^*| f_{\mathcal{K}^*},\ \forall \mathcal{K}^*\in\mathfrak{M}^*, \end{array}\right. \tag{3.3}$$

where $f_{\mathcal{K}}$ and $f_{\mathcal{K}^*}$ denote the mean value of f over κ and κ^*, respectively, and $\varphi_{\mathcal{D}}$ is the mean value of φ over $\mathcal{D}$; that is:

$$\varphi_{\mathcal{D}}(\xi) = \frac{1}{|\mathcal{D}|}\int_{\mathcal{D}} \varphi(z,\xi)\,dz. \tag{3.4}$$

We proved in [3] that this scheme is convergent for any φ satisfying assumptions $(\mathcal{H}_1)$–$(\mathcal{H}_3)$ and any source term $f \in L^{p'}(\Omega)$, and that we can adapt the scheme for source terms in $W^{-1,p'}(\Omega)$, which is the natural space in which source terms can be taken.

We also proved error estimates for the scheme above in the case where the flux φ is assumed to be smooth enough with respect to ξ and to z on the whole domain Ω, and u_e is assumed to belong to $W^{2,p}(\Omega)$.

4. Taking into account discontinuities in the DDFV framework. The point we are concerned with in this paper is that the scheme (3.3) (even though we know that it is convergent) suffers from a lost of consistency in the case where φ presents discontinuities in the space variable z. This behavior is illustrated in section 8 in comparison with the one of the new scheme we propose in the present section. More precisely, we present a way to recover the consistency of the fluxes even when φ presents jumps. The method essentially follows the line described for the very simple toy 1D problem studied in section 2.

4.1. Assumptions on the flux $\boldsymbol{\varphi}$. We first give the precise assumptions we need on the flux φ. First of all, we reinforce the monotonicity assumption $(\mathcal{H}_1)$ in the following way:

- If $1 < p \le 2$: for all $(\xi,\eta)\in\mathbb{R}^2\times\mathbb{R}^2$ and almost every $z\in\Omega$,

$$(\varphi(z,\xi)-\varphi(z,\eta),\xi-\eta) \ge \frac{1}{C_\varphi}|\xi-\eta|^2(1+|\xi|^p+|\eta|^p)^{\frac{p-2}{p}}. \tag{$\mathcal{H}_{1'a}$}$$

- If $p > 2$: for all $(\xi,\eta)\in\mathbb{R}^2\times\mathbb{R}^2$ and almost every $z\in\Omega$,

$$(\varphi(z,\xi)-\varphi(z,\eta),\xi-\eta) \ge \frac{1}{C_\varphi}|\xi-\eta|^p. \tag{$\mathcal{H}_{1'b}$}$$

We also assume that the flux φ is Hölder continuous with respect to ξ:

- If $1 < p \le 2$: for all $(\xi,\eta)\in\mathbb{R}^2\times\mathbb{R}^2$ and almost every $z\in\Omega$,

$$|\varphi(z,\xi)-\varphi(z,\eta)| \le C_\varphi|\xi-\eta|^{p-1}. \tag{$\mathcal{H}_{4a}$}$$

- If $p > 2$: for all $(\xi,\eta)\in\mathbb{R}^2\times\mathbb{R}^2$, and almost every $z\in\Omega$,

$$|\varphi(z,\xi)-\varphi(z,\eta)| \le C_\varphi\left(1+|\xi|^{p-2}+|\eta|^{p-2}\right)|\xi-\eta|. \tag{$\mathcal{H}_{4b}$}$$

The four assumptions above are classical in the error analysis of numerical methods for nonlinear problems and are satisfied by many usual nonlinear operators. We can think, for instance, to p-Laplacian-like operators $\varphi(z,\xi) = k(z)(A(z)\xi,\xi)^{\frac{p-2}{2}}A(z)\xi$, where k (resp. A) is a real-valued (resp. symmetric matrix-valued) bounded map satisfying a uniform coercivity assumption. We also refer to [3] for other examples.

Finally, as we have seen above we want to consider a flux φ which is piecewise smooth with respect to the space variable. The precise meaning of this statement is the following:

- If $1 < p \leq 2$: for all $\xi \in \mathbb{R}^2$, for all $\mathcal{Q} \in \mathfrak{Q}$ and almost every $(z,z') \in \mathcal{Q}^2$,

$$(\mathcal{H}_{5a}) \qquad |\varphi(z,\xi) - \varphi(z',\xi)| \leq C_\varphi(1+|\xi|^{p-1})|z-z'|^{p-1}.$$

- If $p > 2$: φ is Lipschitz on any $\mathcal{Q} \in \mathfrak{Q}$, and for all $\xi \in \mathbb{R}^2$ and almost every $z \in \mathcal{Q}$ we have

$$(\mathcal{H}_{5b}) \qquad \left|\frac{\partial\varphi}{\partial z}(z,\xi)\right| \leq C_\varphi\left(1+|\xi|^{p-1}\right).$$

Contrary to the assumptions we considered in [3], the above hypotheses are localized on each quarter diamond. From a practical point of view, this means that the mesh is built in such a way that the discontinuities with respect to the space variable z of the flux φ are allowed only across edges of the primal mesh and edges of the dual mesh.

4.2. Approximate fluxes on the quarter diamond. From now on, we assume that φ is a given flux satisfying $(\mathcal{H}_2)$,$(\mathcal{H}_3)$ and either $(\mathcal{H}_{1'a})$, $(\mathcal{H}_{4a})$, and $(\mathcal{H}_{5a})$ if $p \leq 2$, or $(\mathcal{H}_{1'b})$, $(\mathcal{H}_{4b})$, and $(\mathcal{H}_{5b})$ if $p > 2$.

Then, we define for each quarter diamond $\mathcal{Q} \in \mathfrak{Q}$, an approximation $\varphi_\mathcal{Q}$ of φ defined as the mean value of φ on $\mathcal{Q}$, that is,

$$\varphi_\mathcal{Q}(\xi) = \frac{1}{|\mathcal{Q}|}\int_\mathcal{Q} \varphi(z,\xi)\,dz. \tag{4.1}$$

Such a map $\varphi_\mathcal{Q}$ inherits the monotonicity, coercivity, and regularity properties of the initial flux φ, that is, for any $\mathcal{Q} \in \mathfrak{Q}$:

$$(\mathcal{P}_2^\mathcal{T}) \qquad (\varphi_\mathcal{Q}(\xi),\xi) \geq \frac{1}{C_\varphi}|\xi|^p - C_\varphi, \quad \forall \xi \in \mathbb{R}^2,$$

$$(\mathcal{P}_3^\mathcal{T}) \qquad |\varphi_\mathcal{Q}(\xi)| \leq C_\varphi(1+|\xi|^{p-1}), \quad \forall \xi \in \mathbb{R}^2.$$

- If $1 < p \leq 2$:

$$(\mathcal{P}_{1'a}^\mathcal{T}) \quad (\varphi_\mathcal{Q}(\xi) - \varphi_\mathcal{Q}(\eta), \xi-\eta) \geq \frac{1}{C_\varphi}|\xi-\eta|^2(1+|\xi|^p+|\eta|^p)^{\frac{p-2}{p}}, \quad \forall \xi,\eta \in \mathbb{R}^2,$$

$$(\mathcal{P}_{4a}^\mathcal{T}) \qquad |\varphi_\mathcal{Q}(\xi) - \varphi_\mathcal{Q}(\eta)| \leq C_\varphi|\xi-\eta|^{p-1}, \quad \forall \xi,\eta \in \mathbb{R}^2.$$

- If $p > 2$:

$$(\mathcal{P}_{1'b}^\mathcal{T}) \qquad (\varphi_\mathcal{Q}(\xi) - \varphi_\mathcal{Q}(\eta), \xi-\eta) \geq \frac{1}{C_\varphi}|\xi-\eta|^p, \quad \forall \xi,\eta \in \mathbb{R}^2,$$

$$(\mathcal{P}_{4b}^\mathcal{T}) \qquad |\varphi_\mathcal{Q}(\xi) - \varphi_\mathcal{Q}(\eta)| \leq C_\varphi\left(1+|\xi|^{p-2}+|\eta|^{p-2}\right)|\xi-\eta|, \quad \forall \xi,\eta \in \mathbb{R}^2.$$

Remark 4.1. It is possible to consider more general definitions of $\varphi_{\mathcal{Q}}$ under the form

$$\varphi_{\mathcal{Q}}(\,\cdot\,) = \int_{\overline{\mathcal{Q}}} \varphi(z,\,\cdot\,)\, d\mu_{\overline{\mathcal{Q}}}(z),$$

where $d\mu_{\overline{\mathcal{Q}}}$ is any probability measure on $\overline{\mathcal{Q}}$. This makes sense since φ is supposed to be Hölder continuous on $\mathcal{Q}$ (see assumptions $(\mathcal{H}_{5a})$–$(\mathcal{H}_{5b})$ above) and hence can be extended to a continuous map on $\overline{\mathcal{Q}}$. This quite general framework includes the case where $\varphi_{\mathcal{Q}}$ is defined by (4.1) but also the case where $\varphi_{\mathcal{Q}}$ is chosen to be the value of φ at a given point in $\overline{\mathcal{Q}}$ or more generally an approximation of the mean value of φ through a quadrature formula with nonnegative coefficients. These three situations are the usual ones that we may use in practice.

For such choices, properties $(\mathcal{P}_2^{\mathcal{T}})$–$(\mathcal{P}_{4b}^{\mathcal{T}})$ still hold and the analysis performed in the remainder of this paper is still valid. The only difference is that a supplementary term $R_{\mathcal{Q}}^z$ appears in the definition 5.5 of the consistency error of the gradient. This term writes

$$R_{\mathcal{Q}}^z = \frac{1}{|\mathcal{Q}|}\int_{\mathcal{Q}} \varphi(z', \nabla_{\mathcal{Q}}^{\mathcal{N}}\mathbb{P}^{\mathcal{T}} u_e)\, dz' - \int_{\overline{\mathcal{Q}}} \varphi(z', \nabla_{\mathcal{Q}}^{\mathcal{N}}\mathbb{P}^{\mathcal{T}} u_e) d\mu_{\overline{\mathcal{Q}}}(z'),$$

and can be estimated by writing

$$|R_{\mathcal{Q}}^z| \le \frac{1}{|\mathcal{Q}|}\int_{\mathcal{Q}}\int_{\overline{\mathcal{Q}}} |\varphi(z, \nabla_{\mathcal{Q}}^{\mathcal{N}}\mathbb{P}^{\mathcal{T}} u_e) - \varphi(z', \nabla_{\mathcal{Q}}^{\mathcal{N}}\mathbb{P}^{\mathcal{T}} u_e)|\, dz\, d\mu_{\overline{\mathcal{Q}}}(z')$$

and using assumptions $(\mathcal{H}_{5a})$ or $(\mathcal{H}_{5b})$.

4.3. Local modification of the discrete gradient operator. As we saw in the 1D case (see section 2), we need to introduce new gradient operators near the discontinuities of the flux and finally define a new approximate flux $\varphi_{\mathcal{D}}^{\mathcal{N}}$ on each diamond cell.

The new gradient operator $\nabla^{\mathcal{N}}$ we propose to consider is built upon the usual DDFV gradient $\nabla^{\mathcal{T}}$. It is chosen to be constant on all the quarter diamonds $\mathcal{Q} \in \mathfrak{Q}$. This new operator has to be thought as the 2D generalization of the definitions of ∇_{N}^{+} and ∇_{N}^{-} in (2.4). In this 1D situation, the place where the artificial unknown $\tilde{u}$ (or $\delta = \tilde{u} - \bar{u}$) must be chosen is clear: it is the point of the mesh where the discontinuity takes place, that is, $x_N = 0$ in the framework of section 2.

In the 2D setting the situation is less straightforward. In order to make a good choice it is useful to remember that the usual DDFV gradient $\nabla_{\mathcal{D}}^{\mathcal{T}} u^{\mathcal{T}}$ on a diamond cell $\mathcal{D}$ can be understood (it is an easy computation) as the gradient of the unique affine function $\Pi_{\mathcal{D}} u^{\mathcal{T}}$ whose value at the middle of each side of the diamond $\mathcal{D}$ is the mean value between the two unknowns associated to the two extremities of this segment. This situation is summed up in Figure 4.1 for a given interior diamond cell $\mathcal{D} \in \mathfrak{D}_{\text{int}}$. In this figure, we introduce x_σ to be the middle of the segment σ for each $\sigma \in \{\sigma_{\mathcal{K}}, \sigma_{\mathcal{L}}, \sigma_{\mathcal{K}^*}, \sigma_{\mathcal{L}^*}\}$.

It seems now natural to define the new discrete gradient operator $\nabla_{\mathcal{Q}}^{\mathcal{N}} u^{\mathcal{T}}$ on each quarter diamond as the gradient of a function $\widetilde{\Pi}_{\mathcal{D}} u^{\mathcal{T}}$ which coincides with $\Pi_{\mathcal{D}} u^{\mathcal{T}}$ in the middle of each side of $\mathcal{D}$ and which is continuous at each point $x_{\sigma_{\mathcal{K}}}$, $x_{\sigma_{\mathcal{L}}}$, $x_{\sigma_{\mathcal{K}^*}}$, and $x_{\sigma_{\mathcal{L}^*}}$, but which is not necessarily continuous on the whole diamond $\mathcal{D}$.

This new function $\widetilde{\Pi}_{\mathcal{D}} u^{\mathcal{T}}$ is then entirely defined, for a given $u^{\mathcal{T}}$, by its four values $\widetilde{\Pi}_{\mathcal{D}} u^{\mathcal{T}}(x_\sigma)$ at each of these four points x_σ. These four values are the artificial unknowns

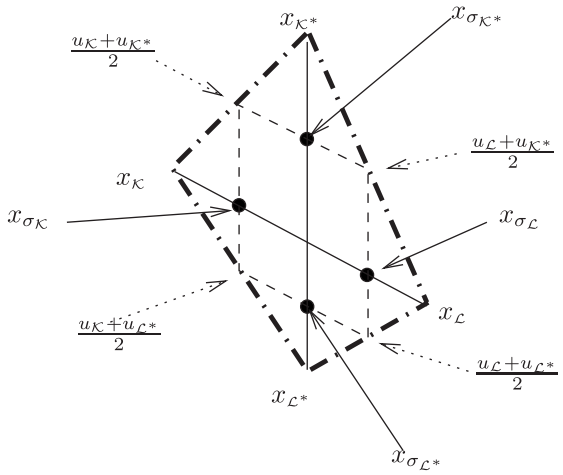


FIG. 4.1. *Affine function whose gradient is* $\nabla_{\mathcal{D}}^{\mathcal{T}} u^{\mathcal{T}}$.

in our problem. Like in the 1D case, it is equivalent and more suitable to work with the new unknowns $\delta_{\mathcal{K}}^{\mathcal{D}}, \delta_{\mathcal{L}}^{\mathcal{D}}, \delta_{\mathcal{K}^*}^{\mathcal{D}}, \delta_{\mathcal{L}^*}^{\mathcal{D}}$ defined to be the differences $\widetilde{\Pi}_{\mathcal{D}} u^{\mathcal{T}}(x_\sigma) - \Pi_{\mathcal{D}} u^{\mathcal{T}}(x_\sigma)$ for each $\sigma \in \{\sigma_{\mathcal{K}}, \sigma_{\mathcal{L}}, \sigma_{\mathcal{K}^*}, \sigma_{\mathcal{L}^*}\}$. Notice that each $\Pi_{\mathcal{D}} u^{\mathcal{T}}(x_\sigma)$ can be computed as an explicit function of $u_{\mathcal{K}}, u_{\mathcal{L}}, u_{\mathcal{K}^*}$, and $u_{\mathcal{L}^*}$.

The situation is simpler in the case of exterior diamond cells $\mathcal{D} \in \mathfrak{D}_{\text{ext}}$, in which case we only need one artificial unknown, $\delta_{\mathcal{K}}^{\mathcal{D}}$. Hence, we define $n_{\mathcal{D}}$ to be the number of artificial unknowns needed on the diamond cell $\mathcal{D}$. From the discussion above we have $n_{\mathcal{D}} = 4$ if $\mathcal{D} \in \mathfrak{D}_{\text{int}}$ and $n_{\mathcal{D}} = 1$ if $\mathcal{D} \in \mathfrak{D}_{\text{ext}}$.

By straightforward computations, the above discussion can be summed up as follows: we define the new discrete gradient operator on each quarter diamond $\mathcal{Q} \in \mathfrak{Q}_{\mathcal{D}}$, to be the gradient of $\widetilde{\Pi}_{\mathcal{D}} u^{\mathcal{T}}$, which reads

$$\nabla_{\mathcal{Q}}^{\mathcal{N}} u^{\mathcal{T}} = \nabla_{\mathcal{D}}^{\mathcal{T}} u^{\mathcal{T}} + B_{\mathcal{Q}} \delta^{\mathcal{D}}, \tag{4.2}$$

where $\delta^{\mathcal{D}} \in \mathbb{R}^{n_{\mathcal{D}}}$ is an artificial $n_{\mathcal{D}}$-tuple of unknowns introduced above and $(B_{\mathcal{Q}})_{\mathcal{Q} \in \mathfrak{Q}_{\mathcal{D}}}$ is a family of matrices whose size is $2 \times n_{\mathcal{D}}$. Existence of such matrices $B_{\mathcal{Q}}$ is clear since $\nabla \widetilde{\Pi}_{\mathcal{D}} u^{\mathcal{T}} - \nabla \Pi_{\mathcal{D}} u^{\mathcal{T}}$ depends linearly on $\delta^{\mathcal{D}}$ on each quarter diamond $\mathcal{Q}$.

In order to implement and to analyze the scheme we need to compute explicitly the matrices $B_{\mathcal{Q}}$. This can be done by integrating (4.2) on $\mathcal{Q}$ and by using the Stokes formula. For instance, for an interior diamond cell, and considering $\mathcal{Q} = \mathcal{Q}_{\mathcal{K},\mathcal{K}^*}$, we find that for any $\delta^{\mathcal{D}}$

$$|\mathcal{Q}_{\mathcal{K},\mathcal{K}^*}| B_{\mathcal{Q}_{\mathcal{K},\mathcal{K}^*}} \delta^{\mathcal{D}} = \int_{\partial \mathcal{Q}_{\mathcal{K},\mathcal{K}^*}} \left(\widetilde{\Pi}_{\mathcal{D}} u^{\mathcal{T}} - \Pi_{\mathcal{D}} u^{\mathcal{T}} \right) \cdot \boldsymbol{\nu}_{\mathcal{Q}} \, ds.$$

We use now the fact that the integral of an affine function over a segment is equal to the value of the function in the middle of the segment multiplied by its length. In particular, since $\widetilde{\Pi}_{\mathcal{D}} u^{\mathcal{T}}$ and $\Pi_{\mathcal{D}} u^{\mathcal{T}}$ are equal in the middle of the sides of $\mathcal{D}$, it only remains the contributions on $\sigma_{\mathcal{K}}$ and $\sigma_{\mathcal{K}^*}$ which reads, by definition of $\delta^{\mathcal{D}}$

$$|\mathcal{Q}_{\mathcal{K},\mathcal{K}^*}| B_{\mathcal{Q}_{\mathcal{K},\mathcal{K}^*}} \delta^{\mathcal{D}} = \delta_{\mathcal{K}}^{\mathcal{D}} |\sigma_{\mathcal{K}}| \boldsymbol{\nu}^* + \delta_{\mathcal{K}^*}^{\mathcal{D}} |\sigma_{\mathcal{K}^*}| \boldsymbol{\nu}.$$

Doing the same computations in the other quarter diamonds we find that, for $\mathcal{D} \in \mathfrak{D}_{\text{int}}$, the four matrices $B_{\mathcal{Q}}$ writes:

$$B_{\mathcal{Q}_{\mathcal{K},\mathcal{K}^*}} = \frac{1}{|\mathcal{Q}_{\mathcal{K},\mathcal{K}^*}|}\left(|\sigma_{\mathcal{K}}|\boldsymbol{\nu}^*, 0, |\sigma_{\mathcal{K}^*}|\boldsymbol{\nu}, 0\right) = \frac{2}{\sin\alpha_{\mathcal{D}}}\left(\frac{\boldsymbol{\nu}^*}{|\sigma_{\mathcal{K}^*}|}, 0, \frac{\boldsymbol{\nu}}{|\sigma_{\mathcal{K}}|}, 0\right), \tag{4.3}$$

$$B_{\mathcal{Q}_{\mathcal{K},\mathcal{L}^*}} = \frac{1}{|\mathcal{Q}_{\mathcal{K},\mathcal{L}^*}|}\left(-|\sigma_{\mathcal{K}}|\boldsymbol{\nu}^*, 0, 0, |\sigma_{\mathcal{L}^*}|\boldsymbol{\nu}\right) = \frac{2}{\sin\alpha_{\mathcal{D}}}\left(-\frac{\boldsymbol{\nu}^*}{|\sigma_{\mathcal{L}^*}|}, 0, 0, \frac{\boldsymbol{\nu}}{|\sigma_{\mathcal{K}}|}\right), \tag{4.4}$$

$$B_{\mathcal{Q}_{\mathcal{L},\mathcal{L}^*}} = \frac{1}{|\mathcal{Q}_{\mathcal{L},\mathcal{L}^*}|}\left(0, -|\sigma_{\mathcal{L}}|\boldsymbol{\nu}^*, 0, -|\sigma_{\mathcal{L}^*}|\boldsymbol{\nu}\right) = \frac{2}{\sin\alpha_{\mathcal{D}}}\left(0, -\frac{\boldsymbol{\nu}^*}{|\sigma_{\mathcal{L}^*}|}, 0, -\frac{\boldsymbol{\nu}}{|\sigma_{\mathcal{L}}|}\right), \tag{4.5}$$

$$B_{\mathcal{Q}_{\mathcal{L},\mathcal{K}^*}} = \frac{1}{|\mathcal{Q}_{\mathcal{L},\mathcal{K}^*}|}\left(0, |\sigma_{\mathcal{L}}|\boldsymbol{\nu}^*, -|\sigma_{\mathcal{K}^*}|\boldsymbol{\nu}, 0\right) = \frac{2}{\sin\alpha_{\mathcal{D}}}\left(0, \frac{\boldsymbol{\nu}^*}{|\sigma_{\mathcal{K}^*}|}, -\frac{\boldsymbol{\nu}}{|\sigma_{\mathcal{L}}|}, 0\right). \tag{4.6}$$

In the case where $\mathcal{D} \in \mathfrak{D}_{\text{ext}}$, there are only two nondegenerate quarter diamonds in $\mathfrak{Q}_{\mathcal{D}}$ and the two corresponding matrices $B_{\mathcal{Q}}$ are given by

$$B_{\mathcal{Q}_{\mathcal{K},\mathcal{K}^*}} = \frac{1}{|\mathcal{Q}_{\mathcal{K},\mathcal{K}^*}|}\left(|\sigma_{\mathcal{K}}|\boldsymbol{\nu}^*\right) = \frac{2}{\sin\alpha_{\mathcal{D}}}\left(\frac{\boldsymbol{\nu}^*}{|\sigma_{\mathcal{K}^*}|}\right), \tag{4.7}$$

$$B_{\mathcal{Q}_{\mathcal{K},\mathcal{L}^*}} = \frac{1}{|\mathcal{Q}_{\mathcal{K},\mathcal{L}^*}|}\left(-|\sigma_{\mathcal{K}}|\boldsymbol{\nu}^*\right) = \frac{2}{\sin\alpha_{\mathcal{D}}}\left(-\frac{\boldsymbol{\nu}^*}{|\sigma_{\mathcal{L}^*}|}\right). \tag{4.8}$$

Notice that these matrices depend only on the geometry of the diamond cell $\mathcal{D}$. Furthermore we easily see from formulas (4.3)–(4.8) above that $\sum_{\mathcal{Q}\in\mathfrak{Q}_{\mathcal{D}}} |\mathcal{Q}| B_{\mathcal{Q}} = 0$ for any diamond cell $\mathcal{D}$. Hence, the following straightforward result holds:

LEMMA 4.2. *For all $\xi \in \mathbb{R}^2$, for all $\mathcal{D} \in \mathfrak{D}$, for all $\delta \in \mathbb{R}^{n_{\mathcal{D}}}$, we have*

$$\xi = \frac{1}{|\mathcal{D}|}\sum_{\mathcal{Q}\in\mathfrak{Q}_{\mathcal{D}}} |\mathcal{Q}|\left(\xi + B_{\mathcal{Q}}\delta\right). \tag{4.9}$$

This Lemma implies that the new gradient has a mean value over $\mathcal{D}$ which equals the usual DDFV gradient $\nabla^{\mathcal{T}}_{\mathcal{D}} u^{\mathcal{T}}$, that is,

$$\nabla^{\mathcal{T}}_{\mathcal{D}} u^{\mathcal{T}} = \frac{1}{|\mathcal{D}|}\sum_{\mathcal{Q}\in\mathfrak{Q}_{\mathcal{D}}} |\mathcal{Q}| \nabla^{\mathcal{N}}_{\mathcal{Q}} u^{\mathcal{T}},$$

which is the 2D equivalent to formula (2.5).

Like in the monodimensional case presented in section 2, we want to eliminate the additional unknowns $\delta^{\mathcal{D}}$ on each $\mathcal{D}$ in such a way that the conservativity of the numerical fluxes on all edges $\sigma \in \mathcal{E}_{\mathcal{D}}$ (the transmission condition) is ensured. More precisely, we want to choose $\delta^{\mathcal{D}}$ such that, setting $\xi = \nabla^{\mathcal{T}}_{\mathcal{D}} u^{\mathcal{T}}$, we have

$$\begin{cases} \left(\varphi_{\mathcal{Q}_{\mathcal{K},\mathcal{K}^*}}(\xi + B_{\mathcal{Q}_{\mathcal{K},\mathcal{K}^*}}\delta^{\mathcal{D}}), \boldsymbol{\nu}^*\right) = \left(\varphi_{\mathcal{Q}_{\mathcal{K},\mathcal{L}^*}}(\xi + B_{\mathcal{Q}_{\mathcal{K},\mathcal{L}^*}}\delta^{\mathcal{D}}), \boldsymbol{\nu}^*\right), \\ \left(\varphi_{\mathcal{Q}_{\mathcal{L},\mathcal{K}^*}}(\xi + B_{\mathcal{Q}_{\mathcal{L},\mathcal{K}^*}}\delta^{\mathcal{D}}), \boldsymbol{\nu}^*\right) = \left(\varphi_{\mathcal{Q}_{\mathcal{L},\mathcal{L}^*}}(\xi + B_{\mathcal{Q}_{\mathcal{L},\mathcal{L}^*}}\delta^{\mathcal{D}}), \boldsymbol{\nu}^*\right), \\ \left(\varphi_{\mathcal{Q}_{\mathcal{K},\mathcal{K}^*}}(\xi + B_{\mathcal{Q}_{\mathcal{K},\mathcal{K}^*}}\delta^{\mathcal{D}}), \boldsymbol{\nu}\right) = \left(\varphi_{\mathcal{Q}_{\mathcal{L},\mathcal{K}^*}}(\xi + B_{\mathcal{Q}_{\mathcal{L},\mathcal{K}^*}}\delta^{\mathcal{D}}), \boldsymbol{\nu}\right), \\ \left(\varphi_{\mathcal{Q}_{\mathcal{K},\mathcal{L}^*}}(\xi + B_{\mathcal{Q}_{\mathcal{K},\mathcal{L}^*}}\delta^{\mathcal{D}}), \boldsymbol{\nu}\right) = \left(\varphi_{\mathcal{Q}_{\mathcal{L},\mathcal{L}^*}}(\xi + B_{\mathcal{Q}_{\mathcal{L},\mathcal{L}^*}}\delta^{\mathcal{D}}), \boldsymbol{\nu}\right), \end{cases} \tag{4.10}$$

in the case where $\mathcal{D} \in \mathfrak{D}_{\text{int}}$. If $\mathcal{D} \in \mathfrak{D}_{\text{ext}}$, the conditions on $\delta^{\mathcal{D}}$ take the simpler form

$$\left(\varphi_{\mathcal{Q}_{\mathcal{K},\mathcal{K}^*}}(\xi + B_{\mathcal{Q}_{\mathcal{K},\mathcal{K}^*}}\delta^{\mathcal{D}}), \boldsymbol{\nu}^*\right) = \left(\varphi_{\mathcal{Q}_{\mathcal{K},\mathcal{L}^*}}(\xi + B_{\mathcal{Q}_{\mathcal{K},\mathcal{L}^*}}\delta^{\mathcal{D}}), \boldsymbol{\nu}^*\right). \tag{4.11}$$

We are now going to show that (4.10) or (4.11) uniquely define $\delta^{\mathcal{D}} \in \mathbb{R}^{n_{\mathcal{D}}}$ as a function of ξ.

PROPOSITION 4.3. *For all $\mathcal{D} \in \mathfrak{D}$ and all $\xi \in \mathbb{R}^2$, there exists a unique $\delta^{\mathcal{D}}(\xi) \in \mathbb{R}^{n_{\mathcal{D}}}$ such that* (4.10) *(resp.* (4.11)*) holds if $\mathcal{D} \in \mathfrak{D}_{\rm int}$ (resp. if $\mathcal{D} \in \mathfrak{D}_{\rm ext}$).*

Proof. We only give the proof for $\mathcal{D} \in \mathfrak{D}_{\rm int}$ (so that $n_{\mathcal{D}} = 4$), since the case of boundary diamond cells can be treated in the same way. Let $\xi \in \mathbb{R}^2$ given and define $F_\xi : \mathbb{R}^4 \mapsto \mathbb{R}^4$ by

$$F_\xi(\delta) = \sum_{\mathcal{Q} \in \mathfrak{Q}_{\mathcal{D}}} |\mathcal{Q}|^t B_{\mathcal{Q}}.\varphi_{\mathcal{Q}}(\xi + B_{\mathcal{Q}}\delta).$$

By using (4.3)–(4.6), we easily see that the conditions (4.10) are equivalent to the equation $F_\xi(\delta^{\mathcal{D}}) = 0$. Hence, the claim will be proved if we show that this nonlinear equation has a unique solution. To this end, we remark that for any $\widetilde{\delta} \in \mathbb{R}^4$, we have

$$\left(F_\xi(\delta), \tilde{\delta}\right) = \sum_{\mathcal{Q} \in \mathfrak{Q}_{\mathcal{D}}} |\mathcal{Q}| \left(\varphi_{\mathcal{Q}}(\xi + B_{\mathcal{Q}}\delta), B_{\mathcal{Q}}\widetilde{\delta}\right). \tag{4.12}$$

Hence, we deduce using properties $(\mathcal{P}_2^{\mathcal{T}})$ and $(\mathcal{P}_3^{\mathcal{T}})$ that there exists C depending only on p and C_φ such that

$$(F_\xi(\delta), \delta) \geq \frac{1}{C} \left(\sum_{\mathcal{Q} \in \mathfrak{Q}_{\mathcal{D}}} |\mathcal{Q}| \, |\xi + B_{\mathcal{Q}}\delta|^p \right) - C|\mathcal{D}|(1 + |\xi|^p). \tag{4.13}$$

Finally, we deduce

$$(F_\xi(\delta), \delta) \geq \frac{1}{C} \left(\sum_{\mathcal{Q} \in \mathfrak{Q}_{\mathcal{D}}} |\mathcal{Q}| \, |B_{\mathcal{Q}}\delta|^p \right) - C|\mathcal{D}|(1 + |\xi|^p),$$

for another constant C. Since $\sum_{\mathcal{Q} \in \mathfrak{Q}_{\mathcal{D}}} |\mathcal{Q}||B_{\mathcal{Q}}\delta|^p \underset{|\delta| \to \infty}{\longrightarrow} \infty$, we deduce that F_ξ is coercive. By the Brouwer theorem (F_ξ is continuous since each $\varphi_{\mathcal{Q}}$ is continuous) we obtain the existence of at least one solution to the problem $F_\xi(\delta) = 0$.

Notice now that if $B_{\mathcal{Q}}\delta = B_{\mathcal{Q}}\widetilde{\delta}, \forall \mathcal{Q} \in \mathfrak{Q}_{\mathcal{D}}$, then $\delta = \widetilde{\delta}$. Hence, we deduce from (4.12) that for all $\delta \neq \widetilde{\delta}$

$$\left(F_\xi(\delta) - F_\xi(\widetilde{\delta}), \delta - \widetilde{\delta}\right) = \sum_{\mathcal{Q} \in \mathfrak{Q}_{\mathcal{D}}} |\mathcal{Q}| \left(\varphi_{\mathcal{Q}}(\xi + B_{\mathcal{Q}}\delta) - \varphi_{\mathcal{Q}}(\xi + B_{\mathcal{Q}}\widetilde{\delta}), B_{\mathcal{Q}}\delta - B_{\mathcal{Q}}\widetilde{\delta}\right) > 0,$$

using property $(\mathcal{P}_{1'a}^{\mathcal{T}})$ or $(\mathcal{P}_{1'b}^{\mathcal{T}})$. This gives the uniqueness of the solution to $F_\xi(\delta) = 0$ and the claim is proved. ∎

Example 4.4. In many situations, it can happen that φ is smooth, for instance, in each primal control volume. In that case, it is possible (see Remark 4.1) to choose the approximations $\varphi_{\mathcal{Q}}$ in such a way that for each $\mathcal{D} \in \mathfrak{D}_{\rm int}$ we have

$$\varphi_{\mathcal{Q}_{\mathcal{K},\mathcal{K}^*}} = \varphi_{\mathcal{Q}_{\mathcal{K},\mathcal{L}^*}}, \quad \text{and } \varphi_{\mathcal{Q}_{\mathcal{L},\mathcal{K}^*}} = \varphi_{\mathcal{Q}_{\mathcal{L},\mathcal{L}^*}}.$$

Then, one can easily show that the solution $\delta^{\mathcal{D}}(\xi)$ of (4.10) appears to satisfy

$$\delta_{\mathcal{K}}^{\mathcal{D}} = 0, \quad \delta_{\mathcal{L}}^{\mathcal{D}} = 0, \quad \text{and } \delta_{\mathcal{K}^*}^{\mathcal{D}} = \delta_{\mathcal{L}^*}^{\mathcal{D}}.$$

Hence, everything happens like in the 1D case and there is in fact only one artificial unknown ($\delta_{\mathcal{K}^*}^{\mathcal{D}}$, for instance) to determine.

A symmetric situation holds if we assume that φ is smooth in each dual control volume and that $\varphi_{\mathfrak{Q}_{\mathcal{K},\mathcal{K}^*}} = \varphi_{\mathfrak{Q}_{\mathcal{L},\mathcal{K}^*}}$, $\varphi_{\mathfrak{Q}_{\mathcal{K},\mathcal{L}^*}} = \varphi_{\mathfrak{Q}_{\mathcal{L},\mathcal{L}^*}}$.

Finally, notice that if we only assume that $\varphi_{\mathfrak{Q}_{\mathcal{K},\mathcal{K}^*}} = \varphi_{\mathfrak{Q}_{\mathcal{K},\mathcal{L}^*}}$, for instance, then in the general nonlinear case there is no reason why $\delta^{\mathcal{D}}_{\mathcal{K}}$ should be 0.

From now on, the new discrete gradient operator $\nabla^{\mathcal{N}}$ is completely determined by

$$\nabla^{\mathcal{N}}_{\mathcal{Q}} u^{\mathcal{T}} = \nabla^{\mathcal{T}}_{\mathcal{D}} u^{\mathcal{T}} + B_{\mathcal{Q}} \delta^{\mathcal{D}}(\nabla^{\mathcal{T}}_{\mathcal{D}} u^{\mathcal{T}}), \text{ for any } \mathcal{D} \in \mathfrak{D} \text{ and } \mathcal{Q} \in \mathfrak{Q}_{\mathcal{D}}, \tag{4.14}$$

where the map $\xi \mapsto \delta^{\mathcal{D}}(\xi)$ is defined in Proposition 4.3. Notice that, in general, this new gradient operator $\nabla^{\mathcal{N}}_{\mathcal{Q}}$ is *nonlinear* contrarily to the operator $\nabla^{\mathcal{T}}_{\mathcal{D}}$ and depends on the flux φ defining the problem (1.1) under study (and more precisely to its approximations $\varphi_{\mathcal{Q}}$).

Furthermore, let us emphasize the fact that the nonlinear map $\xi \mapsto \delta^{\mathcal{D}}(\xi)$ is only defined implicitly through (4.10) or (4.11) (see also the 1D discussion in section 2), which seems to make the new discrete gradient $\nabla^{\mathcal{N}}$ quite difficult to compute. We postpone to section 7 the important discussion on practical ways to solve the scheme we will propose in the sequel.

4.4. Some useful inequalities. The usual DDFV discrete gradient and the modified one can be compared as follows.

LEMMA 4.5. *There exists a constant C that depends only on C_φ and p such that for all $\mathcal{D} \in \mathfrak{D}$, and all $u^{\mathcal{T}} \in \mathbb{R}^{\mathcal{T}}$, we have*

$$\int_{\mathcal{D}} |\nabla^{\mathcal{T}} u^{\mathcal{T}}(z)|^p \, dz \le \int_{\mathcal{D}} |\nabla^{\mathcal{N}} u^{\mathcal{T}}(z)|^p \, dz \le C \int_{\mathcal{D}} (1 + |\nabla^{\mathcal{T}} u^{\mathcal{T}}(z)|^p) \, dz.$$

Proof. Thanks to Lemma 4.2, we have

$$|\mathcal{D}||\nabla^{\mathcal{T}}_{\mathcal{D}} u^{\mathcal{T}}|^p = |\mathcal{D}| \left| \frac{1}{|\mathcal{D}|} \sum_{\mathcal{Q} \in \mathfrak{Q}_{\mathcal{D}}} |\mathcal{Q}| \nabla^{\mathcal{N}}_{\mathcal{Q}} u^{\mathcal{T}} \right|^p = |\mathcal{D}| \left| \frac{1}{|\mathcal{D}|} \int_{\mathcal{D}} \nabla^{\mathcal{N}} u^{\mathcal{T}}(z) \, dz \right|^p.$$

Using the Jensen inequality, we deduce the first inequality.

The second one is a consequence of (4.13) applied to $\xi = \nabla^{\mathcal{T}}_{\mathcal{D}} u^{\mathcal{T}}$, $\delta = \delta^{\mathcal{D}}(\nabla^{\mathcal{T}}_{\mathcal{D}} u^{\mathcal{T}})$ using that $F_{\nabla^{\mathcal{T}}_{\mathcal{D}} u^{\mathcal{T}}}(\delta^{\mathcal{D}}(\nabla^{\mathcal{T}}_{\mathcal{D}} u^{\mathcal{T}})) = 0$ by definition of $\delta^{\mathcal{D}}$. ∎

Finally we can state the following discrete Poincaré inequality. Its proof is sketched in [3] and uses an argument given in [4] (see also [10]).

PROPOSITION 4.6 (discrete Poincaré inequality). *Let $\mathcal{T}$ be a DDFV mesh on Ω. There exists a constant C depending on p, Ω, and $\mathrm{reg}(\mathcal{T})$ such that*

$$\|u^{\mathfrak{M}}\|_{L^p} + \|u^{\mathfrak{M}^*}\|_{L^p} \le C \|\nabla^{\mathcal{T}} u^{\mathcal{T}}\|_{L^p} \le C \|\nabla^{\mathcal{N}} u^{\mathcal{T}}\|_{L^p}, \quad \forall u^{\mathcal{T}} \in \mathbb{R}^{\mathcal{T}}.$$

In particular, if $\nabla^{\mathcal{N}}_{\mathcal{Q}} u^{\mathcal{T}} = 0$ for all $\mathcal{Q} \in \mathfrak{Q}$, then $u^{\mathcal{T}} = 0$.

4.5. The m-DDFV scheme for discontinuous fluxes. We can finally introduce the new approximate flux $\varphi^{\mathcal{N}}_{\mathcal{D}}$ on each diamond cell to be used in the finite volume scheme instead of (3.4). For any diamond cell $\mathcal{D} \in \mathfrak{D}$, we set

$$\varphi^{\mathcal{N}}_{\mathcal{D}}(\xi) = \frac{1}{|\mathcal{D}|} \sum_{\mathcal{Q} \in \mathfrak{Q}_{\mathcal{D}}} |\mathcal{Q}| \varphi_{\mathcal{Q}}(\xi + B_{\mathcal{Q}} \delta^{\mathcal{D}}(\xi)), \quad \forall \xi \in \mathbb{R}^2. \tag{4.15}$$

This definition is nothing but the adaptation to the 2D case of the corresponding formula in the 1D case (see (2.7)).

Applying definition (4.15) to $\xi = \nabla_{\mathcal{D}}^{\mathcal{T}} u^{\mathcal{T}}$ for a given $u^{\mathcal{T}} \in \mathbb{R}^{\mathcal{T}}$, we find that

$$\varphi_{\mathcal{D}}^{\mathcal{N}}(\nabla_{\mathcal{D}}^{\mathcal{T}} u^{\mathcal{T}}) = \frac{1}{|\mathcal{D}|} \sum_{\mathcal{Q} \in \mathfrak{Q}_{\mathcal{D}}} |\mathcal{Q}| \varphi_{\mathcal{Q}}(\nabla_{\mathcal{Q}}^{\mathcal{N}} u^{\mathcal{T}}). \tag{4.16}$$

The m-DDFV scheme that we will study in the sequel of the paper can now be defined by the set of equations

$$\begin{cases} \mathfrak{a}_{\mathcal{K}}^{\mathcal{N}}(u^{\mathcal{T}}) \stackrel{\text{def}}{=} - \displaystyle\sum_{\mathcal{D} \in \mathfrak{D}_{\mathcal{K}}} |\sigma| \left(\varphi_{\mathcal{D}}^{\mathcal{N}}(\nabla_{\mathcal{D}}^{\mathcal{T}} u^{\mathcal{T}}), \boldsymbol{\nu}_{\mathcal{K}}\right) = |\mathcal{K}| f_{\mathcal{K}}, \ \forall \mathcal{K} \in \mathfrak{M}, \\ \mathfrak{a}_{\mathcal{K}^*}^{\mathcal{N}}(u^{\mathcal{T}}) \stackrel{\text{def}}{=} - \displaystyle\sum_{\mathcal{D} \in \mathfrak{D}_{\mathcal{K}^*}} |\sigma^*| \left(\varphi_{\mathcal{D}}^{\mathcal{N}}(\nabla_{\mathcal{D}}^{\mathcal{T}} u^{\mathcal{T}}), \boldsymbol{\nu}_{\mathcal{K}^*}\right) = |\mathcal{K}^*| f_{\mathcal{K}^*}, \ \forall \mathcal{K}^* \in \mathfrak{M}^*, \end{cases} \tag{4.17}$$

that we write under the short form $\mathfrak{a}^{\mathcal{N}}(u^{\mathcal{T}}) = f^{\mathcal{T}}$, with $\mathfrak{a}^{\mathcal{N}}(\,\cdot\,) \stackrel{\text{def}}{=} ((\mathfrak{a}_{\mathcal{K}}^{\mathcal{N}}(\,\cdot\,))_{\mathcal{K}}, (\mathfrak{a}_{\mathcal{K}^*}^{\mathcal{N}}(\,\cdot\,))_{\mathcal{K}^*})$ and $f^{\mathcal{T}} = ((|\mathcal{K}| f_{\mathcal{K}})_{\mathcal{K}}, (|\mathcal{K}^*| f_{\mathcal{K}^*})_{\mathcal{K}^*})$. Note that the only difference between this scheme (4.17) and the previous one (3.3) is the fact that we replaced the previous mean-value approximation $\varphi_{\mathcal{D}}$ of φ over $\mathcal{D}$ by the map $\varphi_{\mathcal{D}}^{\mathcal{N}}$ defined by (4.15).

4.6. Basic properties of the scheme. Before proving existence and uniqueness of the solution for the nonlinear system (4.17), we first give some properties of the map $\mathfrak{a}^{\mathcal{N}}$. The first result is a *summation by parts* result showing that the m-DDFV scheme we propose enjoys the same discrete duality property than the original DDFV scheme.

LEMMA 4.7. *For any $u^{\mathcal{T}}$, $v^{\mathcal{T}}$ in $\mathbb{R}^{\mathcal{T}}$ we have*

$$\begin{aligned} (\mathfrak{a}^{\mathcal{N}}(u^{\mathcal{T}}), v^{\mathcal{T}}) &= 2 \sum_{\mathcal{D} \in \mathfrak{D}} |\mathcal{D}| \left(\varphi_{\mathcal{D}}^{\mathcal{N}}(\nabla_{\mathcal{D}}^{\mathcal{T}} u^{\mathcal{T}}), \nabla_{\mathcal{D}}^{\mathcal{T}} v^{\mathcal{T}}\right) \\ &= 2 \sum_{\mathcal{Q} \in \mathfrak{Q}} |\mathcal{Q}| \left(\varphi_{\mathcal{Q}}(\nabla_{\mathcal{Q}}^{\mathcal{N}} u^{\mathcal{T}}), \nabla_{\mathcal{Q}}^{\mathcal{N}} v^{\mathcal{T}}\right). \end{aligned}$$

Proof. The first equality can be proved in the same way than in [3] by reordering the summation on the primal and dual control volumes as a summation over the diamond set

$$\begin{aligned} &\sum_{\mathcal{K} \in \mathfrak{M}} \mathfrak{a}_{\mathcal{K}}^{\mathcal{N}}(u^{\mathcal{T}}) v_{\mathcal{K}} + \sum_{\mathcal{K}^* \in \mathfrak{M}^*} \mathfrak{a}_{\mathcal{K}^*}^{\mathcal{N}}(u^{\mathcal{T}}) v_{\mathcal{K}^*} \\ &\quad = - \sum_{\mathcal{K} \in \mathfrak{M}} \sum_{\mathcal{D} \in \mathfrak{D}_{\mathcal{K}}} |\sigma| (\varphi_{\mathcal{D}}^{\mathcal{N}}(\nabla_{\mathcal{D}}^{\mathcal{T}} u^{\mathcal{T}}), \boldsymbol{\nu}) v_{\mathcal{K}} - \sum_{\mathcal{K}^* \in \mathfrak{M}^*} \sum_{\mathcal{D} \in \mathfrak{D}_{\mathcal{K}^*}} |\sigma^*| (\varphi_{\mathcal{D}}^{\mathcal{N}}(\nabla_{\mathcal{D}}^{\mathcal{T}} u^{\mathcal{T}}), \boldsymbol{\nu}^*) v_{\mathcal{K}^*} \\ &\quad = \sum_{\mathcal{D} \in \mathfrak{D}_{\text{int}}} |\mathcal{D}| \left(\varphi_{\mathcal{D}}^{\mathcal{N}}(\nabla_{\mathcal{D}}^{\mathcal{T}} u^{\mathcal{T}}), \frac{2}{\sin \alpha_{\mathcal{D}}} \left(\frac{v_{\mathcal{K}} - v_{\mathcal{L}}}{|\sigma^*|} \boldsymbol{\nu} + \frac{v_{\mathcal{K}^*} - v_{\mathcal{L}^*}}{|\sigma|} \boldsymbol{\nu}^* \right) \right) \\ &\qquad + \sum_{\mathcal{D} \in \mathfrak{D}_{\text{ext}}} |\mathcal{D}| \left(\varphi_{\mathcal{D}}^{\mathcal{N}}(\nabla_{\mathcal{D}}^{\mathcal{T}} u^{\mathcal{T}}), \frac{2}{\sin \alpha_{\mathcal{D}}} \frac{v_{\mathcal{K}}}{|\sigma^*|} \boldsymbol{\nu} \right) \\ &\quad = 2 \sum_{\mathcal{D} \in \mathfrak{D}} |\mathcal{D}| \left(\varphi_{\mathcal{D}}^{\mathcal{N}}(\nabla_{\mathcal{D}}^{\mathcal{T}} u^{\mathcal{T}}), \nabla_{\mathcal{D}}^{\mathcal{T}} v^{\mathcal{T}}\right). \end{aligned}$$

To prove the second equality, we use (4.16) and (4.14) to write on each diamond cell $\mathcal{D} \in \mathfrak{D}$

$$\begin{aligned}
|\mathcal{D}|\,(\varphi_{\mathcal{D}}^{\mathcal{N}}(\nabla_{\mathcal{D}}^{\mathcal{T}} u^{\mathcal{T}}), \nabla_{\mathcal{D}}^{\mathcal{T}} v^{\mathcal{T}}) &= \sum_{\mathcal{Q} \in \mathfrak{Q}_{\mathcal{D}}} |\mathcal{Q}|\,(\varphi_{\mathcal{Q}}(\nabla_{\mathcal{Q}}^{\mathcal{N}} u^{\mathcal{T}}), \nabla_{\mathcal{D}}^{\mathcal{T}} v^{\mathcal{T}}) \\
&= \sum_{\mathcal{Q} \in \mathfrak{Q}_{\mathcal{D}}} |\mathcal{Q}|\,(\varphi_{\mathcal{Q}}(\nabla_{\mathcal{Q}}^{\mathcal{N}} u^{\mathcal{T}}), \nabla_{\mathcal{Q}}^{\mathcal{N}} v^{\mathcal{T}} - B_{\mathcal{Q}} \delta^{\mathcal{D}}(\nabla_{\mathcal{D}}^{\mathcal{T}} v^{\mathcal{T}})) \\
&= \sum_{\mathcal{Q} \in \mathfrak{Q}_{\mathcal{D}}} |\mathcal{Q}|\,(\varphi_{\mathcal{Q}}(\nabla_{\mathcal{Q}}^{\mathcal{N}} u^{\mathcal{T}}), \nabla_{\mathcal{Q}}^{\mathcal{N}} v^{\mathcal{T}}) \\
&\qquad - \sum_{\mathcal{Q} \in \mathfrak{Q}_{\mathcal{D}}} |\mathcal{Q}|^t B_{\mathcal{Q}} \varphi_{\mathcal{Q}}(\nabla_{\mathcal{Q}}^{\mathcal{N}} u^{\mathcal{T}}) \cdot \delta^{\mathcal{D}}(\nabla_{\mathcal{D}}^{\mathcal{T}} v^{\mathcal{T}}),
\end{aligned}$$

and this last term vanishes by definition of the map $\delta^{\mathcal{D}}$ (see proposition 4.3). Indeed, we have

$$\sum_{\mathcal{Q} \in \mathfrak{Q}_{\mathcal{D}}} |\mathcal{Q}|^t B_{\mathcal{Q}} \varphi_{\mathcal{Q}}(\nabla_{\mathcal{Q}}^{\mathcal{N}} u^{\mathcal{T}}) \cdot \delta^{\mathcal{D}}(\nabla_{\mathcal{D}}^{\mathcal{T}} v^{\mathcal{T}}) = (F_{\nabla_{\mathcal{D}}^{\mathcal{T}} u^{\mathcal{T}}}(\delta^{\mathcal{D}}(\nabla_{\mathcal{D}}^{\mathcal{T}} u^{\mathcal{T}})), \delta^{\mathcal{D}}(v^{\mathcal{T}})) = 0. \qquad \square$$

LEMMA 4.8. *For any $u^{\mathcal{T}} \in \mathbb{R}^{\mathcal{T}}$ we have*

$$(\mathfrak{a}^{\mathcal{N}}(u^{\mathcal{T}}), u^{\mathcal{T}}) \geq \frac{2}{C_\varphi} \|\nabla^{\mathcal{N}} u^{\mathcal{T}}\|_{L^p}^p - 2C_\varphi |\Omega| \geq \frac{2}{C_\varphi} \|\nabla^{\mathcal{T}} u^{\mathcal{T}}\|_{L^p}^p - 2C_\varphi |\Omega|.$$

Proof. We derive from Lemma 4.7 and property $(\mathcal{P}_2^{\mathcal{T}})$ that

$$\begin{aligned}
(\mathfrak{a}^{\mathcal{N}}(u^{\mathcal{T}}), u^{\mathcal{T}}) &= 2 \sum_{\mathcal{D} \in \mathfrak{D}} \sum_{\mathcal{Q} \in \mathfrak{Q}_{\mathcal{D}}} |\mathcal{Q}|\,(\varphi_{\mathcal{Q}}(\nabla_{\mathcal{Q}}^{\mathcal{N}} u^{\mathcal{T}}), \nabla_{\mathcal{Q}}^{\mathcal{N}} u^{\mathcal{T}}) \\
&\geq 2 \sum_{\mathcal{D} \in \mathfrak{D}} \left(\int_{\mathcal{D}} \frac{1}{C_\varphi} |\nabla^{\mathcal{N}} u^{\mathcal{T}}(z)|^p \, dz - C_\varphi |\mathcal{D}| \right).
\end{aligned}$$

We conclude using Lemma 4.5. $\square$

LEMMA 4.9. *We have*

$$(\mathfrak{a}^{\mathcal{N}}(u^{\mathcal{T}}) - \mathfrak{a}^{\mathcal{N}}(v^{\mathcal{T}}), u^{\mathcal{T}} - v^{\mathcal{T}}) > 0, \quad \forall u^{\mathcal{T}}, v^{\mathcal{T}} \in \mathbb{R}^{\mathcal{T}}, \ u^{\mathcal{T}} \neq v^{\mathcal{T}}.$$

Proof. From Lemma 4.7, we have

$$\begin{aligned}
&(\mathfrak{a}^{\mathcal{N}}(u^{\mathcal{T}}) - \mathfrak{a}^{\mathcal{N}}(v^{\mathcal{T}}), u^{\mathcal{T}} - v^{\mathcal{T}}) \\
&\qquad = 2 \sum_{\mathcal{Q} \in \mathfrak{Q}} |\mathcal{Q}|\,(\varphi_{\mathcal{Q}}(\nabla_{\mathcal{Q}}^{\mathcal{N}} u^{\mathcal{T}}) - \varphi_{\mathcal{Q}}(\nabla_{\mathcal{Q}}^{\mathcal{N}} v^{\mathcal{T}}), \nabla_{\mathcal{Q}}^{\mathcal{N}} u^{\mathcal{T}} - \nabla_{\mathcal{Q}}^{\mathcal{N}} v^{\mathcal{T}}).
\end{aligned} \tag{4.18}$$

By using the monotonicity properties of the nonlinearity $\varphi_{\mathcal{Q}}$, that is, properties $(\mathcal{P}_{1'a}^{\mathcal{T}})$ and $(\mathcal{P}_{1'b}^{\mathcal{T}})$, we deduce that the left-hand side of (4.18) is nonnegative and vanishes if and only if $u^{\mathcal{T}} = v^{\mathcal{T}}$ (by Proposition 4.6). $\square$

THEOREM 4.10. *The scheme* (4.17) *admits a unique solution $u^{\mathcal{T}} \in \mathbb{R}^{\mathcal{T}}$. Furthermore, there exists a constant C depending only on C_φ,* $\mathrm{reg}(\mathcal{T})$*, and p such that this solution $u^{\mathcal{T}}$ satisfies*

$$\|\nabla^{\mathcal{T}} u^{\mathcal{T}}\|_{L^p} + \|\nabla^{\mathcal{N}} u^{\mathcal{T}}\|_{L^p} \leq C \left(1 + \|f\|_{L^{p'}}^{\frac{1}{p-1}} \right). \tag{4.19}$$

Proof. The map $u^{\mathcal{T}} \mapsto \mathfrak{a}^{\mathcal{N}}(u^{\mathcal{T}}) - f^{\mathcal{T}}$ is continuous and coercive thanks to Lemma 4.8. We deduce from the Brouwer theorem the existence of a solution to (4.17). Uniqueness is a consequence of the monotonicity Lemma 4.9.

Estimate (4.19) comes directly from Lemmas 4.5 and 4.8. ∎

We finally show that the numerical solution of the scheme (4.17) depends continuously on the source term f.

THEOREM 4.11 (stability). *There exists a constant $C > 0$ depending only on C_φ, $\mathrm{reg}(\mathcal{T})$, and p such that for any $f, g \in L^{p'}(\Omega)$, we have*

$$\|\nabla^{\mathcal{N}} u^{\mathcal{T}} - \nabla^{\mathcal{N}} v^{\mathcal{T}}\|_{L^p} \leq \begin{cases} C\left(1 + \|f\|_{L^{p'}} + \|g\|_{L^{p'}}\right)^{\frac{2-p}{p-1}} \|f - g\|_{L^{p'}}, & \text{if } 1 < p \leq 2 \\ C\|f - g\|_{L^{p'}}^{\frac{1}{p-1}}, & \text{if } p > 2, \end{cases}$$

where $u^{\mathcal{T}}$ (resp. $v^{\mathcal{T}}$) is the solution of the m-DDFV scheme (4.17) *associated to the data f (resp. g).*

Proof. We apply estimate (4.18) and obtain thanks to property $(\mathcal{P}^{\mathcal{T}}_{1'b})$ for $p > 2$

$$(\mathfrak{a}^{\mathcal{N}}(u^{\mathcal{T}}) - \mathfrak{a}^{\mathcal{N}}(v^{\mathcal{T}}), u^{\mathcal{T}} - v^{\mathcal{T}}) \geq \frac{1}{C} \sum_{\mathcal{D} \in \mathfrak{D}} \sum_{\mathcal{Q} \in \mathfrak{Q}_{\mathcal{D}}} |\mathcal{Q}| |\nabla^{\mathcal{N}}_{\mathcal{Q}} u^{\mathcal{T}} - \nabla^{\mathcal{N}}_{\mathcal{Q}} v^{\mathcal{T}}|^p.$$

For $1 < p \leq 2$, property $(\mathcal{P}^{\mathcal{T}}_{1'a})$ implies

$$(\mathfrak{a}^{\mathcal{N}}(u^{\mathcal{T}}) - \mathfrak{a}^{\mathcal{N}}(v^{\mathcal{T}}), u^{\mathcal{T}} - v^{\mathcal{T}}) \geq \frac{1}{C} \sum_{\mathcal{D} \in \mathfrak{D}} \sum_{\mathcal{Q} \in \mathfrak{Q}_{\mathcal{D}}} |\mathcal{Q}| \left(1 + |\nabla^{\mathcal{N}}_{\mathcal{Q}} u^{\mathcal{T}}|^p + |\nabla^{\mathcal{N}}_{\mathcal{Q}} v^{\mathcal{T}}|^p\right)^{\frac{p-2}{p}} |\nabla^{\mathcal{N}}_{\mathcal{Q}} u^{\mathcal{T}} - \nabla^{\mathcal{N}}_{\mathcal{Q}} v^{\mathcal{T}}|^2.$$

Hence, if $1 < p \leq 2$ we have

$$\|\nabla^{\mathcal{N}} u^{\mathcal{T}} - \nabla^{\mathcal{N}} v^{\mathcal{T}}\|^2_{L^p} \leq C\left(1 + \|\nabla^{\mathcal{N}} u^{\mathcal{T}}\|^{2-p} + \|\nabla^{\mathcal{N}} v^{\mathcal{T}}\|^{2-p}\right) \times \left(\mathfrak{a}^{\mathcal{N}}(u^{\mathcal{T}}) - \mathfrak{a}^{\mathcal{N}}(v^{\mathcal{T}}), u^{\mathcal{T}} - v^{\mathcal{T}}\right).$$

We conclude using the definition of $f^{\mathcal{T}}$ and $g^{\mathcal{T}}$, the discrete Poincaré inequality (Proposition 4.6) and the *a priori* estimate (4.19). ∎

5. Error estimates. We first give some error estimates for the scheme (4.17) in the case where the exact solution u_e of the problem (1.1) is piecewise smooth. More precisely let us introduce for any $q \in [1, +\infty]$ the space

$$W^{2,q}(\mathfrak{Q}) = \left\{ u \in W^{1,p}_0(\Omega), \quad u_{|\mathcal{Q}} \in W^{2,q}(\mathcal{Q}), \ \forall \mathcal{Q} \in \mathfrak{Q} \right\},$$

endowed with the norm

$$\|u\|_{W^{2,q}(\mathfrak{Q})} = \|u\|_{W^{1,p}_0(\Omega)} + \left(\sum_{\mathcal{Q} \in \mathfrak{Q}} \|D^2 u\|^q_{L^q(\mathcal{Q})} \right)^{\frac{1}{q}}.$$

Our first result gives an error estimate for the solution of the m-DDFV scheme. In this result the flux φ is allowed to have discontinuities across all the edges of the primal and the dual meshes.

THEOREM 5.1. *Let $\mathcal{T}$ be a mesh on Ω. Let $f \in L^{p'}(\Omega)$ and assume that the solution u_e to* (1.1) *belongs to $W^{2,p}(\mathfrak{Q})$.*

There exists $C > 0$ depending on $\|u_e\|_{W^{2,p}(\mathfrak{Q})}$, $\mathrm{reg}(\mathcal{T})$, $\|f\|_{L^{p'}}$, C_φ, and p such that the solution $u^{\mathcal{T}} \in \mathbb{R}^{\mathcal{T}}$ of the m-DDFV scheme (4.17) *satisfies*

$$\|u_e - u^{\mathfrak{M}}\|_{L^p} + \|u_e - u^{\mathfrak{M}^*}\|_{L^p} + \|\nabla u_e - \nabla^{\mathcal{N}} u^{\mathcal{T}}\|_{L^p} \leq \begin{cases} C\,\mathrm{size}(\mathcal{T})^{(p-1)^2}, & \text{if } 1 < p \leq 2, \\ C\,\mathrm{size}(\mathcal{T})^{\frac{1}{(p-1)^2}}, & \text{if } p > 2. \end{cases}$$

A typical case for which our method can be applied is the one where the domain Ω can be divided into N disjoint subdomains $(\Omega_i)_{1\leq i\leq N}$ such that

$$\varphi \text{ is smooth over each subdomain } \Omega_i. \tag{5.1}$$

We assume that each domain Ω_i is polygonal and that the mesh is compatible with the subdomains in the sense that, for any i there exists a subset $\mathcal{E}_i$ of $\mathcal{E}$ such that $\partial\Omega_i = \cup_{\sigma\in\mathcal{E}_i}\sigma$. More generally, we may assume that the discontinuities of the flux φ occur only along a finite number of curves in Ω.

In that situation, the diamond cells naturally divide into two subsets defined by

$$\mathfrak{D}_{\mathrm{cont}} = \{\mathcal{D} \in \mathfrak{D},\ \ \exists i \in \{1,\dots,N\},\ \mathcal{D} \subset \Omega_i\},\ \text{ and }\ \mathfrak{D}_{\mathrm{disc}} = \mathfrak{D} \setminus \mathfrak{D}_{\mathrm{cont}}.$$

We propose to use in that case a hybrid DDFV scheme defined as follows:

DEFINITION 5.2. *Under assumption* (5.1)*, we call h-DDFV scheme for the problem* (1.1) *the DDFV scheme still under the form* (4.17)*, but where $\varphi_{\mathcal{D}}^{\mathcal{N}}$ is defined as follows:*

- *For the diamond cells $\mathcal{D} \in \mathfrak{D}_{\mathrm{disc}}$, that is, the ones where the discontinuities of the flux occur, we take $\varphi_{\mathcal{D}}^{\mathcal{N}}$ as defined in* (4.16).
- *For the diamond cells $\mathcal{D} \in \mathfrak{D}_{\mathrm{cont}}$, that is, away from the discontinuities, we take $\varphi_{\mathcal{D}}^{\mathcal{N}}$ to be the usual mean-value $\varphi_{\mathcal{D}}$ of φ over $\mathcal{D}$ defined in* (3.4).

Assuming that u_e is slightly more regular than in the previous result, this approach let us recover the same convergence rate than in the usual continuous flux case (see [3]).

THEOREM 5.3. *Consider the same assumptions than in Theorem* 5.1 *with the additional assumption* (5.1)*. Assume furthermore that the solution u_e of* (1.1) *lies in $W^{2,q}(\mathfrak{Q})$ for $q = p(p-1)^2$ if $p \geq 2$ and $q = \frac{p}{(p-1)^2}$ if $p < 2$. Then, there exists a constant $C > 0$ like in Theorem* 5.1 *such that the solution $u^{\mathcal{T}}$ to the h-DDFV scheme satisfies*

$$\|u_e - u^{\mathfrak{M}}\|_{L^p} + \|u_e - u^{\mathfrak{M}^*}\|_{L^p} + \|\nabla u_e - \nabla^{\mathcal{N}} u^{\mathcal{T}}\|_{L^p} \leq \begin{cases} C\,\mathrm{size}(\mathcal{T})^{p-1}, & \text{if } 1 < p \leq 2, \\ C\,\mathrm{size}(\mathcal{T})^{\frac{1}{p-1}}, & \text{if } p > 2. \end{cases}$$

We are now going to prove these two results. The key ingredients in this analysis are the consistency error estimates on the numerical fluxes across edges of the primal and the dual control volumes.

5.1. Consistency error. In order to evaluate the error between u_e and $u^{\mathcal{T}}$ we need to introduce a projection of the exact solution u_e onto the space of discrete functions $\mathbb{R}^{\mathcal{T}}$. Notice that any function v in $W^{2,p}(\mathfrak{Q})$ is continuous over $\overline{\Omega}$. Hence, it makes sense to consider the center-value projection $\mathbb{P}^{\mathcal{T}}$ defined as follows:

DEFINITION 5.4. *For any $v \in \mathcal{C}^0(\overline{\Omega})$, we define its center-value projection $\mathbb{P}^{\mathcal{T}} v \in \mathbb{R}^{\mathcal{T}}$ as the vector*

$$\mathbb{P}^{\mathcal{T}} v = \Big((v(x_{\mathcal{K}}))_{\mathcal{K}\in\mathfrak{M}}, (v(x_{\mathcal{K}^*}))_{\mathcal{K}^*\in\mathfrak{M}^*} \Big).$$

We refer to [3] for the proofs of the main properties of this projection operator.

As usual in finite volume methods, the error analysis is mainly based on estimates of consistency errors for the fluxes as defined below.

DEFINITION 5.5. *Assume that* $u_e \in W^{2,p}(\mathfrak{D})$. *For any* $\mathcal{Q} \in \mathfrak{D}$, $z \in \overline{\mathcal{Q}}$ *we define*

$$\begin{aligned} R_{\mathcal{Q}}(z) &= \varphi_{\big|\overline{\mathcal{Q}}}(z, \nabla u_e{\big|}_{\overline{\mathcal{Q}}}(z)) - \varphi_{\mathcal{Q}}(\nabla_{\mathcal{Q}}^{\mathcal{N}} \mathbb{P}^{\mathcal{T}} u_e) \\ &= R_{\mathcal{Q}}^{\varphi}(z) + R_{\mathcal{Q}}^{\mathrm{grad}}, \end{aligned}$$

with

$$\begin{aligned} R_{\mathcal{Q}}^{\varphi}(z) &\overset{\mathrm{def}}{=} \varphi_{\big|\overline{\mathcal{Q}}}(z, \nabla u_e{\big|}_{\overline{\mathcal{Q}}}(z)) - \frac{1}{|\mathcal{Q}|} \int_{\mathcal{Q}} \varphi(z', \nabla u_e(z'))\, dz' \\ R_{\mathcal{Q}}^{\mathrm{grad}} &\overset{\mathrm{def}}{=} \frac{1}{|\mathcal{Q}|} \int_{\mathcal{Q}} \left(\varphi(z', \nabla u_e(z')) - \varphi(z', \nabla_{\mathcal{Q}}^{\mathcal{N}} \mathbb{P}^{\mathcal{T}} u_e)\right)\, dz'. \end{aligned}$$

Finally, for any $\mathcal{Q} \in \mathfrak{D}$ *and* $\sigma \in \mathcal{E}_{\mathcal{Q}}$, *we note*

$$R_{\mathcal{Q},\sigma}^{\varphi} = \frac{1}{|\sigma|} \int_{\sigma} (R_{\mathcal{Q}}^{\varphi}(z), \boldsymbol{\nu}_{\mathcal{Q},\sigma})\, dz, \quad R_{\mathcal{Q},\sigma}^{\mathrm{grad}} = \left(R_{\mathcal{Q}}^{\mathrm{grad}}, \boldsymbol{\nu}_{\mathcal{Q},\sigma}\right),$$

and $R_{\mathcal{Q},\sigma} = R_{\mathcal{Q},\sigma}^{\mathrm{grad}} + R_{\mathcal{Q},\sigma}^{\varphi}$, *where* $\boldsymbol{\nu}_{\mathcal{Q},\sigma}$ *is the unit normal to* σ *pointing outward* $\mathcal{Q}$.

It is fundamental to notice that, by definition of the new discrete gradient operator $\nabla^{\mathcal{N}}$ and since $u_e \in W^{2,p}(\mathfrak{D})$, we have the conservativity property

$$R_{\mathcal{Q},\sigma} = -R_{\mathcal{Q}',\sigma}, \text{ if } \sigma = \mathcal{Q}|\mathcal{Q}'. \tag{5.2}$$

The goal is now to estimate each of the two terms involved in this consistency error. The term $R_{\mathcal{Q}}^{\varphi}$ can be controlled by using exactly the same techniques as in [3, Proposition 7.10]. With the present notations, and thanks to assumptions $(\mathcal{H}_{4a})$–$(\mathcal{H}_{4b})$ and $(\mathcal{H}_{5a})$–$(\mathcal{H}_{5b})$, the result reads:

PROPOSITION 5.6. *Let* $\mathcal{T}$ *be a mesh on* Ω *and assume that the solution* u_e *to problem* (1.1) *lies in* $W^{2,p}(\mathfrak{D})$. *There exists a constant* $C > 0$ *depending on* p, $\mathrm{reg}(\mathcal{T})$, *and* C_{φ} *such that*

$$|\mathcal{Q}|\, |R_{\mathcal{Q},\sigma}^{\varphi}|^{\frac{p}{p-1}} \leq C \mathrm{d}_{\mathcal{Q}}^{p\alpha_p} \int_{\mathcal{Q}} (1 + |\nabla u_e|^p + |D^2 u_e|^p)\, dz, \quad \forall \mathcal{Q} \in \mathfrak{D}, \forall \sigma \in \mathcal{E}_{\mathcal{Q}},$$

where $\alpha_p = \max(1, \frac{1}{p-1})$.

The estimate of the term $R_{\mathcal{Q},\sigma}^{\mathrm{grad}}$ will follow from a consistency estimate for the new gradient operator $\nabla^{\mathcal{N}}$ that we have introduced. This is the main difference between the present study and our previous works since the definition of the new discrete gradient depends on the jumps of φ in each diamond cell. Hence, the consistency estimate for this operator cannot be obtained as in the usual way, that is, only by applying well-chosen Taylor formulas. The proof of the estimate is much more involved and makes use of the bound obtained in Proposition 5.6.

We also want to point out the fact that, when $p \neq 2$, we do not obtain the usual first-order consistency estimate as we obtained in [3] for the operator $\nabla_{\mathcal{D}}^{\mathcal{T}}$. Indeed, due to the degeneracy/singularity of the nonlinear operator near the origin, we only recover a consistency property of order less than one.

PROPOSITION 5.7. *Let* $\mathcal{T}$ *be a mesh on* Ω *and assume that the solution* u_e *to problem* (1.1) *lies in* $W^{2,p}(\mathfrak{D})$. *There exists a constant* $C > 0$ *depending on* $p, \mathrm{reg}(\mathcal{T})$,

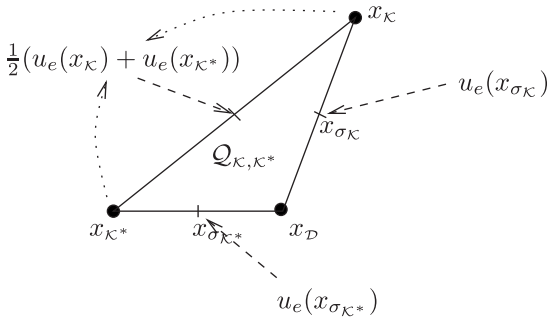


FIG. 5.1. *The affine interpolation* $\mathbb{P}^{\mathfrak{Q}}_{\mathcal{Q}_{\mathcal{K},\mathcal{K}^*}}$ *on the quarter diamond* $\mathcal{Q}_{\mathcal{K},\mathcal{K}^*}$.

and C_φ *such that for any* $\mathcal{D} \in \mathfrak{D}$ *we have*

$$\int_{\mathcal{D}} |\nabla u_e(z) - \nabla^{\mathcal{N}} \mathbb{P}^{\mathcal{T}} u_e(z)|^p \, dz \leq C \mathrm{d}_{\mathcal{D}}^{p(p-1)\alpha_p^2} \sum_{\mathcal{Q} \in \mathfrak{Q}_{\mathcal{D}}} \int_{\mathcal{Q}} \left(1 + |\nabla u_e|^p + |D^2 u_e|^p\right) \, dz,$$

where α_p *is given in Proposition* 5.6.

Proof. Let us give the proof in the case where $\mathcal{D}$ is an interior diamond cell. The case $\mathcal{D} \in \mathfrak{D}_{\mathrm{ext}}$ can be treated in a similar, and in fact, simpler way.

Let us define the projection $\mathbb{P}^{\mathfrak{Q}} u_e$ of u_e on the set of quarter diamonds as follows. For each quarter diamond $\mathcal{Q} \in \mathfrak{Q}$, the restriction of $\mathbb{P}^{\mathfrak{Q}} u_e$ to the triangle $\mathcal{Q}$ is the unique affine function $\mathbb{P}^{\mathfrak{Q}}_{\mathcal{Q}} u_e$ which coincides with u_e at the middle of the semi-edges $\sigma \in \mathcal{E}_{\mathcal{Q}}$ and whose value at the middle of the third side of $\mathcal{Q}$ is the mean value of the values of u_e at the extremities of this side. Notice that this definition makes sense since $u_e\big|_{\overline{\mathcal{Q}}} \in W^{2,p}(\mathcal{Q}) \subset \mathcal{C}^0(\overline{\mathcal{Q}})$. As an example, in the case of the quarter diamond $\mathcal{Q} = \mathcal{Q}_{\mathcal{K},\mathcal{K}^*}$ (see Figure 5.1), this definition reads

$$\begin{aligned}
\mathbb{P}^{\mathfrak{Q}}_{\mathcal{Q}_{\mathcal{K},\mathcal{K}^*}} u_e(x_{\sigma_{\mathcal{K}}}) &= u_e(x_{\sigma_{\mathcal{K}}}), \\
\mathbb{P}^{\mathfrak{Q}}_{\mathcal{Q}_{\mathcal{K},\mathcal{K}^*}} u_e(x_{\sigma_{\mathcal{K}^*}}) &= u_e(x_{\sigma_{\mathcal{K}^*}}), \\
\mathbb{P}^{\mathfrak{Q}}_{\mathcal{Q}_{\mathcal{K},\mathcal{K}^*}} u_e\left(\frac{x_{\mathcal{K}} + x_{\mathcal{K}^*}}{2}\right) &= \frac{u_e(x_{\mathcal{K}}) + u_e(x_{\mathcal{K}^*})}{2}.
\end{aligned}$$

The gradient of $\mathbb{P}^{\mathfrak{Q}}_{\mathcal{Q}_{\mathcal{K},\mathcal{K}^*}} u_e$ is then given by

$$\nabla \mathbb{P}^{\mathfrak{Q}}_{\mathcal{Q}_{\mathcal{K},\mathcal{K}^*}} u_e = \frac{2}{\sin \alpha_{\mathcal{D}}} \left(\frac{u_e(x_{\sigma_{\mathcal{K}^*}}) - \frac{u_e(x_{\mathcal{K}}) + u_e(x_{\mathcal{K}^*})}{2}}{|\sigma_{\mathcal{K}}|} \boldsymbol{\nu} + \frac{u_e(x_{\sigma_{\mathcal{K}}}) - \frac{u_e(x_{\mathcal{K}}) + u_e(x_{\mathcal{K}^*})}{2}}{|\sigma_{\mathcal{K}^*}|} \boldsymbol{\nu}^* \right).$$

Let us now define the consistency error for this projection $\mathbb{P}^{\mathfrak{Q}}$ as follows:

$$T_{\overline{\mathcal{Q}}}(z) = \nabla u_e(z) - \nabla \mathbb{P}^{\mathfrak{Q}}_{\mathcal{Q}} u_e, \quad \forall z \in \mathcal{Q}, \quad \forall \mathcal{Q} \in \mathfrak{Q}. \tag{5.3}$$

By usual Taylor expansions inside each quarter diamond $\mathcal{Q}$ (see [3], for instance), we can easily show that there exists a constant $C > 0$ as in the claim on the proposition such that

$$\int_{\mathcal{Q}} |T_{\overline{\mathcal{Q}}}(z)|^p \, dz \leq C \mathrm{d}_{\mathcal{Q}}^p \int_{\mathcal{Q}} |D^2 u_e(z)|^p \, dz, \quad \forall \mathcal{Q} \in \mathfrak{Q}. \tag{5.4}$$

By the discussion of section 4.3 we remark that, $\mathcal{D}$ being an interior diamond cell, there exists $\widetilde{\delta}_{\mathcal{D}} \in \mathbb{R}^4$ such that

$$\nabla \mathbb{P}^{\mathfrak{Q}}_{\mathcal{Q}} u_e - \nabla^{\mathcal{T}}_{\mathcal{D}} \mathbb{P}^{\mathcal{T}} u_e = B_{\mathcal{Q}} \widetilde{\delta}_{\mathcal{D}}, \, \forall \mathcal{Q} \in \mathfrak{Q}_{\mathcal{D}},$$

and then, by the definition (4.14) of $\nabla^{\mathcal{N}}_{\mathcal{Q}}$, we deduce that there exists $\overline{\delta^{\mathcal{D}}} \in \mathbb{R}^4$ such that

$$\nabla \mathbb{P}^{\mathfrak{Q}}_{\mathcal{Q}} u_e - \nabla^{\mathcal{N}}_{\mathcal{Q}} \mathbb{P}^{\mathcal{T}} u_e = B_{\mathcal{Q}} \overline{\delta^{\mathcal{D}}}, \, \forall \mathcal{Q} \in \mathfrak{Q}_{\mathcal{D}}. \tag{5.5}$$

Since u_e solves (1.1) with $f \in L^{p'}(\Omega)$, we know that the following transmission property holds (just like the fourth equation in (2.1)):

$$\int_{\sigma_{\mathcal{K}}} \varphi|_{\overline{\mathcal{Q}_{\mathcal{K},\mathcal{K}^*}}}(z, \nabla u_e|_{\overline{\mathcal{Q}_{\mathcal{K},\mathcal{K}^*}}}(s)) \cdot \boldsymbol{\nu}^* \, ds = \int_{\sigma_{\mathcal{K}}} \varphi|_{\overline{\mathcal{Q}_{\mathcal{K},\mathcal{L}^*}}}(z, \nabla u_e|_{\overline{\mathcal{Q}_{\mathcal{K},\mathcal{L}^*}}}(s)) \cdot \boldsymbol{\nu}^* \, ds. \tag{5.6}$$

Recall that the gradient operator $\nabla^{\mathcal{N}}$ is built to ensure that the discrete equivalent of this property, that is, the first equation of (4.10), holds. It follows that

$$\begin{aligned}
&\left(\frac{1}{|\sigma_{\mathcal{K}}|} \int_{\sigma_{\mathcal{K}}} \varphi|_{\overline{\mathcal{Q}_{\mathcal{K},\mathcal{K}^*}}}(s, \nabla u_e|_{\overline{\mathcal{Q}_{\mathcal{K},\mathcal{K}^*}}}(s)) \, ds - \varphi_{\mathcal{Q}_{\mathcal{K},\mathcal{K}^*}}(\nabla^{\mathcal{N}}_{\mathcal{Q}_{\mathcal{K},\mathcal{K}^*}} \mathbb{P}^{\mathcal{T}} u_e), \boldsymbol{\nu}^* \right) \\
&\qquad - \left(\frac{1}{|\sigma_{\mathcal{K}}|} \int_{\sigma_{\mathcal{K}}} \varphi|_{\overline{\mathcal{Q}_{\mathcal{K},\mathcal{L}^*}}}(s, \nabla u_e|_{\overline{\mathcal{Q}_{\mathcal{K},\mathcal{L}^*}}}(s)) \, ds - \varphi_{\mathcal{Q}_{\mathcal{K},\mathcal{L}^*}}(\nabla^{\mathcal{N}}_{\mathcal{Q}_{\mathcal{K},\mathcal{L}^*}} \mathbb{P}^{\mathcal{T}} u_e), \boldsymbol{\nu}^* \right) = 0.
\end{aligned}$$

By using Definition 5.5, we get

$$\begin{aligned}
&\left(\frac{1}{|\mathcal{Q}_{\mathcal{K},\mathcal{K}^*}|} \int_{\mathcal{Q}_{\mathcal{K},\mathcal{K}^*}} \Big(\varphi(z, \nabla u_e(z)) - \varphi(z, \nabla^{\mathcal{N}}_{\mathcal{Q}_{\mathcal{K},\mathcal{K}^*}} \mathbb{P}^{\mathcal{T}} u_e) \Big) \, dz, \boldsymbol{\nu}^* \right) \\
&- \left(\frac{1}{|\mathcal{Q}_{\mathcal{K},\mathcal{L}^*}|} \int_{\mathcal{Q}_{\mathcal{K},\mathcal{L}^*}} \Big(\varphi(z, \nabla u_e(z)) - \varphi(z, \nabla^{\mathcal{N}}_{\mathcal{Q}_{\mathcal{K},\mathcal{L}^*}} \mathbb{P}^{\mathcal{T}} u_e) \Big) \, dz, \boldsymbol{\nu}^* \right) = R^{\varphi}_{\mathcal{Q}_{\mathcal{K},\mathcal{K}^*},\sigma_{\mathcal{K}}} - R^{\varphi}_{\mathcal{Q}_{\mathcal{K},\mathcal{L}^*},\sigma_{\mathcal{K}}}.
\end{aligned}$$

Similarly we obtain for the other three semi-edges in the diamond under study the following relations

$$\begin{aligned}
&\left(\frac{1}{|\mathcal{Q}_{\mathcal{L},\mathcal{K}^*}|} \int_{\mathcal{Q}_{\mathcal{L},\mathcal{K}^*}} \Big(\varphi(z, \nabla u_e(z)) - \varphi(z, \nabla^{\mathcal{N}}_{\mathcal{Q}_{\mathcal{L},\mathcal{K}^*}} \mathbb{P}^{\mathcal{T}} u_e) \Big) \, dz, \boldsymbol{\nu}^* \right) \\
&- \left(\frac{1}{|\mathcal{Q}_{\mathcal{L},\mathcal{L}^*}|} \int_{\mathcal{Q}_{\mathcal{L},\mathcal{L}^*}} \Big(\varphi(z, \nabla u_e(z)) \, dz - \varphi(z, \nabla^{\mathcal{N}}_{\mathcal{Q}_{\mathcal{L},\mathcal{L}^*}} \mathbb{P}^{\mathcal{T}} u_e) \Big) \, dz, \boldsymbol{\nu}^* \right) = R^{\varphi}_{\mathcal{Q}_{\mathcal{L},\mathcal{K}^*},\sigma_{\mathcal{L}}} - R^{\varphi}_{\mathcal{Q}_{\mathcal{L},\mathcal{L}^*},\sigma_{\mathcal{L}}}.
\end{aligned}$$

$$\begin{aligned}
&\left(\frac{1}{|\mathcal{Q}_{\mathcal{K},\mathcal{K}^*}|} \int_{\mathcal{Q}_{\mathcal{K},\mathcal{K}^*}} \Big(\varphi(z, \nabla u_e(z)) - \varphi(z, \nabla^{\mathcal{N}}_{\mathcal{Q}_{\mathcal{K},\mathcal{K}^*}} \mathbb{P}^{\mathcal{T}} u_e) \Big) \, dz, \boldsymbol{\nu} \right) \\
&- \left(\frac{1}{|\mathcal{Q}_{\mathcal{L},\mathcal{K}^*}|} \int_{\mathcal{Q}_{\mathcal{L},\mathcal{K}^*}} \Big(\varphi(z, \nabla u_e(z)) - \varphi(z, \nabla^{\mathcal{N}}_{\mathcal{Q}_{\mathcal{L},\mathcal{K}^*}} \mathbb{P}^{\mathcal{T}} u_e) \Big) \, dz, \boldsymbol{\nu} \right) = R^{\varphi}_{\mathcal{Q}_{\mathcal{K},\mathcal{K}^*},\sigma_{\mathcal{K}^*}} - R^{\varphi}_{\mathcal{Q}_{\mathcal{L},\mathcal{K}^*},\sigma_{\mathcal{K}^*}}.
\end{aligned}$$

$$\begin{aligned}
&\left(\frac{1}{|\mathcal{Q}_{\mathcal{K},\mathcal{L}^*}|} \int_{\mathcal{Q}_{\mathcal{K},\mathcal{L}^*}} \Big(\varphi(z, \nabla u_e(z)) - \varphi(z, \nabla^{\mathcal{N}}_{\mathcal{Q}_{\mathcal{K},\mathcal{L}^*}} \mathbb{P}^{\mathcal{T}} u_e) \Big) \, dz, \boldsymbol{\nu} \right) \\
&- \left(\frac{1}{|\mathcal{Q}_{\mathcal{L},\mathcal{L}^*}|} \int_{\mathcal{Q}_{\mathcal{L},\mathcal{L}^*}} \Big(\varphi(z, \nabla u_e(z)) - \varphi(z, \nabla^{\mathcal{N}}_{\mathcal{Q}_{\mathcal{L},\mathcal{L}^*}} \mathbb{P}^{\mathcal{T}} u_e) \Big) \, dz, \boldsymbol{\nu} \right) = R^{\varphi}_{\mathcal{Q}_{\mathcal{K},\mathcal{L}^*},\sigma_{\mathcal{L}^*}} - R^{\varphi}_{\mathcal{Q}_{\mathcal{L},\mathcal{L}^*},\sigma_{\mathcal{L}^*}}.
\end{aligned}$$

Multiplying these equations, respectively, by $|\sigma_{\mathcal{K}}|\overline{\delta_{\mathcal{K}}}$, $|\sigma_{\mathcal{L}}|\overline{\delta_{\mathcal{L}}}$, $|\sigma_{\mathcal{K}^*}|\overline{\delta_{\mathcal{K}^*}}$, and $|\sigma_{\mathcal{L}^*}|\overline{\delta_{\mathcal{L}^*}}$ and summing, we obtain

$$\sum_{\mathcal{Q}\in\mathfrak{Q}_{\mathcal{D}}}\int_{\mathcal{Q}}\left(\varphi(z,\nabla u_e(z))-\varphi(z,\nabla_{\mathcal{Q}}^{\mathcal{N}}\mathbb{P}^{\mathcal{T}}u_e),B_{\mathcal{Q}}\overline{\delta^{\mathcal{D}}}\right)\,dz\le\sum_{\mathcal{Q}\in\mathfrak{Q}_{\mathcal{D}}}|\mathcal{Q}||B_{\mathcal{Q}}\overline{\delta^{\mathcal{D}}}|\sum_{\sigma\in\mathcal{E}_{\mathcal{Q}}}|R^{\varphi}_{\mathcal{Q},\sigma}|,$$

where we used the definitions (4.3)–(4.6). Using (5.3) and (5.5) we finally deduce

$$\begin{aligned}(5.7)\qquad &\sum_{\mathcal{Q}\in\mathfrak{Q}_{\mathcal{D}}}\int_{\mathcal{Q}}(\varphi(z,\nabla u_e(z))-\varphi(z,\nabla_{\mathcal{Q}}^{\mathcal{N}}\mathbb{P}^{\mathcal{T}}u_e),\nabla u_e(z)-\nabla_{\mathcal{Q}}^{\mathcal{N}}\mathbb{P}^{\mathcal{T}}u_e)\,dz\\&\le\sum_{\mathcal{Q}\in\mathfrak{Q}_{\mathcal{D}}}\left(\int_{\mathcal{Q}}|\nabla u_e(z)-\nabla_{\mathcal{Q}}^{\mathcal{N}}\mathbb{P}^{\mathcal{T}}u_e(z)|\,dz+\int_{\mathcal{Q}}|T_{\overline{\mathcal{Q}}}(z)|\,dz\right)\sum_{\sigma\in\mathcal{E}_{\mathcal{Q}}}|R^{\varphi}_{\mathcal{Q},\sigma}|\\&\qquad+\sum_{\mathcal{Q}\in\mathfrak{Q}_{\mathcal{D}}}\int_{\mathcal{Q}}(\varphi(z,\nabla u_e(z))-\varphi(z,\nabla_{\mathcal{Q}}^{\mathcal{N}}\mathbb{P}^{\mathcal{T}}u_e),T_{\overline{\mathcal{Q}}}(z))\,dz.\end{aligned}$$

In the case $p>2$, using assumptions $(\mathcal{H}_{1'b})$ and $(\mathcal{H}_{4b})$ and Young's inequality, we deduce from formula (5.7) that

$$\begin{aligned}\sum_{\mathcal{Q}\in\mathfrak{Q}_{\mathcal{D}}}\int_{\mathcal{Q}}|\nabla u_e(z)-\nabla_{\mathcal{Q}}^{\mathcal{N}}\mathbb{P}^{\mathcal{T}}u_e|^p\,dz&\le C\sum_{\mathcal{Q}\in\mathfrak{Q}_{\mathcal{D}}}\sum_{\sigma\in\mathcal{E}_{\mathcal{Q}}}|\mathcal{Q}||R^{\varphi}_{\mathcal{Q},\sigma}|^{\frac{p}{p-1}}+C\sum_{\mathcal{Q}\in\mathfrak{Q}_{\mathcal{D}}}\int_{\mathcal{Q}}|T_{\overline{\mathcal{Q}}}(z)|^p\,dz\\&\quad+C\left(\sum_{\mathcal{Q}\in\mathfrak{Q}_{\mathcal{D}}}\int_{\mathcal{Q}}|T_{\overline{\mathcal{Q}}}(z)|^p\,dz\right)^{\frac{1}{p-1}}\left(\int_{\mathcal{D}}(1+|\nabla u_e(z)|^p)\,dz\right)^{\frac{p-2}{p-1}}.\end{aligned}$$

From (5.4) and the estimates in Proposition 5.6, it follows that

$$\sum_{\mathcal{Q}\in\mathfrak{Q}_{\mathcal{D}}}\int_{\mathcal{Q}}|\nabla u_e(z)-\nabla_{\mathcal{Q}}^{\mathcal{N}}\mathbb{P}^{\mathcal{T}}u_e|^p\,dz\le C\mathrm{d}_{\mathcal{D}}^{\frac{p}{p-1}}\sum_{\mathcal{Q}\in\mathfrak{Q}_{\mathcal{D}}}\int_{\mathcal{Q}}(1+|\nabla u_e|^p+|D^2u_e|^p)\,dz,$$

and the claim is proved.

In the case $1 < p \le 2$, using assumptions $(\mathcal{H}_{1'a})$ and $(\mathcal{H}_{4a})$, we deduce from formula (5.7) that

$$\begin{aligned}
&\sum_{\mathcal{Q}\in\mathfrak{Q}_\mathcal{D}} \int_\mathcal{Q} |\nabla u_e(z) - \nabla_\mathcal{Q}^\mathcal{N} \mathbb{P}^\mathcal{T} u_e|^p \, dz \\
&\le \Bigg[\sum_{\mathcal{Q}\in\mathfrak{Q}_\mathcal{D}} \left(\int_\mathcal{Q} |\nabla u_e(z) - \nabla_\mathcal{Q}^\mathcal{N} \mathbb{P}^\mathcal{T} u_e(z)| \, dz + \int_\mathcal{Q} |T_{\overline{\mathcal{Q}}}(z)| \, dz \right) \sum_{\sigma\in\mathcal{E}_\mathcal{Q}} |R_{\mathcal{Q},\sigma}^\varphi| \\
&\qquad\qquad + \sum_{\mathcal{Q}\in\mathfrak{Q}_\mathcal{D}} \int_\mathcal{Q} \left(\varphi(z, \nabla u_e(z)) - \varphi(z, \nabla_\mathcal{Q}^\mathcal{N} \mathbb{P}^\mathcal{T} u_e), T_{\overline{\mathcal{Q}}}(z)\right) dz \Bigg]^{\frac{p}{2}} \\
&\times \left(\sum_{\mathcal{Q}\in\mathfrak{Q}_\mathcal{D}} \int_\mathcal{Q} (1 + |\nabla u_e(z)|^p + |\nabla_\mathcal{Q}^\mathcal{N} \mathbb{P}^\mathcal{T} u_e|^p) \, dz \right)^{\frac{2-p}{2}} \\
&\le C \Bigg[\left(\left(\sum_{\mathcal{Q}\in\mathfrak{Q}_\mathcal{D}} \int_\mathcal{Q} |\nabla u_e(z) - \nabla_\mathcal{Q}^\mathcal{N} \mathbb{P}^\mathcal{T} u_e(z)|^p \, dz \right)^{\frac{1}{2}} + \left(\sum_{\mathcal{Q}\in\mathfrak{Q}_\mathcal{D}} \int_\mathcal{Q} |T_{\overline{\mathcal{Q}}}(z)|^p \, dz \right)^{\frac{1}{2}} \right) \\
&\qquad\qquad \times \left(\sum_{\mathcal{Q}\in\mathfrak{Q}_\mathcal{D}} |\mathcal{Q}| \sum_{\sigma\in\mathcal{E}_\mathcal{Q}} |R_{\mathcal{Q},\sigma}^\varphi|^{\frac{p}{p-1}} \right)^{\frac{p-1}{2}} \\
&+ \left(\sum_{\mathcal{Q}\in\mathfrak{Q}_\mathcal{D}} \int_\mathcal{Q} |\nabla u_e(z) - \nabla_\mathcal{Q}^\mathcal{N} \mathbb{P}^\mathcal{T} u_e|^p \right)^{\frac{p-1}{2}} \left(\sum_{\mathcal{Q}\in\mathfrak{Q}_\mathcal{D}} \int_\mathcal{Q} |T_{\overline{\mathcal{Q}}}(z)|^p \, dz \right)^{\frac{1}{2}} \Bigg] \\
&\qquad \times \left(\sum_{\mathcal{Q}\in\mathfrak{Q}_\mathcal{D}} \int_\mathcal{Q} (1 + |\nabla u_e(z)|^p + |\nabla_\mathcal{Q}^\mathcal{N} \mathbb{P}^\mathcal{T} u_e|^p) \, dz \right)^{\frac{2-p}{2}}.
\end{aligned}$$

Using Young's inequality, the bound (5.4) and the estimates in Proposition 5.6, the claim follows. ∎

We can now estimate the consistency error of the scheme due to the approximation of the gradient as follows.

PROPOSITION 5.8. *Let $\mathcal{T}$ be a mesh on Ω and assume that u_e lies in $W^{2,p}(\mathfrak{Q})$. There exists a constant $C > 0$ depending on $p, \mathrm{reg}(\mathcal{T})$, and C_φ such that for any $\mathcal{D} \in \mathfrak{D}$ we have*

$$\sum_{\mathcal{Q}\in\mathfrak{Q}_\mathcal{D}} |\mathcal{Q}| |R_\mathcal{Q}^{\mathrm{grad}}|^{\frac{p}{p-1}} \le C \mathrm{d}_\mathcal{D}^{p(p-1)\alpha_p^3} \sum_{\mathcal{Q}\in\mathfrak{Q}_\mathcal{D}} \int_\mathcal{Q} \left(1 + |\nabla u_e|^p + |D^2 u_e|^p\right) dz.$$

Proof. In the case $1 < p \le 2$, using assumption $(\mathcal{H}_{4a})$ and the consistency estimate of Proposition 5.7, we deduce that

$$\begin{aligned}
\sum_{\mathcal{Q}\in\mathfrak{Q}_\mathcal{D}} |\mathcal{Q}| |R_\mathcal{Q}^{\mathrm{grad}}|^{\frac{p}{p-1}} &\le C_\varphi^{\frac{p}{p-1}} \sum_{\mathcal{Q}\in\mathfrak{Q}_\mathcal{D}} |\mathcal{Q}| \left(\frac{1}{|\mathcal{Q}|} \int_\mathcal{Q} |\nabla u_e(z) - \nabla_\mathcal{Q}^\mathcal{N} \mathbb{P}^\mathcal{T} u_e|^{p-1} \, dz \right)^{\frac{p}{p-1}} \\
&\le C_\varphi^{\frac{p}{p-1}} \sum_{\mathcal{Q}\in\mathfrak{Q}_\mathcal{D}} \int_\mathcal{Q} |\nabla u_e(z) - \nabla_\mathcal{Q}^\mathcal{N} \mathbb{P}^\mathcal{T} u_e|^p \, dz \\
&\le C \mathrm{d}_\mathcal{D}^{p(p-1)\alpha_p^2} \sum_{\mathcal{Q}\in\mathfrak{Q}_\mathcal{D}} \int_\mathcal{Q} \left(1 + |\nabla u_e|^p + |D^2 u_e|^p\right) dz,
\end{aligned}$$

which gives the claim since $\alpha_p = 1$ as soon as $1 < p \le 2$.

When $p > 2$, by Jensen's and Hölder's inequality, we have

$$
\begin{aligned}
&\sum_{\mathcal{Q}\in\mathfrak{Q}_{\mathcal{D}}} |\mathcal{Q}||R_{\mathcal{Q}}^{\mathrm{grad}}|^{\frac{p}{p-1}} \\
&\quad\le C_{\varphi}^{\frac{p}{p-1}} \sum_{\mathcal{Q}\in\mathfrak{Q}_{\mathcal{D}}} |\mathcal{Q}| \Bigg(\frac{1}{|\mathcal{Q}|} \int_{\mathcal{Q}} \left(1 + |\nabla u_e(z)|^{p-2} + |\nabla_{\mathcal{Q}}^{\mathcal{N}} \mathbb{P}^{\mathcal{T}} u_e|^{p-2}\right) \\
&\qquad\qquad \times |\nabla u_e(z) - \nabla_{\mathcal{Q}}^{\mathcal{N}} \mathbb{P}^{\mathcal{T}} u_e| \, dz \Bigg)^{\frac{p}{p-1}} \\
&\quad\le C \left(\sum_{\mathcal{Q}\in\mathfrak{Q}_{\mathcal{D}}} \int_{\mathcal{Q}} \left(1 + |\nabla u_e(z)|^p + |\nabla_{\mathcal{Q}}^{\mathcal{N}} \mathbb{P}^{\mathcal{T}} u_e|^p\right) \, dz \right)^{\frac{p-2}{p-1}} \\
&\qquad\qquad \times \left(\sum_{\mathcal{Q}\in\mathfrak{Q}_{\mathcal{D}}} \int_{\mathcal{Q}} |\nabla u_e(z) - \nabla_{\mathcal{Q}}^{\mathcal{N}} \mathbb{P}^{\mathcal{T}} u_e|^p \, dz \right)^{\frac{1}{p-1}} \\
&\quad\le C \mathrm{d}_{\mathcal{D}}^{p\alpha_p^2} \sum_{\mathcal{Q}\in\mathfrak{Q}_{\mathcal{D}}} \int_{\mathcal{Q}} \left(1 + |\nabla u_e|^p + |D^2 u_e|^p\right) \, dz,
\end{aligned}
\tag{5.8}
$$

and we conclude by noting that $p\alpha_p^2 = p(p-1)\alpha_p^3$ when $p > 2$. ∎

5.2. Proof of Theorem 5.1. We have

$$
\|\nabla u_e - \nabla^{\mathcal{N}} u^{\mathcal{T}}\|_{L^p} \le \|\nabla u_e - \nabla^{\mathcal{N}} \mathbb{P}^{\mathcal{T}} u_e\|_{L^p} + \|\nabla^{\mathcal{N}} \mathbb{P}^{\mathcal{T}} u_e - \nabla^{\mathcal{N}} u^{\mathcal{T}}\|_{L^p}.
$$

Proposition 5.7 gives a bound for the first term. We come back to the proof of the stability Theorem 4.11 to evaluate the second one. That proof shows that the estimate of $\|\nabla^{\mathcal{N}} \mathbb{P}^{\mathcal{T}} u_e - \nabla^{\mathcal{N}} u^{\mathcal{T}}\|_{L^p}$ requires the control of

$$
I \overset{\mathrm{def}}{=} \left(\mathfrak{a}^{\mathcal{N}}(\mathbb{P}^{\mathcal{T}} u_e) - \mathfrak{a}^{\mathcal{N}}(u^{\mathcal{T}}), \mathbb{P}^{\mathcal{T}} u_e - u^{\mathcal{T}}\right).
$$

By classical manipulations (using the conservativity of numerical fluxes) we express I through the consistency errors thanks to

$$
\begin{aligned}
\mathfrak{a}_{\mathcal{K}}^{\mathcal{N}}(u^{\mathcal{T}}) - \mathfrak{a}_{\mathcal{K}}^{\mathcal{N}}(\mathbb{P}^{\mathcal{T}} u_e) &= \sum_{\mathcal{Q}\subset\mathcal{K}} \sum_{\sigma\in\mathcal{E}_{\mathcal{Q}}\cap\partial\mathcal{K}} |\sigma| R_{\mathcal{Q},\sigma}, \ \forall \mathcal{K} \in \mathfrak{M} \\
\mathfrak{a}_{\mathcal{K}^*}^{\mathcal{N}}(u^{\mathcal{T}}) - \mathfrak{a}_{\mathcal{K}^*}^{\mathcal{N}}(\mathbb{P}^{\mathcal{T}} u_e) &= \sum_{\mathcal{Q}\subset\mathcal{K}^*} \sum_{\sigma\in\mathcal{E}_{\mathcal{Q}}\cap\partial\mathcal{K}^*} |\sigma| R_{\mathcal{Q},\sigma}, \ \forall \mathcal{K}^* \in \mathfrak{M}^*.
\end{aligned}
$$

If we define the error $e^{\mathcal{T}} = u^{\mathcal{T}} - \mathbb{P}^{\mathcal{T}} u_e$, whose components are denoted by $((e_{\mathcal{K}})_{\mathcal{K}}, (e_{\mathcal{K}^*})_{\mathcal{K}^*})$, the formulas above yield

$$
I = \sum_{\mathcal{K}\in\mathfrak{M}} \sum_{\mathcal{Q}\subset\mathcal{K}} \sum_{\sigma\in\mathcal{E}_{\mathcal{Q}}\cap\partial\mathcal{K}} |\sigma| R_{\mathcal{Q},\sigma} e_{\mathcal{K}} + \sum_{\mathcal{K}^*\in\mathfrak{M}^*} \sum_{\mathcal{Q}\subset\mathcal{K}^*} \sum_{\sigma\in\mathcal{E}_{\mathcal{Q}}\cap\partial\mathcal{K}^*} |\sigma| R_{\mathcal{Q},\sigma} e_{\mathcal{K}^*}.
$$

Reordering the sum over the diamond cells, we find that

$$
\begin{aligned}
I = \sum_{\mathcal{D}\in\mathfrak{D}} \Big(&|\sigma_{\mathcal{K}^*}| (R_{\mathcal{Q}_{\mathcal{K},\mathcal{K}^*,\sigma_{\mathcal{K}^*}}} e_{\mathcal{K}} + R_{\mathcal{Q}_{\mathcal{L},\mathcal{K}^*,\sigma_{\mathcal{K}^*}}} e_{\mathcal{L}}) \\
&+ |\sigma_{\mathcal{L}^*}| (R_{\mathcal{Q}_{\mathcal{K},\mathcal{L}^*,\sigma_{\mathcal{L}^*}}} e_{\mathcal{K}} + R_{\mathcal{Q}_{\mathcal{L},\mathcal{L}^*,\sigma_{\mathcal{L}^*}}} e_{\mathcal{L}}) \\
&+ |\sigma_{\mathcal{K}}| (R_{\mathcal{Q}_{\mathcal{K},\mathcal{K}^*,\sigma_{\mathcal{K}}}} e_{\mathcal{K}^*} + R_{\mathcal{Q}_{\mathcal{K},\mathcal{L}^*,\sigma_{\mathcal{K}}}} e_{\mathcal{L}^*}) \\
&+ |\sigma_{\mathcal{L}}| (R_{\mathcal{Q}_{\mathcal{L},\mathcal{K}^*,\sigma_{\mathcal{K}}}} e_{\mathcal{K}^*} + R_{\mathcal{Q}_{\mathcal{L},\mathcal{L}^*,\sigma_{\mathcal{K}}}} e_{\mathcal{L}^*}) \Big).
\end{aligned}
\tag{5.9}
$$

Using the conservativity property (5.2), the first term in the sum above reads

$$
\begin{aligned}
(5.10)\quad & |\sigma_{\mathcal{K}^*}|(R_{\mathcal{Q}_{\mathcal{K},\mathcal{K}^*,\sigma_{\mathcal{K}^*}}} e_{\mathcal{K}} + R_{\mathcal{Q}_{\mathcal{L},\mathcal{K}^*,\sigma_{\mathcal{K}^*}}} e_{\mathcal{L}}) \\
&= -|\mathcal{Q}_{\mathcal{K},\mathcal{K}^*}| R_{\mathcal{Q}_{\mathcal{K},\mathcal{K}^*,\sigma_{\mathcal{K}^*}}} \frac{2}{\sin\alpha_{\mathcal{D}}} \frac{e_{\mathcal{L}} - e_{\mathcal{K}}}{|\sigma_{\mathcal{K}}| + |\sigma_{\mathcal{L}}|} + |\mathcal{Q}_{\mathcal{L},\mathcal{K}^*}| R_{\mathcal{Q}_{\mathcal{L},\mathcal{K}^*,\sigma_{\mathcal{K}^*}}} \frac{2}{\sin\alpha_{\mathcal{D}}} \frac{e_{\mathcal{L}} - e_{\mathcal{K}}}{|\sigma_{\mathcal{K}}| + |\sigma_{\mathcal{L}}|} \\
&= -|\mathcal{Q}_{\mathcal{K},\mathcal{K}^*}| R_{\mathcal{Q}_{\mathcal{K},\mathcal{K}^*,\sigma_{\mathcal{K}^*}}} (\nabla_{\mathcal{D}}^{\mathcal{T}} e^{\mathcal{T}}, \boldsymbol{\tau}^*) + |\mathcal{Q}_{\mathcal{L},\mathcal{K}^*}| R_{\mathcal{Q}_{\mathcal{L},\mathcal{K}^*,\sigma_{\mathcal{K}^*}}} (\nabla_{\mathcal{D}}^{\mathcal{T}} e^{\mathcal{T}}, \boldsymbol{\tau}^*).
\end{aligned}
$$

We remark now, by using (4.3) and (4.6), that for any $\delta \in \mathbb{R}^4$ we have

$$
|\mathcal{Q}_{\mathcal{K},\mathcal{K}^*}| \left(B_{\mathcal{Q}_{\mathcal{K},\mathcal{K}^*}} \delta, \boldsymbol{\tau}^*\right) = |\sigma_{\mathcal{K}^*}| \sin\alpha_{\mathcal{D}} \delta_{\mathcal{K}^*} = -|\mathcal{Q}_{\mathcal{L},\mathcal{K}^*}| \left(B_{\mathcal{Q}_{\mathcal{L},\mathcal{K}^*}} \delta, \boldsymbol{\tau}^*\right).
$$

Hence by using once more the conservativity property and the definition (4.14), we can replace $\nabla_{\mathcal{D}}^{\mathcal{T}} e^{\mathcal{T}}$ by the corresponding $\nabla_{\mathcal{Q}}^{\mathcal{N}} e^{\mathcal{T}}$ in the right-hand side of (5.10). It follows

$$
\begin{aligned}
&|\sigma_{\mathcal{K}^*}|(R_{\mathcal{Q}_{\mathcal{K},\mathcal{K}^*,\sigma_{\mathcal{K}^*}}} e_{\mathcal{K}} + R_{\mathcal{Q}_{\mathcal{L},\mathcal{K}^*,\sigma_{\mathcal{K}^*}}} e_{\mathcal{L}}) \\
&\qquad = -|\mathcal{Q}_{\mathcal{K},\mathcal{K}^*}| R_{\mathcal{Q}_{\mathcal{K},\mathcal{K}^*,\sigma_{\mathcal{K}^*}}} (\nabla_{\mathcal{Q}_{\mathcal{K},\mathcal{K}^*}}^{\mathcal{N}} e^{\mathcal{T}}, \boldsymbol{\tau}^*) + |\mathcal{Q}_{\mathcal{L},\mathcal{K}^*}| R_{\mathcal{Q}_{\mathcal{L},\mathcal{K}^*,\sigma_{\mathcal{K}^*}}} (\nabla_{\mathcal{Q}_{\mathcal{L},\mathcal{K}^*}}^{\mathcal{N}} e^{\mathcal{T}}, \boldsymbol{\tau}^*).
\end{aligned}
$$

The other terms in (5.9) being treated in the same way, it follows that

$$
\begin{aligned}
I &\le C \sum_{\mathcal{Q}\in\mathfrak{Q}} \sum_{\sigma\in\mathcal{E}_{\mathcal{Q}}} |\mathcal{Q}| |R_{\mathcal{Q},\sigma}| |\nabla_{\mathcal{Q}}^{\mathcal{N}} e^{\mathcal{T}}| \\
&\le C \left(\sum_{\mathcal{Q}\in\mathfrak{Q}} \sum_{\sigma\in\mathcal{E}_{\mathcal{Q}}} |\mathcal{Q}| |R_{\mathcal{Q},\sigma}|^{\frac{p}{p-1}} \right)^{\frac{p-1}{p}} \|\nabla^{\mathcal{N}} e^{\mathcal{T}}\|_{L^p}.
\end{aligned}
$$

Using properties $(\mathcal{P}_{1'a}^{\mathcal{T}})$ and $(\mathcal{P}_{1'b}^{\mathcal{T}})$, we derive that

$$
\begin{aligned}
\|\nabla^{\mathcal{N}} e^{\mathcal{T}}\|_{L^p}^2 &\le C(1 + \|\nabla^{\mathcal{N}} u^{\mathcal{T}}\|_{L^p}^{2-p} + \|\nabla^{\mathcal{N}} \mathbb{P}^{\mathcal{T}} u_e\|_{L^p}^{2-p}) \\
&\quad \times \left(\sum_{\mathcal{Q}\in\mathfrak{Q}} \sum_{\sigma\in\mathcal{E}_{\mathcal{Q}}} |\mathcal{Q}| |R_{\mathcal{Q},\sigma}|^{\frac{p}{p-1}} \right)^{\frac{p-1}{p}} \|\nabla^{\mathcal{N}} e^{\mathcal{T}}\|_{L^p}, \quad \text{if } 1 < p \le 2,
\end{aligned}
$$

and

$$
(5.11)\qquad \|\nabla^{\mathcal{N}} e^{\mathcal{T}}\|_{L^p}^p \le C \left(\sum_{\mathcal{Q}\in\mathfrak{Q}} \sum_{\sigma\in\mathcal{E}_{\mathcal{Q}}} |\mathcal{Q}| |R_{\mathcal{Q},\sigma}|^{\frac{p}{p-1}} \right)^{\frac{p-1}{p}} \|\nabla^{\mathcal{N}} e^{\mathcal{T}}\|_{L^p}, \quad \text{if } p > 2.
$$

The claim follows by using the estimate (4.19) and Propositions 5.6, 5.7, and 5.8.

5.3. Proof of Theorem 5.3. We give only the proof in the case $p > 2$ since the other case can be treated in the same way. We come back to (5.11) which is still valid for the h-DDFV scheme. It follows

$$
\|\nabla^{\mathcal{N}} e^{\mathcal{T}}\|_{L^p} \le C \left(\sum_{\mathcal{Q}\in\mathfrak{Q}} \sum_{\sigma\in\mathcal{E}_{\mathcal{Q}}} |\mathcal{Q}| |R_{\mathcal{Q},\sigma}|^{\frac{p}{p-1}} \right)^{\frac{1}{p}}.
$$

The terms $R_{\mathcal{Q},\sigma}$ contain the respective contributions of $R^{\varphi}_{\mathcal{Q},\sigma}$ and $R^{\mathrm{grad}}_{\mathcal{Q}}$. As far as $R^{\varphi}_{\mathcal{Q},\sigma}$ is concerned, the estimate of Proposition 5.6 is still valid for the hybrid scheme so that

$$\left(\sum_{\mathcal{Q}\in\mathfrak{Q}}\sum_{\sigma\in\mathcal{E}_{\mathcal{Q}}}|\mathcal{Q}||R^{\varphi}_{\mathcal{Q},\sigma}|^{\frac{p}{p-1}}\right)^{\frac{1}{p}}\leq C\mathrm{size}(\mathcal{T})^{\frac{1}{p-1}}\|u_e\|_{W^{2,p}(\mathfrak{Q})}.$$

We split now the contribution of $R^{\mathrm{grad}}_{\mathcal{Q}}$ into two parts: the one coming from diamond cells in $\mathfrak{D}_{\mathrm{cont}}$ where the usual DDFV approximate flux is used and the one coming from the diamond cells in $\mathfrak{D}_{\mathrm{disc}}$ where we used our new discrete gradient and flux. It follows

$$\left(\sum_{\mathcal{Q}\in\mathfrak{Q}}|\mathcal{Q}||R^{\mathrm{grad}}_{\mathcal{Q}}|^{\frac{p}{p-1}}\right)^{\frac{1}{p}}\leq C\left(\sum_{\mathcal{D}\in\mathfrak{D}_{\mathrm{cont}}}\sum_{\mathcal{Q}\in\mathfrak{Q}_{\mathcal{D}}}|\mathcal{Q}||R^{\mathrm{grad}}_{\mathcal{Q}}|^{\frac{p}{p-1}}\right)^{\frac{1}{p}}+C\left(\sum_{\mathcal{D}\in\mathfrak{D}_{\mathrm{disc}}}\sum_{\mathcal{Q}\in\mathfrak{Q}_{\mathcal{D}}}|\mathcal{Q}||R^{\mathrm{grad}}_{\mathcal{Q}}|^{\frac{p}{p-1}}\right)^{\frac{1}{p}}. \tag{5.12}$$

Near the discontinuities of the flux, that is for each diamond cell $\mathcal{D}\in\mathfrak{D}_{\mathrm{disc}}$, we use the estimate (5.8). Away from the discontinuities, i.e., for $\mathcal{D}\in\mathfrak{D}_{\mathrm{cont}}$ we used the usual DDFV scheme (that is $\nabla^{\mathcal{N}}=\nabla^{\mathcal{T}}$), so that the gradient consistency estimate given by (5.7) reads, since $u_e\in W^{2,p}(\mathcal{D})$:

$$\int_{\mathcal{D}}|\nabla u_e(z)-\nabla^{\mathcal{T}}\mathbb{P}^{\mathcal{T}}u_e(z)|^p\,dz\leq C\mathrm{d}_{\mathcal{D}}^p\int_{\mathcal{D}}\left(1+|\nabla u_e|^p+|D^2u_e|^p\right)\,dz.$$

This estimate is proved, for instance, in [3, Lemma 7.5]. Hence, (5.12) now gives

$$\left(\sum_{\mathcal{Q}\in\mathfrak{Q}}|\mathcal{Q}||R^{\mathrm{grad}}_{\mathcal{Q}}|^{\frac{p}{p-1}}\right)^{\frac{1}{p}}\leq C\mathrm{size}(\mathcal{T})^{\frac{1}{p-1}}\|u_e\|_{W^{2,p}(\mathfrak{Q})}+C\mathrm{size}(\mathcal{T})^{\frac{1}{(p-1)^2}}\left(\int_{\Omega_{\mathrm{disc}}}(1+|\nabla u_e|+|D^2u_e|)^p\,dz\right)^{\frac{1}{p}},$$

where we introduced $\Omega_{\mathrm{disc}}=\bigcup_{\mathcal{D}\in\mathfrak{D}_{\mathrm{disc}}}\mathcal{D}$. Since we assumed that φ is smooth on each subdomain Ω_i, we see that the set Ω_{disc} is a size$(\mathcal{T})$-neighborhood of the union of the boundaries of the Ω_i's. Hence, there exists $C>0$ such that $|\Omega_{\mathrm{disc}}|\leq C\mathrm{size}(\mathcal{T})$. It follows by the Hölder inequality and using the assumption $u_e\in W^{2,q}(\mathfrak{Q})$ that

$$\left(\sum_{\mathcal{Q}\in\mathfrak{Q}}|\mathcal{Q}||R^{\mathrm{grad}}_{\mathcal{Q}}|^{\frac{p}{p-1}}\right)^{\frac{1}{p}}\leq C\mathrm{size}(\mathcal{T})^{\frac{1}{p-1}}\|u_e\|_{W^{2,p}(\mathfrak{Q})}+C\mathrm{size}(\mathcal{T})^{\left(\frac{1}{(p-1)^2}+\frac{q-p}{pq}\right)}\|u_e\|_{W^{2,q}(\mathfrak{Q})},$$

and the claim is proved since $\frac{1}{(p-1)^2}+\frac{q-p}{pq}\geq\frac{1}{p-1}$ as soon as $q\geq p(p-1)^2$.

6. Examples. In the case of a linear problem where $\varphi(z,\xi) = A(z)\xi$, it is easily seen that for any $\mathcal{D} \in \mathfrak{D}$ the numerical flux $\varphi^{\mathcal{N}}_{\mathcal{D}}$ is a linear map of the DDFV gradient $\nabla^{\mathcal{T}}_{\mathcal{D}} u^{\mathcal{T}}$. More precisely, there exists a unique definite positive matrix $\overline{A}_{\mathcal{D}}$ such that $\varphi^{\mathcal{N}}_{\mathcal{D}}(\nabla^{\mathcal{T}}_{\mathcal{D}} u^{\mathcal{T}}) = \overline{A}_{\mathcal{D}} \nabla^{\mathcal{T}}_{\mathcal{D}} u^{\mathcal{T}}$.

In general, it is difficult to give an explicit formula for the matrix $\overline{A}_{\mathcal{D}}$, but it can be evaluated by computing the map $\delta^{\mathcal{D}}$, that is, following Proposition 4.3 and its proof, by computing the inverse of the $n_{\mathcal{D}} \times n_{\mathcal{D}}$ matrix $\sum_{\mathcal{D} \in \mathfrak{Q}_{\mathcal{D}}} |\mathcal{Q}|^t B_{\mathcal{Q}} A_{\mathcal{Q}} B_{\mathcal{Q}}$. This operation has a very low computational cost and has to be made only once.

In some particular cases, it is possible to find an explicit form for $\overline{A}_{\mathcal{D}}$ which is interesting in order to illustrate our approach and to compare the results with the 1D case. Let us consider a given diamond cell $\mathcal{D} \in \mathfrak{D}_{\text{int}}$ whose diagonals are $\sigma = \kappa|\mathcal{L}$ and $\sigma^* = \kappa^*|\mathcal{L}^*$.

- *First example:* We assume that $A(z)$ is constant on each control volume of the primal mesh. We denote by A_κ the value of $A(z)$ on the control volume κ. The matrix $\overline{A}_{\mathcal{D}}$ is then defined by

$$(\overline{A}_{\mathcal{D}}\boldsymbol{\nu}, \boldsymbol{\nu}) = \frac{(|\sigma_\kappa| + |\sigma_\mathcal{L}|)(A_\kappa \boldsymbol{\nu}, \boldsymbol{\nu})(A_\mathcal{L}\boldsymbol{\nu}, \boldsymbol{\nu})}{|\sigma_\mathcal{L}|(A_\kappa \boldsymbol{\nu}, \boldsymbol{\nu}) + |\sigma_\kappa|(A_\mathcal{L}\boldsymbol{\nu}, \boldsymbol{\nu})}, \tag{6.1}$$

$$(\overline{A}_{\mathcal{D}}\boldsymbol{\nu}^*, \boldsymbol{\nu}^*) = \frac{|\sigma_\mathcal{L}|(A_\mathcal{L}\boldsymbol{\nu}^*, \boldsymbol{\nu}^*) + |\sigma_\kappa|(A_\kappa \boldsymbol{\nu}^*, \boldsymbol{\nu}^*)}{|\sigma_\kappa| + |\sigma_\mathcal{L}|} - \frac{|\sigma_\kappa||\sigma_\mathcal{L}|}{|\sigma_\kappa| + |\sigma_\mathcal{L}|} \frac{((A_\kappa \boldsymbol{\nu}, \boldsymbol{\nu}^*) - (A_\mathcal{L}\boldsymbol{\nu}, \boldsymbol{\nu}^*))^2}{|\sigma_\mathcal{L}|(A_\kappa \boldsymbol{\nu}, \boldsymbol{\nu}) + |\sigma_\kappa|(A_\mathcal{L}\boldsymbol{\nu}, \boldsymbol{\nu})}, \tag{6.2}$$

$$(\overline{A}_{\mathcal{D}}\boldsymbol{\nu}, \boldsymbol{\nu}^*) = \frac{|\sigma_\mathcal{L}|(A_\mathcal{L}\boldsymbol{\nu}, \boldsymbol{\nu}^*)(A_\kappa \boldsymbol{\nu}, \boldsymbol{\nu}) + |\sigma_\kappa|(A_\kappa \boldsymbol{\nu}, \boldsymbol{\nu}^*)(A_\mathcal{L}\boldsymbol{\nu}, \boldsymbol{\nu})}{|\sigma_\mathcal{L}|(A_\kappa \boldsymbol{\nu}, \boldsymbol{\nu}) + |\sigma_\kappa|(A_\mathcal{L}\boldsymbol{\nu}, \boldsymbol{\nu})}. \tag{6.3}$$

  We recognize in (6.1) the weighted harmonic mean value of $(A_\kappa \boldsymbol{\nu}, \boldsymbol{\nu})$ and $(A_\mathcal{L}\boldsymbol{\nu}, \boldsymbol{\nu})$ and in the first term of (6.2) the weighted arithmetic mean value of $(A_\kappa \boldsymbol{\nu}^*, \boldsymbol{\nu}^*)$ and $(A_\mathcal{L}\boldsymbol{\nu}^*, \boldsymbol{\nu}^*)$.
- *Second example:* We assume that $A(z) = \lambda(z)\mathrm{Id}$ is isotropic, continuous on each quarter diamond, and we assume that the mesh is orthogonal, that is, $\sigma \perp \sigma^*$ or equivalently $\sin \alpha_{\mathcal{D}} = 1$. Introducing $\lambda_{\mathcal{Q}} = \frac{1}{|\mathcal{Q}|} \int_{\mathcal{Q}} \lambda(z)\, dz$, the mean value of λ over $\mathcal{Q}$, the equivalent matrix $\overline{A}_{\mathcal{D}}$ satisfies in that case:

$$(\overline{A}_{\mathcal{D}}\boldsymbol{\nu}, \boldsymbol{\nu}) = \frac{|\sigma_\kappa| + |\sigma_\mathcal{L}|}{|\sigma_{\kappa^*}| + |\sigma_{\mathcal{L}^*}|} \left(\frac{|\sigma_{\kappa^*}| \lambda_{\mathcal{Q}_{\mathcal{L},\kappa^*}} \lambda_{\mathcal{Q}_{\kappa,\kappa^*}}}{|\sigma_\kappa| \lambda_{\mathcal{Q}_{\mathcal{L},\kappa^*}} + |\sigma_\mathcal{L}| \lambda_{\mathcal{Q}_{\kappa,\kappa^*}}} + \frac{|\sigma_{\mathcal{L}^*}| \lambda_{\mathcal{Q}_{\mathcal{L},\mathcal{L}^*}} \lambda_{\mathcal{Q}_{\kappa,\mathcal{L}^*}}}{|\sigma_\kappa| \lambda_{\mathcal{Q}_{\mathcal{L},\mathcal{L}^*}} + |\sigma_\mathcal{L}| \lambda_{\mathcal{Q}_{\kappa,\mathcal{L}^*}}} \right), \tag{6.4}$$

$$(\overline{A}_{\mathcal{D}}\boldsymbol{\nu}^*, \boldsymbol{\nu}^*) = \frac{|\sigma_{\kappa^*}| + |\sigma_{\mathcal{L}^*}|}{|\sigma_\kappa| + |\sigma_\mathcal{L}|} \left(\frac{|\sigma_\kappa| \lambda_{\mathcal{Q}_{\kappa,\mathcal{L}^*}} \lambda_{\mathcal{Q}_{\kappa,\kappa^*}}}{|\sigma_{\kappa^*}| \lambda_{\mathcal{Q}_{\kappa,\mathcal{L}^*}} + |\sigma_{\mathcal{L}^*}| \lambda_{\mathcal{Q}_{\kappa,\kappa^*}}} + \frac{|\sigma_\mathcal{L}| \lambda_{\mathcal{Q}_{\mathcal{L},\mathcal{L}^*}} \lambda_{\mathcal{Q}_{\mathcal{L},\kappa^*}}}{|\sigma_{\kappa^*}| \lambda_{\mathcal{Q}_{\mathcal{L},\mathcal{L}^*}} + |\sigma_{\mathcal{L}^*}| \lambda_{\mathcal{Q}_{\mathcal{L},\kappa^*}}} \right), \tag{6.5}$$

$$(\overline{A}_{\mathcal{D}}\boldsymbol{\nu}, \boldsymbol{\nu}^*) = 0. \tag{6.6}$$

  Notice that even though $A(z)$ is isotropic, the matrix $\overline{A}_{\mathcal{D}}$ is only diagonal (in the orthogonal frame $(\boldsymbol{\nu}, \boldsymbol{\nu}^*)$) and not isotropic in general. Furthermore, we

see that (6.4) and (6.5) combine the arithmetic mean value of the coefficients in the transverse direction and harmonic mean value of the coefficients along the direction in which we are looking.

Unfortunately, in the nonlinear case there are very few cases where all the computations can be performed explicitly (see, for instance, the 1D Example 2.1). That is the reason why we propose in the following section a fully practical method to solve the m-DDFV and h-DDFV schemes in any situation.

7. Numerical implementation of the scheme. In this section we present a fully explicit algorithm to solve the finite volume scheme under study and we prove its convergence. From now on, we suppose given a DDFV mesh $\mathcal{T}$ on Ω and a source term f.

7.1. Some remarks on the potential case. We assume, only in this paragraph, that φ derives from a potential Φ, that is

$$\left\{ \begin{array}{rcl} \varphi(z,\xi) & = & \nabla_\xi \Phi(z,\xi), \quad \text{for all } \xi \in \mathbb{R}^2 \ \text{ and a.e. } z \in \Omega, \\ \Phi(z,0) & = & 0, \quad \text{for a.e. } z \in \Omega. \end{array} \right. \tag{7.1}$$

We can now define an approximation of Φ on each quarter diamond by $\Phi_{\mathcal{Q}}(\cdot) = \frac{1}{|\mathcal{Q}|}\int_{\mathcal{Q}} \Phi(z,\cdot)\,dz$ that satisfies $\nabla\Phi_{\mathcal{Q}} = \varphi_{\mathcal{Q}}$. Since φ is strictly monotonic, the function Φ is strictly convex.

PROPOSITION 7.1. *The solution $u^{\mathcal{T}}$ of the scheme* (4.17) *is the unique minimizer of the functional defined by*

$$J^{\mathcal{T}}(v^{\mathcal{T}}) = 2\sum_{\mathcal{D}\in\mathfrak{D}} \sum_{\mathcal{Q}\in\mathfrak{Q}_{\mathcal{D}}} |\mathcal{Q}|\Phi_{\mathcal{Q}}(\nabla^{\mathcal{N}}_{\mathcal{Q}} v^{\mathcal{T}}) - \sum_{\mathcal{K}} |\mathcal{K}| f_{\mathcal{K}} v_{\mathcal{K}} - \sum_{\mathcal{K}^*} |\mathcal{K}^*| f_{\mathcal{K}^*} v_{\mathcal{K}^*}, \quad \forall v^{\mathcal{T}} \in \mathbb{R}^{\mathcal{T}}.$$

Proof. By using assumptions $(\mathcal{H}_1)$, $(\mathcal{H}_2)$, and $(\mathcal{H}_3)$, the definition (7.1) and the Poincaré inequality, it is easily seen that $J^{\mathcal{T}}$ is strictly convex and coercive on $\mathbb{R}^{\mathcal{T}}$ and thus has a unique minimizer that we call $u^{\mathcal{T}}$.

Let us now write the Euler–Lagrange equation for this minimization problem. The equation corresponding to the unknown $u_{\mathcal{K}}$ reads

$$2\sum_{\mathcal{D}\in\mathfrak{D}_{\mathcal{K}}} \sum_{\mathcal{Q}\in\mathfrak{Q}_{\mathcal{D}}} |\mathcal{Q}| \left(\varphi_{\mathcal{Q}}(\nabla^{\mathcal{N}}_{\mathcal{Q}} u^{\mathcal{T}}), (\mathrm{Id} + \mathrm{B}_{\mathcal{Q}}.\mathrm{D}\delta^{\mathcal{D}}) \frac{\partial \nabla^{\mathcal{T}}_{\mathcal{D}} \mathrm{u}^{\mathcal{T}}}{\partial \mathrm{u}_{\mathcal{K}}} \right) = |\mathcal{K}| f_{\mathcal{K}}.$$

By definition of $\delta^{\mathcal{D}}$, for any $\mathcal{D}$ we have $\sum_{\mathcal{Q}\in\mathfrak{Q}_{\mathcal{D}}} |\mathcal{Q}|^t B_{\mathcal{Q}} \varphi_{\mathcal{Q}}(\nabla^{\mathcal{N}}_{\mathcal{Q}} u^{\mathcal{T}}) = 0$ so that the term containing the derivative $D\delta^{\mathcal{D}}$ of $\delta^{\mathcal{D}}$ vanishes. Furthermore, by definition of $\nabla^{\mathcal{T}}_{\mathcal{D}} u^{\mathcal{T}}$ (see (3.2)) we have

$$\frac{\partial \nabla^{\mathcal{T}}_{\mathcal{D}} u^{\mathcal{T}}}{\partial u_{\mathcal{K}}} = -\frac{|\sigma|}{2|\mathcal{D}|} \boldsymbol{\nu}_{\mathcal{K}}.$$

Hence it follows, using (4.16),

$$\begin{aligned} |\mathcal{K}| f_{\mathcal{K}} &= -2 \sum_{\mathcal{D}\in\mathfrak{D}_{\mathcal{K}}} \left(\sum_{\mathcal{Q}\in\mathfrak{Q}_{\mathcal{D}}} |\mathcal{Q}| \varphi_{\mathcal{Q}}(\nabla^{\mathcal{N}}_{\mathcal{Q}} u^{\mathcal{T}}), \frac{|\sigma|}{2|\mathcal{D}|} \boldsymbol{\nu}_{\mathcal{K}} \right) \\ &= -2 \sum_{\mathcal{D}\in\mathfrak{D}_{\mathcal{K}}} |\mathcal{D}| \left(\varphi^{\mathcal{N}}_{\mathcal{D}}(\nabla^{\mathcal{T}}_{\mathcal{D}} u^{\mathcal{T}}), \frac{|\sigma|}{2|\mathcal{D}|} \boldsymbol{\nu}_{\mathcal{K}} \right) = - \sum_{\mathcal{D}\in\mathfrak{D}_{\mathcal{K}}} |\sigma| \left(\varphi^{\mathcal{N}}_{\mathcal{D}}(\nabla^{\mathcal{T}}_{\mathcal{D}} u^{\mathcal{T}}), \boldsymbol{\nu}_{\mathcal{K}} \right), \end{aligned}$$

and the claim is proved. ∎

From now on, we denote by $\Delta = \bigoplus_{\mathcal{D}\in\mathfrak{D}} \mathbb{R}^{n_\mathcal{D}}$ the space in which the artificial unknowns $(\delta^\mathcal{D})_\mathcal{D}$ are lying. We introduce the following new functional defined on $\mathbb{R}^\mathcal{T} \times \Delta$

$$J^{\mathcal{T},\Delta}(v^\mathcal{T},\tilde{\delta}) = 2\sum_{\mathcal{D}\in\mathfrak{D}}\sum_{\mathcal{Q}\in\mathfrak{Q}_\mathcal{D}} |\mathcal{Q}|\Phi_\mathcal{Q}\left(\nabla^\mathcal{T}_\mathcal{D} v^\mathcal{T} + B_\mathcal{Q}\tilde{\delta}^\mathcal{D}\right) - \sum_{\mathcal{K}} |\mathcal{K}| f_\mathcal{K} v_\mathcal{K} - \sum_{\mathcal{K}^*} |\mathcal{K}^*| f_{\mathcal{K}^*} v_{\mathcal{K}^*}, \quad \forall v^\mathcal{T}\in\mathbb{R}^\mathcal{T},\ \forall\tilde{\delta}\in\Delta.$$

PROPOSITION 7.2. *The functional $J^{\mathcal{T},\Delta}$ has a unique minimizer which is given by $(u^\mathcal{T}, (\delta^\mathcal{D}(\nabla^\mathcal{T}_\mathcal{D} u^\mathcal{T}))_\mathcal{D})$.*

Proof. It is easily seen that, for any $v^\mathcal{T}\in\mathbb{R}^\mathcal{T}$ fixed, the functional $\tilde{\delta}\in\Delta \mapsto J^{\mathcal{T},\Delta}(v^\mathcal{T},\tilde{\delta})$ decouples into a sum over $\mathfrak{D}$ of independant maps depending only on $\tilde{\delta}^\mathcal{D}$ for a given $\mathcal{D}\in\mathfrak{D}$. Following the proof of Proposition 4.3, each of these maps has a unique minimum exactly given by $\delta^\mathcal{D}(\nabla^\mathcal{T}_\mathcal{D} v^\mathcal{T})$. Hence we proved

$$J^\mathcal{T}(v^\mathcal{T}) = J^{\mathcal{T},\Delta}\left(v^\mathcal{T}, (\delta^\mathcal{D}(\nabla^\mathcal{T}_\mathcal{D} v^\mathcal{T}))_\mathcal{D}\right) \le J^{\mathcal{T},\Delta}(v^\mathcal{T},\tilde{\delta}), \quad \forall\tilde{\delta}\in\Delta,$$

with equality if and only if $\tilde{\delta} = (\delta^\mathcal{D}(\nabla^\mathcal{T}_\mathcal{D} v^\mathcal{T}))_\mathcal{D}$, which gives the claim. ∎

7.2. Derivation of the decomposition-coordination method. In the case of a potential flux φ, we proved in Proposition 7.2 that the solution of the scheme can be obtained by minimizing a functional which can be computed explicitly but depending on much more unknowns than the cardinality of $\mathbb{R}^\mathcal{T}$ in which we look for the approximate solution.

We propose, for this nonquadratic minimization problem, a saddle-point formulation (in the very spirit of [12, 13]). Let us define the set $(\mathbb{R}^2)^\mathfrak{Q}$ of families of vectors of $\mathbb{R}^2$ indexed by the set of quarter diamonds. We suppose given a family $\mathcal{A} = (A_\mathcal{Q})_{\mathcal{Q}\in\mathfrak{Q}}$ of definite positive 2×2 matrices, which is aimed to play the role of heterogeneous and isotropic augmentation parameters. More precisely, we introduce now the augmented lagrangian

$$\begin{aligned} L^{\mathcal{T},\Delta}_\mathcal{A}(v^\mathcal{T},\tilde{\delta},g,\lambda) &= 2\sum_{\mathcal{Q}\in\mathfrak{Q}} |\mathcal{Q}|\Phi_\mathcal{Q}(g_\mathcal{Q}) + 2\sum_{\mathcal{Q}\in\mathfrak{Q}} |\mathcal{Q}|(\lambda_\mathcal{Q}, g_\mathcal{Q} - \nabla^\mathcal{T}_\mathcal{D} v^\mathcal{T} - B_\mathcal{Q}\tilde{\delta}^\mathcal{D}) \\ &\quad + \sum_{\mathcal{Q}\in\mathfrak{Q}} |\mathcal{Q}|\left(A_\mathcal{Q}(g_\mathcal{Q} - \nabla^\mathcal{T}_\mathcal{D} v^\mathcal{T} - B_\mathcal{Q}\tilde{\delta}^\mathcal{D}), (g_\mathcal{Q} - \nabla^\mathcal{T}_\mathcal{D} v^\mathcal{T} - B_\mathcal{Q}\tilde{\delta}^\mathcal{D})\right) \\ &\quad - \sum_{\mathcal{K}} |\mathcal{K}| f_\mathcal{K} v_\mathcal{K} - \sum_{\mathcal{K}^*} |\mathcal{K}^*| f_{\mathcal{K}^*} v_{\mathcal{K}^*}, \quad \forall v^\mathcal{T}\in\mathbb{R}^\mathcal{T}, \forall\tilde{\delta}\in\Delta, \forall g,\lambda\in(\mathbb{R}^2)^\mathfrak{Q}. \end{aligned}$$

If, for any $\mathcal{Q}\in\mathfrak{Q}$, we take $A_\mathcal{Q} = r\mathrm{Id}$ for a given parameter $r>0$ we recover the augmented lagrangian algorithm proposed in [12, 13]. It is easily seen that this lagrangian has a unique saddle-point $(u^\mathcal{T}, \delta^\mathcal{D}, p, \lambda)$ satisfying the equilibrium equations

$$\left\{\begin{array}{l} \varphi_\mathcal{Q}(g_\mathcal{Q}) + \lambda_\mathcal{Q} + A_\mathcal{Q}(g_\mathcal{Q} - \nabla^\mathcal{T}_\mathcal{D} u^\mathcal{T} - B_\mathcal{Q}\delta^\mathcal{D}) = 0, \quad \forall\mathcal{Q}\in\mathfrak{Q}, \\ \displaystyle\sum_{\mathcal{Q}\in\mathfrak{Q}_\mathcal{D}} |\mathcal{Q}|{}^t B_\mathcal{Q} A_\mathcal{Q}(B_\mathcal{Q}\delta^\mathcal{D} + \nabla^\mathcal{T}_\mathcal{D} u^\mathcal{T} - g_\mathcal{Q}) - \sum_{\mathcal{Q}\in\mathfrak{Q}_\mathcal{D}} |\mathcal{Q}|{}^t B_\mathcal{Q}\lambda_\mathcal{Q} = 0, \quad \forall\mathcal{D}\in\mathfrak{D}, \\ g_\mathcal{Q} - \nabla^\mathcal{T}_\mathcal{D} u^\mathcal{T} - B_\mathcal{Q}\delta^\mathcal{D} = 0, \quad \forall\mathcal{Q}\in\mathfrak{Q}, \\ \displaystyle 2\sum_{\mathcal{Q}\in\mathfrak{Q}} |\mathcal{Q}|\left(A_\mathcal{Q}(\nabla^\mathcal{T}_\mathcal{D} u^\mathcal{T} + B_\mathcal{Q}\delta^\mathcal{D} - g_\mathcal{Q}), \nabla^\mathcal{T}_\mathcal{D} v^\mathcal{T}\right) = \sum_{\mathcal{K}} |\mathcal{K}| f_\mathcal{K} v_\mathcal{K} + \sum_{\mathcal{K}^*} |\mathcal{K}^*| f_{\mathcal{K}^*} v_{\mathcal{K}^*} \\ \qquad \displaystyle + 2\sum_{\mathcal{Q}\in\mathfrak{Q}} |\mathcal{Q}|(\lambda_\mathcal{Q}, \nabla^\mathcal{T}_\mathcal{D} v^\mathcal{T}), \quad \forall v^\mathcal{T}\in\mathbb{R}^\mathcal{T}. \end{array}\right. \tag{7.2}$$

These equations are clearly equivalent to

$$
\text{(7.3)}\qquad \begin{cases} \lambda_{\mathcal{Q}} = -\varphi_{\mathcal{Q}}(g_{\mathcal{Q}}), \quad \forall \mathcal{Q} \in \mathfrak{\mathbf{Q}}, \\ \displaystyle\sum_{\mathcal{Q}\in\mathfrak{\mathbf{Q}}_{\mathcal{D}}} |\mathcal{Q}|{}^t B_{\mathcal{Q}} \lambda_{\mathcal{Q}} = 0, \quad \forall \mathcal{D} \in \mathfrak{\mathbf{D}}, \\ g_{\mathcal{Q}} = \nabla_{\mathcal{D}}^{\mathcal{T}} u^{\mathcal{T}} + B_{\mathcal{Q}} \delta^{\mathcal{D}}, \quad \forall \mathcal{Q} \in \mathfrak{\mathbf{Q}}, \\ \displaystyle -2 \sum_{\mathcal{Q}\in\mathfrak{\mathbf{Q}}} |\mathcal{Q}|(\lambda_{\mathcal{Q}}, \nabla_{\mathcal{D}}^{\mathcal{T}} v^{\mathcal{T}}) = \sum_{\mathcal{K}} |\mathcal{K}| f_{\mathcal{K}} v_{\mathcal{K}} + \sum_{\mathcal{K}^*} |\mathcal{K}^*| f_{\mathcal{K}^*} v_{\mathcal{K}^*}, \quad \forall v^{\mathcal{T}} \in \mathbb{R}^{\mathcal{T}}. \end{cases}
$$

The first three equations imply that $\delta^{\mathcal{D}} = \delta^{\mathcal{D}}(\nabla_{\mathcal{D}}^{\mathcal{T}} u^{\mathcal{T}})$, and then the fourth equation is nothing but a different way to write (4.17). As a consequence, the saddle point of $L_{\mathcal{A}}^{\mathcal{T},\Delta}$ gives the unique solution to the finite volume scheme.

From (7.2), we deduce an iterative method to solve our problem following the same idea as [13, ALG 2, p. 170]. In our setting, the algorithm reads as follows: we suppose given $\lambda^0 \in (\mathbb{R}^2)^{\mathfrak{Q}}$, $g^0 \in (\mathbb{R}^2)^{\mathfrak{Q}}$, then for any $n \geq 1$:

1. Find $(u^{\mathcal{T},n}, \delta_{\mathcal{D}}^n) \in \mathbb{R}^{\mathcal{T}} \times \Delta$ solution to the linear problem

$$
\text{(7.4)}\qquad \begin{aligned} &2 \sum_{\mathcal{Q}\in\mathfrak{\mathbf{Q}}} |\mathcal{Q}| \Big(A_{\mathcal{Q}}(\nabla_{\mathcal{D}}^{\mathcal{T}} u^{\mathcal{T},n} + B_{\mathcal{Q}} \delta_{\mathcal{D}}^n - g_{\mathcal{Q}}^{n-1}), \nabla_{\mathcal{D}}^{\mathcal{T}} v^{\mathcal{T}} \Big) \\ &\qquad = \sum_{\mathcal{K}} |\mathcal{K}| f_{\mathcal{K}} v_{\mathcal{K}} + \sum_{\mathcal{K}^*} |\mathcal{K}^*| f_{\mathcal{K}^*} v_{\mathcal{K}^*} + 2 \sum_{\mathcal{Q}\in\mathfrak{\mathbf{Q}}} |\mathcal{Q}| (\lambda_{\mathcal{Q}}^{n-1}, \nabla_{\mathcal{D}}^{\mathcal{T}} v), \quad \forall v^{\mathcal{T}} \in \mathbb{R}^{\mathcal{T}}. \\ &\sum_{\mathcal{Q}\in\mathfrak{\mathbf{Q}}_{\mathcal{D}}} |\mathcal{Q}|{}^t B_{\mathcal{Q}} A_{\mathcal{Q}} (B_{\mathcal{Q}} \delta_{\mathcal{D}}^n + \nabla_{\mathcal{D}}^{\mathcal{T}} u^{\mathcal{T},n} - g_{\mathcal{Q}}^{n-1}) - \sum_{\mathcal{Q}\in\mathfrak{\mathbf{Q}}_{\mathcal{D}}} |\mathcal{Q}|{}^t B_{\mathcal{Q}} \lambda_{\mathcal{Q}}^{n-1} = 0, \, \forall \mathcal{D} \in \mathfrak{\mathbf{D}}. \end{aligned}
$$

   Notice that the second equation explicitly gives, locally on each diamond cell $\mathcal{D}$, the expression of $\delta_{\mathcal{D}}^n$ as an affine function of $\nabla_{\mathcal{D}}^{\mathcal{T}} u^{\mathcal{T},n}$. It is only needed to compute, one time at the beginning of the algorithm, the inverse of all the definite positive symmetric matrices $\sum_{\mathcal{Q}\in\mathfrak{\mathbf{Q}}_{\mathcal{D}}} |\mathcal{Q}|{}^t B_{\mathcal{Q}} \, A_{\mathcal{Q}} \, B_{\mathcal{Q}}$ whose size is $n_{\mathcal{D}} \times n_{\mathcal{D}}$ (which is low since $n_{\mathcal{D}} = 1$ for boundary diamond cells and $n_{\mathcal{D}} = 4$ for interior diamond cells).
   Finally, once the second equation in (7.4) is solved, we can introduce the expression of $\delta_{\mathcal{D}}^n$ as a function of $\nabla_{\mathcal{D}}^{\mathcal{T}} u^{\mathcal{T},n}$ in the first equation of (7.4). This first equation is now an explicit large linear system in the variables $u^{\mathcal{T},n}$. Notice that the matrix of this large sparse linear system is the same at each iteration, which is very important if one wants to use direct linear solvers, for instance. Finally, notice that the size and the stencil of this system is exactly the same as the one of the DDFV matrix for the Laplace equation, for instance.
2. For any $\mathcal{Q} \in \mathfrak{\mathbf{Q}}$, find $g_{\mathcal{Q}}^n$ satisfying

$$
\text{(7.5)}\qquad \varphi_{\mathcal{Q}}(g_{\mathcal{Q}}^n) + \lambda_{\mathcal{Q}}^{n-1} + A_{\mathcal{Q}}(g_{\mathcal{Q}}^n - \nabla_{\mathcal{D}}^{\mathcal{T}} u^{\mathcal{T},n} - B_{\mathcal{Q}} \delta_{\mathcal{D}}^n) = 0.
$$

   This is the unique nonlinear part of the algorithm. The equation is localized on each quarter diamond and consists in solving a nonlinear equation in $\mathbb{R}^2$ defined by the explicit map $\varphi_{\mathcal{Q}} + A_{\mathcal{Q}}$. Notice that $\varphi_{\mathcal{Q}} + A_{\mathcal{Q}}$ is strictly monotonic and coercive on $\mathbb{R}^2$ so that the solution $g_{\mathcal{Q}}^n$ to (7.5) exists and is unique.
   From a practical point of view, one can use here a Newton method to solve all these equations simultaneously (this step can be massively parallelized).

3. Finally compute $\lambda_{\mathcal{Q}}^n$ through

$$\lambda_{\mathcal{Q}}^n = \lambda_{\mathcal{Q}}^{n-1} + A_{\mathcal{Q}}(g_{\mathcal{Q}}^n - \nabla_{\mathcal{D}}^{\mathcal{T}} u^{\mathcal{T},n} - B_{\mathcal{Q}}\delta_{\mathcal{D}}^n), \quad \forall \mathcal{Q} \in \mathfrak{\mathbf{Q}}. \tag{7.6}$$

The choice of the *best* augmentation matrices $\mathcal{A}$ is a complex problem (see the discussion in [13], for instance). We will give some examples of such choices in section 8.

7.3. General fluxes. Convergence of the iterative solver. The above algorithm is deduced from the lagrangian formulation of the scheme which is available only in the variational case (7.1). Nevertheless, the iterative algorithm (7.4)–(7.6) can be used in the general case of any monotonic flux φ. We can now prove the convergence of this algorithm in the non-variational setting.

THEOREM 7.3. *Let $\mathcal{T}$ be a DDFV mesh on Ω and $(\varphi_{\mathcal{Q}})_{\mathcal{Q}}$ a family of strictly monotonic continuous maps from $\mathbb{R}^2$ onto itself. Then for any augmentation matrices family $\mathcal{A}$, the algorithm given by* (7.4)–(7.6) *converges, when n goes to infinity, towards the unique solution to* (7.3)*, that is, the unique solution to the m-DDFV scheme* (4.17)*.*

Remark 7.4. This result still holds if one replaces (7.6) by

$$\lambda_{\mathcal{Q}}^n = \lambda_{\mathcal{Q}}^{n-1} + \gamma A_{\mathcal{Q}}(g_{\mathcal{Q}}^n - \nabla_{\mathcal{D}}^{\mathcal{T}} u^{\mathcal{T},n} - B_{\mathcal{Q}}\delta_{\mathcal{D}}^n), \quad \forall \mathcal{Q} \in \mathfrak{\mathbf{Q}}.$$

for any $\gamma \in \left]0, \frac{1+\sqrt{5}}{2}\right[$. It is possible in some cases to improve the convergence rate of the algorithm by choosing values of $\gamma \neq 1$.

The proof we present here is an adaptation to our framework of the arguments given in [13]. Furthermore, a similar algorithm and result is available for the hybrid h-DDFV scheme.

Proof. For any real-valued or vector-valued families $f = (f_{\mathcal{Q}})_{\mathcal{Q}\in\mathfrak{\mathbf{Q}}}$, $g = (g_{\mathcal{Q}})_{\mathcal{Q}\in\mathfrak{\mathbf{Q}}}$, and any $\alpha \in \{-1, 0, 1\}$, we introduce the following inner products and associated norms:

$$(f,g)_{\alpha,\mathcal{A}} = \sum_{\mathcal{Q}\in\mathfrak{\mathbf{Q}}} |\mathcal{Q}|(A_{\mathcal{Q}}^{\alpha} f_{\mathcal{Q}}, g_{\mathcal{Q}}), \quad \|f\|_{\alpha,\mathcal{A}} = (f,f)_{\alpha,\mathcal{A}}^{\frac{1}{2}}.$$

Let us define the error terms $v^{\mathcal{T},n} = u^{\mathcal{T},n} - u^{\mathcal{T}}$, $h^n = (h_{\mathcal{Q}}^n)_{\mathcal{Q}} \in (\mathbb{R}^2)^{\mathfrak{Q}}$ with $h_{\mathcal{Q}}^n = g_{\mathcal{Q}}^n - g_{\mathcal{Q}}$, $\mu^n = (\mu_{\mathcal{Q}}^n)_{\mathcal{Q}} \in (\mathbb{R}^2)^{\mathfrak{Q}}$ with $\mu_{\mathcal{Q}}^n = \lambda_{\mathcal{Q}}^n - \lambda_{\mathcal{Q}}$, and $\beta^n = (\beta_{\mathcal{D}}^n)_{\mathcal{D}} \in \Delta$ with $\beta_{\mathcal{D}}^n = \delta_{\mathcal{D}}^n - \delta^{\mathcal{D}}$. Finally, we introduce the nonlinear map $\varphi_{\mathfrak{Q}} : (\mathbb{R}^2)^{\mathfrak{Q}} \mapsto (\mathbb{R}^2)^{\mathfrak{Q}}$ defined by $(\varphi_{\mathfrak{Q}}(g))_{\mathcal{Q}} = \varphi_{\mathcal{Q}}(g_{\mathcal{Q}})$ and the linear map $\mathcal{B} : \Delta \mapsto (\mathbb{R}^2)^{\mathfrak{Q}}$ defined by $(\mathcal{B}\beta)_{\mathcal{Q}} = B_{\mathcal{Q}}\beta_{\mathcal{D}}$, where $\mathcal{D}$ is the diamond cell such that $\mathcal{Q} \in \mathfrak{\mathbf{Q}}_{\mathcal{D}}$.

Using those notations, we see from (7.4)–(7.6) that these quantities solve the following equations:

1. Equation for $v^{\mathcal{T},n}$:

$$\begin{aligned}(\nabla^{\mathcal{T}} v^{\mathcal{T},n} + \mathcal{B}\beta^n - h^n, \nabla_{\mathcal{D}}^{\mathcal{T}} w^{\mathcal{T}})_{1,\mathcal{A}} + \left(h^n - h^{n-1}, \nabla^{\mathcal{T}} w^{\mathcal{T}}\right)_{1,\mathcal{A}} \\ = \left(\mu^{n-1}, \nabla^{\mathcal{T}} w^{\mathcal{T}}\right)_{0,\mathcal{A}}, \quad \forall w^{\mathcal{T}} \in \mathbb{R}^{\mathcal{T}}.\end{aligned} \tag{7.7}$$

2. Equation for $\beta_{\mathcal{D}}^n$:

$$\begin{aligned}\sum_{\mathcal{Q}\in\mathfrak{\mathbf{Q}}_{\mathcal{D}}} |\mathcal{Q}|^t B_{\mathcal{Q}} A_{\mathcal{Q}}(B_{\mathcal{Q}}\beta_{\mathcal{D}}^n + \nabla_{\mathcal{D}}^{\mathcal{T}} v^{\mathcal{T},n} - h_{\mathcal{Q}}^n) + \sum_{\mathcal{Q}\in\mathfrak{\mathbf{Q}}_{\mathcal{D}}} |\mathcal{Q}|^t B_{\mathcal{Q}} A_{\mathcal{Q}} \left(h_{\mathcal{Q}}^n - h_{\mathcal{Q}}^{n-1}\right) \\ - \sum_{\mathcal{Q}\in\mathfrak{\mathbf{Q}}_{\mathcal{D}}} |\mathcal{Q}|^t B_{\mathcal{Q}} \mu_{\mathcal{Q}}^{n-1} = 0, \quad \forall \mathcal{D} \in \mathfrak{\mathbf{D}}.\end{aligned} \tag{7.8}$$

3. Equation for $h^n_{\mathcal{Q}}$:

$$\varphi_{\mathcal{Q}}(g^n_{\mathcal{Q}}) - \varphi_{\mathcal{Q}}(g_{\mathcal{Q}}) + \mu^{n-1}_{\mathcal{Q}} + A_{\mathcal{Q}}(h^n_{\mathcal{Q}} - \nabla^{\mathcal{T}}_{\mathcal{D}} v^{\mathcal{T},n} - B_{\mathcal{Q}}\beta^n_{\mathcal{D}}) = 0, \quad \forall \mathcal{Q} \in \mathfrak{Q}. \tag{7.9}$$

4. Equation for $\mu^n_{\mathcal{Q}}$:

$$\mu^n_{\mathcal{Q}} = \mu^{n-1}_{\mathcal{Q}} + A_{\mathcal{Q}}(h^n_{\mathcal{Q}} - \nabla^{\mathcal{T}}_{\mathcal{D}} v^{\mathcal{T},n} - B_{\mathcal{Q}}\beta^n_{\mathcal{D}}), \quad \forall \mathcal{Q} \in \mathfrak{Q}. \tag{7.10}$$

We take $w^{\mathcal{T}} = v^{\mathcal{T},n}$ in (7.7) and get

$$(\nabla^{\mathcal{T}} v^{\mathcal{T},n} + \mathcal{B}\beta^n - h^n, \nabla^{\mathcal{T}} v^{\mathcal{T},n})_{1,\mathcal{A}} + \left(h^n - h^{n-1}, \nabla^{\mathcal{T}} v^{\mathcal{T},n}\right)_{1,\mathcal{A}} = \left(\mu^{n-1}, \nabla^{\mathcal{T}} v^{\mathcal{T},n}\right)_{0,\mathcal{A}}. \tag{7.11}$$

From (7.10), we deduce

$$\frac{1}{2}\|\mu^n\|^2_{-1,\mathcal{A}} - \frac{1}{2}\left\|\mu^{n-1}\right\|^2_{-1,\mathcal{A}} - \left(\mu^{n-1}, h^n - \nabla^{\mathcal{T}} v^{\mathcal{T},n} - \mathcal{B}\beta^n\right)_{0,\mathcal{A}} - \frac{1}{2}\|h^n - \nabla^{\mathcal{T}} v^{\mathcal{T},n} - \mathcal{B}\beta^n\|^2_{1,\mathcal{A}} = 0. \tag{7.12}$$

Taking the $(\cdot,\cdot)_{0,\mathcal{A}}$ inner product of (7.9) with $h^n = g^n - g$, we get

$$(\varphi_{\mathfrak{Q}}(g^n) - \varphi_{\mathfrak{Q}}(g), g^n - g)_{0,\mathcal{A}} + (h^n - \nabla^{\mathcal{T}} v^{\mathcal{T},n} - \mathcal{B}\beta^n, h^n)_{1,\mathcal{A}} + \left(\mu^{n-1}, h^n\right)_{0,\mathcal{A}} = 0.$$

Multiplying (7.8) by $\beta^n_{\mathcal{D}}$, summing over $\mathfrak{D}$ and using (7.11), it follows

$$(\varphi_{\mathfrak{Q}}(g^n) - \varphi_{\mathfrak{Q}}(g), g^n - g)_{0,\mathcal{A}} + \|h^n - \nabla^{\mathcal{T}} v^{\mathcal{T},n} - \mathcal{B}\beta^n\|^2_{1,\mathcal{A}} + \left(\mu^{n-1}, h^n - \nabla^{\mathcal{T}} v^{\mathcal{T},n} - \mathcal{B}\beta^n\right)_{0,\mathcal{A}} + \left(h^n - h^{n-1}, \nabla^{\mathcal{T}} v^{\mathcal{T},n} + \mathcal{B}\beta^n\right)_{1,\mathcal{A}} = 0.$$

Adding this equation to (7.12), we get rid of the term containing $\mu^{n-1}_{\mathcal{Q}}$. We obtain

$$\frac{1}{2}\|\mu^n\|^2_{-1,\mathcal{A}} - \frac{1}{2}\left\|\mu^{n-1}\right\|^2_{-1,\mathcal{A}} + (\varphi_{\mathfrak{Q}}(g^n) - \varphi_{\mathfrak{Q}}(g), g^n - g)_{0,\mathcal{A}} + \frac{1}{2}\|h^n - \nabla^{\mathcal{T}} v^{\mathcal{T},n} - \mathcal{B}\beta^n\|^2_{1,\mathcal{A}} + \left(h^n - h^{n-1}, \nabla^{\mathcal{T}} v^{\mathcal{T},n} + \mathcal{B}\beta^n\right)_{1,\mathcal{A}} = 0. \tag{7.13}$$

From (7.5) we get, for any $\mathcal{Q} \in \mathfrak{Q}$,

$$\varphi_{\mathcal{Q}}(g^n_{\mathcal{Q}}) - \varphi_{\mathcal{Q}}(g^{n-1}_{\mathcal{Q}}) + \left(\mu^{n-1}_{\mathcal{Q}} - \mu^{n-2}_{\mathcal{Q}}\right) + A_{\mathcal{Q}}\left(\left(h^n_{\mathcal{Q}} - h^{n-1}_{\mathcal{Q}}\right) - \left(\nabla^{\mathcal{T}}_{\mathcal{D}} v^{\mathcal{T},n} - \nabla^{\mathcal{T}}_{\mathcal{D}} v^{\mathcal{T},n-1}\right) - B_{\mathcal{Q}}\left(\beta^n_{\mathcal{D}} - \beta^{n-1}_{\mathcal{D}}\right)\right) = 0,$$

so that, using (7.10) it follows

$$\varphi_{\mathcal{Q}}\left(g^n_{\mathcal{Q}}\right) - \varphi_{\mathcal{Q}}\left(g^{n-1}_{\mathcal{Q}}\right) + A_{\mathcal{Q}}\left(h^n_{\mathcal{Q}} - h^{n-1}_{\mathcal{Q}}\right) + A_{\mathcal{Q}} h^{n-1}_{\mathcal{Q}} - A_{\mathcal{Q}}\left(\nabla^{\mathcal{T}}_{\mathcal{D}} v^{\mathcal{T},n} + B_{\mathcal{Q}}\beta^n_{\mathcal{D}}\right) = 0.$$

Taking the inner product $(\cdot,\cdot)_{0,\mathcal{A}}$ of these equations by $g^n - g^{n-1} = h^n - h^{n-1}$ we get

$$\left(\varphi_{\mathfrak{Q}}(g^n) - \varphi_{\mathfrak{Q}}\left(g^{n-1}\right), g^n - g^{n-1}\right)_{0,\mathcal{A}} + \left\|h^n - h^{n-1}\right\|^2_{1,\mathcal{A}} + \left(h^{n-1}, h^n - h^{n-1}\right)_{1,\mathcal{A}} - \left(\nabla^{\mathcal{T}} v^{\mathcal{T},n} + \mathcal{B}\beta^n, h^n - h^{n-1}\right)_{1,\mathcal{A}} = 0.$$

If we add to this equation the following algebraic relation

$$\frac{1}{2}\left\|h^n\right\|_{1,\mathcal{A}}^2-\frac{1}{2}\left\|h^{n-1}\right\|_{1,\mathcal{A}}^2-\frac{1}{2}\left\|h^n-h^{n-1}\right\|_{1,\mathcal{A}}^2=\left(h^{n-1},h^n-h^{n-1}\right)_{1,\mathcal{A}},$$

we get

$$\begin{aligned}&\frac{1}{2}\left\|h^n\right\|_{1,\mathcal{A}}^2-\frac{1}{2}\left\|h^{n-1}\right\|_{1,\mathcal{A}}^2+\frac{1}{2}\left\|h^n-h^{n-1}\right\|_{1,\mathcal{A}}^2\\&\quad+\left(\varphi_{\mathfrak{D}}(g^n)-\varphi_{\mathfrak{D}}(g^{n-1}),g^n-g^{n-1}\right)_{0,\mathcal{A}}-\left(\nabla^{\mathcal{T}}v^{\mathcal{T},n}+\mathcal{B}\beta^n,h^n-h^{n-1}\right)_{1,\mathcal{A}}=0.\end{aligned}$$

Summing this equation with (7.13), it follows:

$$\begin{aligned}(7.14)\quad&\frac{1}{2}\left(\left\|\mu^n\right\|_{-1,\mathcal{A}}^2+\left\|h^n\right\|_{1,\mathcal{A}}^2\right)-\frac{1}{2}\left(\left\|\mu^{n-1}\right\|_{-1,\mathcal{A}}^2+\left\|h^{n-1}\right\|_{1,\mathcal{A}}^2\right)\\&+\left(\varphi_{\mathfrak{D}}(g^n)-\varphi_{\mathfrak{D}}\left(g^{n-1}\right),g^n-g^{n-1}\right)_{0,\mathcal{A}}+\left(\varphi_{\mathfrak{D}}(g^n)-\varphi_{\mathfrak{D}}(g),g^n-g\right)_{0,\mathcal{A}}\\&+\frac{1}{2}\left\|h^n-h^{n-1}\right\|_{1,\mathcal{A}}^2+\frac{1}{2}\left\|h^n-\nabla^{\mathcal{T}}v^{\mathcal{T},n}-\mathcal{B}\beta^n\right\|_{1,\mathcal{A}}^2=0.\end{aligned}$$

Since each map $\varphi_{\mathcal{Q}}$ is monotonic as is $\varphi_{\mathfrak{D}}$, we deduce that the sequence

$$\left(\left\|\mu^n\right\|_{-1,\mathcal{A}}^2+\left\|h^n\right\|_{1,\mathcal{A}}^2\right)_n$$

of nonnegative numbers is nonincreasing and then converges. Coming back to (7.14), we deduce that $(g_{\mathcal{Q}}^n)_n$ converges toward $g_{\mathcal{Q}}$ for any $\mathcal{Q}\in\mathfrak{Q}$, that is $h_{\mathcal{Q}}^n\to 0$, and that

$$h_{\mathcal{Q}}^n-\nabla_{\mathcal{D}}^{\mathcal{T}}v^{\mathcal{T},n}-B_{\mathcal{Q}}\beta_{\mathcal{D}}^n\xrightarrow{n\to\infty}0,\quad\forall\mathcal{Q}\in\mathfrak{Q}.$$

It follows that $\nabla_{\mathcal{D}}^{\mathcal{T}}v^{\mathcal{T},n}+B_{\mathcal{Q}}\beta_{\mathcal{D}}^n$ goes to 0 for any $\mathcal{Q}\in\mathfrak{Q}$. Then we get that $\mu_{\mathcal{Q}}^n\to 0$ by (7.9). Using now (4.9), we find that for any $\mathcal{D}\in\mathfrak{D}$ we have $\nabla_{\mathcal{D}}^{\mathcal{T}}v^{\mathcal{T},n}\to 0$ and then finally $\beta_{\mathcal{D}}^n\to 0$. □

8. Numerical results. We present here some numerical results in the following situation: we consider the domain $\Omega=]0,1[\times]0,1[$, and the flux φ defined by

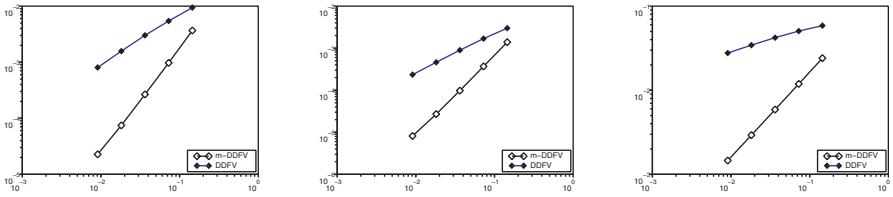
$$(8.1)\qquad\varphi(z,\xi)=\begin{cases}|\xi|^{p-2}\xi, & \text{if } z_1<0.5,\\ (A\xi,\xi)^{\frac{p-2}{2}}A\xi, & \text{if } z_1>0.5,\end{cases}$$

where A is the matrix $A=\left(\begin{smallmatrix}\alpha & 0\\ 0 & \beta\end{smallmatrix}\right)$, $\alpha,\beta\in]1,+\infty[$. Then we construct the source term f, and the boundary data in such a way that the solution of (1.1) is given by

$$u_e(z)=\begin{cases}(\alpha z_1+\gamma z_2)^2, & \text{if } z_1<0.5,\\ \left(z_1+\gamma z_2+\dfrac{\alpha-1}{2}\right)^2, & \text{if } z_1>0.5,\end{cases}$$

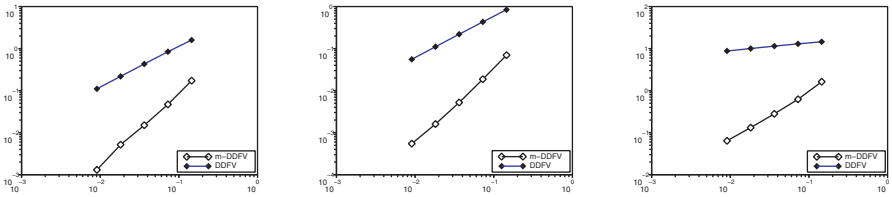
where $\gamma=\sqrt{\alpha\frac{1-\alpha}{1-\beta}}$. It is easily seen that this function u_e satisfies the transmission condition described in the introduction along the line $\{z_1=\frac{1}{2}\}$. This is true for any value of p.

In a first test case, we choose $p=3.0$ and $(\alpha,\beta)=(5.0,2.0)$ so that the problem we consider has discontinuities and is anisotropic. We show in Figure 8.1 the errors in three different norms as a function of the mesh size, in a logarithmic scale, for the



L^∞ error slopes 1.71 and 0.97 | L^p error slopes 1.73 and 0.97 | $W^{1,p}$ error slopes 1.0 and 0.31

FIG. 8.1. *Test case* 1*:* $p = 3.0$*,* $\alpha = 5$*,* $\beta = 2$*.*



L^∞ error slopes 1.98 and 0.98 | L^p error slopes 1.54 and 0.99 | $W^{1,p}$ error slopes 1.04 and 0.19

FIG. 8.2. *Test case* 2*:* $p = 5.0$*,* $\alpha = \beta = 10.0$*.*

original DDFV scheme (3.3)–(3.4) (marked by ⧫) and for the m-DDFV scheme (4.16)–(4.17) (marked by ◇). Notice that, since the operator defined by (8.1) is piecewise constant, the h-DDFV scheme of Definition 5.2 is exactly the same as the m-DDFV scheme.

As predicted by the theory, the m-DDFV scheme provides a much better convergence rate than the original DDFV scheme. Furthermore, and it is an important point, the error (in any of the three norms we consider) obtained by the m-DDFV scheme is better even in the case of coarse meshes.

As a second test case we assume that $p = 5.0$ and $(\alpha, \beta) = (10.0, 10.0)$. In this situation the operator is isotropic, but the jump of the diffusion coefficient is of order $10^{2.5}$. We observe (see Figure 8.2) the same overall behavior of the two schemes.

Finally, we want to illustrate the behavior of the decomposition-coordination algorithm proposed in section 7. First of all, it is shown in [13], for instance, that such algorithms can be applied, in suitable infinite dimensional functional spaces, directly to the continuous problem (1.1). This fact let us hope that the convergence rate of the present method may not depend too much on the size of the mesh we consider. Actually, in our numerical computations, we observed that the number of iterations needed to achieve a given residual norm was essentially the same in each level of refinement of the mesh we considered.

Let us now illustrate the fact that one can take advantage of using the heterogeneous and isotropic augmentation matrices family $\mathcal{A}$. To this end, we consider the first example given above (Figure 8.1) and we plot in Figure 8.3 the evolution during the iterations of the L^p, $W^{1,p}$ errors and the residual norm of the algorithm. The left plot is the one obtained with the classical augmentation term when $A_{\mathcal{Q}} = r\mathrm{Id}$ for any $\mathcal{Q} \in \mathfrak{Q}$, with $r = 1.5$. In the right plot, we have chosen $A_{\mathcal{Q}} = r\mathrm{Id}$ if $\mathcal{Q} \subset \{z_1 < \frac{1}{2}\}$ and $A_{\mathcal{Q}} = rA$ if $\mathcal{Q} \subset \{z_1 > \frac{1}{2}\}$.

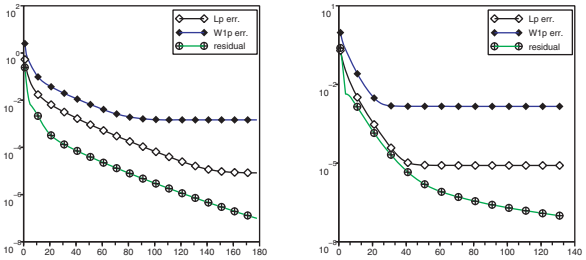


FIG. 8.3. *Convergence of the iterative solver. Isotropic (left) and anisotropic augmentation (right).*

We see that the use of anisotropic and heterogeneous augmention terms lets us achieve the tolerance 10^{-7} on the residual norm in ~ 130 iterations instead of ~ 180 in the isotropic case. Furthermore, we see that the error due to the scheme is achieved after ~ 40 iterations, that is, when the residual norm is $\sim 5.10^{-6}$ in the anisotropic augmentation case, whereas it is only achieved in ~ 150 iterations, for a residual norm of 5.10^{-7} in the isotropic case. This illustrates the fact that the choice of suitable augmentation matrices $\mathcal{A}$ may let us save a significant amount of computational time to solve our scheme.

9. Conclusions. In this paper we provide a modification of the DDFV finite volume scheme for nonlinear elliptic problems on general 2D grids in order to take into account discontinuities in the coefficients. The m-DDFV scheme we obtained is proven to present a better consistency of the fluxes at the discontinuities. The performance of the scheme is illustrated by numerical results on heterogeneous and anisotropic p-Laplacian equations. Moreover, we proposed a generalization of the decomposition-coordination method of Glowinski in order to solve our scheme. We show that the use of heterogeneous and anisotropic augmentation terms in this approach may lead to much better performance of the algorithm.

Finally, some extensions of this work are given in [5] where we analyze the m-DDFV method in the case of the coupling, through transmission conditions, of a linear and a nonlinear elliptic Leray–Lions operators.

Acknowledgments. The authors want to thank the referees whose remarks have improved this paper.

REFERENCES

[1] I. AAVATSMARK, T. BARKVE, O. BØE, AND T. MANNSETH, *Discretization on unstructured grids for inhomogeneous, anisotropic media.* I. *Derivation of the methods*, SIAM J. Sci. Comput., 19 (1998), pp. 1700–1716 (electronic).

[2] I. AAVATSMARK, T. BARKVE, O. BØE, AND T. MANNSETH, *Discretization on unstructured grids for inhomogeneous, anisotropic media.* II. *Discussion and numerical results*, SIAM J. Sci. Comput., 19 (1998), pp. 1717–1736 (electronic).

[3] B. ANDREIANOV, F. BOYER, AND F. HUBERT, *Discrete duality finite volume schemes for Leray-Lions type elliptic problems on general* 2*D-meshes*, Numer. Methods for PDEs, 23 (2007), pp. 145–195.

[4] B. ANDREIANOV, M. GUTNIC, AND P. WITTBOLD, *Convergence of finite volume approximations for a nonlinear elliptic-parabolic problem: A "continuous" approach*, SIAM J. Numer. Anal., 42 (2004), pp. 228–251.

[5] F. BOYER AND F. HUBERT, *Finite volume method for nonlinear transmission problems*, in Domain Decomposition Methods in Science and Engineering XVII, Lecture Notes in Com-

putational Science and Engineering 60, U. Langer, M. Discacciati, D. Keyes, O. Widlund, and W. Zulehner, eds., Heidelberg, 2008, Springer, pp. 443–450. Proceedings of the 17th International Conference on Domain Decomposition Methods held at St. Wolfgang / Strobl, Austria, July 3–7, 2006.

[6] Y. Coudière, C. Pierre, and R. Turpault, *Solving the fully coupled heart and torso problems of electrocardiology with a 3D discrete duality finite volume method*, submitted, (2006). http://hal.archives-ouvertes.fr/hal-00016825/en/.

[7] Y. Coudière, J.-P. Vila, and P. Villedieu, *Convergence rate of a finite volume scheme for a two-dimensional convection-diffusion problem*, M2AN Math. Model. Numer. Anal., 33 (1999), pp. 493–516.

[8] K. Domelevo and P. Omnes, *A finite volume method for the Laplace equation on almost arbitrary two-dimensional grids*, M2AN Math. Model. Numer. Anal., 39 (2005), pp. 1203–1249.

[9] J. Droniou, *Finite volume approximations for fully nonlinear elliptic equations in divergence form*, M2AN Math. Model. Numer. Anal., 40 (2006), pp. 1069–1100 (2007).

[10] R. Eymard, T. Gallouët, and R. Herbin, *Finite volume methods*, in Handbook of Numerical Analysis, Vol. VII, Handb. Numer. Anal., VII, North-Holland, Amsterdam, 2000, pp. 713–1020.

[11] M. Feistauer and V. Sobotíková, *Finite element approximation of nonlinear elliptic problems with discontinuous coefficients*, RAIRO Modél. Math. Anal. Numér., 24 (1990), pp. 457–500.

[12] R. Glowinski and A. Marrocco, *Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité, d'une classe de problèmes de Dirichlet non linéaires*, Rev. Française Automat. Informat. Recherche Opérationnelle, 9 (1975), pp. 41–76.

[13] R. Glowinski, *Numerical methods for nonlinear variational problems*, Springer Series in Computational Physics, Springer-Verlag, New York, 1984.

[14] F. Hermeline, *A finite volume method for the approximation of diffusion operators on distorted meshes*, J. Comput. Phys., 160 (2000), pp. 481–499.

[15] F. Hermeline, *Approximation of diffusion operators with discontinuous tensor coefficients on distorted meshes*, Comput. Methods Appl. Mech. Engrg., 192 (2003), pp. 1939–1959.

[16] J. Leray and J.-L. Lions, *Quelques résultats de Višik sur les problèmes elliptiques nonlinéaires par les méthodes de Minty-Browder*, Bull. Soc. Math. France, 93 (1965), pp. 97–107.

[17] W. B. Liu, *Degenerate quasilinear elliptic equations arising from bimaterial problems in elastic-plastic mechanics*, Nonlinear Anal., 35 (1999), pp. 517–529.

[18] W. B. Liu, *Finite element approximation of a nonlinear elliptic equation arising from bimaterial problems in elastic-plastic mechanics*, Numer. Math., 86 (2000), pp. 491–506.

[19] C. Pierre, *Modélisation et simulation de l'activité électrique du couer dans le thorax, analyse numérique et méthodes de volumes finis*. Ph.D. thesis, Université de Nantes, 2005. http://tel.ccsd.cnrs.fr/tel-00010705.

[20] A. Ženíšek, *The finite element method for nonlinear elliptic equations with discontinuous coefficients*, Numer. Math., 58 (1990), pp. 51–77.

SIAM J. NUMER. ANAL.
Vol. 46, No. 6, pp. 3071–3083
© 2008 Society for Industrial and Applied Mathematics

PROJECTED PSEUDOTRANSIENT CONTINUATION*

C. T. KELLEY†, LI-ZHI LIAO‡, LIQUN QI§, MOODY T. CHU†, J. P. REESE¶, AND C. WINTON†

We dedicate this paper to the memory of H. B. Keller

Abstract. We propose and analyze a pseudotransient continuation algorithm for dynamics on subsets of R^N. Examples include certain flows on manifolds and the dynamic formulation of bound-constrained optimization problems. The method gets its global convergence properties from the dynamics and inherits its local convergence properties from any fast locally convergent iteration.

Key words. pseudotransient continuation, constrained dynamics, gradient flow, bound-constrained optimization, quasi-Newton method

AMS subject classifications. 65H10, 65H20, 65K10, 65L05

DOI. 10.1137/07069866X

1. Introduction. In this paper we extend algorithms and convergence results [9, 12, 13, 15, 24] for the method of pseudotransient continuation (Ψtc) to a class of constrained problems in which projections onto the feasible set are easy to compute. Such constraints arise in inverse eigenvalue and singular value problems [7, 8], where Ψtc is an efficient alternative to a fully time-accurate geometric integration [14], for which theory is only available for fixed time-step methods. The objective of Ψtc is not to stay on the manifold with high precision but rather to move rapidly to the steady state.

The algorithms we propose use the projection and can hence be applied to other classes of problems. Bound-constrained optimization is one example [22] and we explore that in this paper as well. The results in this paper may also be applicable to more general problems on manifolds [1] if the relevant projections can be approximated efficiently, and this will be the subject of future work.

Ψtc was originally designed as a method for finding steady-state solutions to time-dependent differential equations. The idea is to mimic integration to the steady state while managing the "time step" to move the iteration as rapidly as possible to Newton's method. This is different from the standard approach in an algorithm for initial value problems [2], where the time step is controlled with stability and accuracy in mind. Ψtc also differs from traditional continuation methods in that

*Received by the editors July 30, 2007; accepted for publication (in revised form) April 30, 2008; published electronically September 4, 2008.
http://www.siam.org/journals/sinum/46-6/69866.html

†Center for Research in Scientific Computation and Department of Mathematics, North Carolina State University, Box 8205, Raleigh, NC 27695-8205 (Tim_Kelley@ncsu.edu, mtchu@ncsu.edu, cwwinton@ncsu.edu). The work of these authors was partially supported by National Science Foundation grants DMS-0404537 and DMS-0707220 and Army Research Office grants W911NF-04-1-0276, W911NF-06-1-0096, W911NF-06-1-0412, and W911NF-07-1-0112.

‡Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Kowloon, Hong Kong, China (liliao@hkbu.edu.hk). The work of this author was partially supported by the Research Grant Council of Hong Kong.

§Department of Applied Mathematics, Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong, China (maqilq@polyu.edu.hk). The work of this author was partially supported by the Research Grant Council of Hong Kong.

¶School of Computational Science, Florida State University, Dirac Science Library, Tallahassee, FL 32306-4120 (jreese@scs.fsu.edu).

the objective is to find a steady-state solution, not, as is the case for pseudoarclength continuation [20], to track that solution as a function of another parameter. Homotopy methods [31] also introduce an artificial parameter to solve nonlinear equations but not in a way that is intended to capture dynamic properties, such as stability, of the solution.

In the remainder of this section, we will briefly review Ψtc. We refer the reader to [9, 12, 13, 15, 24] for the existing convergence results and references to applications.

In section 2 we will describe an algorithm for constrained Ψtc and prove a convergence theorem which not only allows for constraints but has a weaker stability assumption than was used in [9, 13, 24] and includes more general assumptions on the iteration itself. We close section 2 with some remarks on applications of the theory. In section 3, we will show how the new form of Ψtc can be applied to bound-constrained optimization in a way that maintains superlinear convergence in the terminal phase of the iteration. In section 4 we apply the methods to two example problems.

1.1. Ψtc for nonlinear equations. The formulation for ODE dynamics is the easiest to understand and is sufficient for this paper. Suppose $F : R^N \to R^N$ is Lipschitz continuously differentiable and

$$u^* = \lim_{t\to\infty} u(t), \tag{1.1}$$

where u is the solution of the initial value problem

$$\frac{du}{dt} = -F(u), \qquad u(0) = u_0. \tag{1.2}$$

We will refer to (1.1) as a stability condition in what follows, and it is a property of both F and the initial point u_0. The objective of the algorithms we discuss in this paper is to find u^*.

One might try to find u^* by solving the nonlinear equation $F(u) = 0$ with a globalized version of Newton's method [21, 23]. The danger is that one may find a solution other than u^* and even a solution that is dynamically unstable. One could also use an ODE code to accurately integrate the initial value problem (1.2) to the steady state. The problem with this latter approach is that one is accurately computing transient behavior of u that is not necessarily needed to compute u^*.

The most common form of Ψtc is the iteration

$$u_+ = u_c - \left(\delta_c^{-1} I + F'(u_c)\right)^{-1} F(u_c), \tag{1.3}$$

where, as is standard, u_c is the current iteration, δ_c the current time step, and u_+ the new iteration. The "time step" δ is managed in a way that captures important transients early in the iteration but grows near u^* so that (1.3) becomes Newton's method. One common way to control δ is "switched evolution relaxation" (SER) [29]. In SER the new time step is

$$\delta_+ = \min(\delta_c \|F(u_c)\|/\|F(u_+)\|, \delta_{max}). \tag{1.4}$$

Using $\delta_{max} = \infty$ is common, in which case $\delta_+ = \delta_0 \|F(u_0)\|/\|F(u_+)\|$. We will refer to the updated formula (1.4) as SER-A, to distinguish it from (1.5).

The existing convergence theory applies only to SER-A, but there are other approaches, and we will discuss two. A variation of SER, which we call SER-B, was proposed in [19]. The formula for the time step is

$$\delta_+ = \max(\delta_c / \|u_+ - u_c\|, \delta_{max}). \tag{1.5}$$

In section 2.1.1, we show how SER-B can be modified so that the convergence theory applies in the case of certain gradient flows.

The temporal truncation error approach (TTE) [25] estimates the local truncation error of $(u)_i(t_n)$, the ith component of $u(t_n) \in R^N$, by $\tau_i \equiv \frac{\delta_n^2 (u)_i''(t_n)}{2}$, approximates $(u)_i''$ by

$$\frac{2}{\delta_{n-1}+\delta_{n-2}}\left[\frac{((u)_i)_n - ((u)_i)_{n-1}}{\delta_{n-1}} - \frac{((u)_i)_{n-1} - ((u)_i)_{n-2}}{\delta_{n-2}}\right], \tag{1.6}$$

and computes δ_n by setting $\tau_i = 3/4$ for all i. In section 4 we will compare the three methods for time-step management. The good performance of SER-B and TTE raises interesting research questions.

One way to see how Ψtc and temporal integration are related is to derive Ψtc from the implicit Euler method. The formula for an implicit Euler step is $u^{n+1} = u^n - \delta_n F(u^{n+1})$, where u^n is the approximation to u at t_n and $\delta_n = t_{n+1} - t_n$ is the nth time step. If we determine u^{n+1} by taking a single Newton step for the nonlinear equation $G(u) \equiv u - u^n + \delta_n F(u) = 0$, with u^n as the initial iterate, then we obtain (1.3).

2. Constrained Ψtc . Let F be Lipschitz continuous, and assume that $u(t) \in \Omega$ for all $t \geq 0$, where $\Omega \subset R^N$. Examples of such constrained dynamics are flows where F is the projected gradient onto the tangent space of Ω at u, and our two examples are such flows. One should not expect a general purpose integrator to keep the solutions in Ω, and we use a projection to correct after each step to keep the iterations in Ω.

Let $\mathcal{P}$ be a Lipschitz continuous projection onto Ω. Our assumptions on $\mathcal{P}$ are as follows.

Assumption 2.1.

1. $\mathcal{P}(u) = u$ for all $u \in \Omega$.
2. There are $M_{\mathcal{P}}$ and $\epsilon_{\mathcal{P}}$ such that for all $u \in \Omega$ and v such that $\|v - u\| \leq \epsilon_{\mathcal{P}}$

$$\|\mathcal{P}(v) - u\| \leq \|v - u\| + M_{\mathcal{P}}\|v - u\|^2. \tag{2.1}$$

Assumption 2.1 is trivially true if Ω is convex and $\mathcal{P}$ is the projection onto Ω, for then $\mathcal{P}$ is Lipschitz continuous with Lipschitz constant 1, so $M_{\mathcal{P}} = 0$. If Ω is a smooth manifold of the form $\Omega = \{u \,|\, \mathcal{F}(u) = 0\}$ where $\mathcal{F} : R^N \to R^M$ with $M < N$, and $\mathcal{P}$ is smooth, then $\|\mathcal{P}'(u)\| = 1$, which will imply (2.1). Note that we are making no explicit assumptions on Ω, only assumptions on the existence of $\mathcal{P}$ and its properties.

We will consider a Ψtc iteration of the form

$$u_+ = \mathcal{P}\left[u_c - \left(\delta_c^{-1} I + H(u_c)\right)^{-1} F(u_c)\right], \tag{2.2}$$

where H is an $N \times N$ matrix-valued function of u. We will assume that H is a sufficiently good approximation to F' (or, in the semismooth case, sufficiently close to ∂F) to make the iteration locally convergent. The theory we will develop applies equally well to the inexact formulation $u_+ = u_c + s$, where

$$\|(\delta_c^{-1} I + H(u_c))s + F(u_c)\| \leq \eta_c \|F(u_c)\|. \tag{2.3}$$

The "forcing term" η_c could be, for example, the termination tolerance in a linear iterative method for computing the step [10, 21]. The forcing term has a well-known effect on convergence analysis, which is reflected in (2.11).

Assumption 2.2.

1. There are $M_H, \epsilon_H > 0$ such that

$$\|H(u)\| \le M_H \text{ for all } u \in S(\epsilon_H) = \left\{ z \mid \inf_{t \ge 0} \|z - u(t)\| \le \epsilon_H \right\}. \tag{2.4}$$

   For all $\epsilon > 0$ there is $\bar{\epsilon} > 0$ such that if $u \in S(\epsilon_H)$ and $\|u - u^*\| > \epsilon$, then

$$\|F(u)\| > \bar{\epsilon}. \tag{2.5}$$

2. There is ϵ_L so that if $\|u_c - u^*\| \le \epsilon_L$, then $H(u_c)$ is nonsingular,

$$\|(I + \delta H(u_c))^{-1}\| \le (1 + \beta\delta)^{-1}, \text{ for some } \beta > 0 \text{ and all } \delta \ge 0, \tag{2.6}$$

   and the local Newton iteration

$$u_+^{NL} = u_c - H(u_c)^{-1} F(u_c) \tag{2.7}$$

   reduces the error by a factor $r \in [0, 1)$ for all $u_c \in \Omega$ sufficiently near u^*, i.e.,

$$\|e_+^{NL}\| \le r \|e_c\|. \tag{2.8}$$

One new feature in Assumption 2.2 is the local nonlinear iteration, which is general enough to allow for Gauss–Newton methods (and quasi-Newton methods, see section 2.1.3). Note that the convergence rate for the local iteration is expressed in terms of the underlying unprojected Newton-like method (2.7). The assumption we must make (2.8) on the unprojected iteration can be verified in the examples we consider in section 4.

Theorem 2.1 extends previous work in several ways. The smoothness assumptions on F are relaxed, the projection is introduced to handle constrained dynamics, H is constrained only by the local convergence behavior of the Newton-like iteration (2.7), so superlinear convergence is not required, and (2.6) need only hold in a neighborhood of u^*.

THEOREM 2.1. *Let F be locally Lipschitz continuous, and assume that*

$$\lim_{t \to \infty} u(t) = u^*$$

and that Assumptions 2.1 *and* 2.2 *hold. Let the sequence $\{\delta_n\}$ be updated with* (1.4). *Assume that there is $\delta^* > 0$ such that*

$$M_{\mathcal{P}} \epsilon_L / \beta < \delta^* \le \delta_n \tag{2.9}$$

for all n. Assume that the q-factor r in (2.8) *satisfies*

$$r < ((1 + M_{\mathcal{P}} \epsilon_L) - (1 + \beta\delta^*)^{-1})/2, \tag{2.10}$$

where β is the constant in (2.6). *Then if δ_0 and the sequence $\{\eta_n\}$ are sufficiently small, the inexact Ψtc iteration $u_{n+1} = \mathcal{P}(u_n + s_n)$, where $\|(\delta_n^{-1} I + H(u_n))s_n + F(u_n)\| \le \eta_n \|F(u_n)\|$ converges to u^*. Moreover, there is $K > 0$ such that for n sufficiently large*

$$\|e_{n+1}\| \le \|e_{n+1}^{NL}\| + K \|e_n\| \left(\eta_n + \delta_n^{-1}\right). \tag{2.11}$$

Proof. We will prove the result for the exact ($\eta_n = 0$) iteration with $\delta_{max} = \infty$ in (1.4). The complete proof is based on the same ideas but requires more bookkeeping. The outline of the proof follows those in [9, 13, 24].

We begin with the global phase. We wish to prove that, while u is out of the local convergence region for the iteration (2.7), the iteration remains close to the solution of the differential equation, i.e., in $S(\epsilon)$ for a sufficiently small ϵ. Only (1.1), the lower bound $\delta_n \geq \delta^*$, Lipschitz continuity of F, and Part 2.2 of Assumption 2.2 are needed for this stage of the analysis. Having done this, we address the local phase, where δ is small and u is near u^*. We will use the rest of Assumption 2.2 to prove the local convergence estimate (2.11).

Let $\epsilon < \min(\epsilon_L, \epsilon_H)$. We may reduce ϵ as the proof of the local convergence progresses. The first step is to show that if δ_0 is sufficiently small, then

$$\|u_n - u^*\| < \epsilon \tag{2.12}$$

for sufficiently large n. To do this we need only verify that, for δ sufficiently small,

$$\left(\delta^{-1} I + H(u_n)\right)^{-1} = \delta I + O\left(\delta^2\right), \tag{2.13}$$

and obtain upper and lower bounds on δ_n/δ_0 while (2.12) fails to hold.

The estimate (2.13) follows from (2.4) if $\delta < 1/(2M_H)$. To obtain bounds for δ_n, we can apply the update formula (1.4) and (2.5) to show that while $u_n \in S(\epsilon_H)$ and $\|u_n - u^*\| > \epsilon/2$, then

$$\begin{aligned} \delta^* &\equiv \delta_0 \|F(u_0)\| / \max_{u \in S(\epsilon_H)} \|F(u)\| \\ &\leq \delta_n = \delta_0 \|F(u_0)\| / \|F(u_n)\| \leq 2\delta_0 \|F(u_0)\| / C_F \epsilon. \end{aligned} \tag{2.14}$$

Equation (1.4) with $\delta_{max} = \infty$ and our lower bound on δ imply that

$$\delta^* \leq \delta_n \leq \frac{\|F(u_0)\|\delta_0}{\|F(u_n)\|} \leq \frac{2\delta_0 \|F(u_0)\|}{C_F \epsilon}. \tag{2.15}$$

Hence, for δ_0 sufficiently small, (2.13) holds if (2.12) does not.

With (2.13) in hand, we see that either (2.12) holds or

$$u_{n+1} = \mathcal{P}(u_n - \delta F(u_n)) + O\left(\delta^2\right), \tag{2.16}$$

where the constant in the O-term is independent of n. Now, if $u \in \Omega$, then Lipschitz continuity of $\mathcal{P}$ implies that

$$\mathcal{P}\left(u - \left(\delta^{-1} I + H(u)\right)^{-1}\right) F(u) = \mathcal{P}(u - \delta F(u)) + O\left(\delta^2\right).$$

Euler's method, $u_{n+1} = u_n - \delta F(u)$, has the same local truncation error as $u_{n+1} = \mathcal{P}(u_n - \delta F(u))$, because $u(t) \in \Omega$ for all t [26]. To see this, we note that

$$\mathcal{P}(u(t) - \delta F(u(t))) = \mathcal{P}\left(u(t+\delta) + O\left(\delta^2\right)\right) = u(t+\delta) + O\left(\delta^2\right).$$

Now let T be such that $\|u(t) - u^*\| < \epsilon/2$ for all $t \geq T$, and let N^* be the least integer $\geq T/\delta^*$. The standard analysis for the forward Euler method [2] implies that there is C_E such that

$$\|u_n - u(t_n)\| \leq C_E \max_{1 \leq k \leq n} \delta_k \leq \frac{2C_E \delta_0 \|F(u_0)\|}{C_F \epsilon}$$

for all $n \leq N^*$. In particular, $\|u_n - u^*\| \leq \epsilon$ for all $n \geq N^*$ if $\delta_0 \leq (C_F \epsilon^2)/(4C_E \|F(u_0)\|)$.

For the local phase, assume that $n \geq N^*$. Hence, (2.12) holds. We need to show that $\|e_{n+1}\| < \|e_n\|$ and that $\delta_{n+1} > \delta_n$. Once those things are done, we can complete the proof with a simple calculation.

Define $v_{n+1} = u_n - (\delta_n^{-1} I + H(u_n))^{-1} F(u_n)$, and note that $(H(u_n)^{-1} - (\delta_n^{-1} I + H(u_n))^{-1}) = (I + \delta_n H(u_n))^{-1} H(u_n)^{-1}$. Hence

$$\begin{aligned} v_{n+1} &= u_n - H(u_n)^{-1} F(u_n) + (H(u_n)^{-1} - (\delta_n^{-1} I + H(u_n))^{-1}) F(u_n) \\ &= u_{n+1}^{NL} + (I + \delta_n H(u_n))^{-1} H(u_n)^{-1} F(u_n) \\ &= u_{n+1}^{NL} - (I + \delta_n H(u_n))^{-1} \left(u_{n+1}^{NL} - u_n \right), \end{aligned}$$

and therefore $u_{n+1} = \mathcal{P}(v_{n+1}) = \mathcal{P}(u_{n+1}^{NL} - (I + \delta_n H(u_n))^{-1} (e_{n+1}^{NL} - e_n))$.

We use (2.1) to conclude that

$$\begin{aligned} e_{n+1} &= \mathcal{P}(v_{n+1}) - u^* = \mathcal{P}\left(u_{n+1}^{NL} - (I + \delta_n H(u_n))^{-1} \left(e_{n+1}^{NL} - e_n \right) \right) - u^* \\ &= \mathcal{P}\left(u_{n+1}^{NL} - (I + \delta_n H(u_n))^{-1} \left(e_{n+1}^{NL} - e_n \right) \right) - \mathcal{P}(u^*). \end{aligned}$$

So, since $\epsilon < \epsilon_L$,

$$\begin{aligned} \|e_{n+1}\| &\leq (1 + M_{\mathcal{P}} \epsilon_L) \left(\|e_{n+1}^N\| + \|(I + \delta_n H(u_n))^{-1}\| \left(\|e_{n+1}^{NL}\| + \|e_n\| \right) \right) \\ &\leq (1 + M_{\mathcal{P}} \epsilon_L) \left(2r + (1 + \beta \delta^*)^{-1} \right) \|e_n\|. \end{aligned}$$

This completes the proof since $(1 + M_{\mathcal{P}} \epsilon_L)(2r + (1 + \beta \delta^*)^{-1}) < 1$ by (2.9). Hence, the iteration converges at least locally q-linearly. Formula (1.4) then will imply (2.11). ∎

2.1. Remarks. Even in the unconstrained case (where $\mathcal{P}(u) = u$ for all $u \in R^N$) Theorem 2.1 extends the results from [9, 13, 24] by replacing the semismoothness and the inexact Newton condition with a general condition on the convergence of the local iteration. If the stability condition (1.1) holds, then Ψtc with SER is a convergent iteration for unconstrained optimization, even if one does not use exact Hessians.

2.1.1. Control of δ with f for gradient flows. The key parts to the proof of Theorem 2.1 are showing that the time step remains small until u_n is near the solution and that the step will grow at that time. The way this is done in the case of SER-A is to note that (1.4) implies that if u_n is not close to u^*, then δ_n is bounded from above and below by constant multiples of δ_0, and hence the method is an accurate temporal integration. Then, once near u^*, (2.6) drives u_n to u^* and δ_n to δ_{max}.

One can augment the time-step control method in several ways without affecting the theory. If the dynamics are a gradient flow, then one can reject the step if f is increased, reduce δ, and try again. As long as δ is bounded from above and below by constant multiples of δ_0 and the number of reductions is bounded, the global convergence assertion $u_n \to u^*$ will hold. If one accepts the SER-A formula only when f decreases, then the local theory will be unchanged and Theorem 2.1 will hold. One can also apply this idea to SER-B and TTE but must take care that increases in δ are not too large to maintain the resolution of the dynamics while $u(t)$ is far from u^*. One way to do this is to accept the SER-B or TTE step only if f decreases and limit the size of the increases in δ to factors of two at each iteration.

Another approach is the trust-region method from [15], and the proof of Theorem 2.1 extends to that method. Here $\delta_{max} = \infty$ and the increase in δ at each iteration is limited.

2.1.2. Semismooth nonlinearities. If F is semismooth, which is the case considered in [13], and $H(u_n) \in \partial F(u_n)$, then the local iteration (2.7) is superlinearly convergent, if all matrices in $\partial F(u^*)$ are nonsingular. Hence, we may recover the results from [13] from Theorem 2.1, the extension to DAE dynamics being the same as that in [13] and not relevant to this paper.

2.1.3. Quasi-Newton methods. The proof of Theorem 2.1 can be modified to allow a quasi-Newton [11, 22] model Hessian. This modification simply replaces the dependence of H on u_n with one on the history of the iteration and makes the assumptions of boundedness of H_n and convergence of the local iteration. The resulting modified proof would be somewhat longer, and, since our examples are not quasi-Newton iterations, we used the shorter and simpler proof in this paper.

Unlike Newton or Gauss–Newton iterations, one must assume (rather than verify) the convergence of the local iteration. The reason for this is that local convergence of quasi-Newton methods requires that both the initial iterate and the initial model Hessian be good approximations to the solution and the Hessian at the solution. The global theory [5] requires a line search and convex level sets, neither of which is a part of the Ψtc methods we propose.

2.1.4. Gauss–Newton iteration. If $F(u) = \nabla f(u) = R'(u)^T R(u)$ is the gradient of a nonlinear least squares functional $f(u) = R(u)^T R(u)/2$, where $R : R^N \to R^M$, with $M > N$, we may let H be the Gauss–Newton model Hessian $H(u) = R'(u)^T R'(u)$ and apply Theorem 2.1 if R' has full column rank at the minimizer. Moreover, many of the assumptions can be verified with ease in this case. If $u(t) \to u^*$, which we still must assume, then $H(u^*)$ is symmetric and positive definite, so (2.6) holds. Moreover $\delta^{-1} I + H(u)$ is nonsingular for all $\delta > 0$ and all u, because $H(u)$ is always nonnegative definite. For a zero-residual problem, the estimate (2.10) will hold because the local iteration converges q-quadratically if R is Lipschitz continuously differentiable. In this case, Ψtc is a version of the Levenberg–Marquardt method, where the parameter is selected based on the norm of the gradient rather than with a trust region scheme. Similar ideas for selection of the parameter have been made [22].

3. Bound-constrained optimization. The bound-constrained optimization problem is

$$\min_{\Omega} f(u), \tag{3.1}$$

where $\Omega = \{u \,|\, L \le u \le U\}$, and the inequalities are componentwise.

In order to describe necessary conditions and formulate the algorithms, we must recall some notation from [4, 22, 27].

The l^2 projection onto Ω is $\mathcal{P}$, where

$$\mathcal{P}(u)_i = \max(L_i, \min(U_i, (u)_i)). \tag{3.2}$$

Here $(u)_i$ denotes the ith component of the vector $u \in R^N$. $\mathcal{P}$ trivially satisfies Assumption 2.1 because Ω is convex.

We will assume that f is Lipschitz continuously differentiable. In that case, the first-order necessary conditions for optimality are [4, 22]

$$F(u) = u - \mathcal{P}(u - \nabla f(u)) = 0. \tag{3.3}$$

Equation (3.3) is a semismooth nonlinear equation. Fast locally convergent methods include the semismooth Newton, methods of [30], and the projected Newton or scaled gradient projection methods [4, 22, 27].

Consistent with the unconstrained case, the gradient flow equations are [26]

$$\frac{du}{dt} = -F(u), \qquad u(0) = u_0, \tag{3.4}$$

where in this case, F is defined by (3.3). If we let $u_0 \in \Omega$, then the solution of (3.4) satisfies

$$\lim_{t\to\infty} F(u(t)) = 0. \tag{3.5}$$

We will also assume that (1.1) holds. Since dynamics [26] force $u(t) \in \Omega$ for all t and Ω is bounded, then if there are only finitely many solutions of $F(u) = 0$ in Ω, (1.1) will hold for all $u_0 \in \Omega$ in this case.

We may apply Theorem 2.1 directly, once we describe the maps $H(u)$ and show that (2.8) holds. We use known results [4, 22, 27] from optimization to do this.

One choice of $H(u)$ is the reduced Hessian. Letting $u \in \Omega$ and $0 \le \sigma < \min(U_i - L_i)/2$, we define the set of σ-binding constraints as

$$\begin{aligned}\mathcal{B}^\sigma(u) \quad = \{i \,|\quad & U_i - (u)_i \le \sigma \text{ and } (\nabla f(u))_i < -\sqrt{\sigma} \text{ or} \\ & (u)_i - L_i \le \sigma \text{ and } (\nabla f(u))_i > \sqrt{\sigma}\}.\end{aligned}$$

For $\mathcal{N} \subset \{1, 2, \ldots N\}$ we define $D(\mathcal{N})$ as the diagonal matrix with entries

$$D(\mathcal{N})_{ii} = \begin{cases} 1 & i \in \mathcal{N}, \\ 0 & i \notin \mathcal{N} \end{cases}$$

and define the reduced Hessian as

$$\bar{\mathcal{R}} f(u) = I - D(\mathcal{B}^0(u))(I - \nabla^2 f(u))D(\mathcal{B}^0(u)). \tag{3.6}$$

The second-order sufficiency conditions for a point u^* to be a local minimizer are [27] that $F(u^*) = 0$ and $\bar{\mathcal{R}} f(u^*)$ is positive definite. We will assume that u^* satisfies the second-order sufficiency conditions in what follows.

One must approximate the bounding constraints carefully in order to obtain a superlinearly convergent iteration. One way to do this [22, 27] is to let

$$H(u) = I - D(\mathcal{B}^\sigma(u)) \left(I - \nabla^2 f(u)\right) D(\mathcal{B}^\sigma(u)), \tag{3.7}$$

where $\sigma(u) = \|u - \mathcal{P}(u - \nabla f(u))\|$. With this choice the local iteration satisfies (2.8). In the case of a small residual bound-constrained nonlinear least squares problem, we may replace $\nabla^2 f(u)$ with the Gauss–Newton model Hessian.

4. Examples. In this section we present two examples. The first is a nonlinear equation on a manifold, which is a gradient flow of an inverse singular value problem. The projection is more subtle in this case, and we describe it in detail in section 4.1.1.

The second is a nonlinear bound-constrained least squares problem for which we compare the three variants of Ψtc (SER-A, SER-B, and TTE) with the trust-region method (`lmtr`) from [15], which in the nonlinear least squares case is a classic trust-region method [11, 22], and the damped Levenberg–Marquardt algorithm (`lmls`) from [22]. The projection in this case is trivial to compute, and we use the reduced Gauss–Newton model Hessian instead of $\nabla^2 f$ in (3.7). This example is a small artificial problem, which enables us to consider the cases where the minimizer is in the interior

of the feasible set, on the boundary (and hence degenerate in the sense that the binding constraints are a proper subset of the active constraints), and outside of the feasible set (and therefore a nonzero residual problem). In this example we do not increase δ unless f decreases, and we manage δ with the approach in section 2.1.1. All of the Ψtc methods, and especially SER-B and TTE, work much better if we do that.

In all of the examples, SER-B performs very well.

4.1. Inverse singular value problem. The inverse singular value problem [7] is to find $c \in R^N$ so that the $M \times N$ matrix

$$B(c) = B_0 + \sum_{k=1}^{N} c_k B_k$$

has prescribed singular values $\{\sigma_i\}_{i=1}^N$. This is one example of a wide class of inverse eigenvalue and singular value problems for which a dynamic formulation is useful [8].

One can assume without loss of generality that the matrices $\{B_i\}_{i=1}^N$ are orthonormal with respect to the Frobenius inner product and then formulate the problem as a constrained nonlinear least squares problem

$$\min \Psi(U,V) \equiv \|R(U,V)\|_F^2 \tag{4.1}$$

for $M \times M$ and $N \times N$ matrices U and V, subject to the constraint that U and V be orthogonal. If one finds a solution with a zero residual, then one has solved the original problem. This is not always possible, as the original problem may not have a solution [6]. In (4.1) the residual is

$$R(U,V) = U\Sigma V^T - B_0 - \sum_{k=1}^{N} \langle U\Sigma V^T, B_k\rangle B_k,$$

where $\langle\cdot,\cdot\rangle$ is the Frobenius inner product.

If we let Ω denote the manifold of pairs of $M \times M$ and $N \times N$ orthogonal matrices, then the projection of $\nabla\Psi$ onto the tangent space of Ω at $(U,V) \in \Omega$ is

$$g(U,V) = \frac{1}{2} \begin{pmatrix} \left(R(U,V)V\Sigma^T U^T - U\Sigma V^T R(U,V)^T\right) U \\ \left(R(U,V)^T U\Sigma V^T - V\Sigma^T U^T R(U,V)\right) V \end{pmatrix}.$$

The gradient flow equations for the problem are of the form (1.2) with

$$F(u) = g(U,V), \text{ where } u = \begin{pmatrix} U \\ V \end{pmatrix}. \tag{4.2}$$

Since $F(u)$ is in the tangent space of Ω at u [7], the solution of (1.2) is in Ω if $u_0 \in \Omega$. Since g is analytic in u, the results of [28] will apply, and so (1.1) holds for all initial vectors $u_0 \in \Omega$.

4.1.1. The projection onto Ω. The projection of an $N \times N$ matrix A onto the manifold of orthogonal matrices [17, 18] $A \to U_P$. Here $A = U_P H_P$, with U_P orthogonal and H_P symmetric positive semidefinite, is a polar decomposition of A. H_P is unique. U_P is unique if A is nonsingular. In this case (2.1) will hold. Since $S(\epsilon)$ is near a curve of orthogonal matrices, which have full rank, the possible singularity

of A is not an issue for us. One can compute U_P directly from the singular value decomposition $A = U\Sigma V^T$ of A as $U_P = UV^T$. This is efficient when A is small. For large A, there are several efficient iterative methods [16, 18].

So, in the context of this paper, given a pair of $N \times N$ matrices $w = (A, B)^T$,

$$\mathcal{P}(w) = \begin{pmatrix} U_P^A \\ U_P^B \end{pmatrix},$$

where U_P^A and U_P^B are the orthogonal parts of the polar decompositions of A and B.

4.1.2. Convergence of the local method. In this section we will verify that (2.8) holds. The local method uses the reduced Gauss–Newton model Hessian, which requires the projection $P_T(u)$ onto the tangent space at a point $u \in \Omega$. One can compute that projection by noting that if $w(t)$ is a differentiable orthogonal matrix-valued function, then differentiating $w^T(t)w(t) = I$,

$$\frac{dw(t)^T w(t)}{dt} = \dot{w}(t)^T w(t) + w(t)^T \dot{w}(t) = 0,$$

and hence $w^T\dot{w}$ is skew-symmetric. This implies that the tangent space for the manifold of orthogonal matrices at a point U is the space of matrices W for which U^TW is skew-symmetric. The projection onto the tangent space can then be computed as follows. If $\{S_i\}$ is a Frobenius-orthonormal basis for the skew-symmetric matrices, then a Frobenius-orthonormal basis for the tangent space is $\{US_i\}$, which we can use to compute the projection. If we do this for each component of $u = (U, V)^T$, we obtain $P_T(u)$. Alternatively we could use

$$P_T(u) = \mathcal{P}'(u) \text{ for all } u \in \Omega\ , \tag{4.3}$$

which follows from the fact that $\mathcal{P}$ is a map-to-nearest.

The local method for $F(u) = 0$ uses $H(u) = (I - P_T(u)) + P_T(u)F'(u)P_T(u)$, and we will show that $u_+^{NL} = u_c - H(u_c)^{-1}F(u_c)$ satisfies $\|e_+^{NL}\| = O(\|e_c\|^2)$, which implies (2.8) and will allow us to apply Theorem 2.1. We do this by noting that for all $u \in \Omega$

$$\begin{aligned} F(u) &= P_T(u)F(u) = P_T(u)F'(u)e + O\left(\|e\|^2\right) \\ &= H(u)e - (I - P_T(u))e + P_T(u)F'(u)(I - P_T(u))e + O\left(\|e\|^2\right) \\ &= H(u)e + O\left(\|(I - P_T(u))e\| + \|e\|^2\right). \end{aligned} \tag{4.4}$$

If $u \in \Omega$ is near u^*, then we can use (4.3) and the Lipschitz continuity of $\mathcal{P}$ to conclude that

$$u = \mathcal{P}(u) = \mathcal{P}(u^*) + \mathcal{P}'(u)e + O\left(\|e\|^2\right) = P_T(u)e + O\left(\|e\|^2\right),$$

and so $(I - P_T(u))e = O(\|e\|^2)$. Hence $u_+^{NL} - u^* = u_c - u^* - H(u_c)^{-1}F(u_c) = O(\|e_c\|^2)$.

4.1.3. Computations. We take the example from [7], having orthonormalized the matrices $\{B_j\}_{j=0}^4$ with classical Gram–Schmidt and the Frobenius inner product. In Figure 4.1 we compare the relative performance of the three time-step management strategies SER-A, SER-B, and TTE. While TTE does poorly, it is interesting to see that both versions of SER do well.

We did not reduce δ to respond to increases in $\|\Psi\|$ in this example, and the SER-A and SER-B iterations still converged well.

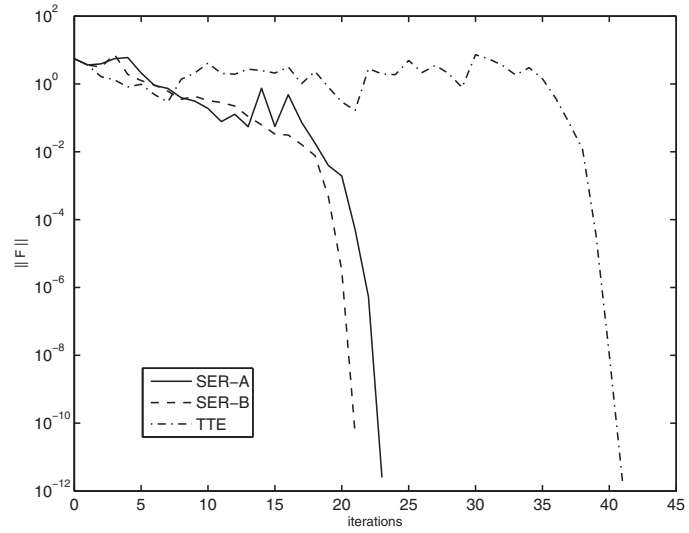


FIG. 4.1. *Inverse singular value problem.*

4.2. Inverse problem. This small ($N = 2$) example is taken from [3, 22]. We seek to identify the damping coefficient c and spring constant k for a simple harmonic oscillator. The governing differential equation is

$$w'' + cw' + kw = 0; \qquad w(0) = w_0, w'(0) = 0 \tag{4.5}$$

on the interval $[0, 1]$. We let $u = (c, k)^T$ and fit samples of the exact solution at 100 equally spaced points. We let $c = k = 1$ be the parameter values for the true solution and use `ode15s` from MATLAB to integrate (4.5) with the approximate parameters. The relative and absolute error tolerances were 10^{-6}.

The function to be minimized is

$$f(u) = \frac{1}{2} R(u)^T R(u) = \frac{1}{2} \sum_{i=1}^{100} (w^{exact}(t_i) - w_i(u))^2,$$

where $t_i = i/100$, $w_i(u)$ is the solution returned by `ode15s` with $u = (c, k)^T$, and w^{exact} is the solution of (4.5) with $(c, k) = (1, 1)$. The upper bounds are $[10, 10]$, and we consider three cases for the lower bounds: $[0, 0]$, placing the global minimizer in the interior of the feasible region, $[1, 0]$, placing the global minimizer on the boundary, and $[2, 0]$, placing the global minimizer outside. In the last case the solution of the unconstrained zero-residual problem does not satisfy the bound constraints. The residual at the optimal point for the constrained problem is 21.5.

The initial iterate in all cases was $(c, k) = (10, 10)$. In this computation we set $\delta_0 = 1/100$ and terminate the continuation when either the norm of the projected gradient $\|F\|$ has been reduced by a factor of 10^3 or $f < 10^{-6}$.

In Figure 4.2 we plot the values of f and $\|F\|$ as functions of the iteration count for several variations of Ψtc : SER-A (`ser-a`), SER-B (`ser-b`), TTE (`tte`), and the trust-region Levenberg–Marquardt method `lmtr` from [15]. We also compare them with the Levenberg–Marquardt line search method (`lmls`) from [22]. For SER-A, SER-B, and TTE, we rejected any step that increased the residual and decreased the time step by factors of 2 until either the residual decreased or $dt = 10^{-4}$. In the latter case we terminated the iteration.

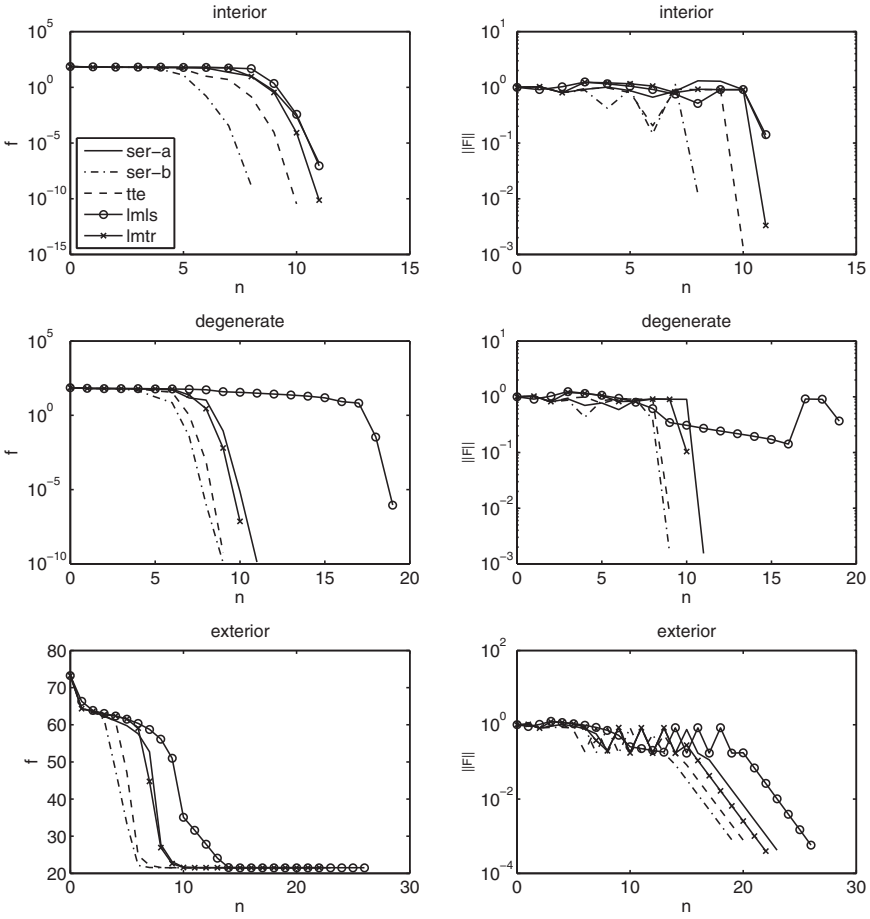


FIG. 4.2. *Parameter ID example.*

The damped Levenberg–Marquardt method from [22] is not as effective as the other four approaches, and SER-B is consistently better than the others.

5. Conclusions. We have described and analyzed a generalization of the pseudo-transient continuation algorithm which can be applied to a class of constrained nonlinear equations. The new approach can be applied to bound-constrained problems and to certain inverse eigenvalue and singular value problems.

We have reported on numerical testing which illustrates the performance of the method.

REFERENCES

[1] P.-A. ABSIL, C. G. BAKER, AND K. A. GALLIVAN, *Trust-region methods on Riemannian manifolds*, Found. Comput. Math., 7 (2007), pp. 303–330.

[2] U. M. ASCHER AND L. R. PETZOLD, *Computer Methods for Ordinary Differential Equations and Differential Algebraic Equations*, SIAM, Philadelphia, 1998.

[3] H. T. BANKS AND H. T. TRAN, *Mathematical and Experimental Modeling of Physical Processes*, Department of Mathematics, North Carolina State University, unpublished lecture notes for Mathematics, 1997, pp. 573–574.

[4] D. P. Bertsekas, *Projected Newton methods for optimization problems with simple constraints*, SIAM J. Control Optim., 20 (1982), pp. 221–246.
[5] R. H. Byrd and J. Nocedal, *A tool for the analysis of quasi-Newton methods with application to unconstrained minimization*, SIAM J. Numer. Anal., 26 (1989), pp. 727–739.
[6] D. Chu and M. T. Chu, *Low rank update of singular values*, Math. Comp., 75 (2006), pp. 1351–1366.
[7] M. T. Chu, *Numerical methods for inverse singular value problems*, SIAM J. Numer. Anal., 29 (1992), pp. 885–903.
[8] M. T. Chu and G. H. Golub, *Inverse Eigenvalue Problems: Theory, Algorithms, and Applications*, Oxford Science, New York, 2005.
[9] T. Coffey, C. T. Kelley, and D. E. Keyes, *Pseudo-transient continuation and differential-algebraic equations*, SIAM J. Sci. Comput., 25 (2003), pp. 553–569.
[10] R. Dembo, S. C. Eisenstat, and T. Steihaug, *Inexact Newton methods*, SIAM J. Numer. Anal., 19 (1982), pp. 400–408.
[11] J. E. Dennis and R. B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Classics Appl. Math. 16, SIAM, Philadelphia, 1996.
[12] P. Deuflhard, *Adaptive Pseudo-Transient Continuation for Nonlinear Steady State Problems*, Technical report 02-14, Konrad-Zuse-Zentrum für Informationstechnik, Berlin, 2002.
[13] K. R. Fowler and C. T. Kelley, *Pseudo-transient continuation for nonsmooth nonlinear equations*, SIAM J. Numer. Anal., 43 (2005), pp. 1385–1406.
[14] E. Hairer, C. Lubich, and G. Wanner, *Geometric numerical integration*, 2nd ed., Springer Ser. Comput. Math. 31, Springer-Verlag, Berlin, 2006.
[15] D. J. Higham, *Trust region algorithms and time step selection*, SIAM J. Numer. Anal., 37 (1999), pp. 194–210.
[16] N. J. Higham, *Computing the polar decomposition – with applications*, SIAM J. Sci. Comput., 7 (1986), pp. 1160–1174.
[17] N. J. Higham, *Matrix nearness problems and applications*, in Applications of Matrix Theory, M. J. C. Glover and S. Barnett, eds., Oxford University Press, London, 1989, pp. 1–27.
[18] N. J. Higham, D. S. Mackey, N. Mackey, and F. Tisseur, *Computing the polar decomposition and the matrix sign decomposition in matrix groups*, SIAM J. Matrix Anal. Appl., 25 (2004), pp. 1178–1192.
[19] H. Jiang and P. A. Forsyth, *Robust linear and nonlinear strategies for solution of the transonic Euler equations*, Comput & Fluids, 24 (1995), pp. 753–770.
[20] H. B. Keller, *Lectures on Numerical Methods in Bifurcation Theory*, Tata Institute of Fundamental Research, Lectures on Mathematics and Physics, Springer-Verlag, New York, 1987.
[21] C. T. Kelley, *Iterative Methods for Linear and Nonlinear Equations*, Frontiers Appl. Math. 16, SIAM, Philadelphia, 1987.
[22] C. T. Kelley, *Iterative Methods for Optimization*, Frontiers Appl. Math. 18, SIAM, Philadelphia, 1987.
[23] C. T. Kelley, *Solving Nonlinear Equations with Newton's Method*, Fundam. Algorithms 1, SIAM, Philadelphia, 1987.
[24] C. T. Kelley and D. E. Keyes, *Convergence analysis of pseudo-transient continuation*, SIAM J. Numer. Anal., 35 (1998), pp. 508–523.
[25] D. E. Keyes and M. D. Smooke, *A parallelized elliptic solver for reacting flows*, in Parallel Computations and Their Impact on Mechanics, A. K. Noor, ed., American Society of Mechanical Engineers, New York, 1987, pp. 375–402.
[26] L.-Z. Liao, H. Qi, and L. Qi, *Neurodynamical optimization*, J. Global Optim., 28 (2004), pp. 175–195.
[27] C.-J. Lin and J. J. Moré, *Newton's method for large bound-constrained optimization problems*, SIAM J. Optim., 9 (1999), pp. 1100–1127.
[28] S. Łojasiewicz and M. A. Zurro, *On the gradient inequality*, Bull. Pol. Acad. Sci. Math., 47 (1999), pp. 143–145.
[29] W. Mulder and B. V. Leer, *Experiments with implicit upwind methods for the Euler equations*, J. Comput. Phys., 59 (1985), pp. 232–246.
[30] L. Qi and J. Sun, *A nonsmooth version of Newton's method*, Math. Program., 58 (1993), pp. 353–367.
[31] L. T. Watson, S. C. Billups, and A. P. Morgan, *Algorithm 652: HOMPACK: A suite of codes for globally convergent homotopy algorithms*, ACM Trans. Math. Software, 13 (1987), pp. 281–310.

SIAM J. NUMER. ANAL.
Vol. 46, No. 6, pp. 3084–3112
© 2008 Society for Industrial and Applied Mathematics

CONVERGENCE OF FOURIER-WAVELET MODELS FOR GAUSSIAN RANDOM PROCESSES*

O. KURBANMURADOV† AND K. SABELFELD‡

Abstract. Mean square convergence and convergence in probability of Fourier-wavelet models (FWM) of stationary Gaussian random processes in the metric of Banach space of continuously differentiable functions and in Sobolev space are studied. Sufficient conditions for the convergence formulated in the frame of spectral functions are given. It is shown that the given rates of convergence of FWM in the mean square obtained in the Nikolskiĭ–Besov classes cannot be improved.

Key words. Fourier-wavelet model, stationary Gaussian random process, Meyer's wavelets, Nikolskiĭ–Besov space, convergence in probability, convergence in mean square

AMS subject classifications. 65C05, 68U20, 62M40

DOI. 10.1137/070699408

1. Introduction.

Random processes and fields provide a convenient tool for a quantitative description of highly inhomogeneous and multiscale media such as, for instance, atmosphere, ocean, and soil (e.g., see [5], [17], [26], [29]).

The random fields are defined by all of their finite-dimensional distributions (e.g., see [12]). Fortunately, the most often used distributions in practice are Gaussian. The Gaussian random fields are defined uniquely by their expectations and covariances which greatly simplifies the measurement design. Along with the Gaussian random fields, one often exploits random fields which are obtained by some transformations of the Gaussian fields either through an explicit transform or implicitly through differential or integral equations where the Gaussian random fields are entered as random coefficients, random source terms, or boundary functions. As an important example we mention the hydraulic conductivity which is often modeled as a lognormal random field; e.g., see [5]. The Darcy flow itself is obtained as a random field which solves the Darcy equation with the lognormal hydraulic conductivity.

Another question about the structure of the random field which in practice is of extremely high importance is whether or not the random field is homogeneous. Unlike gaussianity, the most practically interesting random fields are nonhomogeneous. However, the homogeneous Gaussian random fields can be treated as perfect approximations in many cases, and they are indeed the most frequently used models in all branches of science and technology since the simulation technique based on the spectral representation is well developed.

Simulation of nonhomogeneous random fields is a much more difficult problem. The most completely developed are the Cholesky (singular) decomposition (e.g., see, [8], [32]) and Karhunen–Loeve expansion techniques (e.g., see, [40]). We also mention the wavelet decomposition method, which is applied to many interesting multiscale resolution problems; e.g., see [35], [41].

*Received by the editors August 8, 2007; accepted for publication (in revised form) May 1, 2008; published electronically September 4, 2008. This work was supported partly by the German DFG grant GZ: SA 861/5-1 and RFBR grant N 06-01-00498.

http://www.siam.org/journals/sinum/46-6/69940.html

†Center for Physical and Mathematical Research, Turkmenian State University, Turkmenbashy av. 31, 744000 Ashgabad, Turkmenistan (kurbanmuradov@wias-berlin.de).

‡Weierstrass Institute for Applied Analysis and Stochastics, Mohrenstrasse 39, D – 10117 Berlin, Germany, and Institute of Computational Mathematics and Mathematical Geophysics, Russian Academy of Sciences, Lavrentieva str., 6, 630090 Novosibirsk, Russia (sabelfel@wias-berlin.de).

There is a very extensive literature on simulation of homogeneous random fields, and even for a short discussion of recent publications in this field we would need to write a large review article. So we mention here only some publications, and the reader can search further leading by the references inside these publications: [7], [8], [10], [11], [21], [23], [25], [28], [29], [30], [31], [32], [33], [34].

It should be mentioned that there are also a lot of publications devoted to a comparative analysis of the cost of the simulation algorithms, but as we have concluded in [18], the most competitive simulation method for multiscale random fields are the randomized spectral methods, Fourier-wavelet methods, and the wavelet decomposition technique (e.g., see, [2], [10], [11], [22]).

Having an approximate model u_N for our random field $u(x)$ (say, the parameter N is integer), we have to estimate the error between the statistical characteristics of u_N and $u(x)$. The convergence $u_N \to u$ as $N \to \infty$ can be defined differently: (i) convergence of the relevant finite-difference distributions and (ii) convergence in a relevant functional space. The convergence in function spaces is important for practical applications, and they were considered in many publications (e.g., for the spectral models, see [1], [19], [20], [39]). In [3], [4], the convergence of wavelet decomposition models is studied. As to the Fourier-wavelet models (FWM), the first convergence results were obtained in [21], where the convergence of the correlation function of the model process FWM $u_N(x)$, $B_N(x,r) = E(u_N(x+r)u_N(x))$, to the correlation function $B(r) = E(u(x+r)u(x))$ of the original Gaussian process $u(x)$ has been established. In the language of the spectral function $F(k)$, we have obtained an estimation of $\sup_{x,r\in\mathbb{R}} |B(r) - B_N(x,r)|$ which shows that this quantity can be made arbitrarily small as $N \to \infty$.

In this paper we study the convergence of FWM $u_N \to u$ as $N \to \infty$ in Sobolev spaces $W_2^n[a,b]$ and in the space of n-times continuously differentiable functions $C^n[a,b]$, where n is a nonnegative integer. We study conditions on the spectral function F which are sufficient for $E||\mathcal{E}_N||^2_{W_2^n[a,b]} \to 0$ and $||\mathcal{E}_N||_{C^n[a,b]} \xrightarrow{P} 0$ as $N \to \infty$. Here $[a,b] \subset \mathbf{R}$ is an arbitrary fixed finite interval, and $\mathcal{E}_N(x) = u(x) - u_N(x)$. To make the presentation clear, we first consider the case of a scalar random process $u(x)$ and give sufficient conditions for the convergence $E||\mathcal{E}_N||^2_{L_2[a,b]} \to 0$ and $||\mathcal{E}_N||_{C[a,b]} \xrightarrow{P} 0$ as $N \to \infty$. Then we give a relevant generalization of the results to vector processes and extend the study for some stronger metrics.

The paper is organized as follows. The definition of the FWM for Gaussian random processes is given in section 2. Convergence of FWM in $L_2[0,1]$ and an analysis of the quality of the derived estimates are presented in section 3. Convergence of FWM in the space $C[0,1]$ is given in section 4. In section 5 we study the convergence FWM in some stronger metrics (in spaces $W_p^n[0,1]$ and $C^n[0,1]$) and for vector processes. Conclusions are made in section 6. Finally, in the Appendix we give the proofs of Lemmas 4.1 and 4.2 and present some known results from the theory of Besov's space that we used in our work in the Appendix.

2. Fourier-wavelet models of Gaussian random processes.

2.1. Fourier-wavelet expansions of Gaussian random processes. Let us introduce the notation $\mathbf{u}(x) = (u_1(x), \dots, u_l(x))^T$, $x \in \mathbf{R} = (-\infty, \infty)$, for an l-dimensional real-valued Gaussian stationary (homogeneous) random process with zero mean and a given correlation tensor $B(r)$:

$$B_{ij}(r) = E[u_i(x+r)\, u_j(x)], \quad i,j = 1, \dots, l, \tag{2.1}$$

or the corresponding spectral tensor F:

(2.2)
$$F_{ij}(k) = \int_{-\infty}^{\infty} e^{-\mathrm{i}\,2\pi\,k\cdot r} B_{ij}(r)\,dr, \quad B_{ij}(r) = \int_{-\infty}^{\infty} e^{\mathrm{i}\,2\pi\,r\cdot k} F_{ij}(k)\,dk, \quad i,j = 1,\dots,l.$$

Here and throughout the paper E stands for mathematical expectation. Furthermore, we will assume that $\int_{-\infty}^{\infty} \sum_{j=1}^{l} B_{jj}^2(r)\,dr < \infty$, which ensures the existence of the spectral tensor F_{ij} in the space $L_2(\mathbb{R})$ [26]. Moreover, $B_{ii}(0) = Eu_i^2(x) < \infty$ implies $B_{ii}(0) = \int_{-\infty}^{\infty} F_{ii}(k)\,dk < \infty$, that is, $F \in L_1(\mathbb{R})$. Hence $F \in L_1(\mathbb{R}) \cap L_2(\mathbb{R})$.

For some positive integer n let $Q(k)$ be an $l \times n$-matrix satisfying the condition

$$Q(k)Q^*(k) = F(k), \quad Q(-k) = \bar{Q}(k). \tag{2.3}$$

Here the star stands for the complex conjugate transpose which is equivalent to taking two operations: the transpose ^{T} and the complex conjugation of each entry.

Let $\phi(x)$ and $\psi(x)$, $x \in \mathbb{R}$, be orthonormal scaling and wavelet functions (for the definitions see, e.g., [10], [24], [13]), respectively, and $\hat{\phi}(k)$ and $\hat{\psi}(k)$ are Fourier transformations of these functions:

$$\hat{\phi}(k) = \int_{-\infty}^{\infty} e^{-\mathrm{i}\,2\pi\,kx}\phi(x)dx, \quad \hat{\psi}(k) = \int_{-\infty}^{\infty} e^{-\mathrm{i}\,2\pi\,kx}\psi(x)dx.$$

Assume that $\hat{\phi}(k)$ and $\hat{\psi}(k)$ are even functions

$$\hat{\phi}(-k) = \bar{\hat{\phi}}(k), \quad \hat{\psi}(-k) = \bar{\hat{\psi}}(k). \tag{2.4}$$

Then the Fourier-wavelet expansion of the random process $\mathbf{u}(x)$ reads [10], [21]:

$$\mathbf{u}(x) = \sum_{j=-\infty}^{\infty} \mathcal{F}_{m_0}^{(\phi)}(2^{m_0}x + j)\,\boldsymbol{\xi}_j + \sum_{m=m_0}^{\infty}\sum_{j=-\infty}^{\infty} \mathcal{F}_{m}^{(\psi)}(2^{m}x + j)\,\boldsymbol{\xi}_{mj}, \tag{2.5}$$

where m_0 is an arbitrary (but fixed) integer, $\boldsymbol{\xi}_j$, $\boldsymbol{\xi}_{mj}$ $(j \in \mathbb{Z},\, m \geq m_0)$ is a family of mutually independent standard real-valued Gaussian random vectors of dimension n, and $\mathcal{F}^{(\phi)}(\cdot)$, $\mathcal{F}^{(\psi)}(\cdot)$ are $l \times n$-dimensional matrix functions defined by

$$\begin{aligned} \mathcal{F}_{m}^{(\phi)}(y) &= \int_{-\infty}^{\infty} e^{-\mathrm{i}\,2\pi\,ky}\,2^{m/2}\bar{Q}(2^m k)\hat{\phi}(k)\,dk, \\ \mathcal{F}_{m}^{(\psi)}(y) &= \int_{-\infty}^{\infty} e^{-\mathrm{i}\,2\pi\,ky}\,2^{m/2}\bar{Q}(2^m k)\hat{\psi}(k)\,dk. \end{aligned} \tag{2.6}$$

Since $B(0) = \int F(k)dk < \infty$, and $QQ^* = F$, we conclude that all of the entries of the matrix Q belong to $L_2(\mathbb{R})$. Therefore, the functions given by (2.6) are well-defined.

Let us give a comment concerning some difference between the Fourier-wavelet and wavelet decompositions. In both methods, the random process is represented as a series of deterministic functions weighted by a random Gaussian coefficient. The crucial difference is that, unlike the wavelet decomposition, in the Fourier-wavelet method these random coefficients are independent, which might be quite convenient in practical simulations; see, for instance, [18]. The wavelet decomposition is more general and can be used also to simulate inhomogeneous random fields [41].

Further, in Fourier-wavelet expansion (2.5) we will use Meyer's wavelet functions $\phi(x)$ and $\psi(x)$, which are defined by their Fourier transforms (e.g., see [6]):

$$\phi(x) = \int_{-\infty}^{\infty} e^{\mathrm{i}\,2\pi\,kx}\hat{\phi}(k)\,dk, \quad \psi(x) = \int_{-\infty}^{\infty} e^{\mathrm{i}\,2\pi\,kx}\hat{\psi}(k)\,dk, \tag{2.7}$$

where

$$\hat{\phi}(k)=\begin{cases}1, & |k|\le 1/3,\\ \cos\theta_{1k}, & 1/3\le|k|\le 2/3,\\ 0, & \text{otherwise},\end{cases}\qquad \hat{\psi}(k)=\begin{cases}e^{-\mathrm{i}\,\pi\,k}\sin\theta_{1k}, & 1/3\le|k|\le 2/3,\\ e^{-\mathrm{i}\,\pi\,k}\cos\theta_{2k}, & 2/3x\le|k|\le 4/3,\\ 0, & \text{otherwise},\end{cases}$$

where $\theta_{1k}=\frac{\pi}{2}\nu(3|k|-1)$ and $\theta_{2k}=\frac{\pi}{2}\nu(\frac{3}{2}|k|-1)$. Here $\nu(x)$ is a smooth function satisfying the following conditions: $\nu(x)\equiv 0$ for $x\le 0$, $\nu(x)\equiv 1$ for $x\ge 1$, and $\nu(x)+\nu(1-x)=1$ for $0<x<1$. As an example of such a function, we mention a function $\nu(x)=\nu_p(x)$ depending on a positive parameter p (see [10]):

$$\nu_p(x)=\frac{4^{p-1}}{p}\left\{[x-x_0]_+^p+[x-x_p]_+^p+2\sum_{j=1}^{p-1}(-1)^j\,[x-x_j]_+^p\right\},$$

where $x_j=(1/2)[\cos(((p-j)/p)\pi)+1]$ and $[a]_+=\max(a,0)$. The function ν_p is $p-1$ times continuously differentiable; therefore, by choosing p sufficiently large, we can make the functions $\hat{\phi}$ and $\hat{\psi}$ smooth enough.

2.2. Fourier-wavelet models. In the numerical implementation of (2.5) we have to find a reasonable choice of the cutoff parameters m_1 and b_0, b_m ($m=m_0,\dots,m_1$) in the approximations:

$$\sum_{j=-\infty}^{\infty}\mathcal{F}_{m_0}^{(\phi)}(2^{m_0}x+j)\,\boldsymbol{\xi}_j\simeq\sum_{j=-b-\lfloor 2^{m_0}x\rfloor}^{b-\lfloor 2^{m_0}x\rfloor}\mathcal{F}_{m_0}^{(\phi)}(2^{m_0}x+j)\,\boldsymbol{\xi}_j, \tag{2.8}$$

$$\sum_{m=m_0}^{\infty}\sum_{j=-\infty}^{\infty}\mathcal{F}_{m}^{(\psi)}(2^{m}x+j)\,\boldsymbol{\xi}_{mj}\simeq\sum_{m=m_0}^{m_1}\sum_{j=-b_m-\lfloor 2^{m}x\rfloor}^{b_m-\lfloor 2^{m}x\rfloor}\mathcal{F}_{m}^{(\psi)}(2^{m}x+j)\,\boldsymbol{\xi}_{mj}, \tag{2.9}$$

where $\lfloor a\rfloor$ stands for the integer part of a. The general idea is that b and b_m should be chosen so that supports of the functions $\mathcal{F}_{m_0}^{(\phi)}$ and $\mathcal{F}_{m}^{(\psi)}$ belong essentially to the intervals $[-b,b]$ and $[-b_m,b_m]$, respectively.

In previous papers on Fourier-wavelet models [2], [10], [18], [21], the authors use approximations of type (2.8)–(2.9). Unfortunately, functions staying in the right-hand sides of (2.8)–(2.9) are discontinuous in x (with the jump discontinuity). Therefore, to ensure that the samples of the approximation process are continuous, we have to modify the model. So let us define the following modification of the FWM of the random field $\mathbf{u}(x)$. Let $\{\mu_N\}_{N=1}^{\infty}$ be a sequence of positive integers, $\mu_N\ge m_0$, $N=1,2,\dots$, and $\{b_N\}_{N=1}^{\infty}$, $\{b_{mN}\}_{N=1}^{\infty}$, $m=0,\dots,\mu_N$ be sequences of positive real numbers depending on a positive integer N so that

$$\lim_{N\to\infty}\mu_N=\infty,\quad \lim_{N\to\infty}b_N=\infty,\quad \lim_{N\to\infty}b_{mN}=\infty,\quad m=0,1,\dots,\mu_N. \tag{2.10}$$

For fixed positive number A denote by $\eta_A(\cdot)$ a cutoff function $\eta_A:[0,\infty)\to[0,1]$ defined by

$$\eta_A(x)=\begin{cases}0, & 0\le x\le A,\\ 1, & x\ge A+1,\\ x-A, & A<x<A+1.\end{cases} \tag{2.11}$$

Define functions $\chi_N : \mathbb{R} \to [0,1]$ and $\chi_{mN} : \mathbb{R} \to [0,1]$, $m = 0, 1, \dots, \mu_N$, by assuming that

$$\chi_N(x) = \eta_{b_N}(|x|), \qquad \chi_{mN}(x) = \eta_{b_{mN}}(|x|),$$
$$\chi'_N(x) = 1 - \chi_N(x), \quad \chi'_{mN}(x) = 1 - \chi_{mN}(x).$$

Under FWM of the random process $\mathbf{u}(x)$ we understand the following sequence of random functions:

$$\mathbf{u}_N(x) = \sum_{j\in\mathbb{Z}} \chi'_N(2^{m_0}x + j) \cdot \mathcal{F}^{(\phi)}_{m_0}(2^{m_0}x + j)\, \boldsymbol{\xi}_j + \sum_{m=m_0}^{\mu_N} \sum_{j\in\mathbb{Z}} \chi'_{mN}(2^m x + j) \cdot \mathcal{F}^{(\psi)}_m(2^m x + j)\, \boldsymbol{\xi}_{mj}, \quad N = 1, 2, \dots . \tag{2.12}$$

Our aim is to study the convergence $\mathbf{u}_N \to \mathbf{u}$ as $N \to \infty$ in functional spaces $W_2^n[a,b]$ and $C^n[a,b]$, where n is a nonnegative integer. Here we suggest sufficient conditions on the spectral tensor F which ensure $E||\mathcal{E}_N||^2_{W_2^n[a,b]} \to 0$ and $||\mathcal{E}_N||_{C^n[a,b]} \xrightarrow{P} 0$ as $N \to \infty$. Here $[a,b] \subset \mathbf{R}$ is an arbitrary fixed finite interval, and $\mathcal{E}_N(x) = u(x) - u_N(x)$. Further, not losing in generality, we can take $[a,b] = [0,1]$, and $m_0 = 0$. Moreover, for the simplicity of presentation we first consider the case of a scalar random process $u(x)$ (i.e., $l = 1$) and give sufficient conditions for the convergence $E||\mathcal{E}_N||^2_{L_2[a,b]} \to 0$ and $||\mathcal{E}_N||_{C[a,b]} \xrightarrow{P} 0$ as $N \to \infty$. In section 5 we give relevant generalizations of results for vector processes and stronger metrics.

3. Convergence of FWM in the L_2 metric. Let $u(x)$, $x \in \mathbb{R}$, be a real-valued scalar stationary zero mean Gaussian random process with a spectral function $F(k)$, and let $\phi(x)$ and $\psi(x)$ be orthonormal Meyer's scaling and wavelet functions, respectively. It is clear that their Fourier transforms $\hat{\phi}(k)$ and $\hat{\psi}(k)$ are even functions. In this case the FWM of $u(x)$ is (see (2.12) and recall that $m_0 = 0$):

$$u_N(x) = \sum_{j\in\mathbb{Z}} \chi'_N(x + j) \cdot \mathcal{F}^{(\phi)}_0(x + j)\, \xi_j + \sum_{m=0}^{\mu_N} \sum_{j\in\mathbb{Z}} \chi'_{mN}(2^m x + j) \cdot \mathcal{F}^{(\psi)}_m(2^m x + j)\, \xi_{mj}, \tag{3.1}$$

where ξ_j, ξ_{mj}, $m = 0, 1, \dots$; $j \in \mathbb{Z}$ is a set of mutually independent standard Gaussian random variables.

Let us introduce some notation: $\Delta = [-4/3,\, 4/3]$, $\Delta_m = [-4 \cdot 2^m/3,\, -2^m/3] \cup [2^m/3,\, 4 \cdot 2^m/3]$, $m = 0, 1, 2, \dots$, and $\Delta' = \cup_{m=0}^{\infty} \Delta_m$. For a function $f : \mathbb{R} \to \mathbb{R}$ and $D \subset \mathbb{R}$ we denote by $f|_D$ the restriction of the function f in D. For a measurable $D \subset \mathbb{R}$ we write $|D| = \int_D dx$.

In this paper we use some results from the theory of Nikolskiĭ–Besov spaces (e.g., see [27], [37]). For $r > 0$, a triple (r, j, l) is called admissible if $j \in \mathbb{N}$, $l \in \mathbb{N}_0$, and $j > r - l > 0$. Here $\mathbb{N}_0 = \{0, 1, 2, \dots\}$ and $\mathbb{N} = \{1, 2, \dots\}$. Let us denote by $\Delta_h^{(j)} g$ the jth difference of g:

$$\Delta_h g(\cdot) = g(\cdot + h) - g(\cdot), \dots, \Delta_h^{(j)} g(\cdot) = \Delta_h \Delta_h^{(j-1)} g(\cdot).$$

For $1 \le p, q \le \infty$, and $r > 0$, the Nikolskiĭ–Besov space $B^r_{pq}(\mathbb{R})$ is defined as the set of all functions $f \in L_p$ such that the norm

$$||f||_{B^r_{pq}} = ||f||_{L_p} + ||f||_{b^r_{pq}},$$

where

$$||f||_{b^r_{pq}} = \left(\int_{-1}^{1} \left(\frac{||\Delta_h^{(j)} f^{(l)}||_{L_p}}{|h|^{r-l}} \right)^q \frac{dh}{|h|} \right)^{\frac{1}{q}} \quad (1 \le q < \infty), \quad \text{and}$$

$$||f||_{b^r_{p\infty}} = \sup_{0<|h|\le 1} |h|^{\,l-r}\, ||\Delta_h^{(j)} f^{(l)}||_{L_p}$$

makes sense and is finite for some admissible triple (r, j, l). Here $f^{(l)}(x) = D^l f(x)$ is the lth derivative of the function f. The ambiguity in the choice of triple (r, j, l) is not essential: Different admissible triples correspond to equivalent norms. For a measurable subset $D \subset \mathbb{R}$, the space $B^r_{pq}(D)$ is defined as above but by changing $||\cdot||_{L_p}$ with $||\cdot||_{L_p(D_{jh})}$, where $D_h = \{x \in D : x + \lambda h \in D \text{ for all } \lambda \in [0,1]\}$. For a function $f : \mathbb{R} \to \mathbb{R}$ such that $f|_D \in B^r_{pq}(D)$ we will simply write $||f||_{B^r_{pq}(D)}$ instead of the more exact but complex notation $||(f|_D)||_{B^r_{pq}(D)}$.

We will exploit the following imbeddings (e.g., see [27], [37]):

$$B^r_{pq}(\mathbb{R}) \hookrightarrow B^{\rho}_{p_1 q}(\mathbb{R}), \quad \rho = r\left(1 - \frac{1}{r} \cdot \left[\frac{1}{p} - \frac{1}{p_1}\right]\right), \quad 1 \le p < p_1 < \infty, \tag{3.2}$$

$$B^{r+\varepsilon}_{p\infty}(\mathbb{R}) \hookrightarrow B^r_{pq}(\mathbb{R}) \hookrightarrow B^r_{pq_1}(\mathbb{R}), \quad 1 \le q < q_1 \le \infty, \quad \varepsilon > 0, \tag{3.3}$$

where $X \hookrightarrow Y$ means that $X \subset Y$ for seminormed spaces X and Y and there exists a constant $c > 0$ such that the inequality $||x||_Y \le c||x||_X$ is fulfilled.

3.1. Main results on convergence of FWM in the mean square. Let us formulate two assumptions on the functions $F(k)$ and $Q(k) = (F(k))^{1/2}$.

(H0): For some $s > 0$

$$I_s \stackrel{def}{=} \int_{-\infty}^{\infty} F(k)(1 + |k|^2)^s dk < \infty; \tag{3.4}$$

(H1): there exist positive constants c_0, ε, $\rho_0 > 1/2$, and $\rho_1 > 1/2$ and positive integer m_0 such that

$$\begin{aligned} &\text{(i) } Q|_\Delta \in B^{\rho_0}_{1\infty}(\Delta),\ Q|_{\Delta_0} \in B^{\rho_0}_{1\infty}(\Delta_0),\ \ldots,\ Q|_{\Delta_{m_0-1}} \in B^{\rho_0}_{1\infty}(\Delta_{m_0-1});\\ &\text{(ii) } Q|_{\Delta_m} \in B^{\rho_1}_{1\infty}(\Delta_m), \quad m \ge m_0;\\ &\text{(iii) } ||Q||_{b^{\rho_1}_{1\infty}(\Delta_m)} \le c_0 \cdot 2^{-m(\rho_1 - 1/2 + \varepsilon)}, \quad m \ge m_0. \end{aligned} \tag{3.5}$$

THEOREM 3.1. *Assume that the spectral function $F(k)$ satisfies the condition* (H0). *Let $\{\rho_m\}_{m=0}^{\mu_N}$ be a finite sequence of positive numbers such that* $\min_m \rho_m > 1/2$. *Suppose that $Q|_\Delta \in B^{\rho_0}_{1\infty}(\Delta)$ and $Q|_{\Delta_m} \in B^{\rho_m}_{1\infty}(\Delta_m)$ for $1 \le m \le \mu_N$, and assume that the Meyer wavelet functions ϕ and ψ belong to the class $C^{\nu+1}$ for $\nu =$* $\max\{\lfloor\rho\rfloor, \lfloor s\rfloor\}$, *where $\rho = \max_{0\le m\le\mu_N} \rho_m$. Then*

$$\begin{aligned} E(u(x) - u_N(x))^2 &\le \frac{C_1(s,\psi)}{4^{s(\mu_N+1)}} \cdot I_s + \frac{C_2(\rho_0,\phi)}{b_N^{2\rho_0-1}} ||Q||^2_{B^{\rho_0}_{1\infty}(\Delta)}\\ &+ \sum_{m=0}^{\mu_N} \frac{C_3(\rho_m,\psi)}{b_{mN}^{2\rho_m-1}} \cdot \left(2^{-m}||Q||^2_{L_1(\Delta_m)} + 2^{2m(\rho_m-1/2)}||Q||^2_{b^{\rho_m}_{1\infty}(\Delta_m)}\right), \end{aligned} \tag{3.6}$$

where constants C_i, $i = 1, 2, 3$, depend only on the shown arguments.

Proof. From the definition (3.1) we have

$$u(x) = u_N(x) + v^{(1)}(x) + \sum_{m=0}^{\mu_N} v_m^{(2)}(x) + v^{(3)}(x), \tag{3.7}$$

where

$$\begin{aligned} v^{(1)}(x) &= \sum_{j\in\mathbb{Z}} \chi_N(x+j)\cdot \mathcal{F}_0^{(\phi)}(x+j)\,\xi_j\,, \\ v_m^{(2)}(x) &= \sum_{j\in\mathbb{Z}} \chi_{mN}(2^m x+j)\cdot \mathcal{F}_m^{(\psi)}(2^m x+j)\,\xi_{mj}\,, \\ v^{(3)}(x) &= \sum_{m=\mu_N+1}^{\infty}\ \sum_{j=-\infty}^{\infty} \mathcal{F}_m^{(\psi)}(2^m x+j)\,\xi_{mj}\,. \end{aligned}$$

The random variables ξ_j and ξ_{mj} are mutually independent; hence the terms in the right-hand side of (3.7) are also independent, and therefore

$$E(u(x)-u_N(x))^2 = E(v^{(1)}(x))^2 + \sum_{m=0}^{\mu_N} E(v_m^{(2)}(x))^2 + E(v^{(3)}(x))^2. \tag{3.8}$$

Let us estimate the terms in the right-hand side. First, we have for the last term

$$E(v^{(3)}(x))^2 = \sum_{m=\mu_N+1}^{\infty}\ \sum_{j=-\infty}^{\infty} |\mathcal{F}_m^{(\psi)}(2^m x+j)|^2. \tag{3.9}$$

Further, from the definition (2.6) of the function $\mathcal{F}$, and since the functions $Q(k)$ and $\psi(k)$ are even, we have

$$\mathcal{F}_m^{(\psi)}(y) = \int_{-\infty}^{\infty} e^{-\mathrm{i}2\pi k y} 2^{m/2}\bar{Q}(2^m k)\hat{\psi}(k)dk = \int_{-\infty}^{\infty} e^{\mathrm{i}2\pi k y} 2^{m/2} Q(2^m k)\bar{\hat{\psi}}(k)dk.$$

Therefore,

$$\begin{aligned} \mathcal{F}_m^{(\psi)}(2^m x+j) &= \int_{-\infty}^{\infty} e^{\mathrm{i}2\pi k(2^m x+j)} 2^{m/2} Q(2^m k)\bar{\hat{\psi}}(k)dk \\ &= \int_{-\infty}^{\infty} e^{\mathrm{i}2\pi k x} Q(k)\bar{\hat{\psi}}_{mj}(k)\,dk \\ &= \int_{-\infty}^{\infty} G(x+y)\psi_{mj}(y)dy, \end{aligned} \tag{3.10}$$

where $G(x) = \int_{-\infty}^{\infty} e^{\mathrm{i}2\pi k x} Q(k)\,dk$, and

$$\psi_{mj}(x) = 2^{m/2}\psi(2^m x - j), \quad \hat{\psi}_{mj}(k) = 2^{-m/2}\, e^{-\mathrm{i}\,2\pi\, k\, j\, 2^{-m}}\, \hat{\psi}(2^{-m} k).$$

Consequently for fixed x, the quantities $\mathcal{F}_m^{(\psi)}(2^m x+j)$ $(m,j\in\mathbb{Z})$ are Fourier coefficients in the expansion of $G(x+\cdot)$ with respect to the orthonormal system ψ_{mj} $(m,j\in\mathbb{Z})$.

Notice that the condition (3.4) is equivalent to the fact that the function G belongs to the Sobolev space $H_2^s(\mathbb{R})$. This is because $G\in H_2^s(\mathbb{R})$ means that the function

$Q(k)(1+|k|^2)^{s/2}$ is in $L_2(\mathbb{R})$ and $||G||_{H_2^s} = I_s^{1/2}$ (more details in [37]). Furthermore, we use the fact that $H_2^s(\mathbb{R})$ coincides with the Besov space $B_{22}^s(\mathbb{R})$, and the norms in these spaces are equivalent (see [37, section 2.3.9]). Since $G \in B_{22}^s$, we conclude by virtue of Corollary A.3 that

$$\sum_{j=-\infty}^{\infty} |\mathcal{F}_m^{(\psi)}(2^m x + j)|^2 \le B_s^2\, 4^{-ms} \cdot ||G(x+\cdot)||_{H_2^s}^2 \tag{3.11}$$

for some $B_s = B_s(\psi)$ depending only on s and ψ. From this we get by $||G(x+\cdot)||_{H_2^s}^2 = ||G||_{H_2^s}^2 = I_s$ that

$$E(v^{(3)}(x))^2 = \sum_{m=\mu_N+1}^{\infty} \sum_{j=-\infty}^{\infty} |\mathcal{F}_m^{(\psi)}(2^m x + j)|^2 \le \frac{B_s^2}{4^{(\mu_N+1)s}}\, I_s. \tag{3.12}$$

Now we turn to the estimation of the first two terms of the right-hand side (r.h.s.) of (3.8). It is obvious, due to the independence of random variables $\xi_j,\ j \in \mathbb{Z}$, that

$$E(v^{(1)}(x))^2 = \sum_{j\in\mathbb{Z}} |\chi_N(x+j)|^2 \cdot |\mathcal{F}_0^{(\phi)}(x+j)|^2 \le \sum_{j\in\mathbb{Z}:\, |x+j|\ge b_N} |\mathcal{F}_0^{(\phi)}(x+j)|^2. \tag{3.13}$$

We first estimate each term of this sum. The function $\mathcal{F}_0^{(\phi)}(y)$ has the Fourier transform $\hat{\mathcal{F}}(k) = Q(-k)\hat{\phi}(-k) = Q(k)\hat{\phi}(-k)$ (see (2.6)). Due to the conditions $\phi \in C^{\nu+1}$ and $Q|_\Delta \in B_{1\infty}^{\rho_0}(\Delta)$ it follows from Corollary A.2 that $||\hat{\mathcal{F}}||_{B_{1\infty}^{\rho_0}(\mathbb{R})} \le C(\rho_0,\phi)||Q||_{B_{1\infty}^{\rho_0}(\Delta)}$. Then from Lemma A.1 we get

$$|\mathcal{F}_0^{(\phi)}(y)| \le \frac{C(\rho_0,\phi)}{|y|^{\rho_0}} \cdot ||\hat{\mathcal{F}}||_{B_{1\infty}^{\rho_0}(\mathbb{R})} \le \frac{C(\rho_0,\phi)}{|y|^{\rho_0}} \cdot ||Q||_{B_{1\infty}^{\rho_0}(\Delta)}. \tag{3.14}$$

Note that from (3.13) and (3.14) we get

$$E(v^{(1)}(x))^2 \le \frac{C(\rho_0,\phi)}{b_N^{2\rho_0-1}} \cdot ||Q||_{B_{1\infty}^{\rho_0}(\Delta)}^2. \tag{3.15}$$

Let us turn to estimating the terms $E(v_m^{(2)}(x))^2$, $m = 0,\dots,\mu_N$. Denote by Q_m the function $Q_m : \Delta_0 \to \mathbb{R}$ defined as $Q_m(k) = Q(2^m k)$. Then by the same arguments as we used in the derivation of estimate (3.15), we obtain

$$\begin{aligned} E(v_m^{(2)}(x))^2 &\le \sum_{j\in\mathbb{Z}:\, |2^m x+j|\ge b_{mN}} |\mathcal{F}_m^{(\psi)}(2^m x + j)|^2 \le \frac{C(\rho_m,\psi)}{b_{mN}^{2\rho_m-1}} \cdot 2^m ||Q_m||_{B_{1\infty}^{\rho_m}(\Delta_0)}^2 \\ &\le \frac{C(\rho_m,\psi)}{b_{mN}^{2\rho_m-1}} \cdot 2^m \cdot \left(2||Q_m||_{L_1(\Delta_0)}^2 + ||Q_m||_{b_{1\infty}^{\rho_m}(\Delta_0)}^2\right). \end{aligned} \tag{3.16}$$

Put $l = \lfloor \rho_m \rfloor - 1$. Then

$$||Q_m||_{b_{1\infty}^{\rho_m}(\Delta_0)} = \sup_{|h|\le 1} \frac{1}{|h|^{\rho_m - l}} \int_{\Delta_{0,2h}} |\Delta_h^{(2)} Q_m^{(l)}(k)|\, dk. \tag{3.17}$$

Define a function $g_h : \Delta_m \to \mathbb{R}$ by $g_h(k) = \Delta_h^{(2)} Q^{(l)}(k)$. Then $\Delta_h^{(2)} Q_m^{(l)}(k) = 2^{ml} g_{2^m h}(2^m k)$. Therefore,

$$\begin{aligned} ||Q_m||_{b_{1\infty}^{\rho_m}(\Delta_0)} &= \sup_{|h|\le 1} \frac{1}{|h|^{\rho_m - l}} \int_{\Delta_{0,2h}} 2^{ml} |g_{2^m h}(2^m k)|\, dk \\ &= \sup_{|h|\le 1} \frac{2^{-m}}{|h|^{\rho_m - l}} \int_{\Delta_{m,2^{m+1}h}} 2^{ml} |g_{2^m h}(k)|\, dk\,. \end{aligned} \tag{3.18}$$

By the definition we have

$$\int_{\Delta_{m,2h}} |g_h(k)|\,dk = \int_{\Delta_{m,2h}} |\Delta_h^{(2)} Q^{(l)}(k)| \le ||Q||_{b_{1\infty}^{\rho_m}(\Delta_m)} |h|^{\rho_m - l}, \quad h \in \mathbb{R}. \tag{3.19}$$

Therefore it follows from (3.18) and (3.19) that

$$||Q_m||_{b_{1\infty}^{\rho_m}(\Delta_0)} \le 2^{m(\rho_m-1)} ||Q||_{b_{1\infty}^{\rho_m}(\Delta_m)}. \tag{3.20}$$

Now by taking into account that $||Q_m||_{L_1(\Delta_0)} \le 2^{-m} \cdot ||Q||_{L_1(\Delta_m)}$ and using the estimates (3.16) and (3.20), we obtain

$$E(v_m^{(2)}(x))^2 \le \frac{2C(\rho_m,\psi)}{b_{mN}^{2\rho_m-1}} \cdot \left(2^{-m}||Q||_{L_1(\Delta_m)}^2 + 2^{m(2\rho_m-1)}||Q||_{b_{1\infty}^{\rho_m}(\Delta_m)}^2\right) \tag{3.21}$$

for $m = 0, 1, \ldots, \mu_N$. This completes the proof of Theorem 3.1. □

PROPOSITION 3.1. *Assume that*

$$\lim_{N\to\infty} \mu_N = \infty, \quad \lim_{N\to\infty} b_N = \infty, \quad \lim_{N\to\infty} \min_{0\le m\le \mu_N} \{b_{mN}\} = \infty. \tag{3.22}$$

Then under the assumptions (H0)–(H1)*,*

$$\sup_{x\in[0,1]} E\mathcal{E}_N^2(x) \to 0 \quad as \quad N \to \infty. \tag{3.23}$$

Proof. Indeed, as a simple consequence of Theorem 3.1 we have

$$\begin{aligned}\sup_{x\in[0,1]} E\mathcal{E}_N^2(x) \le{}& \frac{C_1(s,\psi)}{4^{s(\mu_N+1)}} \cdot I_s + \frac{C_2(\rho_0,\phi)}{b_N^{2\rho_0-1}} ||Q||_{B_{1\infty}^{\rho_0}(\Delta)}^2 \\ &+ \sum_{m=0}^{m_0-1} \frac{C_3(\rho_0,\psi)}{b_{mN}^{2\rho_0-1}} \cdot \left(2^{-m}||Q||_{L_1(\Delta_m)}^2 + 2^{2m(\rho_0-1/2)}||Q||_{b_{1\infty}^{\rho_0}(\Delta_m)}^2\right) \\ &+ \sum_{m=m_0}^{\mu_N} \frac{C_3(\rho_1,\psi)}{b_{mN}^{2\rho_1-1}} \cdot \left(2^{-m}||Q||_{L_1(\Delta_m)}^2 + 2^{2m(\rho_1-1/2)}||Q||_{b_{1\infty}^{\rho_1}(\Delta_m)}^2\right). \end{aligned} \tag{3.24}$$

To show (3.23), it is sufficient to check that the last sum in the r.h.s. of this inequality converges to zero as $N \to \infty$. For this sake, for given $\delta > 0$, let us choose $N_0 = N_0(\delta)$ so that

$$\min_{0\le m\le \mu_N} \frac{1}{b_{mN}^{2\rho_1-1}} \le \delta, \quad N \ge N_0.$$

Denote by S_N the last sum in the r.h.s. of (3.24). Then due to the condition (3.5) we get for $N \ge N_0$

$$S_N \le C_3(\rho_1,\psi) \cdot \delta \cdot \left(\sum_{m=0}^{\infty} 2^{-m}||Q||_{L_1(\Delta_m)}^2 + c_0 \cdot \sum_{m=0}^{\infty} 2^{-2\varepsilon m}\right).$$

From

$$||Q||_{L_1(\Delta_m)}^2 \le |\Delta_m| \cdot \int_{\Delta_m} F(k)\,dk$$

and due to the assumption (H0), we get

$$||Q||^2_{L_1(\Delta_m)} \le |\Delta_m|(3\cdot 2^{-m})^{2s}\cdot I_s = 2\cdot 2^m(3\cdot 2^{-m})^{2s}\cdot I_s.$$

This implies that $S_N \le C\cdot\delta$ for $N \ge N_0(\delta)$ with a constant C which is not dependent on N. Since the parameter δ can be chosen arbitrarily small, $S_N \to 0$ as $N \to \infty$. The proof of Proposition 3.1 is complete. □

A straightforward consequence of Theorem 3.1 is the following assertion on the rate of the mean square convergence of FWM (3.1).

COROLLARY 3.1. *Assume that the hypotheses* (H0)–(H1) *are valid. Choose in the model* (3.1)

$$\mu_N = \left\lfloor \frac{(2\rho_0-1)}{2s}\cdot\log_2 N\right\rfloor + 1, \quad b_N = N, \quad b_{mN} = N,\ 0 \le m \le m_0 - 1;$$

$$b_{mN} = N^{\frac{2\rho_0-1}{2\rho_1-1}},\ m_0 \le m \le \mu_N. \tag{3.25}$$

Assume that $\phi,\ \psi \in C^{\nu+1}$, *where* $\nu = \max\{\lfloor s\rfloor,\ \lfloor\rho_0\rfloor,\ \lfloor\rho_1\rfloor\}$. *Then for each positive integer* N *the following estimation is valid:*

$$\sup_{x\in[0,1]} E\mathcal{E}_N^2(x) \le C\cdot N^{-(2\rho_0-1)}, \tag{3.26}$$

where C *is a constant not depending on* N.

3.2. Sufficient conditions for validity of (3.5). From the results given above it follows that the condition (3.5) is crucial in the analysis of the mean square convergence of FWM.

PROPOSITION 3.2. *For a nonnegative* k_0 *and a nonnegative integer* l, *assume that* $Q \in W_1^l(k_0,\infty)$ *and*

$$|\Delta_h^{(2)}Q^{(l)}(k)| \le \frac{C|h|^\gamma}{k^{l+1/2+\gamma+\varepsilon}} \quad \forall h\in[-1,1] \quad \text{and} \quad \forall k \ge k_0 \tag{3.27}$$

for some positive numbers C, ε *and* $\gamma \in (0,2)$. *Then the following estimate is true:*

$$||Q||_{b^{\rho_1}_{1\infty}(\Delta_m)} \le c_0\cdot 2^{-m(\rho_1-1/2+\varepsilon)} \quad \forall\ m \ge \log_2(3k_0), \tag{3.28}$$

where $\rho_1 = l+\gamma$ *and* $c_0 = c_0(\rho_1,\varepsilon)$.

Proof. Since the triple $(l+\gamma, 2, l)$ is admissible,

$$||Q||_{b^{\rho_1}_{1\infty}(\Delta_m)} \le C(\rho_1)\sup_{|h|\le 1}\frac{1}{|h|^\gamma}\int_{\Delta_m}|\Delta_h^{(2)}Q^{(l)}(k)|\,dk.$$

Therefore, for $2^m > 3k_0$ we have

$$||Q||_{b^{\rho_1}_{1\infty}(\Delta_m)} \le C(\rho_1)\sup_{|h|\le 1}\frac{1}{|h|^\gamma}\int_{\Delta_m}\frac{C|h|^\gamma}{k^{l+1/2+\gamma+\varepsilon}}\,dk \le c_0(\rho_1,\varepsilon)\cdot 2^{-m(\rho_1-1/2+\varepsilon)}. \qquad \square$$

COROLLARY 3.2. *Assume that for a nonnegative* k_0 *and positive integer* n, $Q \in W_1^n(k_0,\infty)$ *and*

$$|Q^{(n)}(k)| \le \frac{C}{k^{n+1/2+\varepsilon}} \quad \forall k \ge k_0 \tag{3.29}$$

for some positive numbers C and ε. Then the following estimate is valid:

$$||Q||_{b^n_{1\infty}(\Delta_m)} \le c_0 \cdot 2^{-m(n-1/2+\varepsilon)} \quad \forall\ m \ge \log_2(3(k_0+2)) \tag{3.30}$$

for some positive $c_0 = c_0(n,\varepsilon)$.

Proof. Indeed,

$$\begin{aligned}|\Delta_h^{(2)} Q^{(n-1)}(k)| &\le |\Delta_h Q^{(n-1)}(k+h)| + |\Delta_h Q^{(n-1)}(k+h)| \\ &\le h \cdot |Q^{(n)}(k+\beta h)| + h \cdot |Q^{(n)}(k+\alpha h)|\end{aligned}$$

for some $\alpha \in [0,1]$ and $\beta \in [1,2]$. Hence, by virtue of (3.29)

$$|\Delta_h^{(2)} Q^{(n-1)}(k)| \le \frac{C_1 h}{k^{n+1/2+\varepsilon}} \quad \forall k \ge k_0 + 2\,.$$

Therefore for $l = n-1$ and $\gamma = 1$ all of the conditions of Proposition 3.2 are fulfilled. □

Remark 3.1. Conditions (3.27) and (3.29) can be replaced with the following weaker conditions:

$$\int_{k_0}^{\infty} k^{l+\gamma-1/2+\varepsilon} \cdot \frac{|\Delta_h^{(2)} Q^{(l)}(k)|}{|h|^\gamma} \cdot dk < \infty \text{ and } \int_{k_0}^{\infty} k^{n-1/2+\varepsilon} |Q^{(n)}(k)|\, dk < \infty,$$

respectively.

3.3. Cost estimations for FWM. Now let us discuss the efficiency of FWM. Let us denote by T_N the number of arithmetic operations for calculation of the value of FWM (3.1) in one point $x \in \mathbb{R}$. It is obvious that $T_N \sim b_N + \sum_{m=0}^{\mu_N} b_{mN}$. Here $T_N \sim a_N$ means that there exist constants $0 < C_1 \le C_2$ not depending on N such that $C_1 a_N \le T_N \le C_2 \cdot a_N$. Therefore, if we choose the parameters of the FWM in accordance with (3.25), then under the conditions of Corollary 3.1 it follows that

$$\begin{aligned} T_N &\sim N \quad \text{if} \quad \rho_1 > \rho_0; \\ T_N &\sim N \ln N \quad \text{if} \quad \rho_1 = \rho_0; \\ T_N &\sim N^{(2\rho_0-1)/(2\rho_1-1)} \ln N \quad \text{if} \quad \rho_1 < \rho_0. \end{aligned} \tag{3.31}$$

Let us denote

$$\varepsilon_N = \left(\int_0^1 E(u(x) - u_N(x))^2 dx \right)^{1/2} = \left(E||\mathcal{E}_N||^2_{L_2[0,1]} \right)^{1/2}$$

the root mean square discrepancy of FWM (3.1) in the $L_2[0,1]$ metric. Then under the conditions of Corollary 3.1 the following estimate holds:

$$\varepsilon_N \le C \cdot N^{-(\rho_0 - 1/2)}, \qquad N = 1, 2, \ldots, \tag{3.32}$$

where C is some positive constant not depending on N. Hence under the conditions of Corollary 3.1 the number of operations T_ε required to achieve a given value of the r.m.s. error $\varepsilon_N = \varepsilon$ satisfies the estimates

$$\begin{aligned} T_\varepsilon &\lesssim \varepsilon^{-1/(\rho_0-1/2)} \quad \text{if} \quad \rho_1 > \rho_0; \\ T_\varepsilon &\lesssim \varepsilon^{-1/(\rho_0-1/2)} \ln \varepsilon \quad \text{if} \quad \rho_1 = \rho_0; \\ T_\varepsilon &\lesssim \varepsilon^{-1/(\rho_1-1/2)} \ln \varepsilon \quad \text{if} \quad \rho_1 < \rho_0. \end{aligned} \tag{3.33}$$

Here $T_\varepsilon \lesssim a(\varepsilon)$ means that there exists a constant C not depending on ε such that $T_\varepsilon \le C \cdot a(\varepsilon)$.

3.4. Analysis of the estimates (3.15)–(3.16). In this section we show that the exponents $2\rho_0 - 1$ and $2\rho_m - 1$ appearing in the estimates (3.15)–(3.16) cannot be improved in the Nikolskiĭ–Besov spaces $B^{\rho_0}_{1\infty}$ and $B^{\rho_m}_{1\infty}$, respectively. This will be done by the construction of a function Q which makes it possible to write $E(v^{(1)}(x))^2 \sim b_N^{-(2\rho_0-1)}$ and $E(v^{(2)}(x))^2 \sim b_{mN}^{-(2\rho_m-1)}$.

Let us consider the following example.

Example 3.1. Let $k_0 \geq 0$ and $\rho > 0$. Define

$$Q_0(k) = \begin{cases} 0, & 0 \leq k \leq k_0, \\ (k-k_0)^{\rho-1}, & k > k_0, \end{cases} \tag{3.34}$$

$$Q_0(k) = Q_0(-k) \quad \text{if} \quad k < 0.$$

Let $q : \mathbb{R} \to [0, \infty)$ be an arbitrary $(l+1)$-times differentiable even function with compact support satisfying the condition $q(k_0) \neq 0$. Within this section we denote $l = \lfloor \rho \rfloor$. Define $Q(k) = q(k)Q_0(k)$. From the definitions it follows that $Q \in B^{\rho}_{1\infty}$.

Now let us consider the asymptotic behavior of functions $\mathcal{F}_0^{(\phi)}(y)$, $\mathcal{F}_m^{(\psi)}(y)$ as $y \to \infty$. We will separately consider two cases: (i) ρ is noninteger and (ii) ρ is integer.

3.4.1. Asymptotics of $\mathcal{F}_0^{(\phi)}$ and $\mathcal{F}_m^{(\psi)}$ for noninteger ρ. Let us first assume that $k_0 < 2/3$ and $\hat{\phi} \in C^l$ and consider the function $\mathcal{F}_0^{(\phi)}(y)$. If we denote $g = q\hat{\phi}$, then

$$\begin{aligned} \mathcal{F}_0^{(\phi)}(y) &= \int_{-\infty}^{\infty} e^{-\mathrm{i}\,2\pi\,ky} \cdot Q(k)\hat{\phi}(k)\,dk = \int_{-\infty}^{\infty} e^{-\mathrm{i}\,2\pi\,ky} \cdot Q_0(k)g(k)\,dk \\ &= \frac{1}{(-\mathrm{i}2\pi y)^l} \int_{-\infty}^{\infty} e^{-\mathrm{i}\,2\pi\,ky} \cdot (Q_0 g)^{(l)}(k)\,dk \\ &= \frac{1}{(-\mathrm{i}2\pi y)^l} \int_{-\infty}^{\infty} e^{-\mathrm{i}\,2\pi\,ky} \cdot [Q_0^{(l)}(k)g(k) + \mathcal{R}(k)]\,dk, \end{aligned}$$

where $\mathcal{R}(k) = \sum_{j=0}^{l-1} C_l^j Q_0^{(j)}(k) g^{(l-j)}(k)$. It is obvious that $\mathcal{R}$ is a function from $B^1_{1\infty}$. Therefore, due to Lemma A.1 the integral

$$R(y) = \int_{-\infty}^{\infty} e^{-\mathrm{i}\,2\pi\,ky} \cdot \mathcal{R}(k)\,dk$$

satisfies the following estimate: $|R(y)| \leq \frac{A_1}{1+|y|}, y \in \mathbb{R}$.

Now let us consider the integral

$$\begin{aligned} F_0(y) &= \int_{-\infty}^{\infty} e^{-\mathrm{i}\,2\pi\,ky} \cdot Q_0^{(l)}(k)g(k)\,dk \\ &= \begin{cases} 2 \cdot \int_0^{\infty} \cos(2\pi\,ky) \cdot Q_0^{(l)}(k) \cdot g(k)\,dk, & l \text{ even}, \\ -2\mathrm{i} \cdot \int_0^{\infty} \sin(2\pi\,ky) \cdot Q_0^{(l)}(k) \cdot g(k)\,dk, & l \text{ odd}. \end{cases} \end{aligned}$$

Let us assume that l is even. Then

$$F_0(y) = 2C_1 \cos(2\pi k_0 y) F_c(2\pi y) - 2C_1 \sin(2\pi k_0 y) F_s(2\pi y),$$

where $C_1 = C_1(\rho) = (l-\alpha) \cdot (l-1-\alpha) \cdot \cdots \cdot (1-\alpha)$, $\alpha = 1 - \rho + l$, and

$$F_c(x) = \int_0^{\infty} \cos(kx) \cdot f(k)\,dk, \quad F_s(x) = \int_0^{\infty} \sin(kx) \cdot f(k)\,dk \tag{3.35}$$

are the respective Fourier cosine and sine transformations of the function $f(k) = g(k+k_0)k^{-\alpha}$. Now we will use the following assertion.

THEOREM 3.2 (see [36, Theorem 126]). *Let $f(k) = k^{-\alpha}\varphi(k)$, where $0 < \alpha < 1$ and $\varphi(k)$ is of bounded variation in $(0,\infty)$. Let F_c and F_s be Fourier sine and cosine transformations of f, respectively. Then*

$$F_c(x) = \varphi(+0)\cdot\left(\frac{2}{\pi}\right)^{1/2}\Gamma(1-\alpha)\sin(\pi\alpha/2)\cdot x^{\alpha-1}\cdot(1+o(1)) \quad as \quad x\to\infty$$

and

$$F_c(x) = \varphi(\infty)\cdot\left(\frac{2}{\pi}\right)^{1/2}\Gamma(1-\alpha)\sin(\pi\alpha/2)\cdot x^{\alpha-1}\cdot(1+o(1)) \quad as \quad x\to 0.$$

$F_s(x)$ satisfies similar conditions with $\sin(\pi\alpha/2)$ replaced by $\cos(\pi\alpha/2)$.

For even l it follows from this theorem that

$$F_0(y) = C_2(\rho)g(k_0)\cdot\frac{\sin(\pi\alpha/2 - 2\pi k_0 y)}{y^{1-\alpha}}\cdot(1+o(1)) \quad \text{as} \quad y\to\infty. \tag{3.36}$$

For odd l one can obtain a similar asymptotic:

$$F_0(y) = -\mathrm{i}\cdot C_2(\rho)g(k_0)\cdot\frac{\cos(\pi\alpha/2 + 2\pi k_0 y)}{y^{1-\alpha}}\cdot(1+o(1)) \quad \text{as} \quad y\to\infty. \tag{3.37}$$

Thus we have obtained the following asymptotic result for $\mathcal{F}_0^{(\phi)}$:

$$\mathcal{F}_0^{(\phi)}(y) = C_3(\rho)\cdot\Phi(k_0 y)\cdot\frac{q(k_0)\cdot\hat{\phi}(k_0)}{y^{\rho}}\cdot(1+o(1)) \quad \text{as} \quad y\to\infty, \tag{3.38}$$

where $\Phi(x) = \sin(\pi\alpha/2 - 2\pi x)$ if $l = \lfloor\rho\rfloor$ is even and $\Phi(x) = \cos(\pi\alpha/2 + 2\pi x)$ if l is odd.

Recall that we have assumed that $k_0 \in [0, 2/3]$. If $k_0 \geq 2/3$, then $\mathcal{F}_0^{(\phi)} \equiv 0$. By the same arguments it can be shown that for m such that $2^m < 3k_0 < 2^{m+2}$ the following asymptotic is valid:

$$\mathcal{F}_m^{(\psi)}(y) = C_3(\rho)\cdot 2^{m(\rho-1/2)}\Phi(2^{-m}k_0 y)\cdot\frac{q(2^{-m}k_0)\cdot\hat{\psi}(2^{-m}k_0)}{y^{\rho}}\cdot(1+o(1)) \tag{3.39}$$

as $y\to\infty$, provided $q(2^{-m}k_0)\neq 0$.

3.4.2. Asymptotics of $\mathcal{F}_0^{(\phi)}$ and $\mathcal{F}_m^{(\psi)}$ for integer ρ. Let $l = \rho = 1$. Then

$$\begin{aligned}\mathcal{F}_0^{(\phi)}(y) &= 2\int_{k_0}^{\infty}\cos(2\pi k y)g(k)\,dk = \frac{1}{\pi y}\int_{k_0}^{\infty} g(y)\cdot d(\sin(2\pi k y))\\ &= -\frac{g(k_0)}{\pi y}\cdot(1+O(1/|y|)) \quad \text{as} \quad y\to\infty.\end{aligned} \tag{3.40}$$

If $l \geq 2$, then

$$\mathcal{F}_0^{(\phi)}(y) = \frac{1}{(-\mathrm{i}2\pi y)^{l-1}}\int_{-\infty}^{\infty} e^{-\mathrm{i}\,2\pi\,k y}\cdot[Q_0^{(l-1)}(k)g(k) + \mathcal{R}(k)]\,dk,$$

where $\mathcal{R} \in B^2_{1\infty}$. Therefore

$$\mathcal{F}_0^{(\phi)}(y) = \frac{1}{(-\mathrm{i}2\pi y)^{l-1}} \cdot (F_0(y) + O(1/|y|^2) \quad \text{as} \quad y \to \infty,$$

where

$$F_0(y) = \int_{-\infty}^{\infty} e^{-\mathrm{i}\,2\pi\,ky} \cdot Q_0^{(l-1)}(k) g(k)\,dk = \begin{cases} 2 \cdot \int_{k_0}^{\infty} \cos(2\pi\,ky) \cdot g(k)\,dk, & l \text{ odd}, \\ -2\mathrm{i} \cdot \int_{k_0}^{\infty} \sin(2\pi\,ky) \cdot g(k)\,dk, & l \text{ even}. \end{cases}$$

By using integration by parts and Lemma A.1 we can easily show that

$$\mathcal{F}_0^{(\phi)}(y) = C(\rho) \cdot \Phi(k_0 y) \cdot \frac{q(k_0) \cdot \hat{\phi}(k_0)}{y^l} \cdot (1 + o(1)) \quad \text{as} \quad y \to \infty, \tag{3.41}$$

where $\Phi(x) = \cos(2\pi x)$ if $l = \rho$ is even and $\Phi(x) = \sin(2\pi x)$ if l is odd.

By the same arguments we can establish that (3.39) is valid for integer ρ with $\Phi(x) = \cos(2\pi x)$ if $l = \rho$ is even and $\Phi(x) = \sin(2\pi x)$ if l is odd.

3.4.3. Lower bounds for the estimates (3.15)–(3.16). Now let us come back to the estimates (3.15)–(3.16). Assume that ρ is not integer and $l = \lfloor \rho \rfloor$ is even. Let $k_0 = 1/2$. Then by virtue of (3.38) we get for some positive integer N_0 and for each $N \geq N_0$

$$\begin{aligned} E(v^{(1)}(x))^2 &\geq \sum_{j \in \mathbb{Z}:\, |x+j| \geq b_N + 1} |\mathcal{F}_0^{(\phi)}(x+j)|^2 \\ &\geq C(\rho) \cdot \sum_{j \in \mathbb{Z}:\, |x+j| \geq b_N + 1} \sin^2(\pi\alpha/2 - \pi \cdot (x+j)) \cdot \frac{1}{|x+j|^{2\rho}} \\ &\geq C_1(\rho) \sin^2(\pi\alpha/2 - \pi x) \frac{1}{b_N^{2\rho-1}}, \end{aligned}$$

where $C_1(\rho) > 0$; in the same way the estimate can be established for odd l and noninteger ρ. For integer ρ, one can use (3.41) and establish the analogous result.

If, for nonnegative m_0, one puts $k_0 = 2^{m_0 - 1}$, then by (3.39)

$$E(v_{m_0}^{(2)}(x))^2 \geq C(\rho) \sin^2(\pi\alpha/2 - \pi 2^{m_0} x) 2^{2m_0(\rho - 1/2)} \sum_{j \in \mathbb{Z}:\, |2^{m_0} x + j| \geq b_{m_0 N} + 1} \frac{1}{|2^{m_0} x + j|^{2\rho}}$$

$$\geq C_1(\rho) \sin^2(\pi\alpha/2 - \pi 2^{m_0} x) 2^{2m_0(\rho - 1/2)} \frac{1}{b_{m_0 N}^{2\rho-1}}, \quad N \geq N_0, \tag{3.42}$$

for noninteger ρ with even $l = \lfloor \rho \rfloor$. For other values of ρ ($\lfloor \rho \rfloor$ is odd or ρ is integer) similar estimates can be obtained in an analogous way.

Hence in Nikolskiĭ–Besov classes $B^{\rho_m}_{1\infty}$, $m = 0, 1, \ldots, \mu_N$, the exponents $2\rho_m - 1$, $m = 0, \ldots, \mu_N$, on the r.h.s. of the estimate (3.6) are best possible values. In this connection it should be pointed out that, for Sobolev spaces $W_1^{\rho_m}$, this is not the case. More precisely, if for positive integers $l_0, \ldots, l_{\mu_N}$ we replace the conditions $Q|_\Delta \in B^{\rho_0}_{1\infty}(\Delta)$ and $Q|_{\Delta_m} \in B^{\rho_m}_{1\infty}(\Delta_m)$ in Theorem 3.1 with the conditions $Q|_\Delta \in W_1^{l_0}(\Delta)$ and $Q|_{\Delta_m} \in W_1^{l_m}(\Delta_m)$, then the following estimate holds true:

$$\begin{aligned} E(u(x) - u_N(x))^2 &\leq \frac{C_1(s,\psi)}{4^{s(\mu_N+1)}} \cdot I_s + \frac{C_2(l_0,\phi)}{b_N^{2l_0-1}} ||Q||^2_{W_1^{l_0}(\Delta)} \\ &+ \sum_{m=0}^{\mu_N} \frac{C_3(l_m,\psi)}{b_{mN}^{2l_m-1}} \cdot \left(2^{-m} ||Q||^2_{L_1(\Delta_m)} + 2^{2m(l_m - 1/2)} ||Q||^2_{W_1^{l_m}(\Delta_m)} \right). \end{aligned} \tag{3.43}$$

But exponents $2l_m - 1$, $m = 0, 1, \ldots, \mu_N$ in this estimate might be not best possible. To illustrate this idea let us consider the following example.

Example 3.2. For $\rho \in (1, 2)$ let us consider the function $Q(k)$ constructed as above in Example 3.1 with $k_0 = 1/2$ and $q \in C^2(\mathbb{R})$ such that $\mathrm{supp}(q) \subset (-\frac{2}{3}, \frac{2}{3})$. Then the maximum possible value of an integer l_0 in the condition $Q|_\Delta \in W_1^{l_0}(\Delta)$ is $l_0 = 1$. Therefore by taking into account that $Q|_{\Delta_m} = 0$, $m = 1, 2, \ldots$, we get by (3.43) $E(u(x) - u_N(x))^2 \le C \cdot (1/b_N + 1/b_{0N})$, while the estimate (3.6) ensures more exact inequality is true: $E(u(x) - u_N(x))^2 \le C \cdot (1/b_N^{2\rho-1} + 1/b_{0N}^{2\rho-1})$.

4. Convergence of FWM in the C metric. Now let us turn to the problem of convergence of FWM (3.1) in probability, in the metric $C[0, 1]$. Recall that $\mathcal{E}_N(x) = u(x) - u_N(x)$. Denote by $D = [0, 1]$ the unit interval, and $C(D)$ is the space of continuous scalar functions on D with the uniform norm $||f||_{C(D)} = \max_{x\in D} |f(x)|$. Let us mention that the functional convergence $u_N \xrightarrow{P} u$ in $C(D)$ as $N \to \infty$ means $||\mathcal{E}_N||_{C(D)} \xrightarrow{P} 0$ as $N \to \infty$; that is, for each nonnegative ε and δ there exists $N_0 = N_0(\varepsilon, \delta)$ such that $\mathcal{P}\{||\mathcal{E}_N||_{C(D)} > \varepsilon\} < \delta$ for each $N \ge N_0$. Thus to study the convergence in probability $u_N \xrightarrow{P} u$ in $C(D)$ as $N \to \infty$, we need to estimate the probability $\mathcal{P}\{||\mathcal{E}_N||_{C(D)} > \varepsilon\}$.

Let $\xi(x)$, $x \in D = [0, 1]$, be a Gaussian random process with zero mean. Let

$$\varphi_\xi(h) = \sup_{\substack{x,y\in D \\ |x-y|\le h}} (E|\xi(x) - \xi(y)|^2)^{1/2}, \quad h \in [0, 1].$$

The following assertion is a one-dimensional variant of the well-known Fernique's inequality.

THEOREM 4.1 (see [9]). *If $\int_1^\infty \varphi_\xi(e^{-x^2})dx < \infty$, then almost all samples of the random process $\xi(x)$ are continuous. Moreover, for each $t \ge \sqrt{5}$ the following estimate is valid:*

$$\mathcal{P}\{\sup_{x\in D} |\xi(x)| \ge qt\} \le 10 \cdot \int_t^\infty e^{-x^2/2}\, dx, \tag{4.1}$$

where

$$q = \sup_{x\in D} (E|\xi(x)|^2)^{1/2} + (2 + \sqrt{2}) \cdot \int_1^\infty \varphi_\xi(2^{-x^2})dx.$$

By taking into account the inequality $\int_t^\infty e^{-x^2/2}\, dx \le \frac{e^{-t^2/2}}{t} \le \frac{e^{-t^2/2}}{\sqrt{5}}$ for $t \ge \sqrt{5}$ and the estimate (4.1), one can derive the following estimation:

$$\mathcal{P}\{\sup_{x\in D} |\xi(x)| \ge qt\} \le 2\sqrt{5} \cdot e^{-t^2/2}, \qquad t \ge \sqrt{5}. \tag{4.2}$$

Thus, in order to use this estimate for $\xi(x) = \mathcal{E}_N(x)$ we have to analyze the function

$$\varphi_{\mathcal{E}_N}(h) = \sup_{\substack{x,y\in D \\ |x-y|\le h}} (E|\mathcal{E}_N(x) - \mathcal{E}_N(y)|^2)^{1/2}, \quad h \in [0, 1].$$

Let us turn to estimating this function.

For a fixed positive integer N we denote $\epsilon_N^{(1)} = \frac{||Q||^2_{B^{\rho_0}_{1\infty}(\Delta)}}{b_N^{2\rho_0-1}}$, $\epsilon_N^{(3)} = \frac{I_s}{4^{(\mu_N+1)s}}$, and

$$\epsilon_{mN}^{(2)} = \frac{1}{b_{mN}^{2\rho_m-1}} \cdot \left(2^{-m}||Q||^2_{L_1(\Delta_m)} + 2^{2m(\rho_m-1/2)}||Q||^2_{b^{\rho_m}_{1\infty}(\Delta_m)}\right), \quad m = 0, 1, \dots, \mu_N.$$

THEOREM 4.2. *Assume that all of the conditions of Theorem* 3.1 *are satisfied. Then for each* $\alpha \in (0,1)$ *and* $h \in [0,1]$ *the following inequality is valid:*

$$E|\mathcal{E}_N(x+h) - \mathcal{E}_N(x)|^2 \le C(s) \cdot \left(I_s \cdot |h|^{s\wedge 1}\right)^{\alpha} \cdot \epsilon_N^{1-\alpha}, \quad N = 1, 2, \dots, \tag{4.3}$$

where $s \wedge 1 = \min\{s, 1\}$,

$$\epsilon_N = C_1(\rho_0, \phi) \cdot \epsilon_N^{(1)} + \sum_{m=0}^{\mu_N} C_2(\rho_m, \psi) \cdot \epsilon_{mN}^{(2)} + C_3(s, \psi) \cdot \epsilon_N^{(3)},$$

and C *and* C_i, $i = 1, 2, 3$, *are some constants depending only on the shown arguments.*

Proof. From independence of the random variables ξ_j, ξ_{mj}, $j \in \mathbb{Z}$, $m \ge 0$ it follows that (see (3.7))

$$\begin{aligned} E[\mathcal{E}_N(x+h) - \mathcal{E}_N(x)]^2 &= E[v^{(1)}(x+h) - v^{(1)}(x)]^2 \\ &+ \sum_{m=0}^{\mu_N} E[v_m^{(2)}(x+h) - v_m^{(2)}(x)]^2 + E[v^{(3)}(x+h) - v^{(3)}(x)]^2. \end{aligned} \tag{4.4}$$

The first and second terms in r.h.s. of the last equality are estimated in the following statements. The proof is given in the Appendix.

LEMMA 4.1. *For each* $\alpha \in (0,1)$ *the following estimate holds:*

$$E[v^{(1)}(x+h) - v^{(1)}(x)]^2 \le \left(C(s) I_s \cdot |h|^{(s\wedge 1)}\right)^{\alpha} \cdot (C_1(\rho_0, \phi) \cdot \epsilon_N^{(1)})^{1-\alpha}. \tag{4.5}$$

LEMMA 4.2. *For each* $\alpha \in (0,1)$ *the following estimate is valid for* $N = 1, 2, \dots$:

$$\sum_{m=0}^{\mu_N} E[v_m^{(2)}(x+h) - v_m^{(2)}(x)]^2 \le \left(C(s) \cdot I_s \cdot |h|^{s\wedge 1}\right)^{\alpha} \cdot \left(\sum_{m=0}^{\mu_N} C_2(\rho_m, \psi) \cdot \epsilon_{mN}^{(2)}\right)^{1-\alpha}. \tag{4.6}$$

It is easy to see that

$$E(v^{(3)}(x+h) - v^{(3)}(x))^2 = \sum_{m=\mu_N+1}^{\infty} \sum_{j\in\mathbb{Z}} |a_{mj} - b_{mj}|^2,$$

where $a_{mj} = \mathcal{F}_m^{(\psi)}(2^m \cdot (x+h) + j)$ and $b_{mj} = \mathcal{F}_m^{(\psi)}(2^m \cdot x + j)$. Therefore by using

$$\sum_{j\in\mathbb{Z}} |a_{mj} - b_{mj}|^2 \le ||\Delta_h G||^2_{L_2}$$

for each integer m

$$\sum_{m=\mu_N+1}^{\infty} \sum_{j\in\mathbb{Z}} (|a_{mj}| + |b_{mj}|)^2 \le 2 \cdot [E(v^{(3)}(x+h))^2 + E(v^{(3)}(x))^2],$$

and taking into account (3.12) and (A.7), we get

$$E(v^{(3)}(x+h)-v^{(3)}(x))^2 \le \left(C(s)\cdot I_s\cdot |h|^{s\wedge 1}\right)^{\alpha}\cdot\left(C_3(s,\psi)\cdot\epsilon_N^{(3)}\right)^{1-\alpha}. \tag{4.7}$$

Finally, by using (4.4), (4.5), (4.6), and the last inequality we obtain

$$\begin{aligned} &E|\mathcal{E}_N(x+h)-\mathcal{E}_N(x)|^2 \\ &\le \left(C(s)\cdot I_s\cdot |h|^{s\wedge 1}\right)^{\alpha}\cdot\left[(C_1\cdot\epsilon_N^{(1)})^{1-\alpha}+\left(\sum_{m=0}^{\mu_N} C_{2m}\cdot\epsilon_{mN}^{(2)}\right)^{1-\alpha}+(C_3\cdot\epsilon_N^{(3)})^{1-\alpha}\right] \\ &\le 3^{\alpha}\left(C(s)\cdot I_s\cdot |h|^{s\wedge 1}\right)^{\alpha}\cdot\left[C_1\cdot\epsilon_N^{(1)}+\sum_{m=0}^{\mu_N} C_{2m}\cdot\epsilon_{mN}^{(2)}+C_3\cdot\epsilon_N^{(3)}\right]^{1-\alpha}. \end{aligned} \tag{4.8}$$

Theorem 4.2 is proven. □

Now we are in a position to formulate an assertion on the convergence of FWM in probability.

THEOREM 4.3. *Assume that all of the conditions of the Corollary* 3.1 *are fulfilled. Then for each* $\alpha\in(0,1)$ *and* $t>\sqrt{5}$

$$\mathcal{P}\left\{\sup_{x\in[0,1]}|\mathcal{E}_N(x)|>C\cdot\varepsilon_N(\alpha)\cdot t\right\}\le 2\sqrt{5}\cdot e^{-t^2/2}, \tag{4.9}$$

where C *is some positive constant not depending on* α *and* N *and*

$$\varepsilon_N(\alpha)=\frac{1}{\sqrt{\alpha}}\cdot N^{-\frac{(2\rho_0-1)\cdot(1-\alpha)}{2}}.$$

Proof. From the estimate in (4.3), by taking into account (3.5) and (3.25) we get

$$\sup_{x,y\in[0,1],\,|x-y|\le h} E(\mathcal{E}_N(x)-\mathcal{E}_N(y))^2\le C_1|h|^{\alpha(s\wedge 1)}N^{-(2\rho_0-1)\cdot(1-\alpha)}$$

for each $\alpha\in(0,1)$ and $|h|\le 1$, where C_1 is some positive constant not depending on α, h, and N. Now let $\xi(x)=\mathcal{E}_N(x)$. Then

$$\varphi_\xi(h)=\sup_{\substack{x,y\in D\\ |x-y|\le h}} E^{1/2}(|\xi(x)-\xi(y)|^2)\le C_2\cdot|h|^{\alpha(s\wedge 1)/2}N^{-(2\rho_0-1)\cdot(1-\alpha)/2}.$$

Therefore, by (3.26)

$$\begin{aligned} q=\sup_{x\in D}E^{1/2}|\xi(x)|^2+(2+\sqrt{2})\cdot\int_1^{\infty}\varphi_\xi(2^{-x^2})dx &\le C_3\cdot N^{-\frac{(2\rho_0-1)}{2}} \\ &+C_4\cdot N^{-\frac{(2\rho_0-1)\cdot(1-\alpha)}{2}}\cdot\int_1^{\infty}2^{-x^2\alpha(s\wedge 1)/2)}dx\le C\varepsilon_N(\alpha). \end{aligned}$$

Now (4.9) is a direct consequence of estimation (4.2). Theorem 4.3 is proven. □

COROLLARY 4.1. *Under the conditions of Theorem* 4.3, $u_N\xrightarrow{P}u$ *in the metric of* $C[0,1]$ *as* $N\to\infty$.

Proof. By (4.9), for each $\varepsilon>0$ and $\delta>0$ it can be found a positive integer N_0 such that

$$\mathcal{P}\left\{\sup_{x\in[0,1]}|\mathcal{E}_N(x)|>\varepsilon\right\}\le\delta,\qquad N\ge N_0.$$ □

5. Generalization of results to vector processes and some stronger metrics.

5.1. Convergence in the L_p metric. Recall the well-known relation between different moments of zero mean Gaussian random variables. Let ξ be a zero mean real-valued Gaussian random variable. Then (e.g., see, [16])

$$E|\xi|^p = c_p(E\xi^2)^{p/2}, \quad p \geq 1, \quad c_p = \frac{2^{p/2}}{\sqrt{\pi}}. \tag{5.1}$$

Therefore, under the conditions of Theorem 3.1, by using the estimate in (3.6) for $E\mathcal{E}_N^2(x) = E(u(x) - u_N(x))^2$ it is easy to obtain an estimate for the convergence rate $E||\mathcal{E}_N||^p_{L_p[0,1]} = \int_0^1 E(\mathcal{E}_N(x))^p\,dx$ in the metric of $L_p[0,1]$, $p \geq 1$.

5.2. Convergence in the metric of the Sobolev space W_p^n. For a positive integer n and $p \geq 1$, let $W_p^n[0,1]$ be a Sobolev space with the norm

$$||f||_{W_p^n[0,1]} = ||f||_{L_p[0,1]} + ||f^{(n)}||_{L_p[0,1]}.$$

In the study of convergence $E||\mathcal{E}_N||^p_{W_p^n[0,1]} \to 0$, due to the relation (5.1), we can restrict ourselves to the case $p = 2$. Therefore let us consider the convergence of the FWM (3.1) in the Sobolev space W_2^n. In order to analyze the convergence of FWM, $u_N(x)$ to $u(x)$ as $N \to \infty$ in $W_2^n[0,1]$, it is sufficient to study: (i) convergence $u_N \to u$ in $L_2[0,1]$ and (ii) convergence $u_N^{(n)} \to u^{(n)}$ in $L_2[0,1]$. The first question was already studied in section 3. Therefore we will concentrate on the second question: $u_N^{(n)} \to u^{(n)}$ in $L_2[0,1]$. Recall that for a function $f : \mathbb{R} \to \mathbb{R}$ we denote by $f^{(n)}$ the nth derivative $D^n f$ (if it exists).

First of all note that a necessary and sufficient condition that samples of a random process $u(x)$ belong to $W_2^n[0,1]$ (with probability 1) is (e.g., see, [15])

$$\int_{-\infty}^{\infty} |k|^{2n} F(k)\,dk < \infty. \tag{5.2}$$

If this condition is satisfied, then $D^n u(x) = u^{(n)}$ have the spectral function $k^{2n}F(k)$. In order to ensure that u_N is n-times differentiable we have to choose the cutoff function η_A n-times differentiable. Therefore we will assume that $\eta_A \in C^n$ and

$$\eta_A(x) = \begin{cases} 0, & 0 \leq x \leq A, \\ 1, & x \geq A+1. \end{cases} \tag{5.3}$$

Thus, by using results of section 3 one can formulate sufficient conditions for $u_N^{(n)} \to u^{(n)}$ as $N \to \infty$ in $L_2[0,1]$. Denote by Q_n the function $Q_n(k) = |k|^n Q(k) = |k|^n F^{1/2}(k)$. Then the following assertion is valid.

THEOREM 5.1. *Assume that the spectral function $F(k)$ satisfies the condition*

$$I_{n+s} \stackrel{def}{=} \int_{-\infty}^{\infty} F(k)(1+|k|^2)^{n+s}dk < \infty \tag{5.4}$$

for some $s > 0$. Let $\{\rho_m\}_{m=0}^{\mu_N}$ be a finite sequence of positive numbers such that $\min_m \rho_m > 1/2$. Assume that $Q_n|_\Delta \in B_{1\infty}^{\rho_0}(\Delta)$ and $Q|_{\Delta_m} \in B_{1\infty}^{\rho_m}(\Delta_m)$ for $1 \leq m \leq$

μ_N. Assume that the Meyer wavelet functions ϕ and ψ belong to the class $C^{\nu+1}$ for $\nu = \max\{\lfloor\rho\rfloor, \lfloor s\rfloor\}$, where $\rho = \max_{0\le m\le\mu_N}\rho_m$. Then

$$E(u^{(n)}(x) - u_N^{(n)}(x))^2 \le \frac{C_1(s,\psi)}{4^{s(\mu_N+1)}}\cdot I_{n+s} + \frac{C_2(\rho_0,\phi)}{b_N^{2\rho_0-1}}||Q_n||^2_{B^{\rho_0}_{1\infty}(\Delta)}$$

$$+\sum_{m=0}^{\mu_N}\frac{C_3(\rho_m,\psi)}{b_{mN}^{2\rho_m-1}}\cdot\left(2^{-2ms}\cdot I_{n+s} + 2^{2m(\rho_m-1/2)}||Q_n||^2_{b^{\rho_m}_{1\infty}(\Delta_m)}\right), \tag{5.5}$$

where C_i, $i = 1,2,3$, are some positive constants depending only on the shown arguments.

Proof. Indeed, this assertion is a direct consequence of Theorem 3.1 applied to the random process $u^{(n)}(x)$ with the spectral function $|k|^{2n}F(k)$. Then the estimate (5.5) is a consequence of (3.6) and the following simple inequality:

$$||Q_n||_{L_1(\Delta_m)} = \int_{\Delta_m}|k|^n Q(k)dk \le |\Delta_m|^{1/2}\left(\int_{\Delta_m}|k|^{2n}F(k)dk\right)^{1/2}$$
$$\le 2^{(m+1)/2-ms}I^{1/2}_{n+s}. \qquad \square$$

PROPOSITION 5.1. *Assume that, for a nonnegative k_0 and positive integer l, $Q \in W_1^l(k_0,\infty)$ and, for each nonnegative integer j such that $0 \le j \le n\wedge l$, the following condition is valid:*

$$|Q^{(l-j)}(k)| \le \frac{A_\varepsilon}{k^{n+l-j+1/2+\varepsilon}} \quad \forall k \ge k_0 \tag{5.6}$$

for some positive numbers A_ε and ε. Then

$$||Q_n||_{b^l_{1\infty}(\Delta_m)} \le A_\varepsilon\cdot c_0\cdot 2^{-m(l-1/2+\varepsilon)} \quad \forall\ m \ge \log_2(3(k_0+2)) \tag{5.7}$$

for some positive $c_0 = c_0(n,l,\varepsilon)$.

Proof. Indeed,

$$|\Delta_h^{(2)}Q_n^{(l-1)}(k)| \le |\Delta_h Q_n^{(l-1)}(k+h)| + |\Delta_h Q_n^{(l-1)}(k+h)|$$
$$\le h\cdot|Q_n^{(l)}(k+\beta h)| + h\cdot|Q_n^{(l)}(k+\alpha h)|$$

for some $\alpha\in[0,1]$ and $\beta\in[1,2]$. Since

$$Q_n^{(l)}(k) = \sum_{j=0}^{n\wedge l}C_l^j\cdot n\cdot(n-1)\cdot\cdots\cdot(n-j+1)\cdot k^{n-j}Q^{(l-j)}(k),$$

we find for each $k\in\Delta_m$ and $m\ge\log_2(3(k_0+2))$ that $|Q_n^{(l)}(k)| \le C(n,l)2^{-m(l+1/2+\varepsilon)}$ by virtue of condition (5.6). Therefore,

$$\int_{\Delta_m}|\Delta_h^{(2)}Q^{(l-1)}(k)| \le C(n,l)2^{-m(l+1/2+\varepsilon)}\cdot h\cdot|\Delta_m| \le C_1(n,l)\cdot h\cdot 2^{-m(l-1/2+\varepsilon)}.$$

This completes the proof of (5.7). $\square$

COROLLARY 5.1. *Assume that the spectral function $F(k)$ satisfies the condition (5.4) for some $s>0$ and a positive integer n and that the function $Q_n(k) = |k|^n\cdot F^{1/2}(k) = |k|^n\cdot Q(k)$ satisfies the conditions*

$$Q_n|_\Delta \in B^{\rho_0}_{1\infty}(\Delta), \quad Q|_{\Delta'}\in W_1^l(\Delta') \tag{5.8}$$

for some positive $\rho_0 > 1/2$ and positive integer l. Assume that for each nonnegative integer $j \in [0, n \wedge l]$ the function Q satisfies the condition (5.6)*, and suppose that $\phi, \psi \in C^{\nu+1}$, where $\nu = \max\{\lfloor s \rfloor, \lfloor \rho_0 \rfloor, l\}$. Let us choose in FWM* (3.1)

$$\mu_N = \left\lfloor \frac{(2\rho_0 - 1)}{s \ln 4} \cdot \ln N \right\rfloor, \quad b_N = N, \quad b_{mN} = N^{\frac{2\rho_0 - 1}{2l-1}},\ m = 0, 1, \ldots, \mu_N. \tag{5.9}$$

Then for each positive integer N and $x \in \mathbb{R}$ the following estimate is valid:

$$\begin{aligned} &E(u^{(n)}(x) - u_N^{(n)}(x))^2 \le N^{-(2\rho_0 - 1)} \\ &\quad \times \left[C_1(s, \psi) I_{n+s} + C_2(\rho_0, \phi) \cdot ||Q_n||^2_{B^{\rho_0}_{1\infty}(\Delta)} + C_3(n, \rho_1, \varepsilon, \psi) \cdot C_\varepsilon^2 \right], \end{aligned} \tag{5.10}$$

where C_i, $i = 1, 2, 3$, are some positive constants depending only on the shown arguments.

Thus if the spectral function $F(\cdot)$ satisfies all of the conditions of the last corollary and $Q|_\Delta \in B^{\rho_0}_{1\infty}(\Delta)$, then the estimates (3.32) and (5.10) imply the following result on convergence of FWM in W_2^n.

COROLLARY 5.2. *Let all of the conditions of Corollary* 5.1 *be fulfilled and $Q|_\Delta \in B^{\rho_0}_{1\infty}(\Delta)$. Then*

$$E||\mathcal{E}_N||^2_{W_2^n[0,1]} \le \frac{C}{N^{2\rho_0 - 1}}, \quad N = 1, 2, \ldots,$$

for some positive C not depending on N.

5.3. Convergence in $C^n[0, 1]$. For positive integer n, let $C^n[0, 1]$ be the space of all n-times continuously differentiable real-valued scalar functions $f : [0, 1] \to \mathbb{R}$ with the norm

$$||f||_{C^n[0,1]} = \sup_{x \in [0,1]} |f(x)| + \sup_{x \in [0,1]} |f^{(n)}(x)|.$$

By the definition of this norm it follows that $\{u_N(x)\}_{N=1}^\infty$, $x \in [0, 1]$ converges in the metric of $C^n[0, 1]$ to a random field $u(x)$, $x \in [0, 1]$, as $N \to \infty$ iff: (i) $u_N \to u$ in $C[0, 1]$ as $N \to \infty$ and (ii) $u_N^{(n)} \to u^{(n)}$ in $C[0, 1]$ as $N \to \infty$. Since the first question was studied in section 4, we concentrate on the second one. To ensure that the samples of FWM (3.1) are of $C^n[0, 1]$, we assume below that $\eta_A \in C^n(\mathbb{R})$ and the condition (5.3) is satisfied.

The following assertion is an analog of Theorem 4.3 and can be proved in a similar way.

THEOREM 5.2. *Let all of the conditions of Corollary* 5.1 *be fulfilled. Then for each $\alpha \in (0, 1)$ and $t > \sqrt{5}$,*

$$\mathcal{P}\left\{ \sup_{x \in [0,1]} |\mathcal{E}_N^{(n)}(x)| > C\varepsilon_N(\alpha) \cdot t \right\} \le 2\sqrt{5} \cdot e^{-t^2/2}, \tag{5.11}$$

where C is a positive constant not depending on α and N and

$$\varepsilon_N(\alpha) = \frac{1}{\sqrt{\alpha}} \cdot N^{-\frac{(2\rho_0 - 1) \cdot (1 - \alpha)}{2}}.$$

Now let all of the conditions of Corollary 5.2 be valid. Then by using the simple

inequality $\mathcal{P}\{\xi+\eta>x+y\}\le\mathcal{P}\{\xi>x\}+\mathcal{P}\{\eta>y\}$ and the results of Theorems 4.3 and 5.2 we arrive at the following.

COROLLARY 5.3. *Let all of the conditions of Corollary* 5.2 *be satisfied. Then for each* $\alpha\in(0,1)$ *and* $t>\sqrt{5}$

$$\mathcal{P}\left\{||\mathcal{E}_N||_{C^n[0,1]}>\frac{C\cdot t}{\sqrt{\alpha}}\cdot N^{-\frac{(2\rho_0-1)\cdot(1-\alpha)}{2}}\right\}\le 4\sqrt{5}\cdot e^{-t^2/2},$$

where C *is some positive constant not depending on* α.

Under the conditions of this corollary it follows immediately that $u_N\xrightarrow{P}u$ in the metric of $C^n[0,1]$ as $N\to\infty$.

5.4. Generalizations for vector processes. Let $\mathbf{u}(x)=(u_1(x),\dots,u_l(x))^T$, $x\in\mathbf{R}$, be an l-dimensional real-valued stationary Gaussian random process with mean zero and spectral tensor $F(k)=(F_{ij}(k))$, $i=1,\dots,l$, $j=1,\dots,l$, and $\mathbf{u}_N(x)$ is its FWM given by (2.12). Denote by $|\mathbf{a}|=(\sum_{i=1}^l a_i^2)^{1/2}$ the Euclidean norm of an l-dimensional real-valued vector $\mathbf{a}$ and by $|A|=(\sum_{i=1}^l\sum_{j=1}^n a_{ij}^2)^{1/2}$ the Euclidean norm of a matrix $A=(a_{ij})$, $i=1,\dots l$, $j=1,\dots,n$, which can be realized as the Euclidean norm of the $l\times n$-vector A. For an $l\times n$-dimensional function $Q(k)=(q_{ij}(k))$, $i=1,\dots l$, $j=1,\dots,n$, and $D\subset\mathbb{R}$, denote (if it makes a sense)

$$|||Q|||_{L_1(D)}=\left(\sum_{i=1}^l\sum_{j=1}^n||q_{ij}||^2_{L_1(D)}\right)^{1/2},\quad |||Q|||_{b^\rho_{1\infty}(D)}=\left(\sum_{i=1}^l\sum_{j=1}^n||q_{ij}||^2_{b^\rho_{1\infty}(D)}\right)^{1/2}.$$

Finally, denote by $SpF(k)$ the trace $\sum_{i=1}^l F_{ii}(k)$ of the spectral tensor F. Then by the same arguments as in the scalar case one can establish the following result.

THEOREM 5.3. *Assume that the spectral tensor* $F(k)$ *satisfies the condition*

$$I_s\stackrel{def}{=}\int_{-\infty}^{\infty}SpF(k)(1+|k|^2)^s dk<\infty \tag{5.12}$$

for some $s>0$. *Let* $\{\rho_m\}_{m=0}^{\mu_N}$ *be a finite sequence of positive numbers such that* $\min_m\rho_m>1/2$. *Assume that* $|||Q|||_{L_1(\Delta)}<\infty$, $|||Q|||_{b^{\rho_0}_{1\infty}(\Delta)}<\infty$, *and*

$$|||Q|||_{L_1(\Delta_m)}<\infty,\quad |||Q|||_{b^{\rho_m}_{1\infty}(\Delta_m)}<\infty,\quad m=0,1,\dots,\mu_N.$$

Assume that the Meyer wavelet functions ϕ *and* ψ *belong to the class* $C^{\nu+1}$ *for* $\nu=\max\{\lfloor\rho\rfloor,\lfloor s\rfloor\}$, *where* $\rho=\max_{0\le m\le\mu_N}\rho_m$. *Then*

$$\begin{aligned}E|u(x)-u_N(x)|^2\le{}&\frac{C_1(s,\psi)}{4^{s(\mu_N+1)}}\cdot I_s+\frac{C_2(\rho_0,\phi)}{b_N^{2\rho_0-1}}\cdot\left\{|||Q|||^2_{L_1(\Delta)}+|||Q|||^2_{b^{\rho_0}_{1\infty}(\Delta)}\right\}\\ &+\sum_{m=0}^{\mu_N}\frac{C_3(\rho_m,\psi)}{b_{mN}^{2\rho_m-1}}\cdot\left(2^{-m}|||Q|||^2_{L_1(\Delta_m)}+2^{2m(\rho_m-1/2)}|||Q|||^2_{b^{\rho_m}_{1\infty}(\Delta_m)}\right),\end{aligned} \tag{5.13}$$

where C_i, $i=1,2,3$, *are some constants depending only on the shown arguments.*

Analogous convergence results can be obtained in stronger metrics.

6. Conclusion and discussion. Functional convergence of FWM for stationary Gaussian random processes was studied in Sobolev spaces $W_p^n[0,1]$ and in the space of n-times continuously differentiable functions $C^n[0,1]$. Conditions sufficient for the

convergence of FWM are formulated in terms of spectral functions F (spectral tensors, in the case of vector random processes). Two kinds of assumptions on the behavior of the spectral functions are made: (i) the finiteness of certain spectral moments and (ii) a generalized smoothness of the function $Q = F^{1/2}$ in different wave bands. This is formulated in terms of smoothness in Besov's space $B^{\rho}_{1\infty}$.

Condition (i) is related to the behavior of the high-frequency part of the spectrum which is standard in the convergence studies. Condition (ii) is related to the rate of convergence of the tails of the wavelet function $\mathcal{F}^{\psi}_{m}$ for a given spectral band. For these tails, we obtained upper estimations which are improvable in the sense that for some functions these estimations are exact to within a constant factor.

This analysis is new and provides a constructive algorithm for choosing the cutoff parameters for all spectral bands to ensure a uniform behavior of the error on the whole spectral interval.

We give also estimates of the cost needed to guarantee the desired root mean square error ε in $L_2[0,1]$ depending on the smoothness parameter of the relevant Besov space. The typical behavior of the cost has the form $T_{\varepsilon} \lesssim \frac{\ln \varepsilon}{\varepsilon^{1/(\rho_{min}-0.5)}}$, where ρ_{min} is the minimal smoothness parameter.

The Fourier-wavelet models are well suited for simulation of random processes with smooth spectral functions as is clearly seen from our presentation. However, in case the spectral function is not smooth in an isolated point, it is reasonable to use a hybrid method as follows. The spectral function is decomposed into two parts, the first being smooth, and the second is nonsmooth, and has a compact support in the neighborhood of the isolated point. Then the random process is represented respectively as a sum of two independent processes, the first having the smooth spectral function and the second with the nonsmooth spectral function. The smooth part is simulated by FWM and the second part by a standard deterministic spectral method which takes into account the singularity.

Appendix A. Some technical details. We present here some technical results which we use in the main part of this paper.

Proof of Lemma 4.1. For each $\alpha \in (0,1)$ one has

$$\begin{aligned} &E[v^{(1)}(x+h) - v^{(1)}(x)]^2 \\ &\quad = \sum_{j\in\mathbb{Z}} [\chi_N(x+h+j)\cdot \mathcal{F}^{(\phi)}(x+h+j) - \chi_N(x+j)\cdot \mathcal{F}^{(\phi)}(x+j)]^2 \\ &\quad = \sum_{j\in\mathbb{Z}} |a_j - b_j|^2 \le \left(\sum_{j\in\mathbb{Z}} |a_j-b_j|^2\right)^{\alpha} \cdot \left(\sum_{j\in\mathbb{Z}} (|a_j|+|b_j|)^2\right)^{1-\alpha}. \end{aligned} \tag{A.1}$$

Here we denote $a_j = \chi_N(x+h+j)\cdot \mathcal{F}^{(\phi)}_0(x+h+j)$ and $b_j = \chi_N(x+j)\cdot \mathcal{F}^{(\phi)}_0(x+j)$, and in the second line of (A.1) we use Hölder's inequality:

$$\begin{aligned} \sum_{j\in\mathbb{Z}} |a_j-b_j|^2 &\le \sum_{j\in\mathbb{Z}} \left\{ |a_j-b_j|^{2\alpha} \cdot (|a_j|+|b_j|)^{2\cdot(1-\alpha)} \right\} \\ &\le \left(\sum_{j\in\mathbb{Z}} |a_j-b_j|^{2\alpha p}\right)^{1/p} \cdot \left(\sum_{j\in\mathbb{Z}} (|a_j|+|b_j|)^{2\cdot(1-\alpha)q}\right)^{1/q}, \end{aligned} \tag{A.2}$$

with parameters $p = 1/\alpha$ and $q = 1/(1-\alpha)$. Let us estimate each multiplier in the

right-hand side of (A.1). First note that

$$\begin{aligned}|a_j - b_j| &= |\chi_N(x+h+j)\cdot \mathcal{F}_0^{(\phi)}(x+h+j) - \chi_N(x+j)\cdot \mathcal{F}_0^{(\phi)}(x+j)| \\ &\le |\chi_N(x+h+j)\cdot \mathcal{F}_0^{(\phi)}(x+h+j) - \chi_N(x+h+j)\cdot \mathcal{F}_0^{(\phi)}(x+j)| \\ &\quad + |\chi_N(x+h+j)\cdot \mathcal{F}_0^{(\phi)}(x+j) - \chi_N(x+j)\cdot \mathcal{F}_0^{(\phi)}(x+j)| \\ &\le |\mathcal{F}_0^{(\phi)}(x+h+j) - \mathcal{F}_0^{(\phi)}(x+j)| \\ &\quad + |\mathcal{F}_0^{(\phi)}(x+j)|\cdot|\chi_N(x+h+j) - \chi_N(x+j)|. \end{aligned} \tag{A.3}$$

Therefore,

$$\begin{aligned}\sum_{j\in\mathbb{Z}} |a_j - b_j|^2 &\le 2\cdot \sum_{j\in\mathbb{Z}} [\mathcal{F}_0^{(\phi)}(x+h+j) - \mathcal{F}_0^{(\phi)}(x+j)]^2 \\ &\quad + 2\cdot \sum_{j\in\mathbb{Z}} |\mathcal{F}_0^{(\phi)}(x+j)|^2 \cdot |\chi_N(x+h+j) - \chi_N(x+j)|^2. \end{aligned} \tag{A.4}$$

Note that $\{\mathcal{F}_0^{(\phi)}(x+h+j) - \mathcal{F}_0^{(\phi)}(x+j)\}_{j\in\mathbb{Z}}$ and $\{\mathcal{F}_0^{(\phi)}(x+j)\}_{j\in\mathbb{Z}}$ are the Fourier coefficients of $\Delta_h G(x+\cdot)$ and $G(x+\cdot)$, respectively, in the orthonormal system $\{\phi(\cdot+j)\}_{j\in\mathbb{Z}}$. Therefore

$$\sum_{j\in\mathbb{Z}} |\mathcal{F}_0^{(\phi)}(x+j)|^2 \le ||G(x+\cdot)||_{L_2}^2 = ||G(\cdot)||_{L_2}^2 \le I_s \tag{A.5}$$

and

$$\sum_{j\in\mathbb{Z}} [\mathcal{F}_0^{(\phi)}(x+h+j) - \mathcal{F}_0^{(\phi)}(x+j)]^2 \le ||\Delta_h G(x+\cdot)||_{L_2}^2 = ||\Delta_h G(\cdot)||_{L_2}^2. \tag{A.6}$$

Note that $G \in B_{22}^s \hookrightarrow B_{2\infty}^s \hookrightarrow B_{2\infty}^\gamma$ for each $0 < \gamma \le s$. Therefore, by taking $\gamma = \min\{1/2, s/2\}$ we have

$$\begin{aligned}||\Delta_h G(\cdot)||_{L_2} \le C_1(s)\cdot|h|^\gamma \cdot ||\Delta_h G(\cdot)||_{B_{2\infty}^\gamma} &\le C_2(s)\cdot|h|^\gamma\cdot||\Delta_h G(\cdot)||_{B_{22}^s} \\ &\le C_3(s)\cdot|h|^\gamma\cdot I_s^{1/2} \end{aligned} \tag{A.7}$$

for each $h \in [-1, 1]$. From (A.4)–(A.7), by taking into account that $|\chi_N(x+h+j) - \chi_N(x+j)| \le |h|$, we obtain

$$\sum_{j\in\mathbb{Z}} |a_j - b_j|^2 \le C(s)\, I_s \cdot |h|^{s\wedge 1}, \tag{A.8}$$

where $s \wedge 1 = \min\{1, s\}$.

Now let us estimate the second factor on the r.h.s. of (A.1). From

$$\sum_{j\in\mathbb{Z}} (|a_j| + |b_j|)^2 \le 2\cdot\sum_{j\in\mathbb{Z}} |a_j|^2 + 2\cdot\sum_{j\in\mathbb{Z}} |b_j|^2 = 2\cdot E(v^{(1)}(x+h))^2 + 2\cdot E(v^{(1)}(x))^2$$

and (3.15), we obtain

$$\sum_{j\in\mathbb{Z}} (|a_j| + |b_j|)^2 \le \frac{4\cdot C(\rho_0,\phi)}{b_N^{2\rho_0-1}} \cdot ||Q||_{B_{1\infty}^{\rho_0}(\Delta)}^2. \tag{A.9}$$

Thus by virtue of (A.1) and (A.8)–(A.9) the estimate (4.5) follows. The proof of Lemma 4.1 is complete. □

Proof of Lemma 4.2. For each $\alpha \in (0,1)$, by the same arguments as we used in proof of Lemma 4.1, we arrive at

(A.10)
$$E[v_m^{(2)}(x+h)-v_m^{(2)}(x)]^2 \le \left(\sum_{j\in\mathbb{Z}} |a_{mj}-b_{mj}|^2\right)^{\alpha} \cdot \left(\sum_{j\in\mathbb{Z}} (|a_{mj}|+|b_{mj}|)^2\right)^{1-\alpha},$$

where

$$a_{mj} = \chi_{mN}(2^m\cdot(x+h)+j)\cdot\mathcal{F}_m^{(\psi)}(2^m\cdot(x+h)+j), \quad b_{mj} = \chi_{mN}(2^m\cdot x+j)\cdot\mathcal{F}_m^{(\psi)}(2^m\cdot x+j).$$

From (A.10) it follows that

$$\sum_{m=0}^{\mu_N} E[v_m^{(2)}(x+h)-v_m^{(2)}(x)]^2 \le \left(\sum_{m=0}^{\mu_N}\sum_{j\in\mathbb{Z}} |a_{mj}-b_{mj}|^2\right)^{\alpha}\left(\sum_{m=0}^{\mu_N}\sum_{j\in\mathbb{Z}} (|a_{mj}|+|b_{mj}|)^2\right)^{1-\alpha}.$$

Now obviously

$$\begin{aligned} |a_{mj}-b_{mj}| &\le |\mathcal{F}_m^{(\psi)}(2^m\cdot(x+h)+j)-\mathcal{F}_m^{(\psi)}(2^m\cdot x+j)| \\ &\quad +|\mathcal{F}_m^{(\psi)}(2^m\cdot x+j)|\cdot|\chi_{mN}(2^m\cdot(x+h)+j)-\chi_{mN}(2^m\cdot x+j)|. \end{aligned} \tag{A.11}$$

Since $\{\mathcal{F}_m^{(\psi)}(2^m\cdot(x+h)+j)-\mathcal{F}_m^{(\psi)}(2^m\cdot x+j)\}_{j\in\mathbb{Z}}$, $m=0,\dots,\mu_N$, are Fourier coefficients in the expansion of $\Delta_h G(x+\cdot)$ with respect to the orthonormal system $\{\psi_{mj}\}_{j\in\mathbb{Z}}$, $m=0,\dots,\mu_N$, then (see (A.7)–(A.8))

(A.12)
$$\sum_{m=0}^{\mu_N}\sum_{j\in\mathbb{Z}} |\mathcal{F}_m^{(\psi)}(2^m(x+h)+j)-\mathcal{F}_m^{(\psi)}(2^m x+j)|^2 \le ||\Delta_h G||_{L_2} \le C(s)\, I_s |h|^{s\wedge 1}.$$

When we estimate the second term in the r.h.s. of (A.11) we consider two cases: (i) $2^m|h| \le 1$ and (ii) $2^m|h| > 1$. In the first case, by taking into account that $|\chi_{mN}(2^m\cdot(x+h)+j)-\chi_{mN}(2^m\cdot x+j)|^2 \le (2^m|h|)^2$ (see (A.7) and (3.11)), we have

$$\begin{aligned} &\sum_{m=0}^{\mu_N}\sum_{j\in\mathbb{Z}} |\mathcal{F}_m^{(\psi)}(2^m\cdot x+j)|^2\cdot|\chi_{mN}(2^m\cdot(x+h)+j)-\chi_{mN}(2^m\cdot x+j)|^2 \\ &\quad \le (2^m|h|)^2\cdot\sum_{m=0}^{\mu_N}\sum_{j\in\mathbb{Z}} |\mathcal{F}_m^{(\psi)}(2^m\cdot x+j)|^2 \le (2^m|h|)^{s\wedge 1}\cdot\sum_{m=0}^{\mu_N}\sum_{j\in\mathbb{Z}} |\mathcal{F}_m^{(\psi)}(2^m\cdot x+j)|^2 \\ &\quad \le B_s^2\cdot I_s\cdot|h|^{s\wedge 1} \qquad \text{if} \quad 2^m|h|\le 1. \end{aligned}$$

In the second case (i.e., $2^m|h|>1$) it is obvious that $(2^m|h|)^s > 1$ and $|\chi_{mN}(2^m\cdot(x+h)+j)-\chi_{mN}(2^m\cdot x+j)|^2 \le 1$. Therefore, in this case

$$\begin{aligned} &\sum_{m=0}^{\mu_N}\sum_{j\in\mathbb{Z}} |\mathcal{F}_m^{(\psi)}(2^m\cdot x+j)|^2\cdot|\chi_{mN}(2^m\cdot(x+h)+j)-\chi_{mN}(2^m\cdot x+j)|^2 \\ &\quad \le \sum_{m=0}^{\mu_N}\sum_{j\in\mathbb{Z}} |\mathcal{F}_m^{(\psi)}(2^m\cdot x+j)|^2 \le (2^m|h|)^{2s}\cdot\sum_{m=0}^{\mu_N}\sum_{j\in\mathbb{Z}} |\mathcal{F}_m^{(\psi)}(2^m\cdot x+j)|^2 \\ &\quad \le B_s 2\cdot I_s\cdot|h|^{2s} \qquad \text{if} \quad 2^m|h|>1. \end{aligned} \tag{A.13}$$

Thus in the general case we have

$$\sum_{m=0}^{\mu_N}\sum_{j\in\mathbb{Z}}|\mathcal{F}_m^{(\psi)}(2^m\cdot x+j)|^2\cdot|\chi_{mN}(2^m\cdot(x+h)+j)-\chi_{mN}(2^m\cdot x+j)|^2$$

$$\le B_s^2\cdot I_s\cdot|h|^{s\wedge 1}\qquad\text{if}\quad |h|\le 1. \tag{A.14}$$

By using (A.12) and (A.14) and taking into account (A.11) we obtain

$$\sum_{m=0}^{\mu_N}\sum_{j\in\mathbb{Z}}|a_j-b_j|^2\le C(s)\cdot I_s\cdot|h|^{s\wedge 1}\qquad\text{if}\quad |h|\le 1. \tag{A.15}$$

To estimate the second factor on the r.h.s. of inequality (A.10) we first notice that

$$\sum_{j\in\mathbb{Z}}(|a_j|+|b_j|)^2\le 2\cdot E(v_m^{(2)}(x+h))^2\ +\ 2\cdot E(v_m^{(2)}(x))^2.$$

Then by using the inequality (3.21) we get

$$\sum_{j\in\mathbb{Z}}(|a_j|+|b_j|)^2\le\frac{C(\rho_m,\psi)}{b_{mN}^{2\rho_m-1}}\cdot 2^{-m}||Q||^2_{L_1(\Delta_m)}+\frac{C(\rho_m,\psi)}{b_{mN}^{2\rho_m-1}}\cdot 2^{m(2\rho_m-1)}||Q||^2_{b^{\rho_m}_{1\infty}(\Delta_m)} \tag{A.16}$$

for $m=0,1,\dots,\mu_N$.

From (A.10), (A.15), and (A.16) we get (4.6). The proof of Lemma 4.2 is complete. □

LEMMA A.1. *Let $\hat f\in B^r_{1\infty}(\mathbb{R})$, $r>1/2$. Then f is uniformly continuous, $f\in L_2(\mathbb{R})$, and there exists a positive constant C_r depending only on r such that for all $x\in\mathbb{R}$*

$$|x|^r\,|f(x)|\le C_r||\hat f||_{b^r_{1\infty}}. \tag{A.17}$$

Proof. The uniform continuity of f follows from the fact that $\hat f\in L_1(\mathbb{R})$, since $B^r_{1\infty}(\mathbb{R})\subset L_1(\mathbb{R})$. Then, for a positive $\varepsilon\in(0,r-1/2)$ by (3.2) and (3.3),

$$B^r_{1\infty}(\mathbb{R})\hookrightarrow B^{r-1/2}_{2\infty}(\mathbb{R})\hookrightarrow B^{r-1/2-\varepsilon}_{22}(\mathbb{R}). \tag{A.18}$$

From this we get that $\hat f\in L_2(\mathbb{R})$, and hence $f\in L_2(\mathbb{R})$. Let $l=\lfloor r\rfloor$ be the integer part of r. From

$$\Delta_h^{(l+2)}\hat f(k)=\int_{-\infty}^{\infty}e^{-\mathrm{i}\,2\pi\,kx}(e^{-\mathrm{i}\,2\pi\,hx}-1)^{l+2}f(x)\,dx \tag{A.19}$$

and $f\in L_2(\mathbb{R})$, it follows by the inverse Fourier transform that

$$f(x)(e^{-\mathrm{i}\,2\pi\,hx}-1)^{l+2}=\int_{-\infty}^{\infty}e^{\mathrm{i}\,2\pi\,kx}\Delta_h^{(l+2)}\hat f(k)\,dk. \tag{A.20}$$

By taking the absolute values and dividing this equation by $|h|^r$ we then take the supremum over $h\in\mathbb{R}$. This yields

$$|x|^r\,|f(x)|C'_r\le\sup_{h\in\mathbb{R}}|h|^{-r}\,||\Delta_h^{(l+2)}\hat f||_{L_1},\quad C'_r=\sup_{t\in\mathbb{R}}\left\{\frac{|e^{-\mathrm{i}\,2\pi\,t}-1|^{l+2}}{|t|^r}\right\}.$$

Since the triple $(r, l+2, 0)$ is admissible, we have $\sup_{h\in\mathbb{R}} |h|^{-r} ||\Delta_h^{(l+2)} \hat{f}||_{L_1} \le C_r'' ||\hat{f}||_{b_{1\infty}^r}$ for some C_r'' depending only on r, which completes the proof of Lemma A.1. □

LEMMA A.2. *Assume that $\phi \in C^2(\mathbb{R})$ is chosen so that ϕ, $\phi' = D^1\phi$, $\phi'' = D^2\phi \in L_\infty(\mathbb{R})$. Then for $\varepsilon \in [0,1)$ and $Q \in B_{1\infty}^{1+\varepsilon}(\mathbb{R})$ the product $\phi \cdot Q \in B_{1\infty}^{1+\varepsilon}(\mathbb{R})$ and $||\phi Q||_{b_{1\infty}^{1+\varepsilon}} \le C||Q||_{B_{1\infty}^{1+\varepsilon}}$ for some constant $C = C(\varepsilon, \phi)$ depending only on ε and ϕ.*

Proof. From the obvious equalities

$$\phi(x+h) = \phi(x) + h\phi'(x) + h^2\phi''(x+\alpha\cdot h)/2, \quad \alpha \in [0,1],$$
$$\phi(x+2h) = \phi(x) + 2h\phi'(x) + 2h^2\phi''(x+\beta\cdot 2h), \quad \beta \in [0,1],$$

it follows that

$$\Delta_h^{(2)}(\phi Q)(x) = \phi(x)\cdot\Delta_h^{(2)}Q(x) + 2h\cdot\phi'(x)\Delta_h Q(x+h) + 2h^2\phi''(x+\beta\cdot 2h)Q(x+2h) - h^2\phi''(x+\alpha\cdot h)Q(x+h).$$

Therefore

$$||\Delta_h^{(2)}(\phi Q)||_{L_1} \le ||\phi||_{L_\infty}\cdot||\Delta_h^{(2)}Q||_{L_1} + 2|h|\cdot||\phi'||_{L_\infty}\cdot||\Delta_h Q||_{L_1} + 3h^2||\phi''||_{L_\infty}\cdot||Q||_{L_1}.$$

Since

$$||\Delta_h^{(2)}Q||_{L_1} \le ||Q||_{B_{1\infty}^{1+\varepsilon}}|h|^{1+\varepsilon}, \quad ||\Delta_h Q||_{L_1} \le ||Q||_{B_{1\infty}^{\varepsilon}}\cdot|h|^{\varepsilon} \le C(\varepsilon)||Q||_{B_{1\infty}^{1+\varepsilon}}\cdot|h|^{\varepsilon},$$

then for $|h| \le 1$ we have

$$||\Delta_h^{(2)}(\phi Q)||_{L_1} \le (||\phi||_{L_\infty} + 2C(\varepsilon)||\phi'||_{L_\infty} + 3||\phi''||_{L_\infty})\cdot||Q||_{B_{1\infty}^{1+\varepsilon}}|h|^{1+\varepsilon}.$$

This completes the proof. □

COROLLARY A.1. *For $r > 0$, let $\psi \in C^{\lfloor r\rfloor+1}$ be chosen so that*

$$\max_{n=0,1,\ldots,\lfloor r\rfloor+1} ||D^n\psi||_{L_\infty} < \infty.$$

Then there exists a constant $C = C(r,\psi)$ such that $||\psi F||_{b_{1\infty}^r} \le C||F||_{B_{1\infty}^r}$ for each $F \in B_{1\infty}^r(\mathbb{R})$.

Proof. Since $||\psi F||_{b_{1\infty}^r} = \sup_{h:\,|h|\le 1}\frac{1}{|h|^{1+\varepsilon}}\cdot||\Delta_h^{(2)}(\psi F)^{(l)}||_{L_1}$, where $l = \lfloor r\rfloor - 1$, and $\varepsilon = r - \lfloor r\rfloor$, it is sufficient to show that $||\Delta_h^{(2)}(\psi F)^{(l)}||_{L_1} \le C(r,\psi)||F||_{B_{1\infty}^r}\cdot|h|^{1+\varepsilon}$. Indeed, $\Delta_h^{(2)}(\psi F)^{(l)} = \sum_{n=0}^{l} C_l^n \Delta_h^{(2)}(\psi^{(n)}F^{(l-n)})$. For all n $(0 \le n \le l)$, the functions $\phi = \psi^{(n)}$ and $Q = F^{(l-n)}$ satisfy all of the conditions of Lemma A.2. Therefore, by taking into account that $||\Delta_h^{(2)}Q||_{B_{1\infty}^{1+\varepsilon}} \le c(r)h^{1+\varepsilon}||Q||_{B_{1\infty}^r}$ and using the result of Lemma A.2, one completes the proof of the corollary. □

COROLLARY A.2. *Let $r > 0$ and $\psi \in C^{\lfloor r\rfloor+1}$ be a function with a compact support $\Delta = \mathrm{supp}\{\psi\}$. Then there exists a constant $C = C(r,\psi)$ such that $||\psi f||_{B_{1\infty}^r(\mathbb{R})} \le C||f||_{B_{1\infty}^r(\Delta)}$ for each $f \in B_{1\infty}^r(\Delta)$.*

Proof. Indeed, let us denote by A the extension operator $A : B_{1\infty}^r(\Delta) \to B_{1\infty}^r(\mathbb{R})$, i.e., $(Af)(x) = f(x)$ for $x \in \Delta$ and $f \in B_{1\infty}^r(\Delta)$. The existence of a bounded linear extension operator is well known (e.g., see [27]). The corollary then follows from

$$||\psi f||_{B_{1\infty}^r(\mathbb{R})} = ||\psi\cdot Af||_{B_{1\infty}^r(\mathbb{R})} \le C(r,\psi)||Af||_{B_{1\infty}^r(\mathbb{R})} \le C(r,\psi)\cdot||A||\cdot||f||_{B_{1\infty}^r(\Delta)}.$$ □

Below, we give a result about characterization of the Besov space norms through wavelet coefficients.

For a nonnegative integer L we denote by $\mathcal{R}^L$ the class of $(L+1)$-times continuously differentiable functions $f : \mathbb{R} \to \mathbb{R}$ satisfying the following conditions:

$$\int_{\mathbb{R}} x^n f(x)dx = 0 \quad \text{for} \quad n = 0, 1, 2, \ldots, L;$$

$$\exists \gamma > 0,\, C > 0 \quad \text{such that} \quad |f(x)| \le \frac{C}{(1+|x|)^{2+L+\gamma}} \quad \forall x \in \mathbb{R};$$

$$\exists \varepsilon > 0,\, C_1 > 0 \quad \text{such that} \quad \max_{1 \le n \le L+1} |D^n f(x)| \le \frac{C}{(1+|x|)^{1+\varepsilon}} \quad \forall x \in \mathbb{R}. \tag{A.21}$$

It is known (e.g., see [24], [14], [38]) that, under the assumption $\psi \in \mathcal{R}^{\lfloor s \rfloor}$ on a wavelet function ψ, the norm $||f||_{B^s_{pq}}$ in Besov's space $B^s_{pq}(\mathbb{R})$, $1 \le p, q \le \infty$, $s > 0$, can be equivalently defined through the wavelet coefficients $\beta_{mj}(f) = \int_{\mathbb{R}} f(x)\, \psi_{mj}(x)dx$, m, $j \in \mathbb{Z}$, where $\psi_{mj}(x) = 2^{m/2}\psi(2^m\, x - j)$. Namely, let us introduce the norm $||\cdot||^{(\psi)}_{B^s_{pq}}$ by

$$||f||^{(\psi)}_{B^s_{pq}} = ||f||_{L_p} + \left\{ \sum_{m=-\infty}^{\infty} \left[\sum_{j=-\infty}^{\infty} \left(|\beta_{mj}(f)| \cdot 2^{m(s+\frac{1}{2}-\frac{1}{p})} \right)^p \right]^{q/p} \right\}^{1/q} \qquad (1 \le q < \infty),$$

$$||f||^{(\psi)}_{B^s_{p\infty}} = ||f||_{L_\infty} + \sup_{m \in \mathbb{Z}} \left[\sum_{j=-\infty}^{\infty} \left(|\beta_{mj}(f)| \cdot 2^{m(s+\frac{1}{2}-\frac{1}{p})} \right)^p \right]^{1/p}.$$

Then there exist constants $C_1 = C_1(p, q, s, \psi) > 0$ and $C_2 = C_2(p, q, s, \psi) > 0$ such that

$$C_1 \cdot ||f||^{(\psi)}_{B^s_{pq}} \le ||f||_{B^s_{pq}} \le C_2 \cdot ||f||^{(\psi)}_{B^s_{pq}} \qquad \forall f \in B^s_{pq}(\mathbb{R}). \tag{A.22}$$

As an immediate consequence of the last inequality we arrive at the following.

COROLLARY A.3. *For $s > 0$, let ψ be an orthonormal wavelet function satisfying the condition $\psi \in \mathcal{R}^{\lfloor s \rfloor}$. Then there exists a constant $C = C(p, q, s, \psi)$ such that for each $f \in B^s_{pq}(\mathbb{R})$*

$$\left\{ \sum_{j=-\infty}^{\infty} |\beta_{mj}(f)|^p \right\}^{1/p} \le C \cdot 2^{-m(s+\frac{1}{2}-\frac{1}{p})} \cdot ||f||_{B^s_{pq}}. \tag{A.23}$$

REFERENCES

[1] N. A. BUGLANOVA AND O. KURBANMURADOV, *Convergence of the randomized spectral models of homogeneous Gaussian random fields*, Monte Carlo Methods Appl., 1 (1995), pp. 173–201.

[2] C. CAMERON, *Relative efficiency of Gaussian stochastic process sampling procedures*, J. Comput. Phys., 192 (2003), pp. 546–569.

[3] A. COHEN, J. FROMENT, AND J. ISTAS, *Analyse multiresolution de signaux aleatories*, C.R. Acad. Sci. Paris Ser. I. Math., 312 (1991), pp. 567–570.

[4] A. COHEN AND J.-P. D'ALES, *Nonlinear approximation of random functions*, SIAM J. Appl. Math., 57 (1997), pp. 518–540.

[5] G. DAGAN, *Flow and Transport in Porous Formations*, Springer-Verlag, Berlin, 1989.

[6] I. DAUBECHIES, *Ten Lectures on Wavelets*, CBMS-NSF Regional Conf. Ser. Appl. Math. 61, SIAM, Philadelphia, 1992.

[7] G. Deodatis and M. Shinozuka, *Simulation of stochastic processes by spectral representation*, Appl. Mech. Rev., 44 (1991), pp. 191–204.
[8] C. R. Dietrich and G. N. Newsam, *Fast and exact simulation of stationary Gaussian processes through circulant embedding of the covariance matrix*, SIAM J. Sci. Comput., 18 (1997), pp. 1088–1107.
[9] X. Fernique, *Regularite des Trajectories des Functions Aleatoires Gaussiennes*, Lecture Notes in Math. 480, Springer-Verlag, Berlin, 1975.
[10] F. W. Elliott, Jr., D. J. Horntrop, and A. J. Majda, *A Fourier-wavelet Monte Carlo method for fractal random fields*, J. Comput. Phys., 132 (1997), pp. 384–408.
[11] F. W. Elliott, Jr., and A. J. Majda, *A new algorithm with plane waves and wavelets for random velocity fields with many spatial scales*, J. Comput. Phys., 117 (1995), pp. 146–162.
[12] I. I. Gikhman and A. V. Skorohod, *Introduction to the Theory of Random Processes*, W.B. Saunders, Philadelphia, 1969.
[13] E. Hernandez and G. Weiss, *A First Course on Wavelets*, CRC Press, Boca Raton, FL, 1996.
[14] W. Hardle, G. Kerkyacharian, D. Picard, and A. Tsybakov, *Wavelets, Approximation, and Statistical Applications*, Lecture Notes in Statist. 129, Springer-Verlag, New York, 1998.
[15] I. A. Ibragimov, *Properties of sample functions for stochastic processes and embedding theorems*, Theory Probab. Appl., 18 (1974), p. 442 (Russian version: Teor. Veroyatn. Primen., 18 (1973), pp. 468–480).
[16] V. S. Korolyuk, N. I. Portenko, A. V. Skorohod, and A. F. Turbin, *Handbook of Probability Theory and Mathematical Statistics*, Nauka, Moscow, 1985 (in Russian).
[17] P. R. Kramer, *A review of some Monte Carlo simulation methods for turbulent systems*, Monte Carlo Methods Appl., 7 (2001), pp. 229–244.
[18] P. Kramer, O. Kurbanmuradov, and K. Sabelfeld, *Comparative analysis of multiscale Gaussian random field simulation algorithms*, J. Comput. Phys., 226 (2007), pp. 897–924.
[19] O. Kurbanmuradov, *Weak convergens of approximate models of random fields*, Russian J. Numer. Anal. Math. Modelling, 10 (1995), pp. 500–517.
[20] O. Kurbanmuradov, *Weak convergence of randomized spectral models of Gaussian random vector fields*, Bull. Novosibirsk Computing Center, Numer. Anal., 4 (1993), pp. 19–25.
[21] O. Kurbanmuradov and K. Sabelfeld, *Stochastic spectral and Fourier-Wavelet methods for vector Gaussian random fields*, Monte Carlo Methods Appl., 12 (2006), pp. 395–446.
[22] O. Kurbanmuradov, K. Sabelfeld, and D. Koluhin, *Stochastic Lagrangian models for two-particle motion in turbulent flows. Numerical results*, Monte Carlo Methods Appl., 3 (1997), pp. 199–223.
[23] A. Majda, *Random shearing direction models for isotropic turbulent diffusion*, J. Stat. Phys., 75 (1994), p. 1153.
[24] Y. Meyer, *Ondelettes et Ppérateurs.* I: *Ondelettes*, Hermann, Paris, 1990 (English translation: *Wavelets and Operators*, Cambridge University Press, 1992).
[25] G. A. Mikhailov, *Approximate models of random processes and fields*, Russian J. Comp. Math. Math. Phys., 23 (1983), pp. 558–566 (in Russian).
[26] A. S. Monin and A. M. Yaglom, *Statistical Fluid Mechanics: Mechanics of Turbulence*, Vol. 2. MIT Press, Cambridge, MA, 1981.
[27] S. M. Nikolskiĭ, *Approximation of Function of Several Variables and Imbedding Theorems*, 2nd ed., Nauka, Moscow, 1977 (in Russian).
[28] K. K. Phoon, H. W. Huang, and S. T. Quek, *Comparison betveen Karhunen-Loeve and wavelet expansions for simulation of Gaussian processes*, Comput. & Structures, 82 (2004), pp. 985–991.
[29] L. Piterbarg and A. G. Ostrovskii, *Advection and Diffusion in Random Media. Implification for Sea Surface Temperature Anomalies*, Kluwer Academic, Norwell, MA, 1997.
[30] F. Poirion and C. Soize, *Numerical methods and mathematical aspects for simulation of homogenous and non homogenous Gaussian vector fields*, in Probabilistic Methods in Applied Physics, Lecture Notes in Phys. 451, P. Kree and W. Wedig, eds., Springer-Verlag, Berlin, 1995, pp. 17–53.
[31] K. K. Sabelfeld, *Monte Carlo Methods in Boundary Value Problems*, Springer-Verlag, New York, Berlin, 1991.
[32] M. K. Schneider and A. S. Willsky, *A Krylov subspace method for covariance approximation and simulation of random processes and fields*, Multidimens. Syst. Signal Process., 14 (2003), pp. 295–318.
[33] M. Shinozuka, *Simulation of multivariate and multidimensional random processes*, J. Acoust. Soc. Am., 49 (1971), pp. 357–368.

[34] P. Spanos and R. Ghanem, *Stochastic finite element expansion for random media*, J. Eng. Mech., ASCE, 115 (1989), pp. 1035–1053.
[35] P. D. Spanos, V. Ravi, and S. Rao, *Random field representation in a biorthogonal wavelet basis*, J. Eng. Mech., 127 (2001), pp. 194–205.
[36] E. C. Titchmarsh, *Introduction to the Theory of Fourier Integrals*, 2nd ed., Clarendon Press, Oxford, 1975.
[37] H. Triebel, *Theory of Function Spaces*. Birkhäuser, Basel-Boston-Stuttgart, Academische Verl. Ges., Leipzig, 1983.
[38] P. Wojtaszczyk, *A Mathematical Introduction to Wavelets*, London Math. Soc. Stud. Texts 37, Cambridge University Press, London, 1997.
[39] A. W. Woitishek and S. M. Prigarin, *Moment conditions on functional convergence of randomized numerical spectral models of homogeneous Gaussian fields*, Theory and Applications of Statistical Modeling, Computing Center of Siberian Branch of Acad. Sci. USSR, Novosibirsk, 1988, pp. 41–46 (in Russian).
[40] A. M. Yaglom, *Correlation theory of stationary and related random functions, Volume* I: *Basic results*, Springer-Verlag, Berlin, 1987.
[41] B. A. Zeldin and P. D. Spanos, *Random field representation and synthesis using wavelet basis*, ASME J. Appl. Mech., 63 (1996), pp. 946–952.

SIAM J. NUMER. ANAL.
Vol. 46, No. 6, pp. 3113–3135
© 2008 Society for Industrial and Applied Mathematics

OPERATOR SPLITTING FOR AN IMMUNOLOGY MODEL USING REACTION-DIFFUSION EQUATIONS WITH STOCHASTIC SOURCE TERMS*

TIMOTHY A. LUCAS†

Abstract. When immune cells detect foreign molecules, they secrete soluble factors that attract other immune cells to the site of the infection. In this paper, I study numerical solutions to a model of this behavior proposed by Kepler. In this model the soluble factors are governed by a system of reaction-diffusion equations with sources that are centered on the cells. The motion of the model cells is a Langevin process that is biased toward the gradient of the soluble factors. I have shown that the solution to this system exists for all time and remains positive, the supremum is a priori bounded, and the derivatives are bounded for finite time. I have also developed a first order split scheme for solving the reaction-diffusion stochastic system. This allows us to make use of known first order schemes for solving the diffusion, the reaction, and the stochastic differential equations separately.

Key words. operator splitting, reaction-diffusion, stochastic differential equations

AMS subject classifications. 65C30, 60H35, 65M12, 65N12, 92-08

DOI. 10.1137/070701595

1. Introduction.

The immune system in higher organisms is composed of individual cells called leukocytes which work together to combat pathogens such as bacteria and viruses. These cells are unique because they move throughout the body and signal to each other in order to organize a response to an infection. Upon detecting pathogens in the body, an immune cell will secrete cytokines which are proteins that mediate communication among leukocytes and coordinate their response. Some cytokines called chemokines direct the motion of immune cells along their concentration gradient by a process known as chemotaxis. Other cytokines, such as tumor necrosis factor (TNF), recruit cells to the site of inflammation and activate immune cells. One process by which inflamation is regulated is the shedding of cytokine receptors, such as those for TNF (sTNFr). Those soluble receptors bind to the respective cytokines and limit their activity. Based on these signals, the cells will group together to respond to an infection. The ability of these cells to organize spatially in a timely manner is vital to an effective immune response.

The following sections describe a mathematical model of soluble factors and immune cell motion proposed in [7] and [6]. The cytokines are modeled continuously as soluble factors governed by reaction-diffusion equations with source terms that are centered on activated leukocytes. The motion of the immune cells is inherently stochastic but is biased toward the chemical gradients of the chemokines. We will show that solutions to this system exist and describe a numerical simulation of the model based on operator splitting. Although operator splitting is a well-known technique for solving reaction-diffusion systems; see [4], there is an added degree of difficulty due to the stochastic source terms.

*Received by the editors August 30, 2007; accepted for publication (in revised form) May 6, 2008; published electronically September 4, 2008. This work was supported by the National Institute of Health through grants 1R21-A1058227-02 and HHSN266200500019C.

http://www.siam.org/journals/sinum/46-6/70159.html

†Mathematics Department, Pepperdine University, Malibu, CA 90263 (timothy.lucas@pepperdine.edu).

1.1. Soluble factors. Suppose $U_i(x,t)$ is the local concentration of soluble factor i at the position x and time t. Let x_μ be the position of the center of cell μ in the bounded domain $\Omega \in \mathbb{R}^3$. Then, for $x \in \Omega$, the reaction-diffusion equations that describe the behavior of the soluble factors from [6] are

$$\frac{\partial U_i}{\partial t}(x,t) = \left\{ D_i \Delta - \sum_{j=1}^{N} r_{ij} U_j(x,t) - \lambda_i \right\} U_i(x,t) + \sum_{\mu=1}^{M} J_{i\mu} g_\mu(x - x_\mu(t)). \tag{1.1}$$

Here D_i denotes the diffusion coefficient for soluble factor U_i, r_{ij} is the rate at which U_i is removed by interaction with U_j, and λ_i is the rate of the removal of U_i by other factors. The secretion of soluble factor U_i by cell μ is a smoothly cutoff Gaussian source centered at the cell position x_μ, with support over the volume of the cell. The source term in (1.1) is a sum over the M individual cells indexed by μ, with secretion rates $J_{i\mu}$. The constants D_i, r_{ij}, and $J_{i\mu}$ are all nonnegative.

In this paper we consider a smooth source term with compact support as opposed to the point sources in [6] for two reasons. Point sources do not take into account the volume of the cells which can vary in size and shape. Also, the point sources in [6] are delta functions, and we are currently unaware of any rigorous mathematical treatment of (1.1) with this type of source. We will make use of the smoothness of the source terms in order to show the existence of a solution. In the simulation, there is no qualitative difference between the use of smooth sources versus point sources.

1.2. Cell motion. The system (1.1) is coupled with M systems of stochastic differential equations that describe the motion of each individual cell. For each cell, indexed by μ, we let $x_t \in \mathbb{R}^3$ be the position and $v_t \in \mathbb{R}$ be the velocity. The motion of each cell is a Langevin process

$$dx_t = v_t dt, \quad dv_t = [h(U(x_t)) - \gamma v]\, dt + \sigma\sqrt{\gamma} dW_t, \tag{1.2}$$

where dW_t is standard Brownian motion in $\mathbb{R}^3$. The direction of motion is random, but it is biased toward the direction of the vector $h(U)$, which depends on the local concentration of one or more of the soluble factors by the relation

$$h(U) = \sum_{i=1}^{N} \frac{\chi_i \nabla U_i}{h_0 + |\nabla U_i|}. \tag{1.3}$$

Here χ_i is the chemotactic constant that controls how much the drift is influenced by the gradient of soluble factor U_i, and h_0 is the length of the gradient of U_i at which h attains half of its maximum value. The constants γ, h_0, and χ_i are nonnegative.

In this model we track the motion of each cell to better understand their individual behavior and collective immune response. This is a departure from the mathematical models of pattern formation which are based on the work of Turing in [18], e.g., Othmer and Scriven [14] or Palsson and Othmer [15]. Early models of chemotaxis, such as Keller and Segel in [5], describe the density of immune cells by reaction-diffusion equations but not the motion of individual cells.

1.3. Notation. Throughout we will refer to $C_f(x_1, x_2, \ldots, x_n)$ as a constant that depends on the function f and the values of $x_1, x_2, \ldots, x_n$. We also refer to $L_f(x_1, x_2, \ldots, x_n)$ as a Lipschitz constant for the function f that depends on $x_1, x_2, \ldots, x_n$. We define $B_r(x)$ to be a ball of radius r centered at $x \in \mathbb{R}^n$. For $a, b \in \mathbb{R}$ we define $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$.

Let X and Y be sets and suppose that $f_i : X \to Y$, $i = 1, \ldots, n$. Then $(f_1, \ldots, f_n)$ is an n-tuple whose components are in Y^X. We shall also let $(f_1, \ldots, f_n)$ denote the function from X to Y^n which assigns $(f_1(x), f_2(x), \ldots f_n(x))$ to $x \in X$, i.e.,

$$(f_1, f_2, \ldots, f_n)(x) = (f_1(x), f_2(x), \ldots f_n(x)).$$

1.4. Overview and main results. In the following sections we consider the case where Ω is a cube and the reaction-diffusion equations for the soluble factors have periodic boundary conditions. Section 2 contains a proof that the reaction-diffusion-stochastic system (1.1), (1.2) has a solution for a short time. The contraction defined in section 2 is used in the convergence proofs in the following sections. We consider a first order split scheme in section 3 to make use of known numerical schemes for solving the diffusion, reaction, and stochastic differential equations for the cell motion. In section 4, we carefully add up the local truncation error due to the splitting using a method similar to the convergence proof of higher order numerical schemes for stochastic differential equations by Milstein in [11]. We also show that the solution to the reaction-diffusion system with a nonnegative initial condition is positive. We find in section 5 that the the supremum of the solution to the reaction-diffusion system is a priori bounded, and that derivatives are bounded for finite time. In addition, solutions to the reaction–diffusion-stochastic system exist for all time. The analysis in sections 2–5 is carried out using a single immune cell. The proofs for multiple cells are similar, and section 6 contains a sample simulation with three chemokine sources and multiple cells.

2. Short time existence. The definitions and estimates that follow allow us to define a contraction on the appropriate spaces in order to prove existence of a solution to the model (1.1), (1.2) for a short time.

2.1. Preliminaries. Suppose that the domain Ω is an open cube in $\mathbb{R}^n$, with side length Λ. Let $\mathbb{Z}^n$ be the set of integer frequencies $\omega = (\omega_1, \ldots, \omega_n)$, and define $|\omega|_\infty = \max_{1 \le i \le n} |\omega_i|$. Let $A(k, n)$ be the set of multi-indices $\alpha = (\alpha_1, \alpha_2, \ldots \alpha_n)$, with weight $w(\alpha) = \sum_{i=1}^n |\alpha_i| \le k$. Let $\mathcal{E}$ be the vector space of smooth Λ-periodic complex valued functions on $\mathbb{R}^n$. For each $\omega \in \mathbb{Z}^n$, define $E_\omega(x) \in \mathcal{E}$ by

$$E_\omega(x) = \Lambda^{-\frac{n}{2}} e^{\frac{i2\omega \cdot x}{\Lambda}}, \quad x \in \Omega. \tag{2.1}$$

For a locally integrable function f, define the Fourier coefficient

$$c_\omega(f) = \int_\Omega f(x) E_\omega(x) dx. \tag{2.2}$$

For $s \in \mathbb{R}$ and a locally integrable function f, let $\|f\|_s \in [0, \infty]$ be such that

$$\|f\|_s^2 = \sum_{\omega \in \mathbb{Z}^n} (1 + |\omega|_\infty^2)^s |c_\omega(f)|^2. \tag{2.3}$$

DEFINITION 2.1. *For $s \ge 0$, define $\mathbb{H}_s$ to be the space of measurable, Λ-periodic, complex valued functions f on $\Omega \subset \mathbb{R}^n$ such that $\|f\|_s < \infty$.*

The following facts from [3] about Sobolev spaces are useful throughout.

Remark 2.2. If $k > \frac{n}{2}$, then corresponding to any element $f \in \mathbb{H}_k$ is a continuous function $g \in \mathbb{H}_k$ such that $f = g$ a.e. We may therefore assume when $k > \frac{n}{2}$ that

the elements of $\mathbb{H}_k$ are continuous. Moreover, the set $\{f \in \mathbb{H}_k : f \text{ is continuous}\}$ is a Hilbert space with respect to $\|\cdot\|_k$. For $g_\alpha = \partial_\alpha f$ in the sense of distributions,

$$\|f\|_k^2 = \sum_{w(\alpha)\le k} \|g_\alpha\|_0^2 . \tag{2.4}$$

LEMMA 2.3 (Sobolev's lemma). *Let $k > \frac{n}{2} + l$. Then*

$$\sup_{w(\alpha)\le l} |\partial^\alpha f(x)| \le C(k,l)\, \|f\|_k . \tag{2.5}$$

Remark 2.4. Note that $f \in \mathbb{H}_{k+l}$ if and only if $\partial^\alpha f \in \mathbb{H}_k$ for $w(\alpha) \le l$.

Define the linear map $\Delta : \mathbb{H}_4 \to \mathbb{H}_2$ by $\Delta f = \sum_{i=1}^n \partial_{x_i}^2 f$. Then, for $s \ge 2$,

$$\|\Delta f\|_s \le \|(1-\Delta)(f)\|_s = \|f\|_{s+2} . \tag{2.6}$$

Let $\mathbb{H}_k^N$ be the space of N-tuples $u = (u_1, u_2, \dots, u_N)$ such that $u_i \in \mathbb{H}_k$ for $i = 1, 2 \dots, N$, and set

$$\|u\|_k = \left(\sum_{i=1}^N \|u_i\|_k^2 \right)^{\frac{1}{2}} . \tag{2.7}$$

By our convention $(u_1, \dots, u_N)(x) = (u_1(x), \dots, u_N(x))$. Define $\mathcal{D} : \mathbb{H}_4^N \to \mathbb{H}_2^N$ by

$$\mathcal{D}u = (D_1 \Delta u_1, D_2 \Delta u_2, \dots, D_N \Delta u_N), \tag{2.8}$$

where $D_i > 0$ for $i = 1, 2 \dots, N$ as in (1.1).

We now define the fundamental solution of the diffusion equation. For $t > 0$, define the linear map $K_t : \mathbb{H}_k^N \to \mathcal{E}^N$ by requiring that

$$(K_t(u))_i = \sum_{\omega \in \mathbb{Z}^n} e^{-\frac{2D_i |w|^2 t}{\Lambda}} c_\alpha(u_i) E_\omega . \tag{2.9}$$

Note that

$$\frac{d}{dt} K_t(u) = \mathcal{D}(K_t(u)) = K_t(\mathcal{D}(u)). \tag{2.10}$$

It also follows that

$$\|K_t u\|_s \le \|u\|_s . \tag{2.11}$$

Below we define the space of solutions to the reaction-diffusion equations (1.1).

DEFINITION 2.5. *For any integer $k > \frac{n}{2}$, let $\mathbf{B}_k^N(m_1, m_2)$ be the set of N-tuples $u \in \mathbb{H}_k^N$ such that $\sup|u| \le m_1$ and $\|u\|_k \le m_2$. Note that*

$$\mathbb{H}_k^N = \bigcup_{0 \le m_1, m_2 < \infty} \mathbf{B}_k^N(m_1, m_2).$$

Remark 2.6. If $u \in \mathbf{B}_k^N(m_1, m_2)$ and $t > 0$, then $\sup|K_t(u)| \le \sup|u|$.

DEFINITION 2.7. *For any integer $k > \frac{n}{2}$, let $\mathbf{C}_{k,T}^N(m_1, m_2)$ be the set of functions $U : [0, T] \to \mathbf{B}_k^N(m_1, m_2)$ that are continuous with respect to $\|\cdot\|_k$. Then define*

$$\|U\|_{k,T} = \sup_{0 \le t \le T} \|U(t)\|_k . \tag{2.12}$$

Also, let

$$\mathbf{C}^N_{k,T} = \bigcup_{0\le m_1,m_2<\infty} \mathbf{C}^N_{k,T}(m_1,m_2).$$

Below we define the space of solutions to the stochastic differential equations (1.2).

DEFINITION 2.8. *Suppose $T>0$ and $m>0$. Define $\mathcal{S}_T(m)$ to be the space of continuous stochastic processes on $\mathbb{R}^n$ such that the norm*

$$\|X_t\|_{\mathcal{S}_T} = \left(\mathbb{E}\left[\sup_{0\le t\le T}|X_t|^2\right]\right)^{\frac{1}{2}} \le m. \tag{2.13}$$

We now define the space of functions which will contain the solutions to the reaction–diffusion-stochastic system (1.1), (1.2).

DEFINITION 2.9. *For any integer $k>\frac{n}{2}$, let $\mathcal{B}^N_k(m_1,m_2,m_3)$ be the space of ordered pairs $\phi=(u,\xi)$, where $u\in\mathbf{B}^N_k(m_1,m_2)$ and ξ is a random variable with values in $\mathbb{R}^n$ such that $\mathbb{E}\left[|\xi|^2\right]^{\frac{1}{2}}\le m_3$. Then, for $\psi=(v,\zeta)\in\mathcal{B}^N_k(m_1,m_2,m_3)$, define*

$$(\phi,\psi)_k = (u,v)_k + \xi\cdot\zeta, \tag{2.14}$$

$$|||\phi|||_k = \mathbb{E}\left[(\phi,\phi)_k\right]^{\frac{1}{2}} = \mathbb{E}\left[\|u\|^2_k+|\xi|^2\right]^{\frac{1}{2}}. \tag{2.15}$$

DEFINITION 2.10. *For any integer $k>\frac{n}{2}$ and $T>0$, let $\mathcal{C}^N_{k,T}(M_1,M_2,M_3)$ be the Cartesian product of $\mathbf{C}^N_{k,T}(M_1,M_2)$ and $\mathcal{S}_T(M_3)$. Define the norm $|||\cdot|||_{k,T}$ by*

$$|||(U,X)|||_{k,T} = \mathbb{E}\left[\sup_{0\le t\le T}\left(\|U(t)\|^2_k+|X_t|^2\right)\right]^{\frac{1}{2}}. \tag{2.16}$$

Since our space of solutions is periodic, we must define the Gaussian source terms to be periodic as well. Let $g_\mu:\mathbb{R}^n\to\mathbb{R}$ be a mean zero Gaussian, with standard deviation μ. Define $g_{\mu,i}$ to be the periodification of g_μ, i.e., $g_{\mu,i}(x)=\sum_{j\in\mathbb{Z}} g_\mu\left(x-\frac{2\pi j}{\Lambda}\right)$. Let $\bar{J}_i>0$ for $i=1,2,\dots,N$ and $J=\max\{\bar{J}_i : i=1,2,\dots,N\}$. Then let $\epsilon>0$ and $J_i:\mathbb{R}^n\to\mathbb{R}$ be a smooth cutoff defined by

$$J(X)=\begin{cases}\bar{J}_i & X\in\left(-\frac{\Lambda}{2}+\epsilon,\frac{\Lambda}{2}-\epsilon\right)^3,\\ 0 & X\notin\left(-\frac{\Lambda}{2},\frac{\Lambda}{2}\right)^3.\end{cases}$$

Then define $g_\mu:\mathbb{R}^n\times\mathbb{R}^n\to\mathbb{R}^N$ by $g_\mu(X,x)=(J_1(X)g_{\mu,1}(x),\dots,J_N(X)g_{\mu,N}(x))$.

Let $T>0$, $(U,X)\in\mathcal{C}^N_{k,T}$, and $\mathcal{T}_x$ be translation by $x\in\mathbb{R}^n$. Throughout we will consider the reaction–diffusion-stochastic system with one cell:

$$\begin{aligned}\frac{dU}{dt} &= \mathcal{D}U(t)+f(U)(t)+g(X_t)\circ\mathcal{T}_{X_t},\quad 0<t<T,\\ dX_t &= a(U(X_t),X_t)dt+b(X_t)dW_t,\quad 0<t<T,\\ (U,X)(0) &= (u,\xi),\end{aligned} \tag{2.17}$$

where $u\in\mathbb{H}^N_k$ and $\xi\in\mathbb{R}^n$. We will usually consider the case where $n=3$.

Remark 2.11. We choose to work with a single cell in order to make the proofs more clear. For multiple cells, the definition of g and the estimates that follow would include a sum over all cells indexed by μ.

Remark 2.12. We compose g with $\mathcal{T}_{X_t}$ so that the Gaussian source term is centered on the position of the cell. If the cell leaves the domain, then the cutoff J restricts the source term g from contributing to the reaction-diffusion equation. Since J_i is a function of X, g is still periodic in x.

2.2. Estimates for the reaction-diffusion equations. The following theorem provides estimates for the reaction terms $f(u)$.

THEOREM 2.13. *Let N be a positive integer and $f : \mathbb{R}^N \to \mathbb{R}^N$ be a smooth function such that $f(0) = 0$. Suppose that $0 \le m_1, m_2 < \infty$, k is a nonnegative integer, and $u, v \in \mathbf{B}_k^N(m_1, m_2)$. Then*

$$|f(u)| \le c_{f,1}(m_1)\,|u|\,, \tag{2.18}$$

$$\|f(u)\|_k \le C_{f,k}(m_1)\,\|u\|_k\,, \tag{2.19}$$

where

$$c_{f,k}(m_1) = \sup\left\{\left|\left(\frac{\partial}{\partial u}\right)^{\alpha} f(u)\right| : w(\alpha) \le k, u \in \mathbb{R}^N \text{ and } |u| \le m_1\right\}, \tag{2.20}$$

$$C_{f,k}(m_1) = C(k,N)c_{f,\max\{1,k\}}(m_1)(1+m_1)^{\max\{0,k-1\})}. \tag{2.21}$$

Let $k = 0$ or $k > \frac{n}{2}$. Then we also have

$$\|f(u) - f(v)\|_k \le L_{f,k}(m_1, m_2)\,\|u - v\|_k\,, \tag{2.22}$$

where

$$\begin{aligned} L_{f,0}(m_1) &= c_{f,1}(m_1), \\ L_{f,k}(m_1, m_2) &= C(k,N)c_{f,k+1}(m_1)(1+m_1)^{k-1}m_2. \end{aligned} \tag{2.23}$$

Proof. We can rewrite

$$\begin{aligned} f(u) - f(v) &= \int_0^1 \frac{d}{dt} f((1-t)v + tu)dt \\ &= \left(\int_0^1 \partial f\big((1-t)v + tu\big)\frac{d}{dt}((1-t)v + tu)dt\right) \\ &= \left(\int_0^1 \partial f\big((1-t)v + tu\big)dt\right)(u - v). \end{aligned}$$

It immediately follows from the Cauchy–Schwarz inequality that

$$\|f(u) - f(v)\|_0 \le c_{f,1}(m_1)\,\|u - v\|_0\,,$$

and, for $v = 0$, we have $|f(u)| \le c_{f,1}(m_1)\,|u|$ and $\|f(u)\|_0 \le c_{f,1}(m_1)\,\|u\|_0$.

Let k be a positive integer. Using well-known arguments found in [16, section 13.3], there exist constants $C(k,N)$ and $C(k)$ such that

$$\|f(u)\|_k \le C(k,N)c_{f,k}(m_1)(1+m_1)^{k-1}\,\|u\|_k\,, \tag{2.24}$$

$$\|u_1 u_2\|_k \le C(k)\left(\|u_1\|_k\,\|u_2\|_\infty + \|u_1\|_\infty\,\|u_2\|_k\right). \tag{2.25}$$

Then (2.19) immediately follows from (2.24).

Now consider $k > \frac{n}{2}$. Rewrite $f(u) - f(v) = u_1 u_2$, where $u_1 = \int_0^1 \partial f\big((1-t)v + tu\big)dt$ and $u_2 = u - v$. Then by (2.24),

$$\|u_1\|_k \le C(k+1, N) c_{f,k+1}(m_1)(1+m_1)^k (\|u\|_k \vee \|v\|_k). \tag{2.26}$$

The result (2.22) follows from substituting (2.20) and (2.26) into (2.25) and applying the Sobolev inequality. □

The following estimates on the source terms g can be easily obtained.

LEMMA 2.14. *For $k \ge 0$, there exist constants C_g, $C_{g,k}$, and $L_{g,k}$ such that*

$$|g| \le C_g, \quad \|g\|_k \le C_{g,k}, \quad \text{and} \quad \|g \circ \mathcal{T}_y - g \circ \mathcal{T}_z\|_k \le L_{g,k}|y - z|. \tag{2.27}$$

2.3. Estimates for the cell motion. Let $X_t = (x_t, v_t)$,

$$a(u, X_t) = \begin{pmatrix} v_t \\ h(u) - \gamma v_t \end{pmatrix}, \quad \text{and} \quad b(X_t) = \begin{pmatrix} 0 & 0 \\ 0 & \sigma\sqrt{\gamma} \end{pmatrix}. \tag{2.28}$$

It is immediately clear that b is Lipschitz in X and bounded. We also have

$$|a(u(X), X)| \le C_a(m), \quad X \in \mathcal{S}_T(m), \tag{2.29}$$

since h is bounded by $N(\max_{1\le i\le N} \chi_i)$. Recall from (1.3) that h depends on $\nabla u(X_t)$. Using the Sobolev lemma we can show that a is Lipschitz in both u and X.

LEMMA 2.15. *The drift $a(u(X), X)$ is Lipschitz in u and X, i.e., there exist constants $L_{a,1}(M_2)$ and $L_{a,2}(M_2)$ such that if $u, v \in \mathbf{B}_4^N(M_1, M_2)$, then*

$$|a(u(X), X) - a(v(Y), Y)| \le L_{a,1}(M_2)\,\|u - v\|_3 + L_{a,2}(M_2)\,|X - Y|\,. \tag{2.30}$$

The proof of this lemma depends on the fact that the supremum is bounded by the $\mathbb{H}_k$ norm for $k > \frac{3}{2}$ in three dimensions. Therefore, the supremum of the difference between u and $v \in \mathbf{B}_4^N$ is bounded by the $\mathbb{H}_3$ norm and a Lipschitz constant for ∇u is $\|u\|_4$. Also ∇u is Lipschitz in x, with a constant depending on $\|u\|_4$.

The facts below about Itô integrals from [10] and [13] are useful throughout.

PROPOSITION 2.16 (Itô isometry). *Let $X_t \in \mathcal{S}_T(m)$ and $0 \le t_1 \le t_2 \le T$. Then*

$$\mathbb{E}\left[\left|\int_{t_1}^{t_2} X_t dW_t\right|^2\right] = \mathbb{E}\left[\int_{t_1}^{t_2} |X_t|^2 dt\right]. \tag{2.31}$$

PROPOSITION 2.17 (Doob's martingale inequality for stochastic integrals). *Let $X_t \in \mathcal{S}_T(m)$. Then $\int_0^t X_\tau dW_\tau$ is a martingale, and*

$$\mathbb{E}\left[\sup_{0\le t\le T}\left|\int_0^t X_\tau dW_\tau\right|^p\right] \le \frac{p}{p-1}\mathbb{E}\left[\left|\int_0^T X_\tau dW_\tau\right|^p\right]. \tag{2.32}$$

2.4. Short-time existence. We now set up a contraction to show the existence of a solution to (2.17) for a short time. Define $F : \mathbf{C}_{k,T}^N \times \mathbb{R}^n \times \mathbb{H}_k^N \to \mathbf{C}_{k,T}^N$ by

$$F(U, X, u)(t) = K_t(u) + \int_0^t K_{t-\tau}(f(U) + g \circ \mathcal{T}_{X_\tau})d\tau, \tag{2.33}$$

and $G : \mathbf{C}_{k,t} \times \mathbb{R}^n \to \mathbb{R}^n$ by

$$G(U, X, \xi)(t) = \xi + \int_0^t a(U(X_\tau, \tau), X_\tau)d\tau + \int_0^t b(X_\tau)dW_\tau. \tag{2.34}$$

Then for $X, \xi \in \mathbb{R}^n$, $U \in \mathbf{C}^N_{k,T}$, and $u \in \mathbb{H}^N_k$, define

$$H(U,X,u,\xi)(t) = \big(F(U,X,u)(t), G(U,X,\xi)(t)\big). \tag{2.35}$$

To show H is a contraction, we first let $0 \le m_1 \le M_1 < \infty$, $0 \le m_2 \le M_2 < \infty$, and $0 \le m_3 \le M_3 < \infty$. For an integer $k \ge 4$, assume that $u \in \mathbf{B}^N_k(m_1, m_2)$, $|\xi| \le m_3$, and $(U,X), (V,Y) \in \mathcal{C}^N_{k,T}(M_1, M_2, M_3)$. Then by (2.20) and (2.27),

$$\begin{aligned}
\sup_{0\le t\le T} |F(U,X,u)| &= \sup_{0\le t\le T} \left| K_t(u) + \int_0^t K_{t-\tau}(f(U))d\tau + \int_0^t K_{t-\tau}(g)d\tau \right| \\
&\le m_1 + \sup_{0\le t\le T} \int_0^t |K_{t-\tau}(f(U))|\, d\tau + \sup_{0\le t\le T} \int_0^t |K_{t-\tau}(g)|\, d\tau \\
&\le m_1 + \int_0^T |f(U)|\, d\tau + \int_0^T |g|\, d\tau \le m_1 + T[c_{f,1}(M_1)M_1 + C_g].
\end{aligned}$$

Similarly by (2.21) and (2.27),

$$\|F(U,X,u)\|_{k,T} \le m_2 + T[C_{f,k}(M_1)M_2 + C_{g,k}].$$

Finally, by the Cauchy–Schwarz inequality and Doob's martingale inequality,

$$\begin{aligned}
\|G(U,X,\xi)\|^2_{\mathcal{S}_T} &= \mathbb{E}\left[\sup_{0\le t\le T} \left| \xi + \int_0^t a(U(X_s,s),X_s)ds + \int_0^t b(X_s)dW_s \right|^2 \right] \\
&\le 3|\xi|^2 + 3\mathbb{E}\left[\sup_{0\le t\le T} \left(\left| \int_0^t a(U(X_s,s),X_s)ds \right|^2 + \left| \int_0^t b(X_s)dW_s \right|^2 \right) \right] \\
&\le 3m_3^2 + 3\mathbb{E}\left[T\int_0^T |a(U(X_s,s),X_s)ds|^2 + 4\left| \int_0^t b(X_s)dW_s \right|^2 \right].
\end{aligned}$$

Then by Itô isometry,

$$\|G(U,X,\xi)(t)\|^2_{\mathcal{S}_T} \le 3(m_3 + C_a(M_3)T^2 + 4C_bT).$$

Therefore, we can choose T such that

$$\sup_{0\le t\le T} |F(U,X,u)| \le M_1, \quad \|F(U,X,u)\|_{k,T} \le M_2, \quad \text{and} \quad \|G(U,X,\xi)\|_{\mathcal{S}_T} \le M_3.$$

Thus, H takes $\mathcal{C}^N_{k,T}(M_1,M_2,M_3) \times \mathbf{B}^N_k(m_1,m_2) \times B_{m_3}(0)$ into $\mathcal{C}^N_{k,T}(M_1,M_2,M_3)$. Next we show that H is Lipschitz with respect to the norm $|||\cdot|||_{k,T}$. By (2.22),

$$\begin{aligned}
&\|F(U,X,u)(t) - F(V,Y,u)(t)\|_k \\
&\quad = \left\| \int_0^t K_{t-\tau}(f(U(\tau)) - f(V(\tau)))d\tau + \int_0^t (g\circ\mathcal{T}_{X_\tau} - g\circ\mathcal{T}_{Y_\tau})\, d\tau \right\|_k \\
&\quad \le \int_0^t \|f(U(\tau)) - f(V(\tau))\|_k\, d\tau + \int_0^t \|g\circ\mathcal{T}_{X_\tau} - g\circ\mathcal{T}_{Y_\tau}\|_k\, d\tau \\
&\quad \le L_{f,k}(M_1,M_2)\int_0^t \|U(\tau) - V(\tau)\|_k\, d\tau + L_{g,k}\int_0^t |X_\tau - Y_\tau|\, d\tau.
\end{aligned}$$

By the Cauchy–Schwarz inequality,

$$
\begin{aligned}
&\mathbb{E}\big[\|F(U,X,u)-F(V,Y,u)\|_{k,T}^2\big]\\
&\quad=\mathbb{E}\left[\sup_{0\le t\le T}\left(L_{f,k}(M_1,M_2)\int_0^t\|U(\tau)-V(\tau)\|_k\,d\tau+L_{g,k}\int_0^t|X_\tau-Y_\tau|\,d\tau\right)^2\right]\\
&\quad\le 2\,\mathbb{E}\left[L_{f,k}(M_1,M_2)^2\left(\int_0^T\|U(\tau)-V(\tau)\|_k\,d\tau\right)^2+L_{g,k}^2\left(\int_0^T|X_\tau-Y_\tau|\,d\tau\right)^2\right]\\
&\quad\le 2T\,\mathbb{E}\left[L_{f,k}(M_1,M_2)^2\int_0^T\|U(\tau)-V(\tau)\|_k^2\,d\tau+L_{g,k}^2\int_0^T|X_\tau-Y_\tau|^2\,d\tau\right]\\
&\quad\le 2T^2\,\mathbb{E}\left[L_{f,k}(M_1,M_2)^2\sup_{0\le t\le T}\|U(\tau)-V(\tau)\|_k^2+L_{g,k}^2\sup_{0\le t\le T}|X_\tau-Y_\tau|^2\right]\\
&\quad\le T^2L_{F,k}(M_1,M_2)^2\,|||(U,X)-(V,Y)|||_{k,T}^2\,,
\end{aligned}
$$

where $L_F(M_2)^2=2\max\{L_{f,k}^2(M_2),L_{g,k}\}$.

Define $\tilde a(U(t),X_t,V(t),Y_t)=a(U(X_t,t),X_t)-a(V(Y_t,t),Y_t)$, and $\tilde b(X_t,Y_t)=b(X_t)-b(Y_t)$. By Lemma 2.15, the Cauchy–Schwarz inequality, and the Doob martingale inequality,

$$
\begin{aligned}
\|G(U,X,\xi)(t)-G(V,Y,\xi)(t)\|_{\mathcal{S}_T}^2&=\mathbb{E}\left[\sup_{0\le t\le T}|G(U,X,\xi)(t)-G(V,Y,\xi)(t)|^2\right]\\
&=\mathbb{E}\left[\sup_{0\le t\le T}\left|\int_0^t\tilde a(U(\tau),X_\tau,V(\tau),Y_\tau)d\tau+\int_0^t\tilde b(X_\tau,Y_\tau)dW_\tau\right|^2\right]\\
&\le 2\mathbb{E}\left[\sup_{0\le t\le T}\left|\int_0^t\tilde a(U(\tau),X_\tau,V(\tau),Y_\tau)d\tau\right|^2+\sup_{0\le t\le T}\left|\int_0^t\tilde b(X_\tau,Y_\tau)dW_\tau\right|^2\right]\\
&\le 2\mathbb{E}\left[\sup_{0\le t\le T}t\int_0^t|\tilde a(U(\tau),X_\tau,V(\tau),Y_\tau)|^2\,d\tau+4\left|\int_0^T\tilde b(X_\tau,Y_\tau)dW_\tau\right|^2\right]\\
&\le 2\mathbb{E}\left[T\int_0^T|\tilde a(U(\tau),X_\tau,V(\tau),Y_\tau)|^2\,d\tau+4\int_0^T\left|\tilde b(X_\tau,Y_\tau)\right|^2d\tau\right]\\
&\le 2T\mathbb{E}\left[TL_{a,1}(M_2)^2\|U(t)-V(t)\|_k^2+\big(TL_{a,2}(M_2)^2+4L_b^2\big)\sup_{0\le t\le T}|X_t-Y_t|^2\right]\\
&\le TL_{G,k}(M_2,M_3)^2\,|||(U,X)-(V,Y)|||_{k,T}^2\,,
\end{aligned}
$$

where $L_{G,k}(M_2,M_3)^2=2\max\{TL_{a,1}(M_2),TL_{a,2}(M_2)^2+4L_b^2)\}$. Therefore,

$$
|||H(U,X,u,\xi)-H(V,Y,u,\xi)|||_{k,T}^2\le L_{H,k}^2\,|||(U,X)-(V,Y)|||_{k,T}^2\,, \tag{2.36}
$$

where $L_{H,k}^2=L_{H,k}(M_1,M_2,M_3)^2=T\max\{TL_{F,k}(M_1,M_2)^2,L_{G,k}(M_2,M_3)^2\}$. Thus, H is a contraction for $T<L_{H,k}(M_1,M_2,M_3)^{-1}$. Given $(U,X)(0)=(u,\xi)$, it follows that there is a unique $(U,X)\in\mathcal{C}_{k,T}^N(M_1,M_2,M_3)$ such that $H(U,X,u,\xi)=(U,X)$.

Remark 2.18. Let $k\ge 0$, $(U,X),(V,Y)\in\mathcal{C}_{4,T}^N(M_1,M_2,M_3)$, and $(u,\xi)\in\mathcal{B}_4^N(M_1,M_2,M_3)$. Using the same argument as above, we also have the estimate

$$
|||H(U,X,u,\xi)-H(V,Y,u,\xi)|||_{k,T}^2\le L_{H,k\vee 3}\,|||(U,X)-(V,Y)|||_{k\vee 3,T}\,. \tag{2.37}
$$

The next lemma is the basis of the error estimates that follow.

LEMMA 2.19. *Let k be a nonnegative integer, $(U,X),(V,Y) \in \mathcal{C}^N_{4,T}(M_1,M_2,M_3)$, and $H(U,X,u,\xi) = (U,X)$. Then*

$$|||(U,X)-(V,Y)|||_{k,T} \le \frac{1}{1-L_{H,k}} \, |||H(V,Y,u,\xi)-(V,Y)|||_{k\vee 3,T}\,, \tag{2.38}$$

provided that $L_{H,k} = L_{H,k}(M_1,M_2,M_3) < 1$.

Proof. Since $H(U,X,u,\xi) = (U,X)$,

$$\begin{aligned}
|||(U,X)-(V,Y)|||_{k,T} &= |||H(U,X,u,\xi) - H(V,Y,u,\xi) + H(V,Y,u,\xi) - (V,Y)|||_{k,T} \\
&\le |||H(U,X,u,\xi) - H(V,Y,u,\xi)|||_{k,T} + |||H(V,Y,u,\xi)-(V,Y)|||_{k,T} \\
&\le L_{H,k}(M_1,M_2,M_3)\, |||(U,X)-(V,Y)|||_{k\vee 3,T} + |||H(V,Y,u,\xi)-(V,Y)|||_{k\vee 3,T}\,,
\end{aligned}$$

from which (2.38) immediately follows. ∎

3. Operator splitting. In this section we will apply operator splitting to the reaction–diffusion-stochastic system (2.17). This will aid in both proving the existence of long-time solutions and solving the system numerically.

3.1. Taylor expansions. In order to analyze the error due to operator splitting, we introduce some notation. Suppose X and Y are normed vector spaces. Let f be a J times continuously differentiable function mapping an open convex subset A of X into Y. For each $j = 0,1,\dots,J$, we let $f^{[j]}$ be the jth differential of f. Thus, for each $a \in A$, $f^{(j)}(a)$ is a jth degree polynomial function on X whose value at $v \in X$ is

$$\left(\frac{d}{dt}\right)^j f(a+tv)|_{t=0}.$$

Let $f^{\{J\}}$ be the function on $A \times A$ whose value at $(a,b) \in A \times A$ is the Jth degree polynomial function on X whose value at $v \in X$ is

$$\int_0^1 f^{[j]}(a+\lambda(b-a)(v)d(1-\lambda)^N.$$

Taylor's theorem implies that

$$f(b) = \sum_{j=0}^{J-1} f^{[j]}(a)(b-a) + \frac{1}{J!} f^{\{J\}}(a,b)(b-a), \quad (a,b) \in A \times A.$$

Suppose B is an open subset of $X \times [0,\infty)$, with the property that if $b \in B$, then $\{t : (b,t) \in B\}$ is an interval containing 0. Let f be a J times continuously differentiable function mapping B into Y. For each $j = 0,1,\dots,J$ and $b \in B$, let

$$f^{(j)}(b) = \left(\frac{d}{dt}\right)^j f(b,t)\big|_{t=0} \quad \text{and} \quad f^{<J>}(b,t) = \int_0^1 \left(\frac{d}{d\lambda}\right)^J f(\lambda b) d(1-\lambda)^N.$$

Then Taylor's theorem implies that

$$f(b,t) = \sum_{j=0}^{J-1} \frac{t^j}{j!} f^{(j)}(b) + \frac{t^J}{J!} f^{<J>}(b,t).$$

3.2. Ordinary differential equations. Let $f : \mathbb{R}^N \to \mathbb{R}^N$ be smooth. Suppose R is the flow of f, i.e., the set of $((t, u_0), u_1) \in \mathbb{R} \times \mathbb{R}^N$ such that either $t = 0$ and $u_0 = u_1$ or $t \neq 0$ and there is a continuously differentiable function R carrying an open interval I, with $[0, t] \subset I$ into $\mathbb{R} \times \mathbb{R}^N$ such that

$$\frac{d}{d\tau} R(\tau) = f(R(\tau)) \quad \text{for } \tau \in I.$$

Note that R is a function; the domain of R is an open subset of $\mathbb{R} \times \mathbb{R}^N$ which contains $\{0\} \times \mathbb{R}^N$; and that R is smooth. It follows that

$$R(0, u) = u, \quad \frac{\partial}{\partial t} R(t, u) = f(R(t, u)) \quad \text{whenever } (t, u) \in \mathbb{R} \times \mathbb{R}^N. \tag{3.1}$$

It is easy to show that

$$R^{(0)}(u) = u, \quad R^{(1)}(u) = f(u), \quad R^{(2)}(u) = f^{[1]}(u)(f(u)), \quad u \in \mathbb{R}^N. \tag{3.2}$$

We shall abbreviate $R(t, u)$ by $R_t(u)$.

3.3. A first order split scheme. We can decouple the system (2.17) into three separate systems for the diffusion, reaction, and motion of the cells, respectively, i.e.,

$$\frac{dU}{dt} = \mathcal{D}(U) + g \circ \mathcal{T}_\xi, \qquad U(0) = u, \tag{3.3}$$

$$\frac{dU}{dt} = f(U), \qquad U(0) = u, \tag{3.4}$$

$$dX_t = a(u, X_t)dt + b(X_t)dW_t, \qquad X(0) = \xi. \tag{3.5}$$

Let F, G, and H be defined by (2.33), (2.34), and (2.35), respectively. Let K_t be defined by (2.9), and define

$$L_t(u, \xi) = K_t(u) + \int_0^t K_{t-\tau}(g \circ \mathcal{T}_\xi) d\tau. \tag{3.6}$$

It follows from (2.10) that $L_t(u, \xi)$ solves the inhomogeneous diffusion equation (3.3). Let R_t be the flow of (3.4) as defined by (3.1) and $Y_t(u, \xi)$ be the solution of (3.5) given by

$$Y_t(u, \xi) = \xi + \int_0^t a(u, Y_\tau) d\tau + \int_0^t b(Y_\tau) dW_\tau. \tag{3.7}$$

Then a split-step method to approximate (2.17) is given by the following three steps:

1. Solve the diffusion, (3.3), for one time step T using the initial data u and ξ.
2. Solve the reaction, (3.4), for one time step T using $L_t(u)$ and ξ as initial data.
3. Move the cells according to (3.5) for time T using $R_t(L_t(u))$ and ξ as initial data.

We express this split-step method as

$$\Psi_t(u, \xi) = \begin{pmatrix} R_t(L_t(u, \xi)) \\ Y_t(R_t(L_t(u, \xi)), \xi) \end{pmatrix}. \tag{3.8}$$

The next theorem shows that the local truncation error for the split scheme Ψ_t is $\mathcal{O}(t^{\frac{3}{2}})$. In section 4 we will show that the global error is $\mathcal{O}(t)$.

THEOREM 3.1. *Let (U,X) be the solution of (2.17) on $[0,T)$, where $(U,X)(0) = (u,\xi) \in \mathbb{H}_k^N \times \mathbb{R}^n$. Let the error due to the split scheme (3.8) be given by*

$$\mathbf{e}(t) = (U,X)(t) - \Psi_t(u,\xi).$$

Then, for a nonnegative integer k,

$$|||\mathbb{E}[\mathbf{e}]|||_{k,T} \le C(\sup|u|, \|u\|_{(k+2)\vee 3})T^2, \tag{3.9}$$

$$|||\mathbf{e}|||_{k,T} \le C(\sup|u|, \|u\|_{(k+2)\vee 3})T^{\frac{3}{2}}. \tag{3.10}$$

Proof. Using Lemma 2.19, we estimate the error for the splitting Ψ by estimating the difference between $H(\Psi_t(u,\xi),u,\xi)$ and $\Psi_t(u,\xi)$, i.e.,

$$\begin{aligned}|||\mathbf{e}(t)|||_{k,T} &\le |||H(\Psi_t(u,\xi),u,\xi) - \Psi_t(u,\xi)|||_{k\vee 3,T} \\ &= \mathbb{E}\left[\sup_{0\le t\le T}\left\{\|\mathbf{e}_1(t)\|_{k\vee 3}^2 + |\mathbf{e}_2(t)|^2\right\}\right]^{\frac{1}{2}},\end{aligned}$$

where

$$\begin{aligned}\mathbf{e}_1(t) &= F\left(\Psi_t\begin{pmatrix}u\\ \xi\end{pmatrix}, u\right) - R_t(L_t(u,\xi)), \\ \mathbf{e}_2(t) &= G\left(\Psi_t\begin{pmatrix}u\\ \xi\end{pmatrix}, \xi\right) - Y_t(R_t(L_t(u,\xi)),\xi).\end{aligned}$$

We immediately see from the definitions of G and Y_t that $\mathbf{e}_2(t) = 0$.

We shall now examine $\mathbf{e}_1(t)$. By the definition of F,

$$\begin{aligned}\mathbf{e}_1(t) &= K_t(u) + \int_0^t K_{t-\tau} f(R_\tau(L_\tau(u,\xi)))d\tau \\ &\quad + \int_0^t K_{t-\tau}\left(g\circ\mathcal{T}_{Y_\tau(R_\tau(L_\tau(u,\xi)),\xi)}\right)d\tau - R_t(L_t(u,\xi)).\end{aligned}$$

We proceed using Taylor expansions. A second order expansion of R_t is

$$\begin{aligned}R_t(L_t(u,\xi)) &= L_t(u,\xi) + tR_t^{(1)}(L_t(u,\xi)) + \frac{t^2}{2}R^{<2>}(t,L_t(u,\xi)) \\ &= K_t(u) + \int_0^t K_{t-\tau}\left(g\circ\mathcal{T}_\xi\right)d\tau + tf(L_t(u,\xi)) + \frac{t^2}{2}R^{<2>}(t,L_t(u,\xi)).\end{aligned}$$

Substituting the above expansion,

$$\mathbf{e}_1(t) = \mathbf{e}_{11}(t) + \mathbf{e}_{12}(t) - \frac{t^2}{2}R^{<2>}(t,L_t(u,\xi)),$$

where

$$\begin{aligned}\mathbf{e}_{11}(t) &= \int_0^t K_{t-\tau} f(R_\tau(L_\tau(u,\xi)))d\tau - tf(L_t(u,\xi)), \\ \mathbf{e}_{12}(t) &= \int_0^t K_{t-\tau}\left(g\circ\mathcal{T}_{Y_\tau(R_\tau(L_\tau(u,\xi)),\xi)}\right)d\tau - \int_0^t K_{t-\tau}\left(g\circ\mathcal{T}_\xi\right)d\tau.\end{aligned}$$

A first order Taylor expansion of f is given by

$$f(x+h) = f(x) + f^{\{1\}}(x, x+h)(h).$$

If we substitute a first order Taylor expansion of R_τ in $f(R_\tau(L_\tau(u,\xi)))$, then

$$f(R_\tau(L_\tau(u,\xi))) = f\big(L_\tau(u,\xi)\big) + f^{\{1\}}\big(L_\tau(u,\xi), R_\tau(L_\tau(u,\xi))\big)\big(\tau R^{<1>}(\tau, L_\tau(u,\xi))\big)$$

Thus,

$$\mathbf{e}_{11}(t) = \int_0^t K_{t-\tau} f(L_\tau(u,\xi))d\tau - tf(L_t(u,\xi)) + \mathcal{O}(t^2),$$

where the remainder term is

$$\int_0^t K_{t-\tau} f^{\{1\}}\big(L_\tau(u,\xi), R_\tau(L_\tau(u,\xi))\big)\big(\tau R^{<1>}(\tau, L_\tau(u,\xi))\big)d\tau.$$

Let $\eta(\tau) = K_{t-\tau} f(L_\tau(u,\xi))$. Then

$$\begin{aligned}\|\mathbf{e}_{11}(t)\|_k &= \left\| \int_0^t \eta(\tau)d\tau - t\eta(t) \right\|_k \le \int_0^t \|\eta(\tau) - \eta(t)\|_k \, d\tau \\ &\le \int_0^t \|\eta'(\tau)\|_k \, |t-\tau| d\tau = \sup_{0\le\tau\le t} \|\eta'(\tau)\|_k \, \frac{t^2}{2}.\end{aligned}$$

Then by Theorem 2.13 and Lemma 2.14,

$$\begin{aligned}\|\eta'(\tau)\|_k &= \left\| \frac{\partial}{\partial\tau} K_{t-\tau} f(L_\tau(u,\xi)) \right\|_k = \|\mathcal{D} K_{t-\tau} f(L_\tau(u,\xi))\|_k \le \|f(L_\tau(u,\xi))\|_{k+2} \\ &\le C_{f,k}(\sup|u|) \, \|L_\tau(u,\xi)\|_{k+2} \\ &\le C_{f,k}(\sup|u|) \left\| K_\tau(u) + \int_0^\tau K_{\tau-s} \, (g \circ \mathcal{T}_\xi) \, ds \right\|_{k+2} \\ &\le C_{f,k}(\sup|u|) \, \left(\|u\|_{k+2} + C_{g,k+2} \, \tau \right).\end{aligned}$$

Therefore,

$$\|\mathbf{e}_{11}\|_k \le C_{f,k}(\sup|u|, \|u\|_{k+2}) t^2. \tag{3.11}$$

To find the appropriate bounds on $\mathbf{e}_{12}$ consider a second order Taylor expansion of g;

$$g(x+h) - g(x) = g^{[1]}(x)(h) + \frac{1}{2} g^{\{2\}}(x, x+h)(h).$$

Using (3.7), we can rewrite

$$\begin{aligned}\mathbf{e}_{12}(t) &= \int_0^t K_{t-\tau} \left(g \circ \mathcal{T}_{Y_\tau(R_\tau(L_\tau(u,\xi)),\xi)} - g \circ \mathcal{T}_\xi \right) d\tau \\ &= \int_0^t K_{t-\tau} \left[g^{[1]}(x-\xi) \left(\int_0^\tau a(R_s(L_s(u,\xi)),\xi) ds + \int_0^\tau b(\xi) dW_s \right) \right] d\tau + \mathcal{O}(t^2),\end{aligned}$$

where the remainder term is

$$\frac{1}{2} \int_0^t K_{t-\tau} \left(g^{\{2\}}(x-\xi, x-Y_\tau) \, (Y_\tau - \xi) \right) d\tau.$$

Since the expectation of the Itô integral is zero,

$$\begin{aligned}\|\mathbb{E}\left[\mathbf{e}_{12}(t)\right]\|_k &\le \left\|\mathbb{E}\left[\int_0^t K_{t-\tau}\left(g^{[1]}(x-\xi)\left(\int_0^\tau a(R_s(L_s(u,\xi)),\xi)ds\right)\right)d\tau\right]\right\|_k \\ &\le \int_0^t \left\|\mathbb{E}\left[g^{[1]}(x-\xi)\left(\int_0^\tau a(R_s(L_s(u,\xi)),\xi)ds\right)\right]\right\|_k d\tau \\ &\le C_g C_a(\xi)t^2. \qquad (3.12)\end{aligned}$$

Recall from Lemma 2.14 that g is Lipschitz. It follows that

$$\begin{aligned}\|\mathbf{e}_{12}(t)\|_k &\le \int_0^t \left\|K_{t-\tau}\left(g\circ\mathcal{T}_{Y_\tau(R_\tau(L_\tau(u,\xi)),\xi)} - g\circ\mathcal{T}_{Y_\tau(u,\xi)}\right)\right\|_k d\tau \\ &\le \int_0^t \left\|g\circ\mathcal{T}_{Y_\tau(R_\tau(L_\tau(u,\xi)),\xi)} - g\circ\mathcal{T}_{\xi}\right\|_k d\tau \\ &\le L_{g,k}\int_0^t \left|Y_\tau(R_\tau(L_\tau(u,\xi)),\xi) - \xi\right| d\tau \\ &\le L_{g,k}\int_0^t\int_0^\tau \left|a(R_s(L_s(u,\xi)),\xi)\right| ds d\tau + L_{g,k}\int_0^t \left|\int_0^\tau b(\xi)dW_s\right| d\tau.\end{aligned}$$

By the Cauchy–Schwarz inequality, Doob's martingale inequality, and the Itô isometry,

$$\begin{aligned}\mathbb{E}\left[\sup_{0\le t\le T}\|\mathbf{e}_{12}(t)\|_k^2\right] &\le 2L_{g,k}\mathbb{E}\left[\sup_{0\le t\le T}\left(\int_0^t\int_0^\tau |a(R_s(L_s(u,\xi)),\xi)dsd\tau|\right)^2\right] \\ &\quad + 2L_{g,k}\mathbb{E}\left[\sup_{0\le t\le T}\left(\int_0^t\left|\int_0^\tau b(\xi)dW_s\right|d\tau\right)^2\right] \\ &\le 2L_{g,k}\left(\frac{1}{4}C_a^2(\xi)T^4 + 4T\mathbb{E}\left[\sup_{0\le t\le T}\int_0^t\int_0^\tau |b(\xi)|^2 dsd\tau\right]\right) \\ &\le 2L_{g,k}\left(\frac{1}{4}C_a^2(\xi)T^4 + 2C_b^2T^3\right). \qquad (3.13)\end{aligned}$$

Combining (3.11) and (3.12),

$$\sup_{0\le t\le T}\|\mathbb{E}\left[\mathbf{e}_1(t)\right]\|_k \le C(\sup|u|,\|u\|_{(k+2)})T^2. \qquad (3.14)$$

From (3.11) and (3.13) we have

$$\mathbb{E}\left[\sup_{0\le t\le T}\|\mathbf{e}_1(t)\|_k^2\right]^{\frac{1}{2}} \le C(\sup|u|,\|u\|_{(k+2)})T^{\frac{3}{2}}. \qquad (3.15)$$

Equations (3.9) and (3.10) follow from (3.14), (3.15), and the fact that $\mathbf{e}_2(t)=0$. ∎

4. Convergence. In this section we will show that the splitting outlined in section 3 is a first order method. These ideas can be extended to show the convergence rates of other numerical methods as well.

4.1. Preliminaries. Suppose $k \geq 4$, $t \in [0,T)$, and $\phi = (u,\xi) \in \mathcal{B}_k^N(m_1,m_2,m_3)$ for $0 \leq m_1, m_2, m_3 < \infty$. Let

$$\Phi(t,\phi,\delta) = (U(t,\phi,\delta), X(t,\phi,\delta)) \tag{4.1}$$

be the flow of the reaction–diffusion-stochastic system (2.17), i.e., the set of $((t,\phi,\tau),\eta)$ such that there exists a solution $\mathcal{U}$ of (2.17) whose domain is an interval containing t and $t+\tau$, $\mathcal{U}(t) = \phi$, and $\mathcal{U}(t+\tau) = \eta$. It is a consequence of uniqueness that Φ is a function. By local existence, for any $(t,\phi) \in [0,T) \times \mathcal{B}_k^N(m_1,m_2,m_3)$, there exists $\tau > 0$ such that $(t,\phi,\tau) \in \operatorname{domain}\Phi$. It follows from uniqueness that if δ_1,$\delta_2 > 0$, then

$$\Phi(t,\phi,\delta_1+\delta_2) = \Phi\big(t,\Phi(t,\phi,\delta_1),\delta_2\big). \tag{4.2}$$

Suppose $\Psi(t,\phi,\delta) = (V(t,\phi,\delta), Y(t,\phi,\delta))$ is a *one step method* for solving (2.17) such that $\Psi(t,\phi,0) = \phi$. Then the *local truncation error* for the method Ψ is given by

$$T(t,\phi,\delta) = \Phi(t,\phi,\delta) - \Psi(t,\phi,\delta) = (T_U(t,\phi,\delta), T_X(t,\phi,\delta)), \tag{4.3}$$

where

$$T_U(t,\phi,\delta) = U(t,\phi,\delta) - V(t,\phi,\delta) \ \text{ and } \ T_X(t,\phi,\delta) = X(t,\phi,\delta) - Y(t,\phi,\delta). \tag{4.4}$$

Suppose $t_0, T \in [0,\infty)$ such that $t_0 < T$. Let J be a positive integer such that $\delta = \frac{T}{J} < 1$. Then, for each $j = 1,\dots,J$, let

$$\begin{aligned}
&t_j = t_{j-1} + \delta, && \Phi_j = \Phi(t_0,\Phi_0,t_j - t_0), && \Psi_j = \Psi(t_{j-1},\Psi_{j-1},\delta),\\
&U_j = U(t_0,U_0,t_j - t_0), && X_j = X(t_0,X_0,t_j - t_0),\\
&V_j = V(t_{j-1},V_{j-1},\delta), && Y_j = Y(t_{j-1},Y_{j-1},\delta),\\
&e_j = \Phi_j - \Psi_j, && T_j = T(t_{j-1},\Psi_{j-1},\delta).
\end{aligned}$$

Finally, define

$$\Phi_{t,\phi}(x,\delta) = \Phi(t,\phi,\delta)(x) \quad \text{and} \quad \Psi_{t,\psi}(x,\delta) = \Psi(t,\psi,\delta)(x). \tag{4.5}$$

The next theorem, which gives the error due to the difference in initial conditions, is based on a lemma from [11].

THEOREM 4.1. *Let $Z_t = (Z_U, Z_X)$ be defined by*

$$\Phi(t,\phi,\delta) - \Phi(t,\psi,\delta) = \begin{pmatrix} K_\delta(u-v) \\ \xi - \zeta \end{pmatrix} + \begin{pmatrix} Z_U \\ Z_X \end{pmatrix}, \tag{4.6}$$

where $\phi = (u,\xi)$, $\psi = (v,\zeta)$. Then, for any integer $k \geq 4$, there exists a constant $C_{\Phi,k} = C(\sup|U|, \|U\|_k, \|X\|_{\mathcal{S}_T})$ which is independent of δ, for $\delta < 1$, such that

$$\|\!|\Phi_{t,\phi} - \Phi_{t,\psi}|\!\|_{k,\delta}^2 \leq \|\!|\phi - \psi|\!\|_k^2 \,(1 + C_{\Phi,k}\delta), \tag{4.7}$$

$$\|\!|Z_t|\!\|_{k,\delta}^2 \leq C_{\Phi,k}\, \|\!|\phi - \psi|\!\|_k^2 \,\delta. \tag{4.8}$$

Before we prove this theorem, we must first state several lemmas.

LEMMA 4.2. *Let $\alpha \in \mathbb{B}$, a Banach space with norm $\|\cdot\|$, and β be a function $\beta : [0,T) \to \mathbb{B}$ which is continuous with respect to the norm $\|\cdot\|$. Then*

$$\left\| \alpha + \int_t^{t+\delta} \beta(\tau) d\tau \right\|^2 \leq (1+\delta)\,\|\alpha\|^2 + (1+\delta)\int_t^{t+\delta} \|\beta(\tau)\|^2 \, d\tau. \tag{4.9}$$

The lemma above is proven by squaring the left-hand side and applying the Cauchy–Schwarz inequality. The next lemma from [11] is proven using induction.

LEMMA 4.3. *Let $A, B, \delta > 0$ and $p \geq 1$. If*

$$\mu_{j+1} \leq (1 + A\delta)\mu_j + B\delta^p, \tag{4.10}$$

then

$$\mu_j \leq e^{AT}\mu_0 + \frac{B}{A}\left(e^{AT} - 1\right)\delta^{p-1}. \tag{4.11}$$

Proof of Theorem 4.1. Consider the two components of $\Phi = (U, X)$ separately. In the context of Lemma 4.2, let $\alpha = K_\delta(u - v)$ and $\int_t^{t+\delta} \beta(\tau)d\tau = Z_U$. Then

$$\begin{aligned}
&\|U(t,\phi,\delta) - U(t,\psi,\delta)\|_k^2 \\
&\leq (1+\delta)\left\|K_\delta(u-v)\right\|_k^2 + 2(1+\delta)\int_t^{t+\delta} \left\|K_{t-\tau}\left[f(U(t,\phi,\tau)) - f(U(t,\psi,\tau))\right]\right\|_k^2 d\tau \\
&\quad + 2(1+\delta)\int_t^{t+\delta} \left\|K_{t-\tau}\left[g(X(t,\phi,\tau)) - g(X(t,\psi,\tau))\right]\right\|_k^2 d\tau \\
&\leq (1+\delta)\left\|u-v\right\|_k^2 + 2(1+\delta)L_{f,k}(\sup|U|, \|U\|_k)^2 \int_t^{t+\delta} \|U(t,\phi,\tau) - U(t,\psi,\tau)\|_k^2 d\tau \\
&\quad + 2(1+\delta)L_{g,k}^2 \int_t^{t+\delta} |X(t,\phi,\tau) - X(t,\psi,\tau)|^2 d\tau.
\end{aligned}$$

Define $\tilde{a}(\tau) = a(U(t,\phi,\tau), X(t,\phi,\tau)) - a(U(t,\psi,\tau), X(t,\psi,\tau))$ and $\tilde{b}(\tau) = b(X(t,\phi,\tau)) - b(X(t,\psi,\tau))$. Since the expected value of the Itô integral is zero,

$$\begin{aligned}
&\mathbb{E}\left[\sup_{0\leq s\leq\delta} |X(t,\phi,s) - X(t,\psi,s)|^2\right] \\
&\quad = \mathbb{E}\left[\sup_{0\leq s\leq\delta}\left|\xi - \zeta + \int_t^{t+s} \tilde{a}(\tau)d\tau\right|^2\right] + \mathbb{E}\left[\sup_{0\leq s\leq\delta}\left|\int_t^{t+s} \tilde{b}(\tau)dW_\tau\right|^2\right].
\end{aligned}$$

In the context of Lemma 4.2, let $\alpha = \xi - \zeta$ and $\beta = \tilde{a}$. Then

$$\begin{aligned}
&\mathbb{E}\left[\sup_{0\leq s\leq\delta} |X(t,\phi,s) - X(t,\psi,s)|^2\right] \\
&\quad \leq (1+\delta)\mathbb{E}\left[|\xi-\zeta|^2\right] + (1+\delta)\mathbb{E}\left[\int_t^{t+\delta} \left|\tilde{a}(\tau)\right|^2 d\tau\right] + 4\mathbb{E}\left[\int_t^{t+\delta} \left|\tilde{b}(\tau)\right|^2 d\tau\right] \\
&\quad \leq (1+\delta)\mathbb{E}\left[|\xi-\zeta|^2\right] \\
&\qquad + (1+\delta)L_{a,1}(\|U\|_k)^2 \int_t^{t+\delta} \mathbb{E}\left[\sup_{0\leq\sigma\leq\tau} \|U(t,\phi,\sigma) - U(t,\psi,\sigma)\|_k^2\right] d\tau \\
&\qquad + (1+\delta)L_{a,2}(\|U\|_k, \|X\|_{\mathcal{S}_T})^2 \int_t^{t+\delta} \mathbb{E}\left[\sup_{0\leq\sigma\leq\tau} |X(t,\phi,\sigma) - X(t,\psi,\sigma)|^2\right] d\tau \\
&\qquad + L_b^2 \int_t^{t+\delta} \mathbb{E}\left[\sup_{0\leq\sigma\leq\tau} |X(t,\phi,\sigma) - X(t,\psi,\sigma)|^2\right] d\tau
\end{aligned} \tag{4.12}$$

by the Doob martingale inequality and Itô isometry. Therefore,

$$\begin{aligned}\left|\left|\left|\Phi(t,\phi,s)-\Phi(t,\psi,s)\right|\right|\right|_{k,\delta}^2 &\le (1+\delta)\left|\left|\left|\phi-\psi\right|\right|\right|_k^2 \\ &+ C_{\Phi,k}(\sup|U|,\|U\|_k,\|X\|_{\mathcal{S}_T})\int_t^{t+\delta}\left|\left|\left|\Phi(t,\phi,\tau)-\Phi(t,\psi,\tau)\right|\right|\right|_{k,\delta}^2 d\tau,\end{aligned} \tag{4.13}$$

where

$$\begin{aligned}C_{\Phi,k}(\sup|U|,\|U\|_k,\|X\|_{\mathcal{S}_T}) = \max\{&(1+\delta)\delta(2L_{f,k}(\sup|U|,\|U\|_k)^2+L_{a,1}(\|U\|_k)^2),\\ &(1+\delta)(2L_{g,k}^2+L_{a,2}(\|U\|_k,\|X\|_{\mathcal{S}_T})^2)+L_b^2\}.\end{aligned}$$

Then by the Gronwall inequality,

$$\left|\left|\left|\Phi_{t,\phi}-\Phi_{t,\psi}\right|\right|\right|_{k,\delta}^2 \le \left|\left|\left|\phi-\psi\right|\right|\right|_k^2(1+\delta)e^{C_{\Phi,k}\delta}. \tag{4.14}$$

Equation (4.7) immediately follows, and (4.8) is the result of subtracting $\mathbb{E}\left[|\xi-\zeta|^2\right]$ from (4.12) and (4.14). ∎

Remark 4.4. The proof of this lemma can be modified to yield the same estimate with the $\left|\left|\left|\cdot\right|\right|\right|_k$ norm instead of the $\left|\left|\left|\cdot\right|\right|\right|_{k,\delta}$ norm.

4.2. Positivity and local bounds. It is well known; see [2], that if the initial condition is nonnegative, i.e., $u_i \ge 0$ for $i = 1, 2, \ldots, N$, the solution to the diffusion (3.3) is positive, i.e., for $t > 0$, $U_i(t) > 0$ for all i. The following lemma shows that if the initial condition is positive, the solution to the reaction (3.4) is also positive.

LEMMA 4.5. *Let I be an open interval. Suppose $U : I \to \mathbb{R}^n$ satisfies* (3.4) *and for some $t_0 \in I$, $U_i(t_0) > 0$ for all i. Then, for all $t \in I$, $U_i(t) > 0$ for all i.*

Proof. Note that if $u_i = 0$ for some i, then there exists a solution of (3.4) such that $U_i(t) = 0$ for all $t > 0$. Let $t_0 \in I$ and $U : I \to \mathbb{R}^n$ be a trajectory for (3.4) such that $U_i(t_0) = 0$ for some i. Then $f(U(t_0)) = 0$, and it follows by uniqueness that $U_i(t) = 0$ for all $t \in I$. Thus, if $U_i(t_0) > 0$ for all i, for all $t \in I$, $U_i(t) > 0$ for all i. ∎

Since the split-step method for the reaction-diffusion equation converges to the true solution, the solution to the reaction-diffusion equation is positive if the initial condition is nonnegative.

Before estimating the global error of the split-step method, we must show that the splitting $\Psi_t(u,\xi)$ from (3.8) is bounded in the appropriate norms. Define L_t as in (3.6), and let R_t be the flow of (3.4) where the reaction terms are given by

$$f(U) = -\lambda U - q(U), \tag{4.15}$$

where $\lambda = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_N)$ and $q = (q_1, q_2, \ldots, q_N)$ is given by

$$q_i(U) = \left(\textstyle\sum_{j\ne i} r_{ij}U_j\right)U_i. \tag{4.16}$$

Recall that $\lambda_i \ge 0$ and $r_{ij} \ge 0$ for all i and j. Define $\lambda_{\min} = \min\{\lambda_i : i = 1, 2, \ldots, N\}$.

LEMMA 4.6. *Suppose k is a nonnegative integer, $t \in [0, T)$, and $u \in \mathbb{H}_k^N$ such that $u_i \ge 0$ for $i = 1, 2, \ldots N$. Then*

$$|R_t(L_t(u,\xi))| \le (\sup|u| + C_g t)e^{-\lambda_{\min}t}, \tag{4.17}$$

$$\|R_t(L_t(u,\xi))\|_k^2 \le (\|u\|_k^2 + C_{g,k}t)e^{C_{f,k}(\sup|u|)t}. \tag{4.18}$$

Proof. By Lemma 2.14,

$$|L_t(u,\xi)| \le |K_t(u)| + \int_0^t |K_{t-\tau}(g \circ \mathcal{T}_\xi)|\, d\tau \le \sup|u| + C_g T, \tag{4.19}$$

$$\text{and} \quad \|L_t(u,\xi)\|_k \le \|K_t(u)\|_k + \int_0^t \|K_{t-\tau}(g \circ \mathcal{T}_\xi)\|_k\, d\tau \le \|u\|_k + C_{g,k} T. \tag{4.20}$$

Assume that $u_i > 0$ for all i. It follows from Lemma 4.5 that, for $t > 0$, $R_t(u_i) > 0$ and $q_i(R_t(u)) > 0$ for all i. Thus,

$$\frac{d}{dt} R_t(u) \le -\lambda R_t(u) \quad \text{and} \quad R_t(u) \le u e^{-t\lambda}.$$

Recall that, if $u_i \ge 0$ for all i, the diffusion $L_t(u_i) > 0$ for all $t > 0$. Then by (4.19),

$$|R_t(L_t(u))| \le |L_t(u)| e^{-\lambda_{\min} t} \le (\sup|u| + C_g t) e^{-\lambda_{\min} t}.$$

From Theorem 2.13 we have

$$\begin{aligned} \frac{1}{2}\frac{d}{dt}\|\partial^\alpha R_t(u)\|_0^2 &= \left(\partial^\alpha \left(\frac{d}{dt} R_t(u)(t)\right), \partial^\alpha R_t(u)\right) = (\partial^\alpha f(R_t(u)), \partial^\alpha R_t(u)) \\ &\le \|\partial^\alpha f(R_t(u))\|_0 \, \|\partial^\alpha R_t(u)\|_0 . \end{aligned}$$

Summing over $|\alpha| \le k$,

$$\frac{1}{2}\frac{d}{dt}\|R_t(u)\|_k^2 \le C_{f,k}(\sup|u|)\, \|R_t(u)\|_k^2 .$$

The solution of the differential equation above is

$$\|R_t(u)\|_k^2 \le \|u\|_k^2 \, e^{C_{f,k}(\sup|u|)t}. \tag{4.21}$$

Equation (4.18) follows from substituting $L_t(u)$ for u and applying (4.20). ∎

COROLLARY 4.7. *Let $j = 1, 2, \ldots, J$ and C_1 and C_2 be constants that are independent of δ for $\delta < 1$. Under the assumptions of Lemma* 4.6*,*

$$|V_j| \le \sup|u| e^{-\lambda_{\min} T} + C_1, \tag{4.22}$$

$$\|V_j\|_k^2 \le (\|u\|_k + C_2) e^{C_{f,k}(\sup|u|)T}. \tag{4.23}$$

Proof. Equations (4.22) and (4.23) follow from repeatedly applying (4.17) and (4.18), respectively. ∎

LEMMA 4.8. *If $Y_t(u,\xi)$ is a solution to* (3.5)*, then*

$$\mathbb{E}\left[\sup_{0\le t\le T} |Y_t(u,\xi)|^2\right] \le C_X(\xi, T). \tag{4.24}$$

Proof. Using an integrating factor $e^{\gamma t}$, the solution to the velocity equation in (1.2) is

$$v_t = e^{-\gamma t}\nu + \int_0^t e^{-\gamma(t-\tau)} h(U(x_t)) d\tau + \sigma\sqrt{\gamma}\int_0^t e^{\gamma(t-\tau)} dW_\tau.$$

By the Cauchy–Schwarz inequality, Doob's martingale inequality, and Itô isometry,

$$\begin{aligned}\mathbb{E}\left[\sup_{0\le t\le T}|v_t|^2\right] &\le 3e^{-2\gamma T}\nu + 3\mathbb{E}\left[\sup_{0\le t\le T}\left|\int_0^t e^{-\gamma(t-\tau)}h(U(x_t))d\tau\right|^2\right]\\ &\quad + 3\sigma^2\gamma\mathbb{E}\left[\sup_{0\le t\le T}\left|\int_0^t e^{-\gamma(t-\tau)}dW_\tau\right|^2\right]\\ &\le 3e^{-2\gamma T}\nu + 3T\mathbb{E}\left[\sup_{0\le t\le T}\int_0^t e^{-2\gamma(t-\tau)}|h(U(x_t))|^2d\tau\right]\\ &\quad + 12\sigma^2\gamma\mathbb{E}\left[\sup_{0\le t\le T}\left|\int_0^t e^{-\gamma(t-\tau)}d\tau\right|^2\right]\\ &\le 3e^{-2\gamma T}\nu + \frac{3T}{2\gamma}\left(1-e^{-2\gamma T}\right) + 6\sigma\left(1-e^{-2\gamma T}\right).\end{aligned}$$

The bound on Y_t (4.24) immediately follows. ∎

4.3. Error estimates for numerical methods. The following theorem gives the convergence rate of a general numerical scheme to approximate (2.17). The proof builds on an argument by Milstein in [11].

THEOREM 4.9. *Let $p_2 \ge \frac{1}{2}$ and $p_1 \ge p_2 + \frac{1}{2}$ and $k \ge 3$ be an integer. Assume that there exist constants $C_1, C_2 > 0$ such that*

$$\left|\!\left|\!\left|\mathbb{E}\left[\Phi(t,\xi,\delta) - \Psi(t,\xi,\delta)\right]\right|\!\right|\!\right|_k \le C_1\delta^{p_1}, \tag{4.25}$$

$$\left|\!\left|\!\left|\Phi(t,\xi,\delta) - \Psi(t,\xi,\delta)\right|\!\right|\!\right|_k \le C_2\delta^{p_2}. \tag{4.26}$$

Then, for $j = 0, 1 \ldots, J$, there exists a constant C such that

$$\left|\!\left|\!\left|\Phi(t_0,\Phi_0,t_j) - \Psi(t_0,\Phi_0,t_j)\right|\!\right|\!\right|_k \le C\delta^{p_2-\frac{1}{2}}. \tag{4.27}$$

Proof. To make use of Theorem 4.1, (4.25), and (4.26), we rewrite $\Phi_{j+1} - \Psi_{j+1}$ as the sum of the difference in initial conditions and the local trucnation error, i.e.,

$$\begin{aligned}\Phi_{j+1} - \Psi_{j+1} &= \Phi(t_j,\Phi_j,\delta) - \Psi(t_j,\Psi_j,\delta)\\ &= \Phi(t_j,\Phi_j,\delta) - \Phi(t_j,\Psi_j,\delta) + T(t_j,\Psi_j,\delta).\end{aligned}$$

Then

$$\left|\!\left|\!\left|\Phi_{j+1} - \Psi_{j+1}\right|\!\right|\!\right|_k^2 = A_1 + 2A_2 + A_3, \tag{4.28}$$

where

$$\begin{aligned}&A_1 = \left|\!\left|\!\left|\Phi(t_j,\Phi_j,\delta) - \Phi(t_j,\Psi_j,\delta)\right|\!\right|\!\right|_k^2,\\ &A_2 = \mathbb{E}[(\Phi(t_j,\Phi_j,\delta) - \Phi(t_j,\Psi_j,\delta), T(t_j,\Psi_j,\delta))_k], \quad A_3 = \left|\!\left|\!\left|T(t_j,\Psi_j,\delta)\right|\!\right|\!\right|_k^2.\end{aligned}$$

By Theorem 4.1 and (4.26), we have

$$A_1 \le \left|\!\left|\!\left|\Phi_j - \Psi_j\right|\!\right|\!\right|_k^2 (1 + C_{\Phi,k}\delta) = \left|\!\left|\!\left|e_j\right|\!\right|\!\right|_k^2 (1 + C_{\Phi,k}\delta), \tag{4.29}$$

$$A_3 \le C_2\delta^{2p_2}, \tag{4.30}$$

respectively. Using Theorem 4.1, we write $A_2 = B_1 + B_2$, where

$$B_1 = \mathbb{E}\left[\left(\binom{K_\delta(U_j - V_j)}{X_j - Y_j}, T(t_j, \Psi_j\delta)\right)\right] \text{ and } B_2 = \mathbb{E}\left[(Z_\delta, T(t_j, \Psi_j\delta))\right].$$

Then by the Cauchy–Schwarz inequality and condition (4.25),

$$\begin{aligned} B_1 &= \mathbb{E}\left[\mathbb{E}\left[\left(\binom{K_\delta(U_j - V_j)}{X_j - Y_j}, T(t_j, \Psi_j\delta)\right)\middle|\mathcal{F}_j\right]\right] = \mathbb{E}\left[\left(\binom{K_\delta(U_j - V_j)}{X_j - Y_j}, \mathbb{E}\left[T(t_j, \Psi_j\delta)|\mathcal{F}_j\right]\right)\right] \\ &\le \left|\left|\left|\binom{K_\delta(U_j - V_j)}{X_j - Y_j}\right|\right|\right|_k |||\mathbb{E}\left[T(t_j, \Psi_j\delta)|\mathcal{F}_j\right]|||_k \le |||\Phi_j - \Psi_j|||_k \, |||\mathbb{E}\left[T(t_j, \Psi_j\delta)|\mathcal{F}_j\right]|||_k \\ &\le |||e_j|||_k \left(\sqrt{2}C_1\delta^{p_1}\right). \end{aligned}$$

Using the Cauchy–Schwarz inequality, (4.8), and (4.26), we estimate the second term

$$B_2 \le |||Z_\delta|||_k \, |||T(t_j, \Psi_j, \delta)|||_k \le C_{\Phi,k} \, |||e_j|||_k \, \delta^{\frac{1}{2}}(C_2\delta^{p_2}).$$

Since $p_1 \ge p_2 + 1$, $\delta^{p_1} \le \delta^{p_2+\frac{1}{2}}$ for $\delta < 1$. Thus,

$$A_2 \le C \, |||e_j|||_k \, \delta^{p_2+\frac{1}{2}} \le \frac{1}{2}\delta \, |||e_j|||_k^2 + \frac{1}{2}C^2\delta^{2p_2}. \tag{4.31}$$

Putting the estimates for (4.29), (4.30), and (4.31) together,

$$|||e_{j+1}|||_k^2 \le |||e_j|||_k^2 \left(1 + C_{\Phi,k}\delta\right) + C\delta^{2p_2}. \tag{4.32}$$

The main result (4.27) follows from Lemma 4.3, with $e_0 = 0$. ∎

Remark 4.10. The split scheme Ψ from section 3 satisfies Theorem (4.9), with $p_1 = 2$ and $p_2 = \frac{3}{2}$ and is therefore a first order method.

5. Long-time existence. In the previous section we showed that the splitting outlined in section 3 converges to the true solution, and that the reaction-diffusion equations preserve positivity. In this section we will show that the supremum of the solution is a priori bounded, and its derivatives are bounded for finite time. Assuming the solutions to (2.17) exist for $t \in [0, T)$, the limit as t approaches T also exists, allowing us to solve (2.17) for further time. We begin by showing the solution to (2.17) is bounded in the appropriate norms for $t \in [0, T)$.

THEOREM 5.1. *Let $\Phi(t) = (U(t), X_t)$ be a solution of the reaction–diffusion-stochastic system* (2.17) *on $[0, T)$. Then, for any nonnegative integer k, there exist constants C_U and $C_{U,k}$ such that*

$$\sup|U(t)| \le C_U(\sup|u|), \tag{5.1}$$

$$\|U(t)\|_k \le C_{U,k}(\sup|u|, \|u\|_k, T). \tag{5.2}$$

Proof. Using (4.15) and (4.16), we can rewrite the first equation in (2.17) as

$$\frac{dU}{dt} = (\mathcal{D} - \lambda)U(t) - q(U(t)) + g \circ \mathcal{T}_{X_t}, \quad U(0) = u, \tag{5.3}$$

where we assume that $u \ge 0$. Define $\mathcal{K}_t = e^{-t\lambda}K_t$. Then

$$\frac{d}{dt}\mathcal{K}_t(u) = e^{-t\lambda}\left(\frac{d}{dt}K_t(u) - \lambda K_t(u)\right) = e^{-t\lambda}(\mathcal{D}K_t(u) - \lambda K_t(u)) = \mathcal{K}_t[(\mathcal{D} - \lambda)u].$$

Therefore,

$$U(t) = \mathcal{K}_t(u) + \int_0^t \mathcal{K}_{t-\tau}(-q(U) + g \circ \mathcal{T}_{X_\tau})d\tau. \tag{5.4}$$

Clearly $q_i(U) > 0$ for all i, and thus

$$U(t) \le \mathcal{K}_t(u) + \int_0^t \mathcal{K}_{t-\tau}(g \circ \mathcal{T}_{X_\tau})d\tau.$$

Then (5.1) follows from the fact that

$$|U(t)| \le |\mathcal{K}_t u| + \int_0^t e^{-(t-\tau)\lambda}|K_t(g \circ \mathcal{T}_{X_\tau})|d\tau \le |u| + \frac{1}{\lambda}C_g.$$

Let $u, v \in \mathbb{H}_k^N$, and $|\alpha| \le k$ for a nonnegative integer k. Then,

$$(\partial^\alpha u, \partial^\alpha v) = \sum_{i=1}^N (\partial^\alpha u_i, \partial^\alpha v_i) \le \sum_{i=1}^N \frac{\|\partial^\alpha u_i\|_k^2 + \|\partial^\alpha v_i\|_k^2}{2} = \frac{\|\partial^\alpha u\|_k^2 + \|\partial^\alpha v\|_k^2}{2}.$$

Thus, using Theorem 2.13 and Lemma 2.14,

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\|\partial^\alpha U(t)\|_0^2 &= \left(\partial^\alpha \dot{U}(t), \partial^\alpha U(t)\right) \\ &= (\mathcal{D}\partial^\alpha U(t), \partial^\alpha U(t)) + (\partial^\alpha f(U(t)), \partial^\alpha U(t)) + (\partial^\alpha g, \partial^\alpha U(t)) \\ &\le \|\partial^\alpha f(U(t))\|_0 \|\partial^\alpha U(t)\|_0 + \frac{1}{2}\left(\|\partial^\alpha g\|_0^2 + \|\partial^\alpha U(t)\|_0^2\right).\end{aligned}$$

Summing over $|\alpha| \le k$,

$$\frac{1}{2}\frac{d}{dt}\|U(t)\|_k^2 \le C_{f,k}(C_U(\sup|u|))\|U(t)\|_k^2 + \frac{1}{2}C_{g,k}^2.$$

Then (5.2) follows from the fact that the solution of the differential equation above is

$$\|U(t)\|_k^2 \le \left(\|u\|_k^2 + \frac{C_{g,k}^2}{C_{f,k}(C_U(\sup|u|))}\right)e^{C_{f,k}(C_U(\sup|u|))t} - \frac{C_{g,k}^2}{C_{f,k}(C_U(\sup|u|))}. \tag{5.5}$$

∎

THEOREM 5.2. *The solution of the reaction–diffusion-stochastic system* (2.17) *exists for all* $t \ge 0$.

Proof. Suppose that $[0, T)$ is the largest interval on which $\Phi(t)$ is a solution of the reaction–diffusion-stochastic system (2.17). If $\lim_{t\uparrow T}\Phi(t)$ exists in $\mathcal{C}_k^N$, then there exists a solution on the interval $[0, T+\delta)$ for some $\delta > 0$ that depends on the appropriate bounds on $\Phi(t)$. But this contradicts the assumption that $[0, T)$ is the largest interval of existence, and therefore the solution of (2.17) exists for all time.

In order to show that $\lim_{t\uparrow T}\Phi(t)$ exists, we obtain the Hölder estimate

$$\|\Phi(t+\delta) - \Phi(t)\|_k \le C\delta^{\frac{1}{2}} \tag{5.6}$$

for a constant C. Suppose $U(0) = u$. By Theorem 2.13 and Theorem 5.1,

$$\begin{aligned}\|U(t+\delta) - U(t)\|_k &\le \|(K_\delta - 1)K_t(u)\|_k + \int_t^{t+\delta} \|f(U(\tau))\|_k + \|g \circ \mathcal{T}_{X_\tau}\|_k \, d\tau \\ &\le \delta^{\frac{1}{2}}\|u\|_{k+1} + [C_{f,k}(C_U(\sup|u|), C_{U,k}(\sup|u|, \|u\|_k, T)) + C_{g,k}]\,\delta.\end{aligned}$$

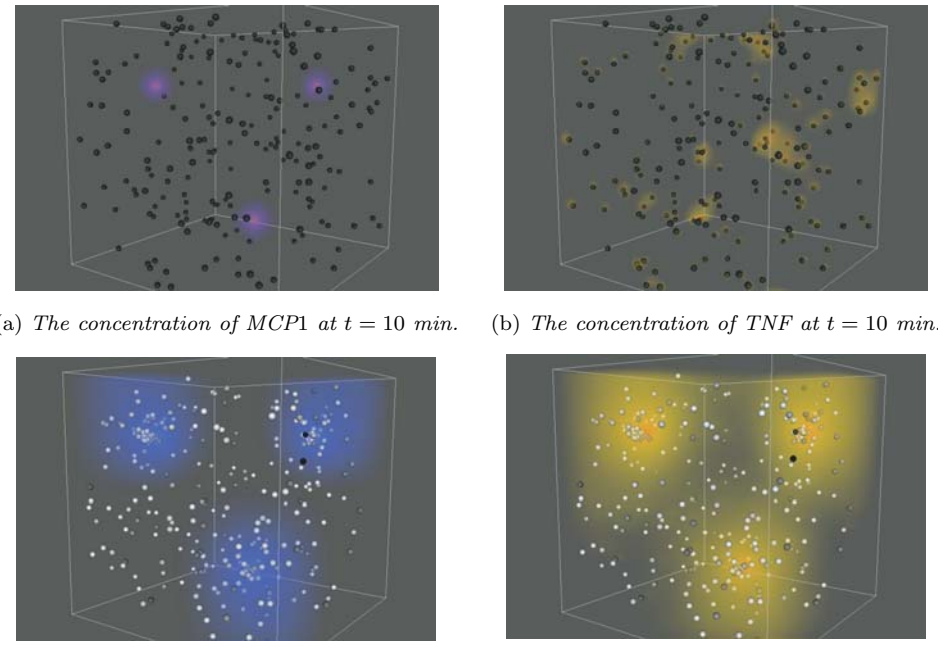
By (2.29) and (4.24), we have

$$\begin{aligned}\mathbb{E}\left[|X_{t+\delta} - X_t|^2\right] &\leq \mathbb{E}\left[2\left|\int_t^{t+\delta} a(U(X_\tau), X_\tau)d\tau\right|^2 + 2\left|\int_t^{t+\delta} b(X_\tau)dW_\tau\right|^2\right] \\ &\leq \mathbb{E}\left[2\delta\int_t^{t+\delta} |a(U(X_\tau), X_\tau)|^2 \, d\tau + 2\int_t^{t+\delta} |b(X_\tau)|^2 \, d\tau\right] \\ &\leq 2\delta^2 C_a(\xi, T) + 2\sigma^2\gamma\delta.\end{aligned}$$

Therefore, (5.6) holds, with $C = C(\sup|u|, \|u\|_k, \xi, T)$. ∎

6. Conclusions. The use of a first order splitting for the reaction–diffusion-stochastic system allows us to make use of known first order numerical methods for the diffusion, reaction, and stochastic differential equations. In particular, the diffusion can be discretized spatially using finite elements. From [1], we know that the resulting error is $\mathcal{O}(h^2)$, where h is the grid spacing. The resulting problem can be discretized in time using a first order backward Euler scheme as described in [17]. From [4], we know that the backward Euler scheme preserves positivity. The reaction can be approximated by a semi-implicit scheme that preserves positivity. General numerical methods for stochastic differential equations are outlined in [8]. In our case the cell motion (1.2) is a Langevin process that can be simulated exactly using ideas from [9]. By combining the numerical methods outlined above, we obtain a numerical simulation that is first order in time and preserves the positivity of the soluble factors.

The numerical methods outlined above have been implemented in a three-dimensional immune simulation with a large number of cells as described in [12]. Figure 6.1 shows a simulation of the immune response to three fixed sources of MCP1



(a) *The concentration of MCP*1 *at* $t = 10$ *min.* (b) *The concentration of TNF at* $t = 10$ *min.*

(c) *The concentration of MCP*1 *at* $t = 120$ *min.* (d) *The concentration of TNF at* $t = 120$ *min.*

FIG. 6.1. *A simulation of an immune response to three sources of MCP*1.

which is known to attract immune cells called macrophages. The macrophages secrete the cytokines TNF and sTNFr depending on their internal state, e.g., in Figure 6.1 the color of each macrophage from black to white represents the rate of secretion on TNF. The domain is a cube with side length 400μm and grid spacing 6.25μm. The radius of each macrophage is 5μm, and the cube initially contains 100 cells, with more cells entering the domain according to a Poisson distribution. The time step is 0.1 min. and in Figure 6.1 you can see the cells clustering around the sources of MCP1 after 120 min. The diffusion of MCP1 and TNF is also apparent in Figure 6.1.

Acknowledgments. The author thanks William K. Allard and Thomas B. Kepler who have served as mentors on this project. The author also thanks Cliburn Chan for his contributions to the simulation code and Eric Monson for his contributions to the vizualization of the simulation.

REFERENCES

[1] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, 2nd ed., Springer-Verlag, New York, 2002.
[2] L. C. Evans, *Partial Differential Equations*, Grad. Stud. Math. 19, American Mathematical Society, Providence, RI, 1998.
[3] G. B. Folland, *Introduction to Differential Equations*, 2nd ed., Princeton University Press, Princeton, N.J., 1995.
[4] W. Hundsdorfer and J. Verwer, *Numerical Solution of Time-Dependent Advection-Diffusion-Reaction Equations*, Springer-Verlag, Berlin, 2003.
[5] E. F. Keller and L. A. Segel, *Model for chemotaxis*, J. Theoret. Biol., 30 (1971), pp. 225–234.
[6] T. B. Kepler and C. Chan, *Spatiotemporal programming of a simple inflammatory process*, Immunol. Rev., 216 (2007), pp. 153–163.
[7] T. B. Kepler, *Microsimulation of Inducible Reorganization in Immunity*, in Complex Systems Science in BioMedicine, T. S. Deisboeck and J. Y. Kresh, eds., Springer-Verlag, New York, 2006, pp. 437–450.
[8] P. Kloeden and E. Platen, *Numerical Solution of Stochastic Differential Equations*, Springer-Verlag, Berlin, 1999.
[9] D. S. Lemons, *An introduction to stochastic processes in physics*, Translated by A. Gythiel, Johns Hopkins University Press, Baltimore, MD, 2002.
[10] X. Mao, *Stochastic Differential Equations and Their Applications*, Horwood Publishing Limited, Chichester, 1997.
[11] G. N. Milstein, *Numerical integration of stochastic differential equations*, Kluwer Academic Publishers Group, Dordrecht, 1995. Translated and revised from the 1988 Russian original.
[12] F. Mitha, T. A. Lucas, F. Feng, T. B. Kepler, and C. Chan, *The multiscale systems immunology project: Software for cell-based immunological simulation*, Source Code for Biology and Medicine, 3 (2008), article 6.
[13] B. Øksendal, *Stochastic Differential Equations*, 6th ed., Universitext, Springer-Verlag, Berlin, 2003.
[14] H. G. Othmer and L. E. Scriven, *Instability and dynamic pattern in cellular networks*, J. Theoret. Biol., 32 (1971), pp. 507–537.
[15] E. Palsson and H. G. Othmer, *A model for individual and collective cell movement in dictyostelium discoideum*, Proc. Natl. Acad. Sci. USA 97, 2000, pp. 10448–10453.
[16] M. E. Taylor, *Partial Differential Equations* III: Nonlinear Equations, Springer-Verlag, New York, 1997, corrected reprint of the 1996 original.
[17] V. Thomée, *Galerkin Finite Element Methods for Parabolic Problems*, 2nd ed., Springer-Verlag, Berlin, 2006.
[18] A. M. Turing, *The chemical basis of morphogenesis*, Bull. Math. Biol., 52 (1990), pp. 153–197; discussion 119–152.

SIAM J. NUMER. ANAL.
Vol. 46, No. 6, pp. 3136–3150
© 2008 Society for Industrial and Applied Mathematics

AN A POSTERIORI CONDITION ON THE NUMERICAL APPROXIMATIONS OF THE NAVIER–STOKES EQUATIONS FOR THE EXISTENCE OF A STRONG SOLUTION*

MASOUMEH DASHTI† AND JAMES C. ROBINSON†

Abstract. In their 2006 paper, Chernyshenko et al. [*J. Math. Phys.*, 48 (2007), 065204, 15 pp]. prove that a sufficiently smooth strong solution of the 3D Navier–Stokes equations is robust with respect to small enough changes in initial conditions and forcing function. They also show that if a regular enough strong solution exists, then Galerkin approximations converge to it. They then use these results to conclude that the existence of a sufficiently regular strong solution can be verified using sufficiently refined numerical computations. In this paper we study the strong solutions with less regularity than those considered in Chernyshenko et al. [*J. Math. Phys.*, 48 (2007), 065204, 15 pp]. We prove a similar robustness result and show the validity of the results relating convergent numerical computations and the existence of the strong solutions.

Key words. Navier–Stokes equations, regularity of solutions, rigorous numerics

AMS subject classifications. 76D05, 74H30

DOI. 10.1137/060677537

1. Introduction. For the three-dimensional (3D) Navier–Stokes equations the existence of strong solutions over an arbitrary time interval is not known. Therefore, the validity of the results of the numerical solutions of 3D Navier–Stokes equations is not obvious. This problem is addressed in [2] where a rigorous relationship between numerical and sufficiently regular exact solutions is given. It is shown there for sufficiently smooth initial conditions and forcing functions (data) that although a priori there is no guarantee of the validity of the numerical solutions, there is an a posteriori condition that if satisfied by the numerical results guarantees the existence of a strong solution. In this paper we show the validity of the results proved in [2] for less regular strong solutions than are covered there.

We will study the Navier–Stokes equations in their functional form. For a bounded domain Ω we let $\mathcal{H}$ be the space of divergence-free smooth vector-valued functions on Ω with compact support and define

$$H = \text{closure of } \mathcal{H} \text{ in } [L^2(\Omega)]^3,$$
$$V = \text{closure of } \mathcal{H} \text{ in } [H^1(\Omega)]^3.$$

In the case of a periodic domain Q, we consider $\mathcal{H}_p$ to be the space of periodic, divergence-free, and smooth vector-valued functions on Q with zero average and use the same notation H and V for the closure of $\mathcal{H}_p$ in $[L^2(Q)]^3$ and $[H^1(Q)]^3$, respectively. The Navier–Stokes equations in their functional form are written as [3, 11]

$$\frac{\mathrm{d}u}{\mathrm{d}t} + \nu Au + B(u,u) = f, \quad \text{with} \quad u(0) = u_0, \tag{1.1}$$

*Received by the editors December 12, 2006; accepted for publication (in revised form) May 12, 2008; published electronically September 4, 2008.
http://www.siam.org/journals/sinum/46-6/67753.html
†Mathematics Institute, University of Wawick, Coventry CV4 7AL, UK (m.dashti@warwick.ac.uk, j.c.robinson@warwick.ac.uk). The first author's research was funded by a Warwick Postgraduate Research Fellowship. The second author's research was supported by a Royal Society University Research Fellowship and the Leverhulme Trust.

where $Au = -\Pi\Delta$, $B(u,u) = \Pi(u.\nabla)u$ with Π the orthogonal projection from L^2 into H. We will consider this equation with the following cases for the data:

(a) "first-order regularity" when $u_0 \in V$ and $f \in L^2(0,T;H) \cap L^1(0,T;V)$

(b) "second-order regularity" when $u_0 \in V^2$ and $f \in L^2(0,T;V) \cap L^1(0,T;V^2)$,

where $V^m = H^m \cap V$. For the periodic case we know that $D(A^{m/2}) = H^m \cap V$ for all m, and therefore we define the norm on V^m as $\|u\|_m = |A^{m/2}u|$. We denote by $(\cdot,\cdot)$ and $|\cdot|$ the inner product and norm on H.

We will show that the strong solution of (1.1) with the data introduced in (a) or (b), if it exists, remains strong if the changes in the initial condition and forcing function are small enough. The exact conditions required for these changes to be "small" are given in Theorems 4.1 and 4.2. For example, in case (a) we require

$$|D(u_0 - v_0)| + \int_0^T |Df(s) - Dg(s)|\,\mathrm{d}s < \frac{1}{k}\left(\frac{\nu^3}{27T}\right)^{1/4} \exp\left(-\frac{k^2}{2}\int_0^T \frac{27k^2}{2}\frac{1}{\nu^3}|Du(s)|^4 + \frac{1}{\nu}|Du(s)|\,|Au(s)|\,\mathrm{d}s\right),$$

where v_0 and g are the nearby initial condition and forcing function and k is a constant. In Theorems 6.1 and 6.2, we use these robustness results to find an a posteriori condition that if satisfied by sufficiently refined numerical approximations, implies the existence of a strong solution.

We then show that if a strong solution exists the Galerkin approximations converge to it. This implies that assuming that the Navier–Stokes equations are a meaningful model of fluid flow, the numerical experiments based on the Galerkin method can be justified mathematically.

Combining the above two results, we prove that the existence of a strong solution can be verified by the Galerkin approximations. In the last section we consider a channel flow as a physical example that can be described by the Navier–Stokes equations with the conditions introduced above. For this example we will show how the results of this paper can be applied to the Galerkin approximations to verify the existence of a strong solution.

The results we prove here for a strong solution with first-order regularity hold in a general bounded domain as well as in the absence of boundaries unlike the results for more regular solutions which are proved only for the equations in a periodic domain or the whole space.

2. General ODE lemma. We first prove an ODE lemma which will be used in dealing with the differential inequalities that appear in the proofs. We consider the differential inequality

$$\frac{\mathrm{d}y}{\mathrm{d}t} \le \delta(t) + \alpha y^n \qquad \text{with} \quad y(0) = y_0 > 0$$

and find the conditions on y_0, $\delta(t)$ and α that ensure that $y(t)$ exists on a finite time interval $[0,T]$. This lemma is a generalization of the result obtained for $n = 2$ in [2].

LEMMA 2.1. *Let $T > 0$, $\alpha > 0$, and $n > 1$ be constants and let $\delta(t)$ be a nonnegative continuous function on $[0,T]$. Let y satisfy the differential inequality*

$$\frac{\mathrm{d}y}{\mathrm{d}t} \le \delta(t) + \alpha y^n \qquad \textit{with} \quad y(0) = y_0 > 0 \tag{2.1}$$

and define

$$\eta = y_0 + \int_0^T \delta(s)\,\mathrm{d}s.$$

(i) *If*

$$\eta < \frac{1}{[(n-1)\alpha T]^{1/(n-1)}}, \tag{2.2}$$

then $y(t)$ remains bounded on $[0,T]$.

(ii) *$y(t) \to 0$ uniformly on $[0,T]$ as $\eta \to 0$.*

Proof. We first consider the following differential inequality

$$\dot{z} \le \alpha z^n \quad \text{with} \quad z(0) = \eta \tag{2.3}$$

and show that

$$\sup_{S_1} y(T) \le \sup_{S_2} z(T), \tag{2.4}$$

where S_1 and S_2 are the sets of all possible solutions of inequalities (2.1) and (2.3), respectively.

Since $\dot{y}$, $\dot{z}$, $y(0)$, and $z(0)$ are nonnegative, the suprema of $y(T)$ and $z(T)$ happen when

$$\begin{aligned}\dot{y} &= \delta(t) + \alpha y^n,\\ \dot{z} &= \alpha z^n\end{aligned}$$

for all $t \in [0,T]$. In this case, for the difference $w = y - z$ we have

$$\dot{w} = \delta(t) + \alpha w \sum_{k=0}^{n-1} \binom{n-1}{k} y^{n-1-k}\, z^k \quad \text{with} \quad w(0) = -\int_0^t \delta(s)\,\mathrm{d}s.$$

Since $y(t)$ and $z(t)$ are greater than zero and assuming they remain finite over $[0,T]$, there exists some $M > 0$ such that

$$\dot{w} \le \alpha M w + \delta(t)$$

and therefore by Gronwall's inequality

$$\begin{aligned} w(t) &\le w(0)\mathrm{e}^{\alpha M t} + \int_0^t \delta(s)\,\mathrm{d}s\\ &= (1 - \mathrm{e}^{\alpha M t}) \int_0^t \delta(s)\,\mathrm{d}s. \end{aligned}$$

This implies $w(t) \le 0$ for any $t \in [0,T]$ and (2.4) follows.

The inequality (2.3) can be written as

$$\frac{\dot{z}}{z^n} \le \alpha.$$

Integrating both sides from 0 to t yields

$$z(T) \le \frac{\eta}{(-\alpha T(n-1)(\eta)^{n-1} + 1)^{1/n-1}}. \tag{2.5}$$

Since $y(t) \le y(T) \le z(T)$, $y(t)$ remains finite on $[0,T]$ provided that

$$\alpha T(n-1)(\eta)^{n-1} < 1,$$

which yields (2.2).

Furthermore, it is clear from (2.5) that $z(T) \to 0$ as $\eta \to 0$, from which it follows that $y(t) \to 0$ uniformly on $[0,T]$. □

3. The inequalities for the nonlinear term. For the nonlinear operator, we will use the following inequalities in this paper:

$$|(B(u,v),Aw)| \le k|Du|\,|Dv|^{1/2}|Av|^{1/2}|Aw| \tag{3.1}$$

$$|(B(w,u),A^2w)| \le c\|u\|_3\|w\|_2^2 \tag{3.2}$$

$$|(B(u,w),A^2w)| \le c'\|u\|_3\|w\|_2^2, \tag{3.3}$$

The last two inequalities only hold in $\mathbb{R}^3$ [9] and the periodic domain [3]. In the latter case where $\Omega = (0,L)^3$, c and c' depend on L [3]. The first inequality holds for both Dirichlet and periodic boundary conditions [3]. The constant k depends on the geometric properties and size of the domain. We show this here for a cubic domain which we will use in Section 7.

Using Hölder inequality, one can write

$$|(B(u,v),Aw)| \le \sum_{i,j=1}^{3} \|u_i\|_{L^6}\|D_iv_j\|_{L^3}|Aw_j|.$$

By the Sobolev embedding theorem and the Poincaré inequality, we have

$$\|u\|_{L^6} \le c_s\|u\|_{H^1} \le c_s(|u| + |Du|) \le c_s(1+c_p)\,|Du|.$$

Also, we can write

$$\begin{aligned}
\|Dv\|_{L^3} &\le \|Dv\|_{L^6}^{1/2}\,|Dv|^{1/2} \\
&\le c_s^{1/2}\,\|Dv\|_{H^1}^{1/2}\,|Dv|^{1/2} \\
&\le c_s^{1/2}\,(1+c_p)^{1/2}\,|D^2v|^{1/2}\,|Dv|^{1/2} \\
&\le c_s^{1/2}\,(1+c_p)^{1/2}\,c_e^{1/2}\,|Av|^{1/2}\,|Dv|^{1/2}.
\end{aligned}$$

Since

$$|(B(u,v),Aw)| \le \sum_{i,j=1}^{3} \|u_i\|_{L^6}\,\|D_iv_j\|_{L^3}\,|Aw_j| \le 9\,\|u\|_{L^6}\,\|Dv\|_{L^3}\,|Aw|,$$

we have

$$k = 9\,c_s^{3/2}\,c_e^{1/2}\,(1+c_p)^{3/2}. \tag{3.4}$$

Now for $\Omega = (0,L_1)\times(0,L_3)\times(0,L_3)$, along the lines of the proof of Theorem (5.10) of [1], one can show that

$$c_s = \sqrt{2}\,(1 + 2\max\{L_1,L_2,L_3\}).$$

Since for a function in a convex domain with zero average, the constant of the Poincaré inequality is d_D/π, where d_D is the diameter of the domain [10], $c_p = (L_1^2 + L_2^2 +$

$L_3^2)^{1/2}/\pi$. The constant c_e also depends on the geometric properties and size of the domain (see [7] or [11] for example)."

In obtaining the robustness conditions in the next section, we keep track of the constants k, c, and c' when they appear, so that the value of the constant coefficients in the robustness conditions could be computed if desired.

The inequalities (3.1)–(3.3) are not as elegant as the inequalities available for the more regular solutions considered in [2], and this is why different proofs are needed in the less regular cases we are studying here.

We note here that assuming a strong solution $u \in L^\infty(0,T;V) \cap L^2(0,T;V^2)$ exists, from (3.1) we have $B(u,u) \in L^2(0,T;H)$ and therefore (1.1) implies that $\mathrm{d}u/\mathrm{d}t \in L^2(0,T;H)$. Having $u_0 \in V^2$ and $f \in L^2(0,T;V)$ in case (b), from the regularity result for periodic domains proved in [3] we know that the strong solution u is, in fact, more regular and an element of $L^\infty(0,T;V^2) \cap L^2(0,T;V^3)$.

4. Robustness of strong solutions. Here we show that as long as a strong solution exists for some specific initial data and forcing function, the equations with sufficiently close data also have a strong solution. A similar result about the robustness of strong solutions with respect to changes in the forcing function is proved by Fursikov [6] in his 1980 paper. For the 3D Navier–Stokes equations with initial condition $u_0 \in V^m$ where $m \geq 1/2$, he shows that the set of forcing functions for which a strong solution exists is open in $L^q(0,T;V^{m-1})$ with $q \geq 2$.

For both the first-order regular solutions of case (a) and the more regular case (b), the condition we obtain here for the data depends explicitly on the viscosity coefficient unlike the more regular cases considered in [2]. The robustness result we prove for the strong solutions with $u_0 \in V$ holds in a general bounded domain. For the second-order regular solutions, however, we need to restrict the domain to be periodic or the whole of $\mathbb{R}^3$ since the inequalities (3.2) and (3.3) hold only in a periodic domain [3] or the whole of $\mathbb{R}^3$ [9].

(a) Strong solutions with first-order regularity. To prove the robustness with respect to the data, we write the governing equation for the difference between the solution of the equations with nearby data and the strong solution and find a bound on the norm of this difference.

THEOREM 4.1. *Consider $f, g \in L^1(0,T,V) \cap L^2(0,T;H)$, $u_0, v_0 \in V$ and let $u \in L^\infty(0,T;V) \cap L^2(0,T;V^2)$ be a strong solution of*

$$\frac{\mathrm{d}u}{\mathrm{d}t} + Au + B(u,u) = f \quad \textit{with} \quad u(0) = u_0.$$

If

$$\begin{aligned} &|D(u_0 - v_0)| + \int_0^T |Df(s) - Dg(s)|\,\mathrm{d}s \\ &< \frac{1}{k}\left(\frac{\nu^3}{27T}\right)^{1/4} \exp\left(-\frac{k^2}{2}\int_0^T \frac{27k^2}{2}\frac{1}{\nu^3}|Du(s)|^4 + \frac{1}{\nu}|Du(s)|\,|Au(s)|\,\mathrm{d}s\right), \end{aligned} \tag{4.1}$$

then the solution of

$$\frac{\mathrm{d}v}{\mathrm{d}t} + \nu Av + B(v,v) = g \quad \textit{with} \quad v(0) = v_0$$

is also a strong solution on $[0,T]$ with the same regularity as u.

Proof. By the local existence of strong solutions ([3], [12]), we know that there exists $T^* > 0$ such that $v \in L^\infty(0,T';V) \cap L^2(0,T';V^2)$ for any $T' < T^*$. We consider T^* the maximal time of existence of the strong solution v, meaning that $\limsup_{t\to T^*} |Dv| = \infty$. In the following argument we assume $T^* \le T$ and deduce a contradiction.

We consider $w = u - v$ which satisfies

$$\frac{\mathrm{d}w}{\mathrm{d}t} + \nu Aw + B(u,w) + B(w,u) - B(w,w) = f - g.$$

Over the time interval $t \in [0,T')$ for any $T' < T^*$, we have $\mathrm{d}v/\mathrm{d}t \in L^2(0,T';H)$ and since $T^* \le T$, $\mathrm{d}u/\mathrm{d}t \in L^2(0,T';H)$. Therefore, taking the inner product of the above equation with Aw and using (3.1) we can write

$$\begin{aligned}\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}|Dw|^2 + \nu|Aw|^2 \le k|Du||Dw|^{1/2}|Aw|^{3/2} + k|Dw||Du|^{1/2}|Au|^{1/2}|Aw| \\ + k|Dw|^{3/2}|Aw|^{3/2} + |D(f-g)|\,|Dw|.\end{aligned}$$

We then use Young's inequality to remove $|Aw|$ (which causes the appearence of ν in the coefficients) and obtain

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}|Dw|^2 \le \left(\frac{27k^4}{4\nu^3}|Du|^4 + \frac{k^2}{2\nu}|Du||Au|\right)|Dw|^2 + \frac{27k^4}{4\nu^3}|Dw|^6 + |D(f-g)|\,|Dw|.$$

Dividing by $|Dw|$ (assuming $|Dw|$ is nonzero; if not, replacing $|Dw|$ by $|Dw|+e_0$ with $e_0 > 0$ will lead us to the same result after tending e_0 to zero at the end) we get

$$\frac{\mathrm{d}}{\mathrm{d}t}|Dw| \le \left(\frac{27k^4}{4\nu^3}|Du|^4 + \frac{k^2}{2\nu}|Du||Au|\right)|Dw| + \frac{27k^4}{4\nu^3}|Dw|^5 + |D(f-g)|.$$

Now letting

$$\beta(t) = \frac{k^2}{2}\int_0^t \left(\frac{27k^2}{2}\frac{1}{\nu^3}|Du|^4 + \frac{1}{\nu}|Du|\,|Au|\right)\,\mathrm{d}s$$

and setting $y(t) = |Dw(t)|\mathrm{e}^{-\beta(t)}$ the above inequality can be written as

$$\frac{\mathrm{d}y}{\mathrm{d}t} \le \alpha y^5 + \delta(t),$$

where $\alpha = \frac{27k^4}{4\nu^3}\mathrm{e}^{4\beta(T)}$ and $\delta(t) = |Df(t) - Dg(t)|$.

By Lemma 2.1, if the condition (4.1) is satisfied, $|Dw(t)|$ is uniformly bounded on $[0,T^*)$. This implies that $|Dv(T^*)| < \infty$, which contradicts the maximality of T^*. It follows that $v(t)$ is a strong solution on $[0,T]$. ∎

(b) Strong solutions with second-order regularity. We follow the same approach as in the previous case to show the robustness for a strong solution with second-order regularity. Here also the inequalities we need to use for the nonlinear operator to find a bound on the V^2-norm of the difference between nearby solutions depend on the V^3-norm of this difference, and therefore the robustness condition on the data depends explicitly on the viscosity coefficient. However, its dependence is weaker in this case.

THEOREM 4.2. *Let $f \in L^1(0,T,V^2) \cap L^2(0,T;V)$, $u_0 \in V^2$ and $u \in L^\infty(0,T;V^2) \cap L^2(0,T;V^3)$ be a strong solution of*

$$\frac{\mathrm{d}u}{\mathrm{d}t} + Au + B(u,u) = f \quad \textit{with} \quad u(0) = u_0.$$

If

$$|A(u_0 - v_0)| + \int_0^T |A(f-g)|\,\mathrm{d}t \;<\; \frac{1}{c}\sqrt{\frac{2\nu}{T}} \exp\left(-\int_0^T (c+c')\|u\|_3\,\mathrm{d}t\right), \tag{4.2}$$

then the solution of

$$\frac{\mathrm{d}v}{\mathrm{d}t} + Av + B(v,v) = g \quad \textit{with} \quad v(0) = v_0$$

is also a strong solution on $[0,T]$ with the same regularity as u.

Proof. As before, the difference $w = u - v$ satisfies

$$\frac{\mathrm{d}w}{\mathrm{d}t} + \nu Aw + B(u,w) + B(w,u) + B(w,w) = f - g.$$

Taking the inner product of the above equation with $A^2 w$ and using (3.2) and (3.3) we obtain

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|w\|_2 + \nu\|w\|_3^2 \le (c+c')\|u\|_3^2\|w\|_2^2 + c\|w\|_3\|w\|_2^2 + \|f-g\|_2\|w\|_2.$$

We apply Young's inequality to the second term on the right-hand side and get

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|w\|_2^2 \le (c+c')\|u\|_3\|w\|_2^2 + \frac{c^2}{4\nu}\|w\|_2^4 + \|f-g\|_2\|w\|_2.$$

Dividing the above inequality by $\|w\|_2$ yields

$$\frac{\mathrm{d}}{\mathrm{d}t}\|w\|_2 \le (c+c')\|u\|_3\,\|w\|_2 + \frac{c^2}{4\nu}\|w\|_2^3 + \|f-g\|_2.$$

Now we consider

$$y(t) = \|w\|_2 e^{-\beta(t)}$$

with $\beta(t) = \int_0^t (c+c')\|u\|_3\,\mathrm{d}t$. The equation with this new variable becomes

$$\frac{\mathrm{d}y}{\mathrm{d}t} \le \delta(t) + \alpha y^3,$$

where $\alpha = \frac{c^2}{4\nu} e^{2\beta(T)}$ and $\delta = \|f-g\|_2$. Using Lemma 2.1 $\|w\|_2$ is bounded if the condition (4.2) is satisfied. ∎

5. Convergence of Galerkin approximations. Here we show that if a strong solution with first- or second-order regularity exists, Galerkin approximations converge to it. Similar results about convergence of various numerical methods, assuming the existence of a strong solution, are given for finite element methods by Heywood and Rannacher [8], for a Fourier collocation method by E [5] and for a nonlinear

Galerkin method by Devulder, Marion, and Titi [4]. Here, as in [2], we make no assumption on the regularity of the Galerkin approximations.

For the first-order regular case the result we prove here holds for the solution of the equations over a general bounded domain. For the second-order regular solutions we need to use inequality (3.3) which is valid only in a periodic domain or the whole of $\mathbb{R}^3$.

In the following theorems we let P_n be the orthogonal projection in H onto the space spanned by the first n eigenfunctions of the Stokes operator A, $\{w_j\}_{j=1}^n$, ordered so that their corresponding eigenvalues satisfy $0 < \lambda_1 \le \lambda_2 \le \dots$. Since the eigenfunctions of A are smooth [3], for any $u \in V^m$, $m \ge 0$ (with $V^0 = H$) we have $u_n = P_n u = \sum_{j=1}^n (u, w_j) w_j \in V^m$.

We note that what we obtain here about the convergence of Galerkin approximations would be useful even if the existence of regular solutions was known.

(a) First-order regularity. The following theorem, like the robustness theorem in the first-order regularity case, holds in sufficiently smooth bounded domains as well as in the absence of boundaries.

THEOREM 5.1. *Let $u_0 \in V$, $f \in L^2(0,T;H)$ and $u \in L^\infty(0,T;V) \cap L^2(0,T;V^2)$ be a strong solution of the Navier–Stokes equations*

$$\frac{\mathrm{d}u}{\mathrm{d}t} + \nu A u + B(u,u) = f(t) \quad \textit{with} \quad u(0) = u_0.$$

Then u_n, the solution of Galerkin system

$$\frac{\mathrm{d}u_n}{\mathrm{d}t} + \nu A u_n + P_n B(u_n, u_n) = P_n f(t) \quad \textit{with} \quad u_n(0) = P_n u_0, \tag{5.1}$$

converges strongly to u in both $L^\infty(0,T;V)$ and $L^2(0,T;V^2)$ as $n \to \infty$.

Proof. We consider $w_n = u - u_n$ which satisfies

$$\frac{\mathrm{d}w_n}{\mathrm{d}t} + \nu A w_n + P_n B(u, w_n) + P_n B(w_n, u) - P_n B(w_n, w_n) = Q_n f(t) - Q_n B(u,u),$$

with $Q_n = I - P_n$. Let $q_n = u - P_n u$, noting that u_n the solution of the Galerkin system (5.1) is different from $P_n u$, the image of u under orthogonal projection P_n. Taking the inner product of the above equation with Aw_n while noting that $(P_n B(u,v), Aw_n) = (B(u,v), P_n A w_n)$, we obtain

$$\begin{aligned}\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}|Dw_n|^2 + \nu |Aw_n|^2 \le k \Big(& |Du|\,|Dw_n|^{1/2}|Aw_n|^{3/2} \\ & + |Dw_n|\,|Du|^{1/2}\,|Au|^{1/2}|Aw_n| + |Dw_n|^{3/2}|Aw_n|^{3/2} \\ & + |Du|^{3/2}|Au|^{1/2}|Aq_n| \Big) + |f|\,|Aq_n|.\end{aligned}$$

Applying Young's inequality and redefining the constant k gives

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}|Dw_n|^2 \le k \Bigg\{ & \left(\frac{1}{\nu^3}|Du|^4 + \frac{1}{\nu}|Du|\,|Au| \right) |Dw_n|^2 + \frac{1}{\nu^3}|Dw_n|^6 \\ & + |Du|^{3/2}|Au|^{1/2}|Aq_n| \Bigg\} + |f|\,|Aq_n|.\end{aligned}$$

Now letting

$$\beta(t) = k \int_0^t \left(\frac{1}{\nu^3}|Du|^4 + \frac{1}{\nu}|Du|\,|Au| \right) \mathrm{d}s$$

and

$$y_n(t) = \mathrm{e}^{-\beta(t)}|Dw_n|^2,$$

the above inequality can be written as

$$\frac{\mathrm{d}y_n}{\mathrm{d}t} \le \alpha y_n^3 + \delta_n(t),$$

with $\alpha = \frac{k}{\nu^3}\mathrm{e}^{2\beta(T)}$ and $\delta_n(t) = |f|\,|Aq_n| + k|Du|^{3/2}|Au|^{1/2}|Aq_n|$.

Using the Hölder inequality we have

$$\int_0^T \delta_n(s)\,\mathrm{d}s \le \left[\left(\int_0^T |f|^2\mathrm{d}s\right)^{1/2} + k\left(\int_0^T |Du|^3|Au|\mathrm{d}s\right)^{1/2}\right]\left(\int_0^T |Aq_n|^2\mathrm{d}s\right)^{1/2}.$$

Since $u \in L^2(0,T;V^2)$, $u(s) \in V^2$ for almost every $s \in [0,T]$ and therefore $q_n(s) = Q_n u(s) \to 0$ in V^2 as $n \to \infty$ and $|Aq_n(s)|^2 \le |Au(s)|^2$ for a.e. $s \in [0,T]$. Hence the Lebesgue dominated convergence theorem for each s implies that $\int_0^T |Aq_n|^2\mathrm{d}s$ tends to zero as $n \to \infty$ and therefore

$$\int_0^T \delta_n(s)\,\mathrm{d}s \to 0$$

as $n \to \infty$. Since $y_n(0) \to 0$ as $n \to \infty$, Lemma 2.1 shows that y_n, and hence $|Dw_n|^2$, converges to zero uniformly on $[0,T]$ as $n \to \infty$. □

(b) Second-order regularity. Here also, in a similar way to the previous case, we try to find a bound on the V^2-norm of the difference between a second-order regular strong solution and Galerkin approximations. However, here the more regular solution space makes the proof easier.

THEOREM 5.2. *Let $u_0 \in V^2$, $f \in L^2(0,T;V) \cap L^1(0,T;V^2)$ and $u \in L^\infty(0,T;V^2) \cap L^2(0,T;V^3)$ be a strong solution of the Navier–Stokes equations*

$$\frac{\mathrm{d}u}{\mathrm{d}t} + \nu Au + B(u,u) = f(t) \quad \textit{with} \quad u(0) = u_0.$$

Then u_n, the solution of Galerkin system

$$\frac{\mathrm{d}u_n}{\mathrm{d}t} + \nu Au_n + P_nB(u_n,u_n) = P_nf(t) \quad \textit{with} \quad u_n(0) = P_nu_0, \tag{5.2}$$

converges strongly to u in both $L^\infty(0,T;V^2)$ and $L^2(0,T;V^3)$ as $n \to \infty$.

Proof. Let $w_n = u - u_n$. Then w_n satisfies

$$\frac{\mathrm{d}w_n}{\mathrm{d}t} + \nu Aw_n + P_nB(u,w_n) + P_nB(w_n,u) + P_nB(w_n,w_n) = Q_nf - Q_nB(u,u).$$

Taking the inner product of the above equation with A^2w_n and using (3.2) and (3.3) we obtain

$$\begin{aligned}\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|w_n\|_2^2 + \nu\|w_n\|_3^2 \le (c+c')\|u\|_3\|w_n\|_2^2 + c\|w_n\|_3\|w_n\|_2^2 \\ + \|Q_nf - Q_nB(u,u)\|_2\|w_n\|_2\end{aligned}$$

Now we use Young's inequality to remove the dependence on $\|w_n\|_3$ and then divide by $\|w\|_2$ to obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}\|w_n\|_2 \le (c+c')\|u\|_3\|w_n\|_2 + \frac{c^2}{4\nu}\|w_n\|_2^3 + \|Q_nf - Q_nB(u,u)\|_2,$$

in which the coefficient of $\|w_n\|_2$ does not depend on n.

Letting

$$y_n = \|w_n\|_2 e^{-\beta(t)},$$

with $\beta(t) = \exp\left(\int_0^t (c+c')\|u(s)\|_3 \, \mathrm{d}s\right)$ yields

$$\dot{y}_n \le \delta(t) + \alpha y_n^3,$$

where $\alpha = \frac{c^2}{4\nu} e^{2\beta(T)}$ and $\delta(t) = \|Q_n f - Q_n B(u,u)\|_2$. So by Lemma 2.1 if

$$\eta = \|Q_n u_0\|_m + \int_0^T \|Q_n f - Q_n B(u,u)\|_2 \, \mathrm{d}s \to 0 \quad \text{as} \quad n \to \infty,$$

then $y_n(t) \to 0$ uniformly on $[0,T]$ as $n \to \infty$.

By (5.2) $\|Q_n u_0\|_m \to 0$ as $n \to \infty$. We know that

$$\|B(u,u)\|_2 \le c\|u\|_2 \|u\|_3 \tag{5.3}$$

([9], [3]), and therefore $f(s) - B(u(s),u(s)) \in V^2$ for almost every $s \in [0,T]$. So since $\{w_j\}_{j=1}^n$ forms a basis for V^2 as well as H, $\|Q_n\left(f(s) - B(u(s),u(s))\right)\|_2 \to 0$ and

$$\|Q_n\left(f(s) - B(u(s),u(s))\right)\|_2 \le \|f(s) - B(u(s),u(s))\|_2$$

for almost every $s \in [0,T]$. Therefore, by the Lebesgue dominated convergence theorem it follows that

$$\int_0^T \|Q_n\left(f(s) - B(u(s),u(s))\right)\|_2 \to 0$$

and the result follows. □

6. Numerical verification of the existence of a strong solution. Here we show that the existence of first- and second-order regular strong solutions can be verified via computations using sufficiently refined Galerkin approximations.

THEOREM 6.1. (i) *Consider the Navier–Stokes equations*

$$\frac{\mathrm{d}u}{\mathrm{d}t} + \nu A u + B(u,u) = f \quad \textit{with} \quad u(0) = u_0 \tag{6.1}$$

with $u_0 \in V$, $f \in L^2(0,T;H) \cap L^1(0,T;V)$, that hold in a bounded domain Ω with sufficiently smooth boundary or in a periodic domain.

Let $v \in L^\infty(0,T;V) \cap L^2(0,T;V^2)$ be a numerical approximation of u satisfying

$$\frac{\mathrm{d}v}{\mathrm{d}t} + \nu A v + B(v,v) \in L^1(0,T;V) \cap L^2(0,T;H)$$

and

$$\begin{aligned} &|Dv(0) - Du_0| + \int_0^T \left\| \frac{dv(s)}{ds} + \nu A v(s) + B(v(s),v(s)) - f(s) \right\|_1 \mathrm{d}s \\ &< \frac{1}{k}\left(\frac{\nu^3}{27T}\right)^{1/4} \exp\left(-\frac{k^2}{2}\int_0^T \frac{27k^2}{2}\frac{1}{\nu^3}|Dv(s)|^4 + \frac{1}{\nu}|Dv(s)||Av(s)|\,\mathrm{d}s\right). \end{aligned} \tag{6.2}$$

Then the Navier–Stokes equations (6.1) *have a strong solution* $u \in L^\infty(0,T;V) \cap L^2(0,T;V^2)$.

(ii) *Let* u *be a strong solution of* (6.1). *Then there exists an* N *such that the Galerkin approximations* u_n *satisfy the inequality* (6.2) *for all* $n > N$. *Therefore in view of part* (i), u *passes the a posteriori test as a solution approximated by* u_n; *i.e., the existence of a strong solution with first-order regularity can be verified by the Galerkin approximations.*

Proof.

(i) Considering

$$g = \frac{\mathrm{d}v}{\mathrm{d}t} + \nu Av + B(v,v),$$

v is a strong solution of

$$\frac{\mathrm{d}z}{\mathrm{d}t} + \nu Az + B(z,z) = g, \quad \text{with } z(0) = v(0).$$

So by Theorem 4.1, if the inequality (6.2) holds, the solution of the equations with nearby data (u_0, f) is a strong solution.

(ii) The strong convergence of u_n to u in $L^\infty(0,T;V)$ and $L^2(0,T;V^2)$ is guaranteed by Theorem 5.1. Therefore, the right-hand side of (6.2) is bounded below for every n. It follows from (5.1) that $|Du_n(0) - Du_0| \to 0$ as $n \to 0$ and

$$\frac{du_n}{dt} + \nu Au_n + B(u_n,u_n) - f = Q_n[B(u_n,u_n) - f].$$

Therefore it only remains to show $\|Q_n[B(u_n(s),u_n(s)) - f]\|_1$ converges to zero as $n \to \infty$ for a.e. $s \in [0,T]$; it is then clear that (6.2) will be satisfied for all n sufficiently large.

For the nonlinear operator we have

$$\begin{aligned}\|B(u_n,u_n)\|_1 &= \sqrt{(D(u_n\cdot\nabla)u_n, D(u_n\cdot\nabla)u_n)} \\ &\le c\left(\|Du_n\|^4_{L^4(\Omega)} + \|u_n\|^2_{L^\infty(\Omega)}|Au_n|^2\right)^{1/2}.\end{aligned}$$

Using the Sobolev inequalities we can write

$$\|Du_n\|_{L^4(\Omega)} \le c\|Du_n\|_{L^6(\Omega)} \le c|Au_n|$$

and

$$\|u_n\|_{L^\infty} \le c|Au_n|.$$

Therefore $\|B(u_n,u_n)\|_1 \le c|Au_n|^2$ which implies

$$f(s) - B(u_n(s),u_n(s)) \in V \quad \text{for a.e. } s \in [0,T]$$

and we will have

$$\|Q_n(B(u_n(s),u_n(s)) - f(s))\|_1 \to 0 \quad \text{as } n \to \infty \text{ and for a.e. } s \in [0,T]$$

and

$$\|Q_n(B(u_n(s),u_n(s)) - f(s))\|_1 \le \|B(u_n(s),u_n(s)) - f(s)\| \quad \text{for a.e. } s \in [0,T].$$

By the Lebesgue-dominated convergence theorem we conclude that

$$\int_0^T \|Q_n(B(u_n(s),u_n(s))-f(s))\|_1 \mathrm{d}s \to 0 \quad \text{as } n\to\infty. \qquad \square$$

A similar result holds for the strong solutions with second-order regularity in a periodic domain.

THEOREM 6.2. (i) *Consider the Navier–Stokes equations*

$$\frac{\mathrm{d}u}{\mathrm{d}t}+\nu Au+B(u,u)=f \quad \textit{with} \quad u(0)=u_0 \tag{6.3}$$

with $u_0\in V^2$, $f\in L^2(0,T;V^1)\cap L^1(0,T;V^2)$, that hold in a periodic domain.

Let $v\in L^\infty(0,T;V^2)\cap L^2(0,T;V^3)$ be a numerical approximation of u satisfying

$$\frac{\mathrm{d}v}{\mathrm{d}t}+\nu Av+B(v,v)\in L^1(0,T;V^2)\cap L^2(0,T;V)$$

and

$$\begin{aligned}\|v(0)-u_0\|_2+\int_0^T \left\|\frac{dv(s)}{ds}(s)+\nu Av(s)+B(v(s),v(s))-f(s)\right\|_2 \mathrm{d}s \\ < \frac{1}{c}\sqrt{\frac{2\nu}{T}}\exp\left(-\int_0^T (c+c')\|u_n\|_3\,\mathrm{d}t\right).\end{aligned} \tag{6.4}$$

Then the Navier–Stokes equations (6.3) *have a strong solution $u\in L^\infty(0,T;V^2)\cap L^2(0,T;V^3)$.*

(ii) *Let u be a strong solution of* (6.3). *Then there exists an N such that the Galerkin approximations u_n satisfy the inequality* (6.4) *for all $n>N$. Therefore in view of part* (i), *u passes a posteriori test as a solution approximated by u_n which means that the existence of the strong solution u can be verified using the Galerkin approximations.*

Using the robustness result for the strong solutions with second-order regularity, Theorem 4.2, the proof of the above theorem closely parallels that of Theorem 6.1.

7. A physical application: channel flow. We considered the Navier–Stokes equations for flow in a domain with nonmoving boundary and with the forcing function in the form of a body force. A physical situation for which the results of this paper and in particular the a posteriori test of Theorem 6.1 can be applied is a channel flow in the domain $0<x<L_x$, $0<y<1$, $0<z<L_z$ with the velocity field $u(x,y,z,t)$ considered to be periodic in x and z with periods L_x and L_z, respectively, and zero at $y=0$ and $y=1$. The body force g is assumed constant and equal to $(1,1,1)$. Here we show how to write the inequality (6.2) for the Galerkin approximations, u_n, of the solution u of this problem.

By Theorem 6.1 we know that if for some n the Galerkin approximation, u_n, satisfies

$$\begin{aligned}|Du_n(0)-Du_0|+\int_0^T \left\|\frac{du_n(s)}{ds}+\nu Au_n(s)+B(u_n(s),u_n(s))-f(s)\right\|_1 \mathrm{d}s \\ < \frac{1}{k}\left(\frac{\nu^3}{27T}\right)^{1/4}\exp\left(-\frac{k^2}{2}\int_0^T \frac{27k^2}{2}\frac{1}{\nu^3}|Du_n(s)|^4+\frac{1}{\nu}|Du_n(s)||Au_n(s)|\,\mathrm{d}s\right),\end{aligned}$$

then a strong solution u exists and the Galerkin approximations converge to it. To check this inequality for a Galerkin approximation of the above channel flow example, we need to compute Au_n, $B(u_n, u_n)$, and $f = \Pi g$.

The functions

$$w_{\mathbf{k}} = \mathrm{e}^{2\pi i(\frac{k_1}{L_x}x + \frac{k_3}{L_z}z)} \sin(\pi k_2 y)$$

with $\mathbf{k} = (k_1, k_2, k_3)$, form an orthogonal basis for the space of L^2-functions on the cubic domain introduced above, which have periodic bounday values in x and z directions and are zero on $y = 0$ and $y = 1$. We can therefore define the Galerkin approximation u_n in terms of $w_{\mathbf{k}}$ as

$$u_n = \sum_{\mathbf{k}=\mathbf{n}_0}^{\mathbf{n}} \alpha_{\mathbf{k}} w_{\mathbf{k}}, \quad \text{with} \quad \nabla \cdot u_n = 0 \tag{7.1}$$

where $\alpha_{\mathbf{k}} = (\alpha_{1\mathbf{k}}(t), \alpha_{2\mathbf{k}}(t), \alpha_{3\mathbf{k}}(t))$, $\alpha_{(-k_1,k_2,-k_3)} = \bar{\alpha}_{(k_1,k_2,k_3)}$ (with $\bar{\alpha}_{\mathbf{k}}$ the complex conjugate of $\alpha_{\mathbf{k}}$), $\mathbf{n}_0 = (-n, 0, -n)$, and $\mathbf{n} = (n, n, n)$. Since u_n is divergence-free we have

$$\sum_{\mathbf{k}=\mathbf{n}_0}^{\mathbf{n}} 2\pi i \left(\frac{k_1}{L_x}\alpha_{1\mathbf{k}} + \frac{k_3}{L_z}\alpha_{3\mathbf{k}}\right) \sin(\pi k_2 y) + \pi k_2\, \alpha_{2\mathbf{k}} \cos(\pi k_2 y) = 0.$$

After expanding $\cos(\pi k_2 y)$ in terms of sine functions we obtain

$$\pi i \left(\frac{k_1}{L_x}\alpha_{1\mathbf{k}} + \frac{k_3}{L_z}\alpha_{3\mathbf{k}}\right) + \sum_{l=\lfloor \frac{k_2}{2} \rfloor}^{\lfloor \frac{k_2+n}{2} \rfloor} (2l + 1 - k_2) \left(\frac{1}{2k_2 - 2l - 1} + \frac{1}{2l+1}\right) \alpha_{2\mathbf{k}} = 0.$$

Defining

$$\hat{\mathbf{k}} = \left(\frac{\pi i k_1}{L_x}, \sum_{l=\lfloor \frac{k_2}{2} \rfloor}^{\lfloor \frac{k_2+n}{2} \rfloor} (2l + 1 - k_2) \left(\frac{1}{2k_2 - 2l - 1} + \frac{1}{2l+1}\right), \frac{\pi i k_3}{L_z} \right)$$

and having (7.1) we obtain

$$Au_n = 4\pi^2 \sum_{\mathbf{k}=\mathbf{n}_0}^{\mathbf{n}} \left(\frac{k_1^2}{L_x^2} + \frac{k_2^2}{4} + \frac{k_3^2}{L_z^2}\right) \alpha_{\mathbf{k}}\, w_{\mathbf{k}}$$

and

$$B(u_n, u_n) = 2 \sum_{\mathbf{k}=2\mathbf{n}_0}^{2\mathbf{n}} \sum_{\mathbf{j}=\mathbf{k}_-}^{\mathbf{k}_+} \sum_{m=\lfloor \frac{k_2+1}{2} \rfloor}^{\lfloor \frac{k_2+2n}{2} \rfloor} \left\{ (\alpha_{\mathbf{k}_m - \mathbf{j}} \cdot \mathbf{j}) \left(\alpha_{\mathbf{j}} - \frac{\hat{\mathbf{k}}}{|\hat{\mathbf{k}}|^2} (\alpha_{\mathbf{j}} \cdot \hat{\mathbf{k}}) \right) \right.$$
$$\left. \frac{1}{\pi} \left(\frac{1}{2m - 2j_2 + 1} + \frac{1}{2m - 2j_2 - 2k_2 + 1} - \frac{1}{2m+1} - \frac{1}{2m - 2k_2 + 1} \right) w_{\mathbf{k}} \right\},$$

where $\mathbf{k}_m = (k_1, 2m + 1 - k_2, k_3)$, $\mathbf{k}_- = (\min\{k_1, 0\}, \min\{k_2, 0\}, \min\{k_3, 0\})$, and $\mathbf{k}_+ = (\max\{k_1, 0\}, \max\{k_2, 0\}, \max\{k_3, 0\})$.

Now we note that $f = (1,0,1)$, since $f - g = (0,1,0)$ is perpendicular (with respect to L^2-norm) to all functions in H. Therefore, using the expansion of u_n and Au_n with respect to $\{w_{\mathbf{k}}\}_{\mathbf{k}=\mathbf{n}_0}^{\mathbf{n}}$ and $B(u_n,u_n)$ with respect to $\{w_{\mathbf{k}}\}_{\mathbf{k}=2\mathbf{n}_0}^{2\mathbf{n}}$, we can write (noting that the coefficients of $w_{\mathbf{k}}$ when k_1, k_2, or k_3 are less than $-n$ or bigger than n are zero for u_n and Au_n)

$$E_n = \frac{\mathrm{d}u_n}{\mathrm{d}t} + \nu A u_n + B(u_n,u_n) - f = \sum_{\mathbf{k}=2\mathbf{n}_0}^{2\mathbf{n}} \beta_{\mathbf{k}} w_{\mathbf{k}} - (1,0,1)$$

and therefore

$$DE_n = \frac{\partial E_n}{\partial x} + \frac{\partial E_n}{\partial y} + \frac{\partial E_n}{\partial z} = 2 \sum_{\mathbf{k}=2\mathbf{n}_0}^{2\mathbf{n}} \beta_{\mathbf{k}} \left(\hat{k}_1 + \hat{k}_2 + \hat{k}_3\right) w_{\mathbf{k}}$$

and

$$\left\| \frac{\mathrm{d}u_n}{\mathrm{d}t} + \nu A u_n + B(u_n,u_n) - f \right\|_1^2 = 2L_xL_z \sum_{\mathbf{k}=2\mathbf{n}_0}^{2\mathbf{n}} \left(\beta_{1\mathbf{k}}^2 + \beta_{2\mathbf{k}}^2 + \beta_{3\mathbf{k}}^2\right) \left(\hat{k}_1 + \hat{k}_2 + \hat{k}_3\right)^2.$$

The norms of Du_n and Au_n are computed as

$$|Du_n|^2 = 2L_xL_z \sum_{\mathbf{k}=\mathbf{n}_0}^{\mathbf{n}} \left(\alpha_{1\mathbf{k}}^2 + \alpha_{2\mathbf{k}}^2 + \alpha_{3\mathbf{k}}^2\right) \left(\hat{k}_1 + \hat{k}_2 + \hat{k}_3\right)^2,$$

$$|Au_n|^2 = 2\pi^2 L_xL_z \sum_{\mathbf{k}=\mathbf{n}_0}^{\mathbf{n}} \left(\alpha_{1\mathbf{k}}^2 + \alpha_{2\mathbf{k}}^2 + \alpha_{3\mathbf{k}}^2\right) \left(\frac{k_1^2}{L_x^2} + \frac{k_2^2}{4} + \frac{k_3^2}{L_z^2}\right)^2.$$

The remaining term on the left-hand side of (6.2), $|Du_n(0) - Du_0|$, is obtained in a similar way to $|Du_n|$.

We note that for the spatial domain and the boundary values of this example, the constant k is obtained from (3.4) by letting $c_s = \sqrt{2}\,(1 + 2\max\{L_x, L_y, 1\})$, $c_p = (L_x^2 + L_z^2 + 1)^{1/2}/\pi$ as explained in Section 3, and $c_e = (\max\{L_x, L_z, 2\})^2/4\pi^2$ [11, Theorem 6.5].

8. Conclusion. We extended the results of [2] to the less regular strong solutions not considered in their paper. Unlike their results, the robustness conditions on the data that we obtained for the first- and second-order regular strong solutions in Theorems 4.1 and 4.2 depended explicitly on the viscosity coefficient in a way that as the viscosity coefficient became smaller, the conditions (4.1) and (4.2) became more restrictive which obviously is not desirable. Moreover, since this is due to the less elegant inequalities available for the trilinear form in the less regular spaces considered here, it does not seem that we can remove the viscosity coefficient from the robustness conditions.

On the other hand, the results proved here for the first-order regular strong solutions are valid in a sufficiently smooth bounded domain as well as in the absence of boundaries, while in the case of the second-order regular solutions considered also here and the more regular ones studied in the above-mentioned paper, the results are restricted to a periodic domain or the whole of $\mathbb{R}^3$. The reason is that the inequalities we need to use in these cases for $|(B(u,v),v)_m|$ when $m \geq 2$ are valid in the absence of boundaries where we know that swapping the action of Π, the orthogonal projection from L^2 onto H, and the Laplacian operator is possible.

It would be interesting to see if there exists a big enough m for which it is possible to swap the action of the Laplacian operator and Π in $D(A^{m/2})$ over a bounded domain. If such a finite m exists we can extend the robustness results for regular enough solutions to the more physically relevant bounded domains. Furthermore, for two-dimensional bounded domains, equality of the Stokes and Laplacian operators in $D(A^{m/2})$ can be used in obtaining a better bound on the attractor dimension in L^2.

REFERENCES

[1] R. A. Adams, *Sobolev Spaces*, Academic Press, New York, 1975.

[2] S. I. Chernyshenko, P. Constantin, J. C. Robinson, and E. S. Titi, *A posteriori regularity of the three-dimensional Navier–Stokes equations from numerical computations*, J. Math. Phys., 48 (2007), 065204, 15 pp.

[3] P. Constantin and C. Foias, *Navier–Stokes Equations*, University of Chicago Press, Chicago, 1988.

[4] C. Devulder, M. Marion, and E. S. Titi, *On the rate of convergence of the nonlinear Galerkin methods*, Math. Comp., 60 (1993), pp. 495–514.

[5] W. E, *Convergence of Fourier methods for the Navier–Stokes equations*, SIAM J. Numer. Anal., 30 (1993), pp. 650–674.

[6] A. V. Fursikov, *Some control problems and results concerning the unique solvability of a mixed boundary value problem for the three-dimentional Navier–Stokes and Euler systems*, Soviet Math. Dokl., 21 (1980), pp. 889–893.

[7] D. Gilbarg and N.S. Trudinger, *Elliptic partial differential equations of second order*, Springer-Verlag, Berlin, 1983.

[8] J. G. Heywood and R. Rannacher, *Finite element approximation of the nonstationary Navier–Stokes problem.* I. *Regularity of solutions and second-order error estimates for spatial discretization*, SIAM J. Numer. Anal., 19 (1982), pp. 275–311.

[9] T. Kato, *Nonstationary flows of viscous and ideal fluids in* $\mathbb{R}^3$, J. Funct. Anal., 9 (1972), pp. 296–305.

[10] L. E. Payne And H. F. Weinberger, *An optimal Poincaré inequality for convex domains*, Arch. Ration. Mech. Anal., 5 (1960), pp. 286–292.

[11] J. C. Robinson, *Infinite-dimensional Dynamical Systems*, Cambridge University Press, Cambridge, 2001.

[12] R. Temam, *Navier–Stokes Equations and Nonlinear Functional Analysis*, SIAM, Philadelphia, 1983.

SIAM J. NUMER. ANAL.
Vol. 46, No. 6, pp. 3151–3180
© 2008 Society for Industrial and Applied Mathematics

TAMING THE CFL NUMBER FOR DISCONTINUOUS GALERKIN METHODS ON STRUCTURED MESHES*

T. WARBURTON† AND T. HAGSTROM‡

Abstract. The upwind discontinuous Galerkin method is an attractive method for solving time-dependent hyperbolic conservation laws. It is possible to use high-order explicit time-stepping methods and high-order spatial approximations without incurring heavy numerical linear algebra overheads. However, the Courant–Friedrichs–Lewy (CFL) condition for these methods depends on the polynomial order used, and there is a somewhat excessive cost for using very high order spatial approximation. We discuss the impact of a covolume mesh based filter on the CFL number for these methods and present an algorithm which has a CFL number independent of the spatial order of approximation. We present computational results for the advection equation and the wave equation on one-dimensional meshes using up to tenth order in space and time.

Key words. discontinuous Galerkin, overlapping, Courant–Friedrichs–Lewy

AMS subject classification. 35G20

DOI. 10.1137/060672601

1. Introduction.

A significant numerical hurdle for wave propagation is the generic need for preservation of phase and amplitude information as rapidly varying wave profiles travel over long distances. The upwind discontinuous Galerkin (DG) method has emerged as a competitive numerical modeling technique for achieving this goal. This method was originally proposed for neutron transport by Reed and Hill [27] and analyzed by many, including Lesaint and Raviart [24], Johnson and Pitkäranta [19], Richter [28], and Peterson [25]. The DG method was revisited for time-dependent conservation laws in a sequence of papers by Cockburn, Shu, et al. (see [9, 8, 7, 11]). The method is particularly attractive for wave propagation because of the ability to use high-order, piecewise polynomial approximations for the solution, which provides an efficient mechanism to manage phase and dissipation errors (see, for example, Ainsworth [1]). The method has been shown to be effective on quadrilateral meshes by Kopriva, Woodruff, and Hussaini [20, 21], on unstructured meshes of triangles with high-order approximation for Maxwell's equations by Warburton [31] and Hesthaven and Warburton [15, 16], and on overlapping triangular meshes by Chung and Engquist [5]. Furthermore, this method automatically controls spurious solutions through selective dissipation of nonphysical solutions (see, for example, Hesthaven and Warburton [17] and Warburton and Embree [32]). In addition, the structure of the numerical method results in computational algorithms which scale efficiently for large scale SIMD computations (see Biswas, Devine, and Flaherty [2]).

The positive aspects of the upwind DG method are accompanied by a particular sour note. As the order of the polynomial approximation n is increased on a fixed mesh of elements, with mesh size h, the spatial differential operator discretized with

*Received by the editors October 18, 2006; accepted for publication (in revised form) May 12, 2008; published electronically September 17, 2008. This work was supported by the AFOSR under contract FA9550-05-1-0473.

http://www.siam.org/journals/sinum/46-6/67260.html

†Computational and Applied Mathematics, Rice University, 6100 Main Street, MS–134, Houston, TX 77005–1892 (timwar@caam.rice.edu).

‡Department of Mathematics and Statistics, University of New Mexico, 1 University of New Mexico, Albuquerque, NM 87131-0001 (hagstrom@math.unm.edu). This author's research was supported by ARO grant DAAD19-03-1-0146 and NSF grant DMS-0610067.

the DG method typically has a spectral radius which grows as n^2/h. This is in contrast to typical finite difference discretized spatial differential operators which have spectral radii which grow as n/h. This translates into the reality that the DG method requires smaller time steps, dt, to scale its spatial derivative operator's spectrum into the stability region of a standard Runge–Kutta (RK) time-stepping method. Thus, the advantages of the DG method are tempered by the extra number of time steps compared with finite difference methods.

In this paper we present a simple modification to the upwind DG method which reduces the spectral radius of a modified version of the discrete DG derivative operator to a more competitive n/h. By way of motivation we note the Hermite–Taylor method analyzed by Goodrich, Hagstrom, and Lorenz [13]. There it is shown that a combination of staggered Hermite interpolation and Taylor expansion based time stepping leads to a method with degree-independent time-step stability constraints. Similarly, we find that for the case of an autonomous system of linear partial differential equations, when combining the modified upwind DG spatial discretization with an $(n+1)$-stage, explicit, RK time-stepping method it is possible to choose a stable time step size, dt, independent of the spatial polynomial order n. In the more general case of partial differential equations with nonlinear terms or time-dependent forcing, we can expect a factor proportional to n increase in the allowable time step size. In the conclusion we discuss the limitations of this approach.

2. Formulation. In this section we describe a standard upwind DG formulation for the one-dimensional, unidirectional, advection equation. We then describe how to introduce a covolume-based filter which allows for larger steps when time stepping the discretized equations.

2.1. Notation. The spatial interval $\Omega = [a, b]$ is divided into a primal grid of K^1 elements, $T_h^1 = \bigcup_{k=1}^{k=K^1} I^{1,k}$, with $I^{1,k} = \left[x^{1,k}, x^{1,k+1}\right]$. We associate the following broken Sobolev spaces with the primal grid:

$$H^m\left(\Omega, T_h^1\right) = \left\{q \in L^2(\Omega) : q|_{I^{1,k}} \in H^m\left(I^{1,k}\right) \text{ for } k = 1, \ldots, K^1\right\}, \tag{2.1}$$

where

$$H^m\left(I^{1,k}\right) = \left\{q \in L^2\left(I^{1,k}\right) : \sum_{\alpha=0}^{\alpha=m} \left\| \frac{\partial^\alpha q}{\partial x^\alpha} \right\|^2_{L^2(I^{1,k})} < \infty \right\}. \tag{2.2}$$

In each element the solution will be represented by polynomials of maximum degree n. The numerical solution on the primal grid is chosen from

$$V^{1,h} = \bigcup_{k=1}^{k=K^1} P^n\left(I^{1,k}\right), \tag{2.3}$$

with no a priori assumption about the continuity of the solution. We also require a covolume mesh whose elements are formed with endpoints in the interior of the primal grid. For definiteness we consider here the special case where these are chosen to be the midpoints of the subintervals of the primal grid, but as seen below such a precise choice is unnecessary. The covolume mesh then consists of K^2 elements, $\left\{I^{2,k}\right\}_{k=1}^{k=K^2}$, with elements $I^{2,k} = [\frac{x^{1,k-1}+x^{1,k}}{2}, \frac{x^{1,k}+x^{1,k+1}}{2}]$ for elements not on the boundary. We will discuss how the covolume mesh is formed at the boundary in section 7.

2.2. Upwind DG derivative operator for advection. We are interested in solving advection and wave type phenomena and consequently consider the prototype one-way advection equation

$$\frac{\partial u}{\partial t} + \frac{\partial u}{\partial x} = 0 \quad \text{for } (x,t) \in [0,2) \times [0,1]\,, \tag{2.4}$$

with periodic boundary conditions $u(2,t) = u(0,t)$.

In this paper we consider the upwind DG discretization of this equation, because of its weighted residual formulation which has guaranteed stability, a block diagonal mass matrix, and potentially high-order accuracy. The upwind DG variational statement involves finding $u \in V^{1,h}$ such that

$$\left(\phi, \frac{\partial u}{\partial t}\right)_{I^{1,k}} + \left(\phi, \frac{\partial u}{\partial x}\right)_{I^{1,k}} + \left(\phi, \frac{n_x - |n_x|}{2}\,[u]\right)_{\partial I^{1,k}} = 0 \tag{2.5}$$

for all $\phi \in V^{1,h}$ and $k = 1, \ldots, K^1$, where the notation $[u]$ denotes the trace of u from inside element k subtracted from the trace of u taken from within the element sharing a boundary vertex with element k. In familiar notation $[u] = u^+ - u^-$, where u^- is the internal trace of u in element k and u^+ is the external trace. Here n_x is the outward facing normal at the element boundary, and the boundary inner product is defined by

$$(f,g)_{\partial I^{1,k}} = f\left(x^{1,k}\right) g\left(x^{1,k}\right) + f\left(x^{1,k-1}\right) g\left(x^{1,k-1}\right). \tag{2.6}$$

We introduce a slightly more general notation which allows us to choose between an upwind boundary treatment and a central treatment via the operator D^1_α, where $\alpha \in [0,1]$ is a parameter which interpolates between each type of treatment. Formally D^1_α is the operator which satisfies for any $u \in V^{1,h}$

$$\left(\phi, D^1_\alpha u\right)_{I^{1,k}} = \left(\phi, \frac{\partial u}{\partial x}\right)_{I^{1,k}} + \left(\phi, \frac{n_x - \alpha|n_x|}{2}\,[u]\right)_{\partial I^{1,k}} \tag{2.7}$$

for $k = 1, \ldots, K^1$ and $\phi \in V^{1,h}$. Now we succinctly represent the upwind DG variational statement by

$$\left(\phi, \frac{\partial u}{\partial t}\right)_{I^{1,k}} + \left(\phi, D^1_\alpha u\right)_{I^{1,k}} = 0. \tag{2.8}$$

This notation highlights the significant fact that despite its apparent complexity, the DG operator is actually just the result of encoding a distributional derivative to account for boundary conditions introduced at the endpoints of each element.

After discretization of the $V^{1,h}$ variational space of functions defined with respect to the primal grid, the advection equation is reduced from a partial differential equation to an ordinary differential equation:

$$\frac{d\mathbf{u}^1}{dt} = -\mathbf{D}^1_\alpha \mathbf{u}^1, \tag{2.9}$$

where $\mathbf{D}^1_\alpha$ is the upwind DG derivative matrix defined on the primal grid by

$$\left(\mathbf{D}^1_\alpha\right)_{ij} = \sum_k \left(\mathbf{M}^{1,k}\right)^{-1}_{in} \left(\left(\phi_n, \frac{\partial \phi_j}{\partial x}\right)_{I^{1,k}} + \left(\phi_n, \frac{n_x - \alpha|n_x|}{2}\,[\phi_j]\right)_{\partial I^{1,k}}\right), \tag{2.10}$$

where $\mathbf{M}^{1,k}_{nm} = (\phi_n, \phi_m)_{I^{1,k}}$. It is useful to view the $\mathbf{D}^1_\alpha$ matrix as representing a discrete distributional derivative operator, and thus later on it will be transparent that modifying this operator by filtering matrices is not a complicated process. Once we establish the method, then we will appeal to the variational statement to establish stability, but in terms of implementation it is most instructive to think in terms of the action of matrices on the solution.

We can immediately establish L^2 stability with the following result:

$$\frac{1}{2}\frac{d}{dt}\sum_k \left(\mathbf{u}^1\right)^t \mathbf{M}^{1,k}\mathbf{u}^1 = -\frac{1}{2}\sum_k \left(\alpha|n_x|\left[u^1\right],\left[u^1\right]\right)_{\partial I^{1,k}} \leq 0, \tag{2.11}$$

which states succinctly that the choice of upwinding the solution at the endpoints of the elements before taking the DG distributional derivatives acts to reduce the L^2 energy of the numerical solution whenever there is a jump in the solution between elements.

We next introduce the discretized form of the transfer operators. These are the matrices $\mathbf{\Pi}^2$, $\mathbf{\Pi}^1$ defined by

$$\left(\phi^2_i, \phi^2_n\right)_{T^2_h} \mathbf{\Pi}^2_{nj} = \sum_k \left(\left(\phi^2_i, \phi^1_j\right)_{I^{1,k}\cap I^{2,k}} + \left(\phi^2_i, \phi^1_j\right)_{I^{1,k-1}\cap I^{2,k}}\right), \tag{2.12}$$

$$\left(\phi^1_i, \phi^1_n\right)_{T^1_h} \mathbf{\Pi}^1_{nj} = \sum_k \left(\left(\phi^1_i, \phi^2_j\right)_{I^{1,k}\cap I^{2,k}} + \left(\phi^1_i, \phi^2_j\right)_{I^{1,k+1}\cap I^{2,k}}\right) \tag{2.13}$$

for all basis members $\left\{\phi^2_i\right\}_{i=1}^{i=N_p}$ on the K covolume mesh elements and all basis members $\left\{\phi^1_i\right\}_{i=1}^{i=N_p}$ on the K primal grid elements. It is immediately observed that these transfer matrices are related by

$$\mathbf{M}^2\mathbf{\Pi}^2 = \left(\mathbf{M}^1\mathbf{\Pi}^1\right)^t. \tag{2.14}$$

2.3. Modified upwind DG method. We now have sufficient notation to propose a new type of DG method which we write in pseudocode.

1. Evaluate initial condition on primal grid.
2. Start time stepping, with an s-stage RK scheme.
3. For each RK stage do the following:
   (a) L^2 project the solution from primal grid to covolume mesh.
   (b) Compute DG distributional derivative on covolume mesh, using upwind fluxes suitable for the advection equation.
   (c) L^2 project the DG distributional derivative to original grid.
   (d) Update solution for RK stage using L^2 projected DG distributional derivative.
4. Continue time stepping until final time reached.

We have added two projection stages. The first stage is designed to project the original piecewise polynomial solution on the primal grid onto the covolume mesh. The purpose of this projection is to diminish the maximum gradients achieved on the covolume mesh elements. We then evaluate the DG distributional gradient on the covolume mesh and project it back on to the primal grid, and again our aim is to reduce the potential for large near boundary gradients.

In matrix notation we are solving the following ordinary differential equation:

$$\frac{d\mathbf{u}}{dt} = -\mathbf{\Pi}^1\mathbf{D}^2_\alpha\mathbf{\Pi}^2\mathbf{u}, \tag{2.15}$$

where the DG derivative matrix $\mathbf{D}_\alpha^2$ is defined on the covolume mesh. We can also relax the stabilization by averaging the stabilized DG derivative with the unstabilized DG derivative:

$$\frac{d\mathbf{u}}{dt} = -\beta \mathbf{\Pi}^1 \mathbf{D}_\alpha^2 \mathbf{\Pi}^2 \mathbf{u} - (1-\beta)\mathbf{D}_\alpha^1 \mathbf{u}, \tag{2.16}$$

where $\beta = 1 - O\left(\frac{1}{n}\right)$. It is straightforward to perform the L^2 stability analysis for this modified DG scheme, with the L^2 energy equation given by

$$\begin{aligned} \frac{d}{dt}\left\|u^1\right\|^2_{L^2(T_h^1)} \leq & -\frac{\beta}{2}\sum_k \left(\alpha|n_x|\left[\Pi^2 u^1\right],\left[\Pi^2 u^1\right]\right)_{\partial I^{2,k}} \\ & -\frac{(1-\beta)}{2}\sum_k \alpha\left(|n_x|\left[u^1\right],\left[u^1\right]\right)_{\partial I^{1,k}}. \end{aligned} \tag{2.17}$$

Thus the numerical energy decreases if there are jumps in u^1 between elements on the primal grid and if there are jumps in $\Pi^2 u^1$ between elements on the covolume mesh. (We will subsequently observe that the filtering process creates an intermediate field with much reduced jumps.)

We discuss further minor variations on this idea in section 6 to address the computational overhead of the additional operations required to compute the filtered solution.

2.4. Motivation for using a covolume mesh based filter. The traditional DG discretization of a first-order distributional derivative has a maximum eigenvalue which asymptotically grows as $O(\frac{n^2}{h})$ when the order of polynomials used, n, is increased or h is decreased (see, e.g., [14, 12]). In the particular case of *hp* type methods, where it is considered desirable to use a high-order polynomial approximation when possible, the derivative operators become artificially stiff. This artificial stiffness is the result of the nature of the piecewise polynomial discretization spaces used. In this section we detail the causes behind the scaling and present a revealing observation that we use to build the modified DG methods.

We first recall the following polynomial inequalities (see, for example, Borwein and Erdélyi [3] for details and generalizations).

THEOREM 2.1 (Bernstein's inequality). *The inequality*

$$\left|\frac{dq}{dx}(x)\right| \leq \frac{n}{\sqrt{1-x^2}}\left\|q\right\|_\infty, \quad -1 < x < 1,$$

holds for every $q \in P^n([-1,1])$.

THEOREM 2.2 (Markov's inequality). *The inequality*

$$\left\|\frac{dq}{dx}\right\|_\infty \leq n^2 \left\|q\right\|_\infty$$

holds for every $q \in P^n([-1,1])$.

We put these together and observe that

$$\left|\frac{dq}{dx}(x)\right| \leq \min\left(n^2, \frac{n}{\sqrt{1-x^2}}\right)\left\|q\right\|_\infty, \quad -1 \leq x \leq 1, \tag{2.18}$$

and since these estimates are sharp, the gradients of normalized nth-order polynomials may be up to $O(\frac{n^2}{h})$ at the near boundary portion of the element and $O\left(\frac{n}{h}\right)$

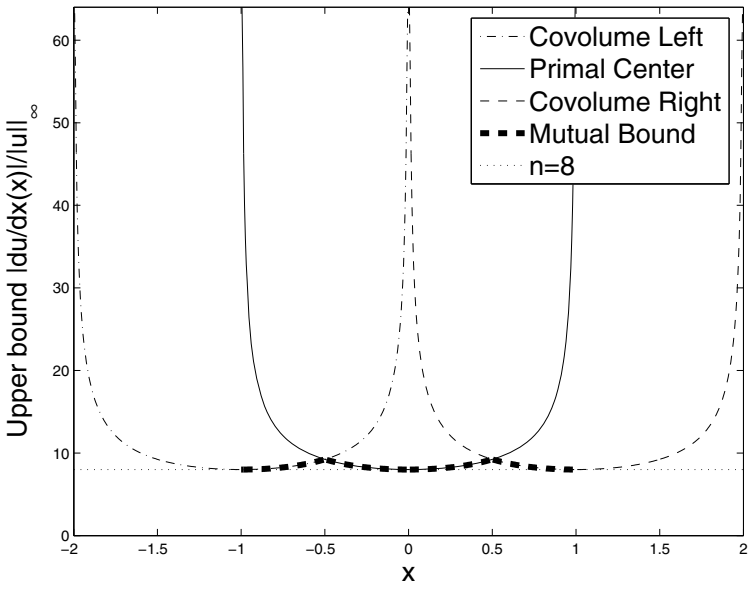


FIG. 2.1. *Combining upper bounds on the relative maximal gradient of polynomials in P^8 for primal and covolume mesh yields a more uniform bound.*

near the center of the element. These inequalities explain the difference between the Courant–Friedrichs–Lewy (CFL) conditions of a typical finite difference method, where a moving interpolation stencil is used and slope information is generally extracted at element centers, and the DG method, where slope information is used throughout.

In this paper we will show how it is possible to achieve CFL numbers within a modified upwind DG method comparable with typical finite difference discretizations, even when high-order polynomial approximations are used. In Figure 2.1, we plot the upper bound (2.18) for three biunit one-dimensional elements which each overlap by half their lengths with their neighbors. This upper bound on the scaled gradient of polynomials grows dramatically near each of the element boundaries. However, we have also highlighted the minimum upper bound of all three in the central element, and this curve is almost constant throughout this element. This observation yields the possibility that we may be able to overcome the artificial stiffness induced in the DG scheme by the nonuniform gradients of polynomial approximations to fields. In the next sections we prove that the covolume derivative operator, $\mathbf{\Pi}^1\mathbf{D}_\alpha^2\mathbf{\Pi}^2$, motivated by these inequalities, in fact satisfies the desired bound.

It is also necessary to study the properties of the projection-based filter that is implicit in our formulation. We note that these projection operators are both contractions, i.e., $\|\Pi^2 u\| \leq \|u\|$ and $\|\Pi^1 u\| \leq \|u\|$. Furthermore, we consider the action of the operator defined by $\Pi^2\Pi^1$. This operator is a filter, and in subsequent sections we establish its properties by deriving bounds on $\|u - \Pi^2\Pi^1 u\|$.

It is instructive to see the consequence of projecting the polynomial lift of the Dirac delta used in the distributional derivative at the center of the DG formulation from the primal grid to the covolume mesh and back again. In Figure 2.2, we show the $n = 8$ polynomial lift of the Dirac delta before and after being filtered by the operator $\beta\Pi^1\Pi^2 + (1 - \beta)$ with $\beta = 1 - \frac{2}{n+1}$ used as a relaxation parameter for the

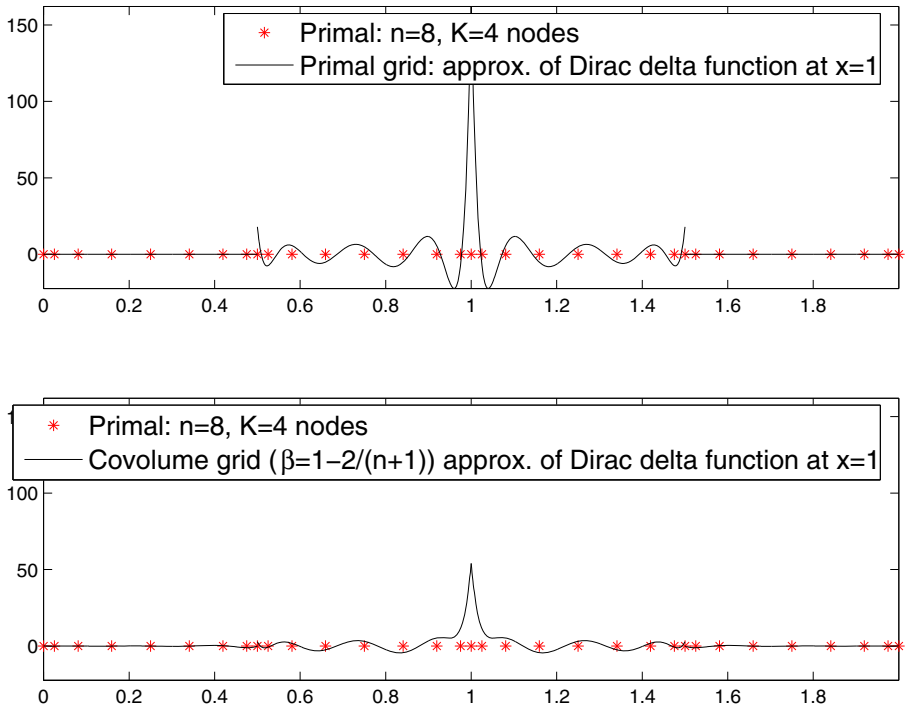


FIG. 2.2. *The polynomial approximation of the Dirac delta used to lift jump data at element boundaries into their interior for* $n = 8$. *Top: without filtering. Bottom: with filtering* $(\beta = 1 - (2/(n+1))$.

amount of filtering to be applied. It is clear that this type of filtering will reduce the contribution to the DG gradient from the jump terms besides de-emphasizing gradients at and near the element boundaries. Taken together, these effects of the filtering are our motivation for using the filtering process to reduce the artificially stiff DG operators and hopefully increase the allowable time step size.

3. Bounds on the covolume derivative operator. Here we prove our main result on the n-dependence of the covolume derivative operator. We work in the framework described above, periodically extending functions beyond $[0, 2)$.

THEOREM 3.1. *Let* q *be a degree* n *piecewise polynomial function relative to the primal grid,* $q \in V^{1,h}$, *and set, for* $\alpha \in [0, 1]$,

$$\mathcal{D}_{S,\alpha} q = \Pi^1 \mathbf{D}^2_\alpha \Pi^2 q. \tag{3.1}$$

Then there exists a universal constant, C, *such that*

$$\|\mathcal{D}_{S,\alpha} q\|_{L^2(\Omega)} \leq C \frac{n}{h} \|q\|_{L^2(\Omega)}. \tag{3.2}$$

Before proving Theorem 3.1 we note that as an immediate corollary we have

$$\| \left(\beta \mathcal{D}_{S,\alpha} + (1-\beta)\mathbf{D}^1_\alpha\right) q\|_{L^2(\Omega)} \leq C \left(\beta \frac{n}{h} + (1-\beta)\frac{n^2}{h}\right) \|q\|_{L^2(\Omega)}, \tag{3.3}$$

so that the desired bound for the matrix appearing in (2.16) is obtained if we take $\beta = 1 - O\left(\frac{1}{n}\right)$.

3.1. Proof of Theorem 3.1. We proceed by proving a number of lemmas. We begin with an L^2 version of Bernstein's inequality.

LEMMA 3.2 (L^2 version of Bernstein's inequality). *The inequality*

$$\left\| \frac{dq}{dx} \right\|_{L^2([-a,a])} \leq C_a n \left\| q \right\|_{L^2([-1,1])}, \tag{3.4}$$

where $C_a = \frac{\sqrt{3}+a}{1-a^2}$, *holds for every* $q \in P^n([-1,1])$.

Proof. Without loss of generality we assume $\left\| q \right\|_{L^2([-1,1])} = 1$. We next represent q using Legendre polynomials:

$$q(x) = \sum_{j=0}^{j=n} c_j P_j(x), \tag{3.5}$$

where we are assuming the standard definition of P_j (see, for example, Sansone [29, Chapter III]). With this normalization, the Legendre polynomials satisfy the following recurrence relationship:

$$\frac{1}{n}\frac{dP_j}{dx}(x) = \frac{1}{1-x^2}P_{j-1}(x) - \frac{x}{1-x^2}P_j(x), \quad j \geq 1. \tag{3.6}$$

Thus

$$\frac{1}{n}\frac{dq}{dx}(x) = \frac{1}{1-x^2}w_1(x) - \frac{x}{1-x^2}w_2(x), \tag{3.7}$$

where

$$w_1(x) = \sum_{j=1}^{j=n} c_j P_{j-1}(x), \quad w_2(x) = \sum_{j=1}^{j=n} c_j P_j(x). \tag{3.8}$$

We then use the triangle inequality and the fact that $(2j+1)/(2j-1) \leq 3$ for $j \geq 1$:

$$\begin{aligned}
\frac{1}{n}\left\| \frac{dq}{dx} \right\|_{L^2([-a,a])} &\leq \max_{|x|\leq a}\left(\frac{1}{1-x^2}\right)\left\| w_1 \right\|_{L^2([-a,a])} + \max_{|x|\leq a}\left(\frac{|x|}{1-x^2}\right)\left\| w_2 \right\|_{L^2([-a,a])} \\
&\leq \frac{1}{1-a^2}\left\| w_1 \right\|_{L^2([-1,1])} + \frac{a}{1-a^2}\left\| w_2 \right\|_{L^2([-1,1])} \\
&\leq \frac{1}{1-a^2}\left(\sum_{j=1}^{j=n}\left(j-\frac{1}{2}\right)^{-1}c_j^2\right)^{\frac{1}{2}} + \frac{a}{1-a^2}\left(\sum_{j=1}^{j=n}\left(j+\frac{1}{2}\right)^{-1}c_j^2\right)^{\frac{1}{2}} \\
&\leq \frac{\sqrt{3}+a}{1-a^2}.
\end{aligned}$$

Noting the normalization of q completes the proof. ∎

Just as with Bernstein's lemma we see the upper bound degenerates when considering the whole interval. We have made no effort to optimize the coefficient C_a, and numerical computations of the bound for $1 \leq n \leq 512$ with $a = 1/2$ reveal that in this case the constant is only suboptimal by approximately a factor of 2.43.

Later on we will also require a polynomial inverse inequality, which is an L^2 analogue of Markov's theorem.

LEMMA 3.3 (inverse inequality). *Assume* $u \in P^n\left(I^{1,k}\right)$*; then*

$$\left\|\frac{\partial u}{\partial x}\right\|_{L^2([-1,1])} \leq \sqrt{3}n^2 \left\|u\right\|_{L^2([-1,1])}.$$

Proof. See, for example, Ciarlet [6] or Schwab [30]. □

In addition to Lemma 3.2 we will need bounds on the pointwise values of degree n polynomials. Precisely we prove the following lemma.

LEMMA 3.4. *The inequality*

$$|q(a)| \leq M_a \sqrt{n} \left\|q\right\|_{L^2([-1,1])}, \tag{3.9}$$

where $M_a = \frac{12}{\sqrt{\pi}}(1-a^2)^{-1/4}$*, holds for every* $q \in P^n([-1,1])$ *with* $n \geq 1$.

(We note that if we apply the result on an interval of width h, the right-hand side scales like $h^{-\frac{1}{2}}$.)

Proof. We again employ the Legendre expansion (3.5) and recall Stieltjes' bound [29, Section III.10]:

$$|P_j(a)| \leq 4\sqrt{\frac{2}{\pi}}(1-a^2)^{-1/4} j^{-1/2}, \tag{3.10}$$

which can obviously be improved by setting the right-hand side to 1 if $j = 0$. We then have using the Cauchy–Schwarz inequality

$$\begin{aligned}
|q(a)| &\leq \sum_{j=0}^{n} |c_j||P_j(a)| \\
&\leq 6\sqrt{\frac{2}{\pi}}(1-a^2)^{-1/4} \sum_{j=0}^{n} \left(j+\frac{1}{2}\right)^{-1/2} |c_j| \\
&\leq 6\sqrt{\frac{2}{\pi}}(1-a^2)^{-1/4} \sqrt{n+1} \left(\sum_{j=0}^{n} \left(j+\frac{1}{2}\right)^{-1} c_j^2\right)^{1/2} \\
&\leq M_a \sqrt{n} \left\|q\right\|_{L^2([-1,1])}. \qquad \square
\end{aligned}$$

Finally, we will make use of the following explicit expression for the lift polynomial, $L(x)$.

LEMMA 3.5. *The polynomial*

$$\begin{aligned}
L^{(k,2)}(x) = (-1)^n \frac{d}{dx}\left(\left(\frac{2(x-x^{1,k})-h}{2h}\right) P_n\left(\frac{2}{h}(x-x^{1,k})\right)\right) \\
+ (-1)^n \frac{4n+1}{2h} P_n\left(\frac{2}{h}(x-x^{1,k})\right)
\end{aligned} \tag{3.11}$$

satisfies

$$\int_{I^{2,k}} L^{(k,2)}(x) Q(x) dx = Q\left(x^{1,k} - \frac{h}{2}\right) \tag{3.12}$$

for all polynomials, Q*, of degree* n*.*

Proof. With the change of variables $z = \frac{2}{h}(x - x^{1,k})$ the statement becomes

$$\int_{-1}^{1} \tilde{L}(z)\tilde{Q}(z)dz = \tilde{Q}(-1) \tag{3.13}$$

for all degree-n polynomials $\tilde{Q}(z)$, where

$$\tilde{L}(z) = (-1)^n \left(\frac{d}{dz} \left(\left(\frac{z-1}{2} \right) P_n(z) \right) + \frac{4n+1}{4} P_n(z) \right). \tag{3.14}$$

To prove it we note that using integration by parts the statement is equivalent to

$$\int_{-1}^{1} P_n(z) \left(\frac{1-z}{2} \frac{d\tilde{Q}}{dz}(z) + \frac{4n+1}{4} \tilde{Q}(z) \right) dz = 0. \tag{3.15}$$

Now we can write $\tilde{Q}(z) = R(z) + \gamma P_n(z)$, where the degree of R is $n-1$. By the orthogonality properties of P_n the integrals involving R and $\frac{dR}{dz}$ are zero. Then integrating by parts once more

$$\begin{aligned} \int_{-1}^{1} \frac{1-z}{2} P_n(z) P_n'(z) dz &= -\frac{1}{2} + \frac{1}{4} \int_{-1}^{1} P_n^2(z) dz \\ &= -\frac{4n+1}{4(2n+1)} \\ &= -\frac{4n+1}{4} \int_{-1}^{1} P_n^2(z) dz. \end{aligned}$$

This completes the proof. ∎

With these results we can prove Theorem 3.1. As the proof involves separate estimates of a number of terms we will avoid keeping close track of the constants. Thus C will be used to denote any absolute constant independent of n and h.

Proof of Theorem 3.1. As $\mathcal{D}_S$ is obviously linear we can consider the special case of q nonzero only on $I^{1,k}$ and denote by q_k the degree n polynomial which is the restriction of q to $I^{1,k}$. Now $\tilde{q} = \Pi_n^2 q$ is supported on $I^{2,k} \cup I^{2,k+1}$, and let $\tilde{q}_j$ denote its restriction to $I^{2,j}$. Last, let $Q = \Pi_n^1 \mathbf{D}^2 \tilde{q}$. We will estimate Q in two parts: Q_d, the contribution of the derivatives, $\tilde{q}_j'$, and Q_l, the contribution of the lifts.

Let $Q_{d,j}$ denote the restriction of Q_d to $I^{1,j}$. Now $Q_{d,j}$ is nonzero only for $j = k$, $k \pm 1$. We begin by estimating the contribution of $\tilde{q}_k'$ to $Q_{d,k}$, which we label $Q_{d,k,k}$. Using the properties of L^2 projection we have

$$\begin{aligned} \int_{I^{1,k}} Q_{d,k,k}^2(x) dx &= \int_{x^{1,k}}^{\frac{x^{1,k}+x^{1,k+1}}{2}} Q_{d,k,k}(x) \tilde{q}_k'(x) dx \\ &= \int_{x^{1,k}}^{x^{1,k}+\frac{h}{4}} Q_{d,k,k}(x) \tilde{q}_k'(x) dx - \int_{x^{1,k}+\frac{h}{4}}^{\frac{x^{1,k}+x^{1,k+1}}{2}} Q_{d,k,k}'(x) \tilde{q}_k(x) dx \\ &\quad + Q_{d,k,k} \left(\frac{x^{1,k}+x^{1,k+1}}{2} \right) \cdot \tilde{q}_k \left(\frac{x^{1,k}+x^{1,k+1}}{2} \right) \\ &\quad - Q_{d,k,k} \left(x^{1,k} + \frac{h}{4} \right) \cdot \tilde{q}_k \left(x^{1,k} + \frac{h}{4} \right). \end{aligned}$$

Proceeding to estimate the four terms on the right-hand side above, we see that the first two can be handled by means of Lemma 3.2. That is, each of the derivative terms is integrated in the strict interior of the domain over which the norm of the polynomial can be bounded. Thus, noting that $\tilde{q}_k$ was computed by an L^2 projection and employing the Cauchy–Schwarz inequality we conclude that these terms are bounded by

$$\text{Term}_1 + \text{Term}_2 \leq C\frac{n}{h}\|Q_{k,k}\|_{L^2(I^{1,k})}\|q_k\|_{L^2(I^{1,k})}. \tag{3.16}$$

Similarly, favorable bounds on the fourth term follow from Lemma 3.4 and the fact that $\frac{x^{1,k}+x^{1,k+1}}{4}$ lies in the interior of both $I^{1,k}$ and $I^{2,k}$. Thus

$$\text{Term}_4 \leq C\frac{n}{h}\|Q_{d,k,k}\|_{L^2(I^{1,k})}\|q_k\|_{L^2(I^{1,k})}. \tag{3.17}$$

The third term is more difficult to estimate, as it involves the evaluation of $\tilde{q}_k$ on the boundary of $I^{2,k}$. (The bound $|Q_{d,k,k}(\frac{x^{1,k}+x^{1,k+1}}{2})| \leq C\sqrt{\frac{n}{h}}\|Q_{d,k,k}\|_{L^2(I^{1,k})}$ is a direct consequence of Lemma 3.4.) To complete the argument we rely on the fact that $\tilde{q}_k$ was computed by projecting q_k, which itself is well behaved at the midpoint of $I^{1,k}$. By Christoffel's formula [29, Section III.5] we have

$$\tilde{q}_k\left(\frac{x^{1,k}+x^{1,k+1}}{2}\right) = \int_0^1 \left(P'_{n+1}(y) + P'_n(y)\right) q_k\left(x^{1,k} + \frac{h}{2}y\right) dy.$$

We again break this integral into two pieces and integrate by parts:

$$\begin{aligned}\tilde{q}_k\left(\frac{x^{1,k}+x^{1,k+1}}{2}\right) = &\int_0^{\frac{1}{2}} \left(P'_{n+1}(y) + P'_n(y)\right) q_k\left(x^{1,k} + \frac{h}{2}y\right) dy \\ &- \frac{h}{2}\int_{\frac{1}{2}}^1 \left(P_{n+1}(y) + P_n(y)\right) q'_k\left(x^{1,k} + \frac{h}{2}y\right) dy \\ &+ 2q_k\left(\frac{x^{1,k}+x^{1,k+1}}{2}\right) \\ &- \left(P_{n+1}\left(\frac{1}{2}\right) + P_n\left(\frac{1}{2}\right)\right) q_k\left(x^{1,k} + \frac{h}{4}\right).\end{aligned}$$

Using Lemma 3.2 and the normalization of the Legendre polynomials we bound the first two terms:

$$\text{Term}'_1 + \text{Term}'_2 \leq C\sqrt{\frac{n}{h}}\|q_k\|_{L^2(I^{1,k})}. \tag{3.18}$$

Similarly, by Stieltjes' bound and Lemma 3.4 we bound the second two terms:

$$\text{Term}'_3 + \text{Term}'_4 \leq C\sqrt{\frac{n}{h}}\|q_k\|_{L^2(I^{1,k})}. \tag{3.19}$$

Combining (3.18) and (3.19) we have

$$\text{Term}_3 \leq C\frac{n}{h}\|Q_{d,k,k}\|_{L^2(I^{1,k})}\|q_k\|_{L^2(I^{1,k})}, \tag{3.20}$$

which combined with (3.16) and (3.17) yields

$$\|Q_{d,k,k}\|_{L^2(I^{1,k})} \leq C\frac{n}{h}\|q_k\|_{L^2(I^{1,k})}, \tag{3.21}$$

with an analogous bound obviously valid for the contribution of $\tilde{q}'_{k+1}$.

Finally, we must estimate $\|Q_{d,k\pm1}\|_{L^2(I^{1,k\pm1})}$. These follow along the same lines as the estimates above. For example,

$$\begin{aligned}\int_{I^{1,k+1}} Q^2_{d,k+1}(x)dx &= \int_{x^{1,k+1}}^{\frac{x^{1,k+1}+x^{1,k+2}}{2}} Q_{d,k+1}(x)\tilde{q}'_{k+1}(x)dx \\ &= \int_{x^{1,k+1}}^{x^{1,k+1}+\frac{h}{4}} Q_{d,k+1}(x)\tilde{q}'_{k+1}(x)dx \\ &\quad - \int_{x^{1,k+1}+\frac{h}{4}}^{\frac{x^{1,k+1}+x^{1,k+2}}{2}} Q'_{d,k+1}(x)\tilde{q}_{k+1}(x)dx \\ &\quad + Q_{d,k+1}\left(\frac{x^{1,k+1}+x^{1,k+2}}{2}\right)\cdot\tilde{q}_{k+1}\left(\frac{x^{1,k+1}+x^{1,k+2}}{2}\right) \\ &\quad - Q_{d,k+1}\left(x^{1,k+1}+\frac{h}{4}\right)\cdot\tilde{q}_{k+1}\left(x^{1,k+1}+\frac{h}{4}\right).\end{aligned}$$

Estimates of the first, second, and fourth terms are exactly analogous to (3.16) and (3.17). For the third term we now still have $|Q_{d,k+1}(\frac{x^{1,k+1}+x^{1,k+2}}{2})| \le C\sqrt{\frac{n}{h}}$ $\|Q_{d,k+1}\|_{L^2(I^{1,k+1})}$ by Lemma 3.4. In this case Christoffel's formula yields

$$\begin{aligned}\tilde{q}_{k+1}\left(\frac{x^{1,k+1}+x^{1,k+2}}{2}\right) &= \int_{-1}^{0}\left(P'_{n+1}(y)+P'_n(y)\right)q_k\left(x^{1,k+1}+\frac{h}{2}y\right)dy \\ &= \int_{-\frac{1}{2}}^{0}\left(P'_{n+1}(y)+P'_n(y)\right)q_k\left(x^{1,k+1}+\frac{h}{2}y\right)dy \\ &\quad - \frac{h}{2}\int_{-1}^{-\frac{1}{2}}\left(P_{n+1}(y)+P_n(y)\right)q'_k\left(x^{1,k+1}+\frac{h}{2}y\right)dy \\ &\quad + \left(P_{n+1}\left(-\frac{1}{2}\right)+P_n\left(-\frac{1}{2}\right)\right)q_k\left(x^{1,k+1}-\frac{h}{4}\right).\end{aligned}$$

Again all of these terms can be estimated as above. With the analogous estimate for $Q_{d,k-1}$ we finally conclude that

$$\sum_{j=k-1}^{j=k+1}\|Q_{d,j}\|^2_{L^2(I^{1,k-1})} \le C^2\frac{n^2}{h^2}\|q_k\|^2_{L^2(I^{1,k})}. \tag{3.22}$$

Summing (3.22) over k and taking the square root shows that the contribution of Q_d satisfies the desired inequality.

We now consider Q_l. We note first that the jumps themselves are bounded by $C\sqrt{\frac{n}{h}}\|q_k\|_{L^2(I^{1,k})}$ following (3.18) and (3.19). Thus we need only estimate projections of the lift polynomials, $L^{(k,2)}(x)$. Now the undifferentiated (second) term in (3.11) has L^2 norm $O(\sqrt{\frac{n}{h}})$, and thus so does its projection. The projection of the differentiated term is then estimated in precisely the same method as the derivatives above. In particular, we estimate the projection of the derivative term of $L^{(k,2)}$ on $I^{(1,k)}$ and $I^{(1,k+1)}$, which repeating the arguments given above yield bounds of the form

$$O\left(\frac{n}{h}\right)\|P_n\left(\frac{2}{h}(x-x^{1,k})\right)\|_{I^{2,k}} + O\left(\sqrt{\frac{n}{h}}\right) = O\left(\sqrt{\frac{n}{h}}\right). \tag{3.23}$$

Therefore the contributions of the lift terms are also bounded by $C\frac{n}{h}\|q_k\|_{L^2(I^{1,k})}$, completing the proof of the theorem. ∎

3.2. Implications for time stepping. In the numerical experiments which follow we employ an nth-order RK–Taylor method suitable for linear autonomous systems such as Maxwell's equations in piecewise uniform media. We find experimentally that with this time-stepping procedure we can use a time step independent of n. However, we as yet lack a complete proof of this fact, which would require a detailed analysis of the stability regions of the RK–Taylor methods. Instead we state more general results which hold for a large class of standard RK methods applicable to a less restrictive class of systems.

Consider now the symmetric-hyperbolic system with smooth, uniformly bounded, 2-periodic coefficients, initial data, and forcing:

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left(A(x,t)u\right) + B(x,t)u + F(x,t), \quad u(x,0) = u_0(x). \tag{3.24}$$

We recall the results of Kreiss and Wu [23]. An RK method is called locally stable if there exists $R > 0$ such that the open half circle $\{z||z| < R, \Re z < 0\}$ is contained within its stability domain. Note that the standard third- and fourth-order RK methods as well as certain higher-order Dormand–Prince pairs are locally stable [22]. We then have the following theorem:

THEOREM 3.6. *Semidiscretize* (3.24) *on* $V^{1,h}$ *using the covolume differentiation operator* $\mathcal{D}_{S,\alpha}$ *and discretize in time using a locally stable RK method. Then there exists a constant* C *independent of* h *and* n *such that the fully discrete method is stable for time steps,* dt*, satisfying*

$$dt \le C\frac{h}{n}. \tag{3.25}$$

Proof. The result is a direct consequence of [23], the L^2 stability of the semidiscrete approximation, and Theorem 3.1. In particular for $F = 0$ the solution of the semidiscrete problem will satisfy

$$\|u_h\|_{L^2(T_h^1)} \le Ce^{\eta t}\|u_0^h\|_{L^2(T_h^1)}, \tag{3.26}$$

which is all that is needed to apply the Kreiss–Wu theory. ∎

Note that the RK–Taylor methods used below are locally stable only for $n = 4k$ and $n = 4k - 1$.

4. Semidiscrete convergence analysis. In this section we prove a suboptimal error bound for the proposed covolume method. We note that, in comparison with error estimates for the standard approach, our estimates are $O(h^{1/2})$ worse. We suspect that, in fact, the proposed method is as accurate as the standard one. We show below that this is true in the constant coefficient case on uniform grids; then we can use Fourier methods as in [18] to show that the error is $O(h^{n+1})$. Our numerical experiments indicate that the modified method is also as accurate as the standard method for nonuniform grids. However, direct application of the method of proof from, e.g., [10] fails. We plan to consider alternative analyses in the future.

To simplify the presentation we consider the $\beta = 1$ fully filtered case, as the extension to the partially filtered case is straightforward. We also consider only the semidiscrete scheme.

THEOREM 4.1. *Suppose u satisfies (2.4) with $u(x,0) = u_0(x) \in H^{r+1}$. Then for all $0 \le s \le \min(n,r)$ there exist constants $C_{n,s}$ such that the solution, u_h, of the semidiscrete problem satisfies the following error estimate:*

$$\begin{aligned}\|u-u_h\|_{L^2(\Omega)} \le C_{n,s} &\sum_{k=1}^{k=K^1} \left(h^{1,k}\right)^{s+1} \|u\|_{H^{s+1}(I^{1,k})} \\ &+ C_{n,s} \sum_{k=1}^{k=K^1} \left(h^{1,k}\right)^{s} T \max_{t\in[0,T]} \left(\|u\|_{H^{s+1}(I^{1,k})}\right) \\ &+ C_{n,s} \sum_{k=1}^{k=K^2} \left(h^{2,k}\right)^{s+1} \|u\|_{H^{s+1}(I^{2,k})} \\ &+ C_{n,s} \sum_{k=1}^{k=K^2} \left(h^{2,k}\right)^{s} T \max_{t\in[0,T]} \left(\|u\|_{H^{s+1}(I^{2,k})}\right).\end{aligned}$$

We begin with a sequence of lemmas estimating the errors induced by the filter. We recall the basic polynomial approximation results on the reference interval, $D = [-1,1]$ (see, for example, Schwab [30] for a comprehensive summary). The first result bounds the L^2 projection error incurred by nth-order polynomial approximation of a function.

LEMMA 4.2 (L^2 polynomial approximation). *Assume that $u \in H^{r+1}\left(I^{1,k}\right)$, $r \ge 0$; then, for every integer s such that $0 \le s \le \min(n+1, r+1)$, the following estimate holds:*

$$\left\|u - \Pi_n^1 u\right\|_{L^2(I^{1,k})} \le C_s \frac{h^s}{(n+1)^s} \left\|\frac{\partial^s u}{\partial x^s}\right\|_{L^2(I^{1,k})}. \tag{4.1}$$

Proof. See Ciarlet [6]. □

Using Lemma 4.2, we are able to establish the following result.

LEMMA 4.3 (bound on filter induced error). *Assume that*

$$u \in H^{r+1}\left(\Omega, T_h^1\right) \bigcap H^{r+1}\left(\Omega, T_h^2\right);$$

then, for every integer s such that $0 \le s \le \min(n+1, r+1)$, the following inequality holds:

$$\left\|u - \Pi^2\Pi^1 u\right\|_{L^2(\Omega)} \le C_{n,s} \sum_{k=1}^{k=K^1} \left(h^{1,k}\right)^s \|u\|_{H^s(I^{1,k})} + C_{n,s} \sum_{k=1}^{k=K^2} \left(h^{2,k}\right)^s \|u\|_{H^s(I^{2,k})}.$$

Proof. Starting from the definition of $\Pi^2\Pi^1$

$$\begin{aligned}\left\|u - \Pi^2\Pi^1 u\right\|_{L^2(\Omega)} &= \left\|u - \Pi^2 u + \Pi^2 u - \Pi^2\Pi^1 u\right\|_{L^2(\Omega)}, \\ &\le \left\|u - \Pi^2 u\right\|_{L^2(\Omega)} + \left\|\Pi^2 u - \Pi^2\Pi^1 u\right\|_{L^2(\Omega)}, \\ &\le \left\|u - \Pi^2 u\right\|_{L^2(\Omega)} + \left\|\Pi^2\left(u - \Pi^1 u\right)\right\|_{L^2(\Omega)}, \\ &\le \left\|u - \Pi^2 u\right\|_{L^2(\Omega)} + \left\|u - \Pi^1 u\right\|_{L^2(\Omega)},\end{aligned}$$

where the last result follows from the contraction property of Π^1. Finally, we observe that each of the two terms on the right-hand side are the remainder terms from

performing an L^2 projection mapping u to piecewise polynomial spaces and apply Lemma 4.2. ∎

We now bound the L^2 norm of the derivative of the error induced by the filter. Recall the following bound for the first derivative of the error incurred in the L^2 projection of a function in terms of its derivatives.

LEMMA 4.4. *Assume that* $u \in H^{r+1}\left(I^{1,k}\right)$, $r \geq 0$; *then, for* $n \geq 1$, *and* $0 \leq s \leq \min(n, r)$, *the following estimate holds:*

$$\left\| \frac{\partial}{\partial x}\left(u - \Pi_n^1 u\right) \right\|_{L^2(I^{1,k})} \leq C_{n,s}\left(h^{1,k}\right)^s \left\| \frac{\partial^{s+1} u}{\partial x^{s+1}} \right\|_{L^2(I^{1,k})}. \tag{4.2}$$

Proof. See Ciarlet [6]. ∎

We use this bound to establish the following minor variation of Lemma 4.3, which bounds the derivative of the filtered component of a function.

LEMMA 4.5 (bound on the derivative of filter induced error). *The inequality*

$$\begin{aligned}\left\| \frac{\partial}{\partial x}\left(u - \Pi^2\Pi^1 u\right) \right\|_{L^2\left(T_h^2\right)} \leq\ & C_{n,s} \sum_{k=1}^{k=K^1} \left(h^{1,k}\right)^s \|u\|_{H^{s+1}(I^{1,k})} \\ & + C_{n,s} \sum_{k=1}^{k=K^2} \left(h^{2,k}\right)^s \|u\|_{H^{s+1}(I^{2,k})}\end{aligned}$$

holds for all $u \in H^{s+1}\left(\Omega, T_h^1\right) \bigcap H^{s+1}\left(\Omega, T_h^2\right)$ *when* $0 \leq s \leq \min(n, r)$.

Proof. Starting from the definition of $\Pi^2\Pi^1$

$$\begin{aligned}\left\| \frac{\partial}{\partial x}\left(u - \Pi^2\Pi^1 u\right) \right\|_{L^2\left(T_h^2\right)} &= \left\| \frac{\partial}{\partial x}\left(u - \Pi^2 u\right) + \frac{\partial}{\partial x}\left(\Pi^2 u - \Pi^2\Pi^1 u\right) \right\|_{L^2\left(T_h^2\right)}, \\ &\leq \left\| \frac{\partial}{\partial x}\left(u - \Pi^2 u\right) \right\|_{L^2\left(T_h^2\right)} + \left\| \frac{\partial}{\partial x}\left(\Pi^2 u - \Pi^2\Pi^1 u\right) \right\|_{L^2\left(T_h^2\right)}, \\ &\leq \left\| \frac{\partial}{\partial x}\left(u - \Pi^2 u\right) \right\|_{L^2\left(T_h^2\right)} + \frac{n^2\sqrt{3}}{h} \left\| \Pi^2 u - \Pi^2\Pi^1 u \right\|_{L^2\left(T_h^2\right)}, \\ &\leq \left\| \frac{\partial}{\partial x}\left(u - \Pi^2 u\right) \right\|_{L^2\left(T_h^2\right)} + \frac{n^2\sqrt{3}}{h} \left\| u - \Pi^1 u \right\|_{L^2\left(T_h^1\right)},\end{aligned}$$

where we have used the the inverse inequality (3.3) and the contraction property of Π^2. The approximation bound follows from Lemmas 4.4 and 4.2. ∎

Next we bound the polynomial lift of the interelement jumps of functions filtered through the covolume mesh. Recall the definition of the lift operator on the reference element as the linear operator $\mathcal{L}$ which satisfies, given $u \in L^2(\partial I)$,

$$(\phi, \mathcal{L}u)_I = (\phi, u)_{\partial I} \quad \forall \phi \in L^2(I). \tag{4.3}$$

We first note the multiplicative trace inequality.

LEMMA 4.6 (multiplicative trace inequality). *There exists a constant* C *independent of* q, h, *and* D *such that*

$$\|q\|^2_{L^2(\partial I^{1,k})} \leq C\left(\|q\|_{L^2(I^{1,k})} \left\| \frac{\partial q}{\partial x} \right\|_{L^2(I^{1,k})} + \frac{1}{h^{1,k}} \|q\|^2_{L^2(I^{1,k})} \right), \quad q \in H^1\left(I^{1,k}\right). \tag{4.4}$$

Proof. See Prudhomme et al. [26] for a succinct derivation of this result. ∎

Also we note the polynomial trace inequality.

LEMMA 4.7 (polynomial trace inequality). *Assume* $u \in P^n\left(I^{1,k}\right)$*; then*

$$\|u\|_{L^2(\partial I^{1,k})} \leq \frac{(n+1)}{\sqrt{h^{1,k}}} \|u\|_{L^2(I^{1,k})}. \tag{4.5}$$

Proof. See Warburton and Hesthaven [33] for a constant free derivation for the d-simplex. ∎

These results allow us to prove the following bound on the polynomial lift of the L^2 norm of the jump in a filtered solution u which is sufficiently smooth.

LEMMA 4.8 (bound on the polynomial lift of filter induced jumps). *The inequality*

$$\sum_{k=1}^{k=K^2} \left\|\mathcal{L}\left[\Pi^2\Pi^1 u\right]\right\|_{L^2\left(I_h^{2,k}\right)} \leq C_{n,s} \sum_{k=1}^{k=K^1} \left(h^{1,k}\right)^s \|u\|_{H^{s+1}(I^{1,k})} + C_{n,s} \sum_{k=1}^{k=K^2} \left(h^{2,k}\right)^s \|u\|_{H^{s+1}(I^{2,k})}$$

holds for all $u \in H^{r+1}\left(\Omega, T_h^1\right) \bigcap H^{r+1}\left(\Omega, T_h^2\right)$ *when* $0 \leq s \leq \min(n, r)$.

Proof. By this definition of the lift operator $\mathcal{L}$,

$$\left(\mathcal{L}\left[\Pi^2\Pi^1 u\right], \mathcal{L}\left[\Pi^2\Pi^1 u\right]\right)_{I^{2,k}} = \left(\mathcal{L}\left[\Pi^2\Pi^1 u\right], \left[\Pi^2\Pi^1 u\right]\right)_{\partial I^{2,k}}; \tag{4.6}$$

thus by the Cauchy–Schwarz inequality and the polynomial trace inequality (Lemma 4.7)

$$\left\|\mathcal{L}\left[\Pi^2\Pi^1 u\right]\right\|_{L^2(I^{2,k})} \leq C\frac{n}{\sqrt{h^{2,k}}} \left\|\left[\Pi^2\Pi^1 u\right]\right\|_{L^2(\partial I^{2,k})}. \tag{4.7}$$

Next, we rely on the solution being in $H^r\left(\Omega, T_h^1\right)$ to guarantee that u is continuous at the midpoints of the elements of the primal grid and therefore continuous at the endpoints of the dual grid. We thus add $[u]$ to the argument of the norm and then break the jump term into two parts:

$$\begin{aligned}\left\|\left[\Pi^2\Pi^1 u\right]\right\|_{L^2(\partial I^{2,k})} &= \left\|\left[\Pi^2\Pi^1 u - u\right]\right\|_{\partial I^{2,k}} \\ &\leq \left\|\Pi^2\Pi^1 u^- - u^-\right\|_{\partial I^{2,k}} + \left\|\Pi^2\Pi^1 u^+ - u^+\right\|_{\partial I^{2,k}}.\end{aligned}$$

Applying the multiplicative trace inequality from Lemma 4.6, summing over both negative and positive traces, for all boundaries of all elements then recalling (4.7), and finally using the filter approximation results from Lemmas 4.3 and 4.5 the result is complete. ∎

We are now in a position to complete the proof of the main theorem of this section.

Proof of Theorem 4.1. We directly establish consistency using the previous lemmas. First, we consider the original advection equation as a residual, projected into $V^{1,h}$:

$$\frac{\partial \Pi^1 u}{\partial t} = -\Pi^1 \frac{\partial u}{\partial x}. \tag{4.8}$$

Next we consider the local truncation error, T_h, which is the residual when the solution is L^2 projected onto V^h and acted on by the numerical operator:

$$\frac{\partial \Pi^1 u}{\partial t} = -\Pi^1 \left(\frac{\partial \Pi^2\Pi^1 u}{\partial x} + \frac{1}{2}\mathcal{L}\left(\tau\left[\Pi^2\Pi^1 u\right]\right)\right) + T_h. \tag{4.9}$$

Finally, we consider the numerical operator acting on the numerical solution

$$\frac{\partial u_h}{\partial t} = -\Pi^1 \left(\frac{\partial \Pi^2 u_h}{\partial x} + \frac{1}{2} \mathcal{L} \left(\tau \left[\Pi^2 u_h \right] \right) \right). \tag{4.10}$$

Combining the first two equations

$$T_h = \Pi^1 \frac{\partial u}{\partial x} - \Pi^1 \left(\frac{\partial \Pi^2 \Pi^1 u}{\partial x} + \frac{1}{2} \mathcal{L} \left(\tau \left[\Pi^2 \Pi^1 u \right] \right) \right), \tag{4.11}$$

we deduce that

$$\|T_h\|_{L^2(\Omega)} \leq \left\| \Pi^1 \left(\frac{\partial u}{\partial x} - \frac{\partial \Pi^2 \Pi^1 u}{\partial x} \right) \right\|_{L^2\left(T_h^1\right)} + \frac{1}{2} \left\| \Pi^1 \mathcal{L} \left(\tau \left[\Pi^2 \Pi^1 u \right] \right) \right\|_{L^2\left(T_h^1\right)}, \tag{4.12}$$

and then recalling the contraction property of Π^1

$$\|T_h\|_{L^2(\Omega)} \leq \left\| \frac{\partial u}{\partial x} - \frac{\partial \Pi^2 \Pi^1 u}{\partial x} \right\|_{L^2\left(T_h^2\right)} + \frac{1}{2} \left\| \mathcal{L} \left(\tau \left[\Pi^2 \Pi^1 u \right] \right) \right\|_{L^2\left(T_h^2\right)}. \tag{4.13}$$

Applying Lemmas 4.3, 4.5, and 4.8 we obtain

$$\|T_h\|_{L^2(\Omega)} \leq C_{n,s} \sum_{k=1}^{k=K^1} \left(h^{1,k}\right)^s \|u\|_{H^{s+1}(I^{1,k})} + C_{n,s} \sum_{k=1}^{k=K^2} \left(h^{2,k}\right)^s \|u\|_{H^{s+1}(I^{2,k})}, \tag{4.14}$$

which requires that the solution u is bounded in a broken Sobolev norm both on the covolume and primal meshes.

To continue the convergence analysis we take the difference between (4.9) and (4.10) to form a local error equation, and using the stability of the upwind DG derivative operator we obtain

$$\frac{d}{dt} \left\| \Pi^1 u - u_h \right\|_{L^2(\Omega)} \leq \|T_h\|_{L^2(\Omega)}. \tag{4.15}$$

Following the approach in Hesthaven and Warburton [15] we now obtain

$$\begin{aligned}
\|u - u_h\|_{L^2(\Omega)} &\leq \left\| u - \Pi^1 u \right\|_{L^2(\Omega)} + \left\| (u - u_h)|_{t=0} \right\|_{L^2(\Omega)} + T \max_{t \in [0,T]} \|T_h\| \\
&\leq C_{n,s} \sum_{k=1}^{k=K^1} \left(h^{1,k}\right)^{s+1} \|u\|_{H^{s+1}(I^{1,k})} \\
&\quad + C_{n,s} \sum_{k=1}^{k=K^1} \left(h^{1,k}\right)^s T \max_{t \in [0,T]} \left(\|u\|_{H^{s+1}(I^{1,k})} \right) \\
&\quad + C_{n,s} \sum_{k=1}^{k=K^2} \left(h^{2,k}\right)^{s+1} \|u\|_{H^{s+1}(I^{2,k})} \\
&\quad + C_{n,s} \sum_{k=1}^{k=K^2} \left(h^{2,k}\right)^s T \max_{t \in [0,T]} \left(\|u\|_{H^{s+1}(I^{2,k})} \right),
\end{aligned}$$

completing the proof. ∎

These estimates can be improved in the specific case when the domain is periodic and the mesh consists of equal sized elements. In section 5 we provide a dispersion analysis of the overlapping DG method and attain optimal order convergence in the limit of small kh for wave numbers k which are small in comparison with the element size h.

5. Fourier analysis for uniform element sizes. In the special case of uniform meshes we can directly analyze the covolume filtered approximation using discrete Fourier series. We recall the results of applying this analysis to the standard DG method, as completed first via symbolic computations by Hu and Atkins [18] and subsequently proven by Ainsworth [1]. Namely, the DG derivative operator admits a discrete eigenfunction, $\hat{u}^{DG}(x)$, which agrees with e^{ikx} within an error of $O(h^{n+1})$, and an eigenvalue, $i\lambda(k)$, which agrees with ik within an error of $O(h^{2n+2})$; precisely, the error has a real part of $O(h^{2n+2})$, corresponding to dissipation, and an imaginary part of $O(h^{2n+3})$, corresponding to a dispersion error.

Our symbolic calculations, carried out for $1 \leq n \leq 14$, show that the covolume filtered DG operator in fact has the same properties. As shown in Table 5.1, the leading-order dissipation error is precisely the same as for the standard DG derivative operator. The constant in the dispersion error is different, however, though of somewhat comparable magnitude.

TABLE 5.1
Errors in the dispersion relation as a function of n and $\zeta = kh$.

n	Standard upwind DG	Covolume filtered upwind DG
1	$1.4 \times 10^{-2}\zeta^4 + i \cdot 3.7 \times 10^{-3}\zeta^5$	$1.4 \times 10^{-2}\zeta^4 + i \cdot 5.4 \times 10^{-3}\zeta^5$
2	$1.4 \times 10^{-4}\zeta^6 + i \cdot 2.4 \times 10^{-5}\zeta^7$	$1.4 \times 10^{-4}\zeta^6 + i \cdot 2.0 \times 10^{-6}\zeta^7$
3	$7.1 \times 10^{-7}\zeta^8 + i \cdot 9.0 \times 10^{-8}\zeta^9$	$7.1 \times 10^{-7}\zeta^8 + i \cdot 9.9 \times 10^{-7}\zeta^9$
4	$2.2 \times 10^{-9}\zeta^{10} + i \cdot 2.2 \times 10^{-10}\zeta^{11}$	$2.2 \times 10^{-9}\zeta^{10} - i \cdot 2.5 \times 10^{-10}\zeta^{11}$
5	$4.5 \times 10^{-12}\zeta^{12} + i \cdot 3.8 \times 10^{-13}\zeta^{13}$	$4.5 \times 10^{-12}\zeta^{12} + i \cdot 2,2 \times 10^{-11}\zeta^{13}$
6	$6.7 \times 10^{-15}\zeta^{14} + i \cdot 4.8 \times 10^{-16}\zeta^{15}$	$6.7 \times 10^{-15}\zeta^{14} - i \cdot 5.7 \times 10^{-16}\zeta^{15}$
7	$7.4 \times 10^{-18}\zeta^{16} + i \cdot 4.7 \times 10^{-19}\zeta^{17}$	$7.4 \times 10^{-18}\zeta^{16} + i \cdot 1.6 \times 10^{-16}\zeta^{17}$
8	$6.4 \times 10^{-21}\zeta^{18} + i \cdot 3.6 \times 10^{-22}\zeta^{19}$	$6.4 \times 10^{-21}\zeta^{18} + i \cdot 1.9 \times 10^{-21}\zeta^{19}$
9	$4.4 \times 10^{-24}\zeta^{20} + i \cdot 2.2 \times 10^{-25}\zeta^{21}$	$4.4 \times 10^{-24}\zeta^{20} + i \cdot 4.5 \times 10^{-22}\zeta^{21}$
10	$2.5 \times 10^{-27}\zeta^{22} + i \cdot 1.1 \times 10^{-28}\zeta^{23}$	$2.5 \times 10^{-27}\zeta^{22} + i \cdot 8.0 \times 10^{-27}\zeta^{23}$
11	$1.2 \times 10^{-30}\zeta^{24} + i \cdot 5.0 \times 10^{-32}\zeta^{25}$	$1.2 \times 10^{-30}\zeta^{24} + i \cdot 6.3 \times 10^{-28}\zeta^{25}$
12	$4.8 \times 10^{-34}\zeta^{26} + i \cdot 1.8 \times 10^{-35}\zeta^{27}$	$4.8 \times 10^{-34}\zeta^{26} + i \cdot 9.4 \times 10^{-33}\zeta^{27}$
13	$1.6 \times 10^{-37}\zeta^{28} + i \cdot 5.8 \times 10^{-39}\zeta^{29}$	$1.6 \times 10^{-37}\zeta^{28} + i \cdot 4.7 \times 10^{-34}\zeta^{29}$
14	$4.9 \times 10^{-41}\zeta^{30} + i \cdot 1.6 \times 10^{-42}\zeta^{31}$	$4.9 \times 10^{-41}\zeta^{30} + i \cdot 5.3 \times 10^{-39}\zeta^{31}$

We note that this result has a number of important consequences for applications on uniform meshes. First, one expects the error to be far less sensitive to the time of integration than other approximations of the same order, as the time dependence of the error is manifested only through the error in the dispersion relation. Second, in terms of formal order of accuracy, one now expects convergence at order $n + 1$. The point again is that the order $n + 1$ error incurred by projecting the initial data onto the piecewise polynomial approximation space is propagated with far more accuracy. Thus we do not experience the usual order reduction over time.

We note that eigenanalysis is straightforward. Without loss of generality we scale x so that the element widths are 1, and as in [13] we represent the solution in the ith element by the coefficients of its Taylor expansion about the element center, P_i. We now simply view the DG method (covolume filtered or standard) as a difference method applied to this vector function. For the covolume filtered derivative operator the difference stencil is seen to involve four elements,

$$(\mathcal{D}P)_i = D_{-2}P_{i-2} + D_{-1}P_{i-1} + D_0P_i + D_1P_{i+1}, \tag{5.1}$$

leading to the following symbol:

$$\hat{\mathcal{D}} = e^{-2ikh} D_{-2} + e^{-ikh} D_{-1} + D_0 + e^{ikh} D_1. \tag{5.2}$$

For $|\zeta| = |kh| \ll 1$ we find that the Taylor coefficients of the exponential function $e^{i\zeta}$,

$$\hat{P}^{(e)} = \begin{pmatrix} 1 \\ i\zeta \\ -\frac{\zeta^2}{2} \\ \vdots \\ i^n \frac{\zeta^n}{n!} \end{pmatrix}, \tag{5.3}$$

satisfy

$$\hat{\mathcal{D}}\hat{P}^{(e)} = i\zeta \hat{P}^{(e)} + O(|\zeta|^{n+1}). \tag{5.4}$$

Moreover, expansion of the coefficients of the characteristic polynomial of $\hat{\mathcal{D}}$ for ζ small reveals that there is a single small eigenvalue, $i\lambda$, which satisfies

$$i\lambda = i\zeta + c_{2n+2}\zeta^{2n+2} + ic_{2n+3}\zeta^{2n+3} + O(|\zeta|^{2n+4}). \tag{5.5}$$

The constants are displayed in Table 5.1.

Of course it would be desirable to prove these results in general using, for example, the techniques in [1]. We will consider this in a later work. Also there will be spurious modes which we have not computed.

6. Algorithmic cost and variations on the method. It is natural to discuss the algorithmic overhead of using the covolume mesh based filtering. For the advection equation example the filter requires approximately two grid transfers as well as two evaluations of the upwind DG derivative per stage in the RK time-stepping scheme, and this approximately triples the cost of each time step. Thus, the minimum n required for breaking even is quite high. We can, however, reduce the computational cost with the following version:

$$\frac{\partial u^1}{\partial t} = -D_1^1 F^\beta u^1, \tag{6.1}$$

where $F^\beta = \beta \Pi^1 \Pi^2 + (1-\beta)$. This formulation requires only one upwind DG derivative evaluation, and the action of the transfers can be consolidated into one operator which uses a three element stencil to filter the upwind DG gradient in each element. This approach requires less than double the operations of the unfiltered method, and the break even n is lower.

The L^2 stability equation for this method is

$$\beta \frac{d}{dt} \left\| \Pi^2 u_h^1 \right\|^2_{L^2(T_h^2)} + (1-\beta) \frac{d}{dt} \left\| u_h^1 \right\|^2_{L^2(T_h^1)} = -\frac{1}{2} \sum_k \left(|n_x| \left[F^\beta u_h^1 \right], \left[F^\beta u_h^1 \right] \right)_{\partial I^{1,k}}.$$

Proceeding as before, we compute a bound on the local truncation error for this

method:

$$\begin{aligned}
\|T_h\|_{L^2(T_h^1)} &\leq \left\| \Pi^1 \frac{\partial u}{\partial x} - \frac{\partial}{\partial x}\left(\beta \Pi^1\Pi^2\Pi^1 u + (1-\beta)\right) u \right\|_{L^2(T_h^1)} \\
&\quad + \left\| \mathcal{L}\left(\tau\left[\beta\Pi^1\Pi^2\Pi^1 u + (1-\beta)\Pi^1 u\right]\right)\right\|_{L^2(T_h^1)} \\
&\leq \left\| \frac{\partial}{\partial x}\left(u - \Pi^1\Pi^2 u\right)\right\|_{L^2(T_h^1)} \\
&\quad + \left\| \frac{\partial}{\partial x}\left((1-\beta)\Pi^1\left(\Pi^2 - 1\right) u\right)\right\|_{L^2(T_h^1)} \\
&\quad + \left\| \mathcal{L}\left(\tau\left[\beta\Pi^1\Pi^2\Pi^1 u + (1-\beta)\Pi^1 u\right]\right)\right\|_{L^2(T_h^1)} \\
&\leq \left\| \frac{\partial}{\partial x}\left(u - \Pi^1\Pi^2 u\right)\right\|_{L^2(T_h^1)} + (1-\beta)\frac{n^2}{h}\left\|\left(\Pi^2 - 1\right) u\right\|_{L^2(T_h^1)} \\
&\quad + \left\| \mathcal{L}\left(\tau\left[\beta\Pi^1\Pi^2\Pi^1 u + (1-\beta)\Pi^1 u\right]\right)\right\|_{L^2(T_h^1)}.
\end{aligned}$$

Following the same strategies as before we obtain the same form for the local truncation error as in (4.14), albeit with different constants; i.e., as one would expect with this type of analysis we lose one order of accuracy for the local truncation error. The global error analysis follows as before, again yielding error bounds with the same form but different constants.

We can also interchange the order of differentiation and filtering to obtain

$$\frac{\partial u^1}{\partial t} = -F^\beta D_1^1 u^1. \tag{6.2}$$

We note that when the right-hand side spatial operator is spectrally equivalent to the filter-then-differentiate operator, from (6.1), and all approximation results follow. This observation is based on operating on both sides of (6.1) and replacing $F^\beta u \to u$.

Finally, an important property of DG methods is that they are locally conservative. As presented none of the modified methods preserves this property, as they do not conserve the mean of the solution or gradients. However, local conservation can be easily restored. For instance, we can restore local conservation of the method in (6.2) by changing the operator F^β so that it filters only functions orthogonal to $\bigcup_{k=1}^{k=K^1} P^0\left(I^{1,k}\right)$. In fact to maintain the spectral scaling of the filtered operator it is sufficient to filter only functions orthogonal to the polynomial space $\bigcup_{k=1}^{k=K^1} P^m\left(I^{1,k}\right)$ with $m \leq \sqrt{n}$.

7. Modified upwind DG method for one-dimensional Maxwell's equations. In this section we consider the transverse mode Maxwell's equations in one dimension, which are a first-order version of the two-way wave equation. The equations are

$$\frac{\partial H_y}{\partial t} = \frac{\partial E_z}{\partial x}, \tag{7.1}$$

$$\frac{\partial E_z}{\partial t} = \frac{\partial H_y}{\partial x}, \tag{7.2}$$

where H_y is the magnetic field and E_z is the electric field. We assume perfectly electrically conducting boundary conditions at $x = 0$ and $x = 2$:

$$0 = \frac{\partial H_y}{\partial x}(0,t) = \frac{\partial H_y}{\partial x}(2,t), \tag{7.3}$$

$$0 = E_z(0,t) = E_z(2,t). \tag{7.4}$$

The appearance of these boundary conditions is in contrast to the previous examples, where the domain is periodic. We are now faced with the question of how to treat these Dirichlet and natural Neumann conditions because we are using a covolume mesh, and thus we require the solution to extend beyond the domain boundary. In the specific circumstance of Maxwell's equations with homogeneous boundary data, we are able to use the method of images. For both cases we extend the primal mesh by two elements at each end which are constructed by reflecting the mesh about each end vertex. We can then extend the primal solution by reflection for the zero Neumann condition and by antireflection for the zero Dirichlet condition. At this point we have extended the primal mesh, and a covolume mesh which encompasses the original domain may be constructed. The last modification of note is that the DG operator is applied on the covolume mesh, and since we have extended the primal grid by two elements at each end we are able to apply arbitrary boundary condition states for the DG derivative on the covolume mesh without affecting the result after projection back onto the primal mesh. With this approach we can recycle the solvers from the modified DG approach for advection.

8. Numerical results. In this section we investigate the sharpness of the analytical prediction of order of accuracy, and in fact we go further and experimentally test the spectral correctness of the discrete modified upwind DG operators. This latter test is an important indicator of potential solution quality and robustness in practical situations. Admittedly, the method as presented is rather limited in geometric flexibility and in representing solutions of low regularity, but it is important to establish the potential for the method before we should even consider further development.

8.1. Test 1: Advection on regular grid. In our first test we chose to discretize the periodic interval $[0, 2]$ with 20 elements using seventh-order polynomials in each element. In Figure 8.1 we show the full spectra of the discretized upwind DG differentiation operator for $\beta = 0, 0.2, 0.4, 0.6, 0.8, 1.0$. It is immediately apparent that the modified operator has a much reduced spectral radius as β increases.

However, it is also apparent that the filtering process is wrapping the spectrum tightly near the imaginary axis, so we must be concerned about the possibility of spurious eigenvalues for the discrete operator. In Figure 8.2 we show the same spectral plots but each zoomed to the limits of the $\beta = 1$ case.

We see that with $\beta = 1$ the normally highly dissipated DG eigenmodes have reduced dissipation and are now in close proximity to the imaginary axis, in particular near the origin, and may present in time-domain simulations as parasitic solutions. This effect can be mitigated in this specific case by choosing $\beta = 0.8$, say. Then there is a clear separation between the spurious and nonspurious eigenvalues. There is clearly a need for a subtle balance between decreasing the spectral radius of the original DG operator and confusing the spectrum near the imaginary axis. The well-represented Fourier modes evident on the imaginary axis are in danger of being inundated with the collapsing modified spurious DG modes.

In Table 8.1 we show the spectral radius for each β value tried. It is clear that this simple modification yields significant reduction in the spectral radius and corre-

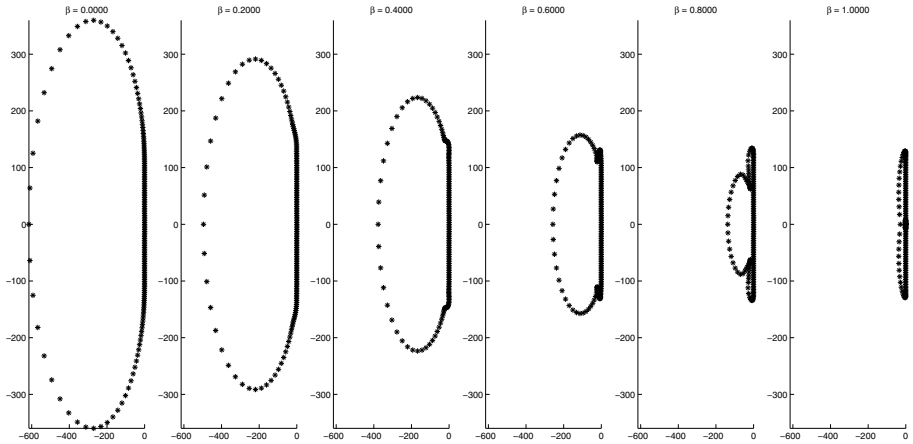


FIG. 8.1. *Full spectra for the* $\beta = 0, 0.2, 0.4, 0.6, 0.8, 1.0$ *version of the discrete upwind DG operator on a mesh of* $K = 20$ *and* $n = 7$. *Notice how the spectral radius diminishes in the* $\beta \to 1$ *limit, i.e., the fully filtered approximation of the derivative.*

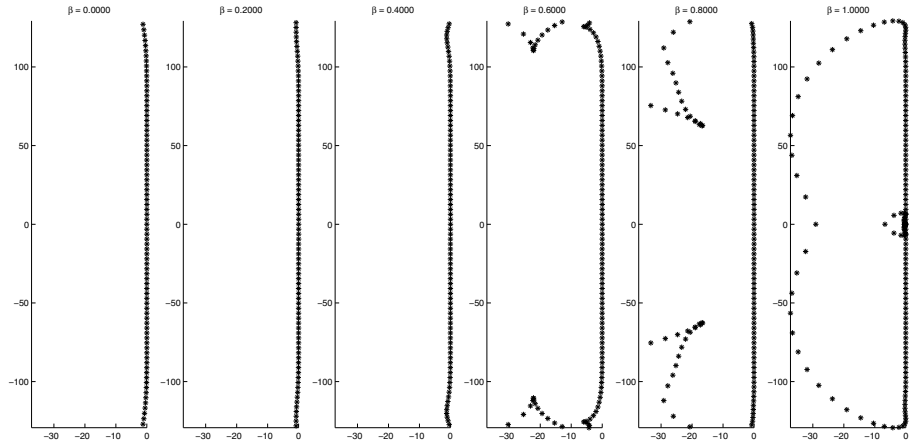


FIG. 8.2. *Close-up of the full spectra for the* $\beta = 0, 0.2, 0.4, 0.6, 0.8, 1.0$ *version of the discrete upwind DG operator on a mesh of* $K = 20$ *and* $n = 7$.

sponding increase in the maximum size of dt. We may deduce that using $\beta = 0.8$ is a reasonable compromise between spectral radius deduction and maintaining separation of the spurious modes from the imaginary axis.

The decrease in spectral radius is an attractive property of this approach, but it is also necessary to examine the impact of the use of the covolume meshes on the accuracy of the solution. We performed h-convergence tests for $K = 32, 64, 128, 256$, $n = 1, 2, \ldots, 9$ with $\beta = 0$ (i.e., unfiltered upwind DG discretization) with $dt = h/(2(n+1))$ and then with $\beta = 1 - 2/(n+1)$ and $dt = h/4$ using an initial condition:

$$u = \exp\left(-40 \sin\left((2\pi x)^2\right)\right). \tag{8.1}$$

We integrated in time using the $(n+1)$-order RK–Taylor method, sometimes referred to as the Jameson–Schmidt–Turkel scheme (see, e.g., [4]), which for a linear autonomous system is simply a recursive evaluation of the Taylor series for the matrix

TABLE 8.1

Dependence of the spectral radius of the partially stabilized operator on the β parameter for the $K = 20$, $n = 7$ upwind DG operator on the $[0, 2]$ periodic interval.

β	0	0.2	0.4	0.6	0.8	1.0
Spectral radius	614.8148	495.7820	376.6496	257.2761	137.1975	129.3503
Decrease factor	1.0000	1.2401	1.6323	2.3897	4.4812	4.7531

TABLE 8.2

h-convergence results for an unfiltered upwind DG discretization of periodic array of pulses advecting in the periodic interval $[0, 2]$ with $dt = h/(2(n + 1))$. Results prone to finite precision effects are marked with $*$.

$n+1$	$h = 2^{-5}$	$h = 2^{-6}$	$h = 2^{-7}$	Estimated order
2	5.636576e-01	3.506396e-01	1.462750e-01	0.973068
3	2.226571e-01	4.608446e-02	4.562249e-03	2.804467
4	4.449859e-02	2.294515e-03	1.286897e-04	4.216860
5	9.280798e-03	2.022760e-04	6.249343e-06	5.268164
6	1.103024e-03	1.267373e-05	2.891989e-07	5.948557
7	1.096652e-04	1.625903e-06	1.288686e-08	6.527459
8	2.025857e-05	7.457799e-08	4.831807e-10	7.677805
9	2.265482e-06	1.065869e-08	1.956979e-11*	8.410415*
10	5.625579e-07	2.223321e-10	2.431499e-12*	8.909898*

exponential. We estimated the rate of convergence by assuming the error has form $\epsilon = C_n h^n$ and using the MATLAB function polyfit to compute a linear fit to the log of the error values: coeffs = polyfit(log([1 .5 .25 .125]),log(errors)); order = coeffs(1).

The results for the unfiltered upwind DG solution of the advection equation are shown in Table 8.2. We clearly see that for $2 \leq n \leq 7$ the order of convergence measured at time $t = 1$ is between n and $n + 1$. It is clear that the $n = 1$ results are preasymptotic for this case, and the $n = 9$ and $n = 10$ cases show convergence rates diminished because the finest resolution error is near finite precision. Here we used a time step $dt = h/(2(n + 1))$ which was experimentally found to be stable for all test cases and close to the maximum dt while retaining stability. The $\beta = 1 - 2/(n + 1)$ results are shown in Table 8.3. Again, the solution includes relatively narrow exponential pulses, which are not well resolved by the linear elements even with $K = 512$ elements, and this is reflected by the preasymptotic order of accuracy obtained for the $n = 1$ test. With the exception of the $n = 1$ and $n = 5$ cases we see that the convergence rates are close to optimal with rates of nearly $n + 1$ reported. The $n = 9$ reported rate may be slightly lower than expected because the errors found on the finest mesh are close to machine precision. The filtered upwind DG results are comparable with the unfiltered upwind DG results shown in Table 8.2 despite the time steps being used for the former being much larger than those used for the latter. To examine if the $n = 5$ slow convergence is caused by an overly aggressive choice of dt we repeated this experiment with $dt = h/5$ instead of $dt = h/4$ and show the results in Table 8.4. It is clear that the near optimal rates of convergence are attained in this case. We can explain this weak dependence on n by considering the stability regions of the higher-order RK–Taylor schemes as shown in Figure 8.3. It is apparent that these regions do not grow homogeneously with increasing order, and we know that the fifth- and sixth-order schemes are not locally stable. This may result in very weak instability for marginally stable time steps and possibly explains

TABLE 8.3
h-convergence results for a periodic array of pulses advecting in the periodic interval $[0,2]$ *with* $dt = h/4$. *Results prone to finite precision effects are marked with* $*$.

$n+1$	$h = 2^{-5}$	$h = 2^{-6}$	$h = 2^{-7}$	Estimated order
2	5.636576e-01	3.506396e-01	1.462750e-01	0.973068
3	2.546519e-01	6.609079e-02	1.067257e-02	2.288273
4	7.765444e-02	7.069154e-03	4.158976e-04	3.772348
5	1.133188e-02	6.106590e-04	1.953812e-05	4.589940
6	2.471067e-03	6.086252e-05	1.665567e-05	3.606489
7	4.749631e-04	4.336478e-06	3.324958e-08	6.901096
8	8.604255e-05	2.632797e-07	1.104381e-09	8.124762
9	8.663317e-06	2.321680e-08	4.581502e-11*	8.764370*
10	1.644355e-06	1.252710e-09	1.913025e-12*	9.856617*

TABLE 8.4
h-convergence results for a periodic array of pulses advecting in the periodic interval $[0,2]$ *with* $dt = h/5$. *Results prone to finite precision effects are marked with* $*$.

$n+1$	$h = 2^{-5}$	$h = 2^{-6}$	$h = 2^{-7}$	Estimated order
2	5.526314e-01	3.321274e-01	1.317145e-01	1.034451
3	2.269038e-01	5.524013e-02	6.671504e-03	2.543963
4	5.991306e-02	3.797743e-03	2.005506e-04	4.111380
5	5.076105e-03	1.905365e-04	7.344839e-06	4.716388
6	7.340714e-04	1.554913e-05	4.416766e-07	5.349357
7	2.419079e-04	2.224695e-06	1.488670e-08	6.994073
8	3.333436e-05	8.429268e-08	4.038341e-10	8.166444
9	4.969031e-06	7.334270e-09	3.320344e-11*	8.595636*
10	5.725566e-07	3.664249e-10	1.267431e-12*	9.392576*

the issue with the $n = 6$ convergence for $dt = h/4$ which disappears for $dt = h/5$. In Figure 8.4 we show the discrete eigenspectra of the fourth- to ninth-order DG operators scaled by $dt = h/4$ and superimposed on their corresponding RK stability regions. Close examination shows that for this choice of dt the $n = 5$ scheme has eigenvalues just outside the stability region of the RK scheme near the ends of the RK stability boundary along the imaginary axis. The spectra, excluding the $n = 5$ case, are more comfortably contained within their RK stability regions, although as usual the limited inclusion of the imaginary axis within these stability regions, for $n > 4$, raises concerns for long time stability. In Figure 8.5 we show results for $dt = h/5$; this more conservative time step comfortably accommodates the discrete spectra within the RK stability regions. Some weak growth in the spectral radius of the scaled operators may be associated with the β weighted sum of filtered and unfiltered gradients.

8.1.1. Filtering and then differentiating. In this section we briefly compare the discrete spectra for an upwind DG discretization of the advection equation where we differentiate on the covolume mesh compared (see (6.1)) with a case where we filter and then differentiate on the primal grid (see (2.15)). In Figure 8.6, we compare these two approaches for a primal grid of $n = 9$, $\beta = 1 - 2/(n+1)$, and $K = 16$. It is immediately apparent that the filter-then-differentiate approach is prone to spurious eigenvalues impinging on the imaginary axis, in contrast to the original approach of differentiating on the primal grid. Furthermore, the filter-then-differentiate approach has a slightly larger spectral radius. These observations may well limit the desirability

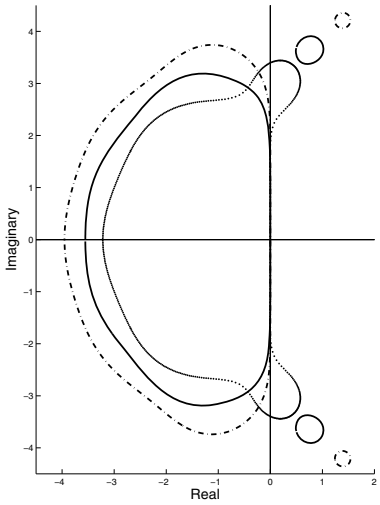


FIG. 8.3. *Margins of absolute stability for the fifth- (points), sixth- (solid line), and seventh-order (dot-dash line) RK–Taylor schemes.*

of the filter-then-differentiate approach, despite its lower algorithmic cost.

8.1.2. Long time integration: Testing the truncation error estimate. To test the sharpness of the L^2 error estimate resulting from the convergence analysis in section 4, we solved the same problem as above to a more challenging final time $T = 100$. We used discretization parameters $n = 4, 5, 6, 7$ and $K = 16, 32, 64, 128$ and set $dt = 0.15h$ and $\beta = 1 - 2/(n + 1)$.

For each simulation, we estimated the slope of the L^2 error, sampled every 10 time steps and fit to a linear function in time. For each order we then fit these slopes as a function of element width h. In Table 8.5 we report estimates for the local truncation error, which range approximately between n and $n+1$ and are in agreement with our estimate.

8.2. Test 2: Advection on variable width elements. Given the above positive results for advection on meshes with equally spaced vertices, we repeated the eigenspectra computations of section 8.1 for advection in a periodic interval with period two. For this experiment we used $K = 32$ elements and $n = 7$ polynomials on a mesh with randomly chosen vertex spacing with element lengths in the range $4.7312e - 02 \leq h \leq 7.7361e - 02$. In Figure 8.7, we show the discrete eigenspectra of the upwind DG discretized derivative operator with $\beta = 1-2/(n+1)$ and with $\beta = 1$. As with the equispaced grid, it is apparent that the $\beta = 1$ pure covolume mesh based filtering introduces spurious solutions onto the imaginary axis. However, relaxing the filter to $\beta = 1 - 2/(n + 1)$ we see that there is convincing separation between the spurious modes and the imaginary axis for all but two of the spurious eigenvalues in this case. Finally, although the filtering is relaxed we see that the spectral radius in this case is not dramatically larger than for the fully filtered case.

8.3. Numerical results: One-dimensional Maxwell's equations. In this section we will present experimental results for the accuracy of the modified DG

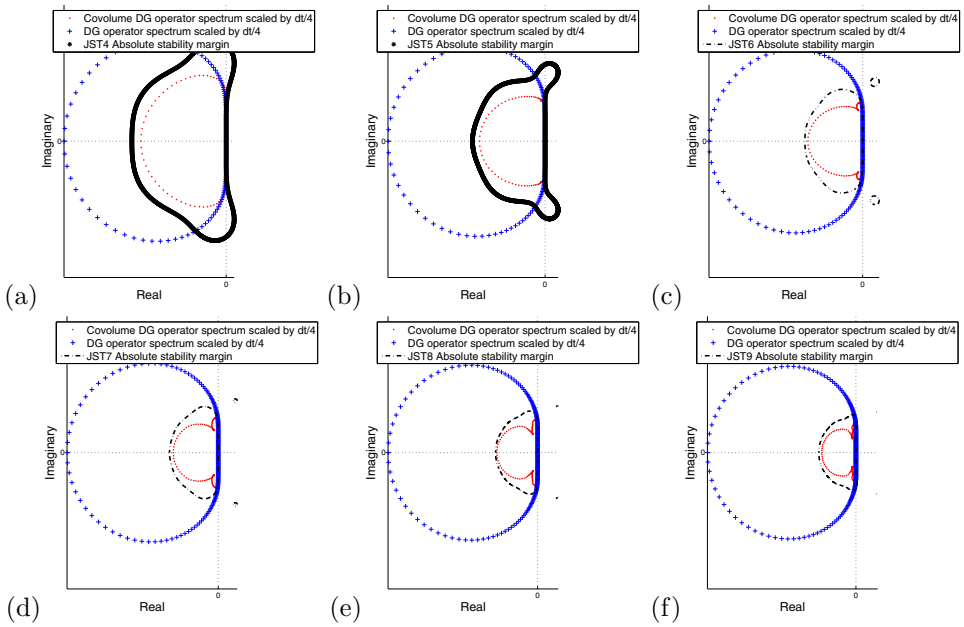


FIG. 8.4. *Discrete eigenvalues of the upwind DG operator scaled by* $dt = h/4$ *with and without covolume mesh filtering, superimposed on the margins of absolute stability of equal-order RK–Taylor schemes.* (a) $n = 3$, (b) $n = 4$, (c) $n = 5$, (d) $n = 6$, (e) $n = 7$, (f) $n = 8$.

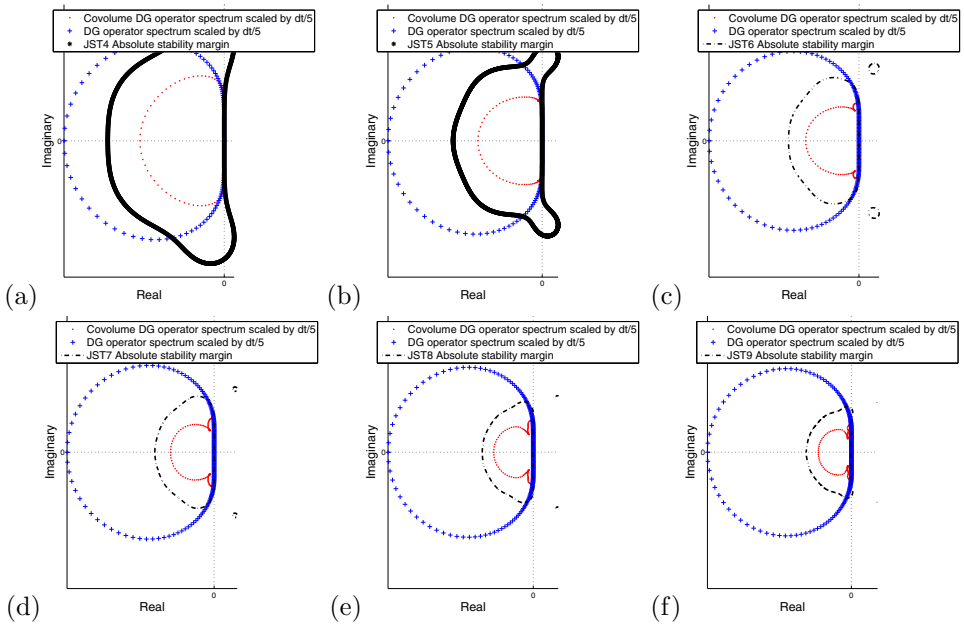


FIG. 8.5. *Discrete eigenvalues of the upwind DG operator scaled by* $dt = h/5$ *with and without covolume mesh filtering, superimposed on the margins of absolute stability of equal-order RK–Taylor schemes.* (a) $n = 3$, (b) $n = 4$, (c) $n = 5$, (d) $n = 6$, (e) $n = 7$, (f) $n = 8$.

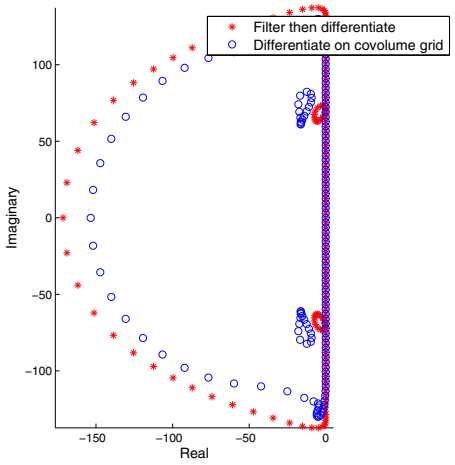


FIG. 8.6. *Comparison of discrete eigenspectra for upwind DG advection* ($K = 16, n = 9$, $\beta = 1 - 2/(n+1)$) *with the filter-then-differentiate scheme and the method where we differentiate on the covolume mesh.*

TABLE 8.5

Estimates of the order of the local truncation error T_h as a function of polynomial order n. Entries marked with $*$ are possibly prone to finite precision effects.

n	4	5	6	7
Estimate of order of T_h	3.8708	5.4041	6.9254	7.9042*

method for Maxwell's equations in section 7.

We first present estimates of the order of accuracy of the method based on simulations run to time $t = 100.23$. The model solution was constructed from a random coefficient linear combination of the first 200 standing wave solutions to the Maxwell's equations which satisfy the boundary conditions

$$H_y = \sum_{m=1}^{m=200} r_m \cos(m\pi x)\cos(m\pi t), \tag{8.2}$$

$$E_z = -\sum_{m=1}^{m=200} r_m \sin(m\pi x)\sin(m\pi t), \tag{8.3}$$

where the r_m are random numbers generated by the MATLAB function rand as $r_m = \text{rand}(1) * 2^{-.2m}$ which decay with increasing mode number to guarantee smoothness. We deliberately chose a relatively long integration time to allow for the truncation error to accumulate to a measurable degree. The time step size dt was chosen according to

$$dt = 0.15h.$$

The results in Table 8.6 show that for $n \geq 4$ we see the expected order of accuracy. The $n = 1$ and $n = 2$ results are likely preasymptotic, and no conclusions are drawn. It should be noted that two different ranges of h were used to estimate the order of the method for $1 \leq n \leq 6$ and $7 \leq n \leq 9$ since the solutions converged to machine precision for the latter range of n before the end of the range of h used for the former

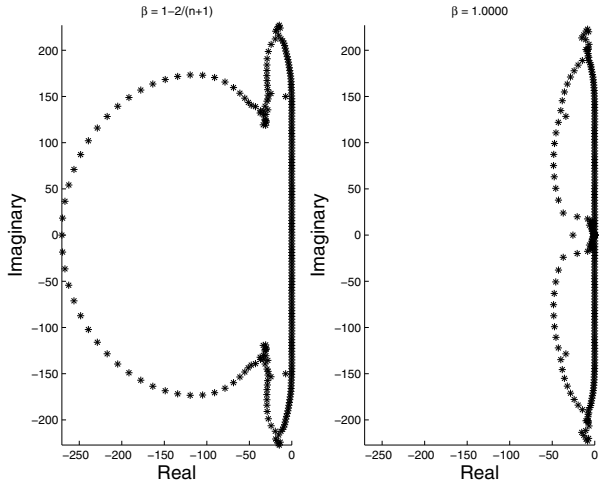


FIG. 8.7. *Full spectra for (left) $\beta = 1 - 2/(n+1)$ and (right) $\beta = 1.0$ versions of the discrete upwind DG operator on a randomly created mesh of $K = 32$ and $n = 7$. Notice the spurious eigenvalues apparent on the imaginary axis for the $\beta = 1$ case.*

TABLE 8.6

Order estimates for error in solving one-dimensional Maxwell's equations to time $t = 100.23$ on a sequence of h-refined meshes with $1 \le n \le 9$. Results which are suspected to be preasymptotic are marked with $*$.

$(n+1)$	$h = 0.03125$	$h = 0.015625$	$h = 0.0078125$	Est. order
2	0.825*	0.479*	0.21*	1.0
3	0.158*	0.0809	0.0237	1.4
4	0.0669	0.00823	0.000595	3.4
5	0.0131	0.000542	1.5e-05	4.9
6	0.00123	2.39e-05	4.1e-07	5.8
$(n+1)$	$h = 0.0625$	$h = 0.03125$	$h = 0.015625$	Est. order
7	0.0125	0.000293	1.69e-06	6.4
8	0.00282	3.35e-05	7.56e-08	7.6
9	0.000937	4.67e-06	3.66e-09	9.0
10	0.000433	9.52e-07	6.2e-10	9.7

tests. Finally, we note that the $n = 1$ case with our choice of $\beta = 1 - 2/(n+1)$ is actually the unmodified upwind DG method.

9. Summary. We have introduced a minor modification to the upwind DG method which increases the allowable time step by a factor proportional to the polynomial order of the spatial discretization, n. Both a basic local truncation error type analysis and Fourier-based dispersion analysis reveal that the modified method yields convergence rates comparable to the unmodified method in L^2. The method was tested for the advection equation and Maxwell's equations on one-dimensional meshes.

The increase in maximum time step was achieved at the expense of additional projection operations between the primal and covolume meshes and additional evaluations of DG derivatives. The break-even point is application dependent and requires further investigation. In addition, the use of approximation on primal and covolume meshes induces extra complexity in handling boundary conditions. This approach also has an impact on the efficiency of the method for parallel computations, but the local

nature of the extra projections will have a relatively minor impact on communication overheads.

In summary, the goal of reducing the spectral radius of the upwind DG spatial operator has been demonstrated and the method analyzed.

Acknowledgments. The authors express their appreciation to Professors David Gottlieb and Jan Hesthaven, Brown University, for many fruitful discussions on this topic.

REFERENCES

[1] M. Ainsworth, *Dispersive and dissipative behaviour of high order discontinuous Galerkin finite element methods*, J. Comput. Phys., 198 (2004), pp. 106–130.

[2] R. Biswas, K. Devine, and J. Flaherty, *Parallel, adaptive finite element methods for conservation laws*, Appl. Numer. Math., 14 (1994), pp. 255–283.

[3] P. Borwein and T. Erdélyi, *Polynomials and Polynomial Inequalities*, 1st ed., Grad. Texts in Math. 161, Springer-Verlag, New York, 1995.

[4] C. Canuto, M. Hussaini, A. Quarteroni, and T. Zang, *Spectral Methods in Fluid Dynamics*, 2nd ed., Springer-Verlag, New York, 1990.

[5] E. T. Chung and B. Engquist, *Optimal discontinuous Galerkin methods for wave propagation*, SIAM J. Numer. Anal., 44 (2006), pp. 2131–2158.

[6] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North–Holland, Amsterdam, 1978.

[7] B. Cockburn, S. Hou, and C.-W. Shu, *The Runge–Kutta local projection discontinuous Galerkin finite element method for conservation laws* IV: *The multidimensional case*, Math. Comp., 54 (1990), pp. 545–581.

[8] B. Cockburn, S.-Y. Lin, and C.-W. Shu, *TVB Runge–Kutta local projection discontinuous Galerkin finite element method for conservation laws* III: *One dimensional systems*, J. Comput. Phys., 84 (1989), pp. 90–113.

[9] B. Cockburn and C.-W. Shu, *TVB Runge–Kutta local projection discontinuous Galerkin finite element method for conservation laws* II: *General framework*, Math. Comp., 52 (1989), pp. 411–435.

[10] B. Cockburn and C.-W. Shu, *The local discontinuous Galerkin method for time-dependent convection-diffusion systems*, SIAM J. Numer. Anal., 35 (1998), pp. 2440–2463.

[11] B. Cockburn and C.-W. Shu, *The Runge–Kutta discontinuous Galerkin method for conservation laws* V: *Multidimensional systems*, J. Comput. Phys, 141 (1998), pp. 199–224.

[12] B. Cockburn and C.-W. Shu, *Runge–Kutta discontinuous Galerkin methods for convection-dominated problems*, J. Sci. Comput., 16 (2001), pp. 173–261.

[13] J. Goodrich, T. Hagstrom, and J. Lorenz, *Hermite methods for hyperbolic initial-boundary value problems*, Math. Comp., 75 (2006), pp. 595–630.

[14] D. Gottlieb and E. Tadmor, *The CFL condition for spectral approximations to hyperbolic initial-boundary value problems*, Math. Comp., 56 (1991), pp. 565–588.

[15] J. S. Hesthaven and T. Warburton, *High-order nodal methods on unstructured grids.* I: *Time-domain solution of Maxwell equations*, J. Comput. Phys., 181 (2002), pp. 186–221.

[16] J. S. Hesthaven and T. Warburton, *High-order accurate methods for time-domain electomagnetics*, Comput. Model. Eng. Sci., 5 (2004), pp. 395–408.

[17] J. S. Hesthaven and T. Warburton, *High-order nodal discontinuous Galerkin methods for the Maxwell eigenvalue problem*, Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 362 (2004), pp. 493–524.

[18] F. Hu and H. Atkins, *Eigensolution analysis of the discontinuous Galerkin method with nonuniform grids. Part* I: *One space dimension*, J. Comput. Phys., 182 (2002), pp. 516–545.

[19] C. Johnson and J. Pitkäranta, *Convergence of a fully discrete scheme for two-dimensional neutron transport*, SIAM J. Numer. Anal., 20 (1983), pp. 951–966.

[20] D. A. Kopriva, S. L. Woodruff, and M. Y. Hussaini, *Discontinuous spectral element approximation of Maxwell's equations*, in Discontinuous Galerkin Methods: Theory, Computation and Applications, Lect. Notes Comput. Sci. Eng. 11, B. Cockburn, G. E. Karniadakis, and C.-W. Shu, eds., Springer-Verlag, New York, 2000, pp. 355–361.

[21] D. A. Kopriva, S. L. Woodruff, and M. Y. Hussaini, *Computation of electromagnetic scattering with a nonconforming discontinuous spectral element method*, Internat. J. Numer.

Methods Engrg., 53 (2002), pp. 105–122.
[22] H.-O. Kreiss and G. Scherer, *Method of lines for hyperbolic differential equations*, SIAM J. Numer. Anal., 29 (1992), pp. 640–646.
[23] H.-O. Kreiss and L. Wu, *On the stability definition of difference approximations for the initial boundary value problem*, Appl. Numer. Math., 12 (1993), pp. 213–227.
[24] P. Lesaint and P. Raviart, *On a finite element method for solving the neutron transport equation*, in Mathematical Aspects of Finite Elements in Partial Differential Equations, Mathematics Research Center, University of Wisconsin-Madison, Madison, WI, 1974, pp. 89–123.
[25] T. E. Peterson, *A note on the convergence of the discontinuous Galerkin method for a scalar hyperbolic equation*, SIAM J. Numer. Anal., 28 (1991), pp. 133–140.
[26] S. Prudhomme, F. Pascal, J. T. Oden, and A. Romkes, *Review of A Priori Error Estimation for Discontinuous Galerkin Methods*, Technical report, TICAM, University of Texas at Austin, Austin, TX, 2000.
[27] W. Reed and T. Hill, *Triangular Mesh Methods for the Neutron Transport Equation*, Technical report, Los Alamos Scientific Laboratory, Los Alamos, NM, 1973.
[28] G. Richter, *An optimal-order error estimate for the discontinuous Galerkin method*, Math. Comp., 50 (1988), pp. 75–88.
[29] G. Sansone, *Orthogonal Functions*, Interscience, New York, 1959.
[30] Ch. Schwab, *p- and hp- Finite Element Methods: Theory and Applications to Solid and Fluid Mechanics*, 1st ed., Numer. Math. Sci. Comput., Oxford University Press, New York, 1998.
[31] T. Warburton, *Application of the discontinuous Galerkin method to Maxwell's equations using unstructured polymorphic hp-finite elements*, in Discontinuous Galerkin Methods: Theory, Computation and Applications, Lect. Notes Comput. Sci. Eng. 11, B. Cockburn, G. E. Karniadakis, and C.-W. Shu, eds., Springer-Verlag, New York, 2000, pp. 451–458.
[32] T. Warburton and M. Embree, *The role of the penalty in the local discontinuous Galerkin method for Maxwell's eigenvalue problem*, Comput. Methods Appl. Mech. Engrg., 195 (2005), pp. 3205–3223.
[33] T. Warburton and J. S. Hesthaven, *On the constants in hp-finite element trace inverse inequalities*, Comput. Methods Appl. Mech. Engrg., 192 (2003), pp. 2765–2773.

SIAM J. NUMER. ANAL.
Vol. 46, No. 6, pp. 3181–3206
© 2008 Society for Industrial and Applied Mathematics

THE COMPOSITE MINI ELEMENT—COARSE MESH COMPUTATION OF STOKES FLOWS ON COMPLICATED DOMAINS*

DANIEL PETERSEIM† AND STEFAN A. SAUTER†

Abstract. We introduce a new finite element method, the *composite mini element*, for the mixed discretization of the Stokes equations on two- and three-dimensional domains that may contain a huge number of geometric details. In standard finite element discretizations of the Stokes problem, such as the classical *mini element*, the approximation quality is determined by the maximal mesh size of the underlying triangulation, while the computational effort is determined by its number of elements. If the physical domain is very complicated, then the minimal number of simplices, which are necessary to resolve the domain, can be very large and distributed in a nonoptimal way with respect to the approximation quality. In contrast to that, the minimal dimension of the composite mini element space is independent of the number of geometric details. Instead of a geometric resolution of the domain and the boundary condition by the finite element mesh the *shape* of the finite element functions is adapted to the geometric details. This approach allows low-dimensional approximations even for problems with complicated geometric details such as holes or rough boundaries. We prove its linear (optimal order) approximability and its inf-sup stability. Further, we will be able to control the nonconformity in the space without increasing the space dimension in such a way that the a priori error estimate $\|\mathbf{u}-\mathbf{u}^{\mathrm{CME}}\|_{1,\Omega}+\|p-p^{\mathrm{CME}}\|_{0,\Omega}\lesssim h\|\mathbf{f}\|_{0,\Omega}$ holds. Thereby, in contrast to the classical methods, the choice of the mesh size parameter h is *not* constrained by the size of geometric details. In addition, it turns out that the method can be viewed as a coarse-scale generalization of the classical mini element approach; i.e., it reduces the computational effort, while the approximation quality depends on the (coarse) mesh size in the usual way.

Key words. finite elements, mixed methods, Stokes equation, nonconforming finite elements, space coarsening, complicated domain

AMS subject classifications. 65N12, 65N15, 65N30, 65D05, 35Q30

DOI. 10.1137/070704356

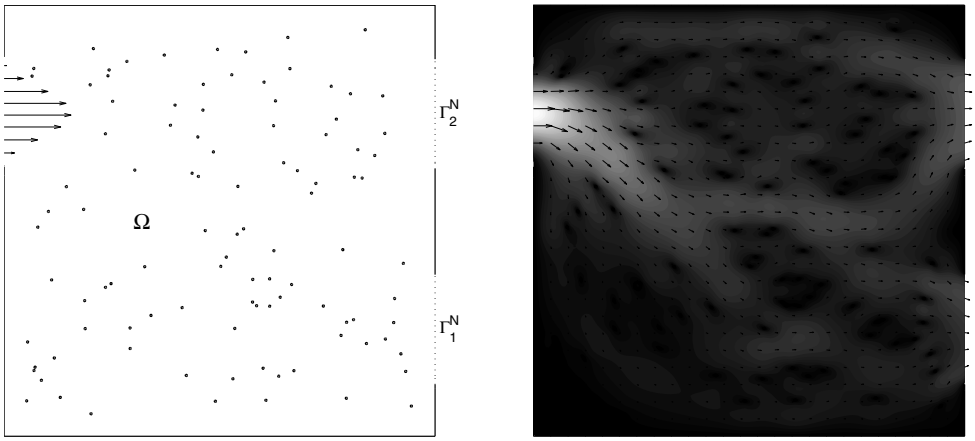
1. Introduction.

The motion of a viscous incompressible fluid in a bounded, connected Lipschitz domain $\Omega\subseteq\mathbb{R}^d$, $d\in\{2,3\}$, can be described by a velocity field $\mathbf{u}:\Omega\to\mathbb{R}^d$ and a pressure distribution $p:\Omega\to\mathbb{R}^d$ which fulfill the Stokes equations under the standard Dirichlet boundary condition

$$\left.\begin{aligned}-\Delta\mathbf{u}+\nabla p &= \mathbf{f}\\ \operatorname{div}\mathbf{u} &= 0\end{aligned}\right\}\quad\text{in }\Omega,\qquad \mathbf{u}=\mathbf{0}\quad\text{on }\partial\Omega, \tag{1.1}$$

where $\mathbf{f}:\Omega\to\mathbb{R}^d$ is a given force density. Problem (1.1) has been studied in detail, both analytically and numerically, in the literature. A wide range of finite element methods for the numerical approximation of (1.1) has been developed. In principle, these methods can also be applied to flow problems on very complicated domains Ω that arise, e.g., in environmental modeling or life sciences. However, the standard requirement that the underlying finite element mesh has to resolve the domain couples the minimal dimension of the finite element space with the number and size of geometric details, e.g., rough boundaries or holes. Such a complicated domain Ω is

*Received by the editors October 2, 2007; accepted for publication (in revised form) May 21, 2008; published electronically September 17, 2008.
http://www.siam.org/journals/sinum/46-6/70435.html
†Institute of Mathematics, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland (dpet@math.uzh.ch, stas@math.uzh.ch). The first author's research was supported by Swiss National Science Foundation grant 200020-113457/1.



(a) Model domain Ω.

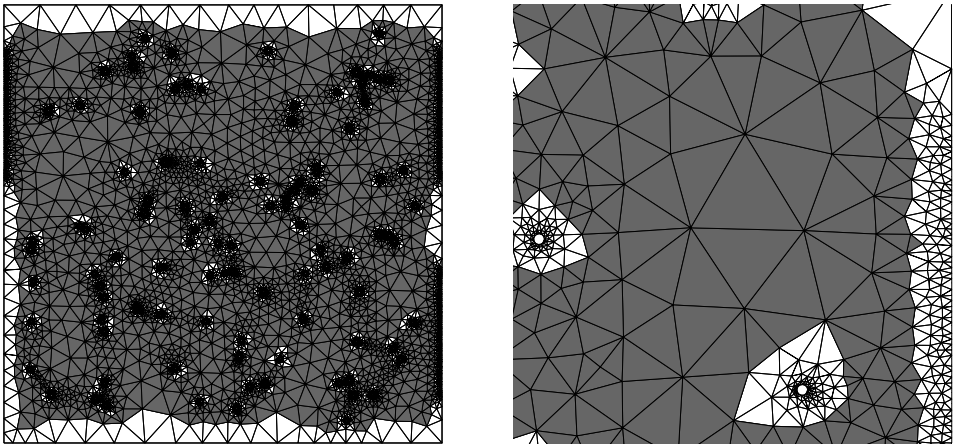
(b) Solution velocity (black = 0, white = 1).

FIG. 1.1. *A model problem: Stokes flow on the unit square with* 100 *randomly distributed circular holes (*$r = 0.005$*), an Dirichlet-inflow boundary (cf. Figure* 1.3(a)*), and two Neumann-outflow boundaries,* $\mathbf{f} = \mathbf{0}$.

depicted in Figure 1.1(a), and a coarse triangulation of Ω is depicted in Figure 1.2(a). This triangulation is a minimal subdivision in the sense that further coarsening can be performed only by violating the shape regularity condition (cf. (2.6)). The application of standard finite elements (cf. section 2) with respect to the depicted triangulation leads to a satisfactory approximation of the corresponding Stokes flow depicted in Figure 1.1(b). However, the gain of putting a huge amount of freedom into the boundary region might be only small since the error is still mainly determined by the maximal mesh size. The use of coarser unfitted meshes is *not* satisfying either since the solution of the model problem behaves critically on perturbations of the model domain, as can be seen in Figure 1.3. That means that a simplification of Ω (allowing coarser subdivisions) typically will lead to poor approximations. In this paper, we will define a method that allows the low-dimensional discretization of such types of problems, while the convergence rates (with respect to the mesh size) are preserved. The main idea is to adapt the *shape* of the finite element functions instead of resolving the geometric details by the finite element mesh. This composite mini element belongs to the class of *composite finite element methods* which have been introduced to Poisson-type problems and to problems in linear elasticity [12], [17]. Here we will generalize composite finite elements to the Stokes problem. Thereby the following difficulties have to be overcome:

(a) the definition of the composite mini element based on new extension operators of finite element functions for Dirichlet, Neumann, or slip-type boundary conditions,
(b) analysis: proof of convergence and discrete stability,
(c) complexity analysis.

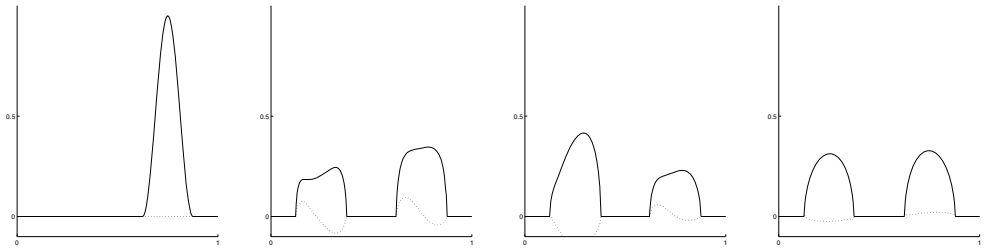
Related approaches in the literature can be found, for instance, in [2], where unfitted meshes are used to approximate elliptic equations with Neumann boundary data.



(a) Triangulation $\mathcal{T}$ with dark-shaded inner zone $\mathcal{T}^{\mathrm{dof}}$ ($\sharp\mathcal{T}^{\mathrm{dof}} \approx 1000$).

(b) $(0.8, 1) \times (0.25, 0.45)$-section of Figure 1.2(a).

FIG. 1.2. *Shape regular triangulation $\mathcal{T}$ of the model domain Ω from Figure* 1.1(a) *resolving the holes as well as the in- and outflow boundaries: max. mesh size $\approx 10^{-1}$, min. mesh size $\approx 10^{-3}$, regularity constant ≈ 2, number of triangles ≈ 9000.*



(a) Inflow velocity of the model problem.

(b) Outflow velocity of the model problem.

(c) Outflow velocity of the model problem with a different random hole distribution.

(d) Outflow velocity of the model problem without holes.

FIG. 1.3. *Consequences of simplification or perturbation of the model problem introduced in Figure* 1.1 *on the outflow behavior: Velocity components (x_1 solid, x_2 dotted) on $\{0\} \times (0, 1)$ and on $\{1\} \times (0, 1)$ for different domain modifications.*

Alternative approaches in [13], [24], and [1] are not restricted to Neumann conditions. However, the coarse spaces introduced are still coupled with the geometry, and the major goal lies in the efficient solution of fine-scale discretization (see, e.g., [24] and [1]). In contrast to our approach, the asymptotic convergence order of the underlying discretization is *not* preserved on coarser meshes. In cases where the geometric details are distributed periodically over the domain, homogenization-based approaches used to solve Stokes problems on complicated domains provide a powerful machinery

to construct problem-adapted finite element spaces (see, e.g., [8], [6]). In contrast, composite finite elements do *not* require any periodicity assumptions on the geometry.

The paper is organized as follows. In section 2 we recall the classical mini element for the weak formulation of the Stokes problem. Then, in section 3, we define the composite mini element space. Its convergence analysis is the topic of section 4, while the final section is devoted to numerical experiments.

2. The weak Stokes problem and the classical mini element. We first introduce some basic notation. By $\mathrm{W}^m_p(\Omega)$ we denote the Sobolev space of L^p-functions with weak derivatives up to order $m \in \mathbb{N} \cup \{0\}$ in $\mathrm{L}^p(\Omega)$, $p \in \mathbb{N} \cup \{\infty\}$. In the special case $p = 2$, these spaces are Hilbert spaces and are denoted by $\mathrm{H}^m(\Omega)$. We will write

$$\begin{array}{lll} & \|\cdot\|_{m,p,\Omega} \quad (|\cdot|_{m,p,\Omega}) & \text{for the norm (seminorm) in } \mathrm{W}^m_p(\Omega), \\ & \|\cdot\|_{m,\Omega} \quad (|\cdot|_{m,\Omega}) & \text{for the norm (seminorm) in } \mathrm{H}^m(\Omega), \\ \text{and} & \langle\cdot,\cdot\rangle_{m,\Omega} & \text{for the scalar product in } \mathrm{H}^m(\Omega). \end{array}$$

We will use bold letters for the function spaces if their elements are vector valued.

The Sobolev space that contains the velocity fields of Stokes flows with a homogeneous Dirichlet boundary condition is denoted by

$$\mathbf{H}^1_{\mathbf{0}}(\Omega) := \left\{ \mathbf{u} \in \mathbf{H}^1(\Omega) : \mathbf{u}|_{\partial\Omega} = \mathbf{0} \quad \text{in the sense of traces} \right\}.$$

The associated pressure space is $\mathrm{L}^2_0(\Omega) := \left\{ p \in \mathrm{L}^2(\Omega) : \int_\Omega p = 0 \right\}$. For a given right-hand side $\mathbf{f} \in \mathbf{H}^{-1}(\Omega) := (\mathbf{H}^1_{\mathbf{0}}(\Omega))'$, the weak formulation of (1.1) reads as follows: *Find a pair* $(\mathbf{u}, p) \in \mathbf{H}^1_{\mathbf{0}}(\Omega) \times L^2_0(\Omega)$ *such that*

$$\begin{array}{lll} \mathfrak{a}(\mathbf{u},\mathbf{v}) + \mathfrak{b}(\mathbf{v},p) = \langle \mathbf{f}, \mathbf{v} \rangle_{0,\Omega} & \forall \mathbf{v} \in \mathbf{H}^1_{\mathbf{0}}(\Omega), \\ \mathfrak{b}(\mathbf{u},q) = 0 & \forall q \in \mathrm{L}^2_0(\Omega). \end{array} \tag{2.1}$$

The bilinear forms $\mathfrak{a} : \mathbf{H}^1(\Omega) \times \mathbf{H}^1(\Omega) \to \mathbb{R}$ and $\mathfrak{b} : \mathbf{H}^1(\Omega) \times \mathrm{L}^2(\Omega) \to \mathbb{R}$ are defined by

$$\mathfrak{a}(\mathbf{u},\mathbf{v}) := 2\int_\Omega \mathbf{Du} : \mathbf{Dv}, \quad \mathbf{Du} := \tfrac{1}{2}\left(\nabla\mathbf{u} + (\nabla\mathbf{u})^\intercal\right), \quad \mathfrak{b}(\mathbf{v},q) := -\int_\Omega q \operatorname{div}\mathbf{v}. \tag{2.2}$$

Both bilinear forms are continuous, and, due to Korn's inequality (cf. [9] and [14]), the bilinear form $\mathfrak{a}$ is coercive with respect to $\mathbf{H}^1_{\mathbf{0}}(\Omega)$; i.e., there exists $\alpha > 0$ such that

$$\mathfrak{a}(\mathbf{u},\mathbf{u}) \geq \alpha \|\mathbf{u}\|^2_{1,\Omega} \quad \forall \mathbf{u} \in \mathbf{H}^1_{\mathbf{0}}(\Omega). \tag{2.3}$$

The coercivity constant α depends only on the diameter of Ω. The bilinear form $\mathfrak{b}$ fulfills the inf-sup condition [10, Lemma 3.2]

$$\inf_{0\neq p\in \mathrm{L}^2_0(\Omega)} \sup_{0\neq \mathbf{u}\in \mathbf{H}^1_{\mathbf{0}}(\Omega)} \frac{\mathfrak{b}(\mathbf{u},p)}{\|\mathbf{u}\|_{1,\Omega}\,\|p\|_{0,\Omega}} \geq \beta > 0. \tag{2.4}$$

It is well known from the framework of mixed variational problems (cf. [10]) that coercivity (2.3) and stability (2.4) are sufficient conditions for the unique solvability of problem (2.1). If the solution has additional smoothness properties, i.e., $(\mathbf{u}, p) \in$

$\big(\mathbf{H}^1_{\mathbf{0}}(\Omega) \times \mathrm{L}^2_0(\Omega)\big) \cap \big(\mathbf{H}^2(\Omega) \times \mathrm{H}^1(\Omega)\big)$, then (cf. [21]) there exists a constant C_{reg} such that

$$\big\|\mathbf{u}\big\|_{2,\Omega} + \big\|p\big\|_{1,\Omega} \leq C_{\mathrm{reg}} \big\|\mathbf{f}\big\|_{0,\Omega}. \tag{2.5}$$

We recall the approximation of the weak solution by the classical mini element (see, e.g., [4]). Its basis is the subdivision of the physical domain Ω into simplices, which can be performed in an exact way only if Ω is a polyhedron. In the general case, Ω needs to be approximated by a polyhedral domain $\Omega_{\mathcal{T}}$. Let $\mathcal{T} := \{T_i : 1 \leq i \leq N\}$ be a subdivision of $\Omega_{\mathcal{T}}$ consisting of (closed) simplices. The mesh size of $\mathcal{T}$ is denoted by $h := \max_{T \in \mathcal{T}} \mathrm{diam}\big(T\big)$. Any two different simplices are supposed to either be disjoint or to share exactly either one face or one side or one vertex. We will refer to Θ as the set of vertices of $\mathcal{T}$ and to $\partial\Theta := \Theta \cap \partial\Omega_{\mathcal{T}}$ as the set of boundary vertices. Furthermore, $\mathcal{T}$ is assumed to be shape regular; i.e., there exists $\rho > 0$ such that

$$\rho_T := \frac{\mathrm{diam}\big(B_T\big)}{\mathrm{diam}\big(T\big)} \geq \rho \quad \forall T \in \mathcal{T}, \tag{2.6}$$

where B_T denotes the largest ball contained in T. Based on the triangulation $\mathcal{T}$ we define the space of continuous piecewise affine functions

$$S_{\mathcal{T}} := \big\{ v \in C^0\big(\overline{\Omega_{\mathcal{T}}}\big) \mid \forall T \in \mathcal{T} : v|_T \in \mathbb{P}_1 \big\}. \tag{2.7}$$

In order to fulfill the discrete analogue of the inf-sup condition (2.4) the velocity space has to be enriched by simplex bubble functions, i.e.,

$$B_{\mathcal{T}} := \mathrm{span}\,\{\psi_T : T \in \mathcal{T}\}, \quad \psi_T := (d+1)^{d+1} \prod_{\mathbf{y} \in V(T)} \lambda_{\mathbf{y}}, \tag{2.8}$$

where $V(T)$ denotes the set of vertices of a simplex T and $\lambda_{\mathbf{y}}$, $\mathbf{y} \in V(T)$, its barycentric coordinates. The unconstrained mini element space on $\mathcal{T}$ is given by

$$\mathbf{X}_{\mathcal{T}} \times \mathrm{M}_{\mathcal{T}} := \big(\mathbf{S}_{\mathcal{T}} \oplus \mathbf{B}_{\mathcal{T}}\big) \times S_{\mathcal{T}}, \tag{2.9}$$

where bold letters mark vector-valued spaces. For problems with a homogeneous Dirichlet boundary condition, the mini element space is given by

$$\mathbf{X}^{\mathbf{0}}_{\mathcal{T}} \times \mathrm{M}_{\mathcal{T}} := \big(\underbrace{\{\mathbf{v} \in \mathbf{S}_{\mathcal{T}} \mid \mathbf{v}|_{\partial\Theta} = \mathbf{0}\}}_{\{\mathbf{v} \in \mathbf{S}_{\mathcal{T}} \,|\, \mathbf{v}|_{\partial\Omega_{\mathcal{T}}} = 0\}} \oplus \mathbf{B}_{\mathcal{T}}\big) \times S_{\mathcal{T}}. \tag{2.10}$$

If $\Omega = \Omega_{\mathcal{T}}$, then the mini element space $\mathbf{X}^{\mathbf{0}}_{\mathcal{T}} \times \mathrm{M}_{\mathcal{T}}$ is conforming, i.e., $\mathbf{X}^{\mathbf{0}}_{\mathcal{T}} \subseteq \mathbf{H}^1_{\mathbf{0}}(\Omega)$. A pair $(\mathbf{u}, p) \in \mathbf{X}^{\mathbf{0}}_{\mathcal{T}} \times \big(\mathrm{M}_{\mathcal{T}} \cap \mathrm{L}^2_0(\Omega)\big)$ is called the mini element approximation if it fulfills the discrete variational system

$$\begin{aligned} \mathfrak{a}(\mathbf{u}, \mathbf{v}) + \mathfrak{b}(\mathbf{v}, p) &= \langle \mathbf{f}, \mathbf{v} \rangle_{0,\Omega} \quad && \forall \mathbf{v} \in \mathbf{X}^{\mathbf{0}}_{\mathcal{T}}, \\ \mathfrak{b}(\mathbf{u}, q) &= 0 && \forall q \in \mathrm{M}_{\mathcal{T}} \cap \mathrm{L}^2_0(\Omega). \end{aligned} \tag{2.11}$$

It is well known that, if the continuous solution satisfies $(\mathbf{u}^*, p^*) \in \mathbf{H}^{1+r}(\Omega) \times \mathrm{H}^r(\Omega)$ for some $r \in (\frac{1}{2}, 1]$, the resulting method fulfills the following a priori error estimate:

$$\big\|\mathbf{u}^* - \mathbf{u}\big\|_{1,\Omega} + \big\|p^* - p\big\|_{0,\Omega} \leq C h^r \big\|\mathbf{f}\big\|_{1-r,\Omega}, \tag{2.12}$$

where $C > 0$ is a constant that is independent of h, $\mathbf{u}^*$, and p^*. As already mentioned in the introduction, estimate (2.12) indeed reflects the error but not necessarily the computational effort, namely the dimension of the finite element space. Our focus is on problems with different scales in the geometry; i.e., the domain contains parts that can be resolved by coarse meshes and (a huge number of) small geometric details. The standard (domain resolving) mini element has two disadvantages for problems of this type:

1. In practice, one is often interested in a moderate accuracy that cannot be achieved at a moderate effort if the mesh has very fine parts in order to resolve the geometry.
2. The mesh density of coarse shape regular triangulations of complicated domains is determined by the geometry and *not* by the smoothness properties of the solution.

The composite mini element allows a flexible handling of multiple scales in the geometry and coarse mesh computations even on very complicated domains.

3. The composite mini element. In this section we define a new mixed finite element for the Stokes equations, the composite mini element (CME). It decouples the minimal space dimension of the approximation space from the domain geometry. In the classical finite element methods this coupling is due to the fact that the boundary condition has to be incorporated into the space, which usually leads to boundary concentrated meshes. The simple idea is now *not* to use every vertex of the mesh as a degree of freedom, as is usually done, but to assign degrees of freedom only to nodal points at a proper distance to the boundary and "smear" the shape functions to the boundary by taking the Dirichlet condition into account. To increase flexibility we will use an overlapping triangulation $\mathcal{T}$, i.e., the condition[1]

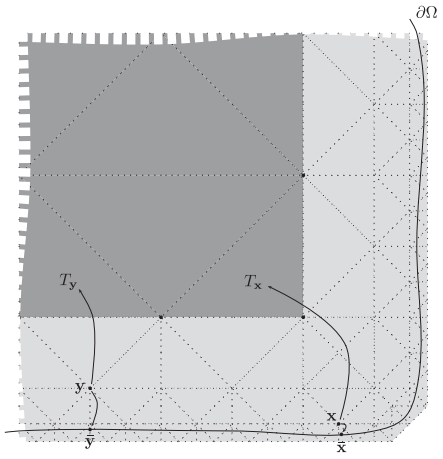
$$\Omega \subseteq \Omega_{\mathcal{T}} := \operatorname{int}\Big(\bigcup_{\tau\in\mathcal{T}} \tau\Big).$$

We define shape functions depending only on a subset of nodes of $\mathcal{T}$; the remaining nodes are *slave nodes*. The latter are used to adapt the shape of the CME functions to the complicated zero boundary. We split the triangulation into two disjoint submeshes, a coarse (inner) part $\mathcal{T}^{\mathrm{dof}}$ and a possibly refined (boundary) part $\mathcal{T}^{\mathrm{slave}}$ (see Figure 1.2):

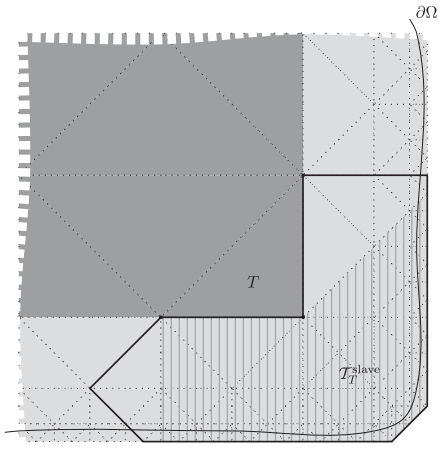
$$\mathcal{T} = \mathcal{T}^{\mathrm{dof}} \cup \mathcal{T}^{\mathrm{slave}}, \quad \emptyset \neq \Omega_{\mathcal{T}^{\mathrm{dof}}} := \bigcup_{T\in\mathcal{T}^{\mathrm{dof}}} T \subseteq \Omega.$$

Accordingly, the set of vertices Θ is decomposed into $\Theta = \Theta^{\mathrm{dof}} \cup \Theta^{\mathrm{slave}}$, where Θ^{dof} is the set of vertices of $\mathcal{T}^{\mathrm{dof}}$ and $\Theta^{\mathrm{slave}} := \Theta \setminus \Theta^{\mathrm{dof}}$. The common vertices of $\mathcal{T}^{\mathrm{dof}}$ and $\mathcal{T}^{\mathrm{slave}}$ have been put into Θ^{dof}, and the degrees of freedom will be associated with Θ^{dof}. We refer to the vertices in Θ^{slave} as slave nodes since later the values therein are defined via extension from the inner degrees of freedom toward the boundary conditions. This approach reduces the number of unknowns in the finite element discretization of the problem to the order $\sharp\mathcal{T}^{\mathrm{dof}}$ compared to $\sharp\mathcal{T}$. Thus, the local mesh refinements located in the slave part of the mesh do not affect the space dimension. This technique will turn out to be very efficient for problems on domains with rough outer boundaries

[1] This inclusion condition can be relaxed in the sense that triangulations are allowed if the errors arising from the domain approximation are negligible.



(a) The choice of closest boundary point and closest inner simplex.



(b) Inner simplex T with its domain of influence $\mathcal{T}_T^{\text{slave}}$ (black bordered). The hatched subset depicts $\hat{\mathcal{T}}_T^{\text{slave}}$.

FIG. 3.1. *Illustration of a part of a composite mesh $\mathcal{T}$ with dark-shaded inner part $\mathcal{T}^{\text{dof}}$.*

and (or) a moderate number of holes. The crucial point in the definition of the CME space is the choice of suitable extension operators for the pressure and for the velocity.

3.1. Construction of extension operators for finite elements. First, we assign a closest boundary point $\bar{\mathbf{x}} \in \partial\Omega$ and a closest inner simplex $T_{\mathbf{x}} \in \mathcal{T}^{\text{dof}}$ to every slave node $\mathbf{x} \in \Theta^{\text{slave}}$ by

$$\mathbf{x} \mapsto \bar{\mathbf{x}} \in \arg\inf_{\mathbf{y}\in\partial\Omega} \operatorname{dist}(\mathbf{x},\mathbf{y}), \quad \mathbf{x} \mapsto T_{\mathbf{x}} \in \arg\min_{T\in\mathcal{T}^{\text{dof}}} \operatorname{dist}(\mathbf{x},T). \tag{3.1}$$

Both mappings are illustrated in Figure 3.1. They can easily be computed in a preprocessing step, for instance, during the assembling of the triangulation (cf. [11], [16]).

3.1.1. Pressure extension. Pressure test functions $q \in S_{\mathcal{T}^{\text{dof}}}$, defined on the inner mesh $\mathcal{T}^{\text{dof}}$, are extended to $\Omega_{\mathcal{T}}$ by

$$\mathcal{E}^{\text{p}} : S_{\mathcal{T}^{\text{dof}}} \to S_{\mathcal{T}}, \quad \big(\mathcal{E}^{\text{p}} q\big)(\mathbf{x}) := \begin{cases} q(\mathbf{x}), & \mathbf{x} \in \Theta^{\text{dof}}, \\ q_{T_{\mathbf{x}}}(\mathbf{x}), & \mathbf{x} \in \Theta^{\text{slave}}, \end{cases} \tag{3.2}$$

where q_T denotes the affine extension (extension by itself) of $q|_T$ to $\mathbb{R}^d$. The operator $\mathcal{E}^{\text{p}}$ is well defined since its image is supposed to be a subspace of $S_{\mathcal{T}}$.

3.1.2. Velocity extension. We will use a modification of $\mathcal{E}^{\text{p}}$ for the extension of the velocity which takes into account the homogeneous Dirichlet boundary condition (cf. [17])

$$\boldsymbol{\mathcal{E}}^{\mathbf{0}} : \mathbf{S}_{\mathcal{T}^{\text{dof}}} \to \mathbf{S}_{\mathcal{T}}, \quad \boldsymbol{\mathcal{E}}^{\mathbf{0}}\mathbf{u}(\mathbf{x}) := \begin{cases} \mathbf{u}(\mathbf{x}) & \forall \mathbf{x} \in \Theta^{\text{dof}}, \\ \mathbf{u}_{T_{\mathbf{x}}}(\mathbf{x}) - \mathbf{u}_{T_{\mathbf{x}}}(\bar{\mathbf{x}}) & \forall \mathbf{x} \in \Theta^{\text{slave}}. \end{cases} \tag{3.3}$$

Note that $\big(\boldsymbol{\mathcal{E}}^{\mathbf{0}}\mathbf{u}\big)(\mathbf{x}) = 0$ if $\mathbf{x} \in \partial\Omega$ and that only the linear part of a mini element velocity field is extended. Using the mean value theorem, $\big(\boldsymbol{\mathcal{E}}^{\mathbf{0}}\mathbf{u}\big)(\mathbf{x})$ can be rewritten in the following form:

$$\big(\boldsymbol{\mathcal{E}}^{\mathbf{0}}\mathbf{u}\big)(\mathbf{x}) = (\nabla \mathbf{u}_{T_{\mathbf{x}}})(\mathbf{x} - \bar{\mathbf{x}}) \quad \forall \mathbf{x} \in \Theta^{\mathrm{slave}}, \tag{3.4}$$

where $\mathbf{u} = (u^1, \dots, u^d)^T$ and $\nabla \mathbf{u}_{T_{\mathbf{x}}}$ is the (constant) Jacobian matrix of $\mathbf{u}$ on $T_{\mathbf{x}}$.

3.2. Space definition. The CME space is defined as the image of the mini element space on $\mathcal{T}^{\mathrm{dof}}$ under the linear mapping $\boldsymbol{\mathcal{E}}^{\mathrm{CME}}$ that is composed of the previous extensions:

$$\boldsymbol{\mathcal{E}}^{\mathrm{CME}} : \mathbf{X}_{\mathcal{T}^{\mathrm{dof}}} \times \mathrm{M}_{\mathcal{T}^{\mathrm{dof}}} \to \mathbf{X}_{\mathcal{T}} \times \mathrm{M}_{\mathcal{T}}, \quad \big(\mathbf{u}^S + \mathbf{u}^B, p\big) \mapsto \big(\boldsymbol{\mathcal{E}}^{\mathbf{0}}\mathbf{u}^S + \mathbf{u}^B, \mathcal{E}^{\mathrm{p}} p\big). \tag{3.5}$$

Thereby, the bubble part $\mathbf{u}^B \in \mathbf{B}_{\mathcal{T}^{\mathrm{dof}}}$ of a velocity field $\mathbf{u} = \mathbf{u}^S + \mathbf{u}^B$, $\mathbf{u}^S \in \mathbf{S}_{\mathcal{T}^{\mathrm{dof}}}$, is simply extended by zero. The operator $\boldsymbol{\mathcal{E}}^{\mathrm{CME}}$ is injective, since functions are not changed on $\mathcal{T}^{\mathrm{dof}}$. If any two elements of $\mathbf{X}_{\mathcal{T}^{\mathrm{dof}}} \times \mathrm{M}_{\mathcal{T}^{\mathrm{dof}}}$ are different, so are their images. Finally, the CME space is defined by

$$\mathbf{X}_{\mathcal{T}}^{\mathrm{CME}} \times \mathrm{M}_{\mathcal{T}}^{\mathrm{CME}} := \boldsymbol{\mathcal{E}}^{\mathrm{CME}}\big(\mathbf{X}_{\mathcal{T}^{\mathrm{dof}}} \times \mathrm{M}_{\mathcal{T}^{\mathrm{dof}}}\big) \subseteq \mathbf{X}_{\mathcal{T}} \times \mathrm{M}_{\mathcal{T}}.$$

Note that, in general, the CME is nonconforming because the Dirichlet boundary condition is satisfied only in an approximate way. This nonconformity can be controlled in an a priori or, respectively, in an a posteriori way by the local mesh size in the near boundary zone $\mathcal{T}^{\mathrm{slave}}$. Note that there is no nonconformity arising from the pressure part of the space. A pair $(\mathbf{u}, p) \in \mathbf{X}_{\mathcal{T}}^{\mathrm{CME}} \times \big(\mathrm{M}_{\mathcal{T}}^{\mathrm{CME}} \cap \mathrm{L}_0^2(\Omega)\big)$ defines the CME approximation if it fulfills the discrete variational system

$$\begin{aligned} \mathfrak{a}(\mathbf{u}, \mathbf{v}) + \mathfrak{b}(\mathbf{v}, p) &= \langle \mathbf{f}, \mathbf{v} \rangle_{0,\Omega} && \forall \mathbf{v} \in \mathbf{X}_{\mathcal{T}}^{\mathrm{CME}}, \\ \mathfrak{b}(\mathbf{u}, q) &= 0 && \forall q \in \mathrm{M}_{\mathcal{T}}^{\mathrm{CME}} \cap \mathrm{L}_0^2(\Omega). \end{aligned} \tag{3.6}$$

4. A priori error analysis. In this section we will present the main theorem on the convergence of the CME method. Our results are based on the general mixed theory presented in [10] and [4]. The main steps are the investigation of the approximation properties as well as discrete coercivity and inf-sup stability.

4.1. Approximability. First, we show that $\mathbf{H}_{\mathbf{0}}^1(\Omega)$-functions can be approximated by the velocity part of the CME space. Usually, a piecewise affine interpolant $\mathcal{I}_{\mathcal{T}}$ with respect to the mesh $\mathcal{T}$ is used to prove this property. However, this is not possible in our situation because the simplices in $\mathcal{T}^{\mathrm{slave}}$ do not contain degrees of freedom. We will use the extensions of the affine interpolants $\mathcal{I}_{\mathcal{T}^{\mathrm{dof}}}$ on the inner mesh $\mathcal{T}^{\mathrm{dof}}$ instead. Therefore we may apply standard interpolation results on $\mathcal{T}^{\mathrm{dof}}$. The approximation quality of the proposed method will depend on the choice of $\mathcal{T}^{\mathrm{dof}}$ and the following two constants:

$$C_1^{\mathcal{T}} := \max_{t \in \mathcal{T}^{\mathrm{slave}}} \frac{\mathrm{dist}(t, \partial\Omega)}{\mathrm{diam}(t)} \quad \text{and} \quad C_2^{\mathcal{T}} := \max_{\mathbf{x} \in \Theta^{\mathrm{slave}}} \frac{\mathrm{dist}(\mathbf{x}, T_{\mathbf{x}})}{\mathrm{diam}(T_{\mathbf{x}})} \tag{4.1}$$

will enter the convergence estimates. We suppose them to be moderately bounded. An algorithm for the generation of meshes such that (4.1) holds with constants of moderate size is presented in [16]. It is well known (cf. [5, Theorem 16.1]) that, for

an arbitrary simplex $T \subseteq \mathbb{R}^d$, $d = 2, 3$, with regularity constant ρ_T, there exists a constant $C_{\mathrm{int}} = C_{\mathrm{int}}(m, p, d)$ such that

$$\big|u - \mathcal{I}_{\mathrm{T}} u\big|_{m,p,T} \leq \frac{C_{\mathrm{int}}}{\rho_T^m} \operatorname{diam}\big(T\big)^{\left(2-\frac{d}{2}+\frac{d}{p}-m\right)} \big|u\big|_{2,T} \quad \forall u \in \mathrm{H}^2(T), \tag{4.2}$$

where $m \in \{0, 1\}$ and $1 \leq p \leq \infty$, provided $\mathrm{W}_p^m(\Omega) \subseteq \mathrm{H}^2(\Omega)$.[2] $\mathcal{I}_{\mathrm{T}} u \in \mathbb{P}_1(\mathbb{R}^d)$ denotes the linear interpolant of u in the vertices of T. The subsequent lemma analyzes the approximation quality of $\mathcal{I}_{\mathrm{T}} u$ in a neighborhood of T. A basic tool of the proof is an inverse estimate of the form

$$\big|q\big|_{m,p,T} \leq \left(\frac{2}{\rho_T}\right)^m h_T^{\left(\frac{d}{p}-m\right)} \big\|q\big\|_{0,\infty,T} \quad \forall q \in \mathbb{P}_1(\mathbb{R}^d), \tag{4.3}$$

where $m \in \{0, 1\}$ and $p \in \mathbb{N} \cup \{\infty\}$.

LEMMA 4.1 (neighborhood property).[3] *Let T be an arbitrary simplex with regularity constant ρ_T, and let t be an arbitrary simplex with regularity constant ρ_t. Let the ratio of the diameters of t and T be denoted by C_{size} and the distance between T and t relative to the size of T by C_{dist}, i.e.,*

$$C_{\mathrm{size}} := \frac{\operatorname{diam}\big(t\big)}{\operatorname{diam}\big(T\big)} \quad \textit{and} \quad C_{\mathrm{dist}} := \frac{\operatorname{dist}(t, T)}{\operatorname{diam}\big(T\big)}.$$

Furthermore, let $u \in H^2(\operatorname{conv}\big(T \cup t\big))$, and let $\mathcal{I}_{\mathrm{T}} u \in \mathbb{P}_1(\mathbb{R}^d)$ denote the affine interpolation of u at the vertices of T. Then, for $m \in \{0, 1\}$ and $1 \leq p \leq \infty$, provided $W_p^m(\Omega) \subseteq H^2(\Omega)$, there exists a constant $C_{\mathrm{np}} = C_{\mathrm{np}}(C_{\mathrm{int}}, d, C_{\mathrm{size}}, C_{\mathrm{dist}}, \rho_t, \rho_T) > 0$ such that

$$\big|u - \mathcal{I}_{\mathrm{T}} u\big|_{m,p,t} \leq C_{\mathrm{np}} \operatorname{diam}\big(T\big)^{\left(2-\frac{d}{2}\right)} \operatorname{diam}\big(t\big)^{\left(\frac{d}{p}-m\right)} \big|u\big|_{2,\operatorname{conv}\left(T\cup t\right)}.$$

Proof. Let $\mathcal{I}_{\mathrm{t}} u \in \mathbb{P}_1(\mathbb{R}^d)$ denote the affine interpolation of u at the vertices of t. We define $h_t := \operatorname{diam}\big(t\big)$ and $h_T := \operatorname{diam}\big(T\big)$. The use of the triangle inequality, (4.2), and (4.3) leads to

$$\begin{aligned} \big|u - \mathcal{I}_{\mathrm{T}} u\big|_{m,p,t} &\leq \big|u - \mathcal{I}_{\mathrm{t}} u\big|_{m,p,t} + \big|\mathcal{I}_{\mathrm{t}} u - \mathcal{I}_{\mathrm{T}} u\big|_{m,p,t} \\ &\leq \frac{C_{\mathrm{int}}}{\rho_t^m} h_t^{\left(2-\frac{d}{2}+\frac{d}{p}-m\right)} \big|u\big|_{2,t} + \left(\frac{2}{\rho_t}\right)^m h_t^{\left(\frac{d}{p}-m\right)} \big\|\mathcal{I}_{\mathrm{t}} u - \mathcal{I}_{\mathrm{T}} u\big\|_{0,\infty,t}. \end{aligned} \tag{4.4}$$

Note that $\big\|\mathcal{I}_{\mathrm{t}} u - \mathcal{I}_{\mathrm{T}} u\big\|_{0,\infty,t} = \max_{\mathbf{x}\in V(t)} \big|\mathcal{I}_{\mathrm{t}} u(\mathbf{x}) - \mathcal{I}_{\mathrm{T}} u(\mathbf{x})\big|$. We will relate the interpolates with respect to T and t by constructing a simplex τ that contains an arbitrary but fixed $\mathbf{x} \in V(t)$ as a vertex, while the other vertices are taken from $V(T)$. More precisely, we chose τ out of the $d + 1$ simplices collected in the set

$$\Lambda := \big\{\tau \mid \tau \text{ is a simplex, } \mathbf{x} \in V(\tau),\ V(\tau) \setminus \{\mathbf{x}\} \subseteq V(T)\big\}.$$

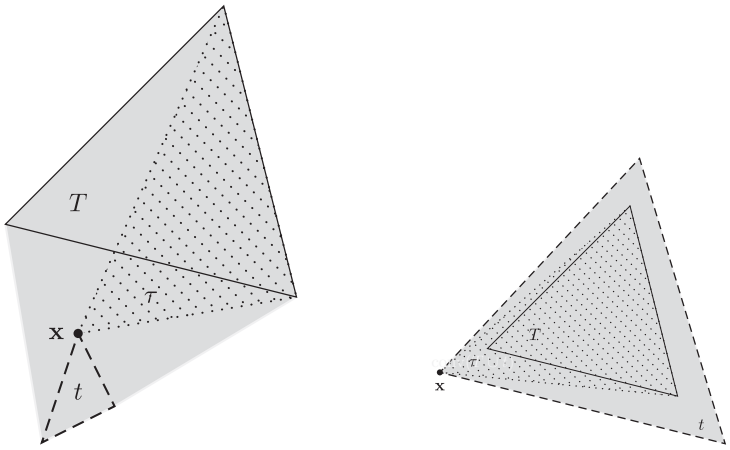
There is at least one $\tau \in \Lambda$ (cf. Figure 4.1) that fulfills the inequality[4]

$$\big|\tau \cap T\big| \geq \frac{1}{d+1} \big|T\big|. \tag{4.5}$$

[2] The condition $\mathrm{W}_p^m(\Omega) \subseteq \mathrm{H}^2(\Omega)$ restricts the choices of m and p depending on the dimension d. The combinations of m and p that will be useful later ($(m, p) \in \{(0, 2), (0, \infty), (1, 2)\}$) are allowed in two as well as three dimensions.

[3] A less general two-dimensional version of this lemma has already been given in [17].

[4] Inequality (4.5) is sharp if T is a regular simplex and $\mathbf{x}$ its midpoint.



(a) $t \cap T = \emptyset$, $C_{\mathrm{dist}} \approx C_{\mathrm{size}} \approx \frac{1}{3}$, $t, T \subsetneq \mathrm{conv}(T \cup t)$.

(b) $T \subseteq t$, $C_{\mathrm{dist}} = 0$, $C_{\mathrm{size}} \approx \frac{3}{2}$, $\mathrm{conv}(T \cup t) = t$.

FIG. 4.1. *Two typical situations for the simplices t and T from Lemma 4.1.*

Since $\tau \cap T$ is again a simplex, we can use Heron's formula to derive

$$\operatorname{diam}\big(B_\tau\big) \geq \operatorname{diam}\big(B_{\tau\cap T}\big) = \frac{d\big|\tau \cap T\big|}{\big|\partial(\tau \cap T)\big|} \geq \frac{1}{d+1}\frac{d\big|T\big|}{\big|\partial T\big|} = \frac{1}{d+1}\operatorname{diam}\big(B_T\big),$$

where, again, B_τ (resp., $B_{\tau\cap T}$) is the maximal ball contained in τ (resp., $\tau \cap T$). For this choice of τ we get

$$\operatorname{diam}\big(\tau\big) \leq (1 + C_{\mathrm{dist}} + C_{\mathrm{size}})h_T, \quad \rho_\tau \geq \frac{\rho_T}{(d+1)(1 + C_{\mathrm{dist}} + C_{\mathrm{size}})}. \tag{4.6}$$

For $\mathbf{y} \in V(\tau) \setminus \{\mathbf{x}\} = (V(\tau) \cap V(T))$ and a Taylor expansion argument, we get

$$\begin{aligned}
\big|\mathcal{I}_{\mathrm{t}}u(\mathbf{x}) - \mathcal{I}_{\mathrm{T}}u(\mathbf{x})\big| &= \big|\mathcal{I}_\tau u(\mathbf{x}) - \mathcal{I}_{\mathrm{T}}u(\mathbf{x})\big| \\
&= \big|\mathcal{I}_\tau u(\mathbf{y}) - \mathcal{I}_{\mathrm{T}}u(\mathbf{y}) + \nabla\left(\mathcal{I}_\tau u - \mathcal{I}_{\mathrm{T}}u\right)\cdot(\mathbf{x}-\mathbf{y})\big| \\
&\leq \big|\nabla\left(\mathcal{I}_\tau u - \mathcal{I}_{\mathrm{T}}u\right)\big|\operatorname{diam}\big(\tau\big) \overset{(4.6)}{\lesssim} \big|\nabla\left(\mathcal{I}_\tau u - \mathcal{I}_{\mathrm{T}}u\right)\big|\big|T\big|^{\frac{1}{2}}h_T^{\left(1-\frac{d}{2}\right)} \\
&\overset{(4.6),(4.5)}{\lesssim} h_T^{\left(1-\frac{d}{2}\right)}\big|\mathcal{I}_\tau u - \mathcal{I}_{\mathrm{T}}u\big|_{1,T\cap\tau} \lesssim h_T^{\left(1-\frac{d}{2}\right)}\left(\big|\mathcal{I}_\tau u - u\big|_{1,\tau} + \big|u - \mathcal{I}_{\mathrm{T}}u\big|_{1,T}\right) \\
&\overset{(4.2)}{\lesssim} h_T^{\left(1-\frac{d}{2}\right)}\left(\operatorname{diam}\big(\tau\big) + h_T\right)\big|u\big|_{2,\mathrm{conv}\left(T\cup\tau\right)} \overset{(4.6)}{\lesssim} h_T^{\left(2-\frac{d}{2}\right)}\big|u\big|_{2,\mathrm{conv}\left(T\cup t\right)}.
\end{aligned} \tag{4.7}$$

Plugging (4.7) into (4.4) finishes the proof. Devoting more effort to the estimation of the constants (see [15]) leads to

$$C_{\mathrm{np}} \leq \frac{4(d+1)^{\frac{3}{2}}C_{\mathrm{int}}}{\rho_t^m \rho_T^{\left(1+\frac{d}{2}\right)}}\left(C_{\mathrm{size}}^{\left(2-\frac{d}{2}\right)} + (1 + C_{\mathrm{dist}} + C_{\mathrm{size}})^3\right). \qquad \square$$

Next, we will show that an arbitrary $\mathbf{H}^2 \cap \mathbf{H}^1_{\mathbf{0}}(\Omega)$-function $\mathbf{u}$, which will be fixed throughout this section, can be approximated sufficiently well by $\boldsymbol{\mathcal{E}}^{\mathbf{0}}\left(\mathcal{I}_{\mathcal{T}^{\mathrm{dof}}}\mathbf{u}\right) \in \mathbf{X}_{\mathcal{T}}^{\mathrm{CME}}$, i.e., by the extension of the piecewise affine interpolation with respect to $\mathcal{T}^{\mathrm{dof}}$. Since in general $\Omega \subseteq \Omega_{\mathcal{T}}$, it will be useful to extend $\mathbf{u}$ to the larger domain Ω_T. It is known that, if Ω is bounded and Lipschitz, there exists a continuous, linear extension operator $\mathfrak{E} : \mathrm{H}^k(\Omega) \to \mathrm{H}^k(\mathbb{R}^d)$, $k \in \mathbb{N}_0$, such that

$$\forall u \in \mathrm{H}^k(\Omega): \quad \mathfrak{E}u|_{\Omega} = u \quad \text{and} \quad \left\|\mathfrak{E}u\right\|_{\mathrm{H}^k(\mathbb{R}^d)} \leq C_{\mathrm{ext}}\left\|u\right\|_{\mathrm{H}^k(\Omega)}, \tag{4.8}$$

with a constant C_{ext} depending only on k and Ω (cf. [20]). It is worth noting that, for domains containing a large number of holes and a possibly rough outer boundary, there exists an extension operator with moderately small norm C_{ext} under mild assumptions on the geometry. For all details, including the characterization of the class of domain geometries, we refer the reader to [18]. In the following we always identify $\mathbf{u}$ with its minimal extension $\mathfrak{E}\mathbf{u}$ without mentioning this explicitly. For $T \in \mathcal{T}^{\mathrm{dof}}$ the approximation results are obvious corollaries of the classical interpolation estimate (4.2). We will concentrate on the boundary region $\mathcal{T}^{\mathrm{slave}}$. Its elements can be grouped according to their closest inner simplices. For $T \in \mathcal{T}^{\mathrm{dof}}$ we define the set of slave simplices by

$$\mathcal{T}_T^{\mathrm{slave}} := \{t \in \mathcal{T}^{\mathrm{slave}} \mid \exists \mathbf{x} \in V(t) \ : \ T_{\mathbf{x}} = T\} \subseteq \mathcal{T}^{\mathrm{slave}}.$$

$\mathcal{T}_T^{\mathrm{slave}}$ is empty for all simplices that do not intersect $\partial\Omega_{\mathcal{T}^{\mathrm{dof}}}$. A simplex $T \in \mathcal{T}^{\mathrm{dof}}$ with $\mathcal{T}_T^{\mathrm{slave}} \neq \emptyset$ is called an *extrapolation simplex* and $\mathcal{T}_T^{\mathrm{slave}}$ its mesh of influence. Additionally, we fix the subsets of $\mathcal{T}^{\mathrm{slave}}$ which are exclusively influenced by a single element $T \in \mathcal{T}^{\mathrm{dof}}$ by

$$\hat{\mathcal{T}}_T^{\mathrm{slave}} := \left\{t \in \mathcal{T}_T^{\mathrm{slave}} \mid \forall \mathbf{x} \in V(t) \ : \ T_{\mathbf{x}} = T\right\} \subseteq \mathcal{T}_T^{\mathrm{slave}}.$$

An example is depicted in Figure 3.1(b). As a first step for the $\mathbf{H}^1$ approximation result we prove a local $\mathbf{L}^\infty$-estimate. Due to the extension $\boldsymbol{\mathcal{E}}^{\mathbf{0}}$ the estimate with respect to a slave simplex t cannot be local in the sense that it depends only on the $\mathbf{H}^2$-norm of $\mathbf{u}$ on t. Therefore we introduce neighborhoods of slave simplices

$$\omega_t := \bigcup_{T \in \mathcal{T}^{\mathrm{dof}} \,:\, t \in \mathcal{T}_T^{\mathrm{slave}}} \left(T \cup \bigcup_{\tau \in \mathcal{T}_T^{\mathrm{slave}}} \tau\right), \quad t \in \mathcal{T}^{\mathrm{slave}},$$

and neighborhoods of extrapolation simplices

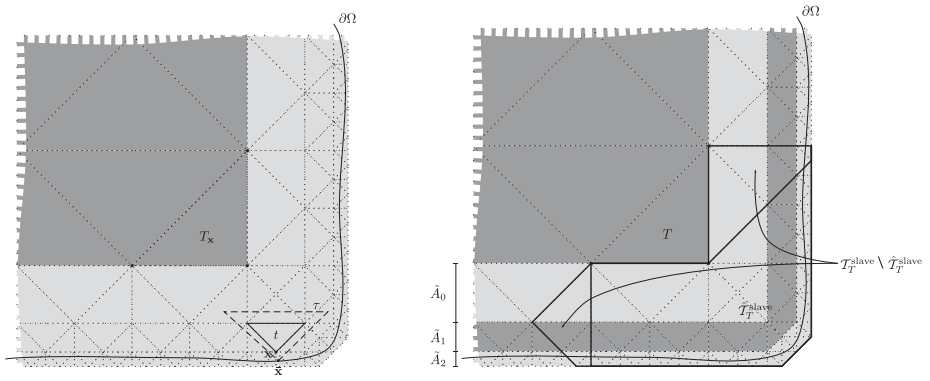
$$\omega_T := \bigcup_{t \in \mathcal{T}_T^{\mathrm{slave}}} \omega_t, \quad T \in \mathcal{T}^{\mathrm{dof}}.$$

LEMMA 4.2. *There is a constant $C = C(C_{\mathrm{int}}, \rho, C_1^{\mathcal{T}}, C_2^{\mathcal{T}}, d) > 0$ that does not depend on $\mathbf{u}$ such that*

$$\left\|\mathbf{u} - \boldsymbol{\mathcal{E}}^{\mathbf{0}}\mathcal{I}_{\mathcal{T}^{\mathrm{dof}}}\mathbf{u}\right\|_{0,\infty,t} \leq C\left[\left(\mathrm{diam}(t)^{(d-1)}h\right)^{\frac{1}{d}}\right]^{\left(2-\frac{d}{2}\right)} \left|\mathbf{u}\right|_{2,\omega_T} \quad \forall t \in \mathcal{T}^{\mathrm{slave}}.$$

Proof. For $t \in \mathcal{T}^{\mathrm{slave}}$ let $h_t := \mathrm{diam}(t)$ and $\mathcal{I}_{\mathrm{t}}\mathbf{u}$ denote the affine interpolant of $\mathbf{u}$ at the vertices of t. Then

$$\left\|\mathbf{u} - \boldsymbol{\mathcal{E}}^{\mathbf{0}}\mathcal{I}_{\mathcal{T}^{\mathrm{dof}}}\mathbf{u}\right\|_{\infty,t} \leq \left\|\mathbf{u} - \mathcal{I}_{\mathrm{t}}\mathbf{u}\right\|_{\infty,t} + \left\|\mathcal{I}_{\mathrm{t}}\mathbf{u} - \boldsymbol{\mathcal{E}}^{\mathbf{0}}\mathcal{I}_{\mathcal{T}^{\mathrm{dof}}}\mathbf{u}\right\|_{\infty,t}. \tag{4.9}$$



(a) Illustration of a slave node $\mathbf{x}$ and the simplex $\tau = \tau_{\mathbf{x}}$ from the proof of Lemma 4.2 in $\mathbb{R}^2$.

(b) The boundary regions $\tilde{A}_i$.

FIG. 4.2. *Some illustrations for the proofs of Lemma* 4.2 *and Theorem* 4.4.

A bound of the first term in (4.9) is given by (4.2). It remains to bound the second term in (4.9) or, equivalently, $\big|\mathbf{u}(\mathbf{x}) - \boldsymbol{\mathcal{E}}^{\mathbf{0}}\mathcal{I}_{\mathcal{T}^{\mathrm{dof}}}\mathbf{u}(\mathbf{x})\big|$ for $\mathbf{x} \in V(t)$. We construct a simplex τ (that does not necessarily belong to $\mathcal{T}$ (see also Figure 4.2(a))) with the following properties:

1. $\mathbf{x} \in \tau$;
2. $\bar{\mathbf{x}}$ is a vertex of τ;
3. $h_\tau := \operatorname{diam}(\tau) \in \mathcal{O}(h_t(\frac{h}{h_t})^{\frac{1}{d}})$, $\rho_\tau \geq \rho$, and $\tau \subseteq \omega_t$.

Due to (2.6) and (4.1) such a simplex always exists. The choice $h_\tau \approx h_t(\frac{h}{h_t})^{\frac{1}{d}}$ is made in order to minimize the bound of the pointwise error (see (4.11) in this proof). Let the closest inner simplex according to $\mathbf{x}$ be denoted by $T_{\mathbf{x}}$, and let its diameter be denoted by $h_{T_{\mathbf{x}}} := \operatorname{diam}(T_{\mathbf{x}})$. Let $\mathcal{I}_\tau$ and $\mathcal{I}_{\mathrm{T}_{\mathbf{x}}}$ denote the affine interpolation operators corresponding to τ and $T_{\mathbf{x}}$. Then we get

$$
\begin{aligned}
\big|\boldsymbol{\mathcal{E}}^{\mathbf{0}}\mathcal{I}_{\mathcal{T}^{\mathrm{dof}}}\mathbf{u}(\mathbf{x}) - \mathbf{u}(\mathbf{x})\big| &\overset{(3.3)}{=} \big|\nabla\mathcal{I}_{\mathrm{T}_{\mathbf{x}}}\mathbf{u}\cdot(\mathbf{x}-\bar{\mathbf{x}}) \underbrace{-\overbrace{(\mathcal{I}_\tau\mathbf{u}(\mathbf{x}) - \mathcal{I}_\tau\mathbf{u}(\bar{\mathbf{x}}))}^{=\nabla(\mathcal{I}_\tau\mathbf{u})\cdot(\mathbf{x}-\bar{\mathbf{x}})} + \mathcal{I}_\tau\mathbf{u}(\mathbf{x})}_{=0} - \mathbf{u}(\mathbf{x})\big| \\
&\leq \big|\nabla\left(\mathcal{I}_{\mathrm{T}_{\mathbf{x}}}\mathbf{u} - \mathcal{I}_\tau\mathbf{u}\right)\cdot(\mathbf{x}-\bar{\mathbf{x}})\big| + \big|\mathbf{u}(\mathbf{x}) - \mathcal{I}_\tau\mathbf{u}(\mathbf{x})\big| \qquad (4.10)\\
&\overset{(4.1),(4.2)}{\lesssim} h_t\big\|\mathcal{I}_{\mathrm{T}_{\mathbf{x}}}\mathbf{u} - \mathcal{I}_\tau\mathbf{u}\big\|_{1,\infty,\tau} + h_\tau^{\left(2-\frac{d}{2}\right)}\big|\mathbf{u}\big|_{2,\tau} \\
&\overset{(4.3)}{\lesssim} \frac{h_t}{h_\tau}\big\|\mathcal{I}_{\mathrm{T}_{\mathbf{x}}}\mathbf{u} - \mathcal{I}_\tau\mathbf{u}\big\|_{0,\infty,\tau} + h_\tau^{\left(2-\frac{d}{2}\right)}\big|\mathbf{u}\big|_{2,\tau} \\
&\lesssim \frac{h_t}{h_\tau}\left(\big\|\mathcal{I}_{\mathrm{T}_{\mathbf{x}}}\mathbf{u} - \mathbf{u}\big\|_{0,\infty,\tau} + \big\|\mathbf{u} - \mathcal{I}_\tau\mathbf{u}\big\|_{0,\infty,\tau}\right) + h_\tau^{\left(2-\frac{d}{2}\right)}\big|\mathbf{u}\big|_{2,\tau} \\
&\overset{L4.1,(4.2)}{\lesssim} \left[\frac{h_t h_{T_{\mathbf{x}}}^{\left(2-\frac{d}{2}\right)}}{h_\tau} + h_t h_\tau^{\left(1-\frac{d}{2}\right)} + h_\tau^{\left(2-\frac{d}{2}\right)}\right]\big|\mathbf{u}\big|_{2,\mathrm{conv}\left(T\cup\tau\right)}. \qquad (4.11)
\end{aligned}
$$

The choice of the diameter of τ as in property 3 leads to the assertion. ∎

LEMMA 4.3. *Let $m \in \{0,1\}$. There is a constant $C = C(C_{\mathrm{int}}, \rho, C_1^{\mathcal{T}}, C_2^{\mathcal{T}}, d) > 0$ such that*

$$\left\|\mathbf{u} - \boldsymbol{\mathcal{E}}^{\mathbf{0}}\mathcal{I}_{\mathcal{T}^{\mathrm{dof}}}\mathbf{u}\right\|_{m,t} \leq Ch^{\left(\frac{2}{d}-\frac{1}{2}\right)}\operatorname{diam}(t)^{\left(\frac{5}{2}-m-\frac{2}{d}\right)}\left|\mathbf{u}\right|_{2,\omega_t} \quad \forall t \in \mathcal{T}^{\mathrm{slave}}.$$

Proof. We separate the linear part of the error by the triangle inequality to estimate

$$\begin{aligned}\left|\mathbf{u} - \boldsymbol{\mathcal{E}}^{\mathbf{0}}\mathcal{I}_{\mathcal{T}^{\mathrm{dof}}}\mathbf{u}\right|_{m,t} &\leq \left|\boldsymbol{\mathcal{E}}^{\mathbf{0}}\mathcal{I}_{\mathcal{T}^{\mathrm{dof}}}\mathbf{u} - \mathcal{I}_{\mathrm{t}}(\mathbf{u})\right|_{m,t} + \left|\mathcal{I}_{\mathrm{t}}(\mathbf{u}) - \mathbf{u}\right|_{m,t}\\ &\overset{(4.3),(4.2)}{\lesssim} \operatorname{diam}(t)^{\left(\frac{d}{2}-m\right)}\left\|\boldsymbol{\mathcal{E}}^{\mathbf{0}}\mathcal{I}_{\mathcal{T}^{\mathrm{dof}}}\mathbf{u} - \mathcal{I}_{\mathrm{t}}(\mathbf{u})\right\|_{\infty,t} + \operatorname{diam}(t)^{(2-m)}\left|\mathbf{u}\right|_{2,t}.\end{aligned}$$

We apply Lemma 4.2 to estimate the first summand and obtain the assertion. The resulting constant C equals $\left(\frac{2}{\rho}\right)^m$ times the constant of Lemma 4.2. ∎

The error estimates in Lemmas 4.2 and 4.3 contain positive powers of $\operatorname{diam}(t)$. Hence, it is possible to localize the approximation error with respect to the elements of $\mathcal{T}^{\mathrm{slave}}$. Unfortunately, the dependence on the $\mathbf{H}^2$-norm of $\mathbf{u}$ is not local with respect to the fine-scale mesh. Thus it is not possible to sum up the local errors to get an optimal global estimate. A deeper analysis is needed to prove the following global approximation property of the velocity space.

THEOREM 4.4 (approximation property of $\mathbf{X}_{\mathcal{T}}^{\mathrm{CME}}$). *Let $m \in \{0,1\}$. There is a constant $C = C(C_{\mathrm{int}}, \rho, C_1^{\mathcal{T}}, C_2^{\mathcal{T}}, d, C_{\mathrm{ext}}) > 0$ which does not depend on h such that*

$$\left\|\mathbf{u} - \boldsymbol{\mathcal{E}}^{\mathbf{0}}\mathcal{I}_{\mathcal{T}^{\mathrm{dof}}}\mathbf{u}\right\|_{m,\Omega} \leq Ch^{(2-m)}\left|\mathbf{u}\right|_{2,\Omega}.$$

Proof. We start with a splitting of the error:

$$\left|\mathbf{u} - \boldsymbol{\mathcal{E}}^{\mathbf{0}}\mathcal{I}_{\mathcal{T}^{\mathrm{dof}}}\mathbf{u}\right|_{m,\Omega}^2 \leq \left|\mathbf{u} - \mathcal{I}_{\mathcal{T}}\mathbf{u}\right|_{m,\Omega}^2 + \sum_{T\in\mathcal{T}^{\mathrm{dof}}}\sum_{t\in\mathcal{T}_T^{\mathrm{slave}}}\left|\mathcal{I}_{\mathcal{T}}\mathbf{u} - \boldsymbol{\mathcal{E}}^{\mathbf{0}}\mathcal{I}_{\mathcal{T}^{\mathrm{dof}}}\mathbf{u}\right|_{m,t}^2. \tag{4.12}$$

As indicated before, the direct application of Lemma 4.3 leads to a suboptimal error bound in the $\mathbf{H}^1$-case. In fact, we use Lemma 4.3 only on those (few) simplices whose nodes are assigned to different inner simplices; i.e., there does not exist a $T \in \mathcal{T}^{\mathrm{dof}}$ such that $t \in \hat{\mathcal{T}}_T^{\mathrm{slave}}$. For all other simplices we will use a more local bound. Let $T \in \mathcal{T}^{\mathrm{dof}}$ and $t \in \hat{\mathcal{T}}_T^{\mathrm{slave}}$. By $\mathcal{I}_{\mathrm{t}}\mathbf{u} \subseteq \mathbb{P}_1(\mathbb{R}^d)$ we denote the affine interpolation of $\mathbf{u}$ at the vertices of t, $h_t := \operatorname{diam}(t)$. Then

$$\begin{aligned}\left|\boldsymbol{\mathcal{E}}^{\mathbf{0}}\mathcal{I}_{\mathcal{T}^{\mathrm{dof}}}\mathbf{u} - \mathcal{I}_{\mathcal{T}}\mathbf{u}\right|_{m,t} &= \left|\boldsymbol{\mathcal{E}}^{\mathbf{0}}\mathcal{I}_{\mathcal{T}^{\mathrm{dof}}}\mathbf{u} - \mathcal{I}_{\mathrm{t}}\mathbf{u}\right|_{m,t} \overset{(4.3)}{\lesssim} h_t^{\left(\frac{d}{2}-m\right)}\left\|\boldsymbol{\mathcal{E}}^{\mathbf{0}}\mathcal{I}_{\mathcal{T}^{\mathrm{dof}}}\mathbf{u} - \mathcal{I}_{\mathrm{t}}\mathbf{u}\right\|_{0,\infty,t}\\ &\overset{(3.3)}{=} h_t^{\left(\frac{d}{2}-m\right)}\max_{\mathbf{x}\in V(t)}\Big|\nabla\mathcal{I}_{\mathrm{T}_{\mathbf{x}}}\mathbf{u}\cdot(\mathbf{x}-\bar{\mathbf{x}}) - \underbrace{\left(\mathcal{I}_{\mathrm{t}}\mathbf{u}(\bar{\mathbf{x}}) + \nabla\mathcal{I}_{\mathrm{t}}\mathbf{u}\cdot(\mathbf{x}-\bar{\mathbf{x}})\right)}_{=\mathcal{I}_{\mathrm{t}}\mathbf{u}(\mathbf{x})}\Big|\\ &\overset{(4.1)}{\lesssim} h_t^{\left(1+\frac{d}{2}-m\right)}\left\|\nabla\left(\mathcal{I}_{\mathrm{T}_{\mathbf{x}}}\mathbf{u} - \mathcal{I}_{\mathrm{t}}\mathbf{u}\right)\right\|_{0,\infty,t} + h_t^{\left(\frac{d}{2}-m\right)}\max_{\mathbf{x}\in V(t)}\left|\mathcal{I}_{\mathrm{t}}\mathbf{u}(\bar{\mathbf{x}}) - \underbrace{\mathbf{u}(\bar{\mathbf{x}})}_{=0}\right|\\ &\overset{L4.1,(4.7)}{\lesssim} h_t^{(1-m)}\left|\mathcal{I}_{\mathrm{T}}\mathbf{u} - \mathcal{I}_{\mathrm{t}}\mathbf{u}\right|_{1,t} + h_t^{(2-m)}\left|\mathbf{u}\right|_{2,\tilde{t}},\end{aligned} \tag{4.13}$$

where $\tilde{t} := \operatorname{conv}\big(t \cup \{\bar{\mathbf{x}} \mid \mathbf{x} \in V(t)\}\big)$. We insert (4.13) into (4.12) and get

$$\begin{aligned}
&\sum_{t\in\mathcal{T}_T^{\mathrm{slave}}} \big|\mathcal{I}_{\mathcal{T}}\mathbf{u} - \boldsymbol{\mathcal{E}}^{\mathbf{0}}\mathcal{I}_{\mathcal{T}^{\mathrm{dof}}}\mathbf{u}\big|_{m,t}^2 \\
&\le \sum_{t\in\hat{\mathcal{T}}_T^{\mathrm{slave}}} \big|\mathcal{I}_{\mathcal{T}}\mathbf{u} - \boldsymbol{\mathcal{E}}^{\mathbf{0}}\mathcal{I}_{\mathcal{T}^{\mathrm{dof}}}\mathbf{u}\big|_{m,t}^2 + \sum_{t\in\mathcal{T}_T^{\mathrm{slave}}\setminus\hat{\mathcal{T}}_T^{\mathrm{slave}}} \big|\mathcal{I}_{\mathcal{T}}\mathbf{u} - \boldsymbol{\mathcal{E}}^{\mathbf{0}}\mathcal{I}_{\mathcal{T}^{\mathrm{dof}}}\mathbf{u}\big|_{m,t}^2 \\
&\overset{\substack{(4.13),\\ L4.3}}{\lesssim} \sum_{t\in\hat{\mathcal{T}}_T^{\mathrm{slave}}} h_t^{(2-2m)}\big|\mathcal{I}_{\mathrm{T}}\mathbf{u} - \mathcal{I}_{\mathrm{t}}\mathbf{u}\big|_{1,t}^2 + h_t^{(4-2m)}\big|\mathbf{u}\big|_{2,\tilde{t}}^2 + \sum_{t\in\mathcal{T}_T^{\mathrm{slave}}\setminus\hat{\mathcal{T}}_T^{\mathrm{slave}}} h^{\left(\frac{4}{d}-1\right)} h_t^{\left(5-2m-\frac{4}{d}\right)}\big|\mathbf{u}\big|_{2,\omega_t}^2 \\
&\le h^{2-2m}\Bigg(\underbrace{\big|\mathcal{I}_{\mathrm{T}}\mathbf{u} - \mathcal{I}_{\mathcal{T}}\mathbf{u}\big|_{1,\hat{\mathcal{T}}_T^{\mathrm{slave}}}^2}_{=:M_{1,T}} + \underbrace{\sum_{t\in\hat{\mathcal{T}}_T^{\mathrm{slave}}} h_t^2\big|\mathbf{u}\big|_{2,\tilde{t}}^2}_{=:M_{2,T}} + \underbrace{\Bigg(\sum_{t\in\mathcal{T}_T^{\mathrm{slave}}\setminus\hat{\mathcal{T}}_T^{\mathrm{slave}}} h_t^2\Bigg)\big|\mathbf{u}\big|_{2,\omega_T}^2}_{=:M_{3,T}} \Bigg).
\end{aligned}$$

If we could show the existence of constants $C_i > 0$, such that $M_{i,T} \le C_i h^2\big|\mathbf{u}\big|_{2,\omega_T}^2$ for all $i \in \{1,2,3\}$, then the proof is finished since

$$\sum_{T\in\mathcal{T}^{\mathrm{dof}}} \big|\mathbf{u}\big|_{2,\omega_T}^2 \le C(\rho, C_{\mathrm{ext}})\big|\mathbf{u}\big|_{2,\Omega}^2.$$

In case of $M_{1,T}$ this can be done in the following way:

$$M_{1,T} \le \big|\mathcal{I}_{\mathrm{T}}\mathbf{u} - \mathbf{u}\big|_{1,\hat{\mathcal{T}}_T^{\mathrm{slave}}}^2 + \big|\mathbf{u} - \mathcal{I}_{\mathcal{T}}\mathbf{u}\big|_{1,\hat{\mathcal{T}}_T^{\mathrm{slave}}}^2 \overset{L4.1,(4.2)}{\lesssim} h^2\big|\mathbf{u}\big|_{2,\operatorname{conv}\left(\hat{\mathcal{T}}_T^{\mathrm{slave}}\right)}^2.$$

For the estimation of $M_{2,T}$ we group simplices from $\mathcal{T}^{\mathrm{slave}}$ not only according to their extrapolation simplices but also according to their boundary distance. We define the sets

$$A_0 := \mathcal{T}^{\mathrm{slave}}, \quad A_k := \big\{t \in \mathcal{T}^{\mathrm{slave}} \mid \max_{\mathbf{x}\in t} \operatorname{dist}(\mathbf{x}, \partial\Omega) < 2^{-k}h\big\}, \quad k \in \mathbb{N},$$

and choose K such that $A_k = \emptyset$ for all $k > K$. Their disjoint versions are given by (cf. Figure 4.2(b))

$$\tilde{A}_k := A_k \setminus A_{k+1}, \quad 0 \le k \le K-1, \quad \tilde{A}_K := A_K. \tag{4.14}$$

The summation in $M_{2,T}$ can be resorted using (4.14):

$$\begin{aligned}
M_{2,T} = \sum_{t\in\hat{\mathcal{T}}_T^{\mathrm{slave}}} h_t^2\big|\mathbf{u}\big|_{2,\tilde{t}}^2 &\overset{(4.1)}{\lesssim} \sum_{k=0}^{K} \sum_{t\in\hat{\mathcal{T}}_T^{\mathrm{slave}}\cap\tilde{A}_k} 2^{-2k}h^2 \Bigg(\sum_{j=0}^{K}\sum_{\tau\in\tilde{A}_j} \big|\mathbf{u}\big|_{2,\tau\cap\tilde{t}}^2\Bigg) \\
&\le h^2\sum_{k=0}^{K} 2^{-2k}\Bigg(\sum_{j=0}^{K}\sum_{\tau\in\tilde{A}_j}\sum_{t\in\hat{\mathcal{T}}_T^{\mathrm{slave}}\cap\tilde{A}_k} 2^{-2k}\big|\mathbf{u}\big|_{2,\tau\cap\tilde{t}}^2\Bigg) \\
&\le h^2\sum_{k=0}^{K} 2^{-2k}\Bigg(\sum_{j=0}^{K}\sum_{\tau\in\tilde{A}_j} \sharp\tilde{A}_{k,\tau}\big|\mathbf{u}\big|_{2,\tau}^2\Bigg).
\end{aligned}$$

Here $\tilde{A}_{k,\tau} := \{t \in \hat{\mathcal{T}}_T^{\text{slave}} \cap \tilde{A}_k \mid \tau \cap \tilde{t} \neq \emptyset\}$. For $\tau \in \tilde{A}_j$ and $j > k$, $\tilde{A}_{k,\tau}$ is an empty set. In the other cases, $j \leq k$, its number of elements is bounded by a constant $C(C_1^{\mathcal{T}}, \rho, d)$ since

$$\operatorname{dist}(t_1, t_2) \leq 2 \cdot 2^{-k} h \quad \forall t_1, t_2 \in \tilde{A}_{k,\tau} \quad \text{and} \quad \big|t_1\big| \geq C(C_1^{\mathcal{T}}, \rho)(2^{-k}h)^d \quad \forall t_1 \in \tilde{A}_{k,\tau}.$$

Furthermore, $\tilde{A}_{k,\tau}$ is an empty set if $\tau \not\subseteq \omega_t$. This leads to

$$M_{2,T} \lesssim h^2 \sum_{k=0}^{K} 2^{-2k} \sum_{j=0}^{k} \big|\mathbf{u}\big|_{2,\omega_t}^2 = h^2 \Big(\underbrace{\sum_{k=0}^{K} 2^{-2k} k}_{\leq 2} \Big) \big|\mathbf{u}\big|_{2,\omega_T}^2.$$

It remains to estimate $M_{3,T}$. Let $N(T)$ denote the set of neighbors of $T \in \mathcal{T}^{\text{dof}}$ in $\mathcal{T}^{\text{dof}}$, i.e., $N(T) := \{\tau \in \mathcal{T}^{\text{dof}} : \bar{\tau} \cap \bar{T} \neq \emptyset\}$. Since $\mathcal{T}_T^{\text{slave}} \setminus \hat{\mathcal{T}}_T^{\text{slave}} \subseteq \bigcup_{\tau \in N(T)} \mathcal{T}_\tau^{\text{slave}}$ we get

(4.15)

$$\begin{aligned} M_{3,T} &\overset{(4.14)}{=} \Big(\sum_{t \in \mathcal{T}_T^{\text{slave}} \setminus \mathcal{T}_\tau^{\text{slave}}} h_t^2 \Big) \big|\mathbf{u}\big|_{2,\omega_t}^2 \leq \Big(\sum_{k=0}^{K} \sum_{\tau \in N(T)} \sum_{t \in \mathcal{T}_T^{\text{slave}} \cap \mathcal{T}_\tau^{\text{slave}} \cap \tilde{A}_k} 2^{-2k} h^2 \Big) \big|\mathbf{u}\big|_{2,\omega_T}^2 \\ &\leq h^2 \big|\mathbf{u}\big|_{2,\omega_T}^2 \Big(\sum_{\tau \in N(T)} \sum_{k=0}^{K} \#\big(\mathcal{T}_T^{\text{slave}} \cap \mathcal{T}_\tau^{\text{slave}} \cap \tilde{A}_k\big) 2^{-2k} \Big). \end{aligned}$$

A simplex t belongs to the set $\mathcal{T}_T^{\text{slave}} \cap \mathcal{T}_\tau^{\text{slave}}$, $\tau \in N(T)$, if it intersects the $(d-1)$-dimensional manifold on which the mapping

$$\mathbb{R}^d \ni \mathbf{x} \mapsto T_{\mathbf{x}} \in \operatorname{argmin}\{\operatorname{dist}(\mathbf{x}, T), \operatorname{dist}(\mathbf{x}, \tau)\}$$

has a jump (see also (3.1)). Therefore and due to (4.1) and shape regularity (2.6) the cardinality of $\mathcal{T}_T^{\text{slave}} \cap \mathcal{T}_\tau^{\text{slave}} \cap \tilde{A}_k$ satisfies

$$\#\big(\mathcal{T}_T^{\text{slave}} \cap \mathcal{T}_\tau^{\text{slave}} \cap \tilde{A}_k\big) \leq C(C_1^{\mathcal{T}}, C_2^{\mathcal{T}}, \rho, d)(2^k)^{d-2} \quad \forall k = 1, \ldots, K. \tag{4.16}$$

In two dimensions, the bounded number of simplices belonging to $\mathcal{T}_T^{\text{slave}} \cap \mathcal{T}^\tau \cap \tilde{A}_k$ is $\mathcal{O}(1)$. This is not true for $d = 3$; however, the growth in k is sufficiently slow to preserve the final estimate:

$$M_{3,T} \overset{(4.15),(4.16)}{\lesssim} \Big(\sum_{k=0}^{K} 2^{-k} \Big) h^2 \big|\mathbf{u}\big|_{2,\omega_t}^2 \leq 2h^2 \big|\mathbf{u}\big|_{2,\omega_T}^2. \qquad \square$$

The derivation of corresponding approximation results for an appropriate pressure function $p \in \mathrm{H}^1(\Omega)$ is less technical. However, the pointwise interpolation operator $\mathcal{I}_{\mathcal{T}^{\text{dof}}}$ which we used up to now is not well defined for H^1-functions and has to be replaced by a quasi-interpolation operator $\Pi_{\mathcal{T}} : \mathrm{H}^1(\Omega_{\mathcal{T}}) \to S_{\mathcal{T}}$ of Scott and Zhang[5] (see, e.g., [19]). Following [19, Theorem 4.1 and Corollary 4.1], the operator $\Pi_{\mathcal{T}}$ is bounded, and there exists a constant C_{qint} that depends only on ρ such that

$$\big\|p - \Pi_{\mathcal{T}} p\big\|_{m,t} \leq C_{\text{qint}} \operatorname{diam}(t)^{1-m} \big\|p\big\|_{1,\omega_t} \quad \forall t \in \mathcal{T}, \quad m \in \{0, 1\}. \tag{4.17}$$

[5]Since we use only the approximation property of this operator and not its special structure it is also possible to use the quasi-interpolation operator as introduced by Clément [7] or its modifications as described in [22] and [23].

As before we let $p \in \mathrm{H}^1(\Omega)$ be fixed and identify it with its extension $\mathfrak{E}p$ (cf. (4.8)). The error estimates with respect to the inner mesh $\mathcal{T}^{\mathrm{dof}}$ are given in (4.17), and we concentrate on the boundary part of the mesh.

THEOREM 4.5 (approximation property of $\mathrm{M}_{\mathcal{T}}^{\mathrm{CME}}$). *There is a constant* $C = C(C_{\mathrm{qint}}, \rho, C_1^{\mathcal{T}}, C_2^{\mathcal{T}}, d)$ *such that*

$$\|p - \mathcal{E}^{\mathrm{p}}\Pi_{\mathcal{T}^{\mathrm{dof}}} p\|_{0,t} \le Ch\|p\|_{1,\omega_t} \quad \forall t \in \mathcal{T}^{\mathrm{slave}}. \tag{4.18}$$

Furthermore, we have the global estimate

$$\|p - \mathcal{E}^{\mathrm{p}}\Pi_{\mathcal{T}^{\mathrm{dof}}} p\|_{0,\Omega} \le Ch\|p\|_{1,\Omega},$$

where the constant C *depends only on* C_{ext}, ρ, *and the constant of the local estimate.*

Proof. We start by estimating the local error. Let $t \in \mathcal{T}^{\mathrm{slave}}$. Then

$$\begin{aligned}\|p - \mathcal{E}^{\mathrm{p}}\Pi_{\mathcal{T}^{\mathrm{dof}}} p\|_{0,t} &\le \|p - \Pi_{\mathcal{T}} p\|_{0,t} + \|\Pi_{\mathcal{T}} p - \mathcal{E}^{\mathrm{p}}\Pi_{\mathcal{T}^{\mathrm{dof}}} p\|_{0,t} \\ &\overset{(4.17),(4.3)}{\lesssim} \operatorname{diam}(t)\|p\|_{1,\omega_t} + |t|^{\frac{1}{2}}\|\mathcal{I}_{\mathcal{T}} p - \mathcal{E}^{\mathrm{p}}\mathcal{I}_{\mathcal{T}^{\mathrm{dof}}} p\|_{\infty,t}.\end{aligned}$$

The infinity norm of the affine function equals the absolute value of $\big(\Pi_{\mathcal{T}} p - \mathcal{E}^{\mathrm{p}}\Pi_{\mathcal{T}^{\mathrm{dof}}} p\big)$ in some vertex $\mathbf{x} \in V(t)$, which can be estimated as follows:

$$\begin{aligned}\big|\Pi_{\mathcal{T}} p(\mathbf{x}) - \mathcal{E}^{\mathrm{p}}\Pi_{\mathcal{T}^{\mathrm{dof}}} p(\mathbf{x})\big| &= \big|\Pi_{\mathcal{T}} p(\mathbf{x}) - \Pi_{\mathrm{T}_{\mathbf{x}}} p(\mathbf{x})\big| \le \|\Pi_{\mathcal{T}} p - \Pi_{\mathrm{T}_{\mathbf{x}}} p\|_{\infty,t} \\ &\lesssim |t|^{-\frac{1}{2}}\|\Pi_{\mathcal{T}} p - \Pi_{\mathrm{T}_{\mathbf{x}}} p\|_{0,t} \overset{L4.1}{\lesssim} \operatorname{diam}(T_{\mathbf{x}})\,|t|^{-\frac{1}{2}}\|p\|_{1,\omega_t}.\end{aligned}$$

Hence, (4.18) is proved. The global estimate follows immediately, since the overlap of ω_t can be controlled in terms of ρ. ∎

The preceding approximation result is one basic ingredient of the error analysis in the next section. Note that there was no restriction concerning the *minimal* mesh size in $\mathcal{T}^{\mathrm{slave}}$, which will allow us to control the nonconformity in the space by adapting the local mesh size in $\mathcal{T}^{\mathrm{slave}}$ and without increasing the space dimension. Furthermore, it is straightforward to generalize Theorems 4.4 and 4.5 to the case of less regular solutions, say $\mathbf{u} \in \mathbf{H}^{1+r}(\Omega)$ and $p \in \mathrm{H}^r(\Omega)$ for $r \in (0,1]$, by using the interpolation theory of Sobolev spaces (cf. [3]).

4.2. Discrete stability and coercivity. In this section, we will investigate the unique solvability of the discrete CME systems. The stability proof makes use of the boundedness of the pressure extension $\mathcal{E}^{\mathrm{p}}$.

LEMMA 4.6. *There is a constant* $C_{\mathcal{E}^{\mathrm{p}}} = C_{\mathcal{E}^{\mathrm{p}}}(\rho, d)$ *such that*

$$\|\mathcal{E}^{\mathrm{p}} p\|_{0,\Omega} \le C_{\mathcal{E}^{\mathrm{p}}}\|p\|_{0,\Omega_{\mathcal{T}^{\mathrm{dof}}}} \quad \forall p \in M_{\mathcal{T}^{\mathrm{dof}}}.$$

Proof. For $p \in \mathrm{M}_{\mathcal{T}^{\mathrm{dof}}}$ there holds

$$\|\mathcal{E}^{\mathrm{p}} p\|_{0,\Omega}^2 \le \sum_{T\in\mathcal{T}^{\mathrm{dof}}} \|\mathcal{E}^{\mathrm{p}} p\|_{0,T}^2 + \sum_{t\in\mathcal{T}^{\mathrm{slave}}} \|\mathcal{E}^{\mathrm{p}} p\|_{0,t}^2 \overset{(4.3)}{\lesssim} \sum_{T\in\mathcal{T}^{\mathrm{dof}}} \|p\|_{0,T}^2 + \sum_{T\in\mathcal{T}^{\mathrm{slave}}} |t|\,\|\mathcal{E}^{\mathrm{p}} p\|_{\infty,t}^2. \tag{4.19}$$

Since $\mathcal{E}^{\mathrm{p}} p|_t$ takes its maximum in a vertex $\mathbf{x} \in V(t)$, there holds

$$\|\mathcal{E}^{\mathrm{p}} p\|_{\infty,t} \le \|p_{T_{\mathbf{x}}}\|_{\infty,t} \overset{L4.1}{\le} \left(1 + \frac{\operatorname{diam}(T_{\mathbf{x}})}{\operatorname{diam}(t)}\right)\|p_{T_{\mathbf{x}}}\|_{\infty,T_{\mathbf{x}}} \lesssim |T_{\mathbf{x}}|^{-\frac{1}{2}}\|p\|_{0,\omega_t\cap\Omega_{\mathcal{T}^{\mathrm{dof}}}},$$

where $p_{T_\mathbf{x}}$ denotes the extension of $p|_{T_\mathbf{x}}$ (by itself) to $\mathbb{R}^d$. We plug this into (4.19), which finishes the proof, since the resulting overlap is bounded by the maximal number of neighbors in $\mathcal{T}^{\mathrm{dof}}$ and therefore in terms of ρ. ∎

Now we can show that the CME is stable for the boundary conditions under consideration.

THEOREM 4.7 (stability). *$\mathbf{X}_{\mathcal{T}}^{\mathrm{CME}} \times M_{\mathcal{T}}^{\mathrm{CME}}$ is a stable pairing; i.e., there is a constant β^{CME} which depends only on d, ρ, and the discrete inf-sup constant of the mini element such that*

$$\inf_{p \in \mathrm{M}_{\mathcal{T}}^{\mathrm{CME}} \cap \mathrm{L}_0^2(\Omega)} \sup_{\mathbf{0} \neq \mathbf{u} \in \mathbf{X}_{\mathcal{T}}^{\mathrm{CME}}} \frac{\mathfrak{b}(\mathbf{u}, p)}{\|\mathbf{u}\|_{1,\Omega} \|p\|_{0,\Omega}} \geq \beta^{\mathrm{CME}}.$$

Proof. We start with the stability of the CME $\mathbf{X}_{\mathcal{T}^{\mathrm{dof}}} \times \mathrm{M}_{\mathcal{T}^{\mathrm{dof}}}$ with respect to the inner mesh $\mathcal{T}^{\mathrm{dof}}$, which is known from [4]:

$$\inf_{0 \neq p \in \mathrm{M}_{\mathcal{T}^{\mathrm{dof}}} \cap \mathrm{L}_0^2(\Omega_{\mathcal{T}^{\mathrm{dof}}})} \sup_{\mathbf{0} \neq \mathbf{u} \in \mathbf{X}_{\mathcal{T}^{\mathrm{dof}}}} \frac{\int_{\Omega_{\mathcal{T}^{\mathrm{dof}}}} p \operatorname{div} \mathbf{u}}{\|p\|_{0,\Omega_{\mathcal{T}^{\mathrm{dof}}}} \|\mathbf{u}\|_{1,\Omega_{\mathcal{T}^{\mathrm{dof}}}}} \geq \beta > 0. \tag{4.20}$$

Thereby β does not depend on the mesh size. We define two mappings to transport this result to the CME space. The first one is just a slight modification of $\mathcal{E}^{\mathrm{p}}$ in order to handle the L_0^2 intersection:

$$\tilde{\mathcal{E}}^{\mathrm{p}} : \mathrm{M}_{\mathcal{T}^{\mathrm{dof}}} \cap \mathrm{L}_0^2(\Omega_{\mathcal{T}^{\mathrm{dof}}}) \to \mathrm{M}_{\mathcal{T}}^{\mathrm{CME}} \cap \mathrm{L}_0^2(\Omega), \quad q \mapsto \mathcal{E}^{\mathrm{p}} q - \frac{1}{|\Omega|} \int_\Omega \mathcal{E}^{\mathrm{p}} q.$$

Due to Lemma 4.6, $\tilde{\mathcal{E}}^{\mathrm{p}}$ is bounded and $\|\tilde{\mathcal{E}}^{\mathrm{p}} q\|_{0,\Omega} \leq 2 C_{\mathcal{E}^{\mathrm{p}}} \|p\|_{0,\Omega_{\mathcal{T}^{\mathrm{dof}}}}$. Furthermore, $\tilde{\mathcal{E}}^{\mathrm{p}}$ is a bijection since $\mathcal{E}^{\mathrm{p}}$ maps constants on constants and the preimage of a constant function is constant. In a second step we construct a bounded mapping $\pi : \mathbf{X}_{\mathcal{T}^{\mathrm{dof}}} \to \mathbf{X}_{\mathcal{T}}^{\mathrm{CME}}$ that satisfies

$$\int_\Omega \tilde{\mathcal{E}}^{\mathrm{p}} q \operatorname{div} \pi \mathbf{u} = \int_{\Omega_{\mathcal{T}^{\mathrm{dof}}}} q \operatorname{div} \mathbf{u} \quad \forall q \in \mathrm{M}_{\mathcal{T}^{\mathrm{dof}}}. \tag{4.21}$$

This step is similar to Fortin's lemma (cf. [4, Proposition II.2.9]), where such a mapping is employed to deduce the discrete stability from the continuous one. Let us suppose for the moment that π exists. From (4.20) we know that

$$\forall q \in \mathrm{M}_{\mathcal{T}^{\mathrm{dof}}} \cap \mathrm{L}_0^2(\Omega_{\mathcal{T}^{\mathrm{dof}}}) \ \exists \mathbf{u}_q \in \mathbf{X}_{\mathcal{T}^{\mathrm{dof}}} : \int_{\Omega_{\mathcal{T}^{\mathrm{dof}}}} q \operatorname{div} \mathbf{u}_q \geq \frac{\beta}{2} \|q\|_{0,\Omega_{\mathcal{T}^{\mathrm{dof}}}} \|\mathbf{u}_q\|_{1,\Omega_{\mathcal{T}^{\mathrm{dof}}}}.$$

The left-hand side can be replaced using (4.21) and the bijectivity of $\tilde{\mathcal{E}}^{\mathrm{p}}$ which leads to

$$\forall p \in \mathrm{M}_{\mathcal{T}}^{\mathrm{CME}} \cap \mathrm{L}_0^2(\Omega) \ \exists q \in \mathrm{M}_{\mathcal{T}^{\mathrm{dof}}} \cap \mathrm{L}_0^2(\Omega_{\mathcal{T}^{\mathrm{dof}}}), \ \mathbf{u}_q \in \mathbf{X}_{\mathcal{T}^{\mathrm{dof}}} : p = \tilde{\mathcal{E}}^{\mathrm{p}} q \quad \text{and}$$
$$\int_\Omega p \operatorname{div} \pi \mathbf{u}_q = \int_{\Omega_{\mathcal{T}^{\mathrm{dof}}}} q \operatorname{div} \mathbf{u}_q \geq \frac{\beta}{2} \|q\|_{0,\Omega_{\mathcal{T}^{\mathrm{dof}}}} \|\mathbf{u}_q\|_{1,\Omega_{\mathcal{T}^{\mathrm{dof}}}}.$$

Let C_π denote the operator norm of π. Recalling $C_{\mathcal{E}^{\mathrm{p}}}$ as in Lemma 4.6 we get

$$\inf_{p \in \mathrm{M}_{\mathcal{T}}^{\mathrm{CME}} \cap \mathrm{L}_0^2(\Omega)} \sup_{\mathbf{0} \neq \mathbf{u} \in \mathbf{X}_{\mathcal{T}}^{\mathrm{CME}}} \frac{\mathfrak{b}(\mathbf{u}, p)}{\|\mathbf{u}\|_{1,\Omega} \|p\|_{0,\Omega}} \geq \frac{\beta}{4 C_{\mathcal{E}^{\mathrm{p}}} C_\pi}.$$

It remains to define π: Let $\mathbf{u} \in \mathbf{X}_{\mathcal{T}^{\mathrm{dof}}}$. The extension of $\mathbf{u} \in \mathbf{X}_{\mathcal{T}^{\mathrm{dof}}}$ by zero (again denoted by $\mathbf{u}$) is mapped onto $\mathbf{X}_{\mathcal{T}}^{\mathrm{CME}}$ by using the $\mathbf{H}^1$-orthogonal projection (denoted by $\mathcal{P}_{\mathbf{X}_{\mathcal{T}}^{\mathrm{CME}}}$). Due to (4.20) and the coercivity of $\mathfrak{a}$ on $\mathbf{H}_{\mathbf{0}}^1(\Omega_{\mathcal{T}^{\mathrm{dof}}})$ there is a unique $\mathbf{u}^* \in \mathbf{X}_{\mathcal{T}^{\mathrm{dof}}}^{\mathbf{0}}$ satisfying the following discrete Stokes problem of Dirichlet type in $\Omega_{\mathcal{T}^{\mathrm{dof}}}$:

$$\begin{aligned}
\int_{\Omega_{\mathcal{T}^{\mathrm{dof}}}} \mathrm{D}(\mathbf{u}^*) : \mathrm{D}(\mathbf{v}) - \int_{\Omega_{\mathcal{T}^{\mathrm{dof}}}} p \ \mathrm{div}\, \mathbf{v} &= \mathbf{0} \qquad \forall \mathbf{v} \in \mathbf{X}_{\mathcal{T}^{\mathrm{dof}}}^{\mathbf{0}}, \\
\int_{\Omega_{\mathcal{T}^{\mathrm{dof}}}} q \ \mathrm{div}\, \mathbf{u}^* &= g(q) \quad \forall q \in \mathrm{M}_{\mathcal{T}^{\mathrm{dof}}} \cap \mathrm{L}_0^2(\Omega_{\mathcal{T}^{\mathrm{dof}}}),
\end{aligned}$$

where the linear form $g : \mathrm{M}_{\mathcal{T}^{\mathrm{dof}}} \cap \mathrm{L}_0^2(\Omega_{\mathcal{T}^{\mathrm{dof}}}) \to \mathbb{R}$ is given by $g(q) := \int_\Omega \tilde{\mathcal{E}}^{\mathrm{p}} q \ \mathrm{div}(\mathbf{u} - \mathcal{P}_{\mathbf{X}_{\mathcal{T}}^{\mathrm{CME}}} \mathbf{u})$. Defining $\pi\mathbf{u} := \mathcal{P}_{\mathbf{X}_{\mathcal{T}}^{\mathrm{CME}}} \mathbf{u} + \mathcal{E}^{\mathbf{0}} \mathbf{u}^*$ condition (4.21) is obviously fulfilled. The operator π is bounded because of the boundedness of the orthogonal projection and Lemma 4.6:

$$\begin{aligned}
\big\|\pi\mathbf{u}\big\|_{1,\Omega} &\le \big\|\mathbf{u}\big\|_{1,\Omega_{\mathcal{T}^{\mathrm{dof}}}} + C^* \Big(\sup_{0 \neq q \in \mathrm{M}_{\mathcal{T}^{\mathrm{dof}}} \cap \mathrm{L}_0^2(\Omega_{\mathcal{T}^{\mathrm{dof}}})} \frac{\big|g(q)\big|}{\big\|q\big\|_{0,\Omega_{\mathcal{T}^{\mathrm{dof}}}}} \Big) \\
&\le \underbrace{\big(1 + C^*(1 + 3C_{\mathcal{E}^{\mathrm{p}}})\big)}_{=C_\pi} \big\|\mathbf{u}\big\|_{1,\Omega_{\mathcal{T}^{\mathrm{dof}}}}. \qquad \square
\end{aligned}$$

It is remarkable that we deduced stability from the mini element with respect to the triangulation $\mathcal{T}^{\mathrm{dof}}$ and not from the continuous result on the complicated domain Ω.

Next, we have to investigate the coercivity of the bilinear form $\mathfrak{a}$ with respect to the discrete space $\mathbf{X}_{\mathcal{T}}^{\mathrm{CME}}$. We know from (2.3) that $\mathfrak{a}$ is coercive on $\mathbf{H}_{\mathbf{0}}^1(\Omega)$. Since $\mathbf{X}_{\mathcal{T}}^{\mathrm{CME}} \not\subseteq \mathbf{H}_{\mathbf{0}}^1(\Omega)$ this result needs to be extended to a certain neighborhood of $\mathbf{H}_{\mathbf{0}}^1(\Omega)$. This neighborhood can be measured in terms of the $\mathbf{L}^2$-norm of the trace, as can be seen in the following lemma, the proof of which (see [15]) is analogous to the proof for the Poincaré inequality.

LEMMA 4.8 (equivalent norms in $\mathbf{H}^1(\Omega)$). *For all* $\mathbf{u} \in \mathbf{H}^1(\Omega)$ *there holds* $\|\mathbf{u}\|_{1,\Omega}^2 \lesssim \mathfrak{a}(\mathbf{u},\mathbf{u}) + \|\mathbf{u}\|_{0,\partial\Omega}^2$.

Lemma 4.8 implies that $\mathfrak{a}$ is coercive on the composite space $\mathbf{X}_{\mathcal{T}}^{\mathrm{CME}}$ if the violation of the zero boundary condition is not too large. We will make this fact precise in the subsequent lemma by showing that the nonconformity in the space can be controlled by the local mesh refinement in the slave part $\mathcal{T}^{\mathrm{slave}}$ of the mesh $\mathcal{T}$, more precisely, by the ratio

$$\mathrm{R}(T) := \max_{t \in \mathcal{T}_T^{\mathrm{slave}} : t \cap \partial\Omega \neq \emptyset} \frac{\mathrm{diam}(t)}{\mathrm{diam}(T)}, \quad T \in \mathcal{T}^{\mathrm{dof}},$$

which can be assigned to every extrapolation simplex.

LEMMA 4.9 (nonconformity). *There is a constant* $C = C(\rho, C_3^{\mathcal{T}}) > 0$ *such that*

$$\big\|\mathbf{u}\big\|_{0,\partial\Omega} \le C\big(\max_{T \in \mathcal{T}^{\mathrm{dof}}} \mathrm{R}(T)\big) h^{\frac{1}{2}} \big|\mathbf{u}\big|_{1,\Omega} \quad \forall \mathbf{u} \in \mathbf{X}_{\mathcal{T}}^{\mathrm{CME}}.$$

Proof. Let $t \in \mathcal{T}^{\text{slave}}$ such that $t \cap \partial\Omega \neq \emptyset$. We start by estimating the value of $\mathbf{u} \in \mathbf{X}_{\mathcal{T}}^{\text{CME}}$ at a vertex $\mathbf{x}$ of t:

$$\big|\mathbf{u}(\mathbf{x})\big| \overset{(3.4)}{\leq} \operatorname{diam}(t)\big|\nabla(\mathbf{u}|_{T_{\mathbf{x}}})\big| = \frac{\operatorname{diam}(t)}{\big|T_{\mathbf{x}}\big|^{\frac{1}{2}}}\big|\mathbf{u}\big|_{1,T_{\mathbf{x}}}. \tag{4.22}$$

Now we can estimate the $\mathbf{L}^2$-norm of $\mathbf{u}$ on $\partial\Omega$:

$$\begin{aligned}
\big\|\mathbf{u}\big\|_{0,\partial\Omega}^2 &\leq \sum_{T\in\mathcal{T}^{\text{dof}}} \sum_{t\in\mathcal{T}_T^{\text{slave}}:t\cap\partial\Omega\neq\emptyset} \big|\partial\Omega\cap t\big|\big|\mathbf{u}\big|_{\infty,t}^2 \\
&\overset{(4.22),(2.6)}{\lesssim} \sum_{T\in\mathcal{T}^{\text{dof}}} \sum_{t\in\mathcal{T}_T^{\text{slave}}:t\cap\partial\Omega\neq\emptyset} \frac{\big|\partial\Omega\cap t\big|\operatorname{diam}(t)^2}{\operatorname{diam}(T)^d}\big|\mathbf{u}\big|_{1,T}^2 \\
&\overset{(2.6)}{\lesssim} \sum_{T\in\mathcal{T}^{\text{dof}}} \big|\partial\Omega\cap\mathcal{T}_T^{\text{slave}}\big|\frac{\operatorname{diam}(t)^2}{\operatorname{diam}(T)^d}\big|\mathbf{u}\big|_{1,T}^2 \\
&\overset{(2.6)}{\lesssim} \underbrace{\left(\max_{T\in\mathcal{T}^{\text{dof}}} \frac{\big|\partial\Omega\cap\mathcal{T}_T^{\text{slave}}\big|}{\operatorname{diam}(T)^{(d-1)}}\right)}_{=:C_3^{\mathcal{T}}} \frac{\operatorname{diam}(t)^2}{\operatorname{diam}(T)}\big|\mathbf{u}\big|_{1,\Omega}^2. \qquad \square
\end{aligned}$$

In Lemma 4.9 we have seen that the nonconformity in the velocity space can be controlled by the ratios $\mathrm{R}(T)$ independent from the mesh size h. This fact will be important for the error estimate in the subsequent section. For the unique solvability we want to avoid constraints on $\mathrm{R}(T)$. The right-hand side in Lemma 4.9 always contains at least a factor $\sqrt{h}$. So, in view of Lemma 4.8, there is an h_0 such that the bilinear form is coercive for all triangulations with mesh size $h \leq h_0$. The case $h > h_0$ is discussed in what follows. The bilinear $\mathfrak{a}$ has a nontrivial kernel given by the set of rigid body motions

$$\mathcal{R} := \{\mathbf{A}\cdot + \mathbf{b} \mid \mathbf{A} \in \mathbb{R}^{d\times d} \text{ skew symmetric}, \mathbf{b} \in \mathbb{R}^d\}, \tag{4.23}$$

and it is therefore coercive on a subspace $\mathbf{U} \subseteq \mathbf{H}^1(\Omega)$ if and only if $\mathbf{U} \cap \mathcal{R} = \{\mathbf{0}\}$. This is the key to prove discrete coercivity.

LEMMA 4.10 (discrete coercivity). *There is a constant α^{CME} that does not depend on h such that $\mathfrak{a}(\mathbf{u},\mathbf{u}) \geq \alpha^{\text{CME}}\big\|\mathbf{u}\big\|_{1,\Omega}^2$ for all $\mathbf{u} \in \mathbf{X}_{\mathcal{T}}^{\text{CME}}$.*

Proof. Let $\mathbf{u} \in \mathbf{X}_{\mathcal{T}}^{\text{CME}}$, $\mathbf{A} \in \mathbb{R}^{d\times d}$ be skew symmetric and $\mathbf{b} \in \mathbb{R}^d$ such that $\mathbf{u}(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{b}$. Then, by definition, $\mathbf{u}(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{b} = \mathbf{A}(\mathbf{x} - \bar{\mathbf{x}})$ in a slave node $\mathbf{x}$, and we get $\mathbf{A}\bar{\mathbf{x}} = -\mathbf{b}$ for all $\bar{\mathbf{x}} \in \{\bar{\mathbf{x}}_i \mid \mathbf{x}_i \in \Theta^{\text{slave}}\}$. Since $\operatorname{span}\{\bar{\mathbf{x}}_i \mid \mathbf{x}_i \in \Theta^{\text{slave}}\} = \mathbb{R}^d$ this can be true only if $\mathbf{A} = \mathbf{0}$ and $\mathbf{b} = \mathbf{0}$. $\square$

4.3. A priori error estimate. This section provides the main result on the convergence of the CME method.

THEOREM 4.11. *The discrete problem* (3.6) *always has a unique solution. Furthermore, if $(\mathbf{u}^*, p^*) \in \big(\mathbf{H}_{\mathbf{0}}^1 \cap \mathbf{H}^{1+r}(\Omega)\big) \times \big(L_0^2(\Omega) \cap H^r(\Omega)\big)$, $r \in (\frac{1}{2}, 1]$, is the solution of* (2.1)*, then we have the following a priori error estimate:*

$$\big\|\mathbf{u}^* - \mathbf{u}\big\|_{1,\Omega} + \big\|p^* - p\big\|_{0,\Omega} \leq C\left(C_{\text{appr}}h^r + \sup_{\mathbf{0}\neq\mathbf{v}\in\mathbf{X}_{\mathcal{T}}^{\text{CME}}} \frac{\big\|\mathbf{v}\big\|_{0,\partial\Omega}}{\big\|\mathbf{v}\big\|_{1,\Omega}}\right)\big\|\mathbf{f}\big\|_{r-1,\Omega},$$

where the constant C depends on α^{CME}, β^{CME}, while C_{appr} depends on the constants in Theorems 4.4 *and* 4.5.

Furthermore, if $\Omega = \Omega_{\mathcal{T}}$ or $\mathrm{R}(T) \lesssim h^r$ for all $T \in \mathcal{T}^{\mathrm{dof}}$, then

$$\|\mathbf{u}^* - \mathbf{u}\|_{1,\Omega} + \|p^* - p\|_{0,\Omega} \lesssim h^r \|\mathbf{f}\|_{r-1,\Omega}.$$

Proof. The unique solvability is a consequence of Theorem 4.7 and Lemma 4.10. Following the abstract mixed theory presented in [4], especially Proposition 2.16, we additionally get the following error bound:

$$\|\mathbf{u}^* - \mathbf{u}\|_{1,\Omega} + \|p^* - p\|_{0,\Omega} \overset{T4.10,T4.7}{\lesssim} \inf_{\mathbf{v}\in\mathbf{X}_{\mathcal{T}}^{\mathrm{CME}}} \|\mathbf{u}^* - \mathbf{v}\|_{1,\Omega} + \inf_{q\in\mathrm{M}_{\mathcal{T}}^{\mathrm{CME}}} \|p^* - q\|_{0,\Omega} + \underbrace{\sup_{\mathbf{0}\neq\mathbf{v}\in\mathbf{X}_{\mathcal{T}}^{\mathrm{CME}}} \frac{\left|\mathfrak{a}(\mathbf{u}^*,\mathbf{v}) + \mathfrak{b}(\mathbf{v},p^*) - \langle \mathbf{f},\mathbf{v}\rangle_{0,\Omega}\right|}{\|\mathbf{v}\|_{1,\Omega}}}_{=:K}.$$

The infima have been estimated in section 4. The supremum which reflects the error due to nonconformity in the space can be estimated by using the identity

$$\int_\Omega \mathbf{D}\mathbf{u}^* : \mathbf{D}\mathbf{v} + \int_\Omega p^* \operatorname{div}\mathbf{v} - \langle \mathbf{f},\mathbf{v}\rangle_{0,\Omega} = \int_{\partial\Omega} \langle (\mathbf{D}\mathbf{u}^* - p^*\mathbf{I_d})\boldsymbol{\nu},\mathbf{v}\rangle\,, \tag{4.24}$$

which holds for the solution $(\mathbf{u}^*, p^*)$ and an arbitrary $\mathbf{v} \in \mathbf{H}^1(\Omega)$, and the trace theorem

$$\|\mathbf{D}\mathbf{u}^*\|_{0,\partial\Omega}^2 + \|p^*\|_{0,\partial\Omega}^2 \lesssim \|\mathbf{u}^*\|_{\frac{3}{2},\Omega}^2 + \|p^*\|_{\frac{1}{2},\Omega}^2. \tag{4.25}$$

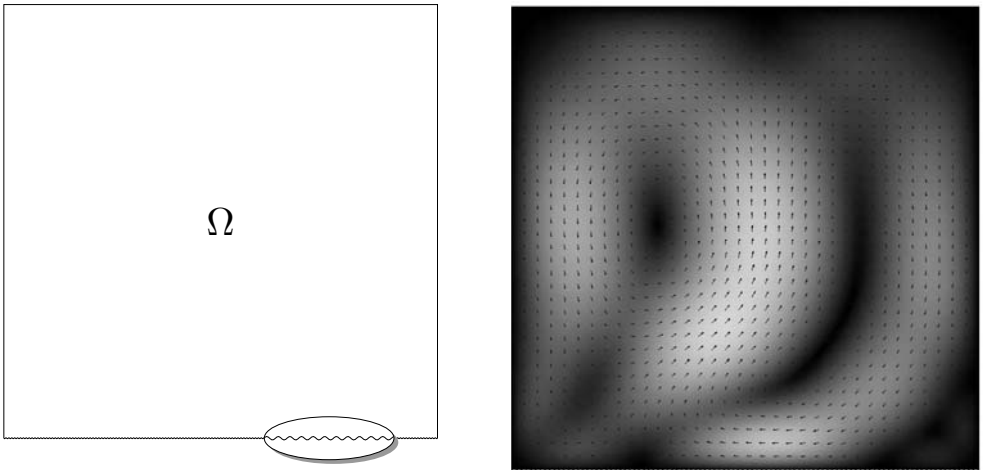
Therefore the nonconformity error K can be estimated as follows:

$$K \overset{(4.24),(4.25)}{\lesssim} \left(\|\mathbf{u}^*\|_{\frac{3}{2},\Omega}^2 + \|p^*\|_{\frac{1}{2},\Omega}^2\right)^{\frac{1}{2}} \sup_{\mathbf{0}\neq\mathbf{v}\in\mathbf{X}_{\mathcal{T}}^{\mathrm{CME}}} \frac{\|\mathbf{v}\|_{0,\partial\Omega}}{\|\mathbf{v}\|_{1,\Omega}}.$$

Finally, the assertion follows from the regularity estimate (2.5) and the interpolation theory of Sobolev spaces (cf. [3]). ∎

5. Numerical experiments. Typical applications of the CME are flow problems on domains with rough outer boundaries which arise, for instance, in the modeling of rivers, lakes, and oceans where the shorelines are rarely smooth. Model problems of this type have been investigated in [15]. The results show that the theoretical estimates of the previous section are sharp for the test problems. Performance tests for composite finite elements for the Poisson equations and elasticity problems can be found in [17] and [16], where especially the use of overlapping meshes has been investigated. The subsequent experiments are addressed to flow problems on complicated domains with mixed boundary conditions. Our first model problem on the unit square with 100 randomly distributes circular holes of diameter 0.005 is depicted in Figure 1.1:

$$\begin{aligned}
-\Delta\mathbf{u} + \nabla p &= \mathbf{0} && \text{in } \Omega,\\
\operatorname{div}\mathbf{u} &= 0 && \text{in } \Omega,\\
\mathbf{u}(\mathbf{x}) &= 0.5\,(1+\cos(8\pi(\mathbf{x}_2 - 0.75))) && \forall \mathbf{x}\in\Gamma^{\mathrm{in}},\\
2(\mathbf{D}\mathbf{u}(\mathbf{x}))\boldsymbol{\nu}(\mathbf{x}) &= p(\mathbf{x})\boldsymbol{\nu}(\mathbf{x}) && \forall \mathbf{x}\in\Gamma^{N},\\
\mathbf{u}(\mathbf{x}) &= \mathbf{0} && \forall \mathbf{x}\in\partial\Omega\setminus(\Gamma^{\mathrm{in}}\cup\Gamma^{N}),
\end{aligned} \tag{5.1}$$



(a) Model domain $\Omega := \{\mathbf{x} \in \mathbb{R}^2 \mid 0 < \mathbf{x}_1 < 1,\ g(\mathbf{x}_1) < \mathbf{x}_2 < 1\}$, $g(\mathbf{x}_1) := a\sin(\omega\pi\mathbf{x}_1)$, $a := 0.001$, $\omega := 199$.

(b) Solution velocity (black = 0, white = 0.5).

FIG. 5.1. *A model problem: Force-driven Stokes flow in a domain with a rough slip bottom boundary.*

where $\Gamma^{\text{in}} := \{0\} \times (\frac{5}{8}, \frac{7}{8})$ and $\Gamma^N := \{1\} \times (\frac{1}{8}, \frac{3}{8}) \cup (\frac{5}{8}, \frac{7}{8})$. The difficulties of problem (5.1) not only stem from the holes but are also related to the mixed boundary conditions. The second test problem (see Figure 5.1) models a force-driven flow in a perturbed unit square (rough bottom boundary) with mixed Dirichlet and slip boundary conditions:

$$
\begin{aligned}
-\Delta\mathbf{u} + \nabla p &= \mathbf{f} && \text{in } \Omega,\\
\operatorname{div}\mathbf{u} &= 0 && \text{in } \Omega,\\
\mathbf{u}(\mathbf{x}) &= \mathbf{0} && \forall \mathbf{x} \in \Gamma^D,\\
\langle \mathbf{u}(\mathbf{x}), \boldsymbol{\nu}(\mathbf{x})\rangle &= 0 && \forall \mathbf{x} \in \partial\Omega \setminus \Gamma^D,\\
(\mathbf{D}\mathbf{u}(\mathbf{x}))\boldsymbol{\nu}(\mathbf{x}) &= \big(\boldsymbol{\nu}(\mathbf{x})^T(\mathbf{D}\mathbf{u}(\mathbf{x}))\boldsymbol{\nu}(\mathbf{x})\big)\,\boldsymbol{\nu}(\mathbf{x}) && \forall \mathbf{x} \in \partial\Omega \setminus \Gamma^D,
\end{aligned}
\tag{5.2}
$$

where $\mathbf{f}(\mathbf{x}) := (\cos(2\pi\mathbf{x}_2)\sin(2\pi\mathbf{x}_1), (\frac{3}{2} - \frac{3}{2}\mathbf{x}_2)^5\sin(\pi\mathbf{x}_2)\cos(2\pi\mathbf{x}_1))^T$, and $\Gamma^D := \{0\} \times [0,1] \cup [0,1] \times \{1\} \cup \{1\} \times [0,1]$. Analytical solutions of (5.1) or (5.2) are not known; approximations on very fine meshes will be used for the error indication instead. We will investigate the mini element spaces with respect to coarse shape-regular meshes (see, e.g., Figures 1.2 and 5.4) for the domains of (5.1) and (5.2). They resolve the geometric details of the domain, i.e., the rough boundaries and holes, as well as the inflow and outflow boundaries. The definition of the method in section 3 has left some freedom in the choice of the inner mesh $\mathcal{T}^{\text{dof}}$ determining the CME space. We will use the simplest possible rule:

$$t \in \mathcal{T}^{\text{dof}} \quad \Leftrightarrow \quad \operatorname{dist}(t, \partial\Omega) > \tfrac{1}{2}h^{\text{slave}}.$$

In the case of boundary concentrated meshes as depicted in Figures 1.2 and 5.4 this choice leads to moderate constants in condition (4.1), at least for $h^{\text{slave}} \in (0, h)$. Note that this choice does not take any special properties of the model problems into

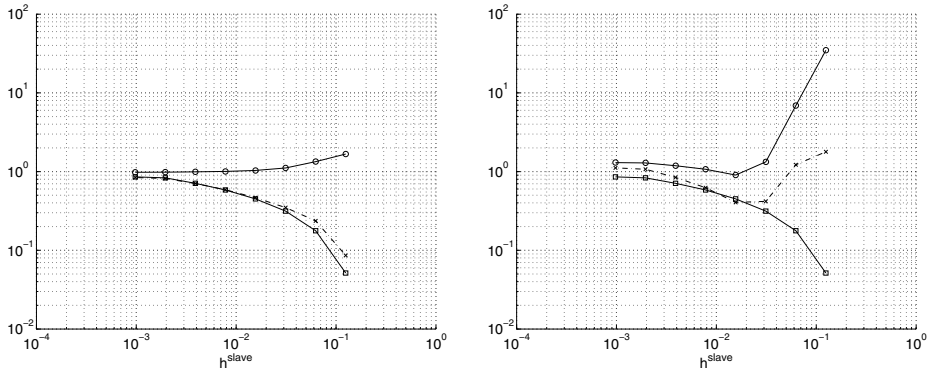
account. The sparse linear system $\mathbf{S}^{\mathrm{CME}}\mathbf{w}^{\mathrm{CME}} = \mathbf{g}$, which arises from the CME discretization of (5.1) and (5.2), can be assembled locally in the usual way (see [11], [16], and [15]). The resulting system matrix is $\mathbf{S}^{\mathrm{CME}}$ fulfills

$$\mathbf{S}^{\mathrm{CME}} = (\mathbf{E}^{\mathrm{CME}})^T \mathbf{S} \mathbf{E}^{\mathrm{CME}},$$

where $\mathbf{E}^{\mathrm{CME}}$ is the matrix representation of the linear operator $\boldsymbol{\mathcal{E}}^{\mathrm{CME}}$ presented in (3.5). $\mathbf{E}^{\mathrm{CME}}$ is a sparse rectangular matrix with order $\sharp\mathcal{T}$ number of rows but only order $\sharp\mathcal{T}^{\mathrm{dof}}$ number of columns. It is obvious that the composite system matrix $\mathbf{S}^{\mathrm{CME}}$ is (much) smaller than the classical mini element matrix $\mathbf{S}$. Sparsity can thereby be preserved since the support of the CME basis functions remains local. As a consequence the storage requirements are reduced, and a (possibly) much smaller system has to be solved.

Before we present numerical results we will comment on the handling of the boundary conditions that are not covered by the theory presented before. The inhomogeneous Dirichlet boundary condition can be handled in the standard way. By introducing a discrete vector field $\mathbf{u}_0 : \Omega \to \mathbb{R}^2$ that approximates the inflow condition sufficiently well the problem can be reduced to a homogeneous problem with a modified right-hand side. Note that $\mathbf{u}_0$ can be chosen in the full mini element space based on a possibly fine mesh close to the inhomogeneous Dirichlet boundary which allows us to handle even complicated Dirichlet data. It is therefore independent from the choice of h^{slave}. From the pressure part of the CME space we know how to deal with functions that are unconstrained on the domain boundary. We simply use $\mathcal{E}^{\mathrm{p}}$ componentwise to define the values of the velocity test functions close to the Neumann boundary. For the slip boundary condition in (5.2) we use $\mathcal{E}^{\mathrm{p}}$ in the tangential direction and $\boldsymbol{\mathcal{E}}^{\mathbf{0}}$ in the normal direction. A theoretical justification of this procedure has been presented in [15], where the use of the slip boundary condition is discussed in detail. The question of which extension has to be used in a slave node $\mathbf{x}$ is decided by its closest boundary point $\bar{\mathbf{x}}$ that belongs to either the Dirichlet, Neumann, or slip boundary.

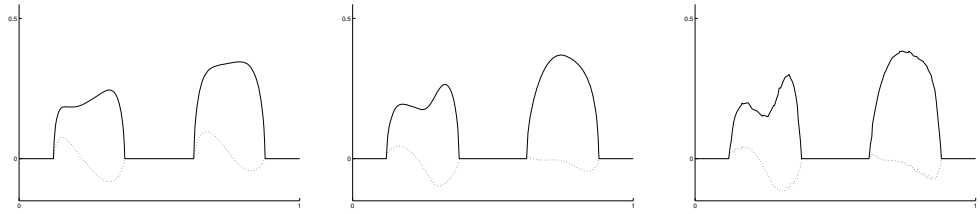
We will use problem (5.1) to investigate the behavior of the error and the system dimension on two coarse triangulations for varying choices of the parameter h^{slave}. Figure 5.2 summarizes the results: In both plots the lower curves ($\square$) represent the compression rates, that is, the quotients between the dimensions of the composite spaces for varying values of h^{slave} and the classical mini element space. The upper curves ($\circ$) show the relative errors (left: velocity, right: pressure), more precisely, the quotient of the errors of the composite approximations and the full mini element approximation. In between, the dotted curves ($\times$) are the product of compression rate and relative error. It can be seen that the errors stay almost constant as h^{slave} tends to h. Since the space dimension rapidly decreases at the same time, the efficiency of the composite method increases. Only for values of $h^{\mathrm{slave}} > h$ does the pressure error become large, which indicates that the pressure extension depends more critically on the choice of h^{slave}. This is not surprising, since information from the boundary conditions is not available. However, the velocity error in Figure 5.2(a) is not affected from the worse pressure approximation, and the behavior is fine if $h^{\mathrm{slave}} < h$. In Figure 5.3 it can be observed that the generalization of the composite method to Neumann boundaries works well for the model problem under consideration. Due to the extension there is an expected loss in smoothness, but the general outflow behavior can be captured without placing additional degrees of freedom. This indicates a certain potential of the method not only for problems with complicated boundaries



(a) Relative $\mathbf{H}^1(\Omega)$ velocity error ($\circ$) and relative space dimension ($\square$) versus h^{slave}, $h = 0.05$.

(b) Relative $\mathrm{L}^2(\Omega)$ pressure error ($\circ$) and relative space dimension ($\square$) versus h^{slave}, $h = 0.05$.

FIG. 5.2. *Relative errors* $\frac{error(CME)}{error(Mini)}$ *in suitable Sobolev norms* ($\circ$) *and relative system dimensions* $\frac{dim(CME)}{dim(Mini)}$ ($\square$) *for resolving triangulations with* $h = 0.05$ *under the variation of the parameter* $h^{\text{slave}} \in (10^{-3}, 1)$. *The product of relative error and relative dimension, depicted as a dashed line, indicates the efficiency of the coarsening process.*



(a) Reference output.

(b) Mini element output, space dimension $\approx 7 \cdot 10^4$.

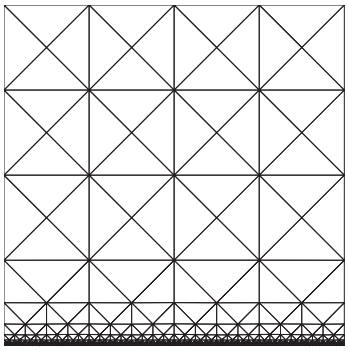
(c) CME output, space dimension $\approx 10^4$.

FIG. 5.3. *Outflow behavior of the model problem of Figure* 1.1(a): *Velocity components* (x_1 *solid,* x_2 *dotted) on* $\{0\} \times (0,1)$ *and on* $\{1\} \times (0,1)$ *for the reference approximation* (a), *the full mini element approximation* (b) *on the triangulation from Figure* 1.2, *and composite approximation* (c) *for* $h^{\text{slave}} \approx \frac{h}{4} \approx 0.025$.

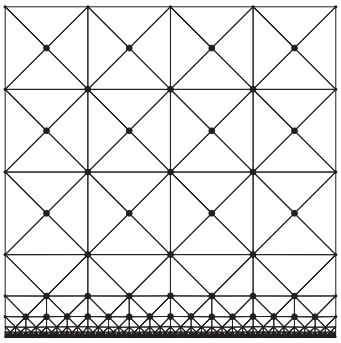
but also with complicated boundary conditions, i.e., complicated Dirichlet data or complicated distribution of mixed Dirichlet–Neumann boundary conditions.

Finally, we use the second model problem (5.2) to investigate the convergence rates as the mesh size h decreases. Thereby we compare the classical mini element on a resolving triangulation (Mini), its composite version (CME) for $h^{\text{slave}} = 0.1h$, and a classical mini element with respect to nonmatching, quasi-uniform triangulations (uniformMini). All approaches and the according parameter choices are summarized in Figure 5.4.

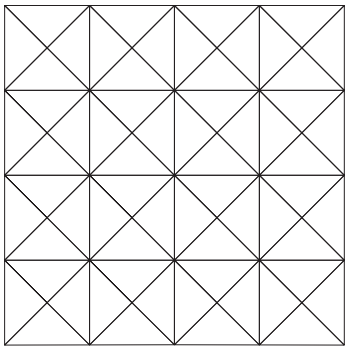
The results of the slip model computations are depicted in Figure 5.5, where the dependence of the velocity and pressure errors on the mesh size and the system



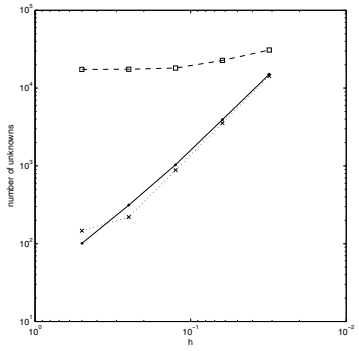
(a) Classical mini element (Mini) on a resolving mesh (minimal mesh size 2^{-10}).



(b) Composite mini element (CME) on a resolving mesh with marked degrees of freedom (•) (minimal mesh size 2^{-10}, $h^{\text{slave}} = h$).



(c) Nonconforming classical mini element (uniformMini) on quasi-uniform mesh.

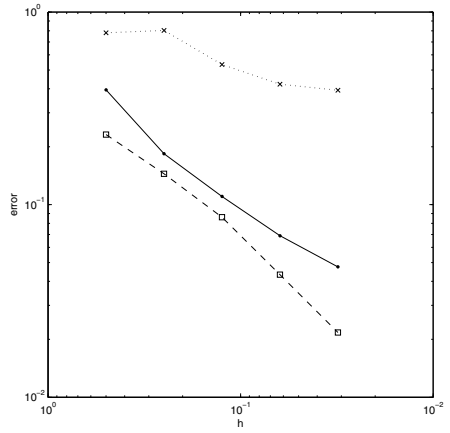


(d) System dimension as a function of h: Mini (dashed), CME (solid), uniformMini (dotted).

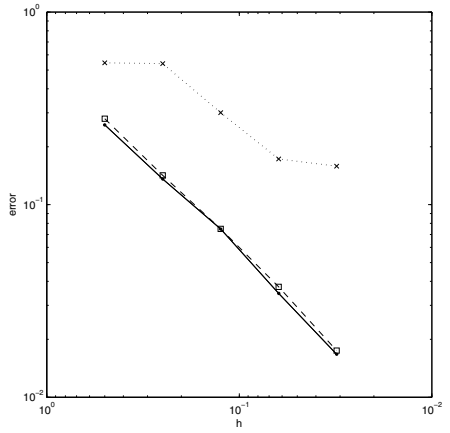
FIG. 5.4. *The grids with degrees of freedom of the different approaches for the solution of the model problem* (5.2).

dimension are depicted. We make the following observations:

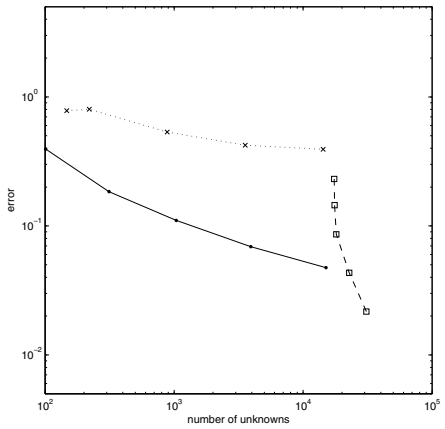
1. The (conforming) mini element and its composite version converge at the predicted optimal (linear) rate (cf. Figure 5.5(a)). The compressed method CME is only slightly worse than the full version (cf. Figure 5.5(a)). Note that the error of the composite method is larger than full mini error only in a near boundary zone. Away from the rough boundary the errors are almost equal; i.e., pollution effects cannot be observed for this test case (cf. Figures 5.5(b)).
2. The quasi-uniform approach is not competitive (cf. Figures 5.5(a)–(b)). The reason is that the very crude approximation of the boundary conditions spoils the overall discretization too significantly. That means that our choice of minimal mesh size for the boundary resolution cannot be weakened without increasing the errors.



(a) Error on Ω as a function of h.



(b) Error on $(0,1)\times(0.1,1)$ as a function of h.



(c) Error on Ω as a function of the system dimension.

FIG. 5.5. *Convergence of the methods applied to model problem* (5.2). *The errors (sum of H^1-velocity and L^2-pressure error) are plotted versus the maximal mesh size h and the system dimension: Mini* (dashed, $\square$), CME (solid, $\circ$), uniformMini (dotted, $\times$).

3. The dimension of the CME space behaves like h^{-d} (cf. Figure 5.4(d)). That means that the CME preserves the accuracy of the full mini element space on the coarser levels, while the cost is comparable with the quasi-uniform approach. This underpins the efficiency of our "fuzzy" treatment of the boundary conditions.
4. The composite method is the most efficient method in this test case. Though the full mini element produces slightly smaller errors than the composite element, the resulting system dimension is up to 100 times bigger than for CME on the coarsest level (see Figure 5.5(c)). In contrast to the CME the less costly quasi-uniform method produces unsatisfying approximations.

The gain of the composite method depends on the maximal mesh size h. If h is small enough to resolve the domain, the composite and the full method are almost equal in cost and accuracy. If one is satisfied with only moderate errors, or one is restricted to smaller dimensions, then the composite method is the better choice. Refinements in the triangulation which are due to geometric issues do not need to be treated by degrees of freedom.

REFERENCES

[1] R. E. Bank and J. Xu, *An algorithm for coarsening unstructured meshes*, Numer. Math., 73 (1996), pp. 1–36.

[2] J. W. Barrett and C. M. Elliott, *A finite-element method for solving elliptic equations with Neumann data on a curved boundary using unfitted meshes*, IMA J. Numer. Anal., 4 (1984), pp. 309–325.

[3] S. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, Springer-Verlag, New York, 1994.

[4] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer-Verlag, New York, 1991.

[5] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North–Holland, Amsterdam, 1978.

[6] D. Cioranescu, P. Donato, and H. I. Ene, *Homogenization of the Stokes problem with non-homogeneous slip boundary conditions*, Math. Methods Appl. Sci., 19 (1996), pp. 857–881.

[7] Ph. Clément, *Approximation by finite element functions using local regularization*, RAIRO Anal. Numér., 9 (1975), pp. 77–84.

[8] A. Damlamian and P. Donato, *Homogenization with small perforations of increasingly complicated shapes*, SIAM J. Math. Anal., 22 (1991), pp. 639–652.

[9] G. Duvaut and J. L. Lions, *Inequalities in Mechanics and Physics*, Springer-Verlag, Berlin, Heidelberg, New York, 1976.

[10] V. Girault and P.-A. Raviart, *Finite Element Approximation of the Navier-Stokes Equations*, Springer-Verlag, Berlin, Heidelberg, New York, 1979.

[11] W. Hackbusch and S. A. Sauter, *Composite finite elements for problems containing small geometric details. Part* II: *Implementation and numerical results*, Comput. Vis. Sci., 1 (1997) pp. 15–25.

[12] W. Hackbusch and S. A. Sauter, *Composite finite elements for the approximation of PDEs on domains with complicated micro-structures*, Numer. Math., 75 (1997), pp. 447–472.

[13] R. Kornhuber and H. Yserentant, *Multilevel methods for elliptic problems on domains not resolved by the coarse grid*, in Domain Decomposition Methods in Scientific and Engineering Computing, Contemp. Math. 180, AMS, Providence, RI, 1994, pp. 49–60.

[14] J. A. Nitsche, *On Korn's second inequality*, RAIRO Anal. Numér., 15 (1981), pp. 237–248.

[15] D. Peterseim, *The Composite Mini Element: A Mixed FEM for the Stokes Equations on Complicated Domains*, Ph.D. thesis, Universität Zürich, Zürich, Switzerland, 2007.

[16] M. Rech, *Composite Finite Elements: An Adaptive Two-Scale Approach to the Nonconforming Approximation of Dirichlet Problems on Complicated Domains*, Ph.D. thesis, Universität Zürich, Zürich, Switzerland, 2006.

[17] M. Rech, S. Sauter, and A. Smolianski, *Two-scale composite finite element method for Dirichlet problems on complicated domains*, Numer. Math., 102 (2006), pp. 681–708.

[18] S. A. Sauter and R. Warnke, *Extension operators and approximation on domains containing small geometric details*, East-West J. Numer. Math., 7 (1999), pp. 61–77.

[19] L. R. Scott and S. Zhang, *Finite element interpolation of nonsmooth functions satisfying boundary conditions*, Math. Comp., 54 (1990), pp. 483–493.

[20] E. M. Stein, *Singular Integrals and Differentiablity Properties of Function*, Princeton University Press, Princeton, NJ, 1970.

[21] R. Temam, *Navier-Stokes Equations, Theory and Numerical Analysis*, North–Holland, Amsterdam, 1984.

[22] R. Verfürth, *A Review of A Posteriori Error Estimation and Adaptive Mesh-Refinement Techniques*, Teubner-Wiley, Stuttgart, 1996.

[23] R. Verfürth, *Error estimates for some quasi-interpolation operators*, M2AN Math. Model. Numer. Anal., 33 (1999), pp. 695–713.

[24] H. Yserentant, *Coarse grid spaces for domains with a complicated boundary*, Numer. Algorithms, 21 (1999), pp. 387–392.

SIAM J. NUMER. ANAL.
Vol. 46, No. 6, pp. 3207–3232
© 2008 Society for Industrial and Applied Mathematics

FULLY COMPUTABLE BOUNDS FOR THE ERROR IN NONCONFORMING FINITE ELEMENT APPROXIMATIONS OF ARBITRARY ORDER ON TRIANGULAR ELEMENTS*

MARK AINSWORTH† AND RICHARD RANKIN†

Abstract. We obtain a fully computable a posteriori error bound on the broken energy norm of the error in the nonconforming finite element approximation on triangles of arbitrary order of a linear second order elliptic problem with variable permeability. The estimator is completely free of unknown constants and provides a guaranteed numerical bound on the broken energy norm of the error. This estimator is shown to be efficient in the sense that it also provides a lower bound for the broken energy norm of the error up to a constant and higher order data oscillation terms.

Key words. robust a posteriori error estimation, nonconforming finite element

AMS subject classifications. 65N15, 65N30

DOI. 10.1137/07070838X

1. Introduction.

A posteriori error estimation is important for determining the accuracy of finite element approximations and implementing adaptive refinement strategies. The papers referenced in [10] provide a recent survey of the state of the art in a posteriori error estimation for nonconforming finite element approximations. However, none of the a posteriori error estimators found in [10] provide actual *computable* upper bounds on the error, apart from those described in [4, 6]. While such error indicators have a role to play, if one really wants to estimate the *value* of the error, then a fresh approach is required.

In [5] computable bounds were obtained for the broken energy norm of the error in the lowest order nonconforming Crouzeix–Raviart finite element approximation on triangles [13] of a linear second order elliptic problem with variable permeability. In this estimator only the constants in front of the oscillation of the data were not specified explicitly but could be obtained using techniques described in [3]. Subsequently, this approach was extended to the nonconforming Fortin–Soulie triangular element of second order [16] with the resulting estimator being completely free of unknown constants [3]. Unfortunately, the approach of [5] could not be applied directly to the Fortin–Soulie elements since, as noted in [16], nonconforming triangular elements of *even* order have a very different structure from the nonconforming triangular elements of *odd* order. The estimators in both [5] and [3] also provided, up to a positive constant independent of the size of any of the elements in the mesh and higher order data oscillation terms, local lower bounds on the broken energy norm of the error.

The objective of the present work is to obtain computable estimators, which are completely free of all unknown constants, for nonconforming triangular elements of arbitrary order [9, 19]. The estimators and techniques used in [3, 5] are not readily generalized to the case of elements of arbitrary order, and a different approach is required. The robustness of the new estimators with respect to jumps in the perme-

*Received by the editors November 16, 2007; accepted for publication (in revised form) May 9, 2008; published electronically October 17, 2008.
http://www.siam.org/journals/sinum/46-6/70838.html
†Mathematics Department, University of Strathclyde, 26 Richmond Street, Glasgow G1 1XH, Scotland (M.Ainsworth@strath.ac.uk, rs.rran@maths.strath.ac.uk). The first author's research was partially supported by the Engineering and Physical Sciences Research Council of Great Britain under grant EP/E040993. The second author was partially supported through a research studentship.

ability coefficient is achieved under the same conditions as in [3, 5], and the estimators provide a local lower bound up to a constant on the broken energy norm of the error plus higher order data oscillation terms.

One advantage of treating elements of arbitrary order together at the same time is that the basic underlying structure of the approach is highlighted. In particular, the fundamental difference in the structure of the odd and even order nonconforming elements also manifests itself in the structure of the estimators. More specifically, in the lowest order case considered in [5] the jumps in the normal derivative of the finite element approximation did not appear, whereas these were found to be necessary in the case of second order elements considered in [3]. Rather surprisingly the relationship between the need to include normal jump terms and the parity of the order extends to the case of arbitrary order nonconforming finite element approximations. This phenomenon is reminiscent of what was witnessed in [8] in the context of biquadratic elements and [1] for higher order elements although in both these cases it was the estimators for the even order elements that were free of terms involving jumps in the normal derivative in contrast to what we find here. The absence of jumps in normal fluxes in the estimator in the case of elements of odd order makes the use of such elements more attractive since the calculation of the jumps in the normal derivative is relatively computationally expensive.

The only other fully computable a posteriori error estimators for nonconforming finite element approximations on triangular elements of which we are aware are those in [3, 5]. The estimators we obtain in this paper for the lowest order odd and even nonconforming finite element approximations differ slightly from those obtained previously in [5], [3]. In particular, in the case of the Crouzeix–Raviart element the treatment of the data oscillation differs from that in [5], although if the data are piecewise constant on the mesh, then the estimators are identical. Similarly, the treatment of the data oscillation for the Fortin–Soulie element differs from that in [3], and in addition the estimator is constructed using a modified projection operator which is more naturally extended to the higher order elements. In practice both estimators gave virtually identical results for smooth data.

The remainder of the paper is organized as follows. In section 2, we describe the finite element schemes and introduce the notation. Following [14] we decompose the broken energy norm of the error into conforming and nonconforming components. Fully computable upper bounds, as well as lower bounds, for each of these components are derived in sections 3 and 4. Finally, in section 5 we summarize the main results and present a numerical example illustrating the theory. An appendix is provided where we give a key result concerning the existence of polynomial solutions of a local boundary value problem with polynomial data.

2. Preliminaries.

2.1. Model problem.

Consider the model problem

$$-\operatorname{div}(a\,\mathbf{grad}\,u) = f \text{ in } \Omega$$

subject to $u = q$ on Γ_D and $a\partial u/\partial \boldsymbol{n} = g$ on Γ_N, where Ω is a simple plane polygonal domain, the disjoint sets Γ_D (nonempty) and Γ_N form a partitioning of the boundary $\Gamma = \partial\Omega$ of the domain, and $\boldsymbol{n}$ is the outward unit normal vector to Γ_N. The data satisfy $f \in L_2(\Omega)$, $g \in L_2(\Gamma_N)$, and $q \in H^1(\Gamma_D)$, and $a \in L_\infty(\Omega)$ is piecewise constant. We will assume that a is positive and that the value of a across the boundary of a subdomain of Ω may undergo a jump of many orders of magnitude, corresponding to a transition between regions of widely differing permeability.

The variational form of the problem consists of finding $u \in H^1(\Omega)$ such that $u = q$ on Γ_D and

$$(a\,\mathbf{grad}\,u, \mathbf{grad}\,v) = (f,v) + (g,v)_{\Gamma_N} \;\; \forall v \in H^1_D(\Omega), \tag{2.1}$$

where $H^1_D(\Omega) = \{v \in H^1(\Omega) : v = 0 \text{ on } \Gamma_D\}$. We shall use the notation $(\cdot,\cdot)_\omega$ to denote the integral inner product over a region or line segment ω and omit the subscript in the case where ω is the physical domain Ω.

Consider a family of partitions $\{\mathcal{P}\}$ of the domain Ω into the union of nonoverlapping, shape-regular triangular elements such that the nonempty intersection of a distinct pair of elements is a single common node or single common edge. Consequently, the family of partitions is locally quasi-uniform in the sense that the ratio of the diameters of any pair of neighboring elements is uniformly bounded above and below over the whole family.

2.2. Nonconforming finite element approximation. Let K denote an individual element in $\mathcal{P}$, let $\partial\mathcal{P}$ denote the set of element edges, and let γ denote an individual edge in $\partial\mathcal{P}$. Let $\mathbb{P}_m(K)$ denote the space of polynomials on $K \in \mathcal{P}$ of total degree at most m, and let $\mathbb{P}_m(\gamma)$ denote the space of polynomials on $\gamma \subset \partial\mathcal{P}$ of total degree at most m (with respect to the arc length parameter).

Henceforth, fix the order $n \geq 1$ of approximation. The order n nonconforming finite element space on triangles is defined by

$$X^n = \Big\{ v : \Omega \to \mathbb{R} : v_{|K} \in \mathbb{P}_n(K) \;\; \forall K \in \mathcal{P}, \\ \Big([v]_\gamma, p \Big)_\gamma = 0 \;\; \forall p \in \mathbb{P}_{n-1}(\gamma) \;\; \text{for } \gamma \subset \partial\mathcal{P}\backslash\Gamma \Big\}$$

with the subspace X^n_D defined by

$$X^n_D = \Big\{ v \in X^n : (v,p)_\gamma = 0 \;\; \forall p \in \mathbb{P}_{n-1}(\gamma) \;\; \text{for } \gamma \subset \Gamma_D \Big\}, \tag{2.2}$$

where, for $v \in X^n$, we define the jump of v on an edge γ by $[v]_\gamma$. The spaces defined above are exactly the same spaces as those defined in [13] for $n = 1$, [16] for $n = 2$, [12] for $n = 3$, [11] for $n = 4, 6$, and more generally [9, 19] for all integer $n \geq 1$. The equivalence of the above definition with the conventional one follows easily by noting that for all $\gamma \subset \partial\mathcal{P}\backslash\Gamma$, if $v \in X^n_D$, then $[v]_\gamma \in \mathbb{P}_n(\gamma)$ is orthogonal to $\mathbb{P}_{n-1}(\gamma)$, and hence v is continuous at the mapped Gauss–Legendre points on γ. Also, for all $\gamma \subset \partial\mathcal{P} \cap \Gamma_D$, if $v \in X^n_D$, then $v_{|\gamma} \in \mathbb{P}_n(\gamma)$ is orthogonal to $\mathbb{P}_{n-1}(\gamma)$, meaning that v is zero at the mapped Gauss–Legendre points on $\gamma \subset \Gamma_D$.

Let $\mathcal{N}$ index a set of points $\{\boldsymbol{x}_m\}_{m\in\mathcal{N}}$ on $\overline{\Omega}$ associated with a Lagrange basis for the conforming finite element space of order n, and let $\mathcal{N}(\omega)$ denote the subset of $\mathcal{N}$ corresponding to points located on $\overline{\omega}$ for $\omega \subset \Omega$.

The nonconforming finite element approximation of order n of problem (2.1) consists of finding $u_{nc} \in X^n$ such that

$$(a\,\mathbf{grad_{nc}}\,u_{nc}, \mathbf{grad_{nc}}\,v) = (f,v) + (g,v)_{\Gamma_N} \;\; \forall v \in X^n_D, \\ (u_{nc},p)_\gamma = (q_I,p)_\gamma \;\; \forall p \in \mathbb{P}_{n-1}(\gamma) \;\; \text{for } \gamma \subset \Gamma_D, \tag{2.3}$$

where $\mathbf{grad_{nc}}$ denotes the operator defined by $(\mathbf{grad_{nc}}\,v)_{|K} = \mathbf{grad}(v_{|K})$ for $K \in \mathcal{P}$. Here, the Dirichlet datum q has been approximated by the piecewise $\mathbb{P}_n$ interpolant q_I which satisfies $q_{I|\gamma} \in \mathbb{P}_n(\gamma)$ and $q_I(\boldsymbol{x}_m) = q(\boldsymbol{x}_m)$ for $m \in \mathcal{N}(\Gamma_D)$.

For convenience, we shall adopt the convention that $\sum_i^j$ indicates a null sum in the case when $i > j$. Likewise, if $i > j$, then $i, \ldots, j$ and $\mathbb{P}_{j-i}$ are taken to be void. For $K \in \mathcal{P}$ let $\{P_{i,K}\}_{i=1}^{\dim \mathbb{P}_{n-3}(K)}$ form a basis for $\mathbb{P}_{n-3}(K)$ such that the value taken by $P_{i,K}$ at all points on K is independent of the size of element K, and for $\gamma \subset \partial\mathcal{P}$ let $L_{i,\gamma}$ denote the Legendre polynomial of degree i mapped onto the edge γ. Let $\beta_K \in \mathbb{P}_3(K)$ denote the function which equals zero on all edges of element K and satisfies $(\beta_K, 1)_K = |K|$. Likewise, for $\gamma \subset \partial K$ let $\beta_\gamma \in \mathbb{P}_2(K)$ denote the function which equals zero on all edges of element K except edge γ and satisfies $(\beta_\gamma, 1)_\gamma = |\gamma|$. Also, if n is a positive even number, let $\tilde{\beta}_K \in \mathbb{P}_n(K)$ be the nonconforming bubble function defined in [19], which we recall is supported on element K and vanishes at the n mapped Gauss–Legendre points on each edge on the boundary of element K.

The next result gives a characterization of the degrees of freedom associated with X^n when n is odd.

LEMMA 2.1. *Suppose that n is a positive odd number. Then the set of degrees of freedom on X^n defined by*

$$v \to \int_K v P_{j,K} \text{ for } j = 1, \ldots, \dim \mathbb{P}_{n-3}\left(\mathbb{R}^2\right) \text{ for all } K \in \mathcal{P} \tag{2.4}$$

and

$$v \to \int_\gamma v L_{j,\gamma} \text{ for } j = 0, \ldots, n-1 \text{ for all } \gamma \in \partial\mathcal{P} \tag{2.5}$$

is unisolvent.

Proof. Let $\sharp\mathcal{P}$ and $\sharp\partial\mathcal{P}$ denote the cardinality of $\mathcal{P}$ and $\partial\mathcal{P}$, respectively. The dimension of the space X^n is $\sharp\mathcal{P}.\dim \mathbb{P}_{n-3}(\mathbb{R}^2) + \sharp\partial\mathcal{P}.\dim \mathbb{P}_{n-1}(\mathbb{R})$ which agrees with the number of degrees of freedom. It therefore suffices to show that if $\rho \in X^n$ satisfies

$$(p, \rho)_K = 0 \ \forall p \in \mathbb{P}_{n-3}(K) \text{ for all } K \in \mathcal{P} \tag{2.6}$$

and

$$(p, \rho)_\gamma = 0 \ \forall p \in \mathbb{P}_{n-1}(\gamma) \text{ for all } \gamma \in \partial\mathcal{P}, \tag{2.7}$$

then $\rho = 0$.

First observe that (2.7) implies that ρ vanishes at the mapped Gauss–Legendre points on all edges $\gamma \in \partial\mathcal{P}$. Since n is odd, it follows that ρ vanishes on all edges $\gamma \in \partial\mathcal{P}$. Consequently, $\rho_{|K}$ must be of the form $\beta_K v$, $v \in \mathbb{P}_{n-3}(K)$. Letting $p = v$ in (2.6) then implies that v must be zero. Repeating this argument for all $K \in \mathcal{P}$ we conclude that the only solution $\rho \in X^n$ to (2.6) and (2.7) is $\rho = 0$. ∎

For n odd we define the functions $\theta_{i,K} \in X^n$ and $\theta_{i,\gamma} \in X^n$ to be such that

$$\int_{K'} \theta_{i,K} P_{j,K'} = \delta_{KK'}\delta_{ij}, \quad \int_{K'} \theta_{i,\gamma} P_{j,K'} = 0,$$

for $j = 1, \ldots, \dim \mathbb{P}_{n-3}(\mathbb{R}^2)$ for all $K' \in \mathcal{P}$, and

$$\int_{\gamma'} \theta_{i,K} L_{j,\gamma'} = 0, \quad \int_{\gamma'} \theta_{i,\gamma} L_{j,\gamma'} = \delta_{\gamma\gamma'}\delta_{ij},$$

for $j = 0, \ldots, n-1$ for all $\gamma' \in \partial\mathcal{P}$. Then, for n odd,

$$X^n = \operatorname{span}\left\{ \{\theta_{i,K}\}_{i=1}^{\dim \mathbb{P}_{n-3}(K)} \text{ for all } K \in \mathcal{P}, \ \{\theta_{i,\gamma}\}_{i=0}^{n-1} \text{ for all } \gamma \in \partial\mathcal{P} \right\}.$$

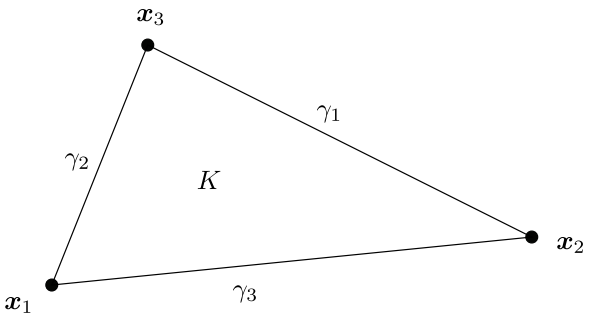


FIG. 2.1. *Enumeration of the vertices and edges of element K.*

The next result is the analogue of Lemma 2.1 in the case of even order elements.

LEMMA 2.2. *Suppose that n is a positive even number. Then the set of degrees of freedom on X^n defined by*

$$v \to \int_K v P_{j,K} \text{ for } j = 1, \ldots, \dim \mathbb{P}_{n-3}\left(\mathbb{R}^2\right) \text{ for all } K \in \mathcal{P}, \tag{2.8}$$

$$v \to \int_\gamma v L_{j,\gamma} \text{ for } j = 0, \ldots, n-1 \text{ for all } \gamma \in \partial\mathcal{P}, \tag{2.9}$$

and

$$v \to \int_{\partial K} v \tilde{\beta}_K \text{ for all } K \in \mathcal{P} \tag{2.10}$$

is unisolvent if and only if the datum v satisfies the compatibility condition

$$\sum_{\gamma \in \partial K} \int_\gamma v \sum_{\substack{1 \le j \le n-1: \\ j \text{ odd}}} \frac{L_{j,\gamma}}{\|L_{j,\gamma}\|^2_{L_2(\gamma)}} = 0 \text{ for all } K \in \mathcal{P}. \tag{2.11}$$

Proof. The dimension of the space X^n is $\sharp\mathcal{P}.\dim \mathbb{P}_{n-3}(\mathbb{R}^2) + \sharp\partial\mathcal{P}.\dim \mathbb{P}_{n-1}(\mathbb{R})$ which is $\sharp\mathcal{P}$ less than the number of degrees of freedom. We shall first show that if $v \in X^n$, then the compatibility condition holds. Now, making use of the fact that if $v \in X^n$, then $v_{|K} \in \mathbb{P}_n(K)$, we can write

$$v_{|\gamma} = \sum_{0 \le j \le n} c_{j,\gamma} L_{j,\gamma} \text{ on } K$$

for some $c_{j,\gamma} \in \mathbb{R}$. We then have

$$\sum_{\gamma \in \partial K} \int_\gamma v \sum_{\substack{1 \le j \le n-1: \\ j \text{ odd}}} \frac{L_{j,\gamma}}{\|L_{j,\gamma}\|^2_{L_2(\gamma)}} = \sum_{\substack{1 \le j \le n-1: \\ j \text{ odd}}} \left(c_{j,\gamma_1} + c_{j,\gamma_2} + c_{j,\gamma_3}\right), \tag{2.12}$$

where the edges and vertices of element K are oriented as shown in Figure 2.1.

Now, since $v_{|K}$ must be continuous at the vertices of element K, we have

$$\left(v_{|\gamma_1}(\boldsymbol{x}_3) - v_{|\gamma_1}(\boldsymbol{x}_2)\right) + \left(v_{|\gamma_2}(\boldsymbol{x}_1) - v_{|\gamma_2}(\boldsymbol{x}_3)\right) + \left(v_{|\gamma_3}(\boldsymbol{x}_2) - v_{|\gamma_3}(\boldsymbol{x}_1)\right) = 0,$$

which implies that

$$\sum_{\substack{1\le j\le n-1:\\ j \text{ odd}}} c_{j,\gamma_1} + \sum_{\substack{1\le j\le n-1:\\ j \text{ odd}}} c_{j,\gamma_2} + \sum_{\substack{1\le j\le n-1:\\ j \text{ odd}}} c_{j,\gamma_3} = 0,$$

which we can insert into (2.12) to give the compatibility condition for element K. Repeating this argument for all elements in $\mathcal{P}$ gives us the $\sharp\mathcal{P}$ linearly independent conditions which make up (2.11).

Now if the compatibility condition holds, then $\sharp\mathcal{P}$ of these degrees of freedom are linearly dependent, and so the dimension of the space X^n is the same as the number of degrees of freedom. We can therefore deduce unisolvency, if the compatibility condition holds, by showing that if $\rho \in X^n$ satisfies

$$(p,\rho)_K = 0 \ \forall p \in \mathbb{P}_{n-3}(K) \text{ for all } K \in \mathcal{P}, \tag{2.13}$$

$$(p,\rho)_\gamma = 0 \ \forall p \in \mathbb{P}_{n-1}(\gamma) \text{ for all } \gamma \in \partial\mathcal{P}, \tag{2.14}$$

and

$$\int_{\partial K} \rho\tilde{\beta}_K = 0 \text{ for all } K \in \mathcal{P}, \tag{2.15}$$

then $\rho = 0$.

As with the odd case we first observe that (2.14) implies that ρ vanishes at the mapped Gauss–Legendre points on all edges $\gamma \in \partial\mathcal{P}$. Since n is even, $\rho_{|K}$ must be of the form $\beta_K v + c\tilde{\beta}_K$, $v \in \mathbb{P}_{n-3}(K)$ for $K \in \mathcal{P}$ and some $c \in \mathbb{R}$. Now, (2.15) then implies that $c = 0$ which allows us to conclude that v must be zero by letting $p = v$ in (2.13). Hence, the only solution $\rho \in X^n$ to (2.13), (2.14), and (2.15) is $\rho = 0$. ∎

The following result gives an alternative formulation of the degrees of freedom in the case of even order elements that will be found more suitable for present purposes.

COROLLARY 2.3. *Suppose that n is a positive even number, and let*

$$X^n_* = \left\{ v \in X^n : \int_\gamma v \sum_{\substack{1\le j\le n-1:\\ j \text{ odd}}} \frac{L_{j,\gamma}}{\|L_{j,\gamma}\|^2_{L_2(\gamma)}} = 0 \text{ for all } \gamma \in \partial\mathcal{P} \right\}. \tag{2.16}$$

Then the set of degrees of freedom defined by (2.8), (2.10), *and*

$$v \to \int_\gamma v L_{j,\gamma} \text{ for } j = 0,\dots,n-2 \text{ for all } \gamma \in \partial\mathcal{P} \tag{2.17}$$

is unisolvent on X^n_*.

Proof. Observe that the degrees of freedom for X^n_* are related to the degrees of freedom for X^n described in Lemma 2.2 with the difference that the edge moment with respect to $L_{n-1,\gamma}$ for all $\gamma \subset \partial\mathcal{P}$ is missing from (2.17). The definition of the space X^n_* (2.16) imposes this missing degree of freedom on every edge $\gamma \subset \partial\mathcal{P}$ while respecting the compatibility condition (2.11). Hence, since X^n_* is a subspace of X^n we can apply Lemma 2.2 to deduce that the degrees of freedom described in Corollary 2.3 are unisolvent on X^n_*. ∎

For n even we define the functions $\theta_{i,K} \in X^n_*$, $\theta_{i,\gamma} \in X^n_*$, and $\theta_{\tilde{\beta},K} \in X^n_*$ to be such that

$$\int_{K'} \theta_{i,K} P_{j,K'} = \delta_{KK'}\delta_{ij}, \quad \int_{K'} \theta_{i,\gamma} P_{j,K'} = 0, \quad \int_{K'} \theta_{\tilde{\beta},K} P_{j,K'} = 0,$$

for $j = 1, \ldots, \dim \mathbb{P}_{n-3}(\mathbb{R}^2)$ for all $K' \in \mathcal{P}$,

$$\int_{\gamma'} \theta_{i,K} L_{j,\gamma'} = 0, \quad \int_{\gamma'} \theta_{i,\gamma} L_{j,\gamma'} = \delta_{\gamma\gamma'}\delta_{ij}, \quad \int_{\gamma'} \theta_{\tilde{\beta},K} L_{j,\gamma'} = 0,$$

for $j = 0, \ldots, n-1$ for all $\gamma' \in \partial\mathcal{P}$, and

$$\int_{\partial K'} \theta_{i,K} \tilde{\beta}_{K'} = 0, \quad \int_{\partial K'} \theta_{i,\gamma} \tilde{\beta}_{K'} = 0, \quad \int_{\partial K'} \theta_{\tilde{\beta},K} \tilde{\beta}_{K'} = \delta_{KK'},$$

for all $K' \in \mathcal{P}$. Then, for n even,

$$X_*^n = \operatorname{span}\left\{ \{\theta_{i,K}\}_{i=1}^{\dim \mathbb{P}_{n-3}(K)},\ \theta_{\tilde{\beta},K} \text{ for all } K \in \mathcal{P},\ \{\theta_{i,\gamma}\}_{i=0}^{n-2} \text{ for all } \gamma \in \partial\mathcal{P} \right\}.$$

We also define the subspace of X_*^n

$$X_{*D}^n = \left\{ v \in X_D^n : \int_\gamma v \sum_{\substack{1 \le j \le n-1: \\ j \text{ odd}}} \frac{L_{j,\gamma}}{\|L_{j,\gamma}\|_{L_2(\gamma)}^2} = 0 \text{ for all } \gamma \in \partial\mathcal{P} \right\}.$$

From the above definitions it follows that in both the odd and even order cases the functions $\theta_{i,K}$ and $\theta_{\tilde{\beta},K}$ are supported on element K with the functions $\theta_{i,\gamma}$ being supported on the two elements having γ as one of their edges.

The essential difference in the structures of the degrees of freedom in the case of odd versus even order elements is well known for nonconforming finite elements. This phenomenon will manifest itself throughout the subsequent analysis.

For further details of the implementation of these nonconforming finite element methods we refer the reader to [19], where the bases defined in [11, 12, 13, 16] for the full spaces X^n are generalized to arbitrary order.

2.3. Projection operators. As seen previously the structure of nonconforming finite elements is different depending on the parity of n. This fact means that we are obliged to consider the cases where n is odd or even separately when constructing the projection operators. We wish to construct operators that project functions from the space $H_D^1(\Omega)$ into the spaces X_D^n for odd $n \ge 1$ and X_{*D}^n for even $n \ge 2$.

LEMMA 2.4. *Suppose that n is a positive odd number. There exists a unique operator $\Pi : H^1(\Omega) \to X^n$ such that, for all $K \in \mathcal{P}$,*

$$(v - \Pi v, p)_K = 0 \ \forall p \in \mathbb{P}_{n-3}(K), \tag{2.18}$$

and, for all $\gamma \subset \partial\mathcal{P}$,

$$(v - \Pi v, p)_\gamma = 0 \ \forall p \in \mathbb{P}_{n-1}(\gamma). \tag{2.19}$$

Moreover, $\Pi : H_D^1(\Omega) \to X_D^n$ and, for all $K \in \mathcal{P}$,

$$\Pi p_{|K} = p_{|K} \text{ for all } p \in H^1(\Omega) \text{ such that } p_{|K} \in \mathbb{P}_n(K). \tag{2.20}$$

Also, there exists a positive constant C independent of the size of any of the elements in the mesh such that for all $K \in \mathcal{P}$

$$\|\Pi v\|_{L_2(K)} \le C \left(\|v\|_{L_2(K)} + h_K^{1/2} \sum_{\gamma \subset \partial K} \|v\|_{L_2(\gamma)} \right), \tag{2.21}$$

$$\sum_{\gamma\subset\partial K}\|\Pi v\|_{L_2(\gamma)}\leq C\sum_{\gamma\subset\partial K}\|v\|_{L_2(\gamma)}\,, \tag{2.22}$$

and

$$\|v-\Pi v\|_{L_2(K)}+h_K^{1/2}\sum_{\gamma\subset\partial K}\|v-\Pi v\|_{L_2(\gamma)}\leq Ch_K\,\|\mathbf{grad}\,v\|_{L_2(K)}\,, \tag{2.23}$$

where h_K is the length of the longest edge of element K.

Proof. The existence of such an operator Π follows from Lemma 2.1. Now, (2.19) implies that, for $v\in H_D^1(\Omega)$ and all $\gamma\subset\partial\mathcal{P}\cap\Gamma_D$, $(\Pi v,p)_\gamma=0$ for all $p\in\mathbb{P}_{n-1}(\gamma)$ and so $\Pi: H_D^1(\Omega)\to X_D^n$. Also, for all $K\in\mathcal{P}$, (2.18) for K and (2.19) for $\gamma\subset\partial K$ uniquely determine $\Pi v_{|K}$. This means that since, for all $p\in H^1(\Omega)$ such that $p_{|K}\in\mathbb{P}_n(K)$, letting $v=p$ and $\Pi p_{|K}=p_{|K}$ satisfies (2.18) for K and (2.19) for $\gamma\subset\partial K$, then (2.20) must hold. Finally, we can write

$$\Pi v_{|K}=\sum_{i=1}^{\dim\mathbb{P}_{n-3}(K)}\theta_{i,K|K}\left(P_{i,K},v\right)_K+\sum_{\gamma\subset\partial K}\sum_{i=0}^{n-1}\theta_{i,\gamma|K}\left(L_{i,\gamma},v\right)_\gamma \tag{2.24}$$

to which we can apply standard scaling arguments to obtain inequalities (2.21), (2.22), and (2.23) upon also noting that the $\theta_{i,K}$ vanish on and outside the boundary of K. ∎

Note that when $n=1$ the projection operator given in Lemma 2.4 reduces to the same projection operator that was defined by (2.7) in [5].

We turn to the case of even order elements.

LEMMA 2.5. *Suppose that n is a positive even number. There exists a unique operator $\Pi: H^1(\Omega)\to X_*^n$ such that, for all $K\in\mathcal{P}$,*

$$\left(v-\Pi v,p\right)_K=0\ \forall p\in\mathbb{P}_{n-3}\left(K\right) \tag{2.25}$$

and

$$\sum_{\gamma'\subset\partial K}\left(v-\Pi v,\tilde{\beta}_K\right)_{\gamma'}=0 \tag{2.26}$$

and, for all $\gamma\subset\partial\mathcal{P}$,

$$\left(v-\Pi v,p\right)_\gamma=0\ \forall p\in\mathbb{P}_{n-2}\left(\gamma\right) \tag{2.27}$$

and

$$\left(\Pi v,\sum_{\substack{1\leq j\leq n-1;\\ j\ odd}}\frac{L_{j,\gamma}}{\|L_{j,\gamma}\|_{L_2(\gamma)}^2}\right)_\gamma=0. \tag{2.28}$$

Moreover, $\Pi: H_D^1\left(\Omega\right)\to X_{*D}^n$ and, for all $K\in\mathcal{P}$,

$$\Pi p_{|K}=p_{|K}\ \textit{for all}\ p\in H^1\left(\Omega\right)\ \textit{such that}\ p_{|K}\in\mathbb{P}_0\left(K\right). \tag{2.29}$$

Also, inequalities (2.21), (2.22), *and* (2.23) *are satisfied.*

Proof. The existence of such an operator Π follows from Corollary 2.3. The same arguments that were used in the proof of Lemma 2.4 can be used to show that the remainder of this lemma is true by noting that we can write

$$\Pi v_{|K} = \sum_{i=1}^{\dim \mathbb{P}_{n-3}(K)} \theta_{i,K|K} \left(P_{i,K}, v\right)_K + \sum_{\gamma \subset \partial K} \left(\theta_{\tilde{\beta},K|K} \left(\tilde{\beta}_K, v\right)_\gamma + \sum_{i=0}^{n-2} \theta_{i,\gamma|K} \left(L_{i,\gamma}, v\right)_\gamma \right). \tag{2.30}$$

∎

While the projection operators that we have defined in Lemma 2.4 are natural extensions of the projection operator defined in [5], it is worth noting that when $n = 2$ the projection operator defined in Lemma 2.5 does not reduce to the projection operator defined in [3].

2.4. Data oscillation. Let $|K|$ denote the area of the element K, and let $|\gamma|$ denote the length of edge γ.

The oscillation of the data f on an element $K \in \mathcal{P}$ is defined to be

$$\operatorname{osc}(f, K) = |K|^{1/2} \inf_{p \in \mathbb{P}_n(K)} \|f - p\|_{L_2(K)}.$$

Since $\mathbb{P}_n(K)$ is finite-dimensional and the norm is strictly convex, it follows that there exists a unique minimizer of the right-hand side, which we shall denote by $P_K f$. Likewise, the oscillation of the Neumann data g on an edge $\gamma \subset \Gamma_N$ is defined to be

$$\operatorname{osc}(g, \gamma) = |\gamma|^{1/2} \inf_{p \in \mathbb{P}_n(\gamma)} \|g - p\|_{L_2(\gamma)},$$

and again we denote the minimizer by $P_\gamma g$. For $K \in \mathcal{P}$ and $\gamma \subset \partial K \cap \Gamma_D$, let the space $H^1_\gamma(K) = \{v : v \in H^1(K) : v = 0 \text{ on } \partial K \setminus \gamma\}$. The oscillation of the Dirichlet data q on an edge $\gamma \subset \partial K \cap \Gamma_D$ is defined to be

$$\operatorname{osc}(q, \gamma) = \inf_{\substack{v \in H^1_\gamma(K): \\ v_{|\gamma} = q - q_I}} \|\mathbf{grad}\, v\|_{L_2(K)}.$$

This definition of oscillation differs from that used in [5], [3]. However, it will transpire that the definition given here is more natural for the general case considered here.

LEMMA 2.6. *Let* $C_K = \frac{h_K}{\pi}$ *and let* $C^K_\gamma = (\frac{|\gamma|}{|K|} \frac{h_K}{\pi} (\frac{h_K}{\pi} + \max_{\gamma' \subset \partial K} |\gamma'|))^{1/2}$ *for* γ, γ' *distinct edges. Then, for* $v \in H^1_D(\Omega)$,

$$(f - P_K f, v - \Pi v)_K \le C_K \|f - P_K f\|_{L_2(K)} \|\mathbf{grad}\, v\|_{L_2(K)} \tag{2.31}$$

for $K \in \mathcal{P}$ *and*

$$(g - P_\gamma g, v - \Pi v)_\gamma \le C^K_\gamma \|g - P_\gamma g\|_{L_2(\gamma)} \|\mathbf{grad}\, v\|_{L_2(K)} \tag{2.32}$$

for $\gamma \subset \partial K \cap \Gamma_N$.

Proof. Since $\Pi v_{|K} \in \mathbb{P}_n(K)$, from the definition of $P_K f$ it follows that

$$(f - P_K f, v - \Pi v)_K = (f - P_K f, v)_K = (f - P_K f, v - c)_K$$

for any $c \in \mathbb{R}$. Now, by [18] there exists a $c \in \mathbb{R}$ such that $\|v - c\|_{L_2(K)} \le C_K \|\mathbf{grad}\, v\|_{L_2(K)}$ from which we can obtain (2.31) after applying the Cauchy–Schwarz inequality to the above. By a similar argument we can also obtain inequality (2.32) by making use of the inequality $\inf_{c \in \mathbb{R}} \|v - c\|_{L_2(\gamma)} \le C^K_\gamma \|\mathbf{grad}\, v\|_{L_2(K)}$, which is proved in [7]. ∎

2.5. Path permeability. Let a_K denote the value taken by the permeability a on element K. For $m \in \mathcal{N}$ let Ω_m denote the patch composed from those elements whose closure contains the point $\boldsymbol{x}_m$, and let $K, K' \subset \Omega_m$ be distinct elements. To define the permeability $a_{KK'}$ between a pair of elements, first observe that there is always at least one connected path $\wp(K, K') \subset \mathcal{P}$ passing from K to K' through adjacent elements belonging to the patch Ω_m. The smallest permeability of all of the elements in the path $\wp(K, K')$ is given by $\min\{a_M : M \in \wp(K, K')\}$. We take $\wp^\star(K, K')$ to be the path which maximizes the value of this quantity and define

$$a_{KK'} = \min\{a_M : M \in \wp^\star(K, K')\}. \tag{2.33}$$

If $m \in \mathcal{N}(\Gamma_D)$, then the element may be linked to $\gamma \subset \Gamma_D$ such that $\boldsymbol{x}_m \subset \overline{\gamma}$ by a connected path $\wp(K, \Gamma_D, m)$ passing through adjacent elements as before. Choosing $\wp^\star(K, \Gamma_D, m)$ to be the path which maximizes the value of $\min\{a_M : M \in \wp(K, \Gamma_D, m)\}$, the permeability between element K and the Dirichlet boundary is then defined by $a_{K\Gamma_D,m} = \min\{a_M : M \in \wp^\star(K, \Gamma_D, m)\}$. We also define

$$\Upsilon_K = \max\left(\max_{\substack{m\in\mathcal{N}(K):\\ m\in\mathcal{N}(\Gamma_D)}} \frac{a_K}{a_{K\Gamma_D,m}}, \max_{\substack{m\in\mathcal{N}(K):\\ m\notin\mathcal{N}(\Gamma_D)}} \left(\sum_{K'\subset\Omega_m\setminus K} \frac{\min(a_K, a_{K'})}{a_{KK'}}\right)\right). \tag{2.34}$$

For more details and examples on path permeability we refer the reader to [5].

2.6. A posteriori error estimation. Let the broken energy norm over a region ω be denoted by $\|\!|\cdot|\!\|_\omega = (a\,\mathbf{grad_{nc}}\,\cdot, \mathbf{grad_{nc}}\,\cdot)_\omega^{1/2}$ where again we shall omit the subscript in the case where $\omega = \Omega$. Our objective is to obtain a computable estimator for the broken energy norm of the error $e = u - u_{nc}$. To achieve this we shall make use of the following lemma, the proof of which can be found in [14].

LEMMA 2.7. *Let $\mathcal{H} = \{w \in H^1(\Omega) : (w, 1)_\Omega = 0$ and $\partial w/\partial \boldsymbol{t} = 0$ on $\Gamma_N\}$, where $\boldsymbol{t}$ is a tangent vector to Γ_N. The error $e = u - u_{nc}$ may be decomposed into the form*

$$a\,\mathbf{grad_{nc}}\,e = a\,\mathbf{grad}\,\phi + \mathbf{curl}\,\psi, \tag{2.35}$$

where the conforming error $\phi \in H^1_D(\Omega)$ satisfies

$$(a\,\mathbf{grad}\,\phi, \mathbf{grad}\,v) = (a\,\mathbf{grad_{nc}}\,e, \mathbf{grad}\,v)\ \forall v \in H^1_D(\Omega) \tag{2.36}$$

and the nonconforming error $\psi \in \mathcal{H}$ satisfies

$$(a^{-1}\,\mathbf{curl}\,\psi, \mathbf{curl}\,w) = (\mathbf{grad_{nc}}\,e, \mathbf{curl}\,w)\ \forall w \in \mathcal{H}. \tag{2.37}$$

Moreover,

$$\|\!|e|\!\|^2 = \|\!|\phi|\!\|^2 + (a^{-1}\,\mathbf{curl}\,\psi, \mathbf{curl}\,\psi). \tag{2.38}$$

The importance of this lemma is that it allows us to write $\|\!|e|\!\|^2$ as the sum of a conforming part $\|\!|\phi|\!\|^2$ and a nonconforming part $(a^{-1}\,\mathbf{curl}\,\psi, \mathbf{curl}\,\psi)$ which reduces the task of obtaining an estimator for $\|\!|e|\!\|$ to that of obtaining separate estimators for each of the two terms in this decomposition.

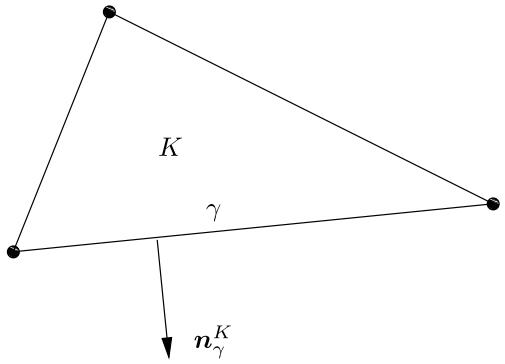


FIG. 3.1. *Orientation of the unit normal vector to an edge γ of element K.*

3. Estimation of the conforming error.

3.1. Upper bound. Let the unit normal vector $\boldsymbol{n}_\gamma^K$ to edge γ of element K be oriented as shown in Figure 3.1.

Our aim is to find a fully computable upper bound for $|||\phi|||^2$. Let $v \in H_D^1(\Omega)$. Then, by (2.36), the definition of e and (2.1),

$$\begin{aligned}(a\,\mathbf{grad}\,\phi, \mathbf{grad}\,v) &= (a\,\mathbf{grad_{nc}}\,e, \mathbf{grad}\,v)\\ &= (a\,\mathbf{grad}\,u, \mathbf{grad}\,v) - (a\,\mathbf{grad_{nc}}\,u_{nc}, \mathbf{grad}\,v)\\ &= (f,v) + (g,v)_{\Gamma_N} - (a\,\mathbf{grad_{nc}}\,u_{nc}, \mathbf{grad}\,v)\\ &= \sum_{K\in\mathcal{P}} \left((f,v)_K - (a_K\,\mathbf{grad}\,u_{nc}, \mathbf{grad}\,v)_K + \sum_{\gamma\subset\partial K} (g,v)_{\gamma\cap\Gamma_N} \right).\end{aligned}$$

Observing that $\Pi v \in X_D^n$, we find

$$0 = (a\,\mathbf{grad_{nc}}\,u_{nc}, \mathbf{grad_{nc}}\,\Pi v) - (f, \Pi v) - (g, \Pi v)_{\Gamma_N}$$

from (2.3), and so

$$\begin{aligned}(a\,\mathbf{grad}\,\phi, \mathbf{grad}\,v) = \sum_{K\in\mathcal{P}} \Bigg(&- (a_K\,\mathbf{grad}\,u_{nc}, \mathbf{grad}\,(v-\Pi v))_K\\ &+ (f, v-\Pi v)_K + \sum_{\gamma\subset\partial K} (g, v-\Pi v)_{\gamma\cap\Gamma_N} \Bigg).\end{aligned}$$

We can then use integration by parts to obtain

$$\begin{aligned}&(a\,\mathbf{grad}\,\phi, \mathbf{grad}\,v)\\ &= \sum_{K\in\mathcal{P}} \Bigg((P_K f + a_K\,\mathrm{div}\,(\mathbf{grad}\,u_{nc}), v-\Pi v)_K + (f - P_K f, v-\Pi v)_K\\ &\quad + \sum_{\gamma\subset\partial K} \Bigg((P_\gamma g, v-\Pi v)_{\gamma\cap\Gamma_N} + (g - P_\gamma g, v-\Pi v)_{\gamma\cap\Gamma_N}\\ &\quad - \left(a_K \frac{\partial u_{nc|K}}{\partial \boldsymbol{n}_\gamma^K}, v-\Pi v \right)_\gamma \Bigg) \Bigg).\end{aligned} \tag{3.1}$$

We shall now consider the two cases n odd and n even separately. For n odd, applying property (2.19) gives

$$
\begin{aligned}
(3.2)\qquad & (a\,\mathbf{grad}\,\phi, \mathbf{grad}\,v) \\
&= \sum_{K\in\mathcal{P}} \left(P_K f + a_K \operatorname{div}\left(\mathbf{grad}\,u_{nc}\right), v-\Pi v\right)_K + \sum_{\gamma\subset\partial\mathcal{P}\cap\Gamma_N} \left(P_\gamma g, v-\Pi v\right)_\gamma \\
&\quad + \sum_{K\in\mathcal{P}} \left(f-P_K f, v-\Pi v\right)_K + \sum_{\gamma\subset\partial\mathcal{P}\cap\Gamma_N} \left(g-P_\gamma g, v-\Pi v\right)_\gamma
\end{aligned}
$$

since $a_K \partial u_{nc|K}/\partial \boldsymbol{n}_\gamma^K \in \mathbb{P}_{n-1}(\gamma)$. We can then rewrite this using the definition of Πv (2.24) and the fact that the $\theta_{i,K}$ vanish on the boundary of element K as

$$
\begin{aligned}
(3.3)\qquad (a\,\mathbf{grad}\,\phi, \mathbf{grad}\,v) = \sum_{K\in\mathcal{P}} \Big(& \left(r_K, v\right)_K + \left(f-P_K f, v-\Pi v\right)_K \\
& + \sum_{\gamma\subset\partial K} \Big(\left(g-P_\gamma g, v-\Pi v\right)_{\gamma\cap\Gamma_N} + \left(R_{\gamma,K}, v\right)_\gamma \Big)\Big),
\end{aligned}
$$

where

$$
r_K = P_K f + a_K \operatorname{div}\left(\mathbf{grad}\,u_{nc}\right) - \sum_{i=1}^{\dim \mathbb{P}_{n-3}(K)} \left(\theta_{i,K}, P_K f + a_K \operatorname{div}\left(\mathbf{grad}\,u_{nc}\right)\right)_K P_{i,K}
$$

and, defining $P_\gamma g$ and $P_{\gamma'} g$ to vanish if $\gamma \cap \Gamma_N$ or $\gamma' \cap \Gamma_N$, respectively, are empty

$$
R_{\gamma,K} = P_\gamma g - \sum_{i=0}^{n-1} \left(\sum_{\gamma'\subset\partial K\cap\Gamma_N} \left(\theta_{i,\gamma}, P_{\gamma'} g\right)_{\gamma'} + \left(\theta_{i,\gamma}, P_K f + a_K \operatorname{div}\left(\mathbf{grad}\,u_{nc}\right)\right)_K \right) L_{i,\gamma}.
$$

We now turn to the case when n is even. Unlike the odd order case we cannot eliminate the term involving the normal derivative of u_{nc} from (3.2). This was possible in the case of odd order elements because of the property that $v - \Pi v$ was orthogonal to $\mathbb{P}_{n-1}(\gamma)$. The corresponding property fails to hold in the case of even order elements owing to the compatibility condition in Lemma 2.2. We shall therefore require the following definitions in order to deal with this term in a way that allows us to prove that the estimator is efficient and robust with respect to the permeability coefficient a. For each $K \in \mathcal{P}$ and each edge $\gamma \subset \partial K$, define

$$
\alpha_\gamma^K = \begin{cases} \dfrac{a_{K'}^{-1/2}}{a_K^{-1/2} + a_{K'}^{-1/2}} & \text{if } \gamma = \partial K \cap \partial K', \\ 1 & \text{if } \gamma \subset \Gamma_N, \\ 0 & \text{if } \gamma \subset \Gamma_D \end{cases}
$$

and

$$
\left[a\frac{\partial u_{nc}}{\partial \boldsymbol{n}}\right]_\gamma = \begin{cases} a_K \dfrac{\partial u_{nc|K}}{\partial \boldsymbol{n}_\gamma^K} + a_{K'} \dfrac{\partial u_{nc|K'}}{\partial \boldsymbol{n}_\gamma^{K'}} & \text{if } \gamma = \partial K \cap \partial K', \\ a_K \dfrac{\partial u_{nc|K}}{\partial \boldsymbol{n}_\gamma^K} & \text{if } \gamma \subset \Gamma. \end{cases}
$$

Now, for n even, since $\alpha_\gamma^K + \alpha_\gamma^{K'} = 1$ for $\gamma = \partial K \cap \partial K'$, we can make use of the properties of the space X_D^n to which Πv belongs and the fact that $v = 0$ on Γ_D to

obtain

$$
\begin{aligned}
&(a\,\mathbf{grad}\,\phi, \mathbf{grad}\,v)\\
&= \sum_{K\in\mathcal{P}} \Bigg((P_K f + a_K \operatorname{div}(\mathbf{grad}\,u_{nc}), v-\Pi v)_K + (f - P_K f, v-\Pi v)_K\\
&+ \sum_{\gamma\subset\partial K} \Bigg((P_\gamma g, v-\Pi v)_{\gamma\cap\Gamma_N} + (g-P_\gamma g, v-\Pi v)_{\gamma\cap\Gamma_N}\\
&- \left(\alpha_\gamma^K \left[a\frac{\partial u_{nc}}{\partial \boldsymbol{n}}\right]_\gamma, v-\Pi v\right)_\gamma\Bigg)\Bigg).
\end{aligned}
$$

In the same way that we did for the odd case we can rewrite this equation using the definition of Πv (2.30) as (3.3) where r_K is defined in the same way that it was for the odd case but

$$
\begin{aligned}
R_{\gamma,K} = P_\gamma g - \alpha_\gamma^K \left[a\frac{\partial u_{nc}}{\partial \boldsymbol{n}}\right]_\gamma + \sum_{i=0}^{n-2} \Bigg(&\left(\theta_{i,\gamma}, \alpha_\gamma^K \left[a\frac{\partial u_{nc}}{\partial \boldsymbol{n}}\right]_\gamma\right)_\gamma\\
&- \sum_{\gamma'\subset\partial K\cap\Gamma_N} (\theta_{i,\gamma}, P_{\gamma'} g)_{\gamma'} - (\theta_{i,\gamma}, P_K f + a_K \operatorname{div}(\mathbf{grad}\,u_{nc}))_K \Bigg) L_{i,\gamma}\\
&- \left(\sum_{\gamma'\subset\partial K\cap\Gamma_N} \left(\theta_{\tilde\beta,K}, P_{\gamma'} g\right)_{\gamma'} + \left(\theta_{\tilde\beta,K}, P_K f + a_K \operatorname{div}(\mathbf{grad}\,u_{nc})\right)_K \right) \tilde\beta_{K|\gamma}
\end{aligned}
$$

when n is even. To arrive at this we have made use of the fact that, for all $p\in\mathbb{P}_{n-1}(\gamma)$, $(\theta_{\tilde\beta,K}, p)_\gamma$ and $(\theta_{i,\gamma'}, p)_\gamma$ vanish for all edges $\gamma\neq\gamma'$.

Note that evaluation of our estimator will require calculation of r_K and $R_{\gamma,K}$. This means that the evaluation of our estimator is simpler when n is odd since in that case we need not calculate $\alpha_\gamma^K \left[a\frac{\partial u_{nc}}{\partial \boldsymbol{n}}\right]_\gamma$. The functions r_K and $R_{\gamma,K}$ are used to calculate an auxiliary function $\boldsymbol{\sigma}_K$ as follows.

THEOREM 3.1. *Consider the problem of minimizing $\|\boldsymbol{\sigma}_K\|^2_{L_2(K)}$ subject to*

$$
(\boldsymbol{\sigma}_K, \mathbf{grad}\,v)_K = (r_K, v)_K + \sum_{\gamma\subset\partial K} (R_{\gamma,K}, v)_\gamma \quad \forall v\in\mathbb{P}_{n+3}(K). \tag{3.4}
$$

A unique solution $\boldsymbol{\sigma}_K\in\mathbb{P}_{n+1}(K)\times\mathbb{P}_{n+1}(K)$ to this problem exists. Moreover, the solution $\boldsymbol{\sigma}_K$ of this problem satisfies

$$
(\boldsymbol{\sigma}_K, \mathbf{grad}\,v)_K = (r_K, v)_K + \sum_{\gamma\subset\partial K} (R_{\gamma,K}, v)_\gamma \quad \forall v\in H^1(K). \tag{3.5}
$$

Also, there exists a positive constant C, independent of the size of any of the elements in the mesh, such that

$$
\|\boldsymbol{\sigma}_K\|_{L_2(K)} \le C\left(|K|^{1/2}\|r_K\|_{L_2(K)} + \sum_{\gamma\subset\partial K} |\gamma|^{1/2}\|R_{\gamma,K}\|_{L_2(\gamma)}\right). \tag{3.6}
$$

Proof. We shall first show that (3.4) is equivalent to

$$
\boldsymbol{n}_\gamma^K\cdot\boldsymbol{\sigma}_K = R_{\gamma,K}\in\mathbb{P}_n(\gamma) \text{ on } \gamma\subset\partial K \text{ for all } \gamma\subset\partial K \tag{3.7}
$$

and

$$-\operatorname{div}\boldsymbol{\sigma}_K = r_K \in \mathbb{P}_n(K) \text{ on } K. \tag{3.8}$$

Now, integration by parts shows that (3.4) is equivalent to

$$\sum_{\gamma\subset\partial K}\left(R_{\gamma,K}-\boldsymbol{n}_\gamma^K\cdot\boldsymbol{\sigma}_K,v\right)_\gamma+\left(r_K+\operatorname{div}\boldsymbol{\sigma}_K,v\right)_K=0\ \forall v\in\mathbb{P}_{n+3}(K), \tag{3.9}$$

which obviously holds when (3.7) and (3.8) are satisfied. Conversely, we first let $v=\beta_K(r_K+\operatorname{div}\boldsymbol{\sigma}_K)\in\mathbb{P}_{n+3}(K)$ in (3.9), where $\beta_K\geq 0$ is the cubic bubble, to see that (3.8) must hold. Furthermore, extending $R_{\gamma,K}$ onto K as a polynomial in $\mathbb{P}_n(K)$ and letting $v=\beta_\gamma(R_{\gamma,K}-\boldsymbol{n}_\gamma^K\cdot\boldsymbol{\sigma}_K)\in\mathbb{P}_{n+3}(K)$ in (3.9) for each $\gamma\subset\partial K$, where $\beta_\gamma\geq 0$ is the quadratic edge bubble, shows that (3.7) is satisfied.

In Theorem 6.1 in the appendix, the existence of a function in $\mathbb{P}_{n+1}(\hat{K})\times\mathbb{P}_{n+1}(\hat{K})$ satisfying (3.7) and (3.8) for a reference element $\hat{K}$ is proven provided that the data satisfy the usual compatibility condition

$$(r_K,1)_K+\sum_{\gamma\subset\partial K}(R_{\gamma,K},1)_\gamma=0. \tag{3.10}$$

To see that this condition does hold we observe that, for all $v\in H^1(K)$,

$$\begin{aligned}&(r_K,v)_K+\sum_{\gamma\subset\partial K}(R_{\gamma,K},v)_\gamma\\&=\left(P_Kf+a_K\operatorname{div}(\mathbf{grad}\,u_{nc}),v-\Pi v\right)_K+\sum_{\gamma\subset\partial K\cap\Gamma_N}(P_\gamma g,v-\Pi v)_\gamma\end{aligned}$$

if n is odd, while for n even

$$\begin{aligned}&(r_K,v)_K+\sum_{\gamma\subset\partial K}(R_{\gamma,K},v)_\gamma\\&=\left(P_Kf+a_K\operatorname{div}(\mathbf{grad}\,u_{nc}),v-\Pi v\right)_K+\sum_{\gamma\subset\partial K}\left(P_\gamma g-\alpha_\gamma^K\left[a\frac{\partial u_{nc}}{\partial\boldsymbol{n}}\right]_\gamma,v-\Pi v\right)_\gamma\end{aligned}$$

from which condition (3.10) then follows since Π preserves constants. Moreover, the strict convexity of the norm implies that $\boldsymbol{\sigma}_K$ exists and is unique. Assertion (3.5) is then a trivial consequence of (3.7) and (3.8). Finally, we can apply standard scaling arguments to inequality (6.5) in the appendix to arrive at inequality (3.6) on recalling $\boldsymbol{\sigma}_K$ has minimal norm. □

Theorem 3.1 shows that, despite the fact that condition (3.4) is enforced with respect to a *finite*-dimensional space of test functions, the condition actually holds over an *infinite*-dimensional space.

We are now in a position to obtain a computable upper bound for $|\!|\!|\phi|\!|\!|$.

LEMMA 3.2. *Define*

(3.11)
$$\Phi_K = a_K^{-1/2}\left(\|\boldsymbol{\sigma}_K\|_{L_2(K)} + C_K\,\|f - P_K f\|_{L_2(K)} + \sum_{\gamma\subset\partial K\cap\Gamma_N} C_\gamma^K\,\|g - P_\gamma g\|_{L_2(\gamma)}\right),$$

where C_K and C_γ^K are the fully computable constants defined in Lemma 2.6. *Then*

$$\|\!|\phi|\!\|^2 \le \sum_{K\in\mathcal{P}} \Phi_K^2. \tag{3.12}$$

Proof. Returning to (3.3) and using Theorem 3.1 we obtain

$$(a\,\mathbf{grad}\,\phi, \mathbf{grad}\,v) = \sum_{K\in\mathcal{P}}\Big((\boldsymbol{\sigma}_K, \mathbf{grad}\,v)_K + (f - P_K f, v - \Pi v)_K + \sum_{\gamma\subset\partial K\cap\Gamma_N} (g - P_\gamma g, v - \Pi v)_\gamma\Big).$$

Inserting (2.31) and (2.32) and then applying the Cauchy–Schwarz inequality yields

$$(a\,\mathbf{grad}\,\phi, \mathbf{grad}\,v) \le \sum_{K\in\mathcal{P}} \Phi_K\,\|\!|v|\!\|_K \le \left(\sum_{K\in\mathcal{P}}\Phi_K^2\right)^{1/2} \|\!|v|\!\|$$

from which we can obtain the result by letting $v = \phi$ and dividing both sides of the inequality by $\|\!|\phi|\!\|$. ∎

3.2. Local lower bounds. We now show that Φ_K also provides a local lower bound for the conforming component of the error plus data oscillation up to a generic constant.

LEMMA 3.3. *Let K^* be the union of those elements, including element K, that share an edge with element K. Then there exists a positive constant c, independent of the size of any of the elements in the mesh and the permeability coefficient a, such that*

(3.13)
$$c\Phi_K \le \begin{cases} \|\!|\phi|\!\|_K + a_K^{-1/2}\,\mathrm{osc}\,(f,K) + a_K^{-1/2}\displaystyle\sum_{\gamma\subset\partial K\cap\Gamma_N} \mathrm{osc}\,(g,\gamma) & \text{if } n \text{ is odd,} \\ \displaystyle\sum_{K'\subset K^*}\Big(\|\!|\phi|\!\|_{K'} + a_K^{-1/2}\,\mathrm{osc}\,(f,K')\Big) + a_K^{-1/2}\sum_{\gamma\subset\partial K\cap\Gamma_N} \mathrm{osc}\,(g,\gamma) & \text{if } n \text{ is even.}\end{cases}$$

Proof. Applying the estimates $C_K \le C|K|^{1/2}$ and $C_\gamma^K \le C|\gamma|^{1/2}$ to the expression for Φ_K allows us to say that

$$\Phi_K \le C a_K^{-1/2}\left(\|\boldsymbol{\sigma}_K\|_{L_2(K)} + \mathrm{osc}\,(f,K) + \sum_{\gamma\subset\partial K\cap\Gamma_N}\mathrm{osc}\,(g,\gamma)\right),$$

and by (3.6) and the equivalence of norms in finite dimensions we can say that

$$\|\boldsymbol{\sigma}_K\|_{L_2(K)} \le C\left(|K|^{1/2}\left\|\beta_K^{1/2} r_K\right\|_{L_2(K)} + \sum_{\gamma\subset\partial K}|\gamma|^{1/2}\left\|\beta_\gamma^{1/2} R_{\gamma,K}\right\|_{L_2(\gamma)}\right).$$

Let $J_{i,\gamma}$ denote the mapped Jacobi polynomial of order i with weight β_γ on edge γ. Then, when n is odd, by property (2.19) of Πv we can write

$$(P_\gamma g, v - \Pi v)_\gamma = \frac{(P_\gamma g, \beta_\gamma J_{n,\gamma})_\gamma}{(J_{n,\gamma}, \beta_\gamma J_{n,\gamma})_\gamma} (J_{n,\gamma}, v - \Pi v)_\gamma .$$

Hence, for all $v \in H^1(K)$, for n odd

$$\begin{aligned} &(P_K f + a_K \operatorname{div}(\mathbf{grad}\, u_{nc}), v - \Pi v)_K \\ &+ \sum_{\gamma \subset \partial K \cap \Gamma_N} \frac{(P_\gamma g, \beta_\gamma J_{n,\gamma})_\gamma}{(J_{n,\gamma}, \beta_\gamma J_{n,\gamma})_\gamma} (J_{n,\gamma}, v - \Pi v)_\gamma \\ &= (r_K, v)_K + \sum_{\gamma \subset \partial K} (R_{\gamma,K}, v)_\gamma , \end{aligned}$$

and for n even

$$\begin{aligned} &(P_K f + a_K \operatorname{div}(\mathbf{grad}\, u_{nc}), v - \Pi v)_K + \sum_{\gamma \subset \partial K} \alpha_\gamma^K \left(P_\gamma g - \left[a \frac{\partial u_{nc}}{\partial \boldsymbol{n}} \right]_\gamma , v - \Pi v \right)_\gamma \\ &= (r_K, v)_K + \sum_{\gamma \subset \partial K} (R_{\gamma,K}, v)_\gamma . \end{aligned}$$

Applying techniques used in [2, 20] to these equations allows us to say, with the aid of inequalities (2.21) and (2.22), that

$$|K|^{1/2} \left\| \beta_K^{1/2} r_K \right\|_{L_2(K)} \le C |K|^{1/2} \left\| \beta_K^{1/2} (P_K f + a_K \operatorname{div}(\mathbf{grad}\, u_{nc})) \right\|_{L_2(K)}$$

with, for n odd and $\gamma \subset \partial \mathcal{P} \cap \partial K$,

$$\begin{aligned} &|\gamma|^{1/2} \left\| \beta_\gamma^{1/2} R_{\gamma,K} \right\|_{L_2(\gamma)} \\ &\le C \Big(|K|^{1/2} \left\| \beta_K^{1/2} (P_K f + a_K \operatorname{div}(\mathbf{grad}\, u_{nc})) \right\|_{L_2(K)} \\ &\quad + \sum_{\gamma' \subset \partial K \cap \Gamma_N} \left| (P_{\gamma'} g, \beta_{\gamma'} J_{n,\gamma'})_{\gamma'} \right| \Big) \end{aligned}$$

and, for n even and all $\gamma \subset \partial \mathcal{P} \cap \partial K$,

$$\begin{aligned} |\gamma|^{1/2} \left\| \beta_\gamma^{1/2} R_{\gamma,K} \right\|_{L_2(\gamma)} \le C \Bigg(&\sum_{\gamma' \subset \partial K} \alpha_{\gamma'}^K |\gamma'|^{1/2} \left\| \beta_{\gamma'}^{1/2} \left(P_{\gamma'} g - \left[a \frac{\partial u_{nc}}{\partial \boldsymbol{n}} \right]_{\gamma'} \right) \right\|_{L_2(\gamma')} \\ &+ |K|^{1/2} \left\| \beta_K^{1/2} (P_K f + a_K \operatorname{div}(\mathbf{grad}\, u_{nc})) \right\|_{L_2(K)} \Bigg). \end{aligned}$$

We can also apply these techniques to (3.1) to obtain, for all $K \in \mathcal{P}$,

$$|K|^{1/2} \left\| \beta_K^{1/2} (P_K f + a_K \operatorname{div}(\mathbf{grad}\, u_{nc})) \right\|_{L_2(K)} \le C \left(a_K^{1/2} |\!|\!| \phi |\!|\!|_K + \operatorname{osc}(f, K) \right)$$

as well as, for n odd and all $\gamma' \subset \partial \mathcal{P} \cap \partial K \cap \Gamma_N$,

$$\left| (P_{\gamma'} g, \beta_{\gamma'} J_{n,\gamma'})_{\gamma'} \right| \le C \left(a_K^{1/2} |\!|\!| \phi |\!|\!|_K + \operatorname{osc}(f, K) + \operatorname{osc}(g, \gamma') \right)$$

and, for n even and all $\gamma' \subset \partial\mathcal{P}$,

$$|\gamma'|^{1/2} \left\| \beta_{\gamma'}^{1/2} \left(P_{\gamma'} g - \left[a \frac{\partial u_{nc}}{\partial \boldsymbol{n}} \right]_{\gamma'} \right) \right\|_{L_2(\gamma')}$$

$$\leq C \left(\sum_{\substack{K' \in \mathcal{P}_i \\ \gamma' \subset \partial K'}} \left(a_{K'}^{1/2} \left\| \phi \right\|_{K'} + \operatorname{osc}(f, K') \right) + \operatorname{osc}(g, \gamma' \cap \Gamma_N) \right),$$

which we can then combine with the above to obtain the result claimed upon noting that

$$\alpha_{\gamma'}^{K} \sum_{\substack{K' \in \mathcal{P}_i \\ \gamma' \subset \partial K'}} a_{K'}^{1/2} = a_K^{1/2}. \qquad \square$$

4. Estimation of the nonconforming error.

4.1. Upper bound. The upper bound for $(a^{-1}\,\mathbf{curl}\,\psi, \mathbf{curl}\,\psi)$ is defined analogously to [3, 5] using the continuous piecewise $\mathbb{P}_n$ interpolant of u_{nc} on $\mathcal{P}$ whose values at the conforming nodes are given by

$$\mathcal{S}(u_{nc})(\boldsymbol{x}_m) = \begin{cases} q(\boldsymbol{x}_m) & \text{if } m \in \mathcal{N}(\Gamma_D), \\ \displaystyle\sum_{K \in \Omega_m} \omega_{K,m} u_{nc|K}(\boldsymbol{x}_m) & \text{if } m \notin \mathcal{N}(\Gamma_D) \end{cases} \tag{4.1}$$

for $m \in \mathcal{N}$ with the weights $\omega_{K,m}$ defined by

$$\omega_{K,m} = \frac{a_K^{1/2}}{\sum_{K' \subset \Omega_m} a_{K'}^{1/2}}.$$

As in [3, 5], it is also essential here that a weighted average be used in order that robustness of the estimator with respect to the permeability coefficient is maintained. Note that these weights satisfy the condition

$$\sum_{K \subset \Omega_m} \omega_{K,m} = 1.$$

LEMMA 4.1. *Let*

$$\Psi_K = \left\| u_{nc} - \mathcal{S}(u_{nc}) \right\|_K + \sum_{\gamma \subset \partial K \cap \Gamma_D} a_K^{1/2} \operatorname{osc}(q, \gamma). \tag{4.2}$$

Then

$$\left(a^{-1}\,\mathbf{curl}\,\psi, \mathbf{curl}\,\psi\right) \leq \sum_{K \in \mathcal{P}} \Psi_K^2. \tag{4.3}$$

Proof. The above result can easily be obtained from the upper bound proven in [5]. $\square$

The $\mathcal{S}(u_{nc})$ defined in (4.1) says that, for $\boldsymbol{x}_m \notin \partial\mathcal{P}$, $\mathcal{S}(u_{nc})(\boldsymbol{x}_m) = u_{nc}(\boldsymbol{x}_m)$, but any choice of $\mathcal{S}(u_{nc})$ at these points will provide an upper bound for the nonconforming component of the broken energy norm of the error when used in the evaluation of $\sum_{K \in \mathcal{P}} \Psi_K^2$. We therefore propose that at these points on the interior of the triangles the values taken by $\mathcal{S}(u_{nc})$ are chosen so that they minimize Ψ_K. Note that the local lower bound proved in the next section will still hold if this is done. Also, the choice of the points $\boldsymbol{x}_m$ on the edge interiors is not unique. In our numerical examples uniformly spaced points were used.

4.2. Local lower bounds. It can also be shown that Ψ_K provides a local lower bound for the nonconforming component of the error plus data oscillation up to a generic constant.

LEMMA 4.2. *Let $\tilde{K}$ denote the patch formed from element K and those elements sharing a common vertex with element K. Then there exists a positive constant c, independent of the size of any of the elements in the mesh and the permeability coefficient a, such that*

$$c\Psi_K^2 \leq \Upsilon_K \left(a^{-1}\,\mathbf{curl}\,\psi, \mathbf{curl}\,\psi\right)_{\tilde{K}} + a_K \sum_{\gamma \subset \partial\tilde{K} \cap \Gamma_D} \operatorname{osc}^2(q,\gamma)\,. \tag{4.4}$$

Proof. This result can be proved using the techniques used to prove its analogue in [3, 5]. ∎

5. Summary and numerical examples.

5.1. Summary of main results. Our fully computable upper bound and local lower bound on the broken energy norm of the error e in the nonconforming finite element approximation are summarized in the following theorem.

THEOREM 5.1. *Let Φ_K and Ψ_K be defined as in Lemmas* 3.2 *and* 4.1*, respectively. Then the broken energy norm of the total error e can be estimated as*

$$|\!|\!|e|\!|\!|^2 \leq \sum_{K\in\mathcal{P}} \left(\Phi_K^2 + \Psi_K^2\right). \tag{5.1}$$

Also, there exists a positive constant c, which is independent of the size of any of the elements in the mesh and the permeability coefficient a, such that

$$\begin{aligned} c\left(\Phi_K^2 + \Psi_K^2\right) \leq \Upsilon_K\, |\!|\!|e|\!|\!|_{\tilde{K}}^2 &+ a_K \sum_{\gamma \subset \partial\tilde{K}\cap\Gamma_D} \operatorname{osc}^2(q,\gamma) \\ &+ a_K^{-1} \sum_{K'\subset\tilde{K}} \operatorname{osc}^2(f,K') + a_K^{-1} \sum_{\gamma\subset\partial K\cap\Gamma_N} \operatorname{osc}^2(g,\gamma)\,, \end{aligned} \tag{5.2}$$

where $\tilde{K}$ denotes the patch formed from element K and those elements sharing a common vertex with element K and Υ_K is defined in (2.34).

Proof. The proof follows from Lemmas 2.7, 3.2, 3.3, 4.1, and 4.2 on observing that $K \subset K^* \subset \tilde{K}$. ∎

5.2. Example. For our example we look at the performance of the estimator for the problem of finding u^* such that $-\operatorname{div}(a\,\mathbf{grad}\,u^*) = f$ in the region $\Omega = (0,1)^2$ with homogeneous Dirichlet data on $\Gamma_D = \partial\Omega$. The function f has been chosen so that the exact solution to this problem is $u^* = (1-x^2)(1-y^2)u$, where u is the solution to the Kellogg problem described in [17], where a is also defined to take the value $161.4\ldots$ in the first and third quadrants and 1 in the second and fourth quadrants. The gradient of the true solution displays highly singular behavior at the origin since $\mathbf{grad}\,u^* = O(r^{-9/10})$. The initial mesh used for this example is shown in Figure 5.1.

Two refinement strategies were implemented.

Strategy 1 (results shown in Figures 5.2(a), 5.3(a), and 5.4(a)): The mesh was uniformly refined whereby each triangle is subdivided into four subtriangles at each step.

Strategy 2 (results shown in Figures 5.2(b), 5.3(b), and 5.4(b)): The mesh was adaptively refined whereby a bulk criterion [15] was used to refine the mesh on the

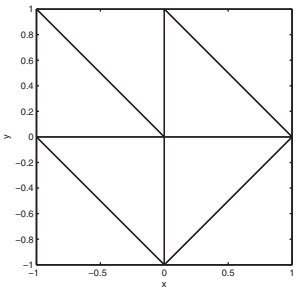
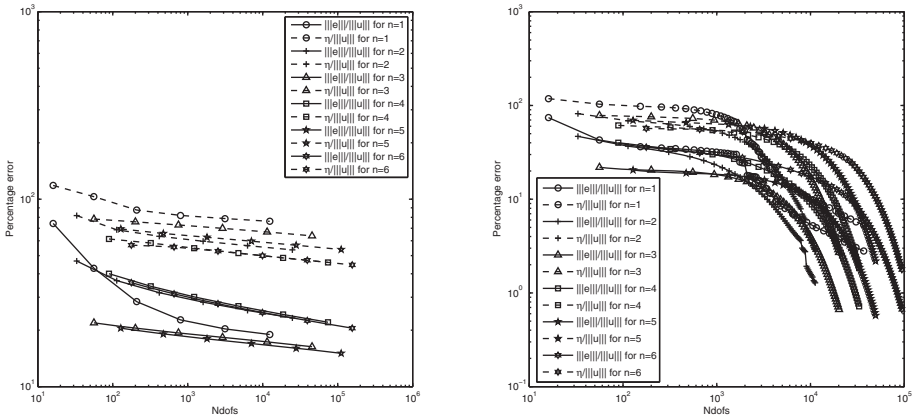


FIG. 5.1. *Initial mesh for Example* 5.2.



(a) (b)

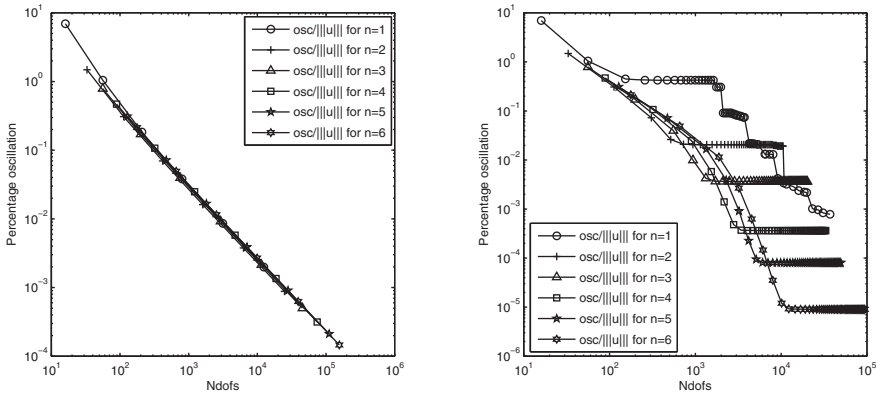
FIG. 5.2. *Performance of estimators for Example* 5.2 *with* (a) *the uniform refinement strategy and* (b) *the adaptive refinement strategy.*

smallest number of elements such that the sum of the contributions toward the estimator from these elements exceeded 50% of the value of the estimator.

When reporting numerical results we let Ndofs denote the number of degrees of freedom, let $\eta^2 = \sum_{K\in\mathcal{P}}(\Phi_K^2 + \Psi_K^2)$, $\mathrm{osc}^2 = \sum_{K\in\mathcal{P}} a_K^{-1} C_K^2 \|f - P_K f\|_{L_2(K)}^2$, and let the effectivity index of the estimator be denoted by $\vartheta = \eta / |||e|||$.

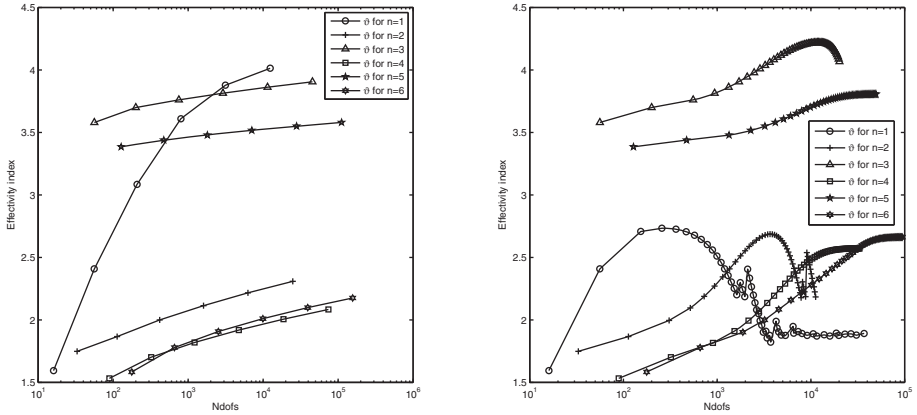
From the graphs in Figure 5.2 it can be seen that, as expected from Theorem 5.1, the estimators provide a numerical upper bound on the broken energy norm of the error. It can also be seen from Figure 5.4 that for this example, which was the one where we witnessed the most overestimation, the estimator overestimates the broken energy norm of the error by a factor less than 4.3 and that the estimator does not appear to be becoming less efficient as n increases. When uniform refinement is used it appears that the nonconforming finite element approximations of odd order outperform those of even order. It is also worth mentioning that in our example the use of an adaptive refinement strategy leads to faster reduction of the broken energy norm of the error than if uniform refinement is used.

In the cases $n = 1$ and $n = 2$ the way we have defined the oscillation, and consequently the estimator, does not reduce to how it was defined in [5] and [3],



(a) (b)

FIG. 5.3. *Data oscillation for Example* 5.2 *with* (a) *the uniform refinement strategy and* (b) *the adaptive refinement strategy.*



(a) (b)

FIG. 5.4. *Effectivity index for Example* 5.2 *with* (a) *the uniform refinement strategy and* (b) *the adaptive refinement strategy.*

respectively. This change was made since when generalized to higher orders the techniques employed in [3] to obtain the constants C_K and C_γ^K in Lemma 2.6 resulted in vast overestimation of the broken energy norm of the error being witnessed since the size of these constants would increase as n did. With the oscillation now defined as it is in Lemma 2.6 these constants are the same for all n.

6. Appendix. Let $\hat{K}$ be the element with vertices, edges, tangents, and normals as shown in Figure 6.1.

We prove the following theorem used in the proof of Theorem 3.1.

THEOREM 6.1. *Suppose* $\boldsymbol{\sigma}^\star \in H(\mathrm{div}; \hat{K})$ *is given such that*

$$\hat{\boldsymbol{n}}_i \cdot \boldsymbol{\sigma}^\star \in \mathbb{P}_n(\hat{\gamma}_i) \quad \text{on } \hat{\gamma}_i \text{ for } i = 1, 2, 3 \tag{6.1}$$

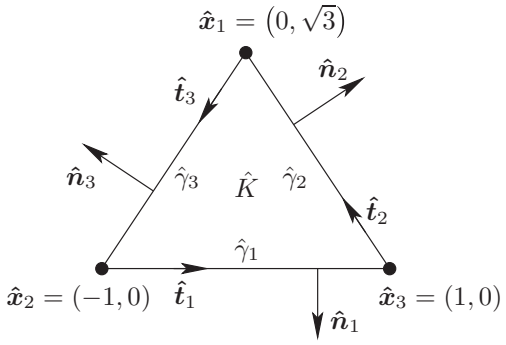


FIG. 6.1. *Direction and enumeration of the vertices, edges, unit tangent, and unit normal vectors of element* $\hat{K}$.

and

$$-\operatorname{div}\boldsymbol{\sigma}^\star \in \mathbb{P}_n\left(\hat{K}\right) \text{ on } \hat{K}. \tag{6.2}$$

Then there exists $\hat{\boldsymbol{\sigma}} \in \mathbb{P}_{n+1}(\hat{K}) \times \mathbb{P}_{n+1}(\hat{K})$ *such that*

$$\hat{\boldsymbol{n}}_i \cdot \hat{\boldsymbol{\sigma}} = \hat{\boldsymbol{n}}_i \cdot \boldsymbol{\sigma}^\star \text{ on } \hat{\gamma}_i \text{ for } i = 1, 2, 3, \tag{6.3}$$

$$-\operatorname{div}\hat{\boldsymbol{\sigma}} = -\operatorname{div}\boldsymbol{\sigma}^\star \text{ on } \hat{K}, \tag{6.4}$$

and there exists a constant C, *independent of* $\boldsymbol{\sigma}^\star$, *such that*

$$\|\hat{\boldsymbol{\sigma}}\|_{H(\operatorname{div};\hat{K})} \leq C\left(\sum_{i=1}^{3} \|\hat{\boldsymbol{n}}_i \cdot \boldsymbol{\sigma}^\star\|_{L_2(\hat{\gamma}_i)} + \|\operatorname{div}\boldsymbol{\sigma}^\star\|_{L_2(\hat{K})}\right). \tag{6.5}$$

The proof of this result will require some preparation. Let $\lambda_i \in \mathbb{P}_1(\hat{K})$ be the usual barycentric coordinates satisfying $\lambda_i(\hat{\boldsymbol{x}}_j) = \delta_{ij}$ for $j = 1, 2, 3$. For ease of exposition, it is convenient to use the convention whereby $\lambda_4 \equiv \lambda_1$, $\hat{\boldsymbol{x}}_4 \equiv \hat{\boldsymbol{x}}_1$, $\hat{\boldsymbol{t}}_4 \equiv \hat{\boldsymbol{t}}_1$, $\lambda_5 \equiv \lambda_2$, $\hat{\boldsymbol{x}}_5 \equiv \hat{\boldsymbol{x}}_2$, and $\hat{\boldsymbol{t}}_5 \equiv \hat{\boldsymbol{t}}_2$. We may parametrize an edge $\hat{\gamma}_i \subset \partial\hat{K}$ by $\boldsymbol{x}(s_i) = \hat{\boldsymbol{x}}_{i+1} + (s_i + 1)\hat{\boldsymbol{t}}_i$, where

$$s_i = \lambda_{i+2} - \lambda_{i+1} \in (-1, 1). \tag{6.6}$$

The following identities will prove useful. For $i = 1, 2, 3$,

$$\left(\hat{\boldsymbol{t}}_i \cdot \hat{\boldsymbol{n}}_i\right) = \lambda_{1|\hat{\gamma}_1} = \lambda_{2|\hat{\gamma}_2} = \lambda_{3|\hat{\gamma}_3} = 0, \tag{6.7}$$

$$\left(\hat{\boldsymbol{t}}_1 \cdot \hat{\boldsymbol{n}}_2\right) = \left(\hat{\boldsymbol{t}}_2 \cdot \hat{\boldsymbol{n}}_3\right) = \left(\hat{\boldsymbol{t}}_3 \cdot \hat{\boldsymbol{n}}_1\right) = -\left(\hat{\boldsymbol{t}}_1 \cdot \hat{\boldsymbol{n}}_3\right) = -\left(\hat{\boldsymbol{t}}_2 \cdot \hat{\boldsymbol{n}}_1\right) = -\left(\hat{\boldsymbol{t}}_3 \cdot \hat{\boldsymbol{n}}_2\right) = \frac{\sqrt{3}}{2}, \tag{6.8}$$

$$\lambda_1 + \lambda_2 + \lambda_3 = (\lambda_2 + \lambda_3)_{|\hat{\gamma}_1} = (\lambda_3 + \lambda_1)_{|\hat{\gamma}_2} = (\lambda_1 + \lambda_2)_{|\hat{\gamma}_3} = 1, \tag{6.9}$$

$$\mathbf{grad}\,\lambda_i = -\frac{1}{\sqrt{3}}\hat{\boldsymbol{n}}_i, \tag{6.10}$$

and

$$\mathbf{curl}\,\lambda_i = \frac{1}{\sqrt{3}}\hat{\boldsymbol{t}}_i. \tag{6.11}$$

LEMMA 6.2. *For $i = 1, 2, 3$, let*

$$\boldsymbol{\theta}_i^{(0)} = \frac{2}{\sqrt{3}} \left(\lambda_{i+1} \hat{\boldsymbol{t}}_{i+2} - \lambda_{i+2} \hat{\boldsymbol{t}}_{i+1} \right). \tag{6.12}$$

Then, for $i, j = 1, 2, 3$,

$$\left(\hat{\boldsymbol{n}}_i \cdot \boldsymbol{\theta}_j^{(0)} \right)_{|\hat{\gamma}_i} = \delta_{ij} \tag{6.13}$$

and

$$-\operatorname{div} \boldsymbol{\theta}_i^{(0)} = -\frac{2}{\sqrt{3}} \text{ in } K. \tag{6.14}$$

Proof. We shall prove this result for $\boldsymbol{\theta}_1^{(0)}$ since the other cases then follow by rotation. The fact that $(\hat{\boldsymbol{n}}_i \cdot \boldsymbol{\theta}_1^{(0)})_{|\hat{\gamma}_i} = 0$ when $i = 2$ or $i = 3$ is easily seen using identity (6.7). Applying identities (6.8) and (6.9) allows us to see that $(\hat{\boldsymbol{n}}_1 \cdot \boldsymbol{\theta}_1^{(0)})_{|\hat{\gamma}_1} = 1$, and making use of the identity (6.10) and simplifying using (6.8) yields (6.14). ∎

LEMMA 6.3. *For integer $m \geq 1$ and $i = 1, 2, 3$, let*

$$\boldsymbol{\theta}_i^{(m)} = \begin{cases} -\dfrac{1}{C_i^{(m,m)}} \left(4\, \mathbf{curl} \left((\lambda_{i+1}\lambda_{i+2})^{(m+1)/2} \right) + \displaystyle\sum_{l=1}^{m-2} C_i^{(l,m)} \boldsymbol{\theta}_i^{(l)} \right) & \text{if } m \text{ is odd,} \\ \dfrac{1}{C_i^{(m,m)}} \left(2\, \mathbf{curl} \left(\lambda_{i+1}\lambda_{i+2}^m - \lambda_{i+1}^m \lambda_{i+2} \right) - \displaystyle\sum_{l=2}^{m-2} C_i^{(l,m)} \boldsymbol{\theta}_i^{(l)} \right) & \text{if } m \text{ is even} \end{cases} \tag{6.15}$$

and

$$C_i^{(l,m)} = \begin{cases} \dfrac{(-1)^{(l-1)/2}\,(m+1)}{2^{m-1}} \begin{pmatrix} \dfrac{m-1}{2} \\ \dfrac{l-1}{2} \end{pmatrix} & \text{if } l \text{ and } m \text{ are both odd,} \\ \dfrac{m-2l-1}{2^{m-1}} \begin{pmatrix} m \\ l \end{pmatrix} & \text{if } l \text{ and } m \text{ are both even and } l \neq m, \\ -\dfrac{m+1}{2^{m-1}} & \text{if } l \text{ and } m \text{ are both even and } l = m, \\ 0 & \text{otherwise.} \end{cases} \tag{6.16}$$

Then, for $i, j = 1, 2, 3$ and integer $m \geq 1$,

$$\left(\hat{\boldsymbol{n}}_i \cdot \boldsymbol{\theta}_j^{(m)} \right)_{|\hat{\gamma}_i} = \begin{cases} \left(s_i^m - \dfrac{1}{2} (1, s_i^m)_{\hat{\gamma}_i} \right) \delta_{ij} = s_i^m \delta_{ij} & \text{if } m \text{ is odd,} \\ \left(s_i^m - \dfrac{1}{2} (1, s_i^m)_{\hat{\gamma}_i} \right) \delta_{ij} = \left(s_i^m - \dfrac{1}{m+1} \right) \delta_{ij} & \text{if } m \text{ is even} \end{cases} \tag{6.17}$$

and

$$-\operatorname{div} \boldsymbol{\theta}_i^{(m)} = 0 \text{ in } K. \tag{6.18}$$

Proof. If we can prove this result for the $\boldsymbol{\theta}_1^{(m)}$, then it shall hold for the $\boldsymbol{\theta}_2^{(m)}$ and $\boldsymbol{\theta}_3^{(m)}$ by rotation. The fact that the $\boldsymbol{\theta}_1^{(m)}$ are the sums of the **curl** of functions which are zero on edges $\hat{\gamma}_2$ and $\hat{\gamma}_3$ implies that $(\hat{\boldsymbol{n}}_2 \cdot \boldsymbol{\theta}_1^{(m)})_{|\hat{\gamma}_2} = (\hat{\boldsymbol{n}}_3 \cdot \boldsymbol{\theta}_1^{(m)})_{|\hat{\gamma}_3} = 0$ and $\operatorname{div} \boldsymbol{\theta}_1^{(m)} = 0$ in K. To prove that $(\hat{\boldsymbol{n}}_1 \cdot \boldsymbol{\theta}_1^{(m)})_{|\hat{\gamma}_1} = s_1^m - \frac{1}{2}(1, s_1^m)_{\hat{\gamma}_1}$ we shall first consider the case that m is an odd number. When $m = 1$, we can apply identity (6.11) to find that

$$\boldsymbol{\theta}_1^{(1)} = -\frac{2}{\sqrt{3}} \left(\lambda_3 \hat{\boldsymbol{t}}_2 + \lambda_2 \hat{\boldsymbol{t}}_3 \right)$$

to which we can apply identity (6.8) along with the definition of s_1 (6.6) to yield $(\hat{\boldsymbol{n}}_1 \cdot \boldsymbol{\theta}_1^{(1)})_{|\hat{\gamma}_1} = s_1$. We shall now assume that the result holds for all odd m up to $m = k$ and complete the proof by induction by showing that the result holds for $m = k + 2$. As with $m = 1$ we can apply identity (6.8) to give

$$-4 \left(\hat{\boldsymbol{n}}_1 \cdot \mathbf{curl} \left((\lambda_2 \lambda_3)^{(k+3)/2} \right) \right)_{|\hat{\gamma}_1} = (k+3) (\lambda_2 \lambda_3)^{(k+1)/2} (\lambda_3 - \lambda_2).$$

By combining the definition of s_1 (6.6) with identity (6.9) we see that $\lambda_{2|\hat{\gamma}_1} = \frac{1}{2} - \frac{1}{2} s_1$ and $\lambda_{3|\hat{\gamma}_1} = \frac{1}{2} + \frac{1}{2} s_1$, which we can use to obtain

$$\begin{aligned}
-4 \left(\hat{\boldsymbol{n}}_1 \cdot \mathbf{curl} \left((\lambda_2 \lambda_3)^{(k+3)/2} \right) \right)_{|\hat{\gamma}_1} &= (k+3) \left(\frac{1}{4} - \frac{1}{4} s_1^2 \right)^{(k+1)/2} s_1 \\
&= (k+3) \sum_{l=0}^{(k+1)/2} \left(\begin{array}{c} \frac{k+1}{2} \\ l \end{array} \right) \frac{(-1)^l}{2^{k+1}} s_1^{2l+1} \\
&= (k+3) \sum_{\substack{\text{odd } l: \\ 1 \le l \le k+2}} \left(\begin{array}{c} \frac{k+1}{2} \\ \frac{l-1}{2} \end{array} \right) \frac{(-1)^{(l-1)/2}}{2^{k+1}} s_1^l \\
&= \sum_{l=1}^{k+2} C_1^{(l,k+2)} s_1^l
\end{aligned}$$

upon applying the binomial theorem and definition (6.16). It then follows that $(\hat{\boldsymbol{n}}_1 \cdot \boldsymbol{\theta}_1^{(k+2)})_{|\hat{\gamma}_1} = s_1^{k+2}$ since $(\hat{\boldsymbol{n}}_1 \cdot \boldsymbol{\theta}_1^{(l)})_{|\hat{\gamma}_1} = s_1^l$ for odd l between 1 and k inclusive.

By applying the same method that we used for the odd case, we can obtain that $(\hat{\boldsymbol{n}}_1 \cdot \boldsymbol{\theta}_1^{(2)})_{|\hat{\gamma}_1} = s_1^2 - \frac{1}{3}$ and

$$2 \left(\hat{\boldsymbol{n}}_1 \cdot \mathbf{curl} \left(\lambda_2 \lambda_3^{k+2} - \lambda_2^{k+2} \lambda_3 \right) \right)_{|\hat{\gamma}_1} = \frac{k+1}{2^{k+1}} + \sum_{l=2}^{k+2} C_1^{(l,k+2)} s_1^l,$$

and so using induction as before we observe that

$$\left(\hat{\boldsymbol{n}}_1 \cdot \boldsymbol{\theta}_1^{(k+2)} \right)_{|\hat{\gamma}_1} = s_1^{k+2} + \nu,$$

where ν is a constant which must take the value $-\frac{1}{2} \left(s_1^{k+2}, 1 \right)_{\hat{\gamma}_1}$ since

$$\left(\hat{\boldsymbol{n}}_1 \cdot \boldsymbol{\theta}_1^{(k+2)}, 1 \right)_{\hat{\gamma}_1} = \sum_{j=1}^{3} \left(\hat{\boldsymbol{n}}_j \cdot \boldsymbol{\theta}_1^{(k+2)}, 1 \right)_{\hat{\gamma}_j} = \left(\operatorname{div} \boldsymbol{\theta}_1^{(k+2)}, 1 \right)_{\hat{K}} = 0. \qquad \square$$

LEMMA 6.4. *For integer $m_1, m_2, m_3 \geq 1$, let $\boldsymbol{\theta}_{\hat{K}}^{(0,0,0)} = 0$,*

$$\boldsymbol{\theta}_{\hat{K}}^{(m_1,0,0)} = \frac{1}{m_1+2}\left(m_1\boldsymbol{\theta}_{\hat{K}}^{(m_1-1,0,0)} + 2\lambda_1^{m_1}\left(\lambda_3\hat{\boldsymbol{t}}_2 - \lambda_2\hat{\boldsymbol{t}}_3\right)\right), \tag{6.19}$$

$$\boldsymbol{\theta}_{\hat{K}}^{(0,m_2,0)} = \frac{1}{m_2+2}\left(m_2\boldsymbol{\theta}_{\hat{K}}^{(0,m_2-1,0)} + 2\lambda_2^{m_2}\left(\lambda_1\hat{\boldsymbol{t}}_3 - \lambda_3\hat{\boldsymbol{t}}_1\right)\right), \tag{6.20}$$

$$\boldsymbol{\theta}_{\hat{K}}^{(0,0,m_3)} = \frac{1}{m_3+2}\left(m_3\boldsymbol{\theta}_{\hat{K}}^{(0,0,m_3-1)} + 2\lambda_3^{m_3}\left(\lambda_2\hat{\boldsymbol{t}}_1 - \lambda_1\hat{\boldsymbol{t}}_2\right)\right), \tag{6.21}$$

$$\begin{aligned}\boldsymbol{\theta}_{\hat{K}}^{(m_1,m_2,0)} = \frac{1}{2\left(m_1+m_2+2\right)}\Big(&m_1\boldsymbol{\theta}_{\hat{K}}^{(m_1-1,m_2,0)} + m_2\boldsymbol{\theta}_{\hat{K}}^{(m_1,m_2-1,0)} \\ &+ 2\lambda_1^{m_1}\lambda_2^{m_2}\left(\lambda_1\hat{\boldsymbol{t}}_3 - \lambda_2\hat{\boldsymbol{t}}_3 + \lambda_3\left(\hat{\boldsymbol{t}}_2 - \hat{\boldsymbol{t}}_1\right)\right)\Big),\end{aligned} \tag{6.22}$$

$$\begin{aligned}\boldsymbol{\theta}_{\hat{K}}^{(0,m_2,m_3)} = \frac{1}{2\left(m_2+m_3+2\right)}\Big(&m_2\boldsymbol{\theta}_{\hat{K}}^{(0,m_2-1,m_3)} + m_3\boldsymbol{\theta}_{\hat{K}}^{(0,m_2,m_3-1)} \\ &+ 2\lambda_2^{m_2}\lambda_3^{m_3}\left(\lambda_2\hat{\boldsymbol{t}}_1 - \lambda_3\hat{\boldsymbol{t}}_1 + \lambda_1\left(\hat{\boldsymbol{t}}_3 - \hat{\boldsymbol{t}}_2\right)\right)\Big),\end{aligned} \tag{6.23}$$

$$\begin{aligned}\boldsymbol{\theta}_{\hat{K}}^{(m_1,0,m_3)} = \frac{1}{2\left(m_3+m_1+2\right)}\Big(&m_3\boldsymbol{\theta}_{\hat{K}}^{(m_1,0,m_3-1)} + m_1\boldsymbol{\theta}_{\hat{K}}^{(m_1-1,0,m_3)} \\ &+ 2\lambda_3^{m_3}\lambda_1^{m_1}\left(\lambda_3\hat{\boldsymbol{t}}_2 - \lambda_1\hat{\boldsymbol{t}}_2 + \lambda_2\left(\hat{\boldsymbol{t}}_1 - \hat{\boldsymbol{t}}_3\right)\right)\Big),\end{aligned} \tag{6.24}$$

$$\begin{aligned}\boldsymbol{\theta}_{\hat{K}}^{(m_1,m_2,m_3)} = \frac{1}{3\left(m_1+m_2+m_3+2\right)}\Big(&m_1\boldsymbol{\theta}_{\hat{K}}^{(m_1-1,m_2,m_3)} \\ &+ m_2\boldsymbol{\theta}_{\hat{K}}^{(m_1,m_2-1,m_3)} + m_3\boldsymbol{\theta}_{\hat{K}}^{(m_1,m_2,m_3-1)} \\ &+ 2\lambda_1^{m_1}\lambda_2^{m_2}\lambda_3^{m_3}\left(\lambda_1\left(\hat{\boldsymbol{t}}_3 - \hat{\boldsymbol{t}}_2\right) + \lambda_2\left(\hat{\boldsymbol{t}}_1 - \hat{\boldsymbol{t}}_3\right) + \lambda_3\left(\hat{\boldsymbol{t}}_2 - \hat{\boldsymbol{t}}_1\right)\right)\Big).\end{aligned} \tag{6.25}$$

Then, for $i = 1, 2, 3$ and integer $m_1, m_2, m_3 \geq 0$,

$$\left(\hat{\boldsymbol{n}}_i \cdot \boldsymbol{\theta}_{\hat{K}}^{(m_1,m_2,m_3)}\right)_{|\hat{\gamma}_i} = 0 \tag{6.26}$$

and

$$-\operatorname{div}\boldsymbol{\theta}_{\hat{K}}^{(m_1,m_2,m_3)} = \lambda_1^{m_1}\lambda_2^{m_2}\lambda_3^{m_3} - \frac{1}{\sqrt{3}}\left(\lambda_1^{m_1}\lambda_2^{m_2}\lambda_3^{m_3}, 1\right)_{\hat{K}} \ \text{in } \hat{K}. \tag{6.27}$$

Proof. By making use of the identity (6.7) it can be seen that (6.26) holds. Now, using the product rule and identity (6.10) yields, for $m_1 \geq 1$,

$$-\operatorname{div}\left(2\lambda_1^{m_1}\left(\lambda_3\hat{\boldsymbol{t}}_2 - \lambda_2\hat{\boldsymbol{t}}_3\right)\right) = \left(m_1+2\right)\lambda_1^{m_1} - m_1\lambda_1^{m_1-1} \text{ in } \hat{K} \tag{6.28}$$

after simplifying with the help of identity (6.9). Now this implies that

$$-\operatorname{div}\boldsymbol{\theta}_{\hat{K}}^{(1,0,0)} = -\frac{1}{3}\operatorname{div}\left(2\lambda_1\left(\lambda_3\hat{\boldsymbol{t}}_2 - \lambda_2\hat{\boldsymbol{t}}_3\right)\right) = \lambda_1 - \frac{1}{3} \text{ in } \hat{K},$$

and so (6.27) is true for $m_2 = m_3 = 0$ and $m_1 = 1$. Now assume that (6.27) holds for $m_2 = m_3 = 0$ and $1 \leq m_1 \leq k$. Hence, from (6.28) and the definition of $\boldsymbol{\theta}_{\hat{K}}^{(m_1,0,0)}$ (6.19) we have

$$-\operatorname{div}\boldsymbol{\theta}_{\hat{K}}^{(k+1,0,0)} = \lambda_1^{k+1} + \nu \text{ in } \hat{K},$$

where ν is a constant. Now since

$$\left(\operatorname{div}\boldsymbol{\theta}_{\hat{K}}^{(k+1,0,0)},1\right)_{\hat{K}}=\sum_{j=1}^{3}\left(\hat{\boldsymbol{n}}_j\cdot\boldsymbol{\theta}_{\hat{K}}^{(k+1,0,0)},1\right)_{\hat{\gamma}_j}=0,$$

the constant ν must be equal to $-\frac{1}{\sqrt{3}}(\lambda_1^{k+1},1)_{\hat{K}}$, and hence (6.27) is true in the case $m_1\geq 1$ and $m_2=m_3=0$ by induction. This means that (6.27) then also holds for the functions defined in (6.20) and (6.21) by rotation. Likewise, for $m_1,m_2\geq 1$, we have

$$\begin{aligned}&-\operatorname{div}\left(2\lambda_1^{m_1}\lambda_2^{m_2}\left(\lambda_1\hat{\boldsymbol{t}}_3-\lambda_2\hat{\boldsymbol{t}}_3+\lambda_3\left(\hat{\boldsymbol{t}}_2-\hat{\boldsymbol{t}}_1\right)\right)\right)\\&=2\left(2+m_1+m_2\right)\lambda_1^{m_1}\lambda_2^{m_2}-m_1\lambda_1^{m_1-1}\lambda_2^{m_2}-m_2\lambda_1^{m_1}\lambda_2^{m_2-1}\text{ in }\hat{K},\end{aligned}$$

from which (6.27) can be proved by induction for the functions defined in (6.22) (and therefore, by rotation, those defined in (6.23) and (6.24)) by also making use of the fact that (6.27) holds for the functions defined in (6.19), (6.20), and (6.21). Finally, the fact that (6.27) is true for all of the other cases allows us to show that (6.27) holds for the functions defined in (6.25) in the same way. ∎

Let $\{P^{(i)}\}_{i=1}^{\dim\mathbb{P}_n(\hat{K})}$ be an orthonormal basis for $\mathbb{P}_n(\hat{K})$ such that $P^{(1)}=3^{-1/4}$. Likewise, for $j=1,2,3$, let $\{P_j^{(i)}\}_{i=0}^{n}$ be an orthonormal basis for $\mathbb{P}_n(\hat{\gamma}_j)$ such that $P_j^{(0)}=2^{-1/2}$.

Corollary 6.5. *For $i=0,\ldots,n$ and $k=1,2,3$ there exist functions $\boldsymbol{\vartheta}_k^{(i)}\in\mathbb{P}_{n+1}(\hat{K})\times\mathbb{P}_{n+1}(\hat{K})$ such that, for $j=1,2,3$,*

$$\left(\hat{\boldsymbol{n}}_j\cdot\boldsymbol{\vartheta}_k^{(i)}\right)_{|\hat{\gamma}_j}=\delta_{jk}P_j^{(i)}\ \text{ and }\ -\operatorname{div}\boldsymbol{\vartheta}_k^{(i)}=-\frac{\sqrt{2}}{\sqrt{3}}\delta_{i0}\ \text{ in }\hat{K}.\tag{6.29}$$

Also, for $i=2,\ldots,\dim\mathbb{P}_n(\hat{K})$ there exist functions $\boldsymbol{\vartheta}_{\hat{K}}^{(i)}\in\mathbb{P}_{n+1}(\hat{K})\times\mathbb{P}_{n+1}(\hat{K})$ such that, for $j=1,2,3$,

$$\left(\hat{\boldsymbol{n}}_j\cdot\boldsymbol{\vartheta}_{\hat{K}}^{(i)}\right)_{|\hat{\gamma}_j}=0\ \text{ and }\ -\operatorname{div}\boldsymbol{\vartheta}_{\hat{K}}^{(i)}=P^{(i)}\ \text{ in }\hat{K}.\tag{6.30}$$

Proof. This follows from Lemmas 6.2, 6.3, and 6.4 upon noting that, for $j=1,2,3$, $\{(\hat{\boldsymbol{n}}_j\cdot\boldsymbol{\theta}_j^{(i)})_{|\hat{\gamma}_j}\}_{i=1}^{n}$ forms a basis for $\{v:v\in\mathbb{P}_n(\hat{\gamma}_j),(v,P_j^{(0)})_{\hat{\gamma}_j}=0\}$ and that, for integer $m_1,m_2,m_3\geq 0$, $\{-\operatorname{div}\boldsymbol{\theta}_{\hat{K}}^{(m_1,m_2,m_3)}\}_{m_1+m_2+m_3=n}$ forms a basis for $\{v:v\in\mathbb{P}_n(\hat{K}),(v,P^{(1)})_{\hat{K}}=0\}$. ∎

Finally, we are in a position to provide the proof of Theorem 6.1.

Proof. Let

$$\hat{\boldsymbol{\sigma}}=\sum_{j=1}^{3}\sum_{i=0}^{n}\left(\hat{\boldsymbol{n}}_j\cdot\boldsymbol{\sigma}^\star,P_j^{(i)}\right)_{\hat{\gamma}_j}\boldsymbol{\vartheta}_j^{(i)}-\sum_{i=2}^{\dim\mathbb{P}_n(\hat{K})}\left(\operatorname{div}\boldsymbol{\sigma}^\star,P^{(i)}\right)_{\hat{K}}\boldsymbol{\vartheta}_{\hat{K}}^{(i)}.\tag{6.31}$$

Then, by Corollary 6.5, we have

$$\hat{\boldsymbol{n}}_j\cdot\hat{\boldsymbol{\sigma}}=\sum_{i=0}^{n}\left(\hat{\boldsymbol{n}}_j\cdot\boldsymbol{\sigma}^\star,P_j^{(i)}\right)_{\hat{\gamma}_j}P_j^{(i)}=\hat{\boldsymbol{n}}_j\cdot\boldsymbol{\sigma}^\star\text{ on }\hat{\gamma}_j\text{ for }j=1,2,3$$

and

$$-\operatorname{div}\hat{\boldsymbol{\sigma}}=-\frac{\sqrt{2}}{\sqrt{3}}\sum_{j=1}^{3}\left(\hat{\boldsymbol{n}}_j\cdot\boldsymbol{\sigma}^\star,P_j^{(0)}\right)_{\hat{\gamma}_j}-\sum_{i=2}^{\dim\mathbb{P}_n(\hat{K})}\left(\operatorname{div}\boldsymbol{\sigma}^\star,P^{(i)}\right)_{\hat{K}}P^{(i)}\text{ in }\hat{K}.\tag{6.32}$$

To see that the right-hand side of (6.32) is equal to $-\operatorname{div}\boldsymbol{\sigma}^\star$, observe that, since $\boldsymbol{\sigma}^\star \in H(\operatorname{div};\hat{K})$,

$$\sum_{j=1}^{3} (\hat{\boldsymbol{n}}_j \cdot \boldsymbol{\sigma}^\star, 1)_{\hat{\gamma}_j} = (\operatorname{div}\boldsymbol{\sigma}^\star, 1)_{\hat{K}}$$

and so

$$-\frac{\sqrt{2}}{\sqrt{3}} \sum_{j=1}^{3} \left(\hat{\boldsymbol{n}}_j \cdot \boldsymbol{\sigma}^\star, P_j^{(0)}\right)_{\hat{\gamma}_j} = -\left(\operatorname{div}\boldsymbol{\sigma}^\star, P^{(1)}\right)_{\hat{K}} P^{(1)}.$$

Finally, we can arrive at inequality (6.5), by applying Cauchy–Schwarz inequalities to the expression obtained for $\|\hat{\boldsymbol{\sigma}}\|_{L_2(\hat{K})}$ using (6.31). ∎

REFERENCES

[1] S. Adjerid, J. E. Flaherty, and I. Babuška, *A posteriori error estimation for the finite element method-of-lines solution of parabolic problems*, Math. Models Methods Appl. Sci., 9 (1999), pp. 261–286.

[2] M. Ainsworth and J. T. Oden, *A Posteriori Error Estimation in Finite Element Analysis*, Pure Appl. Math., Wiley-Interscience, New York, 2000.

[3] M. Ainsworth and R. Rankin, *Robust a posteriori error estimation for the nonconforming Fortin–Soulie finite element approximation*, Math. Comp., 77 (2008), pp. 1917–1939.

[4] M. Ainsworth, *A posteriori error estimation for non-conforming quadrilateral finite elements*, Int. J. Numer. Anal. Model., 2 (2005), pp. 1–18.

[5] M. Ainsworth, *Robust a posteriori error estimation for nonconforming finite element approximation*, SIAM J. Numer. Anal., 42 (2005), pp. 2320–2341.

[6] M. Ainsworth, *Corrigendum: A posteriori error estimation for non-conforming quadrilateral finite elements*, Int. J. Numer. Anal. Model., 4 (2007), pp. 141–142.

[7] M. Ainsworth, *A posteriori error estimation for discontinuous Galerkin finite element approximation*, SIAM J. Numer. Anal., 45 (2007), pp. 1777–1798.

[8] I. Babuška and D. Yu, *Asymptotically exact a posteriori error estimator for biquadratic elements*, Finite Elem. Anal. Des., 3 (1987), pp. 341–354.

[9] Á. Baran and G. Stoyan, *Gauss-Legendre elements: A stable, higher order non-conforming finite element family*, Computing, 79 (2007), pp. 1–21.

[10] C. Carstensen and J. Hu, *A unifying theory of a posteriori error control for nonconforming finite element methods*, Numer. Math., 107 (2007), pp. 473–502.

[11] Y. Cha, M. Lee, and S. Lee, *Stable nonconforming methods for the Stokes problem*, Appl. Math. Comput., 114 (2000), pp. 155–174.

[12] M. Crouzeix and R. S. Falk, *Nonconforming finite elements for the Stokes problem*, Math. Comp., 52 (1989), pp. 437–456.

[13] M. Crouzeix and P.-A. Raviart, *Conforming and nonconforming finite element methods for solving the stationary Stokes equations.* I, Rev. Française Automat. Informat. Recherche Opérationnelle Sér. Rouge, 7 (1973), pp. 33–75.

[14] E. Dari, R. Duran, C. Padra, and V. Vampa, *A posteriori error estimators for nonconforming finite element methods*, RAIRO Model. Math. Anal. Numer., 30 (1996), pp. 385–400.

[15] W. Dörfler, *A convergent adaptive algorithm for Poisson's equation*, SIAM J. Numer. Anal., 33 (1996), pp. 1106–1124.

[16] M. Fortin and M. Soulie, *A nonconforming piecewise quadratic finite element on triangles*, Internat. J. Numer. Methods Engrg., 19 (1983), pp. 505–520.

[17] P. Morin, R. H. Nochetto, and K. G. Siebert, *Data oscillation and convergence of adaptive FEM*, SIAM J. Numer. Anal., 38 (2000), pp. 466–488.

[18] L. E. Payne and H. F. Weinberger, *An optimal Poincaré inequality for convex domains*, Arch. Ration. Mech. Anal., 5 (1960), pp. 286–292.

[19] G. Stoyan and Á. Baran, *Crouzeix-Velte decompositions for higher-order finite elements*, Comput. Math. Appl., 51 (2006), pp. 967–986.

[20] R. Verfürth, *A Review of A Posteriori Error Estimation and Adaptive Mesh-Refinement Techniques*, Wiley-Teubner, Chichester, UK, 1996.

SIAM J. NUMER. ANAL.
Vol. 46, No. 6, pp. 3233–3263
© 2008 Society for Industrial and Applied Mathematics

CONVERGENCE OF A SEMIDISCRETE TWO-GRID ALGORITHM FOR THE CONTROLLABILITY OF THE $1-d$ WAVE EQUATION*

MIHAELA NEGREANU†

Abstract. The problem of exact controllability of elastic strings has been extensively studied during the last years. We consider the problem of computing numerically the boundary control for a finite-dimensional system obtained by discretizing in space the $1-d$ wave equation. More precisely, we analyze whether the controls of numerical approximation schemes converge to the control of the continuous wave equation as the mesh size tends to zero. It is by now well known that, due to high-frequency spurious oscillations, numerical instabilities occur and may lead to the failure of convergence of some apparently natural numerical algorithms. In other words, the classical convergence property of numerical schemes does not guarantee a stable and convergent approximation of controls. Several remedies have been proposed in the literature to compensate for this fact: Tychonoff regularization, Fourier filtering, and mixed finite elements. In this paper we prove that the two-grid method proposed by Glowinski in [*J. Comput. Phys.*, 103 (1992), pp. 189–221] to numerically approximate the control of the wave equation converges in $1-d$. We prove this result in the context of the finite-element space semidiscretization. Our method of proof relies essentially on the particular properties of the Fourier representation of the initial data of the coarse mesh when projected into the fine one. The explicit representation formula of the solutions shows that the high-frequency components are modulated by some weights that diminish the effect of these spurious components. This fact, combined with discrete multipliers techniques, allows us to prove uniform observability inequalities. Classical arguments then allow proving the uniform boundedness of the controls and passing to the limit as the mesh size tends to zero. In this way we prove the convergence of the controls of the finite-element semidiscrete approximation of the $1-d$ wave equation with a boundary control on one extreme. It is important to underline that the controls obtained by this two-grid algorithm are not exact in the sense that they only guarantee the controllability of the projection of the solutions into the coarse mesh.

Key words. wave equation, boundary controllability, uniform observability, two-grid algorithm, finite-element discretization

AMS subject classifications. 65N30, 65N55, 35L05, 34H05, 93B05, 93B07.

DOI. 10.1137/06064915X

1. Introduction.

Consider the $1-d$ wave equation, with constant coefficients and control on one extreme:

$$\begin{cases} u_{tt} - u_{xx} = 0, & 0 < x < 1, \quad 0 < t < T, \\ u(0,t) = 0, \quad u(1,t) = v(t), & 0 < t < T, \\ u(x,0) = u^0(x), \qquad u_t(x,0) = u^1(x), & 0 < x < 1. \end{cases} \tag{1.1}$$

Here (u^0, u^1) are the initial state and v is the control function that acts on the system through the extreme $x = 1$ of the space interval $(0,1)$. System (1.1) describes the vibrations of a controlled string, and it is well posed in appropriate functional spaces.

When no control acts on the boundary $x = 1$ of the string, i.e., when $v = 0$, the energy:

$$E(t) = \frac{1}{2} \int_0^1 \left(\mid u_x(x,t) \mid^2 + \mid u_t(x,t) \mid^2 \right) dx \tag{1.2}$$

*Received by the editors January 5, 2006; accepted for publication (in revised form) May 9, 2008; published electronically October 17, 2008. This research was supported by grant BFM2002-04572-C02-01 of the Spanish MEC and by the Investigation Group 920894 of UCM.

http://www.siam.org/journals/sinum/46-6/64915.html

†Department of Applied Mathematics, Faculty of Mathematics, University Complutense of Madrid, 28040 Madrid, Spain (negreanu@mat.ucm.es).

of the solutions is conserved in time

$$\frac{dE(t)}{dt} = E'(t) = 0 \quad \forall t \in [0,T].$$

The starting point of our study is the following well-known exact controllability result for (1.1) (see [6]): *Given $T \geq 2$ and $(u^0, u^1) \in L^2(0,1) \times H^{-1}(0,1)$, there exists a control function $v \in L^2(0,T)$ such that the solution $u = u(x,t)$ of* (1.1) *satisfies*

$$u(T) = u_t(T) = 0, \qquad 0 < x < 1. \tag{1.3}$$

The goal in this control problem is to drive solutions to equilibrium at time $t = T$. Once the equilibrium configuration is reached at time $t = T$, by taking null control for $t \geq T$, i.e., $v = 0$ for $t \geq T$, the solution remains at rest for all $t \geq T$.

For (u^0, u^1) and $T \geq 2$ as above, there may exist infinitely many controls v with (1.3). The *Hilbert uniqueness method* (HUM) introduced by Lions in 1986 (see [6]) furnishes the one of minimal $L^2(0,T)$-norm, the so-called HUM control. Note that (1.3) makes sense, since in these conditions, (1.1) admits a unique solution $u \in \mathcal{C}([0,T]; L^2(0,1)) \cap \mathcal{C}^1([0,T]; H^{-1}(0,1))$.

Let us now introduce the homogeneous system (the adjoint of (1.1))

$$\begin{cases} \phi_{tt} - \phi_{xx} = 0, & 0 < x < 1, \quad 0 < t < T, \\ \phi(0,t) = \phi(1,t) = 0, & 0 < t < T, \\ \phi(x,T) = \phi^0(x), \qquad \phi_t(x,T) = \phi^1(x), & 0 < x < 1. \end{cases} \tag{1.4}$$

When $(\phi^0, \phi^1) \in H_0^1(0,1) \times L^2(0,1)$, (1.4) admits a unique solution $\phi = \phi(x,t) \in \mathcal{C}^0\left([0,T], H_0^1(0,1)\right) \cap \mathcal{C}^1\left([0,T], L^2(0,1)\right)$.

With the aid of the HUM, the property of the exact controllability of (1.1) is shown to be equivalent to the following observability inequality of the adjoint system (1.4): *Given $T \geq 2$, there exists a positive constant $C(T) > 0$ such that*

$$E(0) \leq C(T) \int_0^T |\, \phi_x(1,t)\,|^2 \, dt \tag{1.5}$$

holds for every solution $\phi = \phi(x,t)$ of the adjoint problem (1.4).

Inequality (1.5), when it holds, guarantees that the total energy of solutions can be observed or estimated from the energy concentrated or measured on the extreme x = 1 of the string during the time interval $(0,T)$, uniformly in the whole class of solutions of (1.4). Here and in what follows, the best constant $C(T)$ in inequality (1.5) will be referred to as the *observability constant*.

It is well known that the problems of observability and controllability are equivalent provided one chooses an appropriate functional setting, which depends on the equation.

When (1.5) holds, the quadratic functional

$$J\left(\phi^0, \phi^1\right) = \frac{1}{2} \int_0^T |\phi_x(1,t)|^2 \, dt + \int_0^1 u^0(x)\phi_t(0)dx - \left\langle u^1(x), \phi(0)\right\rangle_{H^{-1} \times H_0^1} \tag{1.6}$$

is coercive in the space $H_0^1(0,1) \times L^2(0,1)$. Therefore, *when the observability inequality* (1.5) *holds, the functional J has an unique minimizer $(\hat{\phi}^0, \hat{\phi}^1)$ in $H_0^1(0,1) \times L^2(0,1)$. The control $v = -\hat{\phi}_x(1,t)$, with $\hat{\phi} = \hat{\phi}(x,t)$ the solution of* (1.4) *corresponding to $(\hat{\phi}^0, \hat{\phi}^1)$ is such that the solution of* (1.1) *satisfies* (1.3).

Consequently, in this way, the exact controllability problem (1.3) is reduced to the analysis of the so-called *inverse observability inequality* (1.5). Furthermore, the minimization of the functional J provides a useful way of constructing the control of minimal $L^2(0,T)$-norm. Once we know that such a control exists, it is natural to try to construct this control numerically.

The goal of this paper is to obtain a convergent numerical approximation of the exact control.

A possible procedure is to approximate the wave operator by some sequence of discrete or semidiscrete operators and to try to obtain the control of system (1.1) as the limit of the sequence of the controls of the approximating equations.

However, it is well known that the interaction of waves with the numerical mesh produces dispersion phenomena and spurious high-frequency oscillations [14, 15]. In particular, the velocity of propagation of waves, more precisely, the so-called *group velocity*, may converge to zero when the wavelength of solutions is of the order of the size of the mesh and the latter tends to zero. As a consequence of this fact, the time needed to uniformly (with respect to the mesh size) observe (or control) the numerical waves from the boundary tends to infinity as the mesh becomes finer and finer. This is the reason for the unstable behavior of the control and observation properties of most numerical approximation schemes as the mesh size tends to zero. Due to this, the observation and control properties of the discrete model may eventually disappear as the mesh size tends to zero. This occurs for the standard space semidiscrete approximation by finite-differences and finite-elements, and it has some impact when computing the control for the continuous model.

The main reason for the lack of uniform observability/controllability of the numerical high-frequency spurious solutions is that they generate high-frequency wave packets for which the group velocity is of the order of the mesh size (see [17]). Thus, as the mesh size tends to zero, since the velocity becomes smaller and smaller, the time for observability/controllability is increasing in a divergent way. This fact is related with the dispersion diagram associated to the numerical approximation scheme, since, roughly, the slope of the dispersion diagram is the group velocity of the propagation of wave packets and also coincides with the spectral gap.

Several remedies have been proposed in the literature to compensate for this fact: Tychonoff regularization, Fourier filtering, and mixed finite elements. (see [17]).

One way of restoring the uniform observability inequality for the homogeneous adjoint system consists in filtering the high-frequency modes of the solution by truncating the Fourier series. This is equivalent with the uniform controllability of the projection of the solution of the semidiscrete controlled system over the subspaces generated by the remaining eigenmodes, where the high frequencies have been filtered (see [5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16]). However, in practice, it is not a very efficient result to compute these projections of the solutions and can be hard to implement in numerical simulations. Indeed, solving the semidiscrete system provides the nodal values of the solution. Then, one needs to compute its Fourier coefficients and, once these are done, to recalculate the nodal values of the filtered/truncated solution. Therefore, it is convenient to explore other ways of avoiding these high-frequency pathologies that do not require going back and forth from the physical space to the frequency one. In addition, the uniform time T needed for this algorithm to converge as the mesh size tends to zero depends strongly on the value of the filtering parameter, since the latter determines the minimal velocity of the propagation of the remaining numerical waves.

In [13] the authors proposed and analyzed a wavelet-based numerical method for exact controllability of the $1-d$ wave equation, concentrating on the particular case

of *Dirichlet* controls in the context of the finite-element space semidiscretization with spline wavelet basis functions.

The algorithm for the numerical construction of the controls is based on a conjugate gradient method proposed by Glowinski [2] and a two-grid filtering technique in the wavelet basis. The wavelet filtering yields numerical exact controllability in the optimal time of the continuous problem and the convergence of the numerical controls using a multilevel discretization filtering technique. Classical arguments then allow proving the uniform boundedness of the numerical controls and passing to the limit as the mesh size tends to zero. Numerical experiments show that wavelet filtering yields numerical controls with the same control time T as in the continuous case.

Glowinski in [2] proposed an alternative method for efficiently computing the controls. It is a two-grid algorithm based on the use of two numerical meshes: the coarse one G_C^{2h}, with step size $2h$ and the finer one G_F^h, corresponding to h. The bigrid strategy allows us to eliminate the highly oscillating parts of the solution of the uncontrolled semidiscrete system without filtering and performs a part of the computation of the algorithm on a coarse grid. Briefly, this algorithm consists in computing the numerical controls on the fine grid by minimizing a discrete version of the functional J only over the slowly data subspace projected from the coarse grid.

There are numerical evidences of the efficiency of this method introduced at the beginning of the nineties, but the proof of convergence was given only recently in [10] in the case of finite-difference semidiscretizations. The numerical efficiency of the method was illustrated in several experiments in [3] and [1].

This method is closely related to the filtering technique considered in [5]. In fact, it is about replacing a filtered solutions subclass by another one where we consider only the slowly oscillating initial data in the *fine grid* G_F^h, coming from a *coarse grid* G_C^{2h}, with mesh size $2h$. The multigrid method's advantage consists in that it is possible to work directly and completely in the physical space, without the necessity of using the Fourier transform and its inverse, permanently, opposite the filtering method.

The main objective of this paper is to prove that this two-grid method does indeed provide a convergent algorithm. This is done in the context of the finite-element semidiscrete approximation of the $1-d$ wave equation (1.1). It is worth noting that the two-grid technique was introduced by Glowinski in [2] precisely in the context of finite elements in which it can be applied with more flexibility.

The key ingredient of the proof is the uniform observability inequality for any solution of the semidiscrete uncontrolled wave equation defined on the fine grid with slowly oscillating initial data. The proof of the uniform observability combines two ingredients: (a) the Fourier representation of slow data on the fine grid; (b) the multiplier technique.

The paper is organized as follows: in section 2 we present some known properties of the solutions of the finite-element semidiscretization of the adjoint system (1.4), which are used in the proof of the main result. In section 3 we introduce the two-grid method, which permits its application in the finite-element semidiscrete approximation and in other circumstances as finite-difference approximations (see [10]). Section 4 is devoted to the proof of one of our main results, namely, the uniform discrete version of (1.5) for the solutions of the finite-element semidiscretization of (1.4) using the two-grid method. This is precisely the context in which this method was introduced and used (see [2]).

We discuss the consequences of this result concerning the controllability of the discrete controlled system (2.6) in section 5. Moreover, we prove the convergence of

the sequence of controls for system (2.6) that the two-grid algorithm provides as the mesh size h tends to zero.

The convergence results we present in this paper have been announced in [10] for the finite-difference semidiscretization of the $1-d$ wave equation. The same proof is done in that case, and we choose to present the complete proof for the finite-element approximation because it is more complete and it can be applied with more flexibility.

2. The semidiscrete problem: Notations and preliminary results. Take $N \in \mathbb{N}$, set $h = 1/(N+1)$. The finite-element semidiscrete version of system (1.4) (and (1.1), respectively) is obtained by a Galerkin approximation of the solution

$$\phi = \sum_{j=1}^{N} \phi_j(t)\psi_j(x), \tag{2.1}$$

when the basis $(\psi_j(x))_{j=1,\dots,N}$ of $H_0^1(0,1)$ is built by means of P_1 finite elements

$$S_h = \left\{ (\psi_j(x))_{j=1,\dots,N} : \quad \psi_j(x) = \left[1 - \frac{|x-x_j|}{h}\right]^+, \quad j = 1,\dots,N,\ x_j = jh \right\}.$$

The variational formulation of (1.4) in S_h is given by

$$-\int_0^1 \phi'' \psi_k dx = \int_0^1 \phi_x \psi_{k,x} dx, \qquad 0 < t < T \qquad \forall\, k = 1,\dots,N.$$

Taking into account that

$$\int_0^1 \psi_j(x)\psi_k(x) = \frac{h}{6}, \quad \text{if} \quad k = j-1, j+1, \quad \int_0^1 \psi_j(x)\psi_k(x) = 0, \quad \text{if} \quad |k-j| > 2, \tag{2.2}$$

$$\int_0^1 |\ \psi_j(x)\ |^2 = \frac{2h}{3} \quad \forall j = 1,\dots,N, \tag{2.3}$$

$$\int_0^1 \psi_{j,x}(x)\psi_{k,x}(x) = -\frac{1}{h}, \quad \text{if} \quad k = j-1, j+1, \quad \int_0^1 \psi_{j,x}(x)\psi_{k,x}(x) = 0, \quad \text{if} \quad |k-j| > 2, \tag{2.4}$$

$$\int_0^1 |\ \psi_{j,x}(x)\ |^2 = \frac{2}{h} \quad \forall\, j = 1,\dots,N, \tag{2.5}$$

the following simplest finite-element space semidiscretization of (1.1) can be easily derived:

$$\begin{cases} \dfrac{2}{3}u_j''(t) + \dfrac{1}{6}u_{j+1}''(t) + \dfrac{1}{6}u_{j-1}''(t) = \dfrac{u_{j+1}(t) + u_{j-1}(t) - 2u_j(t)}{h^2}, & 0 < t < T,\ j = 1,\dots,N, \\ u_0(t) = 0, \quad u_{N+1}(t) = v_h(t), & 0 < t < T, \\ u_j(0) = u_j^0, \quad u_j'(0) = u_j^1, & j = 1,\dots,N. \end{cases} \tag{2.6}$$

By $'$ we denote the time derivative, and by $\bar{u}(t) = (u_1(t),\dots,u_N(t))$ the vector-valued solution at the time t.

The problem of boundary controllability for (2.6) consists in finding a control $v_h \in L^2(0,T)$ such that the solution of system (2.6) satisfies

$$\bar{u}_h(T) = \bar{u}_h'(T) = \bar{0}.$$

Following the HUM method, for a given h, the existence of a control v_h for (2.6) is equivalent to the existence of an *observability constant* $C_h(T) > 0$ such that

$$E_h(t) \le C_h(T) \int_0^T \left| \frac{\phi_N(t)}{h} \right|^2 dt, \tag{2.7}$$

where $\bar{\phi}(t)$ is the solution of the homogeneous adjoint semidiscrete system

$$\begin{cases} \dfrac{2}{3}\phi_j''(t) + \dfrac{1}{6}\phi_{j+1}''(t) + \dfrac{1}{6}\phi_{j-1}''(t) = \dfrac{\phi_{j+1}(t) + \phi_{j-1}(t) - 2\phi_j(t)}{h^2}, & 0 < t < T,\ j = 1, \dots, N, \\ \phi_0(t) = \phi_{N+1}(t) = 0, & 0 < t < T, \\ \phi_j(T) = \phi_j^0, \quad \phi_j'(T) = \phi_j^1, & j = 1, \dots, N. \end{cases} \tag{2.8}$$

The energy E_h of this system is given by

$$E_h(t) = \frac{h}{6} \sum_{j=1}^{N} | \phi_j'(t) |^2 + \frac{h}{12} \sum_{j=0}^{N} | \phi_j'(t) + \phi_{j+1}'(t) |^2 + \frac{h}{2} \sum_{j=0}^{N} \left| \frac{\phi_{j+1}(t) - \phi_j(t)}{h} \right|^2. \tag{2.9}$$

Observe that the energy E_h represents a natural discretization of the continuous one (1.2).

Now, we rewrite system (2.8) in the following equivalent form:

$$\begin{cases} M_h \bar{\phi}'' = -A_h \bar{\phi}, \\ \phi_0(t) = \phi_{N+1}(t) = 0, \qquad 0 < t < T, \end{cases} \tag{2.10}$$

where $\bar{\phi} = (\phi_1, \dots, \phi_N)$, M_h and A_h are tridiagonal symmetric matrices, the so-called mass and rigidity or stiffness matrices, respectively:

$$M_h = h \begin{bmatrix} 2/3 & 1/6 & 0 & \cdots & 0 & 0 & 0 \\ 1/6 & 2/3 & 1/6 & \cdots & 0 & 0 & 0 \\ 0 & 1/6 & 2/3 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & & \\ 0 & 0 & 0 & \cdots & 2/3 & 1/6 & 0 \\ 0 & 0 & 0 & \cdots & 1/6 & 2/3 & 1/6 \\ 0 & 0 & 0 & \cdots & 0 & 1/6 & 2/3 \end{bmatrix},$$

$$A_h = \tfrac{1}{h} \begin{bmatrix} 2 & -1 & 0 & \cdots & 0 & 0 & 0 \\ -1 & 2 & -1 & \cdots & 0 & 0 & 0 \\ 0 & -1 & 2 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & & \\ 0 & 0 & 0 & \cdots & 2 & -1 & 0 \\ 0 & 0 & 0 & \cdots & -1 & 2 & -1 \\ 0 & 0 & 0 & \cdots & 0 & -1 & 2 \end{bmatrix}. \tag{2.11}$$

We denote by M and A the symmetric matrices

$$M = \frac{1}{h} M_h, \qquad A = \frac{1}{h} A_h.$$

System (2.10) is now

$$M\bar{\phi}'' = -A\bar{\phi}. \tag{2.12}$$

Let us define the Hilbert spaces sequences ℓ^2, $\hbar^1$, and $\hbar^{-1}$ as follows:

$$\ell^2 = \left\{ \{c_k\} \ : \ \|c_k\|_{\ell^2}^2 = \sum_{k\in\mathbb{N}} |\ c_k\ |^2 < \infty \right\}, \tag{2.13}$$

$$\hbar^1 = \left\{ \{c_k\} \in \ell^2 \ : \ \|c_k\|_{\hbar^1}^2 = \sum_{k\in\mathbb{N}} k^2\pi^2 \left|c_k\right|^2 < \infty \right\}, \tag{2.14}$$

$$\hbar^{-1} = \left\{ \{c_k\} \in \ell^2 \ : \ \|c_k\|_{\hbar^{-1}}^2 = \sum_{k\in\mathbb{N}} \left|\frac{c_k}{k\pi}\right|^2 < \infty \right\}. \tag{2.15}$$

DEFINITION 2.1. *Let $\delta = 1/(J+1)$, $J \in \mathbb{N}$, and $(U,V) \in \mathbb{R}^{J+1} \times \mathbb{R}^{J+1}$. We define the following scalar products:*

$$(U,V) := \sum_{j=0}^{J} U_j V_j, \quad \langle U,V\rangle := \delta\,(U,V)\,. \tag{2.16}$$

Taking into account the definitions in Fourier of the discrete norms ℓ^2 and $\hbar^1$, for every function $\bar{\phi} = (\phi_1, \ldots, \phi_N)$, $\phi_0 = 0$, we have

$$|\bar{\phi}|_{\ell^2}^2 = \langle \bar{\phi}, \bar{\phi}\rangle = h\sum_{j=0}^{N} |\phi_j|^2\,, \tag{2.17}$$

$$\|\bar{\phi}\|_{\hbar^1}^2 = \left(A_h^{1/2}\bar{\phi}, A_h^{1/2}\bar{\phi}\right) = \left\langle A^{1/2}\bar{\phi}, A^{1/2}\bar{\phi}\right\rangle = h\sum_{j=0}^{N} \left|\frac{\phi_{j+1}-\phi_j}{h}\right|^2. \tag{2.18}$$

Therefore, the energy (2.9) of the solution of the semidiscrete system (2.8) is given by

$$\begin{aligned} E_h(t) &= \frac{1}{2}\left[\left\langle M^{1/2}\bar{\phi}', M^{1/2}\bar{\phi}'\right\rangle + \left\langle A^{1/2}\bar{\phi}, A^{1/2}\bar{\phi}\right\rangle\right] \\ &= \frac{1}{2}\left[\left|M^{1/2}\bar{\phi}'\right|_{\ell^2}^2 + \left\|\bar{\phi}\right\|_{\hbar^1}^2\right], \end{aligned} \tag{2.19}$$

with the discrete norms $|\cdot|_{\ell^2}$ and $\|\cdot\|_{\hbar^1}$ given by (2.17) and (2.18), respectively.

The following result is well known.

LEMMA 2.2. *The energy E_h of the solution $\bar{\phi}$ of system* (2.8) *is conserved along the time, i.e.,*

$$E_h(t) = E_h(0) \qquad \forall t \in [0,T].$$

Proof. Multiplying equation (2.12) by $\bar{\phi}'$ and integrating in time variable, we have

$$\frac{1}{2}\frac{d}{dt}\left[\left\langle M^{1/2}\bar{\phi}', M^{1/2}\bar{\phi}'\right\rangle + \left\langle A^{1/2}\bar{\phi}, A^{1/2}\bar{\phi}\right\rangle\right] = 0.$$

This is equivalent to $d(E_h(t))/dt = 0$, and therefore $E_h(t)$ is constant in time for all $t \in [0,T]$. ∎

The eigenvalue problem corresponding to (2.8) is of the form

$$\begin{cases} -\dfrac{\varphi_{k+1}+\varphi_{k-1}-2\varphi_k}{h^2} = \mu\left[\dfrac{2}{3}\varphi_k + \dfrac{1}{6}\varphi_{k+1} + \dfrac{1}{6}\varphi_{k-1}\right], \quad k=1,\dots,N, \ (A\bar{\varphi} = \mu M\bar{\varphi}), \\ \varphi_0 = \varphi_{N+1} = 0. \end{cases} \tag{2.20}$$

The eigenvalues and eigenvectors of (2.20) may be computed explicitly. One has

$$\mu_k := \mu_k\left(M^{-1}A\right) = \frac{\lambda_k(h)}{1-\frac{h^2}{6}\lambda_k(h)} = \frac{6}{h^2}\left[\frac{\sin^2(k\pi h/2)}{\frac{3}{2}-\sin^2(k\pi h/2)}\right] = \frac{6}{h^2}\left[\frac{1-\cos(k\pi h)}{2+\cos(k\pi h)}\right] \tag{2.21}$$

and $\{\bar{\varphi}^k\}_k$, $\bar{\varphi}^k = (\varphi_{k,1},\dots,\varphi_{k,N})$, $\varphi_{k,j} = \sin(jk\pi h)$, with the eigenvalues of matrix A

$$\lambda_k(h) = \frac{4}{h^2}\sin^2\left(\frac{k\pi h}{2}\right) \tag{2.22}$$

(see [4], p. 456). The discrete eigenvalues μ_k approximate the eigenvalues of the continuous model $\lambda_k = k^2\pi^2$ for k fixed, when the size of the mesh h tends to zero and its corresponding eigenvectors $\bar{\varphi}^k$ coincide with the eigenfunctions $\varphi^k(x) = \sin(k\pi x)$ of the continuous model on the mesh points.

We have the following orthogonality properties of the eigenvectors of the matrix A in the discrete space:

$$\left\langle \bar{\varphi}^k, \bar{\varphi}^l \right\rangle := h\sum_{j=1}^{N}\varphi_{k,j}\varphi_{l,j} = \delta_{k,l}\frac{(N+1)h}{2} = \frac{1}{2}\delta_{k,l}, \quad k,l = 1,\dots,N, \tag{2.23}$$

where $\delta_{k,l}$ is *Kronecker's delta* (for more details, see [5, 6, 7, 8, 9, 10, 11, 12]). Moreover, for any eigenvectors of the matrix A with corresponding eigenvalues λ, the following identities hold (see (2.5) in [5]):

$$\sum_{j=0}^{N}\left|\frac{\varphi_{j+1}-\varphi_j}{h}\right|^2 = \lambda\sum_{j=1}^{N}|\varphi_j|^2 \tag{2.24}$$

and

$$\sum_{j=1}^{N}\left[\varphi_{k,j+1}\varphi_{l,j} + \varphi_{k,j}\varphi_{l,j+1}\right] = 0, \quad k \neq l. \tag{2.25}$$

The Fourier representation of the solution of (2.8) is as follows:

$$\bar{\phi}(t) = \sum_{k=1}^{N}\left(a_k\sin(\sqrt{\mu_k}t) + b_k\cos(\sqrt{\mu_k}t)\right)\bar{\varphi}^k = \sum_{k=-N}^{N} m_k e^{i\sqrt{\mu_k}t}\bar{\varphi}^k \tag{2.26}$$

for suitable Fourier coefficients a_k, b_k, and m_k depending on the initial data.

The *observability inequality* (2.7) is said to be *uniform* if the constants $C_h(T)$ are bounded uniformly in h as $h \to 0$. However, in [5] it was proved that, for all $T > 0$, the best constant C_h in (2.7) blows-up, i.e., $C_h \to \infty$ as $h \to 0$. As a consequence of this fact, there exist initial data for the wave equation (1.1) such that the controls v_h

of (2.6) are unbounded in $L^2(0,T)$ as $h \to 0$. Thus, computing v_h is not necessarily an efficient way to approximate the control of (1.1).

This nonuniformity turns out to be due to the highly oscillating components of the solutions. In fact, a positive result can be obtained when filtering appropriately the eigenvector components of the solutions. Indeed, as is pointed in [5], we introduce the class $\mathcal{C}_\delta(h)$ of filtered solutions

$$\mathcal{C}_\delta(h) = \left\{ \bar{\phi} \text{ solution of (2.8)} : \bar{\phi} = \sum_{|\mu_k| h \le \sqrt{\delta}} a_k e^{i\sqrt{\mu_k}t} \bar{\varphi}_k \right\}, \tag{2.27}$$

where $\mu_{-k} = -\mu_k$, $k > 0$, and $0 < \delta < 12$. The class $\mathcal{C}_\delta(h)$ consists only of solutions of (2.8) generated by the eigenvectors associated to eigenvalues $\mu_k \le \delta/h^2$, $0 < \delta < 12$, or equivalently, the solutions of the class $\mathcal{C}_\delta(h)$ contain only the eigenvectors corresponding to the $N\varepsilon(\delta)$ smallest eigenvalues for a suitable $0 < \varepsilon(\delta) < 1$. When restricting the solutions to this class, there exists $T(\gamma) > 0$

$$T(\gamma) = \frac{2\sqrt{1 + \dfrac{\gamma}{16}\left(\dfrac{3}{\lambda_1} - \dfrac{h^2}{3}\right)}}{1 - \gamma/12}$$

such that the uniform observability inequality (2.7) holds for all $T > T(\delta)$, with a constant $C(T,\delta)$ independent of h. Moreover, $T(\gamma) \nearrow \infty$ as $\gamma \nearrow 12$, and $T(\gamma) \searrow 2$ as $\gamma \searrow 0$ (see [5]).

Another possible cure is to use the two-grid algorithm introduced by Glowinski [2] to avoid the high-frequency pathologies producing the blow-up phenomena described above. In order to restore the uniformity of the observability inequality with respect to h, we consider and study the convergence of this algorithm.

Our aim is to obtain an uniform observability inequality of type (2.7) for a subclass of solutions $\bar{\phi}$ of (2.8) without filtering.

As we said above, this method is closely related to the presented filtering method. In fact, here we substitute the subclass $\mathcal{C}_\delta(h)$ of solutions for another one where we consider only the slowly oscillating initial data coming from a *coarse grid* G_C^{2h}, with mesh size $2h$. Recall that the multigrid method's advantage consists in that it is possible to work directly and completely in the physical space without the necessity of using the Fourier transform and its inverse, permanently, opposite the filtering method.

3. A two-grid method. Let $N \in \mathbb{N}$ be an odd number. With $h = 1/(N+1)$ we introduce the *coarse grid* G_C^{2h}, an equidistant partition of the interval $(0,1)$, $x_0 = 0 < x_1 = 2h < \cdots < x_{(N+1)/2} = 1$, with $x_j = 2jh$, $j = 0, \ldots, (N+1)/2$ and the *fine grid* G_F^h, $y_0 = 0 < y_1 = h < \cdots < y_N = Nh < y_{N+1} = 1$, with $y_j = jh$, $j = 0, \ldots, N+1$.

Let us introduce the space V_h:

$$V_h = \Bigg\{ \bar{\phi}_h = (\phi_1, \ldots, \phi_N) \in \mathbb{R}^N : \phi_{2j+1} = \frac{\phi_{2j} + \phi_{2j+2}}{2}, \ j = 0, \ldots, \frac{N-1}{2}, \ \phi_0 = \phi_{N+1} = 0 \Bigg\}. \tag{3.1}$$

Note that V_h is constituted by discrete functions defined on the fine grid G_F^h obtained by the interpolation of discrete functions defined on the coarse one G_C^{2h}. Thus, V_h contains only slowly oscillatory functions.

Next, we shall study some general properties of the discrete functions of V_h and the restriction of the matrix A as in (2.11) to this subspace.

PROPOSITION 3.1. *For every function $\bar{\phi}_h \in V_h$, we have*

$$\left(A\bar{\phi}_h\right)_{2j+1} = 0, \qquad j = 0, \dots, \frac{N-1}{2}. \tag{3.2}$$

Proof. By Definition (3.1) of the space V_h, for any function $\bar{\phi}_h \in V_h$, we have

$$\phi_{2j+1} = \frac{\phi_{2j+2} + \phi_{2j}}{2}, \qquad j = 0, \dots, \frac{N-1}{2}, \tag{3.3}$$

which is equivalent to

$$2\phi_{2j+1} - \phi_{2j+2} - \phi_{2j} = 0, \qquad j = 0, \dots, \frac{N-1}{2}.$$

This last identity is exactly (3.2), since

$$\left(A\bar{\phi}_h\right)_j = \frac{2\phi_j - \phi_{j-1} - \phi_{j+1}}{h^2} \qquad \forall j = 1, \dots, N. \tag{3.4}$$

□

DEFINITION 3.2. *Let $\bar{U} \in G_F^h$, $\bar{U} = (U_1, \dots, U_N)$ be a discrete function defined on the fine grid. We define the following orthogonal projection $\Pi_{odd} : G_F^h \to G_F^h$, with*

$$\Pi_{odd}(\bar{U}) = (U_1, 0, \dots, U_{N-1}, 0).$$

Remark 3.3. Note that, as consequence of Proposition 3.1, if $\bar{\phi}_h \in V_h$, one has

$$\Pi_{odd}(A\bar{\phi}_h) = \bar{0}, \tag{3.5}$$

and, that is, $A\bar{\phi}_h$ is orthogonal in the scalar product (2.16) to all of the vectors $v \in \mathbb{R}^N$ for which all of the components corresponding to the even points of the fine grid G_F^h vanish (these are precisely the points of the coarse grid, G_C^{2h}).

Let us denote by $c_k \in \mathbb{R}$ the coefficients of $\bar{\phi}_h \in G_F^h$ in the orthogonal bases $\{\bar{\varphi}^k\}$ of the eigenvectors of the system (2.20), with $\bar{\varphi}^k = (\varphi_{k,1}, \dots, \varphi_{k,N})$, $\varphi_{k,j} = \sin(jk\pi h)$. That is,

$$\bar{\phi}_h = \sum_{k=1}^{N} c_k \bar{\varphi}^k. \tag{3.6}$$

The following holds.

PROPOSITION 3.4. *For every discrete function $\bar{\phi}_h \in V_h$, the coefficients $(c_k)_k$ satisfy*

$$c_{N+1-k} = -c_k \frac{\lambda_k(h)}{\lambda_{N+1-k}(h)}, \qquad k = 1, \dots, N, \tag{3.7}$$

$$c_{\frac{N+1}{2}} = 0, \tag{3.8}$$

where $\lambda_k(h)$ are the eigenvalues given by (2.22) of the symmetric matrix A.

Proof. Recall that $\varphi_{k,j} = \sin(jk\pi h)$. We have

$$\varphi_{N+1-k,j} = \sin\left((N+1-k)\pi jh\right) = (-1)^{j+1} \sin(jk\pi h), \tag{3.9}$$

and then the eigenvectors $\bar{\varphi}^k$ satisfy the symmetry property:

$$\varphi_{k,j} = (-1)^{j+1}\varphi_{N+1-k,j} \quad \forall\, k, j = 1, \dots, N. \tag{3.10}$$

We have

$$\begin{aligned}\left\langle A\bar{\phi}_h, \bar{\varphi}^l \right\rangle &:= h \sum_{j=1}^{N} \left(A\bar{\phi}_h\right)_j \varphi_{l,j} \\ &= \sum_{k=1}^{N} c_k \lambda_k(h) \left\langle \bar{\varphi}^k, \bar{\varphi}^l \right\rangle = h \sum_{j=1}^{N}\sum_{k=1}^{N} c_k \lambda_k(h) \varphi_{k,j}\varphi_{l,j},\end{aligned} \tag{3.11}$$

with the scalar product $\langle \cdot, \cdot \rangle$ defined in (2.16) (recall that $\lambda_k(h)$ are the eigenvalues of the symmetric matrix A, i.e., $A\bar{\varphi}_k = \lambda_k(h)\bar{\varphi}_k$). By Proposition 3.1, relation (3.11) is reduced to

$$\left\langle A\bar{\phi}_h, \bar{\varphi}^l \right\rangle = h \sum_{j=1}^{\frac{N-1}{2}} \sum_{k=1}^{N} c_k \lambda_k(h) \varphi_{k,2j}\varphi_{l,2j}, \qquad l = 1, \dots, N. \tag{3.12}$$

We have

$$\begin{aligned}\left\langle A\bar{\phi}_h, \bar{\varphi}^l \right\rangle &= \sum_{k=1}^{N} c_k \lambda_k(h) \left\langle \bar{\varphi}^k, \bar{\varphi}^l \right\rangle = \sum_{k=1}^{N} c_{N+1-k} \lambda_{N+1-k}(h) \left\langle \bar{\varphi}^{N+1-k}, \bar{\varphi}^l \right\rangle \\ &= h \sum_{j=1}^{N}\sum_{k=1}^{N} c_{N+1-k} \lambda_{N+1-k}(h) \varphi_{N+1-k,j}\varphi_{l,j}.\end{aligned} \tag{3.13}$$

Moreover, due to the symmetry property (3.10), it follows that

$$\begin{aligned}\left\langle A\bar{\phi}_h, \bar{\varphi}^l \right\rangle &= h \sum_{j=1}^{N}\sum_{k=1}^{N} c_{N+1-k} \lambda_{N+1-k}(h) (-1)^{j+1} \varphi_{k,j}\varphi_{l,j} \\ &= h \sum_{j=1}^{\frac{N-1}{2}}\sum_{k=1}^{N} c_{N+1-k} \lambda_{N+1-k}(h) (-1)^{2j+1} \varphi_{k,2j}\varphi_{l,2j} \\ &= -\sum_{k=1}^{N} c_{N+1-k} \lambda_{N+1-k}(h) \left(\langle \bar{\varphi}_k, \bar{\varphi}_l \rangle\right)_{2j} \\ &= -h \sum_{j=1}^{\frac{N-1}{2}}\sum_{k=1}^{N} c_{N+1-k} \lambda_{N+1-k}(h) \varphi_{k,2j}\varphi_{l,2j}, \\ &\qquad\qquad l = 1, \dots, N, \;\; j = 1, \dots, (N-1)/2.\end{aligned} \tag{3.14}$$

Finally, by (3.12) and (3.14), we have

$$h \sum_{j=1}^{\frac{N-1}{2}} \sum_{k=1}^{N} c_k \lambda_k(h) \varphi_{k,2j}\varphi_{l,2j} = -h \sum_{j=1}^{\frac{N-1}{2}} \sum_{k=1}^{N} c_{N+1-k} \lambda_{N+1-k}(h) \varphi_{k,2j}\varphi_{l,2j}.$$

Taking into account the orthogonality properties (2.23) of the eigenvectors of the matrix A, it is immediate to see that the identities (3.7) and (3.8) hold. This concludes the proof. ∎

We consider the following decomposition in low and high frequencies of an arbitrary function $\bar{\phi}_h \in G_F^h$:

$$\bar{\phi}_h = \bar{\phi}_L + \bar{\phi}_H, \quad \bar{\phi}_L = \sum_{k=1}^{\frac{N-1}{2}} c_k \bar{\varphi}^k, \quad \bar{\phi}_H = \sum_{\frac{N+1}{2}}^{N} c_k \bar{\varphi}^k. \tag{3.15}$$

THEOREM 3.5. *For any discrete function $\bar{\phi}_h \in V_h$, we have*

$$\bar{\phi}_h = \bar{\phi}_L + \bar{\phi}_H, \quad \bar{\phi}_L = \sum_{k=1}^{\frac{N-1}{2}} c_k \bar{\varphi}^k, \quad \bar{\phi}_H = -\sum_{k=1}^{\frac{N-1}{2}} c_k \tan^2\left(\frac{k\pi h}{2}\right) \bar{\varphi}^{N+1-k}. \tag{3.16}$$

Proof. Taking into account that

$$\frac{\lambda_k(h)}{\lambda_{N+1-k}(h)} = \frac{\sin^2\left(\frac{k\pi h}{2}\right)}{\sin^2\left(\frac{(N+1-k)\pi h}{2}\right)} = \tan^2\left(\frac{k\pi h}{2}\right),$$

the proof is an immediate consequence of Proposition 3.4 and relations (3.15). ∎

We also deduce the following additional properties.

PROPOSITION 3.6. *For every function $\bar{\phi}_h \in V_h$ as in* (3.6)*, with $\bar{\phi}_h = \bar{\phi}_L + \bar{\phi}_H$ given by* (3.16)*, the following inequalities hold:*

$$\left\|\bar{\phi}_L\right\|_{\hbar^1}^2 \geq \left\|\bar{\phi}_H\right\|_{\hbar^1}^2, \tag{3.17}$$

$$\left|A\bar{\phi}_L\right|_{\ell^2}^2 \geq \left|A\bar{\phi}_H\right|_{\ell^2}^2, \tag{3.18}$$

where $|\cdot|_{\ell^2}$ and $\|\cdot\|_{\hbar^1}$ are the ℓ^2 and $\hbar^1$ discrete norms defined by (2.17) *and* (2.18)*, respectively.*

Proof. By the orthogonality property (2.23) and relation (2.24) of the eigenvectors $\{\bar{\varphi}^k\}_k$, for every function $\bar{\phi}_h \in \mathbb{R}^N$, we have

$$\begin{aligned} \left|A\bar{\phi}_h\right|_{\ell^2}^2 &= \left\langle A\bar{\phi}_h, A\bar{\phi}_h\right\rangle = \sum_{k,l=1}^{N} \lambda_k(h)\lambda_l(h) c_k c_l \left\langle \bar{\varphi}^k, \bar{\varphi}^l\right\rangle \\ &= h\sum_{j=1}^{N}\left|\sum_{k=1}^{N} c_k \lambda_k(h)\bar{\varphi}^k\right|^2 = \frac{1}{2}\sum_{k=1}^{N}\lambda_k^2(h)c_k^2, \end{aligned} \tag{3.19}$$

$$\begin{aligned} \left\|\bar{\phi}_h\right\|_{\hbar^1}^2 &= \left\langle A^{1/2}\bar{\phi}_h, A^{1/2}\bar{\phi}_h\right\rangle = \sum_{k,l=1}^{N} \sqrt{\lambda_k(h)\lambda_l(h)} c_k c_l \left\langle \bar{\varphi}^k, \bar{\varphi}^l\right\rangle \\ &= h\sum_{j=1}^{N}\left|\sum_{k=1}^{N} \sqrt{\lambda_k(h)} c_k \bar{\varphi}^k\right|^2 = \frac{1}{2}\sum_{k=1}^{N}\lambda_k(h)c_k^2. \end{aligned} \tag{3.20}$$

For the component $\bar{\phi}_L$, corresponding to the low frequencies, we obtain

$$\left\|\bar{\phi}_L\right\|_{\hbar^1}^2 = \frac{1}{2}\sum_{k=1}^{\frac{N-1}{2}}\lambda_k(h)c_k^2, \tag{3.21}$$

$$\left|A\bar{\phi}_L\right|_{\ell^2}^2 = \frac{1}{2}\sum_{k=1}^{\frac{N-1}{2}}\lambda_k^2(h)c_k^2, \tag{3.22}$$

and for the component $\bar{\phi}_H$, corresponding to the high frequencies, taking into account Proposition 3.4, we have

$$\|\bar{\phi}_H\|^2_{\hbar^1} = \frac{1}{2}\sum_{k=\frac{N+1}{2}}^{N-1}\left|\frac{\lambda_k(h)}{\lambda_{N+1-k}(h)}\right|^2 c_k^2\lambda_k(h) = \frac{1}{2}\sum_{k=\frac{N+1}{2}}^{N-1}\left|\tan^2\left(\frac{k\pi h}{2}\right)\right|^2 \lambda_k(h)c_k^2, \tag{3.23}$$

$$\left|A\bar{\phi}_H\right|^2_{\ell^2} = \frac{1}{2}\sum_{k=1}^{\frac{N-1}{2}}\left|\frac{\lambda_k(h)}{\lambda_{N+1-k}(h)}\right|^2 \lambda_k^2(h)c_k^2 = \frac{1}{2}\sum_{k=1}^{\frac{N-1}{2}}\left|\tan^2\left(\frac{k\pi h}{2}\right)\right|^2 \lambda_k^2(h)c_k^2. \tag{3.24}$$

Relations (3.17) and (3.18) are obtained immediately taking into account that

$$\frac{\lambda_k^2(h)}{\lambda_{N+1-k}^2(h)} = \frac{\sin^4\left(\frac{k\pi h}{2}\right)}{\sin^4\left(\frac{(N+1-k)\pi h}{2}\right)} = \tan^4\left(\frac{k\pi h}{2}\right) \le 1 \quad \forall\, k = 1,\dots,\frac{N-1}{2}. \tag{3.25}$$

Then, we have (3.17) and (3.18). The proof of Proposition 3.6 is now complete. ∎

4. Two-grid finite-element semidiscrete scheme. In this section we shall see that the two-grid technique for the finite-element semidiscretization of the wave equation allows us to eliminate the highly oscillating components of the control without filtering and to guarantee its convergence. This method is closely related to the filtering mentioned above.

We are able to get a uniform observability inequality for every solution of the wave equation (2.8) defined on the fine grid G_F^h, with initial data in V_h.

A very direct way to eliminate the high frequencies is as follows: take any initial data $(\bar{\phi}^0, \bar{\phi}^1)$ in $V_h \times V_h$ and denote by $\bar{\phi}$ the corresponding solution of system (2.8). We develop the functions $(\bar{\phi}^0, \bar{\phi}^1)$ in high and low frequencies as in (3.15):

$$\bar{\phi}^0 = \bar{\phi}_L^0 + \bar{\phi}_H^0, \qquad \bar{\phi}^1 = \bar{\phi}_L^1 + \bar{\phi}_H^1.$$

First we prove some basic identities.

PROPOSITION 4.1. *Let $\bar{\phi} = \bar{\phi}_L + \bar{\phi}_H$ be the solution of* (2.8)*, with given data $(\bar{\phi}^0, \bar{\phi}^1)$ in time $t = T$ contained in $V_h \times V_h$. For every $t \in [0,T]$, we have*

$$E_{\bar{\phi}}(t) = E_{\bar{\phi}_L}(t) + E_{\bar{\phi}_H}(t) \qquad \text{and} \qquad E_{\bar{\phi}_L}(t) \ge E_{\bar{\phi}_H}(t). \tag{4.1}$$

Proof. The first identity is immediate due to the orthogonality properties of eigenvectors.

Taking into account that $\bar{\phi}_L$ and $\bar{\phi}_H$ are solutions of the semidiscrete wave equation (2.8) in the fine grid G_F^h, where $\bar{\phi} = \bar{\phi}_L + \bar{\phi}_H$ is the development (3.15) of the solution $\bar{\phi}$ in low and high frequencies, the corresponding energies $E_{\bar{\phi}_L}$ and $E_{\bar{\phi}_\hbar}$ are conserved in time, and therefore, to obtain (4.1) for every $t \in [0,T]$, it is sufficient to prove it for $t = T$. That is, we prove

$$\left|M^{1/2}\bar{\phi}_L^1\right|^2_{\ell^2} + \left\|\bar{\phi}_L^0\right\|^2_{\hbar^1} \ge \left|M^{1/2}\bar{\phi}_H^1\right|^2_{\ell^2} + \left\|\bar{\phi}_H^0\right\|^2_{\hbar^1}. \tag{4.2}$$

We have

$$\begin{aligned}\frac{1}{2}|M^{1/2}\bar{\phi}^1|^2_{\ell^2} &= \frac{h}{6}\sum_{j=1}^{N}|\phi_j^1|^2 + \frac{h}{12}\sum_{j=0}^{N}|\phi_j^1 + \phi_{j+1}^1\,|^2 \\ &= \frac{1}{6}|\bar{\phi}^1|^2_{\ell^2} + \frac{h}{12}\sum_{j=0}^{N}|\phi_j^1 + \phi_{j+1}^1\,|^2\,.\end{aligned} \tag{4.3}$$

By the orthogonality properties (2.23), (2.24), and (2.25) of the eigenvectors $\{\bar{\varphi}^k\}_k$ and taking into account that $(\bar{\phi}^0, \bar{\phi}^1) \in V_h \times V_h$, combining (3.17), (4.3), and the property of the conservation of the energy, we deduce that (4.2) holds. The proof of Proposition 4.1 is now complete. ∎

4.1. Uniform observability. This section is devoted to the proof of the uniform observability property, one of the main results of this paper. Our first result guarantees that the observability inequality for (2.8) is uniform within the class $V_h \times V_h$.

THEOREM 4.2. *Let $h > 0$ and $T > 4$. Then, there exist positive constants $C_j(T,h) > 0$, $j = 1, 2$, uniformly bounded in h such that*

$$C_1(T,h)\int_0^T \left[\left|\frac{\phi_N(t)}{h}\right|^2 + \frac{|\phi_N'(t)|^2}{6}\right] dt \le E_h(0) \le C_2(T,h)\int_0^T \left[\left|\frac{\phi_N(t)}{h}\right|^2 + \frac{|\phi_N'(t)|^2}{6}\right] dt \tag{4.4}$$

for the solution of (2.8)*, with initial data in the space $V_h \times V_h$ and all $h > 0$.*

Remark 4.3. An extra term $\int_0^T |\phi_N'(t)|^2/6dt$ appears in the observability inequality (4.4). This is not the case in the context of the finite-difference semidiscretization (see [5, 9, 10]). Filtering the high frequencies by truncating the Fourier series of the solution and using Ingham's inequality, this extra term can be absorbed by the term $\int_0^T |\phi_N(t)|^2 dt$ by increasing the observability constant $C_1(T)$ (see [5]).

We shall see that the uniform inequality (4.4) is sufficient to guarantee the convergence of the sequence of discrete controls v_h.

Proof. (Sketch of the proof of Theorem 4.2.) The proof of this theorem uses discrete multiplier techniques similar to those in [5]. For simplicity, we divide the proof into several steps.

Step 1. Multiplying (2.8) by $j\,(\phi_{j+1} - \phi_{j-1}/2)$, which is a discrete version of the discrete multiplier $x\phi_x$ for the wave equation (1.4), integrating in time $[0,T]$, and adding for $j = 1, \dots, N$, the following identity holds for all the solutions of system (2.8) (see formula (3.26) in [5]):

$$TE_h(0) + X_h(t)\Big|_0^T = \frac{1}{2}\int_0^T \left[\left|\frac{\phi_N(t)}{h}\right|^2 + \frac{|\phi_N'(t)|^2}{6}\right] dt + \frac{h}{12}\sum_{j=1}^N \int_0^T \left|\phi_j'(t) - \phi_{j+1}'(t)\right|^2, \tag{4.5}$$

with

$$X_h(t) = h\sum_{j=1}^N j(\phi_{j+1} - \phi_{j-1})\left(\frac{1}{3}\phi_j' + \frac{1}{12}\phi_{j+1}' + \frac{1}{12}\phi_{j-1}'\right). \tag{4.6}$$

Remark 4.4. Observe that this discrete identity (4.5) is very close to the well-known identity for the $1-d$ wave equation (1.4) (see [6]):

$$TE(0) + X(t)\Big|_0^T = \frac{1}{2}\int_0^T |\phi_x(1,t)|^2 dt, \tag{4.7}$$

with

$$X(t) = \int_0^1 x\phi_x\phi_t dx.$$

Note, however, that in (4.5) in addition to the discrete analogue of the terms entering in the identity (4.7), we also have the extra term

$$\frac{h}{12}\sum_{j=1}^{N}\int_0^T \left|\phi_j'(t)-\phi_{j+1}'(t)\right|^2. \tag{4.8}$$

As indicated in [5], when estimating this term one loses completely the observability inequality if the class of solutions under consideration is not restricted and precisely filtering the high frequencies makes possible its boundary with respect to h.

Step 2. We have the following.

PROPOSITION 4.5 (Equipartition of energy). *For any $h>0$ and $\bar\phi$ solution of system* (2.8)*, the following identity holds:*

$$\frac{h}{12}\int_0^T\sum_{j=0}^{N}\left|\frac{\phi_{j+1}(t)-\phi_j(t)}{h}\right|^2=\frac{1}{12}TE_h-\frac{1}{24}Y_h(t)\Big|_0^T, \tag{4.9}$$

or

$$\begin{aligned} h\int_0^T\sum_{j=0}^{N}&\left|\frac{\phi_{j+1}(t)-\phi_j(t)}{h}\right|^2 \\ &=h\int_0^T\sum_{j=0}^{N}\left[\frac{1}{3}|\phi_j'(t)|^2+\frac{1}{6}|\phi_{j+1}'(t)+\phi_j'(t)|^2\right]dt-Y_h(t)\Big|_0^T, \end{aligned} \tag{4.10}$$

where

$$Y_h(t)=h\sum_{j=1}^{N}\left(\frac{2}{3}\phi_j'+\frac{1}{6}\phi_{j+1}'+\frac{1}{6}\phi_{j-1}'\right)\phi_j. \tag{4.11}$$

Proof. This identity is verified by every solution $\bar\phi$ of system (2.8) (see the formula (3.33) in [5]). It is obtained by multiplying equation (2.8) by $\bar\phi$, taking into account the expression of the energy (2.19). ∎

Taking into account that the equipartition of energy (4.9) holds for every solution $\bar\phi$ of (2.8), we have the following identity for the time derivative $\bar\phi'$ of $\bar\phi$:

$$\frac{h}{12}\int_0^T\sum_{j=0}^{N}\left|\phi_{j+1}'(t)-\phi_j'(t)\right|^2=\frac{h^2}{12}TE_{\bar\phi'}-\frac{h^2}{24}\tilde Y_h(t)\Big|_0^T, \tag{4.12}$$

where

$$\tilde Y_h(t)=h\sum_{j=1}^{N}\left(\frac{2}{3}\phi_j''+\frac{1}{6}\phi_{j+1}''+\frac{1}{6}\phi_{j-1}''\right)\phi_j' \tag{4.13}$$

and $E_{\bar\phi'}$ is the conservative energy of the derivative $\bar\phi'$ of the solution $\bar\phi$.

Step 3. (Some key estimations). The following holds.

PROPOSITION 4.6. *For any $h>0$ and any solution $\bar\phi$ of* (2.8)*, with initial data in the space $V_h\times V_h$, the following relation holds:*

$$\frac{h^2}{12}E_{\bar\phi'}\le\frac{1}{2}E_{\bar\phi_L}\le\frac{1}{2}E_h(t), \tag{4.14}$$

where $E_{\bar\phi_L}$ and $E_h(t)$ are the conservative energies (4.1) *and* (2.9)*, respectively.*

Proof. The time derivative of every solution ϕ of (2.8) is a solution of (2.8), too, with given state $\bar{\phi}'(T) = \bar{\phi}^1$ and $\bar{\phi}''(T) = M^{-1}A\bar{\phi}^0$, A and M being as in (2.11). In virtue of Proposition 4.1, we have

$$\frac{h^2}{12}E_{\bar{\phi}'} = \frac{h^2}{12}\left(E_{\bar{\phi}'_L} + E_{\bar{\phi}'_H}\right) \leq \frac{h^2}{6}E_{\bar{\phi}'_L} \leq \frac{h^2}{6}\Lambda E_{\bar{\phi}_L}, \tag{4.15}$$

where Λ is the largest eigenvalue entering in the Fourier development of the solution $\bar{\phi}_L$ of (2.8), corresponding to the low frequencies, i.e.,

$$\bar{\phi}_L(t) = \sum_{k=1}^{\frac{N-1}{2}} c_k \bar{\varphi}^k. \tag{4.16}$$

That is,

$$\Lambda = \max_{1\leq k\leq (N-1)/2} \mu_k(M^{-1}A) = \max_{1\leq k\leq (N-1)/2} \frac{6}{h^2}\left[\frac{\sin^2\left(\frac{k\pi h}{2}\right)}{\frac{3}{2} - \sin^2\left(\frac{k\pi h}{2}\right)}\right] = \frac{3}{h^2}. \tag{4.17}$$

Thus, using estimation (4.17) of Λ in (4.15), we find exactly inequality (4.14), and the proof is complete. □

Step 4. Now we are able to estimate the extra term (4.8) entering in the identity (4.7).

In view of (4.14) in (4.12), we obtain the following sharp estimate for the solutions of (2.8) in the class $V_h \times V_h$ of slowly varying initial data:

$$\frac{h}{12}\int_0^T \sum_{j=0}^{N} \left|\phi'_{j+1}(t) - \phi'_j(t)\right|^2 \leq \frac{T}{2}E_h(t) - \frac{h^2}{24}\tilde{Y}_h(t)\Big|_0^T. \tag{4.18}$$

In view of this relation, combining (4.5), (4.9), and (4.18) we deduce that

$$\frac{T}{2}E_h(0) + Z_h(t)\Big|_0^T \leq \frac{1}{2}\int_0^T \left[\left|\frac{\phi_N(t)}{h}\right|^2 + \frac{|\phi'_N(t)|^2}{6}\right] dt, \tag{4.19}$$

with

$$Z_h(t) = X_h(t) + \frac{h^2}{24}\tilde{Y}_h(t). \tag{4.20}$$

Recall that the Fourier development of the solution $\bar{\phi}$ is

$$\bar{\phi}(t) = \sum_{|k|\leq N} a_k e^{i\mu_k t}\bar{\varphi}^k, \tag{4.21}$$

with $\mu_k = \sqrt{\mu_k}$ for $k \geq 1$ and $\mu_{-k} = -\mu_k$. Hence, by (4.13),

$$\begin{aligned}\frac{h^2}{24}\tilde{Y}_h(t) &= \frac{h^3}{24}\sum_{j=1}^{N}\left(\frac{2}{3}\phi''_j + \frac{1}{6}\phi''_{j+1} + \frac{1}{6}\phi''_{j-1}\right)\phi'_j \\ &\asymp \frac{h}{2}\sum_{j=1}^{N}\left(\frac{2}{3}\phi'_j + \frac{1}{6}\phi'_{j+1} + \frac{1}{6}\phi'_{j-1}\right)\phi_j = \frac{1}{2}Y_h(t).\end{aligned} \tag{4.22}$$

Equation (4.22) is equivalent to the existence of positive constants K_1, K_2, such that

$$K_1 Y_h(t) \leq \frac{h^2}{24}\tilde{Y}_h(t) \leq K_2 Y_h(t). \tag{4.23}$$

The following provides an estimate on the reminder term Z_h given by (4.20).

PROPOSITION 4.7. *For any $h > 0$ and any solution $\bar{\phi}$ of (2.8), it follows that*

$$|Z_h(t)| \leq \sqrt{1 + \frac{Kh^2}{4} + \frac{3h}{4\mu_1}(-2K + K^2)}E_h(0) \;\; \forall\, 0 < t < T, \tag{4.24}$$

where K is a positive constant.

Proof. Taking into account identity (4.20), with X_h and Y_h given by (4.6) and (4.11), respectively, we obtain

$$\begin{aligned} Z_h(t) &\asymp h\sum_{j=1}^{N} j\,(\phi_{j+1} - \phi_{j-1})\left(\frac{1}{3}\phi'_j + \frac{1}{12}\phi'_{j-1} + \frac{1}{12}\phi'_{j+1}\right) \\ &\quad + \frac{Kh}{2}\sum_{j=1}^{N}\left(\frac{2}{3}\phi'_j + \frac{1}{6}\phi'_{j+1} + \frac{1}{6}\phi'_{j-1}\right)\phi_j \\ &= h\sum_{j=1}^{N}\left[\frac{2}{3}\phi'_j + \frac{1}{6}\phi'_{j+1} + \frac{1}{6}\phi'_{j-1}\right]\left(\frac{j}{2}(\phi_{j+1} - \phi_{j-1}) + \frac{K}{2}\phi_j\right) \\ &= h\sum_{i,j=1}^{N} m_{ij}a_i b_j = \left\langle M\bar{a}, \bar{b}\right\rangle, \end{aligned} \tag{4.25}$$

with

$$a_i = \frac{i}{2}(\phi_{i+1} - \phi_{i-1}) + \frac{K}{2}\phi_i; \;\; b_i = \phi'_i$$

and m_{ij} are the entries of matrix M. Therefore

$$|Z_h(t)| \leq \left|h\sum_{i,j=1}^{N} m_{ij}a_i a_j\right|^{1/2}\left|h\sum_{i,j=1}^{N} m_{ij}b_i b_j\right|^{1/2} = |M^{1/2}\bar{a}|_{\ell^2}\cdot|M^{1/2}\bar{b}|_{\ell^2}. \tag{4.26}$$

On the other hand,

$$\begin{aligned} |M^{1/2}\bar{b}|_{\ell^2} &= h\sum_{j=1}^{N}\left[\frac{2}{3}\mid\phi'_j\mid^2 + \frac{1}{6}(\phi'_{j+1} + \phi'_{j-1})\phi'_j\right] \\ &= h\sum_{j=1}^{N}\left[\frac{1}{3}\mid\phi'_j\mid^2 + \frac{1}{6}\left|\phi'_{j+1} + \phi'_j\right|^2\right] \end{aligned} \tag{4.27}$$

and

(4.28)

$$
\begin{aligned}
|M^{1/2}\bar{a}|_{\ell^2} &\le h\sum_{i=1}^{N} a_i^2 = \frac{h}{4}\sum_{i=1}^{N}[i(\phi_{i+1}-\phi_{i-1})+K\phi_i]^2\\
&= \frac{h}{4}\sum_{i=1}^{N}\big[i^2(\phi_{i+1}-\phi_i)^2+i^2(\phi_i-\phi_{i-1})^2+2i^2(\phi_{i+1}-\phi_i)(\phi_i-\phi_{i-1})\\
&\quad + K^2\phi_i^2+2Ki\phi_{i+1}\phi_i-2Ki\phi_{i-1}\phi_i\big]\\
&\le \frac{h}{4}\sum_{i=1}^{N}\big[2i^2\mid\phi_{i+1}-\phi_i\mid^2+2i^2\mid\phi_i-\phi_{i-1}\mid^2\\
&\quad + 2Ki\phi_i(\phi_{i+1}-\phi_{i-1})+K^2\mid\phi_i\mid^2\big]\\
&\le h(N+1)^2\sum_{i=0}^{N}|\phi_{i+1}-\phi_i|^2+\frac{Kh}{2}\sum_{i=1}^{N}i\phi_i\phi_{i+1}\\
&\quad -\frac{Kh}{2}\sum_{i=1}^{N}i\phi_{i-1}\phi_i+\frac{K^2h}{4}\sum_{i=0}^{N}|\phi_i|^2\\
&= h\sum_{i=0}^{N}\left|\frac{\phi_{i+1}-\phi_i}{h}\right|^2+\frac{Kh}{2}\sum_{i=0}^{N}[i\phi_i\phi_{i+1}-(i+1)\phi_i\phi_{i+1}]+\frac{K^2h}{4}\sum_{i=0}^{N}|\phi_i|^2\\
&= h\sum_{i=0}^{N}\left|\frac{\phi_{i+1}-\phi_i}{h}\right|^2-\frac{Kh}{2}\sum_{i=0}^{N}\phi_i\phi_{i+1}+\frac{K^2h}{4}\sum_{i=0}^{N}|\phi_i|^2\\
&= h\sum_{i=0}^{N}\left|\frac{\phi_{i+1}-\phi_i}{h}\right|^2-\frac{Kh^3}{4}\sum_{i=0}^{N}\left|\frac{\phi_{i+1}-\phi_i}{h}\right|^2+\left(-\frac{Kh}{2}+\frac{K^2h}{4}\right)\sum_{i=0}^{N}|\phi_i|^2\\
&\le \left(1+\frac{Kh^2}{4}+\frac{3h}{4\mu_1}(-2K+K^2)\right)h\sum_{i=0}^{N}\left|\frac{\phi_{i+1}-\phi_i}{h}\right|^2.
\end{aligned}
$$

To obtain the last relation, we have used the inequality

$$
\sum_{j=0}^{N}\left|\frac{\phi_{j+1}(t)-\phi_j(t)}{h}\right|^2 \ge \frac{\mu_1}{3}\sum_{j=0}^{N}\mid\phi_j(t)\mid^2, \tag{4.29}
$$

which is a consequence of the following identity (see [5] for a detailed proof) which holds for any eigenvectors φ, with eigenvalues μ of (2.20):

$$
\sum_{j=0}^{N}\left|\frac{\varphi_{j+1}-\varphi_j}{h}\right|^2=\mu\sum_{j=0}^{N}\left[\frac{1}{3}|\varphi_j|^2+\frac{1}{6}|\varphi_j+\varphi_{j+1}|^2\right]. \tag{4.30}
$$

We deduce that

$$
\begin{aligned}
|Z_h(t)| &\le \left[h\sum_{j=0}^{N}\left(\frac{1}{3}\mid\phi_j'\mid^2+\frac{1}{6}\mid\phi_j'+\phi_{j+1}'\mid^2\right)\right]^{1/2}\\
&\quad \left[\left(1+\frac{Kh^2}{4}+\frac{3h}{4\mu_1}\left(-2K+K^2\right)\right)h\sum_{i=0}^{N}\left|\frac{\phi_{i+1}-\phi_i}{h}\right|^2\right]^{1/2} \qquad (4.31)\\
&\le \sqrt{1+\frac{Kh^2}{4}+\frac{3h}{4\mu_1}(-2K+K^2)}E_h(0). \qquad \square
\end{aligned}
$$

Now, by (4.19) and (4.24), we have

$$\left(T-4\sqrt{1+\frac{Kh^2}{4}+\frac{3h}{4\mu_1}(-2K+K^2)}\right)E_h(0)\le\int_0^T\left[\left|\frac{\phi_N}{h}\right|^2+\frac{|\,\phi'_N\,|^2}{6}\right]dt, \tag{4.32}$$

and (4.4) holds, with

$$C_2(T,h)=1\Big/\left(T-4\sqrt{1+\frac{Kh^2}{4}+\frac{3h}{4\mu_1}(-2K+K^2)}\right)$$

for every

$$T>T(h):=4\sqrt{1+\frac{Kh^2}{4}+\frac{3h}{4\mu_1}(-2K+K^2)}. \tag{4.33}$$

The minimal observability time $T(h)$ obtained in (4.33) is suboptimal.

However, the term $\left(1-h^2/4+3h(1+K^2)/4\mu_1\right)^{1/2}E_h$ is of the order of the discrete L^2-norm of the solutions. Therefore, it is a lower term when compared to the energy. A compactness-uniqueness argument may allow us to get rid of it (see [16]). This compactness-uniqueness principle guarantees that, if an observability inequality holds up to a compact additive remainder, then the remainder may be removed provided a suitable uniqueness or unique-continuation result holds (see [6]). Therefore, we can recover the uniform observability inequality for any $T>4$ with a not explicit constant $C_2(T,h)>0$, uniformly bounded with respect to h.

This concludes the proof of the second (so-called *inverse*) inequality in (4.4).

Let us prove now the first inequality (so-called *direct*) in (4.4). By (4.30), we have (see [5, 6, 7, 8, 9])

$$\begin{aligned}
&\frac{h}{12}\sum_{j=0}^{N}\int_0^T|\phi'_{j+1}-\phi'_j|^2\\
&\qquad\ge\frac{h^3}{6}\mu_1(M^{-1}A)\int_0^T\left[\sum_{j=0}^{N}\frac{|\,\phi'_j\,|^2}{6}+\frac{|\,\phi'_{j+1}+\phi'_j\,|^2}{12}\right]dt\\
&\qquad\ge\frac{h^3}{6}\left(\frac{6}{h^2}\frac{2}{3}\sin^2\left(\frac{\pi h}{2}\right)\right)\int_0^T\left[\sum_{j=0}^{N}\frac{|\,\phi'_j\,|^2}{6}+\frac{|\,\phi'_{j+1}+\phi'_j\,|^2}{12}\right]dt\\
&\qquad\ge\frac{2h^3}{3}\int_0^T\left[\sum_{j=0}^{N}\frac{|\,\phi'_j\,|^2}{6}+\frac{|\,\phi'_{j+1}+\phi'_j\,|^2}{12}\right]dt.
\end{aligned} \tag{4.34}$$

Hence, in (4.5)

$$\begin{aligned}
&\int_0^T\left[\left|\frac{\phi_N}{h}\right|^2+\frac{h^2}{6}\left|\frac{\phi'_N}{h}\right|^2\right]dt\\
&\qquad\le 2TE_h(0)-\frac{2h^3}{3}\sum_{j=0}^{N}\int_0^T\left[\frac{\left|\phi'_j\right|^2}{3}+\frac{\left|\phi'_{j+1}+\phi'_j\right|^2}{6}\right]dt+2X_h(t)\,|_0^T,
\end{aligned} \tag{4.35}$$

with X_h given by (4.6).

Moreover, in (4.9) we obtain

$$2TE_h(0) + Y_h(t)\,|_0^T = 2h\sum_{j=1}^{N}\int_0^T \left(\frac{1}{3}\left|\phi_j'\right|^2 dt + \frac{1}{6}\left|\phi_j' + \phi_{j+1}'\right|^2\right) dt,$$

with Y_h as in (4.13).

Therefore,

(4.36)
$$\begin{aligned}\frac{1}{2}\int_0^T \left[\left|\frac{\phi_N}{h}\right|^2 + \frac{h^2}{6}\left|\frac{\phi_N'}{h}\right|^2\right] dt &\le TE_h(0) - \frac{h^2}{3}\left(TE_h(0) + \frac{1}{2}Y_h(t)\,|_0^T\right) + X_h(t)\,|_0^T \\ &= TE_h(0) - \frac{h^2}{3}TE_h(0) - \frac{h^2}{6}Y_h(t)\,|_0^T + X_h(t)\,|_0^T \\ &= TE_h(0)\left[1 - \frac{h^2}{3}\right] - \frac{h^2}{6}Y_h(t)\,|_0^T + X_h(t)\,|_0^T \\ &= TE_h(0)\left[1 - \frac{h^2}{3}\right] + Z_h(t)\,|_0^T,\end{aligned}$$

where

$$Z_h(t) = X_h(t) - \frac{h^2}{6}Y_h(t). \tag{4.37}$$

The following provides an estimate of the term Z_h.

PROPOSITION 4.8. *For any $h > 0$ and any solution ϕ of* (2.8)*, it follows that*

$$|Z_h(t)| \le \sqrt{1 - \frac{h^4}{12} + \frac{h^4}{2\mu_1}\left(1 + \frac{h^2}{6}\right)}E_h(0) \;\; \forall 0 < t < T. \tag{4.38}$$

The proof of this relation is similar to the proof of (4.24), and it is given in detail in [5] and [9].

Using (4.36) and (4.38), we deduce that

$$\begin{aligned}\frac{1}{2}\int_0^T \left[\left|\frac{\phi_N}{h}\right|^2 + \frac{h^2}{6}\left|\frac{\phi_N'}{h}\right|^2\right] dt \\ \le \left[T\left[1 - \frac{h^2}{3}\right] + 2\sqrt{\left(1 - \frac{h^4}{12} + \frac{h^4}{2\mu_1}\left(1 + \frac{h^2}{6}\right)\right)}\right]E_h(0),\end{aligned} \tag{4.39}$$

with

$$C_1(T,h) = 1 \Big/ \left[T\left(1 - \frac{h^2}{3}\right) + 2\sqrt{\left(1 - \frac{h^4}{12} + \frac{h^4}{2\mu_1}\left(1 + \frac{h^2}{6}\right)\right)}\right].$$

Therefore, we can obtain a constant $C_1(T,h) > 0$ uniformly bounded, with respect to h such that the first inequality in (4.4) holds. The proof of Theorem 4.2 is now complete. □

Remark 4.9. It is important to underline that the inequality in (4.32) is uniform in h precisely because of considering slowly oscillating initial data in the fine grid. Relation (4.14) plays a fundamental role in the above proof.

The time $T > 4$ is twice the observability time for the continuous case. The analysis of the dispersion diagram of filtered solutions predicts a propagation velocity equal to $1/2$ in the class of solutions under consideration and confirms the observability time $T = 4$.

Any time-dependent scalar, linear partial differential equation with constant coefficients admits plane wave solutions

$$\phi(x,t) = e^{i(\omega t - \xi x)}, \quad \xi \in \mathbb{R},\ \omega \in \mathbb{C}, \tag{4.40}$$

where ξ is the *wave number* and ω is the *frequency*. The relationship

$$\omega = \omega(\xi) \tag{4.41}$$

is known as the *dispersion relation* for the equation. By dispersion, one understands the property of a dynamical continuous or discrete (in time) system to propagate, with different velocities, the components of the solution.

Any individual *"monochromatic wave"* of (4.40) moves at the *phase velocity*

$$c(\xi,\omega) = \frac{\omega(\xi)}{\xi}. \tag{4.42}$$

When one superimposes two waves with nearby propagation velocities, there appear wave packets which can propagate with different velocities. The energy of wave packets propagates at the so-called *group velocity*

$$C(\xi,\omega) = \frac{d\omega(\xi)}{d\xi}. \tag{4.43}$$

At the numerical level it is important to distinguish the notions of phase and group velocity. Phase velocity refers to the velocity of the propagation of individual monochromatic waves, while group velocity corresponds to the velocity of the propagation of wave packets, that may significantly differ from the phase velocity when waves of similar frequencies are combined.

• For the continuous wave equation (1.4), we have $\omega(\xi) = \xi$, and therefore $c(\xi) = C(\xi) = 1$.

• For the semidiscrete scheme (2.6), the dispersion relation is

$$\omega(\xi) = \frac{\sqrt{6}}{h}\sqrt{\frac{1-\cos(\xi h)}{2+\cos(\xi h)}}. \tag{4.44}$$

The phase velocity is in this case

$$c(\xi,\omega) = \frac{\sqrt{6}}{\xi h}\sqrt{\frac{1-\cos(\xi h)}{2+\cos(\xi h)}}. \tag{4.45}$$

The corresponding group velocity is

$$C(\xi,\omega) = \frac{\sqrt{6}}{2}\sqrt{\frac{2+\cos(\xi h)}{1-\cos(\xi h)}}\frac{3\sin(\xi h)}{(2+\cos(\xi h))^2} = 3\sqrt{\frac{3}{2}}\sqrt{\frac{1+\cos(\xi h)}{(2+\cos(\xi h))^3}}. \tag{4.46}$$

Note that, as $h \to 0$, for all ξ we have

$$C(\xi,\omega) \le \frac{3\sqrt{3}}{4}\sqrt{1+\cos(\xi h)} \to 0,$$

when $\xi = \pi/h$.

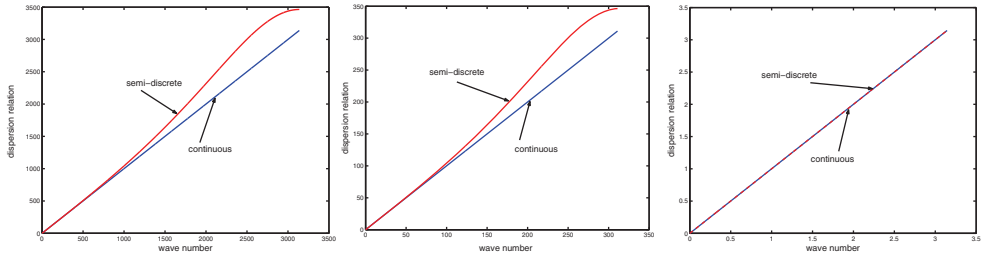


FIG. 1. *Dispersion relation for the continuous and finite-element semidiscrete cases with* $h = 0.001$ *(left),* $h = 0.1$ *(middle),* $h = 0.99999$ *(right).*

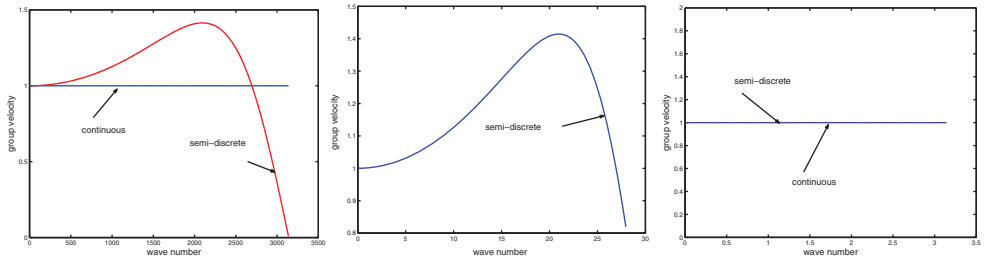


FIG. 2. *Group velocity for the continuous and the finite-element semidiscrete cases with* $h = 0.001$ *(left),* $h = 0.1$ *(middle),* $h = 0.99999$ *(right).*

The group velocity is the derivative of the eigenfrequencies μ_k, and the spectral gap is $\mu_{k+1} - \mu_k$. Both magnitudes are similar, and they become closer as $h \to 0$.

Thus, to efficiently observe at the point $x = 1$ a wave packet (or an initial disturbance concentrated near the extreme $x = 1$) that moves to the left (in the space variable) as t increases, starting at a point sufficiently close to $x = 1$, the time needed is

$$T \geq \frac{2}{\min\limits_{\xi}\{C(\xi,\omega)\}}. \tag{4.47}$$

This is the time that the wave packet needs to, after bouncing at the left extreme $x = 0$, reach the point $x = 1$ (time of going and return or observation time).

In fact, drawing the graph of the group velocity above, one can immediately visualize that it corresponds to a wave propagation of the order of h in space-time. These facts confirm that a suitable filtering of the spurious numerical high frequencies is a good cure for these pathologies.

In Figures 1 and 2 we describe the evolution of the dispersion relation and group velocity diagrams for the semidiscrete case starting with $h = 0.001$ up to $h = 0.99999$.

5. Uniform controllability of the two-grid solutions. In this section, we apply the uniform observability results obtained above to analyze the controllability properties of the semidiscrete system (2.6).

Let us denote by $\mathcal{F}_h$ the subspace of $\mathbb{R}^N \times \mathbb{R}^N$, $\mathcal{F}_h = V_h \times V_h$, with V_h defined by (3.1) endowed with the norm

$$\|(U,V)\|^2_{\mathcal{F}_h} := \|U\|^2_{\hbar^1} + |M^{1/2}V|^2_{\ell^2}, \quad U,\ V \in V_h,$$

and by $\mathcal{F}_h^*$ we denote the dual space of $\mathcal{F}_h$, considered as a subspace of the Euclidian space $\hbar^{-1} \times \ell^2$.

DEFINITION 5.1. *Let $(U,V) \in \mathcal{F}_h^* \subset \hbar^{-1} \times \ell^2$ and $(U',V') \in \mathcal{F}_h$. We define the symmetric products in $\mathcal{F}_h^* \times \mathcal{F}_h$*

$$\big((U,V), (U',V') \big)_{\mathcal{F}_h^* \times \mathcal{F}_h} = \big(MV, U'\big) + \big(- MU, V'\big), \tag{5.1}$$

where M is given by (2.11) *and $\big(\cdot,\cdot\big)$ is the scalar product in $\mathbb{R}^N$,*

$$\big(MV, U'\big) := \frac{1}{6} \sum_{j=1}^{N} (V_{j-1} + V_{j+1} + 4V_j) U'. \tag{5.2}$$

DEFINITION 5.2. *Let $(u',u) \in \hbar^{-1} \times \ell^2$ be discrete functions defined on the fine grid G_F^h. We define a suitable projection $(\bar{U}'_{2h}, \bar{U}_{2h})$ of $(\bar{u}'_h, \bar{u}_h)$ on the coarse grid as follows:*

$$U'_j = \frac{1}{12} \left(\frac{1}{2} u'_{2j-2} + 3u'_{2j-1} + 5u'_{2j} + 3u'_{2j+1} + \frac{1}{2} u'_{2j+2} \right) \tag{5.3}$$

and

$$U_j = \frac{1}{12} \left(\frac{1}{2} u_{2j-2} + 3u_{2j-1} + 5u_{2j} + 3u_{2j+1} + \frac{1}{2} u_{2j+2} \right) \tag{5.4}$$

for any $j = 1, \ldots, (N-1)/2$.

PROPOSITION 5.3. *Set $(\bar{\phi}_h^0, \bar{\phi}_h^1) \in \mathcal{F}_h$ and $(\bar{u}'_h, \bar{u}_h) \in \mathcal{F}_h^* \subset \hbar^{-1} \times \ell^2$. Hence*

$$\big((\bar{u}'_h, \bar{u}_h), (\bar{\phi}_h^0, \bar{\phi}_h^1) \big)_{\mathcal{F}_h^* \times \mathcal{F}_h} = 0 \tag{5.5}$$

if and only if

$$\bar{U}_{2h} = \bar{0}, \quad \text{and} \quad \bar{U}'_{2h} = \bar{0} \qquad \forall\, h > 0, \tag{5.6}$$

where $\bar{U}_{2h}$, $\bar{U}'_{2h}$ are defined by (5.4) *and* (5.3)*, respectively.*

Proof. Let us consider the $\mathbb{R}^N$-vectors in the fine grid

$$\bar{\phi}_h^0 = \left(0, 0, \ldots, \frac{1}{2}, 1, \frac{1}{2}, \ldots, 0, 0 \right), \quad \phi_{2j}^0 = \delta_{2j}(i),$$

$$\phi_{2j+1}^0 = \frac{\phi_{2j}^0 + \phi_{2j+2}^0}{2}, \quad j = 1, \ldots, \frac{N-1}{2},$$

where $\phi_0 = \phi_{N+1} = 0$, $\delta_j(i)$ is the Kronecker's delta, $1 \le i \le N$, i is arbitrarily fixed, and

$$\bar{\phi}_h^1 = (0, 0, \ldots, 0, \ldots, 0, 0).$$

Taking into account Definition (5.1) of the symmetric products in $\mathcal{F}_h^* \times \mathcal{F}_h$, we deduce that (5.5) is equivalent to

$$\big(M\bar{u}'_h, \bar{\phi}_h^0\big) + \big(-M\bar{u}_h, \bar{\phi}_h^1\big) = 0, \tag{5.7}$$

and by (5.2), (5.21) is

(5.8)

$$\left(M\bar{u}_h', \bar{\phi}_h^0\right) + \left(-M\bar{u}_h, \bar{\phi}_h^1\right) := \frac{1}{6}\sum_{j=1}^{N}\left[(u_{j-1}' + u_{j+1}' + 4u_j')\phi_j^0 - (u_{j-1} + u_{j+1} + 4u_j)\,\phi_j^1\right]$$

for all $(\bar{\phi}_h^0, \bar{\phi}_h^1) \in \mathcal{F}_h$. If (5.7) holds for all $(\bar{\phi}_h^0, \bar{\phi}_h^1) \in \mathcal{F}_h$, it follows that it also holds for every $\bar{\phi}_h^0$, $\bar{\phi}_h^1$, defined previously. Therefore,

$$\left(M\bar{u}_h', \bar{\phi}_h^0\right) = \frac{1}{6}\sum_{j=1}^{N}\left(u_{j-1}' + u_{j+1}' + 4u_j'\right)\phi_j^0 = 0. \tag{5.9}$$

We have

$$\sum_{j=1}^{N} u_{j-1}'\phi_j^0 = \sum_{j=1}^{\frac{N-1}{2}}\left[u_{2j-1}'\phi_{2j}^0 + u_{2j}'\phi_{2j+1}^0\right], \tag{5.10}$$

$$\sum_{j=1}^{N} u_{j+1}'\phi_j^0 = \sum_{j=1}^{\frac{N-1}{2}}\left[u_{2j+2}'\phi_{2j+1}^0 + u_{2j+1}'\phi_{2j}^0\right], \tag{5.11}$$

and

$$\sum_{j=1}^{N} u_j'\phi_j^0 = \sum_{j=1}^{\frac{N-1}{2}}\left[u_{2j+1}'\phi_{2j+1}^0 + u_{2j}'\phi_{2j}^0\right]. \tag{5.12}$$

Adding these relations we obtain

$$\sum_{j=1}^{\frac{N-1}{2}}\phi_{2j}^0\left[\frac{1}{2}u_{2j-2}' + 3u_{2j-1}' + 5u_{2j}' + 3u_{2j+1}' + \frac{1}{2}u_{2j+2}'\right] = 0. \tag{5.13}$$

This relation does not imply that $u_j' = 0$ for every $j = 1, \ldots, N$, but, if we consider U_{2h}' as in (5.3), we conclude

$$\bar{U}_{2h}' = \bar{0} \quad \forall\, h \in (0,1).$$

Analogously, we can prove that (5.7) holds if and only if $\bar{U}_{2h} = \bar{0}$, with $\bar{U}_{2h}$ defined as in (5.4). □

5.1. Partial controllability. Set $T > 0$. The partial controllability problem of system (2.6) in the space $\ell^2 \times \hbar^{-1}$ consists in finding a control $v_h \in L^2(0,T)$ such that the solution $\bar{u}_h$ of (2.6) satisfies

$$\mathcal{R}\left(\bar{u}_h'(T), \bar{u}_h(T)\right) = (\bar{0}, \bar{0}). \tag{5.14}$$

Operator $\mathcal{R}$ is the projection from $\hbar^{-1} \times \ell^2$ into $\mathcal{F}_h^*$, with

$$\mathcal{R}\left(\bar{u}_h'(T), \bar{u}_h(T)\right) = \left(\bar{U}'(T), \bar{U}(T)\right),$$

where $\bar{U}(T)$ and $\bar{U}'(T)$ are given by

$$U_j(T) = \frac{1}{12}\left(\frac{1}{2}u_{2j-2}(T) + 3u_{2j-1}(T) + 5u_{2j}(T) + 3u_{2j+1}(T) + \frac{1}{2}u_{2j+2}(T)\right) \tag{5.15}$$

and

(5.16)
$$U_j'(T) = \frac{1}{12}\left(\frac{1}{2}u_{2j-2}'(T) + 3u_{2j-1}'(T) + 5u_{2j}'(T) + 3u_{2j+1}'(T) + \frac{1}{2}u_{2j+2}'(T)\right),$$

for any $j = 1, \dots, (N-1)/2$, where $\bar{u}_h$ is the solution of (2.6).

To simplify the notations, we define

(5.17)
$$\begin{aligned} f(\bar{u}_h, \bar{\phi}_h) &:= \left(M\bar{u}_h^1, \bar{\phi}_h(0)\right) + \left(-M\bar{u}_h^0, \bar{\phi}_h'(0)\right) \\ &= \frac{1}{6}\sum_{j=1}^{N}\left[(u_{j-1}^1 + u_{j+1}^1 + 4u_j^1)\phi_j(0) - (u_{j-1}^0 + u_{j+1}^0 + 4u_j^0)\phi_j'(0)\right], \end{aligned}$$

where $\bar{u}_h$ and $\bar{\phi}_h$ are the solutions of systems (2.6) and (2.8), respectively, and $(\cdot,\cdot)$ is the scalar product (5.2).

Multiplying the first equation in (2.6) by $\{\bar{\phi}_h\}$, integrating by parts in $[0,T]$ and adding in j, $1 \le j \le N$, we deduce that

(5.18)
$$\begin{aligned} &\left(M\bar{u}_h'(T), \bar{\phi}_h^0\right) + \left(-M\bar{u}_h(T), \bar{\phi}_h^1\right) - f\left(\bar{u}_h, \bar{\phi}_h\right) + \int_0^T \sum_{j=1}^{N} u_j(t)\left(M\phi''(t) + A\phi(t)\right)_j \\ &\quad = \frac{1}{h^2}\int_0^T v_h(t)\phi_N(t)dt - \frac{1}{6}\int_0^T v_h(t)\phi_N''(t)dt, \end{aligned}$$

with $f(\bar{u}_h, \bar{\phi}_h)$ defined as in (5.17).

Therefore, it is immediate to see that, for all solution $\bar{\phi}_h$ of the semidiscrete system (2.8), we have

(5.19)
$$\begin{aligned} &\left(M\bar{u}_h'(T), \bar{\phi}_h^0\right) + \left(-M\bar{u}_h(T), \bar{\phi}_h^1\right) - f(\bar{u}_h, \bar{\phi}_h) \\ &\qquad + \frac{1}{6}\int_0^T v_h(t)\phi_N''(t)dt = \frac{1}{h^2}\int_0^T v_h(t)\phi_N(t)dt. \end{aligned}$$

The solution of system (2.6) may be characterized through a transposition argument based on the identity above. Indeed, $\{\bar{u}_h\}_h$ solves (2.6) if for every $s \in [0,T]$, it holds

$$L_s(\bar{\phi}^0, \bar{\phi}^1) = \left(M\bar{u}_h'(s), \bar{\phi}_h(s)\right) + \left(-M\bar{u}_h(s), \bar{\phi}_h'(s)\right)$$

for every solution $\{\bar{\phi}_h\}$ of the discrete problem (2.8), where the functional $L_s : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$ is such that

$$L_s(\bar{\phi}^0, \bar{\phi}^1) = f(\bar{u}_h, \bar{\phi}_h) - \frac{1}{6}\int_0^s v_h(t)\phi_N''(t)dt + \frac{1}{h^2}\int_0^s v_h(t)\phi_N(t)dt.$$

Now, let us consider the functional $L_T : \mathbb{R}^{\mathbb{N}} \times \mathbb{R}^N \to \mathbb{R}$:

$$L_T(\bar{u}_h^0, \bar{u}_h^1) = f(\bar{u}_h, \bar{\phi}_h) - \frac{1}{6}\int_0^T v_h(t)\phi_N''(t)dt + \frac{1}{h^2}\int_0^T v_h(t)\phi_N(t)dt.$$

Proposition 5.4 provides a characterization of the partial controllability property of semidiscrete system (2.6).

PROPOSITION 5.4. *Set $T > 0$. Problem (2.6) is partially controllable in $\ell^2 \times \hbar^{-1}$ if and only if, for each initial state $(\bar{u}_h^0, \bar{u}_h^1) \in \ell^2 \times \hbar^{-1}$, there exists a function $v_h \in L^2(0,T)$ such that*

$$f\left(\bar{u}_h, \bar{\phi}_h\right) = \frac{1}{6}\int_0^T v_h(t)\phi_N''(t)dt - \frac{1}{h^2}\int_0^T v_h(t)\phi_N(t)dt \tag{5.20}$$

for any initial data $(\bar{\phi}_h^0, \bar{\phi}_h^1) \in \mathcal{F}_h$ associated to the solution $\bar{\phi}_h$ of (2.8).

Proof. In view of identity (5.19), it is immediate to see that the control v_h such that (5.21) holds is characterized by the property (5.20). □

Remark 5.5. In the continuous case we have the following characterization of the property of exact controllability: the control v steers the initial data (u^0, u^1) of problem (1.1) to $(0,0)$ if and only if

$$\int_0^T v(t)\phi_x(1,t)dt = \int_0^1 (u^1\phi^0 - u^0\phi^1)dx$$

for every solution ϕ of the homogeneous problem (1.4), with initial data (ϕ^0, ϕ^1).

Set $(\bar{\phi}_h^0, \bar{\phi}_h^1) \in \mathcal{F}_h$. As we said above, (5.20) is equivalent to

$$\left(M\bar{u}_h'(T), \bar{\phi}_h^0\right) + \left(-M\bar{u}_h(T), \bar{\phi}_h^1\right) = 0 \tag{5.21}$$

for all $(\bar{\phi}_h^0, \bar{\phi}_h^1) \in \mathcal{F}_h$.

Note that (5.21) does not imply that the solution of (2.6) vanishes at time $t = T$. Rather the projection (5.14) of the state (u', u) on the coarse grid vanishes. More precisely, $\bar{U}(T), \bar{U}'(T)$ vanish.

5.2. Construction of the two-grid controls. We now briefly describe the algorithm for building an efficient approximation of the control v of (1.1), which is inspired in Glowinski [2]. We apply the uniform observability results obtained in the previous section to analyze the controllability properties of the semidiscrete system (2.6). We can prove now the existence of a bounded sequence of HUM controls for the semidiscrete problem (2.6).

The following uniform (with respect to $h \to 0$) partial controllability property holds in the space $\mathcal{F}_h$.

THEOREM 5.6. *For any $T > 4$, system (2.6) is partially controllable on $\ell^2 \times \hbar^{-1}$ for all $0 < h < 1$. Moreover, the controls $\{v_h\}_h$ of minimal norm are uniformly bounded, more precisely, there exists a positive constant $C = C(T)$, independent of $h \in (0,1)$, such that*

$$\|v_h\|_{L^2(0,T)} \le C\|\mathcal{R}(u_h^1, -u_h^0)\|^2_{\mathcal{F}_h^*} \qquad \forall\, 0 < h < 1. \tag{5.22}$$

Proof. Given $T > 4$ and $\varepsilon > 0$ small enough ($T - 4 > 2\varepsilon$), we consider $\rho_\varepsilon(t) \in \mathcal{C}^\infty(\mathbb{R})$, so that $0 \le \rho_\varepsilon \le 1$, $\rho_\varepsilon = 1$ in $(\varepsilon, T-\varepsilon)$, and the support of ρ_ε is compact in $(0,T)$. We define the functional $J_h : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$ by

$$J_h\left((\bar{\phi}_h^0, \bar{\phi}_h^1)\right) = \frac{1}{2}\int_0^T \left|\frac{\phi_N(t)}{h}\right|^2 dt + \frac{h^2}{12}\int_0^T \rho_\varepsilon(t)\left|\frac{\phi_N'(t)}{h}\right|^2 dt + hf(\bar{u}_h, \bar{\phi}_h), \tag{5.23}$$

where $\bar{\phi}_h$ is the solution of the adjoint problem (2.8), with initial data $(\bar{\phi}_h^0,\ \bar{\phi}_h^1) \in \mathcal{F}_h$, $(\bar{u}_h^0, \bar{u}_h^1)$ are the initial data of system (2.6) to be controlled.

The conservative energy of system (2.8) is

$$E_h(0) = \frac{1}{2} \left\| \left(\bar{\phi}_h^0, \bar{\phi}_h^1 \right) \right\|_{\mathcal{F}_h}^2 . \tag{5.24}$$

We have

$$\left| hf(\bar{u}_h, \bar{\phi}_h) \right| = \left| \left((M\bar{u}_h^1, \bar{\phi}_h(0)), (-M\bar{u}_h^0, \bar{\phi}_h'(0)) \right) \right| \leq \| \mathcal{R} \left(\bar{u}_h^1, -\bar{u}_h^0 \right) \|_{\mathcal{F}_h^*} \| \left(\bar{\phi}_h^0, \bar{\phi}_h^1 \right) \|_{\mathcal{F}_h}. \tag{5.25}$$

The functional J_h is continuous according to (5.25) and the direct observability inequality in (4.4). Moreover, J_h is convex. On the other hand, in view of inequality (4.4), it is uniformly (with respect to $0 < h < 1$) coercive in $\mathcal{F}_h$ provided $T - 2\varepsilon > 4$. Thus, for each $0 < h < 1$ and for any $T - 2\varepsilon > 4$, there is a unique minimizer $(\hat{\phi}_h^0, \hat{\phi}_h^1) \in \mathcal{F}_h$ of the functional J_h:

$$I_h = J_h \left(\hat{\phi}_h^0, \hat{\phi}_h^1 \right) = \min_{\left(\bar{\phi}_h^0, \bar{\phi}_h^1 \right) \in \mathcal{F}_h} J_h \left(\bar{\phi}_h^0, \bar{\phi}_h^1 \right). \tag{5.26}$$

The corresponding Euler equation reads

$$\left(\hat{\phi}_h^0, \hat{\phi}_h^1 \right) \in \mathcal{F}_h : \quad J_h' \left(\left(\hat{\phi}_h^0, \hat{\phi}_h^1 \right) \right) \left(\phi_h^0, \phi_h^1 \right) = 0 \quad \forall \ \left(\phi_h^0, \phi_h^1 \right) \in \mathcal{F}_h.$$

Calculating the Gateaux derivative of J_h in $(\hat{\phi}_h^0, \hat{\phi}_h^1)$, we get that this equality is equivalent to

$$\int_0^T \frac{\hat{\phi}_N(t)}{h} \left[\frac{\phi_N(t)}{h} - \frac{h^2}{6} \left(\rho_\varepsilon(t) \frac{\phi_N'(t)}{h} \right)' \right] dt + hf \left(\bar{u}_h, \bar{\phi}_h \right) = 0 \tag{5.27}$$

for every solution $\bar{\phi}_h$ of the adjoint problem (2.8).

Now, we define

$$v_h(t) = \frac{\hat{\phi}_N(t)}{h}, \tag{5.28}$$

hence, in (5.27) we obtain

$$\int_0^T v_h(t) \left[\frac{\phi_N(t)}{h} - \frac{h^2}{6} \left(\rho_\varepsilon(t) \frac{\phi_N'(t)}{h} \right)' \right] dt + hf \left(\bar{u}_h, \bar{\phi}_h \right) = 0. \tag{5.29}$$

Thus, when $\varepsilon \to 0$, we obtain the partial controllability property (5.20) for every $T > 4$.

We have $I_h \leq J_h(\bar{0}, \bar{0}) = 0$, which is equivalent to

$$\frac{1}{2} \int_0^T \left[\left| \frac{\hat{\phi}_N(t)}{h} \right|^2 dt + \frac{h^2}{6} \int_0^T \rho_\varepsilon(t) \left| \frac{\hat{\phi}_N'(t)}{h} \right|^2 dt \right] \leq \| \mathcal{R} \left(u_h^1, -u_h^0 \right) \|_{\mathcal{F}_h^*} \| \left(\hat{\phi}_h^1, \hat{\phi}_h^1 \right) \|_{\mathcal{F}_h}. \tag{5.30}$$

Combining (5.30) with the uniform observability inequality (4.4) yields

$$\begin{aligned} &\int_0^T \left[\left| \frac{\hat{\phi}_N(t)}{h} \right|^2 dt + \frac{h^2}{6} \int_0^T \left| \frac{\hat{\phi}_N'(t)}{h} \right|^2 dt \right] \\ &\qquad \leq C \| \mathcal{R} \left(u_h^1, -u_h^0 \right) \|_{\mathcal{F}_h^*} \left(\int_0^T \left[\left| \frac{\hat{\phi}_N(t)}{h} \right|^2 dt + \frac{h^2}{6} \int_0^T \left| \frac{\hat{\phi}_N'(t)}{h} \right|^2 dt \right] \right)^{\frac{1}{2}}, \end{aligned} \tag{5.31}$$

or, equivalently,

$$\|v_h\|_{L^2(0,T)} = \int_0^T \left|\frac{\hat\phi_N(t)}{h}\right|^2 dt \le \int_0^T \left[\left|\frac{\hat\phi_N(t)}{h}\right|^2 dt + \frac{h^2}{6}\int_0^T \left|\frac{\hat\phi_N'(t)}{h}\right|^2 dt\right] \le C\|\mathcal{R}\left(u_h^1, -u_h^0\right)\|^2_{\mathcal{F}_h^*}, \tag{5.32}$$

where $C = C(T)$ is independent of h.

In view of (5.32), the sequence of controls we have obtained v_h is bounded in $L^2(0,T)$ when $h \to 0$ as we wanted to prove. □

Remark 5.7. Observe that, in this method, the functional J_h is minimized on $\mathcal{F}_h$, in contrast with the direct application of the minimization method of the functional J_h (that leads to divergent controls as $h \to 0$ for some pathological initial data; see [5] and [17]). The uniform observability inequality (4.4) guarantees the uniform coercivity of J_h when restricted to $\mathcal{F}_h$ and, consequently, the boundedness of the controls.

Note that (5.27) does not imply that the solution of (2.6) vanishes at time $t = T$. Rather a suitable projection of the state (u, u') on the coarse grid vanishes. More precisely, if we define the restriction (5.15)–(5.16), we have

$$\mathcal{R}(\bar u(T), \bar u'(T)) = \left(\bar U_h(T), \bar U_h'(T)\right) = (\bar 0, \bar 0).$$

5.3. Convergence of the controls. The aim of this section is to obtain controls for the continuous system (1.1) as limits as $h \to 0$ of the sequences of controls for the corresponding semidiscrete problems (2.6). We proceed as follows.

Given an initial state $(u^0, u^1) \in L^2(0,1) \times H^{-1}(0,1)$ of the continuous system (1.1), we develop it in Fourier series

$$\left(u^0, u^1\right) = \sum_{k=1}^{\infty} \left(c_k^0, c_k^1\right) \varphi_k(x), \tag{5.33}$$

with $(c_k^0, c_k^1) \in \ell^2 \times \hbar^{-1}$, $\varphi_k(x) = \sin(k\pi x)$. Further, we construct the sequence of discrete initial states:

$$\left(\bar u_h^0, \bar u_h^1\right) = \sum_{k=1}^{N} \left(c_k^0, c_k^1\right) \bar\varphi_k, \tag{5.34}$$

with $\bar\varphi_k = (\varphi_{k,1}, \dots, \varphi_{k,N})$, $\varphi_{k,j} = \sin(k\pi j \Delta x)$, $j, k = 1, \dots, N$.

In view of Theorem 5.6, for every $h \in (0,1)$ there exists a HUM control (v_h) for semidiscrete system (2.6), with initial data (5.34). Then, we prove in Theorem 5.8 that the sequence (v_h) converges (in a sense defined below) to a function $v \in L^2(0,T)$, which is the HUM control for system (1.1), with initial data (5.33).

As the solution u of (1.1), with initial data $(u^0, u^1) \in L^2(0,1) \times H^{-1}(0,1)$ and control $v \in L^2(0,T)$ satisfies $u \in \mathcal{C}([0,T]; L^2(0,1)) \cap \mathcal{C}^1([0,T]; H^{-1}(0,1))$, then there exist functions $c_k(t)$, $k \in \mathbb{N}$ such that

$$u(x,t) = \sum_{k\in\mathbb{N}} c_k(t)\varphi_k(x), \tag{5.35}$$

with

$$\sup_{t\in[0,T]} \left[\| c_k(t) \|_{\ell^2} + \| c_k'(t) \|_{\hbar^{-1}}\right] < \infty. \tag{5.36}$$

The solution of (2.6) may be written in a similar form as $\bar{u}_h(t) = \sum_k c_{k,h}(t)\bar{\varphi}_k$ by putting $c_{k,h}(t) = 0$ for $k > N$.

The control $v \in L^2(0,T)$ of minimal L^2-norm (HUM control) verifies

$$v(t) = -\partial_x \hat{\phi}(1,t), \tag{5.37}$$

where $\hat{\phi}$ is the solution of the adjoint problem (1.4), with initial data $(\hat{\phi}^0, \hat{\phi}^1) \in H_0^1(0,1) \times L^2(0,1)$, the unique minimizer of the functional

$$J\left(\phi^0, \phi^1\right) = \frac{1}{2}\int_0^T \mid \partial_x \phi(1,t) \mid^2 dt - \int_0^1 u^0\phi^1 - < u^1, \phi^0 >_{-1,1}$$

in that space $H_0^1(0,1) \times L^2(0,1)$.

The following convergence result holds.

THEOREM 5.8. *Let $T > 4$. Let $\bar{u}_h$ and u be the solutions of* (2.6) *and* (1.1) *as above. Then*

$$\{c_{k,h}(\cdot)\}_{k\in\mathbb{N}} \to \{c_k(\cdot)\}_{k\in\mathbb{N}} \text{ strongly in } L^p(0,T;\ell^2) \cap W^{1,p}(0,T;\hbar^{-1}) \text{ as } h \to 0, \text{ with } 1 \le p < \infty \tag{5.38}$$

$$v_h(\cdot) \to v(\cdot) \quad \text{strongly in } L^2(0,T) \text{ as } h \to 0, \tag{5.39}$$

where $v_h = v_h(t)$ is the control of (2.6) *constructed by the multigrid algorithm above and $v = v(t)$ is the HUM control for the continuous wave equation which drives the initial data (u^0, u^1) to rest in time T.*

Proof. (Sketch of the proof.) In view of (5.22), it is easy to see that

$$\int_0^T |v_h|^2 \, dt \le C. \tag{5.40}$$

Then, up to the extraction of a subsequence that we still denote by $\{v_h\}_h$, we have

$$v_h(t) \rightharpoonup v(t) \text{ in } L^2(0,T) \text{ as } h \to 0. \tag{5.41}$$

By a Γ-convergence argument, it can also be seen that the limit v is given by

$$v(t) = -\partial_x \hat{\phi}(1,t), \tag{5.42}$$

where $\hat{\phi}$ is the solution of the adjoint problem (1.4), with initial data $(\hat{\phi}^0, \hat{\phi}^1) \in H_0^1(0,1) \times L^2(0,1)$, the unique minimizer of the functional

$$J\left(\phi^0, \phi^1\right) = \frac{1}{2}\int_0^T \mid \partial_x \phi(1,t) \mid^2 dt - \int_0^1 u^0\phi^1 - < u^1, \phi^0 >_{-1,1} \tag{5.43}$$

in the energy space $H_0^1(0,1) \times L^2(0,1)$. By taking limits in (5.29) and thanks to the construction of the initial data to be controlled for the discrete system, we obtain

$$0 = \int_0^1 \left[u^1(x)\phi^0(x) - u^0(x)\phi^1(x)dx\right] + \int_0^T v(t)\partial_x\phi(1,t)dt, \tag{5.44}$$

and this latter condition is equivalent to the fact that v, the limit in (5.41), is a control for system (1.1), driving the initial data (u^0, u^1) to rest , i.e., $v \in L^2(0,T)$ is the control of minimal L^2-norm.

The limit v being identified in a unique way, we deduce that the whole sequence v_h converges.

Moreover, by the hypotheses of Theorem 5.8, the linear term of the discrete functional J_h in (5.23) converges to the linear term of the functional defined in (5.43). Therefore, proving (5.39) is equivalent to proving that

$$J_h\left(\hat{\phi}_h^0, \hat{\phi}_h^1\right) \to J\left(\hat{\phi}^0, \hat{\phi}^1\right) \text{ as } h \to 0,$$

where $(\hat{\phi}_h^0, \hat{\phi}_h^1) \in \mathcal{F}_h$ minimizes (5.23) and $(\hat{\phi}^0, \hat{\phi}^1) \in H_0^1(0,1) \times L^2(0,1)$ minimizes (5.43). Indeed, taking into account the convergence of the linear terms in this functional and the structure of the functionals (5.23) and (5.43), we deduce the convergence of the norms of the controls that, together with the weak convergence, ensure strong convergence.

Thus, the controls v_h and the controlled discrete solutions u_h converge to the control and the controlled solution of the wave equation (1.1). It is important to note that the projections of the solutions of the controlled system end up covering the whole range of frequencies so that, in the limit, we recover the exact controllability property (1.3) of the continuous wave equation. The details of the several steps of the proof are given in [9], and it is similar to the proof of the convergence of the solution of the semidiscrete $1-d$ beam equation (see [7] for more details), and we omit them for brevity. □

Remark 5.9. Equation (5.32) implies that a subsequence of (v_h) can be chosen so that it is weak-convergent to a function $v \in L^2(0,T)$. By taking limits in (5.29) and thanks to the convergence assumptions of the initial data, we obtain

$$0 = \int_0^1 \left[u^1(x)\phi^0(x) - u^0(x)\phi^1(x)dx\right] + \int_0^T v(t)\partial_x\phi(1,t)dt, \tag{5.45}$$

and this condition is equivalent to the fact that v is control for the system (1.1) driving initial data (u^0, u^1) to rest. One can also show by a Γ-convergence argument that v is in fact the HUM control.

Remark 5.10. Similar results can be obtained for the finite-difference semidiscretization of the $1-d$ wave equation (see [9, 10], [17]).

The numerical efficiency of the multigrid algorithm which illustrate our theoretical results was proved in [2] and in [8] by various numerical experiments in the multidimensional and one-dimensional cases, respectively.

Acknowledgment. The author would like to thank Enrique Zuazua for fruitful discussions and his help in making this work possible.

REFERENCES

[1] M. Asch and G. Lebeau, *Geometrical aspects of exact boundary controllability for the wave equation—A numerical study*, ESAIM Control Optim. Calc. Var., 3 (1998), pp. 163–212.

[2] R. Glowinski, *Ensuring well-posedness by analogy; Stokes problem and boundary control of the wave equation*, J. Comput. Phys., 103 (1992), pp. 189–221.

[3] R. Glowinski, C. H. Li, and J.-L. Lions, *A numerical approach to the exact boundary controllability of the wave equation* (I). *Dirichlet controls: Description of the numerical methods*, Japan J. Appl. Math., 68 (1990), pp. 1–76.

[4] E. Isaacson and H. B. Keller, *Analysis of Numerical Methods*, John Wiley & Sons, New York, 1966.

[5] J. A. Infante and E. Zuazua, *Boundary observability for the space discretization of the one-dimensional wave equation*, M2AN Math. Model. Numer. Anal., 33 (1999), pp. 407–438.

[6] J. L. Lions, *Contrôlabilité exacte, stabilisation et perturbations du systemes distribués. Tome 1 Contrôlabilité exacte*, Masson, Paris, RMA8, 1988.
[7] L. León and E. Zuazua, *Boundary controllability of the finite-difference space semi-discretization of the beam equation*, ESAIM Control Optim. Calc. Var., 2 (2002), pp. 827–862.
[8] A. Münch, *Famille de schémas implicites uniformément contrôlables pour l'équation des ondes 1-D.*, C.R. Math. Acad. Sci. Paris, 339 (2004), pp. 733–738.
[9] M. Negreanu, *Métodos numéricos para el análisis de la propagación, observación y control de ondas*, Ph.D. thesis, Universidad Complutense de Madrid, Departamento de Matemática Aplicada, Madrid, 2004.
[10] M. Negreanu and E. Zuazua, *Convergence of a multi-grid method for the controllability of a 1-d wave equation*, C.R. Acad. Sci. Paris Ser. I, 338 (2004), pp. 413–418.
[11] M. Negreanu and E. Zuazua, *Uniform boundary controllability of a discrete $1-D$ wave equation*, Systems Control Lett., 48 (2003), pp. 261–279.
[12] M. Negreanu and E. Zuazua, *Discrete Ingham inequalities and applications*, SIAM J. Numer. Anal., 44 (2006), pp. 412–448.
[13] M. Negreanu, A.-M. Matache, and C. Schwab, *Wavelet filtering for exact controllability of the wave equation*, SIAM J. Sci. Comput., 28 (2006), pp. 1851–1885.
[14] L. N. Trefethen, *Group velocity in finite difference schemes*, SIAM Rev., 24 (1982), pp. 113–136.
[15] R. Vichnevetsky and J. B. Bowles, *Fourier Analysis of Numerical Approximations of Hyperbolic Equations*, SIAM Stud. Appl. Math. 5, SIAM, Philadelphia, 1982.
[16] E. Zuazua, *Boundary observability for the finite-difference space semi-discretizations of the 2-d wave equation in the square*, J. Math. Pures Appl., 78 (1999), pp. 523–563.
[17] E. Zuazua, *Propagation, observation, control of waves approximated by finite difference methods*, SIAM Rev., 47 (2005), pp. 197–243.

SIAM J. NUMER. ANAL.
Vol. 46, No. 6, pp. 3264–3284
© 2008 Society for Industrial and Applied Mathematics

SOLVING ELLIPTIC FINITE ELEMENT SYSTEMS IN NEAR-LINEAR TIME WITH SUPPORT PRECONDITIONERS*

ERIK G. BOMAN†, BRUCE HENDRICKSON‡, AND STEPHEN VAVASIS§

Abstract. We consider linear systems arising from the use of the finite element method for solving scalar linear elliptic problems. Our main result is that these linear systems, which are symmetric and positive semidefinite, are well approximated by symmetric diagonally dominant matrices. Our framework for defining matrix approximation is support theory. Significant graph theoretic work has already been developed in the support framework for preconditioners in the diagonally dominant case, and, in particular, it is known that such systems can be solved with iterative methods in nearly linear time. Thus, our approximation result implies that these graph theoretic techniques can also solve a class of finite element problems in nearly linear time. We show that the support number bounds, which control the number of iterations in the preconditioned iterative solver, depend on mesh quality measures but not on the problem size or shape of the domain.

Key words. finite elements, preconditioning, matrix algorithm, support graph, diagonal dominance

AMS subject classifications. 65F10, 65F50, 65N22

DOI. 10.1137/040611781

1. Introduction.

Finite element discretizations of elliptic partial differential equations (PDEs) give rise to large sparse linear systems of equations. A topic of great interest is preconditioners for the iterative solution of such systems. We focus on scalar boundary value problems of the form $\nabla \cdot (\theta \nabla u) = -f$, in which θ is a scalar conductivity field. See (3.1) below for a more detailed statement of the PDE under consideration. Such PDEs arise in a variety of physical applications listed in section 3. We prove that the stiffness matrix K of this PDE, which is defined precisely by (3.7) below, can be well approximated by a diagonally dominant matrix $\bar{K}$. "Well approximated" means that $\kappa(\bar{K}, K)$ has an upper bound in terms of mesh quality measures only; the notation $\kappa(\cdot, \cdot)$ is defined below. Since significant theory has been developed for diagonally dominant matrices, our result shows that the same theory extends to this class of stiffness matrices. In particular, our approximation result means that the system $K\mathbf{x} = \mathbf{f}$ arising from the finite element method (FEM) can be

*Received by the editors July 20, 2004; accepted for publication (in revised form) May 20, 2008; published electronically October 17, 2008.

http://www.siam.org/journals/sinum/46-6/61178.html

†Scalable Algorithms Department, Sandia National Laboratories, Albuquerque, NM 87185-1318 (egboman@sandia.gov). This author's work was funded by the Applied Mathematical Sciences program, U.S. Department of Energy, Office of Energy Research and performed at Sandia National Labs, a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC-94AL85000.

‡Computer Science and Informatics Department, Sandia National Laboratories, Albuquerque, NM 87185-1318 (bah@cs.sandia.gov). This author's work funded by the Applied Mathematical Sciences program, U.S. Department of Energy, Office of Energy Research and performed at Sandia National Labs, a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC-94AL85000.

§Corresponding author. Department of Combinatorics and Optimization, University of Waterloo, Waterloo, ON N2L 2W2, Canada (vavasis@math.uwaterloo.ca). Part of this work was done while the author was in the Department of Computer Science of Cornell University. This author was supported in part by NSF grant CCF-0085969 and by a grant from NSERC, Canada.

solved in nearly linear time by preconditioned iterative methods.

Our analysis uses the *support theory* framework described in [5] for analyzing condition numbers and (generalized) eigenvalues for preconditioned systems. Our analysis is as follows. In sections 5–7, we state and prove our theorem that K may be factored as $K = A^T \bar{D}^{1/2} H \bar{D}^{1/2} A$. A preliminary general-purpose result in section 2 shows that a matrix of this form can be approximated by $\bar{K} = A^T \bar{D} A$ (a diagonally dominant matrix), with the quality of approximation depending on $\kappa(H)$. The analysis of $\kappa(H)$ in section 7 establishes that this quantity depends on the space dimension d and degree p of the FEM, the quadrature rule, and various mesh quality measures but does not depend on the number of nodes or elements or on the shape of the domain or the size of elements. It also does not depend on the conductivity field θ under certain assumptions to be made below.

The idea of approximating FEM systems by diagonally dominant matrices is not new; see, for example, Gustafsson [13]. In fact, our approach is similar to Gustafsson's in that we approximate each element matrix by a diagonally dominant matrix. In contrast to [13], we are able to rigorously prove bounds on the spectral properties of our approximation (and thus also the preconditioner).

An approach in [3, section 7.3] proposes to precondition general FEM problems with a diagonally dominant preconditioner obtained by using a lower-order method. Again, our result is a departure from theirs because we get precise and rigorous bounds on the quality of the approximation.

In section 10, we specialize the theory to four common cases of the finite element method and report on computational tests for these cases. The computational tests are intended to illustrate the size of the constants in our main theorem and do not measure the practical performance of our solution strategy versus competitors. See further remarks on the practical efficiency of our preconditioning approach in section 9.

2. Support theory. Let A, B be symmetric positive semidefinite matrices of the same size, and let $N(A)$ and $N(B)$ denote their null spaces. We define the *support number* of A with respect to B to be

$$\sigma(A,B) = \sup_{\mathbf{x} \in \mathbb{R}^n - N(B)} \frac{\mathbf{x}^T A \mathbf{x}}{\mathbf{x}^T B \mathbf{x}}. \tag{2.1}$$

Note that this quantity is finite only if $N(B) \subset N(A)$. We follow the convention established in the previous literature of using $\sigma(\cdot,\cdot)$ to denote support numbers. Unfortunately, σ is also commonly used to denote singular values. In this paper, σ with one argument is a singular value, and σ with two arguments is a support number.

The results in this section can be partially generalized to indefinite symmetric matrices. This generalization requires a more general definition of a support number than (2.1) and requires the use of the symmetric product support theorem from [5]. Since this paper focuses on the positive semidefinite case, we omit this generalization.

If A and B are symmetric positive definite (SPD), then $\sigma(A,B) = \lambda_{\max}(A,B)$, the largest generalized eigenvalue. When B is a preconditioner for A in the preconditioned conjugate gradient iterative method [10], the condition number of the preconditioned system is given by $\sigma(A,B)\sigma(B,A)$, which we denote as $\kappa(A,B)$. When A, B are SPD, then $\kappa(A,B) = \kappa(B^{-1}A)$, where κ denotes the standard spectral condition number.

A principal goal of this paper is to propose the construction of a symmetric diagonally dominant matrix $\bar{K}$ that approximates the finite element stiffness matrix K in the sense that $\kappa(K,\bar{K})$ is not too large. The following two lemmas and subsequent theorem define our framework for defining $\bar{K}$ and bounding $\kappa(K,\bar{K})$.

LEMMA 2.1. *Suppose $V \in \mathbb{R}^{n\times p}$, and suppose $H \in \mathbb{R}^{p\times p}$ is SPD. Then*

$$\sigma(VHV^T, VV^T) \le \lambda_{\max}(H),$$

where $\lambda_{\max}$ denotes the largest eigenvalue.

Proof.

$$\begin{aligned}\sigma(VHV^T, VV^T) &= \sup_{\mathbf{x}\in\mathbb{R}^n - N(V^T)} \frac{\mathbf{x}^T VHV^T\mathbf{x}}{\mathbf{x}^T VV^T\mathbf{x}} \\ &= \sup_{\mathbf{y}\in R(V^T)} \frac{\mathbf{y}^T H\mathbf{y}}{\mathbf{y}^T\mathbf{y}} \\ &\le \lambda_{\max}(H).\end{aligned}$$

The second line follows from the first by substituting $\mathbf{y} = V^T\mathbf{x}$ and the third from the Courant–Fischer minimax theorem [10, section 8.1.1]. Here, $R(V^T)$ denotes the range space of V^T. ∎

LEMMA 2.2. *Suppose $V \in \mathbb{R}^{n\times p}$, and suppose $H \in \mathbb{R}^{p\times p}$ is SPD. Then*

$$\sigma(VV^T, VHV^T) \le 1/\lambda_{\min}(H),$$

where $\lambda_{\min}$ denotes the smallest eigenvalue.

Proof. Using the same idea in the previous proof leads to a supremum over the Rayleigh quotient $\mathbf{y}^T\mathbf{y}/(\mathbf{y}^T H\mathbf{y})$, which is bounded above by $1/\lambda_{\min}(H)$. ∎

Combining these two lemmas yields a result for the condition number.

THEOREM 2.3. *Suppose $V \in \mathbb{R}^{n\times p}$, and suppose $H \in \mathbb{R}^{p\times p}$ is SPD. Then*

$$\kappa(VHV^T, VV^T) \le \kappa(H).$$

Proof. The result follows from Lemma 2.1 and 2.2 and by using the fact $\kappa(H) = \lambda_{\max}(H)/\lambda_{\min}(H)$ when H is SPD. ∎

The way we will apply this theorem is to let $V = A^T\bar{D}^{1/2}$, where A is the node-arc incidence matrix of a graph and $\bar{D}$ is a diagonal weight matrix. Then $\bar{K} = VV^T = A^T\bar{D}A$ is diagonally dominant, while $K = VHV^T = A^T\bar{D}^{1/2}H\bar{D}^{1/2}A$ is not (in general). Therefore, we now have a tool to approximate more general positive definite matrices.

3. Finite element analysis. In this section we provide a brief summary of isoparametric finite element approximation as well as an introduction to notation that is crucial for our main theorem. The material in this section is standard [15] in textbooks, except that our description herein uses notation for indexing that is more detailed than usual. This section concludes with a summary of our notation.

The class of problems under consideration consists of finite element discretizations of the following second-order elliptic boundary value problem. Find $u : \Omega \to \mathbb{R}$ satisfying

$$\begin{aligned}\nabla\cdot(\theta\nabla u) &= -f \quad \text{on } \Omega, \\ u &= u_0 \quad \text{on } \Gamma_1, \\ \theta\partial u/\partial n &= g \quad \text{on } \Gamma_2.\end{aligned} \tag{3.1}$$

Here, Ω is a bounded open subset of $\mathbb{R}^d$ (typically $d = 2$ or $d = 3$), Γ_1 and Γ_2 form a partition of $\partial\Omega$, θ is a given scalar field on Ω that is positive-valued everywhere

and is sometimes called the *conductivity*, $f : \Omega \to \mathbb{R}$ is a given function called the *forcing function*, u_0 is a given function called the *Dirichlet boundary condition*, and g is another given function called the *Neumann boundary condition.*

This problem has applications to many problems in mathematical physics. For example, u can represent voltage in a conducting medium, in which case θ stands for electrical conductivity. The two types of boundary conditions stand for, respectively, a boundary point held at fixed voltage or a boundary point electrically insulated (or with prespecified nonzero current). Another application is thermodynamics, in which u stands for the temperature of the body, θ for thermal conductivity, Γ_1 for a boundary held at fixed temperature, and Γ_2 for a boundary insulated (or with prespecified heat flow). Problem (3.1) is also used to model membrane deflection and gravity. It arises as a subproblem in fluid flow modeling.

For clarity, we keep track of our assumptions by explicitly numbering them. One assumption has already been made.

ASSUMPTION 1. *For all* $x \in \Omega$, $\theta(x) > 0$.

Without this assumption, the problem may be ill-posed.

The first step in the isoparametric FEM is to produce a mesh of this domain. For the remainder of the paper, we assume isoparametric elements are defined with respect to the usual polynomial basis on a simplicial reference element, although the results can be generalized to other reference elements and basis families.

In more detail, let T_0 denote the standard unit d-simplex: for $d = 2$, this simplex is the triangle with vertices $(0,0)$, $(1,0)$, $(0,1)$, and for $d = 3$ this simplex is the tetrahedron with vertices $(0,0,0)$, $(1,0,0)$, $(0,1,0)$, $(0,0,1)$. Let p denote the polynomial order of the FEM. In the reference element, position an evenly spaced array of $l = (p+1)(p+2)/2$ *reference nodes* when $d = 2$ or of $l = (p+1)(p+2)(p+3)/6$ nodes when $d = 3$. These nodes have coordinates of the form $(i/p, j/p)$ (in two dimensions) or $(i/p, j/p, k/p)$ (in three dimensions), where i, j (and, in three dimensions, k) are nonnegative integers whose sum is at most p. Let these nodes be enumerated $z_1, \ldots, z_l$.

One generates a *mesh* of Ω composed of m elements to be defined via mapping functions. For each element $t = 1, \ldots, m$, there is a *mapping function* ϕ_t that maps T_0 to $\mathbb{R}^d$. Let $\nabla\phi_t$ denote the derivative (Jacobian) of ϕ_t. The following assumption is nearly universal in the literature.

ASSUMPTION 2. *Mapping function* $\phi_t : T_0 \to \mathbb{R}^d$ *is injective, and* $\det(\nabla\phi_t(z)) > 0$ *for all* $z \in T_0$, $t = 1, \ldots, m$.

The tth *element* is defined to be $\phi_t(T_0)$ and hence has a shape of a curved simplex. The function ϕ_t carries the l reference nodes $z_1, \ldots, z_l$ to l *real-space nodes* $\phi_t(z_1), \ldots, \phi_t(z_l)$, which are often just called *nodes*. These may be denoted $\zeta_{t,1}, \ldots, \zeta_{t,l}$. The nodes are chosen so that the nodes on the boundary of the tth element coincide with the corresponding nodes on the boundaries of its neighbors.

We restrict ϕ_t to be a polynomial of degree p, in which case it is uniquely determined by the positions of its real-space nodes. In more detail, let N_μ, $\mu = 1, \ldots, l$, be a real-valued degree-p polynomial function on T_0 with the property that

$$N_\mu(z_\nu) = \left\{ \begin{array}{ll} 1 & \text{if } \mu = \nu, \\ 0 & \text{if } \mu \neq \nu, \end{array} \right\} \text{ for all } \nu = 1, \ldots, l. \tag{3.2}$$

It is not hard to write down an explicit formula for N_μ and to prove that (3.2) uniquely determines N_μ among degree-p polynomials. These functions $N_1, \ldots, N_l$ are called *shape functions*. For each t, ϕ_t necessarily has the formula $\phi_t = \zeta_{t,1}N_1 + \cdots + \zeta_{t,l}N_l$.

There are many duplicate entries in the list $\zeta_{1,1}, \ldots, \zeta_{m,l}$ because of common nodes at the boundaries of the elements. Let $w_1, \ldots, w_{n'}$ be a listing of the real-space nodes with all duplicates removed. This renumbering is specified by an index mapping function, the *local-to-global* numbering map, denoted LG and carrying an index pair (t, μ) to an index $i \in 1, \ldots, n'$ so that $w_i \equiv \zeta_{t,\mu}$.

Finally, the *mesh* $\mathcal{T}$ is specified by listing the nodes $w_1, \ldots, w_{n'}$ and the local-global mapping defined by LG. From this data, one can deduce all of the ϕ_t's.

Let $\tilde{\Omega}$ be the union of the elements, which may be slightly different from Ω. This is because at the boundary of Ω, the elements may either protrude a bit outside Ω or may fail to cover a small part of the boundary. Let the boundary of $\tilde{\Omega}$ be partitioned into $\tilde{\Gamma}_1$ and $\tilde{\Gamma}_2$ in correspondence with the partition Γ_1, Γ_2 of the boundary of Ω. It is necessary to transfer θ (and several other fields) from Ω to $\tilde{\Omega}$ and to transfer the boundary conditions to $\tilde{\Gamma}_1$ and $\tilde{\Gamma}_2$. We omit the details.

The next step in the isoparametric FEM is to define *basis functions* $\pi_1, \ldots, \pi_{n'}$, which are functions from $\tilde{\Omega}$ to $\mathbb{R}$ satisfying

$$\pi_i(w_k) = \left\{ \begin{array}{ll} 1 & \text{if } i = k, \\ 0 & \text{if } i \neq k, \end{array} \right\} \text{for } k = 1, \ldots, n'. \tag{3.3}$$

We require that π_i, when restricted to an element t, must be of the form $N_\mu \circ \phi_t^{-1}$, where $LG(t, \mu) = i$, or else must be identically 0 if i does not occur among $LG(t, 1), \ldots, LG(t, l)$. This requirement uniquely determines the π_i's. It can be shown that π_i is continuous but fails to be differentiable at interelement boundaries.

Let the global numbering be chosen so that $w_1, \ldots, w_n$, the first n real-space nodes, are the nodes with no Dirichlet boundary condition imposed, and let the final $n' - n$ nodes have Dirichlet boundary conditions. Now we can define the *exact assembled stiffness matrix* K^{exact} of the FEM to be the $n \times n$ matrix whose (i, j) entry is given by

$$K^{\text{exact}}(i, j) = \int_{\tilde{\Omega}} \nabla \pi_i(x) \cdot \theta(x) \nabla \pi_j(x)\, dx, \tag{3.4}$$

where ∇, as usual, denotes the gradient.

Matrix K^{exact} is sparse, symmetric, and positive semidefinite. Symmetry is obvious; semidefiniteness follows from a fairly straightforward argument that we omit, and sparsity follows because $K^{\text{exact}}(i, j)$ is nonzero only if there is an element t that contains both nodes w_i and w_j.

Integral (3.4) is difficult to compute directly because evaluating π_i requires the evaluation of ϕ_t^{-1}. Fortunately, this difficulty is avoided by breaking the integral into a sum over elements, and then carrying out the integral over the reference domain following a change of variables as follows:

$$K^{\text{exact}}(i, j) = \sum_{t=1}^{m} \int_{T_0} \nabla_x \pi_i(\phi_t(z)) \cdot \theta(\phi_t(z)) \nabla_x \pi_j(\phi_t(z)) \det(\nabla \phi_t(z))\, dz, \tag{3.5}$$

where the notation ∇_x means derivative with respect to the coordinates of element t (as opposed to derivative with respect to z, the coordinates of T_0).

The integrand of (3.5) is evaluated using the chain rule for derivatives. Assume w_i is a node of element t (else the above integral is 0). Let μ be the index such that $LG(t, \mu) = i$, so that $\pi_i = N_\mu \circ \phi_t^{-1}$ on element t. Then

$$\begin{aligned} \nabla_x \pi_i(\phi_t(z)) &= \nabla_x N_\mu(z) \\ &= \nabla \phi_t(z)^{-T} \cdot \nabla_z N_\mu(z) \end{aligned} \tag{3.6}$$

and similarly for π_j. Here, $\nabla\phi_t(z)^{-T}$ denotes the transposed inverse of the $d\times d$ matrix $\nabla\phi_t(z)$, which exists by Assumption 2. The '$\cdot$' notation in the previous formula indicates matrix-vector multiplication. Gradients are regarded as length-d column vectors.

We assume that the entries of K are not the exact value of the integral (3.5) but are obtained by a quadrature rule that we now discuss. Let $r_1,\dots,r_q$ be points in the interior of the reference element T_0 called the *Gauss points*. (As is common practice, we use this terminology even if the quadrature rule is not derived from Gaussian quadrature.) Let $\omega_1,\dots,\omega_q$ be corresponding *Gauss weights*. We denote this quadrature rule, i.e., the set of ordered pairs $(r_1,\omega_1),\dots,(r_q,\omega_q)$, by the symbol $\mathcal{Q}$. Then in place of (3.5) we take

$$K(i,j)=\sum_{t=1}^{m}\sum_{k=1}^{q}\nabla_x\pi_i(\phi_t(r_k))\cdot\theta(\phi_t(r_k))\nabla_x\pi_j(\phi_t(r_k))\det(\nabla\phi_t(r_k))\omega_k, \tag{3.7}$$

in which $\nabla_x\pi_i(\phi_t(r_k))$ is evaluated by substituting $z=r_k$ into the right-hand side of (3.6) and similarly for $\nabla_x\pi_j(\phi_t(r_k))$. The symmetry and sparsity of K follow for the same reason as for K^{exact}; positive semidefiniteness will follow from results in the next two sections (under some additional assumptions to be made).

We close this section with a summary of the notation introduced thus far. Integers that define the size of the computation are as follows:

$$\begin{aligned}
d &= \text{the space dimension of (3.1)},\\
p &= \text{the polynomial order of the FEM},\\
l &= \text{the number of reference nodes} = \begin{cases}(p+1)(p+2)/2 & \text{if } d=2,\\ (p+1)(p+2)(p+3)/6 & \text{if } d=3,\end{cases}\\
m &= \text{the number of elements},\\
n &= \text{the number of non-Dirichlet real-space nodes},\\
n' &= \text{the total number of real-space nodes},\\
q &= \text{the number of Gauss points in the quadrature rule } \mathcal{Q}.
\end{aligned}$$

The sets of points in $\mathbb{R}^d$ include

$$\begin{aligned}
z_1,\dots,z_l &= \text{reference nodes},\\
\zeta_{1,1},\dots,\zeta_{m,l} &= \text{real-space nodes (local numbering)},\\
w_1,\dots,w_{n'} &= \text{real-space nodes (global numbering)}.
\end{aligned}$$

The quadrature rule $\mathcal{Q}$ is defined by

$$\begin{aligned}
r_1,\dots,r_q &= \text{Gauss points},\\
\omega_1,\dots,\omega_q &= \text{Gauss weights}.
\end{aligned}$$

Important domains are

$$\begin{aligned}
T_0 &= \text{the reference element},\\
\Omega &= \text{the domain of (3.1)},\\
\phi_1(T_0),\dots,\phi_m(T_0) &= \text{the elements},\\
\tilde{\Omega} &= \text{the approximation to } \Omega \text{ given by } \phi_1(T_0)\cup\cdots\cup\phi_m(T_0).
\end{aligned}$$

Important functions are

$$\begin{aligned}
\phi_1, \ldots, \phi_m &= \text{element mapping functions } (T_0 \to \tilde{\Omega}),\\
N_1, \ldots, N_l &= \text{shape functions } (T_0 \to \mathbb{R}),\\
\theta &= \text{conductivity } (\tilde{\Omega} \to \mathbb{R}),\\
LG &= \text{local-to-global index mapping } \{1, \ldots, m\} \times \{1, \ldots, l\} \to \{1, \ldots, n'\},\\
\pi_1, \ldots, \pi_{n'} &= \text{basis functions } (\tilde{\Omega} \to \mathbb{R}).
\end{aligned}$$

Finally, scalar quantities to be introduced in the next section, but which are included here for completeness, include

$$\begin{aligned}
m_{\mathcal{Q}}, M_{\mathcal{Q}} &= \text{min and max quadrature weights},\\
\hat{\theta}(\mathcal{T}) &= \text{intraelement conductivity variation},\\
\alpha_1, \ldots, \alpha_m &= \text{maximum compression in elements},\\
\beta_1, \ldots, \beta_m &= \text{maximum stretch in elements},\\
\kappa_1(\mathcal{T}), \kappa_2(\mathcal{T}) &= \text{mesh quality measures},\\
\sigma_{\mathcal{Q},p}, \tau_{\mathcal{Q},p} &= \text{the max and min singular values of } S_{\mathcal{Q},p}.
\end{aligned}$$

This section presented in detail the construction of the finite element stiffness matrix K, but it omitted the construction of $\mathbf{f}$, the right-hand side of the linear system $K\mathbf{x} = \mathbf{f}$ under consideration. The construction of $\mathbf{f}$ is not relevant to the main result of this paper, but for the sake of completeness, we give a very brief overview and refer the reader to [15] for the details. The steps involved in obtaining a formula for the ith entry $\mathbf{f}$ are as follows. Multiply the PDE on both sides by π_i, integrate over Ω, and apply integration by parts. Applying integration by parts causes a boundary term involving Neumann data g to enter the equation. Write u as a linear combination of π_j's for $j = 1, \ldots, n'$. The coefficients of this linear combination are the entries of $\mathbf{x}$ for $j = 1, \ldots, n$ and are known from the Dirichlet boundary condition for $j = n+1, \ldots, n'$. Finally, gather terms not involving entries of $\mathbf{x}$ together on the right-hand side and apply quadrature rules to obtain the ith entry of $\mathbf{f}$. Once the linear system $K\mathbf{x} = \mathbf{f}$ is solved, the approximation of the unknown function u of (3.1) is recovered as a linear combination of the π_j's using the entries of $\mathbf{x}$ and the Dirichlet conditions as coefficients.

4. Condition numbers and assumptions. Our main theorem about matrix approximation, which is Theorem 5.1 below, depends on several scalars associated with the FEM and the problem at hand that we define in this section. In addition, this section states further assumptions about the problem and method.

Two constants appearing in our bound are $m_{\mathcal{Q}}$ and $M_{\mathcal{Q}}$, which we define to be the minimum and maximum weights in the quadrature rule, respectively, i.e.,

$$m_{\mathcal{Q}} = \min(\omega_1, \ldots \omega_q);\ M_{\mathcal{Q}} = \max(\omega_1, \ldots, \omega_q). \tag{4.1}$$

Assumption 3. *The quadrature weights are positive, i.e., $m_{\mathcal{Q}} > 0$.*

There is some loss of generality with this assumption because a few popular finite element quadrature schemes (but certainly not all) use negative weights [8].

Assumption 4. *The quadrature scheme is exact for all polynomials of degree up to $2p-2$, i.e., if $\psi : T_0 \to \mathbb{R}$ is a polynomial of degree $2p-2$ or less in the d coordinates of T_0, then*

$$\sum_{k=1}^{q} \psi(r_k)\omega_k = \int_{T_0} \psi(z)\, dz.$$

This assumption is quite reasonable, since it is usually required anyway for an accurate solution by finite element analysis: one wants accurate quadrature of $\nabla\pi_i \cdot \nabla\pi_j$.

We now define a $dq \times (l-1)$ matrix $S_{\mathcal{Q},p}$ according to the following formula:

$$S_{\mathcal{Q},p} = \begin{pmatrix} \nabla N_2(r_1) & \nabla N_3(r_1) & \cdots & \nabla N_l(r_1) \\ \vdots & & & \vdots \\ \nabla N_2(r_q) & \nabla N_3(r_q) & \cdots & \nabla N_l(r_q) \end{pmatrix}. \tag{4.2}$$

Although we introduce this matrix in this section in order to define two associated scalars, the motivation for this definition will be postponed until section 6. Before defining these scalars, we require the following lemma.

LEMMA 4.1. *Under Assumption* 4, *matrix* $S_{\mathcal{Q},p}$ *has full column rank.*

Proof. Let $v \in \mathbb{R}^{l-1}$ be chosen so that $S_{\mathcal{Q},p}v = \mathbf{0}$. Define the function $U : \mathcal{T}_0 \to \mathbb{R}$ according to the formula $U = v_1N_2 + v_2N_3 + \cdots + v_{l-1}N_l$. Observe that, by the definition of $S_{\mathcal{Q},p}$, the first d entries of $S_{\mathcal{Q},p}v$ are precisely $\nabla U(r_1)$, and the next d entries are $\nabla U(r_2)$, etc. Therefore, ∇U vanishes identically at $r_1, \ldots, r_q$. Let $\psi : T_0 \to \mathbb{R}$ be defined by $\psi = \nabla U \cdot \nabla U$. Then ψ has degree at most $2p-2$, is a nonnegative function, and also vanishes identically at $r_1, \ldots, r_q$. By Assumption 4, this means that $\int_{T_0} \psi = 0$. But since ψ is nonnegative-valued, we conclude that $\psi \equiv 0$. Therefore, $\nabla U \equiv 0$ also, and thus U must be a constant function on T_0. But the definition of U omits the N_1 term, and therefore $U(z_1) = 0$. Since U is constant, this means that $U \equiv 0$ on all of T_0, and in particular, $U(z_2) = \cdots = U(z_l) = 0$. But $U(z_2) = v_1$, $U(z_3) = v_2, \ldots, U(z_l) = v_{l-1}$ by the construction of U. Therefore, the entries of v are all zeros. Since v was an arbitrary vector in $N(S_{\mathcal{Q},p})$, this argument proves that this nullspace contains only the $\mathbf{0}$ vector, hence $S_{\mathcal{Q},p}$ has full column rank. ∎

The next constants that appear in our theorem are $\sigma_{\mathcal{Q},p}$ and $\tau_{\mathcal{Q},p}$, defined as follows:

$$\sigma_{\mathcal{Q},p} = \sigma_{\max}(S_{\mathcal{Q},p});\ \tau_{\mathcal{Q},p} = \sigma_{\min}(S_{\mathcal{Q},p}). \tag{4.3}$$

These constants depend only on p and the quadrature scheme. It follows from the lemma that both are positive.

The remaining scalars and assumptions in this section pertain to the mesh. Define for $t = 1, \ldots, m$,

$$\alpha_t = \max\{\|\nabla\phi_t(r_1)^{-1}\|, \ldots, \|\nabla\phi_t(r_q)^{-1}\|\}, \tag{4.4}$$

$$\beta_t = \max\{\|\nabla\phi_t(r_1)\|, \ldots, \|\nabla\phi_t(r_q)\|\}. \tag{4.5}$$

The norms in these equations are matrix 2-norms: $\|B\| \equiv \sigma_{\max}(B)$. Finally, for the whole mesh, a quality measure is

$$\kappa_1(\mathcal{T}) = \max_{t=1,\ldots,m} \alpha_t\beta_t. \tag{4.6}$$

Although the α's and β's are subscripted only by t, it is clear from the definition that they also depend on $\mathcal{Q}$. On the other hand, it is possible to get a $\mathcal{Q}$-independent definition of these by simply taking the upper bounds similar to (4.4), (4.5) over all $z \in T_0$ instead of just $r_1, \ldots, r_q$. Unfortunately, for higher-order elements ($p > 1$), there is no simple method to compute $\max_{z\in T_0} \|\nabla\phi_t(z)^{-1}\|$; a technique for obtaining an upper bound appears in [21].

It should be noted that $\kappa_1(\mathcal{T}) \geq 1$, since for any z, $\|\nabla\phi_t(z)^{-1}\| \cdot \|\nabla\phi_t(z)\| \geq 1$. If all of the elements are well shaped, i.e., not too distorted when compared to the reference element, then $\kappa_1(\mathcal{T})$ will not be much larger than 1.

The second mesh quality measure is

$$\kappa_2(\mathcal{T}) = \max_{t=1,\ldots,m} \frac{\max_{k=1,\ldots,q} \det(\nabla \phi_t(r_k))}{\min_{k=1,\ldots,q} \det(\nabla \phi_t(r_k))}. \tag{4.7}$$

This quantity measures the maximum over elements of the variation in volumetric distortion over the element. This may be regarded as a measure of how much elements depart from linearity (flatness). Measure κ_2 is not completely independent from κ_1 as the following argument shows. It follows from Hadamard's inequality that

$$\alpha_t^{-d} \leq \det(\nabla \phi_t(r_k)) \leq \beta_t^d,$$

and therefore

$$\frac{\max_{k=1,\ldots,q} \det(\nabla \phi_t(r_k))}{\min_{k=1,\ldots,q} \det(\nabla \phi_t(r_k))} \leq (\alpha_t \beta_t)^d,$$

hence $\kappa_2(\mathcal{T}) \leq \kappa_1(\mathcal{T})^d$. This bound is not likely to be tight in practice (see our computation results in Table 10.3), so we prefer to distinguish the roles of κ_1 and κ_2.

The final two assumptions are qualitative in nature and are meant to indicate cases in which our method is expected to work well. The violation of these assumptions does not invalidate our theorem but merely indicates that our bound on $\kappa(K, \bar{K})$ may grow large.

ASSUMPTION 5. *Mesh quality measures $\kappa_1(\mathcal{T}), \kappa_2(\mathcal{T})$ are not too large.*

In the case of linear elements ($p = 1$), $\kappa_1(\mathcal{T})$ is proportional to the reciprocal of the minimum angle (in two dimensions) or the solid angle (in three dimensions) of the mesh, and $\kappa_2(\mathcal{T}) \equiv 1$.

Assumption 5 does not imply that the elements are of a uniform size: A uniform rescaling of element t does not affect the product $\alpha_t \beta_t$.

The final constant and assumption pertain to the conductivity field. Define

$$\hat{\theta}(\mathcal{T}) = \max_{t=1,\ldots,m} \frac{\max_{k=1,\ldots,q} \theta(\phi_t(r_k))}{\min_{k=1,\ldots,q} \theta(\phi_t(r_k))}. \tag{4.8}$$

In other words, $\hat{\theta}(\mathcal{T})$ measures the maximum intraelement variation of the conductivity field θ.

ASSUMPTION 6. *Intraelement conductivity variation $\hat{\theta}(\mathcal{T})$ is not too large.*

This assumption implies that if there are huge jumps in θ in the domain (e.g., because one is modeling a domain composed of two materials with vastly different conductivities), then the mesh boundaries should be aligned with the conductivity jumps. If the mesh boundaries are aligned with the conductivity jumps, then no element will have large variation in θ among its Gauss points, and thus $\hat{\theta}(\mathcal{T})$ will not be large.

Note that, as in the case of $\kappa_1(\mathcal{T})$ and $\kappa_2(\mathcal{T})$, scalar $\hat{\theta}(\mathcal{T})$ also depends on the quadrature rule, but this dependence can be eliminated by overestimating $\hat{\theta}(\mathcal{T})$ as

$$\max_{t=1,\ldots,m} \frac{\max_{z \in T_0} \theta(\phi_t(z))}{\min_{z \in T_0} \theta(\phi_t(z))}.$$

5. The matrix approximation. Our main matrix factorization result is summarized by the following theorem whose proof is explained in upcoming sections.

THEOREM 5.1. *Let K be defined by* (3.7) *above, and let Assumptions* 1–4 *hold. Then K may be factored as $A^T J^T D J A$, where*

- *A is an $(l-1)m \times n$ reduced node-arc incidence matrix of a certain multigraph,*
- *J is a $dqm \times (l-1)m$ matrix that is well conditioned in the sense that*

$$\sigma_{\max}(J) \leq \sigma_{\mathcal{Q},p}, \tag{5.1}$$

and

$$\sigma_{\min}(J) \geq \tau_{\mathcal{Q},p}/\kappa_1(\mathcal{T}). \tag{5.2}$$

- *D is a $dqm \times dqm$ positive definite diagonal matrix.*

Further, $J^T D J$ may be refactored as $\bar{D}^{1/2} H \bar{D}^{1/2}$, where $\bar{D}$ is a $(l-1)m \times (l-1)m$ SPD diagonal matrix and H is a $(l-1)m \times (l-1)m$ SPD matrix whose condition number is bounded as follows:

$$\kappa(H) \leq \hat{\theta}(\mathcal{T})\kappa_1(\mathcal{T})^2\kappa_2(\mathcal{T}) \cdot \frac{M_{\mathcal{Q}}\sigma_{\mathcal{Q},p}^2}{m_{\mathcal{Q}}\tau_{\mathcal{Q},p}^2}. \tag{5.3}$$

This theorem can now be combined with Theorem 2.3 to obtain a good approximation $\bar{K}$ to the stiffness matrix K. In particular, we take $\bar{K} = A^T \bar{D} A$. Note that this matrix, being a weighted graph Laplacian, is symmetric and diagonally dominant. Then in the context of Theorem 2.3, $V = A^T \bar{D}^{1/2}$. The theorem thus implies that the condition number of $\bar{K}$ with respect to K depends only on the condition number of H for which we have a good bound (5.3). Then $\bar{K}$ can be preconditioned using techniques in the previous literature. The complete description of the algorithm is given below in section 8.

The approximation bound clearly depends on the quadrature rule, mesh quality, and the intraelement variation of θ. It is also informative to make a list of quantities on which $\kappa(\bar{K}, K)$ does *not* depend:

- The number of nodes or elements,
- The size of the elements,
- Variation in the size of elements (i.e., the gradation of the mesh),
- The shape of Ω,
- The conductivity field θ (provided that the mesh respects internal boundaries where large conductivity jumps occur).

Although the shape of Ω does not directly affect $\kappa(\bar{K}, K)$, naturally, it affects the mesh generation procedure, which in turn affects mesh quality. A particular case that requires further consideration is when Ω has a sharp corner; a sharp corner of Ω forces the mesh to contain at least one element with a sharp corner, thus causing $\kappa_1(\mathcal{T})$ to be large. Even in this case, however, it may be possible to obtain an upper bound on $\kappa(\bar{K}, K)$ independent of this sharp angle using the analysis of Phillips and Miller (see section 9).

6. The first factorization. In this section we develop the first factorization of K stated in Theorem 5.1. This factorization is related to one proposed in [22]. We state the main result of this section as a lemma.

LEMMA 6.1. *Matrix K defined by* (3.7) *can be factored as $K = A^T S^T R^T D R S A$, where the factors are as follows.*

- *Define matrix $A \in \mathbb{R}^{(l-1)m \times n}$ to be a sparse matrix all of whose entries are -1, 0, or 1. In more detail, A is written in block form*

$$A = \begin{pmatrix} A_1 \\ \vdots \\ A_m \end{pmatrix},$$

where A_t is $(l-1) \times n$ for each element $t = 1, \ldots, m$. The columns are indexed $1, \ldots, n$ in correspondence with nodes $w_1, \ldots, w_n$. Row $\mu - 1$ of A_t has a "1" in column $LG(t, \mu)$ for $\mu = 2, \ldots, l$ and a "-1" in column $LG(t, 1)$. Thus, most rows of A have exactly two nonzero entries. If $LG(t, \mu) > n$ (i.e., node $\zeta_{t,\mu}$ lies in the Dirichlet boundary $\tilde{\Gamma}_1$), then the "1" entry is omitted. Similarly, if $LG(t, 1) > n$, then the "-1" entry is omitted. For this reason, a few rows of A have just one nonzero entry or none at all.

- *Define $S \in \mathbb{R}^{dqm \times (l-1)m}$ by*

$$S = \left(\begin{array}{ccc} S_{\mathcal{Q},p} & & \\ & \ddots & \\ & & S_{\mathcal{Q},p} \end{array} \right) \Bigg\} m \text{ times}, \tag{6.1}$$

where $S_{\mathcal{Q},p}$ was defined by (4.2).

- *Define block diagonal $R \in \mathbb{R}^{dqm \times dqm}$ by*

$$R = \left(\begin{array}{ccc} R_1 & & \\ & \ddots & \\ & & R_m \end{array} \right), \tag{6.2}$$

where each of $R_1, \ldots, R_m \in \mathbb{R}^{qd \times qd}$ is itself block diagonal and given by

$$R_t = \alpha_t^{-1} \left(\begin{array}{ccc} \nabla \phi_t(r_1)^{-T} & & \\ & \ddots & \\ & & \nabla \phi_t(r_q)^{-T} \end{array} \right). \tag{6.3}$$

The scalar α_t used here was defined by (4.4).

- *Finally, $D \in \mathbb{R}^{qdm \times qdm}$ is positive definite diagonal and given by*

$$D = \left(\begin{array}{ccc} D_1 & & \\ & \ddots & \\ & & D_m \end{array} \right), \tag{6.4}$$

where $D_t \in \mathbb{R}^{qd \times qd}$, $t = 1, \ldots, m$, is given by

$$D_t = \alpha_t^2 \left(\begin{array}{ccc} \theta(\phi_t(r_1)) \det(\nabla \phi_t(r_1)) \omega_1 I & & \\ & \ddots & \\ & & \theta(\phi_t(r_q)) \det(\nabla \phi_t(r_q)) \omega_q I \end{array} \right), \tag{6.5}$$

in which I denotes the $d \times d$ identity matrix.

Remarks. Intuitively, this factorization of K decomposes finite element analysis into natural ingredients: A encodes the combinatorial connectivity of the mesh, R encodes the geometry of the mesh, S encodes the quadrature points, and D encodes the quadrature weights and conductivity. The scaling factor α_t, which is necessary for our analysis, cancels out between D_t and R_t.

Note that the positive definiteness of D follows from Assumptions 1, 2, and 3. Note also that A is a reduced node-arc incidence matrix of a multigraph defined on the nodes of $\mathcal{T}$. (*Multi*graph indicates that more than one edge may connect a particular pair of vertices.) Each element gives rise to $l - 1$ arcs in the graph. In particular, for

each element $t = 1, \ldots, m$, there is an arc joining each of its nodes $2, \ldots, l$ to node 1. Columns corresponding to nodes of $\tilde{\Gamma}_1$ are omitted. (This is what is meant by "reduced.")

Proof. Let $\boldsymbol{\xi}$ be an arbitary vector in $\mathbb{R}^n$, and let $u : \tilde{\Omega} \to \mathbb{R}$ be defined by

$$u = \sum_{i=1}^{n} \xi_i \pi_i.$$

For an element t, define function $U_t : T_0 \to \mathbb{R}$ by

$$U_t = \sum_{\mu=1}^{l} \xi_{LG(t,\mu)} N_\mu. \tag{6.6}$$

With these definitions, it is clear that $u \circ \phi_t = U_t$. In (6.6), we follow the convention that $\xi_i \equiv 0$ in the case that $i \in \{n+1, \ldots, n'\}$.

The assembled stiffness matrix is defined by (3.7) so that

$$\begin{aligned}
\boldsymbol{\xi}^T K \boldsymbol{\xi} &= \sum_{i=1}^{n}\sum_{j=1}^{n} \xi_i \xi_j K(i,j) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} \xi_i \xi_j \sum_{t=1}^{m}\sum_{k=1}^{q} \nabla_x \pi_i(\phi_t(r_k)) \cdot \theta(\phi_t(r_k)) \nabla_x \pi_j(\phi_t(r_k)) \det(\nabla \phi_t(r_k)) \omega_k \\
&= \sum_{t=1}^{m}\sum_{k=1}^{q}\sum_{i=1}^{n} \xi_i \nabla_x \pi_i(\phi_t(r_k)) \cdot \theta(\phi_t(r_k)) \sum_{j=1}^{n} \xi_j \nabla_x \pi_j(\phi_t(r_k)) \det(\nabla \phi_t(r_k)) \omega_k \\
&= \sum_{t=1}^{m}\sum_{k=1}^{q} \nabla_x u(\phi_t(r_k)) \cdot \theta(\phi_t(r_k)) \nabla_x u(\phi_t(r_k)) \det(\nabla \phi_t(r_k)) \omega_k \\
&= \sum_{t=1}^{m}\sum_{k=1}^{q} \nabla_x U_t(r_k) \cdot \theta(\phi_t(r_k)) \nabla_x U_t(r_k) \det(\nabla \phi_t(r_k)) \omega_k \\
&= \sum_{t=1}^{m}\sum_{i=1}^{q} (\nabla \phi_t(r_k)^{-T} \nabla_z U_t(r_k)) \cdot \theta(\phi_t(r_k)) \\
&\qquad (\nabla \phi_t(r_k)^{-T} \nabla_z U_t(r_k)) \det(\nabla \phi_t(r_k)) \omega_k \qquad (6.7) \\
&= \mathbf{v}^T D \mathbf{v}. \qquad (6.8)
\end{aligned}$$

In (6.8), we have used the matrix D defined by (6.4). We have also introduced the vector $\mathbf{v} \in \mathbb{R}^{dqm}$ defined in block fashion as follows. Write $\mathbf{v} = [v_1; \cdots; v_m]$, where $v_t \in \mathbb{R}^{dq}$ is itself composed of blocks $v_t = [v_{t,1}; \cdots; v_{t,q}]$. Here $v_{t,k} \in \mathbb{R}^d$ is defined to be

$$v_{t,k} = (\nabla \phi_t(r_k))^{-T} \nabla_z U_t(r_k) / \alpha_t. \tag{6.9}$$

It is clear by the construction of D and $\mathbf{v}$ that $\mathbf{v}^T D \mathbf{v}$ is equal to the expression in (6.7).

Next, we claim that $\mathbf{v} = RSA\boldsymbol{\xi}$. Focus on the block corresponding to a particular element $t \in \{1, \ldots, m\}$, in which case we must show that $v_t = R_t S_{\mathcal{Q},p} A_t \boldsymbol{\xi}$. The product $A_t \boldsymbol{\xi}$ yields a vector with $(l-1)$ entries that contains finite differences of the entries of $\boldsymbol{\xi}$. Specifically, the $\mu - 1$ entry is $\xi_{LG(t,\mu)} - \xi_{LG(t,1)}$ for $\mu = 2, \ldots, l$.

By the definition of $S_{\mathcal{Q},p}$ in (4.2), it follows that the k-block of entries ($k = 1, \ldots, q$) of $S_{\mathcal{Q},p} A_t \boldsymbol{\xi}$ is the d-vector $\nabla_z \bar{U}_t(r_k)$, where

$$\bar{U}_t = (\xi_{LG(t,2)} - \xi_{LG(t,1)})N_2 + \cdots + (\xi_{LG(t,l)} - \xi_{LG(t,1)})N_l.$$

Comparing this equation to (6.6) indicates that $\bar{U}_t$ and U_t differ by

$$\xi_{LG(t,1)}(N_1 + \cdots + N_l).$$

This latter quantity, however, is a constant function (because $N_1 + \cdots + N_l$ is identically 1, a property that follows from (3.2) and the fact that 1 is a polynomial of degree at most p and hence must be expressable as a sum of N_μ's). Therefore, U_t and $\bar{U}_t$ have the same gradient. We conclude that the k block of $S_{\mathcal{Q},p} A_t \boldsymbol{\xi}$ must equal $\nabla_z U_t(r_k)$.

Combining this equation with (6.3) shows that

$$R_t S_{\mathcal{Q},p} A_t \boldsymbol{\xi} = (\nabla \phi_t(r_k))^{-T} \nabla_z U_t(r_k)/\alpha_t$$

and hence is equal to $v_{t,k}$ as defined by (6.9). This concludes the proof that $\mathbf{v} = RSA\boldsymbol{\xi}$.

Since $\mathbf{v}^T D \mathbf{v} = \boldsymbol{\xi}^T K \boldsymbol{\xi}$, this means that $\boldsymbol{\xi}^T A^T S^T R^T D R S A \boldsymbol{\xi} = \boldsymbol{\xi}^T K \boldsymbol{\xi}$. Since this holds for all $\boldsymbol{\xi} \in \mathbb{R}^n$ and since a symmetric matrix C is uniquely determined by the mapping $\boldsymbol{\xi} \mapsto \boldsymbol{\xi}^T C \boldsymbol{\xi}$, this means that $K = A^T S^T R^T D R S A$, which concludes the proof of the lemma. ∎

Now we analyze the singular values of J to finish the first part of Theorem 5.1. Let us define $J = RS$ so that $K = A^T J^T D J A$ as claimed. The block structures of R and S induce a corresponding block structure on J:

$$J = \begin{pmatrix} J_1 & & \\ & \ddots & \\ & & J_m \end{pmatrix},$$

where

$$J_t = R_t S_{\mathcal{Q},p} \tag{6.10}$$

for all $t = 1, \ldots, m$. Because of this structure, the maximum singular value of J is the maximum singular value among any of its blocks and similarly for its minimum singular value. Since in general $\sigma_{\max}(AB) \le \sigma_{\max}(A)\sigma_{\max}(B)$,

$$\begin{aligned} \sigma_{\max}(J_t) &\le \sigma_{\max}(R_t)\sigma_{\max}(S_{\mathcal{Q},p}) \\ &= \sigma_{\max}(\operatorname{diag}(\nabla \phi_t(r_1)^{-T}, \ldots, \nabla \phi_t(r_q)^{-T}))\sigma_{\mathcal{Q},p}/\alpha_t \\ &= \max_{k=1,\ldots,q} \sigma_{\max}(\nabla \phi_t(r_k)^{-1})\sigma_{\mathcal{Q},p}/\alpha_t \\ &= \sigma_{\mathcal{Q},p}. \end{aligned} \tag{6.11}$$

The last line follows from (4.4) and establishes (5.1).

Since $\sigma_{\min}(AB) \ge \sigma_{\min}(A)\sigma_{\min}(B)$ for two matrices A, B with full column rank,

$$\begin{aligned} \sigma_{\min}(J_t) &\ge \sigma_{\min}(R_t)\sigma_{\min}(S_{\mathcal{Q},p}) \\ &= \sigma_{\min}(\operatorname{diag}(\nabla \phi_t(r_1)^{-T}, \ldots, \nabla \phi_t(r_q)^{-T}))\tau_{\mathcal{Q},p}/\alpha_t \\ &= \min_{k=1,\ldots,q} \sigma_{\min}(\nabla \phi_t(r_k)^{-1})\tau_{\mathcal{Q},p}/\alpha_t \\ &= \min_{k=1,\ldots,q} (1/\sigma_{\max}(\nabla \phi_t(r_k)))\tau_{\mathcal{Q},p}/\alpha_t \\ &= (1/\beta_t) \cdot \tau_{\mathcal{Q},p}/\alpha_t \\ &\ge \tau_{\mathcal{Q},p}/\kappa_1(\mathcal{T}). \end{aligned} \tag{6.12}$$

For the last line, we used the fact that $\alpha_t \beta_t \leq \kappa_1(\mathcal{T})$, which follows from (4.6). This establishes (5.2) and concludes the proof of the first factorization in Theorem 5.1.

Our factorization $K = A^T J^T D J A$ is reminiscent of one proposed by Argyris [1] of the form $K = \tilde{A} P \tilde{A}^T$, which he calls the "natural factorization." In Argyris's factorization, however, the matrix $\tilde{A}$ has all $+1$ and 0 entries and therefore is not a node-arc incidence matrix. The purpose of Argyris's matrix $\tilde{A}$ is to assemble the element stiffness matrices, which constitute the blocks of the block-diagonal matrix P.

7. The second factorization. In this section we prove the second part of Theorem 5.1. We state the existence of the second factorization in the form of a lemma.

LEMMA 7.1. *Let J and D be defined as in Lemma 6.1. Then $J^T D J$ can be refactored as $\bar{D}^{1/2} \bar{J}^T \bar{J} \bar{D}^{1/2}$, where*

- *Matrix $\bar{D} \in \mathbb{R}^{m(l-1) \times m(l-1)}$ is positive definite diagonal and is written in block form*

$$\bar{D} = m_{\mathcal{Q}} \begin{pmatrix} f_1 g_1 \alpha_1^2 I & & \\ & \ddots & \\ & & f_m g_m \alpha_m^2 I \end{pmatrix}, \tag{7.1}$$

where I is the $(l-1) \times (l-1)$ identity matrix. In this formula, f_t is the minimum value of θ over the Gauss points of element t:

$$f_t = \min_{k=1,\ldots,q} \theta(\phi_t(r_k)), \tag{7.2}$$

and similarly, g_t is the minimum value of $\det(\nabla \phi_t)$*:*

$$g_t = \min_{k=1,\ldots,q} \det(\nabla \phi_t(r_k)). \tag{7.3}$$

- *Matrix $\bar{J} \in \mathbb{R}^{dqm \times (l-1)m}$ is defined by*

$$\bar{J} = D^{1/2} J \bar{D}^{-1/2}. \tag{7.4}$$

Proof. The fact that $\bar{D}^{1/2} \bar{J}^T \bar{J} \bar{D}^{1/2} = J^T D J$ follows as an immediate consequence of (7.4) regardless of how we have defined $\bar{D}$. ∎

Next we analyze the condition number of H to finish proving the second part of Theorem 5.1. We define $H = \bar{J}^T \bar{J}$ so that $J^T D J = \bar{D}^{1/2} H \bar{D}^{1/2}$. We must estimate the singular values of $\bar{J}$, which are the square roots of the eigenvalues of H.

Let the diagonal block of $\bar{J}$ associated with element t be denoted $\bar{J}_t$ for $t = 1, \ldots, m$. Because of the block structure, the maximum and minimum singular values for $\bar{J}$ are the maximum and minimum singular values among the blocks $\bar{J}_t$, which may be written as

$$\begin{aligned} \bar{J}_t &= D_t^{1/2} J_t f_t^{-1/2} g_t^{-1/2} m_{\mathcal{Q}}^{-1/2} \alpha_t^{-1} \\ &= \hat{D}_t J_t, \end{aligned}$$

where

$$\hat{D}_t = f_t^{-1/2} g_t^{-1/2} m_{\mathcal{Q}}^{-1/2} \alpha_t^{-1} D_t^{1/2}.$$

Examining the constituent parts of D_t given by (6.5), observing that the α_t^2 in (6.5) is cancelled by the α_t^{-1} in the preceding equation, and applying the inequalities

$$1 \leq \omega_k / m_{\mathcal{Q}} \leq M_{\mathcal{Q}} / m_{\mathcal{Q}}$$

for $k = 1, \ldots, q$ (see (4.1)),

$$1 \le \theta(\phi_t(r_k)) f_t^{-1} \le \frac{\max_{k=1,\ldots,q} \theta(\phi_t(r_k))}{\min_{k=1,\ldots,q} \theta(\phi_t(r_k))}$$

(see (7.2)), and

$$1 \le \det(\nabla \phi_t(r_k)) g_t^{-1} \le \frac{\max_{k=1,\ldots,q} \det(\nabla \phi_t(r_k))}{\min_{k=1,\ldots,q} \det(\nabla \phi_t(r_k))}$$

(see (7.3)), we conclude that the diagonal entries of $\hat{D}_t$ satisfy

$$1 \le \hat{D}_t(i,i) \le \left(\frac{\max_k \theta(\phi_t(r_k))}{\min_k \theta(\phi_t(r_k))} \right)^{1/2} \left(\frac{\max_k \det(\nabla \phi_t(r_k))}{\min_k \det(\nabla \phi_t(r_k))} \right)^{1/2} \left(\frac{M_{\mathcal{Q}}}{m_{\mathcal{Q}}} \right)^{1/2}.$$

By (4.7) and (4.8), the first two quantities on the right-hand side of the preceding equation are bounded by $\hat{\theta}(\mathcal{T})^{1/2}$ and $\kappa_2(\mathcal{T})^{1/2}$, respectively. Thus, it is apparent that, for each i,

$$1 \le D_t(i,i) \le (\hat{\theta}(\mathcal{T}) \kappa_2(\mathcal{T}) M_{\mathcal{Q}} / m_{\mathcal{Q}})^{1/2}.$$

Since $\bar{J}_t = \hat{D}_t J_t$, we can combine the inequalities in the previous line with (6.11) and (6.12) to obtain the following:

$$\sigma_{\max}(\bar{J}_t) \le \hat{\theta}(\mathcal{T})^{1/2} \kappa_2(\mathcal{T})^{1/2} M_{\mathcal{Q}}^{1/2} \sigma_{\mathcal{Q},p} / m_{\mathcal{Q}}^{1/2} \tag{7.5}$$

and

$$\sigma_{\min}(\bar{J}_t) \ge \tau_{\mathcal{Q},p} / \kappa_1(\mathcal{T}). \tag{7.6}$$

Since $H = \bar{J}^T \bar{J}$ and $\bar{J}$ has full column rank, (7.5) and (7.6) prove (5.3), which concludes the proof of Theorem 5.1.

8. Preconditioning strategy summary. The main result of this paper is that the stiffness matrix K of (3.1) can be approximated by a symmetric diagonally dominant matrix $\bar{K}$. In this section we discuss several approaches for using this approximation to efficiently solve the linear system $K\mathbf{x} = \mathbf{f}$ by iteration.

The first step in using the approximation is to construct $\bar{K}$. Recall that $\bar{K} = A^T \bar{D} A$, and thus A and $\bar{D}$ must be formed. Forming A means the construction of the multigraph consisting of $l - 1$ edges (a star-tree) per element of $\mathcal{T}$. Matrix A is then the reduced node-arc incidence matrix of this multigraph as specified by Lemma 6.1. The construction of $\bar{D}$ is given by (7.1). Notice that computing A and $\bar{D}$ requires information about the mesh and the original boundary value problem. In other words, our method is not applicable (at least not in a straightforward manner) if the only information about the original problem is the stiffness matrix K.

Once $\bar{K}$ is on hand, there are several ways to proceed. The most straightforward is to suppose that one has an efficient preconditioned conjugate gradient solver for systems of the form $\bar{K}\mathbf{x} = \mathbf{f}$. Efficient preconditioners for symmetric diagonally dominant systems were proposed and analyzed in a graph-theoretic framework by Vaidya. Vaidya's work is described and extended by [4, 6], and [11] contains a related analysis. One of the most recent improvements is due to Spielman and Teng [20], who propose a graph-based preconditioner whose running time is $O(n^{5/4})$. (Spielman and Teng

improve this bound to $O(n^{1+\epsilon})$ for any $\epsilon > 0$, but the algorithm corresponding to the improved bound is no longer preconditioned conjugate gradients.) Other techniques proposed in the literature for symmetric diagonally dominant matrices such as algebraic multigrid could also be used.

If one has a good preconditioner M for $\bar{K}$, then one could also use M as a preconditioner directly for K. This is because of the "triangle inequality" (see [5]), which states that

$$\sigma(K, M) \leq \sigma(K, \bar{K})\sigma(\bar{K}, M). \tag{8.1}$$

This shows that the the overall support number is bounded by the product of the support numbers in each step of the approximation $K \approx \bar{K} \approx M$. In particular. $\kappa(M, K) \leq \kappa(H)\kappa(M, \bar{K})$. Since the number of iterations of preconditioned conjugate gradients is bounded by the square root of the condition number, our analysis shows that the increase in the number of iterations for solving $K\mathbf{x} = \mathbf{f}$ (compared to solving $\bar{K}\mathbf{x} = \mathbf{f}$) is at most a factor of $\sqrt{\kappa(H)}$, which in turn is bounded by

$$\hat{\theta}(\mathcal{T})^{1/2}\kappa_1(\mathcal{T})\kappa_2(\mathcal{T})^{1/2}M_{\mathcal{Q}}\sigma_{\mathcal{Q},p}/(m_{\mathcal{Q}}\tau_{\mathcal{Q},p}).$$

The asymptotically fastest known iterative algorithm for solving symmetric diagonally dominant matrices is due to Elkin et al. [9]. This algorithm, which we denote EEST, extends Spielman and Teng [20] and requires $O(n_e(\log n)^s)$ time to solve any diagonally dominant $\bar{K}\mathbf{x} = \mathbf{f}$, where n_e is the number of nonzero entries in $\bar{K}$ and s is some constant. Our preconditioner $\bar{K}$ is the Laplacian of a graph with ml edges, where m is the number of elements and l is the number of nodes in the reference element (refer to Lemma 6.1), hence $n_e = ml$. We may assume that $m = O(n)$ for the following reasons. First, this relationship holds for all automatic finite element mesh generators of which we are aware. Second, a violation of the relationship $m = O(n)$ implies that the mesh has elements of unbounded aspect ratio, which in turn means $\kappa_1(\mathcal{T})$ tends to infinity so that our results are not as useful anyway. Thus, it is safe to assume that $n_e \leq \text{const} \cdot nl$, and we are regarding l (which depends on p, the polynomial degree, and d, the space dimension) as a constant.

The EEST iteration is more complicated than preconditioned conjugate gradients, and hence it is unclear whether it is possible to simply replace the diagonally dominant coefficient matrix $\bar{K}$ by our stiffness matrix K and expect the algorithm to converge in $O(n(\log n)^s\kappa(H)^{1/2})$ time.

On the other hand, one could obtain this running time $O(n(\log n)^s\kappa(H)^{1/2})$ by using a nested iteration: The outer iteration is preconditioned conjugate gradients in which $\bar{K}$ preconditions K. Thus, $O(\kappa(H)^{1/2})$ outer iterations are required. The inner loop is the EEST algorithm to apply the preconditioner, i.e., to solve $\bar{K}\mathbf{x} = \mathbf{f}$ in $O(n(\log n)^s)$ iterations. The EEST algorithm is fairly complex, and using it in a two-level manner would raise a number of difficulties associated with termination tests, so it is unlikely to be practical currently.

This bound of $O(n(\log n)^s\kappa(H)^{1/2})$ on the number of operations is independent of the condition number of the underlying system because, as noted above, $\kappa(H)$ depends on factors associated with the FEM and on mesh quality measures but not on the conditioning of K.

Our approximation scheme can be rewritten on an element-by-element basis as follows. For some element index $t \in \{1, \ldots, m\}$, consider its element stiffness matrix $K_t = A_t^T J_t^T D_t J_t A_t$, where A_t is defined in Lemma 6.1, J_t is defined by (6.10), and D_t is defined by (6.5). The above proof shows that $\bar{D}_t$, which is the tth block of

(7.1), is a good approximation to $J_t^T D_t J_t$. Thus, we let $\bar{K}_t = A_t^T \bar{D}_t A_t$. The overall approximation is $\bar{K} = \sum_{t=1}^m \bar{K}_i$. Note that the so-called splitting lemma frequently used in support-graph theory (see, e.g., [5]) shows that

$$\sigma(K, \bar{K}) \leq \max(\sigma(K_1, \bar{K}_1), \ldots, \sigma(K_m, \bar{K}_m)),$$

so it suffices to analyze the quality of the approximation of $\bar{K}_t$ to K_t.

Alternatively, we can take a global view as in the above analysis and write $\bar{K} = A^T \bar{D} A$, where $A = (A_1; A_2; \ldots; A_m)$ and $\bar{D} = \mathrm{diag}(\bar{D}_1, \bar{D}_2, \ldots, \bar{D}_m)$. To simplify notation, we have adopted the global view in this paper, but the reader should keep in mind that our approximation can take place element by element, which may be important in an implementation.

9. Notes added in revision. Since the first version of this paper, there have been three related developments of interest. First, R. Gupta's Master's thesis [12] gave a geometric method to approximate element stiffness matrices by diagonally dominant matrices in the $p = 1$, $d = 2$ case. In follow-up work, Wang and Sarin [23] showed how to parallelize the algorithm. Computational experiments are presented. For this particular case $p = 1, d = 2$, their result is essentially equivalent to ours.

Second, Phillips and Miller [18] have shown that, in the $p = 1$, $d = 2$ case, $\kappa(H)$ is still small even if some elements have very small angles, provided that no element has large angles (close to π). In our preceding analysis, the right-hand side of (5.3) would grow large in the presence of small angles, since the factor $\kappa_1(\mathcal{T})$ would be large. Wang and Sarin make a similar observation.

In a third development, Avron et al. [2] have shown how to get an approximation to element stiffness matrices that is optimal up to a constant factor. In other words, for an element stiffness matrix K_t, they find a diagonally dominant matrix $\tilde{K}_t$ such that $\kappa(K_t, \tilde{K}_t) \leq c_1 \kappa(K_t, K_t')$, where K_t' is any other symmetric diagonally dominant matrix of the correct size and c_1 depends only on p, d. Thus, their preconditioner could lead to a faster algorithm than ours, since ours is not optimal in this sense. Their theory, however, does not subsume ours, since they do not obtain any new bounds on $\kappa(K, \tilde{K})$ comparable to (5.3). Their results are competitive and in some cases better than an algebraic multgrid method.

10. Computational tests. There are at least two possible ways to test the effectiveness of our result: Calculate the quality of matrix approximation or measure the speed of an iterative method. Since our theory is primarily about the former issue and since Avron et al. conduct extensive testing on the latter question, we focus on the first question.

We try out four specific commonly occurring cases: $(d, p) = (2, 1), (2, 2), (3, 1)$, and $(3, 2)$, i.e., linear and quadratic triangles and linear and quadratic tetrahedra. Six test meshes were tried, which we denote $\mathbf{T}_1, \ldots, \mathbf{T}_6$; the first two are two-dimensional and the last four are three-dimensional. For quadrature in the $p = 1$ cases, we use the midpoint rule. For the $p = 2$ cases, we use the symmetric $d + 1$-point rule, which is accurate for polynomials up to degree 2. Table 10.1 gives the quadrature rules used.

Table 10.2 gives the values of relevant scalars for these test cases. It should be noted that $\sigma_{\mathcal{Q},p} = \tau_{\mathcal{Q},p} = 1$ for the $p = 1$ cases under the midpoint rule because, in these cases, $S_{\mathcal{Q},p}$ given by (4.2) turns out to be the identity matrix.

Finally, we can tabulate element stiffness approximation bounds. (As mentioned in section 8, it suffices to measure element stiffness approximation, which is much easier to compute than global stiffness approximation.) There are at least three relevant

TABLE 10.1
Quadrature rules used for the six test cases. The rules are written in the form $(r_1, \omega_1), \ldots, (r_q, \omega_q)$*, where* r_k *is a Gauss point in the reference simplex and* ω_k *is its weight. In this table* $\xi_1 = (10 - \sqrt{20})/40$ *and* $\xi_2 = 1 - 3\xi_1$.

Test	$\mathcal{Q}$
$\mathbf{T}_1$	$((1/3, 1/3), 1/2)$
$\mathbf{T}_2$	$((1/6, 1/6), 1/6), ((1/6, 2/3), 1/6), ((2/3, 1/6), 1/6)$
$\mathbf{T}_3, \mathbf{T}_5$	$((1/4, 1/4, 1/4), 1/6)$
$\mathbf{T}_4, \mathbf{T}_6$	$((\xi_1, \xi_1, \xi_1), 1/24), ((\xi_1, \xi_1, \xi_2), 1/24), ((\xi_1, \xi_2, \xi_1), 1/24), ((\xi_2, \xi_1, \xi_1), 1/24)$

TABLE 10.2
Scalars relevant for test cases.

Test	d	p	q	$\sigma_{\mathcal{Q},p}$	$\tau_{\mathcal{Q},p}$	$M_{\mathcal{Q}}/m_q$
$\mathbf{T}_1$	2	1	1	1.00	1.00	1.0
$\mathbf{T}_2$	2	2	3	5.26	0.83	1.0
$\mathbf{T}_3, \mathbf{T}_5$	3	1	1	1.00	1.00	1.0
$\mathbf{T}_4, \mathbf{T}_6$	3	2	4	6.47	0.63	1.0

quantities to tabulate: $\chi_1 = \kappa(K_t, \bar{K}_t)$, $\chi_2 = \kappa(H)$, and $\chi_3 = \hat{\theta}(\mathcal{T})\kappa_1(\mathcal{T})^2\kappa_2(\mathcal{T}) \cdot \frac{M_{\mathcal{Q}}\sigma^2_{\mathcal{Q},p}}{m_{\mathcal{Q}}\tau^2_{\mathcal{Q},p}}$, which is the right-hand side of (5.3). Here, K_t is the stiffness matrix of element t ($t = 1, \ldots, m$), and $\bar{K}_t$ is the diagonally dominant preconditioner given by $A_t^T \bar{D}_t A_t$. It follows from Theorem 2.3 and (5.3) that $\chi_1 \le \chi_2 \le \chi_3$.

We generate six meshes and take the max of χ_1, χ_2, χ_3 over all of the elements of the mesh. For conductivity, we use $\theta \equiv 1$ so that the factor $\hat{\theta}(\mathcal{T})$ does not enter the bound. The meshes are generated as follows. In two dimensions, we use the Triangle mesh generator [19] to generate a mesh of an annulus, which is $\mathbf{T}_1$ and consists of linear ($p = 1$) elements. To generate $\mathbf{T}_2$, a quadratic mesh, we insert midpoint nodes in every element of $\mathbf{T}_1$, and for the edges on the boundaries, we moved the midpoint nodes onto the boundary.

The remaining four meshes are three-dimensional. For $\mathbf{T}_3$, we used the QMG mesh generator [17] to generate a mesh of a unit ball in $\mathbb{R}^3$. QMG's meshes are intended for the $p = 1$ case. For $\mathbf{T}_4$, we use midpoint insertion on the $\mathbf{T}_3$ mesh to obtain a $p = 2$ mesh. Again, the midpoint nodes of edges whose endpoints were on the boundary were displaced outward onto the boundary. For $\mathbf{T}_5$ and $\mathbf{T}_6$, we again mesh a unit ball (actually, one octant of the ball) using a mesh generator tailored for that domain only. The mesh generator for $\mathbf{T}_5$ first generates a highly regular mesh of a unit tetrahedron, and then it projects the mesh nodes radially outward toward one facet so that an octant of the unit ball is covered. Finally, $\mathbf{T}_6$ is obtained from $\mathbf{T}_5$ using midpoint insertion and displacement at the boundary.

Then, for every element in each mesh, we compute the three quantities χ_1, χ_2, and χ_3. We have tabulated the maximum values in Table 10.3. The undesirably large values of χ_1 for $\mathbf{T}_3$ and $\mathbf{T}_4$ appear to be due primarily to poorly shaped elements, i.e., a large value of $\kappa_1(\mathcal{T})$. This is evident from comparing $\mathbf{T}_3$ and $\mathbf{T}_4$ against $\mathbf{T}_5$ and $\mathbf{T}_6$, two meshes for the same domain but with much better shaped tetrahedra.

The value of $\kappa_2(\mathcal{T})$ was quite small in all of the tests; as mentioned earlier, it is identically 1 for the $p = 1$ cases. For the $p = 2$ cases, interior elements are still linear, and elements adjacent to the boundary are fairly close to linear. It is interesting to note that $\chi_1 = \chi_2$ in all cases. This implies that the invariant subspaces of H corresponding to its extremal eigenvalues meet the range space of $\bar{D}^{1/2}A$. We do not have a deeper explanation for this observation.

TABLE 10.3
An approximation of the element stiffness matrix by a diagonally dominant matrix.

Test	m	n	$\kappa_1(\mathcal{T})$	$\kappa_2(\mathcal{T})$	χ_1	χ_2	χ_3
$\mathbf{T}_1$	234	143	4.1	1.0	16.6	16.6	16.6
$\mathbf{T}_2$	234	520	4.7	1.3	185.1	185.1	1117.2
$\mathbf{T}_3$	9078	1913	59.6	1.0	3549.5	3549.5	3549.5
$\mathbf{T}_4$	9078	13608	90.2	2.7	$5.96 \cdot 10^4$	$5.96 \cdot 10^4$	$2.32 \cdot 10^6$
$\mathbf{T}_5$	729	220	5.0	1.0	24.9	24.9	24.9
$\mathbf{T}_6$	729	1330	5.0	1.2	396.6	396.6	2644.5

Although an iterative method was not tested, some conclusions can still be drawn from Table 10.3. For example, if our approximation were used on $\mathbf{T}_6$ in an iterative setting, then the slowdown would be at most a factor of $\sqrt{\kappa(\bar{K}_t, K_t)}$, i.e., a slowdown of at most a factor of 20. In practice, the method seems to work better than that according to Avron et al.

11. Open questions. This work is the first to extend support-tree methods, which previously have been shown to be good preconditioners for diagonally dominant matrices with negative off-diagonal entries, to the class of finite element matrices. We have shown that the scope of the method includes a standard scalar elliptic boundary value problem, but perhaps the scope of finite element problems that can be tackled with this method could be expanded further.

One generalization would be the class of problems $\nabla \cdot (\Theta(x) \nabla u) = -f$, where $\Theta(x)$ is a spatially varying $d \times d$ SPD matrix. This generalization would present problems for our current analysis in the case that $\Theta(x)$ is highly ill conditioned. It would still be straightforward to write $K = A^T J^T D J A$, where D is now block diagonal, but our analysis of the introduction of $\bar{J}$ would run into trouble because the $dq \times dq$ diagonal blocks of D are no longer individually well conditioned.

It would also be interesting to tackle vector problems such as linear elasticity or Stokes' flow, or higher-order equations like the biharmonic equation. It seems likely that our techniques can extend to at least some of these problems, since they all have a SPD weak form. On the other hand, a recent result [7] suggests that the extension of our results to linear elasticity will not be straightforward because of the higher nullity of element stiffness matrices in the case of linear elasticity. A further generalization would be to unsymmetric problems like the convection-diffusion equation. The latter class of problems would require a substantial rethinking of the whole approach, since condition number reduction, which is very relevant for the application of conjugate gradients to SPD systems, is less relevant to the application of GMRES to unsymmetric systems.

Our analysis is based on condition numbers (support numbers). One drawback of this approach is that the convergence and work estimates may be too pessimistic. For instance, the condition number of the preconditioned linear systems depends on $\kappa_1(\mathcal{T})$, the worst aspect ratio of any element in the mesh. If there is only one poorly shaped element in the mesh, we expect iterative solvers will take only a few extra iterations since changing a single element implies a low-rank correction to the assembled stiffness matrix. Any analysis based on condition numbers will be unable to capture this effect. A related open issue is whether we can exploit recent work in mesh quality metrics [16] to show that "good meshes" both have small error in the FEM approximation and also produce linear systems that can be well approximated by diagonally dominant systems.

Another point to make about our method is that, although the condition number of the preconditioned system has an upper bound independent of

$$R_\theta = \max_{x\in\Omega} \theta(x) / \min_{x\in\Omega} \theta(x),$$

there will still be a loss of significant digits due to roundoff error when using our method in the case where R_θ is large. This is because the system matrix and the preconditioner separately are ill-conditioned operators. A special case of (3.1), in which R_θ is extremely large and in which $f = 0$, was considered in [22]. That paper proposed a method based on Gaussian elimination in which the loss of significant digits is avoided. Some of the ideas behind [22] were also extended to solution via conjugate gradient using support preconditioners by [14]. The methodology in [14], however, was for node-arc incidence matrices and for a particular kind of support preconditioner called a *support tree* [11], and it is not clear whether that method would apply to the present setting.

Finally, there is greater effort to couple automatic finite element mesh generation to a detailed analysis of the solver and its error bounds. Our approximation result adds another ingredient to this mix; it would be interesting if an automatic finite element mesh generator could be tailored to reduce specifically the quantity $\kappa(K, \bar{K})$.

Acknowledgments. The authors wish to thank the four anonymous referees of this paper for their very helpful comments.

REFERENCES

[1] J. H. Argyris, *The natural factor formulation of the stiffness for the matrix displacement method*, Comput. Methods Appl. Mech. Engrg., 5 (1975), pp. 97–119.

[2] H. Avron, D. Chen, G. Shklarski, and S. Toledo, *Combinatorial preconditioners for scalar elliptic finite-element problems*, preprint, Tel-Aviv University, Tel-Aviv, 2006.

[3] O. Axelsson and V. A. Barker, *Finite Element Solution of Boundary Value Problems: Theory and Computation*, SIAM Press, Philadelphia, 2001.

[4] M. Bern, J. R. Gilbert, B. Hendrickson, N. Nguyen, and S. Toledo, *Support-graph preconditioners*, SIAM J. Matrix Anal. Appl., 27 (2006), pp. 930–951.

[5] E. G. Boman and B. Hendrickson, *Support theory for preconditioning*, SIAM J. Matrix Anal. Appl., 25 (2003), pp. 694–717.

[6] D. Chen and S. Toledo, *Vaidya's preconditioners: Implementation and experimental study*, Electron. Trans. Numer. Anal., 16 (2003); also available online from http://etna.mcs.kent.edu.

[7] D. Chen and S. Toledo, *Combinatorial characterization of the null spaces of symmetric H-matrices*, Linear Algebra Appl., 392 (2004), pp. 71–90.

[8] R. Cools, *An encyclopaedia of cubature formulas*, J. Complexity, 19 (2003), pp. 445–453.

[9] M. Elkin, Y. Emek, D. A. Spielman, and S.-H. Teng, *Lower-stretch spanning trees*, in Proceedings of the 37th Annual ACM Symposium on Theory of Computing, New York, NY, 2005, Association for Computing Machinery, pp. 494–503.

[10] G. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed., Johns Hopkins University Press, Baltimore, 1996.

[11] K. Gremban, *Combinatorial Preconditioners for Sparse, Symmetric, Diagonally Dominant Linear Systems*, Ph.D. thesis, School of Computer Science, Carnegie-Mellon University, Pittsburg, PA, 1996. Available as Technical Report CMU-CS-96-123.

[12] R. Gupta, *Support Graph Preconditioners for Sparse Linear Systems*, Master's thesis, Texas A&M University, College Station, TX, 2004. Available online at https://txspace.tamu.edu/bitstream/1969.1/1353/1/etd-tamu-2004C-2-CPSC-Gupta.pdf.

[13] I. Gustafsson, *An incomplete factorization preconditioning method based on modification of element matrices*, BIT, 36 (1996), pp. 86–100.

[14] V. E. Howle and S. A. Vavasis, *An iterative method for solving complex-symmetric systems arising in electrical power modeling*, SIAM J. Matrix Anal. Appl., 26 (2005), pp. 1150–1178.

[15] C. Johnson, *Numerical Solution of Partial Differential Equations by the Finite Element Method*, Cambridge University Press, London, 1987.
[16] P. Knupp, *Algebraic mesh quality metrics*, SIAM J. Sci. Comput., 23 (2001), pp. 193–218.
[17] S. A. Mitchell and S. A. Vavasis, *Quality mesh generation in higher dimensions*, SIAM J. Comput., 29 (2000), pp. 1334–1370.
[18] T. Phillips and G. Miller, *private communication*, 2005.
[19] J. R. Shewchuk, *Triangle: Engineering a 2D quality mesh generator and Delaunay triangulator*, in Applied Computational Geometry: Towards Geometric Engineering, Lecture Notes Comput. Sci. 1148, M. C. Lin and D. Manocha, eds., Springer, Berlin, 1996, pp. 203–222. From the First ACM Workshop on Applied Computational Geometry.
[20] D. A. Spielman and S.-H. Teng, *Nearly-linear time algorithms for graph partitioning, graph sparsification, and solving linear systems*, in Proceedings of the 36th Annual ACM Symposium on Theory of Computing, ACM Press, New York, NY, 2004, pp. 81–90.
[21] S. Vavasis, *A Bernstein-Bézier sufficient condition for invertibility of polynomial mappings*, http://arxiv.org/abs/cs.NA/0308021 (2003).
[22] S. A. Vavasis, *Stable finite elements for problems with wild coefficients*, SIAM J. Numer. Anal., 33 (1996), pp. 890–916.
[23] M. Wang and V. Sarin, *Parallel support graph preconditions*, in High Performance Computing – HiPC 2006, Lecture Notes Comput. Sci. 4297, Y. Roberts, M. Parashas, R. Badrinath, and V. Prasanna, eds., Springer, Berlin, 2006, pp. 387–398.

SIAM J. NUMER. ANAL.
Vol. 46, No. 6, pp. 3285–3312
© 2008 Society for Industrial and Applied Mathematics

EFFECTIVE LAGUERRE ASYMPTOTICS*

DAVID BORWEIN†, JONATHAN M. BORWEIN‡, AND RICHARD E. CRANDALL§

Abstract. It is known that the generalized Laguerre polynomials can enjoy subexponential growth for large primary index. In particular, for certain fixed parameter pairs (a, z) one has the large-n asymptotic behavior

$$L_n^{(-a)}(-z) \sim C(a,z) n^{-a/2-1/4} e^{2\sqrt{nz}}.$$

We introduce a computationally motivated contour integral that allows efficient numerical Laguerre evaluations yet also leads to the complete asymptotic series over the full parameter domain of subexponential behavior. We present a fast algorithm for symbolic generation of the rather formidable expansion coefficients. Along the way we address the difficult problem of establishing effective (i.e., rigorous and explicit) error bounds on the general expansion. A primary tool for these developments is an "exp-arc" method giving a natural bridge between converging series and effective asymptotics.

Key words. Laguerre polynomials, effective asymptotic expansions, contour integrals

AMS subject classifications. 34E05, 33C65

DOI. 10.1137/07068031X

1. The challenge of "effectiveness." For primary integer indices $n = 0, 1, 2, \ldots$, we define the Laguerre polynomial thus:

$$L_n^{(-a)}(-z) := \sum_{k=0}^{n} \binom{n-a}{n-k} \frac{z^k}{k!}. \tag{1}$$

Our use of negated parameters $-a, -z$ is intentional, for convenience in our analysis and in connection with related research, as we later explain. We shall work on the difficult problem of establishing asymptotics *with* effective error bounds for the two-parameter domain

$$\mathbb{D} := \{(a,z) \in \mathbb{C} \times \mathbb{C} : z \notin (-\infty, 0]\}.$$

That is, a is any complex number, while z is a complex number not on the negative-closed cut $(-\infty, 0]$.

Herein, we say a function $f(n)$ is of *subexponential* growth if $\log f(n) \sim Cn^D$ as $n \to \infty$ for positive constants C, D, with $D < 1$. It will turn out that $\mathbb{D}$ is the precise domain of subexponential growth of $L_n^{(-a)}(-z)$ as $n \to \infty$, with (a, z) fixed. The reason for the negative-cut exclusion on z is simple: For z negative real, the Laguerre polynomial exhibits oscillatory behavior in large n and is not of subexponential growth. Note also, from the definition (1), that

$$L_n^{(-a)}(0) = \binom{n-a}{n}, \tag{2}$$

which covers the case $z = 0$; again, not subexponential growth.

*Received by the editors January 17, 2007; accepted for publication (in revised form) May 29, 2008; published electronically October 17, 2008.
http://www.siam.org/journals/sinum/46-6/68031.html

†Department of Mathematics, University of Western Ontario, London, ON, N6A 5B7, Canada and School of Mathematical and Physical Sciences, University of Newcastle, Callaghan, NSW 2300, Australia (dborwein@uwo.ca).

‡Faculty of Computer Science, Dalhousie University, Halifax, NS, B3H 2W5, Canada (jborwein@cs.dal.ca). This author's work was funded by NSERC and the Canada Research Chair Program.

§Center for Advanced Computation, Reed College, Portland, OR 97202 (crandall@reed.edu).

For fixed $(a,z) \in \mathbb{D}$, we shall have the large-n behavior (here and beyond we define $m := n+1$ which will turn out to be a natural reassignment):

$$L_n^{(-a)}(-z) \;\sim\; S_n(a,z)\left(1 + O\left(\frac{1}{m^{1/2}}\right)\right), \tag{3}$$

where the subexponential term S_n is

$$S_n(a,z) \;:=\; \frac{e^{-z/2}}{2\sqrt{\pi}}\,\frac{e^{2\sqrt{mz}}}{z^{1/4-a/2}\;m^{1/4+a/2}}. \tag{4}$$

In such expressions, $\Re\,(\sqrt{mz})$ denotes $\sqrt{m|z|}\cos(\theta/2)$, where $\theta := \arg(z) \in (-\pi,\pi]$ (we hereby impose $\arg(-1) := \pi$), and so for $(a,z) \in \mathbb{D}$, the expression (4) involves genuinely diverging growth in n.

1.1. Research motives. There are many interdisciplinary applications of Laguerre asymptotics. The Laguerre functions appear standardly in the quantum theory of the hydrogen atom [46, Chapter 4] and in certain exactly solvable three-body problems of chemical physics [13], [14]. The so-called WKB phase of a quantum eigenstate can, in such cases and for high quantum numbers, be calculated via Laguerre asymptotics. The Hermite polynomials $H_n(x)$—closely related to the Laguerre polynomials $L_n^{(\pm 1/2)}(x^2)$—also figure in quantum analyses; e.g., one may derive Hermite asymptotics from Laguerre asymptotics. In this context there is the fascinating López–Temme representation of *any* $L_n^{(-a)}(-z)$ as a finite superposition of Hermite evaluations [23].

Our own research motive for providing effective asymptotics involves not the Laguerre–Hermite connection, rather it begins with a beautiful link to the incomplete gamma function, namely, [19]

$$\Gamma(a,z) \;=\; z^a e^{-z}\cfrac{1}{z+\cfrac{1-a}{1+\cfrac{1}{z+\cfrac{2-a}{1+\cdots}}}} \;=\; z^a e^{-z}\sum_{n=0}^{\infty}\frac{(1-a)_n}{(n+1)!}\,\frac{1}{L_n^{(-a)}(-z)\,L_{n+1}^{(-a)}(-z)}, \tag{5}$$

where $(c)_n := c(c+1)\cdots(c+n-1)$ is the *Pochhammer symbol*. This series is valid whenever none of the Laguerre denominators has a zero. Thus an interesting sidelight is the research problem of establishing zero-free regions for Laguerre polynomials (see our Open problems section). We note that convergence of the continued-fraction can be highly problematic, especially when the full complex parameter space is allowed. One might encounter smooth convergence, or steady divergence, or even chaotic divergence, all of which possibilities are exemplified in [7, 8, 9, 22].

The problem, then, of general Laguerre asymptotics arises because there is a way to write Riemann zeta-function evaluations $\zeta(s)$ in terms of the incomplete gamma function, as was known to Riemann himself [40, p. 22]. But when s has large imaginary height, one becomes interested in the incomplete gamma's corresponding complex parameters (a,z) also of possibly large imaginary height [11], [12]. For example, in the art of prime-number counting, say, for primes $< 10^{20}$, one might need to know, for example, $\Gamma\left(3/4 + 10^{10}\,i, z\right)$ for various z also of large imaginary height, to good precision, in order to evaluate $\zeta(s)$ high up on the line $\Re(s) = 3/2$. The point is,

exact computations on prime numbers must, of course, employ rigor—hence the need for effective error bounds.

Independent of Riemann-ζ considerations is the following issue: Continued-fraction theory is to this very day incomplete in a distinct sense. The vast majority of available convergence theorems are for Stieltjes, or S-fractions. Now the continued fraction in (5) is an S-fraction only when the a parameter is real [24, p. 138]. Due to the relative paucity of convergence theorems outside the S-fraction class, one encounters great difficulty in estimating the convergence rate for arbitrary $\Gamma(a,z)$; this is what led to our focus on the Laguerre asymptotics. Incidentally, we are aware that subexponential convergence results *might* be attainable via the complicated and profound work of Jacobsen and Thron [21] on oval convergence regions. In any case, our effective-Laguerre approach proves the subexponential convergence of $\Gamma(a,z)$ for any complex pairs $(a,z) \in \mathbb{D}$; moreover, this is done with effective constant factors. Separate research on these superexponential convergence issues for continued fractions is underway.

1.2. Historical results on Laguerre asymptotics. Laguerre asymptotics have long been established for certain restricted domains and usually with noneffective asymptotics.[1] For example, in 1909 Fejér established that, for z on the open cut $(-\infty, 0)$ and any real a, one has [36, Theorem 8.22.1]:

$$L_n^{(-a)}(-z) = \frac{e^{-z/2}}{\sqrt{\pi}(-z)^{1/4-a/2}\ m^{1/4+a/2}} \cos\left(2\sqrt{-mz} + a\pi/2 - \pi/4\right) + O\left(m^{-a/2-3/4}\right), \tag{6}$$

where we again use index $m := n+1$, which slightly alters the coefficients in such classical expansions, said coefficients being for powers $n^{-k/2}$, but we are now using powers $m^{-k/2}$. By 1921 Perron [35] had generalized the Fejér series to arbitrary orders, then, for $z \notin (-\infty, 0]$, established a series consistent with (3) and (4), in essentially the form [36, Theorem 8.22.3]:

$$L_n^{(-a)}(-z) = S_n(a,z)\left(\sum_{k=0}^{N-1} \frac{C_k}{m^{k/2}} + O\left(m^{-N/2}\right)\right), \tag{7}$$

although this was for *real* a and so not for our general parameter domain $\mathbb{D}$. Note that $C_0 = 1$, consistent with our (3) and (4); however, one should take care that because we are using index $m := n+1$, the coefficients C_k in the above formula differ slightly from the historical ones.

A modern literature treatment that is again consistent with the heuristic (3)–(4) is given by Winitzki in [45], where one invokes a formal generating function to yield a contour integral for $L_n^{(-a)}(-z)$. Then a stationary-phase approach yields the correct first-asymptotic term, at least for certain subregions of $\mathbb{D}$. Winitzki's treatment is both elegant and nonrigorous; there is no explicit estimate given on the $O(1/\sqrt{m})$ correction in (3).

There is an interesting anecdote that reveals the difficulty inherent in Laguerre asymptotics. Namely, W. Van Assche in a fine 1985 paper [41] used the expansion

[1] Some researchers use the term "realistic error bound" for big-O terms that have explicit structure. We prefer "effective bound," and when an expansion is bestowed with such a bound, we may say "effective expansion."

(7) for work on zero-distributions, only to find by 2001 that the C_1 term in that 1985 paper had been calculated incorrectly. The amended series is given in his correction note [42] as

$$L_n^{(-a)}(-z) = \frac{e^{-z/2}}{2\sqrt{\pi}} \frac{e^{2\sqrt{nz}}}{z^{1/4-a/2}\, n^{1/4+a/2}} \cdot \left(1 + \left(\frac{3 - 12a^2 + 24(1-a)z + 4z^2}{48\sqrt{z}}\right) \frac{1}{\sqrt{n}} + O\left(\frac{1}{n}\right)\right),$$

or in our own notation with $m := n + 1$,

$$= S_n(a,z) \left(1 + \left(\frac{3 - 12a^2 - 24(1+a)z + 4z^2}{48\sqrt{z}}\right) \frac{1}{\sqrt{m}} + O\left(\frac{1}{m}\right)\right). \tag{8}$$

Note the slight alteration used to obtain our $C_1/\sqrt{m}$ term. Van Assche credits T. Müller and F. Olver for aid in working out the correct $O(1/\sqrt{n})$ component.[2] This story suggests that even a low-order asymptotic development is nontrivial.

The great classical analysts certainly knew in principle how to establish effective error bounds. The excellent treatment of effectiveness for Laplace's method of steepest descent in [27] is a shining example. Also illuminating is Olver's paper [26], which explains effective bounding and shows how unwieldy rigorous bounds can be obtained. However, efficient algorithms for generating *explicit* effective big-O constants have only become practicable in recent times, when computational machinery is prevalent.

2. Contour representation. In this section we develop an efficient—both numerically and analytically—contour-integral representation for $L_n^{(-a)}(-z)$.

2.1. Development of a "keyhole" contour. A well-known integral [36, section 5.4] has contour Γ encircling $s = 1$ and avoiding the branch cut $(-\infty, 0]$:

$$L_n^{(-a)}(-z) = \frac{e^{-z}}{2\pi i} \oint_\Gamma s^{-1-a} \left(1 - \frac{1}{s}\right)^{-n-1} e^{zs}\, ds. \tag{9}$$

This representation holds for all pairs $(a, z) \in \mathbb{C} \times \mathbb{C}$, with n a nonnegative integer.[3]

However—and this is important—we found via experimental mathematical techniques, e.g., extreme-precision evaluations, that a specific kind of contour allows very accurate, efficient, and well-behaved *numerical* Laguerre evaluations. It is not completely understood why the "keyhole" contour we are about to define does so well in numerical Laguerre evaluations; we do know that this new contour consistently provides better numerics than, say, a simple circle surrounding $s = 1$. It may be simply the internal quirks of various modern numerical integrators at work, or it may be the smooth phase behavior along our keyhole's perimeter.

For the desired contour, take $z \neq 0$, $m := n + 1$ and assume $r := \sqrt{m/z}$ has $|r| > 1/2$. Then, use a circular contour centered at $s = 1/2$ with radius $|r|$. This contour will encompass $s = 1$, so the remaining requirement is to avoid the cut $s \in (-\infty, 0]$ as can be done by cutting out a "wedge" from the negative-real arc of the circle, with apex at $s = 1/2$. We tried such schemes with high-precision integration, to

[2] Accordingly, we hereby name the polynomial $C_1(a, z)$ the Perron–van Assche–Müller–Olver (PAMO) coefficient.

[3] The contour representation (9) can easily be continued to noninteger n, with care taken on the $(-n-1)$th power, but our present treatment will only use nonnegative integer n.

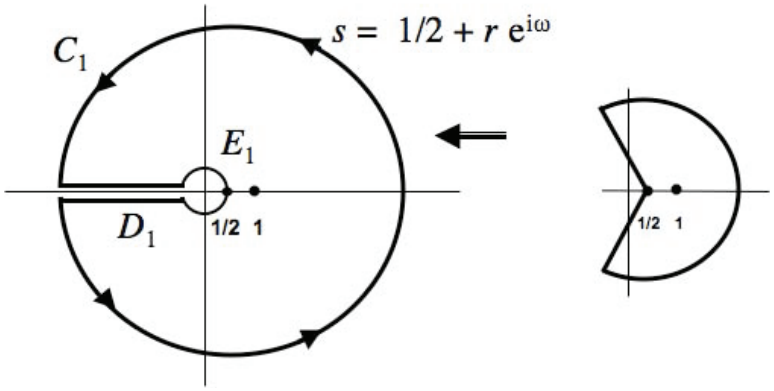


FIG. 1. *A numerically efficient "keyhole contour" for Laguerre evaluations $L_n^{(-a)}(-z)$, valid for all complex a, z, with $z \neq 0$ and $n+1 > |z|/4$. Wedges with center-1/2—as pictured at right—are experimentally accurate, leading to a "keyhole" deformation avoiding the cut $s \in (-\infty, 0]$. It turns that, for any $(a, z) \in \mathbb{D}$, the main arc C_1 gives the predominant contribution for large n, the D_1, E_1 components being subexponentially minuscule.*

settle finally on the contour of Figure 1, where the aforementioned wedge has evolved to a "keyhole" pattern consisting of cut-run D_1 and small, origin-centered circle E_1 of radius $1/2$.

So, adopting constraints and nomenclature

$$z \neq 0, \quad \theta := \arg(z), \quad \omega_\pm := \pm\pi + \theta/2,$$
$$m := n + 1, \quad r := \sqrt{m/z} := \sqrt{m/|z|}e^{-i\theta/2}, \quad R := |r| > 1/2,$$

but no other constraints, we have the following representation:

$$L_n^{(-a)}(-z) = c_1 + d_1 + e_1, \tag{10}$$

where c_1, d_1, e_1 are the respective contributions from contour C_1, cut-discontinuity D_1, and contour E_1 from Figure 1. Exact formulae for said contributions are

$$c_1 = \frac{1}{2\pi} r^{-a} e^{-z/2} \int_{\omega_-}^{\omega_+} \mathcal{H}_m(a, z, e^{-i\omega}) e^{2\sqrt{mz}\cos\omega}\, d\omega, \tag{11}$$

$$d_1 = \frac{e^{-z}}{\pi} \sin(\pi a) \int_{1/2}^{R-1/2} T^{-1-a} \left(1 + \frac{1}{T}\right)^{-m} e^{-zT}\, dT, \tag{12}$$

$$e_1 = -\frac{e^{-z}}{4\pi} \int_{-\pi}^{\pi} \left(2e^{-i\omega}\right)^{1+a} \left(1 - 2e^{-i\omega}\right)^{-m} e^{i\omega + \frac{z}{2}e^{i\omega}}\, d\omega, \tag{13}$$

and when $m > \Re(a)$, we may write this last contribution, by shrinking down the radius-$1/2$ contour segment to embrace the cut $(-1/2, 0]$, as

$$e_1 = \frac{e^{-z}}{\pi} \sin(\pi a) \int_0^{1/2} T^{-1-a} \left(1 + \frac{1}{T}\right)^{-m} e^{-zT}\, dT. \tag{14}$$

For the c_1 contribution above, we have used the function $\mathcal{H}_m$ defined by

$$\mathcal{H}_m(a, z, v) := v^a \left(1 + \frac{v}{2r}\right)^{-1-a} \left[F\left(\frac{v}{r}\right)\right]^m, \tag{15}$$

$$F(t) := \left(\frac{1 + t/2}{1 - t/2}\right) e^{-t}, \tag{16}$$

which for small t can be written $F(t) = 1 + t^3/12 + t^5/80 + \cdots = 1 + O(t^3)$.

In any event, e_1, like d_1, is subexponentially small relative to c_1 for large n and $(a,z) \in \mathbb{D}$. It is to be stressed that decomposition (10) holds for all $(a,z) \in \mathbb{C} \times \mathbb{C}, z \neq 0$, as long as $m := n + 1 > |z|/4$. We remind ourselves of (2) for $z = 0$. This means that such contour calculus applies to both oscillatory (Fejér) cases, where z is negative real, and subexponential (Perron) cases for the stated parameters.

2.2. Numerical success on the keyhole contour. The triumvirate c_1, d_1, e_1 of integrals is suitable for accurate Laguerre computations, which computations do show that *integrals d_1, e_1 tend to be subexponentially small* relative to c_1. This, of course, is our motive for so identifying the contour terms. An example, using the defining series (1) with $(a,z) := (-i, -1-i)$ is

$$L_8^{(i)}(1+i) \;=\; -\frac{137}{288} + \frac{53}{45} i \;\approx\; -0.475694444444444444 + 1.177777777777777777 i,$$

while the contributions from the C_1, D_1, E_1 segments of the contour of Figure 1 are

$$\begin{aligned} c_1 &\approx -0.44406762576110996056 + 0.81722282272705891241 i, \\ d_1 &\approx -0.03169598827425878852 + 0.36065591639721657587 i, \\ e_1 &\approx 0.00006916959092430464 - 0.00010096134649771051 i, \end{aligned}$$

with the sum $c_1 + d_1 + e_1$ giving $L_8^{(i)}(1+i)$ to the implied precision.

Remarkably, this "keyhole-contour" approach has the additional, unexpected feature that, for some parameter regions, the contour evaluation of $L_n^{(a)}(-z)$ is actually *faster* than direct summation of the defining series (1). A typical example is as follows: For the 13-digit evaluation

$$L_{50000}^{(25i/2)}(-30+3i) \approx (0.9275136583293 + 1.7406691595239 i) \times 10^{1056},$$

the defining series (1) was, in our trials, slower than the integral c_1.[4]

3. Effective bounds on the contour components.

3.1. Theta-calculus. We now introduce a notation useful in effective-error analysis. When two functions enjoy

$$|f(z)| \leq |g(z)|$$

over some relevant domain of z values, we shall say that

$$f(z) = \Theta(g(z)), \quad \text{or} \quad f(z) = \Theta : g(z)$$

on said domain. Thus, the Θ-notation is an effective replacement for big-O notation. The reason for allowing the notation $\Theta : \ldots$ is that a long formula g can run arbitrarily to the right of the colon.

Incidentally, there is a literature precedent for such a "theta-calculus." Namely, in computational number theory treatments such as [10], one encounters terms such as, say, $\theta x/\log^2 x$, with a stated constraint $\theta \in [-10, 10]$. This means the term is $O(x/\log^2 x)$ but with effective big-O constant bounded in magnitude by 10. In our nomenclature we would use $\Theta : 10x/\log^2 x$.

[4] Of course, series acceleration as in [5] would give the direct series a "leg up." Still, it is remarkable that contour integration is competitive. For the record, we compared *Mathematica*'s numerical integration to its own `LaguerreL[ ]` function.

3.2. Effective bound for e_1. First we address the integral e_1 as given in relation (13). We need some preliminary lemmas, starting with a collection of polynomial estimates to transcendental functions.

LEMMA 1. *The following inequalities hold for the respective conditions on ω:*

$$\log(5-4\cos\omega) \geq \frac{\log 9}{\pi^2}\omega^2, \quad \omega \in [-\pi,\pi];$$
$$\log(1+\omega) \geq \frac{4}{5}\omega, \quad \omega \in [0,1/2];$$
$$\arcsin\omega \leq \frac{\pi}{\sqrt{8}}\omega, \quad \omega \in [0,1/\sqrt{2}];$$
$$\cos\omega \leq 1-\frac{4}{\pi^2}\omega^2, \quad \omega \in [-\pi/2,\pi/2].$$

Proof. All four are straightforward calculus exercises. □

LEMMA 2. *Consider integrals of error-function class, specifically, for nonnegative real parameter μ,*

$$V_\mu(\alpha,\beta,\gamma) := \int_\gamma^\infty (x^2)^\mu e^{2\alpha x-\beta x^2}\, dx,$$

where each of α,β,γ is real, with $\alpha,\beta>0$. Then we always have the bound

$$V_0(\alpha,\beta,\gamma) = \Theta : \sqrt{\frac{\pi}{\beta}}\; e^{\frac{\alpha^2}{\beta}}.$$

If in addition $\gamma > 2\alpha/\beta$, we also have

$$V_\mu(\alpha,\beta,\gamma) = \Theta : \frac{1}{2}\; \frac{\Gamma(\mu+1/2)}{(\beta-2\alpha/\gamma)^{\mu+1/2}};$$

finally, if $\gamma > 4\alpha/\beta$, we have

$$V_\mu(\alpha,\beta,\gamma) = \Theta :\; 2^{\mu-1/2}\; \frac{\Gamma(\mu+1/2)}{\beta^{\mu+1/2}}.$$

Remark. We use the double-exponential $(x^2)^\mu$ because the first case of the theorem allows $\gamma<0$.

Proof. Completing the exponent's square gives the first bound easily, since

$$V_0 = e^{\alpha^2/\beta}\int_\gamma^\infty e^{-\beta(x-\alpha/\beta)^2}\, dx \;=\; \Theta : \sqrt{\pi/\beta}.$$

For the second bound, it is elementary that, for $x\geq\gamma$, one has $2\alpha x-\beta x^2 \leq x^2(2\alpha/\gamma-\beta)$ so that

$$V_\mu \leq \int_0^\infty x^{2\mu} e^{-(\beta-2\alpha/\gamma)x^2}\, dx,$$

and the resulting bound follows. The final bound follows immediately from the previous bound, because $\gamma > 4\alpha/\beta$ implies $\beta-2\alpha/\gamma > \beta/2$. □

THEOREM 1. *The contour contribution e_1 defined by* (13)*, under conditions $(a,z) \in \mathbb{D}$ and m sufficiently large in the explicit sense $m := n+1 > m_0 := |z|/4,\ \ m > m_1 := 5\left(|\Re(z)| + (|\Im(a)| + |\Im(z)|/2)^2\right)$, is bounded as*

$$e_1 = \Theta : e^{-z/2} 2^{\Re(a)+3} \frac{1}{\sqrt{m}}.$$

Moreover, under alternative conditions $(a,z) \in \mathbb{D}$ and $m > m_0, \quad m > m_2 := \Re(a), \quad m > m_3 := -\frac{5}{4}|z|\cos\theta$ we have a bound

$$e_1 = \Theta : \frac{e^{-z}}{\pi} \sin(\pi a) \frac{2^{\Re(a)-m}}{m - \Re(a)}.$$

Proof. The proof involves careful application of Lemmas 1 and 2 to the integral forms (13), (14); for brevity we omit these details. ∎

3.3. Effective bound for d_1. Again we need an opening lemma.

LEMMA 3. *Consider integrals of the incomplete-Bessel class, specifically*

$$W(\alpha,\beta) := \int_0^1 e^{-\alpha x - \frac{\beta}{x}}\ \frac{dx}{x},$$

where each of α, β is real, with $\beta > 0$. If $\beta > \alpha$, we have a bound

$$W = \Theta : \frac{1}{2} e^{-\beta-\alpha} \sqrt{\frac{\pi}{\beta}}.$$

Proof. We write

$$W(\alpha,\beta) = \int_0^1 e^{\beta x - \alpha x}\, e^{-\beta x - \frac{\beta}{x}}\ \frac{dx}{x} \le e^{\beta-\alpha} \int_0^1 e^{-\beta x - \frac{\beta}{x}}\ \frac{dx}{x}.$$

This last integral is a modified-Bessel term $K_0(2\beta)$ and has the required bound [2, Lemma 1]. ∎

THEOREM 2. *The contour contribution d_1 defined by* (12) *can be bounded, under conditions $(a,z) \in \mathbb{D}$ and $m > m_4 := 4|z|$, as*

$$d_1 = \Theta : \frac{e^{-z/2}}{\sqrt{\pi}} m^{|\Re(a)|/2 - 1/4} |z|^{-|\Re(a)|/2 - 1/4} \sin(\pi a) e^{-2\sqrt{m|z|}\cos^2\frac{\theta}{2}}.$$

Proof. Starting from (12) one applies Lemma 3; again for brevity we omit the details. ∎

3.4. Rigorous estimates on c_1. Having dispensed with d_1, e_1, we next show the integration limits ω_-, ω_+ on the c_1 contribution can be changed—with a subexponentially small error penalty—to $-\pi/2, \pi/2$, respectively.

LEMMA 4. *For any v on the unit circle $\{e^{i\phi} : \phi \in (-\pi, \pi]\}$, any complex a, and any real $R > (1+|a|)/2$ we have*

$$\left| \left(1 + \frac{v}{2R}\right)^{-1-a} \right| \le \frac{1}{1 - \frac{1+|a|}{2R}}.$$

Proof. Via the binomial theorem,

$$\left| \left(1 + \frac{v}{2R}\right)^{-1-a} \right| \le 1 + (1+|a|)\frac{1}{2R} + (1+|a|)^2 \left(\frac{1}{2R}\right)^2 + \cdots = \frac{1}{1 - \frac{1+|a|}{2R}}. \qquad ∎$$

LEMMA 5. *Let v be on the unit circle as in Lemma 4, let $R > 1$ be real, and let m be a positive integer. For the function F appearing in* (16)*, we have the bound*

$$\left| F\left(\frac{v}{R}\right)^m \right| \leq e^{\frac{1}{6}\frac{m}{R^3}}.$$

Proof. From the definition (16) we have

$$F(v/R)^m = e^{\frac{1}{12}\frac{m}{R^3}(1+(3/5)/(2R)^2+(3/7)/(2R)^4+\cdots)}.$$

For $R := 1$, the infinite sum is no larger than 1.18 and is monotonic decreasing in R. □

LEMMA 6. *Let v be on the unit circle as in Lemma 4 and define for nonnegative integer m*

$$K := \left(1+\frac{v}{2R}\right)^{-1-a} F\left(\frac{v}{R}\right)^m.$$

For the assignments $(a, z) \in \mathbb{C} \times \mathbb{C}, z \neq 0, R := \sqrt{m/|z|} > 1, m > m_5 := |z|(1 + |a| + |z|/2)^2$, we have

$$K = \Theta : 2.$$

Proof. From Lemmas 4 and 5 we have

$$|K| \leq \frac{1}{1-\frac{1+|a|}{2(1+|a|+|z|/2)}} e^{\frac{1}{6}\frac{|z|}{1+|a|+|z|/2}}.$$

The fact that the right-hand side is $\Theta(2)$ follows from the observation that the function

$$\frac{1}{1-\frac{Q}{2Q+x}} e^{\frac{1}{3}\frac{x}{2Q+x}}$$

for $Q \geq 1, x \in [0, \infty)$ is itself $\Theta(2)$. This in turn follows easily on substituting $y := x/(2Q + x)$. Now the function to be bounded is $g(y) := (2/(1 + y))e^{y/3}$ on $y \in [0, 1/2]$—as differentiation settles. □

These lemmas in turn allow us to contract the range on the c_1 contour integral.

THEOREM 3. *Decompose c_1 as defined by* (11) *into two terms,*

$$c_1 := c_0 + c_2,$$

with c_0 involving the integral's range contracted to $[-\pi/2, \pi/2]$, namely,

$$c_0 := \frac{1}{2\pi} r^{-a} e^{-z/2} \int_{-\pi/2}^{\pi/2} \mathcal{H}_m(a, z, e^{-i\omega}) e^{2\sqrt{mz}\cos\omega}\, d\omega.$$

Then, under conditions $(a, z) \in \mathbb{D}$ and $m > m_5 := |z|(1 + |a| + |z|/2)^2$, we have a bound

$$c_2 = \Theta : r^{-a} e^{-z/2} e^{\frac{3}{2}\pi|\Im(a)|}.$$

Proof. It is evident that c_2 is obtained from the definition (11) but with the integral replaced via

$$\int_{\omega_-}^{\omega_+} \to \left\{ \int_{\omega_-}^{-\pi/2} + \int_{\pi/2}^{\omega_+} \right\}.$$

Over these domains of integration, we have $e^{2\sqrt{mz}\cos\omega} = \Theta(1)$. Thus, Lemmas 4, 5, and 6 show

$$c_2 = \Theta : \frac{1}{2\pi} r^{-a} e^{-z/2} \left\{ \int_{\omega_-}^{-\pi/2} + \int_{\pi/2}^{\omega_+} \right\} 2 \left| e^{-i\omega a} \right| \, d\omega.$$

As the total support of the integrals cannot exceed π, the desired c_2 bound follows. □

3.5. Summary of the contour decomposition for $L_n^{(-a)}(-z)$. The above manipulations lead to the main result of the present section, namely, a formula that decomposes the Laguerre evaluation, as in the following.

THEOREM 4 (contour decomposition). *Let $(a, z) \in \mathbb{C} \times \mathbb{C}$ be an arbitrary parameter pair, with $z \neq 0$ (with $z = 0$ cases resolved exactly by (2)). If $m := n + 1 = m_0 > |z|/4$, then the contour decomposition*

$$L_n^{(-a)}(-z) = c_0 + \mathcal{E}$$

holds, with c_0, c_2 as defined in Theorem 3 and $\mathcal{E} := c_2 + d_1 + e_1$. If an appropriate set of conditions on m across Theorems 1, 2, and 3 *holds, then we can write*

$$L_n^{(-a)}(-z) = c_0 + S_n(a, z)\mathcal{E}_1,$$

where $\mathcal{E}_1$ is subexponentially small, in the sense that, for any fixed positive ϵ, the large-n behavior is

$$\mathcal{E}_1 = O\left(e^{-(2-\epsilon)\sqrt{m|z|}\cos\frac{\theta}{2}} \right).$$

Moreover, an effective big-O constant is available (the proof exhibits explicit forms).

Proof. The contour calculus holds for all complex pairs (a, z), with $z \neq 0, R := \sqrt{m/|z|} > 1/2$, which assures that the point 1 is contained in the contour. It remains to analyze $\mathcal{E}_1$. Consider, then, an appropriate union of conditions from the cited theorems, say, $(a, z) \in \mathbb{D}$ and $m > \max(m_0, m_1, m_2, m_3, m_4, m_5)$, in which case we have, from Theorems 1, 2, and 3 (in that respective order of Θ terms):

$$\mathcal{E}_1 = \Theta : \frac{e^{-z/2}}{\sqrt{\pi}} \sin(\pi a) \frac{2^{1+\Re(a)-m}}{m - \Re(a)} z^{1/4-a/2} m^{1/4+a/2} e^{-2\sqrt{m|z|}\cos\frac{\theta}{2}}$$

$$+\Theta : 2 \left(\frac{m}{|z|} \right)^{(\Re(a)+|\Re(a)|)/2} e^{\frac{3}{2}\pi|\Im(a)|} e^{-2\sqrt{m|z|}\left(\cos^2\frac{\theta}{2} + \cos\frac{\theta}{2}\right)}$$

$$+\Theta : 2\sqrt{\pi}\, |mz|^{1/4} e^{\frac{3}{2}\pi|\Im(a)|} e^{-2\sqrt{m|z|}\cos\frac{\theta}{2}}.$$

This explicit bounding of the error term $\mathcal{E}_1$ proves the big-O statement of the theorem, while for any choice of ε, an effective big-O constant can be read off at will. □

4. Effective expansion for the $\mathcal{H}$-kernel. Theorem 4 shows the main contribution to $L_n^{(a)}(-z)$, for $(a, z) \in \mathbb{D}$, with $r := \sqrt{m/z}, |r| > 1/2$, is

$$c_0 := \frac{1}{2\pi} r^{-a} e^{-z/2} \int_{-\pi/2}^{\pi/2} \mathcal{H}_m\left(a, z, e^{-i\omega}\right) e^{2\sqrt{mz}\cos\omega} \, d\omega, \tag{17}$$

with the integration kernel $\mathcal{H}_m$ defined, see (15) and (16), as

$$\mathcal{H}_m(a,z,v) := v^a \left(1+\frac{v}{2r}\right)^{-1-a} \left(\frac{1+\frac{v}{2r}}{1-\frac{v}{2r}}\right)^m e^{-mv/r}, \tag{18}$$

where in the integral we assign $v := e^{-i\omega}$. We need to obtain the growth properties of $\mathcal{H}_m$.

4.1. Exponential form for $\mathcal{H}_m$.

LEMMA 7. *For $|v|=1$ and $m > |z|/4$, the $\mathcal{H}$-kernel can be cast in the exponential form*

$$\mathcal{H}_m := v^a \exp\left\{\sum_{k\geq 1} \frac{a_k}{k}\frac{1}{m^{k/2}}\right\}, \tag{19}$$

where

$$a_k := (1+a)(-1)^k \left(\frac{v\sqrt{z}}{2}\right)^k + \left(1-(-1)^k\right)\frac{k}{k+2}\left(\frac{v\sqrt{z}}{2}\right)^{k+2}. \tag{20}$$

Moreover, we have the general coefficient bound

$$|a_k| \leq \left(\frac{\sqrt{z}}{2}\right)^k (1+|a|+|z|/2). \tag{21}$$

Proof. (18) can be recast, with $\rho := v/(2r)$, as

$$\mathcal{H}_m := v^a \exp\left\{-(1+a)\log(1+\rho) - 2m\rho + m\left(\log(1+\rho)-\log(1-\rho)\right)\right\}. \tag{22}$$

Since $|r|>1/2$ and $|v|=1$, the logarithmic series converge absolutely and we have

$$\mathcal{H}_m := v^a \exp\left\{\sum_{k\geq 1}\left[\frac{(1+a)(-1)^k}{k}\left(\frac{v\sqrt{z}}{2}\right)^k g^k + \frac{2}{2k+1}\left(\frac{v\sqrt{z}}{2}\right)^{2k+1} g^{2k-2}\right]\right\}, \tag{23}$$

where $g := 1/\sqrt{m}$, and the precise form (20) for the a_k follows immediately. The given bound on $|a_k|$ is also immediate from (20). ∎

4.2. Exponentiation of series. Though Lemma 7 is progress, we still need to exponentiate a series, in the sense that we want to know, for the following expansion, given the sequence (a_k),

$$\exp\left\{\sum_{k\geq 1}\frac{a_k}{k}x^k\right\} =: \sum_{h\geq 0} A_h x^h,$$

how the A_h depend on the a_k. The combinatorial answer is

$$A_h = \sum_{j=0}^{h}\frac{1}{j!}\,G_h(j;\vec{a}), \qquad G_h(j;\vec{a}) := \sum_{h_1+\cdots+h_j=h}\frac{a_{h_1}\cdots a_{h_j}}{h_1\cdots h_j},$$

with the understanding that $G_h(0,\vec{a}) := \delta_{0h}$ and that such combinatorial sums involve positive integer indices h_i. One useful result is the following.

LEMMA 8. *If all coefficients $a_k = 1$, then, for $j > 0$,*

$$G_h(j;\vec{1}) \;=\; \sum_{h_1+\cdots+h_j=h} \frac{1}{h_1\cdots h_j} \;=\; \Theta : \frac{1}{h}\,(2H_{h-j+1})^{j-1} \;=\; \Theta : \frac{1}{h}\,(2\gamma + 2\log h)^{j-1}\,.$$

Here, $H_p := 1 + 1/2 + \cdots + 1/p$ is the pth harmonic number, $H_0 := 0$, and γ is the Euler constant.

Remark. It turns out that G here enjoys a closed form of sorts, namely,

$$G_h(j;\vec{1}) = \frac{j!}{h!}(-1)^{h-j}\mathcal{S}_h^{(j)},$$

where $\mathcal{S}$ denotes the Stirling number of the first kind, normalized via $x(x-1)\cdots(x-h+1) =: \sum_{j=0}^{h} \mathcal{S}_h^{(j)} x^j$. So one byproduct of our lemma is a rigorous bound on the growth of Stirling numbers; see [1, 24.1.3,III] and [37] for research on Stirling asymptotics.

Proof. The first Θ-estimate arises by induction. For notational convenience we omit the vector $\vec{1}$ and just use the symbol $G_h(j)$. Note that $G_N(1) = 1/N$ and $G_N(2) = \frac{2}{N}H_{N-1}$. Generally we have

$$G_N(J) = \sum_{j=1}^{N-J+1} \frac{1}{j} G_{N-j}(J-1).$$

Now, assume by induction that $G_h(j) = \Theta : \frac{1}{h}(2H_{h-j+1})^{j-1}$ holds for all $j < J$. Then

$$G(N,J) \;\le\; \sum_{j=1}^{N-j+1} \frac{2^{J-2}H_{N-J-j+2}^{J-2}}{j(N-J)} \;\le\; \frac{2^{J-2}}{N} H_{N-J+1}^{J-2} \sum_{j=1}^{N-J+1} \left(\frac{1}{j} + \frac{1}{N-j}\right).$$

Now this last parenthetical term is $H_{N-J+1} + H_{N-1} - H_J$ which, because $H_a - H_b \le H_{a-b}$ for any positive integer indices $a > b$, is bounded above by $2H_{N-J+1}$, which proves the first Θ-bound of the theorem. For the second Θ-bound it suffices, since H_j is increasing, to show that $H_{n-1} > \gamma + \log(n)$, which is an elementary calculus problem. ∎

Though we do not use Lemma 8 directly in what follows, it is useful in proving convergence for various sums $\sum A_h x^h$ and may well matter in future research along our lines.

LEMMA 9. *Let $y \ge 1$ and $x \in (-1,1)$ be real. Then in the expansion*

$$\exp\left\{ y \sum_{k\ge 1} \frac{x^k}{k} \right\} =: \sum_{h\ge 0} Y_h x^h$$

the coefficients Y_h enjoy the bound $Y_h = \Theta : y^h$.

Proof. First, $Y_0 = 1 \le 1$. The left-hand side is $\exp(-y\log(1-x)) = (1-x)^{-y}$ whose binomial expansion has hth coefficient ($h \ge 1$) equal to

$$\frac{y(y+1)\cdots(y+h-1)}{h!} \le y^h\, \frac{1}{1}\frac{1+1/y}{2}\cdots\frac{1+(h-1)/y}{h} \le y^h. \qquad ∎$$

4.3. Effective expansions of exponentiated series. We are now in a position to derive an effective expansion for an exponentiated series, starting with the following.

LEMMA 10. *Assume complex vector $\vec{b}$ of defining coefficients b_k bounded as $|b_k| \leq cd^k$ for positive real c, d with $c \geq 1$. Assume also $|x| < 1/(2cd)$. Then, for any order $N \geq 0$, we have an effective expansion*

$$\exp\left\{\sum_{k\geq 1} \frac{b_k}{k} x^k\right\} = \sum_{h=0}^{N-1} B_h x^h + \Theta : 2c^N d^N x^N,$$

where

$$B_h = \sum_{j=0}^{h} \frac{G_h(j;\vec{b})}{j!}$$

are the usual coefficients of the full formal exponentiation.

Proof. Denoting $f(x) := \exp\left\{\sum_{k\geq 1} \frac{b_k}{k} x^k\right\}$, we have

$$f(x) = \sum_{h=0}^{N-1} B_h x^h + T_N,$$

with remainder $T_N = \sum_{h\geq N} B_h x^h$, having B_h coefficients given by a G-sum as in Lemma 10. Now,

$$|B_h| \leq \sum_{j=0}^{h} \frac{|G_h(j;\vec{b})|}{j!} \leq \sum_{j=0}^{h} \frac{|G_h(j;\vec{f})|}{j!},$$

where $\vec{f} = (cd^k : k \geq 0)$. By Lemma 9 we know that $|B_h| \leq (cd)^h$. Therefore

$$|T_N| \leq \sum_{h\geq N} (cd)^h x^h = \frac{c^N d^N x^N}{1 - cdx} = \Theta : 2(cdx)^N. \qquad \square$$

Finally we arrive at a general expansion—with effective remainder—for the $\mathcal{H}$-kernel.

THEOREM 5 (effective expansion for $\mathcal{H}_m$). *For general complex $(a, z) \in \mathbb{C} \times \mathbb{C}$, assume that $m > m_5 := |z|(1 + |a| + |z|/2)^2$ and $|v| = 1$. Then, for any expansion order $N \geq 0$, we have*

$$\mathcal{H}_m(a, z, v) = v^a \left(\sum_{h=0}^{N-1} \frac{A_h}{m^{h/2}} \; + \; \Theta : 2\left(\frac{m_5}{4m}\right)^{N/2} \right),$$

where, on the basis of the defining coefficients a_k given in (20),

$$A_h := \sum_{j=0}^{h} \frac{G_h(j;\vec{a})}{j!}, \qquad G_h(j;\vec{a}) := \sum_{h_1+\cdots+h_j=h} \frac{a_{h_1} \cdots a_{h_j}}{h_1 \cdots h_j}.$$

Proof. The result follows immediately from Lemma 10, on assigning $\vec{b} = \vec{a}$, with $x := 1/\sqrt{m}$, $c := 1 + |a| + |z|/2$, $d := (1/2)\sqrt{|z|}$. $\square$

Note that Theorem 5 in the instance $N = 0$ implies our previous Lemma 6. It is interesting and suggestive that the threshold $m_5 := |z|(1 + |a| + |z|/2)^2$ appears in these results rather naturally.

4.4. Effective integral form for c_0. To obtain a useful form for c_0, the dominant component of $L_n^{(-a)}(-z)$, we use Theorem 5 to obtain the following.

THEOREM 6. *For $(a,z) \in \mathbb{D}$ and $m > m_5$, the dominant component of Theorems 3 and 4, namely,*

$$c_0 := \frac{1}{2\pi} r^{-a} e^{-z/2} \int_{-\pi/2}^{\pi/2} \mathcal{H}_m(a,z,e^{-i\omega}) e^{2\sqrt{mz}\cos\omega}\, d\omega,$$

can be given an effective form for any order $N \geq 0$, as

$$c_0 = \frac{1}{2\pi} r^{-a} e^{-z/2} \sum_{h=0}^{N-1} \frac{1}{m^{h/2}} \int_{-\pi/2}^{\pi/2} e^{-i\omega a} A_h e^{2\sqrt{mz}\cos\omega}\, d\omega + \; S_n(a,z)\; \mathcal{E}_{2,N},$$

where the error term is bounded as

$$\mathcal{E}_{2,N} = \Theta : \frac{\pi}{\sqrt{2}} \left(\frac{m_5}{4m}\right)^{N/2} \exp\left(\frac{\pi^2 \Im(a)^2 \sec\frac{\theta}{2}}{32\sqrt{m|z|}}\right) \sec^{1/2}\frac{\theta}{2}$$

and the A_h are to be calculated as the first N coefficients of

$$\sum_{h=0}^{\infty} A_h x^h := \exp\left\{ \sum_{k\geq 1} \frac{a_k}{k} x^k \right\},$$

via (20)*, with $v := e^{-i\omega}$.*

Proof. Inserting the effective $\mathcal{H}$-kernel expansion from Theorem 5 directly into the c_0 integral gives the indicated sum over $h \in [0, N-1]$ plus an error term

$$\Theta : \frac{1}{2\pi} r^{-a} e^{-z/2} 2 \left(\frac{m_5}{4m}\right)^{N/2} \int_{-\pi/2}^{\pi/2} e^{\omega\Im(a)} e^{-2\sqrt{m|z|}\cos(\theta/2)\cos\omega}\, d\omega.$$

Using Lemma 1 on $\cos\omega$ and the V_0-part of Lemma 2, we obtain the $\mathcal{E}_{2,N}$ bound of the theorem. ∎

With Theorem 6 we have come far enough to see that a Laguerre evaluation can be obtained—up to a subexponentially small relative error—via the A_h terms in said theorem. To this end, an inspection of the defining relations reveals that, in general, we can decompose an A_h coefficient in terms of powers of $v := e^{-i\omega}$, namely, we define $\alpha_{h,\mu}$ terms via

$$A_h =: \sum_{u=0}^{h} \alpha_{h,u}(a,z) v^{h+2u}. \tag{24}$$

For example,

$$\alpha_{00} = 1, \quad \alpha_{10} = -\frac{1+a}{2} z^{1/2}, \quad \alpha_{11} = \frac{z^{3/2}}{12}, \quad \alpha_{31} = \frac{1}{480}\left(5a^2 + 15a + 16\right) z^{5/2}.$$

The point being, we now have special formulae for the dominant contribution c_0, namely,

$$c_0 = \frac{1}{2\pi} r^{-a} e^{-z/2} \sum_{h=0}^{N-1} \frac{1}{m^{h/2}} \sum_{u=0}^{h} \alpha_{h,u}(a,z)\; \mathcal{I}(2\sqrt{mz}, a+h+2u) \; + \quad S_n(a,z)\; \mathcal{E}_{2,N}, \tag{25}$$

where the integral

$$\mathcal{I}(p,q) := \int_{-\pi/2}^{\pi/2} e^{-iq\omega} e^{p\cos\omega}\, d\omega \tag{26}$$

thus emerges as a fundamental entity for the research at hand.

5. $\mathcal{I}$-integrals and an "exp-arc" method. Having reduced the problem of subexponential Laguerre growth to a study of the $\mathcal{I}$-integrals (26), we next develop a method that is effective for both their numerical and theoretical estimation. This method amounts to the avoidance of stationary-phase techniques, employing instead various forms of *exponential-arcsine* (exp-arc) series, as we see shortly.

Let us first define, for any complex pair (p,q) and $\alpha,\beta \in (-\pi,\pi)$,

$$I(p,q,\alpha,\beta) := \int_{\alpha}^{\beta} e^{-iq\omega} e^{p\cos\omega}\, d\omega \tag{27}$$

so that our special case (26) is simply $\mathcal{I}(p,q) := I(p,q,-\pi/2,\pi/2)$.

Importantly, one may write the Bessel functions J_n of integer order n in terms of $\mathcal{I}$-integrals:

$$J_n(z) = \frac{1}{2\pi}\left(e^{-i\pi n/2}\mathcal{I}(iz,n) + e^{i\pi n/2}\mathcal{I}(-iz,n)\right), \tag{28}$$

and the modified Bessel function, again of integer order n:

$$I_n(z) = \frac{1}{2\pi}\left(\mathcal{I}(z,n) + (-1)^n\mathcal{I}(-z,n)\right), \tag{29}$$

about which representations we shall have more to say in a later section.[5]

5.1. Essentials of the exp-arc method. Now we investigate what we call exp-arc series. First, for any complex τ and $x \in [-1,1]$, one has a remarkable, absolutely convergent expansion (see [6]):

$$e^{\tau \arcsin x} = \sum_{k=0}^{\infty} r_k(\tau)\,\frac{x^k}{k!}, \tag{30}$$

where the coefficients depend on the parity of the index

$$r_{2m+1}(\tau) := \tau \prod_{j=1}^{m}\left(\tau^2 + (2j-1)^2\right), \qquad r_{2m}(\tau) := \prod_{j=1}^{m}\left(\tau^2 + (2j-2)^2\right).$$

By differentiating with respect to x, we obtain

$$\frac{e^{\tau \arcsin x}}{\sqrt{1-x^2}} = \frac{1}{\tau}\sum_{k=0}^{\infty} r_{k+1}(\tau)\,\frac{x^k}{k!},$$

valid for $x \in (-1,1)$. In particular, we have the important expansion (using a function G, in passing)

$$G(\tau,x) := \frac{\cosh(\tau \arcsin x)}{\sqrt{1-x^2}} = \sum_{k=0}^{\infty} g_k(\tau)\frac{x^{2k}}{(2k)!}, \tag{31}$$

[5]The $\mathcal{I}$-integral can also be written in compact terms of an *Anger function* **J** and *Weber function* **E**; see [1].

where

$$g_k(\tau) := \prod_{j=1}^{k} \left((2j-1)^2 + \tau^2\right).$$

These exp-arc expansions may be applied to the $I(p,q,\alpha,\beta)$ integrals as follows.[6] From (27) and its subsequent manipulations we have

$$I(p,q,\alpha,\beta) = e^p \int_\alpha^\beta e^{-iq\omega} e^{-2p\sin^2(\omega/2)}\, d\omega = 2e^p \int_{\sin\frac{\alpha}{2}}^{\sin\frac{\beta}{2}} \frac{e^{-2iq\arcsin x}}{\sqrt{1-x^2}} e^{-2px^2}\, dx$$

$$= i\frac{e^p}{q} \sum_{k=0}^{\infty} \frac{r_{k+1}(-2iq)}{k!} \int_{\sin\frac{\alpha}{2}}^{\sin\frac{\beta}{2}} x^k e^{-2px^2}\, dx. \tag{32}$$

Even though we eventually consider asymptotic expansions, the convergence of such an I-series is the rule.

LEMMA 11. *For $\alpha,\beta \in (-\pi,\pi)$ and any complex pair (p,q), the series* (32) *converges absolutely.*

Proof. Define $\delta := \max(|\sin(\alpha/2)|, |\sin(\beta)/2|)$ so that the integral in (32) is bounded in magnitude by $2\exp(2\Re(p))\delta^{k+1}/(k+1)$. But this means the kth summand in (32) is, up to a k-independent multiplier, bounded in magnitude by $r_{k+1}(2|q|)\delta^{k+1}/(k+1)!$. This summand is of the same form as that in the defining series (30), so absolute convergence is assured. ∎

Likewise, from (31) and (32) we have an absolutely convergent expansion for the special-case $\mathcal{I}$ integral:

$$\mathcal{I}(p,q) = 4e^p \int_0^{1/\sqrt{2}} G(-2iq,x) e^{-2px^2} = 4e^p \sum_{k=0}^{\infty} \frac{g_k(-2iq)}{(2k)!} B_k(p), \tag{33}$$

where B_k is an error–function-class integral

$$B_k(p) := \int_0^{1/\sqrt{2}} x^{2k} e^{-2px^2}\, dx = \frac{1}{2}\frac{1}{(2p)^{k+1/2}} \left\{\Gamma(k+1/2) - \Gamma(k+1/2,p)\right\}. \tag{34}$$

A computationally key recurrence relation accrues: with $B_0(p) = \sqrt{\frac{\pi}{8p}}\,\mathrm{erf}\left(\sqrt{p}\right)$ and $k > 0$, we have

$$B_k(p) := \frac{2k-1}{4p} B_{k-1} - \frac{2^{-k-3/2}}{p} e^{-p}. \tag{35}$$

5.2. Effective expansion for the cosh-arc *G*-series. Our asymptotic analysis of representation (33) begins with a lemma that reveals how the G function (31) has an attractive self-similarity property. Namely, the function appears naturally in modified form within its own error terms.

LEMMA 12. *For any complex τ, the G function* (31) *can be given an effective expansion to any integer order $N \geq 0$, as*

$$G(\tau,x) := \frac{\cosh(\tau\arcsin x)}{\sqrt{1-x^2}} = \sum_{k=0}^{N-1} g_k(\tau)\frac{x^{2k}}{(2k)!} + g_N(\tau)\frac{x^{2N}}{(2N)!}\left(1 + T_N(\tau,x)\right), \tag{36}$$

[6] An equivalent approach to expanding I uses a series for $e^{iq\omega}\sec(\omega/2)$ in terms of $x := \sin(\omega/2)$, as in [20, p. 276].

with the error term T_N conditionally bounded over real $x \in [0, 1/\sqrt{2}]$ in the form

$$\begin{aligned} T_N &= \Theta : 1 && \text{if } N \geq 2x^2|\tau|^2 - 1; \\ &= \Theta : (\sqrt{2}\ x^2 + x|\tau|)e^{|\tau| \arcsin x} && \text{otherwise.} \end{aligned}$$

Proof. The error term is, by the definition of the $g_k(\tau)$, given by the absolutely convergent sum

$$T_N = h_N + h_N h_{N+1} + h_N h_{N+1} h_{N+2} + \cdots,$$

where

$$h_k := \frac{(2k+1)^2 + \tau^2}{(2k+1)(2k+2)} x^2.$$

When $|\tau|^2 x^2 \leq (N+1)/2$ and since $x^2 \leq 1/2$, it is immediate that, for $k \geq N$, we have a bound:

$$|h_k| \leq \frac{(4k^2+4k+1)/2 + (N+1)/2}{4k^2+5k+2} \leq 1/2,$$

whence $T_N = \Theta : 1/2 + 1/4 + 1/8 + \ldots$, settling the first conditional bound of the theorem. In any case, i.e., any complex τ and any $x \in [0, 1/\sqrt{2}]$, we have for $j \geq 0$:

$$\begin{aligned} h_N h_{N+1} \cdots h_{N+j} &= x^{2j} \prod_{k=0}^{j} \frac{(2N+2k+1)^2 \left((2k+1)^2 + \tau^2 \frac{(2k+1)^2}{(2N+2k+1)^2} \right)}{(2N+2k+1)(2N+2k+2)(2k+1)^2} \\ &= \Theta : \frac{g_{j+1}(|\tau|)}{(2j+1)!!^2}. \end{aligned}$$

However, it is elementary that $(2j+1)!!^2 \geq (2j+1)!$ by simple factor-tallying, so

$$|T_N| \leq \frac{\left(1^2 + |\tau|^2\right) \cdot 2}{2!} x^2 + \frac{\left(1^2 + |\tau|^2\right)\left(3^2 + |\tau|^2\right) \cdot 4}{4!} x^4 + \ldots = x \frac{\partial}{\partial x} G(|\tau|, x),$$

where we have noticed that the right-hand series here is itself a differentiated "cosh-arc" series. Thus

$$T_N = \Theta : x \frac{\partial}{\partial x} \frac{\cosh(|\tau| \arcsin x)}{\sqrt{1-x^2}} = \Theta : \frac{x^2}{2\left(1-x^2\right)^{3/2}} \left(e^u + e^{-u}\right) + \frac{x|\tau|}{2\left(1-x^2\right)} \left(e^u - e^{-u}\right),$$

where $u := |\tau| \arcsin x$. Now by excluding $2x^2|\tau|^2 \leq N+1$ for the second conditional bound of the lemma, we have $|\tau| \geq 1$, whence the e^{-u} terms can be ignored over $x \in [0, 1/\sqrt{2}]$, and the second conditional bound follows. ∎

5.3. Effective expansion of the $\mathcal{I}$-integral. We first invoke a classical lemma that bounds the incomplete gamma function for certain parameters.

LEMMA 13. *For integer $M \geq 0$, $\Re(z) \geq 0, z \neq 0$ and $|z| \geq 2M - 1$, we have*

$$\Gamma(M + 1/2, z) = \Theta : 2z^{M-1/2} e^{-z}.$$

Proof. Proofs of results such as these on incomplete-gamma bounds appear in various texts on special functions, e.g., [27, section 2.2, p. 110]. ∎

The results of the present section may now be applied to a general, effective expansion of the $\mathcal{I}$-integral whenever $\Re(p)$ is sufficiently positive.

THEOREM 7 (effective $\mathcal{I}$ expansion). *For the integral*

$$\mathcal{I}(p,q) := I(p,q,-\pi/2,\pi/2) = \int_{-\pi/2}^{\pi/2} e^{-iq\omega} e^{p\cos\omega}\, d\omega,$$

assume an integer expansion order $N \geq 1$. Assume $\phi := \arg(p) \in (-\pi/2, \pi/2)$ and conditions

$$\Re(p) \geq 2N+1, \quad \Re(p) \geq 2\pi|q|^2.$$

Then we have an effective expansion

$$\mathcal{I}(p,q) = \sqrt{\frac{2\pi}{p}}\; e^p \left\{ \sum_{k=0}^{N-1} \frac{g_k(-2iq)}{k!\; 8^k} \frac{1}{p^k} + \Theta : \sqrt{\frac{8}{\pi p}}\; e^{-p} \cosh\left(\frac{\pi}{2}|q|\right) + \frac{g_N(-2iq)}{N!\; 8^N} \frac{1}{p^N} \left(1 + \Theta : u_N \sec^{N+1/2}\phi\right) \right\}, \tag{37}$$

where the g_k are as in the cosh-arc expansion (31)*, and we may take*

$$u_N := 1 + 2^N + \frac{2^{3N+1}}{\pi \binom{2N}{N}}\ ; \tag{38}$$

however, on the extra condition $N \geq 4|q|^2 - 1$, taking $u_N := 1$ suffices.

Proof. The insertion of the series of Lemma 12 into representation (33) results in

$$\mathcal{I}(p,q) = 4e^p \left\{ \sum_{k=0}^{N} \frac{g_k(-2iq)}{(2k)!} B_k(p) + \frac{g_N(-2iq)}{(2N)!} \int_0^{1/\sqrt{2}} x^{2N} T_N(-2iq, x) e^{-2px^2}\, dx \right\},$$

where the $B_k(p)$ are given by (34). The sum over $k \in [0, N]$ here is thus

$$\frac{1}{2} \sum_{k=0}^{N} \frac{g_k(-2iq)}{(2k)!} \frac{\Gamma(k+1/2)}{(2p)^{k+1/2}} + \Theta : e^{-p} \sum_{k=0}^{N} \frac{|g_k(-2iq)|}{(2k)!} \frac{1}{2^{k+1/2} p},$$

where the Θ-term here follows from Lemma 13 on our condition $\Re(p) \geq 2N+1$. But this very Θ-term is bounded above by

$$\frac{e^{-p}}{p\sqrt{2}} \sum_{k=0}^{\infty} \frac{g_k(2|q|)}{(2k)!} \frac{1}{2^k} = \frac{e^{-p}}{p} \cosh\left(2|q| \arcsin\left(1/\sqrt{2}\right)\right),$$

so we have settled the summation and the $\cosh(\pi|q|/2)$ term in (37). Now consider the integral term

$$I_0 := \int_0^{1/\sqrt{2}} x^{2N} T_N(-2iq, x) e^{-2px^2}\, dx.$$

Define $\gamma := |q|^{-1}\sqrt{(N+1)/8}$. If $\gamma \geq 1/\sqrt{2}$, then by Lemma 12, we know that $T_N = \Theta(1)$, and our theorem follows in the $u_N := 1$ case. Otherwise, $\gamma < 1/\sqrt{2}$, and

we bound I_0 using two integrals

$$|I_0| \le \int_0^{1/\sqrt{2}} x^{2N} e^{-2\Re(p)x^2}\, dx + \int_\gamma^{1/\sqrt{2}} \left(\sqrt{2}\, x^{2N+2} + 2|q|\, x^{2N+1}\right) e^{2|q|\arcsin x - 2\Re(p)x^2}\, dx.$$

From Lemma 1 we know that the exponent here can be taken to be $\pi|q|x/\sqrt{2} - 2\Re(p)x^2$. For the assignments $\alpha := \pi|q|/\sqrt{8},\ \beta := 2\Re(p)$, we have $\beta > 4\alpha/\gamma$ so that by Lemma 2 there is a bound

$$\begin{aligned} I_0 &\le \frac{1}{2}\frac{\Gamma(N+1/2)}{(2\Re(p))^{N+1/2}} + \sqrt{2}\, V_{N+1}(\alpha,\beta,\gamma) + 2|q|\, V_{N+1/2}(\alpha,\beta,\gamma) \\ &\le \frac{1}{2}\frac{\Gamma(N+1/2)}{(2\Re(p))^{N+1/2}} + 2^N \frac{\Gamma(N+3/2)}{(2\Re(p))^{N+3/2}} + 2^{N+1}|q|\frac{\Gamma(N+1)}{(2\Re(p))^{N+1}}. \end{aligned}$$

Using $|q|^2 \le \Re(p)/(2\pi)$, $\Re(p) \ge 2N+1$ with $\Re(p) = |p|\cos\phi$ yields the N-dependent u_N form of the theorem. □

6. Effective asymptotics for $L_n^{(-a)}(-z)$. At last we may provide explicit terms for Laguerre expansions in the subexponential-growth regime, which regime turns out to be precisely characterized by the parameter-pair requirement: $(a,z) \in \mathbb{D}$.

First, for convenience, we recapitulate the thresholds for sufficiently large $m := n+1$ from our previous theorems:

$$m_0 := |z|/4, \quad m_1 := 5\left(|\Re(z)| + (|\Im(a)| + |\Im(z)|/2)^2\right),$$
$$m_2 := \Re(a), \quad m_3 := -(5/4)|z|\cos\theta, \quad m_4 = 4|z|, \quad m_5 := |z|(1+|a|+|z|/2)^2,$$
$$m_6 := (2N+1)^2\, \frac{\sec^2(\theta/2)}{|z|}, \quad m_7 := 4\pi^2\, \frac{\sec^2(\theta/2)}{|z|}(|a|+3N-3)^4.$$

Here, $\theta := \arg(z)$ as before, while m_6, m_7 involve an asymptotic expansion order $N \ge 1$. We are aware that the m_i bounds here are interdependent, and some are masked by others (e.g., m_0 is masked by m_4). The important quantity in our central result (Theorem 8, below) is simply $\max_{i\in[0,7]} m_i$.

Before delving into the central result, let us remind ourselves of previous nomenclature:

$$g_k(\tau) := \prod_{j=1}^{k} \left((2j-1)^2 + \tau^2\right), \tag{39}$$

$$a_k := (1+a)(-1)^k \left(\frac{v\sqrt{z}}{2}\right)^k + \left(1-(-1)^k\right)\frac{k}{k+2}\left(\frac{v\sqrt{z}}{2}\right)^{k+2}, \tag{40}$$

$$\sum_{h=0}^{\infty} A_h x^h := \exp\left\{\sum_{k\ge1} \frac{a_k}{k} x^k\right\}, \tag{41}$$

$$A_h := \sum_{u=0}^{h} \alpha_{h,u}(a,z) v^{h+2u}. \tag{42}$$

Note that the $\alpha_{h,u}$ coefficients are thus implicitly defined in terms of the original a_k functions.

6.1. The general subexponential expansion. Using Theorem 7 with $p := 2\sqrt{mz}$ and inserting this into formula (25), we arrive at our desired effective Laguerre expansion.

THEOREM 8 (effective Laguerre expansion). *Assume $(a, z) \in \mathbb{D}$. For asymptotic expansion order $N \geq 1$ and $m := n + 1$ sufficiently large in the sense $m > \max_{i\in[0,7]} m_i$, we have the expansion*

$$L_n^{(-a)}(-z) = \frac{1}{2\sqrt{\pi}} \frac{e^{-z/2} e^{2\sqrt{mz}}}{z^{1/4-a/2}\ m^{1/4+a/2}} \left\{ \sum_{j=0}^{N-1} \frac{C_j}{m^{j/2}} + \frac{\overline{C}_N}{m^{N/2}} + \mathcal{E}_1 + \mathcal{E}_{3,N} \right\}, \tag{43}$$

where the expansion coefficients C_j are given in finite form:

$$C_j := \sum_{k=0}^{j} \frac{1}{16^k\ k!} \frac{1}{z^{k/2}} \sum_{u=0}^{j-k} \alpha_{j-k,u}(a, z)\, g_k(-2i(a + j - k + 2u)), \tag{44}$$

while the error term $\overline{C}_N$ is bounded as

$$\begin{aligned} \overline{C}_N = \Theta : \sum_{v=1}^{N} \frac{1}{|z|^{v/2} 16^v v!} \left(1 + u_v \sec^{v+1/2} \frac{\theta}{2}\right) \\ \sum_{u=0}^{N-v} |\alpha_{N-v,u}(a, z)\, g_v(-2i(a + N - v + 2u))| \\ + \Theta : 4(m_5/4)^{N/2} \sec^{1/2} \tfrac{\theta}{2}, \end{aligned}$$

with u_v taking the v-dependent form of (38) *in Theorem* 7 *(with $q := a$ there). Finally, the term $\mathcal{E}_1$ is subexponentially small (from Theorem* 4*), as is*

$$\mathcal{E}_{3,N} = \Theta : \sqrt{\frac{4}{\pi\sqrt{mz}}} e^{-2\sqrt{mz}} \sum_{h=0}^{N-1} \frac{1}{m^{h/2}} \sum_{u=0}^{h} |\alpha_{h,u}(a, z)| \cosh\left(\frac{\pi}{2}|a + h + 2u|\right).$$

Proof. For brevity we leave out the details—all of which are straightforward, if tedious, applications of the previous theorems and formulae. ∎

We now have a resolution of the domain of subexponential growth as follows.

COROLLARY 1. *For $(a, z) \in \mathbb{D}$, the Laguerre polynomial grows subexponentially in the sense that, for order $N \geq 1$ and any $\varepsilon > 0$,*

$$L_n^{(-a)}(-z) = S_n(a, z) \left\{ \sum_{j=0}^{N-1} \frac{C_j}{m^{j/2}} + O\left(\frac{1}{m^{N/2}}\right) + O\left(e^{-(2-\varepsilon)\sqrt{m|z|}\cos(\theta/2)}\right) \right\},$$

with all coefficients and the implied big-O constant effectively bounded via our previous theorems. Moreover, for $(a, z) \notin \mathbb{D}$, the large-n growth is not subexponential. Thus, the precise domain of subexponential growth is characterized by $(a, z) \in \mathbb{D}$.

Proof. The given subexponential formula is a paraphrase of Theorem 8. Now assume that $z \in (-\infty, 0]$. We already know that $z = 0$ does not yield such growth (see (2)). Now, for z negative real, note that the integrals in (12, 14) are both decaying in large m. Finally, the integral in (11) has phase factor $|\exp(2\sqrt{mz}\cos\omega)| \leq 1$, and Lemma 6 show that c_1 also cannot grow subexponentially in m. ∎

This corollary echoes, of course, the classical Perron result (7), and we again admit that historical efforts derived the C_j coefficients in principle. The new aspects are (a) we have asymptotic coefficients and effective bounds for general $(a, z) \in \mathbb{D}$ parameters, and (b) we can develop in a natural way an algorithm for symbolic generation of said coefficients.

6.2. Algorithm for explicit asymptotic coefficients. Theorem 8 indicates that, to obtain actual C_j coefficients, we need the cosh-arc numbers $g_k(\tau)$ and the $\alpha_{h,u}(a, z)$ coefficients. Observe that the chain of relations starting with (39) is a prescription for the generation of the C_k. All of this can proceed via symbolic processing, noting that v is simply a placeholder throughout.

Remarkably, there is a fast algorithm that bypasses much of the symbolic tedium. First, we have an explicit recursion for A_h, with $A_0 := 1$, as

$$A_k = \frac{1}{k} \sum_{j=0}^{k-1} A_j \, a_{k-j} \tag{45}$$

as follows from differentiating (41) logarithmically and then comparing terms. Second, when we use (42) together with (45), we obtain a recursion *devoid* of the symbolic placeholder v, as

$$\alpha_{k,u} = \frac{1}{k} \sum_{j=0}^{k-1} \left(\alpha_{j,u} b_{k-j} + \alpha_{j,u-1} d_{k-j} \right), \tag{46}$$

where these new recursion coefficients are

$$b_h := (-1)^h (1 + a) \left(\frac{\sqrt{z}}{2} \right)^h, \quad d_h := \left(1 - (-1)^h \right) \frac{h}{h+2} \left(\frac{\sqrt{z}}{2} \right)^{h+2}.$$

In practice we define $\alpha_{0,0} := 1$ and force any $\alpha_{j,u}$, with $u > j$ or $u < 0$ to vanish. In this sense, the collection of $\alpha_{k,u}$ make up a *lower-triangular matrix*, e.g., the entries for $k \le 3$ appear thus:

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ -\frac{(a+1)}{2} z^{1/2} & \frac{1}{12} z^{3/2} & 0 & 0 \\ \frac{a^2+3a+2}{8} z & -\frac{(a+1)}{24} z^2 & \frac{1}{288} z^3 & 0 \\ -\frac{a^3+6a^2+11a+6}{48} z^{3/2} & \frac{5a^2+15a+16}{480} z^{5/2} & -\frac{a+1}{576} z^{7/2} & \frac{1}{10368} z^{9/2} \end{pmatrix},$$

where $\alpha_{3,3}$ is the lower-right element here.

These observations lead to a fast algorithm for the computation of the asymptotic coefficients.

ALGORITHM 1 (fast computation of Laguerre asymptotic coefficients).
For given $(a, z) \in \mathbb{D}$ and desired expansion order N, this algorithm returns the asymptotic coefficients $(C_k : k \in [0, N])$ of relation (44), Theorem 8.

1. Set $\alpha_{0,0} := 1$ and, for desired order N, calculate the lower-triangular matrix elements
$(\alpha_{k,u} : 0 \le u \le k \le N)$ via a recursion such as
$\alpha(k, u)$ {

if($k == 0$) return $\delta_{0,u}$;
return $\frac{1}{k}\sum_{j=0}^{k-1}\left(\alpha_{j,u}b_{k-j} + \alpha_{j,u-1}d_{k-j}\right)$;
}
or use an unrolled, equivalent loop (i.e., one may generate the left-hand column of the α-matrix, then fill in one row at a time, lexicographically).
2. Use the recursion $g_k(\tau) = \left((2k+1)^2 + \tau^2\right) g_{k-1}(\tau)$ and the lower-triangular matrix of α values to generate the C coefficients via (44). □

We observe that the expensive sums in this algorithm are all *acyclic convolutions*. Thus, for numerical input (a, z), the algorithm complexity turns out to be $O\left(N^{2+\epsilon}\right)$ arithmetic operations, with the "2" part of the complexity power arising from the area of the lower-triangular sector.[7]

We employed the algorithm to generate exact asymptotic coefficients as follows:

$$C_0 = 1, \qquad C_1 = \frac{1}{48\sqrt{z}}\left(-12a^2 - 24za + 4z^2 - 24z + 3\right),$$
$$C_2 = \frac{1}{4608z}\left(144a^4 + 576za^3 + 480z^2a^2 + 1728za^2 - 360a^2 - 192z^3a \right.$$
$$\left. + 1152z^2a + 1584za + 16z^4 - 192z^3 + 312z^2 + 432z + 81\right),$$

and so on. Note that the C_1 form here agrees with the PAMO coefficient in (8). We were able to generate the full, symbolic $C_{64}(a, z)$ in about one minute of CPU on a typical desktop computer. To aid future researchers, we report that the numerator of C_{64} has degree 128 in both a, z, while the denominator is an integer divisible by every prime not exceeding 64, namely,

$$2^{319} \cdot 3^{94} \cdot 5^{25} \cdot 7^{13} \cdot 11^7 \cdot 13^5 \cdot 17^4 \cdot 19^3 \cdot 23^3 \cdot 29^2 \cdot 31^2 \cdot 37 \cdot 41 \cdot 43 \cdot 47 \cdot 53 \cdot 59 \cdot 61.$$

While formidable, the bivariate $C_k(a, z)$ coefficients evidently have intricate structure and patterning. Future research into said structure would be of interest.

6.3. Generating and verifying effective bounds. For the first nontrivial effective bound, we can employ the rigorous bound $\overline{C}_1$ in Theorem 8, with $N = 1$, to get an effective version of the original asymptotic (3) as

$$L_n^{(-a)}(-z) = S_n(a, z)\left(1 + \frac{\overline{C}_1}{\sqrt{m}} + \ \mathcal{E}_1 + \mathcal{E}_{3,1}\right),$$

with

$$\overline{C}_1 = \Theta : \frac{\left|1 - 4a^2\right|}{16|z|^{1/2}}\left(1 + 6\sec^{3/2}\frac{\theta}{2}\right) + 2|z|^{1/2}(1 + |a| + |z|/2)\sec^{1/2}\frac{\theta}{2},$$

and we remind ourselves that $\mathcal{E}_{3,1}$ (Theorem 8) and $\mathcal{E}_1$ (Theorem 4) are both subexponentially small.

At last we have an effective numerator, then, for the $1/\sqrt{m}$ asymptotic term. Though this effective numerator is almost surely nonoptimal, we are evidently on the right track, because the exact C_1 asymptotic coefficient above (see (8)) has very much the same form as does the Θ-expression for $\overline{C}_1$ here (i.e., the same degrees

[7]For example, floating-point FFT-based convolutions of length L require $O(L \log L)$ operations, of complexity less than $O\left(L^{1+\epsilon}\right)$.

TABLE 1

Examples of rigorous accuracy versus asymptotic order. For the indices n and parameters $(a, z) \in \mathbb{D}$, numbers $\mathcal{L}_{n,N}$ are computed via the asymptotic series through term $C_{N-1}/(n+1)^{(N-1)/2}$ as in Theorem 8. The relative error (RE) and rigorous upper bound (RRE) on the relative errors in taking these terms is reported. All decimal values are reported as correct to ± 1 in the last displayed digit of $\Re, \Im$ parts. Such rigorous bounds can be used to establish indisputable accuracies for various other functions.

n	a, z	N	Approximation $\mathcal{L}_{n,N}$ to $L_n^{(-a)}(-z)$	RE	RRE
1300	$0, -1 + 4i$	1	$(7 - i) \cdot 10^{37}$	0.0038	0.43
10^7	$0, 1$	1	$1.59 \cdot 10^{2744}$	0.00011	0.00099
		2	$1.595190 \cdot 10^{2744}$	$1.4 \cdot 10^{-9}$	$2.8 \cdot 10^{-7}$
		3	$1.5951907764 \cdot 10^{2744}$	$5.4 \cdot 10^{-13}$	$9.2 \cdot 10^{-11}$
10^6	$\frac{1}{2} + i, \frac{1}{2} + 4i$	1	$(-1.5 - 2.0i) \cdot 10^{1303}$	0.0016	0.021
		2	$(-1.565 - 2.041i) \cdot 10^{1303}$	$1.6 \cdot 10^{-6}$	$9.8 \cdot 10^{-5}$
		3	$(-1.562555 - 2.041423i) \cdot 10^{1303}$	$1.4 \cdot 10^{-9}$	$4.2 \cdot 10^{-7}$
		4	$(-1.56255567 - 2.041423269i) \cdot 10^{1303}$	$1.3 \cdot 10^{-12}$	$1.4 \cdot 10^{-9}$
		5	$(-1.56255567881 - 2.04142326969i) \cdot 10^{1303}$	$1.3 \cdot 10^{-15}$	$5.9 \cdot 10^{-12}$
10^7	$-\frac{3}{2}, 1 - 100i$	1	$(2 + i) \cdot 10^{19520}$	0.027	0.41
		2	$(1.8 + 0.9i) \cdot 10^{19520}$	0.00035	0.034
		3	$(-1.85 + 0.95i) \cdot 10^{19520}$	$3.1 \cdot 10^{-6}$	0.0028
		4	$(-1.851 + 0.957i) \cdot 10^{19520}$	$2.1 \cdot 10^{-8}$	0.00023
		5	$(-1.8510 + 0.9575i) \cdot 10^{19520}$	$1.1 \cdot 10^{-10}$	$2.0 \cdot 10^{-5}$
		10	$(-1.8509993203 + 0.9575616603i) \cdot 10^{19520}$	$7.1 \cdot 10^{-23}$	$7.6 \cdot 10^{-11}$

of appearance for a, z, and similar coefficients). And, it is easy to see that the Θ-expression here is an upper bound on $|C_1|$ itself, as must of course be true; it is necessary that $\overline{C}_N \geq |C_N|$ for every asymptotic order N.

Table 1 shows the numerical instances of our asymptotic series from Theorem 8. It displays various indices n and parameters a, z, together with corresponding asymptotic orders N from the theorem, using the following nomenclature. Denote, for asymptotic order N,

$$T_{n,N} \;:=\; 1 + \frac{C_1}{m^{1/2}} + \cdots + \frac{C_{N-1}}{m^{(N-1)/2}},$$

so, for S_n being the assignment (4), we have an approximation $\mathcal{L}_{n,N}$ defined by

$$\mathcal{L}_{n,N} \;:=\; S_n T_{n,M} \;\approx\; L_n^{(-a)}(-z).$$

Thus, $\mathcal{L}_{n,1}$ is simply S_n, as $T_{n,1} = 1$ always. How good are these $\mathcal{L}_{n,N}$ approximations? The important right-most columns of Table 1 report relative error (call it RE), with

$$\text{RE} := \left| \frac{\mathcal{L}_{n,N} - L_n^{(-a)}(-z)}{L_n^{(-a)}(-z)} \right|,$$

and a *rigorous* upper bound on relative error (call it RRE), expressed in terms of Theorem 8's various error bounds, with

$$\text{RRE} := \left| \frac{\epsilon}{T_{n,N}} \right|,$$

where $\epsilon > 0$ can be taken to be any number not exceeding $\left|\overline{C}_N/m^{N/2} + \mathcal{E}_1 + \mathcal{E}_{3,N}\right|$. Note that the three error components in this last expression are all magnitude-bounded via Theorem 8; this is how the RRE table entry was constructed.

A convenient way to view RRE is to realize that $L_n^{(-a)}(-z)$ is rigorously known to lie in the interval $\mathcal{L}_{n,N} \cdot (1 - \mathrm{RRE}, 1 + \mathrm{RRE})$. The decimal representations in Table 1 are reflexive; i.e., there is ambiguity only in the respective final digits. (When a reported value is $(a + ib) \cdot 10^c$, then both a and b are correct to within ± 1 in their respective last digits.)

One might ask about the "first-missing" asymptotic term $C_N/m^{N/2}$—a term of interest in many asymptotic theories. This term is typically close to, but often greater than, the RE. As is common to many asymptotic developments, rules about the behavior of these first-missing terms are hard to establish when parameters (in our case a, z) roam over the complex plane. In any case, the first-missing term is often an order of magnitude or more greater than the RRE. Yet this discrepancy between first-missing and RRE terms is not as harmful as might appear, as we discuss next.

6.4. Example application of effective bounds: Rigor for incomplete gamma. We have mentioned research motives in section 1.1. A worked example of *rigorous* incomplete gamma evaluations is the following. Let us use our very first entry from Table 1, namely,

$$L_{1300}^{(0)}(1 - 4i) \;\approx\; (7 - i) \cdot 10^{37},$$

with a rather large RRE = 0.43. Though this is an order of magnitude larger than the relative error, the penalty paid is only a few digits of precision for the relevant incomplete gamma. Indeed, from (5),

$$\Gamma(0, -1 + 4i) = e^{1-4i} \sum_{n=0}^{1299} \frac{1}{n+1} \frac{1}{L_n^{(0)}(1 - 4i) L_{n+1}^{(0)}(1 - 4i)} \;+\; \delta,$$

where the error δ is bounded, using RRE = 0.43 and (4), by

$$|\delta| \;<\; \frac{\left|e^{1-4i}\right|}{(1 - \mathrm{RE})^2} \sum_{n \geq 1300} \frac{1}{n+1} \frac{1}{|S_n S_{n+1}|} \;<\; 10^{-76}, \tag{47}$$

the last bound using (4) and careful estimates on the summation over $n \geq 1300$. Thence the sum for $n \in [1, 1299]$ gives a rigorous, 76–digit-accurate value for $\Gamma(0, -1 + 4i)$, $(\approx 0.497 \,+\, 0.415\, i)$. Correspondingly, the actual absolute error in using the sum over $n \in [0, 1299]$ is about 10^{-78}, so the apparently poor RRE bound brings only a 2-digit penalty.

An additional application of Theorem 8 would be to use effective bounds to rule out zeros of L_n in the crossed (a, z)-plane, that is, to establish an m_0 such that the term $1 + \overline{C}_1/\sqrt{m}$ must be positive for all $m > m_0$ (see our Open problems section).

7. Brief remarks on oscillatory regimes. The exp-arc method has led to rigorous asymptotics for subexponential growth but not for the Fejér form (6) for $z \neq 0$ on the cut $(-\infty, 0]$. Such oscillatory behavior can be dealt with, but other techniques come into play. For one thing, contour integrals must be handled differently.

7.1. When z is on the negative real cut. Even on $z \in (-\infty, 0)$ the contour prescription of Figure 1 is valid, and the Laguerre polynomial is exactly the sum

$c_1+d_1+e_1$, with $R := \sqrt{m/|z|} > 1/2$ being the only requirement for contour validity. However—and this is important—the dominant contribution (17) has to change to involve an expanded integration interval; in fact, now we must use the contour integral c_1 itself as the main contribution. For reasons of brevity, we simply state the complete asymptotic result stemming from contour integral c_1 as

$$L_n^{(-a)}(-z) \;\sim\; \frac{e^{-z/2}}{\sqrt{\pi}(-z)^{1/4-a/2}\; m^{1/4+a/2}} \times \left\{ \left(\sum_{k=0}^{\infty} \frac{A_k}{m^k} \right) \cos\left(2\sqrt{-mz} + a\pi/2 - \pi/4\right) + \left(\sum_{k=0}^{\infty} \frac{B_k}{m^{k+1/2}} \right) \sin\left(2\sqrt{-mz} + a\pi/2 - \pi/4\right) \right\}, \tag{48}$$

where these oscillatory-series coefficients are directly related to the coefficients in Theorem 8 by

$$A_k := C_{2k}(a,z), \quad B_k := iC_{2k+1}(a,z).$$

It transpires that, for a real, every A_k, B_k is real, whence the asymptotic has all real terms. The Fejér–Perron–Szegö expansion in [36, Theorem 8.22.2] for the oscillatory Laguerre mode is stated there in a fashion structurally different from our asymptotic (48); notwithstanding this Szegö's own $A_{\text{odd}}, B_{\text{even}}$—also not defined quite like ours—vanish.[8]

7.2. Brief remarks on Bessel functions I_n, J_n. There is vast literature on Bessel asymptotics [44], [39], resulting in the *Hankel asymptotic series* [1, section 9.2.5] and effective bounds on error terms (for certain parameter domains). In many cases these restricted error bounds are nevertheless optimal [44]. It is instructive to explore, at least briefly, the application of our exp-arc method to Bessel expansions. We recall (29) which leads to

$$I_n(z) = \frac{2}{\pi} \sum_{k\geq 0} \frac{g_k(-2in)}{(2k)!} \left\{ e^z B_k(z) + (-1)^n e^{-z} B_k(-z) \right\} \;\sim\; \frac{e^z}{\sqrt{2\pi z}} \sum_{k=0}^{\infty} \frac{g_k(-2in)}{k!\; 8^k} \frac{1}{z^k}, \tag{49}$$

(again we omit proofs) where the first sum is convergent, exact, and the second sum agrees with the classical Hankel asymptotic [44], [1]. For $\arg(z) \in (-\pi/2, \pi/2)$, at least, one may derive effective bounds on I_n error terms using the exp-arc techniques.

Computationalists have known for decades that one way to evaluate Bessel functions uniformly in the argument z is to use the standard ascending series for small $|z|$, but an asymptotic series for large $|z|$. However, via the exp-arc method one can establish a converging series whose evaluation only involves *a single error-function evaluation*, followed by recursion and elementary algebra. In fact, the relations (34), (35) can be used to calculate $J_n(z)$ from relation (28) via the sum

$$J_n(z) = \frac{2}{\pi} \sum_{k=0}^{\infty} g_k(-2in) \left(b_k \cos\chi - c_k \sin\chi \right), \tag{50}$$

[8] Our oscillatory asymptotic (48) has been verified to several terms by Temme; also he verifies our claim that the classical Szegö coefficients do vanish for the parities indicated [38].

with

$$\chi := z - \pi n/2 - \pi/4, \quad b_k := B_k(iz)e^{i\pi/4} + B_k(-iz)e^{-i\pi/4}, \quad ic_k := B_k(iz)e^{i\pi/4} - B_k(-iz)e^{-i\pi/4}.$$

Note that if z is real, then each b_k, c_k is real, whence our series here has all real terms. Note that our recursion (35) likewise ignites a recursion amongst the b_k, c_k.

Note that (50) is actually the Hankel asymptotic if we replace B_k by its first term in (34), namely, $(1/2)(2iz)^{-k-1/2}\Gamma(k+1/2)$; however, we already know that the sum (50) is always convergent. It is remarkable that we are using the same structure as the classical asymptotic, yet convergence for all complex z is guaranteed. Moreover, the $B_k(iz)$ are independent of the order n and so can be reused if multiple $J_n(z)$ are desired for fixed z.

Those acquainted with the intricacies of Bessel theory may observe that our convergent expansion (49) is at least reminiscent of the convergent *Hadamard expansion* found in [44, p. 204] for the modified Bessel function I_ν. Though both expansions are absolutely convergent, there are some important differences between this Hadamard expansion and our exp-arc forms (49, 50). For example, we have given our convergent sum only for integer ν. Moreover, the exp-arc expansion is geometrically convergent, while the Hadamard expansion is genuinely slower.

The research area of convergent expansions related to classical, asymptotic ones has been pioneered in large part by Paris, whose works cover real and complex domains, saddle points, and the like [29], [30], [31], [32], [33], [34]. We should point out that Paris was able to develop within the last decade some similar, linearly convergent Bessel series by modifying the "tails" of Hadamard-class series. Finally,we note that the problem of generalizing such unconditionally convergent series as our (50) to noninteger indices—and with comparison to the recent work of Paris—is analyzed in a new work [3].

8. Open problems.

- How might one proceed with the exp-arc theory to obtain effective error bounds for oscillatory Laguerre modes and-or oscillatory Bessel modes? We know that previous researchers have described how to give effective bounds in these cases (e.g., to our asymptotic (48), as in [36, section 8.72]), but once again we stress, How can this be done explicitly and for full parameter ranges?
- Where are the zeros in the complex z-plane—for fixed a—of $L_n^{(-a)}(-z)$? Are "most of" the zeros along some a-dependent ray, in some sense? There is a considerable literature on this zero–free-region topic, especially for polynomials in real variables. For example, with $a := 0$ the Laguerre zeros are all real and negative; see [24, Chapter X] and references therein. There is also an interesting connection between Laguerre zeros and eigenvalues of certain (large) matrices [15]. Certainly the theorems of the Saff–Varga type are relevant in this context [24, Theorem 11].
- How can Laguerre asymptotics be gleaned from standard recurrence relations amongst the $L_n^{(-a)}(-z)$? One may ask the same question for the Laguerre differential equation as starting point. A promising research avenue for a discrete-iterative approach to asymptotics is [43]; see also [4] for the asymptotic analysis of certain complex continued fractions. As for differential-equation approaches, there is the classical work of Erdélyi and Olver, plus modern work on the combinations of differential, discrete, and saddle-point theory [16], [18].

- The efficient "keyhole" contour of Figure 1 was discovered experimentally. What other analytical problems might be approached in this fashion? For that matter, how might one properly use the celebrated Watson loop-integral lemma with error term [27, Theorem 5.1] on our keyhole contour to obtain similar effective asymptotics?
- Here is a scenario suggested by Temme [38]. When n is small (in some sense not made rigorous here), and, say, $|z| \ll |a|$, one should expect an asymptotic behavior

$$L_n^{(-a)}(-z) \sim \binom{n-a}{n} \left(1 + \frac{z}{1-a}\right)^n.$$

The problem is, How does one make this rigorous, with effective error bounds?

Acknowledgments. We are indebted to J. Buhler, R. Mayer, O-Yeat Chan, R. Paris, and N. Temme for the generous lending of expertise, ideas, and suggestions. We are equally grateful to two referees for their insightful remarks.

REFERENCES

[1] M. Ambramowitz and I. Stegun, *Handbook of Mathematical Functions*, NBS (now NIST), Washington, D.C., 1965.
[2] D. H. Bailey, J. M. Borwein, and R. E. Crandall, *Integrals of the Ising class*, J. Phys. A, 39 (2006), pp. 12271–12302.
[3] D. Borwein, J. Borwein, and O-Y. Chan, *The evaluation of Bessel functions via exp-arc integrals*, J. Math. Anal. Appl., 341 (2008), pp. 478–500.
[4] D. Borwein, J. Borwein, R. Crandall, and R. Mayer, *On the dynamics of certain recurrence relations*, Ramanujan J., 13 (2007), pp. 65–104.
[5] J. M. Borwein and P. B. Borwein, *Pi and the AGM: A Study in Analytic Number Theory and Computational Complexity*, John Wiley, New York, 1987. Paperback edition 1998.
[6] J. Borwein and M. Chamberland, *Integer powers of Arcsin*, Int. J. Math. Math. Sci., 2007, Article ID 19381, http://www.hindawi.com/GetArticle/aspx?doi=10.1155/2007/19381.
[7] J. Borwein, R. Crandall, and G. Fee, *On the Ramanujan AGM fraction. Part* I: *The real-parameter case*, Experiment. Math., 13 (2004), pp. 275–286.
[8] J. Borwein and R. Crandall, *On the Ramanujan AGM fraction. Part* II: *The complex-parameter Case*, Experiment. Math., 13 (2004), pp. 287–296.
[9] C. Bosbach and W. Gawronski, *Strong asymptotics for Laguerre polynomials with varying weights*, J. Comput. Appl. Math., 99 (1998), pp. 77–89.
[10] K. Chao and R. Plymen, *A new bound for the smallest x with $\pi(x) > \mathrm{li}(x)$*, arXiv:math/0509312v5.
[11] R. Crandall, *Unified algorithms for polylogarithm, L-series, and zeta variants*, http://www.reed.edu/∼crandall, 2008, preprint.
[12] R. Crandall, *Alternatives to the Riemann–Siegel formula*, http://www.reed.edu/∼crandall, 2008, preprint.
[13] R. Crandall, R. Bettega, and R. Whitnell, *A class of exactly soluble three-body problems*, J. Chem. Phys., 83 (1985), pp. 698–702.
[14] R. Crandall, R. Whitnell, and R. Bettega, *Exactly soluble two-electron atomic model*, Amer. J. Phys., 52 (1984), pp. 438–442.
[15] H. Dette, *Strong approximation of eigenvalues of large dimensional Wishart matrices by roots of generalized Laguerre polynomials*, J. Approx. Theory, 118 (2002), pp. 290–304.
[16] A. Erdélyi, *Asymptotic forms for Laguerre polynomials*, J. Indian Math. Soc., Golden Jubilee Commemoration Volume, 24 (1960), pp. 235–250.
[17] W. Foster and I. Krasikov, *Explicit bounds for Hermite polynomials in the oscillatory region*, LMS J. Comput. Math., 3 (2000), pp. 307–314.
[18] C. L. Frenzen and R. Wong, *Uniform asymptotic expansions of Laguerre polynomials*, SIAM J. Math. Anal., 19 (1988), pp. 1232–1248.
[19] P. Henrici, *Applied and Computational Complex Analysis: Special Functions, Integral Transforms, Asymptotics, Continued Fractions, Vol.* 2 *(Classics Library)*, John Wiley, New York, 1974 (reprint 1988).

[20] E. Hobson, *A Treatise on Plane Trigonometry*, Cambridge University Press, London, 1925.
[21] L. Jacobsen and W. Thron, *Oval convergence regions for continued fractions* $K(a_n/1)$, preprint, 1986.
[22] A. Kuijlaars and K. McLaughlin, *Riemann–Hilbert analysis for Laguerre polynomials with large negative parameter*, Comput. Methods Func. Theory, 1 (2001), pp. 205–233.
[23] J. López and N. Temme, *Approximations of orthogonal polynomials in terms of Hermite polynomials*, Methods Appl. Anal., 6 (1999), pp. 131–146.
[24] L. Lorentzen and H. Waadeland, *Continued Fractions with Applications*, North-Holland, Amsterdam, 1992.
[25] E. Love, *Inequalities for Laguerre functions*, J. Inequal. Appl., 1 (1997), pp. 293–299.
[26] F. Olver, *Why steepest descents?* SIAM Rev., 12 (1970), pp. 228–247.
[27] F. Olver, *Asymptotics and Special Functions*, Academic Press, New York, 1974.
[28] R. B. Paris, *private communication*, 2007.
[29] R. B. Paris, *On the use of Hadamard expansions in hyperasymptotic evaluation of Laplace-type integrals.* I: *Real variable*, J. Comput. Appl. Math., 167 (2004), pp. 293–319.
[30] R. B. Paris, *On the use of Hadamard expansions in hyperasymptotic evaluation of Laplace-type integrals.* II: *Complex variable*, J. Comput. Appl. Math., 167 (2004), pp. 321–343.
[31] R. B. Paris, *On the use of Hadamard expansions in hyperasymptotic evaluation of Laplace-type integrals.* III: *Clusters of saddles*, J. Comput. Appl. Math., to appear.
[32] R. B. Paris, *On the use of Hadamard expansions in hyperasymptotic evaluation of Laplace-type integrals.* IV: *Poles*, Comput. Appl. Math., to appear.
[33] R. B. Paris, *Exactification of the method of steepest descents: The Bessel functions of large order and argument*, Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 460 (2004), pp. 2737–2759.
[34] R. B. Paris and D. Kaminski, *Hadamard expansions for integrals with saddles coalescing with an endpoint*, Appl. Math. Lett., 18 (2005), pp. 1389–1395.
[35] O. Perron, *Über das Verhalten einer ausgearteten hypergeometrischen Reihe bei unbegrenztem Wachstum eines Parameters*, Journal für die reine und angewandte Math, 151 (1921), pp. 63–78.
[36] G. Szegö, *Orthogonal Polynomials*, Amer. Math. Soc. Colloq. Publ., 4th ed., 23, American Mathematical Society, Providence, RI, 1975.
[37] N. Temme, *Asymptotic estimates of Stirling numbers*, Stud. Appl. Math., 89 (1993), pp. 223–243.
[38] N. Temme, *private communication*, 2006.
[39] S. Thangavelu, *Lectures on Hermite and Laguerre Expansions*, Math. Notes 42, Princeton University Press, Princeton, N.J., 1993.
[40] E. Titchmarsh, *The Theory of the Riemann Zeta Function*, Clarendon Press, Oxford, 1951 & 1967.
[41] W. Van Assche, *Weighted zero distribution for polynomials orthogonal on an infinite interval*, SIAM J. Math. Anal., 16 (1985), pp. 1317–1334.
[42] W. Van Assche, *Erratum to "Weighted zero distribution for polynomials orthogonal on an infinite interval"*, SIAM J. Math. Anal., 32 (2001), pp. 1169–1170.
[43] Z. Wang and R. Wong, *Linear difference equations with transition points*, Liu Bie Ju Center for Math. Sci., preprint, 2004.
[44] G. Watson, *Treatise on the Theory of Bessel Functions*, 2nd ed., Cambridge Math. Lib. Ed., Cambridge University Press, Cambridge, 1966.
[45] S. Winitzki, *Computing the incomplete gamma function to arbitrary precision*, in Lecture Notes in Comput. Sci. 2667, Springer–Verlag, New York, 2003, pp. 790–798.
[46] S. Wong, *Computational Methods in Physics and Engineering*, Prentice–Hall, New Jersey, 1992.

SIAM J. NUMER. ANAL.
Vol. 46, No. 6, pp. 3285–3312
© 2008 Society for Industrial and Applied Mathematics

EFFECTIVE LAGUERRE ASYMPTOTICS*

DAVID BORWEIN†, JONATHAN M. BORWEIN‡, AND RICHARD E. CRANDALL§

Abstract. It is known that the generalized Laguerre polynomials can enjoy subexponential growth for large primary index. In particular, for certain fixed parameter pairs (a, z) one has the large-n asymptotic behavior

$$L_n^{(-a)}(-z) \sim C(a,z) n^{-a/2-1/4} e^{2\sqrt{nz}}.$$

We introduce a computationally motivated contour integral that allows efficient numerical Laguerre evaluations yet also leads to the complete asymptotic series over the full parameter domain of subexponential behavior. We present a fast algorithm for symbolic generation of the rather formidable expansion coefficients. Along the way we address the difficult problem of establishing effective (i.e., rigorous and explicit) error bounds on the general expansion. A primary tool for these developments is an "exp-arc" method giving a natural bridge between converging series and effective asymptotics.

Key words. Laguerre polynomials, effective asymptotic expansions, contour integrals

AMS subject classifications. 34E05, 33C65

DOI. 10.1137/07068031X

1. The challenge of "effectiveness." For primary integer indices $n = 0, 1, 2, \ldots$, we define the Laguerre polynomial thus:

$$L_n^{(-a)}(-z) := \sum_{k=0}^{n} \binom{n-a}{n-k} \frac{z^k}{k!}. \tag{1}$$

Our use of negated parameters $-a, -z$ is intentional, for convenience in our analysis and in connection with related research, as we later explain. We shall work on the difficult problem of establishing asymptotics *with* effective error bounds for the two-parameter domain

$$\mathbb{D} := \{(a,z) \in \mathbb{C} \times \mathbb{C} : z \notin (-\infty, 0]\}.$$

That is, a is any complex number, while z is a complex number not on the negative-closed cut $(-\infty, 0]$.

Herein, we say a function $f(n)$ is of *subexponential* growth if $\log f(n) \sim Cn^D$ as $n \to \infty$ for positive constants C, D, with $D < 1$. It will turn out that $\mathbb{D}$ is the precise domain of subexponential growth of $L_n^{(-a)}(-z)$ as $n \to \infty$, with (a, z) fixed. The reason for the negative-cut exclusion on z is simple: For z negative real, the Laguerre polynomial exhibits oscillatory behavior in large n and is not of subexponential growth. Note also, from the definition (1), that

$$L_n^{(-a)}(0) = \binom{n-a}{n}, \tag{2}$$

which covers the case $z = 0$; again, not subexponential growth.

*Received by the editors January 17, 2007; accepted for publication (in revised form) May 29, 2008; published electronically October 17, 2008.
http://www.siam.org/journals/sinum/46-6/68031.html

†Department of Mathematics, University of Western Ontario, London, ON, N6A 5B7, Canada (dborwein@uwo.ca).

‡Faculty of Computer Science, Dalhousie University, Halifax, NS, B3H 2W5, Canada, and School of Mathematical and Physical Sciences, University of Newcastle, Callaghan, NSW 2300, Australia (jborwein@cs.dal.ca). This author's work was funded by NSERC and the Canada Research Chair Program.

§Center for Advanced Computation, Reed College, Portland, OR 97202 (crandall@reed.edu).

For fixed $(a,z) \in \mathbb{D}$, we shall have the large-n behavior (here and beyond we define $m := n+1$ which will turn out to be a natural reassignment):

$$L_n^{(-a)}(-z) \;\sim\; S_n(a,z)\left(1 + O\left(\frac{1}{m^{1/2}}\right)\right), \tag{3}$$

where the subexponential term S_n is

$$S_n(a,z) \;:=\; \frac{e^{-z/2}}{2\sqrt{\pi}}\,\frac{e^{2\sqrt{mz}}}{z^{1/4-a/2}\;m^{1/4+a/2}}. \tag{4}$$

In such expressions, $\Re\,(\sqrt{mz})$ denotes $\sqrt{m|z|}\cos(\theta/2)$, where $\theta := \arg(z) \in (-\pi,\pi]$ (we hereby impose $\arg(-1) := \pi$), and so for $(a,z) \in \mathbb{D}$, the expression (4) involves genuinely diverging growth in n.

1.1. Research motives. There are many interdisciplinary applications of Laguerre asymptotics. The Laguerre functions appear standardly in the quantum theory of the hydrogen atom [46, Chapter 4] and in certain exactly solvable three-body problems of chemical physics [13], [14]. The so-called WKB phase of a quantum eigenstate can, in such cases and for high quantum numbers, be calculated via Laguerre asymptotics. The Hermite polynomials $H_n(x)$—closely related to the Laguerre polynomials $L_n^{(\pm 1/2)}(x^2)$—also figure in quantum analyses; e.g., one may derive Hermite asymptotics from Laguerre asymptotics. In this context there is the fascinating López–Temme representation of *any* $L_n^{(-a)}(-z)$ as a finite superposition of Hermite evaluations [23].

Our own research motive for providing effective asymptotics involves not the Laguerre–Hermite connection, rather it begins with a beautiful link to the incomplete gamma function, namely, [19]

$$\Gamma(a,z) \;=\; z^a e^{-z}\cfrac{1}{z+\cfrac{1-a}{1+\cfrac{1}{z+\cfrac{2-a}{1+\cdots}}}} \;=\; z^a e^{-z}\sum_{n=0}^{\infty}\frac{(1-a)_n}{(n+1)!}\,\frac{1}{L_n^{(-a)}(-z)\,L_{n+1}^{(-a)}(-z)}, \tag{5}$$

where $(c)_n := c(c+1)\cdots(c+n-1)$ is the *Pochhammer symbol*. This series is valid whenever none of the Laguerre denominators has a zero. Thus an interesting sidelight is the research problem of establishing zero-free regions for Laguerre polynomials (see our Open problems section). We note that convergence of the continued-fraction can be highly problematic, especially when the full complex parameter space is allowed. One might encounter smooth convergence, or steady divergence, or even chaotic divergence, all of which possibilities are exemplified in [7, 8, 9, 22].

The problem, then, of general Laguerre asymptotics arises because there is a way to write Riemann zeta-function evaluations $\zeta(s)$ in terms of the incomplete gamma function, as was known to Riemann himself [40, p. 22]. But when s has large imaginary height, one becomes interested in the incomplete gamma's corresponding complex parameters (a,z) also of possibly large imaginary height [11], [12]. For example, in the art of prime-number counting, say, for primes $< 10^{20}$, one might need to know, for example, $\Gamma\left(3/4 + 10^{10}\,i, z\right)$ for various z also of large imaginary height, to good precision, in order to evaluate $\zeta(s)$ high up on the line $\Re(s) = 3/2$. The point is,

exact computations on prime numbers must, of course, employ rigor—hence the need for effective error bounds.

Independent of Riemann-ζ considerations is the following issue: Continued-fraction theory is to this very day incomplete in a distinct sense. The vast majority of available convergence theorems are for Stieltjes, or S-fractions. Now the continued fraction in (5) is an S-fraction only when the a parameter is real [24, p. 138]. Due to the relative paucity of convergence theorems outside the S-fraction class, one encounters great difficulty in estimating the convergence rate for arbitrary $\Gamma(a,z)$; this is what led to our focus on the Laguerre asymptotics. Incidentally, we are aware that subexponential convergence results *might* be attainable via the complicated and profound work of Jacobsen and Thron [21] on oval convergence regions. In any case, our effective-Laguerre approach proves the subexponential convergence of $\Gamma(a,z)$ for any complex pairs $(a,z)\in\mathbb{D}$; moreover, this is done with effective constant factors. Separate research on these superexponential convergence issues for continued fractions is underway.

1.2. Historical results on Laguerre asymptotics. Laguerre asymptotics have long been established for certain restricted domains and usually with noneffective asymptotics.[1] For example, in 1909 Fejér established that, for z on the open cut $(-\infty,0)$ and any real a, one has [36, Theorem 8.22.1]:

$$L_n^{(-a)}(-z)=\frac{e^{-z/2}}{\sqrt{\pi}(-z)^{1/4-a/2}\ m^{1/4+a/2}}\cos\left(2\sqrt{-mz}+a\pi/2-\pi/4\right)+O\left(m^{-a/2-3/4}\right), \tag{6}$$

where we again use index $m:=n+1$, which slightly alters the coefficients in such classical expansions, said coefficients being for powers $n^{-k/2}$, but we are now using powers $m^{-k/2}$. By 1921 Perron [35] had generalized the Fejér series to arbitrary orders, then, for $z\notin(-\infty,0]$, established a series consistent with (3) and (4), in essentially the form [36, Theorem 8.22.3]:

$$L_n^{(-a)}(-z)=S_n(a,z)\left(\sum_{k=0}^{N-1}\frac{C_k}{m^{k/2}}+O\left(m^{-N/2}\right)\right), \tag{7}$$

although this was for *real* a and so not for our general parameter domain $\mathbb{D}$. Note that $C_0=1$, consistent with our (3) and (4); however, one should take care that because we are using index $m:=n+1$, the coefficients C_k in the above formula differ slightly from the historical ones.

A modern literature treatment that is again consistent with the heuristic (3)–(4) is given by Winitzki in [45], where one invokes a formal generating function to yield a contour integral for $L_n^{(-a)}(-z)$. Then a stationary-phase approach yields the correct first-asymptotic term, at least for certain subregions of $\mathbb{D}$. Winitzki's treatment is both elegant and nonrigorous; there is no explicit estimate given on the $O(1/\sqrt{m})$ correction in (3).

There is an interesting anecdote that reveals the difficulty inherent in Laguerre asymptotics. Namely, W. Van Assche in a fine 1985 paper [41] used the expansion

[1] Some researchers use the term "realistic error bound" for big-O terms that have explicit structure. We prefer "effective bound," and when an expansion is bestowed with such a bound, we may say "effective expansion."

(7) for work on zero-distributions, only to find by 2001 that the C_1 term in that 1985 paper had been calculated incorrectly. The amended series is given in his correction note [42] as

$$L_n^{(-a)}(-z) = \frac{e^{-z/2}}{2\sqrt{\pi}} \frac{e^{2\sqrt{nz}}}{z^{1/4-a/2}\, n^{1/4+a/2}} \cdot \left(1 + \left(\frac{3 - 12a^2 + 24(1-a)z + 4z^2}{48\sqrt{z}}\right) \frac{1}{\sqrt{n}} + O\left(\frac{1}{n}\right)\right),$$

or in our own notation with $m := n + 1$,

$$= S_n(a,z) \left(1 + \left(\frac{3 - 12a^2 - 24(1+a)z + 4z^2}{48\sqrt{z}}\right) \frac{1}{\sqrt{m}} + O\left(\frac{1}{m}\right)\right). \tag{8}$$

Note the slight alteration used to obtain our $C_1/\sqrt{m}$ term. Van Assche credits T. Müller and F. Olver for aid in working out the correct $O(1/\sqrt{n})$ component.[2] This story suggests that even a low-order asymptotic development is nontrivial.

The great classical analysts certainly knew in principle how to establish effective error bounds. The excellent treatment of effectiveness for Laplace's method of steepest descent in [27] is a shining example. Also illuminating is Olver's paper [26], which explains effective bounding and shows how unwieldy rigorous bounds can be obtained. However, efficient algorithms for generating *explicit* effective big-O constants have only become practicable in recent times, when computational machinery is prevalent.

2. Contour representation. In this section we develop an efficient—both numerically and analytically—contour-integral representation for $L_n^{(-a)}(-z)$.

2.1. Development of a "keyhole" contour. A well-known integral [36, section 5.4] has contour Γ encircling $s = 1$ and avoiding the branch cut $(-\infty, 0]$:

$$L_n^{(-a)}(-z) = \frac{e^{-z}}{2\pi i} \oint_\Gamma s^{-1-a} \left(1 - \frac{1}{s}\right)^{-n-1} e^{zs}\, ds. \tag{9}$$

This representation holds for all pairs $(a, z) \in \mathbb{C} \times \mathbb{C}$, with n a nonnegative integer.[3]

However—and this is important—we found via experimental mathematical techniques, e.g., extreme-precision evaluations, that a specific kind of contour allows very accurate, efficient, and well-behaved *numerical* Laguerre evaluations. It is not completely understood why the "keyhole" contour we are about to define does so well in numerical Laguerre evaluations; we do know that this new contour consistently provides better numerics than, say, a simple circle surrounding $s = 1$. It may be simply the internal quirks of various modern numerical integrators at work, or it may be the smooth phase behavior along our keyhole's perimeter.

For the desired contour, take $z \neq 0$, $m := n + 1$ and assume $r := \sqrt{m/z}$ has $|r| > 1/2$. Then, use a circular contour centered at $s = 1/2$ with radius $|r|$. This contour will encompass $s = 1$, so the remaining requirement is to avoid the cut $s \in (-\infty, 0]$ as can be done by cutting out a "wedge" from the negative-real arc of the circle, with apex at $s = 1/2$. We tried such schemes with high-precision integration, to

[2] Accordingly, we hereby name the polynomial $C_1(a, z)$ the Perron–van Assche–Müller–Olver (PAMO) coefficient.

[3] The contour representation (9) can easily be continued to noninteger n, with care taken on the $(-n-1)$th power, but our present treatment will only use nonnegative integer n.

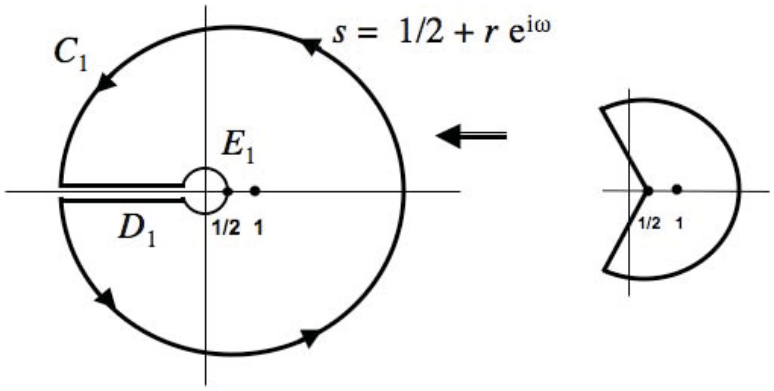


FIG. 1. *A numerically efficient "keyhole contour" for Laguerre evaluations $L_n^{(-a)}(-z)$, valid for all complex a, z, with $z \neq 0$ and $n+1 > |z|/4$. Wedges with center-1/2—as pictured at right—are experimentally accurate, leading to a "keyhole" deformation avoiding the cut $s \in (-\infty, 0]$. It turns that, for any $(a, z) \in \mathbb{D}$, the main arc C_1 gives the predominant contribution for large n, the D_1, E_1 components being subexponentially minuscule.*

settle finally on the contour of Figure 1, where the aforementioned wedge has evolved to a "keyhole" pattern consisting of cut-run D_1 and small, origin-centered circle E_1 of radius 1/2.

So, adopting constraints and nomenclature

$$z \neq 0, \quad \theta := \arg(z), \quad \omega_\pm := \pm\pi + \theta/2,$$
$$m := n + 1, \quad r := \sqrt{m/z} := \sqrt{m/|z|}e^{-i\theta/2}, \quad R := |r| > 1/2,$$

but no other constraints, we have the following representation:

$$L_n^{(-a)}(-z) = c_1 + d_1 + e_1, \tag{10}$$

where c_1, d_1, e_1 are the respective contributions from contour C_1, cut-discontinuity D_1, and contour E_1 from Figure 1. Exact formulae for said contributions are

$$c_1 = \frac{1}{2\pi} r^{-a} e^{-z/2} \int_{\omega_-}^{\omega_+} \mathcal{H}_m(a, z, e^{-i\omega}) e^{2\sqrt{mz}\cos\omega}\, d\omega, \tag{11}$$

$$d_1 = \frac{e^{-z}}{\pi} \sin(\pi a) \int_{1/2}^{R-1/2} T^{-1-a} \left(1 + \frac{1}{T}\right)^{-m} e^{-zT}\, dT, \tag{12}$$

$$e_1 = -\frac{e^{-z}}{4\pi} \int_{-\pi}^{\pi} \left(2e^{-i\omega}\right)^{1+a} \left(1 - 2e^{-i\omega}\right)^{-m} e^{i\omega + \frac{z}{2}e^{i\omega}}\, d\omega, \tag{13}$$

and when $m > \Re(a)$, we may write this last contribution, by shrinking down the radius-1/2 contour segment to embrace the cut $(-1/2, 0]$, as

$$e_1 = \frac{e^{-z}}{\pi} \sin(\pi a) \int_0^{1/2} T^{-1-a} \left(1 + \frac{1}{T}\right)^{-m} e^{-zT}\, dT. \tag{14}$$

For the c_1 contribution above, we have used the function $\mathcal{H}_m$ defined by

$$\mathcal{H}_m(a, z, v) := v^a \left(1 + \frac{v}{2r}\right)^{-1-a} \left[F\left(\frac{v}{r}\right)\right]^m, \tag{15}$$

$$F(t) := \left(\frac{1 + t/2}{1 - t/2}\right) e^{-t}, \tag{16}$$

which for small t can be written $F(t) = 1 + t^3/12 + t^5/80 + \cdots = 1 + O(t^3)$.

In any event, e_1, like d_1, is subexponentially small relative to c_1 for large n and $(a,z) \in \mathbb{D}$. It is to be stressed that decomposition (10) holds for all $(a,z) \in \mathbb{C} \times \mathbb{C}, z \neq 0$, as long as $m := n + 1 > |z|/4$. We remind ourselves of (2) for $z = 0$. This means that such contour calculus applies to both oscillatory (Fejér) cases, where z is negative real, and subexponential (Perron) cases for the stated parameters.

2.2. Numerical success on the keyhole contour. The triumvirate c_1, d_1, e_1 of integrals is suitable for accurate Laguerre computations, which computations do show that *integrals d_1, e_1 tend to be subexponentially small* relative to c_1. This, of course, is our motive for so identifying the contour terms. An example, using the defining series (1) with $(a,z) := (-i, -1-i)$ is

$$L_8^{(i)}(1+i) \;=\; -\frac{137}{288} + \frac{53}{45} i \;\approx\; -0.475694444444444444 + 1.177777777777777777 i,$$

while the contributions from the C_1, D_1, E_1 segments of the contour of Figure 1 are

$$\begin{aligned} c_1 &\approx -0.44406762576110996056 + 0.81722282272705891241 i, \\ d_1 &\approx -0.03169598827425878852 + 0.36065591639721657587 i, \\ e_1 &\approx 0.00006916959092430464 - 0.00010096134649771051 i, \end{aligned}$$

with the sum $c_1 + d_1 + e_1$ giving $L_8^{(i)}(1+i)$ to the implied precision.

Remarkably, this "keyhole-contour" approach has the additional, unexpected feature that, for some parameter regions, the contour evaluation of $L_n^{(a)}(-z)$ is actually *faster* than direct summation of the defining series (1). A typical example is as follows: For the 13-digit evaluation

$$L_{50000}^{(25i/2)}(-30+3i) \approx (0.9275136583293 + 1.7406691595239 i) \times 10^{1056},$$

the defining series (1) was, in our trials, slower than the integral c_1.[4]

3. Effective bounds on the contour components.

3.1. Theta-calculus. We now introduce a notation useful in effective-error analysis. When two functions enjoy

$$|f(z)| \leq |g(z)|$$

over some relevant domain of z values, we shall say that

$$f(z) = \Theta(g(z)), \quad \text{or} \quad f(z) = \Theta : g(z)$$

on said domain. Thus, the Θ-notation is an effective replacement for big-O notation. The reason for allowing the notation $\Theta : \ldots$ is that a long formula g can run arbitrarily to the right of the colon.

Incidentally, there is a literature precedent for such a "theta-calculus." Namely, in computational number theory treatments such as [10], one encounters terms such as, say, $\theta x/\log^2 x$, with a stated constraint $\theta \in [-10, 10]$. This means the term is $O(x/\log^2 x)$ but with effective big-O constant bounded in magnitude by 10. In our nomenclature we would use $\Theta : 10x/\log^2 x$.

[4] Of course, series acceleration as in [5] would give the direct series a "leg up." Still, it is remarkable that contour integration is competitive. For the record, we compared *Mathematica*'s numerical integration to its own `LaguerreL[ ]` function.

3.2. Effective bound for e_1. First we address the integral e_1 as given in relation (13). We need some preliminary lemmas, starting with a collection of polynomial estimates to transcendental functions.

LEMMA 1. *The following inequalities hold for the respective conditions on ω:*

$$\log(5-4\cos\omega) \geq \frac{\log 9}{\pi^2}\omega^2, \quad \omega \in [-\pi,\pi];$$
$$\log(1+\omega) \geq \frac{4}{5}\omega, \quad \omega \in [0,1/2];$$
$$\arcsin\omega \leq \frac{\pi}{\sqrt{8}}\omega, \quad \omega \in [0,1/\sqrt{2}];$$
$$\cos\omega \leq 1-\frac{4}{\pi^2}\omega^2, \quad \omega \in [-\pi/2,\pi/2].$$

Proof. All four are straightforward calculus exercises. □

LEMMA 2. *Consider integrals of error-function class, specifically, for nonnegative real parameter μ,*

$$V_\mu(\alpha,\beta,\gamma) := \int_\gamma^\infty (x^2)^\mu e^{2\alpha x-\beta x^2}\, dx,$$

where each of α,β,γ is real, with $\alpha,\beta>0$. Then we always have the bound

$$V_0(\alpha,\beta,\gamma) = \Theta : \sqrt{\frac{\pi}{\beta}}\; e^{\frac{\alpha^2}{\beta}}.$$

If in addition $\gamma > 2\alpha/\beta$, we also have

$$V_\mu(\alpha,\beta,\gamma) = \Theta : \frac{1}{2}\; \frac{\Gamma(\mu+1/2)}{(\beta-2\alpha/\gamma)^{\mu+1/2}};$$

finally, if $\gamma > 4\alpha/\beta$, we have

$$V_\mu(\alpha,\beta,\gamma) = \Theta : \; 2^{\mu-1/2}\; \frac{\Gamma(\mu+1/2)}{\beta^{\mu+1/2}}.$$

Remark. We use the double-exponential $(x^2)^\mu$ because the first case of the theorem allows $\gamma<0$.

Proof. Completing the exponent's square gives the first bound easily, since

$$V_0 = e^{\alpha^2/\beta}\int_\gamma^\infty e^{-\beta(x-\alpha/\beta)^2}\, dx \;=\; \Theta : \sqrt{\pi/\beta}.$$

For the second bound, it is elementary that, for $x\geq\gamma$, one has $2\alpha x-\beta x^2 \leq x^2(2\alpha/\gamma-\beta)$ so that

$$V_\mu \leq \int_0^\infty x^{2\mu} e^{-(\beta-2\alpha/\gamma)x^2}\, dx,$$

and the resulting bound follows. The final bound follows immediately from the previous bound, because $\gamma > 4\alpha/\beta$ implies $\beta-2\alpha/\gamma > \beta/2$. □

THEOREM 1. *The contour contribution e_1 defined by* (13)*, under conditions $(a,z) \in \mathbb{D}$ and m sufficiently large in the explicit sense $m := n+1 > m_0 := |z|/4,\ \ m > m_1 := 5\left(|\Re(z)| + (|\Im(a)| + |\Im(z)|/2)^2\right)$, is bounded as*

$$e_1 = \Theta : e^{-z/2} 2^{\Re(a)+3} \frac{1}{\sqrt{m}}.$$

Moreover, under alternative conditions $(a,z) \in \mathbb{D}$ and $m > m_0, \quad m > m_2 := \Re(a), \quad m > m_3 := -\frac{5}{4}|z|\cos\theta$ we have a bound

$$e_1 = \Theta : \frac{e^{-z}}{\pi} \sin(\pi a) \frac{2^{\Re(a)-m}}{m - \Re(a)}.$$

Proof. The proof involves careful application of Lemmas 1 and 2 to the integral forms (13), (14); for brevity we omit these details. □

3.3. Effective bound for d_1. Again we need an opening lemma.

LEMMA 3. *Consider integrals of the incomplete-Bessel class, specifically*

$$W(\alpha,\beta) := \int_0^1 e^{-\alpha x - \frac{\beta}{x}}\ \frac{dx}{x},$$

where each of α, β is real, with $\beta > 0$. If $\beta > \alpha$, we have a bound

$$W = \Theta : \frac{1}{2} e^{-\beta-\alpha} \sqrt{\frac{\pi}{\beta}}.$$

Proof. We write

$$W(\alpha,\beta) = \int_0^1 e^{\beta x - \alpha x}\, e^{-\beta x - \frac{\beta}{x}}\ \frac{dx}{x} \le e^{\beta-\alpha} \int_0^1 e^{-\beta x - \frac{\beta}{x}}\ \frac{dx}{x}.$$

This last integral is a modified-Bessel term $K_0(2\beta)$ and has the required bound [2, Lemma 1]. □

THEOREM 2. *The contour contribution d_1 defined by* (12) *can be bounded, under conditions $(a,z) \in \mathbb{D}$ and $m > m_4 := 4|z|$, as*

$$d_1 = \Theta : \frac{e^{-z/2}}{\sqrt{\pi}} m^{|\Re(a)|/2 - 1/4} |z|^{-|\Re(a)|/2 - 1/4} \sin(\pi a) e^{-2\sqrt{m|z|}\cos^2\frac{\theta}{2}}.$$

Proof. Starting from (12) one applies Lemma 3; again for brevity we omit the details. □

3.4. Rigorous estimates on c_1. Having dispensed with d_1, e_1, we next show the integration limits ω_-, ω_+ on the c_1 contribution can be changed—with a subexponentially small error penalty—to $-\pi/2, \pi/2$, respectively.

LEMMA 4. *For any v on the unit circle $\{e^{i\phi} : \phi \in (-\pi, \pi]\}$, any complex a, and any real $R > (1+|a|)/2$ we have*

$$\left| \left(1 + \frac{v}{2R}\right)^{-1-a} \right| \le \frac{1}{1 - \frac{1+|a|}{2R}}.$$

Proof. Via the binomial theorem,

$$\left| \left(1 + \frac{v}{2R}\right)^{-1-a} \right| \le 1 + (1+|a|)\frac{1}{2R} + (1+|a|)^2 \left(\frac{1}{2R}\right)^2 + \cdots = \frac{1}{1 - \frac{1+|a|}{2R}}. \qquad \square$$

LEMMA 5. *Let v be on the unit circle as in Lemma 4, let $R > 1$ be real, and let m be a positive integer. For the function F appearing in* (16)*, we have the bound*

$$\left| F\left(\frac{v}{R}\right)^m \right| \leq e^{\frac{1}{6}\frac{m}{R^3}}.$$

Proof. From the definition (16) we have

$$F(v/R)^m = e^{\frac{1}{12}\frac{m}{R^3}(1+(3/5)/(2R)^2+(3/7)/(2R)^4+\cdots)}.$$

For $R := 1$, the infinite sum is no larger than 1.18 and is monotonic decreasing in R. □

LEMMA 6. *Let v be on the unit circle as in Lemma 4 and define for nonnegative integer m*

$$K := \left(1+\frac{v}{2R}\right)^{-1-a} F\left(\frac{v}{R}\right)^m.$$

For the assignments $(a, z) \in \mathbb{C} \times \mathbb{C}, z \neq 0, R := \sqrt{m/|z|} > 1, m > m_5 := |z|(1 + |a| + |z|/2)^2$, we have

$$K = \Theta : 2.$$

Proof. From Lemmas 4 and 5 we have

$$|K| \leq \frac{1}{1-\frac{1+|a|}{2(1+|a|+|z|/2)}} e^{\frac{1}{6}\frac{|z|}{1+|a|+|z|/2}}.$$

The fact that the right-hand side is $\Theta(2)$ follows from the observation that the function

$$\frac{1}{1-\frac{Q}{2Q+x}} e^{\frac{1}{3}\frac{x}{2Q+x}}$$

for $Q \geq 1, x \in [0, \infty)$ is itself $\Theta(2)$. This in turn follows easily on substituting $y := x/(2Q + x)$. Now the function to be bounded is $g(y) := (2/(1 + y))e^{y/3}$ on $y \in [0, 1/2]$—as differentiation settles. □

These lemmas in turn allow us to contract the range on the c_1 contour integral.

THEOREM 3. *Decompose c_1 as defined by* (11) *into two terms,*

$$c_1 := c_0 + c_2,$$

with c_0 involving the integral's range contracted to $[-\pi/2, \pi/2]$, namely,

$$c_0 := \frac{1}{2\pi} r^{-a} e^{-z/2} \int_{-\pi/2}^{\pi/2} \mathcal{H}_m(a, z, e^{-i\omega}) e^{2\sqrt{mz}\cos\omega}\, d\omega.$$

Then, under conditions $(a, z) \in \mathbb{D}$ and $m > m_5 := |z|(1 + |a| + |z|/2)^2$, we have a bound

$$c_2 = \Theta : r^{-a} e^{-z/2} e^{\frac{3}{2}\pi|\Im(a)|}.$$

Proof. It is evident that c_2 is obtained from the definition (11) but with the integral replaced via

$$\int_{\omega_-}^{\omega_+} \to \left\{ \int_{\omega_-}^{-\pi/2} + \int_{\pi/2}^{\omega_+} \right\}.$$

Over these domains of integration, we have $e^{2\sqrt{mz}\cos\omega} = \Theta(1)$. Thus, Lemmas 4, 5, and 6 show

$$c_2 = \Theta : \frac{1}{2\pi} r^{-a} e^{-z/2} \left\{ \int_{\omega_-}^{-\pi/2} + \int_{\pi/2}^{\omega_+} \right\} 2 \left| e^{-i\omega a} \right| \, d\omega.$$

As the total support of the integrals cannot exceed π, the desired c_2 bound follows. □

3.5. Summary of the contour decomposition for $L_n^{(-a)}(-z)$. The above manipulations lead to the main result of the present section, namely, a formula that decomposes the Laguerre evaluation, as in the following.

THEOREM 4 (contour decomposition). *Let $(a, z) \in \mathbb{C} \times \mathbb{C}$ be an arbitrary parameter pair, with $z \neq 0$ (with $z = 0$ cases resolved exactly by (2)). If $m := n + 1 = m_0 > |z|/4$, then the contour decomposition*

$$L_n^{(-a)}(-z) = c_0 + \mathcal{E}$$

holds, with c_0, c_2 as defined in Theorem 3 and $\mathcal{E} := c_2 + d_1 + e_1$. If an appropriate set of conditions on m across Theorems 1, 2, and 3 *holds, then we can write*

$$L_n^{(-a)}(-z) = c_0 + S_n(a, z)\mathcal{E}_1,$$

where $\mathcal{E}_1$ is subexponentially small, in the sense that, for any fixed positive ϵ, the large-n behavior is

$$\mathcal{E}_1 = O\left(e^{-(2-\epsilon)\sqrt{m|z|}\cos\frac{\theta}{2}} \right).$$

Moreover, an effective big-O constant is available (the proof exhibits explicit forms).

Proof. The contour calculus holds for all complex pairs (a, z), with $z \neq 0, R := \sqrt{m/|z|} > 1/2$, which assures that the point 1 is contained in the contour. It remains to analyze $\mathcal{E}_1$. Consider, then, an appropriate union of conditions from the cited theorems, say, $(a, z) \in \mathbb{D}$ and $m > \max(m_0, m_1, m_2, m_3, m_4, m_5)$, in which case we have, from Theorems 1, 2, and 3 (in that respective order of Θ terms):

$$\mathcal{E}_1 = \Theta : \frac{e^{-z/2}}{\sqrt{\pi}} \sin(\pi a) \frac{2^{1+\Re(a)-m}}{m - \Re(a)} z^{1/4-a/2} m^{1/4+a/2} e^{-2\sqrt{m|z|}\cos\frac{\theta}{2}}$$

$$+\Theta : 2 \left(\frac{m}{|z|} \right)^{(\Re(a)+|\Re(a)|)/2} e^{\frac{3}{2}\pi|\Im(a)|} e^{-2\sqrt{m|z|}\left(\cos^2\frac{\theta}{2} + \cos\frac{\theta}{2}\right)}$$

$$+\Theta : 2\sqrt{\pi}\, |mz|^{1/4} e^{\frac{3}{2}\pi|\Im(a)|} e^{-2\sqrt{m|z|}\cos\frac{\theta}{2}}.$$

This explicit bounding of the error term $\mathcal{E}_1$ proves the big-O statement of the theorem, while for any choice of ε, an effective big-O constant can be read off at will. □

4. Effective expansion for the $\mathcal{H}$-kernel. Theorem 4 shows the main contribution to $L_n^{(a)}(-z)$, for $(a, z) \in \mathbb{D}$, with $r := \sqrt{m/z}, |r| > 1/2$, is

$$c_0 := \frac{1}{2\pi} r^{-a} e^{-z/2} \int_{-\pi/2}^{\pi/2} \mathcal{H}_m\left(a, z, e^{-i\omega}\right) e^{2\sqrt{mz}\cos\omega} \, d\omega, \tag{17}$$

with the integration kernel $\mathcal{H}_m$ defined, see (15) and (16), as

$$\mathcal{H}_m(a,z,v) := v^a \left(1+\frac{v}{2r}\right)^{-1-a} \left(\frac{1+\frac{v}{2r}}{1-\frac{v}{2r}}\right)^m e^{-mv/r}, \tag{18}$$

where in the integral we assign $v := e^{-i\omega}$. We need to obtain the growth properties of $\mathcal{H}_m$.

4.1. Exponential form for $\mathcal{H}_m$.

LEMMA 7. *For $|v|=1$ and $m > |z|/4$, the $\mathcal{H}$-kernel can be cast in the exponential form*

$$\mathcal{H}_m := v^a \exp\left\{\sum_{k\geq 1} \frac{a_k}{k}\frac{1}{m^{k/2}}\right\}, \tag{19}$$

where

$$a_k := (1+a)(-1)^k \left(\frac{v\sqrt{z}}{2}\right)^k + \left(1-(-1)^k\right)\frac{k}{k+2}\left(\frac{v\sqrt{z}}{2}\right)^{k+2}. \tag{20}$$

Moreover, we have the general coefficient bound

$$|a_k| \leq \left(\frac{\sqrt{z}}{2}\right)^k (1+|a|+|z|/2). \tag{21}$$

Proof. (18) can be recast, with $\rho := v/(2r)$, as

$$\mathcal{H}_m := v^a \exp\left\{-(1+a)\log(1+\rho) - 2m\rho + m\left(\log(1+\rho)-\log(1-\rho)\right)\right\}. \tag{22}$$

Since $|r|>1/2$ and $|v|=1$, the logarithmic series converge absolutely and we have

$$\mathcal{H}_m := v^a \exp\left\{\sum_{k\geq 1}\left[\frac{(1+a)(-1)^k}{k}\left(\frac{v\sqrt{z}}{2}\right)^k g^k + \frac{2}{2k+1}\left(\frac{v\sqrt{z}}{2}\right)^{2k+1} g^{2k-2}\right]\right\}, \tag{23}$$

where $g := 1/\sqrt{m}$, and the precise form (20) for the a_k follows immediately. The given bound on $|a_k|$ is also immediate from (20). ∎

4.2. Exponentiation of series. Though Lemma 7 is progress, we still need to exponentiate a series, in the sense that we want to know, for the following expansion, given the sequence (a_k),

$$\exp\left\{\sum_{k\geq 1}\frac{a_k}{k}x^k\right\} =: \sum_{h\geq 0} A_h x^h,$$

how the A_h depend on the a_k. The combinatorial answer is

$$A_h = \sum_{j=0}^{h}\frac{1}{j!}\,G_h(j;\vec{a}), \qquad G_h(j;\vec{a}) := \sum_{h_1+\cdots+h_j=h}\frac{a_{h_1}\cdots a_{h_j}}{h_1\cdots h_j},$$

with the understanding that $G_h(0,\vec{a}) := \delta_{0h}$ and that such combinatorial sums involve positive integer indices h_i. One useful result is the following.

LEMMA 8. *If all coefficients $a_k = 1$, then, for $j > 0$,*

$$G_h(j;\vec{1}) \;=\; \sum_{h_1+\cdots+h_j=h} \frac{1}{h_1\cdots h_j} \;=\; \Theta : \frac{1}{h}\,(2H_{h-j+1})^{j-1} \;=\; \Theta : \frac{1}{h}\,(2\gamma + 2\log h)^{j-1}\,.$$

Here, $H_p := 1 + 1/2 + \cdots + 1/p$ is the pth harmonic number, $H_0 := 0$, and γ is the Euler constant.

Remark. It turns out that G here enjoys a closed form of sorts, namely,

$$G_h(j;\vec{1}) = \frac{j!}{h!}(-1)^{h-j}\mathcal{S}_h^{(j)},$$

where $\mathcal{S}$ denotes the Stirling number of the first kind, normalized via $x(x-1)\cdots(x-h+1) =: \sum_{j=0}^{h} \mathcal{S}_h^{(j)} x^j$. So one byproduct of our lemma is a rigorous bound on the growth of Stirling numbers; see [1, 24.1.3,III] and [37] for research on Stirling asymptotics.

Proof. The first Θ-estimate arises by induction. For notational convenience we omit the vector $\vec{1}$ and just use the symbol $G_h(j)$. Note that $G_N(1) = 1/N$ and $G_N(2) = \frac{2}{N}H_{N-1}$. Generally we have

$$G_N(J) = \sum_{j=1}^{N-J+1} \frac{1}{j} G_{N-j}(J-1).$$

Now, assume by induction that $G_h(j) = \Theta : \frac{1}{h}(2H_{h-j+1})^{j-1}$ holds for all $j < J$. Then

$$G(N,J) \;\le\; \sum_{j=1}^{N-j+1} \frac{2^{J-2}H_{N-J-j+2}^{J-2}}{j(N-J)} \;\le\; \frac{2^{J-2}}{N} H_{N-J+1}^{J-2} \sum_{j=1}^{N-J+1} \left(\frac{1}{j} + \frac{1}{N-j}\right).$$

Now this last parenthetical term is $H_{N-J+1} + H_{N-1} - H_J$ which, because $H_a - H_b \le H_{a-b}$ for any positive integer indices $a > b$, is bounded above by $2H_{N-J+1}$, which proves the first Θ-bound of the theorem. For the second Θ-bound it suffices, since H_j is increasing, to show that $H_{n-1} > \gamma + \log(n)$, which is an elementary calculus problem. □

Though we do not use Lemma 8 directly in what follows, it is useful in proving convergence for various sums $\sum A_h x^h$ and may well matter in future research along our lines.

LEMMA 9. *Let $y \ge 1$ and $x \in (-1,1)$ be real. Then in the expansion*

$$\exp\left\{ y \sum_{k\ge 1} \frac{x^k}{k} \right\} =: \sum_{h\ge 0} Y_h x^h$$

the coefficients Y_h enjoy the bound $Y_h = \Theta : y^h$.

Proof. First, $Y_0 = 1 \le 1$. The left-hand side is $\exp(-y\log(1-x)) = (1-x)^{-y}$ whose binomial expansion has hth coefficient ($h \ge 1$) equal to

$$\frac{y(y+1)\cdots(y+h-1)}{h!} \le y^h\, \frac{1}{1}\frac{1+1/y}{2}\cdots\frac{1+(h-1)/y}{h} \le y^h. \qquad \square$$

4.3. Effective expansions of exponentiated series. We are now in a position to derive an effective expansion for an exponentiated series, starting with the following.

LEMMA 10. *Assume complex vector $\vec{b}$ of defining coefficients b_k bounded as $|b_k| \leq cd^k$ for positive real c, d with $c \geq 1$. Assume also $|x| < 1/(2cd)$. Then, for any order $N \geq 0$, we have an effective expansion*

$$\exp\left\{\sum_{k\geq 1} \frac{b_k}{k} x^k\right\} = \sum_{h=0}^{N-1} B_h x^h + \Theta : 2c^N d^N x^N,$$

where

$$B_h = \sum_{j=0}^{h} \frac{G_h(j;\vec{b})}{j!}$$

are the usual coefficients of the full formal exponentiation.

Proof. Denoting $f(x) := \exp\left\{\sum_{k\geq 1} \frac{b_k}{k} x^k\right\}$, we have

$$f(x) = \sum_{h=0}^{N-1} B_h x^h + T_N,$$

with remainder $T_N = \sum_{h\geq N} B_h x^h$, having B_h coefficients given by a G-sum as in Lemma 10. Now,

$$|B_h| \leq \sum_{j=0}^{h} \frac{|G_h(j;\vec{b})|}{j!} \leq \sum_{j=0}^{h} \frac{|G_h(j;\vec{f})|}{j!},$$

where $\vec{f} = (cd^k : k \geq 0)$. By Lemma 9 we know that $|B_h| \leq (cd)^h$. Therefore

$$|T_N| \leq \sum_{h\geq N} (cd)^h x^h = \frac{c^N d^N x^N}{1 - cdx} = \Theta : 2(cdx)^N. \qquad \square$$

Finally we arrive at a general expansion—with effective remainder—for the $\mathcal{H}$-kernel.

THEOREM 5 (effective expansion for $\mathcal{H}_m$). *For general complex $(a, z) \in \mathbb{C} \times \mathbb{C}$, assume that $m > m_5 := |z|(1 + |a| + |z|/2)^2$ and $|v| = 1$. Then, for any expansion order $N \geq 0$, we have*

$$\mathcal{H}_m(a, z, v) = v^a \left(\sum_{h=0}^{N-1} \frac{A_h}{m^{h/2}} \; + \; \Theta : 2\left(\frac{m_5}{4m}\right)^{N/2} \right),$$

where, on the basis of the defining coefficients a_k given in (20),

$$A_h := \sum_{j=0}^{h} \frac{G_h(j;\vec{a})}{j!}, \qquad G_h(j;\vec{a}) := \sum_{h_1+\cdots+h_j=h} \frac{a_{h_1} \cdots a_{h_j}}{h_1 \cdots h_j}.$$

Proof. The result follows immediately from Lemma 10, on assigning $\vec{b} = \vec{a}$, with $x := 1/\sqrt{m}$, $c := 1 + |a| + |z|/2$, $d := (1/2)\sqrt{|z|}$. $\square$

Note that Theorem 5 in the instance $N = 0$ implies our previous Lemma 6. It is interesting and suggestive that the threshold $m_5 := |z|(1 + |a| + |z|/2)^2$ appears in these results rather naturally.

4.4. Effective integral form for c_0. To obtain a useful form for c_0, the dominant component of $L_n^{(-a)}(-z)$, we use Theorem 5 to obtain the following.

THEOREM 6. *For $(a,z) \in \mathbb{D}$ and $m > m_5$, the dominant component of Theorems 3 and 4, namely,*

$$c_0 := \frac{1}{2\pi} r^{-a} e^{-z/2} \int_{-\pi/2}^{\pi/2} \mathcal{H}_m(a,z,e^{-i\omega}) e^{2\sqrt{mz}\cos\omega}\, d\omega,$$

can be given an effective form for any order $N \geq 0$, as

$$c_0 = \frac{1}{2\pi} r^{-a} e^{-z/2} \sum_{h=0}^{N-1} \frac{1}{m^{h/2}} \int_{-\pi/2}^{\pi/2} e^{-i\omega a} A_h e^{2\sqrt{mz}\cos\omega}\, d\omega + \; S_n(a,z)\; \mathcal{E}_{2,N},$$

where the error term is bounded as

$$\mathcal{E}_{2,N} = \Theta : \frac{\pi}{\sqrt{2}} \left(\frac{m_5}{4m}\right)^{N/2} \exp\left(\frac{\pi^2 \Im(a)^2 \sec\frac{\theta}{2}}{32\sqrt{m|z|}}\right) \sec^{1/2}\frac{\theta}{2}$$

and the A_h are to be calculated as the first N coefficients of

$$\sum_{h=0}^{\infty} A_h x^h := \exp\left\{ \sum_{k\geq 1} \frac{a_k}{k} x^k \right\},$$

via (20)*, with $v := e^{-i\omega}$.*

Proof. Inserting the effective $\mathcal{H}$-kernel expansion from Theorem 5 directly into the c_0 integral gives the indicated sum over $h \in [0, N-1]$ plus an error term

$$\Theta : \frac{1}{2\pi} r^{-a} e^{-z/2} 2 \left(\frac{m_5}{4m}\right)^{N/2} \int_{-\pi/2}^{\pi/2} e^{\omega\Im(a)} e^{-2\sqrt{m|z|}\cos(\theta/2)\cos\omega}\, d\omega.$$

Using Lemma 1 on $\cos\omega$ and the V_0-part of Lemma 2, we obtain the $\mathcal{E}_{2,N}$ bound of the theorem. ∎

With Theorem 6 we have come far enough to see that a Laguerre evaluation can be obtained—up to a subexponentially small relative error—via the A_h terms in said theorem. To this end, an inspection of the defining relations reveals that, in general, we can decompose an A_h coefficient in terms of powers of $v := e^{-i\omega}$, namely, we define $\alpha_{h,\mu}$ terms via

$$A_h =: \sum_{u=0}^{h} \alpha_{h,u}(a,z) v^{h+2u}. \tag{24}$$

For example,

$$\alpha_{00} = 1, \quad \alpha_{10} = -\frac{1+a}{2} z^{1/2}, \quad \alpha_{11} = \frac{z^{3/2}}{12}, \quad \alpha_{31} = \frac{1}{480}\left(5a^2 + 15a + 16\right) z^{5/2}.$$

The point being, we now have special formulae for the dominant contribution c_0, namely,

$$c_0 = \frac{1}{2\pi} r^{-a} e^{-z/2} \sum_{h=0}^{N-1} \frac{1}{m^{h/2}} \sum_{u=0}^{h} \alpha_{h,u}(a,z)\; \mathcal{I}(2\sqrt{mz}, a+h+2u) \; + \quad S_n(a,z)\; \mathcal{E}_{2,N}, \tag{25}$$

where the integral

$$\mathcal{I}(p,q) := \int_{-\pi/2}^{\pi/2} e^{-iq\omega} e^{p\cos\omega}\, d\omega \tag{26}$$

thus emerges as a fundamental entity for the research at hand.

5. $\mathcal{I}$-integrals and an "exp-arc" method. Having reduced the problem of subexponential Laguerre growth to a study of the $\mathcal{I}$-integrals (26), we next develop a method that is effective for both their numerical and theoretical estimation. This method amounts to the avoidance of stationary-phase techniques, employing instead various forms of *exponential-arcsine* (exp-arc) series, as we see shortly.

Let us first define, for any complex pair (p,q) and $\alpha,\beta \in (-\pi,\pi)$,

$$I(p,q,\alpha,\beta) := \int_{\alpha}^{\beta} e^{-iq\omega} e^{p\cos\omega}\, d\omega \tag{27}$$

so that our special case (26) is simply $\mathcal{I}(p,q) := I(p,q,-\pi/2,\pi/2)$.

Importantly, one may write the Bessel functions J_n of integer order n in terms of $\mathcal{I}$-integrals:

$$J_n(z) = \frac{1}{2\pi}\left(e^{-i\pi n/2}\mathcal{I}(iz,n) + e^{i\pi n/2}\mathcal{I}(-iz,n)\right), \tag{28}$$

and the modified Bessel function, again of integer order n:

$$I_n(z) = \frac{1}{2\pi}\left(\mathcal{I}(z,n) + (-1)^n\mathcal{I}(-z,n)\right), \tag{29}$$

about which representations we shall have more to say in a later section.[5]

5.1. Essentials of the exp-arc method. Now we investigate what we call exp-arc series. First, for any complex τ and $x \in [-1,1]$, one has a remarkable, absolutely convergent expansion (see [6]):

$$e^{\tau \arcsin x} = \sum_{k=0}^{\infty} r_k(\tau)\,\frac{x^k}{k!}, \tag{30}$$

where the coefficients depend on the parity of the index

$$r_{2m+1}(\tau) := \tau \prod_{j=1}^{m}\left(\tau^2 + (2j-1)^2\right), \qquad r_{2m}(\tau) := \prod_{j=1}^{m}\left(\tau^2 + (2j-2)^2\right).$$

By differentiating with respect to x, we obtain

$$\frac{e^{\tau \arcsin x}}{\sqrt{1-x^2}} = \frac{1}{\tau}\sum_{k=0}^{\infty} r_{k+1}(\tau)\,\frac{x^k}{k!},$$

valid for $x \in (-1,1)$. In particular, we have the important expansion (using a function G, in passing)

$$G(\tau,x) := \frac{\cosh(\tau \arcsin x)}{\sqrt{1-x^2}} = \sum_{k=0}^{\infty} g_k(\tau)\frac{x^{2k}}{(2k)!}, \tag{31}$$

[5]The $\mathcal{I}$-integral can also be written in compact terms of an *Anger function* **J** and *Weber function* **E**; see [1].

where

$$g_k(\tau) := \prod_{j=1}^{k} \left((2j-1)^2 + \tau^2\right).$$

These exp-arc expansions may be applied to the $I(p,q,\alpha,\beta)$ integrals as follows.[6] From (27) and its subsequent manipulations we have

$$I(p,q,\alpha,\beta) = e^p \int_\alpha^\beta e^{-iq\omega} e^{-2p\sin^2(\omega/2)}\, d\omega = 2e^p \int_{\sin\frac{\alpha}{2}}^{\sin\frac{\beta}{2}} \frac{e^{-2iq\arcsin x}}{\sqrt{1-x^2}} e^{-2px^2}\, dx$$

$$= i\frac{e^p}{q} \sum_{k=0}^{\infty} \frac{r_{k+1}(-2iq)}{k!} \int_{\sin\frac{\alpha}{2}}^{\sin\frac{\beta}{2}} x^k e^{-2px^2}\, dx. \tag{32}$$

Even though we eventually consider asymptotic expansions, the convergence of such an I-series is the rule.

LEMMA 11. *For $\alpha, \beta \in (-\pi, \pi)$ and any complex pair (p,q), the series* (32) *converges absolutely.*

Proof. Define $\delta := \max(|\sin(\alpha/2)|, |\sin(\beta)/2|)$ so that the integral in (32) is bounded in magnitude by $2\exp(2\Re(p))\delta^{k+1}/(k+1)$. But this means the kth summand in (32) is, up to a k-independent multiplier, bounded in magnitude by $r_{k+1}(2|q|)\delta^{k+1}/(k+1)!$. This summand is of the same form as that in the defining series (30), so absolute convergence is assured. ∎

Likewise, from (31) and (32) we have an absolutely convergent expansion for the special-case $\mathcal{I}$ integral:

$$\mathcal{I}(p,q) = 4e^p \int_0^{1/\sqrt{2}} G(-2iq,x) e^{-2px^2} = 4e^p \sum_{k=0}^{\infty} \frac{g_k(-2iq)}{(2k)!} B_k(p), \tag{33}$$

where B_k is an error–function-class integral

$$B_k(p) := \int_0^{1/\sqrt{2}} x^{2k} e^{-2px^2}\, dx = \frac{1}{2}\frac{1}{(2p)^{k+1/2}} \left\{\Gamma(k+1/2) - \Gamma(k+1/2, p)\right\}. \tag{34}$$

A computationally key recurrence relation accrues: with $B_0(p) = \sqrt{\frac{\pi}{8p}}\,\mathrm{erf}\left(\sqrt{p}\right)$ and $k > 0$, we have

$$B_k(p) := \frac{2k-1}{4p} B_{k-1} - \frac{2^{-k-3/2}}{p} e^{-p}. \tag{35}$$

5.2. Effective expansion for the cosh-arc *G*-series. Our asymptotic analysis of representation (33) begins with a lemma that reveals how the G function (31) has an attractive self-similarity property. Namely, the function appears naturally in modified form within its own error terms.

LEMMA 12. *For any complex τ, the G function* (31) *can be given an effective expansion to any integer order $N \geq 0$, as*

$$G(\tau,x) := \frac{\cosh(\tau \arcsin x)}{\sqrt{1-x^2}} = \sum_{k=0}^{N-1} g_k(\tau)\frac{x^{2k}}{(2k)!} + g_N(\tau)\frac{x^{2N}}{(2N)!}\left(1 + T_N(\tau,x)\right), \tag{36}$$

[6] An equivalent approach to expanding I uses a series for $e^{iq\omega}\sec(\omega/2)$ in terms of $x := \sin(\omega/2)$, as in [20, p. 276].

with the error term T_N conditionally bounded over real $x \in [0, 1/\sqrt{2}]$ in the form

$$\begin{aligned} T_N &= \Theta : 1 && \text{if } N \geq 2x^2|\tau|^2 - 1; \\ &= \Theta : (\sqrt{2}\ x^2 + x|\tau|)e^{|\tau| \arcsin x} && \text{otherwise.} \end{aligned}$$

Proof. The error term is, by the definition of the $g_k(\tau)$, given by the absolutely convergent sum

$$T_N = h_N + h_N h_{N+1} + h_N h_{N+1} h_{N+2} + \cdots,$$

where

$$h_k := \frac{(2k+1)^2 + \tau^2}{(2k+1)(2k+2)} x^2.$$

When $|\tau|^2 x^2 \leq (N+1)/2$ and since $x^2 \leq 1/2$, it is immediate that, for $k \geq N$, we have a bound:

$$|h_k| \leq \frac{(4k^2 + 4k + 1)/2 + (N+1)/2}{4k^2 + 5k + 2} \leq 1/2,$$

whence $T_N = \Theta : 1/2 + 1/4 + 1/8 + \ldots$, settling the first conditional bound of the theorem. In any case, i.e., any complex τ and any $x \in [0, 1/\sqrt{2}]$, we have for $j \geq 0$:

$$\begin{aligned} h_N h_{N+1} \cdots h_{N+j} &= x^{2j} \prod_{k=0}^{j} \frac{(2N+2k+1)^2 \left((2k+1)^2 + \tau^2 \frac{(2k+1)^2}{(2N+2k+1)^2} \right)}{(2N+2k+1)(2N+2k+2)(2k+1)^2} \\ &= \Theta : \frac{g_{j+1}(|\tau|)}{(2j+1)!!^2}. \end{aligned}$$

However, it is elementary that $(2j+1)!!^2 \geq (2j+1)!$ by simple factor-tallying, so

$$|T_N| \leq \frac{\left(1^2 + |\tau|^2\right) \cdot 2}{2!} x^2 + \frac{\left(1^2 + |\tau|^2\right)\left(3^2 + |\tau|^2\right) \cdot 4}{4!} x^4 + \ldots = x \frac{\partial}{\partial x} G(|\tau|, x),$$

where we have noticed that the right-hand series here is itself a differentiated "cosh-arc" series. Thus

$$T_N = \Theta : x \frac{\partial}{\partial x} \frac{\cosh(|\tau| \arcsin x)}{\sqrt{1-x^2}} = \Theta : \frac{x^2}{2\left(1-x^2\right)^{3/2}} \left(e^u + e^{-u}\right) + \frac{x|\tau|}{2\left(1-x^2\right)} \left(e^u - e^{-u}\right),$$

where $u := |\tau| \arcsin x$. Now by excluding $2x^2|\tau|^2 \leq N+1$ for the second conditional bound of the lemma, we have $|\tau| \geq 1$, whence the e^{-u} terms can be ignored over $x \in [0, 1/\sqrt{2}]$, and the second conditional bound follows. ∎

5.3. Effective expansion of the $\mathcal{I}$-integral. We first invoke a classical lemma that bounds the incomplete gamma function for certain parameters.

LEMMA 13. *For integer $M \geq 0$, $\Re(z) \geq 0, z \neq 0$ and $|z| \geq 2M - 1$, we have*

$$\Gamma(M + 1/2, z) = \Theta : 2z^{M-1/2} e^{-z}.$$

Proof. Proofs of results such as these on incomplete-gamma bounds appear in various texts on special functions, e.g., [27, section 2.2, p. 110]. ∎

The results of the present section may now be applied to a general, effective expansion of the $\mathcal{I}$-integral whenever $\Re(p)$ is sufficiently positive.

THEOREM 7 (effective $\mathcal{I}$ expansion). *For the integral*

$$\mathcal{I}(p,q) := I(p,q,-\pi/2,\pi/2) = \int_{-\pi/2}^{\pi/2} e^{-iq\omega} e^{p\cos\omega}\, d\omega,$$

assume an integer expansion order $N \geq 1$. Assume $\phi := \arg(p) \in (-\pi/2, \pi/2)$ and conditions

$$\Re(p) \geq 2N+1, \quad \Re(p) \geq 2\pi|q|^2.$$

Then we have an effective expansion

$$\mathcal{I}(p,q) = \sqrt{\frac{2\pi}{p}}\; e^p \left\{ \sum_{k=0}^{N-1} \frac{g_k(-2iq)}{k!\; 8^k} \frac{1}{p^k} + \Theta : \sqrt{\frac{8}{\pi p}}\; e^{-p} \cosh\left(\frac{\pi}{2}|q|\right) + \frac{g_N(-2iq)}{N!\; 8^N} \frac{1}{p^N} \left(1 + \Theta : u_N \sec^{N+1/2}\phi\right) \right\}, \tag{37}$$

where the g_k are as in the cosh-arc expansion (31)*, and we may take*

$$u_N := 1 + 2^N + \frac{2^{3N+1}}{\pi \binom{2N}{N}}\ ; \tag{38}$$

however, on the extra condition $N \geq 4|q|^2 - 1$, taking $u_N := 1$ suffices.

Proof. The insertion of the series of Lemma 12 into representation (33) results in

$$\mathcal{I}(p,q) = 4e^p \left\{ \sum_{k=0}^{N} \frac{g_k(-2iq)}{(2k)!} B_k(p) + \frac{g_N(-2iq)}{(2N)!} \int_0^{1/\sqrt{2}} x^{2N} T_N(-2iq,x) e^{-2px^2}\, dx \right\},$$

where the $B_k(p)$ are given by (34). The sum over $k \in [0,N]$ here is thus

$$\frac{1}{2} \sum_{k=0}^{N} \frac{g_k(-2iq)}{(2k)!} \frac{\Gamma(k+1/2)}{(2p)^{k+1/2}} + \Theta : e^{-p} \sum_{k=0}^{N} \frac{|g_k(-2iq)|}{(2k)!} \frac{1}{2^{k+1/2} p},$$

where the Θ-term here follows from Lemma 13 on our condition $\Re(p) \geq 2N+1$. But this very Θ-term is bounded above by

$$\frac{e^{-p}}{p\sqrt{2}} \sum_{k=0}^{\infty} \frac{g_k(2|q|)}{(2k)!} \frac{1}{2^k} = \frac{e^{-p}}{p} \cosh\left(2|q| \arcsin\left(1/\sqrt{2}\right)\right),$$

so we have settled the summation and the $\cosh(\pi|q|/2)$ term in (37). Now consider the integral term

$$I_0 := \int_0^{1/\sqrt{2}} x^{2N} T_N(-2iq,x) e^{-2px^2}\, dx.$$

Define $\gamma := |q|^{-1}\sqrt{(N+1)/8}$. If $\gamma \geq 1/\sqrt{2}$, then by Lemma 12, we know that $T_N = \Theta(1)$, and our theorem follows in the $u_N := 1$ case. Otherwise, $\gamma < 1/\sqrt{2}$, and

we bound I_0 using two integrals

$$|I_0| \le \int_0^{1/\sqrt{2}} x^{2N} e^{-2\Re(p)x^2}\, dx + \int_\gamma^{1/\sqrt{2}} \left(\sqrt{2}\, x^{2N+2} + 2|q|\, x^{2N+1}\right) e^{2|q|\arcsin x - 2\Re(p)x^2}\, dx.$$

From Lemma 1 we know that the exponent here can be taken to be $\pi|q|x/\sqrt{2} - 2\Re(p)x^2$. For the assignments $\alpha := \pi|q|/\sqrt{8},\ \beta := 2\Re(p)$, we have $\beta > 4\alpha/\gamma$ so that by Lemma 2 there is a bound

$$\begin{aligned} I_0 &\le \frac{1}{2}\frac{\Gamma(N+1/2)}{(2\Re(p))^{N+1/2}} + \sqrt{2}\, V_{N+1}(\alpha,\beta,\gamma) + 2|q|\, V_{N+1/2}(\alpha,\beta,\gamma) \\ &\le \frac{1}{2}\frac{\Gamma(N+1/2)}{(2\Re(p))^{N+1/2}} + 2^N \frac{\Gamma(N+3/2)}{(2\Re(p))^{N+3/2}} + 2^{N+1}|q|\frac{\Gamma(N+1)}{(2\Re(p))^{N+1}}. \end{aligned}$$

Using $|q|^2 \le \Re(p)/(2\pi)$, $\Re(p) \ge 2N+1$ with $\Re(p) = |p|\cos\phi$ yields the N-dependent u_N form of the theorem. □

6. Effective asymptotics for $L_n^{(-a)}(-z)$. At last we may provide explicit terms for Laguerre expansions in the subexponential-growth regime, which regime turns out to be precisely characterized by the parameter-pair requirement: $(a,z) \in \mathbb{D}$.

First, for convenience, we recapitulate the thresholds for sufficiently large $m := n+1$ from our previous theorems:

$$m_0 := |z|/4, \quad m_1 := 5\left(|\Re(z)| + (|\Im(a)| + |\Im(z)|/2)^2\right),$$
$$m_2 := \Re(a), \quad m_3 := -(5/4)|z|\cos\theta, \quad m_4 = 4|z|, \quad m_5 := |z|(1+|a|+|z|/2)^2,$$
$$m_6 := (2N+1)^2\, \frac{\sec^2(\theta/2)}{|z|}, \quad m_7 := 4\pi^2\, \frac{\sec^2(\theta/2)}{|z|}(|a|+3N-3)^4.$$

Here, $\theta := \arg(z)$ as before, while m_6, m_7 involve an asymptotic expansion order $N \ge 1$. We are aware that the m_i bounds here are interdependent, and some are masked by others (e.g., m_0 is masked by m_4). The important quantity in our central result (Theorem 8, below) is simply $\max_{i\in[0,7]} m_i$.

Before delving into the central result, let us remind ourselves of previous nomenclature:

$$g_k(\tau) := \prod_{j=1}^{k} \left((2j-1)^2 + \tau^2\right), \tag{39}$$

$$a_k := (1+a)(-1)^k \left(\frac{v\sqrt{z}}{2}\right)^k + \left(1-(-1)^k\right)\frac{k}{k+2}\left(\frac{v\sqrt{z}}{2}\right)^{k+2}, \tag{40}$$

$$\sum_{h=0}^{\infty} A_h x^h := \exp\left\{\sum_{k\ge1} \frac{a_k}{k} x^k\right\}, \tag{41}$$

$$A_h := \sum_{u=0}^{h} \alpha_{h,u}(a,z) v^{h+2u}. \tag{42}$$

Note that the $\alpha_{h,u}$ coefficients are thus implicitly defined in terms of the original a_k functions.

6.1. The general subexponential expansion. Using Theorem 7 with $p := 2\sqrt{mz}$ and inserting this into formula (25), we arrive at our desired effective Laguerre expansion.

THEOREM 8 (effective Laguerre expansion). *Assume $(a, z) \in \mathbb{D}$. For asymptotic expansion order $N \geq 1$ and $m := n + 1$ sufficiently large in the sense $m > \max_{i\in[0,7]} m_i$, we have the expansion*

$$L_n^{(-a)}(-z) = \frac{1}{2\sqrt{\pi}} \frac{e^{-z/2} e^{2\sqrt{mz}}}{z^{1/4-a/2}\ m^{1/4+a/2}} \left\{ \sum_{j=0}^{N-1} \frac{C_j}{m^{j/2}} + \frac{\overline{C}_N}{m^{N/2}} + \mathcal{E}_1 + \mathcal{E}_{3,N} \right\}, \tag{43}$$

where the expansion coefficients C_j are given in finite form:

$$C_j := \sum_{k=0}^{j} \frac{1}{16^k\ k!} \frac{1}{z^{k/2}} \sum_{u=0}^{j-k} \alpha_{j-k,u}(a, z)\, g_k(-2i(a + j - k + 2u)), \tag{44}$$

while the error term $\overline{C}_N$ is bounded as

$$\begin{aligned} \overline{C}_N = \Theta : \sum_{v=1}^{N} \frac{1}{|z|^{v/2} 16^v v!} \left(1 + u_v \sec^{v+1/2} \frac{\theta}{2}\right) \\ \sum_{u=0}^{N-v} |\alpha_{N-v,u}(a, z)\, g_v(-2i(a + N - v + 2u))| \\ + \Theta : 4(m_5/4)^{N/2} \sec^{1/2} \tfrac{\theta}{2}, \end{aligned}$$

with u_v taking the v-dependent form of (38) *in Theorem* 7 *(with $q := a$ there). Finally, the term $\mathcal{E}_1$ is subexponentially small (from Theorem* 4*), as is*

$$\mathcal{E}_{3,N} = \Theta : \sqrt{\frac{4}{\pi\sqrt{mz}}} e^{-2\sqrt{mz}} \sum_{h=0}^{N-1} \frac{1}{m^{h/2}} \sum_{u=0}^{h} |\alpha_{h,u}(a, z)| \cosh\left(\frac{\pi}{2}|a + h + 2u|\right).$$

Proof. For brevity we leave out the details—all of which are straightforward, if tedious, applications of the previous theorems and formulae. ∎

We now have a resolution of the domain of subexponential growth as follows.

COROLLARY 1. *For $(a, z) \in \mathbb{D}$, the Laguerre polynomial grows subexponentially in the sense that, for order $N \geq 1$ and any $\varepsilon > 0$,*

$$L_n^{(-a)}(-z) = S_n(a, z) \left\{ \sum_{j=0}^{N-1} \frac{C_j}{m^{j/2}} + O\left(\frac{1}{m^{N/2}}\right) + O\left(e^{-(2-\varepsilon)\sqrt{m|z|}\cos(\theta/2)}\right) \right\},$$

with all coefficients and the implied big-O constant effectively bounded via our previous theorems. Moreover, for $(a, z) \notin \mathbb{D}$, the large-n growth is not subexponential. Thus, the precise domain of subexponential growth is characterized by $(a, z) \in \mathbb{D}$.

Proof. The given subexponential formula is a paraphrase of Theorem 8. Now assume that $z \in (-\infty, 0]$. We already know that $z = 0$ does not yield such growth (see (2)). Now, for z negative real, note that the integrals in (12, 14) are both decaying in large m. Finally, the integral in (11) has phase factor $|\exp(2\sqrt{mz}\cos\omega)| \leq 1$, and Lemma 6 show that c_1 also cannot grow subexponentially in m. ∎

This corollary echoes, of course, the classical Perron result (7), and we again admit that historical efforts derived the C_j coefficients in principle. The new aspects are (a) we have asymptotic coefficients and effective bounds for general $(a, z) \in \mathbb{D}$ parameters, and (b) we can develop in a natural way an algorithm for symbolic generation of said coefficients.

6.2. Algorithm for explicit asymptotic coefficients. Theorem 8 indicates that, to obtain actual C_j coefficients, we need the cosh-arc numbers $g_k(\tau)$ and the $\alpha_{h,u}(a, z)$ coefficients. Observe that the chain of relations starting with (39) is a prescription for the generation of the C_k. All of this can proceed via symbolic processing, noting that v is simply a placeholder throughout.

Remarkably, there is a fast algorithm that bypasses much of the symbolic tedium. First, we have an explicit recursion for A_h, with $A_0 := 1$, as

$$A_k = \frac{1}{k} \sum_{j=0}^{k-1} A_j \, a_{k-j} \tag{45}$$

as follows from differentiating (41) logarithmically and then comparing terms. Second, when we use (42) together with (45), we obtain a recursion *devoid* of the symbolic placeholder v, as

$$\alpha_{k,u} = \frac{1}{k} \sum_{j=0}^{k-1} \left(\alpha_{j,u} b_{k-j} + \alpha_{j,u-1} d_{k-j} \right), \tag{46}$$

where these new recursion coefficients are

$$b_h := (-1)^h (1 + a) \left(\frac{\sqrt{z}}{2} \right)^h, \quad d_h := \left(1 - (-1)^h \right) \frac{h}{h+2} \left(\frac{\sqrt{z}}{2} \right)^{h+2}.$$

In practice we define $\alpha_{0,0} := 1$ and force any $\alpha_{j,u}$, with $u > j$ or $u < 0$ to vanish. In this sense, the collection of $\alpha_{k,u}$ make up a *lower-triangular matrix*, e.g., the entries for $k \le 3$ appear thus:

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ -\frac{(a+1)}{2} z^{1/2} & \frac{1}{12} z^{3/2} & 0 & 0 \\ \frac{a^2+3a+2}{8} z & -\frac{(a+1)}{24} z^2 & \frac{1}{288} z^3 & 0 \\ -\frac{a^3+6a^2+11a+6}{48} z^{3/2} & \frac{5a^2+15a+16}{480} z^{5/2} & -\frac{a+1}{576} z^{7/2} & \frac{1}{10368} z^{9/2} \end{pmatrix},$$

where $\alpha_{3,3}$ is the lower-right element here.

These observations lead to a fast algorithm for the computation of the asymptotic coefficients.

ALGORITHM 1 (fast computation of Laguerre asymptotic coefficients).
For given $(a, z) \in \mathbb{D}$ and desired expansion order N, this algorithm returns the asymptotic coefficients $(C_k : k \in [0, N])$ of relation (44), Theorem 8.

1. Set $\alpha_{0,0} := 1$ and, for desired order N, calculate the lower-triangular matrix elements
$(\alpha_{k,u} : 0 \le u \le k \le N)$ via a recursion such as
$\alpha(k, u)$ {

if($k == 0$) return $\delta_{0,u}$;
return $\frac{1}{k}\sum_{j=0}^{k-1}\left(\alpha_{j,u}b_{k-j} + \alpha_{j,u-1}d_{k-j}\right)$;
}
or use an unrolled, equivalent loop (i.e., one may generate the left-hand column of the α-matrix, then fill in one row at a time, lexicographically).
2. Use the recursion $g_k(\tau) = \left((2k+1)^2 + \tau^2\right) g_{k-1}(\tau)$ and the lower-triangular matrix of α values to generate the C coefficients via (44). □

We observe that the expensive sums in this algorithm are all *acyclic convolutions*. Thus, for numerical input (a, z), the algorithm complexity turns out to be $O\left(N^{2+\epsilon}\right)$ arithmetic operations, with the "2" part of the complexity power arising from the area of the lower-triangular sector.[7]

We employed the algorithm to generate exact asymptotic coefficients as follows:

$$C_0 = 1, \qquad C_1 = \frac{1}{48\sqrt{z}}\left(-12a^2 - 24za + 4z^2 - 24z + 3\right),$$
$$C_2 = \frac{1}{4608z}\left(144a^4 + 576za^3 + 480z^2a^2 + 1728za^2 - 360a^2 - 192z^3a \right.$$
$$\left. + 1152z^2a + 1584za + 16z^4 - 192z^3 + 312z^2 + 432z + 81\right),$$

and so on. Note that the C_1 form here agrees with the PAMO coefficient in (8). We were able to generate the full, symbolic $C_{64}(a, z)$ in about one minute of CPU on a typical desktop computer. To aid future researchers, we report that the numerator of C_{64} has degree 128 in both a, z, while the denominator is an integer divisible by every prime not exceeding 64, namely,

$$2^{319} \cdot 3^{94} \cdot 5^{25} \cdot 7^{13} \cdot 11^7 \cdot 13^5 \cdot 17^4 \cdot 19^3 \cdot 23^3 \cdot 29^2 \cdot 31^2 \cdot 37 \cdot 41 \cdot 43 \cdot 47 \cdot 53 \cdot 59 \cdot 61.$$

While formidable, the bivariate $C_k(a, z)$ coefficients evidently have intricate structure and patterning. Future research into said structure would be of interest.

6.3. Generating and verifying effective bounds. For the first nontrivial effective bound, we can employ the rigorous bound $\overline{C}_1$ in Theorem 8, with $N = 1$, to get an effective version of the original asymptotic (3) as

$$L_n^{(-a)}(-z) = S_n(a, z)\left(1 + \frac{\overline{C}_1}{\sqrt{m}} + \ \mathcal{E}_1 + \mathcal{E}_{3,1}\right),$$

with

$$\overline{C}_1 = \Theta : \frac{\left|1 - 4a^2\right|}{16|z|^{1/2}}\left(1 + 6\sec^{3/2}\frac{\theta}{2}\right) + 2|z|^{1/2}(1 + |a| + |z|/2)\sec^{1/2}\frac{\theta}{2},$$

and we remind ourselves that $\mathcal{E}_{3,1}$ (Theorem 8) and $\mathcal{E}_1$ (Theorem 4) are both subexponentially small.

At last we have an effective numerator, then, for the $1/\sqrt{m}$ asymptotic term. Though this effective numerator is almost surely nonoptimal, we are evidently on the right track, because the exact C_1 asymptotic coefficient above (see (8)) has very much the same form as does the Θ-expression for $\overline{C}_1$ here (i.e., the same degrees

[7]For example, floating-point FFT-based convolutions of length L require $O(L \log L)$ operations, of complexity less than $O\left(L^{1+\epsilon}\right)$.

TABLE 1

Examples of rigorous accuracy versus asymptotic order. For the indices n and parameters $(a, z) \in \mathbb{D}$, numbers $\mathcal{L}_{n,N}$ are computed via the asymptotic series through term $C_{N-1}/(n+1)^{(N-1)/2}$ as in Theorem 8. The relative error (RE) and rigorous upper bound (RRE) on the relative errors in taking these terms is reported. All decimal values are reported as correct to ± 1 in the last displayed digit of $\Re, \Im$ parts. Such rigorous bounds can be used to establish indisputable accuracies for various other functions.

n	a, z	N	Approximation $\mathcal{L}_{n,N}$ to $L_n^{(-a)}(-z)$	RE	RRE
1300	$0, -1 + 4i$	1	$(7 - i) \cdot 10^{37}$	0.0038	0.43
10^7	$0, 1$	1	$1.59 \cdot 10^{2744}$	0.00011	0.00099
		2	$1.595190 \cdot 10^{2744}$	$1.4 \cdot 10^{-9}$	$2.8 \cdot 10^{-7}$
		3	$1.5951907764 \cdot 10^{2744}$	$5.4 \cdot 10^{-13}$	$9.2 \cdot 10^{-11}$
10^6	$\frac{1}{2} + i, \frac{1}{2} + 4i$	1	$(-1.5 - 2.0i) \cdot 10^{1303}$	0.0016	0.021
		2	$(-1.565 - 2.041i) \cdot 10^{1303}$	$1.6 \cdot 10^{-6}$	$9.8 \cdot 10^{-5}$
		3	$(-1.562555 - 2.041423i) \cdot 10^{1303}$	$1.4 \cdot 10^{-9}$	$4.2 \cdot 10^{-7}$
		4	$(-1.56255567 - 2.041423269i) \cdot 10^{1303}$	$1.3 \cdot 10^{-12}$	$1.4 \cdot 10^{-9}$
		5	$(-1.56255567881 - 2.04142326969i) \cdot 10^{1303}$	$1.3 \cdot 10^{-15}$	$5.9 \cdot 10^{-12}$
10^7	$-\frac{3}{2}, 1 - 100i$	1	$(2 + i) \cdot 10^{19520}$	0.027	0.41
		2	$(1.8 + 0.9i) \cdot 10^{19520}$	0.00035	0.034
		3	$(-1.85 + 0.95i) \cdot 10^{19520}$	$3.1 \cdot 10^{-6}$	0.0028
		4	$(-1.851 + 0.957i) \cdot 10^{19520}$	$2.1 \cdot 10^{-8}$	0.00023
		5	$(-1.8510 + 0.9575i) \cdot 10^{19520}$	$1.1 \cdot 10^{-10}$	$2.0 \cdot 10^{-5}$
		10	$(-1.8509993203 + 0.9575616603i) \cdot 10^{19520}$	$7.1 \cdot 10^{-23}$	$7.6 \cdot 10^{-11}$

of appearance for a, z, and similar coefficients). And, it is easy to see that the Θ-expression here is an upper bound on $|C_1|$ itself, as must of course be true; it is necessary that $\overline{C}_N \geq |C_N|$ for every asymptotic order N.

Table 1 shows the numerical instances of our asymptotic series from Theorem 8. It displays various indices n and parameters a, z, together with corresponding asymptotic orders N from the theorem, using the following nomenclature. Denote, for asymptotic order N,

$$T_{n,N} \;:=\; 1 + \frac{C_1}{m^{1/2}} + \cdots + \frac{C_{N-1}}{m^{(N-1)/2}},$$

so, for S_n being the assignment (4), we have an approximation $\mathcal{L}_{n,N}$ defined by

$$\mathcal{L}_{n,N} \;:=\; S_n T_{n,M} \;\approx\; L_n^{(-a)}(-z).$$

Thus, $\mathcal{L}_{n,1}$ is simply S_n, as $T_{n,1} = 1$ always. How good are these $\mathcal{L}_{n,N}$ approximations? The important right-most columns of Table 1 report relative error (call it RE), with

$$\text{RE} := \left| \frac{\mathcal{L}_{n,N} - L_n^{(-a)}(-z)}{L_n^{(-a)}(-z)} \right|,$$

and a *rigorous* upper bound on relative error (call it RRE), expressed in terms of Theorem 8's various error bounds, with

$$\text{RRE} := \left| \frac{\epsilon}{T_{n,N}} \right|,$$

where $\epsilon > 0$ can be taken to be any number not exceeding $\left|\overline{C}_N/m^{N/2} + \mathcal{E}_1 + \mathcal{E}_{3,N}\right|$. Note that the three error components in this last expression are all magnitude-bounded via Theorem 8; this is how the RRE table entry was constructed.

A convenient way to view RRE is to realize that $L_n^{(-a)}(-z)$ is rigorously known to lie in the interval $\mathcal{L}_{n,N} \cdot (1 - \mathrm{RRE}, 1 + \mathrm{RRE})$. The decimal representations in Table 1 are reflexive; i.e., there is ambiguity only in the respective final digits. (When a reported value is $(a + ib) \cdot 10^c$, then both a and b are correct to within ± 1 in their respective last digits.)

One might ask about the "first-missing" asymptotic term $C_N/m^{N/2}$—a term of interest in many asymptotic theories. This term is typically close to, but often greater than, the RE. As is common to many asymptotic developments, rules about the behavior of these first-missing terms are hard to establish when parameters (in our case a, z) roam over the complex plane. In any case, the first-missing term is often an order of magnitude or more greater than the RRE. Yet this discrepancy between first-missing and RRE terms is not as harmful as might appear, as we discuss next.

6.4. Example application of effective bounds: Rigor for incomplete gamma. We have mentioned research motives in section 1.1. A worked example of *rigorous* incomplete gamma evaluations is the following. Let us use our very first entry from Table 1, namely,

$$L_{1300}^{(0)}(1 - 4i) \;\approx\; (7 - i) \cdot 10^{37},$$

with a rather large RRE = 0.43. Though this is an order of magnitude larger than the relative error, the penalty paid is only a few digits of precision for the relevant incomplete gamma. Indeed, from (5),

$$\Gamma(0, -1 + 4i) = e^{1-4i} \sum_{n=0}^{1299} \frac{1}{n+1} \frac{1}{L_n^{(0)}(1 - 4i) L_{n+1}^{(0)}(1 - 4i)} \;+\; \delta,$$

where the error δ is bounded, using RRE = 0.43 and (4), by

$$|\delta| \;<\; \frac{\left|e^{1-4i}\right|}{(1 - \mathrm{RE})^2} \sum_{n \geq 1300} \frac{1}{n+1} \frac{1}{|S_n S_{n+1}|} \;<\; 10^{-76}, \tag{47}$$

the last bound using (4) and careful estimates on the summation over $n \geq 1300$. Thence the sum for $n \in [1, 1299]$ gives a rigorous, 76–digit-accurate value for $\Gamma(0, -1 + 4i)$, $(\approx 0.497 \,+\, 0.415\, i)$. Correspondingly, the actual absolute error in using the sum over $n \in [0, 1299]$ is about 10^{-78}, so the apparently poor RRE bound brings only a 2-digit penalty.

An additional application of Theorem 8 would be to use effective bounds to rule out zeros of L_n in the crossed (a, z)-plane, that is, to establish an m_0 such that the term $1 + \overline{C}_1/\sqrt{m}$ must be positive for all $m > m_0$ (see our Open problems section).

7. Brief remarks on oscillatory regimes. The exp-arc method has led to rigorous asymptotics for subexponential growth but not for the Fejér form (6) for $z \neq 0$ on the cut $(-\infty, 0]$. Such oscillatory behavior can be dealt with, but other techniques come into play. For one thing, contour integrals must be handled differently.

7.1. When z is on the negative real cut. Even on $z \in (-\infty, 0)$ the contour prescription of Figure 1 is valid, and the Laguerre polynomial is exactly the sum

$c_1+d_1+e_1$, with $R := \sqrt{m/|z|} > 1/2$ being the only requirement for contour validity. However—and this is important—the dominant contribution (17) has to change to involve an expanded integration interval; in fact, now we must use the contour integral c_1 itself as the main contribution. For reasons of brevity, we simply state the complete asymptotic result stemming from contour integral c_1 as

$$L_n^{(-a)}(-z) \;\sim\; \frac{e^{-z/2}}{\sqrt{\pi}(-z)^{1/4-a/2}\; m^{1/4+a/2}} \times \left\{ \left(\sum_{k=0}^{\infty} \frac{A_k}{m^k} \right) \cos\left(2\sqrt{-mz} + a\pi/2 - \pi/4\right) + \left(\sum_{k=0}^{\infty} \frac{B_k}{m^{k+1/2}} \right) \sin\left(2\sqrt{-mz} + a\pi/2 - \pi/4\right) \right\}, \tag{48}$$

where these oscillatory-series coefficients are directly related to the coefficients in Theorem 8 by

$$A_k := C_{2k}(a,z), \quad B_k := iC_{2k+1}(a,z).$$

It transpires that, for a real, every A_k, B_k is real, whence the asymptotic has all real terms. The Fejér–Perron–Szegö expansion in [36, Theorem 8.22.2] for the oscillatory Laguerre mode is stated there in a fashion structurally different from our asymptotic (48); notwithstanding this Szegö's own $A_{\text{odd}}, B_{\text{even}}$—also not defined quite like ours—vanish.[8]

7.2. Brief remarks on Bessel functions I_n, J_n. There is vast literature on Bessel asymptotics [44], [39], resulting in the *Hankel asymptotic series* [1, section 9.2.5] and effective bounds on error terms (for certain parameter domains). In many cases these restricted error bounds are nevertheless optimal [44]. It is instructive to explore, at least briefly, the application of our exp-arc method to Bessel expansions. We recall (29) which leads to

$$I_n(z) = \frac{2}{\pi} \sum_{k\geq 0} \frac{g_k(-2in)}{(2k)!} \left\{ e^z B_k(z) + (-1)^n e^{-z} B_k(-z) \right\} \;\sim\; \frac{e^z}{\sqrt{2\pi z}} \sum_{k=0}^{\infty} \frac{g_k(-2in)}{k!\; 8^k} \frac{1}{z^k}, \tag{49}$$

(again we omit proofs) where the first sum is convergent, exact, and the second sum agrees with the classical Hankel asymptotic [44], [1]. For $\arg(z) \in (-\pi/2, \pi/2)$, at least, one may derive effective bounds on I_n error terms using the exp-arc techniques.

Computationalists have known for decades that one way to evaluate Bessel functions uniformly in the argument z is to use the standard ascending series for small $|z|$, but an asymptotic series for large $|z|$. However, via the exp-arc method one can establish a converging series whose evaluation only involves *a single error-function evaluation*, followed by recursion and elementary algebra. In fact, the relations (34), (35) can be used to calculate $J_n(z)$ from relation (28) via the sum

$$J_n(z) = \frac{2}{\pi} \sum_{k=0}^{\infty} g_k(-2in) \left(b_k \cos\chi - c_k \sin\chi \right), \tag{50}$$

[8] Our oscillatory asymptotic (48) has been verified to several terms by Temme; also he verifies our claim that the classical Szegö coefficients do vanish for the parities indicated [38].

with

$$\chi := z - \pi n/2 - \pi/4, \quad b_k := B_k(iz)e^{i\pi/4} + B_k(-iz)e^{-i\pi/4}, \quad ic_k := B_k(iz)e^{i\pi/4} - B_k(-iz)e^{-i\pi/4}.$$

Note that if z is real, then each b_k, c_k is real, whence our series here has all real terms. Note that our recursion (35) likewise ignites a recursion amongst the b_k, c_k.

Note that (50) is actually the Hankel asymptotic if we replace B_k by its first term in (34), namely, $(1/2)(2iz)^{-k-1/2}\Gamma(k+1/2)$; however, we already know that the sum (50) is always convergent. It is remarkable that we are using the same structure as the classical asymptotic, yet convergence for all complex z is guaranteed. Moreover, the $B_k(iz)$ are independent of the order n and so can be reused if multiple $J_n(z)$ are desired for fixed z.

Those acquainted with the intricacies of Bessel theory may observe that our convergent expansion (49) is at least reminiscent of the convergent *Hadamard expansion* found in [44, p. 204] for the modified Bessel function I_ν. Though both expansions are absolutely convergent, there are some important differences between this Hadamard expansion and our exp-arc forms (49, 50). For example, we have given our convergent sum only for integer ν. Moreover, the exp-arc expansion is geometrically convergent, while the Hadamard expansion is genuinely slower.

The research area of convergent expansions related to classical, asymptotic ones has been pioneered in large part by Paris, whose works cover real and complex domains, saddle points, and the like [29], [30], [31], [32], [33], [34]. We should point out that Paris was able to develop within the last decade some similar, linearly convergent Bessel series by modifying the "tails" of Hadamard-class series. Finally,we note that the problem of generalizing such unconditionally convergent series as our (50) to noninteger indices—and with comparison to the recent work of Paris—is analyzed in a new work [3].

8. Open problems.

- How might one proceed with the exp-arc theory to obtain effective error bounds for oscillatory Laguerre modes and-or oscillatory Bessel modes? We know that previous researchers have described how to give effective bounds in these cases (e.g., to our asymptotic (48), as in [36, section 8.72]), but once again we stress, How can this be done explicitly and for full parameter ranges?
- Where are the zeros in the complex z-plane—for fixed a—of $L_n^{(-a)}(-z)$? Are "most of" the zeros along some a-dependent ray, in some sense? There is a considerable literature on this zero–free-region topic, especially for polynomials in real variables. For example, with $a := 0$ the Laguerre zeros are all real and negative; see [24, Chapter X] and references therein. There is also an interesting connection between Laguerre zeros and eigenvalues of certain (large) matrices [15]. Certainly the theorems of the Saff–Varga type are relevant in this context [24, Theorem 11].
- How can Laguerre asymptotics be gleaned from standard recurrence relations amongst the $L_n^{(-a)}(-z)$? One may ask the same question for the Laguerre differential equation as starting point. A promising research avenue for a discrete-iterative approach to asymptotics is [43]; see also [4] for the asymptotic analysis of certain complex continued fractions. As for differential-equation approaches, there is the classical work of Erdélyi and Olver, plus modern work on the combinations of differential, discrete, and saddle-point theory [16], [18].

- The efficient "keyhole" contour of Figure 1 was discovered experimentally. What other analytical problems might be approached in this fashion? For that matter, how might one properly use the celebrated Watson loop-integral lemma with error term [27, Theorem 5.1] on our keyhole contour to obtain similar effective asymptotics?
- Here is a scenario suggested by Temme [38]. When n is small (in some sense not made rigorous here), and, say, $|z| \ll |a|$, one should expect an asymptotic behavior

$$L_n^{(-a)}(-z) \sim \binom{n-a}{n} \left(1 + \frac{z}{1-a}\right)^n.$$

The problem is, How does one make this rigorous, with effective error bounds?

Acknowledgments. We are indebted to J. Buhler, R. Mayer, O-Yeat Chan, R. Paris, and N. Temme for the generous lending of expertise, ideas, and suggestions. We are equally grateful to two referees for their insightful remarks.

REFERENCES

[1] M. Ambramowitz and I. Stegun, *Handbook of Mathematical Functions*, NBS (now NIST), Washington, D.C., 1965.
[2] D. H. Bailey, J. M. Borwein, and R. E. Crandall, *Integrals of the Ising class*, J. Phys. A, 39 (2006), pp. 12271–12302.
[3] D. Borwein, J. Borwein, and O-Y. Chan, *The evaluation of Bessel functions via exp-arc integrals*, J. Math. Anal. Appl., 341 (2008), pp. 478–500.
[4] D. Borwein, J. Borwein, R. Crandall, and R. Mayer, *On the dynamics of certain recurrence relations*, Ramanujan J., 13 (2007), pp. 65–104.
[5] J. M. Borwein and P. B. Borwein, *Pi and the AGM: A Study in Analytic Number Theory and Computational Complexity*, John Wiley, New York, 1987. Paperback edition 1998.
[6] J. Borwein and M. Chamberland, *Integer powers of Arcsin*, Int. J. Math. Math. Sci., 2007, Article ID 19381, http://www.hindawi.com/GetArticle/aspx?doi=10.1155/2007/19381.
[7] J. Borwein, R. Crandall, and G. Fee, *On the Ramanujan AGM fraction. Part* I: *The real-parameter case*, Experiment. Math., 13 (2004), pp. 275–286.
[8] J. Borwein and R. Crandall, *On the Ramanujan AGM fraction. Part* II: *The complex-parameter Case*, Experiment. Math., 13 (2004), pp. 287–296.
[9] C. Bosbach and W. Gawronski, *Strong asymptotics for Laguerre polynomials with varying weights*, J. Comput. Appl. Math., 99 (1998), pp. 77–89.
[10] K. Chao and R. Plymen, *A new bound for the smallest x with $\pi(x) > \mathrm{li}(x)$*, arXiv:math/0509312v5.
[11] R. Crandall, *Unified algorithms for polylogarithm, L-series, and zeta variants*, http://www.reed.edu/∼crandall, 2008, preprint.
[12] R. Crandall, *Alternatives to the Riemann–Siegel formula*, http://www.reed.edu/∼crandall, 2008, preprint.
[13] R. Crandall, R. Bettega, and R. Whitnell, *A class of exactly soluble three-body problems*, J. Chem. Phys., 83 (1985), pp. 698–702.
[14] R. Crandall, R. Whitnell, and R. Bettega, *Exactly soluble two-electron atomic model*, Amer. J. Phys., 52 (1984), pp. 438–442.
[15] H. Dette, *Strong approximation of eigenvalues of large dimensional Wishart matrices by roots of generalized Laguerre polynomials*, J. Approx. Theory, 118 (2002), pp. 290–304.
[16] A. Erdélyi, *Asymptotic forms for Laguerre polynomials*, J. Indian Math. Soc., Golden Jubilee Commemoration Volume, 24 (1960), pp. 235–250.
[17] W. Foster and I. Krasikov, *Explicit bounds for Hermite polynomials in the oscillatory region*, LMS J. Comput. Math., 3 (2000), pp. 307–314.
[18] C. L. Frenzen and R. Wong, *Uniform asymptotic expansions of Laguerre polynomials*, SIAM J. Math. Anal., 19 (1988), pp. 1232–1248.
[19] P. Henrici, *Applied and Computational Complex Analysis: Special Functions, Integral Transforms, Asymptotics, Continued Fractions, Vol.* 2 *(Classics Library)*, John Wiley, New York, 1974 (reprint 1988).

[20] E. Hobson, *A Treatise on Plane Trigonometry*, Cambridge University Press, London, 1925.
[21] L. Jacobsen and W. Thron, *Oval convergence regions for continued fractions* $K(a_n/1)$, preprint, 1986.
[22] A. Kuijlaars and K. McLaughlin, *Riemann–Hilbert analysis for Laguerre polynomials with large negative parameter*, Comput. Methods Func. Theory, 1 (2001), pp. 205–233.
[23] J. López and N. Temme, *Approximations of orthogonal polynomials in terms of Hermite polynomials*, Methods Appl. Anal., 6 (1999), pp. 131–146.
[24] L. Lorentzen and H. Waadeland, *Continued Fractions with Applications*, North-Holland, Amsterdam, 1992.
[25] E. Love, *Inequalities for Laguerre functions*, J. Inequal. Appl., 1 (1997), pp. 293–299.
[26] F. Olver, *Why steepest descents?* SIAM Rev., 12 (1970), pp. 228–247.
[27] F. Olver, *Asymptotics and Special Functions*, Academic Press, New York, 1974.
[28] R. B. Paris, *private communication*, 2007.
[29] R. B. Paris, *On the use of Hadamard expansions in hyperasymptotic evaluation of Laplace-type integrals.* I: *Real variable*, J. Comput. Appl. Math., 167 (2004), pp. 293–319.
[30] R. B. Paris, *On the use of Hadamard expansions in hyperasymptotic evaluation of Laplace-type integrals.* II: *Complex variable*, J. Comput. Appl. Math., 167 (2004), pp. 321–343.
[31] R. B. Paris, *On the use of Hadamard expansions in hyperasymptotic evaluation of Laplace-type integrals.* III: *Clusters of saddles*, J. Comput. Appl. Math., to appear.
[32] R. B. Paris, *On the use of Hadamard expansions in hyperasymptotic evaluation of Laplace-type integrals.* IV: *Poles*, Comput. Appl. Math., to appear.
[33] R. B. Paris, *Exactification of the method of steepest descents: The Bessel functions of large order and argument*, Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 460 (2004), pp. 2737–2759.
[34] R. B. Paris and D. Kaminski, *Hadamard expansions for integrals with saddles coalescing with an endpoint*, Appl. Math. Lett., 18 (2005), pp. 1389–1395.
[35] O. Perron, *Über das Verhalten einer ausgearteten hypergeometrischen Reihe bei unbegrenztem Wachstum eines Parameters*, Journal für die reine und angewandte Math, 151 (1921), pp. 63–78.
[36] G. Szegö, *Orthogonal Polynomials*, Amer. Math. Soc. Colloq. Publ., 4th ed., 23, American Mathematical Society, Providence, RI, 1975.
[37] N. Temme, *Asymptotic estimates of Stirling numbers*, Stud. Appl. Math., 89 (1993), pp. 223–243.
[38] N. Temme, *private communication*, 2006.
[39] S. Thangavelu, *Lectures on Hermite and Laguerre Expansions*, Math. Notes 42, Princeton University Press, Princeton, N.J., 1993.
[40] E. Titchmarsh, *The Theory of the Riemann Zeta Function*, Clarendon Press, Oxford, 1951 & 1967.
[41] W. Van Assche, *Weighted zero distribution for polynomials orthogonal on an infinite interval*, SIAM J. Math. Anal., 16 (1985), pp. 1317–1334.
[42] W. Van Assche, *Erratum to "Weighted zero distribution for polynomials orthogonal on an infinite interval"*, SIAM J. Math. Anal., 32 (2001), pp. 1169–1170.
[43] Z. Wang and R. Wong, *Linear difference equations with transition points*, Liu Bie Ju Center for Math. Sci., preprint, 2004.
[44] G. Watson, *Treatise on the Theory of Bessel Functions*, 2nd ed., Cambridge Math. Lib. Ed., Cambridge University Press, Cambridge, 1966.
[45] S. Winitzki, *Computing the incomplete gamma function to arbitrary precision*, in Lecture Notes in Comput. Sci. 2667, Springer–Verlag, New York, 2003, pp. 790–798.
[46] S. Wong, *Computational Methods in Physics and Engineering*, Prentice–Hall, New Jersey, 1992.